Muestreo de conglomerados

Introducción

Hasta ahora se han visto diseños muestrales en los que las unidades muestrales eran las mismas unidades que constituían la población en estudio; por tanto se requería que en el marco poblacional estuvieran listadas todas las unidades de la población.

Sin embargo, en numerosas ocasiones, no es factible o incluso no es posible enumerar todos los elementos que constituyen la población. En este caso, hay que cambiar de estrategia de modo que las unidades muestrales diferirán de las que constituyen explícitamente la población. Es decir, las unidades muestrales van a estar formados por dos o más unidades físicas. Un ejemplo de ello es el estudio de una determinada característica sobre la población de una ciudad grande donde no se puede enumerar a todos los habitantes de la ciudad.

Por ello se plantea el muestreo de conglomerados, en el que la unidad de muestreo, o unidad de selección, contiene más de un elemento de la población. A la unidad de muestreo se le llama **conglomerado**, que es un conjunto de elementos de la población.

A los conglomerados se denominan *unidades primarias*, y a las unidades en que están divididos los conglomerados se denominan *unidades secundarias*. Estos conceptos se pueden generalizar según las etapas en las conste el muestreo por conglomerados: monoetápico, bietápico o polietápico.

a) **Monoetápico**: se seleccionan m conglomerados y se observan *todas* las unidades secundarias en cada conglomerado seleccionado.

Un ejemplo de este diseño muestral es cuando en una ciudad se desea estimar la superficie media de las viviendas, y para ello se consideran las calles como conglomerados y como unidades secundarias todas las viviendas de cada calle.

b) **Bietápico**: se seleccionan m conglomerados y dentro de los conglomerados seleccionados se seleccionan n_i unidades secundarias.

Un ejemplo de este diseño muestral es cuando en una ciudad se desea estimar la superficie media de las viviendas, se realiza una selección de m calles (conglomerados) y

en cada calle seleccionada se realiza una selección de n_i viviendas.

Criterio para formar conglomerados

Un conglomerado no es más que un subconjunto de elementos de la población, de manera que dichos subconjuntos pueden ser naturales, espaciales, geográficos... Para formar conglomerados y su estructura no se tiene una regla fija, sino que depende del estudio a realizar y de los recursos disponibles. Así, las viviendas, las calles o los países pueden ser conglomerados.

Por ejemplo, si queremos estimar el número total de personas en una ciudad que tienen hipertensión, los conglomerados pueden ser las calles de la ciudad, las unidades secundarias los edificios, las unidades terciarias estarán los pisos y las unidades elementales serán las personas.

Ventajas del muestreo por conglomerados

Ventajas de infraestructura: en ocasiones, no se tiene una lista completa de la población o resulta muy costosa poder conseguirla y sin embargo se puede tener información con mayor facilidad o una lista de unidades primarias, como es el caso de calles o viviendas.

Por ejemplo, una lista de todos los niños de educación primaria de la Comunidad de Madrid es muy costosa de conseguir, pero es fácil conseguir una lista de los colegios (conglomerados) y en ellos se puede obtener fácilmente una lista de los alumnos de esos colegios.

Ventajas económicas: entendiendo esta razón en el sentido más amplio, en cuanto a costes económicos y tiempo. Esta es la forma más barata de muestreo, ya que no solo los costes de listado son casi siempre los más bajos (entre los distintos métodos) sino que también los costes de viaje son los más bajos puesto que las unidades, en general, están concentradas.

Inconvenientes del muestreo por conglomerados

Los errores de muestreo en conglomerados son generalmente grandes, comparados con los que se obtienen por otros métodos de muestreo. La razón es que en la mayoría de las poblaciones, las unidades que pertenecen a un mismo conglomerado tienen cierto grado de homogeneidad respecto a muchas características, de manera que las unidades elementales se reparten menos uniformemente sobre la población.

Por ejemplo, las viviendas de un mismo barrio son a menudo bastante similares respecto a las características socio-económicas.

A pesar de este inconveniente, esta técnica muestral es muy útil en poblaciones muy numerosas ya que es de diseño relativamente simple y capaz de reducir los costes de

observación de los elementos de los conglomerados que entran a formar parte de la muestra. Si tenemos que elegir en términos del coste, siempre se elegirá muestreo por conglomerados, pero si tenemos que decidir solo en términos de la precisión, no se tomaría muestreo por conglomerados. En general, se tienen en cuenta ambos puntos de vista y se elige el método más preciso para un coste dado, o el que dé con menor coste la precisión de los estimadores que se fije previamente.

Diferencias entre estratos y conglomerados

A simple vista tanto los estratos como conglomerados parecen semejantes, pero hay diferencias muy relevantes:

i) En el muestreo por conglomerados se seleccionan m de ellos y estos se examinan totalmente o parcialmente, mientras que en el muestreo estratificado se seleccionan todos los estratos y se examinan parcialmente; es decir, en la muestra figuran algunas unidades de cada uno de los estratos.

ii) Aunque tanto estratos como conglomerados son una partición de la población, los estratos suelen ser más grandes que los conglomerados.

iii) En el muestreo estratificado hay homogeneidad dentro de los estratos es decir hay **heterogeneidad entre** los estratos respecto a la característica en estudio de la población.

Sin embargo, en el muestreo por conglomerados hay **heterogeneidad dentro** de los conglomerados, es decir, hay homogeneidad entre los conglomerados respecto a la característica en estudio.

En la práctica, el muestreo por conglomerados se utiliza en diseños muestrales complejos junto con el muestreo estratificado y muestreo aleatorio. Un ejemplo de ello es la encuesta de población activa (*EPA*) que utiliza conjuntamente conglomerados, estratos y selección aleatoria de las últimas unidades.

Se puede observar que si consideramos el ejemplo de conglomerados como las calles, se puede ver que no todas las calles tienen el mismo número de viviendas, con lo cual los conglomerados tienen distintos tamaños. En consecuencia, se estudia el muestreo por conglomerados atendiendo al criterio de las etapas de selección (monoetápico, bietápico) y con tamaño distinto de cada uno de los conglomerados, de modo que el caso de igual tamaño será un caso particular.

Diseño muestral en una etapa y parámetros asociados

Se tiene una población U de tamaño N repartida en M conglomerados que denotamos por C_i, cada uno de tamaño N_i siendo $i = 1, \ldots, M$, de modo que

$$U = \bigcup_{i=1}^{M} C_i$$

y tenemos $N = \sum_{i=1}^{M} N_i$ unidades elementales de la población, donde el número medio de unidades por conglomerado es $\overline{N} = \dfrac{N}{M}$.

El diseño muestral por conglomerados monoetápico consiste en:

a) Seleccionar una muestra probabilística s_c de conglomerados de acuerdo a un diseño $P_c(\cdot)$. El tamaño de s_c lo denotamos por m (para un diseño de tamaño fijo) o por m_{s_c} (para un diseño de tamaño variable).

b) Se observan todos los elementos de la población en los conglomerados seleccionados.

Se puede observar que $P_c(\cdot)$ puede ser la función de probabilidad de un diseño aleatorio simple, sistemático o estratificado. Aquí se particulariza para el muestreo aleatorio simple.

Denotamos por s el conjunto de elementos observados en la muestra

$$s = \bigcup_{i \in s_c} C_i$$

y el tamaño de la muestra s vendrá dado por

$$n_s = \sum_{i \in s_c} N_i = \sum_{i=1}^{m} N_i.$$

Probabilidades de inclusión de los elementos poblacionales

Como la muestra s contiene a todos los elementos de los conglomerados seleccionados entonces:

- La probabilidad de que el conglomerado k se seleccione y pertenezca a s_c se denota como

$$\pi_k^c = P(C_k \in s_c)$$

- La probabilidad de que los conglomerados k, l pertenezcan a la muestra de conglomerados s_c se denota como

$$\pi_{kl}^c = P(C_k, C_l \in s_c)$$

Probabilidades de inclusión de las unidades elementales

Se trata de calcular $P\{u_i \in s\}$. Se tiene que la unidad i pertenecerá a un conglomerado k cuando $\pi_i = \pi_k^c$ para todo $i \in C_k$.

Del mismo modo,

$$\pi_{ij} = P\{u_i, u_j \in s\} = \begin{cases} \pi_k^c & \text{si} \quad u_i, u_j \in C_k \\ \pi_{kl}^c & \text{si} \quad u_i \in C_k \; u_j \in C_l \end{cases}$$

Se define

$$\Delta_{ij}^c = \pi_{ij}^c - \pi_i^c \pi_j^c$$

Entonces, dada la población U y la característica Y tal que Y_{ij} es la observación de la unidad j del conglomerado i, se definen los siguientes parámetros:

- Total poblacional del i-ésimo conglomerado

$$Y_i = \sum_{j=1}^{N_i} Y_{ij}$$

- Total de la población

$$Y = \sum_{i=1}^{M} Y_i$$

- Media poblacional del i-ésimo conglomerado

$$\overline{Y}_i = \frac{Y_i}{N_i}$$

- Media poblacional de totales de conglomerados. Es decir, la media poblacional de Y por conglomerado, o media poblacional por unidad primaria:

$$\overline{Y_c} = \frac{Y}{M}$$

- Media poblacional por unidad secundaria

$$\overline{\overline{Y}} = \frac{Y}{N}$$

- Cuasivarianza poblacional

$$S_Y^2 = \frac{1}{N-1} \sum_{i=1}^{M} \sum_{j=1}^{N_i} (Y_{ij} - \overline{\overline{Y}})^2$$

De la misma manera que en el muestreo estratificado consideramos la descomposición de la variabilidad total para poder definir una medida de homogeneidad dentro de los conglomerados.

$$\sum_{i=1}^{M}\sum_{j=1}^{N_i}(Y_{ij}-\overline{\overline{Y}})^2=\sum_{i=1}^{M}\sum_{j=1}^{N_i}(Y_{ij}-\overline{Y}_i)^2+\sum_{i=1}^{M}\sum_{j=1}^{N_i}(\overline{Y}_i-\overline{\overline{Y}})^2$$

Con respecto a los grados de libertad, se tiene que

$$(N-1)=(N-M)+(M-1)$$

Así

$$(N-1)S_Y^2=\sum_{i=1}^{M}(N_i-1)S_i^2+\sum_{i=1}^{M}N_i(\overline{Y}_i-\overline{\overline{Y}})^2$$

donde

$$S_i^2=\frac{1}{N_i-1}\sum_{j=1}^{N_i}(Y_{ij}-\overline{Y}_i)^2$$

que es la cuasivarianza en el conglomerado i.

