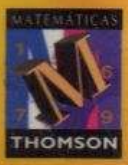
Muestreo:

Diseño y análisis

Sharon L. Lohr



Muestreo: Diseño y análisis

Sharon L. Lohr
Arizona State University

International Thomson Editores

México • Albany • Boston • Johannesburgo • Londres • Madrid • Melbourne • Nueva York
San Francisco • San Juan, PR • Santiago • São Paulo • Singapur • Tokio • Toronto • Washington

Contenido

Prefacio

A veces, las encuestas y las muestras parecen rodearnos. Muchas nos dan información valiosa; otras, por desgracia, están mal concebidas y aplicadas, de tal modo que sería mejor para la ciencia y la sociedad que no se hubieran hecho. Este libro es una guía para ver cuándo una muestra es válida o no, para diseñar y analizar muchas formas diversas de encuestas con muestreo.

El libro se centra en los aspectos estadísticos de la extracción y el análisis de una muestra. La forma de diseñar y verificar, de manera previa, un cuestionario, la construcción de un marco de muestreo y el entrenamiento de los investigadores de campo son temas muy importantes, pero que no serán tratados de manera amplia en este libro.

Escribí este libro pensando en una audiencia amplia, permitiendo cierta flexibilidad al elegir los temas por leer. Para dar lectura a la mayor parte de los capítulos del 1 al 6, usted debe estar familiarizado con las ideas de esperanza, distribuciones muestrales, intervalos de confianza y regresión lineal, temas considerados en la mayor parte de los cursos de introducción a la estadística. Dichos capítulos consideran los diseños básicos de muestreo, como el muestreo aleatorio simple, la estratificación, y el muestreo por conglomerados con probabilidades iguales y distintas de selección. Las secciones opcionales de la teoría estadística para estos diseños están marcadas con asteriscos; para estas secciones, usted debe estar familiarizado con el cálculo o la estadística matemática. El apéndice B proporciona un repaso de los conceptos de probabilidad utilizados en la teoría del muestreo probabilístico.

Los capítulos del 7 al 12 analizan aspectos que no aparecen en muchos otros libros de texto que tratan el tema de muestreo, como el análisis de las encuestas complejas, por ejemplo las realizadas por la Oficina de Censos de Estados Unidos o por Statistics Canada; los distintos enfoques del análisis de las encuestas, qué hacer cuando se presenta una ausencia de respuestas, y cómo realizar pruebas como la ji-cuadrada y el análisis de regresión con datos y encuestas complejas. La National Crime Victimization Survey (encuesta nacional a víctimas de crímenes, NCVS) se analiza, con detalle, como ejemplo de una encuesta compleja. Como en los casos complejos es difícil aplicar muchas de las fórmulas utilizadas para determinar los errores estándar en los diseños más sencillos de muestreo, se analizan algunos métodos con el empleo intensivo de la computadora para estimar las varianzas.

El libro es adecuado para un primer curso de muestreo con encuestas. Puede ser utilizado para un grupo de estudiantes de estadística o para un grupo de alumnos de comercio, sociología, psicología o biología que deseen aprender acerca del diseño y el análisis de datos a partir de las encuestas con muestreo. Los capítulos del 1 al 6 estudian los bloques básicos del muestreo y las secciones sin asterisco de los mismos capítulos proporcionarán el mate-

rial para un curso trimestral sobre muestreo. En mi curso semestral abarco las secciones sin asterisco de los capítulos 1 al 8, y algunos temas selectos de los demás capítulos. El material de los capítulos 9 al 12 puede ser cubierto casi en cualquier orden; los temas de estos capítulos se eligen de acuerdo a las necesidades de los estudiantes.

Existen dos clases de ejercicios en este libro: los que implican la crítica y el análisis de los datos obtenidos de encuestas reales o el diseño de encuestas propias, con esto se presenta al estudiante una amplia variedad de aplicaciones del muestreo y la segunda clase de ejercicios la constituyen los problemas matemáticos (indicados con asteriscos) que desarrollarán en el alumno el conocimiento teórico del tema.

Usted debe saber emplear un paquete estadístico de computadora o una hoja de cálculo para realizar los problemas de este libro. Le recomiendo que utilice un paquete estadístico como Splus, SAS o Minitab o una hoja de cálculo como Excel, Quattro Pro o Lotus 1-2-3 para los ejercicios. El paquete u hoja de cálculo que elija dependerá del tamaño y el nivel del grupo. En un curso trimestral que abarque los conceptos básicos del muestreo, bastará utilizar una hoja de cálculo. Algunos ejercicios de los capítulos posteriores requieren ciertos conocimientos sobre programación en computadora; creo que Splus es ideal para estos ejercicios, pues combina la capacidad de programación con las funciones existentes para el análisis estadístico. Los paquetes de muestreo, como SUDAAN (Shah *et al.*, 1995) y WesVarPC (Brick *et al.*, 1996), aunque valiosos para el practicante del muestreo, ocultan la estructura inherente de los cálculos a las personas que quieren estudiar el material. Por lo tanto, en este libro no me he basado en los paquetes de computadora existentes para el análisis de datos de encuestas, aunque analizamos varios de estos paquetes en la sección 9.6. Una vez que usted comprenda el funcionamiento de los distintos diseños y estimadores utilizados en el muestreo con encuestas, será fácil leer el manual del usuario para un paquete diseñado para realizar encuestas y utilizar el software; por el contrario, si usted confía en los paquetes como si fuesen cajas negras, será difícil saber si está llevando a cabo un análisis adecuado.

Seis características principales distinguen a este libro de otros textos dirigidos a los estudiantes de estadística y otras disciplinas que necesitan conocer métodos de muestreo.

- El libro tiene un contenido y un nivel flexibles. En los cursos que se imparten sobre muestreo se inscriben, por lo general, estudiantes que poseen distintos niveles de conocimiento estadístico. Al elegir las secciones adecuadas, este libro puede servir para una audiencia de alumnos de licenciatura que han llevado un curso de introducción a la estadística o para un primer curso de posgrado para estudiantes de estadística. El libro también es útil para una persona que realiza análisis de encuestas que desee aprender más acerca de los aspectos estadísticos de las encuestas y conocer algunos desarrollos recientes. Los ejercicios también son flexibles. Algunos de ellos enfatizan el dominio de la mecánica. Sin embargo, muchos animan al estudiante a pensar en los detalles del muestreo y a comprender la estructura del diseño de las muestras con mayor profundidad. Otros ejercicios son abiertos y motivan al alumno a continuar explorando estas ideas.

- He tratado de utilizar datos reales en la medida de lo posible; la "Compañía Acme" nunca aparece en este libro. Los ejemplos y ejercicios provienen de las ciencias sociales, la ingeniería, la agricultura, la ecología, la medicina y otras disciplinas; fueron elegidos para ilustrar la amplia gama de aplicaciones de los métodos de muestreo. Varios de los conjuntos de datos tienen variables adicionales a las cuales no se hace referencia en el texto; un instructor puede utilizarlas para ejercicios o variantes adicionales.

- He incorporado al texto la teoría basada en modelos y la teoría basada en la aleatorización, para ubicar los métodos de muestreo dentro del marco de referencia utilizado en otras áreas de la estadística. Muchos de los resultados importantes, logrados en los últimos 20 años de investigación en el área de muestreo, implican el uso de los modelos y la comprensión de ambos puntos de vista es esencial para el profesional encargado de realizar las encuestas. El punto de vista basado en los modelos aparece en la sección 2.8 y se desarrolla en los capítulos posteriores; sin embargo, esas secciones se pueden analizar en cualquier momento posterior al curso.

- Muchos temas de este libro, como la estimación de la varianza y el análisis de regresión de encuestas complejas, no aparecen en otros textos de este nivel. La amplia referencia de muestreo *Model Assisted Survey Sampling*, de Särndal, Swensson y Wretman tiene un nivel matemático mucho mayor.

- Este libro enfatiza la importancia de graficar los datos. El análisis gráfico de los datos de una encuesta se deja de lado con frecuencia, debido al gran tamaño de los conjuntos de datos y el énfasis en la teoría de aleatorización, pero este hecho puede conducir a análisis de datos fallidos.

- El diseño de las encuestas se enfatiza en todo el libro y se relaciona con los métodos para el análisis de los datos de una encuesta. La filosofía de este libro consiste en establecer que el diseño es, por mucho, el aspecto más importante de una encuesta; ninguna cantidad de análisis estadísticos puede compensar el uso de una encuesta mal diseñada. Los modelos motivan los diseños, las gráficas verifican la sensibilidad del diseño a los supuestos del modelo. Por ejemplo, en el capítulo 2 presentamos la fórmula usual para calcular el tamaño de la muestra. Pero también mostramos una gráfica para que el investigador observe la sensibilidad del tamaño de la muestra con respecto al valor de la varianza que se toma para la población de este caso.

Muchas personas han sido muy generosas por el apoyo y las sugerencias brindadas para la elaboración de este libro. Tengo una gran deuda con ellos, aunque debo reservarme el crédito por los defectos de la obra. Las siguientes personas revisaron o utilizaron varias versiones del manuscrito y proporcionaron valiosas sugerencias para mejorarlo: Jon Rao, Elizabeth Stasny, Fritz Scheuren, Nancy Heckman, Ted Chang, Steve MacEachern, Mark Conaway, Ron Christensen, Michael Hamada, Partha Lahiri y varios revisores anónimos: Dale Everson, Universidad de Idaho; James Gentle, George Mason University; Ruth Mickey, Universidad de Vermont; Sarah Nusser, Universidad Estatal de Iowa; N. G. Narasi nha Prasad, Universidad de Alberta, Edmonton; y Deborah Rumsey, Universidad Estatal de Kansas. Tuve valiosas discusiones y el apoyo de Jon Rao, Fritz Scheuren y Elizabeth Stasny. David Hubble y Marshall DeBerry me proporcionaron muchos consejos útiles sobre la National Crime Victimization Survey. Muchas gracias a Alexander Kugushev, Carolyn Crockett y al equipo de producción en Brooks/Cole por su ayuda, consejos y apoyo. Por último, quiero agradecer a Alastair Scott, cuya inspiradora clase de muestreo en la Universidad de Wisconsin me introdujo a las joyas del tema.

Sharon L. Lohr

Contenido

Prefacio

A veces, las encuestas y las muestras parecen rodearnos. Muchas nos dan información valiosa; otras, por desgracia, están mal concebidas y aplicadas, de tal modo que sería mejor para la ciencia y la sociedad que no se hubieran hecho. Este libro es una guía para ver cuándo una muestra es válida o no, para diseñar y analizar muchas formas diversas de encuestas con muestreo.

El libro se centra en los aspectos estadísticos de la extracción y el análisis de una muestra. La forma de diseñar y verificar, de manera previa, un cuestionario, la construcción de un marco de muestreo y el entrenamiento de los investigadores de campo son temas muy importantes, pero que no serán tratados de manera amplia en este libro.

Escribí este libro pensando en una audiencia amplia, permitiendo cierta flexibilidad al elegir los temas por leer. Para dar lectura a la mayor parte de los capítulos del 1 al 6, usted debe estar familiarizado con las ideas de esperanza, distribuciones muestrales, intervalos de confianza y regresión lineal, temas considerados en la mayor parte de los cursos de introducción a la estadística. Dichos capítulos consideran los diseños básicos de muestreo, como el muestreo aleatorio simple, la estratificación, y el muestreo por conglomerados con probabilidades iguales y distintas de selección. Las secciones opcionales de la teoría estadística para estos diseños están marcadas con asteriscos; para estas secciones, usted debe estar familiarizado con el cálculo o la estadística matemática. El apéndice B proporciona un repaso de los conceptos de probabilidad utilizados en la teoría del muestreo probabilístico.

Los capítulos del 7 al 12 analizan aspectos que no aparecen en muchos otros libros de texto que tratan el tema de muestreo, como el análisis de las encuestas complejas, por ejemplo las realizadas por la Oficina de Censos de Estados Unidos o por Statistics Canada; los distintos enfoques del análisis de las encuestas, qué hacer cuando se presenta una ausencia de respuestas, y cómo realizar pruebas como la ji-cuadrada y el análisis de regresión con datos y encuestas complejas. La National Crime Victimization Survey (encuesta nacional a víctimas de crímenes, NCVS) se analiza, con detalle, como ejemplo de una encuesta compleja. Como en los casos complejos es difícil aplicar muchas de las fórmulas utilizadas para determinar los errores estándar en los diseños más sencillos de muestreo, se analizan algunos métodos con el empleo intensivo de la computadora para estimar las varianzas.

El libro es adecuado para un primer curso de muestreo con encuestas. Puede ser utilizado para un grupo de estudiantes de estadística o para un grupo de alumnos de comercio, sociología, psicología o biología que deseen aprender acerca del diseño y el análisis de datos a partir de las encuestas con muestreo. Los capítulos del 1 al 6 estudian los bloques básicos del muestreo y las secciones sin asterisco de los mismos capítulos proporcionarán el mate-

rial para un curso trimestral sobre muestreo. En mi curso semestral abarco las secciones sin asterisco de los capítulos 1 al 8, y algunos temas selectos de los demás capítulos. El material de los capítulos 9 al 12 puede ser cubierto casi en cualquier orden; los temas de estos capítulos se eligen de acuerdo a las necesidades de los estudiantes.

Existen dos clases de ejercicios en este libro: los que implican la crítica y el análisis de los datos obtenidos de encuestas reales o el diseño de encuestas propias, con esto se presenta al estudiante una amplia variedad de aplicaciones del muestreo y la segunda clase de ejercicios la constituyen los problemas matemáticos (indicados con asteriscos) que desarrollarán en el alumno el conocimiento teórico del tema.

Usted debe saber emplear un paquete estadístico de computadora o una hoja de cálculo para realizar los problemas de este libro. Le recomiendo que utilice un paquete estadístico como Splus, SAS o Minitab o una hoja de cálculo como Excel, Quattro Pro o Lotus 1-2-3 para los ejercicios. El paquete u hoja de cálculo que elija dependerá del tamaño y el nivel del grupo. En un curso trimestral que abarque los conceptos básicos del muestreo, bastará utilizar una hoja de cálculo. Algunos ejercicios de los capítulos posteriores requieren ciertos conocimientos sobre programación en computadora; creo que Splus es ideal para estos ejercicios, pues combina la capacidad de programación con las funciones existentes para el análisis estadístico. Los paquetes de muestreo, como SUDAAN (Shah *et al.*, 1995) y WesVarPC (Brick *et al.*, 1996), aunque valiosos para el practicante del muestreo, ocultan la estructura inherente de los cálculos a las personas que quieren estudiar el material. Por lo tanto, en este libro no me he basado en los paquetes de computadora existentes para el análisis de datos de encuestas, aunque analizamos varios de estos paquetes en la sección 9.6. Una vez que usted comprenda el funcionamiento de los distintos diseños y estimadores utilizados en el muestreo con encuestas, será fácil leer el manual del usuario para un paquete diseñado para realizar encuestas y utilizar el software; por el contrario, si usted confía en los paquetes como si fuesen cajas negras, será difícil saber si está llevando a cabo un análisis adecuado.

Seis características principales distinguen a este libro de otros textos dirigidos a los estudiantes de estadística y otras disciplinas que necesitan conocer métodos de muestreo.

- El libro tiene un contenido y un nivel flexibles. En los cursos que se imparten sobre muestreo se inscriben, por lo general, estudiantes que poseen distintos niveles de conocimiento estadístico. Al elegir las secciones adecuadas, este libro puede servir para una audiencia de alumnos de licenciatura que han llevado un curso de introducción a la estadística o para un primer curso de posgrado para estudiantes de estadística. El libro también es útil para una persona que realiza análisis de encuestas que desee aprender más acerca de los aspectos estadísticos de las encuestas y conocer algunos desarrollos recientes. Los ejercicios también son flexibles. Algunos de ellos enfatizan el dominio de la mecánica. Sin embargo, muchos animan al estudiante a pensar en los detalles del muestreo y a comprender la estructura del diseño de las muestras con mayor profundidad. Otros ejercicios son abiertos y motivan al alumno a continuar explorando estas ideas.

- He tratado de utilizar datos reales en la medida de lo posible; la "Compañía Acme" nunca aparece en este libro. Los ejemplos y ejercicios provienen de las ciencias sociales, la ingeniería, la agricultura, la ecología, la medicina y otras disciplinas; fueron elegidos para ilustrar la amplia gama de aplicaciones de los métodos de muestreo. Varios de los conjuntos de datos tienen variables adicionales a las cuales no se hace referencia en el texto; un instructor puede utilizarlas para ejercicios o variantes adicionales.

- He incorporado al texto la teoría basada en modelos y la teoría basada en la aleatorización, para ubicar los métodos de muestreo dentro del marco de referencia utilizado en otras áreas de la estadística. Muchos de los resultados importantes, logrados en los últimos 20 años de investigación en el área de muestreo, implican el uso de los modelos y la comprensión de ambos puntos de vista es esencial para el profesional encargado de realizar las encuestas. El punto de vista basado en los modelos aparece en la sección 2.8 y se desarrolla en los capítulos posteriores; sin embargo, esas secciones se pueden analizar en cualquier momento posterior al curso.

- Muchos temas de este libro, como la estimación de la varianza y el análisis de regresión de encuestas complejas, no aparecen en otros textos de este nivel. La amplia referencia de muestreo *Model Assisted Survey Sampling*, de Särndal, Swensson y Wretman tiene un nivel matemático mucho mayor.

- Este libro enfatiza la importancia de graficar los datos. El análisis gráfico de los datos de una encuesta se deja de lado con frecuencia, debido al gran tamaño de los conjuntos de datos y el énfasis en la teoría de aleatorización, pero este hecho puede conducir a análisis de datos fallidos.

- El diseño de las encuestas se enfatiza en todo el libro y se relaciona con los métodos para el análisis de los datos de una encuesta. La filosofía de este libro consiste en establecer que el diseño es, por mucho, el aspecto más importante de una encuesta; ninguna cantidad de análisis estadísticos puede compensar el uso de una encuesta mal diseñada. Los modelos motivan los diseños, las gráficas verifican la sensibilidad del diseño a los supuestos del modelo. Por ejemplo, en el capítulo 2 presentamos la fórmula usual para calcular el tamaño de la muestra. Pero también mostramos una gráfica para que el investigador observe la sensibilidad del tamaño de la muestra con respecto al valor de la varianza que se toma para la población de este caso.

Muchas personas han sido muy generosas por el apoyo y las sugerencias brindadas para la elaboración de este libro. Tengo una gran deuda con ellos, aunque debo reservarme el crédito por los defectos de la obra. Las siguientes personas revisaron o utilizaron varias versiones del manuscrito y proporcionaron valiosas sugerencias para mejorarlo: Jon Rao, Elizabeth Stasny, Fritz Scheuren, Nancy Heckman, Ted Chang, Steve MacEachern, Mark Conaway, Ron Christensen, Michael Hamada, Partha Lahiri y varios revisores anónimos: Dale Everson, Universidad de Idaho; James Gentle, George Mason University; Ruth Mickey, Universidad de Vermont; Sarah Nusser, Universidad Estatal de Iowa; N. G. Narasi nha Prasad, Universidad de Alberta, Edmonton; y Deborah Rumsey, Universidad Estatal de Kansas. Tuve valiosas discusiones y el apoyo de Jon Rao, Fritz Scheuren y Elizabeth Stasny. David Hubble y Marshall DeBerry me proporcionaron muchos consejos útiles sobre la National Crime Victimization Survey. Muchas gracias a Alexander Kugushev, Carolyn Crockett y al equipo de producción en Brooks/Cole por su ayuda, consejos y apoyo. Por último, quiero agradecer a Alastair Scott, cuya inspiradora clase de muestreo en la Universidad de Wisconsin me introdujo a las joyas del tema.

Sharon L. Lohr

1

Introducción

Cuando la estadística no se basa en cálculos precisos, confunde en vez de guiarnos. La mente se deja llevar fácilmente por la falsa apariencia de exactitud que la estadística mantiene en sus errores, y confiadamente adopta errores escondidos bajo la forma de una verdad matemática.

—Alexis de Tocqueville, *Democracy in America*

1.1 Una controversia muestral

El libro *Women and Love: A Cultural Revolution in Progress* (1987), de Shere Hite, tiene varios resultados ampliamente citados:

- El 84% de las mujeres "no están satisfechas emocionalmente con sus relaciones" (página 804).
- El 70% de las mujeres "con cinco o más años de casadas tienen relaciones sexuales fuera del matrimonio" (página 856).
- El 95% de las mujeres "informan de diversas maneras de acoso emocional y psicológico por parte de los hombres con los que mantuvieron alguna relación sentimental" (página 810).
- El 84% de las mujeres informan de ciertos sentimientos de superioridad por parte de los hombres con los que mantuvieron relaciones sentimentales (página 809).

El libro fue muy criticado en los artículos de periódicos y revistas a lo largo de todo Estados Unidos. Por ejemplo, en la revista *Time*, el artículo "Back Off, Buddy", citado en la portada del 12 de octubre de 1997, denominó a las conclusiones del estudio de Hite como "dudosas" y "de valor limitado".

¿Por qué fue tan criticado el estudio de Hite? ¿Fue incorrecto que citara a las mujeres que sentían que los hombres de sus vidas se resistían a tratarlas como iguales, féminas que posiblemente no habían tenido la oportunidad de hablar anteriormente? ¿Era incorrecto informar de los porcentajes de estas mujeres que no se sentían felices con la relación que llevaban con los hombres?

Por supuesto que no. La investigación de Hite permitió a las mujeres analizar una visión de sus experiencias y reflejó la riqueza de las experiencias de estas mujeres de una forma que no lo lograría un examen de opción múltiple. El error de Hite fue generalizar estos resultados a *todas* las mujeres, hayan participado en la encuesta o no, y afirmar que los porcentajes se aplicaban a todas las mujeres. Las siguientes características de la encuesta no la hacen adecuada para generalizar los resultados a todas las mujeres.

- La muestra fue autoelegida; es decir, las receptoras de los cuestionarios decidieron si debían estar o no en la muestra. Hite envió 100,000 cuestionarios, de los cuales regresó el 4.5%.
- Los cuestionarios fueron enviados a grupos de mujeres profesionistas, centros de asesoría, sociedades eclesiásticas y centros de ciudadanos de edad avanzada. Los miembros podían tener distintas visiones políticas, pero podrían unirse en un grupo "sólo mujeres", de modo que sus puntos de vista podían diferir de los de otras mujeres en Estados Unidos.
- La encuesta tiene 127 preguntas de ensayo y la mayoría de ellas tiene varias partes. ¿Quién tendería a regresar tal encuesta?
- Muchas de las preguntas son vagas, con palabras como *amor*. El concepto de amor tiene tantas interpretaciones como personas existentes, lo que hace imposible establecer una sola interpretación a cualquier estadística que se proponga indicar cuántas mujeres están "enamoradas". Esta redacción serviría para evocar las viñetas de todo el libro, pero dificulta la interpretación de los porcentajes.
- Muchas de las cuestiones estan inducidas; sugieren a la entrevistada la respuesta que debe dar. Por ejemplo "¿Te ve tu esposo/amante como igual? ¿O hay veces en que parece tratarte como inferior? ¿Te deja fuera de las decisiones? ¿Actúa como superior?" (página 795)

Hite escribió lo siguiente: "¿Acaso la investigación que no está basada en una muestra probabilística o aleatoria da el derecho de generalizar los resultados de un estudio a una población en gran escala? Si un estudio es lo bastante grande, si la muestra es lo bastante amplia y si uno generaliza con cuidado, sí" (página 778). La mayoría de los estadísticos encargados de realizar encuestas contestaría a la pregunta de Hite con un rotundo no. En la encuesta de Hite, como las mujeres que recibieron cuestionarios fueron elegidas a propósito y un porcentaje extremadamente pequeño de ellas regresó los cuestionarios, las estadísticas calculadas a partir de estos datos no sirven para indicar la actitud de todas las mujeres en Estados Unidos. La muestra final *no es representativa* de las mujeres de toda la Unión Americana y las estadísticas sólo sirven para describir a las mujeres que contestaron la encuesta.

Hite afirma que los resultados de la muestra se podían generalizar, debido a características como el perfil de edad, nivel educativo y de empleo de las mujeres en la muestra, concordantes con los de la población femenina en Estados Unidos. Pero las mujeres de la muestra diferían en un aspecto importante: estaban dispuestas a tomarse el tiempo de contestar un largo cuestionario relacionado con el acoso masculino y a proporcionar una gran cantidad de información personal para una investigación. Es de esperar que, en cada grupo de edad y clase socioeconómica, las mujeres que optaron por revelar tales datos tengan experiencias distintas a las de las mujeres que optaron por no participar en la encuesta.

1.2 Requisitos de una buena muestra

En la película *Magic Town*, el investigador de opinión pública interpretado por James Stewart, descubrió un pueblo que tenía exactamente las mismas características que todo Estados Unidos. Grandview poseía exactamente la misma proporción de personas que votaban por los republicanos, la misma proporción de personas bajo la línea de pobreza, igual proporción de mecánicos automotrices, etcétera, que Estados Unidos visto como un todo. El personaje de Stewart sólo tenía que entrevistar a las personas de Grandview para saber cuál era la opinión pública en la Unión Americana.

Una muestra perfecta sería como el pueblo de Grandview: una versión a escala de la población, que reflejaría cada una de las características de toda la población. Por supuesto, una muestra perfecta como ésta no puede existir para poblaciones complejas (aunque existiera, no sabríamos que es perfecta sin antes medir a toda la población). Pero una buena muestra reproduce las características de interés que existen en la población de la manera más cercana posible. Esta muestra será **representativa**, en el sentido de que cada unidad muestreada representará las características de una cantidad conocida de unidades en la población.

Necesitamos algunas definiciones para precisar el concepto de una buena muestra.

Unidad de observación Es el objeto sobre el cual se realiza una medición. Ésta es la unidad básica de observación, a veces llamada **elemento**. En los estudios de poblaciones humanas, con frecuencia ocurre que las unidades de observación son los individuos.

Población objetivo Es la colección completa de observaciones que deseamos estudiar. La definición de la población objetivo es una parte importante, y con frecuencia difícil, del estudio. Por ejemplo, en una encuesta política, ¿la población objetivo deberían ser todos los adultos que pueden votar? ¿Todos los votantes registrados? ¿Todas las personas que votaron en la última elección? La elección de la población objetivo afectará profundamente a las estadísticas resultantes.

Muestra Es un subconjunto de una población.

Población muestreada Es la colección de todas las unidades de observación posibles que podrían extraerse en una muestra; en otras palabras, es la población de donde se extrae la muestra.

Unidad de muestreo Es la unidad donde realizamos la muestra. Por ejemplo, podríamos querer estudiar a las personas, pero no tenemos una lista de todos los individuos que pertenecen a la población objetivo. En vez de esto, las familias sirven como las unidades de muestreo y las unidades de observación son los individuos que viven en una familia.

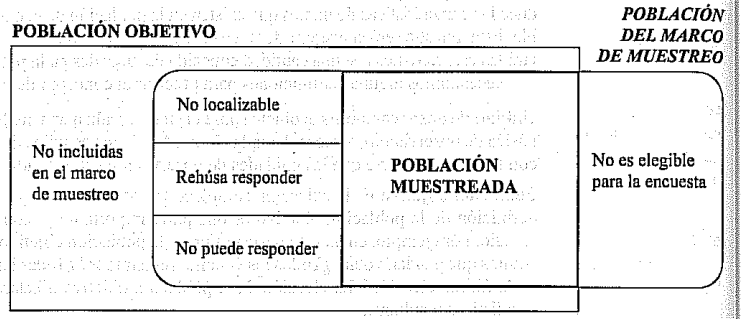
Marco de muestreo Es la lista de las unidades de muestreo. Para las encuestas telefónicas, el marco de muestreo podría ser una lista de todos los números telefónicos residenciales de la ciudad; para las entrevistas personales, una lista de las direcciones de todas las calles; para una encuesta de agricultura, una lista de todas las granjas o un mapa de las áreas que contienen granjas.

En una encuesta ideal, la población muestreada será idéntica a la población objetivo, pero este ideal se cumple muy rara vez. En las encuestas de personas, la población muestreada es, por lo general, menor que la población objetivo. Como muestra la figura 1.1, no todas las personas de la población objetivo se incluyen en el marco de muestreo y varios individuos no contestarán la encuesta.

En el estudio de Hite, una característica de interés consistía en saber el porcentaje de mujeres acosadas en su relación. Una mujer era un elemento. La población objetivo era formada por todas las mujeres adultas en Estados Unidos. La población muestreada por Hite estaba compuesta por las personas pertenecientes a las organizaciones de mujeres que regresarían los cuestionarios. En consecuencia, sólo se podían hacer inferencias con respecto a la población muestreada, no sobre la población de todas las mujeres adultas en la Unión Americana.

La National Crime Victimization Survey [NCVS en español es la Encuesta Nacional a Víctimas de Crímenes] es realizada, en la actualidad, para estudiar los crímenes. Es elaborada

FIGURA 1.1
La población objetivo y la población muestreada en una encuesta telefónica de posibles votantes. No todas las familias tienen teléfono, de modo que varias personas de la población objetivo de posibles votantes no tendrán asociado un número telefónico en el marco de muestreo. En algunas casas con teléfono, los residentes no están registrados para votar y, por lo tanto, no son elegibles para la encuesta. Algunas personas elegibles en la población del marco de muestreo no responden porque no pueden ser contactadas, algunas se rehúsan a contestar la encuesta y algunas podrían estar enfermas o incapacitadas para responder.



por la Oficina de Censos de Estados Unidos y la Oficina de Estadísticas del Departamento de Justicia. Si la característica de interés es la cantidad total de familias en Estados Unidos que han sido víctimas de crímenes durante el año pasado, los elementos son las familias, la población objetivo son todas las familias de Estados Unidos y la población muestreada consta de las familias en el marco de muestreo, construido a partir de la información de censos y permisos de edificación, que están "en casa" y están de acuerdo en contestar las preguntas.

El objetivo de la National Pesticide Survey (en español, la Encuesta Nacional sobre Pesticidas) realizada por la Agencia de Protección Ambiental, era estudiar los pesticidas y nitratos en los pozos de agua potable a nivel nacional. La población objetivo estaba compuesta por todos los pozos pertenecientes a los sistemas comunitarios y domésticos en Estados Unidos. La población muestreada consistía en todos los sistemas comunitarios (enumerados en el sistema de datos federales) y los pozos domésticos identificables fuera de las reservas gubernamentales, pertenecientes a familias dispuestas a cooperar con la encuesta.

Las encuestas de opinión pública se realizan con frecuencia para predecir el candidato que ganará en las próximas elecciones. La población objetivo está formada por las personas que votarán en la próxima elección; la población muestreada está formada con frecuencia por las personas que pueden ser localizadas por teléfono y que dicen estar dispuestas a votar en la próxima elección. Pocas encuestas de las que se realizan en Estados Unidos incluyen a Alaska o Hawaii o a las personas que se encuentran hospitalizadas, viven en dormitorios o se encuentran recluidas en las cárceles; toda esta gente no forma parte del marco de muestreo o de la población muestreada.

1.3 Sesgo de selección

Una buena muestra estará a salvo de presentar un **sesgo de selección**; éste ocurre cuando alguna parte de la población objetivo no está en la población muestreada. Si una encuesta

diseñada para estudiar el ingreso de las familias omite a las personas que se encuentran en una situación transitoria, las estimaciones de la encuesta del ingreso familiar promedio serían probablemente muy grandes. Con frecuencia, una **muestra de conveniencia** es sesgada, pues las unidades más fáciles de elegir o las que más probablemente respondan a la encuesta no son representativas de las unidades más difíciles de elegir o de las unidades que no contesten la encuesta. Los siguientes ejemplos indican algunas dè las formas en que puede ocurrir un sesgo de selección.

- El uso de un procedimiento de selección de la muestra que, sin saberlo los investigadores, dependa de cierta característica asociada a las propiedades de interés. Por ejemplo, unos investigadores extrajeron una muestra de conveniencia de adolescentes, para estudiar la frecuencia con que los adolescentes hablan con sus padres y maestros acerca del SIDA. Pero los adolescentes dispuestos a hablar con los investigadores acerca del SIDA tenían más probabilidad de estar dispuestos a hablar con otras figuras de autoridad acerca del SIDA. Los investigadores sólo promediaron las cantidades de tiempo mencionadas por los adolescentes en la muestra como el tiempo que ocupaban en hablar con sus padres y maestros y probablemente sobreestimaron la cantidad de comunicación existente entre padres, maestros y adolescentes en la población.
- La elección deliberada o que busca una muestra "representativa". Si queremos estimar la cantidad promedio que gasta una persona que va de compras a un centro comercial y extraemos una muestra entre los compradores que parecen haber gastado una cantidad "promedio", habremos elegido de manera deliberada una muestra para confirmar nuestra opinión anterior. Este tipo de muestra se llama a veces una **muestra de juicio**: el investigador emplea su propio juicio para elegir las unidades específicas que debe incluir en la muestra.
- Los errores en la especificación de la población objetivo. Por ejemplo, todas las encuestas en la elección primaria para gobernar Arizona por parte de los demócratas de ese estado en 1994 predecían que el candidato Eddie Basha quedaría atrás de su principal contrincante por al menos 9 puntos porcentuales. En la elección, Basha ganó 37% de los votos; los otros dos candidatos recibieron el 35 y el 28%, respectivamente. Un problema consistió en que muchos votantes estaban indecisos al momento de realizar las encuestas. Otro problema fue que la población objetivo para las encuestas estaba formada por los votantes registrados que habían votado en las elecciones primarias anteriores y que estaban interesados en ésta. Sin embargo, en la elección primaria, Basha obtuvo un fuerte apoyo en las áreas rurales, con grupos demográficos que no habían votado antes y que, por tanto, no eran un objetivo de las encuestas.
- No incluir a toda la población objetivo en el marco de muestreo, lo que se llama **subcobertura**. Muchas encuestas de gran tamaño utilizan los censos decenales de Estados Unidos para construir el marco de muestreo, pero el censo no enumera un gran número de unidades familiares, por lo que se deja de contar a varios grupos de población. Fay *et al.* (1988) estiman que el censo de 1980 no abarcó a 8% de los hombres negros. Así, cualquier encuesta que tome los datos del censo de 1980 como la única base para construir un marco de muestreo, automáticamente no considerará al 8% de los hombres negros; ese error ocurrirá antes de que empiece a realizarse la encuesta.
- La sustitución de un miembro conveniente de una población por un miembro designado que no está disponible. Por ejemplo, si no hay nadie en casa de la familia designada, un representante de campo podría ir a la puerta siguiente. En un estudio de la vida salvaje, el investigador podría sustituir un área aledaña a un camino por un área menos accesible. En cada caso, es muy probable que las unidades muestreadas difieran en varias caracte-

rísticas de las unidades que no entraron en la muestra. Puede ser más probable que la familia sustituida tenga un miembro que no trabaja fuera de casa, que en el caso de la familia originalmente elegida. El área junto al camino podría tener menos ranas que el área más difícil de alcanzar.

- No poder obtener respuestas de toda la muestra elegida. La **ausencia de respuestas** distorsiona los resultados de muchas encuestas, incluso las diseñadas de manera cuidadosa, con lo que se minimizan otras fuentes de sesgo de selección. Con frecuencia, los sujetos que no responden a la encuesta difieren de manera crítica de los que sí responden, pero el alcance de esta diferencia se desconoce, a menos que se obtenga, posteriormente, alguna información relativa de las personas que no contestaron la encuesta. Muchas encuestas presentadas en periódicos o revistas de investigación tienen pésimas tasas de respuesta; en algunos casos, la tasa de respuesta es tan baja como el 10%. Es difícil pensar cómo se pueden generalizar los resultados a la población cuando el 90% de la muestra objetivo no puede ser alcanzada o se rehúsa a participar.

La Adolescent Health Database Survey (en español Encuesta de Salud en Adolescentes) fue diseñada para obtener una muestra representativa de los estudiantes de bachillerato en las escuelas públicas del estado de Minnesota (Remafedi *et al.*, 1992). En total, el 49% de los distritos escolares invitados a participar en la encuesta acordaron participar. La tasa de respuesta varió con el tamaño del distrito escolar:

En cada uno de los distritos escolares participantes, se distribuyeron las encuestas a los estudiantes, cuya participación era voluntaria. De las 52,553 encuestas distribuidas a los alumnos, 36,741 fueron concluidas y regresadas, esto produjo una tasa de respuesta del 70%. La encuesta consistía en hacer preguntas acerca de los hábitos de salud, creencia religiosa, nivel sicosocial y orientación sexual. Es posible que existieran distintos niveles de salud y actividad entre los distritos escolares que respondieron a la encuesta y los que no lo hicieron. Es todavía más probable que los estudiantes que respondieron a la encuesta tuvieran en promedio un perfil de salud distinto al de los estudiantes que no contestaron a la encuesta.

Tipo de distrito escolar	Tasa de participación (%)
Urbano	100
Metropolitano suburbano	25
No metropolitano con más de 2000 estudiantes	62
No metropolitano con 1000-1999 estudiantes	27
No metropolitano con 500-999 estudiantes	61
No metropolitano con menos de 500 estudiantes	53

Muchos de los estudios de comparación entre los sujetos que responden y los que no lo hacen han encontrado diferencias entre los dos grupos. En el Women's Health Study (en español, Estudio de Salud de la Mujer) del estado de Iowa, 41,836 mujeres respondieron a un cuestionario enviado por correo en 1986. Bisgard *et al.* (1994) compararon a las mujeres que respondieron con las 55 323 que no lo hicieron al revisar los datos en el Registro Estatal de Salud; encontraron que la tasa de mortandad ajustada por edades y la tasa de ataques de cáncer era significativamente mayor en las mujeres que no respondieron a la encuesta.

- Permitir que la muestra conste sólo de voluntarios. Tal es el caso de las encuestas que llevan a cabo la radio y la televisión, las estadísticas de tales encuestas no son confiables. La CBS News realizó una encuesta de recepción de llamadas inmediatamente después del discurso State of the Union del presidente Bush el 29 de enero de 1992. Los comentaristas Dan Rather y Connie Chung tuvieron el cuidado de decir que esta muestra "no era científica", pero el canal presentó los porcentajes de espectadores con distintas vi-

siones como si hubieran surgido de una encuesta sólida desde el punto de vista estadístico. Casi 315,000 personas llamaron para responder a lo que el *New York Times* llamó "la más grande muestra sesgada en la historia de las encuestas instantáneas", aunque muchas más trataron de responder; las computadoras de AT&T registraron casi 25 millones de intentos por lograr una conexión. Los ratings de Nielsen estimaron que cerca de 9 millones de familias sintonizaron el programa de CBS, lo que indica que muchos individuos u organizaciones trataron de realizar varias llamadas. En una encuesta de recepción de llamadas siempre existe la posibilidad de que una cierta organización desvíe los resultados al monopolizar el número al cual se llama de manera gratuita.

EJEMPLO 1.1 Muchas encuestas tienen más de uno de estos problemas. *The Literary Digest* (1932, 1936a, b, c) realizaba encuestas para predecir el resultado de la elección presidencial en Estados Unidos desde 1912; sus encuestas alcanzaron una cierta reputación, pues habían predicho al ganador correcto en todas las elecciones entre 1912 y 1932. En 1932, por ejemplo, la encuesta predijo que Roosevelt recibiría el 56% del voto popular y 474 votos en el colegio electoral; en la elección real, recibió el 58% del voto popular y 472 votos en el colegio electoral.

Con tal récord de precisión, no es sorprendente que los editores de *The Literary Digest* tuvieran un alto grado de confianza en sus métodos de elaboración de las encuestas para 1936. Al lanzar la encuesta de 1936, dijeron:

> La encuesta representa una evolución y perfección constantes durante 30 años. Con base en los métodos de "muestreo comercial" utilizados durante más de un siglo por las agencias de publicidad para impulsar las ventas de libros, nuestra lista de correo ha sido extraída de cada directorio telefónico de Estados Unidos, de las listas de socios de los clubes y las asociaciones, de los directorios de cada ciudad, de las listas de votantes registrados, del correo clasificado y de los datos de empleo. (1936a, 3)

El 31 de octubre, la encuesta predijo que el republicano Alf Landon recibiría el 55% del voto popular, comparada con el 41% para el presidente Roosevelt. El artículo "Landon, 1,293,669; Roosevelt, 972,897: Final Returns in The Digest's Poll of Ten Million Voters" contenía esta afirmación: "no afirmamos nuestra infalibilidad. Nosotros no acuñamos la frase 'precisión imprudente' que se ha aplicado de manera tan ligera a nuestras encuestas" (1936b). Es bueno que no afirmaran su infalibilidad; en la elección, Roosevelt recibió 61% de los votos; Landon, 37%.

¿Qué falló? Un problema pudo haber sido una subcobertura en el marco de muestreo, que se basaba fuertemente en los directorios telefónicos y las listas de registro de automóviles; el marco se utilizaba con fines publicitarios, al igual que para las encuestas. Las familias con teléfono o automóvil en 1936 eran más pudientes que las demás y la opinión acerca de la política económica de Roosevelt se relacionaba, por lo general, con la clase económica de quien respondía. Pero el sesgo en el marco de muestreo no explica toda la discrepancia. El análisis post mortem de la encuesta por parte de Squire (1988) y Calahan (1989) indica que incluso las personas con auto y teléfono tenían la tendencia de favorecer a Roosevelt, aunque no al grado con que las personas sin auto ni teléfono lo apoyaron.

Es probable que la baja tasa de respuesta a la encuesta fuese la causante de gran parte del error. Se enviaron por correo *diez millones* de cuestionarios y 2.3 millones lo regresaron; una muestra enorme pero una tasa de respuesta menor al 25%. En Allentown, Pennsylvania, por ejemplo, la encuesta fue enviada por correo a todos los votantes registrados, pero los resultados de la encuesta en Allentown seguían siendo incorrectos, pues sólo se regresó un tercio de las papeletas. Squire (1988) informa que es más probable que las personas que apoyaron a

Landon regresaran los cuestionarios; de hecho, muchos de los que apoyaron a Roosevelt ni siquiera recordaban haber recibido una encuesta, aunque estuvieran en la lista de correo.

Una lección del caso de la encuesta de *The Literary Digest* es que el tamaño de una muestra no garantiza su precisión. Los editores de *Digest* se complacían de haber enviado cuestionarios a más de una cuarta parte de todos los votantes registrados y obtuvieron una enorme muestra de 2.3 millones de personas. Pero las muestras grandes no representativas pueden comportarse tan mal como las muestras pequeñas no representativas. Una muestra grande no representativa puede hacer más daño que una pequeña, pues mucha gente piensa que las muestras grandes siempre son mejores que las pequeñas. El diseño de la encuesta es mucho más importante que el tamaño absoluto de la muestra. ■

¿Qué tan buenas son las muestras con sesgo de selección? Es preferible tener muestras sin sesgo de selección, que sirvan como un microcosmos de la población. Cuando el interés principal es estimar el número total de víctimas de crímenes violentos en Estados Unidos o el porcentaje de posibles votantes en el Reino Unido que pretenden votar por el Partido Laborista en la próxima elección, un sesgo de selección severo puede invalidar las estimaciones de la muestra.

Sin embargo, las muestras a propósito o de juicio pueden proporcionar información valiosa, particularmente en las primeras etapas de una investigación. Teichman *et al.* (1993) tomaron muestras de suelo a lo largo de la carretera interestatal 880 en Alameda County, California, para determinar la cantidad de plomo en las casas y parques cerca de la carretera. Al tomar las muestras, se concentraron en las áreas donde podían jugar los niños y donde el polvo podía ir hacia las casas. El esquema de muestreo a propósito permitió justificar la conclusión del estudio: "la contaminación por plomo del suelo urbano en el área occidental de la bahía, en la zona metropolitana de San Francisco es alta y excede los niveles de desechos peligrosos en muchos sitios". En este estudio, un esquema de muestreo sin sesgo de selección sólo sería necesario si los investigadores quisieran generalizar el porcentaje estimado de sitios contaminados relacionados con toda el área.

1.4 Sesgo de medición

Una buena muestra tiene respuestas precisas para los puntos de interés. El **sesgo de medición** ocurre cuando el instrumento con el que se mide tiene una tendencia a diferir del valor verdadero en alguna dirección. Como en el caso del sesgo de selección, el sesgo de medición debe ser considerado y minimizado en la etapa de diseño de la encuesta; ningún análisis estadístico revelará, por ejemplo, que la pesa añadió de manera errónea 5 kilogramos a cada persona en un estudio de salud.

El sesgo de medición es una preocupación en todas las encuestas y puede ser traicionero. Por ejemplo, en muchos estudios de vegetación, las áreas de muestreo se dividen en terrenos más pequeños. Se eligen algunos terrenos y se registra el número de plantas en cada terreno. Cuando una planta está cerca de la frontera de la región, el investigador de campo debe decidir si la incluye o no en la cuenta. Es probable que una persona que incluya en la cuenta a todas las plantas cerca o en la frontera produzca una estimación demasiado alta del número total de plantas en el área, pues podría contar dos veces algunas plantas. Duce *et al.* (1972) informa de concentraciones de metales, lípidos e hidrocarburos clorados en los 100 micrómetros superiores en la bahía de Narragansett, concentraciones que son de 1.5 a 50 veces más grandes que las existentes en el agua a 20 centímetros de la superficie. Al estudiar el transporte de contaminantes de las aguas costeras a las aguas más profundas en el océano, un esquema de muestreo que ignore este efecto de frontera podría subestimar la cantidad transportada.

A veces es inevitable el sesgo de medición. En el North American Breeding Bird Survey (en español Estudio de Aves Norteamericanas), los observadores se detenían cada media milla en ciertas rutas designadas y contaban todas las aves que oían cantar o hacer ruido o que eran vistas dentro de un radio de un cuarto de milla (Droege 1990). El número de aves es casi siempre una subestimación del número de aves que se encuentran en el área; es posible utilizar modelos estadísticos para ajustar el sesgo de medición. Si los datos se reúnen con el mismo procedimiento y con observadores capacitados de manera similar de un año a otro, el estudio se puede emplear para estimar las tendencias de población de las diversas especies, pues se espera que los sesgos de distintos años sean similares y se puedan cancelar al calcular las diferencias de un año a otro.

La obtención de respuestas precisas es un reto en todo tipo de encuestas, pero en particular en las que se trabaja con personas:

- A veces, las personas no dicen la verdad. En una encuesta agrícola, los granjeros en un área con programas de apoyo alimenticio podrían informar una menor cosecha, esperando un mayor apoyo alimenticio. La obtención de respuestas veraces es un reto importante en las encuestas que implican temas sensibles, como las encuestas en torno al uso de drogas.
- Las personas no siempre comprenden las preguntas. Muchas personas que viven en Estados Unidos se impresionaron con los resultados de una encuesta de Roper en 1993, la cual informaba que el 25% de los americanos no creían que el Holocausto realmente hubiese ocurrido. Al eliminar la estructura de doble negativa de la cuestión y formular de nuevo la pregunta, sólo el 1% pensaba que "posiblemente... el exterminio nazi hacia los judíos nunca ocurrió".
- Las personas olvidan. Un problema con el que se enfrentaron los diseñadores de la NCVS fue con el llamado efecto de **telescopio**: se pregunta a las personas sobre las experiencias que sufrieron como víctimas de un crimen que haya ocurrido en los últimos seis meses, pero algunos incluyen situaciones que sucedieron hace más de seis meses.
- Las personas dan distintas respuestas a diferentes entrevistadores. Schuman y Converse (1971) emplearon encuestadores blancos y negros para entrevistar a los residentes negros en Detroit. A la pregunta "¿cree usted que pueda confiar en la mayoría de las personas blancas, en algunas personas blancas, o en ninguna?", la respuesta de 35% de las personas entrevistadas por un encuestador blanco fue que podían confiar en la mayoría de la gente blanca. El porcentaje fue del 7% para las personas entrevistadas por un encuestador negro.
- Las personas pueden decir lo que piensan que un entrevistador quiere escuchar o lo que piensan que impresionará al entrevistador. En los experimentos realizados con preguntas que inician con "¿está de acuerdo o no con la siguiente afirmación?", se determinó que cierto subconjunto de la población tiende a coincidir con cualquier afirmación sin importar su contenido. Lenski y Leggett (1960) hallaron que cerca de 1/10 de su muestra estaba de acuerdo con las dos afirmaciones siguientes:

  Apenas es justo traer niños al mundo, según parece que será el futuro.

  Los niños recién nacidos tienen un maravilloso futuro por descubrir.

  Algunos comentaristas especulan que el "factor de pena" puede haber jugado cierto papel en las encuestas antes de la elección general en Gran Bretaña de 1992, en la cual el gobierno del Partido Conservador ganó la elección, aunque casi todas las encuestas prede-

cían que el Partido Laborista ganaría. "Las personas pueden *decir* que preferirían mejores servicios públicos, pero al final *votarán* por recortes a los impuestos. Al menos algunos de ellos tuvieron la decencia de sentirse tan apenados como para admitirlo" (Harris 1992).

- Un entrevistador puede afectar la precisión de la respuesta, al leer mal las preguntas, al registrar las respuestas de manera equivocada o al polemizar con el entrevistado. En una encuesta sobre el aborto, un entrevistador con poco entrenamiento y un fuerte sentimiento contra el aborto podría animar al entrevistado a proporcionar una respuesta en vez de otra.
- Ciertas palabras significan cosas distintas para personas diferentes. Una pregunta sencilla como "¿posee usted un auto?" puede ser respondida con un sí o un no, dependiendo de la interpretación del entrevistado a las palabras *posee* (¿cuenta como propiedad el hecho de pagar a una compañía financiera?) o *auto* (¿se permiten camionetas?).
- La formulación y el orden de las preguntas tiene un gran efecto sobre las respuestas obtenidas. A fines de 1993 y a principios de 1994 se realizaron dos encuestas acerca de Elvis Presley. Una encuesta preguntaba: "en los últimos años, ha habido muchos rumores e historias acerca de si Elvis Presley realmente murió. ¿Qué le parece esto? ¿Piensa que existe una posibilidad de que estos rumores sean ciertos y que Elvis siga vivo, o no?" La otra encuesta preguntaba: "un programa de televisión reciente examinó varias teorías sobre la muerte de Elvis Presley. ¿Cree usted que Elvis esté vivo o no?" En la primera encuesta, el 8% de los entrevistados dijo que es posible que Elvis aún esté con vida; en la segunda, el 16% de los entrevistados dijo que es posible que Elvis siga vivo.

Un excelente análisis de estos problemas aparece en Groves (1989) y Asher (1992). En algunos casos, la precisión puede aumentar mediante un diseño cuidadoso de los cuestionarios.

1.5 Diseño de cuestionarios

Esta sección, que es una introducción muy breve al planteamiento y verificación de las preguntas, proporciona ciertas guías y ejemplos generales. Sin embargo, si usted va a escribir un cuestionario, consulte una de las referencias más amplias acerca del diseño de cuestionarios enumeradas en la bibliografía. Actualmente, se realizan muchas investigaciones en el área del uso de resultados de la sicología cognitiva para escribir los cuestionarios; Tanur (1993) y Blair y Presser (1993) son dos referencias útiles sobre este tema.

- *Decida lo que quiere descubrir; éste es el paso más importante para redactar un cuestionario.* Escriba los objetivos de su encuesta. Sea preciso. "Quiero aprender algo acerca de los desamparados" no sirve. En cambio, escriba preguntas específicas como "¿Qué porcentaje de las personas que utilizan los refugios para desamparados en Chicago entre enero y marzo de 1996 tiene menos de 16 años de edad?" Entonces, escriba o elija las preguntas que muestren las respuestas precisas a las cuestiones de la investigación y que motiven a las personas de la muestra a responder las preguntas.
- *Siempre verifique sus preguntas, antes de realizar la encuesta.* Lo ideal es que las preguntas se verifiquen mediante una pequeña muestra de los miembros pertenecientes a la población objetivo. Pruebe con diferentes versiones de las interrogantes y pregunte a los entrevistados en la prueba preliminar la forma en que interpretaron las preguntas.

La NCVS se verificó durante varios años antes de realizarse a escala nacional (Lehnen y Skogan 1981). Las pruebas preliminares sirvieron para decidir el periodo recordado (se decidió preguntar a los entrevistados acerca de las situaciones vividas como víctimas durante los últimos seis meses), para verificar el efecto de los diversos procedimientos y preguntas de la entrevista y para comparar la información de una selección de entrevistas con la información recabada por la policía acerca de las represalias. Como resultado de las pruebas preliminares, algunas preguntas largas y repetitivas se acortaron y se introdujeron formulaciones más específicas.

El cuestionario fue revisado en 1985 y de nuevo en 1991 para emplear investigaciones recientes en sicología cognitiva y para incluir temas no cubiertos en versiones anteriores, como el comportamiento de la víctima y el espectador. Todas las revisiones se probaron ampliamente en el campo antes de ser utilizadas (Taylor 1989). En el pasado, por ejemplo, la NCVS había sido criticada por subinformar del crimen de violación; cuando el cuestionario había sido diseñado a principios de la década de 1970, había la preocupación de que las preguntas directas acerca de la violación serían percibidas como insensibles y embarazosas y provocarían la indignación del Congreso. El cuestionario original de la NCVS planteaba varias preguntas específicas con la intención de hurgar en la memoria de los entrevistados. Incluía preguntas como "¿Alguien tomó algo de usted por la fuerza, como en un asalto o con amenazas?" La última pregunta en la sección de diagnóstico de crímenes violentos del cuestionario era "¿Alguien intentó atacarle de otra forma?" Si la persona entrevistada mencionaba en la respuesta que había sido violada, entonces se informaba de una violación. No debe sorprender que la tasa de víctimas en el caso de una violación registrada en la NCVS de 1990 y de años anteriores era demasiado baja: se informó que cerca de 1 de cada 1000 mujeres mayores de 12 años fueron violadas en 1990. La última versión del cuestionario de la NCVS pregunta directamente acerca de la violación; como resultado, las estimaciones de la cantidad de violaciones se duplicó.

Es probable que las malas interpretaciones de las preguntas no puedan descubrirse al plantear éstas a sus amigos o colegas; éstos podrían tener bases similares a la suya y podrían no entender lo mismo que las personas de la población objetivo. Belson (1981) demuestra que cada una de las 29 preguntas sobre la televisión fueron mal interpretadas por algunos entrevistados. La pregunta "¿cree usted que los programas de noticias de la televisión sean imparciales en política?" fue probada con 56 personas. De éstas, 13 de ellas la interpretaron como se deseaba, 18 redujeron el término *programas de noticias* a "boletines de noticias", 21 lo redujeron a "programas políticos" y 1 lo interpretó como "periódicos". Sólo 25 personas interpretaron *imparciales* como se pretendía; 5 infirieron el sentido opuesto, "parcial"; 11 como "dar demasiada o muy poca atención" y las demás simplemente no estaban familiarizadas con la palabra.

- *Elabore las preguntas de manera sencilla y clara.* Las preguntas que pueden parecerle claras podrían no serlo para alguien que escucha toda la pregunta por teléfono o para una persona con otro idioma materno. Belson (1981, 240) probó la pregunta "¿qué proporción del tiempo que ve la televisión lo dedica a ver programas de noticias?" con 53 personas. Sólo 14 de ellas interpretaron de manera correcta la palabra *proporción* como "porcentaje", "parte" o "fracción". Otras la interpretaron como "cuánto tiempo" o "cuáles programas de noticias observa".

- *Utilice preguntas específicas en vez de preguntas generales, de ser posible.* Strunk y White aconsejan a los redactores "preferir lo específico a lo general, lo definido a lo vago, lo concreto a lo abstracto" (1959, 15). Las buenas preguntas surgen de una buena redacción.

  En vez de preguntar "¿alguien le atacó en los últimos seis meses?", la NCVS plantea una serie de preguntas específicas detallando la forma en que alguien podría ser atacado. La pregunta de la NCVS es "¿alguien le ha atacado o amenazado en alguna de estas formas: (a) Con algún arma, por ejemplo, una pistola o un cuchillo, (b) con algo como un bate de béisbol, una sartén, tijeras o un palo..."

- *Relacione las preguntas que elabore con el concepto de interés.* Esto parece obvio pero se olvida o se ignora en muchas encuestas. En algunas disciplinas, se ha desarrollado y probado un conjunto estándar de preguntas, las cuales son utilizadas posteriormente por otros investigadores. Con frecuencia, el uso de un instrumento común para el estudio permite comparar los resultados de estudios distintos. Sin embargo, en algunos casos, las preguntas estándar no son adecuadas para investigar las hipótesis planteadas.

  Pincus (1993) criticó una de las primeras investigaciones que concluían que es más probable que las personas con artritis tengan problemas sicológicos que las personas sin artritis. En esos estudios, las personas con artritis recibieron el Multiphasic Personality Inventory (en español inventario de personalidad) de Minnesota, una prueba de 566 preguntas cierto/falso utilizada de manera común en la investigación sicológica. Los pacientes con artritis reumatoide tendían a tener una puntuación alta en las escalas de hipocondriasis, depresión e histeria. Parte de la razón de estas altas puntuaciones es clara al analizar las preguntas reales. Una persona con artritis seguramente contesta falso a cuestiones como "actualmente puedo trabajar como siempre lo había hecho", "tengo tan buena salud como la mayoría de mis amigos" y "tengo poco o ningún dolor" sin ser histérica ni hipocondríaca.

- *Decida si debe utilizar preguntas abiertas o cerradas.* Una **pregunta abierta** (el entrevistado no enfrenta categorías en las respuestas) permite a los entrevistados formar sus propias categorías de respuesta; en una **pregunta cerrada** (opción múltiple), el entrevistado elige entre un conjunto de categorías leídas o enumeradas en una tarjeta. Cada tipo tiene sus ventajas. Una pregunta cerrada puede hacer que el entrevistado recuerde respuestas que podría olvidar en caso contrario y va de acuerdo con el principio de que las preguntas específicas son mejores que las generales. Si el tema ha sido ampliamente verificado y se conocen las respuestas de interés, una pregunta cerrada y bien redactada generará por lo general respuestas más precisas, como en la pregunta de la NCVS "¿alguien le ha atacado o amenazado con algo como un bate de béisbol, una sartén, tijeras o un palo?". Por otro lado, si el estudio es exploratorio o las preguntas son sensibles, es mejor utilizar una pregunta abierta. Bradburn y Sudman (1979) observan que los entrevistados informaron de una mayor frecuencia de consumo de bebidas alcohólicas cuando se les planteó una pregunta abierta que una cerrada con categorías desde "nunca" hasta "diariamente".

  La encuesta de Skelly *et al.* (1968) sobre las actitudes de las mujeres con respecto a las telas utilizadas en la ropa dio a aproximadamente la mitad de la muestra una versión abierta del cuestionario y a la otra mitad una versión cerrada, para estudiar las diferencias entre las respuestas. La primera pregunta en el cuestionario abierto era: "¿Qué dificultades y problemas experimenta con más frecuencia al comprar ropa, cualquier tipo de ropa, para usted?"

La pregunta correspondiente de la versión cerrada del cuestionario era: "¿Cuál de estas razones describe mejor las dificultades y problemas que experimenta con más frecuencia al comprar ropa, cualquier tipo de ropa, para usted? ¿Existen otras razones?" La entrevistada debía indicar las afirmaciones de la tarjeta A que se le aplican.

Tarjeta A

1. Tengo una cintura pequeña.
2. Tengo una cintura grande.
3. Necesito algo más corto.
4. Necesito algo más largo.
5. No se me ajusta en los hombros.
6. Tengo caderas muy anchas.
7. Estilos y selecciones limitados.
8. Tengo problemas con los escotes.
9. No puedo encontrar las tallas correctas.
10. Las tallas no están correctas.
11. Mala confección.

De las mujeres que recibieron el cuestionario cerrado, el 25% mencionaron que tenían cintura pequeña, mientras que sólo el 9% de las mujeres que recibieron el cuestionario abierto mencionaron que tenían cintura pequeña. Un mayor porcentaje de mujeres del grupo cerrado mencionaron cada una de las dificultades en la tarjeta. Sin embargo, 10% de las mujeres del grupo abierto mencionaron la dificultad de que el precio era demasiado alto; en el grupo cerrado, sólo el 1% de las entrevistadas mencionaron un alto precio, tal vez porque la tarjeta enfatizaba los problemas de ajuste y se centraba en la figura de la mujer y no en otras dificultades.

Si se utiliza una pregunta cerrada, siempre hay que tener una categoría "otros". En una investigación que estudiaba la actividad sexual entre los adolescentes, se les preguntaba de quién sentían más presión para tener relaciones. Las categorías para la pregunta cerrada eran "amigos del mismo sexo", "novio/novia", "amigos del sexo opuesto", "televisión o radio", "no siento presión" y "otro". La respuesta "padres" o "padre" fue escrita por varios de los adolescentes entrevistados, una respuesta que no fue anticipada por los investigadores.

- *Informe sobre la pregunta que se planteó realmente.* La opinión pública es compleja y usted dejará una impresión distorsionada de ella si comprime los resultados de su cuidadosa investigación en una afirmación sumaria "x% de la población está a favor de la acción afirmativa".

  Los resultados de tres encuestas realizadas en la primavera de 1995, todas con el propósito de estudiar la acción afirmativa, enfatizan la importancia de reportar las preguntas. Una encuesta de *Newsweek* preguntaba lo siguiente: "¿Debería haber una consideración particular para los siguientes grupos con el fin de incrementar su oportunidad para ingresar a la educación superior y obtener trabajos y promociones?" y preguntaba acerca de estos grupos: negros, mujeres, latinos, asiáticos y norteamericanos. La encuesta halló que el 62% de los negros pero sólo el 25% de los blancos contestaron sí a la pregunta relacionada con gente de color. Una encuesta de USA Today, CNN y Gallup preguntaba "¿cuál es su opinión acerca de los programas de acción afirmativa para las mujeres y las minorías: está a favor o en contra?" e informó que el 55% de los entrevistados está a favor de estos programas. Una encuesta de Harris preguntaba "¿está a favor o en contra de una ley que limite los programas de acción afirmativa en su estado?" e informaba que el 51% de los entrevistados estaban a favor de dicha ley. Es claro que estas preguntas se refieren a distintos conceptos, ya que las diferencias entre los porcentajes son demasiado grandes, como para ser adjudicadas a las distintas muestras de personas extraídas por las tres organizaciones. Aun así, los

resultados de las tres encuestas fueron descritas en los periódicos en términos de los porcentajes de las personas que están a favor de la acción afirmativa.

- *Evite preguntas que induzcan o motiven al entrevistado a decir lo que usted quiere escuchar.* Estas preguntas se llaman con frecuencia **preguntas intencionadas**. En la edición del 17 de mayo de 1994, *The Wall Street Journal* informó de la siguiente pregunta planteada por Gallup Organization en una encuesta encomendada por el American Paper Institute: "se estima que los pañales desechables representan menos del 2% de los desechos existentes en los basureros. Por el contrario, las latas de cerveza, el correo de tercera clase y los desperdicios de jardín representan casi el 21% de desechos que se encuentran en los basureros. Dado esto, en su opinión, ¿sería justo establecer un impuesto o prohibir los pañales desechables?"

- *Utilice preguntas de opción forzosa, en vez de a favor/en contra.* Como ya observamos, algunas personas estarán de acuerdo con casi cualquier afirmación. Schuman y Presser (1981, 223) informan de las siguientes diferencias en un experimento que compara las versiones a favor/en contra con las versiones de opción forzosa:

  Pregunta 1: ¿Está a favor o en contra de la siguiente afirmación? La mayoría de los hombres están más capacitados emocionalmente para la política que la mayoría de las mujeres.

  Pregunta 2: ¿Diría usted que la mayoría de los hombres están más capacitados emocionalmente para la política que la mayoría de las mujeres, que los hombres y las mujeres están igual de capacitados o que las mujeres están más capacitadas que los hombres en esta área?

	Años de escolaridad		
	0-11	12	13+
Pregunta 1: porcentaje que está "de acuerdo"	57	44	39
Pregunta 2: porcentaje de "los hombres que están mejor capacitados"	33	38	28

- *Plantee sólo un concepto en cada pregunta.* En particular, evite lo que a veces se llama **preguntas con doble cañón**, llamadas así porque si un cañón de la escopeta no lo alcanza, el otro sí.

  La pregunta "¿está de acuerdo con el préstamo de $50 mil millones otorgado a México por Bill Clinton?" apareció en una encuesta distribuida por un miembro de la Cámara de los Representantes de Estados Unidos a los miembros de ésta. La pregunta confunde dos opiniones de los entrevistados: la opinión de Bill Clinton y la opinión acerca de la política estadounidense hacia México. La desaprobación de uno de estos puntos llevará a una respuesta "en contra" a la pregunta. Observe también que el contenido cargado de la palabra *préstamo*, con toda seguridad producirá más respuestas negativas que las que generaría el término *paquete de ayuda*.

- *Preste atención al efecto del orden de las preguntas.* Si plantea más de una pregunta sobre un tema, por lo general, es mejor (aunque no siempre) elaborar primero la pregunta más general y después las preguntas específicas. McFarland (1981) realizó un experimento donde la mitad de los entrevistados recibieron primero preguntas generales (por ejemplo, "¿qué tan interesado diría que está en la religión: muy interesado, algo interesado o no muy interesado?"), seguidas de preguntas específicas acerca del tema ("¿Ha asistido por propia convicción a la iglesia durante los últimos siete días?"); la otra mitad recibió primero las preguntas específicas y luego las preguntas generales. Cuando la pregunta general se planteó primero, el 56% informó que estaban "muy interesados en la religión"; el porcentaje subió a 64% cuando se planteó primero la pregunta específica.

Serdula *et al.* (1995) encontraron que cuando a una persona entrevistada en una encuesta de salud se le pedía su peso y luego se le preguntaba "¿está tratando de perder peso?", el 28.8% de los hombres y el 48.0% de las mujeres informaron que estaban tratando de perder peso. Cuando a la mitad de la encuesta se les preguntaba "¿está tratando de perder peso?" y el informe del propio peso se preguntaba al final de la encuesta, 26.5% de los hombres y el 40.9% de las mujeres informaron que estaban tratando de perder peso. Los autores especulan que los entrevistados a los que se recuerda su peso podrían informar que tratan de perder peso.

1.6 Errores de muestreo y que no son de muestreo

La mayor parte de las encuestas de opinión informan de un *margen de error*. Muchas simplemente dicen que el margen de error es de 3 puntos porcentuales. Otras dan más detalles, como en una encuesta del *New York Times*: "en teoría, 19 de 20 resultados basados en tales muestras diferirán en no más de 3 puntos porcentuales en cualquier dirección de lo que se hubiera obtenido al entrevistar a todos los norteamericanos". El margen de error dado en las encuestas es una expresión del **error de muestreo**, el cual resulta al considerar una muestra y no al examinar a toda la población. Si consideramos una muestra distinta, es muy probable que obtengamos un porcentaje muestral distinto de las personas que asistieron a una biblioteca pública la semana pasada. Los errores de muestreo se reportan, por lo general, en términos probabilísticos, como en el ejemplo del *New York Times* (analizaremos el cálculo de los errores de muestreo para los distintos diseños de encuesta en los capítulos 2 a 7).

El sesgo de selección y la imprecisión de las respuestas son ejemplos de los **errores que no son de muestreo**, los cuales no se pueden atribuir a la variabilidad entre las muestras. En muchas encuestas, el error de muestreo reportado para esa encuesta puede ser despreciable en comparación con los errores que no son de muestreo; con frecuencia, usted verá encuestas con una tasa de respuesta del 30% que proclaman con orgullo su margen de error del 3%, con esto se ignora el tremendo sesgo de selección en sus resultados.

El objetivo de este capítulo es sensibilizarle sobre las distintas formas de sesgo de selección y las respuestas imprecisas. Podemos reducir algunas formas de sesgo de selección al emplear los métodos de muestreo probabilístico, como se describe en el siguiente capítulo. Con frecuencia, las respuestas precisas se pueden lograr mediante un diseño y una prueba cuidadosos del instrumento de la encuesta, entrenamiento de los entrevistadores y una verificación preliminar de la encuesta. Regresaremos a los errores que no son de muestreo en el capítulo 8.

¿Por qué, en fin, una muestra? Con la abundancia de las encuestas mal realizadas, no debe sorprendernos que algunas personas sean escépticas respecto a las encuestas. "Después de todo", dicen, "nunca me han pedido mi opinión, así que ¿cómo podrían afirmar que los resultados de la encuesta me representan?" El cuestionamiento público de la validez de las encuestas se intensifica después de que una encuesta comete un enorme error al predecir los resultados de una elección, como en la encuesta del *Literary Digest* de 1936 o en la elección presidencial de Estados Unidos en 1948, donde la mayor parte de las encuestas predecían que Dewey derrotaría a Truman. Otro reproche público contra la investigación por medio de las encuestas ocurrió después de la elección general en Gran Bretaña en 1992, cuando el gobierno conservador ganó la reelección a pesar de las predicciones de todas, excepto una, las principales organizaciones de encuestas, en el sentido de que habría un empate o ganaría el Partido Laborista. Un miembro del Parlamento expresó su opinión diciendo que "extrapolar lo que

decenas de millones están pensando a partir de una pequeña muestra de opiniones es una afrenta a la inteligencia humana y niega la verdadera libertad de pensamiento".

Algunas personas insisten en que sólo un censo completo, en donde se mida a cada elemento de la población, será satisfactorio; esta objeción al muestreo tiene una larga historia. Cuando Anders Kiaer (1897), director de estadística en Noruega, propuso el uso del muestreo para reunir las estadísticas oficiales, su propuesta estuvo lejos de ser bien recibida por todos. Los oponentes al muestreo argumentaron que era peligroso y que las muestras no podrían reemplazar a un censo. Sin embargo, después de algunos años, la comunidad estadística internacional fue persuadida de que las muestras representativas son buenas, aunque las muestras probabilísticas sólo se utilizaron con amplitud hasta las décadas de 1930 y 1940.

Para las poblaciones pequeñas, es claro que sería práctico un censo. Por ejemplo, si usted quiere conocer el historial de empleo de los graduados en matemáticas de la Universidad Estatal de Arizona en 1990, podría establecer contacto con ellos. Si todos los graduados responden, entonces las estimaciones de la encuesta no tendrán un error de muestreo. Sin embargo, las estimaciones podrían tener errores que no son de muestreo, si las preguntas están mal redactadas o si los entrevistados tienen información imprecisa. Si alguno de los graduados no regresa el cuestionario, entonces las estimaciones podrían estar sesgadas debido a la ausencia de respuesta.

En general, la elaboración de un censo completo de una población requiere mucho tiempo y dinero y no elimina el error. Con frecuencia, las principales causas de error en una encuesta son la subcobertura, la carencia de respuesta y los descuidos en la recolección de datos. La mayoría de nosotros ha llevado el registro de una chequera en algún momento, que esencialmente es un censo de todos los cheques y depósitos en la cuenta. ¿Quiénes podrían decir que nunca han cometido algún error en la chequera? Por lo general es mucho mejor extraer una muestra de buena calidad y asignar mejor los recursos, por ejemplo, teniendo más cuidado al reunir o registrar los datos, realizar estudios de seguimiento o medir más variables.

Después de todo, la encuesta del *Literary Digest* (véase el ejemplo 1.1) predijo el voto incorrecto aun en condados donde intentó realizar un censo. El censo decenal, que intenta enumerar a cada residente en Estados Unidos, omite algunos segmentos de la población. Para el censo del año 2000, un comité de la National Academy of Sciences (Academia Nacional de Ciencias) recomendó combinar la enumeración con el muestreo para mejorar la precisión del censo. El Congreso discute actualmente esta propuesta.

Existen tres justificaciones principales para el uso del muestreo:

- El muestreo puede proporcionar información confiable con costos mucho menores que los de un censo. Con las muestras probabilísticas (descritas en el siguiente capítulo), usted puede cuantificar el error de muestreo a partir de una encuesta. En algunos casos, una unidad de observación debe ser destruida para ser observada, como cuando una galleta debe pulverizarse para determinar el contenido de grasa. En ese caso, una muestra proporciona información confiable acerca de la población; un censo destruiría a toda la población, y con ello, la necesidad de información relativa a ella.
- Los datos se pueden reunir más rápido, de modo que las estimaciones se pueden publicar de una manera programada. Una estimación de la tasa de desempleo de 1994 no es muy útil si para entrevistar a cada familia se tarda hasta el año 2004.
- Por último, y esta razón no es tan conocida, las estimaciones basadas en las encuestas y sus respectivas muestras son, con frecuencia, más precisas que las basadas en un censo, pues los investigadores pueden tener más cuidado al reunir los datos. Un censo completo necesita, por lo regular, de una gran organización administrativa e implica a muchas personas en la recolección de datos. Con tal complejidad administrativa y la presión por producir las estimaciones a tiempo, se pueden cometer muchos errores en la elaboración del censo. En una muestra, se puede dedicar más atención a la calidad de los datos, al entrenar al personal y realizar un seguimiento de quienes no contestan la encuesta. Es mucho mejor tener buenas mediciones en una muestra representativa que mediciones poco confiables o sesgadas sobre toda la población.

Deming dice: "el muestreo no es una simple sustitución de una cobertura total por una parcial. El muestreo es la ciencia y arte de controlar y medir la confiabilidad de la información estadística útil a través de la teoría de la probabilidad" (1950, 2). En los demás capítulos de este libro exploraremos esta ciencia y arte con detalle.

1.7 Ejercicios

Para cada una de las siguientes encuestas, describa la población objetivo, el marco de muestreo, la unidad de muestreo y la unidad de observación. Analice todas las posibles fuentes de sesgo de selección o imprecisión de las respuestas.

1 El artículo "Lo que los lectores dicen de la mariguana" informa que "más de 75% de los lectores que participaron en una encuesta telefónica informal de PARADE dicen que la mariguana debería ser tan legal como las bebidas alcohólicas" (*Parade*, 31 de julio de 1994, 16). La encuesta telefónica fue anunciada en la página 5 del ejemplar del 12 de junio; los lectores eran instruidos así: "llame al 1-900-773-1200, a 75 centavos la llamada, si desea responder las siguientes preguntas. Utilice solamente un teléfono de tonos. Para participar, llame entre las 8 a.m. EDT (tiempo del este) del sábado 11 de junio hasta la medianoche EDT del miércoles 15 de junio".

2 Una estudiante desea estimar el porcentaje de los fondos mutuos cuyas acciones aumentaron de precio la semana pasada. En una lista de fondos mutuos del periódico, esta alumna elige uno de cada 10 fondos y calcula el porcentaje de aquellos donde el precio de la acción ha aumentado.

3 Los jurados potenciales en ciertas jurisdicciones se eligen de una lista de residentes del país, que sean votantes registrados o conductores con licencia, mayores de 18 años. En el cuarto trimestre de 1994, se enviaron 100 300 citatorios a los residentes del condado de Maricopa, Arizona. Aproximadamente 23 000 de éstos fueron regresados a la oficina de correos como imposibles de entregar. Aproximadamente 7000 personas no quedaron calificadas para el servicio debido a que no eran ciudadanos estadounidenses, tenían menos de 18 años, eran criminales o tenían alguna otra razón que los descalificaba para servir como jurados. Otros 22,000 fueron excusados debido a una enfermedad, problemas financieros, servicio militar o alguna otra razón aceptable. La muestra final consistía de personas que aparecieron para servir como jurados; algunos jurados no excusados no aparecieron.

4 Se extrae una muestra de ocho arquitectos de una ciudad con 14 arquitectos y empresas de arquitectura. Para formar la muestra para una encuesta, cada arquitecto fue contactado por teléfono por orden de aparición en el directorio telefónico. Los primeros ocho que acordaron ser entrevistados conformaron la muestra.

5 Para estimar cuántos libros de la biblioteca deben ser encuadernados de nuevo, un bibliotecario utiliza una tabla de números aleatorios para elegir al azar 100 posiciones en los estantes de la biblioteca. Luego, camina hasta cada posición, busca el libro que se encuentra en ese punto y registra si el libro debe encuadernarse de nuevo o no.

6 Muchos investigadores y creadores de políticas están interesados en la proporción de personas sin hogar que tienen enfermedades mentales. Wright (1988) estima que el 33 por ciento de las personas desamparadas tienen enfermedades mentales, al realizar un muestreo de las personas sin hogar que recibieron atención médica de una de las clínicas del proyecto Health Care for the Homeless (en español Salud para los Desamparados, HCH). Wright argumenta que el sesgo de la selección no es un problema serio, pues las clínicas son bastante accesibles para los desamparados y los perfiles demográficos de los clientes de HCH eran cercanos a los de la población general sin hogar en cada ciudad de la muestra. ¿Está usted de acuerdo?

7 Aproximadamente 16,500 mujeres regresaron la Healthy Women Survey (la Encuesta de Salud en las Mujeres) que apareció en el ejemplar de septiembre de 1992 de *Prevention*. El ejemplar de mayo de 1993, donde se informó de la encuesta, estableció que el "92% de nuestras lectoras calificaron su salud como excelente, muy buena o buena".

8 Se realiza un estudio para determinar el peso promedio de las vacas en una región. De una lista de granjas disponibles en esa región, se eligen al azar 50 de ellas. Luego se registra el peso de cada vaca de las 50 granjas elegidas.

9 El Intrastate Travel Committee (en español, el Comité de Viajes Interestatales) de Arizona realizó un estudio para identificar los patrones de viaje dentro del estado de los residentes en Phoenix y Tucson, para evaluar distintas fuentes de información para planear las vacaciones. Se realizaron 400 entrevistas con residentes de Phoenix y el mismo número con residentes de Tucson. Los números telefónicos de ambas ciudades se generaron al azar, de modo que los números telefónicos enumerados y no enumerados pudieran ser localizados. "Los entrevistados se limitaron a las cabezas de familia y se establecieron cuotas para tener una misma representación de hombres y mujeres. Además, se revisaron los rangos de ingreso y edad para mantener la misma proporción que las bases generales de población de las zonas metropolitanas de Phoenix y Tucson" (oficina de Turismo de Arizona 1991).

10 La siguiente carta al editor apareció el 10 de diciembre de 1995 en *Post-Crescent* de Appleton: "Paul Harvey, Dios lo bendiga, ha iniciado una encuesta a nivel nacional patrocinada por estaciones de radio independientes a través de los programas con intervención del público, para determinar los sentimientos reales de los estadounidenses con respecto al envío de tropas a Bosnia. Hasta ahora, los resultados a lo largo y ancho de la nación promedian un 90% contra esa decisión".

11 Para estudiar el contenido nutricional de los menús en pensiones para ancianos, en el estado de Washington, Goren *et al.* (1993) enviaron encuestas a las 184 pensiones autorizadas en el estado, dirigidas al administrador y al gerente de servicios alimenticios. Los 43 cuestionarios fueron regresados antes de la fecha límite, incluyendo los menús.

12 La edición de junio de 1994 de *PC World* (a la venta en mayo de 1994) incluía un informe sobre la confiabilidad y el soporte técnico para las computadoras personales (PC). Una de las conclusiones, "25% de las PC nuevas tienen problemas", fue el encabezado del 23 de mayo de 1994 en *USA Today*. Cada número de *PC World*, desde octubre de 1993, incluía una forma de la encuesta, la cual planteaba preguntas referentes a los problemas de los usuarios con el hardware. Las personas participantes en la encuesta de cada mes entraban a un sorteo para ganar una nueva PC y se recibieron más de 45 000 respuestas.

13 En un juicio sobre marcas registradas, un demandante que afirme que otra compañía infringe sus marcas registradas debe mostrar con frecuencia que las marcas tienen un *significado secundario* en el mercado; es decir, los usuarios potenciales del producto asocian las marcas registradas con el demandante aun cuando no esté presente el nombre de la compañía. En el caso judicial *Harlequin Enterprises Ltd vs. Gulf & Western Corporation* (503 F. Supp. 647, 1980), el editor de las novelas Harlequin convenció a la corte de que el diseño de la portada de la serie de novelas "Harlequin Presents" había adquirido un significado secundario. Parte de la evidencia presentada era una encuesta de 500 mujeres de tres ciudades, quienes se identificaban como lectoras de novelas. Se les mostraron copias de las novelas de la serie ocultándoles el nombre Harlequin; más del 50% identificaron la novela como un producto Harlequin.

14 En 1976, Ann Landers pidió a los lectores de su columna que respondieran a la siguiente pregunta: "¿si pudiera repetir todo, tendría hijos?" Cerca del 70% de los lectores que respondieron, dijeron no. Ella recibió más de 10,000 respuestas, 80% de las cuales fueron de mujeres.

15 La edición de agosto de 1996 de *Consumer Reports* contenía los ratings de satisfacción para diversas organizaciones de mantenimiento de la salud (HMO) utilizadas por los lectores de la revista. Al describir la encuesta, los editores decían que "los ratings estaban basados en más de 20,000 respuestas a nuestro cuestionario anual de 1995 sobre la experiencia con HMO entre mayo de 1994 y abril de 1995. Estos resultados reflejan las experiencias de los suscriptores de *Consumer Reports*, que son una parte rica y educada de la población de Estados Unidos" (página 40). Responda a las preguntas generales acerca de la población objetivo, marco de muestreo y unidades para esta encuesta. Además, ¿piensa que esta encuesta proporciona información valiosa para comparar los planes de salud? Si fuera a elegir una HMO para usted, ¿qué información preferiría, resultados de esta encuesta o resultados de encuestas de satisfacción de los clientes realizadas por cada HMO?

16 Se ha mostrado que las mutaciones del gen BRCA1 en el cromosoma 17 se asocian con un mayor riesgo de cáncer de pecho y ovarios. Ford *et al.* (1994) estudiaron el riesgo de cáncer en las portadoras de la mutación del BRCA1, al utilizar una muestra de tres familias en América del Norte y Europa Occidental. Las familias fueron seleccionadas por investigadores del cáncer de pecho. Cada familia de la muestra tenía al menos cuatro mujeres, quienes habían recibido un diagnóstico de cáncer de pecho o de ovarios antes de los 60 años de edad. Los investigadores estimaron que el riesgo de cáncer de pecho o de ovarios a partir de la ocurrencia de un segundo cáncer en las mujeres con cáncer de pecho y estimaron "un riesgo acumulado de cáncer de pecho en las portadoras del gen *BRCA1* de 87% a la edad de 70 años". Los científicos concluyeron: "este estudio confirma que las portadoras del gen *BRCA1* tienen un riesgo de por vida de tener cáncer de pecho o de ovarios cercano al 100%, y que las portadoras con un cáncer anterior tienen un alto riesgo de desarrollar un segundo cáncer de pecho o de ovarios y deben ser controladas de acuerdo con esto". (*página 694*). Con base en los altos riesgos calculados a partir de este análisis y de muestras con diseños similares, muchos médicos han recomendado que las mujeres con una historia familiar de cáncer de pecho se realicen pruebas genéticas; algunas féminas se han realizado mastectomías profilácticas después de descubrir que probablemente tuvieran ese gen.

a Responda las preguntas generales acerca de la población objetivo, marco de muestreo y unidades para esta encuesta.

b ¿Proporciona este estudio una estimación de la probabilidad de que una mujer portadora del gen desarrolle un cáncer de pecho o de ovarios? Explique.

17 Las siguientes preguntas, citadas en Kinsley (1981), son parte de una encuesta realizada por Cambridge Reports, financiada por Union Carbide. Critique estas preguntas.

Algunas personas dicen que otorgar créditos fiscales a las compañías por los impuestos que pagan a otras naciones aumentaría la competitividad internacional de estas compañías. Si usted aceptara como hecho que los créditos fiscales para impuestos pagados a otras naciones

aumentarían el dinero disponible para que las empresas estadounidenses ampliaran y modernizaran sus plantas y creasen más empleos, ¿estaría a favor o en contra de tal política fiscal?

¿Está a favor o en contra de modificar los reglamentos ambientales para que sigan protegiendo al público y además cuesten menos a las empresas estadounidenses y reduzcan los costos de producción?

18 El siguiente artículo, "Grupos de derecho al aborto investigan el punto de vista de los votantes", de Jack Coffman, apareció el 26 de diciembre de 1989 en el *St. Paul Pioneer Press Dispatch*[1]. Critique la encuesta descrita en este artículo.

Lo que se ha dado en llamar la mayor encuesta del sentimiento nacional hacia el derecho al aborto se ha convertido en algo más grande de lo esperado por sus organizadores, dicen los líderes del esfuerzo de consulta.

A partir del 20 de noviembre, más de 7000 voluntarios han operado seis centros telefónicos en el área metropolitana de las Ciudades Gemelas y Duluth, con otros 1,000 voluntarios esperando su turno para trabajar en enero, después de dos semanas de vacaciones que iniciaron el 15 de diciembre. Otros 2000 voluntarios comenzarán el próximo mes operaciones telefónicas en sus propias casas en la Minnesota rural.

Desde su inicio, el esfuerzo ha establecido contacto con casi 160 000 familias de Minnesota a quienes se les pregunta su punto de vista sobre el aborto y con los presidentes de 74 de los 87 condados del estado. El objetivo anunciado es establecer contacto con las familias de todos los votantes registrados y con los grupos clave relacionados con la encuesta.

Se espera que la encuesta juegue un papel importante en la sesión legislativa de 1990, cuando los legisladores, quienes tradicionalmente han tenido inclinaciones antiaborto, tendrán que lidiar con este tema tan escabroso. El tema se ha vuelto aún más sensible debido a la decisión, en el verano pasado, de la Corte Suprema de Estados Unidos de mantener una ley en Missouri que incrementa las restricciones al aborto y parece abrir el camino para la acción por parte de las legislaturas locales.

Las fuerzas antiaborto se están moviendo para lograr nuevas restricciones al aborto. Los que respaldan la encuesta sobre el derecho al aborto planean utilizar los resultados, en parte, para detener otras leyes antiaborto.

Hasta ahora, los resultados de la encuesta son "abrumadoramente a favor de la elección", dijo Kayser.

Los resultados de las llamadas realizadas desde noviembre se están organizando en tablas y estarán disponibles durante la próxima sesión legislativa, que comienza el 12 de febrero. Se espera terminar la encuesta de un millón de habitantes de Minnesota el 10 de marzo.

La encuesta, patrocinada por varios grupos de derecho al aborto, es realizada según un contrato con Nancy Brataas Associates Inc., una empresa consultora propiedad de la senadora Nancy Brataas, IR-Rochester. Su costo estimado es de $250 000.

"Es el flujo más maravilloso de voluntarios que haya visto y esto incluye las campañas para presidentes y gobernadores", dijo Brataas (la campaña presidencial en Minnesota del gobernador de Massachusetts Michael Dukakis involucró a 8 000 voluntarios, según un funcionario de la campaña).

Brataas dijo creer que la fuerte respuesta de los voluntarios surge de las "personas que están a favor de una elección y que han dependido de la Corte Suprema, que súbitamente están muy preocupados por el derecho de las mujeres a elegir".

El reclutamiento de los voluntarios "no fue difícil", dijo Mary Stringer, copresidente del centro de la encuesta en St. Paul, en el edificio Griggs-Midway, donde 865 voluntarios han operado 15 teléfonos, con lo que tratan de lograr el objetivo de 1 625 por día.

[1] Reproducido con autorización de *St. Paul Pioneer Press*.

Stringer, quien describió el flujo de los voluntarios como "increíble", señaló dos cajas de solicitudes de los voluntarios que no han sido llamados aún. También se realizan llamadas desde bancos telefónicos en Bloomington, St. Louis Park, White Bear Lake, Minneapolis y Duluth.

Jackie Schwietz, codirectora de Minnesota Citizens Concerned for Life (ciudadanos preocupados por la vida, MMCL), dijo que la encuesta era "sesgada" y "deshonesta", pues las preguntas no mencionan al aborto.

Schwietz dijo que el MMCL tiene un "plan preciso" para impulsar una legislación del aborto "más restrictiva" en el periodo legislativo de 1990. Sin embargo, declinó describir las intenciones del grupo, que mencionó serían el tema de una presentación pública antes del inicio de la sesión.

Cuando los voluntarios llaman a las votantes registradas, preguntan lo siguiente: "¿está de acuerdo o en contra de la siguiente afirmación: la decisión de terminar un embarazo es un tema privado entre una mujer, su familia y su doctor... y no una decisión a ser tomada por el gobierno y los políticos?"

Si la persona entrevistada contesta sí, se le pregunta a continuación: "¿en vista de las amenazas actuales del gobierno al aborto seguro y legal... influirá este tema en su opinión acerca de los políticos en el futuro?"

Si la respuesta a la pregunta original es no, se inquiere: "¿se opone al aborto en los casos de violación... incesto... deformación fetal seria... o para salvar la vida de una mujer?"

19 El 21 de marzo de 1993, NBC transmitió "el primer referéndum nacional: Reforma gubernamental, que fue presentado por Ross Perot". Durante el programa, el candidato presidencial de 1992, Ross Perot, pidió a los espectadores que expresaran sus opiniones por correo al Referéndum Nacional sobre la Reforma Gubernamental, impreso en el ejemplar del 20 de marzo de *TV Guide*. Algunas de las preguntas de la encuesta fueron las siguientes:

¿Cree que por cada dólar de incremento a los impuestos debería haber $2.00 en recorte de gastos, al asignar el ahorro a la reducción del déficit y de la deuda?

¿Debería el presidente presentar un plan general que incluya el recorte de gastos, aumento de gastos y aumento de impuestos y presentar el resultado neto del plano general, de modo que las personas conozcan el resultado neto antes de pagar más impuestos?

¿Debería ser reemplazado el Colegio Electoral por el voto popular para la elección presidencial?

¿Valió la pena este foro televisivo? ¿Quiere seguir participando como miembro con voto de United We Stand America?

2

Muestras de probabilidad simples

[Kennedy] leía una de cada 50 cartas de las treinta mil que llegaban semanalmente a la Casa Blanca, al igual que un resumen estadístico de todo el lote, aunque él sabía que con frecuencia era tan organizado y no representativo como las estacas de Pennsylvania Avenue.

—Theodore Sorensen, *Kennedy*

Los ejemplos de encuestas mal realizadas del capítulo 1 (por ejemplo, la encuesta del *Literary Digest*) tenían fallas tan grandes que producían muestras no representativas. En este capítulo analizaremos la forma de utilizar el **muestreo de probabilidad** para llevar a cabo encuestas. En una muestra de probabilidad, cada unidad de la población tiene una probabilidad de selección conocida; se emplea un método aleatorio (como el uso de una tabla con números aleatorios) para elegir las unidades específicas que se incluirán en la muestra. Si un muestreo de probabilidad se realiza de manera adecuada, un investigador puede utilizar una muestra relativamente pequeña para llevar a cabo inferencias de una población arbitrariamente grande.

En los capítulos del 2 al 6 exploraremos el diseño de las encuestas y las propiedades de las estimaciones de los tres principales componentes del diseño utilizado en una muestra de probabilidad: el muestreo aleatorio simple, el muestreo estratificado y el muestreo por conglomerados. Integraremos todas estas ideas en el capítulo 7 y veremos cómo combinarlas en encuestas complejas, como la NCVS de Estados Unidos. Para facilitar la presentación de los conceptos, supondremos por el momento que la población muestreada es la población objetivo, que el marco de muestreo es completo, que no hay ausencia de respuestas o datos faltantes y que todas las mediciones son exactas. Regresaremos a los errores que no son de muestreo en el capítulo 8.

Por supuesto, usted debe poseer los conocimientos necesarios para entender el muestreo de probabilidad. Tal vez desee revisar el material de las secciones B.1 y B.2 del apéndice B mientras lee este capítulo.

2.1 Tipos de muestras de probabilidad

Los términos *muestra aleatoria simple*, *muestra estratificada* y *muestra por conglomerados* son básicos en cualquier análisis de las encuestas con muestras, de modo que los definiremos enseguida.

- Una **muestra aleatoria simple** es la forma más sencilla de realizar un muestreo probabilístico. Se obtiene una muestra aleatoria simple de tamaño n cuando cualquier subconjunto posible de n unidades en la población tiene la misma probabilidad de ser seleccionada para componer la muestra. Estas muestras son el centro de este capítulo y la base para otros diseños de muestreo más complejos. Al extraer una muestra aleatoria, el investigador mezcla de hecho la población antes de sacar n unidades. Un investigador no necesita examinar a todos los miembros de una población, por la misma razón que un encargado de análisis médicos no tiene que obtener toda la sangre para medir la cantidad de glóbulos rojos: la sangre está bastante bien mezclada, de modo que cualquier muestra sería representativa. Estas muestras se analizan en la sección 2.3, después de presentar el marco de referencia básico para las muestras de probabilidad, en la sección 2.2.
- En una **muestra aleatoria estratificada**, la población se divide en subgrupos llamados *estratos*. Al llevar a cabo esta división, se extrae una muestra aleatoria simple de cada estrato la cual se elige de manera independiente. Los estratos son, con frecuencia, subgrupos de interés para el investigador; por ejemplo, los estratos podrían ser grupos étnicos o de edad en una encuesta que tratara sobre personas; diferentes tipos de terreno en un estudio ecológico o tamaños de empresas en un estudio comercial. Los elementos del mismo estrato tienden, por lo regular, a ser más similares que los elementos elegidos al azar de la población entera, de modo que, a menudo, la estratificación aumenta la precisión, como veremos en el capítulo 4.
- En una **muestra por conglomerados**, las unidades de observación que componen una población se reúnen en unidades de muestreo de mayor tamaño, llamadas *conglomerados*. Suponga que debe realizar una encuesta de los miembros que forman la Iglesia luterana en Minneapolis, pero no cuenta con una lista completa de todos ellos; de modo que no podrá extraer una muestra aleatoria simple de los miembros que componen dicha Iglesia. Sin embargo, posee una lista de todas las Iglesias luteranas. Entonces, extrae una muestra aleatoria simple de las iglesias y, luego, realiza una nueva muestra entre todos o algunos de los miembros de las iglesias elegidas. En este caso, las iglesias forman los conglomerados y los miembros de cada iglesia son las unidades de observación. Es más conveniente realizar un muestreo al nivel de las iglesias; sin embargo, los miembros de la misma iglesia podrían tener más analogías que los luteranos elegidos al azar en Minneapolis, de modo que una muestra de conglomerados de 500 luteranos podría no proporcionar tanta información como una muestra aleatoria simple de 500 luteranos. Revisaremos esta situación con más detalle en el capítulo 5.

Suponga que quiere estimar la cantidad de tiempo promedio que los profesores de la universidad donde estudia ocupan calificando las tareas, en cierta semana. Para extraer una muestra aleatoria simple, construya una lista de todos los profesores y elija al azar n de ellos para formar su muestra. Ahora, pregunte a cada uno de los profesores, que componen la muestra que acaba de formar, la cantidad de tiempo que ocupan al calificar la tarea en esa semana; por supuesto, tendrá que definir, con cuidado, las palabras *tarea* y *calificación* en el cuestionario. En una muestra estratificada, podría clasificar a los profesores por especialidad: ingeniería, ciencias, humanidades, enfermería y bellas artes. Entonces, extraería una muestra aleatoria simple de profesores de ingeniería, una muestra aleatoria simple de profesores de la facultad de ciencias, etcétera. Para una muestra por conglomerados, usted elegiría al azar 10 de los 60 departamentos académicos de la universidad y preguntaría a cada profesor, de esos departamentos, el tiempo que ocupan en calificar la tarea.

Los tres métodos (simple, estratificado y por conglomerados) implican la selección aleatoria de las unidades que formarán parte de la muestra. En una muestra aleatoria simple, las propias unidades de observación se eligen al azar de los elementos que componen la población; en una muestra estratificada, se escogen al azar las unidades de observación dentro de cada estrato; en una muestra por conglomerados, los conglomerados se eligen al azar de entre toda la población. Cada método es una forma de muestreo de probabilidad, que analizaremos en la siguiente sección.

2.2 Marco de referencia para el muestreo de probabilidad

Para mostrar cómo funciona el muestreo de probabilidad, necesitamos enumerar las N unidades que componen la población finita. Denotamos la **población** finita de N unidades, o **universo**, mediante el conjunto de índices

$$\mathcal{U} = \{1,2,\ldots,N\}. \tag{2.1}$$

De esta población podemos extraer varias muestras, que son subconjuntos de $\mathcal{U}$. La muestra particular elegida se denota como $\mathcal{S}$, un subconjunto de n unidades de $\mathcal{U}$.

Suponga que la población tiene cuatro unidades: $\mathcal{U} = \{1,2,3,4\}$. Podemos elegir seis muestras distintas de tamaño 2 de esta población:

$$\mathcal{S}_1 = \{1, 2\} \qquad \mathcal{S}_4 = \{2, 3\}$$
$$\mathcal{S}_2 = \{1, 3\} \qquad \mathcal{S}_5 = \{2, 4\}$$
$$\mathcal{S}_3 = \{1, 4\} \qquad \mathcal{S}_6 = \{3, 4\}$$

En el muestreo de probabilidad, cada muestra posible $\mathcal{S}$ de la población tiene una probabilidad dada $P(\mathcal{S})$ de ser elegida y la suma de las probabilidades de las posibles muestras es igual a 1. Un posible diseño de muestra para obtener una muestra de probabilidad de tamaño 2 sería $P(\mathcal{S}_1) = 1/3$, $P(\mathcal{S}_2) = 1/6$, $P(\mathcal{S}_6) = 1/2$ y $P(\mathcal{S}_3) = P(\mathcal{S}_4) = P(\mathcal{S}_5) = 0$. Las probabilidades $P(\mathcal{S}_1)$, $P(\mathcal{S}_2)$ y $P(\mathcal{S}_6)$ de las muestras posibles se conocen antes de extraer la muestra. Una forma de elegir la muestra consiste en colocar seis bolas etiquetadas dentro de una caja; dos de ellas tienen la etiqueta 1, una la etiqueta 2 y tres con la etiqueta 6. Ahora, elegimos una al azar; si se elige una bola con la etiqueta 6, entonces $\mathcal{S}_6$ es la muestra.

En una muestra de probabilidad, como cada muestra posible tiene una probabilidad dada de ser elegida, conocemos la probabilidad de cada unidad de la población para aparecer en nuestra muestra seleccionada. Calculamos:

$$P(\text{unidad } i \text{ en la muestra}) = \pi_i$$

al sumar, sobre todas las muestras posibles, la probabilidad de que cada muestra contenga a la unidad i. En el muestreo de probabilidad, se conocen los π_i antes de iniciar la encuesta y suponemos que $\pi_i > 0$ para cada unidad de la población. Para el diseño de muestra descrito con anterioridad, $\pi_1 = P(\mathcal{S}_1) + P(\mathcal{S}_2) + P(\mathcal{S}_3) = 1/2$, $\pi_2 = P(\mathcal{S}_1) + P(\mathcal{S}_4) + P(\mathcal{S}_5) = 1/3$, $\pi_3 = P(\mathcal{S}_2) + P(\mathcal{S}_4) + P(\mathcal{S}_6) = 2/3$ y $\pi_4 = P(\mathcal{S}_3) + P(\mathcal{S}_5) + P(\mathcal{S}_6) = 1/2$.

Por supuesto, nunca enumeramos todas las muestras posibles ni calculamos la probabilidad con la que podemos elegir cada muestra posible; esto llevaría mucho tiempo. Pero tal enumeración es subyacente a todo el muestreo de probabilidad. Los investigadores que utilizan una muestra de probabilidad tienen menos discreción acerca de las unidades incluidas en la muestra, de modo que el uso de las muestras de probabilidad nos ayuda a evitar algunos de los sesgos de selección descritos en el capítulo 1. En una muestra de probabilidad, el entrevistador no puede optar por sustituir una persona de aspecto agradable por la persona gruñona elegida para estar en la muestra debido al método de selección aleatoria. Un guardabosques que extrae una muestra de probabilidad de árboles no puede, simplemente, medir los que están cercanos al camino, sino que debe medir los árboles designados para ser incluidos en la muestra. La extracción de una muestra de probabilidad es mucho más difícil que la extracción de una muestra de conveniencia, pero el proce-

dimiento de muestreo de probabilidad garantiza que cada unidad de la población pueda aparecer en la muestra y proporciona información útil para evaluar la precisión del estadístico calculado a partir de la muestra.

Dentro del marco de referencia del muestreo de probabilidad, podemos cuantificar qué tan probable es que nuestra muestra sea "buena". No es posible garantizar que una sola muestra de probabilidad sea representativa de la población con respecto a las características de interés, pero podemos cuantificar la frecuencia con que las muestras cumplirán cierto criterio de representatividad. El concepto es el mismo que el de los intervalos de confianza: no sabemos si el intervalo de confianza al 95% que ha sido construido para la media contiene al valor real de esta última. Sin embargo, sabemos que si repetimos el procedimiento una y otra vez, podemos esperar que en el 95% de los casos los intervalos de confianza resultantes contengan el valor verdadero de la media.

Sea y_i una característica asociada a la i-ésima unidad de la población. Consideramos a y_i como una cantidad fija; si la granja 723 está incluida en la muestra, entonces conocemos con exactitud la cantidad de maíz producida por la granja 723, y_{723}.

EJEMPLO 2.1 Para ilustrar estos conceptos, analicemos una situación artificial en la cual conocemos el valor de y_i para cada una de las $N = 8$ unidades de la población completa. El conjunto de índices para la población es:

$$\mathcal{U} = \{1, 2, 3, 4, 5, 6, 7, 8\}.$$

Los valores de y_i son:

i	1	2	3	4	5	6	7	8
yi	1	2	3	4	5	6	7	8

Hay 70 posibles muestras de tamaño 4 que se pueden extraer sin reemplazo de esta población; las muestras aparecen enumeradas en el archivo samples.dat. Si elegimos la muestra que consta de las unidades {1, 2, 3, 4}, los valores correspondientes de y_i serían 1, 2, 4 y 4. Los valores de y_i para la muestra {2, 3, 6, 7} son 2, 4, 7 y 7. Definimos $P(\mathcal{S}) = 1/70$ para cada subconjunto de tamaño 4 de $\mathcal{U}$. Como verá después de leer la sección 2.3, este diseño es una muestra aleatoria simple sin reemplazo. Cada unidad está exactamente en 35 de las muestras posibles, de modo que $\pi_i = 1/2$ para $i = 1, 2, \ldots, 8$.

Se utiliza un mecanismo aleatorio para elegir una de las 70 muestras posibles. En este ejemplo, como hemos enumerado todas las muestras posibles, la utilización de un mecanismo probable consistiría en generar un número aleatorio entre 1 y 70 y elegir la muestra correspondiente. Con las poblaciones de gran tamaño, la cantidad de muestras es tan grande que en la práctica las propias unidades se eligen al azar de acuerdo con ciertas probabilidades especificadas de antemano. ■

La mayor parte de los resultados en muestreo se basan en la **distribución de muestreo** de una estadística, la distribución de los distintos valores de la estadística obtenidos al considerar todas las muestras posibles de la población. Una distribución de muestreo es un ejemplo de distribución de probabilidad discreta.

Suponga que queremos utilizar una muestra para estimar una cantidad relacionada con la población; por ejemplo, la población total $t = \sum_{\mathcal{U}} y_i$. Una estimación que podemos utilizar para t es $\hat{t}_S = N\bar{y}_S$, donde $\bar{y}_S$ es el promedio de las y_i en $\mathcal{S}$, la muestra elegida. En nuestro ejemplo, $t = 40$. Si la muestra $\mathcal{S}$ consta de las unidades 1, 3, 5 y 6, entonces $\hat{t} = 8 \times (1+4+7+7)/4 = 38$. Como en este caso conocemos la población entera, determinamos $\hat{t}_S$ para cada una

de las 70 muestras posibles. Las probabilidades de selección de las muestras proporcionan la distribución de muestreo de $\hat{t}$:

$$P\{\hat{t} = k\} = \sum_{\mathcal{S}: \hat{t}_S = k} P(\mathcal{S}).$$

La suma se realiza sobre todas las muestras $\mathcal{S}$ para las cuales $\hat{t}_s = k$. Conocemos la probabilidad $P(\mathcal{S})$ con la cual elegimos una muestra $\mathcal{S}$, debido a que consideramos una muestra de probabilidad.

EJEMPLO 2.2 La distribución de muestreo de $\hat{t}$ para la población y el diseño de muestreo del ejemplo 2.1 se deduce completamente de las probabilidades de selección de las diversas muestras. Las cuatro muestras ({3,4,5,6}, {3,4,5,7}, {3,4,6,7} y {1,5,6,7}) producen la estimación $\hat{t} = 44$, de modo que $P\{\hat{t} = 44\} = 4/70$. Para este ejemplo, podemos escribir la distribución de muestreo de $\hat{t}$, pues conocemos los valores de toda la población.

k	22	28	30	32	34	36	38	40	42	44	46	48	50	52	58
$P\{\hat{t} = k\}$	1/70	6/70	2/70	3/70	7/70	4/70	6/70	12/70	6/70	4/70	7/70	3/70	2/70	6/70	1/70

La figura 2.1 muestra la distribución de muestreo. ■

El **valor esperado** de $\hat{t}$, $E[\hat{t}]$, es la media de la distribución de muestreo de $\hat{t}$:

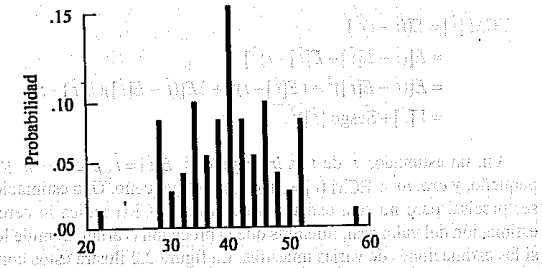
$$E[t] = \sum_s P(S)\hat{t}_s = \sum_k kP(\hat{t} = k). \tag{2.2}$$

El valor esperado de la estadística es el promedio ponderado de los valores de muestra posibles de la estadística, donde el peso es la probabilidad de que aparezca ese valor particular de la estadística.

El **sesgo de estimación** del estimador $\hat{t}$ es:

$$\text{Sesgo}[\hat{t}] = E[\hat{t}] - t. \tag{2.3}$$

FIGURA 2.1
Distribución de muestreo de la muestra total en el ejemplo 2.2.



Si Sesgo$[\hat{t}\,]=0$, decimos que el estimador $\hat{t}$ es **insesgado** para t. Para los datos del ejemplo 2.1, el valor esperado de $\hat{t}$ es

$$E[\hat{t}\,]=\frac{1}{70}(22)+\frac{6}{70}(28)+\cdots+\frac{1}{70}(58)=40.$$

Así, el estimador es insesgado.

Observe que la definición matemática del sesgo, en la ecuación (2.3), *no* es igual a la definición del sesgo de selección o del sesgo de medición descritos en el capítulo 1. Todos indican una desviación sistemática con respecto al valor de la población, pero desde distintas fuentes. El sesgo de selección se debe al método de selección de la muestra: con frecuencia, el investigador actúa como si cada muestra posible S tuviese la misma probabilidad de ser elegida, pero en la realidad algunos subconjuntos de la población podrían tener una probabilidad distinta de selección. Con una subcobertura, por ejemplo, la probabilidad de incluir una unidad que no esté en el marco de muestreo es nula. El sesgo de medición significa que las y_i no son, en realidad, las cantidades de interés, así que aunque $\hat{t}$ pueda estar insesgado en el sentido de (2.3) para $t=\sum_{i=1}^{N}y_i$, el propio t podría no ser el total de interés. El sesgo de estimación significa que el estimador elegido produce un sesgo; por ejemplo, si utilizamos $\hat{t}_S=\sum_{i\in S}y_i$ y no realizamos un censo, $\hat{t}$ sería sesgado. Para ilustrar estas diferencias, suponga que queremos estimar la altura promedio de los actores masculinos que pertenecen a una Asociación de Actores. Habría un sesgo de selección si usted extrae una muestra de conveniencia de los actores en el conjunto; tal vez sea, más o menos, probable que los actores más altos estén trabajando. Habría un sesgo de medición si su cinta métrica agrega, de manera errónea, 3 cm a la altura de cada actor. Habría un sesgo de estimación si usted considera una muestra aleatoria simple a partir de la lista de todos los actores en la Asociación pero estima la altura media al considerar la altura promedio de los seis hombres más bajos de la muestra; el procedimiento de muestreo es bueno, pero el estimador es malo.

La **varianza** de la distribución de muestreo de $\hat{t}$ es

$$\begin{aligned}V[\hat{t}\,]&=E[(\hat{t}-\mathrm{E}[\hat{t}\,])^2]\\&=\sum_{\substack{\text{Todas las muestras}\\ \text{posibles } S}}P(S)(\hat{t}_S-E[\hat{t}\,])^2\end{aligned}\qquad(2.4)$$

Para los datos del ejemplo 2.1,

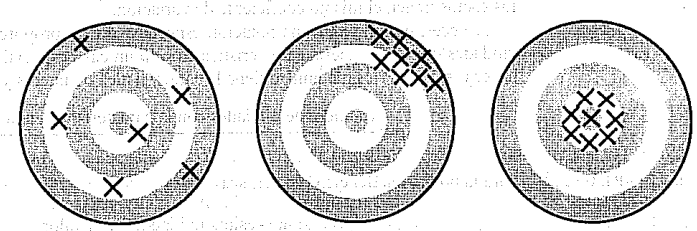
$$V[\hat{t}\,]=\frac{1}{70}(22-40)^2+\cdots+\frac{1}{70}(58-40)^2=\frac{3840}{70}=54.86.$$

Como a veces utilizamos estimadores sesgados, con frecuencia usamos el **error cuadrático medio** (ECM), en vez de la varianza, para medir la exactitud de un estimador:

$$\begin{aligned}\mathrm{ECM}[\hat{t}\,]&=E[(\hat{t}-t)^2]\\&=E[(\hat{t}-E[\hat{t}\,]+E[\hat{t}\,]-t)^2]\\&=E[(\hat{t}-E[\hat{t}\,])^2+(E[\hat{t}\,]-t)^2+2E[(\hat{t}-E[\hat{t}\,])(E[\hat{t}\,]-t)]\\&=V[\hat{t}\,]+\mathrm{Sesgo}[\hat{t}\,])^2,\end{aligned}$$

Así, un estimador $\hat{t}$ de t es *insesgado* si $E[\hat{t}\,]=t$, *preciso* si $V[\hat{t}\,]=E[(\hat{t}-E[\hat{t}\,])^2]$ es pequeño, y *exacto* si $\mathrm{ECM}[\hat{t}\,]=E[(\hat{t}-t)^2]$ es pequeño. Una estimación muy sesgada puede ser precisa, pero no será exacta; la exactitud (ECM) indica lo cerca que se encuentra la estimación del valor real, mientras que la precisión (varianza) mide lo cerca que están entre sí las estimaciones de varias muestras. La figura 2.2 ilustra estos conceptos.

FIGURA 2.2
Arqueros insesgados, precisos y exactos. El arquero A es insesgado: la posición promedio de todas las flechas está en el centro del blanco. El arquero B es preciso pero no insesgado: todas las flechas están cerca entre sí, pero de manera sistemática están alejadas del centro. El arquero C es preciso: todas las flechas están cerca entre sí y cerca del centro del blanco.



Arquero A Arquero B Arquero C

En resumen, la población finita $\mathcal{U}$ consiste en unidades $\{1,2,\ldots,N\}$ cuyos valores medidos son $\{y_1,y_2,\ldots,y_N\}$. Elegimos una muestra $\mathcal{S}$ de n unidades de U al utilizar las probabilidades de selección que definen el diseño del muestreo. Las y son cantidades fijas, aunque desconocidas (a menos que la unidad esté en nuestra muestra $\mathcal{S}$). Excepto por supuestos adicionales, la única información que tenemos acerca del conjunto de y en la población está en el conjunto $\{y_i: i\in\mathcal{S}\}$.

Usted podrá estar interesado en muchas cantidades distintas que están relacionadas con la población. Sin embargo, históricamente, el principal ímpetu para el desarrollo de la teoría de las encuestas con muestras ha sido la estimación de medias y totales de la población. Suponga que queremos estimar el número total de personas que tienen diabetes en Canadá o el número promedio de frutas producidas por cada árbol de naranjas. La población total es:

$$t=\sum_{i=1}^{N}y_i,$$

y la media de la población es:

$$\bar{y}_U=\frac{1}{N}\sum_{i=1}^{N}y_i.$$

Casi todas las poblaciones exhiben cierta variabilidad; por ejemplo, las familias tienen diversos ingresos y los árboles tienen distintos diámetros. Definimos la **varianza** de los valores de la población en torno a la media como:

$$S^2=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{y}_U)^2.\qquad(2.5)$$

La **desviación estándar** de la población es $S=\sqrt{S^2}$.

Con frecuencia, la desviación estándar de la población tiene cierta relación con la media. Una población de árboles podría tener una altura media de 10 metros y una desviación estándar de 1 m. Sin embargo, una población de pequeños cactos, con una altura media de 10 cm, podría tener una desviación estándar de 1 cm. El **coeficiente de variación** (CV) es

una medida de variabilidad relativa, que se puede definir, para $\bar{Y}_U \neq 0$ como

$$CV(y) = \frac{S}{\bar{Y}_U}.$$

Si la altura del árbol se mide en metros, entonces $\bar{Y}_U$ y S también se mide en metros. El coeficiente de variación no depende de la unidad de medida. En este ejemplo, los arbolitos y los cactos tienen el mismo coeficiente de variación.

A veces, es útil tener una notación especial para las proporciones. La proporción de unidades que tienen cierta característica es sólo un caso especial de la media, obtenida al hacer $y_i = 1$ si la i-ésima unidad tiene la característica de interés y $y_i = 0$ si no la tiene. Se

$$p = \frac{\text{(número de unidades con la característica en la población)}}{N}$$

EJEMPLO 2.3 Para la población del ejemplo 2.1, sea

$$y_i = \begin{cases} 1 & \text{si la } i\text{-ésima unidad tiene el valor 7} \\ 0 & \text{si la } i\text{-ésima unidad no tiene el valor 7} \end{cases}$$

Sea $\hat{p}s = \sum_{i=1}^{4} y_i / 4$ la proporción de sietes en la muestra. La lista de todas las posibles muestras en el archivo de datos tiene 5 muestras sin sietes, 30 muestras con un siete, 30 muestras con dos sietes y 5 muestras con tres sietes. Como cada una de las posibles muestras se elige con una probabilidad de 1/70, la distribución de muestreo de $\hat{p}$ es[1]:

k	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$
$P(\hat{p} = k)$	$\frac{5}{70}$	$\frac{30}{70}$	$\frac{30}{70}$	$\frac{5}{70}$

■

23
Muestreo aleatorio simple

El muestreo aleatorio simple es la forma más sencilla de muestreo de probabilidad y propor-ciona la base teórica de las formas más complejas. Existen dos formas de extraer una muestra aleatoria simple: con reemplazo, donde la misma unidad se puede incluir más de una vez en la muestra, y sin reemplazo, donde todas las unidades de la muestra son distintas.

Una **muestra aleatoria simple con reemplazo,** de tamaño n, obtenida a partir de una población de N unidades, se puede pensar como la extracción de n muestras independientes de tamaño 1. Una unidad se extrae de la población al azar, para ser la primera unidad muestreada, con una probabilidad de $1/N$. Luego, la unidad muestreada se reemplaza en la población, y una segunda unidad se elige al azar con una probabilidad de $1/N$. Este procedimiento se repite hasta que la muestra tiene n unidades y puede tener duplicados de la población.

Sin embargo, en el muestreo de poblaciones finitas, el muestreo de una persona que se repite dos veces no proporciona más información. Por lo general preferimos el muestreo sin

[1] En el ejercicio B.2 (p. 427) se deduce de otra manera la distribución de muestreo.

reemplazo, de modo que la muestra no contenga duplicados. Una **muestra aleatoria simple sin reemplazo** de tamaño n se elige de modo que cada subconjunto posible de n unidades distintas en la población tiene la misma probabilidad de ser elegido en la muestra. Existen $\binom{N}{n}$ muestras posibles (véase el Apéndice B), y cada una es igualmente probable, de modo que la probabilidad de elegir cualquier muestra individual $\mathcal{S}$ de n unidades es:

$$P(\mathcal{S}) = \frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}.$$

Como consecuencia de esta definición, la probabilidad de que cualquier unidad dada aparezca en la muestra es de n/N, como se muestra posteriormente en la ecuación (2.18).

Para extraer una muestra aleatoria simple, usted necesita una lista de todas las unidades de observación que pertenecen a la población; esta lista es el marco de muestreo. En este tipo de muestra, la unidad de muestreo y la unidad de observación coinciden. Cada unidad tiene asignado un número, y se elige una muestra de modo que (1) cada unidad tenga la misma posibilidad de aparecer en la muestra y (2) la selección de una unidad no tenga influencia de las demás unidades ya elegidas. Esto se puede ilustrar con la extracción de los números que se encuentran dentro de un sombrero, en la práctica se acostumbra utilizar números seudoaleatorios generados por computadora para elegir la muestra.

EJEMPLO 2.4

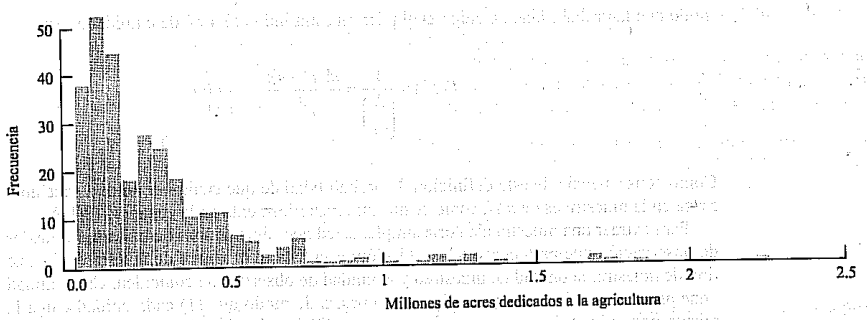
El gobierno de Estados Unidos realiza un censo de agricultura cada cinco años; para ello, reúne datos de todas las granjas (definidas como un lugar donde se producen y venden $1000 o más en productos agrícolas) de los 50 estados que conforman la Unión Americana.[2] El censo de agricultura proporciona los datos sobre el número de granjas, los acres dedicados a las granjas, su tamaño, los resultados de varias cosechas y una amplia variedad de otras medidas en agricultura, para cada uno de los $N = 3078$ condados o equivalentes en Estados Unidos. El archivo agpop.dat contiene la información de 1982, 1987 y 1992 sobre el número de granjas, superficies dedicadas a la agricultura, número de granjas con menos de 9 acres y número de granjas con más de 1000 acres para esta población.

Para extraer una muestra aleatoria simple de tamaño 300 de esta población, el autor de este libro generó 300 números aleatorios entre 0 y 1 en la computadora y los multiplicó cada uno por 3078, redondeando a enteros. Este procedimiento genera una muestra aleatoria simple con reemplazo. Si la población es grande con respecto a la muestra, es probable que cada unidad de la muestra aparezca sólo una vez en la lista. Sin embargo, en este caso, 13 de los 300 números están repetidos. Descartamos los duplicados y los reemplazamos con nuevos números generados en forma aleatoria entre 1 y 3078, hasta que los 300 números sean distintos. En los ejercicios y en el apéndice E se describen otros métodos para elegir una muestra aleatoria simple.

En un principio, podría parecer que los condados elegidos para estar en la muestra no son muy aleatorios. Por ejemplo, los condados 2840, 2841 y 2842 están en la muestra, mientras que ninguno de los condados del 2740 al 2787 aparece. La muestra contiene el 18% de los condados de Virginia, pero ninguno de los condados de Alaska, Arizona, Connecticut, Delaware, Hawaii, Rhode Island, Utah o Wyoming. Existe una tendencia na-

[2] El censo de agricultura era realizado anteriormente por la Oficina de Censos (Bureau of the Census); actualmente es realizado por el Servicio Nacional de Estadísticas Agrícolas (National Agricultural Statistics Service, NASS) de Estados Unidos. Puede consultar más información acerca del censo y algunos datos de Internet, a través del material del NASS en www.fedstats.gov.

FIGURA 2.3
Histograma: número de acres dedicados a la agricultura en 1992, para una muestra aleatoria simple de 300 condados. Observe la asimetría de los datos. La mayor parte de los condados tienen menos de 500,000 acres en granjas; sin embargo, algunos condados tienen más de 1.5 millones de acres en granjas.



tural a querer "ajustar" la lista de números aleatorios, para difundirla un poco más. Si usted quiere una muestra aleatoria, debe resistirse a esta tentación. Varios estudios, comenzando con Neyman (1934), han mostrado que con frecuencia las muestras a propósito no representan la población en variables clave. Si usted sustituye, deliberadamente, otros condados en la muestra generada en forma aleatoria, tal vez haga concordar a la población con alguna característica particular, como la distribución geográfica; sin embargo, es probable que no la haga concordar en características de interés, como el número de granjas o el tamaño promedio de una granja. Si usted quiere garantizar que todos los estados queden representados, no ajuste a propósito la muestra elegida al azar; extraiga una muestra estratificada (que analizaremos en el capítulo 4).

Observemos la variable *acres92*, el número de acres dedicados a la agricultura. Un pequeño número de condados en la población omite esa información (en algunos casos, los datos se omiten para no mostrar datos acerca de las granjas individuales). Primero verificamos el efecto que tiene la ausencia de datos sobre nuestra muestra. Por fortuna, nuestra muestra no tiene datos faltantes (véase en el ejercicio 7 la probabilidad de que esto ocurra). La figura 2.3 muestra un histograma de los acres dedicados a la agricultura en cada uno de los 300 condados.

Para estimar la media de la población $\hat{y}_U$ en una muestra aleatoria simple, utilizamos la media de la muestra

$$\bar{y}S = \frac{1}{n}\sum_{i\in S} y_i \,. \tag{2.6}$$

En lo sucesivo utilizaremos $\bar{y}$ para referirnos a la media de la muestra y eliminaremos el subíndice $\mathcal{S}$ a menos que se necesite para proporcionar una mayor claridad. Como veremos en la sección 2.7, $\bar{y}$ es un estimador insesgado de la media de la población $\bar{y}_U$, y la varianza de $\bar{y}$ es

$$V(\bar{y}) = \frac{S^2}{n}\left(1-\frac{n}{N}\right) \tag{2.7}$$

para S^2 definida en la ecuación (2.5). La varianza $V(\bar{y})$ mide la variabilidad entre las estimaciones de $\bar{y}_U$ de varias muestras.

El factor $(1-n/N)$ se conoce como la **corrección para poblaciones finitas**. Intuitivamente, debemos hacer esta corrección, pues para las poblaciones pequeñas, la **fracción de muestreo** n/N es mayor, la información que tenemos de la población también es mayor y, por lo tanto, la varianza es menor. Si $N = 10$ y elegimos como muestra a las 10 observaciones, esperaríamos que la varianza de $\bar{y}$ sea 0 (lo cual es cierto). Si $N = 10$, sólo existe una muestra posible $\mathcal{S}$ de tamaño 10 sin reemplazo, con $\bar{y}_S = \bar{y}_U$, de modo que no existe variabilidad debida a la extracción de una muestra. Para un censo, la corrección para poblaciones finitas, al igual que $V(\bar{y})$, se anulan. Cuando la fracción de muestreo n/N es grande en una muestra aleatoria simple sin reemplazo, la muestra se parece mucho a un censo, el cual no tiene variabilidad de muestreo.

Para la mayor parte de las muestras extraídas de poblaciones que poseen tamaños muy grandes, la corrección es casi 1. Para las poblaciones grandes, el tamaño de la muestra extraída es el que determina la precisión del estimador (y no el porcentaje de población muestreada): si su sopa está bien revuelta, sólo necesita dos o tres cucharadas para probar el sazón, así tenga uno o veinte litros de sopa. Una muestra de tamaño 100 de una población de 100,000 unidades tiene casi la misma precisión que una muestra de tamaño 100 de una población de 100 millones de unidades:

$$V[\bar{y}] = \frac{S^2}{100}\frac{99{,}900}{100{,}000} = \frac{S^2}{100}(0.999) \qquad \text{para } N = 100{,}000$$

$$V[\bar{y}] = \frac{S^2}{100}\frac{99{,}999{,}900}{100{,}000{,}000} = \frac{S^2}{100}(0.999999) \qquad \text{para } N = 100{,}000{,}000$$

La varianza de la población S^2, que depende de los valores para toda la población, es desconocida. La estimamos mediante la varianza de la muestra:

$$s^2 = \frac{1}{n-1}\sum_{i\in S}(y_i-\bar{y})^2. \tag{2.8}$$

Un estimador insesgado de la varianza de $\bar{y}$ es (véase la sección 2.7)

$$\hat{V}[\bar{y}] = \left(1-\frac{n}{N}\right)\frac{s^2}{n}. \tag{2.9}$$

Por lo general, no reportamos la varianza estimada de $\bar{y}$, sino su raíz cuadrada, el **error estándar** (EE):

$$\text{EE}[\bar{y}] = \sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}. \tag{2.10}$$

El coeficiente de variación de una estimación proporciona una medida de la variabilidad relativa de una estimación. Es igual al error estándar dividido entre la media (definido sólo cuando la media no se anula):

$$\widehat{\text{CV}}(\bar{y}) = \frac{\text{EE}[\bar{y}]}{\bar{y}}. \tag{2.11}$$

Todos estos resultados se aplican a la estimación del total de la población, t, pues

$$t = \sum_{i=1}^{N} y_i = N\bar{y}_U.$$

Para estimar t, utilizamos el estimador insesgado

$$\hat{t} = N\bar{y}. \tag{2.12}$$

Entonces, de la ecuación (2.7),

$$V[\hat{t}] = N^2 V[\bar{y}] = N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n} \tag{2.13}$$

y

$$\hat{V}[\hat{t}] = N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}. \tag{2.14}$$

EJEMPLO 2.5 Para los datos del ejemplo 2.4, $N = 3078$ y $n = 300$, de modo que la fracción de muestreo es 300/3078 = 0.097. Las estadísticas de la muestra son $\bar{y} = 297{,}897, s = 344{,}551.9$, y $\hat{t} = N\bar{y} = 916{,}927{,}110$. Los errores estándar son

$$\text{EE}[\bar{y}] = \sqrt{\frac{s^2}{n}\left(1-\frac{300}{3078}\right)} = 18{,}898.434428$$

y

$$\text{EE}[\hat{t}] = (3078)\,(18{,}898.434428) = 58{,}169{,}381$$

y el coeficiente estimado de variación es

$$\widehat{\text{CV}}[\hat{t}] = \widehat{\text{CV}}[\bar{y}]$$
$$= \frac{\text{EE}[\bar{y}]}{\bar{y}}$$
$$= \frac{18{,}898.434428}{297{,}897}$$
$$= 0.06344.$$

Como los datos son muy asimétricos, también debemos informar de la mediana del número de acres necesarios para desarrollar la actividad agrícola en un condado, que es 196,717. ■

También podríamos querer estimar la proporción de condados, en el ejemplo 2.4, con menos de 200,000 acres de granja. Como la estimación de una proporción es un caso particular de estimación de una media, los resultados de las ecuaciones (2.6) – (2.11) también son válidos para las proporciones y asumen una forma más sencilla. Suponga que queremos estimar la proporción de las unidades en la población que tengan cierta característica; llamemos a esta proporción p. Definimos y_i como 1 si la unidad tiene la característica y 0 en caso contrario. Entonces $p = \sum_{i=1}^{N} y_i / N = \bar{y}_U$ y estimamos p como $\hat{p} = \bar{y}$. En consecuencia, $\hat{p}$ es un estimador insesgado de p. Para la respuesta y_i, que asume los valores 0 o 1,

$$S^2 = \frac{\sum_{i=1}^{N}(y_i - p)^2}{N-1} = \frac{\sum_{i=1}^{N} y_i^2 - 2p\sum_{i=1}^{N} y_i + Np^2}{N-1} = \frac{N}{N-1}p\,(1-p).$$

Así, (2.7) implica que

$$V[\hat{p}] = \left(\frac{N-n}{N-1}\right)\frac{p\,(1-p)}{n}. \tag{2.15}$$

Además,

$$s^2 = \frac{1}{n-1}\sum_{i\in S}(y_i - \hat{p})^2 = \frac{n}{n-1}\hat{p}\,(1-\hat{p}).$$

Así, de (2.9),

$$\hat{V}[\hat{p}] = \left(1-\frac{n}{N}\right)\frac{\hat{p}\,(1-\hat{p})}{n-1}. \tag{2.16}$$

EJEMPLO 2.6 Para la muestra descrita en el ejemplo 2.4, la proporción estimada de condados con menos de 200,000 acres dedicados a la agricultura es:

$$\hat{p} = \frac{153}{300} = 0.51$$

con un error estándar

$$\text{EE}(\hat{p}) = \sqrt{\left(1-\frac{300}{3078}\right)\frac{(0.51)\,(0.49)}{299}} = 0.0275. \blacksquare$$

2.4 Intervalos de confianza

Al realizar una encuesta con muestras, no basta con informar sobre la altura promedio de los árboles o la proporción en la muestra de los votantes con la intención de votar por el candidato B en la siguiente elección. También debe proporcionar una indicación de la exactitud de sus estimaciones. En estadística se utilizan los **intervalos de confianza** (IC) para indicar la exactitud de una estimación.

Un intervalo de confianza al 95% se explica con frecuencia de manera heurística: si extraemos muestras de nuestra población, una y otra vez, y construimos un intervalo de confianza mediante este procedimiento, esperamos que el 95% de los intervalos resultantes incluyan al valor verdadero del parámetro de la población.

En el muestreo de probabilidad, a partir de una población finita, sólo existe un número finito de muestras posibles y conocemos la probabilidad con la que podemos elegir cada una; si pudiéramos generar todas las muestras posibles a partir de la población, calcularíamos el nivel de confianza exacto para un procedimiento de intervalos de confianza.

EJEMPLO 2.7 Regresemos al ejemplo 2.1, donde conocemos a toda la población. Elijamos un procedimiento arbitrario para calcular un intervalo de confianza, al construir las estimaciones de intervalo para t como

$$\text{IC}(\mathcal{S}) = [\hat{t}_{\mathcal{S}} - 4s_{\mathcal{S}}, \hat{t}_{\mathcal{S}} + 4s_{\mathcal{S}}].$$

No hay una razón teórica para elegir este procedimiento, pero ilustrará el concepto de intervalo de confianza. Definimos $u(\mathcal{S})$ como 1 si IC $(\mathcal{S})$ contiene al verdadero valor de la población, 40, y 0 en caso contrario. Como conocemos la población, podemos calcular el intervalo de confianza IC $(\mathcal{S})$ y el valor de $u(\mathcal{S})$ para cada muestra posible $\mathcal{S}$. Algunos de los 70 intervalos de confianza aparecen en la tabla 2.1 (todas las entradas están redondeadas a dos cifras decimales).

TABLA 2.1
Intervalos de confianza para las muestras posibles en una población pequeña

Muestra $\mathcal{S}$	$y_i, i \in \mathcal{S}$	$\hat{t}_{\mathcal{S}}$	$s_{\mathcal{S}}$	IC($\mathcal{S}$)	$u(\mathcal{S})$
(1, 2, 3, 4)	1, 2, 4, 4	22	1.50	[16.00, 28.00]	0
(1, 2, 3, 5)	1, 2, 4, 7	28	2.65	[17.42, 38.58]	0
(1, 2, 3, 6)	1, 2, 4, 7	28	2.65	[17.42, 38.58]	0
(1, 2, 3, 7)	1, 2, 4, 7	28	2.65	[17.42, 38.58]	0
(1, 2, 3, 8)	1, 2, 4, 8	30	3.10	[17.62, 42.38]	1
(1, 2, 4, 5)	1, 2, 4, 7	28	2.65	[17.42, 38.58]	0
(1, 2, 4, 6)	1, 2, 4, 7	28	2.65	[17.42, 38.58]	0
(1, 2, 4, 7)	1, 2, 4, 7	28	2.65	[17.42, 38.58]	0
(1, 2, 4, 8)	1, 2, 4, 8	30	3.10	[17.62, 42.38]	1
(1, 2, 5, 6)	1, 2, 4, 7	34	3.20	[21.19, 46.81]	1
⋮	⋮	⋮	⋮	⋮	⋮
(2, 3, 4, 8)	2, 4, 4, 8	36	2.52	[25.93, 46.07]	1
(2, 3, 5, 6)	2, 4, 7, 7	40	2.45	[30.20, 49.80]	1
(2, 3, 5, 7)	2, 4, 7, 7	40	2.45	[30.20, 49.80]	1
(2, 3, 5, 8)	2, 4, 7, 8	42	2.75	[30.98, 53.02]	1
(2, 3, 6, 7)	2, 4, 7, 7	40	2.45	[30.20, 49.80]	1
(2, 3, 6, 8)	2, 4, 7, 8	42	2.75	[30.98, 53.02]	1
⋮	⋮	⋮	⋮	⋮	⋮
(4, 5, 6, 7)	4, 7, 7, 7	50	1.50	[44.00, 56.00]	0
(4, 5, 6, 8)	4, 7, 7, 8	52	1.73	[45.07, 58.93]	0
(4, 5, 7, 8)	4, 7, 7, 8	52	1.73	[43.00, 58.93]	0
(4, 6, 7, 8)	4, 7, 7, 8	52	1.73	[45.07, 58.93]	0
(5, 6, 7, 8)	7, 7, 7, 8	58	0.50	[56.00, 60.00]	0

Cada intervalo de confianza contiene o no a 40, el total a la población. La afirmación de probabilidad en el intervalo de confianza se hace con respecto de la colección de todas las muestras posibles; para este procedimiento de construcción de intervalos de confianza y población, el nivel de confianza es:

$$\sum_{\mathcal{S}} P(\mathcal{S})u(\mathcal{S}) = 0.77.$$

Esto significa que si extraemos una muestra aleatoria simple de cuatro elementos, sin reemplazo, de esta población formada por ocho elementos, hay un 77% de posibilidades de que nuestra muestra sea una de las "buenas" cuyo intervalo de confianza contenga el valor verdadero, es decir, 40. Así, este procedimiento crea un intervalo de confianza del 77%.

Por supuesto, en la vida real, sólo extraemos una muestra y no conocemos el valor t de la población total. Sin más estudios, no podemos saber si la muestra obtenida es de las "buenas", como $\mathcal{S} = \{2,3,5,6\}$ o de las "malas", como $\mathcal{S} = \{4,6,7,8\}$. El intervalo de confianza sólo es una afirmación de probabilidad acerca de la frecuencia con la que esperamos estar en lo correcto. ■

En la práctica, no conocemos los valores de las estadísticas de todas las muestras posibles, de modo que no podemos calcular el coeficiente exacto de confianza para un procedimiento como el del ejemplo 2.7. En su curso de introducción a la estadística, usted confió en resultados **asintóticos** (cuando el tamaño de la muestra tiende al infinito) para construir intervalos de confianza para una media desconocida μ. El teorema del límite central dice que si tenemos una muestra aleatoria con reemplazo, entonces la distribución de probabilidad de $\sqrt{n}(\bar{y} - \mu)$ converge a una distribución normal cuando n, el tamaño de la muestra, tiende a infinito.

Sin embargo, en la mayor parte de las encuestas con muestras, sólo tenemos una población finita. Para utilizar los resultados asintóticos en el muestreo de poblaciones finitas, suponemos que nuestra población es a su vez parte de una **superpoblación** mayor; esta superpoblación es parte de una superpoblación aún mayor, y así sucesivamente, hasta que las superpoblaciones sean tan grandes como queramos. Nuestra población está contenida en una serie creciente de poblaciones finitas. Esta contención nos puede dar propiedades de consistencia y normalidad asintótica. Uno puede imaginar las superpoblaciones como "universos alternativos" en el sentido de la ciencia ficción: lo que podría haber ocurrido si las circunstancias hubieran sido distintas.

Hájek (1960) demuestra un teorema del límite central para el muestreo aleatorio simple sin reemplazo. En términos prácticos, el teorema de Hájek dice que si se cumplen ciertas condiciones técnicas y si n, N y $N-n$ son "suficientemente grandes", entonces la distribución de muestreo de:

$$\frac{\bar{y} - \bar{y}_U}{\sqrt{\left(1 - \frac{n}{N}\right)}\frac{S}{\sqrt{n}}}$$

es aproximadamente normal (gaussiana) con media 0 y varianza 1. Un intervalo de confianza de $100(1-\alpha)\%$ en una muestra de gran tamaño para la media de la población es:

$$\left[\bar{y} - z_{\alpha/2}\sqrt{1 - \frac{n}{N}}\frac{S}{\sqrt{n}}, \bar{y} + z_{\alpha/2}\sqrt{1 - \frac{n}{N}}\frac{S}{\sqrt{n}}\right],$$

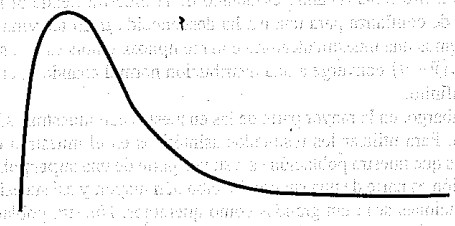
donde $z_{\alpha/2}$ es el percentil $(1-\alpha/2)$ de la distribución normal estándar. Por lo general, S no se conoce, de modo que en las muestras grandes se sustituye s en vez de S, con un ligero cambio en la aproximación; el intervalo de confianza para una muestra grande es:

$$\left[\bar{y} - z_{\alpha/2}\text{EE}(\bar{y}), \bar{y} + z_{\alpha/2}\text{EE}(\bar{y})\right].$$

En el muestreo aleatorio simple sin reemplazo, el 95% de las muestras posibles que pueden elegirse tendrán un intervalo de confianza al 95% para $\bar{y}_U$ que contiene el valor verdadero de $\bar{y}_U$. Cuando $n/N \approx 0$, este intervalo de confianza es igual al intervalo construido en los cursos de introducción a la estadística para el muestreo con reemplazo.

El término impreciso *suficientemente grande* del teorema aparece debido a que la adecuación de la aproximación normal depende de n y de la cercanía con que la población $\{y_i, i = 1,\ldots,N\}$ recuerda a una población generada a partir de la distribución normal. Con frecuencia, el "número mágico" $n = 30$, citado en los libros de introducción a la estadística como un tamaño de muestra "suficientemente grande" para aplicar el teorema del límite central, no basta en los problemas de muestreo de las poblaciones finitas. Muchas de las poblaciones de donde extraemos una muestra son demasiado asimétricas (podríamos medir el ingreso, el número de acres en una granja dedicados al cultivo del maíz, o la concentración de mercurio en los lagos de Minnesota). Para estos ejemplos, esperamos que la mayor

parte de las observaciones sean relativamente pequeñas, pero que unas cuantas sean muy, muy grandes, de modo que un histograma suavizado de toda la población se vería así:



El hecho de pensar las observaciones como generadas a partir de cierta distribución, es útil para decidir si es seguro o no el uso del teorema del límite central. Si piensa que la distribución generatriz se asemeja a la normal, probablemente sea seguro utilizar el teorema con un tamaño de muestra tan pequeño como 50. Si el tamaño de la muestra es demasiado pequeño y la distribución muestral de $\bar{y}$ no es aproximadamente normal, debemos emplear otro método, basado en un supuesto de distribución, para obtener un intervalo de confianza para $\bar{y}_U$. Tales métodos se ajustan a una perspectiva del muestreo basado en modelos (sección 2.8) y son descritos en la sección referente a la Bibliografía, página 460, bajo el título "Estadística matemática y probabilidad".

EJEMPLO 2.8 El histograma de la figura 2.3 exhibe una distribución subyacente para los acres de una granja, alejada de la distribución normal. ¿Es suficientemente grande el tamaño de la muestra como para aplicar el teorema del límite central de Hájek? Para este ejemplo, es probable que la muestra sea lo suficientemente grande como para que la distribución muestral de $\bar{y}$ sea aproximadamente normal (véase el ejercicio 14).

Para los datos del ejemplo 2.4, un intervalo de confianza aproximado al 95% para $\bar{y}_U$ es:

$$[297{,}897-(1.96)(18{,}898.434428),\quad 297{,}897+(1.96)(18{,}898.434428)]$$

$$=[260{,}856,\quad 334{,}938].$$

Para un total de población t, un intervalo de confianza aproximado al 95% es:

$$[916{,}927{,}110-1.96(58{,}169{,}381),\ 916{,}927{,}110+1.96(58{,}169{,}381)]$$

$$=[802{,}915{,}123,\ \ 1{,}030{,}939{,}097].$$

Para estimar las proporciones, el criterio usual de que el tamaño de la muestra es lo suficientemente grande como para usar la distribución normal si $np \geq 5$ y $n(1-p) \geq 5$ es una guía muy útil. Un intervalo de confianza al 95% para la proporción de condados con menos de 200,000 acres para cultivo es

$$0.51 \pm 1.96(0.0275),\ \text{o}\ [0.456, 0.564].$$

Para determinar un intervalo de confianza al 95% para el total de condados con menos de 200,000 acres para cultivo, sólo debemos multiplicar todas las cantidades por N, de modo que la estimación puntual sea $3078(0.51) = 1570$, con un error estándar de $3078 \times \text{EE}(\hat{p}) = 84.65$ y un intervalo de confianza al 95% [1404,1736]. ■

2.5 Estimación del tamaño de la muestra

Con frecuencia, un investigador mide distintas variables y tiene varios objetivos en un estudio. Una persona que quiera diseñar una muestra aleatoria simple debe decidir la cantidad de error de muestreo en las estimaciones que sea tolerable y debe equilibrar la precisión de las estimaciones con el costo del estudio. Aunque se pueden medir muchas variables, con frecuencia un investigador debe centrarse en una o dos respuestas que sean de interés fundamental en el estudio y utilizarlas para estimar el tamaño de la muestra.

Para una única respuesta, siga estos pasos para estimar el tamaño de la muestra:

1. Pregunte: "¿Qué se espera de la muestra? ¿Cuánta precisión necesito?" ¿Cuáles son las consecuencias de los resultados de la muestra? ¿Cuál es la cantidad de error tolerable? Si el estudio que planea pretende medir la tasa mensual de desempleo, usted desearía que sus estimaciones fueran muy precisas, de modo que pueda detectar los cambios en las tasas de desempleo de un mes a otro. Sin embargo, una investigación preliminar necesita, con frecuencia, menos precisión que un estudio en marcha.

   En vez de preguntarse sobre la precisión necesaria, muchas personas se preguntan "¿qué porcentaje de la población debo incluir en mi muestra?" Ésta es una pregunta incorrecta. Excepto en las poblaciones demasiado pequeñas, la precisión se logra mediante el tamaño absoluto de la muestra, no con la proporción de la población cubierta. En la sección 2.3 vimos que la corrección para las poblaciones finitas tiene poco efecto sobre la varianza de la estimación en las poblaciones grandes.
2. Determine una ecuación que relacione el tamaño de muestra n y sus expectativas de la muestra.
3. Estime todas las cantidades desconocidas y despeje n.
4. Si usted tiene poca experiencia en el diseño de encuestas, en este punto verá que el tamaño de muestra calculado en el paso 3 es mucho mayor de lo permisible. Regrese y ajuste algunas de las expectativas de la encuesta y trate de nuevo. En algunos casos, verá que no puede alcanzar la precisión necesaria con los recursos disponibles; en ese caso, deberá decidir incluso si realiza el estudio o no.

Especifique el error tolerable Sólo los investigadores del estudio pueden decir cuál es la precisión necesaria. Con frecuencia, la precisión deseada se expresa en términos absolutos, como

$$P(|\bar{y}-\bar{y}_U| \leq e) = 1-\alpha.$$

El investigador debe decidir cuáles son los valores razonables para α y e; en muchas encuestas, a e se le llama **margen de error**. En muchas de las encuestas con las que se trabaja con personas, donde se mide una proporción, $e = 0.03$ y $\alpha = 0.05$.

A veces, usted querrá lograr cierta precisión relativa. En ese caso, la precisión se puede expresar como:

$$P\left(\left|\frac{\bar{y}-\bar{y}_U}{\bar{y}_U}\right| \leq e\right) = 1-\alpha.$$

Determine una ecuación La ecuación más sencilla que relaciona la precisión y el tamaño de la muestra proviene de los intervalos de confianza de la sección anterior. Para obtener una precisión absoluta, determinamos un valor de n que satisfaga

$$e = z_{\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)}\frac{S}{\sqrt{n}}.$$

Al despejar n, tenemos

$$n = \frac{z_{\alpha/2}^2 S^2}{e^2 + \frac{z_{\alpha/2}^2 S^2}{N}} = \frac{n_0}{1+\frac{n_0}{N}}, \qquad (2.17)$$

donde $n_0 = z_{\alpha/2}^2 S^2 / e^2$. El valor n_0 es el tamaño de muestra para una muestra aleatoria simple con reemplazo.

En los estudios donde una de las principales respuestas de interés es una proporción, con frecuencia, es más fácil utilizar esa respuesta para establecer el tamaño de la muestra. Para las poblaciones grandes, $S^2 \approx p(1-p)$, lo cual alcanza su valor máximo cuando $p = 1/2$. Así, al utilizar $n_0 = 1.96^2/(4e^2)$ obtendremos un intervalo de confianza al 95%, con un ancho de al menos $2e$.

Para calcular un tamaño de muestra con el cual se pueda obtener una precisión relativa dada, sustituimos $e\bar{y}_U$ en vez de e en la ecuación (2.17). Esto produce el tamaño de muestra

$$n = \frac{z_{\alpha/2}^2 S^2}{(e\bar{y}_U)^2 + \frac{z_{\alpha/2}^2 S^2}{N}} = \frac{z_{\alpha/2}^2 \, \mathrm{CV}^2(y)}{e^2 + \frac{z_{\alpha/2}^2 \, \mathrm{CV}^2(y)}{\mathrm{N}}}.$$

Para alcanzar una precisión relativa dada, el tamaño de muestra se puede determinar al utilizar sólo el coeficiente de variación.

EJEMPLO 2.9 Suponga que queremos estimar la proporción de recetas del nuevo recetario de Better Homes & Gardens que no utilizan productos animales. Planeamos extraer una muestra aleatoria simple de las $N = 1251$ recetas y queremos utilizar un intervalo de confianza al 95% con un margen de error de 0.03. Entonces

$$n_0 = \frac{(1.96)^2\left(\frac{1}{2}\right)\left(1-\frac{1}{2}\right)}{(0.03)^2} \approx 1067.$$

El tamaño de la muestra (ignorando la corrección para poblaciones finitas) es grande al compararlo con el tamaño de la población, así que en este caso ajustamos según la corrección indicada para poblaciones finitas y utilizamos

$$n = \frac{n_0}{1+\frac{n_0}{N}} = \frac{1067}{1+\frac{1067}{1251}} = 576. \quad \blacksquare$$

EJEMPLO 2.10 Muchas encuestas de opinión pública especifican el uso de un tamaño de muestra cercano a 110. Ese número proviene del redondeo a centésimos del valor de n_0 del ejemplo 2.9, al observar que el tamaño de la población es muy grande con respecto al de la muestra, por lo que debemos ignorar la corrección para poblaciones finitas. Para las poblaciones grandes, el tamaño de la muestra, y no la proporción de la población que forma la muestra, determina la precisión. ■

Estime las cantidades desconocidas Si estamos interesados en una proporción, podemos utilizar 1/4 como una cota superior para S^2. Para otras cantidades, debemos estimar previamente S^2. Algunos métodos para realizar esta estimación son:

1 Utilice cantidades muestra obtenidas al hacer una prueba preliminar de la encuesta. Éste es probablemente el mejor método, pues la prueba preliminar debe ser similar a la encuesta que se llevará a cabo. Una **muestra piloto**, que es una pequeña muestra extraída para obtener información y que sirve de guía para el diseño de la encuesta principal, puede servir para estimar las cantidades necesarias que establezcan el tamaño de la muestra.

2 Utilice estudios anteriores o datos disponibles en las referencias. Es raro que usted sea la primera persona en estudiar algo relativo a su investigación. Tal vez pueda hallar estimaciones de varianzas ya publicadas en estudios relacionados con los suyos; utilícelas como punto de partida para estimar el tamaño de su muestra. Sin embargo, usted no tiene control sobre la calidad o el diseño de estos estudios, de modo que sus estimaciones podrían no ser confiables o aplicables a su estudio. Además, las estimaciones pueden cambiar con el tiempo y la posición geográfica.

A veces, usted puede utilizar el coeficiente de variación (CV), la razón entre la desviación estándar y la media, para obtener estimaciones de la variabilidad. El CV de una cantidad es una medida de error relativo y tiende a ser más estable en relación con el tiempo y la posición, en comparación con la varianza. Si consideramos una muestra aleatoria de casas en venta el día de hoy en Estados Unidos, veremos que la variabilidad sería mucho mayor que si hubiéramos realizado una encuesta similar en 1930. Pero el precio promedio de una casa también ha aumentado de 1930 a la fecha. Es probable que el CV actual sea parecido al CV de 1930.

3 Si no dispone de nada más, estime la varianza. A veces, una distribución hipotética de los datos nos dará sobre la varianza. Por ejemplo, si usted cree que la población tiene una distribución normal, tal vez no conozca el valor de la varianza, sino que tenga una idea del rango de los datos. Entonces puede estimar S mediante rango/4 o rango/6, pues aproximadamente el 95% de los valores de una población normal están a una distancia de 2 desviaciones estándar de la media y el 99.7% de los valores están a menos de 3 desviaciones estándar de la media.

