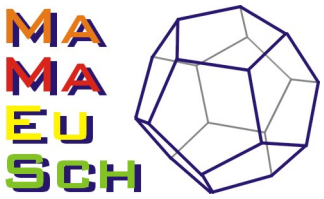
MaMaEuSch

Management **Ma**thematics for
European **Sch**ools
http://www.mathematik.uni-kl.de/~ mamaeusch



Población y muestra. Técnicas de muestreos

Paula Lagares Barreiro[*]
Justo Puerto Albandoz[*]

MaMaEuSch[**]
Management Mathematics for European Schools
94342 - CP - 1 - 2001 - 1 - DE - COMENIUS - C21

[*]Universidad de Sevilla

[**]Este proyecto ha sido llevado a cabo con ayuda parical de la Comunidad Europea en el marco del programa Sócrates. El contenido del proyecto no reflejy necesariamente la posición de la Comunidad Europea, ni implica ninguna responsabilidad por su parte.

Índice general

Capítulo 1

Población y muestra. Técnicas de muestreo

Vamos a ampliar en este capítulo lo que ya vimos al principio de Estadística Descriptiva, incluyendo ahora la definición de algunas técnicas de muestreo y de las nociones suficientes para ser capaces de decidir cuál es la técnica de muestreo mas adecuada a cada situación.

Imagina por ejemplo que tu clase ha sido seleccionada como la muestra de una población. El estudio que se vaya a realizar podría ser de diferentes temas, como los siguientes:

- La opinión sobre la posibilidad de organizar movidas alternativas en tu ciudad, y sobre las propuestas de actividades a realizar en dicha movida.
- Un sondeo sobre la valoración de los diferentes líderes políticos.
- La opinión sobre el destino de un posible viaje de fin de curso de los alumnos de tu nivel.

¿Crees que tu clase sería una buena muestra para cualquiera de estos casos? La respuesta es que, por ejemplo, para el segundo caso, los alumnos de una clase no son la muestra adecuada. Para el primer caso, es razonable pensar que pueden aportar información interesante, aunque la muestra puede resultar "pequeña" y podría faltarle información (chicos de otras edades, de otros barrios...), mientras que para el tercer caso, la muestra puede ser muy adecuada. Es por tanto muy importante la elección de una técnica de muestreo que nos asegure que la muestra escogida es 'adecuada' para el estudio que queremos realizar.

1.1. Motivos para la realización de un muestreo. Consideraciones necesarias

Imagina que vas a realizar estudios para conocer la siguiente información:

- El porcentaje de españoles que tiene acceso a internet.

- La duración media de una determinada marca de pilas.

Para el primer caso, la población a la que debes preguntar es de más de 40 millones de personas. Es obvio que entrevistar a más de 40 millones de personas supone un gran esfuerzo en varios sentidos. Primero, de tiempo, y segundo de dinero, puesto que es necesario contratar a muchos encuestadores, pagarles viajes para que lleguen a todos los pueblos, etc. Además, hay una dificultad añadida: es difícil llegar a todos y cada uno de los españoles, ya que cuando vayamos a entrevistar, habrá gente que esté de viaje fuera del país, habrá gente que esté enferma en el hospital, etc. En este caso, por motivos económicos, de tiempo y de dificultad de acceso a toda la población, sería conveniente entrevistar a una cierta parte de la población, una muestra, elegida convenientemente para poder extraer después conclusiones a toda la población.

En el segundo caso tenemos una problemática diferente. Para poder estudiar la duración de una pila, debemos usarla hasta que se gaste, lo que nos impide volver a usar la pila. Es decir, de alguna manera "destruimos" este elemento de la población. Si quisiéramos probar todas y cada una de las pilas, nos quedaríamos sin ellas. En este caso, de nuevo sería conveniente estudiar sólo un conjunto de esas pilas y luego extraer conclusiones más generales a partir del conjunto que hemos estudiado.

Por las razones anteriores, en muchos casos es conveniente el uso de muestras, pero para que podamos extraer conclusiones, es importante que elijamos bien las muestras para nuestros estudios. Por ejemplo, para el caso de el acceso a internet de los españoles, elegir a 10 personas de 40 millones es insuficiente, no es representativo. Tampoco lo sería preguntarle, por ejemplo a 100 personas de Madrid, o elegir a todos tus amigos y tu familia. Hay cuestiones que debemos especificar a la hora de elegir una muestra:

1. El método de selección de los individuos de la población (tipo de muestreo que se va a utilizar).
2. El tamaño de la muestra.
3. El grado de fiabilidad de las conclusiones que vamos a presentar, es decir, una estimación del error que vamos a cometer (en términos de probabilidad).

Como ya hemos dicho, la selección no adecuada de los elementos de la muestra provoca errores posteriores a la hora de estimar las correspondientes medidas en la población. Pero podemos encontrar más errores: el entrevistador podría no ser imparcial, es decir, favorecer que se den unas respuestas más que otras. Puede ocurrir también que, por ejemplo, la persona que vayamos a entrevistar no quiera contestar a ciertas preguntas (o no sepa contestar). Clasificamos todos estos posibles errores de la siguiente manera:

1. **Error de sesgo o de selección**: si alguno de los miembros de la población tiene más probabilidad que otros de ser seleccionados. Imagina que queremos conocer el grado de satisfacción de los clientes de un gimnasio y para ello vamos a entrevistar a algunos de 10 a 12 de la mañana. Esto quiere decir que las personas que vayan por la tarde no se verán representadas por lo que la muestra no representará a todos los clientes del gimnasio. Una forma de evitar este tipo de error es tomar la muestra de manera que todos los clientes tengan la misma probabilidad de ser seleccionados.
2. **Error o sesgo por no respuesta**: es posible que algunos elementos de la población no quieran o no puedan responder a determinadas cuestiones. O también puede ocurrir, cuando tenemos cuestionarios de tipo personal, que algunos miembros de la población no contesten

sinceramente. Estos errores son, en general, difíciles de evitar, pero en el caso de la sinceridad, se suelen incorporar cuestiones (preguntas filtro) para detectar si se está contestando sinceramente.

