Instituto Nacional de Estadística

OPOSICIONES AL CUERPO SUPERIOR DE ESTADÍSTICOS DEL ESTADO

BOE NÚM. 270, DE 12 DE OCTUBRE DE 2020, PÁG. 87165

Producción Estadística Oficial: Métodos Avanzados

Grupo de Materias Específicas
Especialidad I: Estadística-Ciencia de datos

INE
13 de septiembre de 2021

Índice general

Tema 1

Estimación insesgada en diseños muestrales por conglomerados II. Muestreo de conglomerados con submuestreo: definición, estimadores, varianza y estimador de la varianza.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía.

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

1.1 Introducción

En el tema 7[1] del bloque "Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas"del grupo de materias comunes se introdujo el muestreo de conglomerados monoetápico[2] o sin submuestreo. Este tipo de muestreo surge especialmente en encuestas de tamaño medio o grande cuando las unidades muestrales no son los elementos de la población (unidades de análisis). El muestreo directo de elementos no se usa por una de las siguientes razones o ambas:

- No existe un marco muestral que identifique todos y cada uno de los elementos de la población y el coste de producir un marco muestral así es prohibitivo o prácticamente imposible.
- Los elementos de la población se encuentran tan dispersos que el proceso de medida (recogida de datos) presenta un coste prohibitivo debido a la logística necesaria (desplazamientos para administrar entrevistas, supervisión del trabajo de campo, etc.). Estas dificultades, además, suelen conllevar errores ajenos al muestreo (como errores de medida o falta de respuesta).

[1] Tema 7. Estimación insesgada en diseños muestrales por conglomerados I. Muestreo por conglomerados sin submuestreo: definición, estimadores, varianza y estimador de la varianza.

[2] *One-stage cluster sampling*.

Ejemplo 1. Supóngase que quieren estimarse el número de *tablets* electrónicas por residente en un país. No existe un listado de todos estos dispositivos vendidos y el muestreo directo sobre personas conduciría a una muestra dispersa cuya recogida de datos superaría nuestras restricciones presupuestarias. Por tanto, no puede plantearse un muestreo directo sobre las personas. En su lugar, puede considerarse un listado de secciones censales, seleccionarlas bajo un diseño muestral probabilístico, como se introdujo en el tema 7, y entrevistar a todos los residentes de cada sección censal seleccionada. Esto es el muestreo de conglomerados monoetápico. Las unidades de análisis son las personas residentes y las unidades muestrales son las secciones censales. ■

En este segundo tema sobre muestreo de conglomerados se introducirán posteriores etapas de selección de unidades dentro de cada conglomerado. Para ello, retomamos las mismas definiciones anteriores:

Definición 1

Una *unidad muestral primaria*[a] (a veces también *unidad primaria de muestreo*) es un subconjunto U_i de elementos de la población de una partición disjunta de la población $U = \cup_{i=1}^{N_I} U_i$, con $U_i \cap U_j = \emptyset$ para todo $i \neq j$, donde N_I denota el número total (conocido) de conglomerados.

[a] *Primary Sampling Unit* – PSU.

En el muestreo bietápico o polietápico (más de dos etapas de selección muestral), se seleccionan (sub)conglomerados o unidades en cada conglomerado.

Definición 2

Una *unidad muestral secundaria*[a] (a veces también *unidad secundaria de muestreo*) es cualquier tipo de unidad seleccionada dentro de cada PSU. Si las SSUs son elementos de la población, hablamos de *muestreo bietápico de elementos* y si las SSUs son (sub)conglomerados, hablamos de *muestreo bietápico de conglomerados*, en cuyo caso todos los elementos de cada SSU son seleccionados.

[a] *Secondary Sampling Unit* – SSU.

Por extensión, se define de manera análoga el muestreo polietápico (tanto de elementos como de conglomerados) surgiendo, por tanto, las unidades muestrales terciarias, cuaternarias, etc. Las unidades seleccionadas en la última etapa se denominan las *unidades muestrales últimas*[3] o *unidades muestrales finales*.

Comentario 1. Conglomerados vs. estratos (1/2).
La agrupación de elementos de la población en subconjuntos de la población no es novedoso en el muestreo por conglomerados, pues en la estratificación de la población

[3] *Ultimate sampling units*.

para los diseños muestrales estratificados que se ve en el tema 6 [4] del bloque "Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas" del grupo de materias comunes también se realizan agrupaciones de las unidades de análisis. De hecho, desde el punto de vista estrictamente matemático, las definiciones de PSU y estrato son idénticas (particiones disjuntas de la población). La diferencia estriba en el uso que se hace de estas agrupaciones en el diseño de la muestra. Como veremos más adelante, mientras que todos los estratos se muestrean (algunos incluso exhaustivamente), los conglomerados son seleccionados de acuerdo con un diseño muestral. Esto tiene consecuencias en la estimación de la varianza y, por tanto, en la precisión de las estimaciones. ■

Al igual que indicamos en el tema 7 sobre muestreo por conglomerados monoetápico, para la construcción de estimadores debemos distinguir entre la estimación de totales poblacionales $Y_U = \sum_{k\in U} y_k$ y de medias poblacionales $\bar{Y}_U = \frac{1}{N}\sum_{k\in U} y_k$. En los diseños muestrales de elementos, la construcción del estimador Horvitz-Thompson de la media poblacional $\widehat{\bar{Y}}_U^{\text{HT}}$, de su varianza $\mathbb{V}\left[\widehat{\bar{Y}}_U^{\text{HT}}\right]$ y del estimador de Horvitz-Thompson de ésta $\widehat{\mathbb{V}}^{\text{HT}}\left[\widehat{\bar{Y}}_U^{\text{HT}}\right]$ se reduce a dividir por N en el primer caso y por N^2 en los dos últimos casos las correspondientes expresiones para los totales poblacionales. En los diseños muestrales por conglomerados, cuando el total de número de elementos N de la población no se conoce (p.ej. porque el marco muestral no puede construirse), este método de construcción no puede aplicarse y debe construirse también un estimador para N, lo que conduce al uso de estimadores de razón $\widehat{\bar{Y}}_U^{\text{Rat}} = \frac{\widehat{Y}_U^{\text{HT}}}{\widehat{N}^{\text{HT}}}$ (véase el tema 3).

1.2 Definición y notación

Mantenemos la misma notación que el tema 7 sobre muestreo por conglomerados sin submuestreo del grupo de materias comunes. La población finita $U = \{1, \ldots, k \ldots, N\}$ se divide en una partición disjunta de conglomerados $U = U_1 \cup \ldots U_i \cup \cdots \cup U_{N_I}$, donde el conjunto de conglomerados se denota por $U_I = \{1, \ldots, i, \ldots, N_I\}$. De este modo, el subíndice I denota las entidades asociadas con la primera etapa en el muestreo. El tamaño de cada conglomerado U_i se denotará por N_{U_i} o, simplemente, N_i. Adviértase que se cumple

$$U = \bigcup_{i\in U_I} U_i \qquad N = \sum_{i\in U_I} N_i.$$

Para la muestra, de modo similar, se denotará por n_{s_i} o, simplemente, n_i el tamaño de la submuestra del conglomerado U_i.

En el muestreo por conglomerados monoetápico sucede en general que la varianza del estimador HT es mayor que en el caso de un muestreo aleatorio simple sin repetición.

[4] Tema 6. Estimación insesgada en diseños muestrales sobre unidades elementales IV. Muestreo estratificado: definición, estimadores, varianza y estimador de la varianza. Afijación muestral óptima. Otras afijaciones bajo muestreo aleatorio simple.

Esto se debe a (i) la tendencia habitual de que los elementos de un mismo conglomerado son parecidos entre ellos[5] y (ii) la variación en el tamaño de los conglomerados. La varianza del estimador HT bajo muestreo de conglomerados monoetápico siempre se puede reducir seleccionando más y más conglomerados. Sin embargo, esto conlleva un mayor coste de recogida de datos por ser una muestra mayor resultando a menudo inadmisible debido a las restricciones presupuestarias.

Para mantener bajo control el coste y, al mismo tiempo, aumentar el número de conglomerados seleccionados, podemos extraer una submuestra en los conglomerados seleccionados en lugar de entrevistar a todas las unidades de los conglomerados seleccionados. En tal caso, a continuación debemos estimar el total poblacional de cada conglomerado Y_{U_i} a partir de las submuestras. Si la variación dentro del conglomerado es pequeña, las estimaciones $\widehat{Y}_{U_i}^{\text{HT}}$ tendrán una varianza pequeña, incluso para tamaños de submuestras relativamente moderados. En tal caso, merece la pena usar muestreo bietápico en lugar de muestreo unietápico.

Empezamos con una categoría muy general de diseños muestrales de elementos bietápicos:

Definición 3. Muestreo bietápico

Un diseño muestral de elementos bietápico se define mediante dos etapas de muestreo:

Primera etapa: Se selecciona una muestra s_I de PSUs de U_I ($s_I \subset U_I$) de acuerdo con el diseño $p_I(\cdot)$.
Segunda etapa: Para cada $i \in s_I$, se selecciona una muestra de s_i elementos de U_i ($s_i \subset U_i$) de acuerdo con un diseño $p_i(\cdot|s_I)$ **invariante** e **independiente**.

Comentario 2. La invarianza de los diseños de segunda etapa $p_i(\cdot|s_I)$ quiere decir que, para cada $i \in U_I$ y cada $s_I \subset U_I$, se cumple

$$p_i(\cdot|s_I) = p_i(\cdot).$$

En otras palabras, cada vez que un conglomerado U_i es seleccionado en la primera etapa, debe emplearse el mismo diseño de submuestreo $p_i(\cdot)$. Por ejemplo, se puede establecer que se realice un muestreo aleatorio simple de tamaño muestral n_i cada vez que el conglomerado U_i es seleccionado. ■

Comentario 3. La independencia de los diseños de segunda etapa $p_i(\cdot|s_I)$ quiere decir que, para cada cada $s_I \subset U_I$, se cumple

$$\mathbb{P}\left(\bigcup_{i\in s_I} s_i|s_I\right) = \prod_{i\in s_i} \mathbb{P}\left(s_i|s_I\right).$$

[5]Por ejemplo, porque las personas que viven en un mismo área tienen características similares.

En otras palabras, el submuestreo en cada conglomerado se efectúa independientemente del submuestreo en el resto de conglomerados. ■

Comentario 4. Cuando las propiedades de invarianza y de independencia no se cumplen, nos encontramos en una situación más general y deben aplicarse entonces los métodos de muestreo bifásico (o multifásico).

La muestra de elementos resultante, denotada por s, está compuesta por $n_I = |s_I|$ submuestras $s = \bigcup_{i \in s_I} s_i$ y, por tanto, el tamaño muestral final será la suma de los tamaños muestrales de cada submuestra $n = \sum_{i \in s_I} n_i$.

Debemos fijar también la notación para las probabilidades de inclusión asociadas al muestreo de elementos bietápico. Para el diseño de primera etapa $p_I(\cdot)$ denotaremos las probabilidades de inclusión como π_{Ii} y, π_{Iij}. Denotaremos también $\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$, con $\Delta_{Iii} = \pi_{Ii}(1 - \pi_{Ii})$, y $\check{\Delta}_{Iij} = \frac{\Delta_{Iij}}{\pi_{Iij}}$.

De modo similar, para el diseño de segunda etapa $p_i(\cdot)$, usaremos la notación $\pi_{k|i}$ y $\pi_{kl|i}$. Las cantidades Δ son $\Delta_{kl|i} = \pi_{kl|i} - \pi_{k|i}\pi_{l|i}$, con $\Delta_{kk|i} = \pi_{k|i}(1 - \pi_{k|i})$ y, finalmente, $\check{\Delta}_{kl|i} = \frac{\Delta_{kl|i}}{\pi_{kl|i}}$.

1.3 Estimadores, varianza y estimador de la varianza

Para obtener el estimador HT, su varianza y el estimador HT de esta varianza podemos usar el resultado general sobre la construcción del estimador HT con las probabilidades de inclusión π_k y π_{kl} particulares del muestreo bietápico. Se sigue de las propiedades de invarianza y de independencia que las probabilidades de inclusión de los elementos son

$$\pi_k = \pi_{Ii}\pi_{k|i} \quad \text{si } k \in U_i \tag{1.1a}$$

y

$$\pi_{kl} = \begin{cases} \pi_{Ii}\pi_{k|i} & \text{si } k = l \in U_i, \\ \pi_{Ii}\pi_{kl|i} & \text{si } k, l \in U_i, k \neq l, \\ \pi_{Iij}\pi_{k|i}\pi_{l|j} & \text{si } k \in U_i \text{ y } l \in U_j. \end{cases} \tag{1.1b}$$

Teorema 1

El estimador HT, su varianza y el estimador HT de esta varianza para un muestreo de elementos bietápico están dados por

i. $\widehat{Y}_U^{\mathrm{HT}} = \sum_{k \in U} \frac{y_k}{\pi_k}$,

ii. $\mathbb{V}\left[\widehat{Y}_U^{\mathrm{HT}}\right] = \sum_{k,l\in U}\left(\pi_{kl}-\pi_k\pi_l\right)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$,

iii. $\widehat{\mathbb{V}}^{\mathrm{HT}}\left[\widehat{Y}_U^{\mathrm{HT}}\right] = \sum_{k,l\in U}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$,

donde las probabilidades de inclusión están dados por (1.1).

Demostración 1

Es una aplicación directa del Teorema de Horvitz-Thompson. ■

Aunque este resultado nos permite computar directamente las estimaciones derivadas de este diseño muestral, el Teorema 1 no nos permite comprender cómo influye cada etapa en la construcción de las estimaciones. Para ello, debemos recurrir a un resultado general de la teoría de probabilidad (véase p.ej. Grimmet y Stirzaker 2004, pág. 69). Sean X, Y variables aleatorias. Se cumplen:

- $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X|Y]]$.
- $\mathbb{V}[X] = \mathbb{V}_Y[\mathbb{E}[X|Y]] + \mathbb{E}_Y[\mathbb{V}[X|Y]]$.

En el muestreo bietápico, condicionamos sobre el suceso de que la muestra s_I se seleccione en la primera fase. Sea $\check{y}_{k|i} = \frac{y_k}{\pi_{k|i}}$ y sea[6]

$$\widehat{Y}_{U_i|i}^{\mathrm{HT}} = \sum_{k\in s_i}\check{y}_{k|i} \tag{1.2}$$

el estimador HT con respecto a la segunda etapa del total poblacional de la PSU $Y_{U_i} = \sum_{k\in U_i} y_k$. En caso de submuestrear U_i repetidamente de acuerdo con el diseño $p_i(\cdot)$, $\widehat{Y}_{U_i|i}^{\mathrm{HT}}$ es un estimador insesgado de Y_{U_i}, esto es, es insesgado condicionalmente a seleccionar s_i en la primera etapa. La varianza con respecto a la segunda etapa es

$$V_i \equiv \mathbb{V}\left[\widehat{Y}_{U_i|i}^{\mathrm{HT}}\right] = \sum_{k\in U_i}\sum_{l\in U_i}\Delta_{kl|i}\;\check{y}_{k|i}\;\check{y}_{l|i} \tag{1.3}$$

cuyo estimador insesgado es

$$\widehat{V}_i \equiv \widehat{\mathbb{V}}^{\mathrm{HT}}\left[\widehat{Y}_{U_i|i}^{\mathrm{HT}}\right] = \sum_{k\in s_i}\sum_{l\in s_i}\check{\Delta}_{kl|i}\;\check{y}_{k|i}\;\check{y}_{l|i}. \tag{1.4}$$

Como sucede con el muestreo directo de elementos, se pueden emplear fórmulas alternativas para diseños de tamaño muestral fijo. Si el diseño $p_i(\cdot)$ es de tamaño fijo, V_i también se puede escribir como

$$V_i = -\frac{1}{2}\sum_{k\in U_i}\sum_{l\in U_i}\Delta_{kl|i}\;(\check{y}_{k|i}-\check{y}_{l|i})^2 \tag{1.5}$$

[6] Adviértase la diferencia entre $\widehat{Y}_{U_i}^{\mathrm{HT}}$ y $\widehat{Y}_{U_i|i}^{\mathrm{HT}}$. El primero es el estimador HT del total poblacional de la variable y en el conglomerado U_i, que se construye con las probabilidades de inclusión π_k, $k \in s_i$. El segundo es el estimador HT del total poblacional de la variable y en el conglomerado U_i, condicionado a que se ha escogido el conglomerado U_i en la primera etapa, por tanto, se construye con las probabilidades inclusión $\pi_{k|i}$.

cuyo estimador insesgado viene dado por

$$\widehat{V}_i = -\frac{1}{2}\sum_{k\in s_i}\sum_{l\in s_i}\breve{\Delta}_{kl|i}\ (\breve{y}_{k|i}-\breve{y}_{l|i})^2. \tag{1.6}$$

Para las estimaciones HT en el muestreo bietápico, deben combinarse las aportaciones de ambas etapas:

Teorema 2

En el muestreo de elementos bietápico, el estimador HT del total poblacional $Y_U = \sum_{k\in U} y_k$ viene dado por

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{i\in s_I}\frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}} = \sum_{i\in s_I}\frac{1}{\pi_{Ii}}\sum_{k\in s_i}\frac{y_k}{\pi_{k|i}} \tag{1.7}$$

La varianza de $\widehat{Y}_U^{\mathrm{HT}}$ se puede escribir como la suma de dos componentes,

$$V_{2st} \equiv \mathbb{V}\left[\widehat{Y}_U^{\mathrm{HT}}\right] = V_{PSU} + V_{SSU} \tag{1.8}$$

donde

$$V_{PSU} = \sum_{i\in U_I}\sum_{j\in U_I}\Delta_{Iij}\ \frac{Y_{U_i}}{\pi_{Ii}}\ \frac{Y_j}{\pi_{Ij}}, \tag{1.9a}$$

$$V_{SSU} = \sum_{i\in U_I}\frac{V_i}{\pi_{Ii}} \equiv \sum_{i\in U_I}\frac{1}{\pi_{Ii}}\sum_{k\in U_i}\sum_{l\in U_i}\Delta_{kl|i}\ \breve{y}_{k|i}\ \breve{y}_{l|i} \tag{1.9b}$$

El estimador HT de la varianza V_{2st} se construye estimando cada componente por separado de modo insesgado:

$$\begin{aligned}\widehat{V}_{PSU} &= \sum_{i\in s_I}\sum_{j\in s_I}\breve{\Delta}_{Iij}\ \frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}}\ \frac{\widehat{Y}_{U_j|j}^{\mathrm{HT}}}{\pi_{Ij}} - \sum_{i\in s_I}\frac{1}{\pi_{Ii}}\left(\frac{1}{\pi_{Ii}}-1\right)\widehat{V}_i \\ &\equiv \sum_{i\in s_I}\sum_{j\in s_I}\breve{\Delta}_{Iij}\ \frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}}\ \frac{\widehat{Y}_{U_j|j}^{\mathrm{HT}}}{\pi_{Ij}} \\ &\quad - \sum_{i\in s_I}\frac{1}{\pi_{Ii}}\left(\frac{1}{\pi_{Ii}}-1\right)\sum_{k\in s_i}\sum_{l\in s_i}\breve{\Delta}_{kl|i}\ \breve{y}_{k|i}\ \breve{y}_{l|i},\end{aligned} \tag{1.10a}$$

$$\widehat{V}_{SSU} = \sum_{i\in s_I}\frac{\widehat{V}_i}{\pi_{Ii}^2} = \sum_{k\in s_i}\frac{1}{\pi_{Ii}^2}\sum_{k\in s_i}\sum_{l\in s_i}\breve{\Delta}_{kl|i}\ \breve{y}_{k|i}\ \breve{y}_{l|i}. \tag{1.10b}$$

El estimador HT de $\widehat{Y}_U^{\text{HT}}$ está dado por:

$$\widehat{V}_{2st} \equiv \widehat{\mathbb{V}}^{\text{HT}}\left[\widehat{Y}_U^{\text{HT}}\right] = \widehat{V}_{PSU} + \widehat{V}_{SSU} = \sum_{i\in s_I}\sum_{j\in s_I} \check{\Delta}_{Iij}\,\frac{\widehat{Y}_{U_i|i}^{\text{HT}}}{\pi_{Ii}}\,\frac{\widehat{Y}_{U_j|j}^{\text{HT}}}{\pi_{Ij}} + \sum_{i\in s_I}\frac{\widehat{V}_i}{\pi_{Ii}}. \tag{1.11}$$

Comentario 5. Para construir la demostración del Teorema 2 simplificamos la notación, permitiendo así su generalización al caso multietápico:

$$\begin{aligned}
\mathbb{E}_I\mathbb{E}_{II}\left[\widehat{Y}_U^{\text{HT}}\right] &= \mathbb{E}_{p_I}\left[\mathbb{E}\left[\widehat{Y}_U^{\text{HT}}|s_I\right]\right],\\
\mathbb{V}_I\mathbb{E}_{II}\left[\widehat{Y}_U^{\text{HT}}\right] &= \mathbb{V}_{p_I}\left[\mathbb{E}\left[\widehat{Y}_U^{\text{HT}}|s_I\right]\right],\\
\mathbb{E}_I\mathbb{V}_{II}\left[\widehat{Y}_U^{\text{HT}}\right] &= \mathbb{E}_{p_I}\left[\mathbb{V}\left[\widehat{Y}_U^{\text{HT}}|s_I\right]\right].
\end{aligned}$$

Es decir, el subíndice I indica la esperanza o la varianza con respecto al diseño $p_I(\cdot)$ usado en la primera fase y II indica la esperanza condicionada o la varianza condicionada con respecto al conjunto de diseños $p_i(\cdot)$, $i \in s_I$, usado en la segunda fase, dado s_I. ■

Demostración 2

La ecuación (1.7) equivale al estimador de Horvitz-Thompson toda vez que reconocemos las probabilidades de inclusión de primer orden de cada unidad $k \in U$ de la población (véase la ecuación (1.1a)):

$$\begin{aligned}
\widehat{Y}_U^{\text{HT}} &= \sum_{k\in s}\frac{y_k}{\pi_k} = \sum_{i\in s_I}\sum_{k\in s_i}\frac{y_k}{\pi_{Ii}\pi_{k|i}}\\
&= \sum_{i\in s_I}\frac{\left(\sum_{k\in s_i}\frac{y_k}{\pi_{k|i}}\right)}{\pi_{Ii}} = \sum_{i\in s_I}\frac{\widehat{Y}_{U_i|i}^{\text{HT}}}{\pi_{Ii}}.
\end{aligned}$$

Sin embargo, también es conveniente saber cómo influye cada etapa en la insesgadez del estimador:

$$\begin{aligned}
\mathbb{E}_I\mathbb{E}_{II}\left[\widehat{Y}_U^{\text{HT}}\right] &= \mathbb{E}_{p_I}\left[\mathbb{E}\left[\widehat{Y}_U^{\text{HT}}|s_I\right]\right]\\
&= \mathbb{E}_{p_I}\left[\mathbb{E}\left[\sum_{i\in s_I}\frac{1}{\pi_{Ii}}\sum_{k\in s_i}\frac{y_k}{\pi_{k|i}}|s_I\right]\right]\\
&= \mathbb{E}_{p_I}\left[\sum_{i\in s_I}\frac{1}{\pi_{Ii}}\mathbb{E}\left[\sum_{k\in s_i}\frac{y_k}{\pi_{k|i}}|s_I\right]\right]\\
&= \mathbb{E}_{p_I}\left[\sum_{i\in s_I}\frac{1}{\pi_{Ii}}\left[\sum_{k\in U_i}y_k\right]\right] = \mathbb{E}_{p_I}\left[\sum_{i\in s_I}\frac{Y_{U_i}}{\pi_{Ii}}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i \in U_I} Y_{U_i} \\
&= \sum_{k \in U} y_k = Y_U
\end{aligned}
$$

Para la varianza, usamos los momentos condicionales calculados gracias a las propiedades de invarianza e independencia:

$$
\mathbb{E}_{II}\left[\widehat{Y}_U^{\mathrm{HT}}\right] = \mathbb{E}\left[\widehat{Y}_U^{\mathrm{HT}} | s_I\right] = \sum_{i \in s_I} \mathbb{E}_{p_i}\left[\frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}} | s_I\right] \underbrace{=}_{\text{invarianza}} \sum_{i \in s_I} \mathbb{E}_{p_i}\left[\frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}}\right] = \sum_{i \in s_I} \frac{Y_{U_i}}{\pi_{Ii}},
\tag{1.12a}
$$

$$
\mathbb{V}_{II}\left[\widehat{Y}_U^{\mathrm{HT}}\right] = \mathbb{V}\left[\widehat{Y}_U^{\mathrm{HT}} | s_I\right] \underbrace{=}_{\text{independencia}} \sum_{i \in s_I} \mathbb{V}\left[\frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}} | s_I\right] \underbrace{=}_{\text{invarianza}} \sum_{i \in s_I} \mathbb{V}_{p_i}\left[\frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}}\right] = \sum_{i \in s_I} \frac{V_i}{\pi_{Ii}^2}.
\tag{1.12b}
$$

De este modo, la varianza se escribe inmediatamente como

$$
\begin{aligned}
V_{2st} &= \mathbb{V}\left[\widehat{Y}_U^{\mathrm{HT}}\right] \\
&= \mathbb{V}_I \mathbb{E}_{II}\left[\widehat{Y}_U^{\mathrm{HT}}\right] + \mathbb{E}_I \mathbb{V}_{II}\left[\widehat{Y}_U^{\mathrm{HT}}\right] \\
&= \mathbb{V}_I\left[\sum_{i \in s_I} \frac{Y_{U_i}}{\pi_{Ii}}\right] + \mathbb{E}_I\left[\sum_{i \in s_I} \frac{V_i}{\pi_{Ii}^2}\right] \\
&= \sum_{i \in U_I} \sum_{j \in U_I} \Delta_{Iij} \frac{Y_{U_i}}{\pi_{Ii}} \frac{Y_{U_j}}{\pi_{Ij}} + \sum_{i \in U_I} \frac{V_i}{\pi_{Ii}}
\end{aligned}
$$

que demuestra la expresión de la varianza (1.8).

Ahora debemos demostrar las expresiones para la estimación insesgada de la varianza. Procedemos nuevamente analizando cada componente. Por la propiedad de independencia y la definición de V_i, se cumple

$$
\mathbb{E}_{II}\left[\widehat{Y}_{U_i|i}^{\mathrm{HT}}\, \widehat{Y}_{U_j|j}^{\mathrm{HT}}\right] = \begin{cases} Y_{U_i}^2 + V_i, & \text{si } i = j, \\ Y_{U_i} Y_{U_j}, & \text{si } i \neq j. \end{cases}
$$

Ahora tenemos

$$\begin{aligned}
\mathbb{E}\left[\sum_{i\in s_I}\sum_{j\in s_I}\breve{\Delta}_{Iij}\frac{\widehat{Y}^{\mathrm{HT}}_{U_i|i}}{\pi_{Ii}}\frac{\widehat{Y}^{\mathrm{HT}}_{U_j|j}}{\pi_{Ij}}\right] &= \mathbb{E}_I\left[\sum_{i\in s_I}\sum_{j\in s_I}\breve{\Delta}_{Iij}\frac{\mathbb{E}_{II}\left[\widehat{Y}^{\mathrm{HT}}_{U_i|i}\,\widehat{Y}^{\mathrm{HT}}_{U_j|j}\right]}{\pi_{Ii}\pi_{Ij}}\right] \\
&= \mathbb{E}_I\left[\sum_{i\in s_I}\sum_{j\in s_I}\breve{\Delta}_{Iij}\frac{Y_{U_i}}{\pi_{Ii}}\frac{Y_{U_j}}{\pi_{Ij}}\right] + \mathbb{E}_I\left[\sum_{i\in s_I}\breve{\Delta}_{Iii}\frac{V_i}{\pi^2_{Ii}}\right] \\
&= \sum_{i\in U_I}\sum_{j\in U_I}\Delta_{Iij}\frac{Y_{U_i}}{\pi_{Ii}}\frac{Y_{U_j}}{\pi_{Ij}} + \sum_{i\in U_I}\frac{\pi_{Ii}-\pi^2_{Ii}}{\pi^2_{Ii}}\,V_i \\
&= V_{PSU} + \sum_{i\in U_I}\left(\frac{1}{\pi_{Ii}}-1\right)V_i \qquad (1.13a)
\end{aligned}$$

y

$$\begin{aligned}
\mathbb{E}\left[-\sum_{i\in s_I}\frac{1}{\pi_{Ii}}\left(\frac{1}{\pi_{Ii}}-1\right)\widehat{V}_i\right] &= -\mathbb{E}_I\left[\sum_{i\in s_I}\frac{1}{\pi_{Ii}}\left(\frac{1}{\pi_{Ii}}-1\right)\mathbb{E}_{II}\left[\widehat{V}_i\right)\right] \\
&= -\mathbb{E}_I\left[\sum_{i\in s_I}\frac{1}{\pi_{Ii}}\left(\frac{1}{\pi_{Ii}}-1\right)V_i\right] \\
&= -\sum_{i\in U_I}\left(\frac{1}{\pi_{Ii}}-1\right)V_i \qquad (1.13b)
\end{aligned}$$

La suma de (1.13a) y (1.13b) implica que $\mathbb{E}\left[\widehat{V}_{PSU}\right] = V_{PSU}$.

A continuación demostramos que $\widehat{V}_{SSU}$ es insesgado para V_{SSU}:

$$\begin{aligned}
\mathbb{E}\left[\widehat{V}_{SSU}\right] &= \mathbb{E}_I\left[\sum_{i\in s_I}\frac{\mathbb{E}_{II}\left[\widehat{V}_i\right]}{\pi^2_{Ii}}\right] \\
&= \mathbb{E}_I\left[\sum_{i\in s_I}\frac{V_i}{\pi^2_{Ii}}\right] = \sum_{i\in U_I}\frac{V_i}{\pi_{Ii}} = V_{SSU}.
\end{aligned}$$

Recopilando tenemos que

$$\mathbb{E}\left[\widehat{\mathbb{V}}^{\mathrm{HT}}\left[\widehat{Y}^{\mathrm{HT}}_U\right]\right] = \mathbb{E}\left[\widehat{V}_{PSU}+\widehat{V}_{SSU}\right] = V_{PSU}+V_{SSU} = \mathbb{V}\left[\widehat{Y}^{\mathrm{HT}}_U\right]$$

La demostración queda completada. ■

Es importante remarcar que el Teorema 2 es válido independientemente de la igualdad de los tamaños de los conglomerados y de los tipos de muestreo de ambas etapas (aleatorio simple, Bernoulli, Poisson, sistemático, etc.).

La descomposición de la varianza y su estimador no solo tiene interés teórico. En muchas aplicaciones, como en los estudios piloto previos a una encuesta o en la planificación y mejora de una encuesta en curso (véase la siguiente sección), puede resultar de interés tener una idea de la contribución a la varianza de cada una de las dos etapas de muestreo. Es decir, es necesario estimar V_{PSU} y V_{SSU} de forma separada. Los estimadores proporcionados en (1.10a) y en (1.10b) son útiles para este propósito. Téngase en cuenta que (1.10a) no siempre proporciona estimaciones positivas.

Ejemplo 2. Consideremos el diseño bietápico en el que se usa el muestreo aleatorio simple sin reemplazamiento en ambas etapas. En la primera etapa se toma una muestra s_I de n_I conglomerados a partir de los N_I conglomerados y, a continuación, para cada conglomerado $i \in s_I$, se selecciona una muestra de tamaño n_i sobre los N_i elementos del conglomerado. Aplicando el Teorema 2 obtenemos el estimador HT del total poblacional Y_U que puede escribirse como

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{i \in s_I} \frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}} = \frac{N_I}{n_I} \sum_{i \in s_I} N_i \bar{y}_{s_i}. \tag{1.14}$$

La varianza es

$$V_{2st} = N_I^2 \frac{1-f_I}{n_I} S_{Y U_I}^2 + \frac{N_I}{n_I} \sum_{i \in U_I} N_i^2 \frac{1-f_i}{n_i} S_{y U_i}^2 \tag{1.15}$$

donde

- $f_I = \frac{n_I}{N_I}$ es la fracción de muestreo en primera etapa;
- $f_i = \frac{n_i}{N_i}$ es la fracción de muestreo en segunda etapa;
- $S_{Y U_I}^2 = \frac{1}{N_I - 1} \sum_{i \in U_I} (Y_{U_i} - \bar{y}_{U_I})^2$ es la cuasivarianza en U_I de los totales poblacionales Y_{U_i} de los conglomerados i, con $\bar{y}_{U_I} = \sum_{i \in U_I} \frac{Y_{U_i}}{N_I}$;
- $S_{y U_i}^2 = \frac{1}{N_i - 1} \sum_{k \in U_i} (y_k - \bar{y}_{U_i})^2$ es la cuasivarianza de la variable objetivo y en el conglomerado U_i, con $\bar{y}_{U_i} = \sum_{k \in U_i} \frac{y_k}{N_i}$.

El estimador insesgado de la varianza es

$$\widehat{V}_{2st} = N_I^2 \frac{1-f_I}{n_I} S_{\widehat{Y} s_I}^2 + \frac{N_I}{n_I} \sum_{i \in s_I} N_i^2 \frac{1-f_i}{n_i} S_{y s_i}^2 \tag{1.16}$$

donde

- $S_{\widehat{Y} s_I}^2 = \frac{1}{n_I - 1} \sum_{i \in s_I} \left[\widehat{Y}_{U_i|i}^{\mathrm{HT}} - \left(\frac{1}{n_I} \sum_{i \in s_I} \widehat{Y}_{U_i|i}^{\mathrm{HT}} \right) \right]^2$ es la cuasivarianza en s_I de los totales estimados $\widehat{Y}_{U_i|i}^{\mathrm{HT}} = N_i \bar{y}_{s_i}$ de los conglomerados i;

- $S^2_{ys_i} = \frac{1}{n_i-1}\sum_{k\in s_i}(y_k - \bar{y}_{s_i})^2$ es la cuasivarianza de la variable objetivo y en el conglomerado U_i, con $\bar{y}_{s_i} = \frac{1}{n_i}\sum_{k\in s_i} y_k$.

■

Ejemplo 3. **Conglomerados de igual tamaño y probabilidades iguales**
En la literatura de muestreo (véase p.ej. Kish 1965; Cochran 1977; Lehtonen y Pahkinen 2004; Lohr 2010) a menudo se singulariza el caso de conglomerados de igual tamaño seleccionados con probabilidades iguales en ambas etapas. Supongamos, por tanto, $N_i = N/N_I$, $n_i = m$ para todo $i \in U_I$, de modo que las probabilidades de inclusión de primer orden $\pi_{Ii} = \frac{n_I}{N_I} \equiv f_I$ y $\pi_{k|i} = \frac{N_I \cdot m}{N} \equiv f_{II}$ para todo $i \in U_I$ y todo $k \in U$ se reducen a las fracciones de muestreo en primera y en segunda etapa, respectivamente. Las probabilidades de inclusión de segundo orden equivalen respectivamente a las obtenidas por muestreo aleatorio simple, de modo que nos encontramos en el mismo caso que el Ejemplo 2.

En estas condiciones, el estimador HT del total poblacional se reduce a

$$\widehat{Y}_U^{\mathrm{HT}} = \frac{N_I}{n_I}\sum_{i\in s_I}\frac{N/N_I}{m}\sum_{k\in s_i} y_k = \frac{N}{n_I m}Y_s = N\bar{y}_s.$$

La varianza (1.15) se reduce a

$$V_{2st} = N^2\left[\frac{1-f_I}{n_I}S^2_{yB} + \frac{1-f_{II}}{mn_I}S^2_{yW}\right], \tag{1.17}$$

donde

- $S^2_{yB} = \frac{1}{N_I-1}\sum_{i=1}^{N_I}(\bar{y}_{U_i} - \bar{y}_U)^2$ es la cuasivarianza poblacional entre[7] los conglomerados.
- $S^2_{yW} = \frac{1}{N-N_I}\sum_{i=1}^{N_i}\sum_{k\in U_i}(y_k - \bar{y}_{U_i})^2 = \frac{1}{\sum_{j=1}^{N_I}(N_j-1)}\sum_{i=1}^{N_I}(N_i-1)\left[\frac{1}{N_i-1}\sum_{k\in U_i}(y_k - \bar{y}_{U_i})^2\right]$ es la cuasivarianza poblacional combinada dentro[8] de los conglomerados.

Estas expresiones permiten expresar el efecto de diseño en términos del coeficiente de correlación intraconglomerados poblacional ρ (Lohr 2010). Adviértase que

$$S^2_{yU} = \frac{N_I-1}{N_I}\frac{N}{N-1}S^2_{yB} + \frac{N-N_I}{N-1}S^2_{yW},$$

de modo que, si $\rho = 1 - \frac{N}{N-1}\frac{S^2_{yW}}{S^2_{yU}}$ y $\delta = 1 - S^2_{yW}/S^2_{yU}$, podemos escribir

$$S^2_{yW} = \frac{N-1}{N}(1-\rho)S^2_{yU},$$

[7] *Between*, de ahí el subíndice B.
[8] *Within*, de ahí el subíndice W.

$$
\begin{aligned}
S^2_{yB} &= \left(\frac{N_I}{N}\right)^2 \frac{N-1}{N_I-1}\left[1+\frac{N-N_I}{N_I}\rho\right] S^2_{yU}, \\
V_{2st} &\approx N^2 S^2_{yU}\left[\frac{N_I}{N}\frac{1-f_I}{n_I}\left(1+\frac{N-N_I}{N_I}\rho\right)+\frac{1-f_{II}}{n_I\cdot m}(1-\rho)\right]. \qquad (1.18)
\end{aligned}
$$

El efecto de diseño del diseño bietápico con conglomerados de igual tamaño y probabilidades iguales será, por tanto,

$$
\begin{aligned}
\text{deff}_{2st} &= \frac{V_{2st}}{V_{srswor}} \\
&= \frac{f_{II}(1-f_I)}{1-f_If_{II}}\left(1+\frac{N-N_I}{N_I}\rho\right)+\frac{1-f_{II}}{1-f_If_{II}}(1-\rho), \qquad (1.19)
\end{aligned}
$$

donde $V_{srswor} = N^2\frac{1-f_If_{II}}{n_Im}S^2_{yU}$.

Obsérvese:

- Si $f_{II}=1$, esto es, en muestreo de conglomerados monoetápico, será $\text{deff}_{1st} = \left(1+\frac{N-N_I}{N_I}\rho\right)$, recuperando el resultado conocido.
- Si $\rho > 0$, esto es, si los elementos de la población dentro de un conglomerado tienden a ser similares, entonces el muestreo monoetápico es menos eficiente que el muestreo aleatorio simple. Por ello, es conveniente introducir la segunda etapa de selección, lo que se observa con el segundo término del lado derecho de la ec. (1.19). Esto suele suceder cuando los conglomerados se han formado de manera natural en la población, puesto que los elementos similares tienden a agruparse (mismas áreas geográficas como barrios, ciudades, etc.). El mecanismo de interacción entre los elementos de la población originan la similaridad entre ellos y, por tanto, la construcción de conglomerados naturales con $\rho > 0$. En el caso usualmente genérico en que $f_I \approx 0$, se obtiene $\text{deff}_{2st} \approx 1+\rho(m-1)$ y la afirmación es evidente.
- Si $\rho \approx 0$, esto es, si la variabilidad dentro de los conglomerados es similar a la variabilidad de los conglomerados en la población, entonces $\text{deff}_{2st} \approx 1$ y el muestreo de conglomerados es similar al muestreo aleatorio simple, como cabe esperar.
- Si $\rho < 0$, esto es, si los elementos dentro de cada conglomerado presentan más variabilidad que un grupo de elementos elegidos al azar, entonces el muestreo monoetápico es más eficiente que el muestreo aleatorio simple. Esto ocurre rara vez en los conglomerados formados de modo natural, pero sí puede suceder si se construyen artificialmente. Para el muestreo bietápico, también sucede así cuando $f_I \approx 0$, aunque en general para este caso los costes constituyen un factor muy importante.

El coeficiente de correlación intraconglomerados solo se define para conglomerados de igual tamaño. Una medida alternativa de homogeneidad es el coeficiente de homo-

geneidad $\delta = 1 - \frac{S^2_{yW}}{S^2_{yU}}$ (Särndal, Swensson y Wretman 1992) de los conglomerados (R^2 ajustado (Lohr 2010)). El análisis anterior puede repetirse en términos de δ. ■

Comentario 6. Es ilustrativo señalar las condiciones bajo las cuales cada componente de la varianza es cero:

(a) si $s_I = U_I$ con probabilidad 1, entonces $\pi_{Ii} = \pi_{Iij} = 1$ para todo i y todo j. Entonces, $V_{PSU} = 0$ y $V_{SSU} = \sum_{i \in U_i} V_i$. Esto es la varianza del estimador HT en el muestreo estratificado, donde los N_I PSUs constituyen el conjunto de estratos;

(b) si $s_i = U_i$ con probabilidad 1 para todo i, entonces $V_{SSU} = 0$. En este caso, V_{PSU} es la varianza del estimador HT del total poblacional en muestreo unietápico (véase el tema 7 de Producción Estadística Oficial del grupo de materias comunes).

■

Comentario 7. En la práctica muchas encuestas con diseño bietápico o multietápico hacen uso de los llamados *diseños autoponderados*[9]. Un diseño bietápico autoponderado consiste en lo siguiente. Sea u_i una medida del tamaño de la PSU i-ésima. En primera etapa se escoge un diseño muestral proporcional al tamaño de cada PSU de modo que $\pi_{Ii} \propto u_i$ y, en segunda etapa, se escoge un muestreo aleatorio simple con fracción de muestreo en cada conglomerado dada por $\frac{n_i}{N_i} = \frac{1}{u_i}$. Si en la primera etapa se emplea un diseño de tamaño muestral fijo n_I, entonces

$$\pi_{Ii} = \frac{u_i}{\sum_{i \in U_I} u_i} \cdot n_I.$$

Las probabilidades de inclusión de los elementos de la población se reducen entonces a

$$\pi_k = \pi_{Ii}\pi_{k|i} = \frac{u_i}{\sum_{i \in U_I} u_i} \cdot n_I \cdot \frac{n_i}{N_i} = \frac{n_I}{\sum_{i \in U_I} u_i}.$$

El estimador HT se simplifica notablemente:

$$\widehat{Y}_U^{\text{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k} = \frac{\sum_{i \in U_I} u_i}{n_I} \sum_{k \in s} y_k = \frac{\sum_{i \in U_I} u_i}{n_I} Y_s,$$

esto es, el estimador es proporcional al total muestral de la variable a estimar. Aparte de la simplificación computacional, este tipo de diseños permite un mayor control del trabajo de campo, pues si $\frac{N_i}{u_i} \approx K$, siendo K una constante para todo i, entonces $n_i \approx K$ y el número de entrevistas a realizar en cada conglomerado es básicamente el mismo. Esto permite un mayor control de la logística de la fase de recogida de datos (por ejemplo, asignando un entrevistador por conglomerado de modo que todos tengan la misma carga de trabajo). ■

[9] *Self-weighting design*.

Comentario 8. Conglomerados vs. estratos (2/2).
Visto el efecto que tiene en la varianza el muestreo de conglomerados, podemos ofrecer ahora una perspectiva más amplia de sus diferencias con el muestreo estratificado observando así las diferencias prácticas entre conglomerados y estratos. En el muestreo estratificado se seleccionan elementos de la población de todos los estratos mientras que en el muestreo de conglomerados se seleccionan elementos de una muestra de conglomerados. El objetivo fundamental del muestreo estratificado es aumentar la acuracidad de los estimadores mientras que del muestreo de conglomerados es reducir el coste y aumentar la eficiencia. En este sentido, los estratos se construyen artificialmente por el estadístico mientras que los conglomerados suelen formarse por mecanismos naturales de interacción entre los elementos de la población. Por este motivo, en el muestreo estratificado se observa un alto grado de homogeneidad entre los elementos de la población mientras que en el muestreo de conglomerados se observa homogeneidad entre los conglomerados. Análogamente, en el muestreo estratificado existe heterogeneidad entre los estratos mientras que en el muestreo de conglomerados esta heterogeneidad se da dentro de los conglomerados. ■

1.4 El diseño de una muestra de conglomerados

En la planificación y diseño de una estrategia muestral con conglomerados deben tomarse decisiones sobre cuatro cuestiones fundamentales (Lohr 2010):

1. ¿Qué precisión (acuracidad) quiere alcanzarse?
2. ¿De qué tamaño deben ser los conglomerados (PSUs)?
3. ¿Cuántas SSUs deben ser seleccionadas en segunda etapa en cada PSU seleccionada en primera etapa?
4. ¿Cuántas PSUs deben seleccionarse en primera etapa?

La cuestión 1 debe abordarse en cualquier diseño de una encuesta y, en particular, en su diseño muestral. El resto de cuestiones, como veremos, está muy asociado al coste de muestrear PSUs, al coste de muestrear SSUs en cada PSU y al grado de homogeneidad de los conglomerados (ρ o δ).

1.4.1 Tamaño de las PSUs

Dependiendo de la población de análisis, se pueden presentar dos tipos de conglomerados. Por un lado, aquellas poblaciones donde los conglomerados aparecen de modo natural, como, por ejemplo, en una encuesta de educación a estudiantes los conglomerados naturales pueden ser los colegios o en una encuesta agraria pueden las explotaciones agrícolas. Por otro lado, también existen poblaciones donde el estadístico tiene la oportunidad de escoger una variedad de conglomerados. Por ejemplo, en una encuesta sobre hábitos alimentarios los conglomerados pueden ser barrios, distritos municipales, municipios, etc.

Un principio general en encuestas en áreas geográficas es que a mayor tamño de las

PSUs, mayor variabilidad dentro de cada PSU. Por tanto, tendremos menor valor del coeficiente de correlación intraconglomerados ρ y del coeficiente de homogeneidad δ. Pero si el tamaño es demasiado grande, perdemos entonces el ahorro en costes asociado al muestreo por conglomerados.

Antes de escoger un valor para el tamaño de los conglomerados, es recomendable realizar estudios previos con información preliminar o auxiliar sobre la población. Es recomendable postular algún modelo que relaciona δ o ρ con el tamaño de los conglomerados.

1.4.2 Tamaño de las SSUs

Para escoger el tamaño de las SSUs debe establecerse un criterio. Como en todas las encuestas, el objetivo es obtener el máximo de información con el menor coste posible (en sentido amplio, esto es, no solo monetario, sino con menor carga de respuesta a los informantes o en el menor tiempo de difusión posible). Para ello, por tanto, es necesario considerar una función de coste. Para ser concretos, consideremos el caso del muestreo bietápico con conglomerados de igual tamaño y probabilidades iguales. Denotaremos por $M = N/N_I$ el tamaño de los conglomerados, respetando el resto de la notación introducida en el Ejemplo 3. El primer paso es considerar una función de coste. Por simplicidad, consideramos una función de coste lineal dada por

$$C(n_I, m) = c_I n_I + c_{II} n_I m, \tag{1.20}$$

de modo que c_I expresa el coste por PSU (sin considerar sus unidades secundarias) y c_{II} es el coste de medir cada SSU. Un criterio habitual similar a algunas afijaciones en muestreo estratificado es escoger n_I y m de modo que se minimice la varianza (máxima acuracidad) sujeto al coste expresado por la ec. (1.20). El problema, por tanto, formalmente se plantea como

$$\begin{array}{ll} \underset{n_I,m}{\text{minimizar}} & V_{2st}(n_I, m) = N^2 \left[\dfrac{1 - f_I}{n_I} S_{yB}^2 + \dfrac{1 - f_{II}}{m n_I} S_{yW}^2 \right] \\ \text{sujeto a} & C(n_I, m) = C_0. \end{array}$$

La solución a este problema de optimización está dada por

$$n_I^* = \frac{C_0}{c_I + c_{II} \cdot m^*}, \qquad m^* = \sqrt{\frac{c_I M (N_I - 1)(1 - \delta)}{c_{II}(N - 1)\delta}}.$$

El valor del coeficiente de homogeneidad δ suele estimarse a través de una encuesta piloto. Nótese que en poblaciones muy grandes, $\frac{M(N_I-1)}{(N-1)} \approx 1$, por tanto, podemos simplificar m^*:

$$\widehat{m}^* = \frac{c_I(1 - \widehat{\delta})}{c_{II} \cdot \widehat{\delta}},$$

donde $\widehat{\delta}$ es el valor estimado mediante la encuesta piloto.

Aunque el razonamiento anterior se ha realizado para el sencillo ejemplo de conglomerados de igual tamaño y con probabilidades iguales, también pueden plantearse situaciones más complejas de manera análoga permitiendo tamaños de conglomerados diferentes, así como tamaños muestrales en segunda etapa también diferentes.

1.4.3 Número de PSUs

Para estimar un número adecuado de PSUs nuevamente necesitamos fijar un criterio. Como en otros diseños muestrales, lo más adecuado es relacionar el tamaño muestral con la precisión del estimador. Esto puede abordarse de dos modos. En primer lugar, de modo directo, consideremos la expresión (1.17) para la varianza en muestreo bietápico de conglomerados de igual tamaño y con probabilidades iguales. Podemos entonces acotar

$$V_{2st}(n_I) \leq \frac{1}{n_I}\left[N^2 S_{yB}^2 + \frac{N^2}{m} S_{yW}^2\right] \equiv \frac{\nu}{n_I}.$$

Entonces, una medida del radio del intervalo de confianza para el estimador será $z_{\alpha/2}\sqrt{\frac{\nu}{n_I}}$. Por tanto, si fijamos la precisión mediante un valor e para este radio, el tamaño muestral en primera etapa será

$$n_I = \frac{z_{\alpha/2}^2 \nu}{e^2}.$$

Como es evidente, este enfoque supone tener conocimiento del valor de ν, normalmente a través de una encuesta piloto o de encuestas anteriores.

En segundo lugar, de modo algo más indirecto, la estrategia puede ser relacionar el tamaño muestral, la precisión y el efecto de diseño deff. Para ello, teniendo en mente el significado del efecto de diseño, podemos estimar el tamaño muestral n_{srswor} si el diseño fuese aleatorio simple y, posteriormente, multiplicar este valor por el valor deff.

1.5 La estimación de la media poblacional

Como se adelantó anteriormente, en el caso de muestreo bietápico la estimación de la media poblacional $\bar{y}_U = Y_U/N$ no puede abordarse directamente a partir del Teorema 1 dividiendo el estimador HT por N, puesto que por las condiciones del muestreo N no es conocido.

Tanto para la media como el total poblacional, la construcción de estimadores puede variar dependiendo de la información auxiliar disponible, que puede ser a nivel de los conglomerados, a nivel de los elementos de la población o ambas. El estimador generalizado de regresión y la calibración de pesos de muestreo son técnicas habituales para la incorporación de esta información auxiliar dependiendo de cada circunstancia.

En lo que sigue, tan solo asumiremos que conocemos el tamaño de los conglomerados N_{U_i} para todo $i \in S_I$. La media poblacional es $\bar{y}_U = \frac{Y}{U}N$, donde tanto Y_U como N pueden considerarse totales poblacionales a estimar. En estas condiciones, puede demostrarse el siguiente resultado:

Teorema 3

En un muestreo bietápico de elementos, un estimador aproximadamente insesgado está dado por

$$\widehat{\bar{y}}_U^{\text{Rat}} = \frac{\widehat{Y}_U^{\text{HT}}}{\widehat{N}_{U_I}^{\text{HT}}}, \tag{1.21}$$

donde

$$\begin{aligned} \widehat{Y}_U^{\text{HT}} &= \sum_{i \in s_I} \frac{\sum_{k \in s_i} \frac{y_k}{\pi_{k|i}}}{\pi_{Ii}}, \\ \widehat{N}_{U_I}^{\text{HT}} &= \sum_{i \in s_I} \frac{N_{U_i}}{\pi_{Ii}}. \end{aligned}$$

La varianza aproximada de $\widehat{\bar{y}}_U^{\text{Rat}}$ es

$$\mathbb{V}\left[\widehat{\bar{y}}_U^{\text{Rat}}\right] \approx \frac{1}{N^2}\left(\sum_{i \in U_I}\sum_{j \in U_I} \Delta_{Iij} \frac{N_{U_i}(\bar{y}_{U_i} - \bar{y}_U)}{\pi_{Ii}} \frac{N_{U_j}(\bar{y}_{U_j} - \bar{y}_U)}{\pi_{Ij}} + \sum_{i \in U_I} \frac{V_i}{\pi_{Ii}}\right) \tag{1.22}$$

El estimador aproximadamente insesgado de la varianza es

$$\widehat{\mathbb{V}}\left[\widehat{\bar{y}}_U^{\text{Rat}}\right] \approx \frac{1}{\left(\sum_{i \in s_I} \frac{N_{U_i}}{\pi_{Ii}}\right)^2}\left(\sum_{i \in s_I}\sum_{j \in s_I} \check{\Delta}_{Iij} \frac{\widehat{Y}_{U_i}^{\text{HT}} - N_{U_i}\widehat{\bar{y}}_U^{\text{Rat}}}{\pi_{Ii}} \frac{\widehat{Y}_{U_j}^{\text{HT}} - N_{U_j}\widehat{\bar{y}}_U^{\text{Rat}}}{\pi_{Ij}} + \sum_{i \in s_I} \frac{\widehat{V}_i}{\pi_{Ii}}\right), \tag{1.23}$$

donde V_i está dado por (1.3) y $\widehat{V}_i$ está dado por (1.4).

Demostración 3

La demostración es una aplicación directa del estimador de razón aplicado a los estimadores $\widehat{Y}_U^{\text{HT}}$ y $\widehat{N}_{U_I}^{\text{HT}}$ (véase el tema 3). ■

Las consideraciones realizadas anteriormente para la estimación de totales poblacionales siguen vigentes para la estimación de medias poblacionales. El Teorema anterior también puede demostrarse empleando resultados generales de estimadores generalizados de regresión (GREG) cuando se dispone de información auxiliar $\mathbf{u}_i$ para todos

los conglomerados muestreados $i \in s_I$ y del valor poblacional $\sum_{i \in U_I} \mathbf{u}_i$ (véase Särndal, Swensson y Wretman 1992, pág. 313).

1.6 Simplificación de la estimación de la varianza

Históricamente existen resultados teóricos para simplificar el cálculo de la estimación de la varianza debido a su complejidad computacional cuando estos recursos en las décadas antes de los noventa. El cálculo de la estimación de la varianza a partir de la fórmula (1.11) resultaba complicado, especialmente debido a que hay que calcular la estimación de la varianza $\widehat{V}_i$ para cada $i \in s_I$. Por ello, a veces era deseable disponer de una estimación de la varianza más sencilla. Estas simplificaciones hoy día están menos justificadas desde el punto de vista computacional, pero siguen siendo interesantes desde la perspectiva teórica para ilustrar cómo la introducción de un pequeño sesgo de manera controlada en los estimadores es a menudo beneficioso.

Consideremos tan solo el primer término de (1.11), es decir,

$$\widehat{V}^* = \sum_{i \in s_I} \sum_{j \in s_j} \breve{\Delta}_{Iij} \frac{\widehat{Y}^{\text{HT}}_{U_i|i}}{\pi_{Ii}} \frac{\widehat{Y}^{\text{HT}}_{U_j|j}}{\pi_{Ij}} \tag{1.24}$$

Las cantidades $\breve{\Delta}_{Iij}$ vienen determinadas por el diseño de primera etapa y la otra única información necesaria para este estimador simplificado de la varianza son los totales estimados de los PSU. Tenemos, a partir de (1.13a) en la demostración del Teorema 2

$$\mathbb{E}\left[\widehat{V}^*\right] = \widehat{V}_{2st} - \sum_{i \in U_I} V_i$$

Por tanto, el sesgo de $\widehat{V}^*$ viene dado por

$$\mathbb{B}\left[\widehat{V}^*\right] = -\sum_{i \in U_I} V_i, \tag{1.25}$$

lo que significa que $\widehat{V}^*$ subestima la varianza desconocida de $\widehat{Y}^{\text{HT}}_U$. Esta propiedad no deseable puede llevar a un excesivo optimismo al juzgar la precisión de la estimación. Sin embargo, un vistazo al sesgo relativo del estimador de la varianza

$$\frac{\mathbb{B}\left[\widehat{V}^*\right]}{V_{2st}} = -\frac{\sum_{i \in U_I} V_i}{\sum_{i \in U_I} \sum_{j \in U_I} \Delta_{Iij} \frac{\widehat{Y}^{\text{HT}}_{U_i|i}}{\pi_{Ii}} \frac{\widehat{Y}^{\text{HT}}_{U_j|j}}{\pi_{Ij}} + \sum_{i \in U_I} \frac{V_i}{\pi_{Ii}}} \tag{1.26}$$

muestra que la subestimación obtenida de $\widehat{V}^*$ puede, en muchos casos, no ser importante. El numerador en la expresión (1.26) a menudo será pequeña comparada con el denominador si las probabilidades π_{Ii} son pequeñas y, en consecuencia, la subestimación será despreciable. Por ejemplo, supongamos que los conglomerados se seleccionan con

muestreo aleatorio simple con fracción de muestreo en primera etapa $\frac{n_I}{N_I} = 10\,\% = \pi_{Ii}$. Si $\frac{V_{PSU}}{V_{SSU}} = 4$, entonces el sesgo relativo obtenido de la ecuación (1.26) es sólo $-\frac{1}{50} = -0{,}02$.

1.7 Muestreo con reemplazamiento en primera etapa

Además de la simplificación directa de la varianza considerada en la sección anterior, otro procedimiento para rebajar los costes computacionales relacionados con la estimación de la varianza es emplear diseños muestrales con reemplazamiento en primera etapa. Esto produce ciertamente cierta disminución de la eficiencia del muestreo (por seleccionar conglomerados repetidos), pero conlleva también una reducción de la carga computacional.

Para ello consideremos el siguiente tipo de diseño muestral:

i. En la primera etapa de muestreo, se selecciona una muestra ordenada de conglomerados (PSUs)
$$os_I = \{i_1, \ldots, i_\nu, \ldots, i_{n_I}\}$$
siguiendo un esquema de muestreo con reemplazamiento tal que, en cada extracción[10], la probabilidad de seleccionar el conglomerado i, $i \in \{1, \ldots, N_I\}$, es p_i.

ii. En la segunda etapa, se satisfacen las mismas propiedades de invarianza e independencia que en las secciones anteriores.

iii. Si un conglomerado (PSU) es seleccionado más de una vez, se submuestrea independientemente tantas veces como haya sido seleccionado.

Denotemos por $\widehat{Y}^{\text{HT}}_{U_{i_\nu}|os_I}$ el estimador HT del total poblacional $Y_{U_{i_\nu}}$ en el conglomerado U_{i_ν}; por $V_{i_\nu} = \mathbb{V}\left[\widehat{Y}^{\text{HT}}_{U_{i_\nu}|os_I}|os_I\right]$ la varianza del estimador $\widehat{Y}^{\text{HT}}_{U_{i_\nu}|os_I}$ condicionada al conglomerado seleccionado en primera etapa. Entonces, podemos demostrar el siguiente resultado.

Teorema 4

En un muestreo bietápico bajo las condiciones i, ii y iii anteriores, un estimador insesgado para el total poblacional Y_U está dado por

$$\widehat{Y}^{HH}_U = \frac{1}{n_I}\sum_{\nu=1}^{n_I}\frac{\widehat{Y}^{\text{HT}}_{U_{i_\nu}|os_I}}{p_{i_\nu}}. \tag{1.27}$$

La varianza de $\widehat{Y}^{HH}_U$ está dada por

[10] *Draw.*

$$\mathbb{V}\left[\widehat{Y}_U^{HH}\right] = \frac{1}{n_I}\sum_{i=1}^{N_I} p_i\left(\frac{Y_{U_i}}{p_i} - Y_U\right)^2 + \frac{1}{n_I}\sum_{i=1}^{N_I}\frac{V_i}{p_i}. \tag{1.28}$$

Un estimador insesgado para esta varianza está dado por

$$\widehat{\mathbb{V}}\left[\widehat{Y}_U^{HH}\right] = \frac{1}{n_I(n_I-1)}\sum_{\nu=1}^{n_I}\left[\frac{\widehat{Y}^{\text{HT}}_{U_{i_\nu}|os_I}}{p_{i_\nu}} - \widehat{Y}_U^{HH}\right]^2. \tag{1.29}$$

Demostración 4

Al tratarse de una muestra ordenada en primera etapa podemos definir las variables aleatorias $Z_\nu = \frac{Y_{i_\nu}}{p_{i_\nu}}$ y $\widehat{Z}_\nu = \frac{\widehat{Y}^{\text{HT}}_{U_{i_\nu}}}{p_{i_\nu}}$. Puesto que son variables aleatorias independientes e idénticamente distribuidas con función de masa de probabilidad p_i, se cumplen

$$\mathbb{E}\left[\widehat{Z}_\nu\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{Z}_\nu|os_I\right]\right] = \mathbb{E}\left[Z_\nu\right] = Y_U$$

y

$$\begin{aligned}\mathbb{V}\left[\widehat{Z}_\nu\right] &= \mathbb{V}\left[\mathbb{E}\left[\widehat{Z}_\nu|os_I\right]\right] + \mathbb{E}\left[\mathbb{V}\left[\widehat{Z}_\nu|os_I\right]\right] \\ &= \mathbb{V}\left[Z_\nu\right] + \mathbb{E}\left[\frac{V_{i_\nu}}{p_{i_\nu}^2}\right] \\ &= \sum_{i=1}^{N_I} p_i\left(\frac{Y_{U_i}}{p_i} - Y_U\right)^2 + \sum_{i=1}^{N_I}\frac{V_i}{p_i}.\end{aligned}$$

Como $\widehat{Y}_U^{HH}$ es la media de n_I variables aleatorias $\widehat{Z}_\nu$ independientes distribuidas idénticamente, el Teorema se sigue de resultados genéricos de la media de este tipo de variables aleatorias. ■

El muestreo bietápico con reemplazamiento presenta la ventaja de que resulta fácil seleccionar la muestra y obtener estimaciones para el total poblacional y su varianza. Sin embargo, si el número de estratos es pequeño (como sucede en muchas encuestas complejas altamente estratificadas donde cada estrato tiene pocos conglomerados), el muestreo con reemplazamiento puede ser mucho menos eficiente que el muestreo sin reemplazamiento.

Bibliografía

Cochran, W.G. (1977). *Sampling Techniques*. 3rd. New York: Wiley.

Grimmet, G.R. y D.R. Stirzaker (2004). *Probability and random processes*. 3rd. Oxford Science Publications.
Kish, L. (1965). *Survey Sampling*. New York: Wiley.
Lehtonen, R. y E. Pahkinen (2004). *Practical methods for design and analysis of complex surveys*. 2nd. Chichester: Wiley.
Lohr, S. (2010). *Sampling: Design and Analysis*. 2nd. Duxbury Press.
Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.

Tema 2

Diseños para encuestas a lo largo del tiempo. Encuestas repetidas. Encuestas de panel rotante. Encuestas de panel

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía.

G. Kalton, *Designs for surveys over time*, en D. Pfefferman y C.R. Rao (2009). *Handbook of Statistics 29A*. North-Holland, Amsterdam: Elsevier, cap. 5, pp. 89-108

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

2.1 Diseños para encuestas a lo largo del tiempo

La mayoría de la literatura sobre metodología de encuestas se centra en encuestas que se diseñan para producir una imagen de la población en un momento de tiempo. Sin embargo, en la práctica, los investigadores a menudo están interesados en obtener un vídeo de los cambios que tienen lugar a lo largo del tiempo.

La dimensión temporal se puede introducir reproduciendo la encuesta en distintos momentos de tiempo o usando alguna forma de diseño panel. En este tema veremos las distintas opciones de *diseño* disponibles para las encuestas a lo largo del tiempo y una breve descripción de algunas dificultades que nos podemos encontrar.

Muchas operaciones estadísticas [1] buscan estimar características de una población en un instante específico de tiempo. Por ejemplo, la operación **Cifras oficiales de población de los municipios españoles: Revisión del Padrón Municipal** [2] proporciona una imagen de la población con referencia a 1 de enero de cada año. En la práctica, las

[1] Aunque en la fuente utiliza se habla sobre *surveys* que son encuestas, nosotros nos referiremos a operaciones estadísticas en general

[2] https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177011&menu=resultados&idp=1254734710990

operaciones necesarias para obtener la información sobre el Padrón no se pueden llevar a cabo en un único día, pero el objetivo es que represente a la población en esa fecha.

A menudo los datos recogidos no se refieren a una fecha específica sino a un periodo de tiempo, por ejemplo la situación laboral en una semana, las enfermedades sufridas el mes pasado, o los gastos en los últimos seis meses. Sin embargo, el objetivo de estas *encuestas trasversales* es el de recoger los datos necesarios para describir y analizar características de la población en un instante de tiempo.

Este interés de las operaciones estadísticas trasversales en un determinado instante de tiempo es importante porque tanto las características como la composición de una población cambian con el tiempo. Las estimaciones, por tanto, se refieren a un instante específico de tiempo, un rasgo que es particularmente importante en algunos contextos. Por ejemplo, la tasa de desempleo es un indicador económico fundamental que varía en el tiempo; la tasa puede cambiar de un mes al siguiente debido a un cambio en la economía (con empresas despidiendo a empleados o contratando nuevo personal) y/o debido a un cambio en el mercado laboral (por ejemplo, al finalizar el curso académico, que es cuando los graduados empiezan a buscar trabajo).

Los cambios en las características de una población a lo largo del tiempo plantean cuestiones para el análisis. Por un lado, los políticos necesitan estimar determinadas características de la población de forma reiterada en el tiempo para obtener estimaciones lo más recientes posibles. Pero también están interesados en conocer el cambio que ha tenido lugar en la estimaciones a lo largo del tiempo: ¿La tasa de desempleo ha aumentado o ha disminuido desde la encuesta anterior? Este cambio se denomina *cambio neto* y refleja los cambios tanto en las características como en la composición de la población.

Un análisis más detallado implica conocer las componentes del cambio. ¿Hasta qué punto el cambio (o la ausencia de cambios) se debe a la dinámica de la población, con gente entrando en la población mediante los 'nacimientos' (por ejemplo, con gente que cumple los 16 años o inmigrantes) y saliendo de la población a través de las 'muertes' (por ejemplo, gente que fallece, emigra o se jubila)? ¿Hasta qué punto el cambio se debe a cambios en los estados de las personas de la población? Más aún, ¿cómo funciona el cambio en casos en los que hay un cambio de estado? Por ejemplo, asumiendo que no hay dinámicas poblacionales, si la tasa de desempleo sufre un incremento neto del 1 %, esto se debe a que un 1 % de las personas previamente empleadas perdieron su trabajo o a que, por ejemplo, un 10 % perdió su empleo y un 9 % de estos desempleados encontraron trabajo?

La descomposición del cambio neto en sus dos componentes lleva a medir el *cambio bruto*. Mientras que el cambio neto se puede medir usando muestras separadas para las dos ocasiones, medir el cambio bruto requiere medidas repetidas en la misma muestra, o por lo menos en una submuestra representativa.

Hay dos grandes clases de objetivos para las encuestas a lo largo del tiempo, y dan lugar a distintos enfoques en el diseño de encuestas. En muchos casos, los objetivos se restringen a estimar los parámetros poblacionales o económicos en distintos instantes de tiempo así como los cambios netos, tendencias o valores medios a lo largo de un periodo. Ninguno de estos objetivos requiere mediciones repetidas sobre la misma muestra. En particular, estos objetivos se pueden alcanzar con muestras totalmente independientes en cada instante de tiempo. También con muestras que se construyen para minimizar el solapamiento de la muestra a lo largo del tiempo y la carga de respuesta al informante. Este tipo de *encuestas repetidas* se ven en la Sección 2.2.

Algunos objetivos también se pueden alcanzar con diseños panel, que incluyen a algunos o a todos los miembros de la muestra en distintos instantes de tiempo. De hecho, la precisión de las estimaciones transversales y los cambios netos se pueden mejorar usando un diseño muestral de *panel rotante* que cree algún grado de solapamiento en la muestra a lo largo del tiempo. Los diseños de panel rotante también se pueden usar para eliminar los efectos telescópicos o de memoria que ocurren cuando los informantes proporcionan información errónea sobre cuándo tuvo lugar un evento. Los paneles rotantes se discuten en la Sección 2.3.

Otros objetivos se centran en la estimación del cambio bruto y en otras componentes de cambio individual, y en la agregación de respuestas, por ejemplo, los gastos, de individuos a lo largo del tiempo. Estos objetivos pueden satisfacerse sólo con algunas formas de encuestas panel que recogen datos de los mismos individuos para el periodo de interés. Las *encuestas panel o longitudinales* se discuten en la Sección 2.4.

Otros objetivos de las encuestas a lo largo del tiempo están relacionados con la obtención de estimaciones para poblaciones raras (es decir, un subconjunto de la población general que tiene una característica rara). Uno de esos objetivos es acumular una muestra de casos con esa característica rara a lo largo del tiempo. Si la característica es un evento, como divorciarse, entonces este objetivo puede satisfacerse con cualquiera de los diseños. Sin embargo, si la característica es estable, como por ejemplo pertenecer a un determinado grupo étnico, la acumulación sólo funciona cuando se incluyen nuevas muestras a lo largo del tiempo. En cualquier caso, los analistas necesitan reconocer que las características de una población rara puede variar a lo largo del tiempo.

Un objetivo distinto con una población rara es la producción de estimaciones para esa población en distintos instantes de tiempo. Si la característica rara es estable, se puede identificar una muestra de miembros de esa población en un instante de tiempo y luego volver a esa muestra de forma reiterada en un diseño de panel.

En la última sección del tema, Sección 2.5, se resumirán los aspectos a considerar al elegir el tipo de diseño a adoptar para encuestar una población a lo largo del tiempo.

También se resumirán las dificultades metodológicas que se encuentran al producir conclusiones válidas de encuestas a lo largo del tiempo.

2.2 Encuestas repetidas

En esta sección se discuten varios aspectos y diseños para encuestas a lo largo del tiempo cuando el análisis se centra en la producción de series de estimaciones transversales que se pueden usar para analizar los cambios netos y las tendencias a nivel de agregados. Los diseños que se consideran a continuación no están estructurados de forma que permitan análisis longitudinales a nivel de las unidades elementales.

Una forma común de encuestas repetidas es aquélla en la que se seleccionan muestras separadas de las últimas unidades muestrales cada vez. Cuando el tiempo que transcurre entre las rondas de una encuesta repetida es grande, digamos 5-10 años, la selección de muestras totalmente independientes puede ser una estrategia efectiva.

Sin embargo, en encuestas repetidas con diseños muestrales multietápicos y con periodos de tiempo más pequeños entre rondas se pueden lograr beneficios manteniendo las mismas unidades muestrales de primera etapa (PSUs del inglés *primary sampling units*), y quizás también unidades en etapas posteriores (pero no las unidades de última etapa), en cada ronda [3]. Muestras maestras de PSUs son muy utilizadas para encuestas a hogares, tanto en las encuestas repetidas sobre un determinado tema como para encuestas que tratan diversos temas.

El solapamiento de unidades muestrales para niveles más altos también da lugar a estimaciones más precisas del cambio neto a lo largo del tiempo. Sin embargo, una muestra maestra deja de ser eficiente con el paso del tiempo, al cambiar la población. Cuando se dispone de un marco actualizado y se observa un cambio substancial en la población, es necesario modificar los tamaños de los PSUs y la estratificación. Hay una gran variedad de métodos para mantener tantos PSUs muestrales como sea posible en la nueva muestra mientras se actualizan los tamaños y estratos.

En el caso de encuestas económicas repetidas, se puede usar una metodología basada en alguna forma de números aleatorios permanentes (PRNs del inglés *permanent random numbers*). En esencia, la metodología consiste en asignar un número aleatorio entre 0 y 1 a cada elemento poblacional del marco. A continuación, se puede seleccionar fácilmente una muestra estratificada no proporcionada incluyendo todos los elementos con números aleatorios menores que la fracción muestral en cada estrato. Los valores aleatorios asignados no cambian con el tiempo, dando como resultado que un elemento seleccionado en la muestra en una determinada ronda de una encuesta repetida se mantendrá con una probabilidad bastante alta en las muestras de otras rondas para las cuales la probabilidad de selección no es inferior que aquélla en la que se le asignó el

[3] waves se traducirá por rondas

valor aleatorio.

Este procedimiento flexible automáticamente incluye los cambios en los tamaños de los estratos muestrales y de la muestra en general entre rondas, los elementos que cambian de estratos, los nacimientos y las muertes. Está pensado sobre todo para mejorar la precisión de las estimaciones de cambios entre rondas y a veces para facilitar la recogida de datos, pero también se puede usar para generar una muestra panel para un periodo dado. Los elementos en la muestra para todas las rondas de un periodo dado constituyen la muestra probabilística de elementos que existen a lo largo del periodo, siendo la probabilidad de un elemento de estar en el panel el mínimo de sus probabilidades de selección entre las rondas.

La metodología PRN también se puede modificar para proporcionar una rotación muestral que limite la carga de respuesta de las empresas de la muestra, de forma particular para las pequeñas empresas para las que las probabilidades de selección son pequeñas. Por ejemplo, el PRN de cada empresa se puede aumentar en, por ejemplo, 0,1 en cada ronda y tomando la parte decimal si el resultado excede 1.

Un objetivo crítico cuando se usan las encuestas repetidas es la producción de estimaciones de tendencias válidas a lo largo del periodo de interés, de forma particular cambios de una ronda a la siguiente. Sin embargo, los cambios en el diseño de la encuesta a menudo son convenientes, y desgraciadamente incluso pequeños cambios de diseño pueden afectar a las estimaciones.

Por tanto, los cambios en la redacción de las preguntas o en el contenido del cuestionario, modo de recogida, entrenamiento del encuestador, el marco muestral, procedimientos de codificación, imputación o ponderación pueden afectar y amenazar la validez de las estimaciones de la tendencia.

Está bien documentado el hecho de que cambios en el contenido de un cuestionario pueden implicar efectos en el contexto, que pueden distorsionar las estimaciones de las tendencias (Biemer y col. 1991; Tourangeau, Rips y Rasinski 2000), e incluso el significado de preguntas idénticas puede cambiar a lo largo del tiempo (Kulka 1982).

Incluso un incremento en el tamaño muestral puede afectar a las estimaciones debido a la necesidad de disponer de nuevos entrevistadores y quizá porque sea necesario un esfuerzo para obtener las respuestas. Los que realizan encuestas repetidas a menudo se ven ante el dilema de si mejorar los procedimientos basándose en las experiencias pasadas, en estudios metodológicos, cambios en la población y temas de interés, o no hacer ningún cambio y mantener las estimaciones válidas.

Cuando resulta necesario hacer un cambio metodológico significante, una práctica común es realizar una encuesta puente para uno o más periodos, es decir, realizar

una parte de la encuesta usando los métodos antiguos y otra parte usando los nuevos métodos de forma simultánea. Por ejemplo, si se va a realizar un cambio de modo de recogida para una operación estadística, se puede seleccionar una parte de la muestra que conteste tanto con el antiguo, por ejemplo PAPI y presenciales, como con el nuevo, por ejemplo, CAPI y CATI.

A veces se combinan los datos recogidos en varias rondas de encuestas repetidas para producir muestras más grandes y de esta forma reducir los errores de muestreo, en particular para estimaciones de subgrupos de poblaciones pequeñas (Kish 1999). También se puede repartir la recogida de datos de una encuesta a lo largo del tiempo para facilitar el trabajo de campo. En este caso, la muestra de la encuesta se puede obtener como un conjunto de repeticiones, cada una de las cuales puede producir estimaciones, aunque menos precisas, para la población completa.

2.3 Encuestas de panel rotante

Las encuestas repetidas a hogares pueden mantener el mismo conjunto de PSUs, unidades muestrales de segunda etapa y unidades en otras etapas de muestreo de una ronda a la siguiente, pero seleccionando nuevas muestras de informantes en cada ocasión[4]. Por contra, las encuestas de panel rotante se diseñan para asegurar algún grado de solapamiento en las unidades muestrales finales en rondas específicas. Sin embargo, a diferencia de las encuestas de panel, no todas las mismas unidades muestrales finales se mantienen en todas las rondas. Con un diseño rotante, cada unidad muestral final se mantiene en la muestra sólo durante un periodo limitado de tiempo.

Los diseños de panel rotante son muy utilizados en la encuestas de población activa. Por ejemplo, en la Encuesta de Población Activa realizada por el INE [5], una vez una vivienda ha sido seleccionada para formar parte de la encuesta, permanece en la misma durante seis trimestres consecutivos, al cabo de los cuales se sustituye por otra de la misma sección. Cada trimestre, un sexto de las viviendas entra a formar parte de la muestra y un sexto sale de la misma.

La Tabla 2.1 muestra el esquema de rotación durante un periodo de tres años. La muestra en el primer trimestre está formada por seis grupos de rotación, cada uno de los cuales contiene un sexto del total de la muestra. El grupo de rotación A ha estado en la muestra los cinco trimestres anteriores y sale en el siguiente trimestre, el grupo de rotación B ha estado en la muestra los cuatro trimestres anteriores y permanecerá en la muestra dos trimestres más antes de salir, etcétera.

Varios diseños de panel rotante se usan para encuestas de población activa en distintos

[4] De hecho, las encuestas repetidas se pueden diseñar de forma que minimicen la probabilidad de que una unidad informante sea seleccionada en rondas muy cercanas en el tiempo.

[5] https://ine.es/inebaseDYN/epa30308/docs/resumetepa.pdf

T1A1	T2A1	T3A1	T4A1	T1A2	T2A2	T3A2	T4A2	T1A3	T2A3	T3A3	T4A3
A	G	G	G	G	G	G	M	M	M	M	M
B	B	H	H	H	H	H	H	N	N	N	N
C	C	C	I	I	I	I	I	I	O	O	O
D	D	D	D	J	J	J	J	J	J	P	P
E	E	E	E	E	K	K	K	K	K	K	Q
F	F	F	F	F	F	L	L	L	L	L	L

Tabla 2.1: Ejemplo del esquema de rotación de la EPA durante un periodo de 3 años

países. Por ejemplo, en la encuesta mensual de Población Activa de Canadá cada vivienda permanece en la muestra durante seis meses consecutivos, la encuesta trimestral de Población Activa en Reino Unido usa un panel rotante en cinco etapas y la encuesta mensual de Población Activa de Estados Unidos usa un esquema de rotación más complejo, cada unidad permanece en la muestra durante ocho meses, pero no consecutivos, ya que está en la muestra durante cuatro meses, sale ocho, y vuelve a entrar otros cuatro meses consecutivos, es lo que se denomina un esquema de rotación 4-8-4.

Los objetivos principales de estas encuestas de población activa son las mismas que las de las encuestas repetidas: producir estimaciones trasversales en cada momento de tiempo y medir los cambios netos a lo largo del tiempo. Comparadas con muestras independientes en cada ronda, un diseño rotante induce una correlación entre estimaciones en rondas en las que hay algún solapamiento. Puesto que esta correlación es casi siempre positiva, el solapamiento da como resultado una reducción en el error de muestreo de las estimaciones de cambio neto.

El patrón de rotación 4-8-4 utilizado en la EPA de Estados Unidos, por ejemplo, está diseñado para proporcionar un solapamiento importante de un mes al siguiente y también de un mes determinado en un año con el mismo mes del año siguiente. Es más, con un diseño de panel rotante, la precisión de las estimaciones del nivel actual y del cambio neto pueden 'tomar prestada fortaleza' de los datos recogidos en todas las rondas anteriores de la encuesta, usando la técnica de estimaciones compuestas.

Un diseño de panel rotante puede dar lugar no sólo a mejoras en la precisión de las estimaciones sino también a una reducción de coste ya que entrevistar a las mismas unidades a menudo es más barato que empezar de nuevo. En particular, mientras que la primera entrevista puede ser necesario realizarla presencialmente, las entrevistas posteriores se pueden realizar por teléfono. Esto es lo que ocurre en España, Canadá y Reino Unido.

