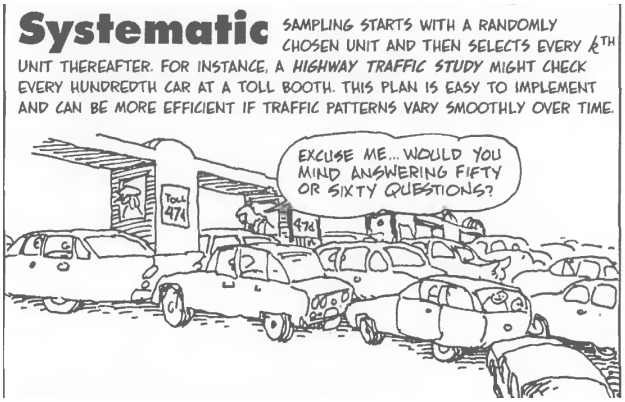
Muestreo sistemático



Autor: **Prof. Christian Salas Eljatib**

E-mail: **cseljatib@gmail.com**

Índice

1. Introducción

- Se entiende por diseños de muestreo sistemático todos aquellos en los cuales en una, varias o todas las etapas, la selección de los elementos de muestreo se ejecuta según un procedimiento sistemático, planteado a priori, de carácter regular y repetitivo, es decir, que elegido un primer elemento al azar, todos los demás quedan automáticamente determinados a partir de dicho primer elemento.
- Como ya sabemos, la idea es seleccionar una muestra de n elementos desde una población de N elementos.
- En un muestreo sistemático, designamos a a como el número entero que cumpla con $a \leq N/n$. Luego, seleccionar aleatoriamente un elemento entre los primeros a de la población. Posteriormente, continue seleccionando cada a-ésimo elemento en la población para ser incluido en la muestra.
- Lo anterior describe el tipo de muestreo sistemático con inicio aleatorio de 1–en–a.
- El contexto de muestreo podría consistir de una población de elementos, tanto que existe una secuencia natural desde el comienzo al final de la lista. Por ejemplo, si uno esta tratando con una población demográfica, la lista podría ser ordenada por orden alfabético, o por dirección, o cronologicamente por fecha de nacimiento, nivel de educación, etc.
- O incluso la lista puede tener ningun orden.
- El muestreo sistemático es frecuentemente usado en el contexto de poblaciones en un área, por ejemplo
  - una ciudad dividida en manzanas
  - un terreno dividido en cuadrantes
  - poligonos exagonales de USA empleado por el programa de monitoreo de bosques de US EPA. Es fácil ver que una muestra sistemática ofrece una más uniforme cobertura (geográfica) que al emplear un muestreo aleatorio simple.

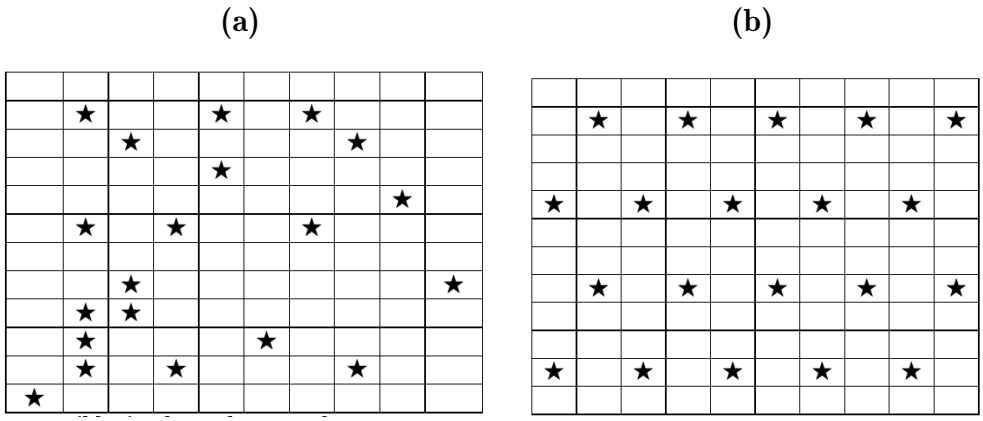


Figura 1: Diferencia entre una posible muestra aleatoria (a) y otra sistemática (b).

