

# Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares

Una aplicación a la  
migración internacional

Andrés Gutiérrez  
Alvaro Fuentes  
Xavier Mancero  
Felipe López  
Felipe Molina



NACIONES UNIDAS

CEPAL

SERIE

**ESTUDIOS ESTADÍSTICOS**

**101**

# **Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares**

Una aplicación a la migración internacional

Andrés Gutiérrez

Álvaro Fuentes

Xavier Mancero

Felipe López

Felipe Molina



NACIONES UNIDAS

**CEPAL**

Este documento fue preparado por Andrés Gutiérrez, Experto Regional en Estadísticas Sociales; Álvaro Fuentes, Estadístico, y Xavier Mancero, Jefe de la Unidad de Estadísticas Sociales, de la División de Estadísticas de la Comisión Económica para América Latina y el Caribe (CEPAL), con el apoyo de Felipe López y Felipe Molina, Consultores de la División.

Las opiniones expresadas en este documento, que no ha sido sometido a revisión editorial, son de exclusiva responsabilidad de los autores y pueden no coincidir con las de la Organización.

Publicación de las Naciones Unidas  
ISSN: 1680-8789 (versión electrónica)  
ISSN: 1994-7364 (versión impresa)  
LC/TS.2020/52  
Distribución: L  
Copyright © Naciones Unidas, 2020  
Todos los derechos reservados  
Impreso en Naciones Unidas, Santiago  
S.20-00293

Esta publicación debe citarse como: A. Gutiérrez y otros, "Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares: una aplicación a la migración internacional" *serie Estudios Estadísticos*, N° 101 (LC/TS.2020/52), Santiago, Comisión Económica para América Latina y el Caribe (CEPAL), 2020.

La autorización para reproducir total o parcialmente esta obra debe solicitarse a la Comisión Económica para América Latina y el Caribe (CEPAL), División de Publicaciones y Servicios Web, publicaciones.cepal@un.org. Los Estados Miembros de las Naciones Unidas y sus instituciones gubernamentales pueden reproducir esta obra sin autorización previa. Solo se les solicita que mencionen la fuente e informen a la CEPAL de tal reproducción.

# Índice

Resumen .....	5
Introducción .....	7
<b>I. Principios básicos de estimación .....</b>	<b>9</b>
A. Análisis en subpoblaciones .....	10
B. Estimación puntual de promedios y proporciones .....	11
C. Aproximación de la varianza.....	11
<b>II. Criterios de calidad .....</b>	<b>15</b>
A. Intervalos de confianza.....	15
B. Coeficientes de variación.....	16
C. Coeficiente de variación logarítmico .....	17
D. El efecto de diseño DEFF .....	19
E. Tamaño de muestra .....	21
F. Tamaño de muestra efectivo.....	21
G. Grados de libertad.....	22
H. Conteo de casos no ponderado .....	23
<b>III. Aplicación de los criterios de calidad.....</b>	<b>25</b>
A. Definición de variables y subpoblaciones.....	25
B. Análisis de bases de datos sin UPM ni estratos .....	27
C. Secuencia lógica para crear reglas de supresión .....	28
<b>IV. Aplicación a la estimación de la población inmigrante .....</b>	<b>33</b>
A. Definición estadística de la migración internacional .....	33
B. La forma de medir la migración en las encuestas de hogares.....	34
C. Precisión estadística de la información sobre inmigración en las encuestas de hogares.....	37
<b>V. Conclusiones.....</b>	<b>43</b>

<b>Bibliografía</b> .....	45	
<b>Anexos</b> .....	47	
Anexo 1 .....	48	
Anexo 2 .....	50	
Serie Estudios Estadísticos: números publicados.....	54	
<b>Cuadros</b>		
Cuadro 1	Estimación de la proporción de personas pobres desagregada por estado migratorio, junto con algunas medidas de precisión, para un país latinoamericano.....	8
Cuadro 2	Estructura de resultados del procesamiento de una proporción en el repositorio BADEHOG.....	30
Cuadro 3	América Latina (18 países): disponibilidad de información para identificar inmigrantes y estimar precisión de los estimadores, alrededor de 2017.....	35
Cuadro 4	América Latina (16 países): variables disponibles para identificar inmigrantes, encuesta alrededor de 2017.....	36
Cuadro 5	América Latina (16 países): porcentaje de personas y límites inferior y superior del intervalo de confianza para migrantes y migrantes recientes, alrededor de 2017.....	38
Cuadro 6	América Latina (16 países): estimadores que no alcanzan los requerimientos mínimos de precisión exigidos por país, según indicador para la subpoblación migrante, alrededor de 2017.....	39
Cuadro 7	América Latina (15 países): estimadores que no alcanzan los requerimientos mínimos de precisión exigidos por país, según indicador para la subpoblación migrante reciente, alrededor de 2017.....	40
Cuadro A1	América Latina (16 países): proporción de personas pobres e indicadores de precisión estadística, según condición de inmigración, alrededor de 2017.....	50
Cuadro A2	América Latina (15 países): tasa de desocupación e indicadores de calidad estadística, según condición de inmigración, alrededor de 2017.....	52
<b>Gráficos</b>		
Gráfico 1	Relación entre el tamaño de muestra y la precisión de un indicador utilizando la transformación Logit.....	18
Gráfico 2	Diagrama de flujo propuesto para la publicación, supresión y revisión de estimaciones de proporciones o razones en encuestas de hogares.....	31

## Resumen

Las encuestas de hogares utilizadas en los países de América Latina para cuantificar los principales indicadores sociales tienen un diseño complejo, que se caracteriza por ser estratificado, multietápico y con probabilidades de inclusión no uniformes. Las estimaciones elaboradas a partir de estas operaciones estadísticas están sujetas al error muestral, lo que hace necesario evaluar su validez estadística mediante diversos indicadores de precisión y confiabilidad.

Este documento describe un conjunto de criterios para evaluar la confiabilidad estadística de los indicadores estimados mediante encuestas de hogares. Su aplicación se ilustra mediante el cálculo de diferentes variables de interés para la población identificada como inmigrante. Se utiliza información reciente de 16 países, proveniente de las encuestas de hogares armonizadas del Banco de Datos de Encuestas de Hogares (BADEHOG) de la CEPAL.



## Introducción

Las encuestas de hogares son una de las fuentes de información más utilizadas para describir y cuantificar las condiciones de vida de las personas y de los hogares en los que habitan. Una característica atractiva de este instrumento es que indaga diversos temas a la vez, lo que permite realizar análisis cruzados entre diversas variables relevantes para una misma población y en un mismo momento. Esto es particularmente importante para producir desagregaciones de indicadores para distintos grupos de población.

Para asegurar que nadie quede atrás en el proceso de desarrollo hacia una mejora en las condiciones de vida, es indispensable obtener cifras representativas para diversos grupos de población. El marco global de indicadores para el seguimiento de la Agenda 2030 para el Desarrollo Sostenible plantea este requerimiento de manera explícita, indicando que “los indicadores deben desglosarse, cuando corresponda, por ingresos, sexo, edad, raza, etnia, estado migratorio, discapacidad y ubicación geográfica, u otras características, de conformidad con los Principios Fundamentales de Estadísticas Oficiales”.

Las estimaciones que es posible producir mediante encuestas de hogares tienen distintos niveles de error y no todas ellas cuentan con validez estadística. Las estadísticas y desagregaciones que han sido consideradas en el diseño de la encuesta alcanzan generalmente niveles adecuados de precisión. Sin embargo, muchos de los cruces de variables y desagregaciones que se realizan con base en los datos de las encuestas no han sido considerados en el diseño de este instrumento. Si bien ello no impide que estas desagregaciones puedan alcanzar niveles adecuados de precisión y confiabilidad, hace indispensable que se evalúe la calidad de los indicadores estimados antes de elaborar conclusiones o validar teorías acerca de la población en estudio.

La forma en que este documento aborda la estimación de la calidad de los indicadores está particularmente motivada por la necesidad de producir estimaciones comparables regionalmente a partir del Banco de Datos de Encuestas de Hogares (BADEHOG) de la CEPAL. A la fecha, este repositorio de información contiene más de 250 encuestas de hogares para 18 países de América Latina, con las cuales se calculan indicadores sobre pobreza, distribución del ingreso, educación, mercado laboral,



entre muchos otros, que se difunden a través de la base de datos CEPALSTAT y de numerosas publicaciones institucionales. En este contexto, se requiere contar con un mecanismo de fácil aplicación y que pueda ser generalizado a los diversos indicadores y desagregaciones que se producen con base en BADEHOG.

Una consideración adicional es que no todas las oficinas nacionales de estadística (ONE) de la región difunden los microdatos de las encuestas de hogares incluyendo la información sobre el diseño muestral (en particular, la identificación de las unidades primarias de muestreo y los estratos), que es indispensable para el cálculo adecuado del error muestral. En dichos casos, se emplea la aproximación propuesta en West y McCabe (2012).

Para ilustrar el cálculo de la confiabilidad estadística, el documento aborda en particular las desagregaciones según la condición migratoria de la población. El interés en este tema proviene no solo de que el fenómeno migratorio ha adquirido relevancia creciente en los últimos años en la región, sino además porque en general no es considerado en el diseño de las encuestas de hogares. Por tanto, cabe esperar que el cálculo de indicadores para la población inmigrante se encuentre, al menos en algunos casos, más allá de las posibilidades de los instrumentos utilizados.

A manera de ejemplo, el cuadro 1 presenta la proporción estimada de personas pobres desagregada en la subpoblación definida por su estado migratorio, en un país de la región. Esta tabla presenta algunas medidas de calidad que permiten determinar si la estimación puntual alcanza una precisión suficiente, indicada en la columna final. En este ejemplo, la estimación de la incidencia de la pobreza entre las personas migrantes no alcanza un nivel adecuado de precisión y, por tanto, esta cifra no debiera ser utilizada para describir la situación de pobreza en dicho grupo.

**Cuadro 1**  
**Estimación de la proporción de personas pobres desagregada por estado migratorio, junto con algunas medidas de precisión, para un país latinoamericano**

Sub-población	Prop. estimada	Límite inf.	Límite sup.	Coef. variación	Efecto diseño	Tamaño muestra	Tamaño muestra efectivo	Grados de libertad	Número de casos	Est. no precisa
Migrante	0,15	0,06	0,30	40,0	3,5	127	37	43	23	*
No migrante	0,35	0,33	0,37	2,6	14,5	38 074	2 620	929	12 698	
Total	0,35	0,33	0,37	2,6	14,5	38 201	2 632	929	12 721	

Fuente: Elaboración propia.

La estructura del documento es la siguiente: El capítulo I aborda los elementos necesarios que, desde la teoría estadística y de muestreo, se necesitan para establecer un análisis apropiado de las estimaciones, a través de conceptos de estratificación, cálculo de estadísticos y consideraciones importantes a la hora de realizar una inferencia apropiada del diseño complejo de las encuestas de hogares. Consecuentemente, el capítulo II describe los criterios de calidad recomendados como estadísticos de precisión para que, a su vez, sean utilizados en la revisión y supresión de cifras (como se ilustra en el cuadro 1). Asimismo, el capítulo III aclara los conceptos de variable de interés y subpoblación, introduce la secuencia lógica de reglas de supresión para estimadores basados en encuestas de hogares, expone cómo procesar estimaciones cuando las variables de diseño muestral no están disponibles siguiendo las recomendaciones de West y McCabe (2012). A modo de ejemplificación, el capítulo IV exhibe una aplicación para la subpoblación de migrantes en los países de América Latina, se aborda la medición del fenómeno y sus resultados para diferentes variables de interés.

## I. Principios básicos de estimación

Suponga que se tiene acceso a una base de datos que representa el levantamiento de una encuesta de hogares. Esta encuesta está inducida por la selección de una muestra  $s$  de tamaño  $n$  sobre la población de interés  $U$  de tamaño  $N$ . Para realizar la medición sobre las unidades pertenecientes a la muestra, se plantea un diseño de muestreo  $P(s)$ , que se asume probabilístico, estratificado y multietápico (Gutiérrez 2016).

El procesamiento de las bases de datos que vienen de encuestas de hogares debe tener en cuenta los criterios del diseño que se estableció para realizar el levantamiento de la información primaria. En particular, la base de datos debe contener como mínimo la siguiente información para cada individuo:

- *Estratos*: son particiones geográficas de la población para las cuales se definen selecciones independientes. Estas agrupaciones son mutuamente excluyentes e inducen  $H$  sub-grupos poblacionales.
- *Unidades primarias de muestreo*: son agregaciones de hogares definidas por un límite cartográfico proveniente del censo. Corresponden a la primera subdivisión de la población de hogares y están anidadas dentro de los estratos.
- *Pesos de muestreo*: corresponden a las ponderaciones utilizadas para representar a la población nacional a partir de los elementos de la muestra.

Aunque el marco de referencia de la teoría de muestreo es la estimación de un parámetro de interés sobre alguna característica de observación, lo cierto es que en la práctica no solo se necesitan estimaciones que abarquen a la población total sino también estimaciones para subgrupos poblacionales. En general, cuando se habla de subgrupos poblacionales se está haciendo referencia a dominios de interés, estratos o postestratos. Es indispensable conocer en qué se diferencian cada uno de ellos, pues de esto depende que la investigación arroje resultados confiables mediante el planteamiento de la mejor estrategia de muestreo.

## A. Análisis en subpoblaciones

Siendo  $U_1, \dots, U_g, \dots, U_G$  los subgrupos poblacionales tales que  $\bigcup_{g=1}^G U_g = U$ , y siendo  $N_g$  el tamaño absoluto del subgrupo  $U_g$ , entonces se tiene que  $\sum_{g=1}^G N_g = N$ . A partir de estas definiciones, se plantean algunas diferencias y similitudes entre los grupos poblacionales, que se resumen a continuación:

- **Dominios de interés:** Este tipo de subgrupos poblacionales son aquellos para los cuales se requieren estimaciones separadas. Estos requerimientos se planean en la etapa de diseño para asegurar que el diseño de la muestra sea tal que al momento de la recolección de la información exista una buena cobertura en cada uno de los dominios de interés. Lo anterior sólo se puede lograr ampliando el tamaño de muestra  $n$  puesto que el marco de muestreo no informa acerca de la pertenencia de los individuos a los dominios de interés. Un aspecto importante de esta clase de subgrupos poblacionales es que el número de individuos en la muestra que pertenecen a un dominio  $n_d$  de interés es siempre aleatorio, y para algunos dominios particulares puede llegar a ser muy pequeño. Por otro lado, el tamaño absoluto de cada dominio  $N_d$  no se conoce ni antes de la etapa de diseño ni después de la etapa de estimación. Un ejemplo claro de estos subgrupos es la condición de ocupación, la condición de pobreza, la rama de actividad, entre otros.
- **Estratos:** Cuando el marco de muestreo permite conocer la pertenencia de todos los individuos de la población a un subgrupo poblacional, se dice que esta clase de subgrupos se llaman estratos. Más aun, cuando se sabe que la característica de interés tiene un comportamiento distinto en cada uno de los estratos y se planea un diseño de muestreo que tenga en cuenta este aspecto mediante la selección aleatoria de unidades en cada uno de los estratos, se dice que el diseño de muestreo es estratificado. El aspecto fundamental de esta clase de subgrupos poblacionales es que el conocimiento de la pertenencia de los individuos a los estratos se incorpora en la etapa de diseño de la muestra. Nótese que, a diferencia de los dominios, en los estratos se conoce tanto  $N_h$  como  $n_h$  antes de la etapa de estimación. Un ejemplo claro de estos subgrupos son las zonas urbanas o rurales, regiones y municipios.
- **Postestratos:** La propiedad que caracteriza a este tipo de subgrupos poblacionales es que, aunque en la etapa de diseño el tamaño del postestrato  $N_g$  es conocido, se desconoce el número de individuos que pertenecerán al postestrato  $n_g$  en la muestra realizada. Un ejemplo claro de estos subgrupos son los grupos etarios, el sexo o la etnia que, si bien no son utilizadas en la fase de diseño, sí se utilizan sus proyecciones demográficas en la fase de análisis para analizar mejor la eficiencia de los estimadores. Al respecto Särndal, Swensson, y Wretman (2003) afirman que existen dos situaciones en las cuales se presenta esta situación, llamada comúnmente postestratificación:
  - Cuando el marco de muestreo es tal que se conoce la pertenencia de todos los elementos a los subgrupos poblacionales pero el investigador decide no utilizar esta información en la etapa de diseño. Las razones para esto son diversas, pero principalmente se decide obviar este tipo de información por facilidad logística. Una vez que se ha realizado la selección de la muestra, se observa la característica de interés  $y_k$  en los individuos tales que  $k \in s$ . El investigador decide utilizar la información auxiliar de pertenencia a los postestratos en la etapa de estimación para mejorar la eficiencia de la estrategia de muestreo, en particular del estimador propuesto.
  - Mediante alguna fuente de información confiable se conocen los tamaños absolutos  $N_g$  de cada subgrupo poblacional, aunque se desconoce la pertenencia de los

individuos a los subgrupos pues el marco de muestreo presenta esta deficiencia. Después de la etapa de diseño, se observa la característica de interés y se pregunta acerca de la pertenencia de los individuos seleccionados en los postestratos de tal forma que en la etapa estimación se utiliza esta información para mejorar la eficiencia de los estimadores de los parámetros de interés.