Después de lo que acabamos de ver, podemos decir que una muestra es sesgada cuando no es representativa de la población.

1.2. Técnicas de muestreo

Ya hemos hecho referencia a la importancia de la correcta elección de la muestra para que sea representativa para nuestra población pero ¿cómo clasificamos las diferentes formas de elegir una muestra? Podemos decir que hay tres tipos de muestreo:

1. **Muestreo probabilístico**: es aquel en el que cada muestra tiene la misma probabilidad de ser elegida.
2. **Muestreo intencional u opinático**: en el que la persona que selecciona la muestra es quien procura que sea representativa, dependiendo de su intención u opinión, siendo por tanto la representatividad subjetiva.
3. **Muestreo sin norma**: se toma la muestra sin norma alguna, de cualquier manera, siendo la muestra representativa si la población es homogénea y no se producen sesgos de selección.

Nosotros siempre haremos muestreo probabilístico, ya que en caso de elegir la técnica adecuada, es el que nos asegura la representatividad de la muestra y nos permite el cálculo de la estimación de los errores que se cometen. Dentro del muestreo probabilístico podemos distinguir entre los siguientes tipos de muestreo:

- Muestreo aleatorio con y sin reemplazo.
- Muestreo estratificado.
- Muestreo por conglomerados.
- Muestreo sistemático.
- Otros tipos de muestreo.

Imagina ahora que ya has seleccionado una muestra de un Centro de Enseñanza Secundaria (CES) en el que hay 560 alumnos. Has elegido una muestra de 28 alumnos para conocer si tienen internet en casa. Pero, ¿qué significa elegir a 28 de 560? ¿Qué proporción de la población estás entrevistando? Y a la hora de obtener conclusiones sobre la población ¿a cuántos alumnos de la población total representa cada uno de los de la muestra?

Para calcular la proporción de alumnos que estamos entrevistando, dividimos el tamaño de la muestra entre el de la población: $28/560 = 0{,}05$, lo que quiere decir que estamos pasando la encuesta al 5 % de la población.

Ahora vamos a calcular a cuántos individuos representa cada uno de los elementos de la muestra. Hacemos la división contraria, dividimos el número de individuos de la población entre los de la

muestra: $560/28 = 20$, lo que querría decir que cada uno de los elementos de la muestra representa a 20 alumnos del CES.

Los dos conceptos que acabamos de ver tienen la siguiente definición formal:

1. **Factor de elevación**: es el cociente entre el tamaño de la población y el tamaño de la muestra, $\frac{N}{n}$. Representa el número de elementos que hay en la población por cada elemento de la muestra.

2. **Factor de muestreo**: es el cociente entre el tamaño de la muestra y el tamaño de la población $\frac{n}{N}$. Si se multiplica por 100, obtenemos el porcentaje de la población que representa la muestra.

1.3. Muestreo aleatorio con y sin reemplazamiento

Ya hemos comentado que en caso de querer hacer muestreo de manera que la muestra sea representativa, debemos realizar muestreo probabilístico. ¿Cómo harías para seleccionar 28 alumnos de 560 dentro de un CES para que tuvieran todos la misma probabilidad de entrar en la muestra? Lo más sencillo sería hacer un sorteo para elegir 28, es decir, escogerlos al azar, así todos tendrían las mismas posibilidades de estar en la muestra.

Este proceso de selección corresponde a un muestreo aleatorio. Diremos que un muestreo es aleatorio cuando, el proceso de selección de la muestra garantice que todas las muestras posibles que se pueden obtener de la población tienen la misma probabilidad de ser elegidas, es decir, todos los elementos de la población tienen la misma posibilidad de ser seleccionados para formar parte de la muestra.

Cuando un elemento es seleccionado, y hemos medido las variables necesarias para el estudio y puede volver a ser seleccionado, se dice que hacemos un muestreo aleatorio con reemplazamiento o reposición. Generalmente recibe el nombre de muestreo aleatorio simple.

En caso de que el elemento no vuelva a formar parte de la población de manera que no puede volver a ser seleccionado se dice que se ha obtenido la muestra mediante un muestreo aleatorio sin reposición o reemplazamiento. En algunos libros, este método recibe también el nombre de muestreo irrestrictamente aleatorio.

Para nuestro ejemplo al elegir la muestra entre los 560 alumnos del CES, si vamos a preguntar por el hecho de que posean internet en casa, no nos interesa preguntarle dos veces a la misma persona, luego una vez elegido un elemento de la muestra no queremos volverlo a seleccionar. Realizaríamos pues un muestreo aleatorio sin reposición o sin reemplazamiento.

Aunque los dos métodos son diferentes, cuando el tamaño de la población es infinito, o tan grande que puede considerarse infinito, ambos métodos nos llevarán a las mismas conclusiones. Sin embargo, si la fracción de muestreo n/N es mayor que 0,1 (muestreamos más del 10 % de la población) la diferencia entre las conclusiones que se obtienen pueden ser importantes.

