Muestreo

PID_00269801

Àngel J. Gil Estallo

Universitat Oberta de Catalunya

Àngel J. Gil Estallo

Doctor en Ciencias Matemáticas por la Universidad de Barcelona desde el año 1996. Profesor titular de escuela universitaria de la Universidad Pompeu Fabra desde 1991. Su actividad docente se centra en temas de matemáticas, estadística e informática en los estudios de Economía de dicha Universidad. Consultor de la Universitat Oberta de Catalunya desde 1998.

La revisión de este recurso de aprendizaje UOC ha sido coordinada por la profesora: Mireia Besalú Mayol (2019)

Cuarta edición: septiembre 2019

Av. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: FUOC

Índice

Muestreo

1. Introducción

En esta sesión introduciremos los aspectos más relevantes que hay que tener en cuenta a la hora de tratar de obtener una muestra a partir de una población. El estudio de las técnicas de muestreo es muy complicado y muy importante, ya que la mayoría de los resultados teóricos se basan en la suposición de que disponemos de una "buena" muestra, representativa de las características globales de la población. Si la muestra no es representativa, las conclusiones que se puedan extraer de la misma serán poco correctas o simplemente nos inducirán a error.

Comenzaremos por recordar la distinción entre **población** y **muestra**. Después trataremos diferentes tipos de técnicas de muestreo, comenzando por la más importante de todas: el **muestreo aleatorio** simple. Esta técnica asegura que todos los individuos de la población tengan la misma probabilidad de ser escogidos y que los individuos se seleccionen de manera independiente unos de otros. También trataremos de forma detallada un método muy importante para obtener muestras aleatorias simples: el basado en **tablas de dígitos aleatorios**.

Una única muestra

De todas las posibles muestras de la población, trabajaremos normalmente con una **única muestra**. A partir de ésta debemos deducir toda la información posible sobre la población global.

A continuación aprenderemos a obtener muestras **por muestreo sistemático**, método más simple que el muestreo aleatorio simple.

Una vez fijados estos dos métodos, introduciremos algunos refinamientos que permiten incluir dentro del proceso de selección de la muestra algunas de las características conocidas de la población; en concreto estudiaremos las muestras **estratificadas** y las muestras por **conglomerados**:

1) En el primer caso (estratificación) se divide la población en grupos de manera que los elementos de cada grupo muestran un comportamiento similar, mientras que individuos de diferentes grupos muestran comportamientos diferentes.

2) En el segundo caso, todos los conglomerados (que pueden ser una agrupación física o geográfica) son similares entre sí, mientras que dentro de cada conglomerado los individuos muestran tanta heterogeneidad como en la población total.

Para acabar, comentaremos un tipo de muestreo llamado **muestreo por cuotas**, en el que se presenta una forma muy pragmática de obtener información.

2. Muestreo: población y muestra

En el estudio de numerosos hechos reales es conveniente considerar y definir con precisión el conjunto de individuos (sean personas, máquinas, motos o lo que sea) relevantes en nuestra investigación.

El conjunto de los individuos objeto de nuestro interés es lo que se denomina **población**.

El censo

En caso de que dispongamos de un listado de todos los individuos de la población, diremos que tenemos un **censo** de la población.

Normalmente, acceder a cada uno de los individuos de la población es imposible, bien porque la población es demasiado grande y resulta inviable económicamente, bien porque el tiempo necesario para recoger todos los datos hace que el estudio resulte inútil.

Motivaciones para el uso de muestras

a) Si estudiamos la duración de cierto tipo de disco duro, no tiene sentido esperar que se estropeen todos para estudiar la media de la duración de los discos.

b) Por otra parte, en el caso de las investigaciones sanitarias, por ejemplo, no podemos pretender suministrar un nuevo fármaco a todos los individuos de Cataluña para saber si tiene las propiedades requeridas.

c) Una situación similar se da en el llamado *control de calidad*: si se quiere controlar la calidad de la producción de cierto producto, por ejemplo, de los yogures producidos en una fábrica, no podemos abrir todos y cada uno de los yogures (ya que destruiríamos el producto e imposibilitaríamos su venta). Incluso en el caso de que queramos controlar sólo el peso de los yogures, no podemos pesarlos todos individualmente, ya que el proceso resultaría muy caro y lento.

En general, se suele seleccionar una muestra y estudiar sobre ésta la característica que nos interesa (la duración de los discos duros, el efecto de un tratamiento, el peso de los yogures, etc.).

Una **muestra** es cualquier subconjunto de la población objeto de nuestro estudio.

***Muestra* en inglés es *sample* y *muestreo* es *sampling*.**

Existen muchos tipos de muestras. Es fácil ver que si la población total está formada por N individuos y queremos muestras de k individuos, podemos formar $\binom{N}{k}$ diferentes.

En general, las muestras se utilizan para obtener información numérica sobre ciertas cantidades relacionadas con la población (y, por tanto, desconocidas *a priori*). Establecer procedimientos que garanticen que la muestra sea lo más representativa posible de la población es, pues, crucial.

Ejemplo de utilización de una muestra

Podemos estar interesados en la media de la altura de todos los catalanes. Resulta evidente que es imposible medir la altura de todos ellos: tendremos que recoger una muestra e intentar deducir el valor de la media poblacional a partir de la media obtenida con los individuos de la muestra.

En caso de que la muestra sea poco representativa de la población global, diremos que la muestra está **sesgada** (o que tiene **sesgo**).

Este sesgo acostumbra a darse cuando algunos sectores de la población están más representados dentro de la muestra que otros.