El diseño y rediseño de las encuestas se basan fundamentalmente en la búsqueda de estos subgrupos en la población. En todas las encuestas de hogares que se planean en América Latina se busca investigar fenómenos asociados a subgrupos poblacionales que se encuentran dispersos en la geografía de los países, con las siguientes características:

- El tamaño de muestra de una encuesta siempre se basa en la incidencia de un fenómeno que clasifica a la población en algún dominio de interés.
- Este tamaño de muestra siempre se reparte entre los diferentes estratos geográficos para mejorar la eficiencia del levantamiento de la información y del diseño de muestreo.
- En muchas ocasiones, las proyecciones demográficas sobre los postestratos son utilizadas en la fase de estimación para mejorar la precisión de los estimadores.

## B. Estimación puntual de promedios y proporciones

Teniendo en cuenta las anteriores definiciones, se asume que la variable de interés  $y$ , que puede ser de naturaleza continua o dicotómica, es observada sobre un individuo  $k \in s$ , el cual se encuentra clasificado como miembro de una subpoblación  $U_d$  ( $d = 1, \dots, D$ ). Por lo tanto, se tiene que:

$$y_{d_k} = y_k * z_{d_k} = \begin{cases} 1, & \text{Si } y_k = 1 \text{ y } k \in U_d \\ 0, & \text{en otro caso} \end{cases}$$

En donde  $y_k$  es el resultado de la observación de la variable de interés  $y$  en el individuo  $k$  y  $z_k$  es la variable indicadora de la subpoblación de interés  $U_d$ , definida de la siguiente manera:

$$z_{d_k} = \begin{cases} 1, & \text{Si } k \in U_d \\ 0, & \text{en otro caso} \end{cases}$$

De esta forma, la estimación de la proporción de personas en el dominio  $U_d$  para la variable  $y$  (que sería un promedio en el caso de que  $y$  fuese de naturaleza continua) es:

$$\hat{P}_d = \frac{\sum_h \sum_i \sum_k w_k y_{d_k}}{\sum_h \sum_i \sum_k w_k z_{d_k}} = \frac{\hat{t}_{y_d}}{\hat{t}_{z_d}}$$

Nótese que los tres sumandos corresponden a los estratos ( $h = 1 \dots, H$ ), las UPMS ( $i = 1, \dots, n_{1h}$ ) y los individuos ( $k = 1, \dots, n_i$ ). Este estimador corresponde a una razón, puesto que tanto el numerador como el denominador son aleatorios, y el ponderador  $w_k$  corresponde al peso de muestreo. La estimación para la proporción (o promedio, si es una variable continua) total definida en toda la población está dada por la siguiente expresión:

$$\hat{P}_U = \frac{\sum_h \sum_i \sum_k w_k y_k}{\sum_h \sum_i \sum_k w_k} = \frac{\hat{t}_y}{\hat{t}_z}$$

## C. Aproximación de la varianza

Debido a las dificultades algebraicas y computacionales, estimar la varianza en encuestas complejas que contemplan esquemas de conglomeración, selección en varias etapas y estratificación, puede tornarse bastante tedioso, costoso y además muy demorado. Por ello, los programas computacionales

existentes utilizan una aproximación conocida como la técnica del *último conglomerado*. Esta aproximación, que sólo tiene en cuenta la varianza de los estimadores en la primera etapa de muestreo, supone que ese muestreo fue realizado con reemplazo. Los procedimientos de muestreo en etapas posteriores de la selección son ignorados a menos que el factor de corrección para poblaciones finitas sea importante a nivel municipal.

En general, el último conglomerado es la primera unidad de muestreo en un diseño complejo. Por ejemplo, considere el siguiente diseño de muestreo en cuatro etapas:

**Sector**  $\Rightarrow$  **Vivienda**  $\Rightarrow$  **Hogar**  
 UPM                      USM                      UFM

En la primera etapa, las unidades primarias de muestreo (UPM) son los sectores cartográficos definidos a partir de los censos de población y vivienda; dentro de cada sector, se seleccionan viviendas, correspondientes a las unidades secundarias de muestreo (USM); de esta forma, el submuestreo continúa hasta seleccionar las unidades finales de muestreo (UFM) que son los hogares. Esta selección multietápica se realiza en cada uno de los estratos en los que se encuentra dividido el territorio nacional. Es necesario tener en cuenta esta particularidad de las encuestas para poder aplicar correctamente esta técnica de aproximación de varianzas. En resumen, para aquellas ciudades que pertenecen al estrato de inclusión forzosa, las UPM serán los sectores cartográficos, y para el resto del país, las UPM serán los municipios cuya probabilidad de inclusión en la muestra de la primera etapa es menor a uno.

En general, la idea es que el software reconozca tres elementos importantes en la base de datos: los estratos ( $h = 1, \dots, H$ ), las UPM ( $i \in s_{I_h}$ ) y los pesos finales de muestreo ( $w_k$ ). En las siguientes secciones, se exploran los detalles de esta técnica de aproximación de varianzas.

En particular considere cualquier estimador del total poblacional dado por la siguiente combinación lineal

$$\hat{t}_y = \sum_{k \in S} w_k y_k = \sum_{k \in U} I_k w_k y_k$$

en donde  $I_k$  son variables indicadoras de la pertenencia del elemento  $k$  a la muestra  $s$ . Ahora, asumamos que el factor de expansión de la encuesta  $w_k$  cumple con los supuestos básicos de un ponderador que hace insesgado a  $\hat{t}_y$ , es decir:

$$E(I_k w_k) = 1$$

Se supone un diseño de muestreo en varias etapas (dos o más) en donde la primera etapa supone la selección de una muestra  $s_{I_h}$  de  $n_{I_h}$  unidades primarias de muestreo ( $i \in s_{I_h}$ ) en el estrato  $h$  de tal forma que si la selección se realizó con reemplazo, la  $i$ -ésima UPM tiene probabilidad de selección  $p_{I_i}$ , mientras que, si la selección se realizó sin reemplazo, la  $i$ -ésima UPM tiene probabilidad de inclusión  $\pi_{I_i}$ .

En las subsiguientes etapas de muestreo, se procede a seleccionar una muestra de elementos para cada una de las UPM seleccionadas en la primera etapa de muestreo. Dentro de la  $i$ -ésima UPM se selecciona una muestra  $s_i$  de elementos; en particular la probabilidad condicional de que el  $k$ -ésimo elemento pertenezca a la muestra, dado que la UPM que la contiene ha sido seleccionada en la muestra de la primera etapa, está dada por la siguiente expresión:

$$\pi_{k|i} = Pr(k \in s_i | i \in s_{I_h})$$

Por ejemplo, dentro del estrato  $h$ , si el muestreo es sin reemplazo en todas sus etapas, la probabilidad de inclusión del  $k$ -ésimo elemento a la muestra  $s$  está dada por

$$\begin{aligned}
\pi_k &= Pr(k \in s) \\
&= Pr(k \in s_i, i \in s_{I_h}) \\
&= Pr(k \in s_i | i \in s_{I_h}) Pr(i \in s_{I_h}) = \pi_{k|i} \times \pi_{I_i}
\end{aligned}$$

Dado que el inverso de las probabilidades de inclusión son un ponderador natural, entonces se definen las siguientes cantidades:

- $w_{I_i} = \frac{1}{\pi_{I_i}}$ , que es el factor de expansión de la  $i$ -ésima UPM.
- $w_{k|i} = \frac{1}{\pi_{k|i}}$ , que es el factor de expansión del  $k$ -ésimo elemento dentro para la  $i$ -ésima UPM.
- $w_k = w_{I_i} \times w_{k|i}$ , que es el factor de expansión final del  $k$ -ésimo elemento para toda la población  $U$ .

Por lo tanto, cuando la muestra de UPM de la primera etapa fue seleccionada con reemplazo, entonces el estimador insesgado (conocido como el estimador de Hansen-Hurwitz) para el total poblacional está dado por la siguiente expresión.

$$\hat{t}_{y,p} = \sum_{h=1}^H \sum_{i \in s_{I_h}} \frac{1}{n_{I_h} p_{I_i}} \hat{t}_{y_i}$$

Y una estimación insesgada de su varianza es:

$$\widehat{Var}(\hat{t}_{y,p}) = \sum_{h=1}^H \sum_{i \in s_{I_h}} \frac{1}{n_{I_h}(n_{I_h} - 1)} \left( \frac{\hat{t}_{y_i}}{p_{I_i}} - \hat{t}_{y,p} \right)^2$$

Suponga ahora que la encuesta tiene un diseño bietápico  $P(s)$  que no contempla reemplazo en la primera etapa. Por lo tanto, algunas cantidades deben ser equiparadas para poder utilizar esta aproximación. En principio, nótese que las cantidades  $\hat{t}_{y_i}$  representan los totales estimados de la variable de interés en la  $i$ -ésima UPM y están dados por la siguiente expresión:

$$\hat{t}_{y_i} = \sum_{k \in s_i} \frac{y_k}{\pi_{k|i}} = \sum_{k \in s_i} w_{k|i} y_k$$

Utilizar la aproximación de la varianza requiere equiparar los términos de manera apropiada. En primer lugar, fijémonos en los estimadores dados por  $\hat{t}_{y,p}$  y  $\hat{t}_y$ . Para realizar esta comparación, se requiere que se asuma la siguiente igualdad en las probabilidades de inclusión de la primera etapa:

$$\pi_{I_i} = p_{I_i} \times n_{I_h}$$

Por lo tanto, el estimador del total poblacional quedaría definido desde  $\hat{t}_y$  como un estimador de tipo Hansen-Hurwitz.

$$\hat{t}_y = \sum_{k \in s} w_k y_k = \sum_{h=1}^H \sum_{i \in s_{I_h}} \sum_{k \in s_i} w_k y_k = \sum_{h=1}^H \sum_{i \in s_{I_h}} \sum_{k \in s_i} \frac{1}{\pi_{I_i} \pi_{k|i}} y_k = \sum_{h=1}^H \sum_{i \in s_{I_h}} \frac{\hat{t}_{y_i}}{\pi_{I_i}} \approx \sum_{h=1}^H \sum_{i \in s_{I_h}} \frac{1}{n_{I_h} p_{I_i}} \hat{t}_{y_i}$$

Ahora, dado que la forma del estimador ha sido equiparada con un estimador tipo Hansen-Hurwitz, es posible utilizar su estimación de varianza que resulta ser mucho más simple de calcular. Aún más, después de un poco de álgebra es posible tener una aproximación que solo hace uso de los factores de expansión finales  $w_k$ , los estratos y las UPM, que suelen ser reportados junto con los micro-datos de las encuestas de hogares, a diferencia de los factores de expansión de la primera etapa o los factores de expansión condicionales dentro de las UPM. Basado en lo anterior, al definir  $\check{t}_{y_i} = \sum_{k \in s_i} w_k y_k$  como la

contribución<sup>1</sup> de la  $i$ -ésima UPM a la estimación del total poblacional y  $\bar{\check{t}}_y = \frac{1}{n_{Ih}} \sum_{i=1}^{n_{Ih}} \check{t}_{yi}$  como la contribución promedio en el muestreo de la primera etapa, entonces el estimador de varianza toma la siguiente forma, conocida como el estimador de varianza del *último conglomerado*.

$$\widehat{Var}(\hat{t}_{y,p}) = \sum_{h=1}^H \sum_{i \in SI_h} \frac{n_{Ih}}{n_{Ih} - 1} \left( \check{t}_{yi} - \frac{1}{n_{Ih}} \sum_{i=1}^{n_{Ih}} \check{t}_{yi} \right)^2 = \sum_{h=1}^H \sum_{i \in SI_h} \frac{n_{Ih}}{n_{Ih} - 1} (\check{t}_{yi} - \bar{\check{t}}_y)^2$$

Utilizar la técnica del último conglomerado es una salida práctica al problema de la estimación de la varianza que, para la mayoría de las encuestas que brindan estadísticas oficiales a los países, puede tornarse bastante complejo. Si bien, la expresión no brinda estimaciones de varianza estrictamente insesgadas, sí constituye una aproximación bastante precisa, que en el peor de los casos sobre-estima ligeramente este parámetro (Wolter 2007).

La varianza de un estimador para un subgrupo de interés no es la misma cuando el subgrupo corresponde a un dominio que cuando el subgrupo corresponde a un estrato. En general, la estimación dentro de los estratos, en los que la membresía de la población es conocida en la etapa de diseño de la encuesta, arroja menores coeficientes de variación que la estimación dentro de los dominios. Este puede asumirse como el precio que se debe pagar ante el desconocimiento de la membresía de las unidades del marco de muestreo a los subgrupos poblacionales. En particular, el estimador de la varianza del último conglomerado cumple con esta regla.

Lumley (2010) plantea que cuando se trata de la estimación en dominios, la varianza involucra una cierta cantidad de ceros para todas aquellas unidades que no pertenecen al dominio. En efecto, operacionalmente, la estimación de la varianza en un subconjunto de la encuesta de hogares debe incluir varias contribuciones iguales a cero en la ecuación de estimación. Esto implica que las observaciones fuera del dominio se descartan con un filtro y las contribuciones nulas se agregan en el momento de la estimación de la varianza.

Cuando se requiere una estimación sobre un subgrupo poblacional que coincide con un estrato  $h'$  (o es una agregación de estratos), y teniendo en cuenta que los individuos de las UPM que no son de este estrato tomarán un valor nulo, tanto para la estimación del total, como para la aproximación de la varianza, entonces la forma de la aproximación se escribe de la siguiente manera:

$$\begin{aligned} \widehat{Var}(\hat{t}_y) &= \sum_{h=1}^H \sum_{i \in SI_h} \frac{n_{Ih}}{n_{Ih} - 1} (\check{t}_{yi} - \bar{\check{t}}_y)^2 \\ &= \frac{n_{Ih'}}{n_{Ih'} - 1} \sum_{i \in SI_{h'}} (\check{t}_{yi} - \bar{\check{t}}_{y_{h'}})^2 + \sum_{h \neq h'} \frac{n_{Ih}}{n_{Ih} - 1} \sum_{i \in SI_h} (\check{t}_{yi} - \bar{\check{t}}_{y_h})^2 \\ &= \frac{n_{Ih'}}{n_{Ih'} - 1} \sum_{i \in SI_{h'}} (\check{t}_{yi} - \bar{\check{t}}_{y_{h'}})^2 \end{aligned}$$

La anterior igualdad se tiene puesto que  $\sum_{h \neq h'} \frac{n_{Ih}}{n_{Ih} - 1} \sum_{i \in SI_h} (\check{t}_{yi} - \bar{\check{t}}_{y_h})^2 = 0$ , dado que  $\check{t}_{yi} = 0$  en estas UPM que no pertenecen al estrato  $h'$ . Sin embargo, esta reducción en la varianza no se tiene cuando los subgrupos poblacionales son dominios, puesto que  $\check{t}_{yi} \neq 0$ . Es decir, la cuantificación del precio que se debe pagar ante el desconocimiento de la membresía de las unidades al subgrupo es del orden de  $\sum_{h \neq h'} \frac{n_{Ih}}{n_{Ih} - 1} \sum_{i \in SI_h} (\check{t}_{yi} - \bar{\check{t}}_{y_h})^2$ .

<sup>1</sup> Note que la suma de estas contribuciones en la muestra de la primera etapa da como resultado la estimación  $\hat{t}_y$ .

