Reflexiones acerca de la construcción de indicadores económicos usando muestras ponderadas

Juan Carlos Martínez Ovando
(Trabajo conjunto con Sergio Olivares y Adriana Roldán)
Banco de México

Seminario Aleatorio, ITAM
Noviembre 8, 2013

Índice

1 Motivación

2 Inferencia en poblaciones finitas

3 Una alternativa predictiva

4 Indicadores salariales

5 Discusión

1. Motivación

1.1 Indicadores económicos que usan *muestreo* y *métodos ponderados*:

- Indicadores de ingreso y gasto (ENIGH-INEGI)
- Indicadores salariales (ENOE-INEGI)
- Indicadores de opinión (INEGI, etc.)
- Otros indicadores económicos (Banxico)

Aspecto común: **Inferencia en poblaciones finitas**.

Dominios planeados = Estratos (e.g., ENIGH - rural/urbano).

Dominios no planeados = Desagregaciones no planeadas
(e.g., ENIGH, ENOE - género, empleo formal y no formal, ciudades).

1.2 Alcances y limitaciones de los métodos ponderados

1. **No son inferencialmente válidos** en ausencia de un marco muestral definido.

2. La construcción de los factores de expansión **no toma en consideración los valores de** (y_j) en $\mathcal{S}$.

3. Supone que **los valores individuales no observados** en $\mathcal{S}$ son iguales a alguno de los valores observados.

   2 y 3 $\Rightarrow$ Son **altamente sensibles a mediciones extremas (o 'outliers').**

4. Se **extenúan** al estudiar **dominios no planeados.**

5. Es **sumamente difícil hacer estimaciones por densidad** de las características de interés, salvo por *"aproximaciones asintóticas"*.

$\Rightarrow$ **Nuestro método permite resolver estas limitaciones, de una manera integral.**

* **Ambos consideran elementos subjetivos, y descansan en aleatorización.**

1.3 Resumen comparativo

Método ponderado

$$\begin{aligned}\hat{T} &= \sum_{j\in\mathcal{S}} \frac{N}{n} y_j \\ &= \sum_{j\in\mathcal{S}} y_j + N^{\widetilde{\mathcal{S}}} \left[\sum_{k=1}^{U} \frac{m_k}{n} y_k^* \right],\end{aligned}$$

Nuestro método

$$\hat{T} = \sum_{j\in\mathcal{S}} y_j + N^{\widetilde{\mathcal{S}}} \left[\sum_{k=1}^{U} \rho_k \cdot y_k^* + \phi \cdot \mu_0 \right],$$

Subyace, para $Y_f \in \widetilde{\mathcal{S}}$:

$$Q(Y_f \in A|\mathcal{S}) = \sum_{k=1}^{U} \frac{m_k}{n} \delta_{y_k^*}(A),$$

$$P(Y_f \in A|\mathcal{S}) = \sum_{k=1}^{U} \rho_k \cdot \delta_{y_k^*}(A) + \phi \cdot G_0(A),$$

Alcances:

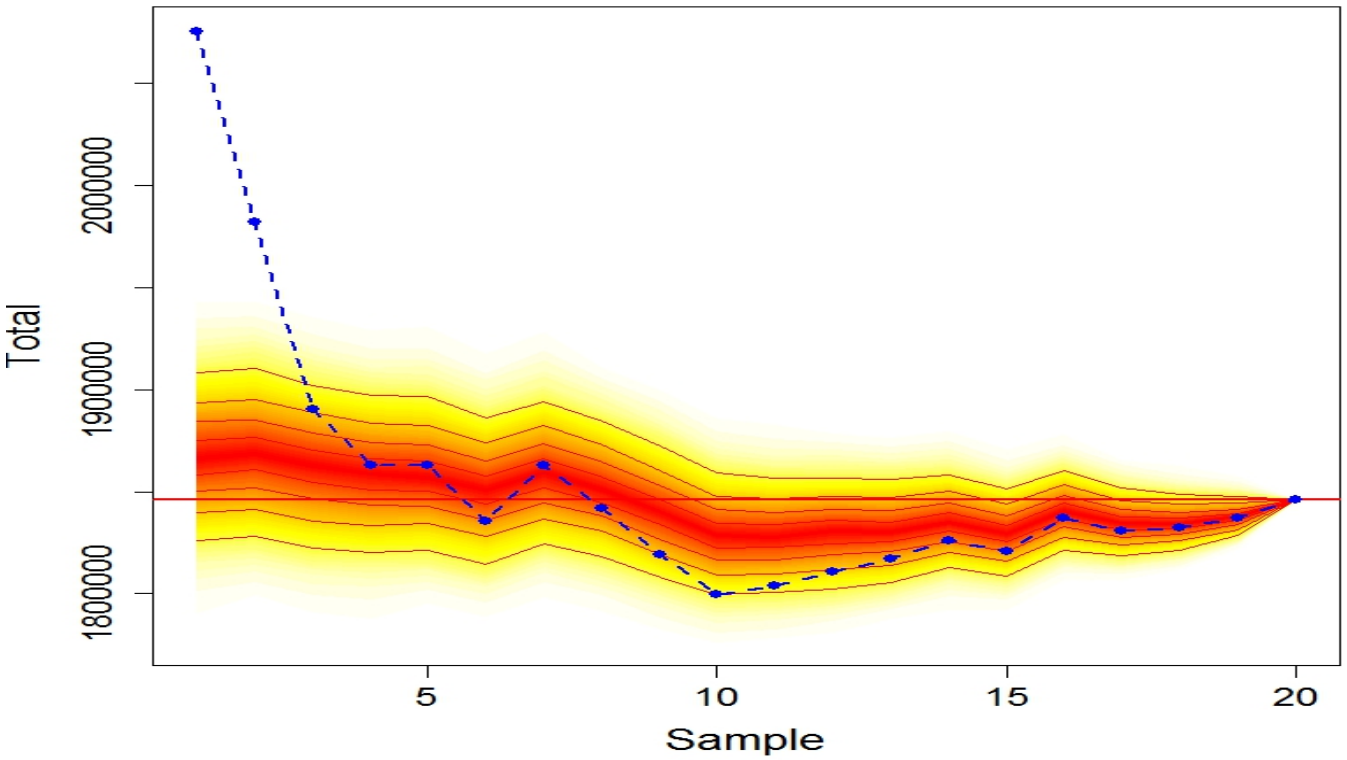
$$\hat{T} \quad var(\hat{T}), \quad sesgo(\hat{T})$$

