Diseño y análisis estadístico de las encuestas de hogares de América Latina

Resumen Orientado al Muestreo

Javier de León
javierdlgomez5@gmail.com

Guatemala

29 de noviembre de 2025

Fuente: CEPAL (2023). Diseño y análisis estadístico de las encuestas de hogares de América Latina.

Contenido

Motivación del Capítulo I

- Toda encuesta es una aproximación a una población finita:

$$U = \{1, 2, \ldots, N\}$$

- Observamos solo una muestra $s \subset U$ con $n = |s|$.
- Queremos estimar un parámetro poblacional θ como la tasa de desempleo en la ENEIC.
- El capítulo introduce el **paradigma del error total**:

$$\text{Error total} = \text{Error de muestreo} + \text{Error no muestral}$$

Parámetros, estimadores y error

- Parámetro poblacional:
$$\theta = \theta(U)$$
- Estimador basado en la muestra:
$$\hat{\theta} = \hat{\theta}(s, \mathbf{y}_s)$$
- Error total en una realización:
$$e = \hat{\theta} - \theta$$
- La función de error usual a considerar es:

$$\mathsf{MSE}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \theta)^2\big] = \underbrace{\mathrm{Var}(\hat{\theta})}_{\text{error de muestreo}} + \underbrace{\mathrm{Sesgo}(\hat{\theta})^2}_{\text{sesgo (suele ser no muestral)}}$$

Error de muestreo

- Surge porque se usa una muestra probabilística.
- Incluso con marco perfecto y medición exacta, $\hat{\theta}$ varía entre muestras.
- Para diseños bien definidos, podemos derivar:
  - varianzas teóricas,
  - aproximaciones; linealización, réplicas, etc.
- Mitigación clásica:
  - aumentar n,
  - optimizar la asignación de estratificación,
  - complejificación de el diseño muestral.

Error no muestral

- Incluye:
  - errores de cobertura del marco,
  - ausencia de respuesta,
  - errores de medición,
  - errores de procesamiento.
- Introduce Sesgo **sesgo**:

$$\text{Sesgo}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta \neq 0$$

- No se corrige simplemente aumentando n.
- Requiere **diseño operativo** y controles de calidad además del diseño estadístico.

Diagrama conceptual del error total

- Los errores pueden clasificarse como:
  - errores de marco y definición,
  - errores del investigador,
  - errores del encuestador,
  - errores del encuestado,
  - error de respuesta y de ausencia.
- El objetivo de la ONE/INE es:

$$\text{Planificar y Coordinar} \Rightarrow \text{minimizar error total}$$

- El muestreo probabilístico es solo una parte del problema.

Proceso de respuesta

- Groves et al. y el libro enfatizan que la respuesta pasa por etapas:
  1. Comprensión de la pregunta
  2. Recuperación de la información de la memoria
  3. Juicio / síntesis de la información
  4. Redacción/comunicación de la respuesta
- En cada etapa puede introducirse:
  - error aleatorio, por ejemplo, ruido Gaussiano,
  - o sesgo sistemático, por ejemplo, subdeclaración de variables sensibles.

Ejemplo: tasa de desempleo

- Sea $Y_k = 1$ si la persona k está desempleada, 0 en caso contrario.
- Parámetro de interés:

$$p = \frac{1}{N} \sum_{k \in U} Y_k$$

- Estimador ponderado con pesos w_k:

$$\hat{p} = \frac{\sum_{k \in s} w_k Y_k}{\sum_{k \in s} w_k}$$

- Fuentes de error:
  - **Muestreo**: varianza de $\hat{p}$ por selección probabilística.
  - **No muestral**: errores en identificar ocupación, recuerdo, no respuesta.

Aplicación INE Guatemala

- Expansión urbana:
  - segmentos que cambian rápidamente,
  - riesgo de que el marco no represente la realidad actual.
- No respuesta:
  - aumento de viviendas cerradas o rechazo,
  - posible sesgo urbano si no se ajusta correctamente.
- Herramientas actuales:
  - CAPI con validaciones y control de tiempos,
  - recontactos, supervisión y auditoría de audio.
- Todo esto encaja en el marco del **error total**.

Universo, población y muestra

- Universo (población objetivo):

$$U = \{1, 2, \ldots, N\}$$

  puede ser:
  - personas residentes,
  - hogares privados,
  - viviendas ocupadas con personas presentes, etc.
- Muestra:

$$s \subset U, \qquad |s| = n$$

  seleccionada mediante un **diseño probabilístico** $p(s)$.
- Para cada elemento k se observa una variable Y_k, o un vector $\mathbf{Y}_k$.

