

ESTE DOCUMENTO SE EXTRAJO
DEL LIBRO:

Schaeffer, R., Mendenhall, W., y
Ott, L. (1987). Elementos de
Muestreo. (3^a. ed.). México:
Iberoámerica

COMO APOYO A LA ASIGNATURA
DE ESTADÍSTICA II SIN
PRETENSIONES DE
REPRODUCIRLO PARA FINES
COMERCIALES.

8

MUESTREO POR CONGLOMERADOS

ESTUDIO DE CASO

¿CUÁLES SON LAS CARACTERÍSTICAS DE LAS PERSONAS QUE VIVEN EN SU BARRIO?

Supóngase que una empresa quiere establecer un negocio en su barrio. ¿Cómo puede esta empresa encontrar información sobre las características de la gente que ahí vive sin realizar su propia encuesta? Una manera es consultar los datos de las estadísticas de bloques o manzanas de la Oficina de Censos de Estados Unidos. Las *estadísticas de manzanas* proporcionan información demográfica —tal como número total de residentes, número en ciertos grupos minoritarios, número de personas mayores de 65 años de edad y número de dueños y arrendatarios— sobre regiones muy pequeñas que suelen concordar con las manzanas de la ciudad. Estos datos son usados por investigadores de mercados, planificadores de viviendas y transporte y asociaciones comunitarias, entre otros.

La empresa que está considerando poner un local en su barrio abastece a los que tienen una edad de 65 años o más. Entonces esta empresa quiere estimar la proporción de residentes de esta edad que viven en un área de 40 manzanas. La empresa decide muestrear 5 de las 40 manzanas y obtener los datos de las estadísticas de manzanas. Las manzanas muestreadas forman conglomerados de personas, y entonces debe utilizarse la técnica de muestreo por conglomerados. (Este problema es una versión de un problema real a menor escala. Usualmente, el número de manzanas y el tamaño de muestra son mucho mayores.)

8.1 INTRODUCCIÓN

Se recordará que el objetivo del diseño de encuestas por muestreo es obtener una cantidad especificada de información acerca de un parámetro poblacional a un costo mínimo. El muestreo aleatorio estratificado es frecuentemente más adecuado para esto que el muestreo irrestricto aleatorio, debido a los tres principios indicados en la Sección 5.1. El muestreo sistemático frecuentemente da resultados al menos tan exactos como el muestreo irrestricto aleatorio y es más fácil de llevar a cabo, según se trató en la Sección 7.1. Este capítulo introduce un cuarto diseño, muestreo por conglomerados, el cual algunas veces proporciona más información por unidad de costo que cualquier otro de los tres diseños estudiados previamente.

DEFINICIÓN 8.1 Una muestra por conglomerados es una muestra aleatoria en la cual cada unidad de muestreo es una colección, o conglomerado, de elementos.

El muestreo por conglomerados es menos costoso que el muestreo aleatorio estratificado o irrestricto, si el costo por obtener un marco que liste todos los elementos poblacionales es muy alto o si el costo por obtener observaciones se incrementa con la distancia que separa los elementos.

Para explicarlo, supóngase que deseamos estimar el ingreso promedio por hogar en una gran ciudad. ¿Cómo debemos seleccionar la muestra? Si usamos muestreo irrestricto aleatorio, se requiere un marco que liste todos los hogares (elementos) en la ciudad, y este marco puede ser muy costoso o imposible de obtener. No podemos evitar

este problema al utilizar muestreo aleatorio estratificado porque incluso se requiere un marco para cada estrato en la población. En lugar de extraer una muestra irrestricta aleatoria de *elementos*, podríamos dividir la ciudad en regiones tales como manzanas (o conglomerados de elementos) y seleccionar una muestra irrestricta aleatoria de ellas. Esta tarea se realiza con facilidad mediante el uso de un marco que liste todas las manzanas de la unidad. Entonces se podría medir el ingreso de cada familia dentro de cada manzana muestreada.

Para ilustrar el segundo principio de la aplicación de muestreo por conglomerados, suponga que se cuenta con una lista de hogares de la ciudad. Podríamos seleccionar una muestra irrestricta aleatoria de hogares, la cual probablemente estará dispersa en toda la ciudad. El costo por realizar entrevistas en los hogares dispersos va a ser grande debido al tiempo de transporte de los entrevistadores y otros gastos relacionados. El muestreo aleatorio estratificado podría reducir estos gastos, pero el uso de muestreo por conglomerados es un método más efectivo para reducir los gastos de transporte. Los elementos dentro de un conglomerado deben estar geográficamente cerca uno de otro, y entonces los gastos de transporte se reducen. Obviamente el transporte dentro de un bloque de la ciudad sería mínimo si se comparara con el transporte asociado al muestreo irrestricto aleatorio dentro de la ciudad.

Para resumir, el muestreo por conglomerados es un diseño efectivo para obtener una cantidad especificada de información al costo mínimo bajo las siguientes condiciones:

1. No se encuentra disponible o es muy costoso obtener un buen marco que liste los elementos de la población, mientras que se puede lograr fácilmente un marco que liste los conglomerados.
2. El costo por obtener observaciones se incrementa con la distancia que separa los elementos.

Las manzanas de la ciudad son usadas frecuentemente como conglomerados de hogares o de personas, porque la Oficina de Censos de Estados Unidos reporta estadísticas de manzana muy detalladas. En los datos censales una manzana puede ser una manzana de ciudad estándar o un área de forma irregular con límites políticos o geográficos identificables. Las estadísticas de manzana contienen información de todas las áreas urbanas y lugares con concentraciones de 10,000 o más personas. En total las estadísticas de manzana cubren el 77% de la población nacional. Los datos reportados para cada manzana incluyen la población total, mezcla racial y número de unidades habitacionales, y pueden incluir el valor en dólares de la propiedad, si la casa es alquilada o propia y si tiene todos los servicios de plomería.

Las estadísticas de manzana de la Oficina de Censos son ampliamente usadas en muestreo por conglomerados por empresas de investigación de mercados, las cuales pueden desear estimar el mercado potencial de un producto, las ventas potenciales si se abre un nuevo almacén en el área, o el número potencial de clientes para un nuevo servicio, tal como una instalación de emergencias médicas.

El gobierno estatal y local muestrean manzanas (conglomerados de unidades habitacionales o personas) a fin de planear nuevos métodos y medios de transporte y además los desarrollos habitacionales. Asimismo organizaciones comunitarias, tales como iglesias, utilizan estadísticas de manzanas para determinar sitios óptimos de ampliación.

Hay muchos otros ejemplos comunes del uso de muestreo por conglomerados. Las mismas unidades habitacionales son conglomerados de personas y pueden formar

unidades de muestreo convenientes al muestrear, por ejemplo, estudiantes universitarios. Los hospitales forman conglomerados convenientes de pacientes con ciertas enfermedades para estudios del tiempo promedio de hospitalización o número promedio de recurrencias de padecimientos.

Otros elementos diferentes de personas son frecuentemente muestreados en conglomerados. Un automóvil forma un buen conglomerado de cuatro llantas para estudios de uso y seguridad de llantas. Un tablero de circuitos fabricado para una computadora forma un conglomerado de semiconductores para prueba. Un naranjo forma un conglomerado de naranjas para la investigación de infestación por insectos. Una parcela en el bosque contiene un conglomerado de árboles para la estimación de volúmenes de madera o proporción de árboles enfermos. Como usted puede ver, la lista de posibles conglomerados, que son unidades convenientes de muestreo, es infinita.

Ahora analizaremos los detalles de la selección de una muestra por conglomerados.

8.2 CÓMO SELECCIONAR UNA MUESTRA POR CONGLOMERADOS

La primera tarea en muestreo por conglomerados es especificar los conglomerados apropiados. Los elementos dentro de un conglomerado están frecuentemente juntos físicamente, por lo que tienden a presentar características similares. Dicho de otra manera, la medición en un elemento en un conglomerado puede estar altamente correlacionada con la de otro elemento. Entonces la cantidad de información acerca de un parámetro poblacional puede no incrementarse sustancialmente al tomar nuevas mediciones dentro de un conglomerado. Ya que las mediciones cuestan dinero, un experimentador podría desperdiciar presupuesto si es que selecciona un conglomerado de gran tamaño. Sin embargo pueden ocurrir situaciones en las cuales los elementos dentro de un conglomerado son muy diferentes entre sí. En tales casos una muestra que contenga pocos conglomerados grandes puede producir una estimación muy buena de un parámetro poblacional, tal como la media.

Por ejemplo supóngase que los conglomerados están formados por cajas de componentes que van saliendo de una línea de producción, un conglomerado de componentes por línea. Si todas las líneas tienen aproximadamente la misma tasa de componentes defectuosos, entonces cada conglomerado (caja) es aproximadamente tan variable con respecto a calidad como la población completa. En este caso se puede obtener un buen estimador de la proporción de productos defectuosos con base en uno o dos conglomerados.

En contraste, supóngase que los distritos escolares se especifican como conglomerados de hogares para estimar la proporción de familias que apoyan un plan de rezonificación. Ya que los conglomerados contienen muchos hogares, los recursos permiten únicamente el muestreo de un número pequeño de conglomerados, dos o tres, por ejemplo. En este caso en un distrito la mayoría de las familias puede estar satisfecha con sus escuelas y no apoyar la rezonificación, mientras que en otro distrito la mayoría puede estar inconforme con sus escuelas y favorecer decididamente la rezonificación. Una muestra pequeña de distritos escolares puede no contener a uno u otro de estos grupos, produciendo por esto un estimador muy deficiente. Se puede obtener mayor información muestreando un número grande de conglomerados de menor tamaño.

