Técnicas de Muestreo I

Patricia Isabel Romero Mares

Departamento de Probabilidad y Estadística
IIMAS UNAM

octubre 2015

Muestreo Sistemático

Es la forma de seleccionar la muestra en la cual solamente la primera unidad de la muestra es tomada al azar y el resto se selecciona sistemáticamente.

$$\text{Población} = \{U_1, U_2, \ldots, U_N\}$$

$$\text{muestra} = \{u_1, u_2, \ldots, u_n\}$$

Sea $k = \frac{N}{n}$ el intervalo de muestreo (suponga que k es entero).

Método: Se selecciona un número aleatorio (arranque aleatorio), i tal que $1 \leq i \leq k$. La muestra consiste de las unidades:

$$U_i, U_{i+k}, U_{i+2k}, \ldots, U_{i+(n-1)k}$$

Se dice que se selecciona una unidad de cada k unidades, "1 en k" ó "1 de cada k".

Lo que se hace, en realidad, es dividir la población de N elementos en k muestras (conglomerados) de tamaño n.

$$
\begin{array}{cccccc}
1 & 2 & \dots & i & \dots & k \\
1+k & 2+k & \dots & i+k & \dots & 2k \\
1+2k & 2+2k & \dots & i+2k & \dots & 3k \\
\vdots & \vdots & \dots & \vdots & \dots & \vdots \\
1+(j-1)k & 2+(j-1)k & \dots & i+(j-1)k & \dots & jk \\
\vdots & \vdots & \dots & \vdots & \dots & \vdots \\
1+(n-1)k & 2+(n-1)k & \dots & i+(n-1)k & \dots & nk=N
\end{array}
$$

Cuántas muestras posibles hay?

$$k = \frac{N}{n}$$

Cada una de estas muestras tiene la misma probabilidad de ser seleccionada:

$$P\,(\text{cualquier muestra}) = \frac{1}{k} = \frac{n}{N}$$

Además,

$$\pi_i = P\,(U_i \text{ en muestra}) = \frac{1}{k} = \frac{n}{N},\ \ i = 1, \ldots, N$$

$$\pi_{ij} = P\,(U_i, U_j \text{ en muestra}) = \begin{cases} \frac{n}{N} & U_i, U_j \in \text{mismo conglomerado} \\ 0 & U_i, U_j \notin \text{mismo conglomerado} \end{cases}$$

Ventajas muestreo sistemático

1. Es más fácil seleccionar la muestra que en m.a.s., especialmente en diseños de muestra polietápicos, donde el encuestador tiene que hacer la selección de unidades de última etapa *in situ*.

2. Una muestra sistemática se dispersa más uniformemente entre toda la población, por lo que es más factible producir una muestra "representativa" que en m.a.s.

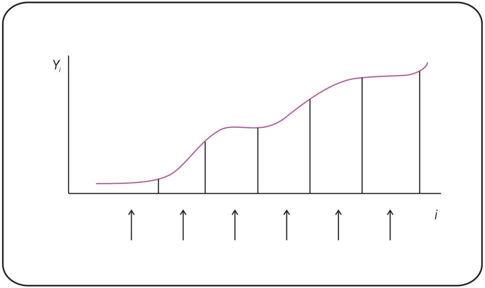
Desventajas muestreo sistemático

1. Un mal arreglo de las unidades en el marco puede producir muestras ineficientes.
2. **No** se pueden calcular estimadores de la varianza con una sola muestra sistemática.

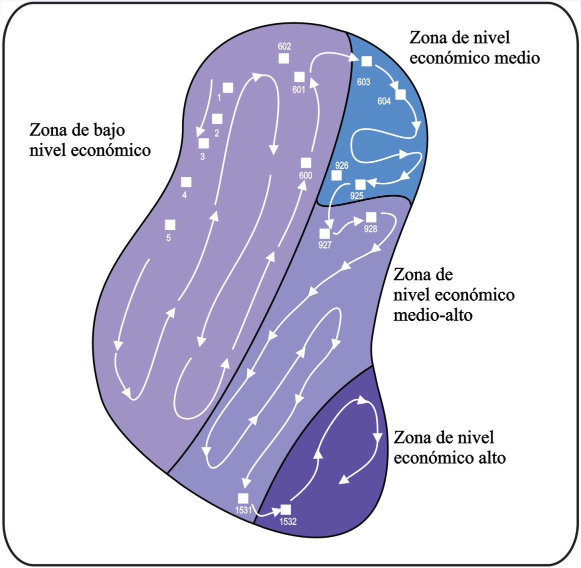
Orden de las unidades en el marco

a. Cuando las unidades de la población están en un orden aleatorio en el marco, con respecto a los valores de Y_i, el muestreo sistemático es equivalente al m.a.s.

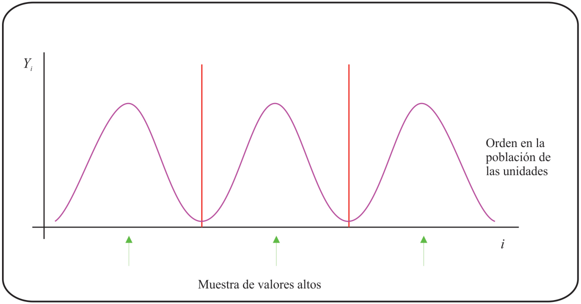
b. Cuando las unidades de la población están ordenados en el marco en relación a los valores de Y_i, el muestreo sistemático produce varianzas de los estimadores menores que los correspondientes en el m.a.s. Esto se debe a que la muestra queda más dispersa sobre la población.