Al preguntar en nuestro ejemplo si los alumnos tienen o no internet en casa, nos interesa conocer tanto el número de alumnos que tiene internet como la proporción que eso supone dentro del centro. Estos dos valores, igual que la media para otros casos (por ejemplo si preguntamos por la altura), son los parámetros más calculados y que habitualmente queremos estimar. Para el caso del muestreo aleatorio tanto con reposición como sin reposición, estos estimadores vienen dados por las expresiones:

Total:

$$\widehat{X} = N \sum_{i=1}^{n} \frac{X_i}{n}.$$

Media:

$$\widehat{\overline{X}} = \sum_{i=1}^{n} \frac{X_i}{n}.$$

Proporción:

$$\widehat{P} = \sum_{i=1}^{n} \frac{P_i}{n}.$$

La proporción sería la media de una variable que toma valores cero o uno. En las anteriores expresiones:

X_i es el valor de la variable que estamos estudiando.

N es el tamaño poblacional.

n es el tamaño muestral.

P_i es una variable que toma los valores 0 ó 1.

La estimación del error para estos estimadores sería:

Total:

Para el muestreo con reposición:

$$\widehat{V}(\widehat{X}) = N^2 \frac{S^2}{n}.$$

Para el muestreo sin reposición:

$$\widehat{V}(\widehat{X}) = N^2 (1 - \frac{n}{N}) \frac{S^2}{n}.$$

Media:

Para el muestreo con reposición:

$$\widehat{V}(\widehat{\overline{X}}) = \frac{S^2}{n}.$$

Para el muestreo sin reposición:

$$\widehat{V}(\widehat{\overline{X}}) = (1 - \frac{n}{N}) \frac{S^2}{n}.$$

Proporción:

Para el muestreo con reposición:

$$\widehat{V}(\widehat{P}) = \frac{\widehat{P}\widehat{Q}}{n-1}.$$

Para el muestreo sin reposición:

$$\widehat{V}(\widehat{P}) = (1 - \frac{n}{N}) \frac{\widehat{P}\widehat{Q}}{n-1}.$$

1.4. Muestreo estratificado

Imagina ahora que queremos hacer una estudio para saber a qué dedican su tiempo libre las personas que viven en tu ciudad. Todos sabemos que los ancianos no realizan el mismo tipo de actividades que los jóvenes, ni tampoco que las personas de mediana edad, como por ejemplo tus padres. Nos interesaría entonces que toda esta información que tenemos de antemano nos ayude a construir una muestra más significativa. De hecho, nos interesa que todos esos colectivos estén representados en nuestra muestra. A los colectivos que hemos definido, en este caso por edad, los llamaremos estratos. Lo que haremos será dividir nuestra muestra de manera que haya representantes de todos los estratos. Vamos a definir rigurosamente la manera de hacer un muestreo en este caso.

Consideramos que tenemos la población de tamaño N dividida en k subpoblaciones de tamaños $N_1, N_2, \ldots, N_k$. Dichas subpoblaciones son disjuntas y cumplen que $N_1 + N_2 + \cdots + N_k = N$. Cada una de las subpoblaciones se denominan estratos. Si deseamos obtener una muestra de tamaño n de la población inicial, seleccionamos de cada estrato una muestra aleatoria de tamaño n_i de manera que $n_1 + n_2 + \cdots + n_k = n$.

¿Qué ventajas e inconvenientes presenta el muestreo estratificado? Las vemos a continuación.

Ventajas:

- Podemos tener información con más precisión dentro de las subpoblaciones sobre la característica objeto del estudio.
- Podemos aumentar la precisión de los estimadores de las características de toda la población.

Inconvenientes:

- La elección del tamaño de las muestras dentro de cada estrato para que el total sea n.
- La división en estratos en algunas poblaciones puede no ser sencilla.

En general, el muestreo estratificado proporciona mejores resultados que el muestreo aleatorio, mientras más diferentes sean los estratos entre sí y más homogéneos internamente.

Podemos considerar 3 métodos para distribuir el tamaño de la muestra entre los estratos:

1. Proporcionalmente al tamaño de cada estrato, es decir, si tomamos el estrato j-ésimo de tamaño N_j, entonces una muestra de dicho estrato será de tamaño $n \cdot (N_j/N)$, siendo N el total de la población y n el tamaño de la muestra.
2. Proporcionalmente a la variabilidad de la característica que estamos considerando en cada estrato. Por ejemplo, si conocemos que la varianza en la altura de los alumnos es de 15 cm y en las alumnas es de 5 cm, la proporción de los alumnos es 3 a 1 y la muestra deber guardar esa proporción.
3. Se asigna el mismo tamaño a cada estrato. Como consecuencia se favorece a los estratos más pequeños y se perjudica a los grandes en cuanto a precisión.

Para el caso del muestreo estratificado, los principales estimadores vendrían dados por las siguientes expresiones:

Total:

$$\widehat{X} = \sum_{h=1}^{k} N_h \overline{X}_h.$$

Media:

$$\widehat{\overline{X}} = \sum_{h=1}^{k} w_h \overline{X}_h = \sum_{h=1}^{k} \frac{N_h}{N} \overline{x}_h.$$

Proporción:

$$\widehat{P} = \sum_{h=1}^{k} w_h \widehat{P}_h,$$

donde

$\overline{X}_h$ es la media muestral de la variable X en el estrato h.

N_h es el tamaño del estrato h.

N es el tamaño poblacional.

n_h es el tamaño muestral en el estrato h.

n es el tamaño muestral.