El problema de elegir un tamaño apropiado del conglomerado puede ser aún más complicado cuando se dispone de un número infinito de posibles tamaños de conglomerados, como en la selección de parcelas forestales para la estimación de la proporción de árboles enfermos. Si existe variabilidad en la densidad de árboles enfermos a lo largo y ancho del bosque, entonces muchas parcelas (conglomerados) pequeñas, localizadas aleatoria o sistemáticamente, pueden ser lo deseable. Sin embargo, localizar aleatoriamente una parcela en el bosque consume mucho tiempo, y una vez localizada, el muestreo de muchos árboles es económicamente conveniente. Entonces muchas parcelas pequeñas son ventajosas para controlar la variabilidad, pero pocas parcelas grandes son económicamente recomendables. Se debe encontrar un equilibrio entre el número y tamaño de las parcelas. No existen buenas reglas que funcionen siempre para tomar esta decisión. Cada problema debe ser estudiado separadamente; pero las encuestas piloto pueden ayudar al experimentador a encontrar la dirección correcta.

Nótese la principal diferencia entre la construcción óptima de estratos (Capítulo 5) y la construcción de los conglomerados. Los estratos deben ser tan homogéneos (semejantes) entre ellos, como sea posible, pero un estrato debe diferir tanto como sea posible de otro con respecto a la característica que está siendo medida. Los conglomerados, por otro lado, deben ser tan heterogéneos (diferentes) entre ellos como sea posible, y un conglomerado debe ser muy similar a otro para poder aprovechar las ventajas económicas del muestreo por conglomerados.

Una vez que los conglomerados han sido especificados se debe conformar un marco que liste todos los conglomerados de la población. Entonces se selecciona una muestra irrestricta aleatoria de conglomerados de este marco mediante el uso de los métodos de la Sección 4.2. Se ilustra con el siguiente ejemplo.

EJEMPLO 8.1

Un sociólogo quiere estimar el ingreso promedio por persona en cierta ciudad pequeña. No existe una lista disponible de adultos residentes. ¿Cómo se debe diseñar la encuesta por muestreo?

SOLUCIÓN

El muestreo por conglomerados parece ser la elección lógica para el diseño de la encuesta porque no se encuentra disponible una lista de elementos. La ciudad es dividida en bloques rectangulares, excepto las dos áreas industriales y los tres parques que contienen pocas casas. El sociólogo decide que cada bloque de la ciudad va a ser considerado como un conglomerado, las dos áreas industriales van a ser consideradas como otro, y, finalmente, los tres parques van a considerarse un conglomerado más. Los conglomerados son numerados sobre un mapa de la ciudad, con los números del 1 al 415. El experimentador tiene tiempo y dinero suficientes para muestrear $n = 25$ conglomerados y entrevistar a cada hogar dentro de cada uno. Entonces se seleccionan 25 números aleatorios entre 1 y 415 de la Tabla 2 del Apéndice, y los conglomerados con esos números son marcados en el mapa. Despues se asignan los entrevistadores a cada uno de los conglomerados seleccionados.

8.3 ESTIMACIÓN DE UNA MEDIA Y UN TOTAL POBLACIONALES

El muestreo por conglomerados es muestreo irrestricto aleatorio con cada unidad de muestreo conteniendo un número de elementos. Por esto los estimadores de la media poblacional μ y el total τ son similares a los de muestreo irrestricto aleatorio. En particular la media muestral \bar{y} es un buen estimador de la media poblacional μ . En esta sección se estudian un estimador de μ y dos estimadores de τ .

En este capítulo se utiliza la siguiente notación:

N = número de conglomerados en la población

n = número de conglomerados seleccionados en una muestra irrestricta aleatoria

m_i = número de elementos en el conglomerado i , $i = 1, \dots, N$

$\bar{m} = \frac{1}{n} \sum_{i=1}^n m_i$ = tamaño promedio del conglomerado en la muestra

$$M = \sum_{i=1}^N m_i = \text{número de elementos en la población}$$

$\bar{M} = \frac{M}{N}$ = tamaño promedio del conglomerado en la población

γ_i = total de todas las observaciones en el i -ésimo conglomerado

El estimador de la media poblacional μ es la media muestral \bar{y} , la cual es dada por

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i}$$

Entonces la media \bar{y} toma la forma de un estimador de razón, como se ha desarrollado en el Capítulo 6, con m_i tomando el lugar de x_i . Entonces la varianza estimada de \bar{y} toma la forma de la varianza de un estimador de razón, dada por la Ecuación (6.2).

Estimador de la media poblacional μ :

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i} \quad (8.1)$$

Varianza estimada de \bar{y} :

$$\hat{V}(\bar{y}) = \left(\frac{N-n}{Nn\bar{M}^2} \right) \frac{\sum_{i=1}^n (y_i - \bar{y}m_i)^2}{n-1} \quad (8.2)$$

Límite para el error de estimación

$$2\sqrt{\hat{V}(\bar{y})} = 2 \sqrt{\left(\frac{N-n}{Nn\bar{M}^2} \right) \frac{\sum_{i=1}^n (\gamma_i - \bar{y}m_i)^2}{n-1}} \quad (8.3)$$

Aquí \bar{M} puede ser estimado por \bar{m} si se desconoce M .

La varianza estimada en la Ecuación (8.2) es sesgada y sería un buen estimador de $V(\bar{y})$ únicamente si n fuera grande, digamos $n \geq 20$. El sesgo desaparece cuando los tamaños de los conglomerados m_1, m_2, \dots, m_N son iguales.

Vamos a ilustrar el uso de la fórmula con un ejemplo.

EJEMPLO 8.2

Se realizan entrevistas en cada uno de los 25 bloques muestreados en el Ejemplo 8.1. Los datos sobre ingresos se presentan en la Tabla 8.1. Use los datos para estimar el ingreso promedio por persona en la ciudad y establezca un límite para el error de estimación.

TABLA 8.1 Ingreso por persona

Conglomerado <i>i</i>	Número de residentes, <i>m_i</i>	Ingreso total por conglomerado <i>y_i</i>	Conglomerado <i>i</i>	Número de residentes <i>m_i</i>	Ingreso total por conglomerado <i>y_i</i>
1	8	\$ 96,000	14	10	\$49,000
2	12	121,000	15	9	53,000
3	4	42,000	16	3	50,000
4	5	65,000	17	6	32,000
5	6	52,000	18	5	22,000
6	6	40,000	19	5	45,000
7	7	75,000	20	4	37,000
8	5	65,000	21	6	51,000
9	8	45,000	22	8	30,000
10	3	50,000	23	7	39,000
11	2	85,000	24	3	47,000
12	6	43,000	25	8	41,000
13	5	54,000			
				$\sum_{i=1}^{25} m_i =$	$\sum_{i=1}^{25} y_i =$
				151	\$1,329,000

SOLUCIÓN

El mejor estimador de la media poblacional μ es dado por la Ecuación (8.1) y se calcula como sigue:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i} = \frac{\$1,329,000}{151} = \$8801$$

Para calcular $\hat{V}(\bar{y})$, necesitamos las siguientes cantidades:

$$\sum_{i=1}^{25} y_i^2 = y_1^2 + y_2^2 + \dots + y_{25}^2$$

$$= (96,000)^2 + (121,000)^2 + \dots + (41,000)^2$$

$$= 82,039,000,000$$

$$\sum_{i=1}^{25} m_i^2 = m_1^2 + m_2^2 + \dots + m_{25}^2$$

$$= (8)^2 + (12)^2 + \dots + (8)^2 = 1,047$$

$$\sum_{i=1}^{25} y_i m_i = y_1 m_1 + y_2 m_2 + \dots + y_{25} m_{25}$$

$$= (96,000)(8) + (121,000)(12) + \dots + (41,000)(8)$$

$$= 8,403,000$$

La siguiente igualdad es fácilmente establecida:

$$\sum_{i=1}^n (y_i - \bar{y}m_i)^2 = \sum_{i=1}^n y_i^2 - 2\bar{y} \sum_{i=1}^n y_i m_i + \bar{y}^2 \sum_{i=1}^n m_i^2$$

Sustituyendo en esta ecuación los datos de la Tabla 8.1 se tiene

$$\begin{aligned} \sum_{i=1}^{25} (y_i - \bar{y}m_i)^2 &= 82,039,000,000 - 2(8801)(8,403,000) \\ &\quad + (8801)^2(1047) \\ &= 15,227,502,247 \end{aligned}$$

Ya que M es desconocido, la \bar{M} que aparece en la Ecuación (8.2) debe ser estimada por \bar{m} , donde

$$\bar{m} = \frac{\sum_{i=1}^n m_i}{n} = \frac{151}{25} = 6.04$$

El Ejemplo 8.1 nos da $N = 415$. Entonces de la Ecuación (8.2)

$$\begin{aligned} \hat{V}(\bar{y}) &= \left(\frac{N-n}{Nn\bar{M}^2} \right) \frac{\sum_{i=1}^n (y_i - \bar{y}m_i)^2}{n-1} \\ &= \left[\frac{415-25}{(415)(25)(6.04)^2} \right] \left(\frac{15,227,502,247}{24} \right) = 653,785 \end{aligned}$$

Entonces la estimación de μ con un límite para el error de estimación, es dada por

$$\bar{y} \pm 2\sqrt{\hat{V}(\bar{y})}, \quad \text{o sea} \quad 8801 \pm 2\sqrt{653,785}, \quad \text{o sea} \quad 8801 \pm 1617$$

La mejor estimación del ingreso promedio por persona es \$ 8801, y el error de estimación debe ser menor que \$ 1617 con una probabilidad cercana a 0.95. Este límite para el error de estimación es bastante grande; podría ser reducido mediante el muestreo de más conglomerados y, consecuentemente, incrementando el tamaño de muestra.

El total poblacional τ es ahora $M\mu$ porque M denota el número total de elementos en la población. Por ende, como en muestreo irrestricto aleatorio, $M\bar{y}$ proporciona un estimador de τ .

Estimador del total poblacional τ :

$$M\bar{y} = M \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i} \quad (8.4)$$

Varianza estimada de $M\bar{y}$:

$$\hat{V}(M\bar{y}) = M^2 \hat{V}(\bar{y}) = N^2 \left(\frac{N-n}{Nn} \right) \frac{\sum_{i=1}^n (y_i - \bar{y}m_i)^2}{n-1} \quad (8.5)$$

Límite para el error de estimación:

$$2\sqrt{\hat{V}(M\bar{y})} = 2 \sqrt{N^2 \left(\frac{N-n}{Nn} \right) \frac{\sum_{i=1}^n (y_i - \bar{y}m_i)^2}{n-1}} \quad (8.6)$$

Nótese que el estimador $M\bar{y}$ es útil únicamente si se conoce el número de elementos M en la población.

