Sociedad Ecuatoriana de Estadística

Análisis de Encuestas por muestreo con R

"Estimación en Áreas Pequeñas"

Andrés Peña Montalvo

andres.pena.montalvo@ciencias.unam.mx

ECUADOR 2025 CLATSE XVI Congreso Latinoamericano de Sociedades de Estadística

Julio 2025

¿Qué es el coeficiente de variación?

El coeficiente de variación es una medida de error relativo a un estimador, se define como:

$$cve(\hat{\theta}) = \frac{se(\hat{\theta})}{\hat{\theta}}$$

Muchas veces se expresa como un porcentaje, aunque no está acotado a la derecha, y por eso es conveniente a la hora de hablar de la precisión de una estadística que viene de una encuesta.

Alertas sobre el coeficiente de variación

Interpretación	Semaforización	Viviendas / Hogares DGES / DGEGSPyJ
Buena		[0%, 15%)
Aceptable		[15%, 25%)
Con reserva		>= 25%

Fuente: INEGI

Alertas sobre el coeficiente de variación

Coeficiente de variación (%)	Número de Observaciones	
	Bajo	**Alto**
[20 , 100]	Estimador no confiable	Estimador no confiable
[15 , 20)	Estimador no confiable	Descriptivo
[5 , 15)	Descriptivo	Estimador confiable
(0 , 5)	Estimador confiable	Estimador confiable

Fuente: INE - Chile

Algunas alertas definidas en la publicación

Cuando se sobrepasa el umbral del coeficiente de variación aparecen algunas de las siguientes alertas:

- No se publica
- Usar con precaución.
- Las estimaciones requieren revisiones, no son precisas y se deben usar con precaución.
- Poco confiable, menos preciso.
- No cumple con los estándares de publicación.
- Con reserva, referencial, cuestionable.
- Valores muy aleatorios, estimación pobre.

Dominios de estudio y subpoblaciones de interés

Una encuesta se planea con el fin de generar información precisa y confiable en los dominios de estudio que se han predefinido. Sin embargo, existen subgrupos poblacionales que la encuesta no abordó en su diseño, y sobre los cuales se quisiera una mayor precisión.

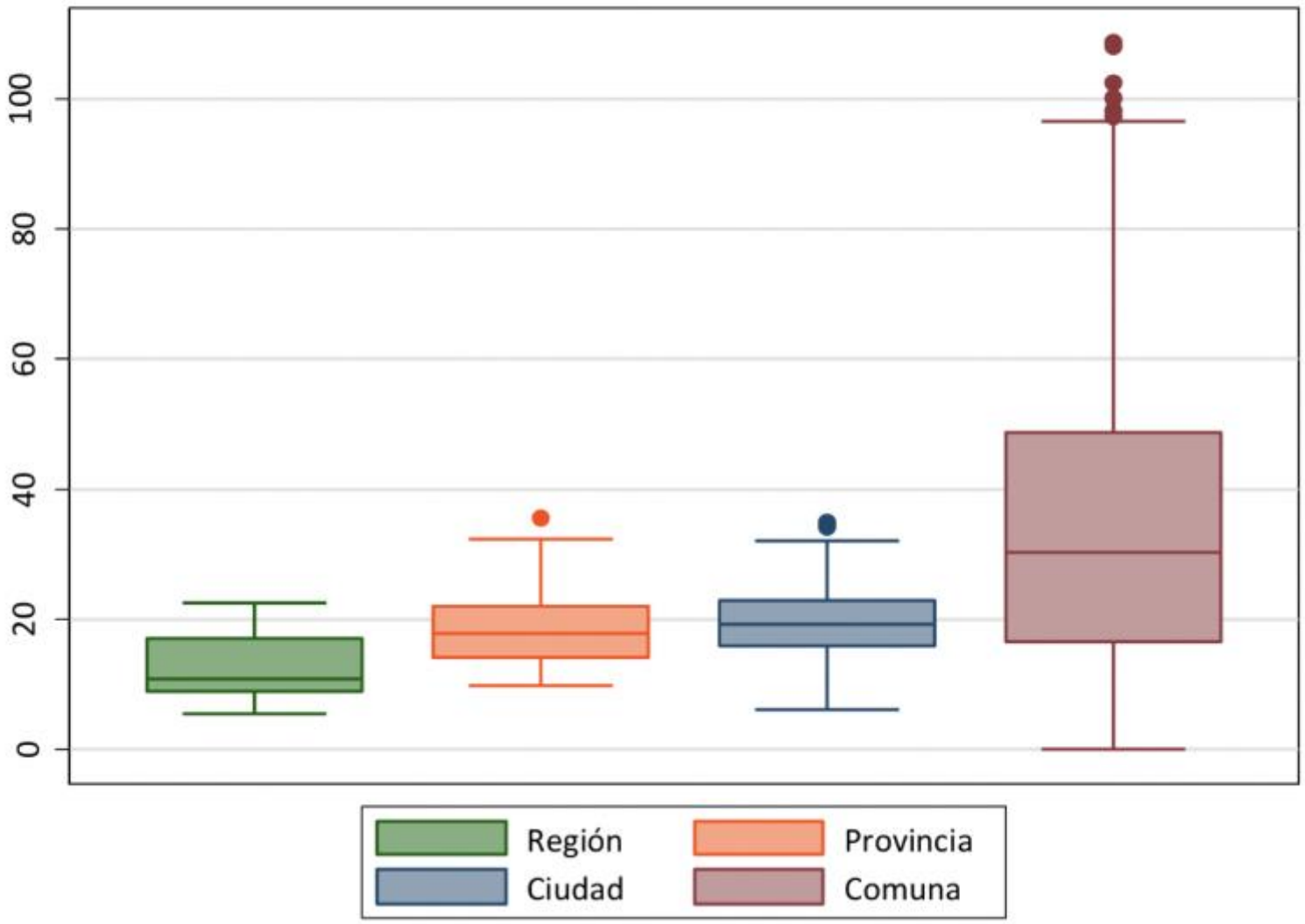
- Incidencia de la pobreza desagregado por departamento o provincia (tamaño de muestra conocido y planificado).
- Tasa de desocupación desagregada por sexo (tamaño de muestra aleatorio, pero planificado).
- Tasa de asistencia neta estudiantil en primaria desagregada por quintiles de ingreso (tamaño de muestra aleatorio).

