

# ESTRATEGIAS DE MUESTREO DISEÑO DE ENCUESTAS Y ESTIMACIÓN DE PARÁMETROS

SEGUNDA EDICIÓN



# ESTRATEGIAS DE MUESTREO DISEÑO DE ENCUESTAS Y ESTIMACIÓN DE PARÁMETROS

SEGUNDA EDICIÓN

Andrés Gutiérrez, PhD.

ISBN: 978-958-631-608-8  
Derechos reservados  
Bogotá, D.C., 2015



# Capítulo 1

## Encuestas y estudios por muestreo

Durante todo el siglo pasado, ha surgido una serie de teorías y principios que ofrecen un marco de referencia unificado en el diseño, implementación y evaluación de encuestas. Este marco de referencia se conoce comúnmente como el paradigma del «error total de muestreo» y ha encaminado la investigación moderna hacia una mejor calidad de las encuestas.

Groves, Fowler, Couper, Lepkowski, Singer & R. (2004)

Este capítulo, a manera de introducción, busca identificar los principios (no matemáticos) del diseño, recolección, procesamiento y análisis de los estudios por muestreo, cuyo crecimiento va en aumento al pasar de los años, pero que sigue teniendo ciertas limitantes de tipo económico y logístico. Un estudio por muestreo involucrará a profesionales de diferentes disciplinas quienes se ocupan de la reducción de costos y el aumento de la calidad de las estimaciones. Un gran campo de la ciencia estadística se preocupa por minimizar los errores muestrales mientras que, por otra parte, otro gran campo de las ciencias sociales se ocupa en minimizar los errores que pueden ser cometidos en el periodo de la recolección de los datos.

### 1.1 Conceptos metodológicos

El muestreo es un procedimiento que responde a la necesidad de información estadística precisa sobre la población y los conjuntos de elementos que la conforman; el muestreo trata con investigaciones parciales sobre la población que apuntan a inferir a la población completa. Es así como en las últimas décadas ha tenido bastante desarrollo en diferentes campos principalmente en el sector gubernamental con la publicación de las estadísticas oficiales que permiten realizar un seguimiento a las metas del gobierno, en el sector académico, en el sector privado y de comunicaciones. Según Lohr (2000) el gasto anual en encuestas por muestreo en Estados Unidos representa de 2 a 5 billones de dólares. Este aumento del uso de las técnicas de muestreo en la investigación es claro porque es un procedimiento que cuesta mucho menos dinero, consume menos tiempo y puede incluso ser más preciso que al realizar una enumeración completa, también llamada censo. Una muestra bien seleccionada de unos cuantos miles de individuos puede representar con gran precisión una población de millones.

Es requisito fundamental de una buena muestra que las características de interés que existen en la población se reflejen en la muestra de la manera más cercana posible, para esto se necesitan definir los siguientes conceptos

- **Población objetivo:** es la colección completa de todas las unidades que se quieren estudiar.

- **Muestra:** es un subconjunto de la población.
- **Unidad de muestreo:** es el objeto a ser seleccionado en la muestra que permitirá el acceso a la unidad de observación.
- **Unidad de observación:** es el objeto sobre el que finalmente se realiza la medición.
- **Variable de interés:** es la característica propia de los individuos sobre la que se realiza la inferencia para resolver los objetivos de la investigación.

En la teoría de muestreo la variable de interés no se supone como una variable aleatoria sino como una cantidad fija o una característica propia de las unidades que componen la población.

### 1.1.1 Encuesta

Por **encuesta** se entiende una investigación estadística con las siguientes características:

1. El objetivo de una encuesta es proveer información acerca de la población finita y/o acerca de subpoblaciones de interés especial.
2. Asociado con cada elemento de la población existe una o más variables de interés. Una encuesta permite conseguir información sobre características poblacionales desconocidas llamadas parámetros. Éstas son funciones de los valores de las variables de interés y son desconocidos y requeridos.
3. El acceso y observación de los elementos de la población se establece mediante un algoritmo de muestreo, que es un mecanismo que asocia los elementos de la población con unidades de muestreo.
4. Una muestra de elementos se escoge. Esto puede ser hecho mediante la selección de las unidades de observación en el esquema. Una muestra es probabilística si se realiza mediante un mecanismo probabilístico y se conoce la probabilidad de selección de todas las posibles muestras.
5. Los elementos seleccionados en la muestra son observados y se realiza el proceso de medición; es decir para cada elemento de la muestra la variable de interés se mide y sus valores se graban.
6. Los valores grabados de las variables son usados para calcular estimaciones de los parámetros de interés.
7. Las estimaciones son finalmente publicadas. Estas sirven para la toma de decisiones.

#### Ciclo de vida de una encuesta

Groves, Fowler, Couper, Lepkowski, Singer & R. (2004) afirman que una encuesta va desde el diseño, pasando por la ejecución hasta, la entrega de las estimaciones. Si no se realiza un buen diseño no habrán buenas estimaciones. En este camino, el investigador debe transitar los siguientes pasos:

1. **Búsqueda de constructores:** los constructores son las ideas abstractas acerca de las cuales el investigador desea inferir. En una encuesta de victimización, se busca medir cuántos incidentes relacionados con crímenes tuvieron lugar en cierto periodo de tiempo; el investigador debe decidir acerca de ¿qué es un crimen?, ¿quién es una víctima?. En una encuesta de calidad de vida, se desea saber cuántas personas pobres hay en una determinada región; por tanto, es necesario decidir acerca de ¿qué es pobreza?

2. **Medición:** la cuestión clave para realizar una buena medición es diseñar preguntas que produzcan respuestas que reflejen perfectamente los constructores que se intentan medir. Por ejemplo, en la encuesta de victimización, se puede preguntar lo siguiente: «en los últimos seis meses ¿ha llamado usted a la policía para reportar algo que le haya sucedido y que usted considere que sea un crimen?». Por otro lado, en la encuesta de calidad de vida, un indicador de pobreza puede estar dado en términos del número de electrodomésticos que posee el hogar. Así, es posible preguntar lo siguiente: «¿cuántos televisores tiene en su hogar?» o también «¿cuántas bombillas eléctricas tiene su hogar?»
3. **Respuesta:** la naturaleza de las respuestas está determinada por la naturaleza de las preguntas. En algunas ocasiones la respuesta puede ser parte de la pregunta, siendo la tarea del respondiente escoger entre las categorías preguntadas; en otras ocasiones, el respondiente genera una respuesta concreta en sus propias palabras.
4. **Edición:** existen relaciones lógicas entre las preguntas de una encuesta. Por ejemplo, si el respondiente declara tener 12 años de edad y haber dado a luz a 5 hijos, debe existir un proceso de edición para este individuo. Este proceso intenta detectar datos atípicos y revisar la información para obtener la mejor medida del constructor buscado.
5. **Análisis y entrega de resultados:** el proceso estadístico arroja estimaciones que permiten la toma de decisiones y la resolución de los objetivos propuestos al comienzo de la investigación.

### 1.1.2 Marco de muestreo

Todo procedimiento de muestreo probabilístico requiere de un dispositivo que permita identificar, seleccionar y ubicar a todos y cada uno de los objetos pertenecientes a la población objetivo y que participarán en la selección aleatoria. Este dispositivo se conoce con el nombre de **marco de muestreo**. En investigaciones por muestreo se consideran dos tipos de objetos:

- **Elementos:** las unidades básicas e individuales sobre las que se realiza la medición.
- **Conglomerado:** agrupación de elementos cuya característica principal es que son homogéneos dentro de sí, y heterogéneos entre sí.

Cuando se dispone de un marco de elementos, se puede aplicar un diseño de muestreo de elementos; en muchas ocasiones se utilizan diseños de muestreo de conglomerados aunque se disponga de un marco de elementos. Si no se dispone de un marco de elementos (o es muy costoso construirlo) se debe recurrir a diseños de muestreo en conglomerados; es decir, que se utilizan marcos de conglomerados. Por ejemplo, al realizar una encuesta cuya unidad de observación sean las personas que viven en una ciudad, es muy difícil poder acceder a un marco de muestreo de las personas. Sin embargo, se puede tener acceso a la división sociodemográfica de la ciudad y así seleccionar algunos barrios de la ciudad, en una primera instancia y luego, seleccionar a las personas de los barrios en una segunda instancia. En el ejemplo anterior, los barrios son un ejemplo claro de conglomerados. Estas agrupaciones de elementos tienen la características de aparecer en el estado de la naturaleza. De esta forma, si se dispone de un marco de elementos, por ejemplo, el listado de empleados de una entidad, es posible aplicar un diseño de muestreo de elementos, realizar la selección aleatoria y de acuerdo a ese mismo diseño realizar las estimaciones necesarias. El lector debe recordar que los elementos son las entidades que componen la población y las unidades de muestreo son las entidades que conforman el marco muestral. Cuando no existe un marco de muestreo disponible es necesario construirlo. Existen dos tipos de marcos de muestreo, a saber:

- **De Lista:** listados físicos o magnéticos, ficheros, archivos de expedientes, historias clínicas que permiten identificar y ubicar a los objetos que participarán en el sorteo aleatorio.

- **De Área:** mapas de ciudades y regiones en formato físico o magnético, fotografías aéreas, imágenes de satélite o similares que permiten delimitar regiones o unidades geográficas en forma tal que su identificación y su ubicación sobre el terreno sea posible.

Es una virtud del marco si contiene **información auxiliar** que permite aplicar diseños muestrales y/o estimadores que conduzcan a estrategias más eficientes con respecto a la precisión de los resultados. O también si la información auxiliar<sup>1</sup> está organizada por órdenes deseables. Se llama información auxiliar **discreta**, si el marco de muestreo permite la desagregación de la población objetivo en categorías o grupos poblacionales más pequeños. Por ejemplo nivel socioeconómico, grupo industrial, etc. Se llama información auxiliar **continua** si existe una o varias características de interés de tipo continuo y positivas. Es deseable que la información auxiliar continua esté altamente relacionada con la característica de interés.

Por otra parte, un marco de muestreo es defectuoso si presenta alguno o varios de los siguientes casos:

- **Sobre-cobertura:** se presenta si en el dispositivo aparecen objetos que no pertenecen a la población objetivo. *No son todos los que están.*
- **Sub-cobertura:** se da cuando algunos elementos de la población objetivo no aparecen en el marco de muestreo o cuando no se ha actualizado la entrada de nuevos integrantes. *No están todos los que son.*
- **Duplicación:** La duplicación en un marco de muestreo se presenta si en el dispositivo aparecen varios registros para un mismo objeto. La razón más frecuente para la presencia de este defecto es la construcción no cuidadosa del marco a partir de la unión de registros administrativos de dos o más fuentes de información.

Estos defectos ocasionan errores en el cálculo de las expresiones que se utilizarán para generar las correspondientes estimaciones, generando sesgo, pérdida de precisión y, en algunos casos, que los resultados del estudio pierdan toda validez.

### Tipos de poblaciones objetivo

Groves, Fowler, Couper, Lepkowski, Singer & R. (2004) consideran que los tipos de poblaciones objetivo que se presentan de manera más frecuente en un estudio por muestreo son las siguientes

- **Hogares y personas:** el marco de muestreo más utilizado en estas poblaciones es de área. Como está basada en zonas geográficas, este tipo de marco requiere la vinculación de los hogares o personas a cada una de las áreas. Cuando se requiere seleccionar personas, este tipo de marcos hace necesarias muchas etapas de muestreo; de esta forma, se selecciona un subconjunto de zonas geográficas. Para cada zona seleccionada, se procede a seleccionar un subconjunto de secciones, luego de manzanas, luego de hogares y, finalmente, para cada hogar se seleccionan las personas; siendo éstas las unidades de observación.
- **Clientes, empleados o miembros de organizaciones:** por lo general, para la selección de miembros de organizaciones se manejan marcos de lista. Es importante que el estadístico esté al tanto de la frecuencia y manera de actualización de la lista pues pueden presentar los tres tipos de defectos vistos anteriormente.

---

<sup>1</sup> Toda información auxiliar disponible para todos y cada uno de los elementos del universo afecta directamente la estrategia empleada para obtener los objetivos de la investigación. Con respecto a la información auxiliar, es deseable que esté bien correlacionada con la característica de interés.

- **Organizaciones:** existen diversos tipos de organizaciones, como por ejemplo, iglesias, prisiones, empresas, hospitales, escuelas, etc. En encuestas a establecimientos comerciales, es frecuente tener acceso a marcos de lista que agrupan a negocios con gran dispersión entre sí. Así, se puede encontrar desde la tienda de barrio, cuyas ventas ascienden a 1000 dólares al mes, hasta un hipermercado que vende 500 millones de dólares al mes.
- **Eventos:** en algunas ocasiones, la población objetivo son eventos. Hay muchos tipos de eventos que clasifican para la realización de una encuesta; entre ellos están los matrimonios, nacimientos, fallecimientos, períodos de depresión, tránsito de un automóvil en un segmento de la vía. Los marcos de muestreo para los eventos, de manera frecuente, son marcos de personas. Así, una persona ya ha experimentado el evento o no. De hecho, puede haber experimentado varios eventos. Sin embargo, otro marco de muestreo para eventos puede estar dado en períodos de tiempo o espacio.
- **Poblaciones poco frecuentes:** cuando la incidencia es muy baja (por ejemplo las poblaciones de incidentes o con alguna enfermedad rara). Generalmente, la manera para acceder a este tipo de poblaciones es mediante un marco de muestreo que contenga a esta población como un subconjunto de elementos que pueden ser ubicados.

**Ejemplo 1.1.1.** Suponga que una entidad oficial del gobierno de su país está interesada en la realización de una encuesta de desempleo con el fin de determinar a) cuántas personas actualmente pertenecen a la fuerza laboral, tanto en el país en cuestión como en sus regiones o subdivisiones geográficas y b) qué proporción de éstas están desempleadas. Con base en lo anterior se tienen los siguientes aspectos para la realización de dicho estudio:

- *Población objetivo:* Todas las personas de Colombia.
- *Dominios o subgrupos de interés:* Grupos de edad, género, grupos ocupacionales y regiones del país.
- *Características de interés:* Pertenencia a la fuerza laboral y estado de empleo. Éstas toman valor uno o cero.
- *Parámetros de interés:* Número total de personas pertenecientes a la fuerza laboral, número total de desempleados, proporción de desempleo.
- *Muestra:* Se selecciona una muestra de la población con la ayuda de mecanismos de identificación y ubicación de las personas en el país.
- *Observaciones:* Cada persona incluida en la muestra es visitada por un encuestador entrenado, quien hará preguntas siguiendo un cuestionario estandarizado y recolectará las respuestas en un instrumento apropiado.
- *Procesamiento:* Los datos se editan y se preparan para la etapa de estimación.
- *Estimación:* Se calculan las estimaciones sobre los parámetros de interés y también indicadores acerca de la incertidumbre de estas estimaciones.

### 1.1.3 Sesgo

En el diseño y puesta en marcha de una encuesta puede ocurrir cierto tipo de situaciones que pueden sesgar las estimaciones finales. Este tipo de sesgos puede ocurrir antes, durante y después de la recolección de los datos. Es tarea del estadístico advertir ante todas las posibles instancias de los problemas que causan los sesgos y procurar que, en todas las etapas de la encuesta, se minimice el error humano y el error estadístico para que al final los resultados del estudio sean tan confiables como sea posible.

### Sesgo de selección

Este tipo de sesgo ocurre cuando parte de la población objetivo no está en el marco de muestreo. Una muestra a conveniencia<sup>2</sup> es sesgada pues las unidades más fáciles de elegir o las que más probablemente respondan a la encuesta no son representativas de las unidades más difíciles de elegir. (Lohr 2000) afirma que se presenta este tipo de sesgo si:

1. La selección de la muestra depende de cierta característica asociada a las propiedades de interés.  
Por ejemplo: Frecuencia con que los adolescentes hablan con los padres acerca del SIDA.
2. La muestra se realiza mediante elección deliberada o mediante un juicio subjetivo. Por ejemplo, si el parámetro de interés es la cantidad promedio de gastos en compras en un centro comercial y el encuestador elige a las personas que salen con muchos paquetes, entonces la información estaría sesgada puesto que no está reflejando el comportamiento promedio de las compras.
3. Existen errores en la especificación de la población objetivo. Por ejemplo, en encuestas electorales, cuando la población objetivo contiene a personas que no están registradas como votantes ante la organización electoral de su país.
4. Existe sustitución deliberada de unidades no disponibles en la muestra. Si, por alguna razón, no fue posible obtener la medición y consecuente observación de la característica de interés para algún individuo en la población, la sustitución de este elemento debe hacerse bajo estrictos procedimientos estadísticos y no debe ser subjetiva en ningún modo.
5. Existe ausencia de respuesta. Este fenómeno puede causar distorsión de los resultados cuando los que no responden a la encuesta difieren críticamente de los que sí respondieron.
6. La muestra está compuesta por respondientes voluntarios. Los foros radiales, las encuestas de televisión y los estudios de portales de internet no proporcionan información confiable.

### Sesgo de medición

Este tipo de sesgo ocurre cuando el instrumento con el que se realiza la medición tiene una tendencia a diferir del valor verdadero que se desea averiguar. Éste sesgo debe ser considerado y minimizado en la etapa de diseño de la encuesta. Nótese que ningún análisis estadístico puede revelar que una pesa añadió a cada persona 2Kg de más en un estudio de salud. (Lohr 2000) cita algunas situaciones en donde se presenta este sesgo de medición:

1. Cuando el respondiente miente. Esta situación se presenta a menudo en encuestas que pregunta acerca del ingreso salarial, alcoholismo y drogadicción, nivel socioeconómico e incluso edad.
2. Difícil comprensión de las preguntas. Por ejemplo: ¿No cree que no este es un buen momento para invertir? La doble negación en la pregunta es muy confusa para el respondiente.
3. Las personas tienden a olvidar. Es bien sabido que las malas experiencias suelen ser olvidadas; esta situación debe acotarse si se está trabajando en una encuesta de criminalidad.
4. Distintas respuestas a distintos entrevistadores. En algunas regiones es muy probable que la raza, edad o género del encuestador afecte directamente la respuesta del entrevistado.

---

<sup>2</sup>A pesar de que las muestras por conveniencia o por juicio no pueden ser utilizadas para estimar parámetros de la población, éstas sí pueden proporcionar información valiosa en las primeras etapas de una investigación o cuando no es necesario generalizar los resultados a la población.

5. Leer mal las preguntas o polemizar con el respondiente. El encuestador puede influir notablemente en las respuestas. Por lo anterior, es muy importante que el proceso de entrenamiento del entrevistador sea riguroso y completo.
6. La muestra está compuesta por respondientes voluntarios. Los foros radiales, las encuestas de televisión y los estudios de portales de internet no proporcionan información confiable.

## 1.2 Marco y Lucy

Este libro toma como base de aplicación una investigación gubernamental que quiere responder al objetivo de *medir el crecimiento económico en el sector industrial*.

Suponga que para completar el objetivo se ha propuesto desarrollar una encuesta a las empresas que hacen parte del sector industrial, para conocer el comportamiento del sector en términos de **constructores** financieros, sociales y fiscales. Una vez termine el proceso de medición, se pueden calcular estimaciones y construir indicadores que permitan inferir acerca del crecimiento del sector en el periodo de interés.

La **población objetivo** la conforman todas las empresas cuya actividad principal esté ligada al sector industrial. El proceso de medición se hará con base en las **características de interés**; a saber: ingresos en el último año fiscal, impuestos declarados en el último año fiscal y número de empleados. Adicionalmente, se requiere conocer si la empresa envía periódicamente algún tipo de material publicitario por correo electrónico porque se sospecha que las empresas obtienen más ingresos cuando utilizan esta estrategia publicitaria, lo cual es favorable para el gobierno porque aumenta la contribución impositiva y aumenta la creación de empleos.

Para obtener las respuestas, un entrevistador visitará las instalaciones físicas de la empresa y realizará las siguientes preguntas:

1. En el último año fiscal, ¿a cuánto ascendieron los ingresos en esta empresa?
2. En el último año fiscal, ¿a cuánto ascendieron los impuestos declarados por esta empresa?
3. Actualmente, ¿cuántos empleados laboran para esta empresa?
4. ¿Esta empresa acostumbra a enviar periódicamente material publicitario por correo electrónico a sus clientes o potenciales clientes?

Se sabe que el tamaño de la población es de 2396 empresas. Dependiendo de la estrategia de muestreo que se vaya a utilizar y de la calidad del marco de muestreo, las unidades de muestreo pueden ser las mismas empresas.

Para abordar la selección de una muestra que permita la inferencia acerca del crecimiento económico del sector, se dispone de un marco de muestreo con las siguientes características para cada empresa que conforma la población.

1. **Identificador:** es una secuencia alfanumérica de dos letras y tres dígitos. Este número de identificación se le otorga a cada empresa en el momento de la constitución legal ante la entidad de registro pertinente.
2. **Ubicación:** es la dirección que se encuentra registrada en la declaración de impuestos.
3. **Zona:** la ciudad está conformada por barrios o zonas geográficas. Dependiendo de la dirección, la empresa pertenece a una y sólo una zona geográfica de la ciudad.

4. **Nivel:** según los registros tributarios, las empresas se catalogan en tres grupos:
- Grandes: empresas que tributan 49 millones de dólares al año o más.
  - Medianas: empresas que tributan más de 11 millones y menos de 49 millones de dólares al año.
  - Pequeñas: empresas que tributan 11 millones de dólares al año o menos.

Nótese que una empresa sólo puede pertenecer a un sólo un nivel industrial.

### Visualización en R

El paquete **TeachingSampling** de R incluye dos archivos de datos. El marco de muestreo llamado **Marco** del cual se extraerá una muestra aleatoria de empresas que deben ser entrevistadas y que contiene la identificación, ubicación, zona y nivel de cada una de las empresas del sector industrial. Por otro lado, incorpora el conjunto de datos llamado **BigLucy** en donde, se encuentran los valores de las características de interés para todos los elementos de la población.

Para tener acceso a los dos conjuntos de datos es necesario cargar el paquete en el entorno de R. El paquete **TeachingSampling** puede ser cargado fácilmente mediante el uso de la siguiente instrucción:

```
library(TeachingSampling)
```

Una vez cargado el paquete **TeachingSampling**, la visualización del marco de muestreo, se realiza de la siguiente forma:

```
data(BigLucy)
BigLucy[1:10,c(1:4,11)]

##           ID      Ubication Level   Zone Segments
## 1 AB0000000001 C0212063K0089834 Small County1 County1 1
## 2 AB0000000002 C0011268K0290629 Small County1 County1 1
## 3 AB0000000003 C0077703K0224194 Small County1 County1 1
## 4 AB0000000004 C0091012K0210885 Small County1 County1 1
## 5 AB0000000005 C0301070K0000827 Small County1 County1 1
## 6 AB0000000006 C0255289K0046608 Small County1 County1 1
## 7 AB0000000007 C0280547K0021350 Small County1 County1 1
## 8 AB0000000008 C0148379K0153518 Small County1 County1 1
## 9 AB0000000009 C0111156K0190741 Small County1 County1 1
## 10 AB0000000010 C0199974K0101923 Small County1 County1 1
```

La instrucción **BigLucy[1:10,c(1:4,11)]** se utiliza para mostrar las diez primeras empresas del marco de muestreo. Si se quiere visualizar todo el conjunto de datos, la instrucción **BigLucy** mostrará la totalidad del marco de muestreo. La función **names** muestra cada uno de los objetos que componen el archivo de datos, mientras que la función **dim** muestra las dimensiones del conjunto de datos.

```
names(BigLucy)

## [1] "ID"          "Ubication"    "Level"        "Zone"         "Income"
## [6] "Employees"   "Taxes"        "SPAM"         "ISO"          "Years"
## [11] "Segments"
```

```
dim(BigLucy)
```

```
## [1] 85296    11
```

La lectura del archivo de datos se hace de la siguiente manera: tomando como referencia la fila número 3 (la tercera empresa del conjunto de datos), es una empresa cuyo número de identificación es AB0000000001, ubicada en la dirección C0212063K0089834, de nivel industrial Small, localizada en la zona County1 y en el segmento County1 1. Esta empresa registró en el último año fiscal un ingreso neto de 281 millones de dólares y realizó un tributo de 3 millones de dólares, actualmente da empleo a 41 empleados, no envía periódicamente publicidad a sus clientes o potenciales clientes mediante correo electrónico, tampoco tiene certificación de calidad ISO y tiene una antigüedad de 14 años.

```
BigLucy[1:10,5:10]
```

	Income	Employees	Taxes	SPAM	ISO	Years
## 1	281	41	3.0	no	no	14.0
## 2	329	19	4.0	yes	no	17.6
## 3	405	68	7.0	no	no	13.6
## 4	360	89	5.0	no	no	44.7
## 5	391	91	7.0	yes	no	23.3
## 6	296	89	3.0	no	no	48.3
## 7	490	22	10.5	yes	yes	17.0
## 8	473	57	10.0	yes	no	7.5
## 9	350	84	5.0	yes	no	38.7
## 10	361	25	5.0	no	no	18.3

Nótese que el conjunto de datos poblacionales BigLucy contiene el valor de las características de interés para cada empresa. Hasta este momento no se ha seleccionado ninguna muestra, pero si se supone hipotéticamente que la muestra seleccionada hubiese sido las diez primeras empresas del marco de muestreo, la base de datos, después de la medición se vería como lo muestra la salida anterior y con estos datos se procede a realizar las estimaciones requeridas para el cumplimiento de los objetivos de la investigación.

Las estadísticas concernientes a las variables en la población se visualizan fácilmente con la función `summary` aplicada al conjunto de datos Lucy.

```
summary(BigLucy[,5:10])
```

	Income	Employees	Taxes	SPAM	ISO
## Min.	: 1	Min. : 1.0	Min. : 0.5	no :33355	no :56896
## 1st Qu.	: 230	1st Qu.: 38.0	1st Qu.: 2.0	yes:51941	yes:28400
## Median	: 388	Median : 62.0	Median : 6.0		
## Mean	: 430	Mean : 63.2	Mean : 11.8		
## 3rd Qu.	: 570	3rd Qu.: 84.0	3rd Qu.: 15.0		
## Max.	: 2510	Max. :263.0	Max. :305.0		
## Years					
## Min.	: 1.0				
## 1st Qu.	:13.1				
## Median	:25.4				
## Mean	:25.4				
## 3rd Qu.	:37.7				
## Max.	:50.0				

Por medio de la función `total`, tenemos acceso al total de las tres características de interés.

```
attach(BigLucy)
total <- function(x){length(x)*mean(x)}

total(Income)

## [1] 36634733

total(Employees)

## [1] 5391992

total(Taxes)

## [1] 1008426
```

El sector industrial tiene altos ingresos que ascienden a 36634733 millones de dólares, aporta al gobierno 1008426 millones de dólares en tarifas impositivas, emplea un total de 5391992 personas. La función `tapply` permite aplicar la función `total` y la función `mean` para calcular el total y el promedio, respectivamente, de las variables de interés en cada categoría de la variable `Level`. La función `table` hace un recuento del total de casos para una o más variables categóricas.

```
tapply(Income,Level,total)

##      Big   Medium   Small
## 3629710 17057285 15947738

table(SPAM,Level)

##      Level
## SPAM   Big Medium Small
##   no    910 10185 22260
##   yes   1995 15610 34336
```

Nótese que la mayoría del ingreso del sector industrial es adquirido por las empresas medianas y pequeñas. Sin embargo, en promedio las empresas grandes doblan el ingreso de las medianas que a su vez es tres veces el ingreso de las empresas pequeñas. En términos absolutos, la estrategia publicitaria de enviar SPAM a los clientes o potenciales clientes se implementa con mayor frecuencia en las empresas pequeñas.

La función `xtabs` permite realizar una tabulación cruzada entre las variables categóricas `Level` y `SPAM` de la base de datos. Los datos de las celdas indican el total de la variable `Income`. Nótese que el ingreso de las empresas que utilizan el SPAM como estrategia de publicidad dobla el ingreso de las empresas que no utilizan SPAM en casi todos los niveles industriales.

```
xtabs(Income~Level+SPAM)

##          SPAM
## Level      no      yes
```

```
##   Big    1116990  2512720
## Medium 6679820 10377465
## Small  6288497  9659241
```

La función `boxplot` permite realizar el diagrama de cajas de cada una de las variables de interés. Nótese que, a excepción de la variable `Years`, existe una dependencia marcada en el comportamiento de las características cuantitativas con el nivel industrial.

Sin embargo, a diferencia del caso anterior, no parece existir una dependencia en el comportamiento de las características cuantitativas con el hábito de enviar publicidad por internet.

Las figuras 1.1 y 1.2 muestran la dispersión y locación de las características de interés por cada nivel industrial. En general, las empresas grandes tienen ingresos más altos, aportan una carga impositiva más alta y emplean a más personas que las empresas medianas y pequeñas. Es deseable que el marco de muestreo contenga la pertenencia al nivel industrial de cada empresa en la población porque es un buen discriminante y permite la implementación de estrategias de muestreo adecuadas que guíen a estimaciones más precisas. La función `n_barplot` muestra un diagrama de barras del total de la variable `Level`.

La figura 1.3 muestra que la distribución de las características de interés no es simétrica y es sesgada a la izquierda. Estos rasgos particulares se deben tener en cuenta al momento de escoger la mejor estrategia de muestreo. La función `hist` permite la creación de los histogramas y la función `pie` permite la creación de un gráfico de torta.

La correlación lineal entre las características de interés es alta; entre `Income` y `Taxes` existe una correlación de 0.91, esto se puede explicar porque las empresas tributan una mayor cantidad de dinero si han obtenido mayores ingresos y viceversa. Se utiliza la función `cor` para obtener la matriz de correlación entre las características de interés.

```
Datos <- data.frame(Income, Employees, Taxes, Years)
cor(Datos)

##           Income Employees      Taxes      Years
## Income    1.0000000  0.643304  0.9166732 -0.0001266
## Employees  0.6433037  1.000000  0.6448609  0.0039724
## Taxes     0.9166732  0.644861  1.0000000  0.0008152
## Years    -0.0001266  0.003972  0.0008152  1.0000000
```

Para visualizar la relación entre las variables de interés, se utiliza la función `pairs` para obtener los diagramas de dispersión para cada par de variables justo como lo muestra la figura 1.4.

La tabla 1.1. resume los parámetros de interés que, mediante una adecuada estrategia de muestreo, se deben estimar para resolver el objetivo principal de la investigación. Si se desean estimaciones discriminadas por nivel industrial, entonces la tabla 1.2. da cuenta del valor de estos parámetros dentro de los subgrupos poblacionales.

Consecuentemente, si se quieren estimaciones discriminadas por comportamiento publicitario, entonces la tabla 1.3. muestra el valor de cada uno de estos parámetros. Por último, si se buscan estimaciones discriminadas tanto por comportamiento publicitario cruzado con nivel industrial, entonces se cuenta con la tabla 1.4. que resume dicha información.

```
par(mfrow=c(2,2))
boxplot(Income~Level)
boxplot(Employees~Level)
boxplot(Taxes~Level)
boxplot(Years~Level)
```

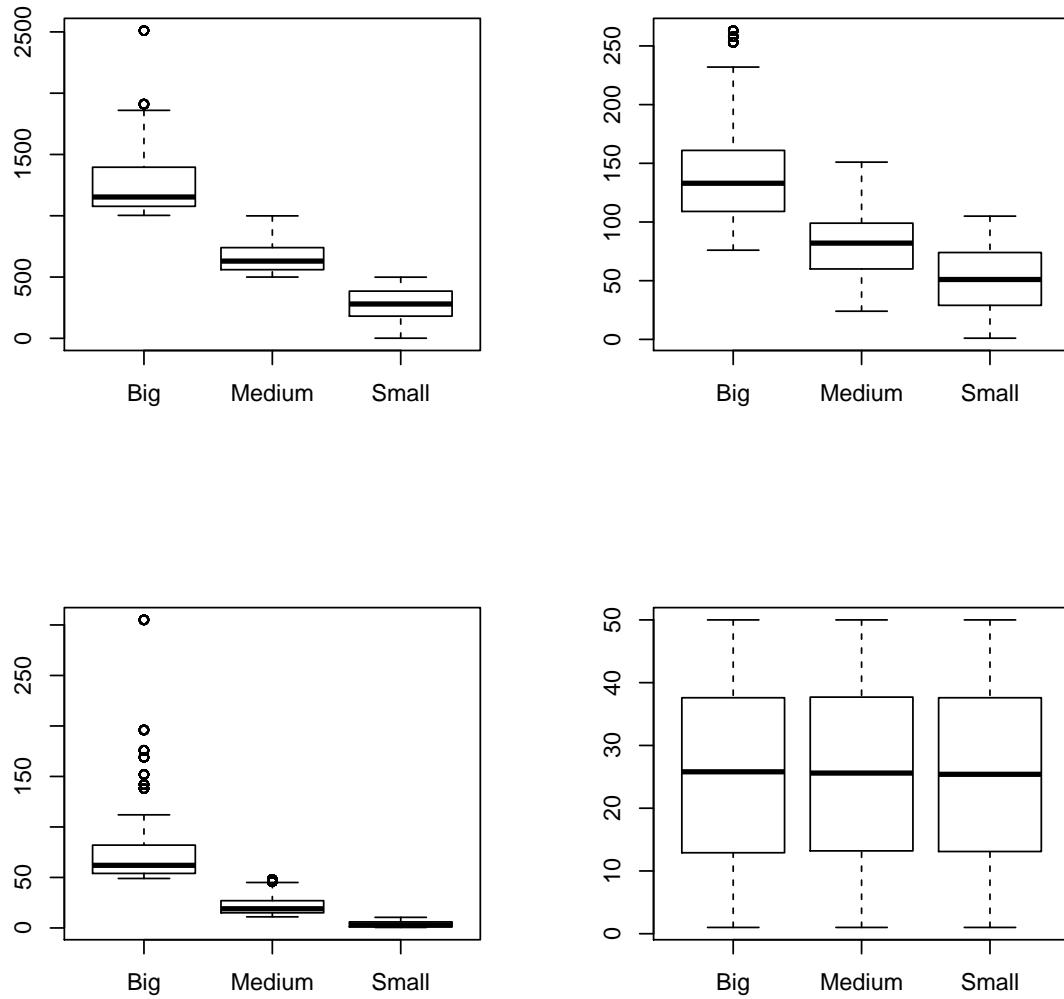


Figura 1.1: Boxplot de las características de interés en cada nivel industrial.

Cuadro 1.1: Parámetros de la población.

	Ingreso	Impuestos	Empleados
N total	2.396	2.396	2.396
Suma	1.035.217	28.654	151.950
Media	432	12	63

```
par(mfrow=c(2,2))
boxplot(Income~SPAM)
boxplot(Employees~SPAM)
boxplot(Taxes~SPAM)
boxplot(Years~SPAM)
```

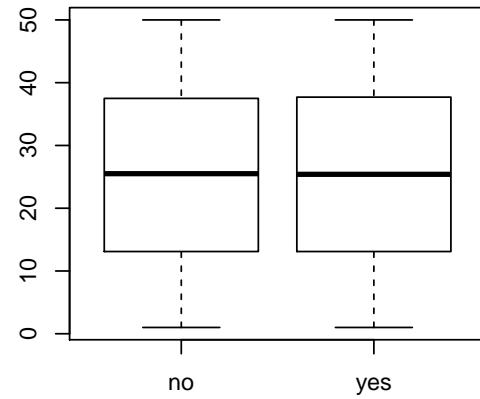
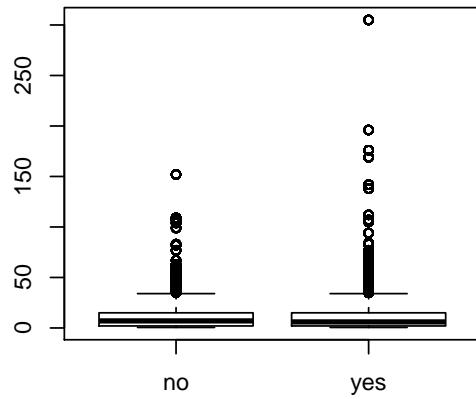
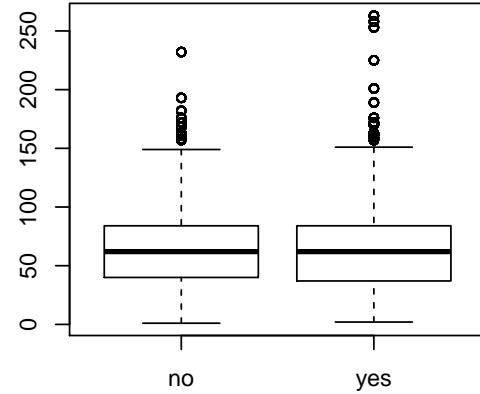
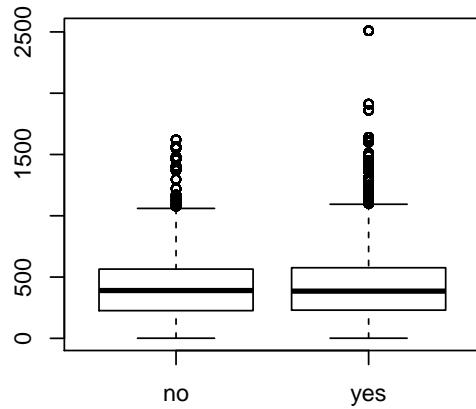


Figura 1.2: Boxplot de las características de interés en cada nivel industrial.

```
par(mfrow=c(2,2))
hist(Income)
hist(Employees)
hist(Taxes)
hist(Years)
```

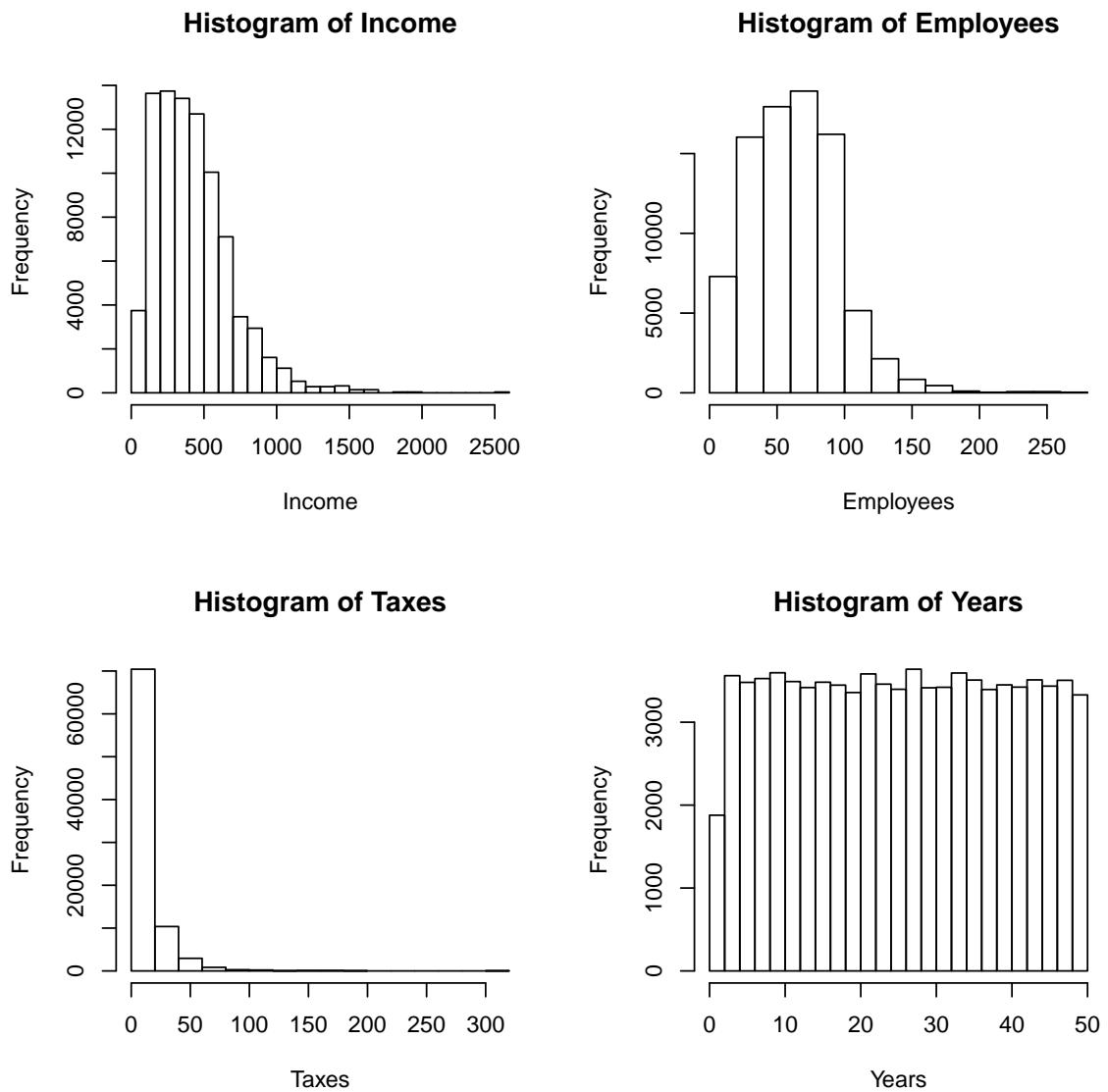


Figura 1.3: *Histograma de las características de interés.*

Cuadro 1.2: *Parámetros de la población discriminados por nivel industrial.*

			Ingreso	Impuestos	Empleados
Nivel	Grande	N total	83	83	83
Mediano		Suma	103.706	6.251	11.461
		Media	1.249	75	138
		N total	737	737	737
Pequeño		Suma	487.351	16.293	59.643
		Media	661	22	81
		N total	1.576	1.576	1.576
		Suma	444.160	6.110	80.846
		Media	282	4	51

Cuadro 1.3: *Parámetros de la población discriminados por comportamiento publicitario.*

			Ingreso	Impuestos	Empleados
SPAM	no	N total	937	937	937
si		Suma	397.952	10.593	59.600
		Media	425	11	64
		N total	1.459	1.459	1.459
		Suma	637.265	18.061	92.350
		Media	437	12	63

Cuadro 1.4: *Parámetros de la población discriminados por nivel industrial y por comportamiento publicitario.*

		SPAM					
		no			sí		
		N total	Suma	Media	N total	Suma	Media
Grande	Ingreso	26	31.914	1.227	57	71.792	1.260
	Impuestos	26	1.844	71	57	4.4.07	77
	Empleados	26	3.587	138	57	7.874	138
Mediano	Ingreso	291	190.852	656	446	296.499	665
	Impuestos	291	6.322	22	446	9.971	22
	Empleados	291	23.745	82	446	35.898	80
Pequeño	Ingreso	620	175.186	283	956	268.974	281
	Impuestos	620	2.427	4	956	3.683	4
	Empleados	620	32.268	52	956	48.578	51



## Capítulo 2

# Muestras probabilísticas y estimadores

La base matemática para el desarrollo del modelo de muestreo se encuentra en la teoría de la inferencia estadística y de manera más directa en la aplicación de los principios básicos de la teoría de probabilidad. Los resultados del modelo de muestreo sólo son válidos si se parte de la certeza de contar con una muestra que satisfaga las condiciones exigidas por la inferencia estadística.