Ejemplos de errores que hay que evitar

Es muy fácil ver que ciertas situaciones académicas producirán claramente muestras sesgadas; pero en las situaciones de la vida real el sesgo puede no ser tan evidente. A continuación presentamos algunos errores que hay que evitar:

1) Imaginemos que queremos obtener información sobre el tiempo que los estudiantes de la UOC dedican a ver la televisión y que queremos obtener datos de 250 individuos. Una opción sería preguntar a los primeros 250 alumnos que llegasen a uno de los encuentros presenciales. Muy probablemente, esta muestra estaría sesgada, ya que: 1) sólo responderán algunos de los alumnos que efectivamente asistan al encuentro y, por tanto, el colectivo de los alumnos que no asistiesen no estaría representado; 2) los alumnos que asistan al encuentro pero lleguen tarde tampoco estarán representados (es posible que lleguen tarde precisamente porque estaban viendo la televisión por la noche, con lo que podría quedar fuera de la muestra un colectivo potencialmente muy interesante para nuestro estudio). Los especialistas en muestreo deben tener en cuenta tales cuestiones.

2) En el caso anterior podríamos limitarnos a enviar un mensaje a todos los estudiantes de la UOC y estudiar todas las respuestas que recibamos. Éste sería un caso de la llamada *respuesta voluntaria*. Las muestras obtenidas de esta manera suelen presentar sesgo a causa de las características de los individuos que contestan, ya que normalmente son los que están más posicionados (a favor o en contra) sobre el tema que se pregunta. En el caso del tiempo que se dedica a ver la televisión, es posible que en el conjunto de individuos que contestan estén sobrerrepresentados los que la ven mucho y subrepresentados los que la ven poco, porque nos les gusta (es posible que no respondan porque no dan importancia al tema, o porque no quieren perder el tiempo en un tema que no les interesa). Estos casos de respuesta voluntaria se dan mucho en las encuestas telefónicas (o virtuales) de la televisión o la radio; puesto que se ignora a los individuos que no escuchan el programa y los que responden suelen ser personas muy interesadas en el tema, las conclusiones que se pueden extraer no son demasiado fiables en general.

3) Un ejemplo clásico y muy real. En 1936 en Estados Unidos se obtuvo una muestra de millones de votantes con la que se pronosticó la derrota de Roosvelt; otra muestra, mucho más modesta, de sólo miles de electores, sirvió para pronosticar su victoria. Finalmente, ¡ganó Roosvelt por mayoría aplastante y con unos resultados similares a los predichos por la segunda muestra! El problema residía en que la primera muestra se obtuvo telefónicamente, en una época en la que disponer de teléfono era sinónimo de un estatus social que favorecía una tendencia específica de voto. Hoy en día, en cambio, las encuestas telefónicas pueden llegar a ser muy precisas.

Así pues, resulta que no todas las muestras que se pueden extraer de una población son útiles desde el punto de vista estadístico. Ahora comenzaremos el estudio, muy descriptivo y resumido, de diferentes técnicas de muestreo que permiten evitar (o al menos reducir) el impacto sobre el resultado final de los errores que acabamos de mencionar.

3. Muestreo aleatorio simple

Por su importancia, comenzaremos por introducir el concepto de muestreo aleatorio simple.

Se dice que se ha obtenido una **muestra aleatoria simple** cuando el proceso para obtenerla garantiza estas dos propiedades:

1) Todos los elementos de la población tienen la misma probabilidad de formar parte de la muestra.

2) Los elementos se seleccionan de uno en uno y con reposición, de manera que las selecciones se hacen siempre sobre el total de la población.

Los efectos de la reposición

El hecho de que haya reposición, es decir, que un individuo escogido pueda volver a ser escogido, tiene mucha importancia desde el punto de vista teórico, aunque habitualmente se considera que si la muestra es menor que el 10% de la medida de la población, da igual que actuemos con reposición o sin ella.

La primera condición asegura que no hay individuos "privilegiados", que tengan más tendencia a estar representados que otros y, por tanto, mejora la representatividad en la muestra. La segunda condición garantiza la independencia de las selecciones, ya que el hecho de tener que escoger a un individuo no modifica las posibilidades de que los otros individuos de la población sean escogidos.

¿Cómo podemos obtener una muestra aleatoria simple a partir de una población dada? Lo primero que tenemos que hacer es conseguir una lista de todos los individuos de la población. Habitualmente se asigna un número a cada individuo para facilitar el trabajo; a continuación podemos, por ejemplo:

1) Escribir cada número (1, 2, ..., *N*) en una papeleta, introducir las papeletas en una urna, mezclarlas perfectamente e ir sacando *k* papeletas, de una en una y cuidando de reintegrar la papeleta a la urna después de cada extracción.

2) También se puede hacer un sorteo con bolas numeradas de 1 a *N*, a condición de que se reintegren al bombo una vez que han salido. En este caso el bombo ayuda (a fuerza de darle vueltas) a que las bolas estén repartidas al azar. Esto es cierto siempre que todas las bolas sean idénticas (excepto en el número), pesen igual y si el bombo es perfecto.

En ambos casos resulta evidente que se satisfacen las dos condiciones de muestra aleatoria simple.

En caso de que no dispongamos de papeletas ni de bolas ni bombos, podemos optar por utilizar las llamadas *tablas de dígitos aleatorios*, que representan el procedimiento más habitual (si no se utiliza un ordenador, claro) para obtener una muestra aleatoria simple dentro de una población finita.

3.1. Elección de una muestra aleatoria: uso de tablas de dígitos aleatorios

Obtención automática de muestras

La mayoría de los programas estadísticos incorporan la posibilidad de extraer muestras del tamaño deseado a partir del conjunto de las observaciones de una variable.

Imaginemos ahora que disponemos de un censo de nuestra población (es decir, de una lista de los *N* individuos de la población) y que queremos extraer una muestra aleatoria de esta población de *k* individuos. Necesitamos buscar un procedimiento que nos ayude a extraer esta muestra de la manera más sencilla posible. La manera habitual consiste en utilizar una tabla de dígitos aleatorios.

A continuación mostraremos un ejemplo de cómo se utiliza este tipo de tablas para obtener muestras aleatorias simples.

