

Capítulo 6

Diseño multietápico

Índice

1. Principio y notaciones
2. Diseño en conglomerados
3. El efecto conglomerados
4. Diseño en dos etapas
5. Nota sobre el diseño autoponderado
6. Consideraciones prácticas

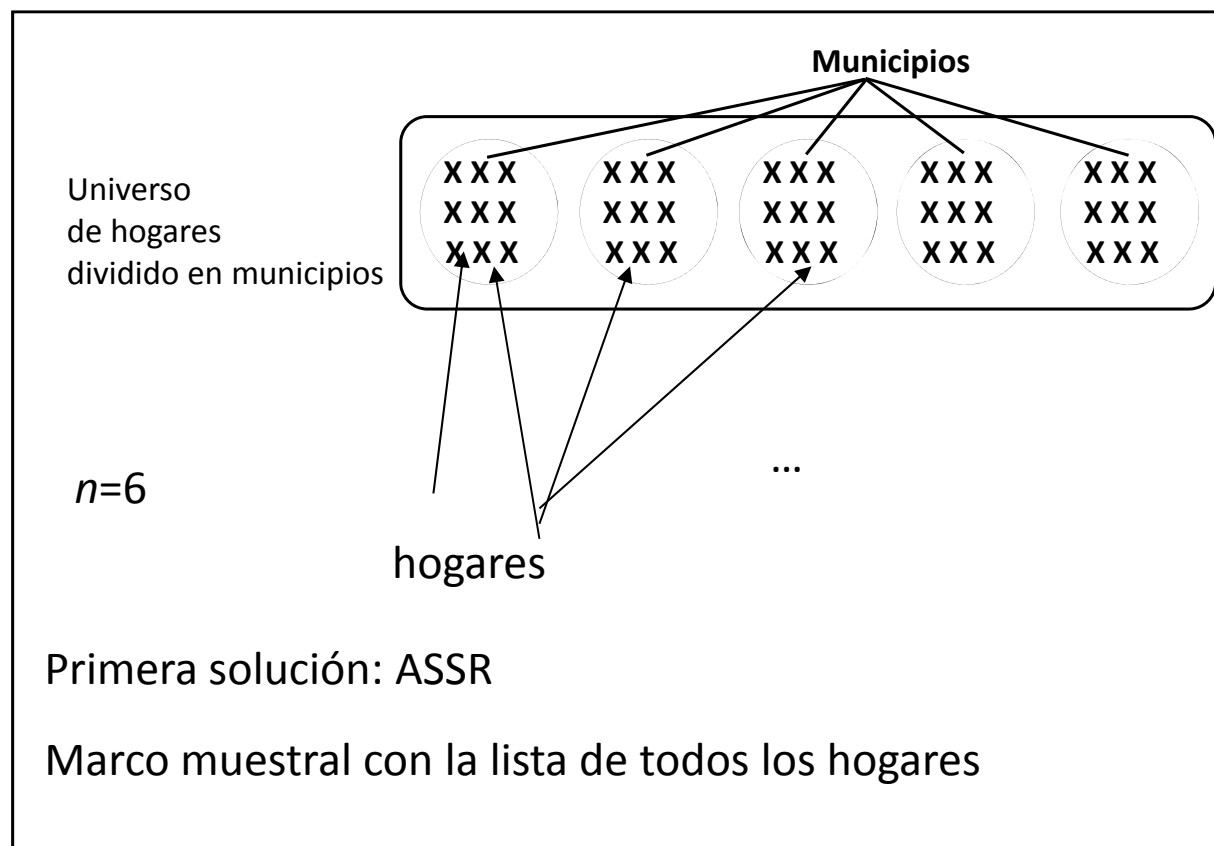
1. Principio y notaciones

a) principio

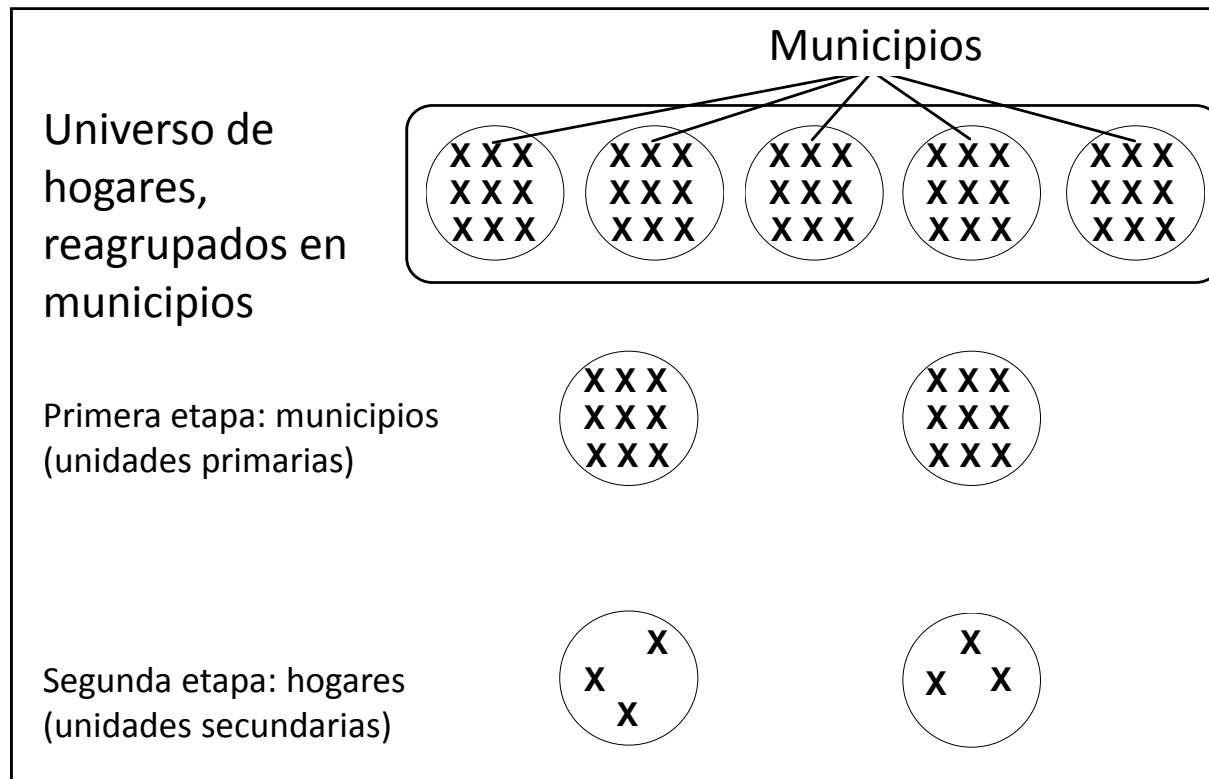
La información auxiliar no se utiliza únicamente para mejorar la precisión de los estimadores. Puede también servir a mejorar la organización de una encuesta.

El muestreo multietápico busca ahorrar.

Ejemplo: seleccionar hogares de una población



Segunda solución: en 2 etapas



Marco muestral con la lista de todos los municipios

Marco muestral (etapa 2) de los hogares en los municipios seleccionados

- En cada etapa se pueden utilizar los métodos presentados en los capítulos anteriores
- Muestreo en conglomerados: en la última etapa, se observan a todas las unidades

b) Justificación, características

Ejemplo: se desea estudiar 2 000 hogares en un país que contiene aproximadamente 500 000, repartidos en 6 000 aldeas. Se dispone únicamente de una lista de las aldeas con una estimación de su población.

- Visitar cada aldea para elaborar una lista de hogares a nivel nacional sería una tarea gigantesca.
- Los hogares de la muestra se encontrarían extremadamente dispersos (enorme pérdida de tiempo en desplazamientos, coste de la operación resultaría prohibitivo).

El muestreo en varias etapas permite resolver los dos problemas mencionados:

- sólo requiere de la lista exhaustiva de las unidades primarias
- habrá economías globales de tiempo y gastos de desplazamiento

El precio a pagar por estas ventajas es que la precisión del muestreo en varias etapas es, en general, menor que la que tendría un muestreo de una sola etapa con el mismo tamaño muestral (en número de unidades estadísticas de última etapa).