Bautista (1998)

### 2.1 Población y muestra aleatoria

El proceso de estimación e inferencia en poblaciones finitas, que finalmente son las que fácilmente encontramos en la realidad y en las que se enfoca el muestreo, es muy diferente al proceso de inferencia de la estadística clásica. Esta última se trata a los valores observados como realizaciones de una variable aleatoria. En contravía con lo anterior, el muestreo asume que los valores observados corresponden a parámetros fijos poblacionales. Partiendo de este hecho formalicemos algunos conceptos que son de vital importancia en el estudio y análisis del muestreo.

#### 2.1.1 Población finita

**Definición 2.1.1.** Una **población finita** es un conjunto de  $N$  elementos  $\{e_1, e_2, \dots, e_N\}$ . Cada unidad puede ser identificada sin ambigüedad por un conjunto de rótulos. Sea  $U = \{1, 2, \dots, N\}$  el conjunto de rótulos de la población finita. El tamaño de la población no es necesariamente conocido.

Es el conjunto de  $N$ , donde  $N < \infty$ , unidades que conforman el universo de estudio.  $N$  es comúnmente llamado el tamaño poblacional. Cada elemento perteneciente a la población puede ser identificado por un rótulo. Sea  $U$  el conjunto de rótulos, tal que

$$U = \{1, \dots, k, \dots, N\}.$$

Se utilizará el subíndice  $k$  para denotar la existencia física del  $k$ -ésimo elemento. Nótese que el **tamaño de la población**,  $N$ , no siempre es conocido y en algunas ocasiones el objetivo de la investigación es poder estimarlo.

### 2.1.2 Muestra aleatoria

Es un subconjunto de la población que ha sido extraído mediante un mecanismo estadístico de selección. Notaremos con una letra mayúscula  $S$  a la muestra aleatoria<sup>1</sup> y con una letra minúscula  $s$  a una realización de la misma. De tal forma que, sin ambigüedad, una muestra seleccionada (realizada) es el conjunto de unidades pertenecientes a

$$s = \{1, \dots, k, \dots, n(S)\}.$$

El número de componentes de  $s$  es llamado el **tamaño de muestra** y no siempre es fijo. Es decir, en algunos casos  $n(S)$  es una cantidad aleatoria. El conjunto de todas las posibles muestras se conoce como **soporte**. Haciendo una analogía con la inferencia estadística clásica, el soporte generado por una muestra aleatoria corresponde al espacio muestral generado por una variable aleatoria.

La anterior definición de muestra, en donde los elementos incluidos se listan dentro de un conjunto, corresponde a la forma clásica de notación. Sin embargo, una muestra también puede ser notada como un vector de tamaño  $N$ . De esta manera, la  $k$ -ésima entrada del vector denotará el número de veces que el elemento fue incluido o seleccionado; si el valor es cero, indica que el elemento no fue incluido en la muestra seleccionada; si el valor es distinto de cero, indica que el elemento sí fue seleccionado. Aunque ambas formas de notación tienen la misma interpretación, para evitar confusiones, se denotará la muestra en forma de vector con una  $\mathbf{s}$  en negrilla, mientras que la muestra en forma de conjunto se denotara con una  $s$  simple sin negrilla. A continuación se dan definiciones más precisas acerca de la muestra aleatoria con o sin reemplazo.

#### Muestra aleatoria sin reemplazo

**Definición 2.1.2.** Una **muestra sin reemplazo** se denota mediante un vector columna

$$\mathbf{s} = (I_1, I_2, \dots, I_N)' \in \{0, 1\}^N \quad (2.1.1)$$

donde

$$I_k = \begin{cases} 1 & \text{si el } k\text{-ésimo elemento pertenece a la muestra,} \\ 0 & \text{en otro caso} \end{cases} \quad (2.1.2)$$

Una muestra aleatoria se dice sin reemplazo si la inclusión de cada uno de los elementos se hace entre los elementos que no han sido escogidos aún; de esta manera el conjunto  $s$  nunca tendrá elementos repetidos. El tamaño de muestra corresponde a la cardinalidad de  $s$ .

$$n(S) = \sum_{k \in U} I_k. \quad (2.1.3)$$

Como  $n(S)$  no es una cantidad fija, es posible que ocurran uno de los siguientes escenarios: a) que la muestra no contenga a ningún elemento, entonces esta muestra se dice vacía; b) que la muestra contenga a todos los elementos de la población, esta muestra se conoce con el nombre de **censo**.

#### Muestra aleatoria con reemplazo

**Definición 2.1.3.** Una **muestra con reemplazo** se denota mediante un vector columna

$$\mathbf{s} = (n_1, n_2, \dots, n_N)' \in \mathbb{N}^N \quad (2.1.4)$$

donde  $n_k$  es el número de veces que el elemento  $k$  está en la muestra

---

<sup>1</sup>Nótese que  $S$  es una variable aleatoria.

En algunos casos, por conveniencia del mecanismo de selección, el usuario prefiere tomar una muestra aleatoria con reemplazo si la inclusión de cada uno de los elementos tiene en cuenta a todos los elementos, ya sea que hayan sido escogidos para pertenecer en la muestra o no. De esta forma, el usuario puede seleccionar una muestra cuyo proceso de selección incluya a un individuo  $m$  veces (nótese que  $m$  puede ser mayor que  $N$ ). Sin embargo, en una muestra aleatoria con reemplazo, dos o más componentes pueden ser idénticos. Un elemento que esté incluido más de una vez en  $s$  es llamado **elemento repetido**.

En principio el tamaño de muestra está dado por

$$n(S) = m = \sum_{k \in U} n_k. \quad (2.1.5)$$

El número de elementos distintos en una muestra aleatoria  $S$  con reemplazo es llamado **tamaño de muestra efectivo** y con probabilidad uno es menor o igual a  $N$ .

### 2.1.3 Soportes de muestreo

En los próximos capítulos empezará el tratamiento particular para estrategias de muestreo específicas; es decir, diseños de muestreo que se ajustan a ciertas situaciones y estimadores que mejoran la eficiencia de la estrategia. Sin embargo, antes de proseguir, es necesario que el lector entienda que las estrategias de muestreo se definen en términos del tipo de muestreo que se utiliza para la selección de muestras. En general, existen dos distinciones básicas.

1. **Tipo de muestreo:** selección de unidades con reemplazo o sin reemplazo.
2. **Tamaño de muestra:** tamaño de muestra fijo o aleatorio.

Como se verá en los capítulos posteriores, dependiendo de las anteriores condiciones, se define la estrategia de muestreo, el tratamiento teórico para la estimación de parámetros y el tipo de soporte. Esta sección trata específicamente sobre las diferentes formas que puede tomar el soporte de un diseño de muestreo dependiendo de las dos distinciones básicas. Para entrar en materia, es necesario enunciar las siguientes definiciones.

**Definición 2.1.4.** Un **soporte**  $Q$  es un conjunto de muestras.

**Definición 2.1.5.** Un **soporte** se llama **simétrico** si para cualquier  $s \in Q$ , todas las permutaciones de  $s$  están también en  $Q$ .

En los siguientes capítulos, a menos que se mencione lo contrario, el término **soporte** hará referencia a un **soporte simétrico**. Algunos soportes simétricos particulares son:

- El *soporte simétrico sin reemplazo* definido como

$$\mathcal{S} = \{0, 1\}^N$$

Nótese que

$$\#(\mathcal{S}) = 2^N$$

Por ejemplo, si  $N = 3$ , entonces  $\mathcal{S}$  queda definido por las siguientes muestras:

$$\mathcal{S} = \{(0, 0, 0)', (1, 0, 0)', (0, 0, 1)', (1, 0, 1)', (0, 1, 0)', (1, 1, 0)', (0, 1, 1)', (1, 1, 1)'\}$$

- El *soporte simétrico sin reemplazo de tamaño fijo* definido como

$$\mathcal{S}_n = \left\{ \mathbf{s} \in \mathcal{S} \mid \sum_{k \in U} s_k = n \right\}$$

Nótese que

$$\#(\mathcal{S}_n) = \binom{N}{n}$$

Por ejemplo, si  $N = 3$  y  $n = 2$ , entonces  $\mathcal{S}_n$  queda definido por las siguientes muestras:

$$\mathcal{S}_n = \{(1, 0, 1)', (1, 1, 0)', (0, 1, 1)'\}$$

- El *soporte simétrico con reemplazo* definido como

$$\mathcal{R} = \mathbb{N}^N$$

donde  $\mathbb{N}$  es el conjunto de los números naturales. Nótese que este soporte es un conjunto contable pero infinito, por tanto

$$\#(\mathcal{R}) = \infty$$

- El *soporte simétrico con reemplazo de tamaño fijo* definido como

$$\mathcal{R}_m = \left\{ \mathbf{s} \in \mathcal{R} \mid \sum_{k \in U} n_k = m \right\}$$

Nótese que

$$\#(\mathcal{R}_m) = \binom{N+m-1}{m}$$

Por ejemplo, si  $N = 3$  y  $m = 2$ , entonces  $\mathcal{R}_m$  queda definido por las siguientes muestras:

$$\mathcal{R}_m = \{(2, 0, 0)', (0, 0, 2)', (0, 2, 0)', (1, 1, 0)', (1, 0, 1)', (0, 1, 1)'\}$$

Tillé (2006) afirma que geométricamente cada vector  $\mathbf{s}$  representa el vértice de un  $N$ -cubo. Además, se tiene el siguiente resultado:

**Resultado 2.1.1.** *Para los soportes definidos anteriormente, se tienen las siguientes propiedades:*

1.  $\mathcal{S}, \mathcal{S}_n, \mathcal{R}, \mathcal{R}_m$  son soportes simétricos.
2.  $\mathcal{S} \subset \mathcal{R}$ .
3. El conjunto  $\{\mathcal{S}_0, \mathcal{S}_1, \dots, \mathcal{S}_N\}$  es una partición de  $\mathcal{S}$ .
4. El conjunto  $\{\mathcal{R}_0, \mathcal{R}_1, \dots, \mathcal{S}_N, \dots\}$  es una partición infinita de  $\mathcal{R}$ .
5.  $\mathcal{S} \subset \mathcal{R}$  para todo  $n = 0, 1, \dots, N$ .

### Muestras probabilísticas

No todas las muestras aleatorias son de tipo probabilístico. Una muestra (con o sin reemplazo) es de tipo probabilístico si:

- Es posible construir (o al menos definir teóricamente) un soporte  $Q$ , tal que  $Q = \{s_1, \dots, s_q, \dots, s_Q\}$ , de todas las muestras posibles obtenidas por un método de selección. En donde  $s_q$ ,  $q = 1, \dots, Q$ , es una muestra perteneciente al soporte  $Q$ .
- Las probabilidades de selección que el proceso aleatorio le otorga a cada posible muestra perteneciente al soporte son conocidas de antemano a la selección de la muestra final.

Nótese que una muestra al azar no necesariamente es una muestra probabilística. En la mala práctica, algunos investigadores utilizan métodos aleatorios de inclusión de elementos sin disponer de un marco de muestreo y sin cumplir las dos condiciones anteriores; de esta manera, aunque los elementos sean escogidos de manera aleatoria o al azar, la muestra resultante no se puede catalogar como una muestra probabilística. Desde aquí en adelante, a menos que se diga lo contrario, el término muestra se refiere a una muestra probabilística. Algunos comentarios de interés son:

1. El universo  $U$  es finito.
2. La muestra probabilística  $s$  puede contener objetos repetidos. Esto sucede cuando el procedimiento de muestreo es con reemplazo.
3. La muestra  $s$  con repeticiones, puede tener un tamaño mayor al de la población.
4. La muestra  $s$  sin repeticiones, puede tener un tamaño máximo igual a  $N$ .
5. Si se presenta la ausencia del marco de muestreo es imposible realizar un procedimiento de muestreo probabilístico. Excepto cuando se realiza un censo.
6. Si la muestra seleccionada no es de tipo probabilístico, entonces no se puede construir ninguna estimación de tipo estadístico.
7. El estadístico deberá responder por los engaños o fraudes, que por ignorancia, mala fe o por la comodidad de mantener un empleo o negocio, para el cual no está capacitado, cometa contra clientes, ciudades y países que confían en la cifras resultantes de sus análisis.

**Ejemplo 2.1.1.** Suponga una población finita de tamaño  $N = 5$ , en donde los integrantes de la población están identificados cada uno con su nombre. La población la conforman los siguientes elementos:

**Yves, Ken, Erik, Sharon, y Leslie,**

En R se utiliza un vector de cadena de texto para indexar la población. Nótese que los elementos pertenecientes al vector son especificados mediante el uso de las comillas. En este caso los identificadores de cada elemento de la población, son asignados al objeto U.

```
U <- c("Yves", "Ken", "Erik", "Sharon", "Leslie")
U[1]
## [1] "Yves"

U[2]
## [1] "Ken"
```

Para obtener el soporte  $Q$ , de todas las posibles muestras de tamaño  $n = 2$  de esta población de tamaño  $N = 5$ , se utiliza la función `Support` del paquete `TeachingSampling`. Esta función contiene tres argumentos: el tamaño de la población  $N$ , el tamaño fijo de cada una de las posibles muestras  $n$  y, por último, una característica  $y$  que puede ser de tipo numérico o puede ser un conjunto de rótulos, la salida de la función será un conjunto de datos contenido todas las posibles muestras de tamaño fijo. Cuando el argumento  $y$  es distinto de `FALSE`, el resultado de la función será la característica poblacional para cada individuo. En el siguiente ejemplo se utiliza la función `Support(N,n,y=FALSE)` para obtener el conjunto de posibles muestras de tamaño dos de la población  $U$ , mientras que la función `Support(N,n,U)` arroja el conjunto de los rótulos en cada una de las 10 posibles muestras.

```

N <- length(U)
N

## [1] 5

n <- 2

Support(N,n)

##      [,1] [,2]
## [1,]    1    2
## [2,]    1    3
## [3,]    1    4
## [4,]    1    5
## [5,]    2    3
## [6,]    2    4
## [7,]    2    5
## [8,]    3    4
## [9,]    3    5
## [10,]   4    5

Support(N,n,U)

##      [,1]      [,2]
## [1,] "Yves"   "Ken"
## [2,] "Yves"   "Erik"
## [3,] "Yves"   "Sharon"
## [4,] "Yves"   "Leslie"
## [5,] "Ken"    "Erik"
## [6,] "Ken"    "Sharon"
## [7,] "Ken"    "Leslie"
## [8,] "Erik"   "Sharon"
## [9,] "Erik"   "Leslie"
## [10,] "Sharon" "Leslie"

```

**Definición 2.1.6.** Un **diseño de muestreo**  $p(\cdot)$  es una distribución de probabilidad multivariante definida sobre un soporte  $Q$ ; es decir,  $p(\cdot)$  es una función que va desde  $Q^2$  hasta  $(0, 1]$  tal que  $p(s) > 0$  para todo  $s \in Q$  y

$$\sum_{s \in Q} p(s) = 1 \quad (2.1.6)$$

---

<sup>2</sup>Nótese que  $Q$  es el espacio muestral cuyos elementos son vectores.

Dado el soporte  $Q$ , un **diseño de muestreo** es una función  $p(\cdot)$ , tal que  $p(s)$  arroja la probabilidad de selección de la muestra realizada  $s$  bajo un esquema de selección particular. En otras palabras, si  $S$  es una muestra aleatoria que toma el valor  $s$  con probabilidad  $p(s)$ , tal que

$$Pr(S = s) = p(s) \quad \text{para todo } s \in Q. \quad (2.1.7)$$

Entonces  $p(\cdot)$  es llamada diseño de muestreo.

El diseño muestreo, es una función que va desde el soporte  $Q$  hasta el intervalo  $]0, 1]$ . Por ser una distribución de probabilidad se tiene que  $p(\cdot)$  cumple que

1.  $p(s) \geq 0$  para todo  $s \in Q$
2.  $\sum_{s \in Q} p(s) = 1$

Nótese que el diseño de muestreo no se refiere a un algoritmo o procedimiento que permite la selección de muestras. Dado un diseño de muestreo, el trabajo del estadístico consiste en encontrar un algoritmo que permita la selección de muestras cuya probabilidad de selección corresponda a la probabilidad inducida por el diseño de muestreo. Para la realización de inferencias acerca de los parámetros de interés, el diseño de muestreo juega un papel muy importante porque las propiedades estadísticas (esperanza, varianza y otros) de las cantidades aleatorias que se calculan basadas en una muestra están determinadas por éste.

Dado un soporte  $Q$ , un diseño de muestreo puede ser:

- **Sin reemplazo** si todas las posibles muestras en  $Q$  son sin reemplazo.
- **Con reemplazo** si todas las posibles muestras en  $Q$  son con reemplazo.
- **De tamaño fijo** si todas las posibles muestras en  $Q$  tienen el mismo tamaño de muestra  $n(S) = n$ .

Cassel, Särndal & Wretman (1976a) explican que la posibilidad de identificar cada una de todas las posibles muestras que pertenecen al soporte  $Q$  es un factor crucial que permite:

- designar un conjunto de muestras a las cuales se les asigna una probabilidad positiva de selección y
- distribuir la totalidad de la masa de probabilidad entre los miembros de  $Q$ .

El rasgo más importante del muestreo probabilístico es que permite conocer, por lo menos teóricamente, la probabilidad de selección de todas las posibles muestras en el soporte  $Q$ . Sin embargo, un diseño de muestreo también deja conocer la probabilidad de inclusión del elemento  $k$  en la muestra  $S$ .

### Algoritmo de selección

Un diseño de muestreo es una distribución de probabilidad sobre un soporte  $Q$ ; pero, de ninguna manera, es un procedimiento que selecciona la muestra per se.

**Definición 2.1.7.** Un **algoritmo de selección** es un procedimiento usado para seleccionar una muestra probabilística.

Tillé (2006) afirma que una forma de seleccionar una muestra es listar todas las posibles muestras, generar una variable aleatoria con distribución uniforme en el intervalo  $[0, 1]$  para luego hacer la correspondiente selección. A este tipo de algoritmos que listan todas las posibles muestras se les conoce con el nombre de **algoritmos de selección enumerativos**; sin embargo, este tipo de algoritmos son ineficientes computacionalmente y sólo son posibles de implementar cuando el diseño de muestreo es conocido y el tamaño poblacional  $N$  es pequeño. A lo largo del libro se incluirán diversos algoritmos de selección específicos para cada diseño de muestreo que permitan la selección de una muestra probabilística.

### 2.1.4 Probabilidad de inclusión

La inclusión del elemento  $k$ -ésimo en una muestra  $s$  particular es un evento aleatorio definido por la función indicadora  $I_k(s)$ , que está dada por

$$I_k(s) = \begin{cases} 1 & \text{si } k \in s \\ 0 & \text{si } k \notin s. \end{cases} \quad (2.1.8)$$

Nótese que la función  $I_k(s)$  es una función de la variable aleatoria  $S$ . Para acortar la notación escribiremos  $I_k = I_k(s)$ , entendiéndose que  $I_k$  es la función indicadora para el elemento  $k$ -ésimo. Bajo un diseño de muestreo  $p(\cdot)$ , una **probabilidad de inclusión** es asignada a cada elemento de la población para indicar la probabilidad de que el elemento pertenezca a la muestra. Para el elemento  $k$ -ésimo de la población, la probabilidad de inclusión se denota como  $\pi_k$  y se conoce como la probabilidad de inclusión de **primer orden** y está dada por

$$\pi_k = Pr(k \in S) = Pr(I_k = 1) = \sum_{s \ni k} p(s). \quad (2.1.9)$$

En donde el subíndice  $s \ni k$  se refiere a la suma sobre todas las muestras que contienen al elemento  $k$ -ésimo. Nótese que de la anterior definición para que una muestra sea considerada probabilística, entonces todos los elementos en la población deben tener probabilidad de inclusión estrictamente mayor a cero.

**Definición 2.1.8.** La **esperanza de una muestra** aleatoria, en el sentido de las definiciones 2.1.2. y 2.1.3., está dada por

$$\mu = E(s) = \sum_{s \in Q} p(s)s \quad (2.1.10)$$

Si el diseño muestral es sin reemplazo, entonces  $\mu = \pi$ , donde  $\pi = (\pi_1, \dots, \pi_N)'$  es el vector de probabilidades de inclusión inducido por el diseño de muestreo. El siguiente resultado provee una manera sencilla para computar y realizar el cálculo de las  $N$  probabilidades de inclusión.

**Resultado 2.1.2.** Dado un soporte  $Q$ , la probabilidad de inclusión  $\pi_k$  es la probabilidad de que el elemento  $k$ -ésimo pertenezca a la muestra aleatoria  $S$  y se puede escribir de la siguiente manera:

$$\pi_k = E(I_k(S)) = \sum_{s \in Q} I_k(s)p(s) \quad (2.1.11)$$

*Demostración.*  $I_k(S)$  es una función de la muestra aleatoria  $S$ , la demostración se sigue de la definición de la esperanza de una función de una variable aleatoria. Por otro lado,  $I_k(S)$  sólo puede tomar dos valores 1 y 0, luego

$$\begin{aligned} E(I_k(S)) &= (1)Pr(I_k(S) = 1) + (0)Pr(I_k(S) = 0) \\ &= Pr(I_k(S) = 1) = Pr(k \in S) = \pi_k \end{aligned}$$

□

Análogamente,  $\pi_{kl}$  se conoce como la probabilidad de inclusión de **segundo orden** y denota la probabilidad de que los elementos  $k$  y  $l$  pertenezcan a la muestra, ésta se denota como  $\pi_{kl}$  y está dada por

$$\pi_{kl} = \Pr(k \in S \text{ y } l \in S) = \Pr(I_k I_l = 1) = \sum_{s \ni k \text{ y } l} p(s). \quad (2.1.12)$$

En donde el subíndice  $s \ni k$  y  $l$  se refiere a la suma sobre todas las muestras que contienen a los elementos  $k$ -ésimo y  $l$ -ésimo.

**Ejemplo 2.1.2.** Considere el siguiente diseño de muestreo  $p(\cdot)$  tal que asigna las siguientes probabilidades de selección a cada una de las 10 posibles muestras de tamaño 2 del soporte  $Q$  de la población  $U$ .

```
p <- c(0.13, 0.2, 0.15, 0.1, 0.15, 0.04, 0.02, 0.06, 0.07, 0.08)
p
## [1] 0.13 0.20 0.15 0.10 0.15 0.04 0.02 0.06 0.07 0.08
```

Es decir, la primera muestra tiene una probabilidad de selección de 0.13, la segunda muestra tiene una probabilidad de selección de 0.15, y así sucesivamente hasta la décima cuya probabilidad de selección es de 0.08. Con las siguientes instrucciones verificamos que las propiedades de diseño muestral sean satisfechas.

```
sum(p)
## [1] 1
p < 0
## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

Mediante el uso de la función `Ik` del paquete `TeachingSampling`, es posible crear las  $N = 5$  funciones indicadoras de los elementos pertenecientes a la población para cada una de las 10 posibles muestras de tamaño fijo y sin reemplazo. Esta función contiene dos argumentos: el tamaño de la población  $N$ , el tamaño fijo de cada una de las posibles muestras  $n$ . Una tabla de datos es creada a partir de los rótulos, la probabilidad de selección y las 5 funciones indicadoras de las posibles muestras contenidas en el soporte  $Q$ .

```
Ind <- Ik(N, n)
Q <- Support(N, n, U)

data.frame(Q, p, Ind)

##      X1     X2   p X1.1 X2.1 X3 X4 X5
## 1  Yves   Ken 0.13    1    1  0  0  0
## 2  Yves  Erik 0.20    1    0  1  0  0
## 3  Yves Sharon 0.15    1    0  0  1  0
## 4  Yves Leslie 0.10    1    0  0  0  1
```

```
## 5   Ken   Erik 0.15   0   1   1   0   0
## 6   Ken   Sharon 0.04  0   1   0   1   0
## 7   Ken   Leslie 0.02  0   1   0   0   1
## 8   Erik  Sharon 0.06  0   0   1   1   0
## 9   Erik  Leslie 0.07  0   0   1   0   1
## 10  Sharon Leslie 0.08  0   0   0   1   1
```

Una vez son calculadas las variables indicadoras para cada elemento y en cada posible muestra, el cálculo de las probabilidades de inclusión se hace muy sencillo al multiplicar las probabilidades de selección con cada una de las variables indicadoras. El resultado se suma por columnas y la salida es un vector de tamaño  $N = 5$  de probabilidades de inclusión.

```
multip <- p * Ind
colSums(multip)

## [1] 0.58 0.34 0.48 0.33 0.27
```

La función `Pik` del paquete `TeachingSampling` arroja el vector de probabilidades de inclusión para todos los elementos de la población. Ésta tiene dos argumentos: un vector `p` de probabilidades de selección de todas las posibles muestras y una matriz `Ind` de  $N$  variables indicadoras. Nótese que la suma de probabilidades de inclusión es el tamaño de muestra esperado, en este caso igual a 2.

```
pik <- Pik(p, Ind)
pik

##      [,1] [,2] [,3] [,4] [,5]
## [1,] 0.58 0.34 0.48 0.33 0.27
```

Luego, el elemento de la población que tiene una mayor probabilidad de ser incluido es **Yves**, mientras que el elemento con una menor probabilidad de inclusión es **Sharon**. Por otra parte, haciendo uso de la función `Pikl` del paquete `TeachingSampling` es posible calcular la matriz de probabilidades de inclusión de segundo orden para el diseño `p` en cuestión. Esta función sólo tiene tres argumentos: `N`, el tamaño de la población, `n`, el tamaño de muestra fijo y `p`, el diseño de muestreo utilizado. La salida de esta función es una matriz cuadrada y simétrica de tamaño  $N \times N$  cuyas entradas corresponden a las probabilidades de inclusión de segundo orden. Para este caso particular tenemos que la función se ejecuta de la siguiente manera.

```
pikl <- Pikl(N, n, p)
pikl

##      [,1] [,2] [,3] [,4] [,5]
## [1,] 0.58 0.13 0.20 0.15 0.10
## [2,] 0.13 0.34 0.15 0.04 0.02
## [3,] 0.20 0.15 0.48 0.06 0.07
## [4,] 0.15 0.04 0.06 0.33 0.08
## [5,] 0.10 0.02 0.07 0.08 0.27
```

Nótese que, bajo este diseño de muestreo, **Yves** y **Erik** corresponden al par de elementos que tienen la más alta probabilidad de inclusión.

### 2.1.5 Característica de interés y parámetros de interés

El propósito de cualquier estudio por muestreo es estudiar una **característica** de interés  $y$  que se encuentra asociada a cada unidad de la población. Es decir, la característica de interés toma el valor  $y_k$  para la unidad  $k$ . Es importante notar que los  $y_k$ s no se consideran variables aleatorias sino cantidades fijas, por tanto la notación de éstas se hace con un letra minúscula  $y$ . El objetivo de la investigación por muestreo es estimar una función de interés  $T$ , llamada **parámetro**, de la característica de interés en la población.

$$T = f\{y_1, \dots, y_k, \dots, y_N\}.$$

Algunos de los parámetros de interés más comunes son:

1. El total poblacional,

$$t_y = \sum_{k \in U} y_k \quad (2.1.13)$$

2. La media poblacional,

$$\bar{y}_U = \frac{\sum_{k \in U} y_k}{N} = \frac{t_y}{N} \quad (2.1.14)$$

3. La varianza poblacional,

$$S_{yU}^2 = \frac{\sum_{k \in U} (y_k - \bar{y}_U)^2}{N-1} \quad (2.1.15)$$

Existen otros parámetros de interés como la mediana poblacional, los percentiles poblaciones, la razón entre dos totales poblacionales o, como se mencionó anteriormente, el tamaño de una población, en cuyo caso estaríamos interesados en  $N$ . Entre otros, algunos ejemplos de investigaciones por muestreo interesadas en los anteriores parámetros son:

- Total de personas que pertenecen a la fuerza laboral.
- Porcentaje de personas que usarían un producto.

Obviamente, estas cantidades poblacionales son desconocidas y ésta es la razón por la que se requiere realizar una investigación por muestreo, porque mediante ésta se pueden estimar estos parámetros poblacionales a partir de una muestra seleccionada.

**Ejemplo 2.1.3.** Suponga que en nuestra población de ejemplo se quiere estimar el total de la variable  $y$ . El valor para cada uno de los elementos de la población es el siguiente:

```
y <- c(32, 34, 46, 89, 35)
y

## [1] 32 34 46 89 35
```

La función `data.frame` crea el conjunto de datos contenido los nombres (rótulos) y el valor de la característica de interés para cada elemento de la población

```
data.frame(U,y)
```

```
##      U  y
## 1   Yves 32
## 2     Ken 34
## 3   Erik 46
## 4 Sharon 89
## 5 Leslie 35
```

Algunos parámetros poblacionales de interés de la característica  $y$  son, el total poblacional y la media dados por  $t_y$  y  $\bar{y}_U$ , respectivamente.

```
ty <- sum(y)
ty

## [1] 236

ybar <- ty / N
ybar

## [1] 47.2
```

### 2.1.6 Estadística y estimador

Una **estadística** es una función  $G$  (que toma valores reales) de la muestra aleatoria  $S$  y sólo depende de los elementos pertenecientes a  $S$ . Cuando una estadística se usa para estimar un parámetro se dice **estimador** y las realizaciones del estimador en una muestra seleccionada  $s$  se dicen **estimaciones**.

Siendo  $G$  una estadística, sus propiedades estadísticas están determinadas por el diseño de muestreo. Es decir, dada la probabilidad de selección de cada muestra  $s \in Q$ , la esperanza, la varianza y otras propiedades de interés están definidas a partir de  $p(s)$ .

La **esperanza** de una estadística  $G$  es

$$E(G) = \sum_{s \in Q} p(s)G(s). \quad (2.1.16)$$

La **varianza** de la estadística  $G$  está definida como

$$Var(G) = E[G - E(G)]^2 \quad (2.1.17)$$

$$= \sum_{s \in Q} p(s)[G(s) - E(G)]^2. \quad (2.1.18)$$

Donde  $G(s)$  es el valor real que toma la estadística  $G$  en la muestra seleccionada (realizada)  $s$  y  $Q$  es el soporte inducido por el diseño muestral. Nótese que las propiedades de las estadísticas  $y$ , por consiguiente, de los estimadores, están definidas con sumas porque el diseño de muestreo induce una distribución de probabilidad discreta sobre todas las posibles muestras  $s$  pertenecientes al soporte  $Q$ .

#### La estadística $I_k$

La cantidad  $I_k$  dada por (2.1.8) es una estadística que toma valores aleatoriamente dependiendo del diseño de muestreo utilizado.

**Resultado 2.1.3.** Las propiedades más importantes de esta estadística son:

- $E(I_k) = \pi_k$
- $Var(I_k) = \pi_k(1 - \pi_k)$
- $Cov(I_k, I_l) = \pi_{kl} - \pi_k\pi_l$  para todo  $k \neq l$

*Demostración.* Por el resultado 2.1.2., la primera propiedad se tiene de inmediato, ahora de la definición de varianza se tiene

$$\begin{aligned} Var(I_k(S)) &= E[I_k(S) - E(I_k(S))]^2 \\ &= Pr(I_k(S) = 1)[1 - \pi_k]^2 + Pr(I_k(S) = 0)[0 - \pi_k]^2 \\ &= \pi_k(1 - \pi_k) \end{aligned}$$

y finalmente, de la definición de covarianza se tiene

$$\begin{aligned} Cov(I_k(S), I_l(S)) &= E[I_k(S)I_l(S)] - E[I_k(S)]E[I_l(S)] \\ &= (1)Pr(I_k(S)I_l(S) = 1) + (0)Pr(I_k(S)I_l(S) = 0) - \pi_k\pi_l \\ &= \pi_{kl} - \pi_k\pi_l \end{aligned}$$

□

A la covarianza de las estadísticas indicadoras para los elementos  $k$  y  $l$ ,  $Cov(I_k, I_l)$ , se le conoce como  $\Delta_{kl}$ . Esta cantidad, dependiendo del diseño, puede tomar valores positivos, negativos o incluso nulos.

#### La estadística $n(S)$ o tamaño de muestra

Como ya se vio, el tamaño de muestra es una cantidad aleatoria, dependiendo del diseño. Nótese que este valor puede ser expresado como función de las estadísticas de inclusión.

$$n(S) = \sum_U I_k. \quad (2.1.19)$$

**Resultado 2.1.4.** Algunas propiedades de interés son:

- $E(n(S)) = \sum_U \pi_k$
- $Var(n(S)) = \sum_U \pi_k - (\sum_U \pi_k)^2 + \sum \sum_{k \neq l} \pi_{kl}$ .

*Demostración.* Para la primera propiedad, se tiene que

$$E[n(S)] = E \left[ \sum_U I_k \right] = \sum_U E[I_k] = \sum_U \pi_k$$

Recordando que las propiedades de la varianza de una suma se tiene

$$\begin{aligned}
 Var[n(S)] &= Var\left[\sum_U I_k\right] \\
 &= \sum_U Var[I_k] + \sum \sum_{k \neq l} Cov[I_k, I_l] \\
 &= \sum_U \pi_k - \sum_U \pi_k^2 - \sum \sum_{k \neq l} \pi_k \pi_l + \sum \sum_{k \neq l} \pi_{kl} \\
 &= \sum_U \pi_k - \left(\sum_U \pi_k\right)^2 + \sum \sum_{k \neq l} \pi_{kl}
 \end{aligned}$$

□

Además, cuando la variación del tamaño de muestra es nula porque se ha decidido utilizar un diseño de tamaño muestral fijo, se tienen las siguientes propiedades.

**Resultado 2.1.5.** Si el diseño de muestreo es de tamaño fijo e igual a  $n$ ,

- $E(n(S)) = \sum_U \pi_k = n$
- $\sum_U \pi_{kl} = n\pi_l$
- $\sum_U \Delta_{kl} = 0$
- $\pi_k(1 - \pi_k) = \sum_{l \neq k} (\pi_k \pi_l - \pi_{kl})$

*Demostración.* La primera propiedad se tiene recordando que la esperanza de una constante es ella misma. Nótese que  $\pi_{kl} = E[I_k(S)I_l(S)]$ , así

$$\begin{aligned}
 \sum_{l \in U} \pi_{kl} &= \sum_{l \in U} E[I_k(S)I_l(S)] = \sum_{l \in U} \sum_{s \in Q} p(s)I_k(s)I_l(s) \\
 &= \sum_{s \in Q} p(s)I_k(s) \sum_{l \in U} I_l(s) \\
 &= n(S) \sum_{s \in Q} p(s)I_k(s) = n\pi_k
 \end{aligned}$$

La tercera propiedad se tiene pues

$$\begin{aligned}
 \sum_U \Delta_{kl} &= \sum_U (\pi_{kl} - \pi_k \pi_l) \\
 &= \sum_U \pi_{kl} - \pi_k \sum_U \pi_l \\
 &= n\pi_k - n\pi_k = 0
 \end{aligned}$$

Para demostrar la última propiedad es necesario redefinir el tamaño de muestra, de tal manera que

$n = \sum_{l \neq k} I_l(S) + I_k(S)$ . Luego,

$$\begin{aligned}\pi_k(1 - \pi_k) &= Var(I_k(S)) \\ &= Cov(I_k(S), I_k(S)) \\ &= Cov\left(I_k(S), n - \sum_{l \neq k} I_l(S)\right) \\ &= - \sum_{l \neq k} Cov(I_k(S), I_l(S)) \\ &= \sum_{l \neq k} (\pi_k \pi_l - \pi_{kl})\end{aligned}$$

□

**Ejemplo 2.1.4.** Continuando con el desarrollo del ejemplo 2.1.3, ahora utilizaremos el vector de probabilidades de inclusión y la matriz de probabilidades de segundo orden para verificar los resultados 2.1.4 y 2.1.5. En primer lugar, nótese que la esperanza del tamaño de muestra, que corresponde a 2 pues el diseño es de tamaño fijo, se obtiene de la siguiente manera.

```
A <- sum(pik)
A

## [1] 2
```

Ahora, el cuadrado de la suma de las probabilidades de inclusión se obtiene así

```
B <- (sum(pik))^2
B

## [1] 4
```

Y la suma de los elementos distintos de la matriz de probabilidades de inclusión de segundo orden es

```
C <- sum(pikl) - sum(diag(pikl))
C

## [1] 2
```

Para comprobar la segunda parte del resultado 2.1.4. basta realizar la siguiente operación  $A-B+C$ . Esta suma es nula y efectivamente corresponde a la varianza del tamaño de muestra en este diseño de muestreo; como, en este caso particular, el tamaño de muestra siempre fue fijo e igual a 2, la varianza debe ser cero.

El siguiente paso de este ejemplo consiste en la verificación de la segunda parte del resultado 2.1.5. En resumidas cuentas, este apartado dice que la suma por filas (o columnas) de la matriz de probabilidades de inclusión de segundo orden debe corresponder exactamente a la multiplicación del tamaño de muestra y el vector de probabilidades de inclusión de primer orden. Lo anterior se corrobora fácilmente por medio del siguiente código.

```

n * pik

##      [,1] [,2] [,3] [,4] [,5]
## [1,] 1.16 0.68 0.96 0.66 0.54

colSums(pikl)

## [1] 1.16 0.68 0.96 0.66 0.54

rowSums(pikl)

## [1] 1.16 0.68 0.96 0.66 0.54

```

Nótese que la suma por filas y por columnas coincide perfectamente con  $n \times \pi_k$  para todo  $k \in U$ . Por otro lado, verificaremos la tercera propiedad que afirma que la suma por filas (o columnas) de la matriz de varianzas-covarianzas de las variables indicadoras de membresía muestral debe dar como resultado un vector de ceros de tamaño cinco. Para esto, se utiliza la función `Deltak1` del paquete `TeachingSampling`. Esta función tiene tres argumentos: `N`, el tamaño de la población, `n`, el tamaño de muestra fijo y `p`, el diseño de muestreo utilizado. La salida de esta función corresponde a una matriz cuadrada y simétrica de tamaño  $N \times N$  cuyas entradas corresponden a las varianzas-covarianzas de las variables indicadoras de membresía muestral. Para este ejemplo, la implementación del siguiente código permite obtener la matriz buscada y la verificación del resultado.