## II. Criterios de calidad

### A. Intervalos de confianza

En general, la precisión de una estadística se debe estudiar a la luz del intervalo de confianza generado por la medida de probabilidad asociada al diseño de muestreo de la encuesta. Por ejemplo, si el parámetro de interés sobre el cual se busca realizar la inferencia es  $\theta$ , y se ha definido una subpoblación de interés  $U_g$ , entonces un intervalo del 95% de confianza sobre esa subpoblación está dado por la siguiente expresión (Heeringa, West, y Berglund 2010):

$$(\hat{\theta} - t_{0,975,gl} * se(\hat{\theta}), \hat{\theta} + t_{0,975,gl} * se(\hat{\theta}))$$

En donde  $\hat{\theta}$  es un estimador por muestreo para el parámetro de interés  $\theta$ ,  $t_{0,975,gl}$  es el percentil 0,975 de una distribución *t-student* con  $gl$  grados de libertad, que están dados por la resta entre el número de UPM seleccionadas menos el número de estratos de muestreo considerados y  $se(\hat{\theta})$  es el error estándar de la estimación, definido por la raíz cuadrada de la varianza del estimador; es decir:

$$se(\hat{\theta}) = \sqrt{\widehat{Var}(\hat{\theta})}$$

En el caso particular de las proporciones, los intervalos de confianza deben estar contenidos dentro del intervalo (0,1). Sin embargo, en algunas ocasiones puede ocurrir que el error estándar de una estimación cercana al 0 o al 1 sea demasiado grande y que el límite inferior, o superior del intervalo de confianza sea menor a cero, o mayor a uno, respectivamente. En este caso, es necesario estimar el intervalo de confianza con una variante que permita considerar estas restricciones. Una solución a este problema es considerar una transformación al estimador. De esta manera, si  $\hat{P}$  es una estimación de la proporción, se define la transformación Logit de la proporción.

$$\hat{L} = \log\left(\frac{\hat{P}}{1 - \hat{P}}\right) = \text{logit}(\hat{P})$$



Note que la aproximación de Taylor de primer orden para  $\hat{L}$  es:

$$\hat{L} \cong L(p) + \frac{\partial \hat{L}}{\partial \hat{P}} \Big|_{\hat{P}=p} (\hat{P} - p) = L(p) + \left( \frac{-1}{p(1-p)} \right) (\hat{P} - p)$$

Luego la varianza de  $\hat{L}$  se puede escribir como:

$$Var(\hat{L}) = AVar(\hat{L}) = \frac{Var(\hat{P})}{p^2(1-p)^2}$$

De esta forma, es posible definir un intervalo de  $(1 - \alpha)100\%$  de confianza para  $L$  como

$$\left( \hat{L} - t_{0,975,gl} \sqrt{Var(\hat{L})}, \hat{L} + t_{0,975,gl} \sqrt{Var(\hat{L})} \right) = (\hat{L}_1, \hat{L}_2)$$

Finalmente, de se tiene que

$$\hat{P} = \text{logit}^{-1}(\hat{L}) = \frac{\exp(\hat{L})}{1 + \exp(\hat{L})}$$

Por tanto, un intervalo de confianza para  $\hat{P}$  está por

$$\left( \text{logit}(\hat{L}_1), \text{logit}(\hat{L}_2) \right) = \left( \frac{\exp(\hat{L}_1)}{1 + \exp(\hat{L}_1)}, \frac{\exp(\hat{L}_2)}{1 + \exp(\hat{L}_2)} \right) \subseteq (0,1)$$

## B. Coeficientes de variación

Esta medida configura un acercamiento al error de muestreo que permite verificar si la inferencia es válida, su definición es como sigue:

$$CV(\hat{\theta}) = \frac{se(\hat{\theta})}{\hat{\theta}} = \frac{\sqrt{Var(\hat{\theta})}}{\hat{\theta}}$$

Esta medida de precisión de las estimaciones se ha consolidado como un estándar de calidad que ha permeado la práctica de las ONE en la publicación de estadísticas oficiales. Su uso es transversal puesto que, por su definición, tiene una naturaleza relativa, liberando al usuario de la unidad de medida inducida por la variable de interés. Además, es posible reformular los intervalos de confianza en términos del coeficiente de variación, de la siguiente manera:

$$\hat{\theta} \pm t_{0,975,gl} * se(\hat{\theta}) = \hat{\theta} (1 \pm t_{0,975,gl} * CV(\hat{\theta}))$$

Como lo afirman Singh, Westlake y Feder (2004), el coeficiente de variación es una medida de fácil interpretación, proporcional a la amplitud del intervalo de confianza, que provee una versión estandarizada y relativa de la precisión alrededor de la estimación puntual que permite comparar dos estimaciones del mismo indicador en diferentes sub-poblaciones, y que además es utilizada en el diseño y el re-diseño de las encuestas, entre otras cualidades. En general, no hay un consenso establecido acerca del valor que debe tomar el coeficiente de variación para que permita establecer si una estimación es precisa. Por ejemplo, desde el punto de vista teórico, Särndal, Swensson, y Wretman (2003) indican que "... un valor del coeficiente de variación del 2% es bueno, considerando las restricciones de la encuesta, mientras que un valor del coeficiente de variación de 9% puede ser considerado inaceptable." De esta forma, muchos institutos nacionales de estadística alrededor del mundo han considerado que las precisiones de las estadísticas resultantes de una encuesta estén

supeditadas al comportamiento de su coeficiente de variación. En el contexto de la calidad de las estimaciones provenientes de encuestas de hogares, mucho se ha discutido acerca del uso del coeficiente de variación en la validación de la confiabilidad y precisión de las cifras que provienen de estudios por muestreo.

Nótese que, cuando se están estimando proporciones, esta medida tiene algunas consideraciones importantes. En primer lugar, fijar un umbral para el coeficiente de variación tiene una interpretación directa sobre la amplitud relativa del intervalo de confianza. Por ejemplo, si la ONE decide fijar como umbral para el coeficiente de variación un 30%, esto implica que la amplitud relativa (AR) del intervalo de confianza estará alrededor de 118%, puesto que:

$$CV(\hat{\theta}) = 30\% \Rightarrow AR = \frac{2 * t_{0,975,gl} * se(\hat{\theta})}{\hat{\theta}} > 118\%$$

Por otro lado, como en todo fenómeno dicotómico resumido en un proporción, la varianza y el error estándar de la proporción obtiene su valor máximo en  $P = 0,5$ . Por lo tanto, en este valor es necesario aumentar el tamaño de muestra para asegurar la precisión definida. También debe tenerse en cuenta que a partir de  $P = 0,5$ , a derecha e izquierda, los fenómenos son simétricos. Por ejemplo, bajo este paradigma, la precisión de una proporción  $P = 0,9$ , es la misma que la de una proporción  $P = 0,1$ ; de la misma manera, la precisión de una proporción  $P = 0,7$ , es la misma que la de una proporción  $P = 0,3$ . Sin embargo, el coeficiente de variación no es una medida simétrica alrededor de  $P = 0,5$ , como sí lo es la varianza y el error estándar y, por su definición, cuando la proporción es pequeña, el coeficiente de variación tiende a ser muy grande, indicando que la precisión es baja.

### C. Coeficiente de variación logarítmico

El coeficiente de variación es una medida que define la precisión de un indicador, pero para el caso de las proporciones no constituye una medida simétrica, como sí lo es el error estándar o la varianza. Por ejemplo, suponga que se está estimando una proporción  $P$ , si la estimación del parámetro de interés es muy cercana a cero, sin importar que tan pequeña sea su varianza, el coeficiente de variación será muy grande y no representará la calidad de la estrategia de muestreo. Sin embargo, el coeficiente de variación del complemento de la proporción  $(1 - P)$  será muy pequeño y confiable. Esto se traduce en una paradoja, puesto que el mismo fenómeno está siendo medido, pero los coeficientes de variación son contradictorios. Debido a lo anterior, las estimaciones que tienen una magnitud pequeña (muy cercana a cero) son automáticamente castigadas por este indicador, incluso si la variabilidad de la cifra es pequeña.

Algunos autores han propuesto la posibilidad de realizar una transformación logarítmica sobre la proporción y utilizar su coeficiente de variación como una medida robusta del error de muestreo en las proporciones cercanas a cero y a uno, que además sea simétrica alrededor de  $P = 0,5$ , que es donde se maximiza la variabilidad de la proporción (Barnett-Walker et al. 2003). Por ende, si  $P \leq 0,5$ , se define  $\hat{L} = -\log(\hat{P})$ . En este caso, la aproximación de Taylor de primer orden es:

$$\hat{L} \cong L + \frac{\partial \hat{L}}{\partial \hat{P}} \Big|_{\hat{P}=P} (\hat{P} - P) = L + \left( \frac{-1}{P} \right) (\hat{P} - P)$$

Luego, la varianza de  $\hat{L}$  será  $Var(\hat{L}) \cong AVar(\hat{L}) = \frac{Var(\hat{P})}{P^2}$ , y por consiguiente el error estándar de la transformación equivaldrá al coeficiente de variación de la proporción, dado por:

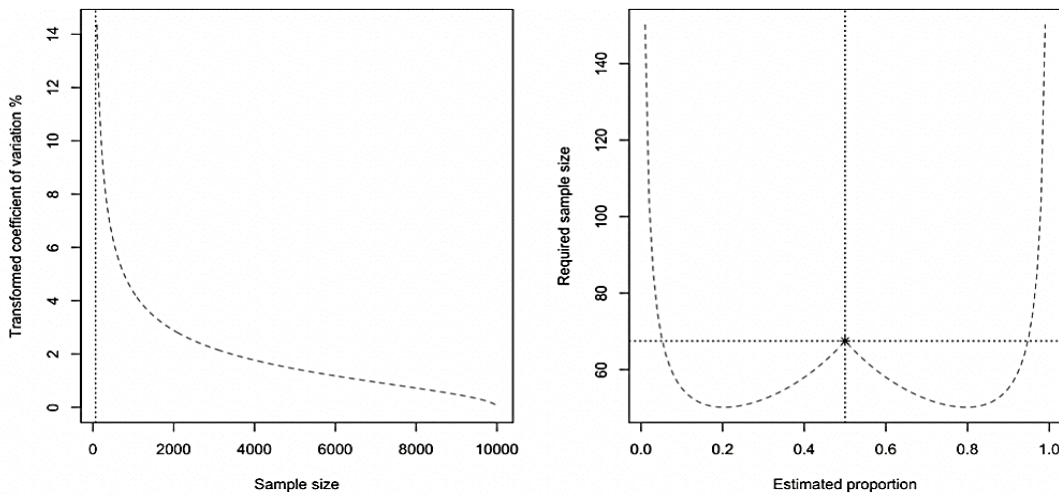
$$SE(\hat{L}) = \sqrt{AVar(\hat{L})} = \frac{\sqrt{Var(\hat{P})}}{\hat{P}} = CV(\hat{P})$$

De esta manera, podemos definir una medida de suavizamiento como el coeficiente de variación asociado a la transformación:

$$CV(\hat{L}) = \frac{SE(\hat{L})}{\hat{L}} = \frac{CV(\hat{P})}{\hat{L}}$$

De manera similar, para mantener la simetría, cuando  $P > 0,5$  se realiza un ajuste definiendo  $\hat{L} = -\log(1 - \hat{P})$ . Por lo tanto, para proporciones centrales, los coeficientes de variación de  $\hat{P}$  y  $\hat{L}$  serán comparables, puesto que  $\hat{L}$  toma valores cercanos a uno cuando  $P \in (0,2, 0,8)$ , y en este caso el  $CV(\hat{L})$  será similar a  $CV(\hat{P})$ .

**Gráfico 1**  
Relación entre el tamaño de muestra y la precisión de un indicador utilizando la transformación Logit



Fuente: Elaboración propia.

La figura muestra que, al igual que con el coeficiente de variación original, el tamaño de muestra aumentará a medida que se requiera mayor precisión en la estimación; pero a diferencia del coeficiente de variación original, el tamaño de muestra será idéntico para los fenómenos que induzcan proporciones simétricas. Además, el tamaño de muestra necesario para estimar eficientemente la proporción  $P$  con una precisión mayor a un determinado umbral del coeficiente de variación  $\delta$  es:

$$n \geq \frac{P(1-P)DEFF}{\frac{P(1-P)DEFF}{N} + \log^2(P)P^2\delta^2}$$

La anterior aproximación, para proporciones menorea a 0,5, se obtiene despejando  $n$  del siguiente desarrollo algebraico; análogamente para proporciones mayores a 0,5.

$$CV(\hat{L}) \leq \delta \Rightarrow \frac{SE(\hat{P})}{-\log(\hat{P})\hat{P}} \leq \delta \Rightarrow \frac{\sqrt{\frac{1}{n}(1-\frac{n}{N})S^2}}{-\log(\hat{P})\hat{P}} \leq \delta$$

## D. El efecto de diseño DEFF

Cuando se selecciona una muestra utilizando un diseño de muestreo complejo es muy improbable que exista independencia entre las observaciones. Además, como el muestreo de las encuestas de hogares es complejo, la distribución de la variable de interés no es la misma para todos los individuos. Por lo anterior, cuando se analizan datos que provienen de encuestas de hogares la inferencia correcta debe tener en cuenta estas grandes desviaciones con respecto al análisis estadístico clásico, que considera muestras aleatorias simples. Por ello, en la mayoría de las ocasiones se necesita aumentar el tamaño de muestra para obtener la precisión deseada. Una forma sencilla de incorporar el efecto del diseño complejo está dada por la siguiente relación, denotada como efecto de diseño (Kish 1965):

$$DEFF = \frac{Var(\hat{\theta})}{Var_{MAS}(\hat{\theta})}$$

En donde  $Var(\hat{\theta})$  denota la varianza de un estimador  $\hat{\theta}$  bajo un diseño de muestreo complejo  $P$  y  $Var_{MAS}(\hat{\theta})$  denota la varianza del este estimador  $\hat{\theta}$  bajo un diseño de muestreo aleatorio simple  $MAS$ . Esta cifra da cuenta del efecto de aglomeración causado por la utilización de un diseño de muestreo complejo ( $p$ ), frente a un diseño de muestreo aleatorio simple  $MAS$ , en la inferencia de un parámetro de la población finita  $\theta$  (que puede ser un total, un promedio, una proporción, una razón, un percentil, etc.). Por ejemplo, suponiendo que el parámetro de interés es la media poblacional ( $\bar{y}$ ) de una variable de interés  $y$  (por ejemplo, el ingreso per cápita mensual), es posible escribir la varianza del estimador bajo el diseño de muestreo complejo como

$$Var(\hat{y}) = \frac{DEFF}{n} \left(1 - \frac{n}{N}\right) S_y^2$$

En donde  $S_y^2$  corresponde a la varianza de las características de interés,  $N$  es el tamaño de la población de interés y  $n$  el tamaño de la muestra de individuos. Por otro lado, suponiendo que el parámetro de interés es la proporción poblacional ( $P$ ) de una variable dicotómica  $y$  (por ejemplo, el porcentaje de individuos de bajo de la línea de pobreza en un país), es posible escribir la varianza del estimador bajo el diseño de muestreo complejo como

$$Var(\hat{P}) = \frac{DEFF}{n} \left(1 - \frac{n}{N}\right) P(1 - P)$$

Por otro lado, en un diseño de muestreo bietápico, el efecto de diseño también se puede expresar de la siguiente manera

$$DEFF = 1 + (\bar{m} - 1)\rho_y$$

donde  $\bar{m}$  corresponde al promedio de hogares seleccionados por UPM y  $\rho_y$  es el coeficiente de correlación intraclase entre la variable de interés y las UPM. En general, nótese que el efecto de diseño será mayor cuando:

- i) El coeficiente de correlación crezca, lo cual no puede ser controlado de antemano, puesto que se trata de la observación de la realidad. En general,  $\rho_y$  será más grande cuando la distribución de la variable de interés sea explicada por las UPM en el país. Por ejemplo, si el indicador de interés es la pobreza y los hogares pobres están aglomerados, segregados y separados de los hogares más acaudalados, entonces  $\rho_y$  será más grande; además, entre más segregación haya, mayor será su valor.
- ii) El promedio de hogares seleccionados por UPM ascienda. Esto es controlado de antemano en la etapa de diseño y será un número fijo y transversal en la encuesta.

La estimación del efecto de diseño es un problema común cuando se trabaja con estimaciones desagregadas en subpoblaciones de interés. Cuando las subpoblaciones constituyen estratos (o agregaciones de estratos) planeados de antemano, para los cuales se conoce previamente su tamaño poblacional, se tiene el siguiente efecto de diseño:

$$DEFF_h = \frac{Var(\hat{\theta}_h)}{Var_{MAS}^h(\hat{\theta}_h)}$$

En donde  $Var_{MAS}^h(\hat{\theta}_h)$  es la varianza restringida al estrato  $h$  ( $h = 1, \dots, H$ ); en el caso en el que  $\hat{\theta}_h$  corresponda al estimador del promedio poblacional en el estrato  $h$ , su valor es el siguiente:

$$Var_{MAS}^h(\hat{\theta}_h) = \frac{1}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{y_h}^2$$

Siendo  $n_h$  el tamaño de la muestra en el estrato  $h$ ,  $N_h$  el tamaño poblacional del estrato  $h$  y  $S_{y_h}^2$  la varianza muestral de la variable de interés restringida al subgrupo  $h$ . Cuando la subpoblación de interés no es un estrato o un post-estrato sino un subgrupo aleatorio—como por ejemplo las personas pobres, las personas ocupadas, o cualquier otro subgrupo no planeado en el diseño de la encuesta o en la etapa de calibración—cuyo tamaño de muestra no es fijo (o condicionalmente fijo por la calibración) sino aleatorio, entonces la estimación correcta del efecto de diseño es la siguiente:

$$DEFF_U = \frac{Var(\hat{\theta}_h)}{Var_{MAS}^U(\hat{\theta}_h)}$$

En donde  $Var_{MAS}^U(\hat{\theta}_h)$  es la varianza poblacional del estimador de interés bajo un muestreo aleatorio simple. En el caso en el que  $\hat{\theta}_h$  corresponda al estimador del promedio poblacional en el estrato  $h$ , entonces su varianza  $Var_{MAS}^U(\hat{\theta}_h)$  estaría dada por la siguiente expresión:

$$Var_{MAS}^U(\hat{\theta}_h) = \frac{1}{n} \left(1 - \frac{n}{N}\right) S_{y_{hu}}^2$$

En donde  $S_{y_{hu}}^2$  es la varianza muestral de la variable de interés calculada en toda la población. Por lo tanto, en ambos efectos de diseño, la estimación de la varianza del diseño de muestreo complejo  $Var(\hat{\theta}_h)$  es la misma, pero el denominador cambia dependiendo de si el subgrupo es un estrato o no. Es por esta razón que en los softwares computacionales las cifras relacionadas con la estimación puntual, errores estándar, intervalos de confianza y coeficientes de variación coinciden plenamente. Sin embargo, tanto los softwares Stata como SPSS estiman por defecto  $DEFF_U$ . Nótese que, en este caso, las estimaciones de  $Var_{MAS}^U(\hat{\theta}_h)$  y  $Var_{MAS}^h(\hat{\theta}_h)$  serán diferentes, puesto que la primera involucra a toda la muestra, mientras que la segunda involucra únicamente a la muestra del estrato.

