UNIVERSIDAD DEL VALLE DE GUATEMALA

Facultad de Ciencias y Humanidades

El uso de técnicas de Estimación en Áreas Pequeñas para el cálculo de indicadores de condiciones de vida en Guatemala

Trabajo de graduación presentado por
Irene Molina Manrique
para optar al grado académico de
Licenciada en Matemáticas

Guatemala
2017

El uso de técnicas de Estimación en Áreas Pequeñas para el cálculo de indicadores de condiciones de vida en Guatemala

UNIVERSIDAD DEL VALLE DE GUATEMALA

Facultad de Ciencias y Humanidades

El uso de técnicas de Estimación en Áreas Pequeñas para el cálculo de indicadores de condiciones de vida en Guatemala

Trabajo de graduación presentado por
Irene Molina Manrique
para optar al grado académico de
Licenciada en Matemáticas

Guatemala
2017

Vo.Bo.

Lic. Dorval Carías

Tribunal examinador:

Lic. Dorval Carías

MA. Nancy Zurita

Lic. Annelisse Balsells

Fecha de aprobación: Guatemala, 24 de noviembre de 2017

PREFACIO

El propósito de este trabajo de graduación es dar una introducción a las técnicas de Estimación en Áreas Pequeñas y presentar un ejemplo de su aplicación en el cálculo de indicadores de condiciones de vida en Guatemala. Actualmente, en Guatemala, las técnicas de Estimación en Áreas Pequeñas no se han estado aplicando, pero el Ministerio de Salud, la SESAN (Secretaría Nacional de Seguridad Alimentaria y Nutrición) y otras organizaciones están muy interesados en tener indicadores de salud a nivel departamental y municipal para la toma de decisiones, distribución de fondos y elaboración de programas. Es por esto que este trabajo debería ser de interés para las personas trabajando en encuestas nacionales, así como técnicos y estudiantes interesados en aprender sobre el tema.

ÍNDICE GENERAL

RESUMEN

Este trabajo de graduación presenta una breve introducción a la teoría general de muestreo y algunas de las técnicas de Estimación en Áreas Pequeñas como: los estimadores directos, indirectos y los modelos básicos a nivel de área y unidad. El informe de la Encuesta de Salud Materno Infantil 2002 presentó indicadores de desnutrición crónica solamente a nivel regional en Guatemala, por lo que este trabajo de graduación también tuvo como objetivo calcular los indicadores a nivel departamental. El trabajo se enfocó específicamente en el modelo a nivel de área Fay-Herriot, con el cual, utilizando indicadores obtenidos del Censo Nacional XI de Población y de VI de Habitación 2002, se obtuvo indicadores departamentales de desnutrición crónica en niños entre 3 y 59 meses de edad. La metodología para el cálculo de estos indicadores se presenta de manera detallada junto con los indicadores obtenidos utilizando estimación directa y el modelo de Fay-Herriot. A partir de los resultados, se puede concluir que los métodos de Estimación en Áreas Pequeñas permiten obtener indicadores con errores estándar menores al hacer uso de información auxiliar y del modelo de regresión en donde se utiliza la información de todos los departamentos para el cálculo de los indicadores departamentales.

CAPÍTULO I: INTRODUCCIÓN

Usualmente, en el análisis de poblaciones, se definen subpoblaciones llamadas **dominios de estudio**. Estos dominios pueden definirse como áreas geográficas, grupos socio-demográficos u otras subpoblaciones. En Guatemala, los dominios geográficos más considerados son: regiones, departamentos y municipios; y los dominios socio-demográficos importantes son el área rural y urbana.

Un dominio es considerado ”grande” si la muestra del dominio es suficientemente grande para obtener estimaciones directas con la precisión adecuada. El dominio es considerado un **área pequeña** si la muestra del dominio no es lo suficientemente grande para dar estimaciones directas con la precisión necesaria. [Rao, 2003] La Encuesta de Salud Materno Infantil de 2002 en Guatemala comprendió 12,119 hogares a nivel nacional, pero solamente 93 en el departamento de Zacapa; esto hace que el departamento de Zacapa sea considerado un ”área pequeña”.

En años anteriores, se han presentado indicadores a nivel nacional, regional y de área rural y urbana. Esto se debe a que la realización de las encuestas de condiciones de vida es una obligación del país con organismos internacionales como la Organización Mundial de la Salud (OMS), el Banco Mundial y el Programa de las Naciones Unidas para el Desarrollo (PNUD). Estas organizaciones solicitan indicadores a nivel nacional y de los principales dominios debido a su interés de comparar los indicadores de diferentes países. En el informe de la Encuesta Nacional de Salud Materno Infantil (ENSMI) 2002 [MSPAS, 2003] solamente se presentan indicadores hasta el nivel regional y los departamentos deben considerarse como áreas pequeñas.

La ENSMI 2014 es la última encuesta de este tipo y en su informe se presentan los indicadores de desnutrición crónica a nivel departamental. En esta ENSMI, 291 municipios fueron visitados, dejando más de 40 municipios sin visitar. En 89 municipios solamente se encuestó 1 sector censal (26 hogares), en 70 municipios se encuestaron 2 sectores censales (52 hogares) y en el resto de municipios se encuestaron más de 2 sectores censales. Es por esto que en mucho municipios es

imposible obtener una estimación directa, ya que no fueron visitados y en los otros, el número de hogares es muy pequeño para obtener estimaciones directas con una precisión adecuada. Ha sido hasta encuestas más recientes donde se han comenzado a calcular indicadores a nivel departamental, y ahora se tienen que empezar a calcular a nivel municipal.

La estimación en áreas pequeñas ha recibido mucha atención en los últimos años debido a la creciente demanda de indicadores confiables para áreas pequeñas. En Guatemala, indicadores a nivel departamental o municipal son necesarios para entidades como el Ministerio de Salud Pública y el Ministerio de Educación. Este tipo de indicadores pueden ayudar en la toma de decisiones, distribución de fondos y elaboración de programas de salud o educación.

En los próximos capítulos se presentan algunas de las técnicas de estimación en áreas pequeñas que se utilizan para el cálculo de indicadores en áreas pequeñas y la metodología para obtener estimaciones directas e indirectas de indicadores en áreas pequeñas utilizando la base de datos de la Encuesta Nacional de Salud Materno Infantil (ENSMI) de 2002 en Guatemala. Se escogió específicamente la ENSMI 2002 debido a que se tienen los indicadores presentados en el informe del Censo Nacional XI de Población y de VI de Habitación 2002 como variables auxiliares. Este trabajo pretende mostrar la importancia de los censos en Guatemala y sus aplicaciones para la estimación en áreas pequeñas.

Los métodos presentados para la estimación en áreas pequeñas son los siguientes:

- Métodos directos
    - Estimador de razón
    - Estimador de regresión
- Métodos indirectos
    - Estimadores indirectos tradicionales
        - Estimadores sintéticos
        - Estimador compuesto
    - Modelos de áreas pequeñas
        - Modelo a nivel de unidad
        - Modelo a nivel de área

CAPÍTULO II: TEORÍA GENERAL DE MUESTREO

A continuación se definen algunos conceptos de teoría de muestreo que presenta Sharon Lohr [Lohr, 1999].

- **Unidad de observación:** el objeto del cual se toma una medida; también es llamado un elemento. En el caso de una encuesta sobre política, las unidades de observación son individuos.
- **Población objetivo:** la colección completa de observaciones que se quiere estudiar. La población objetivo puede ser todas las personas que tienen derecho de votar, las personas empadronadas, o las personas que votaron en las elecciones pasadas.
- **Muestra:** un subconjunto de una población.
- **Población de estudio o población accesible:** es determinada por el marco de muestreo. Es la población de la cual se posee información administrativa o censal que se utiliza para la elaboración del marco de muestreo.
- **Unidad muestral o unidad de muestreo:** la unidad que se utiliza para hacer la muestra; por ejemplo, las **unidades primarias de muestreo** (UPM). Si se quieren estudiar individuos, pero no se tienen una lista de todos los individuos en la población objetivo, se pueden utilizar los sectores censales como las unidades primarias de muestreo y las unidades de observación serían los individuos que viven en cada segmento.
- **Marco de muestreo o marco muestral:** el listado de todas las unidades muestrales.

Una muestra se considera representativa si cada unidad muestreada representa las características de cierto número de unidades en la población. Esto quiere decir que una buena muestra va a reproducir las características de interés de la población lo más cercano posible. Se puede dar el caso en el que las unidades de observación y las de muestreo tengan diferentes unidades de estudio. Por ejemplo, uno de los objetivos principales de la ENSMI 2002 fue proporcionar información acerca de la salud

materno infantil y la salud reproductiva, por lo que la **población objetivo** fue todas las mujeres en edad reproductiva (15 a 49 años) en cada hogar, y las **unidades de observación** fueron las mujeres. En Guatemala, las encuesta utilizan principalmente la información de los censos, por lo que, en el caso de las ENSMI de 1995, 1998/1999 y 2002, las **unidades primaria de muestreo** fueron los sectores o segmentos censales. Puede ser que en algunos de los hogares del **marco muestral** no vivían mujeres en edad reproductiva, por lo que esos hogares no formaron parte de la **muestra**.

A. Las muestras probabilísticas

Para tomar una muestra representativa es conveniente utilizar un método de selección aleatoria que permita registrar la probabilidad de selección de cada unidad. Estas probabilidades de selección son luego utilizadas para el cálculo de los pesos de las unidades en la muestra. En una **muestra probabilística** cada unidad en la población tiene una probabilidad conocida de selección y un método aleatorio de selección es utilizado para seleccionar las unidades que van a ser incluidas en la muestra. Lohr [Lohr, 1999] describe los siguientes tipos de muestras probabilísticas:

- **Muestreo aleatorio simple:** es la forma más simple de muestra probabilística. Una muestra aleatoria simple de tamaño n se toma cuando cada subconjunto posible de n unidades en la población tiene la misma probabilidad de ser la muestra. En este caso, cada miembro de la población tiene la misma probabilidad de ser incluido en la muestra; en otros tipos de muestreo, las probabilidades de selección pueden ser diferentes.

- **Muestreo aleatorio estratificado:** en este caso la población se divide en subgrupos llamados **estratos**. Luego, una muestra aleatoria simple se selecciona en cada estrato. Los estratos usualmente son subgrupos de interés para el investigador. Por ejemplo, los estratos pueden ser diferentes grupos étnicos o tipos de terreno en un estudio ecológico. Los elementos en un mismo estrato tienden a ser más similares que elementos seleccionados aleatoriamente en toda la población, lo que puede aumentar la precisión. En las encuestas nacionales, los departamentos usualmente se consideran como estratos.

- **Muestreo por conglomerados:** en este tipo de muestreo, las unidades de observación en la población son agregadas a unidades de muestreo más grandes llamadas **conglomerados**. Por ejemplo, en una encuesta de estudiantes de primer grado para evaluar su rendimiento en lectura, se toma una muestra de las escuelas en Guatemala. En cada escuela seleccionada

se elige aleatoriamente una sección de primer grado (si hay 2 o más secciones en el grado) y luego se evalúan a todos los estudiantes de la sección seleccionada. En este caso, las escuelas son los conglomerados o las **unidades primarias de muestreo**, las secciones son las **unidades secundarias de muestreo** y los estudiantes son las **unidades de observación**. Un problema que puede surgir es que estudiantes de la misma escuela pueden poseer más características en común que estudiantes seleccionados aleatoriamente; por lo que una muestra por conglomerados de 500 estudiantes no puede proveer suficiente información como una muestra aleatoria simple de 500 estudiantes. El muestreo por conglomerados usualmente se utiliza debido a que resulta menos costoso y no toma mucho tiempo. En las encuestas nacionales de hogares, usualmente se toma en cuenta el uso de estratos y conglomerados.

B. Los estimadores y sus varianzas

Cuando se toma una muestra para una encuesta, las estadísticas obtenidas a partir de la muestra son una estimación del verdadero valor del parámetro poblacional, por lo que no es suficiente reportar solamente los indicadores obtenidos, también se debe reportar una medida que represente la exactitud de estos indicadores. En estos casos se utilizan medidas como la varianza, erorr estándar, el error cuadrático medio y los intervalos de confianza.

Supongamos que queremos calcular el total poblacional $t = \sum_{i=1}^{N} y_i$ utilizando una muestra, donde N es el total de unidades en la población. Una estimación de t es $\hat{t}_S = N\overline{y}_S$, donde $\overline{y}_S$ es el promedio de los valores y_i en la muestra S. La **distribución muestral** de $\hat{t}$, tomando en cuenta las probabilidades de selección para las muestras está dada por

$$P\{\hat{t} = k\} = \sum_{S:\hat{t}_S=k} P(S),$$

donde la suma es sobre todas las muestras S para las cuales $\hat{t}_S = k$. La distribución muestral presenta la frecuencia de todos los posibles resultados que, en este caso, $\hat{t}$ puede tomar en cada posible muestra de la población.

El **valor esperado** de $\hat{t}$, $E[\hat{t}]$, es la media de la distribución muestral de $\hat{t}$.

$$E[\hat{t}] = \sum_S P(S)\hat{t}_S = \sum_k kP(\hat{t} = k).$$

Esto quiere decir que el valor esperado de $\hat{t}$ es el promedio ponderado de todos los valores posibles de $\hat{t}$ para cada muestra, con la probabilidad que ocurra ese valor específico como el peso.

1. Sesgo y varianza. El **sesgo de estimación** del estimador $\hat{t}$ se determina de la siguiente manera

$$Sesgo[\hat{t}] = E[\hat{t}] - t.$$

El estimador $\hat{t}$ es **insesgado** si $Sesgo[\hat{t}] = 0$; pero si se utiliza $\hat{t}_S = \sum_{i \in S} y_i$ y no se realizó un censo, el estimador $\hat{t}$ va a estar sesgado.

La **varianza** de la distribución muestral de $\hat{t}$ está dada por

$$V[\hat{t}] = E[(\hat{t} - E[\hat{t}])^2]$$

$$= E[\hat{t^2} - 2\hat{t}E[\hat{t}] + E[\hat{t}]^2]$$

$$= E[\hat{t^2}] - E[\hat{t}]^2$$

$$= \sum P(S)(\hat{t}_S - E[\hat{t}])^2,$$

donde la suma está sobre todas las posibles muestras S. Usualmente no se reporta la varianza sino su raíz cuadrada, el **error estándar**. El error estándar de una media es el valor esperado de la desviación estándar de las medias de múltiples muestras y para una sola muestra está dado por

$$EE = \frac{s}{\sqrt{n}},$$

donde s es la desviación estándar de la media y n es el tamaño de la muestra.

Es común utilizar estimadores sesgados, por lo que otra medida que se utiliza para medir la exactitud de un estimador es el **error cuadrático medio** (MSE).

$$MSE[\hat{t}] = E[(\hat{t} - t)^2]$$

$$= E[\hat{t}^2] - 2tE[\hat{t}] + t^2$$

$$= E[\hat{t}^2] - E[\hat{t}]^2 + (E[\hat{t}]^2 - 2tE[\hat{t}] + t^2)$$

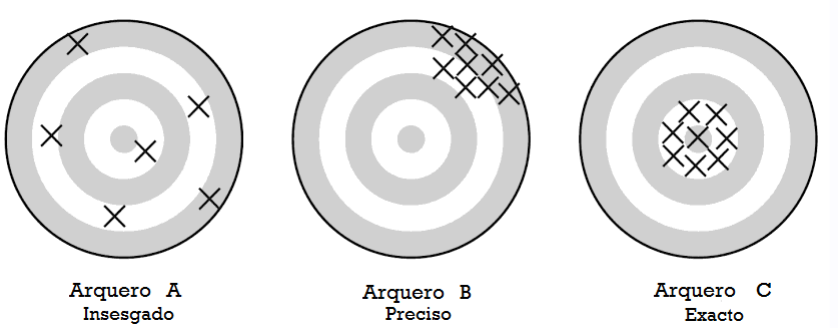
$$= E[\hat{t}^2] - E[\hat{t}]^2 + (E[\hat{t}] - t)^2$$

$$= V[\hat{t}] + (Sesgo[\hat{t}])^2.$$

El MSE mide la exactitud del estimador e indica qué tan cercana está la estimación del valor real; la varianza mide la precisión y también indica qué tan cercanas están las estimaciones de diferentes muestras. En conclusión, el estimador $\hat{t}$ de t es **insesgado** si $E[\hat{t}] = t$, es **preciso** si la varianza $V[\hat{t}]$ es pequeña y es **exacto** si el error cuadrático medio $MSE[\hat{t}]$ es pequeño. Un estimador sesgado puede ser preciso, pero no exacto.

La Figura $II.1$ representa los estimadores insesgados, precisos y exactos como personas que practican tiro con arco. El arquero A es insesgado, pero no es preciso; la posición promedio de todas las flechas que tiró es el centro del blanco de tiro. El arquero B es preciso, pero no insesgado; todas las flechas las tiró juntas, pero lejos del centro del blanco de tiro. Por último, el arquero C es exacto, ya que todas las flechas las tiró juntas y cerca del centro del blanco de tiro.

Figura II.1: Tipos de estimadores



[Lohr, 1999]

Los **intervalos de confianza** son utilizados para indicar la exactitud de una estimación. Usualmente, se utiliza un valor de confianza del 95 %, el cual quiere decir que si se toman distintas muestras de una población una y otra vez, y se calcula el intervalo de confianza para cada muestra posible, se espera que el 95 % de los intervalos obtenidos incluyan el verdadero valor del parámetro poblacional. El radio del intervalo es llamado el margen de error de la estimación y se calcula de la siguiente

manera

$$z_{\alpha/2} EE(\overline{y}),$$

donde $EE(\overline{y})$ es el error estándar de la estimación y $z_{\alpha/2}$ es el percentil $(1-\alpha)/2$ de la distribución normal estándar. Entonces, el intervalo de confianza del 95 % está dado por

$$[\overline{y} - 1.96 EE(\overline{y}), \quad \overline{y} + 1.96 EE(\overline{y})].$$

En ocasiones, por simplicidad, 1.96 se aproxima a 2.

A continuación se presentan los estimadores en diferentes tipos de muestras presentados por Lohr [Lohr, 1999] y se explica cómo el diseño de una muestra determina la varianza y por lo tanto, la precisión de estos estimadores.

2. Muestreo aleatorio simple. Se tiene una población U que contiene N unidades cuyos valores son $\{y_1, y_2, ..., y_N\}$ y se escoge una muestra S de U de n unidades utilizando probabilidades de selección definidas en el diseño de la muestra. Los valores y_i son desconocidos, a menos que sean parte de la muestra. En este caso, el total de la población está dado por

$$t = \sum_{i=1}^{N} y_i,$$

y la media de la población es

$$\overline{y}_U = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

También definimos la varianza de la población como

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \overline{y}_U)^2.$$

Ahora, para estimar la media poblacional $\overline{y}_U$ en una muestra aleatoria simple se utiliza la media muestral

$$\overline{y} = \frac{1}{n} \sum_{i \in S} y_i,$$

el cual es un estimador insesgado de la media poblacional $\overline{y}_U$, y la varianza de $\overline{y}$ es

$$V(\overline{y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right).$$

Como se mencionó en la sección 1, la varianza $V(\overline{y})$ mide las variaciones entre las estimaciones de $\overline{y}$ en diferentes muestras. El factor $\left(1 - \frac{n}{N}\right)$ se llama **factor de corrección para una población finita**. Esta corrección es necesaria, ya que con poblaciones pequeñas, cuanto mayor sea la **fracción muestral** n/N, se tiene más información de la población y la varianza va a ser menor (mayor precisión). Por lo tanto, para un censo, el factor de corrección y la varianza es 0.