```

Delta <- Deltak1(N, n, p)
Delta

##      [,1]     [,2]     [,3]     [,4]     [,5]
## [1,] 0.2436 -0.0672 -0.0784 -0.0414 -0.0566
## [2,] -0.0672  0.2244 -0.0132 -0.0722 -0.0718
## [3,] -0.0784 -0.0132  0.2496 -0.0984 -0.0596
## [4,] -0.0414 -0.0722 -0.0984  0.2211 -0.0091
## [5,] -0.0566 -0.0718 -0.0596 -0.0091  0.1971

rowSums(Delta)

## [1] -0.00000000000000013878 -0.00000000000000008327 -0.00000000000000005551
## [4] -0.00000000000000006939 -0.0000000000000001388

colSums(Delta)

## [1] -0.00000000000000013878 -0.00000000000000008327 -0.00000000000000005551
## [4] -0.00000000000000006939 -0.0000000000000001388

```

De esta manera la suma por filas (o columnas) de la matriz de varianzas-covarianzas de las variables indicadoras de membresía muestral es cero en cada columna (o fila).

Cuando una estadística se construye con la intención de estimar un parámetro, recibe el nombre de **estimador**. Así, las propiedades más comúnmente utilizadas de un estimador  $\hat{T}$  de un parámetro de

interés  $T$  son el sesgo, definido por

$$B(\hat{T}) = E(\hat{T}) - T \quad (2.1.20)$$

y el error cuadrático medio, dado por

$$ECM(\hat{T}) = E[\hat{T} - T]^2 \quad (2.1.21)$$

$$= Var(\hat{T}) + B^2(\hat{T}). \quad (2.1.22)$$

Si el sesgo de un estimador es nulo se dice que el estimador es **insesgado** y cuando esto ocurre el error cuadrático medio se convierte en la varianza del estimador.

Särndal, Swensson & Wretman (1992) afirman que el objetivo en un estudio por muestreo es estimar uno a más parámetros poblacionales. Las decisiones más importantes a la hora de abordar un problema de estimación por muestreo son

- La escogencia de un diseño de muestreo y un algoritmo de selección que permita implementar el diseño.
- La elección de una fórmula matemática o estimador que calcule una estimación del parámetro de interés en la muestra seleccionada.

Las anteriores no son decisiones independientes. Es decir, la escogencia de un estimador dependerá, usualmente, del diseño de muestreo utilizado.

**Definición 2.1.9.** Siendo  $\hat{T}$  un estimador de un parámetro  $T$  y  $p(\cdot)$  un diseño de muestreo definido sobre un soporte  $Q$ , se define una **estrategia de muestreo** como la dupla  $(p(\cdot), \hat{T})$ .

Este libro, como su nombre lo indica, está enfocado en la búsqueda de la mejor combinación de diseño de muestreo y estimador; este problema ha sido considerado a través del desarrollo de la teoría de muestreo. La escogencia de la estrategia de muestreo se lleva a cabo en dos etapas, a saber: **Etapa de diseño**, refiriéndose al periodo durante el cual se decide el diseño de muestreo a utilizar junto con el algoritmo de muestreo que permite la selección de la muestra y finalmente se selecciona la muestra probabilística. Una vez que la información es recogida y grabada entra la **Etapa de estimación** en donde se calculan las estimaciones para la característica de interés utilizando el estimador propio de la estrategia de muestreo escogida.

## 2.2 Estimadores de muestreo

Cada elemento perteneciente a la población tiene una característica de interés asociada  $y$ . Para el elemento  $k$ -ésimo el valor que toma esta característica de interés es  $y_k$ . El objetivo de la investigación por muestreo es estimar un parámetro  $T$  que resulta de interés. El objetivo del estadístico es poder inferir acerca de  $T$  con base en una muestra  $s$ . Un indicador de la precisión de un estimador está dado por el **coeficiente de variación estimado** dado por

$$cve(\hat{T}) = \frac{\sqrt{\widehat{Var}(\hat{T})}}{\hat{T}} \quad (2.2.1)$$

donde  $\widehat{Var}(\hat{T})$  es el estimador de la varianza basado en la muestra seleccionada  $s$ . El coeficiente de variación estimado es una medida comúnmente usada para expresar el error cometido al seleccionar

una muestra y ni utilizar a toda la población en la medición de la variable de interés. Si se realizara un censo y el estimador reprodujera el parámetro poblacional, entonces  $\widehat{Var}(\hat{T})$  sería nula y, por lo tanto, el *cve* también sería nulo.

A continuación, se revisan algunos de los estimados más utilizados en la historia del muestreo. A medida que se avance en la lectura del libro, nuevos estimadores surgirán y, por consiguiente, nuevas estrategias de muestreo que permiten llegar a resultados con una precisión casi clínica. La mayoría de los estimadores presentados en este libro son estimadores de totales o de funciones de totales.

### 2.2.1 El estimador de Horvitz-Thompson

#### Estimador del total poblacional

Narain (1951) descubrió este estimador, aunque su artículo fue editado y publicado por una revista india de poca rotación. Más adelante Horvitz & Thompson (1952) publicaron similares resultados en la revista más importante de estadística en ese tiempo, JASA (Journal of the American Statistical Society). Desde entonces, este estimador se conoce como el estimador de Horvitz-Thompson o estimador  $\pi$ , aunque rigurosamente debería ser llamado estimador de Narain-Horvitz-Thompson . En este libro seguiremos la notación internacional y clásica.

Para un universo  $U$ , se quiere estimar el total poblacional  $t_y$  de la característica de interés  $y$  dado por (2.1.13). Se define el estimador de Horvitz-Thompson(HT) para  $t_y$  como:

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} = \sum_S d_k y_k \quad (2.2.2)$$

Donde  $\pi_k$  es la probabilidad de inclusión para el  $k$ -ésimo elemento, y  $d_k$  es conocido como **factor de expansión** y corresponde al inverso de la probabilidad de inclusión. Nótese que el estimador de Horvitz-Thompson es aleatorio porque está construido con base en una suma sobre la muestra aleatoria  $S$ . La motivación detrás de este estimador, como Brewer (2002) lo indica, descansa en el **principio de representatividad** que afirma que cada elemento incluido en una muestra se representa a sí mismo y a un grupo de unidades que no pertenecen a la muestra seleccionada, cuyas características son cercanas a las del elemento incluido en la muestra. El factor de expansión no es otra cosa que el número de elementos menos uno de la población (no incluidos en la muestra) representados por el elemento incluido.

**Resultado 2.2.1.** Si todas las probabilidades de inclusión de primer orden son mayores a cero ( $\pi_k > 0$  para todo  $k$ ), el estimador de Horvitz-Thompson es insesgado para el total poblacional. Por tanto, se tiene que

$$E(\hat{t}_{y,\pi}) = t_y \quad (2.2.3)$$

*Demostración.* Reescribiendo el estimador de Horvitz-Thompson como  $\hat{t}_{y,\pi} = \sum_S I_k(S) \frac{y_k}{\pi_k}$ , se tiene

$$E(\hat{t}_{y,\pi}) = E \left( \sum_U I_k(S) \frac{y_k}{\pi_k} \right) = \sum_U \frac{y_k}{\pi_k} E(I_k(S)) = \sum_U \pi_k \frac{y_k}{\pi_k} = t_y$$

□

Si el diseño de muestreo es tal que las probabilidades de inclusión de primer orden conservan una buena correlación positiva con la medición de la característica de interés; en otras palabras, si  $\pi_k \propto y_k$ , el estimador de Horvitz-Thompson se reduce a una constante, por lo tanto tendrá varianza nula. En la práctica, una estrategia de muestreo óptima (Cassel, Särndal & Wretman 1976a) es aquella que utiliza el estimador de Horvitz-Thompson junto con un diseño de muestreo que induzca una buena

correlación entre el vector de probabilidades de inclusión y el vector de valores de la característica de interés. Sin embargo, en encuestas multi-propósito, en donde se quiere estimar parámetros para varias características de interés entre las cuales no hay una buena correlación, al utilizar el estimador de Horvitz-Thompson es difícil evadir la débil, e incluso negativa, correlación que existe entre las características de interés y el vector de probabilidades de inclusión. Sin embargo, al incluir información auxiliar en la construcción del estimador se puede palear este hecho.

### Varianza del estimador de Horvitz-Thompson

**Resultado 2.2.2.** *La varianza del estimador de Horvitz-Thompson está dada por la siguiente expresión*

$$Var_1(\hat{t}_{y,\pi}) = \sum_U \sum_{kl} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}. \quad (2.2.4)$$

*Demostración.* De la definición de varianza, se obtiene lo siguiente

$$\begin{aligned} Var_1(\hat{t}_{y,\pi}) &= Var \left( \sum_U I_k(S) \frac{y_k}{\pi_k} \right) \\ &= \sum_U \frac{y_k^2}{\pi_k^2} Var(I_k(S)) + \sum_U \sum_{k \neq l} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} Cov(I_k(S), I_l(S)) \\ &= \sum_U \frac{y_k^2}{\pi_k^2} (\pi_k - \pi_k^2) + \sum_U \sum_{k \neq l} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} (\pi_{kl} - \pi_k \pi_l) \\ &= \sum_U \sum_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} (\pi_{kl} - \pi_k \pi_l) \\ &= \sum_U \sum_{kl} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \end{aligned}$$

□

Sen (1953) y Yates & Grundy (1953) dedujeron el siguiente resultado cuando el diseño de muestreo es de tamaño fijo.

**Resultado 2.2.3.** *Si el diseño  $p(\cdot)$  es de tamaño de muestra fijo, entonces, la varianza del estimador de Horvitz-Thompson se escribe como*

$$Var_2(\hat{t}_{y,\pi}) = -\frac{1}{2} \sum_U \sum_{kl} \Delta_{kl} \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \quad (2.2.5)$$

*Demostración.* Utilizando las propiedades del resultado 2.1.5, se tiene que

$$\begin{aligned}
 Var_2(\hat{t}_{y,\pi}) &= -\frac{1}{2} \sum_U \sum_{kl} \Delta_{kl} \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \\
 &= -\frac{1}{2} \sum_U \sum_{kl} \Delta_{kl} \left( \frac{y_k^2}{\pi_k^2} + \frac{y_l^2}{\pi_l^2} - 2 \frac{y_k y_l}{\pi_k \pi_l} \right) \\
 &= -\frac{1}{2} \left[ \sum_U \sum_{kl} \Delta_{kl} \frac{y_k^2}{\pi_k^2} + \sum_U \sum_{kl} \Delta_{kl} \frac{y_l^2}{\pi_l^2} - 2 \sum_U \sum_{kl} \Delta_{kl} \frac{y_l y_k}{\pi_k \pi_l} \right] \\
 &= -\frac{1}{2} \left[ 2 \sum_U \sum_{kl} \Delta_{kl} \frac{y_k^2}{\pi_k^2} - 2 \sum_U \sum_{kl} \Delta_{kl} \frac{y_l y_k}{\pi_k \pi_l} \right] \\
 &= -\sum_U \frac{y_k^2}{\pi_k^2} \sum_U \Delta_{kl} + \sum_U \sum_{kl} \Delta_{kl} \frac{y_l y_k}{\pi_k \pi_l} \\
 &= \sum_U \sum_{kl} \Delta_{kl} \frac{y_l y_k}{\pi_k \pi_l} = Var_1(\hat{t}_{y,\pi})
 \end{aligned}$$

puesto que  $\sum_U \Delta_{kl} = 0$  para diseños de tamaño fijo. Por lo tanto, en los casos de diseños de muestreo con tamaño fijo, la varianza del estimador de Horvitz-Thompson puede calcularse por medio de  $Var_2(\hat{t}_{y,\pi})$ .  $\square$

### Estimación de la varianza

Es posible construir dos estimadores insesgados para las expresiones (2.2.4) y (2.2.5). Para esto, se requiere que todas las probabilidades de inclusión de segundo orden sean estrictamente positivas ( $\pi_{kl} > 0$  para todo  $k$ ). Con el anterior supuesto, se tienen los siguientes resultados.

**Resultado 2.2.4.** *Un estimador insesgado para la expresión (2.2.4) está dada por*

$$\widehat{Var}_1(\hat{t}_{y,\pi}) = \sum_S \sum_{kl} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \quad (2.2.6)$$

**Resultado 2.2.5.** *Si el diseño es de tamaño de muestra fijo, un estimador insesgado para la expresión (2.2.5) está dado por*

$$\widehat{Var}_2(\hat{t}_{y,\pi}) = -\frac{1}{2} \sum_S \sum_{kl} \frac{\Delta_{kl}}{\pi_{kl}} \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \quad (2.2.7)$$

*Demostración.* Los anteriores resultados son inmediatos al reescribir los estimadores  $\widehat{Var}_1(\hat{t}_{y,\pi})$  y  $\widehat{Var}_2(\hat{t}_{y,\pi})$  en términos de  $U$  y multiplicar por el producto de las funciones indicadoras  $I_k(S)I_l(S)$ . Al aplicar la esperanza se tiene que  $E[I_k(S)I_l(S)] = \pi_{kl}$  y con esto se tiene la demostración.  $\square$

Bautista (1998) resalta los tres siguientes comentarios importantes acerca de las estimaciones arrojadas por anteriores expresiones.

1. Si las probabilidades de inclusión de segundo orden son mayores que cero para todos los elementos en la muestra, pero no para los restantes elementos que no fueron incluidos en la muestra, no se puede garantizar el insesgamiento de las anteriores expresiones.
2. Es posible que las estimaciones de la varianza arrojen resultados negativos, que no pueden ser utilizados ni interpretados. Para evitar esta situación, es necesario garantizar que la covarianza entre las estadísticas de inclusión para cada par de elementos en la población sea negativa ( $\Delta_{kl} < 0 \forall k \neq l$ ).

3. No necesariamente las estimaciones arrojadas por las anteriores expresiones coinciden en todos los casos.

Por su parte, Tillé (2006) agrega que en la práctica, la utilización de las expresiones de los estimadores de la varianza es muy difícil de implementar pues la doble suma hace que el proceso de cálculo computacional sea muy largo e ineficiente. Por lo tanto, para cada diseño de muestreo que se utilice, se deben crear expresiones que pueden ser simplificadas o en algunos casos se deben utilizar aproximaciones.

### Intervalo de confianza para el estimador de Horvitz-Thompson

Hájek (1960) demuestra la convergencia asintótica del estimador de Horvitz-Thompson a una distribución normal. Cuando el tamaño de muestra es suficientemente grande (que dependiendo del comportamiento de la población puede bastar con algunas docenas de individuos), se puede construir un intervalo de confianza de nivel  $(1 - \alpha)$  para el total poblacional  $t_y$  de acuerdo con:

$$IC(1 - \alpha) = \left[ \hat{t}_{y,\pi} - z_{1-\alpha/2} \sqrt{Var(\hat{t}_{y,\pi})}, \hat{t}_{y,\pi} + z_{1-\alpha/2} \sqrt{Var(\hat{t}_{y,\pi})} \right] \quad (2.2.8)$$

donde  $z_{1-\alpha/2}$  se refiere al cuantil  $(1 - \alpha/2)$  de una variable aleatoria con distribución normal estándar. Nótese que

$$1 - \alpha = \sum_{Q_0 \supset s} p(s),$$

donde  $Q_0$  es el conjunto de todas las posibles muestras cuyo intervalo de confianza contiene al total poblacional  $t_y$ . En la práctica muy pocas veces se conoce la varianza del estimador; por lo tanto, el intervalo de confianza estimado de nivel  $(1 - \alpha)$  puede ser obtenido con los datos de la muestra seleccionada reemplazando en (2.2.8) la varianza del estimador por su correspondiente estimación y tomaría la siguiente expresión

$$IC_s(1 - \alpha) = \left[ \hat{t}_{y,\pi} - z_{1-\alpha/2} \sqrt{\widehat{Var}(\hat{t}_{y,\pi})}, \hat{t}_{y,\pi} + z_{1-\alpha/2} \sqrt{\widehat{Var}(\hat{t}_{y,\pi})} \right]. \quad (2.2.9)$$

Al utilizar una estrategia de muestreo en la estimación de un parámetro en poblaciones finitas, las propiedades de la estrategia se estudian en términos de:

- **Confidencialidad:** definida como la suma de las probabilidades de las muestras cuyo intervalo de confianza contiene al parámetro.
- **Precisión:** definida como la longitud del intervalo de confianza.

Nótese que las anteriores propiedades están en función del intervalo de confianza. Para determinar la confidencialidad se debe conocer al parámetro  $T$  (desconocido) por tanto, en términos prácticos la confidencialidad no se puede calcular. Para determinar la precisión y la confidencialidad se requiere conocer la varianza, basada en el diseño de muestreo, del estimador utilizado, digamos  $\hat{T}$ ; sin embargo, el cálculo de esta varianza  $Var(\hat{T})$  implica, casi siempre, el requerimiento de conocer los valores  $y_k$  para todo  $k = 1, \dots, N$ . Luego la precisión tampoco se puede calcular. Sin embargo se debe proponer un estimador de  $Var(\hat{T})$  (ojalá insesgado) que junto con  $\hat{T}$  proporciona una cota para el sesgo y para la precisión.

### Estimación de otros parámetros

Aunque (2.2.2) es un estimador del total poblacional de la característica de interés, se puede utilizar para estimar otras cantidades poblacionales de interés. Si el tamaño poblacional  $N$  es conocido, la media poblacional definida en (2.1.14) puede ser estimada con el estimador de Horvitz-Thompson.

**Resultado 2.2.6.** *La media poblacional es estimada insesgadamente mediante el uso de la siguiente expresión*

$$\hat{y}_\pi = \frac{1}{N} (\hat{t}_{y,\pi}) = \frac{1}{N} \sum_s \frac{y_k}{\pi_k} \quad (2.2.10)$$

La varianza y la varianza estimada del estimador de la media poblacional están dadas por

$$Var(\hat{y}_\pi) = \frac{1}{N^2} Var(\hat{t}_{y,\pi}) \quad (2.2.11)$$

$$Var(\hat{y}_\pi) = \frac{1}{N^2} Var(\hat{t}_{y,\pi}) \quad (2.2.12)$$

respectivamente,

Sin embargo, es la regla más que la excepción que en la mayoría de casos en donde el usuario se enfrenta a una investigación cuyos objetivos están supeditados a la realización de un estudio por muestreo que el tamaño poblacional sea desconocido. En tal caso, podemos usar el estimador de Horvitz-Thompson para estimarlo puesto que  $N$  puede ser escrito de la siguiente manera

$$N = \sum_U 1, \quad (2.2.13)$$

tomando la conocida forma de un total poblacional. Luego, tenemos el siguiente resultado.

**Resultado 2.2.7.** *El tamaño poblacional es estimado insesgadamente mediante el uso de la siguiente expresión*

$$\hat{N}_\pi = \sum_S \frac{1}{\pi_k}. \quad (2.2.14)$$

Cuando se ha estimado el total poblacional de una característica de interés y el tamaño poblacional mediante el uso del estimador de Horvitz-Thompson, surge un estimador para la media poblacional dado por

$$\tilde{y}_S = \frac{\hat{t}_{y,\pi}}{\hat{N}_\pi} \quad (2.2.15)$$

$$= \sum_S \frac{y_k}{\pi_k} / \sum_S \frac{1}{\pi_k}. \quad (2.2.16)$$

La anterior expresión es una razón, o un cociente entre dos totales poblacionales. Las propiedades estadísticas de los anteriores estimadores serán tratados más adelante en las secciones pertinentes del libro.

Tillé (2006) cita que aun al conocer  $N$ , una mala propiedad del estimador de Horvitz-Thompson para la media poblacional se tiene al utilizarlo cuando la característica de interés es constante para todos los elementos de la población ( $y_k = C \forall k \in U$ ). Por supuesto, bajo las anteriores condiciones es claro que

la media poblacional es igual a la constante ( $\bar{y}_U = C$ ). Sin embargo, el estimador  $\hat{y}_\pi$  toma la siguiente forma

$$\hat{y}_\pi = \frac{1}{N} \sum_s \frac{y_k}{\pi_k} = \frac{1}{N} \sum_s \frac{C}{\pi_k} = \frac{C}{N} \sum_s \frac{1}{\pi_k} = C \frac{\hat{N}_\pi}{N}. \quad (2.2.17)$$

Al respecto, Bautista (1998) afirma que en aquellos casos en los que se conoce el valor de  $N$  es preferible ignorarlo y utilizar el estimador  $\tilde{y}_S$  puesto que su variación es menor y cuando  $y_k = C \forall k \in U$  reproduce la media poblacional con varianza nula puesto que

$$\tilde{y}_S = \frac{\hat{t}_{y,\pi}}{\hat{y}_\pi} = \frac{C \hat{y}_\pi}{\hat{y}_\pi} = C.$$

Cuando el tamaño poblacional es conocido y, como se verá más adelante, para algunos diseños de muestreo sin reemplazo, se puede crear un nuevo estimador alternativo del total poblacional inspirado en el siguiente argumento: Si  $\tilde{y}_S$  estima la media poblacional, entonces  $N\tilde{y}_S$  estimará el total poblacional. Por tanto, el estimador alternativo está dado por la siguiente expresión

$$\hat{t}_{y,alt} = N\tilde{y}_S = \hat{t}_{y,\pi} \frac{N}{\hat{N}_\pi} \quad (2.2.18)$$

que se puede ver como una corrección del estimador de Horvitz-Thompson mediante la estimación del tamaño de la población. La varianza y la estimación de la varianza serán tema de capítulos posteriores.

**Ejemplo 2.2.1.** La función HT del paquete **TeachingSampling** arroja la estimación del total poblacional de una o varias características de interés. Esta función tiene dos argumentos: el vector de tamaño  $n$  de probabilidades de inclusión **pik** y el conjunto de valores de la característica o características de interés en los individuos pertenecientes a la muestra, y puede ser un vector en el caso de una sola característica de interés o una matriz en el caso de varias.

Así, si la primera muestra (cuyos elementos son **Yves** y **Ken**) hubiese sido seleccionada y dado que las probabilidades de inclusión de estos dos elementos son 0.58 y 0.34, respectivamente y los valores de la característica de interés son 32 y 34, respectivamente, el estimador de Horvitz-Thompson arrojaría la siguiente estimación:

```
y.s <- c(32, 34)
pik.s <- c(0.58, 0.34)
HT(y.s, pik.s)

##      [,1]
## [1,] 155.2
```

Nótese que el total poblacional para la variable de interés  $y$  es igual a 236. Por otro lado, el cálculo o estimación de la varianza del estimador de Horvitz-Thompson no se encuentra implementado pues la doble suma hace que los procesos computacionales sean muy largos y demorado. Por tanto, si se quieren conocer estos valores, el proceso se debe realizar manualmente. La estimación de la varianza se realiza teniendo en cuenta que  $\pi_{12} = 0.13$ . Así,

$$\begin{aligned}\frac{\Delta_{11}}{\pi_{11}} &= \frac{\pi_{11} - \pi_1\pi_1}{\pi_{11}} = \frac{0.58 - 0.58^2}{0.58} = 0.42 \\ \frac{\Delta_{12}}{\pi_{12}} &= \frac{\pi_{12} - \pi_1\pi_2}{\pi_{12}} = \frac{0.13 - 0.58 * 0.34}{0.13} = -0.52 \\ \frac{\Delta_{21}}{\pi_{21}} &= \frac{\pi_{21} - \pi_2\pi_1}{\pi_{21}} = \frac{0.13 - 0.34 * 0.58}{0.13} = -0.52 \\ \frac{\Delta_{22}}{\pi_{22}} &= \frac{\pi_{22} - \pi_2\pi_2}{\pi_{22}} = \frac{0.34 - 0.34^2}{0.34} = 0.66\end{aligned}$$

Por tanto, utilizando (2.2.6), el estimador de la varianza será

$$\widehat{Var}(\hat{t}_\pi) = \frac{\Delta_{11}}{\pi_{11}} \frac{y_1}{\pi_1} \frac{y_1}{\pi_1} + \frac{\Delta_{12}}{\pi_{12}} \frac{y_1}{\pi_1} \frac{y_2}{\pi_2} + \frac{\Delta_{21}}{\pi_{21}} \frac{y_2}{\pi_2} \frac{y_1}{\pi_1} + \frac{\Delta_{22}}{\pi_{22}} \frac{y_2}{\pi_2} \frac{y_2}{\pi_2}$$

y su respectiva estimación será

$$0.42 \left( \frac{32}{0.58} \right)^2 - 2(0.52) \left( \frac{32}{0.58} \frac{34}{0.34} \right) + 0.66 \left( \frac{34}{0.34} \right)^2 \cong 2140$$

El coeficiente de variación estimado es

$$cve(\hat{t}_\pi) = \frac{\sqrt{2140}}{155.1724} \cong 0.3$$

Y el intervalo de confianza estimado con un nivel de confianza del 95 por ciento para esta estimación es el siguiente:

$$\begin{aligned}IC_s(0.95) &\cong \left[ 155 - (1.96)\sqrt{2140}, 155 + (1.96)\sqrt{2140} \right] \\ &\cong [64, 246]\end{aligned}$$

Continuando con el ejercicio léxico-gráfico de la estimación del total poblacional  $t_y$  en todas las posibles muestras de tamaño 10 de la población  $U$ , tenemos la tabla 2.1 que puede ser reproducida mediante la ejecución del siguiente código computacional.

```
all.pik <- Support(N, n, pik)
all.y <- Support(N, n, y)
all.HT <- rep(0, 10)

for(k in 1:10){
  all.HT[k] <- HT(all.y[,], all.pik[,])
}

all.HT

## [1] 155.2 151.0 324.9 184.8 195.8 369.7 229.6 365.5 225.5 399.3

AllSamples=data.frame(Q, p, all.pik, all.y, all.HT)
```

	1	2	3	4	5	6	7	8
1	Yves	Ken	0.13	0.58	0.34	32.00	34.00	155.17
2	Yves	Erik	0.20	0.58	0.48	32.00	46.00	151.01
3	Yves	Sharon	0.15	0.58	0.33	32.00	89.00	324.87
4	Yves	Leslie	0.10	0.58	0.27	32.00	35.00	184.80
5	Ken	Erik	0.15	0.34	0.48	34.00	46.00	195.83
6	Ken	Sharon	0.04	0.34	0.33	34.00	89.00	369.70
7	Ken	Leslie	0.02	0.34	0.27	34.00	35.00	229.63
8	Erik	Sharon	0.06	0.48	0.33	46.00	89.00	365.53
9	Erik	Leslie	0.07	0.48	0.27	46.00	35.00	225.46
10	Sharon	Leslie	0.08	0.33	0.27	89.00	35.00	399.33

Cuadro 2.1: Estimación para todas las posibles muestras del ejemplo

El vector `all.est` contiene las estimaciones Horvitz-Thompson para cada una de las 10 posibles muestras, su esperanza se calcula como

```
sum(p * all.HT)
## [1] 236
```

Nótese que la esperanza del estimador de Horvitz-Thompson reproduce exactamente el total poblacional. La varianza se calcula de la siguiente manera

$$\begin{aligned} \text{Var}(\hat{t}_\pi) &= (0.13)(155.2 - 236)^2 + (0.2)(151.0 - 236)^2 + \dots \\ &\quad + (0.08)(399.3 - 236)^2 = 7847.2 \end{aligned}$$

Acudiendo a la función `VarHT`, del paquete `TeachignSampling`, es posible reproducir este mismo cálculo de la varianza. Sin embargo, esta función utiliza la expresión teórica de la varianza  $\text{Var}_1(\hat{t}_{y,\pi})$  dada por (2.2.4) para diseños de muestreo de tamaño fijo. Tiene cuatro argumentos: `y`, que es un vector que contiene los valores de la característica de interés en todos y cada uno de los elementos de la población; `N`, el tamaño de la población; `n`, el tamaño de muestra fijo y `p`, el diseño de muestreo utilizado. El resultado de esta función es el cálculo del valor de la varianza teórica del estimador de Horvitz-Thompson para un diseño de muestreo y una configuración de valores poblacionales particular. Siguiendo con el diseño de muestreo dado en el ejemplo 2.1.2 y la configuración de valores de la característica de interés del ejemplo 2.1.3, tenemos que el cálculo de la varianza es exactamente igual al dado por el ejercicio léxico-gráfico.

```
VarHT(y, N, n, p)
## [1] 7847
```

## 2.2.2 El estimador de Hansen-Hurwitz

### Sobre el muestreo con reemplazo

Considere una población finita de  $N$  elementos y un diseño de muestreo que permite la selección de una muestra realizada  $s$ , con reemplazo, de tamaño  $m$ . Como Lohr (2000) lo afirma, la manera más

intuitiva de entender este tipo de diseños muestrales con reemplazo es pensar en la extracción de  $m$  muestras independientes de tamaño 1. Se extrae un elemento de la población para ser incluido en la muestra con una probabilidad  $p_k$ ; sin embargo, ese mismo elemento participa en el siguiente sorteo aleatorio. Este proceso se repite  $m$  veces; es decir, se tiene un total de  $m$  sorteos aleatorios.

Bajo el anterior esquema de selección, es claro que un elemento puede ser seleccionado en la muestra más de una vez; por lo tanto, aunque el tamaño de la muestra seleccionada con reemplazo es  $m$ , el tamaño de muestra efectivo no es necesariamente  $m$ . Nótese que la selección de un elemento que se repite más de una vez no proporciona información nueva. Es por esto que en la práctica, se prefieren los diseños de muestreo que permita la selección de muestras sin duplicados.

Särndal, Swensson & Wretman (1992) afirman que el marco general del muestreo con reemplazo tiene las siguientes características:

- Cada elemento de la población está relacionado directamente con un número positivo  $p_k$  ( $k = 1, \dots, N$ ) de tal forma que

$$\sum_U p_k = 1.$$

A  $p_k$  se le conoce como la **probabilidad de selección** del elemento  $k$ -ésimo. Nótese que estas probabilidades no son necesariamente iguales.

- Para seleccionar el primer elemento que pertenecerá a la muestra de tamaño  $m$ , se lleva a cabo un sorteo aleatorio de tal forma que

$$Pr(\text{Seleccionar el elemento } k) = p_k, \quad k \in U.$$

- El elemento seleccionado es reemplazado en la población y vuelva a ser parte del próximo sorteo aleatorio con la misma probabilidad de selección  $p_k$ .
- El mismo conjunto de probabilidades es usado para seleccionar los restantes elementos. En total se realizan  $m$  sorteos aleatorios independientes.

Ahora, en muestreo con reemplazo la probabilidad de selección de un elemento no es lo mismo que la probabilidad de inclusión<sup>3</sup> del mismo. Se tienen los siguientes resultados.

**Definición 2.2.1.** *Bajo un diseño con reemplazo, se define la variable aleatoria  $n_k(S)$  como el número de veces que el elemento  $k$ -ésimo es seleccionado en la muestra aleatoria  $S$ .*

**Resultado 2.2.8.** *La variable aleatoria  $n_k(S)$  sigue una distribución binomial tal que*

$$E(n_k(S)) = mp_k, \quad Var(n_k(S)) = mp_k(1 - p_k)$$

*Demostración.* Dado que cada una de las  $m$  extracciones inducen eventos estadísticos independientes, la selección en una extracción particular del  $k$ -ésimo elemento sigue una distribución de Bernoulli, con parámetro  $p_k$ . Como se trata de  $m$  extracciones,  $n_k(S)$  sigue una distribución binomial y puede tomar los valores  $0, 1, \dots, m$ ; al definir éxito como la selección del elemento  $k$ -ésimo en la muestra, entonces se tiene la demostración del resultado.  $\square$

**Definición 2.2.2.** *De manera general, un diseño de muestreo con reemplazo se define como*

$$p(s) = \begin{cases} \frac{m!}{n_1(s)! \dots n_N(s)!} \prod_U (p_k)^{n_k(s)} & \text{si } \sum_U n_k(s) = m \\ 0 & \text{en otro caso} \end{cases} \quad (2.2.19)$$

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

---

<sup>3</sup>Nótese que la probabilidad de inclusión se refiere a la probabilidad de que el elemento sea seleccionado al menos una vez en la muestra.

Nótese la diferencia (y a la vez similitud) de la variable  $n_k(S)$  con la variable  $I_k(S)$ , además por la definición anterior se tiene que el diseño de muestreo con reemplazo sigue una distribución multinomial, por lo tanto cumple las condiciones de diseño muestral; es decir,  $\sum_{s \in Q} p(s) = 1$ , donde  $Q$  es el soporte que contiene todas las posibles muestras con reemplazo de tamaño  $m$ . La cardinalidad de  $Q$ , es

$$\#Q = \binom{N+m-1}{m} \quad (2.2.20)$$

**Resultado 2.2.9.** En muestreo con reemplazo, la probabilidad de inclusión de primer orden del elemento  $k$ -ésimo está dada por:

$$\pi_k = 1 - (1 - p_k)^m \quad (2.2.21)$$

*Demostración.* Dado que se trata de eventos independientes los cuales tienen asociada una probabilidad de éxito (éxito equivalente a que el elemento  $k \in s$ )  $p_k$ , entonces cada uno de estos sorteos aleatorios está determinado por una distribución de probabilidad de tipo Bernoulli. Por consiguiente, cuando se realizan  $m$  ensayos independientes, se utiliza la distribución de probabilidad binomial para hallar las probabilidades de inclusión de primer orden de cada uno de los elementos en la población

$$\begin{aligned} \pi_k &= Pr(k \in S) = 1 - Pr(k \notin s) \\ &= 1 - \binom{m}{m} (1 - p_k)^m (p_k)^{m-m} \\ &= 1 - (1 - p_k)^m \end{aligned}$$

□

**Resultado 2.2.10.** En muestreo con reemplazo, las probabilidades de inclusión de segundo orden  $\pi_{kl}$ , están dadas por:

$$\pi_{kl} = 1 - (1 - p_k)^m - (1 - p_l)^m + (1 - p_k - p_l)^m \quad k \neq l = 1 \dots, N \quad (2.2.22)$$

*Demostración.* Para hallar esta probabilidad debemos negar que  $(k \in S \text{ y } l \in s)$ . Esta negación da como resultado  $(k \notin s \text{ ó } l \notin s)$ . Suponga que tenemos dos eventos,  $A = (k \notin s)$  y  $B = (l \notin s)$ ; por tanto,  $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$ . Las probabilidades anteriores se rigen por un modelo binomial, luego:

$$\begin{aligned} \pi_{kl} &= Pr(k \in S \text{ y } l \in s) \\ &= 1 - Pr(k \notin s) - Pr(l \notin s) + Pr(k, l \notin s) \\ &= 1 - (1 - p_k)^m - (1 - p_l)^m + \binom{m}{m} (1 - p_k - p_l)^m (p_k + p_l)^{m-m} \\ &= 1 - (1 - p_k)^m - (1 - p_l)^m + (1 - p_k - p_l)^m \end{aligned}$$

El cuarto sumando en la igualdad anterior se obtiene considerando que cada ensayo se toma como un proceso Bernoulli, donde el éxito es *no escoger ni a k ni a l*. Por tanto

$$\begin{aligned} Pr(\text{Éxito}) &= 1 - Pr(\text{Fracaso}) \\ &= 1 - Pr(\text{Escoger a } k) - Pr(\text{Escoger a } l) + Pr(\text{Escoger a ambos}) \\ &= 1 - p_k - p_l \end{aligned}$$

Puesto que se trata de un sólo ensayo, la probabilidad de escoger a ambos es nula.

□

Esto se nota más claramente con el típico ejemplo del dado. Si el evento es el lanzamiento de un dado y el éxito es *no sacar 3 o 5*, entonces la probabilidad de obtener éxito será:  $1 - Pr(\text{Fracaso})$ , es decir  $1 - Pr(\text{Sale 5}) - Pr(\text{Sale 1}) + Pr(\text{Sale 5 y 1})$ . Es obvio que el último sumando es cero dado que se trata de un sólo lanzamiento.

**Ejemplo 2.2.2.** El lector no debe confundir el concepto de **muestra con reemplazo** con el concepto de **extracción ordenada**. En nuestra población ejemplo el tamaño poblacional es  $N = 5$ . Si se utiliza un diseño de muestreo que induzca muestras de tamaño fijo igual a  $m = 2$ , 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 con reemplazo. Este escenario es evidenciado fácilmente con la ayuda de la variable aleatoria  $n_k(S)$ . Las posibles extracciones ordenadas están dadas de la siguiente manera.

(1,1)	(2,1)	(3,1)	(4,1)	(5,1)
(1,2)	(2,2)	(3,2)	(4,2)	(5,2)
(1,3)	(2,3)	(3,3)	(4,3)	(5,3)
(1,4)	(2,4)	(3,4)	(4,4)	(5,4)
(1,5)	(2,5)	(3,5)	(4,5)	(5,5)

Sin embargo, aunque todas las posibles extracciones ordenadas no constituyen el soporte de muestreo, éstas si ayudan a definirlo. De hecho, el primer paso para la construcción del soporte de muestreo con reemplazo es la determinación de todas las posibles extracciones. La función `OrderWR`<sup>4</sup> del paquete `TeachingSampling` permite conocer todas las posibles extracciones de tamaño fijo para un diseño de muestreo con reemplazo.

Esta función cuenta con tres argumentos: el primer argumento correspondiente al tamaño de la población  $N$ , el segundo, correspondiente al tamaño de las selecciones,  $m$ , que no necesariamente debe ser menor que el tamaño poblacional<sup>5</sup> y, el último corresponde a una característica  $ID$  que puede ser un conjunto de rótulos o cualquier otro tipo de identificador continuo. El resultado de la función `OrderWR` será un conjunto de todas las posibles extracciones ordenadas con tamaño fijo  $m$ . Cuando el argumento  $ID$  es distinto de `FALSE`, la salida de la función corresponderá al rótulo o identificador continuo para cada elemento de la población. En el siguiente ejemplo se utiliza esta función en nuestra población ejemplo  $U$ .

```
N <- length(U)
N

## [1] 5

m <- 2

OrderWR(N, m, ID = FALSE)

##      [,1] [,2]
## [1,]    1    1
## [2,]    1    2
## [3,]    1    3
## [4,]    1    4
## [5,]    1    5
## [6,]    2    1
## [7,]    2    2
## [8,]    2    3
## [9,]    2    4
## [10,]   2    5
```

<sup>4</sup>El autor desea recalcar que el resultado de esta función no corresponde al soporte de muestreo con reemplazo sino al conjunto de todas las posibles extracciones ordenadas con reemplazo y de tamaño fijo.

<sup>5</sup>Se enfatiza que para este tipo de diseños de muestreo con reemplazo es posible que el tamaño de muestra sea mayor al tamaño poblacional.

```

## [11,] 3 1
## [12,] 3 2
## [13,] 3 3
## [14,] 3 4
## [15,] 3 5
## [16,] 4 1
## [17,] 4 2
## [18,] 4 3
## [19,] 4 4
## [20,] 4 5
## [21,] 5 1
## [22,] 5 2
## [23,] 5 3
## [24,] 5 4
## [25,] 5 5

OrderWR(N, m, ID = U)

##      [,1]    [,2]
## [1,] "Yves" "Yves"
## [2,] "Yves" "Ken"
## [3,] "Yves" "Erik"
## [4,] "Yves" "Sharon"
## [5,] "Yves" "Leslie"
## [6,] "Ken"   "Yves"
## [7,] "Ken"   "Ken"
## [8,] "Ken"   "Erik"
## [9,] "Ken"   "Sharon"
## [10,] "Ken"  "Leslie"
## [11,] "Erik" "Yves"
## [12,] "Erik" "Ken"
## [13,] "Erik" "Erik"
## [14,] "Erik" "Sharon"
## [15,] "Erik" "Leslie"
## [16,] "Sharon" "Yves"
## [17,] "Sharon" "Ken"
## [18,] "Sharon" "Erik"
## [19,] "Sharon" "Sharon"
## [20,] "Sharon" "Leslie"
## [21,] "Leslie" "Yves"
## [22,] "Leslie" "Ken"
## [23,] "Leslie" "Erik"
## [24,] "Leslie" "Sharon"
## [25,] "Leslie" "Leslie"

```

Nótese que el conjunto de extracciones ordenadas contiene al soporte de muestreo con reemplazo. Sin embargo, con ayuda de la función `SupportWR` del paquete `TeachingSampling` se define el verdadero soporte inducido por el diseño de muestreo con reemplazo. Los argumentos de esta función son los mismos tres de la función `OrderWR`: `N`, `m` y `ID`. El resultado de la función es el conjunto de todas las posibles muestras con reemplazo de tamaño fijo. Para este ejemplo particular, el soporte está dado por las siguientes muestras y no por todas las posibles extracciones ordenadas.

```
SupportWR(N, m, ID=FALSE)
```

```
##      [,1] [,2]
## [1,]    1    1
## [2,]    1    2
## [3,]    1    3
## [4,]    1    4
## [5,]    1    5
## [6,]    2    2
## [7,]    2    3
## [8,]    2    4
## [9,]    2    5
## [10,]   3    3
## [11,]   3    4
## [12,]   3    5
## [13,]   4    4
## [14,]   4    5
## [15,]   5    5
```

```
SupportWR(N, m, ID=U)
```

```
##      [,1]      [,2]
## [1,] "Yves"   "Yves"
## [2,] "Yves"   "Ken"
## [3,] "Yves"   "Erik"
## [4,] "Yves"   "Sharon"
## [5,] "Yves"   "Leslie"
## [6,] "Ken"    "Ken"
## [7,] "Ken"    "Erik"
## [8,] "Ken"    "Sharon"
## [9,] "Ken"    "Leslie"
## [10,] "Erik"   "Erik"
## [11,] "Erik"   "Sharon"
## [12,] "Erik"   "Leslie"
## [13,] "Sharon" "Sharon"
## [14,] "Sharon" "Leslie"
## [15,] "Leslie" "Leslie"
```

Por supuesto, cada una de las posibles muestras con reemplazo que pertenecen al soporte tiene distintas probabilidades de selección dependiendo de la configuración de las probabilidades de selección individuales para cada elemento,  $p_k$ . Supongamos que cada uno de los cinco elementos de la población tiene probabilidad de selección dadas por

$$p_k = \begin{cases} 1/4, & \text{para } k = \text{Yves, Ken, Leslie,} \\ 1/8, & \text{para } k = \text{Sharon, Erik} \end{cases}$$

Nótese que  $\sum_U p_k = 1$ . Para esta configuración particular, y siguiendo la expresión (2.2.19), las probabilidades de selección  $p(s)$  de las muestras en el soporte y el valor de la variable  $n_k(S)$  estarían dadas por la configuración mostrada en la tabla 2.2, la cual es producida por el siguiente código.