Retomando el ejemplo anterior, cuando los subgrupos de interés son agregaciones de los estratos de diseño, no es correcto utilizar el enfoque que Stata trae por defecto. En efecto, Lumley (2010) afirma que el efecto del diseño compara la varianza de una media o total con la varianza de un estudio del mismo tamaño utilizando un muestreo aleatorio simple sin reemplazo y que su cálculo será incorrecto si los pesos de muestreo se han re-escalado o no son recíprocos a las probabilidades de inclusión. Además, es posible comparar la varianza de la estimación con la varianza de una estimación basada en una muestra aleatoria simple del mismo tamaño que el de la subpoblación. Entonces, por ejemplo, en el muestreo aleatorio estratificado, el efecto de diseño calculado en un estrato será igual a uno. En el caso en que las variables del diseño de muestreo de la encuesta (UPM y estratos) no se encuentren en los microdatos de la base de datos, se recomienda seguir las indicaciones descritas en el capítulo III sección B.

## E. Tamaño de muestra

El tamaño de muestra afecta de manera indirecta la amplitud del intervalo de confianza, a través del error estándar, que generalmente decrece a medida que el tamaño de muestra se hace más grande. Un adecuado tamaño de muestra garantiza la convergencia en distribución de los estimadores a la distribución teórica de donde se calculan los percentiles en el cálculo del intervalo de confianza. En la fase de diseño, es posible mostrar que el tamaño de muestra requerido para estimar el promedio de una variable de interés en una encuesta de hogares, con un error de muestreo relativo menor a  $\delta \in (0,1)$  y una confianza estadística mayor a  $1 - \alpha$ , está dado por la siguiente expresión.

$$n \geq \frac{S_y^2 DEFF}{\frac{\delta^2 \bar{y}^2}{z_{1-\alpha/2}^2} + \frac{S_y^2 DEFF}{N}}$$

En donde  $z_{1-\alpha/2}$  es el percentil  $(1 - \alpha/2)$  asociado a una distribución normal estándar. Nótese que si  $\rho_y$  es grande, entonces el valor del efecto de diseño DEFF también lo será y por consiguiente el tamaño de muestra deberá ser más grande. Por ejemplo, al medir ingresos en la región, debido a la realidad económica de los países, es común encontrar que el tipo de hogar está altamente asociado con el ingreso de los individuos. Esto quiere decir que los ingresos no están uniformemente dispersos a través de todos los hogares, y por ende el coeficiente de correlación intraclase será alto. Por otro lado, si lo que se quiere estimar es una proporción  $P$ , entonces la expresión apropiada para calcular el tamaño de muestra estará dada por

$$n \geq \frac{P(1-P) DEFF}{\frac{\delta^2}{z_{1-\alpha/2}^2} + \frac{P(1-P) DEFF}{N}}$$

Como se puede apreciar, el tamaño de muestra es un indicador de la calidad de la encuesta, el cual resulta ser muy importante en la etapa de planeación y diseño. Sin embargo se tiene que considerar que:

- Si el parámetro de interés **sí** fue tenido en cuenta en la planeación de la encuesta con el propósito de tener representatividad sobre una subpoblación, entonces el tamaño de muestra será apropiado y, por ende, el error de muestreo estará controlado, al igual que el coeficiente de variación, el intervalo de confianza y la precisión de la inferencia será óptima.
- Si el parámetro de interés **sí** fue tenido en cuenta en la planeación de la encuesta, pero hubo una alta tasa de no respuesta, entonces el tamaño de muestra será mucho menor al planeado inicialmente y, por ende, el error de muestreo será más alto, al igual que el coeficiente de variación, y el intervalo de confianza será muy ancho, haciendo que la precisión de la inferencia no sea apropiada.
- Si el parámetro de interés **no** fue contemplado en la planeación y diseño de la encuesta de hogares, entonces es posible que el tamaño de muestra sea menor al necesario y, por ende, el error de muestreo será mayor, el coeficiente de variación será mayor, el intervalo de confianza será más amplio y la precisión de la inferencia será deficiente.

## F. Tamaño de muestra efectivo

El principio general detrás de esta medida está supeditado a que en la inferencia propia de las encuestas de hogares con diseños de muestreo complejos no existe una sucesión de variables que sean independientes e idénticamente distribuidas. Por lo tanto, si se piensa en la muestra  $(y_1, \dots, y_n)$  como un vector en el espacio  $n$ -dimensional, el estándar clásico de la teoría estadística asumiría que cada

componente del vector puede variar por sí mismo. Sin embargo, debido a la forma jerárquica de la selección de los hogares y a la interrelación de la variable de interés con las UPM, la variabilidad de la inferencia en las encuestas complejas tiene un fuerte componente asociados al mismo conglomerado, por lo que la dimensión final del vector  $(y_1, \dots, y_n)$  es mucho menor que  $n$ . De esta forma, se ha definido el tamaño de muestra efectivo (UN 2005, cap. 6) como sigue

$$n_{eff} = \frac{n}{DEFF}$$

En resumen, el diseño clásico de las encuestas de hogares consiste en seleccionar un conjunto de hogares dentro de una misma UPM y repetir esta estrategia de selección sistemáticamente en todo el país. Por lo tanto, se puede pensar en que, si la variable de interés tiene una alta correlación intraclase, entonces la realidad de las personas y de los hogares dentro de una misma UPM será muy homogénea, tanto que se podría interpretar como que la información estuviese repetida, y que los individuos u hogares de una misma UPM no estuvieran aportando de manera diferenciada. Por lo tanto, debido a los efectos del diseño de muestreo complejo, la cantidad de individuos que están aportando a la inferencia del indicador no es el número de personas, ni el número de hogares en la muestra, sino el tamaño de muestra efectivo  $n_{efectivo}$ , que deflacta los efectos de aglomeración.

## G. Grados de libertad

La amplitud del intervalo de confianza de un indicador no sólo está supeditada al error estándar, sino también al percentil de la distribución  $t$  – *student* con sus correspondientes grados de libertad. De esta manera, entre más grados de libertad se consideren, menor será la amplitud del intervalo y mayor será la precisión de la inferencia. En el caso más general en donde la subpoblación sea toda la población objetivo, los grados de libertad se reducen a la siguiente expresión:

$$gl = \#UPM - \#Estratos$$

Los grados de libertad constituyen una medida de cuántas unidades independientes de información se tienen en la inferencia. Nótese que, en el caso extremo de realizar un censo en cada UPM, sin importar el número de individuos que componen el conglomerado, el número de unidades independientes será únicamente el número de UPM seleccionadas en la primera etapa de muestreo puesto que la UPM es la unidad de muestreo que contribuye en mayor medida a la variabilidad de las estimaciones. En las aplicaciones reales de encuestas de hogares, en donde se realiza un submuestreo dentro de la UPM, la variabilidad de la estimación puede verse como la contribución del conglomerado a la gran media, más una contribución (considerada insignificante) de la segunda etapa de muestreo. Nótese la importancia de utilizar la distribución  $t$  – *student* como base inferencial para la construcción de los intervalos de confianza. Por ejemplo, considere el percentil 0.975 para el cual los valores críticos de la distribución  $t$  varían con respecto a sus grados de libertad; en este contexto se tiene que  $t - student_{gl=1} = 12,7$ ,  $t - student_{gl=2} = 4,30$ ,  $t - student_{gl=5} = 2,57$ ,  $t - student_{gl=40} = 2,02$  y  $t - student_{gl=\infty} = Z = 1,96$ .

A nivel desagregado, los grados de libertad son determinantes a la hora de hacer inferencias dentro de subpoblaciones de interés. En este caso los grados de libertad no se consideran fijos sino variables. Korn y Graubard (1999, 209) proponen el siguiente método de cálculo sobre los grados de libertad en una subpoblación  $U_g$ :

$$gl_g = \sum_{h=1}^H v_h * (n_{ih}^g - 1)$$

En donde  $v_h$  es una variable indicadora que toma el valor uno si el estrato  $h$  contiene uno o más casos de la subpoblación de interés y toma el valor cero en otro caso,  $n_{1h}^g$  es el número de unidades primarias de muestreo en el estrato  $h$  ( $h = 1, \dots, H$ ) con uno o más casos de la subpoblación. Al tomar esta definición como criterio de supresión, los funcionarios responsables del diseño y planificación de las encuestas en cada país, deben tener en cuenta que para que una cifra se publicada en un dominio de interés, esté deberá tener suficientes UPM (al menos más de 10) para que la regla no limite su diseminación.

## H. Conteo de casos no ponderado

El número de casos no ponderados en una muestra es simplemente el conteo de los individuos dentro de la muestra que son afectados por un fenómeno de interés en estudio. Este indicador tiene un efecto directo en la determinación de la precisión del estimador de interés y está determinado por la siguiente expresión.

$$n_y = \sum_s \delta_k^y$$

en donde  $\delta_k^y$  es una variable indicadora sobre cada individuo  $k$  de la muestra  $s$ , que toma el valor de uno si el individuo está afectado por el fenómeno representado por la variable de interés  $y$ . Nótese que esta es una cantidad aleatoria por definición, y también puede ser calculada en la muestra de un subgrupo poblacional específico  $U_{g_i}$ , de la siguiente manera:

$$n_y^g = \sum_s z_{gk} \delta_k^y = \sum_{s_g} \delta_k^y$$

Si la incidencia del fenómeno es muy baja (cuando la proporción  $P$  es cercana a cero), tanto el coeficiente de variación original y su transformación logarítmica tendrán magnitudes altas, puesto que:

$$\lim_{n_y \rightarrow 0} CV(\hat{\theta}) = \lim_{n_y \rightarrow 0} CV(\hat{L}) = \infty$$

En muchos países las encuestas de hogares son usadas por las autoridades gubernamentales para asignar recursos a una población potencial. En estos casos, es de particular interés conocer el número de personas que serán susceptible de participar en la repartición de recursos. Por ende, si la estimación de la incidencia total del fenómeno en la población no es precisa, difícilmente se podrá establecer un rubro presupuestal para atender a esta población. Por ejemplo, si la estimación del total de personas afectadas por el fenómeno es del orden de 5% y su margen de error es 5%, entonces el coeficiente de variación será de 100% y el intervalo de confianza de la proporción será (0%, 10%), demasiado amplio para tomar algún tipo de decisión sobre los recursos públicos de un país. Nótese que esta amplitud se magnifica cuando el número de casos no ponderado no es suficiente.





### III. Aplicación de los criterios de calidad

#### A. Definición de variables y subpoblaciones

El análisis apropiado de las estadísticas generadas a partir de las encuestas de hogares debe pasar por una definición clara de la variable que se quiere analizar y de las subpoblaciones para las cuales se calcula. En algunas ocasiones, estas definiciones pueden resultar confusas, ya que hay casos en los que la variable de análisis puede definir una subpoblación. Se presentan a continuación algunos ejemplos de indicadores de interés y las definiciones necesarias para el cálculo de las medidas de calidad correspondientes.

##### a) Promedio del ingreso per cápita en el país

En este caso la variable de interés es una característica continua  $y_k \geq 0$  ( $\forall k \in U$ ) definida sobre toda la población del país. El indicador se escribe como una razón:

$$\hat{y}_{nacional} = \frac{\hat{t}_y}{\hat{N}} = \frac{\sum_h \sum_i \sum_k w_k y_k}{\sum_h \sum_i \sum_k w_k}$$

en donde los subíndices  $h$ ,  $i$  y  $k$ , se refieren a los estratos, las UPM y los individuos, respectivamente. Nótese que la variable que define la población es siempre determinista puesto que  $z_{gk} = 1$  para todos los individuos que residen en el país, es decir para todos los individuos de la muestra. En este caso los grados de libertad corresponden a todas las UPM menos todos los estratos de la encuesta en el país.

##### b) Promedio del ingreso per cápita en una ciudad

En este caso la variable de interés está definida sobre un subgrupo poblacional  $U_g$ , correspondiente a la ciudad de interés. El estimador del indicador se escribe como una razón.

$$\hat{y}_{ciudad} = \frac{\hat{t}_{y_g}}{\hat{N}_g} = \frac{\sum_h \sum_i \sum_k w_k z_{gk} y_k}{\sum_h \sum_i \sum_k w_k z_{gk}}$$

Nótese que la variable que define la subpoblación es dicotómica dada por

$$z_{gk} = \begin{cases} 1, & \text{Si } k \text{ reside en la ciudad } U_g \\ 0, & \text{en otro caso} \end{cases}$$

El tamaño de muestra es  $n_g = \sum_s z_{gk}$ , es decir el tamaño de muestra de la ciudad; los grados de libertad corresponden a todas las UPM en la ciudad menos todos los estratos en la ciudad.

### c) Proporción de personas pobres en el área urbana

En este ejemplo, el estimador del indicador se escribe como una razón sobre el área urbana  $U_g$ .

$$\hat{P}_{urbano} = \frac{\hat{t}_{y_g}}{\hat{N}_g} = \frac{\sum_h \sum_i \sum_k w_k z_{gk} y_k}{\sum_h \sum_i \sum_k w_k z_{gk}}$$

En donde  $y_k$  es la variable de interés que define una característica dicotómica de la siguiente manera

$$y_k = \begin{cases} 1, & \text{si el ingreso per cápita de la persona está por debajo de la línea de pobreza} \\ 0, & \text{en otro caso.} \end{cases}$$

Las mediciones se realizan sobre la subpoblación definida por la siguiente variable

$$z_{gk} = \begin{cases} 1, & \text{si la persona reside en el área urbana } U_g \\ 0, & \text{en otro caso.} \end{cases}$$

En este caso el tamaño de muestra es  $n_g = \sum_s z_{gk}$ , es decir el tamaño de muestra del área urbana; los grados de libertad corresponden a todas las UPM del área urbana menos todos los estratos del área urbana.

### d) Tasa de desocupación nacional

Este indicador está definido como la división entre el total de personas desocupadas sobre el total de personas activas en la fuerza de trabajo. El estimador del indicador está definido como una razón de dos estimadores de totales poblacionales:

$$\widehat{TD}_{nacional} = \frac{\hat{t}_y}{\hat{t}_z} = \frac{\sum_h \sum_i \sum_k w_k z_k y_k}{\sum_h \sum_i \sum_k w_k z_k}$$

En donde las variables de interés toman la siguiente forma

$$y_k = \begin{cases} 1, & \text{si el individuo es desocupado,} \\ 0, & \text{si el individuo no es desocupado,} \\ NA, & \text{si no está en edad de trabajar.} \end{cases}$$

La variable que define la subpoblación es

$$z_k = \begin{cases} 1, & \text{si el individuo es activo,} \\ 0, & \text{si el individuo es inactivo,} \\ NA, & \text{si el individuo no está en edad de trabajar.} \end{cases}$$

En este caso el tamaño de muestra es  $n = \sum_s z_k$ , es decir el número de personas en la muestra que están en edad de trabajar y son activas. Los grados de libertad corresponden a todas las UPM menos todos los estratos de la encuesta en el país en los que se encontraron hogares con individuos en edad de trabajar y activos. Además, el conteo de casos no ponderado corresponde al número de individuos desocupados en la muestra.

### e) Tasa de desocupación masculina en migrantes

Este indicador está definido como la división entre el total de hombres migrantes desocupados sobre el total de hombres migrantes activos. El estimador del indicador está definido como una razón de dos estimadores de totales poblacionales:

$$\widehat{TD}_{\text{hombre-migrante}} = \frac{\hat{t}_{y_g}}{\hat{t}_{z_g}} = \frac{\sum_h \sum_i \sum_k w_k z_{gk} y_k}{\sum_h \sum_i \sum_k w_k z_{gk}}$$

En donde las variables de interés toman la siguiente forma

$$y_k = \begin{cases} 1, & \text{si el individuo es desocupado,} \\ 0, & \text{si el individuo no es desocupado,} \\ NA, & \text{si no está en edad de trabajar.} \end{cases}$$

La variable que define la subpoblación es

$$z_{gk} = \begin{cases} 1, & \text{si el individuo es activo, hombre y migrante,} \\ 0, & \text{si el individuo es inactivo, hombre y migrante} \\ NA, & \text{si el individuo no está en edad de trabajar, o es mujer o es no migrante.} \end{cases}$$

En este caso el tamaño de muestra es  $n = \sum_s z_{gk}$ , es decir el número de personas en la muestra que están en edad de trabajar, son hombres migrantes y están activos. El conteo no ponderado de casos corresponde al número de individuos en la muestra que son hombres migrantes y están desocupados. Los grados de libertad corresponden a todas las UPM menos todos los estratos de la encuesta en el país en los que se encontraron hogares con hombres migrantes y activos en la fuerza de trabajo.