Se resume en que:

$$\text{Var. Total (SCT)} = \text{Var. dentro de Conglomerados (SCD)} + \text{Var. entre Conglomerados (SCE)}$$

Definimos el **coeficiente de homogeneidad** de los conglomerados, δ, como

$$\delta=1-\frac{(N-1)}{(N-M)}\frac{SCD}{SCT}=1-\frac{S_{dentro}^2}{S_Y^2}$$

donde S_Y^2 es la cuasivarianza poblacional.

δ será positivo o negativo según S_{dentro}^2 sea menor o igual que la cuasivarianza total.

El coeficiente de homogeneidad δ satisface que

$$-\frac{M-1}{N-M}\leq\delta\leq 1$$

ya que $S_{dentro}^2\geq 0$ y por la definición de δ se tiene que $\delta\leq 1$.

Por otra parte, cuando SCD es igual a la total SCT, se tiene que

$$\frac{S_{dentro}^2}{S_Y^2}=\frac{N-1}{N-M}$$

por tanto,

$$\delta=1-\frac{S_{dentro}^2}{S_Y^2}\geq 1-\frac{N-1}{N-M}=-\frac{M-1}{N-M}$$

- El valor $\delta=0$ es obtenido cuando $S_{dentro}^2=S_Y^2$.

- El valor $\delta = 1$ se obtiene cuando la homogeneidad dentro de los conglomerados es máxima.
- El valor $\delta = -\frac{M-1}{N-M}$ se obtiene cuando la heterogeneidad dentro de los conglomerados es máxima.

Estimadores de los parámetros poblacionales

Dado un diseño muestral monoetápico por conglomerados y dado el total poblacional $Y = \sum_{i=1}^{M} Y_i$ entonces el estimador del total de la población se puede escribir como

$$\widehat{Y} = \sum_{i \in s_c} \frac{Y_i}{\pi_i^c}$$

cuya varianza es:

$$\begin{aligned} V(\widehat{Y}) &= \sum_{i=1}^{M}\sum_{j=1}^{M} \Delta_{ij}^c \frac{Y_i}{\pi_i^c}\frac{Y_j}{\pi_j^c} = \\ & \sum_{i=1}^{M}\sum_{j=1}^{M} (\pi_{ij}^c - \pi_i^c \pi_j^c)\frac{Y_i}{\pi_i^c}\frac{Y_j}{\pi_j^c} \end{aligned}$$

Un estimador insesgado de la varianza es

$$\begin{aligned} \widehat{V}(\widehat{Y}) &= \sum_{i \in s_c}\sum_{j \in s_c} \widehat{\Delta}_{ij}^c \frac{Y_i}{\pi_i^c}\frac{Y_j}{\pi_j^c} = \\ & \sum_{i \in s_c}\sum_{j \in s_c} \frac{(\pi_{ij}^c - \pi_i^c \pi_j^c)}{\pi_{ij}^c}\frac{Y_i}{\pi_i^c}\frac{Y_j}{\pi_j^c}. \end{aligned}$$

Si el diseño muestral es de tamaño fijo entonces se tiene que

i) La varianza es

$$V(\widehat{Y}) = -\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} (\pi_{ij}^c - \pi_i^c \pi_j^c)\left(\frac{Y_i}{\pi_i^c} - \frac{Y_j}{\pi_j^c}\right)^2$$

ii) Un estimador insesgado de $V(\widehat{Y})$ es

$$\widehat{V}(\widehat{Y}) = -\frac{1}{2}\sum_{i \in s_c}\sum_{j \in s_c} \frac{(\pi_{ij}^c - \pi_i^c \pi_j^c)}{\pi_{ij}^c}\left(\frac{Y_i}{\pi_i^c} - \frac{Y_j}{\pi_j^c}\right)^2$$

Observaciones

Si consideramos un diseño muestral de conglomerados de **tamaño fijo** se tienen las siguientes conclusiones:

En un diseño muestral de conglomerados de tamaño fijo, si para todo $i, j \in U$

$$\frac{Y_i}{\pi_i^c} \approx \frac{Y_j}{\pi_j^c} \approx cte$$

entonces $V(\widehat{Y}) \approx 0$.

Luego si podemos elegir un diseño muestral con probabilidades de inclusión **proporcionales al total** del conglomerado, entonces el muestreo por conglomerados será muy eficiente.

Si conocemos de antemano el tamaño de los conglomerados N_i, un diseño muestral por conglomerados preciso será el que considera las probabilidades de inclusión π_i^c proporcionales a los **tamaños** de los conglomerados Esto es posible siempre y cuando no haya mucha variación entre las medias poblacionales de los conglomerados: condición de homogeneidad **entre** conglomerados y heterogeneidad **dentro** de ellos.

Así, si las medias de los conglomerados son similares, bastará utilizar un diseño muestral con probabilidades de inclusión proporcionales a los tamaños de los conglomerados:

$$\frac{N_i \overline{Y}_i}{\pi_i^c} = cte$$

Si $\overline{Y}_i \approx cte \Rightarrow$

$$\frac{N_i}{\pi_i^c} = cte$$

Es decir,

$$\left(\frac{Y_i}{\pi_i^c} - \frac{Y_j}{\pi_j^c}\right)^2 = \left(\frac{N_i \overline{Y}_i}{\pi_i^c} - \frac{N_j \overline{Y}_j}{\pi_j^c}\right)^2 \propto \left(\frac{N_i}{\pi_i^c} - \frac{N_j}{\pi_j^c}\right)^2$$

Por lo tanto, es esta situación, si se toman las probabilidades de inclusión de los conglomerados proporcionales a sus tamaños, el muestreo por conglomerados será muy eficiente.

Si se toman π_i^c iguales en una situación en la que los tamaños de los conglomerados son distintos, el diseño por conglomerados presentará poca precisión, excepto cuando $Y_i = Y_j$.

En este caso,

$$\left(\frac{Y_i}{\pi_i^c} - \frac{Y_j}{\pi_j^c}\right)^2 = \left(\frac{Y_i - Y_j}{\pi_i^c}\right)^2 = \left(\frac{N_i \overline{Y}_i - N_j \overline{Y}_j}{\pi_i^c}\right)^2$$

Por ello, para que fuera una estimación precisa tendría que:

$$N_i\overline{Y}_i \approx N_j\overline{Y}_j = cte$$

es decir

$$\overline{Y}_i \approx \frac{cte}{N_i} \Rightarrow Y_i \propto \frac{1}{N_i}$$

lo cual es difícil si N_i y N_j son muy distintas.

Resumen

El diseño por conglomerados en una etapa será **eficiente** si:

a) Las π_i^c son proporcionales a los totales Y_i

b) Las π_i^c son proporcionales al tamaño de los conglomerados cuando las $\overline{Y}_i$ son similares.

Muestreo aleatorio simple de conglomerados en una etapa

Supongamos un muestreo de conglomerados con probabilidades de inclusión $\pi_i^c = \frac{m}{M}$ con m iguales, seleccionamos una muestra aleatoria simple de conglomerados y observamos todos los elementos de los conglomerados seleccionados.

Se desea estimar el total poblacional $Y = \sum_{i=1}^{M} Y_i$, distinguiendo entre el estimador directo y el estimador indirecto.

Estimador directo

Utilizando las propiedades del estimador de Horvitz-Thompson para conglomerados se tiene que un estimador del total poblacional es

$$\begin{aligned} \widehat{Y} &= \sum_{i\in s_c} \frac{Y_i}{m/M} \Rightarrow \\ \widehat{Y} &= \frac{M}{m}\sum_{s\in s_c} Y_i = M\widehat{\overline{Y}}_c \end{aligned}$$

con Y_i total del conglomerado i-ésimo, siendo $\widehat{\overline{Y}}_c$ la media de los totales de conglomerados.

La varianza de $\widehat{Y}$ viene dada por

$$\begin{aligned} V(\widehat{Y}) &= M^2\frac{(1-\frac{m}{M})}{m}S_y^2 = \\ & \frac{(M-m)}{m}\frac{M}{M-1}\sum_{i=1}^{M}(Y_i-\overline{Y}_c)^2 \end{aligned}$$

Un estimador insesgado de la varianza de $\widehat{Y}$ viene dado por

$$\begin{aligned}\widehat{V}(\widehat{Y}) &= M^2\frac{(1-\frac{m}{M})}{m}\widehat{S}_y^2 = \\ & M^2\frac{(1-\frac{m}{M})}{m}\frac{1}{m-1}\sum_{i\in s_c}(Y_i-\widehat{\overline{Y}}_c)^2 = \\ & \frac{(M-m)}{m}\frac{M}{m-1}\sum_{i\in s_c}(Y_i-\widehat{\overline{Y}}_c)^2\end{aligned}$$

Análogamente, un estimador de la media poblacional de unidades elementales $\overline{\overline{Y}}$ es

$$\widehat{\overline{\overline{Y}}} = \frac{\widehat{Y}}{N} = \frac{M}{N}\widehat{\overline{Y}}_c$$

y el estimador insesgado de la varianza del estimador es

$$\widehat{V}(\widehat{\overline{\overline{Y}}}) = \frac{(1-\frac{m}{M})}{m\overline{N}^2}\widehat{S}_y^2$$

donde $\overline{N} = \frac{N}{M}$ es el número medio de elementos por conglomerado.