```

pk <- c(0.25, 0.25, 0.125, 0.125, 0.25)
QWR <- SupportWR(N,m, ID=U)
pWR <- p.WR(N, m, pk)
nkWR <- nk(N, m)
SamplesWR <- data.frame(QWR, pWR, nkWR)

```

	1	2	3	n1	n2	n3	n4	n5
1	Yves	Yves	0.06	2.00	0.00	0.00	0.00	0.00
2	Yves	Ken	0.13	1.00	1.00	0.00	0.00	0.00
3	Yves	Erik	0.06	1.00	0.00	1.00	0.00	0.00
4	Yves	Sharon	0.06	1.00	0.00	0.00	1.00	0.00
5	Yves	Leslie	0.13	1.00	0.00	0.00	0.00	1.00
6	Ken	Ken	0.06	0.00	2.00	0.00	0.00	0.00
7	Ken	Erik	0.06	0.00	1.00	1.00	0.00	0.00
8	Ken	Sharon	0.06	0.00	1.00	0.00	1.00	0.00
9	Ken	Leslie	0.13	0.00	1.00	0.00	0.00	1.00
10	Erik	Erik	0.02	0.00	0.00	2.00	0.00	0.00
11	Erik	Sharon	0.03	0.00	0.00	1.00	1.00	0.00
12	Erik	Leslie	0.06	0.00	0.00	1.00	0.00	1.00
13	Sharon	Sharon	0.02	0.00	0.00	0.00	2.00	0.00
14	Sharon	Leslie	0.06	0.00	0.00	0.00	1.00	1.00
15	Leslie	Leslie	0.06	0.00	0.00	0.00	0.00	2.00

Cuadro 2.2: Todas las posibles muestras con reemplazo para el ejercicio

Nótese que la suma de las probabilidades de selección inducidas por el diseño de muestreo es igual a uno y que cada una de ellas es mayor que cero. El lector debe fijarse en que la muestra perteneciente al soporte está dada en términos de  $n_k(S)$ . De esta manera, si se ha seleccionado la séptima muestra dada por 1 0 1 0 0, en realidad, no importa si **Yves** fue seleccionado primero o después que **Erik** y la probabilidad de selección de esta muestra particular es 0.125 pues

$$\begin{aligned}
p(s) &= \frac{2!}{11!0!1!0!0!} \left[ \left(\frac{1}{4}\right)^1 \left(\frac{1}{4}\right)^0 \left(\frac{1}{8}\right)^1 \left(\frac{1}{8}\right)^0 \left(\frac{1}{4}\right)^0 \right] \\
&= 2 \left(\frac{1}{32}\right) = 0.0625
\end{aligned}$$

### Estimador del total poblacional

Hansen, Hurwitz & Madow (1953) proponen un estimador conveniente para el total de una población  $t_y$  cuando el diseño de muestreo es con reemplazo. La lógica que sigue en la construcción de este estimador está dada a continuación. Sea el evento aleatorio:

Seleccionar el elemento  $k$  ( $k \in U$ ) en el  $i$ -ésimo sorteo ( $i = 1, \dots, m$ ).

Este evento define la creación de variables aleatorias, que serán utilizadas más adelante, cuyo comportamiento es posible modelar mediante el siguiente resultado.

**Resultado 2.2.11.** Sean  $U_1, U_2, \dots, U_m$  es una sucesión de variables aleatorias independientes e idénticamente distribuidas con  $E(U_i) = \mu$  y  $\text{Var}(U_i) = \sigma^2$ . Sea  $\bar{U} = \sum_{i=1}^m U_i/m$ . Entonces  $E(\bar{U}) = \mu$ ,

$Var(\bar{U}) = \sigma^2/m$  y un estimador insesgado de  $Var(\bar{U})$  está dado por la siguiente expresión

$$\widehat{Var}(\bar{U}) = \frac{1}{m(m-1)} \sum_{i=1}^m (U_i - \bar{U})^2 \quad (2.2.23)$$

y por consiguiente, un estimador insesgado para  $\sigma^2$  está dado por

$$\hat{\sigma}^2 = \frac{1}{m-1} \sum_{i=1}^m (U_i - \bar{U})^2. \quad (2.2.24)$$

*Demostración.* La esperanza de  $\bar{U}$  es

$$E(\bar{U}) = \frac{1}{m} \sum_{i=1}^m E(U_i) = \mu \quad (2.2.25)$$

La varianza está determinada por

$$Var(\bar{U}) = \frac{1}{m^2} \sum_{i=1}^m Var(U_i) = \sigma^2/m \quad (2.2.26)$$

Nótese que los términos de covarianza son nulos puesto que las variables son independientes entre ellas.  
Ahora como

$$\sum_{i=1}^m (U_i - \bar{U})^2 = \sum_{i=1}^m U_i^2 - m\bar{U}^2 \quad (2.2.27)$$

entonces,

$$E\left(\sum_{i=1}^m (U_i - \bar{U})^2\right) = \sum_{i=1}^m E(U_i^2) - mE(\bar{U}^2) \quad (2.2.28)$$

Por otro lado

$$\begin{aligned} E(U_i^2) &= Var(U_i) + [E(U_i)]^2 = \sigma^2 + \mu^2 \\ E(\bar{U}^2) &= Var(\bar{U}) + [E(\bar{U})]^2 = \sigma^2/m + \mu^2 \end{aligned}$$

Esto conduce a la demostración del teorema puesto que

$$E\left(\sum_{i=1}^m (U_i - \bar{U})^2\right) = (m-1)\sigma^2 \quad (2.2.29)$$

□

El anterior es un resultado muy potente que puede ser utilizado para cualquier tipo de variables aleatorias que sean independientes e idénticamente distribuidas y será la base para la demostración de resultados en la estimación de parámetros que utilicen diseños de muestreo con reemplazo. Siguiendo con el marco teórico del muestreo con reemplazo tenemos la siguiente definición.

**Definición 2.2.3.** Se define la variable aleatoria  $Z_i$  tal que

$$Z_i = y_{k_i}/p_{k_i} \quad k \in U \quad i = 1, \dots, m \quad (2.2.30)$$

donde la cantidad  $y_{k_i}$  es el valor de la característica de interés del  $k$ -ésimo elemento seleccionado en la  $i$ -ésima extracción. Análogamente,  $p_{k_i}$  es el valor de la probabilidad de selección del  $k$ -ésimo elemento seleccionado en la  $i$ -ésima extracción.

**Resultado 2.2.12.** La distribución de la variable aleatoria  $Z_i$  está dada por

$$Pr\left(Z_i = \frac{y_k}{p_k}\right) = p_k, \quad (2.2.31)$$

por tanto la esperanza y varianza de la variable aleatoria  $Z_i$  son

$$E(Z_i) = t_y \quad (2.2.32)$$

y

$$Var(Z_i) = \sum_U p_k \left( \frac{y_k}{p_k} - t_y \right)^2, \quad (2.2.33)$$

respectivamente.

*Demostración.* Dado que se trata de  $m$  sorteos aleatorios independientes, la variable aleatoria  $Z_i$  puede tomar los siguientes valores

$$\frac{y_1}{p_1}, \frac{y_2}{p_2}, \dots, \frac{y_N}{p_N}$$

con probabilidades

$$p_1, p_2, \dots, p_N$$

respectivamente. Luego, acudiendo a la definición genérica del operador esperanza, se tiene

$$E(Z_i) = \sum_U \frac{y_k}{p_k} Pr\left(Z_i = \frac{y_k}{p_k}\right) = \sum_U \frac{y_k}{p_k} p_k = t_y$$

y análogamente se tiene la varianza

$$Var(Z_i) = \sum_U \left( \frac{y_k}{p_k} - E(Z_i) \right)^2 Pr\left(Z_i = \frac{y_k}{p_k}\right) = \sum_U \left( \frac{y_k}{p_k} - t_y \right)^2 p_k$$

□

Dado que las  $m$  extracciones son eventos independientes, también lo son las variables  $Z_i$ <sup>6</sup>. Nótese que la cantidad  $Z_i$  es una estimación del total poblacional con la  $i$ -ésima muestra seleccionada de tamaño 1. Ahora, como existen  $m$  sorteos habrán  $m$  estimaciones del total poblacional; por tanto, como en mucho otros procedimientos estadísticos utilizamos el promedio de estas  $m$  estimaciones para obtener una estimación unificada para  $t_y$ . El estimador de Hansen-Hurwitz toma la siguiente forma

$$\hat{t}_{y,p} = \frac{1}{m} \sum_{i=1}^m \frac{y_{k_i}}{p_{k_i}} \quad (2.2.34)$$

Para tener una estrategia de muestreo que resulte eficiente en la estimación de  $t_y$ , es conveniente utilizar el estimador de Hansen-Hurwitz, cuando las probabilidades de selección son proporcionales a la característica de interés; esto es, cuando  $p_k \propto y_k$ . Si lo anterior sucede, el estimador tendrá una varianza casi nula y la estimación será muy precisa.

**Resultado 2.2.13.** Si  $p_k > 0$ , para todo  $k \in U$ , el estimador  $\hat{t}_{y,p}$  es insesgado

---

<sup>6</sup> $Z_1, \dots, Z_m$  define una sucesión de variables aleatorias independientes e idénticamente distribuidas, o si se quiere, en términos de la inferencia clásica, define una **muestra aleatoria**.

*Demostración.* Las variables aleatorias  $Z_i$  son independientes (porque cada ensayo es independiente) y su distribución está inducida por  $Pr(Z_i = y_k/p_k) = p_k$ ,  $k \in U$ ; es decir, son idénticamente distribuidas. Por tanto, el estimador de Hansen-Hurwitz puede escribirse como:

$$\hat{t}_{y,p} = \frac{1}{m} \sum_{i=1}^m \frac{y_i}{p_i} = \frac{1}{m} \sum_{i=1}^m Z_i = \bar{Z}$$

y así con  $p_k > 0$  para todo  $k \in U$ , tenemos

$$E(\hat{t}_{y,p}) = \frac{1}{m} \sum_{i=1}^m E(Z_i) = \frac{1}{m} \sum_{i=1}^m t_y = t_y$$

□

### Varianza del estimador de Hansen-Hurwitz

Una de las características más importantes del estimador de Hansen-Hurwitz es la sencillez de la expresión de su varianza. Esta misma hace que aunque el muestreo sea con reemplazo, el estimador de Hansen-Hurwitz sea utilizado de manera frecuente por los usuarios de los estudios por muestreo.

**Resultado 2.2.14.** La varianza del estimador de Hansen-Hurwitz está dada por la siguiente expresión

$$Var(\hat{t}_{y,p}) = \frac{1}{m} \sum_{k=1}^N p_k \left( \frac{y_k}{p_k} - t_y \right)^2 \quad (2.2.35)$$

*Demostración.* Por la independencia de las selecciones se tiene que

$$\begin{aligned} Var(\hat{t}_{y,p}) &= Var\left(\frac{1}{m} \sum_{i=1}^m Z_i\right) \\ &= \frac{1}{m^2} \sum_{i=1}^m Var(Z_i) \\ &= \frac{1}{m} Var(Z_i) \\ &= \frac{1}{m} \sum_U \left( \frac{y_k}{p_k} - t_y \right)^2 p_k \end{aligned}$$

□

La anterior expresión hace que el cálculo computacional de la varianza del estimador de Hansen-Hurwitz sea muy sencillo. Sin embargo, esta varianza se puede escribir de varias formas, algunas de ellas muy útiles para el desarrollo teórico de las propiedades del estimador.

**Resultado 2.2.15.** De manera general, la varianza del estimador de Hansen-Hurwitz se puede escribir de la siguiente manera

$$Var(\hat{t}_{y,p}) = \frac{1}{m} \left( \sum_{k=1}^N \frac{y_k^2}{p_k} - t_y^2 \right) \quad (2.2.36)$$

*Demostración.*

$$\begin{aligned}
 Var(\hat{t}_{y,p}) &= \frac{1}{m} \sum_{k=1}^N p_k \left( \frac{y_k}{p_k} - t_y \right)^2 \\
 &= \frac{1}{m} \sum_{k=1}^N p_k \left( \frac{y_k^2}{p_k^2} - 2t_y \frac{y_k}{p_k} + t_y^2 \right) \\
 &= \frac{1}{m} \sum_{k=1}^N \left( \frac{y_k^2}{p_k} - 2t_y y_k + p_k t_y^2 \right) \\
 &= \frac{1}{m} \left( \sum_{k=1}^N \frac{y_k^2}{p_k} - 2t_y \sum_{k=1}^N y_k + t_y^2 \sum_{k=1}^N p_k \right) \\
 &= \frac{1}{m} \left( \sum_{k=1}^N \frac{y_k^2}{p_k} - 2t_y^2 + t_y^2 \right) = \frac{1}{m} \left( \sum_{k=1}^N \frac{y_k^2}{p_k} - t_y^2 \right)
 \end{aligned}$$

□

### Estimación de la varianza

**Resultado 2.2.16.** Un estimador insesgado de la expresión (2.2.35) es

$$\widehat{Var}(\hat{t}_{y,p}) = \frac{1}{m(m-1)} \sum_{i=1}^m \left( \frac{y_i}{p_i} - \hat{t}_{y,p} \right)^2 \quad (2.2.37)$$

*Demostración.* Al desarrollar la varianza del estimador llegamos a que ésta es igual a

$$\frac{1}{m} Var(Z_i).$$

Ahora, utilizando el resultado 2.2.11, como  $Z_1, \dots, Z_m$  conforman una muestra aleatoria de variables con esperanza  $t_y$  e idéntica varianza, entonces un estimador natural e insesgado para la varianza de  $Z_i$  es

$$\frac{1}{m-1} \sum_{i=1}^m (Z_i - \bar{Z})^2 = \frac{1}{m-1} \sum_{i=1}^m \left( \frac{y_i}{p_i} - \hat{t}_{y,p} \right)^2$$

por tanto, un estimador insesgado de la varianza del estimador de Hansen-Hurwitz será

$$\widehat{Var}(\hat{t}_{y,p}) = \frac{1}{m} \frac{1}{m-1} \sum_{i=1}^m \left( \frac{y_{k_i}}{p_{k_i}} - \hat{t}_{y,p} \right)^2$$

□

**Resultado 2.2.17.** Una expresión alternativa para la estimación de la varianza del estimador de Hansen-Hurwitz en muestreo con reemplazo es

$$\widehat{Var}(\hat{t}_{y,p}) = \frac{1}{m(m-1)} \sum_{i=1}^m \left( \frac{y_{k_i}}{p_{k_i}} \right)^2 - m\hat{t}_{y,p}^2$$

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

$$\begin{aligned}
 m(m-1)\widehat{Var}(\hat{t}_{y,p}) &= \sum_{i=1}^m \left( \frac{y_{k_i}}{p_{k_i}} - \hat{t}_{y,p} \right)^2 \\
 &= \sum_{i=1}^m \left( \frac{y_{k_i}^2}{p_{k_i}^2} - 2\hat{t}_{y,p} \frac{y_{k_i}}{p_{k_i}} + \hat{t}_{y,p}^2 \right) \\
 &= \sum_{i=1}^m \left( \frac{y_{k_i}^2}{p_{k_i}^2} \right) - 2\hat{t}_{y,p} \sum_{i=1}^m \frac{y_{k_i}}{p_{k_i}} + m\hat{t}_{y,p} \\
 &= \sum_{i=1}^m \left( \frac{y_{k_i}^2}{p_{k_i}^2} \right) - 2m\hat{t}_{y,p}^2 + m\hat{t}_{y,p} \\
 &= \sum_{i=1}^m \left( \frac{y_{k_i}}{p_{k_i}} \right)^2 - m\hat{t}_{y,p}^2
 \end{aligned}$$

□

Aunque el diseño muestral sea con reemplazo, es posible utilizar el estimador de Horvitz-Thompson, pues conserva su insesgamiento. La comparación entre la precisión del estimador de Horvitz-Thompson y el estimador de Hansen-Hurwitz, en un diseño con repetición depende de la configuración de los valores de la característica de interés en la población  $y_k \forall k = 1, 2, \dots, N$ . Sin embargo, generalmente el estimador de Horvitz-Thompson es más eficiente más eficiente que el estimador de Hansen-Hurwitz, aunque éste último es más fácil de calcular. Cuando el diseño de muestreo es de tamaño fijo, el estimador de Horvitz-Thompson y Hansen-Hurwitz coinciden.

**Ejemplo 2.2.3.** Continuando con el ejercicio léxico-gráfico de la estimación del total poblacional  $t_y$  para todas las posibles muestras con reemplazo de tamaño 2 de la población U, tenemos la siguiente tabla que da cuenta del soporte de muestreo con ayuda de la función `SupportWR`

```

all.y <- SupportWR(N, n, y)
all.pk <- SupportWR(N, n, pk)
all.HH <- rep(0, 15)

for(k in 1:15){
  all.HH[k] <- HH(all.y[k,], all.pk[k,])
}

AllSamplesWR <- data.frame(QWR, all.pk, pWR, all.y, all.HH)

```

El vector `Est` contiene las estimaciones de Hansen-Hurwitz para cada una de las posibles 15 muestras con reemplazo, su esperanza se calcula como

```

sum(all.HH * pWR)

## [1] 236

```

Nótese que la esperanza del estimador equivale al total de la característica de interés, corroborando su insesgamiento. Por otro lado, para seleccionar una muestra con reemplazo, R incorpora la función `sample`, cuyos principales argumentos son

`x, size, replace, prob.`

	1	2	3	4	5	6	7	8
1	Yves	Yves	0.25	0.25	0.06	32.00	32.00	128.00
2	Yves	Ken	0.25	0.25	0.13	32.00	34.00	132.00
3	Yves	Erik	0.25	0.12	0.06	32.00	46.00	248.00
4	Yves	Sharon	0.25	0.12	0.06	32.00	89.00	420.00
5	Yves	Leslie	0.25	0.25	0.13	32.00	35.00	134.00
6	Ken	Ken	0.25	0.25	0.06	34.00	34.00	136.00
7	Ken	Erik	0.25	0.12	0.06	34.00	46.00	252.00
8	Ken	Sharon	0.25	0.12	0.06	34.00	89.00	424.00
9	Ken	Leslie	0.25	0.25	0.13	34.00	35.00	138.00
10	Erik	Erik	0.12	0.12	0.02	46.00	46.00	368.00
11	Erik	Sharon	0.12	0.12	0.03	46.00	89.00	540.00
12	Erik	Leslie	0.12	0.25	0.06	46.00	35.00	254.00
13	Sharon	Sharon	0.12	0.12	0.02	89.00	89.00	712.00
14	Sharon	Leslie	0.12	0.25	0.06	89.00	35.00	426.00
15	Leslie	Leslie	0.25	0.25	0.06	35.00	35.00	140.00

Cuadro 2.3: *Estimaciones de Hansen-Hurwitz para todas las posibles muestras del ejemplo*

`x` es el tamaño de la población, `size` es un número entero que determina el tamaño de la muestra. Para seleccionar una muestra con reemplazo, el argumento `replace` debe tomar el valor `TRUE`, así `replace = TRUE`. Cada elemento perteneciente a la población debe tener asociado un vector de probabilidades de selección cuya suma sea igual a la unidad. En R, el argumento `prob` contiene este vector de probabilidades; cuando se omite este argumento, la función `sample` asume que las probabilidades de selección son idénticas para cada individuo en la población. Así, por ejemplo, para seleccionar una muestra con reemplazo del marco de muestreo de  $U$  de tamaño  $m = 3$ , con las probabilidades de selección dadas por

```
pk
## [1] 0.250 0.250 0.125 0.125 0.250
```

Nótese que la suma de las probabilidades de selección es igual a uno y que los rótulos o nombres para cada individuo en la población están contenidos en el objeto `U`.

```
U
## [1] "Yves"    "Ken"     "Erik"    "Sharon"  "Leslie"
```

Para seleccionar una muestra con reemplazo de tamaño  $m = 3$  se debe escribir el siguiente código

```
sam <- sample(N, 3, replace=TRUE, prob = pk)
sam
## [1] 2 4 3
```

Para la selección anterior, fue escogido dos veces el primer elemento y una vez el tercer elemento. La indexación de los rótulos (nombres) y valores de la característica de interés de los elementos escogidos en la muestra se hace utilizando

```

pkm <- pk[sam]
pkm

## [1] 0.250 0.125 0.125

ym <- y[sam]
ym

## [1] 34 89 46

```

Nótese que el tamaño de muestra es 3, pero el tamaño efectivo de muestra es  $n(S) = 2$ . Siendo `pkm` el vector de probabilidades de selección para los individuos pertenecientes a la muestra y `ym` el vector de valores de la característica de interés para los individuos pertenecientes a la muestra. La función `HH` del paquete `TeachingSampling` realiza la estimación del total poblacional para la característica de interés. Esta función consta de dos argumentos: `y`, el vector de valores de la característica de interés de los individuos en la muestra y `pk` sus correspondientes probabilidades de selección.

```

est <- HH(ym, pkm)[1]
est

## [1] 405.3

```

Para realizar la estimación de la varianza se crea un vector de diferencias `dif` entre  $\frac{y_i}{p_i}$  y la estimación. Luego se procede a elevarlo al cuadrado, sumarlo y dividir por  $m(m - 1)$ .

```

dif <- rep(0, 3)
dif[1] <- (ym[1] / pk[sam][1]) - est
dif[2] <- (ym[2] / pk[sam][2]) - est
dif[3] <- (ym[3] / pk[sam][3]) - est

dif

## [1] -269.33 306.67 -37.33

Var <- (1 / 3) * (1 / 2) * sum(dif^2)
Var

## [1] 27996

sqrt(Var)

## [1] 167.3

```

Luego, el respectivo coeficiente de variación estimado es

$$cve(\hat{t}_p) = \frac{167.3214}{405.3333} \cong 41\%$$

Nótese que utilizando la función `HH`, el resultado que arroja el procedimiento es el mismo.

```
HH(ym, pkm)

##          y
## Estimation 405.33
## Standard Error 167.32
## CVE        41.28
```

Podemos pensar en el coeficiente de variación estimado como una medida de precisión. Así, las anteriores estimaciones se podrían decir inaceptables porque esta medida es muy alta.

El objetivo de este libro es que el lector esté en la capacidad de proponer estrategias de muestreo que permitan estimaciones precisas y confiables. Es decir, estimaciones cuyo coeficiente de variación sea aceptable<sup>7</sup> cuya longitud del intervalo de confianza sea corta con un nivel de confianza satisfactorio.

### 2.2.3 El estimador de Horvitz-Thompson en los diseños con reemplazo

## 2.3 Muestras representativas

La teoría de muestreo se ha visto enriquecida en las últimas décadas por valiosos aportes a nivel mundial; aunque la base de la teoría de muestreo es la teoría de probabilidad, cuyo desarrollo axiomático cuenta varios centenares de años, su desarrollo práctico no sucedió sino hasta comienzos del siglo XX. Sin embargo, en la teoría clásica de inferencia estadística, basados en el pensamiento de Ronald Fisher y otros, asumen que la población es infinita. Un aspecto fundamental de la teoría de muestreo es que está basada en la realidad, en donde las poblaciones por más grandes que sean son de naturaleza finita.

Partiendo de este hecho es posible fundamentar la inferencia basada en una muestra aleatoria pero que proviene de una población finita y desde esta perspectiva los resultados de las inferencias diferirán de una manera significativa. De hecho, el llamado de atención es para que las personas que hacen inferencia con datos provenientes de un estudio por muestreo, se actualicen y no cometan grandes equivocaciones a la hora de presentar los resultados de la inferencia (Chambers & Skinner 2003). Por eso la teoría de muestreo cubre aspectos fundamentales de la estadística, porque desde un experimento controlado, hasta una encuesta por muestreo (Survey sampling), se debe pensar en el mecanismo de recolección de la información, y desde allí en la inferencia.

Un ejemplo común en las aulas de clase es describir la población en el tablero mediante una carita feliz, el profesor dice que una muestra representativa de la población es aquella muestra en donde se sigue viendo la misma carita feliz. Es decir, existe la creencia que una muestra representativa es un modelo reducido de la población y de aquí se desprende un argumento de validez sobre la muestra: una buena muestra es aquella que se parece a la población, de tal forma que las categorías aparecen con las mismas proporciones que en la población. Nada más falso que esta creencia. En algunos casos es fundamental sobre-representar algunas categorías o incluso seleccionar unidades con probabilidades desiguales.

Tillé (2006) cita el siguiente ejemplo: suponga que el objetivo es estimar la producción de hierro en un país y que nosotros sabemos que el hierro es producido, por dos compañías gigantes con miles de empleados y por cientos de pequeñas compañías con pocos empleados. ¿La mejor forma de seleccionar la muestra consiste en asignar la misma probabilidad a cada compañía? Claro que no. Primero averiguamos la producción de las grandes compañías. Después, seleccionamos una muestra de las compañías pequeñas.

La muestra no debe ser un modelo reducido de la población; debe ser una herramienta usada para obtener estimaciones. Es así como el concepto de muestra representativa pierde peso. Más aún, para

<sup>7</sup>En muchos casos un coeficiente de variación aceptable es menor al 3 por ciento.

Hájek (1981), una estrategia de muestreo es una dupla: diseño de muestreo (distribución de probabilidad sobre todas las posibles muestras) y estimador. La teoría de muestreo se ha ocupado de estudiar estrategias óptimas que permitan asegurar la calidad de las estimaciones. Entonces, el concepto de representatividad debería estar asociado con las estrategias de muestreo y no sólo con las muestras.

Siguiendo con Tillé (2006), una estrategia se dice representativa si permite estimar un total poblacional exactamente; es decir, sin sesgo y con varianza nula. Si se utiliza, por ejemplo, el estimador de Horvitz-Thompson junto con un diseño de muestreo apropiado, esta estrategia es representativa sólo sí, junto con la muestra seleccionada, el estimador reproduce algunos totales de la población; tales muestras se llaman muestras balanceadas. Existen también, estimadores que brindan a la estrategia el calificativo de representativa, algunos de ellos son conocidos como estimadores de calibración.

## 2.4 Ejercicios

2.1 Pruebe que bajo un diseño de muestreo  $p(s)$ , el error cuadrático medio de cualquier estimador  $\hat{T}(s)$  de un parámetro  $T$  es igual a la varianza  $Var(\hat{T})$  más el sesgo al cuadrado  $B^2(\hat{T})$ .

$$\text{Sugerencia: } ECM(\hat{T}) = E_p(\hat{T}(s) - T)^2 = \sum_{s \in Q} (\hat{T}(s) - T)^2 p(s).$$

2.2 Demuestre que  $\pi_{kl} = E_p(I_k(s)I_l(s))$ .

2.3 Suponga que tiene acceso a la población finita de tamaño  $N = 5$  del ejemplo 2.2.1. y asuma el siguiente diseño de muestreo sin reemplazo

$$p(S = s) = \begin{cases} 0.2, & \text{para } s = \{Ken, Erik, Sharon\}, s = \{Ken, Leslie\}, \\ 0.3, & \text{para } s = \{Yves, Erik, Leslie\}, s = \{Yves, Sharon\}, \\ 0, & \text{En otro caso.} \end{cases}$$

- Calcule todas las probabilidades de inclusión de primer y de segundo orden.
- ¿Es el anterior un diseño de muestreo de tamaño de muestra fijo? Explique.
- Enumere todos los valores que toma la variable aleatoria  $n(S)$  y verifique las relaciones  $E_p(n(S)) = \sum_U \pi_k$  y  $Var_p(n(S)) = \sum_U \pi_k - (\sum_U \pi_k)^2 + \sum \sum_{k \neq l} \pi_{kl}$ .

2.4 Suponga que tiene acceso a la población finita de tamaño  $N = 5$  del ejemplo 2.2.1. y asuma el siguiente diseño de muestreo sin reemplazo

$$p(S = s) = \begin{cases} 0.1, & \text{Si } n(S) = 3, \\ 0, & \text{En otro caso.} \end{cases}$$

- Defina todas las posibles muestras que pertenecen al soporte inducido por el anterior diseño de muestreo.
- Calcule todas las probabilidades de inclusión de primer y de segundo orden.
- Verifique que  $\sum_U \pi_k = 3$  y que  $\sum_U \pi_k - (\sum_U \pi_k)^2 + \sum \sum_{k \neq l} \pi_{kl} = 0$ . Explique.
- Verifique que  $\sum_U \pi_{k1} = 3 \times \pi_1$ ,  $\sum_U \pi_{k2} = 3 \times \pi_2$ , hasta  $\sum_U \pi_{k5} = 3 \times \pi_5$ .
- Calcule todas las posibles covarianzas  $\Delta_{kl}$  y verifique que  $\sum_U \Delta_{k1} = 0$ , hasta  $\sum_U \Delta_{k5} = 0$ .

2.5 Demuestre o refute la siguiente afirmación: «Bajo cualquier diseño de muestreo, la suma poblacional de las probabilidades de inclusión de primer orden es siempre igual al tamaño de muestra».

2.6 Demuestre o refute la siguiente afirmación: «Bajo cualquier diseño de muestreo, el estimador de Horvitz-Thompson puede ser utilizado para obtener una estimación insesgada del total poblacional».

2.7 Suponga que tiene acceso a la población finita de tamaño  $N = 5$  del ejemplo 2.2.1 y que  $y_k$  denota el valor de la característica de interés en el  $k$ -ésimo individuo. De esta manera, se tiene que:

$$y_{Yves} = 32, \quad y_{Ken} = 34, \quad y_{Erik} = 46, \quad y_{Sharon} = 89, \quad y_{Leslie} = 35$$

- Para el diseño de muestreo del ejercicio 2.3, en cada una de las posibles muestras calcule la estimación de Horvitz-Thompson, la estimación de la varianza, el *cve* y la estimación del intervalo de confianza al 95 %. Por último, muestre que el estimador es insesgado y calcule la varianza del estimador utilizando la expresión (2.2.4).
  - Para el diseño de muestreo del ejercicio 2.4, en cada una de las posibles muestras calcule la estimación de Horvitz-Thompson, la estimación de la varianza, el *cve* y la estimación del intervalo de confianza al 95 %. Por último, muestre que el estimador es insesgado y calcule la varianza del estimador utilizando la expresión (2.2.4) y (2.2.5). ¿Son iguales estas varianzas? Explique.
  - Para el diseño de muestreo del ejercicio 2.3, en cada una de las posibles muestras calcule la estimación de Horvitz-Thompson de la media (expresión 2.2.10), la estimación del tamaño poblacional (expresión 2.2.14), la estimación alternativa de la media (expresión 2.2.15) y la estimación alternativa del total (expresión 2.2.18).
  - Para el diseño de muestreo del ejercicio 2.4, en cada una de las posibles muestras calcule la estimación de Horvitz-Thompson de la media (expresión 2.2.10), la estimación del tamaño poblacional (expresión 2.2.14), la estimación alternativa de la media (expresión 2.2.15) y la estimación alternativa del total (expresión 2.2.18).
- 2.8 Demuestre o refute la siguiente afirmación: «Bajo cualquier diseño de muestreo con reemplazo, el estimador de Hansen-Hurwitz puede ser utilizado para obtener una estimación insesgada del total poblacional».
- 2.9 Demuestre o refute la siguiente afirmación: «La probabilidad de selección de un individuo es siempre igual a su probabilidad de inclusión».
- 2.10 Demuestre o refute la siguiente afirmación: «Cualquier diseño de muestreo con reemplazo se puede ver como un caso particular de la distribución multinomial».
- 2.11 Demuestre o refute la siguiente afirmación: «Para una población de tamaño  $N$ , el número de posibles muestras con reemplazo de tamaño  $m$  es  $N^m$ ».
- 2.12 Suponga que tiene acceso a la población finita de tamaño  $N = 5$  de los anteriores ejercicios y asuma las siguientes probabilidades de selección
- $$p_k = \begin{cases} 0.3, & \text{para } k = Yves, Leslie, \\ 0.2, & \text{para } k = Erik, \\ 0.1, & \text{para } k = Ken, Sharon. \end{cases}$$
- ¿Cuántas muestras con reemplazo de tamaño  $m = 3$  se pueden seleccionar? Especifique explícitamente el diseño de muestreo para estas muestras y compruebe que  $\sum_{s \in Q} p(s) = 1$ .
  - Para este diseño de muestreo, y teniendo en cuenta los valores de la característica de interés del ejercicio 2.7, en cada una de las posibles muestras calcule la estimación de Hansen-Hurwitz, la estimación de la varianza, el *cve* y la estimación del intervalo de confianza al 95 %. Por último, muestre que el estimador es insesgado y calcule la varianza del estimador utilizando la expresión (2.2.35).
  - ¿Es posible utilizar otro tipo de estimadores para obtener estimaciones insesgadas del total poblacional?
- 2.13 Demuestre rigurosamente que el estimador de la varianza del estimador de Hansen-Hurwitz corresponde a la expresión (2.2.36).



## Capítulo 3

# 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.

### 3.1 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 3.1.1.** *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,*

$$p(s) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{si } \#s = n \\ 0 & \text{en otro caso} \end{cases} \quad (3.1.1)$$

**Resultado 3.1.1.** *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  $\#Q = \binom{N}{n}$ .

#### 3.1.1 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 3.1.1.** 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 3.1.2.** *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).

### 3.1.2 El estimador de Horvitz-Thompson

**Resultado 3.1.3.** *Para un diseño de muestreo aleatorio simple, las probabilidades de inclusión de primer y segundo orden están dadas por:*

$$\pi_k = \frac{n}{N} \tag{3.1.2}$$

$$\pi_{kl} = \frac{n(n-1)}{N(N-1)} \tag{3.1.3}$$

respectivamente. La covarianza de las variables indicadoras está dada por

$$\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k \pi_l = -\frac{n}{N^2} \frac{(N-n)}{(N-1)} & \text{para } k \neq l \\ \pi_k (1 - \pi_k) = \frac{n(N-n)}{N^2} & \text{para } k = l \end{cases} \tag{3.1.4}$$

*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 3.1.4.** 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:

$$\hat{t}_{y,\pi} = \frac{N}{n} \sum_S y_k \quad (3.1.5)$$

$$Var_{MAS}(\hat{t}_{y,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{yU}^2 \quad (3.1.6)$$

$$\widehat{Var}_{MAS}(\hat{t}_{y,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{yS}^2 \quad (3.1.7)$$

respectivamente, con

$$S_{yU}^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{y}_U)^2, \quad (3.1.8)$$

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

$$S_{yS}^2 = \frac{1}{n-1} \sum_{k \in S} (y_k - \bar{y}_S)^2 \quad (3.1.9)$$

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

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} = \frac{N}{n} \sum_S y_k. \quad (3.1.10)$$

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  $S_{yU}^2$ .

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

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 3.1.2.** 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.

### 3.1.3 Estimación de la media poblacional

**Resultado 3.1.5.** 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:

$$\hat{y}_\pi = \frac{\hat{t}_{y,\pi}}{N} = \frac{\sum_S y_k}{n} = \bar{y}_S \quad (3.1.11)$$

$$Var_{MAS}(\hat{y}_\pi) = \frac{1}{N^2} Var(\hat{t}_{y,\pi}) = \left(1 - \frac{n}{N}\right) \frac{S_{yU}^2}{n} \quad (3.1.12)$$

$$\widehat{Var}_{MAS}(\hat{y}_\pi) = \frac{1}{N^2} Var(\hat{t}_{y,\pi}) = \left(1 - \frac{n}{N}\right) \frac{S_{ys}^2}{n} \quad (3.1.13)$$

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:

$$\left[ \bar{y}_S - z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_{yU}^2}{n}}, \bar{y}_S + z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_{yU}^2}{n}} \right] \quad (3.1.14)$$

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  $c$ ,

$$c = z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_{yU}}{\sqrt{n}}} \quad (3.1.15)$$

y despejando  $n$ , se tiene:

$$n \geq \frac{n_0}{1 + \frac{n_0}{N}} \quad (3.1.16)$$

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:

$$c|\bar{y}_U| = z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_{yU}}{\sqrt{n}}} \quad (3.1.17)$$

y despejando  $n$ , se tiene:

$$n \geq \frac{k_0}{1 + \frac{k_0}{N}} \quad (3.1.18)$$

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

$$\bar{y}_S(1 \pm A) \quad (3.1.19)$$

Donde  $A$  está dada por

$$A = z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S_{ys}}{\sqrt{n} \bar{y}_S}} = z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{cv}{\sqrt{n}}} \quad (3.1.20)$$

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.

### 3.1.4 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 3.1.2.** 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

$$t_{yd} = \sum_{U_d} y_k \quad (3.1.21)$$

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

$$z_{dk} = \begin{cases} 1 & \text{si } k \in U_d \\ 0 & \text{en otro caso} \end{cases} \quad (3.1.22)$$

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

$$t_{yd} = \sum_U y_{dk}, \quad (3.1.23)$$

el tamaño del dominio  $U_d$  toma la siguiente expresión

$$N_d = \sum_U z_{dk}, \quad (3.1.24)$$

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

$$\bar{y}_{U_d} = \frac{t_{yd}}{N_d} = \frac{\sum_U y_{dk}}{N_d} \quad (3.1.25)$$

### Estimación del total en un dominio

**Resultado 3.1.7.** 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

$$\hat{t}_{yd,\pi} = \frac{N}{n} \sum_S y_{dk} = \frac{N}{n} \sum_{S_d} y_k \quad (3.1.26)$$

$$Var(\hat{t}_{yd,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{y_d U}^2 \quad (3.1.27)$$

$$\widehat{Var}(\hat{t}_{yd,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{y_d S}^2 \quad (3.1.28)$$

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 3.1.8.** 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

$$\hat{N}_{d,\pi} = \frac{N}{n} \sum_S z_{dk} = \frac{N}{n} \sum_{S_d} z_k \quad (3.1.29)$$

$$Var(\hat{N}_{d,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{z_d U}^2 \quad (3.1.30)$$

$$\widehat{Var}(\hat{N}_{d,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{z_d s}^2 \quad (3.1.31)$$

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 3.1.9.** 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

$$\hat{P}_{d,\pi} = \frac{1}{N} \sum_S \frac{N}{n} z_{dk} = \frac{1}{n} \sum_S z_{dk} = \frac{n_d}{n} \quad (3.1.32)$$

$$Var(\hat{P}_{d,\pi}) = \frac{1}{n} \left(1 - \frac{n}{N}\right) S_{z_d U}^2 \quad (3.1.33)$$

$$\widehat{Var}(\hat{P}_{d,\pi}) = \frac{1}{n} \left(1 - \frac{n}{N}\right) S_{z_d s}^2 \quad (3.1.34)$$

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 3.1.10.** 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

$$\hat{y}_{U_d,\pi} = \frac{\frac{N}{n} \sum_S y_{dk}}{N_d} \quad (3.1.35)$$

$$Var(\hat{y}_{U_d,\pi}) = \frac{1}{N_d^2} \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{y_d U}^2 \quad (3.1.36)$$

$$\widehat{Var}(\hat{y}_{U_d,\pi}) = \frac{1}{N_d^2} \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{y_d s}^2 \quad (3.1.37)$$

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

$$\hat{y}_{S_d} = \frac{\hat{t}_{y_d,\pi}}{\hat{N}_{d,\pi}} = \frac{\sum_S y_{dk}}{z_{dk}} = \frac{\sum_{S_d} y_k}{n_d} \quad (3.1.38)$$

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.

### 3.1.5 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.

---

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

```

data(BigLucy)
attach(BigLucy)

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

mean(Inc.pilot)

## [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 (3.1.16) 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 (3.1.18) 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

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

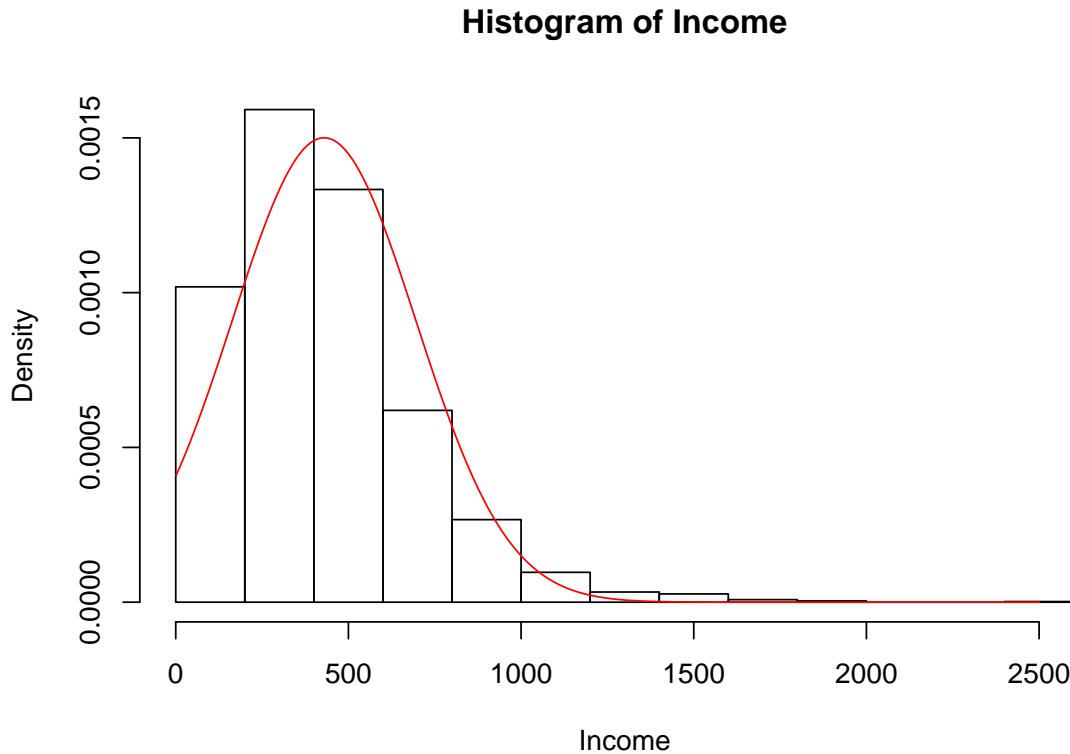


Figura 3.1: Distribución de la característica *Income* y su posible modelamiento bajo la distribución gamma.