$\widehat{P}_h$ es la proporción muestral de la variable en el estrato h.

y la estimación del error que cometemos al estimar los parámetros poblacionales viene dado por:

Total:

$$\widehat{V}(\widehat{X}) = \sum_{h=1}^{k} N_h^2 (1 - f_h) \frac{\widehat{S}_h^2}{n_h},$$

con

$$f_h = \frac{n_h}{N_h} \qquad \text{y} \qquad \widehat{S}_h^2 = \frac{n_h}{n_h - 1} \left[\frac{1}{n_h} \sum_{i=1}^{n_h} X_{hi}^2 - \overline{x}_h \right].$$

Media:

$$\widehat{V}(\widehat{\overline{X}}) = \sum_{h=1}^{k} w_h^2 (1 - f_h) \frac{\widehat{S}_h^2}{n_h},$$

donde w_h, f_h y S_h^2 tienen los mismos significados que antes.

Proporción:

$$\widehat{V}(\widehat{P}) = \sum_{h=1}^{k} w_h^2 (1 - f_h) \frac{\widehat{P}_h \widehat{Q}_h}{n_h - 1},$$

donde $\widehat{Q}_h = 1 - \widehat{P}_h$.

1.5. Muestreo por conglomerados

Nos planteamos hacer un estudio de la altura de los alumnos de Secundaria de tu ciudad. En lugar de hacer un muestreo de todos los chicos de tu ciudad podríamos plantearnos elegir algunos barrios, ya que con respecto a la altura, los barrios son como "pequeñas poblaciones" comparables a la ciudad. En este caso ¿podemos simplificar la elección de la muestra al elegir los barrios sin perder precisión? La respuesta es que en este caso, podríamos elegir barrios y analizar las alturas de los estudiantes de cada barrio sin perder precisión. Vamos a ver el método que nos lo permite.

En el muestreo por conglomerados, la población se divide en unidades o grupos, llamados conglomerados (generalmente son unidades o áreas en los que se ha dividido la población), que deben ser lo más representativas posible de la población, es decir, deben representar la heterogeneidad de la población objeto del estudio y ser entre sí homogéneos.

El motivo para realizar este muestreo es que a veces resultaría demasiado costoso realizar una lista completa de todos los individuos de la población objeto del estudio, o que cuando se terminase de realizar la lista no tendría sentido la realización del estudio.

El principal inconveniente que tiene es que si los conglomerados no son homogéneos entre sí, la muestra final puede no ser representativa de la población.

Suponiendo que los conglomerados sean tan heterogéneos como la población, en relación a las variables estudiadas, y que entre sí sean homogéneos, para obtener una muestra bastará con seleccionar algunos conglomerados. En este caso se habla de muestreo por conglomerados de una etapa.

El muestreo por conglomerados tiene la ventaja de simplificar la recogida de las informaciones muestrales.

Veamos ahora la expresión de los estimadores cuando trabajamos con esta técnica de muestreo.

Total:

$$\widehat{X} = M\frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} M_i}.$$

Media:

$$\widehat{\overline{X}} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} M_i}.$$

Proporción:

$$\widehat{P} = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} M_i},$$

donde

$\widehat{X}_i$ es el total de la variable X en el conglomerado i.

$\widehat{\overline{X}}_i$ es la media muestral de la variable X en el conglomerado i.

N es el número de conglomerados de la población.

M es el tamaño poblacional.

n es el número de conglomerados de la muestra.

M_i es el tamaño del conglomerado i.

A_i es el total de una variable A, que toma el valor 0 ó 1 en el conglomerado i,

y la estimación de los errores que cometemos al hacer estas estimaciones son los siguientes:
Total:

$$\widehat{V}(\widehat{X}) = \frac{N(N-n)}{n}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}M_i)^2.$$

Media:

$$\widehat{V}(\widehat{\overline{X}}) = \frac{N(N-n)}{M^2 n}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}M_i)^2.$$

Proporción:

$$\widehat{V}(\widehat{P}) = \frac{N(N-n)}{M^2 n}\frac{1}{n-1}\sum_{i=1}^{n}(P_i - \overline{P}M_i)^2.$$

1.6. Muestreo sistemático

Se nos puede ocurrir otra manera de muestrear. Imaginemos que en tu centro hay 560 alumnos y hemos decidido elegir una muestra de de 28 personas. En este caso el factor de elevación sería de $560/28 = 20$. Numeramos a los alumnos del 1 al 560. Elegimos entonces un número x al azar entre 1 y 20 y ese sería el primer alumnos seleccionado, el que ocupa el lugar x. Luego tomamos el $x + 20$, $x + 2 \cdot 20$ y así sucesivamente. No es un muestreo aleatorio porque todas las muestras no son igualmente probables. Vamos a definir este tipo de muestreo.

Supongamos que tenemos una población que consta de N elementos, ordenados y numerados del 1 hasta N, y deseamos obtener una muestra de tamaño n. Dicha población la podemos dividir en n subconjuntos, cada uno de ellos con $v = \frac{N}{n}$ elementos, es decir, cada subconjunto consta de tantos elementos como indica el factor de elevación.

Tomamos aleatoriamente un elemento de los enumerados desde 1, 2 hasta $\frac{N}{n}$, y lo llamamos x_0; después se toman los siguientes elementos $x_0 + v$, $x_0 + 2v$, $x_0 + 3v$, $x_0 + 4v \ldots$

En caso de que v no sea entero, se redondea al entero menor, con lo que puede que algunas muestras tengan tamaño $n - 1$. Este hecho introduce una pequeña perturbación en la teoría del muestreo sistemático, que es despreciable si $n > 50$.

Este tipo de muestreo requiere que previamente nos hayamos asegurado de que los elementos ordenados no presentan periodicidad en las variables objeto de estudio, puesto que si hay periodicidad y el período está próximo al valor v, los resultados que se obtengan tendrán grandes desviaciones y no tendrán validez.

