Técnicas de Muestreo I

Patricia Isabel Romero Mares

Departamento de Probabilidad y Estadística
IIMAS UNAM

septiembre 2015

Muestreo estratificado

Muestreo estratificado

Estrato es un subconjunto de unidades muestrales de la población.

Los estratos son subconjuntos de la población que agrupan unidades. Cada estrato se muestrea por separado y se obtienen los estimadores de parámetros (media,total, proporción) para cada estrato, luego se combinan para tener los estimadores de toda la población.

Los estratos forman una partición de la población y se selecciona muestra en cada estrato en forma independiente.

Muestreo estratificado

Razones para utilizar este tipo de diseño de muestra:

1. Estadística. Para reducir la varianza de los estimadores, es decir, tener más precisión.

Cuando la población está constituída por unidades heterogéneas y tenemos una idea previa de los grupos de unidades más homogéneas entre sí, entonces es conveniente formar estratos.

Ejemplo de un caso ideal

Considere una población finita de 20 unidades en las cuales Y toma los valores:

$$\{6,3,4,4,5,3,6,2,3,2,2,6,5,3,5,2,4,6,4,5\}$$

$$\bar{Y} = 4 \quad S^2 = \frac{\sum_{i=1}^{20}\left(Y_i - \bar{Y}\right)^2}{19} = \frac{40}{19} = 2.11$$

Si tomamos una muestra aleatoria simple de tamaño 5 y usamos $\bar{y}$ como estimador de $\bar{Y}$, tenemos:

$$V(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} = \left(1 - \frac{5}{20}\right)\frac{2.11}{5} = 0.316$$

Ejemplo de un caso ideal

Dada la estructura de la población, se puede ordenar como:

$$\underbrace{2,2,2,2}\ \underbrace{3,3,3,3}\ \underbrace{4,4,4,4}\ \underbrace{5,5,5,5}\ \underbrace{6,6,6,6}$$

E S T R A T O S

Suponga que tenemos un mecanismo por el cual podemos seleccionar un elemento al azar de cada grupo para formar nuestra muestra de tamaño 5.
Obtenemos, en cada una de las posibles muestras, los valores:

$$\{2,3,4,5,6\} \text{ cuya } \bar{y} = 4 = \bar{Y}$$

Este estimador tendría varianza **cero** ya que la varianza dentro de cada estrato es cero y no hay fluctuaciones muestrales y, además, el estimador siempre sería igual al parámetro.

Ejemplo uso muestreo estratificado

Suponga un estudio donde interesa conocer alguna característica de los hogares en la Ciudad de México.

Se sabe que esa característica depende fuertemente del nivel socioeconómico de las familias.

Se construyen estratos considerando áreas de la ciudad con niveles socioeconómicos semejantes. Así las colonias se pueden clasificar en relación al nivel socioeconómico como: muy alto, alto, medio, medio-bajo y bajo, formando 5 estratos.

La encuesta se planea para cada estrato por separado.

2. Disponibilidad de marcos. Si la población está identificada a través de dos o más marcos, cada marco define un estrato.

Si para una parte de la población se tiene un buen marco, éste se usa para el muestreo de ese estrato; y las otras partes de la población se muestrean usando otros marcos, tal vez más imprecisos, y posiblemente con otros diseños de muestra.

Por ejemplo, en una encuesta de hogares se cuenta con un buen marco para la zona urbana de construcción antigua, pero las zonas rurales y las urbanas nuevas no tienen un marco adecuado.

Entonces, se podrían usar los planos catastrales para las zonas urbanas antiguas (un estrato), fotografías aéreas para zonas rurales (otro estrato) y en las zonas urbanas nuevas se podría construir un marco de manzanas, seleccionar manzanas y construir el marco de viviendas en las manzanas en muestra (muestreo en dos etapas).

3. Costo. Cuando hay diferentes costos de localizar y levantar la información de las unidades muestrales.

Por ejemplo, en una encuesta en predios agrícolas hay una región cuyo acceso es difícil (solo por avioneta ó a caballo).

Esta región puede constituir un estrato, que será muestreado con un tamaño de muestra más pequeño.