Imaginemos que disponemos de la lista de los 1.400 alumnos matriculados en una universidad y que queremos extraer una muestra de doce estudiantes. Lo primero que hay que hacer es identificar a cada estudiante con un número.

A continuación mostraremos cómo obtener doce números de cuatro dígitos que determinarán los doce individuos seleccionados en la muestra. Lo haremos de la manera siguiente:

Consultad la tabla de dígitos aleatorios en el anexo de esta sesión.

- Escogeremos un punto por donde comenzar a leer la tabla. Podemos hacerlo lanzando un dado y comenzando la tabla por el dígito correspondiente al resultado del dado. Imaginemos que sacamos un tres; esto significa que el dígito que está en la posición tres (indicado en negrita) es el primero que consideramos para calcular cuáles serán los otros individuos de la muestra:

  19**2**23 95034 05756 28713 96409 12531 42544 82853

- Puesto que necesitamos números de cuatro cifras para identificar a nuestros alumnos, iremos formando grupos de cuatro cifras a partir del dígito que hemos obtenido como punto de partida; el primer grupo de cuatro cifras que obtenemos a partir del punto de partida es el 2239. Dado que éste no se corresponde con ningún individuo de la población (sólo tenemos 1.400 estudiantes) nos lo saltamos y miramos la siguiente agrupación de cuatro dígitos, que es el 5034, que también nos saltamos; en cambio, la agrupación de cuatro dígitos siguiente es 0575, que sí se corresponde con cierto estudiante. Así pues, el estudiante de número 575 será el primer individuo de la muestra. Continuaremos el proceso saltando el 6287 e incluyendo en la muestra como segundo estudiante el que tiene el número 1.396.

- Después continuaremos de la misma manera hasta obtener a los doce individuos que necesitamos en la muestra.

La elaboración de tablas

Como curiosidad, podemos mencionar que la tabla de dígitos aleatorios que aparece en el libro de D. Peña y J. Romo, *Introducción a la Estadística para las Ciencias Sociales* (Ed. McGraw-Hill) ha sido construida:

"poniendo los números premiados en los sorteos de la lotería uno detrás de otro" (pág. 268).

Observad que la clave del proceso consiste en la elaboración de las tablas que, en definitiva, son las que deben garantizar la aleatoriedad de todo el proceso. Estas tablas se suelen generar por ordenador y deben superar varias pruebas de aleatoriedad y de independencia entre los dígitos que aparecen en las tablas.

4. Muestreo sistemático

El muestreo sistemático es un procedimiento más simple para obtener una muestra. Supongamos que queremos seleccionar una muestra de tamaño k de una población de n individuos. Este procedimiento se basa en los puntos siguientes:

Recordad que...

... la parte entera de x, denotada por $[x]$, se obtiene truncando el número x a 0 decimales.

1) Se numeran, como en el caso anterior, los individuos de la población, de 1 a N.

2) Se calcula $m = [N / k]$, donde $[x]$ designa la parte entera del número x.

3) Se selecciona al azar un número entre 1 y m, que indicará el primer individuo que formará parte de la muestra.

4) Vamos sumando m tantas veces como sea necesario al número que indica el primer individuo de la muestra e incluimos en la muestra a los individuos que se correspondan con los resultados de estas sumas.

5) Tanto el número que determina el primer individuo que forma parte de la muestra como la cantidad m (que determina los intervalos fijos que sirven para seleccionar a los otros individuos de la muestra), garantizan que se obtenga el número de individuos necesario en la muestra, a la vez que se recorre toda la lista.

Ejemplos de muestreo sistemático

1) Muestra aleatoria de doce individuos con muestreo sistemático

Supongamos que queremos obtener una lista de doce estudiantes entre los 1.400 de cierta universidad utilizando muestreo sistemático. Seguiremos el procedimiento siguiente:

– Numeramos a los estudiantes del 1 al 1.400.

– Seleccionamos al azar un número menor o igual que $m = [1.400 / 12] = 116$; este número corresponde al primer individuo que seleccionaremos para la muestra. Supongamos que utilizando algún sistema que garantice la aleatoriedad obtenemos que el primer individuo que aparece en la muestra es el que ocupa el lugar 10.

– A continuación sumamos tantas veces como sea necesario $m = 116$ a partir del primer individuo. Así, la muestra estará formada por los estudiantes que ocupan las posiciones:

10, 10 + 116 = 126, 10 + 2 * 116 = 242, 10 + 3 * 116 = 358…

Y así hasta obtener a los doce individuos de la muestra.

2) Muestra aleatoria de 400 individuos usando muestreo sistemático sobre la guía de teléfonos.

"Supongamos que hay 834.781 abonados en las guías de teléfonos de Barcelona y que queremos una muestra aleatoria de 400 abonados [...] Puesto que 834.781 dividido por 400 es aproximadamente 2.086, podemos coger cada número de la guía que se encuentre en la posición 2.086., lo que nos dará una muestra de 400 abonados por todas las entradas de la guía. Para comenzar la selección, elegimos un número al azar entre 1 y 2.086 [...]; supongamos que este número es el 731. Buscamos el número que ocupa la posición 731 en la guía y después la entrada número 731 + 2.086 = 2.817, luego 2.817 + 2.086 = 4.903, y así sucesivamente. [...] Podemos añadir algunos atajos de sentido común para facilitar un poco más la tarea.

Contar 2.086 entradas en la guía cada vez es pesado, y un pequeño cambio en el diseño del muestreo anterior no le resta validez, siempre que el cambio se establezca al principio, antes de que el muestreo comience. Por ejemplo, supongamos que, al contar cuántas entrada hay en unas cuantas páginas de la guía, encontramos que la media es de 205 entradas por página, es decir, 2.086 entradas hacen unas 10 páginas, con un restante de 36. Después, desde el punto inicial del muestreo, simplemente contad diez páginas en la misma posición de la página y después contad 36 entradas para llegar a la unidad siguiente de la muestra."