EJEMPLO 2.11 Antes de extraer la muestra de tamaño 300 del ejemplo 2.4, se extrae una muestra piloto de tamaño 30 de la población. Un condado en la muestra piloto de tamaño 30 no tiene el valor de *acres92*; la desviación estándar de la muestra para las otras 29 observaciones es 519,085. Al usar este valor y un margen de error deseado de 60 000,

$$n_0 = (1.96)^2 \frac{519{,}085^2}{60{,}000^2} = 288.$$

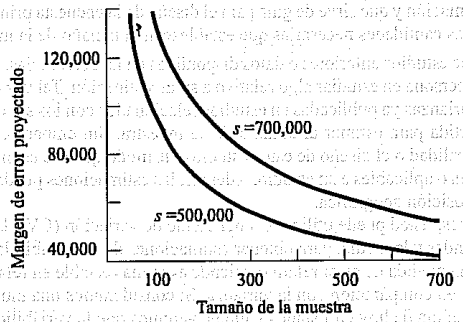
Consideramos una muestra de tamaño 300 en caso de que la desviación estándar estimada, a partir de la muestra piloto, sea demasiado pequeña. Además, ignoramos la corrección para las poblaciones finitas al calcular el tamaño de la muestra; en la mayoría de las poblaciones, esta corrección tendrá poco efecto sobre el tamaño de la muestra.

Usted también puede ver algunas de las posibles consecuencias del uso de diversos tamaños de muestra de manera gráfica. La figura 2.4 enseña el valor de $(1.96)s/\sqrt{n}$, para un rango de tamaños de muestra entre 50 y 700, y para dos valores posibles de la desviación estándar s. La gráfica muestra que si ignoramos la corrección para poblaciones finitas y si la desviación estándar es cercana a 50 0000, una muestra de tamaño 300 dará un margen de error cercano a 60 000. ■

La determinación del tamaño de la muestra es uno de los primeros pasos de una investigación; no existe una fórmula mágica que diga el tamaño de muestra perfecto para su investigación (de hecho, usted sólo lo sabrá al terminar el estudio). La elección de un tamaño de muestra se parece a decidir cuánta comida llevar para un día de campo.

FIGURA 2.4

La gráfica de $(1.96)s/\sqrt{n}$, contra n, para dos valores posibles de la desviación estándar s.



Usted tiene cierta idea de cuántas personas asistirán, pero no sabrá cuánta comida deberá llevar hasta que el día de campo haya terminado. También debe llevar algo más de comida para los eventos inesperados, como cuando el pequeño Freddie (2 años) dé a los patos un tazón de ensalada de papa o cuando el primo Ted lleve a varios invitados. Pero usted no quiere comprar demasiada comida, pues ésta se puede echar a perder y gastaría dinero en vano. Por supuesto, mientras más días de campo organice y mejor conozca a sus convidados, podrá calcular mejor la cantidad correcta de comida. Es reconfortante saber que lo mismo vale para determinar los tamaños de las muestras: la experiencia y el conocimiento de la población le permitirán diseñar mejor las encuestas.

Los resultados de esta sección pueden servirle de guía para elegir el tamaño de la muestra, pero la decisión final es suya. En general, mientras más grande sea la muestra, menor será el error de muestreo. Sin embargo, recuerde que en la mayor parte de las encuestas también deberá preocuparse por dichos errores y deberá programar recursos para controlar los sesgos de selección y de medición. En muchos casos, los errores que no son de muestreo son mayores al considerar una muestra mayor; con una muestra más grande es fácil introducir otras fuentes de error (por ejemplo, es más difícil controlar la calidad de los entrevistadores o hacer un seguimiento a quienes no respondieron) o relajar la exigencia sobre el sesgo de selección.

2.6 Muestreo sistemático

A veces, el muestreo sistemático se utiliza como sustituto del muestreo aleatorio simple, cuando no se dispone de una lista de la población o cuando esta última tiene un orden más o menos aleatorio. Para obtener una muestra sistemática, se elige una muestra de tamaño n y sea k el siguiente entero después de N/n. Luego, determinamos un número aleatorio R entre 1 y k, el cual determina que la muestra esté formada por las unidades numeradas R, $R+k$, $R+2k$,..., $R+(n-1)k$. Por ejemplo, para elegir una muestra de 45 estudiantes de una lista de 45,000 que estudian en la Universidad Estatal de Arizona, el intervalo de muestreo k es 1000. Suponga que el entero aleatorio elegido es 597. Entonces los estudiantes numerados como 597, 1597, 2597...., 44,597 estarían en la muestra.

Si los nombres de los estudiantes están en orden alfabético, es probable que obtengamos una muestra con un comportamiento similar a una muestra aleatoria simple; es poco probable que la posición alfabética de una persona quede asociada con la característica de interés. Sin embargo, el muestreo sistemático no es igual al muestreo aleatorio simple; no tiene la propiedad de que cada grupo posible de n unidades tenga la misma probabilidad de ser la muestra elegida. En el ejemplo anterior, es imposible que los estudiantes 345 y 346 aparezcan en la muestra. Desde el punto de vista técnico, el muestreo sistemático es una forma del muestreo por conglomerados, como veremos en el capítulo 5.

La mayor parte del tiempo, una muestra sistemática proporciona resultados comparables con los de una muestra aleatoria simple, y los métodos de muestreo aleatorio simple se pueden utilizar en el análisis. Si la población tiene un orden aleatorio, la muestra sistemática será muy similar a una muestra aleatoria simple. Se puede pensar que la población está mezclada. En la cita que se encuentra al principio del capítulo, Sorensen informa que el presidente Kennedy leía una muestra sistemática de las cartas que recibía en la Casa Blanca. Esta muestra sistemática se comportaba como una muestra aleatoria. Observe que Kennedy era consciente de que las cartas que leía, aunque representativas de las cartas que recibía en la Casa Blanca, no eran representativas de la opinión pública.

El muestreo sistemático no proporciona, necesariamente, una muestra representativa si la lista de unidades de la población tiene algún orden periódico o cíclico. Por ejemplo, si los hombres y las mujeres se alternan en la lista y k es par, la muestra sistemática sólo tendrá hombres o sólo mujeres, lo que no puede considerarse una muestra sistemática. En los estudios ecológicos realizados en terrenos agrícolas, puede haber una topografía accidentada que produzca un patrón periódico de vegetación. Si un esquema de muestreo sistemático sigue el mismo ciclo, la muestra no se comportará como una muestra aleatoria simple.

Por otro lado, algunas poblaciones tienen un orden creciente o decreciente. Una lista de las cuentas por cobrar puede estar ordenada de mayor a menor cantidad. En este caso, las estimaciones de la muestra sistemática podrían tener una varianza menor (pero inestimable) en relación con las estimaciones respectivas de la muestra aleatoria simple. Una muestra sistemática de una lista ordenada de cuentas por cobrar debe tener algunas cantidades grandes y otras pequeñas. Es posible que una muestra aleatoria simple sólo contenga cantidades pequeñas o sólo cantidades grandes, de modo que haya más variabilidad entre las medias muestrales de todas las muestras aleatorias simples posibles que entre las medias muestrales de todas las muestras sistemáticas posibles.

En el muestreo sistemático, debemos seguir contando con un marco de muestreo y tener cuidado al definir la población objetivo. El muestreo de uno de cada 20 estudiantes que entran a la biblioteca no nos dará una muestra representativa del conjunto de estudiantes. Sin embargo, el muestreo de una de cada 10 personas que salen de un avión probablemente dará una muestra representativa de las personas de ese vuelo. El marco de muestreo para los pasajeros del avión no aparece en forma explícita, pero existe de cualquier modo.

2.7 Resultados de la teoría de aleatorización para el muestreo aleatorio simple*[1]

En esta sección mostraremos que $\bar{y}$ es un estimador insesgado de $\bar{y}_U$: $\bar{y}_U$ es el promedio de todos los valores posibles de $\bar{y}_S$ si pudiéramos examinar todas las muestras aleatorias simples $\mathcal{S}$ con posibilidad de ser elegidas. También calculamos la varianza de $\bar{y}$ dada en la ecuación (2.7) y mostramos que el estimador de la ecuación (2.9) es insesgado en caso de realizar varias veces un muestreo.

[1] Un asterisco (*) indica una sección, capítulo o ejercicio que requiere más conocimientos matemáticos.

No tenemos supuestos de distribución sobre las y_i, para saber si $\bar{y}$ es insesgado para estimar $\bar{y}_U$. Por ejemplo, no suponemos que las y_i tienen una distribución normal con media μ. En el enfoque del muestreo para la **teoría de aleatorización** (también llamada **basada en el diseño**), las y_i son números fijos pero desconocidos; las probabilidades necesarias surgen de las probabilidades de que las unidades estén en la muestra. El enfoque de la teoría de aleatorización proporciona un enfoque **no paramétrico** de la inferencia: no tenemos que establecer supuestos acerca de la distribución de las variables aleatorias.

Veamos cómo funciona la teoría de aleatorización para la deducción de propiedades de la media muestral en el muestreo aleatorio simple. Como en Cornfield (1994), definimos

$$Z_i = \begin{cases} 1 & \text{si la unidad } i \text{ está en la muestra} \\ 0 & \text{en caso contrario} \end{cases}$$

Entonces

$$\bar{y} = \sum_{i \in S} \frac{y_i}{n} = \sum_{i=1}^{N} Z_i \frac{yi}{n}.$$

Las Z_i son las *únicas* variables aleatorias en la ecuación anterior, pues, de acuerdo con la teoría de aleatorización, las y_i son cantidades fijas. Al extraer una muestra aleatoria simple de n unidades entre las N unidades de la población, $\{Z_1, \ldots, Z_N\}$ son variables aleatorias de tipo Bernoulli, que se distribuyen de manera idéntica con:

$$\pi_i = P(Z_i = 1) = P(\text{elegir la unidad } i \text{ en la muestra}) = \frac{n}{N}. \tag{2.18}$$

La probabilidad en (2.18) es consecuencia de la definición de una muestra aleatoria simple. Para ver esto, observe que si la unidad i está en la muestra, entonces las otras $n-1$ unidades de la muestra deben elegirse entre las otras $N-1$ unidades de la población. De una población de tamaño $N-1$ se puede extraer un total de $\binom{N-1}{n-1}$ muestras posibles de tamaño $n-1$, de modo que

$$P(Z_i = 1) = \frac{\text{número de muestras que incluyen a la unidad } i}{\text{número de muestras posibles}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}.$$

Como consecuencia de la ecuación (2.18),

$$E[Z_i] = E[Z_i^2] = \frac{n}{N}$$

y

$$E[\bar{y}] = E\left[\sum_{i=1}^{N} Z_i \frac{y_i}{n}\right] = \sum_{i=1}^{N} \frac{n}{N} \frac{y_i}{n} = \sum_{i=1}^{N} \frac{y_i}{N} = \bar{y}_U.$$

La varianza de $\bar{y}$ también se calcula al utilizar las propiedades de las variables aleatorias $Z_1, \ldots, Z_N$. Observe que

$$V(Z_i) = E[Z_i^2] - (E[Z_i])^2 = \frac{n}{N} - \left(\frac{n}{N}\right)^2 = \frac{n}{N}\left(1 - \frac{n}{N}\right).$$

Para $i \neq j$,

$$\begin{aligned} E[Z_i Z_j] &= P(Z_i = 1 \text{ y } Z_j = 1) \\ &= P(Z_j = 1 \mid Z_j = 1) P(Z_i = 1) \\ &= \left(\frac{n-1}{N-1}\right)\left(\frac{n}{N}\right). \end{aligned}$$

Como la población es finita, las Z_i no son del todo independientes: si sabemos que la unidad i está en la muestra, tenemos una ligera información acerca del hecho de que la unidad j se encuentre en la muestra, lo que se refleja en la probabilidad condicional $P(Z_j = 1 \mid Z_i = 1)$. En consecuencia, para $i \neq j$,

$$\begin{aligned} \text{Cov}(Z_i, Z_j) &= E[Z_i Z_j] - E[Z_i] E[Z_j] \\ &= \frac{n-1}{N-1}\frac{n}{N} - \left(\frac{n}{N}\right)^2 \\ &= -\frac{1}{N-1}\left(1 - \frac{n}{N}\right)\left(\frac{n}{N}\right). \end{aligned}$$

Utilizamos la covarianza (Cov) de Z_i y Z_j para calcular la varianza de $\bar{y}$; consulte las propiedades de la covarianza del apéndice B. La covarianza negativa de Z_i y Z_j es la fuente de la corrección para poblaciones finitas.

$$\begin{aligned} V(\bar{y}) &= \frac{1}{n^2} V\left(\sum_{i=1}^{N} Z_i y_i\right) \\ &= \frac{1}{n^2} \text{Cov}\left(\sum_{i=1}^{N} Z_i y_i, \sum_{j=1}^{N} Z_j y_j\right) \\ &= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 V(Z_i) + \sum_{i=1}^{N}\sum_{j \neq i}^{N} y_i y_j \text{ Cov}(Z_i, Z_j)\right] \\ &= \frac{1}{n^2}\left[\frac{n}{N}\left(1 - \frac{n}{N}\right)\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N}\sum_{j \neq i}^{N} y_i y_j \frac{1}{N-1}\left(1 - \frac{n}{N}\right)\left(\frac{n}{N}\right)\right] \\ &= \frac{1}{n^2}\frac{n}{N}\left(1 - \frac{n}{N}\right)\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j \neq i}^{N} y_i y_j\right] \\ &= \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N(N-1)}\left[(N-1)\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2 + \sum_{i=1}^{N} y_i^2\right] \\ &= \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N(N-1)}\left[N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right] \\ &= \left(1 - \frac{n}{N}\right)\frac{S^2}{n} \end{aligned}$$

Para mostrar que el estimador en (2.9) es un estimador insesgado de la varianza, debemos mostrar que $E[s^2] = S^2$. El argumento es muy similar al anterior. Como $S^2 = \sum_{i=1}^{N}(y_i - \bar{y}_U)^2/(N-1)$, al tratar de determinar un estimador insesgado tiene sentido precisar el valor esperado de $\sum_{i \in S}(y_i - \bar{y})^2$ y luego establecer la constante que proporcione

el carácter insesgado:

$$
\begin{aligned}
E\left[\sum_{i\in S}(y_i-\bar{y})^2\right] &= E\left[\sum_{i\in S}\left\{(y_i-\bar{y}_U)-(\bar{y}-\bar{y}_U)\right\}^2\right] \\
&= E\left[\sum_{i\in S}(y_i-\bar{y}_U)^2 - n(\bar{y}-\bar{y}_U)^2\right] \\
&= E\left[\sum_{i=1}^{N} Z_i(y_i-\bar{y}_U)^2\right] - nV(\bar{y}) \\
&= \frac{n}{N}\sum_{i=1}^{N}(y_i-\bar{y}_U)^2 - \left(1-\frac{n}{N}\right)S^2 \\
&= \frac{n(N-1)}{N}S^2 - \frac{N-n}{N}S^2 \\
&= (n-1)S^2 .
\end{aligned}
$$

Así,

$$
E\left[\frac{1}{n-1}\sum_{i\in S}(y_i-\bar{y})^2\right] = E[s^2] = S^2 .
$$

2.8
Un modelo para el muestreo aleatorio simple*

A menos que usted haya estudiado la teoría de aleatorización en el diseño de experimentos, es probable que las demostraciones de la sección anterior le parezcan extrañas. Las variables aleatorias en la teoría de aleatorización no se ocupan de las respuestas y_i. Simplemente son variables aleatorias que nos dicen si la i-ésima unidad está en la muestra o no. En un enfoque basado en el diseño (o de teoría de aleatorización) de la inferencia de muestreo, la única relación entre las unidades muestreadas y las unidades no muestreadas es que éstas últimas podrían haber estado en la muestra de haber utilizado un valor inicial distinto para el generador de números aleatorios.

En la sección 2.7 encontramos algunas propiedades de la media muestral $\bar{y}$ por medio de la teoría de aleatorización: $y_1, y_2, \ldots, y_N$ eran considerados como valores fijos, y $\bar{y}$ es insesgado porque el promedio de $\bar{y}_S$, para todas las muestras posibles $\mathcal{S}$ es igual a $\bar{y}_U$. Las únicas probabilidades utilizadas para determinar el valor esperado y la varianza de $\bar{y}$ son las probabilidades de que las unidades estén en la muestra.

Seguramente, en el curso básico que usted tomó de estadística, aprendió un enfoque distinto de la inferencia. En dicho curso, usted disponía de variables aleatorias $\{Y_i\}$ que seguían cierta distribución de probabilidad y los valores reales de la muestra fueron realizaciones de esas variables aleatorias. Así, usted suponía, por ejemplo, que $Y_1, Y_2, \ldots, Y_n$ eran independientes e idénticamente distribuidas, con una distribución normal de media μ y varianza σ^2 y utilizaba las propiedades de las variables aleatorias independientes y la distribución normal para determinar los valores esperados de diversos estadísticos.

Podemos extender este enfoque del muestreo al pensar que las variables aleatorias $Y_1, Y_2, \ldots, Y_N$ son generadas a partir de cierto modelo. Los valores reales para la población finita

$y_1, y_2, \ldots, y_N$ son una realización de las variables aleatorias. La distribución de probabilidad conjunta de $Y_1, Y_2, \ldots, Y_N$ proporciona este vínculo entre las unidades que están en la muestra y las que no lo están, en este enfoque **basado en el modelo**, este vínculo es el que falta en el enfoque de aleatorización. En este caso, muestreamos $\{y_i, i \in \mathcal{S}\}$ y utilizamos estos datos para predecir los valores no observados $\{y_i, i \notin \mathcal{S}\}$. Así, los problemas en el muestreo de poblaciones finitas pueden pensarse como problemas de predicción.

En una muestra aleatoria simple, se puede adoptar un modelo sencillo como el siguiente:

$$
Y_1, Y_2, \ldots, Y_N \text{ independientes con } E_M[Y_i] = \mu \text{ y } V_M[Y_i] = \sigma^2 . \qquad \textbf{(2.19)}
$$

El subíndice M indica que la esperanza utiliza la distribución del modelo y no la correspondiente a la aleatorización, utilizada en la sección 2.7. En este caso, μ y σ^2 representan parámetros desconocidos en una población infinita, no las cantidades correspondientes en una población finita, como en la sección 2.7. Extraemos una muestra $\mathcal{S}$ y observamos los valores y_i, para $i \in \mathcal{S}$; es decir, vemos realizaciones de las variables aleatorias Y_i para $i \in \mathcal{S}$. Las otras observaciones en la población $\{y_i, i \notin \mathcal{S}\}$ también son realizaciones de variables aleatorias, pero no vemos éstas. El total de la población finita t se puede escribir como

$$
t = \sum_{i=1}^{N} y_i = \sum_{i\in S} y_i + \sum_{i\notin S} y_i
$$

y es un valor posible de la variable aleatoria

$$
T = \sum_{i=1}^{N} Y_i = \sum_{i\in S} Y_i + \sum_{i\notin S} Y_i .
$$

Conocemos los valores $\{y_i, i \in \mathcal{S}\}$. Para estimar t a partir de la muestra, debemos determinar las estimaciones de los valores y que no están en la muestra. Aquí es donde entra nuestro modelo de la media común μ. El estimador por mínimos cuadrados de μ a partir de la muestra es $\bar{Y}_S = \sum_{i\in S} Y_i / n$ y éste es el mejor predictor lineal insesgado (bajo este modelo) de los valores no observados, de modo que

$$
\hat{T} = \frac{N}{n}\sum_{i\in S} Y_i .
$$

El estimador $\hat{T}$ es *insesgado para el modelo*: si el modelo es cierto, entonces el promedio de $\hat{T} - T$ bajo varias realizaciones de la población es

$$
E_M[\hat{T}-T] = \frac{N}{n}\sum_{i\in S} E_M[Y_i] - \sum_{i=1}^{N} E_M[Y_i] = 0 .
$$

(Observe la diferencia entre determinar las esperanzas bajo el enfoque basado en el modelo y bajo el enfoque basado en el diseño. En el primero, las Y_i son las variables aleatorias y la muestra no presenta información para calcular los valores esperados. En el segundo, las variables aleatorias están contenidas en la muestra $\mathcal{S}$.)

El error cuadrático medio también se calcula como el cuadrado de la desviación promedio entre la estimación y el total de la población finita. Para cualquier realización dada de las variables aleatorias, el error al cuadrado es:

$$
\left[\frac{N}{n}\sum_{i\in S} y_i - \sum_{i=1}^{N} y_i\right]^2 .
$$

Al promediar esta cantidad sobre todas las posibles realizaciones de las variables aleatorias obtenemos el error cuadrático medio bajo los supuestos del modelo:

$$E_M[(\hat{T}-T)^2] = E_M\left[\left(\frac{N}{n}\sum_{i\in S} Y_i - \sum_{i=1}^{N} Y_i\right)^2\right]$$

$$= E_M\left[\left\{\left(\frac{N}{n}-1\right)\sum_{i\in S} Y_i - \sum_{i\notin S} Y_i\right\}^2\right]$$

$$= E_M\left[\left(\frac{N}{n}-1\right)^2\left(\sum_{i\in S} Y_i - nm\right)^2 + \left(\sum_{i\notin S} Y_i - (N-n)m\right)^2\right]$$

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

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

En la práctica, si adopta el modelo de la ecuación (2.19), usted tendrá que estimar σ^2 mediante la varianza de la muestra s^2. Así, los enfoques basados en el diseño o en el modelo [donde el modelo está dado en (2.19)] conducen a la misma estimación de la población total y la misma estimación de la varianza. Sin embargo, si se adopta otro modelo, las estimaciones podrían diferir. En los capítulos 3 y 11 veremos cómo estos dos enfoques pueden conducir a distintas inferencias.

El enfoque basado en el diseño y el enfoque basado en el modelo, donde el modelo está dado en (2.19), también conducen al mismo intervalo de confianza para la media. Sin embargo, estos intervalos de confianza tienen distintas interpretaciones. El intervalo de confianza para $\bar{y}_U$ basado en el diseño se puede interpretar como sigue: si consideramos todas las muestras aleatorias simples posibles de tamaño n a partir de la población finita de tamaño N y construimos un intervalo de confianza al 95% para cada muestra, el 95% de todos los intervalos de confianza construidos de esta forma incluirán al verdadero valor $\bar{y}_U$. Así, el intervalo de confianza basado en el diseño tiene una interpretación de *muestreo repetido*.

El intervalo de confianza basado en el modelo para el parámetro μ se interpreta en términos del modelo (2.19). El procedimiento del intervalo de confianza produce dos variables aleatorias: $LL = \bar{Y}_S - 1.96S/\sqrt{n}$ y $UL = \bar{Y}_S + 1.96S/\sqrt{n}$. Entonces, al utilizar el modelo para inferir que $\bar{Y}_S$ tiene aproximadamente una distribución normal con media μ y varianza S^2/n,

$$P(LL \le \mu \le UL) = 0.95 .$$

Por lo general, este intervalo de confianza basado en el modelo se interpreta en los cursos introductorios de estadística al usar muestras repetidas: si generamos valores para la población una y otra vez al emplear el modelo en (2.19) y construimos un intervalo de confianza para cada muestra resultante, esperamos que el 95% de los intervalos de confianza contengan el verdadero valor de μ. Aunque los dos tipos de intervalos de confianza se pueden interpretar mediante muestras repetidas, existe una diferencia entre ellos. El nivel de confianza basado en el diseño da la proporción esperada del intervalo de confianza que contendrán a $\bar{y}_U$, a partir del conjunto de todos los intervalos de confianza que se pueden construir extrayendo una muestra aleatoria simple de tamaño n de la población finita de valores fijos $\{y_1, y_2, \ldots, y_N\}$. El nivel de confianza basado en el modelo da la proporción esperada de intervalos de confianza que incluirán a μ, a partir del conjunto de todas las muestras que se pueden generar mediante el modelo en (2.19).

Una nota acerca de la notación Algunos libros (por ejemplo, Cochran 1977) y artículos utilizan Y para representar a la población total (t en este libro) y $\bar{Y}$ para representar la media de la población (nuestro $\bar{y}_U$). En este libro reservaremos Y y T para representar variables aleatorias en un enfoque basado en el modelo. Nuestro uso es consistente con otras áreas de la estadística, en donde las letras mayúsculas del final del alfabeto representan, por lo general, variables aleatorias. Sin embargo, usted debe tomar en cuenta que la notación en la bibliografía sobre el muestreo de encuestas no es uniforme.

2.9 ¿Cuándo se debe utilizar una muestra aleatoria simple?

El muestreo aleatorio simple sin reemplazo es el método de muestreo de probabilidad más sencillo; todas las estimaciones se calculan de manera muy similar a como usted aprendió en su curso introductorio de estadística. Las estimaciones son:

Cantidad relativa a la población	Estimación	Error estándar de la estimación
Media de la población, $\bar{y}_U$	$y = \frac{1}{n}\sum_{i\in S} y_i$	$\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}$
Proporción de la población, p	$\hat{p}$	$\sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}}$
Población total, t	$\hat{t} = N\bar{y}$	$N\,\mathrm{EE}(\bar{y})$

La única característica que aparece en las estimaciones para las muestras aleatorias sin reemplazo y que no aparece en las muestras con reemplazo es la corrección para las poblaciones finitas, $(1-n/N)$, la cual disminuye el error estándar cuando el tamaño de la muestra es grande con respecto al tamaño de la población. En la mayor parte de las encuestas realizadas en la práctica, la corrección para las poblaciones finitas es tan cercana a 1 que puede ser ignorada.

Para tamaños de muestra lo "suficientemente grandes", un intervalo de confianza aproximado al 95% está dado por

estimación ± 1.96 EE(estimación).

El margen de error de una estimación es la mitad del ancho del intervalo de confianza; es decir, 1.96 × EE (estimación).

Por lo general, las muestras aleatorias simples son fáciles de diseñar y analizar. Pero no constituyen el mejor diseño que se pueda utilizar en las siguientes situaciones:

- Antes de extraer una muestra aleatoria simple, piense si una muestra con encuestas es el mejor método para analizar su tema de investigación. Si usted quiere determinar la eficacia de cierto aceite como repelente para mosquitos, deberá realizar un experimento controlado, no una encuesta. Usted debe realizar una encuesta en caso de que desee

saber si utilizan ese aceite como repelente o si quiere estimar la cantidad de mosquitos que se encuentran en el área.

- Tal vez usted no disponga de una lista de las unidades de observación o, tal vez, resulte caro invertir horas en un viaje para extraer una muestra aleatoria simple. Si le interesa la proporción de mosquitos portadores del virus de la encefalitis en el suroeste de Wisconsin, no puede construir un marco de muestreo de los mosquitos individuales. Tendrá que obtener muestras en áreas distintas y luego examinar algunos o todos los mosquitos de esas áreas, al utilizar una técnica de muestreo conocida como muestreo por conglomerados (el muestreo por conglomerados será analizado en los capítulos 5 y 6).
- Usted podría tener información adicional que permita diseñar un esquema de muestreo más eficaz en cuanto a costos. En un estudio para estimar la cantidad total de mosquitos en cierta área, un entomólogo desearía saber cuál es la probabilidad de que cierto tipo de terreno tenga la mayor o la menor densidad de mosquitos, antes de extraer cualquier muestra. Podría ahorrar mucho esfuerzo de muestreo al dividir el área en estratos, grupos de unidades similares y, luego, obtener muestras en cada estrato (el muestreo estratificado será analizado en el capítulo 4).

Usted debe utilizar una muestra aleatoria simple en estas situaciones:

- Las personas que analizan datos insisten en el uso de fórmulas de muestreo aleatorio simple, sean adecuadas o no. Algunas personas no desisten de la creencia de que sólo se debe estimar la media al considerar el promedio de la muestra; en ese caso, diseñe una muestra donde lo correcto sea promediar los valores muestrales. Las muestras aleatorias simples se recomiendan, con frecuencia, cuando la evidencia muestral se usa en acciones legales; a veces, al emplear un esquema de muestreo más complejo, el abogado oponente intentará convencer al jurado de que los resultados muestrales no son válidos.
- Se dispone de poca información adicional útil que permita diseñar la encuesta. Si su marco de muestreo es sólo una lista de nombres de estudiantes universitarios en orden alfabético y usted no tiene más información como el área de interés o el semestre que se cursa, el muestreo aleatorio simple o el sistemático son probablemente las mejores estrategias de muestreo de probabilidad.
- El interés principal se encuentra en las relaciones multivariadas, como ecuaciones de regresión que sirvan para toda la población, y no existen razones de peso para realizar una muestra estratificada o por conglomerados. Los análisis multivariados pueden realizarse sobre muestras complejas, pero son más fáciles de realizar e interpretar en una muestra aleatoria simple.

2.10 Ejercicios

1 Sean $N = 6$ y $n = 3$. Para estudiar las distribuciones del muestreo, suponemos que son conocidos todos los valores de la población.

$$y_1 = 98 \quad y_3 = 154 \quad y_5 = 190$$
$$y_2 = 102 \quad y_4 = 133 \quad y_6 = 175$$

Nos interesa $\bar{y}_U$, la media de la población. Se proponen dos planes de muestreo.

- Plan 1: se puede elegir entre ocho muestras posibles.

Número de muestra	Muestra, $\mathcal{S}$	$P(\mathcal{S})$
1	{1, 3, 5}	$\frac{1}{8}$
2	{1, 3, 6}	$\frac{1}{8}$
3	{1, 4, 5}	$\frac{1}{8}$
4	{2, 4, 6}	$\frac{1}{8}$
5	{2, 3, 5}	$\frac{1}{8}$
6	{2, 3, 6}	$\frac{1}{8}$
7	{2, 4, 5}	$\frac{1}{8}$
8	{2, 4, 6}	$\frac{1}{8}$

- Plan 2: se puede elegir entre tres muestras posibles.

Número de muestra	Muestra, $\mathcal{S}$	$P(\mathcal{S})$
1	{1, 4, 6}	$\frac{1}{4}$
2	{2, 3, 6}	$\frac{1}{2}$
3	{1, 3, 5}	$\frac{1}{4}$

a ¿Cuál es el valor de $\bar{y}_U$?

b Sea $\bar{y}$ la media de los valores de la muestra. Para cada plan de muestreo, determine:

- **i** $E[\bar{y}]$
- **ii** $V[\bar{y}]$
- **iii** Sesgo $(\bar{y})$
- **iv** ECM$(\bar{y})$

c ¿Cuál plan de muestreo piensa que sea el mejor? ¿Por qué?

2 Para la población del ejemplo 2.1, considere el siguiente esquema de muestreo:

$\mathcal{S}$	$P(\mathcal{S})$
{1, 3, 5, 6}	$\frac{1}{8}$
{2, 3, 7, 8}	$\frac{1}{4}$
{1, 4, 6, 8}	$\frac{1}{8}$
{2, 4, 6, 8}	$\frac{3}{8}$
{4, 5, 7, 8}	$\frac{1}{8}$

a Determine la probabilidad de selección π_i para cada unidad i.

b ¿Cuál es la distribución muestral de $\hat{t} = 8\bar{y}$?

3 Para la población del ejemplo 2.1, determine la distribución muestral de $\bar{y}$ para:

a Una muestra aleatoria simple de tamaño 3 sin reemplazo.

b Una muestra aleatoria simple de tamaño 3 con reemplazo.

Para cada caso, trace el histograma de la distribución muestral de $\bar{y}$. ¿Cuál distribución tiene menor varianza? ¿Por qué?

4 Una forma de elegir una muestra aleatoria simple es asignar un número a cada unidad de la población y, luego, usar una tabla de números aleatorios para elegir las unidades mediante la lista. En el apéndice E damos una página de una tabla de números aleatorios. Explique el por qué cada uno de los siguientes métodos produce o no una muestra aleatoria simple.

a La población tiene 742 unidades y queremos extraer una muestra aleatoria simple de tamaño 30. Divida la lista aleatoria en segmentos de tamaño 3 y elimine todas las series de tres dígitos que *no estén* entre 001 y 742. Si aparece un número ya incluido, ignórelo. Si utilizamos este método con el primer renglón de números aleatorios del apéndice E, la serie de números de tres dígitos sería la siguiente:

749 700 699 611 136 ...

Incluiríamos las unidades 700, 699, 611 y 136 en la muestra.

b Para la situación que se presenta en la parte (a), cuando un número aleatorio de tres dígitos sea mayor que 742, sólo eliminamos el primer dígito y comenzamos la secuencia con el siguiente dígito. Con este procedimiento, los primeros cinco números serían 497, 006, 611, 136 y 264.

c La población tiene 170 elementos. Al usar los procedimientos de la parte (a) o (b), eliminaríamos muchos números de la lista. Para evitar este desperdicio, divida cada número aleatorio de tres dígitos entre 170 y utilice el residuo redondeado como unidad en la muestra. Si el residuo es 0, utilice la unidad 170. Como en las partes (a) y (b), elimine los duplicados. Para la secuencia del primer renglón en la tabla de números aleatorios, los números generados serían los siguientes:

69 20 19 101 136 ...

d La población tiene 200 elementos. Considere las secuencias de números aleatorios con dos dígitos y coloque un punto decimal al frente para obtener la siguiente secuencia:

.74 .97 .00 .69 .96 ...

Luego multiplique cada decimal por 200 para obtener las unidades de la muestra (convierta .00 en 200):

148 194 200 138 192 ...

e Una escuela tiene 20 grupos; cada grupo tiene entre 20 y 40 estudiantes. Para elegir un estudiante para la muestra, considere un número aleatorio entre 1 y 20 y, luego, escoja un estudiante al azar que pertenezca al grupo seleccionado. No incluya los duplicados en la muestra.

f Para la situación descrita en la parte (e), elija un número aleatorio entre 1 y 20 para seleccionar el grupo. Luego, escoja un segundo número aleatorio entre 1 y 40. Si el número corresponde a un estudiante del grupo, entonces, elija ese estudiante. Si el segundo número aleatorio es mayor que el tamaño del grupo, entonces ignore esta pareja de números aleatorios y comience de nuevo. Como de costumbre, elimine los duplicados de su lista.

5 Mayr *et al.* (1994) extrajeron una muestra aleatoria simple de 240 niños con edades de entre 2 y 6 años, quienes visitaron la clínica pediátrica donde los primeros trabajan. Ellos encontraron la siguiente distribución de frecuencias de los niños que caminan solos (sin ayuda):

Edad (meses)	9	10	11	12	13	14	15	16	17	18	19	20
Número de niños	13	35	44	69	36	24	7	3	2	5	1	1

a Construya un histograma de la distribución de la edad al comenzar a caminar. ¿Se presenta una distribución normal? ¿Cree que la distribución muestral del promedio de la muestra se distribuya normalmente? ¿Por qué sí?, o ¿por qué no?

b Determine la media, el error estándar y un intervalo de confianza al 95% para la edad promedio en que se comienza a caminar solo.

c Suponga que los investigadores desean realizar otro estudio en una región diferente y quieran que el intervalo de confianza al 95% para la edad promedio en que se comienza a caminar solo tenga un margen de error de 0.5. Al utilizar la desviación estándar estimada para estos datos, ¿qué tamaño de muestra necesitan?

6 Una cantidad, con frecuencia, de interés para una clínica es el porcentaje de pacientes retrasados para su vacunación. Algunas clínicas examinan cada registro para determinar el porcentaje; sin embargo, en una clínica grande, la realización de un censo de los registros puede llevar mucho tiempo. Cullen (1994) realizó una muestra de los 580 niños a los que da servicio una clínica familiar en Auckland para estimar la proporción de interés.

a ¿Qué tamaño de muestra sería necesario en una muestra aleatoria simple (sin reemplazo) para estimar la proporción con el 95% de confianza y un margen de error del 0.10?

b En la realidad, Cullen realizó una muestra aleatoria simple con reemplazo de tamaño 120, de los cuales 27 resultaron como *no* retrasados para la vacuna. Dé un intervalo de confianza al 95% para la proporción de niños no retrasados.

***7** (Requiere la aplicación de conocimientos probabilísticos.) En la población que se empleó en el ejemplo 2.4, 19 de los 3078 condados de la población no tenían el valor de *acres92*. ¿Cuál es la probabilidad de que en una muestra aleatoria simple de tamaño 300, no falten datos para esa variable?

8 Una universidad tenía 807 profesores e investigadores en el área de ciencias y humanidades en 1993; la lista del personal académico y sus publicaciones en 1992-1993 estaba disponible en el sistema de cómputo. Para cada miembro del personal académico se registraba el número de publicaciones con arbitraje. Este número no está disponible de manera directa en la base datos, de modo que el investigador debe examinar cada registro por separado. Aquí damos una tabla de frecuencias con el número de publicaciones con arbitraje para una muestra aleatoria simple de 50 académicos.

Publicaciones con arbitraje	0	1	2	3	4	5	6	7	8	9	10
Personal académico	28	4	3	4	4	2	1	0	2	1	1

a Grafique los datos mediante un histograma. Describa la forma de los datos.

b Estime el número medio de publicaciones por cada miembro del personal y dé un error estándar para su estimación.

c ¿Cree que $\bar{y}$ de la parte (b) tendrá aproximadamente una distribución normal? ¿Por qué sí?, o ¿por qué no?

d Estime la proporción de académicos sin publicaciones y dé un intervalo de confianza al 95% para su estimación.

9 Defina un procedimiento para el intervalo de confianza como

$$\text{IC}(\mathcal{S}) = [\hat{t}_S - 1.96\ \text{EE}(\hat{t}_S),\ \hat{t}_S + 1.96\ \text{EE}(\hat{t}_S)].$$

Utilice el método ilustrado en el ejemplo 2.7 y determine el nivel exacto para un intervalo de confianza basado en una muestra aleatoria simple sin reemplazo, de tamaño 4, a partir de la población del ejemplo 2.1. ¿Es igual a 95% el nivel de confianza?

10 Una carta publicada en el número de diciembre de 1995 del *Dell Champion Variety Puzzles* indicaba lo siguiente: "He observado en los últimos números que no hay ganadores del sur en los concursos. Ustedes siempre dicen que los ganadores se eligen al azar. ¿Significa esto que ustedes venden menos en el sur?" En respuesta, los editores realizaron una muestra aleatoria de 1000 datos de los últimos concursos y encontraron que 175 provenían del sur.

a Determine un intervalo de confianza al 95% para el porcentaje de datos provenientes del sur.

b De acuerdo con el resumen estadístico de Estados Unidos, el 30.9% de la población de la Unión Americana vive en estados considerados de la parte sur por los editores. ¿Hay alguna evidencia en el intervalo de confianza de que el porcentaje de datos de la parte sur de Estados Unidos difiere del porcentaje de personas que viven en esa región del país?

11 Para cada una de las siguientes cantidades, grafique los datos y estime la media de la población para esa variable, junto con el error estándar. Dé un intervalo de confianza al 95% para su estimación.

a Número de acres dedicados a la agricultura en 1987.

b Número de granjas, 1992.

c Número de granjas con 1000 acres o más, 1992.

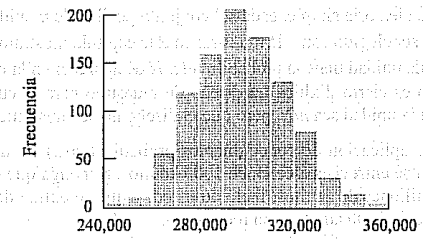
d Número de granjas con 9 acres o menos, 1992.

12 El censo especial del condado de Maricopa, Arizona, en 1995, proporcionó las siguientes poblaciones de las ciudades correspondientes:

Ciudad	Población
Buckeye	4,857
Gilbert	59,338
Gila Bend	1,724
Phoenix	1,149,417
Tempe	153,821

Suponga que quiere estimar el porcentaje de personas que han sido inmunizadas contra la polio en cada ciudad y quiere extraer una muestra aleatoria simple de los habitantes de cada ciudad. ¿Cuál debe ser el tamaño de la muestra en cada una de las cinco ciudades si desea que la estimación de cada ciudad tenga un margen de error de 4 puntos porcentuales? ¿Para cuáles ciudades es significativa la corrección para poblaciones finitas?

FIGURA 2.5
Histograma de las medias de 1000 muestras de tamaño 300, extraídas con reemplazo a partir de los datos del ejemplo 2.4.



Distribución muestral estimada de $\bar{y}$

***13** *Enfoque de la teoría de decisiones para estimar el tamaño de la muestra.* (Requiere del empleo del cálculo.) En un enfoque de la teoría de decisiones, se especifican dos funciones:

$L(n)$, pérdida o "costo" de una mala estimación

$C(n)$, costo de extracción de la muestra

Suponga que existen constantes c_0, c_1 y k tales que

$$L(n) = kV(\bar{y}_S) = k\left(1 - \frac{n}{N}\right)\frac{S^2}{n}$$

$$C(n) = c_0 + c_1 n.$$

¿Qué tamaño de muestra n minimiza el costo total $L(n) + C(n)$?

14 (Se requiere de una computadora para resolver este ejercicio). Si usted tiene una muestra aleatoria simple de gran tamaño, puede estimar la distribución muestral de $\bar{y}_S$ al extraer varias muestras de tamaño n con reemplazo a partir de la lista de valores de la muestra. La figura 2.5 muestra un histograma de las medias de 1000 muestras de tamaño 300 con reemplazo a partir de los datos del ejemplo 2.4; la forma podrá ser un poco asimétrica pero parece aproximadamente normal. ¿Sería lo suficientemente grande una muestra de tamaño 100 de esta población como para aplicar el teorema del límite central? Extraiga 500 muestras de tamaño 100, con reemplazo, de la variable *acre92* en agsrs.dat y trace un histograma de las 500 medias. El enfoque descrito en este ejercicio se conoce como la *agujeta* (véase Efron y Tibshirani 1993); analizaremos este enfoque con detalle en la sección 9.3.

15 El sitio en Internet www.golfcourse.com enumera 14,938 campos de golf por estado. Proporciona información variada relativa a cada campo, como la cuota de ingreso, la clasificación del campo, los pares para cada campo y las instalaciones. Los datos de una muestra aleatoria simple de 120 de los campos aparecen en el archivo golfsrs.dat en el disco de datos.

a Despliegue los datos en un histograma, de tal manera que correspondan con las cuotas de uso semanal de 9 hoyos. ¿Cómo describiría la forma de los datos?

b Determine las cuotas semanales promedio para jugar 9 hoyos y dé el error estándar para la estimación.

16 Repita el ejercicio 15 para el yardaje *back-tee*.

17 Para los datos en golfsrs.dat, estime la proporción de campos de golf que tienen 18 hoyos y dé un intervalo de confianza al 95% para la estimación.

18 En una muestra aleatoria simple, cada subconjunto posible de n unidades tiene la probabilidad $1/\binom{N}{n}$ de ser elegido como la muestra; en este capítulo mostramos que esta definición implica que cada unidad tiene la probabilidad n/N de aparecer en la muestra. Sin embargo, el recíproco no es cierto. Exhiba un diseño de muestreo para el cual la probabilidad de selección de cada unidad sea n/N, pero que el diseño no sea una muestra aleatoria simple.

***19** (Requiere de la aplicación de conocimientos probabilísticos.) En una encuesta típica de opinión pública se entrevista a cerca de 1000 adultos. Suponga que el marco de muestreo contiene 100 millones de adultos, incluido usted, y que se extrae del marco una muestra aleatoria simple de 1000 adultos, sin reemplazo.

a ¿Cuál es la probabilidad de que usted sea elegido para estar en la muestra?

b Ahora suponga que se eligen 2000 de tales muestras, cada una elegida de manera independiente a las demás. ¿Cuál es la probabilidad de que usted *no* esté en ninguna de las muestras?

c ¿Cuántas muestras deben elegirse para tener 0.5 de probabilidad de que usted esté al menos en una muestra?

***20** (Requiere de la aplicación de conocimientos probabilísticos.) En una muestra aleatoria simple con reemplazo, una unidad de población puede aparecer en la muestra de 0 a n veces.

Sea Q_i el número de veces que aparece la unidad i en la muestra

y

$$\hat{t} = \frac{N}{n}\sum_{i=1}^{N} Q_i y_i.$$

a Argumente que la distribución conjunta de $Q_1, Q_2, \ldots, Q_N$ es multinomial con n ensayos y $p_1 = p_2 = \cdots = p_N = 1/N$.

b Utilice la parte (a) para mostrar que $E[\hat{t}] = t$.

c Utilice la parte (a) para determinar $V[\hat{t}]$.

***21** (Requiere de la aplicación de conocimientos probabilísticos.) Suponga que quiere extraer una muestra aleatoria simple de tamaño n de una lista de N unidades, pero que no conoce de antemano N, el tamaño de la población. Considere el siguiente procedimiento:

a Haga $\mathcal{S}_0 = \{1,2,\ldots,n\}$, de modo que la muestra inicial en cuestión conste de las primeras n unidades de la lista.

b Para $k = 1,2,\ldots$, genere un número aleatorio u_k entre 0 y 1. Si $u_k > n/(n+k)$, entonces haga $\mathcal{S}_k$ igual a $\mathcal{S}_{k-1}$. Si $u_k \le n/(n+k)$, entonces elija una de las unidades en $\mathcal{S}_{k-1}$ al azar y reemplácela por la unidad $(n+k)$ para formar $\mathcal{S}_k$.

Muestre que $\mathcal{S}_{N-n}$, obtenida a través de este procedimiento, es una muestra aleatoria simple de tamaño n de la población.

22 Extraiga una pequeña muestra aleatoria simple de algo en lo que esté interesado. Explique aquello que quiere estudiar y describa con cuidado la forma de elegir su muestra aleatoria (dé los números aleatorios generados y explique cómo traducirlos en observaciones), escriba un informe con sus datos y proporcione una estimación puntual y el error estándar para la cantidad o cantidades de interés.

La recolección de los datos de este ejercicio no debe representar mayor esfuerzo, pues usted está rodeado de cosas que esperan ser muestreadas. Algunos ejemplos: datos de fondos mutualistas en la sección financiera del periódico de hoy, pesos reales de bolsas de un kilogramo de zanahorias en el supermercado, costo de un artículo en varias tiendas y el tiempo transcurrido hasta que su módem se conecta al sistema telefónico.

23 ¿Qué tan confiable es la información que está en Internet? Elija un tema controvertido, del cual tenga cierta información. Utilice una herramienta de búsqueda para generar un marco de muestreo de contribuciones al tema. Si está familiarizado con los tratamientos médicos para curar el asma, por ejemplo, podría realizar una búsqueda acerca del "tratamiento contra el asma". Ahora elija una muestra aleatoria simple de estas contribuciones, al utilizar los números asignados por la máquina de búsqueda a las contribuciones, para conformar su muestra. Estime la proporción de contribuciones de la lista que proporcionan información incorrecta y dé un intervalo de confianza al 95% para su proporción.

3

Estimación por razones y por regresión

Los registros de nacimientos, mantenidos en orden para garantizar la condición de los ciudadanos, pueden servir para determinar la población de un gran imperio sin recurrir a un censo de sus habitantes, operación laboriosa y difícil de realizar con exactitud. Pero para esto es necesario conocer la razón entre la población y los nacimientos anuales. El medio más preciso para esto consiste, primero, en elegir subdivisiones del imperio distribuidas de manera casi igual en toda su superficie, para obtener el resultado general independiente de las circunstancias locales; segundo, enumerar con cuidado a los habitantes de varias comunas en cada una de las subdivisiones, durante un tiempo determinado; tercero, determinar el número promedio de nacimientos anuales correspondiente al utilizar la cuenta de nacimientos durante varios años antes y después de este tiempo. Este número, dividido entre el número de habitantes, dará la razón entre los nacimientos anuales y la población, de manera cada vez más confiable conforme aumente la enumeración.

—Pierre-Simon Laplace, *Essai Philosophique sur les Probabilités* (traducción de S. Lohr)

Francia no tenía censos de población en 1802 y Laplace quería estimar el número de personas que ahí vivían (Cochran 1978; Laplace 1814). Laplace obtuvo una muestra de 30 comunas diseminadas por todo el país. Estas comunas tenían un total de 2,037,615 habitantes el 23 de septiembre de 1802. Durante los tres años posteriores a esta fecha se registró un total de 215,599 nacimientos en las 30 comunas. Laplace determinó el número anual de nacimientos registrados en las 30 comunas como 215,599/3 = 71,866.33. Al dividir 2,037,615 entre 71,866.33, Laplace estimó que cada año había un nacimiento registrado por cada 28.352845 personas. Con el razonamiento de que es probable que las comunas con gran población también tuvieran un gran número de nacimientos registrados, al juzgar que la razón entre la población y los nacimientos anuales de la muestra que obtuvo serían similares a lo que ocurre en toda Francia, concluyó que uno podría estimar la población total de Francia al multiplicar la cantidad total de nacimientos anuales en toda Francia por 28.352845 (por alguna razón, Laplace decidió no utilizar en sus cálculos el número real de nacimientos registrados en Francia en el año anterior al 22 de septiembre de 1802, sino que multiplicó la razón por un millón).

Laplace no estaba interesado en la cantidad total de nacimientos registrados en sí, sino que utilizó ésta como información auxiliar para estimar la población total de Francia. Con frecuencia tenemos cierta información auxiliar en las encuestas; pocos investigadores hacen

el gasto de realizar una buena muestra y luego medir sólo una cantidad. Con frecuencia, el marco de muestreo nos proporciona información adicional acerca de cada unidad que puede servir para mejorar la precisión de nuestras estimaciones. Las estimaciones por razones y por regresión utilizan las variables correlacionadas con la variable de interés para mejorar la precisión de las estimaciones de la media y el total de una población.

3.1 Estimación por razones

Para aplicar la estimación por razones, debemos medir dos cantidades y_i y x_i en cada unidad de la muestra; con frecuencia, x_i es una **variable auxiliar** o **subsidiaria**. En la población de tamaño N,

$$t_y = \sum_{i=1}^{N} y_i \quad t_x = \sum_{i=1}^{N} x_i$$

y su razón[1] es

$$B = \frac{t_y}{t_x} = \frac{\bar{y}_U}{\bar{x}_U}.$$

En la forma más sencilla de estimación por razones, se extrae una muestra aleatoria simple de tamaño n y se utiliza la información de x y y para estimar B, t_y o $\bar{y}_U$.

Las estimaciones por razones y por regresión aprovechan la correlación de x y y en la población; mientras mayor es la correlación, mejor funcionan. Definimos el **coeficiente de correlación de la población** como:

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{(N-1)S_x S_y}. \tag{3.1}$$

En este caso, S_x es la desviación estándar de las x_i de la población, S_y es la desviación estándar de las y_i de la población y R no es más que el coeficiente de correlación de Pearson de x y y para las N unidades de la población.

EJEMPLO 3.1 Suponga que la población consta de campos agrícolas de distintos tamaños. Sean

$y_i =$ los *bushels* de granos cosechados en el campo i

$x_i =$ los acres de terreno

Entonces

$B =$ es la producción promedio de bushels por acre

$\bar{y}_U =$ es la producción promedio de bushels por campo

$t_y =$ es la producción total en bushels. ■

[1] ¿Por qué utilizar la letra B para representar la razón? Como veremos en la sección 3.4, la estimación por razones es motivada por un modelo de regresión: $Y_i = \beta x_i + \varepsilon_i$, con $E[\varepsilon_i] = 0$ y $V[\varepsilon_i] = \sigma^2 x_i$. Así, la razón entre t_y y t_x es, en realidad, un coeficiente de regresión.

Si se extrae una muestra aleatoria simple, los estimadores naturales para B, t_y y $\bar{y}_U$ son

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{\hat{t}_y}{\hat{t}_x}$$

$$\hat{t}_{yr} = \hat{B}t_x \tag{3.2}$$

$$\hat{y}_r = \hat{B}\bar{x}_U,$$

donde t_x y $\bar{x}_U$ se suponen que son conocidos.

3.1.1 ¿Por qué utilizar la estimación por razones?

1 A veces, simplemente, queremos estimar una razón. En el ejemplo 3.1, B (la producción promedio por acre) es de interés y se estima mediante la razón entre las medias muestrales $\hat{B} = \bar{y}/\bar{x}$. Si los campos tienen diferente tamaño, el numerador y el denominador son cantidades aleatorias; si se elige otra muestra, es probable que $\bar{y}$ y $\bar{x}$ cambien. En otras situaciones en las que se utilizan encuestas, las razones de interés podrían ser la razón entre pasivos y activos, la razón entre la cantidad de peces atrapados y el número de horas ocupadas en pescar o el ingreso per cápita de los miembros de las familias en Australia.

Algunas estimaciones por razones están ocultas, pues el denominador parece como si fuera un tamaño de muestra común. Para determinar si usted debe usar la estimación por razones para una cantidad, pregúntese lo siguiente: "si considero una muestra distinta, ¿cambiará el denominador?". En caso afirmativo, entonces usted está utilizando la estimación por razones. Suponga que está interesado en el porcentaje de páginas de la revista *Good Housekeeping* que contienen al menos un anuncio. Usted podría extraer una muestra aleatoria simple de diez números de la revista y medir lo siguiente:

$x_i =$ es la cantidad total de páginas en el número i

$y_i =$ es la cantidad total de páginas en el número i

que contienen al menos un anuncio.

La razón de interés se puede estimar como

$$\hat{B} = \frac{\sum_{i \in S} y_i}{\sum_{i \in S} x_i}.$$

El denominador es el número total de páginas en los diez números y es probable que cambie si se toma una muestra distinta de números de la revista.

Desde el punto de vista técnico, utilizamos la estimación por razones cada vez que extraigamos una muestra aleatoria simple y estimemos una media o razón para una subpoblación, como analizaremos en la sección 3.3.

2 A veces, queremos estimar el total de una población, pero desconocemos el tamaño N de la población. Entonces no podemos utilizar el estimador $\hat{t}_y = N\bar{y}$ del capítulo 2. Pero sabemos que $N = t_x/\bar{x}_U$ y podemos estimar N por medio de $t_x/\bar{x}$. Así, utilizamos otra medida del tamaño, t_x, en vez de la población N.

Para estimar la cantidad total de peces, en una redada, que son mayores a 12 cm, usted podría tomar una muestra aleatoria de peces, calcular la proporción de los especímenes que son mayores de 12 cm y multiplicar esa proporción por la cantidad total de peces, N. Tal procedimiento no puede usarse si se desconoce N. Sin embargo, usted puede pesar toda la redada de peces y utilizar el hecho de que tener una longitud mayor de 12 cm (y) está

relacionado con el peso (x), de modo que

$$\hat{t}_{yr} = \bar{y}\frac{t_x}{\bar{x}}.$$

El peso total de la redada, t_x, se mide con facilidad y $t_x/\bar{x}$ estima la cantidad total de peces en la redada.

3 La estimación por razones se utiliza con frecuencia para aumentar la precisión de las medias y totales estimados. Laplace utilizó la estimación por razones con este fin en el ejemplo que se presentó al principio del capítulo. El aumento en la precisión es el principal uso que será analizado en este capítulo.

En el uso de la estimación por razones por Laplace,

y_i = es el número de personas en la comuna i.

x_i = es el número de nacimientos registrados en la comuna i.

Laplace pudo haber estimado la población total de Francia al multiplicar el número promedio de personas en las 30 comunas ($\bar{y}$) por el número total de comunas en Francia (N). Laplace razonó que la estimación por razones lograría mayor precisión: en promedio, mientras mayor sea la población de una comuna, mayor será el número nacimientos registrados. Así, es probable que el coeficiente de correlación de la población R, definido en la ecuación (3.1), sea positivo. Como entonces $\bar{y}$ y $\bar{x}$ también están correlacionados de manera positiva [véase la ecuación (B.11) del apéndice B], la distribución muestral de $\bar{y}/\bar{x}$ tendrá menos variabilidad que la distribución muestral de $\bar{y}/\bar{x}_U$. Así, si se conoce

t_x = (el número total de nacimientos registrados),

es probable que el error cuadrático medio (ECM) de $\hat{t}_{y_r} = \hat{B}t_x$ sea menor que el ECM de $N\bar{y}$, un estimador que no utiliza la información auxiliar de nacimientos registrados.

4 La estimación por razones se usa para ajustar las estimaciones de la muestra de modo que reflejen los totales demográficos. Una muestra aleatoria simple de 400 estudiantes extraídos de una universidad que cuenta con una población de 4000 alumnos puede contener 240 mujeres y 160 hombres, donde 84 de las mujeres de la muestra y 40 de los hombres de la muestra planean seguir la carrera magisterial. Si sólo se utiliza la información de la muestra aleatoria simple, usted estimaría que

$$\frac{4000}{400} \times 124 = 1240$$

estudiantes planean ser maestros. Si sabemos que la universidad tiene 2700 mujeres y 1300 hombres, una mejor estimación de la cantidad de estudiantes que planean seguir la carrera magisterial sería

$$\frac{84}{240} \times 2700 + \frac{40}{160} \times 1300 = 1270.$$

La estimación por razones se utiliza en cada género: en la muestra, el 60% eran mujeres, pero el 67.5% de la población pertenece al sexo femenino, de modo que debemos ajustar la estimación de la cantidad total de estudiantes que planean seguir la carrera magisterial de acuerdo con esto. Para estimar la cantidad de mujeres que quieren seguir la carrera magisterial, hacemos

$$y_i = \begin{cases} 1 & \text{si es mujer y planea seguir la carrera magisterial} \\ 0 & \text{en caso contrario} \end{cases}$$

$$x_i = \begin{cases} 1 & \text{si es mujer} \\ 0 & \text{en caso contrario.} \end{cases}$$

Entonces $(84/240) \times 2700 = (\sum_{i\in S} y_i / \sum_{i\in S} x_i)t_x$ es una estimación por razones de la cantidad total de mujeres que planean seguir la carrera magisterial. De manera análoga, (40/160) × 1300 es una estimación por razones de la cantidad total de hombres que planean seguir la carrera magisterial.

Analizaremos este uso de la estimación por razones (llamado *postestratificación*) en la sección 4.7 y los capítulos 7 y 8.

5 La estimación por razones se usa para ajustar en caso de presentarse una ausencia de respuestas, como analizaremos en el capítulo 8. Suponga que se extrae una muestra de empresas; sea y_i la cantidad gastada en seguros médicos por la empresa i y x_i el número de empleados de la empresa i. Suponga que conocemos x_i para cada empresa de la población. Esperamos que la cantidad que gasta una empresa en seguros médicos esté relacionada con el número de empleados. Sin embargo, algunas empresas podrían no responder a la encuesta. Un método de ajuste por ausencia de respuestas al estimar los gastos totales en seguros consiste en multiplicar la razón $\bar{y}/\bar{x}$ (al utlizar sólo los datos de quienes respondieron) por la población total t_x. Si es menos probable que las empresas con un menor número de empleados respondan a la encuesta y si y_i es proporcional a x_i, entonces será de esperar que la estimación $N\bar{y}$ sobrestime la población total t_y. En la estimación por razones $t_x\bar{y}/\bar{x}$, $t_x/\bar{x}$ seguramente sea menor que N, debido a que es más probable que las compañías con muchos empleados respondan a la encuesta. Así, la estimación por razones del total de gastos en seguros médicos se ajusta por la falta de respuestas de las compañías con menos empleados.

EJEMPLO 3.2 Regresemos a los datos del censo de agricultura que se llevó a cabo en Estados Unidos, descrito en el ejemplo 2.4. El archivo agsrs.dat contiene datos de una muestra aleatoria simple de 300 de los 3078 condados.

Para este ejemplo, suponga que conocemos los totales de la población para 1987, pero que sólo tenemos la información de 1992 para la muestra aleatoria de 300 condados. Al medir la misma cantidad en instantes diferentes, la respuesta de interés en un tiempo anterior con frecuencia es una excelente variable auxiliar. Sean

y_i = los acres totales de las granjas en el condado i en 1992

x_i = los acres totales de las granjas en el condado i en 1987.

En 1987, un total de $t_x = 964{,}470{,}625$ acres estaban dedicados a la agricultura en los Estados Unidos. El promedio de acres por condado para esta población es entonces $\bar{x}_U = 964{,}470{,}625/3078 = 313{,}343{,}3$ acres de granjas por condado. Graficamos los datos y la recta por el origen con pendiente $\hat{B}$ se muestra en la figura 3.1.

Una parte de una hoja de cálculo con los 300 valores de x_i y y_i aparece en la tabla 3.1. Las celdas C304 y D304 contienen la suma de y y x, respectivamente, para esta muestra, de modo que

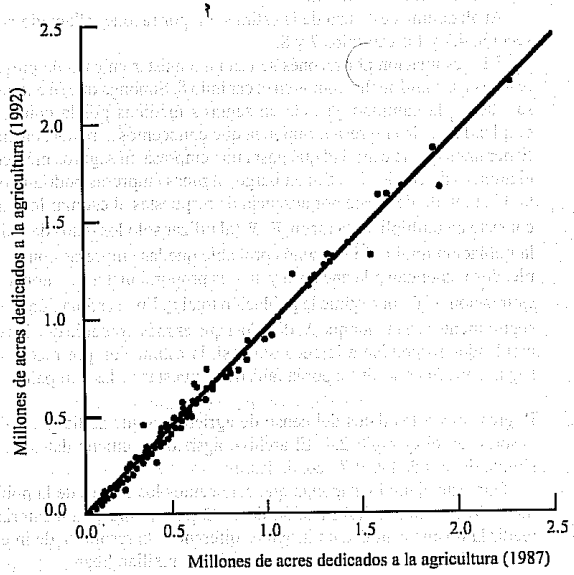
$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{\text{C304}}{\text{D304}} = 0.986565,$$

$$\hat{\bar{y}}_r = \hat{B}\bar{x}_U = (\hat{B})(313{,}343.283) = 309{,}133.6,$$

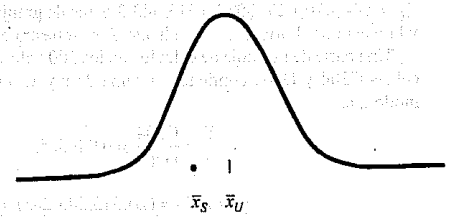
$$\hat{t}_{yr} = \hat{B}t_x = (\hat{B})\,(964{,}470{,}625) = 951{,}513{,}191.$$

Observe que $\bar{y}$ para estos datos es igual a 297,897.0, de modo que $\hat{t}_{y\text{SRS}} = (3078)(\bar{y}) = 916{,}927{,}110$. En este ejemplo, $\bar{x}_S = 301{,}953.7$ es menor que $\bar{x}_U = 313{,}343.3$. Esto significa que nuestra muestra aleatoria simple de tamaño 300 subestima ligeramente la media real de

FIGURA 3.1
Gráfica de los acres, 1992 contra 1987, para una muestra aleatoria simple de 300 condados. La recta en la gráfica pasa por el origen y tiene pendiente $\hat{B} = 0.9866$. Observe que la variabilidad con respecto a la recta aumenta con x.



la población de las x; si la distribución muestral de $\bar{x}$ presenta una forma normal, nuestro valor muestral particular de $\bar{x}_S$ puede estar, aproximadamente, en la posición dada a continuación:

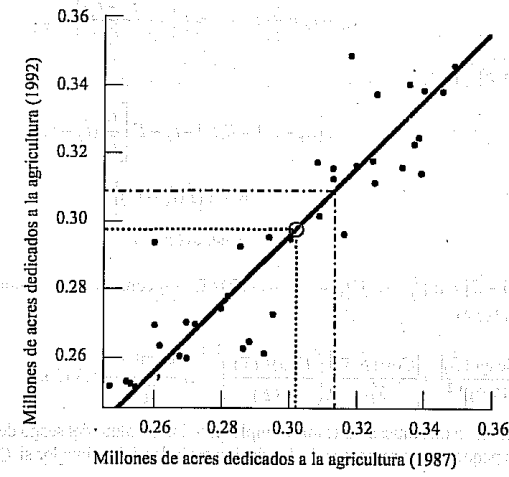


Como las x y las y están correlacionadas en forma positiva, tenemos razón para creer que $\bar{y}_S$ también puede subestimar el valor $\bar{y}_U$ de la población. La estimación por razones proporciona una estimación más precisa de $\bar{y}_U$, desarrollando $\bar{y}_S$ mediante el factor $\bar{x}_U / \bar{x}_S$. La figura 3.2 muestra las estimaciones por razones y por la muestra aleatoria simple de $\bar{y}_U$ sobre una gráfica de la parte central de los datos. ■

TABLA 3.1
Parte de la hoja de cálculo para los datos del censo de agricultura

	A	B	C	D	E
1	Condado	Estado	acres92 (y)	acres87 (x)	Residuo
2					
3	CONDADO COFFEE	AL	175209	179311	−1693.00
4	CONDADO COLBERT	AL	138135	145104	−5019.56
5	CONDADO LAMAR	AL	56102	59861	−2954.78
6	CONDADO MARENGO	AL	199117	220526	−18446.29
7	CONDADO MARION	AL	89228	105586	−14939.48
8	CONDADO TUSCALOOSA	AL	96194	120542	−22728.55
⋮	⋮	⋮	⋮	⋮	⋮
298	CONDADO OZAUKEE	WI	78772	85201	−5284.34
299	CONDADO ROCK	WI	343115	357751	−9829.70
300	CONDADO KANAWHA	WV	19956	21369	−1125.91
301	CONDADO PLEASANTS	WV	15650	15716	145.14
302	CONDADO PUTNAM	WV	55827	55635	939.44
303					
304	Suma por columna		89369114	90586117	3.96176E-09
305	Promedio por columna		297897.0467	301953.7233	
306	Desviación estándar por columna		344551.8948	344829.5964	31657.21817
307	$\hat{B}$ = C304 / D304 =		0.986565237		

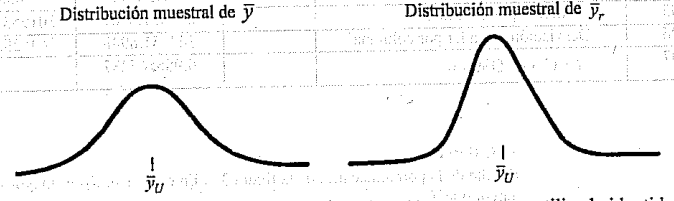
FIGURA 3.2
Detalle de la porción central de la figura 3.1. En este caso, $\bar{x}_U$ es mayor que $\bar{x}_S$, de modo que $\bar{y}_r$ es mayor que $\bar{y}_S$.



3.1.2 Sesgo y error cuadrático medio de los estimadores por razones

A diferencia de los estimadores $\bar{y}$ y $N\bar{y}$ en una muestra aleatoria simple, los estimadores por razones están *sesgados* por lo general para la estimación de $\bar{y}_U$ y t_y. Comenzamos con la estimación insesgada $\bar{y}$; si calculamos $\bar{y}_S$ para cada muestra aleatoria simple posible S, entonces el promedio de todas las medias muestrales de todas las muestras posibles es la media de la población $\bar{y}_U$. El sesgo por estimación en la estimación por razones surge debido a que $\bar{y}$ se multiplica por $\bar{x}_U / \bar{x}$; si calculamos $\hat{\bar{y}}_r$ para todas las muestras aleatorias simples posibles S, entonces el promedio de todos los valores de $\hat{\bar{y}}_r$ a partir de las diferentes muestras será cercano a $\bar{y}_U$ pero por lo general no será igual a $\bar{y}_U$.

Por lo regular la varianza reducida del estimador por razones compensa la presencia del sesgo; aunque $E[\hat{\bar{y}}_r] \neq \bar{y}_U$, el valor de $\hat{\bar{y}}_r$ para cada muestra individual tiene cierta probabilidad de estar más cerca de $\bar{y}_U$ que la *media muestra* $\bar{y}_S$. Después de todo, sólo tomamos una muestra en la práctica; la mayoría de las personas preferirían decir que su estimación particular de la muestra está probablemente, cerca del valor real, en vez de que su valor particular de $\bar{y}_S$ pueda estar un poco lejos de $\bar{y}_U$, pero que la desviación promedio $\bar{y}_S - \bar{y}_U$, promediada sobre todas las muestras S que puedan obtenerse, es cero. Para muestras grandes, las distribuciones muestrales de $\bar{y}$ y $\hat{\bar{y}}_r$ serán aproximadamente normales; si x y y están fuertemente correlacionadas en forma positiva, las siguientes figuras ilustran el sesgo y la varianza relativas de los dos estimadores:



El cálculo del sesgo y la varianza para la estimación por razones utiliza la identidad

$$\hat{t}_{yr} - t_y = \frac{\hat{t}_y}{\hat{t}_x} t_x - t_y = \hat{t}_y \left(1 - \frac{\hat{t}_x - t_x}{\hat{t}_x}\right) - t_y .$$

Como $E[\hat{t}_y] = t_y$,

$$\begin{aligned} E[\hat{t}_{yr} - t_y] &= E[\hat{t}_y] - t_y - E\left[\frac{\hat{t}_y}{\hat{t}_x}(\hat{t}_x - t_x)\right] \\ &= -E\left[\hat{B}(\hat{t}_x - t_x)\right] \\ &= -\operatorname{Cov}(\hat{B}, \hat{t}_x), \end{aligned} \tag{3.3}$$

y $E[\hat{B} - B] = E[\hat{t}_{yr} - t_y]/t_x = -\operatorname{Cov}(\hat{B}, \bar{x})/\bar{x}_U$. En consecuencia, como muestran Hartley y Ross (1954),

$$\frac{|\text{Sesgo}(\hat{B})|}{[V(\hat{B})]^{1/2}} = \left|\frac{\operatorname{Corr}(\hat{B}, \bar{x})}{\bar{x}_U}\right| \left(\frac{V(\hat{B})V(\bar{x})}{V(\hat{B})}\right)^{1/2} \leq \frac{V[\bar{x}]^{1/2}}{\bar{x}_U} + \text{CV}(\bar{x}).$$

Entonces, en una muestra aleatoria simple, el valor absoluto del sesgo del estimador por razones es pequeño con respecto a la desviación estándar del estimador si $\text{CV}(\bar{x})$ es pequeño.

Podemos utilizar un argumento similar al empleado en la sección 2.7 (véase el ejercicio 16 de la página 91) para mostrar que

$$\begin{aligned} E[\hat{B} - B] &\approx \left(1 - \frac{n}{N}\right) \frac{1}{n\bar{x}_U^2} (BS_x^2 - RS_xS_y) \\ &= \frac{1}{\bar{x}_U^2}[BV(\bar{x}) - \operatorname{Cov}(\bar{x}, \bar{y})], \end{aligned} \tag{3.4}$$

donde R es la correlación entre x y y. La última igualdad utiliza la deducción de la covarianza de $\bar{x}$ y $\bar{y}$ en la ecuación (B.10) del apéndice B. Entonces, el sesgo de $\hat{B}$ es pequeño si

- El tamaño de muestra n es grande.
- La fracción de muestreo n/N es grande.
- $\bar{x}_U$ es grande.
- S_x es pequeño.
- La correlación R es cercana a 1.

Para estimar el ECM de $\hat{B}$, la misma identidad utilizada en el cálculo del sesgo implica

$$\begin{aligned} E[(\hat{B} - B)^2] &= E\left[\left(\frac{\bar{y} - B\bar{x}}{\bar{x}}\right)^2\right] \\ &= E\left[\left\{\frac{\bar{y} - B\bar{x}}{\bar{x}_U}\left(1 - \frac{\bar{x} - \bar{x}_U}{\bar{x}}\right)\right\}^2\right] \\ &= E\left[\left(\frac{y - B\bar{x}}{\bar{x}_U}\right)^2 + \left(\frac{\bar{y} - B\bar{x}}{\bar{x}_U}\right)^2 \left\{\left(\frac{\bar{x} - \bar{x}_U}{\bar{x}}\right)^2 - 2\frac{\bar{x} - \bar{x}_U}{\bar{x}}\right\}\right]. \end{aligned}$$

El denominador del primer término es una constante, no una variable aleatoria. Se puede mostrar que el segundo término es pequeño en general, comparado con el primer término, de modo que la varianza y el ECM son aproximados por:

$$E[(\hat{B} - B)^2 \approx E\left[\left(\frac{\bar{y} - B\bar{x}}{\bar{x}_U}\right)^2\right] = \frac{1}{\bar{x}_U^2} E[(\bar{y} - B\bar{x})^2] .$$

Sea

$$d_i = y_i - Bx_i.$$

Entonces, $\bar{y} - B\bar{x} = \bar{d}$, de modo que

$$E\left[(\bar{y} - B\bar{x})^2\right] = V(\bar{d}) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{i=1}^{N} \frac{d_i^2}{N-1} \tag{3.5}$$

y

$$E[(\hat{B} - B)^2] \approx \frac{1}{\bar{x}_U^2} V(\bar{d}).$$

Observe el método que acabamos de utilizar: aproximamos $\hat{B} - B$ por $(\bar{y} - B\bar{x})/\bar{x}_U$, lo cual no contiene cantidades de la muestra en el denominador. Entonces escribimos el nume-

rador como la media muestral de una nueva variable. Una expresión alternativa, algebraicamente equivalente a (3.5), es

$$\frac{1}{\bar{x}_U^2} E\,[\bar{y} - B\bar{x})^2] \doteq \left(1-\frac{n}{N}\right)\frac{1}{n\bar{x}_U^2}(S_y^2 - 2BRS_xS_y + B^2S_x^2). \tag{3.6}$$

(Véase el ejercicio 12).

De (3.5) y (3.6), el ECM aproximado será pequeño cuando

- El tamaño n de la muestra sea grande.
- La fracción de muestreo n/N sea grande.
- Las desviaciones con respecto a la recta $y = Bx$ sean pequeñas.
- La correlación entre x y y sea cercana a +1.
- $\bar{x}_U$ sea grande.

En la práctica, B se desconoce, de modo que no podemos calcular d_i para los valores de la muestra. En vez de esto, usamos lo siguiente:

$$e_i = y_i - \hat{B}x_i,$$

que es el i-ésimo residuo del ajuste de la recta $y = \hat{B}x$. Estimamos la varianza $\hat{B}$ mediante

$$\hat{V}[\hat{B}] = \left(1-\frac{n}{N}\right)\frac{S_e^2}{n\bar{x}_U^2} = \left(1-\frac{n}{N}\right)\frac{1}{n\bar{x}_U^2}\frac{\sum_{i\in S}(y_i - \hat{B}x_i)^2}{n-1} \tag{3.7}$$

Si $\bar{x}_U$ se desconoce, podemos sustituir $\bar{x}_S$ en vez de ella en (3.7).

Como consecuencia de (3.2) y (3.7),

$$\hat{V}[\hat{t}_{yr}] = \hat{V}[t_x\hat{B}] = N^2\left(1-\frac{n}{N}\right)\frac{S_e^2}{n} \tag{3.8}$$

y

$$\hat{V}[\hat{\bar{y}}_r] = \hat{V}[\bar{x}_U\hat{B}] = \left(1-\frac{n}{N}\right)\frac{S_e^2}{n}. \tag{3.9}$$

Si los tamaños de las muestras son lo suficientemente grandes, podemos construir intervalos de confianza del 95% de la siguiente forma:

$$\hat{B} \pm 1.96\text{EE}[\hat{B}], \quad \hat{\bar{y}}_r \pm 1.96\text{EE}[\hat{\bar{y}}_r], \quad \text{o} \quad \hat{t}_{yr} \pm 1.96\text{EE}[\hat{t}_{yr}].$$

En muestras de gran tamaño, el sesgo del estimador es, por lo regular, pequeño con respecto al error estándar (EE), de modo que podemos ignorar el efecto del sesgo en los intervalos de confianza (véase el ejercicio 14).

Observe que si todas las x tuvieran el mismo valor ($S_x = 0$), entonces el estimador del muestreo aleatorio simple sería igual al estimador por razones $\hat{\bar{y}}_r = \bar{y}$ y $EE[\hat{\bar{y}}_r] = \text{EE}[\bar{y}]$.

EJEMPLO 3.3 Regresemos a la muestra obtenida del censo de agricultura. En la hoja de cálculo de la tabla 3.1 creamos la columna E, la cual contiene los residuos $e_i = y_i - \hat{B}x_i$. La desviación estándar de la muestra de la columna E, calculada en la celda E306, es s_e. Así, al utilizar (3.8),

$$\text{EE}(\hat{t}_{yr}) = 3078\sqrt{1-\frac{300}{3078}}\frac{s_e}{\sqrt{300}} = 5{,}344{,}568.$$

Un intervalo de confianza aproximado del 95% para el total de acres agrícolas, al utilizar el estimador por razones, es:

$$951{,}513{,}191 \pm 1.96(5{,}344{,}568) = [941{,}037{,}838,\ 961{,}988{,}544].$$

En contraste, el error estándar de $N\bar{y}_S$ es más de 10 veces mayor:

$$\text{EE}(N\bar{y}_S) = 3078\sqrt{\left(1-\frac{300}{3078}\right)}\frac{s_y}{\sqrt{300}} = 58{,}169{,}381.$$

El coeficiente de variación estimado (CV) para el estimador por razones es 5,344,568/951,513,191 = 0.0056, en comparación con el CV de 0.0634 para el estimador de la muestra aleatoria simple $N\bar{y}$ que no utiliza la información auxiliar. La inclusión de la información de 1987 a través del estimador por razones aumentó en mucho la precisión. Si todas las cantidades por estimar estuvieran altamente correlacionadas con los acres en 1987, podríamos reducir, de manera drástica el tamaño de la muestra y aún así obtener una alta precisión al emplear estimaciones por razones en vez de $N\bar{y}$. ■

EJEMPLO 3.4 Demos otro vistazo a la población hipotética que se empleó en el ejemplo 2.1 para exhibir la distribución muestral de $\hat{t}_{yr}$. Ahora supongamos que también tenemos una medición auxiliar x para cada unidad de la población; los valores para la población son los siguientes:

Número de unidad	x	y
1	4	1
2	5	2
3	5	4
4	6	4
5	8	7
6	7	7
7	7	7
8	5	8

Observe que x y y están correlacionados en forma positiva. Podemos calcular las cantidades para la población, pues conocemos toda la población y la distribución muestral:

$$t_x = 47 \qquad t_y = 40$$
$$S_x = 1.3562027 \qquad S_y = 2.618615$$
$$R = 0.6838403 \qquad B = 0.8510638$$

La tabla 3.2 contiene una parte de la distribución muestral de $\hat{t}_{yr}$. La figura 3.3 proporciona los histogramas de las distribuciones muestrales de dos estimaciones de t_y: $\hat{t}_{MAS} = N\bar{y}$, la estimación utilizada en el capítulo 2 y $\hat{t}_{yr}$. La distribución muestral para la estimación por razones no se difunde tanto como la distribución muestral para $N\bar{y}$; es más asimétrica que simétrica. La asimetría conduce a un ligero sesgo por estimación de la estimación por razones. El total de la población es $t_y = 40$; el valor medio de la distribución muestral de $\hat{t}_{yr}$ es igual a 39.85063.

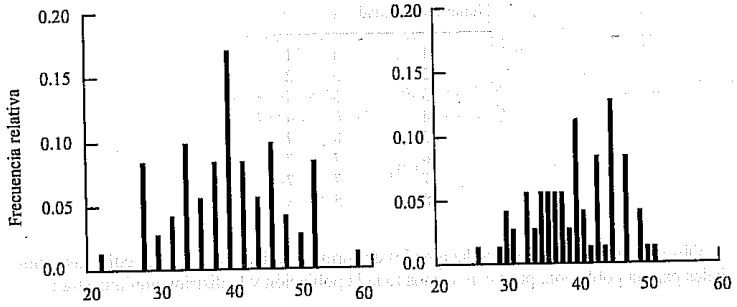
El valor medio de la distribución muestral de $\hat{B}$ es 0.8478857, lo que produce un sesgo de –0.003178. Si usamos las cantidades anteriores para la población, el sesgo aproximado de (3.4) es

$$\left(1-\frac{n}{N}\right)\frac{1}{n\bar{x}_U^2}(BS_x^2 - RS_xS_y) = -0.003126.$$

TABLA 3.2
Distribución muestral para $\hat{t}_{yr}$.

Número de muestra	Muestra, S	$\bar{x}_S$	$\bar{y}_S$	$\hat{B}$	$\hat{t}_{MAS}$	$\hat{t}_{yr}$
1	{1, 2, 3, 4}	5.00	2.75	0.55	22.00	25.85
2	{1, 2, 3, 5}	5.50	3.50	0.64	28.00	29.91
3	{1, 2, 3, 6}	5.25	3.50	0.67	28.00	31.33
4	{1, 2, 3, 7}	5.25	3.50	0.67	28.00	31.33
5	{1, 2, 3, 8}	4.75	3.75	0.79	30.00	37.11
6	{1, 2, 4, 5}	5.75	3.50	0.61	28.00	28.61
⋮	⋮	⋮	⋮	⋮	⋮	⋮
67	{4, 5, 6, 8}	6.50	6.50	1.00	52.00	47.00
68	{4, 5, 7, 8}	6.50	6.50	1.00	52.00	47.00
69	{4, 6, 7, 8}	6.25	6.50	1.04	52.00	48.88
70	{5, 6, 7, 8}	6.75	7.25	1.07	58.00	50.48

FIGURA 3.3
Distribuciones muestrales para (a) $\hat{t}_{MAS}$ y (b) $\hat{t}_{yr}$



Estimación de t mediante una muestra aleatoria simple — Estimación de t por razones

La varianza de la distribución muestral de $\hat{B}$, calculada mediante la definición de varianza en (2.4), es igual a 0.015186446; la aproximación en (3.6) es

$$\left(1-\frac{n}{N}\right)\frac{1}{n\bar{x}_U^2}(S_y^2-2BRS_xS_y+B^2S_x^2)=0.01468762. \quad \blacksquare$$

3.1.2.1 Exactitud de la aproximación del ECM

El ejemplo 3.4 demuestra que la aproximación del ECM en (3.6) sólo es, en realidad, una aproximación; es una buena aproximación en ese ejemplo aunque la población y la muestra sean pequeñas.

Para que (3.6) sea una buena aproximación del ECM, el sesgo debería ser pequeño al igual que los términos descartados en la aproximación de la varianza. Si el coeficiente de variación de $\bar{x}$ es pequeño, es decir, si $\bar{x}_U$ se estima con una precisión relativa alta, entonces, el sesgo es pequeño con respecto a la raíz cuadrada de la varianza. Si formamos un intervalo de confianza al emplear $\hat{t}_{yr} \pm 1.96\text{EE}[\hat{t}_{yr}]$ y usamos (3.8) para determinar la varianza y el error estándar estimados, entonces el sesgo no tendrá un mayor efecto sobre la probabilidad de cobertura del intervalo de confianza. Un $\text{CV}(\bar{x})$ pequeño también significa que $\bar{x}$ es estable de una muestra a otra y que es probable que $\bar{x}$ no se anule, un resultado deseable, pues dividimos entre $\bar{x}$ al formar la estimación por razones. Sin embargo, en algunos de los complejos diseños de muestreo que analizaremos en capítulos posteriores, el sesgo puede ser un punto de preocupación; regresaremos a este aspecto en los capítulos 9 y 12.

Para que (3.6) sea una buena aproximación del ECM, queremos un tamaño de muestra grande (n mayor que 30 o algo así) y $\text{CV}(\bar{x}) \leq .1$, $\text{CV}(\bar{y}) \leq .1$. Si estas condiciones no se cumplen, entonces (3.6) podría subestimar severamente al verdadero ECM.

3.1.2.2 Ventajas de la estimación por razones

¿Qué ganamos al utilizar la estimación por razones? Si las desviaciones de y_i con respecto a $\hat{B}x_i$ son menos que las desviaciones de y_i con respecto a $\bar{y}$, entonces $\hat{V}[\hat{\bar{y}}_r] \leq \hat{V}[\bar{y}]$. Recuerde, del capítulo 2, que

$$\text{ECM}[\bar{y}] = V[\bar{y}] = \left(1-\frac{n}{N}\right)\frac{S_y^2}{n}.$$

Al utilizar la aproximación en (3.6),

$$\text{ECM}[\hat{\bar{y}}_r] \approx \left(1-\frac{n}{N}\right)\frac{1}{n}(S_y^2-2BRS_xS_y+B^2S_x^2).$$

Así,

$$\text{ECM}[\hat{\bar{y}}_r]-\text{ECM}[\bar{y}] \approx \left(1-\frac{n}{N}\right)\frac{1}{n}(S_y^2-2BRS_xS_y+B^2S_x^2-S_y^2)$$

$$=\left(1-\frac{n}{N}\right)\frac{1}{n}S_xB(-2RS_y+BS_x).$$

Así, en cuanto a la exactitud de la aproximación,

$$\text{ECM}[\hat{\bar{y}}_r] \leq \text{ECM}[\bar{y}] \text{ si y sólo si } R \geq \frac{BS_x}{2S_y}=\frac{\text{CV}(x)}{2\text{CV}(y)}.$$

Si los coeficientes de variación son aproximadamente iguales, entonces es mejor utilizar la estimación por razones cuando la correlación entre x y y es mayor que 1/2.

La estimación por razones es más adecuada si una recta por el origen resume la relación entre x_i y y_i y si la varianza de y_i con respecto a la recta es proporcional a x_i. Bajo estas condiciones, $\hat{B}$ es la pendiente de regresión ponderada por mínimos cuadrados para la recta que pasa por el origen, donde los pesos son proporcionales a $1/x_i$; es decir, la pendiente $\hat{B}$ minimiza la suma de cuadrados

$$\sum_{i\in S}\frac{1}{x_i}(y_i-\hat{B}x_i)^2.$$

3.1.3 Estimación por razones y las proporciones

La estimación por razones funciona de la misma manera cuando la cantidad de interés es una proporción.

EJEMPLO 3.5 Peart (1994) reunió los datos de la tabla 3.3 como parte de un estudio para evaluar los efectos de la actividad de los puercos salvajes y la sequía sobre la vegetación nativa de la isla Santa Cruz en California. La investigadora contó el número de brotes de árbol en las áreas protegidas bajo cada uno de los 10 robles de la muestra en marzo de 1992, después de las lluvias que concluyeron con la sequía de 1991. La científica colocó una bandera en cada brote, al determinar posteriormente cuántos de ellos seguían vivos en febrero de 1994. Graficamos los datos (cortesía de Dianne Peart) en la figura 3.4.

Cuando la mayoría de las personas que han asistido a un curso de introducción a la estadística observan datos como éstos, quieren encontrar la proporción muestral de brotes de 1992 que continúan vivos en 1994 y luego utilizar la fórmula para la varianza de una variable aleatoria binomial para calcular el error estándar de la estimación. El uso del error estándar binomial es *incorrecto* para estos datos, pues la distribución binomial requiere que los ensayos sean independientes; en este ejemplo, este supuesto es inadecuado. La supervivencia de los brotes depende de muchos factores, como las precipitaciones locales, la cantidad de luz y la depredación. Es probable que tales factores afecten a los brotes de un mismo terreno de manera similar, lo que conduce a que los distintos terrenos tengan, en general, distintas tasas de supervivencia. El tamaño de muestra en este ejemplo es igual a 10, no a 206.

El diseño es, en realidad, una **muestra de cúmulo**; los cúmulos son los terrenos asociados a cada árbol y las unidades de observación son los brotes individuales en esos terrenos. Para analizar este ejemplo desde el marco de referencia de la estimación por razones, sean

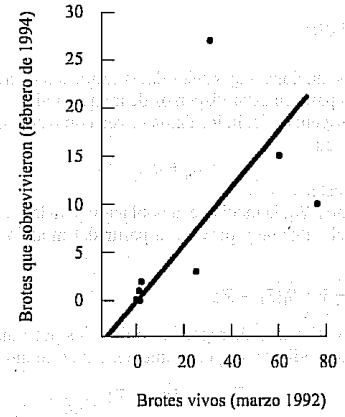
y_i = la cantidad de brotes cerca del árbol i que están vivos en 1994

x_i = la cantidad de brotes cerca del árbol i que están vivos en 1992.

TABLA 3.3
Datos de brotes en la isla Santa Cruz

Árbol	x = Número de brotes, 3/92	y = Brotes vivos, 2/94
1	1	0
2	0	0
3	8	1
4	2	2
5	76	10
6	60	15
7	25	3
8	2	2
9	1	1
10	31	27
Total	206	61
Promedio	20.6	6.1
Desviación estándar	27.4720	8.8248

FIGURA 3.4
La gráfica de brotes que sobrevivieron (febrero de 1994) contra los brotes vivos (marzo 1992), para diez árboles de roble.



Entonces, la estimación por razones de la proporción de brotes que siguen vivos en 1994 es:

$$\hat{B} = \hat{p} = \frac{\bar{y}}{\bar{x}} = \frac{6.1}{20.6} = 0.2961.$$

Si usamos (3.7) e ignoramos la corrección para poblaciones finitas,

$$\text{EE}[\hat{B}] = \sqrt{\frac{1}{(10)(20.6)^2}\frac{\sum_{i\in S}(y_i - 0.2961165x_i)^2}{9}}$$

$$= \sqrt{\frac{56.3778}{(10)(20.6)^2}}$$

$$= 0.115.$$

De haber utilizado la fórmula binomial, habríamos calculado un error estándar de

$$\sqrt{\frac{(0.2961)(0.7039)}{206}} = .0318,$$

que es mucho menor y daría una impresión equivocada de precisión.

La aproximación a la varianza de $\hat{B}$ en este ejemplo podría no ser particularmente buena, pues el tamaño de la muestra es pequeño; aunque es probable que la varianza estimada de $\hat{B}$ sea una subestimación, seguirá siendo mejor que el cálculo de la varianza mediante la distribución binomial, pues los brotes no son independientes. ■

3.2 Estimación por regresión

3.2.1 Uso de un modelo de línea recta

La estimación por razones funciona mejor si los datos se ajustan bien a una línea recta por el origen. A veces, los datos parecen estar dispersos de manera uniforme en torno a una línea recta que no pasa por el origen; es decir, los datos se ven como si el modelo usual de regresión mediante una línea recta

$$y = B_0 + B_1 x$$

proporcionará un buen ajuste.

Suponga que conocemos $\bar{x}_U$, la media de la población para las x. Entonces el estimador por regresión de $\bar{y}_U$ es el valor de y previsto a partir del modelo de regresión ajustado cuando $x = \bar{x}_U$:

$$\hat{\bar{y}}_{\text{reg}} = \hat{B}_0 + \hat{B}_1 \bar{x}_U = \bar{y} + \hat{B}_1(\bar{x}_U - \bar{x}), \tag{3.10}$$

donde $\hat{B}_0$ y $\hat{B}_1$ son los coeficientes de regresión ordinarios por mínimos cuadrados de la ordenada al origen y de la pendiente, respectivamente. Para este modelo,

$$\hat{B}_1 = \frac{\sum_{i\in S}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i\in S}(x_i - \bar{x})^2} = \frac{rs_y}{s_x},$$

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x},$$

y r es el coeficiente de correlación de x y y para la muestra.

Como el estimador por razones, el estimador por regresión es sesgado. Sea B_1 la pendiente de regresión por mínimos cuadrados calculada mediante todos los datos de la población:

$$B_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{\sum_{i=1}^{N}(x_i - \bar{x}_U)^2} = \frac{PR_y}{S_x}.$$

Entonces, usando (3.10), el sesgo de $\hat{\bar{y}}_{\text{reg}}$ está dado por

$$E[\hat{\bar{y}}_{\text{reg}} - \bar{y}_U] = E[\bar{y} - \bar{y}_U] + E[\hat{B}_1(\bar{x}_U - \bar{x})] = -\text{Cov}(\hat{B}_1, \bar{x}). \tag{3.11}$$

Si la recta de regresión pasa por todos los puntos (x_i, y_i) de la población, entonces el sesgo es nulo. En ese caso, $B_1 = B_1$ para cada muestra, de modo que $\text{Cov}(\hat{B}_1, \bar{x}) = 0$.

Como en el caso de la estimación por razones, la estimación por regresión del error cuadrático medio para muestras aleatorias simples es aproximadamente igual a la varianza (véase el ejercicio 18); con frecuencia, el sesgo se puede despreciar en muestras grandes.

El método utilizado para aproximar el error cuadrático medio en la estimación por razones también se puede aplicar a la estimación por regresión. Sea $d_i = y_i - [\bar{y}_U + B_1(x_i - \bar{x}_U)]$; entonces

$$\begin{aligned}\text{ECM}(\hat{\bar{y}}_{\text{reg}}) &= E[\{\bar{y} + \hat{B}_1(\bar{x}_U - \bar{x}) - \bar{y}_U\}^2] \\ &\approx V[\bar{d}] \\ &= \left(1 - \frac{n}{N}\right)\frac{S_d^2}{n}.\end{aligned} \tag{3.12}$$

Si usamos la relación $B_1 = RS_y/S_x$, podemos mostrar que

$$\left(1 - \frac{n}{N}\right)\frac{S_d^2}{n} = \left(1 - \frac{n}{N}\right)\frac{1}{n}\sum_{i=1}^{N}\frac{(y_i - \bar{y}_U - B_1[x_i - \bar{x}_U])^2}{N-1} \tag{3.13}$$

$$= \left(1 - \frac{n}{N}\right)\frac{1}{n}S_y^2(1 - R^2).$$

(Véase el ejercicio 17). Entonces, el error cuadrático medio es pequeño cuando

- n es grande.
- n/N es grande.
- S_y es pequeño.
- La correlación R es cercana a -1 o $+1$.

Podemos calcular el error estándar al determinar la varianza de los residuos para la muestra. Sea $e_i = y_i - (\hat{B}_0 + \hat{B}_1 x_i)$; entonces

$$\text{EE}(\hat{\bar{y}}_{\text{reg}}) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}}. \tag{3.14}$$

EJEMPLO 3.6 Para estimar el número de árboles muertos que se encuentran en una determinada área, dividimos ésta en 100 terrenos cuadrados y contamos el número de árboles muertos en una fotografía de cada terreno. El conteo en las fotos se puede hacer rápidamente, pero a veces un árbol queda mal clasificado o no es detectado. Así, elegimos una muestra aleatoria simple de 25 de los terrenos divididos para el conteo de campo de los árboles muertos. Por el conteo con las fotos, sabemos que la cantidad media de árboles muertos por cada terreno es de 11.3. Los datos, graficados en la figura 3.5, y algunos resultados SAS son los siguientes:

Estadísticas simples

Foto	10	12	7	13	13	6	17	16	15	10	14	12	10
Campo	15	14	9	14	8	5	18	15	13	15	11	15	12
Foto	5	12	10	10	9	6	11	7	9	11	10	10	
Campo	8	13	9	11	12	9	12	13	11	10	9	8	

Variable	N	Desviación estándar	Suma	Mínimo	Máximo
FOTO	25	10.6000	3.0687	265.0000	5.0000	17.0000
CAMPO	25	11.5600	3.0150	289.0000	5.0000	18.0000

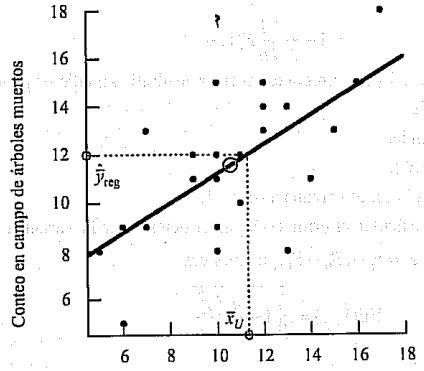
Variable dependiente: CAMPO

Análisis de varianza

Fuente	Grados de libertad	Suma de cuadrados	Cuadrado medio	F Valor	Prob>F
Modelo	1	84.99982	84.99982	14.682	0.0009
Error	23	133.16018	5.78957		
C total	24	218.16000			

(La salida continúa en la página 76).

FIGURA 3.5
Gráfica de datos del conteo de árboles en fotos y en campo, junto con la recta de regresión. Observe que $\hat{\bar{y}}_{reg}$ es el valor previsto a partir de la ecuación de regresión cuando $x = \bar{x}_U$.



Raíz ECM	2.40615	R-cuadrada	0.3896
Dep media	11.56000	R-cuadrada adjunta	0.3631
CV	20.81447		

Estimaciones de los parámetros

Variable	Grados de libertad	Estimación del parámetro	Error estándar	T para H0: Parámetro = 0	Prob>|T|
INTERSECCIÓN	1	5.059292	1.76351187	2.869	0.0087
FOTO	1	0.613274	0.16005493	3.832	0.0009

Si usamos (3.10), la estimación por regresión de la media es:

$$\hat{\bar{y}}_{reg} = 5.06 + 0.613(11.3) = 11.99.$$

Los resultados SAS nos permiten calcular s_e^2 a partir de la suma de cuadrados de los residuos: $s_e^2 = 133.16018/24 = 5.54834$ (en forma alternativa, usted puede utilizar el error estándar medio de los residuos, que divide entre n-1 en vez de n-2). Así, el error estándar es, por (3.14),

$$EE[\hat{\bar{y}}_{reg}] = \sqrt{\left(1 - \frac{25}{100}\right)\frac{5.54834}{25}} = 0.408.$$

De nuevo, el error estándar es menor que el correspondiente a $\bar{y}$:

$$EE[\bar{y}] = \sqrt{\left(1 - \frac{25}{100}\right)\frac{s_y^2}{25}} = 0.522.$$

Esperamos que la estimación por regresión aumente la precisión en este ejemplo, pues las

variables *foto* y *campo* están correlacionadas en forma positiva ($r = 0.62$). Para estimar la cantidad total de árboles muertos, utilizamos

$$\hat{t}_{yreg} = (100)(11.99) = 1199;$$

$$EE[\hat{t}_{yreg}] = (100)(0.408) = 40.8.$$

Un intervalo de confianza aproximado del 95% para la cantidad total de árboles muertos está dado por:

$$1199 \pm (2.07)(40.8) = [1114, 1283].$$

Debido al tamaño relativamente pequeño de la muestra, utilizamos el percentil de la distribución t (con n-2 = 23 grados de libertad) de 2.07 en vez del percentil de la distribución normal de 1.96. ■

3.2.2 Estimación por diferencias

La **estimación por diferencias** es un caso particular de la estimación por regresión, que se utiliza cuando el investigador "sabe" que la pendiente B_1 es 1. Con frecuencia, se recomienda el uso de la estimación por diferencias al usar las muestras aleatorias simples en contabilidad. Una lista de cuentas por pagar consta del valor en libros de cada cuenta; la lista de lo que se debe en cada cuenta. En el esquema de muestreo más sencillo, el auditor realiza un escrutinio mediante una muestra aleatoria de las cuentas para determinar el valor auditado (la cantidad real que se debe) para estimar el error en el total de cuentas por pagar. Las cantidades consideradas son:

y_i = el valor auditado para la compañía i.

x_i = el valor en libros para la compañía i.

Entonces $\bar{y} - \bar{x}$ es la diferencia media para las cuentas auditadas.

La diferencia total estimada es $\hat{t}_y - \hat{t}_x = N(\bar{y} - \bar{x})$; el valor auditado estimado para las cuentas por pagar es $\hat{t}_{ydif} = t_x + (\hat{t}_y - \hat{t}_x)$.

De nuevo, definimos los residuos a partir de este modelo: En este caso, $e_i = y_i - x_i$. La varianza de $\hat{t}_{ydif}$ es:

$$V(\hat{t}_{ydif}) = V[t_x + (\hat{t}_y - \hat{t}_x)] = V(\hat{t}_e),$$

donde $\hat{t}_e = (N/n)\sum_{i \in S} e_i$. Si la variabilidad en los residuos e_i es menor que la variabilidad entre las y_i, entonces la estimación por diferencias aumentará la precisión.

La estimación por diferencias funciona mejor si la población y la muestra tienen una alta proporción de diferencias no nulas que queden separadas casi por partes iguales en excesos y defectos y si la muestra es lo bastante grande como para que la distribución muestral de $(\bar{y} - \bar{x})$ sea aproximadamente normal.

En las auditorías, es posible que todos los valores auditados en la muestra sean iguales a los valores correspondientes en los libros. Entonces $\bar{y} = \bar{x}$ y el error estándar de $\hat{t}_y$ sería calculado como cero. En tal caso, cuando la mayor parte de las diferencias son nulas, se necesita un modelo más sofisticado.

3.3 Estimación en dominios

Con frecuencia queremos separar las estimaciones de alguna subpoblación; las subpoblaciones se llaman **dominios** o **subdominios**. Queremos extraer una muestra aleatoria simple de los visitantes que volaron a Nueva York el 18 de septiembre y estimar la proporción de visitantes

de otro estado que pretenden quedarse más de una semana. Para ese estudio, tenemos dos dominios de estudio: los visitantes provenientes del mismo estado y los que pertenecen a otra entidad federativa. Sin embargo, no sabemos cuáles son las personas de la población que pertenecen a alguno de los dominios hasta obtener la muestra. Así, el número de personas que se encuentren en una muestra aleatoria simple que estén en cada dominio, constituyen una variable aleatoria, con un valor desconocido al momento de diseñar la encuesta.

Suponga que existen D dominios. Sea $\mathcal{U}_d$ el conjunto de índices de las unidades en la población que están en el dominio d y sea $\mathcal{S}_d$ el conjunto de índices en la muestra que están en el dominio d, para $d = 1,2,\ldots, D$. Sea N_d el número de unidades de población en $\mathcal{U}_d$ y n_d el número de unidades muestra en $\mathcal{S}_d$. Suponga que queremos estimar

$$\bar{y}_{U_d} = \sum_{i\in\mathcal{U}_d} \frac{yi}{N_d}.$$

Un estimador natural de $\bar{y}_{U_d}$ es

$$\bar{y}_d = \sum_{i\in\mathcal{S}_d} \frac{y_i}{n_d}, \tag{3.15}$$

que se parece, en principio, a las medias muestrales estudiadas en el capítulo 2.

Sin embargo, la cantidad n_d es una variable aleatoria: Si se extrae una muestra aleatoria simple, es muy probable que tenga un valor distinto de n_d. Las diferentes muestras extraídas en la ciudad de Nueva York tendrían diversas cantidades de visitantes de otro estado. Desde el punto de vista técnico, (3.15) es una estimación por razones. Para ver esto, sea

$$u_i = \begin{cases} y_i & \text{si } i \in \mathcal{U}_d \\ 0 & \text{si } i \notin \mathcal{U}_d \end{cases}$$

$$x_i = \begin{cases} 1 & \text{si } i \in \mathcal{U}_d \\ 0 & \text{si } i \notin \mathcal{U}_d \end{cases}$$

Entonces, $\bar{x}_U = N_d/N, \bar{y}_{U_d} = \Sigma_{i=1}^N u_i / \Sigma_{i=1}^N x_i$ y

$$\bar{y}_d = \hat{B} = \frac{\bar{u}}{\bar{x}} = \frac{\sum_{i\in S} u_i}{\sum_{i\in S} x_i}.$$

Como estamos estimando una razón, utilizamos (3.7) para calcular el error estándar:

$$\begin{aligned} \mathrm{EE}(\bar{y}_d) &= \sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n\bar{x}_U^2}\frac{\sum_{i\in S}(u_i - \hat{B}x_i)^2}{n-1}} \\ &= \sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n\bar{x}_U^2}\frac{\sum_{i\in S_d}(y_i - \hat{B})^2}{n-1}} \\ &= \sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n}\left(\frac{N}{N_d}\right)^2\frac{(n_d-1)s_{yd}^2}{n-1}} \\ &\approx \sqrt{\left(1-\frac{n}{N}\right)\frac{s_{yd}^2}{n_d}}. \end{aligned}$$

La aproximación del último renglón depende de un tamaño de muestra grande en el dominio d; si la muestra es lo bastante grande, entonces esperaríamos que $n_d/n \approx N_d/N$ y $(n_d-1)/(n-1) \approx n_d/n$. En una muestra grande, el error estándar de $\bar{y}_d$ es aproximadamente igual al obtenido al usar la fórmula (2.10). Así, en una muestra que sea lo suficientemente grande, el detalle técnico que consiste en emplear un estimador por razones, establece poca diferencia en la práctica para estimar la media de un dominio.

La situación es un poco más compleja al estimar el total de un dominio. Si conocemos N_d, la estimación es sencilla: Utilizamos $N_d\bar{y}_d$. Empero, si no conocemos N_d, necesitamos estimarlo mediante Nn_d/n. Entonces

$$\hat{t}_{yd} = N\frac{n_d}{n}\frac{\sum_{i\in S} u_i}{n_d} = N\bar{u}.$$

El error estándar es

$$\mathrm{EE}(\hat{t}_{yd}) = N\,\mathrm{EE}(\bar{u}) = N\sqrt{\left(1-\frac{n}{N}\right)\frac{s_u^2}{n}}.$$

EJEMPLO 3.7 En la muestra aleatoria simple de tamaño 300, del censo de agricultura (véase el ejemplo 2.4), 39 condados están en los estados del oeste.[2] ¿Cuál es la cantidad total de acres dedicados a la agricultura en el oeste?

La media muestral de los 39 condados es $\bar{y}_d = 598,680.6$, con una desviación estándar muestral $s_{yd} = 516,157.7$. Así,

$$\mathrm{EE}(\hat{y}_d) = \sqrt{\left(1-\frac{300}{3078}\right)}\frac{516,157.7}{\sqrt{39}} = 78,520.$$

Así, $\widehat{\mathrm{CV}}[\bar{y}_d] = 0.1312$ y un intervalo de confianza aproximado del 95% para la media del total de acres dedicados a la agricultura, para los condados del Oeste de Estados Unidos, es: 444,781, 752,580.

Para estimar el número total de acres dedicados a la agricultura en el Oeste, suponga que no sabemos cuántos condados de la población pertenecen al Oeste. Entonces, se tiene que definir lo siguiente:

$u_i = y_i$ si el condado i está en el Oeste de los Estados Unidos
$u_i = 0$ en caso contrario

Entonces,

$$\hat{t}_{yd} = N\bar{u} = 3078(77,828.48) = 239,556,051.$$

El error estándar es

$$\mathrm{EE}(\hat{t}_{yd}) = 3078\sqrt{\left(1-\frac{300}{3078}\right)}\frac{273,005.4}{\sqrt{300}} = 46,090,460.$$

El coeficiente de variación estimado para $\hat{t}_{yd}$ es $\widehat{\mathrm{CV}}[\hat{t}_{yd}] = 46,090,460/239,556,051 = 0.1924$; de conocer el número de condados en el Oeste de los Estados Unidos y poder utilizar ese valor en la estimación, el coeficiente de variación para el total estimado, sería igual a 0.1312, el coeficiente de variación para la media estimada. ■

EJEMPLO 3.8 Se extrae una muestra aleatoria simple de 1500 dueños de botes con licencia en cierto estado, de una lista de 400,000; 472 de las personas que respondieron a la encuesta, afirmaron poseer un bote con motor fuera de borda con más de 16 pies. Estas 472 personas

[2] Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, Nuevo México, Oregon, Utah, Washington y Wyoming.

informaron tener los siguientes números de hijos:

Número de hijos	Número de personas que respondieron
0	76
1	139
2	166
3	63
4	19
5	5
6	3
8	1
Total	472

Para estimar el porcentaje de dueños de grandes botes con motor fuera de borda con hijos, podemos utilizar $\hat{p} = 396/472 = 0.839$. Este es un estimador por razones, pero en este caso, como ya explicamos, el error estándar es aproximadamente el que debería de ser. Si ignoramos la corrección para poblaciones finitas,

$$\text{EE}(\hat{p}) = \sqrt{\frac{.839(1-.839)}{471}} = 0.017.$$

Para analizar el número promedio de hijos por familia entre los dueños de botes que registraron una embarcación de motor con más de 16 pies, observe que el número promedio de hijos para las 472 personas que respondieron y que están en ese dominio es de 1.667373, con una varianza de 1.398678. Así, un intervalo de confianza aproximado del 95% para el número promedio de hijos en las familias poseedoras de botes grandes es:

$$1.667 \pm 1.96\sqrt{\frac{1.398678}{472}} = [1.56, 1.77].$$

Para estimar el número total de hijos en el estado cuyos padres registran un bote grande, creamos una nueva variable u para las personas que respondieron; este valor es igual al número de hijos si la persona que responde tiene un bote de motor y cero en caso contrario. La distribución de frecuencias para la variable u es, entonces,

Número de hijos	Número de personas que respondieron
0	1104
1	139
2	166
3	63
4	19
5	5
6	3
8	1
Total	1500

Ahora, $\bar{u} = 0.52466$ y $s_u^2 = 1.0394178$, de modo que $\hat{t}_{yd} = 400,000(.524666) = 209,867$ y

$$\text{EE}(\hat{t}_{yd}) = \sqrt{(400,000)^2 \frac{1.0394178}{1500}} = 10,529.5.$$

En este ejemplo, la variable u_i sólo cuenta el número de hijos en la familia i que pertenecen a una familia con un bote grande de motor. ■

En esta sección mostramos que la estimación de medias de un dominio, es un caso particular de estimación por razones, pues el tamaño de la muestra en el dominio varía de una muestra a otra. Si el tamaño de la muestra para el dominio de una muestra aleatoria simple es lo suficientemente grande, podemos utilizar las fórmulas de las muestras aleatorias simples para realizar inferencias con respecto a la media del dominio.

Las inferencias de los totales dependen del conocimiento o no del tamaño de la población del dominio, N_d. Si lo conocemos, entonces, el total estimado es $N_d\bar{y}_d$. Si lo ignoramos entonces, definimos una nueva variable u_i que sea igual a y_i para las observaciones del dominio y cero para las observaciones que no están en el dominio; luego utilizamos $\hat{t}_u$ para estimar el total del dominio.

Los resultados de esta sección sólo se aplican a las muestras aleatorias simples. En la sección 12.3, analizaremos la estimación de las medias del dominio si los datos se reúnen mediante otros diseños de muestreo.

3.4 Modelos para la estimación por razones y por regresión*

Muchos estadísticos han propuesto que (1) si un modelo de regresión proporciona un buen ajuste para el estudio de los datos, entonces el modelo debe usarse para estimar el total de y su error estándar y (2) la forma en que uno obtiene los datos no es tan importante como el modelo que se ajusta. En esta sección analizaremos los modelos que dan las estimaciones puntuales en las ecuaciones (3.2) y (3.10) para el cálculo por razones y por regresión. Sin embargo, las varianzas bajo un enfoque basado en el modelo son ligeramente distintas, como veremos más adelante.

3.4.1 Un modelo para la estimación por razones

Anteriormente, establecimos que la estimación por razones es más adecuada en una muestra aleatoria simple cuando una línea recta por el origen se ajusta bien y cuando la varianza de las observaciones con respecto a esta recta es proporcional a x. Podemos enunciar estas condiciones como un modelo de regresión lineal: Suponga que conocemos $x_1, x_2, \ldots, x_N$ (y todos ellos son mayores que cero) y que $Y_1, Y_2, \ldots, Y_N$ son independientes y siguen el siguiente modelo:

$$Y_i = \beta x_i + \varepsilon_i, \tag{3.16}$$

donde $E_M[\varepsilon_i] = 0$ y $V_M[\varepsilon_i] = \sigma^2 x_i$. La independencia de las observaciones en el modelo es una afirmación explícita de que el diseño de muestreo no proporciona la información que pueda utilizarse para estimar las cantidades de interés: el procedimiento de muestreo no tiene efectos sobre la validez del modelo. Bajo este modelo, $T_y = \sum_{i=1}^{N} Y_i$ es una variable aleatoria, y el total de la población de interés, t_y, es una realización de la variable T_y (esto contrasta con el enfoque de aleatorización, donde t_y se considera como una cantidad fija pero desconocida y las únicas variables aleatorias son los indicadores de la muestra Z_i). Si S representa el conjunto de unidades en nuestra muestra, entonces

$$t_y = \sum_{i \in S} y_i + \sum_{i \notin S} y_i.$$

Observamos los valores de y_i para las unidades de la muestra y predecimos dichos valores para las unidades que no están en dicha muestra como $\hat{\beta}x_i$, donde $\hat{\beta} = \bar{y}/\bar{x}$ es la estimación

de β por mínimos cuadrados ponderados bajo el modelo en (3.16). Entonces, una estimación natural de t_y es:

$$\hat{t}_y = \sum_{i\in S} y_i + \hat{\beta}\sum_{i\in S} x_i = n\bar{y} + \frac{\bar{y}}{\bar{x}}\sum_{i\in S} x_i = \frac{\bar{y}}{\bar{x}}\sum_{i=1}^{N} x_i = \frac{\bar{y}}{\bar{x}}t_x.$$

Ésta es simplemente la estimación por razones de t_y.

En muchos esquemas comunes de muestreo, vemos que si adoptamos un modelo consistente con las razones por las cuales adoptamos cierto esquema de muestreo o método de estimación, entonces, los estimadores puntuales obtenidos mediante tal modelo son muy parecidos a los estimadores basados en el diseño. Sin embargo, la varianza basada en el modelo puede diferir de la varianza que se obtiene por la teoría de aleatorización. En la **teoría de aleatorización o muestreo basado en el diseño,** el *diseño del muestreo* determina la forma de estimar la variabilidad del muestreo. En el **muestreo basado en el modelo,** el *modelo* determina la forma de estimar la variabilidad, y el diseño del muestreo no es importante; mientras se conserve el modelo, usted puede elegir cualesquiera n unidades de la población.

El estimador basado en el modelo

$$\hat{T}_y = \sum_{i\in S} Y_i + \hat{\beta}\sum_{i\in S} x_i$$

es insesgado con respecto al modelo, pues

$$E_M[\hat{T}_y - T] = E_M\left[\hat{\beta}\sum_{i\in S} x_i - \sum_{i\in S} Y_i\right] = 0.$$

La varianza basada en el modelo es

$$V_M[\hat{T}_y - T] = V_M\left[\hat{\beta}\sum_{i\in S} x_i - \sum_{i\in S} Y_i\right]$$

$$= V_M\left[\hat{\beta}\sum_{i\in S} x_i\right] + V_M\left[\sum_{i\in S} Y_i\right]$$

pues $\hat{\beta}$ y $\Sigma_{i\in S}Y_i$ son independientes bajo los supuestos del modelo. El modelo (3.16) no depende de las unidades de población elegidas para estar en la muestra S, de modo que S se puede considerar como fijo. En consecuencia, al usar (3.16),

$$V_M\left[\sum_{i\in S} Y_i\right] = V_M\left[\sum_{i\in S}(\beta x_i + \varepsilon_i)\right] = V_M\left[\sum_{i\in S}\varepsilon_i\right] = \sigma^2\left(\sum_{i\in S} x_i\right),$$

y, de manera análoga,

$$V_M\left[\hat{\beta}\sum_{i\in S} x_i\right] = \left(\sum_{i\in S} x_i\right)^2 V_M\left[\frac{\sum_{i\in S} Y_i}{\sum_{i\in S} x_i}\right] = \left(\sum_{i\in S} x_i\right)^2 \frac{\sigma^2}{\sum_{i\in S} x_i}.$$

Al agrupar los dos términos obtenemos

$$V_M[\hat{T}_y - T] = \frac{\sigma^2\sum_{i\in S} x_i}{\sum_{i\in S} x_i}\left(\sum_{i\in S} x_i + \sum_{i\in S} x_i\right)$$

$$= \frac{\sigma^2\sum_{i\in S} x_i}{\sum_{i\in S} x_i} t_x$$

$$= \left(1 - \frac{\sum_{i\in S} x_i}{t_x}\right)\frac{\sigma^2 t_x^2}{\sum_{i\in S} x_i}. \qquad (3.17)$$

Observe que si el tamaño de la muestra es pequeño con respecto al tamaño de la población, entonces

$$V_M[\hat{T}_y - T] \approx \frac{\sigma^2 t_x^2}{\sum_{i\in S} x_i}.$$

La cantidad $(1 - \sum_{i\in S} x_i / t_x)$ sirve como una corrección para las poblaciones finitas en el enfoque basado en el modelo de la estimación por razones.

EJEMPLO 3.9 Realicemos un análisis basado en el modelo de los datos del censo de agricultura, utilizados en los ejemplos 3.2 y 3.3. Ya hemos graficado los datos en la figura 3.1 y parecía que una línea recta por el origen se ajustaba bien y que la variabilidad con respecto a la recta era mayor para las observaciones con valores mayores de x. Para los puntos dato con x positivo, podemos realizar un análisis de regresión en SAS o S-PLUS sin ordenada al origen y con la variable de peso $1/x$. En SAS, agregamos dos renglones al final del archivo de datos para obtener los valores previstos, como se muestra en el apéndice E.

Modelo: MODELO1
NOTA: No hay ordenada al origen en el modelo. Se define R-cuadrada.
Variable dependiente: ACRES92

Análisis de varianza

Fuente	Grados de libertad	Suma de cuadrados	Media cuadrada	Valor F	Prob>F
Modelo	1	88168461.147	88168461.147	41487.306	0.0001
Error	298	633306.99655	2125.19126		
U. Total	299	88801768.143			

ECM Raíz	46.09980	R-cuadrada	0.9929
Dep media	38097.06433	R-cuadrada adjunta	0.9928
CV	0.12101		

Estimaciones de parámetros

Variable	Grados de libertad	Estimación del parámetro	Error estándar	T para H0: Parámetro=0	Prob>|T|
ACRES87	1	0.986565	0.00484360	203.684	0.0001

(*La salida continúa en la página 84*)

Observación	Peso	Variable dependiente: ACRES92	Valor predicho	Error estándar de la predicción	Media superior 95%	Media inferior 95%	Residuo
1	5.577E-6	175209	176902	868.511	175193	178611	-1693.0
2	6.892E-6	138135	143155	702.826	141771	144538	-5019.6
3	0.000017	56102.0	59056.8	289.943	58486.2	59627.4	-2954.8
4	4.535E-6	199117	217563	1068.140	215461	219665	-18446.3
5	9.471E-6	89228.0	104167	511.416	103161	105174	-14939.5
6	8.296E-6	96194.0	118923	583.857	117774	120072	-22728.5
7	0.000015	57253.0	65414.2	321.155	64782.2	66046.2	-8161.2
8	4.472E-6	210692	220590	1083.000	218459	222721	-9898.1
9	0.000012	78498.0	79188.6	388.781	78423.5	79953.7	-690.6
10	4.262E-6	219444	231453	1136.333	229217	233689	-12009.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
299	0.000064	15650.0	15504.0	76.122	15355.1	15654.7	145.1
300	0.000018	55827.0	54887.0	269.474	54357.2	55417.9	939.4
301	0	.	309134	1517.709	306147	312120	.
302	0	.	9.5151E8	4671509	9.4232E8	9.6071E8	.

La pendiente 0.986565 y la estimación del total basada en el modelo, 9.5151×10^8, son iguales a las estimaciones que se fundamentan en el diseño obtenidas en el ejemplo 3.2. El error estándar basado en el modelo para el total estimado, al utilizar (3.17), es igual a:

$$\sqrt{\hat{\sigma}^2 \frac{t_x - \sum_{i \in S} x_i}{\sum_{i \in S} x_i} t_x}.$$

Podemos emplear los residuos ponderados (para x_i distinto de cero)

$$r_i = \frac{y_i - \hat{\beta} x_i}{\sqrt{x_i}}$$

para estimar σ^2: Si se cumplen los supuestos del modelo, $\hat{\sigma}^2 = \sum r_i^2/(n-1)$ (dada como el ECM en la tabla de análisis de varianza de SAS) estima a σ^2. Así,

$$\mathrm{EE}_M[\hat{T}_y] = \sqrt{(2125.19126)\left(\frac{964,470,625 - 90,586,117}{90,586,117}\right)(964,470,625)}$$

$$= 4,446,719$$

Un análisis basado en el modelo es fácil si ignoramos la corrección para poblaciones finitas. Entonces el error estándar para el total estimado es el error estándar para la respuesta media cuando x es igual a t_x. Si ignoramos la corrección para poblaciones finitas, el error estándar basado en el modelo es exactamente el dado en la columna "Error estándar de la predicción" en la salida SAS (en SAS, éste es el error estándar del valor medio predicho), que es

$$\sqrt{\mathrm{ECM} \frac{t_x^2}{\sum_{i \in S} x_i}} = 4,671,509.$$

Observe que, para este ejemplo, el error estándar basado en el modelo es menor que el error estándar que calculamos al emplear la inferencia por aleatorización, que era 5,344,568.■

Al adoptar un modelo para un conjunto de datos, necesitamos verificar los supuestos del modelo. Los supuestos para cualquier modelo de regresión lineal son los siguientes:

1 El modelo es correcto.

2 La estructura de la varianza es como la dada.

3 Las observaciones son independientes.

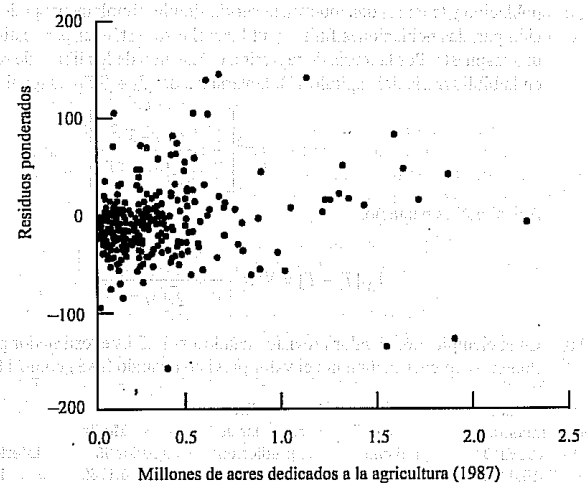
Por lo general, los supuestos 1 y 2 se verifican al graficar los datos y al examinar los residuos del modelo. Sin embargo, el supuesto 3 es difícil de verificar en la práctica y requiere cierto conocimiento de la forma en que se reunieron los datos. A menudo, si usted extrae una muestra aleatoria, entonces puede suponer la independencia de las observaciones.

Podemos realizar ciertas verificaciones acerca de lo adecuado de un modelo para estos datos con una línea recta que pase por el origen: Si la varianza de y_i con respecto a la recta es proporcional a x_i, entonces, una gráfica de los residuos ponderados es igual a:

$$\frac{y_i - \hat{\beta} x_i}{\sqrt{x_i}}$$

contra x_i o log x_i no debe exhibir tales patrones. La gráfica correspondiente a los datos del censo de agricultura aparece en la figura 3.6; que podamos observar en la gráfica nos hace dudar de lo adecuado de este modelo para las observaciones de nuestra muestra.

FIGURA 3.6
La gráfica de residuos ponderados contra x, para la muestra aleatoria del censo de agricultura. Unos cuantos condados podrían estar ausentes, aunque la dispersión parece bastante aleatoria.



3.4.2 Un modelo para la estimación por regresión

Tenemos un resultado similar para la estimación por regresión; para este caso, el modelo es

$$Yi = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

donde las ε_i son independientes e idénticamente distribuidas, con una media 0 y una constante de varianza σ^2. Los estimadores por mínimos cuadrados de β_0 y β_1 son

$$\hat{\beta}_1 = \frac{\sum_{i\in S}(x_i - \bar{x}_S)(Y_i - \bar{Y}_S)}{\sum_{i\in S}(x_i - \bar{x}_S)^2}$$

$$\hat{\beta}_0 = \bar{Y}_S - \hat{\beta}_1 \bar{x}_S.$$

Entonces, al utilizar los valores previstos, en vez de las unidades no muestreadas,

$$\begin{aligned}\hat{T}_y &= \sum_{i\in S} Y_i + \sum_{i\notin S}(\hat{\beta}_0 + \hat{\beta}_1 x_i)\\ &= n\bar{Y}_S + \sum_{i\notin S}(\hat{\beta}_0 + \hat{\beta}_1 x_i)\\ &= n(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_S) + \sum_{i\notin S}(\hat{\beta}_0 + \hat{\beta}_1 x_i)\\ &= \sum_{i=1}^{N}(\hat{\beta}_0 + \hat{\beta}_1 x_i)\\ &= N(\hat{\beta}_0 + \hat{\beta}_1 x_U).\end{aligned}$$

El estimador por regresión de T_y es, entonces, N veces el valor previsto bajo el modelo en $\bar{x}_U$.

En la práctica, si el tamaño de la muestra es pequeño con respecto al tamaño de la población y tenemos una muestra aleatoria simple, simplemente podemos ignorar la corrección para las poblaciones finitas y utilizar el error estándar para estimar el valor medio de una respuesta. Por la teoría de regresión (véase uno de los libros de regresión que aparecen en la bibliografía del capítulo 11), la varianza de $\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_U$ es igual a:

$$\sigma^2\left[\frac{1}{n} + \frac{(\bar{x}_U - \bar{x})^2}{\sum_{i\in S}(x_i - \bar{x})^2}\right].$$

Así, si n/N es pequeño,

$$V_M[\hat{T}_y - T] \approx N^2\sigma^2\left[\frac{1}{n} + \frac{(\bar{x}_U - \bar{x}_S)^2}{\sum_{i\in S}(x_i - \bar{x}_S)^2}\right]. \qquad \textbf{(3.18)}$$

EJEMPLO 3.10 En el ejemplo 3.6, el valor previsto cuando $x = 11.3$ es el estimador por regresión para $\bar{y}_U$. Podemos obtener fácilmente el valor previsto (usando SAS) como 11.9893:

Observación	Variable dependiente: CAMPO	Valor predicho	Error estándar de la predicción	Media superior 95%	Media inferior 95%	Residuo
1	15.0000	11.1920	0.491	10.1769	12.2072	3.8080

2	14.0000	12.4186	0.513	11.3205	13.5167	1.5814
3	9.0000	9.3522	0.751	7.7992	10.9052	-0.3522
⋮	⋮	⋮	⋮	⋮	⋮	⋮
24	9.0000	11.1920	0.491	10.1769	12.2072	-2.1920
25	8.0000	11.1920	0.491	10.1769	12.2072	-3.1920
26		11.9893	0.494	10.9672	13.0114	

Al sustituir las estimaciones en (3.18),

$$\begin{aligned}\mathrm{EE}_M[\hat{\bar{Y}}_{\mathrm{reg}}] &= \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(\bar{x}_U - \bar{x}_S)^2}{\sum_{i\in S}(x_i - \bar{x}_S)^2}\right]}\\ &= \sqrt{5.79\left[\frac{1}{25} + \frac{(11.3-10.6)^2}{226.006}\right]} = 0.494.\end{aligned}$$

Es fácil calcular el valor 0.494 al emplear un software estándar, pero no incorpora la corrección para las poblaciones finitas. El ejercicio 21 examina dicha corrección en la regresión basada en el modelo. ■

3.4.3 Diferencias entre las estimaciones basadas en el modelo y basadas en el diseño

¿Por qué los errores estándar no son iguales en la teoría de aleatorización? Es decir, ¿cómo es posible que haya dos varianzas diferentes para el mismo estimador? La discrepancia se debe a las distintas definiciones de la *varianza*: En el muestreo basado en el diseño, la varianza es la desviación cuadrada promedio de la estimación con respecto a su valor esperado, promediada sobre todas las muestras que se podrían obtener mediante un diseño dado. Si estamos usando un modelo, la varianza es de nuevo la desviación cuadrada promedio de la estimación con respecto a su valor esperado, pero en este caso el promedio es sobre todas las muestras posibles que se podrían generar mediante el modelo de población. Thompson (1997) analiza la inferencia al utilizar estimadores por regresión y proporciona una bibliografía para lecturas posteriores.

Si usted está absolutamente seguro de que su modelo era el correcto, podría minimizar la varianza basada en el modelo del estimador por regresión, al incluir en la muestra sólo los miembros de la población con los valores máximos y mínimos de x y excluir las unidades con valores de x entre esos extremos. Por supuesto, nadie recomendaría ese diseño en la práctica, pues nadie tiene tal certidumbre sobre un modelo. Sin embargo, el modelo no dice que usted deba considerar una muestra aleatoria simple (o cualquier otro tipo de muestra de probabilidad) o que la muestra deba ser representativa de la población, *mientras el modelo sea correcto*.

¿Qué ocurre si el modelo es incorrecto? Las estimaciones basadas en el modelo sólo son insesgadas con respecto al modelo; es decir, son insesgadas sólo dentro de la estructura de ese modelo en particular. Si el modelo es incorrecto, los estimadores basados en el modelo serán insesgados, pero, desde dentro del modelo, no necesariamente podremos decir qué tan grande es el sesgo. Así, si el modelo es incorrecto, la estimación de la varianza basada en el modelo subestimará al ECM. Al utilizar la inferencia basada en el modelo en el muestreo, usted debe tener mucho cuidado en verificar los supuestos del modelo; examinar los residuos y utilizar otras herramientas de diagnóstico. Tenga mucho cuidado con el supuesto de

independencia; ya que, por lo general, esa es la más difícil de verificar. Usted puede, ¡y debe! realizar un diagnóstico para revisar algunos supuestos del modelo para los datos de la muestra; sin embargo, usted está estableciendo un supuesto fuerte, no verificable, en el sentido de que el modelo se aplica a las unidades de población no observadas.

La estimación basada en la aleatorización del ECM se puede utilizar si algún modelo se ajusta a los datos o no, ya que la inferencia de aleatorización sólo depende de la *forma* de elegir la muestra. Pero incluso el más obstinado teórico de la aleatorización se basa en los modelos para la ausencia de respuesta y para el diseño de la encuesta. Hansen *et al.* (1983) señalan que, en general, quienes obtienen muestras según la teoría de aleatorización tienen un modelo en mente al diseñar la encuesta y toman ese modelo en cuenta para mejorar la eficiencia.

Regresaremos a este aspecto cuando revisemos el capítulo 11.

3.5 Comparación

Las estimaciones por razones y por regresión proporcionan una forma de utilizar una variable auxiliar altamente correlacionada con la variable de interés. "Sabemos" que y está correlacionada con x, y sabemos a qué distancia está $\bar{x}$ de $\bar{x}_U$, de modo que utilizamos esa información para ajustar $\bar{y}$ y (esperamos) aumentar la precisión de nuestra estimación. Los estimadores, en ambos cálculos, provienen de modelos que —esperamos— describan a los datos, pero las propiedades de los estimadores, según la teoría de la aleatorización, no dependen de estos modelos.

Como veremos en el capítulo 11, los estimadores por razones y por regresión, analizados en este capítulo, son casos particulares de un estimador por regresión generalizado. Los tres estimadores del total de la población, analizados hasta ahora —$\hat{t}_y$, $\hat{t}_{yr}$ y $\hat{t}_{yreg}$—, se pueden expresar en términos de coeficientes de regresión. Para una muestra aleatoria simple de tamaño n, los estimadores están dados en la siguiente tabla:

	Estimador	e_i
Muestra aleatoria simple	$\hat{t}_y$	$y_i - \bar{y}$
Por razones	$\hat{t}_y\left(\frac{t_x}{\hat{t}_x}\right)$	$y_i - \hat{B}x_i$
Por regresión	$N[\bar{y} + \hat{B}_1(\bar{x}_U - \bar{x})]$	$y_i - \hat{B}_0 - \hat{B}_1 x_i$

Para cada uno, la varianza estimada es:

$$N^2\left(1-\frac{n}{N}\right)\frac{s_e^2}{n}$$

para la e_i particular de la tabla; s_e^2 es la varianza muestral de las e_i.

Los estimadores por razones y por regresión dan mayor precisión que $\hat{t}_y$, cuando el valor de $\sum e_i^2$ para el método es menor que $\sum(y_i - \bar{y})^2$. La estimación por razones es de particular utilidad en el muestreo de cúmulo, como veremos en los capítulos 5 y 6.

En este capítulo, analizamos las estimaciones por razones y por regresión al usar sólo una variable auxiliar x. En la práctica, usted podría utilizar varias variables auxiliares para mejorar la precisión de las estimaciones. Los principios para la utilización de los modelos de regresión múltiple son los mismos; presentaremos la teoría para encuestas generales en la sección 11.6.

3.6 Ejercicios

1 Para cada una de las siguientes situaciones, indique cómo usaría la estimación por razones o la estimación por regresión según sea el caso.

- **a** Estime la proporción de tiempo dedicado a los deportes en las transmisiones de noticias por televisión de la ciudad donde vive.
- **b** Estime la cantidad promedio de peces atrapados por hora por los pescadores que visitan un lago en agosto.
- **c** Estime la cantidad promedio de dinero que los estudiantes de licenciatura gastan en libros de texto, en la universidad donde estudia, durante el semestre de otoño.
- **d** Estime el peso total de la carne útil (descarte los huesos, la grasa y la piel) en un embarque de pollos.

2 El conjunto de datos agsrs.dat también contiene información acerca del número de granjas en 1987 para la muestra de 300 condados. En 1987, Estados Unidos tenía un total de 2,087,759 granjas.

- **a** Grafique los datos.
- **b** Utilice la estimación por razones para calcular el número total de acres dedicados a la agricultura en 1992, al utilizar el número de granjas en 1987 como la variable auxiliar.
- **c** Repita la parte (b), al emplear la estimación por regresión.
- **d** ¿Cuál método proporciona más precisión, la estimación por razones con la variable auxiliar *acres87*, la estimación por razones con la variable auxiliar *farms87* o la estimación por regresión con la variable auxiliar *farms87*? ¿Por qué?

3 Al usar el conjunto de datos agsrs.dat, estime el número total de acres dedicados a la agricultura en 1992 para cada uno de los dos dominios siguientes: (a) condados con menos de 600 granjas y (b) condados con 600 o más granjas. Dé los errores estándar de sus estimaciones.

4 Los leñadores quieren determinar la edad promedio de los árboles que pertenecen a cierto lote. La determinación de la edad es tortuosa, pues hay que contar los anillos del árbol en un corte tomado del árbol. Sin embargo, en general, mientras mayor sea el árbol, mayor será el diámetro, y éste es fácil de medir. Los leñadores miden el diámetro de los 1132 árboles y determinan que la media de la población es igual a 10.3. Luego, se eligen al azar 20 árboles para medir su edad.

Árbol número	Diámetro, x	Edad, y	Árbol número	Diámetro, x	Edad, y
1	12.0	125	11	5.7	61
2	11.4	119	12	8.0	80
3	7.9	83	13	10.3	114
4	9.0	85	14	12.0	147
5	10.5	99	15	9.2	122
6	7.9	117	16	8.5	106
7	7.3	69	17	7.0	82
8	10.2	133	18	10.7	88
9	11.7	154	19	9.3	97
10	11.3	168	20	8.2	99

a Grafique los datos.

b Estime la edad promedio de la población de árboles que pertenecen al lote y dé un error estándar aproximado para su estimación. Etiquete su estimación sobre la gráfica. Usted eligió uno de los métodos de estimación. ¿Por qué?

5 El conjunto de datos counties.dat contiene información sobre el área, población, número de médicos, desempleo, y varias cantidades más para una muestra aleatoria simple de 100 de los 3141 condados de Estados Unidos (Oficina de Censos, 1994). El área total de Estados Unidos es de 3,536,278 millas cuadradas; la población en 1993 fue estimada en 255,077,536.

a Trace un histograma del número de médicos para los 100 condados.

b Estime el número total de médicos en Estados Unidos, junto con el error estándar, al utilizar $N\bar{y}$.

c Grafique el número de médicos contra la población para cada condado. ¿Cuál método cree que sea más adecuado para estos datos, la estimación por razones o la estimación por regresión? ¿Por qué?

d Utilice el método elegido en la parte (c) y use la variable auxiliar de población para estimar el número total de médicos en Estados Unidos, junto con el error estándar.

e El valor "verdadero" del número total de médicos en Estados Unidos es de 532,638. ¿Cuál método de estimación se acercó más?

6 Repita el ejercicio 5, con y como la población de las granjas y x como el área.

7 Repita el ejercicio 5, con y como el número de veteranos y x como la población.

8 Utilice los datos en *golfsrs.dat* para este problema. Al emplear sólo campos con 18 hoyos, estime el costo promedio por jugar 18 hoyos el fin de semana. Dé un error estándar para la estimación.

9 Para los campos de 18 hoyos en *golfsrs.dat*, grafique las tarifas de fin de semana para los 18 hoyos contra la longitud en yardas del backtee. Estime los parámetros de regresión para predecir las tarifas de fin de semana, a partir de la longitud en yardas del backtee. ¿Existe una fuerte relación entre las dos variables?

10 Utilice los datos de golfsrs.*dat* para este problema.

a Estime las tarifas medias para los días entre semana, en los que se juegan 9 hoyos, para los campos que disponen de un profesional del golf.

b Estime las tarifas medias para los días entre semana, en los que se juegan 9 hoyos, para los campos que no disponen de un profesional del golf.

c Realice una prueba de hipótesis para comparar las tarifas medias para los días entre semana para campos con un profesional con el concepto correspondiente para campos sin un profesional.

***11** Consulte la situación del ejercicio 15. Utilice un análisis basado en el modelo para estimar el número total de médicos en Estados Unidos. ¿Cuál modelo eligió? y ¿por qué? ¿Cuáles son los supuestos del modelo? ¿Cree que se satisfagan? Asegúrese de examinar las gráficas de los residuos para tener una evidencia de lo inadecuado del modelo. ¿Cómo difieren sus resultados de los obtenidos en el ejercicio 5?

***12** (Requiere de conocimientos de probabilidad). Utilice las covarianzas deducidas en el apéndice B para mostrar la fórmula (3.6).

13 Algunos libros utilizan la fórmula

$$\hat{V}[\hat{B}] = \left(1-\frac{n}{N}\right)\frac{1}{n\bar{x}_U^2}(s_y^2 - 2\hat{B}rs_xs_y + \hat{B}^2s_x^2),$$

(donde r es el coeficiente de correlación muestral de x y y para los valores en la muestra) para estimar la varianza de una razón.

a Muestre que esta fórmula es algebraicamente equivalente a (3.7).

b Sin embargo, esta fórmula, con frecuencia, no funciona tan bien como (3.7) en la práctica: Si s_x y s_y son grandes, muchos paquetes de computadora truncarán algunas de las cifras significativas, de modo que la resta será imprecisa. Para los datos del ejemplo 3.2, calcule los valores de s_y^2, s_x^2, r y $\hat{B}$. Utilice la fórmula anterior para calcular la varianza estimada de $\hat{t}_{yr}$. ¿Es exactamente el valor de (3.7)?

***14** Recuerde de la sección 2.2 que ECM = (Varianza) + (Sesgo)^2. Utilice (3.12) y otras aproximaciones de la sección 3.1, muestre que $(E[\hat{B}-B])^2$ es pequeño comparado con ECM$[\hat{B}]$, cuando n es grande.

***15** Demuestre que si consideramos las aproximaciones al ECM en (3.6) y (3.12) como exactas, entonces la varianza de $\hat{\bar{y}}_r$ a partir de la estimación por razones es, al menos, tan grande como la varianza de $\hat{\bar{y}}_{\text{reg}}$ a partir de la estimación por regresión. SUGERENCIA: Observe $V(\hat{\bar{y}}_r) - V(\hat{\bar{y}}_{\text{reg}})$ utilice las fórmulas en (3.6) y (3.12) y muestre que la diferencia no es negativa.

***16** Demuestre las ecuaciones (3.4) y (3.11).

***17** Demuestre (3.13).

***18** Sea $d_i = y_i - [\bar{y}_U + B_1(x_i - \bar{x}_U)]$. Muestre que para la estimación por regresión,

$$E[\hat{\bar{y}}_{\text{reg}} - \bar{y}_U] \approx \frac{1-\frac{n}{N}}{nS_x^2}\sum_{i=1}^{N}\frac{d_i(x_i-\bar{x}_U)^2}{N-1}.$$

Como en el ejercicio 14, muestre que $(E[\hat{\bar{y}}_{\text{reg}} - \bar{y}_U])^2$ es pequeño comparado con ECM$[\hat{\bar{y}}_{\text{reg}}]$, cuando n es grande.

***19** (Requiere un conocimiento previo de los modelos lineales). Suponga que tenemos un modelo estocástico

$$Y_i = \beta x_i + \varepsilon_i,$$

donde las ε_i son independientes con media 0 y varianza $\sigma^2 x_i$ y todas las x_i son positivas. Muestre que el estimador por mínimos cuadrados ponderados de β es $\bar{Y}/\bar{x}$ y así, podemos calcular a β al utilizar mínimos cuadrados ponderados. ¿Es el error estándar para β, proveniente de los mínimos cuadrados ponderados igual al de (3.7)?

***20** (Requiere un conocimiento previo de los modelos lineales). Suponga que el modelo en (3.16) especifíca mal la estructura de la varianza y que un modelo mejor tiene $V[\varepsilon_i] = \sigma^2$.

a ¿Cuál es el estimador por mínimos cuadrados ponderados de β si $V[\varepsilon_i] = \sigma^2$? ¿Cuál es el estimador correspondiente del total de población para y?

b Deduzca $V[\hat{T}_y - T_y]$.

c Aplique los estimadores a los datos en agrsrs.dat. ¿Cuál es la relación entre estas estimaciones con las de los ejemplos 3.2 y 3.9?

***21** La ecuación (3.18) proporciona la varianza basada en el modelo para un total de la población, si suponemos que el tamaño de la muestra es pequeño con respecto al tamaño de la población. Deduzca la varianza al incorporar la corrección para las poblaciones finitas.

22 La cantidad B utilizada en la estimación por razones se llama, a veces, el estimador razón de medias. En algunas situaciones, uno preferiría emplear un estimador media de razones: sea $b_i = y_i/x_i$ para la unidad i; entonces el estimador media de razones es:

$$\bar{b} = \frac{1}{n}\sum_{i \in S} b_i$$

con error estándar

$$\text{EE}[\bar{b}] = \sqrt{\left(1 - \frac{n}{N}\right)\frac{S_b^2}{n}}$$

de la teoría de muestras aleatorias simples.

a ¿Cree que el estimador media de razones sea adecuado para los datos del ejemplo 3.5? ¿Por qué sí? o ¿por qué no?

***b** (Requiere un conocimiento previo de modelos lineales). Muestre que $\bar{b}$ es la estimación por mínimos cuadrados ponderados de β bajo el modelo

$$Y_i = \beta x_i + \varepsilon_i$$

donde ε_i tiene media cero y varianza $\sigma^2 x_i^2$.

***23** (Requiere de una computadora para la resolución de este ejercicio).

a Genere 500 conjuntos de datos, cada uno con 30 parejas de observaciones (x_i, y_i). Utilice una distribución normal bivariada con media 0, desviación estándar 1 y correlación 0.5 para generar cada pareja (x_i, y_i). Para cada conjunto de datos, calcule $\bar{y}$ y $\hat{\bar{y}}_{\text{reg}}$ al emplear $\bar{x}_U = 0$. Grafique un histograma de los 500 valores de $\bar{y}$ y otro histograma de los 500 valores de $\hat{\bar{y}}_{\text{reg}}$. ¿Qué es lo que ve?

b Repita la parte (a) para 500 conjuntos de datos, cada uno con 60 parejas de observaciones.

24 Busque el diccionario de un idioma que haya estudiado. Elija 30 páginas de este diccionario al azar. Para cada una de estas páginas, sean:

x = es el número de palabras de la página

y = es el número de palabras de la página que usted conocía (¡Sea honesto!).

¿Cuántas palabras estima que haya en el diccionario? ¿Cuántas estima que conoce? ¿Qué porcentaje de palabras conoce? Dé errores estándar para todas las estimaciones.

4

Muestreo estratificado

Una de las cosas que ella (mamá) me enseñó debería ser evidente para todos, aunque muchos cocineros aún no se dan cuenta de ello. Prepare primero la comida que tarde más en cocinarse.

—Pearl Bailey, *La cocina de Pearl*

4.1

¿Qué es el muestreo estratificado?

Con frecuencia, tenemos información adicional que nos ayuda a diseñar nuestra muestra. Por ejemplo, antes de realizar una encuesta sabemos que los hombres, generalmente comen más que las mujeres; que los residentes de Nueva York pagan más alquiler que los residentes de Des Moines o que los residentes rurales van a la tienda más a menudo que los residentes urbanos.

Si la variable que nos interesa asume distintos valores promedio en diferentes subpoblaciones, podríamos obtener estimaciones más precisas de las cantidades de la población al tomar una **muestra aleatoria estratificada**. La palabra *estratificar* proviene de la palabra latina que significa "formar capas"; dividimos a la población en H subpoblaciones, llamadas **estratos**. Los estratos no se traslapan y conforman la población completa, de modo que cada unidad de muestreo pertenece exactamente un estrato. Extraemos una muestra independiente de cada estrato y, posteriormente, reunimos la información para obtener las estimaciones globales de la población.

Utilizamos el muestreo estratificado por una o más de las siguientes razones:

1 Queremos protegernos contra la posibilidad de obtener una mala muestra. Al extraer una muestra aleatoria simple, de tamaño 100, proveniente de una población de 1000 estudiantes varones y 1000 mujeres, es posible, desde un punto de vista teórico, obtener una muestra con pocos o ningún hombre, aunque no es probable que tal muestra ocurra. La mayoría de las personas no considerarían dicha muestra como representativa de la población y se preocuparían por las posibles respuestas que hombres y mujeres manifestarían en torno al tema en cuestión. En una muestra estratificada, uno podría extraer una muestra aleatoria simple de 50 hombres y una muestra aleatoria simple independiente de 50 mujeres, así se garantizaría que la proporción de hombres en la muestra fuese la misma que en la población. Con este diseño no se puede elegir una muestra con pocos o ningún hombre.

c Aplique los estimadores a los datos en agrsrs.dat. ¿Cuál es la relación entre estas estimaciones con las de los ejemplos 3.2 y 3.9?

***21** La ecuación (3.18) proporciona la varianza basada en el modelo para un total de la población, si suponemos que el tamaño de la muestra es pequeño con respecto al tamaño de la población. Deduzca la varianza al incorporar la corrección para las poblaciones finitas.

22 La cantidad B utilizada en la estimación por razones se llama, a veces, el estimador razón de medias. En algunas situaciones, uno preferiría emplear un estimador media de razones: sea $b_i = y_i/x_i$ para la unidad i; entonces el estimador media de razones es:

$$\bar{b} = \frac{1}{n}\sum_{i \in S} b_i$$

con error estándar

$$\text{EE}[\bar{b}] = \sqrt{\left(1 - \frac{n}{N}\right)\frac{S_b^2}{n}}$$

de la teoría de muestras aleatorias simples.

a ¿Cree que el estimador media de razones sea adecuado para los datos del ejemplo 3.5? ¿Por qué sí? o ¿por qué no?

***b** (Requiere un conocimiento previo de modelos lineales). Muestre que $\bar{b}$ es la estimación por mínimos cuadrados ponderados de β bajo el modelo

$$Y_i = \beta x_i + \varepsilon_i$$

donde ε_i tiene media cero y varianza $\sigma^2 x_i^2$.

***23** (Requiere de una computadora para la resolución de este ejercicio).

a Genere 500 conjuntos de datos, cada uno con 30 parejas de observaciones (x_i, y_i). Utilice una distribución normal bivariada con media 0, desviación estándar 1 y correlación 0.5 para generar cada pareja (x_i, y_i). Para cada conjunto de datos, calcule $\bar{y}$ y $\hat{\bar{y}}_{\text{reg}}$ al emplear $\bar{x}_U = 0$. Grafique un histograma de los 500 valores de $\bar{y}$ y otro histograma de los 500 valores de $\hat{\bar{y}}_{\text{reg}}$. ¿Qué es lo que ve?

b Repita la parte (a) para 500 conjuntos de datos, cada uno con 60 parejas de observaciones.

24 Busque el diccionario de un idioma que haya estudiado. Elija 30 páginas de este diccionario al azar. Para cada una de estas páginas, sean:

x = es el número de palabras de la página

y = es el número de palabras de la página que usted conocía (¡Sea honesto!).

¿Cuántas palabras estima que haya en el diccionario? ¿Cuántas estima que conoce? ¿Qué porcentaje de palabras conoce? Dé errores estándar para todas las estimaciones.

4

Muestreo estratificado

Una de las cosas que ella (mamá) me enseñó debería ser evidente para todos, aunque muchos cocineros aún no se dan cuenta de ello. Prepare primero la comida que tarde más en cocinarse.

—Pearl Bailey, *La cocina de Pearl*

4.1 ¿Qué es el muestreo estratificado?

Con frecuencia, tenemos información adicional que nos ayuda a diseñar nuestra muestra. Por ejemplo, antes de realizar una encuesta sabemos que los hombres, generalmente comen más que las mujeres; que los residentes de Nueva York pagan más alquiler que los residentes de Des Moines o que los residentes rurales van a la tienda más a menudo que los residentes urbanos.

Si la variable que nos interesa asume distintos valores promedio en diferentes subpoblaciones, podríamos obtener estimaciones más precisas de las cantidades de la población al tomar una **muestra aleatoria estratificada**. La palabra *estratificar* proviene de la palabra latina que significa "formar capas"; dividimos a la población en H subpoblaciones, llamadas **estratos**. Los estratos no se traslapan y conforman la población completa, de modo que cada unidad de muestreo pertenece exactamente un estrato. Extraemos una muestra independiente de cada estrato y, posteriormente, reunimos la información para obtener las estimaciones globales de la población.

Utilizamos el muestreo estratificado por una o más de las siguientes razones:

1 Queremos protegernos contra la posibilidad de obtener una mala muestra. Al extraer una muestra aleatoria simple, de tamaño 100, proveniente de una población de 1000 estudiantes varones y 1000 mujeres, es posible, desde un punto de vista teórico, obtener una muestra con pocos o ningún hombre, aunque no es probable que tal muestra ocurra. La mayoría de las personas no considerarían dicha muestra como representativa de la población y se preocuparían por las posibles respuestas que hombres y mujeres manifestarían en torno al tema en cuestión. En una muestra estratificada, uno podría extraer una muestra aleatoria simple de 50 hombres y una muestra aleatoria simple independiente de 50 mujeres, así se garantizaría que la proporción de hombres en la muestra fuese la misma que en la población. Con este diseño no se puede elegir una muestra con pocos o ningún hombre.

