

## Capítulo 1

# Muestras con probabilidades simples

Las muestras no están dadas, las muestras deben ser seleccionadas, asignadas o capturadas. El tamaño de la muestra no siempre es fijo. En estudios por muestreo, el tamaño de muestra es casi siempre una variable aleatoria. Los datos no siempre son independientes o idénticamente distribuidos y usualmente no son seleccionados de una sola población, sino de sub-poblaciones compuestas o complementarias. Más aún, no se produce una sola estimación, se produce un conjunto de estimaciones. Así que la historia que siempre nos han contado está equivocada.

Leslie Kish en Frankel & King (1996)

Cuando el marco de muestreo disponible para la selección de la muestra es una lista conteniendo la identificación y la ubicación de los elementos en la población, se utilizan diseños de muestreo que permitan la inclusión de éstos en la muestra de forma directa. Es decir, en la selección de la muestra, los elementos poblacionales son las mismas unidades de muestreo. Una vez que el procedimiento de muestreo ha seleccionado la muestra de elemento, el siguiente paso a realizar es la medición de la característica de interés  $y_k$  en cada elemento de la muestra seleccionada ( $k \in s$ ).

En este capítulo se describen los diseños de muestreo para elementos más importantes, algunos de los cuales son ampliamente utilizados en la práctica, otros tienen la característica de ser de tamaño de muestra variable o aleatorio. Cuando el marco de muestreo contiene información auxiliar de tipo continuo para cada elemento de la población, se utilizará esta información en la selección de la muestra, incluyendo los diseños proporcionales al tamaño. Cuando el marco de muestreo contiene información auxiliar discreta, se utilizarán diseños de muestra estratificados que permiten, a menudo, mayor precisión cuando la característica de interés presenta comportamientos diferentes en cada estrato o grupo poblacional.

Para cada diseño de muestreo se realiza una descripción teórica, se utilizará la población  $U$  para realizar algunos ejercicios léxico-gráficos que describan el comportamiento de la estrategia de muestreo. Por otro lado, se utilizará la población Lucy y, con ayuda del paquete **TeachingSampling**, se seleccionará una única muestra para la posterior estimación de los parámetros de interés. También habrá ejemplos prácticos de la vida real que permiten una mayor comprensión de las características del diseño y un mayor conocimiento a la hora de decidir qué diseño de muestreo debe ser implementado en determinados casos.

Las estrategias de muestreo implementadas en este capítulo corresponden a la utilización del estimador de Horvitz-Thompson junto con diseños de muestreo sin reemplazo y/o al uso del estimador de Hansen-Hurwitz en diseños de muestra con reemplazo.

## 1.1 Muestreo aleatorio simple sin reemplazo

El muestreo aleatorio simple puede ser visto como la forma más básica de selección de muestras. Supone la existencia de homogeneidad en los valores poblacionales de la característica de interés. Partiendo de esta asunción, este diseño provee probabilidades de selección idénticas para cada una de las posibles muestras pertenecientes al soporte  $Q$ . Lohr (2000) cita un ejemplo al respecto del uso del diseño de muestreo aleatorio simple diciendo que, cuando la población es homogénea, el investigador no necesita examinar todos los elementos de la población así como el encargado del análisis médico no necesita obtener toda la sangre para medir la cantidad de glóbulos rojos.

Una **muestra aleatoria simple sin reemplazo** de tamaño  $n$  se elige de modo que cada posible muestra realizada de tamaño  $n$  tenga la misma probabilidad de ser seleccionada. A diferencia del diseño de muestreo Bernoulli, el diseño de muestreo aleatorio simple sin reemplazo tiene la característica de ser de tamaño fijo. Una **muestra aleatoria simple con reemplazo**, de tamaño  $m$  de una población de  $N$  elementos es la extracción de  $m$  muestras independientes de tamaño 1, en donde cada elemento se extrae de la población con la misma probabilidad.

Lehtonen & Pahkinen (2003) afirman que este diseño de muestreo no es muy común en la práctica y básicamente desempeña dos funciones. Primero, plantean una línea de comparación de la eficiencia relativa con otros diseños de muestreo. Segundo, dentro de los diseños de muestreo más sofisticados como diseños de muestreo estratificado o diseños de muestreo por conglomerados, el muestreo aleatorio simple puede ser utilizado como un método final de selección de unidades primarias.

**Definición 1.1.1.** *Un diseño de muestreo se dice aleatorio simple sin reemplazo si todas las posibles muestras de tamaño  $n$  tienen la misma probabilidad de ser seleccionadas. Así,*

$$p(s) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{si } \#s = n \\ 0 & \text{en otro caso} \end{cases} \quad (1.1.1)$$

**Resultado 1.1.1.** *Definiendo a  $Q$  como el soporte que contiene a todas las posibles muestras de tamaño  $n$ , existen  $\binom{N}{n}$  muestras pertenecientes a  $Q$ . En otras palabras,*

$$\#(Q) = \binom{N}{n}$$

Nótese que  $\sum_{s \in Q} p(s) = 1$  porque  $\#Q = \binom{N}{n}$ .

### 1.1.1 Algoritmos de selección

Durante muchos años, la teoría de muestreo se centró en la parte de la extracción de muestras aleatorias, más que en la construcción de los estimadores. Con la gran ventaja de los nuevos procesadores, lo anterior pasa a un segundo plano. A continuación se presentan dos métodos de selección de una muestra aleatoria simple de tamaño  $n$  de una población de tamaño  $N$ . Existen bastantes métodos de selección de una muestra aleatoria sin reemplazo, en esta sección se abordan dos algoritmos de selección. El primero da una asunción más simple, y puede ser comparado con el conocido método de la extracción de una balota; sin embargo, Tillé (2006) afirma que este método es ineficiente computacionalmente. El segundo método basado en un algoritmo secuencial, permite la selección de la muestra con una sola revisión del marco de muestreo.

#### Método coordinado negativo

Sunter (1977) ha probado que el siguiente método de ordenamiento aleatorio arroja como resultado una muestra aleatoria simple. Para extraer la muestra de tamaño  $n$  de un universo de  $N$  objetos,

1. Generar  $N$  realizaciones de una variable aleatoria  $\xi_k$  ( $k \in U$ ) con distribución uniforme (0,1).
2. Asignar  $\xi_k$  al elemento  $k$ -ésimo de la población.
3. Ordenar la lista de elementos descendente (o ascendente) con respecto a este número aleatorio  $\xi_k$ .
4. A continuación, seleccionar los  $n$  primeros (o los  $n$  últimos) elementos. Esta selección corresponde a la muestra realizada.

Es necesario tener la seguridad de que exista un número grande de décimas en cada  $\xi_k$  para evitar problemas de empates (números aleatorios repetidos).

### Método de selección y rechazo

Fan, Muller & Rezucha (1962) implementaron el siguiente algoritmo de muestreo secuencial (porque se recorre el marco de muestreo, elemento por elemento, y se decide la pertenencia o el rechazo del objeto en la muestra). Es interesante que, más tarde Bebbington (1975) trece años más tarde publica (en un artículo de una página) el mismo método, aunque sin escribir ninguna fórmula.

En general se supone que el marco de muestreo tiene  $N$  individuos, y se quiere seleccionar una muestra aleatoria de  $n$  individuos. Así, para el individuo  $k$  ( $k = 1, 2, \dots, N$ ), se tiene que

1. Realizar  $\xi_k \sim U(0, 1)$

2. Calcular

$$c_k = \frac{n - n_k}{N - k + 1}$$

donde  $n_k$  es la cantidad de objetos seleccionados en los  $k - 1$  ensayos anteriores.

3. Si  $\xi_k < c_k$ , entonces el elemento  $k$  pertenece a la muestra.
4. Detener el proceso cuando  $n = n_k$ .

Dado que este algoritmo se detiene cuando  $n = n_k$ , resulta muy eficiente porque asegura una muestra aleatoria simple y en algunas ocasiones no se requiere recorrer todo el marco de muestreo.

**Ejemplo 1.1.1.** Para seleccionar muestras aleatorias simples, R incorpora la función `sample`. Ésta, por defecto selecciona muestras sin reemplazo. Así, por ejemplo, para seleccionar una muestra aleatoria de tamaño  $n = 2$ , de la población de ejemplo `U` de tamaño  $N = 5$ , sin reemplazo se tiene

```
N <- length(U)
sam <- sample(N, 2, replace=FALSE)
U[sam]

## [1] "Sharon" "Leslie"
```

El algoritmo de selección y rechazo está implementado en la función `S.SI` del paquete `TeachingSampling` cuyos argumentos son el tamaño de la población `N`, el tamaño de muestra deseado `n` y un vector de números aleatorios `e` que, por defecto, se asigna mediante la generación de `N` realizaciones de una variable aleatoria con distribución uniforme en el intervalo  $]0, 1[$ .

Para seleccionar una muestra aleatoria sin reemplazo de tamaño  $n = 2$  por el método de selección y rechazo, de la población de ejemplo `U` de tamaño  $N = 5$ , sólo basta digitar el siguiente código.

```

sam <- S.SI(N, 2)
U[sam]

## [1] "Erik"   "Sharon"

```

Nótese que el resultado de la función `S.SI` es un vector de índices, que aplicados al identificador resulta en una muestra seleccionada que está conformada por los elementos **Erik** y **Leslie**.

La siguiente salida muestra cada uno de los  $N=5$  pasos del algoritmo. Los números aleatorios que se utilizaron están en la columna llamada `ek` y los índices de la muestra seleccionada están en la columna `sam`.

k	Nombre	ek	ck	nk	sam
1	Yves	0.4938	0.4000000	0	0
2	Ken	0.7044	0.5000000	0	0
3	Erik	0.4585	0.6666667	1	3
4	Sharon	0.6747	0.5000000	1	0
5	Leslie	0.8565	1.0000000	2	5

**Resultado 1.1.2.** *El diseño de muestreo Bernoulli coincide con el diseño de muestreo aleatorio simple sin reemplazo cuando el tamaño de muestra se considera fijo e igual a n.*

*Demostración.* Utilizando las propiedades de la probabilidad condicional se tiene que

$$\begin{aligned} Pr(S = s | n(S) = n) &= \frac{Pr(S = s \text{ y } n(S) = n)}{Pr(n(S) = n)} \\ &= \frac{\pi^n (1 - \pi)^{N-n}}{\binom{N}{n} \pi^n (1 - \pi)^{N-n}} = \frac{1}{\binom{N}{n}} \end{aligned}$$

el cual coincide con la expresión (3.2.1). □

Una consecuencia inmediata del anterior resultado es que otro método de selección de muestras para un diseño de muestreo Bernoulli es escoger aleatoriamente el tamaño de muestra de acuerdo a una distribución binomial  $Bin(N, \pi)$  y luego seleccionar una muestra mediante uno de los anteriores algoritmos de selección de muestras aleatorias simples sin reemplazo (Tillé 2006).

### 1.1.2 El estimador de Horvitz-Thompson

**Resultado 1.1.3.** *Para un diseño de muestreo aleatorio simple, las probabilidades de inclusión de primer y segundo orden están dadas por:*

$$\pi_k = \frac{n}{N} \quad (1.1.2)$$

$$\pi_{kl} = \frac{n(n-1)}{N(N-1)} \quad (1.1.3)$$

respectivamente. La covarianza de las variables indicadoras está dada por

$$\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k \pi_l = -\frac{n}{N^2} \frac{(N-n)}{(N-1)} & \text{para } k \neq l \\ \pi_k (1 - \pi_k) = \frac{n(N-n)}{N^2} & \text{para } k = l \end{cases} \quad (1.1.4)$$

*Demostración.* Recurriendo a la definición de probabilidad de inclusión de primer orden, se tiene que

$$\begin{aligned}\pi_k &= \Pr(I_k(S) = 1) \\ &= \frac{\binom{1}{1} \binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}\end{aligned}$$

por otro lado,

$$\begin{aligned}\pi_k l &= \Pr(k \in S \text{ y } l \in s) \\ &= \Pr(I_k(S) = 1 \text{ y } I_l(S) = 1) \\ &= \Pr(I_k(S) = 1 | I_l(S) = 1) \Pr(I_l(s) = 1) \\ &= \frac{n-1}{N-1} \frac{n}{N} = \frac{n(n-1)}{N(N-1)}\end{aligned}$$

□

**Resultado 1.1.4.** Para un diseño de muestreo aleatorio simple, el estimador de Horvitz-Thompson del total poblacional  $t_y$ , su varianza y su varianza estimada están dados por:

$$\hat{t}_{y,\pi} = \frac{N}{n} \sum_S y_k \quad (1.1.5)$$

$$Var_{MAS}(\hat{t}_{y,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{yU}^2 \quad (1.1.6)$$

$$\widehat{Var}_{MAS}(\hat{t}_{y,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{yS}^2 \quad (1.1.7)$$

respectivamente, con

$$S_{yU}^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{y}_U)^2, \quad (1.1.8)$$

la **varianza poblacional** de la característica de interés en el universo  $U$  y con

$$S_{yS}^2 = \frac{1}{n-1} \sum_{k \in S} (y_k - \bar{y}_S)^2 \quad (1.1.9)$$

la **varianza muestral** de los valores de la característica de interés en la muestra aleatoria  $S$ . Además,  $\bar{y}_S = \frac{\sum_S y_k}{n}$ . Por otro lado, nótese que  $\hat{t}_{y,\pi}$  es insesgado para el total poblacional  $t_y$  de la característica de interés  $y$ , y que  $\widehat{Var}_{MAS}(\hat{t}_{y,\pi})$  es insesgado para  $Var_{MAS}(\hat{t}_{y,\pi})$ .

*Demostración.* Por el resultado anterior, tenemos

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} = \frac{N}{n} \sum_S y_k. \quad (1.1.10)$$

La demostración de las varianzas es inmediata al reemplazar las cantidades apropiadas en la expresión genérica del capítulo anterior y teniendo en cuenta que

$$\sum_{k \neq l} y_k y_l = \sum_k \sum_l y_k y_l - \sum_{k=l} y_k y_l = \left( \sum_U y_k \right)^2 - \sum_U y_k^2$$

De tal forma que,

$$\begin{aligned}
Var(\hat{t}_{y,\pi}) &= \frac{N^2}{n^2} Var \left( \sum_U I_k(s) y_k \right) \\
&= \frac{N^2}{n^2} \left( \sum_U Var(I_k(s)) y_k^2 + \sum \sum_{k \neq l} Cov(I_k(S), I_l(s)) y_k y_l \right) \\
&= \frac{N^2}{n^2} \left( \frac{n(N-n)}{N^2} \sum_U y_k^2 - \frac{n}{N^2} \frac{(N-n)}{(N-1)} \sum \sum_{k \neq l} y_k y_l \right) \\
&= \frac{(N-n)}{n} \left( \sum_U y_k^2 - \frac{1}{N-1} \sum \sum_{k \neq l} y_k y_l \right) \\
&= \frac{(N-n)}{n} \frac{1}{N-1} \left( (N-1) \sum_U y_k^2 - \left[ \left( \sum_U y_k \right)^2 - \sum_U y_k^2 \right] \right) \\
&= \frac{N(N-n)}{n} \frac{1}{N-1} \left( \sum_U y_k^2 - \frac{\left( \sum_U y_k \right)^2}{N} \right) \\
&= \frac{N^2}{n} \left( 1 - \frac{n}{N} \right) S_{yU}^2
\end{aligned}$$

Para demostrar el insesgamiento de la varianza estimada es suficiente demostrar que  $S_{yS}^2$  es insesgado para  $S_{yU}^2$ .

$$\begin{aligned}
E(S_{yS}^2) &= E \left( \frac{1}{n-1} \left[ \sum_S y_k^2 - n\bar{y}_S^2 \right] \right) \\
&= \frac{1}{n-1} \left( E \left[ \sum_S y_k^2 \right] - nE \left[ \frac{\hat{t}_{y,\pi}}{N} \right]^2 \right) \\
&= \frac{1}{n-1} \left( \frac{n}{N} \left[ \sum_U y_k^2 \right] - \frac{n}{N^2} E \left[ \hat{t}_{y,\pi} \right]^2 \right) \\
&= \frac{1}{n-1} \left( \frac{n}{N} \left[ \sum_U y_k^2 \right] - \frac{n}{N^2} \left[ \frac{N^2}{n} \left( 1 - \frac{n}{N} \right) S_{yU}^2 - t_y^2 \right] \right) \\
&= \frac{n}{n-1} \left( \frac{1}{N} \left[ \sum_U y_k^2 \right] - \frac{1}{n} \left( 1 - \frac{n}{N} \right) S_{yU}^2 - \frac{t_y^2}{N^2} \right) \\
&= \frac{n}{n-1} \left( \frac{N-1}{N} S_{yU}^2 - \frac{N-n}{nN} S_{yU}^2 \right) \\
&= S_{yU}^2
\end{aligned}$$

En donde se utilizó el hecho de que  $\bar{y}_S = \frac{\hat{t}_{y,\pi}}{N}$  y además

$$E(\hat{t}_{y,\pi})^2 = Var(\hat{t}_{y,\pi}) - t_y^2.$$

□

**Ejemplo 1.1.2.** Para nuestra población de ejemplo  $U$ , existen  $\binom{5}{2} = 10$  posibles muestras de tamaño  $n = 2$ . Realice el cálculo léxico-gráfico del estimador de Horvitz-Thompson y compruebe el insesgamiento y la varianza.

### 1.1.3 Estimación de la media poblacional

**Resultado 1.1.5.** Para un diseño de muestreo aleatorio simple, el estimador de Horvitz-Thompson para la media poblacional  $\bar{y}_U$ , su varianza y su varianza estimada están dados por:

$$\hat{y}_\pi = \frac{\hat{t}_{y,\pi}}{N} = \frac{\sum_S y_k}{n} = \bar{y}_S \quad (1.1.11)$$

$$Var_{MAS}(\hat{y}_\pi) = \frac{1}{N^2} Var(\hat{t}_{y,\pi}) = \left(1 - \frac{n}{N}\right) \frac{S_{yU}^2}{n} \quad (1.1.12)$$

$$\widehat{Var}_{MAS}(\hat{y}_\pi) = \frac{1}{N^2} Var(\hat{t}_{y,\pi}) = \left(1 - \frac{n}{N}\right) \frac{S_{ys}^2}{n} \quad (1.1.13)$$

respectivamente, con  $S_{yU}^2$  y  $S_{ys}^2$  el estimador de la varianza de los valores de la característica de interés  $y$  en el universo  $y$  en la muestra. Nótese que  $\hat{t}_{y,\pi}$  es insesgado para el total poblacional  $t_y$  de la característica de interés  $y$ , y que  $\widehat{Var}_{MAS}(\hat{t}_{y,\pi})$  es insesgado para  $Var_{MAS}(\hat{t}_{y,\pi})$ .

Nótese que la construcción, cálculo y estimación de la varianza son muy intuitivas. Haciendo un símil con la inferencia clásica, suponga que tenemos una muestra aleatoria  $X_1, \dots, X_n$  i.i.d., tal que  $X_i \sim (\mu, \sigma^2)$ . Se sabe que un estimador insesgado para la media  $\mu$  es  $\bar{X}$ , además se sabe que la variación de este estimador es  $\frac{\sigma^2}{n}$ .

