# 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[1] 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.

[1] 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, periodos 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 periodos de tiempo o espacio.
- **Poblaciones poco frecuentes:** cuando la incidencia es muy baja (por ejemplo las poblaciones de invidentes 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[2] 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 si 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 diseñó 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.

[2] 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:
   (a) Grandes: empresas que tributan 49 millones de dólares al año o más.
   (b) Medianas: empresas que tributan más de 11 millones y menos de 49 millones de dólares al año.
   (c) 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 antiguedad 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 las 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)
```

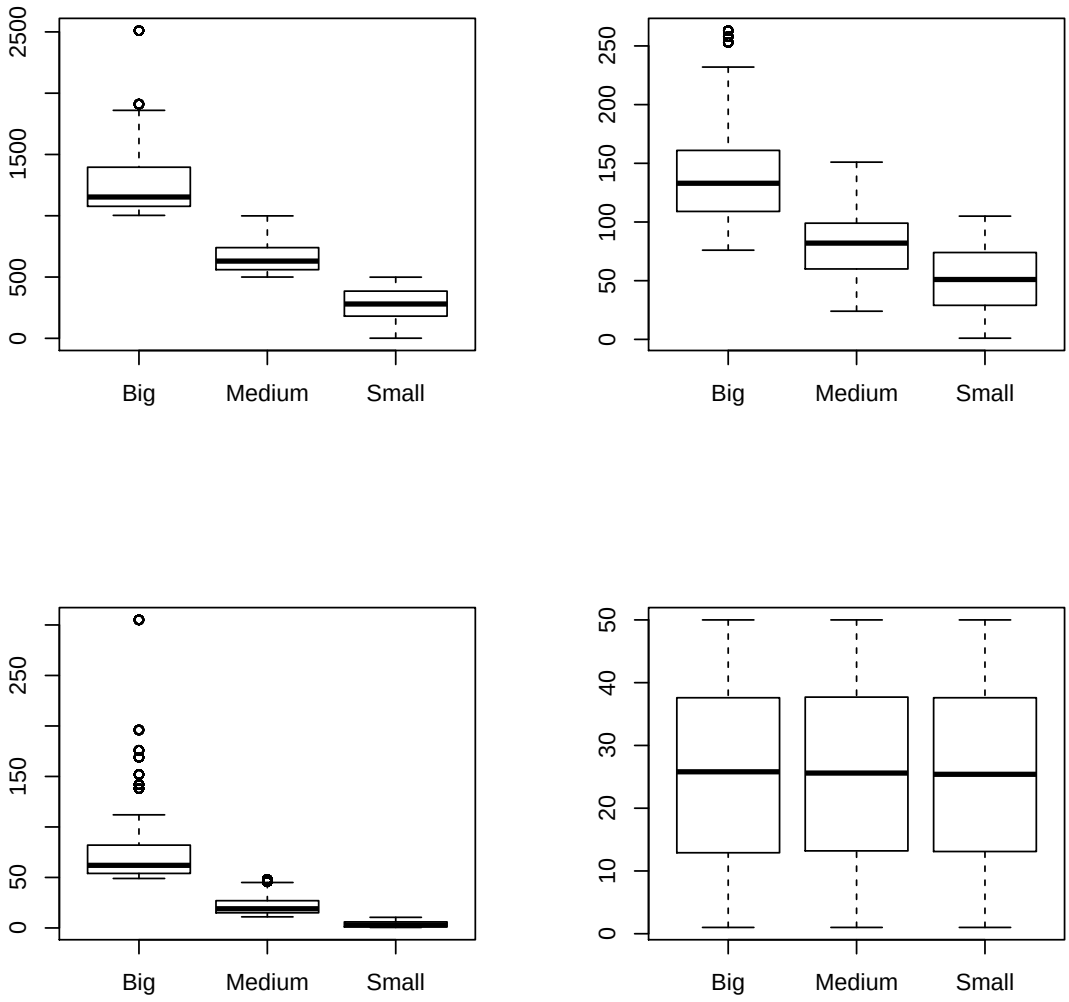


Figura 1.1: *Boxplot de las características de interés en cada nivel industrial.*

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)
```