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 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 AB00000000012 C0033329K0268568 Small County1    419       20     7
## 89 AB00000000089 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. Obtenremos 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)

```

La tabla 3.1 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.

Cuadro 3.1: *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 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 3.2: *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

### 3.1.6 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  $h$

<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.

Cuadro 3.3: *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

compuesto por  $M < n$  individuos, puede modelarse por medio de la distribución hipergeométrica, así:

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

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

- $M = 1$ ,

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

- $M = 2$ ,

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

- $M = 3$ ,

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

## 3.2 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 3.2.1.** Siendo  $n(s)$  el tamaño de muestra, el diseño de muestreo Bernoulli selecciona la muestra  $s$  con probabilidad

$$p(s) = \begin{cases} \pi^{n(s)}(1 - \pi)^{N-n(s)} & \text{si } s \text{ tiene tamaño igual a } n(s) \\ 0 & \text{en otro caso} \end{cases} \quad (3.2.1)$$

### 3.2.1 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  $\varepsilon_k \sim \text{Unif}[0, 1]$ , 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  $\text{Ber}(\pi)$  y se tiene el siguiente resultado.

**Resultado 3.2.1.** 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 3.2.2.** Bajo muestreo Bernoulli, la distribución del tamaño de muestra  $n(S)$  es binomial  $\text{Bin}(N, \pi)$  y

$$\Pr(n(S) = r) = \sum_{s \in Q_r} p(s) = \binom{N}{r} \pi^r (1 - \pi)^{N-r}, \quad (3.2.2)$$

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  $\text{Ber}(\pi)$ , 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 (3.2.1), 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:

$$E(n(S)) = N\pi \quad \text{Var}(n(S)) = N(\pi)(1 - \pi), \quad (3.2.3)$$

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

$$\sum_{s \in Q} p(s) = 1 \quad (3.2.4)$$

*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

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

$\square$

**Resultado 3.2.4.** *Para el diseño de muestreo Bernoulli, las probabilidades de inclusión de primer y segundo orden están dadas por:*

$$\pi_k = \pi \quad (3.2.5)$$

$$\pi_{kl} = \begin{cases} \pi & \text{para } k = l \\ \pi^2 & \text{Para } k \neq l \end{cases} \quad (3.2.6)$$

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

$k$ -ésimo, tenemos

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

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

$$\sum_{r=0}^m \binom{m}{r} a^r b^{m-r} = (a+b)^m. \quad (3.2.7)$$

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

$$Pr(k \in S \text{ y } l \in S) = Pr(I_k = 1)Pr(I_l = 1) = \pi^2 \quad (3.2.8)$$

□

### 3.2.2 El estimador de Horvitz-Thompson

**Resultado 3.2.5.** Para el diseño de muestreo Bernoulli, el estimador de Horvitz-Thompson, su varianza y su varianza estimada están dados por:

$$\hat{t}_{y,\pi} = \frac{1}{\pi} \sum_S y_k \quad (3.2.9)$$

$$Var_{BER}(\hat{t}_{y,\pi}) = \left( \frac{1}{\pi} - 1 \right) \sum_U y_k^2 \quad (3.2.10)$$

$$\widehat{Var}_{BER}(\hat{t}_{y,\pi}) = \frac{1}{\pi} \left( \frac{1}{\pi} - 1 \right) \sum_S y_k^2, \quad (3.2.11)$$

respectivamente

*Demostración.* El resultado es inmediato porque

$$\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k \pi_l = \pi^2 - \pi^2 = 0 & \text{para } k \neq l \\ \pi_{kk} - \pi_k \pi_k = \pi(1-\pi) & \text{para } k = l \end{cases} \quad (3.2.12)$$

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. □

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

$$\hat{t}_{y,\pi} = \frac{1}{\pi} \sum_U y_k = \frac{t_y}{\pi} \neq t_y \quad (3.2.13)$$

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

$$\hat{t}_{y,alt} = N\bar{y}_S = N \frac{\sum_S y_k}{n(S)} = N\bar{y}_S. \quad (3.2.14)$$

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

**Ejemplo 3.2.1.** 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.

### 3.2.3 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:

**Definición 3.2.2.** 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

$$Defeff = \frac{Var_p(\hat{T})}{Var_{MAS}\hat{T}_\pi}. \quad (3.2.15)$$

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

$$Defeff = \frac{Var_p(\hat{t}_{y,\pi})}{\frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{yU}^2}. \quad (3.2.16)$$

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

---

<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  $\pi^N$ .

**Resultado 3.2.6.** Un estimador del efecto de diseño  $D\hat{e}ff$  para el total poblacional  $t_y$  es

$$D\hat{e}ff = \frac{\widehat{Var}_p(\hat{T})}{\frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{ys}^2}. \quad (3.2.17)$$

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

$$deff = \frac{Var_{BER}(\hat{t}_{y,\pi})}{Var_{MAS}(\hat{t}_{y,\pi})} = 1 - \frac{1}{N} + \frac{1}{CV_y^2} \cong 1 + \frac{1}{CV_y^2} \quad (3.2.18)$$

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.

### 3.2.4 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 conteniendo 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 3.4: *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 3.4 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)
```

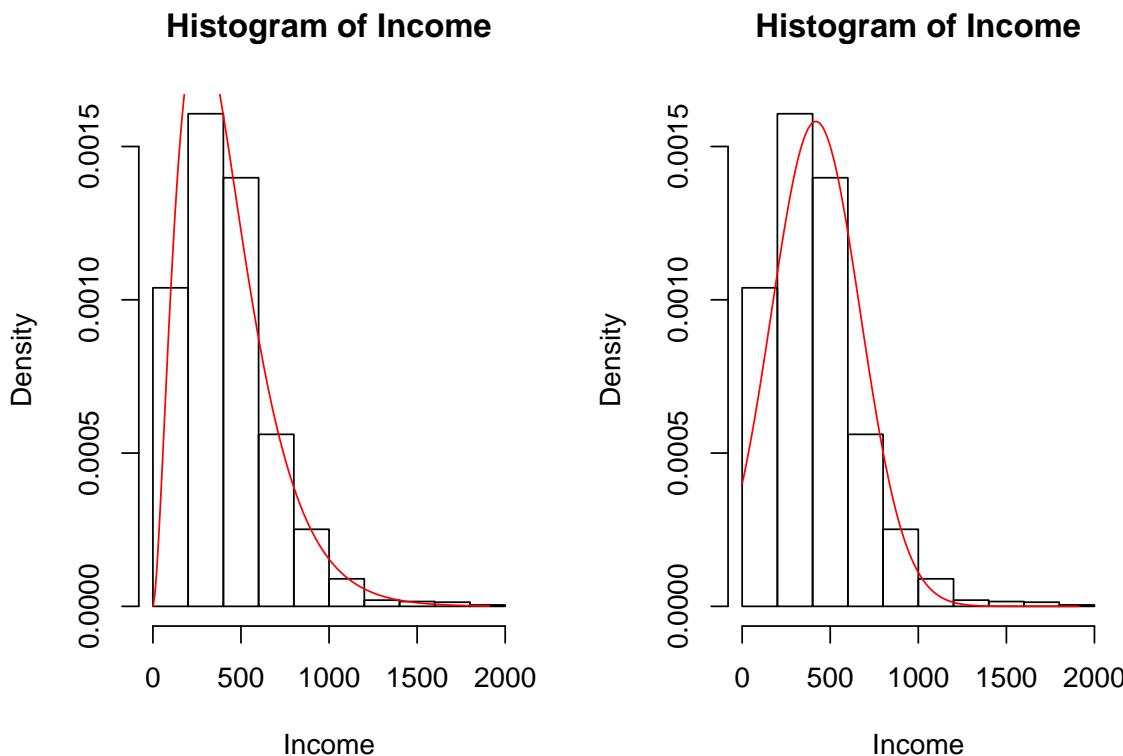


Figura 3.2: Distribución de la característica *Income* y su posible modelamiento bajo la distribución gamma (izquierda) y norma (derecha).

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.

### 3.3 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 3.3.1.** Un diseño de muestreo aleatorio simple con reemplazo se define como

$$p(s) = \begin{cases} \frac{m!}{n_1(s)! \dots n_N(s)!} \prod_U \left(\frac{1}{N}\right)^{n_k(s)} & \text{si } \sum_U n_k(s) = m \\ 0 & \text{en otro caso} \end{cases} \quad (3.3.1)$$

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

**Resultado 3.3.1.** 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 3.3.2.** 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 muestro es una función de densidad multinomial discreta sobre  $Q$ .

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

donde se utiliza el resultado del teorema multinomial que afirma que

$$\sum_{\substack{n_1 \dots n_N \\ \sum_U n_k=S}} \frac{m!}{n_1! \dots n_N!} (p_1)^{n_1} \dots (p_N)^{n_N} = \left( \sum_{k=1}^N p_k \right)^m \quad (3.3.2)$$

□

**Resultado 3.3.3.** Para un diseño aleatorio simple con reemplazo, las probabilidades de inclusión de primer y segundo orden están dadas por

$$\pi_k = 1 - \left(1 - \frac{1}{N}\right)^m \quad (3.3.3)$$

$$\pi_{kl} = 1 - 2\left(1 - \frac{1}{N}\right)^m + \left(1 - \frac{2}{N}\right)^m \quad (3.3.4)$$

respectivamente.

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

**Ejemplo 3.3.1.** 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.

### 3.3.1 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.

$$Bin \left( m - \sum_{i=1}^{k-1} n_i, \frac{1}{N-k+1} \right) \quad (3.3.5)$$

Para todo  $k \in U$ .

**Ejemplo 3.3.2.** 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  $\text{Bin } n$  y  $\text{Bin } p$  son los parámetros de la distribución binomial asociada al algoritmo secuencial. Note que la cantidad  $n_k$  se refiere a la realización de la variable  $n_k(s)$ .

### 3.3.2 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 (2.2.34) que estima de manera insesgada al parámetro de interés  $t_y$ .

**Resultado 3.3.4.** 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:

$$\hat{t}_{y,p} = \frac{N}{m} \sum_{i=1}^m y_i \quad (3.3.6)$$

$$Var_{MRAS}(\hat{t}_{y,p}) = N \frac{(N-1)}{m} S_{yU}^2 \quad (3.3.7)$$

$$\widehat{Var}_{MRAS}(\hat{t}_{y,p}) = \frac{N^2}{m} S_{ysr}^2 \quad (3.3.8)$$

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,

$$\hat{t}_{y,p} = \frac{1}{m} \sum_U n_k(S) \frac{y_k}{p_k} = \frac{N}{m} \sum_U n_k(S) y_k \quad (3.3.9)$$

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

$$\widehat{Var}(\hat{t}_{y,p}) = \frac{1}{m} \frac{1}{m-1} \sum_U n_k(S) (Ny_k - \hat{t}_{y,p})^2 \quad (3.3.10)$$

se tiene el insesgamiento dado por

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

□

**Ejemplo 3.3.3.** 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.

### 3.3.3 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 3.5: *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)
```

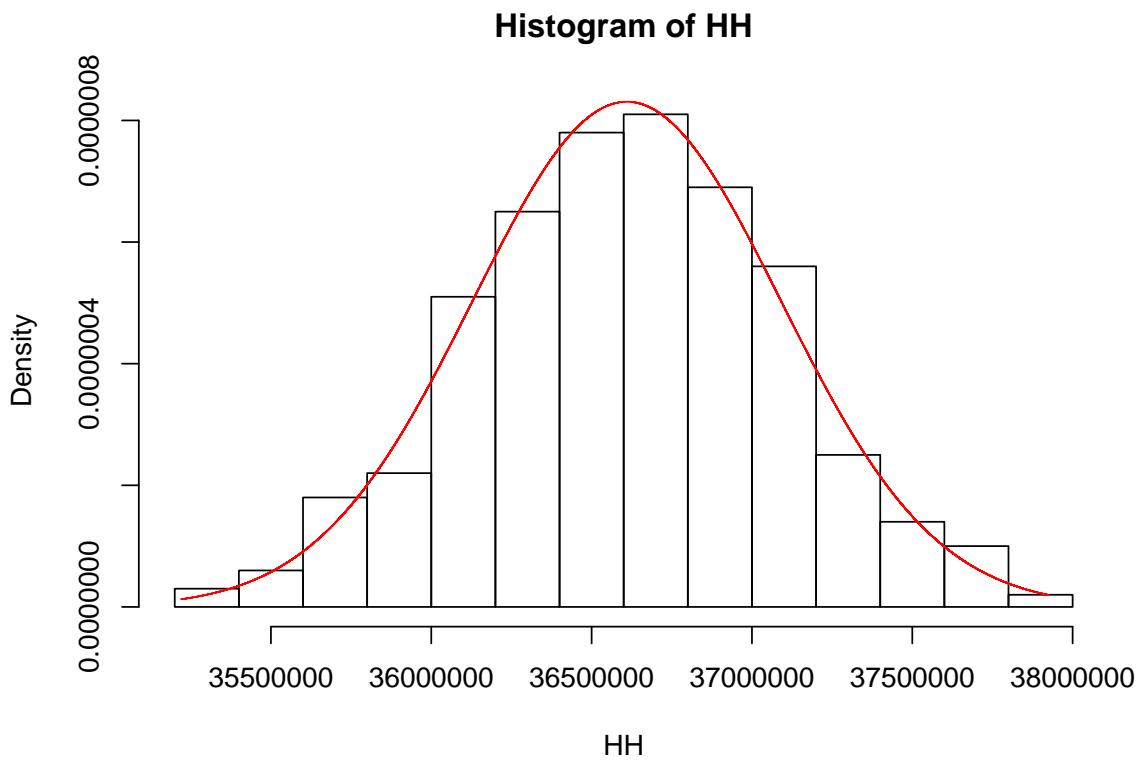


Figura 3.3: Distribución empírica del estimador de Hansen-Hurwitz para el diseño de muestreo aleatorio simple con reemplazo.

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

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.

### 3.4 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

$$N = na + c \quad (3.4.1)$$

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

$$c = N - \left\| \frac{N}{a} \right\| a \quad (3.4.2)$$

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 3.6: Posible configuración del muestreo sistemático.

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

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

$$U = \bigcup_{r=1}^a s_r. \quad (3.4.3)$$

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

$$Q_r = \{s_1, s_2, \dots, s_r, \dots, s_a\}. \quad (3.4.4)$$

**Resultado 3.4.1.** Para este diseño de muestreo, la cardinalidad del soporte es igual al número de grupos formados. Es decir

$$\#Q_r = a$$

**Definición 3.4.1.** 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

$$p(s) = \begin{cases} \frac{1}{a} & \text{si } s \in Q_r \\ 0 & \text{en otro caso} \end{cases} \quad (3.4.5)$$

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

### 3.4.1 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

$$s_r = \{k : k = r + (j - 1)a; j = 1, \dots, n(S)\} \quad (3.4.6)$$

**Ejemplo 3.4.1.** 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 3.4.2.** Para un diseño de muestreo sistemático, las probabilidades de inclusión de primer y segundo orden están dadas por

$$\pi_k = \frac{1}{a} \quad (3.4.7)$$

$$\pi_{kl} = \begin{cases} \frac{1}{a} & \text{si } k \text{ y } l \text{ pertenecen a } s_r \\ 0 & \text{en otro caso} \end{cases} \quad (3.4.8)$$

respectivamente.

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

$$\pi_k = Pr(k \in S) = Pr(\text{seleccionar la muestra } s_r) = \frac{1}{a} \quad (3.4.9)$$

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.  $\square$

### 3.4.2 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 3.4.3.** Para un diseño  $p(\cdot)$  con soporte  $Q$ , la varianza del estimador de Horvitz-Thompson, se puede escribir como

$$Var(\hat{t}_{y,\pi}) = \sum_U \sum_{kl} \frac{\pi_{kl}}{\pi_k \pi_l} y_k y_l - \left( \sum_U y_k \right)^2 \quad (3.4.10)$$

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

$$Var(\hat{t}_{y,\pi}) = \sum_U \sum_{kl} \Delta_{kl} \frac{y_k y_l}{\pi_k \pi_l} \quad (3.4.11)$$

$$= \sum_U \sum_{kl} (\pi_{kl} - \pi_k \pi_l) \frac{y_k y_l}{\pi_k \pi_l} \quad (3.4.12)$$

$$= \sum_U \sum_{kl} \left( \frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) y_k y_l \quad (3.4.13)$$

$$= \sum_U \sum_{kl} \frac{\pi_{kl}}{\pi_k \pi_l} y_k y_l - \sum_U \sum_{kl} y_k y_l \quad (3.4.14)$$

$$= \sum_U \sum_{kl} \frac{\pi_{kl}}{\pi_k \pi_l} y_k y_l - \left( \sum_U y_k \right)^2 \quad (3.4.15)$$

En donde se utiliza el hecho de que

$$\sum_U \sum_{kl} y_k y_l = \sum_U \sum_{k \neq l} y_k y_l + \sum_U y_k^2 = \left( \sum_U y_k \right)^2 \quad (3.4.16)$$

$\square$

**Resultado 3.4.4.** Para el diseño de muestreo sistemático, el estimador de Horvitz-Thompson y su varianza están dados por:

$$\hat{t}_{y,\pi} = at_{sr}, \quad (3.4.17)$$

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

$$Var_{SIS}(\hat{t}_{y,\pi}) = a \sum_{r=1}^a (t_{sr} - t)^2 \quad (3.4.18)$$

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

$$\hat{t}_{y,\pi} = \sum_{sr} \frac{y_k}{\pi_k} = at_{sr} \quad (3.4.19)$$

Utilizando los dos anteriores resultados, se sigue que

$$Var(\hat{t}_{y,\pi}) = \sum_U \sum_U \frac{\pi_{kl}}{\pi_k \pi_l} y_k y_l - \left( \sum_U y_k \right)^2 \quad (3.4.20)$$

$$= a \sum_{r=1}^a \left( \sum_{sr} \sum_{sr} y_k y_l \right) - t^2 \quad (3.4.21)$$

$$= a \sum_{r=1}^a \left( \sum_{k \in s_r} y_k \sum_{l \in s_r} y_l \right) - t^2 \quad (3.4.22)$$

$$= a \sum_{r=1}^a t_{sr}^2 - t^2 \quad (3.4.23)$$

$$= a \sum_{r=1}^a (t_{sr} - \bar{t})^2 \quad (3.4.24)$$

donde

$$\bar{t} = \sum_{r=1}^a \frac{t_{sr}}{a} = \frac{t}{a} \quad (3.4.25)$$

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.

### 3.4.3 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 3.7: 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.

### 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 3.4.5.** *Suponga que la población se divide en  $a$  grupos, de tal forma que existen  $n$  elementos*

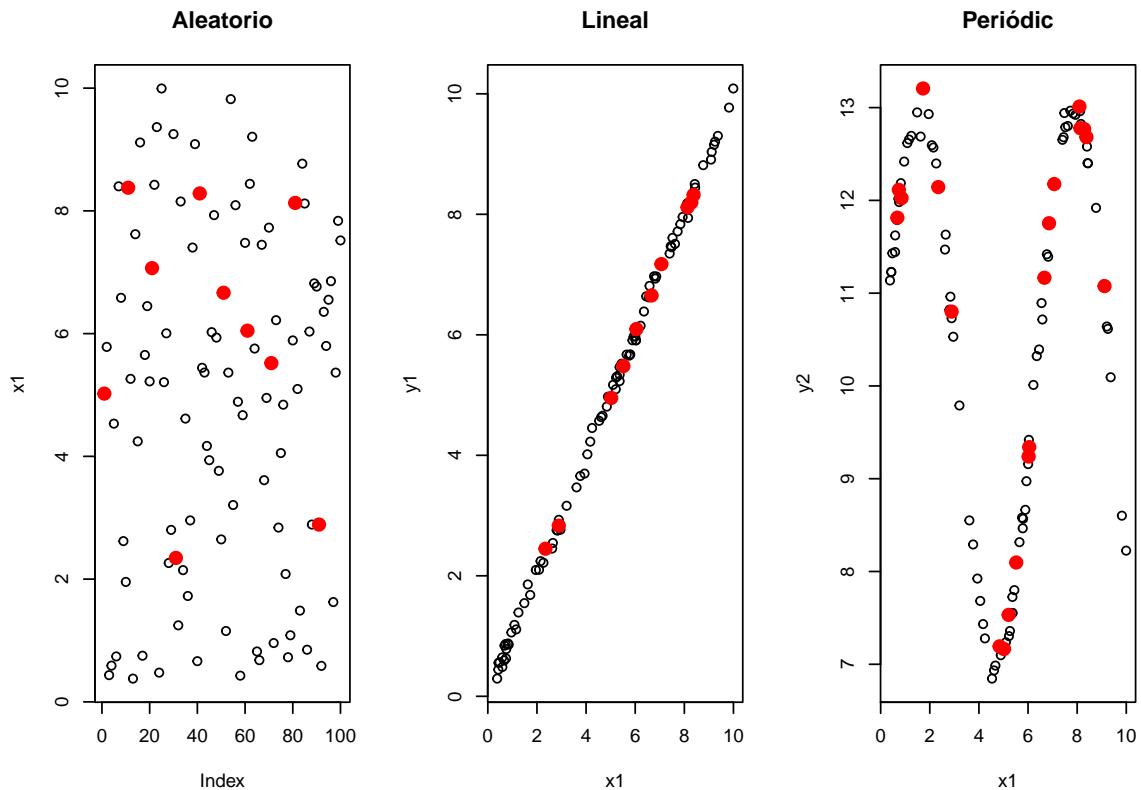


Figura 3.4: Casos de ordenamiento en muestreo sistemático.

por grupo y el tamaño poblacional toma la forma  $N = an$ , entonces

$$(N - 1)S_{y_U}^2 = \underbrace{\sum_U (y_k - \bar{y}_U)^2}_{SCT} = \underbrace{\sum_{r=1}^a \sum_{s_r} (y_{rk} - \bar{y}_{s_r})^2}_{SCD} + \underbrace{\sum_{r=1}^a n (\bar{y}_{s_r} - \bar{y}_U)^2}_{SCE} \quad (3.4.26)$$

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.

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

Cuadro 3.8: Tabla de ANOVA inducida por el muestreo sistemático.

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

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  $t_{sr}$  con  $r = 1, \dots, a$  son similares

$$t_{s_1} \approx t_{s_2} \approx \dots \approx t_{s_a} \approx \bar{t} \quad (3.4.27)$$

**Resultado 3.4.6.** 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

$$Var_{SIS}(\hat{t}_{y,\pi}) = N \sum_{r=1}^a n (\bar{y}_{sr} - \bar{y}_U)^2 = N(SCE) \quad (3.4.28)$$

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

$$\rho = 1 - \frac{n}{n - 1} \frac{SCD}{SCT} \quad (3.4.29)$$

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 3.4.7.** Utilizando la relación 3.4.26  $SCT=SCE+SCD$  se tiene que

$$SCE = SCT \left[ (\rho - 1) \frac{n-1}{n} + 1 \right] \quad (3.4.30)$$

*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.  $\square$

**Resultado 3.4.8.** Con el anterior resultado no es difícil verificar que la varianza del estimador de Horvitz-Thompson bajo muestreo sistemático se puede escribir como

$$Var_{SIS}(\hat{t}_{y,\pi}) = \underbrace{\frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{yU}^2}_{Var_{MAS}(\hat{t}_{y,\pi})} \left\{ \frac{N-1}{N-n} [1 + (n-1)\rho] \right\} \quad (3.4.31)$$

*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  $\square$

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

$$Def = \frac{Var_{SIS}\hat{t}_\pi}{Var_{MAS}\hat{t}_\pi} = \frac{N-1}{N-n} [1 + (n-1)\rho] \quad (3.4.32)$$

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.  $\square$

### 3.4.4 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 3.4.3.** El diseño de muestreo sistemático con  $q$  réplicas está definido como

$$p(s) = \frac{1}{\binom{a}{q}} \quad \text{para todo } s \in Q_r \quad (3.4.33)$$

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

$$\pi_k = \frac{q}{a} \quad (3.4.34)$$

$$\pi_{kl} = \begin{cases} \frac{q}{a} & \text{si } k \text{ y } l \text{ pertenecen a } s_r \\ \frac{q}{a} \frac{q-1}{a-1} & \text{en otro caso} \end{cases} \quad (3.4.35)$$

respectivamente.

**Resultado 3.4.11.** Para el diseño de muestreo sistemático con  $q$  réplicas, el estimador de Horvitz-Thompson y su varianza están dados por:

$$\hat{t}_{y,\pi} = \frac{a}{q} \sum_S t_{sr} \quad (3.4.36)$$

$$VarSIS(\hat{t}_{y,\pi}) = \frac{a^2}{q} \left(1 - \frac{q}{a}\right) S_{t_{sr}U}^2 \quad (3.4.37)$$

$$\widehat{VarSIS}(\hat{t}_{y,\pi}) = \frac{a^2}{q} \left(1 - \frac{q}{a}\right) S_{t_{sr}s}^2 \quad (3.4.38)$$

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  $q$ -sistemático 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  $q$ -sistemático 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.

### 3.4.5 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 característica 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.

```

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

```

Los resultados de la estimación se muestran en la tabla 3.9. 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 3.9: *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

<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.

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 tienen un comportamiento parecido entre ellos. El código necesario para la creación de este gráfico está dado a continuación.

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

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.

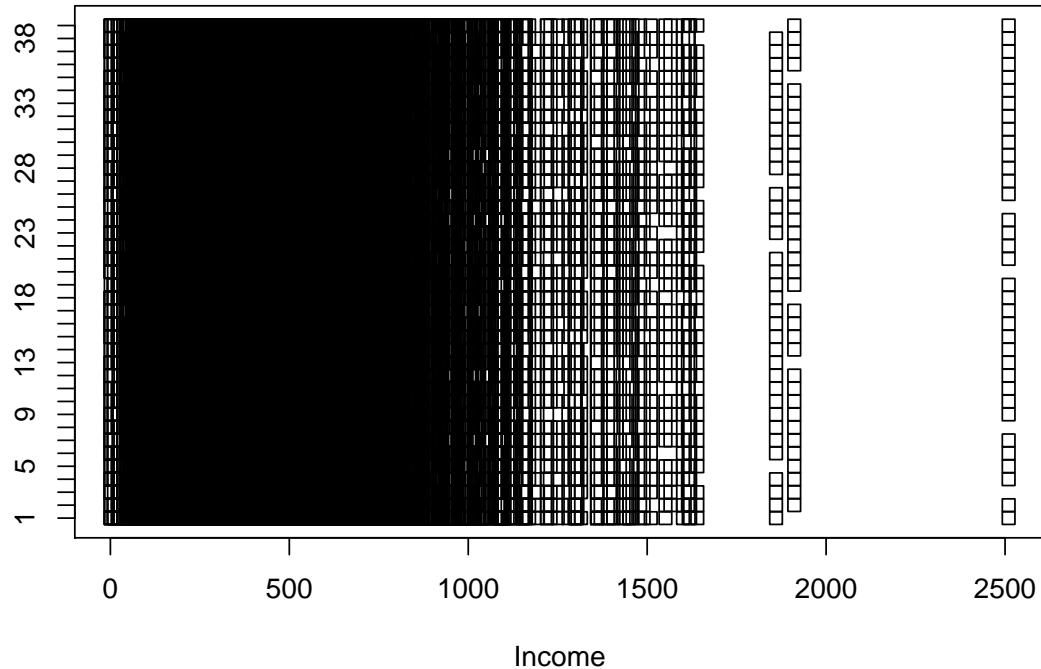


Figura 3.5: Distribución de la característica *Income* con respecto a los grupos creados en el muestreo sistemático.

```
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 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
```

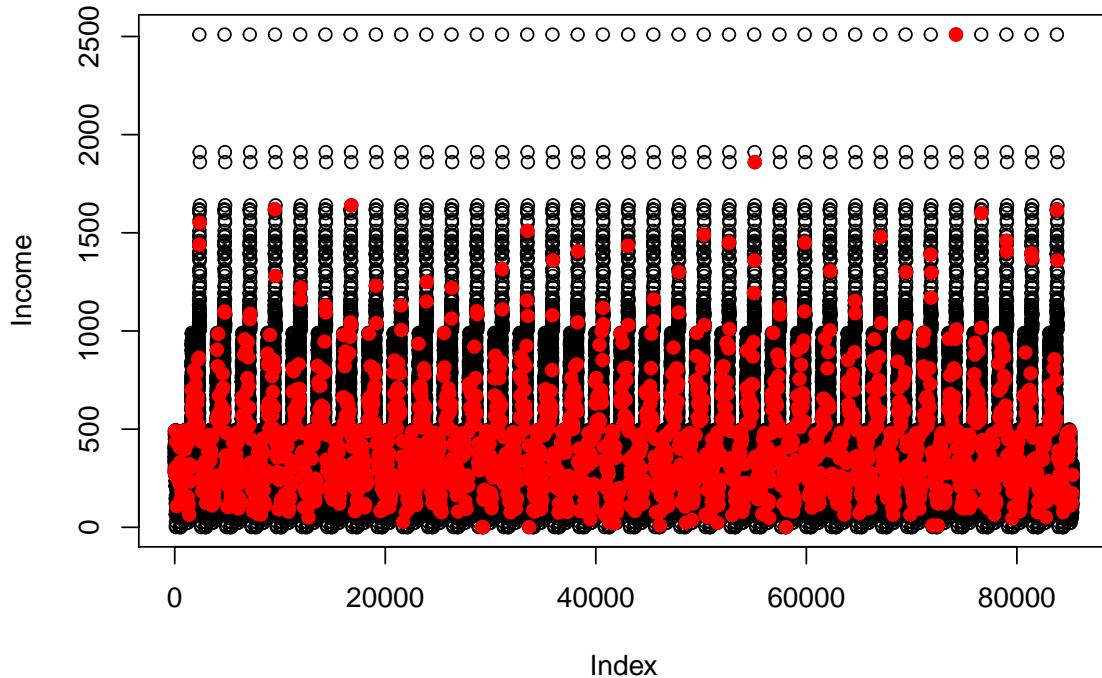


Figura 3.6: Casos seleccionados en muestreo sistemático.

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.

### 3.5 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

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

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

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
[5,]		0.1666667
[6,]	1	
[7,]	1	0.2500000
[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».

## Capítulo 4

# Muestras con probabilidades proporcionalas

Es bien sabido que la estrategia de muestreo que utiliza un diseño de muestreo aleatorio simple con el estimador de Horvitz-Thompson, es una estrategia de muestreo óptima, bajo ciertas formulaciones, si se tiene un conocimiento a priori de que el comportamiento de la población es simétrico con respecto a los rótulos. En tales casos, la incorporación de información auxiliar no mejora la anterior estrategia.

Cassel, Särndal & Wretman (1976b)

Las estrategias de muestreo implementadas en el capítulo anterior, utilizaban métodos de selección tales que la probabilidad de inclusión o probabilidad de selección es idéntica para todos los elementos de la población y se estimaban los parámetros de interés utilizando el estimador de Hansen-Hurwitz, para diseños de muestreo con reemplazo y el estimador de Horvitz-Thompson, para diseños de muestreo sin reemplazo. Las anteriores estrategias no tienen en cuenta la variación innata de las características de interés a través de las unidades poblacionales. Por lo tanto, los anteriores estimadores, dada su construcción genérica y el principio de representatividad, tenderán a poseer una gran variación.

Raj (1968) afirma que, en cuestión de precisión, se puede tener una mayor ganancia cuando se utilizan diseños de muestreo con probabilidades desiguales. En la mayoría de los casos prácticos, la característica de interés no presenta un comportamiento uniforme con respecto a los rótulos de la población. Sin embargo, cuando el marco de muestreo disponible para la selección de la muestra contiene además de la identificación y la ubicación de los elementos en la población, una característica auxiliar continua disponible para todos los elementos de la población  $x_k \quad \forall k \in U$ , es posible utilizar diseños de muestreo que implementen métodos de selección cuyas probabilidades de selección o inclusión, dependiendo del caso, sean proporcionales al total de la característica auxiliar,  $t_x$ .

### 4.1 Diseño de muestreo de Poisson

Este diseño de muestreo es una generalización del diseño de muestreo Bernoulli, en donde las probabilidades de inclusión están dadas a priori de manera independiente para cada individuo. Brewer (2002) indica que este diseño de muestreo no tuvo originalmente ninguna implicación práctica, porque el tamaño de muestra no es fijo, sino que fue utilizado de manera teórica para describir las propiedades de otros estimadores. El primer caso práctico se dio en la selección de muestras de árboles en unidades forestales; más adelante se aplicó en el censo anual manufacturero en Estados Unidos. Aunque este

diseño de muestreo no utiliza información auxiliar para la selección de la muestra, sirve como punto de partida para examinar diseños de muestreo más complejos que sí lo utilizan.

**Definición 4.1.1.** Siendo  $\pi_k$  un número positivo, tal que  $0 < \pi_k \leq 1$ , que representa la probabilidad de inclusión del  $k$ -ésimo elemento, el diseño de muestreo Poisson se define de la siguiente manera

$$p(s) = \prod_{k \in s} \pi_k \prod_{k \notin s} (1 - \pi_k) \quad \text{para todo } s \in Q \quad (4.1.1)$$

con  $Q$ , el soporte que contiene a todas las posibles muestras sin reemplazo.

**Resultado 4.1.1.** Para este diseño de muestreo, el soporte  $Q$  tiene cardinalidad igual a

$$\#(Q) = 2^N$$

**Ejemplo 4.1.1.** En nuestra población ejemplo

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

Las probabilidades de inclusión  $\pi_k$  son 0.2, 0.5, 0.7, 0.5 y 0.9, respectivamente. Las posibles muestra pueden ser de tamaño 0, 1, 2, 3, 4 ó 5. La probabilidad de la muestra de tamaño 0 es

$$(1 - 0.2) \times (1 - 0.5) \times (1 - 0.7) \times (1 - 0.5) \times (1 - 0.9) = 0.006$$

Siguiendo esta misma analogía, a continuación se presenta el cálculo léxico-gráfico para las probabilidades de selección de todas las posible muestras en el soporte de este diseño de muestreo. Para las posibles muestras de tamaño 1, 4 se tiene que sus respectivas probabilidades son:

s	p(s)		s	p(s)
Yves	0.0015		Yves, Ken, Erik, Sharon	0.0035
Ken	0.006		Yves, Erik, Sharon, Leslie	0.0315
Erik	0.014		Yves, Ken, Erik, Leslie	0.0315
Sharon	0.006		Yves, Ken, Sharon, Leslie	0.0135
Leslie	0.054		Ken, Erik, Sharon, Leslie	0.126
Total	0.0815		Total	0.206

Las posibles muestras de tamaño 2, 3 y sus respectivas probabilidades son:

s	p(s)		s	p(s)
Yves, Ken	0.0015		Yves, Ken, Erik	0.0035
Yves, Erik	0.0035		Yves, Ken, Sharon	0.0015
Yves, Sharon	0.0015		Yves, Ken, Leslie	0.0135
Yves, Leslie	0.0135		Yves, Erik, Sharon	0.0035
Ken, Erik	0.014		Yves, Erik, Leslie	0.0315
Ken, Sharon	0.006		Yves, Sharon, Leslie	0.0135
Ken, Leslie	0.054		Ken, Erik, Sharon	0.014
Erik, Sharon	0.014		Ken, Erik, Leslie	0.126
Erik, Leslie	0.126		Ken, Sharon, Leslie	0.054
Sharon, Leslie	0.054		Erik, Sharon, Leslie	0.126
Total	0.288		Total	0.387

Finalmente, la muestra de tamaño 5,  $\{\text{Yves, Ken, Erik, Sharon, Leslie}\}$ , tiene probabilidad 0.0315. Nótese que la suma de todas las posibles muestras es  $\sum p(s) = 1$ .

### 4.1.1 Algoritmo de selección

Bautista (1998) afirma que el conocimiento a priori de las probabilidades de inclusión de los elementos es tal que, en algunas ocasiones, existen elementos de la población que deben ser observados obligatoriamente en la muestra, en estos casos el valor de la probabilidad de inclusión de estos elementos es igual a uno ( $\pi_k = 1$ ). Al subgrupo poblacional cuyos elementos tienen probabilidad de inclusión igual a uno, se le conoce como subgrupo de **inclusión forzosa**. Nótese que el algoritmo de selección de muestra utilizado debe contemplar la inclusión en todas las posibles muestras realizadas de todos los elementos del subgrupo de inclusión forzosa.

La selección de una muestra con diseño de muestreo Poisson se realiza mediante un algoritmo secuencial definido de manera similar que el algoritmo utilizado en la selección de muestras con diseño de muestreo Bernoulli.

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

Dado que  $\varepsilon_k \sim Unif[0, 1]$ , se tiene que  $Pr(\varepsilon_k < \pi_k) = \pi_k$  para  $k \in U$ . Por tanto, la inclusión de los individuos  $k$ -ésimo y  $l$ -ésimo, para  $k \neq l$ , es independiente; sin embargo, la distribución de  $I_k(S)$  no es de tipo Binomial puesto que las variables aleatorias  $I_k(S)$  no son idénticamente distribuidas.

**Resultado 4.1.2.** *Bajo muestreo Poisson, el tamaño de muestra  $n(S)$  es una variable aleatoria, tal que*

$$E(n(S)) = \sum_U \pi_k \quad \text{Var}(n(S)) = \sum_U \pi_k(1 - \pi_k) \quad (4.1.2)$$

*Demostración.* Utilizando el resultado 2.1.4 y las propiedades de una suma de cuadrados es suficiente probar que  $\pi_{kl} = Pr(k \in S, l \in S) = \pi_k \pi_l$  para  $k \neq l$ , lo cual se tiene de inmediato dado que las variables aleatorias  $I_k(S)$  e  $I_l(S)$  son independientes.  $\square$

**Resultado 4.1.3.** *Para el diseño de muestreo Poisson, las probabilidades de inclusión de primer y segundo orden están dadas por:*

$$\pi_k = \pi_k \quad (4.1.3)$$

$$\pi_{kl} = \begin{cases} \pi_k & \text{para } k = l \\ \pi_k \pi_l & \text{en otro caso} \end{cases} \quad (4.1.4)$$

respectivamente.

### 4.1.2 El estimador de Horvitz-Thompson

**Resultado 4.1.4.** *Para el diseño de muestreo Poisson, el estimador de Horvitz-Thompson, su varianza y su varianza estimada están dados por:*

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} \quad (4.1.5)$$

$$Var_{PO}(\hat{t}_{y,\pi}) = \sum_U \left( \frac{1}{\pi_k} - 1 \right) y_k^2 \quad (4.1.6)$$

$$\widehat{Var}_{PO}(\hat{t}_{y,\pi}) = \sum_S (1 - \pi_k) \left( \frac{y_k}{\pi_k} \right)^2 \quad (4.1.7)$$

respectivamente.

*Demostración.* Utilizando el resultado 2.2.2, se sigue que la demostración es inmediata puesto que

$$\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k \pi_l = \pi_k \pi_l - \pi_k \pi_l = 0 & \text{para } k \neq l \\ \pi_{kk} - \pi_k^2 = \pi_k (1 - \pi_k) & \text{para } k = l \end{cases} \quad (4.1.8)$$

luego la doble suma en la varianza del estimador de Horvitz-Thompson pasa a ser una sola suma. La demostración para el estimador de la varianza se lleva a cabo de manera análoga.  $\square$

**Ejemplo 4.1.2.** Para nuestra población de ejemplo  $U$ , suponga que el individuo **Erik** debe estar en la muestra seleccionada; es decir,  $\pi_{Erik} = 1$ . Por tanto, existen  $\binom{1}{1} 2^4 = 16$  posibles muestras. Si el vector de probabilidades de inclusión para cada elemento de la población está dado por  $(0.5, 0.2, 1, 0.9, 0.5)$ . Realice el cálculo léxico-gráfico del estimador de Horvitz-Thompson y compruebe el insesgamiento, la varianza y las propiedades del diseño de muestreo.

#### 4.1.3 Optimalidad en la estrategia de muestreo Poisson

Como se mencionó en capítulos anteriores, una estrategia de muestreo que utilice el estimador de Horvitz-Thompson, es óptima cuando las probabilidades de inclusión inducidas por el diseño de muestreo utilizado están correlacionadas positivamente con la característica de interés; en otras palabras, cuando  $\pi_k \propto y_k$ . En este caso utópico, y si se supone que el diseño de muestreo es de tamaño de muestra fijo ( $n(S) = n$ ), el estimador de Horvitz-Thompson reproduciría el parámetro de interés  $t_y$  con varianza nula cuando las probabilidades de inclusión toman la siguiente forma  $\pi_k = n \frac{y_k}{t_y}$ . De esta forma, la estrategia utilizada sería una estrategia representativa con respecto a la variable de interés, puesto que para cualquier muestra seleccionada, el estimador de Horvitz-Thompson sería igual a  $t_y$ .

**Resultado 4.1.5.** Suponiendo un tamaño de muestra fijo, bajo un diseño de muestreo Poisson, la varianza del estimador de Horvitz-Thompson se minimiza cuando

$$\pi_k = \frac{ny_k}{\sum_U y_k} \quad (4.1.9)$$

*Demostración.* El objetivo es encontrar valores de  $\pi_k$ , tales que  $0 < \pi_k \leq 1$  que minimicen la varianza del estimador de Horvitz-Thompson bajo diseño de muestreo Poisson, lo anterior se tiene cuando se realiza un censo, es decir cuando  $\pi_k = 1$  para todo  $k \in U$ . Sin embargo, en la práctica se desea seleccionar una muestra de tamaño menor a  $N$ . Por tanto, minimizar  $Var_{PO}(\hat{t}_{y,\pi})$  es equivalente a minimizar  $\sum_U \frac{y_k^2}{\pi_k}$  sujeto a la restricción de un tamaño de muestra fijo, tal que  $\sum_U \pi_k = n$ . Luego la cantidad a minimizar está dada por el siguiente producto

$$\left( \sum_U \frac{y_k^2}{\pi_k} \right) \left( \sum_U \pi_k \right)$$

Una solución al anterior problema es utilizar la desigualdad de Cauchy-Schwartz, por tanto

$$\left( \sum_U \frac{y_k^2}{\pi_k} \right) \left( \sum_U \pi_k \right) \geq \left( \sum_U y_k \right)^2$$

Con igualdad cuando  $\frac{y_k}{\pi_k} = c$ , con  $c$  una constante. Ahora, se tiene que

$$n = \sum_U \pi_k = \sum_U \frac{y_k}{c}$$

Luego,

$$c = \sum_U \frac{y_k}{n}$$

Por tanto,

$$\pi_k = \frac{ny_k}{\sum_U y_k}$$

□

El anterior resultado es una ambigüedad puesto que con esa escogencia de las probabilidades de inclusión se asume que la característica de interés es conocida para toda la población. Si lo anterior sucede, no existiría la necesidad de estimar  $t_y$ . Sin embargo, Särndal, Swensson & Wretman (1992) aseguran que como el diseño de muestreo Poisson es de tamaño de muestra variable es ineficiente y utilizar el anterior razonamiento implicaría que el estimador de Horvitz-Thompson tome la siguiente forma

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} = \frac{t_y}{n} \sum_S 1 = t_y \frac{n(S)}{n} \quad (4.1.10)$$

Por tanto, la variación del estimador calculado en cada muestra estaría dada por la variación del tamaño de muestra esperado  $n(S)$ . El anterior razonamiento nos lleva a pensar que el estimador de Horvitz-Thompson tendría un excelente desempeño bajo diseños de muestreo tales que  $\pi_k \propto y_k$  y que induzcan muestras de tamaño fijo. Por otro lado, si el marco de muestreo tiene la virtud de adjuntar información auxiliar continua, por medio de una característica de interés  $x_k$  (en otras palabras, conocer el vector de características auxiliares  $x_1, x_2, \dots, x_N$  antes de realizar el muestreo) que esté muy bien correlacionada con la variable de interés, entonces la varianza de la estrategia de muestreo sería mínima cuando

$$\pi_k = n \frac{x_k}{\sum_U x_k} \quad (4.1.11)$$

Por otro lado, y siguiendo el mismo razonamiento que en el diseño de muestreo Bernoulli, como se tiene un marco de muestreo de elementos, entonces se conoce el tamaño poblacional  $N$ . De esta manera, un estimador para el total poblacional de la característica de interés con menor varianza es el llamado estimador alternativo dado por la expresión (2.2.18), que para el caso particular de muestreo Poisson toma la siguiente forma

$$\hat{t}_{y,alt} = \hat{t}_{y,\pi} \frac{N}{\hat{N}_\pi} \quad (4.1.12)$$

Para estimar la media poblacional, es posible utilizar este mismo razonamiento y junto con la expresión (2.2.15) resulta un estimador menos disperso

$$\tilde{y}_S = \frac{\hat{t}_{y,\pi}}{\hat{N}_\pi} \quad (4.1.13)$$

La forma estructural de los anteriores estimadores es una razón, cociente de dos cantidades aleatorias, y así se reduce parte de la variabilidad del estimador de Horvitz-Thompson que viene del hecho de que el tamaño muestral no es fijo para este diseño.

#### 4.1.4 Marco y Lucy

Aunque esta estrategia de muestreo no fue utilizada en el sentido práctico y tiene una varianza alta dado que el tamaño de muestra es variable, es posible obtener buenos resultados que incentivar el uso de las estrategias de muestreo con probabilidad proporcional al tamaño. En primer lugar, se debe suponer que el marco de muestreo contiene una característica auxiliar continua que será usada en la etapa de diseño y selección de la muestra.

Raj (1968) señala que en el caso concreto de una población agrícola, una característica auxiliar puede ser el área cultivada, para el caso de hogares, una característica auxiliar puede ser el número de personas que habitan en el hogar. Lehtonen & Pahkinen (2003) dan ejemplos claros acerca de las características auxiliares en encuestas de empresas y afirman que para este caso particular una característica auxiliar comúnmente usada es el número de empleados en la empresa; para el caso de encuestas a escuelas, una característica auxiliar es el número de alumnos. En encuestas a hospitales Bautista (1998) afirma que una característica auxiliar es el número de camas por hospital, no así el número de pacientes, pues esta última característica tiene una variación alta y está ligada a la temporada de realización de la encuesta.

Recuérdese que se quieren estimar tres totales de las características de interés Ingreso, Empleados e Impuestos del último periodo fiscal en las empresas del sector industrial. Para efectos prácticos, suponga que el marco de muestreo contiene todos los registros de cada una de las empresas del sector industrial de la característica Ingreso; de esta manera se podrá estimar el total poblacional para las características Empleados e Impuestos. Para efectos académicos, se estimará el total poblacional de la característica Ingreso, resaltando que hacerlo es una ambigüedad porque si se conocen todos los valores poblacionales de la característica de interés no hay necesidad de estimar lo que ya es conocido; sin embargo, como ejercicio académico es completamente admisible.

Con los supuestos anteriores, el marco de muestreo se carga en el ambiente de programación de R, nótese que el marco de muestreo ahora contiene cinco columnas, cuatro que se refieren a la identificación y/o ubicación geográfica y una columna que contiene los registros para la característica Ingreso.

