CAPITULO

VEINTICINCO

MUESTREO DE POBLACIONES FINITAS

25.1 Introducción

El muestreo estudiado hasta ahora se ha referido a experimentos en los cuales las unidades experimentales no se obtuvieron generalmente por procedimientos aleatorios. La aleatorización se usó para asignar tratamientos a unidades y las poblaciones eran hipotéticas en tanto que las unidades diferían por errores aleatorios. Ahora dirigimos nuestra atención a poblaciones que ya no son teóricas sino a poblaciones cuyas unidades experimentales pueden enumerarse y, en consecuencia, se pueden muestrear aleatoriamente. Por ejemplo, los silos se muestrean para determinar los residuos de insecticidas, los suelos para análisis químicos, las poblaciones de plantas, con propósitos taxanómicos, las frutas, para determinar calidad, los campos de trigo, para estimaciones de rendimientos y calidad antes de la cosecha, las razas primitivas, para analizar muchas características, las personas, para conocer sus opiniones, y así sucesivamente. En todos estos ejemplos, las poblaciones que interesan son infinitas.

Un problema nuevo se presenta en el muestreo de poblaciones finitas. Por ejemplo, si deseamos obtener información de muestra de distribuidores de semillas al por mayor en un estado, estaremos muestreando una población finita. Si la muestra es grande, como digamos 25 por ciento de los distribuidores de semillas al por mayor, hay que saber si las técnicas existentes se pueden aplicar o si será necesario desarrollar otras nuevas.

Al muestrear *poblaciones finitas* hay tres maneras bien distintas de hacer la selección. Estas son

1. Muestreo aleatorio
2. Muestreo sistemático
3. Muestreo autoritario

El muestreo aleatorio será nuestra mayor preocupación. La aleatorización puede introducirse en el procedimiento de muestreo de varias maneras que nos dan diversos dise-

ños de muestras. Gracias a la aleatorización pueden obtenerse estimaciones válidas del error. Se puede aplicar la teoría de la probabilidad y se pueden sacar conclusiones válidas.

El muestreo sistemático se usa cuando cada k-ésimo individuo de la población se incluye en la muestra. Tal procedimiento es siempre muy fácil, pero evidentemente insatisfactorio si en la población se presentan tendencias o ciclos no reconocidos aún. Dado que las poblaciones se deben enumerar antes del muestreo, pueden introducirse en forma inconsciente ciertas relaciones entre una o más de las características investigadas y orden de enumeración. En general, no es seguro suponer que no existe tal relación.

El muestreo sistemático puede efectuarse en forma tal, que puede obtenerse una estimación no sesgada del error de muestreo. Esto requiere de más de una muestra sistemática. Para una sola muestra sistemática, las fórmulas de que se dispone para estimar la varianza de una media suponen el conocimiento de la forma de la población.

El muestreo autoritario exige que una persona, bien familiarizada con el material que va a muestrearse, extraiga la muestra sin tener en cuenta la aleatorización. Tal procedimiento depende completamente del conocimiento y pericia de la persona que hace el muestreo. Puede producir buenos resultados en algunos casos, pero rara vez se recomienda.

25.2 Organización del estudio

Un considerable esfuerzo debe dedicarse a la planeación y ejecución de un estudio muestral, además de al muestreo efectivo. Vamos a hacer arbitrariamente una lista de cinco etapas para efectuar un muestreo.

1. Aclaración de objetivos
2. Definición de la unidad de muestreo y de la población
3. Selección de la muestra
4. Realización del estudio
5. Análisis de los datos

Se exponen estas etapas en forma breve.

1. Aclaración de los objetivos Esto consiste, ante todo, en establecer objetivos tan concisamente como sea posible, en que cada objetivo se enuncia como una hipótesis que va a probarse, un intervalo de confianza que ha de calcularse, o una decisión que debe tomarse.

Con este objetivo primario en mente, consideremos qué datos deben recolectarse. Cuando se tienen varios objetivos en mente, tal vez haya que modificar nuestras ideas respecto a qué datos deben recolectarse con el propósito de lograr todos los objetivos. Hasta podemos modificar nuestros objetivos de modo que el estudio no se vuelva demasiado complejo y costoso.

Usualmente, el grupo de estudio contará con un presupuesto fijo y deseará maximizar la cantidad de información por cada peso gastado. O bien, los objetivos incluirán una declaración sobre la cantidad de información deseada, generalmente en la forma del tamaño de un intervalo de confianza y tendremos que minimizar el costo.

2. Definición de la unidad de muestreo y de la población Hasta cierto punto esto se ha hecho en la etapa 1. Para el muestreo, la población debe dividirse en *unidades de mues-*

treo que, en conjunto, constituyen la población. Pueden existir varias posibilidades de escoger las unidades de muestreo. La elección final puede ser un tanto arbitraria, pero debe ser utilizable. Si estamos muestreando personas, podemos escoger el individuo, la familia, o los ocupantes de alguna vivienda específica como unidad de muestreo. Cualquiera que sea la selección hay que localizar e identificar la unidad sobre el terreno.

Para el muestreo aleatorio, debemos poder *enumerar* todas las unidades de muestreo, y tener una lista de todas las unidades. Puede ser necesario revisar las unidades existentes, por ejemplo, listas de niños escolares, granjeros, etc., o bien hacer nuevas listas, lo que parezca más factible y económico. Se pueden juxtaponer cuadrículas sobre los mapas del terreno, o bosques u otras zonas de las cuales se necesite obtener muestras de cultivos o información respecto a la cobertura de la vida silvestre. Aquí, puede ser necesario tener inventiva y hasta arbitrariedad, especialmente si tenemos que hacer el muestreo en un área de forma irregular.