Observación: El problema que presentan estos estimadores es que, en general, serán estimadores con baja precisión. Un caso particular es cuando las medias de los conglomerados $\overline{Y}_i$ varían muy poco de un conglomerado a otro y sin embargo los tamaños de los conglomerados son muy diferentes, con lo cual los totales de los conglomerados $Y_i = N_i\overline{Y}_i$ varían mucho y en consecuencia las varianzas de los estimadores serán muy grandes, obteniéndose precisiones bajas.

Por tanto, se hace necesario buscar estimadores que sean más estables y no presenten esas perturbaciones en la varianza.

Estimador indirecto (de razón)

Si conocemos $N = \sum_{i=1}^{M} N_i$ una alternativa a los estimadores directos es un estimador de razón en el cual N_i es el número de unidades elementales del conglomerado i-ésimo $(i = 1, \ldots, M)$ que se toma como variable auxiliar de Y.

Bajo estos supuestos estamos en las condiciones idóneas para aplicar los estimadores de razón, es decir tenemos una variable en estudio Y y una variable auxiliar N_i de la que conocemos el total poblacional N.

Aplicando los resultados de la estimación de razón

$$R = \frac{Y}{N} = \frac{\sum_{i=1}^{M} Y_i}{\sum_{i=1}^{M} N_i} = \frac{M\overline{Y}_c}{N} = \overline{\overline{Y}}$$

En consecuencia, aplicando la teoría de estimadores de razón se tiene que

$$\widehat{Y}_R = N\frac{\sum_{i\in s_c} Y_i}{\sum_{i\in s_c} N_i} = N\widehat{R}$$

es un estimador del total poblacional de Y.

Por tanto, el estimador de razón de la media poblacional por unidad secundaria es

$$\widehat{\overline{\overline{Y}}}_R = \frac{\sum_{i\in s_c} Y_i}{\sum_{i\in s_c} N_i} = \widehat{R}$$

y el estimador de razón de la media poblacional por unidad primaria será:

$$\widehat{\overline{Y}}^c = \frac{1}{M}\widehat{Y}_R = \frac{N}{M}\widehat{R}$$

La varianza estimada de $\widehat{Y}_R$, teniendo en cuenta el método de aproximación lineal, será:

$$\begin{aligned}\widehat{V}(\widehat{Y}_R) &\approx M^2\frac{(1-\frac{m}{M})}{m}\frac{1}{m-1}\sum_{i\in s_c}(Y_i - \widehat{R}N_i)^2 = \\ & M^2\frac{(1-\frac{m}{M})}{m}\frac{1}{m-1}\sum_{i\in s_c} N_i^2\left(\overline{Y}_i - \widehat{\overline{\overline{Y}}}_R\right)^2\end{aligned}$$

De manera inmediata se obtiene $\widehat{V}(\widehat{\overline{\overline{Y}}}_R)$ dado que $\widehat{\overline{\overline{Y}}}_R = \dfrac{\widehat{Y}_R}{N}$:

$$\widehat{V}(\widehat{\overline{\overline{Y}}}_R) = \frac{(1-\frac{m}{M})M^2}{m(m-1)N^2}\sum_{i\in s_c} N_i^2\left(\overline{Y}_i - \widehat{\overline{\overline{Y}}}_R\right)^2$$

Observación: Considerando la familia de estimadores de razón y sus propiedades respecto a los estimadores directos, se deduce que el estimador de razón para muestreo de conglomerados va a ser más eficiente.

Ejemplo (Elementary Survey Sampling (7ed) - Scheaffer et al.)

Un sociólogo quiere estimar el ingreso per capita en una pequeña ciudad. No hay una lista de adultos residentes disponible. ¿Cómo debería diseñar el muestreo?

El muestreo por conglomerados parece ser la opción lógica para el diseño de la encuesta porque no hay listas de elementos disponible. La ciudad está delimitada en bloques rectangulares, a excepción de dos áreas industriales y tres parques que contienen solo unas pocas casas. El sociólogo decide que cada una de las manzanas de la ciudad se considerará un conglomerado, las dos las áreas se considerarán un solo conglomerado y, finalmente, los tres parques se considerarán otro solo conglomerado. Los conglomerados están numerados en un mapa de la ciudad, con números del 1 al 415. El experimentador tiene suficiente tiempo y dinero para muestrear 25 conglomerados y para entrevistar a todos los hogares dentro de cada conglomerado.

Conglomerado	N_i	Y_i	**Conglomerado**	N_i	Y_i
1	8	96000	14	10	49000
2	12	121000	15	9	53000
3	4	42000	16	3	50000
4	5	65000	17	6	32000
5	6	52000	18	5	22000
6	6	40000	19	5	45000
7	7	75000	20	4	37000
8	5	65000	21	6	51000
9	8	45000	22	8	30000
10	3	50000	23	7	39000
11	2	85000	24	3	47000
12	6	43000	25	8	41000
13	5	54000		$\sum_i N_i = 151$	$\sum_i Y_i = 1329000$

Solución

El estimador de la media es

$$\widehat{\overline{\overline{Y}}}_R = \frac{\sum_{i\in s_c} Y_i}{\sum_{i\in s_c} N_i} = \frac{1329000}{151} = 8801$$

En cuanto a la varianza

$$\begin{aligned}\widehat{V}(\widehat{\overline{\overline{Y}}}_R) &= \frac{(1-\frac{m}{M})M^2}{m(m-1)N^2}\sum_{i\in s_c} N_i^2\left(\overline{Y}_i - \widehat{\overline{\overline{Y}}}\right)^2 = \\ &\frac{(1-\frac{m}{M})}{m(m-1)\overline{N}^2}\sum_{i\in s_c}\left(Y_i - N_i\widehat{\overline{\overline{Y}}}\right)^2\end{aligned}$$

donde $\overline{N}$ es el tamaño medio de los conglomerados.

Pero dado que no se conoce el número de personas de la población N, se sustituye $\overline{N}^2$ por $\widehat{\overline{N}}^2$:

$$\widehat{\overline{N}} = \frac{\sum_{i=1}^m N_i}{m} = \frac{151}{25} = 6{,}04$$

Con lo cual

$$\widehat{V}(\widehat{\overline{\overline{Y}}}_R) = \frac{\left(1-\frac{25}{415}\right)}{25\cdot 24\cdot 6{,}04^2}\sum_{i=1}^{25}(Y_i - N_i\cdot 8801)^2 = 653{,}785$$

Y un intervalo de confianza al 95 % seria

$$\begin{aligned}\widehat{\overline{\overline{Y}}}_R \mp 1{,}96\cdot\sqrt{\widehat{V}(\widehat{\overline{\overline{Y}}}_R)} &= 8801 \mp 1{,}96\cdot\sqrt{653{,}785} = \\ &[7184; 10418]\end{aligned}$$

Lo cual implica un intervalo amplio.

Eficiencia de los conglomerados usando muestreo aleatorio simple

Sea una población U dividida en M conglomerados y se considera en ella un muestreo aleatorio simple de m conglomerados.

Por otro lado, se puede considerar alternativamente, un muestreo aleatorio simple de n unidades secundarias para estimar el total poblacional de unidades secundarias.

Supongamos para ello que la población esta dividida en conglomerados de igual tamaño $N_i = K$ para todo $i = 1, \ldots, M$ y sea $n = mK$.

Suponiendo que los costes de ambos diseños muestrales son despreciables y aplicando ambos diseños muestrales se tiene que:

Bajo m.a.s. de n unidades secundarias (de las N unidades):

$$\begin{aligned} V_{mas}(\widehat{Y}) &= N^2 \frac{(1-f)}{n} S_Y^2 = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S_Y^2 = \\ & K^2 M^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \overline{\overline{Y}})^2 \end{aligned}$$

Bajo m.a.s. de m unidades primarias y teniendo en cuenta la descomposición de la variabilidad total,

$$V_{mcm}(\widehat{Y}) = N^2 \frac{(1-f)}{n} S_{Entre}^2$$

donde

$$S_{Entre}^2 = \frac{K}{M-1} \sum_{i=1}^{M} (\overline{Y}_i - \overline{\overline{Y}})^2$$

Pero se puede demostrar que

$$\begin{aligned} S_{Entre}^2 &= S_y^2 + S_y^2 \frac{(N-M)}{M-1} \left(1 - \frac{S_{Dentro}^2}{S_y^2}\right) = \\ & S_y^2 + S_y^2 \frac{N-M}{M-1} \delta \end{aligned}$$

En consecuencia,

$$\frac{V_{mcm}(\widehat{Y})}{V_{mas}(\widehat{Y})} = 1 + \frac{N-M}{M-1} \delta$$

Por lo tanto, si $\delta < 0$ es más eficiente la realización de muestreo por conglomerados, o lo que es equivalente, si hay heterogeneidad dentro de los conglomerados.

Considerando los resultados obtenidos bajo estimadores directos y generalizando el resultado anterior se tiene que si los conglomerados son internamente homogéneos y/o sus tamaños son muy diferentes, entonces el muestreo por conglomerados no es eficiente.