Para la mayoría de muestras que se toman de poblaciones muy grandes, el factor de corrección es aproximadamente 1. Para poblaciones grandes, lo que determina la precisión (varianza) del estimador es el tamaño de la muestra, no el porcentaje de la población muestreada. Por ejemplo, una muestra de 100 unidades de una población de 100,000 tiene casi la misma precisión que una muestra de 100 unidades de una población de 100 millones.

$$V[\overline{y}] = \frac{S^2}{100}\frac{99,900}{100,000} = \frac{S^2}{100}(0.999), \quad N = 100,000$$

$$V[\overline{y}] = \frac{S^2}{100}\frac{99,999,900}{100,000,000} = \frac{S^2}{100}(0.999999), \quad N = 100,000,000.$$

Como la varianza poblacional S^2 no la conocemos porque depende de los valores de toda la población, la estimamos con la varianza muestral

$$s^2 = \frac{1}{n-1}\sum_{i \in S}(y_i - \overline{y})^2.$$

Un estimador insesgado de la varianza de $\overline{y}$ es

$$\hat{V}[\overline{y}] = \left(1 - \frac{n}{N}\right)\frac{s^2}{n},$$

y el error estándar es

$$EE[\overline{y}] = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}.$$

Todos estos resultados se aplican para estimar el total de una población t, ya que

$$t = \sum_{i=1}^{N} y_i = N\overline{y}_U.$$

Para estimar t utilizamos el estimador insesgado

$$\hat{t} = N\overline{y}.$$

Entonces,

$$V[\hat{t}] = V[N\overline{y}] = N^2 V[\overline{y}] = N^2 \frac{S^2}{n}\left(1 - \frac{n}{N}\right),$$

$$\hat{V}[\hat{t}] = N^2 \frac{s^2}{n}\left(1 - \frac{n}{N}\right).$$

3. Muestreo estratificado. En algunos casos podemos conocer información adicional que puede ayudar a elaborar el diseño de una muestra. Por ejemplo, antes de llevar a cabo una encuesta, se puede anticipar que los residentes de la ciudad de Guatemala pagan más renta que los residentes en las afueras de la ciudad, o las personas que viven en las áreas rurales no compran comida tan frecuentemente como las personas que viven en la ciudad. Si la variable que nos interesa toma diferentes valores promedio en diferentes subgrupos de la población, podemos obtener estimaciones más exactas de toda la población al tomar una **muestra aleatoria estratificada**. En este tipo de muestreo, se divide a la población en H subpoblaciones llamadas **estratos** (los estratos no deben poseer elementos en común). Luego se elige una muestra probabilística independiente de cada estrato y se agrupa la información para obtener estimaciones generales de la población. Usualmente se utilizan muestras aleatorias estratificadas por las siguientes razones:

- Se quiere evitar tomar una mala muestra. Por ejemplo, si se toma una muestra aleatoria simple de 100 individuos de una población de 1000 hombres y 1000 mujeres, es posible que se obtenga una muestra con muy pocos hombres o muy pocas mujeres. En una muestra estratificada se puede obtener una muestra aleatoria simple de 50 hombres y una muestra aleatoria simple de 50 mujeres. De esta manera se garantiza la misma proporción de hombres y mujeres que en la población.

- A veces se quieren datos de precisión conocida para subgrupos. Para tomar una muestra de estudiantes de Ingeniería Mecánica de la Universidad del Valle, conviene separar el marco de muestreo por género y tomar muestras aleatorias separadas para hombres y mujeres. Debido a que en este tipo de carreras existen más estudiantes hombres que mujeres, es mejor incluir una proporción mayor de estudiantes mujeres que hombres para que exista mayor precisión para los dos grupos.

- Una muestra estratificada puede ser más conveniente de administrar y puede tener un menor costo para el estudio. En una encuesta de negocios, una encuesta por medio electrónico puede ser utilizada para los negocios grandes, mientras que una entrevista personal o por teléfono puede utilizarse para negocios pequeños. En algunas encuestas se utilizan diferentes métodos de recopilación en áreas urbanas y rurales.

- Si la muestra estratificada es realizada correctamente, se obtienen estimaciones con mayor precisión (menor varianza) para la población total. Por ejemplo, personas con diferentes edades poseen diferente presión sanguínea, por lo que sería conveniente estratificar dependiendo de la edad. La estratificación funciona para disminuir la varianza, ya que la varianza dentro de cada estrato usualmente es menor que la varianza en toda la población. [Lohr, 1999]

En una muestra estratificada, se divide una población de N unidades muestrales en H estratos con N_h unidades muestrales en el estrato h-ésimo. Para que el muestreo estratificado funcione se necesitan saber los valores de $N_1, N_2, ..., N_H$ y se debe tener

$$N_1 + N_2 + ... + N_H = N,$$

donde N es el número total de unidades en la población. El muestreo aleatorio estratificado es la forma más simple de muestreo estratificado. En él se toma independientemente una muestra aleatoria simple de cada estrato, tal que n_h observaciones son seleccionadas aleatoriamente de las unidades de la población en el estrato h.

Las estimaciones para la población son las siguientes:

y_{hj}, es el valor de la j-ésima unidad en el estrato h

$t_h = \sum_{j=1}^{N_h} y_{hj}$, es el total de la población en el estrato h

$t = \sum_{h=1}^{H} t_h$, es el total de la población

$\overline{y}_{hU} = \frac{\sum_{j=1}^{N_h} y_{hj}}{N_h}$, es la media poblacional en el estrato h

$\overline{y}_U = \frac{t}{N} = \frac{\sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj}}{N}$, la media poblacional

$S_h^2 = \sum_{j=1}^{N_h} \frac{(y_{hj}-\overline{y}_h)^2}{N_h-1}$, es la varianza poblacional en el estrato h

Las estimaciones para la muestra, utilizando las estimaciones para una muestra aleatoria simple en cada estrato son las siguientes:

$\overline{y}_h = \frac{\sum_{j\in S_h} y_{hj}}{n_h}$

$\hat{t}_h = \frac{N_h}{n_h}\sum_{j\in S_h} y_{hj} = N_h\overline{y}_h$

$s_h^2 = \sum_{j\in S_h} \frac{(y_{hj}-\overline{y}_h)^2}{n_h-1}$

Supongamos que solamente se toma una muestra del estrato h. Entonces, se tienen una población de N_h unidades y se toma una muestra aleatoria simple de n_h unidades. Luego, se estimaría $\overline{y}_{hU}$ por $\overline{y}_h$ y t_h por $\hat{t}_h = N_h\overline{y}_h$. El total de la población es $t = \sum_{h=1}^{H} t_h$ entonces, estimamos t por

$$\hat{t}_{str} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} N_h\overline{y}_h. \tag{B.1}$$

Para estimar $\overline{y}_U$ utilizamos

$$\overline{y}_{str} = \frac{\hat{t}_{str}}{N} = \sum_{h=1}^{H} \frac{N_h}{N}\overline{y}_h.$$

Las propiedades de estos estimadores vienen directamente de las propiedades de los estimadores en muestras aleatorias simples.

- **Sesgo** Los estimadores $\overline{y}_{str}$ y $\hat{t}_{str}$ son estimadores no sesgados de $\overline{y}_U$ y t.

$$E\left[\overline{y}_{str}\right] = E\left[\sum_{h=1}^{H} \frac{N_h}{N}\overline{y}_h\right] = \sum_{h=1}^{H} \frac{N_h}{N} E[\overline{y}_h] = \sum_{h=1}^{H} \frac{N_h}{N}\overline{y}_{hU} = \overline{y}_U$$

- **Varianza de estimadores** En este caso, las muestras que se toman de cada estrato son independientes y se conocen $V(\hat{t}_h)$ de la teoría de muestras aleatorias simples, las propiedades de valor esperado y de la ecuación de varianza en muestreo aleatorio simple. Entonces,

$$V(\hat{t}_{str}) = \sum_{h=1}^{H} V(\hat{t}_h) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{S_h^2}{n_h}. \tag{B.2}$$

- **Estimaciones de varianza para muestras estratificadas** Podemos obtener un estimador no sesgado de $V(\hat{t}_{str})$ al sustituir las cantidades poblacionales, S_h^2, por las estimaciones de la muestra, s_h^2. Notemos que para estimar las varianzas se necesitan por lo menos 2 unidades en cada estrato.

$$\hat{V}(\hat{t}_{str}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_h^2}{n_h}$$

$$\hat{V}(\overline{y}_{str}) = \frac{1}{N^2}\hat{V}(\hat{t}_{str}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}.$$

  Como se mencionó anteriormente, el uso de estratos reduce la varianza de las estimaciones, por lo que las hace más precisas. Esto se debe a que la varianza dentro de cada estrato es menor a la varianza de toda la población.

a. Pesos muestrales. El estimador $\hat{t}_{str}$ en B.1 se puede escribir como la siguiente suma ponderada

$$\hat{t}_{str} = \sum_{h=1}^{H} \sum_{j \in S_h} \frac{N_h}{n_h} y_{hj},$$

donde $w_{jh} = N_h/n_h$ es el **peso muestral** y puede ser interpretado como el número de unidades en la población representadas por el miembro de la muestra (h, j) (miembro j-ésimo en el estrato h). Si la población posee 1600 hombres, 400 mujeres y el diseño de la muestra estratificada menciona que se deben muestrear 200 hombres y 200 mujeres entonces, cada hombre en la muestra tienen un peso de $(1600/200) = 8$ y cada mujer un peso de $(400/200) = 2$. Cada mujer en la muestra se representa a ella misma y a otra mujer que no forma parte de la muestra, y cada hombre en la

muestra se representa a él mismo y a 7 hombres que no forman parte de la muestra. Notamos que la probabilidad que la j-ésima unidad en el estrato h sea seleccionada para la muestra es $\pi_{hj} = n_h/N_h$. Entonces, el peso muestral es simplemente el recíproco de la probabilidad de selección:

$$w_{hj} = \frac{1}{\pi_{hj}}.$$

La suma de los pesos muestrales es igual al tamaño de la población N. La estimación del total de la población estratificada está dado por

$$\hat{t}_{str} = \sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj},$$

y la estimación de la media poblacional está dada por

$$\overline{y}_{str} = \frac{\sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j \in S_h} w_{hj}}.$$

b. Asignación proporcional. La **asignación proporcional** es una de las técnicas de muestreo estratificado. En la asignación proporcional, el tamaño de cada estrato en la muestra es proporcional a su tamaño en la población. Por ejemplo, si se tiene una población de 2400 hombres y 1600 mujeres, y se quiere tomar una muestra del $10\,\%$ con asignación proporcional entonces, se escogerían 240 hombres y 160 mujeres. La probilidad que un individuo sea seleccionado para formar parte de la muestra, n/N, es la misma que en una muestra aleatoria simple, pero, en el caso de una muestra estratificada con asignación proporcional, solamente se elige la muestra que contiene 240 hombres y 160 mujeres. En este ejemplo, cada hombre en la muestra representa a 10 hombres de la población y lo mismo ocurre para cada mujer en la muestra.

La muestra se llama **auto ponderada** cuando cada elemento en la muestra posee el mismo peso y representa la misma cantidad de unidades en la población. En este tipo de muestras, el estimador $\overline{y}_{str}$ es el promedio de todas las observaciones de la muestra.

En una muestra estratificada de tamaño n con asignación proporcional, como $n_h/N_h = n/N$, de

la ecuación B.2 se obtiene

$$V_{prop}(\hat{t}_{str}) = \sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)N_h^2\frac{S_h^2}{n_h} = \left(1-\frac{n}{N}\right)\frac{N}{n}\sum_{h=1}^{H}N_hS_h^2$$

$$= \left(1-\frac{n}{N}\right)\frac{N}{n}\sum_{h=1}^{H}\left(S_h^2[(N_h-1)+1]\right)$$

$$= \left(1-\frac{n}{N}\right)\frac{N}{n}\left(\sum_{h=1}^{H}(N_h-1)S_h^2+\sum_{h=1}^{H}S_h^2\right)$$

$$= \left(1-\frac{n}{N}\right)\frac{N}{n}\left(\sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj}-\overline{y}_{hU})^2+\sum_{h=1}^{H}S_h^2\right)$$

$$= \left(1-\frac{n}{N}\right)\frac{N}{n}\left(SSW+\sum_{h=1}^{H}S_h^2\right),$$

donde SSW es la suma de cuadrados dentro de los estratos. Definimos la suma de cuadrados entre estratos como

$$SSB = \sum_{h=1}^{H}\sum_{j=1}^{N_h}(\overline{y}_{hU}-\overline{y}_U)^2 = \sum_{h=1}^{H}N_h(\overline{y}_{hU}-\overline{y}_U)^2,$$

y sabemos que el total, $SSTO$, está definido como

$$SSTO = \sum_{h=1}^{H}\sum_{j=1}^{N_h}(\overline{y}_{hj}-\overline{y}_U)^2 = (N-1)S^2$$

y $SSTO = SSW + SSB$. Entonces, de las muestras aleatorias simples

$$V_{SRS}(\hat{t}) = \left(1-\frac{n}{N}\right)N^2\frac{S^2}{n}$$

$$= \left(1-\frac{n}{N}\right)\frac{N^2}{n}\frac{SSTO}{N-1}$$

$$= \left(1-\frac{n}{N}\right)\frac{N^2}{n(N-1)}(SSW+SSB)$$

$$= V_{prop}(\hat{t}_{str}) + \left(1-\frac{n}{N}\right)\frac{N}{n(N-1)}\left[N(SSB)-\sum_{h=1}^{H}(N-N_h)S_h^2\right].$$

Esto muestra que utilizando estratificación con asignación proporcional, la varianza no sobrepasa la

varianza de una muestra aleatoria simple, a no ser que

$$N(SSB) - \sum_{h=1}^{H}(N - N_h)S_h^2 < 0$$

$$SSB < \sum_{h=1}^{H}\left(1 - \frac{N_h}{N}\right)S_h^2.$$

Lo cual no ocurre si los N_h son suficientemente grandes, ya que $N_h(\overline{y}_{hU} - \overline{y}_U)^2 > S_h^2$. Este resultados solamente se toma en cuenta para varianzas poblacionales. Es posible que una estimación de la varianza utilizando asignación proporcional sea mayor que la estimación de la varianza utilizando muestreo aleatorio simple.

4. Muestreo por conglomerados. Anteriormente vimos que en el muestreo estratificado, la varianza del estimador de t depende de la variabilidad dentro de los estratos y si la variabilidad entre las medias de los estratos es grande a comparación de la variabilidad dentro de los estratos, el muestreo estratificado aumenta la precisión. En el muestreo por conglomerados ocurre lo opuesto. La mayoría de veces, el muestreo por conglomerados proporciona menor precisión (mayor varianza) para los estimadores que si se toma una muestra aleatoria simple con el mismo número de elementos. La razón por la cual se utiliza el muestreo por conglomerados en las encuestas es debido a que resulta menos costoso y no toma mucho tiempo. Los costos reducidos justifican la pérdida de precisión en las estimaciones.

Utilizando el muestreo por conglomerados, la variabilidad de un estimador insesgado de t depende completamente de la variación entre conglomerados, ya que

$$S_t^2 = \sum_{i=1}^{N}\frac{(t_i - \overline{t}_U)^2}{N-1} = \sum_{i=1}^{N}\frac{(M\overline{y}_{iU} - M\overline{y}_U)^2}{N-1} = \sum_{i=1}^{N}\frac{M^2(\overline{y}_{iU} - \overline{y}_U)^2}{N-1} = M(MSB),$$

donde t_i es el total en el conglomerado i, M es el número de elementos en la muestra en cada conglomerado, $\overline{y}_{iU}$ es la media poblacional en el conglomerado i y MSB es la media cuadrática entre los conglomerados. Entonces, para el muestreo por conglomerados,

$$V(\hat{t}_{congl}) = N^2\left(1 - \frac{n}{N}\right)\frac{M(MSB)}{n}.$$

La media cuadrática entre conglomerados, MSB, mide la variación entre conglomerados, y si los elementos en conglomerados distintos varían más entre ellos que los elementos en el mismo conglomerado, la MSB aumenta y el muestreo por conglomerados disminuye la precisión.

CAPÍTULO III: ESTIMADORES DIRECTOS

Un estimador de un dominio se llama **directo** si utiliza los datos obtenidos de la variable de interés, y, solamente de las unidades del dominio que están en la muestra. Los datos de una encuesta son utilizados para proveer estimaciones directas y confiables para toda la población y dominios. Pero en el caso de la estimación en áreas pequeñas, los estimadores directos producen errores estándar muy grandes debido al tamaño reducido de las muestras en los dominios de interés, y en el peor de los casos, puede ser que no se hayan seleccionado unidades muestrales de un dominio en específico. Para una descripción más amplia de estimación directa se puede consultar el libro "Samplying: Design and Analysis" de Sharon Lohr [Lohr, 1999], del cual se obtuvieron los siguientes estimadores y modelos.

A. Estimadores de razón y regresión

En las encuestas es común que se posea información adicional (de la misma encuesta). Esta información adicional de cada una de las unidades puede ser utilizada para mejorar la precisión de nuestras estimaciones. Las estimaciones de razón y regresión utilizan variables que están correlacionadas con la variable de interés para aumentar la precisión de estimaciones como la media o el total poblacional.

1. Estimación de razón. Para realizar una estimación de razón se necesitan dos cantidades y_i y x_i de cada unidad muestral, donde x_i es usualmente llamada una **variable auxiliar** o **subsidiaria**. En una población U de tamaño N,

$$t_y = \sum_{i=1}^{N} y_i, \quad t_x = \sum_{i=1}^{N} x_i$$

y su razón es

$$B = \frac{t_y}{t_x} = \frac{t_y}{t_x} \cdot \frac{N}{N} = \frac{t_y/N}{t_x/N} = \frac{\sum_{i=1}^{N} y_i/N}{\sum_{i=1}^{N} x_i/N} = \frac{\overline{y}_U}{\overline{x}_U}.$$

En la aplicación más simple de estimación de razón se toma una muestra aleatoria simple de tamaño n y la información en x y y es utilizada para estimar B, t_y o $\overline{y}_U$. La estimación de razón

y la estimación de regresión se aprovechan de la correlación entre x y y en la población; a mayor correlación, mejor es la estimación.

Ejemplo 1. ***Estimador de razón.****Supongamos que una población consiste de terrenos de diferente tamaño para agricultura. Sean y_i los kilogramos de granos que se cosecharon en el terreno i y x_i la superficie del terreno i medida en hectáreas. Entonces, B es el rendimiento medio en kilogramos por acre, $\overline{y}_U$ es el rendimiento medio en kilogramos por terreno y t_y es el rendimiento total en kilogramos. Si se toma una muestra aleatoria simple S, los estimadores de B, t_y, y $\overline{y}_U$ son*

$$\hat{B} = \frac{\overline{y}}{\overline{x}} = \frac{\hat{t}_y}{\hat{t}_x} \tag{A.1}$$

$$\hat{t}_{yr} = \hat{B} t_x \tag{A.2}$$

$$\hat{\overline{y}}_r = \hat{B}\overline{x}_U, \tag{A.3}$$

donde t_x y $\overline{x}_U$ se conocen, y $\hat{t}_{yr}$ y $\hat{\overline{y}}_r$ son los estimadores de razón. Cada vez que se obtiene una muestra aleatoria simple y se estima la media o proporción para una subpoblación, se está utilizando la estimación de razón.

a. Sesgo y error estándar . Los estimadores de razón para $\overline{y}_U$ y t_y usualmente son sesgados, a contrario de los estimadores $\overline{y}$ y $N\overline{y}$. La estimación de las varianzas para $\hat{B}, \hat{t}_{yr}$ y $\hat{\overline{y}}_r$ está dada por

$$\hat{V}[\hat{B}] = \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n\overline{x}_U^2}$$

$$\hat{V}[\hat{t}_{yr}] = \hat{V}[t_x \hat{B}] = N^2 \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n}$$

$$\hat{V}[\hat{\overline{y}}_r] = \hat{V}[\overline{x}_U \hat{B}] = \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n},$$

donde si no se conoce $\overline{x}_U$, puede sustituirse por $\overline{x}_S$, y s_e^2, la varianza de la muestra, está dada por

$$s_e^2 = \frac{\sum_{i \in S} (y_i - \hat{B} x_i)^2}{n - 1}.$$

Si la muestra es suficientemente grande, los intervalos con 95 % de confianza pueden ser obtenidos de la siguiente manera

$$\hat{B} \pm 1.96EE[\hat{B}], \quad \hat{\overline{y}}_r \pm 1.96EE[\hat{\overline{y}}_r], \quad \hat{t}_{yr} \pm 1.96EE[\hat{t}_{yr}],$$

donde $EE[\hat{B}]$ es el error estándar del estimador, $EE[\hat{B}] = \sqrt{\hat{V}[\hat{B}]}$, respectivamente para cada estimador y 1.96 puede aproximarse a 2.