```
data(BigLucy)
dim(BigLucy)

## [1] 85296    11
```

Las probabilidades de inclusión deben ser creadas y están dadas por (4.1.9). Nótese que se debe fijar un tamaño esperado de muestra. Para que los resultados sean comparables, se utilizará un tamaño esperado de muestra de  $n(S) = 400$ . Una vez que las probabilidades de inclusión para todas las empresas del sector industrial han sido creadas, se debe verificar que cada una de ellas sea menor a la unidad; para esto, se utiliza la función `which` que R trae implementada en su ambiente básico y cuya salida es un conjunto de índices para los cuales la instrucción dentro del paréntesis es verdadera; cuando no existe ningún índice que cumpla (`pik>1`), la función arroja la siguiente salida `integer(0)`. Sin embargo, si hubiese existido algún registro para el cual la instrucción (`pik>1`) sea cierta, se deben convertir las respectivas probabilidades de inclusión en la unidad.

```
attach(BigLucy)
N <- dim(BigLucy)[1]
n <- 2000
pik <- n * Income / sum(Income)
which(pik>1)

## integer(0)
```

```
sum(pik)

## [1] 2000
```

Nótese que la suma de las probabilidades de inclusión es igual al tamaño de muestra esperado. La correlación entre las probabilidades de inclusión inducidas mediante este diseño de muestreo Poisson es buena. Por supuesto, la correlación entre las  $\pi_k$  y la variable ingreso es uno pues las primeras son función lineal de Ingreso. Ahora, la cantidad de impuestos que las empresas del sector industrial declaran en un año fiscal, es proporcional al ingreso de las mismas; de hecho, si una empresa tiene ganancias nulas, entonces declarará impuestos nulos. Por otro lado, aunque una empresa tenga ganancias nulas, no necesariamente tendrá cero empleados; de hecho, en el sector industrial existen casos en donde una empresa con pocos empleados, tiene ingresos más altos que una empresa con muchos empleados; sin embargo, esta particularidad no se presenta de manera general, si esto fuera así, la correlación sería negativa y la característica de auxiliar Ingreso no debería ser utilizada en la estimación del total de la característica de interés Empleados.

```
cor(pik, cbind(Income, Employees, Taxes))

##      Income Employees Taxes
## [1,]     1     0.6433 0.9167
```

La figura 4.1 muestra el diagrama de dispersión de las tres variables de interés contra el vector de probabilidades de inclusión.

Para seleccionar la muestra bajo un diseño de muestreo Poisson, se utiliza la función **S.PO** del paquete **TeachingSampling**. Esta función consta de dos argumentos, **N**, el tamaño poblacional y **pik**, el vector de probabilidades de inclusión para cada elemento de la población. En nuestro caso, **pik** es el vector de probabilidades creado anteriormente; pero, en general, puede ser utilizado cualquier vector de números entre cero y uno. La función **S.PO** devuelve un conjunto de índices que aplicados a la población resulta en los valores de las características de interés para cada miembro de la muestra seleccionada.

```
sam <- S.PO(N, pik)
muestra <- BigLucy[sam,]
attach(muestra)

sam <- S.PO(N, pik)
muestra <- BigLucy[sam,]
attach(muestra)
head(muestra)

##           ID Ubication Level Zone Income Employees Taxes
## 30 AB0000000030 C0053291K0248606 Small County1   354      33      5
## 68 AB0000000068 C0269132K0032765 Small County1   360      29      5
## 73 AB0000000073 C0061171K0240726 Small County1   330      79      4
## 75 AB0000000075 C0210740K0091157 Small County1   416      32      7
## 193 AB0000000193 C0178986K0122911 Small County1   350      48      5
## 225 AB0000000225 C0197969K0103928 Small County1   330      67      4
##          SPAM ISO Years Segments
## 30      yes  no 44.4 County1 3
## 68      yes  no 21.0 County1 7
## 73      yes  no 20.2 County1 8
```

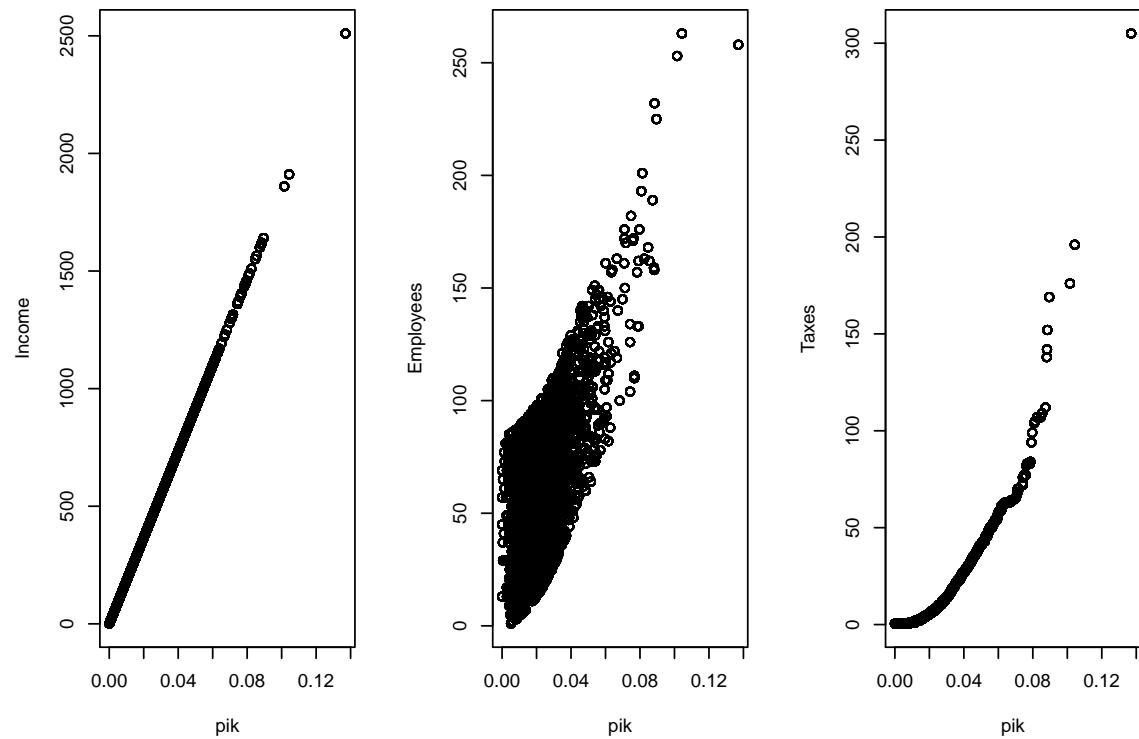


Figura 4.1: Correlación de las probabilidades de inclusión con las características de interés.

```

## 75   yes  no    8.8 County1 8
## 193  yes  no    3.7 County1 20
## 225  yes  no   42.5 County1 23

n.s <- dim(muestra) [1]
n.s

## [1] 2123
  
```

En este caso particular, la primera empresa seleccionada es la identificada con el número AB0000000030. Nótese que el marco de muestreo incluye la característica auxiliar Ingreso y que el tamaño efectivo de muestra es 2123. Una vez que el trabajo de campo ha concluido, comienza la etapa de estimación, en donde se utilizará la función E.PO del paquete **TeachingSampling** que consta de dos argumentos, la matriz o vector de valores de la o las características de interés y pik.s los valores del vector de probabilidad de inclusión de cada uno de los elementos seleccionados en la muestra. En este caso particular se crea un conjunto de datos con la información muestral de las características de interés llamado **estima**. Nótese que la longitud del vector pik.s es de 2123. La función E.PO devuelve las estimaciones del total poblacional, la varianza estimada y el respectivo coeficiente de variación de la(s) característica(s) de interés.

```
pik.s <- pik[sam]
estima <- data.frame(Income, Employees, Taxes)
E.PO(estima, pik.s)
```

La tabla 4.1 muestra los resultados particulares para esta estrategia de muestreo. Nótese que la característica Impuestos, tiene un menor coeficiente de variación porque está mucho mejor correlacionada con el vector de probabilidades de inclusión, mientras que la característica Empleados presenta un mayor coeficiente de variación. Desde un punto de vista completamente académico, está bien afirmar que la estrategia de muestreo utilizada puede ser optimizada si se utiliza un diseño de muestreo con probabilidades de inclusión proporcionales al tamaño de alguna característica auxiliar, pero que induzca muestras de tamaño fijo. Nótese que, aunque el vector de probabilidades de inclusión tiene una correlación de uno con respecto a la característica Income, el coeficiente de variación estimado para esta es de un 2.1348 %, cifra que no es alta, pero que no paga el precio de utilizar esta información auxiliar en la etapa de diseño. Véase que los coeficientes de variación son un poco más bajos que al utilizar un diseño de muestreo Bernoulli, pero no más bajos que los obtenidos al usar un diseño de muestreo aleatorio simple.

Cuadro 4.1: *Estimaciones para el diseño de muestreo Poisson*

	N	Income	Employees	Taxes
Estimation	92131.01	38887769.08	5778632.54	1069292.15
Standard Error	2611.38	830192.20	151520.00	26433.92
CVE	2.83	2.13	2.62	2.47
DEFF	Inf	1.60	3.72	0.20

## 4.2 Diseño de muestreo PPT

Siguiendo con el razonamiento que se introdujo en la sección anterior, Bautista (1998) afirma que en un diseño de muestreo con reemplazo, los valores óptimos de las probabilidades de selección para cada elemento de la población tendrían que estar dados por

$$p_k = \frac{y_k}{t_y}.$$

Por supuesto, con esta escogencia, el estimador de Hansen-Hurwitz estimaría al total poblacional de la característica de interés con varianza nula. De otra forma, el tamaño de muestra necesario para obtener una estimación con sesgo nulo sería de  $m = 1$ . Nótese que por (2.2.34), el estimador de Hansen-Hurwitz, es un promedio de  $m$  estimaciones. Con la escogencia de probabilidades de selección anterior, y con un tamaño de muestra de  $m = 1$ , se tiene que

$$\begin{aligned}\hat{t}_{y,p} &= \frac{1}{1} \sum_{i=1}^1 \frac{y_{k_i}}{p_{k_i}} \\ &= \frac{y_{k_i}}{p_{k_i}} \\ &= t_y \frac{y_{k_i}}{y_{k_i}} = t_y\end{aligned}$$

Por supuesto, desde el punto de vista práctico sería una vez más, una ambigüedad la escogencia de las anteriores probabilidades de selección. Sin embargo, si el marco de muestreo es tal que contiene

el valor de una característica continua auxiliar  $x_k$  bien relacionada con la característica de interés  $y_k$  para cada elemento de la población, es posible mediante el estimador de Hansen-Hurwitz, estimar el parámetro de interés con una varianza pequeña. De hecho, entre mejor correlación exista entre  $y_k$  y  $x_k$  menor varianza tendrá el estimador de Hansen-Hurwitz.

**Definición 4.2.1.** Sea  $x_k$ , el valor de una característica auxiliar continua para el elemento  $k$ -ésimo tal que:

1.  $x_k > 0$  para todo  $k \in U$  y
2.  $x_k$  está disponible y es conocida para todos los elementos de la población.

Entonces, se define un diseño de muestreo con probabilidad de selección proporcional al tamaño de la característica auxiliar, de la siguiente manera

$$p(s) = \begin{cases} \frac{m!}{n_1(s)! \dots n_N(s)!} \prod_U \left( \frac{1}{p_k} \right)^{n_k(s)} & \text{si } \sum_U n_k(s) = m \\ 0 & \text{en otro caso} \end{cases} \quad (4.2.1)$$

Donde  $n_k(s)$  es el número de veces que el elemento  $k$ -ésimo es seleccionado en la muestra realizada  $s$  y  $p_k$  es la probabilidad de selección del elemento  $k$ -ésimo dada por

$$p_k = \frac{x_k}{t_x}. \quad (4.2.2)$$

con  $t_x$  el total poblacional de la característica auxiliar  $x$ .

**Resultado 4.2.1.** Para este diseño de muestreo, el soporte  $Q$  tiene cardinalidad igual a

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

**Resultado 4.2.2.** Dado el soporte  $Q$ , de todas las posibles muestras con reemplazo de tamaño  $m$ , se verifica que el diseño de muestreo con probabilidad de selección proporcional al tamaño de la característica auxiliar es tal que

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

*Demostración.* Dado que

$$\sum_U p_k = \sum_U \frac{x_k}{t_x} = 1$$

entonces la demostración del resultado es inmediata haciendo uso del teorema multinomial.  $\square$

**Resultado 4.2.3.** Para un diseño de muestreo con reemplazo y con probabilidades de selección proporcionales al tamaño de una característica de información auxiliar, las probabilidades de inclusión de primer y segundo orden están dadas por

$$\pi_k = 1 - (1 - p_k)^m \quad (4.2.3)$$

$$\pi_{kl} = 1 - (1 - p_k)^m - (1 - p_l)^m + (1 - p_k - p_l)^m \quad (4.2.4)$$

respectivamente. En donde  $p_k = \frac{x_k}{t_x}$

*Demostración.* Utilizando el resultado 2.2.9 se llega a la demostración inmediata.  $\square$

Cuando se tienen las cantidad 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.

### 4.2.1 Algoritmo de selección

#### Método acumulativo total

Hansen, Hurwitz & Madow (1953) plantearon este método de selección para ser utilizado junto con el estimador que lleva su nombre. Este método es conocido con el nombre de **algoritmo acumulativo total** y consiste en  $m$  selecciones independientes de tamaño 1, tal que:

- Sea

$$p_k = \frac{x_k}{t_x} \quad (4.2.5)$$

- Sea

$$T_k = \sum_{l=1}^k x_l \quad (4.2.6)$$

con  $T_0 = 0$

- Obtener  $\varepsilon$  como una realización de una variable aleatoria con distribución uniforme en el intervalo  $(0,1)$ .
- Seleccionar el  $k$ -ésimo elemento si  $T_{k-1} < \varepsilon T_N \leq T_k$ .

Al repetir  $m$  veces el anterior procedimiento, se ha seleccionado una muestra de un diseño con reemplazo con probabilidades de selección son proporcionales al tamaño de la característica de interés. Como este diseño de muestreo es con reemplazo, cuando existan elementos en la población cuyo valor de la característica auxiliar es muy grande, éstos elementos podrán ser seleccionados muchas veces porque sus probabilidades de selección son grandes con respecto a los demás elementos.

#### Método de Lahiri

En algunas ocasiones, cuando el tamaño poblacional  $N$  es muy grande, el anterior método resulta ineficiente. Lahiri (1951) plantea el siguiente algoritmo de selección: Siendo  $M \geq \max(x_1, \dots, x_N)$ , los siguientes dos pasos se ejecutan para seleccionar un elemento.

1. Seleccione un número  $l$  de manera aleatoria de una distribución de probabilidad uniforme discreta en el intervalo  $[1, N]$ .
2. Seleccione un número  $\eta$  de manera aleatoria de una distribución de probabilidad uniforme discreta en el intervalo  $[1, M]$ .

Si  $\eta \leq x_l$ , entonces el elemento  $l$ -ésimo es seleccionado. Si, por el contrario,  $\eta > x_l$  se repite el procedimiento hasta seleccionar una unidad. Si el tamaño de la muestra a seleccionar es  $m$ , entonces el anterior esquema se realiza  $m$  veces.

**Ejemplo 4.2.1.** Suponga que para la población de ejemplo  $U$  se tiene conocimiento de cada valor de la siguiente característica de información auxiliar correlacionada con la característica de interés.

```
U <- c("Yves", "Ken", "Erik", "Sharon", "Leslie")
x <- c(52, 60, 75, 100, 50)
x

## [1] 52 60 75 100 50
```

Para seleccionar una muestra con probabilidad proporcional a  $x$ , se crean las probabilidades de selección dadas por

```
pk <- x / sum(x)
pk

## [1] 0.1543 0.1780 0.2226 0.2967 0.1484
```

Para seleccionar una muestra con reemplazo de la población  $U$  mediante el método acumulativo total, el paquete **TeachingSampling** implementa la función **S.PPS** que consta de dos argumentos,  $m$  el tamaño de muestra y  $x$  la característica de interés que contiene todos y cada uno de los valores correspondientes a los elementos de la población para la característica auxiliar.

```
sam <- S.PPS(3, x)
U[sam]

## [1] "Erik" "Yves" "Ken"
```

La salida de la función **S.PPS** es un conjunto de índices (no necesariamente distintos) que aplicados a los rótulos poblacionales proporcionan la muestra seleccionada.

#### 4.2.2 El estimador de Hansen-Hurwitz

Hansen & Hurwitz (1943) propusieron el siguiente estimador insesgado para el parámetro de interés  $t_y$  con ayuda de información auxiliar continua en la etapa de diseño.

**Resultado 4.2.4.** Sea  $x_k$ , el valor de una característica auxiliar continua, para un diseño de muestreo aleatorio proporcional al tamaño con reemplazo, el estimador de Hansen-Hurwitz del total poblacional  $t_y$ , su varianza y su varianza estimada están dados por:

$$\hat{t}_{y,p} = \frac{t_x}{m} \sum_{i=1}^m \frac{y_{ki}}{x_{ki}} \quad (4.2.7)$$

$$Var_{PPT}(\hat{t}_{y,p}) = \frac{1}{m} \sum_{k=1}^N p_k \left( \frac{y_k}{p_k} - t_y \right)^2 \quad (4.2.8)$$

$$\widehat{Var}_{PPT}(\hat{t}_{y,p}) = \frac{1}{m(m-1)} \sum_{i=1}^m \left( \frac{y_i}{p_i} - \hat{t}_{y,p} \right)^2 \quad (4.2.9)$$

respectivamente, con  $p_k$  dados por (4.2.2). 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.*

$$\begin{aligned} E \left( \frac{t_x}{m} \sum_{i=1}^m \frac{y_{ki}}{x_{ki}} \right) &= E \left( \frac{t_x}{m} \sum_U n_k(S) \frac{y_k}{x_k} \right) \\ &= \frac{t_x}{m} \sum_U E(n_k(S)) \frac{y_k}{x_k} \\ &= \frac{t_x}{m} \sum_U m \frac{x_k}{t_x} \frac{y_k}{x_k} = t_y \end{aligned}$$

dado que  $E(n(S)) = mp_k$ . Utilizando el resultado 2.2.13 y 2.2.14, se llega a la demostración de las varianzas.  $\square$

**Resultado 4.2.5.** Para el diseño de muestreo PPT, el estimador de Hansen-Hurwitz del total de la característica de información auxiliar reproduce ese total con varianza nula

*Demostración.* De la definición del estimador Hansen-Hurwitz, y de la expresión (4.2.2), se tiene que

$$\hat{t}_{x,p} = \frac{1}{m} \sum_{k \in S} \frac{x_k}{p_k} = \frac{1}{m} \sum_{k \in S} t_x = t_x$$

Por otro lado,

$$Var_{PPT}(\hat{t}_{y,p}) = \frac{1}{m} \sum_{k=1}^N p_k \left( \frac{x_k}{p_k} - t_x \right)^2 \quad (4.2.10)$$

$$= \frac{1}{m} \sum_{k=1}^N p_k (t_x - t_x)^2 = 0 \quad (4.2.11)$$

con lo cual se concluye la demostración  $\square$

**Resultado 4.2.6.** La varianza del estimador de Hansen-Hurwitz también puede ser escrita como

$$Var_{PPT}(\hat{t}_{y,p}) = \frac{1}{m} \sum_U \sum_{k < l} p_k p_l \left( \frac{y_k}{p_k} - \frac{y_l}{p_l} \right)^2 \quad (4.2.12)$$

*Demostración.* Desarrollando términos, se tiene que

$$\begin{aligned} \frac{1}{m} \sum_U \sum_{k < l} p_k p_l \left( \frac{y_k}{p_k} - \frac{y_l}{p_l} \right)^2 &= \frac{1}{2m} \sum_U \sum_{k,l} p_k p_l \left( \frac{y_k}{p_k} - \frac{y_l}{p_l} \right)^2 \\ &= \frac{1}{2m} \sum_{k \in U} p_k \sum_{l \in U} p_l \left( \frac{y_k}{p_k} - \frac{y_l}{p_l} \right)^2 \\ &= \frac{1}{2m} \sum_{k \in U} p_k \sum_{l \in U} \left( p_l \frac{y_k^2}{p_k^2} - 2 \frac{y_k y_l}{p_k} + \frac{y_l^2}{p_l} \right) \\ &= \frac{1}{2m} \sum_{k \in U} p_k \left( \frac{y_k^2}{p_k^2} - 2 \frac{y_k}{p_k} t_y + \sum_{l \in U} \frac{y_l^2}{p_l} \right) \\ &= \frac{1}{2m} \left( \sum_{k \in U} \frac{y_k^2}{p_k} - 2 t_y^2 + \sum_{l \in U} \frac{y_l^2}{p_l} \right) \\ &= \frac{1}{m} \left( \sum_{k \in U} \frac{y_k^2}{p_k} - t_y^2 \right) \\ &= \frac{1}{m} \sum_{k \in U} \left( \frac{y_k^2}{p_k} - p_k t_y^2 \right) \\ &= \frac{1}{m} \sum_{k \in U} p_k \left( \frac{y_k^2}{p_k^2} - 2 \frac{y_k}{p_k} t_y + t_y^2 \right) \\ &= \frac{1}{m} \sum_{k \in U} p_k \left( \frac{y_k}{p_k} - t_y \right)^2 \end{aligned}$$

y esta última expresión coincide con la varianza del estimador de Hansen-Hurwitz en muestreo PPT.  $\square$

Särndal, Swensson & Wretman (1992) afirman que la primera forma que toma la varianza y su estimación insesgada para el estimador de Hansen-Hurwitz es fácil de calcular computacionalmente. Sin embargo, la expresión alternativa de la varianza se utilizará para desarrollos teóricos posteriores.

Esta estrategia de muestreo es con reemplazo, y comparada con una estrategia de muestreo que utilice información auxiliar en la etapa de diseño con el estimador de Horvitz-Thompson es un poco menos eficiente. Sin embargo, en la práctica es más utilizada porque los cálculos computacionales son fáciles de realizar y es preferida porque con un número grande de elementos incluidos en la muestra, el cálculo de la varianza estimada del estimador de Horvitz-Thompson se hace inapropiado por la gran cantidad de productos cruzados.

Esta estrategia de muestreo es utilizada principalmente en la estimación de totales, como se verá más adelante surgen complicaciones, con respecto a la información auxiliar al usar un diseño de muestreo con reemplazo proporcional al tamaño en la estimación de razones. En encuestas de hogares, no resulta adecuado utilizar este diseño de muestreo, puesto que en una población, existe un número de hogares homogéneos por vivienda. Por otro lado, en encuestas de negocios y empresas es útil utilizar diseños proporcionales porque sí existen diferencias marcadas en los tamaños de las mismas; por ejemplo, en el número de empleados, el número de metros cuadrados en las instalaciones, el ingreso, etc. La función de varianza para esta estrategia de muestreo no es monótona decreciente; por la configuración de la información auxiliar, la varianza puede aumentar cuando aumenta el tamaño de muestra.

**Ejemplo 4.2.2.** Para nuestra población de ejemplo  $U$ , existen  $\binom{N+m-1}{m} = 20$  posibles muestras con reemplazo de tamaño  $m = 2$ . Utilizando la característica auxiliar  $x$ , realice el cálculo léxico-gráfico del estimador de Hansen-Hurwitz, compruebe el insesgamiento, calcule la varianza y el insesgamiento del estimador de la varianza.

### 4.2.3 Eficiencia de la estrategia

La regla de oro de una buena muestra reza que para que la inferencia basada en el diseño de muestreo arroje estimaciones que sean (abusando del lenguaje) de varianza mínima e insesgadas, las probabilidades de inclusión (o selección, según sea el caso) que arroje el diseño de muestreo utilizado deben ser directamente proporcionales a los valores que toma la característica de interés en la población. Raj (1954) demuestra el siguiente resultado que condiciona el comportamiento estructural de la información auxiliar que debe cumplir dos condiciones para que la eficiencia de la estrategia PPT sea mayor que la del diseño aleatorio simple con reemplazo.

**Resultado 4.2.7.** *La resta de la varianza de la estrategia aleatoria simple con reemplazo con la varianza de la estrategia PPT da como resultado la siguiente expresión:*

$$Var_{MRAS}(\hat{t}_{y,p}) - Var_{PPT}(\hat{t}_{y,p}) = \frac{N^2}{m} Cov\left(x, \frac{y^2}{x}\right) \quad (4.2.13)$$

*Demostración.* Utilizando la expresión general de la varianza (2.2.36) bajo cualquier diseño de muestreo

con reemplazo se tiene que

$$\begin{aligned}
 Var_{MRAS}(\hat{t}_{y,p}) - Var_{PPT}(\hat{t}_{y,p}) &= \frac{1}{m} \left[ N \sum_{k=1}^N y_k^2 - t_y^2 - t_x \sum_{k=1}^N \frac{y_k^2}{x_k} + t_y^2 \right] \\
 &= \frac{1}{m} \left[ \sum_{k=1}^N \frac{y_k^2}{x_k} (Nx_k - t_x) \right] \\
 &= \frac{N}{m} \left[ \sum_{k=1}^N \frac{y_k^2}{x_k} (x_k - \bar{x}) \right] \\
 &= \frac{N^2}{m} Cov\left(x, \frac{y^2}{x}\right)
 \end{aligned}$$

La última igualdad se tiene puesto que

$$\begin{aligned}
 NCov(x, w) &= \sum_{k=1}^N (x_k - \bar{x})(w_k - \bar{w}) \\
 &= \sum_{k=1}^N (x_k - \bar{x})w_k - \bar{w} \sum_{k=1}^N (x_k - \bar{x}) = \sum_{k=1}^N (x_k - \bar{x})w_k
 \end{aligned}$$

□

El anterior resultado indica que para que la estrategia de muestreo PPT sea más eficiente en términos de varianza que la estrategia de muestreo MRAS, además de que  $p_k \propto x_k$ , es necesario que la correlación entre  $\left(x, \frac{y^2}{x}\right)$  sea positiva. Nótese que si la razón entre  $y$  y  $x$  es constante e igual a  $C$ , se tiene que

$$\begin{aligned}
 Cor\left(x, \frac{y^2}{x}\right) &= Cor\left(x, y \frac{y}{x}\right) \\
 &= Cor(x, yC) \\
 &= Cor(x, y)
 \end{aligned}$$

Por tanto, una condición necesaria para que el diseño de muestreo PPT sea más eficiente que el diseño de muestreo MRAS es que exista una correlación positiva entre la característica de interés y la información auxiliar; pero, una condición suficiente para la optimalidad del diseño PPT, es que la razón  $\frac{y_k}{x_k}$  permanezca constante para todo  $k \in U$ .

Además de la razón constante, Lehtonen & Pahkinen (2003) muestran que la eficiencia del diseño de muestreo PPT está directamente relacionada con el siguiente modelo de regresión

$$y_k = \beta_0 + \beta_1 x_k + E_k \quad (4.2.14)$$

que relaciona la característica de interés con la información auxiliar. Concluye que para que el diseño de muestreo PPT sea más eficiente que el diseño de muestreo MRAS, la cantidad  $\beta_0$  debe ser pequeña. Es decir, que la línea de regresión ajuste cerca del origen. Es más, incluso si la correlación entre la característica de interés y la información auxiliar fuera perfecta e igual a uno, entonces no habría ningún término de error, pero aun así si  $\beta_0$  es grande, entonces la estrategia de muestreo PPT podría arrojar una eficiencia menor a la del diseño de muestreo aleatorio simple con reemplazo.

La eficiencia de la estrategia de muestreo, depende de dos aspectos. Primero, el tipo de parámetro que se quiere estimar. Lehtonen & Pahkinen (2003) afirman que para la estimación de totales, la estrategia de muestreo PPT, funciona mejor, en términos de eficiencia, que para la estimación de razones o medianas. Segundo, que la razón entre  $x_k$  y  $y_k$  sea constante para toda la población.

#### 4.2.4 Marco y Lucy

Una de las características del diseño de muestreo PPT es el uso de información auxiliar en la etapa de diseño. Obviamente, la información auxiliar debe estar presente en el marco de muestreo. En esta sección, de Marco y Lucy, seguiremos la tendencia que comenzamos en el diseño de muestreo Poisson. Suponga que, para todas las empresas del sector industrial, el valor del ingreso en el último año fiscal está disponible en el marco de muestreo.

Se quiere estimar, el total poblacional de las características de interés Empleados e Impuestos, para lo cual, se utilizará una estrategia de muestreo que utiliza un diseño de muestreo con reemplazo y probabilidades de selección de las empresas proporcionales al tamaño de la característica auxiliar Ingreso junto con el estimador de Hansen-Hurwitz. Como se vio antes, para que esta estrategia de muestreo sea óptima con respecto a una que utilice un diseño aleatorio simple con reemplazo se deben cumplir ciertas condiciones. Antes de analizarlas, veamos que, para este caso particular y con un tamaño de muestra igual a  $m = 2000$ , el diseño de muestreo PPT es menos eficiente que el muestreo simple con reemplazo para la estimación del total de empleados, aunque es más eficiente que el muestreo simple con reemplazo para la estimación del total de impuestos declarados. Lo anterior se tiene utilizando la expresión (4.2.13) escrita en código de R.

```
data(BigLucy)
attach(BigLucy)

N <- nrow(BigLucy)
m <- 2000

(N^2 / m) * cov(Income, (Employees^2 / Income))

## [1] -9477162876

(N^2 / m) * cov(Income, (Taxes^2 / Income))

## [1] 897321919
```

Primero, que la correlación entre `Income` y `y2/Income` sea positiva. Aunque la correlación entre `Income` y `Employees` e, `Income` y `Taxes` sea positiva, se debe verificar que la correlación entre `Income` y la nueva variable `Employees2/Income` sea positiva, como también la correlación entre `Income` y `Taxes2/Income`. Mediante el uso de la función `cor` que R incorpora en su ambiente de trabajo, se tiene que para la característica de interés Empleados, la correlación es negativa, aunque casi nula. Mientras que para la característica de interés Impuestos, la correlación buscada es positiva. Esto indica que para la estimación del total de empleados, el uso de la información auxiliar no conlleva a ganancias significativas en la eficiencia de la estrategia. Por otro lado, para la estimación del total de impuestos declarados, sí se tiene un ganancia significativa.

```
cor(Income, (Employees^2 / Income))

## [1] -0.07764
```

```
cor(Income, (Taxes^2/Income))

## [1] 0.7087
```

Otra de las condiciones para la optimalidad de la estrategia es que el cociente entre `Income` y las características de interés `Taxes` y `Employees` sea constante para todo elemento de la población. Mediante el uso de la función `plot` es posible tener un acercamiento gráfico al comportamiento de los respectivos cocientes. Nótese que la función `abline` permite trazar una línea sobre el promedio de los cocientes.

La figura 4.3 muestra que la relación existente entre el cociente `Income` y `Employees` es uniforme en casi toda la población. Por supuesto, se observan algunos datos atípicos que están muy lejos de la línea de referencia, pero en general se observa un comportamiento homogéneo. Esto no ocurre con la relación existente entre el cociente `Income` e `Taxes` donde existe un comportamiento más disperso para todos los elementos de la población. A pesar de lo anterior, se puede afirmar que el comportamiento de la razón es constante.

Un tercer argumento para el uso de la estrategia de muestreo PPT es el examen del ajuste de una línea de regresión entre `Employees` con `Income` y `Taxes` con `Income` respectivamente. Para esto, se ajustan dos modelos. El primero dado por

$$Impuestos_k = \beta_0 + \beta_1 Ingreso + E_k \quad (4.2.15)$$

Para la estimación del total de la característica Impuestos y, el segundo dado por

$$Empleados_k = \beta_0 + \beta_1 Ingreso + E_k \quad (4.2.16)$$

Para la estimación del total de la característica Empleados. Para los modelos anteriores, nos interesa conocer el valor que toma el intercepto de cada línea de regresión. Si el intercepto  $\beta_0$  es cercano a cero, entonces se ha ganado eficiencia al utilizar un diseño de muestreo PPT. R incorpora la función `lm` para el ajuste de modelos lineales. Las estimaciones de  $\beta_0$  y  $\beta_1$  se hacen por medio del método de los mínimos cuadrados. Un análisis de regresión de y contra x es especificado mediante y  $\tilde{x}$ . La salida de la función `lm` está dada por las estimaciones de los coeficientes de los modelos de regresión. Con ayuda de la función `summary` es posible extraer más información respecto a la inferencia de las estimaciones.