Mientras se decide sobre la unidad de muestreo, es necesario considerar qué va a medirse y qué métodos de medida deben usarse. ¿Estamos midiendo estatura, peso u opinión? Si así es, ¿cómo lo vamos a hacer? ¿Puede usarse un cuestionario para medir estrés emocional? Si así es, ¿pueden emplearse como entrevistadores estudiantes universitarios que buscan un trabajo de tiempo parcial? ¿Se les puede adiestrar en unos días? ¿Puede ser de igual utilidad un candidato a ingeniero que un estudiante de premedicina?

3. Selección de la muestra Las formas en que pueden extraerse una muestra se llaman *diseños muestrales*. De ellos se hablará en secciones posteriores de este capítulo.

La selección del tamaño de la muestra se relaciona, en parte, con los recursos disponibles; si son inadecuados para obtener una muestra lo suficientemente grande para lograr los objetivos propuestos, deben revisarse los objetivos o retardar el estudio hasta cuando se tengan los fondos suficientes.

El diseño y el tamaño de la muestra darán una buena idea respecto a la extensión y naturaleza de las tablas y cálculos necesarios.

4. Realización del estudio Probablemente, será necesario adiestrar a parte del personal con el objeto de lograr uniformidad en la localización o identificación de las unidades de muestreo y en el registro de las respuestas a cuestionarios u otros datos. Será necesario un cronograma de actividades. Generalmente, se requiere de un esquema para una verificación temprana de la validez de los datos registrados en los diversos formatos. Debe preverse qué hacer en caso de que haya que tomar decisiones rápidas frente a hechos inesperados.

5. Análisis de los datos Primero, será necesario corregir los datos en cuanto a errores de registro e invalidez. Finalmente, el estudio muestral se deberá revisar en cuanto a maneras posibles de mejorar estudios futuros.

25.3 Muestreo probabilístico

Supóngase que nuestra población se ha definido claramente y que se ha hecho un listado de las unidades de muestreo. Ahora también podemos hacer una lista de todas las posibles muestras. Usamos el término *muestreo probabilístico* cuando

1. Cada unidad de muestreo tiene, o se le ha asignado, una probabilidad conocida de estar en la muestra.
2. Hay selección aleatoria en alguna etapa del procedimiento de muestreo y está directamente relacionado con probabilidades conocidas. La selección aleatoria supone un procedimiento mecánico para seleccionar las unidades que deben incluirse en la muestra.
3. El método de cálculo de una estimación de una media se establece claramente y llevará a un solo valor de la entimación. Esto es parte del análisis de los datos. Al estimar una media, usamos las probabilidades de selección asignadas a las unidades muestrales. Estas suministrarán ponderaciones, cada una de las cuales será cierto múltiplo constante del inverso de la probabilidad.

Cuando se cumplen estos criterios, puede asignarse una probabilidad de selección a cada muestra y a cada estimación. Por tanto, podemos construir una distribución de probabilidades de las estimaciones dadas por nuestro plan de muestreo. De esta manera, podemos evaluar el valor de nuestro plan y compararlo con otros planes de muestreo probabilístico. La evaluación consiste en medir la exactitud de toda estimación por la magnitud de su desviación estándar.

Cuando las probabilidades asignadas a cada unidad de muestreo son iguales, entonces los pesos que han de usarse en el cálculo de las estimaciones de las medias son todas iguales. Realmente no necesitamos pensar concretamente en las ponderaciones, ya que la muestra es *autoponderada*. Aunque tales muestras son fáciles de analizar, carecen de ciertas ventajas que poseen otros planes de muestreo probabilístico, ventajas tales como facilidad y bajo costo de administración por unidad de información y la capacidad de obtener estimaciones para estratos individuales (ver sec. 25.5).

Una muestra probabilística no garantiza que todas nuestras estimaciones sean no sesgadas. Ya hemos visto que en el muestreo aleatorio a partir de una población normal, optamos por utilizar $s = \sqrt{\sum (Y - \bar{Y})^2/(n-1)}$ si bien es una estimación sesgada de σ. También se usan estimaciones sesgadas en estudios muestrales. Deben usarse, naturalmente, con precaución, ya que pueden introducir distorsiones en las unidades probabilísticas. En particular, cuando se promedian estimaciones sesgadas (no necesariamente aritméticamente), el efecto sobre el promedio y su uso posterior puede no ser claro. Ahora se exponen varios tipos de muestreo probabilístico.

25.4 Muestreo aleatorio simple

Para el muestreo aleatorio, se hace un listado de la población y se fija el plan y tamaño de la muestra. Para el *muestreo aleatorio simple*, cada muestra posible tiene la misma probabilidad de ser seleccionada. Este es el criterio importante.

En el proceso efectivo de selección de las unidades muestrales en una población finita, se usa una tabla de números aleatorios y el muestreo se hace sin *reemplazo*. Aparte de esto, las unidades de muestreo se extraen independientemente.

Notación y definiciones Como ahora estamos tratando principalmente con poblaciones finitas, se requieren nueva notación y definiciones. La notación y las definiciones, como

se verá, no son del todo coherentes en la literatura del muestreo. Trataremos de usar letras mayúsculas para cantidades de población y letras minúsculas para cantidades muestrales; también se usa σ^2.

Para comenzar, sea Y_i la observación i-ésima en la población. También usamos Y_i para describir la i-ésima observación muestral cuando no hay confusión posible.

