# Estadística para las ciencias sociales, del comportamiento y de la salud

3a. edición

Haroldo Elorza Pérez-Tejada

CENGAGE Learning™

Australia • Brasil • Corea • España • Estados Unidos • Japón • México • Reino Unido • Singapur

# Capítulo 5

# Muestreo

**Mtra. Yvón Angulo Reyes**
Instituto de Investigaciones Sociales, UNAM

## Propósitos

El objetivo central del presente capítulo es presentar los aspectos básicos de la teoría del muestreo de manera accesible a los lectores que por primera vez se enfrentan a ella.

De igual forma, al término del mismo el lector podrá:

- Comprender los conceptos básicos del muestreo.
- Realizar una lectura crítica de los diseños metodológicos que involucren la realización de encuestas.
- Diferenciar los diferentes tipos de muestreo.
- Describir las principales características del muestreo probabilístico y del muestreo no probabilístico.
- Enunciar para el caso del muestreo no probabilístico en qué consiste el de juicio, cuotas, y bola de nieve.
- Utilizar el muestreo adecuado para llevar a cabo una investigación, experimento o estudio determinado.
- Discriminar las ventajas y desventajas de los diferentes tipos de muestreo.
- Explicar las bases del muestreo probabilístico.
- Describir el muestreo estratificado y el muestreo por conglomerados.
- Identificar cuándo se debe recurrir al diseño de muestras complejas.
- Implementar muestras complejas.
- Enunciar el teorema central del límite.
- Aplicar el teorema central del límite.

## INTRODUCCIÓN

De manera cotidiana se suscita la necesidad de tomar decisiones; como se trata de que sean lo más acertadas posible, se percata de que la información con que se cuenta juega un papel determinante en la toma de decisiones y, por tanto, en su resultado. En algunos casos, esa información será suficiente, pero en otros no, por lo que tendrá que hacerse acopio de información adicional.

Esta situación no es muy distinta cuando se trata de tomar decisiones en otros ámbitos, por ejemplo, el diseño de alguna política, evaluación de la efectividad de algún medicamento, evaluación de una campaña publicitaria o política, efectividad en un tratamiento, calidad en los procesos, estudios de opinión, etc. Al igual que en las decisiones cotidianas, se necesita información, la cual puede ser generada de distintas maneras.

Existen diversas formas de captar la información, la cual depende de los propósitos del estudio. Asimismo, la forma de recabar la información depende de distintos factores, entre los que se encuentran los propósitos con los que se va a utilizar el tipo de información requerida, así como la población a la cual se desea conocer, y los recursos de tiempo y dinero con que se cuente. No está por demás decir que el principal elemento que determina la forma de generar la información es el objetivo para el que se utilizará. Un aspecto más importante de resaltar, es que para los casos referidos, la información será sobre una población específica, la cual conforma nuestro universo o población de estudio.

A manera de introducción se menciona que cuando se tiene la posibilidad de extraer la información de todos y cada uno de los elementos de la población de estudio, se habla de la realización de un censo. En caso contrario, es decir, cuando se toma información de un subconjunto de elementos, más pequeño que la población, entonces se tiene una muestra.

Cuando lo que se desea es contar únicamente con un subconjunto (muestra) de la población a partir de la cual se quiere conocer algo acerca de la población, este subconjunto puede ser seleccionado mediante distintos métodos. Cuando sus elementos son seleccionados de acuerdo con una probabilidad conocida y distinta de cero, se está hablando de un muestreo probabilístico, en caso contrario, será un muestreo no probabilístico.

En este capítulo se abordan estos dos tipos de muestreo, se describen las diferencias entre uno y otro, así como los alcances en cuanto a la información que se genera.



## ALGUNOS CONCEPTOS BÁSICOS

### Censo y muestra

Cuando se está ante un problema de investigación que plantee el conocimiento de determinado fenómeno y se requiera generar información para tal fin, en principio existen dos posibilidades. La primera es tomar información de todos y cada uno de los elementos de la población de interés; es decir, realizar un censo; y la segunda, tomar información de una parte, generalmente pequeña, pero representativa de la población de estudio.

Pero, ¿cómo decidir entre realizar un censo o tomar una muestra? Como se mencionó antes, a diferencia de una muestra, en el censo o enumeración completa se recurre a la "medición" de *toda* la población, con lo cual se esperaría obtener información precisa acerca de lo que se desea conocer. Sin embargo, no siempre es conveniente la realización de un censo, contrario a lo que se pudiera pensar de primera impresión. Es decir, no siempre el censo proporciona mejores resultados que la muestra, debido a que,

en la práctica, existen algunas situaciones que ocasionan que el censo pueda tener fallas importantes, entonces es más confiable la utilización de una muestra. El tipo de errores comunes a un censo se asocian principalmente con problemas de cobertura, porque suelen ser selectivos, o con problemas de control de calidad en el levantamiento de la información. Situaciones que pudieran ser mejor controladas si se tuviera un menor número de casos.

Otros aspectos que pueden ayudar a decidir entre la elaboración de un censo o una muestra, son los costos en uno y otro caso, así como la rapidez y oportunidad con que se genere la información. En cuanto a los costos y la rapidez, es obvio, al tener que llevar a cabo en la muestra un número menor de mediciones, el costo se reduce y se obtiene con mayor rapidez la información. Por consiguiente, se contará con información más oportuna. Este punto es fundamental cuando lo que se requiere es contar con información para la toma de decisiones, o en fenómenos en los cuales su dinámica implique cambios rápidos que deban ser identificados de manera oportuna.

Para ilustrar lo anterior considere el siguiente ejemplo: suponga que desea conocer el tipo de música preferida por los estudiantes de una universidad. Si se decidiera por un censo, tendría que preguntar a todos y cada uno de los estudiantes de la universidad cuál es su música predilecta. Bajo estas circunstancias, el resultado que obtendría sería un listado "exacto" y "exhaustivo" de la música que les gusta a todos los estudiantes.

El hecho de realizar el censo en la universidad implica que debe interrogar a todos y cada uno de los estudiantes inscritos en el momento del censo. Lo que contempla la búsqueda y entrevista de quienes no asistieron o estén ausentes por cualquier otro motivo (intercambios, permisos, etc.). Puesto que se trata de un censo, se tendría que buscar y entrevistar a todos independientemente del lugar o situación en la que se encuentren,† lo cual tendría implicaciones en cuanto a costos y tiempo.

Por otro lado, si se decide por preguntarle sólo a una muestra representativa, el número de casos que debe buscar fuera de la universidad, porque no asistieron, disminuye considerablemente.



## Población objetivo

Una tarea importante en el proceso de investigación es la definición clara y concreta de la población objetivo. Se entiende como población objetivo la conformada por los elementos que cumplan con determinadas características en un tiempo y espacio. Por tanto, un primer paso para llegar a la población objetivo es definir de manera clara y exhaustiva las características de los elementos que harán identificar a aquellos que pertenecen a la población objetivo y a los que no.

Si se retoma el ejemplo de los universitarios. La población objetivo estará conformada por todos y cada uno de los estudiantes inscritos durante el presente ciclo escolar. De igual manera, si únicamente se desean conocer los gustos musicales de los de nuevo ingreso, entonces la población objetivo estará constituida por todos aquellos estudiantes inscritos al primer año durante el actual ciclo escolar.

En algunas ocasiones, por diversos motivos, no es posible acceder de manera completa a la población objetivo, por tanto, se tendrá que hacer referencia a la llamada población investigada.

† Este ejercicio se plantea con el supuesto de que la entrevista tendría que ser cara a cara. ¿De qué otra manera pudiera ser? En una población que de cierta manera puede considerarse como "cautiva", donde es posible tener cierto control de todos sus integrantes, se cuenta con otras alternativas, por ejemplo, las encuestas vía correo electrónico, lo cual disminuye costos y en algunos casos tiempo.

## Marco de muestreo

En la medida en que se tenga bien definida la población objetivo, se podrá contar con un listado que contenga todos los elementos que la integran. A este listado se le conoce como marco de muestreo.

Para el ejercicio que se ha venido siguiendo, el marco de muestreo estará conformado por los nombres de todos y cada uno de los estudiantes de la universidad o los inscritos al primer año, durante el presente ciclo escolar, según sea el caso.

Por tanto, un marco de muestreo será el medio físico a partir del cual pueden identificarse de manera directa o indirecta todos los elementos de la población (Méndez, Eslava; 2004).

En términos ideales se esperaría que el marco de muestreo coincidiera con la población objetivo; sin embargo, esto no siempre ocurre. Es decir, puede suceder que el marco contenga elementos que no sean considerados como parte de la población, o que no contenga todos los que interesan.

La primera situación no representa problema alguno, ya que a partir del marco que contenga toda la información, puede extraerse un listado de todos los que en realidad conforman la población objetivo. Si retoma el caso de los universitarios, podría suceder que como marco se cuente con el listado de todos los inscritos durante el ciclo escolar, y los que interesan son únicamente los de primer ingreso. Entonces, lo que tendría que hacer es extraer del listado de todos los inscritos uno que incluya únicamente a los de recién ingreso.

A diferencia del escenario anterior, se tiene un problema más complicado cuando el marco no contiene a todos los elementos de la población. Ante esta situación existen dos posibilidades, la primera es intentar complementar la información del marco con otras fuentes de información, con otros marcos. En caso de que esto no sea posible, entonces se tendrá que reconocer este problema como una limitación importante y redefinir la población objetivo.