EJEMPLO 8.3

Utilice los datos de la Tabla 8.1 para estimar el ingreso total de todos los residentes de la ciudad, y ponga un límite para el error de estimación. Existen 2500 residentes en la ciudad.

SOLUCIÓN

La media muestral \bar{y} se calcula de \$ 8801 en el Ejemplo 8.2. Entonces la estimación de τ es

$$M\bar{y} = 2500(8801) = \$22,002,500$$

La cantidad $\hat{V}(\bar{y})$ se calcula con el método usado en el Ejemplo 8.2, excepto que M ahora puede ser usado en lugar de \bar{m} . La estimación de τ con un límite para el error de estimación es

$$\bar{M}\bar{y} \pm 2\sqrt{\hat{V}(\bar{M}\bar{y})} = \bar{M}\bar{y} \pm 2\sqrt{M^2\hat{V}(\bar{y})}$$

$$22,002,500 \pm 2\sqrt{(2500)^2(653,785)}$$

$$22,002,500 \pm 4,042,848$$

De nuevo este límite para el error de estimación es grande, y podría ser reducido incrementando el tamaño de muestra.

Frecuentemente el número de elementos en la población no es conocido en problemas donde el muestreo por conglomerados es apropiado. Entonces no podemos usar el estimador $\bar{M}\bar{y}$, pero podemos formar otro estimador del total poblacional que no depende de M . La cantidad \bar{y}_t , dada por

$$\bar{y}_t = \frac{1}{n} \sum_{i=1}^n y_i \quad (8.7)$$

es el promedio de los totales de conglomerados para los n conglomerados muestreados. Es por esto que \bar{y}_t es un estimador insesgado del promedio de los N totales de conglomerados en la población. Por el mismo razonamiento empleado en el Capítulo 4, $N\bar{y}_t$ es un estimador insesgado de la suma de los totales de conglomerados o, equivalentemente, del total poblacional τ .

Por ejemplo es altamente improbable que se conozca el número de adultos varones en una ciudad, por lo que el estimador $N\bar{y}_t$ tendrá que ser usado en lugar de $\bar{M}\bar{y}$ para estimar τ .

Estimador del total poblacional τ , el cual no depende de M :

$$N\bar{y}_t = \frac{N}{n} \sum_{i=1}^n y_i \quad (8.8)$$

Varianza estimada de $N\bar{y}_t$:

$$\hat{V}(N\bar{y}_t) = N^2 \hat{V}(\bar{y}_t) = N^2 \left(\frac{N-n}{Nn} \right) \frac{\sum_{i=1}^n (y_i - \bar{y}_t)^2}{n-1} \quad (8.9)$$

Límite para el error de estimación:

$$2\sqrt{\hat{V}(N\bar{y}_t)} = 2 \sqrt{N^2 \left(\frac{N-n}{Nn} \right) \frac{\sum_{i=1}^n (y_i - \bar{y}_t)^2}{n-1}} \quad (8.10)$$

Si existe una gran cantidad de variación entre los tamaños de los conglomerados y si los tamaños están altamente correlacionados con los totales de conglomerados, la

varianza de $N\bar{y}_t$ [Ecuación (8.9)] es generalmente mayor que la varianza de $\bar{M}\bar{y}$ [Ecuación (8.5)]. El estimador $N\bar{y}_t$ no usa la información proporcionada por los tamaños de los conglomerados m_1, m_2, \dots, m_n y por esto puede ser menos preciso

EJEMPLO 8.4

Use los datos de la Tabla 8.1 para estimar el ingreso total de todos los residentes de la ciudad si M no es conocido. Establezca un límite para el error de estimación.

SOLUCIÓN

El Ejemplo 8.1 nos da $N = 415$. De la Ecuación (8.8) y la Tabla 8.1, la estimación del ingreso total τ es

$$N\bar{y}_t = \frac{N}{n} \sum_{i=1}^n y_i = \frac{415}{25} (1,329,000) = \$22,061,400$$

Esta cantidad es bastante similar a la estimación dada en el Ejemplo 8.3.

Para fijar un límite al error de estimación, primero calculamos

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y}_t)^2 &= \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \\ &= 82,039,000,000 - \frac{1}{25} (1,329,000)^2 \\ &= 11,389,360,000 \end{aligned}$$

Entonces la estimación del ingreso total de todos los residentes de la ciudad, con un límite para el error de estimación, es

$$N\bar{y}_t \pm 2\sqrt{\hat{V}(N\bar{y}_t)}$$

Sustituyendo en la Ecuación (8.10), calculamos

$$\begin{aligned} N\bar{y}_t &\pm 2 \sqrt{N^2 \left(\frac{N-n}{Nn} \right) \frac{\sum_{i=1}^n (y_i - \bar{y}_t)^2}{n-1}} \\ 22,061,400 &\pm 2 \sqrt{(415)^2 \left[\frac{415-25}{(415)(25)} \right] \left(\frac{11,389,360,000}{24} \right)} \\ 22,061,400 &\pm 3,505,920 \end{aligned}$$

El límite para el error de estimación es levemente más pequeño que el límite para el estimador $\bar{M}\bar{y}$ (Ejemplo 8.3), debido parcialmente a que los tamaños de los conglomerados no están altamente correlacionados con los totales de los conglomerados en este ejemplo. En otras palabras, los tamaños de los conglomerados proporcionan poca información referente a los totales de conglomerados; por lo que el estimador insesgado $N\bar{y}_t$ parece ser mejor que el estimador $\bar{M}\bar{y}$.

Los estimadores de μ y τ poseen propiedades especiales cuando todos los tamaños de conglomerados son iguales (esto es, $m_1 = m_2 = \dots = m_N$). Primero, el estimador

\bar{y} , dado por la Ecuación (8.1), es insesgado de la media poblacional μ . Segundo, $\hat{V}(\bar{y})$, dado por la Ecuación (8.2), es un estimador insesgado de la varianza de \bar{y} . Finalmente, los dos estimadores, $M\bar{y}$ y $N\bar{y}_t$, del total poblacional τ son equivalentes.

EJEMPLO 8.5

El gerente de circulación de un periódico desea estimar el número promedio de ejemplares comprados por familia en determinada comunidad. Los costos de transporte de un hogar a otro son sustanciales. Es por eso que se listan los 4,000 hogares de la comunidad en 400 conglomerados geográficos de 10 hogares cada uno, y se selecciona una muestra irrestricta aleatoria de 4 conglomerados. Se realizan las entrevistas con los resultados que se muestran en la tabla anexa. Estime el número promedio de periódicos por hogar en la comunidad y establezca un límite para el error de estimación.

Conglomerado	Número de periódicos										Total
1	1	2	1	3	3	2	1	4	1	1	19
2	1	3	2	2	3	1	4	1	1	2	20
3	2	1	1	1	1	3	2	1	3	1	16
4	1	1	3	2	1	5	1	2	3	1	20

SOLUCIÓN

De la Ecuación (8.1)

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i}$$

Cuando $m_1 = m_2 = \dots = m_n = m$, la ecuación toma la forma

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{nm} = \frac{19 + 20 + 16 + 20}{4(10)} = 1.875$$

También puede mostrarse que

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y}m_i)^2 &= \sum_{i=1}^n y_i^2 - 2\bar{y} \sum_{i=1}^n y_i m_i + \bar{y}^2 \sum_{i=1}^n m_i^2 \\ &= \sum_{i=1}^n y_i^2 - nm^2 \bar{y}^2 \end{aligned}$$

Sustituyendo, obtenemos

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y}m_i)^2 &= (19)^2 + (20)^2 + (16)^2 + (20)^2 - 4(10)^2(1.875)^2 \\ &= 10.75 \end{aligned}$$

Entonces de la Ecuación (8.2),

$$\hat{V}(\bar{y}) = \left(\frac{N-n}{NnM^2} \right) \frac{\sum_{i=1}^n (y_i - \bar{y}m_i)^2}{n-1} = \frac{(400-4)(10.75)}{400(4)(10)^2(3)} = 0.0089$$

Por lo tanto el mejor estimador del número promedio de periódicos por familia, con un límite para el error de estimación, es

$$\bar{y} \pm 2\sqrt{\hat{V}(\bar{y})}, \quad \text{o sea} \quad 1.88 \pm 2\sqrt{0.0089}, \quad \text{o sea} \quad 1.88 \pm 0.19$$

De modo que la mejor estimación del número promedio de periódicos por hogar es 1.88, con una probabilidad alta de que el límite del error de estimación sea menor que 0.19.

8.4 SELECCIÓN DEL TAMAÑO DE MUESTRA PARA LA ESTIMACIÓN DE MEDIAS Y TOTALES POBLACIONALES

La cantidad de información en una muestra por conglomerados es afectada por dos factores, el número y el tamaño relativo de los conglomerados. No se ha presentado el último factor en ninguno de los procedimientos de muestreo ya analizados. En el problema de estimación del número de casas en un estado, con un seguro contra incendios inadecuado, el conglomerado puede ser un municipio, distritos de votación, distritos escolares, comunidades, o cualquier otro agrupamiento conveniente de casas. Como ya hemos visto, el tamaño del límite para el error de estimación depende crucialmente de la variación entre los totales de conglomerados. Entonces, al intentar obtener límites pequeños para el error de estimación, debemos seleccionar conglomerados con la menor variación posible entre estos totales. Ahora vamos a suponer que el tamaño del conglomerado (unidad de muestreo) ha sido elegido y vamos a considerar únicamente el problema de seleccionar el número de conglomerados, n .

De la Ecuación (8.2), la varianza estimada de \bar{y} es

$$\hat{V}(\bar{y}) = \frac{N-n}{NnM^2} (\hat{s}_c^2)$$

donde

$$\hat{s}_c^2 = \frac{\sum_{i=1}^n (y_i - \bar{y}m_i)^2}{n-1} \quad (8.11)$$

La varianza real de \bar{y} es aproximadamente

$$V(\bar{y}) = \frac{N-n}{NnM^2} (\sigma_c^2) \quad (8.12)$$

donde σ_c^2 es la cantidad poblacional estimada por \hat{s}_c^2 .

