Construcción de los factores de expansión

Andrés Gutiérrez

Asesor Regional en Estadísticas Sociales - Comisión Económica para América Latina y el Caribe (CEPAL) - andres.gutierrez@cepal.org

Tabla de contenidos I

Construcción de los Factores de Expansión

Introducción

- Los factores de expansión son fundamentales en el análisis de encuestas de hogares, garantizando estimaciones precisas y representativas de la población objetivo.
- Los procesos de inferencia estadística se basan en el principio de representatividad, donde cada unidad muestral representa a sí misma y a otras similares.
- Los procesos de inferencia estadística en encuestas descansan en las probabilidades de inclusión generadas por el diseño de muestreo.
- La probabilidad de inclusión es un número real contenido en el intervalo $(0, 1]$, entonces su inverso multiplicativo también será un número real mayor o igual que uno.

$$d_k = \frac{1}{Pr(k \in s)}$$

Introducción

- Los diseños de muestreo auto-ponderados aseguran que las unidades finales de muestreo tengan la misma probabilidad de inclusión, independientemente del tamaño de la unidad primaria de muestreo.
- Los pesos de muestreo se utilizan para varios propósitos, como incorporar probabilidades de selección, ajustar casos de elegibilidad incierta, corregir sesgos por ausencia de respuesta, y mejorar las estimaciones mediante información auxiliar externa.
- La construcción de los pesos de muestreo representa un desafío metodológico debido a la expansión urbana y la desactualización de los marcos de muestreo en las áreas geográficas.

Introducción

- Los equipos de los Institutos Nacionales de Estadística deben ser flexibles y adaptarse a la movilidad de las poblaciones, especialmente en áreas urbanas, para garantizar la calidad de las estimaciones en las encuestas de hogares.
- En las encuestas a hogares, no existe una lista exhaustiva de todos los hogares de un país, por lo que se construye el marco de muestreo en varias etapas, seleccionando áreas geográficas, realizando un censo exhaustivo de los hogares en esas áreas y luego seleccionando hogares de la muestra.

Introducción

Para hacerle frente a las imperfeccciones del marco, Valliant y Dever (2017) recomienda:

1. ER (*elegible respondents*): Unidades elegibles que respondieron efectivamente, indicando casos elegibles para los cuales se recolectó suficiente información.
2. ENR (*elegible nonrespondents*): Unidades elegibles que no respondieron, indicando casos elegibles para los cuales no se recolectó información o se recolectó parcialmente.
3. IN (*ineligibles*): Unidades no elegibles que no forman parte de la población de interés.
4. UNK (*unknown elegibility*): Unidades con elegibilidad desconocida, indicando casos donde no se puede determinar si la unidad es elegible o no.

Construcción de los Factores de Expansión

Para construir los factores de expansión de una encuesta se recomienda seguir en este orden los siguientes procesos:

1. Creación de los pesos básicos.
2. Ajuste por elegibilidad desconocida.
3. Descarte de las unidades no elegibles.
4. Ajuste por ausencia de respuesta.
5. Calibración por proyecciones poblacionales y variables auxiliares.
6. Recorte y redondeo de los factores finales (*opcional*).

Creación de los pesos básicos

- Cada esquema de muestreo tiene asociada una única función que vincula a cada elemento con una probabilidad de inclusión en la muestra, representada como

$$\pi_k = Pr(k \in s)$$

.

- El primer paso en la reponderación de los pesos de muestreo es la creación de los pesos básicos d_{1k}, definidos como el inverso multiplicativo de la probabilidad de inclusión:

$$d_{1k} = \frac{1}{\pi_k}$$

para todo $k \in s$.

Ajuste por elegibilidad desconocida

- El segundo paso implica redistribuir el peso de las unidades cuyo estado de elegibilidad es desconocido en la muestra.
- Se redistribuyen los pesos de las unidades con elegibilidad desconocida (UNK) entre las unidades que sí disponen de su estatus de elegibilidad (ER, ENR, IN).
- Si no es posible determinar la elegibilidad de algunas unidades en el marco de muestreo, se tendrá una muestra que incluye subconjuntos de unidades elegibles respondientes (s_{ER}), elegibles no respondientes (s_{ENR}), no elegibles (s_{IN}), y unidades con elegibilidad desconocida (s_{UKN}).

Ajuste por elegibilidad desconocida

- Se recomienda formar B categorías basadas en la información del marco de muestreo, siendo cada categoría una agrupación de al menos 50 casos, según (Valliant y Dever 2017).
- Estas categorías pueden ser estratos o cruces de subpoblaciones.
- Para cada categoría b, se define el factor de ajuste por elegibilidad (a_b)

$$a_b = \frac{\sum_{s_b} d_{1k}}{\sum_{s_b \cap (s_{ER} \cup s_{ENR} \cup s_{IN})} d_{1k}}$$

- Los pesos ajustados por elegibilidad desconocida (d_{2k}) para las unidades cuya elegibilidad pudo ser establecida estarán dados por la siguiente expresión:

$$d_{2k} = a_b * d_{1k} \qquad \forall\ k \in s_b \cap (s_{ER} \cup s_{ENR} \cup s_{IN})$$

Descarte de las unidades no elegibles

- En esta etapa, se descartan tanto las viviendas con elegibilidad desconocida (UNK) como aquellas viviendas que han cambiado su estado de ocupación y ahora no contienen ningún hogar particular (IN) de la población objetivo.
- El tercer paso consiste en ajustar el peso de la etapa anterior de la siguiente manera:

$$d_{3k} = \begin{cases} 0, & \text{si la unidad } k \in (s_{UNK} \cup s_{IN}) \\ d_{2k}, & \text{si la unidad } k \in (s_{ER} \cup s_{ENR}). \end{cases}$$