2 Es probable que queramos datos de precisión conocida sobre los subgrupos. Estos subgrupos serían los estratos, los cuales coinciden, entonces, con los dominios de estudio. McIlwee y Robinson (1992) obtuvieron una muestra de estudiantes graduados en programas de ingeniería mecánica y eléctrica, de las universidades públicas del sur de California. Querían comparar las experiencias educativas y de campo de trabajo de los graduados, hombres y mujeres, de modo que estratificaron el marco de muestreo por género y tomaron muestras aleatorias por separado de graduados, tanto de hombres como de mujeres. Como había muchos más hombres que mujeres, obtuvieron una muestra con mayor proporción de mujeres que de hombres para obtener precisiones comparables en ambos grupos.

3 Una muestra estratificada podría administrarse, de manera más conveniente, a un menor costo. Por ejemplo, se pueden utilizar distintos esquemas de muestreo para diversos estratos. En un estudio comercial, se puede usar una encuesta por correo para empresas grandes y una entrevista personal o telefónica para empresas pequeñas. En otras encuestas, se pueden utilizar distintos esquemas de muestreo en los estratos urbanos o rurales.

4 El muestreo estratificado, si se hace correctamente, dará estimaciones más precisas (con menor varianza) para toda la población. Las personas de distintas edades tienden a tener distintas presiones sanguíneas, de modo que en un estudio de presión sanguínea, sería útil estratificar por grupos de edad. Si se estudia la concentración de plantas en un área, se puede estratificar por tipo de suelo; los pantanos tendrán plantas distintas a las que poseen los bosques. La estratificación permite reducir la varianza, pues es frecuente que esta última en cada estrato sea menor que la varianza en toda la población. Un conocimiento anterior puede servir para ahorrar dinero en el procedimiento de muestreo.

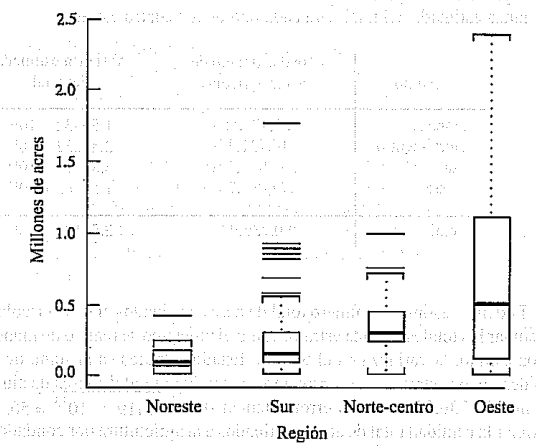
EJEMPLO 4.1 Consulte el ejemplo 2.4, donde obtuvimos una muestra aleatoria simple para estimar la cantidad promedio de acres agrícolas por condado. En dicho ejemplo observamos que, aunque la muestra haya sido obtenida de manera escrupulosa, algunas áreas quedarán representadas en exceso y otras no quedarán representadas. Con una muestra estratificada se obtiene cierto equilibrio sobre la variable de estratificación.

La muestra aleatoria simple del ejemplo 2.4 exhibió un amplio rango de valores para y_i, el número de acres dedicados a la agricultura en el condado i en 1992. Usted podría conjeturar que parte de la gran variabilidad surge debido a que los condados en el oeste de Estados Unidos son mayores, y con ello tienden a tener mayores valores de y, en relación con los condados del este de la Unión Americana.

Para este ejemplo, utilizamos las cuatro regiones del censo de Estados Unidos (noreste, norte-centro, sur y oeste) como estratos. La muestra aleatoria simple del ejemplo 2.4 realizó una muestra de cerca del 10% de la población; para comparar los resultados de la muestra estratificada con la muestra aleatoria simple, también obtenemos una muestra de cerca del 10% de los condados de cada estrato (más adelante analizaremos otros diseños de muestreo estratificado).

Estrato	Número de condados en el estrato	Número de condados en la muestra
Noreste	220	21
Norte-centro	1054	103
Sur	1382	135
Oeste	422	41
Total	3078	300

FIGURA 4.1
Gráfica de bloques para los datos del ejemplo 4.1. La línea gruesa para cada región es la mediana de los datos de la muestra para cada región; las otras líneas horizontales en las cajas son los percentiles números 25 y 75. La región noreste tiene una mediana relativamente baja y una varianza pequeña; la región oeste, por el contrario, tiene una mediana y una varianza mayores. La distribución de acres agrícolas parece tener un sesgo positivo en cada una de las regiones.



Elegimos cuatro muestras aleatorias simples, una para cada uno de los cuatro estratos. Para escoger la muestra aleatoria simple del estrato del noreste, numeramos los condados en dicho estrato, del 1 al 220 y elegimos 21 números de manera aleatoria en {1,..., 220}. Seguimos un procedimiento similar para los otros tres estratos, al seleccionar 103 condados al azar de los 1054 de la región norte-centro, 135 condados de los 1382 que existen en el sur y 41 condados de los 422 que hay en el oeste. Las cuatro muestras aleatorias simples son independientes: aunque sepamos cuáles son los condados que están en la muestra del noreste, esto no nos dice nada sobre cuáles son los condados que están en la muestra del sur.

Los datos reunidos en la muestra de los cuatro estratos aparecen en el archivo de datos agstrat.dat. La figura 4.1 enseña una gráfica de bloques para los datos de cada estrato. Los resúmenes estadísticos para cada estrato son los siguientes:

Región	Tamaño de la muestra	Promedio	Varianza
Noreste	21	97,629.8	7,647,472,708
Norte-centro	103	300,504.2	29,618,183,543
Sur	135	211,315.0	53,587,487,856
Oeste	41	662,295.5	396,185,950,266

Como tomamos una muestra aleatoria simple en cada estrato, podemos utilizar las ecuaciones (2.12) y (2.14) para estimar las cantidades de la población para cada estrato. Utilizamos

$$(220)\,(97{,}629.81) = 21{,}478{,}558.2$$

para estimar el número total de acres dedicados a la agricultura en el noreste, con varianza estimada de:

$$(220)^2\left(1-\frac{21}{220}\right)\frac{7,647,472,708}{21}=1.594316\times10^{13}.$$

La siguiente tabla proporciona las estimaciones del número total de acres agrícolas y la varianza estimada del total para cada uno de los cuatro estratos:

Estrato	Total estimado de acres agrícolas	Varianza estimada del total
Noreste	21,478,558	1.59432×10^{13}
Norte-centro	316,731,379	2.88232×10^{14}
Sur	292,037,391	6.84076×10^{14}
Oeste	279,488,706	1.55365×10^{15}
Total	90,736,034	2.5419×10^{15}

Podemos estimar el número total de acres destinados al uso agrícola en Estados Unidos, al sumar los totales de cada estrato; como el muestreo se realizó de manera independiente en cada estrato, la varianza en el total de Estados Unidos es la suma de las varianzas de los totales de los estratos. Así, estimamos la cantidad total de acres destinados a la agricultura como 909,736,034, con un error estándar de $\sqrt{2.5419 \times 10^{15}} = 50,417,248$. Queremos estimar la cantidad total de acres destinados a la agricultura por condado como 909,736,034/3078 = 295,560.7649, con un error estándar de 50,417,248/3078 = 16,379.87.

Para comparar, la estimación del total en el ejemplo 2.4, al usar una muestra aleatoria simple de tamaño 300, fue igual a 916,927,110, con un error estándar de 58,169,381. Para este ejemplo, el muestreo estratificado garantiza que cada región de Estados Unidos queda representada en la muestra y produce una estimación con un error estándar ligeramente menor que una muestra aleatoria simple con el mismo número de observaciones. La varianza de la muestra en el ejemplo 2.4 fue $s^2 = 1.1872 \times 10^{11}$. Sólo el oeste tuvo una varianza muestral mayor que s^2; la varianza muestral en el noreste fue de sólo 7.647×10^9.

Las observaciones dentro de varios estratos tienden a ser más homogéneas que las observaciones que existen en la población total y la reducción de la varianza en los estratos individuales conduce, a menudo, a una varianza reducida para las estimaciones de la población. En este ejemplo, la ganancia relativa de la estratificación se puede estimar mediante el cociente:

$$\frac{\text{varianza estimada de la estratificación, con } n=300}{\text{varianza estimada de la muestra aleatoria simple con } n=300}=\frac{2.5419\times10^{15}}{3.3837\times10^{15}}=0.75.$$

Si estas cifras fuesen las varianzas de la población, sería de esperar que sólo se necesitarán (300)(0.75) = 225 observaciones con una muestra estratificada para obtener la misma precisión que con una muestra aleatoria simple de 300 observaciones.

Por supuesto, ninguna ley afirma que usted debe realizar una muestra con la misma proporción de observaciones en cada estrato. En este ejemplo, hay más variabilidad de un condado a otro en la región del oeste; si los acres destinados a la agricultura fuesen la variable de interés primaria, usted podría reducir la varianza del total estimado aún más, si tomara una proporción de muestreo en la región del oeste mayor que en las demás regiones. Usted explorará un diseño de muestreo alternativo en el ejercicio 12. ■

4.2 Teoría del muestreo estratificado

Dividimos la población de N unidades de muestreo en H "capas" o estratos, con N_h unidades de muestreo en el estrato h. Para que funcione el muestreo estratificado, tenemos que conocer los valores de $N_1, N_2, \ldots, N_h$ y debemos tener

$$N_1 + N_2 + \cdots + N_H = N,$$

donde N es el número total de unidades en toda la población.

En el **muestreo aleatorio estratificado**, que es la forma más sencilla de muestreo estratificado, tomamos una muestra aleatoria simple de manera independiente en cada estrato, de modo que elijamos al azar n_h observaciones de las unidades de población en el estrato h. Definimos $\mathcal{S}_h$ como el conjunto de n_h unidades en la muestra aleatoria simple para el estrato h.

Notación para la estratificación Las cantidades de la población son:

y_{hj} es el valor de la unidad j en el estrato h

$t_h = \sum_{j=1}^{N_h} y_{hj}$ = es el total de la población en el estrato h

$t = \sum_{h=1}^{H} t_h$ = es el total de la población

$\bar{y}_{hU} = \dfrac{\sum_{j=1}^{N_h} y_{hj}}{N_h}$ = es la media de la población en el estrato h

$\bar{y}_U = \dfrac{t}{N} = \dfrac{\sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj}}{N}$ = es la media global de la población

$S_h^2 = \sum_{j=1}^{N_h} \dfrac{(y_{hj} - \bar{y}_{hU})^2}{N_h - 1}$ = es la varianza de la población en el estrato h

Las cantidades correspondientes para la muestra, al utilizar las estimaciones de la muestra aleatoria simple dentro de cada estrato, son:

$$\bar{y}_h = \frac{\sum_{j\in S_h} y_{hj}}{n_h}$$

$$\hat{t}_h = \frac{N_h}{n_h}\sum_{j\in S_h} y_{hj} = N_h\bar{y}_h$$

$$s_h^2 = \sum_{j\in S_h} \frac{(y_{hj} - \bar{y}_h)^2}{n_h - 1}$$

Suponga que sólo extraemos una muestra del estrato h. Ahí tenemos una población de N_h unidades y obtenemos una muestra aleatoria simple de n_h unidades. Entonces estimamos $\bar{y}_{hU}$ mediante $\bar{y}_h$, y t_h mediante $\hat{t}_h = N_h\bar{y}_h$. El total de la población es $t = \sum_{h=1}^{H} t_h$, de modo que estimamos t mediante la expresión siguiente:

$$\hat{t}_{\text{est}} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} N_h\bar{y}_h. \tag{4.1}$$

Entonces, para estimar $\bar{y}_U$, usamos:

$$\bar{y}_{\text{est}} = \frac{\hat{t}_{\text{est}}}{N} = \sum_{h=1}^{H} \frac{N_h}{N}\bar{y}_h. \tag{4.2}$$

Éste es un promedio ponderado de los promedios de los estratos de la muestra; los pesos son los tamaños relativos de los estratos. Para usar el muestreo estratificado, debemos conocer los tamaños o los tamaños relativos de los estratos.

Las propiedades de estos estimadores son una consecuencia directa de las propiedades de los estimadores en una muestra aleatoria simple:

- **Insesgabilidad.** $\bar{y}_{\text{str}}$ y $\hat{t}_{str}$ son estimadores insesgados de $\bar{y}_U$ y t. Esto es cierto, pues

$$E\left[\sum_{h=1}^{H} \frac{N_h}{N}\bar{y}_h\right] = \sum_{h=1}^{H} \frac{N_h}{N}E[\bar{y}_h] = \sum_{h=1}^{H} \frac{N_h}{N}\bar{y}_{hU} = \bar{y}_U.$$

- **Varianza de los estimadores.** Como realizamos muestras independientes de los estratos y conocemos $V(\hat{t}_h)$ de la teoría de muestras aleatorias simples, las propiedades del valor esperado (página 427) y la ecuación (2.13) implican que

$$V(\hat{t}_{\text{est}}) = \sum_{h=1}^{H} V(\hat{t}_h) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{S_h^2}{n_h}. \tag{4.3}$$

- **Estimaciones de la varianza para muestras estratificadas.** Podemos obtener un estimador insesgado de $V(\hat{t}_{\text{str}})$ al sustituir las estimaciones muestrales s_h^2 por las cantidades poblacionales S_h^2. Observe que, para estimar las varianzas, necesitamos una muestra de al menos dos unidades de cada estrato:

$$V(\hat{t}_{\text{est}}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_h^2}{n_h}, \tag{4.4}$$

$$\hat{V}(\bar{y}_{\text{est}}) = \frac{1}{N^2}\hat{V}(\hat{t}_{\text{est}}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}. \tag{4.5}$$

Como siempre, el error estándar de un estimador es la raíz cuadrada de la varianza del estimador: $\text{EE}(\bar{y}_{\text{est}}) = \sqrt{\hat{V}(\bar{y}_{\text{est}})}$.

- **Intervalos de confianza para muestras estratificadas.** Si (1) los tamaños de las muestras dentro de cada estrato son grandes o (2) el diseño del muestreo tiene una gran cantidad de estratos, un intervalo de confianza aproximado del $100(1-\alpha)\%$ para la media es:

$$\bar{y}_{\text{est}} \pm z_{\alpha/2}\text{EE}(\bar{y}_{\text{est}}).$$

El teorema del límite central, utilizado para construir este intervalo de confianza, aparece en Krewski y Rao (1981). Algunos investigadores emplean el percentil de una distribución t con $n-H$ grados de libertad en vez del percentil de la distribución normal.

EJEMPLO 4.2 Siniff y Skoog (1964) utilizaron el muestreo aleatorio estratificado para estimar el tamaño de la manada Nelchina del caribú de Alaska, en febrero de 1962. En enero y a principios de febrero, del año citado, se realizaron pruebas de campo con varias técnicas de muestreo. Dichas pruebas indicaron a los investigadores que algunas unidades de muestreo propuestas, como "igual tiempo de vuelo", eran difíciles de establecer en la práctica y que una unidad de muestreo conocida como "igual área" de 4 millas cuadradas serviría para este estudio. Los biólogos emplearon las estimaciones preliminares de las densidades del caribú para dividir el área de interés en seis estratos; cada estrato se dividía, entonces, en una retícula de unidades de muestreo de 4 millas cuadradas. Por ejemplo, el estrato A contenía $N_1 = 200$ unidades de muestreo; $n_1 = 98$ de éstas se eligieron de manera aleatoria para entrar en el estudio. Se obtuvieron los siguientes datos:

Estrato	N_h	n_h	$\bar{y}_h$	s_h^2
A	400	98	24.1	5,575
B	30	10	25.6	4,064
C	61	37	267.6	347,556
D	18	6	179.0	22,798
E	70	39	293.7	123,578
F	120	21	33.2	9,795

Con los datos presentados de esta forma, es fácil usar una hoja de cálculo para obtener todos las estimaciones necesarias en el muestreo estratificado. La hoja de cálculo que aparece en la tabla 4.1 muestra que el número total estimado de caribús es de 54,497, con un error estándar de 5840. Un intervalo de confianza del 95% para el número total de caribús es:

$$54{,}497 \pm 1.96(5840) = [43{,}051,\ 65{,}943].$$

TABLA 4.1
Hoja de cálculo para el ejemplo 4.2

	A	B	C	D	E	F	D
1	Estrato	N_h	n_h	$\bar{y}_h$	s_h^2	$\hat{t}_h = N_h\bar{y}_h$	$\left(1 - \frac{n_h}{N_h}\right)N_h^2 \frac{s_h^2}{n_h}$
2	A	400	98	24.1	5,575	9,640	6,872,040.82
3	B	30	10	25.6	4,064	768	243,840.00
4	C	61	37	267.6	347,556	16,324	13,751,945.51
5	D	18	6	179.0	22,798	3,222	820,728.00
6	E	70	39	293.7	123,578	20,559	6,876,006.67
7	F	120	21	33.2	9,795	3,984	5,541,171.43
8	total		211			54,497	34,105,732.43
9	Raíz cuadrada (total)						5,840.01

Por supuesto, este intervalo de confianza sólo refleja la incertidumbre debida a los errores de muestreo; si el procedimiento de campo para contar los caribús tiende a omitir la presencia de algunos de los animales, entonces, el intervalo de confianza será demasiado bajo. ▪

Muestreo estratificado para proporciones Como observamos en la sección 2.3, una proporción es una media de una variable que asume los valores 0 y 1. Para hacer inferencias sobre las proporciones, basta utilizar las ecuaciones (4.1)–(4.5), con $\bar{y}_h = \hat{p}_h$ y $s_h^2 = [n_h/(n_h-1)]\hat{p}_h(1-\hat{p}_h)$. Entonces,

$$\hat{p}_{\text{est}} = \sum_{h=1}^{H} \frac{N_h}{N}\hat{p}_h \tag{4.6}$$

y

$$\hat{V}(\hat{p}_{\text{est}}) = \sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h-1}. \tag{4.7}$$

La estimación del número total de unidades de la población que tienen una característica específica es similar a:

$$\hat{t}_{\text{est}} = \sum_{h=1}^{H} N_h\hat{p}_h.$$

Así, el número total estimado de unidades que pertenecen a la población, con la característica, es la suma de los totales estimados en cada estrato. De manera análoga, $\hat{V}(\hat{t}_{\text{est}}) = N^2\hat{V}(\hat{p}_{\text{est}})$.

EJEMPLO 4.3 El American Council of Learned Societies (ACLS) utilizó una muestra aleatoria estratificada de algunas sociedades ACLS en siete disciplinas, para estudiar los patrones de publicación y el uso de las computadoras y bibliotecas entre los estudiosos que pertenecen a una de las organizaciones que conforman el ACLS (Morton y Price 1989). Los datos aparecen en la tabla 4.2.

Si ignoramos la falta de respuesta por el momento (regresaremos a este aspecto en el ejercicio 9 del capítulo 8) y suponemos que no hay membresías duplicadas, usaremos la muestra estratificada para estimar el porcentaje y número de mujeres que contestaron la encuesta y pertenecen a las principales sociedades, dentro de esas siete disciplinas. En

TABLA 4.2
Datos de la encuesta ACLS

Disciplina	Membresía	Número enviado	Respuestas válidas	Mujeres (%)
Literatura	9,100	915	636	38
Clásicos	1,950	633	451	27
Filosofía	5,500	658	481	18
Historia	10,850	855	611	19
Lingüística	2,100	667	493	36
Ciencias políticas	5,500	833	575	13
Sociología	9,000	824	588	26
Totales	44,000	5,385	3,835	

este caso, sea N_h la cifra de membresía y n_h el número de encuestas válidas. Así,

$$\hat{p}_{\text{est}} = \sum_{h=1}^{7}\frac{N_h}{N}\hat{p}_h = \frac{9100}{44{,}000}0.38 + \cdots + \frac{9000}{44{,}000}0.26 = 0.2465$$

y

$$\text{EE}(\hat{p}_{\text{est}}) = \sqrt{\sum_{h=1}^{7}\left(1-\frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2\frac{\hat{p}_h(1-\hat{p}_h)}{n_h-1}} = 0.0071.$$

El número total estimado de mujeres que pertenecen a las sociedades es: $\hat{t}_{\text{est}} = 44{,}000 \times (0.2465) = 10{,}847$, con $\text{EE}(\hat{t}_{\text{est}}) = 44{,}000(.0071) = 312$. ▪

4.3 Pesos de muestreo

El estimador del muestreo estratificado $\hat{t}_{\text{est}}$ se puede expresar como una suma ponderada de las unidades de muestreo individuales. Usamos (4.1) para escribir

$$\hat{t}_{\text{est}} = \sum_{h=1}^{H}\sum_{j\in S_h}\frac{N_h}{n_h}y_{hj}.$$

El **peso de muestreo** $\omega_{hj} = (N_h/n_h)$ se puede pensar como el número de unidades en la población representadas por el miembro de la muestra (h, j). Si la población tiene 1600 hombres y 400 mujeres y el diseño de la muestra estratificada especifica una muestra de 200 hombres y 200 mujeres, entonces, cada hombre de la muestra tiene un peso de 8 y cada mujer tiene un peso de 2. Cada mujer de la muestra se representa a sí misma y a otra mujer que no ha sido elegida para estar en la muestra, mientras que cada hombre se representa a sí mismo y a otros 7 hombres que no están en la muestra. Observe que la probabilidad de elegir la unidad j del estrato h para estar en la muestra es $\pi_{hj} = n_h/N_h$, la fracción de muestreo en el estrato h. Así, el peso de muestreo no es más que el recíproco de la probabilidad de selección:

$$w_{hj} = \frac{1}{\pi_{hj}}. \tag{4.8}$$

La suma de los pesos de muestreo es igual al tamaño de la población N; cada unidad de la muestra "representa" cierta cantidad de unidades en la población, de modo que la muestra completa "representa" a toda la población. Esta identidad permite verificar la construcción de la variable de peso: si la suma de los pesos de la muestra que quiere obtener es distinta de N, entonces, se ha cometido un error.

La estimación estratificada del total de la población se puede, escribir entonces, como:

$$\hat{t}_{\text{est}} = \sum_{h=1}^{H}\sum_{j\in S_h} w_{hj}y_{hj}, \tag{4.9}$$

y la estimación de la media de la población como:

$$\bar{y}_{\text{est}} = \frac{\sum_{h=1}^{H}\sum_{j\in S_h} w_{hj}y_{hj}}{\sum_{h=1}^{H}\sum_{j\in S_h} w_{hj}}. \tag{4.10}$$

EJEMPLO 4.4 Para el estudio de los caribús del ejemplo 4.2, los pesos son

Estrato	N_h	n_h	w_{hj}
A	400	98	4.08
B	30	10	3.00
C	61	37	1.65
D	18	6	3.00
E	70	39	1.79
F	120	21	5.71

En el estrato A, cada unidad de muestreo de 4 millas cuadradas representa a 4.08 unidades de muestreo en el estrato (incluyéndose a sí misma); en el estrato B, una unidad de muestreo en la muestra se representa a sí misma y a otras 2 unidades de muestreo que no están en la muestra. Para estimar el total de la población, debemos construir una nueva variable de pesos. Esta variable contendrá el valor 4.08 para cada observación en el estrato A, 3.00 por cada observación en el estrato B y así sucesivamente. ■

EJEMPLO 4.5 La muestra del ejemplo 4.1 está diseñada de modo que cada condado de Estados Unidos tenga, aproximadamente, la misma probabilidad de aparecer en dicha muestra. Para estimar la cantidad total de acres destinados a la agricultura en Estados Unidos, creamos una columna en el conjunto de datos con los pesos de muestreo. La columna de los pesos contiene el valor 220/21 para los condados del estrato del noreste, 1054/103 para los condados del norte-centro, 1382/135 para los condados del sur y 422/41 para los condados del oeste. Podemos utilizar (4.9) para estimar el total de la población al formar una nueva columna, con el producto de las variables *weight* y *acres92*, para luego calcular la suma de la nueva columna. Al hacer esto, calculamos $\hat{t}_{\text{est}} = 909,736,035$, la misma estimación (salvo errores de redondeo) obtenida en el ejemplo 4.1.

La variable *weight* de la columna 17 se puede usar para estimar el total de la población para cada variable medida en la muestra. Observe, sin embargo, que usted no puede calcular el error estándar $\hat{t}_{\text{est}}$, a menos que conozca la estratificación. La ecuación (4.4) pide que usted calcule la varianza por separado dentro de cada estrato; los pesos no le indican la membresía de las observaciones de cada estrato. ■

4.4 Asignación de las observaciones en los estratos

Hasta ahora sólo hemos analizado los datos de una encuesta que alguien más ha diseñado. El diseño de la encuesta es la parte más importante del uso de una encuesta en la investigación: si la encuesta está mal diseñada, entonces ninguna cantidad de análisis nos dará la información necesaria. En esta sección analizaremos diversos métodos de asignación de las observaciones en los estratos.

4.4.1 Asignación proporcional

Si usted extrae una muestra estratificada con el fin de que ésta refleje la población con respecto a la variable de estratificación y quisiera que tal muestra fuese una versión miniatura de la población, debe utilizar la asignación proporcional en el diseño que está planeando emplear.

En la **asignación proporcional**, se llama así debido a que la cantidad de unidades en la muestra y en cada estrato es proporcional al tamaño del propio estrato, la probabilidad de selección $\pi_{hj} = n_h/N_h$ es la misma ($= n/N$) para todos los estratos; en una población de 2400 hombres y 1600 mujeres, una distribución proporcional con una muestra del 10% implica el tomar una muestra de 240 hombres y 160 mujeres. Así, la probabilidad de que un individuo sea elegido para estar en la muestra, n/N, es la misma que en una muestra aleatoria simple, pero muchas de las "malas" muestras que podrían aparecer en una muestra aleatoria simple (por ejemplo, una muestra donde las 400 personas sean hombres) no puede ocurrir en una muestra estratificada con asignación proporcional.

Si se utiliza la asignación proporcional, cada unidad de la muestra representa el mismo número de unidades de la población. En nuestro ejemplo, cada hombre de la muestra representa a 10 hombres de la población y cada mujer representa a 10 mujeres de la población. El peso de muestreo de cada unidad de la muestra es, entonces, igual a 10 y la estimación por muestreo estratificado de la media de la población es, simplemente, el promedio de todas las observaciones. Cuando cada unidad de la muestra tiene el mismo peso y representa al mismo número de unidades de la población, la muestra se llama de **autoponderación**. La muestra del ejemplo 4.1 tiene esta característica. En una muestra de este tipo, $\bar{y}_{\text{est}}$ es el promedio de todas las observaciones que existen en dicha muestra.

Cuando los estratos son bastante grandes, en general, la varianza de la población de $\bar{y}_{\text{est}}$ bajo la distribución proporcional es a lo más tan grande como la varianza de la población de $\bar{y}$, al usar el mismo número de observaciones pero reunidas en una muestra aleatoria. Esto es cierto sin importar lo ingenuo que sea el esquema de estratificación. Para ver el porqué de esto último, mostraremos las varianzas dentro de los estratos y entre los estratos, para el caso de la distribución proporcional, en una tabla de análisis de varianza para la población (tabla 4.3).

En una muestra estratificada de tamaño n, con una asignación proporcional, como $n_h/N_h = n/N$, la ecuación (4.3) implica que

$$\begin{aligned} V_{\text{prop}}(\hat{t}_{\text{est}}) &= \sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)N_h^2\frac{S_h^2}{n_h} \\ &= \left(1-\frac{n}{N}\right)\frac{N}{n}\sum_{h=1}^{H}N_hS_h^2 \\ &= \left(1-\frac{n}{N}\right)\frac{N}{n}\left(\text{SSW}+\sum_{h=1}^{H}S_h^2\right). \end{aligned}$$

TABLA 4.3
Tabla de análisis de la varianza de la población

Fuente	gl	Suma de cuadrados
Entre los estratos	$H-1$	$\text{SCE} = \sum_{h=1}^{H}\sum_{j=1}^{N_h}(\bar{y}_{hU}-\bar{y}_U)^2 = \sum_{h=1}^{H}N_h(\bar{y}_{hU}-\bar{y}_U)^2$
Dentro de los estratos	$N-H$	$\text{SCD} = \sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj}-\bar{y}_{hU})^2 = \sum_{h=1}^{H}(N_h-1)S_h^2$
Total, en torno de $\bar{y}_U$	$N-1$	$\text{SCT} = \sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj}-\bar{y}_U)^2 = (N-1)S^2$

Las sumas de los cuadrados demuestran que SCT = SCE + SCD, y

$$V_{\text{MAS}}(\hat{t}) = \left(1-\frac{n}{N}\right)N^2\frac{S^2}{n}$$
$$= \left(1-\frac{n}{N}\right)\frac{N^2}{n}\frac{\text{SCT}}{N-1}$$
$$= \left(1-\frac{n}{N}\right)\frac{N^2}{n(N-1)}(\text{SCD}+\text{SCE})$$
$$= V_{\text{prop}}(\hat{t}_{\text{est}}) + \left(1-\frac{n}{N}\right)\frac{N}{n(N-1)}\left[N(SCE) - \sum_{h=1}^{H}(N-N_h)S_h^2\right].$$

Este resultado muestra que la asignación proporcional con estratificación siempre produce una varianza menor o igual a la de la muestra aleatoria simple (MAS), *a menos que*

$$\text{SCE} < \sum_{h=1}^{H}\left(1-\frac{Nh}{N}\right)S_h^2. \tag{4.11}$$

Esto rara vez ocurre cuando las N_h son grandes; en general, los grandes tamaños de población de los estratos obligan a que $N_h(\bar{y}_{hU}-\bar{y}_U)^2 > S_h^2$. Usualmente, la varianza del estimador de t a partir de la distribución proporcional será menor que la varianza del estimador de t, a partir de una muestra aleatoria simple. Mientras menos parecidas sean las medias del estrato $\bar{y}_{hU}$, se obtendrá una mayor precisión al utilizar la asignación proporcional. Por supuesto, este resultado sólo es válido para las varianzas de la población; es posible que una estimación de la varianza, a partir de la asignación proporcional, sea mayor que la de una muestra aleatoria simple, sólo porque la muestra elegida produjo una mayor varianza muestral.

4.4.2 Asignación óptima

Si las varianzas S_h^2 son más o menos iguales a lo largo de todos los estratos la asignación proporcional, es probablemente la mejor distribución para una mayor precisión. Cuando las S_h^2 varían mucho, la **asignación óptima** puede producir un menor costo. En la práctica, cuando realizamos un muestreo de unidades con distintos tamaños, es probable que las unidades grandes sean más variables que las unidades pequeñas y que las extraigamos en una proporción mayor. Por ejemplo, si queremos obtener una muestra de empresas norteamericanas y nuestro objetivo es estimar la cantidad de comercio con Europa, la variación entre las empresas grandes será mayor que la variación entre las pequeñas. Como resultado, extraeremos un mayor porcentaje de las empresas grandes. La distribución óptima funciona mejor para las unidades de muestreo como empresas, ciudades y hospitales, que varían mucho en tamaño. También es eficaz cuando algunos estratos son más caros de muestrear que otros.

Neter (1978) informa de un estudio realizado por la compañía de ferrocarriles Chesapeake and Ohio (C&O) para determinar los ingresos obtenidos por los embarques de carga entre las distintas rutas, dado que el total de carga de un embarque que pasa por varias rutas se divide entre las diversas rutas. C&O extrajo una muestra estratificada de itinerarios, los documentos que detallan los artículos, la ruta y los cargos del embarque. Los itinerarios fueron estratificados según el cargo por el total de carga y se extrajo una muestra entre todos los itinerarios con cargos mayores de $40, mientras que se obtuvo una muestra sólo entre el 1% de los itinerarios con cobros menores a $5. La justificación de este hecho fue que había poca variabilidad en las deudas a C&O en el estrato de cargos menores, mientras que la variabilidad en el estrato de los cargos arriba de $40 era mucho mayor.

EJEMPLO 4.6 ¿Cómo se les paga a los compositores cuando sus propias obras son interpretadas? En Estados Unidos, muchos compositores están afiliados a la Sociedad Americana de Compositores, Autores y Editores (ASCAP, por sus siglas en inglés). Las redes de televisión, las estaciones locales de radio y televisión, servicios como Muzak, orquestas sinfónicas, restaurantes, clubes nocturnos y otros pagan a la ASCAP una cuota por una licencia anual, basada principalmente en el tamaño de la audiencia, la cual les permite tocar composiciones del catálogo de la ASCAP. Ésta distribuye después las regalías entre los compositores cuyas obras son ejecutadas.

En teoría, un miembro de la ASCAP debería recibir regalías cada vez que sus obras sean ejecutadas. Sin embargo, llevar a cabo un censo de cada obra ejecutada en la Unión Americana no sería práctico; para estimar la cantidad de regalías para sus miembros, la ASCAP utiliza el muestreo. De acuerdo con Dobishinski (1991) y "The ASCAP Advantage" (1992), la ASCAP se basa en los apuntes de los productores de televisión, los cuales tienen los detalles de la música utilizada en un programa, para identificar y clasificar las piezas ejecutadas en las redes de televisión y los principales canales de televisión por cable. Cada año se producen aproximadamente 60,000 horas de cintas en transmisiones de radio y los expertos identifican las composiciones musicales contenidas en esas transmisiones.

El muestreo estratificado se usa para obtener una muestra de estaciones de radio. Éstas se agrupan en estratos con base en la cuota pagada a la ASCAP, el tipo de comunidad donde se encuentra la estación y la región geográfica. Como las estaciones que tienen una mayor cuota contribuyen con más dinero para las regalías, es más probable que éstas participen en la muestra; una vez dentro de la muestra, las estaciones con mayor cuota son grabadas más que las estaciones con una cuota menor. La ASCAP utiliza así una forma de distribución óptima en el grabado: los estratos con mayores cuotas, y por tanto con mayor variabilidad en el monto de las regalías, tienen una mayor proporción de muestreo que los estratos con estaciones de radio con una cuota menor. ■

El objetivo del muestreo consiste en obtener más información a un menor costo. Una función de costo sencilla es la siguiente: sea C el costo total, c_0 los costos adicionales (como el mantenimiento) y c_h el costo de una observación en el estrato h, de modo que

$$C = c_0 + \sum_{h=1}^{H} c_h n_h. \tag{4.12}$$

Queremos distribuir las observaciones en estratos, minimizando $V(\bar{y}_{\text{est}})$ para un costo total dado C, o de manera equivalente, minimizando C para un $V(\bar{y}_{\text{est}})$. Suponga que los costos $c_1, c_2, \ldots, c_H$ son fijos. Para minimizar la varianza para un costo fijo, usamos el cálculo para mostrar que en la asignación óptima, n_h es proporcional a:

$$\frac{N_h S_h}{\sqrt{c_h}} \tag{4.13}$$

para cada h (véase el ejercicio 22). Así, el tamaño óptimo de muestra en el estrato h es:

$$n_h = \left(\frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{l=1}^{H}\frac{N_l S_l}{\sqrt{c_l}}}\right)n.$$

Así, extraemos más elementos de un estrato si se cumple con lo siguiente:

- El estrato representa una gran parte de la población.
- La varianza dentro del estrato es grande; extraemos más elementos para compensar la heterogeneidad.
- El muestreo en el estrato es poco costoso.

EJEMPLO 4.7 La **estratificación por montos** se utiliza, con frecuencia, en la contabilidad. Los montos registrados en los libros permiten estratificar la población. Si usted realiza una auditoría sobre los montos de los préstamos de una institución financiera, el estrato 1 podría constar de todos los préstamos de más de un millón, el estrato 2 podría constar de todos los préstamos entre \$500 000 y \$999 999 y así sucesivamente, hasta el menor estrato, con préstamos menores a los \$10 000. La distribución óptima es, a menudo, una estrategia eficiente para dicha estratificación: S_h^2 será mayor en los estratos con mayores montos de préstamo, de modo que la distribución óptima indicará una mayor proporción de muestreo para tales estratos. Si el objetivo de la auditoría consiste en estimar la discrepancia entre los montos auditados y las cantidades indicadas en los libros de la institución, es probable que un error en el monto registrado para uno de los préstamos de 3 millones contribuya más a la diferencia que un error en el monto registrado para uno de los préstamos de \$3000. En una encuesta como ésta, tal vez quiera utilizar el tamaño de muestra N_1 en el estrato 1, de modo que cada unidad de la población del estrato 1 tenga una probabilidad 1 de estar en la muestra. ■

Si todas las varianzas y los costos son iguales, la asignación proporcional es igual a la asignación óptima. Si conocemos las varianzas que están dentro de cada estrato y sabemos que son distintas, la distribución óptima proporciona una menor varianza para la estimación de $\bar{y}_U$ que la asignación proporcional. Pero la asignación óptima es un esquema más complejo; con frecuencia, la sencillez y la autoponderación de la asignación proporcional valen más que la varianza adicional. Además, la asignación óptima difiere para cada variable medida, mientras que la asignación proporcional sólo depende del número de unidades de población en cada estrato.

La **asignación de Neyman** es un caso particular de distribución óptima, utilizada cuando los costos de los estratos (y no las varianzas) son aproximadamente iguales. En la asignación de Neyman, n_h es proporcional a $N_h S_h$. Si las varianzas S_h^2 se especifican correctamente, la asignación de Neyman proporciona un estimador con una menor varianza que la distribución proporcional.

EJEMPLO 4.8 El estudio de los caribús, en el ejemplo 4.2, utilizó una forma de distribución óptima para determinar n_h. Antes de realizar el estudio, los investigadores obtuvieron aproximaciones de las densidades y la distribución de los caribús y después construyeron estratos, relativamente homogéneos en términos de la densidad de población. Establecieron $n = 225$ como el tamaño total de la muestra. Luego, utilizaron la cantidad estimada en cada estrato como una estimación burda de la desviación estándar, con los resultados de la tabla 4.4. El primer renglón contiene los nombres de las columnas de la hoja de cálculo y el segundo renglón contiene las fórmulas utilizadas para obtener la tabla. Los investigadores querían que la proporción de muestreo fuese de al menos 1/3 en los estratos menores, de modo que utilizaron los tamaños de muestra de la distribución óptima de la columna E como guía para determinar los tamaños de muestra que emplearon realmente, dados en la columna F.

4.4.3 Asignación para una precisión dada dentro de los estratos

A veces, usted está menos interesado en la precisión de la estimación del total o de la media de la población que en la comparación de las medias o los totales entre los distintos estratos. En ese caso, usted determinaría el tamaño de muestra necesario para los estratos individuales, al usar los criterios de la sección 2.5.

TABLA 4.4 Cantidades utilizadas para diseñar el estudio de los caribús del ejemplo 4.8

	A	B	C	D	E	F
1	Estrato	N_h	s_h	$N_h s_h$	n_h	Tamaño de la muestra
2				B*C	D*225/\$D%9	
3	A	400	3,000	1,200,000	96.26	98
4	B	30	2,000	60,000	4.81	10
5	C	61	9,000	549,000	44.04	37
6	D	18	2,000	36,000	2.89	6
7	E	70	12,000	840,000	67.38	39
8	F	120	1,000	120,000	9.63	21
9	Total	699		2,805,000	225	211

EJEMPLO 4.9 El Servicio Postal de Estados Unidos realiza, regularmente, encuestas entre sus clientes para saber qué opinan éstos sobre la calidad de su servicio. La población de los clientes residenciales se estratifica por área geográfica y se desea obtener una precisión de ±3 puntos porcentuales, con un nivel de confianza del 95%, dentro de cada área. Si no hay ausencia de respuestas, tal requisito implica el muestreo de al menos 1067 familias de cada estrato, como calculamos en el ejemplo 2.9. Tal asignación no es proporcional, pues el número de familias varía mucho de un estrato a otro, ni óptima, en el sentido de proporcionar la mayor eficacia para estimar los porcentajes para toda la población. Sin embargo, sí proporciona la precisión deseada dentro de cada estrato. ■

4.4.4 Determinación del tamaño de la muestra

Los distintos métodos para distribuir las observaciones en los estratos proporcionan los tamaños relativos de muestra n_h/n. Después de construir los estratos (véase la sección 4.5) y distribuir las observaciones en los mismos, podemos usar la ecuación (4.3) para determinar el tamaño de muestra necesario para lograr una varianza predeterminada. Como

$$V(\hat{t}_{\text{est}}) \leq \frac{1}{n}\sum_{h=1}^{H} \frac{n}{n_h} N_h^2 S_h^2 = \frac{v}{n},$$

un intervalo de confianza aproximado del 95% sería $\hat{t}_{\text{est}} \pm z_{\alpha/2}\sqrt{v/n}$ si se puede ignorar la corrección para las poblaciones finitas y si es válida la aproximación normal. Haga $n = z_{\alpha/2}^2 v/e^2$ para obtener un intervalo de confianza con una mitad de ancho e.

Este punto de vista requiere conocer los valores S_h^2. En la sección 7.5 veremos un punto de vista distinto que sirve para cualquier diseño de encuesta.

4.5 Definición de los estratos

Se podría pensar que, como el muestreo estratificado casi siempre proporciona una precisión mayor que el muestreo aleatorio simple, no habría necesidad de extraer una muestra aleatoria simple. Sin embargo, la estratificación agrega cierta complejidad a la encuesta. Esta complejidad adicional puede no valer la pena para obtener una pequeña ganancia en la precisión. Además, para realizar una muestra estratificada, necesitamos más información.

Para cada estrato debemos saber cuántos y cuáles miembros de la población pertenecen a ese estrato. En general, queremos que la estratificación sea muy eficiente o que los estratos sean subgrupos en los que estemos interesados, antes de estar dispuestos a incurrir en los gastos administrativos adicionales y la complejidad asociada con la estratificación.

Recuerde que la estratificación es más eficiente cuando las medias del estrato difieren ampliamente; en este caso, la suma de cuadrados entre los estratos es grande y la variabilidad dentro de los estratos será menor. En consecuencia, al construir los estratos queremos que las medias de los mismos sean lo más distintas posible. Lo ideal es que estratifiquemos mediante los valores de y; si nuestra encuesta pretende estimar los gastos totales en publicidad, desearíamos clasificar las empresas que gastan más en publicidad en el estrato 1, las empresas en el siguiente nivel de gastos en publicidad en el estrato 2 y así sucesivamente, hasta el último estrato con las empresas que no gastan en publicidad. El problema con este esquema es que no conocemos los gastos en publicidad de todas las empresas al diseñar la encuesta (si lo supiéramos, ¡no habría necesidad de la encuesta!). En vez de esto, queremos determinar una variable que esté íntimamente relacionada con y. Para estimar los gastos totales de las empresas en publicidad, podríamos estratificar de acuerdo al número de empleados o al tamaño de la empresa, al tipo de producto o de servicio. Para los ingresos de las granjas, podríamos usar el tamaño de la granja como variable de estratificación, pues esperamos que las granjas mayores tengan mayores ingresos.

La mayor parte de las encuestas miden más de una variable, de modo que una variable de estratificación queda relacionada con muchas características de interés. La encuesta de la población actual de la Oficina de Censos de Estados Unidos, la cual mide las características relacionadas con el empleo, estratifica las unidades primarias de muestreo por región geográfica, densidad de población, composición racial, industria principal y variables similares. En la encuesta de empleo, nómina y horas de Canadá, los establecimientos comerciales son estratificados por el tipo de industria, provincia y número estimado de empleados. Los ratings de televisión de Nielsen estratifican según la región geográfica, el tamaño del condado y la penetración del cable, entre otras variables. Si se dispone de varias variables de estratificación, use las variables asociadas con las respuestas más importantes.

El número de estratos elegidos depende de muchos factores; por ejemplo, la dificultad para construir un marco de muestreo con la información estratificada y el costo de la estratificación. Una regla general a tener en mente es la siguiente: mientras menos información exista, menos estratos utilice. Así, usted debe usar una muestra aleatoria simple si de antemano tiene poca información sobre la población objetivo.

Con frecuencia, usted puede reunir datos preliminares que servirán para estratificar el diseño que quiere elaborar. Si realiza una encuesta para estimar la cantidad de peces que existen en una región, puede utilizar las características físicas del área relacionadas con la densidad de los peces, como la profundidad, la salinidad y la temperatura del agua. O bien, puede emplear la información de encuestas de años anteriores o datos de un estudio preliminar como ayuda para construir los estratos. En esta situación, según Saville, "por lo general no es necesario diseñar un esquema de muestreo con más de 2 o 3 estratos, pues nuestro conocimiento de la distribución de los peces será más bien impreciso. Los estratos pueden tener diversos tamaños y cada estrato puede estar compuesto por diversas áreas en distintas partes del área total de la encuesta" (1977,10). En una encuesta que posea información anterior más precisa, utilizaremos más estratos; muchas encuestas son estratificadas hasta el punto de sólo observar dos unidades de muestreo en cada estrato.

Para muchas encuestas, la estratificación puede aumentar la precisión de manera drástica y por lo regular, paga con creces el esfuerzo invertido al construir los estratos. El ejemplo 4.10 describe la construcción de los estratos en una encuesta a gran escala; se trata de la Encuesta Nacional sobre Pesticidas.

EJEMPLO 4.10 Entre 1988 y 1990, la Agencia de Protección Ambiental de Estados Unidos (EPA, por sus siglas en inglés; 1990a,b) realizó una muestra en los pozos de agua potable para estimar la existencia de pesticidas y nitratos. Al diseñar la Encuesta Nacional sobre Pesticidas, los científicos de la EPA querían una muestra representativa de los pozos de agua potable existentes en Estados Unidos. En particular, querían garantizar que los pozos de la muestra tuviesen un amplio rango de niveles de uso de pesticidas y susceptibilidad a la contaminación por el subsuelo. También querían estudiar dos categorías de pozos: los *sistemas comunitarios de agua* (CWS), definidos como "sistemas de agua potable entubada, con al menos 15 conexiones y/o 25 o más residentes permanentes en el área de servicio con al menos un pozo útil para obtener agua potable" y los *pozos domésticos rurales*, "pozos de agua potable que surten unidades habitacionales ocupadas, localizadas en áreas rurales de Estados Unidos, excepto los pozos localizados en reservas federales".

Los siguientes fragmentos, que forman parte de los documentos de la EPA, describen la forma en que se eligieron los estratos de la encuesta:

> Para determinar cuántos pozos deben visitarse para reunir los datos, la EPA necesitó, primero, saber cuántos pozos de agua potable existen en Estados Unidos. Este proceso fue más sencillo de realizar con los sistemas comunitarios que con los pozos domésticos rurales, pues se tiene una lista de todos los sistemas públicos de agua, con sus direcciones, en el Sistema de Datos Federales (FRDS) mantenido por la EPA. Al usar el FRDS, la EPA estimó que había aproximadamente 51000 CWS con pozos en Estados Unidos. La EPA no tenía una lista con los pozos domésticos rurales que sirviera como base para seleccionar los pozos, como en el caso de los CWS. Al emplear los datos de la Oficina de Censos para 1980, la EPA estimó que había aproximadamente 13 millones de pozos domésticos rurales en el país, pero no se conocía a los dueños específicos ni las direcciones de tales pozos.
>
> La EPA eligió una técnica de diseño de encuestas llamada "estratificación", para garantizar que los datos de la encuesta cumplirían con sus objetivos. Esta técnica se utilizó para mejorar la precisión de las estimaciones, al seleccionar más pozos de las áreas con una actividad agrícola sustancial y una mayor susceptibilidad a la contaminación con el agua del subsuelo (vulnerabilidad). La EPA desarrolló criterios para separar a la población de los pozos comunitarios y los pozos rurales en cuatro categorías de uso de pesticidas y tres medidas relativas de vulnerabilidad. Este diseño garantiza que el rango de variabilidad nacional existente, con respecto al uso agrícola de pesticidas y la vulnerabilidad se refleje en la muestra de los pozos.
>
> La EPA identificó cinco grupos de pozos, cuya información estaba interesada en obtener. Estos subgrupos estaban constituidos por los pozos de sistemas comunitarios, en los condados que presentaban una vulnerabilidad promedio relativamente alta; los pozos domésticos rurales, en condados que presentaban una vulnerabilidad promedio relativamente alta; los pozos domésticos rurales, en condados que presentaban un alto uso de pesticidas; los pozos domésticos rurales, en condados que presentaban un alto uso de pesticidas y vulnerabilidad promedio relativamente alta y los pozos domésticos rurales en partes "sembradas y vulnerables" de condados que presentaban un alto uso de pesticidas y una vulnerabilidad promedio relativamente alta.
>
> Dos de las cuestiones de diseño más difíciles de establecer fueron: determinar cuántos pozos incluir en la encuesta y precisar el nivel de precisión a alcanzar para las estimaciones nacionales de la encuesta. Estas dos cuestiones estaban ligadas, pues generalmente se obtiene una mayor precisión al reunir más datos. La solución de estas cuestiones hubiera sido más sencilla si los diseñadores de la encuesta hubieran conocido de antemano la proporción de pozos del país que contenían pesticidas, pero la respuesta a esta pregunta era uno de los propósitos de la encuesta. Aunque se habían realizado muchos estudios estatales para algunos pesti-

cidas, no había estimaciones nacionales confiables acerca de la contaminación del agua potable. La EPA evaluó los requisitos de precisión y los costos de recolección de datos de distintas cantidades de pozos para determinar el tamaño de la encuesta que se ajustaría a los requisitos y presupuesto de la EPA.

En última instancia, los diseñadores de la encuesta eligieron los pozos para la recolección de datos, de modo que la encuesta proporcionara una probabilidad del 90% para detectar la presencia de pesticidas en los pozos CWS de la muestra, al suponer que el 0.5% de todos los pozos comunitarios del país presentarán pesticidas. El diseño de la encuesta para los pozos rurales se estructuró con distintas probabilidades de detección para los diversos subgrupos de interés, al enfatizar las áreas de subcondados "sembradas y vulnerables", donde la EPA quería obtener estimaciones muy precisas de la aparición de pesticidas. La EPA supuso que el 1% de los pozos domésticos rurales, en esas áreas, contendrían pesticidas y diseñó la encuesta de modo que hubiera 97% de probabilidad de detección en las áreas "sembradas y vulnerables" si el supuesto era correcto. La EPA concluyó que una muestra de aproximadamente 1,300 pozos (564 pozos públicos y 734 pozos privados) cumpliría con las especificaciones de precisión de la encuesta y proporcionaría una evaluación nacional representativa del número de pozos que contienen pesticidas.

Selección de pozos para la encuesta. Como no se conocía el número exacto y ubicación de los pozos domésticos rurales, la EPA eligió un diseño de muestra compuesto por varios pasos (etapas) para dichos pozos. El diseño comenzó con el muestreo de los condados y, posteriormente, caracterizó el uso de los pesticidas y la vulnerabilidad por el subsuelo para las áreas de los subcondados. Esto permitió delinear las áreas geográficas, suficientemente pequeñas para realizar el muestreo de los pozos domésticos rurales individuales. Este procedimiento no fue necesario para los pozos de los sistemas comunitarios, pues se conocía su número y ubicación.

El primer paso en la selección del pozo fue común a los casos comunitario y rural. Cada uno de los 3 137 condados (o sus equivalentes) en Estados Unidos fue caracterizado según el uso del pesticida y la vulnerabilidad, para garantizar que la variabilidad en estos dos factores quedaría reflejada en la encuesta. La EPA utilizó los datos relativos al uso de pesticidas de una fuente de investigación de mercado y la información relativa a la parte del área del condado dedicada a la producción agrícola, para clasificar el uso del pesticida agrícola en cada condado como alto, medio, bajo o poco común. La vulnerabilidad al agua del subsuelo de cada condado se estimó mediante un sistema de clasificación llamado DRASTIC, el cual evalúa siete factores: profundidad, recarga, medios acuíferos, medios de suelo, topografía, impacto de la zona no saturada, conductividad del acuífero. El modelo fue modificado para la encuesta, con el fin de evaluar la vulnerabilidad de los acuíferos a la contaminación con pesticidas y nitratos. Uno de los propósitos secundarios de la encuesta era la evaluación de la eficacia de la clasificación DRASTIC. Cada área fue evaluada y recibió una clasificación de alta, moderada o baja, con base en la información obtenida por los mapas de estudios geológicos de Estados Unidos, los mapas de análisis de suelos del Departamento de Agricultura de Estados Unidos y de otros recursos de agencias estatales, asociaciones y universidades. (1990a)

Este procedimiento separó los condados en los 12 estratos dados en la tabla 4.5.

La estratificación proporciona varias ventajas en esta encuesta. Permite obtener estimaciones más precisas de las concentraciones de pesticidas y nitratos de Estados Unidos como un todo, ya que es de esperar que los pozos dentro de un estrato sean más homogéneos que

TABLA 4.5
Estratos para la encuesta nacional sobre pesticidas

Estrato	Uso de pesticidas	Vulnerabilidad al agua del subsuelo (estimado mediante DRASTIC)	Número de condados
1	Alto	Alta	106
2	Alto	Moderada	234
3	Alto	Baja	129
4	Moderado	Alta	110
5	Moderado	Moderada	204
6	Moderado	Baja	267
7	Bajo	Alta	193
8	Bajo	Moderada	375
9	Bajo	Baja	404
10	Poco común	Alta	186
11	Poco común	Moderada	513
12	Poco común	Baja	416

FUENTE: Adaptado de la Agencia de Protección Ambiental de Estados Unidos, 1990a, 3.

la población total de pozos. La estratificación garantiza que la muestra incluya pozos para cada nivel de uso de pesticidas y vulnerabilidad y permite estimar la concentración de pesticidas con un tamaño de muestra predeterminado en cada estrato. El diseño factorial, con cuatro niveles del factor *uso de pesticidas* y tres niveles del factor *vulnerabilidad al agua del subsuelo*, permite analizar los posibles efectos de cada factor por separado, y la interacción de los factores sobre la concentración de los pesticidas. ■

4.6 Un modelo para el muestreo estratificado*

El modelo de análisis de la varianza en un sentido con efectos fijos proporciona una estructura subyacente para el muestreo estratificado. En este caso,

$$Y_{hj} = \mu_h + \varepsilon_{hj},$$

donde los ε_{hj} son independientes, con media 0 y varianza σ_h^2. Entonces, como en la sección 2.8, el estimador por mínimos cuadrados de μ_h obtenido a partir de las unidades de la muestra es el promedio de las observaciones de la muestra en el estrato h.

La variable aleatoria

$$T_h = \sum_{j=1}^{N_h} Y_{hj}$$

representa el total del estrato h, y la variable aleatoria

$$T = \sum_{h=1}^{H} T_h$$

representa el total global.

De la sección 2.8, el mejor estimador lineal insesgado de T_h es

$$\hat{T}_h = \frac{N_h}{n_h} \sum_{j \in S_h} Y_{hj}.$$

Entonces, por los resultados de la sección 2.8 para el muestreo aleatorio simple,

$$E_M[\hat{T}_h - T_h] = 0$$

y

$$E_M[(\hat{T}_h - T_h)^2] = N_h^2 \frac{\sigma_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right).$$

Como obtenemos muestras independientes en los estratos,

$$\begin{aligned} E_M[(\hat{T} - T)^2] &= E_M\left[\left\{\sum_{h=1}^{H}(\hat{T}_h - T_h)\right\}^2\right] \\ &= E_M\left[\sum_{h=1}^{H}(\hat{T}_h - T_h)^2 + \sum_{h=1}^{H}\sum_{k \neq h}(\hat{T}_h - T_h)(\hat{T}_k - T_k)\right] \\ &= E_M\left[\sum_{h=1}^{H}(\hat{T}_h - T_h)^2\right] \\ &= \sum_{h=1}^{H}\left(1 - \frac{n_h}{N_h}\right)N_h^2 \frac{\sigma_h^2}{n_h}. \end{aligned}$$

Podemos estimar la varianza teórica σ_h^2 mediante s_h^2. La adopción de este modelo produce las mismas estimaciones para t y su error estándar que en el caso de la teoría de aleatorización de las ecuaciones (4.1) y (4.4). Sin embargo, si se utiliza otro modelo, se obtienen otras estimaciones.

4.7 Estratificación a posteriori

Suponga que un marco de muestreo enumera todas las familias de un área y que usted quiere estimar la cantidad promedio gastada en comida en un mes. Una buena variable de estratificación sería el tamaño de la familia, pues es de esperar que una familia grande gaste más en comida que una familia pequeña. A partir de los datos de los censos de Estados Unidos, se conoce la distribución del tamaño de las familias en la región:

Número de personas en la familia	Porcentaje de familias
1	25.75
2	31.17
3	17.50
4	15.58
5+	10.00

Sin embargo, el marco de muestreo no incluye información acerca del tamaño de las familias, sino que sólo las enumera.

Sin más información, usted no puede diseñar un plan inteligente de muestreo estratificado. A pesar de ello, puede tomar una muestra aleatoria simple y registrar la cantidad gastada en comida, al igual que el tamaño de cada familia de la muestra. Si n, el tamaño de la muestra, es bastante grande, entonces, es probable que la muestra se parezca a una muestra estratificada con una asignación proporcional: esperaríamos que cerca del 26% de la muestra sea de familias con una persona, 31% de familias con dos personas, etcétera.

Si consideramos los grupos con distintos tamaños de familia como dominios distintos, podemos usar los métodos de la sección 3.3 para estimar la cantidad promedio gastada en provisiones para cada dominio: tomamos una muestra aleatoria simple de tamaño n. Sean $n_1, n_2, \ldots, n_H$ los números de unidades muestreadas en los diversos grupos de tamaños de familia (dominios) y $\bar{y}_1, \ldots, \bar{y}_H$ las medias muestrales para los grupos.

Después de tomar las observaciones, formamos una estimación "estratificada" de $\bar{y}_U$:

$$\bar{y}_{\text{post}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h. \tag{4.14}$$

Si (1) se conoce N_h/N, (2) n_h es razonablemente grande (≥ 30 o por el estilo) y (3) n es grande, entonces podemos utilizar la varianza para la asignación proporcional como una buena aproximación a la varianza:

$$\hat{V}(\bar{y}_{\text{post}}) \approx \left(1 - \frac{n}{N}\right)\sum_{h=1}^{H} \frac{N_h}{N} \frac{s_h^2}{n}. \tag{4.15}$$

ADVERTENCIA: la estratificación *a posteriori* puede ser *peligrosa* si se comete una intrusión de datos: usted puede obtener varianzas arbitrariamente pequeñas si elige los estratos después de ver los datos, al igual que puede obtener una significancia estadística si decide acerca de su hipótesis nula y alternativa después de observar los datos. La estratificación *a posteriori* se utiliza con más frecuencia para corregir los efectos de la ausencia de respuesta diferencial en los estratos definidos **a posteriori** (véase el capítulo 8).

4.8 Muestreo con cuotas

Muchas de las muestras que parecen aleatorias estratificadas son, en realidad, muestras con cuotas. En el **muestreo con cuotas**, la población se divide en distintas subpoblaciones, como en el muestreo aleatorio estratificado, pero con una diferencia importante: el muestreo de probabilidad no se utiliza para elegir a los individuos de la subpoblación para la muestra. En las versiones extremas del muestreo con cuotas, la elección de la unidad en la muestra está basada completamente en el criterio del entrevistador, de modo que se elige una muestra de conveniencia dentro de cada subpoblación.

En el muestreo con cuotas, se necesita un número dado (cuota) de tipos particulares de unidades de población en la muestra final. Por ejemplo, para obtener una muestra con cuotas para $n = 3000$, usted podría especificar que la muestra contenga 1000 hombres 1000 mujeres blancas, 500 hombres de color y 500 mujeres de color, pero podría no dar más

indicaciones acerca de la forma de cumplir con tales cuotas. Así, el muestreo con cuotas no es una forma del muestreo de probabilidad: no conocemos las probabilidades con que cada individuo se puede incluir en la muestra. Se utiliza con frecuencia cuando el muestreo de probabilidad no es práctico, demasiado costoso o innecesario, o cuando las personas que diseñan la muestra no conocen algo mejor.

La gran desventaja del muestreo con cuotas es que no sabemos si las unidades elegidas para la muestra exhiben un sesgo de selección. Si la selección de unidades queda totalmente en manos del entrevistador, éste podría elegir a los miembros más accesibles de la población; por ejemplo, las personas que puedan localizarse fácilmente por teléfono, familias sin perros fieros o áreas del bosque cercanas al camino. Es probable que los miembros más accesibles de una población difieran de manera sistemática de los miembros menos accesibles. Así, a diferencia del muestreo aleatorio estratificado, no podemos decir que la estimación del total de la población a partir de un muestreo con cuotas sea insesgado en un muestreo repetido, que es uno de nuestros criterios usuales de bondad en las muestras de probabilidad. De hecho, en las muestras con cuotas no podemos medir el error de muestreo a través de muestras repetidas y no podemos estimar el sesgo a partir de los datos muestrales. Como la selección de unidades queda a cargo de cada entrevistador, no podemos esperar que la repetición de la muestra arroje resultados similares. Así, cualquier persona que extraiga inferencias de una muestra con cuotas debe asumir necesariamente un enfoque basado en el modelo.

EJEMPLO 4.11 La encuesta de 1945 sobre hábitos de estudio realizado por el Book Manufacturer Institute (Link y Hopf 1946), como muchas encuestas de las décadas de 1940 y 1950 utilizó una muestra con cuotas. Algunas de las clasificaciones utilizadas para definir las clases de cuotas fueron el área, el tamaño de la ciudad, la edad, el sexo y el nivel socioeconómico; un sicólogo supervisor local de cada ciudad determinó las manzanas de la ciudad en donde los entrevistadores debían realizar las encuestas a las personas de cierto grupo socioeconómico dado. Luego, los entrevistadores podían elegir a las familias específicas por entrevistar en las manzanas ya designadas.

El procedimiento de cuotas seguido en la encuesta no produjo una muestra que reflejase las características demográficas de la población de Estados Unidos en 1945. La siguiente tabla compara el nivel educativo de los encuestados con las cifras del censo de 1940, ajustada para reflejar los cambios en la población debido a la guerra:

Distribución por niveles educativos	4000 personas entrevistadas (%)	Censo de Estados Unidos Población urbana y rural no granjera (%)
Hasta 8° grado	28	48
1-3 años de enseñanza media	18	19
4 años de enseñanza media	25	21
1-3 años de enseñanza superior	15	7
4 o más años de enseñanza superior	13	5

FUENTE: Link y Hopf 1946.

El sobremuestreo de las personas con mejor nivel de educación arroja dudas sobre muchos de las estadísticas dadas en el libro. El estudio concluyó que el 31% de los "lectores activos" (quienes habían leído al menos un libro el mes pasado) habían comprado el libro que habían leído y que el 25% de los últimos libros leídos por los lectores activos costaba $1 o menos. ¿Cree que una muestra aleatoria estratificada habría arrojado los mismos resultados? ■

Durante la elección presidencial de Estados Unidos en 1948, las principales encuestas realizadas unos cuantos días antes de la elección predecían que Dewey derrotaría fácilmente a Truman. En realidad, Truman ganó la elección. Según Mosteller *et al.* (1949), un problema de tales encuestas fue que todas usaron el muestreo con cuotas, no un método basado en la probabilidad. Este fracaso de las encuestas en 1948 llevó a muchas organizaciones de encuestas en Estados Unidos a alejarse del muestreo con cuotas, al menos durante unos cuantos años.

Muchas encuestas electorales en Gran Bretaña utilizaron muestras de probabilidad en las décadas de los sesenta y setenta. Sin embargo, las muestras de probabilidad eran mucho más caras que las muestras con cuotas y las muestras con cuotas utilizadas en la década de 1970 dieron predicciones precisas de los resultados de las elecciones, de modo que varias de las organizaciones más grandes de encuestas regresaron al muestreo con cuotas (Taylor 1995). Todas las encuestas que se equivocaron al predecir al ganador en la elección general británica de 1992, usaron métodos con cuotas para elegir a las personas que debían entrevistar en sus casas o en la calle; las clases primarias de cuotas utilizadas fueron el sexo, la edad, el nivel socioeconómico y el nivel de empleo. Aunque nunca sabremos con exactitud lo que falló en esas encuestas (véanse otras explicaciones en Crewe 1992), el uso de muestras con cuotas puede haber jugado cierto papel; si se entrevistaron personas "en la calle", es plausible que las personas de una clase de cuota que estuviesen accesibles difirieran de las personas menos accesibles.

Aunque el muestreo con cuotas no es tan bueno como el muestreo de probabilidad bajo condiciones ideales, por lo general, da mejores resultados que las muestras de conveniencia que se toman con frecuencia, debido a que al menos obliga a incluir a los miembros de los distintos grupos de cuotas. Las muestras con cuotas tienen la ventaja de ser menos caras que las muestras de probabilidad. La calidad de los datos obtenidos mediante muestras con cuotas se puede mejorar al restringir el criterio del entrevistador en la elección de las personas o familias a incluir en la muestra. Muchas organizaciones de encuestas utilizan el muestreo de pro-babilidad junto con el de cuotas; emplean el muestreo de probabilidad para elegir pequeños bloques de encuestados potenciales y luego toman una muestra con cuotas dentro de cada bloque, al usar variables como la edad, el sexo o la raza para definir las clases de las cuotas.

Una muestra con cuotas tiene un desempeño poco favorable en comparación con una muestra aleatoria estratificada, sobre todo cuando esta última no presenta una ausencia de respuestas. Cuando hay una ausencia de respuestas la comparación no es clara. El muestreo con cuotas se puede considerar como un método de sustitución para trabajar con la ausencia de respuestas como veremos en el capítulo 8: un individuo que no responde se sustituye por otra persona en la misma clase de cuota.

Como no conocemos las probabilidades con las que se extraen las unidades, debemos asumir un enfoque basado en el modelo al analizar los datos de una muestra con cuotas. El modelo que generalmente se adopta es el de la sección 4.6; dentro de cada subclase, las variables aleatorias que generan la subpoblación son independientes e idénticamente distribuidas. Tal modelo implica que cualquier selección de unidades de la subclase producirá una muestra representativa; si el modelo es válido, entonces, es probable que el muestreo con cuotas dé buenas estimaciones de la cantidad relativa a la población. Si el modelo no es válido, entonces, las estimaciones del muestreo con cuotas puede quedar demasiado sesgado.

Deville (1991, 177) argumenta que las muestras con cuotas pueden ser útiles para la investigación de mercado, cuando la organización que solicita la encuesta está consciente del modelo utilizado. Sin embargo, las personas que reúnen las estadísticas oficiales de crímenes, desempleo y otros temas que pueden ser tema de debate deben utilizar muestras de probabilidad.

EJEMPLO 4.12 Sanzo *et al.* (1993) utilizaron una combinación de muestreo aleatorio estratificado y de cuotas para estimar la frecuencia de la infección con *Coxiella burnetii* dentro del País Vasco, al norte de España. La *Coxiella burnetii* puede causar fiebre Q, lo que puede traer complicaciones como daños cardiacos y nerviosos. Una revisión de los registros de pacientes con fiebre Q de los hospitales vascos mostró que cerca de ¾ partes de las víctimas eran hombres, cerca de la mitad tenían entre 16 y 30 años de edad y había más víctimas de las áreas con baja densidad de población.

Los autores estratificaron la población objetivo según la densidad de población y, después, seleccionaron de manera aleatoria centros de atención a la salud de los tres estratos. Sin embargo, al elegir personas para una prueba de sangre, "se rechazó un enfoque probabilístico, por considerar que la tasa de rechazo a la prueba de sangre sería alta" (página 1185). En vez de esto, utilizaron el muestreo con cuotas para equilibrar la muestra por edad y género; los médicos preguntaban a los pacientes que necesitaran pruebas de laboratorio si querían participar en el estudio y así reclutaron sujetos para el estudio, hasta alcanzar los tamaños de muestra deseados en los seis grupos de cuotas para cada estrato.

Como se extrajo una muestra con cuotas en vez de una muestra de probabilidad, las personas que analizaron los datos debían establecer supuestos fuertes acerca de la representatividad de la muestra para aplicar los resultados a la población en general del País Vasco. En primer lugar, debía establecerse el supuesto de que las personas que asistían a un centro de salud por pruebas de laboratorio (la población participante en la muestra del estudio) no tenían mayor ni menor probabilidad de estar infectados que las personas que no asistían al centro. En segundo lugar, se debía suponer que las personas a quienes se les pidió participar y que participaron en el estudio son similares (en términos de la infección) a las personas de la misma clase de cuota con pruebas de laboratorio y que no participaron en el estudio. Éstos son supuestos fuertes. Los autores del artículo argumentan que estos supuestos son justificados pero, claro, no pueden demostrar que se cumplen a menos que se realicen estudios de seguimiento.

Si hubieran tomado una muestra de probabilidad de personas y no la muestra con cuotas, ellos no hubieran tenido la necesidad de establecer estos supuestos tan fuertes. Sin embargo, una muestra de probabilidad de personas sería demasiado cara comparada con el diseño de cuotas que fue utilizado; el diseño y establecimiento de una muestra de probabilidad hubiera tomado más tiempo. Con la muestra con cuotas, los autores pudieron reunir información sobre un problema de salud pública; no es claro que los resultados se puedan generalizar a toda la población, pero los datos proporcionan con rapidez mucha información acerca de la frecuencia de la infección, la cual se puede utilizar en investigaciones futuras sobre quién es probable que se infecte y por qué.

4.9 Ejercicios

1 ¿Qué variable (o variables) de estratificación utilizaría para cada una de las siguientes situaciones?

- **a** Una encuesta política para estimar el porcentaje de votantes registrados en Arizona que aprobarían el trabajo realizado actualmente por el gobernador.
- **b** Una encuesta telefónica realizada por alumnos de la universidad donde usted estudia, tiene como objetivo estimar la cantidad total de dinero que los estudiantes gastan en libros.
- **c** Una muestra de escuelas de bachillerato, tomada en la ciudad de Nueva York, para estimar el porcentaje de tales escuelas que ofrece una o más clases de programación de computadoras.
- **d** Una muestra de las bibliotecas públicas en California, que pretende estudiar la disponibilidad de los recursos de cómputo y los gastos per cápita.
- **e** Una encuesta de pescadores que visitan un lago de agua dulce, para saber cuáles son las especies preferidas de peces.
- **f** Un estudio aéreo para estimar el número de morsas, en el banco de hielo cercano a Alaska, comprendido entre los 173° este y los 154° oeste de longitud.
- **g** Una muestra del horario estelar de los programas de televisión en CBS (lunes a sábado, de 7 a 10 y domingos de 6 a 10), para estimar el número promedio de promocionales (anuncios de otros programas de la estación) por hora de transmisión.

2 El conjunto de datos agstrat.dat también contiene información acerca de otras variables. Para cada una de las siguientes cantidades, grafique los datos y estime la media de población para esa variable, junto con el error estándar. Dé un intervalo de confianza del 95% para cada estimación. Compare sus respuestas con las de la muestra aleatoria simple del ejercicio 11 del capítulo 2.

- **a** Número de acres dedicados a las granjas, 1987.
- **b** Número de granjas, 1992.
- **c** Número de granjas con 1000 acres o más, 1992.
- **d** Número de granjas con 9 acres o menos, 1992.

3 Puede obtenerse una muestra de almejas de concha dura mediante una draga. Sin embargo, las almejas no tienden a distribuirse de manera uniforme en un cuerpo de agua, pues algunas áreas proporcionan un mejor hábitat que otros. Así, es probable que la extracción de una muestra aleatoria simple produzca una varianza estimada mayor para el número de almejas del área. Russell (1972) usó el muestreo aleatorio estratificado para estimar el número total de almejas *Mercenaria mercenaria* en la bahía de Narragansett, Rhode Island. El área de interés se dividió en cuatro estratos, con base en estudios preliminares que identificaron las áreas donde abundaban las almejas. Luego se realizaron n_h dragados en el estrato h, para $h = 1, 2, 3, 4$. El número de acres de cada estrato es conocido; Russell calculó que el área barrida durante un dragado común era de 0.039 acres, es decir, 25.6 dragados barren un acre.

- **a** He aquí los resultados del estudio, realizado antes de la temporada comercial. Estime la cantidad total de bushels de almejas en el área y dé el error estándar de su estimación. SUGERENCIA: calcule primero N_h, el número de dragados necesarios para abarcar el estrato h.

Estrato	Área (acres)	Número de dragados realizados	Número promedio de bushels por dragado	Varianza muestral para el estrato
1	222.81	4	0.44	0.068
2	49.61	6	1.17	0.042
3	50.25	3	3.92	2.146
4	197.81	5	1.80	0.794

b Se realizó otra encuesta al final de la temporada comercial. En esta encuesta, los estratos 1, 2 y 3 se colapsaron en un solo estrato, llamado de ahora en adelante como estrato 1. Estime la cantidad total de bushels de almejas (con el error estándar) al final de la temporada.

Estrato	Área (acres)	Número de dragados realizados	Número promedio de bushels por dragado	Varianza muestral para el estrato
1	322.67	8	0.63	0.083
4	197.81	5	0.40	0.046

4 Regrese a la población hipotética del ejemplo 3.4. Ahora, en vez de utilizar x como variable auxiliar en la estimación de la razón, utilícela como variable de estratificación: una unidad de población está en el estrato 1 si $x \leq 5$ y en el estrato 2 si $x > 5$. Con esta estratificación, $N_1 = N_2 = 4$. La población es la siguiente:

Número de unidad	Estrato	y
1	1	1
2	1	2
3	1	4
8	1	8
4	2	4
5	2	7
6	2	7
7	2	7

Considere el diseño de muestreo estratificado en el cual $n_1 = n_2 = 2$.

a Escriba todas las muestras aleatorias simples de tamaño 2 del estrato 1 y determine la probabilidad de cada una. Repita esto para el estrato 2.

b Use el trabajo realizado en la parte (a) para hallar la distribución muestral de $\hat{t}_{est}$.

c Determine la media y la varianza de la distribución muestral de $\hat{t}_{est}$. ¿Cuál es la relación entre la media y la varianza en los ejemplos 2.1 y 3.4?

5 Suponga que una ciudad tiene 90 000 moradas, de las cuales 35 000 son casas, 45 000 son departamentos y 10 000 son condominios. Usted piensa que el uso promedio de la electricidad es casi el doble en las casas que en los departamentos o condominios y que la desviación estándar es proporcional a la media.

a ¿Cómo distribuir una muestra de 900 observaciones si desea estimar el consumo promedio de electricidad de todas las familias en la ciudad?

b Ahora suponga que quiere estimar la proporción global de las familias en donde se practica el ahorro de energía. Usted tiene fuertes razones para creer que el 45% de los moradores usan cierto tipo de ahorro de energía y que los porcentajes correspondientes son de 25% para moradores de apartamentos y 3% para residentes de condominios. ¿Cuál sería la ventaja de la distribución proporcional sobre el muestreo aleatorio simple?

c Alguien más ha realizado una pequeña encuesta, al emplear una muestra aleatoria simple, del uso de la energía en las casas. Con base en la encuesta, cada casa está clasificada con calefacción eléctrica o con algún otro tipo de calefacción. Se registra el consumo de energía, en enero, en kilovatios-hora para cada casa (y_i) y los resultados aparecen a continuación:

Tipo de calefacción	Número de casas	Media muestral	Varianza muestral
Eléctrica	24	972	202,396
No eléctrica	36	463	96,721
Total	60		

Por otros registros, se sabe que 16,450 de las 35,000 casas tienen calefacción eléctrica y que 18,550 poseen calefacción no eléctrica.

i Use la muestra y dé una estimación y el error estándar de la proporción de casas con calefacción eléctrica. ¿Incluye el intervalo de confianza del 95%, que acaba de construir, a la verdadera proporción?

ii Dé una estimación y el error estándar del número promedio de kilovatios-hora utilizados por las casas de la ciudad. ¿Qué tipo de estimador empleó? y ¿por qué eligió dicho estimador?

6 Una analista de la opinión pública tiene un presupuesto de $20 000 para realizar una encuesta. Ella sabe que el 90% de las familias tienen teléfono. Las entrevistas por teléfono cuestan $10 por familia; las entrevistas personales cuestan $30 cada una, si todas las entrevistas se realizan personalmente y $40 cada una si sólo las familias que no tienen teléfono son entrevistadas personalmente (debido a los costos adicionales de viaje). Suponga que las varianzas en los estratos con y sin teléfono son similares y que los costos fijos son $c_0 = \$5000$. Cuántas familias deben ser entrevistadas en cada estrato si:

a Todas las familias son entrevistadas personalmente.

b Las familias que cuentan con servicio telefónico son entrevistadas por teléfono y las familias sin teléfono son entrevistadas personalmente.

7 Para el ejemplo 4.3, construya un conjunto de datos con 3835 observaciones. Incluya tres columnas: la columna 1 es el número de estrato (de 1 a 7), la columna 2 contiene la variable de respuesta del género (0 para hombres y 1 para mujeres) y la columna 3 contiene los pesos del muestreo N_h/n_h para cada observación. Use las columnas 2 y 3 junto con (4.10) y calcule $\hat{p}_{est}$. ¿Es posible calcular $\text{EE}(\hat{p}_{est})$ al usar solamente las columnas 2 y 3, sin más información?

8 La encuesta del ejemplo 4.3 reunió más datos sobre los sujetos a investigación. Otra de las preguntas de la encuesta interrogaba si la persona coincidía con la siguiente afirmación: "cuando leo un ejemplar nuevo de la principal revista de mi disciplina, rara vez encuentro

un artículo que me interese". Los resultados son los siguientes:

Disciplina	Coincide (%)
Literatura	37
Clásicos	23
Filosofía	23
Historia	29
Lingüística	19
Ciencias políticas	43
Sociología	41

a ¿Cuál es la población muestreada en esta encuesta?

b Determine una estimación de la proporción de personas en la población muestreada que coincida con la afirmación y dé el error estándar de la estimación.

9 Construya una pequeña población y estratificación para la cual $V(\hat{t}_{est})$ al usar la asignación proporcional es mayor que la varianza que sería obtenida al tomar una muestra aleatoria simple con el mismo número de observaciones. SUGERENCIA: Use (4.11).

10 En el ejercicio 8 del capítulo 2 se dieron los datos acerca del número de publicaciones para una muestra aleatoria simple de 50 profesores. Sin embargo, no todos los departamentos estuvieron representados en la muestra; ésta contenía varios profesores de sicología y de química, pero ninguno de lenguas extranjeras. Los siguientes datos provienen de una muestra estratificada de profesores, al usar las áreas de ciencias biológicas, ciencias físicas, ciencias sociales y humanidades como estratos. La asignación proporcional se utilizó en esta muestra.

Estrato	Número de profesores en el estrato	Número de profesores en la muestra
Ciencias biológicas	102	7
Ciencias físicas	310	19
Ciencias sociales	217	13
Humanidades	178	11
Total	807	50

La tabla de frecuencias para el número de publicaciones en el estrato es la siguiente:

Número de publicaciones con arbitraje	Número de profesores			
	Biológicas	Físicas	Sociales	Humanidades
0	1	10	9	8
1	2	2	0	2
2	0	0	1	0
3	1	1	0	1
4	0	2	2	0
5	2	1	0	0
6	0	1	1	0
7	1	0	0	0
8	0	2	0	0

a Estime el número total de publicaciones con arbitraje realizadas por los profesores de la institución y dé el error estándar.

b ¿Cuál es la relación que existe entre el resultado que obtuvo de la parte (a) con el resultado de la muestra aleatoria simple del ejercicio 8, capítulo 2?

c Estime la proporción de los profesores sin publicaciones con arbitraje y dé el error estándar.

d ¿Aumentó la precisión en este ejemplo con la estratificación? Explique por qué piensa que sí o que no.

11 Lydersen y Ryg (1991) usaron técnicas de estratificación para estimar las poblaciones de focas con anillos en los fiordos de Svalbard. El área de estudio de 200 km^2 se dividió en tres zonas: la zona 1, el exterior de Sassenfjorden, quedó cubierta con hielo relativamente nuevo durante el periodo de estudio en marzo de 1990 y estaba ligeramente cubierta de nieve; la zona 3, Tempelfjorden, tenía una capa de hielo estable durante todo el año; la zona 2, el interior de Sassenfjorden, era intermedia entre la zona estable 3 y la zona inestable 1. Las focas con anillos necesitan un buen hielo para establecer territorios con respiraderos y la capa de hielo permite a las hembras excavar madrigueras para el nacimiento de sus crías. Así, se pensaba que las tres zonas tendrían distintas densidades de focas.

Para elegir la muestra, los investigadores dividieron toda la región en 200 áreas de 1 km^2; "se construyó una retícula de muestreo que abarcaba el 20% del área total... eligiendo 40 números entre el 1 y el 200 con el generador de números aleatorios". En cada área de la muestra, Imjak, un husky siberiano, rastreó las estructuras de las focas; se registró el número de respiraderos en cada cuadrado de la muestra. Se localizó un total de 199 respiraderos en las zonas 1-3. Los datos (reconstruidos a partir de la información dada en el artículo) están en el archivo seals.dat.

La siguiente tabla da el número de puntos, y el número de puntos muestreados en cada zona:

Zona	Número de puntos	Puntos muestreados
1	68	17
2	84	12
3	48	11
Total	200	40

a ¿Es ésta una muestra aleatoria estratificada o una muestra aleatoria simple estratificada a posteriori? Explique.

b Estime el número total de respiraderos en la región de estudio, junto con el error estándar correspondiente.

c Si usted fuese el diseñador de esta encuesta, ¿cómo distribuiría las observaciones en los estratos si el objetivo consistiese en estimar el número total de respiraderos? ¿Y si el objetivo fuese comparar la densidad de los respiraderos en las tres zonas?

12 La distribución proporcional se usó en la muestra estratificada del ejemplo 4.1. Sin embargo, se observó que la variabilidad fue mucho mayor en el oeste que en las otras regiones. Utilice las varianzas estimadas del ejemplo 4.1 y suponga que los costos de muestreo son los mismos en cada estrato para determinar una distribución óptima para una muestra estratificada de tamaño 300.

13 Elija una muestra aleatoria estratificada de tamaño 300 de los datos del archivo agpop.dat, use la distribución del ejercicio 12. Estime la cantidad total de acres dedicados a la agricul-

tura en Estados Unidos y dé el error estándar de la estimación. ¿Cuál es la relación de este error estándar con el hallado en el ejemplo 4.1?

14 Burnard (1992) envió un cuestionario a una muestra estratificada de tutores y alumnos de enfermería en Gales, para estudiar lo que los tutores y los alumnos entienden por el término *aprendizaje empírico*. El tamaño de la población y el tamaño de la muestra para cada uno de los cuatro estratos son los siguientes:

Estrato	Tamaño de la población	Tamaño de la muestra
Tutores de enfermería general (GT)	150	109
Tutores de enfermería psiquiátrica (PT)	34	26
Estudiantes de enfermería general (GS)	2680	222
Estudiantes de enfermería psiquiátrica (PS)	570	40
Total	3434	397

A los encuestados se les preguntó cuáles de las siguientes técnicas podrían ser identificadas como métodos de aprendizaje empírico; a continuación, se da el número de estudiantes y tutores de cada grupo que identificaron cada método como de aprendizaje empírico:

Método	GS	PS	PT	GT
Desempeño de papeles	213	38	26	104
Actividades para la solución de problemas	182	33	22	95
Simulaciones	95	20	22	64
Ejercicios de construcción de la empatía	89	25	20	54
Ejercicios Gestalt	24	4	5	12

Estime el porcentaje global de estudiantes y tutores de enfermería que identifica a cada una de estas técnicas como de *aprendizaje empírico*. Asegúrese de dar los errores estándar de sus estimaciones.

15 Kruuk *et al.* (1989) usó una muestra estratificada para estimar el número de guaridas de nutrias *Lutra lutra* a lo largo de la línea costera de 1400 km de Shetland, Reino Unido. La línea costera fue dividida en 242 secciones de 5 km (en 237 de ellas no predominaban los edificios) y cada sección fue asignada al estrato cuyo tipo de terreno fuese predominante. Luego, se eligieron de manera aleatoria ciertas secciones de los estratos. En cada sección elegida, los investigadores contaron el número total de guaridas en una franja de 110 m de ancho a lo largo de la costa.

Los datos están en el archivo otters.dat. Los tamaños de población para los estratos son los siguientes:

Estrato	Total de secciones	Secciones contadas
1 Acantilado mayor de 10 m	89	19
2 Agricultura	61	20
3 Ninguna de las anteriores, turba	40	22
4 Ninguna de las anteriores, no turba	47	21

a Estime el número total de guaridas de nutrias a lo largo de la costa de Shetland, junto con el error estándar de la estimación.

b Analice las posibles fuentes de sesgo en este estudio. ¿Cree que sea posible eliminar completamente los sesgos de selección y medición?

16 Las estadísticas de matrimonios y divorcios son compiladas por el Centro Nacional de Estadísticas de la Salud (NCHS, por sus siglas en inglés) y publicadas en volúmenes de *Vital Statistics of the United States*. Los funcionarios locales y estatales proporcionan al NCHS el recuento anual de matrimonios y divorcios en cada condado. Además, algunos estados envían cintas de computadora con datos adicionales o copias en microfilm de las actas de matrimonio o divorcio. Estos datos adicionales se utilizan para calcular las estadísticas de la edad de matrimonio o divorcio, el estado civil anterior de las parejas que se casan, y los hijos implicados en el divorcio. En 1987, si un estado enviaba una cinta de computadora, se incluían todos los registros en las estadísticas de divorcio; si un estado enviaba copias en microfilm, se obtenía una muestra aleatoria de una fracción determinada de las actas de divorcio y se registraban esos datos. Las tasas de muestreo (probabilidades de selección) y el número de registros que fueron muestreados en cada estado en el área de registro de divorcios para 1987, están en el archivo divorce.dat.

a ¿Cuántos divorcios hubo en el área de registro de divorcios durante 1987? SUGERENCIA: use los pesos de muestreo.

b ¿Por qué el NCHS usó distintas tasas de muestreo en diversos estados?

c Estime el número total de divorcios otorgados a hombres con 24 años o menos; a mujeres con 24 años o menos. Dé intervalos de confianza del 95% para las estimaciones que lleve acabo.

d ¿En qué proporción de todos los divorcios tenía el marido entre 40 y 49 años de edad? ¿En qué proporción tenía la mujer entre 40 y 49 años de edad? Dé intervalos de confianza para las estimaciones.

17 Jackson *et al.* (1987) compararon la precisión del muestreo sistemático y estratificado para estimar la concentración promedio de plomo y cobre en el suelo. El área de un kilómetro cuadrado se dividió en cuadrados de 100 m de lado cada uno y se recogió una muestra de suelo en cada una de las 121 intersecciones de la retícula resultante. A continuación, damos un resumen de las estadísticas de esta muestra sistemática:

Elemento	n	Promedio (mg kg^{-1})	Rango (mg kg^{-1})	Desviación estándar (mg kg^{-1})
Plomo	121	127	22-942	146
Cobre	121	35	15-90	16

Los investigadores también estratificaron a posteriori la misma región. El estrato A constaba de tierras agrícolas alejadas de los caminos, villas y bosques. El estrato B contenía áreas a menos de 50 m de los caminos y se esperaba que tuviera mayores concentraciones de plomo. El estrato C contenía los bosques, donde también se esperaba una mayor concentración de plomo, pues el follaje podría capturar partículas suspendidas. Los datos sobre la concentración del plomo y el cobre no fueron usados para determinar los estratos. Los datos

de los puntos de la retícula que caen en cada estrato están en la siguiente tabla:

Elemento	Estrato	n_h	Promedio (mg kg^{-1})	Rango (mg kg^{-1})	Desviación estándar (mg kg^{-1})
Plomo	A	82	71	22-201	28
Plomo	B	31	259	36-942	232
Plomo	C	8	189	88-308	79
Cobre	A	82	28	15-68	9
Cobre	B	31	50	22-90	18
Cobre	C	8	45	31-69	15

a Calcule un intervalo de confianza del 95% para la concentración promedio de plomo en el área, al usar la muestra sistemática (puede suponer que esta muestra se comporta como una muestra aleatoria simple). Repita el procedimiento para la concentración promedio de cobre.

b Ahora utilice la muestra estratificada a posteriori y determine intervalos de confianza del 95% para las concentraciones promedio del plomo y el cobre. ¿Cuál es su relación con los intervalos de confianza de la parte (a)? ¿Cree que el uso de la estratificación en las próximas encuestas mejoraría la precisión?

18 En el ejercicio 17, el tamaño de muestra en cada estrato fue proporcional al área del estrato. Use las desviaciones estándar de la muestra. ¿Cuál sería una asignación óptima para obtener una muestra aleatoria estratificada con 121 observaciones? ¿Es la asignación óptima igual para el cobre que para el plomo?

19 Wilk *et al.* (1977) dieron datos acerca del número y tipos de peces y datos ambientales para el área de la plataforma continental del Atlántico, entre el este de Long Island, Nueva York, y Cape May, Nueva Jersey. El área oceánica bajo estudio se dividió en estratos con base en la profundidad. Se realizó un muestreo con una tasa mayor cerca de la orilla y menor lejos de la orilla. "Se realizó una muestra en los estratos de la orilla (0-28 m) con una tasa de aproximadamente una estación por cada 515 km^2 y en los estratos lejanos a la orilla (29-366 km) se realizó una muestra con una tasa aproximada de una estación por cada 1,030 km^2" (página 1). Así, cada registro en los estratos 3-6 representa el doble de área de un registro en los estratos 1 y 2. Al calcular los promedios de peces capturados y el número de especies, podemos utilizar un peso relativo de muestreo de 1 para los estratos 1 y 2, y el peso 2 para los estratos 3-6.

Estrato	Profundidad	Peso relativo de muestreo
1	0-19	1
2	20-28	1
3	29-55	2
4	56-100	2
5	111-183	2
6	184-366	2

El archivo nybight.dat contiene los datos acerca del total de peces atrapados en las estaciones de muestreo visitadas en junio de 1974 y junio de 1975.

a Construya una gráfica de bloques adyacentes para el número de peces atrapados en las redes en junio de 1974. ¿Parece que hay una gran variación entre los estratos?

b Calcule las estimaciones del número promedio y del peso promedio de los peces atrapados en cada carga en junio de 1974, junto con el error estándar.

c Calcule las estimaciones del número promedio y del peso promedio de peces atrapados en cada carga, en junio de 1975, junto con el error estándar.

d ¿Existe alguna evidencia de que el peso promedio de los peces atrapados en cada carga sea distinto entre junio de 1974 y junio de 1975? Responda al usar una prueba de hipótesis adecuada.

20 En enero de 1995, la Oficina de Evaluación de la Universidad Estatal de Arizona encuestó a los profesores y personal universitario para determinar su reacción ante el cierre de la universidad durante el receso de invierno de 1994. Los profesores y personal de las unidades académicas cerradas durante dicho receso fueron separados en cuatro estratos, de los cuales se obtuvo la siguiente muestra:

Número de estrato	Tipo de empleado	Tamaño de la población (N_h)	Tamaño de la muestra
1	Profesor	1374	500
2	Personal clasificado	1960	653
3	Personal administrativo	252	98
4	Profesional académico	95	95

Mediante el correo del campus, se enviaron cuestionarios a las personas de los cuatro estratos; el tamaño de la muestra en la tabla anterior, es el número de cuestionarios enviados en cada estrato. Regresaremos al punto de ausencia de respuestas, y utilizaremos de nuevo esta encuesta en el capítulo 8; por el momento sólo analizaremos a quienes respondieron en la muestra estratificada de empleados en unidades cerradas; los datos para los 985 individuos que contestaron la encuesta están en el archivo winter.dat. Para este ejercicio, analice las respuestas a la pregunta "¿Quiere usted tener de nuevo un receso de invierno?" (variable *breakaga*).

a No todas las personas de la encuesta contestaron esta pregunta. Determine el número de personas que respondieron la pregunta en cada uno de los cuatro estratos. Para este ejercicio, use estos valores como los n_h.

b Use (4.6) y (4.7) para estimar la proporción de profesores y personal que contestaría sí a la pregunta "¿Quiere usted tener de nuevo un receso de invierno?" y dé el error estándar.

c Defina una nueva variable, la cual asume el valor 1 para las personas que respondieron con un sí a la pregunta y asume el valor 0 para las personas que contestaron con un no; si la persona no responde, la variable queda en blanco (si usted utiliza una hoja de cálculo) o se asigna el código de valor faltante (si usa un software de estadística). Construya una columna de pesos de muestreo N_h/n_h para las observaciones de la muestra (el peso de muestreo será cero o faltante para quienes no contesten). Ahora use (4.10) para estimar la proporción de profesores y personal que contestarían sí a la pregunta "¿Quiere usted tener de nuevo un receso de invierno?"

d Use la columna de ceros y unos construida en la pregunta anterior, para determinar s_h^2 para cada estrato, al calcular la varianza muestral de las observaciones en ese estrato. Ahora use (4.5) para calcular el error estándar de su estimación de la proporción. ¿Por qué su respuesta es la misma que calculó en la parte (b)?

e La estratificación se utiliza, en ocasiones, como un método para trabajar con la ausencia de respuesta. Calcule las tasas de respuesta (el número de personas que responden, dividido entre el número de cuestionarios enviados) para cada estrato. ¿Cuál es el estrato que tiene la menor tasa de respuesta para esta pregunta? ¿Cómo consideran la estratificación quienes no responden?

21 Se está diseñando una muestra estratificada para estimar la frecuencia p de una característica rara (digamos, la proporción de residentes en Milwaukee que tienen la enfermedad de Lyme). El estrato 1, con N_1 unidades, tiene una alta frecuencia de la característica; el estrato 2, con N_2 unidades, tiene baja frecuencia. Suponga que el costo por muestrear una unidad (por ejemplo, el costo por elegir una persona para la muestra y determinar si ella tiene la enfermedad de Lyme) es el mismo para cada estrato y que debe obtenerse una muestra de, a lo más, 2000 unidades.

a Sean p_1 y p_2 las proporciones respectivas en el estrato 1 y 2 con la característica rara. Si $p_1 = 0.10$, $p_2 = 0.03$ y $N_1/N = 0.4$, ¿cuáles son los valores de n_1 y n_2 bajo una asignación óptima?

b Si $p_1 = 0.10$, $p_2 = 0.03$ y $N_1/N = 0.4$, ¿cuál es el valor de $V(\hat{p}_{est})$ bajo una asignación proporcional? ¿Y bajo una asignación óptima? ¿Cuál es la varianza si usted toma una muestra aleatoria simple de 2000 unidades de la población?

c (Use una hoja de cálculo para esta parte del ejercicio). Ahora, fije p como 0.05. Haga variar p_1 de 0.05 a 0.50, y a N_1/N de 0.01 a 0.50 (estos dos valores determinan el valor de p_2). Para cada combinación de p_1 y N_1/N, determine la asignación óptima y la varianza bajo la asignación proporcional y bajo la asignación óptima. Además, determine la varianza de una muestra aleatoria simple de 2000 unidades. ¿En qué caso la asignación óptima proporciona un incremento sustancial en la precisión en relación con la asignación proporcional? ¿Y en relación con una muestra aleatoria simple?

***22** (Requiere de conocimientos del cálculo). Muestre que la varianza de $\hat{t}_{est}$ se minimiza para un costo fijo con la función de costo en (4.12) cuando $n_h \alpha N_h S_h / \sqrt{c_h}$, como en (4.13). SUGERENCIA: use multiplicadores de Lagrange.

23 Suponga que el Departamento de Salud de Arizona desea realizar un estudio de niños de 2 años de edad, cuyas familias reciben atención médica, para determinar la proporción que ha sido vacunada. La atención médica es proporcionada por varias organizaciones y el estado tiene 15 condados. La tabla 4.6 muestra el número de la población de niños de 2 años de edad para cada combinación condado/organización. La muestra será estratificada por condado y organización. Se desea elegir tamaños de muestra para cada combinación, de modo que:

a El margen de error para la estimación del porcentaje de niños vacunados sea 0.05 o menor al tabular los datos de cada condado (al sumar sobre todas las organizaciones de atención de la salud).

b El margen de error para la estimación del porcentaje de niños vacunados sea 0.05 o menor al tabular cada organización (al sumar sobre todos los condados).

c Al menos dos niños (o menos, por supuesto, si la celda no tiene dos niños) se eligen de cada celda.

Observe que para este problema, como para muchos diseños de muestras, es posible tener muchos diseños distintos.

TABLA 4.6
Tabla para el ejercicio 23

	A	B	C	D	E	Otro	Total
Apache	1	13	19	0	0	94	127
Cochise	2	5	0	637	40	0	694
Coconino	1	6	0	125	0	289	421
Gila	0	2	51	151	0	0	204
Graham	0	2	0	63	0	143	208
Greenlee	0	0	0	58	0	0	58
Maricopa	118	169	0	3,732	2,675	5,105	11,799
Mohave	4	6	0	44	0	476	530
Navajo	2	5	132	124	0	0	263
Pima	62	26	0	1,097	727	1,786	3,698
Pinal	5	10	13	22	360	478	888
Santa Cruz	0	5	0	118	150	0	273
Yavapai	7	8	0	173	0	198	386
Yuma	5	5	0	837	0	0	847
La Paz	0	1	0	89	0	0	90
Total	217	263	215	7,270	3,952	8,569	20,486

5

Muestreo por conglomerados con probabilidades idénticas

"Pero los promedios no son reales", objetó Milo, "sólo son imaginarios".

"Es posible", coincidió el niño; sin embargo, el pequeño afirmó: "pero a veces son muy útiles. Por ejemplo, si no tuvieras dinero, pero estuvieras con otras cuatro personas que contasen con diez dólares cada una, entonces, cada uno de ustedes tendría en promedio ocho dólares, ¿no es así?".

"Creo que sí", dijo Milo en voz baja.

"Bueno, piensa en lo bien que te va gracias a los promedios", dijo convencido el niño y éste continuó con la siguiente explicación: "y piensa en el pobre granjero cuando no ha llovido en todo el año. De no ser por una precipitación anual promedio de 37 pulgadas en esta parte del país, toda su cosecha se perdería".

Todo esto era muy confuso para Milo, ya que siempre había tenido problemas en la escuela con este tema.

"Hay más ventajas", continuó el niño. "Por ejemplo, si una rata fuese atrapada por nueve gatos, en promedio, cada gato obtendría el 10 por ciento de rata, mientras que ésta obtendría el 90 por ciento de cada gato. Si fueras una rata, verías cómo todo es más agradable", le dijo el niño a Milo de forma irónica.

—Norton Juster, *The Phantom Tollbooth*

En todos los procedimientos de muestreo analizados hasta el momento, hemos supuesto que la población está dada y que lo único que debemos hacer es estirarnos y obtener una muestra adecuada de unidades. Pero no necesariamente las unidades están definidas de una manera adecuada, aunque la población lo esté. Puede haber varias formas de enumerar las unidades y el tamaño de unidad elegido puede contener subunidades más pequeñas.

Suponga que queremos determinar cuántas bicicletas son propiedad de los residentes de una comunidad de 10 000 familias. Podríamos extraer una muestra aleatoria simple de 400 familias o dividir a la comunidad en bloques que tuviesen, aproximadamente, 20 familias cada uno y analizar a cada familia (o bien obtener una muestra de algunas de las familias) en cada uno de los 20 bloques elegidos al azar de entre los 500 bloques de la comunidad. El último plan es un ejemplo de muestreo por conglomerados. Los bloques son las **unidades de muestreo primario** o **conglomerados**. Las familias son las **unidades de muestreo secundario**; con frecuencia, las unidades de muestreo secundario son los elementos de la población.

Es probable que la muestra por conglomerados, de 400 familias, dé una menor precisión que una muestra aleatoria simple de 400 familias; algunos bloques de la comunidad están compuestos, principalmente, por familias (con más bicicletas), mientras que los residentes

de otros bloques son, casi todos, jubilados (con menos bicicletas). No es muy probable que veinte familias del mismo bloque reflejen la diversidad de la comunidad tan bien como 20 familias elegidas al azar. Así, es probable que el muestreo por conglomerados en esta situación produzca menos información por cada observación que una muestra aleatoria simple del mismo tamaño. Sin embargo, si usted realiza la encuesta personalmente, es mucho más barato y fácil entrevistar a las 20 familias de un bloque que a 20 familias elegidas al azar de entre toda la comunidad, de modo que el muestreo por conglomerados puede dar más información por cada dólar invertido.

En el muestreo por conglomerados, los elementos individuales de la población sólo pueden participar en la muestra si pertenecen a un conglomerado (unidad de muestreo primario) incluido en la muestra. La unidad de muestreo primario no es igual a la unidad de observación (unidad de muestreo secundario) y hay que tomar en cuenta los dos tamaños de unidades experimentales al calcular los errores estándar de las muestras por conglomerados.

¿Por qué deberíamos utilizar las muestras por conglomerados?

1 La construcción de una lista de unidades de observación para el marco de muestreo puede ser difícil, cara o imposible. No podemos enumerar todas las abejas de una región o a todos los clientes de una tienda; pero sí podemos construir una lista con todos los árboles que pertenecen a una localidad específica de un bosque o una lista de individuos de una ciudad para los cuales sólo existe una lista de unidades habitacionales; sin embargo, la construcción de estas listas consumirá mucho tiempo y será muy cara.

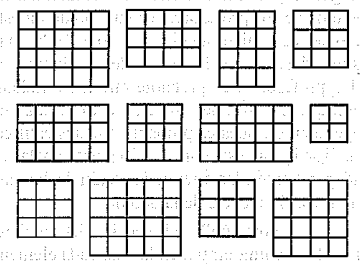
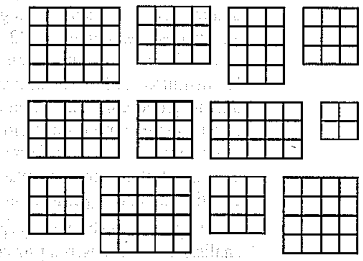
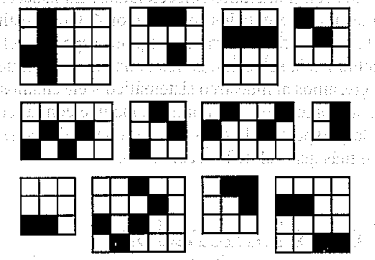
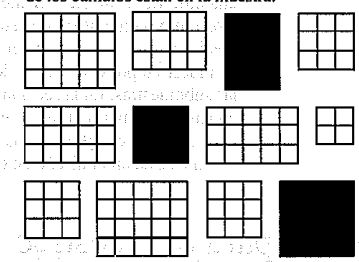
2 La población podría estar muy dispersa geográficamente o aparecer en cúmulos naturales, como las familias o las escuelas. Si la población objetivo son los residentes de los asilos en Estados Unidos, es más barato tomar una muestra de asilos y entrevistar a cada uno de los residentes de los asilos seleccionados, que entrevistar a toda una muestra aleatoria simple de residentes de asilos; con una muestra aleatoria simple de residentes, usted tendría que viajar hasta un asilo sólo para entrevistar a un residente. Al realizar un estudio arqueológico, usted examinaría todos los artefactos en una región y no elegiría puntos al azar ni examinaría sólo los artefactos hallados en esos puntos aislados.

Los conglomerados recuerdan a los estratos, pero sólo de manera superficial: un conglomerado, al igual que un estrato, es una agrupación de los miembros de la población. Sin embargo, el proceso de selección es un poco distinto en ambos métodos. Las analogías y diferencias entre las muestras por conglomerados y las muestras estratificadas se ilustran en la figura 5.1.

Mientras que, por lo general, la estratificación aumenta la precisión en relación con el muestreo aleatorio simple, el muestreo por conglomerados, con frecuencia, la disminuye. Los miembros de un mismo conglomerado tienden a ser más similares que los elementos seleccionados al azar de entre toda la población: los miembros de la misma familia tienden a tener opiniones políticas similares; los peces del mismo lago tienden a presentar concentraciones similares de mercurio; los residentes del mismo asilo tienden a dar opiniones similares sobre la calidad de la atención. Por lo general, estas analogías surgen debido a ciertos factores subyacentes que podrían medirse o no; los residentes del mismo asilo podrían tener opiniones similares debido a que la atención es mala y la concentración de mercurio en los peces reflejaría la concentración de mercurio que existe en el lago. Por tanto, si extraemos una muestra de dos residentes del mismo asilo, no conseguimos tanta información acerca de los residentes de asilos en Estados Unidos como la que obtendríamos al extraer una muestra de dos residentes de asilos distintos, debido a que es probable que los residentes del mismo asilo posean opiniones más similares. Al obtener una muestra de todos los individuos que pertenecen al cúmulo, repetimos parcialmente la misma información en vez de conseguir información nueva y esto implica una menor precisión para las estimaciones de las cantidades de la población. El muestreo por conglomerados se utiliza en la práctica debido a que es más barato y conveniente obtener muestras por conglomerados que al azar entre la población. Casi todas las grandes encuestas familiares realizadas por el gobierno de Estados Unidos, o por instituciones comerciales o académicas utilizan el muestreo por conglomerados debido al ahorro en los costos.

Uno de los más grandes errores cometidos por los investigadores que usan encuestas consiste en analizar una muestra por conglomerados como si fuese una muestra aleatoria

FIGURA 5.1
Analogías y diferencias entre el muestreo por conglomerados y el muestreo estratificado

Muestreo estratificado	Muestreo por conglomerados
Cada elemento de la población está exactamente en un estrato.	Cada elemento de la población está en un solo conglomerado.
Población de H estratos; el estrato h tiene n_h elementos: 	Muestreo por conglomerados en una etapa: población de N conglomerados: 
Se extrae una muestra aleatoria simple de *cada* estrato: 	Se extrae una muestra aleatoria simple de conglomerados; observe que todos los elementos dentro de los cúmulos están en la muestra: 
La varianza de la estimación de $\bar{y}_U$ depende de la variabilidad de los valores *dentro* de los estratos.	El conglomerado es la unidad de muestreo; mientras más cúmulos participen en la muestra, menor será la varianza. La varianza de la estimación de $\bar{y}_U$ depende principalmente de la variabilidad que existe *entre* las medias de los cúmulos.
Para una mayor precisión, los elementos individuales dentro de cada estrato deben tener valores similares, pero las medias por estrato deben diferir entre sí lo más posible.	Para una mayor precisión, los elementos individuales dentro de cada conglomerado deben ser heterogéneos y las medias por cúmulo deben ser similares entre sí.

simple. Por lo general, una confusión como esta hace que los investigadores informen de errores estándar mucho menores de lo debido; esto da la impresión de que los resultados de la encuesta son mucho más precisos de lo que realmente son.

EJEMPLO 5.1 Basow y Silberg (1987) informan de los resultados que obtuvieron de una investigación que realizaron sobre si los estudiantes evalúan de manera distinta a sus maestras que a sus maestros. Los autores formaron parejas de 16 maestras y 16 maestros por materia, años de experiencia docente y nivel de ejercicio; luego, dieron cuestionarios de evaluación a los estudiantes que pertenecían a los grupos de todos estos profesores. El tamaño de muestra para analizar este estudio es $n = 32$, el número de académicos analizados; no es 1029, el número de estudiantes que entregaron los cuestionarios. Las evaluaciones de los alumnos reflejan los distintos estilos de enseñanza de los profesores; es probable que los estudiantes del mismo grupo coincidan en la evaluación que hicieron del profesor y no deban, entonces, considerarse como observaciones independientes, pues es probable que sus calificaciones estén correlacionadas en forma positiva. Si se ignora esta correlación positiva y las calificaciones se tratan como observaciones independientes, las diferencias serán declaradas como estadísticamente significativas con una mayor frecuencia de la debida. ■

Después de un breve viaje por la "tierra de la notación", en la sección 5.1, iniciaremos el análisis con el **muestreo por conglomerados de una etapa**, en la cual cada elemento dentro de un conglomerdo seleccionado forma parte de la muestra. En la sección 5.3, generalizamos los resultados al **muestreo por conglomerados en dos etapas**, en la cual extraemos una muestra de algunos elementos de los conglomerados seleccionados. En la sección 5.4, mostramos cómo usar los pesos de muestreo, presentados en la sección 4.3, para estimar las medias y totales de la población. En la sección 5.5 analizamos algunos aspectos del diseño de muestreo por conglomerados e incluimos la selección de los tamaños de las muestras y las submuestras. En la sección 5.6, regresamos al muestreo sistemático y enseñamos que es un caso particular del muestreo por conglomerados. El capítulo concluye con la teoría del muestreo por conglomerados desde la perspectiva basada en el modelo; deduciremos la teoría basada en el diseño en el marco más general de la sección 6.6.

5.1 Notación para el muestreo por conglomerados

En el muestreo aleatorio simple, las unidades muestreadas son también los elementos observados. En el muestreo por conglomerados, las unidades de muestreo son los conglomerados y los elementos observados constituyen las unidades secundarias dentro de los cúmulos. El universo U es la población de N unidades primarias; S denota la muestra de unidades primarias elegidas entre la población de unidades primarias y S_i es la muestra de unidades secundarias elegidas en la unidad primaria i. Utilizamos la siguiente notación para este capítulo y el siguiente. Las cantidades medidas son:

y_{ij} = medida para el elemento j de la unidad primaria i.

Sin embargo, en el muestreo por conglomerados es más fácil pensar al nivel de las unidades secundarias en términos de los totales del conglomerado. Sin importar cómo se defina, la notación para el muestreo por conglomerados es enredada, pues se necesita una notación para los niveles primario y secundario. En esta sección, presentamos la notación utilizada en este capítulo y en el siguiente como referencia. Observe que en estos capítulos N es el número de unidades primarias y no el número de unidades de observación.

Nivel primario: cantidades de población

N = número de unidades primarias en la población

M_i = número de unidades secundarias en la unidad primaria i

$$K = \sum_{i=1}^{N} M_i = \text{cantidad total de unidades secundarias en la población}$$

$$t_i = \sum_{j=1}^{M_i} y_{ij} = \text{total en la unidad primaria } i$$

$$t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = \text{total de población}$$

$$S_t^2 = \sum_{i=1}^{N} \frac{\left(t_i - \frac{t}{N}\right)^2}{N-1} = \text{varianza en la población de los totales de las unidades primarias}$$

Nivel secundario: cantidades de población

$$\bar{y}_U = \sum_{i=1}^{N}\sum_{j=1}^{M_i} \frac{y_{ij}}{K} = \text{media de la población}$$

$$\bar{y}_{iU} = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i} = \text{media de la población en la unidad primaria } i$$

$$S^2 = \sum_{i=1}^{N}\sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_U)^2}{K-1} = \text{varianza de la población (por unidad secundaria)}$$

$$S_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_{iU})^2}{M_i - 1} = \text{varianza de la población dentro de la unidad primaria } i$$

Cantidades de muestra

n = número de unidades primarias en la muestra

m_i = número de elementos en la muestra de la unidad primaria i

$$\bar{y}_i = \sum_{j \in S_i} \frac{y_{ij}}{m_i} = \text{media muestral (por unidad secundaria) para la unidad primaria } i$$

$$\hat{t}_i = \sum_{j \in S_i} \frac{M_i}{m_i} y_{ij} = \text{total estimado para la unidad primaria } i$$

$$\hat{t}_{\text{unb}} = \sum_{i \in S} \frac{N}{n} \hat{t}_i = \text{estimador insesgado del total de la población}$$

$$s_t^2 = \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N}\right)^2 = \text{varianza estimada de los totales de la unidad primaria } i$$

$$s_i^2 = \sum_{j \in S_i} \frac{(y_{ij} - \bar{y}_i)^2}{m_i - 1} = \text{varianza muestral dentro de la unidad primaria } i$$

5.2 Muestreo por conglomerados en una etapa

En el muestreo por conglomerados en una etapa, todos o ninguno de los elementos que componen un conglomerado (unidad de muestreo primario) están en la muestra. El muestreo por conglomerados en una etapa se utiliza en muchas encuestas donde el costo de muestreo de las unidades secundarias es despreciable en relación con el costo de muestreo de las unidades primarias. Para las encuestas educativas, una unidad primaria natural es el salón de clase; con frecuencia, todos los estudiantes de un determinado salón se incluyen como unidades secundarias, pues se requiere apenas un ligero costo adicional para dar un cuestionario a todos los estudiantes del salón, en vez de a unos cuantos.

En la población de N unidades primarias, la unidad i contiene M_i unidades secundarias (elementos). Extraemos de la población una muestra aleatoria simple de n unidades primarias y medimos nuestra variable de interés en *cada* elemento de la unidad primaria elegida. Así, para el muestreo por conglomerados en una etapa, $M_i = m_i$.

5.2.1 Conglomerados del mismo tamaño: estimación

Consideremos el caso más sencillo, donde cada conglomerado tiene el mismo número de elementos, con $M_i = m_i = M$. Los cúmulos de personas que aparecen con mayor naturalidad no se ajustan a este marco de referencia, pero sí pueden aparecer en el muestreo agrícola e industrial. La estimación de las medias o totales de la población es sencilla: consideramos las medias o totales del cúmulo como las observaciones y sólo ignoramos los elementos individuales.

Así, tenemos una muestra aleatoria simple de n observaciones $\{t_i, i \in S\}$; t_i es el total para todos los elementos de la unidad primaria i. Entonces, $\bar{t}_S$ estima el promedio de los totales por cúmulo. Se realiza una encuesta sobre las familias para estimar el ingreso familiar con dos personas, las observaciones individuales y_{ij} son los ingresos de cada persona dentro de la familia, t_i es el ingreso total de la familia i (*conocemos* t_i para las familias de la muestra, pues se entrevista a ambas personas), $\bar{t}_U$ es el ingreso promedio por familia y $\bar{y}_U$ es el ingreso promedio por persona. Para estimar el ingreso total t podemos usar el siguiente estimador:

$$\hat{t} = \frac{N}{n}\sum_{i\in S} t_i. \tag{5.1}$$

Podemos aplicar los resultados de las secciones 2.3 y 2.7 a $\hat{t}$, pues tenemos una muestra aleatoria simple de n unidades, que ha sido extraída de una población de N unidades. Como resultado, $\hat{t}$ es un estimador insesgado de t, con varianza dada por:

$$V(\hat{t}) = N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} \tag{5.2}$$

y con

$$\mathrm{EE}(\hat{t}) = N\sqrt{\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}}, \tag{5.3}$$

donde S_t^2 y s_t^2 son la varianza de la población y de la muestra, respectivamente, de los totales de la unidad primaria:

$$S_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(t_i - \frac{t}{N}\right)^2$$

y

$$s_t^2 = \frac{1}{n-1}\sum_{i\in S}\left(t_i - \frac{\hat{t}}{N}\right)^2.$$

Para estimar $\bar{y}_U$, dividimos el total estimado entre el número de personas, con lo que obtenemos:

$$\hat{\bar{y}} = \frac{\hat{t}}{NM}, \tag{5.4}$$

con

$$V(\hat{\bar{y}}) = \left(1-\frac{n}{N}\right)\frac{S_t^2}{nM^2} \tag{5.5}$$

y

$$\mathrm{EE}(\hat{\bar{y}}) = \frac{1}{M}\sqrt{\left(1-\frac{n}{M}\right)\frac{s_t^2}{n}}. \tag{5.6}$$

No hay ideas nuevas que se hayan incorporado al caso del muestreo por conglomerados en una etapa; sólo usamos los resultados del muestreo aleatorio simple con los totales por conglomerado como observaciones.

EJEMPLO 5.2 Un estudiante quiere estimar las calificaciones promedio de sus compañeros de dormitorio. En vez de obtener una lista de todos los alumnos que pertenecen al dormitorio y realizar una muestra aleatoria simple, observa que dicho dormitorio consta de 100 cuartos, cada uno con cuatro estudiantes; elige 5 cuartos al azar y pregunta a cada persona sus calificaciones. Los resultados son los siguientes:

Número de persona	Cuarto (Conglomerado) 1	2	3	4	5
1	3.08	2.36	2.00	3.00	2.68
2	2.60	3.04	2.56	2.88	1.92
3	3.44	3.28	2.52	3.44	3.28
4	3.04	2.68	1.88	3.64	3.20
Total	12.16	11.36	8.96	12.96	11.08

Las unidades primarias son los cuartos, de modo que $N = 100$, $n = 5$ y $M = 4$. La estimación del total de la población (la suma estimada de todas las calificaciones de todos los alumnos que pertenecen al dormitorio, cantidad sin sentido en este ejemplo, pero útil para demostrar el procedimiento) es:

$$\hat{t} = \frac{100}{5}(12.16+11.36+8.96+12.96+11.08) = 1130.4,$$

y

$$s_t^2 = \frac{1}{5-1}[(12.16-11.304)^2 + \cdots + (11.08-11.304)^2] = 2.256.$$

En este ejemplo, s_t^2 es sólo la varianza muestral de los totales de los 5 cuartos. Así, al usar (5.4) y (5.6), $\hat{\bar{y}} = 1130.4/400 = 2.826$, y

$$\mathrm{EE}(\hat{\bar{y}}) = \sqrt{\left(1-\frac{5}{100}\right)\frac{2.256}{(5)(4)^2}} = 0.164.$$

Observe que en estos cálculos sólo se utiliza el renglón "total" de la tabla de datos; las calificaciones individuales se emplean sólo en la contribución al total de los cuartos. ■

El muestreo por conglomerados en una etapa, con una muestra aleatoria simple de unidades primarias, produce una muestra autoponderada. El peso de cada unidad de observación es:

$$w_{ij} = \frac{1}{P\{\text{unidad secundaria } j \text{ de la unidad primaria } i \text{ que está en la muestra}\}} = \frac{N}{n}.$$

Entonces, para los datos del ejemplo 5.2,

$$\begin{aligned}\hat{t} &= \sum_{i\in S}\sum_{j\in S_i} w_{ij}y_{ij} \\ &= \frac{N}{n}(3.08+2.60+\cdots+3.28+3.20) \\ &= \frac{100}{5}(56.52) = 1130.4.\end{aligned}$$

Así, como en el muestreo estratificado, podemos estimar un total de la población al sumar el producto de los valores observados y los pesos de muestreo.

De haber tomado una muestra aleatoria simple de nM elementos, cada elemento de la muestra tendría asignado el peso $(NM)/(nM) = N/n$, que es el mismo peso obtenido para el muestreo por conglomerados. Sin embargo, la precisión obtenida para los dos tipos de muestreo puede ser muy diferente; la diferencia de precisión se explora en la siguiente sección.

5.2.2 Conglomerados del mismo tamaño: teoría

En esta sección comparamos el muestreo por conglomerados con el muestreo aleatorio simple. Casi siempre, el muestreo por conglomerados proporciona una menor precisión para los estimadores que en el caso de una muestra aleatoria simple con el mismo número de elementos.

Como en el muestreo estratificado, veamos la tabla de análisis de la varianza (tabla 5.1) para toda la población. En el muestreo estratificado, la varianza del estimador de t depende de la variabilidad que exista *dentro* de los estratos; la ecuación (4.3) y la tabla 4.3 implican que la varianza en el muestreo estratificado es pequeña si la suma de cuadrados dentro de las unidades primarias (SSW) es pequeña con respecto a la suma de cuadrados total (SSTO), o en forma equivalente, si el cuadrado de la media dentro de las unidades primarias (MSW) es pequeño con respecto a S^2. En el muestreo estratificado, usted tiene información de *cada* estrato, así que no tiene que preocuparse por la variabilidad ocasionada por estratos no participantes en la muestra. Si MSB/MSW es grande (es decir, la variabilidad entre las medias por estratos es grande comparada con la variabilidad dentro de los estratos), entonces, el muestreo estratificado aumenta la precisión.

TABLA 5.1
Tabla de análisis de la varianza de la población, muestreo por conglomerados

Fuente	grados de libertad	Suma de cuadrados	Cuadrado de la media
Entre las unidades primarias	$N-1$	$\text{SSB} = \sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{y}_{iU}-\bar{y}_U)^2$	MSB
Dentro de las unidades primarias	$N(M-1)$	$\text{SSW} = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{iU})^2$	MSW
Total en torno a $\bar{y}_U$	$NM-1$	$\text{SSTO} = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_U)^2$	S^2

Ocurre lo contrario en el muestreo por conglomerados. En el muestreo por conglomerados en una etapa, la variabilidad del estimador insesgado de t depende completamente de la parte de la variabilidad que existe *entre* los conglomerados, pues

$$S_t^2 = \sum_{i=1}^{N}\frac{(t_i-\bar{t}_U)^2}{N-1} = \sum_{i=1}^{N}\frac{M^2(\bar{y}_{iU}-\bar{y}_U)^2}{N-1} = M(\text{MSB}).$$

Así, para el muestreo por conglomerados,

$$V(\hat{t}_{\text{cúmulo}}) = N^2\left(1-\frac{n}{M}\right)\frac{M(\text{MSB})}{n}. \quad \textbf{(5.7)}$$

Si MSB/MSW es grande en el muestreo por conglomerados, entonces, este tipo de muestreo reduce la precisión. En este caso, MSB es relativamente grande, pues mide la variabilidad de un conglomerado a otro. Con frecuencia, los elementos de conglomerados distintos varían más que los elementos en el mismo cúmulo, pues conglomerados distintos tienen medias diferentes. Si tomamos una muestra por conglomerados de los grupos y obtenemos una muestra de todos los estudiantes dentro de los grupos seleccionados, es probable que encontremos que las calificaciones promedio de lectura varíen de un grupo a otro. Un excelente maestro de lectura podría elevar las calificaciones de toda la clase; un grupo de estudiantes de un área pobre podría estar descuidado y no tener una buena calificación en la lectura. Los factores no medidos, como la calidad de la enseñanza o la pobreza, pueden afectar a la media global de un conglomerado y hacer que MSB sea grande.

Además, dentro de un grupo, las calificaciones de lectura pueden variar. MSW es el valor combinado de las varianzas dentro de los conglomerados: la varianza de un elemento a otro, presente para todos los elementos de la población. Si los conglomerados son relativamente homogéneos (por ejemplo, cuando los estudiantes del mismo grupo tienen calificaciones similares), MSW será pequeña.

Ahora, compararemos el muestreo por conglomerados con el muestreo aleatorio simple. Si en vez de tomar una muestra por conglomerados de M elementos en cada uno de los n cúmulos, hubiésemos tomado una muestra aleatoria simple (MAS) con nM observaciones, la varianza del total estimado hubiera sido la siguiente:

$$V(\hat{t}_{\text{MAS}}) = (NM)^2\left(1-\frac{nM}{NM}\right)\frac{S^2}{nM} = N^2\left(1-\frac{n}{N}\right)\frac{MS^2}{n}.$$

Al comparar esto con (5.7), vemos que si MSB > S^2, entonces, el muestreo por conglomerados es menos eficiente que el muestreo aleatorio simple.

El **coeficiente de correlación dentro de la clase** (a veces llamado **dentro del conglomerado** o ICC por sus siglas en inglés) nos dice qué tan similares son los elementos del mismo conglomerado. Proporciona una **medida de homogeneidad** dentro de los conglomerados. Este coeficiente se define como el coeficiente de correlación de Pearson para las $NM(M-1)$ parejas (y_{ij}, y_{ik}) para i entre 1 y N y $j \neq k$ (véase el ejercicio 9) y se puede escribir en términos de las cantidades de la tabla de análisis de la varianza como:

$$\text{ICC} = 1 - \frac{M}{M-1}\frac{\text{SSW}}{\text{SSTO}}. \quad \textbf{(5.8)}$$

Como $0 \le \text{SSW/SSTO} \le 1$, (5.8) implica que:

$$-\frac{1}{M-1} \le \text{ICC} \le 1.$$

Si los conglomerados son perfectamente homogéneos y con ello SSW = 0, entonces, ICC = 1. La ecuación (5.8) implica que:

$$\text{MSB} = \frac{NM-1}{M(N-1)}S^2[1+(M-1)\text{ICC}].$$

¿Cuánta precisión perdemos al tomar una muestra por conglomerados? De la ecuación anterior y (5.7), tenemos lo siguiente:

$$\frac{V(\hat{t}_{\text{cúmulo}})}{V(\hat{t}_{\text{MAS}})} = \frac{\text{MSB}}{S^2} = \frac{NM-1}{M(N-1)}[1+(M-1)\text{ICC}]. \qquad \textbf{(5.9)}$$

Si N, el número de unidades primarias que pertenecen a la población, es tan grande que $NM-1 \approx M(N-1)$, entonces, el cociente de las varianzas en (5.9) es aproximadamente $1 + (M-1)$ICC. Así, $1 + (M-1)$ICC unidades secundarias, extraídas en una muestra por conglomerados en una etapa, nos dan aproximadamente la misma cantidad de información que una unidad secundaria de una muestra aleatoria simple. Si ICC = 1/2 y M = 5, entonces $1 + (M-1)\text{ICC} = 3$ y necesitaríamos medir 300 elementos mediante una muestra por cúmulos para obtener la misma precisión de una muestra aleatoria simple de 100 elementos. Sin embargo, esperamos (ya que con frecuencia es más barato y fácil reunir los datos de una muestra por conglomerados) obtener más información por el dinero invertido en el muestreo por conglomerados.

El ICC proporciona una medida de homogeneidad para los conglomerados. El ICC es positivo si los elementos dentro de una unidad primaria tienden a ser similares; en ese caso, SSW será pequeña con respecto a SSTO y el ICC será relativamente grande. Cuando el ICC es positivo, el muestreo por conglomerados es menos eficiente que el muestreo aleatorio simple de los elementos.

Si los conglomerados aparecen de manera natural en la población, usualmente el ICC es positivo. Los elementos dentro del mismo conglomerado tienden a ser más similares que los elementos elegidos al azar entre la población. Esto puede ocurrir debido a que los elementos de un conglomerado comparten un ambiente similar (sería de esperar que los pozos de un mismo cúmulo geográfico tuviesen niveles similares de pesticidas o que un área de una ciudad tuviese una incidencia de sarampión distinta a la de otra área de la ciudad). En las poblaciones humanas, las elecciones personales, así como la interacción entre los miembros de una familia o los vecinos, pueden hacer que el ICC sea positivo; las familias saludables tienden a vivir en barrios similares y las personas del mismo barrio pueden tener opiniones similares.

El ICC es negativo si los elementos que están dentro de un conglomerado se dispersan *más* de lo que se dispersaría un grupo elegido al azar. Esto obliga a las medias del conglomerado a ser casi iguales; como SSTO = SSW + SSB, si SSTO queda fijo y SSW es grande, entonces, SSB debe ser pequeña. Si ICC < 0, el muestreo por conglomerados es más eficiente que el muestreo aleatorio simple de los elementos. Es raro que el ICC sea negativo en los conglomerados que aparecen de manera natural; los valores negativos pueden ocurrir en algunas muestras sistemáticas o en conglomerados artificiales, como veremos en la sección 5.6.

El ICC sólo está definido para conglomerados del mismo tamaño. Una cantidad alternativa que se puede usar como una medida de la homogeneidad en las poblaciones generales es la R^2, llamada R_a^2 y definida como:

$$R_a^2 = 1 - \frac{\text{MSW}}{S^2}. \qquad \textbf{(5.10)}$$

Si todos los conglomerados tienen el mismo tamaño, entonces, el incremento en la varianza debido al muestreo por conglomerados es:

$$\frac{\text{MSB}}{S^2} = 1 + \frac{N(M-1)}{N-1}R_a^2;$$

si compara con (5.9), verá que para muchas poblaciones, R_a^2 está muy cerca del ICC. R_a^2 es una medida razonable de la homogeneidad debido a su interpretación en la regresión lineal: es la cantidad relativa de variabilidad que existe dentro de la población, la cual es explicada por las medias de los conglomerados ajustada para el número de grados de libertad. Si los conglomerados son homogéneos, entonces, las medias por conglomerado son muy variables con respecto a la variación dentro de los conglomerados y R_a^2 será grande.

EJEMPLO 5.3 Consideremos dos poblaciones artificiales, cada una con tres conglomerados y tres elementos por conglomerado.

	Población A			Población B		
Conglomerado 1	10	20	30	9	10	11
Conglomerado 2	11	20	32	17	20	20
Conglomerado 3	9	17	31	31	32	30

Los elementos son los mismos para las dos poblaciones, de modo que éstas comparten los valores $\bar{y}_U = 20$ y $S^2 = 84.5$. En la población A, la mayor parte de la variabilidad aparece dentro de los conglomerados; en la población B, la mayor parte de la variabilidad aparece entre los conglomerados.

	Población A		Población B	
	$\bar{y}_{iU}$	S_i^2	$\bar{y}_{iU}$	S_i^2
Conglomerado 1	20	100	10	1
Conglomerado 2	21	111	19	3
Conglomerado 3	19	124	31	1

Tabla de análisis de la varianza para la población A:

Fuente	gl	SC	CM	F
Entre los conglomerados	2	6	3	0.03
Dentro de los conglomerados	6	670	111.67	
Total en torno a la media	8	676	84.5	

Tabla de análisis de la varianza para la población B:

Fuente	gl	SC	CM	F
Entre los conglomerados	2	666	333	199.8
Dentro de los conglomerados	6	10	1.67	
Total en torno a la media	8	676	84.5	

$$R_a^2 = -0.3215 \text{ e ICC} = 1 - \left(\frac{3}{2}\right)\frac{670}{676} = -0.4867 \text{ para la población A.}$$

$$R_a^2 = 0.9803 \text{ e ICC} = 1 - \left(\frac{3}{2}\right)\frac{10}{676} = 0.9778 \text{ para la población B.}$$

La población A tiene mucha variación entre los elementos que se encuentran dentro de los conglomerados, pero poca variación entre las medias de los conglomerados. Esto se refleja en los grandes valores negativos del ICC y R_a^2. Los elementos del mismo conglomerado son, en realidad, menos similares que los elementos elegidos al azar de toda la población. Para esta situación, el muestreo por conglomerados es más eficiente que el muestreo aleatorio simple.

Lo contrario ocurre en la población B: la mayor parte de la variabilidad ocurre entre los conglomerados, y éstos son relativamente homogéneos. El ICC y el R_a^2 son muy cercanos a 1, lo cual indica que se obtiene poca información nueva al incluir en la muestra a más de un elemento del conglomerado. En este caso, el muestreo por conglomerados en una etapa es mucho menos eficiente que el muestreo aleatorio simple. ■

La mayor parte de las poblaciones que pertenecen al mundo real están en algún punto intermedio de estos extremos. Por lo general, el ICC es positivo, pero no está muy cercano a 1. Así, se pierde un poco de eficiencia al usar el muestreo por conglomerados; esta situación debe ser aprovechada para ahorrar costos.

EJEMPLO 5.4 Cuando todos los conglomerados tienen el mismo tamaño, podemos estimar la varianza de $\hat{t}$ y el ICC a partir de la tabla de análisis de varianza de la muestra. He aquí la tabla de

análisis de la varianza de la muestra para los datos de calificaciones del ejemplo 5.2:

Fuente	gl	SC	CM
Entre los cuartos	4	2.2557	0.56392
Dentro de los cuartos	15	2.7756	0.18504
Total	19	5.0313	0.26480

En un muestreo por conglomerados en una etapa, con conglomerados del mismo tamaño, los cuadrados de las medias para los casos dentro de los cuartos y entre los cuartos son estimadores insesgados de las cantidades correspondientes en la tabla de análisis de varianza de la población (véase el ejercicio 11). Así,

$$E[\widehat{\text{MSB}}] = \text{MSB} = \frac{S_t^2}{M}$$

y, al usar (5.7),

$$\text{EE}(\hat{\bar{y}}) = \sqrt{\left(1-\frac{n}{N}\right)\frac{\widehat{\text{MSB}}}{nM}} = \sqrt{\left(1-\frac{5}{100}\right)\frac{0.56392}{(5)(4)}} = 0.164,$$

como lo calculamos en el ejemplo 5.2.

Sin embargo, el total del cuadrado de la media de la muestra no se debe usar para estimar S^2 cuando n es pequeño. Estos datos se reunieron como una muestra por conglomerados y por tanto, no reflejan de manera adecuada la variabilidad de un conglomerado a otro. En vez de esto, debemos multiplicar las estimaciones insesgadas de MSB y MSW por los grados de libertad de la tabla de análisis de varianza de la población para estimar las sumas de cuadrados de la población en la siguiente tabla. Primero, estimamos las cantidades SSB y SSW para la población y luego las sumamos para estimar SSTO.

Para estos datos, como la población tiene 100 cuartos y, por tanto, 99 grados de libertad para estos cuartos, SSB = 99 × 0.56392 = 55.828. Las estimaciones de las sumas de cuadrados para la población aparecen en la siguiente tabla:

Fuente	gl	$\widehat{\text{SC}}$ (estimado)	CM
Entre los cuartos	99	55.828	0.56392
Dentro de los cuartos	300	55.512	0.18504
Total	399	111.340	0.279

Con estas estimaciones, $\hat{S}^2 = 111.340/399 = 0.279$ (observe la pequeña diferencia que existe entre esta estimación y la de la tabla de análisis de la varianza de la muestra, 0.265). Además,

$$\widehat{\text{ICC}} = 1-\left(\frac{4}{3}\right)\frac{55.512}{111.34} = 0.335$$

y

$$\hat{R}_a^2 = 1-\frac{0.18504}{0.279} = 0.337.$$

Estimamos el incremento en la varianza para el uso del muestreo por conglomerados como:

$$\frac{\widehat{\text{MSB}}}{\hat{S}^2} = \frac{0.56392}{0.279} = 2.02.$$

Este resultado nos indica que necesitamos una muestra de aproximadamente $2.02n$ elementos en una muestra por conglomerados para obtener la misma precisión de una muestra aleatoria simple de tamaño n. Hay cuatro personas en un cúmulo, de modo que en términos de precisión, un conglomerado vale, aproximadamente, 4/2.02 = 1.98 personas en una muestra aleatoria simple. ■

EJEMPLO 5.5 ¿Cuándo se presenta la situación en la que un conglomerado no es tal? Cuando la población es completa.

Considere la situación en la que se lleva acabo un muestreo de robles, en la isla de Santa Cruz, descrita en el ejemplo 3.5. En ese caso, la unidad de muestreo era un árbol y una unidad de observación era un brote cerca del árbol. La población de interés estaba constituida por los brotes de robles en la isla. Como se obtuvo una muestra aleatoria de estos árboles, los tratamos como independientes en el contexto del problema; la independencia era razonable, pues sólo nos interesaba generalizar los resultados del muestreo a toda la población de robles de la isla.

Pero supongamos que el investigador está interesado en la supervivencia de los brotes que existen en toda California, por lo que dividió las regiones con árboles de roble en áreas del mismo tamaño y por lo que eligió al azar cinco de estas áreas para incluirlas en el estudio. Entonces, la unidad de muestreo primario es el área y se obtiene una submuestra de los árboles que pertenecen a cada área. Si la isla de Santa Cruz hubiese sido elegida como una de las cinco áreas, no podríamos considerar a diez árboles que se encuentran en la Isla de Santa Cruz como parte de una muestra aleatoria de árboles representativa de toda la población, sino que tales árboles son parte del cúmulo de la isla de Santa Cruz. Esperaríamos que los diez árboles de la isla de Santa Cruz experimenten, como grupo, factores ambientales distintos (condiciones climáticas, número de depredadores) a los diez árboles elegidos del Valle de Santa Inés, en tierra firme. Así, es probable que el ICC dentro de cada conglomerado (área) sea positivo.

Sin embargo, supongamos que sólo nos interesan los brotes del árbol número 10 de la isla de Santa Cruz. Entonces, la población está formada por todos los brotes de tal árbol y la unidad de muestreo primario es el brote. Así, en este caso, el árbol no es un conglomerado sino toda una población. ■

5.2.3 Conglomerados de distintos tamaños

En los estudios sociales, es raro que los conglomerados tengan el mismo tamaño. En una de las primeras muestras de probabilidad (Converse 1987), el censo de verificación enumerativa de 1937, se eligió una muestra de 2% de las rutas postales y se distribuyeron cuestionarios a todas las familias de cada ruta postal elegida, con el objetivo de verificar las cifras de desempleo. Como las rutas postales tienen distinto número de familias, los tamaños de los conglomerados podían variar en gran medida.

En una muestra por conglomerados de una etapa de n de las N unidades primarias, sabemos cómo estimar los totales y medias de la población de dos formas: al utilizar los estimadores insesgados y la estimación de proporción.

5.2.3.1 Estimación insesgada

Calculamos un estimador insesgado de t exactamente como en (5.1):

$$\hat{t}_{\text{ins}} = \frac{N}{n}\sum_{i\in S} t_i. \tag{5.11}$$

Por (5.3),

$$\text{EE}(\hat{t}_{\text{ins}}) = N\sqrt{\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}}. \tag{5.12}$$

La diferencia entre los cúmulos con el mismo o distinto tamaño se debe a que es probable que la variación entre los totales de los conglomerados t_i individuales sea grande, cuando los conglomerados tengan distintos tamaños. Los investigadores que realizaron el censo de

1937 estaban interesados en el número total de personas desempleadas y t_i sería el número de personas desempleadas en la ruta postal i. Se esperaría un mayor número de personas y, por tanto, más personas desempleadas en una ruta postal con un gran número de familias, que en una ruta postal con un número pequeño de familias. Así, esperaríamos que t_i sea grande cuando el tamaño de conglomerado M_i fuese también grande, y que sea pequeño si M_i también lo es. Así, con frecuencia, s_t^2 es más grande en una muestra por conglomerados cuando las unidades primarias tienen distintos tamaños que cuando todas las unidades primarias tienen el mismo número de unidades secundarias.

La probabilidad de que una unidad primaria esté en la muestra es n/N, cuando se extrae una muestra aleatoria simple de n de las N unidades primarias. Como se utiliza un muestreo por conglomerados en una etapa, una unidad secundaria se incluye en la muestra si su unidad primaria está incluida. Así, como en la página 138,

$$w_{ij} = \frac{1}{P\{\text{unidad secundaria } j \text{ de la unidad primaria } i \text{ está en la muestra}\}} = \frac{N}{n}.$$

Un muestreo por conglomerados de una etapa produce una muestra autoponderada cuando las unidades primarias se eligen con la misma probabilidad. Si usamos los pesos, (5.11) se puede escribir como

$$\hat{t}_{\text{ins}} = \sum_{i\in S}\sum_{j\in S_i} \omega_{ij} y_{ij}. \tag{5.13}$$

Podemos usar (5.11) y (5.12) para obtener un estimador insesgado de $\bar{y}_U$ y su varianza. Definimos

$$K = \sum_{i=1}^{N} M_i$$

como la cantidad total de unidades secundarias en la población; entonces

$$\hat{\bar{y}}_{\text{ins}} = \frac{\hat{t}_{\text{ins}}}{K} \tag{5.14}$$

y

$$\text{EE}(\hat{\bar{y}}_{\text{ins}}) = \frac{\text{EE}(\hat{t}_{\text{ins}})}{K}. \tag{5.15}$$

Sin embargo, para usar (5.14) debemos conocer K; con frecuencia, sólo conocemos M_i para los conglomerados de la muestra. Por ejemplo, en el censo de 1937, el número de familias en una ruta postal se conocería sólo para las rutas postales elegidas para estar en la muestra.

5.2.3.2 Estimación de proporción

Por lo general, esperamos que t_i esté correlacionado con M_i; al usar la estimación de proporción, las M_i son las variables auxiliares, al tomar el papel de las x_i del capítulo 3. Definimos:

$$\hat{\bar{y}}_r = \frac{\sum_{i\in S} t_i}{\sum_{i\in S} M_i}, \tag{5.16}$$

$$\hat{t}_r = K\hat{\bar{y}}_r. \tag{5.17}$$

El estimador $\hat{\bar{y}}_r$ en (5.16) es la cantidad $\hat{B}$ del capítulo 3. El denominador depende de la unidad primaria incluida en la muestra, de modo que el numerador y el denominador varían de una muestra a otra. De (3.7),

$$\text{EE}(\hat{\bar{y}}_r) = \sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n\overline{M}_U^2}\frac{\sum_{i\in S}(t_i - \hat{\bar{y}}_r M_i)^2}{n-1}}$$

$$= \sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n\overline{M}_U^2}\frac{\sum_{i\in S} M_i^2(\bar{y}_i - \hat{\bar{y}}_r)^2}{n-1}} \tag{5.18}$$

y en consecuencia,

$$\text{EE}(\hat{t}_r) = N\sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n}\frac{\sum_{i\in S} M_i^2(\bar{y}_i - \hat{\bar{y}}_r)^2}{n-1}}. \tag{5.19}$$

Si conocemos $\overline{M}_U$, el tamaño promedio de los conglomerados en la población, podríamos sustituir el promedio de los tamaños de las unidades primarias en la muestra, $\overline{M}_S$, en vez de $\overline{M}_U$, en (5.18). Rao y Rao (1971) determinaron que el estimador de la varianza mediante $\overline{M}_S$ tiene menos sesgo que el estimador de la varianza que utiliza $\overline{M}_U$ si la varianza de las y en x_i es proporcional a x_i^t para $0 \le t \le 3/2$, bajo ciertas condiciones.

Observe que $\hat{\bar{y}}_r$ de (5.16) también se puede calcular mediante los pesos w_{ij} como:

$$\hat{\bar{y}}_r = \frac{\sum_{i\in S}\sum_{j\in S_i}\omega_{ij} y_{ij}}{\sum_{i\in S}\sum_{j\in S_i}\omega_{ij}}. \tag{5.20}$$

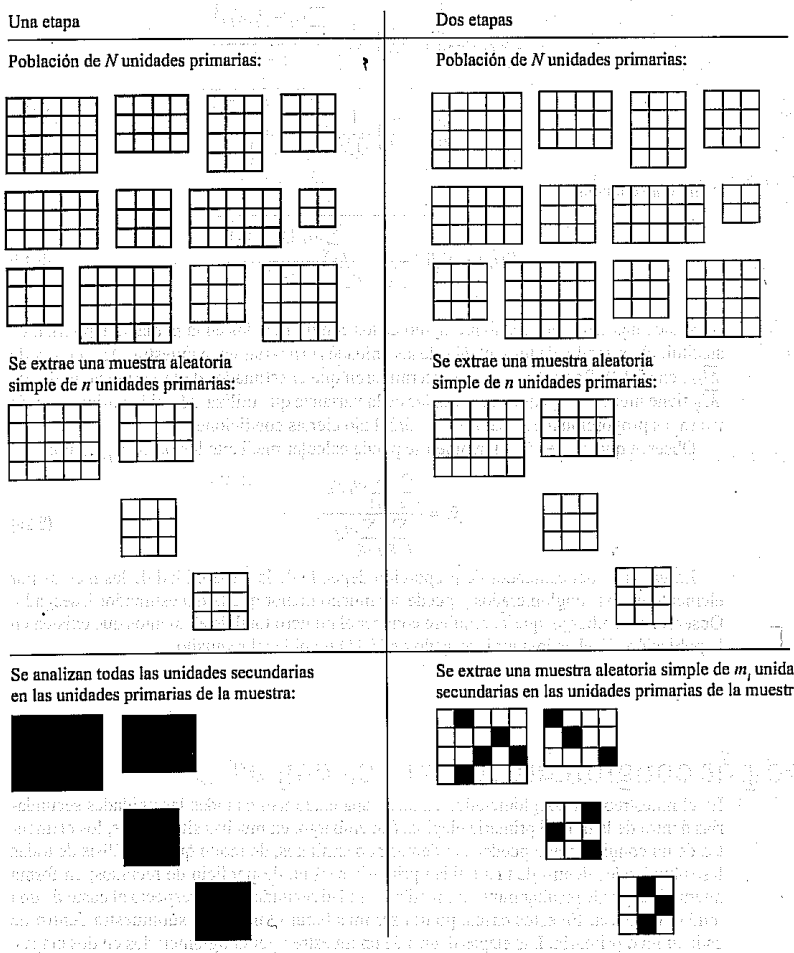
La varianza del estimador de proporción depende de la variabilidad de las medias por elemento en los conglomerados y puede ser mucho menor que la del estimador insesgado. Observe, sin embargo, que $\hat{t}_r$ requiere conocer el número total de elementos que existen en la población, K; el estimador insesgado de (5.11) no pide tal requisito.

5.3 Muestreo por conglomerados en dos etapas

En el muestreo por conglomerados en una etapa examinamos todas las unidades secundarias dentro de la unidad primaria elegida. Sin embargo, en muchas situaciones, los elementos de un conglomerado pueden ser demasiado similares, de modo que el análisis de todas las subunidades dentro de una unidad primaria será un desperdicio de recursos; en forma alternativa, puede resultar muy caro medir la unidad secundaria con respecto al costo de una unidad primaria. En estos casos, podría ser más barato tomar una submuestra dentro de cada unidad primaria. Las etapas dentro de un muestreo por conglomerados en dos etapas, con una muestra de unidades primarias y una submuestra de unidades secundarias con las mismas probabilidades, son las siguientes:

1. Se elige una muestra aleatoria simple $\mathcal{S}$ de n unidades primarias de entre la población de N unidades primarias.
2. Se elige una muestra aleatoria simple de unidades secundarias de cada unidad primaria. La muestra aleatoria simple de m_i elementos del conglomerados i se denota como $\mathcal{S}_i$.

FIGURA 5.2
La diferencia entre el muestreo por conglomerados en una y dos etapas



La diferencia entre el muestreo por conglomerados en una y dos etapas se ilustra en la figura 5.2. La etapa adicional complica la notación y los estimadores, ya que debemos tomar en cuenta la variabilidad que existe entre ambas etapas de recolección de datos. Las

estimaciones puntuales de t y $\bar{y}_U$ son análogas a las del muestreo por conglomerados de una etapa, pero las fórmulas para la varianza son complicadas.

En el muestreo por conglomerados de una etapa, podríamos estimar el total de la población mediante $\hat{t}_{\text{ins}} = (N/n)\sum_{i\in S} t_i$; conocemos los totales t_i de las unidades primarias porque analizamos cada unidad secundaria en la unidad primaria seleccionada. Por otro lado, en el muestreo por cúmulos de dos etapas, debido a que no observamos todas las unidades secundarias en la unidad primaria de la muestra, debemos estimar los totales individuales de la unidad primaria como:

$$\hat{t}_i = \sum_{j\in S_i} \frac{M_i}{m_i} y_{ij} = M_i \bar{y}_i,$$

un estimador insesgado del total de la población es:

$$\hat{t}_{\text{ins}} = \frac{N}{n}\sum_{i\in S} \hat{t}_i = \frac{N}{n}\sum_{i\in S} M_i \bar{y}_i. \tag{5.21}$$

En el muestreo de dos etapas, los $\hat{t}_i$ son variables aleatorias. En consecuencia, la varianza de $\hat{t}$ tiene dos componentes: (1) la variabilidad entre las unidades primarias y (2) la variabilidad de las unidades secundarias dentro de las unidades primarias. No tenemos que preocuparnos por el componente (2) en el muestreo por conglomerados de una etapa.

La varianza de $\hat{t}_{\text{ins}}$ es igual a la varianza de $\hat{t}_{\text{ins}}$ del muestreo por conglomerados en una etapa, más otro término debido a que las $\hat{t}_i$ estiman los totales del cúmulo. Para el muestreo por conglomerados en dos etapas,

$$V(\hat{t}_{\text{ins}}) = N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\sum_{i=1}^{N}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{S_i^2}{m_i}, \tag{5.22}$$

donde S_t^2 es la varianza poblacional de los totales del conglomerado y S_i^2 es la varianza poblacional entre los elementos dentro del conglomerado i. El primer término en (5.22) es la varianza del muestreo por conglomerados en una etapa y el segundo término es la varianza adicional debida al submuestreo. Para demostrar (5.22), tenemos que establecer una condición sobre las unidades incluidas en la muestra. Podemos hacer esto más fácilmente en el marco general del muestreo con probabilidades distintas; para no demostrar dos veces el mismo resultado, demostraremos el resultado general en la sección 6.6.[1]

Para estimar $V(\hat{t}_{\text{ins}})$, sea

$$s_t^2 = \frac{\sum_{i\in S}\left(\hat{t}_i - \frac{\hat{t}_{\text{ins}}}{N}\right)^2}{n-1} \tag{5.23}$$

y

$$s_i^2 = \frac{\sum_{j\in S_i}(y_{ij}-\bar{y}_i)^2}{m_i - 1}. \tag{5.24}$$

Como se verá en la sección 6.6, un estimador insesgado de la varianza en (5.22) es:

$$\hat{V}(\hat{t}_{\text{ins}}) = N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\sum_{i\in S}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}. \tag{5.25}$$

[1]Al trabajar con el nivel adicional de abstracción podemos ver la estructura de la varianza con más claridad, sin tropezar en la notación del caso particular de probabilidades iguales analizado en este capítulo. Si prefiere ver la demostración antes de usar los resultados de varianza, lea la sección 6.6.

El error estándar $\text{EE}(\hat{t}_{\text{ins}})$ es, por supuesto, la raíz cuadrada de (5.25). En muchos casos, N/n será pequeño con respecto a N^2, de modo que la contribución del segundo término en (5.25) al estimador de la varianza será despreciable en comparación con la del primer término.

Si conocemos la cantidad total de los elementos que pertenecen a la población, K, podemos estimar la media de la población mediante

$$\hat{\bar{y}}_{\text{ins}} = \frac{\hat{t}_{\text{ins}}}{K} \tag{5.26}$$

con el siguiente error estándar:

$$\text{EE}(\hat{\bar{y}}_{\text{ins}}) = \frac{\text{EE}(\hat{t}_{\text{ins}})}{K}. \tag{5.27}$$

Como en el caso del muestreo por conglomerados en una etapa que presenta distintos tamaños de conglomerado, el componente de la varianza entre las unidades primarias puede ser muy grande, pues es afectado por las variaciones de tamaño de las unidades (las M_i) y por las variaciones de las $\bar{y}_i$. Si los tamaños de conglomerado son muy diferentes entre sí, este componente es grande, aunque las medias de los conglomerados sean casi constantes.