Uso de métodos SAE

Justificación

- Los estimadores directos, basados solo en unidades de muestreo observadas para cada área pequeña, no son suficientemente confiables.
- Tamaño de muestra pequeño o incluso ninguna unidad observada (falta de información).
- El coeficiente de variación (CV) es demasiado alto para el indicador objetivo a nivel de área.

Incremento del coeficiente de variación



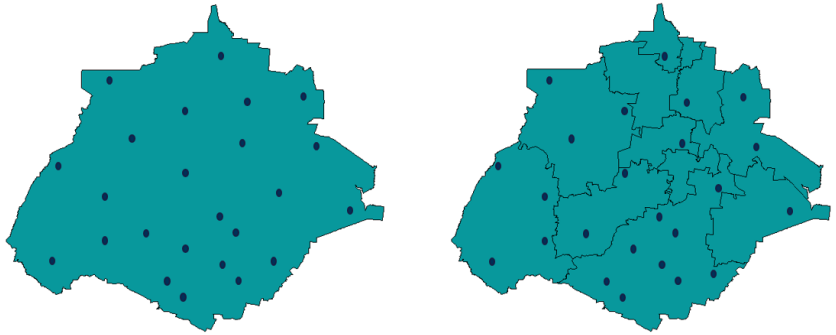
Distribución de los coeficientes de variación en Chile

Justificación

Cuando los estimadores directos no son confiables para algunos dominios de interés, existen dos opciones:

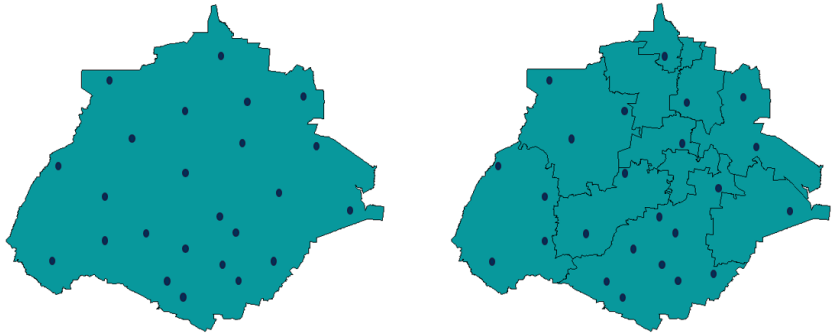
1. Sobremuestreo: aumentar el tamaño de la muestra en los dominios de interés (aumento de los costos).
2. Aplicar técnicas estadísticas que permitan estimaciones confiables en esos dominios, métodos SAE.

¿Qué es un área pequeña?