## B. Análisis de bases de datos sin UPM ni estratos

En algunos casos, las bases de datos de las encuestas de hogares disponibles para los usuarios externos a las ONE no contienen toda la información necesaria para calcular la varianza de los estimadores. Es posible que la base de datos publicada omita estas variables para evitar proveer información geográfica que pudiera permitir la identificación de los informantes, y así mantener su anonimidad.

En este contexto, algunos autores han abordado el problema de estimar los errores de muestreo cuando la base de datos únicamente incorporan los factores de expansión y omite la información de las unidades primarias de muestreo y de los estratos. Una solución parsimoniosa propuesta por West y McCabe (2012) plantea la aproximación del efecto de diseño general a partir de la información disponible en las publicaciones de las ONE. Nótese que el efecto de un diseño complejo cumple la siguiente propiedad:

$$Var(\hat{\theta}) = DEFF \times Var_{MAS}(\hat{\theta})$$

Park (2003) ha propuesto que el efecto de diseño se puede descomponer en tres partes que se relacionan entre sí de forma multiplicativa. En primer lugar está el efecto debido a la ponderación desigual,  $deff^W$ ; en segundo lugar se encuentra el efecto debido a la estratificación,  $deff^S$ ; y por último se tiene el efecto debido al muestreo en varias etapas,  $deff^C$ . Por lo tanto:

$$DEFF = deff^W \times deff^S \times deff^C$$

A su vez, el efecto de diseño debido a la ponderación se expresa de la siguiente manera:

$$Var_{MAS}(\hat{\theta}) = \frac{Var^W(\hat{\theta})}{deff^W}$$

En donde  $Var^W(\hat{\theta})$  se refiere a la varianza de un diseño de muestreo con ponderadores desiguales. Bajo la hipótesis de que se conoce una aproximación  $DEFF^*$  al efecto de diseño total (obtenida de alguna publicación o fuente similar) y teniendo en cuenta de que el efecto de diseño y la varianza debido a la ponderación desigual se pueden calcular fácilmente con los datos de la encuesta, entonces una mejor aproximación a la varianza del estimador es:

$$Var(\hat{\theta}) = Var^w(\hat{\theta}) \times \frac{DEFF^*}{def f^w}$$

En las aplicaciones prácticas descritas en el capítulo IV, se utilizó este acercamiento para aquellos países para los cuales no se dispóní de la información exhaustiva del diseño de muestreo, como las variables de estratifo y UPM. En particular, se decidió que la aproximación  $DEFF^*$  fuese equivalente a la mediana de los efectos de diseño de todas aquellas encuestas que sí proveían esta información en sus bases de datos.

## C. Secuencia lógica para crear reglas de supresión

En este documento se enfatiza que la precisión de una estimación está ligada a los intervalos de confianza, cuya amplitud permitiría establecer si una estimación es precisa o no. Nótese que lo anterior se basa en que la amplitud de los intervalos de confianza muestra el grado de seguridad que es posible tener respecto a la precisión de un estimador. Considere los siguientes ejemplos prácticos:

- La incidencia de la pobreza en un departamento de un país se estimó en 5,2%, con un intervalo de confianza de (5,15%, 5,25%).
- La tasa de desocupación en el país para los hombres se ubicó en 7,5%, con un intervalo de confianza de (7,1%, 7,9%); mientras que para las mujeres se ubicó en 9,2%, con intervalo de confianza de (8,8%, 9,6%).
- La tasa de asistencia neta estudiantil en primaria para el último quintil de ingreso se estimó en 85%, con un intervalo de confianza de (48,2%, 100%).

El primer ejemplo da cuenta de un caso en el que el indicador se estima con una precisión elevada que permite su utilización para diversos fines. En cambio, en el tercer ejemplo claramente no se alcanza la precisión necesaria para que una ONE publique esta cifra confiadamente, o para que un gobierno pueda realizar algún tipo de política pública educativa, y mucho menos para estimar los recursos necesarios para una intervención sobre la población de interés. Sin embargo, las siguientes dos características impiden que la amplitud del intervalo pueda ser usada como criterio estándar en la supresión de cifras:

- En primer lugar, se debe comprender que no todos los indicadores se mueven en el mismo rango de posibles valores y que no es posible comparar la amplitud del intervalo de confianza para la estimación de una proporción (que toma valores entre cero y uno) con la estimación de una media (que toma valores mayores que cero).
- Existen ocasiones en que el resultado de la estimación de una proporción es cero o uno, puesto que las observaciones muestrales siempre detectaron casos en un solo espectro de la variable dicotómica. En este caso, la estimación de la varianza del estimador de la proporción es cero, y por consiguiente la amplitud del intervalo de confianza será nula. Es posible encontrar este tipo de situaciones cuando el tamaño de muestra es insuficiente.

Por otro lado, utilizar únicamente el coeficiente de variación como estándar para la supresión de cifras es un criterio que no tiene en cuenta toda las variantes asociadas a la inferencia en un muestreo complejo. De esta forma, se recomienda estudiar los elementos fundamentales que definen la precisión de las estimaciones para crear una secuencia lógica de revisión, publicación o supresión de cifras. A continuación se concentran estas recomendaciones que incorporan criterios adicionales.

- **Coeficiente de variación:** después de realizar una revisión de las experiencias internacionales, con base en la información publicada en las páginas web de las ONE, para determinar cómo son usados los criterios de supresión de información y los umbrales que se

definen para la validación de las cifras, Gutiérrez (2018) plantea que los países del Mercosur y Estados Unidos utilizan un umbral de  $CV > 30\%$ , Canadá y México usan como referencia un umbral del  $CV > 25\%$ , Chile y Costa Rica utilizan un umbral del  $CV > 20\%$ , Ecuador y Perú utilizan un umbral del  $CV > 15\%$ , mientras que Colombia usa un umbral del  $CV > 10\%$ . De esta forma, cualquier cifra estimada cuyo coeficiente de variación sea mayor al umbral predefinido es suprimida o marcada como una cifra poco confiable. En este contexto, un  $CV > 20\%$  puede asumirse como un valor de referencia útil a nivel regional.

- **Coeficiente de variación logarítmico:** esta medida de suavizamiento toma valores altos cuando las proporciones estimadas están demasiado cercanas a cero o a uno. Barnett-Walker et al. (2003) proponen que la cifra debe ser suprimida si el coeficiente de variación logarítmico es mayor que 17,5%.
- **Tamaño de muestra:** este criterio debe ser considerado como uno de los más importantes a la hora de decidir la ruta de publicación de una cifra, puesto que los desarrollos teóricos en términos de inferencia estadística para encuestas dependen de este término. La cobertura de los intervalos de confianza y la distribución de los estimadores dependen de que tanto el tamaño de la subpoblación como su tamaño de muestra asociado no sean pequeños. En este espíritu, Barnett-Walker et al. (2003) proponen que todas las estimaciones basadas en un tamaño de muestra menor a 100 unidades deberían ser suprimidas o marcadas como no confiables.
- **Tamaño de muestra efectivo:** al igual que con el anterior criterio, el tamaño de muestra efectivo induce que las aproximaciones teóricas, en términos de convergencia de las distribuciones de los estimadores y la cobertura de los intervalos de confianza, se cumplan. Hornik et al. (2002) consideran que, si el tamaño de muestra efectivo no es mayor a 140, entonces la cifra no debería ser considerada para publicación. Por otro lado, teniendo en cuenta el tamaño de muestra inducido por la transformación logarítmica, Barnett-Walker et al. (2003) afirman que cuando la proporción se encuentra entre 0,05 y 0,95, entonces el tamaño de muestra efectivos máximo cuando  $P = 0,5$ , siendo su valor  $n_{eff} = 68$ , tal como se puede ver en el gráfico 1.
- **Conteo de casos no ponderado:** cuando la incidencia de un fenómeno es muy baja y el diseño de la encuesta no lo tuvo en cuenta, entonces es posible que las estimaciones asociadas a tamaños, totales y proporciones sobre este fenómeno no sean confiables. En particular, para las proporciones es posible restringir las estimaciones tales que  $\hat{P} < 0,001$ , pero es más expedito crear una regla a partir del conteo de casos en la muestra. Por ejemplo, National Research Council (2015) plantea que si el número de casos no ponderados es menor a 50 unidades entonces la estimación no es publicada.
- **Grados de libertad:** este criterio apunta a aislar el efecto inflacionario del tamaño de muestra en una encuesta compleja y plantea una aproximación al número de unidades independientes en la inferencia. Además, a medida que crece, la amplitud del intervalo de confianza se estabiliza. Parker, Talih, y Malec (2017) consideran que, si los grados de libertad inducidos por la subpoblación son menos de 8, la cifra debería ser suprimida.

Los criterios que se presentan en esta sección contribuyen a determinar si una estadística puede ser considerada como precisa. En el caso particular de las proporciones, el resultado final del proceso puede resumirse en una tabla con la estructura que se muestra en el cuadro 2. Nótese que es posible que los criterios mencionados den lugar a resultados contradictorios si se aplican de manera independiente. Por ejemplo, para una variable con poca homogeneidad en las UPM, con un tamaño de muestra de  $n = 90$  y un efecto de diseño de  $DEFF = 0,5$ , implicaría un tamaño de muestra efectivo de

$n_{eff} = 180$ . En este caso, si los criterios de supresión se aplicaran de manera independiente, se concluiría que la cifra debería ser suprimida por tener un tamaño de muestra insuficiente, pero a la vez, que la cifra debería ser publicada, por tener un tamaño de muestra efectivo suficiente. Lo anterior, podría llevar a contradicciones y malas interpretaciones por parte los usuarios finales de los datos.

**Cuadro 2**  
**Estructura de resultados del procesamiento de una proporción en el repositorio BADEHOG**

Sub-población	Proporción estimada	Límite inferior	Límite superior	Coficiente de variación	Coficiente de variación logarítmico	Efecto de diseño	Tamaño de muestra	Tamaño de muestra efectivo	Número de casos	Grados de libertad
$U_1$	$\hat{P}_1$	$LI_1$	$LS_1$	$CV_1$	$CVL_1$	$DEFF_1$	$n_1^*$	$n_1^y$	$n_1$	$gl_1$
$U_2$	$\hat{P}_2$	$LI_2$	$LS_2$	$CV_2$	$CVL_2$	$DEFF_2$	$n_2^*$	$n_2^y$	$n_2$	$gl_2$
...	...	...	...	...	...	...	...	...	...	...
$U_G$	$\hat{P}_G$	$LI_G$	$LS_G$	$CV_G$	$CVL_G$	$DEFF_G$	$n_G^*$	$n_G^y$	$n_G$	$gl_G$
$U$	$\hat{P}$	$LI$	$LS$	$CV$	$CVL$	$DEFF$	$n^*$	$n^y$	$n$	$gl$

Fuente: Elaboración propia.

Para evitar que las reglas de supresión generen resultados contradictorios, es recomendable que estas se apliquen siguiendo una secuencia lógica, como la que se presenta en el gráfico 2 que muestra una propuesta de criterios de supresión de cifras para la estimación de proporciones o razones. En una primera instancia se realiza la estimación clásica de los parámetros de interés y se genera una tabla que adjunte el cálculo de todos los criterios descritos anteriormente. Luego, dependiendo de la naturaleza del fenómeno investigado, se deben establecer los criterios que se van a tener en cuenta y los umbrales en cada caso. El próximo paso es decidir, para cada cifra de la tabla generada, si se va a publicar o suprimir, y en algunos casos si se revisará la cifra con mayor detenimiento.

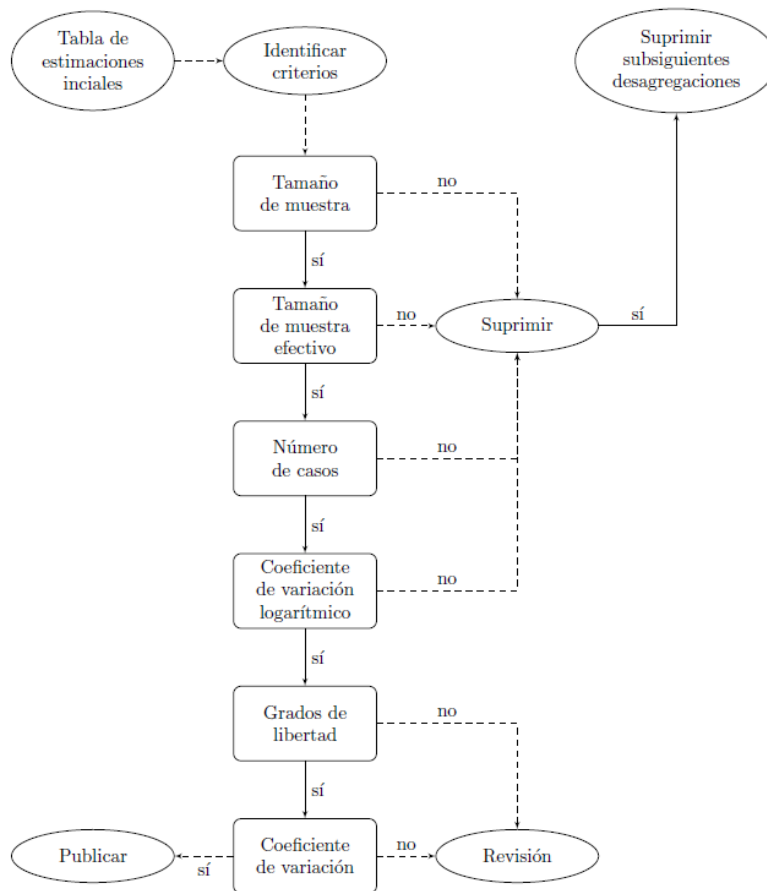
Por ejemplo, en el diagrama propuesto se definen seis criterios como condiciones necesarias para la publicación inmediata de una cifra. Si alguno de los primeros cuatro criterios no se satisface, entonces la cifra es suprimida. Si esos cuatro criterios se satisfacen, pero no se cumple alguno de los dos criterios siguientes, es deseable que la cifra se revise. De manera general, es deseable que las ONE analicen a profundidad sus políticas de supresión, revisión y publicación de cifras para cada una de las encuestas que realiza y, de manera independiente, definan las reglas apropiadas para cada caso. Para cada encuesta se debería considerar un algoritmo particular; es decir, los criterios de supresión no necesariamente deben coincidir entre distintas operaciones estadísticas.

Es de esperar que en el día a día de la producción de estadísticas oficiales en una ONE, el paso de revisión de cifras sea omitido para dar lugar a un algoritmo más expedito que permita automatizar y facilitar la publicación y supresión de las cifras. En general, en este proceso de producción, si una estimación no cumple con los criterios de calidad, no se debiera publicar y se recomienda informar la razón por la cual esta cifra es suprimida. Sin embargo, cuando las bases de datos son puestas a disposición del público general, es posible que en algunas situaciones específicas, bajo una justificación apropiada, los investigadores que hacen uso de estas bases de datos utilicen cifras que cumplan con la mayoría de los criterios presentados en este documento.

De tal forma, existirían dos razones por las cuales una cifra podría entrar en este proceso de revisión: cuando el coeficiente de variación y los grados de libertad no cumplen con los umbrales, aunque los otros cuatro criterios sí lo hagan. En este tipo de situaciones, el analista debe considerar de forma particular el contexto de su investigación y las implicaciones que tendría la utilización de la cifra. Por ejemplo, si la estimación puntual está muy cercana a cero, es posible que el coeficiente de variación sea muy alto, pero que a su vez se tenga un tamaño de muestra adecuado y los demás criterios de calidad se cumplan satisfactoriamente. En cualquier caso, si después del proceso de revisión, se

considera que la cifra debiera ser utilizada por su importancia y adecuada interpretación en el contexto de la investigación, se recomienda hacer explícita la justificación de esta decisión. Por último, si se ha decidido suprimir la publicación de una cifra por no cumplir con los criterios de calidad, se sobreentiende que cualquier desagregación subsiguiente también deberá ser suprimida.

**Gráfico 2**  
**Diagrama de flujo propuesto para la publicación, supresión y revisión de estimaciones de proporciones o razones en encuestas de hogares**



Fuente: Elaboración propia.





## IV. Aplicación a la estimación de la población inmigrante

El objetivo de este capítulo es realizar una aplicación de los conceptos y métodos antes expuestos para la subpoblación de migrantes internacionales en los países de América Latina, utilizando como fuente de información las encuestas de hogares. Cabe mencionar que la sección no apunta a establecer inferencias concluyentes sobre las características de los migrantes en una población, sino más bien exponer los diferentes criterios de calidad y establecer cuándo las cifras estimadas deben ser revisadas o suprimidas.