$$P(T \leq t|\mathcal{S}).$$

$\Rightarrow$ **Ambos métodos son convergentes, el tema es qué tan rápido alcanzan tal convergencia.**

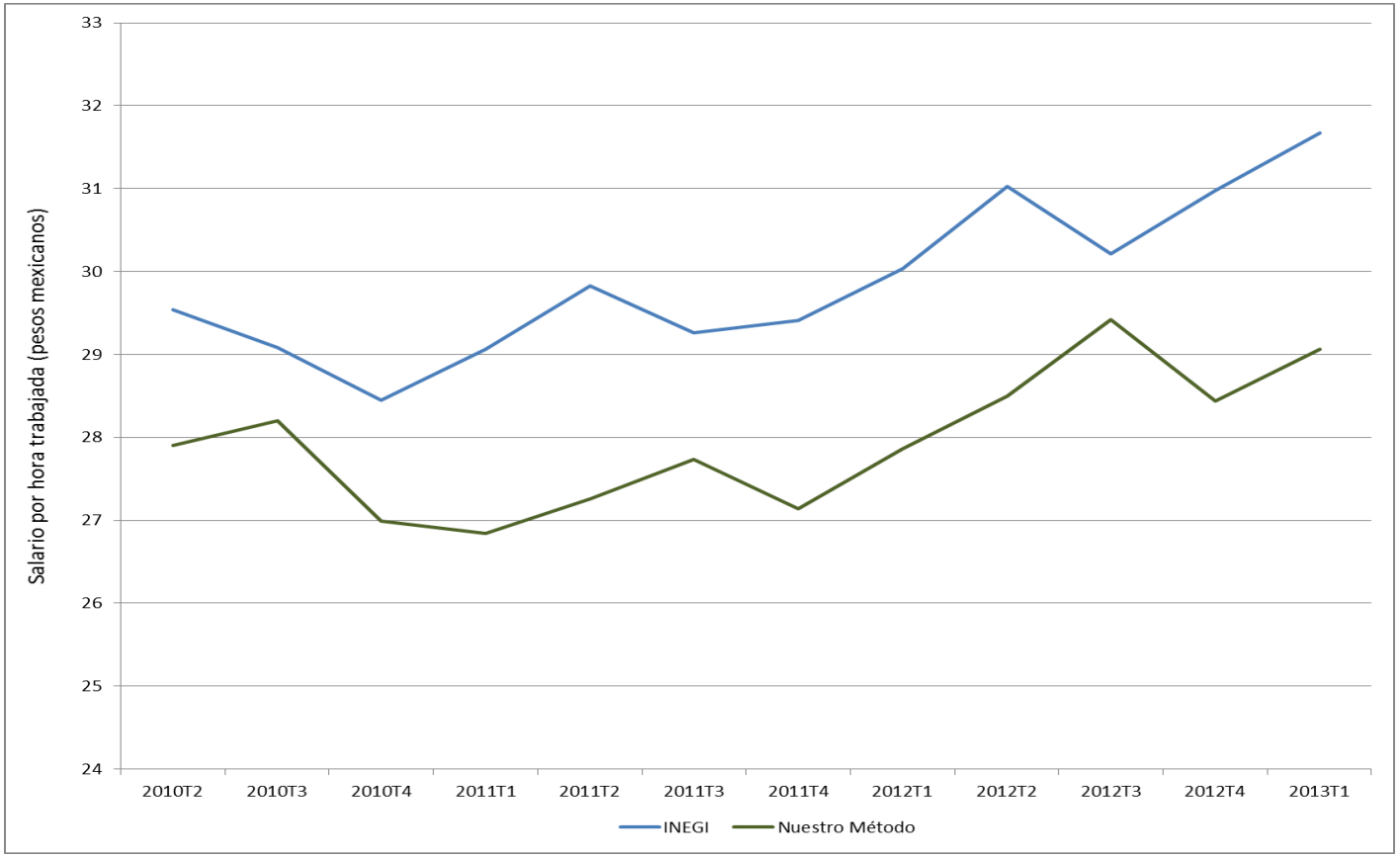
1.4 Implicaciones

Ejercicio de simulación



1.4 Implicaciones

Indicador de salario por hora trabajada (ENOE)



