Resumen Tema 3: Muestreo estratificado.

Hipótesis: Marco perfecto, sin omisiones ni duplicados y elementos bien definido.

En casos en que la población sea muy heterogénea, para aumentar la precisión del M.A.S. podemos recurrir a la estratificación de la población, es decir a la subdivisión de la población de N unidades en L subpoblaciones, de tamaños $N_1, N_2, \cdots, N_L$ respectivamente, que no se superponen y juntas forman la totalidad de la población: $\sum_{h=1}^{L} N_h = N$.

Estas subpoblaciones reciben el nombre de estratos y para realizar el muestreo se extrae una muestra de cada estrato, tomando como muestra final el conjunto de todas las submuestras.

Cada muestreo en cada estrato debe realizarse de forma independiente, pudiéndose aplicar métodos distintos de muestreo en cada uno según se considere oportuno. Si se realiza un M.A.S. en cada estrato, el procedimiento se conoce como muestreo estratificado aleatorio.

Proceso para realizar un Muestreo Estratificado

1- Partimos de U población con N unidades. Dividimos la población en L subgrupos disjuntos de tamaños $N_1, N_2, \cdots, N_L$.

2- Realizamos en cada estrato o subgrupo poblacional un muestreo de tamaño n_h, de tal forma que $\sum_{h=1}^{L} n_h = n$ siendo n el tamaño deseado de muestra.

3- La muestra final de la población será la formada por todas las submuestras obtenidas en cada subpoblación.

Notación específica para este muestreo:

- N_h: Número total de unidades en el estrato h,
- n_h: Número de unidades de la muestra en el estrato h,
- x_{ih}: Valor obtenido en la unidad i del estrato h,
- $W_h = \frac{N_h}{N}$: Ponderación del estrato h,
- $f_h = \frac{n_h}{N_h}$: Fracción de muestreo en el estrato h,
- $\overline{X}_h$: Media poblacional del estrato h,
- $\overline{x}_h$: Media muestral del estrato h,
- S_h^2: Cuasivarianza poblacional del estrato h.

Parámetros	**Estimadores insesgados**	**Varianzas**	**Estimadores insesgados varianza**
$X = \sum_{h=1}^{L} X_h$	$\widehat{X}_{st} = \sum_{h=1}^{L} N_h \overline{x}_h$	$V(\widehat{X}_{st}) = \sum_{h=1}^{L} N_h^2 (1-f_h) \frac{S_h^2}{n_h}$	$\widehat{V}(\widehat{X}_{st}) = \sum_{h=1}^{L} N_h^2 (1-f_h) \frac{s_h^2}{n_h}$
$\overline{X} = \sum_{h=1}^{L} W_h \overline{X}_h$	$\overline{x}_{st} = \sum_{h=1}^{L} w_h \overline{x}_h$	$V(\overline{x}_{st}) = \sum_{h=1}^{L} W_h^2 (1-f_h) \frac{S_h^2}{n_h}$	$\widehat{V}(\overline{x}_{st}) = \sum_{h=1}^{L} W_h^2 (1-f_h) \frac{s_h^2}{n_h}$
$P = \sum_{h=1}^{L} W_h P_h$	$\widehat{P}_{st} = \sum_{h=1}^{L} w_h \widehat{P}_h$	$V(\widehat{P}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h - 1} \frac{P_h Q_h}{n_h}$	$\widehat{V}(\widehat{P}_{st}) = \sum_{h=1}^{L} W_h^2 (1-f_h) \frac{\widehat{P}_h \widehat{Q}_h}{n_h - 1}$

Estimación de los Intervalos de Confianza

- IC para el total con nivel de confianza $1-\alpha$.

$$\left(N\overline{x}_{st} - z_{\alpha/2}\sqrt{\widehat{V}(\widehat{X}_{st})}, N\overline{x}_{st} + z_{\alpha/2}\sqrt{\widehat{V}(\widehat{X}_{st})} \right)$$

- IC para la media con nivel de confianza $1-\alpha$.