Por lo tanto, cuando se realiza una partición de la población en conglomerados debe hacerse de modo que se logren conglomerados con alta variabilidad interna. En la práctica, una forma de hacerlo es usando técnicas de análisis de la varianza.

Observación: Dado un muestreo por conglomerados monoetápico hemos visto que la varianza del estimador del total poblacional se puede poner en función del coeficiente de homogeneidad δ, y en consecuencia podemos obtener un estimador de dicha varianza:

$$\widehat{V}_{mcm}(\widehat{Y}) = \left(1 + \frac{N-M}{M-1}\widehat{\delta}\right)\widehat{V}_{mas}(\widehat{Y})$$

donde

$$\widehat{\delta} = 1 - \frac{\widehat{S}^2_{dentro}}{\widehat{S}^2_y}$$

con

$$\widehat{S}^2_{dentro} = \frac{1}{N-M}\sum_{i=1}^{m}(N_i - 1)S_i^2$$

Se observa que $\widehat{V}_{mcm}(\widehat{Y})$ es sesgado en general, al serlo $\widehat{\delta}$ ya que es un estimador de razón.

Muestreo con probabilidades proporcionales al tamaño

Cuando los N_i son distintos se ha visto que el diseño óptimo se obtiene cuando se muestrean los conglomerados de manera que las probabilidades de inclusión π_i^c son proporcionales a sus tamaños (siempre que las medias $\overline{Y}_i$ no sean muy diferentes).

Pero si las medias $\overline{Y}_i$ no son iguales o similares en diseños muestrales de tamaño fijo, la expresión de Yates-Grundig de la varianza indica que para que esta sea pequeña

$$\frac{Y_i}{\pi_i^c} = cte$$

Sin embargo, esto es difícil de conseguir en principio, ya que se normalmente se desconoce Y_i para todo $i = 1, \ldots, M$.

Por ello se utiliza la variable auxiliar N_i relacionada con Y_i y se desea elegir una manera de muestrear los conglomerados tal que

$$\frac{N_i}{\pi_i^c} = cte$$

es decir $\pi_i^c = KN_i$.

Como en diseños muestrales de tamaño fijo sin reemplazamiento

$$\sum_{i=1}^{M}\pi_i = m$$

que es el tamaño muestral se tiene que

$$\begin{aligned}\sum_{i=1}^{M} KN_i &= m \Rightarrow \\ K &= \frac{m}{\sum_{i=1}^{M} N_i} = \frac{m}{N}\end{aligned}$$

de donde

$$\pi_i^c = \frac{N_i}{N} m$$

Con estas probabilidades se construyen los estimadores de Horvitz-Thomson.

Observación:

En ocasiones puede ocurrir que para algún $i = 1, \ldots, M$ esta expresión produzca que $\pi_i^c > 1$.

En este caso se reajusta y se le asigna probabilidad igual a 1 y el resto de pesos se ajustan para que sumen $m - 1$.

Ejemplo: Dada una población con tres conglomerados A, B y C de tamaños 100, 10 y 1 unidad respectivamente. Supongamos que se van a seleccionar $m = 2$ conglomerados. En consecuencia,

$$\begin{aligned} \pi_A &= \frac{2 \cdot 100}{111} > 1 \\ \pi_B &= \frac{2 \cdot 10}{111} \\ \pi_C &= \frac{2}{111} \end{aligned}$$

por tanto, se tiene que existe un k' tal que $\pi_{k'} > 1$.

Para resolver el problema hacemos

$$\pi'_A = 1$$

y

$$\begin{aligned} \pi'_B &= \alpha \frac{20}{111} \\ \pi'_C &= \alpha \frac{2}{111} \end{aligned}$$

con la condición de que

$$\alpha \left(\frac{20}{111} + \frac{2}{111} \right) = 1$$

lo que implica que

$$\alpha = \frac{111}{22}$$

Por tanto

$$\begin{aligned} \pi'_B &= \frac{20}{22} \\ \pi'_C &= \frac{2}{22} \end{aligned}$$

Observación: Sin embargo, no es fácil encontrar un método de muestreo *sin reemplazamiento* que tenga como propiedades:

a) π_i^c sean proporcionales a N_i.

b) Sea fácil de llevar a cabo.

c) $\pi_{ij}^c > 0$ para todo i, j para que se pueda estimar la varianza.

d) π_{ij}^c sean fáciles de obtener.

e) Se tenga

$$\Delta_{kl} = \pi_{kl} - \pi_k \pi_l < 0.$$

para que la varianza de Yates-Grundig sea positiva.

Esto se puede solucionar usando un esquema de muestreo **con reemplazamiento**.

Muestreo de conglomerados con probabilidades proporcionales al tamaño (*pps*)

El muestreo por conglomerados a menudo proporciona una situación ideal para utilizar el muestreo con probabilidades proporcionales al tamaño (**pps**: *Probability Proportional to Size*) porque el número de elementos en un conglomerado N_i forma una medida natural del tamaño del conglomerado.

El muestreo con probabilidades proporcionales a N_i reduce el error de estimación cuando el total del conglomerado está altamente correlacionado con el número de elementos del conglomerado.

Se tiene que el estimador de la media poblacional $\overline{Y}$ es

$$\widehat{\overline{Y}}_{pps} = \widehat{\overline{\overline{Y}}} = \frac{1}{m} \sum_{i=1}^{m} \widehat{\overline{Y}}_i$$

donde $\widehat{\overline{Y}}_i$ es la media del i-ésimo conglomerado.

La varianza es

$$\widehat{V}\left(\widehat{\overline{Y}}_{pps}\right) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(\widehat{\overline{Y}}_i - \widehat{\overline{Y}}_{pps}\right)^2$$

Y el estimador del total es

$$\widehat{Y}_{pps} = \frac{M}{m} \sum_{i=1}^{m} \widehat{\overline{Y}}_i$$

Y su varianza es

$$\widehat{V}\left(\widehat{Y}_{pps}\right) = \frac{M^2}{m(m-1)} \sum_{i=1}^{m} \left(\widehat{\overline{Y}}_i - \widehat{\overline{Y}}_{pps}\right)^2$$

Estos estimadores son el resultado de la elección de los pesos (N_i/N), sin necesidad de la corrección de población finita, debido al muestreo con reemplazamiento.

Muestreo de conglomerados combinado con estratificación

El muestreo por conglomerados se puede combinar con el muestreo estratificado, en el sentido de que la población se puede dividir en L estratos y luego se puede seleccionar una muestra de conglomerados de cada estrato.

La estimación de la media por el método de conglomerados tiene la forma de un estimador de razón y se puede considerar como la razón entre un estimador del promedio de totales de conglomerado y un estimador del tamaño promedio de conglomerado.

Pensando en términos de estimadores de razón, hay dos formas de calcular el estimador de una media poblacional entre estratos: el estimador separado y el estimador combinado.

Si se emplea el estimador separado, se debe conocer el número total de elementos en cada estrato para poder asignar los pesos de estrato adecuados. Pero debido a que estas cantidades generalmente se desconocen, se suele considerar la forma combinada del estimador de razón en el contexto del muestreo por conglomerados.

Ejemplo

Se toman los datos del ejemplo anterior que constituyen la muestra del estrato 1, donde $M_1 = 415$ y $m_1 = 25$.

Conglomerado	N_i	Y_i	**Conglomerado**	N_i	Y_i
1	8	96000	14	10	49000
2	12	121000	15	9	53000
3	4	42000	16	3	50000
4	5	65000	17	6	32000
5	6	52000	18	5	22000
6	6	40000	19	5	45000
7	7	75000	20	4	37000
8	5	65000	21	6	51000
9	8	45000	22	8	30000
10	3	50000	23	7	39000
11	2	85000	24	3	47000
12	6	43000	25	8	41000
13	5	54000		$\sum_i N_i = 151$	$\sum_i Y_i = 1329000$

Se coge también una ciudad cercana más pequeña que se considera como el estrato 2. En el estrato 2 se toman $m_2 = 10$ conglomerados que se muestrean de $M_2 = 168$.

Cluster	N_i	Y_i
1	2	18000
2	5	52000
3	7	68000
4	4	36000
5	3	45000
6	8	96000
7	6	64000
8	10	115000
9	3	41000
10	1	12000

Hay que estimar el ingreso per cápita promedio en las dos ciudades combinadas y el error de estimación, dados los datos adicionales de la tabla anterior.

Los totales promedio por conglomerado son $\widehat{Y}_1 = \frac{1329000}{25} = 53160$ y $\widehat{Y}_2 = \frac{547000}{10} = 54700$.

Los tamaños promedio de los conglomerados son $\widehat{\overline{N}}_1 = 6{,}04$ y $\widehat{\overline{N}}_2 = 4{,}90$.

La estimación de total promedio por conglomerado es

$$\frac{1}{M}\left(M_1\widehat{Y}_1 + M_2\widehat{Y}_2\right)$$

y el estimador del promedio del tamaño de conglomerado es

$$\frac{1}{M}\left(M_1\widehat{\overline{N}}_1 + M_2\widehat{\overline{N}}_2\right)$$

Un estimador de la media por individuo es entonces

$$\widehat{\overline{\overline{Y}}} = \frac{M_1\widehat{Y}_1 + M_2\widehat{Y}_2}{M_1\widehat{\overline{N}}_1 + M_2\widehat{\overline{N}}_2}$$

y esta ecuación tiene la forma de una estimación de razón combinada. De manera análoga a la varianza utilizada en la parte de razón combinada se puede estimar mediante

$$\widehat{V}(\widehat{\overline{\overline{Y}}}) = \frac{1}{N^2}\left(M_1^2\left(1-\frac{m_1}{M_1}\right)\frac{\widehat{S}^2_{C_1}}{m_1} + M_2^2\left(1-\frac{m_2}{M_2}\right)\frac{\widehat{S}^2_{C_2}}{m_2}\right)$$

donde N es el número total de elementos de la población que se podría estimar mediante $M_1\widehat{\overline{N}}_1 + M_2\widehat{\overline{N}}_2$ si N no es conocida.