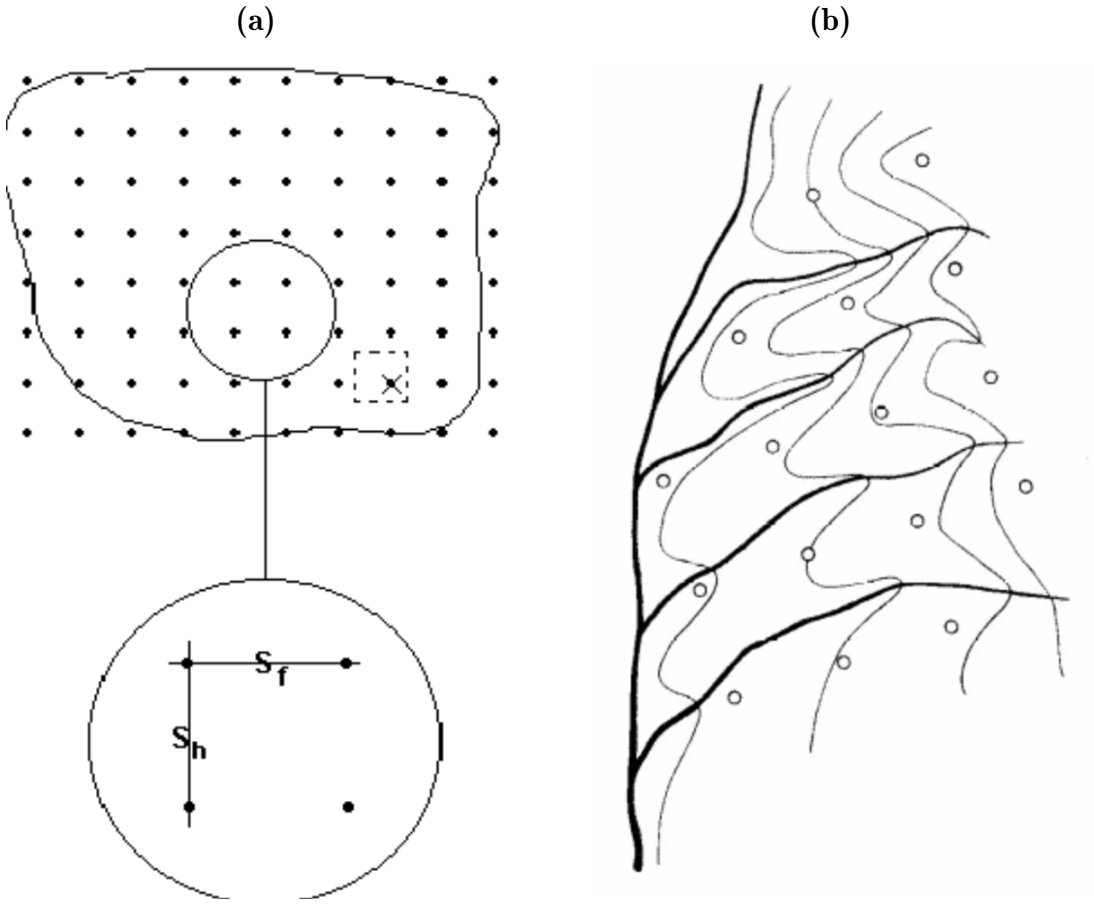


Figura 2: En (a) se muestra una grilla cuadrada de ubicación de parcelas de muestreo y en el recuadro punteado la selección aleatoria de la primera parcela a muestrear. Además, abajo se hace un zoom para indicar la distancia en la fila, y entre hileras. En (b) se representa la distribución potencial sistemática de parcelas en un rodal donde además se representan las curvas de nivel y cursos de agua (en líneas mas gruesas).