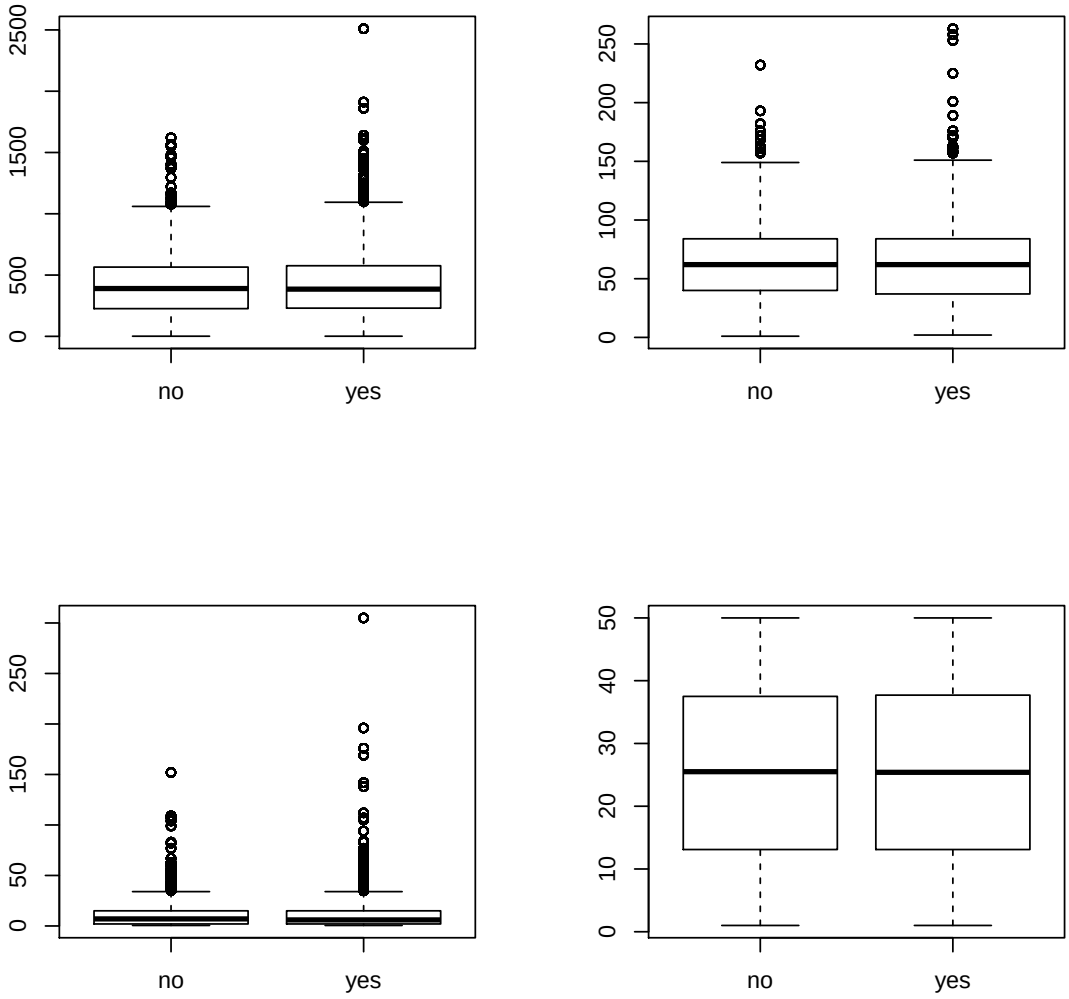


Figura 1.2: *Boxplot de las características de interés en cada nivel industrial.*

