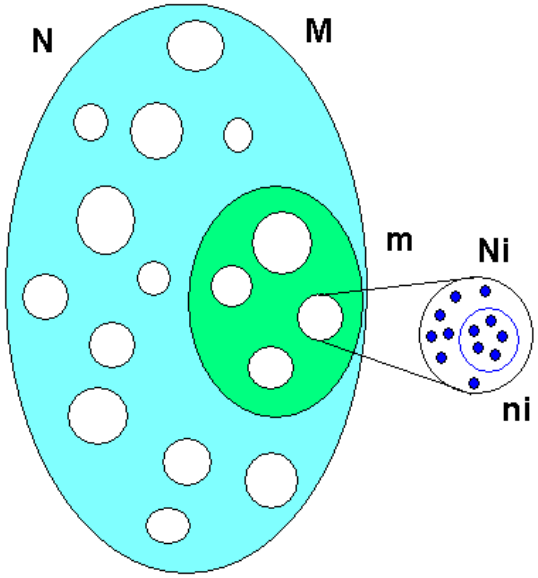
4 MUESTREO POR CONGLOMERADOS

4.1 Definición y Notación

El muestreo por conglomerados se considera como una opción de gran utilidad en situaciones en las que por limitaciones prácticas no se dispone de un marco de unidades elementales de observación o por razones económicas resulta más conveniente recolectar datos en agrupamientos naturales de la población, como lo son empresas, escuelas, hospitales, municipios, localidades, etc.

Se considera que la población se forma de **M** conglomerados como unidades de primera etapa (UPM) de las cuales se toma una muestra aleatoria simple de tamaño **m**. Cada conglomerado tiene **Ni** elementos de los cuales se toma una muestra aleatoria simple de tamaño **ni**.



Notación

M	Número de conglomerados en la población
m	Número de conglomerados en muestra
Ni	Tamaño del conglomerado i.
$N = \sum_{i=1}^{M} N_i$	Tamaño de la población
n_i	Tamaño de la muestra en conglomerado i.

$n = \sum_{i=1}^{m} n_i$ Tamaño total de muestra

y_{ij} Valor de la característica del j del conglomerado i.

$\bar{y}_i = \frac{\sum_{i=1}^{ni} Y_{ij}}{ni}$ Media muestral del conglomerado i.

$\bar{Y}_i = \frac{\sum_{i=1}^{Ni} Y_{ij}}{Ni}$ Media total del conglomerado i.

$\hat{Y}_i = N_i \, \bar{y}_i$ Estimador del total del conglomerado i.

$\bar{Y}_C = \frac{1}{M} \sum_{i=1}^{M} Y_i$ Media de totales por conglomerado en la población

$\hat{\bar{Y}}_C = \frac{1}{m} \sum_{i=1}^{m} \hat{Y}_i$ Media de totales por conglomerado en la muestra.

$S_e^2 = \frac{1}{M-1} \sum_{i=1}^{M} \left(Y_i - \bar{Y}_C\right)^2$ Cuasivarianza entre totales por conglomerados en la población.

$\hat{S}_e^2 = \frac{1}{m-1} \sum_{i=1}^{m} \left(\hat{Y}_i - \hat{\bar{Y}}_C\right)^2$ Cuasivarianza entre totales por conglomerados en la muestra

$S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{Ni} \left(Y_{ij} - \bar{Y}_i\right)^2$ Cuasivarianza de elementos dentro del conglomerado i de la población.

$\hat{S}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{ni} \left(Y_{ij} - \hat{\bar{Y}}_i\right)^2$ Cuasivarianza de elementos dentro del conglomerado i de la muestra.

4.2 Estimador del Total

El estimador se basa en el cálculo del promedio por unidad elemental dentro de cada conglomerado el cual se expande al total del conglomerado al multiplicar por Ni. A continuación se promedian estos totales para los m conglomerados en muestra y luego se expanden por el número M de conglomerados en la población para así tener una estimación del total de la variable.

$$\hat{Y} = \frac{M}{m}\sum_{i=1}^{m}\frac{Ni}{ni}\sum_{j=1}^{ni}Y_{ij}$$

En forma alternativa, cada unidad en muestra se multiplica por un factor de expansión que es igual a los recíprocos de las probabilidades de selección en cada etapa.

$$\hat{Y} = \sum_{i=1}^{m}\sum_{j=1}^{ni}\frac{M}{m}\frac{Ni}{ni}Y_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{ni}F_i Y_{ij}$$

Propiedades del Estimador

El estimador del total es un estimador insesgado.