2. Estimación de regresión. La estimación de razón funciona mejor cuando los datos se comportan como una línea recta que pasa por el origen. Cuando los datos se comportan como una línea recta que no pasa por el origen se utiliza el modelo de regresión usual

$$y = B_0 + B_1 x.$$

Supongamos que se sabe $\overline{x}_U$, la media de la población para la variable x. Entonces, el estimador de regresión de $\overline{y}_U$ es el valor predecido de y por modelo de regresión cuando $x = \overline{x}_U$:

$$B_0 = y - B_1 x$$

$$\hat{B}_0 = \overline{y} - \hat{B}_1 \overline{x}$$

$$\hat{\overline{y}}_{reg} = \hat{B}_0 + \hat{B}_1 \overline{x}_U = \overline{y} + \hat{B}_1(\overline{x}_U - \overline{x}), \tag{A.4}$$

donde $\hat{B}_0$ y $\hat{B}_1$ son los coeficientes de regresión. Para este modelo,

$$\hat{B}_1 = \frac{\sum_{i \in S}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i \in S}(x_i - \overline{x})^2} = \frac{r s_y}{s_x},$$

$$\hat{B}_0 = \overline{y} - \hat{B}_1 \overline{x},$$

r es el coeficiente de correlación muestral de x y y, s_x es la desviación estándar de la muestra de la variable x y s_y es la desviación estándar de la muestra de la variable y.

a. Sesgo y error estándar. De la misma manera que el estimador de razón, el estimador de regresión es sesgado. Sea B_1 el coeficiente de mínimos cuadrados calculado a partir de

todos los datos de la población:

$$B_1 = \frac{\sum_{i=1}^{N}(x_i - \overline{x}_U)(y_i - \overline{y}_U)}{\sum_{i=1}^{N}(x_i - \overline{x}_U)^2} = \frac{RS_y}{S_x},$$

S_y y S_x son las desviaciones estándar poblacionales de las variables y y x respectivamente. Entonces, el sesgo de $\hat{\overline{y}}_{reg}$ está dado por

$$E[\hat{\overline{y}}_{reg} - \overline{y}_U] = E[\overline{y} + \hat{B}_1(\overline{x}_U - \overline{x}) - \overline{y}_U]$$

$$= E[\overline{y} - \overline{y}_U] + E[\hat{B}_1(\overline{x}_U - \overline{x})]$$

$$= -Cov(\hat{B}_1, \overline{x}).$$

Si la línea de regresión pasa por todos los puntos de la población entonces, el sesgo es cero. En ese caso, $\hat{B}_1 = B_1$ para cada muestra entonces, $Cov(\hat{B}_1, \overline{x}) = 0$.

El error estándar puede ser calculado encontrando la varianza de los residuos. Sea $e_i = y_i - (\hat{B}_0 + \hat{B}_1 x_i)$ entonces,

$$EE(\hat{\overline{y}}_{reg}) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}}.$$

B. Modelos para estimación de razón y regresión

En algunos casos se ha mencionado que si un modelo de regresión se ajusta a los datos de una encuesta, el modelo puede ser utilizado para estimar el total para y y su error estándar. También, la manera en la que se obtienen los datos no es tan importante como el modelo al que se ajustan.

1. Modelo para estimación de razón. Como se mencionó anteriormente, la estimación de razón se utiliza en una muestra aleatoria simple cuando los datos se ajustan a una línea recta que pasa por el origen y la varianza de las observaciones en la línea es proporcional a x. Estas condiciones se pueden escribir como el siguiente modelo de regresión lineal: Supongamos que se conocen $x_1, x_2, ..., x_N$ (todas mayores a cero) y $Y_1, Y_2, ..., Y_N$ son independientes y siguen el siguiente modelo

$$Y_i = \beta x_i + \varepsilon_i,$$

donde $E_M[\varepsilon_i] = 0$ y $V_M[\varepsilon_i] = \sigma^2 x_i$. Bajo el modelo, $T_y = \sum_{i=1}^{N} Y_i$ es una variable aleatoria que puede tomar el valor de t_y, el cual es el total poblacional. Si S representa el conjunto de unidades en la muestra entonces,

$$t_y = \sum_{i \in S} y_i + \sum_{i \notin S} y_i.$$

Observamos los valores de y_i para las unidades dentro de la muestra y los predecimos como $\hat{\beta} x_i$ para las unidades que no están en la muestra, donde $\hat{\beta} = \overline{y}/\overline{x}$ es la estimación de mínimos cuadrados ponderada de β bajo el modelo $Y_i = \beta x_i + \varepsilon_i$. Entonces, un estimador de t_y es

$$\hat{t}_y = \sum_{i \in S} y_i + \hat{\beta} \sum_{i \notin S} x_i = n\overline{y} + \frac{\overline{y}}{\overline{x}} \sum_{i \notin S} x_i = \sum_{i \in S} y_i + \frac{\sum_{i \in S} y_i}{\sum_{i \in S} x_i} \sum_{i \notin S} x_i =$$

$$= \frac{\sum_{i \in S} y_i}{\sum_{i \in S} x_i} \left(\sum_{i \in S} x_i + \sum_{i \notin S} x_i \right) = \frac{\sum_{i \in S} y_i}{\sum_{i \in S} x_i} \sum_{i=1}^{N} x_i = \frac{\overline{y}}{\overline{x}} t_x$$

Lo anterior es simplemente la estimación de razón de t_y que se muestra en A.2.

El estimador basado en el modelo

$$\hat{T}_y = \sum_{i \in S} Y_i + \hat{\beta} \sum_{i \notin S} x_i$$

es insesgado bajo el modelo, ya que

$$E_M[\hat{T}_y - T] = E_M \left[\sum_{i \in S} Y_i + \hat{\beta} \sum_{i \notin S} x_i - \sum_{i \in S} Y_i - \sum_{i \notin S} Y_i \right] = E_M \left[\hat{\beta} \sum_{i \notin S} x_i - \sum_{i \notin S} Y_i \right] =$$

$$= E_M \left[\hat{\beta} \sum_{i \notin S} x_i \right] - E_M \left[\sum_{i \notin S} \hat{\beta} x_i + \varepsilon_i \right] = E_M \left[\hat{\beta} \sum_{i \notin S} x_i \right] - E_M \left[\hat{\beta} \sum_{i \notin S} x_i \right] - E_M \left[\varepsilon_i \right] = 0.$$

La varianza basada en el modelo es

$$V_M = [\hat{T}_y - T] = V_M \left[\hat{\beta} \sum_{i \notin S} x_i - \sum_{i \notin S} Y_i \right] = V_M \left[\hat{\beta} \sum_{i \notin S} x_i \right] + V_M \left[\sum_{i \notin S} Y_i \right]$$

porque $\hat{\beta}$ y $\sum_{i\notin S} Y_i$ se consideran independientes por las suposiciones del modelo. El modelo no depende de cuáles unidades de la población se escogen para estar en la muestra S entonces, S puede considerarse como fija. Por esto,

$$V_M\left[\sum_{i\notin S} Y_i\right] = V_M\left[\sum_{i\notin S} \beta x_i + \varepsilon_i\right] = V_M\left[\sum_{i\notin S} \varepsilon_i\right] = \sigma^2\left(\sum_{i\notin S} x_i\right),$$

de igual manera,

$$V_M\left[\hat{\beta}\sum_{i\notin S} x_i\right] = \left(\sum_{i\notin S} x_i\right)^2 V_M\left[\frac{\sum_{i\in S} Y_i}{\sum_{i\in S} x_i}\right] = \left(\sum_{i\notin S} x_i\right)^2 \frac{V_M\left[\sum_{i\in S} Y_i\right]}{\left(\sum_{i\in S} x_i\right)^2} =$$

$$= \left(\sum_{i\notin S} x_i\right)^2 \frac{\sigma^2\left(\sum_{i\in S} x_i\right)}{\left(\sum_{i\in S} x_i\right)^2} = \left(\sum_{i\notin S} x_i\right)^2 \frac{\sigma^2}{\sum_{i\in S} x_i}.$$

Entonces, combinando las dos expresiones anteriores

$$V_M = [\hat{T}_y - T] = \left(\sum_{i\notin S} x_i\right)^2 \frac{\sigma^2}{\sum_{i\in S} x_i} + \sigma^2\left(\sum_{i\notin S} x_i\right) = \frac{\sigma^2\sum_{i\notin S} x_i}{\sum_{i\in S} x_i}\left(\sum_{i\notin S} x_i + \sum_{i\in S} x_i\right) =$$

$$= \frac{\sigma^2\sum_{i\notin S} x_i}{\sum_{i\in S} x_i} t_x = \frac{\sigma^2\left(t_x - \sum_{i\in S} x_i\right)}{\sum_{i\in S} x_i} t_x = \frac{\sigma^2 t_x^2}{\sum_{i\in S} x_i} - \sigma^2 t_x = \sigma^2 t_x\left(\frac{t_x}{\sum_{i\in S} x_i} - 1\right) =$$

$$= \frac{\sigma^2 t_x^2}{t_x}\left(\frac{t_x}{\sum_{i\in S} x_i} - 1\right) = \sigma^2 t_x^2\left(\frac{1}{\sum_{i\in S} x_i} - \frac{1}{t_x}\right) = \frac{\sum_{i\in S} x_i}{\sum_{i\in S} x_i}\sigma^2 t_x^2\left(\frac{1}{\sum_{i\in S} x_i} - \frac{1}{t_x}\right) =$$

$$= \frac{\sigma^2 t_x^2}{\sum_{i\in S} x_i}\left(1 - \frac{\sum_{i\in S} x_i}{t_x}\right).$$

Si el tamaño de la muestra es pequeño en relación con el tamaño de la población entonces,

$$V_M = [\hat{T}_y - T] \approx \frac{\sigma^2 t_x^2}{\sum_{i\in S} x_i}.$$

En este caso, el valor

$$\left(1 - \frac{\sum_{i \in S} x_i}{t_x}\right)$$

sirve como un **factor de corrección para una población finita**.

Cuando se utiliza un modelo para un conjunto de datos se necesitan verificar los supuestos del modelo. Los supuestos para cualquier modelo de regresión lineal son los siguientes:

1. El modelo es correcto.
2. La estructura de la varianza es como se da.
3. Las observaciones son independientes.

Los supuestos (1) y (2) usualmente se pueden verificar al graficar los datos y examinar los residuos del modelo. La suposición (3) puede ser difícil de verificar en la práctica y requiere conocimiento de cómo se recolectaron los datos. Generalmente, si se toma una muestra aleatoria, se asume automáticamente la independencia de las observaciones.

El modelo se puede verificar utilizando una recta a través del origen. Si la varianza de y_i sobre la línea es proporcional a x_i entonces, una gráfica de los residuos ponderados

$$\frac{y_i - \hat{\beta} x_i}{\sqrt{x_i}}$$

con los valores x_i o $\log x_i$ no debería presentar un patrón.

2. Modelo para estimación de regresión. El modelo para la estimación de regresión es el siguiente

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

donde los ε_i son independientes e idénticamente distribuidos con media 0 y varianza constante σ^2. En este modelo, los estimadores de mínimos cuadrados de β_0 y β_1 son

$$\hat{\beta}_0 = \overline{Y}_S - \hat{\beta}_1 \overline{x}_S. \tag{B.1}$$

$$\hat{\beta}_1 = \frac{\sum_{i \in S} (x_i - \overline{x}_S)(Y_i - \overline{Y}_S)}{\sum_{i \in S} (x_i - \overline{x}_S)^2}.$$

Entonces, utilizando los valores predecidos en vez de las unidades no muestreadas

$$\hat{T}_y = \sum_{i\in S} Y_i + \sum_{i\notin S}(\hat{\beta}_0 + \hat{\beta}_1 x_i) = n\overline{Y}_S + \sum_{i\notin S}(\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$= n(\hat{\beta}_0 + \hat{\beta}_1 \overline{x}_S) + \sum_{i\notin S}(\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$= n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i\in S} x_i + \sum_{i\notin S}(\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$= \sum_{i\in S}(\hat{\beta}_0 + \hat{\beta}_1 x_i) + \sum_{i\notin S}(\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$= \sum_{i=1}^{N}(\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$\hat{T}_y = N(\hat{\beta}_0 + \hat{\beta}_1 \overline{x}_U) \tag{B.2}$$

Entonces, el estimador de regresión de $\hat{T}_y$ es N veces el valor predecido bajo el modelo en $\overline{x}_U$. Sabemos que el estimador de regresión de la media $\hat{\overline{Y}}_{reg} = \hat{T}_y/N$. Entonces, por B.2

$$\hat{\overline{Y}}_{reg} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}_U \tag{B.3}$$

Sustituyendo B.1 en B.3 tenemos

$$\hat{\overline{Y}}_{reg} = \overline{Y}_S - \hat{\beta}_1 \overline{x}_S + \hat{\beta}_1 \overline{x}_U$$

$$= \overline{Y}_S + \hat{\beta}_1(\overline{x}_U - \overline{x}_S), \tag{B.4}$$

el cual es el estimador de regresión mostrado en A.4.

La varianza del estimador $\hat{\overline{Y}}_{reg}$ es

$$\sigma^2 \left[\frac{1}{n} + \frac{(\overline{x}_U - \overline{x})^2}{\sum_{i\in S}(x_i - \overline{x})^2}\right].$$

Si la **fracción muestral** n/N es pequeña,

$$V_M[\hat{T}_y - T] \approx N^2\sigma^2 \left[\frac{1}{n} + \frac{(\overline{x}_U - \overline{x}_S)^2}{\sum_{i\in S}(x_i - \overline{x}_S)^2}\right].$$

C. Estimación en dominios

Cuando el tamaño de la muestra de un dominio en una muestra aleatoria simple es suficientemente grande, se pueden utilizar las fórmulas de muestras aleatorias simples para hacer inferencias sobre las medias del dominio. Para el caso de muestras complejas, la estimación en dominios es distinta.

Una **muestra compleja** posee los siguientes componentes: muestreo aleatorio, estimación de razón, estratificación y conglomerados. También es importante mencionar que, debido al diseño complejo de las encuestas, el cálculo de varianzas, errores estándar e intervalos de confianza no puede hacerse en forma directa; para esto se debe utilizar un procedimiento también complejo, como el de Linealización de Taylor. Un ejemplo presentado por Lohr [Lohr, 1999] trata sobre una encuesta realizada en Gambia en 1991 que fue diseñada para estimar el uso de redes de mosquitos alrededor de la cama en la zona rural. El uso de las redes para mosquitos puede disminuir los casos de Malaria.

El marco muestral de la encuesta consistía en todos los pueblos rurales de Gambia que poseían menos de 3,000 habitantes. Los pueblos fueron estratificados en tres regiones geográficas (oriental, central y occidental) y si la región poseía un puesto de salud público (PSP) o no. En cada región se eligieron cinco distritos con probabilidades proporcionales a la población estimada de cada distrito en el censo nacional de 1983. En cada distrito se eligieron cuatro pueblos, nuevamente con probabilidad proporcional a la población del censo: 2 pueblos con PSP y 2 puestos sin PSP. Finalmente, se eligieron seis grupos de hogares de cada pueblo, más o menos aleatoriamente y un investigador registró el número de camas y redes de mosquitos en cada grupo. De manera resumida, el diseño de la muestra fue el siguiente:

Etapa	Unidad muestral	Estratificación
1	Distritos	Región
2	Pueblos	con PSP/sin PSP
3	Villas	

Supongamos que y_i es la variable de interés y sea

$$x_{id} = \begin{cases} 1 & \text{si la unidad de observación } i \text{ está en el dominio } d \\ 0 & \text{si la unidad de observación } i \text{ no está en el dominio } d \end{cases}$$

Entonces, el estimador para el total de la población en el dominio d está dado por

$$\hat{t}_d = \sum_{i \in S} w_i x_{id} y_i,$$

y la media poblacional en el dominio d, asumiendo que la muestra posee algunas observaciones pertenecientes al dominio d, es

$$\hat{\bar{y}}_d = \frac{\hat{t}_d}{\sum_{i \in S} w_i x_{id}}. \tag{C.1}$$

La varianza de $\hat{\bar{y}}_d$ está dada por

$$\hat{V}(\hat{\bar{y}}_d) = \frac{1}{\hat{N}_d^2} \hat{V} \left[\sum_{i \in S} w_i x_{id} (y_i - \hat{\bar{y}}_d) \right]. \tag{C.2}$$

Para que la varianza sea pequeña, el tamaño de la muestra en el dominio d debe ser grande.

Por lo presentado en la sección 2, si ignoramos el factor de corrección para una población finita y se toma una muestra aleatoria simple, la ecuación C.2 pasa a ser

$$\hat{V}(\hat{\bar{y}}_d) \approx \frac{s_d^2}{n_d},$$

donde n_d es el número de observaciones en la muestra que están en el dominio d y s_d^2 es la varianza de esas observaciones.

CAPÍTULO IV: ESTIMADORES INDIRECTOS

La mayoría de métodos para estimación en áreas pequeñas involucran el uso de estimadores **indirectos**. Los métodos de estimación indirecta utilizan datos auxiliares de la población, provenientes de censos u otras fuentes, a través del uso de modelos explícitos o implícitos. Un estimador indirecto posee mayor precisión que un estimador directo, ya que utiliza información adicional, lo que hace que la varianza de la estimación indirecta sea menor a la de la estimación directa.

Los estimadores indirectos se pueden dividir en modelos que incluyen factores aleatorios y los que no los incluyen. Usualmente, los últimos son llamados estimadores tradicionales. Los factores (o efectos) aleatorios y los modelos que los incluyen se definirán más adelante.

A. Estimadores indirectos tradicionales

Dentro de los estimadores indirectos tradicionales, Rao [Rao, 2003] y Marker [A.Marker, 1999] mencionan los estimadores **sintéticos** y los **compuestos**. En la estimación sintética más simple, de primero se obtienen estimaciones nacionales (o regionales) por subgrupos, los cuales pueden construirse tomando en cuenta la edad, raza o sexo, y luego se calcula la estimación del área pequeña tomando un promedio ponderado de estas estimaciones nacionales por subgrupo donde los pesos reflejan la composición del área pequeña en cuanto a los subgrupos. Este método depende fuertemente de qué tan similares son las estimaciones nacionales por subgrupo y las estimaciones por subgrupo del área pequeña. [A.Marker, 1999] En el caso del estimador compuesto, éste se construye a partir de un promedio ponderado de un estimador sintético y un estimador directo. A continuación se presenta una breve introducción de estos dos estimadores.

1. Estimadores sintéticos. Un estimador directo y confiable para un área grande, que cubre varias áreas pequeñas, se utiliza para construir un estimador **sintético** para un área pequeña suponiendo que las áreas pequeñas tienen las mismas características que el área grande. [Rao, 2003]

Estos estimadores fueron propuestos en 1968 por el Centro Nacional de Estadísticas de Salud en Estados Unidos para obtener estimaciones de las tasas de discapacidad de la Encuesta Nacional de Entrevistas de Salud. [A.Marker, 1999]

Supongamos que se necesita estimar la media, $\overline{Y}_i$, de un área pequeña. Por ejemplo, la proporción P_i de personas que viven en pobreza en el área pequeña i ($\overline{Y}_i = P_i$). En este caso, Marker [A.Marker, 1999] define el estimador sintético de $\overline{Y}_i$ como

$$\hat{\overline{Y}}_{iS} = \frac{\sum_j N_{ij}\overline{y}_j}{N_i},$$

donde N_{ij} es el tamaño de la población en el área pequeña i y el subgrupo j, N_i es el tamaño total de la población en el área pequeña i, y $\overline{y}_j$ es la media (estimación directa) de la variable de interés en el subgrupo j. Como el sesgo de $\hat{\overline{Y}}_{iS}$ es aproximadamente igual a $\overline{Y} - \overline{Y}_i$, donde $\overline{Y}$ es la media poblacional, si la media del área pequeña es aproximadamente igual a la media poblacional, $\overline{Y}_i \approx \overline{Y}$, el sesgo va a ser pequeño y el estimador sintético será considerado muy eficiente, ya que su error cuadrado medio (MSE) también va a ser pequeño.