1.1. Ventajas y desventajas del muestreo sistemático

Las principales ventajas del muestreo sistemático son:

- Una ventaja del muestreo sistemático sobre el MAS es la distribución más homogenea a través de la ordenación de la población (en una lista, o en el tiempo, o direccionalmente en una dimensión, o espacialmente en dos dimensiones).
- Además una distribución más uniforme de la muestra a través de una población ordenada, es una ganancia en precisión.
- Es más fácil obtener la muestra y a menudo más fácil ejecutarlo sin errores.
- La muestra sistemática está distribuida más uniformemente sobre la población.
- Es preciso cuando las unidades dentro de una misma muestra son heterogéneas.

Las principales desventajas del muestreo sistemático son:

- No es un muestreo probabilístico.
- Por lo anterior, el "real" error de estimación es desconocido
- Puede provocar una pobre precisión cuando está presente una periocidad insospechada.
- Riesgo de sesgo.

Ejemplo:
Datos satelitales fueron usados para dividir la región de Aysén en $N = 5643$ "poligonos" de tierra relativamente homogenea en uso/cobertura. Estos poligonos tienen forma irregular, y de diferentes tamaños (medidos según su superficie). Luego, los poligonos fueron puestos en una lista ordenada desde el más pequeño al más grande, y aproximadamente $n = 50$ fueron sistemáticamente seleccionados (con un inicio aleatorio).
Por lo tanto, cada posible muestra sistemática tenia garantizada tener una aproximada similar mezcla de pequeños a medianos a grandes poligonos. Esta uniformidad de una a otra posible muestra sistemática es la característica que le permite una ganancia en precisión.
En este caso, el intervalo de muestreo a was

$$a = N/n = 5643/50 = 113 \quad (1)$$

1.2. ¿Por qué el muestreo sistemático no es un muestreo completamente probabilístico?

- El inicio aleatorio entre las primeras a unidades o elementos en el muestreo es lo que incluye lo probabilístico a la selección de la muestra.
- ¿Cómo seleccionar el primero? Seleccione un número uniformemente distribuido entre 0 y 1, luego multipliquelo por a, y trunquelo a un entero.
- Lo anterior, simbolicamente equivale a seleccionar la unidad U_k desde la población, donde k es determinado desde la fórmula

$$k = (a \times U[0,1] + 1)_{giv} \tag{2}$$

  Supongamos que $k = 7$, entonces U_7 es la primera unidad en la muestra. El orden de las otras unidades en la muestra es: $U_{7+a}, U_{7+2a}, U_{7+3a}, \cdots$ hasta que termine el n determinado.

  De igual forma, si U_{33} es la primera unidad de la muestra, las otras unidades seran $U_{33+a}, U_{33+2a}, U_{33+3a}, \cdots$.
- El caso trivial donde $a = 1$ corresponde a un censo
- La fracción de muestreo de 1–en–a muestra sistemática es $f = 1/a$. Por ejemplo, 1–en–20 muestras implica que una de cada 20 unidades de la población estarán en la muestra, y por lo tanto $f = 1/20$, o un 5 %.
- Para el muestreo aleatorio simple, determinamos que el número total posible de muestras de tamaño n en una población de N elementos estaba determinado por la fórmula de combinatoria. En general el resultado, como vimos, es un número muy grande, incluso para tamaños poblacionales bajos.
- Sin embargo en una muestra sistemática de 1–en–a, para lo cual $\Omega = a$ y U_i es un miembro de una sóla de las a posibles muestras sistemáticas.
- Ejemplo: supongamos que $N = 10$, y 1–en–3 muestras sistemáticas es usado.

  Una posible muestra es, digamos s_1, consiste de U_1, U_4, U_7, U_{10}.

  Otra posible muestra es, digamos s_2, consiste de U_2, U_5, U_8

  y finalmente, digamos s_3, consiste de U_3, U_6, U_9

No existe otra 1–en–3 muestra sistemática posible para esta población ordenada desde U_1 a U_{10}.