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

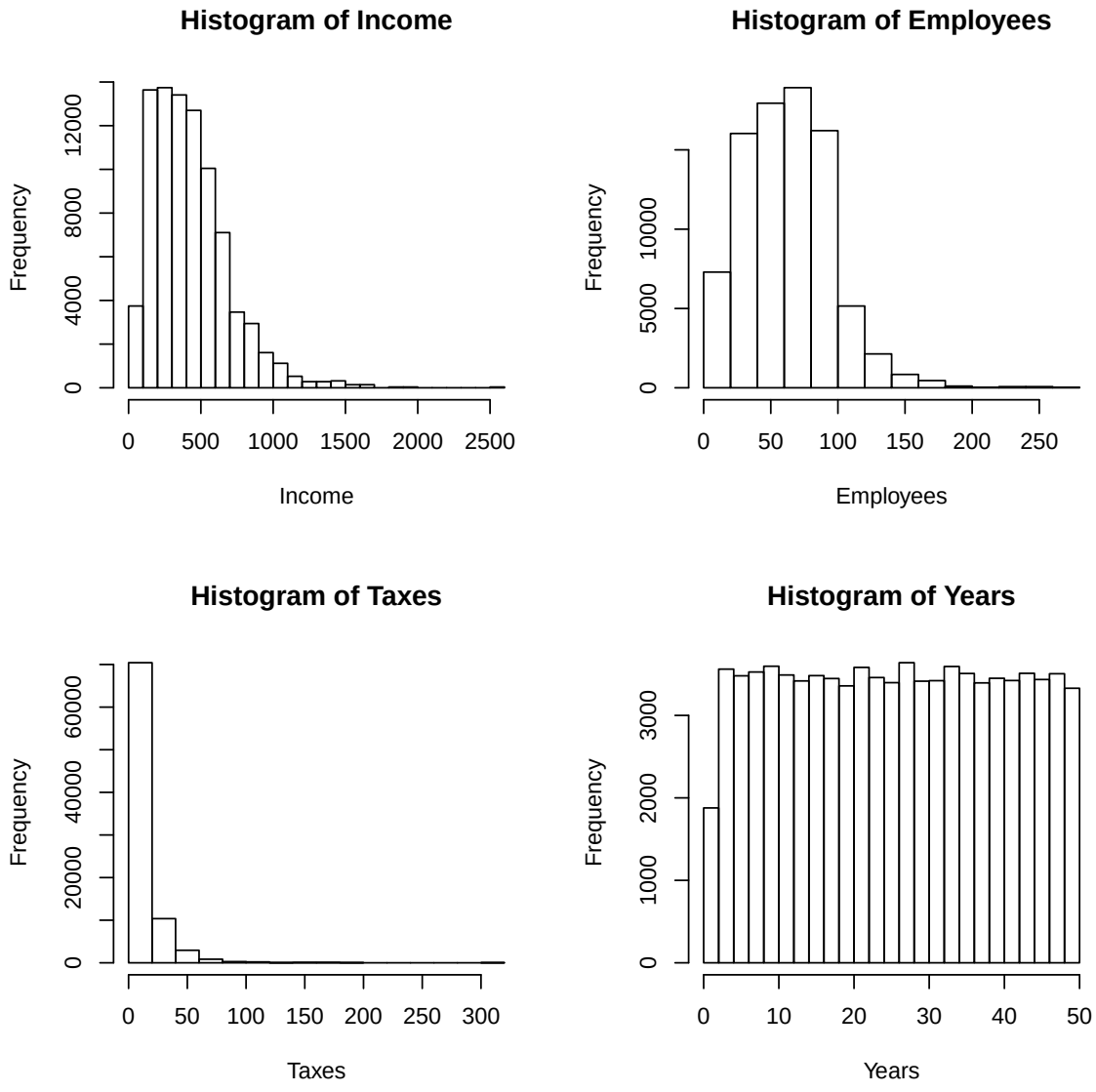


Figura 1.3: *Histograma de las características de interés.*

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

| | | | Ingreso | Impuestos | Empleados |
|---|---|---|---|---|---|
| Nivel | Grande | N total | 83 | 83 | 83 |
| | | Suma | 103.706 | 6.251 | 11.461 |
| | | Media | 1.249 | 75 | 138 |
| | Mediano | N total | 737 | 737 | 737 |
| | | Suma | 487.351 | 16.293 | 59.643 |
| | | Media | 661 | 22 | 81 |
| | Pequeño | 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 |
| | | Suma | 397.952 | 10.593 | 59.600 |
| | | Media | 425 | 11 | 64 |
| | si | 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 | | | si | | |
| | | 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 $\{e_1, e_2, ..., e_N\}$. Cada unidad puede ser identificada sin ambigüedad por un conjunto de rótulos. Sea $U = \{1, 2, ..., N\}$ 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

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

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[1] 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

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

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 **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*

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

*donde*

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

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 $s$.