### A. Definición estadística de la migración internacional

La migración internacional es uno de los principales componentes en los modelos de crecimiento demográfico y se integra al cálculo de una determinada población en un período de interés. Las Recomendaciones de Naciones Unidas sobre estadísticas de migración internacional (Naciones Unidas, 1998) definen un **migrante internacional** como una persona que cambia su país de residencia habitual.

El concepto de residencia habitual está presente en varios marcos conceptuales. En el Sistema de Cuentas Nacionales (CEPAL, 2008) se adopta la definición del Manual de Balanza de Pagos (IMF, 2009). Así, "...una unidad institucional es residente en el territorio económico de un país cuando mantiene un centro de interés económico predominante en ese territorio, es decir, cuando realiza o pretende realizar actividades económicas o transacciones a una escala significativa, indefinidamente o durante un período de tiempo prolongado, que normalmente se interpreta como un año...". La precisión sobre el período de tiempo anual es tomada como parte de la definición operativa de residencia y permite facilitar su comparabilidad entre los distintos países. Una definición similar se utiliza en las estadísticas internacionales de turismo.

En general, la referencia al año puede ser interpretada de dos maneras (Naciones Unidas, 2017), tomando como criterio de definición la mayor parte del año anterior: es decir, más de 6 de los últimos 12 meses (excluyendo salidas por vacaciones o trabajo), o la permanencia continuada en los últimos 12 meses (excluyendo salidas por vacaciones o trabajo). Este último criterio se entiende como el más adecuado para la estimación de la migración internacional y es consistente con el enunciado en la

definición de residencia del sistema de cuentas nacionales. Debido a las dificultades que impone la restricción de permanencia de al menos 12 meses en el país de destino para configurar la situación de residente y dado que existen personas que pueden ser captadas estadísticamente en un país que no es el de su residencia habitual, en las recomendaciones se introduce una *definición operativa de migrante internacional*. Según esta definición, una persona será considerada un *inmigrante* en un determinado país si ha ingresado al país por alguna de sus fronteras, era residente de un tercer país antes de entrar al país en cuestión o no ha alcanzado aún la condición de residente en el mismo, y ha estado o pretende estar en el país por al menos un año. Por otro lado, una persona será considerada un *emigrante* de un determinado país si ha egresado del país por alguna de sus fronteras, era residente del país antes de egresar del mismo, y ha estado o pretende estar en un tercer país por al menos un año.

Estas definiciones no se aplican a grupos especiales, compuestos habitualmente por integrantes de los cuerpos diplomáticos y organismos internacionales cumpliendo su misión en terceros países; integrantes de las fuerzas armadas desarrollando tareas fuera del territorio nacional; personas integrantes de pueblos nómades; trabajadores fronterizos que cuentan con permisos especiales para desarrollar sus actividades laborales en países limítrofes. En el caso de estudiantes internacionales y otros colectivos que viven fuera de fronteras parte del año, suele investigarse si la estancia en el país de destino ha sido continuada en los doce meses anteriores para asignar la residencia a uno u otro país. En el caso en que los períodos de permanencia en ambos países sean similares, suele recurrirse a criterios adicionales a los hasta ahora señalados, pudiendo incluir el país en el cual trabajan, donde abonan sus impuestos, donde vive su familia o donde poseen la mayor parte de sus propiedades.

## **B. La forma de medir la migración en las encuestas de hogares**

En la experiencia latinoamericana, el tema migratorio ha estado presente en forma permanente en los censos y en los registros administrativos que controlan el ingreso y salida de las personas en los puntos de frontera. Adicionalmente, algunas de las encuestas de hogares de la región han incluido una o varias preguntas tendientes a estimar este fenómeno, lo que podría permitir analizar las condiciones de vida y participación en el mercado laboral de las personas migrantes y compararlas con las condiciones equivalentes de las personas nativas del país donde se levanta la encuesta. Además, dada su mayor frecuencia, las encuestas de hogares podrían permitir dar seguimiento a la evolución temporal de la migración, sin necesidad de esperar a que se complete el período intercensal.

Sin embargo, la captación del fenómeno migratorio mediante estos instrumentos tiene sus dificultades. Una de ellas es la posible insuficiencia del tamaño de muestra necesario para estimar en forma precisa un fenómeno que estadísticamente puede ser de escasa relevancia. Esta dificultad se extiende además si, para propósitos analíticos y de mayor precisión en la definición, se pretende distinguir grupos al interior de los inmigrantes, basado en su tiempo de permanencia en el país, su nacionalidad, etc. En esta misma línea, es posible que, de haberse producido una inmigración masiva en el período intercensal, el marco de muestreo de la encuesta no lo recoja adecuadamente, lo que podría llevar a subestimar el número de los inmigrantes, así como otras variables asociadas al perfil sociodemográfico de estas personas. Del mismo modo y dado que muchas veces los inmigrantes se establecen inicialmente en hogares colectivos o pueden habitar incluso en los mismos lugares de trabajo, es posible que sus hogares se encuentren fuera del universo cubierto habitualmente por las encuestas de hogares y no sean por ende parte de sus estimaciones. Finalmente, es posible que se obtenga una subestimación del fenómeno debido al ocultamiento de la condición de migrantes por parte de los entrevistados, sea por desconfianza, desconocimiento o por encontrarse en forma irregular en el país de destino.

De los 18 países cuyas encuestas de hogares están disponibles en BADEHOG, 16 cuentan con preguntas que pueden ser utilizadas para identificar a la población inmigrante e inmigrante reciente (véase el cuadro 3).

**Cuadro 3**  
**América Latina (18 países): disponibilidad de información para identificar inmigrantes y estimar precisión de los estimadores, alrededor de 2017**

País	Nombre de la encuesta	Año encuesta	Migrantes	Migrantes recientes	UPM y estrato
Argentina	Encuesta Permanente de Hogares - EPH	2017	Sí	Sí	No
Bolivia	Encuesta Continua de Hogares - ECH	2017	Sí	Sí	Sí
Brasil	Pesquisa por Amostra de Domicílios - PNAD	2015	Sí	Sí	Sí
Chile	Encuesta de Caracterización Socioeconómica Nacional - CASEN	2017	Sí	Sí	Sí
Colombia	Gran Encuesta Integrada de Hogares	2017	Sí	Sí	No
Costa Rica	Encuesta Nacional de Hogares - ENAHO	2017	Sí	Sí	Sí
Ecuador	Encuesta de Empleo, Subempleo y Desempleo en el Área Urbana y Rural	2017	Sí	No	No
El Salvador	Encuesta de Hogares de Propósitos Múltiples	2017	No	No	No
Guatemala	Encuesta Nacional de Condiciones de Vida - ENCOVI	2014	Sí	Sí	Sí
Honduras	Encuesta Permanente de Hogares de Propósitos Múltiples	2016	Sí	Sí	No
México	Módulo de Condiciones Socioeconómicas de la ENIGH	2016	Sí	Sí	Sí
Nicaragua	Encuesta Nacional de Hogares sobre Medición de Niveles de Vida	2014	Sí	Sí	No
Panamá	Encuesta de Propósitos Múltiples	2017	Sí	Sí	Sí
Paraguay	Encuesta Permanente de Hogares	2017	Sí	Sí	Sí
Perú	Encuesta Nacional de Hogares – Condiciones de Vida y Pobreza	2017	Sí	Sí	Sí
República Dominicana	Encuesta Nacional Continua de Fuerza de Trabajo - ENCFT	2017	Sí	Sí	Sí
Uruguay	Encuesta Continua de Hogares	2017	Sí	Sí	No
Venezuela	Encuesta de Hogares por Muestreo	2014	No	No	No
Disponibles			16	15	10
No disponibles			2	3	8

Fuente: elaboración propia a partir de BADEHOG, Banco de Encuestas de Hogares de CEPAL.

En la mayoría de países la información se ajusta a las recomendaciones internacionales, mientras que en otros es posible realizar una identificación aproximada del carácter de migrante internacional. En 14 países se dispone de información sobre el lugar de nacimiento de los miembros del hogar y en varios de ellos se cuenta con información complementaria sobre el lugar de residencia anterior a la del momento de la encuesta. A su vez, en 15 países se pregunta sobre el lugar de residencia 5 años antes. Esta variable permite estimar movimientos de la población independientemente de su nacionalidad. Combinada con la pregunta sobre el lugar de residencia al momento de nacer, permite establecer qué parte de la población nativa del país fue parte de procesos migratorios (acotado a la referencia

temporal), al tiempo que permite estimar, con limitaciones, si los migrantes internacionales tuvieron algún país de destino intermedio en el período de referencia mencionado (5 años para esta pregunta y la "anterior" sin referencia temporal explícita en forma adicional en la mayoría de las encuestas). Esta información se encuentra resumida en el cuadro 4.

**Cuadro 4**  
**América Latina (16 países): variables disponibles para identificar inmigrantes, encuesta alrededor de 2017**

Pregunta	País															
	ARG	BOL	BRA	CHL	COL	CRI	DOM	ECU	GTM	HND	MEX	NIC	PAN	PER	PRY	URY
Lugar de Nacimiento	*		*		*		*	*	*	*				*		
Nacionalidad/País de origen				*									*			
Lugar de residencia de la madre al momento de nacer el entrevistado				*		*						*			*	
Lugar de residencia inmediato al nacimiento																*
<b>Resumen: Lugar de Nacimiento</b>	*		*	*	*	*	*	*	*	*		*	*	*	*	*
Lugar de residencia anterior a este			*		*			*		*						*
Motivo del cambio					*											
<b>Resumen: Lugar de residencia anterior al actual</b>			*		*			*		*						*
Lugar de residencia 5 años atrás	*	*	*	*	*	*	*		*	*	*	*	*	*	*	*
<b>Resumen: Lugar de residencia 5 años atrás</b>	*	*	*	*	*	*		*		*	*		*	*	*	*
Año en el que llegó a residir en este lugar (País)												*	*			
Tiempo de residencia en este lugar			*			*				*						*
<b>Resumen: Tiempo de residencia</b>			*			*				*		*	*			*
¿A qué nación o pueblo pertenece?		*														

Fuente: elaboración propia a partir de BADEHOG, Banco de Encuestas de Hogares de CEPAL.

En el caso de México, la aproximación a los migrantes internacionales se logra a través del conjunto de personas que residían fuera de ese país 5 años antes del período de la encuesta. Sin embargo, la mayoría de las respuestas obtenidas (alrededor del 8% de la muestra) identifican a ciudadanos de origen mexicano que retornaron de los Estados Unidos. Por ello, se filtra este origen y se establece que los migrantes internacionales son aquellos quienes 5 años antes residían en un país distinto de México y de los Estados Unidos, constituyendo apenas una aproximación. En el caso de Bolivia, la única aproximación al fenómeno se da en forma indirecta, a partir de la pregunta sobre

identificación étnica. Esta permite distinguir entre sus múltiples categorías a aquellos que no pertenecen a ninguno de los grupos considerados porque han nacido fuera de Bolivia. Esto no excluye, sin embargo, que los entrevistados se hayan auto clasificado en alguna de las otras categorías, por lo que la aproximación al fenómeno dista de ser satisfactoria. En los casos de El Salvador y Venezuela, las encuestas de hogares disponibles no permiten identificar, ni siquiera de manera aproximada, a los inmigrantes internacionales.

### **C. Precisión estadística de la información sobre inmigración en las encuestas de hogares**

Con base en las encuestas de hogares descritas en el cuadro 3, es posible identificar a los migrantes internacionales de manera estandarizada, como aquellos entrevistados cuyo país de nacimiento o el país de residencia de la madre al momento de nacer o inmediato posterior era un país diferente de aquel en el que reside en la actualidad. En el caso de Bolivia se da un apartamiento del estándar, identificándose los migrantes internacionales como aquellos que responden “no bolivianos” en la pregunta de identificación étnica de la encuesta. En el caso de México, es posible asumir como migrantes internacionales a quienes residían hace 5 años fuera de México y de Estados Unidos. Adicionalmente, en aquellos países con información disponible, es posible identificar a los migrantes recientes de manera estandarizada como aquellos que llegaron al país de destino en el transcurso de los últimos 5 años anteriores a la encuesta.

Utilizando las definiciones descritas, la información de las encuestas de hogares en torno al año 2017 permite agrupar a los países según el porcentaje de población inmigrante en 3 grupos. En 6 países (Bolivia, Brasil, Honduras, México, Nicaragua y Perú) la población inmigrante es menor al 1%; en 4 países (Colombia, Ecuador, Paraguay y Uruguay) la población inmigrante representa entre 1% y 3%. Por último, hay 6 países (Argentina, Chile, Costa Rica, Guatemala, Panamá y República Dominicana) donde la migración es mayor al 3%. Respecto al porcentaje de inmigrantes recientes, al año 2017, sólo en los casos de Chile y Panamá este superó el 2% del total de la población, en Costa Rica, Guatemala y Uruguay se situó en torno al 0,5% mientras que en el resto de los países el porcentaje estimado fue inferior a este último valor (véase el cuadro 5).

Para estimar la confiabilidad de los indicadores sobre migración, en 9 países se cuenta con información suficiente para la estimación insesgada de las varianzas de los estimadores (es decir, con la identificación del estrato y la unidad primaria de muestreo). En los demás países se estimó la varianza mediante la aproximación descrita en el capítulo III de este documento, que permite evaluar la mayoría de criterios de calidad, con la excepción de los grados de libertad. Para ello, se utilizó como referencia la mediana de los efectos de diseño observados en los países, para cada uno de los indicadores y desagregaciones correspondientes. En concordancia con los criterios de revisión y supresión planteados anteriormente, se considera que una estimación no es precisa cuando el coeficiente de variación es mayor a 20%, el coeficiente de variación logarítmico es mayor a 17,5%, el tamaño de muestra es menor a 100 unidades, el tamaño de muestra efectivo es menor a 68 unidades, los grados de libertad son menores a 10 o el conteo de casos no ponderado es menor a 50 unidades.

De los países analizados, en Bolivia, Honduras y Nicaragua la estimación del porcentaje de inmigrantes en el total de la población no alcanzó el requerimiento mínimo de precisión exigido. En el caso de la inmigración reciente, además de estos países, Guatemala y Perú tampoco alcanzan el umbral mínimo de precisión establecido (véase el cuadro 5).

**Cuadro 5**  
**América Latina (16 países): porcentaje de personas y límites inferior y superior del intervalo de confianza para migrantes y migrantes recientes, alrededor de 2017**  
*(En porcentajes)*

País	Inmigrantes	Límite inferior	Límite Superior	Estimación no precisa	Inmigrantes recientes	Límite inferior	Límite Superior	Estimación no precisa
Argentina	4,7	4,5	5,0		0,4	0,3	0,5	
Bolivia (Estado Plurinacional de)	0,3	0,2	0,4	*	0,1	0,0	0,2	*
Brasil	0,4	0,3	0,4		0,0	0,0	0,0	
Chile	3,9	3,4	4,5		2,6	2,1	3,2	
Colombia	1,1	1,1	1,1		0,7	0,7	0,7	
Costa Rica	8,7	8,0	9,4		0,5	0,4	0,7	
Ecuador	1,2	1,1	1,3		0,0	0,0	0,0	
Guatemala	3,5	2,8	4,4		0,5	0,4	0,8	*
Honduras	0,7	0,6	0,8	*	0,1	0,0	0,2	*
México	...	...	...		0,1	0,1	0,1	
Nicaragua	0,8	0,7	0,9	*	0,2	0,1	0,3	*
Panamá	4,6	4,3	4,8		2,1	1,9	2,4	
Paraguay	2,6	2,2	2,9		0,3	0,2	0,4	
Perú	0,3	0,3	0,4		0,1	0,0	0,1	*
República Dominicana	3,6	3,1	4,2		1,0	0,7	1,2	
Uruguay	2,3	2,2	2,4		0,5	0,4	0,5	

Fuente: Elaboración propia a partir de BADEHOG, Banco de Encuestas de Hogares de CEPAL.

En forma complementaria, se desagregó un conjunto de indicadores habituales de caracterización sociodemográfica, con el objetivo de identificar en qué medida los diseños muestrales de las encuestas permiten generar información según el estatus migratorio de las personas. El cuadro 6 presenta los resultados para un total de 13 indicadores en 12 países que cuentan con la información del estatus migratorio y cuya proporción puede ser estimada con precisión, según el cuadro 5. La información del cuadro 7 corresponde a 10 países en los cuales es posible identificar la migración reciente y estimar el porcentaje de población en esta categoría de manera fiable (nuevamente, según el cuadro 5). En ambos cuadros se representa mediante un asterisco los indicadores para los cuales la desagregación según estatus migratorio no alcanza el nivel de precisión requerido.

De acuerdo con el cuadro 6, solo en 4 países (Argentina, Chile, Colombia y Costa Rica) es posible desagregar todos los indicadores analizados para la población inmigrante. En otros países, dos o más de los indicadores elegidos no pueden ser desagregados para este grupo. De los indicadores analizados, las tasas de desempleo y de pobreza extrema, así como la proporción de población que reside en áreas urbanas, son las que menores posibilidades presentan de ser estimadas para la población inmigrante.