Debido a que no conocemos σ_c^2 o el tamaño promedio \bar{M} del conglomerado, la elección del tamaño de muestra, esto es, el número de conglomerados necesario para comprar una cantidad especificada de información concerniente a un parámetro poblacional, es complicada. Eliminamos esta dificultad utilizando el mismo método usado para la estimación de razón. Esto es, usamos un estimador de σ_c^2 y \bar{M} disponibles de una encuesta previa, o seleccionamos una muestra preliminar de n' elementos. Las estimaciones de σ_c^2 y \bar{M} pueden calcularse de la muestra preliminar y utilizarse para obtener un tamaño de muestra total aproximado n . Entonces, como en todos los problemas de selección de un tamaño de muestra, igualamos dos desviaciones estándar de nuestro estimador, con un límite para el error de estimación B . Este límite es elegido por el experimentador y representa el máximo error que deseé tolerar. Esto es

$$2\sqrt{V(\bar{y})} = B$$

Usando la Ecuación (8.12), podemos despejar n .

Obtenemos resultados similares cuando usamos $M\bar{y}$ para estimar el total poblacional τ , porque $V(M\bar{y}) = M^2 V(\bar{y})$.

Tamaño de muestra aproximado requerido para estimar μ con un límite B para el error de estimación:

$$n = \frac{N\sigma_c^2}{ND + \sigma_c^2} \quad (8.13)$$

donde σ_c^2 es estimado por s_c^2 y

$$D = \frac{B^2 \bar{M}^2}{4}$$

EJEMPLO 8.6

Supóngase que los datos de la Tabla 8.1 representan una muestra preliminar de ingresos en la ciudad. ¿Qué tan grande debe tomarse la muestra en una encuesta futura para estimar el ingreso promedio por persona μ con un límite de \$500 para el error de estimación?

SOLUCIÓN

Para utilizar la Ecuación (8.13), debemos estimar σ_c^2 ; el mejor estimador disponible es s_c^2 , el cual puede ser calculado mediante el uso de los datos de la Tabla 8.1. Usando los cálculos del Ejemplo 8.2, tenemos

$$s_c^2 = \frac{\sum_{i=1}^n (y_i - \bar{y}_m)^2}{n-1} = \frac{15,227,502,247}{24} = 634,479,260$$

La cantidad \bar{M} puede ser estimada por $\bar{m} = 6.04$ calculada con los datos de la Tabla 8.1. Entonces D es aproximadamente

$$\frac{B^2 \bar{M}^2}{4} = \frac{(500)^2 (6.04)^2}{4} = (62,500)(6.04)^2$$

Usando la Ecuación (8.13) tenemos

$$n = \frac{N\sigma_c^2}{ND + \sigma_c^2} = \frac{415(634,479,260)}{415(6.04)^2(62,500) + 634,479,260} = 166.58$$

Entonces se deben muestrear 167 conglomerados.

Tamaño de muestra aproximado requerido para estimar τ , usando $M\bar{y}$, con un límite B para el error de estimación:

$$n = \frac{N\sigma_c^2}{ND + \sigma_c^2} \quad (8.14)$$

donde σ_c^2 es estimada por s_c^2 y

$$D = \frac{B^2}{4N^2}$$

EJEMPLO 8.7

Usando nuevamente los datos de la Tabla 8.1 como una muestra preliminar de ingresos en la ciudad, señale ¿qué tan grande se necesita una muestra para estimar el ingreso total de todos los residentes, τ , con un límite de \$1,000,000 para el error de estimación? Hay 2500 residentes en la ciudad ($M = 2500$)

SOLUCIÓN

Usamos la Ecuación (8.14) y estimamos σ_c^2 mediante

$$s_c^2 = 634,479,260$$

como en el Ejemplo 8.6. Cuando estimamos τ , usamos

$$D = \frac{B^2}{4N^2} = \frac{(1,000,000)^2}{4(415)^2}$$

$$ND = \frac{(1,000,000)^2}{4(415)} = 602,409,000$$

Entonces, usando la Ecuación (8.14) nos da

$$n = \frac{N\sigma_c^2}{ND + \sigma_c^2} = \frac{415(634,479,260)}{602,409,000 + 634,479,260} = 212.88$$

Luego se deben muestrear 213 conglomerados para estimar el ingreso total con un límite de \$1,000,000 para el error de estimación.

El estimador $N\bar{y}_t$, que se muestra en la Ecuación (8.8), se usa para estimar τ cuando M es desconocido. La varianza estimada de $N\bar{y}_t$ que se muestra en la Ecuación (8.9), es

$$\hat{V}(N\bar{y}_t) = N^2 \left(\frac{N-n}{Nn} \right) s_t^2$$

$$s_t^2 = \frac{\sum_{i=1}^n (\bar{y}_i - \bar{y}_t)^2}{n-1} \quad (8.15)$$

donde

Entonces la varianza poblacional de $N\bar{y}_t$ es

$$V(N\bar{y}_t) = N^2 V(\bar{y}_t) = N^2 \left(\frac{N-n}{Nn} \right) \sigma_t^2 \quad (8.16)$$

donde σ_t^2 es la cantidad poblacional estimada por s_t^2 .

La estimación de τ con un límite de B unidades para el error de estimación nos lleva a la siguiente ecuación:

$$2\sqrt{V(N\bar{y}_t)} = B$$

Usando la Ecuación (8.16), podemos despejar n .

Tamaño de muestra aproximado requerido para estimar τ , usando $N\bar{y}_t$ con un límite B para el error de estimación:

$$n = \frac{N\sigma_t^2}{ND + \sigma_t^2} \quad (8.17)$$

donde σ_t^2 se estima mediante s_t^2 , y

$$D = \frac{B^2}{4N^2}$$

EJEMPLO 8.8

Supóngase que los datos de la Tabla 8.1 provienen de un estudio preliminar de ingresos en la ciudad y que no se conoce M . ¿Qué tan grande se debe tomar la muestra para estimar el ingreso total de todos los residentes, τ , con un límite de \$1,000,000 para el error de estimación?

SOLUCIÓN

La cantidad σ_t^2 debe ser estimada por s_t^2 , que se calcula con los datos de la Tabla 8.1. Usando los cálculos del Ejemplo 8.4 nos da

$$s_t^2 = \frac{\sum_{i=1}^n (\bar{y}_i - \bar{y}_t)^2}{n-1} = \frac{11,389,360,000}{24} = 474,556,667$$

El límite para el error de estimación es $B = \$1,000,000$. Por lo que

$$D = \frac{B^2}{4N^2} = \frac{(1,000,000)^2}{4(415)^2}$$

De la Ecuación (8.17)

$$n = \frac{N\sigma_t^2}{ND + \sigma_t^2} = \frac{415(474,556,667)}{415(1,000,000)^2/4(415)^2 + 474,556,667} = 182.88$$

Entonces se debe tomar una muestra de 183 conglomerados para tener un límite de \$1,000,000 en el error de estimación.

8.5 ESTIMACIÓN DE UNA PROPORCIÓN POBLACIONAL

Supóngase que un experimentador desea estimar una proporción poblacional, o fracción, tal como la proporción de casas en un estado con inadecuado servicio de plomería, o la proporción de presidentes de corporación que son universitarios graduados. El mejor estimador de la proporción poblacional p es la proporción muestral \hat{p} . Sea a_i el número total de elementos en el conglomerado i que poseen la característica de interés. Entonces, la proporción de elementos en la muestra de n conglomerados que poseen la característica de interés es dada por

$$\hat{p} = \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n m_i}$$

donde m_i es el número de elementos en el i -ésimo conglomerado, $i = 1, 2, \dots, n$. Nótese que \hat{p} tiene la misma forma de \bar{y} [véase Ecuación (8.1)], excepto que y_i es reemplazado por a_i . La varianza estimada de \hat{p} es similar a la de \bar{y} .

Estimador de la proporción poblacional p :

$$\hat{p} = \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n m_i} \quad (8.18)$$

Varianza estimada de \hat{p} :

$$\hat{V}(\hat{p}) = \left(\frac{N-n}{NnM^2} \right) \frac{\sum_{i=1}^n (a_i - \hat{p}m_i)^2}{n-1} \quad (8.19)$$

Límite para el error de estimación:

$$2\sqrt{\hat{V}(\hat{p})} = 2 \sqrt{\left(\frac{N-n}{Nn\bar{M}^2} \right) \frac{\sum_{i=1}^n (a_i - \hat{p}m_i)^2}{n-1}} \quad (8.20)$$

La fórmula de varianza (8.19), es un buen estimador únicamente cuando la muestra de tamaño n es grande, digamos $n \geq 20$. Si $m_1 = m_2 = \dots = m_N$, entonces \hat{p} es un estimador insesgado de p , y la $\hat{V}(\hat{p})$, que se muestra en la Ecuación (8.19) es un estimador insesgado de la varianza real de \hat{p} para cualquier tamaño de muestra.

EJEMPLO 8.9

Además de la pregunta sobre su ingreso, se interroga a los residentes, de la encuesta muestral del Ejemplo 8.2, acerca de si son dueños o alquilan la casa donde viven. Los resultados se presentan en la Tabla 8.2. Utilice los datos de la tabla 8.2 para estimar la proporción de residentes que viven en casas de alquiler. Establezca un límite para el error de estimación.