De nuevo considere el ejemplo de los universitarios. Suponga que en la universidad se cuenta con sistemas escolarizado y abierto, y que el registro de las inscripciones se realiza por separado. Si lo que desea es conocer el gusto musical de todos los inscritos durante el ciclo escolar, entonces, si se toma únicamente el marco de los inscritos en el sistema escolarizado, la información sólo estará haciendo referencia a este grupo, por lo que para considerar a todos, también tendrá que tomarse en cuenta el marco de los inscritos en el sistema abierto. De esta manera se contará con dos marcos complementarios que cubren a toda la población, sin que existan traslapes.

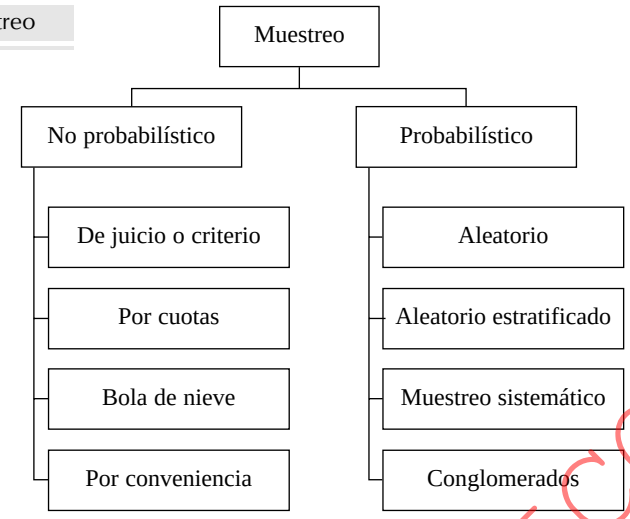


## TIPOS DE MUESTREO

Existen diferentes formas para tomar una muestra, la elección entre una u otra depende de los propósitos para los que se utilizará la información, el conocimiento previo del fenómeno de estudio, así como los recursos con los que se cuente.

De manera general se distingue entre dos tipos de muestreo, los no probabilísticos y los probabilísticos. Cuando se extrae una muestra en donde todos y cada uno de los elementos del marco muestral cuentan con probabilidad de selección distinta de cero, entonces se trata de un muestreo probabilístico, en caso contrario, se tratará de muestreo no probabilístico. Entre los no probabilísticos se cuentan el de juicio, por cuotas, bola de nieve y por conveniencia; en tanto que los probabilísticos son el aleatorio simple, aleatorio estratificado, sistemático, por conglomerados, y las posibles combinaciones que puedan realizarse entre éstos.

**Figura 5.1** Tipos de muestreo




¿Cuáles son las ventajas de un muestreo probabilístico y uno no probabilístico? y, por tanto, ¿cuándo es conveniente utilizar uno y cuándo el otro? La respuesta depende de una serie de factores que serán abordados en los siguientes apartados.

El muestreo probabilístico consiste en tomar una muestra de manera aleatoria a partir de un marco muestral, en donde todos y cada uno de los elementos del marco tienen una probabilidad conocida y distinta de cero de salir en muestra. La selección de la muestra se realiza con base en fundamentos de la teoría de probabilidad, lo cual permite hacer una evaluación objetiva de los resultados y, por ende, se está en posibilidad de conocer el grado de precisión y confianza de los mismos. Por tanto, en el muestreo probabilístico, una vez definida la población de estudio, configurado el marco de muestreo y definida la forma de selección, la conformación de la muestra no depende de los criterios selectivos o preferencias del investigador.

A diferencia del muestreo probabilístico, en el no probabilístico la conformación de la muestra depende en gran medida de los "criterios" del investigador, entonces, no hay una manera objetiva de asignar probabilidades de selección a los elementos del marco de muestreo, en caso de que se conozca. Este último punto es importante, porque en algunas ocasiones, aunque quisiera utilizarse un muestreo probabilístico, no se lograría por la imposibilidad de contar con un marco muestral, así, se recurre a un muestreo no probabilístico.

En resumen, en general puede decirse que se recomienda el uso del muestreo probabilístico cuando se desea conocer de manera objetiva la precisión y confianza de los resultados obtenidos, para lo que deberá contar con marco de muestreo confiable. En tanto que si lo que se desea es obtener información de cierta población para la cual es difícil contar con un listado confiable, o únicamente se desea conocer información de manera exploratoria, el método de muestreo que se utilizará será uno no probabilístico. En este caso, le toca al investigador argumentar la representatividad de la muestra, la cual dependerá en gran medida del objetivo de la investigación.

## MUESTREO NO PROBABILÍSTICO

En el muestreo no probabilístico, la representatividad de la muestra depende de criterios no probabilísticos, es decir, la inclusión o no de un elemento en la muestra se determina en gran medida por el cri-

terio de los investigadores y, en algunos casos, de los propios entrevistadores y no de un proceso aleatorio, por lo que no se está en posibilidades de conocer las probabilidades de inclusión en muestra de cada uno de los elementos de la población objetivo.

Por otro lado, a partir de la definición general, anteriormente mencionada, en la que se considera que un muestreo probabilístico requiere de un marco muestral para conocer la probabilidad de selección de cada uno de los elementos que conformarán la muestra, para el caso del muestreo no probabilístico, este no es un requisito.

## De juicio

La característica principal de este tipo de muestreo es que tanto el tamaño de muestra como la elección de los elementos están sujetos al juicio del investigador, esto es, para realizar un estudio mediante este tipo de muestreo debe recurrirse a la experiencia que se tenga. Es decir, la muestra se forma con los elementos que el investigador considera (según su juicio) que son los más representativos de la población que va a estudiar.

El juicio del investigador se rige por el conocimiento y experiencia que tenga sobre el tema. Por consiguiente, el éxito y la eficacia de la muestra dependen de la opinión del investigador que haya seleccionado los elementos. En virtud de que el conocimiento y experiencia dependen de la persona que realiza el muestreo, entonces, la muestra estará sujeta a estas características. Esta subjetividad en la selección de los elementos de la muestra es lo que hace que se clasifique como un muestreo no probabilístico.

Existen muchas situaciones en las que el muestreo de juicio es útil y aun aconsejable. Un caso puede ser cuando es muy complicado contar con un marco de muestreo confiable que permita obtener una muestra probabilística. Por ejemplo, suponga que desea hacer un estudio acerca de la opinión de los líderes de organizaciones sociales informales acerca de la situación política del país. Es difícil que exista un listado completo de todos estos líderes, y armarlo resultaría muy costoso, en cuanto a tiempo y recursos económicos. Por consiguiente, una opción en este caso es que el investigador realice un listado de los líderes "más representativos" que él considere que deben ser encuestados.

## Por cuotas

Para el caso del muestreo por cuotas, al igual que las muestras generadas por muestreo de juicio, tampoco son probabilísticas. Sin embargo, el muestreo por cuotas permite obtener muestras representativas en cuanto a la distribución de algunas variables relevantes de la población.

Esta representatividad se basa principalmente en la obtención de un número específico de casos, de acuerdo con las variables que se consideren más relevantes. Sin embargo, al igual que en el muestreo de juicio, la selección de los elementos de la muestra no es probabilística, como se verá a continuación.

El procedimiento general del muestreo por cuotas es el siguiente:

1. Variables relevantes. Un primer paso es identificar las variables que van a permitir asignar cuotas. Estas variables se definen de acuerdo con la importancia o la posible influencia que pudieran tener sobre el fenómeno que está estudiándose (por ejemplo, sexo, escolaridad, edad, ingreso, etcétera).
2. Recabar información sobre la distribución de las variables relevantes para la población objetivo. Con esta información se asigna el número de casos de acuerdo con la distribución porcentual de dichas variables.

3. Asignar a cada entrevistador el número de cuestionarios a aplicar, así como la distribución de los mismos de acuerdo con las variables relevantes.

En este último, donde se realiza el proceso de selección[†] de los entrevistados, es donde se introduce el sesgo de los entrevistadores, ya que ellos deciden a quién entrevistar y no hay manera de saber la probabilidad de selección de cada uno de los incluidos en muestra.

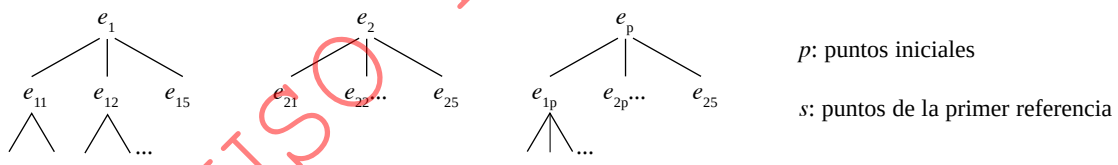
## Bola de nieve

Al igual que en los casos anteriores, cuando no se cuenta con un marco de muestreo confiable o cuando es muy difícil contactar a la población objetivo, es recomendable un muestreo por bola de nieve.

El muestreo de bola de nieve involucra los siguientes pasos:

1. Definición de la población objetivo.
2. Elección de uno o varios elementos para el inicio del levantamiento.[††] En este punto es donde se define, directa (consciente) o indirectamente (inconscientemente), la composición final de la muestra.
3. A cada uno de los elementos encuestados se le solicitará referencia de otro u otros elementos que cumplan con las características de la población objetivo. Estas referencias serán los siguientes integrantes de la muestra. Y así sucesivamente, hasta un determinado número de referencias o hasta que de acuerdo con los criterios del investigador, se cuente con la información suficiente. En consecuencia con este procedimiento, la composición de la muestra se conocerá hasta el final.