Unidades de muestreo

- Unidad primaria de muestreo **UPM**:
  - conglomerado geográfico (segmento, área censal, sector cartográfico),
  - se selecciona en la primera etapa como su nombre lo dice.
- Unidad secundaria:
  - estructuras, viviendas,hogares dentro de la UPM.
- Unidad de análisis:
  - hogar o persona, según el indicador.
- En notación de diseño por etapas:

$$k = (i, j) \quad \text{hogar } j \text{ en la UPM } i$$

Panel Rotativo

Encuestas rotativas:

- Existe un diseño el cual nos permite visitar repetidamente las unidades de análisis,
- se define un esquema con notación a(b)c,
- cada grupo de rotación entra y sale según un calendario fijo.

Ejemplo tomando un esquema 5(0)1:

- cada hogar participa en 5 períodos consecutivos,
- en cada nuevo período entra un grupo nuevo y sale el más antiguo.

Parámetros de interés

- Totales:

$$t_Y = \sum_{k \in U} Y_k$$

- Medias:

$$\mu_Y = \frac{1}{N} \sum_{k \in U} Y_k$$

- Proporciones:

$$p = \frac{1}{N} \sum_{k \in U} I(Y_k \in A)$$

- Razones:

$$R = \frac{\sum_{k \in U} Y_k}{\sum_{k \in U} X_k}$$

Encuestas para indicadores dinámicos

- Para analizar **cambios** entre períodos:

$$\Delta\theta = \theta_t - \theta_{t-1}$$

- Es preferible contar con:
  - panel o panel rotativo,
  - para explotar la correlación entre mediciones de un mismo hogar:

$$\mathrm{Cov}(Y_{k,t}, Y_{k,t-1}) \neq 0$$

- Esto reduce la varianza de las estimaciones de cambios.

Relación diseño–indicador

- Ejemplos:
  - **Empleo**:
    - requiere series de alta frecuencia,
    - diseño rotativo como en ENEIC, o transversal en ENEI.
  - **Ingresos y gastos**:
    - requiere cuestionario largo,
    - se hace cada varios años como la ENIGH y Encovi,
    - diseño transversal con n relativamente grande.
- Una idea de optimización conceptual:

$$\text{Diseño*} = g(\text{parámetro}, \text{dominios}, \text{presupuesto})$$

Aplicación — Encuestas en Guatemala

- ENEIC:
  - indicador central: tasa de participación, ocupación y desocupación,
  - requiere seguimiento temporal $\rightarrow$ diseño rotativo.
- ENIGH / ENIFH:
  - indicadores de pobreza por ingresos y estructura de gasto,
  - periodicidad mayor, pero mayor detalle en módulos.

Definición de marco de muestreo

- Conjunto de unidades que representan operacionalmente el **universo** a estudiar.
- Debe contener:
  - identificación única por unidad,
  - estructura geográfica/jerárquica,
  - cobertura exhaustiva,
  - información auxiliar.
- Idealmente:

$$U \equiv \text{Marco}$$

pero en la práctica hay errores de cobertura.

Marco y definiciones poblacionales

- Población objetivo estadística:

$$U = \{\text{personas en viviendas particulares}\}$$

- Población accesible en el marco puede excluir:
  - construcción nueva no censada,
  - áreas con difícil acceso.
- Error de cobertura $\Rightarrow$ posible sesgo sistemático.

$$\text{Bias}(\hat{\theta}) \propto \frac{\text{población no cubierta}}{\text{población total}}$$

Jerarquía geográfica del marco

$$\text{País} \rightarrow \text{Departamento} \rightarrow \text{Municipio} \rightarrow \text{Área} \rightarrow \text{Segmento/UPM} \rightarrow \text{Vivienda}$$

- Cada nivel define dominios de estudio y estratificación.
- Las UPMs son conglomerados diseñados para:
  - reducir costos del trabajo de campo,
  - minimizar efectos de conglomeración.

UPM y Conglomerados

- UPM = unidades primarias de muestreo.
- Correlación intraclase ρ dentro del conglomerado:

$$\mathrm{Var}_{CP}(\hat{\mu}) \approx \mathrm{Var}_{SRS}(\hat{\mu}) \cdot [1 + (m-1)\rho]$$

- A mayor ρ y tamaño m, mayor pérdida de eficiencia.
- Estratificar UPMs ayuda a reducir ρ.