Tamaño de la población: N
Tamaño de la muestra: n
Media de la población:

$$\bar{Y} = \frac{\sum_i Y_i}{N} = \frac{Y.}{N}, \qquad \text{variable continua}$$

$$P = \frac{A}{N} \qquad \text{proporción}$$

Para una proporción $Y_i = 0$ ó 1; $\sum Y_i/N$ es la proporción de individuos que poseen una característica específica, así que puede servir también como una definición de la media de la población. Es común remplazar $\sum_i Y_i$ por A. Para un porcentaje la media apropiada es $100P$.

Media muestral:

$$\hat{\bar{Y}} = \bar{y} = \frac{\sum_i Y_i}{n} = \frac{y}{n}, \qquad \text{variable continua}$$

$$\hat{P} = p = \frac{a}{n}, \qquad \text{proporción.}$$

Para una proporción, a remplaza a $\sum_i Y_i$. Como los totales de población y sus estimaciones son a menudo de interés, son bien corrientes las cantidades Y, A, $Y.$, a.

Varianza de la población:

$$\sigma^2 = \frac{\sum_i (Y_i - \bar{Y})^2}{N}$$

$$S^2 = \frac{\sum_i (Y_i - \bar{Y})^2}{N - 1} \tag{25.1}$$

Usamos la ec. (25.1) para definir la varianza de la población, ya que nuestra definición de s^2, ec. (25.4), da una estimación no sesgada de S^2.

Varianza de la población de una media:

$$S_{\bar{y}}^2 = \frac{S^2}{n}\left(\frac{N-n}{N}\right) \tag{25.2}$$

$$S_{\bar{p}}^2 = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right) \qquad \text{donde } Q = 1 - P \tag{25.3}$$

Varianza muestral:

$$s^2 = \frac{\sum_i (Y_i - \bar{y})^2}{n-1} \qquad \text{una estimación insesgada de } S^2 \tag{25.4}$$

El numerador se calcula como $\sum Y_i^2 - (\sum Y_i)^2/n$.

Varianza muestral de una media:

$$s_{\bar{y}}^2 = \frac{s^2}{n}\left(\frac{N-n}{N}\right) \tag{25.5}$$

$$s_{\bar{p}}^2 = \frac{pq}{n-1}\frac{N-n}{N} \qquad \text{donde } q = 1 - p \tag{25.6}$$

La ecuación (25.6) da una estimación insesgada de $S_{\bar{p}}^2$, pero generalmente no se usa cuando se calculan intervalos de confianza. La forma más familiar está implícita en la ec. (25.8).

La cantidad $(N - n)/N$ se conoce como *corrección de población finita* o cpf. También puede escribir $1 - n/N$ y a n/N se le llama *fracción de muestreo*. Si la fracción de muestreo es pequeña, digamos menos del 5 por ciento, puede omitirse. Es de interés observar que $qp/(n-1)$ es una estimación no sesgada de la varianza de la población independientemente de que la población sea finita o no, o sea que usamos una estimación insesgada de la varianza de la población en los caps. 20 a 23 cuando usamos $\hat{p}(1-\hat{p})/n$. (Recuérdese que en los caps. 20 a 23 p se usa como parámetro y $\hat{p}$ como estimación).

El intervalo de confianza para una media está dado por la ec. (25.7). Obsérvese que se hace uso de la cpf.

$$\text{IC} = \bar{y} \pm t\left[\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}\right] \tag{25.7}$$

Obviamente estamos suponiendo que $\bar{y}$ se distribuye normalmente, sabiendo que la población de los Y no es normal, ya que es una población finita. Además, se muestrea sin remplazo.

El intervalo de confianza para una proporción exige una distribución hipergeométrica, sec. 23.2, si ha de ser completamente válido. El investigador interesado puede remitirse a los gráficos de Chung y DeLury (25.3). Una aproximación corriente es

$$\text{IC} = p \pm t\sqrt{\frac{pq}{n}\frac{N-n}{N-1}} \tag{25.8}$$

Obsérvese que la desviación estándar estimada no es la que da la ec. (25.6), pero es comparable con la cantidad poblacional dada por la ec. (25.3). La estimación usada en la ec. (25.8) es la más común. La tabla 21.1 puede usarse para juzgar lo apropiado de la ec. (25.8).

El muestreo aleatorio simple se usa cuando se sabe que la población no es muy variable cuando la verdadera población cae entre el 20 y el 80 por ciento. Cuando hay variación considerable, las unidades de muestreo deberán agruparse en estratos de tal modo que puede esperarse que la variación dentro de los estratos sea inferior a la variación entre estratos. Esto lleva al muestreo estratificado. Idea muy parecida es la que lleva a un análisis de la varianza entre grupos y dentro de grupos.

Ejercicio 25.4.1 Considérese la población finita que consiste en los números 1, 2, ..., 6. Calcular la media y la varianza. Considérense todas las posibles muestras de dos observaciones cuando se muestrea esta población sin remplazo. Hacer una tabla de medias y varianzas muestrales y la frecuencia de ocurrencia de cada uno de los valores. Demostrar que la media y varianza muestrales de la ec. (25.5) son estimaciones insesgadas de la media y varianza poblacionales de la ec. (25.1). Si el ejercicio hubiese dicho "muestra con remplazo" ¿qué cambios se requerirían en los cálculos?