Al operador  $\left(1 - \frac{n}{N}\right)$  se le conoce con el nombre de **factor de corrección para poblaciones finitas**. Sólo existe una sola muestra que contiene a todos los elementos de la población, por tanto, si esa muestra es seleccionada, esperamos que no haya variación en el estimador pues reproducirá con exactitud al parámetro, por tanto la varianza del mismo se debe anular. Entre más grande sea el tamaño de muestra  $n$ , al utilizar un diseño de muestreo aleatorio simple, la variabilidad de las estimaciones se debe hacer más pequeña dado que la muestra tenderá a parecerse más a la población finita. Lohr (2000) afirma que el tamaño de muestra es el que determina la precisión de las estimaciones (no así, el porcentaje de la población muestreada):

Si su sopa está bien revuelta, sólo necesita dos o tres cucharadas para probar el sazón, así tenga uno o veinte litros de sopa. Una muestra de tamaño  $n = 100$  de una población de  $N = 100\text{mil}$  elementos, tiene casi la misma precisión que una muestra de tamaño  $n = 100$  de una población de  $N = 100\text{millones}$  de elementos:

1. Para el primer caso,  $Var_{MAS}(\hat{y}_\pi) = \frac{99900}{100000} \frac{S_{yU}^2}{100} = 0.999 \frac{S_{yU}^2}{100}$
2. Para el último caso,  $Var_{MAS}(\hat{y}_\pi) = \frac{9999900}{100000000} \frac{S_{yU}^2}{100} = 0.999999 \frac{S_{yU}^2}{100}$

#### Tamaño de muestra

Bajo muestreo aleatorio simple sin reemplazo, un intervalo de confianza de  $100(1 - \alpha)\%$  para la media de la población es:

$$\left[ \bar{y}_S - z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_{yU}^2}{n}}, \bar{y}_S + z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_{yU}^2}{n}} \right] \quad (1.1.14)$$

y como usualmente no se conoce  $S_{yU}^2$ , lo usual es sustituirlo por el valor muestral  $S_{ys}^2$ . Por lo general, sólo los investigadores del estudio pueden decidir sobre la precisión mínima del mismo. Ésta se expresa como:

$$Pr(|\bar{y}_S - \bar{y}_U| \leq c) = 1 - \alpha$$

Por tanto, la cantidad a minimizar es  $c$ ,

$$c = z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_{yU}}{\sqrt{n}}} \quad (1.1.15)$$

y despejando  $n$ , se tiene:

$$n \geq \frac{n_0}{1 + \frac{n_0}{N}} \quad (1.1.16)$$

con  $n_0 = \frac{z_{1-\alpha/2}^2 S_{yU}^2}{c^2}$ . La desigualdad se tiene porque cuando se aumenta el tamaño de muestra,  $c$  decrece su valor. En algunas ocasiones se quiere lograr una precisión relativa dada por:

$$P\left(\left|\frac{\bar{y}_S - \bar{y}_U}{\bar{y}_U}\right| \leq c\right) = 1 - \alpha$$

que se puede escribir equivalentemente como:

$$P(|\bar{y}_S - \bar{y}_U| \leq c|\bar{y}_U|) = 1 - \alpha$$

la cantidad a minimizar es:

$$c|\bar{y}_U| = z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_{yU}}{\sqrt{n}}} \quad (1.1.17)$$

y despejando  $n$ , se tiene:

$$n \geq \frac{k_0}{1 + \frac{k_0}{N}} \quad (1.1.18)$$

con  $k_0 = \frac{z_{1-\alpha/2}^2 S_{yU}^2}{\bar{y}_U^2 c^2} = \frac{z_{1-\alpha/2}^2 C V^2}{c^2}$ . La desigualdad se tiene porque cuando se aumenta el tamaño de muestra,  $c|\bar{y}_U|$  decrece su valor.

Bajo un diseño aleatorio simple, un intervalo de confianza del  $100(1 - \alpha\%)$  para la media poblacional  $\bar{y}_U$  puede ser escrito como

$$\bar{y}_S(1 \pm A) \quad (1.1.19)$$

Donde  $A$  está dada por

$$A = z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_{ys}}{\sqrt{n} \bar{y}_S}} = z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{cv}{\sqrt{n}}} \quad (1.1.20)$$

Asumiendo que  $CV \doteq cv$  y que  $\frac{n}{N}$  es una cantidad despreciable, podemos determinar un tamaño de muestra para mantener una precisión dada. Por tanto  $A$  se reescribe como

$$A \doteq z_{1-\alpha/2} \frac{CV}{\sqrt{n}}$$

y despejando  $n$ , tenemos que

$$n \geq z_{1-\alpha/2}^2 \frac{CV^2}{A^2}$$

Con un nivel de confianza del  $\alpha = 5\%$ , asumiendo que el coeficiente de variación estimado converge al coeficiente de variación poblacional y que la fracción de muestreo es despreciable para obtener una precisión  $A < 3\%$  si a)  $CV = 0.5$ , el tamaño de muestra debe ser mayor que 1067 unidades; b)  $CV = 1.0$ , el tamaño de muestra debe ser mayor que 4268 unidades y c)  $CV = 1.5$ , el tamaño de muestra debe ser mayor que 9604 unidades. Es decir, entre más dispersa sea la población, con respecto a la media, mayor debe ser el tamaño de muestra para conseguir una precisión dada.

Para poder utilizar las anteriores fórmulas es necesario contar un buen tamaño de muestra, dado que el teorema central del límite clásico (universo infinito) no es el mismo que se ha aplicado aquí. Hájek (1960) demuestra que al utilizar muestreo aleatorio simple (universo finito) y bajo ciertas condiciones de regularidad conocidas como las condiciones de Noether y si  $n$ ,  $N$ , y  $N - n$  son grandes, es decir la fracción muestral  $f = n/N$  se aleja de 0 y de 1, entonces

$$\frac{\bar{y}_S - \bar{y}_U}{\sqrt{(1 - \frac{n}{N}) \frac{S_{yU}}{\sqrt{n}}}} \longrightarrow Normal(0, 1)$$

Cuando se quiere establecer un intervalo de confianza, la confiabilidad del intervalo está garantizada por el insesgamiento del estimador de Horvitz-Thompson. Para asegurar determinada precisión es necesario conocer la varianza poblacional de la característica de interés o el coeficiente de variación del estimador; en estos términos, cuando el coeficiente de variación estimado (cve) es menor del 3% es un caso excelente; entre el 3 y el 5% es bueno; entre el 5 y el 10% es regular; entre el 10 y 15% es apenas presentable; si es más del 15% no es considerado bueno; en este caso algunas agencias de estadísticas oficiales no presentan el coeficiente de variación, aunque se conozca.

Por supuesto, algunas cantidades poblacionales necesarias para estimar el tamaño de muestra no se conocen; de hecho, si se conocieran, no habría necesidad de realizar estudio alguno, porque directamente se conocerían los parámetros poblacionales de interés. Lohr (2000) considera tres escenarios para realizar una estimación previa de los parámetros de interés:

1. Realizar una **prueba piloto**, unas cuantas entrevistas conforman la muestra piloto, seleccionada con el mismo diseño de muestreo genérico. En algunas ocasiones, este método además de servir para estimar las cantidades necesarias para establecer el tamaño de muestra, sirve para confrontar y calibrar el instrumento de medición, ya sea un cuestionario o un instrumento técnico.
2. Utilizar información a priori de estudios anteriores. No siempre el investigador que realiza un estudio por muestreo ha sido el primero en cuestionarse acerca de los objetivos de la investigación. Si esto es así, existen referencias bibliográficas disponibles, en donde se pueden hallar estimaciones de la varianza poblacional o del error estándar. Esta última medida tiende a ser más estable contra el tiempo o posición geográfica.
3. Estimar la varianza ajustando una distribución teórica a la característica de interés. Ospina (2001) afirma que este ajuste se hace con base en supuestos adecuados acerca de la estructura poblacional de la característica de interés (normal, exponencial, uniforme, etc.). La identificación de una distribución apropiada permite hacer uso de sus propiedades para obtener una estimación más realista de la varianza. Cuando el desconocimiento es absoluto, se recomienda utilizar la distribución uniforme. Wu (2003) afirma que las características de interés en poblaciones económicas son sesgadas a la derecha y tienden a ser modeladas mediante distribuciones como la Gamma o la Ji cuadrado.

#### 1.1.4 Estimación en dominios

El primer caso concerniente a la estimación de subgrupo poblacionales es el de las sub-poblaciones llamadas dominios. En muchas investigaciones es necesario llevar a cabo estimaciones sobre la población

en general, y también sobre subgrupos de ella (denominados dominios por la subcomisión en muestreo de las Naciones Unidas). La identificación de los dominios se logra una vez la información de los elementos ha sido registrada. Los dominios tienen que cumplir las siguientes características:

1. Ningún elemento de la población puede pertenecer a dos dominios.
2. Todo elemento de la población debe pertenecer a un único dominio.
3. La reunión de todos los dominios es la población del estudio.

Por ejemplo, al estimar el total de la fuerza laboral en empresas con menos de dos años de funcionamiento. Claramente la población se divide en dos dominios; el primero concerniente a las empresas con menos de dos años de funcionamiento y el segundo dado por las empresas con dos años o más de funcionamiento.

**Definición 1.1.2.** Un dominio  $U_d$  es una sub-población específica o subgrupo poblacional que cumple las siguientes condiciones:

1.  $U_d \subset U$ , tal que  $U = \bigcup_{d=1}^D U_d$
2. Si  $k \in U_l$ , entonces  $k \notin U_d$  para  $d \neq l$
3. El número de elementos en el dominio  $U_d$  es  $N_d$  y es llamado **tamaño absoluto** del dominio.
4. La proporción de elementos en el dominio  $U_d$  con respecto al tamaño poblacional es  $P_d = \frac{N_d}{N}$  y se conoce como **tamaño relativo** del dominio.

La estimación por dominios se caracteriza por el desconocimiento de la pertenencia de las unidades poblacionales al dominio. Es decir, para conocer cuáles unidades de la población pertenecen al dominio, es necesario realizar el proceso de medición.

Fue Hartley (1959) quien desarrolló y unificó la teoría de la estimación en dominios aplicable a cualquier diseño de muestreo. Durbin (1967) obtuvo resultados similares. Las pautas para la estimación en dominios se dan a continuación: para estimar el total de un dominio  $U_d$ , dado por

$$t_{yd} = \sum_{U_d} y_k \quad (1.1.21)$$

es necesario, en primer lugar construir una función indicadora  $z_{dk}$ , para cada elemento de la población, de la pertenencia del elemento al dominio, dada por la siguiente definición.

**Definición 1.1.3.** Sea  $z_{dk}$  la función indicatriz del dominio  $U_d$ , dada por

$$z_{dk} = \begin{cases} 1 & \text{si } k \in U_d \\ 0 & \text{en otro caso} \end{cases} \quad (1.1.22)$$

Ahora, al multiplicar la variable de pertenencia  $z_{dk}$  por el valor de la característica de interés  $y_k$ , se crea una nueva variable  $y_{dk}$  dada por  $y_{dk} = z_{dk}y_k$ , y una vez construida se pueden utilizar los principios del estimador de Horvitz-Thompson para hallar un estimador insesgado del total de la característica de interés en el dominio  $U_d$ .

**Resultado 1.1.6.** El total de la variable de interés en el dominio  $U_d$  está dado por

$$t_{yd} = \sum_U y_{dk}, \quad (1.1.23)$$

el tamaño del dominio  $U_d$  toma la siguiente expresión

$$N_d = \sum_U z_{dk}, \quad (1.1.24)$$

de tal forma que la media de la característica de interés en el dominio  $U_d$  se escribe como

$$\bar{y}_{U_d} = \frac{t_{yd}}{N_d} = \frac{\sum_U y_{dk}}{N_d} \quad (1.1.25)$$

### Estimación del total en un dominio

**Resultado 1.1.7.** Bajo muestreo aleatorio simple sin reemplazo, el estimador de Horvitz-Thompson para el total del dominio  $t_{yd}$ , su varianza y su varianza estimada están dados por

$$\hat{t}_{yd,\pi} = \frac{N}{n} \sum_S y_{dk} = \frac{N}{n} \sum_{S_d} y_k \quad (1.1.26)$$

$$Var(\hat{t}_{yd,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{y_d U}^2 \quad (1.1.27)$$

$$\widehat{Var}(\hat{t}_{yd,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{y_d S}^2 \quad (1.1.28)$$

respectivamente, donde  $S_d = U_d \cap S$  se refiere al conjunto formado por la intersección de la muestra  $S$ . Además,

$$S_{y_d U}^2 = \frac{1}{N-1} \left( \sum_{k \in U} y_{dk}^2 - \frac{(\sum_{k \in U} y_{dk})^2}{N} \right)$$

representa la varianza poblacional de la característica de interés y

$$S_{y_d S}^2 = \frac{1}{n-1} \left( \sum_{k \in S} y_{dk}^2 - \frac{(\sum_{k \in S} y_{dk})^2}{n} \right)$$

la varianza muestral de los valores de la característica de interés.

Nótese que en la expresión  $S_{y_d U}^2$  los valores que intervienen son los de la característica de interés si el elemento pertenece al dominio y ceros si el elemento no pertenece al dominio, lo mismo sucede con  $S_{y_d S}^2$ . Por tanto, las anteriores expresiones van a tomar valores grandes por la inclusión de los ceros; éste es el precio que se debe pagar por el desconocimiento de la pertenencia de los elementos a los dominios.

### Estimación del tamaño absoluto de un dominio

**Resultado 1.1.8.** Bajo muestreo aleatorio simple sin reemplazo, el estimador de Horvitz-Thompson para el tamaño absoluto de un dominio  $N_d$ , su varianza y su varianza estimada están dados por

$$\hat{N}_{d,\pi} = \frac{N}{n} \sum_S z_{dk} = \frac{N}{n} \sum_{S_d} z_k \quad (1.1.29)$$

$$Var(\hat{N}_{d,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{z_d U}^2 \quad (1.1.30)$$

$$\widehat{Var}(\hat{N}_{d,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{z_d S}^2 \quad (1.1.31)$$

respectivamente, con  $S_{z_d U}^2$  y  $S_{z_d S}^2$  la varianza poblacional y la varianza muestral de los valores de la característica de interés  $z_{dk}$ .

Nótese que en la expresión  $S_{z_d U}^2$  los valores que intervienen son unos si el elemento pertenece al dominio y ceros si el elemento no pertenece al dominio, lo mismo sucede con  $S_{y_d s}^2$ .

### Estimación del tamaño relativo de un dominio

**Resultado 1.1.9.** *Bajo muestreo aleatorio simple sin reemplazo, el estimador de Horvitz-Thompson para el tamaño relativo de un dominio  $P_d$ , su varianza y su varianza estimada están dados por*

$$\hat{P}_{d,\pi} = \frac{1}{N} \sum_S \frac{N}{n} z_{dk} = \frac{1}{n} \sum_S z_{dk} = \frac{n_d}{n} \quad (1.1.32)$$

$$Var(\hat{P}_{d,\pi}) = \frac{1}{n} \left(1 - \frac{n}{N}\right) S_{z_d U}^2 \quad (1.1.33)$$

$$\widehat{Var}(\hat{P}_{d,\pi}) = \frac{1}{n} \left(1 - \frac{n}{N}\right) S_{z_d s}^2 \quad (1.1.34)$$

respectivamente, con  $S_{z_d U}^2$  y  $S_{z_d s}^2$  el estimador de la varianza de los valores de la característica de interés  $y_d$  en el universo y en la muestra.

### Estimación de la media de un dominio

**Resultado 1.1.10.** *Bajo muestreo aleatorio simple sin reemplazo, el estimador de Horvitz-Thompson para la media de la característica de interés en un dominio  $\bar{y}_{U_d}$ , su varianza y su varianza estimada están dados por*

$$\hat{y}_{U_d,\pi} = \frac{\frac{N}{n} \sum_S y_{dk}}{N_d} \quad (1.1.35)$$

$$Var(\hat{y}_{U_d,\pi}) = \frac{1}{N_d^2} \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{y_d U}^2 \quad (1.1.36)$$

$$\widehat{Var}(\hat{y}_{U_d,\pi}) = \frac{1}{N_d^2} \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{y_d s}^2 \quad (1.1.37)$$

Para poder utilizar el anterior estimador, es necesario conocer de antemano el valor del tamaño absoluto del dominio  $N_d$ . En la práctica, pocas veces se conoce este valor, por lo tanto un estimador alternativa y completamente intuitivo de la media de la característica de interés en un dominio es la media muestral de la misma en el dominio de interés. De tal forma que el estimador alternativo, toma la siguiente expresión

$$\hat{y}_{S_d} = \frac{\hat{t}_{y_d,\pi}}{\hat{N}_{d,\pi}} = \frac{\sum_S y_{dk}}{z_{dk}} = \frac{\sum_{S_d} y_k}{n_d} \quad (1.1.38)$$

Como las dos cantidades en el numerador y denominador son aleatorias, se está estimando una razón, de tal manera que el cálculo y estimación de la varianza del anterior estimador están fuera del alcance de este capítulo, y serán explicados en los lugares donde sea conveniente.

### 1.1.5 Marco y Lucy

Una de las razones por las que el gobierno realiza la encuesta de crecimiento económico del sector industrial es, no sólo para medir el impacto social e impositivo sino para buscar nuevas estrategias de crecimiento enfocadas en las empresas que conforman este sector. Recientemente, con el boom de la tecnología y el uso masivo de internet, las estrategias de mercadeo han cambiado su forma y su fondo.

Hace unos años, las empresas con un rendimiento muy alto, catalogadas dentro de un nivel industrial grande, podían acceder a pautar un comercial discreto de 900 TRP's<sup>1</sup> en televisión, mientras que las empresas medianas tenían un presupuesto con el cual apenas podían pautar un comercial en la radio. Por supuesto, la estrategia publicitaria de las empresas pequeñas consistía en editar un aviso en las páginas amarillas.

Sin embargo, a medida que cambia y evoluciona la tecnología, también lo hacen los hábitos de las personas. Es muy común que las operaciones financieras, contables y estratégicas de una empresa estén centradas en un servidor conectado a internet. La misma comunicación verbal ha sido reemplazada por altos estándares de tecnología mediante conversaciones virtuales, la comunicación oficial ha desplazado el casillero de correo postal por el correo electrónico que permite la recepción en tiempo real de mensajes sin importar la ubicación espacio temporal del receptor ni de la persona que envía el mensaje. Siendo así, las personas pasan más tiempo frente a un computador que frente al televisor, o escuchando la radio; las páginas amarillas están siendo reemplazadas por los meta-buscadores de la red mundial de información, gigantes como Google, Yahoo y MSN.

Los gerentes de mercadeo (en los casos pertinentes) junto con los presidentes o gerentes de las empresas del sector industrial, han replanteado sus viejas estrategias publicitarias y han hecho, poco a poco, la migración de canal publicitario. Las empresas grandes siguen pautando en televisión, las empresas medianas siguen haciéndolo en la radio y las pequeñas siguen teniendo el mismo viejo aviso clasificado en la sección de las páginas amarillas. Sin embargo, en todos los niveles del sector industrial, se ha empezado a realizar una mejor gestión de sus clientes y/o de sus potenciales clientes.

Las empresas están utilizando listas de correo electrónico masivas para dar a conocer las ventajas competitivas de sus empresas, mediante el envío de portafolios virtuales de los productos y servicios que brindan. Se cree que esta práctica de mercadeo ha aumentado la productividad empresarial porque por medio de la publicidad por internet o SPAM, las empresas consiguen más clientes, por lo tanto consiguen más contratos, por tanto ayudan a la disminución del desempleo y obtienen ventajas fiscales.