**Figura 5.2** Esquema general del muestreo por bola de nieve




Como se puede apreciar a partir de la descripción anterior, si bien la selección de los primeros contactos puede ser hecha a partir de métodos probabilísticos, en caso de que se cuente con un marco de muestreo, la inclusión de los siguientes elementos ya no será probabilística, pues depende de que hayan sido mencionados por los primeros. Es decir, este tipo de muestreo se considera no probabilístico debido a que al ir constituyendo la muestra a partir de referencias de encuestados, no se tiene manera de conocer la probabilidad de selección de todos los elementos de la población; más aún, puede ocurrir que no todos los elementos de la población tengan posibilidad de quedar incluidos en la muestra, debido a que quizá no tengan relación alguna con los puntos de inicio o con los mencionados en referencias posteriores (es decir, que se trate de puntos aislados).

[†] Más que una selección es una elección de personas a encuestar.

[††] En algunos estudios estos puntos de inicio se seleccionan en forma aleatoria, con la finalidad de evitar la introducción de sesgos por parte de los investigadores y darle cierta objetividad a la muestra.

### ■ Ejemplo 1

Suponga que busca realizar un estudio de la opinión que tienen los empresarios acerca de la aprobación de determinada ley arancelaria. A primera vista puede decirse que, aunque es complicado, se podría contar con un marco de muestreo confiable, lo cual sería de gran ayuda. Sin embargo, por otro lado, se percata de que es complicado hacer contacto con los empresarios. Por tanto, un muestreo por bola de nieve es recomendable, ya que únicamente se tendrá que garantizar el contacto con un número reducido de empresarios, los cuales remitirán con empresarios conocidos por éstos, lo que facilitará su inclusión en la muestra para conformar la muestra final.

## Por conveniencia

Cuando una muestra está conformada únicamente por elementos disponibles o con los más dispuestos, entonces se trata de una muestra por conveniencia. A diferencia del muestreo de juicio, en el que se buscan elementos que a juicio del investigador sean los convenientes, una muestra por conveniencia estará constituida por los que se "tengan a la mano".

Obvio, este tipo de muestreo tiene muchas ventajas prácticas; sin embargo, sus resultados sólo harán referencia a los que fueron entrevistados y no a un grupo más grande.

A pesar de esto, el muestreo por conveniencia tiene aplicaciones importantes, por ejemplo en estudios de mercado en los cuales se desee realizar la evaluación de un producto, o cuando se quiere llevar a prueba el diseño de algún cuestionario (prueba piloto en una encuesta), etcétera.



# MUESTREO PROBABILÍSTICO

Parta de que el objetivo básico del muestreo probabilístico es entender el comportamiento de determinado fenómeno, así como el grado de precisión con que se conoce, es decir, se desea estimar lo mejor posible el valor de una determinada variable y conocer la magnitud del posible error que esté cometiéndose.

Por tanto, cuando sea necesario contar con el grado de representatividad de una muestra, así como con los errores de muestreo, es recomendable el uso de un muestreo probabilístico.

Pero, ¿qué implica un muestreo probabilístico? Básicamente el conocimiento de las probabilidades de selección (de estar en la muestra) de cada uno de los elementos de la población objetivo y, por ende, que sea posible conocer las probabilidades de selección de todas y cada una de las posibles muestras que se puedan tomar de la población.† Este hecho es el que da sustento al muestreo probabilístico, ya que a partir de dicha distribución muestral será posible aplicar las leyes de probabilidad en dicho espacio y, por tanto, calcular los estimadores deseados con sus respectivos niveles de confianza y errores de muestreo.

Si bien la aplicación de un muestreo probabilístico requiere el cumplimiento de ciertos requisitos, existen ventajas considerables para su uso en determinadas circunstancias; por ejemplo, el hecho de conocer la precisión con que están obteniéndose los resultados disminuye la posibilidad de tomar decisiones equivocadas.

† Al conjunto de todas estas posibles muestras es lo que se conoce como espacio muestral y su distribución como distribución muestral.

# FUNDAMENTOS DE MUESTREO PROBABILÍSTICO

## Conceptos básicos

Como se mencionó en secciones anteriores, el objetivo del muestreo es estimar los parámetros desconocidos de una población a partir de una muestra. Considere una población en la cual los valores de la variable de estudio están representados por $X_i$, un parámetro se define como una función que resume los valores de esta variable† en el total de la población.

Por ejemplo, si se considera como variable de estudio ($X$), la edad de las personas que asisten a un determinado evento. Entonces un parámetro será la edad promedio de todos los asistentes.

Cuando lo que se tiene es un valor calculado (estimado) a partir de los valores de una muestra, entonces de lo que se trata es de un estimador†† del parámetro poblacional. Para el ejemplo anterior, si la edad de los asistentes se calcula a partir de una muestra, entonces se tendrá una estimación de la edad promedio de los asistentes.

Dos situaciones importantes se desprenden de lo anterior. La primera, que lo deseable es que el estimador fuera lo más parecido al parámetro, y la segunda es que así como se tomó una muestra para estimar el parámetro, bien pudo haberse sacado otra muestra en otros integrantes que proporcionaran otra información y, por tanto, un valor del estimador diferente.

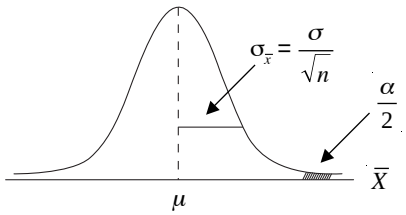
En el primer punto, a lo que está haciéndose referencia es a la precisión con que se desea estimar un parámetro, es decir, qué tan alejado está el estimador del parámetro. En tanto que el segundo punto se refiere a la posibilidad de tener más de una muestra y, entonces, más de una estimación del parámetro. Como cada muestra estará integrada por elementos distintos, es de suponer que algunas estimaciones sean más precisas que otras, es decir, que estén más cercanas al parámetro.

A la distribución de los valores de los estimadores de todas las muestras posibles de tamaño $n$ se le conoce como distribución muestral. Es decir, a partir de cada $n$ (tamaño de muestra) dado, se podrán obtener tantas estimaciones del parámetro como combinaciones posibles de muestras distintas de tamaño $n$ sea posible extraer, las cuales generarán una distribución muestral para cada $n$.

Lo que menciona el Teorema central del límite††† es que a medida que el tamaño de las muestras se va incrementando (es decir, que $n$ vaya siendo más grande), la distribución de la media (distribución muestral) tiende a tener distribución normal, sin que la distribución de la variable que se mide tenga necesariamente distribución normal en la población.

Por tanto, si se tiene un tamaño de muestra $n$ fijo, la distribución muestral del promedio tendrá la siguiente representación.

**Figura 5.3** Distribución muestral del promedio




† Ejemplos de medidas son la media, el total, la proporción, varianza, etcétera.
†† Un estimador puede ser sesgado o insesgado, consistente o no.
††† Teorema fundamental en estadística.

En donde $\mu$ es el valor del parámetro en la población y el resto de los valores son las estimaciones calculadas a partir de cada una de las muestras posibles ($\bar{X}$), las cuales se encuentran alrededor de $\mu$ y se van concentrando más a medida que $n$ aumenta, ya que $\bar{X}$ se distribuye como una normal con parámetros ($\mu$, $\sigma^2_{\bar{X}}$), donde $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$ y $\sigma_{\bar{X}}$ es la desviación estándar de la distribución muestral, conocido como error estándar de la distribución de $\bar{X}$.

Si se retoma la idea de utilizar el muestreo probabilístico para obtener una estimación del parámetro lo más cercano posible, esto con probabilidad de ocurrencia alta, lo anterior se puede poner en los siguientes términos.

$$P(\mu - d \leq \bar{X} \leq \mu + d) = 1 - \alpha \qquad \text{(I)}$$

es decir, lo que se desea es que la estimación de la media se encuentre alejada al parámetro en, a lo más, $d$, con una probabilidad de $(1 - \alpha)$. Al valor $d$ se le conoce como precisión o error de estimación, en tanto que $(1 - \alpha)$ es la confianza con la que esperamos que ocurra esta precisión.

Por tanto, dado que $\bar{X}$ se distribuye como una normal $\left(\mu, \frac{\sigma^2}{n}\right)$, si lo que se quiere es que el estimador cumpla (I) con una confianza del 95%, es decir,

$$P(\mu - d \leq \bar{X} \leq \mu + d) = 0.05 \qquad \text{(II)}$$

como

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma^2}{n}}, \; Z \sim N(0,1)$$

entonces (II) es equivalente a

$$P(-Z_{\frac{\alpha}{2}} \leq Z \leq Z_{\frac{\alpha}{2}}) = 0.05$$

$$P\left(-Z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{\sigma^2}{n}} \leq Z_{\frac{\alpha}{2}}\right) = 0.05$$

$$P\left(-Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 0.05$$

$$P\left(\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 0.05$$

En consecuencia

$$d = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \qquad \sqrt{n} = Z_{\frac{\alpha}{2}} \frac{\sigma}{d}$$

de donde,

$$n = Z^2_{\frac{\alpha}{2}} \frac{\sigma^2}{d^2},$$

y para el caso específico de 95% de confianza,

$$n = 1.96^2 \frac{\sigma^2}{d^2}$$

Por ende, éste será el tamaño de muestra necesario para tener estimaciones con una precisión $d$, y confianza $(1 - \alpha)$ con varianza conocida.† El caso que acaba de esbozarse hace referencia a un tamaño de muestra cuando se desea estimar la media poblacional. Por otra parte, se ve que el tamaño de la muestra depende directamente de la variación del fenómeno estudiado.