Hay, sin embargo, una preocupación que de las respuestas obtenidas por teléfono no sean comparables con aquéllas obtenidas mediante entrevistas presenciales. Y, en general, por los efectos de la recogida de datos usando distintos modos de recogida entre

las etapas de la encuestas panel. Véase de Leeuw 2005 sobre los efectos de recogida de datos con distintos modos en general y Dillman y Christian 2005 sobre el uso de varios modos entre etapas de una encuesta panel.

Se han hecho estudios sobre este tema para analizar los efectos del cambio teniendo también en cuenta el sesgo de los grupos de rotación y se ha concluido que el cambio en los métodos de recogida de datos tiene efectos transitorios en las estimaciones, pero este efecto prácticamente desparece hacia el final del periodo de transición.

Incluso sin tener en cuenta los cambios de modo, hay cierta preocupación de que las respuestas obtenidas en entrevistas repetidas con el mismo informante puedan no ser comparables. Este efecto, que se denomina *condicionamiento de panel*, ocurre cuando las respuestas en etapas posteriores se ven afectadas por la participación del informante en etapas anteriores de la encuesta. Por ejemplo, los informantes pueden cambiar su actitud al verse sensibilizados por el tema de la encuesta y quizá por aprender algo a lo largo de la entrevista (por ejemplo, la existencia de un programa de ayuda social). Algunos informantes pueden cambiar sus conductas de respuestas en entrevistas posteriores, quizá demostrando una mejor memoria después de aprender más sobre los contenidos de la encuesta, estando más motivados a dar respuestas más precisas, dando respuestas menos consideradas por haber perdido interés, o respondiendo a las preguntas filtro de forma que eviten conjuntos largos de preguntas de seguimiento.

Los informantes también pueden pretender ser excesivamente consistentes en las respuestas sobre ítems de opinión. Véase Waterton y Lievesley 1989 y Sturgis, Allum y Brunton-Smith 2009 para una discusión sobre las posibles razones para los efectos de condicionamiento de panel y Cantor 2007 para una revisión más completa sobre el condicionamiento de panel.

Otro asunto a tener en cuenta en los diseños panel, en general, es el *cansancio por el panel*, ver también la Sección 2.4. Mientras que, tanto las encuestas trasversales como las panel están sujetas a falta de respuesta total en la ronda inicial de la recogida de datos, una encuesta de panel también sufre bajas en rondas posteriores. Aunque los pesos de ajuste por falta de respuesta puede ayudar a compensar el cansancio por el panel, las estimaciones derivadas de un grupo de rotación que ha estado en la muestra durante varias rondas, con cansancio por el panel asociado, puede diferir por este motivo de aquellas derivadas de grupos de rotación que han estado en la muestra durante periodos más cortos.

Por último, está la combinación de los efectos de condicionamiento de panel y de cansancio por el panel, que se conoce como el *sesgo por grupo de rotación*, el *sesgo del tiempo en muestra* o el *sesgo por mes en la muestra*. La existencia de este sesgo implica que la estimación en un mes dado está sesgada. Sin embargo, bajo un modelo aditivo para el sesgo por grupo de rotación, las estimaciones de los cambios mensuales están insesgadas ya que el patrón del grupo de rotación está compensado en cada instante

de tiempo (Bailar 1975). Este balance no siempre se consigue; en cualquier caso, no se consigue durante el periodo inicial del diseño de panel rotante. También hay que tener en cuenta que las estimaciones están sesgadas bajo un modelo que incluya un término multiplicativo de sesgo.

Los diseños de panel rotante no se restringen a las encuestas de población activa. También se usan para mejorar la eficiencia en la recogida y la precisión de las estimaciones de cambio y nivel. Una razón importante para usar un diseño de panel rotante es para solucionar el efecto telescópico. Este efecto tiene lugar cuando a los informantes se les pregunta por sucesos que han ocurrido en un periodo determinado y proporcionan información sobre algo ocurrido fuera del periodo de manera inconsciente. Véase Neter y Waksberg 1964a y Neter y Waksberg 1964b para un estudio clásico tanto sobre el efecto telescópico como sobre condicionamiento de panel. Este efecto se puede solucionar con una encuesta de panel rotante, de hecho con cualquier encuesta panel en la cual la muestra es entrevistada nuevamente en intervalos correspondientes con el periodo de referencia: los eventos notificados en la ronda actual que también se modificaron en rondas anteriores se pueden eliminar porque ocurrieron con antelación al periodo de referencia actual. Véase también en la Sección 2.4 las entrevistas dependientes.

Otra aplicación de los diseños de panel rotante es cuando no cabe esperar que los informantes se acuerden de forma precisa de toda la información solicitada para un periodo dado de referencia. Pueden volver a ser entrevistados en intervalos de tiempo para proporcionar la información de periodos de referencia más cortos, siendo luego agregada la información para proporcionar los datos necesarios para el periodo completo.

Un diseño de panel rotante a veces puede proporcionar los datos necesarios para análisis longitudinales durante un periodo de tiempo limitado. Y hay que tener en cuenta la unidad muestral de la encuesta ya que, por ejemplo, en las encuestas de población activa la unidad muestral es la vivienda y no los miembros de la misma ni el hogar. De esta forma, no es necesario entrevistar a los miembros del hogar ni usar los hogares como unidades, pero por contra no sirve para estudios longitudinales de hogares ni de personas.

Un aspecto general de las estimaciones trasversales de encuestas panel es que, a menos que se tomen medidas especiales, la muestra en rondas posteriores no es representativa de los elementos que han entrado en la población después de que la muestra fue seleccionada para la ronda inicial. Raramente es esto un problema para los paneles rotantes ya que la duración de cada grupo de rotación en el panel es relativamente corta.

Por otra parte, en cualquier momento de tiempo, la muestra puede incluir nuevas unidades a través de grupos de rotación más recientes siempre y cuando la cobertura muestral se actualice para cada grupo (por ejemplo, mediante la actualización de las listas de viviendas en el caso de diseños con varias etapas. De esta forma, pueden

estar perfectamente representados en las estimaciones trasversales mediante el uso de un esquema de ponderaciones adecuado que refleje el hecho de que han podido ser seleccionados en alguno de los grupos de rotación en los cuales las estimaciones trasversales están basadas.

2.4 Encuestas de panel

En esta sección se consideran los diseños de encuesta en los cuales los mismos elementos son encuestados a lo largo del tiempo. Estos diseños son conocidos como encuestas panel o encuestas longitudinales. Para estos diseños se usa el término 'encuesta panel', mientras que el término 'longitudinal' se usa para describir los datos que se producen a partir de dicho diseño.

La distinción entre encuestas panel y encuestas de panel rotante se vuelve borrosa en el caso de encuestas panel de duración fija, cuando se introducen paneles nuevos de forma periódica; el nuevo panel se puede solapar con el vigente o puede empezar después de que el actual haya terminado. La distinción que se hace aquí es entre encuestas que se centran principalmente en estimaciones trasversales y estimaciones de cambio neto, como en los paneles rotantes de las encuestas de población activa y las encuestas panel que se centran principalmente en los estudios longitudinales a nivel de unidad individual (por ejemplo, cambio bruto).

En la Sección 2.4.1 se describirán algunos tipos de encuestas panel mientras que en la Sección 2.4.2 se revisarán los problemas metodológicos que surgen con las encuestas panel.

2.4.1 Tipos de encuestas panel

El beneficio de una encuesta panel es que produce los datos necesarios para estudios longitudinales y, de esta forma, expande el potencial analítico de una encuesta trasversal. Aunque las encuestas panel se realizan desde hace mucho tiempo, en los últimos años ha aumentado el interés en ellas gracias al uso de ordenadores más potentes necesarios para manejar grandes ficheros de datos y realizar análisis longitudinales. Hay muchas encuestas panel realizándose actualmente y mucha literatura sobre cómo realizarlas (por ejemplo, Lynn 2009; Trivellato 1999).

Martin y col. 2006 describe de un gran número de encuestas panel sobre condiciones sociales llevadas a cabo en varios países. Estas encuestas abordan temas como la salud física y mental y discapacidad; desarrollo físico, social y educativo; historia laboral; dinámicas familiares; efectos del divorcio; el cambio a la jubilación y los efectos del envejecimiento; y la integración social y cultural de los inmigrantes. La mayoría se centran inicialmente en áreas específicas pero, con el paso del tiempo, tienden a ampliar los temas tratados. De hecho, una de las ventajas de los diseños de panel es que permiten

la recogida de muchos temas en las distintas rondas de la encuesta.

Los datos longitudinales obtenidos a partir de encuestas panel ofrecen la oportunidad de llevar a cabo una gran variedad de análisis que no sería posible con datos trasversales. Por ejemplo, análisis de cambios brutos; la duración de fases (por ejemplo, de pobreza); trayectorias de crecimiento con modelos de curvas de crecimiento (como el desarrollo físico y cognitivo de los niños); indicadores tempranos que predigan resultados posteriores (por ejemplo, resultados en la salud por exposiciones medioambientales en la niñez); y relaciones causales temporales entre 'causas' y efectos, usando la modelización de ecuaciones estructurales longitudinales (la auto eficacia como mediador entre sucesos en una vida estresante y síntomas depresivos entre otras).

A veces, los datos longitudinales necesarios se pueden obtener mediante una encuesta retrospectiva o de registros administrativos. Sin embargo, cuando la calidad de la encuesta retrospectiva es inadecuada y los datos administrativos no están disponibles o son insuficientes, se hace necesaria la recogida directa de datos mediante una encuesta panel.

Aunque las encuestas panel se ocupan principalmente de producir los datos necesarios para análisis longitudinales a nivel de microdato, en la mayoría de los casos también se pueden analizar trasversalmente en cada ronda. Un aspecto importante para el análisis trasversal es que la cobertura muestral sea adecuada en cada una: a no ser que se tomen medidas de forma regular para 'refrescar' las muestras trasversales añadiendo muestras de nuevos miembros de la población desde la última actualización de la muestra, los nuevos miembros no estarán representados en los análisis trasversales. Los nuevos miembros generalmente no están incluidos y por tanto no son necesarios para esos análisis longitudinales que empiezan con datos desde la primera ronda del panel.

Estudios de cohortes

Un tipo de encuestas panel es la que se conoce a menudo como un estudio de cohorte (Bynner 2004). Muchos estudios de cohortes toman muestras de personas de una edad particular y los siguen a través de importantes periodos de sus vidas. Una cohorte puede ser seguida a lo largo de su vida y, de hecho, el estudio puede extenderse para seguir a la descendencia de la cohorte original.

Aunque las cohortes de nacimiento proporcionan datos longitudinales muy valiosos para examinar los efectos de experiencias en la salud a edades muy tempranas y otros factores que aparecen más tarde en la vida, también tiene algunas limitaciones. Los miembros de un estudio de cohorte se ven afectados por los mismos sucesos históricos, o efectos temporales (por ejemplo, guerras y desastres medioambientales) que afectan a la población en ese momento. También se pueden ver afectados por esos eventos de forma distinta debido a su susceptibilidad en la edad en que se ven afectados por los sucesos (efectos de cohorte). Los datos de una única cohorte pueden confundir edad, periodo y efecto de cohorte, y el resultado tiene que ser debidamente interpretado (Yang

2007).

Encuestas panel a hogares

Una segunda categoría de encuestas panel son las encuestas panel a hogares (Rose 2000). Este tipo de encuestas empieza con una muestra de hogares y sigue a sus miembros a lo largo de la duración del panel. Para reflejar las condiciones económicas y sociales de los hogares del panel en cada ronda, la encuesta también recoge datos sobre las personas con las que los miembros del panel viven en rondas posteriores, denominadas *cohabitantes*, *personas externas a la muestra* o *personas asociadas*. Estas encuestas son, en realidad, muestras de personas más que de hogares. Los hogares están cambiando constantemente, con miembros entrando y saliendo, formándose nuevos hogares mientras que otros desaparecen. Por eso, la definición de un hogar longitudinal es bastante problemática, a no ser que el periodo temporal sea muy corto. Son preferibles para análisis longitudinales los análisis a nivel de persona con características específicas de cada ronda.

Algunos ejemplos son la European Community Household Panel (ECHP) (https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/DataPortal/ECHP) o la European Union Statistics on Income and Living Conditions (EU-SILC) que susutituye a la ECHP. (https://ec.europa.eu/eurostat/statistics-explained/index.php?title=EU_statistics_on_income_and_living_conditions_(EU-SILC)_methodology). Un rasgo distintivo de estas encuestas es que están diseñadas para durar un periodo fijo de tiempo, por ejemplo la EU-SILC sigue un diseño con rotación de cuatro años.

Una ventaja de paneles contiguos es que el tamaño muestral total está disponible para análisis longitudinales durante la duración del panel. Sin embargo, los paneles contiguos no pueden tratar los análisis longitudinales durante el periodo que abarcan dos paneles. También hay que tener en cuenta que estimaciones válidas de tendencias de estaciones trasversales, como estimaciones anuales que son importantes para estas encuestas, no se pueden obtener por los sesgos generados por el tiempo de pertenencia en la muestra entre años. Los diseños de rotación con una rotación anual puede producir estimaciones aceptables de la tendencia debido al equilibrio del tiempo de pertenencia en la muestra de entre rondas.

La elección de la duración de las encuestas panel de hogares de duración limitada depende de una combinación de objetivos analíticos y de consideraciones prácticas sobre la recogida de datos, en particular sobre la carga al informante y su efecto en la tasa de respuesta en las rondas posteriores. La tasa de respuesta y otros asuntos prácticos fueron los que influyeron en que se pasase de la ECHP a la EU-SIlC, con un diseño rotante.

Encuestas panel transnacionales

Un proyecto reciente y que es importante en la investigación con encuestas es el uso de datos para comparaciones internacionales, permitiendo la investigación de los efectos de condiciones sociales diferentes en la población estudiada. Estos proyectos se aplican por igual a varias encuestas panel. Algunos ejemplos notables son las encuestas panel sobre el envejecimiento (HRS, ELSA, y SHARE) que se realizan de forma coordinada entre varios países y las encuestas panel sobre ingresos del hogar. Si bien la coordinación es útil, es también necesaria la armonización de la recogida de datos en los distintos países.

2.4.2 Aspectos metodológicos en encuestas panel

Muchos de los aspectos metodológicos (desgaste de la muestra debido al cansancio, condicionamiento de panel, sesgo por el tiempo de permanencia en la muestra, y la necesidad de incluir nuevos miembros) se encuentran tanto en encuestas panel como en encuestas de panel rotante. Sin embargo, los problemas sobre el cansancio o desgaste muestral y la inclusión de nuevos miembros se incrementa con paneles de mayor duración, y los efectos de condicionamiento son un problema serio en el análisis longitudinal.

Mantener la participación en el panel

Éste es un asunto crítico en una encuesta panel. Una encuesta panel está sujeta a *falta de respuesta total*, que ocurre cuando una unidad muestral no participa en ninguna de las rondas del panel. Además, una encuesta panel está sujeta a *falta de respuesta por ronda*, que ocurre cuando una unidad muestral participa en alguna pero no en todas las rondas de la encuesta.

La falta de respuesta por ronda puede consistir en la *falta de respuesta por cansancio* (cuando la unidad abandona el panel en una de las rondas y nunca vuelve) o falta de respuesta por una causa diferente (cuando una unidad muestral no responde en una ronda pero responde en la siguiente o alguna posterior).

Por razones prácticas, muchas encuestas no hacen ningún intento de convertir la falta de respuesta inicial en respuesta en la siguiente o posteriores etaoas. Por tanto, los no informantes iniciales se convierten en no informantes totales. También, los informantes que rotundamente rechazan participar o que no pueden ser localizados en una ronda pueden ser seguidos en rondas posteriores y, por lo tanto, se convierten en casos faltantes debidos al desgaste. Con frecuencia no se hace ningún intento de seguir a aquellos que no han contestado dos rondas consecutivas.

En general, las encuestas panel se enfrentan a la mayor tasa de pérdida en la ronda inicial, después de la cual una alta proporción de los informantes en cada ronda consecutiva también proporciona datos en la siguiente ronda. Sin embargo, la acumulación de falta de respuesta a lo largo del tiempo a menudo da lugar a una tasa de falta de

respuesta alta. En el caso de paneles de largo plazo, esta situación plantea el dilema de si continuar con el panel existente, con su potencial analítico creciente, o terminarlo y empezar con uno nuevo.

Algunos panel, en particular aquellos con una alta tasa de respuesta, se diseñan para que tengan una duración limitada por la preocupación sobre el cansancio. Los posibles efectos en el sesgo de la acumulación de falta de respuesta son una seria preocupación en casi todas las encuestas panel. Véase, por ejemplo, el número especial sobre cansancio en encuestas longitudinales en *Journal of Human Resources* (Volume 33, Number 2, 1998).

Las principales causas de falta de respuesta por ronda son la falta de seguimiento de los miembros del panel que se traslada y los rechazos a contestar por la carga al informante. Se utiliza una gran variedad de métodos con el fin de minimizar la pérdida del seguimiento, de forma particular en los paneles con intervalos grandes entre rondas. Una forma es conseguir información de contacto como el mail o el teléfono móvil.

Prevenir la pérdida de miembros del panel que no quieren participar más es un reto. Lo mismo que en encuestas trasversales, los incentivos pueden aumentar la participación. Con encuestas panel, se puede ofrecer incentivos a aquellos informantes que mostraron reticencia en la ronda previa, es decir, aquellos que no respondieron a muchas de las preguntas. Hay un debate sobre esta forma de proceder, ya que se puede premiar comportamientos que no son deseables.

Minimizar la carga al informante es otro enfoque que limita las negativas en rondas posteriores de la encuesta panel. Una forma de reducir la carga al informante es enlazando con datos administrativos; esto también se puede utilizar si los miembros del panel no puede responder de forma precisa. Por razones éticas, hay que pedir permiso a los miembros del panel para enlazar los datos, y es necesario usar procedimientos que aseguren que el enlace entre los daros no pueda perjudicar a los miembros del panel.

Errores de medida

Los errores de medida son un problema en todas las encuestas, pero son especialmente problemáticos para los análisis longitudinales de los datos de las encuestas de panel.

Kalton, Kasprzyk y McMillen 1989 proporciona una gran variedad de fuentes de errores de medida pueden sesgar las estimaciones de los cambios brutos: los efectos de condicionamiento del panel, cambios en los modos de recogida de datos, cambios en los informantes entre rondas (incluyendo la posibilidad de proxys cuando no se puede contactar con las unidades de la muestra), cambios en el personal (por ejemplo, en los entrevistadores, codificadores de respuestas abiertas), cambios en el cuestionario (con posibles efectos de contexto incluso cuando las preguntas no cambian), cambios en el contenido del cuestionario, cambios en la interpretación de una pregunta, imputación

de respuestas missing, y errores de grabación. Los tipos de efectos que los errores de medida tienen sobre los análisis de los cambios brutos también afectan a muchas otras formas de análisis longitudinal.

Una forma de intentar reducir la sobreestimación del cambio bruto es usar entrevistas dependientes, en las que se recuerda a los informantes sus respuestas en rondas previas. Un riesgo de las entrevistas dependientes es, por supuesto, la generación de falsa consistencia en las respuestas. Se han realizado muchos estudios para evaluar el efecto de las entrevistas dependientes.

En general, se considera que las entrevistas dependientes reducen los errores de respuesta y la sobreestimación del cambio bruto. Las entrevistas dependientes se pueden usar de forma proactiva recordando al informante su estado en rondas previas (por ejemplo que estaba trabajando) y después preguntando por el estado actual, o de forma reactiva preguntando sobre el estado actual y las discrepancias desde la respuesta anterior. Las entrevistas dependientes proactivas también pueden ser útiles para reducir la carga de respuesta. En situaciones en las que los informantes de un hogar pueden cambiar entre rondas, las entrevistas dependientes pueden dar lugar a que se proporcionen las respuestas de un informante a otro. Este problema de la confidencialidad debe ser tratado con un consentimiento que permita compartir las respuestas entre miembros del hogar (Pascale y Mayer 2004).

Ponderación e imputación

Los métodos de ajuste de ponderaciones usados para compensar la falta de respuesta total en las encuestas trasversales se pueden usar también en el caso de falta de respuesta total en encuestas panel. Sin embargo, compensar las falta de respuesta por ronda (particularmente, la falta de respuesta no causada por cansancio) y la falta de respuesta parcial es mucho más difícil.

Gran parte de la información se puede llegar a conocer a través de respuestas proporcionadas en ronda(s) en las que participaron los informantes. Un enfoque para gestionar los datos missing es imputar todos los ítems que son missing, incluyendo los ítems en las rondas en las cuales las unidades muestrales no son informantes. Este enfoque tiene la ventaja de conservar en los ficheros de análisis toda la información proporcionada por las unidades muestrales. Sin embargo, la gran cantidad de imputación necesaria plantea dudas sobre los sesgos que los valores imputados pueden introducir en el análisis.

Otro enfoque alternativo es usar pesos de ajuste para tratar alguna o todas las rondas de missings. Este enfoque limita el fichero de análisis a las unidades muestrales que responden en todas las rondas. Usa un número limitado de respuestas de la encuesta para hacer los ajustes, pero se pierden las respuestas a otros ítems (Kalton 1986; Lepkowski 1989). La imputación es la solución natural cuando una unidad no responde a varios ítems, pero la elección entre imputación y ajustes con pesos es menos directa para la

falta de respuesta por ronda.

Para conservar la estructura de la covarianza en el conjunto de datos, la imputación requiere que se usen todas las otras variables asociadas con la variable a imputar como variables auxiliares en el modelo de imputación. Satisfacer este requerimiento de forma adecuada es muy difícil en el caso de encuestas trasversales, pero lo es incluso más en el caso de encuestas panel porque el modelo tiene que incorporar variables de otras rondas en la encuesta y variables de la ronda en cuestión. Si no se incluyen las respuestas a la misma variable en otras rondas, el cambio bruto estará sobreestimado. Por tanto, es necesario que el conjunto de datos incluya los datos de las rondas previas. Puesto que en el momento de imputar los datos de la ronda actual no se dispone de los datos de rondas futuras, se pueden producir imputaciones preliminares para cada ronda y luego actualizarlas cuando se disponga de los datos de la siguiente ronda.

Las dudas sobre el sesgo que puede producir la imputación masiva por la falta de respuesta por ronda ha dado lugar a que se prefiera el uso de pesos en la imputación. En el caso de falta de respuesta por cansancio, una práctica común es calcular pesos para que los que han informado en cada ronda, basados en los datos recogidos en rondas anteriores (por la definición de cansancio, todos los que han informado en la ronda en curso han informado en rondas previas). Con la gran cantidad de datos disponibles para los informantes y los no informantes por cansancio en una ronda determinada, el cálculo de pesos es más complejo que en encuestas trasversales, pero el proceso es esencialmente el mismo.

Cuestiones sobre el muestreo

Hay muchas cuestiones sobre el muestreo que surgen con las encuestas panel. Una cuestión trata sobre el grado de agrupamiento que se debe usar al seleccionar la muestra en la primera ronda de un panel. La efectividad de agrupación al reducir el tiempo de viaje de los entrevistadores y facilitar las entrevistas personales desaparece en el tiempo a medida que algunos miembros del panel se mudan. Además, una vez que la muestra ha tenido sus primeras encuestas cara a cara se pueden usar otros métodos de recogida de datos que no se benefician del agrupamiento (teléfono o web). Esto hace que el uso de agrupamiento en la primera ronda de una encuesta panel no sea tan importante como en el caso de encuestas trasversales.

La mayoría de las encuestas tienen como objetivo calcular estimaciones para ciertos subgrupos de la población así como para el total poblacional. Subgrupos pequeños a menudo están sobremuestreados para generar tamaños muestrales que produzcan niveles adecuados de precisión para las estimaciones de los subgrupos. Hay que evaluar cuidadosamente el uso del sobremuestreo en el caso de encuestas panel ya que los objetivos y los subgrupos de interés pueden cambiar a lo largo del tiempo y el sobremuestreo puede terminar siendo perjudicial. También hay que tener en cuenta el tipo de subgrupo. La característica del subgrupo puede ser estática, como el grupo racial/étnico, pero también puede cambiar a lo largo del tiempo (por ejemplo, estar

en situación de pobreza o vivir en una determinada provincia). En este último caso el sobremuestreo puede ser problemático ya que las unidades que se muevan de subgrupo en rondas posteriores tendrán pesos distintos a los iniciales de ese subgrupo, dando lugar a una pérdida de precisión de las estimaciones.

Un ejemplo extremo ocurre con paneles de empresas, a menudo se usan muestra de empresas muy desproporcionadas (esto se debe a la asimetría en el tamaño de las empresas, suele haber pocas grandes y muchas medianas y pequeñas), pero con el paso del tiempo, algunas empresas pequeñas pueden crecer. Estas empresas con gran crecimiento pueden mantener los pesos grandes que tenían inicialmente, lo que puede provocar una pérdida de precisión en las estimaciones. Si el sobremuestreo se usa con subgrupos no permanentes en una encuesta panel, se tienen que tener en cuenta algunas consideraciones sobre mantener la variabilidad en las tasas de muestreo dentro de límites razonables, para evitar la pérdida de precisión asociada con movimientos entre subgrupos.

En los tipos de encuestas panel descritas aquí, el objetivo principal es el de proporcionar los datos necesarios para análisis longitudinales. Sin embargo, los datos de una encuesta panel también se pueden usar para análisis trasversales con los datos de cada ronda. Un aspecto importante para este tipo de análisis es que el total de la población esté representado en el momento de la ronda en cuestión, es decir, que se tengan en cuenta en la muestra unidades que han entrado en la población después de que la muestra para la primera ronda ha sido seleccionada. Este mismo problema surge en los análisis longitudinales cuando el punto de inicio de los análisis son posteriores a la puesta en marcha del panel. Hay que tener en cuenta las unidades nuevas de la población y también las pérdidas.

Cuestiones éticas y de confidencialidad

Para concluir esta sección sobre los aspectos metodológicos, es necesario comentar la importancia de las cuestiones técnicas y de los problemas de confidencialidad. El requisito de que los informantes estén informados sobre los motivos del estudio desde el principio puede ser difícil de llevar a cabo en el caso de un estudio panel a largo plazo ya que los motivos pueden cambiar a lo largo del panel. Por eso hay que tener cuidado con estos temas al diseñar los documentos de consentimiento.

Las encuestas panel suelen ser caras de realizar, pero producen mucha información muy valiosa para los análisis de distintos temas. Por tanto, los datos deberían de estar disponibles para tanta gente como sea posible, incluyendo investigadores y público general. Para ello hay que asegurarse de que se cumplen las condiciones de control del secreto estadístico tanto la difusión del conjuntos de datos para el uso público como su uso por parte de investigadores (Béland 1999). Aunque se pueden usar técnicas como la supresión de datos (principalmente de datos geográficos detallados), intercambio de datos, etc., para proteger los ficheros de microdatos que se proporcionen al público en general. En el caso de los investigadores se pueden utilizar otros métodos. Otro aspecto

a tener en cuenta en la difusión de los datos de los estudios panel es que se debe de proporcionar toda la documentación desde el principio, para que se disponga de toda la información necesaria.

2.5 Conclusiones

La recogida de datos en el tiempo permite elegir entre encuestas repetidas, como se ha visto en la Sección 2.2, encuestas de panel rotante Sección en 2.3 y diseños de panel completo visto en Sección 2.4. Si los datos se van a usar para análisis longitudinales, entonces sólo un diseño de panel sirve para este propósito. Sin embargo, si los datos se van a usar sólo para análisis de tendencias, se puede usar cualquiera de los diseños, teniendo en cuenta que hay que refrescar a muestra en cada ronda para incluir nuevos miembros de la población.

Las cuestiones de diseño para una serie de encuestas repetidas podrían parecer las mismas que las de encuestas trasversales, pero de hecho son distintas. Si se quiere realizar una serie de encuestas repetidas, necesitamos reflejar las necesidades de los datos en el futuro con el fin de incluirlos desde el principio.

A lo largo de la serie de encuestas, será necesario tomar decisiones sobre los aspectos que cambiarán en el diseño con el fin de satisfacer las condiciones y de ajustarse a las mejores prácticas, o si es necesario no modificar el diseño con el fin de mantener las estimaciones de las tendencias. Los analistas de encuestas repetidas deben de tener información sobre cualquier cambio que se haga en el diseño que pueda alterar las estimaciones de la tendencia.

Las encuestas panel son más complejas de diseñar y analizar que las encuestas trasversales. Además de los aspectos generales del diseño de encuestas, los que diseñan encuestas panel tienen que prestar mucha atención a asuntos como la cooperación de los miembros del panel, métodos de seguimiento, incluir nuevos miembros en la muestra con el fin de proporcionar estimaciones trasversales válidas, el uso de entrevistas dependientes, y el uso de incentivos.

Los analistas deben de tener información sobre los errores de medida y las condiciones del panel en sus análisis, así como del posible deterioro de la muestra a lo largo del tiempo.

Bibliografía

Bailar, B.A. (1975). "The effects of rotation group bias on estimation from panel surveys". En: *Journal of the American Statistical Association* 70, págs. 23-30.

Béland, Y. (1999). "Release of public use microdata files for NPHS? Mission partially accomplished". En: *Proceedings of the Section on Survey Research Methods. American Statistical Association*, págs. 404-409.

Biemer, P.P., R.M. Groves, L. Lyberg, N.A. Mathiowetz y S. Sudman (1991). *Measurement Errors in Surveys*. New York: John Wiley & Sons.

Bynner, J. (2004). "Longitudinal cohort designs". En: *In: Kempf-Leonard, K. (Ed.), Encyclopedia of Social Measurement. Elsevier/Academic Press, Boston* 2, págs. 591-599.

Cantor, D. (2007). *A review and summary of studies on panel conditioning. In: Menard, S. (Ed.), Handbook of Longitudinal Research: Design, Measurement and Analysis*. San Diego: John Wiley & Sons, págs. 123-139.

Dillman, D.A. y L.M. Christian (2005). "Survey mode as a source of instability in responses across surveys". En: *Field Methods* 17, págs. 30-52.

Kalton, G. (1986). "Handling wave nonresponse in panel surveys". En: *Journal of Official Statistics* 2, págs. 303-314.

Kalton, G., D. Kasprzyk y D.B. McMillen (1989). *Nonsampling errors in panel surveys*. New York: In: Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (Eds.), Panel Surveys. John Wiley & Sons, pp. 249-–270.

Kish, L. (1999). "Survey Methodology". En: *In J. G. Bethlehem and P. G. M. van der Heijden (eds), COMPSTAT – Proceedings in Computational Statistics* 25, págs. 129-138.

Kulka, R.A. (1982). "Monitoring social change via survey replication: prospects and pitfalls from a replication survey of social roles and mental health". En: *Journal of Social Issues* 17—38, págs. 65-76.

Leeuw, E.D. de (2005). "To mix or not to mix data collection modes in surveys". En: *Journal of Official Statistics* 21, págs. 233-255.

Lepkowski, J.M. (1989). *Treatment of wave nonresponse in panel surveys. In: Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (Eds.), Panel Surveys*. New York, pp. 348–374: John Wiley & Sons.

Lynn, P. (2009). *Methodology of Longitudinal Surveys*. Chichester, UK: John Wiley & Sons.

Martin, J., J. Bynner, G. Kalton, P. Boyle, H. Goldstein, V. Gayle, S. Parsons y A. Piesse (2006). "Strategic Review of Panel and Cohort Studies, with Appendices". En: *Report to the U.K. Research Resources Board of the Economic and Social Research Council. Available at: http://www.longviewuk.com/pages/publications.shtml*.

Neter, J. y J. Waksberg (1964a). "A study of response errors in expenditures data from household interviews". En: *Journal of the American Statistical Association* 59, págs. 18-55.

– (1964b). "Conditioning effects from repeated interviews". En: *Journal of Marketing* 28, págs. 51-56.

Pascale, J. y T.S. Mayer (2004). "Exploring confidentiality issues related to dependent interviewing: preliminary findings". En: *Journal of Official Statistics* 20, págs. 357-377.

Pfefferman, D. y C.R. Rao (2009). *Handbook of Statistics 29A*. North-Holland, Amsterdam: Elsevier.

Rose, D. (2000). *Researching Social and Economic Change: The Uses of Household Panel Studies*. Routledge. London https://doi.org/10.4324/9780203501085.

Sturgis, P., N. Allum e I. Brunton-Smith (2009). *Attitudes over time: the psychology of panel conditioning. In: Lynn, P. (Ed.), Methodology of Longitudinal Surveys*. New York: John Wiley & Sons, págs. 113-126.

Tourangeau, R., L.J. Rips y K. Rasinski (2000). *The Psychology of Survey Response*. New York: Cambridge University Press.

Trivellato, U. (1999). "Issues in the design and analysis of panel studies: a cursory review". En: *Quality and Quantity* 33, págs. 339-352.

Waterton, J. y D. Lievesley (1989). *Evidence of conditioning effects in the British Social Attitudes Panel Survey. In: Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (Eds.), Panel Surveys*. New York: John Wiley & Sons, págs. 319-339.

Yang, Y. (2007). "Age-period-cohort distinctions". En: *Markides, K., Blazer, D.G., Studenski, S., Branch, L.G. (Eds.), Encyclopedia of Health and Aging. Sage Publications, Newbury Park, CA*, págs. 20-22.

Tema 3

Introducción a problemas de estimación complejos. El efecto del sesgo en intervalos de confianza de las estimaciones. Consistencia e insesgadez asintótica. La técnica de linealización de Taylor para la estimación de la varianza. Estimador de una razón: varianza y sesgo.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía.

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

3.1 Introducción a problemas de estimación complejos.

Además de parámetros como los totales o medias poblacionales, nos pueden interesar otros parámetros como las razones de totales, coeficientes de regresión, medianas y otros cuantiles poblacionales. Estos parámetros no están estructurados de una manera tan sencilla como el total poblacional y requieren un procedimiento algo más complejo de estimación.

Como ejemplo, supongamos que queremos estimar la razón de dos totales poblacionales desconocidos,

$$R = \frac{\sum_U y_k}{\sum_U z_k} = \frac{Y_U}{Z_U}$$

donde y y z son dos variables de análisis. Por ejemplo, en una población de individuos, y_k puede representar los ahorros y z_k los ingresos del k-ésimo individuo. Por tanto, R es la proporción de ahorros que representa un ingreso de un euro en la población finita. Una forma obvia de obtener un estimador de R es estimar los totales Y_U y Z_U por sus respectivos estimadores de Horvitz-Thompson $\widehat{Y}_U^{\text{HT}} = \sum_{k\in s} \frac{y_k}{\pi_k}$ y $\widehat{Z}_U^{\text{HT}} = \sum_{k\in s} \frac{z_k}{\pi_k}$, donde π_k es la probabilidad de inclusión de primer orden del elemento $k \in U$. El estimador

resultante de R es

$$\widehat{R} = \frac{\widehat{Y}_U^{\text{HT}}}{\widehat{Z}_U^{\text{HT}}} = \frac{\sum_{k\in s} \frac{y_k}{\pi_k}}{\sum_{k\in s} \frac{z_k}{\pi_k}}.$$

Aunque compuesto por dos componentes insesgados, $\widehat{Y}_U^{\text{HT}}$ y $\widehat{Z}_U^{\text{HT}}$, el estimador $\widehat{R}$ no es insesgado para R, sino *aproximadamente* insesgado, bajo ciertas condiciones, como se verá en la Sección 3.5.

Aunque conozcamos las varianzas de $\widehat{Y}_U^{\text{HT}}$ y $\widehat{Z}_U^{\text{HT}}$, no se puede proporciona una fórmula sencilla de $\widehat{R}$. Sin embargo, podemos obtener una varianza aproximada (véase la Sección 3.5) que también nos permite obtener un procedimiento para una estimación de la varianza. De esta forma se pueden obtener intervalos de confianza que se ajustan aproximadamente a un nivel de confianza deseado.

El procedimiento que da lugar al estimador $\widehat{R}$ del parámetro R es un ejemplo sencillo de un principio general importante para la estimación de un parámetro poblacional θ que se puede expresar como una función de varios totales poblacionales[1], $Y_1, Y_2, ..., Y_q$,

$$\theta = f(Y_1, ..., Y_j, ..., Y_q)$$

donde

$$Y_j = \sum_{k\in U} y_j k$$

e $y_{1k}, ..., y_{qk}$ son valores para el k-ésimo elemento de las variables de estudio $y_1, ..., y_q$, respectivamente. En la función $f(\cdot, \ldots, \cdot, \ldots, \cdot)$, se sustituye cada total desconocido Y_j por su correspondiente estimador HT,

$$\widehat{Y}_{jU}^{\text{HT}} = \sum_{k\in s} \frac{y_{jk}}{\pi_k}.$$

El estimador resultante de θ es

$$\widehat{\theta} = f(\widehat{Y}_{1U}^{\text{HT}}, \ldots, \widehat{Y}_{jU}^{\text{HT}}, \ldots, \widehat{Y}_{qU}^{\text{HT}}).$$

Naturalmente, estamos interesados en encontrar las propiedades estadísticas (sesgo, varianza, etc.) de $\widehat{\theta}$. Esto es sencillo en caso de que f sea una función lineal. Pero en ocasiones nos encontramos con funciones f no lineales. Usando la aproximación de primer orden de Taylor, $\widehat{\theta}$ puede aproximarse usando una función lineal. Y, a continuación, se pueden obtener expresiones para el sesgo aproximado y la varianza aproximada. El procedimiento general se explica en la Sección 3.4.

[1] Por sencillez de notación, no escribimos la población finita U respecto a la que se calcula el total.

3.2 El efecto del sesgo en intervalos de confianza de las estimaciones.

La insesgadez es una característica del estimador HT. Aunque se trata de una propiedad deseable, la importancia de la insesgadez exacta tampoco debe exagerarse. Hay dos motivos importantes por los que no es razonable, en muchos casos, aspirar a encontrar un estimador *exactamente* insesgado.

1. Muchos parámetros tienen una estructura que hace difícil encontrar un estimador insesgado.
2. Un estimador con algo de sesgo puede a menudo tener una varianza y un error cuadrático medio más pequeños que un estimador insesgado.

Muchos estimadores útiles en la práctica son de hecho únicamente *aproximadamente* insesgados.

Por otro lado, está generalmente aceptado que deberían evitarse los estimadores con un gran sesgo. Hájek 1971 expresó esta regla:

> ... las estimaciones muy sesgadas son malas independientemente del resto de propiedades que tengan.

Entonces, ¿cuánto sesgo debería aceptarse? Se considera un estimador ideal como aquél cuya distribución muestral se encuentra concentrada en torno al valor desconocido del parámetro. Esto garantiza una alta probabilidad de una estimación cercana. Sea $\widehat{\theta}$ un estimador de θ con varianza $\mathbb{V}(\widehat{\theta})$ y sesgo $\mathbb{B}(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta$. Una medida de acuracidad habitual de $\widehat{\theta}$ es el error cuadrático medio[2] (MSE),

$$\mathbb{MSE}(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta} - \theta)^2] = \mathbb{V}(\widehat{\theta}) + [\mathbb{B}(\widehat{\theta})]^2$$

que depende tanto de la varianza como del sesgo. Si el MSE fuese nuestra única preocupación, podríamos considerar cómo el sesgo y la varianza se compensan para dar lugar a un MSE pequeño. Un sesgo no nulo, y quizá significativo, podría ser compensado con una varianza pequeña.

Pero el MSE no muestra el cuadro completo. Además de un MSE pequeño, también pedimos que el sesgo del estimador sea pequeño en relación con el error cometido. Esto es importante para la validez de los intervalos de confianza, como veremos a continuación.

Definición 4

[2] *Mean square error*.

Definimos la razón de sesgo como

$$\mathrm{BR}(\widehat{\theta}) = \frac{\mathbb{B}(\widehat{\theta})}{[\mathbb{V}(\widehat{\theta})]^{\frac{1}{2}}} \tag{3.1}$$

Esta cantidad es a menudo de gran interés por la siguiente razón. Mientras $\mathrm{BR}(\widehat{\theta})$ sea pequeño, un intervalo de confianza para $\widehat{\theta}$ no cometerá un error muy grande, aunque su sesgo no sea nulo. Para verlo supongamos que

$$Z = \frac{\widehat{\theta} - \mathbb{E}(\widehat{\theta})}{[\mathbb{V}(\widehat{\theta})]^{\frac{1}{2}}}$$

sigue una distribución $N(0,1)$. La probabilidad de que el valor desconocido θ esté incluido en el intervalo

$$(\widehat{\theta} - z_{1-\frac{\alpha}{2}}[\mathbb{V}(\widehat{\theta})]^{\frac{1}{2}}, \widehat{\theta} + z_{1-\frac{\alpha}{2}}[\mathbb{V}(\widehat{\theta})]^{\frac{1}{2}}) \tag{3.2}$$

a menudo se denomina la probabilidad de cobertura y viene dada por

$$\begin{aligned} P_0 &= \mathbb{P}\{\widehat{\theta} - z_{1-\frac{\alpha}{2}}[\mathbb{V}(\widehat{\theta})]^{\frac{1}{2}} < \theta < \widehat{\theta} + z_{1-\frac{\alpha}{2}}[\mathbb{V}(\widehat{\theta})]^{\frac{1}{2}}\} \\ &= \mathbb{P}\{-z_{1-\frac{\alpha}{2}} - \mathrm{BR}(\widehat{\theta}) < Z < z_{1-\frac{\alpha}{2}} - \mathrm{BR}(\widehat{\theta})\} \end{aligned}$$

donde Z es una variable aleatoria $N(0,1)$. Téngase en cuenta que (3.2) es realmente un intervalo de confianza sólo si la varianza $\mathbb{V}(\widehat{\theta})$ es conocida. Normalmente, al calcular el intervalo de confianza, $\mathbb{V}(\widehat{\theta})$ tiene que ser sustituido por su estimación $\widehat{\mathbb{V}}(\widehat{\theta})$.

Se sigue que la probabilidad de cobertura P_0 es igual al nivel de confianza deseado, $1-\alpha$, sólo si $\mathrm{BR}(\widehat{\theta})$ toma el valor 0. Cualquier valor no nulo de $\mathrm{BR}(\widehat{\theta})$ afecta de alguna forma a la probabilidad de cobertura, pero el efecto es menor cuando más cercano a 0 esté la razón de sesgo. Este efecto puede ignorarse si, digamos, $\mathrm{BR}(\widehat{\theta})| \leq \frac{1}{10}$ (véase la tabla 3.1). En la práctica, la razón de sesgo es desconocida, haciendo imposible calcular el verdadero valor de la probabilidad de cobertura P_0.

$\lvert\mathrm{BR}(\widehat{\theta})\rvert$	P_0
0,00	0,9500
0,05	0,9497
0,10	0,9489
0,30	0,9396
0,50	0,9210
1,00	0,8300

Table 3.1: Probabilidad de cobertura P_0 como función de la razón de sesgo $\mathrm{BR}(\widehat{\theta})$.

Pero cuando el tamaño muestral aumenta (en el caso de un diseño de tamaño fijo) o cuando el tamaño muestral esperado aumenta (en el caso de un diseño de tamaño aleatorio), la varianza $\mathbb{V}(\widehat{\theta})$ tenderá a cero y también lo hará el sesgo $\mathbb{B}(\widehat{\theta})$, de forma que la razón de sesgo se aproximará a cero.

Comentario 9. Cuando se use un estimador sesgado, puede que nos interese considerar otros tipos de procedimientos de estimación de intervalos distintos de

$$\widehat{\theta} \pm z_{1-\frac{\alpha}{2}}[\mathbb{V}(\widehat{\theta})]^{\frac{1}{2}}.$$

Por ejemplo, si $\widehat{\theta}$ es insesgado para θ y $\widehat{\mathbb{MSE}}(\widehat{\theta})$ es un buen estimador para

$$\mathbb{MSE}(\widehat{\theta}) = \mathbb{V}(\widehat{\theta}) + [\mathbb{B}(\widehat{\theta})]^2,$$

entonces puede utilizarse como intervalo para θ

$$\widehat{\theta} \pm z_{1-\frac{\alpha}{2}}[\widehat{\mathbb{MSE}}(\widehat{\theta})]^{\frac{1}{2}}. \tag{3.3}$$

■

Naturalmente, nos interesaría conocer las propiedades de cobertura del intervalo dado por (3.3) como función del sesgo.

Para simplificar la cuestión, supongamos que $\mathbb{MSE}(\widehat{\theta})$ es conocido. Estamos interesados en conocer la probabilidad de que el intervalo

$$\widehat{\theta} \pm z_{1-\frac{\alpha}{2}}[\mathbb{MSE}(\widehat{\theta})]^{\frac{1}{2}}$$

contenga el verdadero valor de θ. El valor de la probabilidad se puede expresar en términos de la razón de sesgo $\mathrm{BR}(\widehat{\theta})$. Por ejemplo, con $1-\alpha = 0{,}95$ ($z_{0{,}975} = 1{,}96$), los estudios que asumen que θ se distribuye según una distribución normal muestran que la probabilidad de cobertura se encuentra entre $0,939$ y $0,950$ si $|\mathrm{BR}(\widehat{\theta})| \leq 0,1$.

3.3 Consistencia e insesgadez asintótica.

En primer lugar, recordamos los conceptos de consistencia e insesgadez asintótica de la teoría general de inferencia estadística. Sea τ una parámetro a estimar mediante un estimador $\widehat{\tau}_n$, es decir, una función de n variables aleatorias independientes e idénticamente distribuidas $\xi_1, \xi_2, ..., \xi_n$. El estimador $\widehat{\tau}_n$ se dice que es *asintóticamente insesgado* de τ si

$$\lim_{n\to\infty} \mathbb{E}(\widehat{\tau}_n) = \tau$$

y $\widehat{\tau}_n$ se dice que es *consistente* para τ si, para cualquier valor $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\widehat{\tau}_n - \tau| > \varepsilon) = 0$$

Estrictamente hablando, no es un único estimador $\widehat{\tau}_n$ sino una secuencia de estimadores $\widehat{\tau}_1, \widehat{\tau}_2, ...$ lo que es asintóticamente insesgado o consistente.

En la práctica, n es siempre finito, aunque tome un valor grande. La importancia práctica de la insesgadez asintótica y de la consistencia es la siguiente. Si se sabe que un estimador es insesgado asintóticamente, entonces puede considerarse aproximadamente insesgado cuando n es suficientemente grande. Y si se verifica la consistencia, la distribución muestral del estimador se puede considerar fuertemente concentrado en torno a τ cuando n es suficientemente grande.

Ahora volvamos a la teoría del muestreo. Resulta que las definiciones anteriores de consistencia e insesgadez asintótica no se pueden llevar de forma inmediata al muestreo a partir de una población finita, ya que si $\widehat{\theta}_n$ es un estimador de θ basado en una muestra de tamaño n obtenida a partir de una población finita de tamaño N, entonces no puede ser que $n \to \infty$ sin más, ya que $n \leq N$ y N es fijo y finito.

En este contexto, la cognergencia asintótica debe ser entendida correctamente y estos resultados asintóticos requieren una maquinaria más compleja. Esta maquinaria más compleja consiste en definir una sucesión de poblaciones crecientes de forma que tanto n como N tiendan a infinito. Incluimos simplemente una idea del marco conceptual para el razonamiento asintótico en teoría de muestreo (véase Fuller 2009, para los detalles).

Empezamos con la idea de una sucesión infinita de elementos, etiquetados como $k = 1, 2, 3, ...$ y una sucesión infinita de valores de y asociados, denotados por $y_1, y_2, y_3, ...$, donde y_k es el valor asociado al k-ésimo elemento.

Consideremos una sucesión de poblaciones $U_1, U_2, U_3, ...$, donde U_ν está formado por los N_ν primeros elementos de la sucesión infinita de elementos antes mencionados, es decir, $U_\nu = \{1, 2, ..., N_\nu\}$. Asumimos que $U_1 \subset U_2 \subset U_3 \subset ...$ y, por tanto, $N_1 < N_2 < N_3 < ....$ Sea θ_ν el valor de un determinado parámetro de la población U_ν, es decir, θ_ν es una función de los valores $y_1, y_2, ..., y_{U_\nu}$.

Para cada población U_ν, consideremos un diseño muestral probabilístico $p_\nu(\cdot)$ que asigna una probabilidad determinada $p_\nu(s_\nu)$ a cada muestra posible s_ν de elementos de U_ν. Sean $\pi_{\nu k}$ y $\pi_{\nu kl}$ $(k, l = 1, 2, ..., N_\nu)$ las probabilidades de inclusión determinadas por el diseño $p_\nu(\cdot)$. Por simplicidad asumiremos que el tamaño muestral es fijo y se denota por n_ν. También asumimos que $n_1 < n_2 < n_3 < ....$ Claramente $\nu \to \infty$ significa que tanto $n_\nu \to \infty$ como $N_\nu \to \infty$. Sea $\widehat{\theta}_\nu$ un estimador de θ_ν basado en los valores observados y_k, es decir, aquellos para los que $k \in s_\nu$.

Por ejemplo, el parámetro θ_ν puede ser la media poblacional

$$\theta_\nu = \bar{y}_{U_\nu} = \sum_{k \in U_\nu} \frac{y_k}{N_\nu}$$

y $\widehat{\theta}_\nu$ ser el estimador HT

$$\widehat{Y}_{U_\nu}^{\text{HT}} = \frac{1}{N_\nu} \sum_{k \in s_\nu} \frac{y_k}{\pi_{\nu k}}. \tag{3.4}$$

Definición 5

En relación a la sucesión de poblaciones y de diseños muestrales descritos anteriormente, se define

i. Un estimador $\widehat{\theta}_\nu$ es *asintóticamente insesgado* para θ_ν si $\lim_{\nu\to\infty}[\mathbb{E}_{p_\nu}(\widehat{\theta}_\nu) - \theta_\nu] = 0$

ii. Un estimador $\widehat{\theta}_\nu$ es *consistente* para θ_ν si, para cualquier valor $\epsilon > 0$, $\lim_{\nu\to\infty} \mathbb{P}(|\widehat{\theta}_n - \theta_\nu| > \epsilon) = 0$.

Estas definiciones aún no están claras, porque el proceso del cálculo del límite no está especificada completamente. Que un estimador sea consistente o asintóticamente insesgado depende de cómo se especifique la sucesión $\{y_k\}$ de valores de y y la sucesión $\{p_\nu\}$ de diseños muestrales. Son necesarias algunas condiciones sobre el comportamiento límite de los momentos de la población finita y de las probabilidades de inclusión. Isaki y Fuller 1982 y Robinson y Särndal 1983 enunciaron las condiciones para la consistencia del estimador HT (3.4). Robinson y Särndal 1983 también proporcionaron las condiciones de insesgadez asintótica.

Comentario 10. Si se dispone de estimadores consistentes $\widehat{\theta}_1, ..., \widehat{\theta}_\nu$ para los parámetros $\theta_1, ..., \theta_\nu$, entonces, para muchas funciones f,

$$f(\widehat{\theta}_1, ..., \widehat{\theta}_\nu)$$

es consistente para $f(\theta_1, ..., \theta_\nu)$. En otras palabras, una función de estimadores consistentes es consistente.

Por ejemplo, si $\widehat{\bar{y}}_U$ y $\widehat{\bar{z}}_U$ son estimadores consistentes de las medias poblacionales $\bar{y}_U$ y $\bar{z}_U$, respectivamente, entonces $\frac{\widehat{\bar{y}}_U}{\widehat{\bar{z}}_U}$ será consistente para $\frac{\bar{y}_U}{\bar{z}_U}$. ■

No es difícil demostrar, con la ayuda de la desigualdad de Chebyshev, que si un estimador $\widehat{\theta}_\nu$ es asintóticamente insesgado para θ_ν y su varianza tiende a cero cuando ν tiende a infinito, entonces $\widehat{\theta}_\nu$ es consistente.

Un tipo diferente de consistencia, que también se puede aplicar en el contexto de teoría de muestreo, es la *consistencia de una población finita*. Para definir este tipo de consistencia no necesitamos considerar una sucesión de poblaciones crecientes sino únicamente una población finita fija, para la cual permitiremos que el tamaño muestral aumente hasta que finalmente iguale al tamaño poblacional. La definición es la siguiente.

Definición 6

iii. Un estimador $\widehat{\theta}$ de θ es *consistente* para una población finita bajo una clase de diseños si $s = U$ implica que $\widehat{\theta} = \theta$.

Esta definición muestra el distinto comportamiento de los diseños con tamaño muestral fijo y con tamaño muestral variable. Por ejemplo, si consideramos el muestreo aleatorio simple sin reemplazamiento, es evidente que cuando $n = N$, esto es, $s = U$, entonces $\widehat{Y}_U^{\text{HT}}(s = U) = Y_U$, por lo que el diseño muestral aleatorio simple con reemplazamiento es consistente según esta última definición. Sin embargo, en un muestreo de Bernoulli, existe una probabilidad no nula de que la muestra coincida con toda la población, pero en esta circunstancia $\widehat{Y}_U^{\text{HT}}(s = U) = \frac{Y_U}{\pi} \neq Y_U$. Por tanto, $\widehat{Y}_U^{\text{HT}}$ no es consistente para esta familia de diseños muestrales. Podemos, no obstante, considerar estimadores de la forma $\widehat{Y}_U = \sum_{k \in s} y_k$, que sí son claramente consistentes según la definición (iii). Sin embargo, estos estimadores subestiman sistemáticamente Y_U.

Todos estos criterios de consistencia deben aplicarse, por tanto, con precaución en la práctica. Incluso en la circunstancia de tener un estimador consistente, puede resultar que proporcione estimaciones insatisfactorias cuando el tamaño muestral es pequeño.

3.4 La técnica de linealización de Taylor para la estimación de la varianza.

Examinemos el problema de estimar un parámetro poblacional θ que se puede expresar como una función de q totales poblacionales $Y_1, ..., Y_q$,

$$\theta = f(Y_1, ..., Y_q)$$

donde $Y_j = \sum_{k \in U} y_{jk}$, $j = 1, ..., q$. Asumimos que el vector $(y_{1k}, ..., y_{jk}, ..., y_{qk})^t$ se puede observar para $k \in s$, permitiéndonos obtener los estimadores HT

$$\widehat{Y}_{jU}^{\text{HT}} = \sum_{k \in s} \frac{y_{jk}}{\pi_k}, \quad j = 1, ..., q.$$

Por el principio ya mencionado en la Sección 3.1 el estimador de θ es entonces

$$\widehat{\theta} = f(\widehat{Y}_{1U}^{\text{HT}}, ..., \widehat{Y}_{qU}^{\text{HT}}). \tag{3.5}$$

Analizar las propiedades de $\widehat{\theta}$ es fácil cuando f es una función lineal, es decir, cuando

$$\theta = a_0 + \sum_{j=1}^{q} a_j Y_j.$$

En este caso, el estimador (3.5) es

$$\widehat{\theta} = a_0 + \sum_{j=1}^{q} a_j \widehat{Y}_{jU}^{\text{HT}} \tag{3.6}$$

que es insesgado para θ, con varianza

$$\mathbb{V}(\widehat{\theta}) = \mathbb{V}(\sum_{j=1}^{q} a_j \widehat{Y}_{jU}^{\text{HT}}) = \sum_{j=1}^{q}\sum_{i=1}^{q} a_j a_i \mathbb{C}(\widehat{Y}_{jU}^{\text{HT}}, \widehat{Y}_{iU}^{\text{HT}}) \tag{3.7}$$

donde las covarianzas vienen definidas por $\mathbb{C}(\widehat{Y}_{jU}^{\text{HT}}, \widehat{Y}_{iU}^{\text{HT}}) = \sum_{k\in U}\sum_{l\in U} \Delta_{kl}\frac{y_{jk}}{\pi_k}\frac{y_{il}}{\pi_l}$. Cuando $j = i$, $\mathbb{C}(\widehat{Y}_{jU}^{\text{HT}}, \widehat{Y}_{jU}^{\text{HT}}) = \mathbb{V}\left[\widehat{Y}_{jU}^{\text{HT}}\right]$, es la varianza de $\widehat{Y}_{jU}^{\text{HT}}$. La varianza (3.7) se estima usando las covarianzas estimadas $\widehat{\mathbb{C}}(\widehat{t}_{j\pi}, \widehat{t}_{i\pi}) = \sum_{k\in s}\sum_{l\in s} \check{\Delta}_{kl}\frac{y_{jk}}{\pi_k}\frac{y_{il}}{\pi_l}$, donde $\check{\Delta}_{kl} = \frac{\Delta_{kl}}{\pi_{kl}}$

$$\widehat{\mathbb{V}}(\widehat{\theta}) = \sum_{j=1}^{q}\sum_{i=1}^{q} a_j a_i \widehat{\mathbb{C}}(\widehat{t}_{j\pi}, \widehat{t}_{i\pi}) \tag{3.8}$$

Cuando $j = i$, $\widehat{\mathbb{C}}(\widehat{Y}_{jU}^{\text{HT}}, \widehat{Y}_{iU}^{\text{HT}})$, es la varianza estimada $\widehat{\mathbb{V}}^{\text{HT}}\left[\widehat{Y}_{jU}^{\text{HT}}\right]$.

Comentario 11. Si f es una función lineal, la varianza dada por (3.7), así como su estimador, tienen expresiones alternativas que pueden facilitar los cálculos computacionales. El estimador (3.6) puede escribirse como

$$\widehat{\theta} = a_0 + \sum_{k\in s} \frac{u_k}{\pi_k},$$

con

$$u_k = \sum_{j=1}^{q} a_j y_{jk}$$

Entonces, la varianza (3.7) puede expresarse como

$$\mathbb{V}(\widehat{\theta}) = \mathbb{V}\left(\sum_{k\in s} \check{u}_k\right) = \sum k \in U \sum_{l\in U} \Delta_{kl}\frac{u_k}{\pi_k}\frac{u_l}{\pi_l} \tag{3.9}$$

que se puede estimar por

$$\widehat{\mathbb{V}}(\widehat{\theta}) = \sum_{k\in s}\sum_{l\in s} \widehat{\Delta}_{kl}\frac{u_k}{\pi_k}\frac{u_l}{\pi_l} \tag{3.10}$$

■

El cálculo de $\widehat{\mathbb{V}}(\widehat{\theta})$ usando la ecuación (3.10) requiere el cálculo previo de $\frac{u_k}{\pi_k}$ para $k \in s$, una operación sencilla. Una vez que se dispone de $\frac{u_k}{\pi_k}$, una suma doble no proporciona la estimación deseada (3.10). Claramente, esto es computacionalmente más sencillo que la ecuación (3.8), que implica el cálculo de $\frac{q(q+1)}{2}$ estimaciones de varianzas y covarianzas $\widehat{\mathbb{C}}(\widehat{Y}_{jU}^{\text{HT}}, \widehat{Y}_{iU}^{\text{HT}})$.

Nuestro principal interés en este tema es el caso en que $\theta = f(Y_1, \ldots, Y_q)$ es una función no lineal de los q totales. En este caso a menudo es imposible obtener resultados exactos del sesgo y la varianza del estimador $\widehat{\theta} = f(\widehat{Y}_{1U}^{\text{HT}}, ..., \widehat{Y}_{qU}^{\text{HT}})$. Para evitar dificultades, usamos la *técnica de linealización de Taylor*, que nos proporciona una expresión aproximada para la varianza de $\widehat{\theta}$, un estimador aproximado de esta varianza. Esta técnica también hace posible el cálculo de intervalos de confianza aproximados para θ. La técnica de linealización de Taylor se ha usado en varios campos de la estadística a lo largo del tiempo y data de la época de Gauss por lo menos.

La técnica aproxima el estimador no lineal $\widehat{\theta}$ mediante un pseudo-estimador, denotado por $\widehat{\theta}_0$, que es una función lineal de $\widehat{Y}_{1U}^{\text{HT}}, \ldots, \widehat{Y}_{qU}^{\text{HT}}$, con la que es más sencillo trabajar. $\widehat{\theta}_0$ habitualmente dependerá de determinadas incógnitas, por tanto no es un verdadero estimador. Si la aproximación es buena, $\widehat{\theta}_0$ se comportará aproximadamente como $\widehat{\theta}$ y lo podemos usar en la varianza $\mathbb{V}(\widehat{\theta}_0)$ como una aproximación de $\mathbb{V}(\widehat{\theta})$. También obtendremos un estimador $\widehat{\mathbb{V}}(\widehat{\theta}_0)$.

La técnica para obtener $\widehat{\theta}_0$ consiste en la aproximación de primer orden de Taylor de la función f, desarrollando alrededor del punto $Y_1, \ldots, Y_q$, y eliminando el término del resto de Taylor (sea cual sea su forma). Obtenemos

$$\widehat{\theta} \doteq \widehat{\theta}_0 = \theta + \sum_{j=1}^{q} a_j (\widehat{Y}_{jU}^{\text{HT}} - Y_j) \tag{3.11}$$

donde

$$a_j = \left. \frac{\partial f}{\partial \widehat{Y}_{jU}^{\text{HT}}} \right|_{(\widehat{Y}_{1U}^{\text{HT}}, \ldots, \widehat{Y}_{qU}^{\text{HT}})^t = (Y_1, \ldots, Y_q)} \tag{3.12}$$

De esta forma, para muestras grandes (cuando $\widehat{Y}_{1U}^{\text{HT}}, \ldots, \widehat{Y}_{jU}^{\text{HT}}$ tienen mayores probabilidad de tomar valores cerca de $Y_1, \ldots, Y_q$), el estimador $\widehat{\theta}$ es aproximadamente la variable aleatoria lineal $\widehat{\theta}_0$. La acuracidad numérica de la aproximación (3.11) variará de una muestra s a otra. Asumiremos, entonces, que el sesgo y la varianza de $\widehat{\theta}$ se pueden aproximar por las cantidades correspondientes para el estadístico lineal $\widehat{\theta}_0$.

A partir de ahora usaremos la varianza de un estadístico linealizado como una aproximación a un estimador (no lineal) más complejo. El símbolo $\text{AV}(\widehat{\theta}) = \mathbb{V}(\widehat{\theta}_0)$ indica la forma aproximada de la varianza de $\widehat{\theta}$ y esta varianza aproximada es igual a la varianza

exacta del estadístico linealizado $\widehat{\theta}_0$.

Por conveniencia en los cálculos, usaremos las ecuaciones del Comentario 11. Sean

$$u_k = \sum_{j=1}^{q} a_j y_{jk}. \tag{3.13}$$

Puesto que las ecuaciones 3.7 y 3.9 son equivalentes, la varianza aproximada de $\widehat{\theta}$ se obtienen como

$$\begin{aligned} \mathrm{AV}(\widehat{\theta}) = \mathbb{V}(\widehat{\theta}_0) &= \mathbb{V}\left(\sum_{j=1}^{q} a_j \widehat{Y}_{jU}^{\mathrm{HT}}\right) \\ &= \mathbb{V}(\sum_{k\in s} \frac{u_k}{\pi_k}) \\ &= \sum_{k\in U}\sum_{l\in U} \Delta_{kl} \frac{u_k}{\pi_k}\frac{u_l}{\pi_l}. \end{aligned} \tag{3.14}$$

Comentario 12. La expresión para $\mathbb{V}(\widehat{\theta})$ dada en (3.7) también se puede considerar una aproximación del error cuadrático medio de θ. Como $\mathbb{E}(\widehat{\theta}_0) = \theta$, se sigue que

$$\mathbb{MSE}(\widehat{\theta}) \doteq \mathbb{MSE}(\widehat{\theta}_0) = \mathbb{V}(\widehat{\theta}_0).$$

■

Las cantidades u_k que aparecen en (3.13) se pueden obtener generalmente sin mucha dificultad. Sin embargo, el problema reside en la estimación de (3.14). Los valores u_k dependen de $a_1, ..., a_q$, cantidades que a su vez dependen de los totales poblacionales desconocidos. Por tanto, los u_k son desconocidos. La solución habitual es reemplazar cada total desconocido en el que depende a_j por el correspondiente estimador de Horvitz-Thompson. De esta forma, llegaremos a un estimador $\widehat{a}_j$ de a_j, que nos permita obtener para cada $k \in s$ la variable

$$\widehat{u}_k = \sum_{j=1}^{q} \widehat{a}_j y_{jk}. \tag{3.15}$$

El último paso es entonces tomar la fórmula de la estimación de la varianza asociada con la ecuación (3.14) y sustituir los u_k desconocidos por $\widehat{u}_k$, y así obtenemos

$$\widehat{\mathbb{V}}(\widehat{\theta}) = \sum_{k\in s}\sum_{l\in s} \check{\Delta}_{kl} \frac{\widehat{u}_k}{\pi_k}\frac{\widehat{u}_l}{\pi_l}. \tag{3.16}$$

La justificación de este método es que $\widehat{u}_k$, siendo una función (posiblemente no lineal) de estimadores HT, es consistente para u_k. $\widehat{\mathbb{V}}(\widehat{\theta})$ es una función de estos estimadores consistentes $\widehat{u}_k$ y en muestras grandes debería comportarse como si se hubiese basado

en los u_k reales (desconocidos). De esta forma, se puede asumir que $\widehat{\mathbb{V}}(\widehat{\theta})$ es consistente para $\mathbb{V}(\widehat{\theta})$.

Como la expresión de AV mostrada en (3.14) es el punto de inicio para llegar al estimador (3.16), hemos, estrictamente hablando, conseguido obtener la varianza *aproximada*. Sin embargo, como la varianza aproximada $\mathrm{AV}(\widehat{\theta})$ y la varianza exacta $\mathbb{V}(\widehat{\theta})$ coinciden en muestras grandes, (3.16) también será un buen estimador de $\mathbb{V}(\widehat{\theta})$. Esto se ha demostrado con simulaciones en distintos casos. Esta técnica que se ha expuesto se resume a continuación:

Teorema 5

Para el parámetro poblacional $\theta = f(Y_1, \ldots, Y_q)$ donde $Y_1 = \sum_{k\in U} y_{1k}$, $\ldots$, $Y_q = \sum_{k\in U} y_{qk}$ son totales poblacionales, un estimador insesgado aproximado viene dado por

$$\widehat{\theta} = f(\widehat{Y}_{1U}^{\mathrm{HT}}, \ldots, \widehat{Y}_{qU}^{\mathrm{HT}}),$$

donde $\widehat{Y}_{1U}^{\mathrm{HT}}, \ldots, \widehat{Y}_{qU}^{\mathrm{HT}}$ son los estimadores HT correspondientes.

A través de la linealización de Taylor descrita en las ecuaciones (3.11) y (3.12), la varianza aproximada de $\widehat{\theta}$ se obtiene como

$$\mathrm{AV}(\widehat{\theta}) = \sum_{k\in U}\sum_{l\in U} \Delta_{kl}\frac{u_k}{\pi_k}\frac{u_l}{\pi_l},$$

donde $u_k = \sum_{j=1}^{q} a_j y_{jk}$, estando los coeficientes a_j definidos por (3.12).

Un estimador de la varianza viene dado por

$$\widehat{\mathbb{V}}(\widehat{\theta}) = \sum_{k\in s}\sum_{l\in s} \check{\Delta}_{kl}\frac{\widehat{u}_k}{\pi_k}\frac{\widehat{u}_l}{\pi_l}, \tag{3.17}$$

donde $\widehat{u}_k = \sum_{j=1}^{q} \widehat{a}_j y_{jk}$ con los $\widehat{a}_j$ obtenidos a partir de los a_j sustituyendo el estimador HT apropiado para cada total poblacional desconocido.

Demostración 5

Argumentación incluida más arriba.

Como siempre, cuando el diseño es de tamaño fijo, se puede usar el estimador de la

varianza de Sen-Yates-Grundy. En este caso viene dado por

$$\widehat{\mathbb{V}}(\widehat{\theta}) = -\frac{1}{2}\sum_{k\in s}\sum_{l\in s}\check{\Delta}_{kl}\left(\frac{\widehat{u}_k}{\pi_k} - \frac{\widehat{u}_l}{\pi_l}\right)^2. \tag{3.18}$$

El estimador de la varianza (3.18) siempre se puede usar para un diseño de tamaño fijo. Coincide con 3.17 para los diseños aleatorio simple sin reemplazamiento y estratificado aleatorio simple sin reemplazamiento.

Comentario 13. Tenemos que tener cuidado con el método de linealización de Taylor, ya que da lugar a varianzas subestimadas en el caso de muestras que no sean grandes. En muestras muy grandes, el sesgo del estimador de la varianza es nulo. La complejidad del estadístico es un factor importante. En caso de un estimador sencillo, como el estimador de Hájek para la media poblacional $\widehat{\bar{y}}_U^{\text{Hájek}} = \frac{\sum_{k\in s}\frac{y_k}{\pi_k}}{\sum_{k\in s}\frac{1}{\pi_k}}$, la subestimación del estimador de la varianza de Taylor puede no tener consecuencias incluso para muestras pequeñas, pero en el caso de estadísticas complejas, como el estimador de la varianza, la covarianza, o el coeficiente de correlación poblacional, pueden ser necesarias muestras grandes para que el sesgo sea despreciable. ■

3.5 Estimador de una razón: varianza y sesgo

Volvemos al problema de la estimación de una razón entre dos totales poblacionales desconocidos

$$R = \frac{Y_U}{Z_U} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k}. \tag{3.19}$$

Por ejemplo, si U es una población de hogares, y_k son los ingresos del hogar k-ésimo, y z_k es el número de personas en el k-ésimo hogar, entonces R son los ingresos per cápita por persona de los hogares de la población. O R puede representar las hectáreas dedicadas a trigo, $\sum_{k\in U} y_k$, divididas por el total de hectáreas dedicadas a cultivos en granjas, $\sum_{k\in U} z_k$, para una población de N granjas.

Si los dos totales desconocidos se estiman, respectivamente, por $\widehat{Y}_U^{\text{HT}} = \sum_{k\in s}\frac{y_k}{\pi_k}$ y $\widehat{Z}_U^{\text{HT}} = \sum_{k\in s}\frac{z_k}{\pi_k}$, el estimador resultante (no lineal) de R

$$\widehat{R} = \frac{\widehat{Y}_U^{\text{HT}}}{\widehat{Z}_U^{\text{HT}}}. \tag{3.20}$$

Pasamos ahora a analizar $\widehat{R}$ en profundidad. Hartley y Ross 1954 fijaron una cota superior del sesgo de $\widehat{R}$ para el caso de muestreo aleatorio simple sin reemplazamiento. A continuación incluimos resultados para el caso de un diseño muestral cualquiera.

Teorema 6

El sesgo del estadístico $\widehat{R}$ satisface

$$\frac{[\mathbb{E}\left[\widehat{R}\right] - R]^2}{\mathbb{V}\left(\widehat{R}\right)} \leq \frac{\mathbb{V}\left(\widehat{Z}_U^{\mathrm{HT}}\right)}{\left(Z_U\right)^2}. \tag{3.21}$$

Demostración 6

Consideremos la covarianza $\mathbb{C}\left(\widehat{R}, \widehat{Z}_U^{\mathrm{HT}}\right)$. Como $\widehat{R}\widehat{Z}_U^{\mathrm{HT}} = \widehat{Y}_U^{\mathrm{HT}}$, la covarianza puede escribirse como

$$\begin{aligned}
\mathbb{C}(\widehat{R}, \widehat{Z}_U^{\mathrm{HT}}) &= \mathbb{E}(\widehat{R}\widehat{Z}_U^{\mathrm{HT}}) - \mathbb{E}(\widehat{R})\mathbb{E}(\widehat{Z}_U^{\mathrm{HT}}) \\
&= \mathbb{E}(\widehat{Y}_U^{\mathrm{HT}}) - \mathbb{E}(\widehat{R})\mathbb{E}(\widehat{Z}_U^{\mathrm{HT}}) \\
&= Y_U - \mathbb{E}(\widehat{R})Z_U \\
&= -Z_U[\mathbb{E}(\widehat{R}) - R]
\end{aligned}$$

es decir,

$$\mathbb{E}(\widehat{R}) - R = -\frac{\mathbb{C}\left(\widehat{R}, \widehat{Z}_U^{\mathrm{HT}}\right)}{Z_U}.$$

Como el coeficiente de correlación al cuadrado está acotado superiormente por la unidad,

$$\begin{aligned}
[\mathbb{E}(\widehat{R}) - R]^2 &= \frac{[\mathbb{C}\left(\widehat{R}, \widehat{Z}_U^{\mathrm{HT}}\right)]^2}{\left(Z_U\right)^2} \\
&= \frac{[\rho\left(\widehat{R}, \widehat{Z}_U^{\mathrm{HT}}\right)]^2 \mathbb{V}(\widehat{R})\mathbb{V}\left(\widehat{Z}_U^{\mathrm{HT}}\right)}{\left[Z_U\right]^2} \\
&\leq \frac{\mathbb{V}(\widehat{R})\mathbb{V}\left(\widehat{Z}_U^{\mathrm{HT}}\right)}{\left[Z_U\right]^2}.
\end{aligned}$$

lo que demuestra el resultado. ■

El Teorema 6 nos lleva a la siguiente conclusión. Si

$$\mathrm{BR}(\widehat{R}) = \frac{\mathbb{B}(\widehat{R})}{\{\mathbb{V}(\widehat{R})\}^{\frac{1}{2}}} = \frac{\mathbb{E}(\widehat{R}) - R}{\{\mathbb{V}(\widehat{R})\}^{\frac{1}{2}}}$$

denota la razón de sesgo de $\widehat{R}$ (véase la ecuación (3.1) para la definición de razón de

sesgo), el resultado indica que

$$\left[\mathrm{BR}(\widehat{R})\right]^2 \leq \frac{\mathbb{V}\left(\widehat{Z}_U^{\mathrm{HT}}\right)}{[Z_U]^2}. \tag{3.22}$$

Es decir, si el error estándar relativo de $\widehat{Z}_U^{\mathrm{HT}}$, $\frac{[\mathbb{V}(\widehat{Z}_U^{\mathrm{HT}})]^{\frac{1}{2}}}{|Z_U|}$, se aproxima a cero cuando aumenta el tamaño muestral (que es lo que ocurre normalmente), la razón de sesgo de $\widehat{R}$ también tenderá a cero. Recordamos de la Sección 3.2 que esto es de vital importancia para la construcción de intervalos de confianza válidos. En otras palabras, aquí exponemos un ejemplo en el que se muestra que la razón de sesgo es pequeña para muestras grandes.

Ejemplo 4. Para un diseño muestral aleatorio simple sin reemplazamiento, la desigualdad (3.22) se puede escribir como

$$[\mathrm{BR}(\widehat{R})]^2 \leq \left(\frac{1}{n} - \frac{1}{N}\right)(\mathrm{cv}_{zU})^2,$$

donde $\mathrm{cv}_{zU} = \frac{S_{zU}}{\bar{z}_U}$ es el coeficiente de variación de z. Asumimos que z es siempre positivo. Esto demuestra que la razón de sesgo de $\widehat{R}$ en este caso tiende a cero como $n^{-\frac{1}{2}}$. ■

Aplicaremos ahora la técnica de linealización de Taylor descrita en la Sección 3.4 para encontrar una varianza aproximada de $\widehat{R}$ y encontrar un estimador de la varianza que pueda servir en los cálculos de un intervalo de confianza. El estimador $\widehat{R}$ es una función de dos variables aleatorias $\widehat{Y}_U^{\mathrm{HT}}$ y $\widehat{Z}_U^{\mathrm{HT}}$,

$$\widehat{R} = \frac{\widehat{Y}_U^{\mathrm{HT}}}{\widehat{Z}_U^{\mathrm{HT}}} = f(\widehat{Y}_U^{\mathrm{HT}}, \widehat{Y}_U^{\mathrm{HT}}).$$

Las derivadas parciales necesarias son

$$\frac{\partial \widehat{R}}{\partial \widehat{Y}_U^{\mathrm{HT}}} = \frac{1}{\widehat{Z}_U^{\mathrm{HT}}} \quad ; \quad \frac{\partial \widehat{R}}{\partial \widehat{Z}_U^{\mathrm{HT}}} = -\frac{\widehat{Y}_U^{\mathrm{HT}}}{\left(\widehat{Z}_U^{\mathrm{HT}}\right)^2}.$$

Evaluándolas en el punto $(Y_U, Z_U)^t$, obtenemos

$$a_1 = \frac{\partial \widehat{R}}{\partial \widehat{Y}_U^{\mathrm{HT}}}\Big|_{(Y_U, Z_U)^t} = \frac{1}{Z_U}$$

$$a_2 = \frac{\partial \widehat{R}}{\partial \widehat{Z}_U^{\mathrm{HT}}}\Big|_{(Y_U, Z_U)^t} = \frac{Y_U}{Z_U^2} = \frac{R}{Z_U}.$$

Ahora, a partir de (3.12) y (3.13),

$$u_k = a_1 y_k + a_2 z_k = \frac{1}{Z_U}(y_k - R z_k)$$

y

$$\widehat{u}_k = \frac{1}{\widehat{Z}_U^{\mathrm{HT}}}(y_k - \widehat{R} z_k)$$

y se obtiene el siguiente resultado.

Teorema 7

Usando la linealización de Taylor, el estadístico de razón $\widehat{R} = \frac{\widehat{t}_y}{\widehat{t}_z}$ se aproxima de la siguiente forma

$$\widehat{R} \doteq \widehat{R}_0 = R + \frac{1}{Z_U}\sum_{k\in s}\frac{y_k - R z_k}{\pi_k}. \tag{3.23}$$

El estimador $\widehat{R}$ es aproximadamente insesgado para R, con la varianza aproximada

$$\mathrm{AV}(\widehat{R}) = \mathbb{V}(\widehat{R}_0) = \frac{1}{Z_U^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_k - R z_k}{\pi_k}\frac{y_l - R z_l}{\pi_l}. \tag{3.24}$$

El estimador de la varianza es

$$\widehat{\mathbb{V}}(\widehat{R}) = \frac{1}{\left(\widehat{Z}_U^{\mathrm{HT}}\right)^2}\sum_{k\in s}\sum_{l\in s}\check{\Delta}_{kl}\frac{y_k - \widehat{R} z_k}{\pi_k}\frac{y_l - \widehat{R} z_l}{\pi_l}. \tag{3.25}$$

Demostración 7

Aplíquese el Teorema 5 a la función $R = f(Y_U, Z_U)$.

Comentario 14. Las siguientes expresiones también pueden ser útiles a veces

$$\widehat{R}_0 = R + \frac{1}{Z_U}(\widehat{Y}_U^{\mathrm{HT}} - R\widehat{Z}_U^{\mathrm{HT}}), \tag{3.26}$$

$$\mathrm{AV}(\widehat{R}) = \mathbb{V}(\widehat{R}_0) = \frac{1}{Z_U^2}\left(\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}}) + R^2\mathbb{V}(\widehat{Z}_U^{\mathrm{HT}}) - 2R\mathbb{C}(\widehat{Y}_U^{\mathrm{HT}}, \widehat{Z}_U^{\mathrm{HT}})\right), \tag{3.27}$$

$$\widehat{\mathbb{V}}(\widehat{R}) = \frac{1}{\widehat{t}_z^2}[\widehat{\mathbb{V}}(\widehat{Y}_U^{\mathrm{HT}}) + \widehat{R}^2\widehat{\mathbb{V}}(\widehat{Z}_U^{\mathrm{HT}}) - 2\widehat{R}\widehat{\mathbb{C}}(\widehat{Y}_U^{\mathrm{HT}}, \widehat{Z}_U^{\mathrm{HT}})]. \tag{3.28}$$

Comentario 15. Bajo la aproximación mostrada en (3.23), $\mathbb{E}(\widehat{R}) \doteq \mathbb{E}(\widehat{R}_0) = R$. En otras palabras, el sesgo de $\widehat{R}$, aunque no nulo, se aproxima por cero. El que el sesgo de $\widehat{R}$ no

sea detectable indica una aproximación bastante general, por lo menos, para muestras pequeñas. Se puede obtener una expresión mejorada para el sesgo extendiendo el desarrollo de Taylor para incluir también los términos de segundo orden.

