

Sociedad Ecuatoriana de Estadística

Análisis de Encuestas por muestreo con R

“Muestreo Aleatorio Simple”

Andrés Peña Montalvo

andres.pena.montalvo@ciencias.unam.mx



Julio 2025



X Seminario Internacional
de Estadística Aplicada

Tabla de contenidos

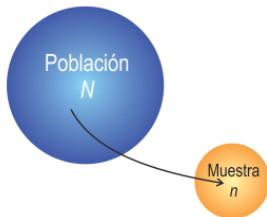
- 1 Muestreo Aleatorio Simple (M.A.S.)
- 2 Estimadores bajo M.A.S.
- 3 Intervalo de confianza
- 4 Estimadores de totales y proporciones
- 5 Tamaño de muestra

Muestreo Aleatorio Simple (M.A.S.)



Muestreo Aleatorio Simple (M.A.S.)

Muestreo Aleatorio Simple (*mas*)



Selección aleatoria de los elementos muestrales con probabilidades de selección en cualquier extracción iguales y sin reemplazo.

Muestreo Aleatorio Simple (M.A.S.)

De una población de N unidades, se selecciona una muestra de tal manera que **todas** las unidades de la población tienen **igual probabilidad** de ser seleccionadas.

- Se mide la unidad seleccionada y se regresa a la población. Si se hace esta operación n veces, se obtiene una muestra aleatoria simple seleccionada **con reemplazo**, las unidades pueden estar mas de una vez en la muestra.
- Se mide la unidad seleccionada y ya no se regresa a la población. Se seleccionan las siguientes unidades con igual probabilidad de las unidades que quedan en la población. Si se hace esta operación n veces, se obtiene una muestra aleatoria simple seleccionada **sin reemplazo**. Este es el procedimiento que vamos a estudiar.

Muestreo Aleatorio Simple (M.A.S.)

Población = $\{U_1, U_2, \dots, U_N\}$

muestra = $\{u_1, u_2, \dots, u_n\}$

muestra \subseteq Población

Características de interés

$\{X_1, X_2, \dots, X_N\}$

$\{Y_1, Y_2, \dots, Y_N\}$

$\{Z_1, Z_2, \dots, Z_N\}$

A cada U_i se le asocia una o varias características de interés X_i, Y_i, Z_i .

Muestreo Aleatorio Simple (M.A.S.)

Una muestra aleatoria simple se define de dos maneras equivalentes:

1. Una muestra aleatoria donde cualquier elemento $U_j, j = 1, \dots, N$ tiene una probabilidad $1/N$ de ser seleccionado en cualquiera de las n extracciones.

Como consecuencia la probabilidad de que un elemento $U_j, j = 1, \dots, N$ esté incluido en muestra es n/N .

$\pi_j = \frac{n}{N}$ es la probabilidad de inclusión de primer orden
 $\frac{1}{\pi_j} = \frac{N}{n}$ es el factor de expansión o peso muestral

Muestreo Aleatorio Simple (M.A.S.)

2. Cualquiera de las $\binom{N}{n}$ muestras posibles tiene la misma probabilidad de ser seleccionada.

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

Muestreo Aleatorio Simple (M.A.S.)

Mediante el proceso de muestreo lo que se desea es hacer inferencia a una población, específicamente se desea calcular una estimación de un parámetro de la población, como:

$$\text{Media} \quad \bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$$

$$\text{Total} \quad Y = \sum_{i=1}^N Y_i$$

$$\text{Proporción} \quad P = \frac{1}{N} \sum_{i=1}^N Y_i = \frac{Y}{N}, \text{ donde}$$

$$Y_i = \begin{cases} 1 & U_i \text{ tiene la característica} \\ 0 & U_i \text{ no tiene la característica} \end{cases}$$

Muestreo Aleatorio Simple (M.A.S.)

$$\text{Razón } R = \frac{Y}{X}$$

$$\begin{aligned} \text{Varianza } \sigma^2 &= \sum_{i=1}^N \frac{(Y_i - \bar{Y})^2}{N} \\ &= \frac{N-1}{N} \sum_{i=1}^N \frac{(Y_i - \bar{Y})^2}{N-1} \\ &= \frac{N-1}{N} S^2 \end{aligned}$$

Con $S^2 = \sum_{i=1}^N \frac{(Y_i - \bar{Y})^2}{N-1}$. Se usa S^2 en lugar de σ^2 por facilidad ya que tenemos un estimador insesgado de S^2 .

Estimadores bajo M.A.S.



Estimadores bajo M.A.S.

Un estimador insesgado de \bar{Y} , la media poblacional de la característica Y , es:

$$\hat{Y} = \sum_{i=1}^n \frac{y_i}{n} = \bar{y} \text{ media muestral}$$

Que sea un estimador insesgado quiere decir que

$$E(\hat{Y}) = \bar{Y}$$

y su varianza es:

$$V(\bar{y}) = E(\bar{y} - \bar{Y})^2 = \left(1 - \frac{n}{N}\right) \frac{S^2}{n}$$

Estimadores bajo M.A.S.