El muestreo sistemático es equivalente al muestreo aleatorio si los elementos se encuentran enumerados de manera aleatoria.

Las ventajas de dicho método son:

1. Extiende la muestra a toda la población.
2. Es de fácil aplicación.

Los inconvenientes que presenta son:

1. Aumento de la varianza si existe periodicidad en la numeración de los elementos, produciéndose sesgo por selección.

2. Problemas a la hora de la estimación de la varianza.

Puede considerarse un caso particular del muestreo por conglomerados, estando cada uno de ellos formado por los siguientes elementos que ocupan en la lista el lugar:

Primer conglomerado: $1,\ 1+v,\ 1+2v,\ 1+3v,\ 1+4v, \ldots$

Segundo conglomerado: $2,\ 2+v,\ 2+2v,\ 2+3v,\ 2+4v, \ldots$

$\ldots$

v-ésimo conglomerado: $v,\ 2v,\ 3v,\ 4v, \ldots\ nv$.

Seleccionar una muestra sistemática equivale a seleccionar al azar un único conglomerado. Para ello es necesario que cada uno de los conglomerados definidos tengan una composición similar a la población.

También puede considerarse como un caso particular de muestreo estratificado con un número de estratos igual a n, cada uno de ellos con v elementos de manera que en cada estrato se elige un único elemento.

En el muestreo estratificado el elemento seleccionado en cada estrato es aleatorio, mientras que en el sistemático se elige de forma aleatoria al primer elemento quedando los restantes determinados por el factor v.

Los estimadores para este tipo de muestreo son:

Total:

$$\widehat{X} = v \sum_{i=1}^{n} X_i.$$

Media:

$$\widehat{\overline{X}} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Proporción:

$$\widehat{P} = \frac{1}{n} \sum_{i=1}^{n} P_i,$$

donde P es una variable que toma los valores 0 ó 1.

1.7. Otros tipos de muestreo

El muestreo bietápico es un caso particular de muestreo por conglomerados en el que en la segunda etapa no se seleccionan todos los elementos del conglomerado, sino que se seleccionan un determinado número de elementos de cada conglomerado de manera aleatoria. Los conglomerados de primera etapa se denominan unidades primarias, los de segunda etapa, secundarias.

El muestreo polietápico es una generalización del anterior, de manera que cada conglomerado puede estar formado a su vez por otros conglomerados y así sucesivamente en varias etapas.

En general, para realizar estudios complejos se utilizan los conceptos de estratificación, conglomerados y muestreo aleatorio. Por ejemplo, la población de un país se podría dividir en conglomerados (provincias, municipios, barrios) que pueden ser bastante heterogéneos internamente (por ejemplo, para estudiar la renta per cápita), pero bastante homogéneos entre sí. Luego es necesario clasificar estas unidades en estratos homogéneos (unidades primarias, por ejemplo los barrios). Cada una de estas unidades primarias se divide en nuevas unidades (bloques de casas) llamadas secundarias, que se dividen en las casas concretas. La muestra se tomaría:

1. Seleccionando una muestra estratificada, de cada estrato (barrios), se toma al menos uno.
2. Se eligen al azar varios bloques de casas dentro de cada barrio seleccionado.
3. Se toman aleatoriamente una o varias casas dentro de los bloques seleccionados.

Capítulo 2

Un ejemplo de aplicación de las técnicas de muestreo

Hemos decidido realizar un estudio en un Centro de Enseñanza Secundaria. Queremos conocer datos sobre el número de alumnos que son zurdos del centro, del número de alumnos que tienen internet en casa, de la altura de los alumnos del centro y de la paga que reciben semanalmente.

El hecho de estudiar el número de zurdos de un centro es una información útil para el propio centro, ya que éste debe disponer del equipamiento adecuado para ellos, por ejemplo, sillas de pala adaptadas.

La conexión a internet en casa es ya, en estos tiempos, un dato fundamental. Esta información puede ser utilizada tanto para sondear la posibilidad de ofrecerle al alumno material a través de internet, tanto para conocer el potencial acceso de éstos a material didáctico en la web.

El estudio de la altura es un clásico. Es interesante, de cualquier forma, conocer si realmente la altura evoluciona con los años y los españoles de hoy son más altos.

La paga es un dato social relevante. Es también interesante conocer de qué dinero disponen habitualmente los chicos de edades adolescentes para comprender a qué dedican su tiempo.

Con estas premisas, decidimos hacer un muestreo para poder obtener conclusiones sobre todos los alumnos del CES sin tener que preguntar a todos y cada uno de ellos. La información de la que partimos es de la distribución de alumnos por grupos y niveles en el centro:

	A	B	C	D	E	Total
1º ESO	33	20				53
2º ESO	20	15	30			65
3º ESO	20	15	26	14		75
4º ESO	27	27	25			79
1º Bach	33	28	30	31	23	145
2º Bach	30	34	32	31		127

Luego estamos trabajando con una población de 544 alumnos de un Instituto de Enseñanza Secundaria.

Partimos de una premisa, vamos a utilizar un tamaño de muestra de alrededor de 60 alumnos, que es el máximo que se nos permite y que nos parece suficiente para el estudio que vamos a realizar.

Ya podemos obtener la primera información entonces, nuestra fracción de muestreo será

$$f = \frac{n}{N} = \frac{60}{558} = 0{,}1102,$$

es decir, vamos a muestrear aproximadamente un 11 % de la población. Podemos calcular nuestra factor de elevación, que sería

$$E = \frac{N}{n} = \frac{544}{60} = 9{,}1,$$

o lo que es lo mismo, cada alumno entrevistado representa aproximadamente a 9 compañeros.