Ejemplo 5. Consideremos el diseño muestral aleatorio simple sin reemplazamiento, con un tamaño muestral $n = fN$, donde f es la fracción de muestreo. Entonces $\widehat{Y}_U^{\text{HT}} = N\bar{y}_s$, $\widehat{Z}_U^{\text{HT}} = N\bar{z}_s$, y $\widehat{R} = \frac{\bar{y}_s}{\bar{z}_s}$. La aproximación lineal dada por (3.23), o equivalentemente por (3.26) es

$$\widehat{R}_0 = R + \frac{1}{\bar{z}_U}\frac{1}{n}\sum_{k\in s}(y_k - Rz_k) = R + \frac{\bar{y}_s - R\bar{z}_s}{\bar{z}_U}.$$

La aproximación de la varianza dada por (3.24) o por (3.27) es

$$\begin{aligned}\text{AV}(\widehat{R}) &= \frac{1}{\bar{z}_U^2}\frac{1-f}{n}\frac{1}{N-1}\sum_{k\in U}(y_k - Rz_k)^2 \\ &= \frac{1}{\bar{z}_U^2}\frac{1-f}{n}(S_{yU}^2 + R^2 S_{zU}^2 - 2RS_{yzU}),\end{aligned}$$

donde S_{yzU} es la covarianza poblacional entre las variables y y z. A partir de (3.25) o de (3.28) obtenemos

$$\begin{aligned}\widehat{\mathbb{V}}(\widehat{R}) &= \frac{1}{\bar{z}_s^2}\frac{1-f}{n}\frac{1}{n-1}\sum_{s}(y_k - \widehat{R}z_k)^2 \\ &= \frac{1}{\bar{z}_s^2}\frac{1-f}{n}(S_{ys}^2 + \widehat{R}^2 S_{zs}^2 - 2\widehat{R}S_{yzs}),\end{aligned}$$

donde

$$S_{ys}^2 = \frac{1}{n-1}\sum_{s}(y_k - \bar{y}_s)^2$$

$$S_{yzs} = \frac{1}{n-1}\sum_{s}(y_k - \bar{y}_s)(z_k - \bar{z}_s)$$

y S_{zs}^2 es análogo a S_{ys}^2. ■

Ejemplo 6. Los resultados de este tema se pueden aplicar al muestreo en dos o más etapas. Consideremos muestreo aleatorio simple sin reemplazamiento en ambas etapas. En primer lugar, se selecciona una muestra aleatoria simple sin reemplazamiento s_I de n_I PSUs de un conjunto de N_I conglomerados. Dentro de cada PSU seleccionada, se obtiene una muestra aleatoria simple sin reemplazamiento s_i de n_i elementos del conglomerado U_i de tamaño N_i. La razón $R = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k} = \frac{Y_U}{Z_U}$ se estima con $\widehat{R} = \frac{\widehat{Y}_U^{\text{HT}}}{\widehat{Z}_U^{\text{HT}}}$, donde $\widehat{Y}_U^{\text{HT}} = \frac{N_I}{n_I}\sum_{i\in s_I} N_i\bar{y}_{s_i}$ y la expresión de $\widehat{Z}_U^{\text{HT}}$ es análoga. Por tanto,

$$\widehat{R} = \frac{\sum_{i\in s_I} N_i\bar{y}_{s_i}}{\sum_{i\in s_I} N_i\bar{z}_{s_i}}.$$

El estimador de la varianza se obtiene de forma análoga. La varianza del estimador HT, para un diseño arbitrario, se estima por $\sum_{k\in s}\sum_{l\in s}\check{\Delta}_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$. Compare esta fórmula

con la varianza del estimador para $\widehat{R}$ dada por (3.25). Para ir de la primera fórmula a la segunda, sustituimos y_k por $e_k = y_k - \widehat{R}z_k$, y multiplicamos por $\frac{1}{\left[\widehat{Z}_U^{\text{HT}}\right]^2}$. Realizamos esta operación para obtener la ecuación de la estimación de la varianza de un muestreo bietápico con muestreo aleatorio simple sin reemplazamiento en ambas etapas[3]. Después de simplificar, el estimador de la varianza resultante para $\widehat{R}$ es

$$\widehat{\mathbb{V}}(\widehat{R}) = \frac{1}{\left(\sum_{s_I} \frac{N_i}{n_I}\bar{z}_{s_i}\right)^2}\left[\left(\frac{1}{n_I} - \frac{1}{N_I}\right)\frac{\sum_{i\in s_I} N_i^2 \bar{e}_{s_i}^2}{n_I - 1} + \frac{1}{n_I N_I}\sum_{i\in s_I} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)\frac{\sum_{k\in s_i}(e_k - \bar{e}_{s_i})^2}{n_i - 1}\right],$$

con $\bar{e}_{s_i} = \bar{y}_{s_i} - \widehat{R}\bar{z}_{s_i}$. ■

Este método para estimar la razón R lleva a uno de los estimadores más utilizados para un total poblacional: el estimador de razón. El estimador de razón necesita información auxiliar. Si y es una variable de interés, el total poblacional $Y_U = \sum_{k\in U} y_k$ se puede escribir como

$$Y_U = Z_U \frac{Y_U}{Z_U} = Z_U R.$$

Siempre y cuando Z_U sea una cantidad conocida, R se puede estimar mediante (3.20) y, por tanto, Y_U se puede estimar mediante el estimador de razón

$$\widehat{Y}_U^{\text{Rat}} = Z_U \widehat{R} = Z_U \frac{\widehat{Y}_U^{\text{HT}}}{\widehat{Z}_U^{\text{HT}}}.$$

Una característica muy importante del estimador de razón es que el total $Z_U = \sum_{k\in U} z_k$ de la variable auxiliar z tiene que ser conocida de forma exacta para que el método funcione. Además, y_k y z_k tienen que ser conocidos para $k \in s$. Enfatizamos que debe ser conocido "de forma exacta", ya que un valor no acurado de Z_U dará lugar a un sesgo no despreciable de $\widehat{Y}_U^{\text{Rat}}$. Cabe mencionar, sin embargo, que en tanto en cuanto Z_U sea conocido, la variable z puede ser considerada una versión "barata" de y. Lo que es imprescindible para una estimación precisa con este método es que y_k sea aproximadamente proporcional a z_k en toda la población (véase el Comentario 16). Por ejemplo, z_k puede ser un proxy[4] de la variable de estudio o una medida obtenida fácilmente y a bajo coste de toda la población (por ejemplo, a partir de un registro administrativo), mientras que y_k es la variable de estudio precisa, pero con un coste de medición mayor, que se obtiene solo a partir de la muestra s.

Las características básicas del estimador de razón se obtienen de forma inmediata del Teorema 7 y se resumen en el siguiente resultado.

[3] Véase el tema 1 de este mismo bloque.
[4] Variable altamente correlacionada.

Teorema 8

Si $Y_U = \sum_{k\in U} y_k$ es el total desconocido de una variable de estudio y y si $Z_U = \sum_{k\in U} z_k$ es el total conocido de una variable auxiliar z, entonces el estimador de razón

$$\widehat{Y}_U^{\text{Rat}} = Z_U \widehat{R} = Z_U \frac{\widehat{Y}_U^{\text{HT}}}{\widehat{Z}_U^{\text{HT}}}$$

es aproximadamente insesgado para Y_U. Su varianza aproximada viene dada por

$$\text{AV}(\widehat{Y}_U^{\text{Rat}}) = \sum_{k\in U}\sum_{l\in U} \Delta_{kl} \frac{y_k - Rz_k}{\pi_k} \frac{y_l - Rz_l}{\pi_l}. \tag{3.29}$$

El estimador de la varianza es

$$\widehat{\mathbb{V}}(\widehat{Y}_U^{\text{Rat}}) = \left(\frac{Z_U}{\widehat{Z}_U^{\text{HT}}}\right)^2 \sum_{k\in s}\sum_{l\in s} \check{\Delta}_{kl} \frac{y_k - \widehat{R}z_k}{\pi_k} \frac{y_l - \widehat{R}z_l}{\pi_l}. \tag{3.30}$$

Demostración 8

Elemental a partir del Teorema 7 y la definición de estimador de razón del total poblacional Y_U.

Comentario 16. La expresión para la varianza aproximada AV dada por (3.29) es cero si los residuos $y_k - Rz_k$ son cero para todos los elementos $k \in U$. Esto no ocurrirá en la práctica, pero frecuentemente podremos tener un conjunto de residuos que, aunque no sean nulos, son pequeños. Entonces, la varianza aproximada AV también será pequeña. Es decir, el estimador de razón es muy preciso cuando los pares poblacionales $(y_k, z_k)^t$ se encuentren dispersos cerca de una línea recta que pase por el origen y con una pendiente R (desconocida). Este modelo de regresión se puede decir que genera el estimador de razón.

Comentario 17. Es importante también hacer notar que el estimador de razón $\widehat{Y}_U^{\text{Rat}}$ para el total poblacional Y_U utilizando información auxiliar Z_U tiene una propiedad muy importante en el uso de la información auxiliar en general: $\widehat{Y}_U^{\text{Rat}}$ es calibrado respecto al total poblacional Z_U. Para verlo, escribamos $\widehat{Y}_U^{\text{Rat}} = \sum_{k\in s} \omega_{ks}(\mathbf{z}) y_k$, donde $\omega_{ks}(\mathbf{z})$ son los pesos de muestreo dados por

$$\omega_{ks}(\mathbf{z}) = \frac{1}{\pi_k} \frac{Z_U}{\widehat{Z}_U^{\text{HT}}(s)}.$$

Es inmediato demostrar que

$$\sum_{k\in s} \omega_{ks}(\mathbf{z}) z_k = Z_U \quad \text{para todo } s.$$

■

3.6 Estimación de otros parámetros poblacionales

La técnica de linealización de Taylor puede emplearse bajo el mismo enfoque para encontrar estimadores de parámetros poblacionales de distinto naturaleza, así como la estimación de sus varianzas. Por mencionar algunos:

- Media poblacional $\bar{y}_U$.
  Considerando la media poblacional $\bar{y}_U = \frac{Y_U}{N} = f(Y_U, N)$ como función de dos totales $Y_U = \sum_{k\in U} y_k$ y $N = \sum_{k\in U} z_k$, con $z_k = 1$ para todo $k \in U$.
- Media poblacional en un dominio $\bar{y}_{U_d}$, con $U_d \subset U$.
  Nuevamente, podemos considerar $\bar{y}_{U_d} = \frac{Y_{U_d}}{N_d} = f(Y_{U_d}, N_{U_d})$ como función de dos totales $Y_U = \sum_{k\in U} y_k\delta_{kU_d}$ y $N_d = \sum_{k\in U} z_k\delta_{kU_d}$, donde $z_k = 1$ para todo $k \in U$ y $\delta_{kU_d} = 1$ si $k \in U_d$ y 0 si $k \notin U_d$.
- Varianza y covarianza poblacional $S_{yzU} = \frac{1}{N-1}\sum_{k\in U}(y_k - \bar{y}_U)(z_k - \bar{z}_U)$.
  Considerando la covarianza poblacional $S_{yzU} = \frac{V_U}{N-1} - \frac{Y_U Z_U}{N(N-1)} = f(V_U, Y_U, Z_U, N)$, donde $v = yz$, $X_U = \sum_{k\in U} x_k$, para $x = y, z, v$, como función de los totales V_U, Y_U, Z_U y N_U.
- Coeficientes de regresión poblacionales $\mathbf{B}_U$.
  Si $\mathbf{y} = (y_1, \ldots, y_N)^t$ es el vector de variables de estudio y $\mathbf{Z} = [z_{ij}]_{\substack{1\le i\le q\\ 1\le j\le N}}$ es la matriz de coeficientes de las variables auxiliares $\mathbf{z}_k = (z_{1k}, \ldots, z_{qk})^t$ para cada $k \in U$, entonces

  $$\mathbf{B}_U = \left(\mathbf{Z}\mathbf{Z}^t\right)^{-1}\mathbf{Z}\mathbf{y} = \left[\sum_{k\in U}\mathbf{z}_k\mathbf{z}_k^t\right]^{-1}\sum_{k\in U}\mathbf{z}_k y_k$$

  puede también considerarse función de totales poblacionales.
- Los cuantiles poblaciones c_p de la variable y.
  Considerando $c_p = F_Y^{*-1}(p)$, donde F_Y^* es la función empírica de distribución de la variable y, $F_Y^*(y) = \frac{|A_y|}{N} = f(T_U(y), N)$, donde $A_y = \{k \in U : y_k \le y\}$, como la función inversa de una función de totales poblacionales.

Como ejemplo ilustrativo, incluimos un procedimiento de estimación de la mediana poblacional (véase Särndal, Swensson y Wretman 1992).

1. Obténgase una función empírica de distribución estimada $\widehat{F}_Y(y)$.
2. Estímese $M = F_Y^{-1}(1/2)$ mediante $\widehat{M} = \widehat{F}_Y^{-1}(1/2)$.

Para estimar la función empírica de distribución F_Y definimos las variables $z_k(y)$ mediante

$$z_k(y) = \begin{cases} 1 & \text{si } y_k \le y, \\ 0 & \text{si } y_k > y. \end{cases}$$

Entonces,

$$F_Y(y) = \frac{\sum_{k\in U} z_k(y)}{N} = \frac{Z_U(y)}{N}.$$

Esta expresión es en realidad una media poblacional, de modo que podemos utilizar el estimador de Hájek

$$\widehat{F}_Y(y) = \widehat{Z}_U^{\text{Hájek}}(y) = \frac{\sum_{k\in s} \frac{z_k(y)}{\pi_k}}{\sum_{k\in U} \frac{1}{\pi_k}}.$$

Esta función $\widehat{F}_Y(y)$ es una función escalonada no decreciente cuyos valores van desde 0 hasta 1 (como la función F_Y). Para encontrar los valores inversos de $\widehat{F}_Y$ es preciso computar sus valores en la zona central del recorrido de valores de y, de modo que $\widehat{M} = \widehat{F}_Y^{-1}(1/2)$.

Bibliografía

Fuller, W.A. (2009). *Sampling Statistics*. Wiley.

Hájek, J. (1971). *Comment on "An essay on the logical foundations of survey sampling by D. Basu", 'The Foundations of Survey Sampling"*.

Hartley, H.O. y A. Ross (1954). "Unbiased ratio estimators". En: *Nature* 174, págs. 270-271.

Isaki, C.T. y W.A. Fuller (1982). "Survey design under the regression superpopulation model". En: *Journal of the American Statistical Association* 77, págs. 89-96.

Robinson, P.M. y C.-E. Särndal (1983). "Asymptotic properties of the generalized regression estimatior in probability sampling". En: *Sankhya* 45, págs. 240-248.

Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.

Tema 4

El estimador lineal de regresión generalizado. Variables auxiliares. estimador en diferencias. Introducción al estimador lineal de regresión generalizado (GREG). Expresiones alternativas para el estimador lineal de regresión generalizado. Varianza y sus estimaciones. El papel del modelo.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía.

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

4.1 El estimador lineal de regresión generalizado: introducción

El énfasis puesto en el uso de información auxiliar para mejorar la precisión de las estimaciones es característica de la teoría de muestreo. El estimador de regresión que se presenta en ese tema es un tipo de estimador que intenta hacer un uso eficiente de la información auxiliar de una población.

4.2 Variables auxiliares

Por lo general, una variable auxiliar es cualquier variable sobre la que tenemos información antes del muestreo. Habitualmente, suponemos que la información a priori de una variable auxiliar es completa, es decir, que el valor de la variable, digamos x, es conocida para cada uno de los N elementos de la población, de forma que los valores $x_1, \ldots, x_N$, están disponibles antes de muestrear. Una variable auxiliar ayuda en la estimación de la variable de análisis. El objetivo es obtener un estimador con mayor acuracidad.

Algunos marcos muestrales, junto con las características de identificación de las unida-

des, contienen una o más variables auxiliares, o información que puede ser transformada en variables auxiliares mediante manipulaciones numéricas sencillas.

Los valores de las variables auxiliares pueden obtenerse de registros administrativos e incluirse en el marco mediante *record linkage* (véase el tema 14[1] del bloque "Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas" del grupo de materias comunes). Hay que tener cuidado con los problemas prácticos que nos podemos encontrar relacionados con la periodicidad de los datos, el uso de codificación distinta, etc.

Ejemplo 7. El Directorio Central de Empresas (DIRCE) [2] es el marco muestral usado en el INE para las encuestas a empresas. Es un marco bastante complejo y está basado en la información de varias fuentes. Por un lado, utiliza información de registros administrativos, como el *Impuesto sobre el Valor Añadido*, el *Impuesto de Sociedades* y el *Impuesto sobre la Renta de las Personas Físicas* de la Agencia Estatal de Administración Tributaria, el *Registro de Cuentas de Cotización* y el *Registro de Trabajadores Activos en Cuenta Propia* de la Seguridad Social, los movimientos del Registro Mercantil y también información de las encuestas estructurales y coyunturales de empresas realizadas por el INE.

Es necesaria la actualización continua para registrar los 'nacimientos' (nuevas empresas que inician su actividad), 'muertes' (finalización de la actividad de la empresa) y cambios en la clasificación basados en el tamaño, la actividad o su ubicación geográfica. Toda esta información genera variables auxiliares que pueden emplearse tanto para la construcción del diseño muestral como para la construcción de estimadores. ■

Supondremos ahora que disponemos de una o más variables auxiliares en el marco muestral. En los temas 3 a 6 del bloque "Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas" del grupo de materias comunes vimos que la información auxiliar se puede usar en la fase de diseño de una encuesta para crear un diseño muestral que mejore la precisión del estimador de Horvitz-Thompson a través de las probabilidades de inclusión de primer orden (por ejemplo, con probabilidades proporcionales al tamaño). Otro ejemplo es el uso de la información auxiliar para construir estratos en el muestreo estratificado.

En este tema la información auxiliar se usará en la fase de estimación y servirá para reducir la varianza en comparación con el estimador HT. La hipótesis básica para el uso de variables auxiliares es que presentan correlación con la variable de análisis. Esta correlación se usa de forma ventajosa en el estimador lineal de regresión generalizado (GREG, por sus siglas en inglés *Generalised REGression*).

[1] Tema 14. Record linkage. Introducción. Visión de conjunto de los métodos. Preparación de los datos.

[2] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736160707&menu=ultiDatos&idp=1254735576550

4.3 Estimador en diferencias

Antes de presentar el estimador GREG, veamos primero el estimador en diferencias. Por dos razones:

a. El estimador en diferencias es más sencillo matemáticamente.

b. La comprensión del estimador en diferencias simplifica la transición al estimador GREG.

Supondremos que disponemos de J variables auxiliares, denotadas por $x_1, \ldots, x_j, \ldots, \ldots, x_J$. El valor de la j-ésima variable x para el k-ésimo elemento poblacional se denota por x_{jk}. Para el k-ésimo elemento definimos el vector $\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})^t$. La variable de estudio y toma el valor y_k para el k-ésimo elemento.

Se supone que los valores $y_1, \ldots, y_N$ son desconocidos e inaccesibles antes del muestreo, mientras que $\mathbf{x}_1, \ldots, \mathbf{x}_N$ son conocidos. El parámetro poblacional a estimar es el total poblacional de y, $Y_U = \sum_{k \in U} y_k$.

Una muestra s se obtiene a partir de U mediante el diseño muestral $p(\cdot)$ con probabilidades de inclusión $\pi_k > 0$ y $\pi_{kl} > 0$ para todo $k, l \in U$. Para cada $k \in s$, observamos y_k y el vector auxiliar asociado $\mathbf{x}_k$. El problema es estimar Y_U cuando hemos observado $(y_k, \mathbf{x}_k)$ para los elementos $k \in s$ y cuando también conocemos $\mathbf{x}_k$ para $k \in U - s$, el conjunto de unidades no muestreadas.

La principal idea es el uso de la información auxiliar para formar un conjunto de N valores y aproximados, denotados por $y_1^0, \ldots, y_N^0$, de forma que y_k^0 sea, por lo menos, una buena aproximación de y_k y preferiblemente un valor que se concentre alrededor de y_k.

El valor aproximado y_k^0 se obtiene como una combinación lineal de los valores conocidos $x_{1k}, \ldots, x_{jk}$

$$y_k^0 = \sum_{j=1}^{J} A_j x_{jk} \equiv \mathbf{A}^t \mathbf{x}_k \tag{4.1}$$

donde suponemos que $A = (A_1, \ldots, A_J)^t$ es un vector de coeficientes *conocido*. Más adelante los coeficientes serán reemplazados por cantidades estimadas a partir de la muestra. Obviamente, y_k^0 se puede calcular para todos los $k \in U$, ya que los valores x son conocidos para toda la población.

Cuando estudios previos o análisis sugieren una relación lineal aproximada para $k = 1, \ldots, N$,

$$y_k \doteq \sum_{j=1}^{J} A_j x_{jk} = \mathbf{A}^t \mathbf{x}_k \tag{4.2}$$

con un vector de coeficientes $A = (A_1, \ldots, A_J)^t$ conocido, es razonable elegir y_k^0 de acuerdo con (4.1). La variable de estudio puede ser explicada en gran medida por las variables auxiliares mediante la relación lineal mostrada en (4.2), donde los A_j son conocidos.

El total poblacional a estimar puede ahora escribirse como

$$Y_U = \sum_{k\in U} y_k = \sum_{k\in U} y_k^0 + \sum_{k\in U}(y_k - y_k^0) = \sum_{k\in U} y_k^0 + \sum_{k\in U} D_k \tag{4.3}$$

denotando $D_k = y_k - y_k^0$ para $k = 1, .., N$. El total aproximado $\sum_{k\in U} y_k^0$ en la ecuación (4.3) es conocido, ya que los $y_1^0, \ldots, y_N^0$ son conocidos, mientras que el total de las diferencias, $\sum_{k\in U} D_k$, es desconocido, ya que los $y_1, \ldots, y_N$ son desconocidos.

Nuestro primer pensamiento es usar el estimador de Horvitz-Thompson para obtener un estimador insesgado del término desconocido en (4.3), pues a priori cometeremos menor error al estimar cantidades intrínsecamente pequeñas que cantidades intrínsecamente grandes. El resultado es conocido como el *estimador en diferencias*:

Definición 7

Se define el estimador en diferencias como

$$\widehat{Y}_U^{\text{Dif}} = \sum_{k\in U} y_k^0 + \sum_{k\in s} \frac{D_k}{\pi_k}, \tag{4.4}$$

donde $D_k = y_k - y_k^0$.

Sus propiedades se establecen de manera casi inmediata:

Teorema 9

El estimador en diferencias $\widehat{Y}_U^{\text{Dif}}$ es insesgado para $Y_U = \sum_{k\in U} y_k$. Su varianza viene dada por

$$\mathbb{V}(\widehat{Y}_U^{\text{Dif}}) = \sum_{k\in U}\sum_{l\in U} \Delta_{kl} \frac{D_k}{\pi_k}\frac{D_l}{\pi_l}, \tag{4.5}$$

cuyo estimador insesgado es

$$\widehat{\mathbb{V}}(\widehat{Y}_U^{\text{Dif}}) = \sum_{k\in s}\sum_{l\in s} \check{\Delta}_{kl} \frac{D_k}{\pi_k}\frac{D_l}{\pi_l} \tag{4.6}$$

Demostración 9

Puesto que $Y_U = \sum_{k\in U} y_k^0 + \sum_{k\in U} D_k$ y el estimador $\widehat{D}_U^{\text{HT}} \sum_{k\in s} \frac{D_k}{\pi_k}$ es insesgado para $D_U = \sum_{k\in U} D_k$, el resultado se sigue fácilmente de las propiedades del estimador de Horvitz-Thompson de un total poblacional.

Vemos en la fórmula 4.5 de la varianza que, siguiendo la motivación original para construir el estimador en diferencias, éste funciona mejor cuando las diferencias D_k son cero o muy próximas a cero para todo k, es decir, cuando los valores aproximados y_k^0 son suficientemente buenos de forma que $D_k = y_k - y_k^0$ son pequeños.

Comentario 18. Bajo un diseño muestral de tamaño muestral fijo, la varianza 4.5 puede escribirse mediante la expresión de Sen-Yates-Grundy como

$$\mathbb{V}(\widehat{Y}_U^{\text{Dif}}) = -\frac{1}{2}\sum_{k\in U}\sum_{l\in U} \Delta_{kl}(\check{D}_k - \check{D}_l)^2, \tag{4.7}$$

cuyo estimador insesgado alternativo es

$$\widehat{\mathbb{V}}(\widehat{Y}_U^{\text{Dif}}) = -\frac{1}{2}\sum_{k\in s}\sum_{l\in s} \check{\Delta}_{kl}(\check{D}_k - \check{D}_l)^2. \tag{4.8}$$

■

Ejemplo 8. Si $y_k^0 = x_k$ y $D_k = y_k - x_k$, no es difícil ver que para el muestreo aleatorio simple sin reemplazamiento con tamaño muestral $n = fN$ se cumple

$$\mathbb{V}(\widehat{Y}_U^{\text{Dif}}) = N^2 \frac{1-f}{n}(S_{yU}^2 + S_{xU}^2 - 2S_{xyU}),$$

donde S_{yU}^2 y S_{xU}^2 son las cuasivarianzas poblacionales de y y x, respectivamente y $S_{xyU} = \frac{1}{N-1}\sum_{k\in U}(x_k - \bar{x}_U)(y_k - \bar{y}_U)$ es la cuasicovarianza poblacional. Si la correlación poblacional $r = \frac{S_{xyU}}{S_{xU}S_{yU}}$ es alta, el estimador en diferencias a menudo conlleva un gran reducción de la varianza comparado con el estimador HT $\widehat{Y}_U^{\text{HT}}$. Tenemos $\text{deff}\left(\widehat{Y}_U^{\text{Dif}}\right) = \frac{\mathbb{V}(\widehat{Y}_U^{\text{Dif}})}{\mathbb{V}(\widehat{Y}_U^{\text{HT}})} = 1 + (\frac{S_{xU}}{S_{yU}})^2 - 2r\frac{S_{xU}}{S_{yU}}$. Por ejemplo, si $r = 0{,}90$ y $S_{xU}/S_{yU} = 1{,}5$, $\text{deff}\left(\widehat{Y}_U^{\text{Dif}}\right) = 0{,}28$, que es una reducción considerable de la varianza.

En general, bajo muestreo aleatorio simple sin reemplazamiento y con $y_k^0 = x_k$, $\mathbb{V}(\widehat{Y}_U^{\text{Dif}}) < \mathbb{V}(\widehat{Y}_U^{\text{HT}})$ si y solo si $r > \frac{1}{2}\frac{S_{xU}}{S_{yU}}$. ■

Una forma alternativa del estimador en diferencias es considerarlo como una posible mejora del estimador de Horvitz-Thompson $\widehat{Y}_U^{\text{HT}} = \sum_{k\in s} \check{y}_k = \sum_{k\in s} \frac{y_k}{\pi_k}$. Supongamos que los valores aproximados son combinaciones lineales de acuerdo con (4.1). Entonces

(4.4) puede escribirse como

$$\widehat{Y}_U^{\text{Dif}} = \widehat{Y}_U^{\text{HT}} + \sum_{j=1}^{J} A_j (X_{jU} - \widehat{X}_{jU}^{\text{HT}}), \tag{4.9}$$

donde $\widehat{X}_{jU}^{\text{HT}} = \sum_{k\in s} \frac{x_{jk}}{\pi_k}$ para $j = 1, \ldots, J$ es el estimador HT del total de x_j, $X_{jU} = \sum_{k\in U} x_{jk}$. Es decir, el estimador en diferencias es igual al estimador HT más un término independiente.

En la situación ideal, cuando (4.2) se verifica exactamente, es decir, cuando $y_k = y_k^0 = \sum_{j=1}^{J} A_j x_{jk}$, para $k = 1, \ldots, N$, el término independiente cancelará el error en el estimador HT de forma que el estimador en diferencias está completamente libre de error. El error en el estimador HT es $\widehat{Y}_U^{\text{HT}} - Y_U = \sum_{k\in s} \frac{y_k}{\pi_k} - \sum_{k\in U} y_k$. El término independiente, cuando se verifica (4.2) de forma exacta, es $\sum_{j=1}^{J} A_j (X_{jU} - \widehat{X}_{jU}^{\text{HT}}) = -(\sum_{k\in s} \frac{y_k}{\pi_k} - \sum_{k\in U} y_k)$. El error del estimador en diferencias $\widehat{Y}_U^{\text{Dif}} - Y_U$ es, por tanto, nulo.

Comentario 19. El estimador en diferencias funciona bien cuando hay una relación aproximadamente lineal entre las variables y y x, y, muy importante, cuando los coeficientes $A_1, \ldots, A_J$, se eligen de forma adecuada. Si estos coeficientes se escogen pobremente, la varianza del estimador en diferencia puede ser considerablemente mayor que la del estimador HT. Un método de cómputo de los coeficientes A_j puede consultarse en (Särndal, Swensson y Wretman 1992, sección 6.8). ■

4.4 Introducción al estimador lineal de regresión generalizado (GREG)

Al igual que con el estimador en diferencias, supondremos que disponemos de J variables auxiliares, denotadas por $x_1, \ldots, x_j, \ldots, \ldots, x_J$. El valor de la j-ésima variable x para el k-ésimo elemento poblacional se denota por x_{jk}. Para el k-ésimo elemento definimos el vector $\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})^t$. La variable de estudio y toma el valor y_k para el k-ésimo elemento.

Se supone que los valores $y_1, \ldots, y_N$ son desconocidos e inaccesibles antes del muestreo, mientras que $\mathbf{x}_1, \ldots, \mathbf{x}_N$ son conocidos. El parámetro poblacional a estimar es el total poblacional de y, $Y_U = \sum_{k\in U} y_k$.

Una muestra s se obtiene a partir de U mediante el diseño muestral $p(\cdot)$ con probabilidades de inclusión $\pi_k > 0$ y $\pi_{kl} > 0$ para todo $k, l \in U$. Para cada $k \in s$, observamos y_k y el vector auxiliar asociado $\mathbf{x}_k$. El problema es ahora también estimar Y_U cuando hemos observado $(y_k, \mathbf{x}_k)$ para los elementos $k \in s$ y cuando también conocemos $\mathbf{x}_k$ para $k \in U - s$, el conjunto de unidades no muestreadas, pero no supondremos que los coeficientes lineales A_j son conocidos, sino que deben estimarse. La herramienta estadística que se empleará para estimar estos coeficientes es el análisis de regresión

lineal. Cambiamos, por ello, ligeramente la notación denotando tales coeficientes por B_j y su estimador por $\widehat{B}_j$.

Definición 8

Se define el estimador lineal de regresión generalizado $\widehat{Y}_U^{\text{GREG}}$ como

$$\widehat{Y}_U^{\text{GREG}} = \sum_{k\in U} \widehat{y}_k + \sum_{k\in s} \frac{e_{ks}}{\pi_k}, \tag{4.10}$$

donde

- $\widehat{y} = \mathbf{x}_k^t \widehat{B}$ es la predicción del valor y_k de acuerdo con el modelo de regresión lineal con coeficientes estimados $\widehat{\mathbf{B}} = (\widehat{B}_1, \ldots, \widehat{B}_J)^t$.
- $e_{ks} = y_k - \widehat{y}_k$ son los residuos del modelo de regresión lineal.
- $\widehat{B}_1, \ldots, \widehat{B}_J$ son las componentes del vector

$$\widehat{\mathbf{B}} = (\widehat{B}_1, \ldots, \widehat{B}_J)^t = \left(\sum_{k\in s} \frac{\mathbf{x}_k \mathbf{x}_k^t}{\sigma_k^2 \pi_k} \right)^{-1} \sum_{k\in s} \frac{\mathbf{x}_k y_k}{\sigma_k^2 \pi_k}.$$

Seguidamente justificaremos esta forma del estimador y el argumento subyacente para construirlo. Obsérvese que la expresión (4.10) del estimador GREG es muy similar a la expresión (4.4) del estimador en diferencias. Ambos estimadores pueden verse bajo el mismo enfoque, solo que ahora los valores aproximados y_k^0 se sustituyen por valores estimados de acuerdo con el modelo de regresión lineal.

La hipótesis básica que se hace para construir el estimador GREG y esperar que sea más eficiente que el estimador HT es la estructura de la población finita en términos de las variables y, $x_1, \ldots, x_J$. Suponemos que los valores y_k parecen haberse generado por un modelo de regresión lineal ξ con y como variable dependiente y x_j como regresores. Más específicamente:

i. Los valores $y_1, \ldots, y_N$ se suponen realizaciones de las variables aleatorias independientes $Y_1, \ldots, Y_N$. A menudo, esto se llama un modelo de *superpoblación*[3];

ii. $\mathbb{E}_\xi [Y_k] = \sum_{j=1}^J \beta_j x_{jk}$ para $k \in U$;

iii. $\mathbb{V}_\xi [Y_k] = \sigma_k^2$ para $k \in U$;

donde $\mathbb{E}_\xi$ y $\mathbb{V}_\xi$ son la esperanza matemática y la varianza con respecto al modelo ξ y β_j, σ_j^2 son parámetros del modelo. En lo que sigue no distinguiremos entre Y_k e y_k siguiendo la costumbre al uso.

[3] *Superpopulation model.*

Es importante señalar que el modelo ξ introduce otro tipo de aleatoriedad diferente a la introducida por la selección aleatoria de la muestra s. Este nueva aleatoriedad se introduce por **hipótesis** y no tiene nada que ver con el mecanismo aleatorio de selección de la muestra a través del diseño muestral.

Adviértase que el modelo es posiblemente heterocedástico, por tanto, pudiendo expresar algún patrón en la varianza. Por ejemplo, podemos expresar que la varianza de y crece con los valores de los regresores x_j.

Ejemplo 9. Dos ejemplos de modelos de una única variable (regresión simple) son el modelo de razón y el modelo de regresión simple:

$$\begin{cases} \mathbb{E}_\xi [y_k] = \beta x_k, \\ \mathbb{V}_\xi [y_k] = \sigma^2 x_k, \end{cases} \tag{4.11a}$$

donde se asume que $x_k > 0$ para todo $k \in U$; y

$$\begin{cases} \mathbb{E}_\xi [y_k] = \beta_1 + \beta_2 x_k, \\ \mathbb{V}_\xi [y_k] = \sigma^2 \end{cases} \tag{4.11b}$$

En ambos modelos se asume, aunque no se diga explícitamente, que $y_1, \ldots, y_N$ son independientes. ■

Comentario 20. Estimación asistida por modelos y dependiente de modelos.
Aunque posteriormente incluiremos comentarios más detallados sobre el rol del modelo de regresión en la construcción de estos estimadores, reseñamos ahora brevemente que el modelo sirve, sobre todo, para describir la relación entre las variables y y x_j.

Se realiza la hipótesis de que los valores poblacionales y parecen realizaciones del modelo, pero en realidad no se asume esta hipótesis para la inferencia, sino solo para la computación de los coeficientes de regresión.

Las propiedades básicas del estimador GREG (insesgadez asintótica, valides de las fórmulas para la varianza, etc.) no dependen de las hipótesis del modelo. Solo la eficiencia del estimador dependerá de la bondad de ajuste. Este tipo de estimación en poblaciones finitas, aún basada en el principio de aleatorización, se denomina *estimación asistida por modelos* en contraposición a la *estimación dependiente de modelos*, donde la validez del modelo es crucial para la calidad de las estimaciones.

Justifiquemos ahora la forma de los coeficientes de regresión estimados $\widehat{B}_j$. Para ello, consideremos una hipotética enumeración de todos los elementos de la población $k \in U$ (esto es, un censo). En este caso, el estimador mínimo-cuadrático ponderado de los

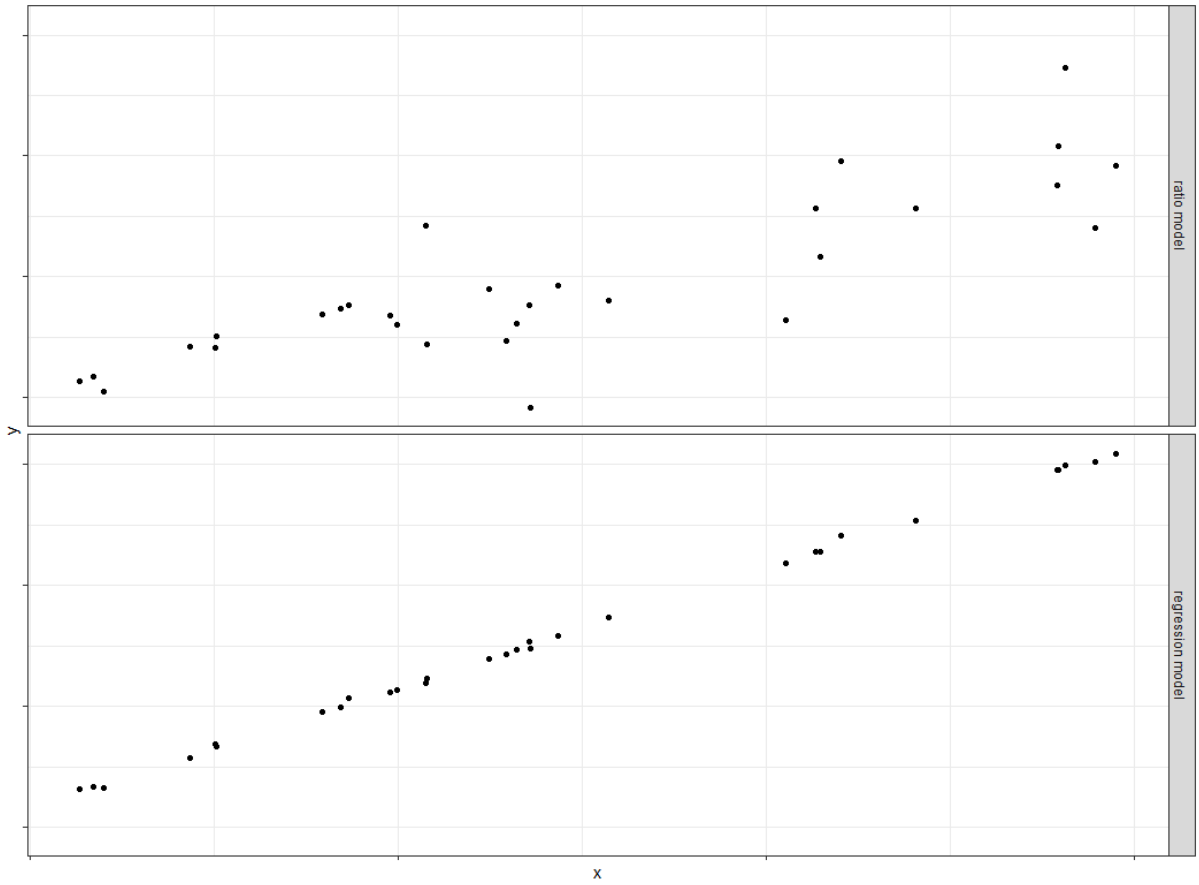

Figura 4.1: Modelos de razón y regresión (4.11a) y (4.11b).

coeficientes de regresión $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^t$ bajo el modelo ξ vendrían dados por

$$\mathbf{B} = (B_1, \ldots, B_J)^t = \left(\sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k^t}{\sigma_k^2} \right)^{-1} \sum_{k \in s} \frac{\mathbf{x}_k y_k}{\sigma_k^2} \tag{4.12a}$$

$$= \left(\mathbf{X} \boldsymbol{\Sigma} \mathbf{X}^t \right)^{-1} \mathbf{X} \boldsymbol{\Sigma}^{-1} \mathbf{Y}, \tag{4.12b}$$

donde

- $\mathbf{X} = [\mathbf{x}_1 \ \ldots \ \mathbf{x}_N] = \begin{pmatrix} x_{11} & \ldots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{J1} & \ldots ; & x_{JN} \end{pmatrix}$ es la matriz de regresores del modelo ξ.
- $\mathbf{Y} = (y_1, \ldots, y_N)^t$ es el vector de la variable independiente.
- $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_N^2 \end{pmatrix}$ es la matriz de covarianzas.

Decimos que $\mathbf{B}$ corresponde a un ajuste hipotético del modelo ξ a toda la población $(\mathbf{x}_k, y_k)$ para todo $k \in U$. Ahora bien, $\mathbf{B}$, en nuestro contexto de estimación en poblaciones finitas, es en realidad un parámetro poblacional de la población finita U y, por tanto, puede ser estimado a partir de los datos de la muestra s (como las medias poblaciones, los cuantiles poblaciones, etc.). Para proceder a su estimación, siguiendo las técnicas de estimación de problemas complejos (véase el tema 3), podemos escribir

$$\mathbf{B} = \mathbf{T}^{-1}\mathbf{t} = f(\mathbf{T}, \mathbf{t}), \tag{4.13}$$

donde

- $\mathbf{T} = \mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^t = \sum_{k\in U} \frac{\mathbf{x}_k\mathbf{x}_k^t}{\sigma_k^2}$;
- $\mathbf{t} = \mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{Y} = \sum_{k\in U} \frac{\mathbf{x}_k y_k}{\sigma_k^2}$.

Un estimador para $\mathbf{B}$ vendrá dado por $\widehat{\mathbf{B}} = f(\widehat{\mathbf{T}}_U^{\text{HT}}, \widehat{\mathbf{t}}_U^{\text{HT}})$, por tanto

$$\begin{aligned}
\widehat{\mathbf{B}} &= (\widehat{B}_1 \ldots, \widehat{B}_J)^t \\
&= \left(\widehat{\mathbf{T}}_U^{\text{HT}}\right)^{-1} \widehat{\mathbf{t}}_U^{\text{HT}} && \text{(4.14a)} \\
&= \left(\sum_{k\in s} \frac{\mathbf{x}_k\mathbf{x}_k^t}{\sigma_k^2\pi_k}\right)^{-1} \sum_{k\in s} \frac{\mathbf{x}_k y_k}{\sigma_k^2\pi_k}, && \text{(4.14b)}
\end{aligned}$$

que es la expresión utilizada en la Definición 8 del GREG. Por tanto, $\widehat{\mathbf{B}}$ estima en una población finita U el parámetro poblacional $\mathbf{B}$, que, a su vez, estima los coeficientes $\boldsymbol{\beta}$ del modelo lineal de regresión subyacente ξ de la superpoblación.

Ejemplo 10. Una vez especificado un modelo, no es difícil construir un estimador GREG. Tomemos como primer ejemplo el modelo de razón (4.11a). En este caso, se obtiene

$$\widehat{B} = \frac{\sum_{k\in s} \frac{y_k}{\pi_k}}{\sum_{k\in s} \frac{x_k}{\pi_k}}$$

y, por tanto,

$$\widehat{Y}_U^{\text{GREG}} = \widehat{Y}_U^{\text{Rat}} = X_U \frac{\widehat{Y}_U^{\text{HT}}}{\widehat{X}_U^{\text{HT}}}.$$

Esto es, se reduce al estimador de razón ya visto en el tema 3 por medios más directos.

Como segundo ejemplo, consideremos el modelo de regresión simple (4.11b) con $\mathbf{x}_k = (1, x_k)^t$ y $\mathbf{B} = (B_1, B_2)^t$. En este caso se obtiene

$$\widehat{\mathbf{B}} = \begin{pmatrix} \widehat{B}_1 \\ \widehat{B}_2 \end{pmatrix} = \begin{pmatrix} \widehat{\bar{y}}_U^{\text{Háj}} - \widehat{B}_2 \widehat{\bar{x}}_U^{\text{Háj}} \\ \frac{\sum_{k\in s} \frac{(x_k - \widehat{\bar{x}}_U^{\text{Háj}})(y_k - \widehat{\bar{y}}_U^{\text{Háj}})}{\pi_k}}{\sum_{k\in s} \frac{(x_k - \widehat{\bar{x}}_U^{\text{Háj}})^2}{\pi_k}} \end{pmatrix},$$

donde $\widehat{\bar{y}}_U^{\text{Háj}} = \frac{\widehat{Y}_N^{\text{HT}}}{\widehat{N}_U^{\text{HT}}}$ y $\widehat{\bar{x}}_U^{\text{Háj}} = \frac{\widehat{X}_N^{\text{HT}}}{\widehat{N}_U^{\text{HT}}}$ denotan los estimadores de Hájek de las medias poblaciones de y y x, respectivamente. El estimador GREG bajo el modelo de regresión simple está dado, por tanto, por

$$\widehat{Y}_U^{\text{GREG}} = N\left[\widehat{\bar{y}}_U^{\text{Háj}} + \widehat{B}_2\left(\bar{x}_U - \widehat{\bar{x}}_U^{\text{Háj}}\right)\right].$$

Adviértase cómo estas derivaciones son válidas para cualquier diseño muestral $p(\cdot)$ que produce las probabilidades de inclusión de primer orden π_k. ■

Comentario 21. Deben hacerse varias observaciones en relación con esta construcción del estimador GREG:

i. Se supone que las matrices $\sum_{k\in U} \frac{\mathbf{x}_k\mathbf{x}_k^t}{\sigma_k^2}$ y $\sum_{k\in s} \frac{\mathbf{x}_k\mathbf{x}_k^t}{\sigma_k^2\pi_k}$ son no singulares, de modo que sus inversas existen.

ii. Adviértase que $\widehat{\mathbf{B}}$ no es un estimador exactamente insesgado para $\mathbf{B}$ puesto que la inversión de matrices no es una operación lineal.

iii. $\widehat{\mathbf{B}}$ debe poder estimarse a partir de su expresión (4.14b), por lo que todas las cantidades en esta fórmula deben conocerse. En particular, las varianzas σ_k^2 deben satisfacer ciertos requisitos. Podemos requerir que sean todas conocidas o que $\sigma_k^2 = v_k\sigma^2$, con v_k conocidos y σ^2 desconocido (p.ej. x_k, como en el GREG con modelo de razón). A menudo, las varianzas se especifican como funciones de parámetros desconocidos que luego se anulan entre sí en las expresiones finales.

iv. En la hipótesis de que todas las varianzas σ_k^2 sean conocidas, la matriz $\sum_{k\in U} \frac{\mathbf{x}_k\mathbf{x}_k^2}{\sigma_k^2}$ podría calcularse de modo exacto, pero es preferible seguir usando la estimación $\sum_{k\in s} \frac{\mathbf{x}_k\mathbf{x}_k^t}{\sigma_k^2\pi_k}$.

■

Comentario 22. El argumento anterior de construcción de un estimador GREG se ha mostrado para un diseño muestral $p(\cdot)$ sin reemplazamiento. Puede repetirse el mismo argumento empleando estimadores de Hansen-Hurvitz si el muestreo es con reemplazamiento en lugar de sin reemplazamiento. ■

Adviértase que si la relación lineal $y_k = \sum_{j=1}^{J} B - jx_{jk}$ fuese perfecta para todo $k \in U$, entonces el segundo término del estimador GREG $\sum_{k\in s} \frac{e_{ks}}{\pi_k}$ sería nulo y el estimador GREG se reduciría a

$$\widehat{Y}_U^{\text{GREG}} = \sum_{k\in U} \widehat{y}_k.$$

La relación lineal perfecta no es la única circunstancia en que se obtiene esta simplificación del estimador GREG, como se demuestra en el siguiente teorema.

Teorema 10

Si existe un vector $\boldsymbol{\lambda} \in \mathbb{R}^J$ tal que, para todo $k \in U$, $\sigma_k^2 = \boldsymbol{\lambda}^t \mathbf{x}_k$, entonces

$$\sum_{k \in s} \frac{e_{ks}}{\pi_k} = 0$$

para todas las muestras s que tienen probabilidades de inclusión de primer orden π_k.

Demostración 10

Utilizando las definiciones:

$$\begin{aligned}
\sum_{k \in s} \frac{\mathbf{x}_k^t}{\pi_k} \widehat{\mathbf{B}} &= \left(\sum_{k \in s} \frac{\boldsymbol{\lambda}^t \mathbf{x}_k \mathbf{x}_k^t}{\sigma_k^2 \pi_k} \right) \widehat{\mathbf{B}} \\
&= \boldsymbol{\lambda}^t \left(\sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k^t}{\sigma_k^2 \pi_k} \right) \widehat{\mathbf{B}} \\
&= \boldsymbol{\lambda}^t \left(\sum_{k \in s} \frac{\mathbf{x}_k y_k}{\sigma_k^2 \pi_k} \right) \\
&= \sum_{k \in s} \frac{\boldsymbol{\lambda}^t \mathbf{x}_k y_k}{\sigma_k^2 \pi_k} \\
&= \sum_{k \in s} \frac{y_k}{\pi_k}.
\end{aligned}$$

Por tanto, $\sum_{k \in s} \frac{e_{ks}}{\pi_k} = \sum_{k \in s} \frac{y_k}{\pi_k} - \left(\sum_{k \in s} \frac{\mathbf{x}_k^t}{\pi_k} \right) \widehat{\mathbf{B}} = \sum_{k \in s} \frac{y_k}{\pi_k} - \sum_{k \in s} \frac{y_k}{\pi_k} = 0.$ ■

Ejemplo 11. Ejemplos sencillos de estructuras de varianza que satisfacen la condición anterior son:

i. $\sigma_k^2 = \sigma^2$ y $x_{1k} = 1$ para todo $k \in U$.

ii. $\sigma_k^2 \propto x_{jk}$ para algún $j \in \{1, \ldots, J\}$ y todo $k \in U$.

iii. $\sigma_k^2 \propto \sum_{j=1}^J a_j x_{jk}$ para todo $k \in U$ y constantes a_j.

Muchas aplicaciones prácticas del estimador GREG cumplen alguna de estas condiciones.

4.5 Expresiones alternativas para el estimador lineal de regresión generalizado.

Al igual que sucedía con el estimador en diferencias, el estimador GREG (8) puede reescribirse de formas alternativas. La Definición (8) pone de manifiesto el mismo argumento original usado para el estimador en diferencias, esto es, para estimar un total poblacional asignamos un valor aproximado (en este caso estimado mediante un modelo de regresión lineal) a cada elemento de la población y utilizamos la muestra para estimar el total del error cometido.

Damos ahora una segunda forma alternativa para el estimador GREG. Teniendo en cuenta que $\widehat{y}_k = \mathbf{x}_k^t \widehat{\mathbf{B}} = \sum_{j=1}^J \widehat{B}_j x_{jk}$, podemos escribir

$$\widehat{Y}_U^{\text{GREG}} = \sum_{k\in U}\sum_{j=1}^J \widehat{B}_j x_{jk} + \sum_{k\in s}\left(\frac{y_k}{\pi_k} - \frac{1}{\pi_k}\sum_{j=1}^J \widehat{B}_j x_{jk}\right) \tag{4.15}$$

$$= \widehat{Y}_U^{\text{HT}} + \sum_{j=1}^J \widehat{B}_j \left(X_{jU} - \widehat{X}_{jU}^{\text{HT}}\right). \tag{4.16}$$

Esta expresión es análoga a la expresión (4.9) para el estimador en diferencias. Pone de manifiesto que el estimador GREG es una mejora sobre el estimador HT mediante un término de ajuste, correlacionado negativamente con el estimador HT.

Además, esta forma muestra que para calcular el estimador GREG solo se necesitan los totales poblacionales X_{jU} de las variables x_j y los valores individuales $\mathbf{x}_k$ *para los elementos de la muestra* $k \in s$. No se necesitan estos valores individuales para toda la población. Si $\mathbf{x}_k$ se recoge junto y_k en la encuesta y se obtienen los totales X_{jU} de una fuente alternativa fiable, el estimador GREG puede seguir utilizándose.

No obstante, adviértase que si los totales se toman de una fuente sin suficiente calidad (por ejemplo, porque esté desactualizada o las definiciones de las variables auxiliares x_j no son idénticas), se corre un riesgo de introducir sesgos severos en el estimador reduciendo notablemente la calidad de las estimaciones.

Damos a continuación otras dos formas alternativas para el estimador GREG obtenidas mediante el uso de álgebra matricial. Sean $\mathbf{X}_U = (X_{1U}, \ldots, X_{JU})^t$ y $\widehat{\mathbf{X}}_U^{\text{HT}} = (\widehat{X}_{1U}^{\text{HT}}, \ldots, \widehat{X}_{JU}^{\text{HT}})^t$. Podemos entonces escribir

$$\widehat{Y}_U^{\text{GREG}} = \widehat{Y}_U^{\text{HT}} + \left(\mathbf{X}_U - \widehat{\mathbf{X}}_U^{\text{HT}}\right)^t \widehat{\mathbf{B}}.$$

Sustituyendo $\widehat{\mathbf{B}}$ usando (4.14a), tenemos

$$\widehat{Y}_U^{\text{GREG}} = \widehat{Y}_U^{\text{HT}} + \left(\mathbf{X}_U - \widehat{\mathbf{X}}_U^{\text{HT}}\right)^t \widehat{\mathbf{B}} = \sum_{k\in s}\frac{y_k}{\pi_k} + \left(\mathbf{X}_U - \widehat{\mathbf{X}}_U^{\text{HT}}\right)^t \widehat{\mathbf{T}}^{-1}\sum_{k\in s}\frac{\mathbf{x}_k y_k}{\sigma_k^2\pi_k} \quad (4.17)$$

$$= \sum_{k\in s}\left[1 + \left(\mathbf{X}_U - \widehat{\mathbf{X}}_U^{\text{HT}}\right)^t \widehat{\mathbf{T}}^{-1}\frac{\mathbf{x}_k}{\sigma_k^2}\right]\frac{y_k}{\pi_k} \quad (4.18)$$

$$\equiv \sum_{k\in s} g_{ks}\frac{y_k}{\pi_k}, \quad (4.19)$$

donde $g_{ks} \equiv 1 + \left(\mathbf{X}_U - \widehat{\mathbf{X}}_U^{\text{HT}}\right)^t \widehat{\mathbf{T}}^{-1}\frac{\mathbf{x}_k}{\sigma_k^2}$ son los llamados pesos generalizados. Adviértase que estos pesos dependen de la muestra. Esta expresión (4.19) pone de manifiesto la linealidad del estimador GREG y muestra las diferencias con el estimador de Horvitz-Thompson introduciendo los pesos g_{ks}.

Ejemplo 12. Retomando los modelos de razón (4.11a) y de regresión simple (4.11b) pueden calcularse los pesos generalizados g_{ks} obteniendo, respectivamente:

$$g_{ks} = \frac{N}{\widehat{N}_U^{\text{HT}}}\left[1 + \frac{\bar{x}_U - \widehat{x}_U^{\text{Háj}}}{\widehat{x}_U^{\text{Háj}}}\right], \quad (4.20a)$$

$$g_{ks} = \frac{N}{\widehat{N}_U^{\text{HT}}}\left[1 + \frac{\bar{x}_U - \widehat{x}_U^{\text{Háj}}}{\tilde{S}_{xx}^2}\left(x_k - \widehat{x}_U^{\text{Háj}}\right)\right], \quad (4.20b)$$

donde

$$\tilde{S}_{xx}^2 = \frac{1}{\widehat{N}_U^{\text{HT}}}\sum_{k\in s}\frac{(x_k - \widehat{x}_U^{\text{Háj}})^2}{\pi_k}.$$

■

La última expresión alternativa es una variante de la propia Definición (8) en la línea de la argumentación tomada prestada del estimador en diferencias. De acuerdo con el modelo de superpoblación, los valores predichos de la variable objetivo y en la población según el modelo ξ cumplen $y_k^0 = \mathbf{x}_k^t\mathbf{B}$ y, por tanto, los residuos poblacionales vienen dados por $E_k = y_k - y_k^0$. Sustituyendo en (4.19), obtenemos

$$\widehat{Y}_U^{\text{GREG}} = \sum_{k\in s}\frac{g_{ks}y_k^0 + g_{ks}E_k}{\pi_k}.$$

Ahora bien,

$$\sum_{k\in s}\frac{g_{ks}y_k^0}{\pi_k} = \sum_{k\in s}\left[1 + \left(\mathbf{X}_U - \widehat{\mathbf{X}}_U^{\text{HT}}\right)^t \widehat{\mathbf{T}}^{-1}\frac{\mathbf{x}_k}{\sigma_k^2}\right]\frac{\mathbf{x}_k^t}{\pi_k}\mathbf{B}$$

$$
\begin{aligned}
&= \left[\left(\widehat{\mathbf{X}}_U^{\mathrm{HT}}\right)^t + \left(\mathbf{X}_U - \widehat{\mathbf{X}}_U^{\mathrm{HT}}\right)^t \widehat{\mathbf{T}}^{-1} \sum_{k\in s} \frac{\mathbf{x}_k}{\sigma_k^2}\frac{\mathbf{x}_k^t}{\pi_k}\right]\mathbf{B} \\
&= \left[\left(\widehat{\mathbf{X}}_U^{\mathrm{HT}}\right)^t + \mathbf{X}_U^t - \left(\widehat{\mathbf{X}}_U^{\mathrm{HT}}\right)^t\right]\mathbf{B} \\
&= \sum_{k\in U} \mathbf{x}_k^t \mathbf{B} \\
&= \sum_{k\in U} y_k^0. \qquad (4.21)
\end{aligned}
$$

Por tanto,

$$\widehat{Y}_U^{\mathrm{GREG}} = \sum_{k\in U} y_k^0 + \sum_{k\in s} g_{ks}\frac{E_{ks}}{\pi_k}. \qquad (4.22)$$

Adviértase que esta expresión no es muy útil computacionalmente, puesto que no conocemos los valores y_k^0, sin embargo, servirá en la sección siguiente para encontrar expresiones para la varianza y su estimación.

Ejemplo 13. La expresión $\widehat{Y}_U^{\mathrm{GREG}} = \sum_{k\in s} g_{ks}\frac{y_k}{\pi_k}$ nos permite mostrar de manera inmediata que el estimador GREG está calibrado para las cantidades auxiliares $\mathbf{X}_U$:

$$\sum_{k\in s} g_{ks}\frac{\mathbf{x}_k}{\pi_k} = \mathbf{X}_U,$$

como está implícito en (4.21). Esto conecta con un enfoque más global de construir estimadores empleando información auxiliar mediante la resolución de determinados problemas de optimización (véase p.ej. Särndal 2007). De hecho, el estimador GREG no es sino la solución a un caso particular de estos problemas de optimización. La calibración de los pesos de muestreo para incorporar información auxiliar y reducir la varianza de los estimadores es una práctica común hoy día en las oficinas de estadística. ■

4.6 Varianza y sus estimaciones

Como sucede con todos los estimadores, deben evaluarse su sesgo y varianza para controlar la calidad de las estimaciones. Por tanto, dadas las diversas expresiones alternativas del estimador GREG anteriores, ahora debemos encontrar expresiones para su sesgo y su varianza, así como la estimación de esta.

Para ello recurriremos a las técnicas de linealización incluidas en el tema 3 de este bloque. Veremos, por tanto, que el estimador GREG no es exactamente insesgado.

Teorema 11

El estimador GREG $\widehat{Y}_U^{\text{GREG}}$, dado equivalentemente por

$$\begin{aligned}
\widehat{Y}_U^{\text{GREG}} &= \sum_{k\in U}\widehat{y}_k + \sum_{k\in s}\frac{y_k-\widehat{y}_k}{\pi_k} && (4.23a)\\
&= \widehat{Y}_U^{\text{HT}} + \left(\mathbf{X}_U - \widehat{\mathbf{X}}_U^{\text{HT}}\right)^t \widehat{\mathbf{B}} && (4.23b)\\
&= \sum_{k\in s} g_{ks}\frac{y_k}{\pi_k} && (4.23c)\\
&= \sum_{k\in U} y_k^0 + \sum_{k\in s} g_{ks}\frac{y_k-y_k^0}{\pi_k}, && (4.23d)
\end{aligned}$$

donde

- $\mathbf{B} = \sum_{k\in U}\frac{\mathbf{x}_k\mathbf{x}_k^t}{\sigma_k^2}$;
- $\widehat{\mathbf{B}} = \sum_{k\in s}\frac{\mathbf{x}_k\mathbf{x}_k^t}{\sigma_k^2\pi_k}$;
- $\widehat{y}_k = \mathbf{x}_k^t\widehat{\mathbf{B}}$;
- $y_k^0 = \mathbf{x}_k^t\mathbf{B}$;
- $g_{ks} = 1 + \left(\mathbf{X}_U - \widehat{\mathbf{X}}_U^{\text{HT}}\right)^t\left(\sum_{k\in s}\frac{\mathbf{x}_k\mathbf{x}_k^t}{\sigma_k^2}\pi_k\right)^{-1}\frac{\mathbf{x}_k}{\sigma_k^2}$;

puede aproximarse a primer orden mediante la linealización de Taylor mediante

$$\widehat{Y}_U^{\text{GREG0}} = \sum_{k\in U} y_k^0 + \sum_{k\in s}\frac{y_k-y_k^0}{\pi_k}. \tag{4.24}$$

El estimador linealizado $\widehat{Y}_U^{\text{GREG0}}$ es aproximadamente insesgado para el total poblacional Y_U y tiene varianza dada por

$$\mathbb{V}\left[\widehat{Y}_U^{\text{GREG}}\right] \approx \mathbb{V}\left[\widehat{Y}_U^{\text{GREG0}}\right] = \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_k-y_k^0}{\pi_k}\frac{y_l-y_l^0}{\pi_l}, \tag{4.25}$$

que puede estimarse mediante

$$\widehat{\mathbb{V}}\left[\widehat{Y}_U^{\text{GREG}}\right] \approx \sum_{k\in s}\sum_{l\in s}\check{\Delta}_{kl}\left[g_{ks}\frac{y_k-\widehat{y}_k}{\pi_k}\right]\left[g_{ls}\frac{y_l-\widehat{y}_l}{\pi_l}\right]. \tag{4.26}$$

Demostración 11

Para aplicar la técnica de linealización de Taylor debemos primero expresar el estimador GREG $\widehat{Y}_U^{\text{GREG}}$ como una función de estimadores de Horvitz-Thompson de totales poblacionales:

$$\begin{aligned}\widehat{Y}_U^{\text{GREG}} &= \widehat{Y}_U^{\text{HT}} + \left(\mathbf{X}_U - \widehat{\mathbf{X}}_U^{\text{HT}}\right)^t \widehat{\mathbf{B}} \\ &= \widehat{Y}_U^{\text{HT}} + \left(\mathbf{X}_U - \widehat{\mathbf{X}}_U^{\text{HT}}\right)^t \widehat{\mathbf{T}}^{-1}\widehat{\mathbf{t}} \\ &= f(\widehat{Y}_U^{\text{HT}}, \widehat{\mathbf{X}}_U^{\text{HT}}, \widehat{\mathbf{T}}, \widehat{\mathbf{t}}),\end{aligned}$$

donde $\widehat{\mathbf{T}}$ y $\widehat{\mathbf{t}}$ están definidos en (4.13). Calculando las derivadas parciales y evaluándolas en $(Y_U, \mathbf{X}_U, \mathbf{T}, \mathbf{t})$ se obtiene a primer orden del desarrollo de Taylor

$$\begin{aligned}\widehat{Y}_U^{\text{GREG}} \approx \widehat{Y}_U^{\text{GREG0}} &= Y_U + (\widehat{Y}_U^{\text{HT}} - Y_U) - \sum_{j=1}^{J} B_j(\widehat{X_{jU}}^{\text{HT}} - X_{jU}) \\ &= \widehat{Y}_U^{\text{HT}} + \left[\mathbf{X} - \widehat{\mathbf{X}}_U^{\text{HT}}\right]^t \mathbf{B} \\ &= \sum_{k\in U} y_k^0 + \sum_{k\in s} \frac{y_k - y_k^0}{\pi_k}. \qquad (4.27)\end{aligned}$$

Entonces,

$$\mathbb{E}\left[\widehat{Y}_U^{\text{GREG}}\right] \approx \mathbb{E}\left[\widehat{Y}_U^{\text{GREG0}}\right] = \sum_{k\in U} y_k^0 + \sum_{k\in U}(y_k - y_k^0) = Y_U.$$

La varianza se obtiene como la de cualquier otro estimador HT:

$$\mathbb{V}\left[\widehat{Y}_U^{\text{GREG0}}\right] = \mathbb{V}\left[\sum_{k\in s}\frac{y_k - y_k^0}{\pi_k}\right] = \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_k - y_k^0}{\pi_k}\frac{y_l - y_l^0}{\pi_l}.$$

Para la estimación de la varianza, sin embargo, no podemos proceder como con los estimadores HT, pues llegaríamos a la expresión

$$\sum_{k\in s}\sum_{l\in s}\check{\Delta}_{kl}\frac{y_k - y_k^0}{\pi_k}\frac{y_l - y_l^0}{\pi_l},$$

en la que no conocemos los valores y_k^0. Una primera opción es estimar y_k^0 mediante $\widehat{y}$ de modo que una primera expresión para la estimación de la varianza sería

$$\widehat{\mathbb{V}}\left[\widehat{y}_U^{\text{GREG0}}\right] = \sum_{k\in s}\sum_{l\in s} \breve{\Delta}_{kl} \frac{y_k - \widehat{y}_k}{\pi_k} \frac{y_l - \widehat{y}_l}{\pi_l}. \tag{4.28}$$

No obstante, si nos fijamos en la expresión exacta (4.23d) (sin aproximación de Taylor), la varianza exacta vendría dada por $\mathbb{V}\left[\sum_{k\in s} g_{ks} \frac{y_k - y_k^0}{\pi_k}\right]$, a la que no podemos aplicar los resultados usuales de los estimadores HT porque g_{ks} depende de la muestra. Sin embargo, si solo retenemos el primer término $g_{ks} \approx 1$ y estimamos $y_k^0 \approx \widehat{y}_k$, obtenemos $\mathbb{V}\left[\sum_{k\in s}\sum_{l\in s} \Delta_{kl} \frac{y_k \widehat{y}_k}{\pi_k} \frac{y_l - \widehat{y}_l}{\pi_l}\right]$ pudiendo ahora estimar este expresión como un estimador HT:

$$\widehat{\mathbb{V}}\left[\widehat{Y}_U^{\text{GREG0}}\right] \approx \sum_{k\in s}\sum_{l\in s} \breve{\Delta}_{kl} \frac{y_k - \widehat{y}_k}{\pi_k} \frac{y_l - \widehat{y}_l}{\pi_l}.$$

Con este resultado podemos construir un intervalo de confianza de nivel de confianza aproximado $100(1-\alpha)\,\%$ dado por

$$\widehat{Y}_U^{\text{GREG}} \pm z_{1-\frac{\alpha}{2}} \left[\widehat{\mathbb{V}}\left[\widehat{Y}_U^{\text{GREG}}\right]\right]^{1/2},$$

donde $z_{1-\frac{\alpha}{2}}$ es el cuantil de orden $1-\frac{\alpha}{2}$ de la distribución normal estandarizada.

Ejemplo 14. Para el modelo de razón (4.11a) la estimación de la varianza del estimador de razón viene dada por

$$\widehat{\mathbb{V}}\left[\widehat{Y}_U^{\text{GREG}}\right] = \widehat{\mathbb{V}}\left[\widehat{Y}_U^{\text{Rat}}\right] \approx \left(\frac{N\bar{x}_U}{\widehat{N}_U^{\text{HT}}\widehat{\bar{x}}_U^{\text{Háj}}}\right)^2 \sum_{k\in s}\sum_{l\in s} \breve{\Delta}_{kl} \frac{y_k \widehat{y}_k}{\pi_k} \frac{y_k \widehat{y}_k}{\pi_k}.$$

Adviértase que esta expresión es válida para cualquier diseño muestral $p(\cdot)$. Para el caso del muestreo aleatorio simple sin reemplazamiento, esta expresión es

$$\widehat{\mathbb{V}}\left[\widehat{Y}_U^{\text{GREG}}\right] = \widehat{\mathbb{V}}\left[\widehat{Y}_U^{\text{Rat}}\right] \approx N^2 \left(\frac{\bar{x}_U}{\bar{x}_s}\right) \frac{1-f}{n} \frac{1}{n-1} \sum_{k\in s} (y_k - \widehat{B}x_k)^2,$$

donde $\widehat{B} = \frac{\bar{y}_s}{\bar{x}_s}$ y f es la fracción de muestreo. ■

Comentario 23. En la demostración del teorema anterior aparecen dos expresiones alternativas ligeramente diferentes para la estimación de la varianza. En un número de casos existe evidencia (Särndal, Swensson y Wretman 1992) de que es preferible la expresión del enunciado. Debe advertirse, no obstante, que existen incluso más alternativas que buscan la insesgadez con respecto al modelo ξ y son consistentes respecto al diseño muestral $p(\cdot)$.

4.7 El papel del modelo

El papel del modelo puede examinarse desde dos puntos de vista, (i) en relación con el sesgo, y (ii) en relación con la varianza. Hemos visto que el estimador GREG $\widehat{Y}_U^{\text{GREG}}$ se comporta aproximadamente como la variable aleatoria linealizada $\widehat{Y}_U^{\text{GREG0}}$ bajo muestreo repetido de una población con valores fijos en todas las variables. Como $\widehat{Y}_U^{\text{GREG0}}$ es insesgado para Y_U independientemente de la distribución de valores en la población finita, se sigue que el estimador de regresión

$$\widehat{Y}_U^{\text{GREG}}$$

es aproximadamente insesgado para Y_U, independientemente de si las hipótesis del modelo ξ son verdaderas o falsas.

Por otro lado, la idoneidad del modelo ξ es fundamental para conseguir una varianza pequeña. Cuanto más se ajuste la dispersión de la población a un patrón lineal $y_k \doteq x_k^t \mathbf{B}$, más pequeños serán los residuos poblacionales $E_k = y_k - x_k^t \mathbf{B}$ y menor será la varianza del estimador de regresión.

Cabe mencionar que las variables x utilizadas en el estimador de regresión tienen que ser elegidas del conjunto de variables auxiliares disponibles. La cuestión fundamental para la eficiencia del estimador de regresión es si un vector x con un total $\sum_{k\in U} \mathbf{x}_k$ también es un vector explicativo de peso para la variable y.

No es necesario que el modelo sea 'verdadero' en el sentido de que verdaderamente describe la generación de los valores y_k de la población. Si los datos poblacionales están bien descritos por el modelo asumido, el estimador de regresión normalmente implicará una reducción grande de la varianza comparado con el estimador HT. Si la población no está bien descrita por el modelo, la mejora en relación con el estimador HT será pequeña, pero el estimador GREG aun así garantiza insesgadez aproximada. Por estas razones se dice que el estimador GREG está asistido por el modelo, pero no es dependiente del modelo.

La estructura de la varianza del modelo (los parámetros σ_k^2) también es importante, pero es especialmente importante en la planificación y diseño de la encuesta. Los valores σ_k^2 tienen (deben tener) un impacto notable en la elección del diseño muestral. De hecho, es recomendable escoger el diseño muestral de modo que $\pi_k \propto \sigma_k$.

Bibliografía

Särndal, C.-E. (2007). "The calibration approach in survey theory and practice". En: *Survey Methodology* 33, págs. 99-119.

Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.

Tema 5

Muestreo Bifásico. Definición. Elección de estimador. El estimador π^*. Muestreo bifásico para la estratificación. Variables auxiliares para la selección de la muestra en dos fases.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía.

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.

5.1 Muestreo Bifásico. Definición.

La teoría y los métodos de muestreo vistos hasta ahora en los temas del bloque "Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas" del grupo de materias comunes y en los cuatro primeros temas de "Producción Estadística Oficial: Métodos Avanzados"forman un conjunto de herramientas básicas en el muestreo. Pero muchas encuestas necesitan técnicas más avanzadas. En este tema veremos el muestreo bifásico, o muestreo seguido de submuestreo. Esta técnica, propuesta originalmente por (Neyman 1938), se presentará de forma general, con diseños muestrales arbitrarios en cada una de las dos fases. El muestreo bifásico es un muestreo adecuado para las encuestas en las que se sabe poco o nada a priori sobre la población.

Existen métodos de estimación muy eficientes que precisan mucha información auxiliar. En muchas encuestas, sin embargo, no se dispone de este tipo de información. Por ejemplo, el estimador combinado de razón de un total poblacional Y_U dado por

$$\widehat{Y}_U^{\text{cRat}} = \sum_{k \in U} x_k \frac{\sum_{h=1}^{H} n_{s_h} \bar{y}_{s_h}}{\sum_{h=1}^{H} n_{s_h} \bar{x}_{s_h}},$$

es a menudo una mejora notable sobre la estrategia muestral directa compuesta por el muestreo aleatorio simple sin reemplazamiento y el estimador de Horvitz-Thompson. Pero el estimador combinado de razón requiere que todos los elementos puedan ser estratificados y que el total $\sum_{x\in U} x_k$ sea conocido. En algunas encuestas, esta información no está disponible.

Cuando el marco contiene poca o ninguna información sobre los elementos de la población, hay dos opciones posibles:

i. Usar un diseño muestral muy sencillo como muestreo aleatorio simple sin reemplazamiento o muestreo monoetápico combinado con el estimador de Horvitz-Thompson. No obstante, solo puede conseguirse una precisión aceptable en las estimaciones mediante un tamaño muestral muy grande, que implica costes normalmente prohibitivos.

ii. Reunir información sobre la población, usarla para construir un nuevo marco más completo y entonces seleccionar una combinación eficiente de diseño muestral y estimador de regresión. Una muestra más pequeña puede ser suficiente para obtener la precisión deseada. Sin embargo, el coste todavía puede ser grande.

Se puede identificar una tercera opción. Es un compromiso entre las opciones anteriores y cosiste en una selección de la muestra en dos fases, de la siguiente forma:

a. En la primera fase, seleccionar una muestra bastante grande de elementos s_a mediante un diseño muestral sencillo $p_a(\cdot)$. Para los elementos de s_a, recoger información barata de una o más variables auxiliares.

b. Con la ayuda de la información auxiliar recogida en la primera fase, seleccionar una muestra en la segunda fase s a partir de s_a usando el diseño muestral $p(\cdot|s_a)$. Este conjunto s es una submuestra.

Esta última técnica se llama muestreo *bifásico* (o a veces muestreo *doble*). Las extensiones a más de dos fases se llaman muestreo *multifásico*. Un punto muy importante para que el muestreo bifásico sea efectivo es la creación de un marco con mucha información, no para toda la población (esto sería muy caro), sino para una parte de la población de la que se extraerá la submuestra.

Ejemplo 15. Frecuentemente, el muestreo en dos etapas se utiliza en los estudios de silvicultura. Se dispone de fotografías aéreas y sistemáticamente se distribuyen puntos en las fotografías. Se estudian las áreas alrededor de los puntos de las fotografías y se clasifican de acuerdo con el tipo de terreno: bosque, bosque improductivo, área no de bosque y agua. Entonces, se extrae una muestra de la primera etapa de puntos en la retícula con una fracción de muestreo mayor para los puntos de la retícula clasificados como bosque que para los clasificados como no de bosque. Las áreas de la muestra de la primera etapa se examinan con más cuidado para clasificarlas según el tamaño y la densidad de los árboles. Luego, se extrae una submuestra de los puntos de la muestra de la primera etapa y se realizan mediciones como uso del suelo, volumen y mortalidad; el porcentaje de área de bosque de la muestra de la segunda etapa puede diferir un

poco de la estimación fotográfica de la primera etapa y la estimación pro proporción se puede usar en la muestra de la segunda etapa para aumentar su precisión.

Una razón adicional para estudiar el muestreo bifásico es que esta teoría es útil para la estimación en caso de falta de respuesta. En una encuesta con falta de respuesta, la selección de una muestra probabilística se puede ver como la primera fase y los informantes se pueden considerar como una subselección derivada del mecanismo (aleatorio) que origina la respuesta/falta de respuesta. Por tanto, esta teoría también resulta interesante para tratar la falta de respuesta.

El muestreo en ocasiones sucesivas (del cual se verá parte en el tema 6 de este mismo bloque) es una técnica que se basa en el muestreo en dos o más fases. El muestreo de conglomerados con submuestreo es un caso particular del muestreo bifásico. La clase de diseños muestrales en dos etapas considerados en el tema 1 (muestreo bietápico) de este bloque se refiere a los diseños que verifican las condiciones de invarianza e independencia definidas en el tema 1[1]. La teoría para el muestreo bifásico que veremos a continuación proporciona un marco más flexible para el muestreo en dos etapas.

Para ilustrarlo, supongamos que nuestro objetivo es obtener una muestra s en dos fases de forma que s tenga un tamaño fijo a priori n. A partir de las N_I PSUs, de tamaños diversos, se selecciona una muestra aleatoria simple sin reemplazamiento s_I de n_I PSUs. El número de elementos incluidos en el conjunto de PSUs seleccionados es

$$N_{s_i} = \sum_{i \in s_I} N_i.$$

Aquí, N_{s_i} variará de una muestra de primera fase a otra. Para el diseño de la segunda fase, mantenemos la condición de independencia, pero no la de invarianza. En una PSU seleccionado, se selecciona una submuestra aleatoria simple sin reemplazamiento s_i de n_{s_i} elementos, donde n_{s_i} es proporcional al tamaño N_i de la PSU, es decir,

$$n_{s_i} = n\frac{N_i}{N_{s_i}}. \tag{5.1}$$

El tamaño muestral total es n, pero como n_{s_i} depende del resultado de la primera fase, ya no se verifica la propiedad de invarianza. El diseño es autoponderado en el sentido de que la probabilidad de inclusión de un elemento

$$\pi_k = \pi_{Ii}\pi_{k|i} = \frac{n_i n}{N_i N_{s_i}}$$

es constante para todo k. Nótese, sin embargo, que depende del resultado del muestreo de primera fase.

[1]En el muestreo bietápico, la invarianza de los diseños de segunda etapa $p_i(\cdot|s_I)$ quiere decir que, para cada $i \in U_I$ y cada $s_I \subset U_I$, se cumple $p_i(\cdot|s_I) = p_i(\cdot)$. La independencia de los diseños de segunda etapa $p_i(\cdot|s_I)$ quiere decir que, para cada cada $s_I \subset U_I$, se cumple $\mathbb{P}\left(\bigcup_{i\in s_I} s_i|s_I\right) = \prod_{i\in s_i} \mathbb{P}\left(s_i|s_I\right)$.

5.2 Elección de estimador

El subíndice a denotará la primera fase. La muestra de primera fase s_a de tamaño n_{s_a} se selecciona de acuerdo con un diseño muestral $p_a(\cdot)$ de forma que $p_a(s_a)$ es la probabilidad de elegir s_a. Las probabilidades de inclusión correspondientes se denotan por

$$\pi_{ak} = \sum_{s_a \ni k} p_a(s_a) \tag{5.2}$$

y

$$\pi_{akl} = \sum_{s_a \ni k \& l} p_a(s_a) \tag{5.3}$$

para k y $l \in U$.

Dada la muestra en primera fase s_a, la muestra de la segunda fase s, de tamaño n_s, se obtiene de acuerdo con el diseño $p(\cdot|s_a)$ de forma que $p(s|s_a)$ es la probabilidad condicional de elegir s. Las probabilidades de inclusión bajo este diseño se denotan por

$$\pi_{k|s_a} = \sum_{s \ni k} p(s|s_a) \tag{5.4}$$

y

$$\pi_{kl|s_a} = \sum_{s \ni k,l} p(s|s_a) \tag{5.5}$$

para k y $l \in s_a$.

La siguiente tarea es encontrar un estimador insesgado para el total poblacional $Y_U = \sum_{x \in U} y_k$. Para simplificar la notación en este tema, escribiremos Y para el total poblacional de la variable y en lugar de Y_U. Un candidato natural es el estimador de Horvitz-Thompson

$$\widehat{Y}_U^{\text{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k}, \tag{5.6}$$

donde π_k es la probabilidad de inclusión del k-ésimo elemento. Obviamente, el estimador (5.6) precisa el cálculo de las probabilidades π_k, pero esto no siempre es posible en un muestreo bifásico. Tenemos

$$\pi_k = \sum_{s \ni k} p(s),$$

donde $p(s)$ es la probabilidad (conjuntamente en las dos fases) de selección de la muestra s,

$$p(s) = \sum_{s_a \supset s} p_a(s_a) p(s|s_a).$$

Por consiguiente,

$$\begin{aligned}\pi_k &= \sum_{s\ni k}\sum_{s_a\supset s} p_a(s_a)p(s|s_a) = \sum_{s_a\ni k}\sum_{\substack{s\supset s_a\\ s\ni k}} p_a(s_a)p(s|s_a)\\ &= \sum_{s_a\ni k} p_a(s_a)\Big[\sum_{\substack{s\supset s_a\\ s\ni k}} p(s|s_a)\Big] = \sum_{s_a\ni k} p_a(s_a)\pi_{k|s_a}\end{aligned} \tag{5.7}$$

Por tanto, para determinar π_k en la práctica, debemos conocer las probabilidades $p_a(s_a)$ para cada s_a, que normalmente conocemos, y también debemos conocer las $\pi_{k|s_a}$ para cada s_a, que normalmente no conocemos, porque $\pi_{k|s_a}$ puede depender del resultado de la primera fase.