- Claramente podemos ver que la diferencia en Ω entre MAS y MSist es enorme para un mismo tamaño muestral n.
- Recordemos que la clase anterior vimos que cuando $N = 100$ hay $\Omega = 17{,}310{,}309{,}456{,}440$ posibles muestras MAS de tamaño muestral $n = 10$
- Sin embargo, el número de posibles muestras en MSist de tamaño $n = 10$ en una población de $N = 100$ es simplemente $\Omega = a = 10$
- Esto implica por lo tanto que el Msist no ofrece la misma posible cantidad de combinaciones de elementos en una muestra como en MAS. Y el tamaño de Ω esta directamente relacionado con la distribución aleatoria de los posibles estimadores de los parámetros de interes.
- Con un muestreo sistemático 1–en–a la probabilidad de inclusión de U_i es $\pi_i = 1/a$ para todos los $i = 1, 2, ..., N$.
- Tal como lo destaca **?**, en un muestreo 1–en–a sistemático es un diseño de muestreo de igual probabilidad, pero no necesariamente un diseño de n fijo. Por lo tanto, es un diseño de tamaño de muestreo aleatorio.

2. Estimación de parámetros

Para estimar μ_y se pueden emplear las siguientes alternativas

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{3}$$

o

$$\widehat{\mu}_{y,sis} = \frac{a}{N}\sum_{i=1}^{n} y_i = \frac{1}{E[N]}\sum_{i=1}^{n} y_i. \tag{4}$$

El estimador (3) será un estimador sesgado de μ_y, debido a la aleatoriedad del denominador n. En contraste (4) estimada insesgadamente μ_y.
Por otro lado, la precisión de $\overline{y}$ como un estimador de μ_y sera usualmente mayor que la precisión del estimador $\widehat{\mu}_{y,sis}$, i.e., $\text{Var}(\overline{y}) < \text{Var}(\widehat{\mu}_{y,sis})$.
De todas maneras el sesgo sera pequeño, y así también la diferencia en precisión entre los dos estimadores.
Denotemos ahora el total de la variable y en la muestra como t_S:

$$t_S = \sum_{i=1}^{n} y_i. \tag{5}$$

Mientras uno puede estimar τ_y con $N\overline{y}$ (como lo realizamos para el MAS), considere ahora lo siguiente en cambio

$$\widehat{\tau}_{y,sis} = N\widehat{\mu}_{y,sis} = a\ t_S. \tag{6}$$

$\widehat{\tau}_{y,sis}$ es un estimador insesgado de τ_y (**?**).
La varianza de $\widehat{\tau}_{y,sis}$ es

$$\text{Var}(\widehat{\tau}_{y,sis}) = \frac{1}{a}\sum_{S=1}^{a}(at_S - \tau_y)^2 \tag{7}$$

y la varianza de $\widehat{\mu}_{y,sis}$ es entonces

$$\text{Var}(\widehat{\mu}_{y,sis}) = \text{Var}(\widehat{\tau}_{y,sis})/N^2 \tag{8}$$

Note que las expresiones de las varianzas 7 y (8), son parámetros, sin embargo como estimar dichas varianzas es complejo en un muestreo sistemático.

Estimando la varianza de los estimadores en un muestreo sistemático

No existe un estimador insesgado de $\text{Var}\left(\widehat{\tau}_{y,sis}\right)$, sin embargo, existen alternativas que se explican a continuación.
(i) Lo más simple es utilizar el estimador de la varianza del MAS, que ya hemos visto antes, como sigue

$$\widehat{\text{Var}}\left(\widehat{\tau}_y\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\widehat{\sigma}_y^2 \tag{9}$$

(ii) La segunda alternativa es emplear el siguiente estimador de la varianza del estimador del total bajo un muestreo sistemático,

$$\widehat{\text{Var}}_{sd}\left(\widehat{\tau}_{y,sis}\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=2}^{n}\frac{\left(\Delta y_i\right)^2}{2(n-1)}, \tag{10}$$

donde Δy_i corresponde a diferencias sucesivas entre los elementos de la muestra, como sigue

$$\Delta y_i = \left(y_i - y_{i-1}\right), \tag{11}$$

y es por eso que el sufijo "sd" en (10) representa "successive differences"
En ambos estimadores de varianza, al dividirlos por N^2 se obtiene el respectivo estimador de la varianza del estimador $\widehat{\mu}_{y,sis}$. Por ejemplo,

$$\widehat{\text{Var}}_{sd}\left(\widehat{\mu}_{y,sis}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=2}^{n}\frac{\left(\Delta y_i\right)^2}{2(n-1)}, \tag{12}$$