M. Greenacre; F. Udina. *Estadística I*. Universitat Oberta de Catalunya.

5. Muestreo estratificado

En los tipos de muestreo estudiados hasta ahora no se ha tenido en cuenta el conocimiento previo que pudiésemos tener de las características de la población. De hecho, tanto en el muestreo aleatorio simple como en el sistemático prevalecen la aleatoriedad de la muestra escogida sobre otras consideraciones que pueden mejorar su representatividad.

Ahora estudiaremos el concepto de estratificación, concepto que integra, en la selección de la muestra, el posible conocimiento de la similitud de los valores de la variable en ciertos colectivos (estratos).

Ejemplo de muestreo estratificado: conexión gratuita a Internet

Supongamos un caso extremo para ilustrar mejor el concepto de muestreo estratificado. Imaginemos una población de 1.000 habitantes en la que sólo 500 individuos disponen de conexión a Internet en casa. Imaginemos que los 500 que tienen conexión están a favor de que las llamadas para conectarse sean gratuitas y que los que no tienen conexión están en contra. Es decir, en el global de la población una mitad está a favor de la gratuidad y la otra mitad, en contra.

Supongamos que extraemos una muestra aleatoria simple de diez individuos de la población; no sería nada extraño obtener seis a favor de la gratuidad y cuatro en contra (de la misma manera que, cuando lanzo al aire una moneda diez veces, no resulta sorprendente obtener seis caras y cuatro cruces). En este caso, en la muestra está a favor de la gratuidad un 60%, mientras que en la población la proporción de los que están a favor de la gratuidad es sólo del 50%. Así pues, resulta evidente que el hecho de disponer de conexión en casa condiciona la respuesta dada a la pregunta, por lo que sería conveniente explorar a la población según estratos. En este caso tenemos dos: el de quienes tienen conexión y el de quienes no la tienen. Así pues, para evitar el sesgo, cada estrato debería aportar el 50% de la muestra, tal como sucede en la población.

Una manera fácil de evitar el sesgo que aparece en los casos de estratificación es hacer que la muestra contenga las mismas proporciones de individuos con características comunes diferentes; de este modo las proporciones de individuos en las muestras serán las mismas que en la población global.

Ejemplo de la necesidad de los estratos

Supongamos que queremos estudiar el tiempo que los estudiantes de la UOC dedican a navegar por Internet. Si aceptamos, como afirman algunos estudios, que los hombres le dedican más tiempo que las mujeres, podríamos desear que en nuestra muestra la proporción de géneros fuese la misma que en el total de la UOC, para poder obtener así una estimación más fiable. Así pues, deberíamos estratificar por género.

En general, si sospechamos que la respuesta a cierta pregunta depende de una característica de los individuos y disponemos de una lista de la población en la que se indica esta característica, el muestreo estratificado intenta reproducir en la muestra la estructura de la población en relación con esta característica.

Reproducción de la estructura probable de la población en la muestra estratificada

Imaginemos que sospechamos que estar a favor o en contra del hecho de que el AVE pase por el aeropuerto de Barcelona depende del lugar de residencia (quienes viven en Barcelona tienen tendencia a estar a favor y quienes viven fuera de Barcelona tienden a estar en contra). En este caso la población está dividida en dos estratos (los que viven en Barcelona y los que viven fuera). Si queremos obtener una muestra estratificada, escogeremos dentro de cada estrato, a partir de un censo de sus miembros, una muestra aleatoria simple para poder completar así una muestra de la población de la medida deseada. Escogeremos las medidas de las muestras aleatorias dentro de cada estrato de manera que en la muestra global estén representados los estratos en la misma proporción que en la población.

En el **muestreo estratificado** los individuos de la población se dividen en grupos disyuntos (llamados **estratos**). La muestra se obtiene seleccionando una muestra aleatoria simple dentro de cada estrato.

En este tipo de muestreo se plantea el problema de cuál debe de ser la medida de la muestra dentro de cada estrato. Habitualmente se suele pedir que, en la muestra, el número de individuos de cada estrato esté en proporción al peso del estrato dentro de la población. Este tipo de muestreo será más preciso que el muestreo aleatorio simple, y más si los individuos de cada estrato son muy similares entre ellos y muy diferentes de los individuos de otros estratos.

Ejemplo de muestreo estratificado: estratificación por curso

Si estudiamos el tiempo de conexión a Internet de los alumnos de secundaria de los institutos públicos de cierto barrio, podemos sospechar que el tiempo de conexión depende del curso en el que están matriculados (bien porque en algún curso se hace una asignatura de informática que incluye el uso de Internet, bien porque la edad lo favorece). Si queremos una muestra de cien estudiantes estratificada por "curso en el que están matriculados", deberíamos conseguir el número de alumnos matriculados en cada curso y distribuir a los individuos de la muestra según la proporción de alumnos en cada curso. Supongamos que los alumnos se distribuyen de esta manera:

		Porcentaje sobre el total
1.º ESO	200	16,19%
2.º ESO	250	20,24%
3.º ESO	260	21,05%
4.º ESO	250	20,24%
1.º bachillerato	150	12,15%
2.º bachillerato	125	10,12%
	1.235	

Para obtener ahora cuántos individuos de la muestra deben ser de cada curso (estrato), repartiremos el total de individuos en la muestra (cien) según el porcentaje que corresponde a cada estrato:

	Porcentaje sobre el total	Individuos en la muestra
1.º ESO	16,19%	16
2.º ESO	20,24%	20
3.º ESO	21,05%	21
4.º ESO	20,24%	20
1.º bachillerato	12,15%	12
2.º bachillerato	10,12%	10
		99

Observamos que a causa del redondeo "hemos perdido" un individuo. Para obtener los cien que necesitamos, añadiremos un alumno a un estrato escogido al azar. Después deberíamos obtener, a partir de la lista de los individuos matriculados en cada curso, una muestra aleatoria simple de la medida correspondiente a cada curso.

Entre los otros criterios para la obtención...