Ejemplo 16. Supongamos que en la primera fase obtenemos una muestra aleatoria simple sin reemplazamiento de tamaño n_a. Entonces

$$p_a(s_a) = \frac{1}{\binom{N}{n_a}}$$

para cada s_a de tamaño fijo n_a (y $p_a(s_a) = 0$ en otro caso), y $\pi_{ak} = \frac{n_a}{N}$ para todo $k \in s_a$. Más aún, supongamos que, para cada $k \in s_a$, grabamos una variable auxiliar x que se considera aproximadamente proporcional a la variable de estudio y.

Supongamos que la segunda fase se lleva a cabo con un muestreo de Poisson, de forma que la probabilidad de inclusión condicional del elemento k es proporcional a x, es decir,

$$\pi_{k|s_a} = n\frac{x_k}{\sum_{k\in s_a} x_k},$$

donde, por simplicidad, se asume que $nx_k < \sum_{x\in s_a} x_k$ para todo k.

Claramente, $p_a(s_a)$ es conocido para cada s_a. Sin embargo, $\pi_{k|s_a}$ depende de s_a y, por tanto, sólo es conocido para la muestra de la primera fase s_a realmente obtenida. En consecuencia, la probabilidad de inclusión π_k necesaria para el estimador (5.6) no se puede obtener. ■

Este ejemplo demuestra que el estimador de Horvitz-Thompson no siempre se puede usar en la práctica. Como consecuencia, buscamos un estimador insesgado que use ponderaciones de una forma más práctica. Para ello, sea $\sum_{k\in s_a} \frac{y_k}{\pi_{ak}}$ el estimador HT de $Y = \sum_{x\in U} y_k$ que podría ser obtenido si y_k y π_{ak} fuesen conocidos para cada $k \in s_a$. Pero y_k se observa únicamente si $k \in s$. Así, dada la muestra s_a, $\sum_{k\in s_a} \frac{y_{ak}}{\pi_{ak}}$ está estimado insesgadamente por el estimador HT condicional

$$\sum_{k\in s} \frac{\frac{y_{ak}}{\pi_{ak}}}{\pi_{k|s_a}} = \sum_{k\in s} \frac{y_k}{\pi_{ak}\pi_{k|s_a}}. \tag{5.8}$$

Este estimador servirá, ya que los $\pi_{k|s_a}$ son conocidos para la muestra s_a obtenido en la primera fase. Si introducimos la cantidad

$$\pi_k^* = \pi_{ak}\pi_{k|s_a} \tag{5.9}$$

y teniendo en cuenta que el peso asociado a y_k en el nuevo estimador (5.8) es $\frac{1}{\pi_k^*}$ podemos decir que (5.8) se obtienen mediante la 'expansión π^*' de los valores y_k en la muestra de la segunda fase. Denotemos el valor de y expandido por π^* mediante

$$\breve{y}_k = \frac{\breve{y}_{ak}}{\pi_{k|s_a}} = \frac{y_k}{\pi_{ak}\pi_{k|s_a}} = \frac{y_k}{\pi_k^*}, \tag{5.10}$$

donde $\breve{}$ nos recuerda que se usa una expansión doble, una para cada fase. Llamaremos a (5.8) el estimador HT* y lo denotaremos por $\widehat{Y}_U^{\mathrm{HT}^*}$; en otras palabras

$$\widehat{Y}_U^{\mathrm{HT}^*} = \sum_{k\in s}\frac{y_k}{\pi_k^*} = \sum_{k\in s}\breve{\breve{y}}_k, \tag{5.11}$$

donde π_k^* viene dado por la ecuación (5.9).

Téngase en cuenta que $\widehat{Y}_U^{\mathrm{HT}^*}$ no coincide en general con el estimador $\widehat{Y}_U^{\mathrm{HT}}$ dado en la ecuación (5.6), ya que $\pi_k \neq \pi_k^*$ (véase el Comentario 24 más abajo), excepto en casos raros. El estimador $\widehat{Y}_U^{\mathrm{HT}^*}$ se ve en la siguiente Sección 5.3, y en la Sección 5.4 se verán ejemplos de su uso en la práctica.

Comentario 24. Nótese que hay sutiles diferencias entre las distintas probabilidades. Por ejemplo, para $k \in s_a$,

$$\mathbb{P}(k \in s | k \in s_a) = \frac{\mathbb{P}(k \in s)}{\mathbb{P}(k \in s_a)} = \frac{\pi_k}{\pi_{ak}} \tag{5.12}$$

que, en general diferirá de $\pi_{k|s_a}$.

Aquí, $\pi_{k|s_a}$ es la probabilidad de incluir k en la muestra de la segunda fase s bajo el diseño particular usado cuando se obtuvo s_a en la primera fase. Por contra, $\frac{\pi_k}{\pi_{ak}}$ es la probabilidad condicional de seleccionar k en la segunda fase, dado que k estaba presente en s_a. Usando la ecuación (5.12), concluimos que π_k y π_k^* no son en general iguales, ya que

$$\begin{aligned}\pi_k &= \mathbb{P}(k \in s) = \mathbb{P}(k \in s_a)\mathbb{P}(k \in s | k \in s_a)\\ &= \pi_{ak}\frac{\pi_k}{\pi_{ak}} \neq \pi_{ak}\pi_{k|s_a} = \pi_k^*.\end{aligned}$$

5.3 El estimador π^* (HT*)

En esta sección veremos las principales características del estimador $\widehat{Y}_U^{\mathrm{HT}^*}$. Para eso usaremos la siguiente notación. Sea

$$\pi_{kl}^* = \pi_{akl}\pi_{kl|s_a} \tag{5.13}$$

$$\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al} \tag{5.14}$$

y

$$\Delta_{kl|s_a} = \pi_{kl|s_a} - \pi_{k|s_a}\pi_{l|s_a} \tag{5.15}$$

Estas expresiones nos ayudarán a escribir el error de la estimación del total del estimador π^* dado por (5.11) como la suma de dos componentes,

$$\widehat{Y}_U^{\text{HT}^*} - Y = \underbrace{(\sum_{k\in s_a}\frac{y_k}{\pi_{ak}} - \sum_{k\in U} y_k)}_{Q_{s_a}} + \underbrace{(\sum_{k\in s}\breve{\breve{y}}_k - \sum_{k\in s_a}\frac{y_k}{\pi_{ak}})}_{R_s}, \tag{5.16}$$

donde Q_{s_a} se puede denominar el error debido a la primera fase de muestreo y R_s el error debido a la segunda fase.

Recordamos que el estimador $\widehat{Y}_U^{\text{HT}^*}$ es insesgado dado s_a, para la estimación $\sum_{k\in s_a}\frac{y_k}{\pi_{ak}}$ que se formaría si solo hubiese una fase en el muestreo. Esto implica que la componente R_s en (5.16) tiene un valor esperado nulo, condicionalmente en s_a. Esta propiedad es muy importante.

Ahora se puede obtener fácilmente el siguiente resultado.

Teorema 12

En el muestreo bifásico, el total poblacional $Y_U = \sum_{k\in U} y_k$ se estima de forma insesgada por el estimador HT*

$$\widehat{Y}_U^{\text{HT}^*} = \sum_{k\in s}\frac{y_k}{\pi_k^*}. \tag{5.17}$$

La varianza de $\widehat{Y}_U^{\text{HT}^*}$ viene dada por

$$\mathbb{V}(\widehat{Y}_U^{\text{HT}^*}) = \sum_{k\in U}\sum_{l\in U}\Delta_{akl}\frac{y_k}{\pi_{ak}}\frac{y_l}{\pi_{al}} + \mathbb{E}_{p_a}\left(\sum_{k\in s_a}\sum_{l\in s_a}\Delta_{kl|s_a}\breve{\breve{y}}_k\breve{\breve{y}}_l\right), \tag{5.18}$$

donde $\breve{\breve{y}}_k = \frac{y_k}{\pi_k^*}$ y las cantidades Δ vienen dadas por las ecuaciones (5.14) y (5.15). Un estimador insesgado de la varianza viene dado por

$$\widehat{\mathbb{V}}(\widehat{Y}_U^{\text{HT}^*}) = \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{akl}}{\pi_{kl}^*}\frac{y_k}{\pi_{ak}}\frac{y_l}{\pi_{al}} + \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}}\breve{\breve{y}}_k\breve{\breve{y}}_l \tag{5.19}$$

Cada componente de (5.19) es insesgado para su equivalente en (5.18).

Demostración 12. La demostración del Teorema 12 es sencilla si recurrimos al mismo resultado general de la teoría de probabilidad (véase p.ej. Grimmet y Stirzaker 2004,

pág. 69) que en el tema 1 de este mismo bloque[2] . En este caso, es adecuado condicionar sobre la muestra s_a obtenida en la primera fase. Para el valor de la esperanza de $\widehat{Y}_U^{\mathrm{HT}^*}$, tenemos

$$\mathbb{E}(\widehat{Y}_U^{\mathrm{HT}^*}) = \mathbb{E}_{p_a}\mathbb{E}(\widehat{Y}_U^{\mathrm{HT}^*}|s_a) = \mathbb{E}_{p_a}\left(\sum_{k\in s_a}\frac{y_{ak}}{\pi_{ak}}\right) = Y.$$

Es decir, el estimador HT* es insesgado. Volviendo a la varianza, tenemos

$$\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}^*}) = \mathbb{V}_{p_a}\mathbb{E}(\widehat{Y}_U^{\mathrm{HT}^*}|s_a) + \mathbb{E}_{p_a}\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}^*}|s_a),$$

que refleja la varianza debida a cada una de las dos fases de muestreo. Pero a partir de la descomposición (5.16)

$$\mathbb{V}_{p_a}\mathbb{E}(\widehat{Y}_U^{\mathrm{HT}^*}|s_a) = \mathbb{V}_{p_a}(Q_{s_a}) = \sum_{k\in U}\sum_{l\in U}\Delta_{akl}\frac{y_k}{\pi_{ak}}\frac{y_l}{\pi_{al}}$$

y

$$\mathbb{E}_{p_a}\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}^*}|s_a) = \mathbb{E}_{p_a}\mathbb{V}(R_s|s_a) = \mathbb{E}_{p_a}\left(\sum_{k\in s_a}\sum_{l\in s_a}\Delta_{kl|s_a}\breve{\breve{y}}_k\breve{\breve{y}}_l\right),$$

donde hemos usado propiedades conocidas del estimador HT.

La varianza sin condicionar (5.18) se sigue inmediatamente. Para el estimador de la varianza usamos $\mathbb{E}(\cdot) = \mathbb{E}_{p_a}\mathbb{E}(\cdot|s_a)$, y obtenemos la primera componente,

$$\begin{aligned}\mathbb{E}\left(\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{akl}}{\pi^*_{kl}}\frac{y_k}{\pi_{ak}}\frac{y_l}{\pi_{al}}\right) &= \mathbb{E}_{p_a}\left(\sum_{k\in s_a}\sum_{l\in s_a}\frac{\Delta_{akl}}{\pi_{akl}}\frac{y_k}{\pi_{ak}}\frac{y_l}{\pi_{al}}\right)\\ &= \sum_{k\in U}\sum_{l\in U}\Delta_{akl}\frac{y_k}{\pi_{ak}}\frac{y_l}{\pi_{al}}\end{aligned}$$

y, para la segunda componente,

$$\mathbb{E}\left(\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}}\breve{\breve{y}}_k\breve{\breve{y}}_l\right) = \mathbb{E}_{p_a}\left(\sum_{k\in s_a}\sum_{l\in s_a}\Delta_{kl|s_a}\breve{\breve{y}}_{ak}\breve{\breve{y}}_{al}\right)$$

como se indica en el enunciado del teorema. ■

Nótese que la varianza (5.18) no se expresa de forma explícita sino como una valor esperado del diseño de primera fase. Esto es necesario ya que $\Delta_{kl|s_a}$ puede depender de la muestra s_a realmente obtenida. Afortunadamente, esto no causa ningún problema en la estimación de la varianza. Vemos que el estimador de la varianza (5.19) está indicado de forma explícita, lo que hace posible los cálculos directos.

[2] Sean X, Y variables aleatorias. Se cumplen:

- $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X|Y]]$.
- $\mathbb{V}[X] = \mathbb{V}_Y[\mathbb{E}[X|Y]] + \mathbb{E}_Y[\mathbb{V}[X|Y]]$.

Comentario 25. El estimador de la varianza mostrado en la ecuación (5.19) también se puede escribir como una expresión compacta de un solo término

$$\widehat{\mathbb{V}}(\widehat{Y}_U^{\mathrm{HT}^*}) = \sum_{k\in s}\sum_{l\in s} \check{\Delta}^*_{kl}\check{y}_k\check{y}_l, \tag{5.20}$$

donde

$$\Delta^*_{kl} = \pi^*_{kl} - \pi^*_k\pi^*_l \tag{5.21}$$

y

$$\check{\Delta}^*_{kl} = \frac{\Delta^*_{kl}}{\pi^*_{kl}} \tag{5.22}$$

Es muy sencillo demostrar la ecuación (5.20). Su simplicidad es de alguna forma engañosa. La varianza del estimador de la varianza puede resulta compleja en muchos casos prácticos. Se verá en el Ejemplo 19. ■

Ejemplo 17. El muestreo bietápico es, de hecho, un caso especial del muestreo bifásico. Consideremos el diseño bietápico aleatorio simple sin reemplazamiento en ambas etapas (véase el tema 1 de este bloque).

Usando muestreo aleatorio simple sin reemplazamiento en cada etapa, se seleccionan n_I PSUs de N_I en la primera etapa, y se submuestrean n_{s_i} de N_i, siempre y cuando se haya seleccionado la i-ésima PSU.

Asumimos que los tamaños n_{s_i} no son necesariamente fijos a priori, pero pueden depender del conjunto de s_I PSUs obtenido en la primera fase. Por ejemplo, los tamaños muestrales n_{s_i} pueden venir determinados por la ecuación (5.1). Sea $f_{s_i} = \frac{n_{s_i}}{N_i}$ y $\widehat{Y}_{U_i}^{\mathrm{HT}} = N_i\bar{y}_{s_i}$. Por el Teorema 12, el estimador $\widehat{Y}_U^{\mathrm{HT}^*}$ viene ahora dado por

$$\widehat{Y}_U^{\mathrm{HT}^*} = \frac{N_I}{n_I}\sum_{i\in s_I}\widehat{Y}_{U_i}^{\mathrm{HT}} \tag{5.23}$$

y su varianza por

$$\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}^*}) = N_I^2\frac{1-f_I}{n_I}S^2_{YU_I} + E_{srswor}\left(\frac{N_I^2}{n_I^2}\sum_{i\in s_I}N_i^2\frac{1-f_{s_i}}{n_{s_i}}S^2_{yU_i}\right), \tag{5.24}$$

donde $S^2_{YU_I} = \frac{1}{N_I-1}\sum_{i\in U_I}\left[Y_{U_i} - \bar{y}_{s_I}\right]^2$ y $S^2_{yU_i} = \frac{1}{N_i-1}\sum_{k\in U_i}\left[y_k - \bar{y}_{U_i}\right]^2$ (véase el tema 1 de este bloque). La segunda componente de la varianza viene dada de forma no explícita como una esperanza con respecto a la selección aleatoria simple sin reemplazamiento de la primera etapa. El estimador insesgado de la varianza es

$$\widehat{Y}_U^{\mathrm{HT}^*} = N_I^2\frac{1-f_I}{n_I}S^2_{\widehat{Y}s_I} + \frac{N_I}{n_I}\sum_{i\in s_I}N_i^2\frac{1-f_{s_i}}{n_{s_i}}S^2_{ys_i}, \tag{5.25}$$

donde $S^2_{\widehat{Y}_{s_I}} = \frac{1}{n_I - 1} \sum_{i \in s_I} \left(\widehat{Y}^{\text{HT}}_{U_i|i} - \frac{1}{n_I} \sum_{i \in s_I} \widehat{Y}^{\text{HT}}_{U_i|i} \right)^2$ y $S^2_{ys_i} = \frac{1}{n_{s_i} - 1} \sum_{k \in s_i} (y_k - \bar{y}_{s_i})^2$ (véase el tema 1 de este mismo bloque). Las conclusiones (5.24) y (5.25) son válidas de una forma más general que sus equivalentes en el muestreo bietápico, ya que ahora se permite fijar el tamaño muestral n_{s_i} convenientemente después de que se ha obtenido la muestra de PSUs.

Si los tamaños muestrales n_{s_i} se fijan de acuerdo con una regla fija (como en el muestreo bietápico), entonces la ecuación (5.25) confirma el estimador de la varianza proporcionado por (1.16) en el tema 1 de este mismo bloque. ■

5.4 Muestreo bifásico para la estratificación

El muestreo estratificado es una técnica que nos permite reducir en gran medida la varianza, siempre que los estratos estén bien construidos. Para obtener buenos estratos es necesario disponer de variables de estratificación correctas. Si no disponemos de esta información desde el principio podemos usar el muestreo bifásico de la siguiente manera.

Se selecciona una muestra grande en la primera fase y se estratifica con la ayuda de características auxiliares observadas, a bajo coste, para los elementos de esta muestra. La segunda fase se lleva a cabo entonces con un muestreo estratificado, con un tamaño muestral más pequeño, y se observa la variable objetivo y para esta muestra estratificada pequeña. El procedimiento se denomina *muestreo bifásico para la estratificación*. Veámoslo a continuación asumiendo que usamos el estimador $\widehat{Y}^{\text{HT}^*}_U$.

El diseño bifásico se define de la siguiente forma:

1. En primera fase se selecciona una muestra grande s_a de tamaño n_{s_a} de acuerdo con un diseño dado $p_a(\cdot)$. Para los elementos de s_a se recoge información que permita una estratificación.

2a. Se usa la información para estratificar s_a en H_{s_a} estratos denotados por s_{ah}, $h = 1, ..., H_{s_a}$ con $n_{s_{ah}}$ elementos en el estrato h. De esta forma

$$s_a = \bigcup_{h=1}^{H_{s_a}} s_{ah} \qquad n_{s_a} = \sum_{h=1}^{H_{s_a}} n_{s_{ah}}.$$

2b. Del estrato h se selecciona una muestra s_h, con $s_h \subset s_{ah}$, de tamaño n_h, de acuerdo con el diseño $p_h(\cdot|s_a)$. El submuestreo se realiza de forma independiente en cada estrato. Para la submuestra total s, la descomposición es

$$s = \bigcup_{h=1}^{H_{s_a}} s_h \qquad n_s = \sum_{h=1}^{H_{s_a}} n_{s_h}.$$

Comentario 26. Téngase en cuenta que los tamaños muestrales $n_{s_{ah}}$ y n_{s_h} son aleatorios. ■

Comentario 27. La interpretación más frecuente del procedimiento definido por (1), (2a) y (2b) es la siguiente. Se concibe un conjunto de experimentos, cada uno de los cuales consiste en la obtención de una muestra en primera fase s_a y, a continuación, la obtención de una submuestra en la segunda fase, de acuerdo con los diseños muestrales especificados en cada fase.

Cuando se obtiene s_a en primera fase, también los estratos están dados correspondientemente de modo que cada vez que se obtiene s_a en primera etapa, se emplea el mismo diseño estratificado (con la misma estratificación) en segunda etapa. Sin embargo, para dos muestras en primera fase que no sean idénticas, las estratificaciones en segunda etapa pueden ser diferentes. Así pues, el número de estratos puede ser diferente. Esto se identifica en la notación H_{s_a}, un número que puede variar de una muestra s_a a otra.

De forma ideal, los estratos se identifican mediante características que no son costosas de observar, pero que aún así son factores de peso para la variable de estudio y. ■

Los principales resultados para el estimador $\widehat{Y}_U^{\mathrm{HT}^*}$ para la estratificación en segunda fase se resumen en el siguiente teorema.

Teorema 13

En muestreo bifásico para la estratificación, el estimador $\widehat{Y}_U^{\mathrm{HT}^*}$ para el total poblacional Y se puede escribir como

$$\widehat{Y}_U^{\mathrm{HT}^*} = \widehat{Y}_U^{\mathrm{strHT}^*} \equiv \sum_{h=1}^{H_{s_a}} \sum_{k \in s_h} \frac{y_k}{\pi_k^*}. \tag{5.26}$$

La varianza de es

$$\mathbb{V}(\widehat{Y}_U^{\mathrm{strHT}^*}) = \sum_{k \in U} \sum_{l \in U} \Delta_{akl} \frac{y_{ak}}{\pi_{ak}} \frac{y_{al}}{\pi_{al}} + \mathbb{E}_{p_a}\Big(\sum_{h=1}^{H_{s_a}} \sum_{k \in s_{ah}} \sum_{l \in s_{ah}} \Delta_{kl|s_a} \breve{\breve{y}}_k \breve{\breve{y}}_l \Big), \tag{5.27}$$

donde $\breve{\breve{y}}_k = \frac{y_k}{\pi_k^*}$.

Un estimador insesgado de la varianza viene dado por

$$\widehat{\mathbb{V}}(\widehat{t}_{\pi^*}) = \sum \sum_{s} \frac{\Delta_{akl}}{\pi_{kl}^*} \breve{y}_{ak} \breve{y}_{al} + \sum_{h=1}^{H_{s_a}} \sum \sum_{s_h} \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}} \breve{\breve{y}}_k \breve{\breve{y}}_l. \tag{5.28}$$

Cada componente de (5.27) es insesgado de su equivalente en (5.28).

Demostración 13

Se sigue directamente del Teorema 12 al estratificar en la segunda etapa en H_{s_a} estratos independientes. ■

Las expresiones dadas por (5.27) y (5.28) son sencillas y compactas porque están enunciadas en términos generales. Cuando se aplican diseños específicos, las fórmulas pueden, sin embargo, volverse un poco largas, como se verá en los siguientes dos ejemplos.

Ejemplo 18. Apliquemos el Teorema 13 especificando que el diseño de segunda fase sea estratificado aleatorio simple sin reemplazamiento, mientras que el diseño de primera fase se mantiene general. Por tanto,

$$\pi_{k|s_a} = \frac{n_h}{n_{s_{ah}}} = f_h \quad \text{para} \quad k \in s_{ah}$$

y

$$\pi_{kl|s_a} = \begin{cases} f_h & \text{para} \quad k = l \in s_{ah}, \\ f_h \frac{n_h - 1}{n_{s_{ah}} - 1} & \text{para} \quad k \in s_{ah}, l \in s_{ah}, k \neq l, \\ f_h f_{h'} & \text{para} \quad k \in s_{ah}, l \in s_{ah'}, h \neq h'. \end{cases} \tag{5.29}$$

Del Teorema 13, el estimador $\widehat{Y}_U^{\text{strHT}^*}$ es ahora

$$\widehat{Y}_U^{\text{strHT}^*} = \sum_{h=1}^{H_{s_a}} n_{s_{ah}} \bar{\breve{y}}_{s_h} \tag{5.30}$$

donde

$$\bar{\breve{y}}_{s_h} = \frac{1}{n_h} \sum_{k \in s_h} \frac{y_k}{\pi_{ak}}. \tag{5.31}$$

Su varianza se puede escribir como

$$\mathbb{V}(\widehat{Y}_U^{\text{strHT}^*}) = \sum_{k \in U} \sum_{l \in U} \Delta_{akl} \frac{y_{ak}}{\pi_{ak}} \frac{y_{al}}{\pi_{al}} + \mathbb{E}_{p_a}\left(\sum_{h=1}^{H_{s_a}} n_{s_{ah}}^2 \frac{1 - f_h}{n_{s_h}} S_{\breve{y}s_{ah}}^2 \right) \tag{5.32}$$

donde $S_{\breve{y}s_{ah}}^2$ es la varianza en el estrato h de los valores elevados $\breve{y}_{ak} = \frac{y_k}{\pi_{ak}}$, es decir,

$$S_{\breve{y}s_{ah}}^2 = \frac{1}{n_{s_{ah}} - 1} \sum_{s_{ah}} (\breve{y}_{ak} - \bar{\breve{y}}_{s_{ah}})^2,$$

con $\bar{\check{y}}_{s_{ah}} = \sum_{k \in s_{ah}} \frac{\check{y}_{ak}}{n_{s_{ah}}}$.

La segunda componente de la ecuación (5.32), que contiene una forma estratificada familiar, se debe dejar como una esperanza, ya que $n_{s_{ah}}$, n_{s_h}, y H_{s_a} pueden estar determinados como una función de la muestra en primera fase s_a verdaderamente obtenida.

El estimador insesgado de la varianza viene dado por

$$\widehat{\mathbb{V}}(\widehat{Y}_U^{\text{strHT}^*}) = \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{akl}}{\pi^*_{kl}} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + \sum_{h=1}^{H_{s_a}} n_{s_{ah}}^2 \frac{1 - f_h}{n_{s_h}} S^2_{\check{y}s_h}, \tag{5.33}$$

donde $\pi^*_{kl} = \pi_{akl}\pi_{kl|s_a}$ con $\pi_{kl|s_a}$ dado por (5.29) y

$$S^2_{\check{y}s_h} = \frac{1}{n_{s_h} - 1} \sum_{s_h} (\check{y}_{ak} - \bar{\check{y}}_{s_h})^2 \tag{5.34}$$

con $\bar{\check{y}}_{s_h}$ dado por (5.31). Téngase en cuenta que si $U = s_a$ con probabilidad uno (es decir, si la población entera es seleccionada en la fase uno), entonces (5.32) y (5.33) se reducen a las fórmulas conocidas para el muestreo estratificado genérico. ■

Ejemplo 19. Continuemos con el ejemplo anterior especificando que la primera fase selecciona una muestra aleatoria simple sin reemplazamiento de n_a elementos de una población de tamaño N, mientras que se emplea un diseño estratificado aleatorio simple sin reemplazamiento en la segunda fase.

Sea $f_a = \frac{n_a}{N}$ la fracción de muestreo en primera fase, $w_{ah} = \frac{n_{s_{ah}}}{n_a}$ el tamaño relativo del estrato h y $f_h = \frac{n_h}{n_{s_{ah}}}$ la fracción de muestreo de la segunda fase del estrato h.

El Teorema 13 nos lleva a las siguientes conclusiones. El estimador $\widehat{Y}_U^{\text{strHT}^*}$ se puede escribir

$$\widehat{Y}_U^{\text{strHT}^*} = N \sum_{h=1}^{H_{s_a}} w_{ah} \bar{y}_{s_h} = N \widehat{\bar{y}}_U. \tag{5.35}$$

Su varianza es

$$\mathbb{V}(\widehat{Y}_U^{\text{strHT}^*}) = N^2 \frac{1 - f_a}{n_a} S^2_{yU} + \mathbb{E}_{srswor} \left(N^2 \sum_{h=1}^{H_{s_a}} w_{ah}^2 \frac{1 - f_h}{n_{s_h}} S^2_{ys_{ah}} \right) = V_1 + V_2. \tag{5.36}$$

Aquí, S^2_{yU} y $S^2_{ys_{ah}}$ denotan la varianza de y en U y en s_{ah}. A partir de la ecuación (5.33) obtenemos los siguientes estimadores insesgados de las componentes de la varianza

$$\widehat{V}_1 = N^2 \frac{1 - f_a}{n_a} \left[\sum_{h=1}^{H_{s_a}} w_{ah}(1 - \delta_h) S^2_{ys_h} + \frac{n_a}{n_a - 1} \sum_{h=1}^{H_{s_a}} w_{ah} (\bar{y}_{s_h} - \widehat{\bar{y}}_U)^2 \right] \tag{5.37}$$

y

$$\widehat{V}_2 = N^2 \sum_{h=1}^{H_{s_a}} w_{ah}^2 \frac{1-f_h}{n_{s_h}} S_{ys_h}^2 \tag{5.38}$$

donde $S_{ys_h}^2$ es la varianza de y en s_h y

$$\delta_h = \frac{1}{n_h}\left[\frac{(n_a - n_{s_{ah}})}{(n_a-1)}\right]$$

Sumando las componentes (5.37) y (5.38) obtenemos, después de algunos cálculos, el estimador insesgado de la varianza

$$\begin{aligned}\widehat{\mathbb{V}}(\widehat{Y}_U^{\text{strHT}^*}) &= \widehat{V}_1 + \widehat{V}_2 \\ &= N(N-1)\sum_{h=1}^{H_{s_a}}\left(\frac{n_{s_{ah}}-1}{n_a-1} - \frac{n_{s_h}-1}{N-1}\right)\frac{w_{ah}S_{ys_h}^2}{n_h} + \frac{N(N-n_a)}{n_a-1}\sum_{h=1}^{H_{s_a}} w_{ah}(\bar{y}_{s_h} - \widehat{\bar{y}}_U)^2\end{aligned} \tag{5.39}$$

Puesto que para calcular $S_{ys_h}^2$ es necesario que $n_h \geq 2$, el diseño de la segunda fase tiene que ser elegido teniendo en cuenta esto.

Cuando N es mucho más grande que n_a y $\frac{n_{s_{ah}}-1}{n_a-1} \doteq w_{ah}$, podemos aproximar el lado derecho de (5.39), lo que nos lleva a

$$\widehat{\mathbb{V}}(\widehat{t}_{\pi^*}) \doteq N^2 \sum_{h=1}^{H_{s_a}} \frac{w_{ah}^2 S_{ys_h}^2}{n_h} + \frac{N^2}{n_a}\sum_{h=1}^{H_{s_a}} w_{ah}(\bar{y}_{s_h} - \widehat{\bar{y}}_U)^2.$$

Aquí el primer término recuerda al estimador de la varianza usado en el muestreo postestratificado. El segundo término añade poco si el muestreo de la primera fase es considerablemente mayor que la muestra de la segunda fase. ■

Comentario 28. En el Ejemplo 19, los estratos se definieron después de examinar la muestra de primera fase s_a. Se permiten distintas estratificaciones para distintas muestras s_a. Una situación algo diferente se presenta cuando los estratos se fijan una vez y para todos, al nivel del total de la población. Es decir, existen H estratos predeterminados, $U_1, ..., U_h, ..., U_H$, que reflejan la subdivisión de la población usando un criterio dado, por ejemplo, un número fijo de grupos edad-sexo.

Estos estratos no están identificados antes del muestreo de la primera fase, pero cada muestra s_a puede estratificarse de acuerdo con el criterio predeterminado. Bajo estas condiciones, consideremos las siguiente selección en dos fases.

a. Sea n_{s_a} el tamaño de la muestra aleatoria simple sin reemplazamiento s_a seleccionada en primera fase. Sea $n_{s_{ah}}$ el número de elementos en s_a que pertenecen al estrato U_h.

b. La segunda fase se lleva a cabo mediante un diseño estratificado aleatorio simple sin reemplazamiento. El tamaño n_{s_h} de la submuestra s_h obtenida a partir de $s_{ah} = s_a \cap U_h$, viene determinada por

$$n_h = \nu_h n_{s_{ah}} \qquad 0 < \nu_h \leq 1,$$

donde los ν_h son constantes fijadas a priori, $h = 1, ..., H$.

Ahora hay una probabilidad no nula de que uno o más de los conjuntos s_{ah} esté vacío. Si n_{s_a} es muy grande, sin embargo, este evento tiene una probabilidad despreciable. Bajo esta hipótesis, obtenemos para el estimador $\widehat{Y}_U^{\text{strHT}^*}$ (5.35) la varianza

$$\mathbb{V}(\widehat{Y}_U^{\text{strHT}^*}) = N^2 \frac{1 - f_a}{n_{s_a}} S_{yU}^2 + N^2 \sum_{h=1}^{H} \frac{W_h S_{yU_h}^2}{n_{s_a}} \left(\frac{1}{\nu_h} - 1 \right) \tag{5.40}$$

donde $W_h = \frac{N_h}{N}$ es el tamaño relativo del estrato U_h y $S_{yU_h}^2$ denota la varianza de y en U_h. La ecuación (5.39) aún es el estimador de varianza apropiado si $\mathbb{P}(n_h \geq 2)$ está muy cerca de la unidad para todo k. Estos resultados fueron obtenidos por Rao 1973. ■

5.5 Variables auxiliares para la selección de la muestra en dos fases.

El estimador $\widehat{Y}_U^{\text{HT}^*}$ mostrado en la ecuación (5.17) depende únicamente de las ponderaciones de los elementos. Las probabilidades de inclusión del diseño muestral en segunda fase puede depender de información recogida en la primera fase, pero las variables auxiliares no aparecen en la fórmula del estimador $\widehat{Y}_U^{\text{HT}^*}$.

Consideramos ahora el uso explícito de variables auxiliares, en especial como parte de estimadores en diferencia y estimadores GREG para muestreo bifásico.

Los valores auxiliares pueden ser de dos tipos:

i. valores obtenido observando los elementos en la muestra de la primera fase s_a, es decir, valores que aparecen en el marco usado en la segunda fase;

ii. valores disponibles desde el principio para todos los N elementos de la población U, es decir, valores dados en el marco inicial.

Como ya hemos mencionado, una idea central en el muestreo bifásico es que se puede obtener la información auxiliar significativa del tipo (i). En muchas aplicaciones no hay información relevante del tipo (ii).

Usemos la siguiente notación:

a. Sea $\mathbf{x}_k$ un vector de J valores auxiliares disponibles para todo $k \in s_a$.

b. Sea $\mathbf{x}_{1k}$ el vector de J_1 valores auxiliares disponibles para todo k en la población U.

Asumimos que $\mathbf{x}_k$ contiene valores de variables conocidas previamente para todo U, así como los valores de las variables conocidas únicamente para $k \in s_a$.

En otras palabras, podemos escribir

$$\mathbf{x}_k = (\mathbf{x}_{1k}^t, \mathbf{x}_{2k}^t)^t,$$

donde $\mathbf{x}_{1k}$ es el vector de J_1 valores conocidos para todo U, si están disponibles, y $\mathbf{x}_{2k}$ es el vector de $J_2 = J - J_1$ valores grabados por observación (relativamente no costosa) de elementos k en la muestra de primera fase s_a únicamente.

En los estimadores GREG que presentaremos, $\mathbf{x}_k$ sirve para obtener valores y predichos, $\widehat{y}_k$ de s a s_a, y $\mathbf{x}_{1k}$ se usa para obtener predicciones $\widehat{y}_{1k}$ de s_a a U. Las cantidades conocidas, predichas u observadas se resumen en la tabla 5.1.

Conjunto de elementos	Valores de vector auxiliares conocidos	Valores de variables de estudio observadas	Valores de variables de estudio predichas
Población, U	$\mathbf{x}_{1k}$	-	$\widehat{y}_{1k}$
Muestra primera fase, s_a	$\mathbf{x}_k = (\mathbf{x}_{1k}^t, \mathbf{x}_{2k}^t)^t$	-	$\widehat{y}_k, \widehat{y}_{1k}$
Muestra segunda fase, s	$\mathbf{x}_k = (\mathbf{x}_{1k}^t, \mathbf{x}_{2k}^t)^t$	y_k	$\widehat{y}_k$

Tabla 5.1: Clasificación de variables objetivo y auxiliares.

El objetivo de la incorporación explícita de esta información auxiliar es mejorar el estimador $\widehat{Y}_U^{\mathrm{HT}^*}$ suponiendo que el diseño muestral de cada una de las dos fases es fijo. La medida en la que se reduce la varianza por el uso de un estimador GREG dependerá de (a) la capacidad predictora de $\mathbf{x}_2$ sobre y, y (b) la capacidad predictora de $\mathbf{x}_1$ sobre y. Se pueden distinguir distintos casos.

El caso típico de una encuesta que usa muestreo bifásico es que $\mathbf{x}_2$ es un predictor potente de y, mientras que $\mathbf{x}_1$ es un predictor débil de y. En la fase de planificación, los estadísticos usan su criterio para identificar variables x que tienen una alta probabilidad de explicar bien y, y estas variables formarán el vector $\mathbf{x}_2$, y sus valores, $\mathbf{x}_{2k}$, se observan para los elementos $k \in s_a$.

Cuando la primera fase consiste en una selección de una muestra y la segunda fase corresponde a una subselección provocada por la falta de respuesta, también es importante distinguir dos categorías de variables auxiliares, $\mathbf{x}_{1k}$ y $\mathbf{x}_{2k}$. El vector $\mathbf{x}_{2k}$ contiene los valores de las variables incluidas en x que son conocidas o están disponibles para todos los elementos de la muestra, tanto los que proporcionan los datos como los que

no. Sólo los que responden proporcionan los valores de y_k.

Los valores de $\mathbf{x}_{1k}$ son conocidos para toda la población. Por razones que no comentamos ahora (véase Särndal, Swensson y Wretman 1992, cap. 15), es importante que $\mathbf{x}_2$ en particular, pero preferiblemente también $\mathbf{x}_1$ sean predictores potentes de y. Pueden ser necesarios esfuerzos y costes para obtener los valores de $\mathbf{x}_{2k}$, pero su presencia mejora la inferencia de forma significativa.

Cabe señalar que además del muestreo bifásico para la estratificación visto en la Sección 5.4, existen otros diseños como el muestreo bifásico de estimadores diferencia (véase la sección 9.6 del Särndal, Swensson y Wretman 1992), que son un paso intermedio a los estimadores de regresión en muestreo bifásico (véase la sección 9.7 del Särndal, Swensson y Wretman 1992).

Bibliografía

Grimmet, G.R. y D.R. Stirzaker (2004). *Probability and random processes*. 3rd. Oxford Science Publications.

Neyman, J. (1938). "Contribution to the theory of sampling human populations". En: *Journal of the American Statistical Association* 33, págs. 101-116.

Rao, J.N.K. (1973). "On double sampling for stratification and anlytic surveys". En: *Biometrika* 60, págs. 125-133.

Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.

Tema 6

Muestreo en dos ocasiones. Estimación del total en cada ocasión. Estimación del cambio absoluto. Estimación de la suma de totales.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía.

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

6.1 Muestreo en dos ocasiones

En muchas encuestas la misma población puede ser muestreada de forma repetida y la misma variable de estudio medida en cada ocasión, de forma que se puede hacer un seguimiento a lo largo del tiempo. Las encuestas repetidas nos permiten estudiar tendencias o cambios en las características de interés a lo largo de un periodo de tiempo.

Por ejemplo, en muchos países, la Encuesta de Población Activa[1] (EPA) se realiza mensual o trimestralmente para estimar el número de ocupados o la tasa de desempleo. Otros ejemplos son las encuestas mensuales en las que se recogen los datos de los precios de bienes para determinar el Índice de Precios de Consumo (IPC) [2] o las encuestas de opinión realizadas con cierta periodicidad para medir las preferencias de los votantes. En el caso de encuestas repetidas se utilizan técnicas especiales.

En el tema 2 de este mismo bloque ya se aborda una primera aproximación a la realización de encuestas a lo largo del tiempo para conocer las características de una población

[1] https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176918&menu=ultiDatos&idp=1254735976595

[2] https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176802&menu=ultiDatos&idp=1254735976607

que evoluciona. Estos cambios en las características de una población a lo largo del tiempo plantean cuestiones para el análisis. Por un lado, se necesitan estimar determinadas características de la población de forma reiterada a lo largo del tiempo para obtener estimaciones lo más recientes posibles. Pero también interesa conocer el cambio que ha tenido lugar en la estimaciones a lo largo del tiempo: ¿la tasa de desempleo ha aumentado o disminuido desde la encuesta anterior? Este cambio se denomina cambio neto y refleja los cambios tanto en las características como en la composición de la población.

Un análisis más detallado implica conocer las componentes del cambio. ¿Hasta qué punto el cambio (o la ausencia de cambios) se debe a la dinámica de la población, con personas formando parte de la población mediante los 'nacimientos' (por ejemplo, con personas que cumplen los 16 años o inmigrantes) y abandonando la población a través de las 'muertes' (por ejemplo, personas que fallecen, emigran o se jubilan)? ¿Hasta qué punto el cambio se debe a cambios en los estados de los individuos de la población? Más aún, ¿cómo funciona el cambio en casos en los que hay un cambio de estado? Por ejemplo, suponiendo que no hay dinámicas poblacionales, si la tasa de desempleo sufre un incremento neto del 1 %, ¿esto se debe a que un 1 % de las personas previamente empleadas perdieron su trabajo o a que, por ejemplo, un 10 % perdió su empleo y un 9 % de estos desempleados encontraron trabajo?

La descomposición del cambio neto en sus dos componentes lleva a medir el cambio bruto. Mientras que el cambio neto se puede medir usando muestras separadas para las dos ocasiones, medir el cambio bruto requiere medidas repetidas en la misma muestra o, por lo menos, en una submuestra representativa.

Hay dos grandes clases de objetivos para las encuestas a lo largo del tiempo, como se vio en el tema 2, que dan lugar a distintos enfoques en el diseño de encuestas. En muchos casos, los objetivos se restringen a estimar los parámetros poblacionales o económicos en distintos instantes de tiempo, así como los cambios netos, tendencias o valores medios a lo largo de un periodo de tiempo. Ninguno de estos objetivos requiere mediciones repetidas sobre la misma muestra. En particular, estos objetivos se pueden alcanzar con muestras totalmente independientes en cada instante de tiempo y también con muestras que se construyen para minimizar el solapamiento de la muestra a lo largo del tiempo y la carga de respuesta al informante.

Algunos objetivos también se pueden alcanzar con diseños panel que incluyen a algunos o a todos los miembros de la muestra en distintos instantes de tiempo. De hecho, la precisión de las estimaciones transversales y los cambios netos se pueden mejorar usando un diseño muestral de panel rotante que cree algún grado de solapamiento en la muestra a lo largo del tiempo. Los diseños de panel rotante también se pueden usar para eliminar los efectos telescópicos que ocurren cuando los informantes informan erróneamente de un evento como si haya ocurrido en un intervalo de tiempo dado.

Otros objetivos se centran en la estimación del cambio bruto y en otras componentes de cambio individual y en la agregación de respuestas, por ejemplo, los gastos de individuos a lo largo del tiempo. Estos objetivos pueden satisfacerse solo con algunas formas de encuestas panel que recogen datos de los mismos individuos para el periodo de interés. Asimismo, otros objetivos de las encuestas a lo largo del tiempo están relacionados con la obtención de estimaciones para poblaciones escasas[3] (es decir, un subconjunto de la población general que tiene una característica escasa, muy poco frecuente). Uno de esos objetivos es acumular una muestra de casos con esa característica escasa a lo largo del tiempo. Si la característica es un evento, como divorciarse, entonces este objetivo puede satisfacerse con cualquiera de los diseños. Sin embargo, si la característica es estable, como, por ejemplo, pertenecer a un determinado grupo étnico, la acumulación solo funciona cuando se incluyen nuevas muestras a lo largo del tiempo.

En cualquier caso, los analistas necesitan reconocer que las características de una población escasa puede variar a lo largo del tiempo. Un objetivo distinto con una población escasa es la producción de estimaciones para esa población en distintos instantes de tiempo. Si la característica escasa es estable, se puede identificar una muestra de miembros de esa población en un instante de tiempo y luego volver a esa muestra de forma reiterada en un diseño de panel.

Otra característica de las encuestas repetidas es que, en general, una gran cantidad de información que está disponible será útil para los diseños que se usarán en el futuro. La idoneidad de varias características del diseño muestral, como la adecuación de las variables de estratificación y los límites de los estratos, el método de afijación muestral y el tamaño de las unidades en las distintas etapas de un diseño multietápico pueden estudiarse a lo largo del tiempo con vistas a mejorar la eficiencia estadística. A menudo la información necesaria para diseñar de forma eficiente una encuesta que se realiza una única vez es muy limitada.

En el diseño de una encuesta repetida, habrá que tener en cuenta circunstancias como nacimientos, muertes o cambios en el tamaño. Los métodos de muestreo y de estimación usados en las encuestas repetidas deberían incorporar estos cambios de una forma estadísticamente eficiente, con tan pocas alteraciones en las operaciones estadísticas en curso como sea posible.

Las encuestas repetidas miden los cambios en las características de una población determinada con mayor precisión de lo que lo hacen una serie de muestras independientes de tamaño comparable. Entre las ventajas de las encuestas repetidas nos encontramos:

i. Reduce la varianza muestral para las estimaciones de cambio, es decir, $\widehat{Y}_2 - \widehat{Y}_1$, donde $\widehat{Y}_1$ es una estimación del total en el instante 1 e $\widehat{Y}_2$ es una estimación del total en el instante 2.

[3] *Rare population*, en inglés.

ii. Se puede usar para obtener información sobre el comportamiento de las unidades estadísticas a lo largo del tiempo.

iii. Puede reducir el error de respuesta (ya que los informantes tienen un mejor conocimiento del cuestionario).

iv. Puede reducir el coste a lo largo del tiempo (desarrollo del cuestionario, programación, formación del personal, etc.).

Aunque este tipo de muestreo también presenta algunos inconvenientes:

i. Su estimación, tratamiento de la falta de respuesta, etc., es más complejo.

ii. Requiere un presupuesto que garantice que la encuesta se mantendrá en el tiempo.

iii. Es más difícil mantener la acuracidad a lo largo de períodos muy largo de tiempo debido a los cambios que tienen lugar en la población.

iv. Puede aumentar el error de respuesta. Por ejemplo, el conocimiento de los informantes sobre el cuestionario puede dar lugar a respuestas incorrectas con el fin de responder con más rapidez al cuestionario.

v. Puede dar lugar a una falta de respuesta más alta a largo plazo, debido al cansancio del informante.

vi. Su organización es más compleja que la de una única encuesta puntual.

vii. Puede dar lugar a un comportamiento inducido por la encuesta. Por ejemplo, un informante al que siempre se le pregunta sobre sus visitas al médico puede empezar a visitar al médico como consecuencia de la encuesta.

viii. Puede ser difícil definir algunos conceptos. Por ejemplo, la composición del hogar puede cambiar a lo largo del tiempo, por tanto, ¿cómo se define un hogar en una encuesta repetida?

ix. Si la muestra inicial seleccionada es 'mala', la oficina estadística tiene que seguir con esa muestra.

En este tema veremos el muestreo en dos ocasiones. Un punto clave es cuántos elementos muestreados en una ocasión anterior deberían mantenerse en la muestra seleccionada en la ocasión actual. La 'proporción de emparejamiento'[4] óptima depende del parámetro a estimar. El problema de emparejamiento[5] ha sido estudiado por varios autores, siendo Patterson 1950 una de las primeras referencias.

En este tema veremos cómo trabajar en casos de diseños muestrales generales. Consideramos el muestreo en dos ocasiones de una población finita $U = \{1, ..., k, ..., N\}$ asumiendo que está compuesta por los *mismos* elementos en dos ocasiones distintas[6].

[4] *Matching proportion*, en inglés.
[5] *Matching problem*, en inglés.
[6] Realmente es muy difícil que los elementos sean exactamente los mismos en dos ocasiones sucesivas debido a las altas y bajas que se producen en todas las poblaciones de manera continua. Pero asumiremos que las diferencias existentes entre las dos ocasiones son despreciables

La variable de estudio, por ejemplo, la tasa de desempleo o el ingreso bruto de los hogares, se estudia en casa ocasión, pero no necesariamente para los mismos elementos. La variable de estudio se denotará por z en la primera ocasión y por y en la segunda.

En la primera ocasión, una muestra s_a se selecciona mediante un diseño muestral $p_a(\cdot)$ y se mide la variable z para todos los elementos en s_a. Las probabilidades de inclusión asociadas con este diseño se denotan por π_{ak} y π_{akl}. Denotamos $\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}$. El estimador de Horvitz-Thompson

$$\widehat{z}_U^{\mathrm{HT}}(s_a) = \sum_{k \in s_a} \frac{z_k}{\pi_{ak}}$$

se usa como estimador del total $Z_U = \sum_{k \in U} z_k$. La muestra s_a seleccionada en la primera ocasión tiene una muestra complementaria, $s_a^c = U - s_a$. La muestra complementaria no es encuestada en la primera ocasión, pero necesitamos las probabilidades de inclusión de la muestra complementaria resultante del diseño $p_a(\cdot)$. Denotaremos por π_{ak}^c la probabilidad de que k es un elemento de s_a^c y por π_{akl}^c la probabilidad de que tanto k como l sean elementos de s_a^c. Denotaremos $\Delta_{akl}^c = \pi_{akl}^c - \pi_{ak}^c\pi_{al}^c$. Por tanto,

$$\begin{aligned}
\pi_{ak}^c &= 1 - \pi_{ak} \\
\pi_{akl}^c &= 1 - \pi_{ak} - \pi_{al} + \pi_{akl} \\
\Delta_{akl}^c &= \Delta_{akl}
\end{aligned}$$

El diseño muestral que se debería usar en la segunda ocasión, en el caso de encuestas que no se repetirán, depende de varios factores, como los parámetros a estimar, la disponibilidad de información auxiliar, el coste y consideraciones de medida (recogida de datos) y otros. Cuando la misma variable de estudio se observa en encuestas repetidas, el interés normalmente recae en estimar tanto *parámetros de nivel* como *parámetros de cambio*. Un gran número de parámetros de interés son de la forma

$$T_U = \phi Z_U + \psi Y_U$$

donde $Z_U = \sum_{k \in U} z_k$, $Y_U = \sum_{k \in U} y_k$ y ϕ y ψ son constantes.

Por ejemplo:

a. $\phi = 0, \psi = 1$ implica $T_U = Y_U$, *el total actual*, que es un parámetro de nivel;

b. $\phi = -1, \psi = 1$ implica $T_U = Y_U - Z_U$, *el cambio absoluto*;

c. $\phi = -\frac{1}{Z_U}, \psi = \frac{1}{Z_U}$ implica $T_U = \frac{Y_U - Z_U}{Z_U}$, *el cambio relativo*; y

d. $\phi = 1, \psi = 1$ implica $T_U = Y_U + Z_U$, que es *la suma de los totales* a lo largo de las dos ocasiones para la característica bajo estudio.

Al elegir un diseño para la segunda ocasión, tenemos más información que en la primera: para cada $k \in s_a$ conocemos el valor z_k. Para el nuevo diseño podemos no considerar ningún solapamiento, considerar solapamiento total o considerar solapamiento parcial

con la primera muestra s_a. Como se ve en este tema, distintos parámetros tienen distintos diseños muestrales óptimos en la segunda ocasión.

Está bastante claro que hay casos en los que la información de la primera ocasión se puede usar para mejorar la estimación. Por tanto, optamos por el solapamiento parcial. En la segunda ocasión, se seleccionan dos muestras independientes, una *muestra aparejada*[7] y una *muestra no aparejada*[8]. La muestra aparejada, denotada por s_m, se selecciona de s_a mediante el diseño $p_m(\cdot|s_a)$. La muestra no aparejada, denotada por s_u, se selecciona de s_a^c mediante el diseño $p_u(\cdot|s_a^c)$ y es independiente de s_m.

Las cantidades $\pi_{k|s_a}$, $\pi_{kl|s_a}$ y $\Delta_{kl|s_a} = \pi_{kl|s_a} - \pi_{k|s_a}\pi_{l|s_a}$ están asociadas con $p_m(\cdot|s_a)$ y $\pi_{k|s_a^c}$, $\pi_{kl|s_a^c}$ y $\Delta_{kl|s_a^c} = \pi_{kl|s_a^c} - \pi_{k|s_a^c}\pi_{l|s_a^c}$ son las cantidades análogas para $p_u(\cdot|s_a^c)$. La variable y se observa para todos los elementos en s_m y s_u. La muestra total en la segunda ocasión es por tanto $s = s_m \cup s_u$.

Ejemplo 20. Este ejemplo se ha extraído de (Hidiroglou y Lavallée 2009) (véase también Pfefferman y Rao 2009). Un problema bastante importante en los marcos de las encuestas económicas es el de la sobrecobertura causada por las unidades muertas [9]. Veamos cómo el muestreo en dos ocasiones puede servir para resolver este problema de las unidades muertas.

Las unidades muertas en una muestra son representativas del total de unidades muestras en el marco de empresas. Estas unidades se mantienen en el marco y son tratadas como ceros en la muestra. Si tuviésemos fuentes que nos permitiesen identificar las unidades muestras, ¿cómo las usamos? Usaremos el muestreo en dos ocasiones para un único estrato para mostrar cómo se puede usar durante la estimación.

En la encuesta de la ocasión 1, se obtiene una muestra s_1 de tamaño n_1 mediante muestreo aleatorio simple sin reemplazamiento a partir de una población U_1 de tamaño N_1. En la segunda ocasión, la población U_2 consistirá de todas las unidades de U_1, así como de un conjunto U_b [10] de altas de tamaño N_b que han tenido lugar entre las dos ocasiones. Supongamos que un subconjunto de U_1 ha muerto entre la creación de U_1 y la recogida de datos de s_1. Este subconjunto denotado por U_d consiste en N_d unidades muertas que son desconocidas. Supongamos que una nueva fuente identifica A_d de las N_d unidades muertas desconocidas, donde $A_d < N_d$. Durante la recogida de datos de s_1 también se observan n_d muertes en la muestra s_1 y a_s ($a_d < n_d$) de estas muertes

[7] *Matched sample*, en inglés.

[8] *Unmatched sample*, en inglés.

[9] Se denominan unidades muertas del inglés *dead units* a aquellas unidades que ya no realizan actividades económicas pero que siguen figurando en los registros administrativos que se usan como fuente de los marcos de unidades económicas. Estas unidades siguen figurando en los registros ya que no realizan las gestiones oportunas para darse de baja en los mismos o por el tiempo que transcurre desde que se dan de baja en el registro administrativo y esa información es recibida en los INEs. En español se denominan bajas. En este ejemplo usaremos muertes y bajas indistintamente.

[10] El subíndice b viene del inglés *birth*, que podemos traducir como alta.

también son identificadas a través de la fuente administrativa. La muestra s_1 se amplía con una muestra s_b de tamaño $n_b = f_1 N_b$, donde $f_1 = \frac{n_1}{N_1}$, seleccionada usando muestreo aleatorio simple sin reemplazamiento a partir de U_b. Si todas las muertes se mantienen en la muestra, entonces la muestra resultante consiste en $n_2 = n_1 + n_b$ unidades, de las cuales n_d se sabe que están muertas.

Supongamos que el parámetro de interés es el total poblacional $Y_{U_2} = \sum_{k \in U_2} y_k$. Un estimador insesgado de Y_2 vienen dado por $\widehat{Y}_{U_2}^{\text{HT}} = \frac{N_1}{n_1} \sum_{k \in s_2} y_k$, donde $s_2 = s_1 \cup s_b$. Cabe señalar que al menos n_d unidades están muertas en la muestra (porque algunas más muertes han tenido lugar durante la recogida de los datos de la segunda ocasión). Un estimador más eficiente para el total Y_{U_2} viene dado por el estimador postestratificado $\widehat{Y}_{U_2}^{\text{PS}} = \frac{N_2}{\widehat{N}_{U_2}^{\text{HT}}} \widehat{Y}_{U_2}^{\text{HT}}$, donde $\widehat{N}_{U_2}^{\text{HT}} = \frac{N_1}{n_1}(n_1 - a_d + n_b)$ y $N_2 = N_1 - A_d + N_b$. Este estimador es un estimador de razón y, por tanto, es aproximadamente insesgado. ■

6.2 Estimación del total en cada ocasión

6.2.1 Estimador del total actual

En muchos casos, hay una buena razón para asumir que y_k está bien aproximado por $y_k^0 = K z_k$, donde K es una constante conocida. El valor de K puede venir sugerida por un estudio anterior o por teoría sobre la cuestión bajo análisis. Usando la primera muestra s_a, la muestra aparejada s_m y las diferencias $D_k = y_k - y_k^0$, podemos construir un estimador diferencia insesgado del total actual, Y_U, concretamente,

$$\widehat{Y}_U^{(1)} = \widehat{Y}_U^0(s_a) + \widehat{T}_D(s_m), \tag{6.1}$$

donde

$$\widehat{Y}_U^0(s_a) = \sum_{k \in s_a} \frac{y_k^0}{\pi_{ak}} \quad \text{y} \quad \widehat{T}_D(s_m) = \sum_{k \in s_m} \frac{D_k}{\pi_{ak} \pi_{k|s_a}}.$$

Un segundo estimador del total actual se puede obtener de la muestra no aparejada, concretamente,

$$\widehat{Y}_U^{(2)} = \widehat{Y}_U^{\text{HT}}(s_u) = \sum_{k \in s_u} \frac{y_k}{\pi_{ak}^c \pi_{k|s_a^c}}. \tag{6.2}$$

Es fácil ver que tanto $\widehat{Y}_U^{(1)}$ como $\widehat{Y}_U^{(2)}$ son insesgados para el total actual. Mediante una combinación lineal obtenemos el nuevo estimador insesgado

$$\widehat{Y}_U = w_1 \widehat{Y}_U^{(1)} + w_2 \widehat{Y}_U^{(2)}, \tag{6.3}$$

donde w_1 y w_2 son ponderaciones constantes no negativas a determinar y tales $w_1 + w_2 = 1$.

Llamaremos a $\widehat{Y}_U$ un *estimador compuesto*; combina el estimador de la muestra aparejada con el estimador de la muestra no aparejada. La elección óptima de $w_1 = 1 - w_2$ será

considerada más adelante.

En primer lugar, veamos el siguiente resultado, en el que denotamos $V_1 = \mathbb{V}\left(\widehat{Y}_U^{(1)}\right)$, $V_2 = \mathbb{V}\left(\widehat{Y}_U^{(2)}\right)$ y $C = \mathbb{C}\left(\widehat{Y}_U^{(1)}, \widehat{Y}_U^{(1)}\right)$.

Teorema 14

La varianza del estimador compuesto (6.3) viene dada por

$$\mathbb{V}(\widehat{T}_y) = w_1^2 V_1 + w_2^2 V_2 + 2w_1 w_2 C, \tag{6.4}$$

donde

$$V_1 = \sum_{k\in U}\sum_{l\in U} \Delta_{akl}\frac{y_k}{\pi_{ak}}\frac{y_l}{\pi_{al}} + \mathbb{E}_a\Big(\sum_{k\in s_a}\sum_{l\in s_a}\Delta_{kl|s_a}\frac{D_k}{\pi_{ak}\pi_{k|s_a}}\frac{D_l}{\pi_{al}\pi_{l|s_a}}\Big) \tag{6.5}$$

$$V_2 = \sum_{k\in U}\sum_{l\in U} \Delta^c_{akl}\frac{y_k}{\pi^c_{ak}}\frac{y_l}{\pi^c_{al}} + \mathbb{E}_a\Big(\sum_{k\in s^c_a}\sum_{l\in s^c_a}\Delta_{kl|s^c_a}\frac{y_k}{\pi^c_{ak}\pi_{k|s^c_a}}\frac{y_l}{\pi^c_{al}\pi_{l|s^c_a}}\Big) \tag{6.6}$$

y

$$C = -\sum\sum_{k\in U}\Delta_{akl}\frac{y_k}{\pi_{ak}}\frac{y_l}{\pi^c_{al}} \tag{6.7}$$

Aquí, la esperanza $\mathbb{E}_a$ en las ecuaciones (6.5) y (6.6) está tomada con respecto al diseño muestral $p_a(\cdot)$.

Demostración 14

Se sigue de $\mathbb{V}(X) = \mathbb{V}(\mathbb{E}(X|Y)) + \mathbb{E}(\mathbb{V}(X|Y))$ y las propiedades elementales de la varianza.

Ejemplo 21. Las expresiones en el Teorema 14 toman formas sencillas cuando se usa el muestreo aleatorio simple. Suponemos que s_a es una muestra aleatoria simple sin reemplazamiento de tamaño n de U, lo que significa que s_a^c es una muestra aleatoria simple sin reemplazamiento de tamaño $N-n$ de U. Sea $f = \frac{n}{N}$. Además, asumimos que s_m es una muestra aleatoria simple sin reemplazamiento de tamaño $m = \mu n$ de s_a y que s_u es una muestra aleatoria simple sin reemplazamiento de tamaño $u = n - m = (1-\mu)n = \nu n$ de s_a^c. Aquí, la cantidad $\mu = 1 - \nu$ se llama la *proporción de emparejamiento*. Luego

$$\widehat{Y}_U^{(1)} = N\bar{y}^0_{s_a} + N\bar{D}_{s_m} = N[\bar{y}_{s_m} + K(\bar{z}_{s_a} - \bar{z}_{s_m})], \tag{6.8}$$

mientras que

$$\widehat{Y}_U^{(2)} = N\bar{y}_{s_u}. \tag{6.9}$$

Se sigue fácilmente que

$$V_1 = N^2\left(\frac{1-f}{n}S_{yU}^2 + \frac{1-\mu}{\mu n}S_{DU}^2\right), \tag{6.10a}$$

$$V_2 = N^2\frac{1-\nu f}{\nu n}S_{yU}^2, \tag{6.10b}$$

$$C = -N^2 S_{yU}^2. \tag{6.10c}$$

La proporción de emparejamiento y el óptimo valor de K se determinarán más tarde. ■

Determinemos ahora $w_1 = 1 - w_2$ y K en el estimador compuesto (6.3) de forma que se minimice la varianza. Obtenemos

$$\begin{aligned}\mathbb{V}(\widehat{Y}_U) &= w_1^2 V_1 + w_2^2 V_2 + 2w_1 w_2 C \\ &= (V_1 + V_2 - 2C)\Big\{w_1 - \frac{V_2 - C}{V_1 + V_2 - 2C}\Big\}^2 + \frac{V_1 V_2 - C^2}{V_1 + V_2 - 2C} \\ &\geq \frac{V_1 V_2 - C^2}{V_1 + V_2 - 2C} = \mathbb{V}_{min}(\widehat{Y}_U)\end{aligned} \tag{6.11}$$

ya que $V_1 + V_2 - 2C = V\left(\widehat{Y}_U^{(1)} - \widehat{Y}_U^{(2)}\right) > 0$. La igualdad se verifica si y sólo si

$$w_1 = 1 - w_2 = \frac{V_2 - C}{V_1 + V_2 - 2C}. \tag{6.12}$$

La mínima varianza

$$\mathbb{V}_{min}\left(\widehat{Y}_U\right) = \frac{V_1 V_2 - C^2}{V_1 + V_2 - 2C} \tag{6.13}$$

es una función estrictamente creciente de V_1 y V_1 es la única componente de (6.13) que depende de K.

Minimizar (6.13) es equivalente a encontrar el valor de K que minimiza V_1. Sean

$$\widehat{Y}_U^{\mathrm{HT}}(s_m) = \sum_{k\in s_m}\frac{y_k}{\pi_{ak}\pi_{k|s_a}}, \quad \widehat{Z}_U^{\mathrm{HT}}(s_m) = \sum_{k\in s_m}\frac{z_k}{\pi_{ak}\pi_{k|s_a}} \quad \text{y} \quad \widehat{Z}_U^{\mathrm{HT}}(s_a) = \sum_{k\in s_a}\frac{z_k}{\pi_{ak}}.$$

Entonces $\widehat{Y}_U^{(1)}$ se puede escribir

$$\widehat{Y}_U^{(1)} = \widehat{Y}_U^{\mathrm{HT}}(s_m) + K\left(\widehat{Y}_U^{\mathrm{HT}}(s_a) - \widehat{Z}_U^{\mathrm{HT}}(s_m)\right) \tag{6.14}$$

y una expresión alternativa para su varianza es

$$\mathbb{V}\left(\widehat{Y}_U^{(1)}\right) = \mathbb{V}(\widehat{Y}_U^{\mathrm{HT}}(s_m)) + K^2\mathbb{V}(\widehat{Z}_U^{\mathrm{HT}}(s_a) - \widehat{Z}_U^{\mathrm{HT}}(s_m)) + 2K\mathbb{C}(\widehat{Y}_U^{\mathrm{HT}}(s_m), \widehat{Z}_U^{\mathrm{HT}}(s_a) - \widehat{Z}_U^{\mathrm{HT}}(s_m))$$

$$
\begin{aligned}
&= \mathbb{V}(\widehat{Y}_U^{\mathrm{HT}}(s_m)) + K^2\mathbb{E}_a\mathbb{V}(\widehat{Z}_U^{\mathrm{HT}}(s_m)\big|s_a) - 2K\mathbb{E}_a\mathbb{C}\left(\widehat{Y}_U^{\mathrm{HT}}(s_m), \widehat{Z}_U^{\mathrm{HT}}(s_m)\big|s_a\right) \\
&= \mathbb{V}(\widehat{Y}_U^{\mathrm{HT}}(s_m)) + \mathbb{E}_a\mathbb{V}\left(\widehat{Z}_U^{\mathrm{HT}}(s_m)\big|s_a\right)\left\{K - \frac{\mathbb{E}_a\mathbb{C}\left(\widehat{Y}_U^{\mathrm{HT}}(s_m), \widehat{Z}_U^{\mathrm{HT}}(s_m)\big|s_a\right)}{\mathbb{E}_a\mathbb{V}\left(\widehat{Z}_U^{\mathrm{HT}}(s_m)\big|s_a\right)}\right\}^2 \\
&- \frac{[\mathbb{E}_a\mathbb{C}(\widehat{Y}_U^{\mathrm{HT}}(s_m), \widehat{Z}_U^{\mathrm{HT}}(s_m)|s_a)]^2}{\mathbb{E}_a\mathbb{V}(\widehat{Z}_U^{\mathrm{HT}}(s_m)\big|s_a)} \qquad (6.15) \\
&\geq \mathbb{V}(\widehat{Y}_U^{\mathrm{HT}}(s_m)) - \frac{[\mathbb{E}_a\mathbb{C}(\widehat{Y}_U^{\mathrm{HT}}(s_m), \widehat{Z}_U^{\mathrm{HT}}(s_m)|s_a)]^2}{\mathbb{E}_a\mathbb{V}(\widehat{Z}_U^{\mathrm{HT}}(s_m)\big|s_a)} = \mathbb{V}_{min}\left(\widehat{Y}_U^{(1)}\right). \qquad (6.16)
\end{aligned}
$$

Aquí los operadores $\mathbb{V}$ y $\mathbb{C}$ están asociados a $p_m(\cdot|s_a)$ y la esperanza $\mathbb{E}_a$ está asociada a $p_a(\cdot)$.

La igualdad se verifica si y sólo si

$$K = \frac{\mathbb{E}_a\mathbb{C}(\widehat{Y}_U^{\mathrm{HT}}(s_m), \widehat{Z}_U^{\mathrm{HT}}(s_m)\big|s_a)}{\mathbb{E}_a\mathbb{V}(\widehat{Z}_U^{\mathrm{HT}}(s_m|s_a))} = K_{opt}. \qquad (6.17)$$

Esta expresión es compleja en general, pero, en casos especiales, se puede evaluar con facilidad, como en el Ejemplo 22.

Ejemplo 22. Volvamos al Ejemplo 21. Sea el coeficiente de correlación entre y y z en la población U

$$r = r_{yzU} = \frac{S_{yzU}}{S_{yU}S_{zU}}.$$

Entonces

$$K_{opt} = r\frac{S_{yU}}{S_{zU}} \qquad (6.18)$$

y con $K = K_{opt}$, la ecuación (6.10a) se convierte en

$$V_1 = N^2\frac{S_{yU}^2}{\mu n}[(1-r^2) + \mu(r^2 - f)] \qquad (6.19)$$

Además,

$$V_2 = N^2\frac{S_{yU}^2}{\mu n}\frac{\mu(1-\nu f)}{\nu} \qquad (6.20)$$

y

$$C = -N^2\frac{S_{yU}^2}{\mu n}(-\mu f) \qquad (6.21)$$

La varianza mínima dada por (6.13) puede, después de algunos cálculos, escribirse como

$$V_{min} = \mathbb{V}_{min}(\widehat{Y}_U) = N^2\frac{S_{yU}^2}{n}\left(\frac{1-\nu r^2}{1-\nu^2 r^2} - f\right). \qquad (6.22)$$

Es sencillo en este caso obtener la proporción de emparejamiento óptima $\mu = 1 - \nu$. Diferenciando V_{min} con respecto a ν e igualando a cero, obtenemos las siguientes proporciones óptimas:

$$\nu_{opt} = \frac{1}{[1 + (1 - r^2)^{\frac{1}{2}}]} \tag{6.23}$$

y

$$\mu_{opt} = \frac{(1 - r^2)^{\frac{1}{2}}}{1 + (1 - r^2)^{\frac{1}{2}}}. \tag{6.24}$$

Si V_{opt} denota el valor de V_{min} cuando $\nu = \nu_{opt}$, tenemos

$$V_{opt} = \mathbb{V}_{opt}(\widehat{Y}_U) = N^2 \frac{S_{yU}^2}{n} \left(\frac{1 - \nu_{opt} r^2}{1 - \nu_{opt}^2 r^2} - f \right) \tag{6.25}$$

$$= N^2 \frac{S_{yU}^2}{n} \left(\frac{1}{2\nu_{opt}} - f \right) = N^2 \frac{S_{yU}^2}{n} \left[\frac{1 + (1 - r^2)^{\frac{1}{2}}}{2} - f \right]. \tag{6.26}$$

Como es natural, queremos conocer si el emparejamiento produce una ganancia en precisión. Si no hay emparejamiento y el estimador de Horvitz-Thompson se usa para el total actual, la varianza es

$$V_{srswor} \equiv \mathbb{V}(\widehat{Y}_U^{\mathrm{HT}}) = N^2 \frac{S_{yU}^2}{n}(1 - f)$$

suponiendo una muestra aleatoria simple sin reemplazamiento de n sobre N.

La reducción relativa de la varianza debida al emparejamiento (que es una medida de la ganancia en precisión) viene expresada por

$$1 - \frac{V_{opt}}{V_{srswor}} = 1 - \frac{\left(\frac{1}{2\nu_{opt}} \right) - f}{1 - f} = \frac{1 - \left(\frac{1}{2\nu_{opt}} \right)}{1 - f}.$$

La tabla 6.1 muestra esta reducción relativa de la varianza, así como la proporción óptima de emparejamiento, para valores seleccionados de r^2 y de $f = \frac{n}{N}$. Para los valores de r^2 en la tabla 6.1 la proporción óptima de emparejamiento nunca excede aproximadamente un $40\,\%$ y decrece significativamente para valores de r^2 cercanos a la unidad. Para fracciones muestrales f del $10\,\%$ o menos, la reducción en la varianza se queda aproximadamente en un rango entre $15\,\%$ y $50\,\%$. Para valores mayores de f, la reducción puede ser mucho mayor. ■

Comentario 29. Es interesante mencionar (pero difícil de verificar) que una derivación alternativa del estimador óptimo del total actual se obtiene mediante el siguiente argumento. Empezamos con una combinación lineal de cuatro estimadores insesgados de los totales poblacionales,

$$\widehat{Y}_U = \alpha \widehat{Z}_U^{\mathrm{HT}}(s_a) + \beta \widehat{Z}_U^{\mathrm{HT}}(s_m) + \gamma \widehat{Y}_U^{\mathrm{HT}}(s_m) + \delta \widehat{Y}_U^{\mathrm{HT}}(s_u) \tag{6.27}$$

		$100\left[1-\frac{V_{opt}}{V(\widehat{T}_\pi)}\right]$ para $f=$				
r^2	Proporción óptima de emparejamiento, %	0,4	0,2	0,1	0,01	0
0,5	41	24	18	16	15	15
0,6	39	31	23	20	19	18
0,7	35	38	28	25	23	23
0,8	31	46	35	31	28	28
0,9	24	57	43	38	35	34
0,95	18	65	49	43	39	39
0,99	9	75	56	50	45	45
0,999	3	81	61	54	49	48

Tabla 6.1: Reducción relativa de la varianza debido al emparejamiento y proporción óptima de emparejamiento para determinados valores de r^2 y de $f=\frac{n}{N}$.

donde α, β, γ y δ son constantes a determinar. Si el estimador (6.27) tiene que ser insesgado para el total actual Y_U debemos tener $\beta=-\alpha$ y $\delta=1-\gamma$, es decir,

$$\widehat{Y}_U=\alpha(\widehat{Z}_U^{\mathrm{HT}}(s_a)-\widehat{Z}_U^{\mathrm{HT}}(s_m))+\gamma(\widehat{Y}_U^{\mathrm{HT}}(s_m)-\widehat{Y}_U^{\mathrm{HT}}(s_u))+\widehat{Y}_U^{\mathrm{HT}}(s_u). \tag{6.28}$$

El siguiente paso es calcular la varianza $\mathbb{V}(\widehat{T}_y)$ y las ecuaciones $\frac{\partial V(\widehat{T}_y)}{\partial\alpha}=0$ y $\frac{\partial V(\widehat{T}_y)}{\partial\gamma}=0$ conducen a los valores óptimos

$$\alpha_{opt}=\gamma_{opt}\frac{\mathbb{E}_a\mathbb{C}(\widehat{Y}_U^{\mathrm{HT}}(s_m),\widehat{Z}_U^{\mathrm{HT}}(s_m)\big|s_a)}{\mathbb{E}_a\mathbb{V}(\widehat{Z}_U^{\mathrm{HT}}(s_m)\big|s_a)} \tag{6.29}$$

y

$$\gamma_{opt}=\frac{\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}}(s_u))-\mathbb{C}(\widehat{Y}_U^{\mathrm{HT}}(s_a),\widehat{Y}_U^{\mathrm{HT}}(s_a^c))}{\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}}(s_m)-\widehat{Y}_U^{\mathrm{HT}}(s_u))-\frac{\left[\mathbb{E}_a\mathbb{C}(\widehat{Y}_U^{\mathrm{HT}}(s_m),\widehat{Z}_U^{\mathrm{HT}}(s_m)\big|s_a)\right]^2}{\mathbb{E}_a\mathbb{V}\left(\widehat{Z}_U^{\mathrm{HT}}(s_m)\big|s_a\right)}} \tag{6.30}$$

donde

$$\widehat{Y}_U^{\mathrm{HT}}(s_a)=\sum_{k\in s_a}\frac{y_k}{\pi_{ak}}\quad\text{y}\quad\widehat{Y}_U^{\mathrm{HT}}(s_a^c)=\sum_{k\in s_a^c}\frac{y_k}{\pi_{ak}^c}.$$

En caso de muestreo aleatorio simple sin reemplazamiento tenemos

$$\alpha_{opt}=r\frac{S_{yU}}{S_{zU}}\frac{\mu}{1-\nu^2r^2}\quad\text{y}\quad\gamma_{opt}=\frac{\mu}{1-\nu^2r^2}.$$

■

Si no se puede especificar un buen valor de K a priori en el estimador diferencia en la ecuación (6.1), se puede usar en su lugar. Supongamos que el modelo

$$\begin{cases}\mathbb{E}_\xi(y_k)=\alpha+\beta z_k\\ \mathbb{V}_\xi(y_k)=\sigma^2\end{cases}$$

para $k \in U$ es una buena descripción de la relación entre y y z.

A partir de la muestra coincidente construimos el estimador de regresión

$$\widehat{Y}_{1r} = \widehat{Y}_U^{\text{HT}}(s_m) + \widehat{A}(\widehat{N}_U^{\text{HT}}(s_a) - \widehat{N}_U^{\text{HT}}(s_m)) + \widehat{B}(\widehat{Z}_U^{\text{HT}}(s_a) - \widehat{Z}_U^{\text{HT}}(s_m)) \tag{6.31}$$

donde

$$\widehat{N}_U^{\text{HT}}(s_a) = \sum_{k \in s_a} \frac{1}{\pi_{ak}}, \quad \widehat{N}_U^{\text{HT}}(s_m) = \sum_{k \in s_m} \frac{1}{\pi_{ak}\pi_{k|s_a}}, \quad \widehat{A} = \tilde{y}_{s_m} - \widehat{B}\tilde{z}_{s_m}$$

y

$$\widehat{B} = \frac{\sum_{k \in s_m} \frac{(z_k - \tilde{z}_{s_m})(y_k - \tilde{y}_{s_m})}{\pi_{ak}\pi_{k|s_a}}}{\sum_{k \in s_m} \frac{(z_k - \tilde{z}_{s_m})^2}{\pi_{ak}\pi_{k|s_a}}}$$

con

$$\tilde{z}_{s_m} = \frac{\widehat{Z}_U^{\text{HT}}(s_m)}{\widehat{N}_U^{\text{HT}}(s_m)} \quad \text{e} \quad \tilde{y}_{s_m} = \frac{\widehat{Y}_U^{\text{HT}}(s_m)}{\widehat{N}_U^{\text{HT}}(s_m)}.$$

$\widehat{Y}_{1r}$ es aproximadamente insesgado para Y_U y su varianza aproximada viene dada por

$$AV_1 = \sum_{k \in U} \sum_{l \in U} \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + \mathbb{E}_a \left\{ \sum_{k \in s_a} \sum_{l \in s_a} \Delta_{kl|s_a} \frac{E_k}{\pi_{ak}\pi_{k|s_a}} \frac{E_l}{\pi_{al}\pi_{l|s_a}} \right\}, \tag{6.32}$$

donde

$$E_k = y_k - \tilde{y}_{s_a} - \widehat{B}_{s_a}(z_k - \tilde{z}_{s_a}), \quad \tilde{y}_{s_a} = \frac{\widehat{Y}_U^{\text{HT}}(s_a)}{\widehat{N}_U^{\text{HT}}(s_a)}, \quad \tilde{z}_{s_a} = \frac{\widehat{Z}_U^{\text{HT}}(s_a)}{\widehat{N}_U^{\text{HT}}(s_a)}$$

y

$$\widehat{B}_{s_a} = \frac{\sum_{k \in s_a} \frac{(z_k - \tilde{z}_{s_a})(y_k - \tilde{y}_{s_a})}{\pi_{ak}}}{\sum_{k \in s_a} \frac{(z_k - \tilde{z}_{s_a})^2}{\pi_{ak}}}.$$

La muestra no aparejada se utiliza como antes. Es decir, mantenemos el estimador insesgado $\widehat{Y}_{U_2} = \widehat{Y}_U^{\text{HT}}(s_u)$ dado por (6.2). La combinación lineal de $\widehat{Y}_{1r}$ y $\widehat{Y}_{U_2}$ conduce a

$$\widehat{Y}_r = w_1 \widehat{Y}_{1r} + w_2 \widehat{Y}_{U_2}, \tag{6.33}$$

donde $w_1 + w_2 = 1$.

Este nuevo estimador es aproximadamente insesgado con varianza aproximada

$$AV(\widehat{Y}_r) = w_1^2 AV_1 + w_2^2 V_2 + 2w_1 w_2 AC \tag{6.34}$$

donde las expresiones de AV_1, $V_2 = \mathbb{V}(\widehat{Y}_{U_2})$ y $AC = AC(\widehat{Y}_{1r}, \widehat{Y}_{U_2}) = C$ vienen dadas por las ecuaciones (6.32), (6.6) y (6.7), respectivamente.

Una derivación análoga a (6.11) muestra que la varianza aproximada en (6.34) está minimizada por

$$w_1 = 1 - w_2 = \frac{V_2 - AC}{AV_1 + V_2 - 2AC}, \tag{6.35}$$

que nos proporciona una varianza mínima aproximada

$$AV_{min}(\widehat{T}_{yr}) = \frac{(AV_1)V_2 - (AC)^2}{AV_1 + V_2 - 2AC} \tag{6.36}$$

Ejemplo 23. Supongamos que se usa un muestreo aleatorio simple sin reemplazamiento como en el Ejemplo 21 y en el Ejemplo 22. Entonces $\widehat{Y}_{1r}$ en el estimador (6.33) viene dado por

$$\widehat{Y}_{1r} = N[\bar{y}_{s_m} + \widehat{B}(\bar{z}_{s_a} - \bar{z}_{s_m})],$$

donde

$$\widehat{B} = \frac{S_{zys_m}}{S^2_{zs_m}}$$

y $AV_1 = V_1$, V_2 y $AC = C$ vienen dadas por (6.19) a (6.21). Esto lleva a

$$AV(\widehat{Y}_{yr})_{min} = N^2 \frac{S^2_{yU}}{n} \left[\frac{1 - \nu r^2}{1 - \nu^2 r^2} - f \right]. \tag{6.37}$$

■

La varianza aproximada es la misma que la varianza mostrada en la ecuación (6.22). Por tanto, la proporción de emparejamiento óptima viene dada en este caso, también, por (6.24). La tabla 6.1 muestra la reducción de la varianza.