2. Inferencia en poblaciones finitas

2.1. Aspectos filosóficos

Inferencia inductiva: Lo 'no observado' es semejante a lo 'observado'.
(Laplace, Johnson, de Finetti, Carnap)

$$\mathcal{P} = \mathcal{S} \cup \widetilde{\mathcal{S}}$$

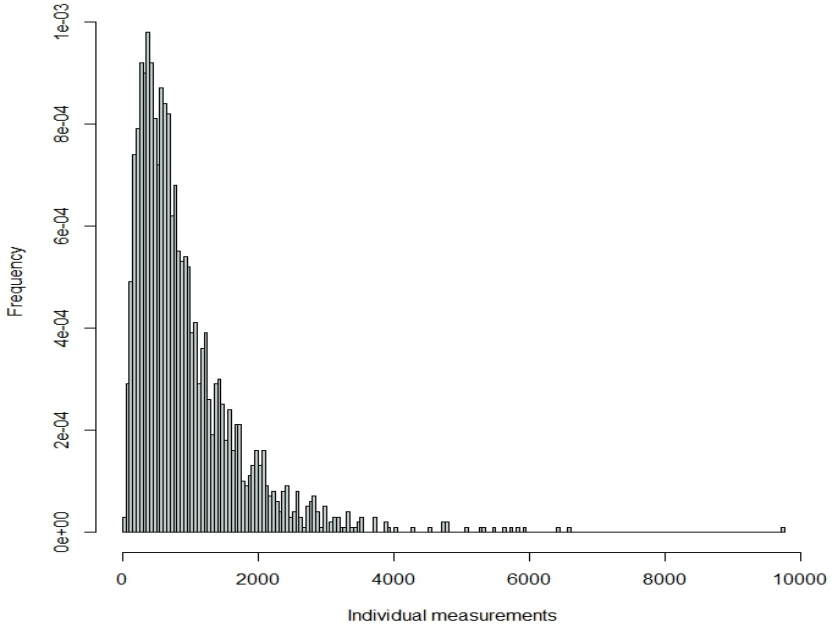
$\mathcal{S}$	$\Rightarrow$	$\widetilde{\mathcal{S}}$
$\{i_1, \ldots, i_n\}$		$\mathcal{P} - \{i_1, \ldots, i_n\}$

Inferencia estadística: Frecuentista, bayesiana, basada en diseño.
(Neyman, Fisher, Lehmann, Good, Savage, Lindley)

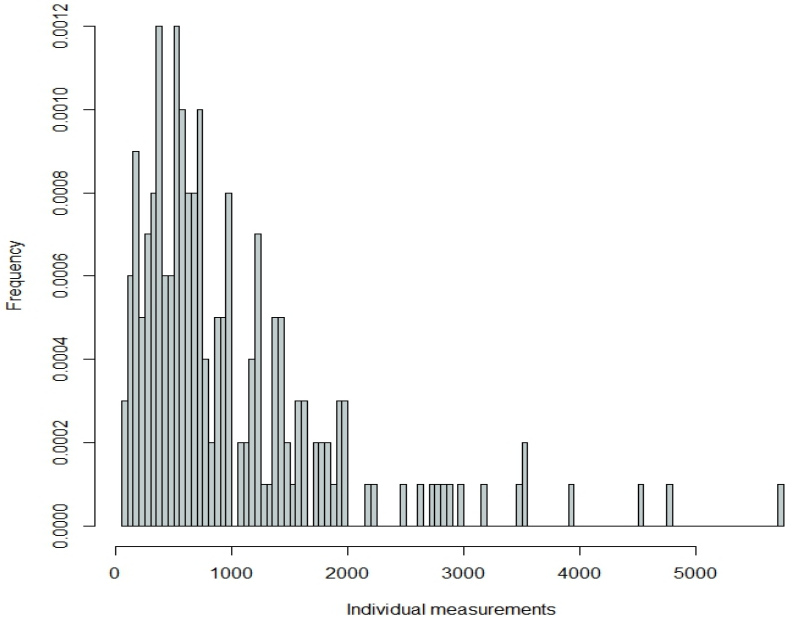
2.1. Aspectos filosóficos

Dificultad: ¿Qué hacer cuando lo 'no observado' es distinto a lo anticipado por lo 'observado'?

Población $\mathcal{P}$



Muestra $\mathcal{S}$ (10 % población)



2.1. Aspectos filosóficos

Implicaciones: El principio basado en diseños (ponderación) *cumple parcialmente* con el principio de inducción.

$$P(Y_f \in A|\mathcal{S}) \quad = \quad \sum_{k=1}^{U} \frac{m_k}{n} \cdot \delta_{y_k^*}(A),$$

donde $\boldsymbol{m} = (m_1, \ldots, m_k)$ son las frecuencias de los datos (y_k^*) observados en $\mathcal{S}$.

* Se descarta la posibilidad de observar los datos **no captados** de la muestra (ya sea por un sesgo no planeado, o por aleatoriedad).

* En muestras pequeñas es común no poder cubrir todos los aspectos relevantes de la población.

2.2. Marco muestral

Supuestos del enfoque basado en diseño:

- Existe una enumeración **unívoca** y **conocida** de la población.
- Las cantidades a medir en los individuos, (y_i), son fijas.
- Aleatoriedad en la selección de $\mathcal{S}$

Así, se define la distribución muestral

$$\pi(\mathcal{S}) \quad = \quad \mathbb{P}\{\, \texttt{seleccionar} : i_1, \ldots, i_n\},$$

para cualquier $\{i_1, \ldots, i_n\} \in \mathcal{P}$, y $\mathcal{S} = \{y_{i_1}, \ldots, y_{i_n}\}$.