Ahora tenemos que decidir qué método utilizamos para muestrear para las diferentes características que vamos a estudiar. Vamos a llamarlas de la siguiente manera:

- X representará a la altura.
- Y representará a la paga.
- Z representará a la variable "ser zurdo"que valdrá 1 en caso de serlo y 0 en caso de ser diestro.
- I representa a la variable "tener internet en casa"que valdrá 1 en caso de que se tenga internet en casa y 0 en caso contrario.

Vamos a diferenciar dos casos de entre las 4 variables. Lo primero que nos hacemos es una pregunta: tenemos la población dividida en niveles y en grupos ¿podemos considerar que esta división tiene influencia en alguna de estas variables? Es decir, ¿podemos considerar que en cada nivel, por ejemplo, la media de las alturas podría variar? La respuesta a esta pregunta es que por lógica, sí que lo hará. A priori, podemos suponer que la edad es una variable que tiene una influencia importante para la altura. ¿Y para la paga? Pues también la edad es importante, puesto que a todos nos han ido aumentando la paga conforme hemos ido creciendo. ¿Ocurre lo mismo con el ser zurdo? Pues no, cuando uno es zurdo, lo es desde que nace, luego la edad no tiene ninguna influencia. Igual ocurre con el hecho de tener internet en casa. Nuestra técnica de muestreo elegida será pues, diferente para estos dos grupos de casos.

Caso I: Variables paga y altura

Ya hemos visto que tenemos la población dividida por cursos y por grupos. Para nosotros, la división en cursos es una división por *estratos* porque los cursos son homogéneos dentro de ellos con respecto a la edad (y podemos pensar que también con respecto a la paga y a la altura), y como hemos visto que la edad tiene influencia en nuestras dos variables, tiene sentido pensar que nos interesa que haya representantes de todos los estratos en nuestra muestra. Luego en estos casos, nuestra elección es un *muestreo aleatorio estratificado*.

Lo siguiente que debemos decidir es el tamaño muestral dentro de cada uno de los estratos, es decir, la afijación.

Tenemos 6 estratos con los siguientes tamaños:

Estrato	Tamaño
1º de ESO (estrato 1)	$N_1 = 53$
2º de ESO (estrato 2)	$N_2 = 65$
3º de ESO (estrato 3)	$N_3 = 75$
4º de ESO (estrato 4)	$N_4 = 79$
1º de Bachillerato (estrato 5)	$N_5 = 145$
2º de Bachillerato (estrato 6)	$N_6 = 127$

Lo más lógico en este caso es utilizar afijación proporcional, es decir, hacemos que los tamaños de los estratos guarden la mismas proporciones que los tamaños de los estratos. Calculamos entonces el tamaño de la muestra en cada estrato a través de la siguiente fórmula:

$$n_i = n \cdot \frac{N_i}{N},$$

luego obtenemos los siguientes tamaños muestrales:

$n_1 = 60 \cdot \frac{53}{544} = 5{,}84$ luego tomamos $n_1 = 6$,
$n_2 = 60 \cdot \frac{65}{544} = 7{,}16$ luego tomamos $n_2 = 8$,
$n_3 = 60 \cdot \frac{75}{544} = 8{,}27$ luego tomamos $n_3 = 8$,
$n_4 = 60 \cdot \frac{79}{544} = 8{,}71$ luego tomamos $n_4 = 8$,
$n_5 = 60 \cdot \frac{145}{544} = 15{,}99$ luego tomamos $n_5 = 16$,
$n_6 = 60 \cdot \frac{127}{544} = 14{,}00$ luego tomamos $n_6 = 14$,

donde los redondeos se han hecho para mantener el tamaño muestral 60 que habíamos acordado.

Luego ya tenemos los tamaños muestrales que necesitamos y podemos hacer un muestreo aleatorio dentro de cada estrato para seleccionar el número de alumnos que indica el correspondiente tamaño muestral del estrato.

Nuestros datos son los siguientes:

Para la altura, obtuvimos:

Estrato 1	165	161	153	150	151	153										
Estrato 2	157	161	168	162	165	171	169	164								
Estrato 3	168	165	175	175	165	163	165	165								
Estrato 4	164	171	177	163	170	165	160	175								
Estrato 5	175	173	161	158	175	164	158	161	158	171	175	170	187	168	170	185
Estrato 6	190	178	194	183	165	170	176	173	168	183	173	183	174	177		

Y para la paga:

Estrato 1	10	0	3.5	0	0	3										
Estrato 2	0	5	0	15	0	3	2	0								
Estrato 3	5	8	8	0	20	5	10	10								
Estrato 4	12	6	5	12	12	6	0	0								
Estrato 5	5	10	12	15	10	12	30	12	30	10	6	5	10	21	40	15
Estrato 6	12	10	9	6	8	9.4	15	0	20	10	15	10	0	0		

Vamos ahora a proceder a hacer las estimaciones. Lo primero que hacemos es calcular las medias de los diferentes estratos, que nos van dando información de cómo se comportan los diferentes estratos. Posteriormente, calcularemos la estimación de la media de altura y de paga de los alumnos

del centro y la acompañaremos de la estimación del error cometido al realizar dicha estimación. Hacemos el proceso independientemente para cada una de las dos variables:

Para la altura tenemos:

Estrato	Media	Cuasivarianza
1	$\overline{x}_1 = 155{,}5$	$S^2_{x1} = 36{,}7$
2	$\overline{x}_2 = 164{,}625$	$S^2_{x2} = 21{,}4107$
3	$\overline{x}_3 = 167{,}625$	$S^2_{x3} = 22{,}5535$
4	$\overline{x}_4 = 168{,}125$	$S^2_{x4} = 36{,}6964$
5	$\overline{x}_5 = 169{,}3125$	$S^2_{x5} = 81{,}6958$
6	$\overline{x}_6 = 177{,}642857$	$S^2_{x6} = 67{,}478$

A primera vista ya observamos un resultado curioso. La media es creciente según aumentamos de curso. Esto nos lleva a pensar que la elección de un muestreo estratificado ha sido adecuada para este caso.