$V(\bar{y})$ se estima inesgadamente con:

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

Para mostrar que $\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{\hat{S}^2}{n}$ es un estimador inesgado de $V(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{S^2}{n}$, basta demostrar que $E(\hat{S}^2) = S^2$.

$$\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2; \quad S^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2$$

Estimadores bajo M.A.S.

$\frac{n}{N}$ es la fracción de muestreo o porcentaje de la población que se muestrea. $(1 - \frac{n}{N})$ es el factor de corrección por finitud, que ajusta por muestrear de una población finita. Toma en cuenta el hecho de que un estimador basado en una muestra con $n = 10$ unidades de una población de $N = 20$ unidades contiene más información acerca de la población que una muestra de tamaño $n = 10$ unidades de una población de $N = 20000$ unidades.

$$\left(1 - \frac{10}{20}\right) = \frac{1}{2} \text{ mitad de la varianza}$$

$$\left(1 - \frac{10}{20000}\right) = 0.9995 \text{ misma varianza que poblaciones infinitas}$$

Si $n = N$ entonces $V(\bar{y}) = 0$ se está haciendo un censo por lo que el estimador del parámetro tiene varianza cero.

Estimadores bajo M.A.S.

Ya tenemos un estimador de la media poblacional \bar{Y} que es la media muestral \bar{y} , que, una vez que tengamos los valores de la muestra de tamaño n nos dará la estimación puntual de \bar{Y} .

Sabemos que este estimador es insesgado y tenemos un estimador de su varianza, es decir, sabemos que

$$E(\bar{y}) = \bar{Y}$$

$\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{\hat{S}^2}{n}$ estimador de la varianza del estimador.

Intervalo de confianza

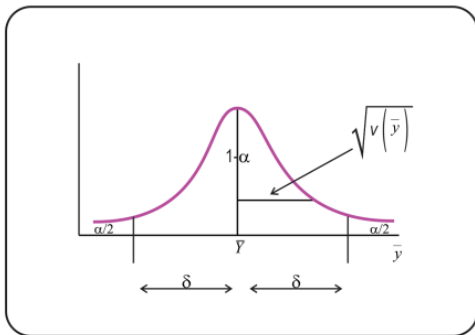


Intervalo de confianza

Por el Teorema Central del Límite podemos suponer que, con n suficientemente grande:

$$\bar{y} \sim N(\bar{Y}, V(\bar{y}))$$

$$\frac{\bar{y} - \bar{Y}}{\sqrt{V(\bar{y})}} \sim N(0, 1)$$



Intervalo de confianza

$$P(|\bar{y} - \bar{Y}| < \delta) = 1 - \alpha$$

$1 - \alpha$ confianza y δ precisión.

$$P(-\delta < \bar{y} - \bar{Y} < \delta) = 1 - \alpha$$

$$P\left(\frac{-\delta}{\sqrt{V(\bar{y})}} < \underbrace{\frac{\bar{y} - \bar{Y}}{\sqrt{V(\bar{y})}}}_{N(0,1)} < \frac{\delta}{\sqrt{V(\bar{y})}}\right) = 1 - \alpha$$

Intervalo de confianza

$$P\left(z_{\alpha/2} < \frac{\bar{y} - \bar{Y}}{\sqrt{V(\bar{y})}} < z_{1-\alpha/2}\right) = 1 - \alpha$$

$$P\left(-z_{1-\alpha/2}\sqrt{V(\bar{y})} < \bar{y} - \bar{Y} < z_{1-\alpha/2}\sqrt{V(\bar{y})}\right) = 1 - \alpha$$

(1)

$$P\left(\bar{y} - z_{1-\alpha/2}\sqrt{V(\bar{y})} < \bar{Y} < \bar{y} + z_{1-\alpha/2}\sqrt{V(\bar{y})}\right) = 1 - \alpha$$

Intervalo de confianza

El intervalo del $(1 - \alpha) \times 100\%$ de confianza para \bar{Y} es:

$$\left(\bar{y} - z_{1-\alpha/2} \sqrt{V(\bar{y})}, \bar{y} + z_{1-\alpha/2} \sqrt{V(\bar{y})} \right)$$

De tablas de la $N(0, 1)$

$$1 - \alpha = 0.99 \quad z_{.995} = 2.57$$

$$1 - \alpha = 0.95 \quad z_{.975} = 1.96$$

$$1 - \alpha = 0.90 \quad z_{.95} = 1.64$$

Más confianza implica intervalos más anchos.