Ejercicio 25.4.2 Se va a muestrear una población consistente en 6,000 compradores de muebles, mediante un cuestionario enviado por correo en relación con su preferencia por determinado mueble. Se toma una muestra aleatoria de 250 personas y se envían los cuestionarios. Como la preferencia se refiere a un accesorio poco costoso, se prevén sólo respuestas sí o no. Todos los cuestionarios fueron devueltos con 187 respuestas sí. Estimar la verdadera proporción de respuestas sí en la población mediante un intervalo de confianza del 95 por ciento.

Ejercicio 25.4.3 Con la misma población del ejercicio 25.4.2, se envió un cuestionario más largo a una muestra aleatoria de 750 compradores. Sólo fueron devueltos 469 cuestionarios. Estimar la verdadera proporción de los que *respondieron* en la población mediante un intervalo de confianza del 90 por ciento.

25.5 Muestreo estratificado

La varianza estimada de una media de población está dada por la ec. (25.5) y, para una proporción, por la ec. (25.6) o la alternativa más frecuente que implica la ec. (25.8). Para disminuir la longitud del intervalo de confianza que estima la media de población, podemos aumentar n o disminuir la varianza de la población. Obviamente, ambas posibilidades deben considerarse.

La forma obvia de disminuir una varianza de población es construir *estratos* con las unidades de muestreo, así que la variación total se particiona de tal manera que la mayor parte posible se asigne a diferencia entre estratos. Así que la variación dentro de los estratos se mantiene baja. La variación entre medias de estratos en la población no contribuye al error de muestreo de la estimación de la media de la población. Ver la ec. (25.15).

La reducción en la variación de la estimación de la media de población es una razón muy importante para la estratificación. Pero en muchos estudios entran varias variables y una buena estratificación para una variable puede no serlo para otra. Así vemos que los estratos a menudo se construyen con base puramente geográfica. En general, esto da resultado, y así encontramos municipios, condados y zonas de recursos de tierra usados como estratos. Este tipo de estratificación a menudo es conveniente por razones administrativas,

ya que es posible obtener la cooperación de organismos, municipios, condados u otros organismos convenientemente localizados.

Además de aumentar la precisión con la cual se miden las medias, la estratificación permite un trabajo eficiente de asignación de recursos ya que podemos usar cualquier método para decidir cuántas unidades de muestreo se han de tomar de cada estrato. Se supone que se hará muestreo en cada estrato. A menudo se desean estimaciones de medias de estratos y, en tales casos, la estratificación es esencial.

Notación y definiciones La notación y definiciones para el muestreo aleatorio estratificado están obviamente relacionadas y son extensiones de las ya dadas en la sec. 25.4 bajo el mismo encabezamiento.

Sea Y_{ki} la observación i-ésima en el estrato k-ésimo, $k = 1, \ldots, s$. Los tamaños de los estratos, medias y varianzas se designarán mediante N_k, $\overline{Y}_k$, o P_k y S_k^2 con los valores muestrales correspondientes n_k, $\bar{y}_k$ o p_k y s_k^2.

La media y la varianza de los estratos son:

$$\overline{Y}_k = \frac{\sum_{i=1}^{N_k} Y_{ki}}{N_k} = \frac{Y_k}{N_k}$$

$$P_k = \frac{A_k}{N_k}$$

$$S_k^2 = \frac{\sum_{i=1}^{N_k} (Y_{ki} - \overline{Y}_k)^2}{N_k - 1}$$

según la ec. (25.1)

La media y la varianza muestral para el estrato k-ésimo:

$$\hat{\overline{Y}}_k = \bar{y}_k = \frac{\sum_{i=1}^{n_k} Y_{ki}}{n_k} = \frac{y_k}{n_k}$$

$$\hat{P}_k = p_k = \frac{a_k}{n_k}$$

$$s_k^2 = \frac{\sum_{i=1}^{n_k} (Y_{ki} - \bar{y}_k)^2}{n_k - 1}$$

según la ec. (25.4)

También se necesitarán parámetros y estadígrafos para la población completa. Sean

$$N = \sum_k N_k \qquad \text{y} \qquad n = \sum_k n_k$$

La razón N_k/N se presenta con bastante frecuencia y se la representa con W_k, esto es $W_k = N_k/N$ donde W corresponde a la ponderación.

Media de población (est significa estratificado):

$$\bar{Y}_{est} = \frac{\sum_k N_k \bar{Y}_k}{N} = \sum_k W_k \bar{Y}_k \tag{25.9}$$

$$P_{est} = \frac{\sum_k N_k P_k}{N} = \sum_k W_k P_k \tag{25.10}$$

Estimación de la media de población

$$\hat{\bar{Y}}_{est} = \bar{y}_{est} = \frac{\sum_k N_k \bar{y}_k}{N} = \sum_k W_k \bar{y}_k \tag{25.11}$$

$$\hat{P}_{est} = p_{est} = \frac{\sum_k N_k p_k}{N} = \sum_k W_k p_k \tag{25.12}$$

(Las medias muestrales son $\bar{y} = \sum n_k \bar{y}_k/n$ y $p = \sum_k n_k p_k/n$.)

La varianza de la estimación de la media de población:

$$\begin{aligned} \sigma^2(\bar{y}_{est}) &= \sum_k \left(\frac{N_k}{N}\right)^2 \frac{S_k^2}{n_k} \frac{N_k - n_k}{N_k} \\ &= \frac{1}{N^2} \sum N_k(N_k - n_k)\frac{S_k^2}{n_k} \end{aligned} \tag{25.13}$$

Compárese la ec. (25.5) con la primera expresión para $\sigma^2(\bar{y}_{est})$.

$$\sigma^2(p_{est}) = \sum_k \frac{N_k^2}{N^2} \frac{P_k Q_k}{n_k} \frac{N_k - n_k}{N_k - 1} \tag{25.14}$$

Compárese la varianza dada en la ec. (25.3) con $\sigma^2(p_{est})$.