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

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*

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

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

[1] 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

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

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

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

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

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

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* $\{\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_N\}$ *es una partición de* $\mathcal{S}$.
4. *El conjunto* $\{\mathcal{R}_0, \mathcal{R}_1, \ldots, \mathcal{S}_N, \ldots\}$ *es una partición infinita de* $\mathcal{R}$.
5. $\mathcal{S} \subset \mathcal{R}$ *para todo* $n = 0, 1, \ldots, N$.

### Muestras probabilísticas

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

- Es posible construir (o al menos definir teóricamente) un soporte $Q$, tal que $Q = \{s_1, \ldots, s_q, \ldots, s_Q\}$, de todas las muestras posibles obtenidas por un método de selección. En donde $s_q$, $q = 1, \ldots, Q$, 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 conteniendo 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*

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

[2]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

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

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 (1976) 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

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

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

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

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*

$$\boldsymbol{\mu} = E(\mathbf{s}) = \sum_{\mathbf{s} \in Q} p(\mathbf{s})\mathbf{s} \tag{2.1.10}$$

Si el diseño muestral es sin reemplazo, entonces $\boldsymbol{\mu} = \boldsymbol{\pi}$, donde $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)'$ 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:*

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

*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

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

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 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.

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

Algunos de los parámetros de interés más comunes son:

1. El total poblacional,
$$t_y = \sum_{k \in U} y_k \tag{2.1.13}$$
2. La media poblacional,
$$\bar{y}_U = \frac{\sum_{k \in U} y_k}{N} = \frac{t_y}{N} \tag{2.1.14}$$
3. La varianza poblacional,
$$S^2_{yU} = \frac{\sum_{k \in U} (y_k - \bar{y}_U)^2}{N - 1} \tag{2.1.15}$$

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 conteniendo 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

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

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

$$\begin{aligned} Var(G) &= E[G - E(G)]^2 && (2.1.17)\\ &= \sum_{s \in Q} p(s)[G(s) - E(G)]^2. && (2.1.18) \end{aligned}$$

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

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

□

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.

$$n(S) = \sum_U I_k. \tag{2.1.19}$$

**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 columnnas) 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 `Deltakl` 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 <- Deltakl(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 -0.00000000000000001388

colSums(Delta)

## [1] -0.00000000000000013878 -0.00000000000000008327 -0.00000000000000005551
## [4] -0.00000000000000006939 -0.00000000000000001388
```

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

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

y el error cuadrático medio, dado por

$$ECM(\hat{T}) = E[\hat{T} - T]^2 \tag{2.1.21}$$
$$= Var(\hat{T}) + B^2(\hat{T}). \tag{2.1.22}$$

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 permita 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

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

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:

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

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*

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

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

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

□

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 1976) 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*

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

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

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

□

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*

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

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

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

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 $Var_2(\hat{t}_{y,\pi})$. □

**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*

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

**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*

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

*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. □

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:

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

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 posible 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

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

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:

- **Confiabilidad:** 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 confiabilidad se debe conocer al parámetro $T$ (desconocido) por tanto, en términos prácticos la confiabilidad no se puede calcular. Para determinar la precisión y la confiabilidad 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 $k = 1, ..., N$. Luego la precisión tampoco se puede calcular. Sin embargo se debe proponer un estimador de $Var(\hat{T})$ (ojalá insesgado) que junto con $\hat{T}$ proporción 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*

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

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

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

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

*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

$$N = \sum_U 1, \tag{2.2.13}$$

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*

$$\hat{N}_\pi = \sum_S \frac{1}{\pi_k}. \tag{2.2.14}$$

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

$$\widetilde{y}_S = \frac{\hat{t}_{y,\pi}}{\hat{N}_\pi} \tag{2.2.15}$$

$$= \sum_S \frac{y_k}{\pi_k} \Big/ \sum_S \frac{1}{\pi_k}. \tag{2.2.16}$$

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 $\hat{\bar{y}}_\pi$ toma la siguiente forma

$$\hat{\bar{y}}_\pi = \frac{1}{N}\sum_s \frac{y_k}{\pi_k} = \frac{1}{N}\sum_s \frac{C}{\pi_k} = \frac{C}{N}\sum_s \frac{1}{\pi_k} = C\frac{\hat{N}_\pi}{N}. \tag{2.2.17}$$

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 $\widetilde{y}_S$ puesto que su variación es menor y cuando $y_k = C\ \forall k \in U$ reproduce la media poblacional con varianza nula puesto que

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

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 $\widetilde{y}_S$ estima la media poblacional, entonces $N\widetilde{y}_S$ estimará el total poblacional. Por tanto, el estimador alternativo está dado por la siguiente expresión

$$\hat{t}_{y,alt} = N\widetilde{y}_S = \hat{t}_{y,\pi}\frac{N}{\hat{N}_\pi} \tag{2.2.18}$$

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,

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

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[k,], all.pik[k,])
}

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

$$Var(\hat{t}_\pi) = (0.13)(155.2 - 236)^2 + (0.2)(151.0 - 236)^2 + \cdots + (0.08)(399.3 - 236)^2 = 7847.2$$

Acudiendo a la función `VarHT`, del paquete `TeachignSampling`, es posible reproducir este mismo calculo de la varianza. Sin embargo, esta función utiliza la expresión teórica de la varianza $Var_1(\hat{t}_{y,\pi})$ 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-Thompsosn 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 calculo 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 $p_k$ ($k = 1, \ldots, N$) 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
$$Pr(\text{Seleccionar el elemento } k) = p_k,\ k \in U.$$
- 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[3] 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*

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

*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 $0, 1, \ldots, m$; al definir éxito como la selección del elemento $k$-ésimo en la muestra, entonces se tiene la demostración del resultado. □

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

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

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

[3] 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

$$\#Q = \binom{N+m-1}{m} \tag{2.2.20}$$

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

$$\pi_k = 1 - (1-p_k)^m \tag{2.2.21}$$

*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:*

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

*Demostración.* Para hallar esta probabilidad debemos negar que ($k \in S$ y $l \in s$). Esta negación da como resultado ($k \notin s$ 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

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

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 ordenas 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`[4] 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[5] 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
```