Sin embargo, este resultado puede ser extendido para el caso en que se desea estimar una proporción poblacional, como caso particular de la media.

Por consiguiente, el cálculo del tamaño de muestra para estimar una proporción con determinada precisión y confianza, se realizará como:

$$n = Z^2_{\frac{\alpha}{2}} \frac{pq}{d^2}, \text{ dado que } \sigma^2 = pq$$

donde para el cálculo se utilizarán igualmente estimaciones de $p$ y $q$.

### ■ Ejemplo 2

Calcular el tamaño de muestra, necesario para estimar la proporción de personas que participan en alguna organización formal, esto con un nivel de confianza de 95% y errores de estimación no mayores a 3 puntos porcentuales. Además, se sabe que en una encuesta anterior se encontró que sólo 25% de la población pertenecía a alguna organización. Entonces:

$$n = \frac{1.96^2(0.25)(0.75)}{0.03^2}$$

$$n = 801 \text{ casos}$$

En caso de que no se tuviese una estimación anterior de la proporción, se podría utilizar como valor de $p$, 0.5, con lo que se obtendría un tamaño de muestra de 1 068 casos, y errores de estimación conservadores.

El cálculo del tamaño de la muestra anterior se realizó con el supuesto de que las muestras serían extraídas en forma probabilística. A continuación, se abordarán distintas maneras en que pueden ser extraídas estas muestras probabilísticas, dependiendo de la información con la que se cuente y el propósito que se tenga, así como para lo que se deseen obtener las estimaciones.

† Cuando no se conoce la varianza será posible reunir estimaciones de estudios anteriores u obtener una aproximación a partir de una prueba piloto, $\frac{n}{N} < 0.1$ poblaciones infinitas.

El diseño de la muestra consiste en la descripción de la forma en que se tomarán los elementos de la población para integrar la muestra, su tamaño, así como la manera en que se calcularán los estimadores. En los siguientes apartados se abordan los diseños básicos de muestreo, se describen la forma de selección de la muestra y la forma de sus estimadores.

## Muestreo aleatorio simple

El muestreo aleatorio simple o muestreo irrestricto aleatorio forma la base de la mayor parte de los diseños de muestreo, así como de las encuestas científicas que se llevan a la práctica. El muestreo aleatorio simple es el más sencillo de los métodos probabilísticos, que permite obtener estimaciones de alguna característica de la población, así como una medida de la confianza y error de las estimaciones hechas. Según Scheaffer: "Si un tamaño de muestra *n* es seleccionado de una población de tamaño *N*, de tal manera que cada muestra posible de tamaño *n* tiene la misma probabilidad de ser seleccionada, el procedimiento de muestreo se denomina irrestricto aleatorio. A la muestra así obtenida se le llama muestra irrestricta aleatoria".†

### Selección de una muestra irrestricta aleatoria (*mia*)

Un método al que se acude para la selección de una muestra irrestricta aleatoria (*mia*) es a través del uso de tablas de números aleatorios. Una de estas tablas consiste en arreglos de números formados a partir de los enteros de 0 a 9, en proporciones aproximadamente iguales, sin tendencias en el patrón en que se generaron los dígitos. Por consiguiente, para contar con una *mia* bastará con tomar de manera aleatoria un conjunto de puntos de la tabla, con la seguridad de que los puntos seleccionados tuvieron la misma probabilidad de salir, que los no seleccionados, puesto que todos fueron generados de manera independiente y con la misma probabilidad de selección. Actualmente, debido a consideraciones prácticas, en una computadora se puede simular este procedimiento para generar las tablas de números aleatorios.

#### *Uso de la tabla de números aleatorios*

Para ejemplificar el uso de la tabla de números aleatorios, suponga que en una escuela desea modificar el plan de estudios si en dicha escuela hay un grupo de 40 profesores y desea obtener una muestra de 4 para conocer su opinión acerca de dicha modificación.

**Paso 1.** Se enumera a los profesores del 1 al 40; en la tabla de números aleatorios, los dígitos se escogerán de dos en dos, porque el tamaño de la población es $N = 40$ (número de dos dígitos).

**Paso 2.** Para dar inicio a la selección de los cuatro profesores, en la tabla de números aleatorios se toman en forma arbitraria una columna y un renglón; suponga que se seleccionan el reglón 60 y la columna 4 y se leen los pares de dígitos en las columnas 4 y 5, lo que da 13, 02, 18, 74, 59, 13, 74, 33; se omiten tanto los números mayores de 40 como los números repetidos.

Cualquier punto de inicio puede ser usado y uno puede moverse en cualquier dirección predeterminada, siguiendo siempre el mismo patrón de movimiento. Si va a utilizarse más de una muestra en cualquier problema, cada una debe tener su propio punto de inicio.

† Sheaffer Richard, L., *Elementos de muestreo*, Iberoamérica, México, 1991.

**Paso 3.** Se continúan leyendo los pares de dígitos hasta que se obtienen cuatro unidades diferentes, es decir, 13, 02, 18, 33. Cada uno de estos números tiene relacionado el nombre de algún profesor, el cual formará parte de la muestra de cuatro profesores seleccionados.

El método anterior deja de ser práctico cuando el número de personas que se quiere seleccionar es muy grande, por lo que es de utilidad el empleo de paquetes estadísticos u hojas de cálculo para la selección de una muestra irrestricta aleatoria.

## Tamaño de muestra para una *mia*

La fórmula para el tamaño de muestra de un muestreo irrestricto aleatorio es la siguiente:

$$n = \frac{Z_{\frac{\alpha}{2}}^2 \sigma^2}{d^2}$$

De acuerdo con la expresión anterior, pareciera que el tamaño de muestra es independiente del tamaño de la población, lo cual llevaría a algunas contradicciones. Por ejemplo, cuando se tienen poblaciones relativamente pequeñas, resulta que el tamaño de muestra es mayor que la población.

### ■ Ejemplo 3

A partir del ejemplo en el que se estimó la proporción de personas que participan en alguna organización social, con una confianza del 95% u error de estimación no mayor a 3 puntos porcentuales, calcular el tamaño de muestra necesario para estimar la proporción en una población de tamaño $N = 300$.

Como se vio en el ejemplo mencionado, el tamaño de muestra necesario es

$$n = \frac{1.96^2(0.25)(0.75)}{0.03^2}$$

$$n = 801 \text{ casos}$$

El tamaño de muestra es casi más del doble de la población total. Este hecho es consecuencia de un supuesto (no mencionado) que se realizó al momento de la deducción de la fórmula y es que están suponiéndose poblaciones infinitas,† por lo que el tamaño de la población no afecta. Pero cuando este supuesto no se cumple tiene que realizarse una corrección "por finitud", la cual implica, *a grosso modo*, la relativización del tamaño de muestra al tamaño de la población. La corrección que debe realizarse es la siguiente:

$$n = \frac{n'}{1 + \frac{n'}{N}}$$

† Este supuesto se hace en el momento del cálculo de la varianza de la muestra.

Por tanto, el tamaño de muestra para el ejemplo anterior será de

$$n = \frac{881}{1 + \frac{881}{300}}$$

$$n = 224 \text{ casos}$$

Es decir, poco más de dos terceras partes de la población total estarán incluidas en la muestra.

## Estimadores para una *mia*

Como se ha mencionado a lo largo del capítulo, el objetivo de una encuesta por muestreo es tratar de determinar (estimar), a partir de un subconjunto de la población total, algunos valores poblacionales (parámetros) que permitan conocer a la población en la medida de lo posible.

Entre los parámetros básicos que ayudan a esta caracterización se encuentran la media o proporción, el total poblacional, la varianza de estos parámetros y los intervalos de confianza.

## Estimador de la media poblacional

Parta de un ejemplo. Suponga que recién se realizó una encuesta sobre la percepción que tienen las personas de 18 años en adelante acerca de si la constitución es adecuada para el país o no,† y que se seleccionó una muestra aleatoria de tamaño 600 de una población total de 10 000. A partir de esta información, ¿cuál es el valor estimado de la proporción de población que está a favor de que la constitución es adecuada y cuáles son sus errores de estimación?

$N = 10\,000$

$n = 600$ casos seleccionados de manera aleatoria

A partir de la muestra se encontró que 400 de las 600 personas mencionaron estar de acuerdo con que la constitución es adecuada para las necesidades del país.

Por tanto, una estimación del parámetro poblacional de las personas que están a favor, viene dado por:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = p$$

donde $x_i$ es el valor observado del $i$-ésimo elemento de la muestra. Por consiguiente, para el caso de este ejercicio se tiene que en 400 de las 600 observaciones $x = 1$, y en 200 $x = 0$. Así,

$$p = \frac{\sum_{i=1}^{n} x_i}{600} = \frac{400}{600} = 0.666$$

† Encuesta del Instituto de Investigaciones Jurídicas. Para este caso en particular fue supuesto un tamaño de población mucho menor al real.

y la varianza estimada de $p$ es:

$$V(p) = \frac{pq}{n}\left(\frac{N-n}{N}\right) = \frac{pq}{n}\left(1-\frac{n}{N}\right)$$

aplicando los valores del ejercicio se tiene

$$V(p) = \frac{(0.66)(0.34)}{600}\left(\frac{10\ 000-600}{10\ 000}\right)$$

$$V(p) = 0.00037037(0.94)$$

$$V(p) = 0.000348148$$

y el intervalo de confianza alrededor de $p$ sería,

$$p \pm 1.96\sqrt{0.000348148}$$

$$p \pm 1.96(0.018658728)$$

$$p \pm 0.03657, \text{ es decir, } 0.667 \pm 0.03657$$

¿Qué se puede decir de los resultados anteriores? Con la muestra de 600 casos obtenida con una muestra irrestricta aleatoria, se estima que aproximadamente el 66.6% de la población está de acuerdo con que la constitución es adecuada. Este dato por sí solo no da información de qué tan confiable es la estimación, para esto se toma el intervalo de confianza

$$(0.630097, 0.703237)$$

es decir, con una confianza del 95% espera que la verdadera proporción (valor poblacional) se encuentre entre los anteriores dos valores, o lo que es lo mismo, que con 95% de confianza la estimación realizada estará alejada del valor poblacional en, a lo sumo, 0.03657 puntos.

