en la Asamblea General de Tenerife. Observad el símil con algunos equipos de fútbol que al subir a primera división tuvieron un aumento de costes que no pudieron asimilar y acabaron en bancarrota. Nosotros no queremos que nos ocurra eso: "no queremos morir de éxito".

Entramos en el siglo XXI, es la era de la globalización, el mercado de revistas ya no funciona por suscripciones en papel a revistas concretas sino por suscripciones electrónicas a bloques de revistas, la mayoría de sociedades de Estadística y de Investigación Operativa de los países occidentales han cedido la edición, impresión, distribución y comercialización de sus revistas a grandes editoriales. Nosotros en algún sentido somos el último mohicano, somos como una hormiga en un mercado de gigantes, y no está nada claro que este romanticismo sea viable a la larga. Mi opinión personal es que tenemos que ser pragmáticos, entrar en el juego a lo grande, y aprovecharnos de la situación de TEST para obtener condiciones económicas ventajosas para TEST y TOP en una futura negociación con editoriales científicas.

En estos momentos nuestros editores, Leandro Pardo Llorente, María Ángeles Gil Álvarez, Marco A. López Cerdá e Ignacio García Jurado, han sido autorizados a iniciar conversaciones con varias editoriales. Previsiblemente este proceso acabará en una negociación que conduzca a un acuerdo que, sin ceder la propiedad y gestión científica de las revistas, permita: (1) desligarnos de la edición técnica, impresión, distribución, comercialización y visibilidad en Internet de nuestras revistas, (2) obtener una compensación económica sustancial que mejore significativamente el balance económico de la SEIO.

Los editores de TEST y TOP, y yo mismo, estamos convencidos de que éste es el camino de la modernidad y es la senda que tenemos que recorrer para seguir manteniendo las altas cotas de calidad de nuestras revistas con un esfuerzo razonable. Los tiempos del amateurismo y del trabajo artesanal se están acabando. En definitiva lo que estamos buscando es un acuerdo temporal (por ejemplo, de cinco años) y renovable a solicitud de ambas partes. También somos conscientes de la relevancia del paso que queremos dar, y en consecuencia todo eventual acuerdo con alguna editorial será previamente sometido a la aprobación del Consejo Ejecutivo y de la Asamblea General de la SEIO.

# 1. ARTÍCULOS DE ESTADÍSTICA

## ESTIMADORES CLÁSICOS DE ÁREAS PEQUEÑAS

**Domingo Morales González**
Centro de Investigación Operativa
Universidad Miguel Hernández de Elche

### 1. Introducción

El muestreo estadístico, en contraposición con los censos, permite obtener información sobre materias muy dispares con un coste reducido. El muestreo se utiliza no solamente para la obtención de estimaciones en la población completa, sino para estimar parámetros en una variedad de subpoblaciones (*dominios*). Los dominios se definen generalmente como áreas geográficas o grupos socioeconómicos. Ejemplos de dominios geográficos (*áreas*) son las comarcas, islas, municipios, distritos sanitarios, etc. Ejemplos de grupos socioeconómicos son grupos sexo-edad, sectores industriales o empresariales, etc.

En el contexto de la estimación en áreas pequeñas, se dice que un estimador de un parámetro en un dominio dado es *directo* si está basado solamente en los datos específicos del dominio. Un estimador directo puede usar también información auxiliar, como por ejemplo el total en el dominio de una variable $x$ relacionada con la variable de interés $y$. Un estimador directo es típicamente un estimador basado en el diseño muestral, aunque en ocasiones su uso se justifique con modelos. Los estimadores basados en el diseño muestral utilizan los pesos muestrales. Las inferencias derivadas de los mismos están basadas en la distribución de probabilidad inducida por el mecanismo aleatorio de

extracción de la muestra, bajo el supuesto de que los valores de la variables en los elementos de la población permanecen fijos. Los estimadores *asistidos por modelos* se introducen a partir de modelos de trabajo, pero optimizando sus propiedades de sesgo y varianza respecto de la distribución del diseño. En la literatura estadística estos estimadores también se consideran basados en el diseño muestral.