La varianza muestral de la estimación de la media de población:

$$\begin{aligned} s^2(\bar{y}_{est}) &= \frac{1}{N^2} \sum_k N_k(N_k - n_k)\frac{s_k^2}{n_k} \\ &= \sum_k W_k^2 \frac{s_k^2}{n_k} - \frac{1}{N} \sum_k W_k s_k^2 \end{aligned} \tag{25.15}$$

La primera forma se obtiene de la ec. (25.13) sustituyendo parámetros por estimadores. Es una estimación no sesgada de $\sigma^2(\bar{y}_{est})$. La segunda forma puede usarse para cálculos.

$$s^2(p_{est}) = \sum_k W_k^2 \frac{p_k q_k}{n_k} \frac{N_k - n_k}{N_k - 1} \tag{25.16}$$

Esta ecuación se obtiene de la ec. (25.14). Da una estimación sesgada de $\sigma^2(p_{est})$, pero es de uso común.

Si la corrección de población finita es pequeña, se omite cuando se calculan los intervalos de confianza.

Al estimar la media de la población, se usan ponderaciones. Por esta razón, la estimación de la media de la población y de la media muestral, $\bar{y}_{est}$ y $\bar{y}$, respectivamente, no tienen que ser las mismas. Sin embargo, $n_1/N_1 = \ldots = n_s/N_s = n/N$, entonces $\bar{y}_{est} = \bar{y}$. A ésto se le llama *asignación proporcional* y se dice que la muestra es *autoponderada*.

Cuando se usa la asignación proporcional y las varianzas *dentro de los estratos* son homogéneas, los resultados relativos a las varianzas se pueden resumir en una tabla de análisis de la varianza con las fuentes de variación para el total, entre y dentro de estratos. El valor de la estratificación particular puede estimarse comparando la desviación estándar de $\bar{y}_{est}$, calculada a partir del cuadrado medio dentro de estratos, con la de $\bar{y}$, calculada a partir del cuadrado medio total. Cochran (25.2) y Hansen y otros (25.4) dan procedimientos exactos.

Ejercicio 25.5.1 El problema de *los que no responden* es general, y de él padecen los usuarios de cuestionarios enviados por correo y aún los entrevistadores. Supóngase que se escogen al azar 900 compradores de una población de 6,000. Se reciben respuestas de 250 y, de éstos, 195 están a favor de una sugerencia. Estimar la proporción de *"los que están a favor"* en la población de los que responden; utilizar un intervalo de confianza del 95 por ciento.

De los que no responden conocidos, se extrae una muestra aleatoria de 50 y se entrevistan. De éstos, "a favor" hay 30 en la muestra de la población de los que no responden. Estimar mediante un intervalo de confianza del 95 por ciento, la proporción de los que responden "a favor" en esta población.

A veces la respuesta y no respuesta a los cuestionarios enviados por correo se usan como criterio de estratificación. Supóngase que los resultados anteriores se consideran como provenientes de tales estratos. Estimar la proporción de la población de los que están "a favor" y la desviación estándar de la estimación. Dar un juicio crítico práctico y dos teóricos sobre el procedimiento.

¿Cómo se podrían usar los resultados de la muestra para contrastar si los dos estratos diferían o no en respuesta a la sugerencia?

25.6 Asignación óptima

La estratificación generalmente produce una disminución en la varianza de la estimación de la media de la población. Sin embargo, la asignación proporcional no siempre es una *asignación óptima* y, por esto, puede necesitarse una *fracción muestral variable*.

Costo fijo En algunos experimentos en muestreo, el costo de obtener una observación a partir de una unidad de muestreo no varía nada de un estrato a otro, y se puede omitir al determinar las fracciones muestrales para los diferentes estratos. El problema está en mi-

nimizar $\sigma^2(\bar{y}_{est})$ tal como aparece en la ec. (25.13) o $\sigma^2(p_{est})$ dada por la ec. (25.14). Las fracciones muestrales se determinan por el tamaño del estrato y su variabilidad, y es bien claro que un estrato mayor implica un número de observaciones mayor, como ocurriría con un estrato con alta variabilidad. Se ha demostrado que la asignación óptima se obtiene cuando el número de observaciones tomadas en un estrato se determina mediante

$$n_k = n\frac{N_k S_k}{\sum_k N_k S_k} \tag{25.17}$$

Obsérvese que el denominador es la suma extendida a todos los estratos.

Aunque la aplicación de esta fórmula necesita de los parámetros S_k, $k = 1, \ldots, s$, a menudo es necesario usar estimaciones. Así, se desea información muestral para los años entre censos, se pueden usar desviaciones estándar con respecto al censo precedente más cercano. En otros casos, puede ser necesario hacer uso de información de otros estudios muestrales relacionados. Cuando no se dispone de estimaciones de S_k, se recomienda la asignación proporcional.