- El efecto de la formación de estratos es **reducir la variabilidad de los estimadores**.
- Ésta se puede reducir mucho si las unidades dentro de cada estrato son muy homogéneas y heterogéneas entre estratos.
- Se pueden usar diferentes diseños de muestra en cada estrato.
- No interesa tener estimaciones por estrato.

A nivel poblacional:

L	No. de estratos
N_h	No. de unidades muestrales estrato h, $h = 1, \ldots, L$
$N = \sum_{h=1}^{L} N_h$	No. de unidades en la población
Y_{hi}	valor de la medición en U_{hi}, $i = 1, \ldots, N_h, h = 1, \ldots, L$
$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} Y_{hi}}{N_h}$	media poblacional estrato h

Notación

$$Y_h = \sum_{i=1}^{N_h} Y_{hi} = N_h \bar{Y}_h \quad \text{total poblacional estrato } h$$

$$Y = \sum_{h=1}^{L} Y_h = \sum_{h=1}^{L} \sum_{i=1}^{N_h} Y_{hi} \quad \text{total poblacional}$$

$$\bar{Y} = \frac{Y}{N} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} Y_{hi}}{\sum_{h=1}^{L} N_h} \quad \text{media poblacional}$$

$$S_h^2 = \frac{\sum_{i=1}^{N_h} \left(Y_{hi} - \bar{Y}_h\right)^2}{N_h - 1} \quad \text{Varianza poblacional estrato } h$$

$$W_h = \frac{N_h}{N} \quad \text{peso del estrato}$$

$$\sum_{h=1}^{L} W_h = 1$$

Consideremos que tenemos una **m.a.s.** en cada estrato.

A nivel muestral:

$$n_h \quad \text{tamaño de muestra estrato } h$$

$$n = \sum_{h=1}^{L} n_h \quad \text{tamaño de muestra}$$

$$\hat{\bar{Y}}_h = \bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \quad \text{estimador media estrato } h$$

$$\hat{Y}_h = N_h \bar{y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi} \quad \text{estimador total estrato } h$$

Estimador del total

El estimador del **total** poblacional es:

$$\begin{aligned}\hat{Y} &= \sum_{h=1}^{L} \hat{Y}_h = \sum_{h=1}^{L} N_h \bar{y}_h \\ &= \sum_{h=1}^{L} N_h \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h} \\ &= \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{N_h}{n_h} y_{hi}.\end{aligned}$$

Donde $\frac{N_h}{n_h}$ es el **factor de expansión**.

La varianza del estimador del total es:

$$\begin{aligned} V\left(\hat{Y}\right) &= \sum_{h=1}^{L} V\left(\hat{Y}_h\right) \text{ muestras independientes en c/estrato} \\ &= \sum_{h=1}^{L} V\left(N_h \bar{y}_h\right) \\ &= \sum_{h=1}^{L} N_h^2 V\left(\bar{y}_h\right). \end{aligned}$$

Como tenemos una m.a.s. en cada estrato,

$$V\left(\hat{Y}\right) = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

El estimador de la varianza del estimador del total es:

$$\hat{V}\left(\hat{Y}\right) = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{S}_h^2}{n_h}$$

donde,

$$\hat{S}_h^2 = \frac{\sum_{i=1}^{n_h} \left(y_{hi} - \bar{y}_h\right)^2}{n_h - 1}$$

Si el tamaño de muestra en cada estrato es grande y podemos hacer la aproximación a la normal del estimador del total, el intervalo aproximado del $(1-\alpha) \times 100\,\%$ de confianza para el total poblacional es:

$$\hat{Y} \pm z_{1-\alpha/2} \sqrt{\hat{V}\left(\hat{Y}\right)}$$

Estimador de la media

El estimador de la **media** poblacional es:

$$\begin{aligned} \hat{\bar{Y}} &= \frac{\hat{Y}}{N} = \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{N} \\ &= \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{L} W_h \bar{y}_h \end{aligned}$$

$\hat{\bar{Y}}$ es una suma ponderada de los promedios muestrales en cada estrato.