$$\left(\overline{x}_{st} - z_{\alpha/2}\sqrt{\widehat{V}(\overline{x}_{st})}, \overline{x}_{st} + z_{\alpha/2}\sqrt{\widehat{V}(\overline{x}_{st})} \right)$$

- IC para proporciones con nivel de confianza $1-\alpha$.

$$\left(\widehat{P}_{st} - z_{\alpha/2}\sqrt{\widehat{V}(\widehat{P}_{st})}, \widehat{P}_{st} + z_{\alpha/2}\sqrt{\widehat{V}(\widehat{P}_{st})} \right)$$

Elección del tamaño muestral en los estratos: Afijación

- Uniforme: $w_h = \frac{1}{N}$, $n_h = \frac{n}{L}$,
- Proprcional: $w_h = W_h = \frac{N_h}{N}$, $n_h = \frac{n}{N} N_h$,
- Neyman: $w_h = \frac{S_h N_h}{\sum_{j=1}^{L} S_j N_j}$, $n_h = \frac{S_h N_h}{\sum_{j=1}^{L} S_j N_j} n$
- Óptima: $w_h = \frac{S_h N_h/\sqrt{C_h}}{\sum_{j=1}^{L} S_j N_j/\sqrt{C_j}}$, $n_h = \frac{S_h N_h/\sqrt{C_h}}{\sum_{j=1}^{L} S_j N_j/\sqrt{C_j}} n$

Determinación del tamaño muestral para un error máximo admisible $e = |\widehat{\theta} - \theta|$ con nivel de confianza p_k:

- Media poblacional $\overline{X}_{st}$: $n = \dfrac{\sum_{h=1}^{L} \dfrac{W_h^2 S_h^2}{w_h}}{\dfrac{e^2}{k^2} + \dfrac{1}{N} \displaystyle\sum_{h=1}^{L} W_h S_h^2}$,
- Total poblacional $\widehat{X}_{st}$: $n = \dfrac{\sum_{h=1}^{L} \dfrac{N_h^2 S_h^2}{w_h}}{\dfrac{e^2}{k^2} + \displaystyle\sum_{h=1}^{L} N_h S_h^2}$,
- Proporción poblacional $\widehat{P}_{st}$: $n = \dfrac{\sum_{h=1}^{L} W_h^2 \dfrac{N_h}{N_h - 1} \dfrac{P_h Q_h}{w_h}}{\dfrac{e^2}{k^2} + \dfrac{1}{N^2} \displaystyle\sum_{h=1}^{L} \dfrac{N_h^2}{N_h - 1} P_h Q_h}$.

Debe tomarse como tamaño muestral n el valor entero más próximo por exceso al obtenido en la fórmula. Como puede verse en las fórmulas, para determinar el tamaño muestral se necesita conocer: los tamaños de los estratos, el error y el nivel de confianza, la variabilidad de cada estrato y el peso correspondiente a cada estrato en la muestra, es decir la afijación usada, la cual me determina que w_h debo sustituir en cada fórmula.

Estratificación a posteriori:

Si no se puede conocer el estrato al cual pertenece una unidad hasta que no se obtienen todos los datos, se debe realizar una estratificación a posteriori que consiste en estudiar un carácter sobre una población, de la cual se conoce otro carácter.

El método consiste en tomar una muestra de la población de la cual se obtienen los valores de los dos caracteres, el que se desea estudiar y el que se conoce y se definen los estratos a posteriori según la variable conocida.

Si el carácter conocido esta relacionado con el que se desea estudiar, el tipo de muestra será representativa y conducirá a mejores resultados que si no hubiera habido estratificación.

Sus estimadores tienen la misma fórmula que los estimadores estratificados, por ejemplo,

$$\overline{x}_{post} = \sum_{h=1}^{L} W_h \overline{x}_h = \overline{x}_{st}$$

Sin embargo, los tamaños n_h ahora son aleatorios, lo que hace la varianza de los estimadores si cambie:

$$V(\overline{x}_{post}) = \frac{1-f}{n} \sum_{h=1}^{L} \frac{N_h}{N} S_h^2 + \frac{1-f}{n} \sum_{h=1}^{L} W_h (1 - W_h) \frac{S_h^2}{n_h}$$