Pasamos a calcular ahora media y cuasivarianza para la paga por estratos:

Estrato	Media	Cuasivarianza
1	$\overline{y}_1 = 2{,}75$	$S^2_{y1} = 4{,}026$
2	$\overline{y}_2 = 3{,}125$	$S^2_{y2} = 26{,}4107$
3	$\overline{y}_3 = 8{,}25$	$S^2_{y3} = 33{,}3571$
4	$\overline{y}_4 = 6{,}625$	$S^2_{y4} = 25{,}4107$
5	$\overline{y}_5 = 15{,}1875$	$S^2_{y5} = 101{,}2291$
6	$\overline{y}_6 = 8{,}8857$	$S^2_{y6} = 35{,}229$

Ahora calculamos la media estimada a partir de la muestra completa y la estimación del error en términos de la estimación de la varianza para las dos variables que estamos estudiando. Para la altura:

$$\widehat{\overline{X}} = \sum_{h=1}^{6} w_h \overline{x}_h = \sum_{h=1}^{6} \frac{N_h}{N} \overline{x}_h = \frac{53}{544} \cdot 155{,}5 + \frac{65}{544} \cdot 164{,}625 + \frac{75}{544} \cdot 167{,}625 + \frac{79}{544} \cdot 168{,}125$$

$$+ \frac{145}{544} \cdot 169{,}3125 + \frac{127}{544} \cdot 177{,}642857 = 168{,}9463.$$

La expresión de la varianza es

$$\widehat{V}(\widehat{\overline{X}}) = \sum_{h=1}^{k} w_h^2 (1 - f_h) \frac{\widehat{S}_h^2}{n_h}.$$

Para nuestro caso

Estrato	w_h	w_h^2	f_h	$1 - f_h$
1	$\frac{53}{544} = 0{,}095$	0.009	$\frac{6}{53} = 0{,}1132$	0.8868
2	$\frac{65}{544} = 0{,}1194$	0.014	$\frac{8}{65} = 0{,}123$	0.8769
3	$\frac{75}{544} = 0{,}1344$	0.018	$\frac{8}{75} = 0{,}1066$	0.8934
4	$\frac{79}{544} = 0{,}1415$	0.02	$\frac{8}{79} = 0{,}1012$	0.8988
5	$\frac{145}{544} = 0{,}2598$	0.0675	$\frac{16}{145} = 0{,}1103$	0.8897
6	$\frac{127}{544} = 0{,}2276$	0.0518	$\frac{14}{127} = 0{,}1102$	0.8898

Ahora sustituimos estas cantidades en la expresión de la estimación de la varianza y nos queda

$$\widehat{V}(\widehat{\overline{X}}) = \sum_{h=1}^{k} w_h^2(1-f_h)\frac{\widehat{S}_h^2}{n_h} = 0{,}009 \cdot 0{,}8868 \cdot \frac{36{,}7}{6} + 0{,}014 \cdot 0{,}8769 \cdot \frac{21{,}4107}{8} + 0{,}018 \cdot 0{,}8934 \cdot \frac{22{,}5535}{8}$$

$$+0{,}02 \cdot 0{,}8988 \cdot \frac{36{,}6964}{8} + 0{,}0675 \cdot 0{,}8897 \cdot \frac{81{,}6958}{16} + 0{,}0518 \cdot 0{,}8898 \cdot \frac{64{,}478}{14} = 0,728.$$

Luego para el caso de la altura ya tenemos nuestras estimaciones. La altura media estimada es 168.9463 y calculamos que cometemos un error de 0.728.

Pasamos ahora a hacer los mismos cálculos para la paga. Empezamos por calcular la media estimada:

$$\widehat{\overline{Y}} = \sum_{h=1}^{6} w_h \overline{y}_h = \sum_{h=1}^{6} \frac{N_h}{N}\overline{y}_h = \frac{53}{544} \cdot 2{,}75 + \frac{65}{544} \cdot 3{,}125 + \frac{75}{544} \cdot 8{,}25 + \frac{79}{544} \cdot 6{,}625$$

$$+\frac{145}{544} \cdot 15{,}1875 + \frac{127}{544} \cdot 8{,}8857 = 8{,}8633.$$

La estimación de la varianza la podemos calcular directamente ya que los valores de w_h y f_h son los mismos

$$\widehat{V}(\widehat{\overline{Y}}) = \sum_{h=1}^{k} w_h^2(1-f_h)\frac{\widehat{S}_h^2}{n_h} = 0{,}009 \cdot 0{,}8868 \cdot \frac{4{,}026}{6} + 0{,}014 \cdot 0{,}8769 \cdot \frac{26{,}4107}{8} + 0{,}018 \cdot 0{,}8934 \cdot \frac{33{,}3571}{8}$$

$$+0{,}02 \cdot 0{,}8988 \cdot \frac{25{,}4107}{8} + 0{,}0675 \cdot 0{,}8897 \cdot \frac{101{,}2291}{16} + 0{,}0518 \cdot 0{,}8898 \cdot \frac{35{,}229}{14} = 0{,}666.$$

Caso II: Variables 'Ser zurdo' y 'Tener internet en casa'

Ahora queremos estudiar las variables 'ser zurdo' y 'tener internet en casa'. Es obvio que la división en estratos no es efectiva en este caso, así que debemos pensar en otro tipo de técnica de muestreo. Seguimos queriendo muestrear alrededor de 60 alumnos. Podríamos pensar que frente a estas variables, los grupos en los que está dividida la población se comportan como pequeñas poblaciones, es decir, podemos considerar que los grupos se comportan aproximadamente como todo el centro. Además, nos resulta interesante la posibilidad de muestrear los grupos porque seleccionar una muestra aleatoria de alumnos, localizarlos y entrevistarlos no es una tarea sencilla.