... de la muestra que aporta cada estrato, podemos destacar aquél según el cual todos los estratos aportan el mismo número de individuos a la muestra final.

Podemos resumir el procedimiento para obtener una muestra estratificada de la población de la forma siguiente:

1) Los individuos de la población se agrupan en estratos disyuntos.

2) La muestra se obtiene asignando un número de individuos a cada estrato y después se seleccionan los individuos necesarios por muestreo aleatorio simple dentro de cada estrato.

3) El número de individuos que cada estrato aporta a la muestra final se puede decidir según diferentes criterios; normalmente se hace de forma proporcional a la medida relativa del estrato en la población.

6. Muestreo por conglomerados

Imaginemos que queremos estudiar el tiempo que los estudiantes de secundaria de **todos** los institutos de Cataluña dedican a navegar por Internet, seleccionando una muestra de trescientos estudiantes. Fácilmente podemos encontrarnos con dos problemas:

- Es difícil disponer de la lista de todos los estudiantes de secundaria de toda Cataluña.

- Aunque tuviéramos la lista y seleccionásemos una muestra aleatoria simple de la población, la recogida de información sería muy compleja, con el coste económico que esto representa. En esta situación se podría dar el caso de tener que visitar trescientos institutos diferentes para entrevistar a un único alumno en cada uno.

En casos como éstos, podemos optar por utilizar muestreo por conglomerados.

Los **conglomerados** son unidades (normalmente físicas o geográficas) en las que se distribuyen los individuos de la población que hay que investigar. En el **muestreo por conglomerados** se selecciona una muestra aleatoria de conglomerados y, dentro de cada conglomerado, se selecciona al azar una muestra de sus individuos.

Normalmente el muestreo por conglomerados se hace en diferentes etapas.

Ejemplo de muestreo por conglomerados

En el caso de los institutos, y con el objetivo de obtener una muestra de estudiantes de toda Cataluña, podríamos proceder de la manera siguiente:

1) Seleccionamos al azar unas cuantas comarcas de Cataluña (conglomerado = comarca).

2) Dentro de cada comarca seleccionamos al azar algunos institutos (conglomerados = institutos).

3) Dentro de cada instituto seleccionamos al azar unas clases (conglomerados = clases).

4) En las clases seleccionadas, y a partir de la lista de clase, seleccionamos una muestra aleatoria simple de sus estudiantes. Observad que la lista de los alumnos de una clase es fácil de conseguir, mientras que la lista de todos los estudiantes de primero de ESO en todos los institutos catalanes será más complicada de obtener.

5) Uniendo las muestras obtenidas en cada clase, obtenemos la muestra global.

Este método simplifica en gran medida la recogida de información muestral, pero presenta también varios inconvenientes:

a) Si los conglomerados son muy diferentes unos de otros, y teniendo en cuenta que no todos los conglomerados están representados en la muestra final, la muestra puede perder representatividad.

b) Cada conglomerado debe contener tanta diversidad como la misma población; dicho de otra manera, en cada conglomerado deben estar representadas (como si se tratase de una muestra aleatoria simple) todas las características de la población. En caso de que en algún conglomerado sólo hubiera individuos de alguna característica particular, esta característica se podría ver sobrerrepresentada o subrepresentada en la muestra (dependiendo de si el conglomerado aporta individuos de este tipo a la muestra final o no).

Podemos resumir el procedimiento para obtener una muestra por conglomerados de la manera siguiente:

1) Se estudia a la población distribuida en conglomerados, que son agrupaciones naturales de los individuos.

2) Se seleccionan al azar algunos de los conglomerados.

3) Dentro de los conglomerados seleccionados se toma una muestra aleatoria simple de sus individuos o bien se escoge a todos los individuos del conglomerado.

4) La muestra final es la reunión de las muestras obtenidas en cada conglomerado.

Finalmente, hay que insistir en la diferencia entre estrato y conglomerado:

Observad...

... la diferencia entre el uso de estratos y de conglomerados.

a) En los estratos se tienen en cuenta grupos dentro de la población que hay que investigar, cuyos individuos presentan características comunes y diferenciadas de los individuos de otros grupos.

b) Los conglomerados representan agrupaciones de la población (por zonas geográficas o por proximidad generalmente) en las que se reproducen todas las características de la población; así pues, los conglomerados deben ser similares unos con otros y cada uno de éstos debe contener individuos tan heterogéneos como si se tratase de la población.

Por otra parte, la construcción de la muestra garantiza que todos los estratos estén representados en ella, mientras que no todos los conglomerados aportan individuos a la muestra final.

7. Muestreo polietápico

A continuación comentaremos un tipo de muestreo que combina los métodos de estratificación y de conglomerados, lo que hace que se mejore la representatividad de la muestra, al mismo tiempo que se mantiene la simplicidad en la recogida de los datos.

El problema se plantea cuando los conglomerados resultan ser muy homogéneos con respecto a la característica que hay que tratar: por ejemplo, agrupaciones físicas muy utilizadas, como barrios o comarcas, agrupan en algunos casos a individuos con características socioeconómicas similares. Esto puede provocar un sesgo en la muestra si por azar sólo seleccionamos barrios o comarcas de cierta categoría socioeconómica.

Ejemplo de sesgo en el muestreo polietápico

Si sólo seleccionamos barrios de clase alta (como los que todos conocemos en nuestra ciudad), los resultados serán muy diferentes que si seleccionamos barrios no tan favorecidos.

Ejemplo de muestreo polietápico

En primer lugar, el acceso a Internet desde casa depende de tener ordenador en casa, y esto depende, a su vez, del nivel socioeconómico y cultural de las familias. Resulta evidente que no todos los barrios serán similares con respecto a las facilidades de conexión en casa.