En el caso de áreas que muestran características individuales muy marcadas, el MSE va a tomar un valor grande; entonces, la condición anterior, $\overline{Y}_i \approx \overline{Y}$, puede ser cambiada a $\overline{Y}_i \approx \overline{Y}(r)$, donde $\overline{Y}(r)$ es la media de un área más grande, llamada región, que cubre el área pequeña. En este caso, utilizamos $\hat{\overline{Y}}_{iS} = \hat{\overline{Y}}(r)$, donde $\hat{\overline{Y}}(r)$ es el estimador directo regional. El sesgo de $\hat{\overline{Y}}(r)$ es aproximadamente igual a $\overline{Y}(r) - \overline{Y}_i$ el cual, bajo la condición planteada anteriormente, va a ser muy pequeño junto con el MSE de $\hat{\overline{Y}}(r)$.

2. Estimador compuesto. Existen distintos estimadores compuestos. J.N.K. Rao [Rao, 2003] los presenta como un promedio ponderado del estimador sintético y el estimador directo para un área pequeña i. Esto se debe a que una manera de balancear el sesgo de un estimador sintético, $\hat{Y}_{i2}$, y la inestabilidad de un estimador directo, $\hat{Y}_{i1}$, es tomar un promedio ponderado de $\hat{Y}_{i1}$ y $\hat{Y}_{i2}$. El estimador compuesto del total del área pequeña Y_i está dado de la siguiente manera

$$\hat{Y}_{iC} = \phi_i \hat{Y}_{i1} + (1 - \phi_i)\hat{Y}_{i2},$$

para un peso $0 \leq \phi_i \leq 1$, el cual Marker [A.Marker, 1999] lo define como

$$\phi_i = \frac{MSE(\hat{Y}_{i2})}{Var(\hat{Y}_{i1}) + MSE(\hat{Y}_{i2})}.$$

B. Modelos de áreas pequeñas

Los métodos tradicionales de estimación indirecta, como los presentados anteriormente, se basan en modelos implícitos que buscan relacionar áreas pequeñas con características similares a través del uso de información adicional. En la actualidad, los métodos de estimación indirecta basados en modelos explícitos son los más utilizados debido a que poseen numerosas ventajas sobre los métodos basados en modelos implícitos. En particular, con estos métodos se pueden obtener medidas de precisión para cada área pequeña, se pueden utilizar modelos de diagnóstico, como el análisis de residuos, para encontrar modelos que se ajusten a los datos, y se pueden manejar casos complejos de datos. [Rao, 2003]

Los individuos dentro de una sociedad son influenciados por los grupos sociales o el contexto al que pertenecen y, de manera contraria, las características de una sociedad son influenciadas por los individuos que la conforman. Las encuestas usualmente involucran problemas que investigan la relación entre el individuo y la sociedad. [Hox, 2002] Es por esto que los modelos consideran una estructura de la población jerárquica (por **niveles**). Un ejemplo puede ser una encuesta de salud en la que los niveles son: departamentos, municipios, hogares y personas. Los modelos que consideran estos niveles o jerarquías son llamados: modelo multinivel, modelo de factores aleatorios, modelo mixto, modelo de coeficientes aleatorios y modelo jerárquico. Todos estos modelos son similares entre sí.

Los modelos básicos de áreas pequeñas se clasifican como modelos de factores aleatorios. Los factores fijos y aleatorios son utilizados en los modelos mencionados anteriormente y también son llamados coeficientes fijos y aleatorios en el caso de los modelos multinivel. Hox [Hox, 2002] describe los coeficientes fijos y aleatorios de la siguiente manera. Los factores aleatorios en un modelo hacen referencia a la variación entre los niveles. Por ejemplo, si se tienen datos de 2000 estudiantes de 100 secciones, en donde cada sección está a cargo de un profesor diferente, y el

modelo

$$popularidad_{ij} = \beta_{0j} + \beta_{1j} sexo_{ij} + e_{ij}$$

se utiliza para predecir la variable ”popularidad” tomando en cuenta la variable binaria ”sexo” a nivel de estudiantes y la variable ”años de experiencia del profesor” (Z_j) a nivel de sección, donde j representa la sección e i al estudiante. En el modelo, β_{0j} es el intercepto, β_{1j} es el coeficiente de regresión para la variable ”sexo” y e_{ij} es el término de error. En este caso, se asume que el intercepto β_{0j} y la pendiente β_{1j} varían dependiendo de la sección, por lo que se les llaman **factores aleatorios**.

La variable a nivel de sección se utiliza para describir la variación de β_{0j} y β_{1j} de la siguiente manera:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \tag{B.1}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}. \tag{B.2}$$

La ecuación B.1 predice la media de la popularidad en una sección a partir de la experiencia del profesor y la ecuación B.2 establece que la relación entre la popularidad y el sexo del estudiante depende de los años de experiencia del profesor. Los u_{ij} son los términos de error a nivel de sección. En las ecuaciones B.1 y B.2, los coeficientes de regresión γ no se asume que varíen entre secciones entonces, estos son llamados **factores fijos**.

Rao [Rao, 2003] clasifica los modelos de áreas pequeñas en dos tipos: (1) los modelos de área que relacionan las medias de las áreas pequeñas con las variables auxiliares a nivel de área, estos modelos son necesarios cuando no se poseen datos a nivel de unidades, y (2) los modelos a nivel de unidad que relacionan los datos de las unidades con las variables auxiliares a nivel de unidad. A continuación se describen los modelos básicos a nivel de área y unidad como fueron presentados por J.N.K. Rao [Rao, 2003] y Hidiroglou y You [Hidiroglou and You, 2016].

1. Modelo básico a nivel de unidad. El modelo básico a nivel de unidad, también llamado ”nested error regression model” en inglés, está dado por

$$y_{ij} = \boldsymbol{x}_{ij}^T \beta + v_i + e_{ij} \quad para \quad j = 1, \ldots, N_i, \quad i = 1, \ldots, m, \tag{B.3}$$

donde y_{ij} es la variable de interés para la j-ésima unidad de la población en el área pequeña i, $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$ es un vector de variables auxiliares de tamaño $p \times 1$ con $x_{ij1} = 1, \boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^T$ es un vector con los parámetros de regresión de tamaño $p \times 1$ y N_i es la cantidad de unidades de la población en el área pequeña i. Se asume que los factores aleatorios v_i son independientes e idénticamente distribuidos (i.i.d) $N(0, \sigma_v^2)$ e independientes de los errores e_{ij}, los cuales son i.i.d $N(0, \sigma_e^2)$. Suponiendo que N_i es grande, la media para el área i es $\overline{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$ y puede ser aproximada por

$$\theta_i = \overline{\boldsymbol{X}}_i^T \boldsymbol{\beta} + v_i,$$

donde $\overline{\boldsymbol{X}}_i = \sum_{j=1}^{N_i} \boldsymbol{x}_{ij}/N_i$ es el vector de medias poblacionales conocidas de $\boldsymbol{x}_{ij}$ para el área i. Se asume que se tomó una muestra independiente en cada área pequeña siguiendo cierto diseño de muestreo. Un diseño de muestreo que no es informativo quiere decir que la distribución de la muestra es la misma que la distribución de la población. En ese caso, los datos de la muestra $(y_{ij}, \boldsymbol{x}_{ij})$ siguen el modelo poblacional

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m,$$

donde n_i es el tamaño de la muestra en el área pequeña i.

2. Modelo básico a nivel de área. El modelo Fay-Herriot es un modelo básico a nivel de área que se utiliza para mejorar las estimaciones directas de encuestas. Este modelo tiene dos componentes. Sea $\hat{\theta}_i^{DIR}$ un estimador directo de θ_i. Asumiendo que $\hat{\theta}_i^{DIR}$ es un estimador insesgado de θ_i, el primer componente está dado por

$$\hat{\theta}_i^{DIR} = \theta_i + e_i, \quad i = 1, ..., m, \tag{B.4}$$

donde e_i es el error de muestreo asociado con el estimador directo $\hat{\theta}_i^{DIR}$ y m es el número de áreas pequeñas. En la práctica, usualmente se asume que los e_i son variables aleatorias normales e independientes con media $E(e_i) = 0$ y varianza de muestreo σ_i^2.

El segundo componente se obtiene al asumir que el parámetro de interés de las áreas pequeñas θ_i se relaciona con variables auxiliares a nivel de área $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})^T$ a través del siguiente

modelo de regresión lineal

$$\theta_i = \boldsymbol{z}_i^T \boldsymbol{\beta} + b_i v_i, \quad i = 1, ..., m, \tag{B.5}$$

donde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ es un vector de coeficientes de regresión de tamaño $p \times 1$, b_i son constantes positivas conocidas y los v_i son los factores aleatorios específicos de área, los cuales se asumen que son i.i.d con $E(v_i) = 0$ y varianza σ_v^2. Si se combinan los modelos B.4 y B.5 se obtiene el modelo lineal mixto a nivel de área

$$\hat{\theta}_i^{DIR} = \boldsymbol{z}_i^T \boldsymbol{\beta} + b_i v_i + e_i. \tag{B.6}$$

Ahora, haciendo las siguientes sustituciones

$$\boldsymbol{y}_i = \hat{\theta}_i^{DIR}, \quad \boldsymbol{X}_i = \boldsymbol{z}_i^T, \quad \boldsymbol{Z}_i = b_i$$

$$\boldsymbol{v}_i = v_i, \quad \boldsymbol{e}_i = e_i, \quad \boldsymbol{V}_i = \sigma_i^2 + \sigma_v^2 b_i^2$$

obtenemos el estimador BLUP (Best Linear Unbiased Predictor) del parámetro θ_i

$$\tilde{\theta}^{FH} = \gamma_i \hat{\theta}_i^{DIR} + (1 - \gamma_i) \boldsymbol{z}_i^T \tilde{\boldsymbol{\beta}}, \tag{B.7}$$

donde $\gamma_i = \sigma_v^2 b_i^2 / (\sigma_i^2 + \sigma_v^2 b_i^2)$ y $\tilde{\boldsymbol{\beta}}$ está dado por

$$\tilde{\boldsymbol{\beta}} = \left[\sum_{i=1}^{m} \frac{\boldsymbol{z}_i \boldsymbol{z}_i^T}{\sigma_i^2 + \sigma_v^2 b_i^2} \right]^{-1} \left[\sum_{i=1}^{m} \frac{\boldsymbol{z}_i \boldsymbol{y}_i}{\sigma_i^2 + \sigma_v^2 b_i^2} \right] \tag{B.8}$$

El estimador BLUP en B.7 depende de la varianza σ_v^2, la cual, en la práctica, no se conoce. Si se reemplaza σ_v^2 por un estimador $\hat{\sigma}_v^2$ obtenemos el estimador EBLUP (Empirical Best Linear Unbiased Predictor)

$$\hat{\theta}^{FH} = \hat{\gamma}_i \hat{\theta}_i^{DIR} + (1 - \hat{\gamma}_i) \boldsymbol{z}_i^T \hat{\boldsymbol{\beta}}, \tag{B.9}$$

donde $\hat{\gamma}_i$ y $\hat{\boldsymbol{\beta}}$ son los valores de γ_i y $\tilde{\boldsymbol{\beta}}$ cuando se reemplaza σ_v^2 por $\hat{\sigma}_v^2$.

Existen diferentes métodos que se utilizan para estimar σ_v^2. Utilizando el método de máxima verosimilitud restringida (REML, restricted maximum likelihood), el estimador de REML $\hat{\sigma}_v^2$ se

obtiene de la siguiente manera:

$$\sigma_v^{2(k+1)} = \sigma_v^{2(k)} + \left[I_R\left(\sigma_v^{2(k)}\right)\right]^{-1} S_R\left(\sigma_v^{2(k)}\right), \quad para \quad k = 1, 2, \ldots,$$

donde $I_R\left(\sigma_v^2\right) = 1/2tr[\boldsymbol{PBPB}]$, $S_R\left(\sigma_v^2\right) = 1/2\boldsymbol{y}^T\boldsymbol{PBPy} - 1/2tr[\boldsymbol{PB}]$,

$$\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1},$$

$$\boldsymbol{B} = diag(b_1^2, \ldots b_m^2).$$

Comenzando con un valor aleatorio para $\sigma_v^{2(1)}$ el algoritmo converge fácilmente.

El estimador del MSE de $\hat{\theta}_i^{FH}$ está dado por

$$mse\left(\hat{\theta}_i^{FH}\right) = g_{1i} + g_{2i} + 2g_{3i}, \tag{B.10}$$

donde g_{2i} representa la variabilidad debida a la estimación del parámetro de regresión β y g_{3i} la variabilidad debida a la estimación de la varianza σ_v^2. Estos términos son definidos de la siguiente manera

$$g_{1i} = \hat{\gamma}_i\sigma_i^2,$$

$$g_{2i} = (1-\hat{\gamma}_i)^2\boldsymbol{z}_i^T var(\hat{\boldsymbol{\beta}})\boldsymbol{z}_i = (1-\hat{\gamma}_i)^2\boldsymbol{z}_i^T\left(\sum_{i=1}^{m}\frac{\boldsymbol{z}_i\boldsymbol{z}_i^T}{\sigma_i^2+\sigma_v^2b_i^2}\right)^{-1}\boldsymbol{z}_i,$$

$$g_{3i} = (\sigma_i^2)^2(\hat{\sigma}_v^2+\sigma_i^2)^{-3}var(\hat{\sigma}_v^2)$$

y la varianza estimada de $\hat{\sigma}_v^2$ es

$$var(\hat{\sigma}_v^2) = 2\left(\sum_{i=1}^{m}(\hat{\sigma}_v^2+\sigma_i^2)^{-2}\right)^{-1}.$$

En la práctica, es común que se posean las estimaciones directas, s_i^2, de las varianzas, σ_i^2. Debido a que estas estimaciones s_i^2 son muy variables, se utilizan modelos externos y funciones para obtener varianzas suavizadas $\tilde{s}_i^2$. Estas varianzas suavizadas son las que se usan en el modelo Fay-Herriot. Para calcular el $mse\left(\hat{\theta}_i^{FH}\right)$ se reemplazan las σ_i^2 por las $\tilde{s}_i^2$ en la ecuación B.10. Debido a esto, se le debe agregar un término extra al estimador del MSE en B.10, debido a la incertidumbre por haber

utilizado las s_i^2. Este término es el siguiente

$$g_{4i} = \frac{4}{n_i - 1} \frac{\hat{\sigma}_v^4 s_i^4}{(\hat{\sigma}_v^2 + s_i^2)^3}.$$

Al utilizar el modelo de Fay-Herriot para calcular las estimaciones indirectas de la variable de interés, se necesitan las estimaciones directas a nivel de área de la variable de interés junto con las estimaciones de las varianzas como parámetros del modelo. Los métodos de estimación directa presentados en la sección anterior pueden ser utilizados para obtener estas estimaciones directas.

CAPÍTULO V: APLICACIÓN DE MÉTODOS DE ESTIMACIÓN EN ÁREAS PEQUEÑAS

En Guatemala, los informes de las encuestas de condiciones de vida en el país, en particular los de la ENCOVI y la ENSMI, son elaborados para reportar indicadores, a nivel nacional, regional y urbano/rural, que solicitan organismos como el Banco Mundial, la Organización Mundial de la Salud, el Centro para el Control y la Prevención de Enfermedades (CDC) y La Agencia de los Estados Unidos para el Desarrollo Internacional (USAID). Estos organismos promueven los programas de encuestas LSMS (Living Standards Measurement Study), RHS (Reproductive Health Surveys) y DHS (Demographic and Health Surveys).

El informe de la ENSMI 2002 presenta indicadores a nivel regional y urbano/rural, pero muchas veces se necesitan calcular indicadores a nivel departamental para poder repartir fondos o elaborar programas de una manera más adecuada. Uno de los objetivos principales de la estimación en áreas pequeñas es proporcionar estadísticas a los gobiernos o instituciones para que puedan planificar la asignación de recursos y tomar decisiones. Es por esto que las instituciones del país, como el Ministerio de Salud Pública y Asistencia Social (MSPAS) y las organizaciones de apoyo a la salud, necesitan contar con información para uso administrativo. Esta información puede obtenerse a partir de las bases de datos de estas encuestas por medio de estudios secundarios en los que se calculen indicadores a nivel departamental o municipal utilizando métodos de estimación en áreas pequeñas.

Este capítulo muestra la metodología para obtener estimaciones directas e indirectas de indicadores en áreas pequeñas utilizando la Encuesta Nacional de Salud Materno Infantil (ENSMI) de 2002 en Guatemala. Se escogió específicamente la ENSMI 2002 debido a que se tienen los indicadores presentados en el informe del Censo Nacional XI de Población y de VI de Habitación 2002 como variables auxiliares. Si se considera otra ENSMI, es necesario complementar la metodología con el

manejo de información de diferentes momentos en el tiempo o utilizar otra fuente de variables auxiliares que sea del mismo año; por ejemplo, una encuesta más grande. En este caso, se consideraron los departamentos como áreas pequeñas y se calcularon algunos indicadores a nivel departamental utilizando métodos indirectos.

A. ENSMI 2002

La Encuesta Nacional de Salud Materno Infantil (ENSMI 2002) es la cuarta encuesta de esta serie que se ha desarrollado desde 1987. Los objetivos de la encuesta fueron proporcionar información acerca de la salud materno infantil y reproductiva, lactancia y nutrición, enfermedades de transmisión sexual y violencia familiar. Esta serie de encuestas ha sido una de las fuentes de información más importantes en Guatemala, ya que proporcionan evidencia acerca de las condiciones de salud reproductiva de la población. La población objetivo de la encuesta fueron todos los miembros residentes en cada uno de los hogares, las mujeres y hombres elegibles en edad reproductiva (15 a 49 años) en cada hogar y los niños y niñas menores de 5 años residentes en el hogar.

La muestra de la encuesta se realizó para obtener indicadores representativos a nivel nacional, regional, urbano/rural e indígena/ladino. El Cuadro V.1 presenta las distintas regiones consideradas. El diseño de la encuesta fue probabilístico, estratificado, multietápico e independiente para cada región. El marco muestral que se utilizó en la primera etapa de selección provenía de la sectorización del X Censo de Población y V de Habitación de 1994.

Cuadro V.1: Regiones de Guatemala

No.	Región	Departamentos
1	Metropolitana	Guatemala
2	Norte	Alta Verapaz, Baja Verapaz
3	Nor-Oriente	El Progreso, Izabal, Zacapa, Chiquimula
4	Sur-Oriente	Santa Rosa, Jalapa, Jutiapa
5	Central	Chimaltenango, Sacatepéquez, Escuintla
6	Sur-Occidente	Sololá, Totonicapán, Quetzaltenango, Suchitepéquez, Retahuleu, San Marcos
7	Nor-Occidente	Huehuetenango, Quiché
8	Petén	Petén

Para esta encuesta se obtuvo dos muestras de hogares. La primera muestra se utilizó para entrevistar a todas las mujeres en edad reproductiva y la segunda para entrevistar a hombres en el mismo segmento cartográfico [1], pero en diferentes hogares. Para la encuesta de hombres se seleccionaron 4,033 hogares y para la de mujeres 12,119. En total, se tienen las entrevistas completas de 2,538 hombres y 9,155 mujeres, todos en edad reproductiva (15 a 49 años). Se entrevistó a una mujer en edad reproductiva por cada hogar y se recopiló información de cada uno de sus hijos nacidos vivos.

1. Diseño de la muestra. La ENSMI 2002 tuvo el mismo marco muestral que se utilizó para la ENSMI 1995 y la ENSMI 1998-1999. Al inicio de la elaboración de la muestra, los hogares fueron considerados parte de las viviendas; esto quiere decir que una vivienda podía poseer uno o más hogares. Luego de hacer la actualización cartográfica de los segmentos censales, las viviendas ya no fueron consideradas y los segmentos censales, hogares y las personas fueron las únicas unidades de muestreo, de las cuales, los segmentos censales fueron las unidades primarias de muestreo. El marco muestral fue estratificado por departamento y el tamaño de la muestra se determinó para obtener indicadores representativos a nivel regional. Dentro de cada estrato, los segmentos fueron seleccionados en forma sistemática con probabilidad proporcional al tamaño (número de hogares). La muestra se seleccionó en tres etapas. En la **primera etapa** se tomó una submuestra de los segmentos censales (UPM) utilizados en la muestra de la ENSMI 1995, y en el caso de Petén, una submuestra de los segmentos censales utilizados en la muestra de la ENSMI 1998-1999. La muestra de la ENSMI 2002 contiene 376 segmentos censales y la fracción de muestreo para cada departamento fue diferente.