Puesto en porcentajes, con 95% de confianza, la estimación que se realizó, de acuerdo con la cual 66.7% de las personas está a favor de que la constitución es adecuada, estará alejada del parámetro (verdadero valor poblacional), en, a lo más, 3.68 puntos porcentuales.

Para el caso de la varianza estimada de la media, se tiene:

$$V(\bar{X}) = \frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)$$

en donde $\sigma^2$ se estima como

$$s^2 = \frac{\sum_{i=1}^{n}(x_1-\bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n}x_i^2 - n\bar{X}^2}{n-1}$$

## Muestreo aleatorio estratificado

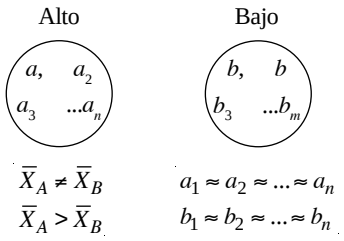
Básicamente, el muestreo estratificado consiste en aprovechar las características de la población para tener estimaciones más precisas. Es decir, la aplicación de un muestreo estratificado supone el conocimiento (o sospecha), por parte del investigador, del comportamiento de algunas características de la población con relación al tema que está investigándose.

**Figura 5.4** Ejemplo de muestreo aleatorio estratificado




Por ejemplo, admita que desea saber el promedio de videojuegos que los niños de entre 6 y 10 años de una escuela conocen. Es claro que la variable de interés, número de videojuegos conocidos, está estrechamente relacionada con el nivel socioeconómico de la familia del niño. Es decir, se esperaría que un niño de familia del estrato socioeconómico alto conociera un mayor número de videojuegos que un niño de estrato bajo y que, a su vez, el número de videojuegos que conocen los niños del estrato alto fuera muy similar. Puesto en términos más formales, lo que se tiene son dos estratos (subconjuntos) muy diferentes entre sí (estrato socioeconómico alto y bajo), pero con integrantes muy similares.

**Figura 5.5** Muestreo aleatorio estratificado



En situaciones como la anterior, es recomendable el empleo del muestreo estratificado, con lo que se esperaría tener mayor precisión en las estimaciones, además de que existe la posibilidad de hacer estimaciones por separado para cada uno de estos grupos (estratos) si el tamaño de muestra en cada estrato permite tener precisión ($d$) aceptable.

En resumen, el muestreo aleatorio estratificado consiste en, primero, identificar una o algunas variables de estratificación. Esta variable debe hacer referencia a determinada característica de la población que esté correlacionada con la variable de estudio y ser capaz de identificar subgrupos heterogéneos entre sí y homogéneos al interior; segundo, en cada uno de estos estratos, seleccionar de manera aleatoria una muestra de tamaño $n_i$, que es el tamaño de muestra de cada uno de los estratos.† Y por último, una vez que se cuente con información de los estratos se procederá a realizar las estimaciones.

## Estimaciones de proporciones

Para el caso del ejemplo anterior, se tienen dos estratos: estratos socioeconómicos alto y bajo. Cada uno de ellos con $N_i$ niños, es decir, el estrato socioeconómico alto con $N_1$ niños y el estrato bajo con $N_2$ niños, de manera que $N_1 + N_2 = N$, donde $N$ es el total de niños de entre 6 y 10 años en la escuela.

Para cada uno de los estratos se asignó $n_i$ niños como tamaño de muestra, $n_i = 15$ niños para el estrato socioeconómico alto y $n_2 = 25$ para el bajo, de manera que $n_i + n_2 = n$, con el tamaño total de la muestra $n$, necesario para obtener estimaciones del *mae* (muestreo aleatorio estratificado).

Por tanto, si supone que

$N$ = 180 niños de entre 6 y 10 años
$N_1 = 70$
$N_2 = 110$

y que se asignó un tamaño de muestra para cada estrato de $n_1 = 15$ y $n_2 = 25$, a partir de los cuales se obtuvieron los siguientes resultados:

| Estrato 1 (alto) | | | | | Estrato 2 (bajo) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 7 | 11 | 9 | 12 | 4 | 7 | 5 | 3 | 6 |
| 9 | 10 | 12 | 11 | 9 | 8 | 5 | 3 | 6 | 4 |
| 13 | 12 | 10 | 9 | 11 | 4 | 6 | 5 | 5 | 4 |
| | | | | | 3 | 6 | 5 | 3 | 4 |
| | | | | | 6 | 7 | 4 | 5 | 7 |

Estimar el número de videojuegos que conocen en promedio los niños de 6 a 10 años en una determinada escuela.

El número promedio estimado de videojuegos está dado por

$$\bar{x}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{x}_i \quad \text{en donde}$$

† Existen diferentes formas de determinar el valor de $n_i$, más adelante son abordadas algunas de ellas.

$\bar{x}_i$ es el promedio de videojuegos que los niños conocen en cada uno de los estratos *i*.

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_i}{n_1} \quad \text{y} \quad \bar{x}_2 = \frac{\sum_{i=1}^{n_2} x_i}{n_2}$$

$$\sum_{i=1}^{n_1} x_i = 155 \qquad \sum_{i=1}^{n_2} x_i = 125$$

$$n_1 = 15 \qquad n_2 = 25$$

de donde

$$\bar{x}_1 = 10.33 \quad \text{y} \quad \bar{x}_2 = 5.0$$

En consecuencia

$$\bar{x}_{st} = \frac{1}{180}\left[(70)(10.33) + (110)(5.0)\right]$$

$$\bar{x}_{st} = 7.07$$

Es decir, los niños de entre 6 y 10 años de cierta escuela, conocen en promedio 7.07 videojuegos. Para completar esta información, se calculará el intervalo de confianza de la siguiente manera:

$$\bar{x}_{st} \pm Z_{\frac{\alpha}{2}} \sqrt{V(\bar{x}_{st})}$$

en donde

$$V(\bar{x}_{st}) = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{s_i^2}{n_i}$$

para el ejercicio anterior

$$s_1^2 = \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i^2)^2}{n_i - 1} = \frac{\sum_{j=1}^{n_i} x_{ij}^2 - n_i \bar{x}_i^2}{n_i - 1}$$

$$s_1^2 = \frac{1637 - 15(10.33)^2}{15 - 1} \quad \text{y} \quad s_2^2 = \frac{673 - 25(5)^2}{25 - 1}$$

$$s_1^2 = 2.5976 \qquad s_2^2 = 2$$

de donde, sustituyendo estos valores en $V(\bar{x}_{st})$ se tiene

$$V(\bar{x}_{st}) = \frac{1}{180^2}\left[70^2\left(1-\frac{15}{70}\right)\frac{2.5976}{15} + 110^2\left(1-\frac{25}{110}\right)\frac{2}{25}\right]$$

$$V(\bar{x}_{st}) = \frac{1}{180^2}(1414.7173)$$

$$V(\bar{x}_{st}) = 0.043664114$$

Por ende, el intervalo de confianza alrededor de $\bar{x}_{st}$ es

$$7.07 \pm 1.96\sqrt{0.04366411}$$

$$7.07 \pm 0.4095608$$

es decir, el verdadero valor del promedio de videojuegos que conocen los niños está entre 6.66 y 7.48, con un 95% de confianza.

Si se hubiera utilizado una *mia*, desaprovechando la información para generar estratos, lo que se hubiera obtenido es que (para esto suponga que la muestra fue obtenida sin estrato)

$$\bar{x} = 7.0$$

$$s^2 = 8.9743$$

$$V(\bar{x}_{st}) = \frac{8.97}{40}\left(1-\frac{40}{180}\right)$$

$$V(\bar{x}_{st}) = 0.1744$$

el intervalo estará dado por (6.18, 7.82). Es decir, al desaprovechar la información para estratificar, se pierde en cuanto a precisión.

## Afijación de la muestra

Por afijación de la muestra se entiende el hecho de distribuir el tamaño de muestra total ($n$) entre los distintos estratos existentes. Puede realizarse por distintos métodos, entre lo cuales se encuentran la afijación uniforme, proporcional, de mínima varianza y óptima. A continuación se hace una breve descripción de cada uno de estos métodos.

La afijación uniforme consiste en asignar el mismo número de casos para cada uno de los estratos, con lo cual se les está dando la misma importancia, independientemente de su tamaño. Cuando la afijación se realiza asignando el número de casos de manera proporcional al tamaño de cada estrato (número de elementos en cada estrato), entonces se tratará de una afijación proporcional.

En caso de que se desee tener estimaciones con varianza mínima, entonces se utilizará el método de Neyman. Finalmente, en la afijación óptima se considera el costo por unidad de muestreo, además de que se pide varianza mínima.

## Muestreo por conglomerados

En los ejercicios propuestos en las secciones anteriores para la *mia* y para el *mae*, se partió del hecho de que era posible contar con un marco muestral para la selección de los elementos de la muestra. En el *mia* para la selección directa de $n$ elementos del total, mientras que en el estratificado para la selección directa de $n_i$ elementos en cada $i$-ésimo estrato.