TABLA 8.2 Número de arrendatarios

Conglomerado	Número de residentes, m_i	Número de arrendatarios a_i	Conglomerado	Número de residentes m_i	Número de arrendatarios a_i
1	8	4	14	10	5
2	12	7	15	9	4
3	4	1	16	3	1
4	5	3	17	6	4
5	6	3	18	5	2
6	6	4	19	5	3
7	7	4	20	4	1
8	5	2	21	6	3
9	8	3	22	8	3
10	3	2	23	7	4
11	2	1	24	3	0
12	6	3	25	8	3
13	5	2			

$$\sum_{i=1}^{25} m_i = 151 \quad \sum_{i=1}^{25} a_i = 72$$

$$\sum_{i=1}^{25} a_i^2 = 262 \quad \sum_{i=1}^{25} m_i^2 = 1047 \quad \sum_{i=1}^{25} a_i m_i = 511$$

SOLUCIÓN

El mejor estimador de la proporción poblacional de arrendatarios es \hat{p} , que se muestra en la Ecuación (8.18), donde

$$\hat{p} = \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n m_i} = \frac{72}{151} = 0.48$$

Para estimar la varianza de \hat{p} , debemos calcular

$$\sum_{i=1}^n (a_i - \hat{p}m_i)^2 = \sum_{i=1}^n a_i^2 - 2\hat{p} \sum_{i=1}^n a_i m_i + \hat{p}^2 \sum_{i=1}^n m_i^2$$

y de la Tabla 8.2

$$\sum_{i=1}^n (a_i - \hat{p}m_i)^2 = 262 - 2(0.477)(511) + (0.477)^2(1047) = 12.729$$

La cantidad \bar{M} es estimada por \bar{m} , donde

$$\bar{m} = \frac{\sum_{i=1}^n m_i}{n} = \frac{151}{25} = 6.04$$

Entonces, de la Ecuación (8.19),

$$\begin{aligned} \hat{V}(\hat{p}) &= \left(\frac{N-n}{Nn\bar{M}^2} \right) \frac{\sum_{i=1}^n (a_i - \hat{p}m_i)^2}{n-1} \\ &= \frac{(415-25)(12.729)}{415(25)(6.04)^2(24)} = 0.00055 \end{aligned}$$

La estimación de p con un límite para el error de estimación es

$$\hat{p} \pm 2\sqrt{\hat{V}(\hat{p})}, \quad \text{o sea} \quad 0.48 \pm 2\sqrt{0.00055}, \quad \text{o sea} \quad 0.48 \pm 0.05$$

Entonces la mejor estimación de la proporción de personas que alquilan casa es 0.48. El error de estimación debe ser menor que 0.05 con probabilidad de aproximadamente 0.95.

8.6 SELECCIÓN DEL TAMAÑO DE MUESTRA PARA LA ESTIMACIÓN DE PROPORCIONES

La estimación de la proporción poblacional p , con un límite de B unidades para el error de estimación, implica que el experimentador quiere

$$2\sqrt{\hat{V}(\hat{p})} = B$$

Esta ecuación puede ser resuelta para n , y la solución es similar a la Ecuación (8.13). Esto es

$$n = \frac{N\sigma_c^2}{ND + \sigma_c^2}$$

donde $D = B^2\bar{M}^2/4$, y σ_c^2 se estima por

$$\sigma_c^2 = \frac{\sum_{i=1}^n (a_i - \hat{p}m_i)^2}{n - 1} \quad (8.21)$$

La Ecuación (8.21) es la misma que la (8.11) con γ_i reemplazada por a_i y \bar{y} por \hat{p} .

EJEMPLO 8.10

Los datos en la Tabla 8.2 son obsoletos. Se va a realizar un nuevo estudio en la misma ciudad con el propósito de estimar la proporción p de residentes que alquilan la casa en que viven. ¿Qué tan grande se debe tomar la muestra para estimar p , con un límite de 0.04 en el error de estimación?

SOLUCIÓN

El mejor estimador de σ_c^2 es s_c^2 , el cual es calculado usando los datos de la Tabla 8.2:

$$s_c^2 = \frac{\sum_{i=1}^n (a_i - \hat{p}m_i)^2}{n - 1} = \frac{12.729}{24} = 0.530$$

La cantidad \bar{M} es estimada por $\bar{m} = 6.04$. También D es aproximada por

$$\frac{B^2\bar{m}^2}{4} = \frac{(0.04)^2(6.04)^2}{4} = 0.0146$$

Entonces

$$n = \frac{N\sigma_c^2}{ND + \sigma_c^2} = \frac{(415)(0.530)}{(415)(0.0146) + 0.530} = 33.40$$

De modo que se deben muestrear 34 conglomerados para estimar p , con un límite de 0.04 para el error de estimación.

8.7 MUESTREO POR CONGLOMERADOS COMBINADO CON ESTRATIFICACIÓN

Así como en el caso de todos los demás métodos de muestreo, el muestreo por conglomerados puede ser combinado con muestreo estratificado, con objeto de que la pobla-

ción pueda ser dividida en L estratos y se pueda seleccionar entonces una muestra por conglomerados en cada estrato.

Recuérdese que la Ecuación (8.1) tiene la forma de un estimador de razón y puede ser considerada como la razón de un estimador del promedio de totales de conglomerados, con respecto al estimador del tamaño promedio de conglomerados. Entonces, pensando en términos de un estimador de razón, tenemos dos modos para formar el estimador de una media poblacional a través de los estratos: el estimador separado y el estimador combinado. Un poco de investigación nos mostrará que si se emplea el estimador separado, se debe conocer el número total de elementos en cada estrato para poder asignar las ponderaciones adecuadas por estrato. Ya que estas cantidades son comúnmente desconocidas, únicamente analizaremos la forma combinada del estimador de razón en el contexto de muestreo por conglomerados.

En lugar de presentar fórmulas generales que parezcan formidables, vamos a ilustrar la técnica con un ejemplo numérico.

EJEMPLO 8.11

Consideremos los datos de la Tabla 8.1 como la muestra del estrato 1, con $N_1 = 415$ y $n_1 = 25$, como en el Ejemplo 8.2. Se toma una ciudad vecina más pequeña como el estrato 2. Para el estrato 2, $n_2 = 10$ bloques se van a muestrear de $N_2 = 168$. Estime el ingreso promedio por persona en las dos ciudades combinadas, y establezca un límite para el error de estimación, dados los datos adicionales que se muestran en la tabla anexa.

Conglomerado <i>i</i>	Número de residentes, <i>m_i</i>	Ingreso total por conglomerado, <i>y_i</i>
1	2	\$ 18,000
2	5	52,000
3	7	68,000
4	4	36,000
5	3	45,000
6	8	96,000
7	6	64,000
8	10	115,000
9	3	41,000
10	1	12,000

SOLUCIÓN

El promedio de los totales de conglomerados en las respectivas muestras son $\bar{y}_{11} = 53,160$ y $\bar{y}_{12} = 54,700$. El promedio de los tamaños de los conglomerados en las respectivas muestras es $\bar{m}_1 = 6.04$ y $\bar{m}_2 = 4.90$. El estimador del promedio poblacional del total por conglomerado es entonces

$$\frac{1}{N} (N_1\bar{y}_{11} + N_2\bar{y}_{12})$$

mientras que el estimador del promedio del tamaño de conglomerados es

$$\frac{1}{N} (N_1 \bar{m}_1 + N_2 \bar{m}_2)$$

Un estimador de la media poblacional por elemento es entonces

$$\bar{y}^* = \frac{N_1 \bar{y}_{t1} + N_2 \bar{y}_{t2}}{N_1 \bar{m}_1 + N_2 \bar{m}_2}$$

y esta ecuación tiene la forma de un estimador de razón combinada. Análogamente a la varianza usada en la Sección 6.6, la varianza de \bar{y}^* puede ser estimada por

$$\hat{V}(\bar{y}^*) = \frac{1}{M^2} \left\{ \frac{N_1(N_1 - n_1)}{n_1(n_1 - 1)} \sum_{i=1}^{n_1} [(\bar{y}_i - \bar{y}_{t1}) - \bar{y}^*(\bar{m}_i - \bar{m}_1)]^2 + \frac{N_2(N_2 - n_2)}{n_2(n_2 - 1)} \sum_{i=1}^{n_2} [(\bar{y}_i - \bar{y}_{t2}) - \bar{y}^*(\bar{m}_i - \bar{m}_2)]^2 \right\}$$

donde M es el número total de elementos en la población y puede ser estimado por $N_1 \bar{m}_1 + N_2 \bar{m}_2$ si no es conocido. La primera suma en la expresión de la varianza es sobre todas las observaciones de la muestra en el estrato 1, y la segunda suma es sobre todas las observaciones del estrato 2.

Para los datos presentados en la tabla,

$$\bar{y}^* = \frac{415(53,160) + 168(54,700)}{415(6.04) + 168(4.90)} = 9385$$

Para el estrato 1

$$\left(\frac{1}{n_1 - 1} \right) \sum_{i=1}^{n_1} [(\bar{y}_i - \bar{y}_{t1}) - \bar{y}^*(\bar{m}_i - \bar{m}_1)]^2 = 675,930,246$$

y para el estrato 2

$$\left(\frac{1}{n_2 - 1} \right) \sum_{i=1}^{n_2} [(\bar{y}_i - \bar{y}_{t2}) - \bar{y}^*(\bar{m}_i - \bar{m}_2)]^2 = 74,934,600$$

Ya que

$$N_1 \bar{m}_1 + N_2 \bar{m}_2 = 3329.8$$

por lo que

$$\hat{V}(\bar{y}^*) = 412,563.8$$

y

$$2\sqrt{\hat{V}(\bar{y}^*)} = 1285$$

Entonces, el ingreso promedio por persona para las dos ciudades combinadas es

$$\$9385 \pm \$1285$$

Vemos que el límite para el error de estimación es un poco más pequeño que el límite para el estrato 1, como se encontró en el Ejemplo 8.2.

8.8 MUESTREO POR CONGLOMERADOS CON PROBABILIDADES PROPORCIONALES AL TAMAÑO

En la Sección 4.6 vimos que algunas veces es posible reducir la varianza de un estimador mediante el muestreo de unidades con probabilidades proporcionales a una medida del tamaño de la unidad. El muestreo por conglomerados suele proporcionar una situación ideal para el uso de muestreo con ppt, ya que el número de elementos en un conglomerado, m_i , representa una medida natural del tamaño del conglomerado. El muestreo con probabilidades proporcionales a m_i paga grandes dividendos en términos de la reducción del límite para el error de estimación, cuando el total del conglomerado y_i está altamente correlacionado con el número de elementos en el conglomerado, lo cual ocurre frecuentemente.