Estimador de la media

La varianza del estimador de la media es:

$$\begin{aligned} V\left(\hat{\bar{Y}}\right) &= V\left(\sum_{h=1}^{L} W_h \bar{y}_h\right) \\ &= \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \end{aligned}$$

El estimador de la varianza del estimador de la media es:

$$\hat{V}\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{S}_h^2}{n_h}$$

Si el tamaño de muestra en cada estrato es grande y podemos hacer la aproximación a la normal del estimador de la media, el intervalo aproximado del $(1-\alpha)\times 100\,\%$ de confianza para la media poblacional es:

$$\hat{\bar{Y}} \pm z_{1-\alpha/2} \sqrt{\hat{V}\left(\hat{\bar{Y}}\right)}$$

Estimador de una proporción

Sea $Y_{hi} = \begin{cases} 1 & U_{hi} \text{ tiene la característica} \\ 0 & U_{hi} \text{ no tiene la característica} \end{cases}$

El estimador de la proporción P de unidades que tienen cierta característica es:

$$\hat{P} = \sum_{h=1}^{L} W_h \hat{p}_h \text{ con } \hat{p}_h = \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h}$$

La varianza de este estimador:

$$V\left(\hat{P}\right) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{P_h(1-P_h)}{n_h}$$

con estimador:

$$\hat{V}\left(\hat{P}\right) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}$$

Si el tamaño de muestra en cada estrato es grande y podemos hacer la aproximación a la normal del estimador de la proporción, el intervalo aproximado del $(1-\alpha)\times 100\,\%$ de confianza para la proporción poblacional es:

$$\hat{P} \pm z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{P}\right)}$$

Distribución de la muestra a los estratos

Suponga que se tiene un tamaño de muestra n determinado.

Cómo se reparte n entre los L estratos?

1. **Distribución óptima.**
Sea C_h el costo de obtener información de una unidad en el estrato h.
Se tiene una función de costo de la forma:

$$\text{costo} = C = C_0 + \sum_h C_h n_h$$

La varianza del estimador $\hat{\bar{Y}}$ se minimiza cuando (Cochran):

$$n_h = n \frac{N_h S_h}{\sqrt{C_h}} \left[\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{C_h}} \right]^{-1}$$

Distribución de la muestra a los estratos

Observe que,

$$n_h \propto \frac{N_h S_h}{\sqrt{C_h}}$$

Esto quiere decir que en un estrato dado, se toma más muestra si:

- El estrato es más grande
- El estrato es más variable
- El costo es menor

2. **Distribución de Neyman.**
Si se considera que los costos C_h son constantes en todos los estratos:

$$n_h = n \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}$$

3. **Distribución proporcional.**
Si se considera que tanto los costos como las varianzas S_h son constantes en todos los estratos, entonces:

$$n_h = n \frac{N_h}{N} = n\, W_h$$

Esta distribución produce muestras autoponderadas:

$$\frac{n_h}{N_h} = \frac{n}{N} \Rightarrow \frac{N_h}{n_h} = \frac{N}{n} \text{ factor de expansión}$$

Tamaño de muestra

1. Consideremos la distribución óptima:

$$n_h = n \frac{N_h S_h}{\sqrt{C_h}} \left[\sum_{i=1}^{L} \frac{N_i S_i}{\sqrt{C_i}} \right]^{-1}$$

1.1 Valor de n que produce varianza mínima para un costo total fijo.

$$C = C_0 + \sum_{h=1}^{L} n_h C_h$$

sustituyendo la expresión para n_h y despejando n:

$$C - C_0 = \sum_{h=1}^{L} C_h n_h$$

$$
\begin{aligned}
C - C_0 &= \sum_{h=1}^{L} C_h \left[n \frac{N_h S_h}{\sqrt{C_h}} \left(\sum_{i=1}^{L} \frac{N_i S_i}{\sqrt{C_i}} \right)^{-1} \right] \\
&= n \sum_{h=1}^{L} \frac{C_h N_h S_h}{\sqrt{C_h}} \left(\sum_{i=1}^{L} \frac{N_i S_i}{\sqrt{C_i}} \right)^{-1} \\
n &= \frac{(C - C_0) \sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} N_h S_h \sqrt{C_h}}
\end{aligned}
$$