Se produce el efecto conglomerado

Efecto de conglomerado:

- las unidades estadísticas reagrupadas en una misma unidad primaria tienen tendencia a parecerse.
- la concentración de la muestra total en una muestra de unidades primarias conduce a una cierta “redundancia” de la información sobre dichas unidades, y a una cierta “falta de representatividad” del conjunto.
- en consecuencia, en el caso de una extracción en varias etapas, la mayor parte de la varianza de los estimadores proviene generalmente de la primera etapa.

c) Notación

muestreo en dos etapas

Unidades primarias:

- M en el universo ($\alpha = 1, \dots, M$)
- m extraídas de la muestra ($i = 1, \dots, m$)

Unidades secundarias:

- N_α en la unidad primaria α ($\beta = 1, \dots, N_\alpha$)
- n_i en la muestra para la unidad primaria i ($j = 1, \dots, n_i$)
- Las unidades se identifican por $\alpha \beta$: unidad secundaria β de la unidad primaria α

Total de la variable en la unidad primaria α :

$$T_{\alpha}(Y) = \sum_{\beta=1}^{N_{\alpha}} Y_{\alpha\beta}$$

Con $Y_{\alpha\beta}$, valor de la variable Y para la unidad secundaria β de la unidad primaria α

El total sobre el universo vale:

$$T(Y) = \sum_{\alpha=1}^M T_{\alpha}(Y)$$

Varianza de los totales:

$$S_T^2 = \frac{1}{M-1} \sum_{\alpha=1}^M \left(T_{\alpha}(Y) - \bar{T} \right)^2$$

con

$$\bar{T} = \frac{1}{M} \sum_{\alpha=1}^M T_{\alpha}(Y)$$

2. Diseño en conglomerados

a) Principio

Se entrevistan todas las unidades de la “última etapa”:

Por ejemplo, se extrae una muestra de aldeas, en el interior de las cuales se entrevistará a todos los hogares.

El interés reside en reducir los costes de desplazamiento y en la no obligación de disponer de un marco muestral completo.

b) Estimación de un total en el caso de extracción de conglomerados con probabilidades iguales

$$\hat{T}(Y) = \frac{M}{m} \sum_{i=1}^m T_i(Y)$$

T_i es el verdadero total de la UP i

La varianza del estimador del total se puede estimar a partir de la muestra por:

$$\hat{V}(\hat{T}(Y)) = M^2 \left(1 - \frac{m}{M}\right) \frac{1}{m} \cdot \left(\frac{1}{m-1} \sum_{i=1}^m \left(T_i(Y) - \hat{T}(Y)\right)^2 \right)$$

$$V(\hat{T}(Y)) = M^2 \left(1 - \frac{m}{M}\right) \frac{S_T^2}{m} \quad \hat{V}(\hat{T}(Y)) = M^2 \left(1 - \frac{m}{M}\right) \frac{s_T^2}{m}$$

c) Estimación de una media (por unidad estadística elemental, es decir, cada unidad secundaria) en el caso de la extracción de conglomerados con probabilidades iguales

Si se conoce el número total N de unidades estadísticas en el universo, se estima la media y la varianza del estimador de la media por:

$$\frac{1}{N} \hat{T}(Y) \qquad \hat{V}\left(\frac{\hat{T}(Y)}{N}\right) = \frac{1}{N^2} \hat{V}(\hat{T}_Y)$$

Cuando no se conoce N , se estima N por:

$$\hat{N} = \frac{M}{m} \sum_{i=1}^m N_i$$

Y, en este caso, el estimador de la media es:

$$\frac{\hat{T}(Y)}{\hat{N}}$$

La varianza de este estimador, más compleja de calcular, es la de una razón (capítulo ASSR).