En la notación de la sección 4.6, sea π_i la probabilidad de que la i -ésima unidad de muestreo aparezca en la muestra, la cual es dada por

$$\pi_i = \frac{m_i}{M} \quad (8.22)$$

Entonces, el estimador de un total poblacional $\hat{\tau}_{ppt}$ toma la forma [véase la Ecuación (4.20)]

$$\begin{aligned} \hat{\tau}_{ppt} &= \frac{1}{n} \sum_{i=1}^n \frac{y_i}{\pi_i} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{(m_i/M)} \\ &= \frac{M}{n} \sum_{i=1}^n \frac{y_i}{m_i} = \frac{M}{n} \sum_{i=1}^n \bar{y}_i \end{aligned}$$

donde \bar{y}_i es el promedio de las observaciones en el i -ésimo conglomerado. La varianza estimada de $\hat{\tau}_{ppt}$ tiene una forma particularmente simple, como se verá después.

Ya que ahora hay M elementos en la población, el estimador de la media poblacional, $\hat{\mu}_{ppt}$, es simplemente

$$\hat{\mu}_{ppt} = \frac{1}{M} \hat{\tau}_{ppt} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i$$

La varianza estimada de $\hat{\mu}_{ppt}$ es también fácil de calcular.

Estimador de la media poblacional μ :

$$\hat{\mu}_{ppt} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i \quad (8.23)$$

donde \bar{y}_i es la media del i -ésimo conglomerado.

Varianza estimada de $\hat{\mu}_{ppt}$:

$$\hat{V}(\hat{\mu}_{ppt}) = \frac{1}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \hat{\mu}_{ppt})^2 \quad (8.24)$$

Límite para el error de estimación:

$$2\sqrt{\hat{V}(\hat{\mu}_{ppt})} = 2\sqrt{\frac{1}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \hat{\mu}_{ppt})^2} \quad (8.25)$$

Estimador del total poblacional τ :

$$\hat{\tau}_{\text{ppt}} = \frac{M}{n} \sum_{i=1}^n \bar{y}_i \quad (8.26)$$

Varianza estimada de $\hat{\tau}_{\text{ppt}}$:

$$\hat{V}(\hat{\tau}_{\text{ppt}}) = \frac{M^2}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \hat{\mu}_{\text{ppt}})^2 \quad (8.27)$$

Límite para el error de estimación:

$$2\sqrt{\hat{V}(\hat{\tau}_{\text{ppt}})} = 2\sqrt{\frac{M^2}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \hat{\mu}_{\text{ppt}})^2} \quad (8.28)$$

Ahora ilustramos la técnica de muestreo con probabilidades proporcionales al tamaño de los conglomerados y el uso de las fórmulas —ya presentadas— en los siguientes ejemplos.

EJEMPLO 8.12

Un auditor desea muestrear los registros de ausencias por enfermedad de una gran empresa, para estimar el número promedio de días de ausencia por enfermedad por empleado en el cuatrimestre pasado. La empresa tiene ocho divisiones, con diferentes números de empleados por división. Ya que el número de días de ausencia por enfermedad dentro de cada división debe estar altamente correlacionado con el número de empleados, el auditor decide muestrear $n = 3$ divisiones con probabilidad proporcional al número de empleados. Muestre cómo seleccionar la muestra si los respectivos números de empleados son 1200, 450, 2100, 860, 2840, 1910, 290, 3200.

SOLUCIÓN

Primero listamos el número de empleados y el intervalo acumulado para cada división, como sigue:

División	Número de empleados	Intervalo acumulado
1	1,200	1-1200
2	450	1201-1650
3	2,100	1651-3750
4	860	3751-4610
5	2,840	4611-7450
6	1,910	7451-9360
7	390	9361-9750
8	3,200	9751-12,950
		12,950

Ya que se van a muestrear $n = 3$ divisiones, debemos seleccionar tres números aleatorios entre 00001 y 12,500. Podemos hacer esta selección empezando en cualquier lugar de la tabla de números aleatorios y seleccionando números de cinco dígitos, pero nosotros elegimos empezar en la línea 1, columna 4 de la Tabla 2 del Apéndice. Los primeros tres números entre 00001 y 12,950 que aparecen al dirigirnos hacia abajo en la columna son, 02011, 07972, y 10281. El primero aparece en el intervalo acumulado de la división 3, el segundo aparece en el intervalo de la división 6 y el tercero aparece en el intervalo de la división 8. Entonces las divisiones 3, 6 y 8 constituyen la muestra. (Nótese que una división puede ser seleccionada más de una vez. En tal caso el dato resultante se trata como dos valores muestrales separados pero iguales.)

EJEMPLO 8.13

Supóngase que el número total de días de ausencia por enfermedad registrados en las tres divisiones muestreadas durante el cuatrimestre pasado son, respectivamente,

$$y_1 = 4320, \quad y_2 = 4160, \quad y_3 = 5790$$

Estime el número promedio de días de ausencia por enfermedad requeridos por persona, de toda la empresa, y establezca un límite para el error de estimación.

SOLUCIÓN

Primero debemos calcular las medias de los conglomerados muestreados, las cuales son

$$\bar{y}_1 = \frac{4320}{2100} = 2.06, \quad \bar{y}_2 = \frac{4160}{1910} = 2.18, \quad \bar{y}_3 = \frac{5790}{3200} = 1.81$$

(Nótese que los números de empleados para los conglomerados muestreados provienen de los datos del Ejemplo 8.12)

Ahora por la Ecuación (8.23)

$$\hat{\mu}_{\text{ppt}} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i = \frac{1}{3} (2.06 + 2.18 + 1.81) = 2.02$$

También, por la Ecuación (8.24)

$$\begin{aligned} \hat{V}(\hat{\mu}_{\text{ppt}}) &= \frac{1}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \hat{\mu}_{\text{ppt}})^2 \\ &= \frac{1}{3(2)} [(2.06 - 2.02)^2 + (2.18 - 2.02)^2 + (1.81 - 2.02)^2] \\ &= 0.0119 \end{aligned}$$

Entonces el límite para el error de estimación es

$$2\sqrt{0.0119} = 0.22$$

Nuestra estimación del número promedio de días de ausencia por enfermedad utilizados por los empleados de la empresa es

$$2.02 \pm 0.22$$

Ahora tenemos tres estimadores del total poblacional en muestreo por conglomerados: el estimador de razón (8.4), el estimador insesgado (8.8) y el estimador ppt (8.26). ¿Cómo sabemos cuál es el mejor? Ahora presentamos algunas pautas acerca de cómo contestar esta pregunta: si y_i no está correlacionado con m_i , entonces el estimador insesgado es mejor que cualquiera de los otros dos. Si y_i está correlacionado con m_i , entonces el estimador de razón y el ppt son más precisos que el estimador insesgado. El estimador ppt es mejor que el estimador de razón si la variación dentro del conglomerado no cambia con un sesgo en m_i . El estimador de razón es mejor que el estimador ppt si la variación dentro del conglomerado se incrementa con el aumento en m_i .

En los Ejemplos 8.12 y 8.13, el número de días de ausencia por enfermedad utilizados debe incrementarse con el número de empleados. Entonces, el estimador insesgado es aquí una elección ineficaz. Pero la variación de días de ausencia por enfermedad dentro de las divisiones puede permanecer relativamente constante a través de las divisiones. En tal caso, el estimador ppt es la mejor elección.

8.9 RESUMEN

Este capítulo introduce un tercer diseño de encuestas por muestreo. En este diseño cada unidad de muestreo es un grupo, o conglomerado de elementos. El muestreo por conglomerados puede proporcionar la máxima información al mínimo costo cuando no se tiene un marco que liste los elementos de la población o cuando el costo por obtener observaciones se incrementa con la distancia entre los elementos.

El estimador de la media poblacional μ es la media muestral \bar{y} , dada por la Ecuación (8.1). La varianza estimada de \bar{y} es dada por la Ecuación (8.2). Se presentan dos estimadores del total poblacional con sus respectivas varianzas estimadas. Se presenta el estimador $M\bar{y}$ en la Ecuación (8.4); el cual se usa cuando se conoce el número de elementos M en la población. El estimador $N\bar{y}$, [véase la Ecuación (8.8)] se usa cuando no se conoce M .

En la Sección 8.4 se estudió un tamaño de muestra apropiado para estimar μ o τ con un límite especificado para el error de estimación.

En muestreo por conglomerados el estimador de una proporción poblacional p es la proporción muestral \hat{p} , dada por la Ecuación (8.18). La varianza estimada de \hat{p} se presenta en la Ecuación (8.19). El problema de la selección de un tamaño de muestra para estimar una proporción es similar al problema de la estimación de una media.

El muestreo por conglomerados se puede usar también dentro de los estratos en una población estratificada, y se presentó un ejemplo en la Sección 8.7.

ANÁLISIS DEL ESTUDIO DE CASO

PROBLEMA DE LAS CARACTERÍSTICAS DEL BARRIO

Al principio de este capítulo se sugirió el uso de los datos de la Oficina de Censos sobre estadísticas de manzana para estimar la proporción de residentes con una edad mayor o igual a 65 años en un área de 40 manzanas. Las $n = 5$ manzanas fueron muestreadas aleatoriamente de las 40 y se obtuvieron los siguientes datos:

Número de residentes, m_i	Personas con 65 años o más, a_i	$\hat{p}m_i$	$a_i - \hat{p}m_i$	$(a_i - \hat{p}m_i)^2$
90	15	21.60	-6.60	43.5600
32	8	7.68	0.32	0.1024
47	14	11.28	2.72	7.3984
25	9	6.00	3.00	9.0000
16	4	3.84	0.16	0.0256
210	50			60.0864

$$\hat{p} = \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n m_i} = \frac{50}{210} = 0.24$$

Así que la mejor estimación de la proporción de personas con edad igual o mayor a 65 años es 0.24.