Orden de las unidades en el marco



c. Si las unidades de la población tienen un orden que se refleja en cambios periódicos de los valores de Y_i y el periodo coincide con el valor de k, el muestreo sistemático puede producir varianzas mayores de los estimadores que el m.a.s. En este caso el problema es que la muestra puede coincidir con todos los valores bajos (o altos) de Y_i, siendo de esta manera poco representativa y con fuertes fluctuaciones de muestra a muestra.



$$\hat{\bar{Y}}_{sis} = \bar{y}$$

Con varianza:

$$V\left(\hat{\bar{Y}}_{sis}\right) = \frac{\sum_{i=1}^{k}\left(\bar{Y}_i - \bar{Y}\right)^2}{k} = \frac{k-1}{k}\frac{\sum_{i=1}^{k}\left(\bar{Y}_i - \bar{Y}\right)^2}{k-1} = \frac{k-1}{k}S_b^2$$

donde S_b^2 es la varianza entre conglomerados (between).

No hay forma de estimar $V\left(\hat{\bar{Y}}_{sis}\right)$, usualmente se utilizan las expresiones del m.a.s.

Se puede demostrar (Teorema 8.1, Cochran) que

$$V\left(\hat{\bar{Y}}_{sis}\right) = \frac{N-1}{N}S^2 - \frac{k\,(n-1)}{N}S_w^2$$

donde,

$$S_w^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_i\right)^2$$

es la varianza dentro de conglomerados.

El estimador de la media de una muestra sistemática es más preciso que el estimador de la media de una m.a.s. si y solo si: (Corolario Teorema 8.1, Cochran)

$$S_w^2 > S^2$$

Demostración:

$$V\left(\hat{\bar{Y}}_{mas}\right) = \frac{N-n}{N}\frac{S^2}{n}$$

entonces

$$V\left(\hat{\bar{Y}}_{sis}\right) < V\left(\hat{\bar{Y}}_{mas}\right) \text{ si y solo si}$$

$$\frac{N-1}{N}S^2 - \frac{k(n-1)}{N}S_w^2 < \frac{N-n}{N}\frac{S^2}{n} \Rightarrow$$

$$
\begin{aligned}
k\,(n-1)\,S_w^2 &> \left(N-1-\frac{N-n}{n}\right)S^2 = \\
&\left(\frac{Nn-n-N+n}{n}\right)S^2 = \frac{N}{n}\,(n-1)\,S^2 = \\
&k\,(n-1)\,S^2 \Rightarrow
\end{aligned}
$$

$$
S_w^2 > S^2
$$

Entonces, el muestreo sistemático es más preciso que el m.a.s. cuando la varianza dentro de las muestras sistemáticas (conglomerados) es **mayor** que la varianza de la población entera.
Es decir, se requieren unidades heterogéneas dentro de la muestra.

Otra forma de la varianza (Cochran secc. 8.3):

$$V\left(\hat{\bar{Y}}_{sis}\right) = \frac{S^2}{n}\left(\frac{N-1}{N}\right)[1+(n-1)\rho]$$

con ρ el coeficiente de correlación intraclase:

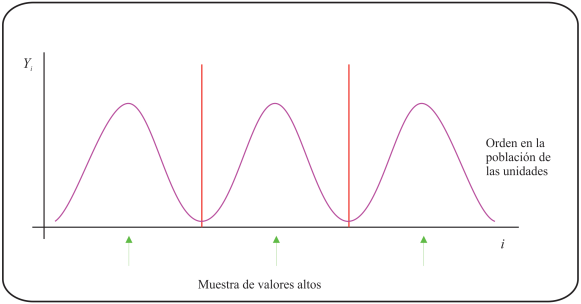
$$\rho = \frac{\sum_{r=1}^{k}\sum_{j' \neq j=1}^{n}\left(Y_{rj}-\bar{Y}\right)\left(Y_{rj'}-\bar{Y}\right)}{(n-1)(N-1)S^2}$$

Ya que $V\left(\hat{\bar{Y}}_{sis}\right) \geq 0 \Rightarrow$

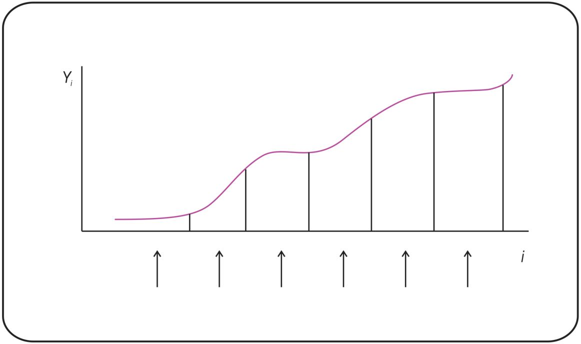
$$\frac{-1}{n-1} \leq \rho \leq 1$$

Comparación con m.a.s.

1. Si ρ está cercano a $1 \Rightarrow$ los elementos de la muestra son muy parecidos $\Rightarrow \; V\left(\hat{\hat{Y}}_{sis}\right) \geq V\left(\hat{\hat{Y}}_{mas}\right)$



2. Si $\rho < 0 \Rightarrow$ los elementos son diferentes
$\Rightarrow V\left(\hat{\hat{Y}}_{sis}\right) \leq V\left(\hat{\hat{Y}}_{mas}\right)$



3. Si $\rho = 0 \Rightarrow V\left(\hat{\bar{Y}}_{sis}\right) \approx V\left(\hat{\bar{Y}}_{mas}\right)$ población con orden aleatorio, por lo tanto el muestreo sistemático es equivalente al m.a.s.

Cuando N no es divisible entre n, es decir,

$$N = nk + r, \;\; r < k$$

el tamaño de muestra será n ó $n+1$ dependiendo de la semilla aleatoria seleccionada.

En este caso la media muestral es un estimador sesgado de la media poblacional, pero el sesgo es negligible.