Un dominio es *grande* si la muestra específica del dominio es suficientemente grande para obtener estimadores directos con una precisión adecuada. Un dominio es *pequeño* en caso contrario. En este texto usaremos el término *área pequeña* para denotar a los dominios pequeños.

En este artículo se realiza una recopilación de estimadores de totales de dominios. Los estimadores que se presentan son directos o asistidos por modelos y se pueden considerar como los "más clásicos" dentro de la Teoría de Estimación en Áreas Pequeñas.

## 2. Notación

Se utiliza la siguiente notación:

- *Índices:* $s$ se usa para la muestra, $d = 1, \ldots, D$ para las áreas pequeñas (comarcas cruzadas con sexo, en este manuscrito), $j = 1, \ldots, N$ para los individuos y $g = 1, \ldots, G$ para los grupos.
- *Población y muestra:* $P = \cup_{d=1}^{D} P_d$ para la población y $s = \cup_{d=1}^{D} s_d$ para la muestra.
- *Tamaños:* $N$ para población y $n$ para muestra. Cuando $N$ o $n$ tienen subíndices entonces denotan el tamaño del correspondiente conjunto indicado.
- *Totales:* $Y$ *o* $X$. Cuando $Y$ o $X$ tienen subíndices entonces denotan el tamaño del correspondiente conjunto indicado.
- *Medias:* $\overline{Y}$ *o* $\overline{X}$. Cuando $\overline{Y}$ o $\overline{X}$ tienen subíndices entonces denotan el tamaño del correspondiente conjunto indicado. Por ejemplo, $\overline{Y}_d$ es la media poblacional del área $d$.
- *Pesos no calibrados:* $w_j$. Son los pesos teóricos del diseño muestral corregidos por la falta de respuesta (factores sin calibrar). En el caso en que la no respuesta se modele mediante variables aleatorias de Bernoulli independientes, los pesos no calibrados serían las inversas de las probabilidades finales de inclusión de individuos; es decir $w_j = 1/\pi_j$.

Para una variable $x_j$, $j = 1, \ldots, N$, interesa en ocasiones reflejar las pertenencias a un área, grupo o al cruce de ambas ($j \in P_d$, $j \in P_g$, $j \in P_d \cap P_g$). En tales casos escribimos $x_{dj}$, $x_{gj}$ y $x_{dgj}$ respectivamente.

## 3. Teoría del diseño muestral

Para el enfoque basado en el diseño muestral $(y_1, \ldots, y_N)$ es el parámetro básico. Un *plan (o diseño) de muestreo probabilístico* es un esquema para elegir las muestras, de forma que cada subconjunto $s$ de la población tenga una probabilidad de selección $p(s)$ conocida. Las definiciones de sesgo y varianza se hacen respecto de $p(s)$.

SESGO: $\mathsf{E}_\pi[\widehat{T} - T] = \sum_s p(s)(\widehat{T}(s) - T)$,
VARIANZA: $\mathsf{V}_\pi[\widehat{T}] = \sum_s p(s)(\widehat{T}(s) - \mathsf{E}_\pi[\widehat{T}])^2$.

Se usa la notación $\mathsf{E}_\pi$ y $\mathsf{V}_\pi$ para recalcar el hecho de que se trata de esperanzas y varianzas respecto de la probabilidad $p(s)$ del diseño (y no respecto de la probabilidad de un modelo: $\mathsf{E}_M$, $\mathsf{V}_M$).

La probabilidad $p(s)$ es en general difícil de calcular. Sin embargo, en la mayoría de los cálculos sólo son necesarias las probabilidades de inclusión $\pi_i = \mathsf{P}(\text{Seleccionar la unidad } i) = \sum_{s \in s(i)} p(s)$, donde $s(i)$ es el conjunto de todas las muestras que contienen la unidad $i$,
$\pi_{ij} = \mathsf{P}(\text{Seleccionar conjuntamente las unidades } i \text{ y } j) = \sum_{s \in s(i,j)} p(s)$, donde $s(i,j) = \{s \subset P \ / \ i, j \in s\}$.