Si queremos muestrear el tiempo que dedican a navegar por Internet desde su casa los estudiantes de cierta gran ciudad (por ejemplo, Barcelona) y tomamos como conglomerados los barrios, podemos encontrar que los individuos dentro de cada barrio tienen comportamientos similares, mientras que el comportamiento de los individuos de barrios diferentes también puede ser muy diferente. En los barrios de clase alta es posible que las mejores condiciones económicas hagan que los estudiantes puedan acceder a Internet con mayor facilidad que los de otros barrios y, por tanto, dedicarle más tiempo.

Para eliminar parcialmente este problema, lo que podemos hacer es agrupar los conglomerados por estratos.

Ejemplo de agrupación de conglomerados por estratos

En el caso de los barrios, podemos agruparlos en estratos, en este caso por nivel socioeconómico. Así pues, si quisiéramos aplicar una estratificación polietápica en el ejemplo del acceso a Internet desde casa, podríamos proceder de la manera siguiente:

- Distribuimos los barrios de Barcelona en *estratos* (por ejemplo: nivel económico alto, nivel económico medio, nivel económico bajo). Dentro de cada estrato escogemos una muestra aleatoria de barrios. En este caso también sería conveniente tener en cuenta a la población de cada barrio, de manera que en cada estrato, barrios más poblados tengan más probabilidad de ser escogidos que los menos poblados. También hay que escoger al menos un conglomerado (barrio en este caso) de cada estrato.

- Después, dentro de cada barrio seleccionamos aleatoriamente institutos (conglomerado = instituto). En este paso se podría proceder directamente sin estratificar si consideramos que los institutos tienden a reproducir la población de cada barrio; es decir, en caso de que tengamos motivos para pensar que los institutos de cada barrio son similares entre sí y que contienen tanta diversidad entre sus estudiantes como la misma población del barrio, podemos optar por no estratificar. (Estratificar o no los institutos depende de la información de que dispongamos y, muchas veces, de la experiencia previa en este tipo de estudios.)

- Acto seguido continuaríamos con las clases y, finalmente, seleccionaríamos una muestra aleatoria dentro de cada clase.

En resumen, en el muestreo polietápico se combinan la idea de estratificación y la de conglomerados, y se estratifican los conglomerados considerados, para después obtener muestras aleatorias de conglomerados dentro de cada estrato.

Acabaremos la presentación de este tipo de muestreo con otro ejemplo:

> "Por ejemplo, se desea tomar una muestra de la población española para estudiar la proporción de personas que están de acuerdo con las relaciones prematrimoniales. Si suponemos que la edad y el sexo pueden influir en la opinión, deberíamos tomar una muestra donde todas las características sean las mismas que en la población base, lo que implica una muestra estratificada. Por otro lado, si suponemos que las provincias son homogéneas respecto a la opinión, podemos ahorrar muchos costes seleccionando al azar 4 provincias y dentro de cada una de ellas una muestra aleatoria, o mejor estratificada. Este procedimiento tiene el inconveniente obvio de que si las provincias no son homogéneas con respecto a la opinión (por ejemplo las provincias más ricas tienen opinión distinta de las más pobres) tendremos sesgos (que evitaremos estratificando las provincias por riqueza)."
>
> D. Peña. *Estadística. Modelos y métodos 1. Fundamentos*. Alianza Universidad Textos.

8. Muestreo por cuotas

Cuando la estratificación no es posible o resulta muy cara, así como en casos en los que no disponemos de una lista de la población que hay que investigar, se puede recurrir al llamado *muestreo por cuotas*.

Imaginemos que ahora queremos estudiar el tiempo de conexión a Internet desde su casa que dedican todos los jóvenes de Barcelona de entre quince y veintitrés años. Fácilmente podemos sospechar que el género y la edad (vinculada a la obtención de trabajo) son características que influyen en el tiempo de conexión. Si quisiéramos estratificar, pues, según el género y la edad, tendríamos que comenzar por conseguir una lista de la población de entre quince y veintitrés años que incluyera, como mínimo, la edad y el sexo. Si bien esta lista puede no estar disponible o resultar demasiado cara, los datos estadísticos referentes a la distribución por edad y sexo de cierta población suelen estar disponibles en los mismos servicios municipales; de esta manera podríamos saber la cantidad de jóvenes en cada una de las edades consideradas y la proporción de géneros dentro de cada grupo de edad. Con estos datos sería muy fácil deducir cuántos individuos de cada combinación (edad-género) debe contener una muestra para reproducir las proporciones reales de la población.

Ejemplo de muestreo por cuotas: cuotas por género y edad

Consideramos el tiempo de conexión a Internet desde sus casas de los jóvenes de Barcelona entre quince y veintitrés años. En el caso que nos ocupa obtenemos datos como éstos:

Edad de año en año de la población de Barcelona por sexo			
Edades	**Total**	**Hombres**	**Mujeres**
15 años	16.362	8.390	7.972
16 años	17.340	8.837	8.503
17 años	18.888	9.663	9.225
18 años	20.338	10.209	10.129
19 años	21.538	10.908	10.630
20 años	22.813	11.614	11.199
21 años	24.098	12.304	11.794
22 años	23.862	12.087	11.775
23 años	23.986	12.131	11.855
Total	189.225	96.143	93.082

Fuente: Padrón Municipal de Habitantes 1996. Departamento de Estadística. Ayuntamiento de Barcelona. Obtenido de la página web www.bcn.es.

A partir de estos datos resulta claro que para obtener una muestra por cuotas (de, por ejemplo, 300 individuos), ésta deberá contener (redondeando convenientemente) 300 * (8.390 / 189.225) = 13 hombres de quince años, 300 * (7.972 / 189.225) = 13 mujeres de quince años, y así sucesivamente.

El paso siguiente será distribuir este número de individuos (o cuotas) entre los entrevistadores, de manera que cada entrevistador deberá conseguir tantos individuos de cada par (edad-género) como le marquen las cuotas asignadas.