Ajuste por ausencia de respuesta

- Se consideran dos variables aleatorias para el manejo de la ausencia de respuesta: I_k y D_k.
- I_k es una variable indicadora que toma el valor de 1 si la unidad k es un respondiente efectivo (ER) o un elegible no respondiente (ENR), y 0 en caso contrario.

$$I_k = \begin{cases} 1, & \text{si } k \in (s_{ER} \cup s_{ENR}) \\ 0, & \text{en otro caso.} \end{cases}$$

Ajuste por ausencia de respuesta

- D_k es una variable indicadora que toma el valor de 1 si la unidad k es un ER, y 0 si es un ENR.

$$D_k = \begin{cases} 1, & \text{si } k \in s_{ER} \\ 0, & \text{si } k \in s_{ENR}. \end{cases}$$

- Estas variables ayudan a distinguir entre las unidades que respondieron efectivamente y las que no respondieron en la encuesta.

Ajuste por ausencia de respuesta

- Se define la probabilidad de respuesta (ϕ_k), también conocida como *propensity score*, como la probabilidad de que una unidad k sea un respondiente efectivo (ER) dado que es un ER o un elegible no respondiente (ENR).

$$Pr[k \in s_{ER} | k \in (s_{ER} \cup s_{ENR})] = Pr[D_k = 1 | I_k = 1] = \phi_k$$

- Si el patrón de ausencia de respuesta es completamente aleatorio o aleatorio (en donde el patrón de la no respuesta puede ser explicado por un conjunto de covariables $\mathbf{z}$), entonces

$$\phi_k = f(\mathbf{z}_k, \beta) \qquad \forall\ k \in (s_{ER} \cup s_{ENR})$$

Ajuste por ausencia de respuesta

- Si fuese plausible tener acceso a las covariables $\mathbf{z}$ para los individuos elegibles en la muestra, entonces:

$$\hat{\phi}_k = f(\mathbf{z}_k, \hat{\beta}) \qquad \forall\ k \in (s_{ER} \cup s_{ENR})$$

- Si el patrón de ausencia de respuesta es no aleatorio, entonces

$$\phi_k = f(\mathbf{y}_k, \beta) \qquad \forall\ k \in (s_{ER} \cup s_{ENR})$$

Ajuste por ausencia de respuesta

- Dado que no se tiene acceso a las variables de interés para todos los individuos en la muestra de unidades elegibles debido a la falta de respuesta, no es posible estimar el patrón de ausencia de respuesta, lo que puede conducir a problemas de sesgo.
- Se puede abordar este problema utilizando un modelo basado en la estimación de las probabilidades de respuesta, también conocido como *propensity score*.
- Siendo la probabilidad de que un individuo responda $\phi_k = Pr(k \in s_{ER})$, se puede estimar utilizando un modelo de regresión logística con un vector de información auxiliar $\mathbf{z}_k$ conocido para todos los $k \in (s_{ER} \cup s_{ENR})$.

Ajuste por ausencia de respuesta

La estimación de $\widehat{\phi}_k$ se realiza mediante un modelo de regresión logística, expresado como:

$$\widehat{\phi}_k = \frac{\exp\{\mathbf{z}_k'\widehat{\beta}\}}{1+\exp\{\mathbf{z}_k'\widehat{\beta}\}} \qquad \forall\ k \in (s_{ER} \cup s_{ENR})$$

donde $\widehat{\beta}$ es el vector de coeficientes estimado de la regresión logística.

Ajuste por ausencia de respuesta

Si la ausencia de respuesta no depende de la variable de interés, es posible definir el siguiente estimador insesgado

$$\hat{t}_y = \sum_{k \in s_{ER}} d_{4k} y_k$$

En donde

$$d_{4k} = \frac{d_{3k}}{\hat{\phi}_k} \qquad \forall\ k \in s_{ER}$$

Ajuste por ausencia de respuesta

- Para aumentar la eficiencia del estimador, se pueden crear categorías homogéneas de individuos que tengan la misma probabilidad de responder, utilizando los valores de covariables disponibles.
- Es crucial obtener un conjunto de covariables que esté disponible tanto para respondientes como para no respondientes.
- Por ejemplo, si se consideran las variables de edad (5 categorías) y sexo (2 categorías), se pueden formar $Q = 10$ categorías según el cruce de estas variables para obtener una estimación de la probabilidad de respuesta en cada clasificación y ajustar el peso de muestreo.

Ajuste por ausencia de respuesta

- La probabilidad de respuesta en la categoría q se estima como:

$$\phi_q = \frac{\sum_{s_{ER} \cap s_q} d_{3k}}{\sum_{s_q} d_{3k}}$$

El nuevo peso ajustado por la ausencia de respuesta se calcula como:

$$d_{4k} = \frac{d_{3k}}{\phi_q} = d_{3k} \frac{\sum_{s_q} d_{3k}}{\sum_{s_{ER} \cap s_q} d_{3k}}$$

Ajuste por ausencia de respuesta

- En un escenario complejo, si las probabilidades de respuesta se estimaron con un modelo de *propensity score*, es posible crear clases de individuos con probabilidades similares de responder.
- Se asume que las unidades dentro de una misma clase tienen la misma configuración de covariables o al menos una probabilidad de respuesta estimada similar $\widehat{\phi}_k$.
- El objetivo es asegurar que cualquier diferencia en las covariables pueda ser ajustada.