* Si la enumeración existe, $\pi(\mathcal{S})$ es completamente **subjetiva.**

2.2. Marco muestral

Ponderación (factores de expansión): Si las condiciones anteriores se cumplen, el ponderador para cada y_i en $\mathcal{S}$ se define como (Kish, 1965; Cochran, 1977),

$$\begin{aligned} w_j &= 1/\pi(\mathcal{S}) \\ &= N/n, \qquad \texttt{(por ejemplo)} \end{aligned}$$

para cualquier $\{i_1, \ldots, i_n\} \in \mathcal{P}$, y $y_j \in \mathcal{S}$.

* En la práctica, en muchos casos no existe la **enumeración** poblacional. Entonces, $\pi(\mathcal{S})$ **no existe**; así ponderar carece de sentido inferencial.

* Aun si existe tal enumeración, el esquema de ponderación **no necesariamente satisface el principio de inducción.**

2.3. Tratamiento de valores extremos

No satisfacer el principio de inducción, conduce a que los estimadores ponderados sean altamente sensibles a valores extremos. (Chambers, 1986, Ghosh, 2008, Beaumont & Rivest, 2009)

- Rao (1971) $\rightarrow$ Fijar $w_j = 1$, si y_j es 'extremo', y redistribuir los pesos restantes).
- Fuller (1991) $\rightarrow \hat{T}_{y^*} = N(\alpha\bar{y}_{\mathcal{S}} + (1-\alpha)\bar{y}_*)$, donde $\bar{y}_*$ es un estimador moderado de $\bar{Y}_{\mathcal{P}}$ poblacional.

* En todos los casos, se emplea un **modelo** para determinar si $y_j \in \mathcal{S}$ es 'extremo'.

* Esto es porque da peso solamente a las observaciones en $\mathcal{S}$, siendo $\mathcal{S}$ no exhaustiva, **los valores extremos son sobre valorados.**

3. Una alternativa predictiva

Predicción: Recordemos que el propósito final es predecir qué pasa en $\widetilde{\mathcal{S}}$, usando la información contenida en la muestra $\mathcal{S}$.

Esquema basado en diseño (ponderación):

$$P(Y_f \in A|\mathcal{S}) \quad = \quad \sum_{k=1}^{U} \frac{m_k}{n} \cdot \delta_{y_k^*}(A),$$

* No existe la posibilidad de observar algo distinto a lo ya observado en $\mathcal{S}$, (i.e., no es flexible).

3.1. Species-sampling models: Derivación bayesiana no paramétrica

Y_f no es observado aun, se supone *desconocido y aleatorio*. Suponemos,

$$\begin{aligned} Y_f|F &= F(dy), \\ F &\sim \Pi(dF) \end{aligned}$$

F - medida de probabilidad aleatoria (i.e. realización de un proceso estocástico en los valores de Y_f),

$\Pi(dF)$ - ley de probabilidad de las trayectorias de F (no es absolutamente continua).

$\Pi(F|\mathcal{S})$: **cálculo muy particular (e.g. análisis combinatorio estocástico).**

Propósito final, **predicción**:

$$P(Y_f \in A|\mathcal{S}) = \int_A F(dy)\Pi(dF|\mathcal{S}).$$

3.1. Species-sampling models (poblaciones finitas) (Nuestra propuesta)

- Si F y $\Pi(dF)$ son un *proceso Dirichlet* (Ferguson, 1973; Kingman, 1977; Pitman, 1996)

Nuestra versión predictiva, para poblaciones finitas:

$$P(y_f \in A|\mathcal{S}) \quad = \quad \sum_{k=1}^{U} \frac{m_k}{\alpha^{DP}_{(N,n)} + n} \cdot \delta_{y_k^*}(A) + \frac{\alpha^{DP}_{(N,n)}}{\alpha^{DP}_{(N,n)} + n} \cdot G_0(A),$$

$\alpha^{DP}_{(N,n)}$ modula la precisión de la predicción,

G_0 modula la posibilidad de observar algo distinto a lo ya observado.

+ Esta alternativa permite corregir por sesgos inducidos por valores extremos. (JCM, SO & AR, 2013b)

3.2. Licitación inicial

Parámetros: α_{DP} y G_0.

- Acelerador de la convergencia (JCM, SO y AR, 2013c):

$$\alpha^{DP}_{(N,n)} \to 0, \text{ cuando } n \to N.$$

- También puede definirse en función de la dispersión en G_0. (Teorema en proceso)

*** En la práctica, puede definirse a través de un estudio de sensibilidad.**

3.2. Licitación inicial Parámetros: α_{DP} y G_0.

Caso 1. Púramente subjetivo (e.g. opiniones de expertos)

Caso 2. Objetivamente, empleando datos previos o de referencia.

Caso 2. (JCM, SO, AR, 2013c)

- Usar datos $\mathcal{S}_{\text{ref}}$.
- Proponer $\mathcal{G} = \{G_1, , G_K\}$, colección de alternativas.
- Elegir G_0, tal que

$$G_0(\cdot|\mathcal{S}_{\text{ref}}) \quad = \quad \text{argmax}_{g_j \in \mathcal{G}} \left\{ \sum_{l=1}^{M} \log d_l^{(j)} \right\}.$$

$d_l^{(j)}$ es una medida de ajuste predictivo de validación cruzada Kullback-Leibler (Gelfand et al, 1992; Gelman et al, 1996).