## 4. Estimador directo del total

El estimador *directo* del total $Y_d$ es

$$\widehat{Y}_d^{direct} = \sum_{j \in s_d} w_j y_j = \sum_{j \in s_d} \frac{y_j}{\pi_j} y_j.$$

**Proposición 4.1.** Si $\pi_j > 0 \ \forall j \in P_d$, entonces

(a) $\mathsf{E}_\pi[\widehat{Y}_d^{direct}] = Y_d$,

(b) $\mathsf{V}_\pi[\widehat{Y}_d^{direct}] = \sum_{i=1}^{N_d} \sum_{j=1}^{N_d} (\pi_{ij} - \pi_i \pi_j) \frac{y_j}{\pi_j} \frac{y_j}{\pi_j}$,

(c) $\widehat{\mathsf{V}}_\pi[\widehat{Y}_d^{direct}] = \sum_{i\in s_d}\sum_{j\in s_d} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}} \frac{y_j}{\pi_j}\frac{y_j}{\pi_j}$ es un estimador insesgado de

$$\mathsf{V}_\pi[\widehat{Y}_d^{direct}].$$

En el caso particular de que $\pi_{ij} = \pi_i\pi_j$ si $i \neq j$ (o equivalentemente, cuando las probabilidades de doble inclusión son sustancialmente inferiores a las de inclusión simple), y teniendo en cuenta que $\pi_{jj} = \pi_j$, se obtiene

$$\begin{aligned}
\mathsf{V}_\pi[\widehat{Y}_d^{direct}] &= \sum_{j\in P_d} \frac{1-\pi_j}{\pi_j}\, y_j^2, \\
\widehat{\mathsf{V}}_\pi[\widehat{Y}_d^{direct}] &= \sum_{j\in s_d} \frac{1-\pi_j}{\pi_j^2}\, y_j^2 \\
&= \sum_{j\in s_d} w_j(w_j-1)y_j^2.
\end{aligned}$$

## 5. Estimador directo de la media

El estimador *directo* de la media es

$$\widehat{\overline{Y}}_d^{direct} = \frac{\widehat{Y}_d^{direct}}{\widehat{N}_d} = \frac{\sum_{j\in s_d} w_j y_j}{\sum_{j\in s_d} w_j}.$$

**Proposición 5.1.** Si $\pi_j > 0\ \forall j \in P_d$, entonces

(a) $\mathsf{E}_\pi[\widehat{\overline{Y}}_d^{direct}] \approx \overline{Y}_d$,

(b) $\mathsf{V}_\pi[\widehat{\overline{Y}}_d^{direct}] \approx \frac{1}{N_d^2}\sum_{i\in P_d}\sum_{j\in P_d} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}(y_i - \overline{Y}_d)(y_j-\overline{Y}_d)$,

(c) $\widehat{\mathsf{V}}_\pi[\widehat{\overline{Y}}_d^{direct}] = \frac{1}{\widehat{N}_d^2}\sum_{i\in s_d}\sum_{j\in s_d} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}(y_i - \widehat{\overline{Y}}_d^{direct})(y_j-\widehat{\overline{Y}}_d^{direct})$.

Teniendo en cuenta que $\pi_{jj} = \pi_j$, en el caso particular de que $\pi_{ij} = \pi_i\pi_j$ $(i \neq j)$ se obtiene

$$\begin{aligned}
\mathsf{V}_\pi[\widehat{\overline{Y}}_d^{direct}] &\approx \frac{1}{N_d^2}\sum_{j\in P_d}\frac{1-\pi_j}{\pi_j}(y_j-\overline{Y}_d)^2, \\
\widehat{\mathsf{V}}_\pi[\widehat{\overline{Y}}_d^{direct}] &= \frac{1}{\widehat{N}_d^2}\sum_{j\in s_d}\frac{1-\pi_j}{\pi_j^2}(y_j-\widehat{\overline{Y}}_d^{direct})^2 \\
&= \frac{1}{\widehat{N}_d^2}\sum_{j\in s_d} w_j(w_j-1)(y_j-\widehat{\overline{Y}}_d^{direct})^2.
\end{aligned}$$

Dado que el estimador directo es aproximadamente insesgado, entonces el error cuadrático medio verifica

$$\begin{aligned}
MSE[\widehat{\overline{Y}}_d^{direct}] &\approx \mathsf{V}_\pi[\widehat{\overline{Y}}_d^{direct}], \\
mse[\widehat{\overline{Y}}_d^{direct}] &= \widehat{\mathsf{V}}_\pi[\widehat{\overline{Y}}_d^{direct}].
\end{aligned}$$

Para más detalles, véase Särndal (1992), pp. 185, 391, o Rao (2003), pp. 12.