Estimación de proporción También podemos usar un estimador de proporción para calcular la media de la población. De nuevo, las y del capítulo 3 son los totales del conglomerado (ahora estimados) y las x son los tamaños del conglomerado M_i:

$$\hat{\bar{y}}_r = \frac{\sum_{i\in S}\hat{t}_i}{\sum_{i\in S}M_i} = \frac{\sum_{i\in S}M_i\bar{y}_i}{\sum_{i\in S}M_i}. \tag{5.28}$$

La fórmula de la varianza se basa, de nuevo, en la aproximación mediante la serie de Taylor en (3.7):

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{\bar{M}^2}\left[\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN}\sum_{i\in S}M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}\right], \tag{5.29}$$

donde las s_i^2 se definen en (5.24),

$$s_r^2 = \frac{\sum_{i\in S}(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2}{n-1},$$

y $\bar{M}$ es el tamaño promedio de conglomerado; para estimar la varianza, podemos usar el promedio en la población o en la muestra.

EJEMPLO 5.6 Los datos del archivo coots.dat provienen del trabajo de Arnold (1991) acerca del tamaño y volumen de los huevos de negreta en Minnedosa, Manitoba. En este conjunto de datos, nos fijamos en los volúmenes de una submuestra de huevos en las nidadas (nidos con huevos) con al menos dos huevos disponibles para su medición.

Los datos se grafican en las figuras 5.3-5.5. Los datos de una muestra por conglomerados se pueden graficar de varias formas; a menudo, usted deberá construir más de una gráfica para ver las características de los datos. Como sólo tenemos dos observaciones por nidada, podemos graficar los datos individuales. Si tuviéramos muchas observaciones por nidada, podríamos construir gráficas de bloques adyacentes, con una gráfica de bloque por cada unidad primaria.[2] Regresaremos al tema de la graficación de datos para encuestas complejas en la sección 7.4.

[2]Hicimos una gráfica similar en la figura 4.1 para una muestra estratificada, construimos una gráfica de bloques para cada estrato.

FIGURA 5.3
Gráfica de datos de volúmenes de huevo. Observe la amplia variación que existe entre las medias de una nidada a otra. Esto indica que los huevos dentro de la misma nidada tienden a ser más similares que dos huevos elegidos al azar de nidadas distintas y que los conglomerados no proporcionan tanta información por huevo como lo haría una muestra aleatoria simple de huevos.

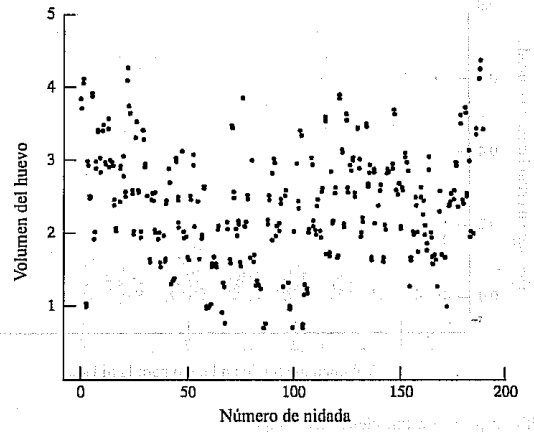


FIGURA 5.4
Otra gráfica de los datos de volumen de los huevos. En este caso, las nidadas se ordenan de la media menor a la media mayor y una recta une las dos medidas de volumen de los huevos en la nidada. La nidada número 88, representada por el segmento largo a la mitad de la gráfica, tiene una gran diferencia, poco usual, entre los dos huevos: un huevo tiene un volumen de 1.85 y el otro de 2.84.

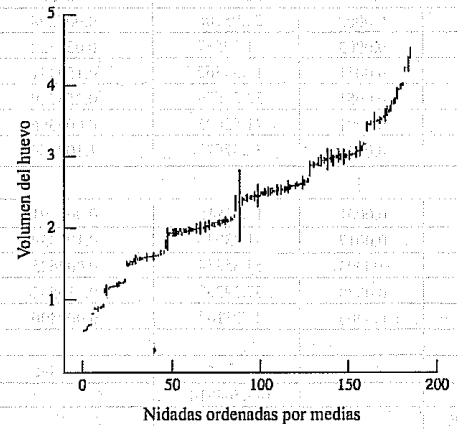


FIGURA 5.5
Una gráfica más para los datos de volumen de los huevos. Esta gráfica muestra la relación entre el volumen medio y la desviación estándar del volumen del huevo dentro de las nidadas. La observación poco usual proviene de la nidada 88. El patrón de montones para las medias merece una investigación más profunda.

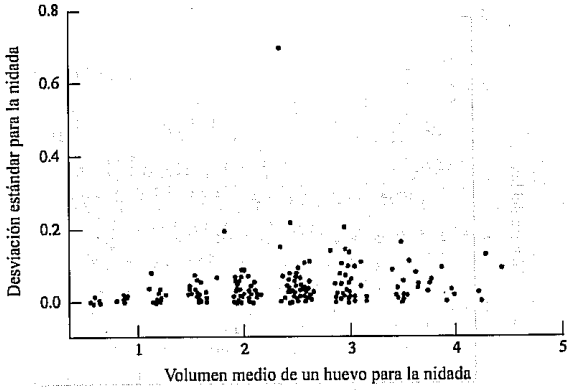


TABLA 5.2
Hoja de cálculo utilizada para las estimaciones del ejemplo 5.6

Nidada	M_i	$\bar{y}_i$	s_i^2	$\hat{t}_i$	$\left(1-\frac{2}{M_i}\right)M_i^2\frac{s_i^2}{m_i}$	$(\hat{t}_i - M_i\hat{\bar{y}}_r)^2$
1	13	3.86	0.0094	50.23594	0.671901	318.9232
2	13	4.19	0.0009	54.52438	0.065615	490.4832
3	6	0.92	0.0005	5.49750	0.005777	89.22633
4	11	3.00	0.0008	32.98168	0.039354	31.19576
5	10	2.50	0.0002	24.95708	0.006298	0.002631
6	13	3.98	0.0003	51.79537	0.023622	377.053
7	9	1.93	0.0051	17.34362	0.159441	25.72099
8	11	2.96	0.0051	32.57679	0.253589	26.83682
9	12	3.46	0.0001	41.52695	0.006396	135.4898
10	11	2.96	0.0224	32.57679	1.108664	26.83682
⋮	⋮	⋮	⋮	⋮	⋮	⋮
180	9	1.95	0.0001	17.51918	0.002391	23.97106
181	12	3.45	0.0017	41.43934	0.102339	133.4579
182	13	4.22	0.00003	54.85854	0.002625	505.3962
183	13	4.41	0.0088	57.39262	0.630563	625.7549
184	12	3.48	0.000006	41.81168	0.000400	142.1994
Suma	1757			4375.947	42.17445	11,439.58
var				149.564814		
$\hat{\bar{y}}_r =$		2.490579				

A continuación, usamos una hoja de cálculo (tabla 5.2) para estimar el resumen de estadísticos para cada nidada. Estos estadísticos se pueden usar, entonces, para estimar el volumen medio de los huevos y su varianza. Hemos redondeado los números para que quepan en la página; en la práctica, por supuesto, usted debe realizar todos los cálculos con la precisión que la máquina que utilice le permita.

Utilizamos el estimador de proporción para determinar el volumen medio de los huevos. En este caso, no podemos usar el estimador insesgado, pues no conocemos K, el número total de huevos de la población. De (5.28),

$$\hat{\bar{y}}_r = \frac{\sum_{i\in S}\hat{t}_i}{\sum_{i\in S}M_i} = \frac{4375.947}{1757} = 2.49.$$

De la hoja de cálculo (tabla 5.2),

$$s_r^2 = \frac{\sum_{i\in S}(\hat{t}_i - M_i\hat{\bar{y}}_r)^2}{n-1} = \frac{11,439.58}{183} = 62.511$$

y $\overline{M}_S = 1757/184 = 9.549$. Así, al emplear (5.29),

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{9.549^2}\left[\left(1-\frac{184}{N}\right)\frac{62.511}{184} + \left(\frac{1}{N}\right)\frac{42.17}{184}\right].$$

Por otro lado, no conocemos N, el número total de nidadas en la población, aunque suponemos que es grande (y sabemos que es mayor que 184). Así, consideramos que la corrección para las poblaciones finitas al nivel primario es 1 y observamos que el segundo término de la varianza estimada será muy pequeño con respecto al primer término. Luego, usamos:

$$\text{EE}(\hat{\bar{y}}_r) = \frac{1}{9.459}\sqrt{\frac{62.511}{184}} = 0.061.$$

El coeficiente estimado de la variación para $\hat{\bar{y}}_r$ es:

$$\frac{\text{EE}(\hat{\bar{y}}_r)}{\hat{\bar{y}}_r} = \frac{0.061}{2.49} = 0.0245. \quad \blacksquare$$

En el ejemplo 5.6, sólo podíamos usar el estimador de proporción, pues no conocíamos N ni K. Sin embargo, las M_i no variaban demasiado, de modo que probablemente el estimador insesgado habría tenido un coeficiente de variación similar. Si todas las M_i son iguales, el estimador insesgado es igual al estimador de proporción (véase el ejercicio 11); si las M_i varían, es frecuente que el estimador insesgado tenga un rendimiento pobre. El siguiente ejemplo ilustra el hecho de que el estimador insesgado de t pueda tener una varianza grande cuando los tamaños de conglomerado son demasiado variables.

EJEMPLO 5.7 *El caso del cachorro con seis patas*

Suponga que queremos estimar el número promedio de patas que pertenecen a los cachorros saludables que se encuentran en las perreras de la apócrifa Ciudad Muestra. Esta ciudad tiene dos perreras: Puppy Palace, y Dog's Life, las cuales alojan a 30 y 10 cachorros, respectivamente. Elegimos una perrera con una probabilidad de 1/2. Después de seleccionar la perrera, elegimos 2 cachorros al azar y usamos $\hat{\bar{y}}_{\text{ins}}$ para estimar el número promedio de patas por cachorro.

Suponga que escogimos a Puppy Palace. No debe sorprendernos que cada uno de los 2 cachorros de la muestra tiene cuatro patas, de modo que $\hat{t}_{\text{pp}} = 30 \times 4 = 120$. Entonces, al

usar (5.21) y (5.26), un estimador insesgado para el número total de patas por cachorro en ambas perreras es:

$$\hat{t}_{\text{ins}} = \frac{2}{1}\hat{t}_{\text{pp}} = 240.$$

Dividimos el total estimado entre el número de cachorros para estimar el número medio de patas por cachorro como 240/40 = 6.

Si hubiésemos elegido a Dog's Life, $\hat{t}_{\text{DL}} = 10 \times 4 = 40$ y

$$\hat{t}_{\text{ins}} = \frac{2}{1}\hat{t}_{\text{DL}} = 80.$$

De haber elegido Dog's Life, el estimador insesgado del número medio de patas por cachorro sería 80/40 = 2.

Éstas no son buenas estimaciones del número de patas por cachorro. Pero el estimador es insesgado desde el punto de vista matemático: (6 + 2)/2 = 4, de modo que el promedio sobre todas las muestras posibles produce el número correcto. La mala calidad del estimador se refleja en la enorme varianza que presenta el estimador, calculada mediante (5.22):

$$V(\hat{t}_{\text{ins}}) = \left(1-\frac{1}{2}\right)2^2\frac{S_t^2}{1} + \frac{2}{1}\sum_{i=1}^{2}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{S_i^2}{m_i}$$

$$= \frac{1}{2}(4)(3200) = 6400.$$

Sin embargo, el estimador de proporción da en el blanco: si se elige a Puppy Palace, $\bar{\bar{y}}_r = 120/30 = 4$; si se escoge a Dog's Life, $\bar{\bar{y}}_r = 40/10 = 4$. Como la estimación es la misma para todas las pruebas posibles, $V(\bar{\bar{y}}_r) = 0$. ■

En general, el estimador insesgado del total de la población es ineficiente si los tamaños de conglomerado son distintos y t_i es, aproximadamente, proporcional a M_i. La varianza de $\hat{t}_{\text{ins}}$ depende de la varianza de t_i y esa varianza puede ser grande si los M_i son distintos.

Sin embargo, el estimador de proporción se comporta bien si t_i es aproximadamente proporcional a M_i. Recuerde de (3.5) que el error cuadrático medio aproximado (ECM) del estimador $\hat{B}$ es proporcional a la varianza de los residuos del modelo: con la notación de este capítulo, el ECM aproximado de $\bar{\bar{y}}_r (= \hat{B})$ es proporcional a $\sum_{i=1}^{N}(t_i - \bar{y}_U M_i)^2$. Cuando t_i (la variable de respuesta) está altamente correlacionada en forma positiva con M_i (la variable auxiliar), los residuos son pequeños. En el ejemplo 5.7, el número total de patas de cachorro en una perrera (t_i) es exactamente cuatro veces el número total de cachorros en la perrera (M_i), de modo que la varianza del estimador de proporción es cero.

Éste es un punto importante, pues gran parte de los conglomerados que aparecen de manera natural tienen distintos tamaños y esperamos que los totales por conglomerados sean proporcionales al número de unidades secundarias. En una muestra por conglomerados de asilos, esperamos que un mayor número de residentes queden satisfechos con la calidad de la atención que proporciona un asilo con 500 residentes que en uno con 20 residentes, aunque las proporciones de residentes satisfechos sean las mismas. El total de calificaciones, en matemáticas, para todos los estudiantes de un grupo será mucho mayor para grupos grandes que para grupos pequeños. En general, esperamos ver más abejas en un área grande que en un área pequeña. Así, para todas estas situaciones, mientras el estimador $\bar{\bar{y}}_r$ funciona bien, la variabilidad del estimador $\hat{t}_{\text{ins}}$ tiende a ser grande. En el capítulo 6, analizaremos un diseño y un estimador alternativos para el muestreo por conglomerados que tendrá una varianza mucho menor para el total de población estimado, cuando t_i es proporcional a M_i.

5.4 Uso de pesos en las muestras por conglomerados

Para estimar las medias y los totales globales de las muestras por conglomerados, la mayor parte de los estadísticos utilizan los pesos de muestreo. Como veremos en las secciones 7.2 y 7.3, los pesos pueden servir para determinar una estimación puntual de casi cualquier cantidad de interés a partir de cualquier diseño de muestreo de probabilidad. Así, los pesos constituyen una herramienta muy valiosa para el análisis de los datos de las encuestas.

Recuerde, del muestreo estratificado, que el peso de un elemento es el recíproco de la probabilidad de su elección. Para el muestreo por conglomerados,

P(seleccionar la unidad secundaria j de la unidad primaria i)
$= P$(seleccionar la unidad primaria i) $\times P$(elegir la unidad secundaria j | se ha seleccionado la unidad primaria i)

$$= \frac{n}{N}\frac{m_i}{M_i}. \tag{5.30}$$

Así,

$$w_{ij} = \frac{NM_i}{nm_i}. \tag{5.31}$$

Por ejemplo, si las unidades primarias son manzanas y las unidades secundarias son familias, entonces la familia j de la manzana i representa a $(NM_i)/(nm_i)$ familias de la población: a sí misma y a $(NM_i)/(nm_i) - 1$ otras familias. Entonces,

$$\hat{t}_{\text{ins}} = \sum_{i\in S}\sum_{j\in S_i} w_{ij}y_{ij} \tag{5.32}$$

y

$$\bar{\bar{y}}_r = \frac{\hat{t}_{\text{ins}}}{\sum_{i\in S}\sum_{j\in S_i} w_{ij}}. \tag{5.33}$$

Observe que $\hat{t}_{\text{ins}}$ es la misma que en (5.21) y que $\bar{\bar{y}}_r$ es la misma que en (5.28). Los pesos de muestreo simplemente proporcionan una forma conveniente de calcular estas estimaciones; no evitan los problemas asociados, como las varianzas grandes. Además, los pesos de muestreo no dan información acerca de cómo determinar los errores estándar; hay que utilizar las fórmulas de este capítulo o un método del capítulo 9.

En el muestreo por conglomerados en dos etapas con un diseño con autoponderación, cada unidad secundaria representa al mismo número de unidades secundarias en la población. Para una muestra autoponderada de personas en Illinois, podríamos tomar una muestra aleatoria simple de condados en Illinois y, luego, tomar una muestra de m_i de las M_i personas del condado i. Para que cada persona de la muestra represente al mismo número de personas de la población, m_i debe ser proporcional a M_i, de modo que m_i/M_i sea aproximadamente constante. Así, en los condados grandes se extraen más personas para la muestra que en los condados pequeños.

EJEMPLO 5.8 En el ejemplo 5.6, los pesos para las observaciones son:

$$\frac{N}{n}\frac{M_i}{m_i} = \frac{N}{184}\frac{M_i}{2}.$$

Como no conocemos N, desplegamos los pesos relativos $M_i/2$ en una hoja de cálculo (tabla

TABLA 5.3
Hoja de cálculo para el cálculo del volumen de los huevos mediante los pesos relativos

nidada	tamaño de la nidada	volumen	peso relativo	peso × volumen
1	13	3.795757	6.5	24.67242
1	13	3.93285	6.5	25.56352
2	13	4.215604	6.5	27.40142
2	13	4.172762	6.5	27.12295
3	6	0.931765	3	2.795294
3	6	0.900736	3	2.702209
4	11	3.018272	5.5	16.6005
4	11	2.978397	5.5	16.38118
⋮	⋮	⋮	⋮	⋮
183	13	4.481221	6.5	29.12794
183	13	4.348412	6.5	28.26468
184	12	3.486132	6	20.91679
184	12	3.482482	6	20.89489
suma	3514		1757	4375.947

5.3). La columna 5 es igual a y_i por el peso relativo; al usar (5.33), $\bar{y}_r = 4375.947/1757 = 2.49$. Sin embargo, los pesos no nos permiten calcular el error estándar; todavía necesitamos emplear (5.29) para esto. ■

5.5 Diseño de una muestra por conglomerados

Las personas y organizaciones que realizan una encuesta costosa a gran escala, deben dedicar mucho tiempo para diseñar un proyecto de este tamaño; por lo general, el diseño y la prueba de las grandes encuestas administradas por la Oficina de Censos de Estados Unidos tarda varios años. Aun en este caso, el principio fundamental del diseño de encuestas sigue siendo válido: usted puede diseñar de mejor manera la encuesta que realizó sólo después de haberla concluido. Al finalizar la encuesta, usted puede evaluar el efecto de los cúmulos sobre las estimaciones y saber dónde debería asignar más recursos para obtener una mejor información.

Mientras más sepa de una población, mejor diseñará un esquema de muestreo eficiente para estudiarlo. Si conoce el valor de y_{ij} para cada persona de la población, entonces, podrá diseñar una encuesta impecable (¡aunque innecesaria, pues usted ya lo sabe todo sobre el objeto de estudio!) para analizar la población. Si sabe poco acerca de la población, es posible que obtenga cierta información sobre ella después de realizar una encuesta, pero es probable que no tenga el diseño más eficiente para llevar a cabo dicha encuesta. Sin embargo, puede utilizar el conocimiento recién adquirido para que la siguiente encuesta sea más eficiente.

Al diseñar una muestra por conglomerados, usted debe resolver cuatro puntos fundamentales:

1 ¿Cuál es la precisión global necesaria?
2 ¿Qué tamaño deben tener las unidades de muestreo primario?

3 ¿Cuántas unidades secundarias deben participar en la muestra, en cada unidad primaria elegida para la muestra?
4 ¿Cuántas unidades primarias deben estar en la muestra?

La pregunta 1 debe contestarse en cualquier diseño de encuesta. Para responder el resto de las preguntas, 2-4, usted debe conocer el costo de muestreo de una unidad primaria para los posibles tamaños de dichas unidades, el costo de muestreo de una unidad secundaria, y una medida de la homogeneidad (R_a^2 o ICC) para los tamaños posibles de unidades secundarias.

5.5.1 Elección del tamaño de la unidad primaria

El tamaño de la unidad primaria es, con frecuencia, una unidad natural. En el ejemplo 5.6, una nidada era, obviamente, una unidad de conglomerado. Una encuesta para estimar la mortandad de los becerros podría usar las granjas como unidad primaria; una encuesta de los estudiantes de sexto grado podría utilizar los grupos o las escuelas como unidad primaria.

Sin embargo, en otras encuestas, el investigador podría tener una amplia gama para elegir la unidad primaria. En una encuesta para estimar la proporción por edad y sexo de unos ciervos que habitan en una región de Colorado (véase un análisis de este problema en Bowden *et al.* 1984), las unidades primarias podrían ser las áreas designadas y las unidades secundarias serían cada uno de los ciervos o grupos de ellos en tales áreas. Pero ¿el tamaño de las unidades primarias debería ser 1 km², 2 km² o 100 m²?

Un principio general en las encuestas de áreas consiste en que mientras mayor sea el tamaño de la unidad primaria, es de esperar que se vea más variabilidad dentro de dicha unidad. Por lo tanto, es de esperar que R_a^2 e ICC sean menores con una unidad primaria grande que con una unidad primaria pequeña. Sin embargo, si el tamaño de la unidad primaria es grande, usted podría perder los beneficios de un menor costo que ofrece el muestreo por comglomerados.

Bellhouse (1984) hace una revisión del diseño óptimo para el muestreo, y la teoría proporciona una útil guía para el diseño de su propia encuesta. Hay formas de "probar" distintos tamaños de unidad primaria antes de realizar la encuesta. Una forma consiste en postular un modelo para la relación entre R_a^2 o el cuadrado de la media dentro de los conglomerados y M y ajustar el modelo mediante datos preliminares o información de otros estudios. Luego, puede utilizar combinaciones diferentes de R_a^2 y M y comparar los costos. Otra forma consiste en desarrollar un experimento y reunir los datos acerca de los costos y las varianzas relativos con distintos tamaños de unidad primaria.

EJEMPLO 5.9 El escarabajo de la papa, en Colorado, es considerado como una plaga mayor por los agricultores. Zehnder *et al.* (1990) estudiaron distintos tamaños de unidades de muestreo que podrían usarse para estimar el número de estos escarabajos. Se eligen 10 sitios al azar de entre 10 campos de papa. Los investigadores inspeccionan visualmente cada sitio en búsqueda de larvas pequeñas, larvas grandes y adultos en un follaje de un único tallo en cada una de las cinco plantas adyacentes.

Luego, los autores consideraron distintos tamaños de unidad primaria, variando de un tallo hasta cinco tallos por sitio. Para estudiar la eficiencia del diseño con un tallo por sitio, examinaron los datos del tallo 1 de cada sitio. De manera análoga, los datos de los tallos 1 y 2 de cada sitio proporcionaron una muestra por conglomerados con dos unidades secundarias en cada unidad primaria, y así sucesivamente. Se necesitan cerca de 30 minutos para caminar entre los sitios de cada campo; el muestreo de un tallo requiere cerca de 10 segundos durante la primera parte de la temporada. Así, el costo total de muestreo de los diez sitios con el diseño de un tallo por sitio se estima como 30 + 100/60 = 31.67 minutos. Los datos para la estimación del número de larvas pequeñas aparecen en la tabla 5.4.

El error estándar relativo se calcula como 1000(EE/costo). Para este ejemplo, como el costo de muestreo de otros tallos en un sitio es pequeño comparado con el tiempo necesario para atravesar el campo, el diseño con cinco tallos por sitio es el más eficiente entre los estudiados. ■

TABLA 5.4
Errores estándar relativos correspondientes al estudio del escarabajo de la papa

Número de tallos por sitio participantes en la muestra	$\hat{\bar{y}}$	$EE(\hat{\bar{y}})$	Costo de muestreo de un campo	Error estándar relativo
1	1.12	0.15	31.67	4.7
2	1.01	0.10	33.33	3.0
3	0.96	0.08	35.00	2.3
4	0.91	0.07	36.67	1.9
5	0.91	0.06	38.33	1.6

5.5.2 Elección de tamaños de submuestreo

El objetivo de diseñar una muestra consiste, por lo general, en obtener más información a un menor costo y con menos inconveniencias. En esta sección nos concentramos en el diseño de una encuesta por conglomerados en dos etapas, donde todos los conglomerados tienen la misma cantidad M de unidades secundarias; el diseño de muestras por conglomerados se analizará de manera general en los capítulos 6 y 7. Un punto de vista para los conglomerados con el mismo tamaño, analizado en Cochran (1977), consiste en minimizar la varianza en (5.22) para un costo fijo. Si $M_i = M$ y $m_i = m$ para todas las unidades primarias, entonces $V(\bar{y}_{ins})$ se puede escribir como (véase el ejercicio 10)

$$V(\hat{\bar{y}}_{ins}) = \left(1-\frac{n}{N}\right)\frac{MSB}{nM} + \left(1-\frac{m}{M}\right)\frac{MSW}{nm}, \qquad (5.34)$$

donde MSB y MSW son los cuadrados de las medias entre y dentro de los conglomerados, respectivamente, en la tabla 5.1, la tabla de análisis de varianza de la población.

Si MSW = 0 y con ello $R_a^2 = 1$, para R_a^2 definido en (5.10), entonces, todos los elementos dentro de un conglomerado tienen el valor de la media del conglomerado. En ese caso, usted puede considerar que $m = 1$; el análisis de más de un elemento por conglomerado sólo cuesta tiempo y dinero adicionales, sin aumentar la precisión. Para otros valores de R_a^2, la distribución óptima depende de los costos relativos de muestreo de las unidades primarias y secundarias.

Si consideramos la fórmula de costo sencilla, la cual es:

$$\text{costo total} = C = c_1 n + c_2 nm, \qquad (5.35)$$

donde c_1 es el costo por unidad primaria (sin incluir el costo por medir las unidades secundarias) y c_2 es el costo por medir cada unidad secundaria. Se puede determinar fácilmente, mediante el cálculo, que los valores

$$n = \frac{C}{c_1 + c_2 m}$$

y

$$m = \sqrt{\frac{c_1 M(MSW)}{c_2(MSB - MSW)}} = \sqrt{\frac{c_1 M(N-1)}{c_2(NM-1)}\left(\frac{1}{R_a^2}-1\right)}$$

minimizan la varianza para un costo total fijo C bajo esta función de costo (véanse los ejercicios 10 y 23); sin embargo, varios valores distintos de m podrían, con frecuencia, servir también; la graficación de la varianza proyectada de la estimación dará más información que el cálculo de una solución fija. Un punto de vista gráfico también le permite desarrollar el siguiente análisis sobre los diseños: ¿Qué ocurre si los costos o la función de costo son ligeramente distintos? ¿O si el valor de R_a^2 se modifica ligeramente? Usted puede analizar distintas funciones de costo con este enfoque.

EJEMPLO 5.10 ¿Será que el submuestreo es más eficiente para el ejemplo 5.2 que el muestreo por conglomerados en una etapa? No conocemos las cantidades de la población, pero tenemos información de la muestra que puede servir para planear los futuros estudios. Recuerde que $\hat{S}^2 = 0.279$ y que estimamos R_a^2 como 0.337. Las figuras 5.6 y 5.7 muestran la varianza estimada que se obtendría para distintos tamaños de la submuestra y para distintos valores de c_1, c_2 y R_a^2. ■

FIGURA 5.6
La varianza estimada que se obtendría para el ejemplo de las calificaciones, para distintos valores de c_1, c_2 y m. La estimación muestral R_a^2 es 0.337. El costo total C es 300 para esta gráfica. Si se necesitan 40 minutos por cuarto y 5 minutos por persona, entonces, hay que usar el muestreo por conglomerados en una etapa; si se necesitan 10 minutos por cuarto y 20 minutos por persona, entonces, sólo se debe incluir en la muestra a una persona por cuarto; si se necesitan 20 minutos por cuarto y 10 minutos por persona, el mínimo se alcanza en $m \approx 2$, aunque lo plano de la curva indica que cualquier tamaño del submuestreo sería aceptable.

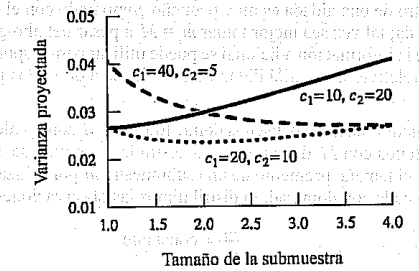
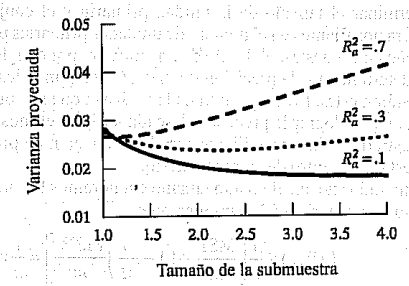


FIGURA 5.7
Varianza estimada que se obtendría para el ejemplo de las calificaciones, para distintos valores de R_a^2 y m. Los costos utilizados para construir esta gráfica son $C = 300$, $c_1 = 20$ y $c_2 = 10$. Al crecer el valor de R_a^2, el tamaño de la submuestra m debe decrecer.



Para los fines del diseño de la muestra, sólo necesitamos una estimación gruesa de R_a^2; por lo general, la R^2 ajustada que pertenece a la tabla de análisis de la varianza obtenida a partir de los datos de la muestra, proporciona un buen punto de partida, aunque a menudo, el valor muestral del total del cuadrado de la media subestima a S^2, cuando el número de unidades primarias en la muestra es pequeño.

EJEMPLO 5.11 He aquí la tabla de análisis de la varianza de la muestra para los datos de las negretas, calculada mediante SAS.

Fuente	GL	Suma de cuadrados	Cuadrado de la media	Valor F
Modelo	183	257.4175336	1.4066532	237.44
Error	184	1.09007825	0.0059243	
Total corregido	367	258.5076118		

R-cuadrada	C.V.	Raíz del ECM	Media de VOLUMEN
0.995783	3.298616	0.076970	2.333394

Si una encuesta futura se planea para estimar el volumen promedio de los huevos, uno podría explorar los tamaños de las submuestras al usar R_a^2 en torno a $1-0.0059243/(258.5/367)=0.99$. Estos datos indican un alto grado de homogeneidad dentro de las nidadas para el volumen de los huevos. Sin embargo, para esta encuesta, el costo marginal por medir otros huevos dentro de una nidada es muy pequeño comparado con el costo de ubicación y acceso a una nidada; tal vez sea mejor tomar $m_i = M_i$ a pesar del alto grado de homogeneidad, debido a que la información adicional se puede utilizar para responder otras cuestiones de investigación relativas a la variabilidad de una nidada a otra o los posibles efectos de *// laying-sequence.* ■

Aunque sólo hemos analizado diseños donde todas las M_i son iguales, también podemos utilizar estos métodos con M_i distintos: basta sustituir $\bar{M}$ en vez de M en nuestro trabajo anterior y decidir el tamaño promedio de una submuestra $\bar{m}$ por utilizar. Luego tomamos $\bar{m}$ observaciones en cada conglomerado o distribuimos las observaciones de modo que

$$\frac{m_i}{M_i} = \text{constante}.$$

Mientras las M_i no varíen demasiado, esto debe producir un diseño razonable. Si las M_i varían bastante y las t_i están correlacionadas con las M_i, una muestra por conglomerados con probabilidades iguales no necesariamente es muy eficiente; en el capítulo 6 presentamos un diseño alternativo.

5.5.3 Elección del tamaño de la muestra (número de unidades primarias)

Después de determinar el tamaño de la unidad primaria y el conjunto o proporción de submuestreo, ahora nos fijamos en el número de unidades primarias por muestrear, n. Como en cualquier diseño de encuestas, el diseño de una muestra por conglomerados es un proceso iterativo: (1) determinamos la precisión deseada, (2) elegimos los tamaños de la unidad primaria y de la submuestra, (3) conjeturamos la varianza que se obtendrá con este diseño, (4) determinamos n para lograr la precisión deseada y (5) repetimos el proceso (al agregar las variables de estratificación y auxiliares para la estimación de proporción) hasta que el costo de la encuesta esté dentro de su presupuesto.

Si los conglomerados tienen el mismo tamaño e ignoramos la corrección para las poblaciones finitas al nivel primario, (5.34) implica que:

$$V(\hat{\bar{y}}_{\text{ins}}) \le \frac{1}{n}\left[\frac{\text{MSB}}{M} + \left(1-\frac{m}{M}\right)\frac{\text{MSW}}{m}\right] = \frac{1}{n}v.$$

Un intervalo de confianza aproximado del $100(1-\alpha)\%$ será el siguiente:

$$\hat{\bar{y}}_{\text{ins}} \pm z_{\alpha/2}\sqrt{\frac{1}{n}v}.$$

Así, para lograr un intervalo de confianza deseado cuya mitad de ancho sea e, hacemos $n = z_{\alpha/2}^2 v/e^2$. Por supuesto, este punto de vista presupone que se tiene cierto conocimiento de v, tal vez por una encuesta anterior. En la sección 7.5 analizaremos la forma de determinar los tamaños de muestra para cualquier situación en la que se conozca la eficiencia del diseño dado, en relación con un diseño para una muestra aleatoria simple.

5.6 Muestreo sistemático

El muestreo sistemático, analizado brevemente en el capítulo 2, es en realidad un caso particular del muestreo por conglomerados. Suponga que queremos tomar una muestra de tamaño 2 de una población con 12 elementos:

1 2 3 4 5 6 7 8 9 10 11 12

Para tomar una muestra sistemática, elegimos un número al azar entre 1 y 4. Extraemos ese elemento y el cuarto a partir de él. Así, la población contiene cuatro unidades primarias (son conglomerados, aunque los elementos no sean adyacentes):

$$\{1, 5, 9\} \quad \{2, 6, 10\} \quad \{3, 7, 11\} \quad \{4, 8, 12\}.$$

Ahora extraemos una muestra aleatoria simple de una unidad primaria.

En una población con NM elementos, existen N elecciones posibles para la muestra sistemática, cada una de tamaño M. Sólo observamos la media del conglomerado correspondiente a nuestra muestra sistemática,

$$\bar{y}_i = \bar{y}_{iU} = \hat{\bar{y}}_{\text{sis}}.$$

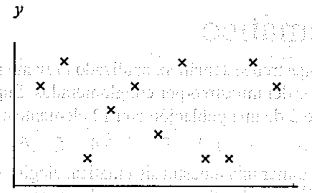
Porque el muestreo por conglomerados en una etapa con conglomerados del mismo tamaño produce estimaciones insesgadas, $E[\hat{\bar{y}}_{\text{sis}}] = \bar{y}_U$. Para una muestra sistemática simple, elegimos $n=1$ de los N conglomerados, de modo que por (5.5) y (5.9), la varianza teórica es:

$$\begin{aligned} V(\hat{\bar{y}}_{\text{sis}}) &= \left(1-\frac{1}{N}\right)\frac{S_t^2}{M^2} \\ &= \left(1-\frac{1}{N}\right)\frac{\text{MSB}}{M} \\ &\approx \frac{S^2}{M}[1+(M-1)\text{ICC}]. \end{aligned} \tag{5.36}$$

Con la notación del muestreo por conglomerados, M es el tamaño de la muestra sistemática. Si ignoramos la corrección para las poblaciones finitas, vemos que el muestreo sistemático es más preciso que una muestra aleatoria simple de tamaño M si el ICC es negativo. El muestreo sistemático es más preciso que el muestreo aleatorio simple cuando la varianza dentro de las posibles muestras sistemáticas (conglomerados) es *mayor* que la varianza general de la población (en ese caso, las medias de los conglomerados son más similares). Si existe poca variación dentro de las muestras sistemáticas con respecto a la población correspondiente (es decir, ICC > 0), entonces todos los elementos de la muestra dan una información similar y es de esperar que el muestreo sistemático tenga una varianza mayor que una muestra aleatoria simple.

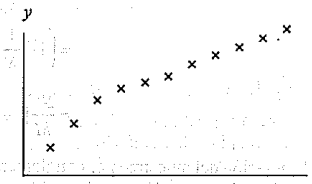
Sin embargo, como $n = 1$, no podemos obtener una estimación insesgada de $V(\hat{\bar{y}}_{sis})$; necesitamos saber algo de la estructura de la población para estimar la varianza. Veamos tres estructuras de población distintas:

1 *La lista tiene un orden aleatorio.* Es probable que el muestreo sistemático produzca una muestra que se comporte como una muestra aleatoria simple. En muchos casos, el orden de la población no está relacionado con las características de interés, como cuando la lista de personas en el marco de muestreo está en orden alfabético. No hay razón para creer que las personas de una muestra sistemática serán más o menos similares a las de una muestra aleatoria de personas; esperamos que ICC ≈ 0. En este caso, el muestreo aleatorio simple y el sistemático darán resultados similares. Podemos utilizar los resultados y las fórmulas de las muestras aleatorias simples para estimar $V(\hat{\bar{y}}_{sis})$.



Posición en el marco de muestreo

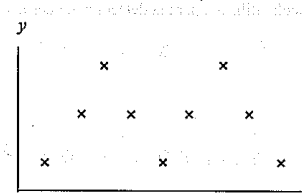
2 *El marco de muestreo tiene un orden creciente o decreciente.* Es probable que el muestreo sistemático sea más preciso que el muestreo aleatorio simple. Los registros financieros pueden estar enumerados con las cantidades más grandes al principio y las más pequeñas al final. Decimos que tal población tiene una **autocorrelación positiva**: los elementos adyacentes tienden a ser más similares que los elementos más alejados. En este caso, $V(\hat{\bar{y}}_{sis})$ es menor que la varianza de la media muestral en una muestra aleatoria simple del mismo tamaño, pues ICC < 0. Una muestra sistemática obliga a que los valores de la muestra se dispersen; es posible que una muestra aleatoria simple conste de todos los valores menores o de todos los valores mayores. Cuando el marco tiene un orden creciente o decreciente, usted puede usar la fórmula de la muestra aleatoria simple para el error estándar, pero es probable que sea una sobrestimación y que los intervalos de confianza construidos mediante el error estándar de una muestra aleatoria simple sean demasiado anchos.



Posición en el marco de muestreo

El muestreo estratificado puede funcionar mejor que el muestreo sistemático para las poblaciones con autocorrelación positiva: si el inicio aleatorio está cerca de cualquiera de los extremos del intervalo de muestreo, una muestra sistemática tenderá a dar una estimación demasiado pequeña o demasiado grande.

3 *El marco de muestreo tiene un patrón periódico.* Si extraemos una muestra con el mismo intervalo que la periodicidad, el muestreo sistemático será menos preciso que el muestreo aleatorio simple. El muestreo sistemático es más riesgoso cuando la población tiene un orden cíclico o periódico y el intervalo de muestreo coincide con un múltiplo del periodo.



Posición en el marco de muestreo

Suponga que los valores de la población (en orden) son:

1 2 3 1 2 3 1 2 3 1 2 3

y que el intervalo de muestreo es 3. Entonces, todos los elementos de la muestra sistemática serán iguales; si usamos la fórmula de la muestra aleatoria simple para estimar la varianza, tendremos que $\hat{V}(\hat{\bar{y}}_{sis}) = 0$. Pero el valor real de $V(\hat{\bar{y}}_{sis})$ para esta población es 2/3; esta muestra no es más precisa que una única observación elegida al azar entre la población.

El muestreo sistemático se utiliza, por lo regular, cuando un investigador desea una muestra representativa de la población, pero no tiene los recursos para construir un marco de muestreo de antemano. Se emplea, comúnmente, para elegir los elementos de la etapa inferior de una muestra por conglomerados. En muchas situaciones donde se usa el muestreo sistemático, la muestra sistemática se puede considerar como una muestra aleatoria simple.

EJEMPLO 5.12 *Muestreo para sitios de desechos tóxicos*

Muchos basureros y rellenos sanitarios que se encuentran en Estados Unidos contienen materiales tóxicos. Estos materiales pueden estar dentro de los recipientes sellados al ser depositados, pero ahora se sospecha que presenten un derrame. Ahora, ya no sabemos dónde fueron depositados esos materiales; los recipientes con desechos tóxicos pueden estar distribuidos al azar en el relleno sanitario o estar concentrados en un área o, incluso, no existir.

Una práctica común consiste en realizar una muestra sistemática de puntos de una retícula y tomar muestras del suelo de cada punto para buscar evidencia de contaminación. Elija un punto al azar en el área y luego construya una retícula que contenga ese punto, de modo que los puntos de la retícula sean equidistantes. Una de estas retículas aparece en la figura 5.8. Las ventajas de tomar una muestra sistemática y no una muestra aleatoria simple son que la primera obliga a cubrir de manera uniforme la región y que es más fácil de establecer. Si a usted no le preocupan los patrones periódicos en la distribución de los materiales tóxicos y tiene poco conocimiento anterior de la ubicación de tales materiales, una muestra sistemática es un buen diseño.

Con cualquier retícula en el muestreo sistemático, usted debe preocuparse por el hecho de que los materiales están dispuestos de manera regular y por ello la retícula no los consiga ubicar, como muestra la figura 5.9. Si esto le preocupa, será mejor tomar una muestra estratificada. Establezca la retícula, pero elija un punto al azar en cada cuadrado para tomar la muestra de suelo. ■

Si la periodicidad es una preocupación sobre la población, una solución consiste en usar las **muestras sistemáticas con interpenetración** (Mahalanobis 1946). En vez de tomar una sola muestra sistemática de la población, tome varias. Entonces, podrá usar las fórmulas para las muestras por conglomerados y estimar las varianzas; cada muestra sistemática actúa como un conglomerado (este punto de vista se explora en el ejercicio 22).

FIGURA 5.8
Una retícula utilizada para la detección de desechos tóxicos.

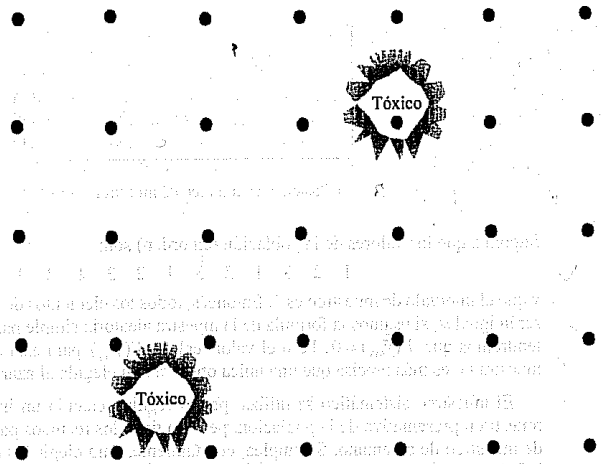
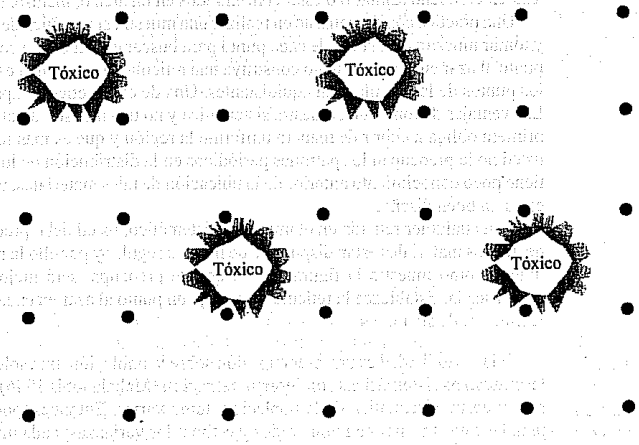


FIGURA 5.9
Una retícula utilizada para la detección de desechos tóxicos: el peor de los escenarios. Como los desechos aparecen con un patrón similar al de la retícula, la muetra sistemática omite cada depósito de desechos tóxicos.



5.7
Modelos para el muestreo por conglomerados*

El modelo de análisis de varianza de un sentido con efectos fijos proporciona un marco de referencia teórico para el muestreo estratificado; un modelo similar que es posible emplear para el muestreo por conglomerados es el modelo de análisis de la varianza de un sentido con efectos aleatorios (Scott y Smith 1969). Analicemos una versión sencilla de este modelo:

$$\text{M1:}\quad Y_{ij} = A_i + \varepsilon_{ij}, \tag{5.37}$$

donde A_i es generada por una distribución con media μ y varianza σ_A^2, ε_{ij} es generada por una distribución con media 0 y varianza σ^2 y todas las A_i y ε_{ij} son independientes.

El modelo M1 implica que el total esperado para un conglomerado crece linealmente con el número de elementos en el conglomerado, pues $E_{\text{M1}}[Y_{ij}] = \mu$ y

$$E_{\text{M1}}[T_i] = E_{\text{M1}}\left[\sum_{j=1}^{M_i} Y_{ij}\right] = M_i\mu.$$

Con frecuencia, este supuesto es adecuado para las muestras por conglomerados en la práctica. Imagine que estamos considerando una muestra por conglomerados en dos etapas para estimar el costo total de un nacimiento en cierto hospital; se eligen los hospitales en la primera etapa y los registros de nacimiento en la segunda etapa (los gemelos y triates cuentan como un registro). Como ilustración, supongamos que μ, el costo promedio nacional por la atención de un nacimiento en un hospital, es cercano a los $10,000. Esperamos que el total de cobros de un hospital sea mayor si el hospital atiende más partos.

Sin embargo, el costo promedio por parto varía de un hospital a otro, pues algunos hospitales pueden tener más gastos por personal, y otros pueden atender a la población de alto riesgo o tener un equipo más caro. Esa variación se refleja en el modelo mediante los efectos aleatorios A_i: A_i es la variable aleatoria que representa el costo promedio por nacimiento en el hospital i y σ_A^2 es la varianza de la población entre las medias de los hospitales. Además, los costos varían de un nacimiento a otro dentro de los hospitales; esa variación se incorpora al modelo mediante el término ε_{ij} con varianza σ^2. Estas ideas se ilustran en la figura 5.10, al suponer que los A_i y los ε_{ij} tienen una distribución normal.

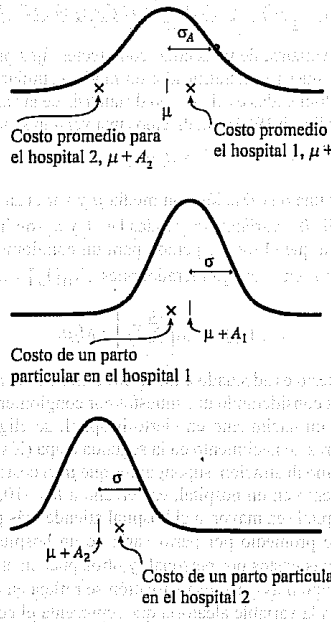
La figura 5.10 ilustra que, de acuerdo con el modelo en (5.37), los costos por los nacimientos en el mismo hospital tienden a ser más similares que los costos por nacimiento elegidos al azar entre toda la población de nacimientos en los hospitales, pues el costo de un nacimiento en un hospital dado incorpora las características del propio hospital, como los costos de personal o la relación enfermera/paciente. El coeficiente de correlación entre las clases para el modelo M1 se define como:

$$\rho = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2}. \tag{5.38}$$

Observe que ρ en el modelo M1 no siempre es negativa, en contraste con el ICC que puede asumir valores negativos.[3] Así, si el modelo M1 describe los datos, el muestreo por conglomerados *debe* ser menos eficiente que una muestra aleatoria simple del mismo

[3] El modelo M1, con $\rho \geq 0$, no sería adecuado si existiera competencia dentro de los conglomerados, de modo que un miembro de un conglomerado se beneficie a expensas de otro. Por ejemplo, si se pueden descontar otros factores ambientales, la competencia dentro del útero podría hacer que algunos gemelos sean más variables que los hermanos que no son gemelos.

FIGURA 5.10
Una ilustración de los efectos aleatorios para los hospitales y los partos.



tamaño. Con el modelo M1,

$$\mathrm{Cov}_{\mathrm{M1}}[Y_{ij}, Y_{kl}] = \begin{cases} \sigma^2 + \sigma_A^2 & \text{si } i = k \text{ y } j = l. \\ \sigma_A^2 & \text{si } i = k \text{ y } j \neq l. \\ 0 & \text{si } i \neq k. \end{cases}$$

5.7.1 Estimación mediante modelos

Ahora veremos las propiedades de varias estimaciones bajo el modelo M1. Para ahorrarnos un poco de trabajo posterior, analizaremos un estimador lineal general de la forma

$$\hat{T} = \sum_{i \in S} \sum_{j \in S_i} b_{ij} Y_{ij}$$

donde b_{ij} son cualquier tipo de constantes. La variable aleatoria que representa al total de la población finita es:

$$T = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij}.$$

Entonces, el sesgo es:

$$E_{\mathrm{M1}}[\hat{T} - T] = E_{\mathrm{M1}}\left[\sum_{i \in S} \sum_{j \in S_i} b_{ij} Y_{ij} - \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij}\right]$$
$$= \mu\left(\sum_{i \in S} \sum_{j \in S_i} b_{ij} - K\right).$$

Así, $\hat{T}$ es insesgado con respecto al modelo cuando $\sum_{i \in S} \sum_{j \in S_i} b_{ij} = K$. La varianza de $\hat{T} - T$ basada en el modelo (para el modelo M1) es:

$$V_{\mathrm{M1}}[\hat{T} - T] = \sigma_A^2 \left[\sum_{i \in S}\left(\sum_{j \in S_i} b_{ij} - M_i\right)^2 + \sum_{i \in S} M_i^2\right]$$
$$+ \sigma^2\left[\sum_{i \in S} \sum_{j \in S_i} (b_{ij}^2 - 2b_{ij}) + K\right]. \qquad (5.39)$$

(Véase el ejercicio 26.)

Ahora, veamos qué ocurre con los estimadores basados en el diseño, bajo el modelo M1. La variable aleatoria para el estimador insesgado con respecto al diseño es:

$$\hat{T}_{\mathrm{ins}} = \sum_{i \in S} \sum_{j \in S_i} \frac{NM_i}{nm_i} Y_{ij};$$

los coeficientes b_{ij} son, simplemente, los pesos de muestreo $(NM_i)/(nm_i)$. Pero

$$\sum_{i \in S} \sum_{j \in S_i} b_{ij} = \frac{N}{n} \sum_{i \in S} \sum_{j \in S_i} \frac{M_i}{m_i} = \frac{N}{n} \sum_{i \in S} M_i,$$

de modo que el sesgo bajo el modelo (5.37) es:

$$\mu\left(\frac{N}{n} \sum_{i \in S} M_i - K\right).$$

Observe que el sesgo depende de la muestra extraída y el estimador es insesgado con respecto al modelo bajo (5.37) sólo cuando el promedio de las M_i en la muestra es igual al promedio de las M_i en la población, como ocurrirá cuando todas las M_i sean iguales. Este resultado nos ayuda a entender por qué el estimador insesgado, con respecto al diseño, funciona tan mal cuando los totales por conglomerado son casi proporcionales a los tamaños de los conglomerados. Es un estimador pobre para un modelo que describe la población.

Para el estimador de proporción, los coeficientes son $b_{ij} = K(M_i/m_i)/\sum_{k \in S} M_k$ y

$$\hat{T}_r = \frac{K \sum_{i \in S} \sum_{j \in S_i} \frac{M_i}{m_i} Y_{ij}}{\sum_{k \in S} M_k}.$$

Para estos b_{ij},

$$\sum_{i \in S} \sum_{j \in S_i} b_{ij} = \sum_{i \in S} \sum_{j \in S_i} \frac{KM_i}{m_i \sum_{k \in S} M_k} = K,$$

de modo que el estimador de proporción es insesgado con respecto al modelo M1. Si el

modelo M1 describe la población, entonces, el estimador de proporción se ajusta para los tamaños de las unidades primarias elegidas en la muestra; utiliza M_i, una cantidad correlacionada con el total de la unidad primaria i, para compensar la posibilidad de que la muestra tenga una proporción de unidades primarias grandes distinta a la de la población.

La expresión de la varianza en (5.39) es compleja; si $M_i = M$ y $m_i = m$ para toda i, entonces $\hat{T}_{\text{ins}} = \hat{T}_r, b_{ij} = (NM)/(nm)$ y la varianza en (5.39) se simplifica como

$$V_{\text{M1}}[\hat{T}_{\text{ins}} - T] = KM(N-n)\frac{\sigma_A^2}{n} + K(MN-mn)\frac{\sigma^2}{mn}. \tag{5.40}$$

EJEMPLO 5.13 Regresemos a las perreras analizadas del ejemplo 5.7. Ciertamente, sigue el modelo M1: todos los cachorros tienen cuatro patas, de modo que $Y_{ij} = \mu = 4$ para toda i, j. En consecuencia, $\sigma_A^2 = \sigma^2 = 0$. Por tanto, la varianza de la estimación $\hat{T}_{\text{ins}}$ basada en el modelo es cero, sin importar la perrera o los cachorros elegidos. Si se elige Puppy Palace para la muestra, el sesgo bajo el modelo (5.37) es $4(2 \times 30 - 40) = 80$; si se escoge Dog's Life, el sesgo es $4\,(2\times10-40) = -80$. La varianza tan grande en el enfoque basado en el diseño se convierte, entonces, en un sesgo al adoptar un enfoque basado en el modelo. No debe sorprendernos que $\hat{T}_{\text{ins}}$ tenga un mal desempeño para las perreras; es un estimador pobre para un modelo que describe bien la situación. Sin embargo, el sesgo y la varianza para $\hat{T}_r$ son iguales a cero. ■

Los resultados anteriores son sólo para el modelo M1. Suponga que un mejor modelo para la población es el siguiente:

$$\text{M2}: \quad Y_{ij} = B_i + \varepsilon_{ij}, \tag{5.41}$$

con $E[B_i] = \mu/M_i, V[M_iB_i] = \sigma_B^2, E[\varepsilon_{ij}] = 0, V[\varepsilon_{ij}] = \sigma^2$, y todos los B_i y ε_{ij} son independientes. Bajo el modelo M2, todos los totales de conglomerado tienen el valor esperado μ, sin importar el tamaño del conglomerado. Los ejemplos descritos por este modelo son más difíciles de aparecer en la práctica, pero construiremos uno con base en el principio de que las tareas se alargan hasta ocupar todo el tiempo asignado para ellas. Todos los estudiantes de la preparatoria El Edén disponen de 100 horas para escribir un artículo para una materia, pero un estudiante debe escribir de uno a cinco artículos. Nunca ha ocurrido que un alumno de El Edén termine un artículo rápidamente y descanse en un tiempo de sobra, de modo que un estudiante con un artículo ocupa las 100 horas para elaborarlo, un alumno con dos artículos ocupa 50 horas en cada uno, etcétera. Así, la cantidad total esperada de tiempo ocupado en escribir los artículos, $E[T_i]$, es 100 para cada estudiante, aunque el número de artículos asignados (M_i) varía.

El estimador $\hat{T}_{\text{ins}}$ es insesgado bajo el modelo M2:

$$E_{\text{M2}}[\hat{T}_{ins} - T] = E_{\text{M2}}\left[\sum_{i\in S}\sum_{j\in S_i}\frac{NM_i}{nm_i}Y_{ij} - \sum_{i=1}^{N}\sum_{j=1}^{M_i}Y_{ij}\right]$$

$$= \sum_{i\in S}\sum_{j\in S_i}\frac{NM_i}{nm_i}\frac{\mu}{M_i} - \sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{\mu}{M_i} = 0.$$

Así, el desempeño de $\hat{T}_{\text{ins}}$ es pobre si el modelo (5.37) es adecuado, aunque con frecuencia es bueno si el modelo (5.41) también es adecuado. Por supuesto, éstos no son los únicos modelos posibles: Royall (1976a) dedujo los resultados para una clase general de modelos posibles que incluyen a (5.37) y (5.41), y permiten varianzas distintas para conglomerados distintos.

Si decide utilizar un punto de vista basado en el modelo para analizar los datos de una muestra por conglomerados, tenga cuidado en verificar que el modelo elegido es el adecuado. En el ejemplo de los cachorros vimos que la varianza del modelo M1 para $\hat{T}_{\text{ins}}$ es cero, pero el sesgo es grande; sin embargo, sólo pudimos evaluar el sesgo porque conocíamos los resultados para toda la población. Una persona que sólo obtuvo una muestra del Puppy Palace y no conocía los resultados de Dog's Life no podría evaluar el sesgo y concluir que los cachorros promedian ¡seis patas cada uno! Así, la evaluación de lo adecuado del modelo es crucial en cualquier análisis basado en el modelo. Usted debe verificar el supuesto de que $V[\varepsilon_{ij}] = \sigma^2$ al graficar las varianzas de cada conglomerado, al igual que cuando verifica el supuesto de la misma varianza en un análisis de varianza. A menudo, una gráfica de $\hat{\imath}_i$ contra M_i es útil para evaluar lo adecuado de un modelo para los datos de la muestra. Como ocurre siempre en la inferencia basada en un modelo, debemos suponer que el modelo también es válido para los elementos de la población que no están en la muestra.

EJEMPLO 5.14 Ajustemos el modelo M1, el de un sentido con efectos aleatorios, a los datos de las negretas. Al observar las figuras 5.4 y 5.5, parece plausible (excepto por una nidada) que la varianza dentro de las nidadas es la misma para cada una de ellas. La figura 5.11 muestra la gráfica de $\hat{\imath}_i$ contra M_i para los datos de las negretas.

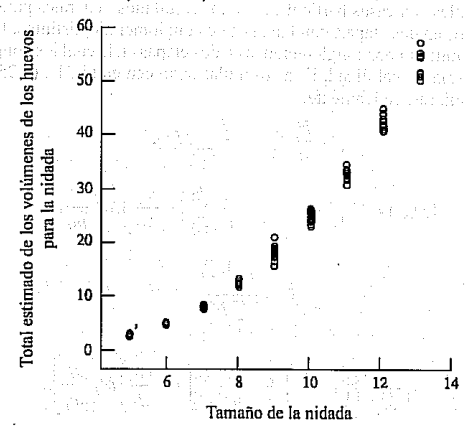
Para estos datos, $\text{Corr}(\hat{\imath}_i, M_i) = 0.97$. Si el modelo M1 es adecuado para los datos, esperamos que $\hat{\imath}_i$ aumente con M_i; si el modelo M2 es adecuado, esperamos que la recta horizontal se ajuste a los puntos graficados. Para estos datos, es claro que $\hat{\imath}_i$ y M_i están relacionados, aunque no parece que la relación sea una línea recta.

Si usamos SAS Proc Mixed, los componentes de la varianza estimada son $\hat{\sigma}_A^2 = 0.70036$ y $\hat{\sigma}^2 = 0.00592$. Si empleamos $b_{ij} = M_i/(m_i\sum_{k\in S}M_k)$, el volumen medio estimado de un huevo es 2.492196; al adaptar (5.39) para ignorar la corrección para las poblaciones finitas (véase el ejercicio 26), la varianza estimada basada en el modelo es:

$$\sum_{i\in S}\left(\frac{M_i}{\sum_{k\in S}M_k}\right)^2\hat{\sigma}_A^2 + \sum_{i\in S}\frac{1}{m_i}\left(\frac{M_i}{\sum_{k\in S}M_k}\right)^2\hat{\sigma}^2 = 0.003944 + 0.000017 = 0.00396.$$

Si se hubiese adoptado otro modelo, la varianza estimada sería diferente. ■

FIGURA 5.11
La gráfica de $\hat{\imath}_i$ contra M_i para los datos de las negretas



5.7.2 Diseño mediante modelos

Los modelos son extremadamente útiles para diseñar una muestra por conglomerados. El uso de un modelo para el diseño no significa que deba emplearse un modelo para el análisis de los datos de una encuesta cuando se vaya a realizar esta última, ya que el modelo proporciona una forma útil de resumir la información que puede servir para que la encuesta sea más eficiente. Se han realizado muchas investigaciónes sobre el uso de modelos para el diseño; consulte las revisiones de la bibliografía en Rao (1979b), Bellhouse (1984) y Royall (1992b).

Suponga que el modelo M1 parece razonable para la población que desea investigar y que todos los tamaños de las unidades primarias que pertenecen a la población son iguales. Entonces, usted quisiera diseñar la encuesta con el fin de minimizar la varianza en (5.40), sujeta a las restricciones de los costos. Entonces, al usar la función de costos en (5.35), la varianza basada en el modelo se minimiza cuando:

$$m=\sqrt{\frac{c_1\sigma^2}{c_2\sigma_A^2}}.$$

Imagine que las M_i no son iguales y que el modelo M1 es válido. Podemos utilizar la varianza en (5.39) para determinar el tamaño óptimo de submuestreo m_i para cada conglomerado. Este enfoque fue utilizado por Royall (1976a) para modelos más generales que los considerados en esta sección. Para $\hat{T}_r, b_{ij}=KM_i/(m_i\sum_{k\in S}M_k)$ y la varianza se minimiza cuando m_i es proporcional a M_i (véase el ejercicio 28).

5.8 Resumen

El muestreo por conglomerados se utiliza, por lo general, en las encuestas de gran tamaño, pero las estimaciones obtenidas de estas muestras tienen, con frecuencia, una varianza mayor que la obtenida si se pudiese medir el mismo número de unidades de observación mediante una muestra aleatoria simple. Sin embargo, si es menos caro tomar muestras de conglomerados que de los elementos individuales, el muestreo por conglomerados puede proporcionar una mayor precisión a un menor costo.

Todas las fórmulas de este capítulo para el muestreo por conglomerados con probabilidades iguales son casos particulares de los resultados generales para el muestreo por conglomerados en dos etapas, con tamaños de conglomerado distintos. Esto se puede aplicar a cualquier muestra por conglomerados en dos etapas en la cual los conglomerados se escojan con la misma probabilidad. Estas fórmulas aparecen en (5.21), (5.25), (5.28) y (5.29) y se repiten aquí, respectivamente:

$$\hat{t}_{\text{ins}}=\frac{N}{n}\sum_{i\in S}\hat{t}_i=\frac{N}{n}\sum_{i\in S}M_i\bar{y}_i, \tag{5.21}$$

$$\hat{V}(\hat{t}_{\text{ins}})=N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}+\frac{N}{n}\sum_{i\in S}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}, \tag{5.25}$$

$$\hat{\bar{y}}_r=\frac{\sum_{i\in S}M_i\bar{y}_i}{\sum_{i\in S}M_i}, \tag{5.28}$$

$$\hat{V}(\hat{\bar{y}}_r)=\left(\frac{1}{M^2}\right)\left[\left(1-\frac{n}{N}\right)\frac{s_r^2}{n}+\frac{1}{nN}\sum_{i\in S}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}\right], \tag{5.29}$$

con

$$s_t^2=\frac{\sum_{i\in S}\left(\hat{t}_i-\frac{\hat{t}_{\text{ins}}}{N}\right)^2}{n-1}$$

y

$$s_r^2=\frac{\sum_{i\in S}(M_i\bar{y}_i-M_i\hat{\bar{y}}_r)^2}{n-1}.$$

Para el muestreo por conglomerados en una etapa, $m_i=M_i$, de modo que el segundo término en (5.25) y (5.29) se anula. De hecho, las fórmulas para el muestreo estratificado son también un caso particular de las que se emplean en este capítulo: para el muestreo estratificado, $n=N$ y obtenemos una muestra de m_i observaciones de las M_i observaciones del estrato i.

En la práctica, las estimaciones puntuales de la media y el total de la población se calculan usualmente mediante pesos. Usted necesita usar las fórmulas anteriores o un método como el de la navaja de bolsillo del capítulo 9, para el cálculo de los errores estándar.

5.9 Ejercicios

1 El ayuntamiento de una pequeña ciudad quiere saber la proporción de votantes que se oponen a tener un quemador de basura de la compañía Phoenix, justo fuera de los límites de la ciudad. Ellos eligen al azar 100 números residenciales del directorio telefónico de la ciudad, el cual contiene 3000 de tales números. Se llama, entonces, a cada residencia elegida y se pregunta (a) el número total de votantes y (b) el número de votantes opuestos al quemador de basura. Se encuestó a un total de 157 votantes; de estos, 23 se rehusaron a contestar la pregunta. De los 134 votantes restantes, 112 se opusieron al quemador, de modo que el ayuntamiento estima la proporción en:

$$\hat{p}=\frac{112}{134}=.83582$$

con

$$\hat{V}[\hat{p}]=\frac{.83582(1-.83582)}{134}=0.00102.$$

¿Son válidas estas estimaciones? ¿Por qué?

2 Senturia *et al.* (1994) describen una encuesta realizada para estudiar el número de niños que tiene a su alcance armas en sus hogares. Se distribuyeron cuestionarios a todos los padres que asistieron a ciertas clínicas en el área de Chicago, durante un periodo de una semana para consultas de niños (sanos o enfermos).

a Suponga que la cantidad de interés es el porcentaje de familias que cuentan con armas. Diga por qué ésta es una muestra por conglomerados. ¿Cuál es la unidad primaria? ¿Cuál es la unidad secundaria? ¿Es una muestra por conglomerados en una o en dos etapas? ¿Cómo estimaría el porcentaje de familias que cuentan con armas y el error estándar de la estimación?

b ¿Cuál es la población de muestreo de este estudio? ¿Cree que el procedimiento de muestreo produce una muestra representativa de las familias con hijos? ¿Por qué?

3 Una empresa de contabilidad está interesada en estimar la tasa de error en una auditoría de acuerdos que está realizando. La población contiene 828 quejas y la empresa audita una muestra aleatoria simple de 85 de estas quejas. En cada una de las 85 quejas de la muestra, se verifican los errores en 215 campos. Una queja tuvo errores en 4 de los 215 campos, 1 queja tuvo 3 errores, 4 quejas tuvieron 2 errores, 22 quejas tuvieron un error y las restantes 57 quejas no tuvieron errores (estos datos fueron una cortesía de Fritz Scheuren).

- **a** Considere las quejas como unidades primarias y las observaciones de cada campo como unidades secundarias y estime la tasa de error de las 828 quejas. Dé un error estándar para la estimación.
- **b** Estime (con error estándar) el número total de errores que existen en las 828 quejas.
- **c** Suponga que en vez de tomar una muestra por conglomerados, la empresa toma una muestra aleatoria simple de $85 \times 215 = 18{,}275$ campos de los 178,020 campos de la población. Si la tasa de error estimada de la muestra aleatoria simple es la misma que en la parte (a), ¿cuál será la varianza estimada $\hat{V}(\bar{p}_{\text{MAS}})$? ¿Cómo se compara con la varianza estimada de la parte (a)?

4 La evidencia mediante encuestas se presenta, con frecuencia, en casos judiciales que implican la violación de las marcas registradas y discriminación en el trabajo. Sin embargo, ha habido cierta controversia acerca de si las muestras que no son de probabilidad son aceptables como evidencia en los litigios. Jacoby y Handlin (1991) seleccionaron 26 de una lista de 1285 revistas de ciencias sociales y del comportamiento. Examinaron todos los artículos publicados durante 1988 para las revistas elegidas y registraron (1) el número de artículos de la revista que describían la investigación empírica de una encuesta (excluyeron los artículos donde los autores analizaban datos de encuestas realizadas por otra persona) y (2) el número total de artículos para cada revista que usaban el muestreo de probabilidad, el muestreo que no es de probabilidad o aquellos para los cuales no se pudo determinar el método de muestreo. Los datos están en el archivo journal.dat.

- **a** Explique por qué ésta es una muestra por conglomerados.
- **b** Estime la proporción de artículos en las 1285 revistas que usan el muestreo que no son de probabilidad y dé el error estándar de la estimación.
- **c** Los autores concluyen que, debido a que "una enorme proporción de... expertos teóricos y prácticos reconocidos se basan en los diseños de muestreo que no son de probabilidad", las cortes "no deben tener problema en admitir encuestas que no son de probabilidad, pero bien realizadas, y su peso correspondiente" (página 175). Comente esta afirmación.

5 Use los datos del archivo coots.dat para estimar la longitud promedio de un huevo, junto con el error estándar. Asegúrese de graficar los datos en forma adecuada.

6 La propietaria de una casa que posee una gran biblioteca necesita estimar el costo de adquisición y el valor de reemplazo de la colección de libros, con fines de seguros. Ella tiene 44 estantes con libros y elige 12 estantes al azar. Para la segunda etapa de muestreo, ella cuenta los libros de los estantes seleccionados. Luego genera cinco números aleatorios entre 1 y M_i para cada estante seleccionado (véase la tabla 5.5) para determinar los libros específicos, numerados de izquierda a derecha, que deben examinarse con detalle. Posteriormente, busca el valor de reemplazo de los libros de la muestra en el catálogo *Books in Print*. Los datos aparecen en el archivo books.dat.

- **a** Construya gráficas de bloques adyacentes para los costos de reemplazo de los libros de cada estante. ¿Parece que las medias son iguales? ¿Y las varianzas?

TABLA 5.5
Tabla para el ejercicio 6

Número de estante	Número de libros (M_i)	Números de los libros seleccionados				
2	26	3	5	6	18	19
4	52	2	15	25	36	37
11	70	19	45	48	56	65
14	47	8	9	16	40	44
20	5	1	2	3	4	5
22	28	1	3	7	14	27
23	27	5	14	16	19	26
31	29	10	14	16	19	23
37	21	8	16	17	18	21
38	31	5	9	17	20	27
40	14	5	6	7	8	14
43	27	4	6	12	16	24

- **b** Estime el costo total de reemplazo de la biblioteca y determine el error estándar de la estimación. ¿Cuál es el coeficiente de variación estimado?
- **c** Estime el costo promedio de reemplazo por libro, junto con el error estándar. ¿Cuál es el coeficiente de variación estimado?

7 Repita el ejercicio 6 para el costo de adquisición de cada libro. Grafique los datos y estime el monto total y el promedio gastado en los libros, junto con los errores estándar.

8 Construya una tabla de análisis de la varianza de la muestra para los datos de costo de reemplazo del ejercicio 6. ¿Cuál es la estimación para R_a^2? ¿Tienden los libros del mismo estante a tener costos de reemplazo similares? Suponga que $c_1 = 10$ y que $c_2 = 4$. Si todos los estantes tienen 30 libros, ¿cuántos libros de cada estante deben estar en la muestra?

***9** Definimos el ICC en la página 139 como el coeficiente de correlación de Pearson para las $NM(M-1)$ parejas (y_{ij}, y_{ik}) para i entre 1 y N y $j \neq k$:

$$\text{ICC} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}^{M}(y_{ij}-\bar{y}_U)(y_{ik}-\bar{y}_U)}{(NM-1)(M-1)S^2}. \qquad \textbf{(5.42)}$$

Muestre que la definición anterior es equivalente a (5.8). SUGERENCIA: Muestre primero que

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}^{M}(y_{ij}-\bar{y}_U)(y_{ik}-\bar{y}_U)+\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_U)^2 = M(\text{SSB}).$$

***10** Suponga que en una muestra por conglomerados en dos etapas, todos los tamaños de cúmulo de la población son iguales ($M_i = M$ para toda i) y que todos los tamaños de muestra para los conglomerados también son iguales ($m_i = m$ para toda i).

- **a** Muestre (5.34).
- **b** Muestre que $\text{MSW} = S^2(1-R_a^2)$ y que

$$\text{MSB} = S^2\left[\frac{N(M-1)R_a^2}{N-1}+1\right].$$

c Use las partes (a) y (b) para expresar $V(\hat{\bar{y}})$ como una función de n, m, N, M, S^2 y R_a^2.

d Muestre que si S^2 y los tamaños de la muestra y de la población son fijos, y si $m/(m-1) > n/N$, entonces $V(\hat{\bar{y}})$ es una función creciente de R_a^2.

***11** Suponga que en una muestra por conglomerados en dos etapas, todos los tamaños de cúmulo de la población son iguales ($M_i = M$ para toda i) y que todos los tamaños de muestra para los conglomerados también son iguales ($m_i = m$ para toda i).

a Muestre que $\hat{t}_{\text{ins}} = \hat{t}_r$ y, por tanto, que $\hat{\bar{y}}_{\text{ins}} = \hat{\bar{y}}_r$.

b Escriba las fórmulas para las sumas de cuadrados en la siguiente tabla de análisis de varianza, para los datos muestrales.

Fuente	gl	Suma de cuadrados	Cuadrado de la media
Entre los conglomerados	$n-1$		$\widehat{\text{MSB}}$
Dentro de los conglomerados	$n(m-1)$		$\widehat{\text{MSW}}$
Total	$nm-1$		

c Muestre que $E[\widehat{\text{MSB}}] = \text{MSW}$ y $E[\widehat{\text{MSB}}] = (m/M)\text{MSB} + [1-(m/M)]\text{MSW}$, donde MSB y MSW son los cuadrados de las medias entre y dentro de los conglomerados, respectivamente, de la tabla de análisis de la varianza de la *población*.

d Muestre, al usar (5.25) o (5.29), que

$$\hat{V}(\hat{\bar{y}}_{\text{ins}}) = \left(1-\frac{n}{N}\right)\frac{\widehat{\text{MSB}}}{nm} + \frac{1}{N}\left(1-\frac{m}{M}\right)\frac{\widehat{\text{MSW}}}{m}.$$

12 Un inspector toma una muestra de un camión que transporta maíz enlatado para estimar el número promedio de fragmentos de gusanos por lata. El camión tiene 580 cajas; cada caja contiene 24 latas. El inspector elige 12 cajas al azar y extrae 3 latas al azar de cada caja seleccionada.

	Caja											
	1	2	3	4	5	6	7	8	9	10	11	12
Lata 1	1	4	0	3	4	0	5	3	7	3	4	0
Lata 2	5	2	1	6	0	7	5	0	3	1	7	0
Lata 3	7	4	2	6	8	3	1	2	5	4	9	0

a Estime el número medio de fragmentos de gusanos por lata, junto con el error estándar de la estimación (puede usar el resultado del ejercicio 11 para calcular el error estándar).

b Suponga que debe revisarse otro camión y se cree que sea similar al anterior. Se necesitan 10 minutos para ubicar y abrir una caja y 8 minutos para ubicar y examinar cada lata específica dentro de una caja. ¿Cuántas latas deben examinarse por caja?

13 El nuevo dulce Green Globules es probado en el mercado, en un área del estado de Nueva York. Una empresa de investigación de mercado decide extraer una muestra de 6 de las 45 ciudades del área y, luego, obtener una muestra de los supermercados de esas ciudades, para conocer el número de cajas vendidas del dulce.

Ciudad	Número de supermercados	Número de cajas vendidas
1	52	146, 180, 251, 152, 72, 181, 171, 361, 71, 186
2	19	99, 101, 51,121
3	37	199, 179, 98, 63, 126, 87, 62
4	39	226, 129, 57, 46, 86, 43, 85, 165
5	8	12, 23
6	14	87, 43, 59

Utilice cualquier paquete de software estadístico para obtener los estadísticos de cada cúmulo. Grafique los datos y estime el número total de cajas vendidas y el número promedio vendido por supermercado, junto con los errores estándar de las estimaciones.

14 El Sistema de Contención de Costos para el Cuidado de la Salud de Arizona (AHCCCS) proporciona asistencia médica a las familias de bajos ingresos en Arizona. Cada condado determina si las familias pueden recibir esta asistencia. Sin embargo, a veces, las familias están certificadas para recibir la asistencia, cuando no deberían estarlo. La sección 36-2905.01 de los estatutos de Arizona ordena la recolección de "una muestra con un control de calidad estadísticamente válido acerca de las certificaciones de elegibilidad realizado por cada condado". La tasa de error de la certificación para cada condado debe determinarse "al dividir el número de miembros de la muestra que fueron erróneamente certificados entre el número total de elementos de la muestra". Sin embargo, se realizan auditorías de control de calidad mediante una muestra de los registros familiares; una vez seleccionado y auditado el registro de una familia, la evaluación de una persona en la familia cuesta lo mismo que evaluar a todos los miembros de la familia.

a Explique la forma de usar el muestreo por conglomerados para estimar la tasa de error de certificación para un condado.

b Suponga que un condado certificó 1572 familias como elegibles para la asistencia médica en 1995. Durante los últimos años, la tasa de error de certificación ha sido cercana al 10%. ¿Cuántas familias se deben incluir en la muestra de modo que la mitad del ancho de un intervalo de confianza del 95%, para estimar la tasa de error de certificación por persona, sea menor a 0.03? ¿Cuál es el supuesto que debe establecer para obtener el tamaño de la muestra?

15 Una investigadora quiere estudiar la frecuencia del uso del cigarro y de otros comportamientos de alto riesgo entre las estudiantes de bachillerato que se encuentran en una región con 35 escuelas de este nivel.

Número de estudiantes	Número de escuelas
0-499	3
500-999	7
1000-1499	18
1500-2000	5

La científica prétende ir en auto hasta n número de escuelas y ahí entrevistar a algunas o a todas las estudiantes del centro de estudios seleccionado. La investigadosra ha realizado un estudio similar con 4 de 29 escuelas que pertenecen a otra región. Los resultados fueron los siguientes:

Escuela	Número de alumnos	Número de alumnas	Número de alumnas entrevistadas	Número de fumadoras
1	1471	792	25	10
2	890	447	15	3
3	1021	511	20	6
4	1587	800	40	27

a Estime el porcentaje de alumnas que fuman, a partir del estudio de las 4 escuelas.

b Use la información del estudio anterior y proponga un diseño para el nuevo estudio. Suponga que se necesitan cerca de 50 horas por escuela para establecer contacto con los funcionarios de la institución, obtener la autorización, conseguir una lista de las estudiantes y viajar de ida y regreso. Aunque las propias entrevistas sólo tardan cerca de 10 minutos, se necesitan cerca de 30 minutos por cada entrevista obtenida, lo que incluye el poder programar la ausencia de las entrevistadas, obtener el permiso de los padres y otras tareas administrativas. La investigadora quisiera ocupar 300 horas o menos en la recolección de los datos.

16 Gnap (1995) realizó una encuesta para estimar la carga de trabajo docente en los distritos escolares públicos del condado de Maricopa en Arizona. La población objetivo estaba compuesta por todos los maestros de tiempo completo de las escuelas públicas, del primero al sexto grado, con al menos un año de experiencia. En 1994, el condado tenía 46 distritos escolares, con 311 escuelas de nivel elemental y 15,086 maestros. Gnap estratificó las escuelas por tamaño del distrito escolar; en este ejercicio consideramos al mayor estrato, formado por las escuelas que pertenecen a los distritos que cuentan con más de 5000 estudiantes. El estrato contenía 245 escuelas; 23 participaron en la encuesta. A todos los maestros de las escuelas seleccionadas se les pidió que respondieran el cuestionario. Sin embargo, debido a la ausencia de respuestas, algunos cuestionarios no fueron regresados (examinaremos los efectos posibles de la ausencia de respuestas en el ejercicio 17 del capítulo 8). Los datos están en el archivo teachers.dat, con la información de las unidades primarias en teachmi.dat.

a ¿Por qué una muestra por conglomerados podría ser un mejor diseño que una muestra aleatoria simple para este estudio? Considere aspectos como el costo, la facilidad de recolección de datos y la confidencialidad de los que respondan a la encuesta. ¿Cuáles son algunas desventajas del uso de una muestra por conglomerados?

b Calcule la media y la desviación estándar de la variable *hrwork* para cada escuela del estrato más "grande". Construya una gráfica de las medias para cada escuela y una gráfica separada de la desviación estándar. ¿Parece haber más variación dentro de una escuela, o bien la mayor parte de la variabilidad ocurre entre escuelas distintas? ¿Cómo trabajó con los valores faltantes (codificados como –9)?

c Construya una gráfica de dispersión de las desviaciones estándar contra las medias para las escuelas y la variable *hrwork*. ¿Existe más variabilidad en las escuelas con mayor carga de trabajo? ¿Menos? ¿No hay una relación aparente?

d Estime el promedio de *hrwork* en el estrato más grande para el condado citado, junto con el error estándar. Use la variable *popteach* del archivo teachmi.dat como las M_i.

17 El archivo measles.dat contiene datos consistentes con los obtenidos en una encuesta de padres cuyos hijos no han sido vacunados contra el sarampión, durante una campaña reciente de vacunación de todos los niños con edades comprendidas entre los 11 y 15 años. Durante la campaña se vacunaron 7633 niños de las 46 escuelas del área; 9962 niños que no tenían una vacuna previa no fueron vacunados. En un estudio de seguimiento para saber por qué los niños no fueron vacunados durante la campaña, Roberts *et al.* (1995) enviaron cuestionarios a los padres de una muestra por conglomerados de los 9962 niños. Se eligieron al azar 10 escuelas; luego, se seleccionaron los niños no vacunados de cada escuela y se envió el cuestionario a los padres de estos niños. No todos los padres respondieron el cuestionario (usted examinará los efectos de la ausencia de respuestas en el ejercicio 18 del capítulo 8).

Escuela	Número de estudiantes no vacunados (M_i)
1	78
2	238
3	261
4	174
5	236
6	188
7	113
8	170
9	296
10	207

a Al usar los datos de los cuestionarios regresados, estime, en forma separada para cada escuela, el porcentaje de padres que regresaron una forma de consentimiento. Para este ejercicio, ignore las respuestas "no contestó".

b Estime el porcentaje global de padres que regresaron una forma de consentimiento y dé un intervalo de confianza del 95% para la estimación.

c ¿Cuál es la relación que existe entre la estimación y el intervalo de la parte (b) con los resultados que habría obtenido al ignorar los conglomerados y analizar los datos como una muestra aleatoria simple? Determine la siguiente proporción:

$$\frac{\text{varianza estimada en la parte (b)}}{\text{varianza estimada si los datos se analizan como una muestra aleatoria simple}}$$

¿Cuál es el efecto de los conglomerados?

18 Repita el ejercicio 17, pero esta vez estime el porcentaje de niños que habían tenido sarampión con anterioridad.

19 Consulte el ejemplo 5.9. Durante la temporada de crecimiento de las papas, se necesita más tiempo para inspeccionar los tallos. Suponga que se necesitan 2 minutos para revisar cada tallo. ¿Cuál es el tamaño de unidad primaria más eficiente?

20 **a** Para la muestra aleatoria simple del censo de agricultura, del archivo agsrs.dat (analizado en el ejemplo 2.4), determine la tabla de análisis de varianza de la muestra para *acres92*, al usar *state* como la variable de conglomerado. ¿Cuánto es R_a^2 para esta muestra? ¿Existe un efecto de conglomerado?

b Suponga que $c_1 = 15c_2$, donde c_1 es el costo de muestreo de un estado y c_2 es el costo de muestreo de un condado dentro de un estado. Si $\overline{M}$ es el tamaño del conglomerado, ¿cuánto debe valer $\bar{m}$ si se quiere extraer una muestra total de 300 condados? ¿Cuántos estados participarán en la muestra (es decir, cuánto vale n)?

21 Use el valor de n determinado en el ejercicio 20 y extraiga una muestra por conglomerados autoponderada de 300 condados del archivo agpop.dat. Grafique los datos mediante una gráfica de bloques adyacentes. Estime el número total de acres dedicados a la agricultura en Estados Unidos, junto con el error estándar, al usar la estimación insesgada y la estimación de proporción. ¿Cuál es la relación entre estos valores, así como su relación con las muestras aleatoria simple y estratificada de los ejemplos 2.4 y 4.1?

22 El archivo ozone.dat contiene las lecturas de ozono por hora en Eskdalemuir, Escocia, para 1994 y 1995.

- **a** Construya un histograma de los valores de la población. Determine la media, la desviación estándar y la mediana de la población.
- **b** Tome una muestra sistemática con periodo 24. Para esto, elija al azar un entero k entre 1 y 24 y seleccione la columna que contiene las observaciones con la hora central de Greenwich (GMT) igual a k. Construya un histograma con los valores de la muestra.
- **c** Ahora suponga que considera la muestra sistemática que acaba de obtener como si fuese una muestra aleatoria simple. Determine la media, la desviación estándar y la mediana de la muestra. Construya una estimación del intervalo de la media de la población, utilice el procedimiento de la sección 2.4. ¿Contiene el intervalo el verdadero valor de la media de la población obtenida en la parte (a)?
- **d** Tome cuatro muestras sistemáticas independientes, cada una con periodo 96. Ahora use las fórmulas del muestreo por conglomerados para estimar la media de la población y construya un intervalo de confianza del 95% para la media.

***23** (Requiere de conocimiento de cálculo.) Muestre que si $M_i = M$ y $m_i = m$ para toda i y si la función de costo es $C = c_1 n + c_2 nm$, entonces,

$$m = \sqrt{\frac{c_1 M(N-1)(1-R_a^2)}{c_2(NM-1)R_a^2}}$$

minimiza la varianza de $\bar{y}_{\text{ins}}$ para un costo total fijo C. SUGERENCIA: use el ejercicio 10.

***24** (Requiere de conocimientos de trigonometría.) En el ejemplo 5.12 se propuso una muestra sistemática para detectar los desechos tóxicos en los rellenos sanitarios. ¿Qué distancia debe haber entre los puntos de muestreo? Suponga que existe un derrame y que se dispersa en una región circular con radio R. Sea $2D$ la distancia entre los puntos de muestreo adyacentes en un mismo renglón o columna.

- **a** Calcule la probabilidad de detección de un contaminante. SUGERENCIA: considere tres casos, con $R < D, D \le R \le \sqrt{2}D$, y $R > \sqrt{2}D$.
- **b** Proponga un diseño de muestreo que dé una mayor probabilidad de detectar un contaminante que la retícula cuadrada, pero que no incremente el número de puntos de muestreo.

***25** (Requiere de conocimientos sobre modelos con efectos aleatorios.) Según el modelo M1 en (5.37), un modelo de un sentido con efectos aleatorios, ρ se puede estimar como:

$$\hat{\rho} = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}^2},$$

donde $\hat{\sigma}_A^2$ y $\hat{\sigma}^2$ estima los componentes de la varianza $\hat{\sigma}_A^2$ y $\hat{\sigma}^2$. Los estimadores del método del momento para un muestreo por conglomerados en una etapa cuando todos los conglomerados tienen el mismo tamaño son $\hat{\sigma}^2 = \text{MSW}$ y $\hat{\sigma}_A^2 = (\text{MSB} - \text{MSW})/M$.

- **a** ¿Cuánto vale $\hat{\rho}$ en el ejemplo 5.4? ¿Cuál es la relación con ICC?
- **b** Calcule $\hat{\rho}$ para las poblaciones A y B del ejemplo 5.3. ¿Por qué difieren del $\widehat{\text{ICC}}$?

***26** (Requiere de conocimientos sobre modelos con efectos aleatorios.)

- **a** Suponga que ignoramos la corrección para las poblaciones finitas de un estimador basado en el modelo. Determine lo siguiente:

$$V_{\text{M1}}\left(\sum_{i\in S}\sum_{j\in S_i} b_{ij}Y_{ij}\right).$$

- **b** Demuestre (5.39). SUGERENCIA: sea

$$c_{ij} = \begin{cases} b_{ij} - 1 & \text{si } i \in S \text{ y } j \in S_i. \\ -1 & \text{en caso contrario} \end{cases}$$

Entonces $\hat{T} - T = \sum_{i=1}^{N}\sum_{j=1}^{M_i} c_{ij}Y_{ij}$.

***27** (Requiere de conocimientos de álgebra lineal y cálculo.) Aunque $\hat{T}_r$ es insesgado para el modelo M1, es posible construir un estimador que posea una menor varianza. Sean:

$$c_k = \frac{m_k}{1 + \rho(m_k - 1)}$$

y

$$\hat{T}_{\text{opt}} = \sum_{i\in S}\sum_{j\in S_i}\frac{c_i}{m_i}\left[\rho M_i + \frac{K - \rho\sum_{k\in S} c_k M_k}{\sum_{k\in S} c_k}\right]Y_{ij}.$$

Muestre que $\hat{T}_{\text{opt}}$ es insesgado y minimiza la varianza en (5.39) entre todos los estimadores insesgados para el modelo (5.37).

***28** (Requiere de conocimientos de cálculo.) Suponga que las M_i no son iguales y que el modelo M1 es válido. El presupuesto le permite realizar un total de L mediciones sobre las subunidades. Muestre que la varianza en (5.39) se minimiza para $\hat{T}_r$ cuando m_i es proporcional a M_i. SUGERENCIA: use los multiplicadores de Lagrange, con la restricción $\sum_{i\in S} m_i = L$.

29 La edición de enero de 1994 de *The Nation* clasificó a 22 columnistas por la frecuencia en que utilizaron las palabras *yo, me, mi*. Seleccione a su columnista favorito. Elija al azar cinco artículos del columnista de este último año y utilice el muestreo por conglomerados de una etapa para estimar la proporción de palabras totales ocupada por *yo, me, mi*. ¿Cuál es su unidad primaria? ¿Su unidad secundaria?

6

Muestreo con probabilidades diferentes

"Personalmente, nunca me han interesado la ficción ni los cuentos. Lo que me gusta leer son hechos y estadísticas de cualquier tipo; aunque sean hechos acerca del cultivo de rábanos, me interesan. Precisamente ahora, por ejemplo, antes de que usted entrara —señala una enciclopedia que se encuentra en un librero— estaba leyendo un artículo sobre 'matemáticas'. Matemáticas perfectamente puras.
"Mi conocimiento matemático termina en '12 por 12', pero disfruté inmensamente con ese artículo. No entendí una palabra, pero los hechos, o lo que el hombre cree que son los hechos, parecen siempre encantadores. Este matemático creía en sus hechos. También yo. Primero obtenga sus propios hechos y —aquí la voz disminuye hasta ser casi imperceptible— luego puede distorsionarlos tanto como desee".

—Mark Twain, citado por Rudyard Kipling, en *From Sea to Sea*

Hasta ahora sólo hemos analizado esquemas de muestreo donde las probabilidades de selección de las unidades de muestreo son iguales. Las probabilidades iguales proporcionan esquemas que, frecuentemente, son fáciles de diseñar y explicar. Sin embargo, estos esquemas no siempre se pueden realizar, no son tan eficientes como los esquemas que utilizan probabilidades diferentes. En el ejemplo 5.7 vimos que un muestreo por conglomerados con probabilidades iguales puede producir una varianza grande para el estimador insesgado respecto al diseño, a la media y al total de la población.

EJEMPLO 6.1 O'Brien *et al.* (1995) extrajeron una muestra de los residentes de casas de asistencia en el área de Filadelfia, para determinar sus preferencias en cuanto a los tratamientos de salud que querían recibir. ¿Desean que se les aplique resucitación cardiopulmonar (CPR) si el corazón deja de latir? ¿Que se les transfiera a un hospital si aparece una enfermedad grave o que se les entube si ya no pueden comer? La población objetivo estaba constituida por todos los residentes de las casas de asistencia autorizados, en el área de Filadelfia. Había 294 de esas instalaciones, con un total de 37,652 camas (antes del muestreo sólo se conocía la cantidad de camas, no el número de residentes).

Puesto que la encuesta debía realizarse personalmente, el muestreo por conglomerados era esencial para mantener los gastos en un nivel aceptable. Si los investigadores hubiesen elegido el muestreo por conglomerados con probabilidades iguales de selección, entonces hubieran tenido que extraer una muestra aleatoria simple de las casas de asistencia y, luego, otra muestra aleatoria simple de residentes dentro de cada casa elegida.

Sin embargo, en una muestra por conglomerados con probabilidades iguales, una casa de asistencia con 20 camas tiene la misma probabilidad de que se elija que una con mil

camas. La muestra es autoponderada sólo si el tamaño de la submuestra es proporcional a la cantidad de camas en la casa de asistencia. Cada cama de la muestra representa la misma cantidad de camas en la población si se utiliza un muestreo por conglomerados de una etapa o si 10% (o cualquier otro porcentaje) de las camas de cada casa participan en la muestra.

El muestreo de las casas de asistencia con probabilidades iguales produciría un estimador válido desde el punto de vista matemático, pero tiene tres desventajas principales. La primera, que se esperaría que la cantidad total de pacientes de una casa que deseen CPR (t_i) sería proporcional al número de camas en la casa (M_i), de modo que los estimadores del capítulo 5 tendrían una varianza grande. En segundo lugar, una muestra autoponderada con probabilidades iguales puede ser difícil de trabajar. Tal vez, habría que ir hasta una casa de asistencia sólo para entrevistar a uno o dos residentes, lo que dificultaría el equilibrio de la carga de trabajo de los entrevistadores. En tercer lugar, el costo de la muestra no se conoce de antemano. Por ejemplo, una muestra aleatoria de 40 casas de asistencia podría constar, principalmente, de instituciones grandes, lo que conduciría a un gasto mayor al anticipado.

En vez de extraer una muestra por conglomerados de casas de asistencia con probabilidades iguales, los investigadores extrajeron al azar una muestra de 57 hogares con probabilidades proporcionales a la cantidad de camas. Luego, obtuvieron una muestra aleatoria simple de 30 camas (y sus ocupantes) de una lista de todas las camas dentro de la casa de asistencia. Si la cantidad de residentes es igual al número de camas y, si al realizar la visita, una casa de asistencia tiene la misma cantidad, número de camas, que el indicado en el marco de muestreo, entonces, el diseño del muestreo hace que cada residente tenga la misma probabilidad de ser incluido en la muestra. El costo se conoce antes de seleccionar la muestra. Se realiza la misma cantidad de entrevistas en cada casa de asistencia y es probable que el estimador del total de la población tenga una menor varianza que los estimadores del capítulo 5.

Como esta muestra es autoponderada, se pueden obtener fácilmente las estimaciones puntuales (pero *no* los errores estándar) de las cantidades deseadas mediante los métodos normales. Puede obtener la edad promedio de los residentes de las casas de asistencia, hallando la mediana muestral de los residentes que componen la muestra, o el percentil 70, al determinar el percentil 70 de la muestra. Aunque una muestra no sea autoponderada, las estimaciones puntuales se pueden calcular fácilmente al usar pesos. Pero cuidado, siempre considere el diseño por conglomerados al calcular la precisión de sus estimaciones. ■

En el capítulo 4 observamos que, en ocasiones, se utiliza el muestreo estratificado para realizar el muestreo de unidades distintas, con probabilidades distintas. En una investigación para estimar los gastos totales de las empresas en el renglón de la publicidad, tal vez quisiéramos estratificar por ventas o ingresos de las compañías. Las empresas más grandes, como IBM, estarían en un estrato, las medianas en uno o varios estratos distintos y las compañías muy pequeñas, como la sastrería de Robin, estarían en un estrato diferente. Un esquema óptimo de distribución extraería una muestra de una proporción muy alta (tal vez 100%) del estrato de las grandes compañías y una fracción muy pequeña de las compañías en el estrato de las empresas más pequeñas; la varianza de una compañía a otra será mucho mayor entre IBM, AT&T y Phillip Morris que entre la sastrería de Robin, la zapatería de Pat y la florería de Leslie. La varianza es mayor en las grandes compañías, tan sólo por el hecho de que las cantidades de dinero implicadas también son mucho más grandes. Así, la varianza del muestreo disminuye al asignar probabilidades diferentes a las unidades de muestreo en estratos distintos.

Para estimar el gasto total en publicidad mediante esta muestra estratificada, asignamos mayores pesos a las compañías con menor probabilidad de selección. Como vimos en la sección 4.3, la probabilidad de incluir en la muestra una compañía del estrato h es n_h/N_h; el peso de muestreo para esa compañía es N_h/n_h. Cada compañía del estrato h incluida en la muestra representa N_h/n_h compañías en la población y $\hat{t}_{str} = \sum_{h=1}^{H} \sum_{j \in S_h} (N_h/n_h) y_{hj}$.

También podemos utilizar las probabilidades de selección diferentes para disminuir las varianzas sin tener que estratificar en forma explícita. Al realizar un muestreo con probabilidades diferentes, variamos deliberadamente las posibilidades de seleccionar distintas unidades primarias de muestreo en la muestra y compensamos este hecho al proporcionar pesos adecuados en la estimación. La clave es que *conocemos* las probabilidades[1] con las que seleccionamos una unidad dada:

$$P(\text{seleccionar la unidad } i \text{ en la primera extracción}) = \psi_i. \quad \textbf{(6.1)}$$

$$P(\text{unidad } i \text{ esté en la muestra}) = \pi_i. \quad \textbf{(6.2)}$$

La selección deliberada de las unidades primarias de muestreo con probabilidades conocidas, aunque distintas, difiere en gran medida del sesgo de selección analizado en el capítulo 1. Muchas encuestas con sesgo de selección realizan muestras con probabilidades diferentes, pero las probabilidades de selección no se conocen ni se pueden estimar, de modo que los realizadores de la encuesta no pueden compensar las probabilidades diferentes en la ponderación. Si se realiza una investigación sobre estudiantes, y se pide la participación de los alumnos que se dirigen hacia la biblioteca, se está realizando un muestreo con probabilidades diferentes; es más probable que se pida la participación de estudiantes que utilizan la biblioteca frecuentemente, que de los que nunca acuden a ella, pero no tiene idea de cuántos estudiantes de la población están representados por un participante en la encuesta y no hay forma de corregir las probabilidades de selección diferentes en la estimación.

Cuando se presenta, por primera vez, la idea del muestreo con probabilidades diferentes, algunas personas suponen que "no es natural" o que es "inventada". Por el contrario, para muchas poblaciones con conglomerados, el muestreo con probabilidades diferentes al nivel de las unidades primarias produce una muestra más representativa de la población que una muestra con probabilidades iguales. En la sección 6.5 daremos algunos ejemplos de muestras con probabilidades diferentes. Para comprender estos ejemplos y diseñar sus propias muestras, es esencial que comprenda la probabilidad. Primero consideraremos el muestreo con reemplazo, partiendo del diseño sencillo con selección de una unidad primaria de muestreo. En la sección 6.4 consideraremos el muestreo con probabilidades distintas sin reemplazo. La notación utilizada en este capítulo se definió en la sección 5.1.

6.1 Muestreo de una unidad primaria de muestreo

Como caso particular, suponga que elegimos una ($n = 1$) de las N unidades primarias de muestreo para incluir en la muestra. El total de la unidad primaria i es t_i y queremos estimar el total de la población, t. El muestreo de una unidad primaria dejará ver las ideas del muestreo con probabilidades diferentes sin presentar más complicaciones.

Comencemos con observar lo que ocurre en una situación en la que conocemos a toda la población. Una ciudad tiene cuatro supermercados, los cuales varían de tamaño entre 100 y 1000 metros cuadrados. Queremos estimar la cantidad total de ventas del último mes, en las cuatro tiendas, con una muestra de una sola tienda. (Por supuesto, esto es un ejemplo; si realmente tuviésemos cuatro tiendas, haríamos un censo.) Supondríamos que una tienda más grande tuviese más ventas que una más pequeña y que la variabilidad en el total de

[1] En este capítulo consideramos dos probabilidades pues, al realizar un muestreo con probabilidades diferentes sin reemplazo (véase la sección 6.4), la selección de una unidad en la primera extracción puede afectar las probabilidades de selección de otras unidades.

ventas entre las diversas tiendas de 1000 m² sea mayor que la variabilidad en el total de ventas entre varias tiendas de 100 m².

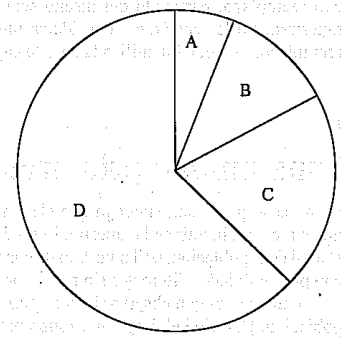
Como la muestra sólo consta de una tienda, tenemos que la probabilidad de elegir una tienda en la primera extracción (ψ_i) es igual a la probabilidad de que la tienda se incluya en la muestra (π_i). Para este ejemplo, supongamos que

$$\pi_i = \psi_i = P(\text{elegir la tienda } i)$$

es proporcional al tamaño de la tienda. Como la tienda A representa 1/16 del área total de las cuatro tiendas, se muestra con probabilidad 1/16. Con fines didáticos, suponga que conocemos los valores de t_i para toda la población:

Tienda	Tamaño (m²)	ψ_i	t_i (en miles)
A	100	$\frac{1}{16}$	11
B	200	$\frac{2}{16}$	20
C	300	$\frac{3}{16}$	24
D	1000	$\frac{10}{16}$	245
Total	1600	1	300

Podríamos elegir una muestra de probabilidad de tamaño 1, con las probabilidades anteriores, al revolver unas cartas numeradas del 1 al 16 y seleccionar una de ellas. Si el número de la carta es 1, elegimos la tienda A; si es 2 o 3, seleccionamos B; si es 4, 5 o 6, elegimos C, y si es del 7 al 16, escogemos D. También podemos girar una ruleta como ésta:



Compensamos las probabilidades diferentes de selección al usar ψ_i también en el estimador. Ya hemos estudiado la compensación de las probabilidades diferentes de selección en el muestreo estratificado: si elegimos 10% de las unidades en el estrato 1 y 20% en el estrato 2, el peso de muestreo es 10 para cada unidad del estrato 1 y 5 para cada unidad del estrato 2. En este caso, elegimos la tienda A con probabilidad 1/16, de modo que el peso de muestreo de la tienda A es 16. Si el tamaño de la tienda es, aproximadamente, proporcional al total de ventas para esa tienda, esperaríamos que la tienda A tuviese también 1/16 de las

ventas totales y que al multiplicar las ventas de la tienda A por 16, estimaríamos las ventas totales para las cuatro tiendas. Como siempre, el peso de muestreo de la unidad i es el recíproco de la probabilidad de selección:

$$w_i = \frac{1}{P(\text{unidad } i \text{ en la muestra})} = \frac{1}{\psi_i}.$$

Así, nuestro estimador del total de la población a partir de una muestra de tamaño 1 con probabilidades diferentes es:

$$\hat{t}_\psi = \sum_{i \in S} w_i t_i = \sum_{i \in S} \frac{t_i}{\psi_i}.$$

Es posible obtener cuatro muestras de tamaño 1 de esta población que se compone por 4 tiendas:

Muestra	ψ_i	t_i	$\hat{t}_\psi$	$(\hat{t}_\psi - t)^2$
{A}	$\frac{1}{16}$	11	176	15,376
{B}	$\frac{2}{16}$	20	160	19,600
{C}	$\frac{3}{16}$	24	128	29,584
{D}	$\frac{10}{16}$	245	392	8,464

Como definimos en el capítulo 2,

$$E[\hat{t}_\psi] = \sum_{\substack{\text{posibles} \\ \text{muestras } S}} P(S)\hat{t}_{\psi S}$$

$$= \frac{1}{16}(176) + \frac{2}{16}(160) + \frac{3}{16}(128) + \frac{10}{16}(392) = 300.$$

Por supuesto, $\hat{t}_\psi$ siempre será insesgado, pues en general,

$$E[\hat{t}_\psi] = \sum_{i=1}^{N} \psi_i \frac{t_i}{\psi_i} = t. \tag{6.3}$$

La varianza de $\hat{t}_\psi$ es

$$V[\hat{t}_\psi] = E[(\hat{t}_\psi - t)^2]$$

$$= \sum_{\substack{\text{posibles} \\ \text{muestras } S}} P(S)(\hat{t}_{\psi S} - t)^2$$

$$= \sum_{i=1}^{N} \psi_i \left(\frac{t_i}{\psi_i} - t \right)^2. \tag{6.4}$$

Para este ejemplo,

$$V[\hat{t}_\psi] = \frac{1}{16}(15{,}376) + \frac{2}{16}(19{,}600) + \frac{3}{16}(29{,}584) + \frac{10}{16}(8{,}464) = 14{,}248.$$

Compare estos resultados con los de una muestra aleatoria simple de tamaño 1, en la cual la probabilidad de selección de cada unidad sea $\psi_i = 1/4$, de modo que $1/\psi_i = 4 = N$. Observe que si todas las probabilidades de selección son iguales, como en el muestreo aleatorio simple, $1/\psi_i$ siempre es igual a N.

Muestra	ψ_i	t_i	$\hat{t}_\psi$	$(\hat{t}_\psi - t)^2$
{A}	$\frac{1}{4}$	11	44	65,536
{B}	$\frac{1}{4}$	20	80	48,400
{C}	$\frac{1}{4}$	24	96	41,616
{D}	$\frac{1}{4}$	245	980	462,400

Como siempre, $\hat{t}_{MAS}$ es insesgado, por lo que tiene una esperanza de 300, aunque para este ejemplo la varianza de la muestra aleatoria simple es mucho mayor que la varianza del esquema de muestreo con probabilidades diferentes:

$$V[\hat{t}_\psi] = \frac{1}{4}(65,536) + \frac{1}{4}(48,400) + \frac{1}{4}(41,616) + \frac{1}{4}(462,400) = 154,488.$$

La varianza del esquema con probabilidades diferentes, 14,248, es mucho menor, pues utiliza información auxiliar: esperamos que el tamaño de la tienda se relacione con las ventas y utilizamos esa información al diseñar el esquema de muestreo.

Creemos que t_i está correlacionado con el tamaño de la tienda, el cual conocemos. Como la tienda D representa 10/16 del área total de los supermercados, es razonable creer que la tienda D represente también 10/16 del total de las ventas. Así, si se elige la tienda D y se supone que ésta represente a 10/16 del total de las ventas, tendríamos una buena estimación de ese total al multiplicar las ventas de la tienda D por 16/10.

¿Qué ocurre si la tienda D sólo representa 4/16 del total de ventas? En este caso, el estimador con probabilidades diferentes $\hat{t}_\psi$ seguirá siendo insesgado al repetir el muestreo varias veces, pero tendrá una varianza grande (ver el ejercicio 5). El método sigue funcionando desde el punto de vista matemático, pero no es tan eficiente como en el caso en que t_i sea, aproximadamente, proporcional a ψ_i.

El muestreo de una sola unidad primaria no es tan extraño como podría pensarse. Muchas encuestas grandes y complejas están tan altamente estratificadas que cada estrato contiene unas cuantas unidades primarias de muestreo. Se utiliza una gran cantidad de estratos para aumentar la precisión de las estimaciones de la encuesta. En este tipo de encuestas podría ser razonable querer elegir una sola unidad primaria de muestreo para cada estrato. Pero, con una sola unidad primaria por estrato en la muestra, no tenemos una estimación de la variabilidad que existe entre las unidades primarias que se encuentran dentro de un estrato. Cuando las grandes empresas encuestadoras extraen como muestra una sola unidad primaria por estrato, frecuentemente dividen dicha unidad de alguna forma para estimar la varianza del estrato; analizaremos este método en el capítulo 9.

6.2 Muestreo en una etapa con reemplazo

Ahora suponga que $n > 1$ y que obtenemos una muestra *con* reemplazo. El muestreo con reemplazo significa que las probabilidades de selección no cambian después de extraer la primera unidad. Sea

$$\psi_i = P \text{ (elegir la unidad } i \text{ en la primera extracción).}$$

Si realizamos un muestreo con reemplazo, entonces ψ_i es también la probabilidad de que la unidad i se seleccione en la primera extracción, en la tercera o en cualquier otra. La probabilidad global de que la unidad i esté en la muestra, al menos una vez, es:

$$\pi_i = 1 - P(\text{unidad } i \text{ no esté en la muestra}) = 1 - (1 - \psi_i)^n.$$

Si $n = 1$, entonces $\pi_i = \psi_i$.

La idea subyacente del muestreo con probabilidades diferentes es sencilla. Extraemos n unidades primarias con reemplazo; luego, estimamos el total de la población, mediante el estimador de la sección anterior, por separado para cada unidad primaria extraída. Algunas unidades primarias podrían extraerse más de una vez. El total estimado de la población, calculada con una unidad primaria dada se incluye tantas veces como se extraiga la unidad primaria. Como las unidades primarias se obtienen con reemplazo, tenemos n estimaciones independientes del total de la población. Luego, estimamos el total de la población t al promediar esas n estimaciones independientes de t. La varianza estimada es la varianza muestral de las n estimaciones independientes de t, dividida entre n.

6.2.1 Selección de las unidades primarias de muestreo

6.2.1.1 El método de tamaño acumulativo

Existen varias formas de realizar una muestra de unidades primarias con probabilidades diferentes. Todas requieren de una medida del tamaño de las unidades primarias de la población. El método de tamaño acumulativo amplía el método utilizado en la sección anterior, en la que se generan números aleatorios y las unidades primarias correspondientes a esos números quedan incluidas en la muestra. En el caso de los supermercados, extrajimos cartas de una baraja, que estaban numeradas del 1 al 16. Si el número de la carta es 1, elegimos la tienda A, si es 2 o 3, elegimos B; si es 4, 5 o 6, elegimos C, y si es un número del 7 al 16, elegimos D. Para una muestra con reemplazo, regresamos la carta después de seleccionar una unidad primaria y realizamos una nueva extracción.

EJEMPLO 6.2 Considere la población de los grupos de la materia "Introducción a la estadística" que se imparte en cierta universidad. Esta población aparece en la tabla 6.1. La universidad tiene 15 grupos que toman esa materia; el grupo i tiene M_i estudiantes, para un total de 647 estudiantes de esta materia. Decidimos extraer una muestra de cinco grupos con reemplazo, con probabilidad proporcional a M_i y, luego, planteamos un cuestionario a cada alumno de los grupos de la muestra. Para este ejemplo, entonces, $\psi_i = M_i/647$.

Para elegir la muestra, generamos cinco números enteros aleatorios con reemplazo, entre 1 y 647. Luego, las unidades primarias por elegir en la muestra son aquellas cuyo rango en la M_i acumulativa incluye los números generados al azar. El conjunto de cinco números aleatorios {487, 369, 221, 326, 282} produce la muestra de las unidades {13, 9, 6, 8, 7}. El método del tamaño acumulativo permite que la misma unidad aparezca más de una vez: los cinco números aleatorios {553, 082, 245, 594, 150} conducen a la muestra {14, 3, 6, 14, 5}; así, la unidad primaria 14 se incluye dos veces en los datos. ■

Por supuesto, podemos extraer una muestra con probabilidades diferentes cuando las ψ_i no son proporcionales a las M_i: basta formar un rango acumulativo de ψ_i y extraer una muestra de números aleatorios uniformes entre 0 y 1. Esta variación del método se analiza en el ejercicio 4.

A menudo, el muestreo sistemático se utiliza para elegir las unidades primarias de muestreo en las muestras complejas y de gran tamaño, en vez de generar números aleatorios con reemplazo. En realidad, el muestreo sistemático produce una muestra sin reemplazo, pero en las grandes poblaciones, los muestreos con o sin reemplazo son muy similares, ya que la probabilidad de que una unidad se seleccione dos veces es pequeña. Para obtener una muestra sistemática de unidades primarias, enumeramos los elementos de la población que

TABLA 6.1
Población de los grupos de introducción a la estadística

Número de grupo	M_i	ψ_i	Rango M_i acumulativo	
1	44	0.068006	1	44
2	33	0.051005	45	77
3	26	0.040185	78	103
4	22	0.034003	104	125
5	76	0.117465	126	201
6	63	0.097372	202	264
7	20	0.030912	265	284
8	44	0.068006	285	328
9	54	0.083462	329	382
10	34	0.052550	383	416
11	46	0.071097	417	462
12	24	0.037094	463	486
13	46	0.071097	487	532
14	100	0.154560	533	632
15	15	0.023184	633	647
Total	647	1		

están en la primera unidad primaria de la muestra, seguidos de los elementos de la segunda unidad primaria y así sucesivamente. Después, extraemos una muestra sistemática de los elementos. Las unidades primarias que deben incluirse en la muestra son aquellas en las que al menos un elemento está en la muestra sistemática de elementos. Mientras mayor sea la unidad primaria, mayor será la probabilidad de que esté en la muestra.

Los grupos de la materia de estadística tienen un total de 647 estudiantes. Para extraer una muestra sistemática (aproximadamente, pues 647 no es múltiplo de 5), elegimos un número aleatorio k entre 1 y 129 y seleccionamos la unidad primaria que contiene al estudiante k, la unidad primaria que posee al estudiante $129 + k$, la unidad con el estudiante $2(129) + k$ y así sucesivamente. Suponga que el número aleatorio que escogimos como valor inicial es 112; entonces, la muestra sistemática de los elementos hace que elijamos las siguientes unidades primarias de muestreo:

Número en la muestra sistemática	Unidad primaria de muestreo elegida
112	4
241	6
370	9
499	13
628	14

Los grupos (unidades primarias) más numerosos tienen mayores posibilidades de estar en la muestra, pues es más probable que un múltiplo del número aleatorio elegido sea uno de los elementos numerados en una unidad primaria grande. Sin embargo, el muestreo sistemático no proporciona una verdadera muestra aleatoria con reemplazo, pues es imposible que los grupos con 129 o menos estudiantes aparezcan en la muestra más de una vez y los grupos con más de 129 estudiantes aparecen en la muestra con una probabilidad de 1. Sin embargo, en muchas poblaciones, es más fácil de implantar que los métodos que proporcionan una muestra aleatoria. Si las unidades primarias se ordenan geográficamente, la extracción de una muestra sistemática puede hacer que las unidades primarias seleccionadas se esparzan más ampliamente sobre la región y por ello se obtendrán mejores resultados que con una muestra aleatoria con reemplazo.

6.2.1.2 Método de Lahiri

El método de Lahiri (1951) puede ser más sencillo de utilizar que el método del tamaño acumulativo, cuando el número de unidades primarias es grande. El método de Lahiri es un ejemplo de método *con rechazo*, pues se generan parejas de números aleatorios para elegir las unidades primarias y, luego, rechaza algunas de ellas si el tamaño de la unidad primaria es demasiado pequeño. Sea N el número de unidades primarias en la población y $\max\{M_i\}$ el tamaño máximo de las unidades primarias. Usted comprobará en el ejercicio 14 que el método de Lahiri produce una muestra con reemplazo, con las probabilidades deseadas.

1. Extraiga un número aleatorio entre 1 y N. Esto indica la unidad primaria en cuestión.
2. Extraiga un número aleatorio entre 1 y $\max\{M_i\}$; si el número aleatorio es menor o igual a M_i, entonces, incluya la unidad primaria i en la muestra; en caso contrario, regrese al paso 1.
3. Repita hasta obtener el tamaño de muestra deseado.

EJEMPLO 6.3 Usemos el método de Lahiri con los grupos del ejemplo 6.2. Para usar el método de Lahiri sólo debemos conocer M_i para cada unidad primaria. El grupo más grande tiene $\max\{M_i\} = 100$ estudiantes, de modo que generamos parejas de números aleatorios, la primera entre 1 y 15 y la segunda entre 1 y 100, hasta que la muestra tenga cinco unidades primarias (vea la tabla 6.2). Las unidades primarias que conforman la muestra son: {12, 14, 14, 5, 1}. ■

6.2.2 Teoría de la estimación

Como realizamos una muestra con reemplazo, la muestra puede contener la misma unidad más de una vez. Para poder llevar un registro de las unidades primarias que aparecen varias veces en la muestra, definimos la variable aleatoria Q_i como:

Q_i = cantidad de veces que la unidad i aparece en la muestra.

TABLA 6.2
Método de Lahiri para el ejemplo 6.3

Primer número aleatorio (unidad primaria i)	Segundo número aleatorio	M_i	Acción
12	6	24	$6 < 24$; incluimos la unidad primaria 12 en la muestra
14	24	100	Incluir en la muestra
1	65	44	$65 > 44$; descartamos la pareja de números e intentamos de nuevo
7	84	20	$84 > 20$; intentamos de nuevo
10	49	34	Intentamos de nuevo
14	47	100	Lo incluimos
15	43	15	Intentamos de nuevo
5	24	76	Lo incluimos
11	87	46	Intentamos de nuevo
1	36	44	Lo incluimos

Entonces $\hat{t}_\psi$ es el promedio de todas las t_i/ψ_i para las unidades seleccionadas que estarán en la muestra:

$$\hat{t}_\psi = \frac{1}{n}\sum_{i=1}^{N} Q_i \frac{t_i}{\psi_i}. \tag{6.5}$$

Si una unidad aparece k veces en la muestra, se cuenta k veces en el estimador. Observe que $\sum_{i=1}^{N} Q_i = n$ y $E[Q_i] = n\psi_i$, de modo que $\hat{t}_\psi$ es insesgado para estimar t.

Para calcular la varianza, observe que el estimador en (6.5) es el promedio de n observaciones independientes, cada una con una varianza $\sum_{i=1}^{N} \psi_i(t_i/\psi_i - t)^2$ [de (6.4)], de modo que:

$$V[\hat{t}_\psi] = \frac{1}{n}\sum_{i=1}^{N} \psi_i \left(\frac{t_i}{\psi_i} - t\right)^2. \tag{6.6}$$

Para estimar $V[\hat{t}_\psi]$ a partir de una muestra, podría pensarse en utilizar una fórmula de la misma forma que (6.6), pero eso no funcionaría. La ecuación (6.6) implica un promedio ponderado de los $(t_i/\psi_i - t)^2$, ponderados por las probabilidades diferentes de selección, pero al extraer la muestra, ya hemos utilizado las probabilidades diferentes; éstas aparecen en las variables aleatorias Q_i de (6.5). Si nuevamente incluimos los ψ_i como multiplicadores al estimar la varianza muestral, estaríamos usando las probabilidades diferentes dos veces. En vez de ello, para estimar la varianza, usamos

$$\hat{V}(\hat{t}_\psi) = \frac{1}{n}\sum_{i=1}^{N} Q_i \frac{\left(\frac{t_i}{\psi_i} - \hat{t}_\psi\right)^2}{n-1}. \tag{6.7}$$

Observe que (6.7) no es más que una variación de la fórmula s^2/n utilizada en los cursos de introducción a la estadística: la suma es simplemente la varianza muestral de los números t_i/ψ_i para las unidades primarias de la muestra. La ecuación (6.7) es un estimador insesgado de la varianza en (6.6), pues

$$\begin{aligned} E[\hat{V}(\hat{t}_\psi)] &= \frac{1}{n(n-1)}\sum_{i=1}^{N} E\left[Q_i\left(\frac{t_i}{\psi_i} - \hat{t}_\psi\right)^2\right] \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{N} E\left[Q_i\left(\frac{t_i}{\psi_i} - t\right)^2 - Q_i(\hat{t}_\psi - t)^2\right] \\ &= \frac{1}{n(n-1)}\left[\sum_{i=1}^{N} n\psi_i\left(\frac{t_i}{\psi_i} - t\right)^2 - nV(\hat{t}_\psi)\right] \\ &= V(\hat{t}_\psi). \end{aligned}$$

Estamos obteniendo una muestra con reemplazo, de modo que la unidad i aparecerá en la muestra con la frecuencia aproximada $n\psi_i$. Una advertencia: si N es pequeño o alguno de los ψ_i es inusualmente grande, es posible que la muestra conste de una unidad primaria, obtenida n veces. En ese caso, la varianza estimada se anula; es mejor utilizar el muestreo sin reemplazo (vea la sección 6.4) si esto puede ocurrir.

EJEMPLO 6.4 Para la situación del ejemplo 6.3, suponga que extraemos una muestra de la unidad primaria seleccionada mediante el método de Lahiri, {12, 14, 14, 5, 1}. La respuesta t_i es la cantidad total de horas que todos los estudiantes del grupo i ocuparon para estudiar la materia de estadística la semana pasada, con los siguientes datos:

Grupo	ψ_i	t_i	t_i/ψ_i
12	$\frac{24}{647}$	75	2021.875
14	$\frac{100}{647}$	203	1313.410
14	$\frac{100}{647}$	203	1313.410
5	$\frac{76}{647}$	191	1626.013
1	$\frac{44}{647}$	168	2470.364

Los números de la última columna de la tabla son las estimaciones de t que se podrían obtener si esa unidad primaria fuese la única seleccionada en una muestra de tamaño 1. El total de la población se calcula al promediar los cinco valores de t_i/ψ_i:

$$\hat{t}_\psi = \frac{2021.875 + 1313.410 + 1313.410 + 1626.013 + 2470.364}{5} = 1749.014.$$

El error estándar (EE) de $\hat{t}_\psi$ es simplemente $s/\sqrt{n}$, donde s es la desviación estándar de la muestra de los cinco números en la última columna de la tabla:

$$\begin{aligned} \text{EE}[\hat{t}_\psi] &= \frac{1}{\sqrt{5}}\sqrt{\frac{(2021.875 - 1749.014)^2 + \cdots + (2470.364 - 1749.014)^2}{4}} \\ &= 222.42. \end{aligned}$$

La cantidad promedio de tiempo que un estudiante ocupó para estudiar estadística es:

$$\hat{\bar{y}}_\psi = \frac{1749.014}{647} = 2.70$$

de $\text{EE}(\hat{\bar{y}}_\psi) = 222.42/647 = 0.34$ horas. ■

6.2.3 Diseño de las probabilidades de selección

Quisiéramos elegir las ψ_i de modo que las varianzas de las estimaciones sean lo más pequeñas posible. Lo ideal es que usemos $\psi_i = t_i/t$ (en este caso, $\hat{t}_\psi = t$ para todas las muestras y $V[\hat{t}_\psi] = 0$), de modo que si t_i fue el ingreso anual de la familia i, ψ_i sería la proporción del ingreso total en la población proveniente de la familia i. Pero, por supuesto, las t_i no se conocen hasta obtener la muestra; aunque el ingreso fuese conocido antes de obtenerla, frecuentemente estamos interesados en más de una cantidad; el uso del ingreso para el diseño de las probabilidades de selección podría no funcionar tan bien para la estimación de otras cantidades.

Como muchos totales en una unidad primaria de muestreo están relacionados con la cantidad de elementos en una unidad primaria. Regularmente, consideramos a ψ_i como la proporción relativa de elementos en la unidad primaria i o el tamaño relativo de la unidad primaria i. En este caso, una unidad primaria mayor tiene más posibilidades de estar en la muestra que una unidad primaria pequeña. Si M_i es el número de elementos en la unidad

primaria i y K es el número de elementos en la población, hacemos $\psi_i = M_i/K$. Con esta elección de las probabilidades ψ_i, tenemos un muestreo con una **probabilidad proporcional al tamaño** (ppt). Utilizamos el muestreo con ppt en el ejemplo 6.2.

Entonces, para el muestreo con ppt de una etapa, $t_i/\psi_i = K\bar{y}_i$, de modo que

$$\hat{t}_\psi = \frac{K}{n}\sum_{i=1}^{N} Q_i \bar{y}_i,$$

$$\hat{\bar{y}}_\psi = \frac{1}{n}\sum_{i=1}^{N} Q_i \bar{y}_i,$$

$$\hat{V}(\hat{t}_\psi) = \frac{1}{n}\sum_{i=1}^{N} Q_i \frac{\left(\frac{t_i}{\psi_i} - \hat{t}_\psi\right)^2}{n-1} = \frac{K^2}{n}\sum_{i=1}^{N} Q_i \frac{(\bar{y}_i - \hat{\bar{y}}_\psi)^2}{n-1}$$

$$\hat{V}(\hat{\bar{y}}_\psi) = \frac{1}{n}\sum_{i=1}^{N} \frac{Q_i(\bar{y}_i - \hat{\bar{y}}_\psi)^2}{n-1}.$$

La suma en las estimaciones de la varianza es, simplemente, la varianza muestral de las medias $\bar{y}_i$' de la unidad primaria.

Todo el trabajo en el muestreo con ppt se realiza en el propio diseño de muestreo. Las estimaciones con ppt se pueden calcular, al considerar los $\bar{y}_i$' simplemente como observaciones individuales y determinar su media y varianza muestral. En la práctica, sin embargo, generalmente hay algunas desviaciones a partir de un esquema estricto con ppt, de manera que usted debe usar (6.5) y (6.7) para estimar el total de la población y su varianza estimada.

EJEMPLO 6.5 El archivo statepop.dat contiene los datos de una muestra con probabilidades diferentes de 100 condados de los Estados Unidos. Los condados fueron elegidos mediante el método de tamaño acumulativo a partir de las listas del *City and County Data Book, 1994*, con probabilidades proporcionales a sus poblaciones. El muestreo se realizó con reemplazo, de modo que los condados muy grandes aparecen varias veces en la muestra: el condado de Los Ángeles, con la mayor población de Estados Unidos, aparece cuatro veces.

Una de las cantidades registradas para cada condado fue el número de médicos en el condado. Es de esperar que los condados más grandes tengan más médicos, de forma que el muestreo con ppt debe funcionar bien para estimar la cantidad total de médicos que hay en Estados Unidos.

Se debe tener cuidado al graficar los datos de una muestra con probabilidades diferentes, pues se deben tomar en cuenta esas probabilidades diferentes al interpretar las gráficas. Una gráfica de t_i contra ψ_i (vea la figura 6.1a) muestra la eficiencia de un diseño con probabilidades diferentes: mientras más se acerque la gráfica a una línea recta, mejor funcionará el muestreo con probabilidades diferentes. Un histograma de t_i en una muestra con ppt no dará una visión representativa de la población de unidades primarias de muestreo, ya que las unidades primarias con ψ_i grande están sobrerrepresentadas en la muestra. Sin embargo, un histograma de t_i/ψ_i puede dar una idea de la difusión implicada en las estimaciones de la población y puede ayudar a identificar las unidades primarias poco usuales (observe la figura 6.1b).

La muestra se seleccionó mediante el método de tamaño acumulativo; la tabla 6.3 presenta los condados que componen la muestra, en orden alfabético por estado. Los ψ_i se calcularon como $M_i/255{,}077{,}536$.

FIGURA 6.1
Dos gráficas para la estimación del número total de médicos en Estados Unidos mediante una muestra con ppt. (a) Gráfica de t_i contra ψ_i; existe una fuerte relación lineal entre las variables, lo que indica que el muestreo con ppt aumenta la eficiencia. La observación inusual es el condado de Nueva York. (b) Histograma de los 100 valores de t_i/ψ_i. Cada valor estima a t.

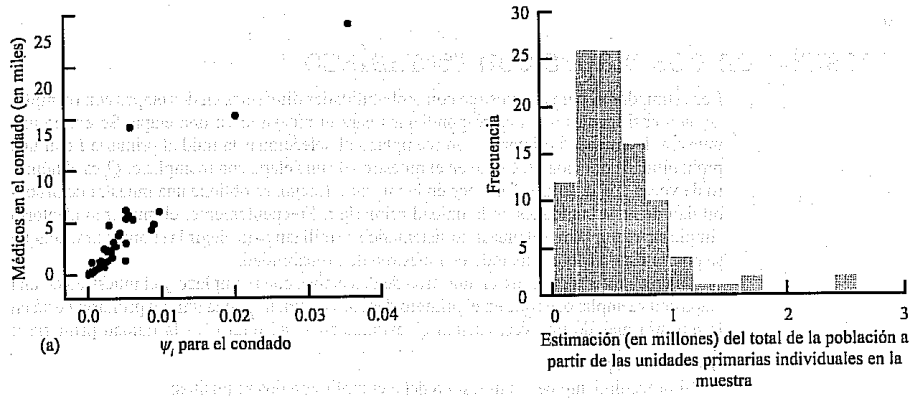


TABLA 6.3
Condados en la muestra del ejemplo 6.5

Estado	Condado	Tamaño de la población, M_i	ψ_i	Número de médicos, t_i	t_i/ψ_i
AL	Wilcox	13,672	0.00005360	4	74,627.72
AZ	Maricopa	2,209,567	0.00866233	4320	498,710.81
AZ	Maricopa	2,209,567	0.00866233	4320	498,710.81
AZ	Pinal	120,786	0.00047353	61	128,820.64
AR	Garland	76,100	0.00029834	131	439,095.36
AR	Mississippi	55,060	0.00021586	48	222.370.54
CA	Contra Costa	840,585	0.00329541	1761	534,379.68
⋮	⋮	⋮	⋮	⋮	⋮
VA	Chesterfield	225,225	0.00088297	181	204,990.72
WA	King	1,557,537	0.00610613	5280	864.704.59
WI	Lincoln	27,822	0.00010907	28	256,709.47
WI	Waukesha	320,306	0.00125572	687	547,096.42
		promedio			570,304.30
		desv. estándar			414,012.30

El promedio de la columna t_i/ψ_i es 570,304.3, la cantidad total estimada de médicos en Estados Unidos. El error estándar de la estimación es $414,012.3/\sqrt{100} = 41,401.23$. Como comparación, el *City and County Data Book* enumera un total de 532,638 médicos, cifra que está a menos de 1 EE de nuestra estimación. ▪

6.3
Muestreo en dos etapas con reemplazo

Los estimadores, para el muestreo con probabilidades diferentes en dos etapas con reemplazo, son casi iguales a los correspondientes para el muestreo en una etapa. Se extrae una muestra de unidades primarias con reemplazo al seleccionar la unidad primaria i con una probabilidad conocida ψ_i. Como en el muestreo de una etapa con reemplazo, Q_i es el número de veces que la unidad i aparece en la muestra. Luego, se obtiene una muestra de probabilidad de m_i subunidades en la unidad primaria i. Frecuentemente, el muestreo aleatorio simple sin reemplazo o el muestreo sistemático se utilizan para elegir la submuestra, aunque se podría usar cualquier método de muestreo de probabilidad.

La única diferencia entre el muestreo de dos etapas con reemplazo y el muestreo de una etapa con reemplazo, es que en el primero debemos estimar t_i. Si la unidad primaria i está en la muestra más de una vez, existen Q_i estimaciones del total para la unidad primaria i; $\hat{t}_{i1}, \hat{t}_{i2}, \cdots, \hat{t}_{iQi}$.

El procedimiento de submuestreo debe cumplir con dos requisitos:

1 Siempre que se elija para estar en la muestra la unidad primaria i, se utiliza el mismo diseño de submuestreo para seleccionar las unidades secundarias de muestreo a partir de esa unidad primaria. Empero, cada submuestra de la misma unidad primaria debe obtenerse independientemente. Así, si decide (antes de realizar el muestreo) que tomará una muestra aleatoria simple de tamaño 5 de la unidad primaria 42, si ésta fue la seleccionada, cada vez que en la muestra aparezca la unidad primaria 42 generará un conjunto distinto de números aleatorios para elegir cinco unidades secundarias en la unidad primaria 42. CUIDADO: si sólo extrae una submuestra de tamaño 5 y la utiliza más de una vez para la unidad primaria 42, no tendrá submuestras independientes y (6.9) no será un estimador insesgado de la varianza.

2 La j-ésima submuestra extraída de la unidad primaria i (para $j = 1, \ldots, Q_i$) se elige de modo que $E[\hat{t}_{ij}] = t_i$. Como se emplea el mismo procedimiento cada vez que la unidad primaria i es elegida para la muestra, podemos definir $V[\hat{t}_{ij}] = V_i$ para toda j.

Los estimadores del muestreo con probabilidades diferentes, de una etapa con reemplazo, se modifican ligeramente para permitir el uso de distintas submuestras de las unidades primarias elegidas más de una vez:

$$\hat{t}_\psi = \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{\hat{t}_{ij}}{\psi_i}. \tag{6.8}$$

$$\hat{V}(\hat{t}_\psi) = \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{\left(\dfrac{\hat{t}_{ij}}{\psi_i} - \hat{t}_\psi\right)^2}{n-1}. \tag{6.9}$$

En el ejercicio 15 se mostrará que (6.9) es un estimador insesgado de la varianza $V(\hat{t}_\psi)$, que aparece en (6.27). Como el muestreo es con reemplazo y, por tanto, es posible tener más de una submuestra de una unidad primaria dada, el estimador de la varianza captura las dos partes de la varianza: la parte que se debe a la variabilidad entre las unidades primarias y la parte que surge debido a que t_i se estima a partir de una submuestra, en lugar de observarse. La media de la población se estima como $\hat{\bar{y}}_\psi = \hat{t}_\psi/K$, con varianza estimada $\hat{V}(\hat{\bar{y}}_\psi) = \hat{V}(\hat{t}_\psi/K^2$.

En una muestra con ppt, en la que la unidad primaria i se elige con probabilidad $\psi_i = M_i/K$, los estimadores se simplifican de nuevo. En este caso, $\hat{\bar{y}}_\psi$ es simplemente el promedio de las medias estimadas de las unidades primarias de la muestra y $\hat{V}(\hat{\bar{y}}_\psi)$ es la varianza muestral de dichas medias estimadas de las unidades primarias, dividida entre n. Los tamaños de las submuestras no aparecen en las estimaciones.

En resumen, he aquí los pasos para extraer una muestra con probabilidades diferentes, de dos etapas con reemplazo:

1 Determinamos las probabilidades de selección ψ_i, la cantidad n de unidades primarias por incluir en la muestra y el procedimiento de submuestreo que se utilizará dentro de cada unidad primaria. Una vez elegido cualquier método para seleccionar las unidades primarias, extraemos una muestra de probabilidad de unidades secundarias dentro de las unidades primarias. Con frecuencia, en un muestreo por conglomerados de dos etapas extraemos una muestra aleatoria simple sin reemplazo de los elementos que se encuetran dentro de la unidad primaria elegida.

2 Elegimos n unidades primarias con probabilidades ψ_i y con reemplazo. Se puede usar el método de tamaño acumulativo o el método de Lahiri para elegir las unidades primarias de la muestra.

3 Utilizamos el procedimiento determinado en el paso 1 para seleccionar las submuestras de las unidades primarias elegidas. Si una unidad primaria aparece en la muestra más de una vez, hay que usar submuestras independientes para cada copia.

4 Estimamos el total de la población t de cada unidad primaria en la muestra como si fuese la única elegida. El resultado son n estimaciones de la forma $\hat{t}_{ij}/\psi_i$.

5 $\hat{t}_\psi$ es el promedio de las n estimaciones del paso 4.

6 $\text{EE}(\hat{t}_\psi) = (1/\sqrt{n})$ (desviación estándar muestral de las n estimaciones del paso 4).

EJEMPLO 6.6 Regresemos al ejemplo 6.4. Ahora supongamos que extraemos una submuestra compuesta por cinco estudiantes de cada grupo en lugar de observar t_i. Veremos que el proceso de estimación es casi el mismo que en el ejemplo 6.4; en este caso, la respuesta y_{ij} es la cantidad total de horas que el alumno j del grupo i ocupó en estudiar la materia de estadística la semana pasada (vea la tabla 6.4). Observe que el grupo 14 aparece dos veces en la muestra; cada vez que lo hace, se reúne una submuestra distinta.

Así, $\hat{t}_\psi = 1617.5$ y $\text{EE}(\hat{t}_\psi) = 521.628/\sqrt{5} = 233.28$. A partir de esta muestra, la cantidad promedio que un alumno ocupó en estudiar estadística es:

$$\hat{\bar{y}}_\psi = \frac{1617.5}{647} = 2.5$$

de $\text{EE}(\hat{\bar{y}}_\psi) = 233.28/647 = 0.36$ horas. ▪

TABLA 6.4
Hoja de cálculo para el ejemplo 6.6

Grupo	M_i	ψ_i	y_{ij}	$\bar{y}_i$	$\hat{t}_i$	t_i/ψ_i
12	24	0.0371	2, 3, 2.5, 3, 1.5	2.4	57.6	1552.8
14	100	0.1546	2.5, 2, 3, 0, 0.5	1.6	160.0	1035.2
14	100	0.1546	3, 0.5, 1.5, 2, 3	2.0	200.0	1294.0
5	76	0.1175	1, 2.5, 3, 5, 2.5	2.8	212.8	1811.6
1	44	0.0680	4, 4.5, 3, 2, 5	3.7	162.8	2393.9
			Promedio			1617.5
			desv. estándar			521.628

En el ejemplo 6.6 elegimos los grupos con una probabilidad proporcional a la cantidad de estudiantes del grupo, de modo que $\psi_i = M_i/K$. Al extraer una submuestra con el mismo número de estudiantes en cada grupo, obtuvimos una muestra autoponderada. Bajo el muestreo con ppt y con reemplazo, con un muestreo aleatorio simple en la segunda etapa, el peso de muestreo para un elemento seleccionado de la unidad primaria i, de (6.8),

$$\omega_i = \frac{1}{n}\frac{M_i}{m_i}\frac{1}{\psi_i}.$$

En el muestreo con ppt, $\psi_i = M_i/K$, tenemos que $w_i = K/(nm_i)$; la muestra es autoponderada si todas las m_i son iguales. Para el ejemplo 6.6, el peso de muestreo es de $647/(5 \times 5) = 25.88$ para cada observación. El total de la población se estima de manera equivalente como:

$$\frac{647}{25}(2+3+2.5+\cdots+3+2+5) = 1617.5.$$

EJEMPLO 6.7 Veamos lo que ocurre si usamos el muestreo con probabilidades diferentes en el caso de las perreras del ejemplo 5.7. Sea ψ_i proporcional la cantidad de cachorros en una perrera, de modo que Puppy Palace, con 30 cachorros, participa en la muestra con una probabilidad de $\frac{3}{4}$ y Dog's Life, con 10, perritas participa en la muestra con una probabilidad de $\frac{1}{4}$. Al igual que antes, y una vez elegida una perrera, extraemos una muestra aleatoria simple de dos cachorros de la perrera en cuestión. Entonces, si se elige Puppy Palace, $\hat{t}_\psi = \hat{t}_{PP}/(3/4) = (30)(4)/(3/4) = 160$. En caso de que se elija Dog's Life, $\hat{t}_\psi = \hat{t}_{DL}/(1/4) = (10)(4)/(1/4) = 160$. Así, cualquiera de las muestras posibles produce un promedio estimado de $\bar{y}_\psi = 160/40 = 4$ patas por cachorro y la varianza del estimador es igual a cero. ■

El muestreo con reemplazo tiene la ventaja de que es muy fácil seleccionar la muestra y obtener estimaciones del total de la población y la varianza correspondiente. Sin embargo, si N es pequeño, como ocurre en muchas encuestas complejas altamente estratificadas con pocos conglomerados en cada estrato, el muestreo con reemplazo es menos eficiente que muchos diseños para el muestreo sin reemplazo. En la siguiente sección analizamos las ventajas y los retos del muestreo sin reemplazo.

6.4 Muestreo con probabilidades diferentes sin reemplazo

A pesar de que generalmente el muestreo con reemplazo es menos eficiente que el muestreo sin reemplazo, se utiliza debido a la facilidad que brinda para elegir y analizar las muestras. Sin embargo, en las grandes encuestas que presentan muchos estratos pequeños, las deficiencias pueden superar las ventajas. Se ha investigado mucho acerca del muestreo con probabilidades diferentes sin reemplazo; empero, la teoría es más complicada, pues la probabilidad de elegir una unidad es distinta para la primera unidad seleccionada que para la segunda, la tercera y las unidades subsecuentes. Sin embargo, cuando comprenda los argumentos probabilísticos implicados en el proceso, podrá determinar las propiedades de cualquier esquema de muestreo.

EJEMPLO 6.8 El ejemplo del supermercado, de la sección 6.1, puede ilustrar algunas características del muestreo con probabilidades diferentes con reemplazo. He aquí la población de nuevo:

Tienda	Tamaño (m^2)	t_i (en miles)
A	100	11
B	200	20
C	300	24
D	1000	245
Total	1600	300

Elijamos dos unidades primarias sin reemplazo y con probabilidades diferentes. Como en las secciones 6.1 a 6.3, sea

$$\psi_i = P(\text{elegir la unidad } i \text{ en la primera extracción})$$

Sin embargo, como estamos extrayendo una muestra sin reemplazo, la probabilidad de elegir la unidad j en la segunda extracción dependerá de la unidad elegida en la primera extracción.

Una forma de elegir las unidades con probabilidades diferentes es emplear ψ_i como la probabilidad de seleccionar la unidad i en la primera extracción, para luego ajustar las probabilidades de selección de las demás tiendas en la segunda extracción. Si se eligió la tienda A en la primera extracción, entonces, para escoger la segunda tienda, podríamos hacer girar la ruleta y bloquear la sección de la tienda A o revolver la baraja y extraer una carta sin la carta 1. Así,

$$P(\text{elegir la tienda A en la primera extracción}) = \psi_A = 1/16$$

y

$$P(\text{elegir B en la segunda extracción} \mid \text{A fue elegida en la primera extracción}) = \frac{\frac{2}{16}}{1-\frac{1}{16}} = \frac{\psi_B}{1-\psi_A}.$$

El denominador es la suma de las ψ_i para las tiendas B, C y D. En general,

$P(\text{elegir primero la unidad } i \text{ y luego la unidad } j) =$

$= P(\text{elegir primero la unidad } i)\, P(\text{elegir en segundo lugar la unidad } j \mid \text{la unidad } i \text{ fue elegida primero})$

$$= \psi_i \frac{\psi_j}{1-\psi_i}.$$

De manera análoga,

$$P(\text{elegir primero la unidad } j \text{ y luego la unidad } i) = \psi_j \frac{\psi_j}{1-\psi_j}.$$

Observe que la P(elegir primero la unidad i y luego la unidad j) no es igual que la P(elegir primero la unidad j y luego la unidad i): ¡El orden de selección establece una diferencia! Sin embargo, al sumar las probabilidades de las dos opciones, podemos determinar la probabilidad de que una muestra de tamaño 2 conste de las unidades primarias i y j:

$$\text{Para } n = 2,\ P(\text{las unidades } i \text{ y } j \text{ estén en la muestra}) = \pi_{ij} = \psi_i \frac{\psi_j}{1-\psi_i} + \psi_j \frac{\psi_i}{1-\psi_j}.$$

Para una muestra de tamaño 2, la probabilidad de que la unidad primaria i esté en la muestra es, entonces, la suma sobre j de las probabilidades de que las unidades i y j se encuentren en

la muestra:

$$\text{Para } n=2,\ P(\text{la unidad } i \text{ esté en la muestra}) = \pi_i = \sum_{\substack{j=1 \\ j\neq i}}^{N} \pi_{ij}.$$

La siguiente tabla muestra los valores de π_i y π_{ij} para los supermercados. Las entradas de la tabla son π_{ij} para cada pareja de tiendas (al redondear a cuatro cifras decimales); los márgenes dan los valores de π_i para las cuatro tiendas.

		Tienda j				
		A	B	C	D	π_i
Tienda i	A	—	.0173	.0269	.1458	.1900
	B	.0173	—	.0556	.2976	.3705
	C	.0269	.0556	—	.4567	.5393
	D	.1458	.2976	.4567	—	.9002
	π_j	.1900	.3705	.5393	.9002	2.0000

■

6.4.1 El estimador de Horvitz-Thompson

En el muestreo sin reemplazo, π_i es la *probabilidad de inclusión*; esto es, la probabilidad de que la unidad i esté en la muestra; π_{ij} es la probabilidad de que las unidades i y j se encuentren en la muestra. La probabilidad de inclusión π_i se calcula como la suma de las probabilidades de todas las muestras que contienen la unidad i y tiene la propiedad de que

$$\sum_{i=1}^{N} \pi_i = n. \tag{6.10}$$

Para las π_{ij}, como mostraremos en el teorema 6.1 de la sección 6.6,

$$\sum_{\substack{j=1 \\ j\neq i}}^{N} \pi_{ij} = (n-1)\pi_i. \tag{6.11}$$

Para los supermercados, las π_i resultantes no son proporcionales a los tamaños de las tiendas; de hecho, no pueden serlo, pues la tienda D representa más de la mitad del área total, pero no se puede formar parte de la muestra con una probabilidad mayor a 1. Las π_i resultantes de este método de extracción por extracción, formulado por Yates y Grundy (1953) pueden o no ser las probabilidades de inclusión que desean tener en la muestra; tal vez, se necesite ajustar las ψ_i para obtener un conjunto predeterminado de π_i.

En el ejemplo 6.8, $\pi_A = P(\text{la tienda A esté en la muestra}) = 0.19$, y la suma de las π_i es igual a 2. Así, π_i/N es la *probabilidad promedio* de que una unidad sea seleccionada en una de las extracciones: es la probabilidad que asignaríamos a la unidad i elegida en la extracción k ($k = 1, \ldots, n$) si no conociéramos las probabilidades reales.

Recuerde que para el muestreo con reemplazo $\hat{t}_\psi$ es el promedio de $\hat{t}_{ij}/\psi_i$ para las unidades primarias de la muestra. Pero cuando las muestras se extraen sin reemplazo, las probabilidades de selección dependen de las unidades que se eligieron anteriormente. En vez de dividir el total estimado para la unidad primaria i entre ψ_i, dividimos entre la probabilidad *promedio* de elegir esa unidad en una extracción, π_i/n. Tenemos, entonces, el **estimador de Horvitz-Thompson (HT)** del total de la población (Horvitz y Thompson 1952):

$$\hat{t}_{\text{HT}} = \frac{1}{n}\sum_{i\in S}\frac{\hat{t}_i}{\pi_i/n} = \sum_{i\in S}\frac{\hat{t}_i}{\pi_i} = \sum_{i=1}^{N} Z_i \frac{\hat{t}_i}{\pi_i}, \tag{6.12}$$

donde $Z_i = 1$ si la unidad i está en la muestra y 0 en caso contrario.

Se muestra fácilmente que el estimador de Horvitz-Thompson es insesgado para t, al usar el teorema 6.2, que se demostrará en la sección 6.6. En este caso, $P(Z_i = 1) = \pi_i$, de modo que por (6.19),

$$E[\hat{t}_{\text{HT}}] = \sum_{i=1}^{N} \pi_i \frac{t_i}{\pi_i} = t.$$

Si usamos (6.20) a (6.22), la varianza del estimador de Horvitz-Thompson es

$$\begin{aligned} V(\hat{t}_{\text{HT}}) &= \sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}t_i^2 + \sum_{i=1}^{N}\sum_{k\neq i}^{N}\frac{\pi_{ik}-\pi_i\pi_k}{\pi_i\pi_k}t_i t_k + \sum_{i=1}^{N}\frac{V(\hat{t}_i)}{\pi_i} \\ &= \sum_{i=1}^{N}\sum_{k>i}^{N}(\pi_i\pi_k - \pi_{ik})\left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k}\right)^2 + \sum_{i=1}^{N}\frac{V(\hat{t}_i)}{\pi_i}. \end{aligned} \tag{6.13}$$

La segunda expresión en (6.13) es la forma Sen-Yates-Grundy (Sen 1973; Yates y Grundy 1953).

El teorema 6.3 de la sección 6.6 implica que

$$\hat{V}_1[\hat{t}_{\text{HT}}] = \sum_{i\in S}(1-\pi_i)\frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i\in S}\sum_{\substack{k\in S \\ k\neq i}}\frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\frac{\hat{t}_i}{\pi_i}\frac{\hat{t}_k}{\pi_k} + \sum_{i\in S}\frac{\hat{V}(\hat{t}_i)}{\pi_i} \tag{6.14}$$

y la forma de Sen-Yates-Grundy,

$$\hat{V}_2[\hat{t}_{\text{HT}}] = \sum_{i\in S}\sum_{\substack{k\in S \\ k>i}}\frac{\pi_i\pi_k - \pi_{ik}}{\pi_{ik}}\left(\frac{\hat{t}_i}{\pi_i} - \frac{\hat{t}_k}{\pi_k}\right)^2 + \sum_{i\in S}\frac{\hat{V}(\hat{t}_i)}{\pi_i} \tag{6.15}$$

son estimadores insesgados de la varianza en (6.13). Sin embargo, puede surgir un problema al estimar la varianza de $\hat{t}_{\text{HT}}$. ¡Los estimadores insesgados en (6.14) o (6.15) pueden producir una estimación negativa de la varianza en ciertos diseños con probabilidades diferentes! A veces, la estabilidad se puede mejorar mediante una elección cuidadosa del diseño de muestreo; pero, en general, los cálculos son complicados.

Una alternativa, que evita algo de la inestabilidad potencial y la complejidad de los cálculos, consiste en emplear el estimador de la varianza con reemplazo en (6.9) en vez de (6.14) o (6.15), como sugirió Durbin (1953). Si el muestreo sin reemplazo es más eficiente que el muestreo con reemplazo, se espera que el estimador de la varianza con reemplazo en (6.9) sobrestime la varianza y produzca intervalos de confianza conservadores; aunque, en muchos casos, el sesgo es pequeño. Los métodos que utilizan ampliamente las computadoras, descritos en el capítulo 9, calculan la varianza con reemplazo.

Observe que para usar el estimador de Horvitz-Thompson cuando $n > 1$, debemos conocer la probabilidad de inclusión π_i para cada unidad primaria. El procedimiento de extracción por extracción, utilizado en el ejemplo de los supermercados (determinar la probabilidad de que cualquier pareja de unidades primarias esté en la muestra para luego calcular la probabilidad general de que la unidad i esté en la muestra) se vuelve algo tedioso para las poblaciones

grandes y tamaños de muestra mayores que 2. El muestreo sistemático se puede emplear para extraer una muestra sin reemplazo y se puede implantar con relativa facilidad (de ahí su uso tan extendido), pero muchas de las π_{ij} para la población son iguales a cero. Brewer y Hanif (1983) presentan más de 50 métodos para seleccionar muestras con probabilidades diferentes sin reemplazo. La mayoría de estos métodos son para $n = 2$. Algunos métodos son más fáciles de utilizar, otros son más adecuados para aplicaciones específicas mientras que otros dan una estimación más estable de la varianza del estimador de Horvitz-Thompson para t.

6.4.2 Pesos en las muestras con probabilidades diferentes

Todos los esquemas de muestreo sin reemplazo, analizados hasta ahora, pueden considerarse como casos particulares del muestreo por conglomerados de dos etapas, (posiblemente) con probabilidades diferentes.

En el estimador de Horvitz-Thompson, el peso de muestreo para la unidad primaria i es:

$$\omega_i = \frac{1}{\pi_i}.$$

Así, el estimador de Horvitz-Thompson para el total de la población es

$$\hat{t}_{\mathrm{HT}} = \sum_{i \in S} \omega_i \hat{t}_i.$$

Para una muestra de probabilidad sin reemplazo de las unidades secundarias que se encuentran dentro de las unidades primarias, podemos definir, al seguir la notación de Särndal *et al.* (1992),

$\pi_{j|i} = P$(la unidad secundaria j de la unidad primaria i se incluye en la muestra | la unidad primaria i está en la muestra).