La primera varianza $\widehat{S}^2_{C_1}$ es la varianza de los términos $\left(Y_i - N_i\widehat{\overline{\overline{Y}}}\right)$ del estrato 1. La segunda varianza $\widehat{S}^2_{C_2}$ es la misma correspondiente al segundo estrato.

De este modo,

$$\widehat{\overline{\overline{Y}}} = \frac{415\cdot 53160 + 168\cdot 54700}{415\cdot 6{,}04 + 168\cdot 4{,}90} = 9385$$

Para el estrato 1

$$\widehat{S}^2_{C_1} = 25998^2$$

y para el estrato 2

$$\widehat{S}^2_{C_2} = 8657^2$$

Como la estimación de N no conocida es:

$$M_1\widehat{\overline{N}}_1 + M_2\widehat{\overline{N}}_2 = 3328{,}8$$

entonces

$$\widehat{V}(\widehat{\overline{\overline{Y}}}) = 412563{,}8$$

De modo que el intervalo de confianza es

$$9385 \mp 1{,}96\sqrt{412563{,}8} = [8126{,}1; 10644]$$

Se obtiene que el intervalo es ligeramente menor que el obtenido para el estrato 1 aislado como se vio en el ejemplo previo.

Ejemplo en R

```
cluster.mu = function(M, m.vec, y, total=T, N=NA) {
# M = número de conglomerados en la población
# N = número de elementos en la población
# m.vec = vector de los tamaños de los conglomerados en la muestra
# y = Un vector de totales por conglomerado, o una lista de las
# observaciones por conglomerado (esto se fija por 'total')

m = length(m.vec)

# Si N es deconocido, m.bar es estimado con la media de m.vec
if(is.na(N)) {mbar = mean(m.vec)}
else {mbar = N/M}

# Si no hay totales de observaciones entonces se calculan
if(total==F) {y = unlist(lapply(y,sum))}

# Denominas mu.hat al R.hat de la formula
mu.hat = sum(y)/sum(m.vec)
s2.c = sum((y-(mu.hat*m.vec))^2)/(m-1)
var.mu.hat = ((M-m)/(M*m*mbar^2))*s2.c

B = 2*sqrt(var.mu.hat)
cbind(mu.hat,s2.c,var.mu.hat,B)
}
```

Se toma un ejemplo:

```
m = c(55,60,63,58,71,78,69,58,52,71,73,64,69,58,63,75,78,51,67,70)

y = c(2210,2390,2430,2380,2760,3110,2780, 2370,1990,2810,2930,
      2470,2830,2370,2390,2870,3210,2430,2730,2880)

cluster.mu(60, m.vec=m, y, total=T, N=NA)
```

```
   mu.hat     s2.c var.mu.hat         B
 40.16884 13057.14  0.1025411 0.6404408
```

Ejemplo con la librería `survey`

En California se requiere que todos los estudiantes de los centros públicos hagan un examen cada año. Así, el *Departamento de Educación* del Estado elabora un índice de rendimiento académico *API* anual, que califica el desempeño general de una escuela en términos de puntuaciones en los exámenes.

Los datos contienen las calificaciones *API* e información demográfica sobre 6194 escuelas en 757 distritos escolares. En este ejemplo, se utilizan los distritos escolares como los conglomerados PSU (las unidades de muestreo primarias).

Tomamos al azar una muestra de 189 distritos escolares y se registran los datos de todas las escuelas seleccionadas.

En este caso, se tienen los identificadores:

- `sname`: School name.
- `snum`: School number.
- `dname`: District name.
- `dnum`: District number.

```r
# Importas los datos desde un fichero tipo Stata
library(foreign)
cosa = read.dta("C:/.../Eje1Cluster.dta")

library(survey)
dclus1 = svydesign(ids=~dnum, weight=~pw, data=cosa, fpc=~fpc)
summary(dclus1)
```

```
1 - level Cluster Sampling design
With (189) clusters.
svydesign(ids = ~dnum, weight = ~pw, data = cosa, fpc = ~fpc)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2497  0.2497  0.2497  0.2497  0.2497  0.2497
Population size (PSUs): 757
Data variables:
 [1] "cds"      "stype"    "name"     "sname"    "snum"     "dname"
 [7] "dnum"     "cname"    "cnum"     "flag"     "pcttest"  "api00"
[13] "api99"    "target"   "growth"   "sch_wide" "comp_imp" "both"
[19] "awards"   "meals"    "ell"      "yr_rnd"   "mobility" "acs_k3"
[25] "acs_46"   "acs_core" "pct_resp" "not_hsg"  "hsg"      "some_col"
[31] "col_grad" "grad_sch" "avg_ed"   "full"     "emer"     "enroll"
[37] "api_stu"  "_merge"   "pw"       "fpc"
```

```r
svymean(~api00, dclus1)
svyquantile(~api00, dclus1, quantile=c(0.25,0.5,0.75), ci=TRUE)
svytotal(~stype, dclus1)
```

```
        mean     SE
api00 670.52 11.097

#.................................

$quantiles
        0.25 0.5   0.75
api00 572.75 670 767.25

$CIs
, , api00

       0.25 0.5 0.75
(lower  538 647  737
upper)  607 693  798

#.................................

         total      SE
stypeE 4145.48 430.359
stypeH  765.01  82.319
stypeM  949.25 105.042
```

```
# Calculamos la proporción de estudiantes de inglés y la proporción
# de estudiantes seleccionados para comidas subvencionadas en
# las escuelas primarias, intermedias y secundarias

svyby(~ell+meals, ~stype, design=dclus1, svymean)
```

```
  stype      ell    meals   se.ell se.meals
E     E 23.95749 50.82899 1.868107 2.886000
H     H 12.98429 25.94241 1.235414 2.017627
M     M 18.14768 41.56962 1.761447 3.065678
```

```
# Los modelos de regresión muestran que estas variables socio-económicas
# predicen la valoración de API

regmodel = svyglm(api00 ~ ell + meals, design=dclus1)
summary(regmodel)
```

```
Call:
svyglm(formula = api00 ~ ell + meals, design = dclus1)

Survey design:
svydesign(ids = ~dnum, weight = ~pw, data = cosa, fpc = ~fpc)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 829.7578     6.9248 119.824  < 2e-16 ***
ell          -0.6693     0.1693  -3.953  0.00011 ***
meals        -3.1422     0.1542 -20.376  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 4898.252)
```

Muestreo por conglomerados bietápico

El muestreo por conglomerados en dos etapas es una extensión del concepto de muestreo por conglomerados.

Un conglomerado a menudo contiene demasiados elementos para obtener una medición en cada uno, o contiene elementos tan casi iguales que la medición de sólo unos pocos elementos proporciona información sobre un conglomerado completo.

Cuando ocurre cualquiera de las situaciones, se puede seleccionar una muestra probabilística de conglomerados y luego tomar otra muestra probabilística de elementos dentro de cada conglomerado. El resultado es una muestra de conglomerados de dos etapas.

Definición: Una muestra de conglomerados de dos etapas se obtiene seleccionando primero una muestra probabilística de conglomerados y luego seleccionando una muestra probabilística de elementos de cada conglomerado.

En este tema nos limitamos a muestras aleatorias simples en cada etapa. Por ejemplo, se puede realizar una encuesta nacional de opiniones de estudiantes universitarios seleccionando una muestra aleatoria simple de universidades de todas las del país y luego seleccionando una muestra aleatoria simple de estudiantes de cada universidad. Así, cada universidad corresponde a un conglomerado de estudiantes.

De manera similar, la cantidad total de cuentas por cobrar de una cadena de tiendas se puede estimar tomando primero una muestra aleatoria simple de tiendas y luego seleccionando una muestra aleatoria simple de cuentas de cada una. Por lo tanto, cada cadena de tiendas proporciona un conglomerado de cuentas.

El muestreo por conglomerados en dos etapas se usa comúnmente en encuestas grandes que involucran el muestreo de unidades de vivienda.

Las ventajas del muestreo por conglomerados en dos etapas sobre otros diseños son las mismas que las del muestreo por conglomerados en una etapa. Primero, un marco que enumere todos los elementos de la población puede ser imposible o costoso de obtener, mientras que obtener una lista de todos los conglomerados puede ser fácil. Por ejemplo, compilar una lista de todos los estudiantes universitarios del país sería costoso y requeriría mucho tiempo, pero se puede obtener fácilmente una lista de universidades.

En segundo lugar, el costo de obtener datos puede aumentar con los costos de viaje si los elementos muestreados se distribuyen en un área geográfica extensa. Por lo tanto, el muestreo de grupos de elementos que están físicamente juntos es a menudo más económico.

El primer problema al seleccionar una muestra de conglomerados de dos etapas es la elección de los conglomerados adecuados. Son deseables dos condiciones:

– Proximidad geográfica de los elementos dentro de un grupo.

– Tamaños de conglomerados que sean adecuados para administrar.

La selección de los conglomerados apropiados también depende de si queremos muestrear algunos conglomerados y muchos elementos de cada uno, o muchos conglomerados y algunos elementos de cada uno.

En última instancia, la elección se basa en los costos. Los conglomerados grandes tienden a poseer elementos heterogéneos y, por lo tanto, se requiere una muestra grande de cada uno para adquirir estimaciones precisas de los parámetros de la población.