- La mayoría de las encuestas nacionales están planificadas para entregar estimaciones confiables a nivel nacional y regional pero a niveles más bajos se reduce la precisión.
- Un área pequeña es un dominio para el cual el tamaño de muestra específico no es suficientemente grande para obtener estimaciones confiables.
- Habitualmente son dominios no planificados y su tamaño de muestra esperado es aleatorio y es más grande a medida que aumenta el tamaño de la población del área.

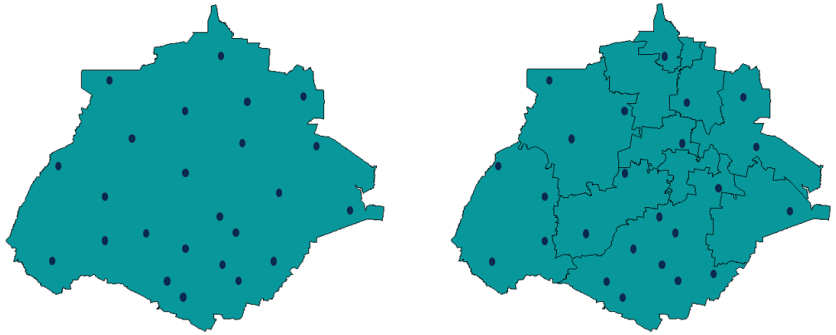
¿Qué es un área pequeña?



La subpoblación de interés puede ser una zona geográfica o subgrupos socioeconómicos.

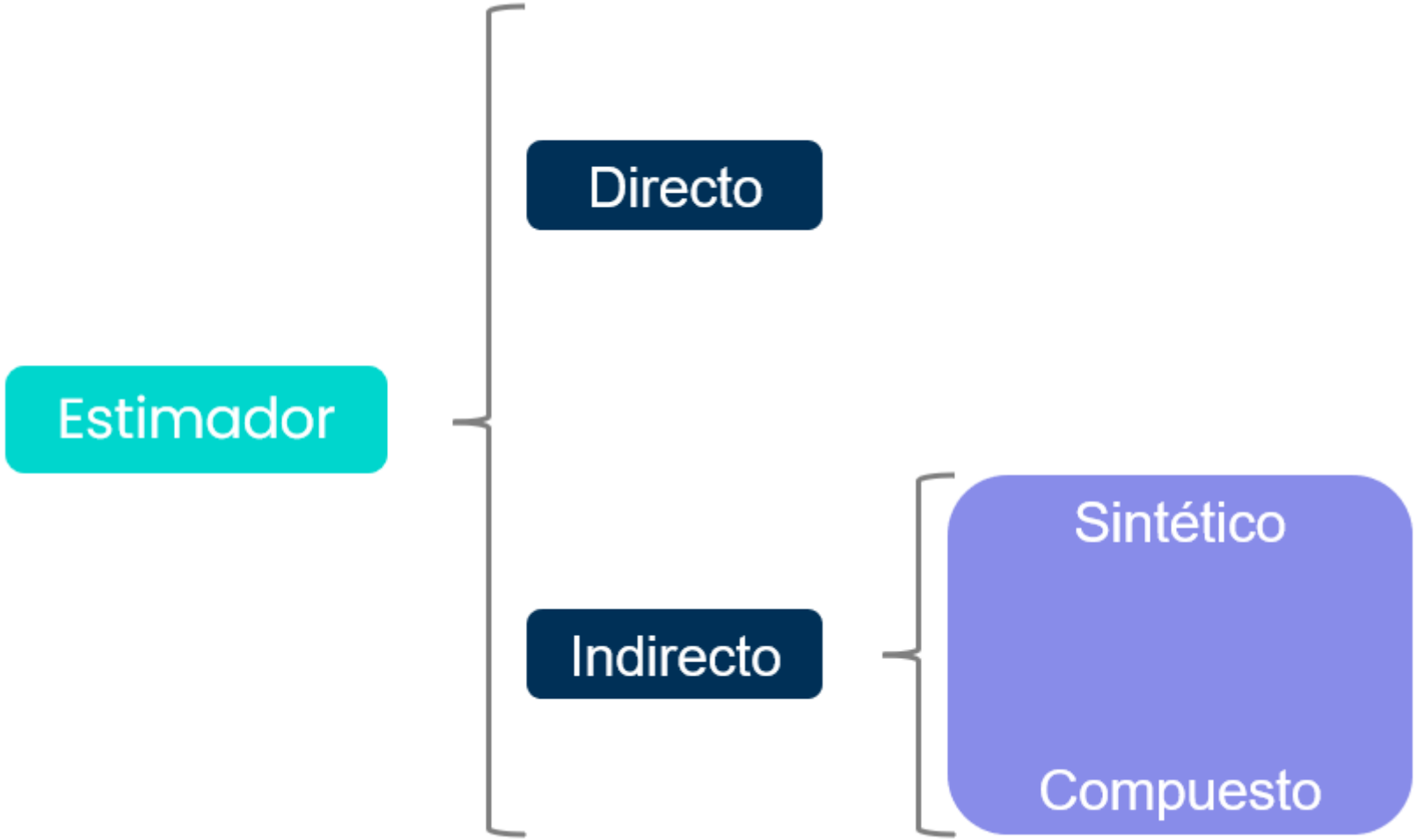
- Geográfico: provincias, áreas del mercado de trabajo, municipios, sectores censales para medir por ejemplo la tasa de desempleo a nivel comunal.
- Dominio de subgrupos específicos: edad $\times$ sexo $\times$ raza dentro del ámbito geográfico de una zona, para medir por ejemplo la tasa de desempleo por sexo o edad específica en las zonas urbanas.