Entonces,

$$\hat{t}_i = \sum_{j \in S_i} \frac{y_{ij}}{\pi_{j|i}}.$$

La probabilidad global de elegir el elemento (i, j) es $\pi_{j|i}\pi_i$. Así, podemos definir el peso de muestreo para el elemento (i, j) como:

$$\omega_{ij} = \frac{1}{\pi_{j|i}\pi_i} \tag{6.16}$$

y el estimador de Horvitz-Thompson para el total de la población como:

$$\hat{t}_{\mathrm{HT}} = \sum_{i \in S} \sum_{j \in S_i} \omega_{ij} y_{ij}. \tag{6.17}$$

Estimamos la media de la población como:

$$\hat{\bar{y}}_{\mathrm{HT}} = \frac{\sum_{i \in S} \sum_{j \in S_i} \omega_{ij} y_{ij}}{\sum_{i \in S} \sum_{j \in S_i} \omega_{ij}}. \tag{6.18}$$

6.4.3 El estimador de Horvitz-Thompson para diseños generales sin reemplazo

En la sección 5.8 estudiamos que las fórmulas para el muestreo estratificado son un caso particular de las fórmulas para el muestreo por conglomerados en dos etapas. De hecho, todas las fórmulas para la estimación insesgada de los totales en el muestreo sin reemplazo (que se presentaron en los capítulos 2, 4, 5 y 6) son casos particulares de (6.12) a (6.15).

Por ejemplo, en el muestreo aleatorio simple, una unidad primaria es un elemento individual y $t_i = y_i$. En el apéndice B mostraremos que

$$\pi_i = P(Z_i = 1) = \frac{n}{N}$$

y

$$\pi_{ij} = P(Z_i = 1, Z_j = 1) = \frac{n(n-1)}{N(N-1)}.$$

Así, para el muestreo aleatorio simple,

$$\hat{t}_{\mathrm{HT}} = \sum_{i=1}^{N} Z_i \frac{N}{n} y_i = N\bar{y},$$

y

$$V(\hat{t}_{\mathrm{HT}}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_{i=1}^{N} \sum_{k \neq i}^{N} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} y_i y_k$$

$$= \sum_{i=1}^{N} \frac{1 - \frac{n}{N}}{\frac{n}{N}} y_i^2 + \sum_{i=1}^{N} \sum_{k \neq i}^{N} \frac{\frac{n-1}{N-1} - \frac{n}{N}}{\frac{n}{N}} y_i y_k$$

$$= \cdots = N^2 \left(1 - \frac{n}{N}\right) \frac{S^2}{n} = V(N\bar{y}).$$

En el ejercicio 18, se demostrará que las fórmulas para el muestreo estratificado son un caso particular de la estimación de Horvitz-Thompson.

6.5 Ejemplos de muestras con probabilidades diferentes

Muchas situaciones de muestreo son adecuadas para las muestras con probabilidades diferentes. En esta sección daremos tres ejemplos comunes de diseños de muestreo.

EJEMPLO 6.9 *Marcado de dígitos aleatorios*

En las encuestas telefónicas es importante tener un procedimiento eficiente y bien definido para generar los números telefónicos que aparecerán en la muestra. Cuando se empezaron a utilizar las encuestas telefónicas, muchas empresas simplemente tomaban números del directorio telefónico. Sin embargo, ese punto de vista conducía a un sesgo de selección, pues hay números telefónicos que no aparecen en el directorio, que tampoco contiene los números agregados después de su publicación. Se han sugerido modificaciones al muestreo, que toma como base al directorio telefónico, para poder incluir los números que no aparecen en él, pero la mayoría de estas modificaciones tienen algunas dificultades con la subcobertura.

Muestreo de elementos de marcado de dígitos aleatorios La generación de números telefónicos al azar a partir del marco de todos los números telefónicos posibles evita la subcobertura de los números no enlistados. En Estados Unidos, los números telefónicos constan de

código de área	+	prefijo (o código de intercambio)	+	sufijo
(3 dígitos)		(3 dígitos)		(4 dígitos)

Así, podemos elegir una muestra aleatoria de los números telefónicos en Estados Unidos seleccionando al azar una combinación de código de área y prefijo, agregando un número de cuatro dígitos elegido al azar, de 0000 a 9999. Si el número aleatorio elegido no pertenece a una familia, lo descartamos e intentamos con una nueva combinación de 10 dígitos.

Este método es fácil de comprender y explicar; suponiendo que no hay ausencia de respuestas, produce una muestra aleatoria simple a partir del marco de todos los números telefónicos posibles. El método es autoponderado, pues esperamos marcar números residenciales a partir de un prefijo, con una tasa proporcional a la frecuencia relativa de los números que comiencen con ese prefijo. En la práctica, el método puede ser caro: Lepkowski (1988) informa que menos de 25% de todos los números telefónicos potenciales generados por este método pertenecen a familias. Tal vez se regresan varias llamadas para poder comprobar si el número es o no residencial.

El método de Mitofsky-Waksberg Mitofsky (1970) y Waksberg (1978) desarrollaron un método por conglomerados para el muestreo de los números telefónicos residenciales. A continuación describimos el procedimiento "utopía muestral", donde todos contestan el teléfono; Lavrakas (1993) y Potthoff (1994) dan sugerencias sobre la forma de emplear el método de Mitofsky-Waksberg cuando los residentes no son tan cooperativos.

1 Construya un marco de todos los códigos de área y prefijos en el área de interés.

2 Extraiga una muestra aleatoria de números telefónicos de 10 dígitos a partir del conjunto de números con código de área y prefijo en el marco de referencia y sufijo entre 0000 y 9999. Después del paso 2, usted tendrá una muestra de números telefónicos, exactamente como en el muestreo de elementos de marcado de dígitos aleatorios.

3 Marque cada número seleccionado en el paso 2. Si el número seleccionado es residencial, entreviste a la familia y elija su unidad primaria para incluirla en la muestra; la unidad primaria asociada es el bloque de 100 números telefónicos que tienen los mismos primeros ocho dígitos que el número seleccionado. Por ejemplo, si se determina que el número telefónico elegido al azar (202)456-1414 es residencial, entonces se incluye en la muestra la unidad primaria de todos los números de la forma (202)456-14xx. Los números telefónicos conservados son una muestra aleatoria simple de las casas de la región. Si el número seleccionado no es residencial, lo descartamos junto con su unidad primaria. Continuamos el muestreo en la primera etapa hasta seleccionar la cantidad deseada de unidades primarias.

4 Para la segunda etapa del muestreo, elegimos al azar otros números telefónicos sin reemplazo a partir de cada unidad primaria de la muestra, hasta alcanzar el tamaño de muestra deseado para cada unidad primaria.

El método de Mitofsky-Waksberg aumenta dramáticamente el porcentaje de llamadas realizadas que logran llegar a las casas. Lepkowski (1988) determinó que 60% de los números telefónicos elegidos en la segunda etapa eran residenciales, en comparación con 25% para el muestreo mediante dígitos aleatorios. El método funciona debido a que las unidades primarias de 100 números telefónicos forman conglomerados: algunas unidades primarias no están asignadas, otras tienden a pertenecer a establecimientos comerciales y otras son mayoritariamente residenciales. El procedimiento de dos etapas elimina del muestreo las unidades primarias no asignadas en la segunda etapa y reduce la probabilidad de seleccionar unidades primarias con pocos números telefónicos residenciales.

En condiciones ideales, el procedimiento de Mitofsky-Waksberg extrae la muestra de las unidades primarias con probabilidades proporcionales a la cantidad de números telefónicos residenciales en las unidades primarias. Si la segunda etapa indica que deben elegirse

otros $(k-1)$ números telefónicos residenciales en cada unidad primaria de la muestra y si todas las unidades primarias de la muestra tienen al menos k números telefónicos residenciales, entonces el procedimiento de Mitofsky-Waksberg da la misma probabilidad de selección a cada número telefónico residencial; el resultado es una muestra autoponderada de números telefónicos residenciales.

Nuestro procedimiento de muestreo con ppt requiere que conozcamos las probabilidades de selección de todas las unidades primarias de la muestra. En el procedimiento de Mitofsky-Waksberg se desconocen las probabilidades antes del muestreo, aunque se pueden calcular los pesos. Si M_i es la cantidad de números telefónicos residenciales en la i-ésima unidad primaria y k la cantidad de números telefónicos residenciales en cada unidad primaria seleccionados para estar en la muestra. Entonces

$$P(\text{seleccionar un número}) = P(\text{seleccionar la } i\text{-ésima unidad primaria})P(\text{seleccionar el número} \mid \text{la unidad primaria se seleccionó})$$

$$\alpha \frac{M_i}{K}\frac{k}{M_i} = \frac{k}{K}.$$

Para estimar el total de una población, hay que conocer K, la cantidad total de números telefónicos residenciales en la población y usar los pesos de muestreo $u_{ij} = K/k$.

Para estimar un promedio o proporción, que es el objetivo típico de las encuestas telefónicas, no hay que conocer K, sino únicamente un "peso relativo" w_{ij} para cada respuesta y_{ij} de la muestra y se puede estimar la media de la población como

$$\hat{\bar{y}} = \frac{\sum_{i\in S}\sum_{j\in S_i}\omega_{ij}y_{ij}}{\sum_{i\in S}\sum_{j\in S_i}\omega_{ij}}.$$

En este caso, con una muestra autoponderada, puede usar pesos relativos $w_{ij} = 1$.

Observe que aunque en condiciones ideales el método de Mitofsky-Waksberg conduce a una muestra autoponderada de números telefónicos residenciales, *no* proporciona una muestra autoponderada de familias, pues algunas pueden tener más de un número telefónico y otras pueden no tener teléfono. En la práctica, una persona que utilice el método de Mitofsky-Waksberg podría ajustar los pesos para compensar las líneas telefónicas múltiples y la ausencia de respuesta, como analizaremos en el capítulo 8. ■

EJEMPLO 6.10 *Muestreo PPP*

El muestreo con probabilidad proporcional a la predicción (PPP), descrito por Schreuder *et al* (1968) generalmente se recomienda como esquema de muestreo en silvicultura. Suponga que un investigador desea estimar el volumen total de madera en un área. Hay varias opciones disponibles: (1) estimar el volumen de cada árbol en el área. Sin embargo, podría haber miles de árboles y esto podría llevar mucho tiempo. (2) Utilizar un muestreo por conglomerados en donde se elijan terrenos con la misma área y se mida el volumen de cada árbol en el área seleccionada. (3) Utilizar un esquema de muestreo con probabilidades diferentes donde los puntos del área se elijan al azar y se incluyan en la muestra los árboles más cercanos a los puntos. En este diseño, un árbol se elige con probabilidad proporcional al área de la región que sea más cercana a ese árbol que a cualquier otro árbol. (4) Calcular aproximadamente el volumen de cada árbol y luego elegir los árboles con una probabilidad proporcional al volumen estimado. Al hacer esto de una sola vez, seleccionando los árboles

al estimar el volumen, se tiene un muestreo PPP: la predicción P representa al volumen predicho (estimado) utilizado para determinar los π_i.

Como una forma de muestreo con probabilidades diferentes en la que se desconoce la probabilidad de seleccionar una unidad primaria antes de obtener la muestra, el muestreo PPP es un caso particular de **muestreo de Poisson**. Los árboles más grandes tienden a producir más madera y contribuyen en gran medida a la variabilidad de la estimación del total de volumen. Así, es de esperar que el muestreo con probabilidades diferentes implique un menor esfuerzo de muestreo. En teoría, se puede estimar, aproximadamente el volumen de cada uno de los N árboles del bosque, y obtener un valor x_i para el árbol i. Entonces, podría revisar algunos árboles elegidos al azar con probabilidades proporcionales a x_i y medir cuidadosamente el volumen t_i. Sin embargo, ese procedimiento requiere dos viajes por el bosque y añade mucho trabajo al proceso de muestreo. En el muestreo PPP sólo se realiza un recorrido en el bosque y los árboles se seleccionan para estar en la muestra a la vez que se miden los x_i. El procedimiento es el siguiente:

1 Estime o conjeture el valor máximo probable de x_i para los árboles. Defina un valor L mayor que el valor máximo estimado de x_i.

2 Vaya hasta un árbol del bosque y determine x_i para ese árbol. Genere un número aleatorio u_i en $[0, L]$. Si $u_i \le x_i$, mida entonces el volumen y_i de ese árbol; en caso contrario, pase al siguiente.

3 Repita el paso 2 para cada árbol del bosque.

En este caso, el muestreo con probabilidades diferentes otorga a cada árbol maderable la misma posibilidad de selección para la muestra. Observe que el tamaño de la muestra con probabilidades diferentes se desconoce hasta concluir el muestreo. La probabilidad de incluir el árbol i en la muestra es $\pi_i = x_i/L$. El estimador de Horvitz-Thompson es

$$\hat{t}_{\mathrm{HT}} = \sum_{i\in S}\frac{y_i}{\pi_i} = L\sum_{i\in S}\frac{y_i}{x_i} = \sum_{i=1}^{N} Z_i\frac{y_i}{\pi_i},$$

donde $Z_i = 1$ si el árbol i está en la muestra y 0 en caso contrario. Entonces, el tamaño de la muestra es la variable aleatoria $\sum_{i=1}^{N} Z_i$ con valor esperado $\sum_{i=1}^{N} x_i/L$.

Como el tamaño de la muestra es variable y no fijo, el muestreo de Poisson proporciona otro método de muestreo con probabilidades diferentes además de los analizados en las secciones 6.1 a 6.4. Brewer y Hanif (1983) brinda más teoría y bibliografía acerca del muestreo de Poisson. ■

En el muestreo de recursos naturales, el muestreo PPP es un ejemplo del uso de probabilidades diferentes. Otros ejemplos aparecen en Overton y Stehman (1995).

EJEMPLO 6.11 *Muestreo de unidades monetarias*

Un contador que revisa las cuentas por cobrar de una empresa, frecuentemente extrae una muestra para estimar el balance real de esas cuentas. Se conoce el valor contable x_i de cada cuenta de la población; el valor auditado t_i se conoce sólo para las cuentas de la muestra. En la sección 3.2 vimos cómo se podía utilizar la información adicional x_i en la estimación de diferencias para mejorar la precisión de una muestra aleatoria simple de cuentas. De manera análoga, se podría usar la estimación por regresión o proporción.

Los valores contables se podrían usar en el diseño de la muestra. En lugar de emplearlas en el análisis. Sería posible estratificar las cuentas por el valor de x_i o bien, extraer una muestra con probabilidades diferentes, con probabilidades de selección proporcionales a x_i

(o hacer las dos cosas: primero estratificar y luego extraer la muestra con probabilidades diferentes dentro de cada estrato). Si extrae una muestra de las cuentas con probabilidades proporcionales a x_i, entonces cada unidad monetaria individual en los valores contables tiene la misma probabilidad de incluirse en la muestra (de ahí el nombre **muestreo de unidades monetarias**). Como cada unidad monetaria tiene la misma probabilidad de ser incluida en la muestra, una cuenta con valor contable, 10,000 dólares tiene 10 veces más posibilidades de estar en la muestra que una con valor contable de 1000 dólares.

Considere un cliente con 87 cuentas por cobrar, con un balance de 612,824 dólares. El auditor decidió que una muestra de tamaño 25 bastará para estimar el error en las cuentas por cobrar y extrae una muestra aleatoria con reemplazo de los $612,824 en la población de valores contables. Como los dólares individuales sólo se pueden auditar como parte de toda la cuenta, cada dólar seleccionado sirve como "gancho" para considerar toda la cuenta para auditarla. En este ejemplo se usa el método de tamaño acumulativo para seleccionar las unidades primarias (cuentas); frecuentemente, en la práctica los auditores extraen una muestra sistemática de dólares y sus unidades primarias acompañantes, que garantiza que se incluyen en la muestra las cuentas con valores contables mayores al intervalo de muestreo. La tabla 6.5 muestra las primeras líneas de la selección de cuentas; la tabla completa está en el archivo audit.dat. En este caso, las cuentas 3 y 13 se incluyen una vez, mientras que la cuenta 9 se incluye dos veces (aunque sólo hay que auditarla una vez, pues ésta es una muestra por conglomerados de una etapa). Así, éste es un ejemplo de muestreo con ppt de una etapa con reemplazo, según lo analizado en la sección 6.2.

Las cuentas seleccionadas se auditan y los valores contables se registran en la tabla 6.6. Usando los resultados de la sección 6.2, la declaración exagerada total se estima en 4334 dólares, con un error estándar de $13,547/\sqrt{25} = 2709$ dólares. Sin embargo, en muchas situaciones de auditoría, la mayoría de los valores auditados coinciden con los valores contables, de modo que las diferencias son nulas. Un intervalo de confianza basado en una aproximación normal no funciona adecuadamente en este caso, de modo que generalmente los auditores utilizan cuotas de confianza basadas en la distribución Poisson o multinomial (consulte Neter *et al* 1978) en vez de un intervalo de confianza de la forma (promedio ± 1.96 EE).

TABLA 6.5
Selección de cuentas para una muestra de auditoría

Cuenta (Unidad de auditoría)	Valor contable	Valor contable acumulativo	Número aleatorio	
1	2,459	2,459		
2	2,343	4,802		
3	6,842	11,644	11,016	
4	4,179	15,823		
5	750	16,573		
6	2,708	19,281		
7	3,073	22,354		
8	4,742	27,096		
9	16,350	43,446	31,056	38,500
10	5,424	48,870		
11	9,539	58,409		
12	3,108	61,517		
13	3,935	65,452	63,047	
14	900	66,352		

TABLA 6.6
Resultados de la auditoría a las cuentas de la muestra

Cuenta (Unidad de auditoría)	Valor contable (VC)	ψ_i	Valor auditado (VA)	Diferencia VC - VA	Diferencia / ψ_i	Diferencia por dólar
3	6,842	0.0111647	6,842	0	0	0.00000
9	16,350	0.0266798	16,350	0	0	0.00000
9	16,350	0.0266798	16,350	0	0	0.00000
13	3,935	0.0064211	3,935	0	0	0.00000
24	7,090	0.0115694	7,050	40	3,457	0.00564
29	5,533	0.0090287	5,533	0	0	0.00000
34	2,163	0.0035296	2,163	0	0	0.00000
36	2,399	0.0039147	2,149	250	63,862	0.10421
43	8,941	0.0145898	8,941	0	0	0.00000
44	3,716	0.0060637	3,716	0	0	0.00000
45	8,663	0.0141362	8,663	0	0	0.00000
46	69,540	0.1134747	69,000	540	4,759	0.00777
46	69,540	0.1134747	69,000	540	4,579	0.00777
46	69,540	0.1134747	69,000	540	4,579	0.00777
49	6,881	0.0112283	6,881	0	0	0.00000
55	70,100	0.1143885	70,100	0	0	0.00000
55	70,100	0.1143885	70,100	0	0	0.00000
55	70,100	0.1143885	70,100	0	0	0.00000
56	6,467	0.0105528	6,467	0	0	0.00000
61	21,000	0.0342676	21,000	0	0	0.00000
70	3,847	0.0062775	3,847	0	0	0.00000
74	2,422	0.0039522	2,422	0	0	0.00000
75	2,291	0.0037384	2,191	100	26,749	0.04365
79	4,667	0.0076156	4,667	0	0	0.00000
81	31,257	0.0510049	31,257	0	0	0.00000
		Promedio			4,334	0.007071874
		Desviación estándar			13,547	0.2210527

Otra forma de buscar la estimación con probabilidades diferentes es determinar la declaración exagerada en cada dólar individual de la muestra. Por ejemplo, la cuenta 24 tiene un valor contable de 7090 dólares y un error de 40 dólares. El error se prorratea por cada dólar del valor contable, lo que conduce a una declaración exagerada de 0.00564 dólares por cada uno de los 7090 dólares. La exageración promedio para los dólares individuales de la muestra es 0.007071874 dólares, de modo que la exageración total para la población se estima en (0.007071874)(612,824) = 4334. ■

6.6
Resultados y demostraciones de la teoría de aleatorización*

En el muestreo por conglomerados en dos etapas, siempre elegimos primero la unidad primaria y luego las subunidades dentro de la unidad primaria extraída. Un enfoque para calcular una varianza teórica para cualquier estimador en el muestreo de varias etapas es condicionar las unidades primarias por incluir en la muestra. Para hacerlo necesitamos usar las propiedades 4 (condicionamiento sucesivo) y 5 (cálculo condicional de las varianzas) de la esperanza condicional, establecidas en la sección B.4.

En esta sección establecemos y demostramos el teorema 6.2, de Horvitz-Thompson (Horvitz y Thompson, 1952), que presenta las propiedades del estimador de Horvitz-Thompson en (6.12). En el teorema 6.3 determinamos los estimadores insesgados de la varianza. Luego, mostramos que la varianza para el muestreo por conglomerados con probabilidades iguales en (5.22) se obtiene como caso particular de estos teoremas. Sin embargo, primero demostramos (6.10) y (6.11).

TEOREMA 6.1

Para una muestra de probabilidad de n unidades sin reemplazo, sea

$$Z_i = \begin{cases} 1 & \text{si la unidad primaria } i \text{ está en la muestra} \\ 0 & \text{si la unidad primaria } i \text{ no está en la muestra} \end{cases}$$

y definamos

$$P(Z_i = 1) = \pi_i$$

y

$$P(Z_i = 1 \text{ y } Z_k = 1) = \pi_{ik}.$$

Entonces

$$\sum_{i=1}^{N} \pi_i = n$$

y

$$\sum_{\substack{k=1 \\ k \neq i}}^{N} \pi_{ik} = (n-1)\pi_i.$$

Demostración Como el tamaño de la muestra es n, $\sum_{i=1}^{N} Z_i = n$. Además,

$$E[Z_i] = E[Z_i^2] = \pi_i$$

pues $P(Z_i = 1) = \pi_i$. En consecuencia,

$$n = E\left[\sum_{i=1}^{N} Z_i\right] = \sum_{i=1}^{N} \pi_i.$$

Además,

$$\sum_{\substack{k=1 \\ k \neq i}}^{N} \pi_{ik} = \sum_{\substack{k=1 \\ k \neq i}}^{N} E[Z_i Z_k] = E[Z_i(n - Z_i)] = \pi_i(n-1),$$

lo cual concluye la demostración.

TEOREMA 6.2 *Horvitz-Thompson*

Sean Z_i, π_i y π_{ik} como en el teorema 1. Suponga que el muestreo de la segunda etapa se realiza de modo que el muestreo en cualquier unidad primaria sea independiente del muestreo en cualquier otra unidad primaria, y que $\hat{t}_i$ es independiente de $(Z_1, \ldots, Z_N)$, con $E[\hat{t}_i] = E[\hat{t}_i \mid Z_1, \ldots, Z_N] = t_i$. Entonces

$$E\left[\sum_{i=1}^{N} Z_i \frac{\hat{t}_i}{\pi_i}\right] = \sum_{i=1}^{N} \pi_i \frac{t_i}{\pi_i} = t \qquad \textbf{(6.19)}$$

y

$$V\left[\sum_{i=1}^{N} Z_i \frac{\hat{t}_i}{\pi_i}\right] = V_{\text{up}} + V_{\text{us}} \tag{6.20}$$

donde

$$V_{\text{up}} = V\left[\sum_{i=1}^{N} Z_i \frac{\hat{t}_i}{\pi_i}\right] = \sum_{i=1}^{N}(1-\pi_i)\frac{t_i^2}{\pi_i} + \sum_{i=1}^{N}\sum_{\substack{k=1 \\ k\neq i}}^{N}(\pi_{ik}-\pi_i\pi_k)\frac{t_i}{\pi_i}\frac{t_k}{\pi_k} \tag{6.21}$$

y

$$V_{\text{ssu}} = \sum_{i=1}^{N}\frac{V(\hat{t}_i)}{\pi_i}. \tag{6.22}$$

Demostración Observe primero que

$$\text{Cov}(Z_i, Z_k) = \begin{cases} \pi_i(1-\pi_i) & \text{si } i = k. \\ \pi_{ik} - \pi_i\pi_k & \text{si } i \neq k. \end{cases}$$

Utilizamos el condicionamiento sucesivo para mostrar (6.19):

$$\begin{aligned} E\left[\sum_{i=1}^{N} Z_i \frac{\hat{t}_i}{\pi_i}\right] &= E\left\{E\left[\sum_{i=1}^{N} Z_i \frac{\hat{t}_i}{\pi_i}\middle| Z_1, \ldots, Z_N\right]\right\} \\ &= E\left[\sum_{i=1}^{N} Z_i \frac{t_i}{\pi_i}\right] \\ &= \sum_{i=1}^{N}\pi_i\frac{t_i}{\pi_i} \\ &= t. \end{aligned}$$

El primer paso aplica simplemente el condicionamiento sucesivo; en el segundo paso utilizamos la independencia de $\hat{t}_i$ y $(Z_1, \ldots, Z_N)$.

Para determinar la varianza, empleamos la expresión para su cálculo condicional, en la propiedad 5 de la sección B.4, y de nuevo utilizamos la independencia de $\hat{t}_i$ y $(Z_1, \ldots, Z_N)$:

$$\begin{aligned} V\left[\sum_{i=1}^{N} Z_i \frac{\hat{t}_i}{\pi_i}\right] &= V\left[E\left(\sum_{i=1}^{N} Z_i \frac{\hat{t}_i}{\pi_i}\middle| Z_1,\ldots,Z_N\right)\right] \\ &\quad + E\left[V\left(\sum_{i=1}^{N} Z_i \frac{\hat{t}_i}{\pi_i}\middle| Z_1,\ldots,Z_N\right)\right] \\ &= V\left[\sum_{i=1}^{N} Z_i \frac{t_i}{\pi_i}\right] + E\left[\sum_{i=1}^{N} Z_i^2 \frac{V(\hat{t}_i)}{\pi_i^2}\right] \\ &= \sum_{i=1}^{N}\sum_{k=1}^{N}\frac{t_i}{\pi_i}\frac{t_k}{\pi_k}\text{Cov}(Z_i, Z_k) + \sum_{i=1}^{N}\pi_i\frac{V(\hat{t}_i)}{\pi_i^2} \\ &= \sum_{i=1}^{N}\pi_i(1-\pi_i)\frac{t_i^2}{\pi_i^2} + \sum_{i=1}^{N}\sum_{\substack{k=1 \\ k\neq i}}^{N}(\pi_{ik}-\pi_i\pi_k)\frac{t_i}{\pi_i}\frac{t_k}{\pi_k} + \sum_{i=1}^{N}\frac{V(\hat{t}_i)}{\pi_i}. \quad \blacksquare \end{aligned}$$

La ecuación (6.19) establece que el estimador de Horvitz-Thompson es insesgado y las ecuaciones (6.20) a (6.22) muestran que (6.13) es la varianza del estimador de Horvitz-Thompson.

El teorema 6.3 proporciona un estimador insesgado para la varianza en (6.13).

TEOREMA 6.3

Suponga que las condiciones del teorema 6.2 y que $\hat{V}(\hat{t}_i)$ es un estimador insesgado de $V(\hat{t}_i)$. Entonces,

$$E\left[Z_i \frac{\hat{V}(\hat{t}_i)}{\pi_i^2}\right] = \frac{V(\hat{t}_i)}{\pi_i}, \tag{6.23}$$

$$E\left[\sum_{i=1}^{N} Z_i \frac{\hat{V}(\hat{t}_i)}{\pi_i^2}\right] = V_{\text{us}}, \tag{6.24}$$

y

$$\begin{aligned} &E\left[\sum_{i=1}^{N} Z_i(1-\pi_i)\frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i=1}^{N}\sum_{\substack{k=1 \\ k\neq i}}^{N} Z_i Z_k \frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\frac{\hat{t}_i}{\pi_i}\frac{\hat{t}_k}{\pi_k}\right] \\ &= E\left[\sum_{i=1}^{N}\sum_{k=i+1}^{N} Z_i Z_k \frac{\pi_i\pi_k - \pi_{ik}}{\pi_{ik}}\left(\frac{\hat{t}_i}{\pi_i} - \frac{\hat{t}_k}{\pi_k}\right)^2\right] \\ &= V_{\text{up}} + \sum_{i=1}^{N}(1-\pi_i)\frac{V(\hat{t}_i)}{\pi_i}. \end{aligned} \tag{6.25}$$

Demostración Demostramos (6.23) y (6.24) utilizando de nuevo el condicionamiento sucesivo:

$$E\left[Z_i \frac{\hat{V}(\hat{t}_i)}{\pi_i^2}\right] = E\left[E\left(Z_i \frac{\hat{V}(\hat{t}_i)}{\pi_i^2}\middle| Z_1,\ldots,Z_N\right)\right] = E\left[Z_i \frac{V(\hat{t}_i)}{\pi_i^2}\right] = \frac{V(\hat{t}_i)}{\pi_i^2}.$$

(6.24) es una consecuencia inmediata.

Para demostrar (6.25), observamos que como $\hat{t}_i$ y $(Z_1, \ldots, Z_N)$ son independientes,

$$E[\hat{t}_i^2 \mid Z_1,\ldots,Z_N] = E[\hat{t}_i^2] = t_i^2 + V(\hat{t}_i).$$

Así,

$$\begin{aligned} E\left[\sum_{i=1}^{N} Z_i(1-\pi_i)\frac{\hat{t}_i^2}{\pi_i^2}\right] &= E\left[E\left(\sum_{i=1}^{N} Z_i(1-\pi_i)\frac{\hat{t}_i^2}{\pi_i^2}\middle| Z_1,\ldots,Z_N\right)\right] \\ &= E\left[\sum_{i=1}^{N} Z_i \frac{1-\pi_i}{\pi_i^2}\{t_i^2 + V(\hat{t}_i)\}\right] \\ &= \sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}[t_i^2 + V(\hat{t}_i)]. \end{aligned}$$

Como el submuestreo se realiza de manera independiente en distintos conglomerados, $E[\hat{t}_i\hat{t}_k] = t_i t_k$ para $k \neq i$, de modo que

$$E\left[\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} Z_i Z_k \frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\frac{\hat{t}_i}{\pi_i}\frac{\hat{t}_k}{\pi_k}\right]$$

$$= E\left[E\left(\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} Z_i Z_k \frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\frac{\hat{t}_i}{\pi_i}\frac{\hat{t}_k}{\pi_k}\middle| Z_1,\ldots,Z_N\right)\right]$$

$$= E\left[\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} Z_i Z_k \frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\frac{t_i}{\pi_i}\frac{t_k}{\pi_k}\right]$$

$$= \sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N}(\pi_{ik}-\pi_i\pi_k)\frac{t_i}{\pi_i}\frac{t_k}{\pi_k}.$$

Al combinar los dos resultados, vemos que

$$E\left[\sum_{i=1}^{N} Z_i(1-\pi_i)\frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} Z_i Z_k \frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\frac{\hat{t}_i}{\pi_i}\frac{\hat{t}_k}{\pi_k}\right]$$

$$= V_{\text{up}} + \sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}V(\hat{t}_i),$$

lo cual demuestra la primera parte de (6.25). Análogamente demostramos la segunda parte de (6.25):

$$E\left[\sum_{i=1}^{N}\sum_{k=i+1}^{N} Z_i Z_k \frac{\pi_i\pi_k-\pi_{ik}}{\pi_{ik}}\left(\frac{\hat{t}_i}{\pi_i}-\frac{\hat{t}_k}{\pi_k}\right)^2\right]$$

$$= E\left\{E\left[\sum_{i=1}^{N}\sum_{k=i+1}^{N} Z_i Z_k \frac{\pi_i\pi_k-\pi_{ik}}{\pi_{ik}}\left(\frac{\hat{t}_i}{\pi_i}-\frac{\hat{t}_k}{\pi_k}\right)^2\middle| Z_1,\ \ldots,\ Z_N\right]\right\}$$

$$= E\left[\sum_{i=1}^{N}\sum_{k=i+1}^{N} Z_i Z_k \frac{\pi_i\pi_k-\pi_{ik}}{\pi_{ik}}\left(\frac{t_i^2+V(\hat{t}_i)}{\pi_i^2}-2\frac{t_i}{\pi_i}\frac{t_k}{\pi_k}+\frac{t_k^2+V(\hat{t}_k)}{\pi_k^2}\right)\right]$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N}(\pi_i\pi_k-\pi_{ik})\left(\frac{t_i^2+V(\hat{t}_i)}{\pi_i^2}-2\frac{t_i}{\pi_i}\frac{t_k}{\pi_k}+\frac{t_k^2+V(\hat{t}_k)}{\pi_k^2}\right)$$

$$= \sum_{i=1}^{N}[\pi_i(n-\pi_i)-(n-1)\pi_i]\left(\frac{t_i^2+V(\hat{t}_i)}{\pi_i^2}\right)+\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N}(\pi_{ik}-\pi_i\pi_k)\frac{t_i}{\pi_i}\frac{t_k}{\pi_k}.$$

El último paso es consecuencia del teorema 6.1 y es fácil ver que la última expresión es equivalente a $V_{\text{up}} + \sum_{i=1}^{N}(1-\pi_i)V(\hat{t}_i)/\pi_i$. Esto concluye la demostración del teorema 6.3. ■

El teorema 6.3 implica que (6.14) y (6.15) son estimadores insesgados de la varianza del estimador de Horvitz-Thompson.

Si se eligen las unidades primarias con probabilidades iguales, como en el capítulo 5, entonces

$$P(Z_i = 1) = \pi_i = \frac{n}{N},$$

$$P(Z_i = 1 \text{ y } Z_j = 1) = \pi_{ij} = \frac{n}{N}\frac{n-1}{N-1},$$

$$\hat{t}_{\text{ins}} = \sum_{i\in S}\frac{N}{n}\hat{t}_i = \sum_{i=1}^{N} Z_i \frac{N}{n}\hat{t}_i,$$

de modo que podemos aplicar el teorema 6.2 con $\pi_i = n/N$. Entonces,

$$E[\hat{t}_{\text{ins}}] = \sum_{i=1}^{N}\frac{n}{N}\frac{N}{n}t_i = t,$$

y de (6.21),

$$V_{\text{up}}[\hat{t}_{\text{ins}}] = \sum_{i=1}^{N}\left(1-\frac{n}{N}\right)\left(\frac{N}{n}\right)t_i^2 + \sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N}\left[\frac{n}{N}\frac{n-1}{N-1}-\left(\frac{n}{N}\right)^2\right]\left(\frac{N}{n}\right)^2 t_i t_k$$

$$= \frac{N}{n}\left(1-\frac{n}{N}\right)\left[\sum_{i=1}^{N} t_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} t_i t_k\right]$$

$$= \frac{N}{n(N-1)}\left(1-\frac{n}{N}\right)\left[(N-1)\sum_{i=1}^{N} t_i^2 - \sum_{i=1}^{N}\sum_{k=1}^{N} t_i t_k + \sum_{i=1}^{N} t_i^2\right]$$

$$= \frac{N}{n(N-1)}\left(1-\frac{n}{N}\right)\left[N\sum_{i=1}^{N} t_i^2 - t^2\right]$$

$$= N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n}.$$

Por el resultado (2.7) de la teoría de muestras aleatorias simples,

$$V(\hat{t}_i) = M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{S_i^2}{m_i},$$

de modo que, usando (6.22),

$$V_{\text{us}} = \sum_{i=1}^{N}\frac{N}{n}M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{S_i^2}{m_i}.$$

Esto concluye la demostración de (5.22).

Después de un poco de álgebra, se puede mostrar que al extraer una muestra por conglomerados con probabilidades iguales,

$$\sum_{i=1}^{N} Z_i(1-\pi_i)\left(\frac{N}{n}\right)^2 \hat{t}_i^2 + \sum_{i=1}^{N}\sum_{\substack{k=1\\k\neq i}}^{N} Z_i Z_k \frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\left(\frac{N}{n}\right)^2 \hat{t}_i\hat{t}_k = N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}.$$

Así, por el teorema 6.3,

$$E\left[N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}\right] = N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\left(1-\frac{n}{N}\right)\sum_{i=1}^{N} V(\hat{t}_i); \qquad \textbf{(6.26)}$$

y en consecuencia,

$$E[s_t^2] = S_t^2 + \frac{1}{N}\sum_{i=1}^{N} V(\hat{t}_i).$$

Observe que el valor esperado de s_t^2 es mayor que S_t^2: Incluye la variación del total de las unidades primarias al total de las unidades secundarias, más la variación por el hecho de desconocer el total de las unidades primarias.

Como

$$\hat{V}(\hat{t}_i) = \left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}$$

es un estimador insesgado de $V(\hat{t}_i)$, el teorema 6.3 implica que

$$E\left[\sum_{i=1}^{N} Z_i\left(\frac{N}{n}\right)^2 \hat{V}(\hat{t}_i)\right] = E\left[\sum_{i\in S}\left(\frac{N}{n}\right)^2 \hat{V}(\hat{t}_i)\right] = V_{\text{us}}.$$

Entonces, usando (6.26),

$$\begin{aligned} &E\left[N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\sum_{i\in S}\hat{V}(\hat{t}_i)\right] \\ &\quad = N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\left(1-\frac{n}{N}\right)\sum_{i=1}^{N} V(\hat{t}_i) + \sum_{i=1}^{N} V(\hat{t}_i) \\ &\quad = N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\sum_{i=1}^{N} V(\hat{t}_i) \end{aligned}$$

Así, (5.25) es un estimador insesgado de (5.22).

Los métodos utilizados en estas demostraciones se pueden aplicar a cualquier cantidad de niveles de conglomerados. Tal vez desee extraer una muestra de escuelas, luego de grupos dentro de las escuelas y posteriormente de los estudiantes dentro de los grupos. En el ejercicio 28 pedimos determinar una expresión para la varianza en el muestreo por conglomerados de tres etapas. Rao (1979a) presenta un enfoque alternativo y elegante, basado en propiedades de las matrices definidas no negativas para deducir los errores cuadráticos medios y los estimadores de varianza para los estimadores lineales de los totales de una población.

6.7 Modelos y muestreo con probabilidades diferentes*

En general, los datos de un buen diseño de muestreo deben producir inferencias razonables a partir de un enfoque basado en el modelo o un enfoque de aleatorización. Veamos cómo funciona el estimador con ppt para el modelo M1 de (5.37). El modelo es

$$\text{M1:}\quad Y_{ij} = A_i + \varepsilon_{ij},$$

con los A_i generados con una distribución con media μ y varianza σ_A^2, los ε_{ij} se generan mediante una distribución con media 0 y varianza 1, y todos los A_i y los ε_{ij} son independientes.

Como hicimos con los estimadores del capítulo 5, podemos escribir el estimador con ppt como una combinación lineal de las variables aleatorias Y_{ij}. Para un diseño con ppt, $\psi_i = M_i/K$, de modo que

$$\hat{T}_p = \sum_{i\in S}\frac{K}{nM_i}\hat{T}_i = \sum_{i\in S} K\frac{\overline{Y}_{SI}}{n} = \sum_{i\in S}\sum_{j\in S_i}\frac{K}{nm_i}Y_{ij}.$$

Observe que $\sum_{i\in S}\sum_{j\in S_i} K/(nm_i) = K$, de modo que $\hat{T}_P$ es insesgado bajo el modelo M1 en (5.37). Además, de (5.39),

$$\begin{aligned} V_{\text{M1}}[\hat{T}_P - T] &= \sigma_A^2\left[\sum_{i\in S}\left(\sum_{j\in S_i}\frac{K}{nm_i} - M_i\right)^2 + \sum_{i\notin S} M_i^2\right] \\ &\quad + \sigma^2\left[\sum_{i\in S}\sum_{j\in S_i}\left\{\left(\frac{K}{nm_i}\right)^2 - 2\frac{K}{nm_i}\right\} + K\right] \\ &= \sigma_A^2\left[\frac{K^2}{n} - 2\frac{K}{n}\sum_{i\in S} M_i + \sum_{i=1}^{N} M_i^2\right] + \sigma^2\left[\sum_{i\in S}\frac{K^2}{n^2 m_i} - K\right]. \end{aligned}$$

La varianza basada en el modelo $\hat{T}_E$ tiene implicaciones para el diseño. Suponga que se desea una muestra que minimice $V_{\text{M1}}[\hat{T}_P - T]$. Los tamaños de las unidades primarias M_i de las unidades de la muestra sólo aparecen en el término $-2\sigma_A^2(K/n)\sum_{i\in S} M_i$, de modo que para n fija, la varianza es mínima cuando las n unidades con M_i mayor se incluyen en la muestra. Si además se establece una restricción sobre la cantidad de subunidades que pueden examinarse, $\sum_{i\in S}(1/m_i)$ es mínimo cuando todas las m_i son iguales.

La inferencia en el enfoque basado en el modelo no depende del diseño de muestreo. Mientras el modelo M1 sea válido para la población, $\hat{T}_P$ es insesgado respecto del modelo, con la varianza dada anteriormente. En un enfoque basado en el modelo, el investigador, con confianza, podrá elegir simplemente las unidades primarias con valores mayores de M_i para incluirlas en la muestra. Sin embargo, en la práctica esto no es posible: nadie tiene confianza plena en el modelo, particularmente antes de recoger los datos. Royall y Eberhardt (1975) sugieren el uso del muestreo equilibrado, en el que la muestra se selecciona de tal modo que las inferencias sean robustas con respecto de ciertas formas de especificación incorrecta del modelo.

Como se describió en la sección 6.2, el muestreo con ppt se puede pensar como una manera de introducir el carácter aleatorio dentro del diseño óptimo para el modelo M1 y el estimador $\hat{T}_P$. El diseño autoponderado donde todos los m_i son iguales minimiza, además, la varianza en el enfoque basado en el modelo. Así, si se supone que el modelo M1 describe

los datos, el muestreo y la estimación con ppt deben tener un buen desempeño en la práctica. Särndal (1978) y Thompson (1997) analizan las diferencias entre la inferencia basada en el diseño y la basada en el modelo en las muestras para encuestas.

Concluimos nuestro análisis con un ejemplo ampliamente citado de Basu, que se utiliza con frecuencia para demostrar que las estimaciones de Horvitz-Thompson pueden ser tan "tontas" como cualquier otro procedimiento estadístico aplicado en forma inadecuada.

> El dueño de un circo quiere embarcar sus 50 elefantes adultos, para lo que necesita un cálculo aproximado del peso total de los animales. Como es difícil pesar un elefante, el dueño quiere estimar el peso total pesando sólo un elefante. ¿Cuál debe pesar? El dueño busca en sus registros y encuentra una lista de los pesos de los animales, elaborada hace tres años. En ella ve que "Sambo", un elefante de peso medio, era el promedio (en peso) de su manada. Verifica con el entrenador, quien le asegura que "Sambo" puede seguir considerándose como el elefante promedio del grupo. Por lo tanto, el dueño planea pesar a "Sambo" y considerar a $50y$ (donde y es el peso actual de "Sambo") como una estimación del peso total $Y = Y_1 + Y_2 + \cdots + Y_{50}$ de los 50 elefantes. Pero el estadístico del circo se horroriza al conocer el plan de muestreo del dueño. "¿Cómo puede obtener una estimación insesgada de Y de esta forma?", protesta el estadístico. Así, juntos trabajan para elaborar un plan de muestreo. Con la ayuda de una tabla de números aleatorios, diseñan un plan que asigna una probabilidad de selección de 99/100 a "Sambo" y probabilidades de selección iguales de 1/4900 a cada uno de los otros 49 elefantes. Naturalmente, "Sambo" se selecciona y el dueño queda feliz. "¿Cómo va a estimar Y?", pregunta el estadístico. "¿Por qué? La estimación será $50y$, por supuesto", responde el dueño. "¡Oh, no! Eso no puede ser correcto", replica el estadístico, "leí recientemente un artículo en *Annals of Mathematical Statistics* donde se demuestra que el estimador de Horvitz-Thompson es el único estimador hiperadmisible en la clase de todos los estimadores insesgados polinomiales generalizados". "¿Cuál es el estimador de Horvitz-Thompson en este caso?", pregunta el dueño, impresionado. "Como la probabilidad de selección de Sambo en nuestro plan fue de 99/100, la estimación adecuada de Y es $100y/99$ y no $50y$", contesta el estadístico. El dueño, incrédulo, pregunta "¿y cuál sería la estimación de Y si nuestro plan de muestreo hubiese elegido, digamos, al gran elefante 'Jumbo'?" "De acuerdo con lo que entiendo del método de estimación de Horvitz-Thompson, contesta tristemente el estadístico "la estimación adecuada de Y sería entonces $4900y$, donde y es el peso de 'Jumbo'". De esta forma, el estadístico perdió su empleo en el circo (¡y tal vez se convirtió en profesor de estadística!). (1971, 212-213).

¿Fue correcto despedir al estadístico? Un profesional que desee utilizar un modelo para analizar los datos de una encuesta debería responder afirmativamente: El empleado del circo utilizaba el modelo $y_i \propto 99/100$ para "Sambo" y $y_i \propto 1/4900$ para los demás elefantes de la manada, lo que en realidad no es un modelo que se ajuste bien a los datos. Un estadístico que utiliza las inferencias con aleatorización también respondería que sí: Aunque los modelos no se utilizan explícitamente en la teoría de Horvitz-Thompson, el estimador es más eficiente (tiene varianza mínima) cuando el total de las unidades primarias (en este caso, y_i) es proporcional a la probabilidad de selección. El ingenuo diseño utilizado por el estadístico del circo conduce a una enorme varianza del estimador de Horvitz-Thompson. Si esas no fuesen razones suficientes, el estadístico propone una muestra de tamaño 1; ¡no puede verificar la validez del modelo en un enfoque basado en el modelo, ni estimar la varianza del estimador de Horvitz-Thompson!

6.8 Ejercicios

1 Para cada una de las siguientes situaciones, ¿qué unidad podría usarse como unidad primaria de muestreo? ¿Cree que habría un fuerte efecto de acumulación? ¿Extraería la muestra de unidades primarias con probabilidades iguales o diferentes?

a Usted quiere estimar el porcentaje de personas que utilizan lentes de contacto entre todos los pacientes de optometristas y oftalmólogos de la Fuerza Aérea de Estados Unidos.

b La teniasis se adquiere ingiriendo larvas de tenia en carne de cerdo preparada en forma inadecuada. Es necesario diseñar una encuesta para estimar el porcentaje de habitantes con teniasis en un poblado. Se necesita un examen médico para diagnosticar la enfermedad.

c Usted quiere estimar la cantidad total de vacas y novillos en todas las granjas lecheras de Ontario; además, quiere estimar la tasa de natalidad y la tasa de partos de fetos muertos.

d Usted quiere estimar los porcentajes de estudiantes de licenciatura de las universidades de Estados Unidos registrados como votantes y de afiliados a algún partido político.

e Una empresa pesquera está interesada en la distribución del ancho del caparazón de ciertos cangrejos. La red de un barco pesquero puede pescar un límite de 30 cangrejos.

f Usted desea realizar una encuesta de satisfacción del cliente con las personas que han participado en recorridos guiados en autobús por el área del borde del Gran Cañón. Los grupos de recorrido varían en tamaño, de ocho a 44 personas.

2 Los historiadores que desean utilizar los datos de los censos de Estados Unidos realizados antes de la era de las computadoras enfrentaron la nada agradable tarea de examinar carretes de registros manuscritos microfilmados ordenados geográficamente. Las muestras de microfichero de uso público (PUMS, por sus siglas en inglés) se hicieron tomando muestras de los registros y transcribiéndolos en computadora. Ruggles describe la construcción de los PUMS del censo de 1940:

> Los datos de población del censo de 1940 se preservan en 4,576 carretes de microfilms. Cada página del censo contiene información de 40 individuos. Se tomaron los renglones de cada página como "renglones muestra" por la Oficina de Censos: los individuos que caían en esos renglones (5% de la población) contestaron un conjunto de preguntas adicionales que aparecen en la parte inferior de la página de censo.
> Dos de cada cinco páginas de censo se seleccionaron de manera sistemática para su análisis; en cada uno de ellos, se eligió al azar uno de los dos renglones de muestra ya designados, los capturistas contaron entonces el tamaño de la unidad de muestra que contenía el renglón señalado. Las unidades de tamaño 6 o menor se incluyeron en la muestra, en proporción inversa a su tamaño. Así, se incluyó en la muestra cada unidad de una persona, cada segunda unidad de dos personas, cada tercera unidad de tres personas y así sucesivamente. Las unidades con siete o más personas se incluyeron con una probabilidad de 1 en 7: cada séptima familia de tamaño 7 o mayor se seleccionó para la muestra. (1995, 44)

a Explique por qué es una muestra por conglomerados. ¿Cuáles son las unidades primarias? ¿Cuáles las secundarias?

b ¿Qué efecto supone que tenga la acumulación en las estimaciones de raza? ¿De edad? ¿De empleo?

c Construya una tabla para la probabilidad de selección de personas en las unidades de una persona, unidades de dos personas y así sucesivamente.

d ¿Qué ocurre si se estima la edad promedio de la población mediante la edad promedio de las personas de la muestra? ¿Qué estimador debe emplear?

e ¿Cree que la extracción de una muestra sistemática fue una buena idea para esta muestra? ¿Por qué?

f ¿Proporciona este método una muestra representativa de las familias? ¿Por qué?

g ¿Qué tipo de muestra se obtiene de los individuos con la información adicional? Explique.

3 Ruggles también describe los PUMS de 1950:

> Los datos del censo de 1950 están contenidos en 6,278 carretes de microfilm. Cada página de censo contiene información de 30 personas. Cada quinto renglón de la página de censo se tomó como renglón muestra y las preguntas adicionales para los individuos de ese renglón aparecen en la parte inferior de la forma. Para el censado del último renglón muestra de cada página, había un bloque de preguntas adicionales. Así, 20% de las personas contestó un conjunto básico de preguntas adicionales y 3.33% respondió un conjunto completo de preguntas adicionales.
>
> Se tomó al azar una de cada 11 páginas dentro de los distritos enumerados. En cada una el individuo del sexto renglón muestra (el correspondiente al conjunto completo de preguntas) se incluyó en la muestra. Los demás miembros de la unidad de muestra que contenía al individuo seleccionado también fueron incluidos. (1995, 45)

Para los PUMS de 1950, responda las mismas preguntas del ejercicio 2.

4 Un investigador quiere extraer una muestra con probabilidades diferentes de 10 de las 25 unidades primarias de la población enumerada a continuación y desea extraer las unidades con reemplazo.

Unidad primaria	ψ_i	Unidad primaria	ψ_i
1	0.000110	14	0.014804
2	0.018556	15	0.005577
3	0.062999	16	0.070784
4	0.078216	17	0.069635
5	0.075245	18	0.034650
6	0.073983	19	0.069492
7	0.076580	20	0.036590
8	0.038981	21	0.033853
9	0.040772	22	0.016959
10	0.022876	23	0.009066
11	0.003721	24	0.021795
12	0.024917	25	0.059185
13	0.040654		

a Adapte el método de tamaño acumulativo para extraer una muestra de tamaño 10 con reemplazo, con probabilidades ψ_i.

b Adapte el método de Lahiri para extraer una muestra de tamaño 10 con reemplazo, con probabilidades ψ_i.

5 Para el ejemplo de los supermercados de la sección 6.1, suponga que los ψ_i están dados, pero que cada tienda tiene $t_i = 75$. ¿Cuál es el valor de $E[\hat{t}_\psi]$? ¿$V[\hat{t}_\psi]$?

6 Para el ejemplo de los supermercados de la sección 6.1, suponga que los ψ_i son 7/16 para la tienda A y 3/16 para cada una de las tiendas B, C y D. Muestre que $\hat{t}_\psi$ es insesgado y determine su varianza. ¿Cree que el esquema de muestreo con estos ψ_i es bueno?

7 Regrese al ejemplo de los supermercados de la sección 6.1. Ahora elija dos supermercados, con reemplazo. Enumere las 16 muestras posibles (A,A), (A,B), etcétera, y determine las probabilidades de seleccionar cada muestra. Calcule $\hat{t}_\psi$ para cada muestra. ¿Cuál es el valor de $E[\hat{t}_\psi]$? ¿$V[\hat{t}_\psi]$?

8 El archivo statpps.dat enumera la cantidad de condados, área y población en 1992 para los 50 estados de Estados Unidos, más el Distrito de Columbia.

a Utilice el método de tamaño acumulativo para extraer una muestra de tamaño 10 con reemplazo, con probabilidades proporcionales al área. ¿Cuál es el valor de ψ_i para cada estado de la muestra?

b Utilice el método de tamaño acumulativo para extraer una muestra de tamaño 10 con reemplazo, con probabilidades proporcionales a la población. ¿Cuál es el valor de ψ_i para cada estado de la muestra?

c ¿En qué difieren las dos muestras? ¿Qué estados tienden a estar en cada muestra?

9 Para este problema, utilice la muestra de estados extraídos con probabilidad proporcional a la población del ejercicio 8.

a Use la muestra para estimar la cantidad total de condados de Estados Unidos y determine el error estándar de su estimación. ¿Cuál es la relación de su estimación con el valor verdadero de la cantidad total de condados (que se puede calcular, pues el archivo statepps.dat contiene los datos de toda la población)?

b Ahora suponga que su amigo Tom encuentra los 10 valores de los números de los condados en su muestra, pero no sabe que eligió estos estados con probabilidades proporcionales a la población. Tom estima entonces el área total usando las fórmulas para una muestra aleatoria simple. ¿Cuáles valores para el total estimado y su error estándar obtiene Tom? ¿Cuál es la diferencia entre estos valores y los suyos? ¿Es insesgada la estimación de Tom para el total de la población?

10 En el ejemplo 2.4 extrajimos una muestra aleatoria simple para estimar el total de acres dedicados a la agricultura en Estados Unidos en 1992. En el ejemplo 3.2 utilizamos la estimación por proporciones, con una variable auxiliar dada por la cantidad de acres de granjas en 1987, para aumentar la precisión de la estimación. Ahora, use la muestra de estados extraída con probabilidad proporcional al área del ejercicio 8, y obtenga una submuestra aleatoria de cinco condados de cada estado, mediante el archivo agpop.dat. Estime el total de acres dedicados a la agricultura en 1992, junto con su error estándar.

11 El archivo statepop.dat, utilizado en el ejemplo 6.5, también contiene información acerca de la cantidad total de granjas, la cantidad de veteranos y otros elementos.

a Grafique la cantidad total de granjas contra las probabilidades de selección ψ_i. ¿Indica esta gráfica que el muestreo con probabilidades diferentes es útil en ese caso?

b Estime la cantidad total de granjas en Estados Unidos, junto con su error estándar.

12 Para este problema, utilice el archivo statepop.dat.

a Grafique la cantidad total de veteranos contra las probabilidades de selección ψ_i. ¿Indica su gráfica que el muestreo con probabilidades diferentes será útil en este caso?

b Estime la cantidad total de veteranos en Estados Unidos y determine el error estándar de su estimación.

c Estime la cantidad total de veteranos de Vietnam en Estados Unidos y determine el error estándar de su estimación.

13 Regresemos a la situación del ejercicio 8 del capítulo 2, donde obtuvimos una muestra aleatoria simple para estimar la cantidad promedio y total de publicaciones con arbitraje de los profesores e investigadores. Ahora, consideremos una muestra con ppt de los profesores: el tamaño de las 27 unidades académicas varía de 2 a 92. Utilizamos el método de Lahiri para elegir 10 unidades primarias con probabilidades proporcionales al tamaño, con reemplazo, y obtenemos una muestra aleatoria simple de cuatro (o menos, si $M_i < 4$) miembros de cada unidad primaria. Observe que la unidad académica 14 aparece tres veces en la muestra; cada vez que aparece, obtenemos una submuestra distinta.

Unidad académica	M_i	ψ_i	y_{ij}
14	65	0.0805452	3, 0, 0, 4
23	25	0.0309789	2, 1, 2, 0
9	48	0.0594796	0, 0, 1, 0
14	65	0.0805452	2, 0, 1, 0
16	2	0.0024783	2, 0,
6	62	0.0768278	0, 2, 2, 5
14	65	0.0805452	1, 0, 0, 3
19	62	0.0768278	4, 1, 0, 0
21	61	0.0755886	2, 2, 3, 1
11	41	0.0508055	2, 5, 12, 3

Determine la cantidad total estimada de publicaciones, junto con su error estándar.

***14** (Requiere probabilidad.)

a Demuestre que el método de Lahiri produce una muestra con ppt con reemplazo.

b Suponga que la población tiene N unidades primarias, con tamaños $M_1, M_2, \ldots, M_N$. Sea X la cantidad de parejas de números aleatorios que pueden generarse para obtener una muestra de tamaño n. Determine $E[X]$.

***15** (Requiere probabilidad.) En la sección 6.3, observe que las variables aleatorias $Q_1, \ldots, Q_N$ tienen una distribución multinomial conjunta con probabilidades $\psi_1, \psi_2, \ldots, \psi_N$. Utilice las propiedades de la distribución multinomial para mostrar que $\hat{t}_\psi$ en (6.8) es un estimador insesgado de t con varianza dada por

$$V(\hat{t}_\psi) = \frac{1}{n}\sum_{i=1}^{N}\psi_i\left(\frac{t_i}{\psi_i} - t\right)^2 + \frac{1}{n}\sum_{i=1}^{N}\frac{V_i}{\psi_i}. \qquad (6.27)$$

Muestre además que (6.9) es un estimador insesgado de la varianza en (6.27). SUGERENCIA: Utilice las propiedades de la esperanza condicional dadas en el apéndice B y escriba

$$V(\hat{t}_\psi) = V(E[\hat{t}_\psi \mid Q_1, \ldots, Q_N]) + E(V[\hat{t}_\psi \mid Q_1, \ldots, Q_N]).$$

16 Muestre que las dos expresiones para la varianza en (6.13) son equivalentes. SUGERENCIA: Use (6.10) y (6.11).

17 Muestre que (6.14) y (6.15) son equivalentes cuando las unidades primarias y secundarias se eligen con probabilidades iguales, como en el capítulo 5. ¿Son iguales si las unidades primarias se eligen con probabilidades diferentes?

18 Muestre que las fórmulas para el muestreo estratificado en (4.3) y (4.5) son consecuencia de las fórmulas para el estimador de Horvitz-Thompson.

19 Utilice la población del ejemplo 3.4 para este ejercicio. Sea ψ_i proporcional a x_i.

a Utilice el método de extracción por extracción ilustrado en el ejemplo 6.8 para calcular π_i para cada unidad y π_{ij} para cada pareja de unidades, para una muestra sin reemplazo de tamaño 2.

b ¿Cuál es el valor de $V(\hat{t}_{\text{HT}})$? ¿Cuál es su relación con la varianza con reemplazo que utiliza (6.27)?

***20** (Requiere probabilidad.) *Procedimiento de Brewer* (1963, 1975) *para el muestreo sin reemplazo con probabilidades diferentes.* Para una muestra de tamaño $n = 2$, sea π_i la probabilidad deseada de inclusión para la unidad primaria i, con la restricción usual de que $\sum_{i=1}^{N}\pi_i = n$. Sea $\psi_i = \pi_i/2$ y

$$a_i = \frac{\psi_i(1-\psi_i)}{1-\pi_i}.$$

Extraiga la primera unidad primaria con la probabilidad $a_i / \sum_{k=1}^{N} a_k$ de seleccionar la unidad primaria i. Suponga que la unidad primaria i se seleccionó en la primera extracción y seleccione la segunda unidad primaria entre las restantes $N-1$ unidades primarias con probabilidades $\psi_j/(1-\psi_i)$.

a Muestre que

$$\pi_{ij} = \frac{\psi_i\psi_j}{\sum_{k=1}^{N} a_k}\left(\frac{1}{1-\pi_i} + \frac{1}{1-\pi_j}\right).$$

b Muestre que P (seleccionar la unidad primaria i en la muestra) = π_i. SUGERENCIA: Muestre primero que

$$2\sum_{k=1}^{N} a_k = 1 + \sum_{k=1}^{N}\frac{\psi_k}{1-\pi_k}.$$

c El estimador de Sen-Yates-Grundy (SYG) de la varianza en (6.15) para el muestreo de una etapa es

$$\hat{V}(\hat{t}_{\text{HT}}) = \sum_{i\in S}\sum_{\substack{j\in S \\ j>i}}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{t_i}{\pi_i} - \frac{t_j}{\pi_j}\right)^2.$$

Muestre que $\pi_i\pi_j - \pi_{ij} \geq 0$ para el método de Brewer, de modo que el estimador SYG de la varianza es siempre no negativo.

21 La siguiente tabla proporciona los valores de población para una pequeña población de conglomerados:

psu, i	M_i	Valores, y_{ij}	t_i
1	5	3, 5, 4, 6, 2,	20
2	4	7, 4, 7, 7,	25
3	8	7, 2, 9, 4, 5, 3, 2, 6	38
4	5	2, 5, 3, 6, 8,	24
5	3	9, 7, 5	21

Usted quiere seleccionar dos unidades primarias sin reemplazo, con probabilidades de inclusión proporcionales a M_i. Use el método de Brewer del ejercicio 20 para construir una tabla de π_{ij} para las muestras posibles. ¿Cuál es la varianza del estimador de Horvitz-Thompson?

***22** (Requiere probabilidad.) Rao (1963) analiza el siguiente método de rechazo para seleccionar una muestra con ppt sin reemplazo: Seleccione n unidades primarias con probabilidades ψ_i y con reemplazo. Si cualquiera de las unidades primarias aparece más de una vez en la muestra, rechace toda la muestra y seleccione otras n unidades primarias con reemplazo. Repita el procedimiento hasta obtener una muestra de n unidades primarias sin repeticiones.

Determine π_i y π_{ij} para este procedimiento, con $n = 2$.

***23** (Requiere probabilidad.) *Método de Rao-Hartley-Cochran* (1962) *para la selección de unidades primarias con probabilidades diferentes*. Para extraer una muestra de tamaño n, dividimos la población en n grupos aleatorios de unidades primarias, $U_1, U_2, \ldots, U_n$. Luego elegimos una unidad primaria de cada grupo (de manera independiente) con probabilidad proporcional al tamaño. Si la unidad primaria i está en el grupo k, se selecciona con probabilidad $x_{ki} = M_i / \sum_{j \in U_k} M_j$. Sea $\alpha(k)$ la etiqueta de la unidad primaria seleccionada del grupo k. Entonces, condicionalmente sobre los grupos, $t_{\alpha(k)} / x_{k,\alpha(k)}$ estima el total en el grupo k. El estimador del total de la población es

$$\hat{t}_{\text{RHC}} = \sum_{k=1}^{n} \frac{t_{\alpha(k)}}{x_{k,\alpha(k)}}.$$

Muestre que $\hat{t}_{\text{RHC}}$ es insesgado para t y determine su varianza. SUGERENCIA: Use dos conjuntos de variables indicatrices. Sea $I_{ki} = 1$ si la unidad primaria i está en el grupo k y 0 en caso contrario; sea $Z_i = 1$ si la unidad primaria i es seleccionada para estar en la muestra. Entonces $\hat{t}_{\text{RHC}} = \sum_{k=1}^{n} \sum_{i=1}^{N} I_{ki} Z_i t_i / x_{ki}$.

***24** (Requiere cálculo.) En (6.27), suponga que la varianza del estimador del total en la unidad primaria i es $V_i = M_i^2 S_i^2 / m_i$. Si usted sólo puede extraer una submuestra de un total esperado de $C = E\left[\sum_{i \in S} m_i\right]$ unidades secundarias, ¿qué valores de m_i minimizan (6.27)?

25 En el ejemplo 6.9 se mostró que el método de Mitofsky-Waksberg produce una muestra autoponderada si cualquier unidad primaria de ella tiene al menos k números telefónicos residenciales. Suponga que una unidad primaria de la muestra tiene $x < k$ números residenciales. ¿Cuál es el peso relativo de un número telefónico en esa unidad primaria?

26 Una desventaja del método de Mitofsky-Waksberg descrito en el ejemplo 6.9 es que el procedimiento de muestreo secuencial, eligiendo los números de la unidad primaria hasta tener un total de k números residenciales puede ser difícil de implantar. Suponga que en la segunda etapa usted marca otros $(k-1)$ números, sean residenciales o no, y sea x el número de líneas residenciales entre los $(k-1)$ números. ¿Cuáles son los pesos relativos para los números telefónicos residenciales?

27 El método de Mitofsky-Waksberg, descrito en el ejemplo 6.9, proporciona una muestra autoponderada de números telefónicos bajo circunstancias ideales. ¿Proporciona una muestra autoponderada de los adultos? ¿Por qué? En caso negativo, ¿qué pesos relativos deben usarse?

***28** (Requiere probabilidad.) Suponga que se extrae una muestra por conglomerados de tres etapas de una población con N unidades primarias, M_i unidades secundarias en la unidad primaria i, y L_{ij} unidades terciarias de muestreo en la unidad secundaria j de la unidad primaria i. Para extraer la muestra, se eligen al azar n unidades primarias, luego m_i unidades secundarias de la unidad primaria elegida y luego l_{ij} unidades terciarias de la unidad secundaria seleccionada.

a Muestre que los pesos de la muestra son

$$w_{ijk} = \frac{N}{n} \frac{M_i}{m_i} \frac{L_{ij}}{l_{ij}}.$$

b Sea

$$\hat{t} = \sum_{i \in S} \sum_{j \in S_i} \sum_{k \in S_i} w_{ijk} y_{ijk}.$$

Muestre que

$$E[\hat{t}] = t = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \sum_{k=1}^{L_{ij}} y_{ijk}.$$

c Utilice las propiedades de la esperanza condicional de la sección B.4 y determine una expresión para $V(\hat{t})$.

***29** (Basado en el modelo.) Suponga que la población completa se observa en la muestra, de modo que $n = N$ y $m_i = M_i$. Examine los tres estimadores $\hat{T}_{\text{ins}}, \hat{T}_r$ (de la sección 5.7) y $\hat{T}_P$ (de la sección 6.7). Si se observa toda la población, ¿cuáles de estos estimadores son iguales a $T = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij}$?

7

Encuestas complejas

No hay remedio más efectivo para aplacar la rabiosa susceptibilidad de la gente que las cifras. Para estar segura, debe prepararse cuidadosamente, abarcar el caso y no sólo su cuarta parte, y debe reunirse por sí misma, no para comprobar alguna teoría. Este tipo de preparación puede encontrarla en el censo nacional.

—Ida Tarbell, *The Ways of Woman* (1915)

La mayoría de las encuestas grandes comprende varias de las ideas que hemos analizado: una encuesta debe estar estratificada con varias etapas de formación de conglomerados y basarse en estimaciones de proporción o regresión para ajustarse a las otras variables. Las fórmulas para estimar los errores estándar pueden ser horrorosas, particularmente si hay varias etapas de formación de conglomerados sin reemplazo. Generalmente, los pesos de muestreo y los efectos del diseño se utilizan en las encuestas complejas para simplificar las cosas. En este capítulo analizamos estos aspectos y las gráficas para los datos de encuestas complejas. Concluimos con una descripción del diseño de la encuesta nacional de víctimas de crímenes y con los aspectos paralelos de las muestras de encuestas y los experimentos diseñados.

7.1 Integración de los componentes del diseño

Hemos visto la mayoría de los componentes de una encuesta compleja: el muestreo aleatorio, la estimación de proporción, la estratificación y la formación de conglomerados. Ahora veremos cómo integrarlos en un diseño de muestreo. Aunque en la práctica frecuentemente se emplean los pesos (sección 7.2) para determinar las estimaciones puntuales y los métodos con uso intensivo de computadoras (capítulo 9) para calcular las varianzas de las estimaciones, es importante comprender los principios básicos del funcionamiento integrado de los componentes. Aquí aparecen los conceptos que usted ya conoce, en forma modular, listos para integrarse.

7.1.1 Bloques de construcción de las encuestas

1 *Muestreo por conglomerados con reemplazo.* Seleccione una muestra de n conglomerados con reemplazo; el conglomerado i se selecciona con probabilidad ψ_i. Estime el total

del conglomerado i mediante una estimador insesgado $\hat{t}_i$. Luego, considere los n valores (la muestra es con reemplazo, de modo que algunos de los valores del conjunto pudieran ser de la misma unidad primaria de muestreo) de $u_i = \hat{t}_i/\psi_i$ como observaciones. Estime el total de la población mediante $\overline{u}$ y estime la varianza del total estimado mediante s_u^2/n.

2 *Muestreo por conglomerados sin reemplazo.* Seleccione una muestra de n conglomerados sin reemplazo; la probabilidad de elegir el conglomerado i para la muestra es π_i. Estime el total para el conglomerado i mediante un estimador insesgado $\hat{t}_i$ y calcule una estimación insesgada de la varianza de $\hat{t}_i$, $\hat{V}(\hat{t}_i)$. Luego, estime el total de la población mediante el estimador de Horvitz-Thompson[1] a partir de la ecuación (6.12):

$$\hat{t}_{HT} = \sum_{i \in S} \frac{\hat{t}_i}{\pi_i}.$$

Utilice una fórmula exacta de los capítulos 5 o 6, o un método del capítulo 9, para estimar la varianza.

3 *Estratificación* Sean $\hat{t}_1,\ldots,\hat{t}_H$ estimadores insesgados de los totales de los estratos $t_1,\ldots,t_H$ y sean $\hat{V}(\hat{t}_1),\ldots,\hat{V}(\hat{t}_H)$ estimadores insesgados de las varianzas. Entonces, estime el total mediante

$$\hat{t} = \sum_{h=1}^{H} \hat{t}_h$$

y su varianza mediante

$$\hat{V}(\hat{t}) = \sum_{h=1}^{H} \hat{V}(\hat{t}_h).$$

4 *Estimación de proporción.* Sean $\hat{t}_y$ y $\hat{t}_x$ estimadores insesgados de t_y y t_x, respectivamente. Entonces, la proporción se estima mediante

$$\hat{B} = \frac{\hat{t}_y}{\hat{t}_x}$$

y su varianza mediante

$$\hat{V}(\hat{B}) = \frac{\hat{B}^2}{t_x^2}\hat{V}(\hat{t}_x) + \frac{1}{t_x^2}\hat{V}(\hat{t}_y) - 2\frac{\hat{B}}{t_x^2}\widehat{\mathrm{Cov}}(\hat{t}_x,\hat{t}_y), \tag{7.1}$$

como se mostrará en la sección 9.1. El estimador de proporción del total es $\hat{B}t_x$ con varianza estimada $t_x^2\hat{V}(\hat{B})$.

Frecuentemente empleamos las razones para estimar las medias, haciendo que la variable auxiliar x_i sea igual a 1 si la unidad i está en la muestra y 0 en caso contrario. Entonces $\hat{t}_x$ estima el tamaño de la población y la razón divide el total estimado de la población entre su tamaño estimado.

Generalmente, la estratificación forma la clasificación más gruesa: los estratos pudieran ser, por ejemplo, áreas del país, códigos de área o tipos de hábitat. Se extraen muestras de los conglomerados (a veces, varios niveles de conglomerados) de cada estrato del diseño y puede haber una estratificación adicional en ellos. Si hay varios niveles de formación de

[1] Recuerde que el estimador de Horvitz-Thompson abarca a los otros estimadores insesgados sin reemplazo del total como casos particulares, como analizamos en la sección 6.4.

conglomerados y estratificación, es conveniente trazar un diagrama o una tabla del diseño de la encuesta, como se muestra en el siguiente ejemplo.

EJEMPLO 7.1 La malaria es un serio problema de salud en Gambia. La morbilidad por la enfermedad se puede reducir usando mosquiteros para cama impregnados con insecticida, pero eso sólo es eficaz si el uso de mosquiteros es amplio. En 1991 se diseñó una encuesta de cobertura nacional para estimar la frecuencia del uso de los mosquiteros en las áreas rurales. La descripción de ese trabajo, junto con sus resultados, aparece en D'Alessandro *et al.* (1994)

El marco de muestreo comprendió de todas las villas rurales con menos de 3000 personas en Gambia. Los pobladores se estratificaron en tres regiones geográficas (este, centro y oeste) y dependiendo de si el poblado tenía centro de salud o no. En cada región se eligieron cinco distritos, con probabilidad proporcional a su población según las estimaciones del censo nacional de 1983. En cada distrito se seleccionaron cuatro villas, de nuevo con probabilidad proporcional a la población del censo: dos poblaciones con centro de salud y dos sin centro. Por último, de manera más o menos aleatoria se eligieron seis complejos de cada villa y un investigador registró la cantidad de camas y mosquiteros, junto con otras informaciones, para cada complejo.

En resumen, el diseño de la muestra es el siguiente:

Etapa	Unidad de muestreo	Estratificación
1	Distrito	Región
2	Población	Con/sin centro de salud
3	Complejo	

Para calcular las estimaciones o los errores estándar mediante las fórmulas de los capítulos anteriores, se comenzaría con el nivel 3 e iría subiendo. Los siguientes son los pasos que se emplearían para estimar la cantidad total de mosquiteros (sin utilizar la estimación de razón):

1. Registre la cantidad total de mosquiteros por cada complejo.
2. Estime la cantidad total de mosquiteros para cada pueblo mediano (cantidad de complejos en el poblado) × (cantidad promedio de mosquiteros por complejo). Determine la varianza estimada de la cantidad total de mosquiteros para cada pueblo.
3. Estime la cantidad total de mosquiteros para los poblados con centro de salud en cada distrito. La muestra de villas del distrito se obtuvo con probabilidades proporcionales a la población, de modo que debemos usar las fórmulas del capítulo 6 para estimar el total y la varianza del total estimado. Repita el procedimiento para los pueblos sin centro de salud en cada distrito.
4. Sume las estimaciones de los dos estratos (con/sin centro de salud) para estimar la cantidad de mosquiteros en cada distrito; sume las varianzas estimadas de los dos estratos para estimar la varianza del distrito.
5. Ya tiene la cantidad total estimada de mosquiteros y la varianza estimada para cada distrito. Ahora, utilice las fórmulas del muestreo por conglomerados en dos etapas para estimar la cantidad total de mosquiteros para cada región.
6. Por último, sume los totales estimados para cada región para estimar la cantidad total de mosquiteros para cama en Gambia. Sume las varianzas de las regiones, como en el caso del muestreo estratificado.

Parece un poco complicado, ¿no? Y todavía no hemos incluido la estimación de razón, que casi seguramente tendremos que incorporar, porque conocemos las cantidades aproximadas de la población para la cantidad de camas en cada nivel. Como veremos más adelante en este capítulo y en el capítulo 9, podemos utilizar los pesos de muestreo y los métodos mediante el empleo de computadoras para evitar buena parte de este esfuerzo. ■

7.1.2 Estimación de razón en encuestas complejas

La estimación de razón es parte del análisis, no del diseño, y no aparece en un diagrama del diseño. La estimación de razón se puede utilizar en casi cualquier nivel de la encuesta, aunque generalmente se emplea cerca de la parte superior.

Parte del interés en el estudio de mosquiteros para cama era la proporción de camas con ellos. La razón utilizada para las proporciones se podría calcular en casi cualquier nivel de la encuesta; para simplificar nuestro análisis, supongamos que sólo nos interesan los pueblos con centros de salud. En lo sucesivo, x se refiere a las camas y y a los mosquiteros.

1 *Nivel de complejo*. Calcule la proporción de camas en el complejo que tiene mosquiteros y utilice estas proporciones como las observaciones. Luego, la estimación por pueblo sería el promedio de las proporciones de los seis complejos, la estimación por distrito se calcularía a partir de las cinco estimaciones de los pueblos, y así sucesivamente. Esto es análogo al estimador "media de razones" del ejercicio 22 del capítulo 3.

2 *Nivel poblados*. Para cada población, calcule (cantidad total de mosquiteros)/(cantidad total de camas). La varianza estimada por poblado se calcula mediante (7.1). Entonces, por distrito, promedie los cocientes obtenidos para las poblaciones del distrito.

3 *Nivel distrito*. Esto es análogo al nivel poblado, excepto que se forman los cocientes para cada distrito.

4 *Nivel región*. Use las fórmulas ppt para estimar la cantidad total de camas y la cantidad total de mosquiteros para las regiones C (centro), E (este) y W (oeste). El resultado son seis estimaciones de los totales $\hat{t}_{xC}, \hat{t}_{xE}, \hat{t}_{xW}, \hat{t}_{yC}, \hat{t}_{yE}, \hat{t}_{yW}$ y estimaciones de las varianzas y covarianzas asociadas con los totales estimados. Ahora, calcule los tres cocientes $\hat{t}_{yC}/\hat{t}_{xC}, \hat{t}_{yE}/\hat{t}_{xE}$, y $\hat{t}_{yW}/\hat{t}_{xW}$ y utilice la fórmula para estimación de cocientes para estimar la varianza de cada cociente. Luego, combine las tres estimaciones de cocientes utilizando la estratificación:

$$\hat{B} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \frac{\hat{t}_{yh}}{\hat{t}_{xh}}.$$

$$\hat{V}(\hat{B}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \hat{V}\left(\frac{\hat{t}_{yh}}{\hat{t}_{xh}}\right).$$

5 *Arriba del nivel región*. Use la estratificación para estimar $\hat{t}_y$ y $\hat{t}_x$ para toda la población, junto con las varianzas y covarianzas estimadas. Ahora, estime el cociente $\hat{t}_y/\hat{t}_x$ y use (7.1) para calcular la varianza.

Recuerde, del capítulo 3, que el estimador de *proporción* es sesgado y el sesgo puede ser importante para muestras pequeñas. El tamaño de muestra es pequeño para varios niveles, de modo que deberá tener mucho cuidado con el estimador: sólo se extrae una muestra de seis complejos por poblado y cinco de ellos por distrito, de modo que el sesgo es una cuestión importante en esos niveles. Cassel *et al* (1977, capítulo 7) compara varias estrategias relativas a los estimadores de proporciones.

En el nivel región, una estimación comparable del total de la población es el **estimador de proporción por separado**:

$$\sum_{h=1}^{H} \frac{\hat{t}_{xh}\hat{t}_{yh}}{\hat{t}_{xh}}.$$

La estimación de proporción, realizada por separado en cada estrato, puede mejorar la eficiencia si los $\hat{t}_{yh}/\hat{t}_{xh}$ varían de un estrato a otro. *No* debe usarse cuando los tamaños de muestra de los estratos son pequeños, pues cada cociente es sesgado y el sesgo se puede propagar a través de los estratos.

Por arriba del nivel región, el **estimador de proporción combinado** $t_x\hat{t}_y/\hat{t}_x$ brinda una estimación comparable del total de la población. El estimador combinado tiene un sesgo menor cuando se extraen pocas unidades primarias por cada estrato. Sin embargo, cuando las proporciones varían mucho de un estrato a otro, el estimador combinado no aprovecha la eficiencia adicional debida a la estratificación, como hace el estimador de proporción por separado.

7.1.3 Sencillez en el diseño de una encuesta

Todos los componentes de un diseño aumentan su eficiencia de un estudio a otro. Sin embargo, en ocasiones el diseñador de encuestas poco experimentado puede estar tentado a usar un diseño de muestreo complejo sólo porque existe o porque se ha usado anteriormente, no porque haya demostrado ser más eficiente. Se debe garantizar, por medio de pruebas previas o investigaciones anteriores, que un diseño complejo realmente es más eficiente y práctico; casi siempre hay que preferir un diseño más sencillo que proporcione la misma cantidad de información por unidad monetaria invertida, a un diseño más complejo: frecuentemente, es más fácil de administrar y analizar, y es menos probable que analistas posteriores procesen incorrectamente los datos de la investigación. Un diseño complejo debe ser eficiente al estimar *todas* las cantidades de interés primario; una distribución óptima en un muestreo estratificado para estimar la cantidad total invertida en planes de salud por las empresas de Estados Unidos puede ser demasiado ineficiente para estimar el porcentaje de empresas que se declaran en bancarrota en un año.

7.2 Pesos de muestreo

7.2.1 Construcción de pesos de muestreo

En muchas encuestas muestrales de gran tamaño, los pesos se utilizan para trabajar con los efectos de la estratificación y la formación de conglomerados sobre las estimaciones puntuales. Ya hemos visto la manera de utilizar los pesos de muestreo en el muestreo estratificado y el muestreo por conglomerados. El peso de muestreo para una unidad de observación es siempre el recíproco de la probabilidad de que se elija la unidad de observación para la muestra.

Recuerde que, para el muestreo estratificado,

$$\hat{t}_{\text{est}} = \sum_{h=1}^{H} \sum_{j \in S_h} \omega_{hj} y_{hj},$$

donde el peso de muestreo $w_{hj} = (N_h/n_h)$ se puede pensar como la cantidad de observaciones en la población representadas por la observación muestral y_{hj}. La probabilidad de seleccionar la unidad j en el estrato h para estar en la muestra es $\pi_{hj} = n_h/N_h$, de modo que el peso de muestreo no es más que el inverso de la probabilidad de selección: $w_{hj} = 1/\pi_{hj}$.

La suma de los pesos de muestreo en el muestreo estratificado es igual al tamaño de la población, N; cada unidad de la muestra "representa" cierta cantidad de unidades en la población, de modo que toda la muestra "representa" toda la población. La estimación de $\bar{y}_U$ para el muestreo estratificado es

$$\bar{y}_{\text{est}} = \frac{\sum_{h=1}^{H} \sum_{j \in S_h} \omega_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j \in S_h} \omega_{hj}}.$$

En la sección 5.4 utilizamos las mismas formas de los estimadores en el muestreo por conglomerados, y en la sección 6.4 mostramos la forma general de los estimadores ponderados. En el muestreo por conglomerados con probabilidades iguales,

$$w_{ij} = \frac{NM_i}{nm_i} = \frac{1}{\text{probabilidad de que la unidad secundaria } j \text{ de la unidad primaria } i \text{ esté en la muestra}}$$

De nuevo,

$$\hat{t} = \sum_{i \in S} \sum_{j \in S_i} \omega_{ij} y_{ij},$$

y la estimación de la media de la población es

$$\frac{\hat{t}}{\sum_{i \in S} \sum_{j \in S_i} w_{ij}}.$$

Para el muestreo por conglomerados con probabilidades diferentes, cuando π_i es la probabilidad de que la unidad primaria i esté en la muestra y $\pi_{j|i}$ es la probabilidad de que la unidad secundaria j esté en la muestra dado que la unidad primaria i está en la muestra, los pesos de muestreo son $w_{ij} = 1/(\pi_i \pi_{j|i})$.

Para el muestreo por conglomerados en tres etapas, el principio se extiende: Sea w_p el peso para la unidad primaria, $w_{s|p}$ el peso para la unidad secundaria y $w_{t|s,p}$ el peso asociado a la unidad terciaria de muestreo. Entonces, el peso general del muestreo para una unidad de observación es

$$w = w_p \times w_{s|p} \times w_{t|s,p}.$$

Toda la información necesaria para construir las estimaciones puntuales se incluye en los pesos de muestreo; al calcular las estimaciones puntuales, las probabilidades de selección de las unidades primarias, secundarias y terciarias (que pueden tener expresiones complicadas) sólo aparecen a través de los pesos. Pero no obstante, los pesos de muestreo no dan información acerca de la forma de determinar los errores estándar de las estimaciones, por lo que su conocimiento no permite establecer inferencias estadísticas. Las varianzas de las estimaciones dependen de las probabilidades de que cualquier pareja de unidades sea seleccionada para estar en la muestra y requiere más conocimiento del diseño de muestreo que el dado simplemente por los pesos.

Con frecuencia, los pesos muy grandes se truncan para que ninguna observación individual tenga una contribución demasiado grande a la estimación global. Aunque esto sesga el estimador, puede reducir el error cuadrático medio (ECM). El truncamiento se utiliza con frecuencia cuando se emplean los pesos para ajustar la ausencia de respuesta, como se describe en el capítulo 8.

Como en el resto del libro consideraremos los diseños estratificados en varias etapas, a partir de este momento adoptaremos una notación unificada para las estimaciones del total de la población. Consideramos a y_i como una medición sobre la unidad de observación i y w_i como el peso de muestreo de la unidad de observación i. Así, para una muestra estratificada, y_i es una unidad de observación dentro de un estrato particular y $w_i = N_h/n_h$, donde la unidad i está en el estrato h. Esto nos permite escribir el estimador general del total de la población como

$$\hat{t}_y = \sum_{i \in S} w_i y_i, \tag{7.2}$$

donde todas las mediciones se realizan en el nivel de la unidad de observación. El estimador general de la media de la población es

$$\hat{\bar{y}} = \frac{\hat{t}_y}{\sum_{i \in S} w_i}; \tag{7.3}$$

$\sum_{i \in S} w_i$ estima la cantidad de unidades de observación, N, en la población.

EJEMPLO 7.2 El estudio de los mosquiteros para camas en Gambia, del ejemplo 7.1, se diseñó de modo que dentro de cada región, cada complejo tendría casi la misma probabilidad de incluirse; las probabilidades sólo variaban debido a que distritos diferentes tenían distinta cantidad de personas en los poblados con centros de salud y debido a que la cantidad de complejos podría no ser exactamente proporcional a la población de la comunidad; para los que tenían con centros de salud en la región central, por ejemplo, la probabilidad de incluir un complejo dado en la investigación era

$$P(\text{seleccionar al distrito}) \times P(\text{seleccionar el poblado} \mid \text{se seleccionó el distrito}) \times P(\text{seleccionar el complejo} \mid \text{se seleccionó el distrito y el poblado})$$

$$\propto \frac{D1}{R} \times \frac{V}{D2} \times \frac{1}{C},$$

donde

C es la cantidad de complejos en el poblado

V es la cantidad de personas en el poblado

$D1$ es la cantidad de personas en el distrito

$D2$ es la cantidad de personas en el distrito, en poblados con centros de salud

R es la cantidad de personas en los poblados con centros de salud en todos los distritos del centro

Como la cantidad de complejos en un poblado será aproximadamente proporcional a la cantidad de personas en una población, V/C será aproximadamente igual para todos los complejos. R también es igual para todos los complejos dentro de una región. Los pesos para cada región, los recíprocos de las probabilidades de inclusión, difieren en gran medida, debido a la variabilidad en $D1/D2$. Sin embargo, conforme R varía de un estrato a otro, los complejos en estratos más populosos tendrán mayores pesos que los correspondientes a estratos menos poblados. ■

7.2.2 Muestras autoponderadas y no autoponderadas

Los pesos de muestreo para todas las unidades de observación son los mismos en las encuestas autoponderadas. Las muestras autoponderadas pueden, en ausencia de errores no debidos al muestreo, considerarse como representativas de la población, debido a que cada unidad observada representa la misma cantidad de unidades no observadas de la población; entonces, se pueden aplicar los métodos estadísticos comunes a la muestra, para obtener las estimaciones puntuales. Un histograma con los valores de la muestra despliega las frecuencias de ocurrencia aproximadas en la población; la media de la muestra, la mediana y otras estadísticas de la muestra estiman las cantidades correspondientes de la población. Además, las muestras autoponderadas con frecuencia tienen una varianza menor y las estadísticas de la muestra son más robustas (Kish 1992).

Sin embargo la mayoría de las muestras autoponderadas grandes utilizadas en la práctica no son muestras aleatorias simples; la estratificación se utiliza para reducir las varianzas y obtener estimaciones individuales para los dominios de interés; la formación de conglomerados, normal con ppt, se emplea para reducir costos. El software estadístico común (programas para análisis de datos que satisfagan la hipótesis estadística adecuada, en el sentido de que las observaciones sean independientes y estén idénticamente distribuidas) proporciona estimaciones correctas de la media, los percentiles y demás cantidades en una encuesta compleja autoponderada. Sin embargo, los errores estándar, las estadísticas de prueba de hipótesis y los intervalos de confianza construidos mediante ese software son incorrectos, como ya hemos mencionado. Al leer un artículo o libro en el que los autores analicen los datos de una encuesta compleja, observe si tomaron en cuenta la estructura de los datos en el análisis o si simplemente pasaron los datos en bruto por un procedimiento normal de SAS o SPSS y dieron esos resultados. Si es así, sus inferencias deben tomarse con desconfianza; es posible que sólo tengan significado estadístico debido a que no participaron en el diseño de la encuesta, con relación a los errores estándar.

Por supuesto, muchas encuestas extraen deliberadamente unidades de observación con probabilidades diferentes. Las probabilidades de muestreo desproporcionadas aparecen frecuentemente en la estratificación: se utiliza una fracción de muestreo mayor para un estrato de grandes empresas que de pequeñas. La encuesta nacional de salud y nutrición de Estados Unidos (NHANES) extrae una muestra excesiva, voluntariamente, de las áreas con grandes poblaciones de negros y mexicano-estadounidenses (Ezzati-Rice y Murphy 1995); el exceso de muestreo en estas poblaciones permite comparar la salud de las minorías.

7.2.3 Pesos y un análisis de los datos de una encuesta con base en un modelo

Se podría pensar que un estadístico con una perspectiva basada en el modelo podría ignorar por completo los pesos. Después de todo, para un estadístico de investigaciones basadas en un modelo, el diseño de la muestra es irrelevante y la parte importante del análisis consiste en encontrar un modelo que resuma la estructura de la población; como los pesos de muestreo son funciones de las probabilidades de selección en el diseño, tal vez también sean irrelevantes.

Sin embargo, los enfoques basados en el modelo y en la aleatorización no están tan alejados como podría sugerir parte de la literatura sobre el tema. Recuerde que el estadístico que diseña una encuesta que será analizada mediante pesos visualiza implícitamente un modelo para los datos; NHANES está estratificada y las subpoblaciones se representan de manera excesiva porque los investigadores suponen que habrá una diferencia entre las

subpoblaciones. Estas diferencias también deben incluirse en el modelo. Si se ignoran los pesos al analizar los datos de NHANES, se por ejemplo, se estará suponiendo implícitamente que los blancos, los negros y los mexicano-estadounidenses pueden intercambiarse totalmente en cuanto a su estado de salud. Ignorar la formación de conglomerados en la inferencia supone que las observaciones del mismo conglomerado no están correlacionadas, lo que generalmente es falso. Un analista de datos que ignore las variables de estratificación y la dependencia entre las observaciones no está ajustando un buen modelo a los datos, sino que simplemente es indolente. Un buen análisis de los datos de una encuesta con modelos es difícil y requiere de una amplia validación del modelo. El libro de Skinner *et al* (1989) incluye varios capítulos sobre el modelado de datos a partir de encuestas complejas.

Muchos investigadores han descubierto que los pesos de muestreo contienen información que se puede utilizar en un análisis basado en el modelo. Little (1991) desarrolla una clase de modelos que producen estimadores cuyo comportamiento es similar al de los obtenidos usando pesos en la encuesta. Pfeffermann (1993) describe un marco de referencia para decidir si deben utilizarse los pesos de muestreo en los modelos de regresión de los datos de una encuesta.

7.3 Estimación de una función de distribución

Hasta ahora, nos hemos concentrado en la estimación de las medias, totales y proporciones de una población. Desde un punto de vista histórico, la teoría de muestreo se desarrolló principalmente para determinar tales estadísticas básicas y contestar preguntas como "¿qué porcentaje de hombres adultos está desempleado?", "¿cuál es la cantidad total de dinero destinado al cuidado de la salud en Estados Unidos?" o "¿cuál es la proporción entre la cantidad de aves exóticas y nativas en cierta área?".

Determinadas estadísticas distintas de las medias o los totales pueden ser interesantes. Tal vez desee estimar el ingreso mediano en Canadá, el percentil 95 de las calificaciones o construir un histograma para mostrar la distribución de la longitud de ciertos peces. Una compañía de seguros podría establecer los reembolsos por un procedimiento médico mediante el percentil 75 de los costos del procedimiento. Podemos estimar cualquiera de estas cantidades (aunque no sus errores estándar) con los pesos de muestreo, que nos permiten construir una distribución empírica para la población.

Suponga que conocemos los valores para toda la población de N unidades. Entonces, cualquier cantidad de interés puede calcularse mediante la **función de masa de probabilidad**,

$$f(y) = \frac{\text{cantidad de unidades cuyo valor es } y}{N},$$

o la **función de distribución**,

$$f(y) = \frac{\text{cantidad de unidades con valor } \leq y}{N} = \sum_{x \leq y} f(x).$$

En la teoría de probabilidad, éstas son la función de masa de población y la función de distribución para la variable aleatoria Y, donde Y es el valor obtenido de una muestra aleatoria de tamaño 1 de la población. Entonces $f(y) = P\{Y = y\}$ y $F(y) = P\{Y \leq y\}$. Por supuesto, $\sum f(y) = F(\infty) = 1$.

Cualquier cantidad relativa a la población se puede calcular mediante la función de masa de probabilidad o la función de distribución. La media de la población es

$$\bar{y}_U = \sum y f(y).$$

Una mediana de la población es cualquier valor m que satisfaga $F(m) \geq 1/2$ y $P(Y \geq m) \geq 1/2$; en general, x es un percentil r si $F(x) \geq r$ y $P(Y \geq x) \geq 1 - r$. La varianza de la población también se puede escribir mediante la función de masa de probabilidad:

$$\begin{aligned} S^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_U)^2 \\ &= \frac{N}{N-1}\sum_y f(y)\left[y - \sum_x x f(x)\right]^2 \\ &= \frac{N}{N-1}\left[\sum_y y^2 f(y) - \left\{\sum_y y f(y)\right\}^2\right]. \end{aligned}$$

EJEMPLO 7.3 Consideremos una población artificial de 1000 hombres y 1000 mujeres, en el archivo htpop.dat. Medimos la altura de cada persona, redondeada a centímetros. La tabla de frecuencias (tabla 7.1) proporciona la función de masa de probabilidad y la función de distribución para las 2000 personas de la población. Las figuras 7.1 y 7.2 muestran las gráficas de $F(y)$ y $f(y)$. La media de la población es $\sum y f(y) = 168.6$.

Ahora, extraemos una muestra aleatoria simple de tamaño 200 de la población (archivo htsrs.dat). Una muestra aleatoria simple es autoponderada: cada persona de la muestra representa 10 personas de la población. Por lo tanto, el histograma de la muestra se parecerá a $f(y)$ de la población; la figura 7.3 muestra que esto realmente ocurre.

Pero suponga que se extrae una muestra estratificada de 160 mujeres y 40 hombres (archivo htstrat.dat) en lugar de una muestra autoponderada. Un histograma de los datos en bruto distorsionará la distribución de la población, como muestra la figura 7.4. La media y mediana de la muestra son demasiado pequeñas, pues los hombres están subrepresentados en la muestra. ■

Los pesos de muestreo nos permiten construir funciones empíricas de masa de probabilidad y de función de distribución para los datos que no permiten calcular cualquier estadística. Definimos la **función de masa de probabilidad empírica** como la suma de los pesos para todas las observaciones que asumen el valor y, dividida entre la suma de todos los pesos:

$$\hat{f}(y) = \frac{\sum_{i \in S; y_i = y} w_i}{\sum_{i \in S} w_i}.$$

La **función de distribución empírica** $\hat{F}(y)$ es la suma de todos los pesos para las observaciones con valores $\leq y$, dividida entre la suma de todos los pesos:

$$\hat{F}(y) = \sum_{x \leq y} \hat{f}(x).$$

TABLA 7.1
Tabla de frecuencias para la población del ejemplo 7.3

Valor, y	Frecuencia	$f(y)$	$F(y)$	Valor, y	Frecuencia	$f(y)$	$F(y)$
136	1	0.0005	0.0005	172	57	0.0285	0.6540
140	1	0.0005	0.0010	173	45	0.0225	0.6765
141	2	0.0010	0.0020	174	52	0.0260	0.7025
142	1	0.0005	0.0025	175	57	0.0285	0.7310
143	6	0.0030	0.0055	176	49	0.0245	0.7555
144	3	0.0015	0.0070	177	54	0.0270	0.7825
145	4	0.0020	0.0090	178	57	0.0285	0.8110
146	3	0.0015	0.0105	179	40	0.0200	0.8310
147	14	0.0070	0.0175	180	35	0.0175	0.8485
148	11	0.0055	0.0230	181	43	0.0215	0.8700
149	13	0.0065	0.0295	182	29	0.0145	0.8845
150	20	0.0100	0.0395	183	26	0.0130	0.8975
151	15	0.0075	0.0470	184	29	0.0145	0.9120
152	18	0.0090	0.0560	185	23	0.0115	0.9235
153	28	0.0140	0.0700	186	21	0.0105	0.9340
154	38	0.0190	0.0890	187	19	0.0095	0.9435
155	38	0.0190	0.1080	188	17	0.0085	0.9520
156	57	0.0285	0.1365	189	15	0.0075	0.9595
157	53	0.0265	0.1630	190	10	0.0050	0.9645
158	49	0.0245	0.1875	191	14	0.0070	0.9715
159	55	0.0275	0.2150	192	10	0.0050	0.9765
160	77	0.0385	0.2535	193	9	0.0045	0.9810
161	72	0.0360	0.2895	194	7	0.0035	0.9845
162	66	0.0330	0.3225	195	2	0.0010	0.9855
163	62	0.0310	0.3535	196	7	0.0035	0.9890
164	61	0.0305	0.3840	197	8	0.0040	0.9930
165	60	0.0300	0.4140	198	4	0.0020	0.9950
166	75	0.0375	0.4515	199	2	0.0010	0.9960
167	79	0.0395	0.4910	200	4	0.0020	0.9980
168	62	0.0310	0.5220	201	1	0.0005	0.9985
169	79	0.0395	0.5615	204	1	0.0005	0.9990
170	72	0.0360	0.5975	206	2	0.0010	1.0000
171	56	0.0280	0.6255				

FIGURA 7.1
La función $F(y)$ para la población de alturas

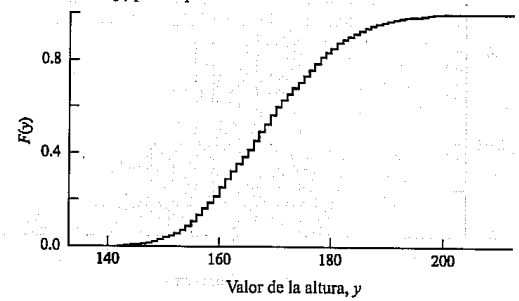


FIGURA 7.2
La función $f(y)$ para la población de alturas

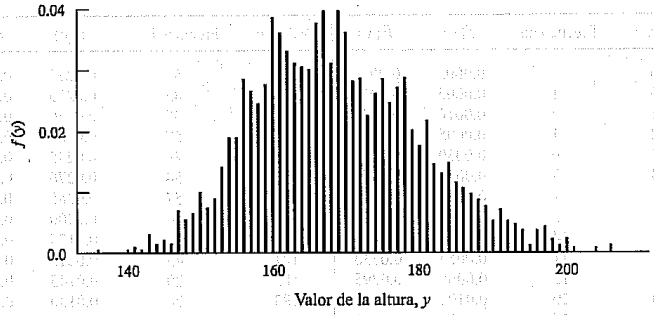


FIGURA 7.3
Un histograma de los datos en bruto a partir de una muestra aleatoria simple de tamaño 200. La forma general es similar a la de $f(y)$ para la población, pues la muestra es autoponderada.

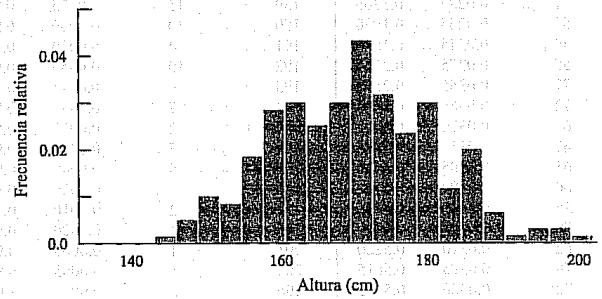


FIGURA 7.4
Un histograma de datos en bruto de una muestra estratificada de 160 mujeres y 40 hombres. Las personas altas están subrepresentadas en la muestra.

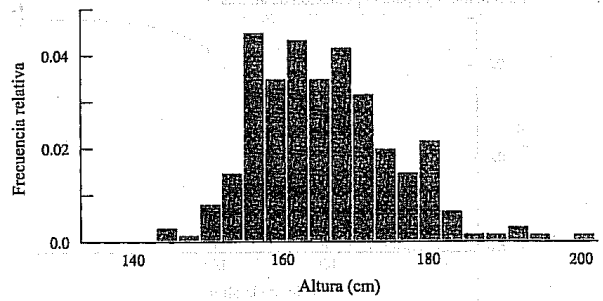
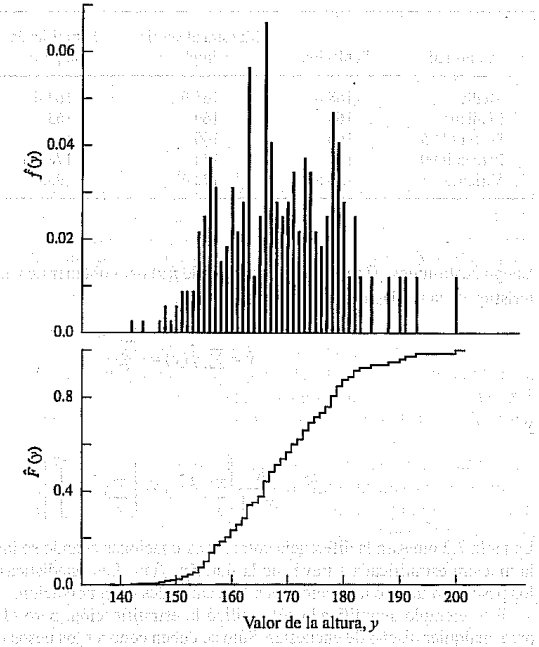


FIGURA 7.5
Las estimaciones $\hat{f}(y)$ y $\hat{F}(y)$ para la muestra estratificada de 160 mujeres y 40 hombres.



Para una muestra autoponderada, $\hat{f}(y)$ se reduce a la frecuencia relativa de y en la muestra. Para una muestra no autoponderada, $\hat{f}(y)$ y $\hat{F}(y)$ son intentos por reconstruir las funciones f y F de la población, con base en la muestra. El peso w_i es la cantidad de unidades de población representadas por la unidad i, de modo que $\sum_{i \in S; y_i = y} w_i$ estima la cantidad total de unidades en la población que tienen el valor y.

Cada mujer de la muestra estratificada tiene peso de muestreo 6.25; cada hombre tiene peso de muestreo 25. Las funciones empíricas de masa de probabilidad y de distribución obtenidas a partir de la muestra estratificada son como en la figura 7.5. Los pesos corrigen la subrepresentación de las personas más altas, visible en el histograma de la figura 7.4. Sin embargo, la escasez de hombres en la muestra tiene su precio: la cola derecha de $\hat{f}(y)$ tiene unos cuantos picos de tamaño 25/2000, en lugar de varios valores que la vayan disminuyendo poco a poco.

La función de masa de probabilidad empírica $\hat{f}(y)$ se puede usar para determinar estimaciones de las cantidades de la población. En primer lugar, expresamos la característica de la población en términos de $f(y)$: $\bar{y}_U = \sum y f(y)$ o

$$S^2 = \frac{N}{N-1}\left\{\sum_y f(y)\left[y - \sum_x x f(x)\right]^2\right\} = \frac{N}{N-1}\left\{\sum_y y^2 f(y) - \left[\sum_y y f(y)\right]^2\right\}.$$

TABLA 7.2
Estimaciones de las muestras en el ejemplo 7.3

Cantidad	Población	Muestra aleatoria simple	Estratificada sin pesos	Estratificada con pesos
Media	168.6	168.9	164.6	169.0
Mediana	168	169	163	168
Percentil 25	160	160	157	161
Percentil 90	184	184	178	182
Varianza	124.5	122.6	93.4	116.8

Luego sustituimos $\hat{f}(y)$ en cada aparición de $f(y)$ para obtener una estimación de la característica de la población. Así, con este método,

$$\hat{\bar{y}} = \sum_y y\hat{f}(y) = \frac{\sum_{i \in S} y_i w_i}{\sum_{i \in S} w_i}$$

y

$$\hat{S}^2 = \frac{N}{N-1}\left\{\sum_y y^2 \hat{f}(y) - \left[\sum_y y\hat{f}(y)\right]^2\right\}. \qquad (7.4)$$

La tabla 7.2 muestra la diferencia entre las estimaciones cuando se incorporan los pesos de la muestra estratificada a través de la función $\hat{f}(y)$. Las estadísticas calculadas mediante los pesos son mucho más cercanas a las cantidades de la población.

Este ejemplo simplificado sólo utilizó la estratificación, pero el método es el mismo para cualquier diseño de encuestas. Sólo se deben conocer los pesos de muestreo para estimar casi todo mediante la función de distribución empírica. Si se desea, puede suavizar la función de distribución empírica antes de estimar las cantidades; consulte Silverman (1986), Scott (1992) o Venables y Ripley (1994, sección 5.5). Gill *et al* (1988) muestran que esta función de distribución empírica es uniforme y asintóticamente consistente; sin embargo, en muestras pequeñas, las colas de la función de masa de probabilidad empírica $\hat{f}(y)$ frecuentemente son muy cortas, con muestras autoponderadas o no, pues los valores extremos podrían no quedar incluidos en la muestra.[2] Nusser *et al* (1996) utilizan un enfoque semiparamétrico para estimar el consumo diario de diversos nutrientes en la encuesta continua de consumo de alimentos, una encuesta estratificada de varias etapas.

Aunque los pesos se pueden utilizar para determinar las estimaciones puntuales mediante la función de distribución empírica, el cálculo de los errores estándar es mucho más complejo y requiere conocimiento del diseño de muestreo. Las varianzas de las estadísticas calculadas a partir de la función de distribución empírica se analizarán en el capítulo 9.

[2] Pueden aparecer más problemas al estimar las funciones de distribución, pues quienes responden podrían redondear sus respuestas. Por ejemplo, algunas personas podrían redondear su altura a 165 o 170 cm, provocando la aparición de picos en esos valores. Si usted suaviza la función de masa de probabilidad empírica, podría elegir un ancho de banda que incremente el suavizado o tal vez desee adoptar un modelo para el efecto de redondeo por parte de los encuestados.

7.4 Graficación de datos de una encuesta compleja

Las gráficas sencillas revelan mucha información acerca de los datos de una muestra aleatoria simple pequeña o una muestra sistemática representativa. Las estimaciones de los histogramas o las densidades suavizadas muestran la forma de los datos; las gráficas de dispersión y las matrices de gráficas de dispersión muestran las relaciones entre las variables; otras gráficas, analizadas en Chambers *et al* (1983) y Cleveland (1994) enfatizan otras características de los datos. Sin embargo, en un diseño complejo de muestreo, una única gráfica no muestra toda la riqueza de los datos. Como vemos en la figura 7.4, las gráficas generalmente utilizadas para las muestras aleatorias simples pueden llevar a confusiones al aplicarse a los datos en bruto de las muestras no autoponderadas. La formación de conglomerados causa varias dificultades al graficar los datos de una encuesta compleja, como se observó en el ejemplo 5.6, pues la estructura de los conglomerados y los pesos (posiblemente distintos) deben exhibirse en las gráficas. Los problemas se complican debido a que los conjuntos de datos de las encuestas suelen ser muy grandes e implican varios niveles de los conglomerados.

Los datos se deben graficar con y sin pesos para observar su efecto. Además, los datos se deben graficar por separado para cada estrato y para cada unidad primaria, de ser posible, para examinar la variabilidad en las respuestas. Usted ya sabe cómo graficar los datos en bruto sin los pesos; en esta sección daremos algunos ejemplos de su incorporación a las gráficas.

EJEMPLO 7.4 La investigación de jóvenes bajo custodia de 1987 (Beck *et al.* 1988; Departamento de Justicia de Estados Unidos, 1989) extrajo una muestra de adolescentes y adultos jóvenes en instituciones juveniles de largo plazo operadas por el estado. A fines de 1987 se entrevistó sobre sus antecedentes familiares a los residentes de esas instalaciones, historia criminal previa y el uso de drogas y alcohol. Las variables elegidas para esta encuesta están en el archivo syc.dat.

Las instalaciones forman una unidad de conglomerado natural para una encuesta personal; el marco de muestreo de 206 instalaciones se construyó mediante el censo de muchachos en custodia de 1985 (CIC, por sus siglas en inglés). Las unidades primarias (las instalaciones) se dividieron en 16 estratos por la cantidad de residentes según el CIC de 1985. Cada una de las 11 instalaciones con 360 o más jóvenes formaba su propio estrato (estratos 6 a 16); cada una de éstas se incluyó en la muestra y se extrajo una submuestra de los residentes de las 11 instalaciones. En los estratos 1 a 5, se extrajo una muestra de las instalaciones con probabilidad proporcional al tamaño de las 195 instalaciones restantes; luego se extrajo una submuestra de los residentes con fracciones de muestreo predeterminadas. La tabla 7.3 contiene la información relativa a los estratos.

TABLA 7.3
Información de los estratos en la encuesta de jóvenes bajo custodia

Estrato	Tamaño CIC (cantidad de residentes)	Cantidad de unidades primarias en el marco	Cantidad de residentes en CIC	Cantidad de unidades primarias elegibles en la muestra
1	1–59	99	2881	11
2	60–119	39	3525	7
3	120–179	30	4355	7
4	180–239	13	2594	7
5	240–359	14	4129	7

Las fronteras de los estratos se eligieron de modo que las cantidades de residentes en cada estrato fuesen comparables entre sí. Originalmente se pretendía que cada residente tuviera una probabilidad, 1/8 de inclusión en la muestra, lo que produciría una muestra autoponderada con peso constante 8. Sin embargo, las instalaciones en los estratos 14 y 16 tuvieron un gran crecimiento entre 1985 y 1987, de modo que las fracciones de muestreo en esos estratos cambiaron a 1/11 y 1/12, respectivamente. En los estratos 1-5, los pesos varían de 5 a 15 (aproximadamente), dependiendo de la probabilidad de selección de la instalación y de la fracción de muestreo predeterminada en ella. Los pesos se ajustaron también debido a la ausencia de respuestas y para que los datos de la muestra se ajustasen a las cantidades del censo de 1987 para jóvenes en instalaciones estatales de retención a largo plazo. Después de todos los ajustes de pesos, éstos variaban de 5 (en el estrato 4) a 58 (para algunos adolescentes en estados que requerían autorización de los padres, por lo que tuvieron menor tasa de respuesta).

Veamos algunas gráficas de la edad de los residentes. Algunos jóvenes tienen más de 18 años, pues las instalaciones de California se incluyeron en la muestra. Como se pretendía que la muestra fuese aproximadamente autoponderada, el histograma de los datos sin pesos de la figura 7.6 y la función de masa de probabilidad empírica con pesos de la figura 7.7 tienen una forma similar en lo global, aunque bajo un análisis más cuidadoso aparecen algunas discrepancias: los pesos indican que los adolescentes de 15 años fueron un poco submuestreados debido a las probabilidades diferentes de selección y a la ausencia de respuestas, mientras que hubo un cierto exceso de muchachos de 17 años en la muestra.

Si sólo nos interesase la distribución de toda la población, podríamos concentrarnos en las gráficas del tipo de las figuras 7.6 y 7.7 y en otras gráficas similares con información relativa a las distribuciones univariadas como las gráficas de cuantiles contra cuantiles (vea Chambers *et al* 1983). Pero también quisiéramos explorar las diferencias de un estrato a otro con respecto de la distribución de las edades. La figura 7.8 incorpora los pesos en las gráficas de bloques de los datos; como la variable de respuesta (la edad) es discreta, podemos mostrar un poco más de detalle en cada estrato. La figura 7.9 muestra la suma de los pesos para cada edad dentro del estrato. La frecuencia relativa estimada de los jóvenes con tal edad en cada estrato se indica mediante un círculo cuya área es proporcional a la suma de los pesos.

FIGURA 7.6
Un histograma de todos los datos, sin incorporar los pesos. El histograma exhibe la distribución de las edades en la muestra.

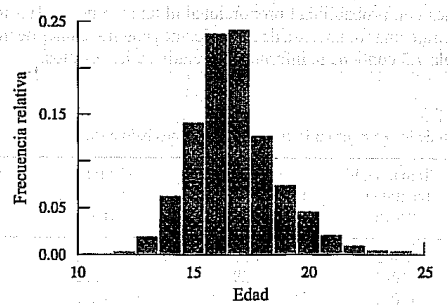


FIGURA 7.7
Una función de masa de probabilidad estimada para la edad, $\hat{f}(y)$. La forma es similar a la del histograma de los datos en bruto, pero existen relativamente más jóvenes de 15 años y relativamente menos de 17 años.

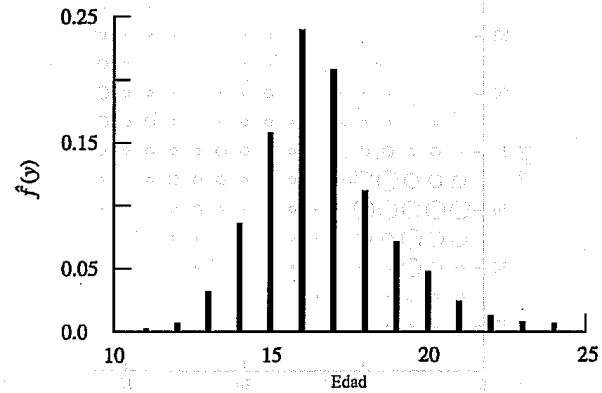
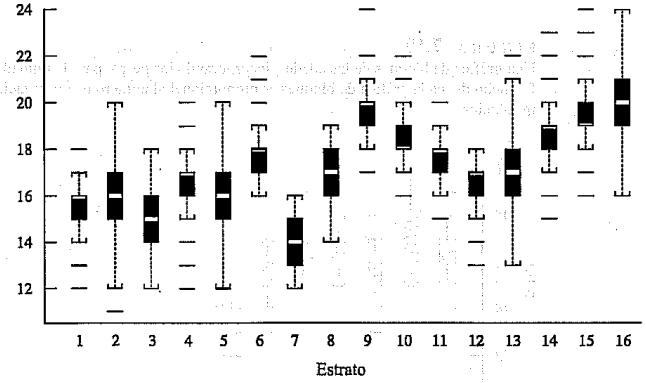


FIGURA 7.8
Una gráfica de bloques de las distribuciones de edad para cada estrato, incorporando los pesos. Observe la amplia variabilidad de un estrato a otro.



También podríamos estar interesados en la variabilidad de una instalación a otra. Las figuras 7.10 y 7.11 muestran gráficas similares para las unidades primarias del estrato 5. Estas gráficas pueden construirse para cada estrato para mostrar las diferencias en variabilidad entre los estratos. ■

FIGURA 7.9
La distribución de edades para cada estrato. El área de cada círculo es proporcional a la suma de los pesos para las observaciones muestrales en ese estrato y grupo de edades. La máxima cantidad de jóvenes menores de 18 años está en los estratos 1–5.

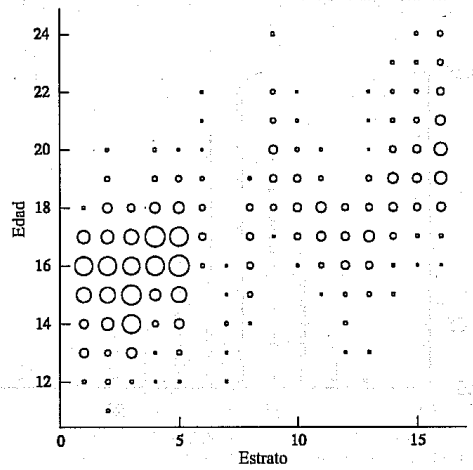


FIGURA 7.10
Una gráfica de bloques de las edades, incorporando los pesos, para la unidad primaria del estrato 5. El ancho de cada gráfica de bloques es proporcional al número de observaciones muestrales en esa instalación.

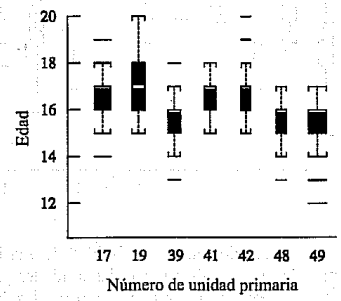
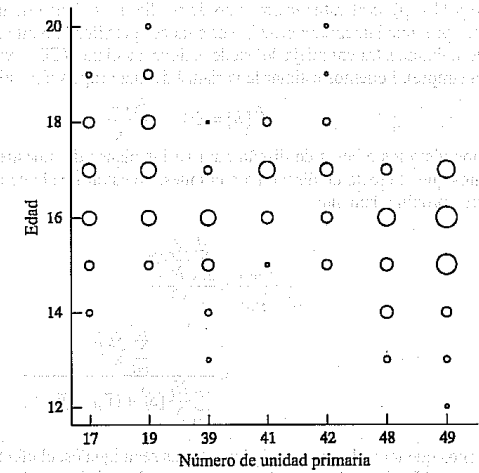


FIGURA 7.11
La distribución de edades para cada unidad primaria del estrato 5. El área de cada círculo es proporcional a la suma de los pesos para las observaciones muestrales en esa unidad primaria y grupo de edad.



7.5 Efectos del diseño

Cornfield (1951) sugirió medir la eficiencia de un plan de muestreo mediante el cociente de la varianza que se obtendría mediante una muestra aleatoria simple de k unidades de observación, entre la varianza obtenida mediante el plan de muestreo complejo con k unidades de observación. Kish (1965) llamó al recíproco del cociente de Cornfield **efecto de diseño** (ED) de un plan de muestreo y un estimador, y lo utilizó para resumir el efecto del diseño sobre la varianza de la estimación:

$$\text{ED(plan, estadística)} = \frac{V(\text{estimación del plan de muestreo})}{V(\text{estimación de una muestra aleatoria simple con el mismo número de unidades de observación})} \quad (7.5)$$

Para estimar una media a partir de una muestra con n unidades de observación,

$$\text{ED}(\text{plan}, \hat{\bar{y}}) = \frac{V(\hat{\bar{y}})}{\left(1-\frac{n}{N}\right)\frac{S^2}{n}}.$$

El efecto de diseño proporciona una media de la precisión ganada o perdida por el uso del diseño más complejo en vez de una muestra aleatoria simple. Aunque es un concepto útil, no es una forma de evitar el cálculo de las varianzas: Se requiere una estimación de la varianza a partir del diseño complejo para determinar el efecto de diseño. Por supuesto, distintas cantidades de la misma encuesta pueden tener distintos efectos de diseño. Kish

muestra la forma en que el efecto de diseño permite utilizar conocimientos anteriores para diseñar la encuesta.

Generalmente, la varianza de una muestra aleatoria simple es más fácil de obtener que $V(\hat{y})$: Al estimar una proporción, la varianza de una muestra aleatoria simple es aproximadamente $p\ (1-p)/n$; al estimar otro tipo de media, la varianza de una muestra aleatoria simple es aproximadamente S^2/n. Así, si se conoce aproximadamente el efecto de diseño, la varianza de la muestra compleja se puede estimar mediante (ED × varianza de la muestra aleatoria simple). Podemos estimar la varianza de una proporción estimada $\hat{p}$ mediante

$$\hat{V}[\hat{p}] = \text{ED} \times \frac{\hat{p}(1-\hat{p})}{n}.$$

Hemos visto los efectos de diseño para varios planes de muestreo. En la sección 4.4 mostramos que el efecto de diseño para el muestreo estratificado con distribución proporcional era aproximadamente

$$\frac{V_{\text{prop}}}{V_{\text{MAS}}} \approx \frac{\sum_{h=1}^{H} \frac{N_h}{N} S_h^2}{S^2} \approx \frac{\sum_{h=1}^{H} \frac{N_h}{N} S_h^2}{\sum_{h=1}^{H} \frac{N_h}{N} [S_h^2 + (\bar{y}_{hU} - \bar{y}_U)^2]}. \tag{7.6}$$

A menos que todas las medias de los estratos sean iguales, el efecto de diseño para una muestra estratificada por lo general será menor que 1; casi siempre la estratificación proporciona mayor precisión por unidad de observación que una muestra aleatoria simple.

También analizamos ampliamente los efectos de diseño en el muestreo por conglomerados, particularmente en la sección 5.2.2. De (5.9), el efecto de diseño para un muestreo por conglomerados de una etapa, cuando todas las unidades primarias tienen M unidades secundarias, es aproximadamente

$$1 + (M-1)\text{ICC}$$

El coeficiente de correlación dentro de la clase (ICC) generalmente es positivo, en el muestreo por conglomerados, de modo que por lo común el efecto de diseño es mayor que 1; las muestras por conglomerados comúnmente proporcionan menos precisión por unidad de observación que una muestra aleatoria simple.

Antes de calcular las varianzas de nuestra muestra, en las encuestas con estratificación y conglomerados, no podemos asegurar que el efecto de diseño para una cantidad dada será mayor o menor que 1. La estratificación tiende a aumentar la precisión y los conglomerados tienden a disminuirla, de modo que el efecto de diseño total dependerá de si se pierde más precisión con los conglomerados que la ganada con la estratificación.

EJEMPLO 7.5 Para el estudio de los mosquiteros del ejemplo 7.1, calculamos que el efecto de diseño para la proporción de camas con mosquiteros era 5.89. Esto significa que se necesitan aproximadamente seis veces más observaciones con el diseño de muestreo complejo utilizado en la encuesta para obtener la misma precisión que se lograría con una muestra aleatoria simple. El elevado efecto de diseño en esta encuesta se debe a la formación de los conglomerados: los poblados tienden a ser homogéneos en cuanto al uso de mosquiteros. Si se ignoran los conglomerados y se analiza la muestra como si fuese una muestra aleatoria simple, los errores estándar estimados serían demasiado pequeños y podría suponerse que se ha logrado más precisión de la real. ■

7.5.1 Efectos de diseño e intervalos de confianza

Si se conoce el efecto de diseño de cada estadística, podría emplearse junto con el software estándar para obtener intervalos de confianza para las medias y los totales. Si se extrae una muestra de n unidades de observación a partir de una población de N unidades de observación posibles y si $\hat{p}$ es la estimación de la proporción de interés en la encuesta, un intervalo de confianza aproximado de 95% para p es (suponiendo que la corrección para poblaciones finitas es cercana a 1):

$$\hat{p} \pm 1.96\sqrt{\text{ED}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{7.7}$$

Al estimar una media en vez de una proporción, si la muestra es lo bastante grande como para aplicar el teorema central del límite, un intervalo de confianza aproximado del 95% es

$$\hat{\bar{y}} \pm 1.96\sqrt{\text{ED}}\sqrt{\frac{\hat{S}^2}{n}},$$

donde $\hat{S}^2$ se puede calcular mediante (7.4).

Kish (1995) y otros autores utilizan ahora el término efecto de diseño de Tukey[3] para referirse a la cantidad

$$\text{EDT}(\bar{y}) = \frac{\text{EE}(\bar{y}_{plan})}{\frac{s}{\sqrt{n}}},$$

de modo que EDT será un factor adecuado para un error estándar de la mitad del ancho del intervalo de confianza. En la práctica, como señala Kish, la elección entre ED o EDT tiene poca importancia, pero debemos ser cuidadosos en la definición utilizada en una encuesta dada.

7.5.2 Efectos de diseño y tamaños de muestra

Los efectos de diseño son muy útiles para estimar el tamaño de muestra necesario para una encuesta, razón por la que Cornfield (1951) los introdujo, para estimar el tamaño de muestra necesario si la unidad de muestreo en una encuesta para estimar la frecuencia de la tuberculosis era un opúsculo o un bloque de censo, en vez de un individuo. Se especificó que el error máximo permisible era de 20% de la frecuencia real, o $0.2 \times p$. Si la frecuencia de la tuberculosis era $p = 0.01$, el tamaño de la muestra para una muestra aleatoria simple debería ser

$$n = \frac{1.96^2 p(1-p)}{(0.2p)^2} = 9508.$$

Cornfield recomendó aumentar el tamaño de muestra de la muestra aleatoria simple a 20,000, para tener mayor precisión en las estimaciones por separado de las subpoblaciones. Estimó el efecto de diseño para el muestreo de los opúsculos del censo en vez de los individuos en 7.4 y concluyó que, si se utilizaban, con un promedio de 4600 individuos cada uno, como unidad de muestreo, se necesitaría un tamaño de muestra de 148,000 y no de 20,000 adultos.

Si conoce el efecto de diseño para una encuesta similar, deberá estimar sólo el tamaño de muestra necesario con una muestra aleatoria simple. Luego, multiplique el tamaño de

[3] El término se debe a Turkey (1968).

muestra por ED para obtener la cantidad de unidades que deberán observarse con el diseño complejo. Para los fines del tamaño de muestra, tal vez querrá emplear efectos de diseño por separado para cada estrato.

7.6 La encuesta nacional de víctimas de delitos

La mayoría de las estadísticas criminales de los periódicos estadounidenses provienen de los informes de delitos (UCR) compilados por la FBI a partir de informes proporcionados por las oficinas policiacas. Sin embargo, los UCR subestiman la cantidad de crímenes en Estados Unidos, principalmente porque no todos los delitos son informados a la policía.

La encuesta nacional de víctimas de delitos (NCVS)[4] es una gran investigación nacional administrada por la oficina de estadísticas de justicia, con entrevistas realizadas por la oficina de censos. Como la encuesta de la población actual (CPS), la NCVS sigue un diseño estratificado con conglomerados y varias etapas. La información acerca del diseño de la CPS aparece en Hanson (1978) y McGuiness (1994); la información adicional acerca de la NCVS está en la documentación publicada por la oficina de estadísticas judiciales. La NCVS encuesta a familias en todo el territorio de Estados Unidos y pregunta a los miembros de la familia con 12 o más años de edad acerca de sus experiencias como víctimas de delitos durante los últimos seis meses. La NCVS es la única fuente nacional de información acerca de las víctimas de delitos.

La CPS y la NCVS alguna vez utilizaron diseños similares; de hecho, el diseño de la NCVS iniciado en 1970 utilizó un subconjunto de las unidades primarias de muestreo de la CPS para ahorrar gastos de administración y de entrevistas. Para los diseños de muestra de la NCVS en 1980 y 1990, el traslape con el diseño anterior NCVS se maximizó en la medida de lo posible. Aquí describiremos el diseño iniciado en 1980, utilizado para producir las estimaciones de la NCVS para 1990; las características básicas del diseño son las mismas que para la NCVS posterior a 1990. Regresaremos a la NCVS en el capítulo 8, para mostrar la forma en que se ajustaron los pesos para la ausencia de respuestas y la subcobertura en esta encuesta grande y compleja.

Una unidad primaria en la NCVS es un condado, un grupo de condados adyacentes o un área estadística metropolitana (AEM). Un AEM es una gran ciudad, junto con las comunidades adyacentes, económica y socialmente integradas a ella. Algunos ejemplos son el AEM de Montgomery, Alabama (incluida en los diseños de muestra de 1980 y 1990), que incluye los condados de Autauga, Elmore y Montgomery; el AEM de Columbus, Ohio, con los condados Delaware, Fairfield, Franklin, Madison, Pickaway y Union, y el AEM de Albany-Schenectady-Troy, en Nueva York, con los condados de Albany, Greene, Montgomery, Rensselaer, Saratoga y Schenectady.

Cualquier unidad primaria con una población aproximada de 550,000 o más habitantes (de acuerdo con el censo de 1980) se incluye automáticamente en la muestra. Se dice que tal unidad primaria es *autorrepresentada* (AR), pues no representa ninguna unidad primaria distinta de sí misma. La probabilidad de seleccionar esta unidad primaria es 1.

Las demás unidades primarias quedan agrupadas en estratos, de modo que cada grupo del estrato tenga una población cercana a 650,000. En la NCVS, las unidades primarias se agrupan en estratos con base en la ubicación geográfica, la información demográfica disponible mediante el censo de 1980 y las tasas de delitos de los UCR. Se elige una unidad

[4] La encuesta se llamaba Encuesta Nacional de Delitos. Utilizamos NCVS para referirnos a ambos nombres.

primaria de cada uno de estos estratos, con una probabilidad proporcional al tamaño de la población; esta unidad primaria se denomina *no autorrepresentada* (NAR), pues se supone que no sólo se representa a sí misma, sino a todas las unidades primarias de ese estrato. Dentro de un estrato, una unidad primaria con una población de 100,000 tiene dos veces más probabilidad de ser elegida para la muestra que una con población de 50,000. Para la NCVS de 1990, había 84 unidades primarias AR y 153 unidades primarias NAR. Como las tasas de víctimas varían de una región a otra, un mayor número de estratos en la NCVS aumenta la precisión de las estimaciones.

La segunda etapa del muestreo implica la selección de los distritos de enumeración (DE), áreas geográficas utilizadas en el censo del decenio de 1980; por lo general, un DE contiene entre 300 y 400 familias, pero la población y el área de los DE varían considerablemente.[5] Los DE se eligen con probabilidad proporcional al tamaño de su población en 1980; la cantidad de DE seleccionados dentro de una unidad primaria queda determinada de modo que la muestra de DE sea aproximadamente autoponderada. En el listado del censo, los DE se ordenan según su ubicación geográfica y se eligen mediante el muestreo sistemático, como se describe en la sección 6.2, de modo que los DE de la muestra se distribuyen geográficamente en la unidad primaria seleccionada. Si la tasa global de muestreo es $1/x$, en las unidades primarias AR el intervalo de muestreo es x. Si se utilizan los registros del censo para el marco de muestreo, las direcciones se numeran desde 1 hasta el número de familias en la unidad primaria. Se elige un número aleatorio entre 1 y k y los DE elegidos para estar en la muestra son aquellos contenidos en las direcciones k, $k + x$, $k + 2x$, y así sucesivamente. En las unidades primarias NAR, el intervalo de muestreo es (probabilidad de elegir la unidad primaria) $\times$ (x).

En la tercera etapa del muestreo, cada DE seleccionado se divide en conglomerados de aproximadamente cuatro hogares cada uno. El censo enumera los hogares dentro de un DE en orden geográfico y, en la medida de lo posible, se utilizó esa lista. Se extrae una muestra de esos conglomerados y la muestra incluye cada hogar de un conglomerado seleccionado o aproximadamente cuatro hogares. Para la encuesta se entrevista a todas las personas con 12 o más años de edad de cada hogar.

En algunas regiones se empleó el *muestreo por área*. Si la lista de hogares del censo fuese la única utilizada durante esa década, habría una subcobertura esencial de la población, pues no incluiría los hogares de reciente creación. Para incluir estos hogares, la NCVS utiliza una muestra de permisos de construcción de unidades residenciales y extrae una muestra de ellas. En el muestreo por área, un investigador de campo enumera todos los hogares o manzanas habitadas dentro de un área seleccionada en un DE, y esa lista sirve como marco de muestreo para el área.

En resumen, las etapas para la NCVS de 1990 aparecen en la tabla 7.4.

Las entrevistas para la NCVS, con las personas de 12 o más años se realizan cada mes, y abarcan los hogares seleccionados para la muestra en un periodo de seis meses; lo que permite distribuir uniformemente durante el año la carga de trabajo de las entrevistas. Para un análisis longitudinal de los datos y una relativa seguridad de que los delitos reportados en el periodo de seis meses realmente ocurrieron en esos meses y no anteriormente, se entrevista cada seis meses a los residentes de cada hogar durante un periodo de tres años, para un total de siete entrevistas. La primera entrevista no se utiliza para estimar la proporción de víctimas, sino sólo como cuota que establece un marco de tiempo para informar de las víctimas, de modo que una víctima reportada en dos periodos sucesivos de entrevistas sólo se cuenta una vez. El hecho de convertirse en víctima es una experiencia que queda muy grabada en la memoria de la mayoría de las personas (de hecho, tanto que es fácil creer que el suceso ocurría más recientemente de cuando sucedió, y se transporta al periodo de refe-

[5] Para el censo de 1990, los DE recibieron el nombre de áreas de registro de direcciones.

TABLA 7.4
Etapas de muestreo para la NCVS de 1990

Etapa	Unidad de muestreo	Estratificación
1	Unidad primaria (condado, conjunto de condados adyacentes o un AEM)	Ubicación, información demográfica y características relacionadas con el delito
2	Distrito de enumeración	
3	Conglomerados de cuatro hogares	
4	Hogar	
5	Persona dentro del hogar	

rencia de seis meses. El uso de un panel permite a la oficina de estadísticas acotar cada entrevista por la anterior; se pregunta detalladamente a una persona si cree que el incidente se repitió desde la última entrevista.

Para 1990, cerca de 62 600 hogares (incluyendo las manzanas) estaban en la muestra. De éstos, 56 800 recibieron el cuestionario principal (los ocupantes de los demás hogares recibieron un nuevo cuestionario por fases). Cerca de 8 200 de los 56 800 hogares seleccionados no fueron elegibles para la NCVS, pues estaban vacíos, fueron demolidos o ya no eran utilizados como lugares de residencia. Sin embargo, en cerca de 1600 de los hogares no se efectuaron entrevistas, pues no se localizó a los residentes o se negaron a participar en la encuesta. La NCVS de 1990 tuvo una tasa de ausencia de respuesta de 1600/48600, cerca de 3.3%. En total, cerca de 95 000 personas respondieron al cuestionario.

Es claro que se trata de un diseño complejo de encuesta; se utilizaron los pesos para calcular estimaciones de las tasas de víctimas y la cantidad total de delitos. La encuesta está diseñada para ser (aproximadamente) autoponderada, de modo que al principio cada individuo recibe el mismo peso base de (1/probabilidad de selección del hogar). Para la NCVS a fines de 1980, cada persona representa aproximadamente 1658 habitantes de Estados Unidos, de modo que el peso base es 1658.

Aunque la NCVS está diseñada para ser autoponderada, a veces un conglomerado elegido dentro de un DE tiene más hogares que lo supuesto. Por ejemplo, un edificio de departamentos podría haberse erigido en lugar de varias casas. En ese caso, sólo se entrevistó a las personas de una submuestra del conglomerado. Si se utilizó el submuestreo, las unidades de la submuestra reciben un factor de control del peso (FCP). Por ejemplo, si sólo se incluye en la muestra la tercera parte, las unidades de la muestra reciben un FCP de 3, pues representan al triple de unidades. Si un hogar está en un conglomerado donde no se necesita el submuestreo, se le asigna un FCP de 1. En este nivel, un hogar de la muestra representa

$$\text{peso base} \times \text{FCP}$$

hogares de la población. Éste es el peso de muestreo para un hogar en la muestra de la NCVS; como la encuesta pretende entrevistar a todas las personas con 12 o más años en las casas de la muestra, el peso de muestreo de una persona de la muestra es igual al peso para el hogar en cuestión.

Los demás ajustes de peso en la NCVS corresponden a la ausencia de respuesta o se utilizan posteriormente en la estratificación. No se entrevistó a algunas personas elegidas para estar en la muestra por estar ausentes o rehusarse a participar. El entrevistador reúne la información demográfica acerca de las personas que no contestaron y la emplea para ajustar los pesos, como un intento por contrarrestar la ausencia de respuesta. (Éste es un ejemplo de ajuste de las clases de peso contra la ausencia de respuesta, analizado en la sección 8.5.) Se utilizaron dos ajustes de peso para la ausencia de respuesta: el factor de ajuste para la falta de entrevistas dentro de los hogares (FADH) y el factor de ajuste para la falta de entrevistas de cada hogar (FAH). En cada factor de ajuste, el objetivo es aumentar los pesos de las unidades entrevistadas más similares a las unidades que no pudieron entrevistarse.

El FADH se utiliza para compensar las personas que no respondieron la encuesta en los hogares donde al menos un miembro sí lo hizo. Se calcula por separado para cada una de las regiones de Estados Unidos (noreste, este medio, sur y oeste). Dentro de cada región, las personas de los hogares donde contestó al menos una persona se clasifican en 24 celdas, usando la raza de la persona designada como persona de referencia, la edad y sexo del miembro de la familia que no contestó y la relación de la persona que no contestó con la persona de referencia. Cualquiera de las 24 celdas que contuviese menos de 30 casos entrevistados o que produzca un FADH mayor o igual a 2 se combinan con las celdas similares; la fusión de las celdas evita que ciertos individuos tengan pesos muy grandes. Entonces,

$$\text{FADH} = \frac{\text{suma de pesos de todas las personas en la celda}}{\text{suma de pesos de todas las personas entrevistadas en la celda}}$$

Los pesos utilizados para calcular el FADH son los asignados hasta este momento en el procedimiento de pesos (es decir, peso base × FCP). Así, los pesos de quienes contestaron se incrementan de modo que representen a quienes no contestaron y a las personas de la población que serían representadas por quienes no contestaron, además de su representación original. Después de aplicar el FADH, el peso de un individuo es

$$\text{peso base} \times \text{FCP} \times \text{FADH}.$$

Parte de la ausencia de respuestas se debe a individuos que no responden, en hogares que sí; otras ausencias se deben a que toda la familia no responde. Cerca de 3% a 4% de los hogares elegibles para la encuesta no se alegaron o se rehusaron a responder; el FAH se utiliza para tratar de compensar la ausencia de respuestas a nivel de los hogares. Para el FAH, los hogares se agrupan en celdas mediante el nivel del AEM, urbano o rural, y la raza de la persona de referencia. Entonces

$$\text{FAH} = \frac{\text{suma de pesos de todas las personas en la celda}}{\text{suma de pesos de todas las personas entrevistadas en la celda}}$$

Como en el caso del FADH, los pesos utilizados para calcular el FAH son los empleados hasta ahora: peso base × FCP × FADH. Las celdas se fusionan hasta que el FAH sea menor que 2.

En este punto de la construcción de los pesos, el peso asignado a un individuo es

$$\text{peso base} \times \text{FCP} \times \text{FADH} \times \text{FAH}.$$

Los pesos de muestreo para los individuos que responden se incrementan, de modo que representen también a quienes no respondieron y que son similares desde el punto de vista demográfico.

Como la NCVS es una muestra, su información demográfica difiere por lo general de la información de la población de Estados Unidos como un todo. Se usan dos etapas de la

estimación de las proporciones para ajustar los valores de la muestra, de modo que coincidan mejor con la información actualizada de los censos. Se espera que este ajuste reduzca la varianza de las estimaciones de las tasas de víctimas.

La primera etapa de la estimación de proporciones se usa sólo en las unidades primarias no autorrepresentadas y pretende reducir la variabilidad que surge del uso de una unidad primaria para representar al estrato. La estimación de proporciones se utiliza para asignar distintos pesos a las celdas, estratificadas por región, nivel de AEM y raza. El factor en la primera etapa,

$$\text{FPE} = \frac{\text{conteo independiente de la cantidad de personas en la celda}}{\text{estimación muestral (suma de pesos) de la cantidad de personas en la celda}}$$

ajusta las diferencias entre las características censales de las unidades primarias no autorrepresentadas de la muestra y las características de todo el conjunto de unidades primarias no autorrepresentadas. El FPE es igual a 1 para las unidades primarias autorrepresentadas y se trunca en 1.3 para las unidades primarias no autorrepresentadas.

El factor de la segunda etapa (FSE) de la estimación de proporciones se aplica a todos los elementos de la muestra. Las personas se clasifican en 72 grupos, según su edad, raza y sexo. Las celdas deben tener una cantidad de al menos 30 personas entrevistadas y el FSE debe estar entre 0.5 y 2.0; las celdas se fusionan hasta cumplir estas condiciones.

$$\text{FSE} = \frac{\text{conteo independiente de la cantidad de personas en la celda}}{\text{estimación muestral (suma de pesos) de la cantidad de personas en la celda}}$$

El FSE es una forma de estratificación posterior: pretende ajustar la distribución muestral de edad, raza o sexo de modo que la clasificación cruzada coincida con los conteos independientes, que pretenden ser más precisos. Si la suma de pesos de las mujeres blancas ancianas de la muestra es mayor que la "mejor" estimación actual del número de mujeres blancas ancianas en la población, obtenida mediante la información actualizada de un censo, entonces el FSE será menor que 1 para todas las mujeres blancas en la muestra.

Después de todos los ajustes, el peso final de la persona i es

$$w_i = \text{peso base} \times \text{FCP} \times \text{FADH} \times \text{FAH} \times \text{FPE} \times \text{FSE}.$$

El peso w_i se utiliza como si en realidad hubiese w_i personas en la población exactamente iguales a aquella persona a la que se asignó el peso. En la NCVS de 1990, los pesos de las personas varían de 1100 a 9000, con la mayoría de los pesos entre 1500 y 2500. La figura 7.12 muestra los diagramas de bloques de los pesos de las personas entrevistadas entre julio y diciembre de 1990. Los pesos se incluyen en las cintas de uso público de la NCVS; para utilizarlos y estimar la cantidad total de asaltos con agravantes reportados por las mujeres blancas, defina

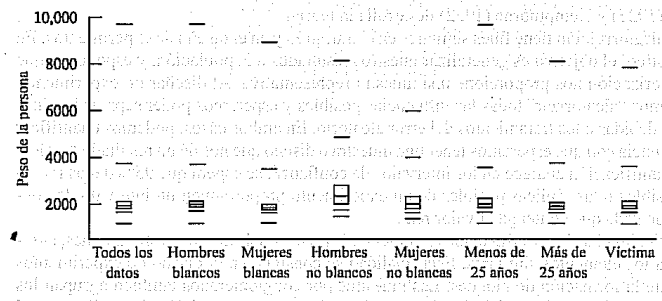
$$y_i = \begin{cases} k & \text{si la persona } i \text{ es una mujer blanca que ha informado de } k \text{ asaltos con agravantes} \\ 0 & \text{en caso contrario} \end{cases}$$

y use $\sum_{i \in S} w_i y_i$ como estimación.

Aunque la ausencia de respuesta es relativamente baja en la NCVS, los pesos establecen una diferencia al calcular las tasas de víctimas. Por lo general, las estimaciones de estas tasas son mayores al utilizar los pesos que cuando no se utilizan. Los hombres jóvenes negros que respondieron a la encuesta tienen una probabilidad desproporcionada de ser víctimas de un delito y la subcobertura y ausencia de respuestas entre los negros es alta.

Como el diseño de muestreo y el esquema de pesos son tan complicados en la NCVS, la determinación de los efectos de diseño requiere mucho esfuerzo. Ahora, calculamos las

FIGURA 7.12
Gráfica de bloques de los pesos para la NCVS de 1990, para todas las personas, los hombres blancos, las mujeres blancas, los hombres no blancos, las mujeres no blancas, las personas con menos de 25 años, las personas con más de 25 años y las víctimas de delitos violentos. Las líneas horizontales representan el máximo, los percentiles 95, 75, la mediana, los percentiles 25 y 5 y el mínimo. Observe que los pesos son mucho mayores para los hombres no blancos, lo que indica una mayor ausencia de respuesta y subcobertura en ese grupo.



varianzas mediante métodos de réplica, como los descritos en el capítulo 9. El diseño de muestreo afecta las estimaciones de la varianza en varios niveles distintos:

1. En los estratos no autorrepresentados, sólo se elige una unidad primaria entre varias unidades del estrato, de modo que existe una varianza entre las unidades primarias en ese estrato.
2. Dentro de un distrito de enumeración, se selecciona un conglomerado de aproximadamente cuatro hogares para estar en la muestra. Es probable que estos hogares tengan una correlación positiva.
3. Todas las personas dentro de los hogares de la muestra fueron entrevistadas, lo que da un efecto de conglomerado para las personas.
4. El muestreo sistemático se utiliza para elegir los distritos de enumeración en lugar del muestreo aleatorio simple. El efecto del muestreo sistemático sobre la varianza es difícil de determinar, aunque se supone que, con frecuencia, el muestreo sistemático produce una menor varianza con respecto del muestreo aleatorio simple, pues las unidades muestrales en una muestra sistemática son obligadas a dispersarse en el marco de muestreo.

Los ajustes de peso, en especial el FSE, también afectan la varianza de las estimaciones. Se cree que el FSE disminuye la varianza de las estimaciones, como sería de esperar, pues el ajuste es en realidad una forma de estratificación posterior. El efecto de diseño general para la NCVS y para encuestas oficiales similares en Estados Unidos, es cercano a 2.

7.7 Muestreo y diseño de experimentos*

Fienberg y Tanur (1987) y Yates (1981) analizan varias similitudes entre las encuestas muestrales y los experimentos diseñados. En esta sección anotamos algunos de ellos.

El muestreo aleatorio simple, donde el universo $\mathcal{U}$ tiene N unidades, es similar al enfoque de aleatorización de la comparación de dos tratamientos con un total de N unidades experimentales. Para probar la hipótesis H_0: $\mu_1 = \mu_2$, asignamos al azar n de las N unidades al primer tratamiento y las demás $N-n$ unidades al tratamiento 2. El valor observado de la estadística de prueba se compara con la distribución de referencia, basada en las $\binom{N}{n}$ asignaciones posibles de las unidades experimentales a los tratamientos. El valor p proviene de la distribución de aleatorización. El uso de la aleatorización para la inferencia aparece ya en Fisher (1925) y Kempthorne (1952) desarrolla la teoría.

La aleatorización tiene fines similares en el muestreo y en el diseño de experimentos. En el muestreo, el objetivo es generalizar nuestros resultados a la población, y esperamos que la aleatorización nos proporcione una muestra representativa. Al diseñar un experimento, intentamos "aleatorizar" todas las influencias posibles y esperamos poder separar las diferencias debidas a los tratamientos del error aleatorio. En ambos casos, podemos cuantificar la frecuencia con que esperamos tener una muestra o diseño que nos dé un resultado "malo". Esta cuantificación aparece en los intervalos de confianza: se espera que 95% de las muestras posibles o las réplicas posibles de un experimento proporcionen un intervalo de confianza de 95% que contenga el valor real.

La finalidad de la estratificación es aumentar la precisión de nuestras estimaciones, agrupando a los elementos similares. Esta finalidad se consigue en el diseño de experimentos mediante la formación de bloques. Las muestras por conglomerados también agrupan los elementos similares, pero la finalidad es la conveniencia, no la precisión. Un análogo en el diseño de experimentos sería el diseño por fragmentos, que generalmente proporciona una mayor precisión en la estimación de los subelementos que en la estimación de todos los elementos.

La analogía estructural entre las encuestas y el diseño de experimentos fue aprovechada usando las tablas de análisis de varianza para desarrollar la teoría de estratificación y el muestreo por conglomerados. Utilizamos un análisis de varianza de un sentido, con efectos fijos, para un enfoque de la estratificación basado en el modelo y un análisis de varianza de un sentido con efectos aleatorios para un enfoque del muestreo por conglomerados, basado en el modelo. Gran parte de la teoría del muestreo por conglomerados es similar a la teoría de los modelos con efectos aleatorios. En los modelos de los capítulos 5 y 6 nos basamos en los componentes de la varianza para explicar la dependencia de los datos.

La estratificación posterior y la estimación de proporciones y de regresión en el muestreo nos permite aumentar la precisión de nuestras estimaciones aprovechando la relación entre la variable de interés y otras variables de clasificación; este mismo objetivo se logra en el diseño de experimentos utilizando el ajuste covariado, como en el análisis de la covarianza.

El diseño de experimentos y el muestreo están implicados en debates similares entre el uso del enfoque de la teoría de aleatorización o el uso de un enfoque basado en el modelo. En las secciones 2.8, 3.4, 4.6, 5.7 y 6.7 hemos expuesto algunos aspectos de la estimación de funciones de totales, pero hay mucho más al respecto. El lector interesado en el tema puede comenzar con los artículos de Smith (1994) y Hansen *et al* (1983), y el libro de Thompson (1997). Royall (1992a) resume un enfoque del muestreo basado en el modelo.

Por último, en las encuestas muestrales y el diseño de experimentos, es crucial realizar un esfuerzo adecuado al diseñar el estudio. Ninguna cantidad de análisis estadísticos, por más sofisticados que sean, puede compensar un diseño pobre. El capítulo 1 presentó algunos ejemplos de resultados desastrosos debidos al sesgo de selección resultante del diseño o ejecución pobre de la encuesta. Una encuesta telefónica no sólo es inútil para generalizarse a toda una población, sino que es dañina, pues las personas podrían creer que las estadísticas son precisas. De manera análoga, poco puede concluirse acerca de la eficacia de los tratamientos A y B para una condición médica si la mayoría de los pacientes enfermos se escogen para el tratamiento A: Si la duración media de los síntomas es significativamente menor para el tratamiento B que para el tratamiento A, ¿tal diferencia se debe al tratamiento o a diferencias entre los pacientes?

Por supuesto, a veces es posible ajustar un diseño imperfecto durante el análisis. Si se dispone de una medida de la severidad de la enfermedad al inicio del estudio, podría utilizarse como covariante al comparar los dos tratamientos, aunque subsistiría la preocupación por confundirla con otras cantidades no medidas. Los valores de las celdas faltantes en un diseño de análisis de varianza de dos sentidos pueden estimarse mediante un modelo. Del mismo modo, la información disponible acerca de las personas que no contestan una encuesta se puede usar para mejorar la estimación cuando haya ausencia de respuestas, como veremos en el siguiente capítulo.

7.8 Ejercicios

1 Consiga uno de los artículos enumerados a continuación o algún otro que utilice un diseño complejo de encuestas y redacte un breve ensayo que incluya:

a Un breve resumen del diseño y el análisis.

b El análisis de la efectividad del diseño y lo adecuado del análisis.

c Sus recomendaciones para estudios futuros de ese tipo.

Stewart, R.E., y H, A. Kantrud. 1973. Ecological distribution of breeding waterfowl populations in North Dakota. *Journal of Wildlife Management* 37 (1):39–50.

Matson, R.G., y W. D. Lipe. 1975. Regional sampling: A case study of Cedar Mesa, Utah. En *Sampling in archaeology*, 124-143. Editado por J. Mueller. Tucson: University of Arizona Press.