El gobierno quiere corroborar esta hipótesis y dependiendo de los resultados del estudio implementar un programa de capacitación gratuita a las empresas que aún no han entrado en el ámbito de la información mediante el uso masivo de la red informática internet. El presupuesto del gobierno es de unos cuantos millones de dólares, por lo tanto se necesitan estimaciones muy precisas que respondan al objetivo de la investigación.

### Estimación del tamaño de muestra

La estrategia de muestreo que se va a utilizar es la siguiente: el estimador de Horvitz-Thompson aplicado a un diseño de muestreo aleatorio simple sin reemplazo. Se selecciona una muestra piloto de tamaño 30 de la población. Para esto, una vez cargado el archivo de datos Lucy, utilizamos la función `sample` para extraer la muestra piloto. Como la característica de interés es el ingreso de las empresas, tomamos los valores de la varianza y de la media como estimaciones que servirán para el cálculo del tamaño de la muestra.

```
data(BigLucy)
attach(BigLucy)

N <- dim(BigLucy)[1]
sam <- sample(N, 100)
Inc.pilot <- Income[sam]

mean(Inc.pilot)
```

---

<sup>1</sup>Puntos acumulados de rating del grupo objetivo obtenidos considerando sólo consumidores viendo el comercial de televisión de una marca dada

```
## [1] 441
var(Inc.pilot)
## [1] 67280
```

Los valores que se utilizarán en la estimación del tamaño de muestra son la varianza muestral igual a 66.952, el promedio muestral igual a 455; con estos valores se tiene una estimación del coeficiente de variación igual a 0,57. Se debe escoger un tamaño de muestra que proporcione estimaciones precisas, el tamaño de muestra depende de la precisión que se requiera para cumplir con los objetivos del estudio.

- Error absoluto: el margen de error para este estudio es de 25 millones de dólares.
- Nivel de confianza del 95 %.
- Mediante (1.1.16) se tiene que  $n_0 = 411$ .
- Al utilizar el factor de corrección de poblaciones finitas, llegamos a que  $n \geq 351$ .

Sin embargo, este cálculo se puede cotejar restringiendo las estimaciones mediante un error relativo.

- Error relativo: se requieren estimaciones con menos del 7 % de error.
- Nivel de confianza del 95 % y una estimación de  $CV = 0.57$ .
- Mediante (1.1.18) se tiene que  $k_0 = 446$ .
- Al utilizar el factor de corrección de poblaciones finitas, llegamos a que  $n \geq 376$ .

Suponga que mediante fuentes oficiales se ha tenido acceso a información de estudios pasados que han modelado la característica de interés `Income` utilizando la familia de distribuciones Gamma con parámetro de forma 2,7 y parámetro de escala 180. Haciendo una simulación de  $N = 2396$  valores provenientes de una distribución gamma con los anteriores parámetros, se pueden estimar los valores de la varianza para la característica de interés y así una estimación del tamaño de muestra.

```
bary <- mean(Income)
sdy <- sd(Income)
x <- seq(min(Income), max(Income), by=10)
a <- 2.7
b <- 180
```

La determinación del tamaño de muestra para esta investigación utilizando la estrategia de muestreo mencionada al principio de la sección y consideraciones respecto a que la estimación de la varianza de la muestra piloto puede ser pequeña, da como resultado una muestra de tamaño  $n = 400$  empresas del sector industrial. Como el tamaño de la población es  $N = 2396$ , entonces el valor de la probabilidad de inclusión para todos los elementos es de  $\pi_k = \frac{400}{2396} \cong 0.17$ .

R incorpora la función `sample` para la selección de muestras con o sin reemplazo. En este caso puede ser utilizada como en la selección de la muestra piloto. Sin embargo, para seleccionar una muestra mediante el algoritmo de selección y rechazo, el paquete `TeachingSampling` adjunta la función `S.SI` que se utilizará en la selección de 400 empresas del sector industrial.

Primero se carga en R el archivo `Marco` que contiene el marco de muestreo para la selección de la muestra. Se fijan los parámetros de la función, `N` y `pik`. Esta función devuelve un vector contenido

```
hist(Income, freq=FALSE, breaks=10)
lines(x, dnorm(x, bary, sdy), col=2)
```

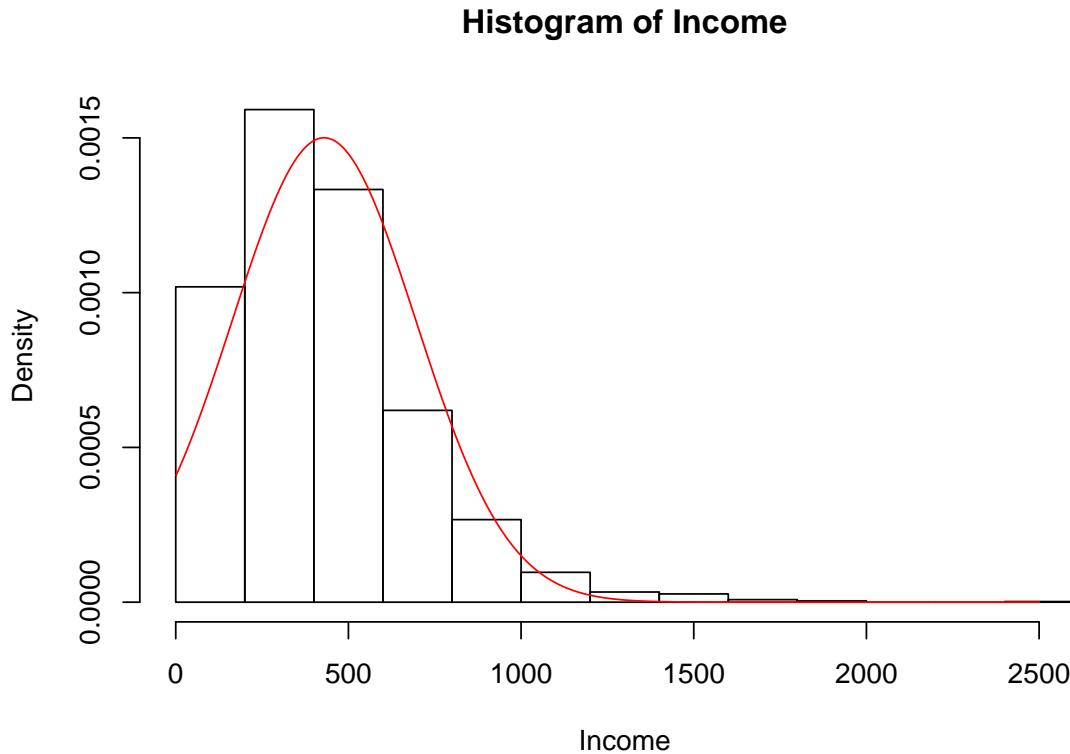


Figura 1.1: Distribución de la característica *Income* y su posible modelamiento bajo la distribución gamma.

el índice de los elementos seleccionados en la muestra. En este caso particular, el primer elemento seleccionado es el número 7 y el último el número 2395.

```
data(BigLucy)
attach(BigLucy)
N <- dim(BigLucy)[1]
n <- 2000
sam <- S.SI(N,n)
muestra <- BigLucy[sam,]
attach(muestra)
```

```
head(muestra)

##           ID Ubication Level   Zone Income Employees Taxes
## 12 AB0000000012 C0033329K0268568 Small County1     419      20     7
## 89 AB0000000089 C0016430K0285467 Small County1     491      26    10
## 150 AB0000000150 C0241162K0060735 Small County1    384      70     6
```

```

## 177 AB0000000177 C0063734K0238163 Small County1    319      55     4
## 193 AB0000000193 C0178986K0122911 Small County1    350      48     5
## 221 AB0000000221 C0158483K0143414 Small County1    295      57     3
##          SPAM ISO Years Segments
## 12      no   no  41.5 County1 2
## 89      no   no  20.3 County1 9
## 150     no   no  21.7 County1 15
## 177     yes  no   3.1 County1 18
## 193     yes  no   3.7 County1 20
## 221     no   no 13.0 County1 23

n <- dim(muestra)[1]
n

## [1] 2000

```

Aplicando los índices obtenidos por la función **S.SI** al marco de muestreo obtenemos la identificación y ubicación de las empresas seleccionadas en la muestra. Una vez que la etapa de recolección de datos se haya realizado; es decir, la medición de todos y cada uno de los elementos seleccionados ya ha sido realizada, se realiza la estimación. Obtendremos un archivo de datos de **Lucy** conteniendo los valores de las características de interés para las empresas seleccionadas que será adjuntado a R mediante la función **attach**.

La etapa de estimación de resultados se hace utilizando la función **E.SI(N,n,y)** del paquete **TeachingSampling** cuyos argumentos son **y**, un vector conteniendo los valores de la característica de interés en la muestra, **N** el tamaño de la población y **n** el tamaño de la muestra seleccionada. En este caso la longitud de cada vector es de  $n = 400$ . Esta función arroja la estimación del total poblacional de **y** usando el estimador de Horvitz-Thompson, la estimación de la varianza y el coeficiente de variación del mismo. Por ejemplo, la variable **Income** dentro del objeto **estima** contiene los valores del ingreso declarado en el último año por 400 empresas del sector industrial pertenecientes a la muestra. La estimación para esta característica se hace mediante el siguiente código:

```

estima <- data.frame(Income, Employees, Taxes)
E.SI(N,n,estima)

```

Cuadro 1.1: *Estimaciones para el diseño de muestreo aleatorio simple sin reemplazo*

	N	Income	Employees	Taxes
Estimation	85296.00	37120648.61	5436212.62	1031335.26
Standard Error	0.00	503724.19	61441.76	30446.62
CVE	0.00	1.36	1.13	2.95
DEFF		1.00	1.00	1.00

La tabla 1.1 muestra los resultados obtenidos para este caso particular. Nótese que se obtienen mejores resultados que al utilizar un diseño de muestreo Bernoulli. Sin embargo, comparar estos resultados de ingreso total en el sector industrial con el de las mediciones pasadas, no es suficiente y se desea tener estimaciones para el dominio o subgrupo de las empresas que utilizan el envío de SPAM como estrategia publicitaria.

La función **Domains** contenida en el paquete **TeachingSampling** es utilizada para obtener las variables indicadoras  $z_{dk}$  para cada dominio, el único argumento de la función es un vector de pertenencia de cada individuo. En este caso, el vector de pertenencia es **SPAM**, la salida de esta función es una matriz

de unos y ceros, en donde cada columna está dicotomizada. Existen tantas columnas como subgrupos poblacionales, y en cada columna el número uno implica la pertenencia del elemento al dominio y cero la no pertenencia del elemento al dominio.

```
Dominios <- Domains(SPAM)
head(Dominios)

##      no yes
## [1,]  1   0
## [2,]  1   0
## [3,]  1   0
## [4,]  0   1
## [5,]  0   1
## [6,]  1   0
```

Para estimar el tamaño absoluto de cada dominio, lo único que se debe hacer es multiplicar la matriz de características de interés (en este caso, la matriz llamada `estima`) por cada columna de la matriz resultante de la dicotomización. La siguiente salida lo muestra claramente para el dominio de la población que sí utiliza el SPAM como método publicitario.

```
SPAM.si <- Dominios[,2]*estima
head(SPAM.si)

##    Income Employees Taxes
## 1      0          0     0
## 2      0          0     0
## 3      0          0     0
## 4    319         55     4
## 5    350         48     5
## 6      0          0     0
```

Mientras que para el dominio que no utiliza el SPAM se tiene la siguiente salida

```
SPAM.no <- Dominios[,1]*estima
head(SPAM.no)

##    Income Employees Taxes
## 1    419        20     7
## 2    491        26    10
## 3    384        70     6
## 4      0          0     0
## 5      0          0     0
## 6    295        57     3
```

Utilizando la función `E.SI` en la matriz resultante de la dicotomización obtenemos las estimación de los tamaños absolutos de cada dominio. En este caso, se estima que 1420 empresas ya están utilizando otras técnicas radicales de publicidad, mientras que las restantes 976 no lo hacen. Nótese que la varianza de cada estimación es la misma, esto es claro porque los valores de esta característica de interés son ceros y uno y por tanto la estructura de varianza resulta idéntica en cada caso.

```
E.SI(N,n,Dominios)
```

	N	no	yes
## Estimation	85296	34758.1	50537.9
## Standard Error	0	926.4	926.4
## CVE	0	2.7	1.8
## DEFF	NaN	1.0	1.0

Está claro que existe una tendencia en el sector industrial de publicidad virtual mediante el envío de SPAM por correo electrónico. Las siguientes cifras son las verdaderamente importantes pues muestran que las empresas que utilizan SPAM tienen mayores ingresos, emplean a más gente y contribuyen con una mayor cantidad de dinero en cuanto a impuestos se refiere, esto se da porque hay más empresas que utilizan el SPAM de las que no lo hacen.

```
E.SI(N, n, SPAM.no)
E.SI(N, n, SPAM.si)
```

Como  $N_d$  es desconocido, podemos utilizar el estimador alternativo dado por la expresión (3.2.38), para obtener una estimación (aunque no la varianza ni el c.v.e) de la media de la característica de interés en cada dominio. Simplemente tomamos las estimaciones  $t_{yd}$  y las dividimos por la estimación de  $N_d$ . Las siguientes tablas resumen las estimaciones para cada uno de los dominios de interés<sup>2</sup>.

Cuadro 1.2: *Estimaciones para el diseño de muestreo aleatorio simple en el dominio que no envía SPAM*

	N	Income	Employees	Taxes
Estimation	85296.00	15146479.85	2225543.23	412064.98
Standard Error	0.00	508712.72	71010.05	20887.82
CVE	0.00	3.36	3.19	5.07
DEFF		1.00	1.00	1.00

Cuadro 1.3: *Estimaciones para el diseño de muestreo aleatorio simple en el dominio que sí envía SPAM*

	N	Income	Employees	Taxes
Estimation	85296.00	21974168.76	3210669.38	619270.28
Standard Error	0.00	565808.43	75591.61	27203.24
CVE	0.00	2.57	2.35	4.39
DEFF		1.00	1.00	1.00

### 1.1.6 Probabilidades de inclusión en unidades de muestreo

En Särndal, Swensson & Wretman (1992) se considera una encuesta para medir los ingresos de los hogares. El marco de muestreo es una lista de individuos y una muestra de tamaño  $n$  se selecciona mediante muestreo aleatorio simple sin reemplazo, el hogar correspondiente al individuo es identificado y se procede a realizar la medición correspondiente. La probabilidad de inclusión de un hogar  $h$  compuesto por  $M < n$  individuos, puede modelarse por medio de la distribución hipergeométrica, así:

<sup>2</sup>Nótese que el anterior procedimiento asegura la estimación de los parámetros de dominios no sólo en MAS sino para cualquier diseño de muestreo.

$$\begin{aligned}
\pi_H &= \Pr(H \in s) \\
&= 1 - \Pr(H \notin s) \\
&= 1 - \Pr(\text{Ninguno de los } M \text{ salió en la muestra de tamaño } n) \\
&= 1 - \frac{\binom{M}{0} \binom{N-M}{n}}{\binom{N}{n}} \\
&= 1 - \frac{(N-M)!/n!(N-M-n)!}{N!/(N-M)!n!} \\
&= 1 - \frac{(N-M)!}{N!} \frac{(N-n)!}{(N-M-n)!} \\
&= 1 - \frac{(N-n) \dots (N-n-M+1)}{N \dots (N-M+1)}
\end{aligned}$$

Asumiendo que  $N$  y  $n$  son grandes ( $f > 0$ ), se obtienen las siguientes aproximaciones:

- $M = 1$ ,

$$\begin{aligned}
\pi_H &= 1 - \frac{N-n}{N} \\
&= 1 - \left(1 - \frac{n}{N}\right) = 1 - (1-f)
\end{aligned}$$

- $M = 2$ ,

$$\begin{aligned}
\pi_H &= 1 - \frac{(N-n)(N-n-1)}{N(N-1)} \\
&= 1 - \left(1 - \frac{n}{N}\right) \left(1 - \frac{n}{N-1}\right) \doteq 1 - (1-f)^2
\end{aligned}$$

- $M = 3$ ,

$$\begin{aligned}
\pi_H &= 1 - \frac{(N-n)(N-n-1)(N-n-2)}{N(N-1)(N-2)} \\
&= 1 - \left(1 - \frac{n}{N}\right) \left(1 - \frac{n}{N-1}\right) \left(1 - \frac{n}{N-2}\right) \doteq 1 - (1-f)^3
\end{aligned}$$

## 1.2 Diseño de muestreo Bernoulli

En el diseño de muestreo Bernoulli se fija a priori (por experiencia o alguna otra razón) la probabilidad de inclusión de todos los individuos, la cual permanece constante para todo el universo. Es decir,  $\pi_k = \pi$  para todo  $k \in U$ . Un típico ejemplo de la implementación de este diseño en la práctica es la revisión de equipajes de pasajeros por los funcionarios de la aduana en un aeropuerto; se fija la probabilidad de inclusión para cada pasajero y mediante cierto mecanismo de selección (muy simple) se selecciona la muestra, conforme las personas van ingresando al sitio. Nótese que el tamaño de muestra  $n(S)$  es aleatorio porque una muestra realizada mediante este mecanismo de selección puede incluir a todos los pasajeros o a ningún pasajero de la población.

**Definición 1.2.1.** Siendo  $n(s)$  el tamaño de muestra, el diseño de muestreo Bernoulli selecciona la muestra  $s$  con probabilidad

$$p(s) = \begin{cases} \pi^{n(s)} (1-\pi)^{N-n(s)} & \text{si } s \text{ tiene tamaño igual a } n(s) \\ 0 & \text{en otro caso} \end{cases} \quad (1.2.1)$$

### 1.2.1 Algoritmo de selección

La selección de una muestra con diseño Bernoulli conlleva los siguientes pasos:

1. Fijar el valor de  $\pi$  tal que  $0 < \pi < 1$ .
2. Obtener  $\varepsilon_k$  para  $k \in U$  como  $N$  realizaciones independientes de una variable aleatoria con distribución uniforme sobre el intervalo  $[0, 1]$ .
3. El elemento  $k$ -ésimo pertenece a la muestra con probabilidad  $\pi$ . Es decir, si  $\varepsilon_k < \pi$  el individuo  $k$ -ésimo es seleccionado.

Dado que  $\varepsilon_k \sim Unif[0, 1]$ , se tiene que  $Pr(\varepsilon_k < \pi) = \pi$  para  $k \in U$ . Por tanto, la inclusión de los individuos  $k$ -ésimo y  $l$ -ésimo, para  $k \neq l$ , es independiente. Esto implica que la distribución de  $I_k(S)$  es Bernoulli  $Ber(\pi)$  y se tiene el siguiente resultado.

**Resultado 1.2.1.** Definiendo a  $Q_r$  como el soporte que contiene a todas las posibles muestras de tamaño  $r$ , existen  $\binom{N}{r}$  muestras pertenecientes a  $Q_r$ . En otras palabras

$$\#(Q_r) = \binom{N}{r} \quad r = 0, \dots, N$$

Sin embargo, al definir  $Q$  como el soporte general de todas las posibles muestras de tamaños entre  $r = 0$  y  $r = N$ , se tiene que

$$\#(Q) = \sum_{r=1}^N \binom{N}{r} = 2^N$$

**Resultado 1.2.2.** Bajo muestreo Bernoulli, la distribución del tamaño de muestra  $n(S)$  es binomial  $Bin(N, \pi)$  y

$$Pr(n(S) = r) = \sum_{s \in Q_r} p(s) = \binom{N}{r} \pi^r (1 - \pi)^{N-r}, \quad (1.2.2)$$

con  $r = 1, \dots, N$  y  $Q_r$  el soporte que contiene a todas las posibles muestras de tamaño  $r$ , donde  $Q_r \subset Q$ .