Por el contrario, los conglomerados pequeños suelen contener elementos relativamente homogéneos, en cuyo caso se puede obtener información precisa sobre las características de un conglomerado seleccionando una pequeña muestra de cada conglomerado.

Se puede considerar el problema de tomar muestras de los ingresos por persona en una gran ciudad. La ciudad se puede dividir en grandes conglomerados, por ejemplo distritos, que contienen una variedad heterogénea de ingresos. Por lo tanto, una pequeña cantidad de distritos puede producir una sección representativa de los ingresos dentro de la ciudad, pero se requerirá una muestra bastante grande de elementos de cada conglomerado para estimar con precisión su media (debido a la heterogeneidad de los ingresos dentro del grupo).

Por el contrario, la ciudad se puede dividir en grupos pequeños y relativamente homogéneos, por ejemplo, manzanas. Por tanto, una pequeña muestra de personas de cada bloque proporcionará información adecuada sobre la media de cada conglomerado, pero para obtener información precisa sobre el ingreso medio de toda la ciudad se necesitarán muchos bloques.

Como otro ejemplo, se puede considerar la encuesta de opinión de estudiantes universitarios. Si los estudiantes de una universidad tienen opiniones similares sobre la cuestión de interés, pero las opiniones difieren mucho de una universidad a otra, entonces la muestra debe contener algunos representantes de muchas universidades diferentes. Pero si las opiniones varían mucho dentro de cada universidad, entonces la encuesta debe incluir a muchos representantes de cada una de unas pocas universidades.

Para seleccionar la muestra, primero obtenemos un marco que enumera todos los conglomerados de la población. Luego, extraemos una muestra aleatoria simple de conglomerados, utilizando los procedimientos de muestreo aleatorio. En tercer lugar, obtenemos marcos que enumeran todos los elementos en cada uno de los conglomerados muestreados. Finalmente, elegimos una muestra aleatoria simple de elementos de cada uno de estos marcos.

Así, en resumen:

- En el método de conglomerados bietápico se obtiene la muestra de elementos como resultado de dos etapas de muestreo.
- Los elementos de la población se agrupan en subpoblaciones disjuntas, llamadas *unidades primarias de muestreo* (PSU). Así, en una primera etapa de muestreo se extrae una muestra de PSU.
- En la segunda etapa, se toman las *unidades secundarias de muestreo* (SSU) que pueden ser conglomerados de elementos, para cada PSU de la muestra de la primera etapa.
- Se toma una muestra de SSU (segunda etapa de muestreo) de cada PSU de la muestra de la primera etapa. Cuando los SSU son conglomerados, todos los elementos de la SSU son muestreados.

Ejemplo de muestreo por conglomerados

El comité sueco de *Calidad en la Educación* toma muestras anuales para medir el consumo de drogas entre los adolescentes.

Los datos se recogen a través de cuestionarios anónimos de los estudiantes de *noveno grado*.

Es un ejemplo de muestreo bietápico, donde los colegios son PSU y las clases son SSU:

(i) Se toma una muestra de colegios de un marco que recoge todos los colegios del país.

(ii) Se toma una muestra de cada colegio seleccionado, de clases de noveno grado y se seleccionan todos los estudiantes de las clases que se han tomado.

Fórmulas aplicadas en muestreo de conglomerados bietápico

Se usa la siguiente **notación**:

- $M \equiv$ el número de conglomerados de la población.
- $m \equiv$ el número de conglomerados *seleccionados* en una muestra aleatoria simple.
- $N_i \equiv$ el número de elementos en el conglomerado i.
- $n_i \equiv$ el número de elementos seleccionados en una muestra aleatoria simple del conglomerado i.
- $N = \sum_{i=1}^{M} N_i \equiv$ el número de elementos de la población.

- $\overline{N} = \dfrac{N}{M} \equiv$ el tamaño medio del conglomerado de la población.
- $Y_{ij} \equiv$ la j-ésima observación de la muestra del i-ésimo conglomerado.
- $\widehat{\overline{Y}}_i = \dfrac{1}{n_i} \sum\limits_{j=1}^{n_i} Y_{ij} \equiv$ la media muestral del i-ésimo conglomerado.

El **estimador de la media poblacional** $\overline{\overline{Y}}$ es

$$\widehat{\overline{\overline{Y}}} = \left(\frac{M}{N}\right) \frac{\sum\limits_{i=1}^{m} N_i \widehat{\overline{Y}}_i}{m}$$

asumiendo un muestreo aleatorio simple en cada etapa.

La **varianza estimada** de $\widehat{\overline{\overline{Y}}}$ es

$$\widehat{V}(\widehat{\overline{\overline{Y}}}) = \left(\frac{M-m}{M}\right) \frac{1}{m\overline{N}^2} \widehat{S}_b^2 + \frac{1}{mM\overline{N}^2} \sum_{i=1}^{m} N_i^2 \left(\frac{N_i - n_i}{N_i}\right) \frac{\widehat{S}_i^2}{n_i}$$

donde

$$\begin{aligned} \widehat{S}_b^2 &= \frac{\sum\limits_{i=1}^{m} \left(N_i \widehat{\overline{Y}}_i - \overline{N} \cdot \widehat{\overline{\overline{Y}}}\right)^2}{m-1} \\ \widehat{S}_i^2 &= \frac{\sum\limits_{j=1}^{n_i} \left(Y_{ij} - \widehat{\overline{Y}}_i\right)^2}{n_i - 1}, \qquad i = 1, 2, \ldots, m \end{aligned}$$

Para estimar el **total de la población** $\widehat{Y}$ se obtiene de manera trivial que

$$\widehat{Y} = N\widehat{\overline{\overline{Y}}} = \frac{M}{m} \sum_{i=1}^{m} N_i \widehat{\overline{Y}}_i$$

de modo que la estimación de la varianza es:

$$\widehat{V}(\widehat{Y}) = \left(1 - \frac{m}{M}\right) \frac{M^2}{m} \widehat{S}_b^2 + \frac{M}{m} \sum_{i=1}^{m} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{\widehat{S}_i^2}{n_i}$$

Cuando el número total de elementos de la población N **no es conocido** se estima mediante

$$\widehat{N} = M \frac{\sum\limits_{i=1}^{m} N_i}{m}$$

Es decir, el tamaño medio del conglomerado multiplicado por el número de conglomerados de la población.

En este caso, el estimador de la media $\overline{\overline{Y}}$ de la población es

$$\widehat{\overline{\overline{Y}}}_R = \frac{\sum_{i=1}^{m} N_i \widehat{\overline{Y}}_i}{\sum_{i=1}^{m} N_i}$$

La varianza estimada de $\widehat{\overline{\overline{Y}}}$ es la misma que antes, pero sustituyendo $\widehat{S}_b^2$ por $\widehat{S}_R^2$

$$\widehat{S}_R^2 = \frac{\sum_{i=1}^{m} N_i^2 \left(\widehat{\overline{Y}}_i - \widehat{\overline{\overline{Y}}}_R \right)^2}{m-1}$$

Ejemplo (Elementary Survey Sampling (7ed) - Scheaffer et al.)

Un fabricante de prendas de vestir tiene 90 plantas ubicadas en todo el país y desea estimar el número promedio de horas que las máquinas de coser estuvieron inactivas por reparaciones en los últimos meses.

Debido a que las plantas están muy dispersas, decide utilizar el muestreo por conglomerados, especificando cada planta como un conglomerado de máquinas. Cada planta contiene muchas máquinas y comprobar el registro de reparación de cada máquina llevaría mucho tiempo.

Por lo tanto, utiliza un muestreo en dos etapas. Se dispone de suficiente tiempo y dinero para recoger 10 plantas y aproximadamente el 20 % de las máquinas en cada planta. Hay que calcular el tiempo de inactividad promedio por máquina y el intervalo de confianza correspondiente al 95 %. Se obtienen los datos de la siguiente tabla y se sabe que hay un total de $N = 4500$ máquinas en todas las plantas.

Plantas	N_i	n_i	**Tiempo Inactivas**	$\widehat{\overline{Y}}_i$	$\widehat{S}_i^2$
1	50	10	5, 7, 9, 0, 11, 2, 8, 4, 3, 5	5.4	11.38
2	65	13	4, 3, 7, 2, 11, 0, 1, 9, 4, 3, 2, 1, 5	4.0	10.67
3	45	9	5, 6, 4, 11, 12, 0, 1, 8, 4	5.67	16.75
4	48	10	6, 4, 0, 1, 0, 9, 8, 4, 6, 10	4.8	13.29
5	52	10	11, 4, 3, 1, 0, 2, 8, 6, 5, 3	4.3	11.12
6	58	12	12, 11, 3, 4, 2, 0, 0, 1, 4, 3, 2, 4	3.83	14,88
7	42	8	3, 7, 6, 7, 8, 4, 3, 2	5.0	5.14
8	66	13	3, 6, 4, 3, 2, 2, 8, 4, 0, 4, 5, 6, 3	3.85	4.31
9	40	8	6, 4, 7, 3, 9, 1, 4, 5	4.88	6.13
10	56	11	6, 7, 5, 10, 11, 2, 1, 4, 0, 5, 4	5.0	11.8

Solución:

El estimador de la media poblacional $\overline{\overline{Y}}$ es

$$\widehat{\overline{\overline{Y}}} = \left(\frac{M}{N}\right) \frac{\sum_{i=1}^{m} N_i \widehat{\overline{Y}}_i}{m} = \frac{1}{\overline{N}} \frac{1}{m} \sum_{i=1}^{m} N_i \widehat{\overline{Y}}_i = \frac{1}{50} 240{,}02 = 4{,}80$$