3.3. Total de la población (Nuestra propuesta)

Total:

$$\underbrace{T}_{\texttt{Incierto}} = \underbrace{T^{\mathcal{S}}}_{\texttt{Observado}} + \underbrace{T^{\widetilde{\mathcal{S}}}}_{\texttt{Incierto}},$$

Predicción:

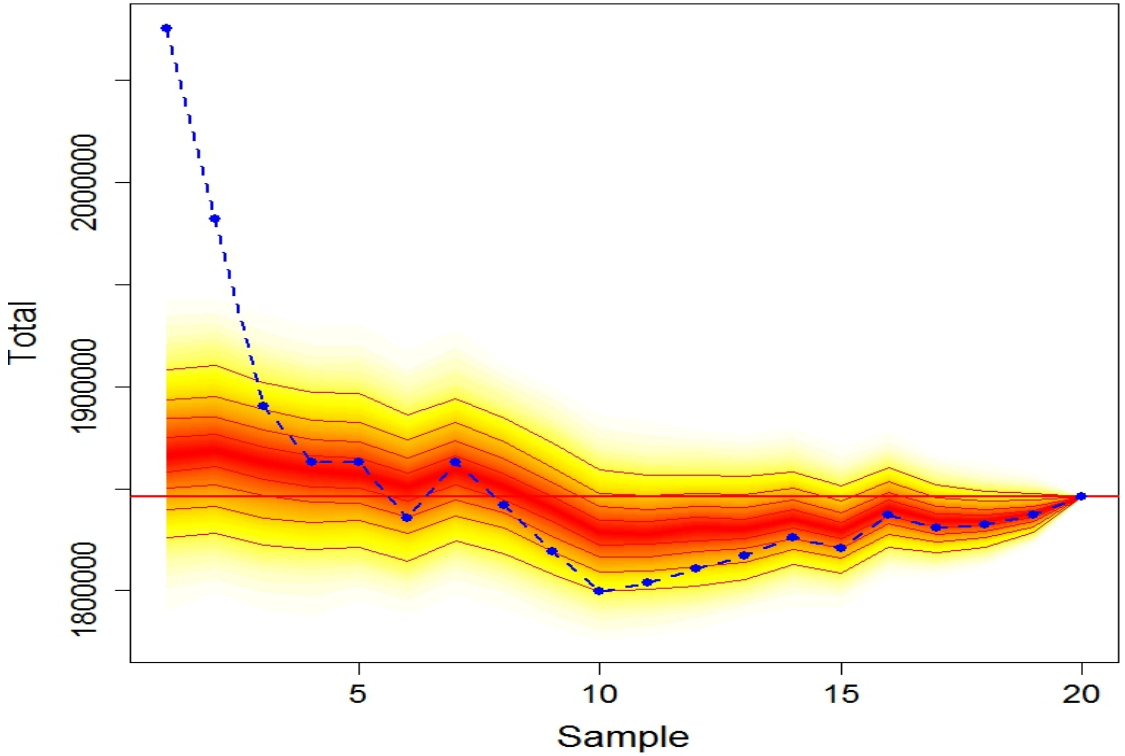
$$\Pr(T \leq t|\mathcal{S}) = P^{*N^{\widetilde{\mathcal{S}}}}\left(T^{\widetilde{\mathcal{S}}} \leq t - T^{\mathcal{S}}\right),$$

Estimador puntual:

$$\widehat{T} = \sum_{g\in\mathcal{S}} y_g + N^{\widetilde{\mathcal{S}}} \cdot \left[\sum_{k=1}^{U} \frac{m_k}{\alpha^{DP}_{(N,n)} + n} \cdot y^*_k + \frac{\alpha^{DP}_{(N,n)}}{\alpha^{DP}_{(N,n)} + n} \cdot \mu_0\right], \text{ con } \mu_0 = \mathbb{E}_{G_0}(Y_f).$$

* Incluye como caso particular al estimador ponderado tradicional, cuando $\alpha^{DP}_{(N,n)} = 0$, (JCM, SO & AR, 2013a,b)

3.3. Total de la población (Nuestra propuesta)



* Población simulada. Línea punteada muestra la trayectoria del estimador basado en ponderación.

3.3. Total de la población (dominios no planeados) (Nuestra propuesta)

Definición: Partición $(\mathcal{D}_d)_{d=1}^{D}$ de $\mathcal{P}$ que **NO** coincide con dominios planeados.

Totales:

$$\underbrace{T}_{\texttt{Incierto}} = \underbrace{T^{\mathcal{S}}}_{\texttt{Observado}} + \underbrace{T^{\widetilde{\mathcal{S}}}}_{\texttt{Incierto}},$$

Desagregación:

$$\underbrace{T^{\mathcal{S}}}_{\texttt{Observado}} = \underbrace{T^{\mathcal{S}\cap\mathcal{D}_1} + \cdots + T^{\mathcal{S}\cap\mathcal{D}_d}}_{\texttt{Observado}},$$

$$\underbrace{T^{\widetilde{\mathcal{S}}}}_{\texttt{Incierto}} = \underbrace{T^{\widetilde{\mathcal{S}}\cap\mathcal{D}_1} + \cdots + T^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d}}_{\texttt{Incierto}}.$$