A veces, la ecuación (25.17), dará uno o más valores de n_k mayores que sus correspondientes N_k. En tales casos, el 100 por ciento del muestreo se hace en estos estratos y las restantes fracciones muestrales se ajustan de modo que la muestra total sea del tamaño planeado originalmente. Por ejemplo, si $n_s > N_s$ al aplicar la ec. (25.17), entonces se hace $n_s = N_s$ para propósitos del muestreo y recalculamos los restantes n_k con la ecuación

$$n_k = \frac{(n - N_s)N_k S_k}{\sum_{k=1}^{s-1} N_k S_k} \qquad k = 1, \ldots, s-1$$

Cuando el muestreo estratificado es para proporciones, el n_k se puede determinar mediante

$$n_k = n\frac{N_k\sqrt{P_k Q_k}}{\sum_k N_k\sqrt{P_k Q_k}} \tag{25.18}$$

Esta ecuación es una aproximación; es similar a la ec. (25.17), si bien ésta no es una aproximación.

Lo que se gana en precisión como resultado del uso de una asignación óptima en vez de una proporcional, probablemente no es tanto para estimar proporciones como para estimar medias de variables continuas. Como la asignación proporcional ofrece la comodidad de las muestras autoponderadas, frecuentemente se recomienda cuando se han de estimar proporciones.

Costo variable Cuando el costo para obtener una observación varía de estrato en estrato, se necesita cierta *función de costo* que dé el costo total. Una sencilla función de costo es la dada por

$$\text{Costo} = C = a + \sum c_k n_k \tag{25.19}$$

donde a es un costo fijo, independientemente del tipo de asignación del muestreo a los estratos, y c_k representa el costo por observación en el estrato k. Para esta función de costo, la $\sigma^2(\bar{y}_{\text{est}})$ mínima se obtiene si tomamos una muestra grande en un estrato grande, una muestra grande cuando la varianza del estrato es alta, y una muestra pequeña cuando el costo del estrato es alto. Es decir, el tamaño de la muestra para todo estrato es proporcional a $N_k S_k / \sqrt{c_k}$.

En un estudio efectivo tal vez tengamos que operar con un presupuesto fijo o bien haya que estimar la varianza de la media de población con una precisión especificada. El último requisito determina el tamaño de la muestra y, a su turno, el presupuesto.

Para un *presupuesto fijo*, el tamaño de muestra óptimo para cada estrato es

$$n_k = \frac{N_k S_k / \sqrt{c_k}(C - a)}{\sum_k N_k S_k \sqrt{c_k}} \tag{25.20}$$

Para una *varianza fija*, la ec. (25.21) da el tamaño óptimo de la muestra para cada estrato. En este caso, minimizamos el costo para una varianza fija o predeterminada.

$$\begin{aligned} n_k &= \frac{N_k S_k}{\sqrt{c_k}} \frac{\sum_k N_k S_k \sqrt{c_k}}{N^2 \sigma^2(\bar{y}_{st}) + \sum_k N_k S_k^2} \\ &= \frac{W_k S_k}{\sqrt{c_k}} \frac{\sum_k W_k S_k \sqrt{c_k}}{\sigma^2(\bar{y}_{st}) + \sum_k W_k S_k^2 / N} \end{aligned} \tag{25.21}$$

donde $W_k = N_k / N$.

Las estimaciones aproximadas de los S_k^2 y c_k suelen ser bastante adecuadas para estimar tamaños de muestra óptimos para los estratos. Cuando el muestreo es para estimar *proporciones*, S_k puede remplazarse por $\sqrt{P_k Q_k}$ en las ecs. (25.20) y (25.21) para tener n_k aproximadamente óptimos.

Ejercicio 25.6.1 Demuéstrese que las ecs. (25.17) y (25.20) dan los mismos resultados que la asignación proporcional cuando las varianzas de los estratos son homogéneos y el costo c_k no varía.

Ejercicio 25.6.2 Los siguientes datos provienen de R. J. Jessen (25.5). Los estratos son tipos de áreas de cultivo, N_k el número de granjas rurales, $S_k^2(1)$ es una varianza para el número de porcinos y $S_k^2(2)$ es una varianza para el número de ovejas. Los N_k corresponden a datos del censo de 1939, mientras que los S_k^2 se han obtenido de una muestra tomada en 1939 y son solamente estimaciones.

Para cada conjunto de datos, calcular los tamaños de muestras utilizando asignaciones proporcional y óptima para un tamaño de muestra total de 800. Comparar los resultados.

Estrato	1	2	3	4	5	Estado
N_i	39,574	38,412	44,017	36,935	41,832	200,770
$S_i^2(1)$	1,926	2,352	2,767	1,967	2,235	2,303
$S_i^2(2)$	764	20	618	209	87	235

25.7 Muestreo multietápico o por conglomerados

En ciertos esquemas de muestreo, las unidades de muestreo están en grupos de igual o desigual tamaño y los grupos, en vez de unidades, son los que se muestran aleatoriamente. A tales grupos se les llama *unidades primarias de muestreo* o upm. Las observaciones pueden obtenerse sobre todas las unidades elementales o éstas, a su vez, pueden ser muestreadas. Por ejemplo, podemos estar interesados en individuos, las unidades elementales, y podemos obtenerlas extrayendo una muestra aleatoria de familias, las unidades primarias de muestreo, y observar dentro de ellas todas las unidades. Esto sería un plan de *muestreo por conglomerados simple* o un plan de *muestreo de una etapa*. Al muestrear el suelo en un terreno para llevar a cabo un experimento, podemos dividir el terreno en parcelas experimentales, colocar una cuadrícula encima de cada parcela para definir las unidades de muestreo, y luego obtener varias observaciones de cada parcela. Esto sería un muestreo en *dos etapas o submuestreo*, en el que la primera etapa era esencialmente un censo.