Ajuste por ausencia de respuesta

- Se pueden crear C clases y s_c es la muestra de n_c unidades elegibles en la clase c $(c = 1, 2, ..., C)$.
- Se pueden utilizar diversas medidas para ajustar las clases (Valliant y Dever 2017):

Promedio no ponderado:

$$\hat{\phi}_c = \frac{\sum_{k \in s_c} \hat{\phi}_k}{n_c}$$

Promedio ponderado:

$$\hat{\phi}_c = \frac{\sum_{k \in s_c} d_{3k} \hat{\phi}_k}{n_c}$$

Mediana no ponderada:

$$\hat{\phi}_c = mediana[\hat{\phi}_k]$$

Ajuste por ausencia de respuesta

Tasa de respuesta no ponderada:

$$\hat{\phi}_c = \frac{\#(s_{ER} \cap s_c)}{n_c}$$

Tasa estimada de respuesta:

$$\hat{\phi}_c = \frac{\sum_{s_c \cap s_{ER}} d_{3k}}{\sum_{s_c} d_{3k}}$$

- La elección de la medida depende de la similitud en las probabilidades de respuesta dentro de cada clase y la variabilidad de los pesos de muestreo.

Calibración de los factores de expansión

- Después de establecer los pesos de muestreo en la encuesta, se ajustan utilizando la información auxiliar disponible a nivel nacional o por estratos específicos.
- El objetivo es minimizar el sesgo y los errores estándar, especialmente en estudios con ausencia de respuesta.
- Esto se logra calibrando los pesos para que la suma ponderada de las características de la muestra sea igual a los totales poblacionales conocidos o estimados.

$$\sum_{k \in s} w_k \mathbf{x}_k = \mathbf{t}_{\mathbf{X}}$$

Calibración de los factores de expansión

- La calibración de pesos busca que éstos reproduzcan el tamaño conocido de las regiones o del país.
- Se utiliza la metodología de calibración para garantizar que la suma ponderada de los pesos para cada estrato sea igual al tamaño conocido de ese estrato:

$$\sum_{s_h} w_k = N_h$$

- Una vez ejecutado el proceso de calibración, los nuevos pesos se expresan en términos de los pesos originales y un factor de calibración g_k aplicado a cada unidad de la muestra.

$$w_k = g_k * d_{4k} \qquad \forall\ k \in s$$

Calibración de los factores de expansión

- Los valores g_k en la calibración de pesos dependen de la muestra seleccionada s y de la función de optimización utilizada.
- Los valores g_k no tienen una forma cerrada general, pero pueden tomar valores específicos según la información auxiliar disponible.
- Por ejemplo, bajo la distancia Ji-cuadrado, el estimador de calibración se expresa como:

$$\hat{t}_{y,cal} = \sum_{k \in s} w_k y_k = \hat{t}_{y,\pi} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x},\pi})\hat{\mathbf{B}}_s$$

- Donde $\hat{\mathbf{B}}_s$ es un vector de coeficientes de regresión que depende de la muestra s y de constantes q_k.

Calibración de los factores de expansión

- En este caso, los ponderadores g_k se calculan como:

$$g_k = 1 + (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x},\pi}) \left(\sum_s w_k q_k \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \sum_s w_k q_k \mathbf{x}_k$$

- Los estimadores de calibración son **aproximadamente insesgados**, pero el sesgo puede ser cuantificado mediante la expresión:

$$Bias(\hat{t}_{y,cal}) = E \left[\sum_{k \in s} (w_k - d_k) y_k \right]$$

Medidas de calidad en la calibración

- Las medidas de calidad en la calibración, según Silva (2004), se centran en proteger contra el sesgo generado por restricciones excesivas y garantizar la precisión de las variables utilizadas para calibrar.
- La teoría sugiere que más variables de calibración reducen la varianza de las estimaciones, pero demasiadas restricciones pueden causar problemas computacionales y aumentar la varianza y el sesgo.
- Se recomienda un análisis para determinar el número óptimo de variables de calibración.

Medidas de calidad en la calibración

- La interpretación de los pesos de calibración puede volverse difícil si son menores que uno, lo que puede llevar a estimaciones negativas en algunos dominios, especialmente cuando el tamaño de muestra es pequeño.
- Para garantizar que los pesos estén dentro de un intervalo específico, se deben minimizar las distancias y satisfacer las ecuaciones de calibración, aunque puede no haber una solución exacta y el algoritmo de calibración puede no converger.

Medidas de calidad en la calibración

- Es importante analizar los pesos g_k en cada dominio, estrato y postestrato, identificando aquellos que son potencialmente grandes o influyentes.
- Se recomienda postestratificar la muestra y aplicar la calibración solo a las unidades con pesos estables.
- Los pesos de calibración pueden estar restringidos a un espacio predefinido por el usuario mediante límites (L, U). Por ejemplo, $L = 1$ y U se calcula como $Q_3 + 1.5 * (Q_3 - Q_1)$, donde Q_3 y Q_1 son el tercer y primer cuartil de la distribución de g_k.

Medidas de calidad en la calibración

- La ausencia de respuesta puede generar sesgo en las estimaciones finales si el mecanismo subyacente no es aleatorio (MAR) o completamente aleatorio (MCAR).
- En presencia de ausencia de respuesta, es más probable que aparezcan pesos de calibración negativos y que estos no converjan a los pesos originales.
- La varianza de los estimadores de calibración puede diferir de la de los estimadores de regresión convencionales.