*Demostración.* La distribución de  $I_k(S)$  es Bernoulli  $Ber(\pi)$ , las inclusiones de los individuos en la muestra son eventos independientes, entonces  $n(S) = \sum_U I_k$  sigue una distribución binomial. Ahora, dado el diseño de muestreo (1.2.1), para cualquier  $s \in Q_r$ , se cumple que  $p(s) = \pi^r (1 - \pi)^{N-r}$ . Como existen  $\binom{N}{r}$  maneras de seleccionar una muestra de  $r$  elementos de una población de tamaño  $N$ , se tiene que  $\#(Q_r) = \binom{N}{r}$ . Luego, al sumar  $p(s)$  sobre todas las muestras del soporte  $Q_r$  se obtiene el resultado.  $\square$

Como  $n(S)$  es aleatorio, existen  $2^N$  posibles muestras en el soporte  $Q$ . Nótese que  $n(S)$  tiene una distribución Binomial y, por tanto, su esperanza y varianza están dadas por:

$$E(n(S)) = N\pi \quad Var(n(S)) = N(\pi)(1 - \pi), \quad (1.2.3)$$

Aunque el investigador haya fijado las probabilidades de inclusión, se puede verificar que realmente el diseño de muestreo Bernoulli cumple las condiciones establecidas en el capítulo anterior y también que las probabilidades de inclusión, inducidas por el diseño de muestreo, son idénticas para cada elemento en la población  $\pi_k = \pi$ .

**Resultado 1.2.3.** Bajo el diseño de muestreo Bernoulli, se verifica que

$$\sum_{s \in Q} p(s) = 1 \quad (1.2.4)$$

*Demostración.* Para una población de tamaño  $N$ , el tamaño de muestra puede ser  $r$  con  $r = 0, 1, \dots, N$ . Es suficiente probar que  $\sum_{r=0}^N Pr(n(S) = r) = 1$ , utilizando el teorema binomial se tiene de inmediato porque  $n(S) \sim Bin(N, \pi)$ . Más aún, se tiene que

$$\begin{aligned} \sum_{s \in Q} p(s) &= \sum_{s \in Q_0} p(s) + \sum_{s \in Q_1} p(s) + \dots + \sum_{s \in Q_N} p(s) \\ &= \binom{N}{0} \pi^0 (1 - \pi)^{N-0} + \dots + \binom{N}{N} \pi^N (1 - \pi)^{N-N} \\ &= \sum_{r=0}^N \binom{N}{r} \pi^r (1 - \pi)^{N-r} = (\pi + 1 - \pi)^N = 1 \end{aligned}$$

□

**Resultado 1.2.4.** Para el diseño de muestreo Bernoulli, las probabilidades de inclusión de primer y segundo orden están dadas por:

$$\pi_k = \pi \quad (1.2.5)$$

$$\pi_{kl} = \begin{cases} \pi & \text{para } k = l \\ \pi^2 & \text{Para } k \neq l \end{cases} \quad (1.2.6)$$

*Demostración.* Teniendo en cuenta que existen  $\binom{N-1}{r-1}$  muestras de tamaño  $r$  que contienen al elemento  $k$ -ésimo, tenemos

$$\begin{aligned} \pi_k &= \sum_{\substack{s \ni k \\ s \subset Q}} p(s) \\ &= \sum_{\substack{s \ni k \\ s \subset Q_0}} p(s) + \sum_{\substack{s \ni k \\ s \subset Q_1}} p(s) + \dots + \sum_{\substack{s \ni k \\ s \subset Q_N}} p(s) \\ &= 0 + \binom{N-1}{0} \pi (1 - \pi)^{N-1} + \dots + \binom{N-1}{N-1} \pi (1 - \pi)^{N-1} \\ &= \sum_{r=0}^{N-1} \binom{N-1}{r} \pi^{r+1} (1 - \pi)^{N-1-r} \\ &= \pi \sum_{r=0}^{N-1} \binom{N-1}{r} \pi^r (1 - \pi)^{N-1-r} = \pi(\pi + (1 - \pi))^{N-1} = \pi \end{aligned}$$

Donde se utiliza el resultado del teorema binomial (Mood, Graybill & Boes 1974) que afirma que

$$\sum_{r=0}^m \binom{m}{r} a^r b^{m-r} = (a + b)^m. \quad (1.2.7)$$

Ahora como las inclusiones de los elementos de la población en la muestra son eventos independientes, entonces

$$Pr(k \in S \text{ y } l \in S) = Pr(I_k = 1) Pr(I_l = 1) = \pi^2 \quad (1.2.8)$$

□

### 1.2.2 El estimador de Horvitz-Thompson

**Resultado 1.2.5.** Para el diseño de muestreo Bernoulli, el estimador de Horvitz-Thompson, su varianza y su varianza estimada están dados por:

$$\hat{t}_{y,\pi} = \frac{1}{\pi} \sum_S y_k \quad (1.2.9)$$

$$Var_{BER}(\hat{t}_{y,\pi}) = \left( \frac{1}{\pi} - 1 \right) \sum_U y_k^2 \quad (1.2.10)$$

$$\widehat{Var}_{BER}(\hat{t}_{y,\pi}) = \frac{1}{\pi} \left( \frac{1}{\pi} - 1 \right) \sum_S y_k^2, \quad (1.2.11)$$

respectivamente

*Demostración.* El resultado es inmediato porque

$$\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k \pi_l = \pi^2 - \pi^2 = 0 & \text{para } k \neq l \\ \pi_{kk} - \pi_k \pi_k = \pi(1 - \pi) & \text{para } k = l \end{cases} \quad (1.2.12)$$

luego la doble suma en la varianza del estimador de Horvitz-Thompson pasa a ser una sola suma; lo anterior sucede análogamente con la expresión de la estimación de la varianza.  $\square$

Nótese que en caso de que la muestra realizada o seleccionada esté compuesta por todas las unidades de la población, es decir se deba realizar un censo<sup>3</sup>, la probabilidad de inclusión para cada elemento de la población estaría dada por  $\pi_k = \pi$ . En este caso, el estimador de Horvitz-Thompson estaría dado por la siguiente expresión

$$\hat{t}_{y,\pi} = \frac{1}{\pi} \sum_U y_k = \frac{t_y}{\pi} \neq t_y \quad (1.2.13)$$

En este caso, el estimador de Horvitz-Thompson es deficiente para la estimación del total poblacional  $t_y$  y se sugiere la utilización del estimador alternativo para el total poblacional que, para el caso particular del diseño de muestreo Bernoulli, estaría dado por

$$\hat{t}_{y,alt} = N\tilde{y}_S = N \frac{\sum_S y_k}{n(S)} = N\bar{y}_S. \quad (1.2.14)$$

Fácilmente se verifica que si  $s = U$ , entonces  $\hat{t}_{y,alt} = t_y$ .

**Ejemplo 1.2.1.** Para nuestra población de ejemplo  $U$ , existen  $2^5 = 32$  posibles muestras. Si la probabilidad de inclusión es fija para cada elemento e igual a 0,3, realice el cálculo léxico-gráfico del estimador de Horvitz-Thompson y compruebe el insesgamiento y la varianza.

### 1.2.3 El efecto de diseño

Una medida que compara la eficiencia entre dos estrategias de muestreo es el efecto de diseño. Ésta herramienta práctica muestra la ganancia o pérdida, de precisión, al utilizar una estrategia de muestreo más compleja que un diseño aleatorio simple sin reemplazo junto con el estimador de Horvitz-Thompson y está definida de la siguiente manera:

<sup>3</sup>En el diseño de muestreo Bernoulli, la probabilidad de seleccionar todas las unidades de la población en la muestra es equivalente a  $\pi^N$ .

**Definición 1.2.2.** Siendo  $(\hat{T}, p(\cdot))$  y  $(\hat{T}_\pi, MAS)$  dos estrategias de muestreo utilizadas para la estimación del parámetro  $T$ , se define el efecto de diseño como

$$Deff = \frac{Var_p(\hat{T})}{Var_{MAS}\hat{T}_\pi}. \quad (1.2.15)$$

en particular, el efecto de diseño, restringido a la estimación de un total poblacional y al usar el estimador de Horvitz-Thompson en ambas estrategias, toma la siguiente forma

$$Deff = \frac{\widehat{Var}_p(\hat{t}_{y,\pi})}{\frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{yU}^2}. \quad (1.2.16)$$

Cuando el efecto de diseño es más grande que la unidad, la varianza de la estrategia del numerador es más grande que la denominador, por tanto, se ha perdido precisión al utilizar una estrategia de muestreo más compleja; si el cociente es menor que uno, se ha ganado precisión. Fue Cornfield (1951) quien sugirió evaluar la eficiencia de una estrategia de muestreo al hacer el cociente entre la varianza de la misma y la del diseño aleatorio simple sin reemplazo con el estimador de Horvitz-Thompson. Más adelante Kish (1965) lo llamo DEFF (efecto de diseño, por sus siglas en inglés).

Sin embargo, en la mayoría de ocasiones, el cálculo de este cociente no es sencillo. Lehtonen & Pahkinen (2003) plantea una estimación del efecto de diseño para totales mediante la estimación de las varianzas que intervienen en la expresión. De esta forma, se tiene

**Resultado 1.2.6.** Un estimador del efecto de diseño  $Deff$  para el total poblacional  $t_y$  es

$$\hat{Deff} = \frac{\widehat{Var}_p(\hat{T})}{\frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{ys}^2}. \quad (1.2.17)$$

No todos los parámetros tienen el mismo comportamiento, por lo tanto, los efectos de diseño para estos no tendrán un mismo criterio de optimalidad. Es decir, si existe un criterio de optimalidad con respecto a un parámetro, digamos el total poblacional  $t_y$ , no necesariamente se cumplirá ese criterio con un parámetro distinto, digamos la mediana poblacional.

Dado que el tamaño de muestra en diseños diferentes al muestreo aleatorio simple sin reemplazo puede ser variable, es necesario asegurarse que  $n = E_{MAS}(n(S)) = E_p(n(S))$  para que exista un punto objetivo de comparación. Por ejemplo, para comparar la eficiencia del estimador de Horvitz-Thompson en el diseño de muestreo Bernoulli, es necesario fijar el tamaño de muestra, dado que este diseño no es de tamaño fijo; es decir que  $n = E_{MAS}(n(S)) = E_{BER}(n(S)) = N\pi$ . Por lo que resulta que  $\pi = n/N$ .

De esta manera podemos introducir la medida de eficiencia del diseño de muestreo Bernoulli con respecto al MAS, así

$$deff = \frac{Var_{BER}(\hat{t}_{y,\pi})}{Var_{MAS}(\hat{t}_{y,\pi})} = 1 - \frac{1}{N} + \frac{1}{CV_y^2} \cong 1 + \frac{1}{CV_y^2} \quad (1.2.18)$$

Por tanto, si el efecto de diseño  $deff$  es igual a 1.8, esto implica que la varianza del  $\pi$  estimador bajo diseño de muestreo Bernoulli es 1.8 veces la varianza del  $\pi$  estimador bajo MAS.

#### 1.2.4 Marco y Lucy

Suponga que se debe seleccionar una muestra con un diseño de muestreo Bernoulli. Se quiere que el tamaño esperado de muestra sea de  $N\pi = 400$  empresas del sector industrial. Como el tamaño de la población es  $N = 2396$ , entonces el valor que se fija para  $\pi$  es de 0.1669. Para seleccionar la muestra

se utiliza la función `S.BE(N,prob)` del paquete `TeachingSampling` cuyos parámetros son `N`, el tamaño poblacional y `prob` el valor de la probabilidad de inclusión para cada elemento de la población. Esta función utiliza el algoritmo secuencial descrito en la anterior sección.

Primero se carga en R el archivo `Marco` que contiene el marco de muestreo para la selección de la muestra. Se fijan los parámetros de la función, `N` y `prob`. Esta función devuelve un vector contenido el índice de los elementos seleccionados en la muestra. En este caso particular, el primer elemento seleccionado es el número 2 y el último el número 2394.

```

data(BigLucy)
N <- dim(BigLucy)[1]
pik <- 0.025
sam <- S.BE(N,pik)
muestra <- BigLucy[sam,]
attach(muestra)
head(muestra)

##           ID      Ubication Level   Zone Income Employees Taxes
## 86 AB00000000086 C0246025K0055872 Small County1    456     75     9
## 118 AB0000000118 C0140163K0161734 Small County1   436     77     8
## 159 AB0000000159 C0045680K0256217 Small County1   230     10     2
## 200 AB0000000200 C0035648K0266249 Small County1   310     54     4
## 325 AB0000000325 C0059021K0242876 Small County1   208     22     1
## 373 AB0000000373 C0079681K0222216 Small County1   270     72     3
##          SPAM ISO Years Segments
## 86   yes  no   22 County1  9
## 118  yes  no   49 County1 12
## 159  yes  no    7 County1 16
## 200  yes  no   22 County1 20
## 325  no   no   28 County1 33
## 373  yes  no   26 County1 38

n <- dim(muestra)[1]
n

## [1] 2228

```

Aplicando los índices obtenidos por la función `S.BE` al marco de muestreo obtenemos la identificación y ubicación de las empresas seleccionadas en la muestra. Nótese que el tamaño de muestra efectivo es de 2228 empresas. Una vez que la etapa de recolección de datos se haya realizado, obtendremos un archivo de datos de `Lucy` conteniendo los valores de las características de interés para las empresas seleccionadas que será adjuntado a R mediante la función `attach`.

La etapa de estimación de resultados se hace utilizando la función `E.BE(y,prob)` del paquete `TeachingSampling` cuyos argumentos son `y`, un vector o matriz conteniendo los valores de las características de interés en la muestra y `prob`, la probabilidad de inclusión. En este caso la longitud de cada vector es de  $n = 2228$ . Esta función arroja la estimación del total poblacional de `y` usando el estimador de Horvitz-Thompson, la estimación de la varianza y el coeficiente de variación del mismo. Por ejemplo, la variable `Income` contiene los valores del ingreso declarado en el último año por 396 empresas del sector industrial pertenecientes a la muestra. La estimación para esta característica se hace mediante el siguiente código:

```
estima <- data.frame(Income, Employees, Taxes)
E.BE(estima,pik)
```

Cuadro 1.4: *Estimaciones para el diseño de muestreo Bernoulli*

	N	Income	Employees	Taxes
Estimation	89120.00	37281880.00	5541240.00	988720.00
Standard Error	1864.32	910626.84	130608.04	35971.57
CVE	2.09	2.44	2.36	3.64
DEFF	Inf	3.75	4.71	1.49

La tabla 1.4 muestra los resultados obtenidos para este caso particular, donde la desviación relativa de una estimación, medida en porcentaje está definida como

Por otro lado, nótese que, aunque la distribución asintótica del estimador de Horvitz-Thompson es normal, es necesario verificar el comportamiento del estimador con el tamaño de muestra esperado. Se realizaron varios experimentos de Monte Carlo con el propósito de tener un examen más cercano del estimador de Horvitz-Thompson del total de la característica `Income` en la población `Lucy`. El resultado de la simulación se muestra en los histogramas de la figura 3.1. Se espera que el promedio de las estimaciones en cada experimento coincida con el total poblacional y la varianza de éstas debe acercarse a la varianza basada en el diseño de muestreo Bernoulli.

```
bary <- mean(Income)
sdy <- sd(Income)
x <- seq(min(Income),max(Income),by=10)
a <- (bary/sdy)^2
b <- sdy^2/bary

par(mfrow=c(1,2))
hist(Income,freq=FALSE, breaks=10)
lines(x, dgamma(x, shape=a, scale=b), col=2)
hist(Income, freq=FALSE, breaks=10)
lines(x, dnorm(x, bary, sdy), col=2)
```

La media de las estimaciones de  $t_y$  es 1035176 que ajusta bien con el parámetro correspondiente  $t_y = 1035217$ . La distribución parece ser simétrica con forma de campana (los valores de la distribución teórica se muestran en la curva sólida y roja) y no se notan grandes discrepancias entre lo observado y lo teórico. En algunos casos, en donde el tamaño de muestra no es lo suficientemente grande, se debe verificar el comportamiento normal del estimador.

### 1.3 Muestreo aleatorio simple con reemplazo

Una **muestra aleatoria simple con reemplazo**, de tamaño  $m$  de una población de  $N$  elementos es la extracción de  $m$  muestras independientes de tamaño 1, en donde cada elemento se extrae de la población con la misma probabilidad

$$p_k = \frac{1}{N} \quad \forall k \in U$$

**Definición 1.3.1.** Un diseño de muestreo aleatorio simple con reemplazo se define como

$$p(s) = \begin{cases} \frac{m!}{n_1(s)! \dots n_N(s)!} \prod_U \left(\frac{1}{N}\right)^{n_k(s)} & \text{si } \sum_U n_k(s) = m \\ 0 & \text{en otro caso} \end{cases} \quad (1.3.1)$$

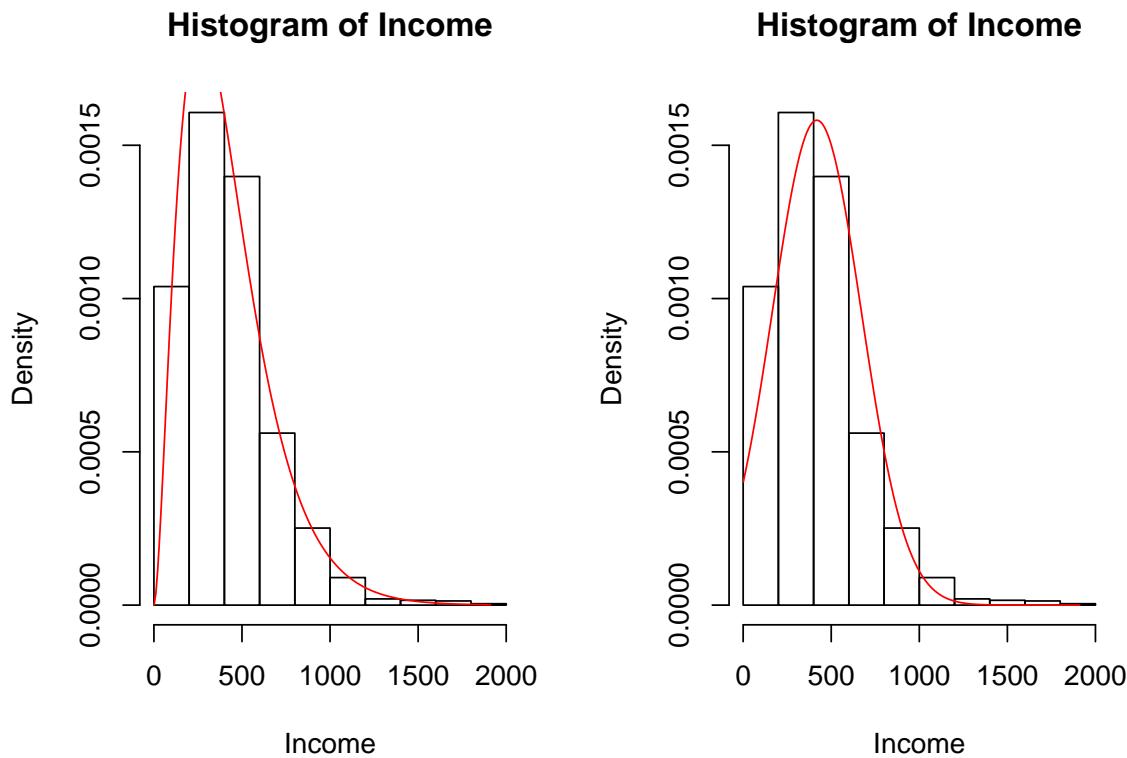


Figura 1.2: Distribución de la característica *Income* y su posible modelamiento bajo la distribución gamma (izquierda) y norma (derecha).