3.4. Total de la población (dominios no planeados) (Nuestra propuesta)

- A la incertidumbre de los totales, le **subyace** la incertidumbre sobre la composión de los dominios no planeados. (Meeden, 2005, Lehtonen & Veijanen, 2009)

Conteos:

$$\underbrace{N}_{\texttt{Observado}} = \underbrace{N^{\mathcal{S}}}_{\texttt{Observado}} + \underbrace{N^{\widetilde{\mathcal{S}}}}_{\texttt{Observado}},$$

Desagregación:

$$\underbrace{N^{\mathcal{S}}}_{\texttt{Observado}} = \underbrace{N^{\mathcal{S}\cap\mathcal{D}_1} + \cdots + N^{\mathcal{S}\cap\mathcal{D}_d}}_{\texttt{Observado}},$$

$$\underbrace{N^{\widetilde{\mathcal{S}}}}_{\texttt{Observado}} = \underbrace{N^{\widetilde{\mathcal{S}}\cap\mathcal{D}_1} + \cdots + N^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d}}_{\texttt{Incierto}}.$$

3.4. Total de la población (dominios no planeados) (Nuestra propuesta)

Predicción:

$$\Pr\{\boldsymbol{T}^{\mathcal{D}}, \boldsymbol{N}^{\mathcal{D}} | \mathcal{S}\} \quad = \quad \Pr\{\boldsymbol{T}^{\mathcal{D}} | \boldsymbol{N}^{\mathcal{D}}, \mathcal{S}\} \times \Pr\{\boldsymbol{N}^{\mathcal{D}} | \mathcal{S}\}.$$

Conteos

$$\Pr\{\boldsymbol{N}^{\mathcal{D}} \leq \boldsymbol{n} | \mathcal{S}\} \quad = \quad \texttt{Mult} - \texttt{Dir}(\boldsymbol{N}^{\widetilde{\mathcal{S}} \cap \mathcal{D}} \leq \boldsymbol{n} - \boldsymbol{N}^{\mathcal{S} \cap \mathcal{D}} | N^{\widetilde{\mathcal{S}}}, \boldsymbol{\alpha}(N, \mathcal{S})),$$

$\boldsymbol{\alpha}(N, \mathcal{S})$ - vector de frecuencias esperadas entre dominios no planeados.

$\boldsymbol{\alpha}(N, \mathcal{S}) \to \mathbf{0}$, cuando $n \to N$.

3.4. Total de la población (dominios no planeados) (Nuestra propuesta)

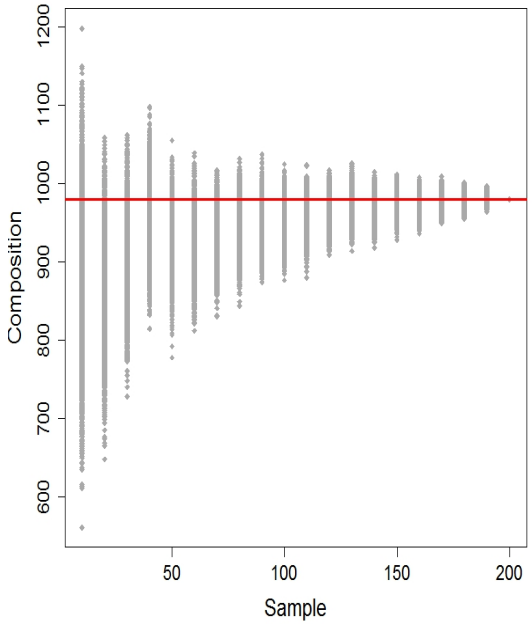
Predicción de los totales parciales:

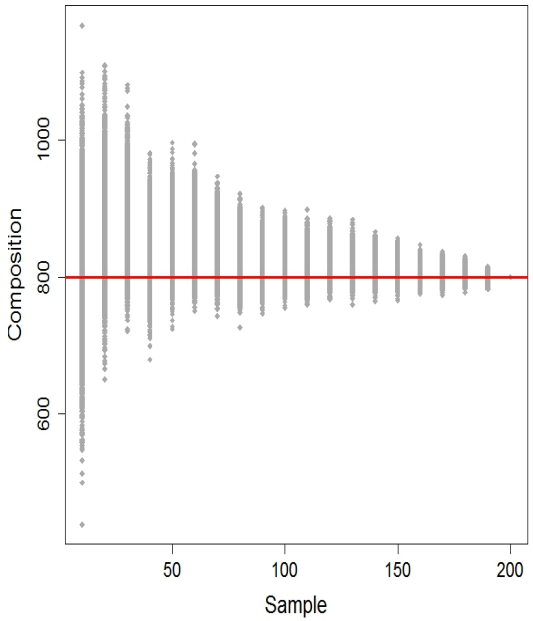
$$\Pr\{\boldsymbol{T}^{\mathcal{D}} \leq \boldsymbol{t} | \boldsymbol{N}^{\widetilde{\mathcal{S}}\cap\mathcal{D}}, \mathcal{S}\} = \\ = \left(P^{*N^{\widetilde{\mathcal{S}}\cap\mathcal{D}_1}} \left(T^{\widetilde{\mathcal{S}}\cap\mathcal{D}_1} \leq t_1 - T^{\mathcal{S}\cap\mathcal{D}_1} \right), \right. \\ , \ldots, \\ \left. , P^{*N^{\widetilde{\mathcal{S}}\cap\mathcal{D}_D}} \left(T^{\widetilde{\mathcal{S}}\cap\mathcal{D}_D} \leq t_D - T^{\mathcal{S}\cap\mathcal{D}_D} \right), \right.$$