Ahora bien, ¿qué son los grupos para nosotros? Pues ya hemos dicho que interiormente se comportan como pequeñas poblaciones con respecto a nuestras variables, mientras que entre ellos son similares. Estamos hablando de que tenemos la población dividida en conglomerados, luego para este caso aplicaremos el muestreo por conglomerados.

Lo siguiente que tenemos que decidir es el número de grupos que vamos a muestrear. Como los grupos no son regulares en tamaño, 2 ó 3 grupos nos asegurarían estar rondando los 60. Para evitar la posibilidad de muestrear dos de los grupos pequeños y que nos quede una muestra excesivamente pequeña para lo que pretendemos, decidimos seleccionar 3 grupos de entre todos los del centro.

Así pues, los datos que hemos obtenido son los siguientes. Para la variable 'ser zurdo':

Grupo 1:
1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0,
Grupo 2:
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
Grupo 3:
0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0,
donde 1 significa que es zurdo y 0 que no lo es. Ahora, para la variable tener internet en casa, hemos obtenido:
Grupo 1:
1 0 0 1 0 1 0 1 0 1 1 1 1 0 1 0 0 1 0 0,
Grupo 2:
1 1 1 0 1 1 0 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 0,
Grupo 3:
1 1 0 1 0 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1 1,
donde, en este caso, 1 significa que tienen internet y 0 significa que no tienen.

Pasamos ahora a calcular las estimaciones del total de zurdos y de la proporción de zurdos, así como del total de alumnos que tienen internet en casa y la proporción de alumnos que tienen internet en casa.

Calculamos el total y la proporción para cada grupo y cada variable:

	Zurdos		Internet	
Grupo	Total	Proporción	Total	Proporción
1	3	0.15	10	0.5
2	0	0	17	0.7391
3	2	0.08	20	0.8

Ahora ya podemos calcular las estimaciones de la proporción y del total de las variables Z e I. Comenzamos por la variable Z:

$$\widehat{Z} = M \cdot \frac{\sum_{i=1}^{n} \widehat{Z}_i}{\sum_{i=1}^{n} M_i} = 544 \cdot \frac{\sum_{i=1}^{3} \widehat{Z}_i}{\sum_{i=1}^{3} M_i} = 544 \cdot \frac{3+0+2}{20+23+25} = 544 \cdot \frac{5}{68} = 40,$$

$$\widehat{P_Z} = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} M_i} = \frac{3+0+2}{20+23+25} = \frac{5}{68} = 0{,}0735,$$

y ahora repetimos los cálculos para la variable I:

$$\widehat{I} = M \cdot \frac{\sum_{i=1}^{n} \widehat{I}_i}{\sum_{i=1}^{n} M_i} = 544 \cdot \frac{\sum_{i=1}^{3} \widehat{I}_i}{\sum_{i=1}^{3} M_i} = 544 \cdot \frac{10+17+20}{20+23+25} = 544 \cdot \frac{47}{68} = 376,$$

$$\widehat{P_I} = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} M_i} = \frac{10+17+20}{20+23+25} = \frac{47}{68} = 0{,}6911.$$

Pasamos ahora a calcular la estimación del error que estamos cometiendo en nuestros cálculos para la variable 'ser zurdo':

$$\widehat{V}(\widehat{Z}) = \frac{N(N-n)}{n} \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \overline{Z} M_i)^2 = \frac{21(21-3)}{3} \frac{1}{2} \left[(3 - 0{,}0735 \cdot 20)^2 + (0 - 0{,}0735 \cdot 23)^2 + (2 - 0{,}0735 \cdot 25)^2 \right]$$

$$= 329{,}18.$$

$$\widehat{V}(\widehat{P_Z}) = \frac{N(N-n)}{nM^2}\frac{1}{n-1}\sum_{i=1}^{n}(P_{Zi}-\widehat{P}M_i) = \frac{21(21-3)}{3\,(544)^2}\frac{1}{2}\left[(3-0{,}0735\cdot 20)^2+(0-0{,}0735\cdot 23)^2+(2-0{,}0735\cdot 25)^2\right]$$

$$= 0{,}00111.$$

Pasamos ahora a calcular los errores estimados para la variable 'tener internet en casa',

$$\widehat{V}(\widehat{I}) = \frac{N(N-n)}{n}\frac{1}{n-1}\sum_{i=1}^{n}(I_i-\overline{I}M_i) = \frac{21(21-3)}{3}\frac{1}{2}\left[(10-0{,}6911\cdot 20)^2+(17-0{,}6911\cdot 23)^2+(20-0{,}6911\cdot 25)^2\right]$$

$$= 1464{,}123.$$

$$\widehat{V}(\widehat{P_I}) = \frac{N(N-n)}{nM^2}\frac{1}{n-1}\sum_{i=1}^{n}(P_{Ii}-\widehat{P}M_i) = \frac{21(21-3)}{3\,(544)^2}\frac{1}{2}\left[(10-0{,}6911\cdot 20)^2+(17-0{,}6911\cdot 23)^2+(20-0{,}6911\cdot 25)^2\right]$$

$$= 0{,}00049.$$