Intervalo de confianza

Cuando no se conoce $V(\bar{y})$ y se estima con $\hat{V}(\bar{y})$ entonces,

$$\frac{\bar{y} - \bar{Y}}{\sqrt{\hat{V}(\bar{y})}} \sim t_{n-1}$$

y el intervalo aproximado del $(1 - \alpha) \times 100\%$ de confianza para \bar{Y} es:

$$\bar{y} \pm t_{n-1}^{1-\alpha/2} \sqrt{\hat{V}(\bar{y})}$$

En general, como n es grande, el valor de la t se aproxima a la normal y se usa como intervalo de confianza:

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

Estimadores de totales y proporciones



Estimador del total

$$Y = \sum_{i=1}^N Y_i = N\bar{Y}$$

$$\hat{Y} = N\hat{\bar{Y}} = N\bar{y} = \sum_{i=1}^n \frac{N}{n} y_i \left\{ \text{Note que } \frac{N}{n} = \frac{1}{\frac{n}{N}} \right\}$$

$$E(\hat{Y}) = Y$$

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

$$\hat{V}(\hat{Y}) = N^2 \left(1 - \frac{n}{N}\right) \frac{\hat{S}^2}{n} \text{ es insesgado para } V(\hat{Y})$$

Intervalo del $100(1 - \alpha)\%$ de confianza para Y es:

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

Estimador de una proporción

Sea:

$$Y_i = \begin{cases} 1 & U_i \text{ tiene la característica } A \\ 0 & U_i \text{ no tiene la característica } A \end{cases}$$

$$P = \frac{\# \text{ de elementos que tienen la característica } A}{\text{total de elementos}} = \frac{\sum_{i=1}^N Y_i}{N}$$

Un estimador insesgado de P es:

$$\hat{P} = \frac{\sum_{i=1}^n Y_i}{n} = \bar{y}$$

Con varianza:

$$V(\hat{P}) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{N}{N-1} P(1-P)$$

Estimador de una proporción

y su estimador es:

$$\hat{V}(\hat{P}) = \left(1 - \frac{n}{N}\right) \frac{\hat{P}(1 - \hat{P})}{n - 1}$$

Suponiendo normalidad, el intervalo del $100(1 - \alpha)\%$ de confianza es:

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

Determinación del tamaño de muestra



Determinación del tamaño de muestra

$$n = ?$$

n pequeña

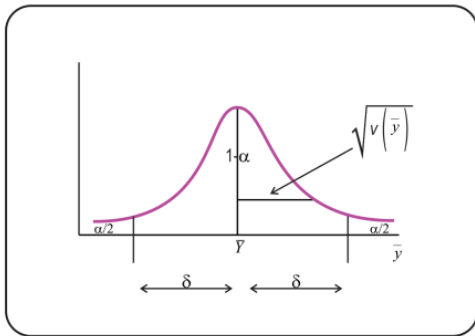
- inferencias inútiles
- intervalos de confianza muy grandes
- poca precisión

n grande

- costos elevados
- se puede descuidar la calidad de la información

Determinación del tamaño de muestra

Suponiendo normalidad en el estimador:



n para estimar un promedio

Se fija una precisión δ y una confianza $1 - \alpha$. De la gráfica anterior,

$$P(|\bar{y} - \bar{Y}| < \delta) = 1 - \alpha$$

$$P(\bar{y} - \delta < \bar{Y} < \bar{y} + \delta) = 1 - \alpha$$

Por otro lado, sabemos que:

$$P\left(-z_{1-\alpha/2} < \frac{\bar{y} - \bar{Y}}{\sqrt{V(\bar{y})}} < z_{1-\alpha/2}\right) = 1 - \alpha$$

$$P\left(\bar{y} - z_{1-\alpha/2}\sqrt{V(\bar{y})} < \bar{Y} < \bar{y} + z_{1-\alpha/2}\sqrt{V(\bar{y})}\right) = 1 - \alpha$$

Por lo tanto,

$$\delta = z_{1-\alpha/2}\sqrt{V(\bar{y})}$$

$$\delta = z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S^2}{n}} = z_{1-\alpha/2} \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S^2}$$

$$\delta^2 = z_{1-\alpha/2}^2 \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

Despejando n

$$n = \frac{1}{\frac{\delta^2}{S^2 z_{1-\alpha/2}^2} + \frac{1}{N}}$$

Si N es grande

$$n_0 = \frac{S^2 z_{1-\alpha/2}^2}{\delta^2}$$

Si N no es grande

$$n = \frac{1}{\frac{1}{n_0} + \frac{1}{N}} = \frac{n_0}{1 + \frac{n_0}{N}}$$

δ es el error absoluto.

n para estimar un promedio

Necesitamos conocer S^2 para calcular el tamaño de muestra.

Opciones:

1. Usar estimadores de S^2 de encuestas similares anteriores o de censos.
2. Estimar S^2 usando una encuesta piloto.

n para estimar un total

Suponiendo normalidad en el estimador:

$$n_0 = \frac{z_{1-\alpha/2}^2 N^2 S^2}{\delta^2}$$

$$n = \begin{cases} n_0 & \text{si } N \text{ es grande} \\ \frac{n_0}{1 + \frac{n_0}{N}} & \text{si } N \text{ no es grande} \end{cases}$$

n para estimar una proporción

Recordemos que con la definición de la variable a medir Y_i como 0 o 1 , tenemos que $P = \bar{Y}$, entonces, suponiendo normalidad en el estimador de P

$$n_0 = \frac{z_{1-\alpha/2}^2 P(1 - P)}{\delta^2}$$

$$n = \begin{cases} n_0 & \text{si } N \text{ es grande} \\ \frac{n_0}{1 + \frac{n_0}{N}} & \text{si } N \text{ no es grande} \end{cases}$$

Gracias!!!