En la **segunda etapa**, se eligió un número predeterminado de hogares en cada segmento censal elegido en la primera etapa. De manera aleatoria (todos los hogares en el segmento tuvieron igual probabilidad de ser seleccionados), se eligió un hogar como punto de partida entre el hogar número 1 y la n, n era el total de hogares dentro del segmento. Los hogares incluidos en la muestra fueron el hogar del inicio y los consecutivos en el listado de hogares. En total se eligieron 30 hogares por segmento para la entrevista de mujeres y 10 por segmento para la entrevista de hombres. En el departamento de Guatemala, para cada segmento en la muestra, se seleccionaron dos conglomerados de 24 hogares para la entrevista de mujeres y 8 hogares para la entrevista de hombres. Esta selección fue realizada de manera aleatoria.

[1] A partir de la ENSMI 2008 se les llamó sectores cartográficos. Los segmentos cartográficos pasaron a ser áreas comprendidas dentro de los sectores cartográficos.

Para ambas áreas, rural y urbana, no se tomaron en cuenta los hogares deshabitados, destruidos o en construcción. Se definió un hogar como una edificación o inmueble que tuviera acceso independiente y estuviera habitado. Si el hogar era habitado por más de una persona, todas las personas debían compartir la misma alimentación. Tampoco se tomaron en cuenta edificaciones utilizadas exclusivamente para: fines productivos, comerciales o oficina, ni hogares colectivos como: conventos, internados, guarniciones militares, hoteles, etcétera.

En los hogares seleccionados se llenó un cuestionario del hogar, y en cada hogar en donde se completó este cuestionario, se llevó a cabo la **tercera etapa** de selección de la muestra. La tercera etapa consistió en elegir de manera aleatoria una mujer en edad reproductiva por cada hogar. En el caso de los hogares para hombres, también se seleccionó de manera aleatoria un hombre entre 15 y 49 años para ser entrevistado. El proceso de selección fue el siguiente: de primero se registró la información de las personas que residían en el hogar. Luego, se enumeraron todas las mujeres de 15 a 49 años de edad comenzando con la de mayor edad y terminando con la menor (lo mismo para los hogares para la entrevista de hombres). Al final del cuestionario del hogar se encontraba un Cuadro en donde el eje horizontal contenía las posibles cantidades de mujeres de 15 a 49 años en el hogar y el eje vertical contenía los números del 1 al 9, los cuales representaban el último dígito del número de cuestionario. Los números con los que se enumeraron a cada mujer del hogar estaban anotados de manera aleatoria dentro del área comprendida entre los ejes. Al intersectar el total de mujeres de 15 a 49 años en un hogar con el último dígito del número de cuestionario del hogar, se obtenía el número de la mujer o hombre al que se iba a entrevistar.

La probabilidad de selección de cada mujer entrevistada es inversamente proporcional a la cantidad de mujeres en edad fértil en el hogar. El peso utilizado para análisis de las variables en el cuestionario de las mujeres (PesoMEF) es el peso del hogar multiplicado por el número de mujeres en edad fértil en el hogar. En el caso de los análisis para obtener indicadores de los hijos de las mujeres entrevistadas, el peso que se utiliza es el mismo (PesoMEF), ya que se midieron a todos los hijos de cada mujer encuestada. [MSPAS, 2003]

2. Desnutrición crónica. La ENSMI 2002 fue una encuesta RHS, el cual es un programa a cargo del CDC. Los lineamientos que se siguen en las encuestas RHS y DHS son muy similares,

por lo que la metodología que se siguió para obtener el indicador de niños con desnutrición crónica fue la de las encuestas DHS [Rutstein and Rojas, 2006].

Las DHS mencionan tres tipos de desnutrición: el retardo en el crecimiento, basado en la altura (o talla) y edad de un niño, es una medida de **desnutrición crónica**. La emaciación, basada en el peso y altura de un niño, es una medida de **desnutrición aguda**; y el bajo peso, basado en el peso para la edad del niño, es una medida de **desnutrición global**. Los indicadores antropométricos que se utilizan para identificar si un niño posee alguno de estos tipos de desnutrición son: la relación de Talla/Edad, Peso/Talla y Peso/Edad. Al obtener las mediciones de talla y peso de los niños, se asigna un puntaje z expresado en desviaciones estándar (DE). Los puntajes z se utilizan para describir la distancia que existe entre una medición y el promedio esperado según una población de referencia, y se les asignan a cada una de las tres relaciones: Talla/Edad, Peso/Talla y Peso/Edad.

En la metodología de las DHS, para la asignación de puntajes z se utiliza el Estándar de Referencia Internacional NCHS/CDC/WHO, el cual se basa en niños bien nutridos en Estados Unidos. A partir del 2006 la Organización Mundial de la Salud introdujo una nueva población de referencia, la cual se obtuvo a partir de un estudio realizado a finales de la década de los 90 en 6 diferentes países del mundo (Brasil, Estados Unidos, Ghana, India, Noruega y Omán), el cual tenía como objetivo describir la forma en que los niños y niñas deben crecer si están bien nutridos. [2]

La asignación de puntajes z se realiza a través de una función de interpolación que toma en cuenta el sexo, edad (medida como la diferencia entre la fecha de nacimiento y la fecha de la entrevista, tiene que ser precisa al día del mes), altura en centímetros y el peso en kilogramos. Debido a las variaciones naturales en una población bien nutrida, 2.2 % de los niños van a estar entre -2.0 y -2.99 desviaciones estándar por debajo de la media y 0.1 % van a estar -3.0 o más desviaciones estándar por debajo de la media. El grado de malnutrición de una población se mide en base a si se exceden estos porcentajes que ocurren en una población de niños bien nutridos.

Los puntajes z son muy sensibles a cambios de edad, por lo que a los niños que tienen fechas de nacimiento incompletas no se les asignan. Los niños con puntaje z de Talla/Edad por debajo de -6 DE o por arriba de +6 DE, con puntaje z de Peso/Edad por debajo de -6 DE o por arriba de +6

[2] El informe de la ENSMI 2008 presenta una comparación de las poblaciones de referencia de NCHS y OMS.

DE, o con puntaje z de Peso/Talla por debajo de -4 DE o por arriba de +6 DE se consideran que poseen datos inválidos. También se consideran inválidas las siguientes combinaciones: puntaje z de Talla/Edad menor que -3.09 DE y puntaje z de Peso/Edad mayor que +3.09 DE, o puntaje z de Talla/Edad mayor que +3.09 DE y puntaje z de Peso/Edad menor que -3.09 DE.

Las DHS consideran las siguientes clasificaciones de desnutrición en base a los puntajes z de los niños:

- **Desnutrición crónica severa:** el puntaje z de Talla/Edad es menor que -3.0 DE por debajo de la media basada en el Estándar de Referencia Internacional NCHS/CDC/WHO.
- **Desnutrición crónica moderada:** el puntaje z de Talla/Edad está entre -2.0 y -2.99 DE por debajo de la media basada en el Estándar de Referencia Internacional NCHS/CDC/WHO.
- **Desnutrición aguda severa:** el puntaje z de Peso/Talla es menor que -3.0 DE por debajo de la media basada en el Estándar de Referencia Internacional NCHS/CDC/WHO.
- **Desnutrición aguda moderada:** el puntaje z de Peso/Talla está entre -2.0 y -2.99 DE por debajo de la media basada en el Estándar de Referencia Internacional NCHS/CDC/WHO.
- **Bajo peso severo:** el puntaje z de Peso/Edad es menor que -2.0 DE por debajo de la media basada en el Estándar de Referencia Internacional NCHS/CDC/WHO.
- **Bajo peso moderado:** el puntaje z de Peso/Edad está entre -2.0 y -2.99 DE por debajo de la media basada en el Estándar de Referencia Internacional NCHS/CDC/WHO.

[Rutstein and Rojas, 2006]

El informe de la ENSMI 2002 reportó el porcentaje de niños con desnutrición crónica total, el cual se calculó sobre los niños vivos entre 3 y 59 meses de edad. La desnutrición crónica total comprende la desnutrición crónica severa y moderada, por lo que el puntaje z de Talla/Edad debe ser menor que -2 DE.

B. Indicador de niños con desnutrición crónica

A continuación se explica la metodología que se empleó para el cálculo del indicador de niños con desnutrición crónica en cada departamento y un análisis de los resultados departamentales en el cual se discuten (1) la relación entre los indicadores departamentales en una misma región y (2) los márgenes de error (intervalos de confianza) de los indicadores departamentales y regionales. Los cálculos se realizaron utilizando la base de datos (gt02f_hijos.sav) de la ENSMI 2002 proporcionada en la página Global Health Data Exchange, el lenguaje R y el software RStudio. Las funciones que se utilizaron fueron obtenidas del paquete $survey$, que se utiliza para el análisis de muestras complejas de encuestas, y el paquete sae, que se utiliza para estimación en áreas pequeñas. Para una descripción mucho más detallada de la metodología y funciones empleadas se puede consultar el Apéndice A.

1. Metodología. La metodología para el cálculo del indicador de desnutrición crónica consistió en lo siguiente:

- Calcular las estimaciones directas del indicador para cada región y departamento utilizando las funciones del paquete $survey$ de R, las cuales dan los mismos resultados que la ecuación C.1 de la sección C de Estimación en Dominios, y luego compararlas con las presentadas en los informes de la ENSMI 2002 y ENSMI 2014.
- Utilizar las estimaciones directas mencionadas anteriormente y el modelo Fay-Herriot para calcular las estimaciones indirectas del indicador con su error error estándar e intervalo de confianza para cada departamento.

a. Estimación directa. Las estimaciones directas a nivel regional y departamental se hicieron sobre todos los niños y niñas entre 3 y 59 meses de edad tomando en cuenta lo presentado en la sección 2 de Desnutrición Crónica. Se eligió utilizar las funciones $svymean$ y $svyby$, del paquete $survey$ de R, debido al diseño complejo de la muestra y el uso de los departamentos (áreas pequeñas) como estratos. También se puede utilizar la ecuación C.1 de la sección C de Estimación en Dominios, ya que da los mismos resultados que la función $svymean$. Los detalles de la implementación de estas funciones se pueden ver en el Apéndice A y en el código 5 del Apéndice E.

Los indicadores regionales y departamentales obtenidos se muestran en los Cuadros V.2 y V.3. El Cuadro 5 contiene los porcentajes regionales reportados en el informe de la ENSMI 2002 y el Cuadro 7 contiene los reportados por la ENSMI 2014. Los errores estándar, intervalos de confianza y

efectos del diseño reportados por la ENSMI 2002 se pueden observar en el Cuadro 6.

Cuadro V.2: Estimaciones directas de los indicadores de desnutrición crónica por región para niños y niñas entre 3 y 59 meses de edad

Región	Valor estimado (V)	Error estándar (EE)	No. de casos		Efecto del diseño (EDIS)	Intervalo de confianza	
			Sin ponderar (SP)	Ponderados (P)		V-2EE	V+2EE
Metropolitana	0.361	0.044	590	1085.2	2.231	0.275	0.448
Norte	0.607	0.034	918	461.2	2.147	0.539	0.674
Nor-Oriente	0.397	0.048	503	465.0	2.209	0.303	0.491
Sur-Oriente	0.463	0.035	449	448.6	1.509	0.394	0.532
Central	0.421	0.045	702	491.9	2.455	0.332	0.510
Sur-Occidente	0.586	0.022	1559	948.1	1.758	0.543	0.628
Nor-Occidente	0.683	0.028	1072	529.4	1.957	0.629	0.737
Petén	0.461	0.052	543	188.6	2.471	0.359	0.563

b. Estimación indirecta. Los métodos de estimación indirecta hacen uso de información adicional en la forma de variables auxiliares para obtener estimaciones más precisas. Los indicadores presentados en el informe del Censo Nacional XI de Población y de VI de Habitación 2002 [INE, 2003] se utilizaron como variables auxiliares y estos se muestran en el Cuadro 4 del Apéndice D. La función $mseFH$ se utilizó para calcular las estimaciones indirectas del indicador de desnutrición crónica; esta función hace estimaciones en dominios basándose en el modelo Fay-Herriot y también calcula los MSE.

Para escoger cuáles variables auxiliares incluir en el modelo se implementaron tres métodos de selección de variables: **selección hacia adelante**, **eliminación hacia atrás** y **selección por pasos**. En cada uno de los métodos se crearon diferentes modelos con diferentes variables auxiliares utilizando la función $mseFH$ y se analizaron los valores-p de los coeficientes de cada modelo. Luego se eligieron los modelos con las varibles auxiliares cuyos coeficientes tenían los menores valores-p. También es importante mencionar que, al momento de construir un modelo, las variables auxiliares no deben correlacionarse entre sí. Por último, para comparar modelos y elegir el que mejor se adecuaba a los datos, se utilizó el Criterio de Información de Akaike (Akaike Information Criterion, AIC) y el Criterio de Información Bayesiano (Bayesian Information Criterion, BIC). Las secciones 2 y 3 del

Cuadro V.3: Estimaciones directas de los indicadores de desnutrición crónica por departamento para niños y niñas entre 3 y 59 meses de edad

Departamento	Valor estimado (V)	Error estándar (EE)	No. de casos		Efecto del diseño (EDIS)	Intervalo de confianza		Varianza (Var)
			Sin ponderar (SP)	Ponderados (P)		V-2EE	V+2EE	
Guatemala	0.361	0.044	590	1085.2	2.231	0.275	0.448	0.0019
El Progreso	0.319	0.074	119	49.7	1.753	0.173	0.464	0.0055
Sacatepequez	0.506	0.092	98	69.7	1.833	0.325	0.687	0.0085
Chimaltenango	0.627	0.060	277	171.3	2.069	0.510	0.744	0.0036
Escuintla	0.256	0.054	327	250.9	2.232	0.151	0.361	0.0029
Santa Rosa	0.398	0.069	94	175.6	1.360	0.263	0.533	0.0047
Solola	0.689	0.048	183	85.2	1.420	0.594	0.783	0.0023
Totonicapan	0.821	0.041	310	153.0	1.885	0.741	0.900	0.0017
Quetzaltenang	0.470	0.042	248	225.8	1.324	0.388	0.552	0.0017
Suchitepequez	0.457	0.044	269	136.5	1.445	0.372	0.543	0.0019
Retalhuleu	0.449	0.046	178	77.2	1.238	0.360	0.539	0.0021
San Marcos	0.621	0.053	371	270.4	2.110	0.517	0.725	0.0028
Huehuetenang	0.690	0.034	567	309.3	1.745	0.625	0.756	0.0011
Quiche	0.673	0.046	505	220.1	2.245	0.582	0.764	0.0021
Baja Verapaz	0.453	0.082	194	86.2	2.310	0.293	0.613	0.0067
Alta Verapaz	0.642	0.036	724	375.0	2.044	0.571	0.713	0.0013
Petén	0.461	0.052	543	188.6	2.471	0.359	0.563	0.0027
Izabal	0.422	0.093	149	192.2	2.301	0.240	0.605	0.0087
Zacapa	0.082	0.040	28	60.9	0.758	0.004	0.160	0.0016
Chiquimula	0.510	0.053	207	162.3	1.536	0.406	0.615	0.0028
Jalapa	0.560	0.064	249	126.6	2.062	0.434	0.686	0.0041
Jutiapa	0.457	0.044	106	146.4	0.911	0.370	0.543	0.0019

Apéndice A describen la implementación de estos métodos y el análisis de los valores-p y los criterios.

Al tener el modelo final se prosiguió a calcular los indicadores departamentales de desnutrición crónica en niños y niñas de 3 a 59 meses de edad. El Cuadro V.4 contiene los resultados obtenidos junto con los errores estándar, intervalos de confianza y varianzas.

Cuadro V.4: Estimaciones indirectas por departamento de los indicadores de desnutrición crónica para niños y niñas entre 3 y 59 meses de edad

Departamento	Valor estimado (V)	Error estándar (EE)	No. de casos		Intervalo de confianza		Varianza (Var)
			Sin ponderar (SP)	Ponderados (P)	V-2EE	V+2EE	
Guatemala	0.367	0.043	590	1085.2	0.281	0.453	0.0019
El Progreso	0.328	0.066	119	49.7	0.197	0.459	0.0043
Sacatepequez	0.451	0.077	98	69.7	0.297	0.605	0.0059
Chimaltenango	0.606	0.055	277	171.3	0.496	0.716	0.0030
Escuintla	0.281	0.050	327	250.9	0.181	0.381	0.0025
Santa Rosa	0.396	0.062	94	175.6	0.273	0.519	0.0038
Solola	0.666	0.045	183	85.2	0.575	0.756	0.0021
Totonicapan	0.793	0.039	310	153.0	0.715	0.871	0.0015
Quetzaltenango	0.473	0.040	248	225.8	0.393	0.554	0.0016
Suchitepequez	0.467	0.042	269	136.5	0.384	0.550	0.0017
Retalhuleu	0.450	0.044	178	77.2	0.362	0.537	0.0019
San Marcos	0.619	0.050	371	270.4	0.519	0.718	0.0025
Huehuetenango	0.691	0.033	567	309.3	0.625	0.757	0.0011
Quiche	0.678	0.044	505	220.1	0.590	0.767	0.0020
Baja Verapaz	0.460	0.070	194	86.2	0.321	0.600	0.0048
Alta Verapaz	0.654	0.035	724	375.0	0.584	0.725	0.0013
Petén	0.471	0.049	543	188.6	0.373	0.569	0.0024
Izabal	0.424	0.076	149	192.2	0.273	0.576	0.0058
Zacapa	0.112	0.038	28	60.9	0.035	0.189	0.0015
Chiquimula	0.496	0.050	207	162.3	0.397	0.595	0.0025
Jalapa	0.528	0.058	249	126.6	0.412	0.644	0.0034
Jutiapa	0.450	0.042	106	146.4	0.366	0.534	0.0018

C. Análisis de los resultados

Los indicadores directos de desnutrición crónica por región que se calcularon utilizando las funciones mencionadas anteriormente fueron muy similares a los presentados en el informe de la ENSMI 2002 (Cuadro 6), por lo que se puede decir que se utilizó básicamente el mismo procedimiento para calcular los indicadores directos que el utilizado por los investigadores de la ENSMI 2002. Por otro lado, los indicadores directos regionales sí varían de los presentados en el informe de la ENSMI 2014 (Cuadro 7), siendo los últimos menores en la mayoría de regiones. Los indicadores

departamentales obtenidos (Cuadro V.3) también varían un poco de los presentados en el informe de la ENSMI 2014 (Cuadro 7).

Las diferencias entre los indicadores de desnutrición crónica de la ENSMI 2002 y la ENSMI 2014 pueden ser debido a la muestra de la ENSMI 2014, ya que fue diseñada para obtener indicadores confiables a nivel departamental, lo que la hace más grande que la de la ENSMI 2002 comprendiendo 864 segmentos censales, mientras que la ENSMI 2002 solamente tomó en cuenta 376 [MSPAS, 2017]. También es importante recordar que, cómo se mencionó anteriormente, a partir del 2006 hubo un cambio en la población de referencia y, debido a que en la nueva población de referencia se tomaron en cuenta más países, los indicadores de desnutrición en Guatemala se esperan que sean menores a los calculados tomando en cuenta la población de referencia del 2002, ya que esta estaba basada en niños bien nutridos de Estados Unidos.