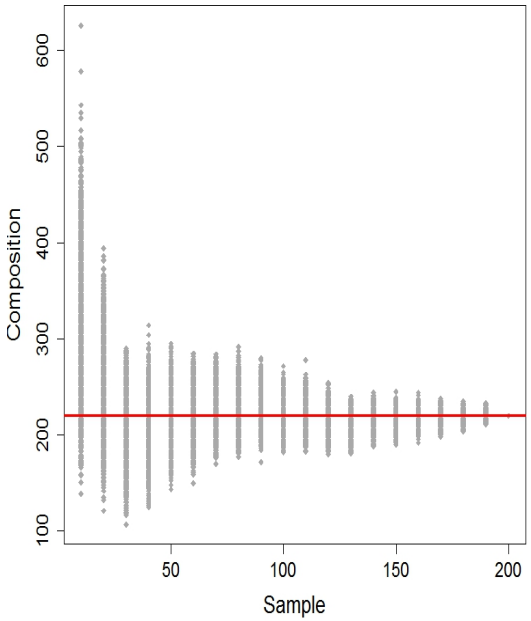
+ Este objeto es esencialmente un vector de convoluciones D-dimensional, condicional en $\boldsymbol{N}^{\widetilde{\mathcal{S}}\cap\mathcal{D}}$ que es desconocido.

3.4. Total de la población (dominios no planeados) (Nuestra propuesta)

Composición



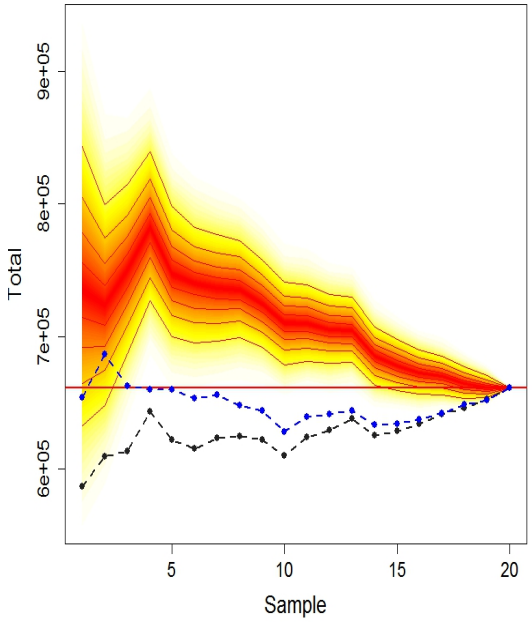


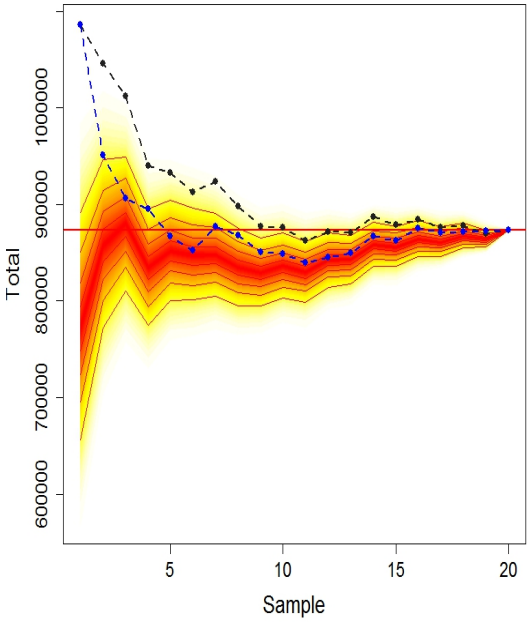


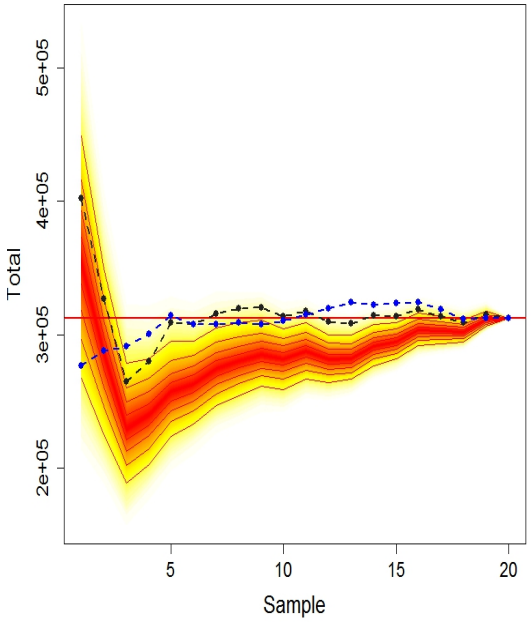
* Población simulada. Línea punteada muestra la trayectoria del estimador basado en ponderación.

3.4. Total de la población (dominios no planeados) (Nuestra propuesta)

Sub-totales







* Población simulada. Las líneas punteadas muestra la trayectoria de dos estimadores, uno basado en ponderación y otro 'naive'.