1.2 Valor de n que produce costo mínimo para una varianza fija, o equivalentemente para un error de estimación fijo

$$\delta = z_{1-\alpha/2}\sqrt{V\left(\hat{\bar{Y}}\right)}$$

a) Para estimar la **media**

$$\begin{aligned} V\left(\hat{\bar{Y}}\right) &= \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \\ &= \sum_{h=1}^{L} \frac{N_h^2}{N^2} \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2 \ldots\ldots (1) \end{aligned}$$

La asignación óptima es:

$$n_h = n \frac{N_h S_h}{\sqrt{C_h}} \left[\sum_{i=1}^{L} \frac{N_i S_i}{\sqrt{C_i}} \right]^{-1} \ldots\ldots (2)$$

$$\delta = z_{1-\alpha/2} \sqrt{V\left(\hat{\bar{Y}}\right)} \Rightarrow V\left(\hat{\bar{Y}}\right) = \frac{\delta^2}{z^2_{1-\alpha/2}} \ldots\ldots (3)$$

sustituyendo (2) y (3) en (1) y despejando n:

$$n = \frac{\sum_{h=1}^{L} N_h S_h \sqrt{C_h} \left[\sum_{i=1}^{L} N_i S_i / \sqrt{C_i} \right]}{N^2 \frac{\delta^2}{z^2_{1-\alpha/2}} + \sum_{h=1}^{L} N_h S_h^2}$$

b)Para estimar el **total**

$$V\left(\hat{Y}\right) = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \ldots\ldots (4)$$

$$\delta = z_{1-\alpha/2}\sqrt{V\left(\hat{Y}\right)} \Rightarrow V\left(\hat{Y}\right) = \frac{\delta^2}{z_{1-\alpha/2}^2} \ldots\ldots (5)$$

$$n_h = n \frac{N_h S_h}{\sqrt{C_h}} \left[\sum_{i=1}^{L} \frac{N_i S_i}{\sqrt{C_i}}\right]^{-1} \ldots\ldots (6)$$

Sustituyendo (5) y (6) en (4) y despejando n:

$$n = \frac{\sum_{h=1}^{L} N_h S_h \sqrt{C_h} \left[\sum_{i=1}^{L} N_i S_i / \sqrt{C_i}\right]}{\frac{\delta^2}{z_{1-\alpha/2}^2} + \sum_{h=1}^{L} N_h S_h^2}$$

2. Considerando la asignación de Neyman (costos C_h constantes):

$$n_h = n \frac{N_h S_h}{\sum_{i=1}^{L} N_i S_i}$$

2.1 Para estimar la **media:**

$$n = \frac{\left[\sum_{h=1}^{L} N_h S_h\right]^2}{N^2 \frac{\delta^2}{z^2_{1-\alpha/2}} + \sum_{h=1}^{L} N_h S_h^2}$$

2.2 Para estimar el **total:**

$$n = \frac{\left[\sum_{h=1}^{L} N_h S_h\right]^2}{\frac{\delta^2}{z^2_{1-\alpha/2}} + \sum_{h=1}^{L} N_h S_h^2}$$

3. Si consideramos la distribución proporcional:

$$n_h = n\,\frac{N_h}{N}$$

3.1 Para estimar la **media:**

$$V\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{L} \frac{N_h^2}{N^2}\left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$$

$$\delta = z_{1-\alpha/2}\sqrt{V\left(\hat{\bar{Y}}\right)} \Rightarrow V\left(\hat{\bar{Y}}\right) = \frac{\delta^2}{z_{1-\alpha/2}^2}$$

$$n = \frac{N\sum_{h=1}^{L} N_h S_h^2}{N^2\frac{\delta^2}{z_{1-\alpha/2}^2} + \sum_{h=1}^{L} N_h S_h^2}$$

3.2 Para estimar el **total:**

$$n = \frac{N \sum_{h=1}^{L} N_h S_h^2}{\frac{\delta^2}{z_{1-\alpha/2}^2} + \sum_{h=1}^{L} N_h S_h^2}$$

Nota. Para estimar proporciones utilice las expresiones de tamaño de muestra para estimar la media con $S_h^2 = P_h(1 - P_h)$.

Muestreo estratificado

Se puede demostrar (Cochran) que:

$$V_{opt}\left(\hat{\bar{Y}}\right) \leq V_{prop}\left(\hat{\bar{Y}}\right) \leq V_{m.a.s.}\left(\hat{\bar{Y}}\right)$$