Distribución de individuos entre los entrevistadores

Si tenemos tres entrevistadores, enviaremos a cada uno a diferentes zonas y encargaremos al primero que entreviste a seis hombres de quince años; al segundo, a cuatro hombres de quince años y al tercero, a tres, por ejemplo. Haremos esta distribución por cada par edad-género. Los entrevistadores irán preguntando hasta que rellenen las correspondientes cuotas. Así, una vez que el primer entrevistador haya conseguido la opinión de seis hombres de quince años, ya no anotará ninguna respuesta más de este colectivo.

En el **muestreo por cuotas** se distribuyen los individuos de la población en diferentes categorías y se asigna un número de individuos a cada categoría, de manera que la proporción de individuos de cada categoría en la muestra sea similar a la proporción dentro de la población. Una vez calculadas estas proporciones, el entrevistador recibe instrucciones sobre el número de individuos que debe entrevistar a cada categoría (cuota). Cuando ha cubierto los individuos de cierta categoría que tiene asignados, deja de recoger datos de ésta y continúa con las otras categorías.

Características del muestreo por cuotas

En los casos estudiados hasta ahora, el entrevistador (que es quien formula las preguntas) recibía una lista de las personas a las que tenía que entrevistar. En el muestreo por cuotas el entrevistador interviene en la selección de las personas de la muestra, respetando siempre las cuotas que se le asignan.

El muestreo por cuotas es muy utilizado en estudios de hábitos de consumo, de *marketing*, etc. por su simplicidad y por la facilidad en la recogida de datos (que también hace que sea mucho más barato). En general, la información re-

cogida no es lo bastante fiable para hacer un estudio estadístico detallado, pero sí resulta útil para hacer una primera aproximación al tipo de respuesta que podemos obtener y para poder diseñar una recogida de datos más fiable.

9. Resumen

En esta sesión hemos estudiado diferentes formas de obtener muestras a partir de una población fijada. En la tabla siguiente recogemos los diferentes métodos introducidos, sus principales características y un recordatorio de los ejemplos que se han tratado en cada método:

Tipo de muestreo	Breve descripción	Algún ejemplo
Aleatorio simple	**1.** Todos los individuos de la población tienen la misma probabilidad de ser escogidos. **2.** Los individuos se seleccionan de manera independiente unos de otros.	**1.** Una muestra de doce estudiantes de una universidad.
Aleatorio simple usando tablas de dígitos aleatorios	Se selecciona al azar al primer individuo de la muestra y los siguientes se seleccionan a partir de la tabla.	**1.** Una muestra de doce estudiantes de una universidad.
Sistemático	Se selecciona al azar al primer individuo de la muestra y los siguientes se seleccionan a intervalos fijos.	**1.** Una muestra de doce estudiantes de la UOC. **2.** Muestra aleatoria de cuatrocientos individuos sobre la guía de teléfonos.
Estratificado	Los individuos de la población se dividen en grupos disyuntos (estratos). La muestra se obtiene seleccionando una muestra aleatoria simple dentro de cada estrato.	Tiempo de conexión a Internet de los alumnos de los institutos públicos de cierto barrio. Estratificación por cursos.
Por conglomerados	Los conglomerados son unidades (normalmente físicas o geográficas) en las que se distribuyen los individuos de la población que hay que investigar. En el muestreo por conglomerados se selecciona una muestra aleatoria de conglomerados y, dentro de cada conglomerado, se selecciona al azar una muestra de sus individuos.	Tiempo de conexión a Internet de los alumnos de los institutos públicos de toda Cataluña. Conglomerados: comarca, instituto y clase.
Polietápico	Se aplica estratificación a los conglomerados.	Tiempo de conexión a Internet desde su casa de los alumnos de los institutos públicos de Barcelona. Estratificamos los conglomerados (barrios) e institutos, si es preciso.
Por cuotas	Se distribuyen los individuos de la población en diferentes categorías y se asigna un número de individuos a cada categoría, de manera que la proporción de individuos de cada categoría en la muestra sea similar a la proporción dentro de la población. El entrevistador va seleccionando a los individuos de la muestra hasta llenar las cuotas.	Tiempo de conexión a Internet desde casa de los jóvenes de Barcelona de entre quince y veintitrés años. Cuotas por género y edad.

Ejercicios

1. Acabad de completar la muestra aleatoria simple a partir de la tabla de dígitos aleatorios de los doce estudiantes matriculados en la universidad.

2. Acabad de completar la muestra sistemática de los doce estudiantes matriculados en la universidad.

3. A partir de los datos de la tabla "edad de año en año", explicad cómo obtendríamos una muestra de quinientos hombres de entre dieciocho y veintiún años (ambos incluidos).

a) Estratificada por edad

b) Por cuotas según la edad

4. Acabad de completar la muestra por cuotas del ejemplo del tiempo de conexión a Internet de los jóvenes de Barcelona de entre quince y veintitrés años.

5. En la edición para Cataluña del día 28-4-2001, el diario *El Periódico* publicaba la siguiente ficha técnica de un sondeo (de ámbito estatal) efectuado con motivo de las elecciones autonómicas en el País Vasco del día 13-5-2001.

> "Tipo de muestreo: estratificado por autonomías y dimensión del municipio. Selección aleatoria de las viviendas. Selección de individuos según cuotas de sexo y edad representativas de la población de cada comunidad."

Comentad los tipos de muestreo que aparecen.