Medidas de calidad en la calibración

Para evaluar los escenarios de calibración, se pueden utilizar medidas como el error relativo promedio (M1) y el coeficiente de variación relativo promedio (M2).

$$M1 = \frac{1}{P}\sum_{p=1}^{P}\frac{|\hat{t}_{x_p,cal} - t_{x_p}|}{t_{x_p}}$$

$$M2 = \frac{1}{P}\sum_{p=1}^{P}\frac{\sqrt{Var(\hat{t}_{x,\pi})}}{t_x}$$

Medidas de calidad en la calibración

- La proporción de ponderadores extremos menores a un límite inferior (L) predefinido se calcula como

$$M3 = \frac{1}{n} \sum_{k \in s} I(g_k < L)$$

- De manera similar, la proporción de ponderadores extremos mayores a un límite superior (U) predefinido se calcula como

$$M4 = \frac{1}{n} \sum_{k \in s} I(g_k > U)$$

- Valores muy alejados de la unidad, puede ser señal de sesgo en las estimaciones.

Medidas de calidad en la calibración

- El coeficiente de variación de los ponderadores se calcula como

$$M5 = \frac{\sigma(g)}{\bar{g}}$$

  Siendo $\sigma(g)$ la desviación estandar muestral de los ponderadores g_k y $\bar{g}$ su promedio muestral.

- Una alta dispersión de los ponderadores puede ser indeseable, ya que podría indicar la presencia de valores influyentes que alejarían a los pesos calibrados de los pesos muestrales.

Medidas de calidad en la calibración

- Otra medida relevante es la distancia Ji-cuadrado entre los pesos de calibración y los pesos originales, dada por

$$M6 = \frac{1}{n} \sum_{k \in s} \frac{(w_k - d_k)^2}{d_k}$$

.

- Una distancia Ji-cuadrado pequeña indicaría una cercanía entre los pesos calibrados y los pesos originales, lo que señalaría que el insesgamiento se mantiene.

Medidas de calidad en la calibración

- La eficiencia de los estimadores de calibración puede evaluarse mediante

$$M7 = \frac{1}{P}\sum_{p=1}^{P}\frac{Var(\hat{t}_{x_p,cal})}{Var(\hat{t}_{x_p,\pi})}$$

.

- Se espera que esta medida sea pequeña y siempre menor a uno, indicando que la inferencia estadística con los pesos calibrados es mayor que con el estimador HT.

Medidas de calidad en la calibración

- El efecto de diseño debido a la ponderación desigual $DEFF^w$ se define como

$$M8 = 1 + cv^2(w_k)$$

.

- Se busca que esta medida sea cercana a uno, lo que indica que la dispersión de los pesos finales está controlada.

Calibración integrada para hogares y personas

- La calibración puede realizarse a nivel de personas o de hogares, cada opción con ventajas y consideraciones específicas.
- Calibrar a nivel de hogares implica que todos los miembros del mismo hogar tendrán el mismo peso de muestreo, independientemente de sus características individuales.
- Calibrar a nivel de personas permite que los pesos de muestreo de los hogares también pueden verse alterados, y que los pesos finales de muestreo de las personas sean diferentes dentro de los hogares.

Calibración integrada para hogares y personas

- Calibrar a nivel de personas puede introducir factores de expansión diferentes para los miembros de un mismo hogar, lo que puede generar problemas en la interpretación de los resultados.
- Es importante considerar el nivel de agregación adecuado para la calibración, dependiendo de los objetivos de la encuesta y la naturaleza de los datos.
- La calibración a nivel de personas puede llevar a ajustes diferentes para los miembros de un mismo hogar, lo que puede complicar la integración de los resultados a nivel de hogar.

Calibración integrada para hogares y personas

- Un enfoque más integrado, como el *integrated household weighting*, busca establecer pesos consistentes tanto a nivel de persona como de hogar, permitiendo una transición suave entre unidades de análisis.
- Sea $w_{k|i}$ el factor de expansión de la persona k que pertenece al hogar i, y a $w_{II,i}$ como el factor de expansión del hogar i, se debe cumplir la siguiente restricción:

$$w_{k|i} = w_{II,i} \text{ para toda persona } k \text{ en el hogar } i.$$

- Este enfoque garantiza que los pesos de muestreo sean coherentes con las restricciones de calibración a nivel de persona, al tiempo que facilita la agregación de los datos a nivel de hogar.

Enfoque de Estevao & Sarndal

Para calibrar por variables como el sexo, se utiliza un enfoque de post-estratificación, donde las ecuaciones de calibración se basan en la comparación entre las características de la muestra y las proyecciones demográficas.

$$\left(\sum_{k \in s} w_k x_{1k}, \sum_{k \in s} w_k x_{2i} \right) = (t_{x1}, t_{x2})$$

- La suma se hace sobre las personas en la muestra s.
- x_{k1} toma el valor de uno, si el individuo k es mujer, y cero en otro caso.
- t_{x1} y t_{x2} son las proyecciones demográfica para mujeres y hombres

Enfoque de Estevao & Sarndal

Es posible calibrar utilizando la muestra de hogares, donde las ecuaciones de calibración se basan en covariables continuas, como características socioeconómicas de los hogares.

$$\left(\sum_{i \in s_{II}} w_{II,i} z_{1i}, \sum_{i \in s_{II}} w_{II,i} z_{2i} \right) = (t_{z1}, t_{z2})$$

- La suma se realiza sobre la muestra de hogares s_{II}.
- $z_{1i} = \sum_{k \in s_i} x_{k1}$ se refiere al número de hombres en el hogar i.
- $z_{2i} = \sum_{k \in s_i} x_{k2}$ es el número de mujeres en el hogar i.
- Los totales de calibración $t_{z1} = t_{x1}$ y $t_{z2} = t_{x2}$ siguen siendo el número de hombres y mujeres en la población.