Cuando para algún problema de investigación no existe un marco de muestreo o su construcción sea demasiado costosa, o cuando las unidades de muestreo se encuentran muy dispersas, al grado de impactar de manera importante en el costo del levantamiento, entonces se puede recurrir al uso del muestreo por conglomerados.

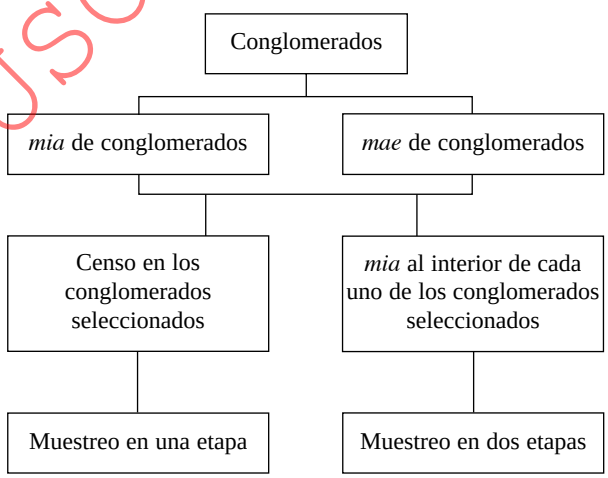
El muestreo por conglomerados consiste en considerar como unidades de muestreo agrupaciones (grupos, conjuntos) de elementos. De esta manera, sólo será necesario un listado (marco de muestreo) de estos conglomerados y no de todos los elementos.

Este tipo de muestreo puede hacerse en una o más etapas. Cuando en un muestreo se realiza la selección de $n$ conglomerados y los $m_i$ elementos de los $n$ conglomerados son medidos, entonces se trata de un muestreo por conglomerados en una etapa, puesto que únicamente hubo una etapa de selección. Este procedimiento se puede hacer cuando los conglomerados no son muy grandes;† en caso contrario, a lo que se recurre es a la selección de una muestra de cada uno de los conglomerados seleccionados, en tal caso, se tratará de un muestreo por conglomerados bietápico.





### Selección por conglomerados




† Pero esto no es recomendable cuando las unidades al interior de los conglomerados son muy parecidas entre sí, debido a que se pueden tener problemas de autocorrelación intraclase.

Por tanto, la selección de una muestra por conglomerados consiste en lo siguiente:

Primero, identificar agrupaciones (conglomerados) de elementos en la población, que puedan ser considerados como unidades de muestreo. Puesto que sólo se seleccionará un subconjunto de estos conglomerados, lo deseable es que sean lo más homogéneos entre sí, y lo más heterogéneos al interior, esto para garantizar que los que no salgan en muestra contengan la misma información que los que sí salen.

Se debe garantizar que todos y cada uno de los elementos de la población pertenezcan a uno y sólo a uno de los conglomerados, de manera que la intersección de estos subconjuntos sea vacía y su unión sea el total de la población.

Segundo, a partir del marco de conglomerados, se selecciona una muestra de conglomerados, ya sea con un *mia* o con un *mae*.†

Tercero, en caso de que el muestreo sea en una etapa, entonces se "medirán" (encuestarán) todas las unidades de los conglomerados seleccionados. En caso contrario, al interior de cada uno de los conglomerados se tomará nuevamente una muestra, como en la primera etapa de selección puede ser a través de un *mia* o un *mae*.

El uso del muestreo por conglomerados ofrece algunas ventajas, pero también desventajas. Entre las primeras se puede contar el hecho de que no se necesita un marco de muestreo de las unidades últimas (unidades de observación). Por otro lado, cuando se definen unidades geográficas como conglomerado, el costo de desplazamientos disminuye, puesto que sólo será necesario visitar algunas áreas para recolectar la información.

Esas ventajas son muy atractivas, sin embargo también tiene sus desventajas, las cuales se ven reflejadas principalmente en el descenso de la precisión de los estimadores.



### ■ Ejemplo 4

En una ciudad se desea conocer cuántos libros en promedio leen los niños que se encuentran inscritos en 6º grado de primaria en escuelas públicas.

En este caso podría decirse que se cuenta con un marco de muestreo (listado) de todos los alumnos de 6º grado de las escuelas públicas, entonces, ¿por qué no elegir directamente un *mia*? Si se hiciera un *mia*, habría la posibilidad de que la muestra tuviera niños de todas las escuelas, o de la mayoría de ellas, por lo que se tendría que ir a todas, y el costo en cuanto a tiempo y recursos económicos sería alto.

Por tanto, puesto que en general se puede considerar que las escuelas del gobierno son similares entre sí, entonces, cada una de las escuelas puede tomarse como un conglomerado y, a partir del listado de todas las escuelas, hacer una selección de algunas de ellas.

Suponga que en total se tienen 120 escuelas de gobierno en la ciudad, y que se seleccionan de manera aleatoria 10 de ellas. Estimar el promedio de libros leídos por los niños. Se supone que una vez seleccionada la escuela, se pregunta a todos los niños de 6º grado cuántos libros leyeron el año pasado.

† Tanto en la *mia* como en el *mae* se asignan probabilidades iguales de salir en muestra a cada una de las unidades de muestreo. Para el caso de muestreo por conglomerados, se puede seguir el mismo procedimiento. Sin embargo, puesto que en el muestreo por conglomerados, éstos pueden tener tamaños distintos, entonces existe la posibilidad de realizar su selección con probabilidades distintas, en particular se puede hacer con probabilidades proporcionales al tamaño de los conglomerados (número de integrantes).

Por tanto,

$N$ = 120 Total de escuelas en la ciudad
$n$ = 10 Total de escuelas en muestra
$m_i$ = Número total de niños de 6º grado en la escuela $i$, donde $i$ = 1, 2, ... $n$
$x_i$ = Número total de libros leídos por los niños de la escuela $i$

La información del número de libros leídos por los niños de 6º grado de las escuelas en muestra es la siguiente:

| Conglomerado | $m_i$ | $x_i$ |
|---|---|---|
| 1 | 60 | 180 |
| 2 | 50 | 125 |
| 3 | 53 | 190 |
| 4 | 70 | 132 |
| 5 | 65 | 151 |
| 6 | 58 | 92 |
| 7 | 62 | 75 |
| 8 | 74 | 120 |
| 9 | 45 | 104 |
| 10 | 56 | 62 |

Como $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} m_i}$, entonces $\bar{x} = \dfrac{1,31}{593} = 2.076$

es decir, el promedio de libros leídos por niño se calcula considerando el total de libros en los conglomerados en la muestra entre el total de niños de los conglomerados en ella.

La varianza de $\bar{x}$ se calcula como

$$V(\bar{x}) = \frac{N - n\sum_{i=1}^{n}(x_i - \bar{x}m_i)}{Nn\bar{M}^2 \quad n-1}$$

con $M$ igual al número total de niños en todas las escuelas. Por tanto, $\bar{M}$ = si 56.7, entonces

$$V(\bar{x}) = \frac{120-10}{120(10)(56.7)^2} = \frac{18\ 231.82}{9}$$

$$V(\bar{x}) = 0.05782873$$

de donde se obtiene el intervalo de confianza

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \sqrt{V(\bar{x})}$$

$$\bar{x} \pm 1.96 \sqrt{0.05783}$$

$$2.076 \pm 0.4713$$

es decir, en promedio, los niños de 6º grado de las escuelas públicas leen entre 1.61 y 2.55 libros en el año, esto con un 95% de confianza.

En este caso se contó con información acerca del número total de niños en las escuelas, cuando no se tiene esta información, una buena aproximación se realiza como:

$$\overline{M} = \frac{\sum_{i=1}^{n} m_i}{n}$$

## Muestras complejas

De acuerdo con lo que se ha visto a lo largo del apartado, para la aplicación de cada uno de los tipos de muestreo que se han descrito es necesario el cumplimiento de determinadas condiciones, las cuales son principalmente en cuanto al tipo de información necesaria, es decir, en cuanto al marco de muestreo. Según lo planteado hasta este momento, para realizar un muestreo aleatorio simple es necesario contar con un marco de muestreo (listado) de todos y cada uno de los elementos que integran la población, de entre los cuales se realizará la selección directa de una muestra; lo mismo sucede en el caso del muestreo aleatorio estratificado, para el cual es necesario contar con información que permita realizar una estratificación (clasificación) de todos y cada uno de los elementos de la población, así como el listado de todos los elementos pertenecientes a cada uno de los estratos, para posteriormente realizar la selección directa de una muestra en cada uno de los estratos; en tanto que para el caso del muestreo por conglomerados, el cual pareciera ser el menos exigente en cuanto a cantidad de información. Se requiere del listado de todos y cada uno de los conglomerados, a partir del cual se realizará la selección directa de una muestra de ellos.

Como se puede observar, la aplicación de cada uno de estos tipos de muestreo requiere de determinada información. Pero, ¿qué pasa cuando no se cuenta con esa información? Por ejemplo, considere que desea realizar una encuesta a nivel nacional sobre la percepción que tienen los mexicanos, de 18 años y más, acerca de lo adecuada que resulta la constitución política del país.† Vea las posibilidades de muestreo con las que contaría hasta este momento.