## 6. Estimador sintético básico

El estimador *sintético básico* del total es

$$\widehat{Y}_d^{synth} = \sum_{g=1}^{G} N_{dg}\widehat{\overline{Y}}_g^{direct}.$$

Se trata de un estimador sesgado (véase Särndal (1992), pp. 410). Su esperanza aproximada es

$$\mathsf{E}_\pi[\widehat{Y}_d^{synth}] \approx \sum_{g=1}^{G} N_{dg}\overline{Y}_g$$

y su sesgo es

$$\begin{aligned}
\mathsf{B}_\pi[\widehat{Y}_d^{synth}] &= \mathsf{E}_\pi[\widehat{Y}_d^{synth}] - \sum_{j\in P_d} y_j \\
&\approx \sum_{g=1}^{G} N_{dg}(\overline{Y}_g - \overline{Y}_{dg}).
\end{aligned}$$

Si $\pi_{ij} = \pi_i\pi_j$ $(i \neq j)$ y $\pi_{jj} = \pi_j$, entonces

$$\begin{aligned}
V &= \mathsf{V}_\pi[\widehat{Y}_d^{synth}] = \sum_{g=1}^{G} N_{dg}^2\, \mathsf{V}_\pi[\widehat{\overline{Y}}_g^{direct}] \\
&\approx \sum_{g=1}^{G}\frac{N_{dg}^2}{N_g^2}\sum_{j\in P_g}\frac{1-\pi_j}{\pi_j}(y_j-\overline{Y}_g)^2, \\
\widehat{V} &= \widehat{\mathsf{V}}_\pi[\widehat{Y}_d^{synth}] \\
&= \sum_{g=1}^{G}\frac{N_{dg}^2}{\widehat{N}_g^2}\sum_{j\in s_g} w_j(w_j-1)(y_j-\widehat{\overline{Y}}_g^{direct})^2.
\end{aligned}$$

## 7. Estimador post-estratificado

El estimador *post-estratificado* del total es

$$\widehat{Y}_d^{pst} = \sum_{g=1}^{G} N_{dg}\widehat{\overline{Y}}_{dg}^{direct}.$$

Se trata de un estimador aproximadamente insesgado (véase Särndal (1992), pp. 406-407). Si $\pi_{ij} = \pi_i\pi_j$ $(i \neq j)$ y $\pi_{jj} = \pi_j$, entonces

$$\begin{aligned} V &= \mathsf{V}_\pi[\widehat{Y}_d^{pst}] = \sum_{g=1}^{G} N_{dg}^2 \mathsf{V}_\pi[\widehat{\overline{Y}}_{dg}^{direct}] \\ &\approx \sum_{g=1}^{G} \sum_{j\in P_g\cap P_d} \frac{1-\pi_j}{\pi_j}(y_j - \overline{Y}_{dg})^2, \\ \widehat{V} &= \widehat{\mathsf{V}}_\pi[\widehat{Y}_d^{pst}] \\ &= \sum_{g=1}^{G} \frac{N_{dg}^2}{\widehat{N}_{dg}^2} \sum_{j\in s_g} w_j(w_j-1)(y_j - \widehat{\overline{Y}}_{dg}^{direct})^2. \end{aligned}$$

## 8. Estimador dependiente del tamaño de la muestra

Drew et al. (1982) propusieron el estimador dependiente del tamaño de la muestra

$$\widehat{Y}_d^{ssd} = \gamma_d \widehat{Y}_d^{pst} + (1-\gamma_d)\widehat{Y}_d^{synth},$$

donde

$$\gamma_d = \begin{cases} 1 & \text{si } \widehat{N}_d^{direct} \geq \delta N_d \\ \frac{\widehat{N}_d^{direct}}{\delta N_d} & \text{en caso contrario.} \end{cases}$$

La constante $\delta$ se elige para controlar la contribución del componente sintético. Algunas veces su valor es $\delta = 1$ y otra veces es $\delta = 2/3$ (por ejemplo, en la Encuesta de Población Activa Canadiense).