6.2.2 Estimación del total previo

Los datos recogidos de la muestra en la segunda ocasión también se pueden usar para mejorar el estimador original del total previo Z_U.

En lugar del estimador de Horvitz-Thompson $\widehat{Z}_U^{\text{HT}}$, consideramos

$$\widehat{Z}_U = \alpha \widehat{Z}_U^{\text{HT}}(s_a) + \beta \widehat{Z}_U^{\text{HT}}(s_m) + \gamma \widehat{Y}_U^{\text{HT}}(s_m) + \delta \widehat{Y}_U^{\text{HT}}(s_u), \tag{6.38}$$

donde α, β, γ y δ son constantes que habrá que determinar.

Si queremos que el estimador (6.38) sea insesgado para el total previo Z_U, entonces debemos tener $\alpha = 1 - \beta$ y $\gamma = -\delta$, es decir,

$$\widehat{Z}_U = \widehat{Z}_U^{\mathrm{HT}}(s_a) + \beta(\widehat{Z}_U^{\mathrm{HT}}(s_m) - \widehat{Z}_U^{\mathrm{HT}}(s_a)) + \delta(\widehat{Y}_U^{\mathrm{HT}}(s_u) - \widehat{Y}_U^{\mathrm{HT}}(s_m)). \tag{6.39}$$

Los valores óptimos de β y δ se obtienen, respectivamente, igualando las derivadas parciales de $\mathbb{V}(\widehat{Z}_U^{\mathrm{HT}})$ respecto a β y δ a cero. Esto nos lleva a

$$\beta_{opt} = \delta_{opt} \frac{\mathbb{E}_a \mathbb{C}(\widehat{Y}_U^{\mathrm{HT}}(s_m), \widehat{Z}_U^{\mathrm{HT}}(s_m) \big| s_a)}{\mathbb{E}_a \mathbb{V}(\widehat{Z}_U^{\mathrm{HT}}(s_m) \big| s_a)} \tag{6.40}$$

y

$$\delta_{opt} = \frac{\mathbb{C}(\widehat{Z}_U^{\mathrm{HT}}(s_a), \widehat{Y}_U^{\mathrm{HT}}(s_a)) - \mathbb{C}(\widehat{Z}_U^{\mathrm{HT}}(s_a), \widehat{Y}_U^{\mathrm{HT}}(s_a^c))}{\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}}(s_m) - \widehat{Y}_U^{\mathrm{HT}}(s_u)) - \frac{\left[\mathbb{E}_a \mathbb{C}(\widehat{Y}_U^{\mathrm{HT}}(s_m), \widehat{Z}_U^{\mathrm{HT}}(s_m) | s_a)\right]^2}{\mathbb{E}_a \mathbb{V}(\widehat{Z}_U^{\mathrm{HT}}(s_m) | s_a)}} \tag{6.41}$$

Ejemplo 24. Usando muestreo aleatorio simple sin reemplazamiento como en el Ejemplo 21, obtenemos

$$\beta_{opt} = \frac{\mu\nu r^2}{1 - \nu^2 r^2} \quad \text{y} \quad \delta_{opt} = \frac{S_{zU}}{S_{yU}} \frac{\mu\nu r}{1 - \nu^2 r^2} \tag{6.42}$$

y

$$\mathbb{V}(\widehat{T}_z)_{min} = N^2 \frac{S_{zU}^2}{n} \left\{ \frac{1 - \nu r^2}{1 - \nu^2 r^2} - f \right\} \tag{6.43}$$

La expresión que figura entre paréntesis es la misma que en la ecuación (6.22). La proporción de emparejamiento es, por tanto, la misma que para la estimación del total actual. ■

6.3 Estimación del cambio absoluto

Para estimar el cambio absoluto $\Delta = Y_U - Z_U$, podemos usar dos estimadores

$$\widehat{\Delta}_1 = \widehat{Y}_{1r} - \widehat{Z}_U^{\mathrm{HT}}(s_a) \tag{6.44}$$

y

$$\widehat{\Delta}_2 = \widehat{Y}_2 - \widehat{Z}_U^{\mathrm{HT}}(s_a), \tag{6.45}$$

donde $\widehat{Y}_{1r}$ es el estimador de regresión del total en cada ocasión dado por (6.31) y donde $\widehat{Y}_2 = \widehat{Y}_U^{\mathrm{HT}}(s_u)$ viene dado por (6.2). La combinación lineal de $\widehat{\Delta}_1$ y $\widehat{\Delta}_2$ da

$$\widehat{\Delta} = \delta_1 \widehat{\Delta}_1 + \delta_2 \widehat{\Delta}_2 \tag{6.46}$$

donde $\delta_1 + \delta_2 = 1$. Esto es un estimador aproximadamente insesgado del cambio absoluto Δ.

Su varianza aproximada viene dada por

$$AV(\widehat{\Delta}) = \delta_1^2 AV(\widehat{\Delta}_1) + \delta_2^2 AV(\widehat{\Delta}_2) + 2AC(\widehat{\Delta}_1, \widehat{\Delta}_2). \tag{6.47}$$

El razonamiento usado anteriormente en la sección precedente nos lleva directamente a

$$\delta_{1opt} = 1 - \delta_{2opt} = \frac{\mathbb{V}(\widehat{\Delta}_2) - AC(\widehat{\Delta}_1, \widehat{\Delta}_2)}{AV(\widehat{\Delta}_1) + \mathbb{V}(\widehat{\Delta}_2) - 2AC(\widehat{\Delta}_1, \widehat{\Delta}_2)} \tag{6.48}$$

y así tenemos

$$AV(\widehat{\Delta})_{min} = \frac{AV(\widehat{\Delta}_1)\mathbb{V}(\widehat{\Delta}_2) - [AC(\widehat{\Delta}_1, \widehat{\Delta}_2)]^2}{AV(\widehat{\Delta}_1) + \mathbb{V}(\widehat{\Delta}_2) - 2AC(\widehat{\Delta}_1, \widehat{\Delta}_2)} \tag{6.49}$$

Ejemplo 25. De nuevo, consideramos el muestreo aleatorio simple sin reemplazamiento como en el Ejemplo 21. Entonces

$$\widehat{\Delta}_1 = N[\bar{y}_{s_m} + \widehat{B}(\bar{z}_{s_a} - \bar{z}_{s_m}) - \bar{z}_{s_a}] \quad \text{y} \quad \widehat{\Delta}_2 = N(\bar{y}_{s_u} + \bar{z}_{s_a})$$

y

$$AV(\widehat{\Delta_1}) = \frac{N^2}{n}\left[\frac{S_{yU}^2}{\mu}(1 - \mu f - \nu r^2) + (1-f)S_{zU}^2 - 2(1-f)S_{zyU}\right]$$

$$\mathbb{V}(\widehat{\Delta}_2) = \frac{N^2}{n}\left[\frac{1-\nu f}{\nu}S_{yU}^2 + (1-f)S_{zU}^2 + 2fS_{zyU}\right]$$

y

$$AC(\widehat{\Delta}_1, \widehat{\Delta}_2) = \frac{N^2}{n}[(1-f)(S_{zU}^2 - S_{zyU}) - f(S_{yU}^2 - S_{zyU})].$$

Una hipótesis que simplifica los cálculos y que a menudo es realista, es que $S_{yU}^2 = S_{zU}^2 = S^2$. En tal caso tendríamos

$$AV(\widehat{\Delta_1}) = \frac{N^2S^2}{n}\left[\frac{1}{\mu}(1 - \nu r^2) + 1 - 2r - 2f(1-r)\right]$$

$$\mathbb{V}(\widehat{\Delta}_2) = \frac{N^2S^2}{n}\left[\frac{1}{\nu} + 1 - 2f(1-r)\right]$$

y

$$AC(\widehat{\Delta}_1, \widehat{\Delta}_2) = \frac{N^2S^2}{n}\left[(1 - r - 2f(1-r))\right].$$

A partir de (6.49), tenemos

$$AV_{min}(\widehat{\Delta}) = 2\frac{N^2S^2}{n}(1-r)\left[\frac{1}{1-\mu r} - f\right], \tag{6.50}$$

que es una función creciente de μ para cada r tal que $0 < r < 1$. Por lo tanto, si $0 < r < 1$, la proporción óptima de emparejamiento es el $100\,\%$; es decir, la mejor política para estimar el cambio absoluto es el uso de la misma muestra en ambas ocasiones. Por contra, para estimar el nivel (el total de y, por ejemplo), la proporción óptima de emparejamiento raramente excede el $40\,\%$, como se puede ver en la tabla 6.1. ■

6.4 Estimación de la suma de totales

Para estimar la suma de totales, $T_U = Y_U + Z_U$, formamos los estimadores

$$\widehat{T}_1 = \widehat{Y}_{1r} + \widehat{Z}_U^{\mathrm{HT}}(s_a) \tag{6.51}$$

y

$$\widehat{Y}_2 = \widehat{Y}_2 - \widehat{Z}_U^{\mathrm{HT}}(s_a) \tag{6.52}$$

donde $\widehat{Y}_{1r}$ es el estimador de regresión del total actual Y_U dado por (6.31) y donde $\widehat{Y}_2 = \widehat{T}_{s_u}$ es el estimador de Horvitz-Thompson mostrado en la ecuación (6.2).

Se puede ver que la varianza aproximada mínima es

$$AV_{min} = 2\frac{N^2 S^2}{n}\frac{1+r}{1-\mu r} \tag{6.53}$$

si asumimos muestreo aleatorio simple sin reemplazamiento, $S_{yU}^2 = S_{zU}^2 = S^2$ y $f = 0$. Esto implica que la proporción óptima de emparejamiento es cero cuando $r > 0$. Es decir, la mejor política para estimar la suma de los dos totales es elegir una muestra totalmente nueva en la segunda ocasión.

Comentario 30. Un gran número de encuestas a gran escala están diseñadas para medir los cambios de la población a lo largo del tiempo. Ejemplos bien conocidos son las Encuestas de Población Activa, realizadas en muchos países de forma regular, mensual o trimestralmente. Este tipo de encuestas a menudo usan algún tipo de solapamiento de la muestra. El diseño y la estimación de este tipo de encuestas puede requerir métodos especiales, por ejemplo, el uso de análisis de series temporales combinado con herramientas de muestreo probabilístico. Para más información sobre este tipo de muestreo véase (Duncan y Kalton 1987; Binder e Hidiroglou 1988). ■

Bibliografía

Binder, D.A. y M.A. Hidiroglou (1988). *Sampling in time. In: P.R. Krishnaiah and C.R. Rao (eds), Handbook of Statistics, Vol 6*. Amsterdam: North-Holland, págs. 187-211.

Duncan, G.J. y G. Kalton (1987). "Issues of design and analysis of surveys across time". En: *International Statistical Review* 55, págs. 97-117.

Hidiroglou, M.A. y P. Lavallée (2009). *Sampling and Estimation in Business Surveys*. North-Holland, Amsterdam: Elsevier, págs. 441-470.

Patterson, H.D. (1950). "Sampling on seccessive occasions with partial replacement of units". En: *Journal of the Royal Statistical Society* 12, págs. 241-255.
Pfefferman, D. y C.R. Rao (2009). *Handbook of Statistics 29A*. North-Holland, Amsterdam: Elsevier.
Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.

Tema 7

Métodos indirectos de estimación de la varianza. Método de los grupos aleatorios. Método de las semimuestras equilibradas. Método Jackknife. Método Bootstrap.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía.

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

7.1 Introducción

La precisión de un estimador $\widehat{\theta}$ a menudo se discute en términos de la varianza $\mathbb{V}(\widehat{\theta})$. Normalmente el valor exacto de la varianza es desconocido, porque depende de cantidades poblacionales desconocidas. Después de obtener los datos de una encuesta, sin embargo, se puede calcular una estimación de la varianza. Cuando se difunden los resultados de una encuesta, es una buena práctica (véase, p.ej., el tema 16[1] del bloque "Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas"del grupo de materias comunes) proporcionar las estimaciones de la varianza $\widehat{\mathbb{V}}(\widehat{\theta})$ para los estimadores $\widehat{\theta}$ usados en la encuesta. A menudo se usa una estimación de la varianza para calcular un intervalo de confianza

$$\widehat{\theta} \pm \text{constante} \times [\widehat{\mathbb{V}}]^{\frac{1}{2}},$$

suponiendo que la distribución muestral de $\widehat{\theta}$ es aproximadamente normal.

[1]Tema 16. La calidad en la estadística oficial y el Código de Buenas Prácticas de las Estadísticas Europeas. El concepto de calidad en la estadística oficial. El Código de Buenas Prácticas de las Estadísticas Europeas. El marco de garantía de la calidad del Sistema Estadístico Europeo. La calidad en los productos y en los procesos estadísticos. Sistemas de evaluación global de la calidad: auditorías, autoevaluación y revisiones por homólogos en las oficinas de Estadística.

Este tema presenta algunos métodos especiales de estimación de la varianza, añadidos a los vistos en temas anteriores. Hasta ahora se han visto tres métodos generales de estimación de la varianza.

1. El tema 2[2] del bloque "Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas"del grupo de materias comunes considera el caso en que $\widehat{\theta}$ es un estimador HT. Se vieron dos estimadores insesgados de $\mathbb{V}(\widehat{\theta})$, el estimador de Horvitz-Thompson y el estimador de Yates-Grundy-Sen. Ambos estimadores pueden ser modificados y usados cuando $\widehat{\theta}$ es una función lineal de varios estimadores HT.
2. Cuando $\widehat{\theta}$ es una función no lineal de estimadores HT, no se puede encontrar normalmente una expresión exacta para la varianza $\mathbb{V}(\widehat{\theta})$. En el tema 3 de este mismo bloque, se usa la linealización de Taylor para obtener una varianza aproximada $AV(\widehat{\theta}) = \mathbb{V}\left[\widehat{\theta}_0\right]$. La Sección 3.4 soluciona la estimación de esta varianza aproximada.
3. El tema 4 de este mismo bloque presenta un estimador de la varianza residual ponderada para el estimador GREG. Este estimador de la varianza tiene ventajas sobre el estimador de linealización de Taylor.

Este tema presenta algunas técnicas aproximadas para la estimación de la varianza, incluyendo grupos aleatorios, semimuestras equilibradas, jackknife y bootstrap. Estas técnicas son capaces de manejar tanto diseños complejos de muestreo como estimadores complejos. Por tanto, se pueden usar en casos en los que los métodos (1) y (3) anteriores no son fácilmente aplicables. Puesto que normalmente requieren el uso extensivo de computación, a veces se denominan computacionalmente intensivos.

Estos estimadores de la varianza son complicados en teoría, porque es difícil obtener resultados generales sobre sus propiedades estadísticas, por ejemplo, su sesgo. Los pocos resultado teóricos de este tema se limitan a los casos en los que la varianza de un estimador HT es estimada. Estas técnicas se usan principalmente con estimadores más complejos. En estos casos, no se dispone de resultado teóricos y nuestro conocimiento sobre las propiedades estadísticas de los estimadores de la varianza se limitan a conclusiones obtenidas de estudios simulados o otras fuentes de evidencia empírica.

A lo largo de este tema, suponemos muestreo sin reemplazamiento.

Una explicación más detallada de estos métodos (y del método de linealización de Taylor) está dada extensamente por Wolter 2007. Este tema se basa en esa fuente y en el resumen incluido en (Särndal, Swensson y Wretman 1992, cap. 11). Para un breve

[2]Tema 2. Ideas básicas sobre estimación en muestreo probabilístico. Diseño muestral. Probabilidades de inclusión. La noción de estadístico. Indicadores de pertenencia a la muestra. Estimadores y sus propiedades básicas. El estimador Horvitz-Thompson (estimador π) y sus propiedades. Muestreo con reemplazamiento. Efecto de diseño. Intervalos de confianza.

resumen de las técnicas de estimación de la varianza, véase también (Rust 1985).

Un estimador de la varianza $\widehat{\mathbb{V}}(\widehat{\theta})$ debería de ser:

a. Insesgado, o aproximadamente, es decir, debería satisfacer $\mathbb{E}[\widehat{\mathbb{V}}(\widehat{\theta})] \approx \mathbb{V}(\widehat{\theta})$.

b. Estable, es decir, la varianza del estimador de la varianza debería de ser pequeño.

c. No negativo, es decir, tener siempre valores no negativos.

d. Producir intervalos de confianza que abarquen θ con una probabilidad aproximadamente igual al nivel de confianza indicado.

¿Hasta qué punto los estimadores de la varianza que se presentan en este tema cumplen los requisitos anteriores? Todos son no negativos. Cuando $\widehat{\theta}$ es un estimador HT, a menudo tiene sesgo positivo. Por tanto, sobreestimarán $\mathbb{V}(\widehat{\theta})$. Cuando $\widehat{\theta}$ es una función no lineal de los estimadores HT no existen resultados analíticos exactos. Sin embargo, a menudo se espera que estos estimadores de la varianza sigan funcionando bien en este caso. Para la estabilidad y los niveles de confianza, debe recurrirse a estudios con datos simulados.

Presentamos ahora un resultado que se usará varias veces más tarde en el tema. Supongamos que $\widehat{\theta}$ es un estimador de θ, basado en los datos de una muestra probabilística. A menudo consideraremos un número (digamos A) de subconjuntos de la muestra original y un estimador separado de θ calculado a partir de cada subconjunto,

$$\widehat{\theta}_1, \ldots, \widehat{\theta}_a, \ldots, \widehat{\theta}_A.$$

Se puede calcular un estimador alternativo con la muestra completa obteniendo la media de $\widehat{\theta}_a$,

$$\widehat{\theta}^* = \frac{1}{A}\sum_{a=1}^{A}\widehat{\theta}_a.$$

A veces $\widehat{\theta}^*$ y $\widehat{\theta}$ se pueden definir de forma que sean idénticos.

Nuestro mayor interés aquí es estimar la varianza de $\widehat{\theta}^*$, es decir, $\mathbb{V}(\widehat{\theta}^*)$. Consideraremos las dos siguientes expresiones,

$$\widehat{V}_1 = \frac{1}{A(A-1)}\sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta}^*)^2 \tag{7.1}$$

y

$$\widehat{V}_2 = \frac{1}{A(A-1)}\sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta})^2 \tag{7.2}$$

El uso de la expresión (7.1) como un estimador de $\mathbb{V}(\widehat{\theta})$ está justificada por el siguiente resultado

$$\mathbb{E}(\widehat{V}_1) = \mathbb{V}(\widehat{\theta}^*) - \frac{1}{A(A-1)} \sum_{\substack{a=1 \\ a \neq b}}^{A} \sum_{b=1}^{A} \mathbb{C}(\widehat{\theta}_a, \widehat{\theta}_b) + \frac{1}{A(A-1)} \sum_{a=1}^{A} [\mathbb{E}(\widehat{\theta}_a) - \mathbb{E}(\widehat{\theta}^*)]^2. \tag{7.3}$$

Se sigue de (7.3) que, si $\widehat{\theta}_1 \ldots, \widehat{\theta}_A$ estuviesen incorrelados y tuviesen la misma esperanza, entonces $\widehat{V}_1$ dado por (7.1) sería insesgado para $\mathbb{V}(\widehat{\theta}^*)$.

El uso de la expresión (7.2) como un estimador de $\mathbb{V}(\widehat{\theta}^*)$ está justificado por la igualdad

$$\sum_{a=1}^{A} (\widehat{\theta}_a - \widehat{\theta})^2 = \sum_{a=1}^{A} (\widehat{\theta}_a - \widehat{\theta}^*)^2 + A(\widehat{\theta}^* - \widehat{\theta})^2 \tag{7.4}$$

de lo que se sigue que

$$\widehat{V}_2 \geq \widehat{V}_1.$$

De hecho, tanto $\widehat{V}_2$ como $\widehat{V}_1$ también se usan para estimar la varianza de $\widehat{\theta}$, $\mathbb{V}(\widehat{\theta})$, bajo la hipótesis de que $\mathbb{V}(\widehat{\theta}^*)$ y $\mathbb{V}(\widehat{\theta})$ son aproximadamente iguales.

La mayoría de los estimadores aproximados de la varianza que veremos en este tema son de la forma mostrada en las ecuaciones (7.1) y (7.2). En los casos que consideraremos a continuación, $\widehat{\theta}_1, \ldots, \widehat{\theta}_A$ normalmente estarán correlados, y tanto $\widehat{V}_1$ como $\widehat{V}_2$ serán estimadores sesgados de $\mathbb{V}(\widehat{\theta}^*)$ y $\mathbb{V}(\widehat{\theta})$.

7.2 Método de los grupos aleatorios

7.2.1 Grupos aleatorios independientes

La técnica de grupos aleatorios independientes es una técnica especial para seleccionar una muestra que conduzca a una estimación de la varianza sencilla. La idea de los grupos aleatorios independientes tiene sus orígenes en trabajos de Mahalanobis 1939; Mahalanobis 1944; Mahalanobis 1946 y Deming 1956. La terminología en este área varía bastante. Muestras interpenetrantes (Mahalanobis) y muestras replicadas (Deming) son términos alternativos para lo que llamamos grupos aleatorios.

En lugar de seleccionar una única muestra s de tamaño n, seleccionamos A muestras independientes $s_1, \ldots, s_A$, de igual tamaño $m = \frac{n}{A}$. Asumimos que se selecciona la primera muestra s_1 y luego se devuelve a la población. Entonces se selecciona la segunda muestra s_2, muestreando sobre el total de la población, con el mismo diseño con el que se obtuvo s_1 y de forma independiente a s_1. Entonces s_2 se devuelve a la población,

una tercera muestra s_3 se selecciona con el mismo diseño que s_1 y s_2, y así sucesivamente, hasta que se seleccionan A muestras. El diseño común con el que se selecciona cada muestra s_a puede ser sin reemplazamiento. Lo que es importante, sin embargo, es que se devuelva cada muestra s_a antes de que la siguiente muestra s_{a+1} sea seleccionada.

Bajo estas condiciones, las muestras, o grupos aleatorios como también se llaman, $s_1, \ldots, s_A$, se pueden considerar resultados de A repeticiones independientes del mismo experimento aleatorio, concretamente, seleccionando m elementos de una población dada mediante un diseño de muestreo fijo. Las A muestras no tienen por qué ser necesariamente disjuntas. (Por el contrario, si las muestras no fuesen devueltas a la población, entonces serían disjuntas, pero ya no serían independientes).

Para cada $a = 1, \ldots, A$, se calcula un estimador $\widehat{\theta}_a$ de θ a partir únicamente de los datos de s_a. La misma fórmula del estimador se utiliza de principio a fin. La media de los $\widehat{\theta}_a$, aquí denotada por $\widehat{\theta}_{IRG}$ [3] es un candidato natural para estimar θ

$$\widehat{\theta}_{IRG} = \frac{1}{A}\sum_{a=1}^{A}\widehat{\theta}_a. \tag{7.5}$$

Como estimador de $\mathbb{V}(\widehat{\theta}_{IRG})$, la varianza de $\widehat{\theta}_{IRG}$, consideremos el estimador de la varianza de los grupos aleatorios independientes,

$$\widehat{V}_{IRG1} = \frac{1}{A(A-1)}\sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta}_{IRG})^2. \tag{7.6}$$

Bajo las hipótesis de grupos aleatorios independientes, es decir, las A muestras se seleccionan de forma independiente y con el mismo diseño muestral, los estimadores $\widehat{\theta}_1, \ldots, \widehat{\theta}_A$ son variables aleatorias independientes e idénticamente distribuidas. Se sigue inmediatamente a partir de (7.3) que $\widehat{V}_{IRG1}$ es insesgado,

$$\mathbb{E}(\widehat{V}_{IRG1}) = \mathbb{V}(\widehat{\theta}_{IRG}). \tag{7.7}$$

Comentario 31. Para que se verifique (7.7), no es necesario que los $\widehat{\theta}_a$ estén idénticamente distribuidos. Podemos ver de la ecuación (7.3) que (7.7) se verifica bajo la hipótesis más débil de que $\widehat{\theta}_1, \ldots, \widehat{\theta}_a, \ldots, \widehat{\theta}_A$ tengan la misma esperanza. Por tanto, no necesitan tener la misma varianza, ni necesitan ser insesgados para θ mientras sean independientes y tengan la misma esperanza. ■

El estimador de la varianza mostrado en la ecuación (7.6) es sorprendentemente fácil de calcular. Una vez que se han obtenido los valores $\widehat{\theta}_1, \ldots, \widehat{\theta}_a, \ldots, \widehat{\theta}_A$, es poco el trabajo que queda para llegar a $\widehat{V}_{IRG1}$. Otro detalle es que la insesgadez de (7.7) se verifica independientemente de la complejidad del estimador $\widehat{\theta}_a$ e independientemente de la

[3] Del inglés *Independent Random Groups*.

complejidad del diseño muestral usado para obtener s_a. Por ejemplo, $\widehat{\theta}_a$ puede ser un estimador del coeficiente de correlación, y el diseño puede ser uno que no permita la estimación de la varianza insesgada mediante técnicas tradicionales, como el muestreo sistemático con un arranque aleatorio simple.

Sin embargo, también existen ciertos problemas asociados con la técnica de grupos aleatorios independientes:

1. La selección y recogida de datos para una serie de muestras independientes puede resultar más costoso y complejo que seleccionar y observar una única muestra grande. Además, hay que ser cuidadoso de no crear un dependencia no deseada entre los $\widehat{\theta}_a$. Esta dependencia podría introducirse mediante el efecto entrevistador, o durante el procesamiento de los datos por el personal del instituto encargado de esta fase.
2. Para obtener un estimador de la varianza estable, el número A de muestras independientes debería de ser grande. En la práctica, sin embargo, A puede que no sea tan grande, lo que hace que el estimador de la varianza sea inestable. Por ejemplo, Mahalanobis propuso usar sólo cuatro grupos, mientras que Deming sugiere diez.

Esta deficiencias pueden explicar por qué la técnica de grupos aleatorios independientes no se usa a menudo en la práctica.

El estimador $\widehat{\theta}_{IRG}$ se obtiene como media de las estimaciones para las distintas muestras independientes $s_1, \ldots, s_A$. De forma alternativa, θ podría estimarse a partir de la muestra combinada. Aplicando la fórmula del estimador a los datos correspondientes a

$$s = \bigcup_{a=1}^{A} s_a$$

obtenemos un estimador de la muestra combinada, denotado por $\widehat{\theta}$. Claramente, $\widehat{\theta}$ puede diferir de $\widehat{\theta}_{IRG}$. Por ejemplo, una estimación de un coeficiente de correlación basada en la s combinada no es en general igual a la media de las estimaciones de los A coeficientes de correlación obtenidos de las distintos subconjuntos $s_1, \ldots, s_A$.

Una pregunta relacionada es: ¿cómo se debería de estimar la varianza de $\widehat{\theta}$? A menudo, $\widehat{V}_{IRG1}$ definida por (7.6) se usa de hecho para estimar $\mathbb{V}(\widehat{\theta})$ cuando $\widehat{\theta}$ y $\widehat{\theta}_{IRG}$ no son idénticos.

Un estimador alternativo de $\mathbb{V}(\widehat{\theta})$ que se usa a veces es

$$\widehat{V}_{IRG2} = \frac{1}{A(A-1)} \sum_{a=1}^{A} (\widehat{\theta}_a - \widehat{\theta})^2 \tag{7.8}$$

Se sabe poco sobre si $\widehat{V}_{IRG1}$ o $\widehat{V}_{IRG2}$ debería ser preferible como estimador de $\mathbb{V}(\widehat{\theta})$. Sin embargo, a partir de (7.4), se cumple

$$\widehat{V}_{IRG2} \geq \widehat{V}_{IRG1}. \tag{7.9}$$

Por tanto, $\widehat{V}_{IRG2}$ es más conservador que $\widehat{V}_{IRG1}$. Cuando se trata de estimar $\mathbb{V}(\widehat{\theta}_{IRG})$, se sigue a partir de (7.9) que $\widehat{V}_{IRG2}$ tendrá sesgo positivo.

7.2.2 Grupos aleatorios dependientes

La técnica de grupos aleatorios dependientes es un intento de adaptar la técnica de grupos aleatorios independientes a una muestra que no verifica los requisitos de grupos aleatorios independientes. Supongamos que primero seleccionamos una muestra grande del total poblacional mediante un diseño de muestreo probabilístico. Después de que esta muestra se ha obtenido, se usa un mecanismo aleatorio para dividir la muestra en un número de submuestras disjuntas, esto es, en grupos aleatorios. Éstos no serán independientes, pero una característica importante del procedimiento a describir es que serán tratados como si fuesen independientes. Estas ideas fueron descritas en primer lugar por Hansen, Hurwitz y Madow 1953.

Sea s la muestra seleccionada de la población U; la llamaremos la muestra completa. A continuación se divide s en A grupos aleatorios disjuntos $s_1, \ldots, s_a \ldots, s_A$, es decir,

$$s = \bigcup_{a=1}^{A} s_a.$$

Sea s de tamaño fijo n y supongamos por simplicidad que los grupos son de igual tamaño, $m = \frac{n}{A}$. Suponemos que s se divide en en grupos mediante una estrategia de aleatorización de forma que "cada grupo aleatorio tiene esencialmente el mismo diseño muestral que la muestra de la que procede" (Wolter 2007). No siempre es obvio cómo se puede realizar. En algunos casos hay varias formas posibles, como se verá más tarde en esta subsección.

Sean $\widehat{\theta}_1, \ldots, \widehat{\theta}_a, \ldots, \widehat{\theta}_A$ estimadores de θ, donde $\widehat{\theta}_a$ está basado en los datos del grupo s_a únicamente; $a = 1, \ldots, A$. Se asume que los $\widehat{\theta}_a$ son insesgados, o casi, pero no necesitan ser estimadores HT. Podemos considerar dos estimadores de θ de muestra completa alternativos: (a) el estimador $\widehat{\theta}_{\text{DRG}}$ [4] obtenido mediante una media,

$$\widehat{\theta}_{\text{DRG}} = \frac{1}{A} \sum_{a=1}^{A} \widehat{\theta}_a \tag{7.10}$$

y (b) el estimador $\widehat{\theta}$ basado en los datos de la muestra completa s, ignorando la división en grupos aleatorios. En algunos casos $\widehat{\theta}_{\text{DRG}}$ y $\widehat{\theta}$ se pueden definir de forma que sean

[4] DRG viene del inglés *dependent random groups*

idénticos. Como antes en el tema, podemos formar dos estimadores de la varianza alternativos

$$\widehat{V}_{\mathrm{DRG1}} = \frac{1}{A(A-1)} \sum_{a=1}^{A} (\widehat{\theta}_a - \widehat{\theta}_{\mathrm{DRG}})^2 \tag{7.11}$$

y

$$\widehat{V}_{\mathrm{DRG2}} = \frac{1}{A(A-1)} \sum_{a=1}^{A} (\widehat{\theta}_a - \widehat{\theta})^2. \tag{7.12}$$

Ambos son posibles estimadores de $\mathbb{V}(\widehat{\theta}_{\mathrm{DRG}})$ así como de $\mathbb{V}(\widehat{\theta})$, pero ninguno es insesgado para cualquiera de estos dos objetivos.

Algunas conclusiones inmediatas son:

1. Se sigue de (7.3) que si los grupos aleatorios se forman tal que $\widehat{\theta}_1, \ldots, \widehat{\theta}_A$ tienen la misma esperanza, no necesariamente igual a θ, entonces el sesgo de $\widehat{V}_{\mathrm{DRG1}}$ como estimador de $\mathbb{V}(\widehat{\theta}_{\mathrm{DRG}})$ viene dado por

$$\mathbb{E}(\widehat{V}_{\mathrm{DRG1}}) - \mathbb{V}(\widehat{\theta}_{\mathrm{DRG}}) = -\frac{1}{A(A-1)} \sum_{\substack{a=1 \\ a \neq b}}^{A} \sum_{b=1}^{A} \mathbb{C}(\widehat{\theta}_a - \widehat{\theta}_b)$$

   En particular, si todos los pares $\widehat{\theta}_a$ y $\widehat{\theta}_b$ tienen la misma covarianza γ, entonces

$$\mathbb{E}(\widehat{V}_{\mathrm{DRG1}}) - \mathbb{V}(\widehat{\theta}_{\mathrm{DRG}}) = -\gamma \tag{7.13}$$

2. Se sigue de (7.4) que $\widehat{V}_{\mathrm{DRG2}} \geq \widehat{V}_{\mathrm{DRG1}}$.

Ejemplo 26. Sea una muestra s de tamaño $n = Am$ extraída de una población U de tamaño N mediante un diseño muestral aleatorio simple sin reemplazamiento. Divídase a continuación la muestra s de modo aleatorio en A grupos aleatorios disjuntos $s_1, \ldots, s_A$ de igual tamaño m, de tal modo que cada s_a es una muestra aleatoria simple sin reemplazamiento de la población U[5]. Supongamos que quiere estimarse el total poblacional $Y_U = \sum_{k \in U} y_k$. Sea $\widehat{Y}_U^{\mathrm{HT}} = N\bar{y}_s$ e $\widehat{Y}_{aU}^{\mathrm{HT}} = N\bar{y}_{s_a}$. En este caso, $\widehat{Y}_U^{\mathrm{HT}}$ y $\widehat{Y}_{\mathrm{DRG}}$ son idénticos:

$$\widehat{Y}_U^{\mathrm{HT}} = \widehat{Y}_{\mathrm{DRG}} = N\bar{y}_U,$$

con varianza

$$\mathbb{V}\left[\widehat{Y}_U^{\mathrm{HT}}\right] = N^2 (1-f) \frac{S_{yU}^2}{n}.$$

El estimador de la varianza mediante grupos aleatorios es, por tanto,

[5]Esto puede conseguirse extrayendo s_1 como una muestra aleatoria simple sin reemplazamiento de tamaño m de s. A continuación, se extrae otra muestra aleatoria simple sin reemplazamiento de tamaño m del resto de la muestra $s - s_1$. Seguidamente, se extrae igualmente s_3 de $s - (s_1 \cup s_2)$ y así sucesivamente.

$$\widehat{V}_{\mathrm{DRG}} \equiv \widehat{V}_{\mathrm{DRG1}} = \widehat{V}_{\mathrm{DRG2}} = N^2 \frac{1}{A(A-1)} \sum_{a=1}^{A} \left(\bar{y}_{s_a} - \bar{y}_s\right)^2 .$$

Ahora bien, como para $a \neq b$ se cumple $\mathbb{C}\left(\bar{y}_{s_a}, \bar{y}_{s_b}\right) = \mathbb{V}\left(\bar{y}_s\right) - \frac{S_{yU}^2}{n}$, se sigue de (7.13) que el sesgo relativo de $\widehat{V}_{\mathrm{DRG}}$ está dado por

$$\frac{\mathbb{E}\left(\widehat{V}_{\mathrm{DRG}}\right) - \mathbb{V}\left(N\bar{y}_s\right)}{\mathbb{V}\left(N\bar{y}_s\right)} = \frac{f}{1-f},$$

que es irrelevante cuando la fracción de muestreo f es pequeña y, en el caso extremo cuando $A = n$ (esto es, $m = 1$), entonces $\widehat{V}_{\mathrm{DRG}} = N^2 \frac{S_{ys}^2}{n}$ y el sesgo podría corregirse tan solo con el factor $1 - f$. ■

Ejemplo 27. Considérese el muestreo proporcional al tamaño sin reemplazamiento para la estimación del total poblacional $Y_U = \sum_{k \in U} y_k$. Una muestra s de tamaño fijo $n = Am$ se selecciona proporcionalmente al tamaño con probabilidades de inclusión π_k, $k \in U$. La muestra extraída s se divide en A grupos aleatorios $s_1, \ldots, s_A$ de igual tamaño m, como en el ejemplo anterior. Por tanto, cada grupo s_a se comporta como si hubiese sido seleccionada con un diseño muestral con probabilidades de inclusión $\pi_{ka} = \frac{m}{n}\pi_k$, $k \in U$. Los estimadores HT basados en s y s_a serán, por tanto,

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k}, \tag{7.14}$$

$$\widehat{Y}_{aU}^{\mathrm{HT}} = \sum_{k \in s_a} \frac{y_k}{\pi_{ka}}. \tag{7.15}$$

Entonces, será

$$\widehat{Y}_{\mathrm{DRG}} = \frac{1}{A} \sum_{a=1}^{A} \widehat{Y}_{aU}^{\mathrm{HT}} = \widehat{Y}_U^{\mathrm{HT}},$$

cuyas varianzas son por tanto idénticas y se reducen a

$$\widehat{V}_{\mathrm{DRG}} \equiv \widehat{V}_{\mathrm{DRG1}} = \widehat{V}_{\mathrm{DRG2}} = \frac{1}{A(A-1)} \sum_{a=1}^{A} \left(\widehat{Y}_{aU}^{\mathrm{HT}} - \widehat{Y}_U^{\mathrm{HT}}\right)^2 . \tag{7.16}$$

No es difícil demostrar que el sesgo de $\widehat{V}_{\mathrm{DRG}}$ es

$$\mathbb{E}\left[\widehat{V}_{\mathrm{DRG}}\right] - \mathbb{V}\left[\widehat{Y}_U^{\mathrm{HT}}\right] = \frac{n}{n-1}\left[V_0 - \mathbb{V}\left[\widehat{Y}_U^{\mathrm{HT}}\right]\right],$$

donde $V_0 \equiv \frac{1}{n}\sum_{k \in U} \frac{\pi_k}{n}\left[\frac{ny_k}{\pi_k} - Y_U\right]^2$. El ejemplo anterior es, de hecho, un caso particular de este ejemplo. ■

Ejemplo 28. Cuando la muestra s se obtiene mediante un muestreo estratificado, existen dos maneras fundamentalmente diferentes de aplicar la técnica de los grupos aleatorios dependientes, que conducen a estimadores de la varianza diferentes (sin una preferencia clara por ninguna de ellos). Por simplificar, estimemos el total poblacional $Y_U = \sum_{h=1}^{H} Y_{U_h}$ mediante el estimador HT dado por $\widehat{Y}_U^{\text{strHT}} = \sum_{h=1}^{H} \widehat{Y}_{U_h}^{\text{strHT}}$. El primer enfoque es centrarse en la varianza $\mathbb{V}\left(\widehat{Y}_U^{\text{strHT}}\right) = \sum_{h=1}^{H} \mathbb{V}\left(\widehat{Y}_{U_h}^{\text{HT}}\right)$ y aplicar la técnica de los grupos aleatorios dependientes a cada estrato h por separado, de modo que puede escribirse

$$\widehat{\mathbb{V}}_{\text{DRG}}\left(\widehat{Y}_U^{\text{HT}}\right) = \sum_{h=1}^{H} \widehat{\mathbb{V}}_{\text{DRG}}\left(\widehat{Y}_{U_h}^{\text{HT}}\right).$$

El segundo enfoque consiste en formar los grupos aleatorios transversalmente a todos los estratos, de modo que cada s_a es en realidad una muestra estratificada de la población U. Si la muestra completa s tiene n_h elementos del estrato h, entonces el grupo aleatorio s_a tendrá $m_h = \frac{n_h}{A}$ elementos del estrato h. Tenemos así A grupos aleatorios dependientes de igual tamaño $m = \sum_{h=1}^{H} m_h$ con una afijación por estrato idéntica a la de la muestra completa. Como estimador de la varianza de $\widehat{Y}_U^{\text{strHT}}$ podemos usar $\widehat{V}_{\text{DRG1}}$ o $\widehat{V}_{\text{DRG2}}$. No existe evidencia clara sobre qué opción es preferible (Wolter 2007). ■

Ejemplo 29. Cuando el diseño muestral incluye dos o más etapas de muestreo, una práctica extendida es aplicar la técnica de los grupos aleatorios a las PSUs únicamente. Así, la muestras s_I de PSUs se divide en grupos aleatorios s_{Ia} ($a = 1, \ldots, A$). Suponemos que estos grupos de PSUs se forman del mismo modo que los grupos de elementos en los ejemplos previos. Considérese, por ejemplo, un muestreo en primera etapa proporcional al tamaño sin reemplazamiento. Sea el estimador para el total poblacional $Y_U = \sum_{k \in U} y_k$ dado por $\widehat{Y}_U^{\text{HT}} = \sum_{i \in s_I} \frac{\widehat{Y}_{U_i}^{\text{HT}}}{\pi_{Ii}}$, donde π_{Ii} denota la probabilidad de inclusión en primera etapa del conglomerado i. El diseño muestral en etapas posteriores es arbitrario. La muestra en primera etapa s_I de n_I conglomerados se divide aleatoriamente en grupos aleatorios $s_{I1}, \ldots, s_{IA}$ de igual tamaño $m_I = n_I/A$. Se sigue que, para una muestra s_I dada, el grupo s_{Ia} se comporta como una muestra aleatoria simple de s_I. El estimador de Y_U basado en s_{Ia} será, por tanto,

$$\widehat{Y}_{aU}^{\text{HT}} = \frac{n_I}{m_I} \sum_{k \in s_{Ia}} \frac{\widehat{Y}_{U_i}^{\text{HT}}}{\pi_{Ii}}.$$

Nuevamente, $\widehat{Y}_U^{\text{HT}}$ e $\widehat{Y}_{\text{DRG}} = \frac{1}{A} \sum_{a=1}^{A} \widehat{Y}_{aU}^{\text{HT}}$ coinciden y la estimación de su varianza común será

$$\widehat{V}_{\text{DRG}} = \frac{1}{A(A-1)} \sum_{a=1}^{A} \left(\widehat{Y}_{aU}^{\text{HT}} - \widehat{Y}_U^{\text{HT}}\right)^2.$$

■

Comentario 32. Los estimadores de la varianza (7.11) y (7.12) pueden sufrir inestabilidad en el sentido de que están basados en un pequeño número de grupos. Norlèn

y Waller 1979 sugirieron un procedimiento mejorado, concretamente, repetir el procedimiento completo de grupos aleatorios independientemente un número de veces, por ejemplo R veces, sobre la misma muestra s. De esta forma, obtenemos una secuencia de R estimadores de la varianza independientes $\widehat{V}_1, \ldots, \widehat{V}_r, \ldots, \widehat{V}_R$, donde cada $\widehat{V}_r$ es un estimador de la varianza de grupos aleatorios dependientes. Calculando la media da un estimador de la varianza

$$\widehat{V} = \frac{1}{R} \sum_{r=1}^{R} \widehat{V}_r.$$

Este procedimiento de grupos aleatorios repetido proporcionará obviamente un estimador de la varianza con variabilidad reducida. Sea

$$\mathbb{V}\left(\widehat{V}_r\right) = c_1 + c_2,$$

donde $c_1 = \mathbb{E}\left[\mathbb{V}\left(\widehat{V}_r|s\right)\right]$ y $c_2 = \mathbb{V}\left[\mathbb{E}\left(\widehat{V}_r|s\right)\right]$. Entonces

$$\mathbb{V}\left(\widehat{V}\right) = \left(\frac{c_1}{R}\right) + c_2 < c_1 + c_2.$$

Cabe señalar que los grupos aleatorio repetidos están formados a partir de una única misma muestra, por eso el único coste adicional, normalmente pequeño, se debe al esfuerzo computacional mayor.

7.3 Método de las semimuestras equilibradas

El método de semimuestras equilibradas se desarrolló inicialmente para el caso de un gran número de estratos y una muestra compuesta únicamente por dos elementos (o dos PSUs) por estrato. En este caso, la técnica de grupos aleatorios dependientes tendrá una estabilidad pobre, porque solo se pueden formar dos grupos disjuntos. La idea de usar semimuestras para la estimación de la varianza fue introducida por la Oficina de Censos de EE.UU. en torno a 1960. La técnica de semimuestras equilibradas que describiremos fue desarrollada por McCarthy 1966; McCarthy 1969. Otros nombres para esta técnica son replicaciones repetidas balanceadas (BRR del inglés *balanced repeated replications*) y pseudoreplicación.

Describimos la técnica tal y como se usa en el caso de la estimación de la varianza de un estimador HT. Se han explicado varios métodos de estimación de la varianza de un estimador HT. Sin embargo, el estimador HT ilustra bien la forma en que funciona esta técnica. El caso en que $\widehat{\theta}$ es un estimador más complejo es de más interés en la práctica y se verá al final de la sección.

Sea una población U de N elementos dividida en H estratos $U_1, \ldots, U_H$ de tamaños $N_1, \ldots, N_H$ respectivamente. Sea s_h una muestra de tamaño fijo $n_h = 2$ seleccionada

independientemente en cada estrato ($h = 1, \ldots, H$) mediante un diseño con probabilidades de inclusión π_k y π_{kl}. El total poblacional se puede escribir como

$$Y_U = \sum_{t=1}^{H} Y_{U_h} = \sum_{h=1}^{H} \sum_{k \in U_h} y_k.$$

El correspondiente estimador HT viene dado por

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{h=1}^{H} \widehat{Y}_{U_h}^{\mathrm{HT}} = \sum_{t=1}^{H} \sum_{k \in s_h} \frac{y_k}{\pi_k}. \tag{7.17}$$

Un estimador simplificado de la varianza para el caso del muestreo estratificado es

$$\widehat{V}_0 = \sum_{t=1}^{H} \frac{1}{n_h(n_h-1)} \sum_{k \in s_h} \left(\frac{y_k}{p_k} - \widehat{Y}_{U_h}^{\mathrm{HT}} \right)^2,$$

donde $p_k = \pi_k / n$ y, en este caso (con $n_h = 2$), queda

$$\widehat{V}_0 = \frac{1}{2} \sum_{h=1}^{H} \sum_{k \in s_h} \left(\frac{y_k}{p_k} - \widehat{Y}_{U_h}^{\mathrm{HT}} \right)^2. \tag{7.18}$$

La técnica de semimuestras equilibradas ofrece una forma alternativa de calcular el estimador de la varianza (7.18). Consideremos una muestra dada $s = \bigcup_{h=1}^{H} s_h$, donde cada s_h consiste de exactamente dos elementos. Una semimuestra es un conjunto que consiste de exactamente uno de los dos elementos de cada s_h. Hay, por tanto, 2^H posibles semimuestras.

La idea básica de la técnica de semimuestras equilibradas es seleccionar un conjunto de semimuestras del conjunto de todas las 2^H semimuestras, calcular una estimación $\widehat{Y}_{aU}^{\mathrm{HT}}$ de Y_U para cada semimuestra seleccionada y entonces usar estos valores $\widehat{Y}_{aU}^{\mathrm{HT}}$ para estimar $\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}})$). Veremos que la selección de las semimuestras de una forma ingeniosa (concretamente como un 'conjunto balanceado') hace posible calcular $\widehat{V}_0$ dado por (7.18) a partir de los valores observados $\widehat{Y}_{aU}^{\mathrm{HT}}$.

Necesitamos notación adicional para describir la técnica. Para cada s_h, sean sus dos elementos renombrados de forma temporal (aleatoriamente) como $h1$ y $h2$. De esta forma, una semimuestra consiste únicamente en uno solo de los dos elementos $h1$ y $h2$ para cada estrato h. Para cada semimuestra posible ($a = 1, \ldots, 2^H$), definimos además variables indicadoras δ_{ah} y ε_{ah} ($h = 1, \ldots, H$) tales que

$$\delta_{ah} = \begin{cases} 1 & \text{si la semimuestra } a \text{ contiene al elemento } h_1, \\ 0 & \text{si la semimuestra } a \text{ contiene al elemento } h_2, \end{cases}$$

y

$$\varepsilon_{ah} = \begin{cases} 1 & \text{si la semimuestra } a \text{ contiene al elemento } h_1, \\ -1 & \text{si la semimuestra } a \text{ contiene al elemento } h_2. \end{cases}$$

Por tanto, cada semimuestra posible ($a = 1, \ldots, 2^H$) está caracterizada por el vector $(\delta_{a1}, \ldots, \delta_{aH})$, o, alternativamente, por el vector $(\varepsilon a1, \ldots, \varepsilon_{aH})$. Claramente, $\varepsilon_{ah} = 2\delta_{ah} - 1$.

Definición 9

Un conjunto de A semimuestras (etiquetadas $a = 1, \ldots, A$) se dice que está *equilibrada* si

$$\sum_{a=1}^{A} \varepsilon_{ah}\varepsilon_{ah'} = 0 \tag{7.19}$$

para todos los pares de estratos h y h' ($h \neq h'$). Un conjunto de semimuestras que es equilibrada y también verifica

$$\sum_{a=1}^{A} \varepsilon_{ah} = 0 \tag{7.20}$$

para cada $h = 1, \ldots, H$ se dice que es *equilibrada ortogonal completa* o está en *equilibrio ortogonal completo*.

La técnica de semimuestras equilibradas de estimación de la varianza se puede ahora describir de la siguiente forma.

i. Obtener un conjunto equilibrado de semimuestras s_a ($a = 1, \ldots, A$) a partir de s.

ii. Para cada semimuestra ($a = 1, \ldots, A$), se calcula

$$\widehat{Y}_{aU}^{\text{HT}} = \sum_{h=1}^{H} \left[\frac{\delta_{ah} y_{h1}}{p_{h1}} + \frac{(1 - \delta_{ah}) y_{h2}}{p_{h2}} \right], \tag{7.21}$$

donde $p_{hi} = \frac{\pi_{hi}}{n_h} = \frac{\pi_{hi}}{2}$ ($i = 1, 2$).

iii. El estimador de la varianza de semimuestras equilibradas $\widehat{V}_{BH}$ [6] se obtiene como

$$\widehat{V}_{BH} = \frac{1}{A} \sum_{a=1}^{A} (\widehat{Y}_{aU}^{\text{HT}} - \widehat{Y}_{U}^{\text{HT}})^2, \tag{7.22}$$

donde $\widehat{Y}_{U}^{\text{HT}}$ es el estimador HT definido en (7.17).

Este estimador verifica la siguiente propiedad importante: si el conjunto de semimuestras está equilibrado, entonces

$$\widehat{V}_{BH} = \widehat{V}_0, \tag{7.23}$$

[6] del inglés *balanced half-samples*

donde $\widehat{V}_0$ está definido por (7.18). Por tanto, el procedimiento de semimuestras equilibradas es simplemente una forma alternativa de calcular $\widehat{V}_0$. Además, si el conjunto de semimuestras es equilibrado ortogonal completo, entonces la media de los valores $\widehat{Y}_{aU}^{\text{HT}}$ es igual a $\widehat{Y}_U^{\text{HT}}$, es decir,

$$\widehat{Y}_{BH} \equiv \frac{1}{A}\sum_{a=1}^{A}\widehat{Y}_{aU}^{\text{HT}} = \widehat{Y}_U^{\text{HT}}. \tag{7.24}$$

Por tanto, bajo el equilibrado ortogonal completo, el estimador de la varianza $\widehat{V}_{BH}$ se puede obtener de forma alternativa como

$$\frac{1}{A}\sum_{a=1}^{A}(\widehat{Y}_{aU}^{\text{HT}} - \widehat{Y}_{BH})^2.$$

Ejemplo 30. Considérese un diseño muestral estratificado aleatorio simple con dos elementos muestreados por estrato. Entonces, $\pi_{h1} = \pi_{h2} = 2p_{h1} = 2p_{h2} = 2/N_h$, el estimador HT es

$$\widehat{Y}_U^{\text{strHT}} = \sum_{h=1}^{H} N_h \bar{y}_{s_h}$$

y

$$\mathbb{V}\left(\widehat{Y}_U^{\text{strHT}}\right) = \frac{1}{2}\sum_{h=1}^{H} N_h^2(1-f_h)S_{yU_h}^2.$$

Supongamos que disponemos de un conjunto equilibrado de semimuestras. Entonces

$$\widehat{Y}_{aU}^{\text{HT}} = \sum_{h=1}^{H} N_h\left[\delta_{ah}y_{h1} + (1-\delta_{ah})y_{h2}\right]$$

y se verifica fácilmente que se cumple (7.23), de modo que

$$\widehat{V}_{BH} = \frac{1}{2}\sum_{h=1}^{H} N_h^2 S_{ys_h}^2.$$

Además, si el conjunto está en equilibrio ortogonal completo se cumple $\widehat{Y}_{BH} = \widehat{Y}_U^{\text{strHT}}$. ■

Veamos ahora cómo puede obtenerse un conjunto equilibrado de semimuestras. El conjunto que consiste en todas las 2^H semimuestras posibles es equilibrado y también equilibrado ortogonal completo. Sin embargo, por motivos computacionales, queremos un conjunto mucho más pequeño de semimuestras. En todos los casos de interés práctico, resulta que podemos encontrar un conjunto que consiste en no más de $H+3$ semimuestras equilibradas, por motivos que se discuten más adelante.

Se sigue de la ecuación (7.19) que encontrar un conjunto equilibrado de A semimuestras es lo mismo que encontrar una matriz $A \times H$ con elementos ε_{ah} iguales a 1 o a -1 y con

todas las columnas ortogonales a pares. Tales matrices de orden $c \times c$, originalmente destinadas a su uso en el diseño experimental, fueron creadas por Plackett y Burman 1946 para $c = 2, 4, 8, 12, 16, \ldots, 200$. Wolter 2007 también presenta estas matrices para $c = 2, 4, 8, 12, 16, \ldots, 100$. Las matrices presentadas por estos autores no son únicas, es decir, para una c dada, existen otras matrices $c \times c$ de valores de ε_{ah} que también satisfacen la condición de equilibrado. Hay que aclarar que las matrices de Plackett y Burman y de Wolter no verifican la condición de equilibrio ortogonal completo. Estas matrices tienen exactamente una columna de solo 1 o solo -1, por lo que (7.20) no se verifica. El resto de columnas en cada matriz, sin embargo, verifican la condición de equilibrio ortogonal completo.

Ejemplo 31. Veamos un ejemplo concreto de construcción de un conjunto de semimuestras equilibradas. Supongamos que tenemos $H = 8$ estratos. Podemos entonces usar la siguiente matriz 8×8 dada por Wolter 2007:

$$\begin{pmatrix} +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\ +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & -1 & +1 & +1 & -1 & -1 \\ +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\ +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ +1 & -1 & -1 & +1 & -1 & +1 & +1 & -1 \\ +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \end{pmatrix}.$$

Cada file de la matriz definirá una semimuestra. De este modo, la primera semimuestra consistirá de los primeros elementos de cada estrato. La segunda semimuestra consistirá de los primeros elementos de los estratos 1, 3, 5 y 7 y de los segundos elementos para los estratos 2, 4, 6 y 8. Y así para el resto de filas. El conjunto de semimuestras es equilibrado ya que la matriz es ortogonal por construcción. Si H fuese 7, podemos utilizar la misma matriz seleccionando cualesquiera 7 columnas para obtener 8 semimuestras balanceadas. Similarmente sucede si $H = 6, 5$. ■

El principio que ilustra este ejemplo consiste en que una matriz $c \times c$ (siendo c múltiplo de 4) puede emplearse para construir conjuntos de $A = c$ semimuestras balanceadas cuando $H = c, c-1, c-2, c-3$. De este modo, el número de semimuestras balanceadas no necesita ser mayor que $H + 3$.

Si se precisa equilibrio ortogonal completo, el procedimiento del siguiente ejemplo ilustra cómo obtenerlo.

Ejemplo 32. Si partimos de la misma matriz 8×8 anterior y suprimimos la primera columna, las restantes siete columnas satisfarán la propiedad de equilibrio ortogonal completo por construcción. De este modo, estas 7 columnas son ortogonales a pares y la

suma por columnas es 0. Por tanto, para $H = 7$ estratos obtenemos un conjunto de $A = 8$ semimuestras en equilibrio ortogonal completo. Para $H = 6$ estratos, suprimiendo la primera columna y cualquier de las siete restantes obtendríamos $A = 8$ semimuestras en equilibrio completo para el caso de $H = 6$ estratos. Similarmente para $H = 5$ estratos. Para el caso $H = 8$ la matriz anterior no sirve y deberíamos partir de una matriz 12×12 y aplicar este mismo procedimiento. Por tanto, este principio muestra que una matriz $c \times c$ (donde c es múltiplo de 4) puede usarse para obtener $A = c$ semimuestras en equilibrio ortogonal completo para $H = c - 1, c - 2, c - 3, c - 4$. ■

Comentario 33. La técnica de semimuestras equilibradas se puede ampliar a muestreo multietápico con muestreo estratificado de dos PSUs en cada estrato de PSUs. La técnica se aplica entonces a las PSUs. Consideremos la estimación del total poblacional Y_U por el estimador insesgado

$$\widehat{Y}_U^{nst} = \sum_{h=1}^{H} \frac{1}{\pi_{Ii}} \sum_{i \in s_{Ih}} \widehat{Y}_{U_i}^{\mathrm{HT}},$$

donde s_{Ih} es la muestra de dos PSUs del h-ésimo estrato de PSUs, π_{Ii} es la probabilidad de inclusión del i-ésimo PSU y $\widehat{Y}_{U_i}^{\mathrm{HT}}$ es un estimador insesgado del total Y_{U_i} del i-ésimo PSU, basado en el submuestreo dentro de la PSU mediante un diseño arbitrario en una o más etapas. Podemos proceder como en el muestreo en una etapa. Las dos PSUs seleccionadas del estrato h-ésimo se renombran como $h1$ y $h2$. Procediendo entonces como en los ejemplos anteriores para cada semimuestra ($a = 1, \ldots, A$) calculamos

$$\widehat{Y}_{aU}^{\mathrm{HT}} = \sum_{h=1}^{H} \left[\frac{\delta_{Iah} \widehat{Y}_{h1}}{p_{Ih1}} + \frac{(1 - \delta_{Iah}) \widehat{Y}_{h2}}{p_{Ih2}} \right],$$

donde $\delta_{Iah} = 1$ si la semimuestra a-ésima contiene la PSU etiquetado $h1$, (y $\delta_{Iah} = 0$ en otro caso), $p_{Ihj} = \frac{\pi_{Ihj}}{2}$ para $j = 1, 2$, y $\widehat{Y}_{hj}$ es el estimador de $Y_{U_{hj}}$ basado en los datos del submuestreo dentro de la PSU hj. Ahora se puede calcular el siguiente estimador de la varianza

$$\widehat{V}_{BH} = \frac{1}{A} \sum_{a=1}^{A} (\widehat{Y}_{aU}^{\mathrm{HT}} - \widehat{Y}_U^{\mathrm{HT}})^2.$$

■

Comentario 34. Una limitación de la técnica de semimuestras equilibradas descrita anteriormente es la necesidad de que exactamente dos elementos (o dos PSUs) se seleccionen de cada estrato. Se han sugerido modificaciones de la técnica para casos en los que el tamaño muestral de los estratos n_h sea superior a dos. Véase (Wolter 2007). ■

Comentario 35. La técnica de semimuestras equilibradas descrita aquí da lugar a un estimador de la varianza aproximado idéntico a (7.18) mediante un método alternativo de cálculo. El estimador de la varianza resultante habría sido apropiado bajo muestreo con reemplazamiento. El enfoque se olvida del hecho de que nuestra muestra ha sido seleccionada sin reemplazamiento, pero se puede modificar para tener en cuenta la

característica de la falta de reemplazamiento. Para cada semimuestra ($a = 1, \ldots, A$) calculamos

$$\widehat{Y}_{aU}^{\mathrm{HT}^*} = \widehat{Y}_U^{\mathrm{strHT}} + \sum_{h=1}^{H} \left(\frac{\pi_{h1}\pi_{h2}}{\pi_{h1h2}} - 1 \right)^{\frac{1}{2}} \left[\frac{\delta_{ah} 2 y_{h1}}{\pi_{h1}} + \frac{(1-\delta_{ah}) 2 y_{h2}}{\pi_{h2}} - \widehat{Y}_{U_h}^{\mathrm{HT}} \right],$$

donde $\widehat{Y}_U^{\mathrm{strHT}} = \sum_{h=1}^{H} \widehat{Y}_{U_h}^{\mathrm{HT}}$ y $\widehat{Y}_{U_h}^{\mathrm{HT}} = \frac{y_{h1}}{\pi_{h1}} + \frac{y_{h2}}{\pi_{h2}}$. Claramente, esto requiere conocer las probabilidades de inclusión de segundo orden $\pi_{h1,h2}$ tales que

$$\frac{\pi_{h1}\pi_{h2}}{\pi_{h1,h2}} - 1 \geq 0$$

para $h = 1, \ldots, H$. Un estimador modificado de la varianza se define ahora como

$$\widehat{V}_{BH}^* = \frac{1}{A} \sum_{a=1}^{A} (\widehat{Y}_{aU}^{\mathrm{HT}^*} - \widehat{Y}_U^{\mathrm{HT}})^2. \tag{7.25}$$

Si el conjunto de semimuestras es equilibrado ortogonal completo, entonces $\widehat{V}_{BH}^*$ es idéntico al estimador de la varianza de Sen-Yates-Grundy (véase el tema 2 del bloque de Producción Estadística Oficial (Materias Comunes), que en este caso (con todos los $n_h = 2$) es

$$\widehat{V} = \sum_{h=1}^{H} \left(\frac{\pi_{h1}\pi_{h2}}{\pi_{h1,h2}} - 1 \right) \left(\frac{y_{h1}}{\pi_{h1}} - \frac{y_{h2}}{\pi_{h2}} \right)^2$$

Por ejemplo, con muestreo aleatorio simple sin reemplazamiento en cada estrato, tenemos

$$\pi_{h1} = \pi_{h2} = \frac{2}{N_h}$$

$$\pi_{h1h2} = \frac{2}{N_h(N_h - 1)}$$

y

$$\frac{\pi_{h1}\pi_{h2}}{\pi_{h1h2}} - 1 = \frac{N_h - 2}{N_h} > 0,$$

por tanto, el método modificado funciona y nos da el estimador insesgado

$$\widehat{V}_{BH}^* = \widehat{V} = \frac{1}{2} \sum_{h=1}^{H} N_h^2 (1 - f_h) S_{ys_h}^2.$$

■

Comentario 36. En nuestra discusión anterior, hemos usado semimuestras equilibradas para estimar la varianza del estimador HT $\widehat{Y}_U^{\mathrm{HT}}$. En este caso, sin embargo, los métodos estándares discutidos en el tema 2 del bloque de materias comunes sobre Producción Estadística Oficial se pueden aplicar fácilmente y la técnica de semimuestras equilibradas no ofrece ninguna ventaja particular. Su principal interés reside en aplicaciones a estimaciones de la varianza para parámetros más complejos, por ejemplo, regresión de poblaciones finitas y coeficientes de correlación. En tales casos, no tenemos resultados

exactos para las propiedades (como el sesgo) del estimador de la varianza de semimuestras equilibradas. Pero como se sabe que la técnica funciona bien para el estimador HT sencillo, el supuesto tácito es que continúa dando resultados satisfactorios, incluso si la situación es más compleja. Un conjunto considerable de evidencias empíricas pueden respaldar estos supuestos.

Cuando el parámetro θ y el estimador $\widehat{\theta}$ son complejos, la literatura sugiere varios estimadores de semimuestras equilibradas alternativos para la varianza $\mathbb{V}(\widehat{\theta})$. Seguiremos asumiendo un muestreo estratificado de dos elementos por estrato, pero no necesariamente muestreo aleatorio simple sin reemplazamiento dentro del estrato. Sea $\widehat{\theta}$ un estimador de θ de la muestra completa. Sea $\widehat{\theta}_a$ un estimador de θ basado en los datos de la a-ésima semimuestra y de la misma estructura que $\widehat{\theta}$. Sea $\widehat{\theta}_a^c$ un estimador de θ basado en el complementario de la a-ésima semimuestra. Se han sugerido cuatro estimadores de la varianza en esta situación:

$$\widehat{V}_{BH1} = \frac{1}{A}\sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta})^2, \tag{7.26}$$

$$\widehat{V}_{BH2} = \frac{1}{A}\sum_{a=1}^{A}(\widehat{\theta}_a^c - \widehat{\theta})^2, \tag{7.27}$$

$$\widehat{V}_{BH3} = \frac{1}{2}\sum_{a=1}^{A}(\widehat{V}_{BH1} + \widehat{V}_{BH2}) \tag{7.28}$$

y

$$\widehat{V}_{BH4} = \frac{1}{4A}\sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta}_a^c)^2 \tag{7.29}$$

Sustituyendo

$$\widehat{\theta}_{BH} = \frac{1}{A}\sum_{a=1}^{A}\widehat{\theta}_a$$

en lugar de $\widehat{\theta}$ en las ecuaciones (7.26) a (7.29) obtenemos incluso más fórmulas. Las comparaciones teóricas entre estos estimadores alternativos de la varianza son escasos. Aclaramos que si $\widehat{\theta}$ es un estimador HT o una combinación lineal de estimadores HT y si el conjunto de semimuestras es equilibrado, entonces (7.26) a (7.29) son idénticos. Estos cuatro estimadores están incluidos en un estudio sobre Monte Carlo de Frankel 1971. No hay grandes diferencias en el comportamiento (sesgo, error cuadrático medio, tasa de cobertura) entre los cuatro estimadores. Desde un punto de vista computacional, $\widehat{V}_{BH1}$ y $\widehat{V}_{BH2}$ son ligeramente más sencillos que los otros dos estimadores. ■

7.4 Método *jackknife*

El método jackknife surgió fuera del campo del muestreo para encuestas. La idea inicial, desarrollada por Quenouille 1949; Quenouille 1956, fue usar *jackknifing* [7] para reducir el sesgo de un estimador en el contexto de una población infinita. Quenouille 1958 posteriormente sugirió que la técnica también debería usarse para calcular estimaciones de la varianza. Para poblaciones finitas, el método *jackknife* fue considerado por primera vez por Durbin 1959. En este tema veremos una revisión del método *jackknife* tal y como se usa normalmente para estimar la varianza en muestreo. Para más detalle, véase (Wolter 2007).

Sea s una muestra (la muestra completa) de n elementos obtenidos mediante un diseño muestral no estratificado de elementos. El muestreo estratificado se discutirá más tarde. Sea θ el parámetro poblacional a estimar mediante $\widehat{\theta}$, un estimador basado en los datos de la muestra total s. El objetivo es estimar $\mathbb{V}(\widehat{\theta})$.

El método *jackknife* comienza con una partición de la muestra s en A grupos aleatorios dependientes de igual tamaño m $(= \frac{n}{A})$, como se ha descrito en la Sección 7.2.2. Asumimos que, para cualquier s dada, cada grupo es una muestra aleatoria simple sin reemplazamiento obtenida a partir de s, incluso si s en sí misma no es una muestra aleatoria simple sin reemplazamiento. A continuación, para cada grupo $(a = 1, \ldots, A)$, calculamos $\widehat{\theta}_{(a)}$, un estimador de θ con la misma forma funcional que $\widehat{\theta}$, pero basado únicamente en los datos que hay en la muestra después de eliminar el grupo a. Para $a = 1, \ldots, A$, definimos entonces

$$\widehat{\theta}_a = A\widehat{\theta} - (A-1)\widehat{\theta}_{(a)}, \tag{7.30}$$

que a veces se denomina el a-ésimo pseudovalor. El estimador *jackknife* de θ (un estimador alternativo a $\widehat{\theta}$) es

$$\widehat{\theta}_{JK} = \frac{1}{A}\sum_{a=1}^{A}\widehat{\theta}_a \tag{7.31}$$

y el estimador de la varianza *jackknife* se define como

$$\widehat{V}_{JK1} = \frac{1}{A(A-1)}\sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta}_{JK})^2. \tag{7.32}$$

En la práctica, $\widehat{V}_{JK1}$ se utiliza tanto como estimador de $\mathbb{V}(\widehat{\theta})$ como de $\mathbb{V}(\widehat{\theta}_{JK})$. Una alternativa a $\widehat{V}_{JK1}$ que se usa a veces es

$$\widehat{V}_{JK2} = \frac{1}{A(A-1)}\sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta})^2. \tag{7.33}$$

[7] El verbo *jackknife* se utilizará en inglés, no se traduce dado su uso extendido en la comunidad estadística.

Señalamos que $\widehat{V}_{JK2} \geq \widehat{V}_{JK1}$, que se sigue de (7.4).

Además, se sigue de (7.3) que si los $\widehat{\theta}_a$ ($a = 1, \ldots, A$) fuesen variables aleatorias incorreladas con la misma esperanza, entonces $\widehat{V}_{JK1}$ sería insesgado para $\mathbb{V}(\widehat{\theta}_{JK})$. Sin embargo, los estimadores $\widehat{\theta}_a$ están en general correlados, por eso no se verifica la insesgadez. No hay resultados exactos (tamaño muestral finito) para las propiedades (sesgo, etc.) de los estimadores de la varianza *jackknife* cuando $\widehat{\theta}$ es más complejo que un estimador HT. En el caso en que θ es un total poblacional y $\widehat{\theta}$ el estimador HT correspondiente, el método *jackknife* funciona bien en lo que se refiere al sesgo. Este resultado, junto con la evidencia empírica, parece ser la principal justificación para usar el método *jackknife* en situaciones más complejas.

Ejemplo 33. Sea $\theta = Y_U = \sum_{k \in U} y_k$ el total poblacional. Sea s una muestra de U de tamaño fijo n obtenida mediante muestreo proporcional al tamaño sin reemplazamiento con probabilidades de inclusión $\pi_1, \ldots, \pi_N$. Sea $\widehat{\theta} = \widehat{Y}_U^{\text{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k}$. Divídase s en A grupos aleatorios de igual tamaño como se ha descrito anteriormente. Definimos

$$\widehat{\theta}_{(a)} = \widehat{Y}_{U_{(a)}}^{\text{HT}} = \frac{A}{A-1} \sum_{k \in s - s_a} \frac{y_k}{\pi_k}. \tag{7.34}$$

Se sigue de (7.30), (7.31) y (7.34) que

$$\widehat{\theta}_a = \widehat{Y}_{U_{(a)}}^{\text{HT}} = A \sum_{k \in s_a} \frac{y_k}{\pi_k}$$

y

$$\widehat{\theta}_{JK} = \widehat{Y}_U^{\text{JK}} = \frac{1}{A} \sum_{a=1}^{A} \widehat{Y}_{U_{(a)}}^{\text{HT}} = \widehat{Y}_U^{\text{HT}}.$$

Por tanto, los estimadores de la varianza $\widehat{V}_{JK1}$ y $\widehat{V}_{JK2}$ son idénticos:

$$\widehat{V}_{JK1} = \widehat{V}_{JK2} = \frac{1}{A(A-1)} \sum_{a=1}^{A} \left(\widehat{Y}_{U_{(a)}}^{\text{HT}} - \widehat{Y}_U^{\text{HT}} \right)^2 \equiv \widehat{V}_{JK}. \tag{7.35}$$

Obsérvese que en este caso $\widehat{V}_{JK1} = \widehat{V}_{JK2} = \widehat{V}_{\text{DRG}}$. es decir, con el estimador HT, la técnica *jackknife* es un método alternativo para calcular $\widehat{V}_{\text{DRG}}$ y, por tanto, tienen el mismo sesgo. ■

Ejemplo 34. En caso de muestreo aleatorio simple sin reemplazamiento, los resultados del ejemplo anterior se reducen a

$$\begin{aligned}
\widehat{Y}_U^{\text{HT}} &= N\bar{y}_s, \\
\widehat{Y}_{U_{(a)}}^{\text{HT}} &= N\bar{y}_{s-s_a}, \\
\widehat{Y}_{U_{(a)}}^{\text{HT}} &= N\bar{y}_{s_a}, \\
\widehat{Y}_{JK} &= N\bar{y}_s, \\
\widehat{V}_{JK} &= \frac{N^2}{A(A-1)} \sum_{a=1}^{A} \left(\bar{y}_{s_a} - \bar{y}_s\right)^2, \\
\mathbb{E}\left(\widehat{V}_{JK}\right) &= \mathbb{E}\left[N^2 \frac{S_{ys}^2}{n}\right], \\
\mathbb{E}\left[\widehat{V}_{JK}\right] - \mathbb{V}\left(N\bar{y}_s\right) &= NS_{yU}^2.
\end{aligned}$$

Por tanto, en este caso, para corregir el sesgo tan solo hace falta incluir el factor $1-f$. Si $A = n$ y $m = 1$, obtenemos

$$\widehat{V}_{JK} = N^2 \frac{S_{ys}^2}{n}$$

y

$$(1-f)\widehat{V}_{JK} = N^2(1-f)\frac{S_{ys}^2}{n},$$

que es el estimador usual para este tipo de muestreo. ■

Comentario 37. Cuando el método *jackknife* se aplica a muestreo estratificado, usaremos estimadores de la varianza distintos de (7.32) y de (7.33). De acuerdo con Wolter 2007, pág. 175, se debe ser 'especialmente cuidadoso de no aplicar los estimadores *jackknife* clásicos . . . a los problemas de muestreo estratificado'. Asumamos que la muestra del estrato h ($h = 1, \ldots, H$) está particionada aleatoriamente en A_h grupos. para un total de $A = \sum_{h=1}^{H} A_h$ grupos. Como antes, sea $\widehat{\theta}$ el estimador de θ con la muestra completa. Sea $\widehat{\theta}_{(ha)}$ el estimador de θ basado en lo que queda en el estrato h después de omitir el a-ésimo grupo. Un estimador sugerido para $\mathbb{V}(\widehat{\theta})$ es (véase (Rust 1985))

$$\widehat{V}_{JK3} = \sum_{h=1}^{H} \left[\frac{(A_h - 1)}{A_h}\right] \sum_{a=1}^{A_h} (\widehat{\theta}_{(ha)} - \widehat{\theta})^2, \tag{7.36}$$

que sería insesgado si la selección muestral fuese con reemplazamiento y si $\widehat{\theta}$ fuese el estimador HT del total poblacional $\theta = Y_U = \sum_{k \in U} y_k$.

En particular, supongamos que se usa muestreo aleatorio simple sin reemplazamiento en cada estrato. Sea

$$\widehat{Y}_U^{\text{strHT}} = \sum_{h=1}^{H} N_h \bar{y}_{s_h}$$

y

$$\begin{aligned}\widehat{Y}^{\mathrm{HT}}_{U_{(ha)}} &= N_1\bar{y}_{s_1} + \cdots + N_{h-1}\bar{y}_{s_{h-1}} + N_h\bar{y}_{s_h - s_{ha}} \\ &\quad + N_{h+1}\bar{y}_{s_{h+1}} + \cdots + N_H\bar{y}_{s_H}.\end{aligned}$$

Entonces obtenemos

$$\widehat{V}_{JK3} = \sum_{h=1}^{H} \frac{N_h^2}{A_h(A_h-1)} \sum_{a=1}^{A_h} (\bar{y}_{s_{ha}} - \bar{y}_{s_h})^2.$$

Bajo la hipótesis adicional de que $A_h = n_h$, obtenemos

$$\widehat{V}_{JK3} = \sum_{h=1}^{H} \frac{N_h^2 S^2_{ys_h}}{n_h},$$

que es el estimador de la varianza, en el que se ha omitido la corrección por población finita. Véase (Wolter 2007) para otros estimadores para el caso del muestreo estratificado. ■

Comentario 38. En el muestreo multietápico, el método *jackknife* normalmente se aplica a nivel de las PSUs. Supongamos una muestra en primera etapa s_I que contiene n_I PSUs se selecciona a partir del conjunto U_I compuesto por N_I PSUs. Sea s_I particionado aleatoriamente en A grupos de PSUs, con m PSUs en cada grupo. Sea $\widehat{\theta}$ es estimador de θ con la muestra completa, y denotemos por $\widehat{\theta}_{(a)}$ el estimador de θ que está basado en los datos que quedan tras eliminar el a-ésimo grupo de PSUs. Usando las ecuaciones (7.30) a (7.33) en este nuevo contexto, obtenemos

$$\widehat{\theta}_a = A\widehat{\theta} - (A-1)\widehat{\theta}_{(a)}; \quad a = 1, \ldots, A,$$

$$\widehat{\theta}_{JK} = \frac{1}{A}\sum_{a=1}^{A} \widehat{\theta}_a$$

y los estimadores de la varianza son

$$\widehat{V}_{JK1} = \frac{1}{A(A-1)} \sum_{a=1}^{A} (\widehat{\theta}_a - \widehat{\theta}_{JK})^2,$$

$$\widehat{V}_{JK2} = \frac{1}{A(A-1)} \sum_{a=1}^{A} (\widehat{\theta}_a - \widehat{\theta})^2,$$

que sirven para el propósito dual de estimar $\mathbb{V}(\widehat{\theta})$ así como $\mathbb{V}(\widehat{\theta})_{JK}$. Por ejemplo, sea $\theta = Y_U = \sum_{i\in U_I} Y_{U_i}$, y

$$\widehat{\theta} = \widehat{t} = \sum_{i\in s_I} \frac{\widehat{Y}_{U_i}}{\pi_{Ii}},$$

donde $\widehat{Y}_{U_i}$ es un estimador del i-ésimo total Y_{U_i} de la PSU i basado en submuestrear dentro de la i-ésima PSU y π_{Ii} es la probabilidad de inclusión de primera etapa de esta PSU. Además

$$\widehat{\theta}_{(a)} = \widehat{Y}_{(a)} = \left[\frac{1}{A(A-1)}\right] \sum_{i \in s_I - s_{Ia}} \frac{\widehat{Y}_i}{\pi_{Ii}}$$

lo que da

$$\widehat{Y}_{aU}^{\text{HT}} = A \sum_{i \in s_{Ia}} \frac{\widehat{Y}_i}{\pi_{Ii}},$$

$$\widehat{\theta}_{JK} = \widehat{Y}_{JK} = \widehat{Y}_U^{\text{HT}}$$

y el estimador de la varianza

$$\widehat{V}_{JK1} = \widehat{V}_{JK2} = \frac{1}{A(A-1)} \sum_{a=1}^{A} (\widehat{Y}_{aU}^{\text{HT}} - \widehat{Y}_U^{\text{HT}})^2 \equiv \widehat{V}_{JK}$$

En el muestreo estratificado de PSUs, la técnica descrita en el Comentario 37 se podría aplicar a las PSUs. ■

Comentario 39. Para el método *jackknife* necesitamos fijar un número adecuado de grupos A. Para tener una buena acuracidad en los resultados del estimador de la varianza, lo ideal sería tener tantos grupos como sea posible, lo que significa $A = n$, es decir, $m = 1$. Por otro lado, por motivos computacionales, lo ideal sería tener tan pocos grupos como fuese posible, el caso más extremo sería $A = 2$, y $m = \frac{n}{2}$. En la práctica, la elección más frecuente parece ser $A = n$, o, cuando el coste computacional es una cuestión seria, un compromiso entre los dos extremos $A = n$ y $A = 2$. ■

Comentario 40. Kovar, Rao y Wu 1988 encontraron en un estudio empírico que el método *jackknife* funcionaba mal para estimar los estimadores de la varianza de cuantiles poblacionales. ■

7.5 Método *bootstrap*

Como en el caso del método *jackknife*, el método *bootstrap* se desarrolló fuera de la teoría de muestreo como una forma de obtener estimaciones aproximadas de la varianza e intervalos de confianza. Su creador fue Efron (Efron 1979; Efron 1981; Efron 1982).

De forma gradual, el método *bootstrap* empezó a llamar la atención como una alternativa a las otras técnicas de estimación aproximada de la varianza vistas en este tema. El método *bootstrap* se diseñó inicialmente para su uso con observaciones independientes, la hipótesis inicial de la teoría estadística tradicional. Un problema básico, que todavía no se ha resuelto, es cómo el método debería de modificarse de manera correcta para incorporar las características especiales del muestreo, incluyendo la falta de independencia que aparece en el muestreo sin reemplazamiento y otras complejidades de diseños y

de estimadores.

El objetivo de esta sección es indicar el simple uso de la idea del *bootstrap* en el muestreo. Para más información sobre el *bootstrap* en muestreo véase (Bickel y Freedman 1984; Bondensson y Holm 1985; Gross 1980; Kovar, Rao y Wu 1988; McCarthy 1985; Rao y Wu 1984; Rao y Wu 1987; Wolter 2007).