U.S. Veterans Administration. 1980. *Study of former prisoners of war*. Washington, D.C.: Government Printing Office. El diseño de muestreo se analiza en las páginas 16–21.

Carra, J.S. 1984. Lead levels in blood of children around smelter sites in Dallas. In *Environmental sampling for hazardous wastes*. Editado por E. G. Schweitzer y J. A. Santolucito, ACS Symposium Series 267. Washington, D.C.: American Chemical Society.

Gerbert, B., B. T. Maguire, y T.J. Coates. 1990. Are patients talking to their physicians about AIDS? *American Journal of Public Health* 80:467-468.

Langley, G. R., D.L. Tritchler, H. A. Llewellyn-Thomas, y J. E. Till. 1991. Use of written cases to study factors associated with regional variations in referral rates. *Journal of Clinical Epidemiology* 44(4/5):391-402.

Oppliger, R. A., G.L. Landry, S. W. Foster, y A.C. Lambrecht. 1993. Bulimic behaviors among interscholastic wrestlers: A statewide survey. *Pediatrics* 91 (4): 826-831.

Tanfer, K. 1993. National Survey of Men: Design and execution. *Family Planning Perspectives* 25:83-86.

Wadsworth, J., J. Field, A.M. Johnson, S. Bradshaw, y K. Wellings. 1993. Methodology of the National Survey of Sexual Attitudes and Lifestyles. *Journal of the Royal Statistical Society, Ser. A*, 156:407-421.

Benson, V., y M.A. Marano. 1994. Current estimates from the National Health Interview Survey. *Vital and Health Statistics* 10 (189). El diseño de muestreo se describe en el apéndice I, a partir de la página 132.

Guyon, A.B, A. Barman, J. U. Ahmed, A.U. Ahmed, y M.S. Alam. 1994. A baseline survey on use of drugs at the primary health care level in Bangladesh. *Bulletin of the World Health Organization* 72 (2): 265-271.

Heneman, H. G., D. L. Huett, R. J. Lavigna, y D. Ogsten. 1995. Assessing managers' satisfaction with staffing services. *Personnel Psychology*. 48:163-172.

Kellermann, A. L., L. Westphal, L. Fischer, y B. Harvard. 1995. Weapon involvement in home invasion crimes. *Journal of the American Medical Association* 273 (22): 1759-1762.

Tielsch, J. M., J. Katz, H. A. Quigley, J. C. Javitt, y A. Sommer. 1995. Diabetes, intraocular pressure, and primary open-angle glaucoma in the Baltimore Eye Survey. *Ophthalmology* 102 (1):48-54.

2 Muchas oficinas gubernamentales de estadística y otras organizaciones que reúnen datos de encuestas tienen ahora sitios en Internet, donde proporcionan información acerca del diseño de encuestas. En la tabla 7.5 aparecen algunas direcciones de Internet (sujetas a cambios, aunque usted podrá encontrar la organización mediante una búsqueda). El primer sitio de la lista, www.fedstats.gov, proporciona vínculos con las agencias gubernamentales de Estados Unidos que gastan al menos 500,000 dólares anuales en actividades estadísticas. Muchas de las oficinas realizan encuestas. El sitio www.lib.umich.edu/libhome/Docu-ments.center/stats.html proporciona vínculos hacia información relativa a encuestas sobre varios temas, desde las finanzas hasta la agricultura.

TABLA 7.5
Sitios de Internet con información sobre encuetas de gran tamaño

Organización	Dirección
Consejo Federal de Política Estadística	www.fedstats.gov
Oficina de Censos de Estados Unidos	www.census.gov
Estadísticas de Canadá	www.statcan.ca
Estadísticas de Noruega	www.ssb.no
Estadísticas de Suecia	www.scb.se
Oficina de Estadísticas Nacionales del Reino Unido	www.ons.gov.uk
Oficina de Estadísticas de Australia	wwwstatistics.gov.au
Estadísticas de Nueva Zelanda	www.stats.govt.nz
Estadísticas de los Países Bajos	www.cbs.nl
Organización Gallup	www.gallup.com
Investigación de medios Nielsen	www.nielsenmedia.com
Centro de investigación de la opinión pública (NORC)	www.norc.uchicago.edu
Consorcio interuniversitario para la investigación social y política (ICPSR)	www.icpsr.umich.edu

Revise un sitio en Internet que describa una encuesta compleja. Escriba un resumen con los objetivos, el diseño y el método utilizado en el diseño. ¿Cree que se podría mejorar el diseño? En caso afirmativo, ¿cómo?

3 Se le ha pedido que diseñe una encuesta para estimar la cantidad total de autos sin permiso que se estacionan en los lugares de estacionamiento para discapacitados en el campus. ¿Qué variables (si existen) tomaría en cuenta para la estratificación? ¿Para la formación de conglomerados? ¿Cuál sería la información necesaria para apoyar el diseño de la encuesta? Describa un diseño de encuesta que piense que funcione en esta situación.

4 Repita el ejercicio 3 para una encuesta que estime la cantidad de libros de una biblioteca que necesiten encuadernarse nuevamente.

5 Repita el ejercicio 3 para una encuesta que estime el porcentaje de personas de su ciudad que hablen más de un idioma.

6 Repita el ejercicio 3 para una encuesta que estime la distribución de los huevos puestos por los gansos canadienses.

7 Demuestre que, en una muestra estratificada, $\sum yf(y)$ produce el estimador en (4.2).

8 ¿Cuál es el valor de $\hat{S}^2$ en (7.4) para una muestra aleatoria simple? ¿Cuál es su relación con la varianza muestral s^2?

9 En una muestra por conglomerados de dos etapas de las áreas rurales y urbanas en Nepal, Rothenberg *et al* (1985) determinaron que el efecto de diseño para las enfermedades contagiosas comunes era mucho mayor que para las enfermedades contagiosas raras. En las áreas urbanas, el sarampión, con una incidencia estimada de 123.9 casos por cada 1000 niños por año, tenía un efecto de diseño de 7.8; la difteria, con una incidencia estimada de 2.1 casos por cada 1000 niños por año, tenía un efecto de diseño de 1.9.

Explique por qué sería de esperar tal disparidad entre los efectos de diseño. SUGERENCIA: Suponga que extrae una muestra de 1000 niños, en 50 conglomerados de 20 niños cada uno. Suponga además que la enfermedad está concentrada lo más posible, de modo que si la incidencia estimada fuese de 40 por cada 1000, todos los niños de dos conglomerados tendrían la enfermedad, y ningún niño de los otros 38 conglomerados tendría la enfermedad. Ahora calcule ED para las incidencias que van de 1 por cada 1000 hasta 200 por cada 1000.

10 Use los datos del archivo nybight.dat (vea el ejercicio 19 del capítulo 4) para determinar la función de masa de probabilidad empírica del número de especies capturadas por cada red arrojada en 1974. Asegúrese de utilizar los pesos de muestreo.

11 Use los datos del archivo teachers.dat (vea el ejercicio 16 del capítulo 5) y los pesos de muestreo para determinar la función de masa de probabilidad empírica del número de horas trabajadas. ¿Cuál es el efecto de diseño?

12 Use los datos del archivo measles.dat (vea el ejercicio 17 del capítulo 5). ¿Cuál es el efecto de diseño para el porcentaje de padres que regresaron una forma de consentimiento? ¿Para el porcentaje de niños que tuvieron sarampión con anterioridad?

13 La encuesta de jóvenes bajo custodia extrajo una muestra de los jóvenes residentes en instalaciones preventivas a largo plazo al final de 1987. ¿Es esta muestra representativa de los jóvenes que estuvieron confinados a largo plazo en 1987? ¿Por qué?

14 El archivo syc.dat contiene más información acerca de la encuesta de jóvenes en custodia de 1987. Grafique los datos de la edad de los jóvenes al momento de su primer arresto. ¿Cuál es la edad promedio en el primer arresto? ¿La mediana? ¿El percentil 25? (Use el "peso final" para estimar estas cantidades. *No* calcule los errores estándar por el momento.) ¿Cómo se comparan sus estimaciones con las estimaciones obtenidas sin pesos?

15 Use el archivo syc.dat y los pesos finales para estimar la proporción de jóvenes que

a Tienen 14 años o menos.

b Fueron recluidos por una acción violenta.

c Crecieron con sus dos padres.

d Son hombres.

e Son latinoamericanos.

f Crecieron en una familia con un único padre.

g Han utilizado drogas.

16 El archivo ncvs.dat incluye algunas variables de los incidentes con víctimas reportados entre julio y diciembre de 1989 en la NCVS. Los pesos de los incidentes son los pesos de las personas, divididos entre el número de víctimas implicadas en el incidente. Use estos datos para estimar el porcentaje de

a Incidentes violentos con víctimas.

b Situaciones de delitos violentos con víctimas lesionadas.

c Situaciones de delitos violentos con víctimas, reportados a la policía.

Realice los cálculos con y sin pesos. ¿Establecen estos pesos una diferencia? (*No* determine los errores estándar, pues no tiene la información suficiente para hacerlo.)

17 La encuesta británica de delitos (BCS) también es una encuesta estratificada con varios niveles (Aye Maung 1995). En contraste con la NCVS, la BCS no se diseñó para ser aproximadamente autoponderada, ya que las áreas céntricas de una ciudad participan en la muestra con casi el doble de proporción de las áreas no céntricas. En la BCS, los hogares se eligen usando el muestreo de probabilidad, pero sólo se entrevista a un adulto (elegido al azar) en cada hogar que contesta la encuesta. Haga igual a 1 el peso relativo de muestreo de un hogar céntrico.

a Considere la BCS como una muestra de hogares. ¿Cuál es el peso relativo de muestreo de un hogar no céntrico?

b Considere la BCS como una muestra de adultos. Construya una tabla de pesos relativos de muestreo para la muestra de adultos.

Cantidad de adultos	Áreas céntricas	Áreas no céntricas
1		
2		
3		
4		
5		

***18** (Requiere probabilidad.) *Estimadores de proporción combinados.* En una muestra estratificada, el estimador de proporción combinado del total de la población se define como

$\hat{t}_{\text{ycomb}} = t_x \hat{t}_y / \hat{t}_x$, , donde $\hat{t}_y = \sum_{h=1}^{H} \hat{t}_{yh}$, $\hat{t}_{yh}$ es un estimador insesgado del total de la población para y en el estrato h, $\hat{t}_x = \sum_{h=1}^{H} \hat{t}_{xh}$, $\hat{t}_{xh}$ es un estimador insesgado del total de la población para x en el estrato h y t_x es el total de la población para x.

a Muestre que

$$\frac{|\text{sesgo}\,[\hat{t}_{\text{ycomb}}]|}{\sqrt{V(\hat{t}_{\text{ycomb}})}} \leq \text{CV}(\hat{t}_x).$$

SUGERENCIA: Véase la página 66.

b En una muestra aleatoria estratificada, determine el sesgo aproximado y el ECM de $\hat{t}_{\text{ycomb}}$.

***19** (Requiere probabilidad.) *Estimadores de proporción por separado.* En una muestra estratificada, el estimador de proporción por separado del total de la población se define como

$$\hat{t}_{\text{ysep}} = \sum_{h=1}^{H} \frac{t_{xh} \hat{t}_{yh}}{\hat{t}_{xh}},$$

donde $\hat{t}_{yh}$ es un estimador insesgado del total de la población para y en el estrato h, $\hat{t}_{xh}$ es un estimador insesgado del total de la población para x en el estrato h y t_{xh} es el total de la población para x en el estrato h.

Use los resultados de la sección 3.1 para determinar el sesgo y una aproximación del ECM de $\hat{t}_{\text{ysep}}$ en una muestra aleatoria estratificada. Permita el uso de cocientes distintos, B_h, en cada estrato. ¿Cuándo será pequeño el sesgo?

8

Ausencia de respuesta

La señorita Schuster-Slatt comentó que imaginaba que los maridos ingleses eran cariñosos y que estaba preparando un cuestionario para circularlo entre los hombres jóvenes del Reino Unido, para intentar averiguar sus preferencias en cuanto al matrimonio.

"Pero los ingleses no responderán los cuestionarios", aseguró Harriet.

"¿No lo harán?", preguntó la señorita Schuster-Slatt, decepcionada.

"No", repuso Harriet, "no lo harán. Como nación, no nos importan las encuestas".

—Dorothy Sayers, *Gaudy Night*

La mejor manera de enfrentar la ausencia de respuestas es prevenirla. En ocasiones, ante la ausencia de respuesta, es posible modelar los datos faltantes, pero predecir esos datos nunca es tan bueno como observarlos de primera mano. Las personas que no responden con frecuencia difieren de manera crucial de las personas que sí lo hacen. Si la tasa de ausencia de respuesta no es despreciable, la inferencia basada tan sólo en quienes contestaron puede tener fallas serias.

En este capítulo analizamos dos tipos de ausencia de respuesta: **ausencia de respuesta por unidad**, en donde falta toda la unidad de observación, y **ausencia de respuesta por elemento**, en donde se dispone de algunas mediciones para la unidad de observación, pero donde falta al menos un elemento. En una encuesta de personas, la ausencia de respuesta por unidad significa que la persona no proporciona información para la encuesta; la ausencia de respuesta por elemento significa que la persona no responde a un punto particular del cuestionario. En la encuesta sobre la población actual (CPS) y la encuesta nacional a víctimas de delitos (NCVS), la ausencia de respuesta por unidad puede surgir por varias razones. Tal vez, el entrevistador no estableció contacto con la familia; la persona podría estar enferma, o el encuestado podría rehusarse a participar. En estas encuestas, el entrevistador intenta obtener cierta información demográfica de la unidad de habitación, como su nivel urbano/rural, para emplearla posteriormente para los ajustes por ausencia de respuesta. Esta ausencia principalmente se da por rechazo. Por ejemplo, una familia puede decidir no dar información acerca de sus ingresos.

En los estudios de agricultura o de la vida silvestre, generalmente se utiliza el término *datos faltantes* en lugar de *ausencia de respuesta*, pero los conceptos y las soluciones son similares. Por ejemplo, en un estudio de crías de patos, es posible que los investigadores no puedan encontrar las aves, por lo que, en cierto sentido, no responden. Tal vez el nido sufrió

el ataque de depredadores antes de que el investigador pudiera determinar la cantidad de huevos en él; lo que es comparable a la ausencia de respuesta por elemento.

En este capítulo analizaremos cuatro puntos de vista para enfrentar la ausencia de respuesta:

1 Prevención. Diseñe la encuesta de modo que la ausencia de respuesta sea pequeña. Éste es, con mucho, el mejor método.

2 Extraiga una submuestra representativa de quienes no responden; y úsela para establecer inferencias acerca de quienes no contestaron.

3 Utilice un modelo para predecir los valores de quienes no responden. Los pesos utilizan implícitamente un modelo para los ajustes debidos a la ausencia de respuestas. Con frecuencia, la imputación corrige la ausencia de respuesta por elemento y se pueden utilizar algunos modelos paramétricos para cualquiera de los tipos de ausencia de respuesta.

4 Ignore la ausencia de respuesta (No lo recomendamos, pero desgraciadamente es una práctica muy común).

8.1 Efectos por ignorar la ausencia de respuesta

EJEMPLO 8.1 Thomsen y Siring (1983) reportan los resultados de una encuesta de 1969 acerca del comportamiento electoral, realizada por la Oficina Central de Estadísticas de Noruega. En esta investigación, después de tres llamadas telefónicas se enviaba un cuestionario por correo. La tasa final de ausencia de respuesta fue de 9.9%, que generalmente se considera pequeña. ¿Diferían las personas que no contestaron de las que sí?

En el registro de votantes en Noruega era posible determinar si una persona había votado. El porcentaje de electores que votaron se podía comparar entre quienes contestaron la encuesta y quienes no; en la tabla 8.1 aparecen los resultados. La muestra seleccionada está formada por todas las personas elegidas para integrarla, incluyendo los datos del registro de votantes para ambos tipos (los que responden y los que no).

La diferencia en la tasa de votación entre quienes no contestaron y la muestra seleccionada era más grande en los grupos de menor edad. Entre quienes no contestaron, la tasa de votación varió con el tipo de ausencia de respuesta. La tasa global de votación para las personas que se rehusaron a participar fue 81%, la tasa de votación para quienes no estaban en casa fue 65% y la tasa de votación para quienes padecían enfermedades mentales y físicas fue 55%, lo que implica que la ausencia o la enfermedad fueron las principales causas del sesgo por ausencia de respuesta. ■

TABLA 8.1
Porcentaje de personas que votaron

		Edad				
	Todos	20-24	25-29	30-49	50-69	70-69
Personas que no contestaron	71	59	56	72	78	74
Muestra seleccionada	88	81	84	90	91	84

FUENTE: Adaptado de la tabla 8 en Thomsen y Siring 1983.

Se ha demostrado varias veces que la ausencia de respuesta puede tener grandes efectos en los resultados de una encuesta; en el ejemplo 8.1, una tasa de ausencia de respuesta de menos de 10% condujo a una sobreestimación de la tasa de votación en Noruega. Holt y Elliot analizan los resultados de una serie de estudios realizados sobre la ausencia de respuesta en el Reino Unido, y señalan que "las bajas tasas de ausencia de respuesta están asociadas con las siguientes características: residentes de Londres, familias sin automóvil, personas que viven solas; ancianos, divorciados o viudos originarios de algún país de la Comunidad Británica, bajo nivel de escolaridad o autoempleados" (1991, 334).

Además, incrementar el tamaño de la muestra sin enfrentar la ausencia de respuesta no reduce el sesgo por esa razón; una muestra mayor sólo proporciona más observaciones de la clase de personas que contestarían la encuesta. De hecho, al aumentar el tamaño de la muestra, el sesgo por la ausencia de respuesta podría empeorar, ya que tal vez se desperdiciaron recursos que hubiesen servido para reducir o remediar la ausencia de respuesta, o provocar una recolección de datos menor cuidadosa. Recuerde que la terrible encuesta de 1936 analizada en la página 7 tuvo 2.4 millones de personas que contestaron, pero una tasa de respuesta de menos de 25%. El propio censo de Estados Unidos, realizado cada 10 años, no incluye toda la población, y la tasa de subcobertura varía para los diversos grupos demográficos. A principios de la década de 1990, la ausencia de respuesta y la subcobertura de este censo provocaron un litigio de ciertas ciudades para obligar a la oficina de censos a realizar ajustes por la ausencia de respuesta y el debate continúa.

La mayoría de las encuestas pequeñas ignoran las ausencias de respuesta restantes después de telefonear de nuevo y de los seguimientos, y reportan los resultados con base sólo en registros completos. Hite (1987) lo hizo así en la encuesta analizada en el capítulo 1, y gran parte de las críticas a sus resultados se debieron a su baja tasa de respuesta. La ausencia de respuesta también se ignora en muchas encuestas que aparecen en los periódicos.

El análisis de los registros completos tiene como hipótesis subyacente la idea de que quienes no responden son similares a quienes sí, y que las unidades con elementos faltantes son similares a las unidades que tienen las respuestas de todas las preguntas. Muchas evidencias indican que estas hipótesis son erróneas. Por ejemplo, si se ignora la ausencia de respuesta en la encuesta sobre víctimas de delitos, se subestima la proporción de víctimas. Biderman y Cantor (1984) encontraron una baja proporción de éstan entre quienes contestaron en tres entrevistas consecutivas, comparados con quienes no respondieron en al menos una de esas entrevistas o que se mudaron antes de completar el estudio.

Los resultados de un análisis exclusivo de registros completos deben considerarse como representativos únicamente de la población de personas que respondieron a la encuesta, lo que rara vez equivale a la población objetivo. Si usted insiste en estimar las medias y totales de la población usando sólo los registros completos sin realizar ajustes por ausencia de respuestas, al menos deberá informar de la tasa correspondiente.

El principal problema de la ausencia de respuesta es el sesgo potencial de las estimaciones de la población. Imagine que la población está dividida en dos estratos, un tanto artificiales, de personas que responden y personas que no. Los individuos que responden entre la población son las unidades que responderían si fueran elegidos para estar en la muestra; la cantidad de personas que responden, N_R, no es conocida. De manera análoga, quienes no responden, N_M, son las unidades que no lo harán. Entonces tenemos las siguientes cantida-

des de la población:

Estrato	Tamaño	Total	Media	Varianza
Quienes responden	N_R	t_R	$\bar{y}_{RU}$	S_R^2
Quienes no responden	N_M	t_M	$\bar{y}_{MU}$	S_M^2
Población total	N	t	$\bar{y}_U$	S^2

La población, como un todo, tiene varianza $S^2 = \sum_{i=1}^{N}(y_i - \bar{y}_U)^2/(N-1)$, media $\bar{y}_U$, y total t. Es posible que una muestra de probabilidad de la población contenga algunas personas que respondan y otras que no pero, por supuesto, a la primera llamada no observamos y_i para ninguna de las unidades en el estrato de los que no responden. Si la media de la población en el estrato de los que no responden difiere de la media en el estrato de los que sí, la estimación de la media de la población utilizando sólo a quienes responden produce un sesgo.[1]

Sea $\bar{y}_R$ un estimador aproximadamente insesgado de la media en el estrato de los que sí responden, utilizando sólo a quienes lo han hecho. Como

$$\bar{y}_U = \frac{N_R}{N}\bar{y}_{RU} + \frac{N_M}{N}\bar{y}_{MU},$$

el sesgo es aproximadamente

$$E[\bar{y}_R] - \bar{y}_U \approx \frac{N_M}{N}(\bar{y}_{RU} - \bar{y}_{MU}).$$

El sesgo es pequeño si (1) la media de los que no responden es cercana a la media de los que sí, o (2) N_M/N es pequeña (hay poca ausencia de respuesta). Pero nunca podemos garantizar (1), pues generalmente carecemos de datos de quienes no responden. La única forma segura para controlar el sesgo por la ausencia de respuesta es minimizando la tasa de tal ausencia.

8.2 Diseño de encuestas para reducir errores que no son de muestreo

Una característica común de las encuestas pobres es la falta de tiempo invertido en su diseño y en el seguimiento de la ausencia de respuesta. Muchas personas poco experimentadas en las encuestas (y, por desgracia, algunas con no poca experiencia) simplemente comienzan a reunir datos sin tomar en cuenta los problemas potenciales en su proceso de recolección; envían cuestionarios a todos los elementos de la población objetivo y analizan los que regresan. No debe asombrarnos que tales encuestas tengan pobres tasas de respuesta. Muchos de los trabajos que aparecen en revistas académicas acerca del poder adquisitivo, por ejemplo, tienen tasas de respuesta de entre 10% y 15%. Es difícil concluir algo acerca de la población en una encuesta de este tipo.

El investigador que conoce bien la población objetivo podría anticipar algunas de las razones de la ausencia de respuesta y evitar algunas de ellas. Sin embargo, la mayoría de los investigadores no saben tanto de las razones de la ausencia de respuestas como creen saber. Necesitan descubrir por qué ocurren y resolver tantos problemas como sea posible antes de iniciar la encuesta.

[1] Con frecuencia, la varianza también es demasiado pequeña. Por ejemplo, en las encuestas sobre ingresos, es más probable que los ricos y los pobres no respondan a una pregunta sobre el ingreso. En ese caso, S_R^2, para el estrato de los que responden, es menor que S^2. La estimación puntual de la media puede estar sesgada, al igual que la estimación de la varianza.

Estas razones se pueden descubrir mediante algunos experimentos y la aplicación de métodos de mejoramiento de calidad a la recolección y procesamiento de los datos. ¿No sabe por qué algunas encuestas anteriores relacionadas con la suya tuvieron una baja tasa de respuesta? Diseñe un experimento para averiguarlo. ¿Cree que se introdujeron errores en el registro y procesamiento de los datos? Utilice un diseño anidado para descubrir las fuentes de error. Cualquier libro sobre control de calidad o diseño de experimentos le dirá cómo reunir los datos.

Por supuesto, puede basarse en los experimentos de otros investigadores como ayuda para minimizar los errores que no son de muestreo. La bibliografía sobre diseño de experimentos y control de calidad al final de este libro son un buen punto de partida. Hidiroglou *et al.* (1993) establecen un marco de referencia general para la ausencia de respuesta.

EJEMPLO 8.2 El censo de la década de 1990 en Estados Unidos intentó entrevistar a cada una de las más de 100 millones de familias de esa nación. La tasa de respuesta para la encuesta por correo fue de 65% y debía entrevisarse personalmente a las familias que no recibieron el cuestionario por correo, agregando millones de dólares al costo del censo. El incremento de la tasa de respuesta por correo para los próximos censos produciría grandes ahorros.

Dillman *et al* (1995a) informa de los resultados de un experimento factorial utilizado en la prueba de implantación de un censo de 1992, diseñado para explorar los efectos individuales y la interacción de tres factores experimentales sobre las tasas de respuesta. Los tres factores eran (1) una carta previa, avisando a la familia de la próxima llegada de la forma del censo, (2) un sobre con porte pagado, incluido con el formato del censo, y (3) una tarjeta postal con un recordatorio, enviada unos días después del formato del censo. Los resultados fueron asombrosos, como muestra la figura 8.1. El experimento estableció que, aunque los tres factores influyen en la tasa de respuesta, la carta y la tarjeta postal produjeron mayores tasas que el sobre con porte pagado. ■

La ausencia de respuesta puede tener varias causas por lo que no se recomienda un único método para cada encuesta. Platek (1977) clasifica las fuentes de ausencia de respuesta de

FIGURA 8.1
Tasas de respuesta logradas para cada combinación de los factores *carta*, *sobre* y *tarjeta*. La tasa de respuesta observada al utilizar los tres apoyos fue de 64.3% y al no hacerlo fue de 50%.

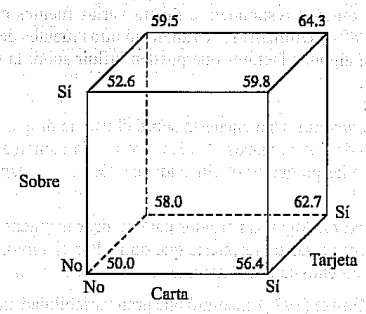
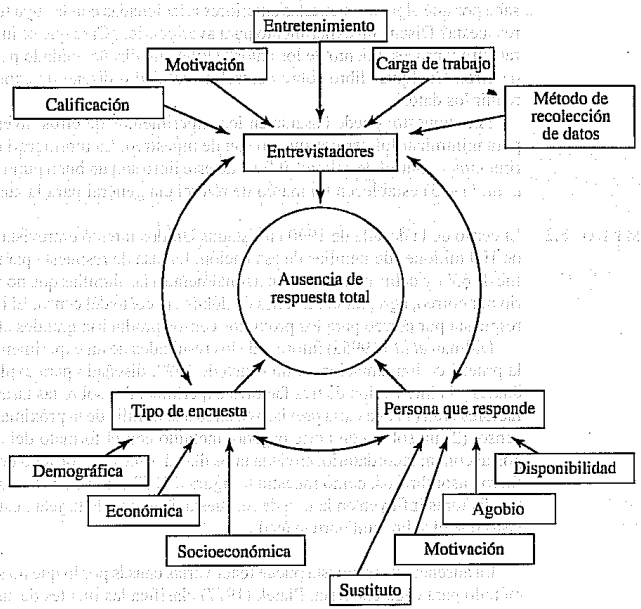


FIGURA 8.2
Factores que afectan la ausencia de respuesta



FUENTE: "Some Factors Affecting Non-Response", by R. Platek, 1977, *Survery Methodology*, 3, 191-214. Copyright© 1977 Survey Methodology. Reproducido con autorización.

acuerdo con (1) el contenido de la encuesta, (2) métodos de recolección de datos y (3) características de quienes responden, e ilustra varias fuentes con el diagrama de la figura 8.2. Groves (1989) y Dillman (1978) analizan otras fuentes de ausencia de respuesta. Los siguientes son algunos factores que pueden influir sobre la tasa de respuesta y la precisión de los datos.

- *Contenido de la encuesta.* Una encuesta sobre el uso de drogas o asuntos financieros puede tener gran cantidad de rechazos. A veces, se puede aumentar la tasa de respuesta ordenando con cuidado las preguntas o utilizando una técnica de respuesta aleatoria (vea la sección 12.5).

- *Tiempo de la encuesta.* Algunos periodos para realizar llamadas o temporadas del año pueden producir mayores tasas de respuesta que otras. Por ejemplo, el mes vacacional de agosto sería un mal momento para realizar una encuesta de casa en casa en Alemania.

- *Entrevistadores.* Gower (1979) encontró una gran variabilidad en las tasas de respuesta logradas por diversos entrevistadores, de modo que cerca del 15% de los éstos informaron

de una ausencia de respuesta casi nula. Algunos investigadores de campo en una encuesta de aves podrían ser mejores para ubicar e identificarlas que otros. Se pueden aplicar los métodos estándar de mejoramiento de calidad para aumentar la tasa de respuesta y la precisión de los entrevistadores. Los mismos métodos se pueden aplicar al proceso de codificación de datos.

- *Método de recolección de datos.* En general, las encuestas telefónicas y por correo tienen una tasa de respuesta menor que las encuestas personales (aunque también son más baratos). El sistema de entrevistas telefónicas apoyado por computadora (CATI) ha mostrado que mejora la precisión de los datos reunidos mediante encuestas telefónicas; con CATI, todas las preguntas se despliegan en una computadora y el entrevistador codifica las respuestas en la computadora al ir planteando las preguntas. CATI es particularmente útil en las encuestas donde una respuesta determina la pregunta que debe plantearse a continuación (Catlin e Ingram 1988).

Con frecuencia, las encuestas por correo, fax o Internet tienen bajas tasas de respuesta. Hay que analizar las posibles razones de la ausencia de respuesta en una encuesta por correo antes de enviar el cuestionario: ¿Se envió la encuesta a la dirección equivocada? ¿Descartaron los receptores al sobre como correo inútil incluso antes de abrirlo? ¿Llegará la encuesta al presunto receptor? ¿Creerá el receptor que el llenado de la encuesta vale la pena?

- *Diseño del cuestionario.* Ya hemos visto que la formulación de las preguntas tiene un efecto importante sobre las respuestas recibidas, pero también puede afectar al hecho de que se conteste a un elemento del cuestionario. El libro de Tanur (1993) explora estudios recientes acerca de la aplicación de la investigación cognoscitiva en el diseño de las preguntas. En una encuesta por correo, una forma bien diseñada para quien responde puede incrementar la precisión de los datos.

- *Agobio de las personas que responden.* Quienes contestan una encuesta le hacen un favor inmenso, y por lo que la encuesta debe ser lo menos entrometida posible. Un cuestionario breve, que requiera menos detalles, puede reducir el agobio de la persona que responde, que es una preocupación fundamental en las encuestas de panel como la NCVS, donde, cada seis meses durante tres años y medio se entrevista a las familias de la muestra. DeVries *et al.* (1966) analizan los métodos utilizados para reducir el agobio de las personas, que responden en los Países Bajos. Técnicas como la estratificación pueden reducir ese agobio pues una muestra menor basta para obtener la precisión requerida.

- *Presentación de la encuesta.* La presentación de la encuesta proporciona el primer contacto entre el entrevistador y la persona que potencialmente contestará; una buena presentación, que motive la respuesta, puede incrementar drásticamente las tasas de respuesta. La empresa Nielsen Media Research enfatiza a las familias seleccionadas en la muestra que su participación en las clasificaciones de Nielsen afecta los programas de televisión que se transmiten. La persona que responde debe saber la finalidad con la que se utilizarán los datos (personas sin escrúpulos frecuentemente pretenden estar realizando una encuesta cuando en realidad tratan de atacar a los clientes o neófitos) y garantizar la confidencialidad.

- *Incentivos y antiincentivos.* Los incentivos, financieros o de otros tipos, pueden incrementar las tasas de respuesta. Los antiincentivos también pueden ser útiles. Los médicos que se negaron a ser evaluados por sus colegas después de su selección en una muestra estratificada de la lista del Colegio de Médicos y Cirujanos de Ontario verían suspendidas sus licencias médicas. No asombra que la ausencia de respuesta sea baja (McAuley *et al* 1990).

- *Seguimiento.* El contacto inicial de la muestra es, por lo general, menos costoso por unidad que el seguimiento de quienes respondieron inicialmente. Si la encuesta inicial se realiza por correo, un recordatorio puede incrementar la tasa de respuesta. Pero no todos

contestaron las llamadas de seguimiento; algunas personas se negarán a responder la encuesta sin importar la frecuencia con que se establezca contacto con ellas. Usted debe decidir la cantidad de llamadas de seguimiento que debe realizar antes de que los resultados marginales no justifiquen el dinero invertido.

Debe tratar de obtener al menos cierta información acerca de quienes no respondan, que puede emplearse para realizar ajustes por la ausencia de respuesta e incluir elementos sustitutos que se puedan utilizar en la ausencia de respuesta por elementos. En realidad, no hay una compensación completa por no tener los datos, pero la información parcial puede ser mejor que nada. La información acerca de raza, sexo o edad de una persona que no contesta puede usarse posteriormente para los ajustes por ausencia de respuesta. Las preguntas sobre el ingreso pueden conducir a un rechazo, pero es posible que respondan las preguntas sobre autos, empleo o escolaridad y que éstas sirvan para predecir el ingreso. Si las pruebas preliminares de la encuesta indican un problema de ausencia de respuesta que usted no sepa cómo evitar, trate de diseñarla de modo que al menos se pueda reunir alguna información para cada unidad de observación.

La calidad de los datos de una encuesta queda determinada en gran medida en la etapa de diseño. Las palabras de Fisher (1938) acerca de los experimentos se pueden aplicar también al diseño de encuestas con muestras: "Llamar a un estadístico después de realizar un experimento, sería como pedirle que realice una autopsia: él sólo puede decir de qué murió el experimento". Cualquier presupuesto para una encuesta debe asignar recursos suficientes para el diseño de la encuesta y para el seguimiento de la ausencia de respuestas. No ahorre en el diseño de la encuesta; cada hora invertida en el diseño puede ahorrar semanas de remordimiento posterior.

8.3 Callbacks* y muestreo de dos etapas

Virtualmente todas las buenas encuestas se basan en callbacks para obtener las respuestas de personas que no están en casa durante el primer intento. El análisis de los datos que brindan pueden proporcionar cierta información acerca de los sesgos que pueden esperarse a partir de las personas restantes que no contestan a la encuesta.

EJEMPLO 8.3 Traugott (1987) analizó los datos de callbacks de dos encuestas realizadas en Michigan en 1984, acerca de la preferencia por algún candidato presidencial. Las tasas globales de respuesta fueron aproximadamente de 65%, que es un valor típico para las grandes encuestas políticas. Cerca de 21% de la muestra entrevistada respondió a la primera llamada; se realizaron hasta 30 intentos para localizar a las personas que no contestaron a la primera llamada. Traugott determinó que las personas que respondieron después de varios intentos eran más probablemente hombres, de más edad y republicanos, en comparación con los que contestaron pronto, mientras que 48% de quienes lo hicieron en la primera llamada apoyaban a Reagan y 45% apoyaban a Mondale; 59% de la muestra total apoyaron a Reagan y 39% a Mondale. La diferencia entre los procedimientos para el seguimiento de quienes no responden y la insistencia en las llamadas pueden explicar algunas inconsistencias de las encuestas políticas.

Si quienes no contestan se parecen a quienes lo hacen tarde, se podría especular que es más probable que quienes no responden estén a favor de Reagan pero quienes no lo hacen no necesariamente se parecen a los difíciles de alcanzar; las personas que se rehúsan terminantemente a participar pueden diferir en gran medida de las personas que no pudieron

*N. del R.T. Callbacks se refiere a efectuar una segunda llamada al encuestado.

contestar de inmediato y es más probable que quienes no respondieron estén enfermos o alguna otra circunstancia impida su participación. Tampoco conocemos la probabilidad con la que quienes no contestaron la encuesta votarán en la elección. Incluso, si especulamos que es más probable que estén a favor de Reagan, no necesariamente es más probable que voten por Reagan. ■

Con frecuencia, cuando la encuesta se diseña de modo que se usen callbacks, el contacto inicial es mediante una encuesta por correo; las llamadas de seguimiento utilizan un método más caro, como la entrevista personal.

Hansen y Hurwitz (1946) propusieron un submuestreo de las personas que no responden, así como el uso del **muestreo en dos fases** (también llamado **muestreo doble**) para la estratificación, para estimar la media o el total de la población. La población se divide en dos estratos, como se describe en la sección 8.1. Los dos estratos son quienes responden y quienes no responden al inicio, las personas que no contestaron a la primera llamada. En la sección 12.1 veremos la teoría del muestreo en dos fases para los diseños generales de encuestas; aquí ilustraremos la forma de emplarlo para la ausencia de respuesta.

En la forma más sencilla del muestreo en dos fases, seleccionamos al azar n unidades de la población. De éstas, n_R responden y n_M no. Los valores n_R y n_M son variables aleatorias; cambiarán si elegimos otra muestra aleatoria simple. Luego, hacemos una segunda llamada sobre una submuestra aleatoria de $100\nu\%$ de las n_M personas que no han respondido en la muestra, donde la fracción de submuestreo ν no depende de los datos reunidos.

Suponga que mediante un esfuerzo sobrehumano se logra establecer un contacto con todas las personas objetivo que no han contestado. Sea $\bar{y}_R$ el promedio muestral de quienes respondieron originalmente y $\bar{y}_M$ el promedio de quienes no respondieron de la submuestra. Las estimaciones de muestreo en dos fases para la media y el total de la población son

$$\hat{\bar{y}} = \frac{n_R}{n}\bar{y}_R + \frac{n_M}{n}\bar{y}_M \tag{8.1}$$

y

$$\hat{t} = N\hat{\bar{y}} = \frac{N}{n}\sum_{i\in S_R} y_i + \frac{N}{n}\frac{1}{\nu}\sum_{i\in S_M} y_i, \tag{8.2}$$

donde S_R representa las unidades de la muestra que están en el estrato de quienes responden y S_M representa las unidades de la muestra en el estrato de quienes no responden. Observe que $\hat{t}$ es una suma ponderada de las unidades observadas; los pesos son N/n para quienes responden y $N/(n\nu)$ para quienes no atienden en la submuestra. Como sólo se extrajo una submuestra en el estrato de quienes no responden, cada unidad de la submuestra de ese estrato representa más unidades en la población de lo que representa una unidad en el estrato de quienes no responden.

En la sección 12.1 calculamos el valor esperado y la varianza de estos estimadores. Como $\hat{t}$ es un estimador con probabilidades diferentes y pesos adecuados, el teorema 6.2 implica que $E[\hat{t}] = t$. De (12.5), si podemos ignorar la corrección para poblaciones finitas, podemos estimar la varianza como

$$\hat{V}(\hat{\bar{y}}) = \frac{n_R - 1}{n-1}\frac{s_R^2}{n} + \frac{n_M - 1}{n-1}\frac{s_M^2}{\nu n} + \frac{1}{n-1}\left[\frac{n_R}{n}(\bar{y}_R - \hat{\bar{y}})^2 + \frac{n_M}{n}(\bar{y}_M - \hat{\bar{y}})^2\right].$$

Si todos responden en la submuestra, el muestreo en dos fases no sólo elimina el sesgo por la ausencia de respuesta, sino que también toma en cuenta la ausencia de respuesta original en la varianza estimada.

8.4 Mecanismos para la ausencia de respuesta

La mayoría de las encuestas tienen cierta ausencia de respuesta residual, aun después de un diseño cuidadoso y un seguimiento de la ausencia de respuesta. Todos los métodos utilizados para tratar la ausencia de respuesta se basan en el modelo. Si quisiéramos establecer inferencias acerca de quienes no responden, deberíamos suponer que de alguna manera ellos están relacionados con quienes responden. Una buena referencia no técnica para los métodos para enfrentar la ausencia de respuesta es Groves (1989); la obra en tres volúmenes editada por Madow *et al.* (1983) contiene mucha información actualizada sobre la investigación estadística acerca de la ausencia de respuesta.

La separación de los miembros de la población en dos estratos fijos, de los que contestarían y los que no, está bien para pensar en el sesgo potencial por la ausencia de respuesta y para los métodos de dos etapas. Para ajustar la ausencia de respuesta que resta después de tomar las demás medidas, necesitamos una configuración más elaborada, haciendo que la respuesta o no respuesta de la unidad i sea una variable aleatoria. Definimos la variable aleatoria

$$R_i = \begin{cases} 1 \text{ si la unidad } i \text{ responde.} \\ 0 \text{ si la unidad } i \text{ no responde.} \end{cases}$$

Después del muestreo podemos conocer las realizaciones del indicador de respuesta para las unidades elegidas en la muestra. Registramos el valor de y_i si r_i, la realización de R_i, es 1. Por supuesto, desconocemos la probabilidad de que responda una unidad seleccionada en la muestra,

$$\phi_i = P(R_i = 1),$$

aunque suponemos que es positiva. Rosenbaum y Rubin (1983) llaman a ϕ_i la **calificación de propensión** para la unidad i.

Suponga que y_i es una respuesta de interés y que $\mathbf{x}_i$ es un vector de información conocida acerca de la unidad i de la muestra. La información utilizada en el diseño de la encuesta está incluida en $\mathbf{x}_i$. Consideramos tres tipos de datos faltantes, usando la terminología de Little y Rubin (1987) para clasificar la ausencia de respuesta.

Faltante completamente al azar Si ϕ_i no depende de $\mathbf{x}_i$, y_i, o del diseño de la encuesta, los datos faltantes son **faltantes completamente al azar** (MCAR, missing completely at random). Tal situación ocurre si, por ejemplo, una persona del laboratorio tira un tubo de ensayo con la muestra de sangre de uno de los participantes de la encuesta; no hay razón para creer que el hecho de tirar el tubo de ensayo tiene algo que ver con la cantidad de glóbulos blancos.[2] Si los datos son MCAR, las personas que responden son representativas de la muestra seleccionada.

Los datos faltantes en la NCVS serían MCAR si la probabilidad de la ausencia de respuesta no tiene relación alguna con la región de Estados Unidos, la raza, el sexo, la edad o cualquier otra variable medida en la muestra y si la probabilidad de la ausencia de respuesta no está relacionada con variable alguna relativa a la calidad de víctima. Las personas que no responden serían elegidas esencialmente al azar entre la muestra.

[2] Sin embargo, incluso en este caso, las mentes suspicaces podrían crear un escenario donde la ausencia de respuesta podría estar relacionada con las cantidades de interés: Tal vez los trabajadores del laboratorio tengan menor probabilidad de tirar los tubos de ensayo que crean que contenga HIV.

Si las probabilidades de respuesta ϕ_i son todas iguales y los eventos $\{R_i = 1\}$ son condicionalmente dependientes entre sí e independientes del proceso de selección de la muestra dado n_R, entonces los datos son MCAR. Si se extrae una muestra aleatoria simple de tamaño n bajo este mecanismo, entonces las personas que responden serán una submuestra aleatoria simple de tamaño variable n_R. La media muestral de quienes responden, $\bar{y}_R$, es aproximadamente insesgada para la media de la población. El mecanismo MCAR se adopta implícitamente cuando se ignora la ausencia de respuesta.

Faltante al azar dadas las covariantes o ausencia de respuesta ignorable Si ϕ_i depende de $\mathbf{x}_i$ pero no de y_i, los datos son **faltantes al azar** (MAR, missing at random); la ausencia de respuesta depende sólo de las variables observadas. Podemos modelar con éxito la ausencia de respuesta, pues conocemos los valores de $\mathbf{x}_i$ para todas las unidades de la muestra. Las personas de la NCVS serían MAR si la probabilidad de responder a la encuesta dependiese de la raza, la edad y el sexo (todas son cantidades conocidas), aunque invariante con respecto de la experiencia de ser una víctima, dentro de cada clase edad/raza/sexo. Esto se conoce a veces como **ausencia de respuesta ignorable:** ignorable significa que un modelo puede explicar el mecanismo de ausencia de respuesta y que ésta se puede ignorar después de que el modelo la toma en cuenta, no que la ausencia de respuesta pueda ser completamente ignorada para usar métodos con datos completos.

Ausencia de respuesta no ignorable Si la probabilidad de la ausencia de respuesta depende del valor de una variable de respuesta y no puede explicarse por completo mediante los valores de las $\mathbf{x}_i$, entonces la ausencia de respuesta es **no ignorable**. Ésta es probablemente la situación para la NCVS; se sospecha que una persona que ha sido víctima de un delito tiene menor probabilidad de responder a la encuesta que una persona que no lo ha sido incluso si comparten los valores de todas las variables conocidas, como raza, edad y sexo. Es más probable que las víctimas se muden después del acto y con ello no queden incluidas en las entrevistas posteriores de la encuesta sobre víctimas de delitos. Los modelos pueden ser útiles, pues la probabilidad de la ausencia de respuesta también puede depender de variables conocidas, pero que no pueden ajustarse por completo para la ausencia de respuesta.

Las probabilidades de responder, ϕ_i, son útiles para reflexionar acerca del tipo de ausencia de respuesta. Por desgracia se desconoce, de modo que no sabemos qué tipo de ausencia de respuesta está presente. A veces podemos distinguir entre MCAR y MAR ajustando un modelo que prediga las probabilidades observadas de respuesta para subgrupos de covariados conocidos; si los coeficientes en un modelo de regresión logística son significativamente distintos de cero, es probable que los datos faltantes no sean MCAR. La distinción entre MAR y la ausencia de respuesta no ignorable es más difícil. En la siguiente sección analizamos un método para estimar las ϕ_i.

8.5 Métodos de ponderación para la ausencia de respuesta

En los capítulos anteriores hemos visto cómo utilizar los pesos para calcular las estimaciones de diversos esquemas de muestreo (vea las secciones 4.3, 5.4 y 7.2). Los pesos de muestreo son los recíprocos de las probabilidades de selección, de modo que una estimación del total de la población es $\sum_{i \in S} \omega_i y_i$. Para la estratificación, los pesos son $w_i = N_h/n_h$ si la unidad i está en el estrato h; para los elementos de muestreo con probabilidades diferentes, $w_i = 1/\pi_i$.

Los pesos también son útiles para ajustar la ausencia de respuesta. Sea Z_i la variable indicadora de la presencia en la muestra seleccionada, donde $P(Z_i = 1) = \pi_i$. Si R_i es inde-

pendiente de Z_i, entonces la probabilidad de que la unidad i sea medida es

$$P(\text{seleccionar la unidad } i \text{ en la muestra y que responda}) = \pi_i \phi_i.$$

La probabilidad de responder, f_i, se estima para cada unidad de la muestra, usando la información auxiliar conocida para todas las unidades de la muestra seleccionada. El peso final para una persona que responde es entonces $1/(\pi_i \hat{\phi}_i)$. El método de ponderación supone que podemos estimar las probabilidades de respuesta a partir de variables conocidas para todas las unidades, suponiendo datos MAR. La bibliografía para mayor información acerca de la ponderación es Oh y Scheuren (1983) y Holt y Elliot (1991).

8.5.1 Ajuste de clases de ponderación

Hemos interpretado los pesos de muestreo w_i como la cantidad de unidades en la población representadas por la unidad i de la muestra. Los métodos de clases de ponderación amplían este punto de vista para compensar los errores no de muestreo: se utilizan variables conocidas para todas las unidades de la muestra seleccionada para formar clases de ajuste del peso y se espera que quienes responden y quienes no lo hacen en la misma clase de ajuste del peso sean similares. Los pesos de quienes contestan en la clase de ajuste del peso se incrementan de modo que quienes lo hacen representen la parte de la población de quienes no responden, así como su parte correspondiente.

EJEMPLO 8.4 Suponga que conocemos la edad para cada miembro de la muestra seleccionada y que la persona i de la muestra seleccionada tiene el peso de muestreo $w_i = 1/\pi_i$. Entonces, se pueden formar las clases de ponderación dividiendo la muestra seleccionada entre los distintos grupos de edad, como muestra la tabla 8.2.

Estimamos la probabilidad de respuesta de cada clase mediante

$$\hat{\phi}_c = \frac{\text{suma de pesos para quienes responden en la clase } c}{\text{suma de pesos para la muestra seleccionada en la clase } c}.$$

Entonces, el peso de muestreo de cada persona que responde en la clase c es multiplicado por $1/\hat{\phi}_c$, el factor de ponderación de la tabla 8.2. El peso de cada persona que responde, con edad entre 15 y 24, por ejemplo, se multiplica por 1.622. Como no hubo ausencia de respuesta en el grupo de más de 65 años, sus pesos no cambian.

TABLA 8.2
Ilustración de los factores de ajuste de clases de ponderación

	Edad					
	15-24	25-34	35-44	45-64	65+	Total
Tamaño de la muestra	202	220	180	195	203	1,000
Personas que responden	124	187	162	187	203	863
Suma de pesos para la muestra	30,322	33,013	27,046	29,272	30,451	150,104
Suma de pesos para quienes responden	18,693	28,143	24,371	28,138	30,451	
$\hat{\phi}_c$	0.6165	0.8525	0.9011	0.9613	1.0000	
Factor de ponderación	1.622	1.173	1.110	1.040	1.000	

Suponemos que la probabilidad de respuesta es la misma dentro de cada clase de ponderación, lo que implica que dentro de una clase de ponderación, la probabilidad de respuesta no depende de y. Como ya hemos dicho, los métodos de clases de ponderación suponen datos MAR. El peso de una persona que responde en la clase de ponderación c es $1/(\pi_i \hat{\phi}_c)$.

Para estimar el total de la población usando los ajustes de clases de ponderación, sea $x_{ci} = 1$ si la unidad i está en la clase c y cero en caso contrario. Entonces el nuevo peso para la persona i que responde es

$$\tilde{\omega}_i = \sum_c \frac{\omega_i x_{ci}}{\hat{\phi}_c},$$

donde w_i es el peso de muestreo para la unidad i; $\tilde{\omega}_i = \omega_i / \hat{\phi}_c$ si la unidad i está en la clase c. Asignamos $\tilde{\omega}_i = 0$ si la unidad i es alguien que no responde. Entonces,

$$\hat{t}_{\omega c} = \sum_{i \in S} \tilde{\omega}_i y_i$$

y

$$\hat{\bar{y}}_{\omega c} = \frac{\hat{t}_{\omega c}}{\sum_{i \in S} \tilde{\omega}_i}.$$

Por ejemplo, en una muestra aleatoria simple, si n_c es la cantidad de unidades de la muestra que están en la clase c, n_{cR} es la cantidad de personas que responden y que están en la clase c y $\bar{y}_{cR}$ es el promedio para quienes responden y que están en la clase c, entonces $\hat{\phi}_c = n_{cR}/n_c$ y

$$\hat{t}_{\omega c} = \sum_{i \in S} \sum_c \frac{N}{n} \frac{n_c}{n_{cR}} x_{ci} y_i = N \sum_c \frac{n_c}{n} \bar{y}_{cR}.$$

EJEMPLO 8.5 *La encuesta nacional a víctimas de delitos*

Para ajustar la ausencia de respuesta individual en la NCVS, utilizamos el factor de ajuste para la falta de entrevistas dentro de los hogares (FADH) del capítulo 7. Los entrevistadores de la NCVS reúnen la información demográfica de quienes no responden y esta información se usa para clasificar a todas las personas en 24 celdas de ajuste de peso. Las celdas dependen de la edad de la persona, la relación de la persona con la persona de referencia (la cabeza de la familia) y su raza.

Para cualquier celda, sea W_R la suma de los pesos para quienes responden y W_M para quienes no responden. Entonces, el nuevo peso para una persona que contesta y que está en cierta celda será el peso anterior, multiplicado por el factor de ajuste de peso $(W_M + W_R)/W_R$. Así, los pesos que se asignarían a quienes no responden son redistribuidos entre quienes responden y que tienen (esperamos) características similares.

Hay un problema cuando $(W_M + W_R)/W_R$ es demasiado grande. Si $(W_M + W_R)/W_R > 2$, la celda contiene más personas que no responden que personas que sí. En este caso, la varianza de la estimación aumenta; si el número de quienes responden en la celda es pequeño, el peso podría ser inestable. La oficina de censos de Estados Unidos une las celdas para obtener factores de ajuste de peso menores o iguales a 2. Si hay menos de 30 personas entrevistadas en una celda o si el factor de ajuste de peso es mayor que 2, la celda se une (colapsa) con las celdas cercanas, hasta que la nueva celda tiene más de 30 observaciones y un factor de ajuste de peso menor o igual a 2. ■

Construcción de clases de ponderación Las clases de ajuste de peso deben construirse como si fuesen estratos; como se muestra en la siguiente sección, el ajuste de peso es similar a la estratificación posterior. Las clases deben formarse de modo que las unidades dentro de cada clase sean lo más similares posible con respecto a las principales variables de interés, de modo que las tasas de respuesta varíen de una clase a otra.

Little (1986) sugiere estimar las probabilidades de respuesta ϕ_i como función de las variables conocidas (tal vez mediante una regresión logística) y agrupar las observaciones en clases con base en $\hat{\phi}_i$. Este punto de vista es preferible a sólo usar los valores estimados de ϕ_i en pesos individuales, ya que las probabilidades de respuesta estimadas pueden ser extremadamente variables y hacer que las estimaciones finales sean inestables.

8.5.2 Estratificación posterior

La estratificación posterior es similar al ajuste de clases de ponderación, excepto que los datos de la población se usan para ajustar los pesos. Suponga que se extrae una muestra aleatoria simple; después de reunirla, las unidades se agrupan en H estratos, generalmente a partir de variables demográficas, como la raza o el sexo. La población tiene N_h unidades en el estrato posterior h; de éstas, n_h se seleccionaron para la muestra y n_{hR} respondieron. El estimador estratificado posterior para $\bar{y}_U$ es

$$\bar{y}_{\text{post}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{hR};$$

el estimador de clases de ponderación para $\bar{y}_U$ si las clases de ponderación son los estratos posteriores, es

$$\bar{y}_{wc} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_{hR}.$$

Los dos estimadores tienen una forma similar. La única diferencia es que en la estratificación posterior se conocen los N_h, mientras que en los ajustes de clases de ponderación no se conocen y se estiman mediante Nn_h/n.

Para el estimador estratificado posteriormente, a menudo se emplea la varianza condicional dados los n_{hR}. Para una muestra aleatoria simple,

$$V(\bar{y}_{\text{post}} \mid n_{hR}, h=1,\ldots,H) = \sum_{h=1}^{H} \left(\frac{Nh}{N}\right)^2 \left(1-\frac{n_{hR}}{N_h}\right)\left(\frac{S_h^2}{n_{hR}}\right). \quad \textbf{(8.3)}$$

La varianza no condicional de $\bar{y}_{\text{post}}$ es ligeramente mayor, con algunos términos adicionales de orden $1/n_{hR}^2$, de acuerdo con Oh y Scheuren (1983). En el ejercicio 5 del capítulo 9 aparece un estimador de la varianza para la estratificación posterior.

8.5.2.1 Estratificación posterior usando pesos

En el diseño general de encuestas, se supone que la suma de los pesos en el subgrupo h estima las cifras de la población N_h para ese subgrupo. La estratificación posterior utiliza el estimador de proporción dentro de cada subgrupo para los ajustes debidos a las cifras reales de la población.

Sea $x_{hi} = 1$ si la unidad i es alguien que responde posteriormente en el estrato h, y 0 en caso contrario. Sea entonces

$$\omega_i^* = \sum_{h=1}^{H} \omega_i x_{hi} \frac{N_h}{\sum_{j\in S} \omega_j x_{hj}}.$$

Usando los pesos modificados,

$$\sum_{i\in S} \omega_i^* x_{hi} = N_h,$$

y el estimador estratificado posteriormente del total de la población es

$$\hat{t}_{\text{post}} = \sum_{i\in S} \omega_i^* y_i.$$

La estratificación posterior puede ajustar la subcobertura y la ausencia de respuesta si las cifras de la población N_h incluyen en la encuesta personas que no están en el marco de muestreo de la misma.

EJEMPLO 8.6 El factor de la segunda etapa en la NCVS (véase la sección 7.6) utiliza la estratificación posterior para ajustar los pesos. Después de realizar los demás ajustes de pesos, incluyendo los ajustes de clases de ponderación para la ausencia de respuesta, se usa la estratificación posterior para que las cifras de la muestra coincidan con las estimaciones de las cifras de la población de la oficina de censos. Cada persona se asigna a algunos de los 72 estratos posteriores con base en la edad, raza y sexo de la persona. La cantidad de integrantes de la población que están en ese estrato posterior, N_h, se conoce por otras fuentes. Entonces, el peso para una persona en el estrato posterior h se multiplica por

$$\frac{N_h}{\text{suma de pesos para todos los que responden en el estrato posterior } h}.$$

Con los tipos de ponderación, el factor de ponderación para ajustar la ausencia de respuesta por unidad siempre es al menos 1. Con la estratificación posterior, como los pesos se ajustan de modo que la suma sea un total conocido de la población, el factor de ponderación puede ser cualquier número positivo, aunque son deseables factores de ponderación menores o iguales a 2. ■

La estratificación posterior supone que (1) dentro de cada estrato posterior, cada unidad seleccionada para estar en la muestra tiene la misma probabilidad de ser alguien que responda, (2) la respuesta o falta de respuesta de una unidad es independiente del comportamiento de las demás unidades y (3) las personas que no responden en un estrato posterior son como las que sí lo hacen. Los datos son MCAR dentro de cada estrato posterior. Éstas son hipótesis importantes; para que sean un poco más plausibles, los investigadores frecuentemente utilizan muchos estratos posteriores. Sin embargo, un gran número de estratos posteriores pueden crear problemas adicionales, pues si hay pocas personas que respondan en algún estrato posterior se pueden obtener estimaciones inestables y problemas para aplicar el teorema central de límite. Al enfrentar estratos posteriores con pocas observaciones, la mayoría de los practicantes integra los estratos posteriores con otros que tengan medias similares en observaciones clave, hasta tener una cantidad razonable de observaciones en cada estrato posterior. Para la encuesta de población actual CPS, un número "razonable" significa que cada grupo tiene al menos 20 observaciones y que la tasa de respuesta para cada grupo es de al menos 50%.

8.5.2.2 Ajuste de rastrilleo

El *rastrilleo* es un método de estratificación posterior que se puede usar cuando los estratos posteriores se forman usando más de una variable, pero sólo se conocen los totales margina-

les de la población. El rastrilleo se utilizó por vez primera en el censo de 1940 para garantizar que los datos completos del censo y las muestras extraídas de él tenían resultados consistentes. Lo introdujeron Deming y Stephan (1940); Brackstone y Rao (1976) ampliaron la teoría. Oh y Scheuren (1983) describen las estimaciones de proporciones con rastrilleo para la ausencia de respuesta.

Considere la siguiente tabla de sumas de pesos de una muestra; cada entrada es la suma de los pesos de muestreo para las personas de la muestra que caen en esa clasificación (por ejemplo, la suma de los pesos de muestreo para las mujeres de raza negra es 300).

	Negro	Blanco	Asiático	Indígena	Otros	Suma de pesos
Mujer	300	1200	60	30	30	1620
Hombre	150	1080	90	30	30	1380
Suma de pesos	450	2280	150	60	60	3000

Ahora, suponga que conocemos las cifras reales de la población para los totales marginales. Sabemos que la población tiene 1510 mujeres y 1490 hombres, 600 negros, 2120 blancos, 150 asiáticos, 100 indígenas y 30 personas en la categoría "otros". Sin embargo, las cifras de la población para cada celda de la tabla son desconocidas; no conocemos la cantidad de mujeres negras en esta población y no podemos suponer que haya independencia. El rastrilleo nos permite ajustar los pesos para que las sumas de los pesos en los márgenes sean iguales a las cifras de la población.

Primero, ajustamos los renglones. Multiplicamos cada entrada por (población real del renglón)/(población estimada del renglón). Al multiplicar las celdas del renglón de las mujeres por 1510/1620 y las celdas del renglón de los hombres por 1490/1380 obtenemos la siguiente tabla:

	Negro	Blanco	Asiático	Indígena	Otros	Suma de pesos
Mujer	279.63	1118.52	55.93	27.96	27.96	1510
Hombre	161.69	1166.09	97.17	32.39	32.39	1490
Total	441.59	2284.61	153.10	60.35	60.35	3000

Ahora, los totales por renglón están bien, pero los totales por columna no son iguales a los totales de la población. Repetimos el mismo procedimiento con las columnas de la nueva tabla. Cada una de las entradas de la primera columna se multiplica por 600/441.59. Obtenemos la siguiente tabla:

	Negro	Blanco	Asiático	Indígena	Otros	Suma de pesos
Mujer	379.94	1037.93	54.79	46.33	13.90	1532.90
Hombre	220.06	1082.07	95.21	53.67	16.10	1467.10
Total	600.00	2120.00	150.00	100.00	30.00	3000.00

Pero esto nuevamente ha echado a perder los totales por renglón. Repita el procedimiento hasta que los totales por renglón y por columna sean iguales a las cifras de la población. El procedimiento converge siempre que todas las cifras de las celdas sean positivas. En este ejemplo, la tabla final de cifras ajustadas es

	Negro	Blanco	Asiático	Indígena	Otros	Suma de pesos
Mujer	375.59	1021.47	53.72	45.56	13.67	1510
Hombre	224.41	1098.53	96.28	54.44	16.33	1490
Total	600.00	2120.00	150.00	100.00	30.00	3000

Las entradas de la última tabla podrían ser mejores estimaciones de las poblaciones de la celda (es decir, con menor varianza) que las estimaciones ponderadas originales, tan sólo porque usan más información relativa a la población. El factor de ajuste de peso para cada hombre blanco en la muestra es 1098.53/1080; el peso de cada hombre blanco se incrementa un poco para ajustar la ausencia de respuesta y la subcobertura. De manera análoga, los pesos de las mujeres blancas disminuyen, pues están representadas en exceso en la muestra.

Las hipótesis para el rastrilleo son las mismas que para la estratificación posterior, con la hipótesis adicional de que las probabilidades de respuesta dependen sólo del renglón y la columna, y no de la celda particular. Si los tamaños de muestra en cada celda son bastante grandes, el estimador de rastrilleo es aproximadamente insesgado.

El rastrilleo tiene algunas dificultades: el algoritmo puede no converger si algunas de las estimaciones por celda se anulan. Hay también un cierto peligro de "ajuste excesivo": si existe poca relación entre la dimensión adicional en el rastrilleo y las medias de la celda, el rastrilleo puede incrementar la varianza en lugar de disminuirla.

8.5.3 Estimación de la probabilidad de respuesta: otros métodos

Algunos métodos de clases de ponderación usan pesos que son los recíprocos de la probabilidad de respuesta estimada. Un ejemplo famoso es el método de Politz-Simmons para ajustar la indisponibilidad de los miembros de la muestra.

Suponga que todas las llamadas telefónicas se realizan en las tardes, de lunes a viernes. A cada persona que responde se le pregunta si estaba en casa, en el momento de la entrevista, en cada una de las tardes de los cuatro días entre semana anteriores. El entrevistado responde que estaba en casa k de esas cuatro noches. Entonces, se supone que la probabilidad de respuesta es proporcional a la cantidad de noches en casa durante las horas de entrevista, de modo que la probabilidad de respuesta se estima como $\hat{\phi}_i = (k_i + 1)/5$. El peso de muestreo w_i para cada persona que responde se multiplica entonces por $5/(k_i + 1)$. Las personas que contestan, con $k = 0$, sólo estuvieron en casa una de las cinco tardes y se les asigna la representación de su parte de la población, más la parte de las cuatro personas de la muestra que se llamaron durante una de sus tardes "no disponibles". Las personas que atienden y que tienen mayor probabilidad de estar en casa tienen $k = 4$; se supone que se estableció contacto con todas las personas de la muestra y que estuvieron en casa todas las tardes, de modo que sus pesos no varían. La estimación de la media de la población es

$$\hat{\bar{y}} = \frac{\sum_{i \in S} \frac{5\omega_i y_i}{k_i + 1}}{\sum_{i \in S} \frac{5\omega_i}{k_i + 1}}.$$

Este método de ponderación, descrito por Hartley (1946) y Politz y Simmons (1949), se basa en la premisa de que las personas más accesibles tenderán a estar representadas en exceso en los datos de la encuesta. El método es fácil de usar, atractivo desde el punto de

vista teórico y se puede emplear junto con callbacks. Aún así, omite a las personas que no estuvieron en casa en las cinco tardes o que se rehusaron a participar en la encuesta. Como la ausencia de respuesta se debe principalmente al rechazo en algunas encuestas telefónicas, el método de Politz-Simmons podría ser inútil para enfrentar la ausencia de respuesta. Los valores de k también pueden tener cierto error, pues las personas pueden equivocarse al recordar el número de tardes que estuvieron en casa.

Potthoff *et al* (1993) modificaron y ampliaron el método de Politz-Simmons para determinar los pesos con base en la cantidad necesaria de rellamadas, suponiendo que las ϕ_i siguen una distribución beta.

8.5.4 Una advertencia en cuanto a los pesos

Los modelos para los ajustes de pesos para la ausencia de respuesta son fuertes: en cada celda de ponderación, se supone que quienes responden y quienes no, son similares. Se supone que cada individuo de una clase de ponderación tiene la misma probabilidad de responder la encuesta, sin importar el valor de la respuesta. Estos modelos nunca describen en forma exacta el estado real de las cosas y siempre se debe evaluar su plausibilidad e implicaciones. Una tendencia desafortunada de muchos practicantes de encuestas es considerar al ajuste de pesos como un remedio completo y actuar entonces como si no hubiese ausencia de respuestas. Los pesos pueden mejorar gran parte de las estimaciones, pero es raro que eliminen todos los sesgos por ausencia de respuesta. Si se realizan los ajustes de peso (recuerde que el hecho de no realizar ajustes es también un modelo acerca de la naturaleza de la ausencia de respuesta), los practicantes siempre deben establecer el modelo de respuesta supuesto y dar una evidencia para justificarlo. Por lo general, los ajustes de peso se usan para la ausencia de respuesta por unidad y no para la ausencia de respuesta por elemento (que requiere un peso distinto para cada elemento).

8.6 Imputación

La omisión de elementos en las encuestas puede ocurrir por varias razones: un entrevistador puede olvidarse de preguntar algo; quien responde puede rehusarse a contestar alguna pregunta o tal vez no pueda proporcionar la información; el funcionario que captura los datos podría omitir el valor correspondiente. A veces, los elementos con respuestas cambian elementos faltantes cuando el conjunto de datos se edita o limpia: un editor de datos podría no poder resolver la discrepancia entre una persona de tres años de edad que haya votado en la última elección y puede cambiar la categoría de ambos datos como faltantes.

La **imputación** generalmente se utiliza para asignar valores a los elementos faltantes. Frecuentemente se asigna un valor de reemplazo al valor faltante mediante un valor de otra persona en la encuesta, similar a la que no responde al elemento con respecto a otras variables. Al usar la imputación, hay que crear una variable adicional en el conjunto de datos que indique si la respuesta fue medida o imputada.

Los procedimientos de imputación se utilizan no sólo para reducir el sesgo por la ausencia de respuestas, sino para producir un conjunto de datos rectangular y "limpio", sin huecos en los valores faltantes. Si queremos analizar las tablas de ciertos subgrupos de la población, la imputación nos permite hacerlo sin considerar la ausencia de respuesta por separado cada vez que construyamos una tabla. Algunas referencias para la imputación son Sande (1983) y Kalton y Kasprzyk (1982; 1986).

EJEMPLO 8.7 La encuesta de población actual CPS tiene una tasa de respuesta familiar global alta (por lo general arriba de 90%), aunque algunas familias se niegan a contestar ciertas preguntas. La tasa de respuesta es cercana a 20% en muchas preguntas de ingresos. Esta ausencia de respuestas crearía un sesgo sustancial en cualquier análisis, a menos que se tome alguna acción correctiva: varios estudios sugieren que la ausencia de respuesta por elemento, para los elementos de ingreso, es mayor para las familias de bajos y altos ingresos. La imputación de los datos faltantes permite utilizar técnicas estadísticas estándar, como la regresión, sin que el analista deba enfrentar la ausencia de respuesta con métodos desarrollados de manera particular. Si hay que hacer alguna imputación, para las encuestas como CPS, la oficina que reúne los datos tiene más información que la guíe para llenar los valores faltantes de la que tendría un analista independiente, pues la identificación de la información no aparece en las cintas de uso público.

La CPS utiliza la ponderación para ajustar la falta de entrevista y la imputación hot-deck para la ausencia de respuestas por elemento. La muestra se divide en clases usando las variables sexo, edad, raza y otras características demográficas. Si falta un elemento, se sustituye un elemento correspondiente de otra unidad en esa clase. Por lo general, la imputación hot-deck se realiza tomando el valor del elemento faltante de una familia similar a la familia con el elemento faltante respecto de alguna otra variable explicativa, como el tamaño de la familia. ■

En la tabla 8.3 usamos un pequeño conjunto de datos para ilustrar algunos de los métodos para la imputación. Este conjunto artificial de datos sólo se usa para fines de ilustración; en la práctica, se necesita un conjunto de datos mayor para la imputación. Un "1" indica que la persona respondió "sí" a la pregunta.

TABLA 8.3
Pequeño conjunto de datos utilizado para ilustrar los métodos de imputación

Persona	Edad	Sexo	Años de escolaridad	¿Víctima de delito?	¿Víctima de delito violento?
1	47	M	16	0	0
2	45	F	?	1	1
3	19	M	11	0	0
4	21	F	?	1	1
5	24	M	12	1	1
6	41	F	?	0	0
7	36	M	20	1	?
8	50	M	12	0	0
9	53	F	13	0	?
10	17	M	10	?	?
11	53	F	12	0	0
12	21	F	12	0	0
13	18	F	11	1	?
14	34	M	16	1	0
15	44	M	14	0	0
16	45	M	11	0	0
17	54	F	14	0	0
18	55	F	10	0	0
19	29	F	12	?	0
20	32	F	10	0	0

8.6.1 Imputación deductiva

Algunos valores se pueden asignar en la edición de datos, mediante las relaciones lógicas entre las variables. En la tabla 8.9, la persona 9 no respondió si había sido víctima de algún delito violento, pero como respondió que no había sido víctima de delito alguno, la respuesta de delito violento debe cambiar a 0.

En ocasiones, la *imputación deductiva* puede utilizarse en las encuestas longitudinales. Si una mujer tiene dos hijos en cierto año y dos hijos dos años después, pero falta el valor en el año intermedio, el valor lógico por asignar sería 2.

8.6.2 Imputación de la media por celda

Las personas que responden se dividen en clases (celdas) con base en variables conocidas, como en los ajustes de clases de ponderación. Entonces, sustituimos el promedio de los valores de las unidades que responden y que están en la celda c, $\bar{y}_{cR}$, en cada valor faltante. La *imputación de la media por celda* supone que los elementos faltantes son faltantes completamente al azar (MCAR) dentro de las celdas.

EJEMPLO 8.8 Construimos las cuatro celdas de nuestro ejemplo usando las variables edad y sexo. (En la práctica, por supuesto, usted querrá tener mucho más individuos en cada celda.)

		Edad	
		≤ 34	≥ 35
Edad	M	Personas 3, 5, 10, 14	Personas 1, 7, 8, 15, 16
	F	Personas 4, 12, 13, 19, 20	Personas 2, 6, 9, 11, 17, 18

Las personas 2 y 6, que no tienen el valor para los años de escolaridad, recibirán el valor promedio de las cuatro personas con 35 o más años de edad y que contestaron la pregunta: 12.25. La media para cada celda después de la imputación es igual a la media de quienes respondieron. Sin embargo, el valor imputado no es una de las respuestas posibles a la pregunta sobre la escolaridad. ■

La imputación de la media proporciona las mismas estimaciones puntuales para las medias, los totales y las proporciones relacionadas con los ajustes de clases de ponderación. Sin embargo, los métodos de imputación de la media no reflejan la variabilidad de quienes no responden, pues todas las observaciones faltantes en una clase reciben el mismo valor imputado. La distribución de y tendrá una distorsión, debido a un "pico" en el valor de la media muestral de quienes responden; en consecuencia, la varianza estimada en la subclase también será muy pequeña.

Para evitar ese pico, se podría utilizar una imputación estocástica de la media por celda. Si la variable de respuesta tuviese aproximadamente una distribución normal, los valores faltantes podrían asignarse mediante un valor generado de manera aleatoria a partir de una distribución normal con media $\bar{y}_{cR}$ y desviación estándar s_{cR}.

La imputación de la media, estocástica o asignada de cualquier otra forma, distorsiona las relaciones entre las diversas variables, pues la imputación se realiza por separado para cada elemento faltante. Las correlaciones y otras estadísticas de la muestra se modifican. Jinn y Sedransk (1980a; 1989b) analizan el efecto de los distintos métodos de imputación sobre el análisis secundario de los datos; por ejemplo, para estimar una pendiente de regresión.

8.6.3 Imputación hot-deck

En la *imputación hot-deck*, al igual que en la imputación de la media por celda y los métodos de ajuste con ponderación, las unidades de la muestra se dividen en clases. El valor de una de las unidades de la clase y que responde se sustituye en cada respuesta faltante. Con frecuencia, los valores de un conjunto de elementos faltantes relacionados entre sí se toman del mismo donante para preservar algunas de las relaciones multivariadas. El nombre *hot deck* proviene de los días en que los programas de computadora y los conjuntos de datos se perforaban en tarjetas: el conjunto de éstas con el conjunto de datos por analizar se "calentaba" en la lectora de tarjetas, de modo que se empleaba el término *hot deck* para referirse a las imputaciones realizadas mediante el mismo conjunto de datos. Fellegi y Holt (1976) analizan los métodos de edición de datos y la imputación hot-deck con encuestas grandes.

¿Cómo elegir la unidad donante? Existen varios métodos.

Imputación hot-deck secuencial Algunos procedimientos de imputación hot-deck asignan el valor del mismo subgrupo leído por última vez por la computadora. Esto es consecuencia de las días de uso de tarjetas en las computadoras (la imputación se puede realizar con un único paso) y también es, parcialmente, la suposición de que, si los datos se ordenan de alguna manera geográfica, las unidades adyacentes del mismo subgrupo tenderán a ser más similares que las unidades elegidas al azar en el subgrupo. Un problema del uso del valor en la "tarjeta" anterior es que con frecuencia, quienes no responden también tienden a aparecer en conglomerados, de modo que alguien podría ser donante varias veces, de una manera no controlable por quien extraiga la muestra. Por lo general, se usa alguno de los otros métodos de imputación hot-deck en la mayoría de las encuestas.

En nuestro ejemplo, la persona 19 no tiene la respuesta para saber si fue víctima de algún delito. La persona 13 tiene la última respuesta registrada en su subclase, de modo que se asigna el valor 1.

Imputación hot-deck aleatoria Se elige un donante al azar entre las personas de la celda, con la información de todos los elementos faltantes. Para preservar las relaciones multivariadas, generalmente se emplean los valores del mismo donante en todos los elementos faltantes de una persona.

En nuestro pequeño conjunto de datos, el individuo 10 no tiene los valores de ser o no víctima. Las personas 3, 5 y 14 de su celda tienen respuestas en ambas preguntas, de modo que una de ellas se elige al azar como donante. En este caso, se elige la 14 y sus valores se asignan a ambas variables faltantes.

Imputación hot-deck del vecino más cercano Se define una medida de distancia entre las observaciones y se asigna el valor de la persona que respondió y está "más cerca" del individuo con el elemento faltante, donde la cercanía se define mediante la función distancia.

Si la edad y el sexo se utilizan para la función distancia, de modo que se elija como donante la persona con edad más cercana y el mismo sexo. Las respuestas de ser víctima o no de la persona 3 se asignan a la persona 10.

8.6.4 Imputación por regresión

La *imputación por regresión* predice el valor faltante usando una regresión del elemento de interés sobre las variables observadas para todos los casos. Una variante es la *imputación por regresión estocástica*, donde el valor faltante se reemplaza mediante el valor predicho a partir del modelo de regresión, más un término de error generado aleatoriamente.

Sólo tenemos 18 observaciones completas para las respuestas de calidad de víctima (en realidad, no son bastantes para ajustar un modelo a nuestro conjunto de datos), pero una regresión logística de la respuesta con la edad como variable de explicación proporciona el siguiente modelo para la probabilidad predicha de ser una víctima, $\hat{p}$:

$$\log\frac{\hat{p}}{1-\hat{p}} = 2.5643 - 0.0896 \times \text{edad}.$$

La probabilidad predicha de ser víctima de un delito para una persona de 17 años de edad es 0.74; como esto es mayor que un truncamiento predeterminado de 0.5, se asigna el valor 1 a la persona 10.

EJEMPLO 8.9 Paulin y Ferraro (1994) analizan los modelos de regresión para la asignación del ingreso en la encuesta de gastos del consumidor de Estados Unidos. Para la parte de entrevistas de la encuesta se entrevista a las familias seleccionadas cada trimestre, durante cinco trimestres consecutivos. En cada entrevista, se les pide que recuerden sus gastos durante los tres meses anteriores. Los datos se utilizan para relacionar los gastos del consumidor con características tales como el tamaño y los ingresos de las familias; son la fuente de informes en el sentido de que los gastos exceden al ingreso en determinadas clases de ingreso.

La encuesta de gastos del consumidor realiza casi 5000 entrevistas anualmente, en comparación con casi 60,000 para la NCVS. Este tamaño de muestra es demasiado pequeño para los métodos de imputación hot-deck, pues en una muestra pequeña es menos probable hallar donantes adecuados para quienes no responden. Hay que adoptar un modelo paramétrico si debe hacerse una imputación. Paulin y Ferraro usaron modelos de regresión múltiple para predecir el logaritmo del ingreso familiar (se utilizan logaritmos pues la distribución del ingreso es asimétrica) a partir de variables de explicación, entre las que se encuentran los gastos totales y las variables demográficas. Estos modelos suponen que los elementos de ingreso son MAR, dados los covariados. ■

8.6.5 Imputación cold-deck

En la *imputación cold-deck*, los valores se asignan a partir de una encuesta anterior o de otras informaciones, como datos históricos. (ya que el conjunto de datos que sirve como fuente para la asignación no es el que está actualmente en la computadora, el conjunto está "frío".) Hay poca teoría para este método. Como en el caso de la imputación hot-deck, la imputación cold-deck no garantiza la eliminación del sesgo de selección.

8.6.6 Sustitución

Los métodos de *sustitución* son similares a la imputación cold-deck. En ocasiones, se permite a los entrevistadores elegir un sustituto cuando está en el campo; si la familia seleccionada para la muestra no está en casa, se pasa a la siguiente puerta. La sustitución puede ayudar a reducir parte del sesgo por la ausencia de respuesta, ya que la siguiente familia podría ser más similar a la que no responde que otra elegida al azar entre la población. Pero la familia de la siguiente puerta es alguien que responde; si la ausencia de respuesta está relacionada con las características de interés, seguirá habiendo un sesgo por ausencia de respuesta. Un problema adicional es que, como el entrevistador puede elegir la familia, la muestra ya no tendrá probabilidades conocidas de selección.

La encuesta de abuso de drogas realizada en Michigan en 1975 debía estimar la cantidad de personas que utilizaron 16 tipos de sustancias durante el año anterior. El diseño de la muestra era el de una muestra estratificada de varias etapas, con 2100 familias. Se realizaron tres llamadas telefónicas a una casa; luego se intentaba con la de la derecha y posteriormente con la de la izquierda. Los datos muestran que la tasa de uso de drogas aumenta al incrementar la cantidad de llamadas necesarias.

Algunas encuestas seleccionan a los sustitutos designados al mismo tiempo en que se seleccionan las unidades de la muestra. Si alguna unidad no responde, entonces se elige al azar a uno de los sustitutos designados. El estudio longitudinal de Estados Unidos (véase National Center of Educational Statistics 1977) utilizó este método. Esta muestra estratificada de varias etapas de los grupos que concluían el bachillerato de 1972 pretendía proporcionar datos acerca de las experiencias, planes y actitudes relacionadas con la educación en este sector de estudiantes. Se eligieron al azar cuatro bachilleratos en cada uno de los 600 estratos. Se designaron dos para la muestra y los otros dos se conservaron como respaldo en caso de ausencia de respuesta. De las 1200 escuelas designadas para la muestra, 948 participaron, 21 no tenían estudiantes en el último año y 231 se rehusaron o no pudieron participar. Los investigadores eligieron 122 instituciones del grupo de respaldo para sustituir a las que no respondieron. Los estudios de seguimiento mostraron un sesgo consistente de 5% en varios de los totales estimados, lo que se atribuyó al uso de escuelas sustituto y a la ausencia de respuesta.

La sustitución tiene el peligro adicional de que los esfuerzos por establecer contacto con las unidades designadas pueden no ser tan grandes como cuando no hay opciones. Es necesario informar en los resultados si se emplea sustitución.

8.6.7 Imputación múltiple

En la *imputación múltiple*, cada valor faltante se asigna m (≥ 2) veces distintas. Generalmente, se usa el mismo modelo estocástico para cada imputación, lo que crea m conjuntos de "datos" distintos, sin valores faltantes. Cada uno de los m conjuntos de datos se analiza como si no hubiera imputación; los diversos resultados dan al analista una medida de la varianza adicional debida a la imputación. La imputación múltiple con diversos modelos para ausencia de respuesta puede dar una idea de la sensibilidad a la ausencia de respuesta de los resultados a los modelos particulares. Vea Rubin (1987; 196) para los detalles de implantación de la imputación múltiple.

8.6.8 Ventajas y desventajas de la imputación

La imputación crea un conjunto rectangular y "limpio" de datos que puede analizar el software estándar. Los análisis de los diversos subconjuntos de datos producirán resultados consistentes. Si la ausencia de respuesta es MAR dados los covariados utilizados en el procedimiento de imputación, la imputación reduce sustancialmente el sesgo debido a la ausencia de respuesta. Si algunas partes de los datos son confidenciales, quien los reúne puede realizar la imputación. Esta persona tiene más información de la muestra y la población que el público (por ejemplo, podría saber la dirección exacta de cada miembro de la muestra) y frecuentemente, mediante esa información puede hacer una mejor asignación.

El mayor peligro del uso de la imputación es que un análisis futuro de los datos no distinguirá entre los valores originales y los asignados. Lo ideal es que quien realice la imputación registre cuáles fueron las observaciones asignadas, cuántas veces se empleó

como donante cada registro (no imputado) y qué donante se utilizó para una respuesta específica asignada a un receptor. Los valores imputados pueden ser buenas estimaciones, pero no son datos reales.

Las varianzas calculadas usando los datos junto con los valores asignados son siempre demasiado pequeñas, en parte debido al incremento artificial del tamaño de la muestra y en parte debido a que los valores asignados se consideran como realmente obtenidos en la recolección de datos. La varianza real será mayor que la estimada mediante un paquete normal de software. Rao (1996) y Fay (1996) analizan métodos para estimar las varianzas después de la imputación.

8.7 Modelos paramétricos para la ausencia de respuesta*

La mayoría de los métodos para enfrentar la ausencia de respuesta suponen que ésta es *ignorable*; es decir, que la ausencia de respuesta es independiente de las variables de interés, condicionalmente sobre los covariados medidos. En esta situación, en vez de simplemente dividir las unidades entre las distintas subclases y ajustar los pesos, se puede ajustar un modelo de superpoblación. Así, de este modelo, se predicen los valores de las y que no están en la muestra. Con frecuencia, el ajuste del modelo es iterativo.

En un enfoque completamente basado en el modelo, desarrollamos un modelo para los datos completos y le agregamos componentes para tomar en cuenta el mecanismo propuesto de ausencia de respuesta. Este punto de vista tiene ciertas ventajas sobre otros métodos: El enfoque de modelado es flexible y puede servir para incluir cualquier dato relativo al mecanismo de ausencia de respuesta, el modelador debe establecer explícitamente en el modelo las hipótesis relativas a la ausencia de respuesta y es posible evaluar algunas de estas hipótesis. Además, si el modelo es bueno las estimaciones de la varianza resultantes del ajuste del modelo toman en cuenta la ausencia de respuesta.

EJEMPLO 8.10 Muchas personas creen que los búhos moteados de Washington, Oregon y California están en peligro de extinción debido a que la explotación de los bosques de coníferas reduce su hábitat. Se necesitan buenas estimaciones del tamaño de la población de estos búhos para un debate razonado del tema.

En el plan de muestreo descrito por Azuma *et al* (1990), una región de interés se divide en N regiones de muestreo (las unidades primarias) y se selecciona una muestra aleatoria simple de n unidades primarias. Sea $Y_i = 1$ si la unidad i la ocupa una pareja de búhos y 0 en caso contrario. Suponga que las Y_i son independientes y que $P(Y_i = 1) = p$, la verdadera proporción de unidades primarias ocupadas. Si fuera posible determinar con exactitud la ocupación de cada unidad primaria, la proporción de unidades primarias podría estimarse mediante la proporción muestral $\bar{y}$. Sin embargo, aunque una cantidad fija de visitas puede establecer si una unidad primaria está ocupada; la determinación de que una unidad primaria está desocupada podría ser incorrecta pues algunas parejas de búhos podrían "no responder", de modo que al ignorar la ausencia de respuesta es probable que se obtenga una estimación demasiado baja del porcentaje de ocupación.

Azuma *et al* (1990) proponen el uso de una distribución geométrica para la cantidad de visitas necesarias para descubrir los búhos en una unidad ocupada, modelando así la ausencia de respuesta. Las hipótesis del modelo son (1) la probabilidad de determinar la ocupación en la primera visita, η, es la misma para todas las unidades primarias, (2) cada visita a una unidad primaria es independiente y (3) las visitas pueden continuar hasta avistar un búho. La distribución geométrica se utiliza generalmente para la cantidad de llamadas telefónicas necesarias en las encuestas a personas (vea Potthoff *et al* 1993).

Sea X_i la cantidad de visitas necesarias para determinar si la unidad primaria i está ocupada o no. Bajo el modelo geométrico,

$$P(X_i = x \mid Y_i = 1) = \eta(1-\eta)^{x-1} \quad \text{para } x = 1,2,3,\ldots$$

Sin embargo, el presupuesto del servicio forestal de Estados Unidos no permite muchas visitas. Suponga que se permite un máximo de s visitas a cada unidad primaria. La variable aleatoria Y_i no puede observarse; las variables aleatorias observables son

$$V_i = \begin{cases} k \text{ si } Y_i = 1, X_i = k, \text{ y } X_i \le s. \\ 0 \text{ en caso contrario.} \end{cases}$$

$$U_i = \begin{cases} 1 \text{ si } Y_i = 1 \text{ y } X_i \le s. \\ 0 \text{ en caso contrario.} \end{cases}$$

En este caso, $\sum_{i\in S} U_i$ cuenta la cantidad de unidades primarias que se observaron ocupadas y $\sum_{i\in S} V_i$ cuenta el número total de visitas realizadas a las unidades ocupadas. Bajo el modelo geométrico, la probabilidad de que un búho sea vea por primera vez en la unidad primaria i durante la visita k ($\le s$) es

$$P(V_i = k) = \eta(1-\eta)^{k-1} p,$$

y la probabilidad de distinguir un búho en una de las s visitas a la unidad primaria i es

$$P(U_i = 1) = E[U_i] = [1-(1-\eta)^s]p.$$

Así, el valor esperado de la proporción muestral de unidades ocupadas, $E[\overline{U}]$, es igual a $[1-(1-\eta)^s]\,p$ y es menor que la proporción de interés si $\eta < 1$. El modelo geométrico coincide con la suposición de que los búhos no se visitan en las s visitas.

Determinamos las estimaciones de máxima verosimilitud de p y η bajo la hipótesis de que todas las unidades primarias son independientes. La función de verosimilitud

$$(\eta p)^{\sum_i u_i}(1-\eta)^{\sum_i (v_i - u_i)}[1-p+p(1-\eta)^2]^{n-\sum_i u_i}$$

se maximiza cuando

$$\hat{p} = \frac{\overline{u}}{1-(1-\hat{\eta})^2}$$

y cuando $\hat{\eta}$ es solución de

$$\frac{\overline{v}}{\overline{u}} = \frac{1}{\eta} - \frac{s(1-\eta)^s}{1-(1-\eta)^s};$$

se necesitan métodos numéricos para calcular $\hat{\eta}$. La teoría de máxima verosimilitud también permite calcular la muestra de covarianza asintótica de las estimaciones de los parámetros.

Una muestra aleatoria simple de 240 unidades primarias del hábitat en California tuvo los siguientes resultados:

Número de visita	1	2	3	4	6	5
Cantidad de unidades primarias ocupadas	33	17	12	7	7	5

Se vio que un total de 81 unidades primarias estaban ocupadas en seis visitas, de modo que $\overline{u} = 81/240 = 0.3375$. La cantidad promedio de visitas realizadas a las unidades ocupadas fue de $\overline{v}/\overline{u} = 196/81 = 2.42$. Así, las estimaciones de máxima verosimilitud son $\hat{\eta} = 0.334$ y $\hat{p} = 0.370$; usamos la matriz de covarianza asintótica de la teoría de máxima

verosimilitud para estimar la varianza de $\hat{p}$ como 0.00137. De esta manera, un intervalo de confianza aproximado de 95% para la proporción de unidades ocupadas es 0.370 ± 0.072.

La incorporación del modelo geométrico para la cantidad de visitas dio una mayor estimación de la proporción de unidades ocupadas. Sin embargo, si el modelo no describe los datos, la estimación $\hat{p}$ seguirá siendo sesgada; si el modelo es pobre, $\hat{p}$ puede ser una estimación de la tasa de ocupación peor que $\bar{u}$. El modelo geométrico sería inadecuado si, por ejemplo, los investigadores de campo tuvieran mayor probabilidad de encontrar los búhos en visitas posteriores debido a que acumulan más información de dónde buscar.

Necesitamos verificar si el modelo geométrico describe adecuadamente la cantidad de visitas necesarias para determinar la ocupación. Desgraciadamente no podemos determinar si el modelo describe la situación de las unidades donde no se detectan búhos en seis visitas, pues nos faltan datos. Sin embargo, podemos utilizar una prueba χ^2 de bondad de ajuste para ver si los datos de las seis visitas realizadas se ajustan por el modelo. De acuerdo con él, esperamos que en $n\eta(1-\eta)^{k-1}p$ de las unidades primarias se hayan observado búhos en la visita k, e introducimos nuestras estimaciones de p y h para calcular las cifras esperadas:

Visita	Cifra observada	Cifra esperada
1	33	29.66
2	17	19.74
3	12	13.14
4	7	8.75
5, 6	12	9.71
Total	81	80.99

Combinamos las visitas 5 y 6 en una categoría, de modo que la cifra esperada en la celda sea mayor que 5. La estadística de la prueba χ^2 es 1.75, con valor $p > 0.05$. No hay indicios de que el modelo sea inadecuado para los datos que tenemos, aunque tampoco podemos verificar que sea adecuado para los datos faltantes. El modelo geométrico supone que las observaciones son independientes y que es posible determinar la ocupación de una unidad primaria ocupada después de un número suficiente de visitas. No podemos verificar si esta hipótesis del modelo es razonable o no: si algunos taimados búhos nunca pudieran detectarse en cualquier número de visitas, $\hat{p}$ seguirá siendo demasiado pequeña. ■

Para utilizar modelos con ausencia de respuesta, se necesita (1) un profundo conocimiento de la estadística matemática, (2) una computadora poderosa y (3) un conocimiento de métodos numéricos para la optimización. Por lo general, los métodos de máxima verosimilitud se utilizan para estimar los parámetros y las ecuaciones de verosimilitud rara vez tienen soluciones cerradas. El cálculo de las estimaciones requiere de métodos numéricos, aun para el sencillo modelo adoptado para los búhos; fue una muestra aleatoria simple con un modelo geométrico sencillo para el mecanismo de respuesta lo que nos permitió escribir la función de verosimilitud. Las funciones de verosimilitud para diseños de muestreo más complejos o los mecanismos de ausencia de respuesta son mucho más difíciles de construir (en particular si las observaciones del mismo conglomerado se consideran como dependientes) y el cálculo de las estimaciones requiere gran cantidad de cálculos. Little y Rubin (1987) analizan los métodos basados en la verosimilitud para los datos faltantes en general. Stasny (1991) da un ejemplo del uso de modelos para tomar en cuenta la ausencia de respuesta.

8.8 ¿Qué es una tasa de respuesta aceptable?

Con frecuencia, el investigador dirá: "Espero una tasa de respuesta de 60% en mi encuesta. ¿Es esto aceptable? ¿Me dará la encuesta resultados válidos?" Como hemos visto en este capítulo, la contestación depende de la naturaleza de la ausencia de respuesta: si quienes no responden son MCAR, entonces podemos ignorar ampliamente la ausencia de respuesta y utilizar a quienes contestan como una muestra representativa de la población. Si quienes no responden tienden a diferir de quienes sí, entonces los sesgos de los resultados al usar sólo a quienes sí lo hicieron hará que toda la encuesta pierda su valor.

Muchas referencias proporcionan consejos acerca de las cuotas para la aceptación de las tasas de respuesta. Babbie, por ejemplo, opina: "creo que una tasa de respuesta de al menos 50% es *adecuada* para el análisis y los informes; una respuesta de al menos 60% es *buena* y una tasa de respuesta de 70% es *muy buena*" (1973, 165). Creo que el establecimiento de tales criterios absolutos para las tasas de respuesta aceptables es peligroso y ha llevado a muchos investigadores de encuestas a una complacencia infundada acerca de la ausencia de respuesta; hay muchos ejemplos de encuestas con una tasa de respuesta de 70% cuyos resultados son erróneos. La NCVS necesita corregir el sesgo por la ausencia de respuesta incluso con una tasa de respuesta cercana a 95%.

Se debe tomar en cuenta que las tasas de respuesta se pueden manipular al definirlas de otra manera. Con frecuencia, los investigadores no dicen cómo se calculó la tasa de respuesta o podrían utilizar una estimación de dicha tasa que sea menor de lo que debería. Muchas encuestas inflan la tasa de respuesta eliminando las unidades del denominador que no pudieron localizarse. Se acumulan muchos resultados distintos para la tasa de respuesta, según la definición utilizada; todas las definiciones siguientes se han empleado en las encuestas:

$$\frac{\text{cantidad de entrevistas concluidas}}{\text{cantidad de unidades en la muestra}}$$

$$\frac{\text{cantidad de entrevistas concluidas}}{\text{cantidad de unidades contactadas}}$$

$$\frac{\text{entrevistas concluidas} + \text{unidades inelegibles}}{\text{unidades contactadas}}$$

$$\frac{\text{entrevistas concluidas}}{\text{unidades contactadas} - (\text{unidades inelegibles})}$$

$$\frac{\text{entrevistas concluidas}}{\text{unidades contactadas} - (\text{unidades inelegibles}) - \text{rechazos}}$$

Observe que una "tasa de respuesta" calculada con la última fórmula será mucho mayor que la calculada mediante la primera fórmula, pues el denominador es menor.

Los criterios para reportar las tasas de respuesta en Statistics Canada (1993) y Hidiroglou *et al* (1993) proporcionan una solución sensible para reportar las tasas de respuesta. Definen las *unidades al alcance* como aquellas que pertenecen a la población objetivo y las *unidades resueltas* como aquellas unidades para las que se sabe pertenecen o no a la población objetivo.[3] Ellos sugieren que se informe de varias tasas de respuesta para una encuesta,

[3] Si, por ejemplo, la población objetivo son los números telefónicos residenciales, podría ser imposible decir si un teléfono que suena pero no es contestado pertenece a la población objetivo; tal número sería una *unidad no resuelta*.

incluyendo lo siguiente:

- Tasa fuera de alcance: el cociente del número de unidades fuera de alcance entre la cantidad de unidades al alcance y no resueltas.
- Tasa de no contactos: el cociente del número de no-contactos y las unidades no resueltas entre la cantidad de unidades al alcance y no resueltas
- Tasa de rechazo: el cociente del número de rechazos entre la cantidad de unidades al alcance
- Tasa de ausencia de respuesta: el cociente del número de quienes no responden y las unidades no resueltas entre la cantidad de unidades al alcance y no resueltas

Las distintas medidas de la ausencia de respuesta pueden ser adecuadas para distintas encuestas y yo dudaría en recomendar una definición de tasa de respuesta que se ajuste a todos los casos. Sin embargo, las cantidades utilizadas para calcular la tasa de respuesta deben quedar definidas en cada encuesta. Las siguientes recomendaciones de la Oficina de Administración y Presupuesto del Comité Federal sobre Metodología Estadística de Estados Unidos, incluidas en González *et al* (1994), son de utilidad:

> Recomendación 1. Los equipos de la encuesta deben calcular las tasas de respuesta de una manera uniforme con respecto del tiempo y documentar los componentes de la tasa de respuesta en cada edición de una encuesta.
>
> Recomendación 2. Para encuestas repetidas, los equipos de la encuesta deben revisar los componentes de la tasa de respuesta (como los rechazos, no estar en casa, fuera del alcance, dirección no localizable, regreso del correo, etcétera), junto con la documentación de rutina acerca de los cambios de costo y diseño.
>
> Recomendación 3. Los componentes de la tasa de respuesta deben publicarse en los informes de la encuesta; los lectores pueden recibir las definiciones de las tasas de respuesta utilizadas, incluyendo las cifras reales, así como comentarios sobre la importancia de las tasas de respuesta para la calidad de los datos de la encuesta.
>
> Recomendación 4. Un poco de investigación acerca de la ausencia de respuesta puede tener sus dividendos. Los administradores de la encuesta deben apoyarla como una forma de mejorar la eficacia de las operaciones de recolección de datos.

8.9 Ejercicios

1 Ryan *et al* (1991) reportan los resultados de una encuesta nacional de madres realizada por correo por Ross Laboratories, para investigar la alimentación infantil en Estados Unidos. Se enviaron cuestionarios a madres con hijos de seis meses de edad, preguntándoles el tipo de leche que dieron a sus hijos durante los primeros seis meses de vida y de ciertas variables socioeconómicas. Los autores afirman que la cantidad de cuestionarios enviados aumentó de 1984 a 1989: "En 1989 se enviaron 56,894 cuestionarios y regresaron 30,694. En 1989 se enviaron 196,000 y se recibieron 89,640". Las familias con bajos ingresos tienen un exceso de muestreo en el diseño de la muestra, dada su menor tasa de respuesta. Las personas que respondieron se dividieron en subclases definidas por la región, las cuestiones étnicas, la edad y la escolaridad; se calcularon los pesos con base en la información de la Oficina de Censos.

a ¿Qué se utilizó? ¿Ajustes de clases de ponderación o estratificación posterior?

b El exceso de muestreo de las familias con bajos ingresos es una forma de sustitución. ¿Cuáles son las ventajas y las desventajas del uso de la sustitución en esta encuesta?

c Las cifras ponderadas son "comparables con las publicadas por la Oficina de Censos de Estados Unidos y el Centro Nacional de Estadísticas de Salud" en cuanto a las cuestiones étnicas, la edad de la madre, el ingreso, la escolaridad, el empleo, el peso al nacer, la región y participación en el programa de apoyo nutricional a mujeres, bebés y niños. Con estas cifras ponderadas, los investigadores estimaron que cerca de 53% de las madres tenían un hijo, mientras que los datos gubernamentales indican que cerca de 43% de las madres tiene un hijo. ¿La coincidencia de las cifras ponderadas con las estadísticas oficiales indica que la ponderación corrige el sesgo por la ausencia de respuesta? Explique.

d Analice el uso de la ponderación en esta encuesta. ¿Podría pensar en algunas formas de mejorar esto?

2 Unos investigadores seleccionaron una muestra aleatoria simple de 200 estudiantes del último año de bachillerato de una población de 2000 para una encuesta de hábitos de uso de la televisión, con una tasa de respuesta global de 75%. Al verificar los registros escolares, pudieron determinar el promedio de calificaciones (GPA) para quienes no responden y clasificaron la muestra de acuerdo con esto:

GPA	Tamaño de la muestra	Número de personas que respondieron	Horas de televisión $\bar{y}$	Horas de televisión s_y
3.00–4.00	75	66	32	15
2.00–2.99	72	58	41	19
Menor de 2.00	53	26	54	25
Total	200	150		

a ¿Cuál es la estimación para el número promedio de horas de televisión empleadas semanalmente si sólo se analiza a quienes respondieron? ¿Cuál es el error estándar de la estimación?

b Realice una prueba χ^2 para la hipótesis nula de que los tres grupos de GPA tienen las mismas tasas de respuesta. ¿Qué puede concluir? ¿Qué dicen sus resultados acerca del tipo de datos faltantes? ¿Cree que sus datos son MCAR? ¿MAR? ¿No ignorables?

c Realice un análisis de varianza en un sentido para verificar la hipótesis nula de que los tres grupos de GPA tienen el mismo nivel medio de uso de la televisión. ¿Qué puede concluir? ¿Indica este análisis de varianza que GPA puede ser una buena variable para construir las celdas de ponderación? ¿Por qué?

d Utilice la clasificación GPA para ajustar los pesos de quienes responden en la muestra. ¿Cuál es la estimación por clase de ponderación del tiempo promedio de uso de la televisión?

e Las cifras de la población son de 700 estudiantes con un GPA entre 3 y 4; 800 estudiantes con un GPA entre 2 y 3; y 500 estudiantes con un GPA menor a 2. Use estas cifras de la población para construir una estimación estratificada posteriormente del tiempo promedio de uso de la televisión.

f ¿Qué otros métodos podría emplear para realizar ajustes debidos a la ausencia de respuesta?

g ¿Qué otras variables podría reunir y utilizar en los modelos con ausencia de respuesta?

3 La siguiente descripción y evaluación de la ausencia de respuesta es de un estudio de Hamilton, Ontario, acerca de la actitud de los dueños de casas en relación con sanitarios de composta:

> La encuesta se realizó por medio de un cuestionario autoadministrado, enviado por correo. Se mandaron 1200 cuestionarios a una muestra elegida al azar de propietarios de casas. Una

semana después se giraron cartas de agradecimiento como seguimiento. En total, regresaron 329 cuestionarios, lo que representa una tasa de respuesta de 27%. Eso se consideró satisfactorio, pues muchos encuestadores por correo consideran una tasa de respuesta de 15 a 20% como buena. (Wynia *et al* 1993, 362)

¿Está de acuerdo en que la tasa de respuesta de 27% es satisfactoria? Suponga que los investigadores se acercan a usted en búsqueda de una asesoría estadística para analizar estos datos y diseñar una encuesta de seguimiento. ¿Qué les diría?

4 Kosmin y Lachman (1993) incluyeron una pregunta sobre la afiliación religiosa en 56 encuestas familiares realizadas cada semana; el tema de las encuestas varió de una semana a otra, desde el uso de la televisión por cable, hasta la preferencia por ciertos artículos de consumo, pasando por temas políticos. Después de cuatro llamadas, la tasa de respuesta por unidad fue de 50%; un 2.3% adicional se rehusó a contestar la pregunta sobre religión. Los autores sostienen:

> "Desde una perspectiva nacional, la cantidad de entrevistas transparentes y el cuidadoso diseño de la investigación garantizaron un alto nivel de precisión... Las estimaciones del error estándar para nuestra muestra nacional global indica que podemos tener 95% de confianza en que las cifras obtenidas tienen un margen de error, más o menos, de menos del 0.2%. Esto significa, por ejemplo, que estamos más que 95% seguros de que la cifra de los católicos está en el rango de 25.0% a 26.4% para la población de Estados Unidos. (página 284)

a Critique la afirmación anterior.

b Si usted hubiera anticipado la ausencia de respuesta por elementos, ¿cree que sería mejor insertar la pregunta de interés en distintas encuestas cada semana, como aquí se hizo, o usar el mismo conjunto de preguntas adicionales en cada encuesta? Justifique su respuesta. ¿Cómo diseñaría un experimento para probar su conjetura?

5 Encuentre un ejemplo de encuesta en un periódico o revista popular. ¿Se da la tasa de respuesta? En tal caso, ¿cómo se calculó? ¿Cómo cree que afectaría la ausencia de respuesta a las conclusiones de la encuesta? Dé sugerencias acerca de la forma en que el periodista podría enfrentar los problemas de la ausencia de respuestas en el artículo.

6 Encuentre un ejemplo de encuesta en una revista académica. ¿Cómo calcularon los autores la tasa de ausencia de respuesta? ¿Cómo enfrenta la encuesta a la ausencia de respuesta? ¿Cómo cree que afectaría la ausencia de respuesta a las conclusiones del estudio? ¿Cree que los autores toman en cuenta, en forma adecuada, los sesgos potenciales por la ausencia de respuesta? ¿Qué sugerencias daría para estudios posteriores?

7 En el ejercicio 20 del capítulo 4 se mencionó brevemente el tema de la ausencia de respuesta en la encuesta de clausura del periodo vacacional de invierno (en el archivo winter.dat). ¿Qué modelo se adopta para la ausencia de respuesta cuando las fórmulas del muestreo estratificado se usan para estimar la proporción de empleados de la universidad que contestarían afirmativamennte a la pregunta "¿Quiere tener de nuevo una clausura del periodo vacacional de invierno?" ¿Cree que éste sea un modelo razonable? ¿Cómo modelaría adicionalmente los efectos de la ausencia de respuesta en esta encuesta? ¿Qué información adicional podría reunir para ajustar la ausencia de respuesta por unidad?

8 Un tema de discusión entre la comunidad estadística de Estados Unidos en los últimos años es si la Asociación Estadística de EU (ASA) debería ofrecer un proceso de certificación para sus miembros, de modo que los estadísticos que cumplan las condiciones puedan ser designados como "estadísticos certificados". En 1994, la ASA realizó una encuesta entre sus miembros acerca de este tema (los datos están en el archivo certify.dat). La encuesta se

envió a la totalidad de 18,609 miembros, recibiendo 5001 respuestas. Los resultados de la encuesta se reportaron en la edición de octubre de 1994 de *Amstat News*.

Suponga que en 1994, la membresía de la ASA tenía las siguientes características: 55% tienen doctorado y 38% maestría; 29% trabajan en la industria, 34% en la academia y 11% en el gobierno. No se dispone de la clasificación cruzada entre escolaridad y lugar de trabajo.

a ¿Cuáles son las tasas de respuesta para las distintas subclases de membresía a la ASA? ¿Son MCAR las personas que no responden? ¿Cree que sean MAR?

b Use el rastrilleo para ajustar los pesos de las seis celdas definidas por la escolaridad (doctorado o no doctorado) y el lugar de trabajo (industria, academia, otros). Comience con un peso inicial de 18,609/5001 para cada persona que responde. ¿Qué hipótesis debe establecer para usar el rastrilleo?

Estime la proporción de miembros de ASA que respondieron a cada una de las categorías 0 a 5 (variable *certify*), con y sin los pesos de rastrilleo. Para este ejercicio, puede clasificar los valores faltantes en las categorías "sin doctorado" u "otro lugar de trabajo".

c ¿Cree que los opositores a la certificación tienen la justificación de usar los resultados de esta encuesta para afirmar que la mayoría de los miembros de la ASA se oponen a la certificación? ¿Por qué?

9 La encuesta ACLS del ejemplo 4.3 no tuvo ausencia de respuestas. Calcule la tasa de respuesta en cada estrato de la encuesta. ¿Qué modelo se adoptó para la ausencia de respuesta en el ejemplo 4.3? ¿Existe alguna evidencia de que la tasa de ausencia de respuesta varía entre los estratos o que está relacionada con el porcentaje de membresía femenina?

10 Se usan pesos en la encuesta de jóvenes en custodia (analizado en el ejemplo 7.4) para ajustar la ausencia de respuesta por unidad. Use un procedimiento hot-deck para asignar valores a la variable que mide con quién vivía el joven. ¿qué variables utilizaría para agrupar los datos en clases?

11 Repita el ejercicio 10 con un modelo de imputación por regresión.

12 Repita el ejercicio 10, para la variable *ha utilizado drogas ilegales*.

13 Repita el ejercicio 11, para la variable *ha utilizado drogas ilegales*.

14 La National Science Foundation publicó los resultados de la encuesta de doctores en 1995.[4] ¿Cómo enfrenta esta encuesta la ausencia de respuesta? ¿Cree que el sesgo por ausencia de respuesta sea un problema para esta encuesta?

15 ¿Cómo enfrenta la encuesta criticada en el ejercicio 1 del capítulo 7 a la ausencia de respuestas? En su opinión, ¿enfrentaron adecuadamente los investigadores los problemas debidos a la ausencia de respuesta? ¿Qué sugerencias haría para mejorar el estudio?

16 Conteste las preguntas del ejercicio 15 para la encuesta analizada en el ejercicio 2 del capítulo 7.

[4] "Characteristics of Doctoral Scientists and Engineers in the United States: 1995", NSF Publication 97-319. Es posible conseguir copias gratuitas en Division of Science Resources Studies, National Science Foundation, Arlington, VA, 22230; por correo electrónico mediante pubs@nsf.gov; o a través de Internet (www.nsf.gov/sbe/srs).

17 Gnap (1995) realizó una encuesta acerca de la carga de trabajo de los profesores, utilizada en el ejercicio 16 del capítulo 5.

- **a** La encuesta original pretendía tener una muestra por conglomerados de una etapa. ¿Cuál fue la tasa de respuesta original?
- **b** ¿Esperaría que hubiese un sesgo por ausencia de respuestas en este estudio? En tal caso, ¿en qué dirección esperaría dicho sesgo? ¿Cuáles profesores cree que con menor probabilidad responderían a la encuesta?
- **c** Gnap también reunió los datos de una submuestra aleatoria de quienes no responden en el estrato "grande", en el archivo teachnr.dat. ¿Cuál es la diferencia entre quienes responden y quienes no?
- **d** ¿Existe alguna evidencia de ausencia de respuestas al comparar la submuestra de quienes no respondieron con los que respondieron en la encuesta original?

18 No todos los padres encuestados en el estudio analizado en el ejercicio 17 del capítulo 5 regresaron el cuestionario. En el diseño de muestreo original, se enviaron 50 cuestionarios a los padres de niños en cada escuela, para un tamaño de muestra total planeado de 500. Sabemos que de los 9962 niños que no se vacunaron durante la campaña, la forma de consentimiento no se devolvió para 6698 de tales niños, la forma de consentimiento se devolvió, rechazando la vacuna, para 2061 niños, y 1203 niños, cuyos padres aprobaron la vacunación, no estuvieron presentes ese día.

- **a** Calcule la tasa de respuesta para cada conglomerado. ¿Cuál es la correlación de la tasa de respuesta y el porcentaje de quienes respondieron en la escuela y que regresaron la forma de consentimiento? ¿De la tasa de respuesta y el porcentaje de quienes respondieron en cada escuela y que rechazaron la vacunación?
- **b** En general, 67% (6698/9962) de los padres de la población objetivo no regresaron la forma de consentimiento. Use los datos de quienes respondieron para calcular un intervalo de confianza de 95% para la proporción de padres en la muestra que no regresaron la forma de consentimiento. Calcule otras dos estimaciones del intervalo para esta cantidad: una suponiendo que los valores faltantes son todos 0 y otra suponiendo que los valores faltantes son todos 1. ¿Cuál es la relación entre sus estimaciones y la cantidad de la población?
- **c** Repita la parte (b), examinando el porcentaje de padres que regresaron la forma pero que se negaron a vacunar a sus hijos.
- **d** ¿Cree que la ausencia de respuestas sea un problema para esta encuesta?

9

Estimación de la varianza en encuestas complejas*

Regocíjate, pues debajo de nubes y estrellas
nuestro planeta es más que Maine o Texas.
Bendice los grandiosos hechos de tener
doce meses, nueve musas y dos sexos,
y en los señoríos de la Tierra, infinidad de
artes, climas, maravillas e ideas.

—Phyllis McGinley, "*In Praise of Diversity*"[1]

Utilizando pesos, es fácil estimar las medias y los totales de la población. Usando la técnica de los pesos la estimación de las varianzas es más compleja: en el capítulo 7 observamos que en una encuesta compleja con varios niveles de estratificación y conglomerados, las varianzas de las medias y totales estimados se calculan en cada nivel y luego se combinan conforme avanzamos en el diseño de la encuesta. La estratificación posterior y los ajustes por ausencia de respuesta también afectan a la varianza.

En los capítulos anteriores presentamos y dedujimos fórmulas de la varianza para varios planes de muestreo; algunas de ellas, como las correspondientes a las muestras aleatorias simples, son relativamente sencillas. Otras, como $\hat{V}(\hat{t})$ para una muestra con conglomerados, en dos etapas y sin reemplazo, son más complicadas. Todo esto sirve para estimar las varianzas de los totales estimados pero, con frecuencia, queremos estimar otras cantidades a partir de los datos de la encuesta, para los que no hemos presentado ninguna fórmula para la varianza. Por ejemplo, en el capítulo 3 obtuvimos una varianza aproximada para una razón entre dos medias al extraer una muestra aleatoria simple. ¿Qué pasaría si quisiéramos estimar una razón, pero la encuesta no fuese una muestra aleatoria simple? ¿Cómo estimar la varianza?

Este capítulo describe varios métodos para estimar las varianzas de los totales estimados y otras estadísticas en encuestas complejas. La sección 9.1 describe el método de linealización, generalmente empleado para calcular las varianzas de estadísticas no lineales. Las secciones 9.2 y 9.3 presentan los métodos de grupos aleatorios y remuestreo para calcular las varianzas de estadísticas lineales y no lineales. La sección 9.4 describe el

[1] De *The Love Letters of Phyllis McGinley*, de Phyllis McGinley, Derechos reservados 1951, 1952, 1953, 1954 por Phyllis McGinley. Renovación de derechos reservados © 1979, 1980, 1981, 1982 por Phyllis Hayden Blake. Reproducido con autorización de Viking Penguin, división de Penguin Books USA Inc.

cálculo de las funciones generalizadas de varianza y la sección 9.5 describe la construcción de intervalos de confianza. Estos métodos se describen con más detalle en Wolter (1985) y Rao (1988); Rao (1997) y Rust y Rao (1996), resumen los trabajos recientes en el área.

9.1 Métodos de linealización (series de Taylor)

La mayoría de las fórmulas para la varianza en los capítulos 2 a 6 correspondían a estimaciones de las medias y de los totales y se pueden utilizar para determinar la varianza de cualquier combinación lineal de las medias y totales estimados. Si $\hat{t}_1,\ldots,\hat{t}_k$ son estimaciones insesgadas de k totales en la población, entonces

$$V\left(\sum_{i=1}^{k} a_i\hat{t}_i\right)=\sum_{i=1}^{k} a_i^2 V(\hat{t}_i)+2\sum_{i=1}^{k}\sum_{j=i+1}^{k} a_i a_j \mathrm{Cov}(\hat{t}_i,\hat{t}_j). \qquad (9.1)$$

Este resultado también se puede expresar mediante las estimaciones insesgadas de k medias de la población:

$$V\left(\sum_{i=1}^{k} a_i\hat{\bar{y}}_i\right)=\sum_{i=1}^{k} a_i^2 V(\hat{\bar{y}}_i)+2\sum_{i=1}^{k}\sum_{j=i+1}^{k} a_i a_j \mathrm{Cov}(\hat{\bar{y}}_i,\hat{\bar{y}}_j).$$

Así, si t_1 es la cantidad total declarada como robada por las víctimas de asaltos, t_2 es la cantidad de días de trabajo perdidos por las víctimas a causa del delito y t_3 son los gastos médicos totales realizados por las víctimas, una medida de las consecuencias económicas del robo (suponiendo que se pierden 150 dólares por cada día de trabajo) sería $\hat{t}_1+150\hat{t}_2+\hat{t}_3$. Por (9.1), la varianza es

$$V(\hat{t}_1+150\hat{t}_2+\hat{t}_3)=V(\hat{t}_1)+150^2 V(\hat{t}_2)+V(\hat{t}_3)$$
$$+300\mathrm{Cov}(\hat{t}_1,\hat{t}_2)+2\mathrm{Cov}(\hat{t}_1,\hat{t}_3)+300\mathrm{Cov}(\hat{t}_2,\hat{t}_3).$$

En esta expresión debemos calcular seis varianzas y covarianzas. Para los cálculos, es más fácil definir una nueva variable en el nivel de observación,

$$q_i=y_{i1}+150y_{i2}+y_{i3},$$

y determinar $V(\hat{t}_q)=V\left(\sum_{i\in S}\omega_i q_i\right)$ en forma directa.

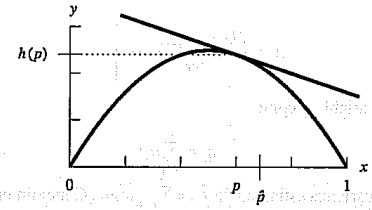
Suponga que nos interesa la proporción de pérdida total correspondiente al bien robado, t_1/t_q. Ésta no es una estadística lineal, pues no se puede expresar en la forma $a_1t_1+a_2t_q$ para ciertas constantes a_i. Sin embargo, el teorema de Taylor del cálculo nos permite **linealizar** una función suave, no lineal $h(t_1,t_2,\ldots,t_k)$ de los totales de la población. El teorema de Taylor nos proporciona las constantes $a_0, a_1,\ldots, a_k$ tales que

$$h(t_1,\ldots,t_k)\approx a_0+\sum_{i=1}^{k} a_i t_i.$$

Así, podemos aproximar $V[h(\hat{t}_1,\ldots,\hat{t}_k)]$ mediante $V\left(\sum_{i=1}^{k} a_i\hat{t}_i\right)$, que sabemos calcular mediante (9.1).

Las aproximaciones con serie de Taylor se han utilizado en estadística durante mucho tiempo para calcular las varianzas aproximadas. Woodruff (1971) ilustra su uso en las encuestas complejas; Binder (1983) da un tratamiento más riguroso de los métodos de serie de Taylor para encuestas complejas e indica cómo usar la linealización cuando el parámetro de interés θ es solución de $h(\theta,t_1,\ldots,t_k)=0$, aunque θ no se pueda expresar como una función explícita de $t_1,\ldots,t_k$.

FIGURA 9.1
La función $h(x)=x(1-x)$, junto con la tangente a la función en el punto p. Si $\hat{p}$ está cerca de p, entonces $h(\hat{p})$ estará cerca de la recta tangente. La pendiente de la recta tangente es $h'(p)=1-2p$.



EJEMPLO 9.1 La cantidad $\theta=p(1-p)$, donde p es una proporción de la población, se puede estimar mediante $\hat{\theta}=\hat{p}(1-\hat{p})$. Suponga que $\hat{p}$ es un estimador insesgado de p y que $V(\hat{p})$ es conocido. Sea $h(x)=x(1-x)$, de modo que $\theta=h(p)$ y $\hat{\theta}=h(\hat{p})$. Ahora h es una función no lineal de x, pero podemos aproximarla en cualquier punto cercano a mediante la recta tangente a la función. La pendiente de la recta tangente está dada por la derivada, como se muestra en la figura 9.1.

La versión de primer orden del teorema de Taylor establece que si la segunda derivada de h es continua, entonces

$$h(x)=h(a)+h'(a)(x-a)+\int_a^x (x-t)h''(t)dt;$$

en condiciones que generalmente se satisfacen en estadística, el último término es pequeño con respecto a los dos primeros, por lo que usamos la aproximación

$$h(\hat{p})\approx h(p)+h'(p)(\hat{p}-p)$$
$$=p(1-p)+(1-2p)(\hat{p}-p).$$

Entonces,

$$V[h(\hat{p})]\approx(1-2p)^2 V(\hat{p}-p),$$

y conocemos $V(\hat{p})$, de modo que podemos calcular la varianza aproximada de $h(\hat{p})$. ■

A continuación damos los pasos básicos para construir un estimador de linealización de la varianza de una función no lineal de las medias o de los totales:

1. Exprese la cantidad de interés como una función de las medias o los totales de las variables medidas o calculadas en la muestra. En general, $\theta=h(t_1,t_2,\ldots,t_k)$ o $\theta=h(\bar{y}_{1U},\ldots,\bar{y}_{kU})$. En el ejemplo 9.1, $\theta=h(\bar{y}_U)=h(p)=p(1-p)$.
2. Determine las derivadas parciales de h con respecto a cada argumento. Estas derivadas, evaluadas en las cantidades de la población, son las constantes de linealización a_i.

3 Aplique el teorema de Taylor para linealizar la estimación:

$$h(\hat{t}_1,\hat{t}_2,\ldots,\hat{t}_k) \approx h(t_1,t_2,\ldots,t_k)+\sum_{j=1}^{k} a_j(\hat{t}_j - t_j),$$

donde

$$a_j = \left.\frac{\partial h(c_1,c_2,\ldots,c_k)}{\partial c_j}\right|_{t_1,t_2,\ldots,t_k}$$

4 Defina la nueva variable q como

$$q_i = \sum_{j=1}^{k} a_j t_{ij}.$$

Ahora, determine la varianza estimada de $\hat{t}_q = \sum_{i\in S}\omega_i q_i$. Generalmente esto aproxima a la varianza de $h(\hat{t}_1,\ldots,\hat{t}_k)$.

EJEMPLO 9.2 Usamos los métodos de linealización para aproximar la varianza de los estimadores de proporción y de regresión en el capítulo 3. En ese capítulo emplemos una muestra aleatoria simple, el estimador $\hat{B} = \bar{y}/\bar{x} = \hat{t}_y/\hat{t}_x$, y la aproximación

$$\hat{B} - B = \frac{\bar{y} - B\bar{x}}{\bar{x}} \approx \frac{\bar{y} - B\bar{x}}{\bar{x}_U} = \sum_{i\in S}\frac{y_i - Bx_i}{n\bar{x}_U}.$$

La aproximación resultante de la varianza era

$$V[\hat{B} - B] \approx \frac{1}{n^2\bar{x}_U^2} V\left[\sum_{i\in S}(y_i - Bx_i)\right].$$

En esencia, usamos el teorema de Taylor para obtener esta aproximación. Los siguientes pasos dan el mismo resultado.

1 Expresamos B como función de los totales de la población. Sea $h(c,d) = d/c$, de modo que

$$B = h(t_x,t_y) = \frac{t_y}{t_x} \text{ y } \hat{B} = h(\hat{t}_x,\hat{t}_y) = \frac{\hat{t}_y}{\hat{t}_x}.$$

Suponga que las estimaciones muestrales $\hat{t}_x$ y $\hat{t}_y$ son insesgadas.

2 Las derivadas parciales son

$$\frac{\partial h(c,\,d)}{\partial c} = \frac{-d}{c^2} \text{ y } \frac{\partial h(c,\,d)}{\partial d} = \frac{1}{c};$$

evaluadas en $c = t_x$ y $d = t_y$, son iguales a $-t_y/t_x^2$ y $1/t_x$.

3 Por el teorema de Taylor,

$$\hat{B} = h(\hat{t}_x,\hat{t}_y)$$
$$\approx h(t_x,t_y) + \left.\frac{\partial h(c,\,d)}{\partial c}\right|_{t_x,t_y}(\hat{t}_x - t_x) + \left.\frac{\partial h(c,\,d)}{\partial d}\right|_{t_x,\,t_y}(\hat{t}_y - t_y).$$

Si usamos las derivadas parciales del paso 2,

$$\hat{B} - B \approx \frac{t_y}{t_x^2}(\hat{t}_x - t_x) + \frac{1}{t_x}(\hat{t}_y - t_y).$$

4 El error cuadrático medio de $\hat{B}$ es

$$E[(\hat{B}-B)^2] \approx E\left[\left\{-\frac{t_y}{t_x^2}(\hat{t}_x - t_x) + \frac{1}{t_x}(\hat{t}_y - t_y)\right\}^2\right]$$
$$= \frac{t_y^2}{t_x^4}V(\hat{t}_x) + \frac{1}{t_x^2}V(\hat{t}_y) - 2\frac{t_y}{t_x^3}\mathrm{Cov}(\hat{t}_x,\hat{t}_y) \qquad (9.2)$$
$$= \frac{1}{t_x^2}\{B^2V(\hat{t}_x) + V(\hat{t}_y) - 2B\mathrm{Cov}(\hat{t}_x,\hat{t}_y)\}.$$

Podemos sustituir los valores estimados para B, las varianzas y la covarianza y tal vez para t_x a partir del esquema particular de muestreo utilizado en (9.2). O bien, podemos definir

$$q_i = \frac{1}{t_x}[y_i - \hat{B}x_i]$$

y determinar $\hat{V}(\hat{t}_q)$.

Si el diseño de muestreo es una muestra aleatoria simple de tamaño n, entonces $V(\hat{t}_x) = N^2(1-n/N)S_x^2/n, V(\hat{t}_y) = N^2(1-n/N)S_y^2/n$, y $\mathrm{Cov}(\hat{t}_x,\,\hat{t}_y) = N^2(1-n/N)RS_xS_y/n$. ■

Ventajas Si conocemos las derivadas parciales, la linealización proporciona casi siempre una estimación de la varianza para una estadística y se puede aplicar en los diseños generales de muestreo. Los métodos de linealización se han utilizado por años en estadística y la teoría está bastante desarrollada. Hay software para el cálculo de las estimaciones de la varianza mediante linealización para muchas funciones no lineales de interés, como los coeficientes de proporción y de regresión. Analizaremos algún software en la sección 9.6.

Desventajas Los cálculos pueden ser complicados y el método puede ser difícil de aplicar para funciones complejas que impliquen la actualización de pesos. Usted debe determinar expresiones analíticas para las derivadas parciales de h o calcular las derivadas en forma numérica. Se necesita otra fórmula de la varianza para cada estadística no lineal estimada, lo que puede requerir mucha programación especial y es necesario un método para cada estadística. Además, no todas las estadísticas se pueden expresar como una función suave de los totales de la población. Por ejemplo, la mediana y otras cantidades no se ajustan a este marco de referencia. La exactitud de la aproximación por linealización depende del tamaño de la muestra; la estimación de la varianza con frecuencia tiene un sesgo hacia abajo si la muestra no es bastante grande.

9.2 Métodos de grupos aleatorios

9.2.1 Réplicas del diseño de la encuesta

Suponga que el diseño básico de la encuesta se copia de manera independiente R veces. En este caso, la *independencia* significa que cada vez que se extraiga una muestra, sus unidades se reemplazan en la población, de modo que estén disponibles para muestras posteriores. Entonces, las R réplicas de la muestra producen R estimaciones independientes de la cantidad de interés; la variabilidad entre esas estimaciones puede servir para estimar la varianza

de $\hat{\theta}$. Mahalanobis (1946) describe los primeros usos del método, que llamó "réplica de redes de unidades de la muestra" y "muestreo interpenetrante".

Sea

θ = parámetro de interés

$\hat{\theta}_r$ = estimación insesgada de θ calculada a partir de la réplica r

$$\tilde{\theta} = \sum_{r=1}^{R} \frac{\hat{\theta}_r}{R}.$$

Si $\hat{\theta}_r$ es una estimación insesgada de θ, también lo es $\tilde{\theta}$, y

$$\hat{V}_1(\tilde{\theta}) = \frac{1}{R} \frac{\sum_{r=1}^{R} (\hat{\theta}_r - \tilde{\theta})^2}{R-1} \tag{9.3}$$

es una estimación insesgada de $V(\tilde{\theta})$. Observe que $\hat{V}_1(\tilde{\theta})$ es la varianza muestral de las R estimaciones independientes de θ, divididas entre R, la estimación usual de la varianza de una media muestral.

EJEMPLO 9.3 *The 1991 Information Please Almanac* indica los costos de inscripción, colegiaturas y hospedaje de todas las instituciones de educación superior de Estados Unidos. Suponga que queremos estimar la proporción entre la colegiatura para no residentes y la colegiatura para residentes de las universidades públicas de Estados Unidos. En una típica puesta en práctica del método de grupos aleatorios, se extraerían muestras independientes usando el mismo diseño, determinando $\hat{\theta}$ para cada muestra. Consideremos cuatro muestras aleatorias simples, cada una de tamaño 10 (tabla 9.1). Las cuatro muestras son sin reemplazo, aunque la misma institución puede aparecer en más de una de las cuatro muestras.

Para este ejemplo,

$$\hat{\theta}_r = \frac{\text{promedio de colegiaturas para no residentes en la muestra } r}{\text{promedio de colegiaturas para residentes en la muestra } r}.$$

Así, $\hat{\theta}_1 = 2.3288, \hat{\theta}_2 = 2.5802, \hat{\theta}_3 = 2.4591$, y $\hat{\theta}_4 = 3.1110$. El promedio muestral de las cuatro estimaciones independientes de θ es $\tilde{\theta} = 2.6198$. La desviación estándar muestral de las cuatro estimaciones es 0.343, de modo que el error estándar de $\tilde{\theta}$ es $0.343/\sqrt{4} = 0.172$. La varianza estimada se basa en cuatro observaciones independientes, de modo que un intervalo de confianza de 95% para la razón en cuestión es

$$2.6198 \pm 3.18(0.172)$$

donde 3.18 es el valor t crítico adecuado con tres grados de libertad. Observe que la pequeña cantidad de réplicas hace que el intervalo de confianza sea más ancho de lo que sería si se considerasen más réplicas, pues la estimación de la varianza con tres grados de libertad no es muy estable. ■

9.2.2 Separación de la muestra en grupos aleatorios

En la práctica, las submuestras no se extraen de manera independiente, aunque la muestra completa se selecciona de acuerdo con el diseño de la encuesta. Entonces, la muestra completa se separa en R grupos, de modo que cada uno de ellos forma una versión en miniatura de la encuesta, que refleja el diseño de la muestra. Luego, los grupos se consideran como réplicas independientes del diseño básico de la encuesta.

TABLA 9.1
Cuatro muestras aleatorias simples de universidades utilizadas en el ejemplo 9.3

Institución	Inscripción	Colegiatura para residentes	Colegiatura para no residentes
Columbus College	3,482	1,348	3,747
Southeastern Massachusetts University	5,354	1,677	4,983
U.S. Naval Academy	4,500	1,500	1,500
Athens State College	1,392	1,080	2,160
University of South Alabama	9,195	1,875	2,475
Virginia State University	3,308	3,071	5,135
SUNY College of Technology–Farmingdale	10,802	1,542	3,950
University of Houston	18,684	930	4,050
CUNY–Lehman College	7,841	1,340	4,140
Austin Peay State University	4,784	1,210	4,166
Promedio	6,934.2	1,559	3,630.6

Institución	Inscripción	Colegiatura para residentes	Colegiatura para no residentes
SUNY–New Paltz	4,696	1,495	4,095
Indiana University–Southeast	4,931	1,350	3,342
University of Wisconsin–Platteville	5,080	1,658	4,740
University of California–Santa Barbara	16,853	1,578	5,799
Weber State College	12,783	1,308	3,513
Kennesaw College	8,404	1,296	3678
South Dakota State University	6,366	1,835	3,363
Dickinson State University	1,402	1,659	4,731
Chadron State College	2,143	1,361	2,036
University of Alaska–Fairbanks	7,028	1,512	3,540
Promedio	6,968.6	1,505.2	3,883.7

Institución	Inscripción	Colegiatura para residentes	Colegiatura para no residentes
University of Alaska–Anchorage	4,091	941	2,765
University of Maine–Fort Kent	594	1,710	4,140
Southern University–Baton Rouge	9,448	1,354	2,876
University of Oregon	13,786	1,782	5,043
Virginia State University	3,308	3,071	5,135
Glenville State College	2,185	1,150	2,900
Winston-Salem State University	2,532	896	4,268
Framingham State College	3,359	1,701	4,729
SUNY–Old Westbury	3,999	1,350	3,292
Northwest Missouri State University	4,600	1,320	2,415
Promedio	4,790.2	1,527.5	3,756.3

Institución	Inscripción	Colegiatura para residentes	Colegiatura para no residentes
Central Washington University	6,398	1,674	5,712
Worcester State College	3,600	1,296	3,792
University of California–Davis	17,202	1,676	7,592
Sam Houston State University	12,359	1,060	4,180
University of Texas–Tyler	2,335	861	3,695
Southeastern Oklahoma State University	3,616	804	1,992
University of Southern Colorado	3,909	1,536	5,275
Penssylvania State University	31,251	3,754	7,900
East Central University	3,606	1,200	4,140
Univ of Arkansas–Monticello	1,854	1,410	3,230
Promedio	8,613	1,527.1	4,750.8

Si tenemos una muestra aleatoria simple de tamaño n, los grupos se forman dividiendo al azar las n observaciones en R grupos, cada uno de tamaño n/R. Estos grupos seudoaleatorios no son totalmente réplicas independientes, pues una unidad de observación sólo puede aparecer en uno de los grupos. No obstante, los grupos pueden considerarse como réplicas independientes si el tamaño de la población es relativamente grande con respecto al tamaño de la muestra. En una muestra por conglomerados, las unidades primarias de muestreo se separan al azar entre los R grupos. La unidad primaria se lleva todas sus unidades de observación al grupo aleatorio, de modo que cada grupo aleatorio sigue siendo una muestra por conglomerados. En una muestra estratificada de varias etapas, un grupo aleatorio contiene una muestra de unidades primarias de cada estrato. Observe que si se extrae una muestra de k unidades primarias en el estrato más pequeño, cuando mucho se pueden formar k grupos aleatorios. Si θ es una cantidad no lineal, en general $\tilde{\theta}$ no será igual a $\hat{\theta}$, el estimador calculado directamente a partir de la muestra completa. Por ejemplo, en la estimación de proporción, $\tilde{\theta} = (1/R)\sum_{r=1}^{R} \bar{y}_r/\bar{x}_r$, mientras que $\hat{\theta} = \bar{y}/\bar{x}$. Por lo general, $\hat{\theta}$ es un estimador más general que $\tilde{\theta}$. En ocasiones se utiliza $\hat{V}_1(\tilde{\theta})$ de (9.3) para estimar $V(\hat{\theta})$, aunque es una estimación con exceso. Otro estimador de la varianza es ligeramente mayor, pero se usa frecuentemente:

$$\hat{V}_2(\hat{\theta}) = \frac{1}{R}\frac{\sum_{r=1}^{R}(\hat{\theta}_r - \hat{\theta})^2}{R-1}. \tag{9.4}$$

EJEMPLO 9.4 La encuesta de jóvenes en custodia de 1987, analizada en el ejemplo 7.4, se dividió en siete grupos aleatorios; el diseño de la encuesta tenía 16 estratos. Cada uno de los estratos 6 a 16 consistía de una instalación de seguridad (unidad primaria); cada una de ellas participaba en la muestra con probabilidad 1. En los estratos 1 a 5, las instalaciones se seleccionaron con una probabilidad proporcional a la cantidad de residentes en el censo de 1985 para menores en custodia.

Se deseaba que cada grupo aleatorio fuese una miniatura del diseño de muestreo. Para cada instalación autorrepresentada en los estratos 6 a 16, los números de grupo aleatorio se asignaron así: el primer residente seleccionado en la instalación recibía un número entre 1 y 7. Digamos que al primer residente correspondía el 6. Entonces, el segundo residente de esa instalación recibía el 7; el tercero, el 1; el cuatro, el 2, y así sucesivamente. En los estratos 1 a 5, todos los residentes de una instalación (unidad primaria) se asignaban al mismo grupo aleatorio. Así, para las siete instalaciones de la muestra en el estrato 2, todos los residentes de la instalación 33 se asignaban al grupo aleatorio número 1, todos los residentes de la instalación 9 al grupo aleatorio número 2, etcétera. Se formaron siete grupos aleatorios, pues cada uno de los estratos 2 a 5 tienen siete unidades primarias.

Después de realizar todas las asignaciones a los grupos aleatorios, cada "uno de ellos tenía el mismo diseño básico de la muestra original. Por ejemplo, el grupo aleatorio forma una muestra estratificada en la que se extrae una muestra (casi) aleatoria de residentes entre las instalaciones autorrepresentadas en los estratos 6 a 16 y se extrae una muestra con ppt (probabilidad proporcional al tamaño) de las instalaciones en cada uno de los estratos 1 a 5.

Para usar el método de grupos aleatorios y estimar una varianza, calculamos $\hat{\theta}$ para cada grupo aleatorio. Las siguiente tabla muestra las estimaciones de la edad promedio de los residentes en cada grupo aleatorio; cada estimación se calculó mediante

$$\hat{\theta}_r = \frac{\sum w_i y_i}{\sum w_i},$$

donde w_i es el peso final para el residente i y la suma se realiza sobre todas las observaciones en el grupo aleatorio r.

Número de grupo aleatorio	Estimación de la edad promedio, $\hat{\theta}_r$
1	16.55
2	16.66
3	16.83
4	16.06
5	16.32
6	17.03
7	17.27

Las siete estimaciones $\hat{\theta}_r$ se consideran como observaciones independientes, de modo que

$$\tilde{\theta} = \frac{1}{7}\sum_{r=1}^{7}\hat{\theta}_r = 16.67$$

y

$$\hat{V}_1(\tilde{\theta}) = \frac{1}{7}\frac{\sum_{r=1}^{7}(\hat{\theta}_r - \tilde{\theta})^2}{6} = \frac{0.1704}{7} = 0.024.$$

Usando todo el conjunto de datos, calculamos $\hat{\theta} = 16.64$ con

$$\hat{V}_2(\hat{\theta}) = \frac{1}{7}\frac{\sum_{r=1}^{7}(\hat{\theta}_r - \hat{\theta})^2}{6} = \frac{0.1716}{7} = 0.025.$$

Podemos utilizar $\tilde{\theta}$ o $\hat{\theta}$ para calcular los intervalos de confianza; si usamos $\hat{\theta}$, un intervalo de confianza de 95% para la edad promedio es

$$16.64 \pm 2.45\sqrt{0.025} = [16.3,\ 17.0]$$

(2.45 es el valor crítico t con seis grados de libertad). ■

Ventajas No se requiere software especial para estimar la varianza y es fácil calcular la estimación de la varianza. El método es adecuado para los problemas multiparamétricos o no paramétricos; puede servir para estimar varianzas de percentiles y de funciones no suaves, así como para varianzas de funciones suaves de los totales de la población. Los métodos de grupos aleatorios se utilizan fácilmente después de los ajustes de ponderación para la ausencia de respuesta y la subcobertura.

Desventajas Frecuentemente, la cantidad de grupos aleatorios es pequeña, lo que da estimaciones imprecisas de las varianzas. Por lo general, habrá que tener por lo menos 10 grupos aleatorios para obtener una estimación más estable de la varianza y para no inflar el intervalo de confianza usando la distribución t en lugar de la distribución normal. El establecimiento de los grupos aleatorios puede ser difícil en los diseños complejos, ya que cada grupo debe tener la misma estructura de diseño que la encuesta completa. El diseño de la encuesta puede limitar la cantidad de grupos aleatorios que puedan construirse; si se seleccionan dos unidades primarias en cada estrato, entonces sólo se pueden formar dos grupos aleatorios.

9.3 Métodos de remuestreo y réplicas

Los métodos de grupos aleatorios son fáciles de calcular y explicar, pero son inestables si una encuesta compleja sólo se puede separar en una pequeña cantidad de grupos. Los métodos de remuestreo consideran la muestra como si en sí misma fuese una población; extraemos distintas muestras de esta nueva "población" y usamos las submuestras para estimar una varianza. Todos los métodos de esta sección calculan las estimaciones de la varianza para una muestra en la cual se extrae una muestra de unidades primarias con reemplazo. Si las unidades primarias se extraen sin reemplazo, los métodos pueden ser utilizados pero es de esperar que sobrestimen la varianza y produzcan intervalos de confianza conservadores.

9.3.1 Réplica repetida balanceada (RRB)

Algunas encuestas están tan estratificadas que sólo se seleccionan dos unidades primarias de cada estrato. Esto es el máximo grado de estratificación posible que permite el cálculo de las estimaciones de la varianza en cada estrato.

9.3.1.1 RRB en una muestra aleatoria estratificada

Ilustraremos la RRB para un problema que ya sabemos resolver: el cálculo de la varianza para $\bar{y}_{est}$ de una muestra aleatoria estratificada. En la sección 9.3.1.2 analizamos estadísticas más complicadas a partir de muestras estratificadas de varios niveles.

Suponga que elegimos una muestra aleatoria simple de dos unidades de observación en cada uno de siete estratos. Etiquetamos arbitrariamente una de las unidades de la muestra en el estrato h como y_{h1} y la otra como y_{h2}. Los valores de la muestra aparecen en la tabla 9.2.

La estimación estratificada de la media de la población es

$$\bar{y}_{est} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = 4451.7.$$

Si ignoramos la corrección para poblaciones finitas en la ecuación (4.5) obtenemos la estimación de la varianza

$$\hat{V}_{est}(\bar{y}_{est}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h};$$

donde $n_h = 2$, como aquí, $s_h^2 = (y_{h1} - y_{h2})^2/2$, de modo que

$$\hat{V}_{est}(\bar{y}_{est}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{(y_{h1} - y_{h2})^2}{4}.$$

Aquí, $\hat{V}_{est}(\bar{y}_{est}) = 55{,}892.75$. Esto puede estimar en exceso la varianza, si el muestreo es sin reemplazo.

TABLA 9.2
Una pequeña muestra aleatoria estratificada, utilizada para ilustrar RRB

Estrato	$\frac{N_h}{N}$	y_{h1}	y_{h2}	$\bar{y}_h$	$y_{h1} - y_{h2}$
1	.30	2,000	1,792	1,896	208
2	.10	4,525	4,735	4,630	−210
3	.05	9,550	14,060	11,805	−4,510
4	.10	800	1,250	1,025	−450
5	.20	9,300	7,264	8,282	2,036
6	.05	13,286	12,840	13,063	446
7	.20	2,106	2,070	2,088	36

Para emplear el método de grupos aleatorios, seleccionaríamos al azar una de las observaciones de cada estrato para el grupo 1 y la otra al grupo 2. Los grupos de esta situación son mitades de muestra. Por ejemplo, el grupo 1 podría constar de $\{y_{11}, y_{22}, y_{32}, y_{42}, y_{51}, y_{62}, y_{71}\}$ y el grupo 2 de las otras siete observaciones. Entonces,

$$\hat{\theta}_1 = (.3)(2000) + (.1)(4735) + \cdots + (.2)(2106) = 4824.7,$$

y

$$\hat{\theta}_2 = (.3)(1792) + (.1)(4525) + \cdots + (.2)(2070) = 4078.7.$$

La estimación de la varianza mediante grupos aleatorios (en este caso, 139, 129) sólo tiene un grado de libertad para el diseño de dos unidades primarias por estrato y en la práctica es inestable. Si hacemos una asignación distinta de las observaciones a los grupos; por ejemplo, si el grupo 1 consta de y_{h1} para los estratos 2, 3 y 5 y y_{h2} para los estratos 1, 4, 6 y 7, entonces $\hat{\theta}_1 = 4508.6$, $\hat{\theta}_2 = 4394.8$, y la estimación de la varianza mediante grupos aleatorios hubiera sido 3238.

McCarthy (1966; 1969) observa que en total pueden formarse 2^H mitades de muestra posibles y sugiere el uso de una muestra balanceada de las 2^H mitades de muestra posibles para estimar la varianza. La **réplica repetida balanceada** utiliza la variabilidad entre R réplicas de mitades de muestra seleccionadas de manera equilibrada para estimar la varianza de $\hat{\theta}$.

Para definir el equilibrio, introducimos la siguiente notación. La mitad de muestra r puede definirse mediante un vector $\alpha_r = (\alpha_{r1}, \ldots, \alpha_{rH})$: sea

$$y_h(\alpha_r) = \begin{cases} y_{h1} \text{ si } \alpha_{rh} = 1. \\ y_{h2} \text{ si } \alpha_{rh} = -1. \end{cases}$$

En forma equivalente,

$$y_h(\alpha_r) = \frac{\alpha_{rh} + 1}{2} y_{h1} - \frac{\alpha_{rh} - 1}{2} y_{h2}.$$

Si, como antes, el grupo 1 contiene las observaciones $\{y_{11}, y_{22}, y_{32}, y_{42}, y_{51}, y_{62}, y_{71}\}$, entonces $\alpha_1 = (1,-1,-1,-1,1,-1,1)$. De manera análoga, $\alpha_2 = (-1,1,1,1,-1,1,-1)$. El conjunto de R réplicas de mitades de muestra está **balanceado** si

$$\sum_{r=1}^{R} \alpha_{rh} \alpha_{rl} = 0 \quad \text{para toda } l \neq h.$$

Sea $\hat{\theta}(\alpha_r)$ la estimación de interés, calculada de la misma forma que $\hat{\theta}$ pero usando sólo las observaciones de la mitad de la muestra seleccionadas mediante α_r. Para estimar la

media de una muestra estratificada, $\hat{\theta}(\alpha_r) = \sum_{h=1}^{H}(N_h/N)y_h(\alpha_r)$. Definimos el estimador de varianza RRB como

$$\hat{V}_{\text{RRB}}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}[\hat{\theta}(\alpha_r) - \hat{\theta}]^2.$$

Si el conjunto de mitades de muestra está balanceado, entonces $\hat{V}_{\text{RRB}}(\bar{y}_{\text{est}}) = \hat{V}_{\text{est}}(\bar{y}_{\text{est}})$. (Dejamos la demostración de este hecho para el ejercicio 6.) Si, además, $\sum_{r=1}^{R}\alpha_{rh} = 0$ para $h = 1,\ldots,H$, entonces $\frac{1}{R}\sum_{r=1}^{R}\bar{y}_{\text{est}}(\alpha_r) = \bar{y}_{\text{est}}$.

Para nuestro ejemplo, el conjunto de a de la siguiente tabla cumple la condición de equilibrio $\sum_{r=1}^{8}\alpha_{rh}\alpha_{rl} = 0$, para toda $l \neq h$. La matriz de 8 × 7 de 1 y −1 tiene columnas ortogonales; de hecho, es la matriz de diseño (excluyendo la columna de unos) para un diseño factorial fraccionario (Box *et al* 1978). Los diseños descritos por Plackett y Burman (1946) dan matrices con k columnas ortogonales, donde k es un múltiplo de 4; Wolter (1985) enumera explícitamente algunas de estas matrices.

		Estrato (h)						
		1	2	3	4	5	6	7
Mitad de muestra (r)	α_1	−1	−1	−1	1	1	1	−1
	α_2	1	−1	−1	−1	−1	1	1
	α_3	−1	1	−1	−1	1	−1	1
	α_4	1	1	−1	1	−1	−1	−1
	α_5	−1	−1	1	1	−1	−1	1
	α_6	1	−1	1	−1	1	−1	−1
	α_7	−1	1	1	−1	−1	1	−1
	α_8	1	1	1	1	1	1	1

La estimación para cada mitad de muestra, $\hat{\theta}(\alpha_r) = \bar{y}_{\text{est}}(\alpha_r)$ se calcula a partir de los datos de la tabla 9.2.

Mitad de muestra	$\hat{\theta}(\alpha_r)$	$[\hat{\theta}(\alpha_r) - \hat{\theta}]^2$
1	4732.4	78,792.5
2	4439.8	141.6
3	4741.3	83,686.2
4	4344.3	11,534.8
5	4084.6	134,762.4
6	4592.0	19,684.1
7	4123.7	107,584.0
8	4555.5	10,774.4
Promedio	4451.7	55,892.8

El promedio de $[\hat{\theta}(\alpha_r) - \hat{\theta}]^2$ para las ocho réplicas de mitades de muestra es 55,892.75, que es igual a $\hat{V}_{\text{est}}(\bar{y}_{\text{est}})$ para el muestreo con reemplazo. Observe que podemos establecer la estimación RRB anterior creando una nueva variable de pesos para cada réplica de mitad de muestra. El peso de muestreo para la observación i en el estrato h es $w_{hi} = N_h/n_h$, y

$$\bar{y}_{\text{est}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{2}w_{hi}y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{2}w_{hi}}.$$

En la RRB con una muestra aleatoria estratificada, eliminamos una de las dos observaciones del estrato h para calcular $y_h(\alpha_r)$. En compensación, duplicamos el peso de la observación restante. Definimos

$$w_{hi}(\alpha_r) = \begin{cases} 2w_{hi} & \text{si la observación } i \text{ del estrato } h \text{ está en} \\ & \text{la mitad de muestra seleccionada por } \alpha_r. \\ 0 & \text{en caso contrario.} \end{cases}$$

Entonces

$$\bar{y}_{\text{est}}(\alpha_r) = \frac{\sum_{h=1}^{H}\sum_{i=1}^{2}w_{hi}(\alpha_r)y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{2}w_{hi}(\alpha_r)}.$$

Del mismo modo, para cualquier estadística $\hat{\theta}$ calculada mediante los pesos w_{hi}, calculamos $\hat{\theta}(\alpha_r)$ exactamente de la misma forma, pero empleando los nuevos pesos $w_{hi}(\alpha_r)$. Si usamos las nuevas variables ponderadas en lugar de seleccionar el subconjunto de observaciones, facilitamos los cálculos de encuestas con muchas variables de respuesta: podemos utilizar la misma columna $w_{hi}(\alpha_r)$ para determinar la r-ésima estimación de la mitad de la muestra para todas las cantidades de interés. Los pesos modificados también facilitan la extensión del método a las muestras estratificadas de varias etapas.

9.3.1.2 RRB en una encuesta estratificada de varias etapas

Cuando $\bar{y}_U$ es la única cantidad de interés en una muestra aleatoria estratificada, RRB simplemente es un buen método de calcular la varianza en la ecuación (4.5) y añade poco al procedimiento del capítulo 4. La importancia de RRB en una encuesta compleja proviene de su habilidad para estimar la varianza de una cantidad general de la población θ, donde θ puede ser la razón entre dos variables, un coeficiente de correlación, un cuantil o alguna otra cantidad de interés.