## 9. Estimador GREG

Consideremos $p$ variables explicativas evaluadas en las $N$ unidades de la población; es decir, $\boldsymbol{x}_j = (x_{j,1}, \ldots, x_{j,p})$, $j = 1, \ldots, N$. Sean las medias poblacionales y directas

$$\overline{\boldsymbol{X}}_d = \frac{1}{N_d}\sum_{j\in P_d} \boldsymbol{x}_j \text{ y } \widehat{\overline{\boldsymbol{X}}}_d^{direct} = \frac{1}{\widehat{N}_d}\sum_{j\in s_d} w_j \boldsymbol{x}_j.$$

Sea el modelo

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

donde $\boldsymbol{X}$ es una matriz $n \times p$ con filas $\boldsymbol{x}_j$, $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma_e^2 \boldsymbol{W}^{-1})$ y $\boldsymbol{W} = diag(w_1, \ldots, w_n)$. El estimador mínimo cuadrático de $\boldsymbol{\beta}$ es

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{y} \\ &= \left(\sum_{j\in s} w_j \boldsymbol{x}_j^t \boldsymbol{x}_j\right)^{-1} \left(\sum_{j\in s} w_j \boldsymbol{x}_j^t y_j\right). \end{aligned}$$

El estimador *GREG* del total es

$$\widehat{Y}_d^{greg} = N_d \widehat{\overline{Y}}_d^{direct} + N_d(\overline{\boldsymbol{X}}_d - \widehat{\overline{\boldsymbol{X}}}_d^{direct})\widehat{\boldsymbol{\beta}}.$$

Se trata de un estimador aproximadamente insesgado (véase Särndal (1992), pp. 401). La varianza aproximada de $\widehat{Y}_d^{greg}$ es

$$\mathsf{V} \approx \sum_{i\in P_d}\sum_{j\in P_d} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}(E_i - \overline{E}_d)(E_j - \overline{E}_d),$$

donde

$$\begin{aligned} \overline{E}_d &= \frac{1}{N_d}\sum_{j\in P_d} E_j, \quad E_j = y_j - \boldsymbol{x}_j\boldsymbol{B}, \\ \boldsymbol{B} &= \left(\sum_{j\in P} w_j \boldsymbol{x}_j^t \boldsymbol{x}_j\right)^{-1} \left(\sum_{j\in P} w_j \boldsymbol{x}_j^t y_j\right). \end{aligned}$$

Un estimador de $\mathsf{V}_\pi[\widehat{Y}_d^{greg}]$ es

$$\widehat{\mathsf{V}} = \sum_{i\in s}\sum_{j\in s} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j} g_{di}g_{dj}(y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})(y_j - \boldsymbol{x}_j\widehat{\boldsymbol{\beta}}),$$

Si $\pi_{ij} = \pi_i\pi_j$ $(i \neq j)$ y $\pi_{jj} = \pi_j$, entonces

$$\begin{aligned} \mathsf{V}_\pi[\widehat{Y}_d^{greg}] &\approx \sum_{j\in P_d} \frac{1-\pi_j}{\pi_j}(E_j - \overline{E}_d)^2, \\ \widehat{\mathsf{V}}_\pi[\widehat{Y}_d^{greg}] &= \sum_{j\in s} w_j(w_j-1)g_{dj}^2(y_j - \boldsymbol{x}_j\widehat{\boldsymbol{\beta}})^2. \end{aligned}$$

## Referencias


Drew, D., Singh, M.P., and Choudhry, G.H. (1982). Evaluation of small area estimation techniques for the Canadian labour force survey. *Survey Methodology*, **8**, 17–47.

Rao, J.N.K. (2003). *Small area estimation.* John Wiley.

Särndal C.E., Swensson B. and Wretman J. (1992). *Model assisted survey sampling.* Springer-Verlag.