Enfoque de Estevao & Sarndal

- Calibrar sobre la base de hogares reproduce los totales auxiliares sobre la base de hogares, mientras que calibrar sobre la base de personas reproduce los totales auxiliares sobre la base de personas.
- En un hogar, la probabilidad de inclusión de las personas es de uno debido a la inclusión forzosa, lo que facilita la generación de factores de expansión para las personas en el segundo escenario de calibración.

$$w_{k|i} = \frac{w_{II,i}}{Pr(k \in U_i | i \in sI)} = \frac{w_{II,i}}{1} = w_{II,i}$$

Enfoque de Estevao & Sarndal

- La calibración conjunta para hogares y personas implica construir los pesos a nivel de persona luego de calibrar la base de hogares, utilizando la expresión:

$$w_k = d_{k|i} w_{II,i} \quad \forall k \in s_i$$

- Dado que todos los individuos pertenecientes a un hogar son seleccionados para responder la encuesta de hogares, se tiene que $d_{k|i} = 1$, por lo que el peso del individuo será igual al peso calibrado del hogar.
- Además, como el muestreo es de conglomerados en la última etapa y todos los individuos del hogar son seleccionados, el peso de muestreo del hogar será el promedio de los pesos individuales.

Enfoque de Lemaitre & Dufour

- Se crean nuevas variables de calibración a nivel de persona, como el promedio de las variables originales en el hogar.
- Variables definidas:
  - $z_{ik} = \sum_{i \in s_{II}} x_{ik}$
  - $\bar{z}_{ik} = \frac{z_{ik}}{N_i}$
  - Donde z_{ik} es la agregación a nivel de hogar de las covariables originales de calibración a nivel de persona, y N_i es el tamaño del i-ésimo hogar.

Ejecución del algoritmo de calibración

- Al usar las variables z en lugar de las variables x, se reproducen las ecuaciones de calibración satisfactoriamente.
- Todos los individuos que comparten las mismas covariables en la calibración tienen pesos idénticos dentro de un mismo hogar.
- Esta calibración se realiza con la base de datos a nivel de personas.

Conclusiones del estudio de Neethling y Galpin (2006)

- Ambos enfoques reducen el sesgo y aumentan la precisión de las estimaciones.
- Proporcionan un único conjunto de ponderaciones para los datos de las encuestas.
- El segundo enfoque permite actualizar las restricciones de calibración y ejercer un mayor control sobre los tamaños de los subgrupos de interés.

Calibración sobre razones, medias y proporciones

- Consideremos Q subgrupos de interés con razones conocidas.
- Restricción de calibración:

$$\hat{\mathbf{R}}_{cal} = (\hat{R}_{1,cal}, ..., \hat{R}_{Q,cal})' = (R_1, ..., R_Q)' = \mathbf{R}$$

- Donde $\hat{R}_{q,cal} = \frac{\sum_{k \in s} w_k y_{qk}}{\sum_{k \in s} w_k x_{qk}}$.

Restricciones Sobre Razones

- Gutierrez, Zhang, y Rodriguez (2016), imponen la restricción $\hat{\mathbf{R}}_{cal} = \mathbf{R}_U$.
- Se define la variable de información auxiliar z_{qk} para cada subgrupo q como:

$$z_{qk} = \begin{cases} y_{qk} - R_q x_{qk} & \text{si } k \in s_q \\ 0 & \text{en otro caso} \end{cases}$$

Restricciones Sobre Medias

- Si las medias de los subgrupos son conocidas, la restricción se define como $\bar{\mathbf{y}}_{cal} = \bar{\mathbf{y}}$ para cada $q = 1, ..., Q$.
- Variable de calibración z_{qk}:

$$z_{qk} = \begin{cases} y_{qk} - \bar{y}_q & \text{si } k \in s_q \\ 0 & \text{en otro caso} \end{cases}$$

En donde

$$t_{z_q} = \sum_{k \in U} z_{qk} = \sum_{k \in U} y_{qk} - R_q x_{qk} = 0$$

Caso Particular

Si las medias de los subgrupos son conocidas, la restricción queda como

$$\bar{\mathbf{y}}_{cal} = (\bar{y}_{1,cal}, ..., \bar{y}_{Q,cal})' = (\bar{y}_1, ..., \bar{y}_Q)' = \bar{\mathbf{y}}$$

Así, la restricción para las ecuaciones de calibración, $\bar{\mathbf{y}}_{cal} = \bar{\mathbf{y}}$, para cada $q = 1, ..., Q$, se define a partir de la siguiente variable de calibración:

$$z_{qk} = \begin{cases} y_{qk} - \bar{y}_q & \text{si } k \in s_q \\ 0 & \text{en otro caso} \end{cases}$$

Calibración con Valores Perdidos y Totales Estimados

- Para utilizar los estimadores de calibración en las encuestas de hogares, la información de las covariables de calibración bebe estar completa en la base de datos de la encuestas.
- La actualización del sistema de calibración requiere covariables completas en la base de datos de la encuesta.
- Ausencia de datos completos puede dificultar la inclusión de nuevas covariables en el sistema de calibración.