Se prueba fácilmente al tomar esperanzas condicionales en cada etapa.

$$\begin{aligned}
E(\hat{Y}) &= E_i Ej(\hat{Y}) \\
&= E_i E_j\left[\frac{M}{m}\sum_{i=1}^{m}\frac{Ni}{ni}\sum_{j=1}^{ni}y_{ij}\right] \\
&= E_i\left[\frac{M}{m}\sum_{i=1}^{m}Ni\ E_j\left(\frac{\sum_{j=1}^{ni}y_{ij}}{ni}\right)\right] \\
&= E_i\left[\frac{M}{m}\sum_{i=1}^{m}Ni\,\bar{Y}_i\right] \\
&= E_i\left[\frac{M}{m}\sum_{i=1}^{m}Y_i\right] \\
&= M\bar{Y}_C \\
&= Y
\end{aligned}$$

La varianza del estimador del total se puede descomponer fácilmente en sus dos fuentes: la varianza entre conglomerados y la varianza dentro de conglomerados. Se suma la varianza de la muestra aleatoria simple de conglomerados en primera etapa y la varianza de las muestras aleatorias dentro de conglomerados en la segunda etapa.

$$V(\hat{Y}) = \underbrace{M^2 \frac{M-m}{M}\frac{S_e^2}{m}}_{\substack{Entre \\ Conglomerados}} + \underbrace{\frac{M}{m}\sum_{i=1}^{M} N_i^2\left(\frac{Ni-ni}{Ni}\right)\frac{S_i^2}{ni}}_{\substack{Dentro\ de \\ Conglomerados}}$$

La varianza del estimador del total está en función de las cuasivarianzas entre y dentro de conglomerados:

$$S_e^2 = \frac{1}{M-1}\sum_{i=1}^{M}\left(Y_i - \overline{Y}_C\right)^2$$

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

La siguiente fórmula permite su estimación insesgada.

$$\hat{V}(\hat{Y}) = \underbrace{M^2 \frac{M-m}{M}\frac{\hat{S}_e^2}{m}}_{\substack{Entre \\ Conglomerados}} + \underbrace{\frac{M}{m}\sum_{i=1}^{m} N_i^2\left(\frac{Ni-ni}{Ni}\right)\frac{\hat{S}_i^2}{ni}}_{\substack{Dentro\ de \\ Conglomerados}}$$

El mayor aporte a la varianza se da entre conglomerados, usualmente más del 90%. Ello sugiere que entre más conglomerados tenga la muestra, el estimador resulta más eficiente. Si se requiere incrementar la muestra, conviene más incrementar el número de conglomerados en muestra que incrementar el número de unidades elementales en muestra dentro de los conglomerados**.**

4.3 Relación del Muestreo por Conglomerados y el Muestreo Estratificado

Si el tamaño de muestra de conglomerados, (unidades de primera etapa UPM) es igual al total de conglomerados en la población, esto es m = M, lo cual equivale a un censo de UPMs, el estimador del total por conglomerados coincide con el estimador del total por muestreo estratificado.

$$\hat{Y} = \frac{M}{M}\sum_{i=1}^{M}\frac{N_i}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

$$= N\sum_{i=1}^{M}\frac{N_i}{N}\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

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

$$= N\sum_{i=1}^{M} W_i y_i$$

La varianza del estimador del total obviamente coincidirá con la forma general de la varianza para muestreo estratificado.

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

$$= M^2\left(1-\frac{M}{M}\right)\frac{\hat{S}_e^2}{m}+\frac{M}{M}\sum_{i=1}^{M}N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{\hat{S}_i^2}{n_i}$$

$$= \frac{N^2}{N^2}\sum_{i=1}^{M}N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{\hat{S}_i^2}{n_i}$$

$$= N^2\sum_{i=1}^{M}\frac{N_i^2}{N^2}\left(1-\frac{n_i}{N_i}\right)\frac{\hat{S}_i^2}{n_i}$$

$$\hat{V}(\hat{Y}) = N^2\sum_{i=1}^{M}W_i^2V(y_i)$$

Se concluye que el muestreo estratificado es equivalente a un censo de conglomerados.

Por otra parte, si todas las unidades dentro de los conglomerados en muestra son seleccionadas, entonces la contribución de la varianza se reduce a la contribución entre conglomerados.

$$\hat{V}\left(\hat{Y}\right) = M^2\,\frac{M-m}{M}\frac{\hat{S}_e^2}{m}$$

Se mencionó que la proporción de la varianza entre conglomerados es mayor que la varianza dentro de conglomerados y que por ello es más eficiente incrementar el número de conglomerados en muestra que incrementar el número de elementos dentro de conglomerados. Esto se debe a la redundancia informativa que se presenta dentro de cada conglomerado al estar formados por unidades con alto grado de homogeneidad.