Donde  $n_k(s)$  es el número de veces que el elemento  $k$ -ésimo es seleccionado en la muestra realizada  $s$ .

**Resultado 1.3.1.** Para este diseño de muestreo, existen  $\binom{N+m-1}{m}$  posibles muestras de tamaño  $m$ ; es decir

$$\#(Q) = \binom{N + m - 1}{m}$$

**Resultado 1.3.2.** Dado el soporte  $Q$ , de todas las posibles muestras con reemplazo de tamaño  $m$ , se verifica que el diseño de muestreo aleatorio simple con reemplazo es tal que

$$\sum_{s \in Q} p(s) = 1$$

*Demostración.* La demostración es inmediata porque este diseño de muestreo es una función de densidad

multinomial discreta sobre  $Q$ .

$$\begin{aligned}
 \sum_{s \in Q} p(s) &= \sum_{s \in Q} \frac{m!}{n_1(s)! \dots n_N(s)!} \prod_U \left(\frac{1}{N}\right)^{n_k(s)} \\
 &= \sum_{s \in Q} \frac{m!}{n_1(s)! \dots n_N(s)!} \left(\frac{1}{N}\right)^{n_1(s)} \dots \left(\frac{1}{N}\right)^{n_N(s)} \\
 &= \sum_{\substack{n_1(s) \dots n_N(s) \\ \sum_U n_k(S)=m}} \frac{m!}{n_1(s)! \dots n_N(s)!} \left(\frac{1}{N}\right)^{n_1(s)} \dots \left(\frac{1}{N}\right)^{n_N(s)} \\
 &= \underbrace{\left(\frac{1}{N} + \dots + \frac{1}{N}\right)_N^m}_{N \text{ veces}} \\
 &= 1
 \end{aligned}$$

donde se utiliza el resultado del teorema multinomial que afirma que

$$\sum_{\substack{n_1 \dots n_N \\ \sum_U n_k=m}} \frac{m!}{n_1! \dots n_N!} (p_1)^{n_1} \dots (p_N)^{n_N} = \left( \sum_{k=1}^N p_k \right)^m \quad (1.3.2)$$

□

**Resultado 1.3.3.** Para un diseño aleatorio simple con reemplazo, las probabilidades de inclusión de primer y segundo orden están dadas por

$$\pi_k = 1 - \left(1 - \frac{1}{N}\right)^m \quad (1.3.3)$$

$$\pi_{kl} = 1 - 2 \left(1 - \frac{1}{N}\right)^m + \left(1 - \frac{2}{N}\right)^m \quad (1.3.4)$$

respectivamente.

*Demostración.* Utilizando los resultados 2.2.9. y 2.2.10., respectivamente, se llega a la demostración.

□

**Ejemplo 1.3.1.** En nuestra población ejemplo el tamaño poblacional es  $N = 5$ . Si se quisiera seleccionar una muestra aleatoria simple con reemplazo de tamaño  $m = 2$ , entonces existirían  $N^m = 5^2 = 25$  posibles extracciones ordenadas. Sin embargo, sólo existen  $\binom{N+m-1}{m} = \binom{6}{2} = 15$  posibles muestras. Cada una de las posibles muestras que pertenecen al soporte con reemplazo tienen las siguientes probabilidades de selección.

	V1	V2	p	n1	n2	n3	n4	n5
1	Yves	Yves	0.04	2	0	0	0	0
2	Ken	Ken	0.04	0	2	0	0	0
3	Erik	Erik	0.04	0	0	2	0	0
4	Sharon	Sharon	0.04	0	0	0	2	0
5	Leslie	Leslie	0.04	0	0	0	0	2
6	Yves	Ken	0.08	1	1	0	0	0
7	Yves	Erik	0.08	1	0	1	0	0
8	Yves	Sharon	0.08	1	0	0	1	0

9	Yves	Leslie	0.08	1	0	0	0	1
10	Ken	Erik	0.08	0	1	1	0	0
11	Ken	Sharon	0.08	0	1	0	1	0
12	Ken	Leslie	0.08	0	1	0	0	1
13	Erik	Sharon	0.08	0	0	1	1	0
14	Erik	Leslie	0.08	0	0	1	0	1
15	Sharon	Leslie	0.08	0	0	0	1	1

Nótese que la suma de las probabilidades inducidas por el diseño de muestreo es igual a uno y que cada una de ellas es mayor que cero.

### 1.3.1 Algoritmo de selección

Tillé (2006) presenta dos algoritmos para seleccionar una muestra aleatoria simple con reemplazo. El primero, de manera general induce  $m$  selecciones individuales y el segundo, es un método secuencial que implementa la selección mediante la distribución binomial.

#### Método de $m$ selecciones

El siguiente método de selección se implementa en  $m$  pasos, y aunque no es eficiente computacionalmente, es muy conocido.

- Seleccionar un primer elemento con probabilidad  $\frac{1}{N}$  de todo el conjunto de datos.
- Seleccionar un segundo elemento con probabilidad  $\frac{1}{N}$  de todo el conjunto de datos.
- ...
- Seleccionar un  $m$ -ésimo elemento con probabilidad  $\frac{1}{N}$  de todo el conjunto de datos.

Hace unas pocas décadas, cuando no existía la ayuda tecnológica de ahora, no imagino como los encargados de la selección de la muestra pudieron haber utilizado este algoritmo. Imagine seleccionar una muestra de 3000 elementos sin la facilidad de un computador.

#### Método secuencial

Tillé (2006) afirma que este procedimiento es mejor que el anterior porque permite seleccionar una muestra de tamaño  $m$  en una sola pasada por el conjunto de datos.

- Seleccionar  $n_k$  veces el elemento  $k$ -ésimo de acuerdo a una distribución binomial.

$$Bin\left(m - \sum_{i=1}^{k-1} n_i, \frac{1}{N - k + 1}\right) \quad (1.3.5)$$

Para todo  $k \in U$ .

**Ejemplo 1.3.2.** Como se ha visto en los capítulos anteriores, R incorpora en la función `sample`, la selección de muestras aleatorias simples con reemplazo, simplemente el argumento `replace` debe ser activado mediante, `replace=TRUE`. Así, para seleccionar una muestra con reemplazo de tamaño  $m = 3$ , sólo es necesario escribir el siguiente código.

```
N <- length(U)
sam <- sample(N, 3, replace=TRUE)
U[sam]

## [1] "Yves"   "Sharon" "Leslie"
```

El procedimiento de selección de una muestra aleatoria con reemplazo de tamaño  $m$  mediante el uso del algoritmo secuencial está implementado en la función `S.WR(N,m)` cuyos argumentos son  $N$ , el tamaño de la población y  $m$ , el tamaño de la muestra con reemplazo. Así, para seleccionar una muestra aleatoria simple con reemplazo de la población  $U$  de tamaño  $N = 5$ , se tiene

```
m <- 3
sam <- S.WR(N,m)
U[sam]

## [1] "Ken"    "Leslie" "Leslie"
```

Una vez más, la salida de la función es un vector de índices (no necesariamente distintos) de los elementos pertenecientes a la muestra seleccionada  $s$ . Este algoritmo utiliza la distribución binomial en cada uno de sus pasos, de tal forma que para la selección de la anterior muestra conformada por **Ken**, **Leslie** y **Leslie** cada uno de los  $N = 5$  pasos del algoritmo arrojaron los siguientes resultados.

k	Nombre	Bin n	Bin p	nk
1	Yves	3	0.2000	0
2	Ken	3	0.2500	1
3	Erik	2	0.3333	0
4	Sharon	2	0.5000	2
5	Leslie	0	1.0000	0

Donde `Bin n` y `Bin p` son los parámetros de la distribución binomial asociada al algoritmo secuencial. Note que la cantidad `nk` se refiere a la realización de la variable  $n_k(s)$ .

### 1.3.2 El estimador de Hansen-Hurwitz

Cuando se tienen las cantidades del resultado 3.3.3 se pueden implementar los principios del estimador de Horvitz-Thompson para estimar el total poblacional  $t_y$ ; sin embargo, el cálculo y estimación de la varianza de esta estrategia de muestreo resulta ser muy compleja (computacionalmente). Por esta razón, utilizaremos el estimador de Hansen-Hurwitz dado por (??) que estima de manera insesgada al parámetro de interés  $t_y$ .

**Resultado 1.3.4.** *Para un diseño de muestreo aleatorio simple con reemplazo, el estimador de Hansen-Hurwitz del total poblacional  $t_y$ , su varianza y su varianza estimada están dados por:*

$$\hat{t}_{y,p} = \frac{N}{m} \sum_{i=1}^m y_i \quad (1.3.6)$$

$$Var_{MRAS}(\hat{t}_{y,p}) = N \frac{(N-1)}{m} S_{yU}^2 \quad (1.3.7)$$

$$\widehat{Var}_{MRAS}(\hat{t}_{y,p}) = \frac{N^2}{m} S_{ysr}^2 \quad (1.3.8)$$

respectivamente, con  $S_{yU}^2$  el estimador de la varianza de los valores de la característica de interés  $y$  en el universo y  $S_{ysr}^2$  el estimador de la varianza de los valores  $y_i$  que pertenecen a la muestra seleccionada ( $\forall i \in m$ ) (no necesariamente distintos) en la muestra. Esto es,

$$S_{ysr}^2 = \frac{1}{m-1} \sum_{i=1}^m (y_i - \bar{y}_S)^2.$$

Nótese que  $\hat{t}_{y,p}$  es insesgado para el total poblacional  $t_y$  de la característica de interés  $y$ , y que  $\widehat{Var}_{MRAS}(\hat{t}_{y,p})$  es insesgado para  $Var_{MRAS}(\hat{t}_{y,p})$ .

*Demostración.* Los resultados se obtienen escribiendo el estimador de Hansen-Hurwitz de la siguiente manera,

$$\hat{t}_{y,p} = \frac{1}{m} \sum_U n_k(S) \frac{y_k}{p_k} = \frac{N}{m} \sum_U n_k(S) y_k \quad (1.3.9)$$

Por tanto, utilizando el resultado 2.2.8., se tiene que

$$\begin{aligned} E(\hat{t}_{y,p}) &= \frac{N}{m} \sum_U E(n_k(S)) y_k \\ &= \frac{N}{m} \sum_U \frac{m}{N} y_k = t_y \end{aligned}$$

Por otro lado, asumiendo que las variables  $Z_i$  son independientes e idénticamente distribuidas

$$\begin{aligned} Var(\hat{t}_{y,p}) &= Var\left(\frac{1}{m} \sum_i^m Z_i\right) \\ &= \frac{1}{m^2} \sum_i^m Var(Z_i) \\ &= \frac{1}{m^2} \sum_i^m \left( \sum_U \frac{1}{N} (Ny_k - t)^2 \right) \\ &= \frac{1}{m} \left( \frac{N^2}{N} \sum_U (y_k - \bar{y}_U)^2 \right) \\ &= N \frac{(N-1)}{m} S_{yU}^2 \end{aligned}$$

Escribiendo el estimador de la varianza como

$$\widehat{Var}(\hat{t}_{y,p}) = \frac{1}{m} \frac{1}{m-1} \sum_U n_k(S) (Ny_k - \hat{t}_{y,p})^2 \quad (1.3.10)$$

se tiene el insesgamiento dado por

$$\begin{aligned}
 E\left(\widehat{Var}(\hat{t}_{y,p})\right) &= \frac{1}{m} \frac{1}{m-1} \sum_U E(n_k(S)(Ny_k - \hat{t}_{y,p})^2) \\
 &= \frac{1}{m} \frac{1}{m-1} \sum_U E(n_k(S)(Ny_k - t_y)^2 - n_k(S)(\hat{t}_{y,p} - t_y)^2) \\
 &= \frac{1}{m} \frac{1}{m-1} E\left(\sum_U n_k(S)(Ny_k - t_y)^2\right) \\
 &\quad - \frac{1}{m} \frac{1}{m-1} E\left((\hat{t}_{y,p} - t_y)^2 \sum_U n_k(S)\right) \\
 &= \frac{1}{m} \frac{1}{m-1} \left[ E\left(\sum_U n_k(S)(Ny_k - t_y)^2\right) - mE((\hat{t}_{y,p} - t_y)^2) \right] \\
 &= \frac{1}{m} \frac{1}{m-1} \left[ m \left( \sum_U \frac{m}{N} (Ny_k - t_y)^2 \right) - mVar(\hat{t}_{y,p}) \right] \\
 &= \frac{1}{m} \frac{1}{m-1} [m^2 Var(\hat{t}_{y,p}) - mVar(\hat{t}_{y,p})] \\
 &= Var(\hat{t}_{y,p})
 \end{aligned}$$

□

**Ejemplo 1.3.3.** Para nuestra población de ejemplo  $U$ , existen  $\binom{N+m-1}{m} = 20$  posibles muestras con reemplazo de tamaño  $m = 2$ . Realice el cálculo léxico-gráfico del estimador de Hansen-Hurwitz y compruebe el insesgamiento y la varianza.

### 1.3.3 Marco y Lucy

Suponga que se quiere seleccionar una muestra aleatoria simple con reemplazo de tamaño  $m = 400$  empresas del sector industrial. Para la selección de la muestra es posible usar la función `sample` que viene integrada con R. En primer lugar se debe cargar el marco de muestreo que permite la selección, identificación y posterior ubicación de cada individuo en la muestra con reemplazo. Para la selección de la muestra es necesario ingresar los parámetros de la función, en este caso  $N=2396$ , el tamaño poblacional, está dado por la cantidad de filas (registros de empresas del sector industrial) del marco de muestro y  $m=400$  empresas que se seleccionaran con reemplazo.

```

data(BigLucy)
attach(BigLucy)
N <- dim(BigLucy)[1]
m <- 2000
sam <- sample(N, m, replace=TRUE)

```

Sin embargo, para seleccionar la muestra con reemplazo utilizando el método secuencial, el paquete `TeachingSampling` adjunta la función `S.WR` cuyos argumentos son `N`, el tamaño de la población y `m`, el tamaño de la muestra con reemplazo. El resultado de la función es un conjunto de índices (no necesariamente distintos) que aplicados a la población resulta en los valores de la característica de interés para las empresas (no necesariamente distintas) seleccionadas. Nótese que una empresa seleccionada se tendrá en cuenta en la etapa de estimación tantas veces como haya sido seleccionada.

```

sam <- S.WR(N,m)
muestra <- BigLucy[sam,]
attach(muestra)

head(muestra)

##           ID      Ubication Level     Zone Income Employees Taxes
## 62    AB0000000062 C0196110K0105787 Small County1    456      71   9.0
## 63    AB0000000063 C0242126K0059771 Small County1    340      28   5.0
## 63.1   AB0000000063 C0242126K0059771 Small County1    340      28   5.0
## 93    AB0000000093 C0159050K0142847 Small County1    441      66   8.0
## 115   AB0000000115 C0123025K0178872 Small County1     10      65   0.5
## 296   AB0000000296 C0129476K0172421 Small County1    245      67   2.0
##          SPAM ISO Years Segments
## 62     yes  no    12 County1 7
## 63     yes  no    20 County1 7
## 63.1   yes  no    20 County1 7
## 93     no   no    11 County1 10
## 115   no   no    28 County1 12
## 296   no   no     2 County1 30

dim(muestra)

## [1] 2000   11

```

La primera empresa en ser seleccionada mediante el método secuencial es la empresa que ocupa la segunda posición en el marco de muestreo; es decir, la empresa cuyo número único de identificación corresponde a **AB002**, la segunda y tercera empresa en ser seleccionadas corresponde a la empresa identificada con el número único **AB015**. Si un elemento ha sido seleccionada más de una vez, R codifica automáticamente las posteriores selecciones con un punto seguido de un número que indica el número de veces menos uno que ha sido seleccionada la misma unidad.

Una vez que las empresas son seleccionadas, se programa la visita del encuestador en la cual se registran los valores de las características de interés. Cuando se tiene la base de datos con la información pertinente para todas las empresas seleccionadas en la muestra con reemplazo, se procede a estimar los totales de las características de interés. La función **E.WR** del paquete **TeachingSampling** permite la estimación de una o varias características de interés simultáneamente. Para ello, se debe crear un conjunto de datos con la información recolectora para cada una de las 400 empresas en las características de interés. En este caso creamos un conjunto de datos con las tres características de interés **Income**, **Employees** y **Taxes**.

La función **E.WR** del paquete **TeachingSampling** tiene tres argumentos, **N**, el tamaño de la población y **m**, el tamaño de la muestra con reemplazo y el conjunto de datos (conteniendo los valores para la(s) característica(s) de interés). El resultado de la función es la estimación del total, la varianza estimada y el respectivo coeficiente de variación de la(s) característica(s) de interés.

```

estima <- data.frame(Income, Employees, Taxes)
E.WR(N, m, estima)

```

La tabla **??.** muestra los resultados particulares de esta estrategia de muestreo. Nótese que con un menor tamaño de muestra, se obtienen mejores resultados que al utilizar una estrategia de muestreo que contempla un diseño Bernoulli y el estimador de Horvitz-Thompson.

Cuadro 1.5: *Estimaciones para el diseño de muestreo aleatorio simple con reemplazo*

	N	Income	Employees	Taxes
Estimation	85296.00	36809829.98	5404439.86	1021547.54
Standard Error	0.00	512237.40	62252.04	31822.36
CVE	0.00	1.39	1.15	3.12
DEFF		1.02	1.02	1.02

### El efecto de diseño

Sin embargo, utilizando el efecto de diseño podemos comparar la eficiencia de la anterior estrategia utilizada en Lucy mediante el efecto de diseño. Utilizando la definición podemos aproximar la medida mediante

$$\begin{aligned} Deff &= \frac{Var_{MRAS}(\hat{t}_{y,p})}{Var_{MAS}(\hat{t}_{y,\pi})} \\ &= \frac{1}{1-f} \left(1 - \frac{1}{N}\right) \cong \frac{1}{1-f} \end{aligned}$$

Por tanto, para la estrategia de muestreo utilizada anteriormente, tenemos  $Deff = \frac{1}{1 - \frac{2000}{85296}} = 1.02$ .

Lo anterior indica que existe una pérdida del 2% de precisión al utilizar la estrategia de muestreo con reemplazo y el estimador de Hansen-Hurwitz. En general se tiene que, para tamaños de muestra muy pequeños, en comparación a  $N$ , las dos estrategias arrojan resultados muy similares. Sin embargo, a medida que el tamaño de muestra crece, en comparación a  $N$ , la medida  $Deff$  aumenta significativamente; es decir, existe una pérdida muy grande de eficiencia.