```
M.I <- lm(Taxes ~ Income)
summary(M.I)

##
## Call:
## lm(formula = Taxes ~ Income)
##
## Residuals:
##     Min      1Q Median      3Q     Max
## -5.58  -3.99  -1.60   2.62 169.65
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)    
## (Intercept) -13.6780825  0.0447706  -306 <0.0000000000000002 ***
## Income       0.0593729  0.0000886    670 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
par(mfrow=c(1,2))

plot(Employees / Income)
abline(h = mean(Employees / Income), col = 2)

plot(Taxes / Income)
abline(h = mean(Taxes / Income), col = 2)
```

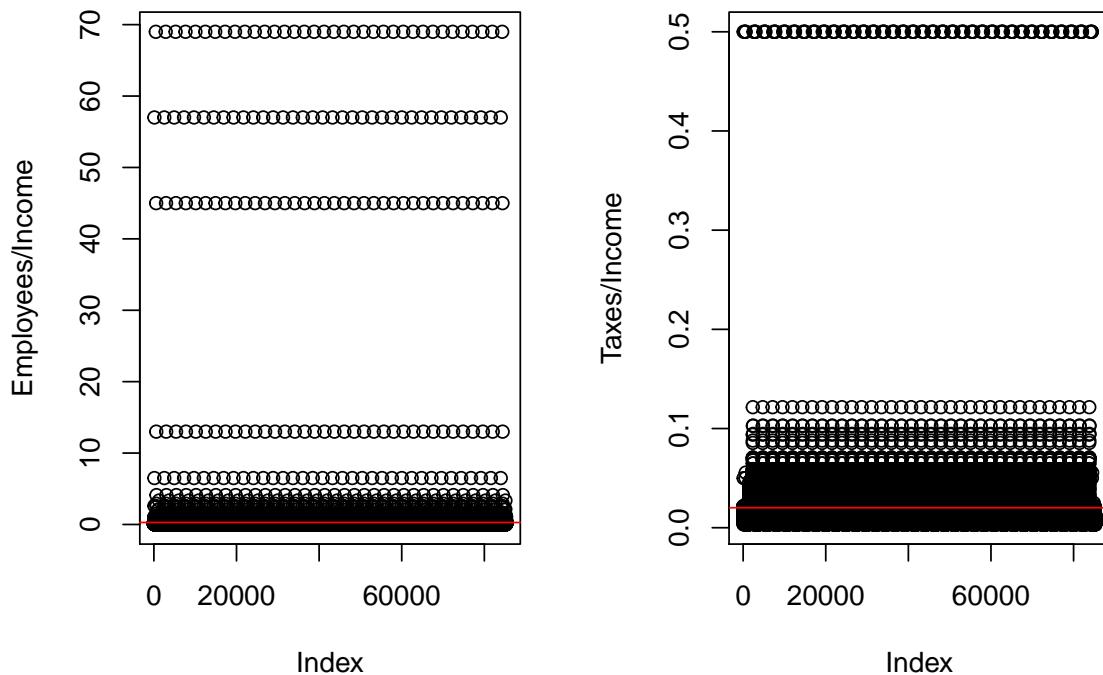


Figura 4.2: Comportamiento del cociente de la información auxiliar con las características de interés.

```
## Residual standard error: 6.88 on 85294 degrees of freedom
## Multiple R-squared:  0.84, Adjusted R-squared:  0.84
## F-statistic: 4.49e+05 on 1 and 85294 DF,  p-value: <0.0000000000000002
```

Para el primer modelo, se nota que la estimación del intercepto está dada por  $-13.6781$  y, a juzgar por las tres estrellas, es una cantidad significativa. Aunque para nuestro análisis está cerca del origen, por tanto se gana en eficiencia al utilizar esta estrategia de estimación para el total poblacional de la característica de interés Impuestos.

```
M.E <- lm(Employees ~ Income)
summary(M.E)

##
```

```

## Call:
## lm(formula = Employees ~ Income)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -46.35 -21.99   0.31  21.36  82.19 
## 
## Coefficients:
##             Estimate Std. Error t value     Pr(>|t|)    
## (Intercept) 29.124429  0.163386   178 <0.0000000000000002 *** 
## Income       0.079373  0.000323    245 <0.0000000000000002 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 25.1 on 85294 degrees of freedom 
## Multiple R-squared:  0.414, Adjusted R-squared:  0.414 
## F-statistic: 6.02e+04 on 1 and 85294 DF,  p-value: <0.0000000000000002

```

El intercepto del segundo modelo ha sido estimado como 29.1244, a diferencia del modelo anterior, no se puede decir que está cerca del origen. Además, por la magnitud de la escala de medición de las características, se puede decir que es una cantidad importante y no despreciable.

La figura 4.3 muestra la línea de regresión ajustada para los dos modelos anteriores; es claro que el intercepto del modelo con impuestos declarados se puede considerar nulo, pero el intercepto del modelo con número de empleados es grande. Los tres anteriores argumentos permiten estar confiados al utilizar la estrategia de muestreo PPT para la estimación del total de impuestos declarados, pero se sabe que para la estimación del total de número de empleados, este diseño muestral no es más eficiente que el diseño simple con reemplazo.

Una vez se ha decidido usar la estrategia de muestreo PPT, es necesario seleccionar la muestra. En este caso, se ha querido utilizar el mismo tamaño de muestra, que en las anteriores estrategias de muestreo. En primer lugar, se adjunta el marco de muestreo que no sólo contiene la ubicación e identificación sino además el valor de la información auxiliar *Ingreso* para cada empresa del sector industrial. La selección de la muestra se hace mediante el uso de la función *S.PPS* para la cual los argumentos introducido son *m* = 2000 junto con la información auxiliar *Income*. Esta función utiliza el algoritmo de selección acumulativo total.

```

pk <- Income / sum(Income)
sam <- S.PPS(m, Income)
muestra <- BigLucy[sam,]
attach(muestra)

head(muestra)

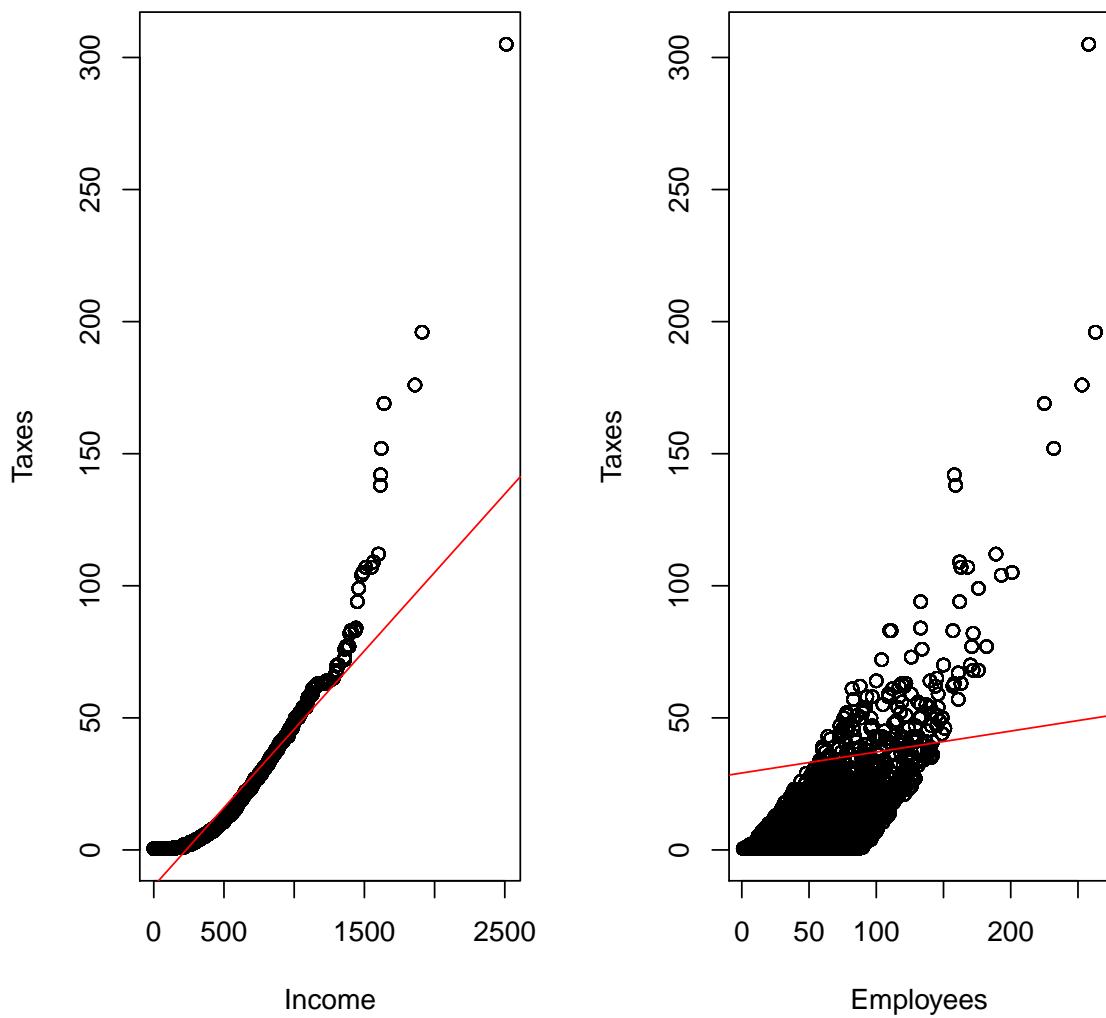
##           ID Ubication Level Zone Income Employees Taxes
## 78383 AB0000078383 C0175657K0126240 Medium County9    500      91     12
## 35440 AB0000035440 C0056667K0245230 Medium County47    530      69     13
## 80432 AB0000080432 C0296907K0004990 Small County93    214      38      1
## 21382 AB0000021382 C0200702K0101195 Medium County33    879     103     38
## 56293 AB0000056293 C0121122K0180775 Small County63    458      55      9
## 9575 AB0000009575 C0118488K0183409   Big County23   1360     104     72
##          SPAM ISO Years Segments

```

```
par(mfrow=c(1,2))

plot(Income, Taxes)
abline(M.I, col="red")

plot(Employees, Taxes)
abline(M.E, col="red")
```

Figura 4.3: *Líneas de regresión.*

```
## 78383 no yes 46.7 County9 11
## 35440 no yes 30.1 County47 51
## 80432 no no 7.4 County93 52
```

```
## 21382 yes yes 39.1 County33 76
## 56293 no no 17.4 County63 119
## 9575 yes yes 41.2 County23 82
```

El método acumulativo total no tiene en cuenta ningún ordenamiento. En este caso particular, la última empresa en ser seleccionada fue la empresa con número de identificación AB0000078383, aunque esta empresa ya había sido seleccionada en la muestra en dos ocasiones. Es decir, fue seleccionada en tres ocasiones.

Una vez seleccionada la muestra con reemplazo, se utiliza la función `E.PPS` del paquete `TeachingSampling` cuyos argumentos son la(s) característica(s) de interés y un vector de probabilidades de selección `pk`. Por supuesto, el vector de probabilidades de selección en la población está dado por `pk <- Income / sum(Income)`. Sin embargo, en la función `E.PPS`, el vector de probabilidades debe corresponder a las probabilidades de selección de cada uno de los elementos elegidos en la muestra. En este caso la longitud del vector `pk.s` es de `m = 2000`.

```
pk.s <- pk[sam]
estima <- data.frame(Income, Employees, Taxes)
E.PPS(estima, pk.s)
```

Los resultados de aplicar la estrategia de muestreo son muy favorables. Nótese, que a diferencia de la estrategia de muestreo Poisson, el total poblacional de la característica auxiliar ingreso, es estimada exactamente con varianza casi nula. El total poblacional de las características de interés Empleados e Impuestos tienen coeficientes de variación menores a 2 %. La tabla 4.2 muestra los resultados obtenidos en este ejercicio particular.

Cuadro 4.2: Muestreo PPT: estimación de los totales de las características de interés.

	N	Income	Employees	Taxes
Estimation	80970.23	36634733.00	5274355.17	1048536.36
Standard Error	1371.31	0.00	80797.15	13855.51
CVE	1.69	0.00	1.53	1.32
DEFF	Inf	0.00	1.23	0.07

Asimismo, una estrategia alternativa es utilizar un diseño de muestreo con reemplazo y probabilidad de selección proporcional al tamaño junto con el estimador de Horvitz-Thompson, el cual es también insesgado. Särndal, Swensson & Wretman (1992) se preguntan cuál es el mejor estimador y llegan a la conclusión que dependiendo de la configuración de los valores de las características de interés y de información auxiliar un estimador tendrá menor varianza que el otro. Por tanto, no es posible generalizar. De lo que sí se puede estar seguro, es de la simplicidad, en materia de cálculos del estimador de Horvitz-Thompson. En la práctica, este es un argumento muy fuerte que incentiva el uso del estimador de Hansen-Hurwitz.

Utilizando el resultado 4.2.3., es posible estimar los parámetros de interés mediante el uso del estimador de Horvitz-Thompson. Para esto, se calculan las probabilidades inclusión. Nótese que la suma de éstas es de 358. Se extraen las probabilidades de inclusión de los elementos en la muestra y se utiliza la forma genérica del estimador de Horvitz-Thompson.

```
pik <- 1 - (1 - pk)^2000
sum(pik)

## [1] 1968
```

```
pik.s <- pik[sam]
sum(1 / pik.s)

## [1] 81975

colSums(estima/pik.s)

##      Income Employees      Taxes
## 37253068   5353738   1071778
```

Las estimaciones resultantes no son mejores, en el sentido práctico, a las obtenidas mediante el uso del estimador de Hansen-Hurwitz. Ahora, la estimación de la varianza supondría un esfuerzo computacional demasiado grande.

### 4.3 Diseño de muestreo $\pi$ PT

Como se vio en la sección anterior, utilizar un esquema de muestreo con probabilidades proporcionales a alguna característica de información auxiliar puede resultar en ganancia de precisión. Sin embargo, utilizar una estrategia de muestreo que contemple un diseño de muestreo con reemplazo es menos eficiente que implementar una estrategia de muestreo que contemple un diseño de muestreo sin reemplazo y de tamaño muestral fijo.

En la sección anterior, se utilizó un diseño de muestreo con probabilidades proporcionales, con reemplazo y, sin embargo, arrojó muy buenos resultados en términos de eficiencia comparado con los diseños de muestreo de probabilidades simples. Esta sección se concentra en la implementación de diseños de muestreo con probabilidades de inclusión proporcionales a una característica de interés y cuya estructura general sea sin reemplazo. De esta forma, es posible aumentar dramáticamente la eficiencia de la estrategia que involucra al estimador de Horvitz-Thompson.

Lohr (2000) afirma que el muestreo de probabilidades simples, proporciona esquemas que, frecuentemente, son fáciles de explicar y diseñar. Sin embargo, estos esquemas no siempre pueden ser realizados puesto que las probabilidades simples no siempre reflejan el comportamiento de la característica de interés en la población.

Este diseño de muestreo induce probabilidades de inclusión proporcionales al tamaño de una característica de información auxiliar<sup>1</sup>. De esta manera, se supone que el marco de muestreo tiene la bondad de poseer información auxiliar de tipo continuo y positiva disponible para todo elemento perteneciente a la población finita. Asimismo, el diseño de muestreo  $\pi$ PT<sup>2</sup>, de tamaño de muestra fijo e igual a  $N$ , se basa en la construcción de probabilidades de inclusión que obedezcan la siguiente relación:

$$\pi_k = \frac{nx_k}{t_x} \quad 0 < \pi_k \leq 1 \quad (4.3.1)$$

Además se busca que:

- El algoritmo de selección de muestras bajo este diseño sea de fácil implementación computacional.
- Las probabilidades de inclusión de segundo orden sean positivas,  $\pi_{kl} > 0$ . De lo contrario el estimador de la varianza podría ser sesgado.

<sup>1</sup>El requisito indispensable de la información auxiliar es que sea aproximadamente proporcional a la característica de interés.

<sup>2</sup>Nótese que la sigla  $\pi$ PT se refiere a los diseños de muestreo que inducen probabilidades de inclusión proporcionales a una característica de información auxiliar.

- El cálculo de estas probabilidades de inclusión de segundo orden,  $\pi_{kl}$ , sea sencillo.
- $\Delta_{kl} < 0 \quad \forall k \neq l$  para que la estimación de la varianza no sea negativa.

Este diseño de muestreo se puede considerar como una generalización de la mayoría de diseños de muestreo sin reemplazo. Por ejemplo: si la característica de información auxiliar es constante e igual a  $C$ , entonces para un tamaño de muestra fijo, las probabilidades de inclusión de primer orden estarían dadas por:

$$\begin{aligned}\pi_k &= \frac{n x_k}{t_x} \\ &= \frac{nC}{NC} = \frac{n}{N}\end{aligned}$$

Con lo que se tiene un diseño de muestreo caracterizado por probabilidades simples. En ciertas ocasiones, cuando la población tiene un comportamiento muy variable, irregular y sesgado, algunas de las  $p_{ik}$  inducidas por la expresión (4.3.1) pueden ser mayores a uno para ciertos elementos. En tal caso, estos elementos son incluidos en todas las posibles muestras y toman el nombre de **elementos de inclusión forzosa**. Sin embargo, para calcular la probabilidad de inclusión de los elementos restantes, se debe excluir estos elementos de inclusión forzosa y volver a calcular las probabilidades de inclusión mediante una reformulación de la expresión (4.3.1) dada por

$$\pi_k = \frac{(n - n^*)x_k}{\sum_{k \in U^*} x_k} \quad 0 < \pi_k \leq 1; \quad k \in U^* \quad (4.3.2)$$

donde  $n^*$  corresponde al número de elementos de inclusión forzosa y  $U^*$  la población finita excluyendo a estos elementos de inclusión forzosa. Al final del proceso, deberían existir dos grupos de elementos:

1. Un grupo de elementos de inclusión forzosa con probabilidades de inclusión iguales a uno.
2. Un grupo de elementos con probabilidades de inclusión  $0 < \pi_k < 1$  y proporcionales a  $x_k$ .

Por tanto, el problema se reduce a la selección de  $n$  unidades con probabilidades de inclusión tales que

$$\sum_{k \in U} \pi_k = n$$

El siguiente resultado da cuenta de la forma estructural que toma el estimador de Horvitz-Thompson, de su varianza y de su varianza estimada.

**Resultado 4.3.1.** Para el diseño de muestreo  $\pi$ PT, el estimador de Horvitz-Thompson, su varianza y su varianza estimada están dados por:

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} \quad (4.3.3)$$

$$Var_{\pi PT}(\hat{t}_{y,\pi}) = -\frac{1}{2} \sum_U \sum_{kl} \Delta_{kl} \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \quad (4.3.4)$$

$$\widehat{Var}_{\pi PT}(\hat{t}_{y,\pi}) = -\frac{1}{2} \sum_S \sum_{kl} \frac{\Delta_{kl}}{\pi_{kl}} \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \quad (4.3.5)$$

**Resultado 4.3.2.** Para el diseño de muestreo  $\pi$ PT, el estimador de Horvitz-Thompson del total de la característica de información auxiliar reproduce ese total con varianza nula

*Demostración.* De la definición del estimador de Horvitz-Thompson, y de la expresión (4.3.1), se tiene que

$$\hat{t}_{x,\pi} = \sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in S} t_x \frac{1}{n} = t_x$$

Por otro lado,

$$Var_{\pi PT}(\hat{t}_{x,\pi}) = -\frac{1}{2} \sum_U \sum_{kl} \Delta_{kl} \left( \frac{x_k}{\pi_k} - \frac{x_l}{\pi_l} \right)^2 \quad (4.3.6)$$

$$= -\frac{1}{2} \sum_U \sum_{kl} \Delta_{kl} \left( \frac{t_x}{n} - \frac{t_x}{n} \right)^2 = 0 \quad (4.3.7)$$

con lo cual se concluye la demostración  $\square$

**Ejemplo 4.3.1.** Suponga que para la población de ejemplo  $U$  se tiene conocimiento de cada valor de la siguiente característica de información auxiliar correlacionada con la característica de interés. Por tanto, un primer paso para el cálculo de las probabilidades de inclusión es aplicar la expresión (4.3.1).

```
n <- 4
x <- c(52, 60, 75, 100, 50)
pik <- n * x / sum(x)
pik

## [1] 0.6172 0.7122 0.8902 1.1869 0.5935
```

Nótese que el cuarto elemento de la población, correspondiente a **Sharon** es un elemento de inclusión forzosa; es decir que está presente en todas las posibles muestras. El siguiente paso es separar a **Sharon** de los restantes elementos y proseguir con el cálculo de las probabilidades de inclusión inducidas por la expresión (4.3.2)

```
n <- 3
x <- c(52, 60, 75, 50)
pik <- n * x / sum(x)
pik

## [1] 0.6582 0.7595 0.9494 0.6329
```

Por tanto el vector de probabilidades de inclusión para toda la población  $U$  está dado por

$$\boldsymbol{\pi} = (\underbrace{0.6582278}_{\text{Yves}}, \underbrace{0.7594937}_{\text{Ken}}, \underbrace{0.9493671}_{\text{Erik}}, \underbrace{1.0000}_{\text{Sharon}}, \underbrace{0.6329114}_{\text{Leslie}})'$$

## 4.4 Selección de muestras $\pi$ PT

Existen varios métodos de selección de muestras  $\pi$ PT. Sin embargo, todos ellos están basados en una teoría fuerte y complicada y, en algunas ocasiones, son muy difíciles de implementar en la práctica. A continuación, se exponen dos métodos de selección de muestras de tamaño  $n = 1$  y  $n = 2$ . Särndal, Swensson & Wretman (1992) comentan que a simple vista parecería irreal considerar tamaños de muestra tan pequeños. Sin embargo, en muestreo estratificado y muestreo para conglomerados (ver siguientes capítulos) tiene sentido seleccionar solamente una o dos unidades primarias de muestreo.

**Tamaño de muestra  $n = 1$** 

Para  $n = 1$  se utiliza el método acumulativo total, que consiste en:

1. Definir  $T_0 = 0$  y  $T_k = T_{k-1} + x_k$  ( $k \in U$ ).
2. Calcular un número aleatorio  $\varepsilon$  con distribución uniforme en el intervalo  $[0, 1]$ .
3. Si  $T_{k-1} < \varepsilon T_N < T_k$ , el elemento  $k$ -ésimo se selecciona.

Nótese que este algoritmo de selección garantiza que el diseño de muestreo es un autentico  $\pi$ PT puesto que

$$\pi_k = Pr(k \in S) = Pr(T_{k-1} < \varepsilon T_N < T_k) = \frac{T_k - T_{k-1}}{T_N} = \frac{x_k}{t_x}$$

Por supuesto, no es posible obtener un estimador insesgado de la varianza del estimador de Horvitz-Thompson puesto que la muestra sólo considera la inclusión de un elemento de la población finita.

**Tamaño de muestra  $n = 2$** 

En este escenario es preciso garantizar que las probabilidades de inclusión de primer orden estén dadas por

$$\pi_k = \frac{2x_k}{t_x}$$

para todo elemento de la población finita. En este caso, los dos elementos de la muestra son seleccionados uno por uno. Para tal fin, se debe seguir el siguiente algoritmo (Brewer 1963, Brewer 1975) que utiliza el método acumulativo total en cada una de las dos selecciones, así:

1. En la primera extracción, el elemento  $k$ -ésimo es seleccionado con probabilidad

$$p_k = \frac{c_k}{\sum_{k \in U} c_k}$$

donde

$$c_k = \frac{x_k(T_N - x_k)}{T_N(T_N - 2x_k)}$$

2. En la segunda extracción, el elemento seleccionado en el paso anterior, digamos el elemento  $k^*$ , es retirado del sorteo. El segundo elemento es seleccionado con probabilidad

$$p_l|_{k^*} = \frac{x_l}{T_N - x_{k^*}}$$

**Resultado 4.4.1.** Bajo el esquema de selección de Brewer las probabilidades de inclusión de primer orden satisfacen la siguiente relación

$$\pi_k = \frac{2x_k}{t_x}$$

Las probabilidades de inclusión de segundo orden están dadas por

$$\pi_{kl} = \frac{2x_k x_l}{T_N(\sum_{k \in U} c_k)} \frac{T_N - x_k - x_l}{(T_N - 2x_k)(T_N - 2x_l)}$$

*Demostración.* La probabilidad de inclusión de primer orden del  $k$ -ésimo elemento está dada por

$$\begin{aligned} \pi_k &= Pr(k \in S) \\ &= Pr(k \text{ sea seleccionado en la primera extracción}) \\ &\quad + Pr(k \text{ sea seleccionado en la segunda extracción}) \\ &= p_k + p_{k|j} \sum_{\substack{j \in U \\ j \neq k}} p_j \\ &= \frac{x_k(T_N - x_k)/T_N(T_N - 2x_k)}{D} \\ &\quad + \sum_{\substack{j \in U \\ j \neq k}} \frac{x_j(T_N - x_j)/T_N(T_N - 2x_j)}{D} \frac{x_k}{T_N - x_j} \\ &= \frac{x_k/T_N}{D} \left( \frac{T_N - x_k}{T_N - 2x_k} + \sum_{\substack{j \in U \\ j \neq k}} \frac{x_j}{T_N - 2x_j} \right) \\ &= \frac{x_k/T_N}{D} \left( \frac{T_N}{T_N - 2x_k} - \frac{2x_k}{T_N - 2x_k} + \sum_{j \in U} \frac{x_j}{T_N - 2x_j} \right) \\ &= \frac{x_k/T_N}{D} \left( 1 + \sum_{j \in U} \frac{x_j}{T_N - 2x_j} \right) = \frac{x_k/T_N}{D} (2D) = \frac{2x_k}{T_N} \end{aligned}$$

Donde

$$\begin{aligned} D &= \sum_{k \in U} \frac{x_k(T_N - x_k)}{T_N(T_N - 2x_k)} \\ &= \frac{1}{2} \sum_{k \in U} \frac{x_k(2T_N - 2x_k)}{T_N(T_N - 2x_k)} \\ &= \frac{1}{2} \left( 1 + \sum_{k \in U} \frac{x_k}{T_N - 2x_k} \right) \end{aligned}$$

La última relación se tiene puesto que

$$\sum_{k \in U} \frac{x_k(T_N - x_k)}{T_N(T_N - 2x_k)} - \sum_{k \in U} \frac{x_k}{T_N - 2x_k} = 1$$

Análogamente para las probabilidades de inclusión de segundo orden.  $\square$

**Resultado 4.4.2.** Bajo muestreo  $\pi PT$  con el algoritmo de selección de Brewer, se tiene que.

1.  $Var_{\pi PT}(\hat{t}_{y,\pi})$  es menor que  $Var_{PPT}(\hat{t}_{y,p})$ .
2. La estimación de la varianza es siempre positiva.

Lohr (2000) afirma que generalmente el muestreo con reemplazo es menos eficiente que el muestreo sin reemplazo. Sin embargo, el muestreo con reemplazo se utiliza con mucha más frecuencia debido a la facilidad que brinda para elegir y analizar las muestras. Se ha investigado mucho acerca del muestreo con probabilidades proporcionales sin reemplazo; hay que notar que la teoría de estos tipos de muestreo es mucho más complicada. Existen varios algoritmos que permiten la selección de muestras de tamaño  $n > 2$  con probabilidades de inclusión desiguales; en particular, con probabilidades proporcionales a una característica de información auxiliar<sup>3</sup>. En esta sección, revisaremos algunos de estos esquemas que permiten la selección de muestras para tamaños de muestra fijos y mayores que dos.

#### 4.4.1 Método de Sunter

En Sunter (1977) y en Sunter (1986) se propone un procedimiento secuencial que, en general, no es aplicable a cualquier vector de probabilidades de inclusión de primer orden. Este algoritmo de muestreo sólo funciona cuando los elementos de la población son ordenados descendente y cuando los elementos con valores más pequeños comparten las mismas probabilidades de inclusión. Este método, que en realidad es una modificación del algoritmo de Fan-Muller-Rezucha para la selección de muestras simples, asume la existencia de una variable auxiliar que induce probabilidades de inclusión de primer orden dadas por la expresión (4.3.1) y consiste en:

1. Ordenar descendente la población de acuerdo con los valores que toma la característica de información auxiliar  $x_k$ .
2. Realizar  $\xi_k \sim U(0, 1)$ .
3. Para  $k = 1$ , el primer elemento de la lista ordenada es incluido en la muestra sí y solamente si  $\xi_1 < \pi_1$ .
4. Para  $k \geq 2$ , el  $k$ -ésimo elemento de la lista ordenada es incluido en la muestra sí y solamente si

$$\xi_k \leq \frac{n - n_{k-1}}{n - \sum_{i=1}^{k-1} \pi_i} \pi_k$$

donde  $n_{k-1}$  representa el número de elementos que ya han sido seleccionados al final del paso  $k - 1$ .

**Resultado 4.4.3.** *Bajo el esquema de selección de Sunter, las probabilidades de inclusión de primer orden están dadas por*

$$\pi_k = \begin{cases} \frac{nx_k}{T_N} & \text{si } k = 1, \dots, k^* - 1 \\ \frac{n\bar{x}_{k^*}}{T_N} & \text{si } k = k^*, \dots, N \end{cases}$$

donde  $k^* = \min\{k_0, N - n + 1\}$  con  $k_0$  equivalente al menor  $k$  para el cual se cumple que  $nx_k/T_k > 1$ ,  $T_k = \sum_{j=1}^k x_j$  y

$$\bar{x}_{k^*} = \frac{T_{k^*}}{N - k^* + 1}$$

Por otra parte, se cumple que para todo  $k \neq l$ ,  $\pi_{kl} > 0$  y  $\Delta_{kl} < 0$ .

<sup>3</sup>El lector interesado en conocer aún más acerca de estos algoritmos de selección puede referirse a los siguientes tres libros: Brewer & Hanif (1983), Hájek (1981) y Tillé (2006)).

Con el anterior resultado se establece que este método de selección de muestras no induce probabilidades de inclusión estrictamente proporcionales a la característica de información auxiliar. Särndal, Swensson & Wretman (1992) afirman que relajar un poco este supuesto es un precio menor que debe pagarse para que el esquema de selección sea ejecutable en la práctica.

**Ejemplo 4.4.1.** Volviendo con la población ejemplo  $U$ . Suponga que se tiene acceso a los valores de la característica de información auxiliar  $x$  para todos los elementos de la población. Es posible seleccionar una muestra  $\pi$ PT de tamaño  $n = 3$  con el método de Sunter. Para tal fin, es necesario recurrir a la función **S.piPS** del paquete **TeachingSampling**.

Esta función consta de tres argumentos: el primero, **x**, hace referencia al vector de información auxiliar continua para toda la población. El segundo, **n**, determina el tamaño de la muestra. Con estos dos argumentos, la función **S.piPS** construye las probabilidades de inclusión proporcionales a la característica de información auxiliar. El tercer argumento, **e**, que es opcional, corresponde a un vector de números aleatorios con el que se procede a ejecutar el esquema de selección de Sunter.

```

U <- c("Yves", "Ken", "Erik", "Sharon", "Leslie")
N <- length(U)
n <- 3
x <- c(52, 60, 75, 100, 50)
pik <- (n*x)/sum(x)

pik

## [1] 0.4629 0.5341 0.6677 0.8902 0.4451

sum(pik)

## [1] 3

sam <- S.piPS(n, x, e=runif(N))
U[sam]

## [1] "Sharon" "Erik"   "Ken"

x[sam]

## [1] 100  75  60

```

La función **S.piPS** devuelve un conjunto de índices (distintos por definición) que aplicados a los rótulos poblacionales proporcionan la muestra realizada o seleccionada. Para el anterior ejercicio particular, la muestra realizada estuvo conformada por **Sharon**, **Erik** y **Ken**. Es importante recalcar que esta función no necesita de ningún ordenamiento previo sobre la característica de información auxiliar; en otras palabras, los resultados serán idénticos si se realiza un ordenamiento previo o si no se realiza tal ordenamiento.

#### 4.4.2 Método de escisión

Desde la publicación de Brewer & Hanif (1983) se han propuesto numerosas técnicas de muestreo con probabilidades de inclusión desiguales. Sin embargo, en el artículo de Deville & Tillé (1998), se habla de ocho nuevos métodos; entre ellos, el método de escisión. Este método es considerado como un

nuevo enfoque que presenta de manera más simple los restantes métodos de selección de muestras con probabilidades desiguales. Tillé (2006) comenta que el método de escisión es un medio para integrar la presentación de los demás métodos y para hacerlos comparables.

En palabras de uno de los autores (Tillé 2006), el método de escisión propuesto por Deville & Tillé (1998) es:

...un marco de referencia de los métodos de muestreo sin reemplazo, con tamaño muestral fijo y con probabilidades desiguales, en particular con probabilidades proporcionales al tamaño de una característica de información auxiliar.

La idea básica del método consiste en dividir el vector de probabilidades de inclusión en dos o más vectores nuevos. A continuación, uno de estos vectores se selecciona aleatoriamente, de tal manera que el promedio de los vectores de como resultado el vector de probabilidades de inclusión. Este simple paso se repite hasta que se obtenga una muestra.

Con el planteamiento anterior, el método de escisión se puede considerar como un algoritmo de Martingalas que incluye todos los procedimientos de selección individual y secuencial y permite derivar un gran número de algoritmos de muestreo de probabilidades desiguales. Más aun, muchos procedimiento bien conocidos de probabilidades desiguales pueden ser formulados bajo la forma de una partición del vector de probabilidades de inclusión. Por tanto, la presentación puede ser estandarizada, lo cual permite una comparación más simple de procedimientos.

### Escisión en dos partes

Este método consiste en seleccionar una muestra, de tamaño  $n(S) = n$ , de probabilidades desiguales mediante la partición de la probabilidad de inclusión del  $k$ -ésimo elemento en dos partes  $\pi_k^a$  y  $\pi_k^b$  tal que

$$\pi_k = \lambda\pi_k^a + (1 - \lambda)\pi_k^b \quad (4.4.1)$$

De tal forma que  $0 \leq \pi_k^a \leq 1$  y  $0 \leq \pi_k^b \leq 1$  y que

$$\sum_{k \in U} \pi_k^a = \sum_{k \in U} \pi_k^b = n \quad (4.4.2)$$

Donde  $0 < \lambda < 1$ . La esencia del método es la selección de  $n$  elementos con probabilidades desiguales mediante la transformación iterativa del vector de probabilidades de inclusión. Si la escisión es tal que uno o varios de los  $\pi_k^a$  y de los  $\pi_k^b$  son equivalentes a cero o uno, entonces el problema de muestreo se verá reducido en el siguiente paso. De hecho, un vez que un componente del vector de probabilidades de inclusión converja a cero o uno, es deberá permanecer en este estado hasta que se seleccione una muestra<sup>4</sup>. En general, el algoritmo de muestreo de este esquema es el siguiente:

1. Definir  $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$ .
2. Construir un par de vectores  $\boldsymbol{\pi}^a(t)$  y  $\boldsymbol{\pi}^b(t)$  y definir un número  $\lambda(t) \in (0, 1)$  tales que

$$\boldsymbol{\pi}(t) = \lambda(t)\boldsymbol{\pi}^a(t) + (1 - \lambda(t))\boldsymbol{\pi}^b(t) \quad (4.4.3)$$

---

<sup>4</sup>Una muestra es seleccionada cuando todas las entradas del vector de probabilidades de inclusión se conviertan en ceros o unos.

3. Definir para el siguiente paso al vector de probabilidades de inclusión de tal forma que

$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}^a(t) & \text{con probabilidad } \lambda(t) \\ \boldsymbol{\pi}^b(t) & \text{con probabilidad } 1 - \lambda(t) \end{cases} \quad (4.4.4)$$

4. Iterar hasta obtener convergencia; es decir, hasta que todas las entradas del vector de probabilidades de inclusión sean cero o uno en ambas particiones. De esta forma, para cada tiempo  $t$ , existe una posible muestra correspondiente a  $S = \boldsymbol{\pi}(t)$ .

### Esquema de soporte mínimo

**Definición 4.4.1.** Si para un vector fijo de probabilidades de inclusión es posible plantear un diseño de muestreo cuyo soporte contenga a lo más  $N$  muestras  $s$ , tales que  $p(s) > 0$ . En tal caso, el diseño de muestreo se dice de soporte mínimo.

A continuación se presenta el esquema de soporte mínimo que permite seleccionar una muestra en a lo más  $N$  pasos.

Paso 1 Ordenar el vector de probabilidades de inclusión en orden ascendente, denotado como  $(\pi_{(1)}, \dots, \pi_{(k)}, \dots, \pi_{(N)})$

Paso 2 (Primera iteración,  $t = 1$ ) Calcular

$$\lambda(1) = \min\{1 - \pi_{(N-n)}, \pi_{(N-n+1)}\}$$

Luego, computar las siguientes particiones del vector de probabilidades de inclusión

$$\pi_{(k)}^a(1) = \begin{cases} 0 & \text{si } k \leq N - n \\ 1 & \text{si } k > N - n \end{cases} \quad (4.4.5)$$

$$\pi_{(k)}^b(1) = \begin{cases} \frac{\pi_{(k)}}{1 - \lambda(1)} & \text{si } k \leq N - n \\ \frac{\pi_{(k)} - \lambda(1)}{1 - \lambda(1)} & \text{si } k > N - n \end{cases} \quad (4.4.6)$$

Paso 3 ( $t$ -ésima iteración,  $t \geq 2$ ) Definir los siguientes conjuntos

$$\begin{aligned} A(t) &= \{k | 0 < \pi_{(k)}^b(t-1) < 1\} \\ B(t) &= \{k | \pi_{(k)}^b(t-1) = 1\} \end{aligned}$$

y las siguientes cantidades:

$$\begin{aligned} N^*(t) &= \#A(t) \\ n^*(t) &= \#B(t) \end{aligned}$$

Luego, para los elementos  $k \in A(t)$  calcular

$$\lambda(t) = \min\{1 - \pi_{(N^*(t)-n^*(t))}^b(t-1), \pi_{(N^*(t)-n^*(t)+1)}^b(t-1)\}$$

A continuación, para los elementos  $k \in A(t)$  computar las siguientes particiones del vector de probabilidades de inclusión

$$\pi_{(k)}^a(t) = \begin{cases} 0 & \text{si } k \leq N^*(t) - n^*(t) \\ 1 & \text{si } k > N^*(t) - n^*(t) \end{cases} \quad (4.4.7)$$

$$\pi_{(k)}^b(t) = \begin{cases} \frac{\pi_{(k)}^b(t-1)}{1-\lambda(t)} & \text{si } k \leq N^*(t) - n^*(t) \\ \frac{\pi_{(k)}^b(t-1)-\lambda(t)}{1-\lambda(t)} & \text{si } k > N^*(t) - n^*(t) \end{cases} \quad (4.4.8)$$

Paso 4 Iterar hasta obtener convergencia; es decir, hasta que  $\pi_{(k)}^b(t) \in \{0, 1\}$ .

**Ejemplo 4.4.2.** En este apartado se muestra paso a paso cómo trabaja el algoritmo de mínimo soporte basado en el método de escisión. Volvemos entonces a nuestra población ejemplo

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

El cálculo de las probabilidades de inclusión se hace con respecto a la expresión (4.3.1) donde la característica de información auxiliar corresponde a

$$\mathbf{x} = (52, 60, 75, 100, 50)$$

Por tanto, el vector de probabilidades de inclusión está dado por

$$\boldsymbol{\pi} = (0.46, 0.53, 0.67, 0.90, 0.44)$$

El método exige el ordenamiento del vector de probabilidades de inclusión en orden ascendente. Luego de esto, se tiene que el procedimiento converge en cuatro etapas. La tabla 4.3 muestra la convergencia del método y todas las posibles muestras que surgen del diseño muestral con soporte mínimo. Los cálculos en cada etapa se dan a continuación:

Cuadro 4.3: Diseño de mínimo soporte para la población  $U$ .

		Etapa 1 $\lambda(1) = 0.53$		Etapa 2 $\lambda(2) = 0.06$		Etapa 3 $\lambda(3) = 0.02$		Etapa 4 $\lambda(4) = 0.78$	
$k$	$\pi_k$	$\pi_k^a$	$\pi_k^b$	$\pi_k^a$	$\pi_k^b$	$\pi_k^a$	$\pi_k^b$	$\pi_k^a$	$\pi_k^b$
<b>Leslie</b>	0.44	0	0.94	0	1	1	1	1	1
<b>Yves</b>	0.46	0	0.98	1	0.98	0	1	1	1
<b>Ken</b>	0.53	1	0	0	0	0	0	0	0
<b>Erik</b>	0.67	1	0.29	1	0.24	1	0.22	0	1
<b>Sharon</b>	0.90	1	0.79	1	0.78	1	0.78	1	0

Etapa 1  $N = 5, n = 3, \lambda = \min\{1 - \pi_{(2)}, \pi_{(3)}\} = 0.53$

Etapa 2  $N^*(2) = 4, n^*(2) = 3, \lambda(2) = \min\{1 - \pi_{(1)}(1), \pi_{(2)}(1)\} = 0.06$

Etapa 3  $N^*(3) = 3, n^*(3) = 2, \lambda(3) = \min\{1 - \pi_{(1)}(2), \pi_{(2)}(2)\} = 0.02$

Etapa 4  $N^*(4) = 2, n^*(4) = 1, \lambda(4) = \min\{1 - \pi_{(1)}(3), \pi_{(2)}(3)\} = 0.78$

Por tanto, el diseño muestral de mínimo soporte está dado por

$$p(s) = \begin{cases} 0.53 & \text{si } s = \{\text{Ken, Erik, Sharon}\} \\ 0.0282 = (1 - 0.53) \times 0.06 & \text{si } s = \{\text{Yves, Erik, Sharon}\} \\ 0.0088 = (1 - 0.53 - 0.0282) \times 0.02 & \text{si } s = \{\text{Leslie, Erik, Sharon}\} \\ 0.3377 = (1 - 0.53 - 0.0282 - 0.008) \times 0.78 & \text{si } s = \{\text{Leslie, Yves, Sharon}\} \\ 0.0953 = (1 - 0.53 - 0.0282 - 0.008 - 0.3377) & \text{si } s = \{\text{Leslie, Yves, Erik}\} \end{cases}$$

#### 4.4.3 Estimación de la varianza

Existe un número muy grande de diseños y algoritmos de muestreo que trabajan bajo el supuesto de probabilidades de inclusión desiguales. En el caso particular del diseño de muestreo sin reemplazo y proporcional al tamaño de una característica de interés, las probabilidades de inclusión siguen el comportamiento dado por la expresión (4.3.1). Cada uno de estos métodos de muestreo inducen probabilidades de inclusión de primer y segundo orden. Las probabilidades de inclusión de primer orden son esenciales al momento de completar la estrategia de muestreo con el estimador de Horvitz-Thompson. Sin embargo, las probabilidades de inclusión de segundo orden, aunque servirían teóricamente para calcular y estimar la varianza del estimador de Horvitz-Thompson, son inefficientes pues cuando el tamaño de muestra crece, su cálculo se vuelve una total aventura, en muchos casos imposible de finiquitar.

Al respecto Tillé (2006) comenta, en el prefacio de su libro de algoritmos de muestreo, que «tiene la convicción de que las probabilidades de inclusión de segundo orden no son usadas para nada» y añade que «en la práctica el uso de las probabilidades de inclusión de segundo orden es muchas veces irreal porque son muy difíciles de calcular computacionalmente y  $n^2$  términos deben ser sumados para calcular la estimación».

Para evitar el cálculo y estimación de la varianza del estimador de Horvitz-Thompson con dobles sumas, Deville & Tillé (2005) proponen una aproximación de la varianza<sup>5</sup> y su respectiva estimación para un diseño exponencial<sup>6</sup> dada por el siguiente resultado

**Resultado 4.4.4.** *Para la familia de diseños exponenciales, la aproximación de la varianza del estimador de Horvitz-Thompson está dada por*

$$Var(\hat{t}_{y,\pi}) = \sum_{k \in U} \frac{b_k}{\pi_k^2} (y_k - y_k^*)^2 \quad (4.4.9)$$

donde

$$y_k^* = \pi_k \frac{\sum_{l \in U} b_l y_l / \pi_l}{\sum_{l \in U} b_l} \quad (4.4.10)$$

Hájek (1981) ha propuesto la siguiente escogencia de  $b_k$

$$b_k = \frac{N\pi_k(1 - \pi_k)}{(N - 1)} \quad (4.4.11)$$

Un estimador de la anterior aproximación de la varianza está dada por

$$\widehat{Var}(\hat{t}_{y,\pi}) = \sum_{k \in S} \frac{c_k}{\pi_k^2} (y_k - \hat{y}_k^*)^2 \quad (4.4.12)$$

<sup>5</sup>Existe mucha literatura escrita alrededor del tema de aproximaciones y simplificaciones de la varianza del estimador de Horvitz-Thompson. Para una mejor comprensión del tema Matei & Tille (2005) han escrito un excelente artículo de revisión.

<sup>6</sup>Los diseños de muestreo exponenciales son una gran familia que incluyen diseños tales como muestreo aleatorio simple, muestreo multinomial, muestreo de probabilidades desiguales con reemplazo y algunos diseños de probabilidades desiguales sin reemplazo. Para más información acerca de los diseños de muestreo exponenciales el lector deberá remitirse a Tillé (2006).

donde

$$\hat{y}_k^* = \pi_k \frac{\sum_{l \in S} c_l y_l / \pi_k}{\sum_{l \in S} c_l} \quad (4.4.13)$$

Deville (1993) ha propuesto la siguiente escogencia de  $c_k$

$$c_k = (1 - \pi_k) \frac{n}{(n - 1)} \quad (4.4.14)$$

**Ejemplo 4.4.3.** Para nuestra población de ejemplo  $U$ , existen  $\binom{N}{n} = 10$  posibles muestras  $\pi$ PT de tamaño  $n = 3$ . Utilizando las probabilidades de inclusión del ejemplo 4.4.1, realice el cálculo léxico-gráfico del estimador de Horvitz-Thompson, calcule la aproximación de la varianza dada por la expresión (4.4.9) y para cada muestra estime esta varianza usando la expresión (4.4.12) y compruebe su insesgamiento.

#### Acerca del muestreo $\pi$ PT

En general, la familia de diseños de muestreo  $\pi$ PT son utilizados cuando el comportamiento de la característica de interés en la población finita es bastante asimétrico. Para la estimación de totales, este diseño es más eficiente, en términos de reducción de la varianza. Sin embargo, cuando se quiere estimar otro tipo de parámetros poblacionales, como razones o medianas, los diseños de muestreo proporcionales al tamaño no son muy apetecidos, pues es difícil encontrar una característica de información auxiliar bien correlacionada con la razón entre las dos características de interés. En resumen, se tiene que:

- Se utiliza esencialmente para la estimación de totales poblacionales.
- Al seleccionar hogares no vale la pena utilizar este diseño pues, en general, en cada vivienda hay una misma cantidad de hogares.
- En encuestas de negocios es bueno utilizar diseños proporcionales porque sí existen diferencias en los tamaños considerados (por ejemplo en total de ventas mensuales, número de empleados contratados al año, etc.).
- Debido a que este diseño de muestreo involucra información auxiliar, entonces es más eficiente que el diseño de muestreo aleatorio simple, siempre y cuando la característica de interés esté relacionada positivamente con la información auxiliar.
- Un defecto de este diseño de muestreo es que su varianza no es una función monótona decreciente. Debido a la configuración particular de la información, la varianza puede crecer si se aumenta el tamaño de muestra.

#### 4.4.4 Marco y Lucy

En este apartado de Marco y Lucy suponga que se tienen las mismas condiciones que en el apartado de Marco y Lucy del diseño de muestreo PPT (ver la sección 4.2.4). Siendo así, el marco de muestreo permite conocer los valores poblacionales de una característica de información auxiliar. En este caso ésta es la variable `Income`. Dadas las bondades del marco de muestreo, se quiere seleccionar una muestra de tamaño  $n=2000$  mediante un diseño de muestreo sin reemplazo que induzca probabilidades de inclusión proporcionales a esta característica de información auxiliar.

La selección de la muestra se realiza haciendo uso de la función `S.piPS` del paquete `TeachingSampling` para la cual los argumentos introducidos son: el vector de valores poblacionales de la característica de información auxiliar `Income` y el tamaño de la muestra sin reemplazo  $n=2000$ . Nótese que esta función utiliza el algoritmo de selección de Sunter.

