Técnicas de Muestreo I

Patricia Isabel Romero Mares

Departamento de Probabilidad y Estadística
IIMAS UNAM

noviembre 2015

Muestreo con probabilidad proporcional al tamaño

Cuando las unidades muestrales varían considerablemente de tamaño y la variable bajo estudio está relacionada con el tamaño de la unidad, el m.a.s. podría no ser un diseño adecuado.

Lo adecuado es considerar esta información de tamaño asignando las probabilidades de selección de forma proporcional al tamaño de la unidad.

Muestreo con probabilidad proporcional al tamaño

(ppt ó pps)

Hay dos formas:

Con reemplazo. La probabilidad de selección de una unidad específica en cualquier extracción es la misma $\Rightarrow$ cálculo de varianza sencillo.

Sin reemplazo. La probabilidad de selección de una unidad específica varía de acuerdo al número de extracción $\Rightarrow$ cálculo de varianza difícil o imposible.

Algoritmo

$N =$ tamaño de la población
$n =$ tamaño de la muestra
X_i medida de tamaño de la U_i, $i = 1, \ldots, N$ (es conocida)

Algoritmo.

1. Se forman los acumulados sucesivos.

$$\begin{array}{ll} U_1 & X_1 \\ U_2 & X_1 + X_2 \\ U_3 & X_1 + X_2 + X_3 \end{array}$$

$$\begin{array}{ll} U_N & X_1 + X_2 + X_3 + \ldots + X_N = X \end{array}$$

Algoritmo

2. Se selecciona un número aleatorio R tal que

$$1 \leq R \leq \sum_{i=1}^{N} X_i = X$$

3. Se selecciona la U_i si

$$X_1 + X_2 + \ldots + X_{i-1} < R \leq X_1 + X_2 + \ldots + X_i$$

4. Se repiten los pasos 2 y 3 hasta completar n unidades en muestra.

Una comunidad tiene 10 huertos de diferentes tamaños. Se desea tomar una muestra ppt con reemplazo de tamaño 4.

No. huerto	Tamaño	Acumulado	Rango
1	150	150	1-150
2	50	200	151-200
3	80	280	201-280
4	100	380	281-380
5	200	580	381-580
6	160	740	581-740
7	40	780	741-780
8	220	1000	781-1000
9	60	1060	1001-1060
10	140	1200	1061-1200

Se seleccionan 4 números aleatorios R tales que

$$1 \leq R \leq 1200$$

Para $R = 600$ se selecciona U_6

$R = 2$ se selecciona U_1

$R = 796$ se selecciona U_8

$R = 901$ se selecciona U_8

Muestra=$\{U_1, U_6, U_8, U_8\}$

Estimador de total

Y_i valor de la característica de interés en U_i, $i = 1, \ldots, N$

X_i valor de la medida de tamaño en U_i, $i = 1, \ldots, N$ (conocida)

P_i probabilidad de extracción de U_i, $i = 1, \ldots, N$

$$P_i = \frac{X_i}{X}, \text{ donde } X = \sum_{i=1}^{N} X_i$$

Sea

$$Z_i = \frac{Y_i}{P_i} = \frac{Y_i}{\frac{X_i}{X}} = \frac{Y_i}{X_i} X \quad i = 1, \ldots, N$$

Entonces,

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{P_i} = \frac{1}{n} \sum_{i=1}^{n} z_i = \bar{z}$$

A cada elemento de la población se le asocia el valor

$$Z_i = \frac{Y_i}{P_i} = \frac{Y_i}{X_i} X$$

Al tomar la muestra, los valores obtenidos serán:

$$z_i = \frac{y_i}{P_i}$$

$$\{z_1, z_2, \ldots, z_n\} \text{ v.a.i.i.d.}$$

La probabilidad de elegir en la primera extracción la U_i, es decir, que el valor de z_1 sea Z_i es:

$$P(z_1 = Z_i) = \frac{X_i}{X}, \; i = 1, \ldots, N$$

La probabilidad de elegir en la j-ésima extracción la U_i es:

$$P(z_j = Z_i) = \frac{X_i}{X},\ i = 1, \ldots, N$$

Entonces,

$$\begin{aligned} E(z_j) &= \sum_{i=1}^{N} Z_i P(z_j = Z_i) \\ &= \sum_{i=1}^{N} Z_i \frac{X_i}{X} = \sum_{i=1}^{N} \frac{Y_i}{X_i} X \frac{X_i}{X} \\ &= \sum_{i=1}^{N} Y_i = Y \end{aligned}$$