Supongamos que se selecciona una muestra probabilística s de una población U mediante un diseño de muestreo arbitrario sin reemplazamiento. El parámetro poblacional θ se estima mediante $\widehat{\theta}$, y buscamos un estimador de $\mathbb{V}(\widehat{\theta})$. Veamos una breve descripción de cómo funciona el método *bootstrap*.

i. Usando los datos muestrales, construimos una población artificial U^*, que asumimos que es una réplica de la población U real, pero desconocida.

ii. Seleccionamos un conjunto de muestras independientes, 'remuestras' o 'muestras bootstrap', de U^* mediante un diseño idéntico al que se usó para seleccionar s a partir de U. La independencia implica que cada muestra *bootstrap* debe devolverse a U^* antes de seleccionar la siguiente. Para cada muestra *bootstrap*, se calcula una estimación $\widehat{\theta}_a^*$ $(a = 1, \ldots, A)$ de la misma forma en que se calculó $\widehat{\theta}$.

iii. Se considera la distribución observada de $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_A^*$ como una 'estimación' de la distribución muestral del estimador $\widehat{\theta}$ y $\mathbb{V}(\widehat{\theta})$ se estima mediante

$$\widehat{V}_{BS} = \frac{1}{A-1} \sum_{a=1}^{A} (\widehat{\theta}_a^* - \widehat{\theta}^*)^2,$$

donde

$$\widehat{\theta}^* = \frac{1}{A} \sum_{a=1}^{A} \widehat{\theta}_a^*.$$

Comentario 41. En algunas aplicaciones del método *bootstrap*, cuando el cálculo de un intervalo de confianza es el principal objetivo, no se calcula una estimación de la varianza $\widehat{V}_{BS}$, sino que el intervalo se obtiene directamente a partir de las distribución observada de $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_A^*$. Por ejemplo, para un intervalo de confianza al $95\,\%$ se toma el intervalo entre los puntos menores y mayores al $2{,}5\,\%$ de esa distribución. ■

El siguiente ejemplo muestra cómo pueden usarse los principios del método *bootstrap* para estimar una varianza $\mathbb{V}(\widehat{\theta})$, cuando $\widehat{\theta}$ es el estimador HT del total poblacional.

Ejemplo 35. Sea $\theta = Y_U = \sum_{k \in U} y_k$ y $\widehat{\theta} = \widehat{Y}_U^{\mathrm{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k}$, donde π_k son las probabilidades de inclusión bajo el diseño utilizado para obtener s. Por simplicidad, supongamos que el diseño tiene un tamaño muestral fijo n. Buscamos estimar $\mathbb{V}(\widehat{\theta}) = V$.

El primer paso es construir la población artificial U^*. Una posibilidad es que U^* esté compuesto de réplicas de los elementos en s. Para cada $k \in s$ creamos $\frac{1}{\pi_k}$ elementos artificiales de U^*, compartiendo todas el mismo valor y, concretamente, y_k y el mismo valor de π, π_k. Por simplicidad asumiremos que $\frac{1}{\pi_k}$ es un entero para todo $k \in s$. Usaremos el subíndice i para etiquetar a los elementos artificiales de U^*. Para el i-ésimo elemento, sean y_i^* y π_i^* los valores respectivos de y y π. Entonces, a cada $k \in s$ le corresponden $\frac{1}{\pi_k}$ elementos en U^* y todos ellos tienen $y_i^* = y_k$ y $\pi_i^* = \pi_k$. Mediante esta construcción, U^* será del tamaño

$$N^* = \sum_s \frac{1}{\pi_k} = \widehat{N}_U^{\text{HT}}$$

y el total poblacional de U^* será

$$Y_{U^*} = \sum_{i \in U^*} y_i^* = \sum_{k \in s} \frac{y_k}{\pi_k} = \widehat{Y}_U^{\text{HT}}.$$

El siguiente paso es seleccionar las remuestras de U^*. Para simplificar, sea cada remuestra una muestra proporcional al tamaño con reemplazamiento de n muestras, con probabilidades de selección $p_i^* = \frac{\pi_k^*}{n}$ para cada $i \in U^*$. Entonces, las remuestras son con reemplazamiento aunque la muestra original s fuese sin reemplazamiento. Obsérvese que $\sum_{i \in U^*} p_i^* = 1$.

Para cada muestra *bootstrap* ($a = 1, \ldots, A$) calculamos el estimador de Hansen-Hurwitz

$$\widehat{\theta}_a^* = \widehat{Y}_a^{HH*} = \frac{1}{n} \sum_{j=1}^{n} \frac{y_{i_{j(a)}}^*}{p_{i_{j(a)}}^*},$$

donde el valor de y dado por $y_{i_{j(a)}}^*$ está asociado con el elemento de U^* que se obtuvo en la j-ésima selección para la a-ésima muestra *bootstrap*. Definiendo

$$\widehat{Y}^{HH*} = \frac{1}{A} \sum_{a=1}^{A} \widehat{Y}_a^{HH*}$$

definimos el estimador de la varianza como

$$\widehat{V}_{BS} = \frac{1}{A-1} \sum_{a=1}^{A} \left(\widehat{Y}_a^{HH*} - \widehat{Y}^{HH*} \right)^2.$$

Su sesgo se obtiene básicamente. Dada la muestra s, los estimadores $\widehat{Y}_1^{HH*}, \ldots, \widehat{Y}_A^{HH*}$ son variables aleatorias independientes e idénticamente distribuidas. Entonces:

$$\mathbb{E}\left(\widehat{V}_{BS} | s \right) \;=\; \mathbb{V}\left(\widehat{Y}_a^{HH*} \middle| s \right)$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i\in U^*} p_i^* \left(\frac{y_i^*}{p_i^*} - Y_{U*}\right)^2 \\
&= \frac{1}{n^2}\sum_{k\in s} \left(\frac{y_k}{p_k} - \widehat{Y}_U^{\mathrm{HT}}\right)^2 \\
&= \frac{n-1}{n}\left[\frac{1}{n(n-1)}\sum_{k\in s}\left(\frac{ny_k}{\pi_k} - \widehat{Y}_U^{\mathrm{HT}}\right)^2\right].
\end{aligned}
$$

Por tanto, para el estimador $\widehat{Y}_U^{\mathrm{HT}}$ obtenemos

$$
\mathbb{E}\left(\widehat{V}_{BS}\right) = \mathbb{E}_p\left[\mathbb{E}\left(\widehat{V}_{BS}|s\right)\right] = \frac{n-1}{n}\mathbb{E}_p\left[\frac{1}{n(n-1)}\sum_{k\in s}\left(\frac{ny_k}{\pi_k} - \widehat{Y}_U^{\mathrm{HT}}\right)^2\right].
$$

Es decir, el estimador *bootstrap* proporciona con el mismo valor esperado que los estimadores de las técnicas anteriores. ■

Bibliografía

Bickel, P.J. y D.A. Freedman (1984). "Asymptotic normality and the bootstrap in stratified sampling". En: *The Annals of Statistics* 12, págs. 470-482.

Bondensson, L. y S. Holm (1985). "Bootstrap-estimation of the mean-square error of the ratio estimate for sampling without replacement". En: *Contributions to Probability and Statistics in Honour of Gunnar Blom*, págs. 85-96.

Deming, W.E. (1956). "On simplifications of sampling design through replication with equal probabilities and without stages". En: *J. Amer. Stat. Assoc.* 51, págs. 24-53.

Durbin, J. (1959). "A note on the application of Quenouille's method of bias reduction to the estimation of ratios". En: *Biometrika* 46, págs. 477-480.

Efron, B. (1979). "Bootstrap methods: Another look at the Jackknife". En: *Annals of Statistics* 7, págs. 1-26.

– (1981). "Nonparametric standard errors and confidence intervals". En: *Canadian Journal of Statistics* 9, págs. 139-172.

– (1982). "The jackknife, the bootstrap and other resampling plans". En: *Philadelphia: SIAM monograph no 38*.

Frankel, M.R. (1971). *Inference from Survey Samples: An Empirical Investigation*. Institute for Social Research, University of Michigan.

Gross, S.T. (1980). "Median estimation in sample surveys". En: *Proceeding of the Section on Survey Research, America Statistical Association*, págs. 181-184.

Hansen, M.H., W.N. Hurwitz y W.G. Madow (1953). *Sample survey Methods and Theory*. 7th. Vol. I and II. New York: Wiley.

Kovar, J.G., J.N.K. Rao y C.F.J. Wu (1988). "Bootstrap and other methods to measure errors in survey estimates". En: *Canadian Journal of Statatistics* 16, págs. 25-45.

Mahalanobis, P.C. (1939). "A sample survey of the acreage under jute in Bengal". En: *Sankhya* 4, págs. 511-531.

– (1944). "On large-scale sample surveys". En: *Philosophical Transactions of the Royal Society of London B* 231, págs. 329-345.

Mahalanobis, P.C. (1946). "Recent experiments in statistical sampling in the Indian Statistical Institute". En: *Journal of the Royal Statistical Society* 109, págs. 325-370.
McCarthy, P.J. (1966). "Replication: An approach to the analysis of data from complex surveys". En: *Vital and Heallth Statistics*. Ed. por National Center for Health Statistics. 2 14. Washington DC: Public Health Service.
– (1969). "Pseudo-replication: half-samples". En: *Review of the International Statistical Institute* 37, págs. 239-264.
– (1985). "The bootstrap and finite population sampling". En: *Vital and Health Statistics, Series 2 No. 95*.
Norlèn, U. y T. Waller (1979). "Estimation in a complex survey - experiences from a survey of buildings with regard to energy usage". En: *Statistisk Tidskrift* 17, págs. 109-124.
Plackett, R.L. y J.P. Burman (1946). "The design of optimum multifactorial experiments". En: *Biometrika* 33, págs. 305-325.
Quenouille, M.H. (1949). "Problems in plane sampling". En: *Annals of Mathematical Statistics* 20, págs. 355-375.
– (1956). "Notes on bias in estimation". En: *Biometrika* 43, págs. 353-360.
– (1958). "Bias and confidence in not-quite large samples (abstract)". En: *Annals of Mathematical Statistics* 29, pág. 614.
Rao, J.N.K. y C.F.J. Wu (1984). "Bootstrap inference for sample surveys". En: *Proc. ASA Secfion on Survey Research Methods*, págs. 106-112.
– (1987). "Methods for standard errors and confidence intervals from sample survey data: some recent work". En: *Bulletin of the International Statistical Institute* 52, págs. 5-21.
Rust, K. (1985). "Variance estimation for complex estimators in sample surveys". En: *J. Official Stat.* 1, págs. 381-397.
Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.
Wolter, K. (2007). *Introduction to variance estimation*. 2nd. New York: Springer.

Tema 8

Estimación en dominios. Los métodos básicos de estimación en dominios. Condicionamiento sobre el tamaño muestral del dominio. Dominios pequeños: estimadores sintéticos.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía.

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

8.1 Introducción

En la mayoría de las operaciones estadísticas, sino en todas, las estimaciones se necesitan no sólo para la población total, sino también para varias subpoblaciones. Por ejemplo, en las encuestas económicas se suelen publicar los datos por actividad económica y/o regiones, y en las las encuestas sociales por tipo de hogar, por sexo y/o por edad.

A menudo, los datos de base (el tamaño muestral) para para la población total es considerable, y por eso se pueden obtener estimaciones con una buena precisión a nivel nacional, mientras que cuando la muestra total se desagrega a un nivel local detallado, algunas áreas tendrán tan pocas observaciones (o ninguna) que es imposible obtener estimaciones de una precisión aceptable. Por otro lado, otras subpoblaciones no presentan ningún problema de carencia de datos. En una muestra aleatoria simple sin reemplazamiento de tamaño $n = 1000$ de una población de individuos, se espera que en torno a 500 observaciones sean de "hombres"y otras tantas de "mujeres", una base sólida para obtener estimaciones separadas por sexo.

Definición 10

Las subpoblaciones para las que se necesitan estimaciones puntuales e intervalos de confianza separados se llaman *dominios*. A veces se usa el término *dominios de estudio*, en particular sobre una subpoblación diseñada en la fase de planificación como aquélla para la que es necesario una estimación separada. En tal caso se pueden reservar recursos adecuados para garantizar que la muestra es suficientemente grande en ese dominio y así obtener una estimación de una precisión aceptable.

Sin embargo, en general, un dominio es cualquier subpoblación para la cual se puede necesitar una estimación separada, antes o después de la fase de planificación. A menudo, la necesidad de estimaciones para ciertos dominios se manifiesta únicamente después de que se ha decidido el diseño muestral o después de que el muestreo y el trabajo de campo han sido completados. Para responder a estas necesidades, el estadístico debe de aprovechar lo mejor posible los datos que tenga a mano. El número de observaciones que caen en el dominio es normalmente aleatorio y a veces muy pequeño. Éstas son las características que le dan a la estimación en dominios sus rasgos particulares.

Definición 11

Un *dominio pequeño* es aquél que representa únicamente una pequeña fracción del total de la población. En este dominio, el estadístico normalmente tiene muy pocas observaciones (o ninguna). Esta complicación crea el problema de "la estimación en dominios pequeños"también conocido como *la estimación en pequeñas áreas*. Los pequeños dominios suelen están definidos geográficamente.

Para más información sobre la estimación en pequeñas áreas véase (Rao y Molina 2015).

En este tema se verán técnicas que asumen que el número de observaciones de un dominio es escaso, pero no extremadamente pequeño.

Consideremos una partición de la población $U = \{1, ..., k, ..., N\}$ en D subconjuntos, $U_1, ..., U_d, ..., U_D$, llamados dominios. Sea N_d el tamaño de U_d. Tenemos las ecuaciones de particiones

$$U = \bigcup_{d=1}^{D} U_d; \quad N = \sum_{d=1}^{D} N_d. \tag{8.1}$$

El objetivo es estimar los totales de los dominios

$$Y_d = \sum_{U_d} y_k; \quad d = 1, ..., D,$$

o las medias de los dominios

$$\bar{y}_{U_d} = \frac{Y_d}{N_d}; \quad d = 1, ..., D.$$

Comentario 42. En este tema, usaremos Y_d para el total del dominio e $\bar{y}_{U_d}$ para la media del dominio. La notación general para los estimadores de estos parámetros es $\widehat{Y}_d$ y $\widehat{\bar{y}}_{U_d}$, respectivamente.

Si N_d es desconocido (es el caso más común en la práctica), entonces el parámetro $\bar{y}_{U_d} = \frac{Y_d}{N_d}$ es un cociente de dos parámetros desconocidos. Cabe señalar que el estimador de $\bar{y}_{U_d}$ analizado en la sección *3.3.2 Estimadores, varianza y estimador de la varianza* del tema 3[1] del bloque "Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas"del grupo de materias comunes es el resultado de considerar $\bar{y}_{U_d}$ como una razón de dos totales desconocidos.

Comentario 43. La división en dominios representa un nuevo tipo de partición de la población en subconjuntos. Algunas particiones importantes que se han visto en temas anteriores son: (a) un conjunto de estratos (con motivo del muestreo estratificado); (b) un conjunto de PSUs (para el muestreo de conglomerados bietápico o multietápico); (c) un conjunto de grupos (para los modelos de grupos en la estimación por regresión).

Si los elementos en u_d (digamos, un área geográfica) pueden ser identificados de antemano e incluidos en una lista, el estadístico puede seleccionar una muestra directamente del dominio de acuerdo con un diseño adecuado. El dominio es entonces designado en efecto como un estrato, y la selección de la muestra en el dominio se controla mediante la elección del diseño en el estrato. Sin embargo, incluso si U_d pudiese ser seleccionado como un estrato, no siempre se elige hacerlo así en la práctica. Si el número de dominios D es grande, se puede incurrir en un coste alto si se decide realizar una selección controlada en cada dominio. En este tema veremos el caso en el que es imposible (por ejemplo, por falta de marco) o no es práctico (por ejemplo, por razones de coste) considerar cada dominio como un estrato.

Asumamos que una encuesta se lleva a cabo sobre la población U. Es decir, una muestra probabilística s de tamaño n_s se selecciona a partir de U de acuerdo con un diseño muestral especificado $p(\cdot)$ con probabilidades de inclusión π_k, π_{kl} y, como de costumbre, fijamos $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ y $\check{\Delta}_{kl} = \frac{\Delta_{kl}}{\pi_{kl}}$.

Denotemos por s_d la parte de s que cae en U_d, es decir,

$$s_d = s \cap U_d$$

Denotamos por n_{s_d} el tamaño de s_d. Los análogos muestrales de las ecuaciones (8.1) son entonces

$$s = \bigcup_{d=1}^{D} s_d; \quad n_s = \sum_{d=1}^{D} n_{s_d}. \tag{8.2}$$

Aquí, n_{s_d} es aleatorio y posiblemente bastante pequeño. Con este número aleatorio de observaciones, la tarea del estadístico es obtener la mejor estimación posible dentro del

[1] Tema 3. Estimación insesgada en diseños muestrales sobre unidades elementales I. Muestreo de Bernoulli: definición, estimadores, varianza y estimador de la varianza. Muestreo aleatorio simple: sin y con reemplazamiento: definición, estimadores, varianza y estimador de la varianza.

dominio.

Es útil, como en la sección *3.3.3 Estimación en dominios* del tema 3 del bloque "Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas"del grupo de materias comunes, trabajar con variables indicadoras de dominio. Para el dominio d-ésimo, definimos

$$z_{dk} = \begin{cases} 1 & \text{si } k \in U_d, \\ 0 & \text{en caso contrario.} \end{cases} \tag{8.3}$$

El vector

$$z_k = (z_{1k}, ..., z_{dk}, ..., z_{Dk})$$

consiste en $D-1$ valores "0"y un único valor "1"especificando el dominio del k-ésimo elemento. Los z_{dk} son constantes, no variables aleatorias, y a menudo son desconocidos antes del muestreo.

El tamaño del dominio se puede expresar como

$$N_d = \sum_U z_{dk}.$$

El estimador de Horvitz-Thompson (por tanto insesgado) de N_d viene dado por

$$\widehat{N}_d^{\text{HT}} = \sum_d \frac{z_{dk}}{\pi_k} = \sum_{s_d} \frac{1}{\pi_k}.$$

Por otro lado, el tamaño muestral del dominio se puede representar como

$$n_{s_d} = \sum_U z_{dk} I_k = \sum_{U_d} I_k,$$

donde I_k es la variable indicadora de pertenencia a la muestra, es decir, $I_k = 1$ si $k \in s$ e $I_k = 0$ en caso contrario. Bajo el diseño muestral dado $p(\cdot)$, el tamaño muestral del dominio esperado es por tanto

$$\mathbb{E}(n_{s_d}) = \sum_U z_{dk}\pi_k = \sum_{U_d} \pi_k. \tag{8.4}$$

De esta forma, el diseño, o más específicamente los π_k inducidos por el diseño, determinan si se puede esperar un tamaño muestral bueno o malo en el dominio.

En particular, bajo un diseño aleatorio simple sin reemplazamiento, con n elementos elegidos a partir de N,

$$\mathbb{E}_{srswor}(n_{s_d}) = \frac{N_d n}{N} = f P_d N \tag{8.5}$$

donde $f = \frac{n}{N}$ y

$$P_d = \frac{N_d}{N}$$

es el tamaño relativo del dominio. En la ecuación (8.5), dos fracciones normalmente pequeñas, f y P_d, son multiplicadas por un número normalmente grande, el tamaño poblacional N. Cambios menores en los tres términos, f, P_d y N pueden causar cambios considerables en $\mathbb{E}_{srswor}(n_{s_d})$. Por ejemplo, si $P_d = 0{,}3\,\%$, $f = 1\,\%$ y $N = 100000$, tenemos

$$\mathbb{E}_{srswor}(n_{s_d}) = 3$$

lo que obviamente ofrece una mala perspectiva para obtener una buenas estimación. Si aumentamos cada uno de los tres valores, las perspectivas son mucho mejores. Si $P_d = 0{,}9\,\%$, $f = 2\,\%$ y $N = 400000$, tenemos

$$\mathbb{E}_{srswor}(n_{s_d}) = 72.$$

Así, si es posible disponer de alguna información auxiliar de peso, un número de observaciones cercano a 70 puede proporcionar una estimación bastante precisa del dominio.

Purcell y Kish 1979 señalaron que la elección del método de estimación dependerá de la situación, en función del tamaño relativo del dominio, P_d. Introducen cuatro tipos de dominios calificados como *mayor*, *menor*, *mini* y *raro* dependiendo de si el tamaño relativo $P_d = \frac{N_d}{N}$ satisface, respectivamente, $P_d \geq 0{,}1$, $0{,}01 \leq P_d < 0{,}1$, $0{,}0001 \leq P_d < 0{,}01$, y $P_d < 0{,}0001$. Para un dominio mini o raro, incluso una muestra total muy grande puede que no incluya ninguna observación en ese dominio. En tales casos puede ser necesario usar métodos de estimación especiales, que no consideraremos en este tema.

Comentario 44. El problema de la estimación en dominios está relacionado con los marcos imperfectos. Supongamos que el marco poblacional U_F tiene un único problema que es la sobrecobertura. Entonces, la población marco U es un subconjunto (un dominio) de U_F. Por ejemplo, para una población de establecimientos de empresas, el marco puede estar un poco desactualizado y la sobrecobertura $U_F - U$ se corresponde con empresas que han cerrado desde que el marco se elaboró.

Comentario 45. Si las proporciones de los dominios P_d ($d = 1, ..., D$) son muy variables, los tamaños muestrales de los dominios $n_{s_1}, ..., n_{s_D}$ también pueden ser significativamente diferentes. Es posible entonces que en muchos dominios, se obtengan estimaciones muy precisas, mientras que en los dominios más pequeños las estimaciones sean pobres. Esto sugiere un enfoque en dos fases. En primer lugar la muestra grande de primera fase s se divide en las partes de los dominios s_d, y a continuación tiene lugar el submuestreo, con pequeñas tasas de submuestreo en los dominios grandes, y cerca del $100\,\%$ de submuestra en los dominios más pequeños.

8.2 Los métodos básicos de estimación en dominios

Examinemos en primer lugar las técnicas básicas para la estimación del total del dominio

$$Y_d = \sum_{U_d} y_k$$

de la media del dominio

$$\bar{y}_{U_d} = \frac{Y_d}{N_d} = \frac{1}{N_d} \sum_{U_d} y_k$$

y, al final de la sección, la diferencia entre las medias de dos dominios.

Sea

$$y_{dk} = z_{dk} y_k = \begin{cases} y_k & \text{si } k \in U_d, \\ 0 & \text{en caso contrario,} \end{cases} \tag{8.6}$$

donde z_{dk} es la variable indicadores definida por (8.3). Entonces, el total del dominio es

$$Y_d = \sum_{U_d} y_{dk}$$

y el estimador de Horvitz-Thompson correspondiente es

$$\widehat{Y}_d^{\text{HT}} = \sum_{s} \check{y}_{dk} = \sum_{s_d} \check{y}_k.$$

Cuando N_d es conocido, un estimador mejor de Y_d se obtiene usando los resultados de la sección *3.3.2 Estimadores, varianza y estimador de la varianza* del tema 3 del bloque "Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas"del grupo de materias comunes. Consideremos la media muestral ponderada por π del dominio

$$\tilde{y}_{s_d} = \frac{1}{\widehat{N}_d^{\text{HT}}} \sum_{s_d} \check{y}_k \tag{8.7}$$

con

$$\widehat{N}_d^{\text{HT}} = \sum_{s_d} \frac{1}{\pi_k}.$$

Multiplicando $\tilde{y}_{s_d}$ por N_d obtenemos un estimador alternativo de t_d que denotamos por $\tilde{Y}_d$. Esto es,

$$\tilde{y}_d = N_d \tilde{y}_{s_d}.$$

Para ello es necesario que $n_{s_d} \geq 1$.

En resumen, las técnicas básicas para la estimación en dominios son:

1. Para estimar el total del dominio cuando N_d es desconocido, se usa el estimador de Horvitz-Thompson
$$\widehat{Y}_d^{\text{HT}} = \sum_{s_d} \check{y}_k.$$
Veremos sus propiedades en el Teorema 15.

2. Para estimar el total del dominio cuando N_d es conocido, usamos
$$\tilde{Y}_d = N_d \tilde{y}_{s_d} = \frac{N_d}{\widehat{N}_d^{\text{HT}}} \sum_{s_d} \check{y}_k$$
El Teorema 15 muestra sus propiedades fundamentales.

3. Para estimar la media del dominio, tanto si N_d es conocido como si no, usamos
$$\widehat{\bar{y}}_{U_d} = \tilde{y}_{s_d}$$
como se han descrito los resultados de la sección *3.3.2 Estimadores, varianza y estimador de la varianza* del tema 3 del bloque "Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas"del grupo de materias comunes.

Teorema 15

El estimador de Horvitz-Thompson del total del dominio $t_d = \sum_{U_d} y_k$ es
$$\widehat{Y}_d^{\text{HT}} = \sum_{s_d} \check{y}_k \tag{8.8}$$
con varianza
$$\mathbb{V}(\widehat{Y}_d^{\text{HT}}) = \sum \sum_{U_d} \delta_{kl} \check{y}_k \check{y}_l. \tag{8.9}$$
Un estimador insesgado de la varianza es
$$\widehat{\mathbb{V}}(\widehat{Y}_d^{\text{HT}}) = \sum \sum_{s_d} \check{\delta}_{kl} \check{y}_k \check{y}_l. \tag{8.10}$$
Cuando el tamaño del dominio N_d es conocido, un estimador más habitual es
$$\tilde{Y}_d = N_d \tilde{y}_{s_d}, \tag{8.11}$$
donde $\tilde{y}_{s_d}$ es la media ponderada por π dada por (8.7). La varianza aproximada es
$$AV(\tilde{y}_{s_d}) = \sum \sum_{U_d} \check{\delta}_{kl} \left(\frac{y_k - \bar{y}_{U_d}}{\pi_k} \right) \left(\frac{y_l - \bar{y}_{U_d}}{\pi_l} \right). \tag{8.12}$$

Un estimador de la varianza viene dado por

$$\widehat{\mathbb{V}}(\tilde{y}_{s_d}) = \left(\frac{N_d}{\widehat{N}_d^{\mathrm{HT}}}\right)^2 \sum\sum_{s_d} \check{\delta}_{kl} \left(\frac{y_k - \tilde{y}_{s_d}}{\pi_k}\right)\left(\frac{y_l - \tilde{y}_{s_d}}{\pi_l}\right). \tag{8.13}$$

Un intervalo de confianza construido de la forma habitual con la ayuda de (8.10) o de (8.13) da aproximadamente un cobertura del $100(1-\alpha)\,\%$ cuando se seleccionan repetidamente muestras s del total poblacional U, con el diseño dado $p(\cdot)$.

Aquí, las expresiones (8.9) y (8.10) se siguen fácilmente si aplicamos el teorema 4 del tema 2 [2] del bloque "Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas"del grupo de materias comunes a la variable del dominio y_d cuyo valor y_{dk} viene definido por (8.6). Las igualdades (8.12) y (8.13) se siguen de los resultados de la sección *3.3.2 Estimadores, varianza y estimador de la varianza* del tema 3 del bloque "Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas"del grupo de materias comunes, simplemente multiplicando por N_d^2. Cabe señalar que (8.12) se expresa por medio de diferencias a partir de la media del dominio, $y_k - \bar{y}_{U_d}$, mientras que (8.9) involucra los valores y_k brutos. En consecuencia, la varianza de $\tilde{Y}_d$ es generalmente más pequeña que la de $\widehat{Y}_d^{\mathrm{HT}}$. La mejora es más obvia si nos centramos en un diseño específico, como en el siguiente ejemplo.

Ejemplo 36. Bajo el diseño aleatorio simple sin reemplazamiento con n elementos seleccionados a partir de N, el estimador de Horvitz-Thompson de t_d es

$$\widehat{Y}_d^{\mathrm{HT}} = \left(\frac{N}{n}\right)\sum_{s_d} y_k.$$

Si $P_d = \frac{N_d}{N}$ y $Q_d = 1 - P_d$, la varianza mostrada en la ecuación (8.9) se puede expresar como

$$\begin{aligned} \mathbb{V}_{srswor}(\widehat{Y}_d^{\mathrm{HT}}) &= N^2 \frac{1-f}{n} \frac{(N_d - 1)S_{yU_d}^2 + N_d Q_d \bar{y}_{U_d}^2}{N-1} \\ &\doteq N^2 \frac{1-f}{n} P_d (S_{yU_d}^2 + Q_d \bar{y}_{U_d}^2), \end{aligned} \tag{8.14}$$

donde

$$S_{yU_d}^2 = \frac{1}{N_d - 1}\sum_{U_d}(y_k - \bar{y}_{U_d})^2$$

es la varianza del dominio. El estimador de la varianza en (8.10) tiene una expresión

[2] Tema 2. Ideas básicas sobre estimación en muestreo probabilístico. Diseño muestral. Probabilidades de inclusión. La noción de estadístico. Indicadores de pertenencia a la muestra. Estimadores y sus propiedades básicas. El estimador Horvitz-Thompson (estimador π) y sus propiedades. Muestreo con reemplazamiento. Efecto de diseño. Intervalos de confianza.

análoga,

$$\begin{aligned}\widehat{\mathbb{V}}_{srswor}(\widehat{Y}_d^{\mathrm{HT}}) &= N^2\frac{1-f}{n}\frac{(n_d-1)S_{ys_d}^2+n_dq_d\bar{y}_{s_d}^2}{n-1}\\ &\doteq N^2\frac{1-f}{n}p_d(S_{ys_d}^2+q_d\bar{y}_{s_d}^2),\end{aligned} \tag{8.15}$$

donde

$$S_{ys_d}^2=\frac{1}{n_{s_d}-1}\sum_{s_d}(y_k-\bar{y}_{s_d})^2 \tag{8.16}$$

$$\bar{y}_{s_d}=\sum_{s_d}\frac{y_k}{n_{s_d}},\quad p_d=\frac{n_{s_d}}{n},\quad \text{y}\quad q_d=1-p_d.$$

Volviendo ahora al estimador $\tilde{y}_d$, que requiere conocer el tamaño del dominio, obtenemos a partir de la ecuación (8.12)

$$\begin{aligned}AV_{srswor}(\tilde{y}_d) &= N^2\frac{1-f}{n}\frac{(N_d-1)S_{yU_d}^2}{N-1}\\ &\doteq N^2\frac{1-f}{n}P_dS_{yU_d}^2.\end{aligned} \tag{8.17}$$

mientras que el estimador de la varianza dado en (8.13) toma la forma

$$\begin{aligned}\widehat{\mathbb{V}}_{srswor}(\tilde{y}_d) &= \left(\frac{N_d}{\widehat{N}_d^{\mathrm{HT}}}\right)^2N^2\sum_{h=1}^{H}\frac{(n_{s_d}-1)S_{ys_d}^2}{n-1}\\ &\doteq N^2\frac{1-f}{n}\left(\frac{1}{n_{s_d}}-\frac{1}{\widehat{N}_d^{\mathrm{HT}}}\right)S_{ys_d}^2,\end{aligned} \tag{8.18}$$

con $\widehat{N}_d^{\mathrm{HT}}=N\frac{n_{s_d}}{n}$. Obviamente, el cálculo de (8.18) requiere que $n_{s_d}\geq 2$.

El incremento en la varianza debido a la falta de conocimiento de N_d se puede expresar mediante el cociente de las ecuaciones (8.14) y (8.17),

$$\frac{\mathbb{V}_{srswor}(\widehat{Y}_d^{\mathrm{HT}})}{AV_{srswor}(\tilde{y}_d)}\doteq 1+\frac{Q_d}{(cv_{yU_d})^2}, \tag{8.19}$$

donde $cv_{yU_d}=\frac{S_{yU_d}}{\bar{y}_{U_d}}$ es el coeficiente de variación de y en el d-ésimo dominio. Por ejemplo, si $cv_{yU_d}=0{,}5$, la varianza de $\widehat{Y}_d^{\mathrm{HT}}$ es aproximadamente cinco veces la de $\tilde{Y}_d$ cuando el dominio cuenta sólo con un pequeño porcentaje de la población (Q_d cerca de la unidad). Si el dominio cuenta con el $50\,\%$ de la población (con cv_{yU_d} invariante en $0{,}5$) la varianza de $\widehat{Y}_d^{\mathrm{HT}}$ no crece tanto con respecto a la de $\tilde{Y}_d$, pero aún así sigue siendo mayor; el cociente de varianzas decrece a tres aproximadamente.

Ejemplo 37. Supongamos que el dominio se puede identificar previamente y que podemos muestrear de forma controlada. ¿Obtendríamos entonces un estimador mejor que (8.11)? La intuición podría indicar que sí, para una respuesta más completa en un caso específico, consideremos el diseño aleatorio simple sin reemplazamiento. Si se puede identificar el dominio previamente, el estadístico puede seleccionar una muestra aleatoria simple sin reemplazamiento de n_d (ahora un número fijo) a partir de los N_d elementos del dominio. El estimador insesgado de t_d es entonces

$$N_d \bar{y}_{s_d} = N_d \sum_{s_d} \frac{y_k}{n_d},$$

con varianza exacta

$$\mathbb{V}'_{srswor}(N_d \bar{y}_{s_d}) = N_d^2 \left(\frac{1}{n_d} - \frac{1}{N_d} \right) S^2_{yU_d}. \tag{8.20}$$

Comparemos esto con el enfoque de la estimación del dominio no controlada. La expresión de la AV mostrada en (8.17) se puede escribir

$$AV_{srswor}(\tilde{y}_d) \doteq N_d^2 \left(\frac{1}{n_d^0} - \frac{1}{N_d} \right) S^2_{yU_d}, \tag{8.21}$$

donde

$$n_d^0 = n \frac{N_d}{N}$$

es el total muestral esperado en el dominio.

Es decir, si el tamaño muestral n_d bajo condiciones controladas es igual al total muestral esperado en el dominio $\mathbb{E}(n_{s_d}) = n_d^0$ bajo condiciones sin controlar, las dos varianzas serán aproximadamente la misma, que es un resultado bastante intuitivo. Por tanto, al grado de aproximación usado aquí no hay pérdida de precisión usando el enfoque de estimación en dominios teniendo en cuenta que N_d es conocido. Para obtener una aproximación más cercana a $AV_{srswor}(\tilde{y}_d)$, de forma que se pierda menos precisión, ver la Sección 8.3.

Ejemplo 38. Consideremos el diseño aleatorio simple sin reemplazamiento, con H estratos que tienen unidades en los D dominios. El estimador de la media del dominio mostrado en la ecuación (8.7) es entonces

$$\widehat{\bar{y}}_{U_d} = \tilde{y}_{s_d} = \sum_{h=1}^{H} \frac{N_h}{n_h} \frac{\sum_{s_{dh}} y_k}{\sum_{h=1}^{H} \frac{N_h}{n_h} n_{s_{dh}}}$$

donde $\frac{N_h}{n_h}$ es el inverso de la fracción de muestreo en el estrato h, s_{dh} es la intersección del dominio U_d y la muestra aleatoria simple seleccionada en el estrato h, y $n_{s_{dh}}$ es el tamaño aleatorio de esta intersección. La varianza estimada (8.13) dividida por N_d^2 se puede escribir, después de algunos cálculos, como

$$\widehat{\mathbb{V}}_{STSI}(\widehat{\bar{y}}_{U_d}) = \frac{1}{(\widehat{N}_d^{\mathrm{HT}})^2} \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} \frac{\sum_{s_{dh}} (y_k - \bar{y}_{s_{dh}})^2 + n_{s_{dh}} (1 - p_{dh}) (\bar{y}_{s_{dh}} - \widehat{\bar{y}}_{U_d})^2}{n_h - 1}$$

donde

$$p_{dh} = \frac{n_{s_{dh}}}{n_h}, \quad \widehat{N}_d^{\text{HT}} = \sum_{h=1}^{H} N_h \left(\frac{n_{s_{dh}}}{n_h} \right)$$

e $\bar{y}_{s_{dh}}$ es la media directa de y en la celda dh.

Los términos en $(\bar{y}_{s_{dh}} - \widehat{\bar{y}}_{U_d})^2$ de la expresión resultan molestos ya que suponen hacer las diferencias entre las medias de los estratos en el dominio. Una estratificación que sea altamente eficiente para estimar la población total o un dominio grande puede dejar de serlo para estimaciones en dominios pequeños, como se puede ver en Durbin 1958.

En muchas aplicaciones, es de interés analizar si dos dominios son diferentes. Si la población muestral consiste en individuos, nos puede interesar comparar, por ejemplo, los hombre y las mujeres, o el grupo de edad de los mayores de 65 con el grupo de edad entre 30 y 50, o los residentes en áreas rurales con los residentes en áreas urbanas. Es de particular importancia la comparación entre las medias de dos dominios. Un científico social o un político a menudo puede estar interesados no sólo en un estimación puntual de la diferencia entre entre la media de dos dominios, sino también en un valor que indique si la diferencia es significativamente grande. Por tanto, es de interés obtener un intervalo de confianza para la diferencia desconocida de dos medias. Si el valor cero no está contenido en el intervalo a un nivel del $95\,\%$, hay evidencias suficientes para rechazar la hipótesis de que los dominios sean iguales.

Si consideramos los dos dominios a comparar U_1 y U_2 (es decir, $d = 1$ y $d = 2$), la estimación de la diferencia vendrá dada por

$$D = \bar{y}_{U_1} - \bar{y}_{U_2} = \frac{t_1}{N_1} - \frac{t_2}{N_2}.$$

Un estimador natural es la diferencia de las medias muestrales ponderadas por π correspondientes,

$$\widehat{D} = \tilde{y}_{s_1} - \tilde{y}_{s_2}$$

con $\tilde{y}_{s_d}$ dado en (8.7) para $d = 1, 2$. La varianza es

$$\mathbb{V}(\widehat{D}) = \mathbb{V}(\tilde{y}_{s_1}) + \mathbb{V}(\tilde{y}_{s_2}) - 2\mathbb{C}(\tilde{y}_{s_1}, \tilde{y}_{s_2}).$$

Se pueden obtener aproximaciones de las dos varianza del lado derecho de la ecuación dividiendo la expresión en la ecuación (8.12) por N_d^2, $d = 1, 2$, mientras que una expresión aproximada para la covarianza correspondiente se calcula como

$$AC(\tilde{y}_{s_1}, \tilde{y}_{s_2}) = \frac{1}{N_1 N_2} \sum_{k \in U_1} \sum_{l \in U_2} \Delta_{kl} \frac{y_k - \bar{y}_{U_1}}{\pi_k} \frac{y_l - \bar{y}_{U_2}}{\pi_l}. \tag{8.22}$$

La varianza aproximada resultante de $\widehat{D}$ se puede expresar como

$$AV(\widehat{D}) = \sum_{U_1 \cup U_2} \sum \Delta_{kl} \check{A}_k \check{A}_l, \tag{8.23}$$

con $\check{A}_k = \frac{A_k}{\pi_k}$, donde

$$A_k = \frac{z_{1k}(y_k - \bar{y}_{U_1})}{N_1} - \frac{z_{2k}(y_k - \bar{y}_{U_2})}{N_2}.$$

El estimador de la varianza correspondiente, necesario para el cálculo del intervalo de confianza, es

$$\widehat{\mathbb{V}}(\widehat{D}) = \sum_{s_1 \cup s_2} \sum \check{\Delta}_{kl} \check{a}_k \check{a}_l. \tag{8.24}$$

con $\check{a}_k = \frac{a_k}{\pi_k}$, donde

$$a_k = \frac{z_{1k}(y_k - \tilde{y}_{s_1})}{\widehat{N}_1} - \frac{z_{2k}(y_k - \tilde{y}_{s_2})}{\widehat{N}_2}.$$

Ejemplo 39. Se puede comprobar que la covarianza aproximada en (8.22) es cero para un diseño tal que π_k sea constante para todos los k y Δ_{kl} sea constante para todos los $k \neq l$. Si consideramos el diseño muestreo aleatorio simple sin reemplazamiento (n elementos de N), tenemos

$$\widehat{D} = \bar{y}_{s_1} - \bar{y}_{s_2},$$

con $\bar{y}_{s_d} = \sum_{s_d} \frac{y_k}{n_{s_d}}$, $d = 1, 2$. En ausencia del término de covarianza, tenemos a partir de (8.21),

$$AV_{srswor}(\widehat{D}) \doteq \sum_{d=1}^{2} \left(\frac{1}{n_d^0} - \frac{1}{N_d} \right) S_{yU_d}^2$$

y el estimador de la varianza mostrado en (8.24) es en consecuencia aditivo,

$$\widehat{\mathbb{V}}_{srswor}(\widehat{D}) = \sum_{d=1}^{2} \left(\frac{1}{n_{s_d}} - \frac{1}{\widehat{N}_d^{\mathrm{HT}}} \right) S_{ys_d}^2$$

donde $\widehat{N}_d^{\mathrm{HT}} = N \frac{n_{s_d}}{n}$, y donde hemos aproximado $\frac{n(n_{s_d}-1)}{n_{s_d}(n-1)}$ por la unidad.

Existen técnicas para mejorar el estimador de la diferencia de las medias de dos dominios usando información auxiliar.

8.3 Condicionamiento sobre el tamaño muestral del dominio

En el Ejemplo 37 se ha visto que la estimación en dominios, cuando se usa el muestreo aleatorio simple sin reemplazamiento a partir de toda la población y $\tilde{Y}_d = N_d \tilde{y}_{s_d}$, es esencialmente tan eficiente como el muestreo muestreo aleatorio simple sin reemplazamiento directo dentro de un dominio a priori identificado. Cuando se desconoce el tamaño muestral del dominio se hace necesaria una aproximación de la varianza mejor que (8.17) para poder apreciar esta pérdida limitada de eficiencia. Con este fin, es útil en condicionamiento sobre n_{s_d}.

Denotaremos por A_d el evento $n_{s_d} \geq 1$. Si n es muy grande, $\mathbb{P}(A_d)$ es cercano a la unidad, incluso si el tamaño relativo del dominio P_d es bastante pequeño. Para un valor fijo

de n_{s_d} tal que $n_{s_d} \geq 1$, la parte del dominio de la muestra, $s_d = s \cap U_d$, se comporta como una selección aleatoria simple de n_{s_d} a partir de N_d. Para $\tilde{Y}_d = N_d \bar{y}_{s_d}$ tenemos, por tanto,

$$\begin{cases} \mathbb{E}_{srswor}(\tilde{Y}_d | A_d, n_{s_d}) = t_d & (8.25) \\ \mathbb{V}_{srswor}(\tilde{Y}_d | A_d, n_{s_d}) = N_d^2 \left(\frac{1}{n_{s_d}} - \frac{1}{\widehat{N}_d^{\text{HT}}} \right) S_{yU_d}^2 & (8.26) \end{cases}$$

Así, $\tilde{Y}_d$ es condicionalmente insesgado, dado cualquier tamaño muestral en el dominio tal que $n_{s_d} \geq 1$, y la varianza condicional viene dada por (8.26). Calculando la media sobre todos los valores $n_{s_d} \geq 1$,

$$\begin{cases} \mathbb{E}_{srswor}(\tilde{Y}_d | A_d) = t_d & (8.27) \\ \mathbb{V}_{srswor}(\tilde{Y}_d | A_d) = N_d^2 \left[\mathbb{E}_{srswor} \left(\frac{1}{n_{s_d}} | A_d \right) - \frac{1}{\widehat{N}_d^{\text{HT}}} \right] S_{yU_d}^2. & (8.28) \end{cases}$$

Para obtener (8.28) hemos usado

$$\mathbb{V}[\mathbb{E}_{srswor}(\tilde{Y}_d | A_d, n_{s_d})] = \mathbb{V}(t_d | A_d) = 0,$$

donde $\mathbb{V}(\cdot)$ denota la varianza con respecto a la distribución de n_{s_d}. En otras palabras, dado que el dominio contiene al menos una observación, $\tilde{Y}_d$ es insesgado para Y_d bajo el muestreo aleatorio simple sin reemplazamiento a partir de U.

Los momentos (8.27) y (8.28) todavía tienen una condición, concretamente, el evento A_d. Suponiendo que n es suficientemente grande como para que A_d ocurra casi seguro, concluimos, a partir de (8.27) y de (8.28), que $\tilde{Y}_d$ es insesgado para Y_d con la varianza sin condicionar.

$$\mathbb{V}_{srswor}(\tilde{Y}_d) = N_d^2 \left[\mathbb{E}_{srswor} \left(\frac{1}{n_{s_d}} \right) - \frac{1}{\widehat{N}_d^{\text{HT}}} \right] S_{yU_d}^2 \tag{8.29}$$

Al hacer uso del signo de igualdad estamos haciendo un abuso de notación: la ecuación (8.29) es válida bajo la hipótesis de que $n_{s_d} = 0$ con probabilidad despreciable.

Mediante el desarrollo en serie de Taylor obtenemos

$$\mathbb{E}_{srswor} \left(\frac{1}{n_{s_d}} \right) \doteq \frac{1}{n_d^0} + \frac{(1-f)(1-P_d)}{(n_d^0)^2} \tag{8.30}$$

donde $f = \frac{n}{N}$, y

$$n_d^0 = \mathbb{E}(n_{s_d}) = n \frac{N_d}{N} = n P_d.$$

A partir de (8.29) y de (8.30), y asumiendo $Q_d = 1 - P_d$

$$\mathbb{V}_{srswor}(\tilde{Y}_d) = N_d^2 \left(\frac{1}{n_d^0} - \frac{1}{N_d}\right)\left(1 + \frac{Q_d}{n_d^0}\right) S_{yU_d}^2. \tag{8.31}$$

Comparando la varianza muestral sin controlar (8.31) con la varianza muestral controlada (8.20) obtenemos

$$\frac{\mathbb{V}_{srswor}(\tilde{Y}_d)}{\mathbb{V}'_{srswor}(N_d \bar{y}_{s_d})} = 1 + \frac{Q_d}{n_d^0}$$

que es aproximadamente $1 + \frac{1}{n_d^0}$ si n es mucho más grande que $n_d^0 = nP_d$. Por ejemplo, si el tamaño muestral del dominio esperado es $n_d^0 = 10$ elementos, la varianza aumenta aproximadamente un $10\,\%$. En otras palabras, hay una pérdida de precisión no despreciable causada por la pérdida de control del tamaño muestral del dominio.

Comentario 46. La varianza condicional mostrada en (8.26) se estima insesgadamente (dado n_{s_d} tal que $n_{s_d} \geq 2$) por

$$\widehat{V}^*_{srswor} = N_d^2 \left(\frac{1}{n_{s_d}} - \frac{1}{N_d}\right) S_{ys_d}^2$$

donde $S_{ys_d}^2$ viene dado por (8.16). Este estimador de la varianza condicional en su esencia coincide con la expresión (8.18). La diferencia entre $\frac{1}{\widehat{N}_d^{\mathrm{HT}}}$ y $\frac{1}{N_d}$ es insignificante, con lo que no tiene consecuencias significativas en la práctica.

8.4 Dominios pequeños: estimadores sintéticos

Se pueden obtener estimadores mejorados en los dominios si se dispone de más información auxiliar (más allá de conocer N_d). Una posibilidad es usar un método de estimación de regresión. Otra es usar un estimador de razón por dominio.

En el caso del estimador de regresión, denotaremos por $\mathbf{x}_k$ el valor del k-ésimo elemento del vector auxiliar $\mathbf{x}$, de dimensión, por ejemplo, J. A continuación se ajustará un modelo que describa la presunta relación entre la variable de estudio y las variables auxiliares. Los valores ajustados $\widehat{y}_k$ se usan para construir un estimador de regresión adecuado. Los valores resultantes predichos son

$$\widehat{y}_k = \mathbf{x}_k^{'} \widehat{\mathbf{B}}$$

donde

$$\widehat{\mathbf{B}} = \left(\sum_s \frac{\mathbf{x}_k \mathbf{x}_k^{'}}{\sigma_k^2 \pi_k}\right)^{-1} \sum_s \frac{\mathbf{x}_k \mathbf{x}_k^{'}}{\sigma_k^2 \pi_k}$$

y los residuos vienen dados por

$$e_{ks} = y_k - \widehat{y}_k = y_k - \mathbf{x}_k^{'} \widehat{\mathbf{B}}.$$

Teorema 16

Asumiendo que el tamaño del dominio N_d es conocido, el estimador de regresión del total del dominio $t_d = \sum_{U_d} y_k$ es

$$\widehat{Y}_{dr} = \sum_{U_d} \widehat{y}_k + \left(\frac{N_d}{\widehat{N}_d^{\text{HT}}}\right) \sum_{s_d} \check{e}_{ks} = \sum_{s} g_{dks} \check{y}_k \tag{8.32}$$

donde

$$\check{e}_{ks} = \frac{e_{ks}}{\pi_k} = \frac{(y_k - \widehat{y}_k)}{\pi_k}, \quad \check{y}_k = \frac{y_k}{\pi_k}$$

y las ponderaciones g (que dependen del dominio d, de la muestra total s, y del elemento k) son

$$g_{dks} = \frac{N_d}{\widehat{N}_d^{\text{HT}}} z_{dk} + \left(\sum_{U_d} \mathbf{x}_k - \frac{N_d}{\widehat{N}_d^{\text{HT}}} \sum_{s_d} \check{\mathbf{x}}_k\right)' \left(\sum_{s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2 \pi_k}\right)^{-1} \frac{\mathbf{x}_k}{\sigma_k^2},$$

con z_{dk} el indicador del dominio dado por (8.3).

La varianza aproximada vendrá dada por

$$AV(\widehat{Y}_{dr}) = \sum \sum_{U_d} \delta_{kl} \left(\frac{E_k - \bar{E}_{U_d}}{\pi_k}\right) \left(\frac{E_l - \bar{E}_{U_d}}{\pi_l}\right) \tag{8.33}$$

donde $E_k = y_k - \mathbf{x}_k' \widehat{\mathbf{B}}$ son los residuos ajustados poblacionales, con

$$\mathbf{B} = \left(\sum_{U} \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2}\right)^{-1} \sum_{U} \frac{\mathbf{x}_k y_k}{\sigma_k^2},$$

y $\bar{E}_{U_d} = \sum_{U_d} \frac{E_k}{N_d}$. Un estimador de la varianza es

$$\widehat{\mathbb{V}}(\widehat{Y}_{dr}) = \sum \sum_{s} \check{\delta}_{kl} \frac{g_{dks} e_{ks}}{\pi_k} \frac{g_{dls} e_{ls}}{\pi_l}. \tag{8.34}$$

Un tercer enfoque es modelizar en términos de grupos de elemento de la población que se perciben como homogéneos. Esto lleva a un estimador de dominio postestratificado. La idea es que los grupos son un factor de peso para explicar la varianza de la variable y, mientras que los dominios quizá no lo son. Por ejemplo, los grupos por edad/sexo a menudo explicarán una buena parte de la variación entre individuos, pero una partición en dominios geográficos puede resultar, en comparación, un factor explicativo débil.

Particionamos la población en G grupos que se denotan $U_{\cdot 1}, ..., U_{\cdot g}, ..., U_{\cdot G}$. Los tamaños, normalmente desconocidos, se denotan por $N_{\cdot 1}, ..., N_{\cdot g}, ..., N_{\cdot G}$. Consideraremos el caso

en que los G grupos poblacionales intesecan los D dominios para formar una matriz de DG celdas denotadas por U_{dg}, $d = 1, ..., D$; $g = 1, .., G$. Sea N_{dg} el tamaño de U_{dg}. Para una mayor claridad, los dominios se denotan ahora por $U_{d\cdot}$, $d = 1, ..., D$, y sus respectivos tamaños por $N_{d\cdot}$, $d = 1, ..., D$.

Si aplicamos el Teorema 16 a esta metodología tenemos el estimador de dominio postestratificado

$$\widehat{Y}_{dpos} = \sum_{g=1}^{G} N_{dg} \tilde{y}_{s_{dg}}, \tag{8.35}$$

donde los totales poblacionales N_{dg} tienen que ser conocidos.

Este modelo contiene, sin embargo, DG parámetros β a estimar, posiblemente un número muy grande. Esto puede causar dificultades, incluyendo la imposibilidad de calcular algún estimador cuando los totales de algunas celdas $n_{s_{dg}}$ ($g = 1, ..., G$) son cero. A menudo es sensato reducir el número de parámetros del modelo; una opción es trabajar con un modelo expresado únicamente en términos de los efectos del grupo. Esto estabiliza las estimaciones de los parámetros y causa poca pérdida de eficiencia si los grupos (más que los dominios) son el principal factor para explicar la variación en y. Como se ha mencionado anteriormente, los grupos por edad/sexo pueden ser un factor dominante para explicar la variación en los individuos, mientras que los dominios definidos geográficamente pueden ser, comparativamente, un factor explicativo débil.

Esta simplificación del modelo pretende limitar al máximo (aunque sin pérdida de información auxiliar valiosa) el número de parámetros en el modelo usando para generar el estimador de la media o el total del dominio.

Consideremos entonces el modelo de grupo tal que, para $g = 1, ..., G$,

$$\begin{cases} \mathbb{E}_\xi(y_k) & = \beta_g \\ \mathbb{V}_\xi(y_k) & = \sigma_g^2 \end{cases} \tag{8.36}$$

para todos los $k \in U_g$. En este modelo ANOVA unidireccional, $U_{\cdot 1}, ..., U_{\cdot g}, ..., U_{\cdot G}$ son los grupos poblacionales indicados anteriormente.

A partir del Teorema 16, podemos obtener el estimador aproximadamente insesgado del total del dominio t_d generado por este modelo más parsimonioso

$$\widehat{Y}_{dr} = \sum_{g=1}^{G} N_{dg} \tilde{y}_{s_{\cdot g}} + \left(\frac{N_{d\cdot}}{\widehat{N}_{d\cdot}^{\text{HT}}} \right) \sum_{g=1}^{G} \widehat{N}_{dg}^{\text{HT}} (\tilde{y}_{s_{dg}} - \tilde{y}_{s_{\cdot g}}), \tag{8.37}$$

con

$$\tilde{y}_{s_{\cdot g}} = \sum_{s_{\cdot g}} \frac{\check{y}_k}{\widehat{N}_{\cdot g}^{\text{HT}}}; \quad \tilde{y}_{s_{dg}} = \sum_{s_{dg}} \frac{\check{y}_k}{\widehat{N}_{dg}^{\text{HT}}}$$

y

$$\widehat{N}_{\cdot g}^{\mathrm{HT}} = \sum_{s_{\cdot g}} \frac{1}{\pi_k}; \quad \widehat{N}_{dg}^{\mathrm{HT}} = \sum_{s_{dg}} \frac{1}{\pi_k}; \quad \widehat{N}_{d\cdot}^{\mathrm{HT}} = \sum_{s_{d\cdot}} \frac{1}{\pi_k}.$$

Las cantidades N_{dg} deben proceder de fuentes precisas (registros administrativos, etc.). Cabe señalar que el término

$$\sum_{g=1}^{G} \frac{\widehat{N}_{dg}^{\mathrm{HT}} \tilde{y}_{s_{dg}}}{\widehat{N}_{d\cdot}^{\mathrm{HT}}} = \sum = \sum_{s_{d\cdot}} \frac{\check{y}_k}{\widehat{N}_{d\cdot}^{\mathrm{HT}}} = \tilde{y}_{s_{d\cdot}}$$

no conlleva cálculos difíciles en tanto en cuanto el dominio tenga por lo menos una observación. Por tanto, podemos escribir el estimador (8.37) como

$$\widehat{Y}_{dr} = N_{d\cdot} \left[\tilde{y}_{s_{d\cdot}} + \sum_{g=1}^{G} \left(\frac{N_{dg}}{N_{d\cdot}} - \frac{\widehat{N}_{dg}^{\mathrm{HT}}}{\widehat{N}_{d\cdot}^{\mathrm{HT}}} \right) \tilde{y}_{s\cdot g} \right].$$

Los residuos para la fórmula de la AV mostrada en la ecuación (8.33) son en este caso

$$E_k = y_k - \bar{y}_{U\cdot g} \tag{8.38}$$

para $k \in U_{dg}$, donde $\bar{y}_{U\cdot g}$ es la media de y en $U_{\cdot g}$. Las cantidades necesarias para el estimador de la varianza (8.34), para $k \in s_{dg}$, vienen dadas por

$$e_{ks} = y_k - \tilde{y}_{s\cdot g}$$

y

$$g_{dks} = N_{d\cdot} \left\{ \frac{z_{dk}}{\widehat{N}_{d\cdot}^{\mathrm{HT}}} + \left(\frac{N_{dg}}{N_{d\cdot}} - \frac{\widehat{N}_{dg}^{\mathrm{HT}}}{\widehat{N}_{d\cdot}^{\mathrm{HT}}} \right) \frac{1}{\widehat{N}_{\cdot g}^{\mathrm{HT}}} \right\}.$$

Comentario 47. Resulta de interés comparar la eficiencia del estimador (8.37) con el estimador postestratificado dado por (8.35). Este último debería de tener una varianza menor si hay muchos datos y existen diferencias entre dominios así como diferencias entre grupos. Los residuos (8.38) que determinan la varianza aproximada del estimador (8.37) se puede escribir como

$$E_k = E_k^* + (\bar{y}_{U_{dg}} - \bar{y}_{U\cdot g})$$

donde la barra denota una media directa sobre los conjuntos indicados, y

$$E_k^* = y_k - \bar{y}_{U_{dg}}$$

es el residuo asociado con el estimador postestratificado (8.35). Un caso de especial interés se da cuando

$$\bar{y}_{U_{dg}} = \bar{y}_{U\cdot g} \tag{8.39}$$

para todos los grupos $g = 1, ..., G$. Esto indica que, en cada grupo, la media del grupo para todos los dominios es igual a la media del grupo para la población en conjunto, es decir, en un sentido, el dominio es como la población. Entonces se tiene la igualdad de

los residuos; $E_k = E_k^*$ para todos los k. En consecuencia, la varianza aproximada de los estimadores (8.35) y (8.37) son iguales, y ambos son casi insesgados. En la práctica, no se puede calcular (8.39) de forma exacta. Pero en presencia de efectos de grupo fuertes y sin efectos de dominio pronunciados, se pierde poca eficiencia usando el estimador obtenido a partir del modelos más sencillo en (8.36). Además, el estimador (8.37) tiene la ventaja de que su cálculo es posible aunque haya algunas celdas con totales nulos. En raras ocasiones, (8.37) puede dar lugar a estimaciones imposibles. Si y es siempre una variable no negativa, sólo se pueden tolerar estimaciones no negativas de Y_d, pero existe una remota posibilidad para dominios muy pequeños de que (8.37) sobrecorrija y dé lugar a estimaciones negativas.

Definición 12

El primer término del estimador (8.37),

$$\widehat{Y}_{dsy} = \sum_{g=1}^{G} N_{dg} \tilde{y}_{s\cdot g} \tag{8.40}$$

se puede considerar un estimador por sí mismo; sin embargo, es sesgado. Se llama *estimador sintético* y ha sido estudiado a fondo, especialmente para el diseño aleatorio simple sin reemplazamiento, en cuyo caso $\tilde{y}_{s\cdot g}$ se convierte en $\bar{y}_{s\cdot g}$, la media sin ponderar del grupo. Para más información véase Gonzalez 1973.

Se verifica que

$$\mathbb{E}(\widehat{Y}_{dsy}) \doteq \sum_{g=1}^{G} N_{dg} \bar{y}_{U\cdot g}$$

por lo tanto el sesgo de $\widehat{Y}_{dsy}$ como estimador de t_d es aproximadamente

$$\mathbb{B}(\widehat{Y}_{dsy}) = \mathbb{E}(\widehat{Y}_{dsy}) - t_d \doteq -\sum_{g=1}^{G} N_{dg}(\bar{y}_{U_{dg}} - \bar{y}_{U\cdot g}).$$

Si se cumple la igualdad dada en (8.39) para $g = 1, .., G$, entonces $\widehat{Y}_{dsy}$ es insesgado. Para que se verifique (8.39) en la práctica, sería necesario un golpe de suerte; generalmente, el sesgo desconocido no es cero, quizá pueda tomar un valor considerable, y un intervalo de confianza obtenido en torno a la estimación puntual no estaría centrado y sería inválido. Entonces, ¿por qué los estadísticos están interesados en el estimador (8.40)? La respuesta es que la varianza de $\widehat{Y}_{dsy}$ es a menudo extremadamente pequeña, y en la mayoría de los casos mucho más pequeña que la del estimador casi insesgado mostrado en la ecuación (8.37). Esto no es una sorpresa ya que las medias de los grupos $\tilde{y}_{s\cdot g}$ en (8.40) vienen determinadas con mucha precisión como resultado de tamaños muestrales de grupo considerables y quizá varianza pequeñas dentro de los grupos. Por otro lado,

el término de corrección del sesgo,

$$\left(\frac{N_{d\cdot}}{\widehat{N}_{d\cdot}}\right)\sum_{g=1}^{G}\widehat{N}_{dg}^{\mathrm{HT}}(\tilde{y}_{s_{dg}}-\tilde{y}_{s\cdot g})$$

aportará una varianza grande al estimador (8.37). Por tanto $\widehat{Y}_{dsy}$ será mucho más preciso (varianza baja), pero puede ser altamente inacurado (sesgo y MSE considerables). Usar $\widehat{Y}_{dsy}$ es arriesgado; el estadístico confía en tener un sesgo pequeño o despreciable. Si la hipótesis se verifica, el estimador sintético es interesante. Por eso el estimador sintético (8.40) no es recomendable excepto si el estadístico está dispuesto a correr el riesgo del valor que puedan tomar el sesgo y el intervalo de confianza.

Comentario 48. De forma general, un estimador sintético se define usando el término de predicción del estimador de regresión (8.32)

$$\widehat{Y}_{dsy}=\sum_{U}\widehat{y}_k=(\mathbf{x}_k)'\widehat{\mathbf{B}}$$

en donde se mejora el estimador al usar información externa al dominio al obtener las predicciones $\widehat{y}_k$. Por ejemplo, un estimador sintético obtenido al imponer el modelo de razón simple de un parámetro, en el que $\mathbf{x}_k=x_k$ y $\mathbf{B}=\beta$, sobre el dominio d viene dado por

$$\widehat{Y}_{dsy}=(\sum_{U_d}x_k)\widehat{B} \tag{8.41}$$

donde

$$\widehat{B}=\frac{\sum_s \check{y}_k}{\sum_s \check{x}_k}$$

Aquí, $\widehat{B}$ estima el parámetro de la pendiente

$$B=\frac{\sum_U y_k}{\sum_U x_k} \tag{8.42}$$

para la población en su conjunto. La varianza de $\widehat{B}$ es a menudo pequeña porque la estimación usa la muestra completa s. En consecuencia, la varianza del estimador (8.41) es pequeña, pero el sesgo es considerable si la razón del dominio

$$B_d=\frac{\sum_{U_d} y_k}{\sum_{U_d} x_k}$$

difiere considerablemente de la razón de la población en conjunto mostrada en la ecuación (8.42).

Se han construido un gran número de estimadores como combinación lineal ponderada de un término sintético sesgado y de varianza pequeña, y de un término que es un

estimador estándar insesgado de varianza grande, como la media muestral ponderada del dominio. Se pueden usar distintos principios estadísticos a la hora de construir las ponderaciones. Para más información véase R. Platek y C. Särndal 1987.

Los estimadores combinados normalmente tienen sesgo considerable en muestras pequeñas y medianas, por lo que el MSE contiene un término bastante importante de sesgo al cuadrado. Por tanto, normalmente es imposible obtener un intervalo de confianza válido. Sin embargo, el MSE puede ser más pequeño que el de un estimador insesgado. Además, la ponderación de los términos se puede hacer de forma que el sesgo tienda a cero a medida que el tamaño muestral aumenta. Cuando el tamaño muestral del dominio es muy pequeño, se le puede dar un peso considerable al término sintético. De esta forma, a medida que el tamaño muetral aumenta, la ponderación se va desplazando hacia el estimador estándar, de forma que el estimador combinado sea consistente.

Bibliografía

Durbin, J. (1958). "Sampling theory for estimates based on fewer individual than the number selected". En: *Bulletin of the International Statistical Institute* 36, págs. 113-119.

Gonzalez, M.E. (1973). "Use and evaluation os synthetic estimates". En: *Proceeding of the Section on Survey Research, America Statistical Association*, págs. 33-36.

Purcell, N.J. y L. Kish (1979). "Estimation for small domains". En: *Biometrics* 35, págs. 365-384.

R. Platek M.P. Singh, J.N.K. Rao y C.E. Särndal (1987). *Small Area Statistics: An International Symposium*. New York: Wiley.

Rao, J.N.K. e I. Molina (2015). *Small area estimation*. New York: Wiley.

Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.

Tema 9

Reponderación de datos en presencia de falta de respuesta. Tratamientos tradicionales de la falta de respuesta. Vectores auxiliares e información auxiliar. El enfoque de calibrado. Estimación puntual bajo calibrado.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía.

C.-E. Särndal y S. Lundström (2005). *Estimation in Surveys with Nonresponse*. Chichester: Wiley

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

9.1 Tratamientos tradicionales de la falta de respuesta

La teoría basada en diseño, también llamada teoría de la aleatoriedad, es aplicable cuando la encuesta tiene el total de las respuestas (respuesta completa). Esta teoría de inferencia se basa en la aleatoriedad en la selección muestral. El diseño muestral puede ser el muestreo aleatorio estratificado, el muestreo por conglomerados, etc. Sea el que sea, cada elemento posee su propia probabilidad de inclusión π_k. Estas probabilidades juegan un papel decisivo en la inferencia estadística con respuesta completa.

Cuando la falta de respuesta entra en escena, hecho que es inevitable en la práctica, es conveniente pensar en cada elemento como si tuviese su propia probabilidad de respuesta. A diferencia de la fase muestral, la fase de respuesta está fuera del control de los estadísticos. La falta de respuesta ocurre con probabilidades desconocidas. Sin embargo, el concepto de probabilidades de respuesta individuales genera un marco productivo para los métodos de tratamiento de la falta de respuesta, como se verá en este tema.

Nuestra principal preocupación en este tema serán las encuestas afectadas por la falta

de respuesta. Es decir, los valores de una variable de estudio *y* se observan sólo para los *k* elementos de un subconjunto *r* de la muestra completa *s*. Llamaremos a *r* el *conjunto respuesta*. En este punto asumimos que y se ve afectada sólo por falta de respuesta total. Con cualquier técnica que se use, un estimador del total $Y = \sum_U y_k$ estará más o menos insesgado (excepto si la falta de respuesta tiene lugar de una forma totalmente aleatoria, lo que casi nunca ocurre). Por tanto el sesgo es inevitable. Buscamos métodos que limiten el sesgo lo más posible.

Antes de 1980, los métodos de ajuste de la falta de respuesta generalmente se basaban en un *modelo determinístico* de respuesta de la encuesta. Se asumía que la población finita consistía de dos partes disjuntas, un estrato con respuesta y otro sin respuesta. Cada elemento del primero respondía con certeza si era seleccionado en la muestra, y cada elemento del segundo estrato tenía probabilidad nula de responder. Una crítica obvia de este modelo es que es muy simplista y no realista. Más aún, los tamaños de los dos estratos a menudo no se podían asumir como conocidos, por lo que, a menudo, se estimaba el total para el estrato con respuesta y se añadía un término que compensase de alguna forma la contribución del estrato con falta de respuesta.

En los años 80 se hizo popular un método que resultó más satisfactorio. Este consideraba el conjunto respuesta *r* como el resultado de dos selecciones probabilísticas. La muestra *s* se selecciona en primer lugar a partir de la población *U*, entonces el conjunto respuesta *r* se obtenía como un subconjunto de *s*. Este es más realista y más general que el determinístico en el sentido de que permite que cada elemento *k* tenga su propia probabilidad de respuesta θ_k donde $0 \leq \theta_k \leq 1$ para todo k. Esta generalización tiene un precio: las probabilidades de respuesta θ_k son siempre desconocidas. Un método que se ha utilizado es reemplazar las θ_k por estimaciones basada en información auxiliar.

El que el elemento *k* responda o no es un resultado binario con probabilidad desconocida, como un lanzamiento de una moneda no equilibrada que puede dar cara o cruz con probabilidad desconocida. Para el elemento k, podemos definir una variable aleatoria binaria R_k con valor $R_k = 1$ si k responde y $R_k = 0$ si no lo hace. El valor esperado del indicador de respuesta R_k, condicionado a la muestra obtenida *s*, es $\mathbb{E}_q(R_k|s) = \theta_k$, donde $\mathbb{E}_q$ es la esperanza con respecto al mecanismo de respuesta $q(r|s)$. Asumiremos que la probabilidad de respuesta θ_k depende de *k* pero no de la muestra *s* a la cual *k* pertenece.

No está claro cómo se originó esta idea tan importante de asociar a cada elemento un número único, la probabilidad de respuesta (o proporcionar el valor solicitado y_k de la variable de estudio y). Para ver el cálculo de la esperanza y la varianza de estimadores tradicionales en este contexto, se puede consultar Lindström y Lundström 1974 y R. Platek y Tremblay 1978.

Hay una teoría que basa la estimación en torno a la idea de que el elemento *k* está equipado con una probabilidad de inclusión individual conocida, π_k, y una proba-

bilidad de respuesta individual desconocida, θ_k. Esto la convierte en una 'teoría de cuasi-aleatorización término adecuado acuñado por Oh y Scheuren 1978.

En torno a 1980 se realizó un importante trabajo por parte del *Panel on Incomplete Data*, promovido por *US National Science Foundation*. El resultado son tres volúmenes con contribuciones de distintos autores, algunas teóricas y algunas empíricas y prácticas. Varias contribuciones están en línea con la 'teoría de cuasi-aleatorización' se pueden consultar, por ejemplo, el capítulo de Platek y Gray 1983. Esta teoría ha servido como guía en las últimas décadas para los metodólogos en los institutos de estadística.

Otras de las contribuciones introducen y explican conceptos, bastantes nuevos en ese momento, tales como 'ignorado', en oposición a 'no ignorado', 'falta de respuesta' (o 'missingness'), *missing* aleatoriamente (MAR del inglés *missing at random*), completamente *missing* aleatoriamente (MCAR del inglés *missing completely at random*). Estos conceptos han contribuido, desde entonces, de forma significativa a entender la falta de respuesta y los datos missing. Para ver el enfoque Bayesiano, consultar Rubin 1987 y Little y Rubin 1987.

Otro manual que se puede consultar es Lundström y Särndal 2001, *a Current Best Methods manual* producido por y usado en *Statistics Sweden* para el tratamiento de los datos de encuestas afectados por falta de respuesta y deficiencias del marco.

La idea de la cuasi-aleatorización nos lleva al método de *reponderación en dos fases para la falta de respuesta*. La idea aquí es seleccionar una muestra s y luego obtener el un conjunto de respuesta r como un subconjunto de s. Särndal, Swensson y Wretman 1992 es un ejemplo de referencia reciente en la que se ha discutido el método de las dos fases para ajustar la falta de respuesta. El Capítulo 9 de ese libro trata sobre la formulación tradicional del muestreo bifásico, en ausencia de falta de respuesta. Consiste en lo siguiente: se selecciona una primera muestra de U, se observan determinadas variables útiles (aunque no la(s) variable(s) de estudio), a continuación se selecciona de la primera muestra una submuestra más pequeña, y se observan la(s) variable(s) de estudio en los elementos de la submuestra. En este contexto, todas las probabilidades de inclusión son conocidas, tanto las de la primera fase como las de la segunda. En el Capítulo 15 del libro, se adapta la teoría del muestreo bifásico a la situación en la que existe falta de respuesta. Comentaremos brevemente este método.

En primer lugar consideramos la hipótesis de que la distribución de la respuesta $q(r|s)$ es conocida (aunque en la práctica no sea el caso). Esto implica que las probabilidades de respuesta de primer y segundo orden, dadas por

$$\mathbb{P}(k \in r|s) = \theta_k, \quad \mathbb{P}(k \,\&\, \ell \in r|s) = \theta_{k\ell}, \tag{9.1}$$

son conocidas y se pueden usar para estimar $Y = \sum_U y_k$. Podríamos entonces calcular los pesos combinados $\frac{d_k}{\theta_k} = \frac{1}{\pi_k} \times \frac{1}{\theta_k}$, para $k \in r$, y usarlos para construir el estimador

bifásico insesgado

$$\widehat{Y} = \sum_r \frac{d_k}{\theta_k} y_k. \tag{9.2}$$

Esto extiende el estimador básico de Horvitz–Thompson a una selección en dos fases. Pero el θ_k siempre es desconocido en la práctica, por eso no se puede calcular (9.2). Para hacerlo operativo debemos primero estimar el θ_k. Debemos suponer que existe información auxiliar para que esto sea posible. Sea $\widehat{\theta_k}$ la estimación de θ_k, para $k \in r$. Un estimador ajustado a la falta de respuesta en dos fases se obtiene sustituyendo θ_k por $\widehat{\theta_k}$ en (9.2). De esta forma podemos calcular

$$\widehat{Y} = \sum_r \frac{d_k}{\widehat{\theta_k}} y_k. \tag{9.3}$$

Se han propuesto distintas formas de estimar $\widehat{\theta_k}$. Los estimadores del tipo (9.3) han sido estudiados y usados de forma extensiva en los últimos 30 años.

Una tendencia más reciente enfatiza el uso de información auxiliar para alcanzar dos objetivos: una reducción de la varianza del estimador y una reducción de su sesgo debido a la falta de respuesta. Cuando hablemos de información auxiliar distinguiremos dos tipos: vectores con información a nivel de la población, de forma que el total de un vector $\mathbf{x_k^*}$ es conocido; y vectores con información únicamente a nivel de muestra, de forma que se especifica un vector de valores $\mathbf{x}_k^o$ para cada $k \in s$.

Podemos extender el uso de la técnica de estimación GREG (Tema 4). Denotemos por $\sum_U \mathbf{x_k^*}$ el total conocido del vector auxiliar $\mathbf{x_k^*}$. Si el θ_k fuese conocido, podríamos construir el estimador GREG en dos fases, que es aproximadamente insesgado, y que viene dado por

$$\widehat{Y} = \sum_r \frac{d_k}{\theta_k} g_{k\theta} y_k. \tag{9.4}$$

donde $d_k = \frac{1}{\pi_k}$, y

$$g_{k\theta} = 1 + \Big(\sum_U \mathbf{x_k^*} - \sum_r \frac{d_k}{\theta_k} \mathbf{x_k^*} \Big)' \Big(\sum_r \frac{d_k}{\theta_k} c_k \mathbf{x_k^*} (\mathbf{x_k^*})' \Big)^{-1} (c_k \mathbf{x_k^*}). \tag{9.5}$$

donde las c_k son constantes especificadas. Sustituyendo los θ_k desconocidos por estimaciones adecuadas $\widehat{\theta_k}$, basadas en valores de variables auxiliares conocidos para $k \in s$ obtenemos, a partir de (9.4),

$$\widehat{Y} = \sum_r \frac{d_k}{\widehat{\theta_k}} g_{k\widehat{\theta}} y_k. \tag{9.6}$$

donde $g_{k\widehat{\theta}}$ viene dado por (9.5) con θ_k en lugar de $\widehat{\theta_k}$.

En general, es preferible trabajar con (9.6) a hacerlo con (9.3), y tanto más si $\mathbf{x}_\mathbf{k}^*$ es un vector auxiliar con información útil. Se han obtenido buenos resultados usando (9.3) y (9.6) en encuestas. Muchos de los estimadores obtenidos a partir de estas formulas son casos particulares de una familia más amplia que son los estimadores de calibrado.

En Kott 1994 se discuten varias cuestiones en relación a (9.2) y (9.4). En estas fórmulas aparece θ_k lo que hace que no sean estimadores prácticos, ya que los θ_k son desconocidos. Además, para ser insesgados o aproximadamente insesgados, ambos requieren que todos los θ_k sean estrictamente positivos. Como se señala en Kott 1994, este requisito es quizá poco realista, porque la mayoría de las encuestas tienen una fracción de núcleo duro de falta de respuesta, compuesta de individuos que no responden bajo ninguna circunstancia.

Un método alternativo ha consistido en usar las fórmulas (9.2) y (9.4) como punto de partida y llegar a los estimadores (9.3) y (9.6) mediante una estimación de los θ_k desconocidos. Como resultado, $\frac{d_k}{\widehat{\theta_k}}$ jugará el papel de peso para los k elementos que responden. Este enfoque se estructura en tres pasos:

i. se asume la existencia de un *mecanismo de respuesta*;

ii. se formula, para este mecanismo, un modelo realista en el cual los θ_k desconocidos aparezcan como parámetros desconocidos; y

iii. se estiman los θ_k, usando variables auxiliares cualesquiera que sean relevantes y la información del subconjunto de la muestra s que respondió.

Una crítica que se puede hacer a este método es que es difícil defender un modelo propuesto como más realista que cualquier otro competidor. Nuestro conocimiento sobre el comportamiento de la verdadera respuesta es limitado. El tiempo y los recursos necesario para realizar un estudio en profundidad del comportamiento de la respuesta son también limitados en los institutos de estadística.

Se han usado distintos modelos para el mecanismo de respuesta. Un modelo que se usa con mucha frecuencia es asumir que la población consiste en grupos que no se solapan de forma que todos los elementos dentro del mismo grupo responden con la misma probabilidad, y de forma independiente. Estos grupos se conocen como *grupos homogéneos de respuesta* (RHGs del inglés *response homogeneity groups*). Aquí, la información auxiliar necesaria es que podemos clasificar de forma única cada elemento muestral, si responde o no responde, en uno de los grupos. A menudo, en encuestas dirigidas a individuos, se usan grupos formados por un cruce de categorías de edad con sexo, de forma que, por ejemplo, cinco categorías de edad cruzadas con las dos de sexo nos da diez grupos. La experiencia ha demostrado que esta forma tan sencilla de agrupar raramente logrará la hipótesis inherente de probabilidad de respuesta constante dentro de cada grupo. Una agrupación por edad y sexo puede puede ser fácil de establecer pero a menudo es ineficiente para explicar un comportamiento a la hora de responder que, por lo general, es mucho más complejo. Se pueden obtener modelos RHG más eficientes, siempre que la información esté disponible, agrupando vía otros factores distintos a la edad y el

sexo.

Supongamos ahora que sustituimos los θ_k desconocidos en (9.4) por las estimaciones $\widehat{\theta}_k$ obtenidas a partir del modelo RHG. Este caso particular para (9.6) se discute en detalle en Särndal, Swensson y Wretman 1992, Capítulo 15, en el que también se proporciona un estimador de la varianza. Este estimador de la varianza tiene dos componentes, una que mide la varianza muestral y otra la varianza por la falta de respuesta. El estimador puntual es (casi) insesgado si el modelo RHG asumido es la verdadera representación del mecanismo de respuesta. Sin embargo, como ya se ha mencionado anteriormente, no importa qué grupos podemos establecer para los elementos muestrales, probablemente no se ajustarán de forma perfecta a las necesidades de igual probabilidad de respuesta para todos los elementos dentro de un grupo. Otras alternativas consisten en la modelización mediante la regresión logística y la modelización exponencial, como se explica en en Ekholm y Laaksonen 1991 y Folsom 1991.

Para resumir, el método de reponderación en dos fases para tratar la falta de respuesta necesita modelizar el mecanismo de respuesta como paso inicial para obtener los estimadores en (9.3) y (9.6). Para ello, uno debe (i) proporcionar la formulación matemática del modelo, y (ii) seleccionar las variables explicativas para este a partir de un conjunto más grande de variables auxiliares disponibles. Esto precisa análisis y toma de decisiones.

9.2 Vectores auxiliares e información auxiliar

El uso eficiente de información auxiliar permitirá realizar una estimación fiable en presencia de falta de respuesta. El de calibrado es más sencillo y directo que el en dos fases que se ha visto anteriormente.

La información auxiliar se transmite a través de un vector auxiliar. El *vector auxiliar* es un vector cuyos valores son conocidos para cada elemento informante, $k \in r$. Además, existe información en este vector para un *conjunto mayor* que *r*. Esta información ayudará tanto ante la falta de respuesta como en la reducción de la varianza. La notación genérica para un vector auxiliar será $\mathbf{x}$ y su valore para el elemento *k*-ésimo se denotará por x_k. La información para el conjunto mayor proporciona la *información input*, necesaria para el cálculo de los pesos de calibrado.