3. Ejemplo de muestreo sistemático

Asumamos la siguiente población de número de plantas de *Eucalyptus globulus* atacada por el hongo *Botritis cinerea* en un vivero forestal.

LINEA	PLATABANDA						
(MTS)	A	B	C	D	E	F	G
1	81	80	99	70	65	91	72
2	90	79	82	86	62	97	91
3	99	84	80	87	54	51	67
4	98	87	89	80	90	100	54
5	83	73	96	86	53	92	56
6	74	95	96	85	96	94	77
7	95	98	76	97	58	70	78
8	82	76	76	75	86	71	51
9	85	94	72	84	51	79	66
10	81	91	70	79	86	67	51
11	100	99	95	94	87	53	79
12	97	83	98	75	65	72	73
13	79	96	82	96	60	86	63
14	71	90	81	83	96	51	86
15	90	97	83	97	96	100	97
16	84	95	77	89	65	77	93
17	74	99	72	98	69	80	97
18	93	87	89	78	76	68	65
19	97	71	81	88	78	88	62
20	90	82	83	85	66	63	92

En este caso $N = 140$ y empleemos un tamaño muestral $n = 28$. Tal como hemos visto anteriormente, el intervalo de muestreo se determina por $a = N/n = 140/28 = 5$.

Se inicia el muestreo sistemático en la platabanda A y se selecciona al azar entre el metro lineal 1 al 5. En este ejemplo fue seleccionado el número 4, por lo tanto el valor de la variable Y en ese metro lineal es seleccionado como el primer elemento de la muestra, por lo tanto $y_1 = 98$. El segundo elemento de la muestra es por lo tanto el noveno $(4 + a)$ elemento en la platabanda A ($y_2 = 85$) y así sucesivamente. Todos los elementos de la muestra sistemática se han representado en amarillo.

Por lo tanto el vector y con nuestros elementos de la muestra en R se representa como

```
> y <- c(98, 85, 71, 97, 87, 94,
     90, 71, 89, 72, 81, 81, 80,
     84, 83, 88, 90, 51, 96, 78,
     100, 79, 51, 88, 54, 66, 86,
     62)
> n <- length(y)
> N <- 140
> a <- 5
```

a. Ahora calculamos el estimador del parámetro de interes, mediante Eq. (4) como sigue,

$$\widehat{\mu}_{y,sis} = \frac{a}{N}\sum_{i=1}^{n} y_i$$

en R esto se puede calcular como sigue

```
> m.y <- (a/N) * sum(y)
> m.y
```

```
[1] 80.429
```

b. $\widehat{\mathrm{SE}}\,[\widehat{\mu}_y]$: Error estándar estimado de $\widehat{\mu}_y$

Recordemos que este error estándar se puede realizar mediante dos alternativas: usando el del muestreo aleatorio simple (MAS) o el de las diferencias sucesivas.

Procederemos primero a calcular este error asumiendo MAS. Para calcular la desviación estándar de la variable y en la muestra, definida como

$$\widehat{\sigma}_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \tag{13}$$

```
> var.y <- var(y)
> sd.y <- sd(y)
> sd.y
```

```
[1] 13.764
```

Ahora podemos calcular el error estándar estimado del estimador $\bar{y}$, asumiendo un muestreo aleatorio simple, el cual representamos por $\widehat{\mathrm{SE}}\,(\bar{y})$, empleando la siguiente expresión

$$\begin{aligned}\widehat{\mathrm{SE}}\,[\bar{y}] &= \sqrt{\frac{\widehat{\sigma}_y^2}{n}(1-f)} \\ &= \sqrt{\frac{\widehat{\sigma}_y^2}{n}\left(1-\frac{n}{N}\right)}\end{aligned} \tag{14}$$

y en R

```
> f <- n/N
> se.med.y <- sqrt((var.y/n) * (1 -
```

```
    f))
> se.med.y

[1] 2.3265
```

La unidad de medición es número de plantas atacadas por el hongo por metro lineal.