El límite para el error de estimación es

$$\begin{aligned} 2\sqrt{\hat{V}(\hat{p})} &= 2\sqrt{\left(\frac{N-n}{Nn\bar{m}^2}\right)\left(\frac{1}{n-1}\right)\sum_{i=1}^n (a_i - pm_i)^2} \\ &= 2\sqrt{\left[\frac{35}{(40)(5)(42)^2}\right]\left(\frac{1}{4}\right)(60.0864)} \\ &= 0.08 \end{aligned}$$

Entonces la estimación de la proporción verdadera para el área de 40 manzanas es 0.24 ± 0.08 o bien 0.16 a 0.32. Tenemos confianza en que más del 16% de los residentes tiene una edad igual o mayor a 65 años.

EJERCICIOS

- 8.1 Una experimentadora que trabaja en un área urbana desea estimar el valor promedio de una variable altamente correlacionada con raza. Ella piensa que debe usar muestreo por conglomerados, con manzanas como conglomerados y adultos dentro de manzanas como elementos.

Explique por qué se debería o no usar muestreo por conglomerados en cada una de las siguientes situaciones.

- La mayoría de los adultos en ciertas manzanas son blancos y la mayoría son no blancos en otras manzanas.
- La proporción de no blancos es la misma en cada bloque y no está cercana a 1 o a 0.
- La proporción de no blancos difiere de manzana a manzana en la manera que se podría esperar si los conglomerados fueran hechos asignando aleatoriamente los adultos de la población a los conglomerados.

- 8.2** Un fabricante de sierras de cinta quiere estimar el costo de reparación promedio mensual para las sierras que ha vendido a ciertas industrias. El fabricante no puede obtener un costo de reparación para cada sierra, pero puede obtener la cantidad total gastada en reparación y el número de sierras que tiene cada industria. Entonces decide usar muestreo por conglomerados, con cada industria como un conglomerado. El fabricante selecciona una muestra irrestricta aleatoria de $n = 20$ de $N = 96$ industrias a las que da servicio. Los datos sobre costo total de reparaciones por industria y el número de sierras por industria se presentan en la tabla anexa. Estime el costo promedio de reparación por sierra para el mes pasado, y establezca un límite para el error de estimación.

Industria	Número de sierras	Costo total de reparación para el mes pasado (en dólares)	Industria	Número de sierras	Costo total de reparación para el mes pasado (en dólares)
1	3	50	11	8	140
2	7	110	12	6	130
3	11	230	13	3	70
4	9	140	14	2	50
5	2	60	15	1	10
6	12	280	16	4	60
7	14	240	17	12	280
8	3	45	18	6	150
9	5	60	19	5	110
10	9	230	20	8	120

- 8.3** Para los datos en el Ejercicio 8.2, estime la cantidad total gastada por las 96 industrias en la reparación de sierras. Establezca un límite para el error de estimación.

- 8.4** Después de verificar sus registros de ventas, el fabricante del Ejercicio 8.2 se percata de que ha vendido un total de 710 sierras a esas industrias. Usando esta información adicional, estime la cantidad total gastada en reparación de sierras por estas industrias, y establezca un límite para el error de estimación.

- 8.5** El mismo fabricante (Ejercicio 8.2) quiere estimar el costo de reparación promedio por sierra para el mes siguiente. ¿Cuántos conglomerados debe seleccionar en la muestra si quiere que el límite para el error de estimación sea menor que \$2.00?

- 8.6** Un politólogo desarrolla una prueba para medir el grado de conocimiento sobre acontecimientos actuales. Él quiere estimar la calificación promedio que obtendrán en su prueba todos los estudiantes de una escuela preparatoria. La administración de la escuela no le permitirá seleccionar aleatoriamente a los estudiantes fuera de clases, pero si interrumpir un pequeño número de clases con el propósito de aplicar la prueba a cada miembro de la clase. Entonces el experimentador selecciona al azar 25 clases de un total de 108 a una hora determinada. Se aplica la prueba a cada miembro de las clases muestreadas, con los resultados que se presentan en la tabla anexa.

Estime la calificación promedio que sería obtenida para esta prueba por todos los estudiantes en la escuela. Establezca un límite para el error de estimación.

Clase	Número de estudiantes	Calificación total	Clase	Número de estudiantes	Calificación total
1	31	1590	14	40	1980
2	29	1510	15	38	1990
3	25	1490	16	28	1420
4	35	1610	17	17	900
5	15	800	18	22	1080
6	31	1720	19	41	2010
7	22	1310	20	32	1740
8	27	1427	21	35	1750
9	25	1290	22	19	890
10	19	860	23	29	1470
11	30	1620	24	18	910
12	18	710	25	31	1740
13	21	1140			

- 8.7** El politólogo del Ejercicio 8.6 quiere estimar la calificación promedio en la prueba para una escuela preparatoria similar. Él quiere que el límite para el error de estimación sea menor que 2 puntos. ¿Cuántas clases debe tomar en la muestra? Supóngase que la escuela tiene 100 clases durante cada hora en este periodo escolar.

- 8.8** Una industria está considerando la revisión de su política de jubilación y quiere estimar la proporción de empleados que apoyan la nueva política. La industria consiste de 87 plantas separadas localizadas en todo Estados Unidos. Ya que los resultados deben ser obtenidos rápidamente y con poco dinero, la industria decide usar muestreo por conglomerados, con cada planta como un conglomerado. Se selecciona una muestra irrestricta aleatoria de 15 plantas y se obtienen las opiniones de los empleados en estas plantas a través de un cuestionario. Los resultados se presentan en la tabla anexa. Estime la proporción de empleados en la industria que apoyan la nueva política de jubilación y establezca un límite para el error de estimación.

Planta	Número de empleados	Número de empleados que apoyan la nueva política	Planta	Número de empleados	Número de empleados que apoyan la nueva política
1	51	42	9	73	54
2	62	53	10	61	45
3	49	40	11	58	51
4	73	45	12	52	29
5	101	63	13	65	46
6	48	31	14	49	37
7	65	38	15	55	42
8	49	30			

- 8.9** La industria del Ejercicio 8.8 modificó su política de jubilación después de obtener los resultados de la encuesta. Ahora se quiere estimar la proporción de empleados a favor de la política modificada. ¿Cuántas plantas deben ser muestreadas para tener un límite de 0.08 para el error de estimación? Use los datos del Ejercicio 8.8 para aproximar los resultados de la nueva encuesta.

- 8.10 Se diseña una encuesta económica para estimar la cantidad promedio gastada en servicios para el hogar en una ciudad. Ya que no se encuentra disponible una lista de hogares, se usa muestreo por conglomerados, con divisiones (barrios) formando los conglomerados. Se selecciona una muestra aleatoria de 20 barrios de la ciudad de un total de 60. Los entrevistadores obtienen el costo de los servicios de cada hogar dentro de los barrios seleccionados; los costos totales se muestran en la tabla anexa. Estime la cantidad promedio de gastos en servicios por hogar en la ciudad y establezca un límite para el error de estimación.

Barrio muestreado	Número de hogares	Cantidad total gastada en servicios	Barrio muestreado	Número de hogares	Cantidad total gastada en servicios
1	55	\$2210	11	73	\$2930
2	60	2390	12	64	2470
3	63	2430	13	69	2830
4	58	2380	14	58	2370
5	71	2760	15	63	2390
6	78	3110	16	75	2870
7	69	2780	17	78	3210
8	58	2370	18	51	2430
9	52	1990	19	67	2730
10	71	2810	20	70	2880

- 8.11 En la encuesta del Ejercicio 8.10 se desconoce el número de hogares en la ciudad. Estime la cantidad total gastada en servicios por todos los hogares de la ciudad y establezca un límite para el error de estimación.
- 8.12 La encuesta económica del Ejercicio 8.10 se va a llevar a cabo en una ciudad vecina de estructura similar. El objetivo es estimar la cantidad total gastada en servicios por los hogares de la ciudad, con un límite de \$ 5000 para el error de estimación. Use los datos del Ejercicio 8.10 para encontrar el número aproximado de conglomerados que se necesitan para obtener este límite.
- 8.13 Un inspector quiere estimar el peso promedio de llenado para cajas de cereal empacadas en una fábrica. El cereal está en paquetes que contienen 12 cajas cada uno. El inspector selecciona aleatoriamente 5 y mide el peso de llenado de cada caja en los paquetes muestreados, con los resultados (en onzas) que se muestran en la tabla anexa. Estime el peso promedio de llenado para las cajas empacadas por esta fábrica, y establezca un límite para el error de estimación. Suponga que el número total de cajas empacadas por la fábrica es lo suficientemente grande para que no se tome en cuenta la corrección por población finita.

Paquete	Onzas de llenado											
1	16.1	15.9	16.1	16.2	15.9	15.8	16.1	16.2	16.0	15.9	15.8	16.0
2	15.9	16.2	15.8	16.0	16.3	16.1	15.8	15.9	16.0	16.1	16.1	15.9
3	16.2	16.0	15.7	16.3	15.8	16.0	15.9	16.0	16.1	16.0	15.9	16.1
4	15.9	16.1	16.2	16.1	16.1	16.3	15.9	16.1	15.9	15.9	16.0	16.0
5	16.0	15.8	16.3	15.7	16.1	15.9	16.0	16.1	15.8	16.0	16.1	15.9

- 8.14 Un periódico quiere estimar la proporción de votantes que apoyan a cierto candidato, candidato A, en una elección estatal. Ya que la selección y entrevista de una muestra irrestricta aleatoria de votantes registrados es muy costosa, se utiliza muestreo por conglomerados, con distritos como conglomerados. Se selecciona una muestra irrestricta aleatoria de 50 distritos de un total de 497 que tiene el estado. El periódico quiere hacer la estimación el día de la elección, pero antes de que se haya hecho la cuenta final de los votos. Es por eso que los reporteros son enviados a los lugares de votación de cada distrito en la muestra, para obtener la información pertinente directamente de los votantes. Los resultados se muestran en la tabla anexa. Estime la proporción de votantes que apoyan al candidato A, y establezca un límite para el error de estimación.