[4]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.

[5]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

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

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

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

### 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:

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

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 $U_1, U_2, \ldots, U_m$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas con $E(U_i) = \mu$ y $Var(U_i) = \sigma^2$. 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*

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

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

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

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

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

La varianza está determinada por

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

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

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

entonces,

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

Por otro lado

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

Esto conduce a la demostración del teorema puesto que

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

□

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*

$$Z_i = y_{k_i}/p_{k_i} \qquad k \in U \quad i = 1, \ldots, m \tag{2.2.30}$$

*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*

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

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

$$E(Z_i) = t_y \tag{2.2.32}$$

*y*

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

*respectivamente.*

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

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

con probabilidades

$$p_1, p_2 \ldots, p_N$$

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 $Z_i$[6]. 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

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

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*

[6] $Z_1, \ldots, Z_m$ 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*

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

*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*

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

*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*

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

*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 $Z_1, \ldots, Z_m$ 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 $y_k \ \forall k = 1, 2, ..., N$. 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] / pkm[1]) - est
dif[2] <- (ym[2] / pkm[2]) - est
dif[3] <- (ym[3] / pkm[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[7] 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

[7] 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 $B^2(\hat{T})$.

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

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

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

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 $\sum_U \pi_k - \left(\sum_U \pi_k\right)^2 + \sum\sum_{k\neq l}\pi_{kl} = 0$. 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 com 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).

# Bibliografía


Bautista, J. (1998), *Diseños de muestreo estadístico*, Universidad Nacional de Colombia.

Brewer, K. (2002), *Combined sampling inference, weighting Basu's elephants*, London: Arnorld.

Cassel, C., Särndal, C. & Wretman, J. (1976), *Foundations of Inference in Survey Sampling*, Wiley.

Chambers, R. L. & Skinner, C. J., eds (2003), *Analysis of Survey Data*, Wiley.

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 poulation', *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, M., Hurwitz, W. & Madow, W, G. (1953), *Sample survey methods and theory. Vols. I and II*, John Wiley and Sons.

Horvitz, D. & Thompson, D. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association* **47**, 663–685.

Lohr, S. (2000), *Sampling: Design and Analysis*, Thompson.

Narain, R. (1951), 'On sampling without replacement with varying probabilities', *Journal of Indian Society of Agricultural Statistics* **3**, 169–175.

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.

Tillé, Y. (2006), *Sampling Algorithms*, Springer.

Yates, F. & Grundy, P. (1953), 'Selecting withou replacement from within estrata with probability proportional to size', *Journal of the Royal Statitical Society* **B15**, 235–261.