Calibración con Valores Perdidos y Totales Estimados

- Los estimadores de calibración se basan en un marco inferencial de diseño de muestreo, no en modelos estadísticos.
- La calibración es un problema de optimización matemática con restricciones sobre totales auxiliares.
- Los INE utilizan variables estructurales de alta calidad como covariables de calibración.

Ejemplo

Un país está evaluando la actualización de las covariables de calibración en su encuesta de fuerza de trabajo. Con la llegada de nuevos flujos de migración internacional, se considera que la nacionalidad de los encuestados está relacionada con su condición laboral. Incluir esta variable en el sistema de calibración podría reducir los sesgos causados por la ausencia de respuesta de los extranjeros en la encuesta.

Calibración con Valores Perdidos y Totales Estimados

- La actualización del sistema de calibración busca consistencia con cifras oficiales de migración y trata la ausencia de respuesta.
- La calibración se enfoca en la muestra efectiva donde las variables auxiliares están completas.
- Si suponemos que la información de la covariable no está completa en las bases, entonces:

  1 Descartar la variable auxiliar si la información está incompleta en la base de datos de la encuesta.

  2 Imputar o rellenar valores faltantes antes de la calibración para mantener la variable auxiliar.

Calibración con Valores Perdidos y Totales Estimados

- La imputación implica la introducción de algún tipo de modelo, lo que afecta la base de los estimadores de calibración.
- Se pierde el enfoque en el diseño de muestreo y la comparabilidad temporal si se utiliza imputación.
- Vincular unidades de muestra a registros de alta calidad permite una inferencia robusta y fiel al paradigma de las encuestas de hogares.

Calibración con Valores Perdidos y Totales Estimados

- Es recomendable esforzarse en obtener covariables de calibración durante la recolección de información primaria.
- Si no es posible obtener estas covariables, utilizar registros estadísticos para obtener información como la nacionalidad es una solución viable.
- Se debe evitar la adopción de modelos de imputación en las covariables de calibración.
- La calibración con totales de control estimados es cada vez más utilizada debido a la frecuencia limitada de los censos.

Calibración con Valores Perdidos y Totales Estimados

- Las estructuras poblacionales y demográficas observadas en censos pueden desactualizarse rápidamente debido a cambios como explosiones migratorias.
- Ejemplo: *American Community Survey* en los Estados Unidos proporciona estimaciones actualizadas y oportunas con información anual detallada acerca del ingreso, educación, empleo, cobertura en salud, costos del hogar y condiciones para los residentes del país.

Calibración con Valores Perdidos y Totales Estimados

- La calibración con totales de control estimados permite que una encuesta pequeña se apoye en totales de control estimados de una encuesta más grande.
- Los estimadores derivados se denotan con un asterisco y buscan nuevos ponderadores w_k^* que satisfagan la restricción:

$$\sum_{k \in s_A} w_k^* \mathbf{x}_k = \sum_{j \in s_B} w_j \mathbf{x}_j = \hat{\mathbf{t}}_{\mathbf{x},cal}$$

Calibración con Valores Perdidos y Totales Estimados

El estimador de un total con las observaciones de la muestra pequeña tendría la siguiente forma funcional:

$$\hat{t}^{*}_{y,cal} = \sum_{k \in s_A} w^{*}_{k} y_k = \hat{t}_{y,\pi} + (\hat{\mathbf{t}}_{\mathbf{x},cal} - \hat{\mathbf{t}}_{\mathbf{x},\pi})\hat{\mathbf{B}}_{s_A}$$

En donde

$$\hat{\mathbf{B}}_{s_A} = \left(\sum_{s_A} w_k q_k \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \sum_{s_A} w_k q_k \mathbf{x}_k y_k$$

Calibración con Valores Perdidos y Totales Estimados

Dever y Valliant (2016) muestran que, al utilizar la metodología de calibración con totales de control estimados, existe sesgo para los estimadores de razón, definidos así:

$$\hat{R}^*_{y,cal} = \frac{\hat{t}^*_{y,cal}}{\hat{t}^*_{z,cal}}$$

En donde $\hat{t}^*_{y,cal}$ y $\hat{t}^*_{z,cal}$ denotan dos estimadores de calibración con totales de control estimados. Nótese que esta es la misma forma que tomaría cualquier promedio estimado

$$\bar{y}^*_{cal} = \frac{\hat{t}^*_{y,cal}}{\hat{N}^*_{cal}}$$

Calibración con Valores Perdidos y Totales Estimados

El denominador de $\hat{R}^*_{cal}$ sería $\hat{t}^*_{z,cal} = \hat{N}^*_{cal}$. Dever y Valliant (2016) presentan la siguiente expresión para el sesgo de un promedio $\bar{y}^*_{cal}$:

$$Bias(\bar{y}^*_{cal}) \approx \frac{1}{E(\hat{N}^*_{cal})} \left[Bias(\hat{t}^*_{y,cal}) - \bar{y}\ Bias(\hat{N}^*_{cal}) \right]$$

Con $Bias(\hat{t}^*_{cal})$ y $Bias(\hat{N}^*_{cal})$, los sesgos de los estimadores de calibración con totales de control estimados del total poblacional (t_y) y del tamaño poblacional (N), respectivamente.