3. El efecto de conglomerados

a) Principio

El hecho de extraer en dos etapas, o de extraer conglomerados, induce a menudo a una pérdida de precisión (respecto del muestreo simple con la misma cantidad de unidades encuestadas) debido a que las unidades situadas en el interior de una misma unidad primaria tienden a parecerse.

b) El coeficiente de correlación intraconglomerado

$$\delta = \frac{\sum_{\alpha=1}^M \sum_{\substack{\beta=1 \\ \gamma \neq \beta}}^{N_{\alpha}} \sum_{\gamma=1}^{N_{\alpha}} (Y_{\alpha\beta} - \bar{Y})(Y_{\alpha\gamma} - \bar{Y})}{\sum_{\alpha=1}^M \sum_{\beta=1}^{N_{\alpha}} (Y_{\alpha\beta} - \bar{Y})^2} * \frac{1}{\bar{N} - 1}$$

c) Consecuencias sobre la precisión del muestreo

Si se procede a una extracción en dos etapas, o por conglomerados, si todas las unidades primarias tienen el mismo tamaño \bar{N}

Si el tamaño de la muestra de unidades secundarias por unidad primaria es constante e igual a \bar{n}

entonces:

$$V(\hat{T}(Y)) = (1 + \delta \cdot (\bar{n} - 1)) V_{ASSR}(\hat{T}(Y))$$

La magnitud $DEFF$ permite estimar la pérdida de precisión debida al paso de un plan de muestreo a otro. Se le denomina "efecto de diseño" ("*Design Effect*")

$$DEFF = 1 + \delta(\bar{n} - 1)$$

Problema

Un banco tiene 39800 clientes-empresas en sus ficheros informáticos repartidos en 3980 agentes. Cada agente gestiona exactamente 10 clientes. Se desea estimar la proporción de clientes a los cuales el banco ha hecho un préstamo.

Al azar (MAS), se seleccionan 40 agentes. Para cada agente seleccionado, se observa el número de clientes que han pedido un prestamos. Se obtiene:

nr	1		40	Suma
efec	185
Efec ²	1263

Estimar la proporción de clientes a los cuales el banco ha hecho un préstamo.

Problema 6

Un banco tiene 39800 clientes-empresas en sus ficheros informáticos repartidos en 3980 agentes. Cada agente gestiona exactamente 10 clientes. Se desea estimar la proporción de clientes a los cuales el banco ha hecho un préstamo.

Al azar (MAS), se seleccionan 40 agentes. Para cada agente seleccionado, se observa el número de clientes que han pedido un prestamos. Se obtiene:

nr	1		40	Suma
efec	185
Efec ² _{$\hat{p} = \frac{\hat{T}}{N} = 46.25\%$}			...	1263

Estimar la proporción de clientes a los cuales el banco ha hecho un préstamo.

$$N=39800 \quad M=3980 \quad N_a=10 \quad m=40 \quad n=400 \quad N_i=10$$

$$S_T^2 = 10.445$$

$$\hat{T}(Y) = \frac{M}{m} \sum_{i=1}^m T_i(Y) = \frac{3980}{40} 185 = 18407.5$$

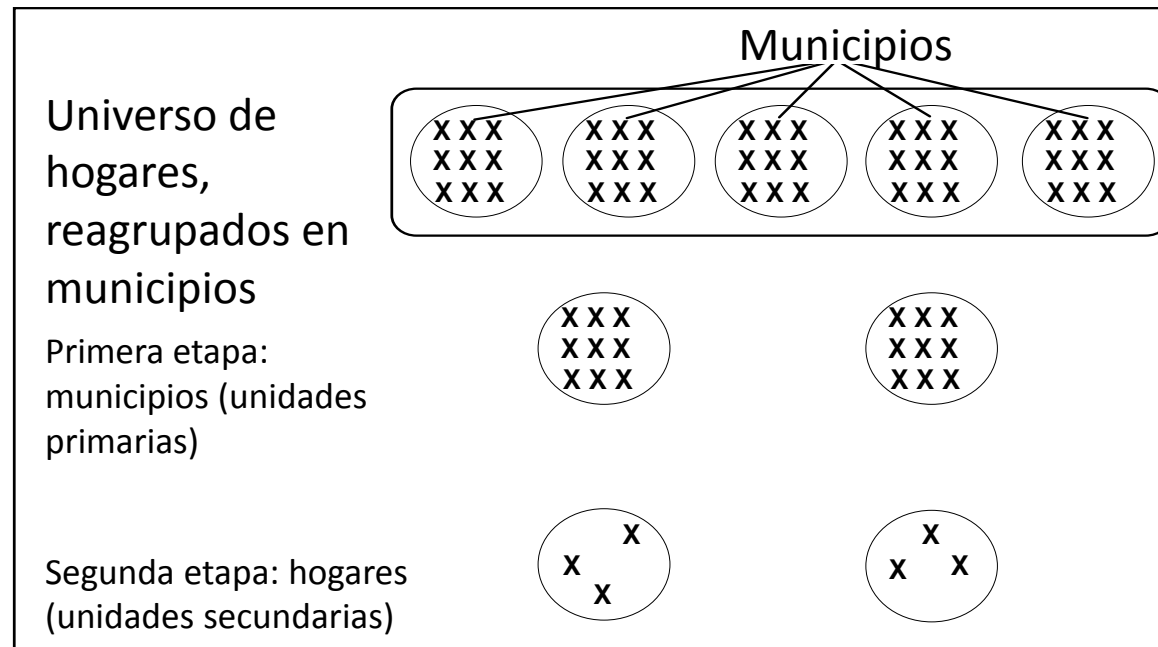
$$\hat{p} = \frac{\hat{T}}{N} = 46.25\%$$