Obviamente pueden idearse muchos tipos de muestreo por conglomerados. La mayoría tendrá ciertas ventajas obvias en relación con el costo y la aplicabilidad, ya que el costo de pasar de una upm a otra es probablemente mayor que el de pasar de una subunidad a otra, porque la identificación de los upm puede ser más simple que identificar la subunidad. Cuando un conglomerado se define por asociación con un área, tenemos muestreo de área. Por ejemplo, podemos muestrear secciones de a cuarto de milla cuadrada de área, el conglomerado o upm, y enumerar todas las granjas en la upm.

Supóngase que una población consiste en N upm, de la cual sacamos una muestra aleatoria de tamaño n; cada upm consiste en M subunidades, de las cuales extraemos una muestra de tamaño m para cada una de las n upm. (Las letras M y N, y m y n se intercambian a menudo en la literatura de muestreo). Ahora hay MN *elementos* en la población y mn en la muestra. El plan es de muestreo en dos etapas o de submuestreo.

Los cálculos usualmente se efectúan por elementos, tal como se acostumbra en el análisis de la varianza. Una observación se denota mediante Y_{ij}, donde j se refiere al elemento e i a la upm. La media de todos los elementos en una upm se designa $\overline{Y}_{i\cdot}$ o simplemente $\overline{Y}_i$, y la media de población por $\overline{Y}_{\cdot\cdot}$ o simplemente $\overline{Y}$. Las medias muestrales correspondientes están dadas por los símbolos $\bar{y}_{i\cdot}$ o $\bar{y}_i$ y $\bar{y}_{\cdot\cdot}$ o $\bar{y}$.

Supongamos ahora que N y M son infinitos y definimos un elemento por

$$Y_{ij} = \overline{Y} + \delta_i + \varepsilon_{ij} \tag{25.22}$$

Este es un modelo lineal tal como se expuso en las secs. 7.6 y 7.7 con notación algo diferente. Si denotamos la varianza de los δ por S_δ^2 y las de los ε por S_w^2, entonces los cuadrados medios muestrales definidos como en la tabla 25.1 son estimaciones de las cantidades

en la columna de valor esperado. Obsérvese que no se intenta que s_a^2 sea una estimación de S_a^2.

Las sumas de cuadrados se calculan usualmente a partir de las siguientes fórmulas de cálculo y no por las fórmulas de definición de la tabla 25.1.

$$\text{Entre las upm:} \qquad (n-1)s_a^2 = \frac{\sum_i Y_{i\cdot}^2}{m} - \frac{Y^2}{nm}$$

$$\text{Dentro de las upm:} \qquad n(m-1)s_w^2 = \sum_i \left(\sum_{j=1}^{m} Y_{ij}^2 - \frac{Y_{i\cdot}^2}{m} \right)$$

$$\text{o} \qquad = \text{SC(total)} - \text{SC(dentro de las upm)}$$

$$\text{SC(total)} = \sum_{i,j} Y_{ij}^2 - \frac{Y^2}{nm}$$

Si relacionamos el análisis de la varianza dado en la tabla 25.1 con las secs. 7.6 y 7.8, vemos que el cuadrado medio dentro de las upm puede llamarse *error muestral* y que el cuadrado medio entre upm puede llamarse *error experimental*.

El error experimental en lugar del error muestral es el apropiado en cuanto a la estimación de $\bar{Y}$ mediante un intervalo de confianza. El error experimental se basa en la unidad escogida al azar en la primera etapa de muestreo y corresponde a la parcela a la cual se aplica un tratamiento en forma aleatoria en un experimento en el terreno o de laboratorio. En el curso corriente de los sucesos, sería de esperar que el error muestral fuese menor que el error experimental ya que esperaríamos más homogeneidad dentro de las upm que entre las upm. Así, el error muestral no sería apropiado para calcular un intervalo de confianza para $\bar{Y}$.

La varianza de la media muestral, $s^2(\bar{y})$, se estima mediante s_a^2/nm, estimación no sesgada de la verdadera varianza. Ahora podemos estimar tanto S_w^2 como S_a^2 y construir

Tabla 25.1 Análisis de la varianza y valores esperados en muestreo en dos etapas

Fuente de variación	gl	Cuadrado medio	Valores esperados del cuadrado medio
Entre las ump	$n-1$	$s_a^2 = \dfrac{m\sum_i (\bar{y}_i - \bar{y})^2}{n-1}$	$S_w^2 + mS_a^2$
Dentro las ump	$n(m-1)$	$s_w^2 = \dfrac{\sum_i \sum_j (Y_{ij} - \bar{y}_i)^2}{n(m-1)}$	S_w^2
Total	$nm-1$	$\dfrac{\sum_{i,j} (Y_{ij} - \bar{y})^2}{nm-1}$	

estimaciones de las varianzas de las medias de tratamientos para las diferentes asignaciones de nuestros esfuerzos. Así, para el esquema presente,

$$\sigma^2(\bar{y}) = \frac{S_w^2}{nm} + \frac{S_a^2}{n}$$

No se disminuye S_a^2/n tomando más submuestras, pero S_a^2 probablemente sea lo que más contribuya a $\sigma^2(\bar{y})$. Si fuésemos a muestrear n, entonces disminuiríamos ambas contribuciones. Así, en teoría, la mejor asignación a nuestro esfuerzo es tomar tantas upm como sea posible y tomar muy pocos elementos dentro de cada upm si ello implica un esfuerzo considerable; naturalmente, necesitamos dos de tales elementos de cada upm si tenemos que estimar bien sea S_w^2 o S_a^2 conservando facilidad de cálculo.