Dado que el diseño de muestreo es con reemplazo, se quiere verificar que la distribución asintótica del estimador de Hansen-Hurwitz sea normal. Se realiza una simulación de Monte Carlo, con los mismos lineamientos utilizados en la sección 3.1.3 en donde se realizaron varios experimentos de Monte Carlo para examinar el comportamiento del estimador de Hansen-Hurwitz en la característica ingreso. El resultado de la simulación se muestra en los histogramas de la figura 3.3. En este experimento de Monte Carlo el promedio de las estimaciones de cada experimento coincide con el total poblacional y se espera que la varianza de las estimaciones debe acercarse a la varianza basada en el diseño de muestreo aleatorio simple.

```
HH <- c()
for(i in 1:500){
  sam <- sample(N, m, replace=TRUE)
  HH[i] = E.WR(N, m, BigLucy$Income[sam])[1,2]
}

barHH <- mean(HH)
sdHH <- sd(HH)
x <- seq(min(HH),max(HH),by=10)

hist(HH, freq=FALSE)
lines(x, dnorm(x, barHH, sdHH), col=2)
```

La media de las estimaciones de  $t_y$  es 36609714.14 que ajusta bien con el parámetro correspondiente  $t_y = 1035217$ . Nótese que la varianza del estimador (mediante este experimento de Monte Carlo) no es

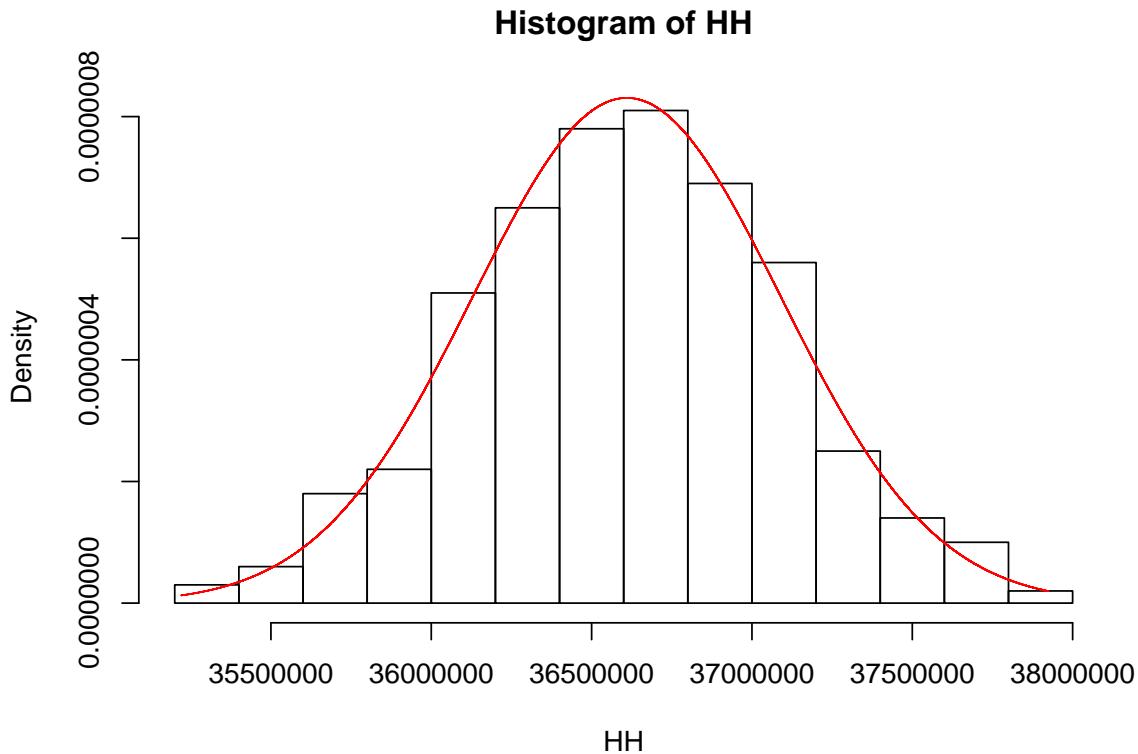


Figura 1.3: Distribución empírica del estimador de Hansen-Hurwitz para el diseño de muestreo aleatorio simple con reemplazo.

muy grande y que la distribución del estimador no muestra valores atípicos. Hay que tener cuidado con las afirmaciones acerca de normalidad en este caso pues la distribución, aunque parece ser simétrica y con forma de campana, en realidad puede estar sesgada a derecha o a izquierda.

## 1.4 Diseño de muestreo sistemático

En algunas ocasiones, cuando no se dispone de un marco de muestreo, por lo menos no de forma explícita, o cuando el marco disponible está ordenado de forma particular, con respecto a los rótulos del mismo, es posible utilizar el diseño de muestreo sistemático como una opción para la selección de muestras. La característica más particular de este diseño de muestreo es que todas las unidades se suponen enumeradas del 1 al  $N$ , al menos implícitamente, y se tiene conocimiento de que la población se encuentra particionada en  $a$  grupos poblacionales latentes. En este orden de ideas el tamaño poblacional  $N$  puede ser escrito como

$$N = na + c \quad (1.4.1)$$

en donde  $0 \leq c < a$  y  $n$ , el tamaño de muestra esperado, se define como la parte entera del cociente  $N/a$ . Nótese que  $c$  es un entero que representa el residuo algebraico del total poblacional y se puede ver fácilmente que toma la siguiente forma

$$c = N - \left\| \frac{N}{a} \right\| a \quad (1.4.2)$$

En donde  $\left\| \frac{N}{a} \right\|$  representa la parte entera del cociente  $N/a$ . Una vez que los grupos han sido conformados, se procede a escoger de manera aleatoria, un número entre 1 y  $a$ , por ejemplo  $r$ . La muestra estará conformada sistemáticamente por los elementos  $r, r+a, r+2a, \dots, r+(n-1)a$ . Nótese que en el caso en donde  $c = 0$ , el tamaño de muestra estará dado por  $n = N/a$ ; de otra forma, si  $c > 0$ , el tamaño de muestra puede ser  $n = \left\| \frac{N}{a} \right\|$  ó  $n = \left\| \frac{N}{a} \right\| + 1$ . Como lo señala Raj (1968) este diseño de muestreo es un caso especial de un muestreo por conglomerados, como se verá en los siguientes capítulos.

Cuadro 1.6: *Possible configuración del muestreo sistemático.*

Grupo	$s_1$	$\dots$	$s_r$	$\dots$	$s_a$
$n = 1$	1	$\dots$	$r$	$\dots$	$a$
$n = 2$	$1+a$	$\dots$	$r+a$	$\dots$	$2a$
$n = 3$	$1+2a$	$\dots$	$r+2a$	$\dots$	$3a$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$n = \left\  \frac{N}{a} \right\ $	$1+(n-1)a$	$\dots$	$r+(n-1)a$	$\dots$	$na$
$n = \left\  \frac{N}{a} \right\  + 1$	$1+na$	$\dots$	$\square$	$\dots$	$\square$

El anterior esquema permite una mejor comprensión del funcionamiento del diseño de muestreo sistemático. Nótese el ordenamiento por grupos de las unidades que pertenecen a la población. En particular, esta tabla corresponde a una población, en donde, si se seleccionara el último grupo  $s_a$ , entonces el tamaño de muestra sería  $n = \left\| \frac{N}{a} \right\|$ , mientras que si se escogiera el primer grupo  $s_1$ , el tamaño de muestra estaría dado por  $n = \left\| \frac{N}{a} \right\| + 1$ .

Por otro lado, nótese que cada grupo  $s_r$  constituye una posible muestra, de tal forma que

$$U = \bigcup_{r=1}^a s_r. \quad (1.4.3)$$

El soporte  $Q$  de todas las posibles muestras sistemáticas, queda entonces definido como

$$Q_r = \{s_1, s_2, \dots, s_r, \dots, s_a\}. \quad (1.4.4)$$

**Resultado 1.4.1.** *Para este diseño de muestreo, la cardinalidad del soporte es igual al número de grupos formados. Es decir*

$$\#Q_r = a$$

**Definición 1.4.1.** *Suponga que el tamaño poblacional es tal que  $N = na + c$ , con  $0 \leq c < a$ . Se define un diseño de muestreo sistemático de la siguiente manera*

$$p(s) = \begin{cases} \frac{1}{a} & \text{si } s \in Q_r \\ 0 & \text{en otro caso} \end{cases} \quad (1.4.5)$$

Dado que sólo existen  $a$  posibles muestras, el diseño de muestreo sistemático cumple que  $\sum_{s \in Q} p(s) = 1$ .

### 1.4.1 Algoritmo de selección

El siguiente algoritmo secuencial permite la extracción de una muestra mediante el diseño de muestreo sistemático.

1. Seleccionar con probabilidad  $\frac{1}{a}$  un arranque aleatorio. Es decir un entero  $r$ , tal que  $1 \leq r \leq a$ .

2. La muestra estará definida por el siguiente conjunto

$$s_r = \{k : k = r + (j - 1)a; j = 1, \dots, n(S)\} \quad (1.4.6)$$

**Ejemplo 1.4.1.** Nuestra población ejemplo  $U$  está ordenada de la siguiente forma

$$U = \{\text{Yves, Ken, Erik, Sharon, Leslie}\}$$

Suponga que sistemáticamente se divide en  $a = 2$  grupos. El primero dado por:

$$s_1 = \{\text{Yves, Erik, Leslie.}\}$$

y el segundo conformado por:

$$s_2 = \{\text{Ken, Sharon}\}$$

De tal forma que  $N = (2)(2) + 1$ . Para seleccionar un arranque aleatorio  $r$  se utilizará un dado, de tal forma que si el resultado de un lanzamiento es par, entonces la muestra seleccionada será  $s_1$ , de lo contrario la muestra seleccionada será  $s_2$ .

**Resultado 1.4.2.** *Para un diseño de muestreo sistemático, las probabilidades de inclusión de primer y segundo orden están dadas por*

$$\pi_k = \frac{1}{a} \quad (1.4.7)$$

$$\pi_{kl} = \begin{cases} \frac{1}{a} & \text{si } k \text{ y } l \text{ pertenecen a } s_r \\ 0 & \text{en otro caso} \end{cases} \quad (1.4.8)$$

respectivamente.

*Demostración.* considerando que el elemento  $k$ -ésimo sólo puede pertenecer a una y sólo una muestra  $s_r$ , tenemos que

$$\pi_k = Pr(k \in S) = Pr(\text{seleccionar la muestra } s_r) = \frac{1}{a} \quad (1.4.9)$$

Por otra parte, suponga que los elementos  $k$ -ésimo y  $l$ -ésimo pertenecen al grupo  $s_r$ . De esta manera, estos elementos son incluidos en la muestra sí y sólo sí se selecciona el grupo  $s_r$ , por tanto, la probabilidad de inclusión de segundo orden está dada por la probabilidad de selección del grupo  $s_r$  igual a  $\frac{1}{a}$ . Si los elementos  $k$ -ésimo y  $l$ -ésimo pertenecen a grupos distintos, la probabilidad de ser incluidos en la muestra realizada es nula.  $\square$

### 1.4.2 El estimador de Horvitz-Thompson

Una vez que el diseño de muestreo es definido, la estrategia se completa con el uso del estimador de Horvitz-Thompson, por ser este un diseño sin reemplazo. El siguiente resultado será útil para definir las propiedades de varianza del estimador.

**Resultado 1.4.3.** *Para un diseño  $p(\cdot)$  con soporte  $Q$ , la varianza del estimador de Horvitz-Thompson, se puede escribir como*

$$Var(\hat{t}_{y,\pi}) = \sum_U \sum_U \frac{\pi_{kl}}{\pi_k \pi_l} y_k y_l - \left( \sum_U y_k \right)^2 \quad (1.4.10)$$

*Demostración.* Partiendo del resultado 2.2.2., se tiene que

$$Var(\hat{t}_{y,\pi}) = \sum_U \sum_{kl} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \quad (1.4.11)$$

$$= \sum_U \sum_{kl} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \quad (1.4.12)$$

$$= \sum_U \sum_{kl} \left( \frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) y_k y_l \quad (1.4.13)$$

$$= \sum_U \sum_{kl} \frac{\pi_{kl}}{\pi_k \pi_l} y_k y_l - \sum_U \sum_{kl} y_k y_l \quad (1.4.14)$$

$$= \sum_U \sum_{kl} \frac{\pi_{kl}}{\pi_k \pi_l} y_k y_l - \left( \sum_U y_k \right)^2 \quad (1.4.15)$$

En donde se utiliza el hecho de que

$$\sum_U \sum_{kl} y_k y_l = \sum_{k \neq l} y_k y_l + \sum_U y_k^2 = \left( \sum_U y_k \right)^2 \quad (1.4.16)$$

□

**Resultado 1.4.4.** Para el diseño de muestreo sistemático, el estimador de Horvitz-Thompson y su varianza están dados por:

$$\hat{t}_{y,\pi} = at_{sr}, \quad (1.4.17)$$

con  $t_{sr} = \sum_{k \in S_r} y_k$ , y

$$Var_{SIS}(\hat{t}_{y,\pi}) = a \sum_{r=1}^a (t_{sr} - t)^2 \quad (1.4.18)$$

En este caso no existe estimador de la varianza.

*Demostración.* De la definición del estimador de Horvitz-Thompson y dado que las probabilidades de inclusión de primer orden son todas iguales al valor  $1/a$ , entonces

$$\hat{t}_{y,\pi} = \sum_{Sr} \frac{y_k}{\pi_k} = at_{sr} \quad (1.4.19)$$

Utilizando los dos anteriores resultados, se sigue que

$$Var(\hat{t}_{y,\pi}) = \sum_U \sum_{kl} \frac{\pi_{kl}}{\pi_k \pi_l} y_k y_l - \left( \sum_U y_k \right)^2 \quad (1.4.20)$$

$$= a \sum_{r=1}^a \left( \sum_{sr} y_k y_l \right) - t^2 \quad (1.4.21)$$

$$= a \sum_{r=1}^a \left( \sum_{k \in S_r} y_k \sum_{l \in S_r} y_l \right) - t^2 \quad (1.4.22)$$

$$= a \sum_{r=1}^a t_{sr}^2 - t^2 \quad (1.4.23)$$

$$= a \sum_{r=1}^a (t_{sr} - \bar{t})^2 \quad (1.4.24)$$

donde

$$\bar{t} = \sum_{r=1}^a \frac{t_{s_r}}{a} = \frac{t}{a} \quad (1.4.25)$$

Por la definición 3.4.1, algunas probabilidades de inclusión de segundo orden son nulas, por ello no se tiene un estimador de la varianza.  $\square$

Más allá de que los principios del estimador de Horvitz-Thompson no permitan estimar la varianza para este diseño, la razón genérica radica en que, de una forma u otra, se está seleccionando uno y sólo un grupo de elementos y se calcula un sólo total para el grupo. Como la selección es de sólo un grupo, no se tiene un marco de comparación y no se puede llegar a una estimación de la varianza.

### 1.4.3 Optimalidad de la estrategia

Una vez que la estrategia de muestreo queda definida, es indispensable tocar el tema de la configuración de los valores de la característica de interés mediante el ordenamiento particular que se tiene en el marco de muestreo. Bautista (1998) utiliza el siguiente esquema para explicar la eficiencia de esta estrategia de muestreo.

Cuadro 1.7: Configuración de totales por grupo.

Grupo	$s_1$	$\dots$	$s_r$	$\dots$	$s_a$
Valor de la característica	$y_1$		$y_r$		$y_k$
	$y_{1+a}$		$y_{r+a}$		$y_{2a}$
	$y_{1+2a}$		$y_{r+2a}$		$y_{3a}$
	$\dots$		$\dots$		$\dots$
Total de grupo	$y_{1+(n-1)a}$		$y_{r+(n-1)a}$		$y_{na}$
	$t_{s_1}$	$\dots$	$t_{s_r}$	$\dots$	$t_{s_a}$

Este diseño de muestreo puede resultar más eficiente que el diseño de muestreo aleatorio simple, dependiendo del ordenamiento del marco de muestreo. Es usado para palear las posibles imperfecciones generadas por un diseño de muestreo aleatorio simple. Por ejemplo, puede resultar que en una muestra simple, todos los elementos de la muestra seleccionada compartan una característica latente que perjudique la precisión de las estimaciones. En el caso de una población de personas, puede resultar que una muestra simple sólo incluya hombres. Cuando se sabe que el marco de muestreo está ordenado de manera aleatoria, es recomendable utilizar el diseño de muestreo aleatorio simple, porque asegura una muestra bien mezclada. Por ejemplo, si el marco de muestreo está ordenado alfabéticamente, es casi seguro que se obtendrá una muestra que sea representativa de la población, puesto que la posición alfabética no debería estar asociada con la característica de interés.

Además, mediante este diseño de muestreo, no es necesario poseer un marco de muestreo de forma física para poder realizar una muestra probabilística. Sin embargo, se debe tener cuidado con la especificación del diseño, pues como lo afirma Lohr (2000) no es lo mismo seleccionar una de cada 10 personas que entran a una biblioteca que seleccionar una de cada 10 personas que salen de un avión. En el segundo caso, existe de forma implícita, un marco de muestreo.

Como se verá más adelante, el diseño de muestreo sistemático puede ser más preciso que el diseño de muestreo aleatorio simple cuando los grupos  $s_r$  poseen mucha variación interna. De manera contraria, si el valor de los elementos dentro de los grupos proporciona la misma información, entonces la eficiencia del diseño se verá disminuida significativamente con respecto al diseño aleatorio simple.

La figura 3.4 muestra los tres casos más particulares en el uso de esta estrategia de muestreo cuyas características son las siguientes:

1. **Ordenamiento aleatorio:** cuando el ordenamiento del marco de muestreo no está relacionado con la característica de interés, la eficiencia de este diseño es comparable con la de muestreo aleatorio simple. Ordenamiento por orden alfabético.
2. **Ordenamiento lineal:** cuando el ordenamiento del marco de muestreo es tal que se puede observar una tendencia lineal, entonces la selección de una muestra sistemática obliga a que los valores de los elementos incluidos tengan una alta dispersión haciendo que el comportamiento de los grupos formados sea heterogéneo con respecto al valor de la característica de interés. Ordenamiento de registros contables.
3. **Ordenamiento periódico:** si la población es tal que se observa un patrón de tipo periódico, el muestreo sistemático puede arrojar peores resultados que una muestra aleatoria simple pues si el intervalo de muestreo coincide con el patrón de periodicidad, la muestra seleccionada incluiría elementos cuyos valores de la característica de interés serían muy parecidos. Una muestra seleccionada de esta manera no sería representativa de la población. En algunos casos es posible encontrar poblaciones con este tipo de comportamiento periódico; por ejemplo, el flujo vehicular durante las 24 horas del día o las ventas en negocios durante cierta temporada del año.