Para calcular la varianza de $\widehat{\overline{\overline{Y}}}$ se calcula primero

$$\widehat{V}(\widehat{\overline{\overline{Y}}}) = \left(1 - \frac{m}{M}\right) \frac{1}{m\overline{N}^2} \widehat{S}_b^2 + \frac{1}{mM\overline{N}^2} \sum_{i=1}^{m} N_i^2 \left(\frac{N_i - n_i}{N_i}\right) \frac{\widehat{S}_i^2}{n_i}$$

donde

$$\widehat{S}_b^2 = \frac{\sum_{i=1}^{m} \left(N_i \widehat{\overline{Y}}_i - \overline{N} \cdot \widehat{\overline{\overline{Y}}}\right)^2}{m-1} = 27{,}72^2$$

$$\sum_{i=1}^{m} N_i^2 \left(\frac{N_i - n_i}{N_i}\right) \frac{\widehat{S}_i^2}{n_i} = 21{,}985$$

Con lo que

$$\widehat{V}(\widehat{\overline{\overline{Y}}}) = \left(1 - \frac{10}{90}\right) \frac{1}{10 \cdot 50^2} 27{,}72^2 + \frac{1}{10 \cdot 90 \cdot 50^2} 21{,}985 = 0{,}0371$$

De este modo, un intervalo de confianza al 95 % es:

$$\widehat{\overline{\overline{Y}}} \mp 1{,}96 \sqrt{\widehat{V}(\widehat{\overline{\overline{Y}}})} = 4{,}80 \mp 1{,}96 \sqrt{0{,}0371} = 4{,}80 \mp 0{,}39$$

El estimador del **total** es:

$$\widehat{Y} = N \widehat{\overline{\overline{Y}}} = 90 \cdot 240{,}02 = 21{,}602$$

y la varianza del total es:

$$\widehat{V}(\widehat{Y}) = N^2 \widehat{V}(\widehat{\overline{\overline{Y}}}) = 4500^2 \cdot 0{,}0371 = 7{,}5128 \times 10^5$$

Y el intervalo de confianza es:

$$\widehat{Y} \mp 1{,}96 \cdot \sqrt{\widehat{V}(\widehat{Y})} = 21{,}602 \mp 1{,}96 \cdot \sqrt{7{,}5128 \times 10^5} = 21{,}602 \mp 1698{,}9$$

Supongamos ahora que el número total de elementos de la población N **no es conocido**.

En este caso, el estimador de la media $\overline{\overline{Y}}$ de la población es

$$\widehat{\overline{\overline{Y}}}_R = \frac{\sum_{i=1}^{m} N_i \widehat{\overline{Y}}_i}{\sum_{i=1}^{m} N_i} = \frac{\frac{1}{m}\sum_{i=1}^{m} N_i \widehat{\overline{Y}}_i}{\frac{1}{m}\sum_{i=1}^{m} N_i} = \frac{240{,}02}{52{,}2} = 4{,}60$$

En este caso en la varianza estimada de $\widehat{\overline{\overline{Y}}}$ que es la misma que antes, hay que sustituir $\widehat{S}_b^2$ por $\widehat{S}_R^2$:

$$\widehat{S}_R^2 = \frac{\sum_{i=1}^{m} N_i^2 \left(\widehat{\overline{Y}}_i - \widehat{\overline{\overline{Y}}}_R\right)^2}{m-1} = 35{,}1^2$$

Se puede estimar $\overline{N}$ mediante el tamaño medio de los conglomerados de la muestra:

$$\overline{N} = \frac{\sum_{i=1}^{m} N_i}{m} = \frac{522}{10} = 52{,}2$$

Y así

$$\widehat{V}(\widehat{\overline{\overline{Y}}}) = \left(1 - \frac{10}{90}\right)\frac{1}{10 \cdot 52{,}2^2} 35{,}1^2 + \frac{1}{10 \cdot 90 \cdot 52{,}2^2} 21{,}985 = 0{,}0492$$

De este modo, un intervalo de confianza al 95 % es:

$$\widehat{\overline{\overline{Y}}} \mp 1{,}96\sqrt{\widehat{V}(\widehat{\overline{\overline{Y}}})} = 4{,}60 \mp 1{,}96\sqrt{0{,}0492} = 4{,}60 \mp 0{,}44$$

Ejemplo

El gasto en ocio es importante para los negocios de una ciudad con una residencia para estudiantes. Una empresa está interesada en estimar el promedio mensual gastado en ocio por estudiante. El procedimiento de muestreo que utilizan es tomar habitaciones seleccionadas al azar y muestrear un subconjunto de estudiantes en esas habitaciones. El número de habitaciones es de 112 con un total de 306 estudiantes. Los datos se muestran en la tabla:

M_i	3	2	4	3	6	3	5	4	6	3
m_i	2	2	2	2	2	2	2	2	3	2
y_i	23, 27	18, 3	37, 12	22, 21	18, 62	21, 25	3, 12	7, 36	2, 12, 51	18, 17

Se considera una función en R para hacer un muestreo por conglomerados bietápico.

```
dos.etapas.conglomerado = function(M, m, y, N=NA, Ni, ni) {

 # M = número de conglomerados de la población
 # m = número de conglomerados muestreados
 # y = una lista de observaciones de los conglomerados muestreados
 # N = número de elementos de la población, a menudo desconocido
```

```
# Ni = número de elementos en cada conglomerado seleccionado
# ni = número de elementos muestrado en cada conglomerado
#      seleccionado

  yi.bar = unlist(lapply(y,mean))
  s2i    = unlist(lapply(y,var))

    if(is.na(N)) {   # Caso donde N es desconocido
     N.bar = mean(Ni)
     mu.hat = sum(Ni*yi.bar)/(m*N.bar)
     s2.r = sum((Ni^2)*(yi.bar-mu.hat)^2)/(m-1)
     cosa.1 = ((M-m)/M)*(1/(m*N.bar^2))*s2.r
     cosa.2 = (1/(m*M*(N.bar^2)))*sum(Ni^2*((Ni-ni)/Ni)*(s2i/ni))

     var.mu = cosa.1 + cosa.2
     B = 2*sqrt(var.mu)
     cat(" mu.hat =",mu.hat,"\n", " var.mu =",var.mu,"\n",
      " B =",B,"\n", " yi.bar =",yi.bar,"\n", " s2i =",s2i,"\n",
      " s2.r =",s2.r,"\n")

   } else {   # Caso donde N es conocido
     N.bar = N/M
     mu.hat = (1/N.bar)*(sum(Ni*yi.bar)/m)
     s2.b = sum((Ni*yi.bar-N.bar*mu.hat)^2)/(m-1)
     cosa.1 = ((M-m)/M)*(1/(m*N.bar^2))*s2.b
     cosa.2 = (1/(M*m*(N.bar^2)))*sum(Ni^2*((Ni-ni)/Ni)*(s2i/ni))

     var.mu = cosa.1 + cosa.2
           B = 2*sqrt(var.mu)
      cat("  mu.hat =",mu.hat,"\n", " var.mu =",var.mu,"\n","
      B =",B,"\n", " yi.bar =",yi.bar,"\n", " s2i =",s2i,"\n"
      ," s2.b =",s2.b,"\n")
     }
}
```

```r
# EJEMPLO
M = 112
m = 10
N = 306

Ni = c(3,2,4,3,6,3,5,4,6,3)
ni = c(2,2,2,2,2,2,2,2,3,2)

y = list(y1=c(23,27),y2=c(18,3),y3=c(37,12),y4=c(22,21),y5=c(18,62),
         y6=c(21,25),y7=c(3,12),y8=c(7,36),y9=c(2,12,51),y10=c(18,17))

dos.etapas.conglomerado(M,m,y,N,Ni,ni)
```

```
mu.hat = 31.97124
var.mu = 48.82493

     B = 13.97497
yi.bar = 25 10.5 24.5 21.5 40 23 7.5 21.5 21.66667 17.5
s2i = 8 112.5 312.5 0.5 968 8 40.5 420.5 670.3333 0.5
s2.b = 3816.614
```

EPA

En la práctica se utilizan muchas veces diseños muestrales mixtos, como el muestreo estratificado de conglomerados en r-etapas.

El criterio que se sigue para definir los estratos es, o bien la situación geográfica, o bien el tamaño de los conglomerados, o bien una variable auxiliar X relacionada con Y.

El ejemplo típico es la elaboración de la *EPA*. Esta tiene como objetivo principal el conocimiento de la actividad económica en España en lo relativo a su componente humana. Pretende dar información de las principales categorías laborales en relación al mercado de trabajo.

La encuesta se realiza a las personas de todo el territorio nacional que residen en viviendas familiares principales, es decir, las utilizadas todo el año como vivienda habitual. Las unidades de análisis son las personas de las viviendas. Se excluyen del análisis los llamados *hogares colectivos* (hospitales, cuarteles, hoteles, centros religiosos).

Teniendo en cuenta la población a la que va dirigida la encuesta, es necesario disponer de un listado de todas las viviendas familiares existentes en el territorio nacional. Esto es difícil de conseguir, ya que por una parte esta relación viene afectada por el paso del tiempo, y por otra hay errores tales como omisiones, duplicidades, etc.

Como consecuencia de esa situación, se seleccionan, en lugar de viviendas, áreas geográficas que son más estables en el tiempo.

España se encuentra dividida en 17 Comunidades Autónomas (C.A.) y a su vez en 50 provincias de las cuales 47 son peninsulares y 3 insulares, además de las ciudades de Ceuta y Melilla. Las provincias se encuentran divididas en municipios y estas en distritos municipales. El INE, junto con los ayuntamientos, divide los distritos en secciones censales.