Cualquier z_j es un estimador insesgado de Y, de hecho, un estimador de razón:

$$z_j = \frac{Y_j}{X_j} X$$

Estimador del total

$$
\begin{aligned}
V(z_j) &= E\left[z_j - E(z_j)\right]^2 \\
&= \sum_{i=1}^{N} (Z_i - Y)^2 P(z_j = Z_i) \\
&= \sum_{i=1}^{N} (Z_i - Y)^2 \frac{X_i}{X} \\
&= \sum_{i=1}^{N} \left(\frac{Y_i}{X_i} X - Y\right)^2 \frac{X_i}{X} \left\{\frac{X^2}{X^2}\right\} \\
V(z_j) &= \sum_{i=1}^{N} \left(\frac{Y_i}{X_i} - \frac{Y}{X}\right)^2 X_i X = S_z^2
\end{aligned}
$$

Entonces,

$$\begin{aligned} V\left(\hat{Y}\right) &= V\left(\bar{z}\right) = \frac{1}{n^2}\sum_{i=1}^{n} V\left(z_i\right) \ \{\text{son independientes}\} \\ &= \frac{1}{n^2} n V\left(z_j\right) = \frac{1}{n} V\left(z_j\right) = \frac{1}{n} S_z^2 \\ V\left(\hat{Y}\right) &= \frac{X}{n}\sum_{i=1}^{N}\left(\frac{Y_i}{X_i} - \frac{Y}{X}\right)^2 X_i \end{aligned}$$

Note que si se tiene una proporcionalidad perfecta entre Y_i y X_i entonces

$$\begin{aligned} \frac{Y_i}{X_i} &= k, \;\; i = 1, \ldots, N \\ \frac{Y}{X} &= k \end{aligned}$$

entonces,

$$\left(\frac{Y_i}{X_i} - \frac{Y}{X} \right) = 0 \Rightarrow V\left(\hat{Y}\right) = 0$$

y $\hat{Y} = Y$, ya que

$$E\left(\bar{z}\right) = \frac{1}{n} \sum_{i=1}^{n} E\left(z_i\right) = \frac{1}{n} nY = Y$$

El estimador de la varianza del estimador del total es:

$$\hat{V}\left(\hat{Y}\right) = \hat{V}\left(\bar{z}\right) = \frac{1}{n}\hat{S}_z^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(z_i - \bar{z}\right)^2}{n-1}$$

$$\begin{aligned}
\hat{V}\left(\hat{Y}\right) &= \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i}{P_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}\right)^2 \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(X\frac{y_i}{X_i} - \frac{X}{n}\sum_{i=1}^{n}\frac{y_i}{X_i}\right)^2 \\
\hat{V}\left(\hat{Y}\right) &= \frac{X^2}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{X_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{X_i}\right)^2
\end{aligned}$$

Estimador de la media poblacional

$$\hat{\bar{Y}} = \frac{\hat{Y}}{N}$$

$$V\left(\hat{\bar{Y}}\right) = \frac{1}{N^2} V\left(\hat{Y}\right) = \frac{1}{N^2} \frac{S_z^2}{n}$$

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{1}{N^2} \hat{V}\left(\hat{Y}\right) = \frac{1}{N^2} \frac{\hat{S}_z^2}{n}$$

Considerando que $\hat{Y}$ tiene distribución normal, el tamaño de muestra para una precisión δ y confianza $1-\alpha$ usando muestreo ppt con reemplazo es:

$$n = \frac{z_{\alpha/2}^2 S_z^2}{\delta^2}$$

Existe un algoritmo de selección ppt sin reemplazo, llamado ppt sistemático.

Algoritmo

1. Forme los totales acumulados sucesivos

$$
\begin{array}{ll}
U_1 & X_1 \\
U_2 & X_1 + X_2 \\
U_3 & X_1 + X_2 + X_3 \\
 & \\
U_N & X_1 + X_2 + X_3 + \ldots + X_N = X
\end{array}
$$

2. Seleccione un número aleatorio R tal que

$$1 \leq R \leq k \text{ con } k = \sum_{i=1}^{N} X_i/n = X/n$$

3. Selecione las unidades cuyos índices satisfagan

$$X_1 + X_2 + \cdots + X_{i-1} < R + jk \leq X_1 + X_2 + \cdots + X_i$$

para $j = 0, 1, 2, \ldots, (n-1)$.

El algoritmo asegura que ninguna unidad será seleccionada más de una vez si

$$X_i \leq k = \frac{X}{n}, \;\; i = 1, \ldots, N$$

Se utiliza este algoritmo para seleccionar muestras ppt sin reemplazo, pero se utilizan los estimadores del ppt con reemplazo.