3.5. `predfinitepop` (Librería para R)

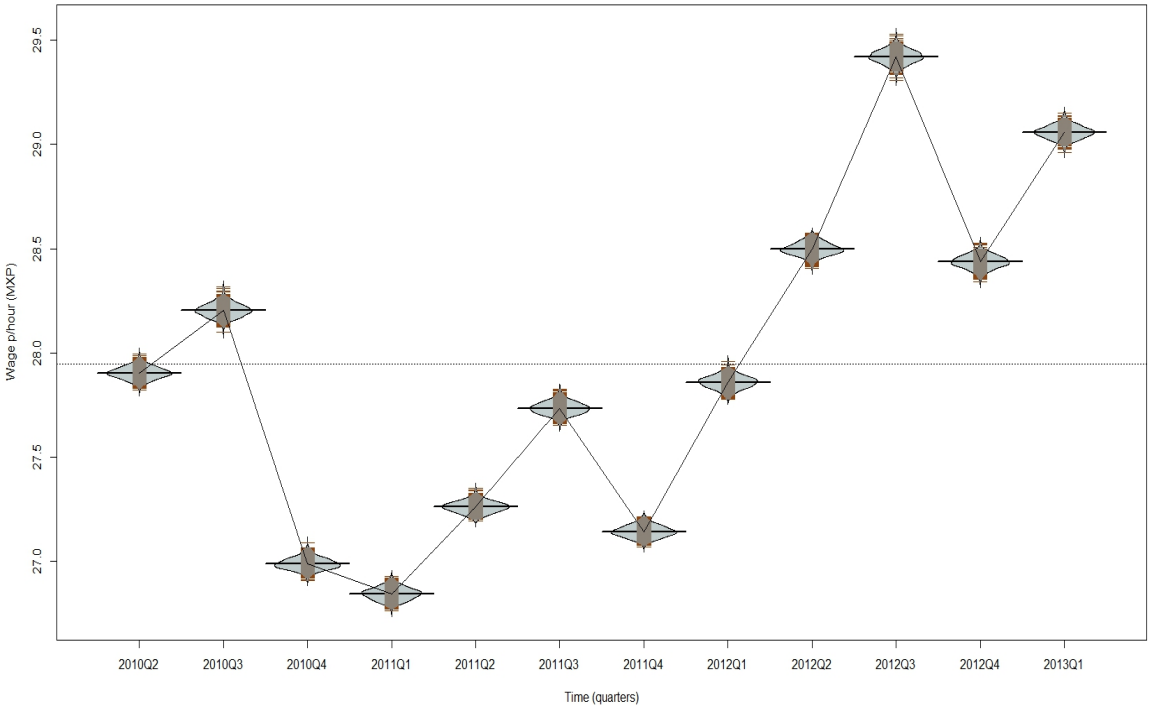
Genera simulaciones de Monte Carlo de las distribuciones predictivas descritas arriba.

Funciones principales:

- `domplan.R` - Predicciones en dominios planeados, calibrando G_0 con datos de referencia.
- `domplan_g0.R` - Predicciones en dominios planeados, con una distribución G_0 dada.
- `domnoplan.R` - Predicciones en dominios planeados y no planeados, calibrando G_0 con datos de referencia.
- `domnoplan_g0.R` - Predicciones en dominios planeados y no planeados, con una distribución G_0 dada.
- `g0_licitacion.R` - Licitación inicial de G_0 usando datos $\mathcal{S}_{\text{ref}}$.

4. Indicadores salariales

Salario p/ hora de la población ocupada



Fuente: ENOE, 2010T1 a 2013T2.

5. Discusión

Hemos desarrollado una metodología nueva para inferencia en poblaciones finitas:

- Se adapta a esquemas de muestreo no informativos bajo estratificación.
- Es robusta ante valores extremos.
- Permite incorporar coherentemente información relevante.
- Incorpora la incertidumbre originada por los dominios planeados.
- Produce estimaciones consistentes entre dominios planeados y no planeados.
- Produce predicciones por densidades.
- Es implementable, y está desarrollada **'en casa'.**

* La inferencia en **dominios no planeados** es un tema no resuelto aún. El problema es la falta de información.

Temas de investigación

- Cálculo del tamaño de la muestra
- Rapidez de convergencia
- Marco inferencial para un muestreo informativo-multietápico
- Robustez
- Extensiones de la libreria para `R`

Referencias

- Chambers (1986) "A Unified Approach to Robust Estimation in Finite Population Sampling," JASA.
- Beaumont y Rivest (2009) "Dealing with outliers in survey data," Sample Surveys: Design, Methods and Applications.
- Beaumont et al (2013), "A new approach to weighting and inference in sample surveys," Biometrika.
- Fuller (1991) "Simple estimators for the mean of skewed populations," Statistica Sinica.
- Hájek (1971). Comment on: "An essay on the logical foundations of survey sampling, part one." The Foundations of Survey Sampling.
- Horvitz y Thompson (1952) "A generalization of sampling without replacement from a finite universe," JASA.
- Pitman (1996) "Some developments of the Blackwell-MacQueen urn scheme." Statistics, Probability and Game Theory.

Contribuciones

- "A Bayesian nonparametric framework to inference on totals of finite populations," JCM, SO & AR (2013a).
- "Predictive inference on finite populations segmented in planned and unplanned domains," JCM, SO & AR (2013b).
- "Robust model-based inference of Mexico's currency flows related to expenditures of international travelers," JCM, SO & AR (2013c).
- `predfinitepop:` Una librería para `R-project`.

Gracias por su atención.

juan.martinez@banxico.org.mx