Suponga que la población tiene H estratos y que se seleccionan dos unidades primarias del estrato h, con probabilidades diferentes y con reemplazo. (En los métodos con réplicas, queremos utilizar el muestreo con reemplazo, pues el diseño de submuestreo no afecta al estimador de la varianza, como vimos en la sección 6.3.) Podemos emplear el mismo método al realizar el muestreo sin reemplazo en cada estrato, pero es de esperar que la varianza estimada de $\hat{\theta}$, calculada bajo la hipótesis de muestreo con reemplazo, sea mayor que la varianza sin reemplazo.

El archivo de datos para una encuesta compleja con dos unidades primarias por estrato se parece al de la tabla 9.3, después de organizar por estrato y unidad primaria.

El vector α_r define la mitad de muestra r: si $\alpha_{rh} = 1$, entonces, todas las unidades de observación de la unidad primaria 1 del estrato h están en la mitad de muestra r; si $\alpha_{rh} = -1$, entonces, todas las unidades de observación de la unidad primaria 2 del estrato h están en la mitad de muestra r. Los vectores α_r se seleccionan de manera balanceada, exactamente como en el muestreo aleatorio estratificado. Ahora, para la mitad de muestra r, creamos una nueva columna de pesos $w(\alpha_r)$:

$$w_i(\alpha_r) = \begin{cases} 2w_i & \text{si la unidad de observación } i \text{ está en la mitad de muestra } r. \\ 0 & \text{en caso contrario.} \end{cases} \quad (9.5)$$

TABLA 9.3
Estructura de datos después de ordenar

Número de observación	Número de estrato	Número de unidad primaria	Número de unidad secundaria	Peso, w_i	Variable de respuesta 1	Variable de respuesta 2	Variable de respuesta 3
1	1	1	1	w_1	y_1	x_1	u_1
2	1	1	2	w_2	y_2	x_2	u_2
3	1	1	3	w_3	y_3	x_3	u_3
4	1	1	4	w_4	y_4	x_4	u_4
5	1	2	1	w_5	y_5	x_5	u_5
6	1	2	2	w_6	y_6	x_6	u_6
7	1	2	3	w_7	y_7	x_7	u_7
8	1	2	4	w_8	y_8	x_8	u_8
9	1	2	5	w_9	y_9	x_9	u_9
10	2	1	1	w_{10}	y_{10}	x_{10}	u_{10}
11	2	1	2	w_{11}	y_{11}	x_{11}	u_{11}
Etc.							

Para la estructura de datos de la tabla 9.3 y $\alpha_{r1} = -1$ y $\alpha_{r2} = -1$, la columna $w(\alpha_r)$ será
$(0,0,0,0,2\omega_5,2\omega_6,2\omega_7,2\omega_8,2\omega_9,2\omega_{10},2\omega_{11},\ldots)$.

Ahora usamos la columna $w(\alpha_r)$ en vez de w para estimar las cantidades para la mitad de muestra r. La estimación del total de la población de y para la muestra completa es $\sum w_i y_i$; la estimación del total de la población de y para la mitad de muestra r es $\sum w_i(\alpha_r) y_i$. Si $\theta = t_y / t_x$, entonces $\hat{\theta} = \sum w_i y_i / \sum w_i x_i$, y $\hat{\theta}(\alpha_r) = \sum w_i(\alpha_r) y_i / \sum w_i(\alpha_r) x_i$. En la sección 7.3 vimos que la función de distribución empírica se calcula mediante los pesos

$$\hat{F}(y) = \frac{\text{suma de } w_i \text{ para todas las observaciones con } y_i \le y}{\text{suma de } w_i \text{ para todas las observaciones}}.$$

Entonces, la distribución empírica usando la mitad de muestra r es

$$\hat{F}_r(y) = \frac{\text{suma de } w_i(\alpha_r) \text{ para todas las observaciones con } y_i \le y}{\text{suma de } w_i(\alpha_r) \text{ para todas las observaciones}}.$$

Si θ es la mediana de la población, entonces podemos definir $\hat{\theta}$ como el menor valor de y para el cual $\hat{F}(y) \ge 1/2$, y $\hat{\theta}(\alpha_r)$ es el menor valor de y para el cual $\hat{F}_r(y) \ge 1/2$.

Para cualquier cantidad θ, definimos

$$\hat{V}_{\text{RRB}}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}[\hat{\theta}(\alpha_r) - \hat{\theta}]^2. \qquad \textbf{(9.6)}$$

La RRB también puede servir para estimar covarianzas de estadísticas: si θ y η son dos cantidades de interés, entonces

$$\widehat{\text{Cov}}_{\text{RRB}}(\hat{\theta},\hat{\eta}) = \frac{1}{R}\sum_{r=1}^{R}[\hat{\theta}(\alpha_r) - \hat{\theta}][\hat{\eta}(\alpha_r) - \hat{\eta}].$$

En el ejercicio 7 se describen otros estimadores de varianza RRB, como variantes de (9.6).

Aunque la equivalencia exacta de $\hat{V}_{\text{RRB}}[\bar{y}_{\text{est}}(\alpha)]$ y $\hat{V}_{\text{est}}(\bar{y}_{\text{est}})$ no se extiende a las estadísticas no lineales, Krewski y Rao (1981) y Rao y Wu (1985) muestran que si h es una función

suave de los totales de la población, la estimación de la varianza a partir de la RRB es asintóticamente equivalente a la estimación correspondiente a la linealización. Al extraer una muestra aleatoria estratificada (Shao y Wu 1992), RRB también proporciona un estimador consistente de la varianza para cuantiles.

EJEMPLO 9.5 Bye y Gallicchio (1993) describen las estimaciones de varianza RRB en la encuesta de ingreso y participación en programas (SIPP) en Estados Unidos. SIPP, como NCVS, tiene un diseño estratificado, por conglomerados, de varias etapas. Los estratos autorrepresentados (AR) constan de una unidad primaria que se extrae en la muestra con probabilidad 1 y para cada estrato no autorrepresentado (NAR) se selecciona una unidad primaria con ppt. Estrictamente, RRB no se aplica, pues sólo se elige una unidad primaria en cada estrato, mientras que RRB requiere dos unidades primarias por estrato. Para usar RRB, se formaron "seudoestratos" y "seudounidades primarias". Un seudoestrato típico se forma combinando un estrato AR con dos estratos NAR similares. La unidad seleccionada en cada estrato NAR fue asignada al azar a una de las dos seudounidades primarias, y los segmentos de la unidad primaria AR se dividieron al azar entre las dos seudounidades primarias. Este procedimiento creó 72 seudoestratos, cada uno con dos seudounidades primarias.

Las 72 mitades de muestra, cada una con las observaciones de una seudounidad primaria de cada seudoestrato, se forman mediante un diseño de Plackett-Burman (1946) de 71 factores. Este diseño es ortogonal, de modo que el conjunto de réplicas de mitades de muestra está balanceado.

Cerca de 8500 de las 54,000 personas de la muestra de 1990 aseguraron recibir los beneficios de la seguridad social; Bye y Gallicchio querían estimar la media y la mediana del monto de beneficio mensual para las personas que recibían beneficios, para varias subpoblaciones. La media de beneficio mensual para los hombres casados fue estimada como

$$\frac{\sum_{i \in S_M} w_i y_i}{\sum_{i \in S_M} w_i},$$

donde y_i es el monto de beneficio mensual para la persona i de la muestra, w_i es el peso asignado a la persona i, y S_M es el subconjunto de la muestra formado por los hombres casados que reciben los beneficios de la seguridad social. La mediana del pago de beneficios se puede estimar a partir de la función de distribución empírica para los hombres casados de la muestra:

$$\hat{F}(y) = \frac{\text{suma de pesos para los hombres casados con } 0 < y_i \le y}{\text{suma de pesos para todos los hombres casados que reciben beneficios}}.$$

La estimación de la mediana de la muestra, $\hat{\theta}$, satisface $\hat{F}(\hat{\theta}) \ge 1/2$, pero $\hat{F}(x) < 1/2$ para toda $x < \hat{\theta}$.

El cálculo de $\hat{\theta}_r$ para una réplica es sencillo: simplemente definimos una nueva variable de peso $w(\alpha_r)$, como ya hemos descrito, y lo usamos en vez de w para estimar la media y la mediana. ■

Ventajas RRB proporciona una estimación de la varianza asintóticamente equivalente a la correspondiente a métodos de linealización para funciones suaves de los totales de la población y para los cuantiles. Requiere relativamente pocos cálculos en comparación con los métodos de la navaja y de *bootstrap*.

Desventajas Según lo definido anteriormente, RRB requiere un diseño de dos unidades primarias por estrato. Sin embargo, en la práctica, con frecuencia se extiende a otros diseños de muestreo mediante esquemas de equilibrio más complejos. RRB, como los métodos de la navaja y *bootstrap*, estima la varianza con reemplazo y puede sobrestimar la varianza si los

9.3.2 La navaja

N_h, la cantidad de unidades primarias en el estrato h de la población, son pequeños.

El método de la **navaja** (jackknife), al igual que RRB, amplía el método de grupos aleatorios permitiendo que las réplicas de los grupos se traslapen. La navaja fue introducida por Quenouille (1949; 1956) como un método para reducir el sesgo; Tukey (1958) lo usó para estimar varianzas y calcular intervalos de confianza. En esta sección describimos el método *navaja con una eliminación*; Shao y Tu (1995) analizan otras formas de la navaja y dan los resultados teóricos.

Para una muestra aleatoria simple, sea $\hat{\theta}_{(j)}$ el estimador de la misma forma que $\hat{\theta}$, pero sin utilizar la observación j. Así, si $\hat{\theta} = \bar{y}$, entonces $\hat{\theta}_{(j)} = \bar{y}_{(j)} = \sum_{i \neq j} y_i/(n-1)$. Para una muestra aleatoria simple, definimos el estimador de navaja con una eliminación (llamado de esta forma pues eliminamos una observación en cada réplica) como

$$\hat{V}_{\text{JK}}(\hat{\theta}) = \frac{n-1}{n}\sum_{j=1}^{n}(\hat{\theta}_{(j)} - \hat{\theta})^2. \tag{9.7}$$

¿Por qué el factor $(n-1)/n$? Veamos qué ocurre con $\hat{V}_{\text{JK}}(\hat{\theta})$ cuando $\hat{\theta} = \bar{y}$:

$$\bar{y}_{(j)} = \frac{1}{n-1}\sum_{i \neq j} y_i = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i - y_j\right) = \bar{y} - \frac{1}{n-1}(y_j - \bar{y}).$$

Entonces,

$$\sum_{j=1}^{n}(\bar{y}_{(j)} - \bar{y})^2 = \frac{1}{(n-1)^2}\sum_{j=1}^{n}(y_j - \bar{y})^2 = \frac{1}{n-1}s_y^2.$$

EJEMPLO 9.6 Así, $\hat{V}_{\text{JK}}(\bar{y}) = s_y^2/n$, la estimación con reemplazo de la varianza de $\bar{y}$.

Usemos la navaja para estimar la razón entre la colegiatura para no residentes y la colegiatura para residentes para el primer grupo de instituciones de la tabla 9.1. En este caso, $\hat{\theta} = \bar{y}/\bar{x}$, $\hat{\theta}_{(j)} = \hat{B}_{(j)} = \bar{y}_{(j)}/\bar{x}_{(j)}$, y

$$\hat{V}_{\text{JK}}(\hat{B}) = \frac{n-1}{n}\sum(\hat{B}_{(j)} - \hat{B})^2.$$

Para cada grupo de navaja, omitimos una observación. Así, $\bar{x}_{(1)}$ es el promedio de todas las x, excepto x_1: $\bar{x}_{(1)} = (1/9)\sum_{i=2}^{9} x_i$ (tabla 9.4).

En este caso, $\hat{B} = 2.3288$, $\sum(\hat{B}_{(j)} - \hat{B})^2 = 0.1043$, y $\hat{V}_{\text{JK}}(\hat{B}) = 0.0938$. ■

TABLA 9.4
Cálculos del método de la navaja para el ejemplo 9.6

j	x	y	$\bar{x}_{(j)}$	$\bar{y}_{(j)}$	$\hat{B}_{(j)}$
1	1365	3747	1580.6	3617.7	2.2889
2	1677	4983	1545.9	3480.3	2.2513
3	1500	1500	1565.6	3867.3	2.4703
4	1080	2160	1612.2	3794.0	2.3533
5	1875	2475	1523.9	3759.0	2.4667
6	3071	5135	1391.0	3463.4	2.4899
7	1542	3950	1560.9	3595.1	2.3032
8	930	4050	1628.9	3584.0	2.2003
9	1340	4140	1583.3	3574.0	2.2573
10	1210	4166	1597.8	3571.1	2.2350

¿Cómo podemos extender esto a una muestra por conglomerados? Podría suponerse que bastaría eliminar una unidad de observación a la vez, pero eso no servirá de nada; pues se destruiría la estructura de conglomerados y daría una estimación de la varianza que sólo es correcta si la correlación entre las clases es igual a cero. En cualquier método de remuestreo y en el método de grupos aleatorios, conserve juntas las unidades de observación dentro de una unidad primaria mientras construye las réplicas, lo que preserva la dependencia entre las unidades de observación dentro de la misma unidad primaria. Así, para una muestra por conglomerados, aplicaríamos el estimador de varianza de navaja en (9.7) de modo que n sea la cantidad de unidades primarias y $\hat{\theta}_{(j)}$ la estimación de θ que se obtendría al eliminar todas las observaciones de la unidad primaria j.

En una muestra por conglomerados, estratificada y con varias etapas, la navaja se aplica por separado en cada estrato en la primera etapa de muestreo, eliminando una unidad primaria a la vez. Suponga que existen H estratos y que se eligen n_h unidades primarias para la muestra del estrato h. Suponga que estas unidades primarias se eligen con reemplazo.

Para aplicar la navaja, eliminamos una unidad primaria a la vez. Sea $\hat{\theta}_{(hj)}$ el estimador de la misma forma que $\hat{\theta}$ al omitir la unidad primaria j del estrato h. Para calcular $\hat{\theta}_{(hj)}$, definimos una nueva variable de ponderación; sea

$$w_{i(hj)} = \begin{cases} w_i & \text{si la unidad de observación } i \text{ no está en el estrato } h. \\ 0 & \text{si la unidad de observación } i \text{ está en la unidad primaria } j \text{ del estrato } h. \\ \dfrac{n_h}{n_h - 1}w_i & \text{si la unidad de observación } i \text{ está en el estrato } h, \text{ pero no en la unidad primaria } j. \end{cases}$$

Entonces usamos los pesos $w_{i(hj)}$ para calcular $\hat{\theta}_{(hj)}$:

$$\hat{V}_{\text{JK}}(\hat{\theta}) = \sum_{h=1}^{H}\frac{n_h - 1}{n_h}\sum_{j=1}^{n_h}(\hat{\theta}_{(hj)} - \hat{\theta})^2. \tag{9.8}$$

EJEMPLO 9.7 Aquí usamos la navaja para calcular la varianza de la media del volumen de un huevo, del ejemplo 5.6. Obtenemos $\hat{\theta} = \hat{y}_r = 4375.947/1757 = 2.49$. En ese ejemplo, como no conocíamos la cantidad de nidadas en la población, calculamos la varianza con reemplazo.

En primer lugar, determinamos el vector de ponderación para cada una de las 184 iteraciones de la navaja. Sólo tenemos un estrato, de modo que $h = 1$ para todas las observaciones. Para $\hat{\theta}_{(11)}$, eliminamos la primera unidad primaria. Así, los nuevos pesos para las observaciones en la primera unidad primaria son 0; los pesos en las restantes unidades primarias son los pesos anteriores multiplicados por $n_h/(n_h - 1) = 184/183$. Al usar los pesos del ejemplo 5.8, las nuevas columnas de pesos según el método de la navaja aparecen en la tabla 9.5. Observe que las sumas de pesos de navaja varían de una columna a otra, pues la muestra original no era autoponderada. Calculamos $\hat{\theta}$ como $\left(\sum w_i y_i\right)/\sum w_i$; para determinar $\hat{\theta}_{(hj)}$, seguimos el mismo procedimiento pero utilizamos $w_{i(hj)}$ en vez de w_i. Así, $\hat{\theta}_{(1,\,1)} = 4349.348/1753.53 = 2.48034$; $\hat{\theta}_{(1,\,2)} = 4345.036/1753.53 = 2.47788$; $\hat{\theta}_{(1,\,184)} = 4357.819/1754.54 = 2.48374$. Usando (9.8), tenemos que $\hat{V}_{\text{JK}}(\hat{\theta}) = 0.00373$. Esto produce un error estándar de 0.061, igual al calculado en el ejemplo 5.6. ■

Ventajas Éste es un método de utilidad general. El mismo procedimiento se utiliza para estimar la varianza de cada estadística donde se pueda usar la navaja. La navaja funciona en muestras estratificadas de varias etapas donde no se puede aplicar la RRB, si en cada estrato se extraen más de dos unidades primarias. La navaja proporciona un estimador consistente

TABLA 9.5
Pesos del método de la navaja para el ejemplo 9.7

nidada	tamaño de la nidada	peso relativo	w(1, 1)	w(1, 2)	...	w(1, 184)
1	13	6.5	0	6.535519	...	6.535519
1	13	6.5	0	6.535519	...	6.535519
2	13	6.5	6.535519	0	...	6.535519
2	13	6.5	6.535519	0	...	6.535519
3	6	3	3.016393	3.016393	...	3.016393
3	6	3	3.016393	3.016393	...	3.016393
4	11	5.5	5.530055	5.530055	...	5.530055
4	11	5.5	5.530055	5.530055	...	5.530055
⋮	⋮	⋮	⋮	⋮		⋮
183	13	6.5	6.535519	6.535519	...	6.535519
183	13	6.5	6.535519	6.535519	...	6.535519
184	12	6	6.032787	6.032787	...	0
184	12	6	6.032787	6.032787	...	0
Suma	3514	1757	1753.53	1753.53	...	1754.54

de la varianza cuando θ es una función suave de los totales de la población (Krewski y Rao 1981).

Desventajas La navaja tiene un desempeño pobre al estimar las varianzas de algunas estadísticas. Por ejemplo, la navaja produce un estimador pobre de la varianza de cuantiles en una muestra aleatoria simple. En general se sabe poco del desempeño de la navaja en los diseños de muestreo sin reemplazo con probabilidades diferentes.

9.3.3 La técnica de *bootstrap*

Como en el caso de la técnica de la navaja, los resultados teóricos de la técnica de **bootstrap** fueron desarrollados para áreas de estadísticas distintas del muestreo de encuestas; Shao y Tu (1995) resumen los resultados teóricos de la técnica de *bootstrap* en muestras de encuestas complejas. Primero describiremos la técnica de *bootstrap* para una muestra aleatoria simple con reemplazo, siguiendo el desarrollo de Efron (1979, 1982) descrito en Efron y Tibshirani (1993). Suponga que $\mathcal{S}$ es una muestra aleatoria simple de tamaño n. Al extraer la muestra, esperamos que reproduzca las propiedades de la población completa. Entonces, consideramos la muestra $\mathcal{S}$ como si fuese una población y obtenemos nuevas muestras a partir de $\mathcal{S}$. Si la muestra realmente es similar a la población (si la función de masa de probabilidad empírica de la muestra es semejante a la función de masa de probabilidad de la población), entonces las muestras generadas a partir de la función de masa de probabilidad empírica se comportarán como muestras de la población.

EJEMPLO 9.8 Utilicemos la técnica de *bootstrap* para estimar la varianza de la altura mediana, θ, en la población de alturas del ejemplo 7.3, usando la muestra del archivo ht.srs. La altura mediana de la población es $\theta = 168$; la media de la muestra en ht.srs es $\hat{\theta} = 169$. La figura 7.2, con la función de masa de probabilidad de la población y la figura 7.3, con el histograma de la muestra, tienen una forma similar (en gran medida porque el tamaño de la muestra aleatoria simple es grande), de modo que sería de esperar que el extraer una muestra aleatoria simple de tamaño n con reemplazo a partir de $\mathcal{S}$ sería similar a extraer una muestra aleatoria simple con reemplazo a partir de la población. Sin embargo, una nueva muestra extraída de $\mathcal{S}$ no sería exactamente como $\mathcal{S}$, pues la nueva muestra utiliza el reemplazo: algunas observaciones en $\mathcal{S}$ pueden aparecer dos o más veces en la nueva muestra, mientras que otras observaciones en $\mathcal{S}$ podrían no aparecer nunca.

Extraemos una muestra aleatoria simple de tamaño 200 con reemplazo de $\mathcal{S}$ para formar la primera remuestra. La primera remuestra de $\mathcal{S}$ tiene una función de masa de probabilidad empírica similar, aunque no idéntica, a la de $\mathcal{S}$; la mediana de la remuestra es $\hat{\theta}_1^* = 170$. Al repetir el proceso, la segunda remuestra de S tiene mediana $\hat{\theta}_2^* = 169$. Extraemos un total de $R = 2000$ remuestras de $\mathcal{S}$ y calculamos la mediana muestra de cada una, obteniendo $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_R^*$. Obtenemos la siguiente tabla de frecuencias para las 2000 medianas muestrales:

Mediana de la remuestra	165.0	166.0	166.5	167.0	167.5	168.0	168.5	169.0	169.5	170.0	170.5	171.0	171.5	172.0
Frecuencia	1	5	2	40	15	268	87	739	111	491	44	188	5	4

La media muestral de estos 2000 valores es 169.3 y la varianza muestral de los 2000 valores es 0.9148; éste es el estimador de la técnica de *bootstrap* para la varianza. Podemos usar la distribución obtenida mediante la técnica de *bootstrap* para calcular en forma directa un intervalo de confianza: como estima la distribución muestral de $\hat{\theta}$, calculamos un intervalo de confianza de 95% determinando los percentiles 2.5 y 97.5 de la distribución dada por la técnica de *bootstrap*. Para esta distribución, un intervalo de confianza de 95% para la mediana es [167.5, 171]. ■

Si la muestra aleatoria simple original es sin reemplazo, Gross (1980) propone la creación de N/n copias de la muestra para formar una "seudopoblación", para luego extraer R muestras aleatorias simples sin reemplazo a partir de la seudopoblación. Si n/N es pequeño, las distribuciones de la técnica de *bootstrap* con y sin reemplazo deben ser similares.

Sitter (1992) describe y compara tres métodos *bootstrap* para encuestas complejas. En todos estos métodos, la técnica de *bootstrap* se aplica dentro de cada estrato. He aquí los pasos para usar una versión de la técnica de bootstrap con reescalamiento de Rao y Wu (1988) para una muestra aleatoria estratificada.

1 Para cada estrato, extraemos una muestra aleatoria simple de tamaño $n_h - 1$ con reemplazo, a partir de la muestra en el estrato h. Haga esto individualmente para cada estrato.

2 Para cada remuestra r ($r = 1, 2, \ldots, R$), cree una nueva variable de ponderación

$$w_i(r) = w_i \times \frac{n_h}{n_h - 1} m_i(r)$$

donde $m_i(r)$ es la cantidad de veces que se selecciona la observación i para estar en la remuestra. Calcule $\hat{\theta}_r^*$, usando los pesos $w_i(r)$.

3 Repita los pasos 1 y 2, R veces, para un número grande R.

4 Calcule

$$\hat{V}_B(\hat{\theta}) = \frac{1}{R-1} \sum_{r=1}^{R} (\hat{\theta}_r^* - \hat{\theta})^2.$$

Ventajas La técnica de *bootstrap* servirá para funciones no suaves (como los cuantiles) en los diseños generales de muestreo. Esta técnica es adecuada para determinar intervalos de confianza en forma directa. Para obtener un intervalo de confianza de 90%, sólo considere los percentiles 5 y 95 a partir de $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_R^*$ o utilice un método t con *bootstrap*, como se describe en Efron (1982).

Desventajas La técnica de *bootstrap* requiere más cálculos que las técnicas de RRB o dè la navaja, pues generalmente R es un número muy grande. En comparación con estas otras dos técnicas, se ha desarrollado menos trabajo teórico acerca de la técnica de bootstrap en diseños generales de muestreo.

9.4 Funciones generalizadas de varianza

En encuestas gubernamentales grandes, como la de población actual (CPS) o la de fuerza de trabajo de Canadá, se calculan y publican cientos y miles de estimaciones. Las agencias que analizaron los resultados podían calcular errores estándar para cada estimación publicada y emitir tablas adicionales de los errores estándar pero eso aumentaría en gran medida la carga de trabajo implicada en la publicación de estimaciones periódicas a partir de las encuestas. Además, otros analistas de las cintas de uso público podrían desear calcular otras estimaciones y esas cintas podrían no proporcionar la información suficiente para calcular los errores estándar.

Varias encuestas proporcionan **funciones generalizadas de varianza** para calcular errores estándar, que se han utilizado en la CPS desde 1947. En este caso, describiremos varias de estas funciones para la NCVS de 1990.

Victimas de delitos en Estados Unidos, 1990 (Departamento de Justicia de Estados Unidos 1992, 146) proporciona fórmulas para las funciones generalizadas de varianza para el cálculo de errores estándar. Si $\hat{t}$ es un número estimado de personas o familias que han sido víctimas de un delito en particular o si $\hat{t}$ estima un número total de incidentes con víctimas,

$$\hat{V}(\hat{t}) = a\hat{t}^2 + b\hat{t}. \tag{9.9}$$

Si $\hat{p}$ es una proporción estimada,

$$\hat{V}(\hat{p}) = \left(\frac{b}{\hat{x}}\right)\hat{p}(1-\hat{p}), \tag{9.10}$$

donde $\hat{x}$ es una población base estimada para la proporción. Para la NCVS de 1990, los valores de a y b eran $a = -0.00001833$ y $b = 3725$. Por ejemplo, se estimó que 1.23% de las personas de entre 20 y 24 años de edad fueron asaltadas en 1990 y que 18,017,100 personas estaban en ese grupo de edad. Así, la estimación mediante la función generalizada de varianza de $\mathrm{EE}(\hat{p})$ es

$$\sqrt{\frac{3725}{18,017,100}(.0123)(1-.0123)} = .0016.$$

Suponiendo que podemos aplicar los resultados asintóticos, esto da un intervalo de confianza aproximado de 95% de $0.0123 \pm (1.96)(0.0016)$, o [0.0091, 0.0153].

En 1990 se realizaron un total estimado de 800, 510 asaltos. Usamos (9.9) para ver que el error estándar de esta estimación es

$$\sqrt{(-.00001833)(800,510)^2 + 3725(800,510)} = 54,499.$$

¿De dónde provienen estas fórmulas? Suponga que t_i es el número total de unidades de observación pertenecientes a una clase (digamos que es la cantidad total de personas en Estados Unidos que fueron víctimas de delitos violentos en 1990). Sea $p_i = t_i/N$ la proporción de personas de la población pertenecientes a esa clase. Si d_i es el efecto de diseño para la estimación de p_i en la encuesta (vea la sección 7.5), entonces

$$V(\hat{p}_i) \approx d_i \frac{p_i(1-p_i)}{n} = \frac{b_i}{N} p_i(1-p_i), \tag{9.11}$$

donde $b_i = d_i \times (N/n)$. De manera análoga,

$$V(\hat{t}_i) \approx d_i N^2 \frac{p_i(1-p_i)}{n} = a_i t_i^2 + b_i t_i,$$

donde $a_i = -d_i/n$. Si al estimar la proporción en un dominio (digamos, la proporción de personas en el grupo de edad 20-24 que fueron víctimas de asaltos), el denominador en (9.11) cambia por el tamaño estimado de la población del dominio (vea la sección 3.3).

Si los efectos de diseño son similares para estimaciones distintas, de modo que $a_i \approx a$ y $b_i \approx b$, entonces podemos estimar las constantes a y b para que (9.9) y (9.10) sean aproximaciones de la varianza para diversas cantidades. El procedimiento general para construir una función generalizada de varianza es el siguiente:

1 Usando las réplicas o cualquier otro método, estime las varianzas de los k totales de la población en cuestión, $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_k$. Sea v_i la varianza relativa para $\hat{V}(\hat{t}_i)/\hat{t}_i^2 = \hat{t}_i, v_i = \widehat{\mathrm{CV}}(\hat{t}_i)^2$, para $i = 1, 2, \ldots, k$.

2 Postule un modelo que relacione v_i con $\hat{t}_i$. Muchas encuestas utilizan el modelo

$$v_i = \alpha + \frac{\beta}{\hat{t}_i}.$$

Éste es un modelo de regresión lineal, con variable de respuesta v_i y variable de explicación $1/\hat{t}_i$. Valliant (1987) determinó que este modelo produce estimaciones consistentes de las varianzas para las clases de modelos de superpoblación estudiados por él.

3 Use técnicas de regresión para estimar α y β. Valliant (1987) sugiere el uso de mínimos cuadrados ponderados para estimar los parámetros, dando mayor peso a los elementos con v_i pequeña. La estimación de la varianza mediante la función generalizada de varianza es entonces el valor predicho a partir de la ecuación de regresión, $a + b/\hat{t}$.

Reemplazamos los valores a_i y b_i para los elementos individuales mediante las cantidades a y b, que se calculan para los k elementos. Para la NCVS de 1990, $b = 3725$. La mayoría de los pesos para la NCVS de 1990 están entre 1500 y 2500; b es aproximadamente igual a (peso promedio) × (efecto de diseño), si el trabajo global de diseño es aproximadamente 2.

Valliant (1987) determinó que si los efectos de diseño para los k totales estimados son similares, las varianzas para las funciones generalizadas de varianza frecuentemente son más estables que la estimación directa, ya que suavizan algunas de las fluctuaciones de un elemento a otro. Sin embargo, si no se pone suficiente interés en el modelo del paso 2, es probable que la estimación de la varianza mediante la función generalizada de varianza sea pobre, y usted sólo puede saber que es pobre si calcula la varianza directamente.

Ventajas Las funciones generalizadas de varianza se pueden utilizar cuando la información de las cintas de uso público es insuficiente para el cálculo directo de los errores estándar. La persona que reúne los datos puede calcular la función generalizada de varianza, pues generalmente posee más información que el público para estimar las varianzas. Una función generalizada de varianza ahorra mucho tiempo y acelera la producción de los informes anuales; también sirve para diseñar encuestas similares en el futuro.

Desventajas El modelo que relaciona v_i con $\hat{t}_i$ podría no ser adecuado para la cantidad de interés, lo que produce una estimación poco confiable de la varianza. Se debe tener cuidado al utilizar las funciones generalizadas de varianza para las estimaciones no incluidas al calcular los parámetros de regresión. Si una subpoblación tiene un grado alto, poco usual de conglomeración (y por tanto un alto efecto de diseño), la estimación de la varianza mediante las funciones generalizadas de varianza puede ser demasiado pequeña.

9.5 Intervalos de confianza

9.5.1 Intervalos de confianza para funciones suaves de los totales de la población

Hay resultados teóricos acerca de la mayoría de los métodos de estimación de la varianza analizados en este capítulo, partiendo de que, bajo ciertas hipótesis, $(\hat{\theta}-\theta)/\sqrt{\hat{V}(\hat{\theta})}$ sigue asintóticamente una distribución normal canónica. Estos resultados y condiciones aparecen en Binder (1983), para las estimaciones por linealización; en Krewski y Rao (1981) y Rao y Wu (1985), para los métodos de RRB y de la navaja; en Rao y Wu (1988) y Sitter (1992) para la técnica de *bootstrap*. En consecuencia, cuando se cumplen esas hipótesis, podemos construir un intervalo de confianza aproximado de 95% para θ como

$$\hat{\theta} \pm 1.96\sqrt{\hat{V}(\hat{\theta})}.$$

Alternativamente, podemos sustituir un percentil t_{gl} por 1.96, donde gl es la cantidad de grupos menos uno, para el método de grupos aleatorios. Rust y Rao (1996) proporcionan criterios para los grados de libertad adecuados para otros métodos.

En forma general, las hipótesis para los métodos de linealización, de la navaja, RRB y de *bootstrap* son los siguientes:

1. La cantidad de interés θ se puede expresar como una función suave de los totales de la población; más precisamente, $\theta = h(t_1, t_2, \ldots, t_k)$, donde las segundas derivadas parciales de h son continuas.
2. Los tamaños de muestra son grandes. Puede ser que la cantidad de unidades primarias en la muestra de cada estrato sea grande o que la encuesta tenga un gran número de estratos (véase Rao y Wu 1985 para las condiciones técnicas precisas necesarias). Además, para construir un intervalo de confianza mediante la distribución normal, los tamaños de muestra deben ser lo bastante grandes como para que la distribución de muestreo de $\hat{\theta}$ sea aproximadamente normal.

Además, varios estudios de simulación indican que estos intervalos de confianza se comportan bien en la práctica. Wolter (1985) resume algunos de los estudios de simulación; otros aparecen en Kovar *et al* (1988) y Rao *et al* (1992). Estos estudios indican que los métodos de linealización y de la navaja tienden a dar estimaciones similares de la varianza, mientras que los procedimientos RRB y *bootstrap* producen estimaciones ligeramente mayores. A veces, se puede usar una transformación de modo que la distribución muestral de una estadística se acerque a una distribución normal: por ejemplo, al estimar el ingreso total, se puede usar una transformación logarítmica debido a que la distribución del ingreso es demasiado asimétrica.

9.5.2 Intervalos de confianza para cuantiles de una población

Los resultados teóricos arriba descritos para los métodos de RRB, de la navaja, de *bootstrap* y de linealización no se aplican a los cuantiles de la población, pues éstos no son funciones suaves de los totales de la población. Se han desarrollado métodos especiales para construir intervalos de confianza para los cuantiles; McCarthy (1993) compara varios intervalos de confianza para la mediana y su análisis se aplica también a otros cuantiles.

Sea q un valor entre 0 y 1. Defina entonces el cuantil θ_q como $\theta_q = F^{-1}(q)$, donde $F^{-1}(q)$ se define como el menor valor y que satisface $F(y) \geq q$. De manera análoga, definimos $\hat{\theta}_q = \hat{F}^{-1}(q)$. Ahora, F^{-1} y $\hat{F}^{-1}$ *no* son funciones suaves, pero suponemos que la población y la muestra son lo bastante grandes como para ser aproximadas mediante funciones continuas.

Algunos de los métodos ya analizados funcionan bastante bien para construir intervalos de confianza para los cuantiles. El método de grupos aleatorios es efectivo si el número de grupos aleatorios, R, es moderado. Sea $\hat{\theta}_q(r)$ el cuantil estimado a partir del grupo aleatorio r. Entonces, un intervalo de confianza para θ_q es

$$\hat{\theta}_q \pm t\sqrt{\frac{\sum_r [\hat{\theta}_q(r) - \hat{\theta}_q]^2}{(R-1)R}}$$

donde t es el percentil adecuado a partir de una distribución t con $R-1$ grados de libertad. De manera análoga, los estudios empíricos de McCarthy (1993), Kovar *et al* (1988), Sitter (1992) y Rao *et al.* (1992) indican que, en ciertos diseños, podemos formar intervalos de confianza usando

$$\hat{\theta}_q \pm 1.96\sqrt{\hat{V}(\hat{\theta}_q)}$$

donde calculamos la estimación de la varianza mediante el método de RRB o de *bootstrap*.

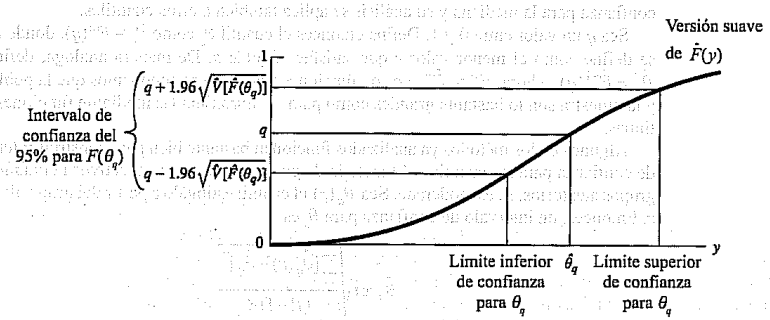
Podemos construir un intervalo alternativo mediante un método introducido por Woodruff (1952). Para cualquier y, $\hat{F}(y)$ es una función de los totales de la población: $\hat{F}(y) = \sum w_i u_i / \sum w_i$, donde $u_i = 1$ si $y_i \leq y$ y $u_i = 0$ si $y_i > y$. Así, podemos usar un método de este capítulo para estimar $V[\hat{F}(y)]$ para cualquier valor y, y un intervalo de confianza aproximado de 95% para $F(y)$ está dado por

$$\hat{F}(y) \pm 1.96\sqrt{\hat{V}[\hat{F}(y)]}.$$

Ahora, usemos el intervalo de confianza de $q = F(\theta_q)$ para obtener un intervalo de confianza aproximado para θ_q. Como tenemos un intervalo de confianza de 95%,

$$\begin{aligned}
0.95 &\approx P\left\{\hat{F}(\theta_q) - 1.96\sqrt{\hat{V}[\hat{F}(\theta_q)]} \leq q \leq \hat{F}(\theta_q) + 1.96\sqrt{\hat{V}[\hat{F}(\theta_q)]}\right\} \\
&= P\left\{q - 1.96\sqrt{\hat{V}[\hat{F}(\theta_q)]} \leq \hat{F}(\theta_q) \leq q + 1.96\sqrt{\hat{V}[\hat{F}(\theta_q)]}\right\} \\
&\approx P\left(\hat{F}^{-1}\left\{q - 1.96\sqrt{\hat{V}[\hat{F}(\theta_q)]}\right\} \leq \theta_q \leq \hat{F}^{-1}\left\{q + 1.96\sqrt{\hat{V}[\hat{F}(\theta_q)]}\right\}\right).
\end{aligned}$$

FIGURA 9.2
Intervalos de confianza de Woodruff para el cuantil θ_q, si la función de distribución empírica es continua. Como $F(y)$ es una proporción, podemos calcular fácilmente un intervalo de confianza para cualquier valor de y, lo que se muestra sobre el eje vertical. Luego buscamos los puntos correspondientes sobre el eje horizontal para formar un intervalo de confianza para θ_q.



Así, un intervalo de confianza aproximado de 95% para el cuantil θ_q es

$$\left[\hat{F}^{-1}\left\{q-1.96\sqrt{\hat{V}[\hat{F}(\hat{\theta}_q)]}\right\},\hat{F}^{-1}\left\{q+1.96\sqrt{\hat{V}[\hat{F}(\hat{\theta}_q)]}\right\}\right].$$

Ilustramos la deducción de este intervalo de confianza en la figura 9.2.

Necesitamos varias hipótesis técnicas para usar el intervalo del método de Woodruff. Estas hipótesis las establecieron Rao y Wu (1987), y Francisco y Fuller (1991), quienes estudiaron un intervalo de confianza similar. Básicamente, el problema es que F y $\hat{F}$ son funciones escalonadas; tienen saltos en valores de y de la población y la muestra. Las condiciones técnicas dicen básicamente que los saltos en F y en $\hat{F}$ deben ser pequeños y que la distribución muestral de $\hat{F}(y)$ es aproximadamente normal.

EJEMPLO 9.9 Utilicemos el método de Woodruff para construir un intervalo de confianza de 95% para la altura mediana en el archivo ht.srs, analizado en los ejemplos 7.3 y 9.8. Observe que $\hat{F}(\theta_q)$ es la proporción muestral de las observaciones en la muestra aleatoria simple que asumen un valor de a lo más θ_q; de modo que al ignorar la corrección para poblaciones finitas,

$$\hat{V}[\hat{F}(\theta_q)] \approx \frac{1}{n}F(\theta_q)[1-F(\theta_q)] = \frac{1}{n}q(1-q).$$

Así, para esta muestra,

$$1.96\sqrt{\hat{V}[\hat{F}(\theta_{0.5})]} \approx 1.96\sqrt{\frac{(.5)(.5)}{200}} = 0.0693.$$

Así, el límite inferior de confianza para la mediana es $\hat{F}^{-1}(.5-0.0693)$, y el límite superior de confianza para la mediana es $\hat{F}^{-1}(.5+0.0693)$. Como las alturas sólo se miden redondeando a centímetros, usaremos la interpolación lineal para suavizar la función escalonada

$\hat{F}$. Obtenemos los siguientes valores para la función de distribución empírica:

y	$\hat{F}(y)$
167	0.405
168	0.440
170	0.515
171	0.550
172	0.605

Luego, al interpolar,

$$\hat{F}^{-1}(0.4307) = 167 + \frac{.4307-.405}{.44-.405}(168-167) = 167.7,$$

y

$$\hat{F}^{-1}(0.5693) = 171 + \frac{.5693-.55}{.605-.55}(172-171) = 171.4.$$

Así, un intervalo de confianza aproximado de 95% para la mediana es [167.7, 171.4]. ■

9.5.3 Intervalos de confianza condicionales

Los intervalos de confianza presentados hasta este punto del capítulo se han desarrollado mediante el enfoque basado en el diseño. Un intervalo de confianza de 95% se puede interpretar, en el sentido del muestreo repetido, como sigue: si las muestras se extrajeran repetidamente en la población finita, sería de esperar que 95% de los intervalos de confianza resultantes incluyan el verdadero valor de la cantidad en la población.

A veces, en particular en situaciones donde se usa la estimación de proporciones o la estratificación posterior, tal vez sea mejor construir un intervalo de confianza condicional. En la estratificación utilizada para la ausencia de respuesta (sección 8.5.2), los tamaños de muestra de las personas que responden n_{hR} en los estratos posteriores son desconocidos al seleccionar la muestra; así, son variables aleatorias, que pueden ser distintas al tomar otra muestra. En (8.3) presentamos la varianza condicional, condicional sobre los valores de n_{hR}. La interpretación de un intervalo de confianza condicional de 95%, construido mediante la varianza en (8.3), es que sería de esperar que 95% de todas las muestras con tales valores específicos de n_{hR} proporcionen intervalos de confianza que contengan a $\bar{y}_U$.

La teoría de los intervalos de confianza condicional está más allá de los objetivos de este libro; referimos al lector interesado a Särndal *et al* (1992, sección 7.10), Casady y Valliant (1993) y Thompson (1997, sección 5.12) para un análisis más profundo y bibliografía.

9.6 Resumen y software

En este capítulo le presentamos brevemente algunos tipos básicos de métodos de estimación de la varianza utilizados en la práctica: linealización, grupos aleatorios, réplicas y funciones generalizadas de varianza. Esto es sólo una introducción; lea alguna de las referencias mencionadas en este capítulo antes de aplicar estos métodos a su propia encuesta

compleja. Desde 1980, se ha realizado gran parte de la investigación para explorar las propiedades y comportamiento de estos métodos y los métodos de estimación de la varianza siguen siendo un tema de estudio de los estadísticos.

Los métodos de linealización son tal vez los más investigados en términos de propiedades teóricas y se han empleado ampliamente para determinar estimaciones de la varianza en encuestas complejas. Sin embargo, la principal desventaja de la linealización es que debemos calcular las derivadas para cada estadística de interés y esto complica los programas para estimación de las varianzas. Si el software no considera la estadística de interés, usted deberá escribir su propio código.

El método de grupos aleatorios es intuitivamente atractivo para estimar las varianzas: es fácil de explicar y calcular, se puede utilizar para casi cualquier estadística de interés. Su principal desventaja es que generalmente necesitamos bastantes grupos aleatorios para tener una estimación estable de la varianza y la cantidad de grupos aleatorios que podemos formar queda limitado por el número de unidades primarias de cada estrato participantes en la muestra.

Los métodos de remuestreo para las encuestas estratificadas de varias etapas evitan el uso de las derivadas parciales para el cálculo de las estimaciones de las submuestras de la muestra compleja. Sin embargo, deben construirse con cuidado, de modo que la correlación de las observaciones en el mismo conglomerado se preserve en el remuestreo. Los métodos de remuestreo requieren de más tiempo de cómputo que la linealización, pero menos tiempo de programación: se usa el mismo método en todas las estadísticas. Se ha mostrado que estos métodos son equivalentes a la linealización para muestras grandes, cuando la característica de interés es una función suave de los totales de la población.

El método RRB se puede emplear con casi cualquier estadística, pero por lo general sólo se usa para los diseños con dos unidades primarias por estrato o para los diseños que se pueden reformular de modo que haya dos unidades por estrato. Los métodos de la navaja y *bootstrap* también se pueden utilizar para la mayoría de los estimadores que probablemente se emplean en las encuestas (excepción: el método de la navaja con una eliminación podría no funcionar muy bien para estimar la varianza de los cuantiles) y se pueden usar en las muestras estratificadas de varias etapas, donde se eligen más de dos unidades primarias en cada muestra, pero requieren más cálculos que la RRB.

Las funciones generalizadas de varianza son baratas y fáciles de usar, aunque tienen una desventaja importante: a menos que usted calcule la varianza mediante alguno de los demás métodos, no podrá garantizar que su estadística sigue el modelo utilizado para desarrollar dichas funciones.

Todos los métodos (excepto las funciones generalizadas) suponen que la información en los conglomerados está disponible para el analista de los datos. En muchas encuestas, no se libera tal información pues podría llevar a identificar a las personas que responden. Vea en Dippo *et al* (1984) un análisis de este problema.

Se han desarrollado varios paquetes de software como apoyo para analizar los datos de encuestas complejas. Cohen (1997), Lepkowski y Bowles (1996) y Carlson *et al* (1993) evalúan los paquetes de computadoras personales para el análisis de los datos de encuestas complejas.[1] SUDAAN (Shah *et al* 1995), OSIRIS (Lepkowski 1982), Stata (StataCorp 1996) y PC-CARP (Fuller *et al* 1989) utilizan métodos de linealización para estimar las varianzas de estadísticas no lineales. SUDAAN, por ejemplo, calcula las varianzas de los totales estimados de la población para diversos diseños de muestreo estratificado de varias etapas con H estratos, para el muestreo por conglomerados con probabilidades diferentes, con o sin reemplazo en la primera etapa de muestreo, y para muestras aleatorias simples con o sin reemplazo en las etapas posteriores. La fórmula en (6.9) se usa para estimar la varianza en cada estrato en el muestreo con reemplazo y la forma de Sen-Yates-Grundy en (6.15) se usa para la varianza sin reemplazo. Luego, se suman las varianzas para los totales de los estratos para estimar la varianza del total estimado de la población. SUDAAN utiliza la linealización para determinar las varianzas de las razones, los coeficientes de regresión y otras estadísticas no lineales. Las versiones recientes de SUDAAN también implantan los métodos RRB y de la navaja.

OSIRIS también implementa estos métodos. Los paquetes de software para encuestas WesVarPC (Brick *et al* 1996, que pueden bajarse en forma gratuita de www.westat.com) y VPLX (Fay 1990) utilizan métodos de remuestreo para calcular las estimaciones de la varianza. En el apéndice D damos una función S-PLUS sencilla para el método de la navaja; esto no pretende sustituir el software comercial ya probado, sino darle una idea de la forma de hacer esos cálculos. Después de aprender los principios de los métodos, puede utilizar el software comercial para sus encuestas complejas.

[1] Lepkowski y Bowles (1996) indican cómo tener acceso a los paquetes gratuitos (o casi gratuitos) CENVAR, CLUSTERS, Epi Info, VPLX y WesVarPC mediante el correo electrónico o Internet. El software para el análisis de los datos de encuestas cambia rápidamente; la sección de métodos de investigación de encuestas de la American Statistical Association (www.amstat.org) es una buena fuente de información actualizada.

9.7 Ejercicios

1 ¿Cuáles de los métodos de estimación de la varianza de este capítulo serían adecuados para estimar la proporción de camas que tienen mosquiteros en la encuesta respectiva de Gambia del ejemplo 7.1? Explique por qué cada método es o no adecuado.

2 Como en el ejemplo 9.1, sea $h(p) = p(1-p)$.

a Determine el término del residuo en el desarrollo de Taylor, $\int_a^x (x-t)h''(t)dt$, y úselo para determinar una expresión exacta para $h(\hat{p})$.

b ¿Es probable que el término del residuo sea menor que los demás términos? Justifique su respuesta.

c Determine una expresión exacta para $V[h(\hat{p})]$ para una muestra aleatoria simple con reemplazo. ¿Cuál es su relación con la aproximación del ejemplo 9.1?

3 La pendiente de la recta de regresión para la población es

$$B_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(yi - \bar{y}_U)}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x}_U)^2}}.$$

a Exprese B_1 como una función de los totales de la población $t_1 = \sum_{i=1}^{N} x_i, t_2 = \sum_{i=1}^{N} y_i, t_3 = \sum_{i=1}^{N} x_i^2$, y $t_4 = \sum_{i=1}^{N} x_i y_i$, de modo que $B_1 = h(t_1, t_2, t_3, t_4)$.

b Sea $\hat{B}_1 = h(\hat{t}_1, \hat{t}_2, \hat{t}_3, \hat{t}_4)$ y suponga que $E[\hat{t}_i] = t_i$, para $i = 1, 2, 3, 4$. Use el método de linealización para determinar una aproximación a la varianza de $\hat{B}_1$. Exprese su respuesta en términos de $V(\hat{t}_i)$ y $\text{Cov}(\hat{t}_i, \hat{t}_j)$.

c ¿Cuál es la aproximación por linealización de la varianza de una muestra aleatoria simple de tamaño n?

d Determine un variado linealizado q_i de modo que $\hat{V}(\hat{B}_1) = \hat{V}(\hat{t}_q)$.

4 El coeficiente de correlación para la población es

$$R=\frac{\sum_{i=1}^{N}(x_i-\bar{x}_U)(y_i-\bar{y}_U)}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x}_U)^2\sum_{i=1}^{N}(y_i-\bar{y}_U)^2}}.$$

a Exprese R como función de los totales de la población $t_1=\sum_{i=1}^{N}x_i, t_2=\sum_{i=1}^{N}y_i, t_3=\sum_{i=1}^{N}x_i^2, t_4=\sum_{i=1}^{N}x_iy_i$, y $t_5=\sum_{i=1}^{N}y_i^2$, de modo que $R=h(t_1,t_2,t_3,t_4,t_5)$.

b Sea $\hat{R}=h(\hat{t}_1,\ldots,\hat{t}_5)$ y suponga que $E[\hat{t}_i]=t_i$, para $i=1,\ldots,5$. Use el método de linealización para determinar una aproximación a la varianza de $\hat{R}$.

c ¿Cuál es la aproximación por linealización de la varianza de una muestra aleatoria simple de tamaño n?

5 *Estimación de la varianza con estratificación* a posteriori. Suponga que estratificamos *a posteriori* la muestra en L estratos, con cifras de población $N_1, N_2,\ldots, N_L$. Entonces, el estimador estratificado posterior para el total de la población es

$$\hat{t}_{\text{post}}=\sum_{l=1}^{L}\frac{N_l}{\hat{N}_l}\hat{t}_l=h(\hat{t}_1,\ldots,\hat{t}_L,\hat{N}_1,\ldots,\hat{N}_L),$$

donde

$$\hat{t}_l=\sum_{i\in S_l}w_iy_i,\qquad \hat{N}_l=\sum_{i\in S_l}w_i,$$

y S_l es el conjunto de unidades de la muestra que están en el estrato posterior l. Muestre mediante la linealización que

$$V(\hat{t}_{\text{post}})\approx V\left(\hat{t}-\sum_{l=1}^{L}\frac{t_l}{N_l}\hat{N}_l\right).$$

6 Suponga que se extrae una muestra aleatoria estratificada con dos observaciones por cada estrato. Muestre que si $\sum_{r=1}^{R}\alpha_{rh}\alpha_{rl}=0$, para $l\neq h$, entonces

$$\hat{V}_{\text{RRB}}(\bar{y}_{\text{est}})=\hat{V}_{\text{est}}(\bar{y}_{\text{est}}).$$

SUGERENCIA: observe primero que

$$\bar{y}_{\text{est}}(\alpha_l)-\bar{y}_{\text{est}}=\sum_{h=1}^{H}\frac{N_h}{N}\alpha_{lh}\frac{y_{h1}-y_{h2}}{2}.$$

Y luego exprese $\hat{V}_{\text{RRB}}(\bar{y}_{\text{est}})$ de manera directa mediante y_{h1} y y_{h2}.

7 Otros estimadores RRB de la varianza son

$$\frac{1}{4R}\sum_{r=1}^{R}[\hat{\theta}(\alpha_r)-\hat{\theta}(-\alpha_r)]^2$$

y

$$\frac{1}{2R}\sum_{r=1}^{R}\{[\hat{\theta}(\alpha_r)-\hat{\theta}]^2+[\hat{\theta}(\alpha_r)-\hat{\theta}]^2\}.$$

Para una muestra aleatoria estratificada con dos observaciones por estrato, muestre que si $\sum_{r=1}^{R}\alpha_{rh}\alpha_{rl}=0$ para $l\neq h$, entonces cada uno de estos estimadores de la varianza es equivalente a $\hat{V}_{\text{est}}(\bar{y}_{\text{est}})$.

8 Suponga que el parámetro de interés es $\theta=h(t)$, donde $h(t)=at^2+bt+c$ y t es el total de la población. Sea $\hat{\theta}=h(\hat{t})$. Demuestre, en una muestra aleatoria estratificada con dos observaciones por estrato, que si $\sum_{r=1}^{R}\alpha_{rh}\alpha_{rl}=0$ para $l\neq h$, entonces

$$\frac{1}{4R}\sum_{r=1}^{R}\left[\hat{\theta}(\alpha_r)-\hat{\theta}(-\alpha_r)\right]^2=\hat{V}_L(\hat{\theta}),$$

que es la estimación de la varianza por linealización (vea Rao y Wu 1985).

9 Use los grupos aleatorios en el archivo de datos syc.dat para estimar las varianzas de las estimaciones de la proporción de jóvenes que:

a Tienen 14 o menos años de edad.

b Fueron detenidos por un delito con violencia.

c Vivían con ambos padres.

d Son hombres.

e Son latinoamericanos.

f Crecieron principalmente en una familia con uno solo de sus padres.

g Han utilizado drogas.

10 El método de linealización de la sección 9.1 es el utilizado históricamente para determinar las varianzas. Binder (1996) propone pasar directamente a estimar la varianza, evaluando las derivadas parciales en las estimaciones de la muestra en vez de utilizar las cantidades de la población. ¿Cuál es la estimación de Binder para la varianza del estimador de la razón? ¿Difiere del de la sección 9.1?

11 Determine una estimación mediante el método de la navaja de la edad media de la población de árboles en un terreno, para los datos del ejercicio 4 del capítulo 3 y calcule la estimación de la varianza mediante el mismo método. ¿Cuál es la relación de estos estimadores con los basados en el método de serie de Taylor? Asegúrese de incluir los detalles de cálculo de sus estimaciones con el método de la navaja.

12 Utilice el método de la navaja para estimar las varianzas de sus estimaciones en las partes (a) y (b) del ejercicio 17 del capítulo 5.

13 Use el método de la navaja para estimar la varianza del estimador de la razón usada en el ejemplo 3.2. ¿Cuál es su relación con el estimador mediante linealización?

14 Emplee el método de Woodruff para construir un intervalo de confianza para la mediana de la cuota de uso de los *greens* entre semana, para nueve hoyos, usando la muestra aleatoria simple en el archivo golfsrs.dat.

10

Análisis de datos categóricos en encuestas complejas*

No obstante, la estadística no debe hacerse para demostrar una idea preconcebida.

—Florence Nightingale, nota en *Physique Sociale*, de A. Quetelet

Hasta ahora hemos tratado de estimar cantidades de resumen, como las medias, los totales y los porcentajes, en diversos diseños de muestreo. Los totales y los porcentajes son importantes para muchas encuestas, ya que proporcionan una descripción de la población: por ejemplo, el porcentaje de la población que fue víctima de algún delito o la cantidad total de personas desempleadas en Estados Unidos. Sin embargo, con frecuencia, los investigadores están interesados en cuestiones multivariadas: ¿está asociada la raza con el hecho de ser víctima de algún delito? ¿Podemos predecir el estado de desempleo a partir de variables demográficas? Generalmente, en estadística tales cuestiones se contestan usando técnicas del análisis de datos categóricos o la regresión (que analizaremos en el capítulo 11). No obstante, las técnicas que usted ha aprendido en un curso introductorio de estadística suponían que todas las observaciones eran independientes e idénticamente distribuidas a partir de cierta distribución de la población. Esta hipótesis ya no es válida en los datos de encuestas complejas; en este capítulo y el siguiente examinaremos los efectos del diseño de un muestreo complejo sobre los análisis estadísticos de uso común.

Puesto que gran parte de la información de las encuestas muestrales se reúne en forma de porcentajes, en este análisis se emplean ampliamente los métodos de datos categóricos. De hecho, muchos de los conjuntos de datos utilizados para ilustrar la prueba ji cuadrada en los cursos de introducción a la estadística surgen de encuestas complejas. Nos preocupan principalmente los efectos de la formación de conglomerados en las pruebas de hipótesis de uso común y los modelos para los datos categóricos, ya que, generalmente, la formación de conglomerados disminuye la precisión. Comenzaremos revisando varias pruebas ji cuadrada al extraer una muestra aleatoria simple de una población grande.

10.1 Pruebas ji cuadrada con muestreo multinomial

EJEMPLO 10.1 A cada pareja de una muestra aleatoria simple de 500 parejas casadas en una población grande se le pregunta (1) si la familia posee al menos una computadora personal, y (2) si la familia tiene televisión por cable. La siguiente tabla de contingencias presenta los

resultados:

		¿Computadora? Sí	No	
¿Cable?	Sí	119	188	307
	No	88	105	193
		207	293	500

¿Es más probable que las familias con computadora se suscriban a la televisión por cable? Con frecuencia, para este tipo de preguntas se utiliza una prueba ji cuadrada para la independencia. Bajo la hipótesis nula de que poseer una computadora y suscribirse a la televisión por cable son independientes, las cifras esperadas para celda de la tabla de contingencias son las siguientes:

		¿Computadora? Sí	No	
¿Cable?	Sí	127.1	179.9	307
	No	79.9	113.1	193
		207	293	500

La **estadística de la prueba ji cuadrada de Pearson** es

$$X^2 = \sum_{\substack{\text{todas las}\\ \text{celdas}}} \frac{(\text{cifra observada} - \text{cifra esperada})^2}{\text{cifra esperada}} = 2.281.$$

La **estadística de la prueba ji cuadrada de la razón de verosimilitud** es

$$G^2 = 2 \sum_{\substack{\text{todas las}\\ \text{celdas}}} (\text{cifra observada}) \ln\left(\frac{\text{cifra observada}}{\text{cifra esperada}}\right) = 2.275.$$

Las dos estadísticas de prueba son asintóticamente equivalentes; para muestras grandes, cada una sigue aproximadamente una distribución ji cuadrada (χ^2) con un grado de libertad bajo la hipótesis nula. El valor p de cada estadística es 0.13, lo que no da razón para dudar de que la hipótesis de poseer computadora y suscribirse a la televisión por cable son independientes.

Si tener computadora y suscribirse a la televisión por cable son eventos independientes, las posibilidades de que quien adquiere televisión por cable posea una computadora deben ser iguales a las posibilidades de que una persona sin televisión por cable posea computadora. Estimamos las posibilidades de tener computadora si la familia está suscrita al cable como 119/188 y estimamos las posibilidades de tener computadora si la familia no está suscrita como 88/105. Entonces, el **cociente de posibilidades** se estima como

$$\frac{\frac{119}{188}}{\frac{88}{105}} = 0.755.$$

Si la hipótesis nula de independencia es cierta, esperamos que el cociente de posibilidades sea cercano a 1. En forma equivalente, esperamos que el logaritmo de este cociente sea cercano a 0. La posibilidad logarítmica es -0.28, con un error estándar asintótico de

$$\sqrt{\frac{1}{119}+\frac{1}{88}+\frac{1}{188}+\frac{1}{105}} = 0.186;$$

un intervalo de confianza aproximado de 95% para la posibilidad logarítmica es $-0.28 \pm 1.96(0.186) = [-0.646, 0.084]$. Este intervalo de confianza incluye a cero, lo que confirma el resultado de la prueba de hipótesis en el sentido de que no hay evidencia en contra de la independencia. ■

Las pruebas ji cuadrada se utilizan de manera común en estas situaciones; cada una supone una forma de muestreo aleatorio. Estas pruebas se analizan con más detalle en Lindgren (1993, capítulo 10), Agresti (1990) y Christensen (1990).

10.1.1 Prueba de independencia de factores

Cada una de las n observaciones independientes se clasifica mediante dos factores: el factor por renglón R con r niveles y el factor por columna C con c niveles. Cada observación tiene la probabilidad p_{ij} de caer en la categoría por renglón i y categoría por columna j, con lo que obtenemos la siguiente tabla de probabilidades. En este caso, $p_{i+} = \sum_{j=1}^{c} p_{ij}$ es la probabilidad de que una unidad seleccionada al azar caiga en la categoría por renglón i y $p_{+j} = \sum_{i=1}^{r} p_{ij}$ es la probabilidad de que una unidad seleccionada al azar caiga en la categoría por columna j:

		C 1	2	...	c	
R	1	$P11$	$P12$	...	$P1c$	$P1+$
	2	$P21$	$P22$	...	$P2c$	$P2+$
	⋮	⋮	⋮		⋮	⋮
	r	$Pr1$	$Pr2$	...	Prc	$Pr+$
		$P+1$	$P+2$	...	$P+c$	1

La cifra esperada en la celda (i, j) de la muestra es x_{ij}. Si todas las unidades de la muestra son independientes, las x_{ij} son de una distribución multinomial con rc categorías; este esquema de muestreo se conoce como **muestreo multinomial**. En encuestas, las hipótesis para el muestreo multinomial se cumplen en una muestra aleatoria simple con reemplazo; lo hacen aproximadamente en una muestra aleatoria simple sin reemplazo cuando el tamaño de la muestra es pequeño comparado con el tamaño de la población. Esta última situación apareció en el ejemplo 10.1: el muestreo multinomial independiente significa que tenemos una muestra de 500 familias (aproximadamente) independientes y nos fijamos a qué categoría pertenece cada una.

La hipótesis nula de independencia es

$$H_0 : p_{ij} = p_{i+}p_{+j} \quad \text{para } i = 1,\ldots,r \text{ y } j = 1,\ldots,c. \tag{10.1}$$

Sean $m_{ij} = np_{ij}$ las cifras esperadas. Si H_0 es cierta, $m_{ij} = np_{i+}\,p_{+j}$, y podemos estimar m_{ij} como

$$\hat{m}_{ij} = n\hat{p}_{i+}\hat{p}_{+j} = n\frac{x_{i+}}{n}\frac{x_{+j}}{n},$$

donde $\hat{p}_{ij} = x_{ij}/n$ y $\hat{p}_{i+} = \sum_{j=1}^{c} \hat{p}_{ij}$. La estadística de la prueba ji cuadrada de Pearson es

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(x_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} = n\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(\hat{p}_{ij} - \hat{p}_{i+}p_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}}. \tag{10.2}$$

La estadística de prueba de la razón de verosimilitud es

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij} \ln\left(\frac{x_{ij}}{\hat{m}_{ij}}\right) = 2n\sum_{i=1}^{r}\sum_{j=1}^{c} \hat{p}_{ij} \ln\left(\frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}}\right). \tag{10.3}$$

Si usamos el muestreo multinomial con un tamaño de muestra suficientemente grande, X^2 y G^2 se distribuyen aproximadamente como una variable aleatoria ji cuadrada con $(r-1)(c-1)$ grados de libertad bajo la hipótesis nula. Qué tan grandes "suficientemente grande" depende de la cantidad de celdas y de las probabilidades esperadas; Fienberg (1979) argumenta que los valores p serán aproximadamente correctos si (1) la cifra esperada en cada celda es mayor que 1 y (b) $n \geq 5 \times$ (cantidad de celdas).

Una afirmación equivalente a (10.1) es que todos los cocientes de posibilidades sean iguales a 1:

$$H_0 : \frac{p_{11}p_{ij}}{p_{1j}p_{i1}} = 1 \text{ para toda } i \geq 2 \text{ y } j \geq j.$$

Podemos estimar cualquier cociente de posibilidades $p_{ij}p_{kl}/p_{il}p_{kj}$ sustituyendo las proporciones estimadas: $\hat{p}_{ij}\hat{p}_{kl}/\hat{p}_{il}\hat{p}_{kj}$. Si la muestra es suficientemente grande, el *logaritmo* del cociente estimado tiene aproximadamente una distribución normal con error estándar

$$\sqrt{\frac{1}{x_{ij}} + \frac{1}{x_{kl}} + \frac{1}{x_{il}} + \frac{1}{x_{kj}}}.$$

10.1.2 Prueba de la homogeneidad de las proporciones

Las estadísticas de la prueba ji cuadrada de Pearson y de la razón de verosimilitud en (10.2) y (10.3) también sirven cuando cada una de las muestras aleatorias independientes de r poblaciones se clasifica en una de c categorías. El muestreo multinomial se realiza dentro de cada población, de modo que el esquema de muestreo se llama **muestreo multinomial producto**, y es equivalente al muestreo aleatorio estratificado cuando la fracción de muestreo para cada estrato es pequeña o cuando el muestreo es con reemplazo.

La diferencia entre el muestreo multinomial producto y el muestreo multinomial es que los totales por renglón p_{i+} y x_{i+} son cantidades fijas en el muestreo multinomial producto; x_{i+} es el tamaño de muestra predeterminado para el estrato i. La hipótesis nula en el sentido de que la proporción de observaciones que caen en la clase j es la misma para todos los estratos es

$$H_0 : \frac{p_{1j}}{p_{1+}} = \frac{p_{2j}}{p_{2+}} = \cdots = \frac{p_{rj}}{p_{r+}} = p_{+j} \quad \text{para todo } j = 1, \ldots, c. \tag{10.4}$$

Si la hipótesis nula en (10.4) es cierta, entonces $m_{ij} = np_{i+}p_{+j}$, como antes, y las cifras esperadas bajo H_0 son $\hat{m}_{ij} = n\hat{p}_i + \hat{p}_{+j}$, exactamente como en la prueba de independencia.

EJEMPLO 10.2 Los tamaños de muestra utilizados en el ejercicio 14 del capítulo 4, la muestra estratificada de estudiantes de enfermería y tutores, fueron los tamaños de muestra de quienes respondieron. Usemos una prueba ji cuadrada para homogeneidad de proporciones, para probar la hipótesis nula de que la tasa de ausencia de respuesta es la misma para cada estrato. Los cuatro estratos forman los renglones de la siguiente tabla de contingencias:

	No responden	Sí responden	
Estudiante general	46	222	268
Tutor general	41	109	150
Estudiante de psiquiatría	17	40	57
Tutor de psiquiatría	8	26	34
	112	397	509

Las dos estadísticas para la prueba ji cuadrada son $X^2 = 8.218$, con valor p 0.042, y $G^2 = 8.165$, con valor p 0.043. Así, hay evidencias de tasas distintas de ausencia de respuesta entre los cuatro grupos. Sin embargo, la siguiente tabla muestra que la diferencia no puede atribuirse al efecto principal de general/psiquiátrico o estudiante/tutor:

	Tasa de ausencia de respuesta	
	Estudiante	Tutor
General	17%	27%
de psiquiatría	30%	24%

Se necesitará una investigación más profunda para explorar el patrón de ausencia de respuesta. ■

10.1.3 Prueba de bondad de ajuste

Suponemos de nuevo un muestreo multinomial, con las observaciones independientes clasificadas en k categorías. La hipótesis nula es

$$H_0 : p_i = p_i^{(0)} \quad \text{para } i = 1, \ldots, k,$$

donde $p_i^{(0)}$ está determinada de antemano o es una función de los parámetros θ que se estimarán a partir de los datos.

EJEMPLO 10.3 Webb (1995) examinó los registros de seguridad de 17,952 pilotos de la Fuerza Aérea de Estados Unidos durante un periodo de ocho años en torno de la Segunda Guerra Mundial y construyó la siguiente tabla de frecuencias.

Cantidad de accidentes	Cantidad de pilotos
0	12,475
1	4,117
2	1,016
3	269
4	53
5	14
6	6
7	2

Si los accidentes ocurren al azar (y si ningún piloto es más o menos "propenso a los accidentes" que los demás) una distribución de Poisson ajusta bien a los datos. Estimamos la media de la distribución de Poisson mediante la cantidad media de accidentes por piloto en la muestra, 0.40597. Las probabilidades observada y esperada bajo la hipótesis nula de

que los datos siguen una distribución de Poisson aparecen en la siguiente tabla. Las probabilidades esperadas se calculan mediante las probabilidades de Poisson $e^{-\lambda}\lambda^x/x!$, con $\lambda = 0.40597$.

Cantidad de accidentes	Proporción observada, $\hat{p}_i$	Probabilidad esperada bajo H_0, $\hat{p}_i^{(0)}$
0	.6949	.6663
1	.2293	.2705
2	.0566	.0549
3	.0150	.0074
4	.0030	.0008
5+	.0012	.0001

Las dos estadísticas de la prueba ji cuadrada son

$$X^2 = \sum_{\text{todas las celdas}} \frac{(\text{cifra observada} - \text{cifra esperada})^2}{\text{cifra esperada}}$$

$$= \sum_{i=1}^{k} \frac{\left(n\hat{p}_i - n\hat{p}_i^{(0)}\right)^2}{n\hat{p}_i^{(0)}} \qquad (10.5)$$

$$= n\sum_{i=1}^{k} \frac{\left(\hat{p}_i - \hat{p}_i^{(0)}\right)^2}{\hat{p}_i^{(0)}}$$

y

$$G^2 = 2n\sum_{i=1}^{k} \hat{p}_i \ln\left(\frac{\hat{p}_i}{\hat{p}_i^{(0)}}\right). \qquad (10.6)$$

Para los pilotos, $X^2 = 756$ y $G^2 = 400$. Si la hipótesis nula es cierta, ambas estadísticas seguirán aproximadamente una distribución χ^2 con cuatro grados de libertad (se utilizan dos grados de libertad en n y $\hat{\lambda}$). Ambos valores p son menores que 0.0001, lo que proporciona una evidencia de que un modelo de Poisson no se ajusta a los datos. Más pilotos no tienen accidentes, o tienen más de dos accidentes, de lo que sería de esperar con un modelo Poisson. Así, la evidencia muestra que algunos pilotos son más propensos a los accidentes que lo que ocurriría bajo el modelo Poisson. ■

Todas las estadísticas de la prueba ji cuadrada en (10.2), (10.3), (10.5) y (10.6) crecen con n. Si la hipótesis nula no es precisamente cierta en la población (si las familias con cable tienen una probabilidad infinitesimalmente mayor de poseer una computadora personal en comparación con las familias sin cable), casi podemos garantizar el rechazo de la hipótesis nula al extraer una muestra aleatoria de gran tamaño. Esta propiedad de la prueba de hipótesis significa que será sensible a la inflación artificial del tamaño de la muestra ignorando los conglomerados.

10.2 Efectos del diseño de la muestra sobre las pruebas ji cuadrada

El diseño de la muestra puede afectar a las probabilidades estimadas de las celdas y las pruebas de asociación y bondad de ajuste. En los diseños complejos de encuestas ya no tenemos el muestreo aleatorio que da a X^2 y G^2 una distribución aproximada χ^2. Así, si sólo empleamos un paquete estadístico estándar para hacer nuestras pruebas ji cuadrada, los

niveles de significado y los valores p serán incorrectos. En particular, los conglomerados pueden tener un fuerte efecto sobre los valores p de las pruebas ji cuadrada. En una muestra por conglomerados con un coeficiente de correlación entre clases (ICC) positivo, frecuentemente ocurrirá que el valor p verdadero será mucho mayor que el valor p reportado por el paquete estadístico bajo la hipótesis de muestreo multinomial independiente. Veamos qué puede ocurrir con las pruebas de hipótesis si se ignora el diseño de la encuesta en una muestra por conglomerados.

EJEMPLO 10.4 Suponga que a la esposa y al esposo se les cuestiona acerca de la suscripción a la televisión por cable y la computadora, para la encuesta analizada en el ejemplo 10.1, y que ambos dan la misma respuesta. Aunque las hipótesis de muestreo multinomial se cumplen para la muestra aleatoria simple de parejas, no se cumplen para la muestra por conglomerados de personas; lejos de ser unidades independientes, el esposo y la esposa de la misma familia coinciden totalmente en sus respuestas. El ICC para la muestra por conglomerados es 1.

¿Qué ocurre si ignoramos los conglomerados? La tabla de contingencias para las frecuencias observadas es la siguiente:

		¿Computadora? Sí	No	
¿Cable?	Sí	238	376	614
	No	176	210	386
		414	586	1000

Las proporciones y el cociente de posibilidades estimados son idénticos a los del ejemplo 10.1: $\hat{p}_{11} = 238/1000 = 119/500$, y el cociente de posibilidades es

$$\frac{\frac{238}{376}}{\frac{176}{210}} = 0.755.$$

Pero $X^2 = 4.562$ y $G^2 = 4.550$ son el doble de los valores para las estadísticas de prueba del ejemplo 10.1. Si ignora los conglomerados y compara estas estadísticas con las de una distribución χ^2 con un grado de libertad, reportará un "valor p" de 0.033 y concluirá que los datos proporcionan evidencias de que tener una computadora y suscribirse a la televisión por cable no son independientes. Si seguimos este camino, el "valor p" podría disminuir aún más, entrevistando también a los niños de cada familia, multiplicando así la estadística de la prueba original por 4.

¿Podría obtener un valor p arbitrariamente pequeño observando más unidades secundarias dentro de cada unidad primaria? Absolutamente no. Las estadísticas X^2 y G^2 tienen una distribución χ_1^2 nula *si se utiliza el muestreo multinomial*. Cuando se extrae una muestra por conglomerados en vez de esto y el ICC es positivo, X^2 y G^2 *no* siguen una distribución χ_1^2 bajo la hipótesis nula. Para los mil esposos y esposas, $X^2/2$ y $G^2/2$ siguen una distribución χ_1^2 bajo H_0, lo cual da el mismo valor p hallado en el ejemplo 10.1. ■

10.2.1 Tablas de contingencia para datos de encuestas complejas

Las cifras observadas x_{ij} no reflejan necesariamente las frecuencias relativas de las categorías de la población, a menos que la muestra sea autoponderada. Suponga que extraemos

una muestra aleatoria simple de salones de escuelas de nivel básico en Denver y que se evalúa a cada uno de los 10 estudiantes elegidos al azar en cada salón con respecto a su autoconcepto (alto o bajo) y depresión clínica (presente o no). Se elige a los estudiantes para la muestra con probabilidades diferentes; los estudiantes de grupos pequeños tienen más probabilidad de estar en la muestra que los de grupos grandes. Una tabla con las cifras observadas a partir de la muestra, ignorando las probabilidades de selección, no daría una imagen precisa de la asociación entre su autoconcepto y la depresión en la población, si el grado de asociación difiere con el tamaño de la clase. Aunque la asociación entre el autoconcepto y la depresión fuese la misma para distintos tamaños de clase, las estimaciones de la cantidad de estudiantes deprimidos usando los márgenes de la tabla de contingencias estarían equivocados.

Sin embargo, recuerde que podemos usar pesos de muestreo para estimar cualquier cantidad de la población. En este caso, los empleamos para estimar las proporciones por celda. Estimamos p_{ij} como

$$\hat{p}_{ij} = \frac{\sum_{k \in S} w_k y_{kij}}{\sum_{k \in S} w_k}, \qquad \textbf{(10.7)}$$

donde

$$y_{kij} = \begin{cases} 1 & \text{si la unidad de observación } k \text{ está en la celda } (i, j) \\ 0 & \text{en caso contrario} \end{cases}$$

y w_k es el peso para la unidad de observación k. Así,

$$\hat{p}_{ij} = \frac{\text{suma de pesos para las unidades de observación en la celda } (i, j)}{\text{suma de pesos para todas las unidades de observación en la muestra}}.$$

Si la muestra es autoponderada, $\hat{p}_{ij}$ será la proporción de unidades de observación que caen en la celda (i, j). Usamos las estimaciones $\hat{p}_{ij}$ y construimos la tabla

		C				
		1	2	...	c	
R	1	$\hat{p}_{11}$	$\hat{p}_{12}$	...	$\hat{p}_{1c}$	$\hat{p}_{1+}$
	2	$\hat{p}_{21}$	$\hat{p}_{22}$	...	$\hat{p}_{2c}$	$\hat{p}_{2+}$
	⋮	⋮				⋮
	r	$\hat{p}_{r1}$	$\hat{p}_{r2}$	...	$\hat{p}_{rc}$	$\hat{p}_{r+}$
		$\hat{p}_{+1}$	$\hat{p}_{+2}$	...	$\hat{p}_{+c}$	1

para examinar las asociaciones; estimamos los cocientes de posibilidades mediante $\hat{p}_{ij}\hat{p}_{kl} / \hat{p}_{il}\hat{p}_{kj}$. Podemos construir un intervalo de confianza para p_{ij} usando cualquier método de estimación de la varianza analizado hasta ahora, o utilizar un efecto de diseño para modificar el intervalo de confianza de la muestra aleatoria simple, como en (7.7).

Sin embargo no debe tirar las cifras observadas. Si los cocientes de posibilidades calculados mediante las $\hat{p}_{ij}$ difieren de manera apreciable de los cocientes de posibilidades calculados mediante las cifras observadas x_{ij}, usted deberá investigar por qué difieren. Tal vez, los cocientes de posibilidades para la depresión y la autoconcepción difieren para clases más grandes o dependen de factores socioeconómicos relacionados con el tamaño de la clase. Si eso ocurre, usted debe incluir estos otros factores en un modelo de los datos o tal vez deba probar separadamente la asociación para las clases grandes y pequeñas.

10.2.2 Efectos sobre las pruebas de hipótesis y los intervalos de confianza

Podemos estimar las proporciones de una tabla de contingencia y los cocientes de posibilidades usando los pesos. Sin embargo, estos pesos no son de mucha ayuda al construir las pruebas de hipótesis y los intervalos de confianza, pues dependen de los conglomerados y, a veces, de la estratificación del diseño de la encuesta.

Observemos primero el efecto de la estratificación. Si los estratos son las categorías en bruto, la estratificación no representa problema alguno; esencialmente tenemos un muestreo multinomial producto, como lo describimos en la sección 10.1.2, y podemos probar la homogeneidad de las proporciones de la manera usual.

Sin embargo, en las encuestas altamente estratificadas, la asociación entre los estratos y otros factores podría no ser de interés. Por ejemplo, en la encuesta nacional a víctimas de delitos, podríamos estar interesados en la asociación entre el género y el hecho de ser víctima de algún delito violento y quisiéramos incluir los datos de todos los estratos en el análisis. En general, la estratificación aumenta la precisión de las estimaciones sobre las muestras aleatorias simples. Para una muestra aleatoria simple, (10.2) implica

$$X^2 = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}}.$$

Una muestra estratificada con n unidades de observación brinda la misma precisión para la estimación de p_{ij} que una muestra aleatoria simple con n/d_{ij} unidades de observación, donde d_{ij} es el efecto de diseño para la estimación de p_{ij}. Si la estratificación vale la pena, los efectos de diseño generalmente serán menores que 1. En consecuencia, si empleamos las estadísticas de prueba de una muestra aleatoria simple en (10.2) y (10.3) con los $\hat{p}_{ij}$ de la muestra estratificada, X^2 y G^2 serán menores de lo que deberían para seguir una distribución $\chi^2_{(r-1)(c-1)}$ nula; los "valores p" calculados ignorando la estratificación serán demasiado grandes y H_0 no se rechaza con la frecuencia debida. Así, mientras el procedimiento FREQ de SAS u otros paquetes normales de estadística podrían darle un valor p de 0.04, el valor p real podría ser 0.02. Al ignorar la estratificación se obtiene una prueba conservadora. De manera similar, un intervalo de confianza construido para un cociente de posibilidades logarítmicas casi siempre es demasiado grande si se ignora la estratificación. En realidad, sus estimaciones son más precisas de lo que indica el intervalo de confianza de la muestra aleatoria simple.

Por lo general, los conglomerados tienen el efecto opuesto. Los efectos de diseño para $\hat{p}_{ij}$ con una muestra por conglomerados generalmente son mayores que 1; una muestra por conglomerados con n unidades de observación da la misma precisión que una muestra aleatoria simple con menos de n observaciones. Si ignoramos los conglomerados, es de esperar que X^2 y G^2 sean mayores que al extraer la muestra aleatoria simple con tamaño equivalente y es probable que los "valores p" calculados ignorando los conglomerados sean demasiado pequeños. SAS podría darle un valor p de 0.04, aunque el verdadero valor p sea 0.25. Al ignorar los conglomerados, usted podría decidir que una asociación es significativa desde el punto de vista estadístico cuando en realidad sólo se debe a una variación aleatoria de los datos. Los intervalos de confianza para los cocientes de posibilidades logarítmicas serán más anchos de lo que deberían ser; las estimaciones no son tan precisas como los intervalos de confianza de SAS podrían inducirle a creer.

Con frecuencia, ignorar los conglomerados en las pruebas ji cuadrada es más peligroso que ignorar la estratificación. Una prueba ji cuadrada basada en una muestra aleatoria simple con datos estratificados podría exhibir las asociaciones fuertes pero no descubrirá todas las asociaciones débiles. Sin embargo, al ignorar los conglomerados, algunas asociaciones se considerarán importantes desde el punto de vista estadístico cuando en realidad no lo son. Al ignorar los conglomerados en las pruebas de bondad de ajuste, podría adoptarse un modelo innecesariamente complicado para describir los datos.

Con frecuencia, algún investigador ignorante de la teoría de muestreo analizará correctamente una muestra estratificada, usando los estratos como una de las variables de clasificación. Pero tal vez el investigador ni siquiera registraría los conglomerados y, con frecuencia, simplemente pasará las cifras observadas por los procedimientos FREQ de SAS o CROSSTABS de SPSS y aceptará el valor p indicado como verdadero. Como ilustración, considere un investigador que desee reproducir el estudio de Basow y Silberg (1987) acerca de si los alumnos de bachillerato evalúan de modo diferente a los profesores que a las profesoras. (Analizamos el estudio original en el ejemplo 5.1.) El investigador selecciona una muestra estratificada de profesores y profesoras de bachillerato y pide a cada estudiante en los grupos de estos docentes que los evalúen. Se obtuvieron más de 2000 respuestas de estudiantes y el investigador clasifica las respuestas por sexo del maestro y dependiendo de si el estudiante le otorga una calificación alta o baja. El investigador, comparando la estadística X^2 de Pearson sobre las cifras observadas con una distribución χ^2_1, declara una asociación estadísticamente significativa entre el género del profesor y la calificación dada por el estudiante. La variable de estratificación *sexo del profesor* es una de las variables de clasificación, de modo que no hay necesidad de realizar ajustes a la estratificación. Sin embargo, el valor p reportado casi seguramente es incorrecto, por varias razones: (1) se ignoran los conglomerados de estudiantes dentro de un grupo; en realidad, el investigador ni siquiera registra a qué profesor evalúa el estudiante, sino sólo su sexo, de modo que la investigación no puede tomar en cuenta los conglomerados. Si las evaluaciones de los estudiantes reflejan la calidad de la enseñanza, es de esperar que los estudiantes de un "buen" profesor den mejores calificaciones que los estudiantes de uno "malo". El ICC para los estudiantes es positivo y el tamaño de muestra equivalente en una muestra aleatoria simple es menor de 2000. El valor p reportado por el investigador es demasiado pequeño y el investigador se puede equivocar al afirmar que las profesoras reciben una calificación media distinta al ser evaluadas por los estudiantes. (2) Varios estudiantes podrían dar respuestas para más de un maestro en la muestra. No es claro el efecto que estas respuestas múltiples podrían tener sobre la prueba de independencia. (3) No todos los estudiantes asisten a clase o participan en la evaluación. Parte de la ausencia de respuesta podría ser faltante completamente al azar (un estudiante podría estar enfermo el día del estudio), pero otra parte podría estar relacionada con la calidad de enseñanza percibida (los estudiantes faltan a clase porque el profesor es confuso).

Las implicaciones sociales por el reporte de resultados positivos falsos debido al hecho de ignorar los conglomerados pueden ser costosas. El administrador de una universidad podría decidir dar a las profesoras una ventaja injusta al determinar aumentos con base (parcial) en las calificaciones otorgadas por los estudiantes; un investigador médico podría concluir que un nuevo medicamento con más efectos colaterales que el tratamiento estándar es más eficaz para combatir una enfermedad, aunque el significado estadístico se deba a la inflación por conglomerados del tamaño de la muestra; un funcionario del gobierno podría decidir que se necesita un nuevo programa social para remediar una "desigualdad" demostrada en la prueba de hipótesis. El mismo problema ocurre fuera de las encuestas con muestras, particularmente en la bioestadística. Los conglomerados pueden corresponder a parejas de ojos, a pacientes del mismo hospital o a mediciones repetidas de la misma persona.

¿Es serio el problema de los conglomerados en las encuestas reales? Varios estudios han encontrado que puede serlo. Holt *et al* (1980) encontraron que el nivel de significación real para las pruebas realizadas en forma nominal en el nivel $\alpha = 0.05$ variaron de 0.05 a 0.50. Fay (1985) hace referencia a varios estudios, demostrando que las estadísticas de prueba basadas en una muestra aleatoria simple "pueden arrojar resultados en extremo erróneos al aplicarse a datos que surgen de un diseño complejo de encuesta". El estudio de simulación en Thomas *et al* (1996) calculó los niveles de significación reales alcanzados por X^2 y G^2 cuando el nivel de significación nominal se estableció en $\alpha = 0.05$; ellos encontraron niveles de significación reales entre 0.30 y 0.40.

10.3 Correcciones a las pruebas ji cuadrada

Se han propuesto varios métodos para tomar en cuenta el diseño de muestreo al realizar pruebas de bondad de ajuste, homogeneidad de las poblaciones e independencia de variables. Thomas *et al* (1996) describen más de 25 métodos desarrollados para probar la independencia en tablas de dos sentidos y proporcionan una útil bibliografía. Algunos de estos métodos y variaciones están descritos más detalladamente en Rao y Thomas (1988; 1989). Fay (1985) describe un método alternativo que implica el uso del método de la navaja (*jackknife*) sobre la propia estadística de prueba.

En esta sección bosquejaremos algunos de los métodos básicos para probar la independencia de las variables; la teoría para las pruebas de bondad de ajuste y las pruebas para la homogeneidad de las proporciones es similar. Sin embargo, en las encuestas complejas, a diferencia del muestreo multinomial y multinomial producto, las pruebas para la independencia y la homogeneidad de las proporciones no son necesariamente las mismas. Holt *et al* (1980) observan que con frecuencia (pero no siempre) los conglomerados tienen menos efecto sobre las pruebas de independencia que sobre las pruebas de bondad de ajuste o de homogeneidad de las proporciones.

Recuerde de (10.1) que la hipótesis nula de independencia es

$$H_0 : p_{ij} = p_{i+}p_{+j} \quad \text{para } i = 1,\ldots,r \text{ y } j = 1,\ldots,c.$$

Para una tabla 2×2, $p_{ij} = p_{i+}p_{+j}$ es equivalente a $p_{11}p_{22} - p_{12}p_{21} = 0$ para toda i y j, de modo que la hipótesis nula se reduce a una sola ecuación. En general, la hipótesis nula se puede expresar como $(r-1)(c-1)$ ecuaciones distintas, lo que conduce a $(r-1)(c-1)$ grados de libertad para el muestreo multinomial. Sea

$$\theta_{ij} = p_{ij} - p_{i+}p_{+j}.$$

Entonces la hipótesis nula de independencia es

$$H_0 : \theta_{11} = 0, \quad \theta_{12} = 0,\ldots,\theta_{r-1,c-1} = 0.$$

10.3.1 Pruebas de Wald

La prueba de Wald (1943) fue la primera que se utilizó para probar la independencia en encuestas complejas (Koch *et al* 1975). Para la tabla de 2×2, la hipótesis nula implica una cantidad,

$$\theta = \theta_{11} = p_{11} - p_{1+}p_{+1} = p_{11}p_{22} - p_{12}p_{21},$$

y estimamos a θ como

$$\hat{\theta} = \hat{p}_{11}\hat{p}_{22} - \hat{p}_{12}\hat{p}_{21}.$$

La cantidad θ es una función suave de los totales de la población, de modo que podemos encontrar una estimación de $V(\hat{\theta})$ usando uno de los métodos del capítulo 9. Si los tamaños

de muestra son lo bastante grandes y H_0: $\theta = 0$ es cierta, entonces

$$\frac{\hat{\theta}}{\sqrt{\hat{V}(\hat{\theta})}}$$

sigue aproximadamente una distribución normal estándar. En forma equivalente, bajo H_0, la **estadística de Wald**

$$X^2_W = \frac{\hat{\theta}^2}{\hat{V}(\hat{\theta})} \tag{10.8}$$

sigue aproximadamente una distribución χ^2 con un grado de libertad.

EJEMPLO 10.5 Analicemos la asociación entre "¿hay alguien en su familia que haya sido encarcelado?" (variable *famtime*) y "¿ha sido puesto bajo libertad condicional o enviado a una institución correccional por un delito con violencia?" (variable *everviol*) usando los datos de la encuesta a jóvenes bajo custodia. Un total de $n = 2588$ jóvenes en la encuesta tuvieron respuestas a ambas preguntas. La siguiente tabla muestra la suma de los pesos de cada categoría. Observe que podemos calcular esta tabla usando SAS con la variable de ponderación, pero la prueba ji cuadrada de SAS está completamente equivocada, pues actúa como si hubiese 24,699 observaciones. En este caso, SAS, con los pesos, da $X^2 = G^2 = 11.6$, con el "valor p" incorrecto < 0.001.

		¿Alguna vez violento? No	Sí	
¿Algún miembro de la familia encarcelado?	No	4,761	7,154	11,915
	Sí	4,838	7,946	12,784
		9,599	15,100	24,699

Esto produce la siguiente tabla de proporciones estimadas:

		¿Alguna vez violento? No	Sí	
¿Algún miembro de la familia encarcelado?	No	.1928	.2896	.4824
	Sí	.1959	.3217	.5176
		.3887	.6113	1.0000

Así,

$$\hat{\theta} = \hat{p}_{11}\hat{p}_{22} - \hat{p}_{12}\hat{p}_{21} = \hat{p}_{11} - \hat{p}_{1+}\hat{p}_{+1} = 0.0053$$

Una forma de estimar la varianza de $\hat{\theta}$ es calcular $\hat{p}_{11}\hat{p}_{22} - \hat{p}_{12}\hat{p}_{21}$ para cada uno de los siete grupos aleatorios, como analizamos en el ejemplo 9.4, y determinar la varianza de las siete estimaciones, casi independientes, de θ. Las siete estimaciones, con su promedio y desviación estándar, son

Grupo aleatorio	$\hat{\theta}$
1	0.0132
2	0.0147
3	0.0252
4	−0.0224
5	0.0073
6	−0.0057
7	0.0135
Promedio	0.0065
Desviación estándar	0.0158

Usamos el método de grupos aleatorios para ver que el error estándar de $\hat{\theta}$ es $0.0158/\sqrt{7} = 0.0060$, de modo que la estadística de prueba es

$$\frac{\hat{\theta}}{\sqrt{\hat{V}(\hat{\theta})}} = 0.89.$$

Como nuestra estimación de la varianza a partir del método de grupos aleatorios tiene sólo seis grados de libertad, comparamos la estadística de prueba con una distribución t_6 en vez de una distribución normal estándar. Esta prueba no muestra evidencias de una asociación entre los dos factores cuando analizamos la población como un todo, pero la prueba de hipótesis no dice nada acerca de las posibles asociaciones entre las dos variables en las subpoblaciones; por ejemplo, podría ocurrir que la violencia y el encarcelamiento de algún miembro de la familia estén asociadas positivamente entre los menos jóvenes y negativamente entre los más jóvenes; necesitaríamos estudiar las subpoblaciones por separado o ajustar un modelo log-lineal para ver si éste es el caso. ■

Para tablas mayores, sea $\theta = [\theta_{11}\theta_{12}\ldots\theta_{r-1,c-1}]^T$ (donde T indica la "transpuesta") $(r-1)\,(c-1)$ el vector de θ_{ij}, de modo que la hipótesis nula es

$$H_0\colon \theta = 0.$$

La estadística de Wald es entonces

$$X^2_W = \hat{\theta}^T \hat{V}(\hat{\theta})^{-1} \hat{\theta}$$

donde $\hat{V}(\hat{\theta})$ es la matriz de covarianzas estimadas de $\hat{\theta}$. En muestras muy grandes y bajo H_0, X^2_W sigue aproximadamente una distribución $\chi^2_{(r-1)(c-1)}$. Pero "grande" en una encuesta compleja se refiere a un gran número de unidades primarias de muestreo, no necesariamente a una gran cantidad de unidades de observación. En una tabla de contingencias de 4×4, $\hat{V}(\hat{\theta})$ es una matriz de 9×9 y requiere el cálculo de 45 varianzas y covarianzas distintas. Si una muestra por conglomerados sólo tiene 50 unidades primarias, la matriz de covarianzas estimadas será muy inestable. En la práctica, la prueba de Wald para tablas de contingencias grandes frecuentemente se desempeña de manera pobre y no recomendamos su uso. Algunas modificaciones de la prueba de Wald se desempeñan mejor; los detalles aparecen en Thomas *et al* (1996).

10.3.2 Pruebas de Bonferroni

La hipótesis nula de independencia,

$$H_0 : \theta_{11} = 0,\ \theta_{12} = 0, \ldots,\ \theta_{r-1,c-1} = 0,$$

tiene $m = (r-1)(c-1)$ componentes:

$$H_0(1): \theta_{11} = 0$$
$$H_0(2): \theta_{12} = 0$$
$$\vdots$$
$$H_0(m): \theta_{(r-1)(c-1)} = 0.$$

En lugar de utilizar la covarianza estimada de todas las $\hat{\theta}_{ij}$ como en la prueba de Wald, podemos utilizar la desigualdad de Bonferroni para probar cada componente $H_0(k)$ por separado con un nivel de significación α/m (Thomas 1989). El procedimiento de Bonferroni proporciona una prueba conservadora. H_0 será rechazada en el nivel α si cualquiera de las $H_0(k)$ es rechazada en el nivel α/m; es decir, si

$$\frac{|\hat{\theta}_{ij}|}{\sqrt{\hat{V}(\hat{\theta}_{ij})}} > t_\kappa\left(\frac{\alpha}{2m}\right)$$

para cualquier i y j. Cada estadística de prueba se compara con una distribución t_k, donde el estimador de la varianza tiene k grados de libertad. Si se usa el método de grupos aleatorios para estimar la varianza, entonces k es igual a (número de grupos) − 1; si se usa otro método, k es igual a (número de unidades primarias) − (número de estratos).

Aunque ésta es una prueba conservadora, aparentemente funciona bien en la práctica. Además, es fácil de implantar, en particular si se usa un método de remuestreo para estimar las varianza, ya que cada prueba se puede realizar por separado.

EJEMPLO 10.6 En la encuesta a jóvenes en custodia, analicemos la relación entre la edad y el hecho de que el joven haya sido enviado a la institución por un delito con violencia (usando la variable *crimtype*, *currviol* se define como 1 si *crimtype* = 1; y 0 en caso contrario). Usamos los pesos para estimar la proporción de la población que cae en cada celda:

		Grupo de edad			
		≤ 15	16 o 17	≥ 18	
¿Delito con violencia?	No	.1698	.2616	.1275	.5589
	Sí	.1107	.1851	.1453	.4411
		.2805	.4467	.2728	1.0000

La hipótesis nula es

$$H_0: \theta_{11} = p_{11} - p_{1+}p_{+1} = 0$$
$$\theta_{12} = p_{12} - p_{1+}p_{+2} = 0.$$

Primero veamos qué ocurre si ignoramos los conglomerados y pretendemos que la estadística de prueba en (10.2) sigue una distribución χ^2 con dos grados de libertad. Con n = 2621 jóvenes en la tabla, la estadística X^2 de Pearson es

$$X^2 = n\sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}} = 34.$$

Al comparar esto con una distribución χ^2_2 obtenemos un "valor p" incorrecto de 4×10^{-8}.

Ahora, usemos la prueba de Bonferroni. Para estos datos, $\hat{\theta}_{11} = 0.0130$ y $\hat{\theta}_{12}0.0119$. Usamos el método de grupos aleatorios para estimar las varianzas, como en el ejemplo 10.5,

para obtener las siete estimaciones:

Grupo aleatorio	$\hat{\theta}_{11}$	$\hat{\theta}_{12}$
1	−0.0195	0.0140
2	0.0266	−0.0002
3	0.0052	0.0159
4	0.0340	0.0096
5	0.0197	0.0202
6	0.0025	0.0298
7	−0.0103	0.0143

Así, $\text{EE}(\hat{\theta}_{11}) = 0.0074$, $\text{EE}(\hat{\theta}_{12}) = 0.0035$, $\hat{\theta}_{11}/\text{EE}(\hat{\theta}_{11}) = 1.8$, y $\hat{\theta}_{12}/\text{EE}(\hat{\theta})_{12} = 3.4$. El percentil 0.9875 de una distribución t con seis grados de libertad es 2.97; como la estadística de prueba para $H_0(2)$: $\theta_{12} = 0$ excede este valor crítico, rechazamos la hipótesis nula en el nivel 0.05. ■

10.3.3 Comparación de los momentos con los momentos de la distribución ji cuadrada

Las estadísticas de prueba X^2 y G^2 no siguen una distribución $\chi^2_{(r-1)(c-1)}$ en una encuesta compleja bajo la hipótesis nula de independencia, pero ambas estadísticas tienen una distribución simétrica y un múltiplo de X^2 o G^2 podría seguir aproximadamente una distribución χ^2.

Podemos obtener una **corrección de primer orden** comparando la media de la estadística de prueba con la media de la distribución $\chi^2_{(r-1)(c-1)}$ (Rao y Scott 1981; 1984). La media de una distribución $\chi^2_{(r-1)(c-1)}$ es $(r-1)(c-1)$; podemos calcular $E[X^2]$ o $E[G^2]$ bajo el diseño de muestreo complejo cuando H_0 es verdadera y comparar la estadística de prueba

$$X_F^2 = \frac{(r-1)(c-1)X^2}{E[X^2]}$$

o

$$G_F^2 = \frac{(r-1)(c-1)G^2}{E[G^2]}$$

con una distribución $\chi^2_{(r-1)(c-1)}$. Bedrick (1983) y Rao y Scott (1984) muestran que bajo H_0,

$$E[X^2] \approx E[G^2]$$
$$\approx \sum_{i=1}^{r}\sum_{j=1}^{c}(1-p_{ij})d_{ij} - \sum_{i=1}^{r}(1-p_{i+})d_i^R - \sum_{j=1}^{c}(1-p_{+j})d_j^C, \qquad (10.9)$$

donde d_{ij} es el efecto de diseño para la estimación de p_{ij}, d_i^R es el efecto de diseño para la estimación de p_{i+} y d_j^C es el efecto de diseño para la estimación de p_{+j}. En la práctica, si el estimador de las varianzas por celda tiene k grados de libertad, es un poco mejor comparar $X_F^2/(r-1)(c-1)$ o $G_F^2/(r-1)(c-1)$ con una distribución F con $(r-1)(c-1)$ y $(r-1)(c-1)\kappa$ grados de libertad.

Frecuentemente, la corrección de primer orden se puede usar con tablas ya publicadas, pues sólo hay que estimar las varianzas de las proporciones en la tabla de contingencias; no hay que estimar toda la matriz de covarianza de las $\hat{p}_{ij}$, como en la prueba de Wald. Sólo estamos ajustando la estadística de prueba de modo que su media bajo H_0 sea $(r-1)(c-1)$; los valores p de interés provienen de la cola de la distribución de referencia y no necesariamente ocurre que la cola de la distribución de X_F^2 concuerda con la cola de la distribución

$\chi^2_{(r-1)(c-1)}$. Rao y Scott (1981) muestran que X_F^2 y G_F^2 tienen una distribución χ^2 nula si, y sólo si, los efectos de diseño para las varianzas y covarianzas de las $\hat{p}_{ij}$ son iguales. En caso contrario, la varianza de X_F^2 es mayor que la varianza de una distribución $\chi^2_{(r-1)(c-1)}$ y los valores p de X_F^2 son con frecuencia un poco menores que lo debido (pero más cercanos a los valores p reales si no se realizó corrección alguna).

EJEMPLO 10.7 También podemos realizar la prueba de hipótesis en el ejemplo 10.6 usando la corrección de primer orden. Estimamos los siguientes efectos de diseño usando el método de grupos aleatorios para estimar las varianzas por celda:

		Grupo de edad			
		≤ 15	16 o 17	≥ 18	
¿Crimen con violencia?	No	20.2	1.9	2.8	5.7
	Sí	5.3	8.4	2.4	5.7
		22.0	9.7	4.3	

Varios de los efectos de diseño son muy grandes, como sería de esperar, pues algunas instalaciones tienen principalmente delincuentes violentos o principalmente no violentos. Por ejemplo, todos los residentes de la Instalación 31 están ahí por un delito violento. Además, las instalaciones que tienen principalmente delincuentes no violentos tienden a ser más grandes. Entonces, es de esperar que los conglomerados ejerzan un papel sustancial sobre la prueba de hipótesis.

Usamos (10.9) para estimar $E[X^2]$ como 4.2 y $X_F^2 = 2X^2/4.2 = 16.2$. Al comparar 16.2/2 con una distribución $F_{2,12}$ (la estimación de la varianza mediante grupos aleatorios tiene seis grados de libertad) da un valor p aproximado de 0.006. Es probable que este valor p siga siendo demasiado pequeño debido a la gran disparidad en los efectos de diseño. ■

Rao y Scott (1981; 1984) proponen también una **corrección de segundo orden**—comparar la media y la varianza de la estadística de prueba con la media y la varianza de una distribución χ^2, como en las pruebas de modelos de análisis de la varianza de Satterthwaite (1946)—. Satterthwaite quien comparó una estadística de prueba T con distribución simétrica con una distribución de referencia χ^2, eligiendo una constante k y ν grados de libertad, de modo que $E[kT] = \nu$ y $V[kT] = 2\nu$ (ν y 2ν son la media y la varianza de una distribución χ^2 con ν grados de libertad). En este caso, si $m = (r-1)(c-1)$, sabemos que $E[kX_F^2] = km$ y

$$V[kX_F^2] = V\left(\frac{kmX^2}{EX^2}\right) = \frac{V[X^2]k^2m^2}{[E(X^2)]^2},$$

así que al comparar los momentos tenemos que

$$\nu = 2\frac{[E(X^2)]^2}{V[X^2]} \quad \text{y} \quad k = \frac{\nu}{m}.$$

Entonces,

$$X_S^2 = \frac{\nu X_F^2}{(r-1)(c-1)} \tag{10.10}$$

se compara con una distribución χ^2 con ν grados de libertad. La estadística G_S^2 se forma de manera análoga. De nuevo, si el estimado de las varianzas de las $\hat{p}_{ij}$ tiene κ grados de libertad, es un poco mejor comparar X_S^2/ν o G_S^2/ν con una distribución F con ν y $\nu\kappa$ grados de libertad.

En general, la estimación de $V[X^2]$ es un tanto complicada y requiere toda la matriz de covarianzas de las $\hat{p}_{ij}$, de modo que, con frecuencia, la corrección no se puede utilizar cuando sólo se dispone de datos en tablas ya publicadas. Si todos los efectos de diseño son similares, las correcciones de primero y segundo orden se comportarán de manera análoga. Sin embargo, cuando los efectos de diseño varían apreciablemente, los valores p obtenidos mediante X_F^2 podrían ser demasiado pequeños y X_S^2 puede ser mejor. El ejercicio 11 indica la forma de calcular la corrección de segundo orden.

10.3.4 Métodos basados en el modelo para las pruebas ji cuadrada

Todos los métodos analizados utilizan las estimaciones de covarianza de las proporciones para ajustar las pruebas ji cuadrada. También se puede utilizar un punto de vista basado en el modelo. Describiremos el modelo de Cohen (1976) para una muestra por conglomerados con dos unidades de observación por conglomerado. Algunas extensiones y otros modelos utilizados para el muestreo por conglomerados aparecen en Altham (1976), Brier (1980), Rao y Scott (1981), y Wilson y Koehler (1991). Estos modelos suponen que el efecto de diseño es el mismo para cada celda y margen.

EJEMPLO 10.8 Cohen (1976) presenta un ejemplo que explora la relación entre el sexo y el diagnóstico de esquizofrenia. Los datos fueron 71 parejas hospitalizadas de hermanos. Muchas enfermedades mentales tienden a aparecer por familias, de modo que podríamos esperar que si a un hermano se le diagnostica esquizofrenia, es más probable que el otro hermano tenga el mismo diagnóstico. Así, es probable que cualquier análisis que ignore la dependencia entre hermanos dé valores p demasiado pequeños. Si sólo clasificamos los 142 pacientes por género y diagnóstico e ignoramos la correlación entre hermanos, obtenemos la siguiente tabla. En este caso, S indica que al paciente se le diagnosticó esquizofrenia, y N significa que al paciente no se le diagnosticó esquizofrenia.

	S	N	
Masculino	43	15	58
Femenino	32	52	84
	75	67	142

Si analizamos estos datos en un paquete estadístico estándar (yo usé SAS), $X^2 = 17.89$ y $G^2 = 18.46$. Sin embargo, recuerde que SAS supone que todas las observaciones son independientes, de modo que el "valor p" de 0.00002 es incorrecto.

Sin embargo, conocemos la estructura de conglomerado de los 71 conglomerados. Usted puede ver en la tabla 10.1 que la mayoría de las parejas caen en los bloques diagonales: si un hermano padece esquizofrenia, es más probable que el otro también. En 52 de las parejas de hermanos, ambos fueron diagnosticados como esquizofrénicos o ambos fueron diagnosticados como no esquizofrénicos.

Sea q_{ij} la probabilidad de que una pareja caiga en la celda (i, j) en la clasificación de las parejas. Así, q_{11} es la probabilidad de que ambos hermanos sean esquizofrénicos y hombres, y q_{12} es la probabilidad de que el hermano más joven sea una mujer esquizofrénica y el mayor sea un hombre esquizofrénico, etcétera. Entonces, modelamos los q_{ij} mediante

$$q_{ij} = \begin{cases} aq_i + (1-a)q_i^2 & \text{si } i = j \\ (1-a)q_iq_j & \text{si } i \neq j \end{cases} \tag{10.11}$$

donde a es un efecto del conglomerado y q_i es la probabilidad de que un individuo esté en la clase i (i = SM, SF, NM, NF). Si $\alpha = 0$, los miembros de una pareja son independientes y

TABLA 10.1
Información por conglomerados de las 71 parejas de hermanos

		Hermano menor SM	SF	NM	SM	
Hermano mayor	SM	13	5	1	3	22
	SF	4	6	1	1	12
	NM	1	1	2	4	8
	NF	3	8	3	15	29
		21	20	7	23	71

simplemente podemos realizar la prueba ji cuadrada regular usando los individuos; compararíamos la X^2 usual de Pearson, calculada ignorando los conglomerados, con una distribución $\chi^2_{(r-1)(c-1)}$. Si $a = 1$, los dos hermanos tienen una correlación perfecta, de modo que esencialmente tenemos una única pieza de información de cada pareja; $X^2/2$ compararíamos con una distribución $\chi^2_{(r-1)(c-1)}$. Para a entre 0 y 1, si el modelo es válido, $X^2/(1+a)$ seguirá aproximadamente una distribución $\chi^2_{(r-1)(c-1)}$ si la hipótesis nula es verdadera.

El modelo se puede ajustar mediante máxima verosimilitud (vea los detalles en Cohen 1976). Entonces $\hat{a} = .3006$, y las probabilidades estimadas para las cuatro celdas son las siguientes:

	S	N	
Masculino	0.2923	0.1112	0.4035
Femenino	0.2330	0.3636	0.5966
	0.5253	0.4748	1.0000

Podemos verificar el modelo usando una prueba de bondad de ajuste para los datos por conglomerados de la tabla 10.1. Este modelo no exhibe una carencia de ajuste significativa, mientras que el modelo que supone la independencia sí la exhibe. Para verificar si el género y la esquizofrenia son independientes en la tabla 1.3006 = 2 × 2, X^2/13.76, lo que comparamos con una distribución χ^2_1. El valor p resultante es 0.0002, cerca de 10 veces más grande que el valor p del análisis que pretendía que los hermanos fuesen independientes. ■

10.4 Modelos log-lineales

Si hay más de dos variables de clasificación, con frecuencia nos interesa comprobar si existen relaciones más complejas entre los datos. Los modelos log-lineales generalmente se utilizan para estudiar estas relaciones.

10.4.1 Modelos log-lineales con muestreo multinomial

En una tabla de dos sentidos, si la variable por renglón y la variable por columna son independientes, entonces $p_{ij} = p_{i+}\, p_{+j}$. En forma equivalente,

$$\ln p_{ij} = \ln p_{i+} + \ln p_{+j}$$
$$= \mu + \alpha_i + \beta_j,$$

donde

$$\sum_{i=1}^{r} \alpha_i = 0 \quad \text{y} \quad \sum_{j=1}^{c} \beta_j = 0.$$

Esto se llama **modelo log-lineal** debido a que los logaritmos de las probabilidades por celda siguen un modelo lineal: el modelo para la independencia en una tabla 2 × 2 se puede escribir como

$$\mathbf{y} = \mathbf{X}\beta,$$

donde

$$\mathbf{y} = \begin{bmatrix} \ln(p_{11}) \\ \ln(p_{12}) \\ \ln(p_{21}) \\ \ln(p_{22}) \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix}$$

$$\beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \beta_1 \end{bmatrix}.$$

Estimamos los parámetros β mediante las probabilidades estimadas $\hat{p}_{ij}$. Para los datos del ejemplo 10.1, las probabilidades estimadas son las siguientes:

		¿Computadora? Sí	No	
¿Cable?	Sí	0.238	0.376	0.614
	No	0.176	0.210	0.386
		0.414	0.586	1.000

Las estimaciones de los parámetros son $\hat{\mu} = -1.428$, $\hat{\alpha}_1 = 0.232$, y $\hat{\beta}_1 = -0.174$. Los valores ajustados de $\hat{p}_{ij}$ para el modelo de independencia son entonces

$$\hat{p}_{ij} = \exp(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j)$$

y están dados en la siguiente tabla:

		¿Computadora? Sí	No	
¿Cable?	Sí	0.254	0.360	0.614
	No	0.160	0.226	0.386
		0.414	0.586	1.000

También quisiéramos ver qué tan bien se ajusta este modelo a los datos. Podemos hacer esto de dos formas:

1 Probar la bondad de ajuste del modelo, usando X^2 en (10.5) o G^2 en (10.6): para una tabla de contingencias de dos sentidos, estas estadísticas son equivalentes a las estadísticas para probar la independencia. Para el ejemplo computadora/cable, la estadística de razón de verosimilitud para la bondad de ajuste es 2.27. En el muestreo multinomial, X^2 y G^2 siguen aproximadamente una distribución $\chi^2_{(r-1)(c-1)}$ si el modelo es correcto.
2 Podemos escribir el siguiente modelo completo, o saturado, para los datos:

$$\ln p_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

con $\sum_{i=1}^{r}(\alpha\beta)_{ij} = \sum_{j=1}^{c}(\alpha\beta)_{ij} = 0$. El último término es análogo al término de interacción en un modelo de análisis de la varianza de dos sentidos. Este modelo dará un ajuste perfecto a las probabilidades observadas por celda, pues tiene $r\,c$ parámetros. La hipótesis nula de independencia es equivalente a

$$H_0 : (\alpha\beta)_{ij} = 0 \quad \text{para } i = 1,\ldots,r-1;\ j = 1,\ldots,\ c-1.$$

Los paquetes estándar de estadística, como SAS, proporcionan estimaciones de los $\alpha\beta_{ij}$ y sus errores estándar asintóticos bajo el muestreo multinomial. Para el modelo saturado en el ejemplo computadora/cable, el procedimiento CATMOD de SAS da el siguiente resultado:

Efecto	Parámetro	Estimado	Error estándar	Ji-cuadrada	Prob
CABLE	1	0.2211	0.0465	22.59	0.0000
COMP	2	–0.1585	0.0465	11.61	0.0007
CABLE*COMP	3	–0.0702	0.0465	2.28	0.1313

Los valores en la columna *Ji-cuadrada* son las estadísticas de prueba de Wald para la prueba, si ese parámetro se anula. Así, bajo el muestreo multinomial, el valor p para verificar si el término de interacción se anula es 0.1313; de nuevo, para este ejemplo, éste es exactamente igual al valor p de la prueba para independencia.

10.4.2 Modelos log-lineales en una encuesta compleja

¿Qué ocurre en una encuesta compleja? Obtenemos estimaciones puntuales de los parámetros del modelo, como siempre, usando pesos. Así, estimamos las p_{ij} mediante (10.7) y usamos las estimaciones $\hat{p}_{ij}$ del software estándar para estimar los parámetros del modelo. Sin embargo, como es usual, las estadísticas de prueba para la bondad de ajuste y los errores estándar asintóticos para las estimaciones de parámetros dadas por SAS son incorrectas. Scheuren (1973) analiza algunos de los retos al ajustar modelos log-lineales a datos de CPS.

Muchas de las correcciones utilizadas para las pruebas ji cuadrada de independencia también se pueden usar para la prueba de hipótesis en modelos log-lineales. Rao y Thomas (1988; 1989) y Fay (1985) describen varias pruebas de bondad de ajuste para tablas de contingencia de encuestas complejas, incluyendo las pruebas de Wald, de la navaja y las correcciones de primero y segundo orden a X^2 y G^2.

También podemos usar la desigualdad de Bonferroni para comparar los modelos log-lineales anidados. Por ejemplo, para probar la independencia en una tabla de dos sentidos,

comparamos el modelo saturado con el modelo reducido de independencia y verificamos cada una de las $m = (r-1)(c-1)$ hipótesis nulas

$$H_0(1) : (\alpha\beta)_{11} = 0$$
$$\vdots$$
$$H_0(m) : (\alpha\beta)_{(r-1)(c-1)} = 0$$

por separado en el nivel α/m.

Más en general, mediante este método podemos comparar cualquiera de dos modelos log-lineales anidados. Para una tabla tridimensional $r \times c \times d$, sea

$$\mathbf{y} = [\ln(p_{111}), \ln(p_{112}), \ldots, \ln(p_{rcd})]^T.$$

Suponga que el modelo menor es

$$\mathbf{y} = \mathbf{X}\beta,$$

y que el modelo mayor es

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\theta$$

donde θ es un vector de longitud m. Entonces podemos ajustar el modelo mayor y realizar m pruebas de hipótesis por separado de las hipótesis nulas

$$H_0 : \theta_i = 0,$$

cada una al nivel α/m, comparando $\hat{\theta}_i / \mathrm{EE}(\hat{\theta}_i)$ con una distribución t.

EJEMPLO 10.9 Analicemos una tabla tridimensional obtenida a partir de la encuesta a jóvenes en custodia, para examinar las relaciones entre las variables "¿hay alguien en su familia que haya sido encarcelado?", (*famtime*), y "¿ha sido puesto en libertad condicional o enviado a alguna institución correccional por un delito violento?" (*everviol*), y *age* (edad), para las observaciones sin datos faltantes. Las probabilidades por celda son p_{ijk}. Las probabilidades estimadas $\hat{p}_{ijk}$, estimadas usando pesos, están en la siguiente tabla:

		¿Algún miembro de la familia encarcelado? No, ¿Alguna vez violento? No	No, ¿Alguna vez violento? Sí	Sí, ¿Alguna vez violento? No	Sí, ¿Alguna vez violento? Sí	
Grupo de edad	≤ 15	0.0588	0.0698	0.0659	0.2801	0.2801
	16-17	0.0904	0.1237	0.0944	0.4461	0.4461
	≥ 18	0.0435	0.0962	0.0355	0.2738	0.2738
		0.1928	0.2896	0.1959	0.3217	1.0000

El modelo saturado para la tabla de tres sentidos es

$$\log p_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}.$$

El procedimiento CATMOD de SAS, haciendo uso de los pesos, da las siguientes estima-

ciones de los parámetros para el modelo saturado:

Efectos	Parámetro	Estimado	Error estándar	Ji-cuadrada	Prob
AGECLASS	1	–0.1149	0.00980	137.45	0.0000
	2	0.3441	0.00884	1515.52	0.0000
EVERVIOL	3	–0.2446	0.00685	1275.26	0.0000
AGECLASS*EVERVIOL	4	0.1366	0.00980	194.27	0.0000
	5	0.0724	0.00884	67.04	0.0000
FAMTIME	6	0.0242	0.00685	12.51	0.0004
AGECLASS*FAMTIME	7	0.0555	0.00980	32.03	0.0000
	8	0.0128	0.00884	2.10	0.1473
EVERVIOL*FAMTIME	9	–0.0317	0.00685	21.42	0.0000
AGECLASS*EVERVIOL*FAMTIME	10	0.0089	0.00980	0.82	0.3646
	11	0.0161	0.00884	3.33	0.0680

Como ésta es una encuesta compleja y como SAS actúa como si el tamaño de la muestra fuese Σw_i al usar los pesos, los errores estándar y los valores p de los parámetros son completamente incorrectos. Sin embargo, podemos estimar la varianza de cada parámetro reajustando el modelo log-lineal en cada uno de los grupos aleatorios y usando la estimación de la varianza mediante los grupos aleatorios para realizar pruebas de hipótesis sobre los parámetros individuales. Los errores estándar bajo los grupos aleatorios para los 11 parámetros del modelo aparecen en la tabla 10.2. La hipótesis nula de no interacción entre las variables es

$$H_0 : (\alpha\beta)_{ij} = (\alpha\gamma)_{ik} = (\beta\gamma)_{jk} = (\alpha\beta\gamma)_{ijk} = 0;$$

o bien, usando la numeración de parámetros de SAS,

$$H_0 : \beta_4 = \beta_5 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0.$$

Esta hipótesis nula tiene siete componentes. Para usar la prueba de Bonferroni, probamos cada parámetro individual en el nivel 0.05/7. El percentil $(1 - 0.05/14)$ de una distribución

TABLA 10.2
Errores estándar por grupos aleatorios para el ejemplo 10.9

Parámetro	Estimación	Error estándar	Estadística de prueba
1	–0.1149	0.1709	–0.67
2	0.3441	0.0953	3.61
3	–0.2446	0.0589	–4.15
4	0.1366	0.0769	1.78
5	0.0724	0.0379	1.91
6	0.0242	0.0273	0.89
7	0.0555	0.0191	2.91
8	0.0128	0.0218	0.59
9	–0.0317	0.0233	–1.36
10	0.0089	0.0191	0.47
11	0.0161	0.0167	0.96

t_6 es 4.0; ninguna de las estadísticas de prueba $\hat{\beta}_i/\text{EE}(\hat{\beta}_i)$, para $i = 4,5,7,8,9,10,11$, excedió ese valor crítico, de modo que no rechazaríamos la hipótesis nula en el sentido de que las tres variables sean independientes. Sin embargo, tal vez deberíamos explorar con más detalle la interacción *ageclass*famtime*. ■

Los paquetes para encuestas SUDAAN, PC CARP y WesVarPC, entre otros, realizan pruebas de hipótesis usando datos de encuestas complejas. En la sección 9.6 analizamos brevemente estos paquetes.

10.5 Ejercicios

1 Encuentre un ejemplo o ejercicio en algún libro de texto de introducción a la estadística que realice una prueba ji cuadrada sobre los datos de una encuesta. ¿Qué diseño cree que se haya utilizado para la encuesta? ¿Es adecuada una prueba ji cuadrada para muestreo multinomial para estos datos? ¿Por qué?

2 Lea uno de los siguientes artículos o algún otro artículo de investigación en donde se realice un análisis de datos categóricos sobre datos de una encuesta compleja. Describa el diseño de muestreo y el método de análisis. ¿Tomaron los autores en cuenta el diseño en su análisis de los datos? ¿Deberían haberlo hecho?

Gold, M.R., R. Hurley, T. Lake, T. Ensor, y R. Berenson, 1995. A national survey of the arrangements managed-care plans make with physicians. *New England Journal of Medicine* 333: 1689-1683.

Koss, M.P., C.A. Gidycz, y N. Wisniewski. 1987. The scope of rape: Incidence and prevalence of sexual aggression and victimization in a national sample of higher education students. *Journal of Consulting and Clinical Psychology* 55:162-170.

Lipton, R. B., W.F. Stewart, D.F. Celentano, y M.L. Reed. 1992. Undiagnosed migraine headaches: A comparison of symptom-based and reported physician diagnosis. *Archives of Internal Medicine* 152: 1273-1278.

Sarti, E., P.M. Schantz, A. Plancarte, M. Wilson, I. Gutiérrez, A. López, J. Roberts, y A. Flisser. 1992. Prevalence and risk factors for *Taenia Solium* taeniasis and cysticercosis in humans and pigs in a village in Morelos, Mexico. *American Journal of Tropical Medicine and Hygiene* 46: 677-685.

3 Schei y Bakketeig (1989) extrajeron una muestra aleatoria simple de 150 mujeres de entre 20 y 49 años de edad de la ciudad de Trondheim, Noruega. Su objetivo era investigar la relación entre el abuso sexual y físico por parte del esposo y ciertos síntomas ginecológicos en la mujer. De las 150 mujeres seleccionadas para la muestra, 15 se habían mudado, una había muerto, tres fueron excluidas por no ser elegibles para el estudio y 13 se rehusaron a participar.

De las 118 mujeres que participaron en el estudio, 20 reportaron algún tipo de abuso sexual o físico por parte de su pareja. Ocho reportaron haber sido golpeadas, dos pateadas o mordidas, siete recibieron una golpiza y tres fueron amenazadas o cortadas con un cuchillo; 17 de las mujeres en el estudio reportaron un síntoma ginecológico de sangrado irregular o dolor pélvico. Las cantidades de mujeres en cada una de las cuatro categorías de síntomas

ginecológicos y abuso por parte de su marido están en la siguiente tabla:

		Abuso		
		No	Sí	
¿Síntomas ginecológicos presentes?	No	89	12	101
	Sí	9	8	17
		98	20	118

a Si el abuso y la presencia de síntomas ginecológicos no están asociados, ¿cuáles son las probabilidades esperadas en cada una de las cuatro celdas?

b Realice una prueba ji cuadrada de asociación de las variables *abuso* y *presencia de síntomas ginecológicos*.

c ¿Cuál es la tasa de respuesta de este estudio? ¿Qué definición de tasa de respuesta utilizó? ¿Cree que la ausencia de respuesta podría afectar las conclusiones del estudio? Justifique su respuesta.

4 Samuels (1996) reunió datos para examinar el desempeño de los estudiantes en cursos de seguimiento si el curso de prerrequisitos lo imparte un instructor de tiempo completo o de tiempo parcial. La siguiente tabla muestra los resultados de los estudiantes de los cursos Matemáticas I y Matemáticas II.

Instructor de Matemáticas I	Instructor de Matemáticas II	Calificación en Matemáticas II 8, 9 o 10	6, 7 o abandono	
Tiempo completo	Tiempo completo	797	461	1258
Tiempo completo	Tiempo parcial	311	181	492
Tiempo parcial	Tiempo completo	570	480	1050
Tiempo parcial	Tiempo parcial	909	449	1358
		2587	1571	4158

a La hipótesis nula aquí es que la proporción de estudiantes que obtuvieron 8, 9 o 10 es la misma para cada una de las cuatro combinaciones de tipo de instructor. ¿Es ésta una prueba de independencia, de homogeneidad o de bondad de ajuste?

b Realice una prueba de hipótesis para la hipótesis nula en la parte (a), suponiendo que los estudiantes son independientes.

c ¿Cree que la hipótesis de que los estudiantes son independientes sea válida? Justifique su respuesta.

5 Use el archivo winter.dat para este ejercicio. Analizamos por primera vez los datos en el ejercicio 20 del capítulo 4.

a Pruebe la hipótesis nula de que *class* no está asociado con *breakaga*. En el contexto de la sección 10.1, ¿qué tipo de muestreo se realizó?

b Ahora, construya una tabla de contingencias de 2 × 2 para las variables *breakaga* y *work*. Use los pesos de muestreo para estimar las probabilidades p_{ij} para cada celda.

c Calcule los cocientes de posibilidades usando las $\hat{p}_{ij}$ de la parte (b). ¿Cuál es la relación de éstos con los cocientes de posibilidades calculados mediante las cifras observadas (ignorando los pesos de muestreo)?

d Estime $\theta = p_{11}p_{22} - p_{21}p_{12}$ usando las $\hat{p}_{ij}$ calculadas en la parte (b).

e Pruebe la hipótesis nula H_0: $\theta = 0$.

f ¿Cómo afecta la estratificación a la prueba de hipótesis?

6 Use el archivo teachers.dat para este ejercicio. Analizamos los datos por primera vez en el ejercicio 16 del capítulo 5.

a Construya una nueva variable *zassist* que asuma el valor 1 si un ayudante de profesor ocupa tiempo en apoyar al profesor, y 0 en caso contrario. Construya también la nueva variable *zprep*, que asume los valores bajo, medio y alto con base en la cantidad de tiempo que el maestro ocupa para preparar su clase.

b Construya una tabla de contingencias 2 × 3 para las variables *zassist* y *zprep*. Use los pesos de muestreo para estimar las probabilidades p_{ij} para cada celda.

c Utilice el método de Bonferroni para probar la hipótesis nula de que *zassist* no está asociada con *zprep*.

7 Algunos investigadores han utilizado el siguiente método para realizar pruebas de asociación en tablas de dos sentidos. En lugar de utilizar los pesos originales de observación w_k, definimos

$$w_k^* = \frac{nw_k}{\sum_{l \in S} w_l},$$

donde n es el número de unidades de observación en la muestra. Así, la suma de los nuevos pesos w_k^* es n. La cifra "observada" para la celda (i, j) es

$$x_{ij} = \text{suma de las } w_k^* \text{ para las observaciones en la celda } (i, j),$$

y la cifra "esperada" para la celda (i, j) es

$$\hat{m}_{ij} = \frac{x_{i+}x_{+j}}{n}.$$

Compare la estadística de prueba

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(x_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

con una distribución $\chi^2_{(r-1)(c-1)}$.

¿Dará esta prueba valores p correctos para los datos de una encuesta compleja? ¿Por qué?

SUGERENCIA: trate con los datos de los ejemplos 10.1 y 10.4.

***8** (Requiere cálculo.) Considere X^2_{ip} en (10.8).

a Use el método de linealización de la sección 9.1 para aproximar $V(\hat{\theta})$ en términos de $V(\hat{p}_{ij})$ y $\text{Cov}(\hat{p}_{ij}, \hat{p}_{kl})$.

b ¿Cuál es la estadística de Wald al emplear la estimación de $V(\hat{\theta})$ por linealización de la parte (a), usando el muestreo multinomial? (Bajo el muestreo multinomial, $V(\hat{p}_{ij}) = p_{ij}(1 - p_{ij})/n$ y $\text{Cov}(\hat{p}_{ij}, \hat{p}_{kl}) = -p_{ij}\,p_{kl}/n$. ¿Es lo mismo que la estadística X^2 de Pearson?

9 (Requiere cálculo.) *Estimación del cociente de posibilidades logaritmicas en una encuesta compleja.* Sean

$$\theta = \log\left(\frac{p_{11}p_{22}}{p_{12}p_{21}}\right) \text{ y } \hat{\theta} = \log\left(\frac{\hat{p}_{11}\hat{p}_{22}}{\hat{p}_{12}\hat{p}_{21}}\right).$$

a Use el método de linealización de la sección 9.1 para aproximar $V(\hat{\theta})$ en términos de $V(\hat{p}_{ij})$ y $\text{Cov}(\hat{p}_{ij}, \hat{p}_{kl})$.

b Utilice la parte (a). ¿Cuál es el valor de $V(\hat{\theta})$ bajo el muestreo multinomial?

10 Pruebe que para el muestreo multinomial, $X_F^2 = X^2$. SUGERENCIA: ¿cuál es el valor de $E[X^2]$ en (10.9) para una muestra multinomial?

***11** (Requiere estadística matemática y álgebra lineal.) *Deducción de las correcciones de primer y segundo orden a la X^2 de Pearson* (vea Rao y Scott 1981).

a Suponga que el vector aleatorio $\mathbf{Y}$ se distribuye normalmente con media $\mathbf{0}$ y matriz de covarianza Σ. Entonces, si $\mathbf{C}$ es simétrica, muestre que $\mathbf{Y}^T\mathbf{C}\mathbf{Y}$ tiene la misma distribución que $\Sigma\lambda_i W_i$, donde las W_i son variables aleatorias independientes χ_1^2 y las λ_i son los valores propios de $\mathbf{C}\Sigma$.

b Sea $\hat{\theta} = (\hat{\theta}_{11},\ldots,\hat{\theta}_{1,(c-1)},\ldots,\hat{\theta}_{(r-1),1},\ldots,\hat{\theta}_{(r-1)(c-1)})^T$, donde $\hat{\theta}_{ij} = \hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j}$. Sea $\mathbf{A}$ la matriz de covarianza de $\hat{\theta}$ si se extrae una muestra multinomial de tamaño n y la hipótesis nula es verdadera. Use la parte (a) para argumentar que, asintóticamente, $\hat{\theta}^T\mathbf{A}^{-1}\hat{\theta}$ tiene la misma distribución que $\Sigma\lambda_i W_i$, donde las W_i son variables aleatorias independientes χ_1^2 y las λ_i son los valores propios de $\mathbf{A}^{-1}V(\hat{\theta})$.

c ¿Cuáles son los valores de $E\left[\hat{\theta}^T\mathbf{A}^{-1}\hat{\theta}\right]$ y $V\left[\hat{\theta}^T\mathbf{A}^{-1}\hat{\theta}\right]$ en términos de los λ_i?

d Determine $E\left[\hat{\theta}^T\mathbf{A}^{-1}\hat{\theta}\right]$ y $V\left[\hat{\theta}^T\mathbf{A}^{-1}\hat{\theta}\right]$ para una tabla 2×2. Tal vez desee usar su respuesta al ejercicio *8.

12 Conocemos la estructura de conglomerados de los datos del ejemplo 10.8. Use los resultados del capítulo 5 (suponiendo un muestreo por conglomerados de una etapa) para estimar la proporción para cada celda y margen de la tabla 2×2 y determine la varianza de cada proporción estimada. Ahora utilice los efectos de diseño estimados para realizar una prueba de hipótesis de independencia usando X_F^2. ¿Cuál es la relación de estos resultados con la prueba basada en el modelo?

13 Los siguientes datos son de la encuesta de salud en Canadá, dados en Rao y Thomas (1989, 107). Relacionan las categorías de uso del cigarrillo (fumador, fumador ocasional, nunca ha fumado) con la condición física de 2505 personas. Los fumadores que renunciaron no fueron incluidos en el análisis. Las proporciones estimadas en la siguiente tabla fueron determinadas aplicando los pesos muestrales a la muestra. Los efectos de diseño aparecen entre corchetes. Queremos probar si la categoría de uso del cigarrillo y la condición física son independientes.

		Condición física: Recomendada	Mínima aceptable	Inaceptable	
Uso del cigarrillo:	Fumador	.220 [3.50]	.150 [4.59]	.170 [1.50]	.540 [1.44]
	Fumador ocasional	.023 [3.45]	.010 [1.07]	.011 [1.09]	.044 [2.32]
	Nunca	.203 [3.49]	.099 [2.07]	.114 [1.51]	.416 [2.44]
		.446 [4.69]	.259 [5.96]	.259 [1.71]	1

a ¿Cuál es el valor de X^2 si usted supone que las 2505 personas se reunieron en una muestra multinomial? ¿Y el de G^2? ¿Cuál es el valor p para cada estadística bajo el muestreo multinomial? ¿Por qué son incorrectos estos valores p?

b Use (10.9) para determinar el valor esperado aproximado de X^2 y G^2.

c Calcule las estadísticas corregidas X_F^2 y G_F^2 para estos datos y determine los valores p para las pruebas de hipótesis. ¿Afectan los conglomerados de la encuesta al valor p obtenido?

14 Los siguientes datos son de Rao y Thomas (1988) y se reunieron en la encuesta sobre estructura de clases en Canadá, una muestra estratificada de varias etapas realizada en 1982-1983 para estudiar el nivel de empleo y la estructura social. Canadá fue dividido en 35 estratos, por región y tamaño de población; se extrajeron dos unidades primarias en 34 de estos estratos, y una unidad primaria en el estrato 35. Las varianzas se estimaron mediante réplicas repetidas balanceadas (RRB) usando los 34 estratos con dos unidades primarias. Los efectos de diseño estimados aparecen en corchetes, después de la proporción estimada para cada celda.

	Hombres	Mujeres	
Gerentes con poder de decisión	0.103 [1.20]	0.038 [1.31]	0.141 [1.09]
Gerentes asesores	0.018 [0.74]	0.016 [1.95]	0.034 [1.95]
Supervisores	0.075 [1.81]	0.043 [0.92]	0.118 [1.30]
Trabajadores semiautónomos	0.105 [0.71]	0.085 [1.85]	0.190 [1.44]
Trabajadores	0.239 [1.42]	0.278 [1.15]	0.516 [1.86]
	0.540 [1.29]	0.460 [1.29]	

a ¿Cuál es el valor de X^2 si supone que las 1463 personas surgen de una muestra aleatoria simple? ¿Y el de G^2? ¿Cuál es el valor p de cada estadística bajo un muestreo multinomial? ¿Por qué son incorrectos estos valores p?

b Use (10.9) para determinar el valor esperado aproximado de X^2 y G^2.

c ¿Cuántos grados de libertad están asociados con las estimaciones RRB de la varianza?

d Calcule las estadísticas con corrección de primer orden X_F^2 y G_F^2 para estos datos y determine valores p aproximados para las pruebas de hipótesis. ¿Afectan los conglomerados de la encuesta al valor p obtenido?

e La corrección de segundo orden de Rao-Scott dio la estadística de prueba $X_S^2 = 38.4$, con 3.07 grados de libertad. ¿Cuál es la relación del valor p obtenido usando X_S^2 con el valor p de X_F^2?

11

Regresión con datos de encuestas complejas*

Ahora se había dado cuenta que no sabía nada fundamental y, como un monje agobiado con la conciencia del pecado, se lamentó: "¡Si sólo supiera más!... Sí, ¡y si tan sólo pudiera recordar la estadística!"

—Sinclair Lewis, *It Can't Happen Here*

EJEMPLO 11.1 ¿Cuál es la relación entre el uso de medicamentos y cigarrillo por la madre con el peso al nacer y la mortalidad infantil? ¿Cómo se relaciona el peso al nacer de un bebé con el de los hermanos mayores?

En casi todo este libro hemos enfatizado la estimación de medias y totales de una población; por ejemplo, ¿cuántos bebés con bajo peso nacen en Estados Unidos cada año? Sin embargo, las preguntas acerca de las relaciones entre las variables a menudo se contestan en estadística usando alguna forma del **análisis de regresión**. Una variable de respuesta (por ejemplo, *peso al nacer*) se relaciona con diversas variables explicativas (por ejemplo, *madre que fuma o que no fuma, ingreso familiar* y *edad de la madre*). Quisiéramos utilizar la ecuación de regresión resultante no sólo para identificar la relación entre las variables para nuestros datos, sino también para predecir el valor de la respuesta para futuros bebés o para los bebés no incluidos en la muestra.

Usted sabe cómo ajustar los modelos de regresión si se cumplen las "hipótesis comunes", revisadas en la sección 11.1. Sin embargo, es frecuente que estas hipótesis no se cumplan para datos de encuestas complejas. Por ejemplo, para responder las preguntas anteriores, se podrían usar los datos de la encuesta de salud de madres e hijos (MIHS) en 1988 en Estados Unidos. La encuesta, realizada por la oficina de censos para el Centro Nacional de Estadísticas de Salud, proporciona datos sobre varios factores relacionados con el embarazo y la salud del bebé, incluyendo la ganancia de peso, el uso del cigarrillo y medicamentos durante el embarazo, la exposición materna a los desperdicios tóxicos, y las complicaciones durante el embarazo y el parto (Sanderson *et al* 1991). Pero, como la mayoría de las encuestas a gran escala, la MIHS no es una muestra aleatoria simple. Se extrajeron muestras aleatorias estratificadas de los registros vitales de 1988 de los 48 estados continentales y del Distrito de Columbia de Estados Unidos. Las muestras incluyeron 10,000 de nacimiento de los 3,909,510 nacimientos de 1988, 4000 reportes de muerte fetal a partir de las 15,000 muertes estimadas de fetos con 28 semanas o más de gestación, y 6000 certificados de defunción para bebés con menos de un año de edad a partir de la población de 38,910 de tales muertes. Como los bebés negros tienen mayor incidencia de bajo peso al nacer y mortalidad

infantil que los bebés blancos, tuvieron una mayor fracción de muestreo que los no negros. Los bebés con bajo peso al nacer también se sobremuestrearon. A las madres de los registros de la muestra se les envió un cuestionario acerca de los cuidados prenatales; del consumo de cigarrillo, del alcohol y medicamentos, ingreso familiar, hospitalización y salud del bebé, además de otras variables relacionadas. Después de recibir el permiso de la madre, los investigadores también enviaron cuestionarios a proveedores de cuidados prenatales y de hospitales, preguntando acerca de la salud de la madre y del bebé antes y después del nacimiento. ■

Como vimos para el caso de las tablas de contingencia del capítulo 10, las diferentes probabilidades de selección, los conglomerados y la estratificación de la muestra complican un análisis estadístico. En la MIHS, las probabilidades con selecciones diferentes para bebés en estratos diferentes deben tomarse en cuenta al ajustar los modelos de regresión. Si alguna encuesta implica el uso de conglomerados, como la encuesta nacional a víctimas de crímenes (NCVS), entonces los errores estándar para los coeficientes de regresión calculados bajo la hipótesis de que las observaciones son independientes, serán incorrectos.

En este capítulo analizaremos la forma de hacer la regresión en encuestas de muestras complejas. Revisamos el método tradicional del análisis de regresión, basado en el modelo, como se enseña en los cursos de introducción a la estadística, en la sección 11.1. En la sección 11.2 analizamos un método basado en el diseño para la regresión y da los métodos para calcular los errores estándar de los coeficientes de regresión. La sección 11.3 compara los métodos basado en el diseño y basado en el modelo. La sección 11.4 presenta un método basado en el modelo y la sección 11.5 aplica estas ideas a la regresión logística.

En el capítulo 3 ya usamos la estimación por regresión; aunque en ese capítulo enfatizamos el uso de la información en una variable auxiliar para aumentar la precisión de la estimación del total de la población, $t_y = \sum_{i=1}^{N} y_i$. En las secciones 11.1 a 11.5 nuestro interés principal será explorar la relación entre distintas variables, y estimar los coeficientes de regresión. En la sección 11.6 volvemos al uso de la regresión para mejorar la precisión de los totales estimados.

11.1 Regresión basada en el modelo para muestras aleatorias simples

Como se expone generalmente en áreas de la estadística distintas del muestreo, la inferencia por regresión se basa en un modelo que se supone describe la relación entre la variable explicativa x y la variable de respuesta y. El modelo lineal que generalmente se utiliza para una sola variable explicativa es

$$Y_i/x_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{11.1}$$

donde Y_i es una variable aleatoria para la respuesta, x_i es una variable explicativa y β_0 y β_1 son parámetros desconocidos. Las Y_i son variables aleatorias; los datos reunidos en la muestra son una realización de esas variables aleatorias, y_i, $i \in \mathcal{S}$. Suponemos que las ε_i, las desviaciones de la variable de respuesta con respecto de la recta descrita por el modelo, satisfacen las condiciones (A1) a (A3):

(A1) $E[\varepsilon_i] = 0$ para toda i. En otras palabras, $E[Y_i/x_i] = \beta_0 + \beta_1 x_i$.

(A2) $V[\varepsilon_i] = \sigma^2$ para toda i. La varianza respecto de la línea de regresión es la misma para todos los valores de x.

(A3) $\text{Cov}[\varepsilon_i, \varepsilon_j] = 0$ para $i \neq j$. Las observaciones no están correlacionadas.

Frecuentemente, también se supone (A4). Implica (A1) a (A3) y agrega la hipótesis adicional de que las ε_i se distribuyen normalmente.

(A4) Condicionalmente sobre las x_i, las ε_i son independientes e idénticamente distribuidas a partir de una distribución normal con media 0 y varianza σ^2.

Las **estimaciones ordinarias por mínimos cuadrados** (OLS) de los parámetros son los valores de β_0 y β_1 que minimizan la suma residual de cuadrados $\sum[y_i - (\beta_0 + \beta_1 x_i)]^2$. Obtenemos los estimadores de la pendiente β_1 y ordenada al origen β_0 resolviendo las **ecuaciones normales**: Para el modelo en (11.1), éstas son

$$\begin{aligned} \beta_0 n \quad &+ \beta_1 \sum x_i = \sum y_i \\ \beta_0 \sum x_i \; &+ \beta_1 \sum x_i^2 = \sum x_i y_i. \end{aligned}$$

Al resolver las ecuaciones normales obtenemos las estimaciones de los parámetros

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum x_i y_i - \frac{\left(\sum x_i \sum y_i\right)}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} \\ \hat{\beta}_0 &= \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n}. \end{aligned} \tag{11.2}$$

$\hat{\beta}_1$ y $\hat{\beta}_0$ son estimaciones lineales en y, ya que cada una se puede escribir en la forma $\sum a_i y_i$ para a_i constantes conocidas. Aunque por lo general no se enseña de esta forma, es equivalente a (11.2) lo siguiente:

$$\hat{\beta}_1 = \sum_i \left[\frac{x_i - \left(\sum x_j\right)/n}{\sum x_j^2 - \left(\sum x_j\right)^2/n} \right] y_i$$

y

$$\hat{\beta}_0 = \sum_i \frac{1}{n} \left[1 - \frac{x_i \sum x_j - \left(\sum x_j\right)^2/n}{\sum x_j^2 - \left(\sum x_j\right)^2/n} \right] y_i.$$

Si se cumplen las hipótesis (A1) a (A3), entonces $\hat{\beta}_0$ y $\hat{\beta}_1$ son las **mejores estimaciones lineales insesgadas**; es decir, entre todas las estimaciones insesgadas bajo el modelo (11.1), $\hat{\beta}_0$ y $\hat{\beta}_1$ tienen la menor varianza. Si se cumple la hipótesis (A4), podemos utilizar la distribución t para construir intervalos de confianza y pruebas de hipótesis para la pendiente y la

ordenada al origen de la recta de regresión "verdadera". Bajo la hipótesis (A4),

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{V}_M(\hat{\beta}_1)}}$$

sigue una distribución t con $n-2$ grados de libertad. El subíndice M se refiere al uso del modelo para estimar la varianza; para el modelo (11.1), un estimador de la varianza, insesgado con respecto del modelo, es

$$\hat{V}_M(\hat{\beta}_1) = \frac{\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2/(n-2)}{\sum(x_i - \bar{x})^2}. \quad (11.3)$$

El coeficiente de determinación R^2 en la regresión lineal es

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Éstos son los resultados obtenidos a partir de cualquier buen paquete de software para estadística.

EJEMPLO 11.2 Para ilustrar la regresión en el marco recién analizado, usamos los datos de Macdonell (1901), con la longitud (en cm) del dedo medio de la mano izquierda y la altura (en pulgadas) de 3000 criminales. Al final del siglo XIX, se creía que las tendencias criminales podrían expresarse en características físicas distinguibles de las características físicas de las clases no criminales. Macdonell comparó las medias y correlaciones de mediciones antropométricas de los criminales con las de personas de Cambridge (que supuestamente provienen de otra clase de la sociedad). Éste es un conjunto de datos importante en la historia de la estadística; es el que usó Student (1908) para demostrar la distribución t. El conjunto completo de datos para los 3000 criminales está en el archivo anthrop.dat.

Se extrajo una muestra aleatoria simple de 200 individuos (archivo anthsrs.dat) de las 3000 observaciones. Al ajustar un modelo lineal con S-PLUS, llamando y a la altura y x a la longitud del dedo medio izquierdo, obtenemos la siguiente salida:

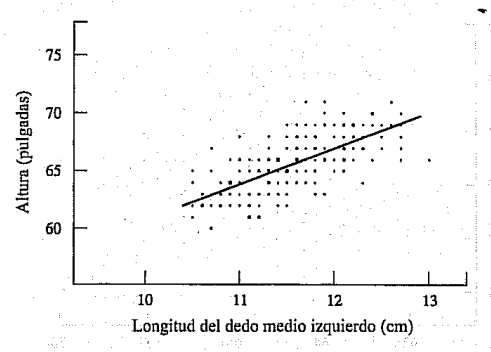
	Valor	Error estándar	Valor t	Pr(>|t|)
Ordenada al origen	30.3162	2.5668	11.8109	0.0000
x	3.0453	0.2217	13.7348	0.0000

Graficamos los datos de la muestra junto con la recta de regresión OLS de la figura 11.1. El modelo parece ser un buen ajuste a los datos ($R^2 = 0.49$) y usando un análisis basado en el modelo, un intervalo de confianza de 95% para la pendiente de la recta es

$$3.0453 \pm 1.972(0.2217) = [2.61,\ 3.48].$$

Si generamos una y otra vez muestras de tamaño 200 a partir del modelo en (11.1) y construimos un intervalo de confianza para la pendiente de cada muestra, esperaríamos que 95% de los intervalos de confianza resultantes incluyan el verdadero valor de β_1. ■

FIGURA 11.1
Una gráfica de altura contra longitud del dedo para una muestra aleatoria simple de 200 observaciones. El área de cada círculo es proporcional al número de observaciones en ese valor de (x, y). La recta de regresión OLS, aquí trazada, tiene la ecuación $y = 30.32 + 3.05x$.



He aquí algunas observaciones importantes para la aplicación de la regresión a los datos de una encuesta:

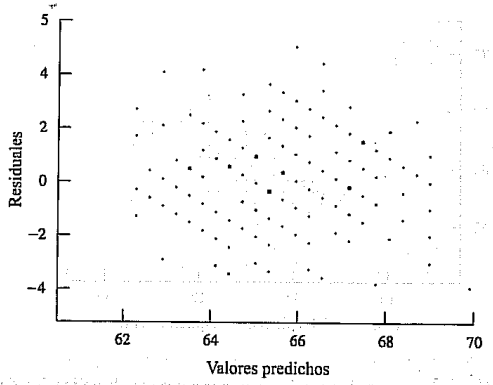
1 No se necesitan suposiciones adicionales para calcular las estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ a partir de los datos; sólo son fórmulas. Necesitamos suponer (A1) a (A4) para hacer *inferencias* en torno de los parámetros "reales" pero desconocidos β_0 y β_1, y acerca de los valores predichos de la variable de respuesta. Así, sólo utilizamos las hipótesis cuando construimos un intervalo de confianza para β_1 o para un valor predicho, o cuando queremos decir, por ejemplo, que $\hat{\beta}_1$ es la mejor estimación lineal insesgada de β_1.

Esto también es cierto para otras estadísticas que calculemos. Si extraemos una muestra de conveniencia de 100 personas, siempre podemos calcular el promedio de los ingresos de esas personas, pero no podemos evaluar la precisión de esa estadística, a menos que establezcamos hipótesis para el modelo con respecto de la población y de la muestra. Sin embargo, con una muestra de probabilidad, podemos usar el propio diseño de la muestra para hacer inferencias y no tenemos que establecer hipótesis para el modelo.

2 Si las suposiciones no se satisfacen al menos aproximadamente, es muy probable que las inferencias basadas en el modelo acerca de los parámetros y los valores predichos sean incorrectas. Por ejemplo, si las observaciones tienen una correlación positiva en vez de ser independientes, es probable que la estimación de la varianza a partir de (11.3) sea menor que lo debido. En consecuencia, es probable que se piense que los coeficientes de regresión son estadísticamente significativos con más frecuencia de lo debido, como se demuestra en Kish y Frankel (1974).

3 Podemos verificar parcialmente las hipótesis del modelo, graficando los residuales y usando varias estadísticas de diagnóstico, como se describe en los libros de regresión en la bibliografía. Una gráfica utilizada en forma común es la de residuales contra los valores predichos, usada para verificar (A1) y (A2). Para los datos del ejemplo 11.2, esta gráfica aparece en la figura 11.2 y no muestra indicios de que los datos en la muestra violen las hipótesis (A1) o (A2). (Esto no significa que las hipótesis sean ciertas, sólo que no vemos nada en la gráfica que indique que no son válidas. Algunas de las hipótesis, en particular la

FIGURA 11.2
Gráfica de residuales para un análisis basado en el modelo de datos de altura de criminales, usando la muestra aleatoria simple graficada en la figura 11.1. No se ven patrones aparentes, distintos de las rectas diagonales causadas por la variable de respuesta con valores enteros.



independencia, en la práctica son difíciles de verificar.) Sin embargo, no tenemos forma de saber si las observaciones que no están en la muestra se ajustan a este modelo, a menos que realmente las veamos.

4 La regresión no se limita a las variables relacionadas mediante una línea recta. Sea y el peso al nacer y sea x igual a 1 si la madre es negra y 0 si la madre no es negra. En este caso, la pendiente de regresión estima la diferencia en el peso medio al nacer de madres negras y no negras, y la estadística de prueba para H_0: $\beta_1 = 0$ es la estadística de prueba t combinada para la hipótesis nula de que el peso medio al nacer de madres negras es igual al peso medio al nacer de madres no negras. Así, la comparación de medias para las subpoblaciones se puede considerar como un caso particular de análisis de regresión.