Solucionario

1. Si continuamos explorando la tabla de dígitos aleatorios y vamos marcando los grupos de cuatro dígitos a partir del individuo inicial (alternando subrayado y cajas), obtenemos lo siguiente:

19223 95034 05756 28713 96409 12531 45544 82853

73676 47150 99400 01927 27754 42648 82425 36290

45467 71709 77558 00095 32863 29485 82226 90056

52711 38889 93074 60227 40011 85848 48767 52573

De estas agrupaciones de cuatro dígitos, las siguientes son menores que 1.400 y, por tanto, corresponden a los cuatro primeros alumnos que son escogidos para formar la muestra:

575 1396 19 118

Continuando este procedimiento, obtenemos que la muestra final estará formada por los estudiantes correspondientes a los números siguientes:

575 1396 19 118 895 1181 656 91 881 1238 463 206

2. Si vamos sumando 126 al número correspondiente al primer individuo de la muestra (que era el número 10), obtenemos que la muestra estará formada por los individuos correspondientes a los números siguientes:

10 126 242 358 474 590 706 822 938 1054 1170 1286

3. Los datos que interesan se encuentran en la tabla siguiente, obtenida a partir del original:

Edades	Hombres	Número de individuos en la muestra de 500 correspondientes a la edad
18 años	10.209	114
19 años	10.908	121
20 años	11.614	129
21 años	12.304	137
Total	45.035	501

Redondeando, pasamos de quinientos: podemos eliminar un individuo de la muestra o conservar los quinientos uno. Con estos cálculos tenemos el número de hombres de cada edad que debe haber en la muestra. Ahora:

a) Si queremos estratificar, debemos conseguir una lista con los nombres de todos los hombres de entre dieciocho y veintiún años de Barcelona, en la que se indique su edad, y seleccionar una muestra aleatoria simple de la medida necesaria para cada estrato (es decir, para cada edad); podemos extraer las muestras aleatorias con ayuda del ordenador o de la tabla de dígitos aleatorios.

b) Enviaremos a los entrevistadores a diferentes zonas (intentando obtener la máxima representatividad) con instrucciones precisas de cuántos hombres de cada edad deben entrevistar. La unión de las respuestas obtenidas por los entrevistadores debe contener tantos hombres de cada edad como se ha encontrado en la tabla anterior.

4. Completamos la tabla para obtener el número de individuos de cada par (género-edad) que debe contener la muestra. Estas cuotas se distribuyen entre los entrevistadores.

Edades	Total	Hombres	Mujeres	En la muestra	
				Hombres	Mujeres
15 años	16.362	8.390	7.972	13	13
16 años	17.340	8.837	8.503	14	13
17 años	18.888	9.663	9.225	15	15
18 años	20.338	10.209	10.129	16	16
19 años	21.538	10.908	10.630	17	17
20 años	22.813	11.614	11.199	18	18
21 años	24.098	12.304	11.794	20	19
22 años	23.862	12.087	11.775	19	19
23 años	23.986	12.131	11.855	19	19
Total	189.225	96.143	93.082	151	149

5. Se ha querido que en la muestra final la proporción de individuos de cada comunidad autónoma y de cada medida de municipio fuese igual a la proporción con respecto a la población total. Se han utilizado conglomerados (= viviendas) seleccionadas de manera aleatoria dentro de cada estrato (comunidad y municipio). Finalmente, los entrevistadores han ido a las viviendas seleccionadas con unas cuotas de sexo y edad para garantizar que en cada comunidad la muestra contiene la misma proporción según género y edad que el total de la población de la comunidad.

Anexo

Dígitos aleatorios							
19223	95034	05756	28713	96409	12531	42544	82853
73676	47150	99400	01927	27754	42648	82425	36290
45467	71709	77558	00095	32863	29485	82226	90056
52711	38889	93074	60227	40011	85848	48767	52573
95592	94007	69971	91481	60779	53791	17297	59335
68417	35013	15529	72765	85089	57067	50211	47487
82739	57890	20807	47511	81676	55300	94383	14893
60940	72024	17868	24943	61790	90656	87964	18883
36009	19365	15412	39638	85453	46816	83485	41979
38448	48789	18338	24697	39364	42006	76688	08708
81486	69487	60513	09297	00412	71238	27649	39950
59636	88804	04634	71197	19352	73089	84898	45785
62568	70206	40325	03699	71080	22553	11486	11776
45149	32992	75730	66280	03819	56202	02938	70915
61041	77684	94322	24709	73698	14526	31893	32592
14459	26056	31424	80371	65103	62253	50490	61181
38167	98532	62183	70632	23417	26185	41448	75532
73190	32533	04470	29669	84407	90785	65956	86382
95857	07118	87664	92099	58806	66979	98624	84826
35476	55972	39421	65850	04266	35435	43742	11937
71487	09984	29077	14863	61683	47052	62224	51025
13873	81598	95052	90908	73592	75186	87136	95761
54580	81507	27102	56027	55892	33063	41842	81868
71035	09001	43367	49497	72719	96758	27611	91596
96746	12149	37823	71868	18442	35119	62103	39244
96927	19931	36809	74192	77567	88741	48409	41903
53909	99477	25330	64359	40085	16925	85117	36071
15689	14227	06565	14374	13352	49367	81982	87209
36759	58984	68288	22913	18638	54303	00795	08727
69051	64817	87174	09517	84534	06489	87201	97245
05007	16632	81194	14873	04197	85576	45195	96565
68732	55259	84292	08796	43165	93739	31685	97150
45740	41807	65561	33302	07051	93623	18132	09547
27816	78416	18329	21337	35213	37741	04312	68508
66925	55658	39100	78458	11206	19876	87151	31260
08421	44753	77377	28744	75592	08563	79140	92454
53645	66812	61421	47836	12609	15373	98481	14592
66831	68908	40772	21558	47781	33586	79177	06928
55588	99404	70708	41098	43563	56934	48394	51719
12975	13258	13048	45144	72321	81940	00360	02428
96767	35964	23822	96012	94591	65194	50842	53372
72829	50232	97892	63408	77919	44575	24870	04178
88565	42628	17797	49376	61762	16953	88604	12724
62964	88145	83083	69453	46109	59505	69680	00900
19687	12633	57857	95806	09931	02150	43163	58636
37609	59057	66967	83401	60705	02384	90597	93600
54973	86278	88737	74351	47500	84552	19909	67181
00694	05977	19664	65441	20903	62371	22725	53340
71546	05233	53946	68743	72460	27601	45403	88692
07511	88915	41267	16853	84569	79367	32337	03316