En el método de calibrado distinguiremos tres casos distintos, llamados InfoU, InfoS e InfoUS, dependiendo de la información de que dispongamos. Son los siguientes:

InfoU. Se dispone de información a nivel de la población *U*. Denotamos por $\mathbf{x}_k^*$ al vector de dimensión $J^* \geq 1$ tal que:

(i) el vector poblacional total $\sum_U \mathbf{x}_k^*$ es conocido;

(ii) para cada $k \in r$, el vector de valores $\mathbf{x}_k^*$ es conocido.

El vector auxiliar en este caso es $\mathbf{x}_k = \mathbf{x}_k^*$. A menudo nos referimos a $\mathbf{x}_k^*$ como el 'vector estrella'.

InfoS. Se dispone de información a nivel de la muestra s, pero no se dispone de ninguna a nivel de la población. Denotamos por $\mathbf{x}_k^o$ al vector de dimensión $J^o \geq 1$ tal que:

(i) para cada $k \in s$, el vector de valores $\mathbf{x}_k^o$ es conocido, mientras que $\sum_U \mathbf{x}_k^o$ es desconocido;

(ii) para cada $k \in r$, el vector de valores $\mathbf{x}_k^o$ es conocido.

El vector auxiliar para este caso es $\mathbf{x}_k = \mathbf{x}_k^o$. A menudo nos referimos a $\mathbf{x}_k^o$ como el 'vector luna'. Tiene información a un nivel inferior que el vector estrella.

InfoUS. Se dispone de los dos tipos de información, InfoU y InfoS, y se usan combinados para el cálculo de pesos. Una opción es formular el vector auxiliar como el vector $\mathbf{x}_k = \binom{\mathbf{x}_k^*}{\mathbf{x}_k^o}$ de dimensión $J^* + J^o$.

Las condiciones (i) y (ii) en InfoU y InfoS son las *mínimamente requeridas* para los procedimientos de estimación puntual. Veamos de dónde salen en la práctica InfoU y InfoS. Las posibles fuentes de información son la encuesta en sí misma, un censo de la población en cuestión, un registro administrativo o la unión de varios registros.

Empecemos considerando InfoU. Las condiciones (i) y (ii) prevalecen bajo dos escenarios importantes en la práctica: (a) El vector total auxiliar es 'importado' de una fuente fiable externa a la encuesta. (b) Existe un registro que enumera los elementos de la población de *1* a *N* de forma que el vector de valor $\mathbf{x}_k^*$ está especificado para cada elemento $K = 1, 2, ..., N$. La lista puede ser el resultado de unir varios registros administrativos usando una clave o un identificador único. Por ejemplo, personas o empresas. La lista también puede servir como un marco muestral (o para crear un marco) para la población objetivo en la encuesta.

En el caso de un *total importado*, la información $\sum_U \mathbf{x}_k^*$ proviene de una fuente externa, mientras que los valores individuales $\mathbf{x}_k^*$ son conocidos o medidos en la encuesta para $k \in s$. Es importante entonces que el vector tras el total importado $\sum_U \mathbf{x}_k^*$ en la condición (i) de InfoU mida el mismo concepto que $\mathbf{x}_k^*$ en (ii). En otras palabras, $\sum_U \mathbf{x}_k^*$ no puede estar desactualizado o tendremos medida erróneas del total poblacional real del vector $\mathbf{x}_k^*$ medido en la encuesta actual para cada $k \in r$.

El vector $\mathbf{x}_k^*$ también puede contener variables distintas de las categóricas, teniendo en cuenta que los totales poblacionales correspondientes se pueden importar de una fuente fiable.

Por otra parte nos encontramos con un problema cuando los valores de $\mathbf{x}_k^*$ propor-

cionados por los informantes contienen errores de medida. Esto es particularmente problemático si estos errores no son aleatorios sino sistemáticos, como cuando hay variables sensibles. Por ejemplo, supongamos que a los individuos encuestados se les pide cuantificar su consumo de alcohol durante un mes de referencia. Puede haber una tendencia a subestimar el consumo que hace uno mismo. En el caso del total poblacional, supongamos que podemos importar los datos mensuales de las ventas de una fuente considerada fiable, como por ejemplo de algún registro del estado sobre bebidas alcohólicas. Aunque esto nos pueda dar una imagen razonablemente precisa del consumo total, el problema es que las medidas obtenidas de los informantes están subestimados de forma sistemática. Esta disparidad puede perturbar de forma severa los pesos calculados para el calibrado.

Consideremos ahora el caso de un *listado poblacional* existente y que se usa para satisfacer las necesidades de InfoU. Esta situación es típica en las encuestas a individuos y a hogares en varios países europeos, principalmente en los países nórdicos, en los que un registro poblacional contiene un listado de todos los individuos k de la población (el marco) U. Los errores de cobertura de estos registros son escasos.

El registro puede ser el resultado de cruzar varias fuentes administrativas usando para ello un único elemento identificador, tal como el D.N.I. o N.I.E. en el caso de España para las encuestas sociales, y el N.I.F. en el caso de las encuestas económicas.

Un registro de la población típicamente contiene valores de una gran variedad de variables. Los valores de las variable están por tanto disponibles para todos los elementos de la población, no sólo para los elementos muestreados. Ejemplos de estas variables son la edad y el sexo, la situación laboral, la región en la que reside y el tipo de residencia, el nivel de estudios, y la nacionalidad. Como resultado, se puede especificar un vector estrella de valores $\mathbf{x}_k^*$ para cada individuo $k \in U$. Por tanto $\mathbf{x}_k^*$ es conocido para cada $k \in s$, de igual forma que lo es para cada $k \in r$. Sumando los valores $\mathbf{x}_k^*$ del marco podemos obtener el total deseado $\sum_U \mathbf{x}_k^*$, satisfaciendo por tanto la condición (i).

Consideremos ahora la InfoS, caracterizada por un vector luna de valores $\mathbf{x}_k^o$ medido o conocido de alguna otra manera para cada elemento muestral, $k \in s$. Es especialmente importante incluir en $\mathbf{x}_k^o$ las variables que explican el comportamiento de la falta de respuesta. El ingenio y el buen juicio ayudan en la identificación de estas variables. Un ejemplo es la identidad de los entrevistadores a los que se les asigna el elemento $k \in s$.

Se pueden obtener valores de la variable auxiliar para $k \in s$ a través del método de preguntas básicas propuesto por Bethlehem y Kersten 1985. Por experiencia, las personas que rehúsan responder a un cuestionario entero pueden, sin embargo, responder de buen grado a una o dos 'preguntas básicas'. Las variables detrás de las preguntas básicas deberían reflejar, o estar relacionadas con, temas que son fundamentales para la encuesta en cuestión, de forma que estén bien correladas con variables importantes de la encuesta y con la propensión a la respuesta.

Las respuestas a las preguntas básicas también se pueden obtener por mail o por encuesta telefónica. Las variables de las preguntas básicas se pueden incluir en el vector $\mathbf{x}_k^o$, asumiendo que las respuestas a esas variables se han obtenido para todas, o virtualmente todos, los elementos $k \in s$.

Case	Auxiliary vector, $\mathbf{x}_k$	Information input, $\mathbf{X}$
InfoU	$\mathbf{x}_k^*$	$\mathbf{X}^* = \sum_U \mathbf{x}_k^*$
InfoS	$\mathbf{x}_k^o$	$\widehat{\mathbf{X}}^o = \sum_s d_k \mathbf{x}_k^o$
InfoUS	$\binom{\mathbf{x}_k^*}{\mathbf{x}_k^o}$	$\binom{\mathbf{X}^*}{\mathbf{X}^o}$

Tabla 9.1: Formulación del vector auxiliar y del input de información correspondiente para los casos InfoU, InfoS e InfoUS.

La notación para InfoU, InfoS y InfoUS se resume en la Tabla (9.1), donde $\mathbf{X}$ es la notación genérica para el input de información que acompaña al vector auxiliar $\mathbf{x}_k$.

Podemos vee InfoU como un caso especial de InfoUS, que se obtiene cuando existe un vector estrella pero ningún vector luna; el vector auxiliar, por tanto, se reduce a $\mathbf{x}_k = \mathbf{x}_k^*$. De forma similar, InfoS es un caso especial de InfoUS obtenido cuando no hay un vector estrella, de forma que $\mathbf{x}_k = \mathbf{x}_k^o$.

El input de información $\widehat{\mathbf{X}}^o = \sum_s d_k \mathbf{x}_k^o$ es la estimación insesgada de Horvitz–Thompson del total desconocido $\mathbf{X}^o = \sum_U \mathbf{x}_k^o$. De ser posible, se puede reemplazar $\sum_s d_k \mathbf{x}_k^o$ por una mejor estimación (insesgada o casi). Una posible alternativa es usar una del tipo GREG (Tema 4), $\widehat{\mathbf{X}}^o = \sum_s d_k g_k \mathbf{x}_k^o$. Esto requiere que $\mathbf{x}_k^*$ sea conocido individualmente para cada $k \in s$.

9.3 El enfoque de calibrado

Como ya hemos comentado, abordar el problema de la falta de respuesta es importante. Nuestro objetivo es obtener un estimador de $Y = \sum_U y_k$ que cumpla las siguientes propiedades:

(i) que tenga un sesgo pequeño;

(ii) que tenga una varianza pequeña;

(iii) que tenga un sistema de pesos que reproduzca el input de información auxiliar cuando se aplique a las variables auxiliares;

(iv) que tenga un sistema de pesos que sea útil para estimar el total de cada variable y en una encuesta multi-propósito.

La propiedad (i) es particularmente importante. En relación con la varianza, consistirá en la suma de una componente de la varianza muestral y una componente de la varianza por falta de respuesta. Lo ideal sería que ambas componentes fuesen pequeñas. La propiedad expresada en (iii) es la pieza clave en el de calibrado. La propiedad (iv) expresa la conveniencia de un único sistema de pesos. Este aspecto es importante para la producción rutinaria y puntual de estadísticas, especialmente en encuestas grandes.

Veamos ahora el *de calibrado* para reponderar encuestas con falta de respuesta. El primer paso es fijar un vector auxiliar adecuado mediante la selección de variables de un conjunto grande de variables disponibles. A continuación veremos unos principios básicos sobre esta selección. El siguiente paso es el cálculo de los *pesos de calibrado*, usando software existente. Existen distintos programas que pueden ser usados para esta tarea, y algunos también estiman la varianza para diseños de muestreo particulares.

Los principios que se pueden usar para la selección de las variables auxiliares que nos permitan tener un vector auxiliar eficiente son:

Principio 1

El vector auxiliar (o el vector instrumento, si es distinto del vector auxiliar) debería explicar la probabilidad de respuesta inversa, llamada la influencia de respuesta.

Principio 2

El vector auxiliar debería explicar las principales variables de estudio.

Principio 3

El vector auxiliar debería identificar los dominios más importantes.

El Principio 1 se ve corroborado por los resultados de W. A. Fuller y Baker 1994, y el Principio 2 fundamentalmente confirma las conclusiones de Bethlehem 1985.

Cuando se cumple el Principio 1, se reduce el sesgo en las estimaciones calibradas para todas las variables de estudio. Esto es importante, ya que una encuesta grande tendrá normalmente muchas variables de estudio, y necesitamos asegurarnos de la eliminación efectiva del sesgo en todas las estimaciones. Por tanto, el Principio 1 es particularmente relevante.

Si se cumple el Principio 2, el sesgo se reduce en las estimaciones de las principales variables de estudio, pero quizá no en las estimaciones (obtenidas con los mismos pesos) para otras variables de estudio. Las principales variables de estudio en una encuesta se pueden identificar generalmente, pero basado en razones subjetivas. Cuando se

satisface el Principio 2, la varianza de las estimaciones también se reducirá.

Los residuos que determinan la varianza es probable que sean más pequeños si el vector auxiliar se puede formular de forma que identifique los principales dominios. También es deseable que los residuos sean pequeños en los dominios, desde la perspectiva de reducir el sesgo. Esta es la motivación detrás del Principio 3.

El de calibrado produce un *estimador calibrado* de Y , denotado por $\widehat{Y}_W$ y su correspondiente *estimador de varianza*, que denotamos $\widehat{\mathbb{V}}(\widehat{Y}_W)$. El subíndice W se usa para indicar ponderado (del inglés *weighting*). Este enfoque es muy general. En presencia de información auxiliar de peso, alcanza el doble objetivo de reducir el error de muestreo y el error de no respuesta. Se puede aplicar a cualquiera de los diseños de muestreo comunes y para cualquier vector auxiliar. El no recurre a ningún modelo(s) de forma explícita. Es muy adecuado para la producción rutinaria y atemporal de estimaciones, por ejemplo, en un instituto nacional de estadística. En la siguiente sección veremos los aspectos teóricos y prácticos de este enfoque. Los desarrollos teóricos están principalmente en Lundstrom y Särndal 1999 y Deville 2000.

9.4 Estimación puntual bajo calibrado

Presentamos ahora el *estimador calibrado* de $Y = \sum_U y_k$, denotado por $\widehat{Y}_W$. Los pesos calibrados se denotan por w_k para $k \in r$. El estimador calibrado viene dado por

$$\widehat{Y}_W = \sum_r w_k y_k. \tag{9.7}$$

Dependiendo de lo buena que sea la información auxiliar podría ocurrir que los w_k protejan satisfactoriamente al estimador frente al sesgo de no respuesta. Nos centraremos en el caso general, InfoUS, en el que vector auxiliar es $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^o \end{pmatrix}$ de dimensiones $J^* + J^o$ y el input de información es $\mathbf{X} = \begin{pmatrix} \mathbf{X}^* \\ \widehat{\mathbf{X}}^o \end{pmatrix}$. InfoU sería un caso especial en el que $\mathbf{x}_k = \mathbf{x}_k^*$, e InfoS cuando $\mathbf{x}_k = \mathbf{x}_k^o$. Los tres casos dan lugar a sistemas de peso calibrado diferentes.

Buscamos un sistema de pesos w_k para $k \in r$ que satisfaga la *ecuación de calibración*

$$\sum_r w_k \mathbf{x}_k = \mathbf{X}. \tag{9.8}$$

La ecuación (9.8) implica para $\mathbf{x}_k^*$ que $\sum_r w_k \mathbf{x}_k^* = \sum_U \mathbf{x}_k^*$, y para $\mathbf{x}_k^o$ que $\sum_r w_k \mathbf{x}_k^o* = \sum_s d_k \mathbf{x}_k^o$. Los pesos que satisfacen (9.8) se dice que están *calibrados* al input de información $\mathbf{X}$. Reproducirán exactamente la información dada $\mathbf{X}$ cuando se apliquen a los

valores del vector auxiliar $\mathbf{x}_k$ y se sumen en el conjunto de respuesta *r*. Con otro término a menudo utilizado, el sistema de pesos w_k para $k \in r$ se dice que es *consistente* con la información $\mathbf{X}$.

Como consecuencia de la selección de muestreo, el elemento *k* lleva asociado el peso $d_k = \frac{1}{\pi_k}$. Cuando hay falta de respuesta, los pesos d_k para $k \in r$ son por ellos mismos demasiado pequeños en media para producir estimaciones aceptables. La suma ponderada $\sum_r d_k y_k$ es una subestimación de $\sum_U y_k$. El d_k debe aumentar. Se deben buscar nuevos pesos que sean mayores que los d_k para todos, o al menos una mayoría, de los elementos que han respondido. Denotamos por ν_k el factor, o peso, por el cual multiplicamos d_k para obtener el nuevo peso para el elemento *k*. Es decir, $w_k = d_k \nu_k$.

¿Qué expresión debería de tener el peso ν_k? Debería reflejar las características individuales conocidas del elemento $k \in r$, resumida en el vector de valores $\mathbf{x}_k$. Una forma sencilla es por tanto $\nu_k = 1 + \lambda' \mathbf{x}_k$, donde λ es un vector a determinar. En esta fórmula, dejamos que ν_k dependa linealmente del valor conocido $\mathbf{x}_k$ para el elemento *k*.

Es un ejercicio sencillo determinar el valor de λ que satisface el requisito de calibrado. Si introducimos $\nu_k = 1 + \lambda' \mathbf{x}_k$ en (9.8) y resolvemos para λ', obtenemos $\lambda' = \lambda_k^{'}$, donde

$$\lambda_k^{'} = \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k \right)^{'} \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k^{'} \right)^{-1}, \tag{9.9}$$

asumiendo que existe la inversa de la matriz $\sum_r d_k \mathbf{x}_k \mathbf{x}_k^{'}$. El nuevo peso es $w_k = d_k + d_k \lambda_k^{'} \mathbf{x}_k$, donde el término añadido $d_k \lambda_k^{'} \mathbf{x}_k$ será positivo para la mayoría de (pero no necesariamente para todos) los elementos. Este término modifica el peso de muestreo d_k, que no es suficientemente grande, por un valor más razonable, habitualmente mayor.

Hemos determinado los pesos que se tienen en cuenta para la falta de respuesta y están calibrados a la información dada. Estos pesos se denominan *pesos estándares* y vienen dados por

$$w_k = d_k \nu_k, \nu_k = 1 + \lambda_k^{'} \mathbf{x}_k, \tag{9.10}$$

donde $\lambda_k^{'} = (\mathbf{X} - \sum_r d_k \mathbf{x}_k)^{'} (\sum_r d_k \mathbf{x}_k \mathbf{x}_k^{'})^{-1}$. El estimador calibrado resultante es

$$\widehat{Y}_W = \sum_r w_k y_k. \tag{9.11}$$

Además del input de información $\mathbf{X}$, el cálculo de w_k necesita dos sumas sobre el conjunto de respuesta *r*. Ambas sumas se pueden calcular porque $\mathbf{x}_k$ es conocido para cada $k \in r$. La matriz $(\sum_r d_k \mathbf{x}_k \mathbf{x}_k^{'})$, que necesita ser invertible, es de dimensión $(J^* + J^o) \times (J^* + J^o)$, donde J^* y J^o son las dimensiones de $\mathbf{x}_k^*$ y $\mathbf{x}_k^o$, respectivamente.

Al sistema de pesos definido por (9.10) lo llamaremos *pesos estándar*. Veremos algunos pesos alternativos en la Sección (9.6).

Ejemplo 40. **El vector auxiliar más sencillo**
El vector auxiliar más sencillo es uno que es constante para todo *k*. Como no reconoce diferencias entre individuos, es ineficiente para el tratamiento de la falta de respuesta. Pero nos da el ejemplo más básico de estimación por calibrado. Con $\mathbf{x}_k = \mathbf{x}_k^* = 1$ para todo *k*, obtenemos a partir de (9.10) $w_k = d_k(\frac{N}{\sum_r d_k})$ y $\widehat{Y}_W = N\bar{y}_{r;d}$, donde $\bar{y}_{r;d} = (\frac{\sum_r d_k y_k}{\sum_r d_k})$ es la media de *y* elevada con los pesos de diseño de los informantes. En particular, para muestreo aleatorio simple sin reemplazamiento, $w_k = \frac{N}{m}$ para todo *k* y $\widehat{Y}_W = N\bar{y}_r$, donde $\bar{y}_r = (\frac{\sum_r y_k}{m})$ y *m* es el tamaño del conjunto de respuestas *r*. Veremos más ejemplos en la Sección 9.7

Comentario 49. Para un vector dado $\mathbf{x}_k$ y el correspondiente input de información $\mathbf{X}$, el cálculo de w_k definido por (9.10) puede dar lugar a unos pocos valores negativos o excesivamente grandes ν_k. Los pesos negativos son considerados por muchos como no deseables, lo mismo que pesos positivos excesivamente grandes. Hay distintas formas de abordar esta cuestión. Alguno de los software existente tiene incorporados procedimientos que permiten que los pesos se restrinjan a unos intervalos previamente especificados. Se debería de tener en mente esta cuestión cuando construya el vector auxiliar. Antes de tomar una decisión final sobre x_k, se debería de estar seguro de que el vector no contiene variables *x* que puedan causar problemas con pesos no deseables.

Comentario 50. Hemos argumentado que la forma lineal, $\nu_k = 1 + \lambda'\mathbf{x}_k$, es razonable para el ajuste. Las formas no lineales también se pueden considerar. Podríamos suponer $\nu_k = F(\lambda'\mathbf{x}_k)$ para alguna función $F(\cdot)$ con propiedades adecuadas, por ejemplo, $F(\lambda'\mathbf{x}_k) = exp(\lambda'\mathbf{x}_k)$. Buscaríamos entonces un vector λ que satisfaga la ecuación de calibrado $\sum_r d_k F(\lambda'\mathbf{x}_k) = \mathbf{X}$. Si ese vector es λ_k', los pesos calibrados son $w_k = d_k F(\lambda_k'\mathbf{x}_k)$. Para nuestros propósitos, la forma lineal será suficiente; tiene considerables ventajas computacionales.

9.5 Comentarios sobre el calibrado

En la Sección (9.3), se formularon cuatro propiedades deseables para el estimador. ¿Se verifican estas propiedades en el estimador calibrado $\widehat{Y}_W = \sum_r w_k y_k$ con los pesos datos por (9.10)? Las propiedades (i) y (ii) relativas al sesgo y varianza se discuten a fondo en el tema siguiente. La propiedad (iii) se satisface por construcción. La propiedad (iv) expresa la conveniencia de tener un conjunto de pesos que son aplicables, con buenos resultados, a *todas* las variables de interés en una encuesta. Los pesos calibrados w_k cumplen este objetivo en el sentido de que ya incorporan la información considerada: la mejor posible bajo las condiciones de la encuesta. El siguiente comentario señala una propiedad interesante del estimador calibrado.

Comentario 51. Consideremos el caso InfoU, donde $\mathbf{x}_k = \mathbf{x}_k^*$ y $\mathbf{X} = \mathbf{X}^* = \sum_U \mathbf{x}_k^*$. Asumamos que existe una relación lineal perfecta en la población entre la variable de estudio y_k y el vector auxiliar $\mathbf{x}_k^*$, de forma que, para cada $k \in U$,

$$y_k = (\mathbf{x}_k^*)' \beta^* \tag{9.12}$$

donde β es un vector columna de constantes (desconocidas). Entonces el estimador calibrado $\widehat{Y}_W$ nos da una estimación exacta del total objetivo Y. Esto se sigue a partir de

$$\widehat{Y}_W = \sum_r w_k y_k = \left(\sum_r w_k \mathbf{x}_k^* \right)' \beta^* = \left(\sum_U \mathbf{x}_k^* \right)' \beta^* = \sum_U y_k = Y.$$

Todo lo necesario para establecer la igualdad $\widehat{Y}_W = Y$ es la propiedad de calibración de los pesos, dada por (9.8).

En la práctica, la relación lineal perfecta expresada por (9.12) no se verifica. Si lo hiciese, difícilmente sería necesario que la variable y fuese parte de la encuesta. Pero el resultado sugiere que cuando existe una relación lineal fuerte entre y_k y $\mathbf{x}_k^*$, el estimador calibrado $\widehat{Y}_W$ debería de acercarse en valor al objetivo Y. Tanto el error de muestreo como el error por falta de respuesta deberían esencialmente de desaparecer.

Algunas características adicionales del método de calibrado son:

- *Generalidad*. Aunque $\widehat{Y}_W$ se llama 'el estimador calibrado', en realidad es la expresión para una familia más amplia de estimadores, que se corresponden con las distintas formulaciones de $\mathbf{x}_k$. Se obtienen una gran generalidad y flexibilidad a partir del hecho de que el vector $\mathbf{x}_k$ en (9.11) puede tener virtualmente cualquier forma, mientras está disponible el correspondiente input de información $\mathbf{X}$, y en la medida en que la matriz sea invertible.
- *Aspectos computacionales*. Cuando se usa en la práctica el método de calibrado, no es necesario una desarrollar una fórmula o expresión algebraica para un vector $\mathbf{x}_k$ elegido. Se lleva a cabo a través de software existente. Podemos contar con un número de programas de software disponibles. El usuario necesita especificar el vector $\mathbf{x}_k$ y el correspondiente input de información. El software producirá los pesos calibrados w_k y las correspondientes estimaciones calibradas $\widehat{Y}_W = \sum_r w_k y_k$. Algún software también calcula la varianza asociada.
- *Técnicas convencionales*. Hay técnicas para la reponderación de la falta de respuesta que se puede denominar 'convencional', considerando la atención que han tenido en los últimos años en la literatura. Muchas de éstas se pueden obtener como casos especiales del estimador calibrado $\widehat{Y}_W$.

9.6 Pesos de calibrado alternativos

Los pesos de calibrado no son únicos. La ecuación de calibrado (9.8) impone sólo una pequeña restricción sobre los pesos. Como aclaramos ahora, existen muchos conjuntos de pesos calibrados, para un vector $\mathbf{x}_k$ dado con input de información $\mathbf{X}$.

Necesitamos algunos conceptos adicional con el fin de ver la imagen completa de la técnica de calibrado: los *pesos iniciales*, los *pesos finales* y el *vector instrumento*. Los d_k son los pesos iniciales en (9.11), en el sentido de que los cálculos definidos por (9.11) transformarán los d_k en los nuevos pesos w_k, que son los pesos finales.

Primero aclaremos que podemos trabajar con otros pesos iniciales distintos de los d_k y aún así conseguir que se satisfagan las condiciones de los pesos finales de calibrado a la información $\mathbf{X}$ dada. Sea $d_{\alpha k}$, para $k \in r$, cualquier conjunto de pesos positivos. En (9.11) y en el correspondiente vector $\lambda_r^{'}$, sustituimos d_k por $d_{\alpha k}$. Los pesos finales resultantes w_k aún satisfarán la ecuación de calibrado (9.8). Por ejemplo, podríamos considerar $d_{\alpha k} = Cd_k$, donde C es un valor positivo que no depende de k, como $\frac{n}{m}$, el inverso de la tasa de respuesta en la encuesta.

Ahora señalamos un segundo cambio en (9.11) que también dejará la propiedad de calibrado (9.8) intacta. Sea $\mathbf{z}_k$ cualquier vector de valores especificado para $k \in r$ y con la misma dimensión que $\mathbf{x}_k$. El vector $\mathbf{z}_k$ puede ser una función concreta de $\mathbf{x}_k$ o de otros datos observados con anterioridad sobre k. En (9.11), sustituye $\nu_k = 1 + \lambda_r^{'}\mathbf{x}_k$ y $(\sum_r d_k \mathbf{x}_k \mathbf{x}_k^{'})^{-1}$ por $\nu_k = 1 + \lambda_r^{'}\mathbf{z}_k$ y $(\sum_r d_k \mathbf{z}_k \mathbf{x}_k^{'})^{-1}$, respectivamente. Los nuevos pesos resultantes todavía están calibrados al input de información $\mathbf{X}$. Llamamos a z_k un *vector instrumento* para el calibrado.

Usando los conceptos de pesos iniciales y vector instrumento, obtenemos el sistema de pesos calibrados

$$w_k = d_{\alpha k}\nu_k, \quad \nu_k = 1 + \lambda_r^{'}\mathbf{z}_k \tag{9.13}$$

donde $\lambda_r^{'} = (\mathbf{X} - \sum_r d_{\alpha k}\mathbf{x}_k)^{'}(\sum_r d_{\alpha k}\mathbf{z}_k\mathbf{x}_k^{'})^{-1}$. Estos w_k satisfacen la ecuación de calibrado (9.8) para *cualesquiera* pesos iniciales positivos $d_{\alpha k}$ y *cualquier* instrumento $\mathbf{z}_k$, mientras la matriz $\sum_r d_{\alpha k}\mathbf{z}_k\mathbf{x}_k^{'}$ pueda ser invertible. Usando estos pesos, el estimador calibrado es $\widehat{Y}_W = \sum_r w_k y_k$.

El *peso estándar* definido por (9.11) es el caso especial de (9.13) obtenido para $d_{\alpha k} = d_k$ y $\mathbf{z}_k = \mathbf{x}_k$. A no ser que se especifique otra cosa, asumimos a partir de ahora que se utilizarán las *especificaciones estándar* $d_{\alpha k} = d_k$ y $\mathbf{z}_k = \mathbf{x}_k$.

En resumen, para calcular los pesos w_k dados por (9.13) necesitamos especificar:

- los pesos iniciales $d_{\alpha k}$;
- el vector auxiliar de valores $\mathbf{x}_k$ y el correspondiente input de información $\mathbf{X}$;

- el vector instrumento de valores $\mathbf{x}_k$, si es diferente de $\mathbf{x}_k$.

El procedimiento de calibrado transforma los pesos iniciales a través de un input de información en los pesos finales dados por (9.11) o más general por (9.13). Los pesos finales también se conocen por *pesos calibrados*, y tienen en cuenta la falta de respuesta de una forma apropiada.

Comentario 52. Una propiedad deseable (pero no obligatoria) de los pesos calibrados es que su suma sea el tamaño poblacional N. Cuando esto ocurre, el sistema de pesos estima correctamente el tamaño poblacional. Es una condición justificada, pero por sí misma da poca información para la elección de $d_{\alpha k}$ y $\mathbf{z}_k$. Un gran número de sistemas de pesos calibrados verifican esta propiedad. Independientemente de los pesos iniciales $d_{\alpha k}$ y del vector instrumento $\mathbf{z}_k$ en (9.13), la ecuación $\sum_r w_k = N$ se verifica para InfoU e InfoUS cuando el vector estrella $\mathbf{x}_k^*$ contenga la constante 1 para todo k. No es un inconveniente serio si $\sum_U w_k = N$ no se verifica, considerando, por ejemplo, que en algunas encuestas, el tamaño poblacional N es desconocido.

Debido a la libertad para elegir $d_{\alpha k}$ y $\mathbf{z}_k$, existen muchos sistemas de pesos calibrados para dar input de información, $\mathbf{X}$. A cada sistema de pesos le corresponde un estimador calibrado $\widehat{Y}_W = \sum_r w_k y_k$. Esto plantea la cuestión de cómo elegir $d_{\alpha k}$ y $\mathbf{z}_k$. La mayoría de las veces usaremos $d_{\alpha k} = d_k$ como pesos iniciales. Una razón es que los pesos finales w_k son en muchos casos invariantes a un ajuste preliminar de los d_k, como $d_{\alpha k} = Cd_k$, donde C es un valor que no depende de k, por ejemplo, $C = \frac{n}{m}$, el inverso de la tasa de respuesta.

Consideremos ahora la elección de $\mathbf{z}_k$. En la mayoría de los casos tomamos $\mathbf{z}_k$ idéntico a $\mathbf{x}_k$, la *elección estándar*. Veremos más ejemplos en la Sección 9.7. Pero en la práctica, podemos especificar las componentes de $\mathbf{z}_k$ como otras funciones de las componentes de $\mathbf{x}_k$. Si $\mathbf{x}_k = (x_{1k}, x_{2k})'$, donde x_{1k} y x_{2k} son positivos, podemos obtener pesos calibrados a partir de (9.13) tomando, por ejemplo,$\mathbf{x}_k = (\sqrt{x_{1k}}, \sqrt{x_{2k}})'$. No queremos decir que un $\mathbf{z}_k$ diferente de $\mathbf{x}_k$ dé pesos, en algún sentido mejores que los que se obtienen con $\mathbf{z}_k = \mathbf{x}_k$, ni tampoco recomendamos ninguna forma específica de $\mathbf{z}_k$. Simplemente queremos alertar sobre una generalidad del de calibrado que quizá puede ser explorado y usado como ayuda en algunos escenarios.

Comentario 53. Para el caso de respuesta completa, existe una relación cercana entre el estimador calibrado $\widehat{Y}_W = \sum_r w_k y_k$ con pesos definidos por (9.13) y el estimador GREG (Tema 4). Si $r = s$ y si $\mathbf{x}_k = \mathbf{x}_k^*$, $\mathbf{z}_k = c_k \mathbf{x}_k^*$, los dos estimadores son idénticos. Ésta es una propiedad atractiva, ya que el estimador GREG es conocido por tener propiedades favorables: un sesgo muy cercano a cero y una varianza pequeña cuando y_k está bien explicada por $\mathbf{x}_k^*$. Cuando la falta de respuesta es muy limitada, el estimador calibrado es cercano al estimador GREG.

Comentario 54. El estimador en dos fases $\widehat{Y} = \sum_r \frac{d_k}{\widehat{\theta}_k} g_{k\widehat{\theta}} y_k$ dado por (9.6) es un estima-

dor calibrado con pesos de la forma (9.13). Los pesos iniciales son $d_{\alpha k} = \frac{d_k}{\widehat{\theta}_k}$, el vector auxiliar es $\mathbf{x}_k = \mathbf{x}_k^*$, el instrumento es $\mathbf{z}_k = \mathbf{x}_k^*$, el input de información es $\sum_U \mathbf{x}_k^*$, y los pesos finales son

$$w_k = \frac{d_k}{\widehat{\theta}_k}\left\{1 + \left(\sum_U \mathbf{x}_k^* - \sum_r \frac{d_k}{\widehat{\theta}_k}\mathbf{x}_k^*\right)' \left(\sum_r \frac{d_k}{\widehat{\theta}_k}\mathbf{x}_k^*(\mathbf{x}_k^*)'\right)^{-1} \mathbf{x}_k^*\right\}.$$

Es fácil ver que satisfacen la ecuación de calibrado (9.8) con $\mathbf{X} = \sum_U \mathbf{x}_k^*$.

Comentario 55. Incluso si $\sum_r d_{\alpha k}\mathbf{z}_k\mathbf{x}_k'$ es singular, los pesos calibrados podrían todavía venir dados por (9.13) si sustituimos una *inversa generalizada*.

9.7 Ejemplos de estimadores calibrados

Veamos algún ejemplo más de estimadores calibrados.

9.7.1 Clasificación unidireccional

La información sobre una clasificación de elementos es muy útil para la estimación con falta de respuesta. Consideremos un modelo de clasificación que asigna un elemento k a uno de los grupos o categorías que forman el conjunto P, donde cada grupo o categoría es mutuamente exclusivo y exhaustivo – por ejemplo, grupos de edades, o los grupos definidos por la clasificación cruzada de grupo de edad por sexo por situación laboral. El identificador del grupo para el elemento k se define como

$$\gamma_k = (\gamma_{1k}, ..., \gamma_{pk}, ..., \gamma_{Pk})' \tag{9.14}$$

donde, para $p = 1, ..., P$, $\gamma_{pk} = 1$ si k pertenece al grupo p y $\gamma_{pk} = o$ en caso contrario. Es decir, el vector γ_k tiene $P - 1$ componentes cero y una única componente que vale la unidad. Este último identifica la pertenencia al grupo de k. La información sobre γ_k para InfoU y la información sobre γ_k para InfoS representan dos casos diferentes; consideremos ambos. La condición $\mu'\mathbf{x}_k = 1$ para todo k se verifica tomando $\mu = (1, 1, ..., 1)$.

En el caso de InfoU, el vector auxiliar es $\mathbf{x}_k = \mathbf{x}_k^* = \gamma_k$ con la información asociada $\sum_U \mathbf{x}_k^* = (N_1, .., N_p, ..., N_P)'$, donde N_p es el tamaño conocido del grupo U_p. Esto se utiliza en muchas encuestas, en particular en los países nórdicos, en los que se dispone de registros de la población con sus características demográficas, educativas y económicas. En otros países, N_p se puede obtener de los totales censales (actualizados). Sea r_p el conjunto de respuestas del grupo p; el conjunto total de respuestas es $r = \bigcup_{p=1}^P r_p$. A partir de la fórmula de pesos (9.10), con $d_{\alpha k} = d_k$ y $\mathbf{z}_k = \mathbf{x}_k$, obtenemos $\nu_k = \mathbf{F}_p^*$ para todo $k \in r_p$, donde $\mathbf{F}_p^* = \frac{N_p}{\sum_{r_p} d_k}$. El estimador calibrado (9.11) se convierte en

$$\widehat{Y}_W = \sum_{p=1}^P N_p \bar{y}_{r_p;d} = \widehat{Y}_{PWA}. \tag{9.15}$$

donde $\bar{y}_{r_p;d} = \frac{\sum_{r_p} d_k y_k}{\sum_{r_p} d_k}$ es la media ponderada de las respuestas dentro del grupo. Si la falta de respuesta ocurre de forma aleatoria en cada grupo, entonces $\bar{y}_{r_p;d}$ estima la media del grupo $\bar{y}_{U_p} = \sum_{U_p} \frac{y_k}{N_p}$ prácticamente sin sesgo, y entonces (9.15) casi no tiene sesgo para $\widehat{Y}_{PWA} = \sum_U y_k$.

En particular, para muestreo aleatorio simple sin reemplazamiento e InfoU, (9.15) se convierte en

$$\widehat{Y}_{PWA} = \sum_{p=1}^{P} N_p \bar{y}_{r_p}. \tag{9.16}$$

donde $\bar{y}_{r_p} = \sum_{r_p} \frac{y_k}{m_p}$ es la media de y para los m_p informantes que responden en el grupo p. La fórmula (9.16) a veces se denomina *estimador postestratificado*. El término es un poco confuso ya que el estimador postestratificado tradicional se refiere a una única fase de muestreo. En consecuencia, algunos autores como Kalton y Kasprzyk 1986 hacen una distinción entre el estimador postestratificado, tal y como se usa para la respuesta completa en el muestreo en una sola fase, y (9.16), al que ellos llaman el *estimador ajustado con el peso poblacional*, como se refleja en la notación $\widehat{Y}_{PWA}$. En el último, el término más apropiado reconoce una fase de muestreo seguida por una fase de falta de respuesta.

9.7.2 Una única variable auxiliar cuantitativa

Este ejemplo no es muy común ya que el objetivo de llegar a una fórmula bien conocida requiere un vector instrumento $\mathbf{z}_k$ distinto de la elección estándar $\mathbf{z}_k = \mathbf{x}_k$. Consideremos una variable auxiliar cuantitativa, x, cuyo valor para el elemento k es x_k. Puede ser una medida del tamaño de k, como el número de empleados de la empresa k en una encuesta a empresas. Para cada $k \in r$, se observa el valor x_k o es conocido. Podemos encontrarnos con dos casos: podemos trabajar únicamente con x_k, o, suponiendo que N es conocido, con el vector $(1, x_k)'$. Consideraremos ambas posibilidades.

Consideremos InfoU con $\mathbf{x}_k = \mathbf{x}_k^* = x_k$ y el correspondiente total poblacional conocido $\sum_U x_k$. De (9.13) con $d_{\alpha k} = d_k$ y $\mathbf{z}_k = 1$ para todo k, obtenemos

$$\widehat{Y}_W = \left(\sum_U x_k\right) \frac{\sum_r d_k y_k}{\sum_r d_k x_k} = \widehat{Y}_{RA}. \tag{9.17}$$

Esto tiene la forma del *estimador de razón*, de ahí el índice RA (del inglés *ratio estimator*), Tema **??**. Si ocurre la falta de respuesta con igual probabilidad en toda la población, entonces (9.17) casi no tiene sesgo para $Y = \sum_U y_k$.

Los rasgos del estimador de razón se hacen incluso más explícitos bajo muestreo aleatorio simple sin reemplazamiento, en cuyo caso

$$\widehat{Y}_{RA} = N \bar{x}_U \frac{\bar{y}_r}{\bar{x}_r}, \tag{9.18}$$

donde $\bar{x}_U = \sum_U \frac{x_k}{N}$, $\bar{y}_r = \sum_r \frac{y_k}{m}$, y $\bar{x}_r$ se define de forma análoga. Nótese que el estimador de razón definido en el Tema **??** coincide con el (9.17) cuando $r = s$, es decir, cuando no hay falta de respuesta. Bajo muestreo aleatorio simple sin reemplazamiento, el estimador de razón con respuesta total es casi insesgado, pero esta propiedad no se puede be aplicar a (9.18) a no ser que la falta de respuesta ocurra de forma completamente aleatoria.

En el caso InfoS, conocemos el valor de x_k para cada $k \in s$. Obtenemos los pesos de (9.13) con $\mathbf{x}_k = \mathbf{x}_k^o = x_k$, $d_{\alpha k} = d_k$ y $\mathbf{z}_k = 1$ para todo k. En particular, para muestreo aleatorio simple sin reemplazamiento e InfoS, la media muestral de x, $\bar{x}_s = \sum_s \frac{x_k}{n}$, se puede calcular, y el estimador se convierte en

$$\widehat{Y}_W = N\bar{x}_s \frac{\bar{y}_r}{\bar{x}_r}. \tag{9.19}$$

En presencia de una variable x continua, podemos formular el vector auxiliar alternativamente como $\mathbf{x}_k = (1, x_k)'$. Para InfoU, tenemos $\mathbf{x}_k = \mathbf{x}_k^* = (1, x_k)'$ con el input de información $\mathbf{X} = \sum_U \mathbf{x}_k = (N, \sum_U xk)'$, asumiendo que el tamaño poblacional N es conocido. Con los pesos estándar (9.10) obtenemos

$$\widehat{Y}_W = N\{\bar{y}_{r;d} + (\bar{x}_U - \bar{x}_{r;d})B_{r;d}\} = \widehat{Y}_{REG}. \tag{9.20}$$

la forma del *estimador de regresión*, donde $\bar{x}_{r;d} = \frac{\sum_r d_k x_k}{\sum_r d_k}$ y

$$B_{r;d} = \frac{\sum_r d_k (x_k - \bar{x}_{r;d})(y_k - \bar{y}_{r;d})}{\sum_r d_k (x_k - \bar{x}_{r;d})^2}. \tag{9.21}$$

Los beneficios de $\widehat{Y}_{REG}$ como un estimador ajustado para la falta de respuesta se discuten, por ejemplo, en Bethlehem 1985. El clásico estimador de regresión se obtiene a partir de (9.20) cuando $r = s$. El estimador (9.20) normalmente da mejor protección contra el sesgo por falta de respuesta que el estimador de razón (9.17).

Bibliografía

Bethlehem, J. G. (1985). "Reduction of nonresponse bias through regression estimation". En: *Journal of Official Statistics* 4, págs. 251-260.

Bethlehem, J. G. y H. M. P. Kersten (1985). "On the treatment of nonresponse in sample surveys". En: *Journal of Official Statistics* 1, págs. 287-300.

Deville, J. C. (2000). "Generalized calibration and application to weighting for nonresponse." En: *In J. G. Bethlehem and P. G. M. van der Heijden (eds), COMPSTAT – Proceedings in Computational Statistics*, págs. 65-76.

Ekholm, A. y S. Laaksonen (1991). "Weighting via response modeling in the Finnish Household Budget Survey". En: *Journal of Official Statistics* 3, págs. 325-337.

Folsom, R.E. (1991). "Exponential and logistic weight adjustments for sampling and nonresponse error reduction". En: *Proceedings of the Social Statistics Section, American Statistical Association*, págs. 197-202.

Kalton, G. y D. Kasprzyk (1986). "The treatment of missing data". En: *Survey Methodology* 12, págs. 1-16.
Kott, P. S. (1994). "A note on handling nonresponse in sample surveys". En: *Journal of the American Statistical Association* 89, págs. 693-696.
Lindström, H. y S. Lundström (1974). "A method to discuss the magnitude of the non-response error". En: *Statistisk Tidskrift* 12, págs. 505-520.
Little, R. J. A. y D. B. Rubin (1987). *Statistical Analysis with Missing Data*. New York: Wiley.
Lundstrom, S. y C.-E. Särndal (1999). "Calibration as a standard method for treatment of nonresponse". En: *J. Official Stat.* 15, págs. 305-327.
Lundström, S. y C.-E. Särndal (2001). *Estimation in the Presence of Nonresponse and Frame Imperfections*. Örebro: Statistics Sweden.
Oh, H. L. y F. J. Scheuren (1978). *Weighting adjustment for unit nonresponse. W. G. Madow, I. Olkin and D. B. Rubin (eds), Incomplete Data in Sample Surveys, Vol. 2*. New York: Academic Press.
Platek, R. y G.B. Gray (1983). *Imputation methodology. In W. G. Madow, I. Olkin and D. B. Rubin (eds), Incomplete Data in Sample Surveys, Vol. 2.* New York: Academic Press.
R. Platek, M. P. Singh y V. Tremblay (1978). *Adjustments for nonresponse in surveys. In N. K. Namboodiri (ed.), Survey Sampling and Measurement.* New York: Academic Press.
Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
Särndal, C.-E. y S. Lundström (2005). *Estimation in Surveys with Nonresponse*. Chichester: Wiley.
Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.
W. A. Fuller, M. M. Loughin y H. D. Baker (1994). "Regression weighting in the presence of nonresponse with application to the 1987–1988 nationwide Food Consumption Survey". En: *Survey Methodology* 20, págs. 75-85.

Tema 10

Estimación basada en modelos estadísticos. Aspectos generales de la estimación basados en modelos. Teoría de la predicción. Comparación con la teoría del muestreo probabilístico en poblaciones finitas.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía.

R. Valliant, A.H. Dorfmann y R.M. Royall (2000). *Finite population sampling and inference. A prediction approach*. New York: Wiley

R.G. Clark R.R. Chambers (2012). *An introduction to model-based survey sampling with applications*. New York: Oxford University Press

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

10.1 Aspectos generales de la estimación basados en modelos

El conjunto de temas expuestos con anterioridad contiene las ideas centrales de la estimación basada en diseños muestrales o muestreo probabilístico, que es el enfoque seguido en la producción estadística oficial. No obstante, esta no es la única opción para calcular estimaciones y realizar inferencia sobre una población finita. Para entender adecuadamente las sutilezas entre estas variantes comentaremos, en primer lugar, un ejercicio propuesto por Deming 1950 y, en segundo lugar, las definiciones de muestra empleadas en muestreo y en inferencia estadística.

Centrémonos en la siguiente versión adaptada del ejercicio propuesta por Deming 1950, pág. 254. Considérese una máquina industrial que fabrica tuercas según un conjunto dado de especificaciones técnicas (forma geométrica, resistencia térmica, peso, dureza, etc.). Estas tuercas se empaquetan en cajas de capacidad fija (digamos, N tuercas), que se distribuyen posteriormente para su venta al por menor. Distinguimos dos cuestiones estadísticamente diferentes (aunque relacionadas). Por un lado, podemos interesarnos

por conocer el número de tuercas defectuosas en una caja dada. Por otro lado, podemos también interesarnos por conocer el ratio de producción de tuercas defectuosas por la máquina. Ambas cuestiones tienen interés por sí mismas. El vendedor al por menor tendrá interés en la primera cuestión, que se formula estadísticamente como un problema de estimación en poblaciones finitas (Cassel, Särndal y Wretman 1977). Este es el problema que se resuelve con el muestreo probabilístico contenido en el resto de temas sobre producción estadística oficial. La segunda cuestión se trata en realidad de un problema clásico de inferencia (véase, p.ej. Casella y Berger 2002). El funcionamiento de la máquina puede modelizarse como un fenómeno aleatorio, lo que justifica la definición de una variable aleatoria de Bernoulli $P \simeq \mathrm{Ber}(p)$, cuyo parámetro p expresa la ratio de tuercas defectuosas y que se estima, bien puntualmente bien por intervalos, con cualquiera de las técnicas contenidas en los temas de inferencia estadística.

Adviértase que en el primer problema no existen elementos estocásticos en su definición mientras que en el segundo sí. Es la propia definición del problema la que introduce la sutil diferencia: en la formulación de un problema de estimación en una población finita no existe ningún elemento aleatorio, las probabilidades se introducen como parte de la solución (principio de aleatorización[1]).

Las definiciones del concepto de muestra en teoría de muestras y en inferencia estadística ponen claramente en evidencia las diferencias entre ambos tipos de problemas. La definición de muestra en el muestreo probabilístico puede formularse así (Cassel, Särndal y Wretman 1977):

Definición 13

Dada una población finita $U = \{1, \ldots, N\}$ de tamaño finito N, se define una ***muestra ordenada*** como una sucesión $os = (k_1, \ldots, k_{n(s)})$ tal que $k_i \in U$ para $i = 1, \ldots, n(s)$.

Dada una población finita $U = \{1, \ldots, N\}$ de tamaño finito N, se define una ***muestra no ordenada*** como un subconjunto no vacío $s \subseteq U$.

Comentario 56. Adviértase cómo en esta definición no existe ningún elemento que haga referencia a un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ subyacente. ■

En los problemas de inferencia estadística, la definición de muestra, a pesar de usar el mismo término, es matemáticamente muy diferente (Casella y Berger 2002):

Definición 14

Las variables aleatorias $X_1, \ldots, X_n$ se denominan una ***muestra aleatoria de muestra*** n ***de una población*** F_X si $X_1, \ldots, X_n$ son variables aleatorias mutuamente indepen-

[1] *Randomization principle*, en inglés.

dientes y la función marginal de distribución de cada variable X_i es la misma función F_X.

Comentario 57. Nótese cómo en la propia definición de muestra ahora sí es necesario disponer de un espacio de probabilidad subyacente (sobre el que están definidas las variables aleatorias X_i). ■

Esto tiene consecuencias directas en las expresiones de los estimadores que se usan para generar las estimaciones. Consideremos, por ejemplo, el muestreo aleatorio simple sin reemplazamiento de modo que el estimador HT viene dado por $\widehat{Y}_U^{\text{HT}} = \frac{N}{n}\sum_{k\in s} y_k$. En esta expresión, el único elemento aleatorio es la muestra s. Alternativamente, si expresamos $\widehat{Y}_U^{\text{HT}} = \frac{N}{n}\sum_{k\in s} y_k = \frac{N}{n}\sum_{k\in U} y_k I_k$, ahora los elementos aleatorios son las variables I_k. Los valores y_k de la variable Y son valores numéricos fijos. En los problemas de inferencia, el estimador media muestral viene dado por $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, donde ahora cada elemento Y_i es una variable aleatoria, que al realizar el experimento aleatorio, tomará valores numéricos y_i (diferentes en cada realización del experimento de acuerdo con la distribución F_Y).

Como se explica en los temas sobre estándares internacionales de producción (véase el tema 14 [2] del bloque "Producción Estadística Oficial: Principios básicos del ciclo de producción de operaciones estadísticas"del grupo de materias comunes), el proceso de producción estadística presenta varias fases: identificar las necesidades estadísticas, diseñar el proceso, construir las herramientas necesarias, recoger los datos, procesar los datos, analizar los datos, difundir los resultados y evaluar el proceso. De todas estas fases, la estimación basada en los modelos estadísticos atañe sobre todo a la selección de la muestra y a la construcción de los estimadores y su varianza para la inferencia.

10.2 Teoría de la predicción

10.2.1 Cuestiones generales

En esta sección planteamos los principios generales de la estimación basada en modelos estadísticos.

Supongamos que conocemos el número N de unidades en una población finita y que cada unidad tiene asociado un valor y_k. El problema general es elegir algunas de las unidades como una muestra, observar los valores y_k de las unidades muestrales y, después, usar estas observaciones para estimar el valor de alguna función $h(y_1, y_2, \ldots, y_N)$ de todos los valores y_k de la población. La función $h(y_1, y_2, \ldots, y_N)$ puede ser una simple combinación de los valores y_k, como la media o el total, o algo más complejo como los cuantiles. Nos centraremos en funciones que son combinaciones lineales de los valores

[2] Tema 14. Metadatos de la producción estadística. I. GSBPM. Introducción. El modelo. Relaciones con otros modelos y estándares. Niveles 1 y 2 del GSBPM. Descripciones de fases y subprocesos. Procesos generales. Otros usos del GSBPM.

y_k, como el total poblacional $\sum_{k\in U} y_k$.

El enfoque de la predicción trata a los valores $y_1, y_2, \ldots, y_N$ como ***valores de variables aleatorias*** $Y_1, Y_2, \ldots, Y_N$. Una vez se ha observado la muestra, estimar $h(y_1, y_2, \ldots, y_N)$ implica ***predecir*** una función de los valores y_k no observados. Para ello, las relaciones entre las variables aleatorias se expresan a través de un modelo de su función de probabilidad conjunta y las predicciones se calculan tomando como referencia este modelo. Usamos el término 'predicción' en el sentido de hacer una ***hipótesis*** sobre los valores y_k no observados, no en el sentido literal de predecir valores futuros. En la mayoría de las aplicaciones, los valores de Y se realizarán para todas las unidades de la población finita antes de que se seleccione la muestra (por ejemplo, todas las empresas tienen una cifra de negocios aunque solo observemos la de aquellas seleccionadas en la muestra). Después de seleccionar y observar una muestra, conoceremos los valores y_k de las unidades muestrales, pero los valores y_k de las unidades no muestrales siguen siendo desconocidas. Nuestro desconocimiento de los valores y_k no muestreados significa que debemos predecir matemáticamente alguna función de esos valores con el fin de tener un estimador o predictor del total de la población.

Para introducir el método de estimación, consideremos un ejemplo muy sencillo. Supongamos una población U de $N = 33$ unidades, de la cual se extrae una muestra s de $n_s = 32$ unidades y de la que queremos estimar el total poblacional $Y_U = \sum_{k\in U} y_k$. Denotamos el conjunto de unidades muestrales (muestra) por s y la unidad no seleccionada como k^*. El total poblacional se puede escribir como

$$Y_U = \sum_{k\in s} y_k + y_{k^*}.$$

La primera componente es conocida (se recogen los datos de toda la muestra[3]) y estimar Y_U equivale básicamente a predecir el valor y_{k^*}. Si la variable Y está correlacionada con otra variable X, cuyos valores conozcamos, entonces una forma natural es usar un modelo de regresión para predecir. Podemos ajustar el modelo de regresión a los valores muestrales observados y usar el modelo ajustado y los valores conocidos de X para predecir el valor desconocido de la variable aleatoria Y.

Ejemplo 41. Consideremos la tabla 10.1 de valores x e y. Ajustemos un sencillo modelo de regresión a los 32 elementos de la muestra y usemos este modelo para predecir el valor desconocido y_{k^*}. La recta de regresión lineal tiene coeficientes $\widehat{\beta}_0 = 1{,}620716$ y $\widehat{\beta}_1 = 1{,}231947$, de modo que $\widehat{Y}_{k^*} = 22{,}88411$. Esto nos permite estimar, según el enfoque esbozado anteriormente, mediante

$$\sum_{k\in s} y_k + \widehat{Y}_{k^*} = 617{,}93 + 22{,}88411 = 640{,}8141,$$

[3]No consideramos errores ajenos al muestreo como la falta de respuesta.

que supone un error relativo frente al total verdadero de $\frac{640{,}8141-638{,}58}{638{,}58} \approx 0{,}0035$, esto es, menos del $1\,\%$. ■

ID	x	y	sample
1	26,49	15,09	1
2	11,70	14,96	1
3	14,45	15,03	1
4	7,06	9,39	1
5	17,66	32,91	1
6	20,60	47,17	1
7	15,12	28,06	1
8	22,81	42,98	1
9	23,86	34,81	1
10	7,88	12,71	1
11	10,35	18,31	1
12	9,92	14,44	1
13	29,07	32,65	1
14	18,45	27,89	1
15	3,00	4,28	1
16	12,75	14,03	1
17	13,99	16,27	1
18	18,40	13,12	1
19	5,49	4,17	1
20	12,77	13,78	1
21	12,66	10,85	1
22	17,26	20,65*	0
23	1,62	2,57	1
24	9,84	14,19	1
25	5,44	9,08	1
26	6,60	3,32	1
27	23,00	32,29	1
28	1,21	1,92	1
29	26,19	21,26	1
30	16,45	33,45	1
31	8,28	11,12	1
32	16,85	28,98	1
33	29,53	36,85	1
	$Y_U =$	638,58	

Tabla 10.1: Estimación por modelos. Ejemplo.

Este mismo razonamiento se extiende al caso con más de una unidad no seleccionada en la muestra (como es el caso habitual). Si denotamos por r el conjunto de

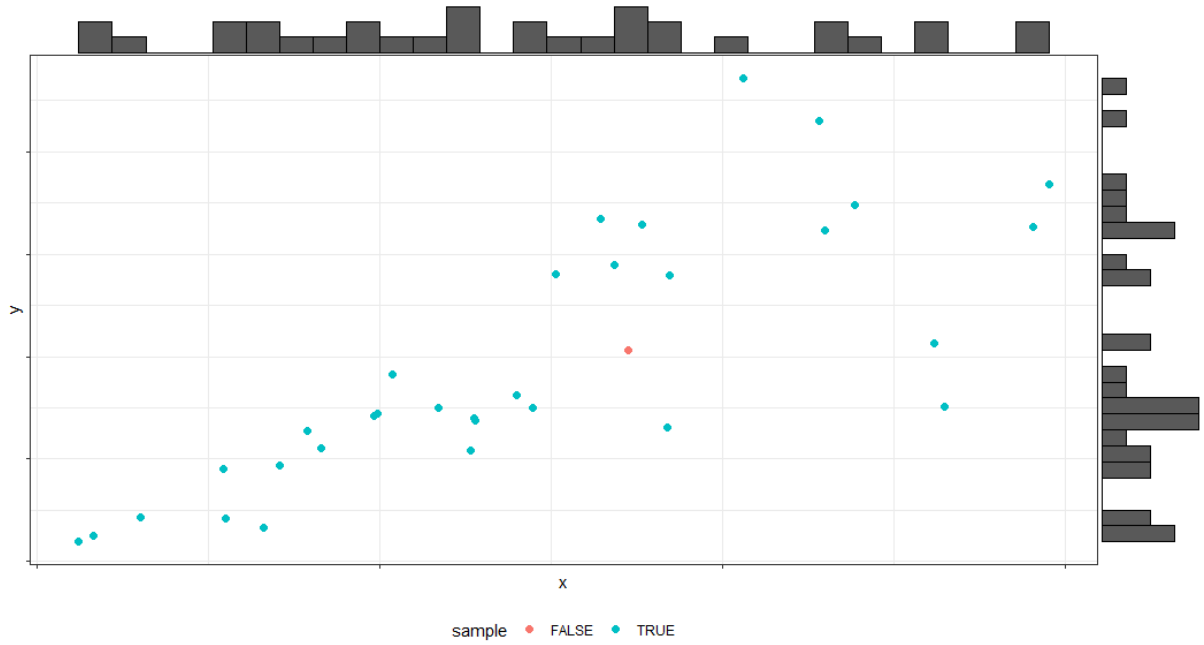

Figura 10.1: Relación entre x e y.

unidades no seleccionadas en la muestra (r de resto), el total poblacional está dado por $Y_U = \sum_{k\in s} y_k + \sum_{k\in r} y_k$. La tarea de estimar Y_U se convierte en la de predecir el valor $\sum_{k\in r} y_k$ para la variable aleatoria no observada $\sum_{k\in r} Y_k$. Adviértase que estamos suponiendo que los valores y_k son realizaciones de las variables aleatorias Y_k. Si denotamos por $\widehat{Y}_k$ el predictor para la unidad no muestral k, entonces un estimador de Y_U es $\widehat{Y}_U = \sum_{k\in s} y_k + \sum_{k\in r} \widehat{Y}_k$. El error de estimación es entonces la diferencia entre la predicción del total no muestral y su valor verdadero $\widehat{Y}_U - Y_U = \sum_{k\in r} \widehat{Y}_k - \sum_{k\in r} Y_k = \sum_{k\in r} \left(\widehat{Y}_k - Y_k\right)$.

Adviértase que el error absoluto $\widehat{Y}_U - Y_U$ es una variable aleatoria y, como tal, tiene momentos. En particular, pueden calcularse bajo el modelo m escogido su valor esperado $\mathbb{E}_m\left[\widehat{Y}_U - Y_U\right]$ y varianza $\mathbb{V}_m\left[\widehat{Y}_U - Y_U\right]$. La cantidad Y_U, en contra de lo supuesto en muestreo probabilístico, no es ahora una cantidad fija. Esto permite seguir una estrategia habitual en la construcción de estimadores: buscar un estimador óptimo en el sentido de ser insesgado ($\mathbb{E}_m\left[\widehat{Y}_U - Y_U\right] = 0$) y con mínima varianza ($\mathbb{V}_m\left[\widehat{Y}_U - Y_U\right]$). Esta búsqueda, en general, se realiza entre los estimadores que pueden escribirse como combinaciones lineales de las variables aleatorias Y_k:

$$\widehat{Y}_U = \sum_{k\in s} y_k + \sum_{k\in r} a_k Y_k.$$

Bajo hipótesis de normalidad, pueden entonces construirse intervalos de confianza para $\frac{\widehat{Y}_U - Y_U}{\sqrt{\mathbb{V}_m\left(\widehat{Y}_U - Y_U\right)}}$.

En el enfoque basado en modelos, no obstante, existen otros métodos adicionales de predicción. Podemos usar las técnicas de predicción bayesianas, que calculan la distribución a posteriori de Y_U dada la muestra de valores y_k. O se podrían usar métodos basados en la inferencia fiduciaria o, incluso, en algoritmos de aprendizaje automático. En cualquier caso, todos estos métodos tienen en común que reconocen que:

(i) una vez que las unidades muestrales han sido elegidas y sus valores y_k son conocidos, estimar el total poblacional Y_U es equivalente a predecir el valor de la suma de los valores y_k que no pertenecen a la muestra;

(ii) el modelo de la distribución de probabilidad conjunta de $Y_1, Y_2, \ldots, Y_N$ proporciona la base para la predicción.

Por contra, el enfoque que ha dominado el muestreo en poblaciones finitas desde mediados del siglo XX no está basado en modelos de predicción, sino en las distribuciones de probabilidad creadas cuando la selección de las unidades que componen la muestra se deja al azar. Este enfoque, como el enfoque predicción basado en modelos de regresión lineal, da lugar a una teoría en la que las inferencias se hacen en términos de sesgo, varianza y normalidad aproximada.

10.2.2 Predicción óptima

En línea con estas cuestiones generales, vamos a encontrar una expresión genérica para el estimador óptimo para el total poblacional $Y_U = \sum_{k \in U} y_k$ de una variable y en una población finita U de N unidades estadísticas. Como es habitual, emplearemos el error cuadrático medio como criterio de optimalidad, de modo que construiremos un estimador basado en un modelo m que haga que el error cuadrático medio $\mathbb{MSE}_m\left(\widehat{Y}_U - Y_U\right)$ sea mínimo. Recuérdese que bajo el modelo m tanto $\widehat{Y}_U$ como Y_U son variables estadísticas. Este error cuadrático medio puede descomponerse del modo habitual:

$$\mathbb{MSE}_m\left(\widehat{Y}_U - Y_U\right) = \mathbb{V}_m\left(\widehat{Y}_U - Y_U\right) + \left(\mathbb{E}_m\left(\widehat{Y}_U - Y_U\right)\right)^2.$$

Si nos restringimos a los estimadores insesgados respecto al modelo, entonces debemos encontrar el estimador $\widehat{Y}_U$ minimiza la varianza $\mathbb{V}_m\left(\widehat{Y}_U - Y_U\right)$. Al tratarse de variables aleatorias podemos usar el siguiente resultado en teoría de la probabilidad:

Teorema 17

Sean V, W variables aleatorias reales. El predictor de mínima varianza de W condicionado a la variable aleatoria V está dado por la esperanza condicionada $\mathbb{E}\left(W\middle|V\right)$.

Demostración 15

Véase, p.ej., Grimmet y Stirzaker 2004.

En nuestro caso de estimación en poblaciones finitas, este resultado se reduce a

Corolario 18

El predictor de mínimo error cuadrático medio del total poblacional Y_U en una población finita U de tamaño N es

$$\widehat{Y}_U = \mathbb{E}_m\left(Y_U \middle| y_k, k \in s; x_k, k \in U\right) = Y_s + \mathbb{E}_m\left(\sum_{k\in r} Y_k \middle| y_k, k \in s; x_k, k \in U\right), \quad (10.1)$$

donde s y r denotan la muestra y el resto de la población $r = U - s$ y x_k denota el valor de la variable auxiliar (posiblemente multidimensional) x para el elemento k de la población.

Demostración 16

Aplíquese el teorema anterior a las variables $\widehat{Y}_U$ e Y_U.

Obsérvese que la esperanza condicionada (10.1) dependerá de parámetros desconocidos (que definen el modelo estadístico m), por tanto, en la práctica no puede computarse. Bajo condiciones suficientemente generales, sin embargo, estos parámetros pueden ser estimados a partir de los elementos de la muestra $k \in s$, de modo que estas estimaciones se sustituyen en (10.1) produciendo un predictor *empirical best* $\widehat{Y}_U^{\text{EB}}$.

Ejemplo 42. **Población homogénea**
Un modelo general para una población homogénea empieza con el concepto de *intercambiabilidad*[4]. Las variables aleatorias cuya realización arroja los valores poblacionales y_k, $k \in U$, se dicen *intercambiables* hasta orden K si la distribución conjunta de $\{y_k\}_{k\in A}$ es la misma para cualquier permutación A de $k = 1, 2, \ldots, K$ etiquetas de la población.

Es fácil comprobar que en una población intercambiable todos los momentos de productos de los valores Y de la población hasta orden K son los mismos. En particular, si $K \geq 2$, entonces todas las unidades de la población tienen valores y con la misma media y la misma varianza y todos los pares de unidades diferentes de la población tienen la misma varianza. Una población así se denominará *homogénea de segundo orden* (a menudo solo *homogénea*). Se supondrán que los valores de unidades diferentes son independientes de modo que todas las covarianzas son 0.

El modelo para poblaciones homogéneas de segundo orden representan la unidad básica para construir modelos más complejos más cercanos a situaciones prácticas. Bajo este modelo, los valores y se modelizan mediante las especificaciones:

[4] *Exchangeability*, en inglés.

$$\begin{aligned} \mathbb{E}_m(Y_k) &= \mu, && (10.2a)\\ \mathbb{V}_m(Y_k) &= \sigma^2, && (10.2b)\\ Y_k, Y_l & \text{ independientes para todo } k \neq l. && (10.2c) \end{aligned}$$

Bajo estas hipótesis, podemos construir el estimador $\widehat{Y}_U$ para el total del población $Y_U = \sum_{k\in U} y_k$. Obsérvese que en este modelo carecemos de variables auxiliares x, por tanto, la esperanza condicionada (10.1) se reduce a

$$\widehat{Y}_U = Y_s + \mathbb{E}_m\left(\sum_{k\in s} Y_k | y_k, k \in s\right).$$

Para el modelo 10.2 esta expresión se reduce a

$$\widehat{Y}_U = Y_s + (N-n)\,\mu.$$

Como se indicó anteriormente, el parámetro μ necesita ser estimado. Bajo la hipótesis de intercambiabilidad parece natural estimar $\widehat{\mu} = \bar{y}_s$ de modo que

$$\widehat{Y}_U^{\text{EB}} = Y_s + (N-n)\,\bar{y}_s = N\bar{y}_s. \tag{10.3}$$

Este estimador también recibe el nombre de *estimador de expansión* y coincide con el estimador homónimo en el muestreo probabilístico. ■

Adviértase que el estimador EB no es único, puesto que no existe una única manera de estimar los parámetros desconocidos como μ. Proporciona un método sencillo de construir estimadores y será estadísticamente eficiente si se emplea un método de estimación razonable. Una alternativa más compleja consiste en encontrar el Mejor Predictor Lineal Insesgado (BLUP[5]). En muchas ocasiones, el estimador BLUP también es un estimador EB.

Para definir el estimador BLUP, se define primero un estimador linear como aquél que es combinación lineal de los variables aleatorias Y de las unidades muestrales. El estimador $\widehat{Y}_U^{\text{BLUP}}$ del total poblacional Y_U satisface las siguientes condiciones:

- Es un predictor lineal; esto es, se escribe de la forma $\widehat{Y}_U^{\text{BLUP}} = \sum_{k\in s} \omega_k Y_k$, donde los pesos ω_k están por determinar. No existen restricciones sobre estos pesos más allá de su no dependencia de las variables Y. Pueden depender, como suele suceder, de las unidades muestrales $k \in s$ y de las variables auxiliares X_k de estas unidades.
- Es insesgado para Y_U; esto es, cumple $\mathbb{E}_m\left(\widehat{Y}_U^{\text{BLUP}} - Y_U\right) = 0$.

[5] *Best Linear Unbiased Estimator*, en inglés.

- Para cualquier muestra s el error tiene mínima varianza; esto es, cumple

$$\mathbb{V}_m\left(\widehat{Y}_U^{\text{BLUP}} - Y_U\right) \leq \mathbb{V}_m\left(\widehat{Y}_U - Y_U\right)$$

  para cualquier otro predictor lineal $\widehat{Y}_U$.

Para llegar a una expresión del estimador BLUP, tan solo es necesaria la incorrelación (no necesariamente la independencia) de las variables Y_k. La derivación es sencilla. Por la linealidad, tenemos $\widehat{Y}_U = \sum_{k\in s}\omega_k Y_k = \sum_{k\in s} Y_k + \sum_{k\in s}(\omega_k - 1)Y_k = \sum_{k\in s} Y_k + \sum_{k\in s} u_k Y_k$ de modo que el error puede escribirse como

$$\widehat{Y}_U - Y_U = \sum_{k\in s} u_k Y_k - \sum_{k\in r} Y_k.$$

Imponemos ahora la condición de insesgadez, de modo que

$$\mathbb{E}_m\left(\widehat{Y}_U - Y_U\right) = \mu\left[\sum_{k\in s} u_k - (N-n)\right] = 0,$$

esto es,

$$\sum_{k\in s} u_k - (N-n) = 0.$$

Al minimizar la varianza llegamos a

$$\begin{aligned}
\mathbb{V}_m\left(\widehat{Y}_U - Y_U\right) &= \mathbb{V}_m\left(\widehat{Y}_{U-s} - Y_{U-s}\right) \\
&= \mathbb{V}_m\left(\widehat{Y}_{U-s}\right) - 2\cdot\mathbb{C}_m\left(\widehat{Y}_{U-s}, Y_{U-s}\right) + \mathbb{V}_m\left(Y_{U-s}\right) \\
&= \sigma^2\sum_{k\in s} u_k^2 - 0 + (N-n)\sigma^2.
\end{aligned}$$

Esta expresión será mínima para aquellos valores de u_k (por tanto, de ω_k) que minimicen $\sum_{k\in s} u_k^2$ sujetos a la restricción $\sum_{k\in s} u_k - (N-n) = 0$. Con las técnicas usuales de optimización se obtiene que $u_k = \frac{N-n}{n}$ y, por tanto, $\omega_k = \frac{N}{n}$. El estimador BLUP es, por tanto, nuevamente el estimador de expansión

$$\widehat{Y}_U^{\text{BLUP}} = N\bar{y}_s.$$

En la misma línea puede encontrarse una expresión para la varianza y construir, por tanto, intervalos de confianza. Usando $u_k = \frac{N-n}{n}$, siguiendo las mismas expresiones anteriores se llega inmediatamente a

$$\mathbb{V}_m\left(\widehat{Y}_U^{\text{BLUP}} - Y_U\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sigma^2.$$

Para construir los intervalos de confianza es preciso estimar σ^2, lo que puede hacerse de modo insesgado mediante la cuasivarianza muestral, de modo que un estimador insesgado de la varianza de predicción está dado por

$$\widehat{V}_m\left(\widehat{Y}_U^{\text{BLUP}}\right) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)S_{ys}^2,$$

donde $S_{ys}^2 = \frac{1}{n-1}\sum_{k\in s}(y_k-\bar{y}_s)^2$. Para muestras de tamaño grande puede aplicarse como es común el teorema central de límite de modo que el estadístico z definido por

$$Z = \frac{\widehat{Y}_U^{\text{BLUP}} - Y_U}{\sqrt{\widehat{\mathbb{V}}_m\left(\widehat{Y}_U^{\text{BLUP}}\right)}}$$

se distribuye como una variable aleatoria normal $N(0,1)$. Por tanto, un intervalo de confianza aproximado del $100(1-\alpha)\,\%$ para Y_U está dado por

$$\widehat{Y}_U^{\text{BLUP}} \pm q_{\alpha/2}\sqrt{\widehat{\mathbb{V}}_m\left(\widehat{Y}_U^{\text{BLUP}}\right)},$$

donde $q_{\alpha/2}$ es el cuantil $(1-\alpha/2)$ de la distribución normal estándar. Como Y_U es una variable aleatoria, a menudo nos referimos a este intervalo como un *intervalo de predicción*.

Este procedimiento de construcción de estimadores EB y BLUP puede extenderse con relativa facilidad para modelizar poblaciones más complejas dando lugar a los modelos de poblaciones homogéneas estratificadas, de poblaciones con estructura de regresión, de poblaciones conglomeradas. Véase (R.R. Chambers 2012) para los detalles.

10.3 Comparación con la teoría del muestreo probabilístico en poblaciones finitas

10.3.1 La teoría del muestreo probabilístico

Un *diseño muestral probabilístico* es un esquema de selección de muestra de forma que cada subconjunto s de unidades de la población U tiene una probabilidad $p(s)$ conocida de selección. Para un vector poblacional $\mathbf{y} = (y_1,\ldots,y_N)^t$ y un estimador $\widehat{Y}_U(s;\mathbf{y})$ dados, cada posible muestra proporciona un valor del estimador. La probabilidad de que el estimador tome un valor particular es la probabilidad de que una de las muestras dando ese valor es seleccionada. Las definiciones de sesgo y varianza se establecen en función de las probabilidades de selección $p(s)$:

$$\mathbb{E}_p\left(\widehat{Y}_U\right) - Y_U = \sum_{s\in\mathcal{S}} p(s)\widehat{Y}_U(s;\mathbf{y}) - \sum_{k\in U} y_k, \tag{10.4a}$$

$$\mathbb{V}_p\left(\widehat{Y}_U\right) = \sum_{s\in\mathcal{S}} p(s)\left(\widehat{Y}_U(s;\mathbf{y}) - \mathbb{E}_p\left(\widehat{Y}_U\right)\right)^2. \tag{10.4b}$$

Puesto que el sesgo, la varianza y otras cantidades estadísticas se calculan obteniendo la media de ***todas las muestras que podrían obtenerse*** para un diseño de muestreo particular, a veces se refieren a ellas como propiedades *basadas en el diseño*. En este enfoque el único elemento aleatorio es el conjunto s, que indica qué unidades están incluidas en la muestra.

Ejemplo 43. Consideremos una población U de $N = 4$ elementos. Seleccionamos una muestra probabilística mediante un diseño muestral de Bernouilli de parámetro $\pi = \frac{1}{2}$. Supongamos que al realizar la selección se obtiene la muestra $s = U$, lo que puede ocurrir con probabilidad $\pi^N = \left(\frac{1}{2}\right)^4$. Según la teoría general de muestreo vista en el resto de temas, el estimador de Horvitz-Thompson arroja una estimación $\widehat{y}^{HT} = \pi * y_s = 2 * Y_U$, que tiene un error del $100\,\%$. La varianza de este estimador es $\mathbb{V}\left(\widehat{Y}_U^{\mathrm{HT}}\right) = \left(\frac{1}{\pi} - 1\right) \sum_{k \in U} y_k^2 > 0$, que es estrictamente positiva. ¿Cómo es posible que teniendo los valores y_k de todos los elementos de la población la estimación del total sea tan mala y, además, la varianza sea estrictamente positiva (conocemos todos los valores)? La respuesta es que la inferencia con los diseños muestrales se basan en ***todas las muestras que podrían obtenerse*** y no únicamente en la muestra seleccionada. ■

Ejemplo 44 (**Los elefantes de Basu**). Un ejemplo histórico que muestra las limitaciones del muestreo probabilístico es la siguiente historia debida a Basu 1971:

> El propietario de un circo está planificando embarcar sus 50 elefantes, para lo que necesita tener una estimación aproximada del peso total de los elefantes. Como pesar un elefante es un proceso engorroso, el propietario quiere hacer la estimación del peso total pesando únicamente un elefante. ¿Qué elefante elegir? Así pues, el propietario consulta sus registros y descubre una lista con los pesos de todos los elefantes de hace 3 años. Descubre que hace 3 años, Sambo, el elefante de tamaño medio, tenía el peso medio de la manada. Consulta con el entrenador de los elefantes, quien le asegura que Sambo puede aún considerarse el elefante medio de la manada. Por tanto, el propietario planifica pesar a Sambo y estimar el peso total $Y_U = \sum_{k \in U} y_k$ mediante $50 \times y_{\text{Sambo}}$ (donde y_{Sambo} denota el peso de Sambo).
>
> Pero el estadístico del circo se horroriza cuando conoce el plan de muestreo intencionado[6]. "¿Cómo puedes obtener una estimación insesgada de Y_U de este modo?", protesta el estadístico. De este modo, juntos trabajan en un plan de muestreo de compromiso. Con la ayuda de una tabla de números aleatorios[7] idean un plan que afija una probabilidad de selección $\pi_{\text{Sambo}} = \frac{99}{100}$ para Sambo y una probabilidad de selección igual $\pi_k = \frac{1}{49}\frac{1}{100}$ para el resto de elefantes. Naturalmente, Sambo es seleccionado y el propietario es feliz. "¿Cómo vas a estimar Y_U?", pregunta el estadístico. "¿Por qué? La estimación debe ser $50 \times y_{\text{Sambo}}$, por supuesto", dice el propietario. "¡Oh! ¡No! Eso no puede ser correcto", dice el estadístico, "He leído recientemente

[6] *Purposive sampling*, en inglés.
[7] Hoy día esta parte se haría con un generador de números pseudoaleatorios en una computadora.

un artículo en *Annals of Mathematical Statistics* donde se demuestra que el estimador de Horvitz-Thompson es el único estimador hiperadmisible en la clase de todos los estimadores insesgados polinomiales generalizados". "¿Cuál es la estimación de Horvitz-Thompson en este caso?", pregunta el propietario, ciertamente impresionado. "Como la selección de probabilidad de Sambo en nuestro diseño muestral es $\frac{99}{100}$", dice el estadístico, "la estimación apropiada para Y_U es $\widehat{Y}_U^{\text{HT}} \frac{y_{\text{Sambo}}}{\pi_{\text{Sambo}}} = \frac{100}{90} \times y_{\text{Sambo}}$"."Y ¿cómo habrías estimado Y_U", pregunta el incrédulo propietario, "si con nuestro plan de muestreo hubiese sido seleccionado, digamos, el gran elefante Jumbo?". "De acuerdo con mi entendimiento del método de estimación de Horvitz-Thompson", dice el infeliz estadístico, "la estimación apropiada de Y_U hubiese sido $\widehat{Y}_U^{\text{HT}} = \frac{y_{\text{Jumbo}}}{\frac{1}{100} \times \frac{1}{49}} = 4900 \times y_{\text{Jumbo}}$", donde y_{Jumbo} es el peso de Jumbo. Así es como el estadístico perdió su empleo en el circo (¡y quizá se convirtió en profesor de Estadística!).

■

10.3.2 ¿Qué enfoque usar?

Si tenemos opciones para la inferencia (teoría de predicción, teoría de muestreo probabilístico, o quizá, una mezcla de ambas), ¿cuál deberíamos usar? No hay ninguna duda sobre la validez matemática de cualquiera de las dos teorías. La cuestión clave es cuál es más apropiada para la inferencia a partir de una muestra observada y una estimación calculada. En el ejemplo anterior se empleó un modelo de regresión lineal como un modelo tentativo y aproximado, pues nunca conocemos el modelo estadístico que generan los datos y_k. En contraste, nuestro control y conocimiento de una probabilidad de selección muestral $p(\cdot)$ en el enfoque basado en diseños es completo (al menos, en principio). Por tanto, es más convincente a priori basar la inferencia en esta última opción, manteniendo los resultados tan independientes como sea posible de modelos de predicción más sujetos a errores (de especificación).

La cuestión de la falibilidad inevitable de los modelos estadísticos es una cuestión central en la teoría de predicción para el muestreo y la estimación basada en modelos en general. Existen cuestiones fundamentales que se encuentran en la controversia entre la inferencia basada en modelos estadísticos y en diseños muestrales. Existe importante literatura al respecto (Royall 1968; Royall y Herson 1973a; Royall y Herson 1973b; Royall 1976b; Royall 1976a; Royall 1976c; Royall y Cumberland 1978; Cumberland y Royall 1981; Hansen, Madow y Tepping 1983; Royall y Cumberland 1981; Royall 1992; Brewer 1994; Brewer 1999; Smith 1976; Smith 1983; Smith 1994; Smith 1999; Smith 2001). Incluimos algunas de estas cuestiones importantes a continuación.