Las secciones censales son áreas geográficas con límites perfectamente definidos, cuyo tamaño de población viene determinado por las secciones electorales (un máximo de 2000 electores y un mínimo de 500). En cada sección censal se confecciona una lista de viviendas familiares con sus direcciones postales obtenidas del último Censo o Padrón Municipal. Como las secciones censales varían con el tiempo, con cada Censo o Padrón al inicio de cada año se realiza una actualización de las mismas y si es necesario se dividen o se fusionan las secciones censales.

En consecuencia, el marco poblacional esta determinado por las 32000 secciones censales en que se encuentra dividido el territorio nacional y la relación de viviendas familiares que se conforma para cada sección censal seleccionada para la muestra

Así, en esquema, la división administrativa oficial en España es:

Comunidades Autónomas $\Rightarrow$ Provincias $\Rightarrow$ Municipios $\Rightarrow$

Distritos Municipales ⇒ Secciones Censales ⇒ Viviendas ⇒ Personas.

Las **unidades primarias** son la secciones censales

Las **unidades secundarias** son las viviendas.

En las viviendas seleccionadas se pregunta a todas las personas. El muestreo utilizado es el muestreo bietápico con estratificación de las unidades de primera etapa. Toda la información necesaria para realizar la estratificación se obtiene de los Censos de Población.

Las secciones censales (conglomerados) se estratifican según un doble criterio: geográfico y socio-económico. Se hace una primera estratificación de las secciones, de acuerdo a la provincia y tipo de municipio (según su número de habitantes) a las que pertenecen.

Dentro de cada uno de los estratos anteriormente formados, las secciones censales se agrupan en subestratos según la categoría socio-económica de los hogares ubicados en la sección.

Ejemplo con la librería `survey`

Se considera un muestreo estratificado bietápico con un ejemplo de centros de enseñanza en California.

Primero se estratifican los institutos basándonos en su media de `api99score`. Se hace un muestreo en cada estrato de manera independiente, tomando una muestra de los distritos (*conglomerados*), y luego se toma una muestra aleatoria simple de institutos (*elementos*).

En concreto, se selecciona aleatoriamente el 25 % de distritos que hay dentro de cada estrato, y se seleccionan aleatoriamente 3 institutos de cada distrito seleccionado.

Los datos en formato `Stata`, se pueden descargar desde

```
https://stats.idre.ucla.edu/other/examples
```

Aunque parece que ahora los datos originales no están disponibles.

O bien alternativamente se pueden usar los datos, que son semejantes, de la librería `survey` que tiene el fichero de ejemplo `apiclus2`.

```
# Importas los datos desde formato Stata
library(foreign)
cosa = read.dta("C:/.../Eje2Cluster.dta")

# head(cosa)

library(survey)

# Diseño estratificado de conglomerados bietápicos
dclus2 = svydesign(ids=~dnum+~snum, strata=~strata, data=cosa,
weight=~pwt, fpc=~fpc+N)
summary(dclus2)
```

```
Stratified 2 - level Cluster Sampling design
With (189, 466) clusters.

svydesign(ids = ~dnum + ~snum, strata = ~strata, data = cosa,
    weight = ~pwt, fpc = ~fpc + N)

Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00748 0.06250 0.15000 0.15065 0.24934 0.25000

First-level Stratum Sizes:
             1   2
obs        227 239
design.PSU  94  95
actual.PSU  94  95

Population stratum sizes (PSUs):
  1   2
377 380

Data variables:
 [1] "cds"       "stype"     "name"      "sname"     "snum"      "dname"
 [7] "dnum"      "cname"     "cnum"      "flag"      "pcttest"   "api00"
[13] "api99"     "target"    "growth"    "sch_wide"  "comp_imp"  "both"
[19] "awards"    "meals"     "ell"       "yr_rnd"    "mobility"  "acs_k3"
[25] "acs_46"    "acs_core"  "pct_resp"  "not_hsg"   "hsg"       "some_col"
[31] "col_grad"  "grad_sch"  "avg_ed"    "full"      "emer"      "enroll"
[37] "api_stu"   "mean"      "strata"    "_freq"     "x"         "_merge"
[43] "n"         "xx"        "number"    "N"         "nn"        "p1"
[49] "p2"        "pwt"       "fpc"       "comp_imp1" "meals3"
```

```r
eso = svymean(~api00, dclus2)
supe = eso[[1]] + qt(0.975,187)*sqrt(attr(eso,"var"))
infe = eso[[1]] - qt(0.975,187)*sqrt(attr(eso,"var"))

print(eso)
cat("Intervalo de Confianza para api00: [",infe,";",supe,"]","\n")
```

```
        mean     SE
api00 681.84 10.855

Intervalo de Confianza para api00: [ 660.4258 ; 703.2543 ]
```

```r
# Intervalo de confianza
confint(eso)
```

```
         2.5 %   97.5 %
api00 660.5644 703.1157
```

```r
eso = svymean(~growth, dclus2)
supe = eso[[1]] + qt(0.975,187)*sqrt(attr(eso,"var"))
infe = eso[[1]] - qt(0.975,187)*sqrt(attr(eso,"var"))
print(eso)
cat("Intervalo de Confianza para growth: [",infe,";",supe,"]","\n")
```

```
         mean     SE
growth 30.718 2.4127

Intervalo de Confianza para growth: [ 25.95797 ; 35.47729 ]
```

```r
# Intervalo de confianza
confint(eso)
```

```
          2.5 %   97.5 %
growth 25.98877 35.44649
```

```r
eso = svytotal(~yr_rnd, dclus2)
supe = eso[[1]] + qt(0.975,187)*sqrt(attr(eso,"var"))
infe = eso[[1]]- qt(0.975,187)*sqrt(attr(eso,"var"))
print(eso)

cat("Intervalo de Confianza para yr_rnd: [",infe,";",supe,"]","\n")
```

```
        total     SE
yr_rnd 718.91 217.84

Intervalo de Confianza para yr_rnd: [ 289.1694 ; 1148.66 ]
```

```r
# Intervalo de confianza
confint(eso)
```

```
          2.5 %   97.5 %
yr_rnd 291.9506 1145.879
```

```r
# Calcula la proporción promedio de estudiantes de inglés y
# y estudiantes elegibles para comedor subencionado para colegios
# de los dos estratos considerados

svyby(~ell+meals, ~strata, design=dclus2, svymean, vartype="ci")
```

```
  strata      ell    meals  ci_l.ell ci_l.meals ci_u.ell ci_u.meals
1      1 26.35245 57.39307 21.773880   49.69776 30.93102   65.08838
2      2  8.59442 22.43398  6.624235   18.52256 10.56460   26.34540
```

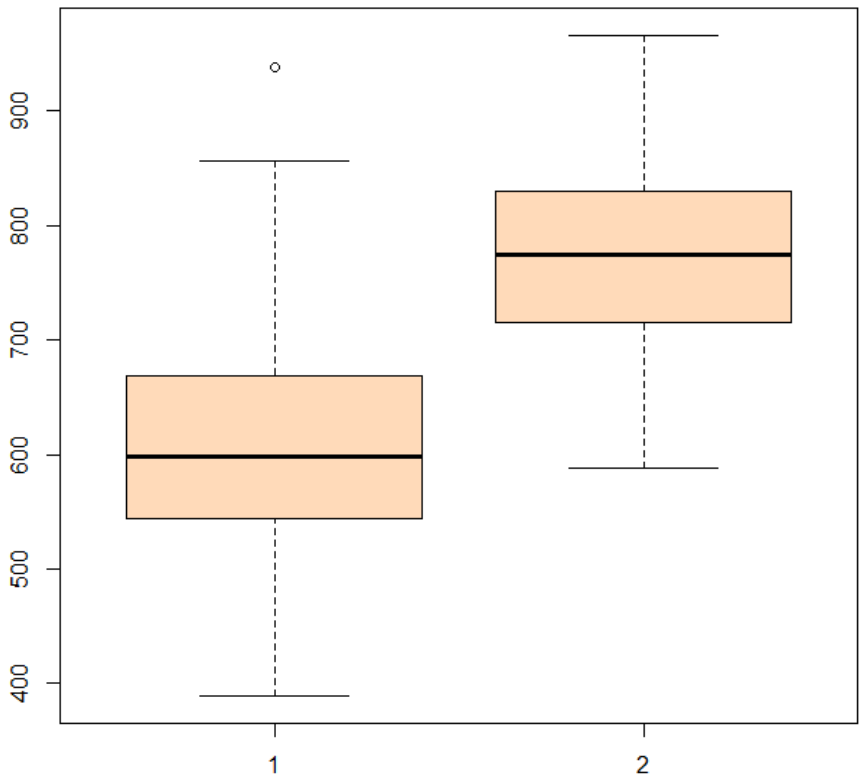
```r
# Modelos de regresión para mostrar que estas variables
# socioeconómicas variables predicen la puntuación API
# y si los colegios alacanzaron los objetivos de API

regmodel = svyglm(api00 ~ awards + meals, design=dclus2)
summary(regmodel)
```

```
Call:
svyglm(formula = api00 ~ awards + meals, design = dclus2)

Survey design:
svydesign(ids = ~dnum + ~snum, strata = ~strata, data = cosa,
    weight = ~pwt, fpc = ~fpc + N)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 772.7654    11.6604   66.27  < 2e-16 ***
awards       66.1988    10.1227    6.54 5.85e-10 ***
meals        -3.1923     0.1367  -23.34  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 4077.376)
```

```
svyboxplot(~api00 ~ factor(strata), dclus2, col="peachpuff")
```