Cada uno de los indicadores regionales del Cuadro V.2 puede ser considerado como un promedio de los indicadores de los departamentos de una misma región. La región Central tiene un indicador de desnutrición del 0.421, pero los departamentos que la conforman, Sacatepéquez, Chimaltenango y Escuintla, tienen indicadores muy diferentes (0.451, 0.606 y 0.281) entre sí. Lo mismo ocurre en las regiones Nor-Oriente y Sur-Occidente. Los indicadores de las regiones Nor-Oriente y Sur-Occidente son 0.397 y 0.586, respectivamente; pero el departamento de Zacapa, el cual pertenece a la región Nor-Oriente junto con Izabal, Chiquimula y El Progreso, tiene un valor muy alejado de los demás departamentos. En el caso de la región Sur-Occidente, los departamentos que la conforman, Sololá, Totonicapán, Quetzaltenango, Suchitepéquez, Retalhuleu y San Marcos, poseen indicadores entre 0.450 y 0.793. El informe de la ENSMI 2002 solamente reporta indicadores a nivel regional, por lo que un director departamental del MSPAS sólo conoce el promedio regional de desnutrición crónica, pero desconoce si su departamento está por arriba o abajo de ese promedio. Esto muestra la variación que puede existir entre los indicadores de una misma región y la importancia de obtener estimaciones en áreas pequeñas para poder implementar programas de ayuda de una mejor manera.

Los intervalos de confianza son considerados los márgenes de error de las estimaciones e indican qué tan confiables son los indicadores. Los indicadores departamentales tienen un margen de error mayor a los regionales. Esto se debe a que el diseño de la muestra de la ENSMI 2002 fue realizado para obtener estimaciones confiables a nivel regional y urbano/rural. En el diseño de la muestra de la

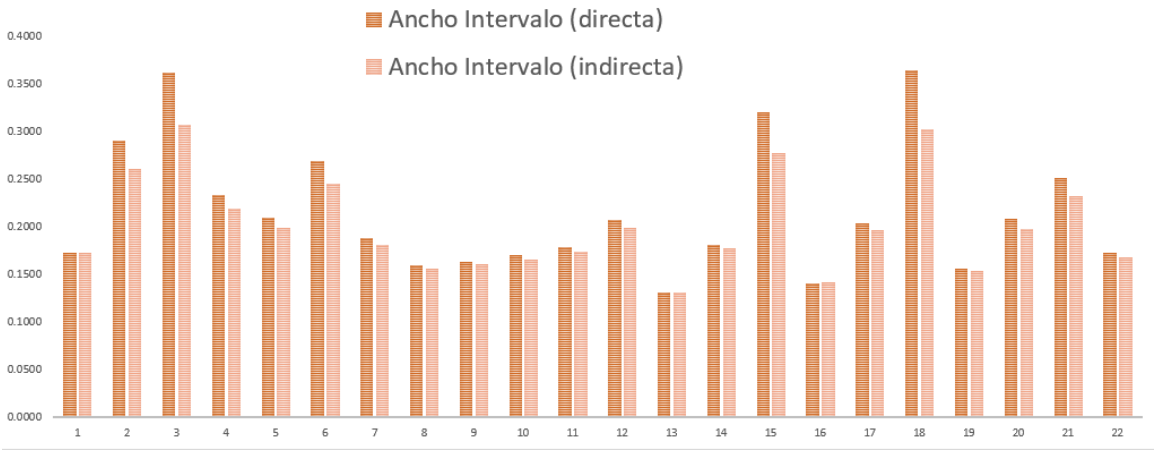
ENSMI 2002 se consideraron los departamentos (áreas pequeñas) como estratos, lo cual aseguró contar con una cantidad determinada de segmentos censales en cada departamento, pero el número fue tan reducido que no se consideró reportar indicadores departamentales. Esta particularidad de la ENSMI 2002 hace posible calcular estimaciones directas para todos los departamentos con márgenes de error que no son tan malos. Sin embargo, los métodos de Estimación en Áreas Pequeñas permiten obtener indicadores con errores estándar menores al hacer uso de información auxiliar y del modelo de regresión en donde se utiliza la información de todos los departamentos para el cálculo de los indicadores departamentales.

El Cuadro V.5 contiene los resultados de ambos tipos de estimaciones y la Figura V.1 muestra la diferencia entre los márgenes de error de ambas. De esto se puede observar que los errores estándar y varianzas son menores para las estimaciones indirectas obtenidas, lo que causa que el ancho de los intervalos de confianza se reduzca y que las estimaciones indirectas sean más precisas y exactas.

Cuadro V.5: Estimaciones directas e indirectas de los indicadores de desnutrición crónica para niños y niñas entre 3 y 59 meses de edad

Departamento	Valor estimado (V) (directo)	Valor estimado (V) (indirecto)	Error estándar (EE) (directo)	Error estándar (EE) (indirecto)	Ancho de intervalo (directo)	Ancho de intervalo (indirecto)
Guatemala	0.361	0.367	0.044	0.043	0.173	0.173
El Progreso	0.319	0.328	0.074	0.066	0.291	0.262
Sacatepequez	0.506	0.451	0.092	0.077	0.362	0.308
Chimaltenango	0.627	0.606	0.060	0.055	0.234	0.219
Escuintla	0.256	0.281	0.054	0.050	0.210	0.200
Santa Rosa	0.398	0.396	0.069	0.062	0.270	0.246
Solola	0.689	0.666	0.048	0.045	0.189	0.182
Totonicapan	0.821	0.793	0.041	0.039	0.160	0.156
Quetzaltenango	0.470	0.473	0.042	0.040	0.164	0.161
Suchitepequez	0.457	0.467	0.044	0.042	0.171	0.166
Retalhuleu	0.449	0.450	0.046	0.044	0.179	0.174
San Marcos	0.621	0.619	0.053	0.050	0.207	0.200
Huehuetenango	0.690	0.691	0.034	0.033	0.132	0.131
Quiche	0.673	0.678	0.046	0.044	0.182	0.178
Baja Verapaz	0.453	0.460	0.082	0.070	0.321	0.278
Alta Verapaz	0.642	0.654	0.036	0.035	0.142	0.142
Petén	0.461	0.471	0.052	0.049	0.204	0.197
Izabal	0.422	0.424	0.093	0.076	0.365	0.303
Zacapa	0.082	0.112	0.040	0.038	0.156	0.154
Chiquimula	0.510	0.496	0.053	0.050	0.208	0.198
Jalapa	0.560	0.528	0.064	0.058	0.252	0.233
Jutiapa	0.457	0.450	0.044	0.042	0.173	0.168

Figura V.1: Ancho de los intervalos de las estimaciones directas e indirectas



CAPÍTULO VI: CONCLUSIONES

A pesar que se pudo obtener estimaciones directas con errores estándar relativamente pequeños debido al diseño de la muestra de la ENSMI 2002, los métodos de Estimación en Áreas Pequeñas hacen posible el cálculo de estimaciones con errores estándar significativamente menores.

Al momento de hacer estimaciones indirectas es conveniente utilizar una fuente de variables auxiliares que sea del mismo año. Si se considera otra ENSMI junto con el Censo 2002, es necesario complementar la metodología con el manejo de información de diferentes momentos en el tiempo o utilizar otra fuente de variables auxiliares que sea del mismo año, como una encuesta más grande. Si en caso se utiliza una encuesta más grande en vez de un censo como fuente de variables auxiliares, también hay que tomar en cuenta los errores, ya que, a diferencia de los censos, los indicadores poseen errores adjuntos debido a las muestras de las encuestas.

Este trabajo puede ser utilizado por los investigadores que trabajan en encuestas nacionales para conocer la aplicación de los métodos de Estimación en Áreas Pequeñas y su importancia. De esta manera, ellos pueden recrear los cálculos presentados en el apéndice A utilizando encuestas y censos más recientes para así obtener indicadores en áreas pequeñas, como son los municipios.

CAPÍTULO VII: BIBLIOGRAFÍA

[A.Marker, 1999] A.Marker, D. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, Vol. 15.

[Brant, 2004] Brant, R. (2004). Automatic variable selection procedures. `https://www.stat.ubc.ca/~rollin/teach/643w04/lec/node40.html`.

[Fabozzi *et al.*, 2014] Fabozzi, F. J., Focardi, S. M., Rachev, S. T., and Arshanapalli, B. G. (2014). *The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications*, chapter Appendix E. John Wiley and Sons, Inc., Hoboken, New Jersey.

[Hidiroglou and You, 2016] Hidiroglou, M. A. and You, Y. (2016). Survey methodology: Comparison of unit level and area level small area estimators. `http://www.statcan.gc.ca/pub/12-001-x/2016001/article/14540-eng.pdf`.

[Hox, 2002] Hox, J. (2002). *Multilevel Analysis: Techniques and Applications*. Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey.

[INE, 2003] INE (2003). *Censos Nacionales XI de Población y VI de Habitación 2002: Características de la Población y de los Locales de Habitación Censados*. Fondo de Población de las Naciones Unidas (UNFPA), Guatemala.

[Lohr, 1999] Lohr, S. (1999). *Samplying: Design and Analysis*. Brooks/Cole Publishing Company, Pacific Grove, California.

[Molina and Marhuenda, 2015a] Molina, I. and Marhuenda, Y. (2015a). Package 'sae'. `https://cran.r-project.org/web/packages/sae/sae.pdf`.

[Molina and Marhuenda, 2015b] Molina, I. and Marhuenda, Y. (2015b). R package sae: Methodology. `https://cran.r-project.org/web/packages/sae/vignettes/sae_methodology.pdf`.

[MSPAS, 2003] MSPAS (2003). *Encuesta Nacional de Salud Materno Infantil 2002*. Guatemala.

[MSPAS, 2017] MSPAS (2017). *Encuesta Nacional de Salud Materno Infantil 2014*. Guatemala.

[Rao, 2003] Rao, J. (2003). *Small Area Estimation*. John Wiley and Sons, Inc., Hoboken, New Jersey.

[Rutstein and Rojas, 2006] Rutstein, S. O. and Rojas, G. (2006). *Guide to DHS Statistics*. Calverton, Maryland.

CAPÍTULO VIII: APÉNDICE

A. Descripción de cálculos y funciones utilizadas

Todos los cálculos se realizaron utilizando la base de datos (gt02f_hijos.sav) de la ENSMI 2002 proporcionada en la página Global Health Data Exchange, algunos de los indicadores presentados en el informe del Censo Nacional XI de Población y de VI de Habitación 2002 y el lenguaje R y software RStudio. R cuenta con dos paquetes: el paquete $survey$, que se utiliza para el análisis de muestras complejas de encuestas, y el paquete sae, que se utiliza para estimación en áreas pequeñas. Las funciones que se pueden utilizar para el cálculo de estimaciones directas son: $svymean$ del paquete $survey$, o la función $direct$ del paquete sae. Para la estimación indirecta, las funciones que se pueden utilizar son: $eblupFH$ y $mseFH$ del paquete sae.

B. Estimación directa

La base de datos de niños (gt02f_hijos.sav) contiene la información de todos los hijos nacidos vivos de las mujeres en edad reproductiva entrevistadas. Para el cálculo de este indicador se utilizaron las siguientes variables de la base de datos:

- vivo: ”Si”, si el niño está vivo; ”No”, si el niño murió
- mnac: Mes de nacimiento, ”98” si es inválido
- anac: Año de nacimiento, ”9998” si es inválido
- hedad: Meses cumplidos desde nacido
- HAZ: puntaje z de Talla/Edad
- WAZ: puntaje z de Peso/Edad
- WHZ: puntaje z de Peso/Talla
- mpaquete: número de paquete, representa el segmento censal (unidades primarias de muestreo)

- PesoMEF: la variable de peso que se debe utilizar en todos los cálculos para mujeres con entrevista completa. El mismo peso se utiliza para los cálculos de niños, ya que se recopiló información de todos los hijos de las mujeres entrevistadas.
- MHDEPTO: departamento

De primero se comenzó reduciendo la base de datos a todos los niños vivos entre 3 y 59 meses de edad con variables $mnac$, $anac$, HAZ, WAZ y WHZ válidas. Como se mencionó anteriormente, se consideró desnutrido a cualquier niño con puntaje z de Talla/Edad (HAZ) menor que -2 y el resultado se codificó en la nueva variable $desnu$ (1 si el niño está desnutrido y 0 si no lo está).

Para obtener las estimaciones directas del indicador de niños con desnutrición crónica total se pueden utilizar dos funciones: $svymean$ del paquete $Survey$ de R, o la función $direct$ del paquete Sae de R. La función $svymean$ considera los pesos y la estructura compleja de la muestra, mientras que la función $direct$ asume un muestreo aleatorio simple. En la práctica, las áreas pequeñas pueden ser muy reducidas, por lo que no son consideradas en el diseño de la muestra (un ejemplo de esto son los municipios en Guatemala). En el caso de la ENSMI 2002, los departamentos fueron considerados como estratos y las probabilidades de selección y los pesos fueron calculados de acuerdo a ellos, por lo que se decidió utilizar la función $svymean$, como se indica en el código 5 del Apéndice E.

Para hacer uso de la función $svymean$ fue necesario definir el diseño de la encuesta utilizando la función $svydesign$ del mismo paquete. Esta función recibe los siguientes parámetros: id = unidades primarias de muestreo, $strata$ = los estratos (departamentos), $weights$ = pesos y $data$ = la tabla de datos. Esta función guarda la información del diseño de la encuesta para luego hacer estimaciones.

La ecuación C.1 de la sección C Estimación en Dominios también puede utilizarse para obtener las estimaciones directas de niños con desnutrición crónica. Para esto, tenemos el estimador de razón

$$\hat{\bar{y}}_d = \frac{\sum_{i \in S} w_i x_{id} y_i}{\sum_{i \in S} w_i x_{id}}, \tag{B.1}$$

donde los w_i van a ser los pesos de cada niño (PesoMEF), x_{id} va a ser 1 si el niño se encuentra en el dominio d y 0 si está en otro dominio, y y_i va a ser 1 si el niño es considerado desnutrido y 0 si no lo está. El código 6 del Apéndice E muestra la implementación de esta ecuación. Este método da los

mismos resultados que la función $svymean$.

La función $svyby$ reporta la estimación obtenida en subconjuntos utilizando la función $svymean$, el error estándar, el intervalo de confianza, la varianza de la estimación y el efecto del diseño. El efecto del diseño (EDIS) reportado en el informe de la ENSMI 2002 es la razón entre el error estándar correspondiente al diseño de la muestra empleado y el error estándar que se obtiene tratando la muestra como si hubiera sido aleatoria simple. El DEFF que reporta R es la razón entre la varianza correspondiente al diseño empleado de la muestra y la varianza que se obtiene tratando la muestra como si hubiera sido aleatoria simple. Por lo tanto, como el error estándar es la raíz de la varianza entonces,

$$EDIS = \sqrt{DEFF}.$$

C. Estimación indirecta

1. Variables auxiliares. Para obtener las estimaciones indirectas, se utilizaron indicadores presentados en el informe del Censo Nacional XI de Población y de VI de Habitación 2002 como variables auxiliares. Es importante mencionar que ambos, el Censo Nacional como la ENSMI, son del mismo año. Los indicadores que se consideraron fueron los siguientes:

- TIERRA: Hogares cuyo material predominante en el piso es tierra
- EXAGUA: Hogares con chorro de agua de uso exclusivo
- SANI: Hogares que poseen servicio sanitario
- TSANI: Hogares que poseen servicio sanitario conectado a una red de drenajes o a una fosa séptica
- EXSANI: Hogares que poseen servicio sanitario de uso exclusivo
- COCINA: Hogares con un cuarto exclusivo para cocinar
- DORMI: Número promedio de personas del hogar por dormitorio
- MCOCINA: Hogares que regularmente utilizan leña o carbón para cocinar
- BASURA: Hogares que regularmente eliminan la basura con un servicio municipal o privado

- INDIG: Personas indígenas
- MALFAB: Mujeres que saben leer y escribir
- HALFAB: Hombres que saben leer y escribir
- PRIMARIA: Personas mayores de 7 años que aprobaron la primaria

[INE, 2003] y los valores para cada departamento se muestran en el Cuadro 4.

2. Modelo eblupFH en R. Para la estimación indirecta del indicador de desnutrición crónica, se pueden utilizar las funciones $eblupFH$ o $mseFH$ del paquete Sae de R. La función $eblupFH$ hace estimaciones en dominios basándose en el modelo Fay-Herriot. La función $mseFH$ devuelve los mismos resultados que la función $eblupFH$ más el MSE de las estimaciones. [Molina and Marhuenda, 2015b] Por facilidad, se utilizó la función $mseFH$ mostrada en el código 1.

Código 1 Funciones $mseFH$ y $eblupFH$

```
1 mseFH(formula, vardir, method = "REML", data)
2 eblupFH(formula, vardir, method = "REML", data)
```

El parámetro $formula$ recibe una descripción simbólica del modelo que se quiere ajustar a los datos. Esta consiste en el vector de estimaciones directas (las que se obtuvieron en la sección anterior con la función $svymean$, Cuadro V.3) del lado izquierdo de la fórmula y las variables auxiliares a nivel de área (los indicadores del Censo Nacional mencionados anteriormente) del lado derecho de la fórmula separadas por ”+”; por ejemplo:

$$\textit{EstimaciónDirecta} \sim VariableAuxiliar1 + \cdots + VariableAuxiliarN.$$

El parámetro $vardir$ es un vector que contiene las varianzas de las estimaciones directas para cada dominio (las varianzas del Cuadro V.3) y $method$ se refiere al método utilizado para estimar la varianza, el cual puede ser ML (Maximum Likelihood), $REML$ (Restricted Maximum Likelihood) o FH (Fay-Herriot). En este caso se utilizó el método predeterminado (default), el cual es $REML$. Por último, el parámetro $data$ contiene una tabla de datos con las variables mencionadas en $formula$ y $vardir$. Ambas funciones, $eblupFH$ y $mseFH$, devuelven un vector con las estimaciones de los dominios, los coeficientes del modelo con sus respectivos valores-p para cada variable y un vector que contiene medidas que representan qué tan bien se ajusta el modelo a los datos. [Molina and

Marhuenda, 2015a]

Los valores-p de los coeficientes son utilizados para determinar cuáles variables deben incluirse en el modelo. Estos indican si se cumple la hipótesis nula: el coeficiente es igual a 0. Un valor-p menor que 0.05 indica que se puede rechazar la hipótesis nula, lo cual significa que la variable que acompaña a ese coeficiente juega un papel importante en el modelo. Esto quiere decir que cambios en esa variable están relacionados a cambios en la estimación. Las variables cuyo coeficiente tiene un valor-p mayor a 0.05 no son consideradas significativas, por lo tanto, no se incluyen en el modelo.

También existen criterios como el R cuadrado (utilizado para modelos de regresión), el Criterio de Información de Akaike (Akaike Information Criterion, AIC) y el Criterio de Información Bayesiano (Bayesian Information Criterion, BIC) que se utilizan para determinar el modelo que mejor se ajusta a nuestros datos. Estos criterios asignan puntajes a los modelos y se utilizan para elegir el modelo con el mejor puntaje. Las funciones $eblupFH$ y $mseFH$ devuelven los puntajes del modelo de los siguientes criterios: Criterio de Información de Akaike (AIC), Criterio de Información Bayesiano (BIC) y el Criterio de Información de Kullback (KIC).

Los criterios más utilizados para comparar modelos son el AIC y el BIC. El criterio AIC se considera como el primer criterio que se debe utilizar. Ambos criterios están dados por

$$-2\log L(\hat{\theta}) + kp$$

donde θ es el conjunto de parámetros del modelo, $L(\hat{\theta})$ es la función de verosimilitud evaluada en la estimación de máxima verosimilitud de θ, p es el número de parámetros en el modelo y k es 2 para el AIC y $\log(n)$ para el BIC, donde n es el número de observaciones. [Fabozzi *et al.*, 2014]

Para elegir el mejor modelo entre un grupo de modelos propuestos, se calculan los valores de AIC y BIC para cada modelo y se considera mejor el que menores valores de AIC y BIC posea. El componente, $\log L(\hat{\theta})$, es la probabilidad de obtener los datos dado el modelo propuesto. Como la función de verosimilitud, $\log L(\hat{\theta})$, se multiplica por -2, el modelo con el menor AIC o BIC es para el cual la función de verosimilitud alcanza mayor valor. El componente kp es un factor de ajuste basado en el número de parámetros. Mientras más parámetros tenga el modelo, mayor va a ser el

valor del AIC y BIC. La diferencia entre el AIC y BIC es el factor de ajuste. El BIC impone una penalización mayor para el modelo con mayor cantidad de parámetros. [Fabozzi *et al.*, 2014]

3. Selección de variables. Ya que se conocen las variables auxiliares, la función ($mseFH$) y los criterios (AIC y BIC) que se utilizan para obtener un modelo y calcular las estimaciones indirectas, es necesario seleccionar cuáles variables auxiliares se van a incluir como parámetros del modelo. Los indicadores del Censo Nacional XI de Población y de VI de Habitación 2002 mencionados anteriormente, se utilizaron como estas variables auxiliares y se consideraron ciertos métodos para seleccionar las variables más significativas.