También se dispone de la teoría para poblaciones finitas y cuando las upm difieren en el número de elementos que contienen. El lector interesado puede consultar Cochran (25.2) y a Hansen et al. (25.4).

Al muestrear para proporciones con N conglomerados y M elementos por conglomerado, extráiganse n conglomerados y enumérense completamente. Entonces una proporción observada es una proporción verdadera P_i para el conglomerado i-ésimo y no está sujeta a variación muestral. Estimamos la proporción de población por

$$\hat{P} = p_{nM} = \frac{\sum_i P_i}{n}$$

y su varianza por

$$s^2(p_{nM}) = \frac{N-n}{N}\frac{1}{n}\frac{\sum_i (P_i - p_{nM})^2}{n-1}$$

Cuando el esquema de muestreo supone tomar sólo m de los M elementos en un conglomerado, entonces sólo se estima P_i y debe introducirse un término para la variación muestral dentro de conglomerados en la varianza de la estimación de la proporción de población P. Ahora tenemos

$$\hat{P} = p_{nm} = \bar{p} = \frac{\sum_i p_i}{n}$$

y

$$s^2(\bar{p}) = \frac{M-m}{M-1}\frac{m}{m-1}\frac{1}{Nnm}\sum_i p_i q_i + \frac{N-n}{N}\frac{1}{n}\frac{\sum_i (p_i - \bar{p})^2}{n-1}$$

Ladell (25.6) y Cochran (25.1) describen un interesante experimento de muestreo donde se impone un control local, en forma de una restricción sobre el submuestreo. En

Tabla 25.2 Análisis de la varianza de los datos de cienpiés

Fuente	gl	Suma de cuadrados	Cuadrado medio
Filas	4	515.44	128.86
Columnas	4	523.44	130.86
Error experimental	16	712.16	44.51
Entre mitades de parcelas	25	2,269.00	90.76
Error muestral	100	3,844.00	38.44
Totales	149	7,864.04	

el ejemplo particular intervino la superimposición de un diseño de cuadrado latino, inicialmente sin tratamientos, sobre un área experimental y la obtención de seis muestras de suelos de cada parcela. En estas muestras de suelos se hicieron recuentos de cienpiés. Se dispuso de manera que se tomaron tres muestras de cada una de las mitades norte y sur de la parcela. En consecuencia, las diferencias no aleatorias en el número de cienpiés entre las mitades no influyen en las comparaciones de tratamientos o en el error experimental. Los resultados se presentan en la tabla 25.2; aquí combinados el "error experimental" usual y "tratamientos" ya que no hubo verdaderos tratamientos. Es claro que "el control local" aumentó apreciablemente la precisión del experimento.

Ejercicio 25.7.1 Suponga que deseamos hacer una estimación antes de la cosecha del rendimiento de trigo en un estado donde se cultiva trigo. El área sembrada de trigo se divide, para los fines del muestreo, en parcelas de un acre. Se toma una muestra aleatoria de 250 parcelas y se obtienen dos submuestras de cada una de las 250 parcelas. Cada submuestra es de 2 pies cuadrados, aproximadamente 1/10,000 de acre, así que no se necesita aplicar la teoría del muestreo finito.

El análisis de la varianza da un error muestral de 20 (dentro de las ump) y un error experimental de 70 (entre las ump). (Los rendimientos fueron convertidos a bushels por acre). Efectuar el análisis de la varianza. Estimar las componentes de la varianza. Calcular la varianza de una media de tratamiento. Estimar la varianza de una media de tratamiento suponiendo que la tasa de submuestreo se dobla (cuatro submuestreos, en vez de dos). ¿Da esto una apreciable ganancia en precisión? (Expresar la varianza estimada como un porcentaje de la observada).

Ejercicio 25.7.2. Utilizar los datos de la tabla 25.2 para calcular el error experimental como si no se hubiese superimpuesto un diseño a las parcelas. (Usar una media ponderada de cuadrados medios de filas, columnas y error experimental). ¿Cuál habría sido el error de muestreo si no se hubiera usado ningún control local? (Usar una media ponderada de los cuadrados medios de las mitades de las parcelas y el error de muestreo). ¿Cuál habría sido el error experimental sin el diseño y sin el control local? (Sumar los valores que se acaban de calcular para los errores experimental y muestral. Este incluye el error muestral sin control local dos veces de modo que el error muestral con control total debe restarse ahora). Calcular la desviación estándar de cada una de las tres varianzas que se acaba de calcular.

Ejercicio 25.7.3 ¿Cuál es el mínimo número de submuestras por media parcela de control local si se necesita estimar el error de muestreo y, al mismo tiempo, no perder la facilidad de cálculo?

Referencias

25.1. Cochran, W. G.: "The information supplied by the sampling results," *Ann. Appl. Biol.*, 25: 383-389 (1938).

25.2. Cochran, W. G.: *Sampling Techniques*, Wiley, Nueva York, 1953.

25.3. Chung, J. H., y D. B. DeLury: *Confidence Limits for the Hypergeometric Distribution*, University of Toronto Press, Toronto, Ontario, 1950.

25.4. Hansen, M. H. W. N. Hurwitz, y W. G. Madow: *Sample Survey Methods and Theory*, 2 vols., Wiley, Nueva York, 1953.

25.5. Jessen, R. J.: "Statistical investigation of a sample survey for obtaining farm facts," *Iowa Agr. Exp. Sta. Res. Bull.* 304, 1942.

25.6. Ladell, W. R. S.: "Field experiments on the control of wireworms," *Ann. Appl. Biol.*, 25: 341-382 (1938).