Las limitaciones para la desagregación de datos son particularmente evidentes en el caso de la población migrante reciente. Aun restringiendo el análisis a los países que satisfacen los criterios de calidad para la proporción de población inmigrante reciente a nivel nacional, en varios casos no es posible estimar más que unos pocos indicadores. Solo las encuestas de Argentina, Chile y Colombia, que son algunas de las que cuentan con mayores tamaños de muestra en la región, ofrecen la posibilidad de estimar la mayoría de indicadores para este grupo.

**Cuadro 6**  
**América Latina (12 países): indicadores que no alcanzan los requerimientos mínimos de precisión exigidos por país para la población migrante, alrededor de 2017**

	ARG	BRA	CHL	COL	CRI	ECU	GTM	PAN	PRY	PER	DOM	URY	Est. no precisas
Residentes en áreas urbanas							*						3/12
Residentes en áreas rurales	...	*				*	*	*		*		*	8/12
Personas de 0 a 14 años													1/12
Personas de 15 a 64 años													0/12
Personas de 65 años y más						*				*	*		5/12
Hombres													0/12
Mujeres													0/12
Personas con educación primaria completa													1/12
Personas con educación secundaria													2/12
Personas con educación terciaria							*				*		4/12
Personas en situación de pobreza								*		*		*	4/12
Personas en situación de pobreza extrema		*				*	*	*	*	*		*	9/12
Personas desempleadas		*				*	*	*	*	*	*		9/12
<b>Estimaciones no precisas</b>	0/13	3/13	0/13	0/13	0/13	4/13	5/13	4/13	2/13	5/13	3/13	3/13	

Fuente: Elaboración propia a partir de BADEHOG, Banco de Encuestas de Hogares de CEPAL.



**Cuadro 7**  
**América Latina (10 países): indicadores que no alcanzan los requerimientos mínimos de precisión exigidos por país, para la población migrante reciente, alrededor de 2017**

	ARG	BRA	CHL	COL	CRI	MEX	PAN	PRY	DOM	URY	Est. no precisas
Residentes en áreas urbanas		*			*			*			3/10
Residentes en áreas rurales	...	*			*	*	*	*	*	*	7/10
Personas de 0 a 14 años	*	*			*	*	*	*			6/10
Personas de 15 a 64 años	*							*			2/10
Personas de 65 años y más	*	*	*	*	*	*		*	*	*	9/10
Hombres	*										1/10
Mujeres	*	*									2/10
Personas con educación primaria completa	*	*						*			3/10
Personas con educación secundaria	*	*			*	*		*	*		6/10
Personas con educación terciaria	*	*			*	*		*	*	*	7/10
Personas en situación de pobreza	*	*			*	*	*	*		*	7/10
Personas en situación de pobreza extrema	*	*	*		*	*	*	*	*	*	6/10
Personas desempleadas	*	*			*	*	*	*	*	*	7/10
<b>Estimaciones no precisas</b>	11/13	11/13	2/13	1/13	9/13	8/13	5/13	11/13	6/13	5/13	

Fuente: Elaboración propia a partir de BADEHOG, Banco de Encuestas de Hogares de CEPAL.

De manera complementaria, en los cuadros A1 y A2 del anexo del documento se presenta en forma detallada para cada país la estimación de la tasa de pobreza y de la tasa de desocupación, respectivamente, según condición migratoria. Además de la estimación puntual, los cuadros proveen información sobre diversos indicadores de precisión estadística, incluyendo los límites inferior y superior del intervalo de confianza, el coeficiente de variación, el coeficiente de variación logarítmico, el efecto de diseño, el tamaño de la muestra, el tamaño efectivo de la muestra, los grados de libertad, el número de casos efectivos y, por último, la evaluación de si la estimación alcanza o no los requerimientos mínimos de precisión. De un total de 15 países, en 10 la tasa de pobreza entre los migrantes alcanza el requerimiento de precisión establecido. Al caso ya señalado de Bolivia, cuya estimación general sobre migrantes ya no cumplía con este requisito, se agregan los casos de Honduras, Panamá, Perú y Uruguay. En estos casos los coeficientes de variación estimados son superiores a 0,20 o el número de casos efectivos es insuficiente, por lo que sus cifras no debieran ser utilizadas en el análisis conjunto de pobreza e inmigración. Por su parte, el cuadro A2 muestra los resultados referidos a la tasa de desocupación por país. De un total de 15 países, solamente en 7 la tasa de desocupación entre los migrantes alcanza el requerimiento de precisión establecido.

Con base en los resultados descritos, el análisis de la inmigración a nivel regional mediante las encuestas de hogares de los países de la región está sujeto a las siguientes restricciones:

- De los 18 países latinoamericanos cuyas encuestas están disponibles en BADEHOG, es posible estimar la proporción y nivel de los inmigrantes en 12 países. Se excluye a Bolivia, Honduras y Nicaragua por la imprecisión del estimador, a México por problemas de comparabilidad de la definición, y a El Salvador y Venezuela, por falta de la información necesaria en las encuestas.
- Es posible estimar algunos indicadores para la población inmigrante en 8 países, que corresponden a aquellos donde la proporción de inmigrantes supera el 3% (Argentina, Chile, Costa Rica, Guatemala, Panamá y República Dominicana) y a Colombia y Brasil (hasta 2015), que por el tamaño de muestra permiten algunas desagregaciones.
- De acuerdo a los resultados reseñados en el cuadro 6, en pocos casos es posible desagregar la información para el área rural o presentar la información sobre pobreza extrema y desocupación. Debe trabajarse también con cautela al desagregar variables representativas del nivel educativo y del grupo de edad.
- Dados los problemas de precisión ya reseñados, no es pertinente desagregar la información de la subpoblación migrante por más de una variable a la vez.

Además, no resulta recomendable utilizar la información sobre migración reciente excepto en los casos de Chile y Colombia, tomando los mismos recaudos que se han señalado en los puntos anteriores.



## V. Conclusiones

Las estimaciones elaboradas a partir de encuestas de hogares están sujetas al error muestral y, por tanto, se requiere evaluar su validez estadística mediante diversos indicadores de calidad. En este documento se plantea un conjunto de criterios a considerar para poder obtener estimaciones válidas, que permitan su uso por parte de investigadores, hacedores de política y otros usuarios interesados.

Una vez obtenido el indicador de interés (por ejemplo, la proporción de personas en situación de pobreza y de pobreza extrema), se estiman los intervalos de confianza y otros indicadores de calidad con base en la información sobre el diseño muestral complejo, resumida en el factor de expansión, los estratos y las unidades primarias de muestreo (UPM). Cuando las encuestas de hogares sólo proveen los pesos de muestreo se utiliza el ajuste de West y McCabe (2012). Los criterios de calidad estimados para cada cifra están supeditados al intervalo de confianza, que describe un conjunto de valores en donde es posible que el indicador de interés se encuentre. Este intervalo está determinado por el estimador de muestreo para el indicador, su error estándar y el percentil 0,975 de una distribución *t-student* con grados de libertad igual a la diferencia entre estratos y UPM. Sin embargo, a partir de los elementos que conforman los intervalos de confianza, es posible establecer criterios de precisión como una manera más expedita de publicación, revisión o supresión de estimaciones. En resumen, estos criterios son los siguientes:

- Coeficiente de variación: este indicador modela el error de muestreo de un estimador. Se calcula como la razón entre el error estándar de la estimación y su estimación puntual. Esta medida es ampliamente utilizada para validar la precisión de una estimación, un coeficiente de variación elevado puede implicar que el error estándar de la estimación es relativamente grande.
- Coeficiente de variación logarítmico: como el coeficiente de variación no es simétrico alrededor de 0,5, esta medida se define como una transformación logarítmica sobre la proporción, la cual evita que las estimaciones cercanas a cero sean castigadas con un coeficiente de variación alto aun cuando su variación sea pequeña.

- Efecto de diseño: da cuenta de la correlación entre la variable de interés y la distribución de los hogares en las unidades primarias de muestreo. Esta medida depende del promedio de hogares seleccionados por UPM y del coeficiente de correlación intraclase entre la variable de interés y las UPM.
- Tamaño de muestra: el número de unidades que comprenden el diseño de muestreo afecta indirectamente la amplitud del intervalo de confianza mediante el error estándar, el cual depende de manera inversamente proporcional al tamaño de muestra. Con un mayor tamaño de muestra tendremos estimaciones más precisas y un intervalo de confianza más angosto.
- Tamaño de muestra efectivo: se calcula como la razón entre el tamaño de muestra y el efecto de diseño. El tamaño de muestra efectivo tiene por objeto deflactar el tamaño de muestra inicial por medio de la correlación intraclase de las UPM seleccionadas; de esta forma se evita contabilizar el exceso de información repetida debido a la aglomeración de los hogares en las UPM.
- Conteo de casos no ponderado: está determinado por el número de casos en la muestra afectados por el fenómeno de estudio, sin considerar el factor de expansión, ni el diseño complejo de la encuesta.
- Grados de libertad: se definen como la diferencia entre el número de UPM y el número de estratos. Estos definen el valor del percentil de la distribución a la hora de calcular el intervalo de confianza; mientras más grados de libertad, el intervalo de confianza será más estrecho y la estimación será más precisa.

La aplicación de estos criterios puede hacerse siguiendo una secuencia lógica, que permita determinar si una cifra en particular se considera suficientemente precisa. En este documento se presenta un ejercicio empírico en el que se considera que una estimación no es confiable si el coeficiente de variación es mayor a 20%, el coeficiente de variación logarítmico es mayor a 17,5%, el tamaño de muestra es menor a 100 unidades, el tamaño de muestra efectivo es menor a 68 unidades, los grados de libertad son menores a 10 o el conteo de casos no ponderado es menor a 50 unidades.

Para ejemplificar lo anterior, se calculan diferentes variables de interés para las personas en condición de migratoria internacional en 16 países de América Latina. Los resultados muestran que, siguiendo los criterios establecidos, la tasa de pobreza para la población inmigrante resulta confiable en 10 países, mientras que la tasa de desocupación para la población inmigrante es válida solo en 7 países. Por otra parte, en ningún país es posible calcular estos indicadores para la población inmigrante en áreas rurales, debido a que no satisfacen los criterios de supresión. Tampoco es recomendable desagregar la subpoblación de migrantes por más de una variable a la vez para obtener estimaciones válidas.

De esta manera, el documento ilustra que los indicadores calculados a partir de encuestas de hogares no siempre son representativos de la población bajo estudio y que, para obtener cifras válidas, es necesario evaluar la confiabilidad estadística considerando los elementos asociados al diseño muestral de la encuesta. La rigurosidad estadística es esencial para producir evidencias que permitan a las investigaciones y políticas públicas abordar de manera efectiva los problemas en cuestión.

## Bibliografía

- Barnett-Walker, Kortnee C., James R. Chromy, Teresa R. Davis, Steven L. Emrich, Dawn M. Odom, y Lisa E. Packer, (2003), «2001 National Household Survey on Drug Abuse».
- Comisión Económica para América Latina y el Caribe (CEPAL) (2008), Sistema de cuentas nacionales. United Nations Economic Commission for Latin America and the Caribbean.
- Gutiérrez, A. (2018), Limitaciones de las encuestas de hogares en la medición de indicadores sociales. Presentación en «Taller regional sobre desagregación de estadísticas sociales mediante metodologías de estimación en áreas pequeñas». (<https://www.cepal.org/sites/default/files/courses/files/limitaciones-encuestas-hogare-medicion-indicadores-sociales-cepal-andres-gutierrez.pdf>).
- \_\_\_\_\_(2016), Estrategias de muestreo: diseño de encuestas y estimación de parámetros. Segunda edición. Ediciones de la U.
- Hansen, M. H., Hurwitz, W., Madow, M. (1953), Sample Survey Methods and Theory. Wiley.
- Heeringa, Steven G., Brady T. West, y Patricia A. Berglund. 2010, Applied survey data analysis. Chapman & Hall/CRC statistics in the social and behavioral sciences series. CRC Press.
- Hornik, R., Maklan, D., Cadell, Prado, A., Barmada, C., Jacobsohn, L., Orwin, R., et al. (2002), Evaluation of the National Youth Anti-Drug Media Campaign: Fourth Semi-Annual Report of Findings, 740.
- IBM (2017), IBM SPSS Complex Samples. URL [ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/23.0/en/client/Manuals/IBM\\_SPSS\\_Complex\\_Samples.pdf](ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/23.0/en/client/Manuals/IBM_SPSS_Complex_Samples.pdf).
- IMF (2009), Balance of payments and international investment position manual (6th ed. (BPM6)). Washington, DC: Internat. Monetary Fund.
- Kish, L. (1965), Survey Sampling. John Wiley; Sons.
- Korn, E.L., Graubard, B (1999), Analysis of health surveys. Wiley.
- Lumley, T. (2016), Survey: analysis of complex survey simples.
- \_\_\_\_\_(2010), Complex surveys: a guide to analysis using R. Wiley series in survey methodology. Wiley.
- National Research Council (2015), Realizing the Potential of the American Community Survey: Challenges, Tradeoffs, and Opportunities. National Academies Press. <https://doi.org/10.17226/21653>.
- Park, I. (2003), Design Effects and Survey Planning, 8.
- Parker, JD, Talih, M., Malec, D.J. (2017), National Center for Health Statistics Data Presentation Standards for Proportions». Vital Health Stat 2 (175).

- R Core Team (2019), R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Särndal, C.E., Swensson B., Wretman, J. (2003), Model Assisted Survey Sampling. Springer Science & Business Media.
- SAS. (2010), SAS/STAT 9.22 User's Guide-Survey Sampling and Analysis Procedures. <https://support.sas.com/documentation/cdl/en/statugsurveysamp/63778/PDF/default/statugsurveysamp.pdf>.
- Singh, A. C., M. Westlake, M. Feder. (2004), A generalization of the Coefficient of variation with application to suppression of imprecise estimates. En.
- STATA (2013), STATA Survey Data. URL. <https://www.stata.com/manuals13/svy.pdf>.
- UN (2017), Handbook on Measuring International Migration through Population Censuses. Recuperado de <https://unstats.un.org/unsd/statcom/48th-session/documents/BG-4a-Migration-Handbook-E.pdf>.
- \_\_\_\_\_ (2005), Household surveys in developing and transition countries. Studies in methods / United Nations, Department of Economic and Social Affairs, Statistics Division Series F.
- \_\_\_\_\_ (1998), Recommendations on Statistics of International Migration. [https://unstats.un.org/unsd/publication/SeriesM/SeriesM\\_58rev1E.pdf](https://unstats.un.org/unsd/publication/SeriesM/SeriesM_58rev1E.pdf).
- West, B. T. (2012), Accounting for Multi-stage Sample Designs in Complex Sample Variance Estimation.
- West, B T., McCabe S.E. (2012), Incorporating complex sample design effects when only final survey weights are available. *The Stata Journal* 12 (4): 718-25.
- Westat (2007), WesVar 4.3. Users guide.
- Wolter, K. (2007), Introduction to Variance Estimation. Springer Science & Business Media.

## Anexos



## Anexo 1

### Software: procesamiento de las cifras

Esta sección presenta los procedimientos necesarios para calcular los criterios de calidad de las estimaciones que hacen uso de las bases de datos de encuestas de hogares BADEHOG, utilizando el software computacional existente y revisa con detalle las aproximaciones computacionales utilizadas para realizar los procesos estadísticos que se necesitan para lograr el cometido de la publicación de cifras con altos niveles de precisión y confiabilidad. Además, se muestran algunas de las características más importantes que los paquetes estadísticos computacionales incorporan en el manejo de datos que provienen de estructuras de muestreo complejas, como las encontradas en las encuestas de hogares. Una revisión más exhaustiva y detallada que adjunta sintaxis y código computacional puede encontrarse en Heeringa, West, y Berglund (2010 Apéndice A). En general, estas herramientas computacionales están pensadas para hacer más eficiente el uso de las aproximaciones de varianza en muestras complejas, así como las técnicas de replicación para obtener los estimativos de varianza inducidos por el diseño de muestreo (Westat, 2007). Algunos de estos softwares son de uso libre, aunque la mayoría corresponde a productos comerciales. En general estos softwares, además de proveer estadísticas descriptivas (como medias, totales, proporciones, percentiles y razones), permiten ajustar modelos de regresión lineales y logísticos. Todas las estadísticas resultantes están basadas en el diseño de muestreo de la encuesta.