Calidad del marco

- Requisitos fundamentales:
  1. **Actualizado**
  2. **Exhaustivo**
  3. **Libre de duplicados**
  4. **Información suficiente para selección y expansión**
- Importancia:

$$\Rightarrow \text{evitar sub/sobre cobertura} \Rightarrow \text{evitar sesgos}$$
$$\Rightarrow \text{garantizar probabilidades de inclusión válidas}$$

Errores de cobertura y sesgo

- Si la población excluida difiere sistemáticamente,

$$\mathrm{Bias}(\hat{\theta}) = (\overline{Y}_{\text{fuera}} - \overline{Y}_{\text{dentro}}) \cdot \frac{N_{\text{fuera}}}{N}$$

- Mitigación:
  - actualización de marco,
  - ampliación de UPMs para incluir nueva urbanización,
  - ajuste por calibración posterior.

Aplicación — INE Guatemala

- Marco Maestro derivado del Censo 2018.
- UPM $\approx$ segmentos cartográficos:

$$[50 - 250] \text{ viviendas} \subset \text{UPM}$$

- Desafíos críticos:
  - Expansión urbana $\rightarrow$ omisiones crecientes.
  - Viviendas desocupadas o para alquiler temporal $\rightarrow$ riesgo operativo.
- Uso de imágenes satelitales para verificación en campo.

Conexión con el FC01

El FC01 es un listado de viviendas realizado luego de un levantamiento cartográfico en una UPM seleccionada, es instrumento de actualización del marco.

- Importante para:
  - ubicar viviendas reales dentro del segmento,
  - detectar crecimiento o contracción de UPMs,
  - situación real vs. marco original.

Objetivo de la estratificación

- **Controlar la variabilidad** dividiendo la población en grupos homogéneos.
- Se garantiza **representatividad** en dominios clave.
- Se reduce la varianza del estimador:

$$\mathrm{Var}(\hat{\mu}_{strat}) = \sum_{h=1}^{H} W_h^2 \frac{S_h^2}{n_h}$$

- Comparar con SRS:

$$\mathrm{Var}_{SRS}(\hat{\mu}) = \frac{1 - \frac{n}{N}}{n} S^2$$

SRS vs Estratificado: Varianzas

Supuestos:

- H estratos con pesos $W_h = \frac{N_h}{N}$,
- Media del estrato h: μ_h; varianza: S_h^2,
- Asignación proporcional: $n_h = nW_h$,
- Ignoramos la corrección de población finita.

Descomposición de la Varianza Total

Descomposición clásica:

$$S^2 = \underbrace{\sum_{h=1}^{H} W_h S_h^2}_{\text{varianza dentro de estratos}} + \underbrace{\sum_{h=1}^{H} W_h (\mu_h - \mu)^2}_{\text{varianza entre estratos}}.$$

La segunda parte mide cuán diferentes son los estratos entre sí. Si los estratos están bien construidos, esta cantidad es grande.

Demostración: ¿Por qué Estratificar?

$$\bar{Y}_{\mathsf{SRS}} - \bar{Y}_{\mathsf{st}} = \frac{1}{n}\Big[S^2 - \sum_{h=1}^{H} W_h S_h^2\Big].$$

Reemplazando la descomposición:

$$= \frac{1}{n}\sum_{h=1}^{H} W_h(\mu_h - \mu)^2 \ \geq 0.$$

$$\implies \bar{Y}_{\mathsf{st}} \leq \bar{Y}_{\mathsf{SRS}}.$$

Criterios comunes de estratificación

- Geográfico
- Áreas
- Socioeconómico
- Tamaño de UPM
- Combinaciones:

Estratos: Área $\times$ Departamento

Estratificación + Conglomerados

- En encuestas de hogares:

  Estratos $\Rightarrow$ dónde muestrear

  UPMs $\Rightarrow$ cómo muestrear

- La estratificación **mitiga** el exceso de varianza por conglomeración.

$$DEFF = 1 + (m - 1)\rho$$

- La estratificación ayuda a reducir ρ dentro de estratos.

Asignación de muestra — Regla de Neyman

- Objetivo: minimizar varianza para tamaño n.
- Asignación óptima:

$$n_h = n \cdot \frac{N_h S_h}{\sum_{j=1}^{H} N_j S_j}$$

- Intuición:
  - Más muestra donde N_h es mayor en conjunto con,
  - la variabilidad interna S_h.