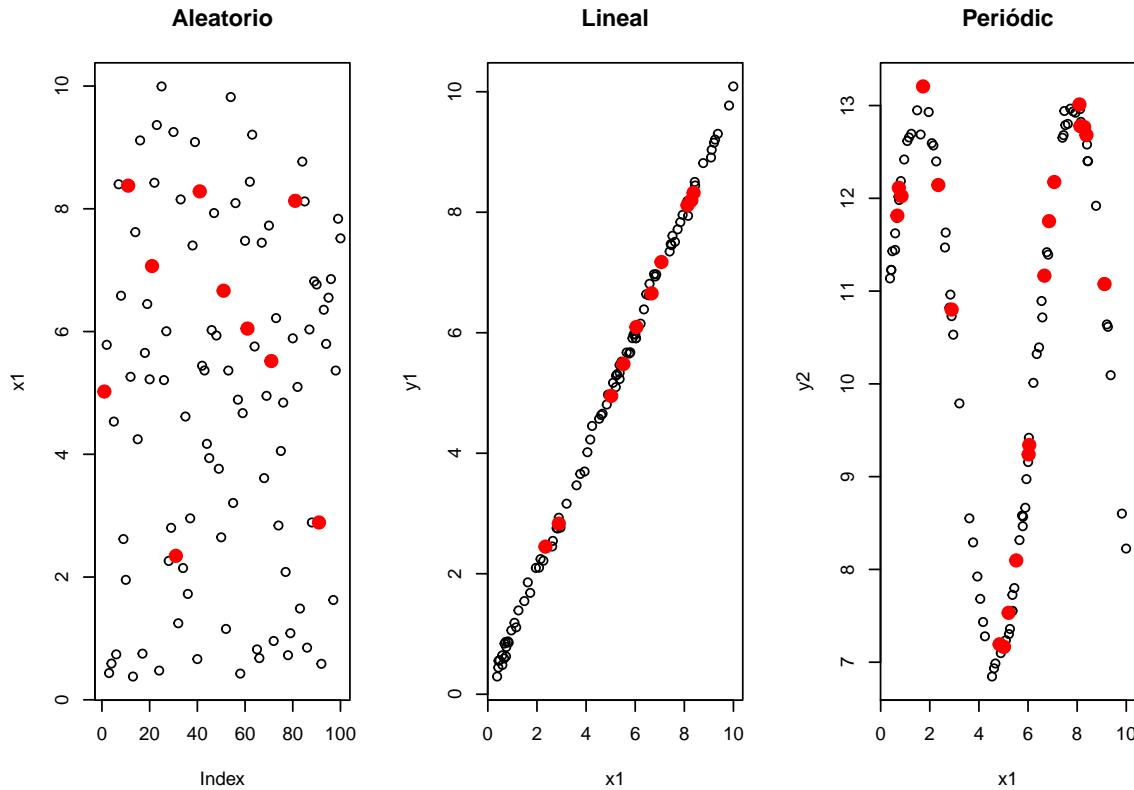


Figura 1.4: Casos de ordenamiento en muestreo sistemático.

### Descomposición de la varianza

Algunos críticos de la teoría del muestreo han querido separar el pensamiento estadístico de la metodología de estudios por muestreo. Lo anterior sumado a la falta de preparación del usuario del muestreo

ha abierto una brecha entre dos mundos. La verdad es que la estadística sin muestreo no está completa y viceversa Kish (1965). En estos apartes, debemos considerar uno de los resultados más importantes de la estadística que ha permitido el desarrollo de la misma en diversos campos de la vida práctica.

**Resultado 1.4.5.** *Suponga que la población se divide en  $a$  grupos, de tal forma que existen  $n$  elementos por grupo y el tamaño poblacional toma la forma  $N = an$ , entonces*

$$(N - 1)S_{y_U}^2 = \underbrace{\sum_U (y_k - \bar{y}_U)^2}_{SCT} = \underbrace{\sum_{r=1}^a \sum_{s_r} (y_{rk} - \bar{y}_{s_r})^2}_{SCD} + \underbrace{\sum_{r=1}^a n (\bar{y}_{s_r} - \bar{y}_U)^2}_{SCE} \quad (1.4.26)$$

La sigla **SCT** se refiere a la suma de cuadros del total de la población y no es otra cosa que el numerador en la fórmula del estimador de la varianza. El anterior resultado es importante porque permite descomponer la suma de cuadrados total en dos cantidades. Primero, **SCD** que denota la suma de cuadrados dentro (al interior) de los grupos y segundo, **SCE** que hace referencia a la suma de cuadrados entre los grupos. Por supuesto, la varianza como parámetro poblacional es fija, por tanto si

1. **SCE** es alta, entonces **SCD** es baja, indicando así que los grupos están construidos de tal forma que resultan ser muy heterogéneos entre sí, pero dentro de ellos existe homogeneidad.
2. **SCE** es baja, entonces **SCD** es alta, lo que quiere decir que los grupos son muy disímiles en su interior, pero entre ellos tienen un comportamiento similar.

Esta representación de la descomposición de la varianza, se puede ver claramente en una tabla de ANOVA (análisis de varianza, por sus siglas en inglés), de la siguiente manera.

Cuadro 1.8: Tabla de ANOVA inducida por el muestreo sistemático.

Fuente	gl	Suma de cuadrados	Cuadrado medio
Entre	$a - 1$	$SCE = \sum_{r=1}^a n (\bar{y}_{s_r} - \bar{y}_U)^2$	$\frac{SCE}{a - 1}$
Dentro	$N - a$	$SCD = \sum_{r=1}^a \sum_{s_r} (y_{rk} - \bar{y}_{s_r})^2$	$\frac{SCD}{N - a}$
Total	$N - 1$	$SCT = \sum_U (y_k - \bar{y}_U)^2$	$s_{y_U}^2$

Desde un punto de vista totalmente pragmático, la estrategia de muestreo tendrá un mejor desempeño cuando la variabilidad total entre los grupos sea mínima y la variabilidad dentro de los grupos sea máxima. El siguiente resultado da una mejor comprensión de la descomposición de la varianza en los grupos. Es decir, la varianza del estimador de Horvitz-Thompson, bajo muestreo sistemático, será cercana a cero cuando el ordenamiento de los grupos en la población es tal que los totales  $t_{s_r}$  con  $r = 1, \dots, a$  son similares

$$t_{s_1} \approx t_{s_2} \approx \dots \approx t_{s_a} \approx \bar{t} \quad (1.4.27)$$

**Resultado 1.4.6.** *Sin pérdida de generalidad, considere que el tamaño muestral es tal que  $N = na$ , entonces la varianza del estimador de Horvitz-Thompson bajo un diseño de muestreo sistemático toma la siguiente forma*

$$Var_{SIS}(\hat{t}_{y,\pi}) = N \sum_{r=1}^a n (\bar{y}_{s_r} - \bar{y}_U)^2 = N(SCE) \quad (1.4.28)$$

*Demostración.* Partiendo de la definición de la varianza del estimador de Horvitz-Thompson en muestreo sistemático, se tiene que

$$\begin{aligned} Var_{SIS}(\hat{t}_{y,\pi}) &= a \sum_{r=1}^a (t_{sr} - \bar{t})^2 \\ &= \frac{N}{n} \sum_{r=1}^a (n\bar{y}_{sr} - n\bar{y}_U)^2 \\ &= \frac{N}{n} \sum_{r=1}^a n^2 (\bar{y}_{sr} - \bar{y}_U)^2 \\ &= N \sum_{r=1}^a n (\bar{y}_{sr} - \bar{y}_U)^2 = N(SCE) \end{aligned}$$

□

Por tanto, se quiere que toda la variabilidad esté por dentro de cada uno de los grupos.

**Definición 1.4.2.** Se define el coeficiente de correlación intra-clase como

$$\rho = 1 - \frac{n}{n-1} \frac{SCD}{SCT} \quad (1.4.29)$$

Esta medida de correlación entre los pares de elementos de los grupos formados toma una valor máximo igual a uno cuando **SCE** es nula y toma un valor mínimo igual a  $-\frac{1}{n-1}$  cuando **SCE** es máxima. En particular, es deseable para esta estrategia que  $\rho$  tome valores cercanos a cero.

**Resultado 1.4.7.** Utilizando la relación 1.4.26  $SCT=SCE+SCD$  se tiene que

$$SCE = SCT \left[ (\rho - 1) \frac{n-1}{n} + 1 \right] \quad (1.4.30)$$

*Demostración.* De la definición del coeficiente de correlación intra-clase se tiene que

$$\begin{aligned} (\rho - 1) \frac{n-1}{n} + 1 &= 1 - \frac{SCD}{SCT} \\ &= \frac{SCE}{SCT} \end{aligned}$$

por tanto al despejar  $SCE$  se tiene el resultado. □

**Resultado 1.4.8.** Con el anterior resultado no es difícil verificar que la varianza del estimador de Horvitz-Thompson bajo muestreo sistemático se puede escribir como

$$Var_{SIS}(\hat{t}_{y,\pi}) = \underbrace{\frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{yU}^2}_{Var_{MAS}(\hat{t}_{y,\pi})} \left\{ \frac{N-1}{N-n} [1 + (n-1)\rho] \right\} \quad (1.4.31)$$

*Demostración.* Partiendo de la última expresión tenemos que

$$\begin{aligned} \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{yU}^2 \left\{ \frac{N-1}{N-n} [1 + (n-1)\rho] \right\} &= \frac{N}{n} SCT [1 + (n-1)\rho] \\ &= N(SCT) \left[ 1 - \frac{SCD}{SCT} \right] \\ &= N(SCE) \\ &= Var_{SIS}(\hat{t}_{y,\pi}) \end{aligned}$$

que coincide con la varianza del estimador de Horvitz-Thompson en muestreo sistemático □

Nótese que la primera parte de la anterior ecuación se refiere al valor del estimador de Horvitz-Thompson bajo un diseño de muestreo aleatorio simple sin reemplazo. Siguiendo esta idea, el efecto de diseño está dado por el siguiente resultado.

**Resultado 1.4.9.** *El efecto de diseño de la estrategia de muestreo que utiliza un diseño sistemático y el estimador de Horvitz-Thompson está dado por*

$$Deff = \frac{Var_{SIS} \hat{t}_\pi}{Var_{MAS} \hat{t}_\pi} = \frac{N-1}{N-n} [1 + (n-1)\rho] \quad (1.4.32)$$

Dado el efecto de diseño, se concluye que esta estrategia de muestreo es

1. Igual de eficiente al muestreo aleatorio simple si  $\rho = \frac{1}{1-N}$ .
2. Menos eficiente que el muestreo aleatorio simple si  $\rho > \frac{1}{1-N}$ .
3. Más eficiente que el muestreo aleatorio simple si  $\rho < \frac{1}{1-N}$ .

*Demostración.* La demostración es inmediata teniendo en cuenta el anterior resultado.  $\square$

#### 1.4.4 Diseño de muestreo $q$ -sistemático

Cuando la periodicidad es un problema o cuando se quiere tener un estimativo insesgado de la varianza del estimador de Horvitz-Thompson, Mahalanobis (1946) propone el uso de muestras sistemáticas interpenetradas. Este método consiste en seleccionar, no una, sino  $q$  muestras sistemáticas. De esta manera se seleccionan  $q$  arranques aleatorios en grupos de tamaño  $aq$ , de tal manera que el tamaño poblacional se escribe como  $N = a \frac{n}{q} + c$ .

**Definición 1.4.3.** *El diseño de muestreo sistemático con  $q$  réplicas está definido como*

$$p(s) = \frac{1}{\binom{a}{q}} \quad \text{para todo } s \in Q_r \quad (1.4.33)$$

con  $Q_r$  definido en 3.4.4.

Por supuesto, la cardinalidad del soporte es  $\#Q_r = \binom{a}{q}$ , por tanto este diseño de muestreo cumple las propiedades del capítulo anterior. Teniendo en cuenta que se han formado  $a$  grupos, entonces el diseño de muestreo  $q$ -sistemático puede ser visto como un diseño MAS de tamaño de muestra igual a  $q$  de los totales de todos los grupos. Una vez más, estos grupos también pueden ser vistos como conglomerados.

**Resultado 1.4.10.** *Para un diseño de muestreo sistemático, las probabilidades de inclusión de primer y segundo orden están dadas por*

$$\pi_k = \frac{q}{a} \quad (1.4.34)$$

$$\pi_{kl} = \begin{cases} \frac{q}{a} & \text{si } k \text{ y } l \text{ pertenecen a } s_r \\ \frac{q(q-1)}{a(a-1)} & \text{en otro caso} \end{cases} \quad (1.4.35)$$

respectivamente.

**Resultado 1.4.11.** *Para el diseño de muestreo sistemático con  $q$  réplicas, el estimador de Horvitz-Thompson y su varianza están dados por:*

$$\hat{t}_{y,\pi} = \frac{a}{q} \sum_S t_{sr} \quad (1.4.36)$$

$$VarSIS(\hat{t}_{y,\pi}) = \frac{a^2}{q} \left(1 - \frac{q}{a}\right) S_{t_{sr}U}^2 \quad (1.4.37)$$

$$\widehat{VarSIS}(\hat{t}_{y,\pi}) = \frac{a^2}{q} \left(1 - \frac{q}{a}\right) S_{t_{sr}s}^2 \quad (1.4.38)$$

respectivamente, con  $S_{t_{sr}U}^2$  y  $S_{t_{sr}s}^2$  el estimador de la varianza de los totales de la característica de interés  $y$  en cada grupo  $s_r$  del universo  $y$  en la muestra. Nótese que  $\hat{t}_{y,\pi}$  es insesgado para el total poblacional  $t_y$  de la característica de interés  $y$ , y que  $\widehat{VarSIS}(\hat{t}_{y,\pi})$  es insesgado para  $VarSIS(\hat{t}_{y,\pi})$ .

Al respecto de esta estrategia, el lector debe notar que:

- La varianza del estimador de Horvitz-Thompson bajo el diseño de muestro  $q$ -sistématico crece cuando se aplica a un universo que está ordenado igualmente de forma sistemática.
- La varianza del estimador de Horvitz-Thompson bajo el diseño de muestro  $q$ -sistématico depende del ordenamiento de los valores de la característica de interés por lo que puede suceder que ésta no sea monótonamente decreciente en función del tamaño de muestra.
- El efecto de la correlación intra-clase tiene una gran repercusión en el tamaño de muestra; si existe una alta correlación intra-clase entonces el tamaño de muestra debe ser mayor para tener un *c.v.e* pequeño y viceversa.
- En estudios de tipo electoral se dice que un candidato tiene alta correlación intra-clase (por ejemplo en los barrios) cuando la imagen del candidato está polarizada. Es decir, la mayoría de votación en determinado barrio es muy alta por el candidato o muy baja. Por otro lado, se dice que la campaña electoral tiene baja correlación intra-clase cuando la votación en los barrios no es ni muy baja ni muy alta.

#### 1.4.5 Marco y Lucy

En nuestro intento de obtener estimaciones precisas para la evaluación del comportamiento del sector industrial en lo corrido del último año fiscal, hemos notado que el marco de muestreo está ordenado de manera alfanumérica en orden ascendente por el rótulo de identificación industrial. Además, se sabe que el número de identificación de cada empresa no tiene una secuencia específica, sino que es asignado de acuerdo a la fecha de registro de la empresa. De tal forma, la primera empresa en ser registrada ante el organismo gubernamental competente es la identificada con el número de identificación **AB001** y la última empresa en ser registrada es la identificada con el número **AB987**.

Nótese que las características de interés son Ingreso, número de empleados e impuestos declarados en el último año fiscal y se supone, de manera correcta, que estas características no tienen ninguna relación con la fecha de registro de la empresa. Así, puede suceder que una empresa joven, tenga unos altos réditos, pocos empleados y una alta declaración de impuestos, pero puede suceder lo contrario; de hecho, este comportamiento está sujeto a la estrategia de *marketing* utilizada en cada periodo comercial y no a la antigüedad del negocio. Por las anteriores razones, se supone que el ordenamiento del marco de muestreo es completamente aleatorio.

Se ha decidido que la población va a ser particionada en seis grupos, de tal forma que el tamaño efectivo de muestra será 399 o 400. El marco de muestreo es cargado en el ambiente de R.

```
data(BigLucy)
attach(BigLucy)
```

```
N <- dim(BigLucy)[1]
a <- 40
floor(N/a)

## [1] 2132
```

El procedimiento que se sigue es la creación de los grupos sistemáticos. Esto puede realizarse con la función (`array(1:a,N)`) que permite la creación de la secuencia **1,2,3,4,5,6,1,2,3,4,5,6,1,2,...**; sin embargo, es indispensable definir este arreglo como un factor, es decir como una variable de tipo categórica nominal cuyos rótulos significan la pertenencia de un individuo a un grupo.

La selección de la muestra se realiza mediante la función `S.SY` del paquete `TeachingSampling` cuyos argumentos son `N`, el tamaño de la población y `a`, el número de grupos. Esta función sigue el algoritmo secuencial descrito en esta estrategia de muestreo y lo que hace es aleatoriamente asignar un arranque aleatorio y saltar, en este caso, de seis en seis elementos hasta barrer toda la lista. El resultado de la función es un listado de índices que aplicados a la población resulta en los valores de las características de interés de los elementos incluidos en la muestra realizada.

```
sam <- S.SY(N, a)
muestra <- BigLucy[sam,]
attach(muestra)

head(muestra)

##           ID Ubication Level Zone Income Employees Taxes
## 12 AB00000000012 C0033329K0268568 Small County1    419      20     7
## 52 AB00000000052 C0038888K0263009 Small County1    380      90     6
## 92 AB00000000092 C0208289K0093608 Small County1    460      79     9
## 132 AB0000000132 C0100864K0201033 Small County1   304      18     4
## 172 AB0000000172 C0299521K0002376 Small County1   310      86     4
## 212 AB0000000212 C0189164K0112733 Small County1   280      77     3
##          SPAM ISO Years Segments
## 12      no   no    42 County1 2
## 52     yes  no    18 County1 6
## 92      no  no    39 County1 10
## 132     no  no    23 County1 14
## 172     no  no    37 County1 18
## 212     no  no    48 County1 22

n <- dim(muestra)[1]
n

## [1] 2133
```

En el anterior caso particular, el arranque aleatorio fue igual a tres; por tanto, la muestra está conformada por los elementos **3, 9, ..., 2385 y 2391** del marco de muestreo. Una vez recolectada la información de la muestra, se procede a realizar la estimación mediante el uso de la función<sup>4</sup> `E.SY` del paquete `TeachingSampling` cuyos argumentos son `N`, `a` y un conjunto de datos contenido la información de las características de interés para cada elemento en la muestra.

<sup>4</sup>Dado que no existe el estimador genérico para la varianza del estimador de Horvitz-Thompson, esta función utiliza una aproximación conservadora de la varianza suponiendo que se realizó un muestreo aleatorio simple.

```
estima <- data.frame(Income, Employees, Taxes)
E.SY(N, a, estima)
```

Los resultados de la estimación se muestran en la tabla 1.9. Es de considerar que la eficiencia de esta estrategia de muestreo es mucho mayor a la de una estrategia que utilice un diseño de muestreo aleatorio simple. Nótese que los coeficientes de variación son mucho menores y también, aunque este es un argumento un poco más débil, la desviación relativa es menor.

Cuadro 1.9: *Estimaciones para el diseño de muestreo sistemático*

	N	Income	Employees	Taxes
Estimation	85320.00	36772240.00	5378960.00	1024500.00
Standard Error	0.00	494734.35	61139.85	32063.31
CVE	0.00	1.35	1.14	3.13
DEFF		1.00	1.00	1.00

Es hora de preguntarse, ¿por qué los resultados de las estimaciones son mejores que en otro tipo de estrategias de muestreo? Vamos a realizar un procedimiento de evaluación, puramente académico, y vamos a suponer que tenemos acceso a la información de la característica de interés a nivel poblacional.

En primer lugar, se realiza un análisis de varianza para obtener la descomposición de las sumas de cuadrados para la característica de interés `Income`. Para esto usamos la función `lm` que relaciona a la variable de interés con un factor de agrupamiento. La variable `grupo` fue creada como un vector de cinco niveles y puede ser usada en este caso. Aplicando la función `anova` al modelo, se obtiene una tabla de sumas de cuadrados.

```
data(BigLucy)
attach(BigLucy)
```

```
N<-dim(BigLucy)[1]
n<-2133
a<-floor(N/n)
c<-N-floor(N/n)*n
a*n+c

## [1] 85296

grupo<-as.factor(array(1:a,N))
anova(lm(Income~grupo))

## Analysis of Variance Table
##
## Response: Income
##              Df    Sum Sq Mean Sq F value Pr(>F)
## grupo          38    58913   1550     0.02      1
## Residuals 85257 6029937065   70727
```

Siguiendo a Dalgaard (2008), en la mayoría de textos estadísticos (incluyendo el que el lector tiene en sus manos) las sumas de cuadrados son rotuladas como SCD, SCE y SCT. Sin embargo, R usa una rotulación diferente. La variación **entre** los grupos es rotulada con el nombre del factor de agrupación, en este caso `grupo`. La variación **dentro** de los factores de agrupación es rotulada como `Residuals`.