$$\hat{V}(\hat{p}) = \frac{\hat{V}(\hat{T}(Y))}{N^2} = \frac{1}{N^2} \frac{M^2}{m} \left(1 - \frac{m}{M}\right) s_T^2$$

IC para P:

$$[46.2\% \pm 10.2\%]$$

4. Extracción en dos etapas



Marco muestral con la lista de todos los municipios

Marco muestral (etapa 2) de los hogares en los municipios seleccionados

- En cada etapa se pueden utilizar los métodos presentados en los capítulos anteriores
- Muestreo en conglomerados: en la última etapa, se observan a todas las unidades

Extracción de las unidades primarias con probabilidades iguales (extracción en dos etapas)

a) Estimador del total

- Estimador del total en unidad primaria i seleccionada

$\hat{T}_i(Y)$ corresponde al plan de muestreo escogido en la segunda etapa

- Estimador del total $\hat{T}(Y) = \frac{M}{m} \sum_{i=1}^m \hat{T}_i(Y)$ sin sesgo si $\hat{T}_i(Y)$ sin sesgo

Por ejemplo, si en la segunda etapa se ha efectuado una extracción aleatoria simple,

$$\hat{T}_i(Y) = \frac{N_i}{n_i} \sum_{j=1}^{n_i} y_{ij} = N_i \bar{y}_i$$

b) Varianza del estimador del total de Y

$$V(\hat{T}(Y)) = M^2 \left(1 - \frac{m}{M} \right) \frac{S_T^2}{m} + \frac{M}{m} \sum_{\alpha=1}^M Z_{\alpha}$$


varianza del estimador

$\hat{T}_{\alpha}(Y)$ del total $T_{\alpha}(Y)$

Si en la segunda etapa se han extraído n_{α} unidades en cada unidad primaria α , con probabilidades iguales y sin reposición:

$$Z_{\alpha} = N_{\alpha}^2 \left(1 - \frac{n_{\alpha}}{N_{\alpha}} \right) \frac{S_{\alpha}^2}{n_{\alpha}}$$

c) Estimación de la varianza del estimador del total de Y

$$\hat{V}(\hat{T}(Y)) = M^2 \left(1 - \frac{m}{M} \right) \frac{s_T^2}{m} + \frac{M}{m} \sum_{i=1}^m \hat{Z}_i$$


con
$$s_T^2 = \frac{1}{m-1} \sum_{i=1}^m \left(\hat{T}_i(Y) - \frac{\hat{T}(Y)}{M} \right)^2$$

Estimador de la varianza del
estimador $\hat{T}_i(Y)$

Si ASSR en la segunda etapa

$$\hat{Z}_i = N_i^2 \left(1 - \frac{n_i}{N_i} \right) \frac{s_i^2}{n_i}$$

d) Notas

$$V(\hat{T}(Y)) = \frac{M^2}{m} \left(1 - \frac{m}{M}\right) S_T^2 + \frac{M}{m} \sum_{\alpha=1}^M Z_{\alpha}$$