Por simplicidad en los siguientes análisis, considérese el caso de tamaños de conglomerados iguales y tamaños de muestra iguales dentro de cada conglomerado.

$$N_i = \bar{N} = \frac{N}{M} \qquad n_i = \bar{n} = \frac{n}{m}$$

El estimador de la media poblacional se resulta ser equivalente al promedio simple de todas las unidades en muestra.

$$\hat{\bar{Y}} = \frac{1}{N}\frac{M}{m}\sum_{i=1}^{m}\frac{\bar{N}}{\bar{n}}\sum_{j=1}^{\bar{n}}y_{ij}$$

$$= \frac{1}{N}\frac{M}{m}\frac{\bar{N}}{\bar{n}}\sum_{i=1}^{m}\sum_{j=1}^{\bar{n}}y_{ij}$$

$$= \frac{1}{m\bar{n}}\sum_{i=1}^{m}\sum_{j=1}^{\bar{n}}y_{ij}$$

El estimador de la varianza también experimenta un cambio sustancial.

$$\hat{V}(\hat{\bar{Y}}) = \frac{1}{N^2}\left[M^2\left(1-\frac{m}{M}\right)\frac{\hat{S}_e^2}{m} + \frac{M}{m}\sum_{i=1}^{M} N^2\left(1-\frac{n}{N}\right)\frac{\hat{S}_i^2}{n}\right]$$

$$= \frac{1}{N^2}M^2\left(1-\frac{m}{M}\right)\frac{\hat{S}_e^2}{m} + \frac{1}{N^2}\frac{M}{m}\left(\frac{N^2}{M^2}\right)\left(\frac{N-n}{N}\right)\sum_{i=1}^{M}\frac{\hat{S}_i^2}{n}$$

$$= \frac{1}{N^2}M^2\left(1-\frac{m}{M}\right)\frac{\hat{S}_e^2}{m} + \frac{1}{N^2}\frac{M}{mn}\left(\frac{N^2}{M}\right)\left(\frac{N-n}{N}\right)\sum_{i=1}^{M}\frac{\hat{S}_i^2}{M}$$

$$= \frac{1}{N^2}\left(1-\frac{m}{M}\right)\frac{\hat{S}_e^2}{m} + \frac{1}{nm}\left(\frac{N-n}{N}\right)\sum_{i=1}^{M}\frac{\hat{S}_i^2}{M}$$

$$= \left(1-\frac{m}{M}\right)\frac{S_{1e}^2}{m} + \left(\frac{N-n}{N}\right)\frac{S_{2i}^2}{nm} \qquad \text{Donde} \qquad S_{2i}^2 = \sum_{i=1}^{M}\frac{\hat{S}_i^2}{M} \qquad S_{1e}^2 = \frac{S_e^2}{N^2}$$

La varianza de la media queda expresada en otros términos por:

$$V\left(\hat{\bar{Y}}\right) = \left(1-\frac{m}{M}\right)\frac{S_{1e}^2}{m} + \left(\frac{N-n}{N}\right)\frac{S_{2i}^2}{nm}$$

Pero alternativamente se puede expresar:

$$V\left(\hat{\bar{Y}}\right) = \left(1-\frac{m}{M}\right)\frac{S_{1e}^2}{m} + \left(\frac{N-n}{N}\right)\frac{1}{M}\sum_{i=1}^{M}\frac{S_i^2}{nm}$$

4.4 Coeficiente de Correlación Intraclase.

Las unidades de análisis dentro de un mismo conglomerado presentan semejanzas que son medidas a través del coeficiente de correlación calculado entre todas las parejas posibles de unidades dentro de un mismo conglomerado $\binom{N_i}{2}$=Ni(Ni-1)/2. El coeficiente de correlación intraclase mide la relación lineal entre las parejas de un mismo conglomerado, pero también se puede interpretar como el incremento en la probabilidad de que dos unidades seleccionadas al azar dentro de un mismo conglomerado tengan el mismo valor para la variable de análisis respecto de una selección no conglomerada.