Retomemos el Ejemplo 41 y supongamos que para hacer la estimación del total poblacional Y_U usamos un diseño muestral aleatorio simple sin reemplazamiento. Empleamos el correspondiente estimador de Horvitz-Thompson, que en este caso se reduce al estimador de expansión $\widehat{Y}_U^{\text{HT}} = N\widehat{y}_s = 33 \times \bar{y}_s$, donde $\bar{y}_s$ denota la media muestral. Este estima-

dor, como sabemos, es insesgado. Y es insesgado independientemente de la existencia entre las variables x e y. Supongamos que la muestra seleccionada es de tamaño $n = 5$ y corresponde a las unidades con los valores x más pequeños (unidades 28, 23, 15, 25 y 19). Para esta muestra, tenemos $\bar{y}_s = 4{,}404$ y, por tanto, $\widehat{Y}_U^{\text{HT}} = 145{,}332 \ll Y_U = 638{,}58$, que es una muy mala estimación para el total poblacional. En este caso, aun sabiendo que hemos tenido mala suerte en la selección (aleatoria) de la muestra, mantenemos que el estimador $\widehat{Y}_U^{\text{HT}}$ es insesgado para Y_U y, por tanto, tiene buena calidad.

Ejemplo 45. Quizá pueda pensarse que escoger un ejemplo extremo como el caso de la muestra con las 5 unidades con valor x más bajo es demasiado extremo para emplearlo como argumento en liza frente a la insesgadez del estimador de Horvitz-Thompson y la calidad de la estimación final. Pero si extraemos 10^6 muestras aleatorias simples sin reemplazamiento de la misma población del Ejemplo 41, calculamos $\widehat{Y}_U^{\text{HT}}$ para cada muestra y representamos la distribución en el muestreo de este valor (véase la Figura 10.2), observamos que la mitad de los valores se encuentran fuera del intervalo $(520{,}344, 749{,}694)$, todas estimaciones muy pobres del valor $Y_U = 638{,}58$. Aun así, se sigue aludiendo a la insesgadez respecto al diseño muestral $p(\cdot)$ como un argumento de calidad suficiente. ■

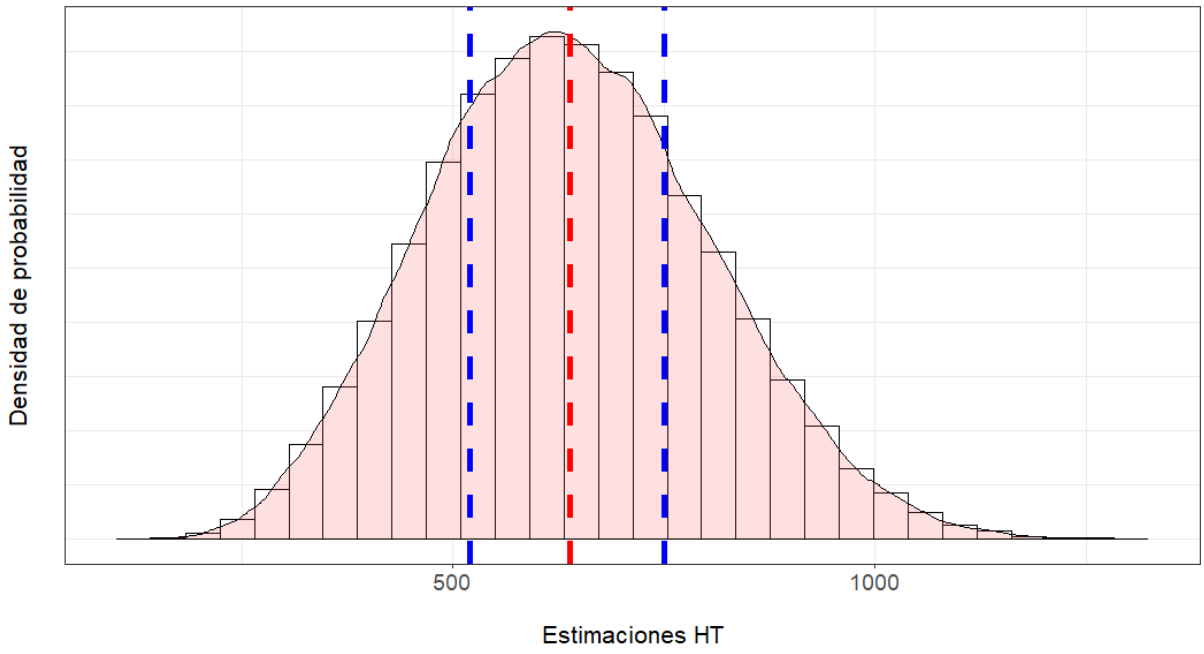

Figura 10.2: Estimaciones de Horvitz-Thompson para 10^6 muestras.

Al considerar, por tanto, la cuestión del sesgo como medida de calidad, parece evidente que $\widehat{Y}_U^{\text{HT}} = N\bar{y}_s$ puede presentar sesgo en cierta manera a definir a continuación. A partir de la Figura 10.1, parece evidente que existe una relación lineal entre x e y, de modo que podemos escribir $y = \beta x$ para algún valor positivo de β. Por tanto, podemos escribir

$$\mathbb{E}_m \left(N\bar{y}_s - N\bar{y}_U \right) = N\beta \left(\bar{x}_s - \bar{x}_U \right).$$

Por tanto, si la muestra s seleccionada tiene una media muestral de la variable auxiliar

x superior al valor de la media poblacional, el sesgo *respecto al modelo* será positivo. Complementariamente, si la media muestral es inferior, el sesgo *respecto al modelo* será negativo. Esto puede explicar la mala calidad de las estimaciones fuera del intervalo $(520{,}344, 749{,}694)$ indicadas más arriba. Luego, aparentemente el concepto de sesgo respecto al modelo parece más relevante que el de sesgo respecto al diseño muestral. El ejemplo extremo de las 5 unidades seleccionadas con el valor x más pequeño se reproduce también con cualquier otra muestra con valores $\bar{x}_s$ muy alejados de $\bar{x}_U$.

Un principio fundamental que rige la inferencia estadística es el llamado *Principio de Condicionalidad*, esto es, que la inferencia sobre un fenómeno aleatorio se base en las variables aleatorias observadas cuya distribución de probabilidad es conocida y, por tanto, no dependa de parámetros sobre los que debemos establecer la inferencia. Para ilustrar este concepto, supongamos, de acuerdo con el muestreo probabilístico, que extraemos una muestra aleatoria simple sin reemplazamiento de tamaño $n = 100$ de una población U de $N = 1000$ unidades. La varianza del estimador HT de la media poblacional está dada por

$$\mathbb{V}(\bar{y}_s) = \left(1 - \frac{100}{1000}\right) \frac{S^2_{yU}}{100}, \tag{10.5}$$

donde S^2_{yU} denota la cuasivarianza poblacional de la variable y. Supongamos que, antes de extraer la muestra, se considera la posibilidad de realizar un censo de toda la población, pero por las restricciones presupuestarias no somos capaces de decidir si invertir esta cantidad de dinero en este estudio o en otro alternativo igualmente importante. Para decidir, lanzamos una moneda al aire. El tamaño muestral en estas condiciones es claramente aleatorio: $n = 100$ con probabilidad $1/2$ y $n = 1000$ con probabilidad $1/2$. La varianza del estimador $\bar{y}_s$ es, por tanto,

$$\mathbb{V}(\bar{y}_s) = \frac{1}{2}\mathbb{V}(\bar{y}_s|n = 100) + \frac{1}{2}\mathbb{V}(\bar{y}_s|n = 1000) = \frac{1}{2} \times \left(\left(1 - \frac{100}{1000}\right) \frac{S^2_{yU}}{100}\right). \tag{10.6}$$

Las dos varianzas son matemáticamente correctas, pero ¿qué varianza debe usarse? Debemos usar la varianza condicional (10.5) porque realiza la inferencia sobre el tamaño verdadero de la muestra observada. La varianza incondicional (10.6) subestima la incertidumbre de la estimación porque se basa en información que no tiene lugar (el censo que nunca se realiza). La varianza incondicional es probabilística correcta pero inferencialmente errónea. No es correcto emplearla como una herramienta para interpretar y comunicar la incertidumbre de nuestra estimación de la media poblacional. No es correcto porque no distingue entre la incertidumbre cuando $n = 100$ y cuando $n = 1000$, algo que, sin embargo, sí conocemos.

Por tanto, parece evidente que la inferencia sobre una población finita debe estar basada condicionalmente sobre los valores de la muestra observada, es decir, el sesgo, el error estándar, el coeficiente de variación, etc. deben estar basados sobre la muestra realmente

observada y no sobre un promedio de todas las posibles muestras.

Otro ejemplo que ilustra las limitaciones del muestreo probabilístico en relación con el Principio de Condicionalidad es el siguiente. Consideremos un estadístico A que emplea un diseño muestral aleatorio simple con reemplazamiento de tamaño muestral $n = 25$ de una población U de $N = 100$ unidades. La muestra seleccionada resulta no tener unidades duplicadas a pesar del reemplazamiento. El estadístico A emplea el estimador de Hansen-Hurwitz para estimar el total poblacional de modo que su varianza es

$$\mathbb{V}_p \widehat{Y}_U^{\mathrm{HH}} = \frac{N^2}{n} S_{yU}^2.$$

Supongamos que otro estadístico B emplea independientemente un diseño muestral aleatorio simple sin reemplazamiento, obteniendo exactamente la misma muestra que el estadístico A. Ahora, este segundo estadístico utiliza el estimador de Horvitz-Thompson cuya varianza está dada por

$$\mathbb{V}_p \left(\widehat{Y}_U^{\mathrm{HT}} \right) = \left(1 - \frac{n}{N} \right) \frac{N^2}{n} S_{yU}^2.$$

Adviértase que la estimación en ambos casos es la misma: $N\bar{y}_s$. Sin embargo, la varianza en el caso con reemplazamiento es $1{,}33$ veces mayor que en el caso sin reemplazamiento, incluso a pesar de usar la misma muestra y resultar la misma estimación. ¿Realmente es la estimación de B más precisa que la de A? Este tipo de resultado, nuevamente, se produce por la violación del Principio de Condicionalidad.

Existen otros dos argumentos de índole más técnica que ilustran igualmente las limitaciones del muestreo probabilístico. Por un lado, Basu 1971 demostró que para el problema de estimación en poblaciones finitas no existe ningún estimador lineal insesgado óptimo (de mínima varianza), en contra de las situaciones comunes en otros campos de la Estadística. Por otro lado, si se calcula la función de verosimilitud para el problema de estimación en poblaciones finitas, se obtiene una función constante (Cassel, Särndal y Wretman 1977), por tanto, no permite realizar inferencia válida alguna sobre los valores de y para las unidades no seleccionadas en la muestra. Si se pesan bien un hacha, un asno o una caja de herraduras, su peso no permite realizar inferencia alguna sobre el peso de los otros dos objetos. Y esto es así independientemente de cómo se han empleado los números aleatorios para seleccionar qué objeto pesar. Si debemos aprender algo sobre las unidades no seleccionadas en la muestra a partir de las observaciones de las unidades muestrales, entonces ambos grupos de unidades deben tener una conexión lógica más fuerte que la que pueda obtenerse por efecto del azar (de selección de la muestra).

10.3.3 ¿Por qué usar muestreo aleatorio?

El Principio de Aleatorización establece que el muestreo aleatorio es la condición *sine qua non* para la inferencia en poblaciones finitas. Si rechazamos este principio, entonces

surge la cuestión: "¿por qué usar muestreo aleatorio?". Existen varios argumentos para la aleatorización. Algunos son razonables mientras otros, no tanto.

Una posición extrema (y poco razonable) es la establecida por el Principio de Aleatorización, esto es, que la aleatorización artificial proporciona la única base para la rigurosa inferencia probabilística y que, en ausencia de la aleatorización, es imposible establecer inferencias probabilísticas válidas.

En contraste con esta visión, hemos visto con anterioridad algunos ejemplos de que la distribución de probabilidad determinada por la aleatorización artificial no es apropiada incluso cuando está disponible. Afirmar que, en general, las inferencias probabilísticas no son válidas cuando la distribución de la aleatorización no está disponible es simplemente erróneo. Esto no implica negar el valor de la aleatorización, sino tan solo negar que representa la base para toda la inferencia probabilística válida y rigurosa. Una de las razones acertada para la aleatorización es asegurar la imparcialidad. Por ejemplo, los árbitros en fútbol lanzan una moneda al aire para decir quién tiene la primera posesión del balón. El otro equipo puede maldecir su suerte, pero en cualquier caso el árbitro está protegido de acusaciones de parcialidad. La aleatorización para la selección de muestras o la afijación de tratamientos puede tener una función similar, especialmente cuando los resultados de la encuesta se usan en decisiones controvertidas. La aleatorización puede proporcionar protección contra la acusación de que una muestra se ha seleccionado deliberadamente para apoyar determinado punto de vista. La aleatorización puede también ser un escudo efectivo contra las múltiples fuentes de sesgo personal inconsciente que podrían aparecer en una selección de muestra basada en el juicio personal, incluso, de expertos. Estos sesgos son notoriamente insidiosos y a menudo imposibles de corregir mediante tratamientos estadísticos cuando son detectados.

Uno de los roles más importantes de la aleatorización en las encuestas por muestreo es similar al del diseño de experimentos. En un experimento, si los grupos de placebo y tratamiento difieren con respecto a una variable objetivo clave, entonces queremos atribuir tal diferencia al efecto del tratamiento. Pero si los grupos difieren con respecto a alguna otra variable importante, entonces nuestra conclusión se debilita. Esta otra variable importante, y no el tratamiento, puede explicar la diferencia observada. Estas "otras variables" están siempre presentes en los estudios observacionales, donde por definición siempre hay una auto-selección de unidades en la composición de los grupos. Los grupos de fumadores y no fumadores pueden ser comparables con respecto a la edad, el sexo y la ocupación, pero inevitablemente difieren en relación con cualquier factor social, física y psicológica que conduce al tabaquismo.

Los experimentos presentan la ventaja de que pueden evitar las diferencias de grupo que se crean cuando hay auto-selección (como en los estudios observacionales). Pero cuando las unidades estadísticas se asignan a los grupos de comparación (digamos, grupos de placebo y tratamiento) por el experimentador que observa que los grupos son comparables con respecto a las variables A y B, no es necesario que un escéptico esté de

acuerdo en que las diferencias observadas representen efectos del tratamiento. Puede plantear la hipótesis que los grupos difieren con respecto a otra variable C y desarrollar una explicación plausible de los resultados en términos de C, no del tratamiento. Si los datos disponibles no nos permiten examinar los grupos en términos de las diferencias respecto a C, entonces las conclusiones se tornan una disputa.

El experimentador se encuentra en una situación ventajosa si las unidades se han dividido en los grupos de placebo y tratamiento de modo aleatorio de modo que existe una alta probabilidad de obtener comparabilidad entre ambos en términos de C. En estas circunstancias, es el escéptico el que debe aportar argumentos a su favor: en ausencia de evidencia contraria, hay buenas razones para creer que los grupos son comparables. Hay buenas razones para cree que lo que suele pasar (comparabilidad en términos de C) también ha sucedido en esta ocasión con los grupos establecidos, a menos que exista una evidencia clara que indique lo contrario. El escéptico debe proporcionar argumentos en contra.

En el caso de muestreo en poblaciones finitas, se necesita comparabilidad entre las unidades estadísticas seleccionadas en la muestra y las no seleccionadas para ser capaces de establecer inferencias sobre la población completa. Si existe evidencia de que la muestra y el resto de la población difieren en algunas características conocidas, esto debe ser tenido en cuenta al construir los estimadores incluso si se ha empleado la aleatorización para seleccionar la muestra.

En la estimación en poblaciones finitas, debemos usar estimadores como los estimadores de razón (véase el tema 3) y de regresión (GREG) (véase el tema 4) o estimadores estratificados para dar cabida a la estructura de la población y algunos desequilibrios entre la muestra y el resto de la población.

Si dependemos por completo de la distribución generada por la aleatorización para hacer inferencias, entonces estamos inclinados a aceptar la posición insostenible de ignorar las propiedades particulares de la muestra seleccionada, violando, por tanto, el Principio de Condicionalidad. Si existen variables que afectan a la variable objetivo a estimar Y, pero carecemos de conocimiento sobre estas variables, entonces no podemos tener en cuenta tales peculiaridades. Pero no podemos ignorar aquellas variables que sí conocemos.

La aleatorización es deseable, pero no es ni necesaria ni suficiente para una inferencia estadística rigurosa. La inferencia válida puede llevarse a cabo sin ella y, cuando está presente, no crea necesariamente el marco probabilístico apropiado para la inferencia. La aleatorización puede proporcionar una garantía razonable de equilibrio con respecto a los factores incontrolados, pero no garantiza equilibrio y no justifica ignorar la evidencia de que existan desequilibrios. No hay nada en la naturaleza de la aleatorización que libere al estadístico de la responsabilidad de identificar las peculiaridades en la muestra y entender por qué están ahí.

Para terminar, incluimos algunas reflexiones finales sobre la cuestión de la inferencia y las técnicas empleadas en la producción estadística oficial.

En primer lugar, debemos mencionar que ya se estableció un debate a través de la literatura científica (véase, p.ej., Smith 1976; Hansen, Madow y Tepping 1983; Smith 1999; Brewer 1999, y múltiples referencias allí citadas). La posición final de la comunidad de estadísticos oficiales básicamente se resume en que los diseños muestrales liberan al estadístico de realizar hipótesis sobre la generación de los valores de la población, hipótesis que luego deben ser justificadas dada la utilidad y finalidad de estas estadísticas. Se reconoce la superioridad de los modelos en términos de la acuracidad *si el modelo escogido para la población es correcto* (Hansen 1987), extremo que a menudo es difícil de asegurar con suficiente certidumbre (Hansen, Madow y Tepping 1983). No obstante, esto no debe tomarse como una garantía de los diseños muestrales frente a los modelos estadísticos tampoco. El ejemplo de los elefantes de Basu muestra que las estimaciones basadas en diseños muestrales pueden igualmente tener muy baja calidad *si el diseño muestral escogido para la muestra no es bueno*. Decía Box 1976 que "todos los modelos son incorrectos, pero algunos son útiles"; bien puede afirmarse lo mismo de los diseños muestrales.

En segundo lugar, las nuevas fuentes de datos tanto digitales como administrativas limitan el uso de los diseños muestrales puesto que el estadístico no puede seleccionar las unidades muestrales (mecanismo de selección de la muestra desconocido[8]), por lo que los modelos estadísticos aparecen como el principal recurso para establecer la inferencia. Adviértase que las técnicas de aprendizaje automático (Murphy 2012) son en realidad un uso extendido de modelos estadísticos.

En tercer lugar, la estadística bayesiana ha realizado avances notables tanto desde el punto de vista del análisis de datos y la modelización como desde el punto de vista computacional. Su uso en la producción de estadísticas oficiales empieza a ser objeto de análisis (Fienberg 2011; Little 2012).

Por último, la abundancia de datos, muy conectada con la reciente Ciencia de Redes[9], está permitiendo la creación de planes de muestreo donde se tenga en cuenta la relación entre los elementos de la población, relaciones que se expresan mediante la teoría de grafos. Técnicas como el muestreo en grafos[10] (Zhang y Patone 2017) o el muestreo de conglomerados adaptativo (Turk y Borkowski 2005) deberán ser tenidas en cuenta para la producción de estadísticas oficiales.

[8] Por ejemplo, la información de la web obtenida por técnicas de *web scraping*.
[9] *Network Science*, en inglés.
[10] *Graph sampling*, en inglés.

Bibliografía

Basu, D. (1971). "An essay on the logical foundations of survey sampling. Part I". En: *in* Foundations of Statistical Inference, *Toronto: Holt, Rinehart and Winston, pp. 203-242, 1971*.

Box, G.E.P. (1976). "Science and Statistics". En: *Journal of the American Statistical Association* 71, págs. 791-799.

Brewer, K.R.W. (1994). "Survey sampling inference: some past perspectives and present prospects". En: *Pakistan J. Stat.* 10, págs. 15-30.

– (1999). "Design-based or prediction inference? Stratified random vs. stratified balanced sampling". En: *International Statistical Review* 67, págs. 35-47.

Casella, G. y R.L. Berger (2002). *Statistical Inference*. Duxbury Press.

Cassel, C.-M., C.-E. Särndal y J.H. Wretman (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley.

Cumberland, W.G. y R.M. Royall (1981). "Prediction models and unequal probability sampling". En: *Journal of the Royal Statistical Society B* 43, págs. 353-367.

Deming, W.E. (1950). *Some theory of sampling*. New York: Wiley.

Fienberg, S.E. (2011). "Bayesian models and methods in public policy and government settings". En: *Statistical Science* 26, págs. 212-226.

Grimmet, G.R. y D.R. Stirzaker (2004). *Probability and random processes*. 3rd. Oxford Science Publications.

Hansen, M.H. (1987). "Some history and reminiscences on survey sampling". En: *Statistical Science* 2, págs. 180-190.

Hansen, M.H., W.G. Madow y B.J. Tepping (1983). "An evaluation of model-dependent and probability sampling inferences in sample surveys". En: *Journal of the American Statistical Association* 78, págs. 776-793.

Little, R.J. (2012). "Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics". En: *Journal of Official Statistics* 28, págs. 309-334.

Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

R.R. Chambers, R.G. Clark (2012). *An introduction to model-based survey sampling with applications*. New York: Oxford University Press.

Royall, R.M. (1968). "An old approach to finite population sampling theory". En: *J. Amer. Stat. Assoc.* 63, págs. 1269-1279.

– (1976a). "Current advances in sampling theory: Implications for human observational studies". En: *Am. J. Epidemiology* 104, págs. 463-473.

– (1976b). "Likelihood functions in finite population sampling theory". En: *Biometrika* 63, págs. 605-614.

– (1976c). "The linear least squares prediction approach to two-stage sampling". En: *Journal of the American Statistical Association* 71, págs. 657-664.

– (1992). "The model based (prediction) approach to finite population sampling theory". En: *Lecture Notes-Monograph Series. Current Issues in Statistical Inference: Essays in Honor of D. Basu* 17, págs. 225-240.

Royall, R.M. y W.G. Cumberland (1978). "Variance estimation in finite population sampling". En: *J. Amer. Stat. Assoc.* 73, págs. 351-358.

– (1981). "An empirical study of the ratio estimator and estimators of its variance". En: *Journal of the American Statistical Association* 76, págs. 66-77.

Royall, R.M. y J. Herson (1973a). "Robust estimation in finite populations I". En: *Journal of the American Statistical Association* 68, págs. 880-889.
– (1973b). "Robust estimation in finite populations II". En: *Journal of the American Statistical Association* 68, págs. 890-893.
Smith, T.M.F. (1976). "The foundations of survey sampling: a review". En: *J. R. Stat. Soc. A* 139, págs. 183-204.
– (1983). "On the validity of inferences from non-random sample". En: *J. R. Stat. Soc. Series A* 146, págs. 394-403.
– (1994). "Sample Surveys 1975-1990: An age of reconciliation?" En: *International Statistical Review* 62, págs. 5-19.
– (1999). "Recent developments in sample survey theory and their impact of Official Statistics". En: *Bulletin of the International Statistical Institute LVIII*.
– (2001). "Biometrika century: sample surveys". En: *Biometrika* 88, págs. 167-194.
Turk, P. y J.J. Borkowski (2005). "A review of adaptive cluster sampling: 1990–2003". En: *Environmental and Ecological Statistics* 2005, págs. 55-94.
Valliant, R., A.H. Dorfmann y R.M. Royall (2000). *Finite population sampling and inference. A prediction approach*. New York: Wiley.
Zhang, L-C. y M. Patone (2017). "Graph sampling". En: *METRON* 75, págs. 277-299.

Tema 11

Métodos para el desarrollo, testeo y evaluación de instrumentos de recogida de datos. Un marco para el desarrollo, testeo y evaluación. Desarrollo de contenido, medidas y cuestiones en encuestas. Testeo de preguntas y cuestionarios. Evaluación de preguntas y cuestionarios. Desarrollo, testeo y evaluación de instrumentos de recogida electrónica de datos. Análisis de datos cuantitativos. Enfoques multimétodo para el desarrollo, testeo y evaluación. Organización y logística.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía.

G. Snijkers, G. Haraldsen y col. (2013). *Designing and Conducting Business Surveys*. New York: Wiley.

11.1 Un marco para el desarrollo, testeo y evaluación

Nuestro marco para llevar a cabo los procesos de desarrollo, testeo y evaluación se basa en el enfoque del *Total Survey Error* (TSE) (Groves y col. 2004), en el que los pasos para desarrollar medidas de encuestas están asociadas con las fuentes de los errores de la encuesta en cada fase. La Figura 11.1 ilustra este marco.

La recogida de datos puede considerarse como mediciones de algún concepto subyacente de interés para los investigadores. Por tanto, el desarrollo del cuestionario empieza definiendo los conceptos asociados con los objetivos de la investigación. Los conceptos se pueden descomponer en atributos específicos que pueden o no ser medibles en un sentido práctico. Las mediciones se especifican con más detalle y estas especificaciones se convierten en el material bruto para las preguntas de las encuestas. Las respuestas a las preguntas de las encuestas se convierten en datos que son resumidos o analizados para examinar los objetivos de investigación de una recogida de datos particular.

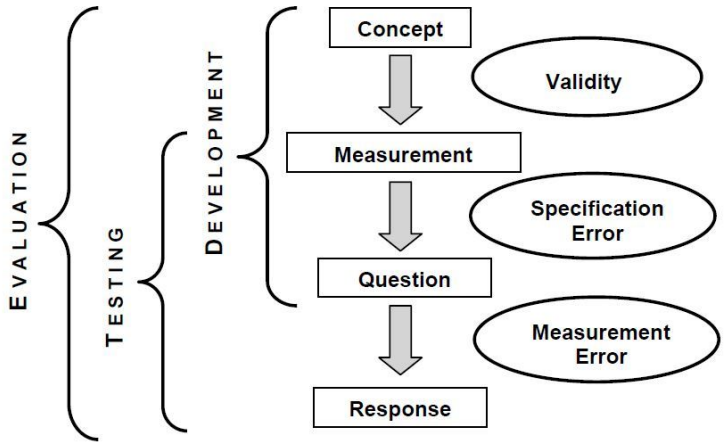

Figura 11.1: Pasos y fuentes de error en el desarrollo, testeo y evaluación de cuestionarios

Estos pasos, que muestran cómo pasamos de los conceptos a los datos vía encuestas, se muestran en la Figura 11.1, presentados como una ligera variación del lado de 'medida' del enfoque TSE. También en la Imagen 11.1 podemos ver los tipos de fuentes de errores ajenos al muestreo asociados a cada paso - validez, errores de especificación, y errores de medida. Algunas preguntas sobre la *validez* son, '¿Qué se puede medir para reflejar el concepto?' o '¿Cómo de bien la medida refleja el concepto?'. El *error de especificación* se asocia a la pregunta '¿Cómo de bien capta la pregunta la medida deseada?'. Finalmente, el *error de medida* se asocia a la pregunta '¿Cómo de bien se ajusta la respuesta al propósito de la pregunta?'.

Finalmente, la Figura 11.1 enlaza los pasos - empezando por conceptos, definiendo su medida, especificando las preguntas de la encuesta, y obteniendo las respuestas - a las fases de desarrollo, testeo y evaluación de cuestionarios en los cuales estos errores ajenos al muestreo pueden ser encontrados, investigados y examinados.

Definimos el *desarrollo* del cuestionario como las actividades y procesos en los cuales se definen los conceptos, se especifican los atributos medibles, y se formulan las preguntas del cuestionario para recoger estas medidas. Durante el desarrollo, nuestro objetivo es identificar y definir medidas válidas de los conceptos subyacentes de interés, y después especificar preguntas que son traducciones de las definiciones de medida. Por tanto, durante el desarrollo, intentamos asegurar la validez y minimizar los errores de especificación.

Definimos el *testeo* del cuestionario como la investigación realizada antes de la recogida de datos con el fin de determinar el grado con el cual las respuestas obtenidas a partir del cuestionario alcanzan el objetivo de la pregunta tal y como se especifica por las medidas deseadas. En otras palabras, el motivo de las pruebas del cuestionario es asegurar que las medidas se han operacionalizado de forma correcta en las preguntas del cuestionario

con el fin de identificar y abordar los problemas en las preguntas que contribuyen a los errores de medida en los datos, y, si es necesario, determinar si medidas pobremente especificadas dan lugar a preguntas problemáticas.

Definimos la *evaluación* del cuestionario como las investigaciones que conducen a evaluar la capacidad del cuestionario de medir el concepto reflejado en la calidad de los datos recogidos. La evaluación se realiza durante o después la recogida de datos, o al menos, implica los datos recogidos de los informantes reales, como los test de campo o experimentos realizados como parte del desarrollo y testeo.

El modelo de respuesta de la encuesta y el error de medida

La recogida de datos difiere entre las encuestas sociales y las económicas. Las encuestas económicas recogen en su mayoría datos numéricos, mientras que las sociales recogen información autobiográfica o de comportamiento de personas o atributos de hogares. Los informantes de encuestas sociales normalmente responden a preguntas sobre ellos mismos, sus hogares u otros individuos del hogar, y las preguntas a menudo (no siempre, por eso hay que tener cuidado con el sesgo de memoria en las encuestas sociales) se pueden responder de memoria. Las encuestas económicas precisan un informante que actúe en nombre de la unidad (empresa/establecimiento), a menudo proporcionando datos que se encuentra grabados en una base de datos.

Como consecuencia, el proceso de respuesta de una encuesta puede ser un poco más complejo en el caso de encuestas económicas. Los pasos del modelo de respuesta de una encuesta son los siguientes (algunos se aplican sólo a las encuestas económicas por lo explicado anteriormente) (D. Willimack y Nichols 2010; Sudman y col. 2000; Bavdaz 2010):

1. Registro de la información y codificación de la información en la memoria;
2. Organización de las tareas de respuesta:
    a Selección y/o identificación del informante(s);
    b Planificación del trabajo;
    c Establecer las prioridades y la motivación.
3. Comprensión de lo solicitado en la encuesta;
4. Recuperación de la información de la memoria y/o registros;
5. Evaluación de la adecuación de la información recuperada para alcanzar el propósito percibido de la pregunta;
6. Comunicación de la respuesta;
7. Revisión y envío de los datos a la organización.

Variaciones en la realización de estos pasos contribuye al error de medida. Para especificar mejor las fuentes del error de medida en el proceso de respuesta empecemos por el paso 1, *grabación de la información*, junto con el conocimiento de esos registros *codificados* en la memoria de la persona. En el caso de encuestas económicas, los datos grabados en bases de datos de empresas tienen un propósito no estadístico, y pueden estar distribuidos en distintos departamentos. Como consecuencia pueden haber potenciales errores de medida.

El relación con la *selección del informante*, en las encuestas sociales los proxys pueden ser fuentes de errores de medida. En las encuestas económicas influye no sólo esto, sino también el hecho de que esta actividad no sea prioritaria.

En relación con los cuatro pasos cognitivos - comprensión, recuperación, evaluación y comunicación - contribuyen a potenciales errores de medida no sólo las variaciones en las habilidades cognitivas del informante, sino también el contexto organizativo, tanto dentro de la empresa como en el hogar. Veamos algunos ejemplos:

- *Comprensión*. Los informantes pueden tener distintos grados de familiaridad con la terminología, y pueden interpretar las preguntas de forma distinta a las personas de su entorno (hogar/empresa).
- *Recuperación*. Grado en que los informantes tienen conocimiento o acceso a los registros, o su habilidad para obtener o extraer datos de fuentes alternativas. El error de medida puede aumentar.
- *Evaluación*. Cuando los datos disponibles no se ajustan a los datos solicitados, el error de medida se puede asociar con variaciones entre informantes (y sistemas de registro en caso de económicas) en relación con las estrategias usadas para estimar, aproximar o manipular los datos disponibles en un esfuerzo para proporcionar la respuesta.
- *Comunicación*. Los diseños y las instrucciones sobre cómo incluir las respuestas en los cuestionarios pueden no ser adecuadas para distintos tipos de datos, o poco claras, permitiendo que la misma información se pueda proporcionar de más de una forma posible, contribuyendo al error de medida.

El último paso en el modelo devuelve la respuesta al nivel organizativo, donde los *datos están disponibles* para la organización que realiza la encuesta. En este paso hay que tener en cuenta temas como la consistencia, confidencialidad, y seguridad. Variaciones en los criterios para la difusión de datos, y su impacto en los datos publicado, pueden contribuir al error de medida.

Restricciones y limitaciones en el desarrollo, testeo y evaluación de cuestionarios

La complejidad del proceso de respuesta, la naturaleza de las encuestas y el contexto en que se realizan interactúan de tal forma que tienen consecuencias en el diseño del cuestionario y en los métodos usados para el desarrollo, testeo y evaluación de

instrumentos de recogida de datos. D. K. Willimack y col. 2004 identificaron algunos de los más importantes temas que influyen:

- La *naturaleza de los datos solicitados* - medida de conceptos técnicos con definiciones precisas - tiene consecuencias en muchos aspectos de la recogida de datos, incluyendo el personal implicado en el diseño del cuestionario, la elección del (los) modo(s) de recogida de datos, diseño de cuestionarios, y la habilidad de proporcionar información por parte del informante. Los modos de recogida autocumplimentados son los que más se usan en las encuestas económicas, mientras que en las sociales se está produciendo un cambio de face to face a CATI y autocumplimentados. En los cuestionarios autoadministrados es necesario proporcionar instrucciones detalladas para igualar los distintos grados de conocimiento y motivación de los informantes, y para transmitir la información necesaria de forma consistente a todos los informantes.
- La *respuesta de la encuesta requiere mucho trabajo* y representa un coste tangible, pero no productivo para la organización. Más aún, esta naturaleza de requerir mucho trabajo puede ser acentuada por la necesidad de múltiples informantes y fuentes de datos. Como consecuencia, el completar un cuestionario largo o complejo en el caso de encuestas económicas puede llevar horas, días o incluso semanas, mientras el informante dedica el tiempo en identificar, contactar, solicitar y esperar por los datos de otros departamentos, y fuentes de datos de la empresa.
- *Carga de respuesta al informante* exhaustiva, porque las necesidades relacionadas con las características de la población objetivo y el uso de datos representa unas restricciones para la organización que realiza la encuesta en la implementación de los métodos de testeo y evaluación de cuestionarios. Además, involucra a informantes (en todo tipo de encuestas, pero en las económicas más aún) en el testeo y evaluación de cuestionarios añade carga sobre la carga que esas personas ya tienen al responder a las encuestas en las que ya están incluidos.

11.2 Desarrollo de contenido, medidas y cuestiones en encuestas[1]

Muchas encuestas económicas realizadas por los INEs proporcionan datos que pueden servir como input de contabilidad nacional, además de servir para describir la situación económica. En el caso de las encuestas sociales, además de servir para describir la situación social, sus microdatos pueden ser utilizados por los investigadores para realizar múltiples estudios. Independientemente de su uso, los datos necesitan empezar con algún tipo de concepto o constructo subyacente asociado con teorías o hipótesis que requieren medidas específicas. Así es como las cuestiones empiezan.

Dentro de una teoría, los conceptos subyacentes o postulados son descritos en una red de relaciones entre atributos que forman los componentes de los conceptos. Un atributo o indicador puede ser considerado como la pieza más pequeña de información medible que puede ser identificada. Un algoritmo específico asocia uno o más atributos como inputs para definir un concepto. De forma más específica, un concepto C_1, puede

[1] La mayor parte de esta sección se ha extraído de Snijkers y D.K. Willimack 2011

ser expresado como una función de atributos que van desde a n (A_1, ..., A_n), donde la definición funcional denota el mapeo:

$$C_1 = f(A_1, ..., A_n).$$

Los conceptos pueden ser sencillos, consistiendo de un único atributo, y por tanto, unidimensional. Sin embargo, muchos conceptos suelen ser multidimensionales, indicando que un concepto está compuesto de más de un atributo. Mientras que los conceptos subyacentes y sus atributos pueden parecer básicos y sin complicaciones, sus medidas puede que no sean tan sencillas. Consideremos, por ejemplo, el concepto de empleo, uno de los constructos más fundamentales en la descripción de la economía. Aunque parece muy sencillo, 'empleo' tiene una gran variedad de dimensiones: ¿El investigador está interesado en personas o en equivalentes a jornada completa (EJCs)? ¿Qué periodo de referencia debería de considerar el informante - un día en particular, una semana al mes, a final de mes? ¿Hay que incluir a los trabajadores a tiempo parcial? ¿Y a los temporales, subcontratados, contratistas? Y los asalariados? Teniendo en cuenta estos atributos, y siguiendo un algoritmo específico, se define el concepto 'empleo'.

Especificar los atributos es el primer paso hacia la vinculación de los conceptos en el proceso de recogida de datos. En el caso de una encuesta, las preguntas del cuestionario se diseñan para medir estos atributos. Las preguntas son la operacionalización de estos atributos; son las definiciones operativas de acuerdo con las cuales se miden los atributos. Esto incluye la redacción de una pregunta, las opciones de respuesta, y las instrucciones. Una pregunta puede basarse en un único atributo, y ser el resultado de una pregunta unidimensional; también puede basarse en una combinación de atributos, resultando una pregunta multidimensional.

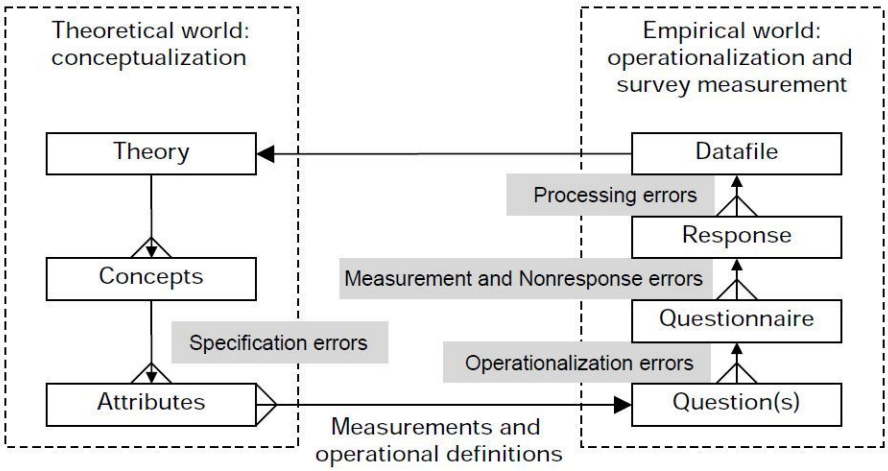

Figura 11.2: De la teoría al dato. Fuente: Snijkers y D.K. Willimack 2011

Los pasos para ir de la teoría al dato se muestra en la Figura 11.2. Ir de los conceptos a los atributos es lo que se llama *conceptualización* (o *especificación del concepto*); ir de los atributos a las preguntas y al cuestionario se llama *operacionalización*. Si la conceptualización y la operacionalización se hacen de forma correcta, obtenemos datos válidos en relación

con la validez de constructo (Groves y col. 2004) reflejando conceptos subyacentes. Una pobre conceptualización da lugar a errores de especificación (Biemer y Cantor 2007); del mismo modo, una pobre operacionalización da lugar a errores de operacionalización o de diseño (Snijkers 2002).

Para obtener datos válidos, la validez de constructo es un criterio importante y necesario. Sin embargo, un cuestionario válido en el sentido de que mide los conceptos subyacentes puede no ser válido en el sentido de que la operacionalización puede llevar a otro tipo de errores ajenos al muestreo (Snijkers 2002; Groves 1989), como falta de respuesta parcial o errores de medida. Para completar nos podemos encontrar con errores de procesamiento, que ocurren al codificar las respuestas (Groves y col. 2004).

Para conseguir un cuestionario que proporcione datos válidos es necesario realizar un proceso iterativo entre la conceptualización y la operacionalización. Esto se puede conseguir usando los métodos que veremos el la Sección 11.2. Pero antes veamos los roles de las distintas partes de este proceso.

Los roles de los expertos en la materia, las partes interesadas y los usuarios [2]

A menudo las distintas partes interesadas y los usuarios pueden jugar un papel importante en el desarrollo de los cuestionarios como proporcionar los fundamentos conceptuales o las necesidades de las encuestas, incluso pueden elaborar borradores de preguntas o colaborar en el diseño del cuestionario con los expertos. Se puede buscar input de asociaciones de empresarios, que indiquen la terminología correcta o informen sobre la disponibilidad de los datos solicitados, o de investigadores o expertos en la materia.

Métodos para el desarrollo del contenido y la especificación de medidas

Se pueden usar distintos métodos para evaluar las necesidades de datos, investigar conceptos, identificar medida y especificar preguntas para conseguir la validez del constructo durante el diseño. Hox 1997 distingue dos enfoques, el top-down, basado en la teoría, y el bottom-up, basado en los datos. El enfoque basado en la teoría está representado en el lado izquierdo de la Figura 11.2, empieza con constructos y avanza hacia variables observables; el enfoque basado en los datos es el del lado derecho, parte de las operacionalizaciones y las observaciones y avanza hacia los conceptos teóricos.

- **Top-down basados en la teoría**

[2] Los usuarios y partes interesadas engloban desde unidades dentro del propio INE como contabilidad nacional (usuario de la mayoría de las operaciones económicas), organismos internacionales como Eurostat (que a través de los Reglamentos establece desagregaciones, periodicidades o clasificaciones a utilizar), o otro tipo de organizaciones de usuarios.

Análisis de dimensión/atributo. Un indicador es similar a lo que llamamos un atributo. Atributos empíricos (o indicadores) se especifican para los conceptos en una teoría. El resultado es una red de conceptos que están vinculados de forma lógica, y de forma colectiva construyen la teoría. El proceso de especificación de conceptos termina cuando uno o más atributos se pueden identificar para todos los conceptos. Los atributos con la base para las preguntas de la encuesta.

Evaluación de las necesidades de los usuarios. Se pueden usar varios métodos que involucran a usuarios, investigadores, y otras partes interesadas.

- *Grupos de expertos/usuarios/asesores*. Los miembros de estos grupos priorizan nuevas preguntas y cambios en las existentes, y son los responsables de la aprobación final del cuestionario.
- *Seminarios con usuarios*. Se pueden usar para (1) identificar cómo y por qué los datos de interés serán usados, (2) identificar carencias en los datos existentes, (3) entender mejor las necesidades de datos para temas específicos, y (4) crear un conjunto preliminar de prioridades.
- *Rondas iterativas de consultas*. Los usuarios preparan una lista de datos solicitados, que son priorizados por los expertos en la materia. A continuación el resto de *stakeholders* realizan otra priorización. El resultado es un cuestionario que equilibra las necesidades de nuevos datos con los datos obtenibles.
- *Revisión de contenido a larga escala*. Se realizan revisiones de borradores de cuestionarios y se mandan a los usuarios, asociaciones de empresarios (en caso de encuestas económicas) y otras asociaciones.
- *Grupos exploratorios con personal de las encuestas y con usuarios*. Sirve para clarificar los conceptos a medir, las tablas de resultados, y las publicaciones, ayudando en la especificación de preguntas, definiciones y terminología.

- **Bottom-up basados en los datos**

Los siguientes métodos utilizan datos proporcionados por los informantes, o información recogida desde su perspectiva.

Minería de cuestionario. Se comienza con cuestionarios ya existentes y se revisan las preguntas con la ayuda de expertos en la materia. Se puede usar informes de estudios pretest, el análisis de paradatos, de falta de respuesta parcial o de los tipos y la frecuencia de los edits que no se verifican. El resultado es una selección de preguntas que pueden ser usadas en instrumentos de recogida de datos posteriores, en la misma operacionalización o en modificadas.

Grupos de discusión. Si es totalmente incierto cómo se pueden medir los conceptos, se pueden realizar entrevistas exhaustivas con un grupo pequeño representativo de la población objetivo, que se puede acompañar de grupos de discusión o de

entrevistas personales (Snijkers 2002). Un primer objetivo es entender los conceptos desde el punto de vista del informante. También se investiga la disponibilidad en la base de datos y la periodicidad en caso de encuestas económicas. Otros objetivos incluyen evaluar la carga al informante y la calidad.

Mapeado de conceptos. Este proceso implica 6 pasos. En el primero, se especifica el tema del grupo de discusión, y se seleccionan los participantes de la población objetivo. El segundo es una sesión de *brainstorming* con los participantes para generar frases que describan los aspectos más relevantes. En el tercero los participantes ordenan las frases relacionadas en pilas de acuerdo con su propia perspectiva. En el cuarto se analizan las matrices de similaridad producidas por las distintas pilas, el resultado es un mapa de conceptos. En el quinto los participantes discuten posibles significados y etiquetas del mapa. En el último se traducen las frases en preguntas.

Estudios de viabilidad. Se escoge un pequeño grupo de informantes para que proporcionen la información solicitada en el instrumento de recogida. Sirve para saber si los conceptos son medibles, la estructura de las preguntas y del cuestionario. En este método no es necesario tener el cuestionario finalizado.

Revisión por parte de los expertos. Los expertos pueden ayudar en la definición de los conceptos, evaluar las fuentes de datos esperadas, ayudar a asociar conceptos con medidas. El resultado son sugerencias para redacciones alternativas a las preguntas o especificaciones.

Estudios de recordkeeping. En el caso de las encuestas económicas se selecciona un pequeño grupo de unidades de la población objetivo y se analiza la disponibilidad de la información necesaria y cómo está estructurada en el sistema de gestión de la empresa.

11.3 Testeo de preguntas y cuestionarios

Las pruebas empiezan cuando hay un borrador de cuestionario elaborado por los metodólogos de encuestas y con el que los otros *stakeholders* están de acuerdo. Es importante una fase de desarrollo en la que se usen métodos descritos en la Sección 11.2. Veamos algunos métodos de pretesteo. Primero consideraremos cómo el conocimiento del proceso de respuesta puede guiar los planes y protocolos de pretesteo. Después, a medida que describimos varios métodos, destacaremos algunas modificaciones que se han hecho en la práctica en el caso de encuestas económicas.

El modelo de respuesta como marco para el testeo

El modelo presentado en la Sección 11.1 es una herramienta útil para planificar y realizar el testeo de los cuestionarios. El modelo no ofrece soluciones a los problemas, sino que ofrece un marco para investigar las actividades y comportamientos de los informantes que pueden contribuir a errores de medida. Evaluar si las soluciones funcionan requiere métodos complementarios de evaluación como los descritos en la Sección 11.4.

Existen diferencias entre las encuestas sociales, que pueden desarrollarse en un medio controlado como es un laboratorio que pueda simular el entorno en el que se realizarían las encuestas, y las encuestas económicas, las que las entrevistas cognitivas muchas veces no se pueden realizar en un laboratorio (no resultaría sencillo simular el entorno habitual de una empresa), ni tampoco en la empresa (no resulta sencillo que mientras el resto del personal de la empresa realiza sus labores habituales una persona realice un entrevista cognitiva al informante de la encuesta).

Los protocolos diseñados en torno al modelo de respuesta a menudo comienzan con cuestiones globales, determinando el criterio de selección de los informantes, y otras tareas de las que el informante es responsable. Los protocolos usados para conducir entrevistas pretest en campo son:

1. *Presentación de la visita*, que creen las expectativas de los informantes con motivo de la visita y los procedimientos a utilizar.
2. *Preguntas generales sobre el informante y el proceso de respuesta*, para captar la atención del informante permitiendo también a los investigadores crear un ambiente familiar y obtener información para luego realizar preguntas más específicas.
3. *Observación o reconstrucción del proceso de respuesta*, para examinar asuntos cognitivos, procedimentales, y de carga asociados con las preguntas del cuestionario a testear.
4. *Preguntas de evaluación*, para obtener detalles adicionales que evalúen la calidad de las respuestas a las preguntas del cuestionario, y para conocer la opinión del informante sobre la encuesta y sugerencias de mejoras.
5. *Corrección de datos y respuestas a las preguntas del informante*, aprovechando el encuentro cara a cara para proporcionar *feedback* constructivo e instrucciones al informante para mejorar la calidad de futuros datos, y también creando una relación entre la oficina estadística y los informantes.

En el caso del proceso de respuesta de encuestas económicas Bavdaz 2010 ha desarrollado el modelo que se puede ver en la Figura 11.3.

Las definiciones asociadas a este modelo son las siguientes:

(a) Un dato es *accesible* - la respuesta necesaria puede estar disponible.

(b) Un dato es *generable* - la respuesta necesaria no está disponible pero con los datos disponibles se puede generar.

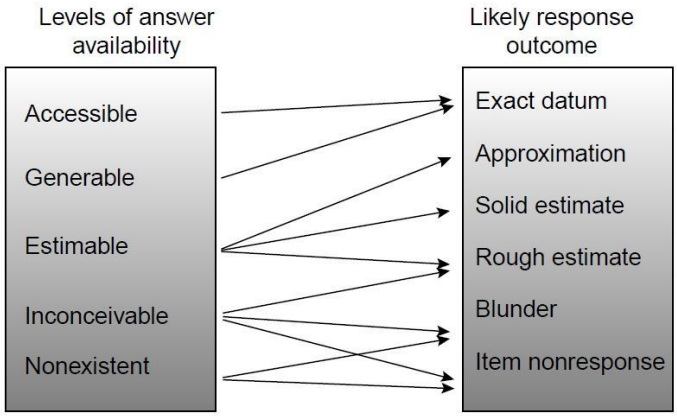

Figura 11.3: Niveles de disponibilidad de las respuestas y posibles resultados de las respuestas. Fuente: Bavdaz 2010

(c) Un dato es *estimable* - la respuesta necesaria no está disponible y con los datos disponibles se puede obtener una estimación.

(d) Un dato es *inconcebible* - no hay datos disponibles para obtener o estimar la respuesta necesaria, el esfuerzo necesario para la estimación es inimaginable

(e) Un dato es *inexistente* - no se puede estimar la respuesta necesaria.

Cualquier resultado distinto de 'dato exacto' implica error de medida.

Métodos

Veamos a continuación algunos métodos para testear preguntas y cuestionarios extraídos del D. K. Willimack y col. 2004.

Revisiones de expertos metodólogos. En estas revisiones, los expertos en diseño de cuestionarios evalúan el cuestionario, aplicando principios generales de diseño junto con otras experiencias de pretesteo. Así se identifican potenciales problemas y se sugieren soluciones.

Estudios exploratorios/de viabilidad y visitas. En el caso de encuestas económicas en las que los datos están disponibles en bases de datos se pueden realizar visitas *in-situ*. Los tamaños suelen ser pequeños y se pueden realizar también telefónicamente para reducir el coste.

Grupos de discusión con informantes. Los grupos de discusión se pueden usar en los procesos de desarrollo del cuestionario, testeo y evaluación. Se basan en discusiones entre un pequeño grupo de personas (8-12 miembros de la población objetivo) sobre un (conjunto de) tema(s). Se centra en evaluar la claridad de la redacción de las preguntas,

el uso y la utilidad de las instrucciones, reacciones generales al cuestionario y a su diseño.

Métodos de pretesteo cognitivos. Estos métodos incluyen entrevistas en profundidad, pensar en voz alta y/o técnicas de investigación, reuniones y *vignettes*. Los métodos tradicionales cognitivos se han adaptado para reducir el trabajo de los informantes.

Vignettes. Escenarios hipotéticos, conocidos como *vignettes*, permiten a los investigadores evaluar los errores de respuesta, ya que se conocen las respuestas asociadas a las viñetas.

Métodos etnográficos. Los métodos etnográficos usados en el pretesting consisten en *encuestas en profundidad sin estructura*, y, en menor medida, la *observación*. La investigación etnográfica emplea dos niveles de análisis:

- *Significados/conocimiento* - las interpretaciones de los informantes a las preguntas, su conocimiento sobre la materia y sobre aspectos relevantes para la encuesta.
- *Proceso* - las interacciones de los informantes con los instrumentos de la encuesta, sistemas de información (en el caso de encuestas económicas), etc.

Las *entrevistas en profundidad* facilitan la discusión de las interpretaciones de las preguntas de la encuesta de los informantes y cómo las responden. La *observación* permite a los investigadores recoger información sobre el proceso que explica la interacción de los informantes con los instrumentos de recogida de información, las relaciones con los demás miembros de la familia/hogar/empresa o la apariencia y estructura de la interface con los sistemas de información (en caso de encuestas económicas).

Métodos para las instrucciones de testeo. Se puede estudiar el uso de las instrucciones por parte de los informantes, dónde están colocadas en el cuestionario y su formato. Uno de los métodos para probar la comprensión de las instrucciones consiste en pedirle al informante que explique a otra persona, de su hogar o de su empresa, cómo tendrían que contestar las preguntas.

Estudios piloto. Los estudios piloto formales permiten la evaluación de datos proporcionados por informantes a los cuestionarios, y son particularmente importantes de cara a la realización de nuevas encuestas o para grandes rediseños. Se usa el término *piloto* para referirnos a pruebas de campo con informantes en los que la recogida de datos simula el proceso de recogida. Otra característica de los estudios piloto, independientemente del tamaño muestral o del alcance, es que los datos recogido no se usan para estimaciones oficiales, aunque se pueden calcular estimaciones para evaluar su calidad. Los tamaños muestrales y los diseños de estudios piloto varían mucho. Los estudios piloto pueden ser iterativos o involucrar múltiples paneles, que en el caso de encuestas económicas pueden dirigirse a empresas de distintos tamaños y actividades económicas. Los estudios piloto a menudo se aprovechan de otros métodos de evaluación.

11.4 Evaluación de preguntas y cuestionarios

Métodos usados durante la recogida de datos

Codificación de conducta. Codificación de conducta (también conocida por *análisis de interacción*) se centra en comportamientos observables de los entrevistadores e informantes mientras interactúan durante la cumplimentación del cuestionario. Las entrevistas realizadas en persona o por teléfono son directamente observables, o se pueden grabar y revisar a posteriori. Puede incluir descripciones del grado en el cual los entrevistadores leen o no las preguntas tal y como están redactadas o, en el caso de los informantes, incluye acciones como interrumpir, preguntar por aclaraciones, proporcionar una respuesta adecuada o inadecuada o rechazar responder. Estos resultados permiten localizar preguntas problemáticas (véase Groves y col. 2004). En el caso de cuestionarios autocompletados incluye el análisis del cuestionario en busca de patrones o notas para preguntas específicas.

Experimentos con parte de la muestra. Experimentos con parte de la muestra (también conocidos por *paneles divididos*) permiten la evaluación de preguntas, cuestionarios o diseños alternativos. También se han utilizado para evaluar la posición de las instrucciones. En este tipo de experimentos partes de la muestra son asignados de forma aleatoria a uno de dos o más grupos en los que algunos atributos del diseño han sido modificados. Uno de estos grupos, el *grupo control*, mantiene las condiciones/diseño. Se analizan los resultados del resto de grupos para evaluar el efecto de las variaciones (Keppel y Wickens 2013).

Métodos usados después de la recogida de datos

Evaluaciones empíricas post recogida. Actividades de evaluación está asociadas con encuestas periódicas. Se analiza la falta de respuesta parcial, las tasas de imputación, los outliers y los registros que presentan problemas en la depuración. Esto puede llevar a la eliminación de preguntas o a redefinir las categorías de respuesta.

Reuniones con informantes y encuestas de análisis de respuesta. En las reuniones con informantes (o segundas entrevistas) se recontacta con informantes y se les pregunta sobre las estrategias de respuesta y las fuentes de datos que han usado para proporcionar las respuestas de preguntas específicas. De esta forma se puede evaluar si los datos recogidos cumplen los propósitos de las preguntas. Las reuniones formales con informantes, también conocidas como encuestas de análisis de respuesta (RASs del inglés *response analysis surveys*) se realizan después de la recogida de datos usando cuestionarios estructurados. Aunque lo normal es que se usen en encuestas pilotos o encuestas rediseñadas, también se pueden usar para evaluar la calidad de los datos de encuestas periódicas (Goldenberg 1994). El tamaño de las muestras puede variar desde 20 personas a varios cientos.

Estudios de reentrevistas y evaluaciones de contenido. Los estudios de reentrevistas es una variación de las reuniones con informantes en las que el primer objetivo es estimar el sesgo en estadísticas de síntesis debido a los errores al proporcionar la información, por tanto, los tamaños de muestra y las metodologías de la recogida de datos deberían de ser suficientes para este propósito. Las evaluaciones de contenido son estudios de reentrevistas que examinan las componentes de los datos proporcionados así como sus fuentes y su fiabilidad. Se suelen utilizar para cambiar cuestionarios e instrucciones.

Estudios de validación. Los errores de medida se pueden evaluar directamente si existe un 'valor real' conocido que se pueda comparar con las respuestas de los informantes. Como fuentes para estos valores se pueden usar los registros administrativos (Groves y col. 2004). Este tipo de estudios se pueden usar para validar los datos. Se puede utilizar sobre todo en encuestas económicas.

Métodos para identificar problemas en los cuestionarios de encuestas periódicas

Muchas encuestas, en particular las realizadas por los INEs, se repiten con cierta periodicidad para generar series temporales (por ejemplo, mensualmente, trimestralmente, anualmente) ofreciendo la posibilidad de evaluaciones posteriores de respuestas pasadas que pueden ayudar a identificar preguntas o diseños problemáticos. Cualquiera de los métodos descritos anteriormente son útiles para este objetivo. Veamos algunos más para encuestas periódicas.

Feedback del personal encuestador. A menudo se puede usar la experiencia del personal de recogida que tienen contacto con los informantes. El personal de recogida puede transmitir el *feedback* de recogidas previas, y pueden actuar como *proxis* de los informantes. A pesar de su falta de rigor, estas evaluaciones informales son útiles para sugerir áreas que precisan más investigación. Algunos métodos de *feedback* incluyen:

- *Reuniones con entrevistadores*. Se puede utilizar para identificar preguntas problemáticas y asegurar su correcta comprensión.
- *Preguntas, comentarios y quejas de los informantes*. El *feedback* de las interacciones del personal de recogida con los informantes se puede usar de forma informal para evaluar los cuestionarios. Se pueden usar las preguntas de los informantes, y las llamadas sobre los edits que han fallado.
- *Entrevistas o grupos de discusión con personal*. Se puede realizar con personal que revisa los datos y que resuelve dudas en la depuración (Tuttle, Morrison y D. Willimack 2010).

Examinar los cuestionarios completos. Examinar los cuestionarios completos obtenidos durante la recogida a menudo proporciona pistas que pueden ayudar a los metodólogos a identificar las preguntas o los diseños que pueden ser problemáticos para los informantes. En el caso de encuestas económicas los informantes pueden indicar apartados que fueron difíciles de responder o en los que se tuvieron que hacer estimaciones. Apartados en los que se proporcionó una respuesta y posteriormente se cambió puede

indicar problemas con preguntas, instrucciones o formato. Los informantes pueden escribir observaciones o aclaraciones donde se puedan escribir dentro del cuestionario, o especificar respuestas alternativas cuando se pueda elegir 'Otros'.

Indicadores estadísticos de procesos. Varios indicadores se pueden usar para evaluar la calidad de los datos recogidos y para evaluar el resultado de preguntas individuales. Algunos de estos indicadores son la tasa de falta de respuesta total y parcial, índices de verosimilitud, y tasas de incumplimiento de edits, pero hay que tener cuidado con su interpretación:

- La *tasa de falta de respuesta total* puede ayudar a identificar las características de la población muestral que suelen tener problemas al responder.
- La *tasa de falta de respuesta parcial* es un poco más difícil de interpretar ya que puede significar 'no aplicable', 'aplicable pero cero', 'reuso a contestar' o 'no lo sé'.
- Un *índice de verosimilitud* se define como un índice de desviación de valores esperados que son calculados de datos administrativos, datos de años anteriores o datos de fuentes comparables. Hay que interpretarlo con cuidado porque su variabilidad se puede deber a las condiciones económicas.
- Las *tasas de incumplimiento de edits* puede ayudar a identificar preguntas problemáticas. Tasas altas en algún edit puede indicar qué apartados pueden ser difíciles de contestar. Hay que tener cuidado en su interpretación porque algunos edits pueden ser el resultados de varias preguntas, lo que dificulta aislar la fuente de error (Tuttle, Morrison y D. Willimack 2010).

11.5 Desarrollo, testeo y evaluación de instrumentos de recogida electrónica de datos

En este tema nos centraremos en el desarrollo, testeo y evaluación de instrumentos de recogida electrónica de datos, tanto a través de la web como a través de hojas de cálculo o el intercambio electrónico de datos [3].

Estrategias de desarrollo para instrumentos electrónicos

Las estrategias de desarrollo para instrumentos electrónicos siguen algunos de los procedimientos asociados con el desarrollo de aplicaciones de software. Veamos cuatro de ellos.

Diseño centrado en el usuario. Es importante usar el enfoque del diseño centrado en el usuario cuando se desarrollen los instrumentos electrónicos, prestando atención a las necesidades de los informantes. Este enfoque se centra en cómo los informantes quieren usar los instrumentos de encuesta, más que en que los informantes aprendan las

[3] EDI del inglés *Electronic Data Interchange* que incluye ficheros XML o XBRL)

características de los instrumentos. Los investigadores abogan por este tipo de enfoque asegurando que el instrumento es intuitivo y fácil de usar, con el fin de reducir una potencial carga al informante y, con suerte, aumentar la tasa de respuesta. Se deberían de realizar test de usabilidad para asegurar que se ha alcanzado un diseño centrado en el usuario.

Recopilación de necesidades/Análisis de tareas.

- *Recopilación de necesidades*. Los objetivos de esta fase ocurre durante la fase de desarrollo. Las necesidades son especificaciones para todos los aspectos del sistema y vienen de los objetivos y propósitos del sistema, las necesidades del informante, y un análisis funcional del sistema en el que se encuentra la encuesta electrónica. En este paso los miembros del equipo de diseño obtienen ideas de distintas fuentes, como los usuarios, las quejas o las peticiones de los informantes. Las necesidades deberían de estar documentadas, ser medibles, testeables, trazables y relacionadas con las necesidades de los usuarios o de los informantes.
- *Análisis de tareas*. Es el proceso de identificar las tareas de los informantes e identificar cómo pueden ser satisfechas usando un cuestionario electrónico. Es importante conocer qué necesitan los informantes para responder el cuestionario y qué acciones tiene que realizar en la encuesta electrónica para conseguirlo. Además es importante entender cómo los informantes evalúan la calidad del proceso de respuesta electrónica. Aunque el informante pueda tener un resultado exitoso, es importante asegurar que el proceso de conseguir ese resultado es satisfactorio. Este análisis de tareas debería ser realizado tan temprano como sea posible en el proceso de desarrollo para que pueda proporcionar un input en el diseño del sistema. Se deberían de identificar las necesidades de los informantes de forma que sea posible una transición fácil entre un entorno basado en el papel y uno electrónico. La incorporación de métodos como la recopilación de necesidades o el análisis de tareas en el proceso de desarrollo de encuestas electrónicas ayuda a asegurar que la organización desarrolla un instrumento que conoce las expectativas de los informantes sobre cómo debería funcionar una encuesta electrónica. Cuando los informantes se enfrentan a un instrumento que funciona de una forma distinta a la esperada los informantes pueden sentirse defraudados y minimizar el uso del instrumento.

Evaluaciones heurísticas de los instrumentos electrónicos, guías de estilo y estándares. Los instrumentos electrónicos también están sujetos a revisiones de expertos, que se llaman *evaluaciones heurísticas* porque normalmente son evaluadas en relación con los principios de interacción humano-ordenador o heurísticos. Las revisiones usando esquemas de codificación heurísticos pueden ser realizadas por expertos en usabilidad, paneles de expertos o miembros del equipo de desarrollo. La carga adicional de probar instrumentos electrónicos con informantes puede ser minimizada adhiriéndose a una guía de estilo o unos estándares para la interface de usuario cuando se desarrollen los instrumentos electrónicos. Estas reglas de navegación, formato, diseño de pantalla, gráficos, ayuda, tratamiento de errores, *feedback* con el usuario, y otros mecanismos electrónicos se basan en estándares de usabilidad y en buenas prácticas, así como en pruebas anteriores y

experiencias operativas con los instrumentos electrónicos.

Tests funcionales y de ejecución. Hacia el final de la fase de desarrollo de encuestas electrónicas, los programadores prueban el sistema de varias formas para asegurar que todas las componentes necesarias funcionen correctamente. Durante los tests, los programadores trabajan con el sistema para resolver los problemas relacionados con el diseño, la velocidad y la capacidad de respuesta de la encuesta electrónica. Para hacer esto, pueden utilizar varios escenarios. Además, los programadores realizan una *prueba de carga*, en la cual simulan la demanda esperada en un sistema para determinar cómo se comporta el sistema bajo condiciones normales o de máxima demanda. De esta forma pueden identificar la capacidad operativa máxima del sistema y anticipar cuellos de botella.

Pruebas de usabilidad

Los cuestionario autocumplimentados a través de internet añaden otra capa a la interacción entre informantes y el cuestionario, el de la interface del usuario. *Pruebas de usabilidad* es el nombre que se le da las metodologías para probar las interfaces gráficas, tales como la navegación o los edits integrados. Algunos ejemplos son:

- *Uso de prototipos*. Puesto que el plazo para el desarrollo de instrumentos electrónicos debe incluir la programación, los tests pueden empezar con prototipos en papel o capturas de pantalla que puedan ayudar a los desarrolladores. Las pruebas de usabilidad se realizan normalmente con prototipos que están funcionando total o parcialmente. Las entrevistas con los usuarios pueden realizarse usando instrumentos existentes, para detectar problemas de usabilidad y poder realizar el rediseño.
- *Observación*. Las pruebas de usabilidad dependen en la observación directa de los informantes. Los investigadores asignan tareas halladas al usar el instrumento, y después observan a los informantes cuando llevan a cabo acciones como proporcionar los datos, corregir los errores, o enviar el cuestionario completado. Los comportamientos de los informantes al interaccionar con el instrumento también pueden ser codificado, para servir de ayuda en el análisis empírico de problemas de usabilidad. Estos tests se podrían grabar con el permiso de los informantes.
- *Métodos cognitivos*. Los investigadores pueden adaptar e incorporar métodos cognitivos en los tests de usabilidad, como pruebas simultáneas, entrevistas retrospectivas y valoraciones de los usuarios y los informantes. Se han usado viñetas u otros hipotéticos escenarios para probar opciones de diseño alternativas. Después de interactuar con el instrumento, se pide a los usuarios que rellenen cuestionarios o que participen en grupos que proporcionen el *feedback* sobre sus experiencias.
- *Rastreo ocular*. Los investigadores han estado usando cada vez más equipos con reastreo ocular para completar la información aprendida durante las pruebas tradicionales de trazabilidad. El rastreo ocular consiste es seguir el recorrido de lo que una persona mira. Este tipo de equipamiento está incluido en el monitor del ordenador, y el software de rastreo ocular capta la ruta de lo que mira el

informante en la pantalla. El resultado muestra el recorrido y la duración del tiempo empleado en mirar los objetos de la pantalla. Estas pruebas se tienen que llevar a cabo en los laboratorios donde se hacen las pruebas de usabilidad.

Evaluación de instrumentos de recogida electrónica de datos

El uso de los instrumentos de recogida electrónica también se puede evaluar empíricamente una vez que el instrumento está en campo. Las decisiones de diseño que se tomen tienen que tener en cuenta el modo seleccionado.

Paradatos. Para encuestas online, los *paradatos* se refieren a los datos de proceso que se han creado como un subproducto de la recogida de datos con este tipo de encuestas (Couper 2008). Estos datos pueden reunirse y analizarse y proporcionar información sobre el proceso por el que pasan los informantes cuando rellenan un cuestionario online. Los paradatos describen cómo los informantes completan la encuesta frente al contenido de sus respuestas.

Recoger paradatos que consisten en la información sobre el número de visitas a la página web donde se encuentran los cuestionarios hace posible controlar el progreso de los informantes en la cumplimentación de la encuesta, en relación con cómo iniciaron la encuesta, cuántos la han completado y en qué punto los informantes interrumpen la cumplimentación. También es posible examinar el comportamiento del informante al completar la encuesta; por ejemplo, si los informantes tienden a entrar y completarlo de una vez, o si vuelven varias veces antes de enviarlo. Examinar el punto en el que los informantes tienden a abandonar la cumplimentación puede ayudar a identificar preguntas problemáticas.

Los paradatos recogidos sobre el informante hacen posible identificar el proceso de respuesta de los informantes en su medio natural. Cada acción 'significativa' es registrada, de forma que puede ser identificada y situada en el tiempo. Puesto que uno de los mayores desafíos al analizar los paradatos es manejar ficheros de datos muy grandes y extraer la información útil de ellos, es aconsejable decidir qué acciones tienen significado dependiendo de los asuntos de interés. Algunas posibilidades son:

- Clicar un botón de selección;
- Clicar y seleccionar una opción de respuesta en un menú desplegable;
- Clicar en una casilla (tanto para seleccionar como para deseleccionar);
- Escribir texto en una casilla habilitada para ello;
- Clicar en un *hyperlink*;
- Enviar la página.

Algunos ejemplos de paradatos recogidos y analizados incluyen:

1. Tiempos de respuesta - el tiempo que pasa entre que se carga la pantalla en el ordenador del informante y la respuesta se envía.
2. Si, y en qué sentido, el informante cambia su respuesta.
3. El orden en que las preguntas son respondidas.

Analizando estos paradatos se pueden mejorar los instrumentos de recogida electrónica de datos.

11.6 Análisis de datos cualitativos

En todo tipo de encuestas se usan métodos cualitativos de investigación. La investigación cualitativa se caracteriza típicamente por diseños muestrales deliberados con un número pequeño de casos. Hemos visto que los métodos cualitativos varían y pueden incluir contacto personal entre investigadores y sujetos, observaciones abiertas o encubiertas, y análisis de textos y documentos. Aunque es inapropiado hacer inferencias estadísticas generalizadas a una población objetivo, los datos resultantes son ricos en detalles, proporcionando contexto subyacente y una visión de la información empírica. Veamos ahora métodos de análisis, resumen e informes de resultados de métodos cualitativos.

Los datos cualitativos recogidos durante el desarrollo y las pruebas de los instrumentos de encuestas consistirán en conversaciones, observaciones y documentos. Los datos cualitativos proporcionan información sobre el 'qué' y el 'por qué', y están caracterizados por proporcionar información rica y detallada. Como en cualquier otra recogida de datos y análisis de proceso, pueden ocurrir errores , incluyendo el sesgo del investigador. Todas estas cuestiones han dado lugar a la necesidad de métodos sistemáticos para el análisis cualitativo (Miles y Huberman 1994).

El análisis de datos cualitativos no es una tarea sencilla, ya que generalmente se recoge una cantidad enorme de datos cualitativos. El volumen de datos puede resultar abrumador, dejando al investigador no seguro de por dónde empezar. La tarea es convertir la cantidad de datos en un resumen y asegurar que éste es un fiel reflejo de los datos.

Captura de datos

En primer lugar, los datos necesitan ser capturados de los medios en los que se encuentran. Las entrevistas grabadas en audio o en vídeo tiene que ser resumidas o transcritas. Si este trabajo es realizado por personal que no pertenece al instituto de estadística, entonces debe firmar un documento que asegure la confidencialidad. Además es necesario usar métodos que aseguren la calidad con el fin de que los datos sean acurados.

Tipos de análisis

Distintos tipos de técnicas de análisis son necesarias para distintos tipos de investigaciones cualitativas.

Los datos recogidos de entrevistas, grupos de debate, u observaciones son típicamente analizados usando *análisis de contenidos*. Este enfoque consiste en identificar temas a partir de los datos, o categorizar la información de acuerdo con temas previamente identificados. Algunos de los pasos son:

- Añadir códigos a un conjunto de notas de campo extraídas de observaciones o entrevistas;
- Anotar reflexiones u otros comentarios en los márgenes;
- Ordenar o cribar este material para identificar frases similares, relaciones entre variables, patrones, temas, etc.;
- Aislar estos patrones y procesos, coincidencias y diferencias, y llevarlas a campo en la siguiente ola de recogida de datos;
- Elaborar gradualmente un pequeño conjunto de generalizaciones que cubra las consistencias detectadas en los datos de base;
- Confrontar estas generalizaciones con un conjunto de información formalizado en forma de constructos o teorías.

La mayoría de los análisis de datos cualitativos tendrán tres flujos simultáneos de actividad: reducción de datos, visualización de datos, y extracción de conclusiones/verificación.

Reducción de datos

Una dificultad del análisis de datos cualitativos es tener una posición neutral, sin prejuicios, con el fin de asegurar la validez y la fiabilidad. Para conseguirlo los estadísticos se debe sumergir en los datos, con el objetivo de identificar los temas clave que surgen de los datos.

Durante la reducción, los detalles contextuales también se deben de tener en cuenta. Por ejemplo, si el informante tuvo acceso a los datos requeridos en el cuestionario o las condiciones en la oficina durante la realización de la entrevista cognitiva si ésta se realizó en la oficina.

El proceso de reducción de datos también depende del objetivo de la investigación y de los métodos usados para recoger los datos.

Display de datos

A continuación de la reducción de datos, hay que examinar los temas para identificar las asociaciones existentes entre ellos. Por ejemplo, quizá muchos informantes que no pudieron proporcionar los datos de una determinada pregunta son empresas pequeñas u hogares con determinadas características. Esto relaciona el tema y el contexto. Esta

actividad se puede hacer de forma visual, usando diagramas de flujo. Sin este paso los resultados pueden consistir sólo en temas y subtemas, pero sin las conexiones entre ellos.

Extracción de conclusiones y verificación

Una vez que se ha realizado el análisis, las conclusiones se pueden alcanzar centrándose en los patrones comunes percibidos a partir de los datos. A medida que se alcancen las conclusiones a menudo será necesario verificarlas volviendo a los datos brutos.

11.7 Enfoques multimétodo para el desarrollo, testeo y evaluación

Puesto que muchos métodos de pretesting son cualitativos, usar más de método para desarrollar, pretestear, y/o evaluar un cuestionario puede incrementar la confianza en los resultados de los estadísticos. Los distintos métodos tienen sus fortalezas y deficiencias, y suelen ser complementarios. Algunos métodos ayudan a revelar o entender la naturaleza de problemas con las preguntas o los cuestionarios, mientras que otros métodos sugieren soluciones, mientras que otros métodos comparan las alternativas o evalúan la efectividad. Véase (Bavdaz 2009) para más información.

11.8 Organización y logística

Estructuras organizativas de la encuesta para el desarrollo, testeo y evaluación de instrumentos

En muchos INEs el personal está organizado por encuesta o por dominio, por eso adquieren experiencia en el tema tratado en la encuesta. En la mayoría de los casos obtienen su experiencia en fases posteriores para sugerir modificaciones en los cuestionarios. Y los expertos en los cuestionarios trabajan fuera de estas áreas de dominios, aunque algunas organizaciones asignan especialistas en el diseño de encuestas a expertos en temas. Es importante que el experto en el diseño del cuestionario forme parte del equipo que desarrolla una encuesta desde el principio. El uso de su conocimiento sobre las perspectivas de los informantes y buenas prácticas en el diseño de cuestionarios puede contribuir al proceso de conceptualización y de operacionalización, junto con el resto de interesados, expertos en la materia, estadísticos a cargo de la encuesta, ayudando a resolver problemas comunes y a evitar obstáculos.

Logística de prestests

Procedimientos para entrevistas pretests, incluyendo entrevistas cognitivas, tienen diferentes características dependiendo de si las encuestas son sociales o económicas. En el caso de las sociales se suelen desarrollar en un laboratorio, mientras que en el caso de las económicas pueden realizarse directamente en la empresa, para que los informantes tengan acceso a sus datos. Esto implica un mayor coste en tiempo y dinero, pero a cambio el contexto es más realista. Debido a las limitaciones de recursos, alguna vez

los pretest cognitivos se han desarrollado mediante entrevistas telefónicas o con un cuestionario auto-administrado. Sin embargo, las comparaciones entre estos métodos abreviados y las entrevistas cognitivas tradicionales face-to-face han mostrado que las últimas proporcionan información más completa y detallada.

En relación con la duración, también se observan diferencias entre las encuestas sociales, más largas, y las económicas, 60-90 minutos. Por este motivo en el caso de las encuestas sociales se suele usar todo el cuestionario, mientras que en las económicas se reduce a un conjuntos de preguntas.

Los tamaños de las muestras varían mucho debido a las condiciones de las organizaciones. Los estudios pueden usar grupos de tan sólo 2-5 personas hasta grupos de 60, dependiendo de los objetivos del estudio y de los recursos.

Aunque no es inusual realizar sólo una ronda de entrevistas con un pequeño número de casos en los estudios cognitivos, es preferible rondas múltiples o fases en las entrevistas pretest, en los que el cuestionario es rediseñado entre fases. Habitualmente, las entrevistas cognitivas pretest son realizadas por expertos en el desarrollo de cuestionarios y por metodólogos.

Bibliografía

Bavdaz, M. (2009). "Conducting research on the response process in business surveys". En: *Statistical Journal of the International Association of Official Statistics* 26, págs. 1-14.

– (2010). "The multidimensional integral business survey response model". En: *Survey Methodology* 36, págs. 81-93.

Biemer, P.P. y D. Cantor (2007). "Introduction to survey methods for businesses and organizations". En: *short course presented at the 3rd International Conference on Establishment Surveys (ICES-III), Montreal, June 18, American Statistical Association, Alexandria, VA.*

Couper, M.P. (2008). *Designing Effective Web Surveys*. Cambridge: Cambridge University Press.

Goldenberg, K.L. (1994). "Answering questions, questioning answers: Evaluating data quality in an establishment survey". En: *Proceeding of the Section on Survey Research, America Statistical Association*, págs. 1357-1362.

Groves, R.M. (1989). *Survey errors and survey costs*. New York: Wiley.

Groves, R.M., F.J. Fowler Jr., M.P. Couper, J.M. Lepkowski, E. Singer y R. Tourangeau (2004). *Survey Methodology*. New York: Wiley.

Hox, J. (1997). *From theoretical concepts to survey questions, in Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D. (Ed.), Survey Measurement and Process Quality*. New York: Wiley, págs. 47-69.

Keppel, G. y T.D. Wickens (2013). *Design and Analysis: A Researcher's Handbook*. 5th. Upper Saddle River, NJ: Pearson Prentice Hall.

Miles, M.B. y A.M. Huberman (1994). *Qualitative Data Analysis: An Expanded Sourcebook*. Sage, Thousand Oaks.

Snijkers, G. (2002). "Cognitive Laboratory Experiences: On Pre-Testing Computerized Questionnaires and Data Quality". En: *PhD thesis Utrecht University, Statistics Netherlands, Heerlen.*

Snijkers, G., G. Haraldsen, J. Jones y D.K. Willimack (2013). *Designing and Conducting Business Surveys*. New York: Wiley.

Snijkers, G. y D.K. Willimack (2011). "The missing link: From concepts to questions in economic surveys". En: *paper presented at the 2nd European Establishment Statistics Workshop (EESW11)*, págs. 12-14.

Sudman, S., D. Willimack, E. Nichols y T. L. Mesenbourg (2000). "Exploratory reasearch at the U.S. Census Bureau on the survey response process in large companies". En: *Proceedings of the 2nd International Conference on Establishment Surveys (ICES-II)* 26, págs. 327-337.

Tuttle, A.D., R.L. Morrison y D. Willimack (2010). "From start to pilot: A multimethod approach to the comprehensive redesign of an economic survey questionnaire". En: *Journal of Official Statistics* 26, págs. 87-103.

Willimack, D. y E. Nichols (2010). "A hybrid response process model for business surveys". En: *Journal of Official Statistics* 26, págs. 3-24.

Willimack, D. K., L. Lyberg, J. Martin, L. Japec y P. Whitridge (2004). *Evolution and adaptation of questionnaire development, evaluation, and testing methods for establishment surveys. In: Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., and Singer, E. (Eds.), Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: Wiley, págs. 385-407.