- En la expresión de la varianza, el primer término es en general el más importante. Los dos términos de esta fórmula se refieren a las dos etapas de extracción y permiten descomponer la varianza en las partes correspondientes a cada una.
- Si aumenta m , los dos términos disminuyen; si se aumentan los tamaños de muestra en la segunda etapa, únicamente el segundo término disminuye.

e) Estimación de una media

Habitualmente, el número total N de unidades estadísticas en el universo es desconocido, (de hecho, no se tiene un marco muestral para las unidades secundarias, sino únicamente la lista de unidades primarias).

Por lo tanto, para estimar la media de Y por unidad estadística en el universo a partir del total, hay que estimar este número total desconocido a partir de la muestra de unidades primarias.

Llamando a la estimación \hat{N} de N ; se estima la media por .

$$\hat{T}(Y) / \hat{N}$$

Caso particular: muestreo autoponderado

Si las unidades primarias se extraen con probabilidades iguales y si, además, la tasa de muestreo es la misma para la segunda etapa de muestreo (también con probabilidades iguales) en el interior de todas las unidades primarias extraídas:

$$\hat{T}(Y) = \frac{M}{m} \sum_{i=1}^m \frac{N_i}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

↑
constante

Misma ponderación para todas las unidades estadísticas de la muestra: muestreo autoponderado.

En este caso, la media simple calculada sobre el conjunto de las unidades extraídas se utiliza como estimador de la media sobre el universo

Inconveniente: El tamaño de la muestra depende de las UP seleccionadas

Problema 1

Se quiere conocer el tiempo medio que pasan delante de su televisor los clientes de una cadena de televisión por cable. Las entrevistas, que comportan otras preguntas, se harán cara a cara

Hay 100000 abonados. Interesa un tipo de muestreo que ocasione pocos desplazamientos del entrevistador.

Se decide entrevistar a 100 personas, seleccionadas en dos etapas. Se forma 1000 UP según la zona geográfica, de forma que cada UP tenga los mismos efectivos. Se decide seleccionar 10 UP en la primera etapa y emplear una tasa de muestreo constante en la segunda etapa.

1. ¿cuál es el tamaño de cada submuestra?

Se indican los resultados observados en la muestra en la tabla 1.

2. Estimar el tiempo medio y el tiempo total frente al televisor, por punto y por intervalo

UP	1	2	3	4	5	6	7	8	9	10
\bar{y}	20	30	20	20	10	30	10	10	20	40
s^2	12	16	10	10	2	20	4	4	4	10

6. Consideraciones prácticas

¿Cuándo utilizar muestreos multietápicos?

Los métodos de este tipo son eficientes en ciertos dominios de estudio, absolutamente ineficientes en otros (cuando existe una cierta “concentración” del fenómeno estudiado en ciertas unidades primarias, por ejemplo). La eficiencia dependerá de la capacidad de construir unidades primarias heterogéneas (se debe buscar que cada unidad primaria contenga individuos diferentes entre sí).

Además, es interesante **estratificar las unidades primarias**: por ejemplo, si éstas son unidades geográficas, se crearán estratos basados en criterios administrativos o agroecológicos. Es en esta etapa donde se gana precisión. La estratificación de las unidades primarias está ligada a menudo con la voluntad de proporcionar resultados para cada estrato.

La estratificación de las unidades secundarias, por el contrario, se revela mucho menos rentable (y además puede transformar el empadronamiento en una operación demasiado pesada y propensa a errores).

Problema 3

Se busca estimar la renta media de los hogares de un distrito de una ciudad compuesta de 60 secciones censales. Se sabe que hay 5000 hogares en el distrito. Se seleccionan al azar 3 secciones y se entrevistan a todos los hogares de las secciones seleccionadas

Sección	1	2	3
Nr hogares en la sección	120	100	80
Suma de la renta (en miles de euros)	2100	2000	1500

Estimen el total de las rentas y la renta media por hogar, en los dos casos por punto y por intervalo a partir de los resultados.