$$r = \frac{E\left((Y_{ij}-\bar{Y})(Y_{ik}-\bar{Y})\right)}{E(Y_{ij}-\bar{Y})^2}$$

$$r = \frac{\sum_{i=1}^{M}\sum_{j<}^{\bar{N}-1}\sum_{k}^{\bar{N}}\left(Y_{ij}-\bar{Y}\right)\left(Y_{ik}-\bar{Y}\right)/(M\bar{N}(\bar{N}-1)/2)}{\sum_{i=1}^{M}\sum_{j=1}^{\bar{N}}(Y_{ij}-\bar{Y})^2/M\bar{N}}$$

$$= \frac{\sum_{i=1}^{M}\sum_{j<}^{\bar{N}-1}\sum_{k}^{\bar{N}}\left(Y_{ij}-\bar{Y}\right)\left(Y_{ik}-\bar{Y}\right)/(M\bar{N}(\bar{N}-1)/2)}{(M\bar{N}-1)\sum_{i=1}^{M}\sum_{j=1}^{\bar{N}}(Y_{ij}-\bar{Y})^2/(M\bar{N}(M\bar{N}-1))}$$

$$= \frac{\sum_{i=1}^{M}\sum_{j<}^{\bar{N}-1}\sum_{k}^{\bar{N}}\left(Y_{ij}-\bar{Y}\right)\left(Y_{ik}-\bar{Y}\right)/(M\bar{N}(\bar{N}-1)/2)}{(M\bar{N}-1)S^2/M\bar{N}}$$

$$r = \frac{2\sum_{i=1}^{M}\sum_{j<}^{\bar{N}-1}\sum_{k}^{\bar{N}}\left(Y_{ij}-\bar{Y}\right)\left(Y_{ik}-\bar{Y}\right)}{\left(\bar{N}-1\right)\left(M\bar{N}-1\right)S^2}$$

4.5 Relación del Muestreo por Conglomerados y el Muestreo Aleatorio Simple

Se ha argumentado que el muestreo por conglomerados resulta más económico que el muestreo aleatorio simple, pues abarata costos al tomar unidades cercanas dentro de un mismo conglomerado. El conglomerado puede ser una manzana como conglomerado de viviendas, una escuela como conglomerado de estudiantes, una fábrica como conglomerado de obreros, etc. ¿Pero cuál es el costo en eficiencia estadística?. Para dar respuesta, se Inicia con la revisión de la varianza total, la cual queda reflejada en la siguiente fórmula:

$$S^2 = \sum_{i=1}^{M}\sum_{i=1}^{N_i}\frac{\left(y_{ij}-\bar{Y}\right)^2}{N-1}$$

Esta S^2 se relaciona con S_{1e}^2 y con S_{2i}^2 de la siguiente forma:

$$S_{1e}^2 = \frac{M\bar{N}-1}{\bar{N}^2(M-1)}S^2\left[1+(\bar{N}-1)r\right] \qquad S_{2i}^2 = \frac{M\bar{N}-1}{M\bar{N}}S^2(1-r)$$

Donde r es el coeficiente de correlación intraclase.

Al sustituir estas expresiones en la fórmula de la varianza de la media se obtiene la relación de la varianza del muestreo aleatorio simple con la del muestreo por conglomerados. Se observa fácilmente que la varianza del estimador de la media por conglomerados es mayor que la del muestreo aleatorio simple y que la diferencia se incrementa con el valor del coeficiente de correlación y el tamaño medio de muestra dentro de cada conglomerado.

$$V(\hat{\bar{Y}}) = \left(1 - \frac{m}{M}\right)\frac{S_{1e}^2}{m} + \left(\frac{\bar{N} - \bar{n}}{\bar{N}}\right)\frac{S_{2i}^2}{\bar{n}m}$$

$$\hat{V}(\hat{\bar{Y}}) = \hat{V}(\hat{\bar{Y}}_{MAS})[1 + (\bar{n} - 1)r]$$

El efecto de diseño fue definido por Kish (Design Effect abreviado Deff) como el cociente de la varianza del estimador con el modelo seleccionado, entre la varianza del estimador correspondiente con muestreo aleatorio simple . Algunos autores hacen referencia a la raíz cuadrada del Deff identíficándolo como Deft y que daría como consecuencia el cociente análogo entre errores estándares.

$$Deff = \frac{\hat{V}(\hat{\bar{Y}})}{\hat{V}(\hat{\bar{Y}}_{MAS})} = [1 + (\bar{n} - 1)r]$$

De donde $V(\hat{Y}) = V(\hat{Y}_{mas})Deff$ esta relación es frecuentemente aprovechada para calcular el tamaño de muestra por conglomerados a partir del cálculo del tamaño por muestreo aleatorio simple y posteriormente multiplicándolo por el efecto de diseño.