Número de votantes	Número que vota por A	Número de votantes	Número que vota por A	Número de votantes	Número que vota por A
1290	680	1893	1143	843	321
1170	631	1942	1187	1066	487
840	475	971	542	1171	596
1620	935	1143	973	1213	782
1381	472	2041	1541	1741	980
1492	820	2530	1679	983	693
1785	933	1567	982	1865	1033
2010	1171	1493	863	1888	987
974	542	1271	742	1947	872
832	457	1873	1010	2021	1093
1247	983	2142	1092	2001	1461
1896	1462	2380	1242	1493	1301
1943	873	1693	973	1783	1167
798	372	1661	652	1461	932
1020	621	1555	523	1237	481
1141	642	1492	831	1843	999
1820	975	1957	932		

- 8.15 El periódico del Ejercicio 8.14 quiere realizar una encuesta similar durante la siguiente elección. ¿Qué tan grande se necesitará la muestra para estimar la proporción de votantes que favorecen un candidato similar, con un límite de 0.05 para el error de estimación?
- 8.16 Un guardabosques desea estimar la altura promedio de los árboles en una plantación. La plantación se divide en parcelas de un cuarto de acre. Se selecciona una muestra irrestricta aleatoria de 20 parcelas de un total de 386 parcelas en la plantación. Se miden todos los árboles en las parcelas muestreadas, con los resultados que se muestran en la tabla anexa. Estime la altura promedio de los árboles en la plantación y establezca un límite para el error de estimación. (Sugerencia: el total para el conglomerado i se puede encontrar tomando m_i veces el promedio del conglomerado.)

Número de árboles	Altura promedio (en pies)	Número de árboles	Altura promedio (en pies)
42	6.2	60	6.3
51	5.8	52	6.7
49	6.7	61	5.9
55	4.9	49	6.1

47	5.2	57	6.0
58	6.9	63	4.9
43	4.3	45	5.3
59	5.2	46	6.7
48	5.7	62	6.1
41	6.1	58	7.0

- 8.17 Para reafirmar la seguridad, una compañía de taxis quiere estimar la proporción de llantas inseguras en sus 175 taxis. (No considere las llantas de refacción.) La selección de una muestra aleatoria de llantas es impráctica, así que se usa muestreo por conglomerados, con cada taxi como un conglomerado. Una muestra irrestricta aleatoria de 25 taxis nos da los siguientes números de llantas inseguras por taxi:

$$\begin{array}{l} 2, 4, 0, 1, 2, 0, 4, 1, 3, 1, 2, 0, 1, \\ 1, 2, 2, 4, 1, 0, 0, 3, 1, 2, 2, 1 \end{array}$$

Estime la proporción de llantas inseguras que se están usando en la compañía de taxis, y establezca un límite para el error de estimación.

- 8.18 Los comercios solicitan frecuentemente a los contadores la realización de inventarios. Ya que un inventario completo es costoso, a través del muestreo se pueden realizar inventarios cada cuatro meses. Supóngase que una empresa abastecedora de artículos de plomería desea un inventario para muchos artículos pequeños en existencia. La obtención de una muestra aleatoria de artículos es muy difícil. Sin embargo, los artículos se encuentran dispuestos en anaqueles, y la selección de una muestra aleatoria de anaqueles es relativamente fácil, considerando a cada anaquel como un conglomerado de artículos. Una muestra de 10 anaqueles de un total de 48 dio los resultados que se muestran en la tabla siguiente. Estime la cantidad total de dólares de los artículos en los anaqueles y establezca un límite para el error de estimación.

Conglomerado	Número de artículos, m_i	Cantidad total de dólares, y_i
1	42	83
2	27	62
3	38	45
4	63	112
5	72	96
6	12	58
7	24	75
8	14	58
9	32	67
10	41	80

- 8.19 Una empresa especializada en la fabricación y venta de ropa de descanso tiene 80 almacenes en Florida y 140 en California. Con cada estado como un estrato, la empresa desea estimar el tiempo promedio de ausencia por enfermedad por empleado durante el año pasado. Cada almacén puede ser considerado como un conglomerado de empleados, y se puede determinar de los registros el tiempo total de ausencia por enfermedad para cada almacén. Muestras irrestrictas aleatorias de 8 almacenes de Florida y 10 almacenes de California nos dan los resultados que se muestran en la tabla acompañante (m_i denota el número de empleados y y_i denota el total de días de ausencia por enfermedad para el i -ésimo almacén). Estime la cantidad promedio de ausencia por enfermedad por empleado, y calcule un estimador de la varianza de su estimador.

Florida		California	
m_i	y_i	m_i	y_i
12	40	16	51
20	52	8	32
8	30	4	11
14	36	3	10
24	71	12	33
15	48	17	39
10	39	24	61
6	21	30	37
		21	40
		9	41

- 8.20 Las estadísticas de manzana reportan el número de unidades habitacionales, el número de residentes y el número total de cuartos dentro de las unidades habitacionales para una muestra aleatoria de ocho manzanas seleccionadas de una gran ciudad. (Suponga que el número de manzanas en la ciudad es muy grande.) Los datos se presentan en la tabla acompañante.

Manzana	Número de unidades habitacionales	Número de residentes	Número de cuartos
1	12	40	58
2	14	39	72
3	3	12	26
4	20	52	98
5	12	37	74
6	8	33	57
7	10	41	76
8	6	14	48

- (a) Estime el número promedio de residentes por unidad habitacional y establezca un límite para el error de estimación.
(b) Estime el número promedio de cuartos por residente y establezca un límite para el error de estimación.

- 8.21 Ciertos tipos de tableros de circuitos fabricados para su instalación en computadoras tienen 12 microcircuito defectuosos por tablero. Durante la inspección de control de calidad de 10 de esos tableros, el número de microcircuito defectuosos por tablero fue como sigue:

$$2, 0, 1, 3, 2, 0, 0, 1, 3, 4$$

Estime la proporción de microcircuito defectuosos en la población de la cual se extrajo la muestra y establezca un límite para el error de estimación.

- 8.22 Considere la situación del Ejercicio 8.21. Suponga que la muestra utilizada proviene de un embarque de 50 de tales tableros. Estime el número total de microcircuito defectuosos en este embarque y establezca un límite para el error de estimación.

- 8.23 Una empresa grande tiene sus inventarios de equipo listados separadamente por departamento. De los 15 departamentos en la empresa, se van a muestrear aleatoriamente 5, por un auditor que

va a verificar que todo el equipo esté identificado y localizado apropiadamente. La proporción de artículos del equipo que no estén identificados propiamente es de interés al auditor. Los datos se dan en la tabla siguiente. Estime la proporción de artículos del equipo en la empresa que no están identificados propiamente y establezca un límite para el error de estimación.

Departamento	Número de artículos del equipo	Número de artículos identificados inapropiadamente
1	15	2
2	27	3
3	9	1
4	31	1
5	16	2

- 8.24 Suponga que para la empresa del Ejercicio 8.23, los 15 departamentos tienen el número de artículos del equipo que se da en la tabla acompañante. Seleccione una muestra de 3 departamentos, con probabilidades proporcionales al número de artículos del equipo.

Departamento	Número de artículos	Departamento	Número de artículos
1	12	9	31
2	9	10	26
3	27	11	22
4	40	12	19
5	35	13	16
6	15	14	33
7	18	15	6
8	10		

- 8.25 Suponga que los tres departamentos seleccionados en el Ejercicio 8.24 tienen cada uno dos artículos del equipo identificados inapropiadamente. Estime el número total de artículos inapropiadamente identificados en la empresa y establezca un límite para el error de estimación.
- 8.26 Un gran embarque de mariscos congelados es empaquetado en cajas, conteniendo cada uno veinticuatro paquetes de 5 libras. Hay cien cajas en el embarque. Un inspector del gobierno determina el peso total (en libras) de mariscos dañados para cada una de cinco cajas muestreadas. Los datos son como sigue:

9, 6, 3, 10, 2

Estime el peso total de mariscos dañados en el embarque y establezca un límite para el error de estimación.

- 8.27 Usando los datos del Ejercicio 8.26, estime la cantidad promedio de mariscos dañados por paquete de 5 libras y establezca un límite para el error de estimación.
- 8.28 Un politólogo desea muestrear a los estudiantes residentes de una universidad. Las unidades habitacionales pueden ser convenientemente usadas como conglomerados de estudiantes, o co-

lecciones de unidades habitacionales (dormitorios para estudiantes de primer año, casas de fraternidad, y así sucesivamente) pueden ser usadas como estratos. Analice los méritos de muestreo por conglomerados contra muestreo aleatorio estratificado, si el objetivo es estimar la proporción de estudiantes que favorecen a cierto candidato en los siguientes tipos de elecciones.

- (a) Una elección de dirigentes estudiantiles.
 (b) Una elección del presidente del país.

- 8.29 ¿En qué condiciones el muestreo por conglomerados produce un límite más pequeño para el error de estimación de una media que el muestreo irrestricto aleatorio?
- 8.30 Sin considerar los costos de muestreo, ¿qué criterio usaría usted para seleccionar conglomerados apropiados en un problema de muestreo por conglomerados?

EXPERIENCIAS CON DATOS REALES

- 8.1 En la Tabla 3 del Apéndice se muestra el ingreso por persona en Estados Unidos (durante 1977). Se presentan también valores para la población de 1980. Tratando a cada estado como un conglomerado de personas, seleccione una muestra aleatoria de estados y estime el ingreso personal total para Estados Unidos. Establezca un límite para el error de estimación.
- 8.2 Trate de realizar un estudio económico, tal vez considerando los hogares en cierta área geográfica fija (tal vez unas cuantas manzanas de la ciudad) como conglomerados de personas. Seleccione una muestra de n hogares y, después de obtener el permiso para la entrevista, registre la cantidad total semanal que se gasta en alimentos por todos los individuos en el hogar, y el número de individuos. Entonces estime la cantidad promedio gastada en alimentos por persona entre los hogares de esta población. Aun si todo el dinero es realmente gastado por una persona (digamos la madre), la cantidad total es la misma que se hubiera registrado si cada individuo hubiera comprado su propia alimentación. Entonces, se dispone del total para el conglomerado, aun cuando puede no contarse con las observaciones por elemento.