Recorte y redondeo

Recorte de pesos extremos

- Multitud de ajustes en factores de expansión puede generar pesos extremos en la distribución.
- Recorte de pesos mayores a un umbral predefinido, generalmente alrededor de 3.5 veces la mediana de los pesos

$$U = 3.5 \times mediana(\mathbf{w_k})$$

.

- Truncar pesos mayores a U para evitar valores extremadamente grandes

$$w_k^* = \begin{cases} U, \ si \ w_k \geq U \\ w_k, \ en \ otro \ caso. \end{cases}$$

.

Recorte de pesos extremos

- Calcular la cantidad neta perdida debido al recorte de pesos extremos

$$K = \sum_{s_r} (w_k^* - w_k)$$

.

- Distribuir equitativamente K entre las unidades no recortadas.
- Iterar el proceso hasta que todos los nuevos pesos estén por debajo del umbral U.
- Este procedimiento ayuda a controlar la varianza del estimador y mejorar la precisión de la inferencia.

El problema del redondeo de los factores de expansión

- Los factores de expansión no enteros generan complejidades prácticas, aunque no tengan impacto negativo teóricamente.
- Redondear los factores de expansión al entero más cercano es una práctica perjudicial que introduce sesgo en la inferencia.
- El redondeo puede causar problemas de sobre o subestimación en diferentes dominios de estudio.

Ejemplos

Ejemplos prácticos muestran cómo el redondeo afecta encuestas.

- En encuestas de establecimientos, el redondeo puede sesgar los resultados en unidades con grandes flujos de ventas.
- En encuestas agropecuarias, el redondeo puede ser desastroso para unidades que contribuyen significativamente a la producción nacional.
- En encuestas de hogares con diseños auto-ponderados, el redondeo puede sesgar por completo un estrato entero.

El problema del redondeo de los factores de expansión

- La muestra probabilística s fue seleccionada de una población finita U utilizando un diseño de muestreo que genera probabilidades de inclusión π_k para cada individuo k en U.
- Los estimadores de muestreo $\hat{t}_y = \sum_s d_k \, y_k$ son insesgados teóricamente cuando el factor de expansión d_k es igual al inverso de la probabilidad de inclusión π_k.

$$E(\hat{t}_y) = E\left(\sum_s \frac{y_k}{\pi_k}\right) = E\left(\sum_U I_k \frac{y_k}{\pi_k}\right) = \sum_U E(I_k)\frac{y_k}{\pi_k} = \sum_U \pi_k \frac{y_k}{\pi_k} = t_y$$

- El redondeo determinístico de los factores de expansión puede introducir sesgo en los estimadores de muestreo.

El problema del redondeo de los factores de expansión

- Para evitar el sesgo de redondeo, es necesario utilizar un método aleatorio que mantenga la propiedad de insesgamiento en los estimadores.
- Una estrategia es redondear los factores de expansión tomando la parte entera y luego añadir aleatoriamente una unidad a algunos factores para corregir el sesgo.
- Este enfoque probabilístico garantiza que la suma de los factores de expansión redondeados sea igual a la original, preservando la propiedad de insesgamiento en los estimadores de muestreo.

Proceso de Redondeo

1. Para $k \in s$, definir

$$\phi_k = d_k - \lfloor d_k \rfloor$$

2. Seleccionar una submuestra $s_a = (c_1, \ldots, c_k, \ldots, c_n)'$ de s con probabilidades de inclusión ϕ_k, para $k \in s$.
3. Si $c_k = 0$, entonces $\tilde{d}_k = \lfloor d_k \rfloor$; en otro caso, si $c_k = 1$, entonces $\tilde{d}_k = \lfloor d_k \rfloor + 1$.

- Así, $E(\hat{t}_y) = t_y$, asegurando la precisión de la inferencia basada en los estimadores de muestreo.

El problema del redondeo de los factores de expansión

- La submuestra s_a puede no tener un tamaño fijo debido a que $\sum_s \phi_k$ no siempre es entera; se puede emplear un algoritmo de muestreo Poisson (Gutiérrez 2016, sec. 4.1) en este caso.
- Cuando la suma $\sum_s \phi_k$ es entera, se puede utilizar un algoritmo más eficiente como el método de Brewer (Tillé 2006a) para generar una submuestra de tamaño fijo.

El problema del redondeo de los factores de expansión

- La esperanza de los factores redondeados condicionados a la submuestra s_a es igual a los factores de expansión originales, lo cual se demuestra con la ecuación

$$E(\tilde{d}_k|s_a) = \lfloor d_k \rfloor + E(c_k|s_a) = \lfloor d_k \rfloor + \phi_k = d_k$$

.

- El método aleatorio de redondeo siempre induce insesgamiento en los estimadores de muestreo, ya que

$$\begin{aligned} E\left(\sum_s \tilde{d}_k y_k\right) &= E\left[E\left(\sum_s \tilde{d}_k y_k|s_a\right)\right] \\ &= E\left(\sum_s E(\tilde{d}_k|s_a) y_k\right) \\ &= E\left(\sum_s d_k y_k\right) = t_y \end{aligned}$$

El problema del redondeo de los factores de expansión

- La calibración de los factores de expansión en la encuesta introduce un problema de optimización más complejo al utilizar el redondeo aleatorio, ya que esto puede hacer que los factores de expansión pierdan su propiedad de calibración.
- Diferentes soluciones han sido propuestas para abordar este problema, siendo la más fácil de implementar en el software estadístico `R` la presentada por Tillé (2019) y Sartore et al. (2019).