¿Qué es un área pequeña?



- La solución es **tomar prestada fuerza** de otras áreas y/o en diferentes ocasiones mediante modelos explícitos o implícitos que explotan la relación entre variables aumentando el tamaño efectivo de la muestra.
- El modelo proporciona un enlace a áreas relacionadas y/o períodos de tiempo a través de información complementaria tales como recuentos de censos (recientes o actuales) o registros administrativos relacionados con la variable objetivo.

Algunos métodos



Algunos métodos

- Estimador sintético: En el contexto de subpoblaciones, los estimadores se llaman sintéticos cuando éstos se basan en un estimador directo y se estiman a partir de información auxiliar a través de un modelo.
- Estimador compuesto: es una combinación lineal entre un estimador directo y un estimador sintético. Representa un buen compromiso entre las características de los dos componentes.

$$\hat{\bar{Y}}_d^c = \phi_d \hat{\bar{Y}}_d^{DIR} + (1-\phi_d)\hat{\bar{Y}}_d^{SYN}, \qquad 0 \leq \phi_d \leq 1$$

Algunos métodos

- El estimador compuesto está dado por una combinación lineal de estimador sintético y estimador directo equilibrando el sesgo potencial del estimador sintético contra la inestabilidad del estimador directo (compensación entre precisión y sesgo).
- Las estimaciones más grandes de áreas pequeñas están más cerca de las estimaciones directas mientras que las más pequeñas están más cerca de las estimaciones sintéticas.

Algunos métodos

Los modelos de áreas pequeñas se clasifican en dos grandes tipos:

BLUP/EBLUP basado en el modelo Fay-Herriot

BLUP/EBLUP basado en el modelo Fay-Herriot

$$\tilde{\delta}_d^{FH} = \gamma_d \hat{\delta}_d^{DIR} + (1 - \gamma_d)\mathbf{x}_d'\tilde{\boldsymbol{\beta}},$$

es una combinación lineal convexa del estimador directo y del estimador sintético de regresión a nivel de área.

- Si la varianza muestral ψ_d es pequeña comparada con la heterogeneidad no explicada σ_u^2, $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d)$ es cercano a uno.
- Entonces, cuando el tamaño muestral del área es grande (ψ_d pequeña), el BLUP $\tilde{\delta}_d^{FH}$ se acerca al estimador directo.
- Por tanto, no necesitamos saber si el área es pequeña para usar este estimador.

BLUP/EBLUP basado en el modelo Fay-Herriot

- Habitualmente, no sabemos el verdadero valor de σ_u^2 de los efectos aleatorios u_d.
- Sea $\hat{\sigma}_u^2$ un estimador consistente para σ_u^2.
- Entonces, obtenemos el BLUP empírico (*empirical BLUP, EBLUP*) de δ_d,

$$\hat{\delta}_d^{FH} = \hat{\gamma}_d \hat{\delta}_d^{DIR} + (1 - \hat{\gamma}_d)\mathbf{x}_d'\hat{\boldsymbol{\beta}}$$

donde

$$\hat{\gamma}_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \psi_d)$$

BLUP/EBLUP basado en el modelo Fay-Herriot

- En un área no muestreada, la varianza del estimador directo ψ_d tiende a infinito y γ_d tiende a cero
- Tomando el valor límite $\gamma_d = 0$, obtenemos el estimador sintético de regresión,

$$\hat{\delta}_d^{FH} = \mathbf{x}_d' \hat{\boldsymbol{\beta}}$$

Gracias por su atención