Existen tres métodos de selección de variables: **selección hacia adelante**, **eliminación hacia atrás** y **selección por pasos**. En la **selección hacia adelante** se van agregando variables una por una al modelo. La más significativa de las variables se va agregando al modelo, siempre y cuando su valor-p esté por debajo de cierto nivel. Generalmente este nivel se coloca por arriba de 0.05, como 0.10 o 0.15, debido a la naturaleza exploratoria de este método. La selección termina hasta que las variables restantes no posean valores-p significativos. En la **eliminación hacia atrás** se comienza con el modelo con todas las variables de interés. Luego, se van eliminando una por una las variables que no son significativas (las que posean los valores-p más altos), ajustando el modelo cada vez que se elimina una variable. Este ciclo termina hasta que las variables que hayan quedado en el modelo tengan un valor-p significativo (menor que 0.05). Por último, la **selección por pasos** permite movimientos en cualquiera de las dos direcciones, se pueden eliminar o agregar variables. Usualmente se comienza como en la eliminación hacia atrás y luego se pueden ir agregando variables que más adelante resulten significativas. [Brant, 2004]

Utilizando el método de eliminación hacia atrás, se comenzó incluyendo todas las variables en el modelo como se presenta en el código 2. Los resultados de la función se muestran en la Figura 1. Debido a que la variable EXAGUA fue la que poseía el valor-p más grande, se eliminó del modelo y luego se corrió el nuevo modelo sin ésta variable. se concluyó que las variables INDIG y SANI son las más significativas.

Código 2. Modelo inicial para el método de eliminación hacia atrás

```
1 mseFH(EstDir~TIERRA+EXAGUA+SANI+TSANI+EXSANI+COCINA+DORMI+MCOCINA+BASURA+INDIG+MALFAB+
      HALFAB+PRIMARIA,vardir, data)
```

Figura 1: Resultados de la función $mseFH$ para el modelo inicial del método de eliminación hacia atrás

```
$est$fit$estcoef
                    beta  std.error       tvalue     pvalue
X(Intercept)  -0.24056820  0.7513283 -0.320190534 0.74882390
XTIERRA        1.35317300 16.7748503  0.080666770 0.93570696
XEXAGUA       -0.04714626 14.7881617 -0.003188108 0.99745626
XSANI         -4.32695478 83.9185263 -0.051561377 0.95887820
XTSANI       -31.59451926 22.4237346 -1.408976684 0.15884206
XEXSANI      -15.56692294 84.6101011 -0.183984214 0.85402584
XCOCINA       13.25954640 32.0113183  0.414214319 0.67871716
XDORMI         0.22617207  0.2480942  0.911637875 0.36195938
XMCOCINA     -21.14002032 25.7551204 -0.820808444 0.41175539
XBASURA       14.36149285 31.3428038  0.458207024 0.64680371
XINDIG        11.11429390  5.2575830  2.113954990 0.03451911
XMALFAB      -10.86795692 52.2853094 -0.207858709 0.83533929
XHALFAB        9.30024338 53.5574020  0.173650010 0.86214053
XPRIMARIA     88.12718738 87.6212265  1.005774410 0.31452410
```

Para el método de selección hacia adelante de primero se obtuvieron los valores-p de cada variable auxiliar utilizando la función. Con este método se llegó a que las variables INDIG y DORMI son las más significativas. Por último, con el método de selección por pasos se llegaron a las mismas conclusiones anteriores. Por lo tanto, los modelos que se consideraron fueron los siguientes

$$Modelo \quad A: \quad \textit{Estimación directa} \sim INDIG + DORMI$$

$$Modelo \quad B: \quad \textit{Estimación directa} \sim INDIG + SANI,$$

y sus implementaciones se muestran en los códigos 3 y 4.

Código 3. Implementación del modelo A

```
mseFH(EstDir~INDIG+DORMI, vardir, data)
```

Código 4. Implementación del modelo B

```
mseFH(EstDir~INDIG+SANI, vardir, data)
```

La correlación entre las variables de un modelo también debe tomarse en cuenta. Usualmente conviene utilizar variables que no estén correlacionadas entre sí. El Cuadro 1 presenta el coeficiente de correlación entre cada una de las variables del Censo Nacional. Podemos observar que el coefi-

ciente de correlación entre INDIG y DORMI es mayor que el coeficiente de correlación entre INDIG y SANI, pero ambos son relativamente bajos.

Cuadro 1: Correlación entre las variables auxiliares obtenidas del Censo Nacional XI de Población y de VI de Habitación 2002)

	TIERRA	EXAGUA	SANI	TSANI	EXSANI	COCINA	DORMI	MCOCINA	BASURA	INDIG	MALFAB	HALFAB	PRIMARIA
TIERRA	1.000	0.280	0.304	0.106	0.325	0.353	0.693	0.930	0.094	0.837	0.234	0.303	0.263
EXAGUA	0.280	1.000	0.990	0.977	0.989	0.992	0.372	0.244	0.973	0.302	0.994	0.993	0.992
SANI	0.304	0.990	1.000	0.972	1.000	0.994	0.344	0.277	0.967	0.334	0.994	0.998	0.995
TSANI	0.106	0.977	0.972	1.000	0.968	0.958	0.516	0.064	0.998	0.160	0.986	0.972	0.978
EXSANI	0.325	0.989	1.000	0.968	1.000	0.994	0.324	0.296	0.962	0.350	0.992	0.998	0.994
COCINA	0.353	0.992	0.994	0.958	0.994	1.000	0.299	0.339	0.949	0.362	0.990	0.997	0.994
DORMI	0.693	0.372	0.344	0.516	0.324	0.299	1.000	0.680	0.521	0.508	0.403	0.340	0.380
MCOCINA	0.930	0.244	0.277	0.064	0.296	0.339	0.680	1.000	0.035	0.809	0.211	0.285	0.253
BASURA	0.094	0.973	0.967	0.998	0.962	0.949	0.521	0.035	1.000	0.150	0.981	0.965	0.970
INDIG	0.837	0.302	0.334	0.160	0.350	0.362	0.508	0.809	0.150	1.000	0.251	0.318	0.280
MALFAB	0.234	0.994	0.994	0.986	0.992	0.990	0.403	0.211	0.981	0.251	1.000	0.997	0.998
HALFAB	0.303	0.993	0.998	0.972	0.998	0.997	0.340	0.285	0.965	0.318	0.997	1.000	0.998
PRIMARIA	0.263	0.992	0.995	0.978	0.994	0.994	0.380	0.253	0.970	0.280	0.998	0.998	1.000

Para elegir cuál modelo utilizar entre el *Modelo A* y el *Modelo B*, se deben comparar los valores de AIC y BIC de cada modelo y se escoge el que tenga los menores valores (tomando en cuenta el signo). Estos valores se pueden observar en el Cuadro 2. Ambos modelos tienen valores muy parecidos por lo que se compararon los resultados de ambos modelos, los cuales se muestran en el Cuadro 3. Debido a que ambos modelos producen resultados muy similares, se podría elegir cualquiera de los dos y en este caso se eligió el modelo A.

Cuadro 2: Valores de AIC y BIC para los modelos A y B

Modelo	AIC	BIC
A	-23.98747	-19.62330
B	-23.95668	-19.59251

Cuadro 3: Indicadores obtenidos utilizando los modelos A y B

Departamento	Valor estimado Modelo A (V)	Valor estimado Modelo B (V)	Error estándar Modelo A (EE)	Error estándar Modelo B (EE)	Ancho de intervalo Modelo A	Ancho de intervalo Modelo B
Guatemala	0.367	0.359	0.043	0.044	0.173	0.178
El Progreso	0.328	0.343	0.065	0.065	0.262	0.261
Sacatepequez	0.451	0.485	0.077	0.076	0.308	0.302
Chimaltenango	0.606	0.618	0.055	0.055	0.219	0.219
Escuintla	0.281	0.279	0.050	0.050	0.200	0.200
Santa Rosa	0.396	0.397	0.061	0.062	0.246	0.246
Solola	0.665	0.671	0.045	0.046	0.182	0.182
Totonicapan	0.793	0.795	0.039	0.039	0.156	0.157
Quetzaltenango	0.473	0.480	0.040	0.040	0.161	0.160
Suchitepequez	0.467	0.464	0.042	0.042	0.166	0.166
Retalhuleu	0.449	0.446	0.044	0.043	0.174	0.174
San Marcos	0.618	0.600	0.050	0.049	0.200	0.197
Huehuetenango	0.691	0.690	0.033	0.033	0.131	0.131
Quiche	0.678	0.680	0.044	0.044	0.178	0.178
Baja Verapaz	0.460	0.459	0.070	0.070	0.278	0.279
Alta Verapaz	0.654	0.654	0.035	0.035	0.142	0.142
Petén	0.471	0.460	0.049	0.049	0.197	0.195
Izabal	0.424	0.425	0.076	0.076	0.303	0.304
Zacapa	0.112	0.115	0.038	0.038	0.154	0.154
Chiquimula	0.496	0.495	0.050	0.050	0.198	0.199
Jalapa	0.528	0.528	0.058	0.058	0.233	0.233
Jutiapa	0.450	0.449	0.042	0.042	0.168	0.168

4. Errores estándar e intervalos de confianza. Luego, se calcularon las estimaciones indirectas del indicador de desnutrición crónica por departamento junto con sus errores estándar, intervalos de confianza y varianza. Implementando las funciones mostradas en los códigos 3 y 4 se obtuvieron las estimaciones indirectas por departamento junto con sus MSE.

Recordando lo que se mencionó en la sección 1 Sesgo y Varianza y suponiendo que el sesgo es 0, tenemos

$$MSE[\hat{t}] = Var[\hat{t}] + (Sesgo[\hat{t}])^2 = Var[\hat{t}].$$

Entonces, podemos decir que los valores de MSE que se obtuvieron con la función $mseFH$ son las varianzas de las estimaciónes. A partir de las varianzas podemos calcular los errores estándar, ya que

$$EE = \sqrt{Var},$$

y a partir de los errores estándar podemos calcular los intervalos de confianza. El cógido 7 mostrado en el Apéndice E es el que se utilizó para calcular las estimaciones indirectas por departamento junto con los errores estándar e intervalos de confianza.

D. Cuadros de indicadores

Cuadro 4: Indicadores obtenidos del informe del Censo Nacional XI de Población y de VI de Habitación 2002)

Departamento	TIERRA	EXAGUA	SANI	TSANI	EXSANI	COCINA	DORMI	MCOCINA	BASURA	INDIG	MALFAB	HALFAB	PRIMARIA
Guatemala	0.0177	0.1936	0.2467	0.1817	0.2262	0.194	2.16	0.0277	0.1928	0.0305	0.2065	0.2111	0.0637
El Progreso	0.0026	0.0109	0.0119	0.0044	0.0117	0.0116	2.67	0.0075	0.0023	0.0001	0.0095	0.01	0.0037
Sacatepequez	0.0018	0.0167	0.022	0.0145	0.0192	0.0198	2.5	0.0096	0.0108	0.0093	0.0173	0.0199	0.0068
Chimaltenango	0.0094	0.0254	0.0354	0.0155	0.0337	0.034	3.04	0.0278	0.009	0.0314	0.0266	0.0312	0.0102
Escuintla	0.0066	0.0253	0.0448	0.0237	0.0407	0.0394	2.93	0.0251	0.014	0.0036	0.0342	0.0396	0.0137
Santa Rosa	0.0076	0.0181	0.023	0.007	0.022	0.024	2.95	0.0199	0.0042	0.0007	0.0189	0.0208	0.0076
Solola	0.0084	0.0214	0.0218	0.005	0.0206	0.0216	3.24	0.0217	0.0054	0.0264	0.0139	0.0184	0.006
Totonicapan	0.0133	0.0197	0.0227	0.004	0.0223	0.0255	3.41	0.0234	0.0017	0.0297	0.0169	0.0206	0.0067
Quetzaltenango	0.0102	0.0371	0.0506	0.0208	0.0463	0.0486	2.85	0.0335	0.0157	0.0301	0.0401	0.0446	0.0142
Suchitepequez	0.0096	0.0203	0.0268	0.0151	0.0257	0.0293	3.5	0.0255	0.0081	0.0185	0.0219	0.0268	0.0082
Retalhuleu	0.006	0.0101	0.0192	0.0063	0.0183	0.0179	3.29	0.0153	0.004	0.0049	0.0146	0.0171	0.0053
San Marcos	0.0256	0.0417	0.0582	0.0113	0.0565	0.0567	3.92	0.0533	0.0058	0.0221	0.0431	0.0534	0.0153
Huehuetenango	0.032	0.0452	0.0508	0.0143	0.0494	0.0537	3.78	0.0567	0.0062	0.0491	0.0378	0.0485	0.0148
Quiche	0.0314	0.034	0.0367	0.0076	0.036	0.0429	3.78	0.0464	0.0032	0.0518	0.0237	0.032	0.0088
Baja Verapaz	0.0088	0.0138	0.0154	0.0032	0.0152	0.0161	3.27	0.0156	0.0018	0.0113	0.0104	0.0126	0.0041
Alta Verapaz	0.0356	0.0255	0.0541	0.0083	0.053	0.0418	3.94	0.0527	0.0055	0.0641	0.0276	0.0407	0.0109
Petén	0.0147	0.0154	0.0219	0.0037	0.0215	0.0228	3.64	0.0247	0.0024	0.0101	0.019	0.023	0.0068
Izabal	0.0075	0.0185	0.0237	0.0091	0.0229	0.0203	3.06	0.0162	0.0062	0.0065	0.0181	0.0202	0.0068
Zacapa	0.0038	0.015	0.0162	0.0077	0.0154	0.0159	2.79	0.0109	0.0044	0.0001	0.0126	0.0132	0.0045
Chiquimula	0.011	0.0173	0.0173	0.0086	0.0165	0.0212	3.13	0.0193	0.0053	0.0045	0.0161	0.0164	0.0059
Jalapa	0.0104	0.0147	0.0148	0.0052	0.0141	0.018	3.11	0.0162	0.0027	0.0042	0.013	0.0145	0.0046
Jutiapa	0.0117	0.0229	0.0208	0.0099	0.0202	0.031	3.04	0.0259	0.0047	0.0012	0.024	0.0258	0.0095

Cuadro 5: Indicadores de desnutrición para niños y niñas entre 3 y 59 meses de edad. **Nota.** Recuperado de la ENSMI 2002.

Cuadro 9.7 Indicadores de desnutrición para niño/as de 3 a 59 meses de edad

Porcentaje (*) de niño/as de 3 a 59 meses de edad, clasificados como desnutridos según tres indicadores antropométricos: talla para la edad, peso para la talla y peso para la edad, según características seleccionadas. ENSMI-2002

Característica	Porcentaje con desnutrición crónica (Talla para la edad)		Porcentaje con desnutrición aguda (Peso para la talla)		Porcentaje con desnutrición global (Peso para la edad)		No. de casos no ponderados	No. de casos ponderados
	Severa(1)	Total(2)	Severa(1)	Total(2)	Severa(1)	Total(2)		
Área								
Urbana	14.6	36.5	0.4	1.2	1.5	16.2	(1,542)	1,500
Rural	24.4	55.5	0.3	1.8	4.7	25.9	(4,766)	3,093
Región								
Metropolitana	14.3	36.1	0.6	1.1	1.1	15.1	(586)	1,078
Norte	24.7	61.0	.	1.2	4.2	23.7	(912)	457
Nor-Oriente	13.1	39.7	0.9	3.6	5.3	17.7	(503)	465
Sur-Oriente	16.6	46.6	.	1.3	2.5	26.0	(446)	446
Central	17.3	42.1	0.7	1.8	3.0	21.7	(702)	492
Sur-Occidente	28.0	58.5	0.2	1.7	5.1	28.5	(1,551)	942
Nor-Occidente	37.4	68.3	0.1	1.3	6.6	31.5	(1,064)	525
Petén	14.0	46.1	.	2.1	2.4	18.0	(544)	189
Grupo étnico								
Indígena	35.5	69.5	0.2	1.7	5.6	30.4	(3,055)	1,844
Ladino	11.6	35.7	0.4	1.6	2.4	17.5	(3,253)	2,750
Nivel de educación								
Sin educación	31.6	65.6	0.3	2.2	5.7	29.9	(2,482)	1,660
Primaria	18.3	46.4	0.3	1.4	3.1	21.6	(3,171)	2,267
Secundaria o más	5.3	18.6	0.4	1.1	0.6	8.5	(655)	667
Total	**21.2**	**49.3**	**0.3**	**1.6**	**3.7**	**22.7**	**(6,308)**	**4,594**

Nota: Las estimaciones se refieren a los niño/as de 3 a 59 meses de edad (se excluyen lo/as menores de 3 meses). Cada índice se expresa en términos del número de desviaciones estándar (DE) de la media del patrón de referencia internacional utilizado por NCHS/CDC/WHO. Lo/as niño/as se clasifican como desnutrido/as si están 2 o más desviaciones estándar (DE) por debajo de la población de referencia.
(1) Niño/as que están 3 (DE) o más por debajo de la media.
(2) Niño/as que están 2 (DE o más por debajo de la media. Incluye a los niño/as que están 3 (DE) o más por debajo de la media.
* Incluye sólo la respuesta afirmativa para cada columna

Cuadro 6: Indicadores regionales de desnutrición crónica para niños y niñas entre 3 y 59 meses de edad **Nota.** Recuperado de la ENSMI 2002.

Región	Valor estimado (V)	Error estándar (EE)	No. de casos		Efecto del diseño (EDIS)	Intervalo de confianza	
			Sin ponderar (SP)	Ponderados (P)		V-2EE	V+2EE
Metropolitana	0.361	0.044	586	1078.0	2.224	0.273	0.449
Norte	0.610	0.034	912	456.7	2.094	0.542	0.678
Nor-Oriente	0.397	0.048	503	465.0	2.197	0.301	0.493
Sur-Oriente	0.466	0.036	446	446.0	1.514	0.394	0.538
Central	0.421	0.045	702	491.9	2.437	0.331	0.511
Sur-Occidente	0.585	0.022	1551	942.5	1.732	0.541	0.629
Nor-Occidente	0.683	0.027	1064	524.8	1.897	0.629	0.737
Petén	0.461	0.052	544	188.7	2.438	0.357	0.565

Cuadro 7: Indicadores de desnutrición para niños y niñas entre 3 y 59 meses de edad. **Nota.** Recuperado de la ENSMI 2014.