El problema del redondeo de los factores de expansión

- La calibración de los factores de expansión genera nuevos pesos w_k que satisfacen la propiedad de igualar un conjunto de totales auxiliares $\mathbf{t_x}$ disponibles para toda la población, es decir,

$$\sum_s w_k \mathbf{x}_k = \mathbf{t_x}$$

.

- El algoritmo de muestreo balanceado (Tillé 2006b, cap. 8) es una forma de calibración desde el diseño de muestreo que permite seleccionar la submuestra s_a de manera óptima y conservar la consistencia de los pesos calibrados con los totales auxiliares.

Algotirmos para la Calibración Desde el Diseño de Muestreo

1. Definir $\phi_k = w_k - \lfloor w_k \rfloor$ y calcular $\tilde{\mathbf{x}}_k = \phi_k \, \mathbf{x}_k$ para $k \in s$.
2. Seleccionar una submuestra balanceada $s_a = (c_1, \ldots, c_k, \ldots, c_n)'$ de s con probabilidades de inclusión ϕ_k, tal que

$$\sum_{k \in s_a} \frac{\tilde{\mathbf{x}}_k}{\phi_k} \cong \sum_{k \in s} \tilde{\mathbf{x}}_k$$

3. Asignar los nuevos pesos redondeados $\tilde{w}_k$ de acuerdo con las reglas: si $c_k = 0$, entonces $\tilde{w}_k = \lfloor w_k \rfloor$; si $w_k = 1$, entonces $\tilde{w}_k = \lfloor w_k \rfloor + 1$.

El problema del redondeo de los factores de expansión

La submuestra balanceada implica que los pesos redondeados cumplan la siguiente relación

$$\sum_s c_k \, \mathbf{x}_k \cong \sum_U \mathbf{x}_k - \sum_U \lfloor w_k \rfloor \, \mathbf{x}_k$$

Lo cual conlleva inmediatamente a que los nuevos pesos, además de estar redondeados, también estén calibrados; es decir

$$\sum_s \tilde{w}_k \mathbf{x}_k \cong \mathbf{t}_\mathbf{x}$$

El problema del redondeo de los factores de expansión

- El redondeo aleatorio depende de la selección de la submuestra s_a para completar los restos de la parte entera.
- Se pueden aplicar diferentes algoritmos de muestreo utilizando la librería `sampling`, como el algoritmo de Brewer (Gutiérrez 2016), para seleccionar la submuestra.

Ejemplo

- Un ejemplo numérico donde se calculan los factores de expansión y los excesos ϕ_k para una muestra de tamaño $n = 200$ de una población de tamaño $N = 9200$.
- Asuma que el vector de probabilidades de inclusión en la muestra toman la siguiente forma

$$\pi_s = (\underbrace{15/500}_{50\ veces}, ..., \underbrace{15/800}_{80\ veces}, ..., \underbrace{15/700}_{70\ veces})'$$

- El vector de pesos de muestreo estará definido de la siguiente manera:

$$\mathbf{d}_s = (\underbrace{33.33333}_{50\ veces}, ..., \underbrace{53.33333}_{80\ veces}, ..., \underbrace{46.66667}_{70\ veces})'$$

Ejemplo

- El vector de excesos $\phi_k = d_k - \lfloor d_k \rfloor$ estará dado por la siguiente expresión:

$$\phi_s = (\underbrace{0.33333}_{130\ veces}, ..., \underbrace{0.66667}_{70\ veces})'$$

- La selección de la submuestra s_a se basa en el algoritmo de Brewer debido a que $\sum_s \phi_k$ es entero en este caso.

Ejemplo

- El redondeo aleatorio garantiza que la suma de los nuevos factores coincida con la suma de los factores originales al final del proceso.
- Si los pesos están calibrados mediante covariables de calibración, se puede utilizar el método del cubo para asegurar que la submuestra esté balanceada y que los pesos redondeados cumplan las restricciones de calibración dentro de una tolerancia predefinida.

¡Gracias!

Email: andres.gutierrez@cepal.org

Referencias

Dever, Jill A., y Richard Valliant. 2016. «General Regression Estimation Adjusted for Undercoverage and Estimated Control Totals». *Journal of Survey Statistics and Methodology* 4 (3): 289-318. https://doi.org/10.1093/jssam/smw001.

Gutierrez, Hugo Andres, Hanwen Zhang, y Nelson Rodriguez. 2016. «The Performance of Multivariate Calibration on Ratios, Means and Proportions». *Revista Colombiana de Estadística* 39 (2): 281. https://doi.org/10.15446/rce.v39n2.55424.

Gutiérrez, H. A. 2016. *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*. Segunda edición. Ediciones de la U.

Neethling, A, y J S Galpin. 2006. «Weighting of Household Survey Data: A Comparison of Various Calibration, Integrated and Cosmetic Estimators». *South African Statistical Journal*, 23.

Sartore, Luca, Kelly Toppin, Linda Young, y Clifford Spiegelman. 2019. «Developing Integer Calibration Weights for Census of Agriculture». *Journal of Agricultural, Biological and Environmental Statistics* 24 (1): 26-48. https://doi.org/10.1007/s13253-018-00340-4.

Silva, PL. d N. 2004. «Calibration estimation: when and why, how much and how». *Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística*.

Tillé, Yves. 2006b. *Sampling Algorithms*. Springer Series en Statistics. Springer-Verlag. https://doi.org/10.1007/0-387-34240-0.

———. 2006a. *Sampling Algorithms*. Springer Series en Statistics. Springer-Verlag.