Por tanto, se observa que la variación total se encuentra dentro de los grupos; mientras que existe una baja variación entre los grupos. Esto es bueno para efectos de la eficiencia de la estrategia.

Por un lado, al observar la gráfica de la característica de interés con respecto al ordenamiento natural del marco de muestreo, no es posible identificar un patrón lineal o de periodicidad, cuando realizamos el gráfico con respecto a los grupos, nos damos cuenta de que dentro de ellos existe una muy alta variabilidad y más aún, los cinco grupos tiene un comportamiento parecido entre ellos. El código necesario para la creación de este gráfico está dado a continuación.

```
stripchart(Income ~ grupo)
```

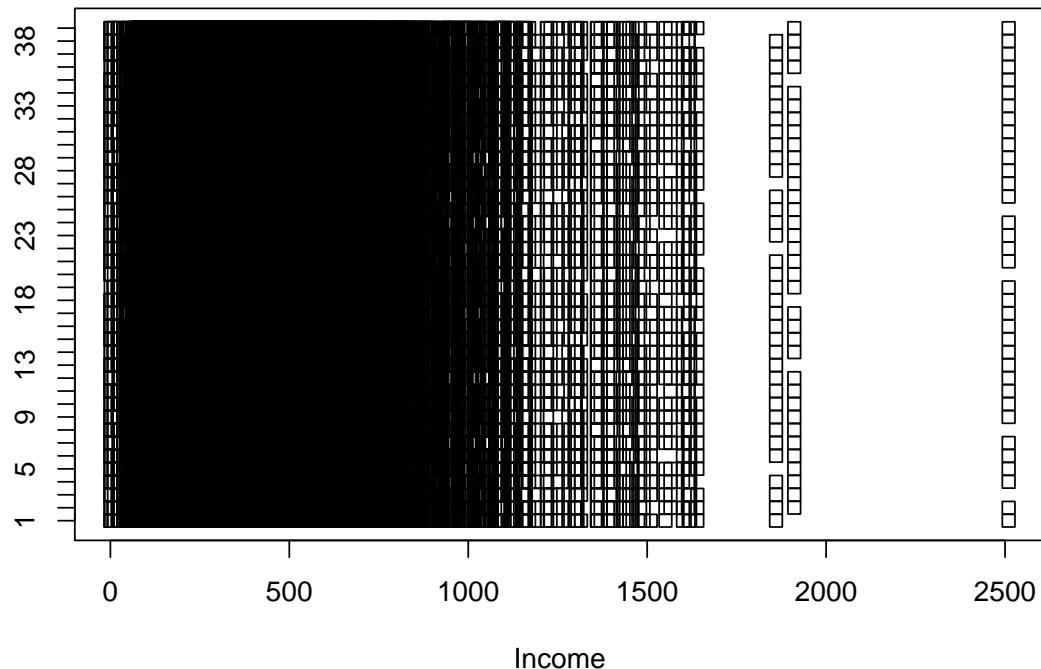


Figura 1.5: Distribución de la característica *Income* con respecto a los grupos creados en el muestreo sistemático.

Por otro lado, el ordenamiento aleatorio se observa muy claramente en la figura 3.6., en dónde los puntos marcados corresponden a los elementos seleccionados. Nótese la buena dispersión de la muestra en la población, haciéndola representativa. El código necesario para la creación de este gráfico es el siguiente.

```
sam <- seq(1, N, by=a)
plot(Income)
points(sam, Income[sam], col="red", pch=19)
```

Es claro que esta estrategia de muestreo resultó más eficiente que la estrategia de muestreo aleatorio simple. Pero, ¿cuánto más eficiente?. Con unos simples cálculos algebraicos se obtiene un coeficiente de

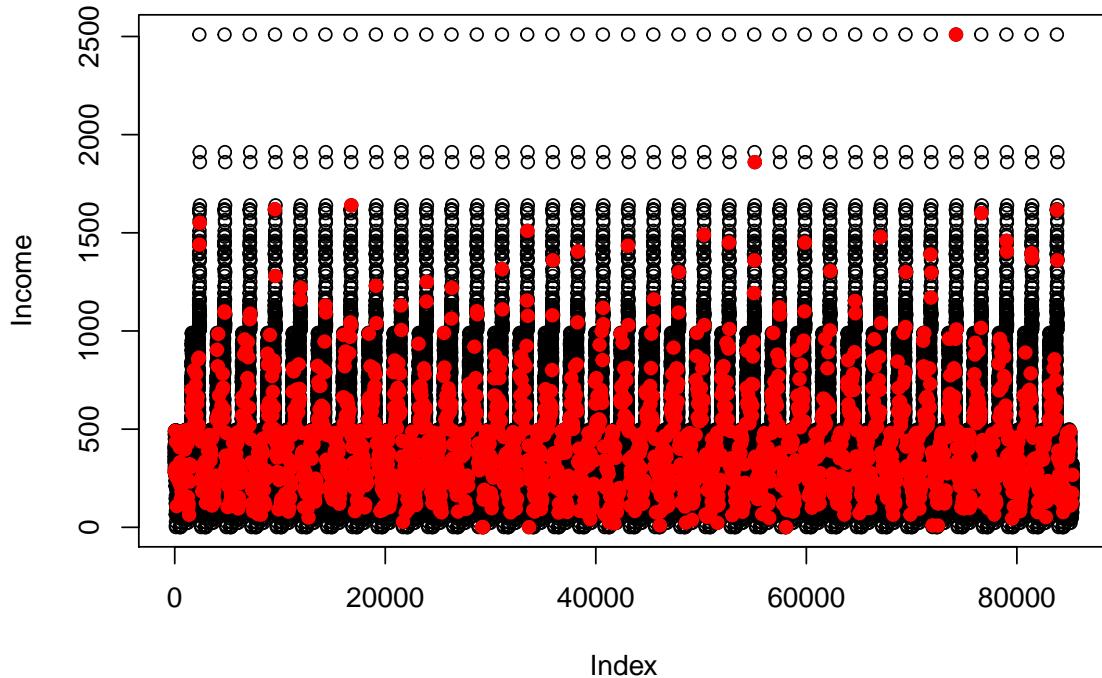


Figura 1.6: Casos seleccionados en muestreo sistemático.

correlación intra-clase muy cercano a cero y esto es bueno puesto que cumple con los requerimientos en la definición de  $\rho$ .

```
SCD <- anova(lm(Income~grupo))$Sum[1]
SCE <- anova(lm(Income~grupo))$Sum[2]
rho <- 1 - (n / (n-1)) * (SCE / (SCD + SCE))
rho

## [1] -0.00046

rho > 1 / (1 - N)

## [1] FALSE
```

Sin embargo, lo verdaderamente asombroso es que la ganancia en eficiencia al usar este diseño es de veintinueve veces puesto que el efecto de diseño es aproximadamente 0.02.

```
VarHT <- N * SCD
VarHT

## [1] 5025031348
```

```
Deff <- (N - 1) * (1 + (n - 1) * rho) / (N - n)
Deff

## [1] 0.021
```

Los anteriores diseños de muestreo pertenecen al grupo de los diseños de probabilidad de inclusión constante. En el siguiente capítulo veremos diseños con probabilidad de inclusión proporcional al tamaño que hace uso de información auxiliar continua en el marco de muestreo.

## 1.5 Ejercicios

3.1 Suponga una población de 10 elementos  $U = \{e_1, e_2, \dots, e_{10}\}$ .

- Seleccione una muestra mediante un diseño Bernoulli con probabilidad de inclusión  $\pi = 0.4$ , utilizando el algoritmo de la sección 3.1.1. y teniendo en cuenta que para cada elemento en la población se obtuvo el siguiente conjunto de números aleatorios uniformes

$$\varepsilon = \{0.152, 0.158, 0.614, 0.593, 0.140, 0.851, 0.803, 0.996, 0.433, 0.790\}$$

- Otra manera de seleccionar una muestra Bernoulli es generando un sólo número aleatorio de una distribución  $Binomial(N, \pi)$ ; este valor generado es el tamaño de muestra  $n(S)$  y con ayuda del marco de muestreo se selecciona una muestra aleatoria simple de tamaño  $n(S)$ . Suponiendo que la realización de  $Binomial(10, 0.4)$  fue  $n(s) = 5$ , utilice el algoritmo coordinado negativo para la selección de una muestra, teniendo en cuenta que para cada elemento en la población se obtuvo el siguiente conjunto de números aleatorios uniformes

$$\xi = \{0.370, 0.561, 0.064, 0.412, 0.952, 0.461, 0.256, 0.275, 0.213, 0.443\}$$

3.2 Complete el cálculo léxico-gráfico del ejemplo 3.1.1.

3.3 En un estudio de calidad de vida en cárceles, se utilizó un diseño de muestreo Bernoulli con probabilidad de inclusión  $\pi = 0.15$  para seleccionar una muestra de reclusos. En la penitenciaría hay 1243 reclusos y se observaron las características de interés **CVDP** y **OTMA** para los presos incluidos en la muestra. Además se obtuvieron los siguientes resultados

Característica	$\sum_s y_k$	$\sum_s y_k^2$
CVDP	5412	95299
OTMA	82503	604926

- Utilice el estimador de Horvitz-Thompson para calcular una estimación del total poblacional, el coeficiente de variación estimado y un intervalo de confianza al 95 % para estas características de interés.
- Utilice el estimador de Horvitz-Thompson para calcular una estimación de la media poblacional, el coeficiente de variación estimado y un intervalo de confianza al 95 % para estas características de interés.
- Si el tamaño de muestra efectivo fue 191, utilice el estimador alternativo para calcular una estimación del total poblacional y de la media poblacional.

3.4 Suponga una población de 12 elementos  $U = \{e_1, e_2, \dots, e_{12}\}$ . Seleccione una muestra aleatoria simple sin reemplazo de tamaño  $n = 4$  utilizando el algoritmo de Fan-Muller-Rezucha teniendo

en cuenta que para cada elemento en la población se obtuvo el siguiente conjunto de números aleatorios uniformes

$$\xi = \{0.787, 0.946, 0.766, 0.338, 0.520, 0.849, 0.828, 0.165, 0.416, 0.105, 0.069, 0.853\}$$

3.5 Complete el cálculo léxico-gráfico del ejemplo 3.2.2.

3.6 Demuestre o refute la siguiente afirmación: «En muestreo aleatorio simple, para la estimación de un total poblacional, el estimador de Horvitz-Thompson coincide con el estimador alternativo».

3.7 Demuestre o refute la siguiente afirmación: «En muestreo aleatorio simple, para la estimación de un total en dominios de interés, se cumple siempre que  $\sum_{d=1}^D \hat{t}_{y_d, \pi} > \hat{t}_{y, \pi}$ ».

3.8 Demuestre o refute la siguiente afirmación: «En muestreo aleatorio simple, el coeficiente de variación estimado del estimador de Horvitz-Thompson para el total poblacional es menor que el coeficiente de variación estimado del estimador de Horvitz-Thompson para la media poblacional».

3.9 En un estudio de satisfacción empresarial en una entidad prestadora de salud que sirve a 748 asociados, se quiere averiguar el promedio del número de horas al mes (**NHM**) que los asociados permanecen en consulta médica. Para esto se planea un muestreo aleatorio simple pues se conoce que, para este caso particular, una aproximación para la varianza de esta característica de interés es de 3.4839 y para el coeficiente de variación es de 0.5324.

- Con una confianza del 95 %, determine el tamaño de muestra mínimo para estimar el parámetro de interés con un error absoluto no mayor 15 minutos.
- Con una confianza del 95 %, determine el tamaño de muestra mínimo para estimar el parámetro de interés con un error relativo no mayor a 2 %.

3.10 Demuestre las siguientes igualdades

$$(n - 1)S_{y_S}^2 = \sum_{k \in S} (y_k - \bar{y}_S)^2 = \sum_{k \in S} y_k^2 - \frac{(\sum_{k \in S} y_k)^2}{n}$$

$$(N - 1)S_{y_U}^2 = \sum_{k \in U} (y_k - \bar{y}_U)^2 = \sum_{k \in U} y_k^2 - \frac{(\sum_{k \in U} y_k)^2}{N}$$

3.11 Demuestre rigurosamente los resultados 3.2.7 y 3.2.8.

3.12 Para el ejercicio 3.9, suponga que se deciden realizar  $n = 50$  entrevistas y que se obtuvo que  $\sum_s y_k = 178$  y  $\sum_s y_k^2 = 826$ . A continuación se presenta una tabla de frecuencias de las observaciones

NHM	0	1	2	3	4	5	6	7	8
Frecuencia	1	5	13	9	7	4	6	4	1

- Obtenga una estimación de Horvitz-Thompson para el total de horas mensuales que los asociados permanecen en consulta médica, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.
- Obtenga una estimación de Horvitz-Thompson para el promedio de horas mensuales que los asociados permanecen en consulta médica, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.

- Obtenga una estimación de Horvitz-Thompson para el total de asociados que permanecen en consulta médica menos (estrictamente) de cuatro horas, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.
- Obtenga una estimación de Horvitz-Thompson para la proporción de asociados que permanecen en consulta médica, más (estrictamente) de seis horas, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.

3.13 Complete el cálculo léxico-gráfico del ejemplo 3.3.3.

3.14 Para una población de  $N = 10$  elementos se planeó diseño aleatorio simple con reemplazo de tamaño de muestra  $m = 6$ . Complete la siguiente salida del algoritmo secuencial utilizado para la extracción de la muestra

k	nbin	pbin	nk
[1,]			0
[2,]	6	0.1111111	3
[3,]			1
[4,]	2	0.1428571	0
[5,]		0.1666667	1
[6,]	1		
[7,]	1	0.2500000	0
[8,]			0
[9,]	1		0
[10,]	1		1

3.15 Suponga que se realizó un muestreo aleatorio simple con reemplazo para la población del ejercicio 3.3.

- Utilice el estimador de Hansen-Hurwitz para obtener una estimación del total poblacional para características de interés **CVDP** y **OTMA**, reporte el coeficiente de variación estimado y un intervalo de confianza del 95 %.
- Bajo el supuesto de muestreo aleatorio simple con reemplazo, construya las probabilidades de inclusión de primer y segundo orden y utilice el estimador de Horvitz-Thompson para calcular una nueva estimación del total poblacional para las características de interés.

3.16 Demuestre o refute la siguiente afirmación: «Para tamaños de muestra iguales, la estrategia de muestreo aleatorio simple con reemplazo junto con el estimador de Hansen-Hurwitz es siempre de menor varianza que la estrategia de muestreo aleatorio simple sin reemplazo junto con el estimador de Horvitz-Thompson».

3.17 Demuestre o refute la siguiente afirmación: «El diseño de muestreo sistemático es de tamaño de muestra fijo».

3.18 Demuestre o refute la siguiente afirmación: «Aunque no existe la estimación de la varianza del estimador de Horvitz-Thompson en muestreo sistemático, es siempre conveniente reemplazarla por la expresión de la varianza estimada en un diseño aleatorio simple».

3.19 Para estimar el total de horas diarias que los estudiantes permanecen en la biblioteca de una universidad, se utilizó un diseño de muestreo sistemático con dos arranques aleatorios. La población fue dividida en siete grupos latentes y se seleccionó una muestra simple de dos enteros entre el uno y el siete. Los enteros seleccionados son el 3, y 7. Lo anterior implica que la muestra de estudiantes, que serán entrevistados a la salida de la biblioteca, está conformada por dos grupos. A saber el

grupo  $s_3$  conformado por los estudiantes 3, 10, 17, ... y el grupo  $s_7$  conformado por los estudiantes 7, 14, 21, ... Los resultados del sondeo para los dos grupos se dan a continuación

$$t_{s_3} = \sum_{s_3} y_k = 3574 \quad t_{s_7} = \sum_{s_7} y_k = 5024$$

Calcule una estimación insesgada para el número total de horas de permanencia en la biblioteca, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.

- 3.20 Suponga una población de 9 elementos cuyos valores para la característica de interés se dan a continuación

$$\mathbf{y} = \{23, 20, 24, 31, 24, 29, 25, 33, 21\}$$

- Utilice el análisis de varianza (ANOVA) para calcular la varianza del estimador de Horvitz-Thompson en un diseño de muestreo sistemático simple con  $a = 2$  grupos.
- Calcule el coeficiente de variación intra-clase y el efecto de diseño. Decida si, para este caso particular, el diseño sistemático es más eficiente que el diseño de muestreo aleatorio simple.

- 3.21 Demuestre o refute la siguiente afirmación: «En un diseño de muestreo sistemático, si hay homogeneidad dentro de los grupos y heterogeneidad entre sus medias, entonces este diseño es menos eficiente que el diseño de muestreo aleatorio simple».



# Bibliografía

- Bautista, J. (1998), *Diseños de muestreo estadístico*, Universidad Nacional de Colombia.
- Bebington, A. (1975), ‘A simple method of drawing a sample without replacement’, *Applied Statistics* **24**, 136.
- Cornfield, J. (1951), ‘The determination of sampling size’, *American journal of public health* **41**, 654–661.
- Dalgaard, P. (2008), *Introductory Statistics with R*, 2 edn, Springer.
- Durbin, J. (1967), ‘Design of multi-stage surveys for the estimation of sampling errors’, *Applied statistics* **16**, 152–164.
- Fan, C., Muller, M. & Rezucha, I. (1962), ‘Development of sampling plans by using sequential (item by item) selection techniques and digital computer’, *Journal of the American Statistical Association* **57**, 387–402.
- Frankel, M. & King, B. (1996), ‘A conversation with leslie kish’, *Statistical Science* **11**, 65–87.
- Hájek, J. (1960), ‘Limiting distributions in simple random sampling from a finite poulation’, *Publication of Mathematical Institute of the Hungarian Academy of Science* **5**, 361–374.
- Hartley (1959), ‘Analytic studies of survey data’, *Instituto di Statistica Volume in honor of Corrado Gini*.
- Kish, L. (1965), *Survey Sampling*, Wiley.
- Lehtonen, R. & Pahkinen, E. (2003), *Practical methods for design and analysis of complex surveys*, 2 edn, New York: Wiley.
- Lohr, S. (2000), *Sampling: Design and Analysis*, Thompson.
- Mahalanobis, P. (1946), ‘Recent experiment in statistical sampling in the indian statitical institute’, *Journal of the Royal Statistical Society* **109**, 325–370.
- Mood, A. M., Graybill, F. A. & Boes, D. C. (1974), *Introduction to the Theory of Statistics*, 3 edn, McGraw Hill.
- Ospina, D. (2001), *Introducción al muestreo.*, Universidad Nacional de Colombia.
- Raj, D. (1968), *Sampling theory*, McGraw Hill.
- Särndal, C., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer, New York.
- Sunter, A. (1977), ‘List sequential sampling with equal or unequal probabilities without replacement’, *Applied Statistics* **26**, 261–268.

- Tillé, Y. (2006), *Sampling Algorithms*, Springer.
- Wu, C. (2003), ‘Optimal calibration estimators in survey sampling’, *Biometrika* **90**, 937–951.