Si deseara aplicar un muestreo aleatorio simple, lo que se requiere es un listado de todos los mexicanos de 18 años y más residentes en el país, en el momento de la aplicación de la encuesta. Este listado

† El diseño del esquema de muestreo completo de la Encuesta Nacional de actitudes, percepciones y valores, se puede consultar en http://www.bibliojuridica.org/libros/3/1324/16.pdf. Esta encuesta fue realizada por especialistas del Instituto de Investigaciones Jurídicas en colaboración con el Instituto de Investigaciones Sociales de la UNAM.

deberá contar con información que permita identificar de manera única a cada uno de los mexicanos, así como información acerca de su ubicación. De acuerdo con lo que se ha visto hasta este momento, a partir de este listado de cerca de 59 millones de mexicanos,† se realizará la selección de manera directa a partir de la generación de números aleatorios o muestro sistemático.†† ¡Demasiado complicado! Ante la imposibilidad de contar con un listado con estas características y lo poco práctico que resulta realizar de esta manera la selección de los integrantes de la muestra,††† se tiene que recurrir a métodos "indirectos" de selección (diferentes etapas de selección).

Si lo que se quisiese es hacer un muestreo aleatorio estratificado, clasificando de acuerdo al tamaño de localidad, o por estratos socioeconómicos, por ejemplo, la necesidad de un marco de muestreo por individuos no se salva. Es decir, puesto que la selección de la muestra se tendría que realizar al interior de cada uno de los estratos, entonces, se requiere de un listado exhaustivo de todos los mexicanos incluidos en cada uno de los estratos, y realizar la selección directa en cada uno de ellos.

Por otro lado, a diferencia de los dos casos anteriores, para un muestreo por conglomerados, el marco muestral estará conformado por todos y cada uno de los conglomerados, los cuales para este caso en particular podrían ser los municipios, localidades, AGEB's,†††† etc. Los conglomerados generalmente son unidades mayores que las unidades de observación. A partir del listado de todos los conglomerados, se selecciona de manera directa una muestra de conglomerados, al interior de los cuales se realiza la recolección de la información de todos los elementos que los conformen. Por tanto, la información con la cual se construyó la muestra fue el listado de todos los conglomerados y el listado de los elementos de un número limitado de conglomerados.



## Esquema de muestreo en una sola etapa

Como se puede observar en la descripción anterior, la información necesaria para realizar la selección de la muestra depende de las especificaciones del tipo de muestreo que se desea realizar, el cual, por lo demás, está determinado por los objetivos de la investigación.

Una particularidad de cada uno de los esquemas de muestreo mencionados en los apartados anteriores es que se llega a la muestra a través de una etapa de selección. Es decir, en cada uno de los esquemas seleccionados, la muestra es consecuencia de la selección directa de los elementos. A este tipo de esquema de muestreo se le llama muestreo en una etapa.

En situaciones reales de muestras con alcance geográfico muy grande, es poco factible que se pueda aplicar el muestreo en una sola etapa debido principalmente a la imposibilidad de contar con un marco de muestreo con las características para hacer un muestreo de este tipo.††††† Entonces, a lo que se recurre



† De acuerdo con información del Censo General de Población y Vivienda del 2000, la población de 18 años y más en el país fue de 58 772 635.

†† Suponiendo que el listado no tienen algún orden en particular, es decir, que su arreglo es aleatorio.

††† En caso de que se pudiera contar con un listado con estas características, el problema de la selección de la muestra quedaría resuelto con los paquetes estadísticos que tienen la opción de extraer muestras simples de manera muy sencilla.

†††† Las Áreas Geoestadísticas Básicas (AGEB) son áreas geográficas delimitadas de acuerdo con criterios establecidos por el INEGI, las AGEB son la extensión territorial, que corresponde a la subdivisión de las AGEM (Áreas Geoestadísticas Municipales), constituye la unidad básica del Marco Geoestadístico Nacional y, dependiendo de sus características, se clasifican en dos tipos; Áreas Geoestadísticas Básicas Urbanas y Áreas Geoestadísticas.

††††† Además, debido a la forma en la que se selecciona a los elementos de la muestra, su dispersión es muy grande, lo cual, si bien tiene ventajas en cuanto a los estimadores, en la parte logística del levantamiento de la información, se vuelve demasiado complejo y costoso.

en estos casos es al diseño de muestras que consideren más de una etapa de selección y, en algunos casos, se combinan estos tres tipos de muestreo. Es decir, antes de llegar a la selección directa de la unidad de observación, se realiza la selección de unidades mayores.

## Esquemas complejos de muestreo

De manera general, los muestreos complejos hacen referencia a que antes de la selección de las unidades de observación (unidades últimas de muestreo), se hizo la selección de unidades de muestreo más grandes, en una o varias etapas. Por tanto, para realizar un muestreo complejo se requiere definir cuántas etapas de selección se realizarán, así como las unidades de muestreo que se tomarán en cada una de las etapas, y la forma en la que se llevará a cabo la selección de cada una de estas unidades en cada una de las etapas.[†]

### ■ Ejemplo 5

Suponga que se desea conocer la percepción que tienen los universitarios acerca de la Constitución de los Estados Unidos Mexicanos, y con lo que se cuenta únicamente es con el listado de todas las universidades del país.

De acuerdo con el enunciado anterior, no se tiene a la mano el listado de todos los alumnos inscritos en todas las universidades y su recolección resulta por demás compleja. Pero con lo que sí se cuenta es con un listado de todas las universidades. Por consiguiente, las etapas y unidades de muestreo para un esquema de muestreo complejo podrían quedar especificadas de la siguiente manera.

| Etapa de selección | Unidad de muestreo |
|---|---|
| 1a. etapa | **Unidad primaria de muestreo:** Universidades<br>**Marco de muestreo:** Listado de todas las universidades del país.<br>**Estratificación:** En este nivel se puede aplicar una estrategia de estratificación, considerando dos estratos (universidades públicas o privadas); esto siempre y cuando se considere que las opiniones vertidas por los estudiantes están relacionadas con el tipo de universidad a la que asisten. |
| 2a. etapa | **Unidad secundaria de muestreo:** Grupos de alumnos.[††]<br>**Marco de muestreo:** Listado de todos los grupos de las universidades seleccionadas. |
| 3a. etapa | **Unidad terciaria de muestreo:** Unidad última de muestreo, alumnos inscritos.<br>**Marco de muestreo:** Listado de todos los alumnos inscritos en las universidades seleccionadas en la segunda etapa de muestreo. |

[†] La Encuesta Nacional de Prácticas y Consumo Culturales realizada en 2003 empleó un diseño de muestra complejo, el cual se puede consultar en la dirección http://sic.conaculta.gob.mx/encuesta/encuesta.zip.

[††] En este caso se está suponiendo que las universidades tienen manera de clasificar a cada uno de los alumnos en uno y sólo un grupo. Otra manera en la que se puede realizar el muestro es a través de la selección de clases, aunque esta forma tiene más inconvenientes que la selección por grupos debido a que si bien todos los alumnos están al menos en una clase, algunos pertenecen a más de una, con lo que se estaría corriendo el riesgo de estar considerando más de una vez a un mismo alumno. Por tanto, para este ejercicio considere que es posible asignar a cada uno de los alumnos a un grupo.

Para que un universitario sea seleccionado, con anterioridad debió haber sido seleccionado su grupo, así como la universidad en la que se encuentra inscrito; es decir, se tiene una especie de selección condicionada.

*Estrategia de selección*. La estrategia de selección empleada en cada una de las etapas de muestreo no necesariamente es la misma, sino más bien depende en gran medida del tipo y cantidad de información con la que se cuente. Por consiguiente, para el caso anterior será posible tener los siguientes esquemas de muestreo.

| Etapa de selección | Estrategia de selección |
|---|---|
| 1a. etapa | • Selección aleatoria de conglomerados (universidades). En caso de que se haya realizado estratificación, esta selección se realizaría en cada uno de los estratos de manera independiente.<br>• Selección con probabilidad proporcional al tamaño del conglomerado (es decir, al número de alumnos inscritos en cada universidad).<br>• En caso de que se haya decidido estratificar, selección de conglomerados al interior de cada uno de los estratos, ya sea con muestreo aleatorio simple o con probabilidad proporcional. |
| 2a. etapa | • Selección aleatoria de grupos.<br>• Selección de grupos con probabilidad proporcional. |
| 3a. etapa | • Selección aleatoria de estudiantes al interior de cada uno de los grupos.<br>• Selección sistemática de estudiantes en cada grupo. |



## Efecto de diseño

Obvio esta forma de hacer la selección de la muestra (esquema de muestreo) tiene costos en cuanto a la precisión de los estimadores. Estos costos se denominan *efecto de diseño*. El efecto de diseño es el costo asociado a la estimación de una muestra, debido a que no se utilizó un muestreo aleatorio simple. El efecto de diseño† no es otra cosa que el cociente entre la varianza del diseño de muestra utilizado y la varianza que se hubiese obtenido si se hubiera aplicado un muestreo aleatorio simple.

$$DEF = \frac{V(diseño)}{V(mia)}$$

Dependiendo del esquema de muestreo que se haya utilizado el factor se alejará o se acercará a uno. Este factor también se puede calcular para los casos en que la muestra es obtenida en una sola etapa, con un muestreo distinto al *mia*.††

† Generalmente para hacer referencia al *efecto de diseño* se utiliza la abreviatura *DEF* por su escritura en inglés, *Design effect*.
†† Si la muestra se obtuvo con un muestreo aleatorio simple, el valor del *DEF* es 1.