- **R:** es un software de uso libre cuyo uso es cada vez más frecuente en la investigación social, puesto que es muy probable encontrar los más recientes hallazgos científicos programados en este software (R Core Team, 2019). Al ser de uso libre, los investigadores pueden subir sus propias colecciones de funciones computacionales al repositorio oficial ('CRAN') y ponerlas a disposición de la comunidad. El paquete 'samplesize4surveys' permite determinar el tamaño de muestra de individuos y hogares en encuestas de hogares repetidas, tipo panel y con rotación. Los paquetes 'sampling' y 'TeachingSampling' permiten seleccionar muestras probabilísticas desde los marcos de muestreo bajo un gran variedad de diseños y algoritmos de muestreo. El paquete 'survey', una vez que el diseño de muestreo ha sido predefinido mediante la función 'svydesign()', permite analizar datos provenientes de encuestas de hogares y obtener estimaciones apropiadas de los errores estándar.
- **STATA:** El entorno 'svy' provee un conjunto de herramientas para hacer una inferencia apropiada de las estadísticas oficiales provenientes de encuestas de hogares (STATA, 2013). El comando 'svyset' permite especificar las variables que identifican las características del diseño de muestreo de la encuesta, como los pesos de muestreo, los conglomerados y los estratos. El comando 'svydescribe' proporciona tablas que describen los estratos y las unidades de muestra para una determinada etapa de la encuesta. Una vez cargadas las definiciones del diseño de muestreo, cualquier modelo puede ser estimado y sus estadísticos resultantes estarán basados en el diseño de muestreo de la encuesta. El entorno 'svy' también permite la ejecución de algunos comandos predictivos.
- **SAS:** este software estadístico incluye un procedimiento para la selección de muestras probabilísticas llamado 'SURVEYSELECT' que integra los métodos de selección más comunes como muestreo aleatorio simple, muestreo sistemático, muestreo con probabilidad proporcional al tamaño, y algunas herramientas de afijación en los estratos. Para analizar los datos provenientes de muestras complejas se han programado algunos procedimientos (SAS, 2010). 'SURVEYMEANS', que estima totales, medias, proporciones, y percentiles, junto con sus respectivos errores estándar, límites de los intervalos de confianza y pruebas de hipótesis. 'SURVEYFREQ', estima las estadísticas descriptivas (como totales y proporciones) de interés

en tablas de una y dos vías, brinda las estimaciones del error de muestreo, y realiza un análisis de la bondad del ajuste de las estimaciones, independencia, riesgos y razones de odds. `SURVEYREG` y `SURVEYLOGISTIC` ajustan modelos de regresión lineal y logísticas, respectivamente. Estos procedimientos estiman los coeficientes de regresión, con sus respectivos errores, y adjunta un análisis exhaustivo de las propiedades de los modelos. Por último, `SURVEYPHREG` ajusta modelos de riesgos utilizando técnicas de máxima-pseudo verosimilitud.

- **SPSS:** el módulo `complex samples` de `SPSS` (IBM, 2017) incorpora la selección de muestras complejas mediante la definición de un esquema de muestreo establecido por el usuario. Luego, es necesario crear un plan de análisis mediante la asignación de variables de diseño, métodos de estimación y tamaños de las unidades de muestreo. Una vez definido el plan de muestreo, el módulo integra la posibilidad de estimar conteos, estadísticas descriptivas y celdas de tablas cruzadas. También es posible realizar estimaciones de razones y de coeficientes de regresión en modelos lineales, junto con las respectivas estadísticas de pruebas de hipótesis. Por último el módulo permite estimar modelos no lineales, como regresiones logísticas, regresiones ordinales o regresiones de Cox.

## Anexo 2

### Resultados de la estimación de algunos indicadores sociales por estado

**Cuadro A1**  
**América Latina (16 países): proporción de personas pobres e indicadores de precisión estadística, según condición de inmigración, alrededor de 2017**

Subgrupo	Pers. en situación pobreza (en porcentajes)	Límite inferior	Límite superior	Coef. de variación	Coef. de variación logarítmico	Efecto de diseño	Tamaño de muestra	Tamaño de muestra efectivo	Grados de libertad	Número de casos	Precisión
Argentina (2017)											
Migrante	19,9	17,2	22,6	6,7	4,2	3,2	1 902	564	-	284	
No migrante	18,7	17,6	19,8	6,8	1,8	11,7	56 279	24 347	-	10 645	
Total	18,7	17,6	19,8	6,7	1,8	11,9	58 181	24 599	-	10 929	
Bolivia (2017)											
Migrante	14,5	3,1	25,9	40,0	20,8	2,8	127	45	43	23	*
No migrante	35,2	33,4	37,0	2,6	2,5	14,5	38 074	2 628	929	12 698	
Total	35,1	33,3	37,0	2,6	2,5	14,5	38 201	2 641	929	12 721	
Brasil (2015)											
Migrante	7,5	5,3	9,8	15,0	5,8	2,5	1 199	484	452	101	
No migrante	18,8	18,4	19,2	1,1	0,6	9,1	355 705	38 923	5 616	70 004	
Total	18,8	18,4	19,2	1,1	0,6	9,1	356 904	39 019	5 616	70 105	
Chile (2017)											
Migrante	13,5	10,8	16,1	10,2	5,1	13,6	5 783	426	644	947	
No migrante	10,6	10,1	11,1	2,2	1,0	12,1	210 656	17 469	1 295	23 270	
Total	10,7	10,3	11,2	2,2	1,0	12,4	216 439	17 460	1 295	24 217	
Colombia (2017)											
Migrante	40,2	38,4	42,1	2,7	2,6	3,2	9 008	3 605	-	3 813	
No migrante	29,7	29,3	30,0	1,2	0,5	11,7	758 859	252 826	-	209 646	
Total	29,8	29,4	30,1	1,2	0,5	11,9	767 867	256 483	-	213 459	
Costa Rica (2017)											
Migrante	20,2	17,5	22,9	6,9	4,3	3,6	2 959	820	735	639	
No migrante	14,9	13,9	15,9	3,5	1,8	6,6	31 884	4 796	1 118	5 010	
Total	15,3	14,3	16,4	3,4	1,8	7,5	34 843	4 661	1 119	5 649	
Ecuador (2017)											
Migrante	20,4	16,5	24,4	13,3	6,2	3,2	1 655	960	-	412	
No migrante	23,7	22,8	24,5	4,1	1,3	11,7	108 628	46 033	-	28 459	
Total	23,6	22,8	24,5	4,2	1,3	11,9	110 283	46 864	-	28 871	

Subgrupo	Pers. en situación de pobreza (en porcentajes)	Límite inferior	Límite superior	Coef. de variación	Coef. de variación logarítmico	Efecto de diseño	Tamaño de muestra	Tamaño de muestra efectivo	Grados de libertad	Número de casos	Precisión
Guatemala (2014)											
Migrante	21,0	15,2	26,7	14,0	9,0	10,1	1 237	122	379	305	
No migrante	51,6	48,9	54,2	2,6	3,8	38,1	53 585	1 405	1 036	28 647	
Total	50,5	47,8	53,2	2,7	3,9	40,6	54 822	1 350	1 036	28 952	
Honduras (2016)											
Migrante	49,0	36,4	61,6	21,5	18,4	3,2	185	155	-	94	*
No migrante	53,1	51,1	55,2	6,3	2,9	11,7	27 112	23 947	-	14 632	
Total	53,1	51,1	55,1	6,4	2,9	11,9	27 297	24 100	-	14 726	
Nicaragua (2014)											
Migrante	39,7	28,7	50,8	18,3	15,4	3,2	247	129	-	102	
No migrante	46,3	44,4	48,3	5,0	2,8	11,7	29 134	13 205	-	11 665	
Total	46,3	44,3	48,3	5,0	2,8	11,9	29 381	13 324	-	11 767	
Panamá (2017)											
Migrante	6,4	4,5	8,4	22,4	5,6	3,2	1 436	964	-	124	*
No migrante	16,2	15,0	17,4	12,0	2,1	11,7	42 313	36 104	-	8 041	
Total	15,8	14,6	16,9	12,1	2,1	11,9	43 749	37 050	-	8 165	
Paraguay (2017)											
Migrante	14,3	10,6	18,0	13,3	6,8	2,6	1 071	406	452	142	
No migrante	21,7	20,0	23,4	4,0	2,6	15,1	34 144	2 266	949	8 286	
Total	21,5	19,8	23,2	4,0	2,6	15,3	35 215	2 295	950	8 428	
Perú (2017)											
Migrante	6,6	1,6	11,7	38,8	14,3	4,5	355	79	231	27	*
No migrante	18,9	18,2	19,6	2,0	1,2	11,3	124 505	11 004	5 350	27 916	
Total	18,9	18,1	19,6	2,0	1,2	11,3	124 860	11 025	5 351	27 943	
República Dominicana (2017)											
Migrante	35,1	29,5	40,7	8,1	7,8	2,7	817	305	239	311	
No migrante	24,7	23,2	26,1	3,0	2,1	5,9	20 060	3 414	1 028	5 427	
Total	25,0	23,6	26,5	2,9	2,1	5,9	20 877	3 512	1 029	5 738	
Uruguay (2017)											
Migrante	1,8	0,9	2,7	43,7	6,3	3,2	2 797	2 608	-	47	*
No migrante	2,7	2,4	3,0	19,8	1,7	11,7	115 471	105 054	-	2 881	
Total	2,7	2,3	3,0	20,0	1,7	11,9	118 268	107 640	-	2 928	

Fuente: Elaboración propia a partir de BADEHOG, Banco de Encuestas de Hogares de CEPAL.

**Cuadro A2**  
**América Latina (15 países): tasa de desocupación e indicadores de calidad estadística, según condición de inmigración, alrededor de 2017**

Subgrupo	Personas en situación de desocupación (en porcentajes)	Límite inferior	Límite superior	Coef. de variación	Coef. de variación logarítmico	Efecto de diseño	Tamaño de muestra	Tamaño de muestra efectivo	Grados de libertad	Número de casos	Precisión
Argentina (2017)											
Migrante	8,2	6,6	9,8	6,8	4,0	1,4	962	316	-	58	
No migrante	7,2	6,7	7,6	3,3	1,0	2,2	25 009	10 645	-	1 530	
Total	7,2	6,7	7,7	3,6	1,3	2,5	25 971	10 801	-	1 588	
Bolivia (2017)											
Migrante	0,5	-0,5	1,5	100,6	19,1	0,2	62	252	43	1	*
No migrante	3,3	3,0	3,7	5,4	1,6	1,8	18 202	9 962	929	724	
Total	3,3	3,0	3,7	5,4	1,6	1,8	18 264	10 003	929	725	
Brasil (2015)											
Migrante	5,5	3,5	7,5	18,4	6,4	1,4	619	457	452	37	*
No migrante	9,6	9,4	9,8	1,1	0,5	2,5	182 602	74 283	5 616	17 859	
Total	9,6	9,4	9,8	1,1	0,5	2,5	183 221	74 192	5 616	17 896	
Chile (2017)											
Migrante	7,9	5,5	10,4	15,6	6,1	12,8	3 926	307	644	305	
No migrante	8,6	8,4	8,9	1,5	0,6	2,0	97 162	48 223	1 295	8 366	
Total	8,6	8,3	8,9	1,7	0,7	2,7	101 088	36 994	1 295	8 671	
Colombia (2017)											
Migrante	12,6	11,4	13,9	3,9	2,4	1,4	4 146	1 835	-	558	
No migrante	9,3	9,2	9,5	0,7	0,3	2,2	389 407	142 646	-	40 953	
Total	9,4	9,2	9,5	0,8	0,3	2,5	393 553	144 551	-	41 511	
Costa Rica (2017)											
Migrante	7,4	6,0	8,7	9,2	3,5	1,3	1 862	1 418	735	143	
No migrante	7,5	7,0	8,0	3,4	1,3	1,3	14 000	10 795	1 118	1 036	
Total	7,5	7,0	7,9	3,2	1,2	1,3	15 862	11 940	1 119	1 179	
Ecuador (2017)											
Migrante	3,7	2,0	5,3	23,3	7,1	1,4	873	613	-	35	*
No migrante	3,9	3,7	4,2	2,8	1,0	2,2	54 494	19 158	-	1 833	
Total	3,9	3,7	4,2	3,1	1,0	2,5	55 367	19 618	-	1 868	
Guatemala (2014)											
Migrante	4,1	2,2	6,0	24,1	7,5	2,5	642	253	379	24	*
No migrante	2,2	1,9	2,5	7,1	1,9	2,4	21 993	9 006	1 036	493	
Total	2,3	2,0	2,6	7,1	1,9	2,7	22 635	8 452	1 036	517	

Subgrupo	Personas en situación de desocupación (en porcentajes)	Límite inferior	Límite superior	Coefficiente de variación	Coefficiente de variación logarítmico	Efecto de diseño	Tamaño de muestra	Tamaño de muestra efectivo	Grados de libertad	Número de casos	Precisión
Honduras (2016)											
Migrante	1,2	-1,6	3,9	143,5	26,8	1,4	79	81	-	1	*
No migrante	4,5	4,0	5,1	8,9	2,0	2,2	11 889	10 848	-	546	
Total	4,5	3,9	5,1	10,1	2,1	2,5	11 968	10 919	-	547	
Nicaragua (2014)											
Migrante	4,3	-0,8	9,3	72,3	19,2	1,4	76	77	-	4	*
No migrante	5,3	4,8	5,9	6,4	1,8	2,2	13 665	8 808	-	865	
Total	5,3	4,7	5,9	7,2	1,9	2,5	13 741	8 881	-	869	
Panamá (2017)											
Migrante	5,7	4,2	7,2	11,1	4,7	1,4	927	450	-	47	*
No migrante	5,2	4,8	5,7	5,6	1,5	2,2	19 775	13 791	-	917	
Total	5,3	4,8	5,7	6,0	1,6	2,5	20 702	13 902	-	964	
Paraguay (2017)											
Migrante	2,3	0,9	3,7	31,4	8,3	1,0	558	541	452	15	*
No migrante	5,3	4,8	5,8	4,9	1,7	2,3	17 370	7 408	949	842	
Total	5,2	4,7	5,7	4,9	1,7	2,4	17 928	7 604	950	857	
Perú (2017)											
Migrante	17,3	6,5	28,0	31,8	18,1	3,3	124	37	231	15	*
No migrante	3,9	3,6	4,1	3,2	1,0	2,7	66 829	24 866	5 350	2 023	
Total	3,9	3,7	4,2	3,1	1,0	2,7	66 953	24 846	5 351	2 038	
República Dominicana (2017)											
Migrante	5,1	2,8	7,4	22,7	7,6	1,4	560	401	239	25	*
No migrante	6,3	5,7	6,9	4,8	1,7	1,4	9 047	6 304	1 028	559	
Total	6,3	5,7	6,9	4,7	1,7	1,4	9 607	6 659	1 029	584	
Uruguay (2017)											
Migrante	7,8	6,2	9,4	12,1	4,1	1,4	1 437	1,333	-	110	
No migrante	7,9	7,6	8,2	3,0	0,8	2,2	57 821	52 969	-	4 292	
Total	7,9	7,5	8,2	3,4	0,9	2,5	59 258	54 441	-	4 402	

Fuente: Elaboración propia a partir de BADEHOG, Banco de Encuestas de Hogares de CEPAL.



NACIONES UNIDAS

Serie

CEPAL

Estudios Estadísticos

## Números publicados

Un listado completo, así como los archivos pdf están disponibles en  
[www.cepal.org/publicaciones](http://www.cepal.org/publicaciones)

101. Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares: una aplicación a la migración internacional, Andrés Gutiérrez, Álvaro Fuentes, Xavier Mancero, Felipe López, Felipe Molina, (LC/TS.2020/52), 2020.
100. Desafíos en el diseño de medidas de pobreza multidimensional, María Emma Santos, (LC/TS.2019/5), 2019.
99. Non monetary indicators to monitor SDG targets 1.2 and 1.4: Standards, availability, comparability and quality, María Emma Santos, (LC/TS.2019/4), 2018.
98. Una propuesta de estimación del producto interno bruto trimestral de América Latina y el Caribe, Giannina López, Alberto Malmierca, (LC.TS.2018/88), 2018.
97. Desagregación de datos en encuestas de hogares. Metodologías de estimación en áreas pequeñas, Isabel Molina, (LC/TS.2018/82), 2018.
96. ¿Cuál es el alcance de las transferencias no contributivas en América Latina?: discrepancias entre encuestas y registros, Pablo Villatoro, Simone Cecchini, (LC/TS.2018/46), 2018.
95. Avances y desafíos de las cuentas económico-ambientales en América Latina y el Caribe, Franco Carvajal, (LC/TS.2017/148), 2018.
94. La situación de las estadísticas, indicadores y cuentas ambientales en América Latina y el Caribe, (LC/TS.2017/135), 2017.
93. Indicadores no monetarios de carencias en las encuestas de los países de América Latina: disponibilidad, comparabilidad y pertinencia, Pablo Villatoro, (LC/TS.2017/130), 2017.
92. Un índice de pobreza multidimensional para América Latina, María Emma Santos, Pablo Villatoro, Xavier Mancero Pascual Gerstenfeld, (LC/L.4129), 2015.

## ESTUDIOS ESTADÍSTICOS

### Números publicados:

- 101 Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares  
Una aplicación a la migración internacional  
*Andrés Gutiérrez, Álvaro Fuentes, Xavier Mancero, Felipe López, Felipe Molina*
- 100 Desafíos en el diseño de medidas de pobreza multidimensional,  
*María Emma Santos*
- 99 Indicadores no monetarios para el seguimiento de las metas 1.2 y 1.4 de los Objetivos de Desarrollo Sostenible  
Estándares, disponibilidad, comparabilidad y calidad  
*María Emma Santos*
- 98 Una propuesta de estimación del producto interno bruto trimestral de América Latina y el Caribe  
*Giannina López, Alberto Malmierca*