Ahora calculemos el error estándar pero usando la varianza de diferencias sucesivas (Eq. 12), que volvemos a escribir acá para representar la fórmula en R

$$\widehat{\mathrm{Var}}_{sd}\left(\widehat{\mu}_{y,sis}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=2}^{n} \frac{\left(\Delta y_i\right)^2}{2(n-1)}.$$

Para el cálculo de las diferencias sucesivas de esta fórmula emplearemos la función `diff()`, como sigue

```
> var.sd <- ((1/n) - (1/N)) * sum((diff(y))^2)/(2 *
    (n - 1))
> se.sd <- sqrt(var.sd)
> se.sd

[1] 2.4701
```

Como es de esperarse, los errores estándares son distintos. Para representar de mejor forma el diseño de muestreo realizado, se utilizará para los siguientes cálculos el error estándar derivado de las diferencias sucesivas, es decir 2,4701 plantas atacadas/metro lineal.

c. Error de muestreo o margen de error para un 90 % de confianza estadística.

El error de muestreo (EM) se calcula como sigue

$$\mathrm{EM}_{(\widehat{\mu}_{y,sis},\alpha)} = \widehat{\mathrm{SE}}\left[\widehat{\mu}_{y,sis}\right] t_{(1-\frac{\alpha}{2},n-1)}, \tag{15}$$

donde $t_{(1-\frac{\alpha}{2},n-1)}$ es el valor de la distribución de t-student empleando un nivel de significancia α. Para obtener el valor de t podemos ocupar R o una tabla de valores para la distribución de t-student

```
> df <- n - 1
> conf <- 90
> alpha <- 1 - (conf/100)
```

```
> alpha.2 <- alpha/2
> t.value <- abs(qt(1 - alpha.2,
    df))
> t.value
```

```
[1] 1.7033
```

Entonces $t_{(1-\frac{0,1}{2},28-1)} = 1{,}7033$. Ahora podemos aplicar (15) y calcular el error de muestreo

$$\text{EM}_{(\bar{y},0,1)} = 2{,}47 \times 1{,}7033 = 4{,}207 \tag{16}$$

este valor se expresa en las mismas unidades que la variable aleatoria y. Este error de muestreo se puede expresar en porcentaje al dividirlo por la media aritmética, como sigue

$$\text{EM}_{(\bar{y},0,1)}\,\% = 100 \times \frac{4{,}207}{80{,}429} = 5{,}2\,\% \tag{17}$$

d. Estimar un intervalo de confianza al 90 % estadístico para el estimador $\widehat{\mu}_{y,sis}$.

Un intervalo de confianza para $\widehat{\mu}_y$ se calcula como sigue

$$\widehat{\mu}_{y,sis} \pm \widehat{\text{SE}}\,[\widehat{\mu}_{y,sis}]\, t_{(1-\frac{\alpha}{2},n-1)} \tag{18}$$

$$\pm\, EM \tag{19}$$

lo cual es

$$80{,}429 \pm 4{,}207 \tag{20}$$

$$\pm\, 4{,}207 \tag{21}$$

$$[76{,}221\ ;\ 84{,}636], \tag{22}$$

Ejercicio:

- Calcule el error de muestreo para esta misma muestra, pero ahora asumiendo un error estándar del estimador como si fuera muestreo aleatorio simple.
- Realice un muestreo sistemático pero ahora empleando $a = 7$. Compare sus resultados con los obtenidos en este ejemplo.

Referencias

Gregoire TG, HT Valentine. 2008. Sampling Strategies for Natural Resources and the Environment. New York, USA. Chapman & Hall/CRC. 474 p.