## Muestras extraídas en una sola etapa

i) Si la muestra fue extraída en una sola etapa y se utilizó un muestreo estratificado, entonces:

$$V(mae) \leq V(mia)$$

Por lo que el *DEF* será menor que 1, es decir, se tiene una ganancia en cuanto a precisión.†

ii) En tanto, si lo que se utiliza es un muestreo por conglomerados, entonces la relación entre las varianzas de los dos muestreos será la siguiente.

$$V(mia) < V(mconglomerados)$$

Es decir, a diferencia del caso anterior, al utilizar un muestreo por conglomerados,†† la varianza es mayor, por lo que el *DEF* asociado a este esquema de muestreo será mayor que uno.

Este hecho es de gran importancia, puesto que como se vio en el apartado del cálculo del tamaño de muestra, la magnitud de la varianza tendrá repercusiones en el tamaño de la muestra.

La expresión para el cálculo del tamaño de muestra considerando el efecto de diseño es la siguiente:

$$n = \frac{Z^2_{\alpha/2}\sigma^2}{d^2} * DEF$$

Por tanto, si se emplea una *mia*, el tamaño de muestra no se verá afectado, puesto que el valor del *DEF* será uno. Manteniendo los mismos niveles de confianza y precisión, si se aplica un *mae*, el tamaño de muestra suficiente, será menor al del *mia*, en virtud de que el *DEF* será menor que uno. En tanto que si se aplica un muestreo por conglomerados, el tamaño de muestra necesario para obtener estimaciones con los mismos niveles de confianza y precisión, será mayor que el requerido para la *mia* o el *mae*, puesto que el *DEF* será mayor que uno.

## Muestras extraídas en más de una etapa

En virtud de que el efecto de diseño no es más que la varianza asociada al esquema de muestreo aplicado, para el caso de las muestras extraídas en dos etapas es de esperar que el efecto de diseño sea mayor que uno, debido a que en cada una de las etapas de selección se tienen pérdidas en cuanto a precisión.

El cálculo de las varianzas para esquemas de muestreo complejo no es sencillo, sin embargo, en caso de que se utilice un diseño complejo, será necesario calcular el *DEF* asociado a éste y utilizarlo para el análisis, en caso contrario, se corre el riesgo de calcular errores estándar demasiados pequeños, con lo que se estaría sobreestimando su precisión.

De manera general, el cálculo de la varianza de un esquema de muestreo complejo se realiza partiendo de la última etapa de selección, calculando las estimaciones de los totales y varianzas, de acuerdo con la estrategia de selección empleada en cada una de las etapas.

† Para mayor explicación vaya al apartado de "Muestreo aleatorio estratificado" de este capítulo.

†† Es decir, el costo asociado a no contar con información suficiente para poder realizar otro tipo de muestreo se ve reflejado en la mayor varianza obtenida en las estimaciones.

Cuando se esté considerando la posibilidad de utilizar un esquema de muestreo complejo debe evaluarse muy bien la conveniencia del mismo. La sencillez del esquema de muestreo se verá retribuida cuando se realice el análisis de la encuesta. Cada etapa de muestreo que se agregue, cada estratificación que se realice, se verá reflejada en la complejidad para el análisis, así como en el incremento del tamaño de la muestra. Por tanto, debe sopesarse muy bien la utilización de un diseño complejo, y si aún así se decide que es la mejor opción, se deberá dar mucha importancia a la definición de las etapas de selección y a las unidades de muestreo.

## Resumen

Generar información no es una tarea fácil. Requiere de atención y dedicación de la toma de decisiones a partir del objetivo de la investigación. Se requiere tener conocimiento acerca de los distintos métodos, de sus virtudes y desventajas. En este capítulo se presentaron diversas estrategias de muestreo, tanto probabilísticas como no probabilísticas, las cuales por sí solas o combinadas, representan excelentes herramientas para la generación de información. La construcción del esquema de muestreo adecuado depende de diversos factores entre los que se encuentran el conocimiento que se tenga del fenómeno (aspecto) que se desea estudiar, el tipo de información con que se cuenta para realizar el diseño de la muestra, los recursos financieros, así como los conocimientos acerca de los tipos de muestras que se pueden realizar. En este último aspecto, a lo largo del capítulo se proporcionaron los elementos básicos para tener un panorama general del diseño de distintas estrategias de muestreo.



## Ejercicios

* **5.1** ¿Cuál es la principal características de un muestreo de tipo probabilístico y uno no probabilístico?

* **5.2** ¿En qué casos es recomendable utilizar un muestreo probabilístico?

* **5.3** ¿Cuándo es preferible utilizar una muestra en lugar de un censo?

* **5.4** ¿En qué consiste el muestreo irrestricto aleatorio?

* **5.5** ¿Qué es un marco de muestreo?

* **5.6** ¿En un muestreo irrestricto aleatorio, qué elementos determinan el tamaño de muestra?

**5.7** Suponga que desea saber qué porcentaje de los niños de una escuela primaria toman cursos extraclase. En total la escuela tiene 450 alumnos inscritos. Calcular los tamaños de muestra asociados con un nivel de confianza del 95% y diferentes niveles de precisión (2, 3, 4 y 5 puntos porcentuales). ¿Cuál es el comportamiento del tamaño de muestra conforme va disminuyendo la precisión?

Nota: Los ejercicios marcados con un * son para discusión en clase.

**5.8** Repita el ejercicio anterior, pero ahora manteniendo fijo el nivel de precisión 5 puntos porcentuales, y variando el nivel de confianza (80%, 85%, 90%, 95% y 99%). ¿Cuál es el comportamiento del tamaño de la muestra?

*** 5.9** ¿Cómo es la varianza de un muestro aleatorio por conglomerados comparada con la de un muestreo estratificado?, ¿por qué?

**5.10** Suponga que se cuenta con la información del número de hijos vivos de una muestra de 100 mujeres. Y que esta muestra fue tomada de una población total de 550 mujeres, 260 del área rural y 290 del área urbana. Calcular la media y varianza del número promedio de hijos, *a*) suponiendo que la muestra fue extraída con un *mia*, y *b*) suponiendo que se utilizó estratificación de acuerdo con el lugar de residencia, es decir, urbano o rural.

| ***Rural (47)*** | | | | | ***Urbano (53)*** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 7 | 4 | 5 | 1 | 2 | 1 | 3 | 3 | 1 | 3 |
| 5 | 8 | 3 | 6 | 3 | 3 | 2 | 4 | 4 | 2 | 3 |
| 7 | 5 | 5 | 7 | 4 | 4 | 1 | 7 | 1 | 2 | 4 |
| 4 | 4 | 3 | 3 | 7 | 1 | 2 | 5 | 2 | 4 | |
| 6 | 6 | 4 | 2 | 3 | 2 | 1 | 4 | 2 | 2 | |
| 3 | 6 | 5 | 3 | 6 | 3 | 2 | 5 | 2 | 4 | |
| 6 | 2 | 2 | 3 | 8 | 1 | 2 | 2 | 3 | 1 | |
| 7 | 4 | 1 | 4 | | 2 | 2 | 3 | 4 | 4 | |
| 3 | 3 | 7 | 5 | | 3 | 3 | 4 | 5 | 2 | |
| 7 | 8 | 8 | 8 | | 4 | 5 | 5 | 1 | 2 | |



*** 5.11** En una escuela secundaria se desea estimar el número promedio de horas al día que los estudiantes de primer grado ven la televisión. Suponga que en total se tienen 140 escuelas en la ciudad. ¿Qué tipo de muestreo se recomienda para estimar el promedio de horas diarias que los estudiantes ven la televisión? Defina la población objetivo y el esquema de muestreo que se proponga.

**5.12** Siguiendo con el ejercicio anterior. Suponga que de las 140 escuelas se toman 10 de manera aleatoria, y al interior de cada una de ellas se pregunta a todos los niños de primer grado el número de horas al día que ven televisión. La información para cada uno de las escuelas es la siguiente.

| Escuela | Número de niños en primer grado en la escuela *i* | Número total de horas que los niños de primer grado de la escuela *i* ven la televisión |
|---|---|---|
| 1 | 60 | 180 |
| 2 | 50 | 75 |
| 3 | 53 | 132 |
| 4 | 70 | 185 |
| 5 | 65 | 240 |
| 6 | 58 | 180 |
| 7 | 62 | 98 |
| 8 | 74 | 170 |
| 9 | 45 | 150 |
| 10 | 56 | 65 |

¿Qué tipo de muestreo es el que se está aplicando? Calcule el número promedio de horas que los niños de primer grado ven la televisión, así como el intervalo de confianza alrededor de este valor.

*** 5.13** Suponga que desea realizar un estudio para conocer los hábitos de juego de los asistentes a las casas de juegos. ¿Cómo definiría la población de estudio?, ¿qué tipo de muestreo recomienda?, ¿por qué?

*** 5.14** A partir de una encuesta nacional se estimó que el número promedio de libros que leen los mexicanos es de 2.93, con un 90% de confianza y margen de error no mayor a 1.75 puntos porcentuales. ¿Cómo deben interpretarse estos valores?

**5.15** Un profesor de una facultad desea estimar el tiempo que debe darle a sus estudiantes para resolver un examen estándar que está elaborando. Con tal propósito toma una muestra aleatoria de 15 estudiantes y les aplica el examen. Los resultados obtenidos se muestran en la siguiente tabla. Estime el tiempo promedio para terminar el examen y, por tanto, el tiempo que el profesor debe darles a sus 135 alumnos, así como el límite de error de estimación.

| *Tiempo (en minutos)* | | |
|---|---|---|
| 65 | 80 | 65 |
| 83 | 60 | 83 |
| 60 | 70 | 73 |
| 75 | 90 | 74 |
| 92 | 90 | 80 |



Nota: Los ejercicios marcados con un * son para discusión en clase.