Cuadro 11.1b Estado nutricional de niñas y niños menores de 5 años según lugar de residencia

Porcentaje de niñas y niños de facto menores de 5 años clasificados como malnutridos (desnutridos o con sobrepeso) según los 3 índices del estado nutricional: talla-para-edad, peso- para-talla, y peso-para-edad, según lugar de residencia, Guatemala 2014-2015

	Porcentaje con desnutrición crónica (Talla para la edad) [1]			Porcentaje con desnutrición aguda (Peso para la talla)				Porcentaje con desnutrición global (Peso para la edad)				
Lugar de residencia	Porcentaje por debajo de -3 DE	Porcentaje por debajo de -2 DE [2]	Promedio valor Z (DE)	Porcentaje por debajo de -3 DE	Porcentaje por debajo de -2 DE [2]	Porcentaje por arriba de +2 DE	Promedio valor Z (DE)	Porcentaje por debajo de -3 DE	Porcentaje por debajo de -2 DE [2]	Porcentaje por arriba de +2 DE	Promedio valor Z (DE)	Número de niñas y niños
Área de residencia												
Urbana	9.7	34.6	-1.6	0.1	0.8	5.2	0.3	1.5	9.5	1.2	-0.7	4,431
Rural	20.4	53.0	-2.1	0.1	0.7	4.4	0.3	2.4	14.3	0.4	-1.0	8,135
Región												
Metropolitana	4.8	25.3	-1.3	0.2	1.0	5.3	0.3	0.9	7.7	1.5	-0.5	1,881
Norte	17.4	50.0	-2.0	0.2	0.7	3.6	0.5	2.0	10.7	0.4	-0.9	1,436
Suroriente	14.1	40.3	-1.8	0.4	0.8	5.3	0.4	1.8	11.0	0.8	-0.8	1,143
Nororiente	14.4	39.9	-1.7	0.0	0.8	5.8	0.3	2.5	12.6	1.7	-0.8	1,145
Central	13.6	41.1	-1.7	0.1	0.8	5.0	0.3	1.2	10.1	0.7	-0.8	1,340
Suroccidente	18.2	51.9	2.0	0.1	0.8	4.3	0.3	2.3	14.2	0.4	-1.0	2,988
Noroccidente	30.8	68.2	-2.5	0.0	0.3	4.8	0.4	3.6	19.8	0.0	-1.2	2,109
Petén	8.6	36.1	-1.5	0.0	0.7	3.5	0.3	0.7	7.6	0.7	-0.6	523
Departamento												
Guatemala	4.8	25.3	-1.3	0.2	1.0	5.3	0.3	0.9	7.7	1.5	-0.5	1,881
...Guatemala municipio	*2.5*	*18.7*	*-1.1*	*0.0*	*0.9*	*3.5*	*0.3*	*0.7*	*4.1*	*1.1*	*-0.4*	*474*
...Guatemala resto	*5.6*	*27.6*	*-1.4*	*0.3*	*1.1*	*5.9*	*0.3*	*1.0*	*8.9*	*1.6*	*-0.6*	*1,407*
El Progreso	5.2	29.1	-1.3	0.0	1.6	4.5	0.2	2.3	8.9	2.2	-0.6	139
Sacatepéquez	11.8	42.4	-1.7	0.5	0.9	8.5	0.6	1.3	7.3	0.5	-0.6	222
Chimaltenango	21.6	56.5	-2.2	0.0	0.4	5.9	0.5	1.1	12.6	0.4	-0.9	527
Escuintla	7.1	26.9	-1.4	0.0	1.1	2.9	0.1	1.1	8.9	1.0	-0.7	591
Santa Rosa	9.7	33.6	-1.6	0.2	0.6	4.8	0.3	0.9	9.4	0.9	-0.7	357
Sololá	24.7	65.6	-2.3	0.0	0.0	4.6	0.4	1.4	15.5	0.4	-1.0	271
Totonicapán	30.8	70.0	-2.5	0.0	0.5	4.9	0.4	2.6	18.5	0.4	-1.2	347
Quetzaltenango	13.8	48.8	-1.9	0.3	1.0	4.8	0.4	1.2	12.1	0.4	-0.9	712
Suchitepéquez	10.4	39.6	-1.7	0.0	1.1	3.9	0.2	1.9	12.2	0.1	-0.8	420
Retalhuleu	9.8	34.2	-1.6	0.0	1.1	3.9	0.1	1.9	12.3	0.7	-0.8	302
San Marcos	21.3	54.8	-2.1	0.0	0.7	3.9	0.3	3.5	15.2	0.3	-1.0	935
Huehuetenango	34.0	67.7	-2.5	0.0	0.4	4.9	0.4	4.8	21.4	0.0	-1.2	1,100
Quiché	27.3	68.7	-2.4	0.0	0.2	4.6	0.4	2.3	18.0	0.0	-1.2	1,010
Baja Verapaz	17.2	50.2	-2.0	0.0	0.6	5.8	0.4	2.3	13.2	0.4	-0.9	257
Alta Verapaz	17.5	50.0	-2.0	0.3	0.7	3.1	0.5	1.9	10.2	0.3	-0.8	1,179
Petén	8.6	36.1	-1.5	0.0	0.7	3.5	0.3	0.7	7.6	0.7	-0.6	523
Izabal	5.3	26.4	-1.3	0.0	1.2	6.2	0.3	1.4	6.2	2.1	-0.5	362
Zacapa	15.6	40.0	-1.8	0.0	0.5	8.2	0.4	0.9	13.1	2.4	-0.7	235
Chiquimula	24.9	55.6	-2.2	0.0	0.4	4.6	0.3	4.3	19.2	0.6	-1.1	409
Jalapa	22.0	53.8	-2.1	0.5	0.6	5.9	0.5	2.9	16.0	0.9	-0.9	332
Jutiapa	11.9	35.7	-1.6	0.6	0.9	5.3	0.4	1.7	8.5	0.7	-0.7	454
Total	**16.6**	**46.5**	**-1.9**	**0.1**	**0.7**	**4.7**	**0.3**	**2.1**	**12.6**	**0.7**	**-0.9**	**12,567**

Nota: El cuadro está basado en niñas y niños que durmieron en el hogar la noche anterior a la entrevista. Cada índice se expresa en desviaciones estándar (DE) de la mediana de los Estándares de Crecimiento de los Niños de la Organización Mundial de la Salud (OMS) adoptados en 2006. El cuadro está basado en niñas y niños con fechas válidas de nacimiento (mes y año) y mediciones válidas tanto de talla como de peso.

[1] Para los menores de 2 años la talla se mide acostados, y también en los pocos casos en los que la edad de niñas y niños no se conoce o mide menos de 85 cm. Para todos las demás niñas y niños la talla se mide de pie.

[2] Incluye niñas y niños que están por debajo de -3 desviaciones estándar (DE) de la mediana de la población para los Estándares de Crecimiento de los Niños de la OMS.

E. Códigos

Código 5. Implementación de las funciones *svyby* y *svymean* para obtener las estimaciones directas del porcentaje de niños entre 3 y 59 meses de edad con desnutrición crónica

```
# ESTIMACIÓN DIRECTA DEL INDICADOR DE DESNUTRICIÓN CRÓNICA

library(Hmisc)
library(survey)
library(utils)

# ENSMI 2002 BASES DE DATOS
# Contiene los datos de los hijos nacidos vivos de las mujeres en edad reproductiva entrevistadas, 28731 nacimientos
ensmi_h = spss.get("gt02f_hijos.sav", use.value.labels=TRUE)

# Variables que se utilizan:

# vivo = ¿está vivo?
# mnac = mes de nacimiento
# anac = año de nacimiento
# hedad = meses cumplidos desde nacidos (todos los niños, vivos y fallecidos)
# HAZ = valor z de talla para edad del niño
# WAZ = valor z de peso para edad del niño
# WHZ = valor z de peso para talla del niño
# mpaquete = número de paquete (representa el sector censal, unidades primarias de muestreo)
# MHOGAR = número de hogar
# PesoMEF = la variable de peso que se debe utilizar en todos los cálculos para mujeres con entrevista completa. La suma de todos estos pesos debe ser 9155 (la probabilidad de selección de un niño es la misma que la probabilidad de selección de su madre)

# Se identifican los registros de niños vivos con variables mnac y anac válidas
ensmi_h2 = subset(ensmi_h, ensmi_h$vivo=="Si" & ensmi_h$mnac != 98 & ensmi_h$anac != 9998)

# Se reduce la base ensmi_h2 a sólo los niños entre 3 y 59 meses de edad
ensmi_h2.1 = subset(ensmi_h2, ensmi_h2$hedad >= 3 & ensmi_h2$hedad <= 59)

# Se identifican los registros de niños con variables HAZ, WAZ, y WHZ válidas
# Niños con HAZ menor a -6 DE o mayor a 6 DE tienen datos inválidos, 0=inválido, 1=válido
ensmi_h2.1$validHAZ = ifelse(ensmi_h2.1$HAZ< -6 | ensmi_h2.1$HAZ>6,0,1)
# Niños con WAZ menor a -6 DE o mayor a 6 DE tienen datos inválidos, 0=inválido, 1=válido
ensmi_h2.1$validWAZ = ifelse(ensmi_h2.1$WAZ< -6 | ensmi_h2.1$WAZ>6,0,1)
# Niños con WHZ menor a -6 DE o mayor a 6 DE tienen datos inválidos, 0=inválido, 1=válido
```

```
ensmi_h2.1$validWHZ = ifelse(ensmi_h2.1$WHZ< -4 | ensmi_h2.1$WHZ>6,0,1)

# Se hace una tabla de niños con registros de HAZ, WAZ y WHZ válidos
ensmi_h2.2 = subset(ensmi_h2.1,ensmi_h2.1$validHAZ==1 & ensmi_h2.1$validWAZ==1 & ensmi_h2.1$validWHZ==1)

# Se revisa si hay datos con las siguientes combinaciones inválidas:
# HAZ menor a -3.09 DE y WAZ mayor a 3.09 DE
nrow(subset(ensmi_h2.2, ensmi_h2.2$HAZ< -3.09 & ensmi_h2.2$WAZ>3.09))
# HAZ mayor a 3.09 DE y WA menor a -3.09 DE
nrow(subset(ensmi_h2.2, ensmi_h2.2$HAZ > 3.09 & ensmi_h2.2$WAZ< -3.09))
# (No hay combinaciones inválidas)

# BASE FINAL: ensmi_h2.2, tabla con todos los registros de niños vivos entre 3 y 59 meses de edad con valores HAZ, WAZ y WHZ válidos

# Se considera que un niño tiene desnutrición total si HAZ <= -2
# Se crea una nueva columna "desnu", donde se registra si el niño está desnutrido o no, 1 = desnutrido
ensmi_h2.2$desnu = ifelse(ensmi_h2.2$HAZ <= -2, 1,0)

# Para que R no considere que los pesos están estandarizados, creamos una nueva variable de peso
ensmi_h2.2$PesoMEF2 = ensmi_h2.2$PesoMEF*100

# Se determina el diseño de la encuesta
# id (unidades primarias de muestreo) = la variable mpaquete representa los segmentos censales
# strata (estratos) = los departamentos
diseño1 = svydesign(id=~mpaquete,strata=~MHDEPTO,weights=~PesoMEF2, data=ensmi_h2.2)
# Resumen del diseño
summary(diseño1)

# Estimaciones por región
# Devuelve la estimación, error estándar (se), intervalo de confianza (ci), varianza (var) y efecto del diseño (deff)
desnu_svy_reg = svyby(~desnu,~REGION,diseño1,svymean,deff=TRUE,vartype = c("se","ci","var"))

# Se elimina la columna repetida de región
desnu_svy_reg$REGION = NULL

# El EDIS reportado en el informe de la ENSMI 2002 es la razón entre el error estándar correspondiente al diseño empleado y el error estándar que se obtiene tratando la muestra como si hubiera sido aleatoria simple
```

```
# El DEFF que reporta R es la razón entre la varianza correspondiente al diseño empleado y la varianza que se obtiene tratando la muestra como si hubiera sido aleatoria simple sin reemplazo.
# Entonces, EDIS = sqrt(DEFF)
desnu_svy_reg$DEff.desnu = sqrt(desnu_svy_reg$DEff.desnu)

# Se cambian los nombres de las columnas
colnames(desnu_svy_reg) = c("Valor estimado (V)", "Error estándar (EE)","V-2EE", "V+2EE","Varianza (Var)","Efecto del diseño (EDIS)")

# La tabla desnu_svy_reg se pasa a un archivo csv
write.csv(desnu_svy_reg, "desnu_svy_reg.csv")

# Estimaciones por departamento
# Devuelve la estimación, error estándar (se), intervalo de confianza (ci), varianza (var) y efecto del diseño (deff)
desnu_svy_depto = svyby(~desnu,~MHDEPTO,diseño1,svymean,deff=TRUE,vartype = c("se","ci","var"))

# Se elimina la columna repetida de departamento
desnu_svy_depto$MHDEPTO = NULL

# El EDIS reportado en el informe de la ENSMI 2002 es la razón entre el error estándar correspondiente al diseño empleado y el error estándar que se obtiene tratando la muestra como si hubiera sido aleatoria simple
# El DEFF que reporta R es la razón entre la varianza correspondiente al diseño empleado y la varianza que se obtiene tratando la muestra como si hubiera sido aleatoria simple sin reemplazo.
# Entonces, EDIS = sqrt(DEFF)
desnu_svy_depto$DEff.desnu = sqrt(desnu_svy_depto$DEff.desnu)

# Se cambian los nombres de las columnas
colnames(desnu_svy_depto) = c("Valor estimado (V)", "Error estándar (EE)","V-2EE", "V+2EE","Varianza (Var)","Efecto del diseño (EDIS)")

# La tabla desnu_svy_depto se pasa a un archivo csv
write.csv(desnu_svy_depto, "desnu_svy_depto.csv")
```

Código 6. Implementación de la ecuación C de Estimación en Dominios para obtener las estimaciones directas del porcentaje de niños entre 3 y 59 meses de edad con desnutrición crónica

```
# ESTIMACIÓN DIRECTA CON ECUACIONES DE ESTIMACIÓN EN DOMINIOS

# Lista con el nombre de las regiones
regiones = c(levels(ensmi_h2.2$MHREG))
```

```
# Tabla vacía para luego ingresar las estimaciones
desnu_reg = matrix(nrow=length(regiones),ncol=1,dimnames =list(regiones,c("Valor estimado (V)")))
# Utilizando las ecuaciones de estimación en dominios
for (n in 1:length(regiones)) {
  desnu_reg[n,1] = sum(subset(ensmi_h2.2, ensmi_h2.2$MHREG==regiones[n])$PesoMEF*subset(ensmi_h2.2, ensmi_h2.2$MHREG==regiones[n])$desnu,na.rm = TRUE)/sum(subset(ensmi_h2.2, ensmi_h2.2$MHREG==regiones[n])$PesoMEF,na.rm = TRUE)
}
# La tabla se pasa a un archivo .csv
write.csv(desnu_reg,"desnu_reg.csv")

# Lista con el nombre de los departamentos
deptos = c(levels(ensmi_h2.2$MHDEPTO))
# Tabla vacía para luego ingresar las estimaciones
desnu_deptos = matrix(nrow=length(deptos),ncol=1,dimnames =list(deptos,c("Valor estimado (V)")))
# Utilizando las ecuaciones de estimación en dominios
for (n in 1:length(deptos)) {
  desnu_deptos[n,1] = sum(subset(ensmi_h2.2, ensmi_h2.2$MHDEPTO==deptos[n])$PesoMEF*subset(ensmi_h2.2, ensmi_h2.2$MHDEPTO==deptos[n])$desnu,na.rm = TRUE)/sum(subset(ensmi_h2.2, ensmi_h2.2$MHDEPTO==deptos[n])$PesoMEF,na.rm = TRUE)
}
# La tabla se pasa a un archivo .csv
write.csv(desnu_deptos,"desnu_deptos.csv")
```

Código 7. Implementación de la función $mseFH$ para obtener las estimaciones indirectas del porcentaje de niños entre 3 y 59 meses de edad con desnutrición crónica

```
# ESTIMACIÓN INDIRECTA DEL INDICADOR DE DESNUTRICIÓN CRÓNICA

# Lista con el nombre de los departamentos
deptos = c(levels(ensmi_h2.2$MHDEPTO))
# Tabla con departamentos, estimaciones directas, varianzas de las estimaciones directas e indicadores del censo
tabla1 = data.frame(deptos,desnu_svy_depto$`Valor estimado (V)`,desnu_svy_depto$`Varianza (Var)`,censo[,2:ncol(censo)])
# Se cambian los nombres de las columnas de la tabla
colnames(tabla1)[1:3] = c("Depto","V","Var")

# Se obtienen las estimaciones y sus MSE con la función mseFH
# Modelo A
desnu_eblup_deptoA = mseFH(V~INDIG+DORMI,Var, data = tabla1)
# Modelo B
desnu_eblup_deptoB = mseFH(V~INDIG+SANI,Var, data = tabla1)
```

```
15
16 # Como MSE = Varianza + Sesgo, suponemos que Sesgo = 0 y calculamos el EE a partir de la
       varianza, ya que EE = sqrt(Var)
17 desnu_eblup_deptoA$est$EE = sqrt(desnu_eblup_deptoA$mse)
18 desnu_eblup_deptoB$est$EE = sqrt(desnu_eblup_deptoB$mse)
19
20
21 # Se agregan columnas con los límites de los intervalos de confianza
22 # Límite inferior
23 desnu_eblup_deptoA$est$ci1 = desnu_eblup_deptoA$est$eblup-(2*desnu_eblup_deptoA$est$EE)
24 desnu_eblup_deptoB$est$ci1 = desnu_eblup_deptoB$est$eblup-(2*desnu_eblup_deptoB$est$EE)
25 # Límite superior
26 desnu_eblup_deptoA$est$ci2 = desnu_eblup_deptoA$est$eblup+(2*desnu_eblup_deptoA$est$EE)
27 desnu_eblup_deptoB$est$ci2 = desnu_eblup_deptoB$est$eblup+(2*desnu_eblup_deptoB$est$EE)
28
29 # Tabla con la estimación indirecta, error estándar, intervalo de confianza y varianza (
       MSE)
30 desnu_eblupA = data.frame(cbind(deptos, desnu_eblup_deptoA$est$eblup, desnu_eblup_deptoA$
       est$EE, desnu_eblup_deptoA$est$ci1, desnu_eblup_deptoA$est$ci2, desnu_eblup_deptoA$mse
       ))
31 desnu_eblupB = data.frame(cbind(deptos, desnu_eblup_deptoB$est$eblup, desnu_eblup_deptoB$
       est$EE, desnu_eblup_deptoB$est$ci1, desnu_eblup_deptoB$est$ci2, desnu_eblup_deptoB$mse
       ))
32
33 # Se cambian los nombres de las columnas de la tabla
34 colnames(desnu_eblupA) = c("Departamento", "Valor estimado (V)", "Error estándar (EE)","V
       -2EE", "V+2EE","Varianza (Var)")
35 colnames(desnu_eblupB) = c("Departamento", "Valor estimado (V)", "Error estándar (EE)","V
       -2EE", "V+2EE","Varianza (Var)")
36
37 # Tabla de comparación entre estimación directa e indirecta
38 comparacionA = data.frame(deptos,desnu_svy_depto$`Valor estimado (V)`,desnu_eblupA$`Valor
       estimado (V)`,
39                         desnu_svy_depto$`Error estándar (EE)`, desnu_eblupA$`Error
                               estándar (EE)`,
40                         desnu_svy_depto$`V-2EE`, desnu_svy_depto$`V+2EE`,
41                         desnu_eblupA$`V-2EE`, desnu_eblupA$`V+2EE`,
42                         desnu_svy_depto$`Varianza (Var)`, desnu_eblupA$`Varianza (Var)`)
43 comparacionB = data.frame(deptos,desnu_svy_depto$`Valor estimado (V)`,desnu_eblupB$`Valor
       estimado (V)`,
44                          desnu_svy_depto$`Error estándar (EE)`, desnu_eblupB$`Error
                                estándar (EE)`,
45                          desnu_svy_depto$`V-2EE`, desnu_svy_depto$`V+2EE`,
46                          desnu_eblupB$`V-2EE`, desnu_eblupB$`V+2EE`,
47                          desnu_svy_depto$`Varianza (Var)`, desnu_eblupB$`Varianza (Var)`)
```

```
48 # Se le cambian los nombres de las columnas de la tabla
49 colnames(comparacionA) = c("Departamentos","V (directa)","V (indirecta)","EE (directa)","
       EE (indirecta)",
50                          "V-2EE (directa)","V+2EE (directa)","V-2EE (indirecta)","V+2EE (
                                indirecta)",
51                          "Var (directa)","Var (indirecta)")
52 colnames(comparacionB) = c("Departamentos","V (directa)","V (indirecta)","EE (directa)","
       EE (indirecta)",
53                           "V-2EE (directa)","V+2EE (directa)","V-2EE (indirecta)","V+2EE
                                 (indirecta)",
54                           "Var (directa)","Var (indirecta)")
55 # La tabla se pasa a un archivo .csv
56 write.csv(comparacionA, "comparacionA.csv")
57 write.csv(comparacionB, "comparacionB.csv")
```