Si se conoce el efecto de diseño y el tamaño de muestra medio por conglomerado, se puede calcular el coeficiente de correlación intraclase en forma más simple que a partir del Deff.

$$r = \frac{Deff - 1}{\bar{n} - 1}$$

4.6 Asignación de Muestra

Una vez que se ha decidido efectuar un muestreo por conglomerados se debe determinar cuántas unidades primarias de muestreo (UPM) o conglomerados (m) hay que seleccionar en una primera etapa y cuántas unidades de segunda etapa $\bar{n}$ hay que seleccionar como promedio en cada conglomerado. Para tener una menor varianza y costo.

$$V(\hat{\bar{Y}}) = \left(1 - \frac{m}{M}\right)\frac{S_{1e}^2}{m} + \left(\frac{\bar{N} - \bar{n}}{\bar{N}}\right)\frac{S_{2i}^2}{\bar{n}m}$$

Como función de costo se supondrá una que involucre los costos unitarios por acceder a una unidad de primera etapa y a una de segunda etapa multiplicados por los respectivos tamaños. El total será igual al costo variable del proyecto.

$$c_V = c_1 m + c_2 m\bar{n}$$

Se procede a definir una función que involucra la varianza y la restricción de la función de costos para minimizarla con la técnica del Multiplicador de Lagrange ya utilizada en muestreo estratificado.

$$\varphi = \left(\frac{M-m}{M} \right) \frac{S_{1e}^2}{m} + \left(\frac{\bar{N}-\bar{n}}{\bar{N}} \right) \frac{S_{2i}^2}{\bar{n}m} + \lambda \left(c_1 m + c_2 m\bar{n} - c_V \right)$$

Se procede a derivar la función respecto de m y n media e igualar a cero.

$$\frac{\partial \varphi}{\partial m} = -\frac{S_{1e}^2}{m^2} - \frac{S_{2i}^2}{m^2 \bar{n}} + \frac{S_{2i}^2}{\bar{N}m^2} + \lambda \left(c_1 + c_2 \bar{n} \right) = 0$$

$$\frac{\partial \varphi}{\partial \bar{n}} = -\frac{S_{2i}^2}{m\bar{n}^2} + \lambda c_2 m = 0$$

La primera ecuación se multiplica por m^2 y la segunda por m

$$-S_{1e}^2 - \frac{S_{2i}^2}{\bar{n}} + \frac{S_{2i}^2}{\bar{N}} + \lambda m^2 \left(c_1 + c_2 \bar{n} \right) = 0$$

$$-\frac{S_{2i}^2}{\bar{n}^2} + \lambda c_2 m^2 = 0$$

Se despeja λm^2 en la segunda ecuación y se sustituye en la primera

$$\lambda m^2 = \frac{S_{2i}^2}{c_2 \bar{n}^2}$$

$$-S_{1e}^2 - \frac{S_{2i}^2}{\bar{n}} + \frac{S_{2i}^2}{\bar{N}} + \frac{S_{2i}^2}{c_2 \bar{n}^2} \left(c_1 + c_2 \bar{n} \right) = 0$$

Se distribuye el producto y se efectúan las cancelaciones necesarias

$$-S_{1e}^2 - \frac{S_{2i}^2}{\bar{n}} + \frac{S_{2i}^2}{\bar{N}} + \frac{S_{2i}^2 c_1}{c_2 \bar{n}^2} + \frac{S_{2i}^2}{\bar{n}} = 0$$

$$-S_{1e}^2 + \frac{S_{2i}^2}{\bar{N}} + \frac{S_{2i}^2 c_1}{c_2 \bar{n}^2} = 0$$

Finalmente se despeja el tamaño medio de muestra en cada conglomerado.

$$\bar{n} = \sqrt{\frac{\frac{c_1}{c_2} S_{2i}^2}{S_{1e}^2 - \frac{S_{2i}^2}{\bar{N}}}}$$

Una forma alternativa de expresar el tamaño medio por conglomerado es mediante el coeficiente de correlación intraclase. Se observa fácilmente que a mayor costo por UPM se incrementa la muestra dentro de los conglomerados y que al aumentar el coeficiente de correlación intraclase, se tiende a disminuir el tamaño de muestra dentro de los conglomerados.

$$\bar{n} = \sqrt{\frac{c_1}{c_2} \frac{1-r}{r}} \qquad c_T = c_1 m + c_2 m \bar{n} \qquad m = \frac{c_T}{(c_1 + c_2 \bar{n})}$$

El ahorro económico en los diseños por conglomerados trae aparejada la disminución en la precisión que provoca el efecto de conglomeración. Ello da lugar para que los diseñadores mezclen el muestreo por conglomerados con otros recursos como el muestreo estratificado y el uso de variables con información adicional incorporada vía procedimientos de selección con probabilidades no homogéneas, estimadores de razón, regresión, etc.