Aplicación — INE Guatemala

- Estratos típicos:

$$\text{Estratos} = \underbrace{22}_{\text{Deptos}} \times \underbrace{2}_{\text{Área}} = 44$$

- UPM seleccionados PPT .
- Tamaño objetivo por UPM en campo:

$$m \approx 10 \text{ viviendas}$$

- Meta: precisión departamental.

Conclusiones de Diseños Complejos

- Estratificación $\downarrow$ Varianza de estimadores
- Conglomeración $\uparrow$ Varianza pero $\downarrow$ costo operativo
- Diseño óptimo: balance costo vs precisión

Selección probabilística

- Cada unidad $k \in U$ debe tener:

$$\pi_k = \mathbb{P}(k \in s) > 0$$

- Propósito:
  - estimadores insesgados,
  - varianza calculable,
  - representatividad estadística.
- Generalmente por etapas:

$$\text{UPM} \Rightarrow \text{Vivienda} \Rightarrow \text{Hogar/persona}$$

Probabilidades de inclusión

Tres niveles típicos

$$\pi_i = \mathbb{P}(\text{UPM } i \in s)$$

$$\pi_{j|i} = \mathbb{P}(\text{hogar } j \in s \mid i \text{ seleccionado})$$

$$\pi_{ij} = \pi_i \cdot \pi_{j|i}$$

- Los pesos iniciales o persos base del muestreo son:

$$d_{ij}^0 = \frac{1}{\pi_{ij}}$$

- Son la base para el cálculo de los factores de expansión.

Selección Primera Etapa

- UPM seleccionadas con probabilidad proporcional al tamaño **PPT**:

$$\pi_i = \frac{n_I \cdot N_{II_i}}{\sum_{i=1}^{U} N_{II_i}}$$

- Intuición:
  - UPM más grandes tienen mayor chance de ser seleccionadas.
  - Reduce varianza asociada al tamaño de conglomerados.

Selección MAS o SRS dentro de UPM

- Listado FC01 $\rightarrow N_{II}$ viviendas en la UPM especificada.
- Se requiere seleccionar m viviendas:

$$\pi_k = \frac{m}{N_{II}}$$

Proporciona el diseño conceptual canónico o base, este es uno de los modelos de selección más comunes implementados recientemente por el INE.

Selección sistemática dentro de UPM

- Listado FC01 $\rightarrow N_{II}$ viviendas en la UPM especificada.
- Se requiere seleccionar m viviendas:

$$k = 1 + \lfloor r + (j-1) \cdot \frac{N_{II}}{m} \rfloor$$

  donde $r \sim U(0, \frac{N_{II}}{m})$
- simple de implementar en campo,
- buena dispersión espacial,
- requiere aproximación de la varianza $\downarrow$

Selección de hogar dentro de vivienda

- Dependiendo de la encuesta:
  - todos los hogares de la vivienda.
  - Selección aleatoria de 1 hogar o persona.
- Ejemplo persona aleatoria:

$$\pi_{p|h} = \frac{1}{H_h}$$

si hay H_h personas elegibles en el hogar.

Ejemplo Aplicado

1. Selección de UPM con PPT usando el Marco Maestro de Muestreo.
2. En cada UPM seleccionada:
    1. Listado FC01 actualizado
    2. Selección de 10 viviendas
    3. Si hogar único $\rightarrow$ se encuesta completo
    4. Si múltiples hogares $\rightarrow$ selección aleatoria interna
3. Se calculan:

$$\pi_i,\ \pi_{j|i},$$
$$d_{ij}^0 = 1/\pi_{ij}$$

Control de la variabilidad del tamaño

- El PPT busca compensar:

$$\text{UPM grandes} \Rightarrow \text{más probabilidad}$$

$$\text{UPM pequeñas} \Rightarrow \text{menos probabilidad}$$

- Se evita que el número de viviendas observadas por habitante varíe demasiado.
- Necesario para eficiencia y para pesos estables.

Conclusiones del Capítulo V

- Selección probabilística asegura validez inferencial.
- PPT + selección sistemática/MAS es estándar en la región.
- El FC01 es clave para garantizar cobertura y selección válida.
- El PRN preserva aleatoriedad y seguimiento longitudinal.

Gracias por su Atención