```

data(BigLucy)
attach(BigLucy)

N <- dim(BigLucy)[1]
n <- 2000
res <- S.piPS(n, Income)
sam <- res[,1]
muestra <- BigLucy[sam,]
attach(muestra)
head(muestra)

##           ID      Ubication Level     Zone Income Employees Taxes
## 83802 AB0000083802 C0088740K0213157 Big County98   2510    258   305
## 57446 AB0000057446 C0188504K0113393 Big County64   2510    258   305
## 45466 AB0000045466 C0251334K0050563 Big County55   2510    258   305
## 31090 AB0000031090 C0193894K0108003 Big County43   2510    258   305
## 23902 AB0000023902 C0146522K0155375 Big County36   2510    258   305
## 14318 AB0000014318 C0025721K0276176 Big County28   2510    258   305
##          SPAM ISO Years Segments
## 83802 yes yes 19.9 County98 46
## 57446 yes yes 46.7 County64 39
## 45466 yes yes 49.9 County55 11
## 31090 yes yes 47.6 County43 11
## 23902 yes yes 19.4 County36 23
## 14318 yes yes 38.7 County28 44

```

El resultado de la función `S.piPS` es una muestra ordenada de forma descendente por los valores de la característica de información auxiliar. El siguiente paso es recolectar la información de las características de interés `Employees` e `Taxes` para los elementos incluidos en la muestra realizada.

Después de recolectar la información, es necesario estimar los totales de las características de interés. En esta etapa se utiliza la función `E.piPS` del paquete `TeachingSampling` cuyos argumentos son: `estima`, correspondiente a la lista que contiene los valores observados en la muestra para cada una de las características de interés y `pik.s`, correspondiente al vector de probabilidades de inclusión (proporcionales a la característica de información auxiliar) de los elementos en la muestra.

```

pik.s <- res[,2]
estima <- data.frame(Income, Employees, Taxes)
E.piPS(estima, pik.s)

```

Los resultados para este ejercicio particular son excelentes. Nótese que los estimativos de la varianza no son exactos, pues están dados por el resultado 4.4.2, aunque sí aproximados. Por otra parte, el resultado 4.3.4 asegura que éstos serían menores a los arrojados por la estrategia de muestreo que utiliza un diseño PPT con reemplazo y el estimador de Hansen-Hurwitz. Por supuesto, este diseño de muestreo es más eficiente que el de Poisson, no es de extrañar que los resultados para la variable Ingreso sean tan exactos. Recuérdese que ésta fue la variable utilizada como característica de información auxiliar. La siguiente tabla muestra los resultado para un ejercicio particular. Una vez más, la característica Impuestos tiene un menor coeficiente de variación estimado puesto que está mucho mejor correlacionada con la variable Ingreso.

Véase que para obtener estos resultados, fue necesario conocer el valor de  $N$  dado por la longitud del vector de información auxiliar. Nótese que no siempre se puede asegurar el conocimiento del total

Cuadro 4.4: Muestreo  $\pi$ PT: estimación de los totales de las características de interés

	N	Income	Employees	Taxes
Estimation	87293.74	36634733.00	5507948.11	1001150.91
Standard Error	1588.19	0.00	90066.04	13235.55
CVE	1.82	0.00	1.64	1.32
DEFF	Inf	0.00	1.44	0.06

poblacional. Sin embargo, aunque no se conociera, con la función HT se hubiera llegado a los mismos resultados, en términos de la estimación de los totales, pero no se obtendrían los estimativos concernientes a la varianza, tal y como se ilustra a continuación.

```
HT(estima, pik.s)
```

```
##           [,1]
## Income    36634733
## Employees 5507948
## Taxes     1001151
```

## 4.5 Ejercicios

4.1 Demuestre o refute la siguiente afirmación: «Cuando el comportamiento de la característica de interés es uniforme en la población es más conveniente utilizar diseños de muestreo proporcionales al tamaño de una característica de información auxiliar».

4.2 Demuestre o refute la siguiente afirmación: «En muestreo Poisson, cuando las probabilidades de inclusión son tales que  $\pi_k = ny_k/t_y$  la varianza del estimador de Horvitz-Thompson es nula».

4.3 Complete el cálculo léxico-gráfico del ejemplo 4.1.2.

4.4 Suponga una población de 10 elementos  $U = \{e_1, \dots, e_{10}\}$  cuyo marco de muestreo contiene una característica de información auxiliar dada por

$$\mathbf{x} = (62, 151, 76, 77, 80, 60, 194, 78, 74, 61)$$

- Si se desea seleccionar una muestra sin reemplazo de tamaño esperado  $n(S) = 6$ , utilice la expresión (4.3.2) para construir un vector de probabilidades de inclusión proporcionales a  $\mathbf{x}$  tales que  $0 < \pi_k \leq 1$  para todo  $k \in U$  y verifique  $\sum_U \pi_k = 6$
- Utilice el algoritmo de la sección 4.1.1 para seleccionar una muestra Poisson teniendo en cuenta que se obtuvo el siguiente conjunto de números aleatorios uniformes

$$\boldsymbol{\varepsilon} = \{0.858, 0.698, 0.541, 0.320, 0.965, 0.497, 0.208, 0.006, 0.340, 0.206\}$$

- Utilice el método de Sunter para seleccionar una muestra  $\pi$ PT teniendo en cuenta que se obtuvo el siguiente conjunto de números aleatorios uniformes

$$\boldsymbol{\xi} = \{0.322, 0.542, 0.032, 0.141, 0.453, 0.668, 0.174, 0.318, 0.691, 0.006\}$$

4.5 (Särndal, Swensson & Wretman 1992, p. 117) Para estimar el total de la característica de interés  $y$  de una población de  $N = 284$  elementos, se utilizó un diseño de muestreo Poisson de tamaño de muestra esperado  $n(S) = 10$ . Las probabilidades de inclusión fueron proporcionales a una característica de información auxiliar  $x$  cuyo total poblacional es  $t_x = 8182$ . Luego, el algoritmo de selección arrojó una muestra de tamaño efectivo de 12 elementos, para las cuales se obtuvo la siguiente información

$x_k$	$y_k$
54	5246
671	59877
28	2208
27	2546
29	2903
62	6850
42	3773
48	4055
33	4014
446	38945
12	1162
46	4852

- Calcule una estimación insesgada para el total poblacional de la característica de interés, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.
- Calcule una estimación insesgada para la media poblacional de la característica de interés, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.
- Utilice el estimador alternativo para calcular estimaciones tanto del total como de la media poblacional.

4.6 Complete el cálculo léxico-gráfico del ejemplo 4.4.3.

4.7 Suponiendo que los datos del ejercicio 4.5 provienen de un diseño de muestreo  $\pi$ PT, calcule una estimación para el total de la característica de interés. Utilizando la aproximación de la varianza dada en (4.4.12), reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %.

4.8 Utilice el esquema de mínimo soporte para especificar un diseño de muestreo  $\pi$ PT de tamaño  $n = 3$  para una población de tamaño  $N = 6$  cuyo vector de probabilidades de inclusión de primer orden es

$$\boldsymbol{\pi} = (0.07, 0.17, 0.41, 0.61, 0.83, 0.91)'$$

Demuestre que el procedimiento converge en cuatro pasos que inducen cinco muestras y calcule la probabilidad de selección de cada muestra.

4.9 Demuestre o refute la siguiente afirmación: «En muestreo PPT es posible utilizar los estimadores de Horvitz-Thompson y de Hansen-Hurwitz, al comparar las dos estrategias se tiene que las dos aportan la misma precisión pero diferente confiabilidad».

4.10 Complete el cálculo léxico-gráfico del ejemplo 4.2.2.

4.11 Suponga una población de 12 elementos  $U = \{e_1, \dots, e_{12}\}$  cuyo marco de muestreo contiene una característica de información auxiliar dada por

$$\mathbf{x} = (674, 802, 829, 726, 709, 789, 742, 791, 805, 797, 771, 692)$$

- Si se desea seleccionar una muestra con reemplazo de tamaño  $m = 6$ , construya un vector de probabilidades de selección proporcionales a  $\mathbf{x}$  tales que  $0 < p_k \leq 1$  para todo  $k \in U$  y verifique  $\sum_U p_k = 6$
- Utilice el método acumulativo total para seleccionar una muestra PPT teniendo en cuenta que para cada una de las seis extracciones se generaron los siguientes números aleatorios uniformes

$$\boldsymbol{\varepsilon} = \{0.075, 0.397, 0.280, 0.407, 0.982, 0.782\}$$

- Utilice el método de Lahiri para seleccionar una muestra PPT usando sus propios números aleatorios  $\eta$  y  $l$  en cada una de las extracciones.
- 4.12 Demuestre o refute la siguiente afirmación: «Para la estimación de totales, el diseño PPT es preferido sobre el diseño  $\pi$ PPT porque permiten agilizar los cálculos computacionales de varianza y coeficiente de variación».
- 4.13 Demuestre o refute la siguiente afirmación: «Para la estimación de totales, el diseño PPT siempre es más eficiente que el diseño de muestreo aleatorio simple con reemplazo».
- 4.14 Suponga una población de  $N = 12$  elementos cuyos valores observados para la característica de interés son

$$y = \{50, 53, 44, 45, 53, 31, 35, 45, 34, 44, 52, 52\}$$

y los valores observados para la característica de información auxiliar son

$$x = \{1005, 1072, 884, 907, 1068, 625, 705, 909, 692, 891, 1046, 1052\}$$

- Calcule la correlación entre  $y^2/x$  y  $x$ .
  - Realice un gráfico de dispersión para  $y/x$  y explique si se puede afirmar que la razón es constante para los elementos de la población.
  - Utilice el análisis de regresión simple para estimar el valor del intercepto y decida si este es estadísticamente diferente de cero.
  - Para un tamaño de muestra  $m = 6$ , utilice la expresión (4.2.13) y los anteriores argumentos para justificar o descalificar la escogencia del diseño de muestreo PPT para esta población.
- 4.15 Asumiendo que los datos del ejercicio 4.5 provienen de un diseño de muestreo PPT, calcule la estimación de Hansen-Hurwitz para el total de la característica de interés, reporte el coeficiente de variación estimado y un intervalo de confianza al 95 %. También calcule la estimación de Horvitz-Thompson para el total de la característica de interés.



## Capítulo 5

# Muestreo estratificado

La estratificación es una de las técnicas más difundidas y usadas en muestreo puesto que tiene funcionalidades estadísticas y administrativas que la hacen atractiva: permite tratar con subpoblaciones, aumenta la eficiencia de las estimaciones y contribuye a la administración eficiente de grandes encuestas.

Richard Valliant (2000)

En algunas ocasiones, la característica de interés tiende a tomar distintos valores promedio con respecto a subgrupos poblacionales. De alguna manera, si la población tiene un comportamiento diferente en estos subgrupos, es posible mejorar la precisión de las estimaciones tomando muestras independientes en cada uno de los subgrupos poblacionales. Lo anterior es intuitivo cuando entre los subgrupos existe mucha variabilidad, pero dentro de ellos la variabilidad es constante.

En general, cuando existe en el marco de muestreo información auxiliar que permite la división de la población en  $H$  subgrupos con el objetivo de seleccionar una muestra en cada subgrupo, se dice que la estrategia de muestreo utiliza un **diseño de muestreo estratificado** y el nombre de los subgrupos, formados antes de la recolección de la información, se denomina **estratos**. Nótese la diferencia con los subgrupos poblacionales llamados **dominios**, en donde la partición de la población se realiza después de la recolección de la información.

Con frecuencia, tenemos información adicional que nos ayuda a diseñar la estrategia de muestreo. Cuando esta información se refiere a la pertenencia de cada uno de los elementos a un subgrupo, podemos aplicar una estrategia que utilice un diseño de muestreo estratificado. No es solamente la disponibilidad de esta información auxiliar la que nos lleva a utilizar un diseño de muestreo estratificado, además de esto:

1. La variable de interés asume distintos valores promedio en diferentes sub-poblaciones.
2. De una u otra forma (proceso logístico y/o de recolección de datos) es mejor estratificar y dividir la población en particiones. Lehtonen & Pahkinen (2003) afirman que algunas variables típicas de estratificación son de tipo regional (municipio, estado o provincia), demográfico (género o grupo de edad) y socioeconómico (grupo de ingresos). Existen censos, en períodos anteriores que pueden contener esta valiosa información.

La necesidad de estratificar<sup>1</sup> la población surge por una o más de las siguientes razones:

---

<sup>1</sup>Dividir la población en  $H$  estratos disjuntos.

- Por razones administrativas. Existen marcos de muestreo que ya tienen dividida la población en subgrupos formados naturalmente.
- Se desea garantizar que la muestra seleccionada sea representativa con respecto al comportamiento de la población según la información auxiliar. Al seleccionar una muestra aleatoria simple de una población de personas, podría suceder que la muestra seleccionada no incluyera a ningún hombre.
- Se requieren estimativos con alta precisión discriminados para cada sub-población. Aumentar el tamaño de muestra en los estratos menos representados.
- Menor Coste. Distintos esquemas operativos para diversos estratos. Encuestas por correo para empresas grandes. Menor tamaño de muestras en zonas de tolerancia o zonas de difícil manejo del orden público.
- Reducción de la varianza en la estimación. Personas de distintas edades con distintas presiones sanguíneas (estratificar por grupos de edad). Se reduce la varianza pues los estratos son homogéneos por dentro, pero heterogéneos entre sí.

El objetivo del diseño estratificado es dar un tratamiento particular a cada subgrupo, ya sea por razones económicas, administrativas o logísticas. Es indispensable delimitar bien los subgrupos en la etapa de diseño. Por ejemplo, en un estudio dentro de una universidad, si se quiere averiguar el número de horas que los estudiantes permanecen enfrente de un computador, no es una buena idea (defecto técnico) dividir la población en cursos porque los cursos no brindan una partición de la población, dado que en distintos cursos pueden estar los mismos estudiantes.

## 5.1 Fundamentos teóricos

Suponga que el marco de muestreo es tal que permite conocer la pertenencia de cada elemento de la población  $U$  en  $H$  sub-grupos poblacionales separados  $U_h$  ( $h = 1, 2, \dots, H$ ) también llamados estratos. Éstos se definen como grupos de elementos mutuamente excluyentes. Cada elemento puede pertenecer a uno y sólo a un estrato. De tal forma que

- $\bigcup_{h=1}^H U_h = U$
- $U_h \cap U_i = \emptyset \quad h \neq i$

Cada estrato  $U_h$  es de tamaño  $N_h$ , por tanto

$$\sum_{h=1}^H N_h = N \tag{5.1.1}$$

Con la población dividida en  $H$  estratos, el objetivo sigue siendo estimar los siguientes parámetros poblacionales

1. El total poblacional,

$$t_y = \sum_{k \in U} y_k = \sum_{h=1}^H \sum_{k \in U_h} y_k = \sum_{h=1}^H t_{yh} \tag{5.1.2}$$

donde  $t_{yh} = \sum_{k \in U_h} y_k$

2. La media poblacional,

$$\bar{y} = \frac{\sum_{k \in U} y_k}{N} = \frac{1}{N} \sum_{h=1}^H \sum_{k \in U_h} y_k = \frac{1}{N} \sum_{h=1}^H N_h \bar{y}_h \quad (5.1.3)$$

$$\text{donde } \bar{y}_h = \frac{1}{N_h} \sum_{k \in U_h} y_k$$

Sampath (2001) afirma que dependiendo de la naturaleza de los estratos, diferentes estrategias de muestreo pueden ser utilizadas en diferentes estratos. De tal forma que, en ausencia de información auxiliar, se utilice una estrategia aleatoria simple en algunos estratos, mientras que para aquellos sub-grupos tales que el marco de muestreo permite el conocimiento de información auxiliar continua, es posible aplicar una estrategia de muestreo proporcional al tamaño, e incluso para aquellos sub-grupos en los que, por obligación (logística o técnica), se deba aplicar un censo.

Es importante aclarar que la selección de las  $H$  muestras es realizada de manera independiente en cada estrato.<sup>2</sup> De tal forma que la muestra aleatoria  $S$ <sup>3</sup> queda definida por

$$S = \bigcup_{h=1}^H S_h. \quad (5.1.4)$$

En particular, si la muestra seleccionada es  $s$ , entonces

$$s = \bigcup_{h=1}^H s_h. \quad (5.1.5)$$

Nótese que si el tamaño de muestra en cada estrato es igual a  $n_h$ , entonces el tamaño de la muestra seleccionada mediante un diseño de muestreo estratificado es

$$n = \sum_{h=1}^H n_h. \quad (5.1.6)$$

Así, para cada estrato  $h$   $h = 1, \dots, H$  existe un conjunto de todas las posibles muestras denotado como soporte del estrato  $h$ , o  $Q_h$ . Cada uno de los soportes  $Q_h$  induce la definición del soporte general de la siguiente manera

$$Q^H = \bigtimes_{h=1}^H Q_h. \quad (5.1.7)$$

En donde  $\bigtimes$  denota el operador de producto cartesiano<sup>4</sup>. La cardinalidad de cada soporte  $Q_h$  depende del diseño de muestreo utilizado en la selección de la muestra del estrato  $h$ . Así

$$\#Q^H = \prod_{h=1}^H \#Q_h. \quad (5.1.8)$$

Por supuesto, el diseño de muestreo estratificado es un autentico diseño de muestreo como lo enuncian los siguientes resultados.

<sup>2</sup>Esto se debe a la independencia entre las selecciones. Aunque se conozcan qué unidades serán incluidas en la muestra de algún estrato, este conocimiento no afecta, de ninguna manera, la inclusión de cualquier otra unidad en los restantes estratos.

<sup>3</sup>Nótese que  $S$  es una variable aleatoria y que las medidas de probabilidad utilizadas para la selección de muestras en cada estrato son distintas.

<sup>4</sup>Por ejemplo, en presencia de dos conjuntos  $A = \{a, b\}$  y  $B = \{1, 2\}$ , entonces el producto cartesiano entre  $A$  y  $B$  es  $A \times B = \{(a, 1), (a, 2), (b, 1), (b, 2)\}$ .

**Resultado 5.1.1.** Siendo  $p_1(s_1), p_2(s_2), \dots, p_H(s_H)$  los diseños de muestreo utilizados en cada estrato  $h \quad h = 1, \dots, H$ , entonces el diseño de muestreo estratificado se define como

$$p(s) = \prod_{h=1}^H p_h(s_h) \quad (5.1.9)$$

*Demostración.* Se tiene que

$$\begin{aligned} p(s) &= Pr(\text{Seleccionar } s_1 \text{ de } U_1, \dots, \text{Seleccionar } s_H \text{ de } U_H,) \\ &= p_1(s_1) \cdots p_H(s_H), \end{aligned}$$

puesto que el proceso de selección es independiente en cada estrato.  $\square$

**Resultado 5.1.2.** El diseño de muestreo estratificado cumple que

$$1. \quad p(s) \geq 0 \text{ para todo } s \in Q$$

$$2. \quad \sum_{s \in Q} p(s) = 1$$

*Demostración.* La primera propiedad se tiene de inmediato puesto que todas las expresiones en 5.1.9 son mayores o iguales a cero. La segunda propiedad se tiene por inducción matemática sobre el número de estratos.

- Si  $H = 2$  existen dos soporte, uno para cada estrato,  $Q_1$  definido como

$$Q_1 = \{s_{11}, s_{12}, \dots, s_{1H_1}\} \quad (5.1.10)$$

y  $Q_2$  definido como

$$Q_2 = \{s_{21}, s_{22}, \dots, s_{2H_2}\} \quad (5.1.11)$$

tales que

$$Q^2 = \left\{ s_{11} \bigcup s_{21}, s_{11} \bigcup s_{22}, \dots, s_{11} \bigcup s_{2H_2}, \dots, s_{1H_1} \bigcup s_{2H_2} \right\} \quad (5.1.12)$$

Ahora, como la selección de las muestras se realiza en forma independiente, en particular se tiene que

$$p\left(s_{11} \bigcup s_{21}\right) = p(s_{11})p(s_{21}) \quad (5.1.13)$$

de manera análoga para el elemento que pertenezca al soporte. Ahora,

$$\begin{aligned} \sum_{s \in Q} p(s) &= p(s_{11})p(s_{21}) + p(s_{11})p(s_{22}) + \dots + p(s_{11})p(s_{2H_2}) + \\ &\quad \dots + p(s_{1H_1})p(s_{21}) + p(s_{1H_1})p(s_{22}) + \dots + p(s_{1H_1})p(s_{2H_2}) \\ &= p(s_{11})[\underbrace{p(s_{21}) + p(s_{22}) + \dots + p(s_{2H_2})}_1] + \\ &\quad \dots + p(s_{1H_1})[\underbrace{p(s_{21}) + p(s_{22}) + \dots + p(s_{2H_2})}_1] \\ &= p(s_{11}) + \dots + p(s_{1H_1}) \\ &= 1 \end{aligned}$$

- Si  $H = k$ , se supone que

$$\sum_{s \in Q^k} p(s) = 1 \quad (5.1.14)$$

donde

$$Q^k = \left\{ \bigcup_{h=1}^k s_h \mid s_h \in Q_h \right\}. \quad (5.1.15)$$

- Si  $H = k + 1$ , se tienen  $k + 1$  soportes tales que

$$\begin{aligned} Q_1 &= \{s_{11}, s_{12}, \dots, s_{1H_1}\} \\ &\vdots \\ Q_k &= \{s_{k1}, s_{k2}, \dots, s_{kH_k}\} \\ Q_{k+1} &= \{s_{k+1,1}, s_{k+1,2}, \dots, s_{k+1,H_{k+1}}\} \end{aligned} \quad (5.1.16)$$

Por consiguiente se tiene que

$$\begin{aligned} \sum_{s \in Q} p(s) &= p(s_{k+1,1}) \underbrace{\left[ \sum_{s \in Q^k} p(s) \right]}_1 + \dots + p(s_{k+1,1H_{k+1}}) \underbrace{\left[ \sum_{s \in Q^k} p(s) \right]}_1 \\ &= p(s_{k+1,1}) + \dots + p(s_{k+1,H_{k+1}}) \\ &= 1 \end{aligned}$$

□

### 5.1.1 Estimación en el muestreo estratificado

Si uno de los propósitos de la estratificación es obtener estimaciones más precisas, cabe preguntarse qué forma toman los estimadores y cómo definirlos a través de los estratos; pero aun más ¿qué forma toma la varianza del estimador en los estratos y su varianza estimada?. Los siguientes resultados, responden a los anteriores cuestionamientos.

**Resultado 5.1.3.** Si  $\hat{t}_{yh}$  estima insesgadamente el total de la característica de interés  $t_{yh}$  del subgrupo poblacional  $h$  con varianza igual a  $Var(\hat{t}_{yh})$ , entonces un estimador insesgado para el total poblacional  $t_y$  está dado por

$$\hat{t}_y = \sum_{h=1}^H \hat{t}_{yh} \quad (5.1.17)$$

el cual tiene una varianza igual a

$$Var(\hat{t}_y) = \sum_{h=1}^H Var(\hat{t}_{yh}) \quad (5.1.18)$$

*Demostración.* Dado que  $\hat{t}_{yh}$  es insesgado, tenemos que

$$\begin{aligned} E \left( \sum_{h=1}^H \hat{t}_{yh} \right) &= \sum_{h=1}^H E(\hat{t}_{yh}) \\ &= \sum_{h=1}^H t_{yh} = t_y \end{aligned}$$

Por otro lado, acudiendo a la independencia de la selección de muestras en cada estrato

$$\begin{aligned} Var\left(\sum_{h=1}^H \hat{t}_{yh}\right) &= \sum_{h=1}^H Var(\hat{t}_{yh}) + \sum_{h=1}^H \sum_{i=1}^H \underbrace{Cov(\hat{t}_{yh}, \hat{t}_{yi})}_0 \\ &= \sum_{h=1}^H Var(\hat{t}_{yh}) \end{aligned}$$

□

**Resultado 5.1.4.** Si  $\widehat{Var}(\hat{t}_{yh})$  estima insesgadamente a  $Var(\hat{t}_{yh})$ , entonces un estimador insesgado para  $Var(\hat{t}_y)$  está dado por

$$\widehat{Var}(\hat{t}_y) = \sum_{h=1}^H \widehat{Var}(\hat{t}_{yh}) \quad (5.1.19)$$

*Demostración.* La demostración es inmediata por el insesgamiento en cada uno de los estratos. □

### 5.1.2 El estimador de Horvitz-Thompson

**Resultado 5.1.5.** Para el diseño de muestreo estratificado, el estimador de Horvitz-Thompson, su varianza y su varianza estimada están dados por:

$$\hat{t}_{y,\pi} = \sum_{h=1}^H \hat{t}_{yh,\pi} \quad (5.1.20)$$

$$Var_{EST}(\hat{t}_{y,\pi}) = \sum_{h=1}^H Var_{p_h}(\hat{t}_{yh,\pi}) \quad (5.1.21)$$

$$\widehat{Var}_{EST}(\hat{t}_{y,\pi}) = \sum_{h=1}^H \widehat{Var}_{p_h}(\hat{t}_{yh,\pi}) \quad (5.1.22)$$

donde

$$\hat{t}_{yh,\pi} = \sum_{k \in S_h} \frac{y_k}{\pi_k} \quad (5.1.23)$$

Con  $Var_{p_e}(\hat{t}_{yh,\pi})$  es la varianza de  $\hat{t}_{yh,\pi}$  en el  $h$ -ésimo estrato y  $\widehat{Var}_{p_h}(\hat{t}_{yh,\pi})$  es la estimación de  $Var_{p_h}(\hat{t}_{yh,\pi})$  en el  $h$ -ésimo estrato.

**Ejemplo 5.1.1.** Nuestra población ejemplo  $U$  dada por

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

se divide en dos estratos de la siguiente forma

$$U_1 = \{\text{Erik, Sharon}\}$$

y el segundo conformado por:

$$U_2 = \{\text{Yves, Ken, Leslie.}\}$$

En el primer estrato se selecciona una muestra aleatoria de tamaño  $n_1 = 1$  de acuerdo a un diseño de muestreo aleatorio simple sin reemplazo. Por otra parte, en el segundo estrato se selecciona una muestra de tamaño  $n_2 = 2$  de acuerdo al siguiente diseño de muestreo

$$p_2(s) = \begin{cases} 1/4, & \text{si } s = \{\text{Yves, Ken}\}, \\ 1/4, & \text{si } s = \{\text{Yves, Leslie}\}, \\ 1/2, & \text{si } s = \{\text{Ken, Leslie}\}. \end{cases}$$

Realice el cálculo léxico-gráfico para comprobar el insesgamiento del estimador de Horvitz-Thompson para todas las posibles muestras de tamaño  $n = 3$ . Defina los soporte  $Q_1$  y  $Q_2$  así como el soporte general  $Q^2$  para cada estrato.

En las próximas secciones se estudiarán los diseños estratificados más utilizados en la práctica.

## 5.2 Diseño de muestreo aleatorio estratificado

Al igual que el muestreo aleatorio simple sin reemplazo, el diseño de muestreo aleatorio estratificado (EST-MAS) es el más sencillo de los diseños estratificados. En este caso particular se selecciona una muestra aleatoria simple en cada estrato, de tal forma que las selecciones sean independientes. Este diseño de muestreo es utilizado cuando la variabilidad de la característica de interés dentro de los estratos es similar; en otras palabras, cuando se sabe que el comportamiento de la característica de interés al interior de los estratos es homogéneo. Sin embargo, también se utiliza cuando no se dispone de ninguna información auxiliar continua que permita hacer uso de diseños de muestreo, en cada estrato, que permitan mejorar la eficiencia de una muestra aleatoria simple.

En cada estrato  $h$  una muestra aleatoria simple sin reemplazo de tamaño  $n_h$  es seleccionada, de manera independiente, de la población del estrato de tamaño  $N_h$ . Aunque el diseño de muestreo aleatorio simple es utilizado como un método final de selección de elemento, en conjunto el diseño estratificado puede resultar dramáticamente más eficiente que utilizar un diseño de muestreo aleatorio simple sin dividir la población.

**Definición 5.2.1.** Para tamaños de muestra fijos en cada estrato, denotados como  $n_1, \dots, n_H$ , un diseño de muestreo se dice estratificado aleatorio simple sin reemplazo si la probabilidad de seleccionar una muestra de tamaño  $n$  está dada por

$$p(s) = \begin{cases} \prod_{h=1}^H \frac{1}{\binom{N_h}{n_h}}, & \text{si } \sum_{h=1}^H n_h = n \\ 0, & \text{en otro caso} \end{cases} \quad (5.2.1)$$

Nótese que  $\sum_{s \in Q^H} p(s) = 1$  porque  $\#Q^H = \prod_{h=1}^H \binom{N_h}{n_h}$ .

### 5.2.1 Algoritmos de selección

En la selección de las muestras aleatorias simples sin reemplazo en cada estrato es posible utilizar los algoritmos de muestreo dados en el capítulo 3, de tal forma que los siguientes pasos se deben realizar.

- Separar la población en  $H$  subgrupos o estratos mediante la caracterización poblacional de información auxiliar.

- En cada estrato seleccionar una muestra aleatoria simple sin reemplazo. Los algoritmos utilizados en la selección de la muestra dentro de cada estrato pueden ser los métodos coordinado negativo o el método de selección y rechazo de Fan, Muller & Rezucha (1962).
- Cada una de las  $H$  selecciones es realizada de manera independiente

**Ejemplo 5.2.1.** Suponga que nuestra población de ejemplo  $U$  está particionada de acuerdo a la sección anterior. Es necesario definir los dos estratos en R, de manera tal que ningún elemento tenga una doble pertenencia a algún estrato.

```
U1 <- c("Erik", "Sharon")
N1 <- length(U1)
U2 <- c("Yves", "Ken", "Leslie")
N2 <- length(U2)
```

R permite realizar operaciones entre conjuntos de datos. En particular, el operador `union` es utilizado para verificar que la unión de los estratos dé como resultado la población de ejemplo  $U$ . Nótese que el tamaño poblacional es la suma de los tamaños de los dos estratos.

```
U <- union(U1,U2)
N <- N1+N2

U

## [1] "Erik"   "Sharon" "Yves"   "Ken"    "Leslie"

N

## [1] 5
```

Se ha decidido seleccionar una muestra aleatoria simple sin reemplazo de tamaño  $n_1 = 1$  para  $U_1$  y una muestra aleatoria simple sin reemplazo de tamaño  $n_2 = 2$  para  $U_2$ . De tal forma que la muestra general será de tamaño  $n = n_1 + n_2 = 3$ .

```
sam1 <- sample(N1, 1, replace=FALSE)
U1[sam1]

## [1] "Sharon"

sam2 <- S.SI(N2,2)
U2[sam2]

## [1] "Ken"    "Leslie"

sam <- union(U1[sam1],U2[sam2])
sam

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

Por supuesto, es posible utilizar la función `sample` que viene incorporada en el ambiente genérico de R o también es posible utilizar la función la función `S.SI` del paquete `TeachingSampling`. Sin importar

el algoritmo de selección de las muestras aleatorias simples sin reemplazo, es importante notar que se han seleccionado tantas muestras como estratos existen en la población.



# 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.
- Brewer, K. (1963), ‘A model of systematic sampling with unequal probabilities’, *Australasian Journal of Statistics* **5**, 93–105.
- Brewer, K. (1975), ‘A simple procedure for  $\pi$ pswor’, *Australian Journal of Statistics* **17**, 166–172.
- Brewer, K. (2002), *Combined sampling inference, weighting Basu’s elephants*, London: Arnorld.
- Brewer, K. & Hanif, M. (1983), *Sampling with unequal probabilities*, New York: Springer-Verlag.
- Cassel, C., Särndal, C. & Wretman, J. (1976a), *Foundations of Inference in Survey Sampling*, Wiley.
- Cassel, C., Särndal, C. & Wretman, J. (1976b), ‘Some results on generalized difference estimation and generalized regression estimation for finite populations’, *Biometrika* **63**, 615–620.
- Chambers, R. L. & Skinner, C. J., eds (2003), *Analysis of Survey Data*, Wiley.
- 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.
- Deville, J. (1993), ‘Estimation de la variance pour les enquêtes en deux phases’, *Note Interne Manuscrite. France: INSEE* .
- Deville, J.-C. & Tillé, Y. (1998), ‘Unequal probability sampling without replacement through a splitting method’, *Biometrika* **85**, 89–101.
- Deville, J.-C. & Tillé, Y. (2005), ‘Variance approximation under balanced sampling’, *Journal of Statistical Planning and Inference* **128**, 411–425.
- 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.
- Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E. & R., T. (2004), *Survey Methodology*, Wiley.

- Hájek, J. (1960), 'Limiting distributions in simple random sampling from a finite population', *Publication of Mathematical Institute of the Hungarian Academy of Science* **5**, 361–374.
- Hájek, J. (1981), *Sampling from a finite population*, New York: Marcel Dekker.
- Hansen, H. M. & Hurwitz, W. N. (1943), 'On the theory of sampling from finite populations', *Annals of Mathematical Statistics* **14**, 333–362.
- Hansen, M., Hurwitz, W. & Madow, W. G. (1953), *Sample survey methods and theory. Vols. I and II*, John Wiley and Sons.
- Hartley (1959), 'Analytic studies of survey data', *Instituto di Statistica Volume in honor of Corrado Gini*.
- Horvitz, D. & Thompson, D. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association* **47**, 663–685.
- Kish, L. (1965), *Survey Sampling*, Wiley.
- Lahiri, D. (1951), 'A method for sample selection providing unbiased ratio estimates', *Bulletin of the International Statistical Institute*. **33,2**, 133–140.
- 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 Statistical Institute', *Journal of the Royal Statistical Society* **109**, 325–370.
- Matei, A. & Tille, Y. (2005), 'Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size', *Journal of Official Statistics*. **4**, 543–570.
- Mood, A. M., Graybill, F. A. & Boes, D. C. (1974), *Introduction to the Theory of Statistics*, 3 edn, McGraw Hill.
- Narain, R. (1951), 'On sampling without replacement with varying probabilities', *Journal of Indian Society of Agricultural Statistics* **3**, 169–175.
- Ospina, D. (2001), *Introducción al muestreo*, Universidad Nacional de Colombia.
- Raj, D. (1954), 'On sampling with probabilities proportional to size', *Ganita* **5**, 175–182.
- Raj, D. (1968), *Sampling theory*, McGraw Hill.
- Sampath, S. (2001), *Sampling Theory and Methods*, Narosa Publishing House.
- Särndal, C., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer, New York.
- Sen, A. (1953), 'On the estimate of the variance in sampling with varying probabilities', *Journal of the Indian Society of Agricultural Statistics* **5**, 119–127.
- Sunter, A. (1977), 'List sequential sampling with equal or unequal probabilities without replacement', *Applied Statistics* **26**, 261–268.
- Sunter, A. (1986), 'Solutions to the problem of unequal probabilities sampling without replacement', *International Statistical Review* **54**, 33–50.
- Tillé, Y. (2006), *Sampling Algorithms*, Springer.

- Wu, C. (2003), ‘Optimal calibration estimators in survey sampling’, *Biometrika* **90**, 937–951.
- Yates, F. & Grundy, P. (1953), ‘Selecting without replacement from within strata with probability proportional to size’, *Journal of the Royal Statistical Society B* **15**, 235–261.