11.2 Regresión en encuestas complejas

Muchos de los investigadores que realizan análisis de regresión en datos de encuestas complejas sólo pasan los datos a través de un programa estándar de análisis estadístico como SAS o SPSS y reportan el modelo y los errores estándar indicados por el software. Se podría debatir acerca de asumir un punto de vista basado en el modelo o en el diseño (lo que haremos en la sección 11.3), pero la estructura de datos debe tomarse en cuenta en ambos puntos de vista.

¿Qué puede ocurrir en las encuestas complejas?

1 Las observaciones pueden tener distintas probabilidades de selección π_i. Si la probabilidad de selección está relacionada con la variable de respuesta y_i, entonces un análisis que no tome en cuenta las diferentes probabilidades de selección puede conducir a sesgos en los

parámetros de regresión estimados. Este problema se analiza con detalle en Nathan y Smith (1989), quienes proporcionan una bibliografía sobre el tema.

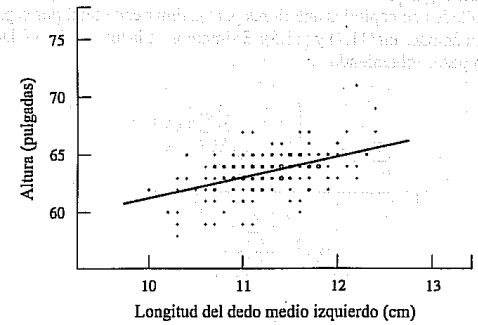
Por ejemplo, suponga que extraemos una muestra con probabilidades diferentes de 200 hombres de la población descrita en el ejemplo 11.2 y que las probabilidades de selección son mayores para los hombres más bajos. (Con fines de ilustración, utilicé las y_i para establecer las probabilidades de selección, con π_i proporcional a 24 para $y < 65$, 12 para $y = 65$, 2 para $y = 66$ o 67 y 1 para $y > 67$, con los datos en el archivo anthuneq.dat.) La figura 11.3 muestra una gráfica de dispersión de los datos a partir de esta muestra, junto con la recta de regresión OLS descrita en la sección 11.1. La ecuación de regresión OLS es $y = 43.41 + 1.79x$, comparada con la ecuación $y = 30.32 + 3.05x$ para la muestra aleatoria simple del ejemplo 11.2. Ignorar las probabilidades de selección en este ejemplo conduce a una estimación muy diferente de la recta de regresión.

Las personas que no responden, que pueden suponerse que tienen probabilidad nula de selección, pueden distorsionar la relación por una razón muy similar. Si las personas que no responden en la MIHS tienen mayor probabilidad de tener hijos con peso bajo al nacer, entonces un modelo de regresión que prediga el peso al nacer a partir de variables explicativas podría no ajustarse a las personas que no responden. La ausencia de respuesta por elemento puede tener efectos similares.

La estratificación de la MIHS también debe tomarse en cuenta. La encuesta se estratificó debido a que los investigadores querían garantizar un tamaño de muestra adecuado para los bebés negros y con bajo peso al nacer. Es plausible que cada estrato tenga su propia línea de regresión y la postulación de una única línea recta que se ajuste a todos los datos puede ocultar alguna información.

2 Aunque los estimadores de los parámetros de regresión sean aproximadamente insesgados con respecto del diseño, los errores estándar dados por SAS o SPSS tendrán una alta probabilidad de ser incorrectos si el diseño de la encuesta utiliza conglomerados. Por lo general, al usar los conglomerados, el efecto de diseño para los coeficientes de regresión será mayor que 1.

FIGURA 11.3
Una gráfica de y contra x para una muestra de 200 criminales con probabilidades diferentes. El área de cada círculo es proporcional al número de observaciones en ese punto dato. La recta OLS es $y = 43.41 + 1.79x$. La pendiente menor de esta recta, comparada con la pendiente 3.05 de la muestra aleatoria simple en la figura 11.1, refleja el submuestreo de los hombres altos.



11.2.1 Estimación puntual

Tradicionalmente, la teoría de muestreo basada en el diseño se ha preocupado por estimar cantidades a partir de una población finita, cantidades como $t_y = \sum_{i=1}^{N} y_i$ o $\bar{y}_U = t_y/N$. Así, en ese espíritu descriptivo, las cantidades de interés para la regresión en una población finita son los coeficientes de mínimos cuadrados para la población, B_0 y B_1, que minimicen

$$\sum_{i=1}^{N}(y_i - B_0 - B_1 x_i)^2$$

sobre toda la población finita. Sería bueno que la ecuación $y = B_0 + B_1 x$ resumiese información útil acerca de la población (en caso contrario, ¿para qué nos interesarían B_0 y B_1?), pero no necesitamos más hipótesis para decir que éstas podrían ser las cantidades de interés. Como en la sección 11.1, las ecuaciones normales son

$$B_0 N + B_1 \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i$$

$$B_0 \sum_{i=1}^{N} x_i + B_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i,$$

y B_0 y B_1 se pueden expresar como funciones de los totales de la población:

$$B_1 = \frac{\sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)\Big/N}{\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\Big/N} = \frac{t_{xy} - \frac{t_x t_y}{N}}{t_{x^2} - \frac{(t_x)^2}{N}} \qquad (11.4)$$

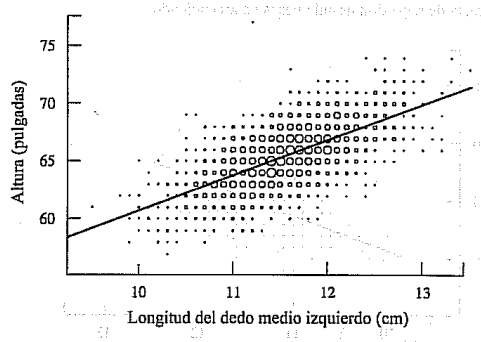
$$B_0 = \frac{\sum_{i=1}^{N} y_i - B_1 \sum_{i=1}^{N} x_i}{N} = \frac{t_y - B_1 t_x}{N}. \qquad (11.5)$$

Conocemos los valores para toda la población, para la muestra del ejemplo 11.2. Estos valores de la población aparecen graficados en la figura 11.4, junto con la recta de mínimos cuadrados de la población $y = 30.179 + 3.056x$.

Como B_0 y B_1 son funciones de los totales de la población, podemos utilizar los métodos desarrollados en capítulos anteriores para estimar cada total por separado y luego sustituir las estimaciones en (11.4) y (11.5). Estimamos cada total de la población en (11.4) y (11.5) usando pesos, obteniendo

$$\hat{B}_1 = \frac{\sum_{i\in S} \omega_i x_i y_i - \frac{\left(\sum_{i\in S} \omega_i x_i\right)\left(\sum_{i\in S} \omega_i y_i\right)}{\sum_{i\in S} \omega_i}}{\sum_{i\in S} \omega_i x_i^2 - \frac{\left(\sum_{i\in S} \omega_i x_i\right)^2}{\sum_{i\in S} \omega_i}} \qquad (11.6)$$

FIGURA 11.4
Gráfica de la población de 3000 criminales. El área de cada círculo es proporcional al número de observaciones de la población en esas coordenadas. La recta de regresión OLS de la población es $y = 30.18 + 3.06x$.



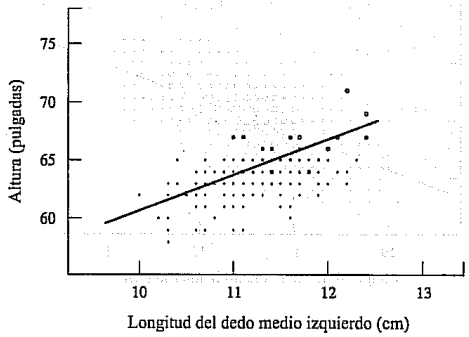
$$\hat{B}_0 = \frac{\sum_{i\in S} w_i y_i - \hat{B}_1 \sum_{i\in S} w_i x_i}{\sum_{i\in S} w_i} \qquad (11.7)$$

Nota sobre los cálculos Aunque (11.6) y (11.7) son expresiones correctas para los estimadores, están sujetas a un error de redondeo y no son tan buenas para los cálculos como otros algoritmos ya desarrollados. En la práctica, se debe utilizar un software profesional diseñado para estimar los parámetros de regresión en encuestas complejas. Si no tiene acceso a tal software, use cualquier paquete de regresión estadística que calcule estimaciones con mínimos cuadrados ponderados. Si utiliza los pesos w_i en la estimación con mínimos cuadrados ponderados, usted obtendrá las mismas estimaciones puntuales de (11.6) y (11.7); sin embargo, en las encuestas complejas, los errores estándar y las pruebas de hipótesis proporcionadas por el software serán incorrectos y deben ignorarse.

Graficación de los datos En cualquier análisis de regresión se *deben* graficar los datos. La graficación de datos multivariados es un reto incluso para datos de una muestra aleatoria simple (Cook y Weisberg 1994 analizan detalladamente las gráficas de regresión.) Los datos obtenidos a partir del diseño de una encuesta compleja (con estratificación, pesos diferentes y conglomerados) tiene más características que incorporar a las gráficas. En la figura 11.5 indicamos los pesos mediante el área de los círculos. Sin embargo, la muestra con probabilidades diferentes usada en la página 353 y en el ejemplo 11.3 no tiene conglomerados ni estratificación. Si una encuesta tiene relativamente pocos conglomerados por estrato, usted puede graficar los datos por separado para cada uno o indicar la pertenencia a un conglomerado mediante un color. La graficación para los datos de encuestas es un área de investigación actual. Korn y Graubard (1998) desarrollaron de manera independiente algunas de las gráficas aquí mostradas y analizan otras gráficas posibles.

EJEMPLO 11.3 Estimemos las cantidades para poblaciones finitas B_0 y B_1 para la muestra con probabilidades diferentes graficada en la figura 11.3. Las estimaciones puntuales, utilizando los pesos, son $\hat{B}_0 = 30.19$ y $\hat{B}_1 = 3.05$. Si ignoramos los pesos y simplemente pasamos los datos ob-

FIGURA 11.5
Una gráfica de datos a partir de una muestra con probabilidades diferentes. El área de cada círculo es proporcional a la suma de los pesos para las observaciones con esos valores de x y y. Observe que los hombres más altos en la muestra también tienen pesos mayores, de modo que la pendiente de la recta de regresión usando pesos va ascendiendo.



servados a través de un programa estándar de regresión como el procedimiento REG de SAS, obtenemos estimaciones muy diferentes: $\hat{B}_0 = 43.41$ y $\hat{B}_1 = 1.79$.

La figura 11.5 muestra por qué los pesos, que estaban relacionados con y, establecen una diferencia en este caso. Los hombres más altos tienen menores probabilidades de selección por lo que, no tantos de ellos aparecían en la muestra con probabilidades diferentes. Sin embargo, los hombres altos seleccionados tenían pesos de muestreo mayores; un hombre de 1.75m en la muestra representó 24 veces más unidades de población que un hombre de 1.52m en la muestra. Al incorporar los pesos, calculamos las estimaciones de los parámetros como si fuesen en realidad w_i puntos dato con valores (x_i, y_i). ■

11.2.2 Errores estándar

Ahora analizaremos el efecto del diseño de muestreo complejo sobre los errores estándar. Como $\hat{B}_0$ y $\hat{B}_1$ son funciones de los totales estimados de la población, podemos usar los métodos del capítulo 9 para calcular las estimaciones de la varianza.

Para cualquier método de estimación de la varianza, bajo ciertas condiciones de regularidad, un intervalo de confianza aproximado del $100(1-\alpha)\%$ para B_1 está dado por

$$\hat{B}_1 \pm t_{\alpha/2}\sqrt{\hat{V}(\hat{B}_1)},$$

donde $t_{\alpha/2}$ es el punto superior $\alpha/2$ de una distribución t con los grados de libertad asociados a la estimación de la varianza. Para la linealización, la navaja (jackknife) o la réplica repetida balanceada (RRB) en una muestra estratificada de varias etapas, usaríamos (número de unidades primarias en la muestra) − (número de estratos) como los grados de libertad. Para el método de estimación de la varianza mediante grupos aleatorios, los grados de libertad adecuados serían (cantidad de grupos) − 1.

11.2.2.1 Errores estándar usando la linealización

Podemos emplear el estimador de varianza por linealización para la pendiente, pues B_1 es una función de cuatro totales de la población t_{xy}, t_x, t_y, y t_{x^2}. Así, usando la linealización, como se mostró en el ejercicio 3 del capítulo 9,

$$V_L(\hat{B}_1) \approx V\left[\frac{\partial B_1}{\partial t_{xy}}(\hat{t}_{xy}-t_{xy}) + \frac{\partial B_1}{\partial t_x}(\hat{t}_x - t_x) + \frac{\partial B_1}{\partial t_y}(\hat{t}_y - t_y) + \frac{\partial B_1}{\partial t_{x^2}}(\hat{t}_{x^2} - t_{x^2})\right]$$

$$= V\left[\left(t_{x^2} - \frac{(t_x)^2}{N}\right)^{-1}\sum_{i\in S}\omega_i(y_i - B_0 - B_1 x_i)\left(x_i - \frac{t_x}{N}\right)\right].$$

Defina

$$q_i = (y_i - \hat{B}_0 - \hat{B}_1 x_i)(x_i - \hat{\bar{x}}),$$

donde $\hat{\bar{x}} = \hat{t}_x / \hat{N}$. Entonces podemos utilizar

$$\hat{V}_L(\hat{B}_1) = \frac{\hat{V}\left(\sum_{i\in S} w_i q_i\right)}{\left[\sum_{i\in S} w_i x_i^2 - \frac{\left(\sum_{i\in S} w_i x_i\right)^2}{\sum_{i\in S} w_i}\right]^2} \qquad (11.8)$$

para estimar la varianza de $\hat{B}_1$.

Observe que el estimador de la varianza basado en el diseño en (11.8) difiere del estimador de la varianza basado en el modelo en (11.3), aunque se extraiga una muestra aleatoria simple. En una muestra aleatoria simple de tamaño n, si ignoramos la corrección para poblaciones finitas,

$$\hat{V}\left(\sum_{i\in S} w_i q_i\right) = \hat{V}(\hat{t}_q) = \frac{N^2 s_q^2}{n},$$

con

$$s_q^2 = \frac{\sum_{i\in S}(x_i - \bar{x}_S)^2(y_i - \hat{B}_0 - \hat{B}_1 x_i)^2}{n-1}.$$

Así, si ignoramos la corrección para poblaciones finitas, (11.8) implica

$$\hat{V}_L(\hat{B}_1) = \frac{n\sum_{i\in S}(x_i - \bar{x}_S)^2(y_i - \hat{B}_0 - \hat{B}_1 x_i)^2}{(n-1)\left[\sum_{i\in S}(x_i - \bar{x}_S)^2\right]^2}.$$

Sin embargo, de (11.3),

$$\hat{V}_M(\hat{\beta}_1) = \frac{\sum_{i\in S}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{(n-2)\sum_{i\in S}(x_i - \bar{x})^2}.$$

¿Por qué la diferencia? El estimador de la varianza basado en el diseño, $\hat{V}_L$, proviene de las probabilidades de selección del diseño, mientras que $\hat{V}_M$ proviene de la desviación cuadrática promedio sobre todas las realizaciones posibles del modelo. Los intervalos de confianza construidos a partir de las dos estimaciones de la varianza tienen diferentes interpretaciones. Con el intervalo de confianza basado en el diseño,

$$\hat{B}_1 \pm t_{\alpha/2}\sqrt{\hat{V}_L(\hat{B}_1)},$$

el nivel de confianza es $\sum u(\mathcal{S})P(\mathcal{S})$, donde la suma se realiza sobre todas las muestras posibles $\mathcal{S}$ que se pueden elegir mediante el diseño de muestreo, $P(\mathcal{S})$ es la probabilidad de seleccionar la muestra $\mathcal{S}$ y $u(\mathcal{S}) = 1$ si el intervalo de confianza construido a partir de la muestra $\mathcal{S}$ contiene la característica de la población B_1 y $u(\mathcal{S}) = 0$ en caso contrario. En una muestra aleatoria simple, el nivel de confianza basado en el diseño es la proporción de muestras posibles que producen un intervalo de confianza que contiene a B_1, a partir del conjunto de todas las muestras aleatorias simples de tamaño n de la población finita de valores fijos $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$.

Para el intervalo de confianza basado en el modelo,

$$\hat{\beta}_1 \pm t_{\alpha/2}\sqrt{\hat{V}_M(\hat{\beta}_1)}$$

el nivel de confianza es la proporción esperada de intervalos de confianza que incluirán a β_1, a partir del conjunto de todas las muestras que se podrían generar a partir del modelo en (A1) a (A3). Así, el estimador basado en el modelo supone que (A1) a (A3) son válidas para el mecanismo de población infinita que genera los datos. El diseño del tipo de una muestra aleatoria simple hace que la hipótesis (A3) (observaciones no correlacionadas) sea razonable. Si un modelo lineal describe la relación entre x y y, entonces (A1) también es plausible. Sin embargo, cualquier violación de la hipótesis (A2) (varianzas iguales) puede tener un gran efecto sobre las inferencias. El estimador de la varianza por linealización es más robusto con respecto de la hipótesis (A2), como veremos en el ejercicio 16.

EJEMPLO 11.4 Para la muestra aleatoria simple del ejemplo 11.2, las estimaciones de la varianza con base en el modelo y en el diseño son bastante similares, ya que parece que las hipótesis del modelo son válidas para la muestra y la población. Para estos datos, $\hat{B}_1 = \hat{\beta}_1$, pues $w_i = 3000/200$ para toda i; $\hat{V}_L(\hat{B}_1) = 0.048$; y $\hat{V}_M(\hat{\beta}1) = (0.2217)^2 = 0.049$. Sin embargo, en otras situaciones, las estimaciones de la varianza pueden ser algo distintas; por lo general, si hay alguna diferencia, la estimación de la varianza por linealización es mayor que la estimación de la varianza con base en el modelo.

Para la muestra de 200 criminales con probabilidades diferentes definimos la nueva variable

$$q_i = (y_i - \hat{B}_0 - \hat{B}_1 x_i)(x_i - \hat{\bar{x}}) = (y_i - 30.1859 - 3.0541x_i)(x_i - 11.51359).$$

(Observe que $\hat{\bar{x}}$ es la estimación de $\bar{x}_U$ calculada mediante las probabilidades diferentes; el promedio muestral de las 200 x_i en la muestra es 11.2475, que es un poco menor.) Entonces $V(\sum_{i\in S} w_i q_i) = 238{,}161$, y

$$\left[\sum_{i\in S} w_i x_1^2 - \frac{\left(\sum_{i\in S} w_i x_i\right)^2}{\sum_{i\in S} w_i}\right]^2 = 688{,}508,$$

de modo que $\hat{V}_L(\hat{B}_1) = 0.346$. Si ignoramos los pesos, entonces el análisis OLS da $\hat{\beta}_1 = 1.79$ y $\hat{V}_M(\hat{\beta}_1) = 0.05121169$. La varianza estimada es mucho menor usando el modelo, pero $\hat{\beta}_1$ es insesgada como estimador de B_1. ■

11.2.2.2 Errores estándar usando la navaja (jackknife)

Suponga que tenemos una muestra estratificada de varias etapas, con pesos w_i y H estratos. Se extrae una muestra de n_h unidades primarias en el estrato h. Recordemos (vea la sección 9.3.2) que para la iteración j en el estrato h para el método de la navaja, omitimos todas las unidades de observación en la unidad primaria j y volvemos a calcular la estimación mediante las unidades restantes. Definimos

$$w_{i(hj)} \begin{cases} w_i & \text{si la unidad de observación } i \text{ no está en el estrato } h. \\ 0 & \text{si la unidad de observación } i \text{ está en la unidad primaria } j \text{ del estrato } h. \\ \dfrac{n_h}{n_h - 1} w_i & \text{si la unidad de observación } i \text{ está en el estrato } h \text{ pero no en la unidad primaria } j. \end{cases}$$

Entonces el estimador de la varianza con reemplazo de $\hat{B}_1$ según el método de la navaja es

$$\hat{V}_{\text{JK}}(\hat{B}_1) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (\hat{B}_{1(hj)} - \hat{B}_1)^2, \tag{11.9}$$

donde $\hat{B}_1$ se define en (11.6) y $\hat{B}_{1(hj)}$ tiene la misma forma, pero reemplazando $w_{i(hj)}$ en cada aparición de w_i en (11.6).

EJEMPLO 11.5 Para nuestras dos muestras de tamaño 200 de los 3000 criminales,

$$\hat{V}_{\text{JK}}(\hat{B}_1) = \frac{199}{200} \sum_{j=1}^{200} (\hat{B}_{1(J)} - \hat{B}_1)^2,$$

donde $\hat{B}_{1(J)}$ es la pendiente estimada al eliminar la observación j y cambiar el peso de las demás observaciones en consecuencia. La diferencia entre la muestra aleatoria simple y la muestra con probabilidades diferentes está en los pesos. Para la muestra aleatoria simple, los pesos originales son $w_i = 3000/200$, por lo que $w_{i(j)} = 200w_i/199 = 300/199$ para $i \neq j$. Así, para la muestra aleatoria simple, $\hat{B}_{1(J)}$ es la estimación OLS de la pendiente al omitir la observación j. Para la muestra aleatoria simple, obtenemos $\hat{V}_{\text{JK}}(\hat{B}_1) = 0.050$.

Para la muestra con probabilidades diferentes, los pesos originales son $w_i = 1/\pi_i$ y $w_{i(j)} = 200w_i/199$ para $i \neq j$. Usamos los nuevos pesos $w_{i(j)}$ para calcular $\hat{B}_{1(J)}$ para cada iteración del método de la navaja, obteniendo $\hat{V}_{\text{JK}}(\hat{B}_1) = 0.461$. La varianza estimada mediante el método de la navaja es mayor que la varianza por linealización, como frecuentemente ocurre en la práctica. ■

11.2.3 Regresión múltiple usando matrices

Ahora veremos algunos resultados para la regresión múltiple en general. Nos basaremos fuertemente en resultados matriciales que aparecen en los textos sobre modelos lineales y regresión enumerados en la bibliografía al final del libro. Si no tiene mucha experiencia en la teoría de regresión, deberá estudiar ese material antes de leer esta sección.

Suponga que queremos determinar una relación entre y_i y un vector de dimensión p de variables explicativas $\mathbf{x}_i$, donde $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T$. Queremos estimar el vector de dimen-

sión p de parámetros de población, $\mathbf{B}$, en el modelo $y = \mathbf{x}^T\mathbf{B}$. Definimos

$$\mathbf{y}_U = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \text{y} \quad X_U = \begin{bmatrix} \mathbf{X}_1^T \\ \mathbf{X}_2^T \\ \vdots \\ \mathbf{X}_N^T \end{bmatrix}.$$

Las ecuaciones normales para la población completa son

$$\mathbf{X}_U^T\mathbf{X}_U\mathbf{B} = \mathbf{X}_U^T y_U,$$

y las cantidades de interés para poblaciones finitas son, suponiendo que $(\mathbf{X}_U^T\mathbf{X}_U)^{-1}$ existe,

$$\mathbf{B} = (\mathbf{X}_U^T\mathbf{X}_U)^{-1}\mathbf{X}_U^T y_U,$$

que son las estimaciones por mínimos cuadrados para la población completa.

$\mathbf{X}_U^T\mathbf{X}_U$ y $\mathbf{X}_U^T y_U$ son matrices de totales de la población. El elemento (j,k) de la matriz $p \times p$ $\mathbf{X}_U^T\mathbf{X}_U$ es $\sum_{i=1}^N x_{ij}x_{ik}$, y el elemento k del p-vector $\mathbf{X}_U^T y_U$ es $\sum_{i=1}^N x_{ik}y_i$.

Así, podemos estimar las matrices $\mathbf{X}_U^T\mathbf{X}_U$ y $\mathbf{X}_U^T y_U$. Sea $\mathbf{X}_s$ la matriz de valores explicativos para la muestra, $\mathbf{y}_s$ el vector de respuesta de las observaciones muestrales, y $\mathbf{W}_s$ una matriz diagonal con los pesos w_i de la muestra. Entonces, el elemento (j, k) de la matriz $p \times p$ $\mathbf{X}_S^T\mathbf{W}_S\mathbf{X}_S$ es $\sum_{i\in S} w_i x_{ij}x_{ik}$, lo que estima el total de la población $\sum_{i=1}^N x_{ij}x_{ik}$; el elemento k del p-vector $\mathbf{X}_S^T\mathbf{W}_S y_S$ es $\sum_{i\in S} w_i x_{ik}y_i$, lo cual estima el total de la población $\sum_{i=1}^N x_{ik}y_i$. Entonces, de manera similar a (11.6) y (11.7), definimos el estimador de $\mathbf{B}$ como

$$\hat{\mathbf{B}} = (\mathbf{X}_S^T\mathbf{W}_S\mathbf{X}_S)^{-1}\mathbf{X}_S^T\mathbf{W}_S y_S. \qquad \textbf{(11.10)}$$

Sea

$$\mathbf{q}_i = \mathbf{x}_i^T(yi - \mathbf{x}_i^T\hat{\mathbf{B}}).$$

Entonces, usando la linealización, como se muestra en Shah et al (1977),

$$\hat{V}(\hat{\mathbf{B}}) = (\mathbf{X}_S^T\mathbf{W}_S\mathbf{X}_S)^{-1}\hat{V}\left(\sum_{i\in S} w_i\mathbf{q}_i\right)(\mathbf{X}_S^T\mathbf{W}_S\mathbf{X}_S)^{-1}. \qquad \textbf{(11.11)}$$

Podemos construir intervalos de confianza para los parámetros individuales:

$$\hat{B}_k \pm t\sqrt{\hat{V}(\hat{B}_k)},$$

donde t es el percentil adecuado de la distribución t. Korn y Graubard (1990) sugieren el uso del método de Bonferroni para la inferencia simultánea con respecto de m parámetros de regresión, construyendo un intervalo de confianza del $100(1 - \alpha/m)$ % para cada uno de los parámetros.

11.2.4 Regresión con pesos contra mínimos cuadrados ponderados

Muchos libros de texto de regresión analizan la estimación por regresión usando mínimos cuadrados ponderados como un remedio para las varianzas distintas. Si el modelo que genera los datos es

$$Y_i = \mathbf{x}_i^T\beta - \varepsilon_i$$

con ε_i independientes y normalmente distribuidos, con media 0 y varianza σ_i^2, entonces ε_i/σ_i sigue una distribución normal con media 0 y varianza 1. La estimación por mínimos cuadrados ponderados (WLS) es

$$\hat{\beta}_{\text{WLS}} = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\mathbf{y}$$

con $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$. La estimación por mínimos cuadrados ponderados minimiza $\sum(y_i - \mathbf{x}_i^T\beta)^2/\sigma_i^2$ y da a las observaciones con menor varianza más peso al determinar la ecuación de regresión. Si el modelo es válido entonces, según la teoría de los mínimos cuadrados ponderados,

$$V(\hat{\beta}) = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}.$$

No estamos usando los mínimos cuadrados ponderados en este sentido, aunque nuestro estimador puntual sea el mismo. Nuestros pesos provienen del diseño de muestreo, no de una estructura supuesta de covarianzas. Nuestra varianza estimada de los coeficientes no es $(\mathbf{X}^T\hat{\Sigma}^{-1}\mathbf{X})^{-1}$, la varianza estimada según la teoría de mínimos cuadrados ponderados, sino que es

$$(\mathbf{X}_S^T\mathbf{W}_S\mathbf{X}_S)^{-1}\hat{V}\left[\sum_{i\in S} w_i\mathbf{x}_i^T(y_i - \mathbf{x}_i^T\hat{\mathbf{B}})\right](\mathbf{X}_S^T\mathbf{W}_S\mathbf{X}_S)^{-1}.$$

Por supuesto, es posible combinar el punto de vista de mínimos cuadrados ponderados de los cursos de regresión con el enfoque de poblaciones finitas, definiendo las cantidades de interés de la población como

$$\mathbf{B} = (\mathbf{X}_U^T\Sigma_U^{-1}\mathbf{X}_U)^{-1}\mathbf{X}_U^T\Sigma_U^{-1}\mathbf{y}_U,$$

lo que generaliza el modelo de regresión. Esto es en esencia lo que se hace en la estimación de cocientes, usando $\Sigma_U = \text{diag}(x_1, x_2, \ldots, x_N)$, como mostraremos en el ejemplo 11.9.

11.2.5 Software para regresión en encuestas complejas

Varios paquetes de software de los analizados en la sección 9.6 calculan los coeficientes de regresión y sus errores estándar para datos de encuestas complejas. SUDAAN y PC CARP utilizan la linealización para calcular las varianzas estimadas de las estimaciones de los parámetros. OSIRIS y WesVarPC utilizan métodos de réplicas para estimar las varianzas.

Antes de usar software escrito por otros para realizar un análisis de regresión sobre los datos de una encuesta, revise cómo enfrenta el problema de datos faltantes. Por ejemplo, si una observación omite uno de los valores x, SUDAAN, al igual que SAS, excluye la observación del análisis. Si su encuesta tiene gran cantidad de ausencias de respuesta por elemento en distintas variables, es posible que termine haciendo su análisis de regresión con sólo 20 de las observaciones de la muestra. Tal vez deba considerar la cantidad de ausencia de respuesta por elemento a la vez que cuestiones científicas al elegir los covariados para su modelo.

Muchas encuestas realizadas por organizaciones gubernamentales no liberan bastante información en las cintas de uso público como para calcular las varianzas estimadas para los coeficientes de regresión. Por ejemplo, el conjunto de datos de uso público de la NCVS de 1990 contiene pesos para cada familia y persona de la muestra, pero no proporciona información sobre los conglomerados. Sin embargo, tales encuestas proporcionan con frecuencia información acerca de los efectos de diseño para estimar los totales de la población. En este caso, estime los parámetros de regresión mediante los pesos proporcionados. Luego, estime la

varianza para los coeficientes de regresión como si se hubiese extraído una muestra aleatoria simple y multiplique cada varianza estimada por un efecto de diseño general para los totales de la población. En general, los efectos de diseño para los coeficientes de regresión tienden a (pero no tienen que) ser menores que los efectos de diseño para estimar las medias y los totales de la población, de modo que al multiplicar las varianzas estimadas de los coeficientes de regresión por el efecto de diseño produce con frecuencia una estimación conservadora de la varianza (vea Skinner 1989). Intuitivamente, podemos explicar esto porque un buen modelo de regresión puede controlar parte de la variabilidad en una variable de respuesta de un conglomerado a otro. Por ejemplo, si parte de la razón por la que las familias del mismo conglomerado tienden a tener experiencias similares como víctimas de crímenes es el nivel promedio de ingresos en el vecindario, entonces sería de esperar que el ajuste del ingreso en la regresión podría tomar en cuenta parte de la variabilidad de un conglomerado a otro. Entonces, los residuales del modelo mostrarían un menor efecto debido a los conglomerados.

11.3 ¿Hay que utilizar los pesos en la regresión?

En la mayoría de las áreas de la estadística, un análisis de regresión tiene por lo general uno de tres propósitos:

1 Describe la relación entre dos o más variables. Puede ser de interés la relación entre el ingreso familiar y el peso de un bebé al nacer o la relación entre el nivel de instrucción y el ingreso con la posibilidad de ser víctima de un delito violento. El interés está simplemente en una estadística de resumen que describe la asociación entre las variables explicativas y de respuesta.

2 Predice el valor de y para una observación futura. Si conocemos los valores de varias variables demográficas y de salud para una mujer embarazada, ¿podemos predecir el peso del bebé al nacer o la probabilidad de que el bebé sobreviva?

3 Nos permite controlar los valores futuros de y cambiando los valores de las variables explicativas. Para esto, quisiéramos que la ecuación de regresión nos diese una relación de causa y efecto entre x y y.

Los datos de una encuesta pueden servir para los dos primeros propósitos, pero generalmente no se pueden utilizar para establecer relaciones causales definitivas entre las variables.[1] Las encuestas de muestra generalmente brindan datos de observación no experimentales. Observamos un subconjunto de posibles variables explicativas, que no necesariamente incluyen las variables que son las causas fundamentales de los cambios en y. En una encuesta de salud que pretende estudiar la relación entre la nutrición, el ejercicio y la incidencia de cáncer, se podía preguntar a los participantes su dieta y hábitos de ejercicio (o bien, que lo observara el investigador) y hacer un seguimiento posterior para ver si han contraído cáncer. Suponga que un análisis de regresión indica una asociación significativa negativa entre el consumo de vitamina E y la incidencia de cáncer, después de ajustar otras variables, como la edad. El análisis sólo establece una asociación, no la causalidad; usted no puede concluir que la incidencia de cáncer disminuirá si comienza a alimentar a las personas con vitamina E. Aunque la vitamina E podría ser la causa de la disminución en la incidencia del cáncer, la causa podría ser también una de las variables no medidas asociadas con el consumo de vitamina E y la incidencia del cáncer. Para concluir que la vitamina E afecta a la incidencia de cáncer, se debe

[1] Muchos estadísticos dirían que los datos de encuestas no se pueden usar para establecer afirmaciones causales de forma alguna. Las unidades experimentales deben asignarse en forma aleatoria a los tratamientos para inferir la causalidad. Sin embargo, algunas encuestas, como el estudio del ejemplo 8.2, incluyen una experimentación, y para estos casos podemos concluir con frecuencia que un cambio en el tratamiento causó un cambio en la respuesta.

realizar un experimento: distribuir aleatoriamente a los participantes en el estudio en grupos con y sin vitamina E, y observar posteriormente la incidencia de cáncer.

Con frecuencia, el propósito de un análisis de regresión difiere del de un análisis para estimar las medias y totales de una población. Al estimar la cantidad total de personas desempleadas mediante una encuesta, estamos interesados en la cantidad de población finita t_y; queremos estimar cuántas personas de la población en agosto de 1994 estaban sin empleo, pero en un análisis de regresión, ¿está interesado en B_0 y B_1, las estadísticas de resumen para la población finita? ¿O está interesado en descubrir una "verdad universal"? ¿Será capaz de afirmar, por ejemplo, que no sólo ha encontrado una asociación positiva entre la cantidad de grasa en una dieta y la presión sanguínea de sístole para la población en estudio, aunque esperaría una asociación similar en otras poblaciones? Al comparar las medias por dominio, Cochran indica: "pocas veces tiene interés científico saber si las medias por dominio de población finita son iguales, debido a que estas medias no serían exactamente iguales en una población finita, excepto por una lejana posibilidad. En vez de esto, verificamos la hipótesis nula en el sentido de que ambos dominios fueron extraídos de poblaciones *infinitas* con la misma media". (1977, 39). La comparación de medias por dominio es un caso particular de regresión lineal (vea el ejercicio 13) y el comentario de Cochran también se aplica a la regresión lineal en general.

Muchos estadísticos de encuestas han debatido sobre la importancia de los pesos de muestreo para la inferencia en regresión; algunos de los artículos implicados en el debate aparecen en la bibliografía de este capítulo. Brewer y Mellor (1973) presentan un entretenido y revelador diálogo entre un estadístico basado en el modelo y otro basado en el diseño, quienes llegan finalmente a un acuerdo; este diálogo es un excelente punto de partida para un estudio posterior. Las referencias proporcionan un análisis más profundo de los aspectos implicados en esta cuestión y que hemos venido presentando; tratamos de resumir los distintos puntos de vista y presentamos las contribuciones de cada uno para un buen análisis de los datos de una encuesta.

Se defienden dos puntos de vista básicos:

1 *Basado en el diseño*. Presentamos la posición basada en el diseño en la sección anterior. Las cantidades de interés son las características $\mathbf{B}$ para la población finita sin importar qué tan bien se ajusta el modelo a la población. Las inferencias se basan en el muestreo repetido en la población finita, y la estructura de probabilidad utilizada para la inferencia es la definida mediante las variables aleatorias que indican la inclusión en la muestra. Puede haber un modelo que genere los datos, aunque no necesariamente sabemos cómo es, de modo que el análisis no se basa en un modelo teórico. Los pesos se necesitan para estimar las medias y totales de una población y, por analogía, deben usarse también en la regresión lineal.

2 *Basado en el modelo*. Un modelo estocástico describe la relación entre y_i y x_i válida para cada observación en la población. Un modelo posible es $Y_i \mid \mathbf{x}_i = \mathbf{x}_i^T\beta + \varepsilon_i$, con los ε_i independientes y distribuidos de manera normal, con varianza constante. Si las observaciones en la población siguen realmente el modelo, entonces el diseño de la muestra no debe tener ningún efecto, mientras las probabilidades de selección dependan sólo de y por medio de las x. El valor $\mathbf{B}$ es simplemente la estimación de β por mínimos cuadrados, si se conocieran los valores de toda la población; como sólo se conoce una muestra, use las estimaciones OLS

$$\hat{\beta}_{OLS} = (\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T\mathbf{y}_S.$$

Busque un modelo que suponga genera la población y luego estime los parámetros de ese modelo.

Särndal *et al* (1992) adoptan un punto de vista *apoyado por el modelo*, para el que se usa un modelo que especifique los parámetros de interés, pero toda la inferencia se basa en el

diseño de la encuesta. Así, se ajusta un modelo particular porque cree que es un candidato plausible para generar la población, aunque utiliza los pesos de muestreo para estimar los parámetros y el diseño de la muestra para estimar las varianzas de la estimación. Al hacer inferencias mediante el diseño de la muestra, en esta sección consideramos al enfoque apoyado por el modelo como parte del punto de vista basado en el diseño.

La distinción entre los dos puntos de vista es importante para el analista de encuestas, pues la mayoría de los paquetes de software se basan en el diseño o en el modelo. El software estadístico general, como SAS, S-PLUS, BMDP o SPSS suponen un punto de vista de la regresión basado en el modelo, como vimos en la sección 11.1. Los paquetes de encuestas como SUDAAN, PC CARP y WesVarPC se basan en la estimación de los parámetros de población finita usando el enfoque de la sección 11.2. Así, es importante conocer el punto de vista que se desea adoptar. Al pasar ciegamente los datos a través del software, sin comprender lo que está estimando, puede conducir a resultados mal interpretados.

La mayoría de los estadísticos concuerdan en que es bueno que un modelo de regresión describa el verdadero estado de la naturaleza. Así, si supiéramos que un modelo describirá cualquier observación posible que implique a x y y, entonces hay que adoptar ese modelo. En las ciencias físicas, muchos modelos como fuerza = masa × aceleración se pueden deducir de manera teórica. Mientras se permanezca lejos de las velocidades cercanas a la de la luz, cualquier observación en la que se midan con precisión la fuerza, la masa y la aceleración debe corresponder con este modelo. El diseño de la cantidad de observaciones que deben extraerse en la muestra deberá representar poca diferencia para determinar las estimaciones puntuales de los coeficientes de regresión, ya que cada observación posible queda descrita por el modelo.[2]

Por desgracia, casi nunca hay modelos deducidos teóricamente y que sean válidos para todas las observaciones para las situaciones de encuestas. Algún economista podría conjeturar una relación entre la cantidad de hijos, el ingreso y el dinero gastado en comida, pero no hay garantías de que ese modelo sea adecuado para cualquier subgrupo de la población. Otras variables pueden estar relacionadas con la cantidad gastada en comida, (como el nivel educativo o la cantidad de tiempo fuera de casa), pero no ser medidas en la encuesta. Además, la verdadera relación entre las variables podría no ser perfectamente lineal. Así, el principal reto de la inferencia basada en el modelo es su especificación.

Entonces, al asumir un punto de vista basado en el modelo, se debe examinar cuidadosamente la hipótesis y hacer todo lo posible por verificar lo adecuado del modelo para sus datos. Esto incluye graficar los datos y los residuales, realizar exámenes de diagnóstico y emplear diseños de muestreo que permitan evaluar métodos alternativos que puedan proporcionar una mejor descripción de la relación entre las variables. (Por supuesto, también debe graficar los datos si asume un enfoque basado en el diseño.) La inferencia acerca de observaciones que no están en la muestra se basa simplemente en la suposición de que el modelo adoptado se les puede aplicar y debe tenerse mucho cuidado en las generalizaciones fuera de los datos de la muestra. Debe suponer que las unidades de la población que no están en la muestra también pueden ser descritas por el modelo, lo que es una suposición muy aventurada.

El punto de vista de la regresión basada en el modelo tiene varios atractivos: se vincula con las teorías sociológicas del investigador, es consistente con otras áreas de la estadística y proporciona un mecanismo para tomar en cuenta la ausencia de respuesta. El enfoque basado en el modelo brinda un marco de referencia para comparar teorías acerca de las relaciones estructurales. Además, las estimaciones basadas en el modelo se pueden emplear con muestras relativamente pequeñas y con muestras no de probabilidad. Aunque las inferencias basadas en el diseño no dependen de las hipótesis del modelo, en la práctica se necesitan tamaños

[2] Sin embargo, el diseño de muestreo puede afectar las varianzas de las estimaciones puntuales.

de muestra grandes para poder construir intervalos de confianza. Los errores estándar de las estimaciones de parámetros basadas en el modelo generalmente son menores que los correspondientes a las estimaciones basadas en el diseño y que incorporan los pesos.

Sin embargo, la mala especificación del modelo y los covariados omitidos son una preocupación importante en un análisis basado en él y la falta de covariados puede no ser evidente en los análisis residuales normales. Además, en un diseño de encuesta compleja, los predictores faltantes necesarios pueden estar relacionados con el diseño y los pesos de la encuesta. Por ejemplo, para nuestra muestra con probabilidades diferentes de la figura 11.3, las probabilidades de selección que utilizamos dependen del valor de y. Usted puede suponer que la altura queda determinada por muchas, muchas variables x_1, x_2,..., pero el conjunto de datos sólo tiene una de estas posibles variables explicativas. Si todas las demás variables se incluyeran en el modelo, entonces las probabilidades diferentes de selección no serían importantes; sin embargo, como esto no ocurre, las probabilidades de selección tienen información útil para estimar la pendiente de regresión.

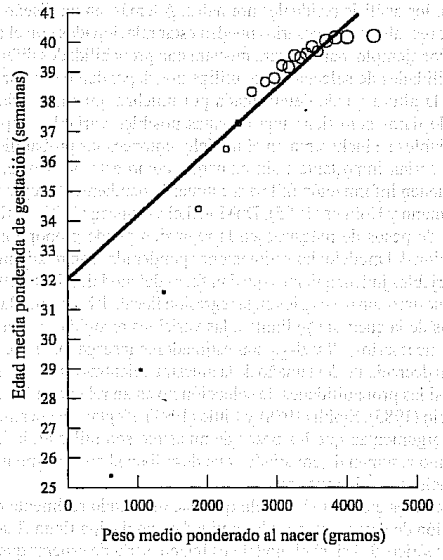
Pfeffermann y Holmes (1985), DuMouchel y Duncan (1983), y Kott (1991) argumentan que el uso de pesos de muestreo en la regresión puede proporcionar robustez a la mala especificación del modelo: las estimaciones ponderadas son relativamente poco afectadas si algunas variables independientes quedan fuera del modelo.[3] Kott (1991) argumenta que los pesos de muestreo son necesarios en la regresión lineal, debido a que la elección de covariados en los datos de la encuesta se limita a las variables recogidas por la encuesta: si se omiten covariados necesarios, $\hat{B}$ y β_{OLS} son estimadores insesgados de β, pero el sesgo de $\hat{B}$ es una función decreciente del tamaño de la muestra, mientras que β_{OLS} sólo es asintóticamente insesgado si las probabilidades de selección no están relacionadas con los covariados faltantes. Rubin (1985), Smith (1988) y Little (1991) adoptan una perspectiva basada en el modelo, pero argumentan que los pesos de muestreo son útiles en la inferencia basada en el modelo como resumen de covariados que describen el mecanismo mediante el que las unidades se incluyen en la muestra.

Un aspecto es claro: si el modelo que esté utilizando realmente describe el mecanismo de generación de datos, entonces la cantidad de población finita **B** debe ser muy cercana al parámetro teórico β. Así, si el modelo es bueno, sería de esperar que la estimación puntual de β mediante este modelo fuese similar a la estimación puntual B calculada usando los pesos de muestreo. Sugerimos el ajuste de un modelo con y sin pesos. Si las estimaciones de los parámetros son distintas, deberá explorar algunas alternativas al modelo adoptado. Una diferencia entre las estimaciones con o sin peso puede indicar que el modelo propuesto no se ajusta bien para una parte de la población. Lohr y Liu (1994) exploran este aspecto para la NCVS.

EJEMPLO 11.6 Korn y Graubard (1995b) ilustran la diferencia que la inclusión de los pesos puede traer a un análisis de regresión, usando los datos del componente "nacido vivo" del MIHS de 1988. Como mencionamos en el ejemplo 11.1, los bebés negros y los bebés con bajo peso al nacer están sobremuestreados, de modo que sus pesos de muestreo son menores que los pesos para los bebés blancos, con un peso normal al nacer. La figura 11.6 muestra una gráfica de los datos y la recta de regresión estimada, al usar los pesos en el cálculo de los parámetros de regresión; la figura 11.7 ignora los pesos. La regresión ponderada jala la recta de regresión hacia donde se estima que está la población; en la regresión no ponderada, la recta proporciona el mejor ajuste por mínimos cuadrados a los datos de la mues-

[3] Sin embargo, esta robustez tiene su precio: como ya mencionamos antes, la varianza basada en el diseño, usando los pesos, es por lo general mayor que la varianza basada en el modelo. Kish (1992) proporciona un buen panorama de la inflación de la varianza debida al uso de estimaciones ponderadas en vez de estimaciones sin pesos.

FIGURA 11.6
Gráfica de la edad media ponderada de gestación contra el peso medio ponderado al nacer para grupos sucesivos de aproximadamente 500 observaciones. Las áreas de burbujas son proporcionales a los tamaños estimados de población de los grupos. La línea recta es el ajuste de regresión lineal ponderado a los datos originales (no agrupados).



FUENTE: De "Examples of Differing Weighted and Unweighted Estimates from a Sample Survey", de E.L. Korn y B.I. Graubard, 1995, *The American Statistician*, vol. 49, páginas 291-295. Copyright © 1995 American Statistical Association. Reproducido con autorización.

tra, pero no describe tan bien la población. Al examinar las gráficas es claro que las líneas de regresión difieren en esa medida debido a que un modelo lineal no es adecuado para los datos; si ajustáramos una regresión cuadrática, entonces los modelos de la regresión ponderada y no ponderada se parecerían más. Así, en este ejemplo, la diferencia entre las estimaciones de parámetros con y sin pesos surge debido a que el modelo lineal adoptado no es adecuado. ■

Cada uno de los puntos de vista de la inferencia acerca de los parámetros de regresión en encuestas complejas puede ser adecuado, dependiendo del uso que quiera darse al modelo de regresión. Tal vez desee considerar las siguientes preguntas al decidir sobre el enfoque a utilizar:

1 ¿Está realizando una regresión para generar estadísticas oficiales que se utilizarán para determinar una política pública? En tal caso, tal vez desee usar los pesos para estimar los parámetros y el diseño para hacer inferencias sobre tales parámetros. Si emplea los pesos para estimar las medias de población y de dominio, tal vez también quiera usarlos para estimar los parámetros de regresión, de modo que los resultados de distintos análisis sean consistentes (vea Alexander 1991). Como observamos antes, **B** debe ser cercano a β para un buen modelo y una población finita de gran tamaño, de modo que una estimación de **B** basada en el diseño también debería estimar β.

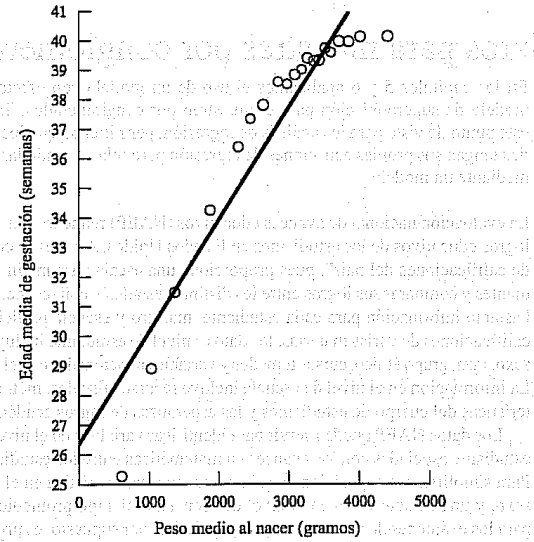
2 ¿Se tomó una muestra de probabilidad? En caso contrario, debe utilizar un enfoque basado en el modelo.

3 ¿De qué tamaño era la muestra? La teoría basada en el diseño se basa en tamaños de muestra grandes para hacer inferencias sobre los parámetros. Si tiene una muestra pequeña, es probable que deba usar un enfoque basado en el modelo.

4 ¿Cuánto se ha estudiado el tema con anterioridad? Si la teoría científica y las investigaciones empíricas anteriores apoyan el modelo propuesto, podría confiar un poco más en su modelo y tener más confianza en un enfoque basado en él.

Sin embargo, los investigadores que escucharon el mensaje de que los pesos de muestreo no son importantes en el análisis de regresión, pero ignoraron el resto de la discusión, cometen a menudo un error: ignoran los pesos y los conglomerados en los datos pasando simplemente los datos de la encuesta por el software de regresión estándar. Esto es incorrecto desde cualquier punto de vista: ya sea que se usen o no los pesos para construir un estimador, la depen-

FIGURA 11.7
Gráfica de la edad media de gestación contra el peso medio al nacer para grupos sucesivos de aproximadamente 500 observaciones. Las áreas de burbujas son proporcionales a los tamaños de muestra de los grupos. La línea recta es el ajuste de regresión lineal no ponderado a los datos originales (no agrupados).



FUENTE: De "Examples of Differing Weighted and Unweighted Estimates from a Sample Survey", de E.L. Korn y B.I. Graubard, 1995, *The American Statistician*, vol. 49, páginas 291-295. Copyright © 1995 American Statistical Association. Reproducido con autorización.

dencia con respecto de los datos reflejada en los conglomerados *debe* tomarse en cuenta al calcular los errores estándar. En la siguiente sección analizamos un enfoque basado en el modelo que incorpora la correlación positiva entre observaciones del mismo conglomerado.

11.4 Modelos mixtos para muestras por conglomerados

En los capítulos 5 y 6 analizamos el uso de un modelo con efectos aleatorios como un modelo de superpoblación para el muestreo por conglomerados. También podemos usar este punto de vista para los análisis de regresión, permitiendo que los diversos conglomerados tengan sus propias ecuaciones de regresión pero relacionando las diferentes ecuaciones mediante un modelo.

EJEMPLO 11.7 La evaluación nacional de avances educativos (NAEP) reúne los datos sobre las bases y los logros educativos de los estudiantes en Estados Unidos. A veces se conoce como "la boleta de calificaciones del país", pues proporciona una escala para medir el avance de los estudiantes y comparar sus logros entre los distintos estados y con respecto del tiempo. Se reúne bastante información para cada estudiante, maestro y escuela participante. Además de las calificaciones de varias materias, los datos a nivel de estudiante incluyen información como sexo, raza, grupo étnico, cursos tomados y variables relacionadas con el nivel socioeconómico. La información en el nivel de escuela incluye recursos fiscales, métodos educativos, características del cuerpo de estudiantes y las esperanzas de logros académicos.

Los datos NAEP pueden servir para identificar variables en el nivel de escuela y en el de estudiante asociadas con los avances en matemáticas entre los estudiantes de octavo grado. Para simplificar la exposición, consideremos una característica en el nivel de estudiante, el sexo, y una característica en el nivel de escuela, el tiempo promedio ocupado en el grupo para los exámenes de matemáticas. En la práctica, por supuesto, es probable que se incluyan más variables en el modelo, ya que sería de esperar que varias características estén asociadas con los avances en matemáticas que se estén evaluando. Sea Y_{ij} la calificación en matemáticas del estudiante j en la escuela i de la muestra y sea $x_{ij} = 1$ si el estudiante j en la escuela i es mujer y 0 si es hombre.

Esperamos un efecto de conglomerados en estos datos; medir todas las variables que podrían estar asociadas con las calificaciones de los estudiantes en matemáticas es imposible, y las características de las escuelas, los maestros y los vecindarios no incluidas en el modelo inducen una correlación positiva en las calificaciones dentro de una escuela. Por ejemplo, el maestro de séptimo y octavo grado de una escuela podría ser muy bueno para inculcar a los estudiantes el gusto por el estudio de las matemáticas, pero esa excelencia no está registrada en la encuesta. Los estudiantes de esos grupos podrían tener entonces un mejor desempeño promedio en los exámenes, de modo que sus calificaciones fuesen más similares, incluso después de ajustar los covariados conocidos, que las calificaciones de una muestra aleatoria de estudiantes de la población. Al considerar características no medidas como éstas sobre todas las escuelas, el resultado es un coeficiente positivo de correlación entre clases.

Así, es probable que un modelo $Y_{ij} = \beta_0 + x_{ij}\beta_1 + \varepsilon_{ij}$, donde las ε_{ij} son variables aleatorias independientes con media 0 y varianza σ^2, sea inadecuado para estos datos. Si se adopta este modelo erróneo y los datos de todos los estudiantes se pasan por el procedimiento REG de SAS, entonces los valores p para las estimaciones de los parámetros serán demasiado pequeños. Además, el modelo no permite distintas relaciones entre el género y la calificación

en distintas escuelas, lo que realmente podría ocurrir, ya que algunas escuelas podrían animar más a los estudiantes de un sexo que a los del otro.

Un modelo que incorpora los efectos de conglomerados y permite que las escuelas tengan distintas pendientes para el sexo es

$$Y_{ij} = \beta_{0i} + (x_{ij} - \bar{x}_i)\beta_{1i} + \varepsilon_{ij}.$$

En este caso, suponemos que las ε_{ij} son variables aleatorias independientes $N(0,\sigma^2)$; restamos la media de x_{ij} para la escuela i, $\bar{x}_i$, de cada x_{ij}, de modo que podemos interpretar β_{0i} como la calificación promedio en la escuela i. La escuela i tiene su propio modelo de regresión lineal, con ordenada al origen β_{0i} y pendiente β_{1i}, pero las pendientes y las ordenadas al origen de las distintas escuelas también están relacionadas mediante un modelo. Un modelo sencillo para las pendientes y las ordenadas al origen les permite estar esencialmente distribuidas en forma aleatoria en torno de una media:

$$\beta_{0i} = \beta_0 + \delta_{0i};\ \beta_{1i} = \beta_1 + \delta_{1i},$$

donde δ_{0i} y δ_{1i} sigue una distribución normal bivariada con $E_M[\delta_{0i}] = E_M[\delta_{1i}] = 0, V_M[\delta_{0i}] = \tau_{00}, V_M[\delta_{1i}] = \tau_{11}$, y $\mathrm{Cov}_M(\delta_{0i},\delta_{1i}) = \tau_{01}$. En esta situación, podemos escribir el modelo como

$$Y_{ij} = \beta_0 + (x_{ij} - \bar{x}_i)\beta_1 + \delta_{0i} + (x_{ij} - \bar{x}_i)\delta_{1i} + \varepsilon_{ij}. \quad \textbf{(11.12)}$$

El parámetro β_0 representa la calificación media de las escuelas; β_1 representa la pendiente media para el género en las escuelas. Los efectos aleatorios δ_{0i} y δ_{1i} representan la diferencia en la ordenada al origen y la pendiente entre la escuela i y los valores promedio de cada una para todas las escuelas; miden el efecto de la escuela. Por último, ε_{ij} se refiere a la desviación adicional con respecto de la media debida al estudiante individual, después de tomar en cuenta el efecto del género y la escuela.

Observe que si $\tau_{00} = \tau_{11} = 0$, no hay efecto de la escuela sobre la calificación en el examen y en ese caso el modelo se reduce a un modelo de regresión lineal regular. Sin embargo, en la mayoría de las aplicaciones, las pendientes y las ordenadas al origen varían de una escuela a otra. ■

En estadística, el modelo en (11.12) es un ejemplo de **modelo lineal mixto**; tiene efectos fijos (β_0 y β_1) y aleatorios (δ_{0i}, δ_{1i} y ε_{ij}). En econometría, (11.12) se conoce con frecuencia como un **modelo de regresión con coeficientes aleatorios**; en las ciencias sociales se conoce como un **modelo lineal de varios niveles o jerárquico**. La edición de verano de 1995 del *Journal of Educational and Behavioral Statistics* se dedicó a los modelos de varios niveles; estos artículos contienen una útil bibliografía y son un buen punto de partida para lecturas posteriores. Otras referencias que proporcionan una buena introducción al tema son deLeeuwy Kreft (1986), Goldstein (1987), Goldstein y Silver (1989), y Bryk y Raudenbush (1992). Estos modelos se pueden ajustar en el procedimiento MIXED de SAS o en paquetes especializados como HLM (Bryk *et al* 1988) o ML3 (Prosser *et al* 1992).

El modelo mixto en (11.12) es un modelo de superpoblación y se supone válido para todas las escuelas y estudiantes de la población. Una ventaja del uso de tal modelo es que no requiere que las escuelas se elijan al azar, siempre que el modelo describa la población. Un enfoque de modelo mixto también congenia con la verificación de diversas teorías en torno de la educación matemática.

El modelo en (11.12) también es útil como punto de partida para investigaciones posteriores. Podemos estimar los efectos aleatorios δ_{0i} y δ_{1i} para cada escuela; es posible que el

investigador quiera examinar escuelas con valores inusualmente altos o bajos para ver por qué esas escuelas pudieran ser distintas. También es posible que quiera incluir otras variables de predicción al estimar las ordenadas al origen y las pendientes para las distintas escuelas. Por ejemplo, podría conjeturar que el hecho de realizar más exámenes de matemáticas en una escuela podría implicar mejores calificaciones o conducir a una pequeña diferencia de sexo en la escuela. Este predictor adicional puede incluirse fácilmente en el modelo mixto. Sea z_i la cantidad promedio de tiempo invertido en exámenes de matemáticas en la escuela i. Entonces podemos modelar la ordenada al origen y la pendiente en la escuela i como

$$\beta_{0i} = \beta_0 + \gamma_0 z_i + \delta_{0i} \text{ y } \beta_{1i} = \beta_1 + \gamma_1 z_i + \delta_{1i},$$

donde γ_0 representa el efecto del tiempo invertido en los exámenes de matemáticas sobre la ordenada al origen y δ_{0i} representa el efecto restante de la escuela después de ajustar con respecto de z_i.

11.5 Regresión logística

Generalmente, en la regresión lineal, se considera la variable de respuesta como aproximadamente continua; por ejemplo, el peso al nacer, el ingreso o el área de una hoja. Sin embargo, en las encuestas hay muchas variables dicotómicas, donde y_i sólo asume los valores 1 (sí) o 0 (no). Frecuentemente se emplea la **regresión logística** (vea una bibliografía general en Hosmer y Lemeshow 1989) para predecir las probabilidades de tener la respuesta 1 para las variables dicotómicas.

Sea $\mathbf{x}$ un vector de variables independientes y β el vector de parámetros desconocidos. Entonces, el modelo estándar de regresión logística asume la forma

$$p(\mathbf{x}) = \frac{\exp(\mathbf{x}^T\beta)}{1+\exp(\mathbf{x}^T\beta)}, \tag{11.13}$$

donde $p(\mathbf{x})$ representa la probabilidad de que una unidad con covariados $\mathbf{x}$ tenga 1 como respuesta. En forma alternativa, podemos expresar el modelo en escala logit, donde $\text{logit}(p) = \ln[p/(1-p)]$:

$$\text{logit}[p(\mathbf{x})] = \mathbf{x}^T\beta. \tag{11.14}$$

EJEMPLO 11.8 Para los datos del ejemplo 10.1, sea $y_i = 1$ si la familia i tiene una computadora y $y_i = 0$ en caso contrario. Sea $x_i = 1$ si la familia i está suscrita a la televisión por cable y 0 en caso contrario. El modelo de regresión logística ajustado es

$$\widehat{\text{logit}}[p_i] = -0.177 - 0.281x_i.$$

Observe que la pendiente, -0.28, es el cociente de posibilidades logarítmicas del ejemplo 10.1. Es fácil regresar la transformación a las probabilidades condicionales predichas: Cuando $x = 1$, entonces $\ln[\hat{p}/(1-\hat{p})] = -0.4573184$, de modo que

$$\hat{p} = \frac{\exp(-0.4573184)}{1+\exp(-0.4573184)} = 0.388 = \frac{119}{307}. \quad \blacksquare$$

Gran parte del análisis anterior sobre regresión lineal en este capítulo también se aplica a la regresión logística: un diseño de encuesta complejo afectará a los errores estándar de los coeficientes de regresión logística, así como afecta a los errores estándar de los coeficientes de regresión lineal. La regresión logística con una variable independiente dicotómica es esencialmente equivalente a determinar el cociente de posibilidades en una tabla de contingencias 2×2, de modo que el análisis del capítulo 10 acerca de la forma en que el diseño de muestreo afecta las pruebas comunes de bondad de ajuste también se aplica a la verificación de la significación de los coeficientes de la regresión logística.

Binder (1983), Chambless y Boyle (1985) y Roberts et al (1987) establecen una teoría basada en el diseño para la estimación de los parámetros de la regresión logística. Así como la teoría basada en el diseño para la regresión lineal comenzaba con la definición de las cantidades de interés en la población usando las ecuaciones normales, en este caso definimos las cantidades de interés en términos de la función de verosimilitud que adoptaríamos si dispusiéramos de toda la población para el estudio. Si hay N unidades en la población, esta verosimilitud (suponiendo independencia) es

$$\pounds(\beta) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i} \tag{11.15}$$

donde $p_i = \exp(\mathbf{x}_i^T\beta)/[1+\exp(\mathbf{x}_i^T\beta)]$ representa la probabilidad de que una unidad con covariados $\mathbf{x}_i$ tenga 1 como respuesta. El parámetro de población finita $\mathbf{B}$ se define entonces como la estimación de máxima verosimilitud de β usando (11.15). El parámetro $\mathbf{B}$ es la solución del sistema de ecuaciones

$$\sum_{i=1}^{N} x_{ij}\left[y_i - \frac{\exp(\mathbf{x}_i^T\mathbf{B})}{1+\exp(\mathbf{x}_i^T\mathbf{B})}\right] = 0 \quad \text{para } j = 1,\ldots,p \tag{11.16}$$

si pudiéramos observar a todos los elementos de la población.

Ahora que hemos definido $\mathbf{B}$, lo estimamos sustituyendo las estimaciones para los totales de la población. Una estimación de $\mathbf{B}$ basada en el diseño está dada por la solución $\hat{\mathbf{B}}$ de

$$\sum_{i\in S} w_i x_{ij}\left[y_i - \frac{\exp(\mathbf{x}_i^T\hat{\mathbf{B}})}{1+\exp(\mathbf{x}_i^T\hat{\mathbf{B}})}\right] = 0 \quad \text{para } j = 1,\ldots,p, \tag{11.17}$$

donde $\mathcal{S}$ denota las unidades incluidas en la muestra. La observación i en la muestra representa w_i observaciones en la población.

Para una estimación de β, basada en el modelo, simplemente omitimos los pesos: $\hat{\beta}$ es la solución de

$$\sum_{i\in S} x_{ij}\left[y_i - \frac{\exp(\mathbf{x}_i^T\hat{\beta})}{1+\exp(\mathbf{x}_i^T\hat{\beta})}\right] = 0 \quad \text{para } j = 1,\ldots,p. \tag{11.18}$$

La estimación de la varianza para la regresión logística se analiza en las referencias anteriores. Rao *et al* (1998) presenta una versión modificada de las calificaciones para verificar la significación de los coeficientes de regresión logística.

La regresión logística tiene una diferencia importante con la regresión lineal. En la sección 11.2 observamos el sesgo que puede aparecer al estimar los parámetros de regresión lineal si las probabilidades de selección están relacionadas con la variable de respuesta, pero las probabilidades diferentes de selección no se toman en cuenta en el análisis. Por ejemplo, en una encuesta de salud, la presión sanguínea podría emplearse como variable de estratificación y usar una mayor fracción de muestreo en el estrato de presión alta que en el estrato de presión baja. Si ignoramos las probabilidades de selección y ajustamos un modelo de regresión lineal para predecir la variable continua *presión sanguínea* a partir de

covariados como *edad, dieta* e *historia de uso del cigarrillo*; los coeficientes de regresión pueden tener un sesgo importante para la estimación de **B**.

Sin embargo, Prentice y Pyke (1979) muestran que si un modelo de regresión logística es válido y contiene un término de ordenada al origen, entonces la ordenada al origen es la única estimación de parámetro afectada por un diseño de muestra que dependa de los y. Tales diseños de muestra son particularmente comunes en epidemiología y economía, donde se conocen como *estudios de control de casos* y *muestreo con base en las opciones*. En una aplicación a la epidemiología, la población se podría dividir en dos estratos: personas con cáncer de pulmón y personas sin cáncer de pulmón. Se elige una muestra de cada estrato; como el cáncer de pulmón es raro, la muestra estratificada tiene una fracción de muestreo mucho mayor (y menores pesos de muestreo) en el estrato con cáncer que en el estrato sin cáncer. Pero si el interés principal es estimar los coeficientes de edad, dieta e historia de uso del cigarro en una regresión logística, el muestreo desproporcionado no establece una diferencia en un análisis basado en el modelo. Sería de esperar que si el modelo es bueno, la única diferencia entre un análisis ponderado o no aparecería en los términos de la ordenada al origen. Evidentemente, si se emplea una muestra por conglomerados, tendremos que tomar en cuenta la dependencia de los datos inducida por los conglomerados en el modelo de regresión logística para la estimación de la varianza, como se analiza en Scott y Wild (1989).

11.6 Estimación generalizada por regresión para los totales de la población

En el capítulo 3 presentamos la estimación por proporción y por regresión en el marco de las muestras aleatorias simples, con estimadores

$$\hat{t}_{yr} = \frac{\hat{t}_y}{\hat{t}_x} t_x$$

$$\hat{t}_{\text{yreg}} = \hat{t}_y + \hat{B}_1(t_x - \hat{t}_x).$$

Ahora extenderemos estas estimaciones a las muestras de encuestas complejas. Queremos mejorar el estimador $\hat{t}_y = \sum_{i\in S}\omega_i y_i$ incluyendo información auxiliar mediante el modelo

$$Y_i/\mathbf{x}_i = \mathbf{x}_i^T\beta + \varepsilon_i, \tag{11.19}$$

con $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip})$ y $V_M(\varepsilon_i) = \sigma_i^2$. Suponemos conocidos los totales reales de la población $\mathbf{t}_x$ y que por ello se pueden usar para ajustar la estimación $\hat{t}_y$. Permitimos que las varianzas difieran de modo que la estimación por proporción y la estratificación posterior también se ajusten a este marco de referencia general. Estamos usando el punto de vista apoyado por el modelo, descrito con más detalle en Särndal *et al* (1992, capítulos 6 y 7).

Definimos

$$B = (\mathbf{X}_U^T\Sigma_U^{-1}\mathbf{X}_U)^{-1}\mathbf{X}_U^T\Sigma_U^{-1}y_U,$$

donde Σ_U es una matriz diagonal, con σ_i^2 como elemento i en la diagonal. El parámetro de población finita **B** es la estimación por mínimos cuadrados ponderados de β para las observaciones en la población, usando el modelo en (11.19). Así, la forma de **B** está inspirada por (11.19), pero consideramos a **B** como una cantidad de población finita. La entrada (jk) de

$(\mathbf{X}_U^T\Sigma_U^{-1}\mathbf{X}_U)$ es $\sum_{i=1}^N x_{ij}x_{ik}/\sigma_i^2$. Ahora estimamos **B** mediante

$$\hat{B} = (\mathbf{X}_S^T W_S \Sigma_S^{-1}\mathbf{X}_S)^{-1}\mathbf{X}_S^T W_S \Sigma_S^{-1} y_S. \tag{11.20}$$

El **estimador generalizado por regresión** del total de la población es

$$\hat{t}_{\text{ygreg}} = t_y + (\mathbf{t}_x - \hat{\mathbf{t}}_x)^T\hat{\mathbf{B}}. \tag{11.21}$$

Usando la linealización,

$$V(\hat{t}_{\text{ygreg}}) = V[\hat{t}_y + (\mathbf{t}_x - \hat{\mathbf{t}}_x)^T\hat{\mathbf{B}}] \approx V(\hat{t}_y - \hat{\mathbf{t}}_x^T\mathbf{B}).$$

Sea $e_i = y_i - x_i^T\hat{\mathbf{B}}$ el residuo i. Entonces podemos estimar la varianza mediante

$$\hat{V}(\hat{t}_{\text{ygreg}}) = \hat{V}\left(\sum_{i\in S} w_i e_i\right).$$

Si el modelo es bueno, esperamos que la variabilidad en los residuales sea menor que la variabilidad en las observaciones originales, de modo que el estimador generalizado por regresión será más eficiente que $\hat{t}_y$. En una muestra aleatoria simple (MAS), por ejemplo,

$$\hat{V}_{\text{MAS}}(\hat{t}_y) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{\sum_{i\in S}(y_i - \bar{y})^2}{n-1},$$

pero

$$\hat{V}_{\text{MAS}}(\hat{t}_{\text{ygreg}}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{\sum_{i\in S} e_i^2}{n-1};$$

si los residuales tienden a ser menores que la desviación de y_i en torno de la media, entonces la varianza estimada es menor para el estimador generalizado por regresión.

EJEMPLO 11.9 *Estimación por proporción*

Adopte el modelo

$$y_i = \beta x_i + \varepsilon_i, \quad V_M(\varepsilon_i) = \sigma^2 x_i.$$

Entonces,

$$\hat{B} = \left(\sum_{i\in S}\frac{w_i x_i^2}{x_i}\right)^{-1}\sum_{i\in S}\frac{w_i x_i y_i}{x_i} = \frac{\sum_{i\in S} w_i y_i}{\sum_{i\in S} w_i x_i} = \frac{\hat{t}_y}{\hat{t}_x}.$$

El estimador generalizado por regresión del total de la población es

$$\hat{t}_{\text{ygreg}} = \hat{t}_y + (t_x - \hat{t}_x)\frac{\hat{t}_y}{\hat{t}_x} = \frac{t_x\hat{t}_y}{\hat{t}_x},$$

que es el estimador por proporción estándar. ■

EJEMPLO 11.10 *Estratificación posterior*

Suponga que conocemos las cifras de población N_c para C estratos posteriores, $c = 1,\ldots,C$. Defina las variables $x_{ic} = 1$ si la unidad de observación i está en el estrato posterior c y 0 en caso contrario. Considere el modelo

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_C x_{iC} + \varepsilon_i,$$

con $V_M(\varepsilon_i) = \sigma^2$. Entonces

$$\sigma^2 \mathbf{X}_U^T \Sigma_U^{-1} \mathbf{X}_U = \mathbf{X}_U^T \mathbf{X}_U = \text{diag}\,(N_1,\ldots,N_C),$$

y

$$\sigma^2 \mathbf{X}_S^T \mathbf{W}_S \Sigma_S^{-1} \mathbf{X}_S = \mathbf{X}_S^T \mathbf{W}_S \mathbf{X}_S = \text{diag}\,(\hat{N}_1,\ldots,\hat{N}_C).$$

Como resultado, $\hat{B}_c = \hat{t}_{yc} / \hat{N}_c$, donde $\hat{t}_{yc} = \sum_{i\in S} w_i x_{ic} y_i$ es el total estimado de la población en el estrato posterior c y $\hat{N}_c = \sum_{i\in S} w_i x_{ic}$ es la cifra estimada de la población en el estrato posterior c. El estimador generalizado por regresión es

$$\hat{t}_{y\text{greg}} = \hat{t}_y + \sum_{c=1}^{C} (N_c - \hat{N}_c)\frac{\hat{t}_{yc}}{\hat{N}_c} = \sum_{c=1}^{C} \frac{N_c \hat{t}_{yc}}{\hat{N}_c}. \quad \blacksquare$$

Con frecuencia, las variables auxiliares son útiles para muchas de las variables de respuesta de interés. Tal vez quiera estratificar posteriormente por grupos de edad, raza y sexo al estimar cada total de población en su encuesta. Esto puede implantarse fácilmente, pues el estimador generalizado por regresión es un estimador lineal en y. Para ver esto, definimos

$$g_i = 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_x)^T (\mathbf{X}_S^T \mathbf{W}_S \Sigma_S^{-1} \mathbf{X}_S)^{-1} \frac{\mathbf{x}_i}{\sigma_i^2}.$$

Entonces,

$$\hat{t}_{y\text{greg}} = \sum_{i\in S} w_i g_i y_i,$$

donde los g_i no dependen de valores de la variable de respuesta. Para estimar totales con el estimador generalizado por regresión, formamos una nueva columna en los datos, con los valores $a_i = w_i g_i$. Entonces usamos el vector de a_i como el vector de ponderación para estimar el total de población de cualquier variable.

11.7 Ejercicios

1 Lea uno de los siguientes artículos, o algún otro en donde se utilice la regresión o la regresión logística en datos de una encuesta compleja.

Stevens, R. G., D. Y. Jones, M.S. Micozzi, y P.R. Taylor. 1988. Body iron stores and the risk of cancer. *New England Journal of Medicine* 319: 1047-1052.

Martorell, R., F. Mendoza, y R. O. Castillo. 1989. Genetic and environmental determinants of growth in Mexican-Americans. *Pediatrics* 84: 864-871.

Patterson, C.J., J.S. Kupersmidt, y N.A. Vaden. 1990. Income level, gender, ethnicity, and household composition as predictors of children's school-based competence. *Child Development* 61: 485-494.

Tymms, P.B., y C.T. Fitz-Gibbon. 1992. The relationship between part-time employment and A-level results. *Educational Research* 34: 193-199.

Breen, N., y L. Kessler. 1994. Changes in the use of screening mammography: Evidence from the 1987 and 1990 National Health Interview Surveys. *American Journal of Public Health* 84: 62-67.

Subar, A. F., R.G. Ziegler, B. H. Patterson, G. Ursin, y B. Graubard. 1994. US dietary patterns associated with fat intake: The 1987 National Health Interview Survey. *American Journal of Public Health* 84: 359-366.

Bachman, R., y A.L. Coker. 1995. Police involvement in domestic violence: The interactive effects of victim injury, offender's history of violence, and race. *Violence and Victims* 10:91-106.

Flegal, K.M., R.P. Troiano, E. R. Pamuk, R.J. Kuczmarksi, y S.M. Campbell. 1995. The influence of smoking cessation on the prevalence of overwigh in the United States. *New England Journal of Medicine* 333: 1165-1170.

Sashi, C.M., y L. W. Stern. 1995. Product differentiation and market performance in producer goods industries. *Journal of Business Research* 33: 115-127.

Singhapakdi, A., K.L. Kraft, S.J. Vitell, y K. C. Rallapalli. 1995. The perceived importance of ethics and social responsibility on organizational effectiveness: A survey of marketers. *Journal of the Academy of Marketing Science* 23:49-56.

Wang, X., B. Zuckerman, G.A. Coffman, y M.J. Corwin. 1995. Familial aggregation of low birth weight among whites and blacks in the United States. *New England Journal of Medicine* 333: 1744-1749.

Escriba una crítica del artículo. ¿Cuál es la finalidad y el diseño de la encuesta? ¿Cuál es el objetivo del análisis? ¿Cómo usan los autores la información del diseño de la encuesta en el análisis? ¿Cree que el análisis de datos se realiza correctamente? ¿Por qué? En caso contrario, ¿cómo se podría mejorar? ¿Son justificadas las conclusiones extraídas en el artículo?

2 Una investigadora quiere estudiar la relación entre la edad de un niño, el número de hermanos y la cantidad en dólares de la lista de regalos solicitados por el niño a Santa Claus. Ella quiere estimar también la cantidad total de niños que visitan a Santa Claus y la cantidad total en dólares de todas las solicitudes de los niños. Sería difícil construir un marco de muestreo de los niños que visitan a Santa Claus entre el primero y el 24 de diciembre, pero tiene una lista de centros comerciales y tiendas de la ciudad donde Santa Claus estará presente, así como las veces que estará en cada lugar. Los sitios de Santa Claus se dividen en cuatro categorías: 23 tiendas departamentales, 19 tiendas de descuento, 15 tiendas de juguetes y 5 centros comerciales. La investigadora quiere que usted le ayude a diseñar la muestra de niños.

a ¿Cuáles preguntas plantearía a la investigadora para aclarar el problema?

b Suponiendo que tiene las respuestas a todas sus preguntas, sugiera un diseño para la encuesta.

c ¿Cómo afectará su diseño de la encuesta al análisis de los datos? ¿Cuál es su propuesta para analizar los datos? ¿Existen otras variables explicativas que sugeriría a la investigadora?

3 Use los datos en el archivo anthrop.dat para este problema.

- **a** Construya una población a partir de las 3000 observaciones en anthrop.dat de la cual se hayan eliminado los 1000 individuos con los valores máximos de y. Ahora, extraiga una muestra aleatoria simple de tamaño 200 de los 2000 individuos restantes y grafique los datos junto con la línea de regresión OLS. ¿Cuál es la relación de esta línea con la línea de regresión de la población?
- **b** Repita la parte (a), pero use como población los 2000 individuos con los menores valores de x.
- **c** ¿Existe una diferencia entre las ecuaciones de regresión en las partes (a) y (b)? Explique y relacione sus hallazgos con el modelo en (11.1).

4 Use los datos del archivo nybight.dat (vea el ejercicio 19 del capítulo 4) para este problema. Utilice los datos de 1974 para estimar los coeficientes en un modelo de regresión lineal que prediga el peso de la pesca a partir del número de peces capturados. Proporcione los errores estándar de sus estimaciones. (¡Asegúrese de graficar los datos!)

5 Realice un análisis basado en el modelo para el marco del ejercicio 4. Examine los residuales y postule una estructura adecuada de varianza para el modelo.

6 Repita el ejercicio 4 para predecir la cantidad de especies capturadas a partir de la temperatura en la superficie.

7 Repita el ejercicio 5 para predecir el número de especies capturadas a partir de la temperatura en la superficie.

8 Use los datos del archivo teachers.dat (descrito en el ejercicio 16 del capítulo 5) para este problema.

- **a** Estime los coeficientes en un modelo de regresión lineal para predecir *preprmin* a partir de *size*. Proporcione los errores estándar de sus estimaciones. ¿Hay alguna evidencia de que las dos variables estén relacionadas entre sí? (¡Asegúrese de graficar los datos!)
- **b** Realice un análisis basado en el modelo de los mismos datos. Examine los residuales y postule una estructura adecuada de varianza para el modelo.

9 Use los datos del archivo books.dat (descrito en el ejercicio 6 del capítulo 5) para este problema.

- **a** Grafique *replace* contra *purchase* para los datos en bruto.
- **b** Grafique *replace* contra *purchase* usando los pesos de muestreo.
- **c** Utilice un enfoque basado en el diseño para estimar la ecuación de regresión para predecir *replace* a partir de *purchase*, junto con los errores estándar. ¿Cuántos grados de libertad utilizaría al construir un intervalo de confianza para la pendiente?

10 Para la situación del ejercicio 9, postule un modelo para la estructura de varianza. Use ese modelo para estimar la pendiente de la línea de regresión para predecir *replace* a partir de *purchase*. ¿Cuál es la relación de su estimación y su error estándar con sus respuestas en el ejercicio 9?

11 Emplee el conjunto de datos del ejercicio 13 del capítulo 4 para este problema. Use los pesos para ajustar un modelo de regresión para predecir *acres92* a partir de *largef92*. Proporcione un error estándar para la pendiente estimada. Ahora ignore el diseño de muestreo y calcule la estimación OLS de la pendiente. ¿Difieren en algo las estimaciones puntuales? Explique por qué examinando las gráficas de los datos.

12 Lush (1945, 95) analiza distintas estimaciones de la heredabilidad del porcentaje de grasa en la leche para rebaños de vacas lecheras. La *heredabilidad* se define como el porcentaje de variabilidad en el porcentaje de grasa que puede atribuirse a diferencias en la herencia de distintos individuos; el resto de la variabilidad se atribuye a las diferencias en el ambiente. Observa que cuando el rebaño se consideró como una muestra aleatoria simple, la estimación de la heredabilidad era aproximadamente 0.8; al realizar una regresión del porcentaje de grasa para las hijas con respecto del porcentaje de grasa de las madres, donde cada madre era representada por un único registro, la estimación de la heredabilidad disminuía por debajo de 0.3.

Desde una perspectiva de muestreo, ¿por qué son tan distintas estas estimaciones? Discuta la forma en que analizaría los datos de los rebaños completos desde una perspectiva basada en el diseño y una basada en el modelo.

13 *Comparación de medias por dominio.* Suponga que podemos dividir la población en dos grupos, con tamaños respectivos N_1 y N_2 y medias de población $\bar{y}_{1U}$ y $\bar{y}_{2U}$. La media global de la población es $\bar{y}_U = (N_1\bar{y}_{1U} + N_2\bar{y}_{2U}/N$, con $N = N_1 + N_2$. Sea $x_i = 1$ si la unidad de observación i está en el grupo 1 y $x_i = 0$ si está en el grupo 2. El peso de la unidad de observación i es w_i.

Muestre que $B_1 = \bar{y}_{1U} - \bar{y}_{2U}$ y $B_0 = \bar{y}_{2U}$. Muestre también que

$$B_1 = \frac{\sum_{i\in S} w_i x_i y_i}{\sum_{i\in S} w_i x_i} - \frac{\sum_{i\in S} w_i (1-x_i) y_i}{\sum_{i\in S} w_i (1-x_i)} = \hat{\bar{y}}_1 - \hat{\bar{y}}_2$$

y $\hat{B}_0 = \hat{\bar{y}}_2$.

14 Considere los datos de una muestra aleatoria simple en el archivo uneqvar.dat.

- **a** Grafique y contra x.
- **b** Determine la línea de regresión ajustada, bajo la hipótesis de varianzas iguales.
- **c** Calcule $\hat{V}_M(\hat{\beta}_1)$ y $\hat{V}_L(\hat{\beta}_1)$. ¿Cuál es su relación?

15 Muestre que (11.10) es equivalente a (11.6) y (11.7) para la regresión lineal.

***16** (Requiere teoría de modelos lineales.) Suponga que el modelo "verdadero" que describe la relación entre x y y es

$$Y_i \mid x_i = \beta_0 = \beta_1 x_i = \varepsilon_i,$$

donde las ε_i se generan de manera independiente a partir de una distribución $N(0,\sigma_i^2)$. ¿Cuál es la matriz de covarianza para las estimaciones OLS de los parámetros? ¿Cómo se relaciona esto con el análisis de los distintos estimadores de la varianza en las páginas 357-358?

17 Frecuentemente, el coeficiente de determinación R^2 se cita en los análisis de regresión. Para una regresión lineal, la cantidad de población finita R^2 se define como

$$R^2 = \frac{B_1 \sum_{i=1}^{N} (x_i - \bar{x}_U)(y_i - \bar{y}_U)}{\sum_{i=1}^{N} (y_i - \bar{y}_U)^2}.$$

a Muestre que R^2 es el cuadrado del coeficiente de correlación de la población, R, definido en (3.1).

b Escriba R^2 como función de los totales de la población.

c Proporcione un estimador $\hat{R}^2$ de R^2 para datos de una encuesta compleja, usando los pesos.

18 Fienberg (1980) dice: "No conocemos una justificación para la aplicación de los métodos multivariados estándar a los datos ponderados... la inserción automática de una matriz de pesos basados en la muestra dentro de un análisis por mínimos cuadrados ponderados conduce con bastante frecuencia a confusiones e incluso puede ser incorrecto". ¿Cuál es el punto de vista de la inferencia por regresión defendido por Fienberg? ¿Cuál es su opinión?

19 Suponga un modelo

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

con $V_M(\varepsilon_i) = \sigma^2$. ¿Cuál es el estimador generalizado por regresión de t_y? Muestre que $\hat{t}_{x\text{greg}} = t_x$.

12

Otros temas de muestreo*

Casi todos los estados han devuelto sus censos. Le envío los resultados, con tinta negra, si están basados (en la medida de lo posible) en los datos reales, y con tinta roja si no son los datos regresados, aunque bastante conocidos. Con un pequeño margen para las omisiones, somos más de cuatro millones, aunque de hecho sabemos que las omisiones han sido muy grandes.

—Thomas Jefferson, carta a David Humphreys, 23 de agosto de 1791

12.1 Muestreo en dos etapas

En ocasiones, tal vez se desee utilizar la estratificación, el muestreo con probabilidades diferentes o la estimación por proporción para aumentar la precisión de su estimador, pero el marco de muestreo no tiene información de variables auxiliares útiles. Por ejemplo, suponga que quiere obtener una muestra de empresas con probabilidad proporcional al ingreso, pero que no tiene información del ingreso en el marco de muestreo. O bien, tal vez quiera estimar el volumen total de madera cortado en un bosque, midiendo el volumen total en una muestra de carga de camiones con troncos. El volumen de madera en un camión está relacionado con el peso de la carga, de modo que esperaría tener mayor precisión usando la estimación por proporción con y_i igual al volumen de madera en el camión i y x_i igual al peso del camión i. Pero la estimación por proporción $\hat{t}_{yr} = t_x \hat{t}_y / \hat{t}_x$ requiere conocer el peso total de todas las cargas y no es práctico pesar cada camión de la población.

El **muestreo en dos etapas**, también llamado **muestreo doble**, brinda una solución. El muestreo en dos etapas, introducido por Neyman (1938), es útil cuando es relativamente caro medir la variable de interés y, pero es posible medir fácilmente una variable correlacionada x y usarla para mejorar la precisión del estimador de t_y.

Suponga que la población tiene N unidades de observación. La muestra se extrae en dos etapas:

1 *Muestra de la primera etapa.* Extraemos una muestra de probabilidad de $n^{(1)}$ unidades, conocida como muestra de la primera etapa. Medimos las variables auxiliares $\mathbf{x}$ para cada unidad de la muestra de la primera etapa. En la encuesta de empresas, se debería extraer una muestra aleatoria de declaraciones de impuestos y registrar el ingreso reportado por cada empresa de la muestra. Para medir el volumen de madera se podría pesar una muestra de

camiones seleccionados al azar o con probabilidad proporcional al volumen de madera estimado. La muestra de la primera etapa es por lo general relativamente grande (y puede serlo, pues no es caro obtener la información auxiliar) y debe brindar información precisa acerca de la distribución de las x.

2 *Muestra de la segunda etapa.* Ahora actúe como si la muestra de la primera etapa fuese una población y seleccione una muestra de probabilidad de la muestra de la primera etapa. Mida las variables de interés para cada unidad de la submuestra, conocida como muestra de la segunda etapa. Puesto que se está considerando la muestra de la primera etapa como la población de la cual se extrae la muestra de la segunda etapa, se puede utilizar la información auxiliar recogida en la primera etapa al diseñar la muestra de la segunda etapa. Se podría seleccionar las empresas con las cuales establecer contacto con probabilidad proporcional al ingreso medido en la muestra de la primera etapa, o bien, utilizar la información de ingreso para estratificar las empresas de la muestra de la primera etapa y luego establecer contacto con un subconjunto seleccionado al azar de las empresas en cada estrato de ingreso para obtener la información deseada sobre variables como los gastos totales. Se podrían seleccionar las cargas en las que el volumen de madera se medirá con probabilidad proporcional al peso, o bien, utilizar la información de la muestra de la primera etapa para obtener una mejor estimación del peso total y emplear la estimación por proporción. En cada caso, es relativamente caro medir las variables y, pero y está correlacionada con x.

El muestreo en dos etapas puede ahorrar tiempo y dinero si es barato obtener la información auxiliar y si contar con información auxiliar puede aumentar la precisión de las estimaciones para las cantidades de interés.

EJEMPLO 12.1 Stockford y Page (1984) utilizaron el muestreo en dos etapas para estimar el porcentaje de veteranos de la guerra de Vietnam residentes de los hospitales de la Oficina de Veteranos (VA) de Estados Unidos que realmente combatieron en Vietnam.

El Censo Anual de Pacientes (APC) de la VA en 1982 incluyó una muestra aleatoria de 20% de los internos en hospitales de la VA. Se incluyó la siguiente pregunta: "Si el periodo de servicio es la guerra de Vietnam, ¿estaba de servicio en Vietnam?" con las categorías de respuestas "sí", "no" y "no disponible". Se obtuvieron las respuestas a esta pregunta mediante los registros médicos de los pacientes. Sin embargo, esa respuesta podría ser imprecisa por varias razones: (1) Gran parte de los registros médicos fueron llenados por el propio paciente, quien podría no recordar la ubicación de su servicio debido a problemas médicos, o estar confundido acerca de la definición de servicio en Vietnam (algunos pilotos cuya estación de trabajo estaba oficialmente en Tailandia realizaron misiones sobre Vietnam); (2) un paciente podría afirmar erróneamente un servicio en Vietnam al suponer que la respuesta podría afectar los beneficios de la VA; o (3) puede haber errores al registrar la respuesta en el registro médico. Además, muchos pacientes no estuvieron "disponibles" para la respuesta. Así, la respuesta a la pregunta sobre el servicio en Vietnam de la encuesta APC fue insatisfactoria para estimar el porcentaje de veteranos de la guerra de Vietnam en hospitales de la VA que estuvieron de servicio en ese país.

Stockford y Page verificaron los registros militares para una submuestra estratificada de veteranos hospitalizados para determinar la clasificación real del servicio en Vietnam. La información de la encuesta original se utilizó para la estratificación, ya que se esperaban distintos porcentajes de servicio en los grupos "sí", "no" y "no disponible" en la encuesta APC. Se verificaron los registros militares de todos los pacientes en el estrato "no disponible". Se esperaba que las varianzas dentro de los estratos fuesen relativamente bajas en los

estratos "sí" y "no" pues, aunque los datos de la encuesta APC eran imprecisos, sería de esperar que un mayor porcentaje de personas que respondieron "sí" haya realizado su servicio en Vietnam que quienes respondieron "no"; se verificaron los registros militares de una submuestra de 10% para cada uno de estos dos estratos.

Los resultados de la pregunta "¿Estuvo de servicio en Vietnam?" fueron las siguientes:

Grupo APC	Clasificación en la encuesta APC	Tamaño de la submuestra	Servicio en Vietnam en la submuestra
Sí	755	67	49
No	804	72	11
No disponible	505	505	211
Total	2064	644	271

Como era de esperar, el porcentaje de veteranos con servicio en Vietnam difería en los tres grupos: de los veteranos con respuesta "sí" a la pregunta de la encuesta APC, 73% realmente sirvieron allí, en comparación con 15% para el grupo "no" y 42% para los veteranos cuya información no estaba disponible. ■

EJEMPLO 12.2 Frecuentemente, el muestreo en dos etapas se utiliza en los estudios de silvicultura. Se dispone de fotografías aéreas y sistemáticamente se distribuyen puntos en las fotografías. Se estudian las áreas alrededor de los puntos de las fotografías y se clasifican de acuerdo con el tipo de terreno: bosque, bosque improductivo, área no de bosque y agua. Entonces, se extrae una muestra de la primera etapa de puntos en la retícula con una fracción de muestreo mayor para los puntos de la retícula clasificados como bosque que los clasificados como no de bosque. Las áreas de la muestra de la primera etapa se examinan con más cuidado, para clasificarlas según el tamaño y densidad de los árboles. Luego, se extrae una submuestra de los puntos de la muestra de la primera etapa y se realizan mediciones como uso del suelo, volumen y mortalidad; el porcentaje de área de bosque de la muestra de la segunda etapa puede diferir un poco de la estimación fotográfica de la primera etapa y la estimación por proporción se puede usar en la muestra de la segunda etapa para aumentar su precisión. ■

EJEMPLO 12.3 En la sección 8.3 estudiamos el uso del muestreo en dos etapas en el ajuste por la ausencia de respuestas. Se extrae una muestra de probabilidad de la población; las unidades de la muestra se dividen en los dos estratos de las personas que responden y las que no. Luego, se extrae una submuestra de las personas que no responden. La muestra de la primera etapa es la muestra de probabilidad original. La variable

$$x_i = \begin{cases} 1 & \text{si la observación } i \text{ responde} \\ 0 & \text{si la observación } i \text{ es una persona que no responde} \end{cases}$$

se analiza para cada elemento de la muestra de la primera etapa. Luego, se emplea la información relativa a x_i en la muestra de la segunda etapa. Observamos el valor de interés y_i para todas las observaciones con $x_i = 1$; se extrae una submuestra para las observaciones con $x_i = 0$. ■

12.1.1 Teoría de muestreo en dos etapas

Primero estableceremos los resultados en general y luego para el caso en que las muestras de la primera y la segunda etapa son aleatorias simples. Un marco de referencia general para el muestreo en dos etapas aparece en Särndal y Swensson (1987).

Sea $\mathcal{S}^{(1)}$ la muestra de la primera etapa; las unidades seleccionadas para la muestra quedan determinadas por las variables aleatorias

$$Z_i = \begin{cases} 1 & \text{si la unidad } i \text{ está en la muestra de la primera etapa.} \\ 0 & \text{si la unidad } i \text{ no está en la muestra de la primera etapa.} \end{cases}$$

Sea $w_i^{(1)}$ el peso de muestreo para la muestra de la primera etapa: $w_i^{(1)} = 1/[P(Z_i = 1)]$. Observamos un vector de características auxiliares $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ik}]^T$ para cada unidad de observación en la muestra de la primera etapa. Podemos usar la teoría desarrollada en capítulos anteriores para estimar el total de la población para la variable auxiliar j como

$$\hat{t}_{x_j}^{(1)} = \sum_{i \in S^{(1)}} w_i^{(1)} x_{ij} = \sum_{i=1}^{N} Z_i w_i^{(1)} x_{ij}.$$

Ahora indicamos la pertenencia a la muestra de la segunda etapa $\mathcal{S}^{(2)}$ mediante la variable aleatoria

$$D_i = \begin{cases} 1 & \text{si la unidad } i \text{ está en la muestra de la segunda etapa.} \\ 0 & \text{si la unidad } i \text{ no está en la muestra de la segunda etapa.} \end{cases}$$

La probabilidad de que una unidad esté en la muestra de la segunda etapa depende de si está en la muestra de la primera etapa y también puede depender de información auxiliar reunida en la muestra de la primera etapa; denotamos esta independencia escribiendo $P(D_i = 1 \mid \mathbf{Z})$, donde $\mathbf{Z}$ es el vector $(Z_1, Z_2, \ldots, Z_N)^T$. Así, cuando determinamos una esperanza condicional sobre $\mathbf{Z}$, considerando que se conoce la información de la muestra de la primera etapa. Los pesos de submuestreo para la muestra final, la de la segunda etapa, también dependen de las unidades seleccionadas en la muestra de la primera etapa:

$$w_i^{(2)} = w_i^{(2)}(\mathbf{Z}) = \begin{cases} \dfrac{1}{P(Di = 1/\mathbf{Z})} & \text{si } Z_i = 1. \\ 0 & \text{si } Z_i = 0. \end{cases}$$

Un análogo del estimador de Horvitz-Thompson para el muestreo en dos etapas es

$$\hat{t}_y^{(2)} = \sum_{i \in S^{(2)}} w_i^{(1)} w_i^{(2)} y_i = \sum_{i=1}^{N} Z_i D_i w_i^{(1)} w_i^{(2)} y_i. \qquad \textbf{(12.1)}$$

Usamos el siguiente mecanismo para determinar las propiedades de los estimadores en dos etapas. Definimos

$$\hat{t}_y^{(1)} = \sum_{i \in S^{(2)}} w_i^{(1)} y_i = \sum_{i=1}^{N} Z_i w_i^{(1)} y_i.$$

Ahora, no sabemos lo que sea $\hat{t}_y^{(1)}$, pues sólo observamos las y_i en la muestra de la segunda etapa. Pero $\hat{t}_y^{(1)}$ sirve como el "total de la población" estimado en la segunda etapa (si conociéramos y_i para todas las unidades de la muestra de la primera etapa, estimaríamos y_i como $\hat{t}_y^{(1)}$. Al considerar la muestra de la primera etapa como conocida, tenemos

$$E[\hat{t}_y^{(2)} \mid \mathbf{Z}] = \sum_{i=1}^{N} Z_i w_i^{(1)} w_i^{(2)} y_i E[D_i \mid \mathbf{Z}] = \sum_{i=1}^{N} Z_i w_i^{(1)} y_i = \hat{t}_y^{(1)}.$$

Luego, al usar el condicionamiento sucesivo (consulte la sección B.4, página 434),

$$E[\hat{t}_y^{(2)}] = E\{E[\hat{t}_y^{(2)} \mid \mathbf{Z}]\} = E\left[\sum_{i=1}^{N} Z_i w_i^{(1)} y_i\right] = t_y.$$

Además, de la propiedad 5 de la sección B.4,

$$V(\hat{t}_y^{(2)}) = V(E[\hat{t}_y^{(2)} \mid \mathbf{Z}]) + E(V[\hat{t}_y^{(2)} \mid \mathbf{Z}]) = V(\hat{t}_y^{(1)}) + E(V[\hat{t}_y^{(2)} \mid \mathbf{Z}]).$$

El primer término es la varianza que se obtendría si y_i hubiese sido observado para cada observación en $\mathcal{S}^{(1)}$; el segundo término es la varianza adicional del submuestreo en la segunda etapa.

12.1.2 Muestreo en dos etapas con estimación de proporción

Definimos $\mathcal{S}^{(1)}$, $\mathcal{S}^{(2)}$, Z_i y D_i como antes. Medimos la variable auxiliar x_i para cada observación en la muestra de la primera etapa; de esa muestra, podemos estimar el total de población $t_x = \sum_{i=1}^{N} x_i$ como

$$\hat{t}_x^{(1)} = \sum_{i \in S^{(1)}} w_i^{(1)} x_i = \sum_{i=1}^{N} Z_i w_i^{(1)} x_i.$$

Ahora, seleccionamos la muestra de la segunda etapa y medimos y_i en las unidades de la submuestra. De la muestra de la segunda etapa $\mathcal{S}^{(2)}$, podemos calcular $\hat{t}_y^{(2)}$ usando (12.1) y}

$$\hat{t}_y^{(2)} = \sum_{i \in S^{(2)}} w_i^{(1)} w_i^{(2)} x_i = \sum_{i=1}^{N} Z_i D_i w_i^{(1)} w_i^{(2)} x_i.$$

Entonces,

$$\hat{t}_{yr}^{(2)} = \frac{\hat{t}_x^{(1)} \hat{t}_y^{(2)}}{\hat{t}_x^{(2)}}.$$

Observe que este estimador es muy similar al estimador por proporción en (3.2); usamos $\hat{t}_x^{(1)}$ de la muestra de la primera etapa en vez de la cantidad no conocida t_x.

Si linealizamos,

$$\hat{t}_{yr}^{(2)} \approx t_y + \frac{t_x}{t_x}(\hat{t}_y^{(2)} - t_y) + \frac{t_y}{t_x}(\hat{t}_y^{(1)} - t_x) - \frac{t_y t_x}{t_x^2}(\hat{t}_x^{(2)} - t_x).$$

Entonces,

$$\begin{aligned} V(\hat{t}_{yr}^{(2)}) &\approx V\left[\hat{t}_y^{(2)} + \frac{t_y}{t_x}(\hat{t}_x^{(1)} - \hat{t}_x^{(2)})\right] \\ &= V\left\{E\left[\hat{t}_y^{(2)} + \frac{t_y}{t_x}(\hat{t}_x^{(1)} - \hat{t}_x^{(2)}) \,\middle|\, \mathbf{Z}\right]\right\} + E\left\{V\left[\hat{t}_y^{(2)} + \frac{t_y}{t_x}(\hat{t}_x^{(1)} - \hat{t}_x^{(2)}) \,\middle|\, \mathbf{Z}\right]\right\} \\ &= V[\hat{t}_y^{(1)}] + E\left[V\left(\hat{t}_y^{(2)} - \frac{t_y}{t_x}\hat{t}_x^{(2)} \,\middle|\, \mathbf{Z}\right)\right] \\ &= V[\hat{t}_y^{(1)}] + E[V(\hat{t}_d^{(2)} \mid \mathbf{Z})], \end{aligned}$$

donde $d_i = y_i - (t_y/t_x)x_i$. Así, la varianza del estimador por proporción en dos etapas es la varianza que se calcularía para $\hat{t}_y^{(1)}$ si observáramos y_i para cada unidad en la muestra de

la primera etapa, junto con un término adicional que implica la varianza de los residuales del modelo de proporción. El ejercicio 2 proporciona la varianza y un estimador de la varianza si el diseño de la muestra en ambas etapas es el de una muestra aleatoria simple. Rao y Sitter (1995) y Sitter (1997) dedujeron otros estimadores de la varianza para los estimadores por proporción y regresión en el muestreo en dos etapas.

12.1.3 Muestreo en dos etapas para estratificación

Para mayor sencillez, supongamos que se extrae una muestra aleatoria simple en la primera etapa y que se utiliza un muestreo aleatorio simple para las submuestras de la segunda etapa. (Särndal *et al* 1992 dan un tratamiento más general, permitiendo un muestreo con probabilidades diferentes en cualquiera de las etapas.) Definimos $\mathcal{S}^{(1)}$, $\mathcal{S}^{(2)}$, Z_i y D_i como antes. Si extraemos una muestra aleatoria simple de tamaño n en la primera etapa,

$$P(Z_i = 1) = \frac{n}{N}.$$

Separamos las unidades de observación en H estratos, pero sabemos a qué estrato pertenece cada unidad hasta que se selecciona en la primera etapa. Sin embargo, en la población, el estrato h tiene N_h unidades (no conocemos N_h) y $N = \sum_{h=1}^{H} N_h$ (suponga que conocemos N).
Sea

$$x_{ih} = \begin{cases} 1 & \text{si la unidad } i \text{ está en el estrato } h. \\ 0 & \text{si la unidad } i \text{ no está en el estrato } h. \end{cases}$$

Observe x_{ih}, $h = 1, \ldots, H$ para cada unidad de la muestra de la primera etapa; suponga que extraemos al menos dos unidades de cada estrato. La cantidad de unidades de la muestra de la primera etapa que pertenecen al estrato h es una variable aleatoria:

$$n_h = \sum_{i=1}^{N} Z_i x_{ih}.$$

Ahora, extraiga una submuestra aleatoria simple de tamaño m_h en el estrato h; m_h puede depender de la primera etapa del muestreo. Las submuestras de los distintos estratos se seleccionan de manera independiente, dada la información en la muestra de la primera etapa. Con el submuestreo aleatorio,

$$P(D_i = 1 \mid \mathbf{Z}) = Z_i \sum_{h=1}^{H} x_{ih} \frac{m_h}{n_h}.$$

Aunque $P(D_i = 1 \mid \mathbf{Z})$ se escribe como una suma, todas excepto una de las x_{ih} para $h = 1, \ldots, H$, se anulan, pues cada unidad pertenece exactamente a un estrato. El peso de muestreo para una unidad de la segunda etapa en el estrato h es $w_i^{(2)} = n_h/m_h$; en general, $\omega_i^{(2)} = Z_i \sum_{h=1}^{H} x_{ih} n_h / m_h$.

El estimador estratificado del total de la población para el muestreo en dos etapas es

$$\begin{aligned} \hat{t}_{\text{est}}^{(2)} &= \sum_{i=1}^{N} Z_i D_i w_i^{(1)} w_i^{(2)} y_i \\ &= \sum_{i=1}^{N} \sum_{h=1}^{H} Z_i D_i \frac{N}{n} \frac{n_h}{m_h} x_{ih} y_i \\ &= N \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h^{(2)}, \end{aligned} \tag{12.2}$$

donde $\bar{y}_h^{(2)} = \sum_{i \in S^{(2)}} x_{ih} y_i / m_h$ es el promedio de las unidades de la segunda etapa en el estrato h. El estimador correspondiente de la media de la población es

$$\hat{\bar{y}}_{est}^{(2)} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h^{(2)}. \tag{12.3}$$

Recuerde que un estimador del total de la población para un muestreo aleatorio estratificado, (4.1), es

$$\hat{t}_{\text{est}} = N \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h;$$

el estimador de muestreo en dos etapas simplemente sustituye n_h/n en vez de N_h/N. Como mostramos para el estimador en (12.1), $E[\hat{t}_{\text{est}}^{(2)} \mid \mathbf{Z}] = \hat{t}_y^{(1)}$, y así $E[\hat{t}_{\text{est}}^{(2)}] = t_y$.

De nuevo calculamos la varianza en forma condicional,

$$\begin{aligned} V(\hat{t}_{\text{est}}^{(2)}) &= V\left(E\left[\hat{t}_{\text{est}}^{(2)} \mid \mathbf{Z}\right]\right) + E\left(V\left[\hat{t}_{\text{est}}^{(2)} \mid \mathbf{Z}\right]\right) \\ &= V(\hat{t}_y^{(1)}) + N^2 E\left(V\left[\sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h^{(2)} \mid \mathbf{Z}\right]\right) \\ &= N^2 \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n} + N^2 E\left[\sum_{h=1}^{H} \left(\frac{n_h}{n}\right)^2 \left(1 - \frac{m_h}{n_h}\right) \frac{s_h^{2(1)}}{m_h}\right]. \end{aligned}$$

El primer término es la varianza de la muestra aleatoria simple de la primera etapa; el segundo término es la varianza adicional del submuestreo de la segunda etapa. En este caso, $S_y^2 = \sum_{i=1}^{N} (y_i - \bar{y}_U)^2 / (N-1)$ es la varianza de población de las y;

$$s_h^{2(1)} = \frac{\sum_{i \in S^{(1)}} x_{ih} (y_i - \bar{y}_h^{(1)})^2}{n_h - 1}$$

sería la varianza muestral de las y_i en el estrato h de la muestra de la primera etapa si observáramos todas. La varianza de $\hat{t}_{\text{est}}^{(2)}$ se deja como esperanza, pues n_h y m_h son variables aleatorias.

Rao (1973) proporciona la varianza estimada en el muestreo en dos etapas como

$$\begin{aligned} \hat{V}(\hat{t}_{\text{est}}^{(2)}) = N(N-1) \sum_{h=1}^{H} \left(\frac{n_h - 1}{n-1} - \frac{m_h - 1}{N-1}\right) \frac{n_h}{n} \frac{s_h^{2(2)}}{m_h} \\ + \frac{N^2}{n-1} \left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} \frac{n_h}{n} \left(\bar{y}_h^{(2)} - \hat{\bar{y}}_{\text{est}}^{(2)}\right)^2 \end{aligned} \tag{12.4}$$

donde $s_h^{2(2)}$ es la varianza muestral de las y_i en el estrato h. Si podemos ignorar la corrección para poblaciones finitas,

$$\hat{V}(\hat{\bar{y}}_{\text{est}}^{(2)}) \approx \sum_{h=1}^{H} \frac{n_h - 1}{n-1} \frac{n_h}{n} \frac{s_h^{2(2)}}{m_h} + \frac{1}{n-1} \sum_{h=1}^{H} \frac{n_h}{n} \left(\bar{y}_h^{(2)} - \hat{\bar{y}}_{\text{est}}^{(2)}\right)^2. \tag{12.5}$$

EJEMPLO 12.4 Apliquemos estos resultados a los datos del ejemplo 12.1. Como $\bar{y}_h^{(2)} = \hat{p}_h$ es una proporción, $s_h^{2(2)} = m_h \hat{p}_h (1-\hat{p}_h)/(m_h - 1)$. Las estadísticas de la muestra de la segunda etapa son como sigue:

Estrato	n_h	m_h	$\hat{p}_h$	$s_h^{2(2)}$
Sí	755	67	0.7313	0.1995
No	804	72	0.1528	0.1313
No disponible	505	505	0.4178	0.2437
Total	2064	644		

El porcentaje estimado de pacientes en hospitales de la VA que son veteranos de la guerra de Vietnam y que estuvieron de servicio en allí es, de (12.3),

$$\hat{\bar{y}}_{\text{est}}^{(2)} = \left(\frac{755}{2064}\right)(0.7313) + \left(\frac{804}{2064}\right)(0.1528) + \left(\frac{505}{2064}\right)(0.4178) = 0.4293.$$

La muestra de la primera etapa es una muestra aleatoria simple con $n/N = 0.2$, de modo que debemos incluir la corrección para poblaciones finitas en la estimación de la varianza. Al calcular los términos en (12.4),

$$\sum_{h=1}^{H}\left(\frac{n_h - 1}{n-1} - \frac{m_h - 1}{N-1}\right)\frac{n_h}{n}\frac{s_h^{2(2)}}{m_h} = 0.000391 + 0.000271 + 0.0000231 = 0.000686,$$

y

$$\frac{1}{n-1}\left(1 - \frac{n}{N}\right)\sum_{h=1}^{H}\frac{n_h}{n}\left(\bar{y}_h^{(2)} - \hat{\bar{y}}_{\text{st}}^{(2)}\right)^2$$

$$= (1.29 \times 10^{-5}) + (1.16 \times 10^{-5}) + (1.24 \times 10^{-8}) = 0.0000245.$$

Así, $\hat{V}(\hat{\bar{y}}_{\text{est}}^{(2)}) = 0.000686 + 0.0000245 = 0.00071$, y EE $(\hat{\bar{y}}_{\text{est}}^{(2)}) = 0.027$.

¿Fue más eficiente en este caso el muestreo en dos etapas? De haber extraído una muestra aleatoria simple de tamaño 644 directamente de los registros y haber observado $\hat{p} = 0.429$, el error estándar sería $\text{EE}(\hat{p}) = 0.019$, que en realidad es menor que el error estándar del diseño de muestreo en dos etapas. Si observa los términos individuales de las estimaciones de la varianza, verá por qué el muestreo en dos etapas no aumenta la eficiencia en este ejemplo. Todas las unidades de la primera etapa en el estrato "no disponible" participaron en la submuestra, dando un valor muy bajo a $s_h^{2(2)}/m_h$ para ese estrato, pero los tamaños de muestra en los otros dos estratos fueron demasiado pequeños, lo que condujo a contribuciones relativamente grandes de estos dos estratos a la varianza global.

Suponga que ahora usamos una asignación proporcional en la muestra de la segunda etapa y que observamos las mismas proporciones de la muestra. Entonces, podría extraerse una submuestra de 236 registros en el estrato "si", 251 registros en el estrato "no" y 157 registros en el estrato "no disponible". En ese caso, si las proporciones de la muestra siguen siendo las mismas, el error estándar de la muestra en dos etapas sería 0.017, una modesta disminución del error estándar de una muestra aleatoria simple de tamaño 644. Se podría ahorrar un poco más si se utiliza algún tipo de asignación óptima (véase el ejercicio 5). ■

12.2
Estimación por captura y recaptura

EJEMPLO 12.5 Suponga que queremos estimar N, la cantidad de peces en un lago. Un método es el siguiente: atrapamos y marcamos 200 peces del lago para luego soltarlos. Permitimos que los marcados y liberados se mezclen con los otros peces del lago. Luego, extraemos una segunda muestra independiente de 100 peces. Supongamos que 20 de los animales de la segunda muestra están marcados. Entonces, suponiendo que la población no ha cambiado en el intervalo entre ambas muestras y que cada recolección proporciona una muestra aleatoria simple de peces en el lago, estimamos que 20% de los peces en el lago están marcados y que por lo tanto los 200 marcados en la muestra original representaban 20% de la población. Así, estimamos el tamaño de la población N como 1000 peces aproximadamente. ■

Este método de estimación del tamaño de una población se llama **estimación por captura y recaptura** en dos muestras. Otros nombres son: captura con señas o marcas, método de Petersen (1986) o índice de Lincoln (1930). El método se basa en las siguientes hipótesis:

1. La población es *cerrada*: ningún pez entra o sale del lago en el intervalo entre las muestras. Esto significa que N es la misma para cada muestra.
2. Cada muestra de peces es una muestra aleatoria simple de la población. Esto significa que cada pez tiene la misma probabilidad de inclusión en una muestra; no ocurre, por ejemplo, que los más pequeños o menos saludables tengan más posibilidades de ser capturados. Además, no existen "peces ocultos" en la población, imposibles de atrapar.
3. Las dos muestras son independientes. Los peces marcados de la primera muestra se vuelven a mezclar en la población, de modo que el estado de marcación de un pez no está relacionado con la probabilidad de que se seleccione en la segunda muestra. Además, los peces incluidos en la primera muestra no son "tímidos" ni "felices" ante una red, la probabilidad de que un pez sea atrapado en la segunda muestra no depende de su historia de captura.
4. Los peces no pierden sus marcas y los marcados pueden identificarse como tales. La pintura soluble en agua, por ejemplo, no sería una buena elección para marcar el material.

En esta forma sencilla, la captura y recaptura es un caso particular de la estimación de cocientes del total de una población, y podemos usar los resultados del capítulo 3 cuando las muestras y la población son grandes. Sea n_1 el tamaño de la primera muestra, n_2 el tamaño de la segunda muestra y m la cantidad de peces atrapados en la segunda muestra. En el ejemplo 12.5, $n_1 = 200$, $n_2 = 100$, $m = 20$ y usamos la estimación $\hat{N} = n_1 n_2 / m$. Para ver por qué esta estimación cae en el marco de referencia del capítulo 3, sean

$$y_i = 1 \text{ para cada pez del lago.}$$

$$x_i = \begin{cases} 1 & \text{si el pez } i \text{ está marcado.} \\ 0 & \text{si el pez } i \text{ no está marcado.} \end{cases}$$

Entonces estimamos $N = t_y = \sum_{i=1}^{N} y_i$ como $\hat{t}_{yr} = t_x \hat{B}$, donde $t_x = \sum_{i=1}^{N} x_i = n_1$ y

$\hat{B} = \bar{y}/\bar{x} = n_2/m$. Esta estimación de cocientes,

$$\hat{N} = \hat{t}_{yr} = \frac{n_1 n_2}{m} \tag{12.6}$$

es también la estimación de máxima verosimilitud (consulte los ejercicios 8 y 9). Al aplicar (3.7) a la segunda muestra aleatoria simple e ignorar la corrección para poblaciones finitas,

$$\hat{V}(\hat{N}) = t_x^2 \hat{V}(\hat{B}) = \left(\frac{n_1 n_2}{m}\right)^2 \frac{n_2 - m}{m(n_2 - 1)} \approx \frac{n_1^2 n_2 (n_2 - m)}{m_3}.$$

Para los datos del ejemplo 12.5, $\hat{V}(\hat{N}) = 40{,}000$.

Sin embargo, al ser una estimación de cociente, $\hat{N}$ es sesgado y el sesgo puede ser grande en las aplicaciones a la vida salvaje con tamaños pequeños de muestra. De hecho, es posible que la segunda muestra conste totalmente de animales no marcados, haciendo infinita la estimación en (12.6). Chapman (1951) propone la estimación menos sesgada

$$\tilde{N} = \frac{(n_1 + 1)(n_2 + 1)}{m + 1} - 1. \tag{12.7}$$

Una estimación de la varianza de $\tilde{N}$ (Seber 1970) es

$$\hat{V}(\tilde{N}) = \frac{(n_1 + 1)(n_2 + 1)(n_1 - m)(n_2 - m)}{(m + 1)^2 (m + 2)}. \tag{12.8}$$

Las estimaciones en (12.7) y (12.8) frecuentemente se emplean en aplicaciones a la vida salvaje. Para los datos de los peces, $\tilde{N} = (201)(101)/21 - 1 = 966$ y $\hat{V}(\tilde{N}) = 30{,}131$.

Muchos investigadores han construido intervalos de confianza para el tamaño de la población utilizando

$$\hat{N} \pm 1.96\sqrt{\hat{V}(\hat{N})} \quad \text{o} \quad \tilde{N} \pm 1.96\sqrt{\hat{V}(\tilde{N})}.$$

Sin embargo, esto no es totalmente satisfactorio, pues ambos requieren que $\hat{N}$ o $\tilde{N}$ tengan aproximadamente una distribución normal y la distribución normal podría no ser una buena aproximación a la distribución de $\hat{N}$ o $\tilde{N}$ para poblaciones y muestras pequeñas. Analizaremos los intervalos de confianza en la sección 12.2.2; pero primero revisaremos otro enfoque de los datos que será útil al desarrollar los intervalos de confianza.

12.2.1 Tablas de contingencia para experimentos con captura y recaptura

Fienberg (1972) sugiere que los datos de captura y recaptura se vean en una tabla de contingencias incompleta. Para los datos del ejemplo 12.5, la tabla es la siguiente:

		¿Está en la segunda muestra? Sí	No	
¿Está en la primera muestra?	Sí	20	180	200
	No	80	?	?
		100	?	N

En general, si x_{ij} es la cifra observada en la celda (i,j), la tabla de contingencia se ve así. El asterisco indica que no observamos esa celda.

		¿Está en la segunda muestra? Sí	No	
¿Está en la primera muestra?	Sí	x_{11} $(=m)$	x_{12}	x_{1+} $(=n_1)$
	No	x_{21}	x^*_{22}	x^*_{2+}
		x_{+1} $(=n_2)$	x^*_{+2}	x^*_{++}

Las cifras esperadas son:

		¿Está en la segunda muestra? Sí	No	
¿Está en la primera muestra?	Sí	m_{11}	m_{12}	m_{1+}
	No	m_{21}	m^*_{22}	m^*_{2+}
		m_{+1}	m^*_{+2}	$m^*_{++} = N$

Para estimar las cifras esperadas usaríamos $\hat{m}_{11} = x_{11}, \hat{m}_{12} = x_{12}$, y $\hat{m}_{21} = x_{21}$. Si la presencia en la primera muestra es independiente de la presencia en la muestra 2, entonces las posibilidades de estar en la segunda muestra son las mismas para los peces marcados y los que no: $m_{11}/m_{21} = m_{12}/m_{22}$. En consecuencia, bajo la independencia, la cifra esperada en la celda o pez no incluido en cualquiera de las muestras es

$$\hat{m}_{22} = \frac{\hat{m}_{12}\hat{m}_{21}}{\hat{m}_{11}} = \frac{x_{12}x_{21}}{x_{11}},$$

y

$$\hat{N} = \hat{m}_{11} + \hat{m}_{12} + \hat{m}_{21} + \hat{m}_{22} = \frac{x_{+1}x_{1+}}{x_{11}}.$$

Calculamos la estimación $\hat{N}$ con base en la hipótesis de que las dos muestras son independientes; por desgracia, esa hipótesis no se puede verificar pues sólo se observan tres de las cuatro celdas de la tabla de contingencias.

12.2.2 Intervalos de confianza para N

En muchas aplicaciones de captura y recaptura se han construido intervalos de confianza usando

$$\hat{N} \pm 1.96\sqrt{\hat{V}(\hat{N})} \quad \text{o} \quad \tilde{N} \pm 1.96\sqrt{\hat{V}(\tilde{N})}.$$

Si empleamos el primer intervalo para los datos del ejemplo 12.5, $\hat{V}(\hat{N}) = 40{,}000$, y un intervalo de confianza asintótico de 95% sería $1000 \pm 1.96(200) = [608, 1392]$. El intervalo de confianza usando la distribución normal y $\tilde{N}$ es [626, 1306]. Desafortunadamente, los intervalos de confianza basados en la hipótesis de que $\hat{N}$ o $\tilde{N}$ siguen una distribución normal generalmente tienen poca probabilidad de cobertura en muestras pequeñas, pues la distribución de $\hat{N}$ y $\tilde{N}$ es un poco asimétrica en la realidad, como veremos en el ejercicio 13. En general, desaconsejamos la utilización de estos intervalos de confianza.

Un defecto adicional de los intervalos de confianza basados en la distribución normal puede ocurrir en las muestras pequeñas. Por ejemplo, supongamos que $n_1 = 30, n_2 = 20$,

y $m = 15$. Entonces $\hat{N} = (30)(20)/15 = 40$, y $\hat{V}(\hat{N}) = 26.7$. Al usar una aproximación normal a la distribución de $\hat{N}$ obtenemos el intervalo de confianza [30, 50]. Sin embargo, la cota inferior de 30 es un tanto ingenua; se observó un total de 35 animales distintos en las dos muestras, de modo que sabemos que N debe ser al menos 35.

Cormack (1992) analiza el uso de la prueba de ji cuadrada de Pearson o de razón de verosimilitud para la independencia para construir un intervalo de confianza. Con este método, llenamos la observación faltante x_{22} con algún valor u y realizamos una prueba ji cuadrada para la independencia sobre el conjunto de datos completado artificialmente. El intervalo de confianza de 95% para m_{22} está formado entonces por todos los valores de u para los que no se rechazaría la hipótesis de independencia para las dos muestras, al nivel 0.05. Para los datos del ejemplo 12.5, tratemos con el valor $u = 600$. Con este valor, la tabla de contingencias "completa" es

		¿Está en la segunda muestra?		
		Sí	No	
¿Está en la primera muestra?	Sí	20	180	200
	No	80	600	680
		100	780	880

Podemos realizar fácilmente una prueba ji cuadrada de Pearson para la independencia en esta tabla, obteniendo un valor p de 0.49. Como $0.49 > 0.05$, el valor 600 estaría dentro del intervalo de confianza de 95% para u, y el valor 880 estaría dentro del intervalo de confianza de 95% para N. Sin embargo, si u es igual a 1500, obtenemos el valor $p = 0.0043$, de modo que 1500 queda fuera del intervalo de confianza de 95% para u, y con ello 1780 queda fuera del intervalo de confianza de 95%, para N. Al continuar de esta manera, vemos que los valores de u entre 430 y 1198 son los únicos que producen un valor $p > 0.05$, de modo que [430, 1198] es un intervalo de confianza de 95% para m_{22}. El intervalo de confianza correspondiente para N se obtiene sumando la cantidad de animales observados en las demás celdas, 280, a los extremos del intervalo de confianza para m_{22}, resultando el intervalo [710,1478].

La prueba de razón de verosimilitud se puede usar de manera análoga, incluyendo en el intervalo de confianza a todos los valores de u para los cuales el valor p de la prueba de razón de verosimilitud es mayor que 0.05. Si usamos el código S-PLUS del apéndice D, vemos que los valores de u entre 437 y 1233 dan una razón de verosimilitud mayor que 0.05. Así, el intervalo de confianza para N usando la prueba de razón de verosimilitud es [717,1513].

Otra alternativa para los intervalos de confianza consiste en usar la técnica de bootstrap (Buckland 1984); para aplicarla, volvemos a extraer una muestra en los individuos observados de la segunda muestra. Extraemos R muestras de tamaño 100 con reemplazo, de los 20 peces marcados y los 80 no marcados ya observados. Calculamos $\hat{N}^*$ para cada una de las R nuevas muestras y determinamos los puntos porcentuales 2.5 y 97.5 de los R valores. Con $R = 999$, el intervalo de confianza de 95% fueron los valores 25 y 875 de la lista ordenada de los $\hat{N}^*$,[714, 1538].

Observe que los tres intervalos de confianza, obtenidos mediante la prueba ji cuadrada de Pearson, la prueba ji cuadrada con razón de verosimilitud y la técnica de bootstrap son similares, pero los tres difieren de los intervalos de confianza basados en la normalidad asintótica de $\hat{N}$ o $\tilde{N}$.

12.2.3 Uso de captura y recaptura en listas

La estimación por captura y recaptura no se limita a la estimación de poblaciones de la vida salvaje. También se puede utilizar cuando las dos muestras son listas de individuos, siempre

que se cumplan las hipótesis del método. Suponga que quiere estimar la cantidad de estadísticos en Estados Unidos y que obtiene la lista de miembros de la American Statistical Association (ASA) y del Institute for Mathematical Statistics (IMS). Cada estadístico es miembro o no de ASA, o bien es miembro o no de IMS. (Hay otras organizaciones importantes, pero para simplificar la exposición limitaremos nuestro análisis a estas dos.) Entonces, n_1 es la cantidad de miembros de ASA, n_2 la cantidad de miembros de IMS y m es el número de personas en ambas listas. Podemos estimar la cantidad de estadísticos mediante $\hat{N} = n_1 n_2 / m$, exactamente como si los estadísticos fuesen peces. Las hipótesis de esta estimación son como antes, pero con implicaciones ligeramente distintas al caso de la vida salvaje:

1 La población es cerrada. En los estudios de la vida salvaje, esta hipótesis podría no cumplirse, pues con frecuencia los animales mueren o emigran en el intervalo entre las muestras. Sin embargo, al considerar las listas como las muestras, por lo general podemos actuar como si la población fuese cerrada, si las listas son del mismo periodo.

2 Cada lista proporciona una muestra aleatoria simple de la población de estadísticos. Esta hipótesis es un poco más problemática: implica que la probabilidad de pertenecer a ASA es la misma para todos los estadísticos, al igual que pertenecer a IMS. No admite la posibilidad de que un grupo se rehúse a pertenecer a cualquiera de estas organizaciones o la posibilidad de que ciertos subgrupos tengan distintas probabilidades de pertenecer a una organización.

3 Las dos listas son independientes. En este caso, esto quiere decir que la probabilidad de que un estadístico esté en ASA no depende de su membresía en IMS. Frecuentemente, esta hipótesis tampoco se cumple; tal vez los estadísticos tienden a ingresar sólo a una organización y por lo tanto los miembros de ASA tienen menos probabilidad de pertenecer a IMS que quienes no lo son.

4 Los individuos pueden hacerse corresponder en las listas. Esto parece fácil, pero muchas veces es asombrosamente difícil. ¿Es J. Smith de la primera lista la misma persona que Jonquil Smith en la segunda?

EJEMPLO 12.6 La oficina de censos de Estados Unidos trata de enumerar tantas personas como sea posible en su censo decenal. Sin embargo, la omisión de personas es inevitable, por lo que las estimaciones de la población a partir del censo sea una subestimación de las cifras reales de la población. Además, se supone que la subcobertura no es uniforme, probablemente es mayor en las áreas urbanas del interior y en las minorías y varía entre las diversas regiones de Estados Unidos. Debido a que los representantes en el Congreso, los miles de millones de dólares de fondos federales y otros recursos se proporcionan con base en los resultados de los censos, muchos gobiernos estatales y locales se preocupan de que las cifras sean precisas. La estimación por captura y recaptura, en este contexto conocida como **estimación con sistema dual**, se ha utilizado desde 1950 para evaluar la cobertura de los censos decenales. En los últimos años se ha desarrollado una controversia considerable, llegando hasta los juzgados, sobre el uso de estos métodos para ajustar las estimaciones de la población a partir del censo. Fienberg (1992) proporciona bibliografía para la estimación con sistema dual; los artículos del número de noviembre de 1994 de *Statistical Science* analizan esta controversia.

Hogan (1993) describe la encuesta de enumeración posterior (PES) de 1990 utilizada por la oficina de censos. En Canadá se utiliza un procedimiento similar, llamado verificación inversa de registros. Se extraen dos muestras; la muestra P se extrae directamente de la

población, de manera independiente al censo, y se emplea para estimar el número de personas omitidas por el censo. La muestra E se extrae de la enumeración del propio censo y se usa para estimar errores del censo, como las personas no existentes o duplicadas.

Se obtienen estimaciones de la población por separado para cada uno de los 1392 estratos posteriores, donde la población se estratifica posteriormente por región, raza, propiedad de su hogar, edad, y otras variables. Se usan los estratos posteriores porque se espera que la hipótesis 2 (probabilidades iguales de recaptura) se satisfaga aproximadamente dentro de cada estrato posterior; sabemos que esto no se logra considerando la población como un todo debido a las tasas diferenciales de subconteo en el censo. La tabla de la población para un estrato posterior es la siguiente:

		¿Está en la enumeración del censo? Sí	No	
¿Está en PES?	Sí	N_{11}	N_{12}	N_{1+}
	No	N_{21}	N^*_{22}	N^*_{2+}
		N_{+1}	N^*_{+2}	N

Se usa la enumeración del censo, la muestra P y la muestra E para llenar las celdas de la tabla. Entonces

$$\hat{N} = \frac{\hat{N}_{+1}\hat{N}_{1+}}{\hat{N}_{11}}.$$

Las cantidades $\hat{N}_{+1}$ y $\hat{N}_{11}$ son estimaciones de la muestra P; $\hat{N}_{1+}$ es la estimación del total del estrato posterior, usando pesos, a partir de la muestra P y $\hat{N}_{11}$ es una estimación ponderada de las concordancias entre la muestra P y la enumeración del censo. En este caso, $\hat{N}_{+1}$ no es la cifra real del censo, sino la cifra ajustada mediante la muestra E para eliminar las personas duplicadas y ficticias. Muchos tamaños de muestra de los estratos posteriores fueron pequeños, lo que implica grandes varianzas para las estimaciones de las cifras de la población, de modo que las estimaciones se suavizaron y ajustaron mediante modelos de regresión.

Las anteriores hipótesis deben cumplirse para la estimación con sistema dual para dar una estimación de la población mejor que los datos originales del censo. Se espera que la hipótesis 2 se cumpla dentro de los estratos posteriores. Sin embargo, la hipótesis 3 también implica cierta preocupación, ya que la muestra P también muestra ausencia de respuesta. Freedman y Navidi (1992) y Breiman (1994) analizan este problema, así como las preocupaciones en torno del ajuste de las estimaciones por regresión. Otra preocupación es la capacidad para hacer corresponder las personas de la muestra P con las personas del censo. Como es de suponer que los individuos de la muestra P no correspondientes se han omitido en el censo, los errores en el proceso de correspondencia en las dos muestras pueden conducir a sesgos en las estimaciones de la población. Ding y Fienberg (1994; 1996) deducen modelos para concordar los errores en la estimación con sistema dual.

El debate acerca del uso del muestreo para mejorar la precisión de las cifras de un censo continúa. Para el censo del año 2000, un comité de la Academia Nacional de Ciencias recomendó la enumeración de la población en cada condado hasta alcanzar una tasa de respuesta de 90%, para luego extraer una muestra del 10% restante. Sin embargo, una solicitud presentada al Congreso norteamericano prohibiría el uso de cualquier fondo "para planear o preparar de cualquier forma el uso del muestreo al realizar el censo del decenio 2000". ■

12.2.4 Estimación con varias recapturas

Las hipótesis para la estimación por captura y recaptura con dos muestras descrita anteriormente son fuertes: la población debe ser cerrada y las dos muestras aleatorias independientes. Además, estas hipótesis no se pueden probar, pues sólo observamos tres de las cuatro celdas de la tabla de contingencias; necesitamos las cuatro celdas para probar la independencia de las muestras.

Se pueden ajustar modelos más complicados si se extraen $K > 2$ muestras aleatorias y en particular si se utilizan diferentes marcas para los individuos atrapados en las diversas muestras. Con los peces, por ejemplo, podemos marcar la aleta pectoral izquierda para los animales atrapados en la primera muestra, la aleta pectoral derecha para los capturados en la segunda y una aleta dorsal para los atrapados en la tercera muestra. Sabríamos entonces que un pez capturado en la muestra 4 con marcas en la aleta pectoral izquierda y en la aleta dorsal habría sido capturado en las muestras 1 y 3, pero no en la 2.

Schnabel (1938) analizó primero la forma de estimar N al extraer K muestras y determinó que la estimación de máxima verosimilitud de N es la solución de

$$\sum_{i=1}^{K} \frac{(n_i - r_i)M_i}{N - M_i} = \sum_{i=1}^{K} r_i,$$

donde n_i es el tamaño de la muestra i, r_i es la cantidad de peces recapturados en la muestra i y M_i es el número de peces marcados al extraer la muestra i.

Si se usan marcas individuales, podemos también analizar aspectos de la inmigración y la emigración de la población y probar algunas de las hipótesis de independencia.

EJEMPLO 12.7 Domingo-Salvany *et al* (1995) usaron la captura y la recaptura para estimar la frecuencia de adicción al opio en Barcelona, España. Uno de sus conjuntos de datos consistía de tres muestras de 1989: (1) una lista de adictos al opio de las salas de urgencia (lista E); (2) una lista de personas que comenzaron un tratamiento contra la adicción al opio durante 1989, reportada al sistema de información sobre uso de drogas en Cataluña (lista T); (3) una lista de muertes por sobredosis de heroína registradas por el instituto forense en 1989 (lista D). Había un total de 2864 personas distintas en las tres listas. Las integrantes de las tres se compararon con los siguientes resultados:

		¿Está en la lista D? Sí		No	
		¿Está en la lista T? Sí	No	¿Está en la lista T? Sí	No
¿Está en la lista E?	Sí	6	27	314	1728
	No	8	69	712	?

No es claro que estos datos cumplan las hipótesis para el método de captura y recaptura con dos muestras. La hipótesis de independencia entre las muestras puede no cumplirse; si el tratamiento es útil, es menos probable que las personas tratadas aparezcan en una de las otras dos muestras. Además, es mucho menos probable que las personas de la lista de muertos aparezcan en las otras listas; la hipótesis de la población cerrada tampoco se cumple, pues una de las muestras es una lista de fallecidos. Sin embargo, un análisis utilizando las hipótesis sin que se cumplan perfectamente pueden proporcionar cierta información acerca

de la cantidad de adictos al opio. Como hay más de dos muestras, podemos evaluar las hipótesis de independencia mediante modelos loglineales. Sin embargo, hay una hipótesis que *nunca* podemos probar: la celda faltante sigue el mismo modelo que el resto de los datos. ■

Si se extraen tres muestras, las cifras esperadas son:

		¿Está en la muestra 3? Sí: ¿Está en la muestra 2? Sí	Sí: ¿Está en la muestra 2? No	No: ¿Está en la muestra 2? Sí	No: ¿Está en la muestra 2? No
¿Está en la muestra 1?	Sí	m_{111}	m_{121}	m_{112}	m_{122}
	No	m_{211}	m_{221}	m_{212}	$m^{\bullet}_{22}$

En la sección 10.4 analizamos los modelos loglineales. El modelo saturado para tres muestras es:

$$\ln m_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk},$$

Sin embargo, no podemos ajustar este modelo, pues requiere ocho grados de libertad y sólo tenemos siete celdas. Podemos ajustar los siguientes modelos, donde α se refiere a la lista E, β se refiere a la lista T y γ se refiere a la lista D.

1 *Completa independencia.*

$$\ln m_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k.$$

Este modelo implica que la presencia en cualquiera de las listas es independiente de la presencia en cualquiera de las demás listas. El modelo de independencia siempre debe adoptarse en la captura y recaptura con dos muestras.

2 *Una lista es independiente de las otras dos.*

$$\ln m_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}.$$

La presencia en la lista E se relaciona con la probabilidad de que alguien esté en la lista T, pero la presencia en la lista D es independiente de la presencia en otras listas. Hay tres versiones de este modelo; las otras dos sustituyen $(\alpha\gamma)_{ik}$ o $(\beta\gamma)_{jk}$ en vez de $(\alpha\beta)_{ij}$.

3 *Dos muestras son independientes dada la tercera.*

$$\ln m_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik}.$$

Hay tres modelos de este tipo; los otros dos sustituyen $(\alpha\beta)_{ij} + (\beta\gamma)_{ik}$ o $(\alpha\gamma)_{ij} + (\beta\gamma)_{ik}$ en vez de $(\alpha\beta)_{ij} + (\alpha\gamma)_{ik}$. La presencia en las listas de muertos o de tratamiento son condicionalmente independientes dado el estado en la lista E; una vez que sabemos que alguien está en la lista de la sala de urgencias, el hecho de saber si está en la lista de muertos no proporciona información adicional acerca de la probabilidad de que esté en la lista de tratamiento.

4 *Todas las interacciones son de dos sentidos.*

$$\ln m_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}.$$

Este modelo siempre se ajusta perfectamente a los datos: tiene la misma cantidad de parámetros como celdas hay en la tabla de contingencia.

Por desgracia, en ninguno de estos modelos podemos probar la hipótesis de que la celda faltante siga el modelo, pero al menos podemos examinar las hipótesis de independencia por parejas entre las muestras. Para los datos de adicción se ajustaron los siguientes modelos loglineales a partir de los datos, usando la función glim de S-PLUS (se puede usar cualquier programa con un modelo loglineal y que determine estimaciones mediante máxima verosimilitud):

Modelo	G^2	grados de libertad	Valor p	$\hat{m}_{222}$	$\hat{N}$	Intervalo de confianza del 95%
1 Independencia	1.80	3	0.62	3,967	6,831	[6,322, 7,407]
2a E*T	1.09	2	0.58	4,634	7,499	[5,992, 9,706]
2b E*D	1.79	2	0.41	3,959	6,823	[6,296, 7,425]
2c T*D	1.21	2	0.55	3,929	6,793	[6,283, 7,373]
3a E*T, E*D	0.19	1	0.67	6,141	9,005	[5,921, 16,445]
3b E*T, T*D	0.92	1	0.34	4,416	7,280	[5,687, 9,820]
3c E*DT*D	1.20	1	0.27	3,918	6,782	[6,253, 7,388]
4 E*T, E*D, T*D	—	0	—	7,510	10,374	[4,941, 25,964]

En este caso, G^2 es la estadística de la prueba de razón de verosimilitud para ese modelo. De manera un tanto asombrosa, el modelo de independencia se ajusta bien a los datos. Las cifras esperadas por celda bajo el modelo 1, completa independencia, son las siguientes:

		¿Está en la lista D? Sí: ¿Está en la lista T? Sí	Sí: ¿Está en la lista T? No	No: ¿Está en la lista T? Sí	No: ¿Está en la lista T? No
¿Está en la lista E?	Sí	5.1	28.3	310.8	1730.7
	No	11.7	64.9	712.4	3966.7

Estas cifras predichas por celda conducen a la estimación

$$\hat{N} = 2864 + 3967 = 6831$$

si se adopta el modelo de independencia. Análogamente se pueden calcular los valores de $\hat{N}$ para los otros modelos; estimando el valor de la celda faltante a partir del modelo y sumando esa estimación al total conocido para las demás celdas, 2864.

Podemos usar una prueba inversa de razón de verosimilitud (Cormack 1992) para construir un intervalo de confianza para N, usando cualquiera de los modelos. Un intervalo de confianza de 95% para la celda faltante consta de aquellos valores u para los que no se rechazaría una prueba de hipótesis de nivel 0.05 de H_0: $m_{222} = u$ para el modelo loglineal adoptado. Sea $G^2(u)$ la estadística de la prueba de razón de verosimilitud (desviación) para la tabla completa, con u sustituida en la celda omitida; sea t el total de las siete celdas observadas y sea $\hat{u}$ la estimación de la celda faltante usando ese modelo loglineal. Cormack muestra que el conjunto

$$\left\{u : G^2(u) - G^2(\hat{u}) + \log\left(\frac{u}{t+u}\right) - \log\left(\frac{\hat{u}}{t+\hat{u}}\right) < q_1(\alpha)\right\}$$

(donde $q_1(\alpha)$ es el percentil de la distribución χ^2_1 con área α en la cola derecha) es un intervalo de confianza aproximado del $100(1-\alpha)\%$ para m_{222}. En el apéndice D aparece una función S-PLUS para calcular el intervalo de confianza de Cormack. Este intervalo de confianza es adicional sobre el modelo seleccionado y no incluye la incertidumbre asociada con

la elección del modelo. Cormack también analiza la extensión de la prueba ji cuadrada invertida de Pearson para la bondad de ajuste, que produce un intervalo similar. Buckland y Garthwaite (1991) analizan el uso de la técnica de bootstrap para determinar intervalos de confianza para la recaptura múltiple usando modelos loglineales que incorporan el procedimiento de selección del modelo en cada iteracción de la técnica de bootstrap.

Para estos datos, la estimación puntual y el intervalo de confianza parecen basarse fuertemente en el ajuste del modelo particular, aunque todos parecen ajustarse a las celdas observadas. Note que la estimación $\hat{N}$ es mayor y los intervalos de confianza son más anchos para los modelos que incluyen la interacción E*T, aunque esa interacción no es significativa desde el punto de vista estadístico. De alguna manera asombra el buen ajuste del modelo de independencia, pues no sería de esperar que se cumplan las hipótesis de independencia. Además, la población no es cerrada, pero tenemos poca información acerca de la migración hacia adentro o afuera de ella.

En esta sección sólo hemos presentado una introducción a la estimación del tamaño de la población, bajo la hipótesis de que la población es cerrada. También se ha realizado mucha investigación acerca de la estimación bajo captura y recaptura, incluyendo modelos para poblaciones con nacimientos, muertes y migraciones; unas fuentes recomendables para lecturas posteriores son Seber (1982), Pollock (1991) y el artículo panorámico del Grupo de Trabajo Internacional para el Estudio y Predicción de Enfermedades (International Working Group for Disease Monitoring and Forecasting, 1995).

12.3 Revisión de la estimación en dominios

12.3.1 Medias de dominios en encuestas complejas

En la mayoría de las encuestas no sólo se desea obtener estimaciones para la población como un todo, sino también para subpoblaciones, considerados como **dominios** en el muestreo de encuestas. En la sección 3.3 analizamos la estimación en las subpoblaciones para las muestras aleatorias simples y mostramos que la estimación de medias en dominios era un caso particular de la estimación por proporciones porque el tamaño de muestra en el dominio varía de una muestra a otra, pero observamos que si el tamaño de muestra para el dominio en una muestra aleatoria simple era lo bastante grande, podríamos actuar esencialmente como si su tamaño fuese fijo para las inferencias respecto de la media del dominio.

La cuestión no es tan sencilla en las encuestas complejas con muchos dominios. Una preocupación importante es que el tamaño de muestra para un dominio dado sea demasiado pequeño como para proporcionar una estimación útil. El investigador que utilice la Encuesta Nacional a Víctimas de Delitos (NCVS) para estimar las tasas de incidencia para los grupos raza × género por separado en cada estado encontrará varias celdas vacías, aun con una muestra de 90,000 personas. Además, aunque el dominio no esté completamente vacío, en una encuesta compleja es posible que algunas unidades primarias, e incluso algunos estratos, no contengan elementos del dominio, de modo que las estimaciones de la varianza deben calcularse cuidadosamente.

Sea y_i la variable de interés y sea

$$x_{id} = \begin{cases} 1 & \text{si la unidad de observación } i \text{ está en el dominio } d. \\ 0 & \text{si la unidad de observación } i \text{ no está en el dominio } d. \end{cases}$$

Entonces, usando la teoría desarrollada en este libro, estimamos el total de la población para el dominio d como

$$\hat{t}_d = \sum_{i \in S} w_i x_{id} y_i$$

y la media de la población para el dominio d, suponiendo que la muestra tiene ciertas observaciones en el dominio d, como

$$\bar{\hat{y}}_d = \frac{\hat{t}_d}{\sum_{i \in S} w_i x_{id}}.$$

Como $\bar{\hat{y}}_d$ es un cociente, estimamos la varianza usando la linealización (vea el ejemplo 9.2) como

$$\hat{V}(\bar{\hat{y}}_d) = \frac{1}{\hat{N}_d^2} \hat{V}\left[\sum_{i \in S} \omega_i x_{id}(y_i - \bar{\hat{y}}_d)\right] \qquad \textbf{(12.9)}$$

El tamaño de la muestra en el dominio d debe ser grande para que la varianza de linealización sea precisa.

Como analizamos en el capítulo 3, si ignoramos la corrección para poblaciones finitas y extraemos una muestra aleatoria simple, (12.9) implica

$$\hat{V}(\bar{\hat{y}}_d) \approx \frac{s_d^2}{n_d},$$

donde n_d es la cantidad de observaciones de la muestra en el dominio d y s_d^2 es la varianza muestral para las observaciones muestrales en el dominio d.

Advertencia En una muestra aleatoria simple, si se crea un nuevo conjunto de datos que conste sólo de observaciones de la muestra en el dominio d y luego aplica la fórmula común para la varianza, su estimación estará aproximadamente insesgada. *No* adopte este método para estimar la varianza de medias de dominios en encuestas complejas. Es común que una unidad primaria de una muestra no contenga observaciones en el dominio d; si usted elimina tal unidad primaria y luego aplica la fórmula común de la varianza, probablemente la subestimará.

A veces, al usar tablas o archivos de datos públicos, no se pueden calcular los errores estándar de cada dominio, pues se carece de información suficiente acerca del diseño de la muestra. Una posible solución consiste en multiplicar el error estándar bajo el muestreo aleatorio simple por $\sqrt{\text{ed}}$ (efecto de diseño) para la media global. Como observan Kish y Frankel (1974), a menudo este método puede sobrestimar el error estándar, ya que el efecto de los conglomerados puede reducirse dentro del dominio. Para dominios pequeños, y en especial para las diferencias, los efectos de diseño tienden a 1.

12.3.2 Estimación de áreas pequeñas

En el análisis anterior usamos la linealización para aproximar la varianza del cociente $\hat{t}_d / \hat{N}_d$. La validez de esta aproximación depende de contar con un tamaño de muestra suficientemente grande en el dominio. En la práctica, el tamaño de muestra en el dominio d puede ser demasiado pequeño y la varianza de $\bar{\hat{y}}_d$ demasiado grande. Algunos dominios de interés pueden no contar con observación alguna.

Muchas encuestas oficiales grandes proporcionan estimaciones muy precisas de todo el país. Por ejemplo, la NCVS brinda información confiable acerca de la incidencia de distintos tipos de víctimas en Estados Unidos. Sin embargo, si está interesado en estimaciones de

tasas de delitos violentos en el nivel estatal para utilizarlas en la distribución de fondos federales para más oficiales de policía, los tamaños de muestra para algunos estados son tan pequeños que las estimaciones directas de la tasa de delitos violentos tienen poca utilidad. Sin embargo, podría pensarse que las tasas de delitos son tan similares en los estados vecinos con características parecidas y usar la información de otros estados para mejorar la estimación de la tasa de crímenes violentos para el estado con tamaño de muestra pequeño. También podría incorporar la información sobre tasa de criminalidad de otras fuentes, como las estadísticas de la policía, para mejorar su estimación.

Análogamente, los datos de la Evaluación Nacional de Avances Educativos (NAEP; vea el ejemplo 11.7) reunidos para estudiantes de Nueva York podrían bastar para estimar los avances en matemáticas de octavo grado para los estudiantes del estado, pero no para una evaluación directa de los logros en matemáticas en las ciudades individuales, como Rochester. Sin embargo, los datos de encuesta de Rochester se pueden combinar con estimaciones de otras ciudades y con datos administrativos escolares (por ejemplo, calificaciones en otros exámenes estandarizados o información acerca de la enseñanza de las matemáticas en las escuelas) para producir una estimación de los logros en matemáticas de octavo grado en Rochester, la cual esperamos que tenga un menor error cuadrático medio.

Las técnicas de estimación de áreas pequeñas, en las que se obtienen estimaciones para dominios con tamaños de muestra pequeños, han sido un centro de atención reciente e investigación intensa en estadística. Se han propuesto varias técnicas; Ghosh y Rao (1994) dan una descripción detallada de ellas y una bibliografía para lecturas posteriores. Aquí resumimos algunos de los métodos propuestos. En lo sucesivo, las cantidades de interés son los totales por dominio t_d, para $d = 1, \ldots, D$; las variables indicativas de la membresía en el dominio d son x_{id}, como definimos anteriormente.

1 *Estimadores directos.* Un estimador directo de t_d depende sólo de las observaciones de la muestra en el dominio d; como ya hemos señalado,

$$\hat{t}_d(\text{dir}) = \sum_{i \in S} w_i x_{id} y_i.$$

Este estimador directo es insesgado, pero el tamaño de muestra pequeño puede conducir a una varianza grande inaceptable (¡en particular si el dominio d no tiene observaciones en la muestra!).

2 *Estimadores sintéticos.* Suponga que tenemos cierta cantidad asociada con t_d para cada dominio d. Para estimar las tasas de víctimas de crímenes violentos, podríamos usar u_d, la cantidad total de crímenes violentos en el dominio d obtenida a partir de reportes de la policía. Entonces, si los cocientes t_d/u_d son similares en dominios diferentes y si cada cociente es similar al cociente de los totales de población t_y/t_u, entonces una forma simple de estimador sintético

$$\hat{t}_d(\text{sin}) = \left(\frac{\hat{t}_y}{\hat{t}_u}\right) u_d$$

puede ser más preciso que $\hat{t}_d(\text{dir})$. Ciertamente, la varianza de $\hat{t}_d(\text{sin})$ será relativamente pequeña, ya que $(\hat{t}_y/\hat{t}_u)$ se estima de toda la muestra y se espera que sea precisa. Sin embargo, si los cocientes no son homogéneos (por ejemplo, si la proporción de víctimas de delitos violentos reportados a la policía varía mucho de un dominio a otro), entonces el estimador sintético puede tener un sesgo grande.

Se puede emplear la estimación sintética en subconjuntos de la población y luego combinar los estimadores sintéticos para cada subconjunto. Para estimar la cantidad de víctimas de delitos violentos en áreas pequeñas, podría dividir la población en diferentes clases edad/raza/género. Luego, determine una estimación sintética de la cantidad total de víctimas de delitos violentos en el dominio d para cada clase edad/raza/género y sumamos las estimaciones para las clases edad/raza/género para estimar la cantidad total de víctimas de delitos violentos en un área pequeña d. Se espera que los cocientes (cantidad de víctimas de delitos violentos en el dominio d para la clase edad/raza/género c de la NCVS)/(cantidad de víctimas de delitos violentos en el dominio d para la clase edad/raza/género c a partir de los informes de la policía) son más homogéneos que los cocientes t_d/u_d.

3 *Estimadores compuestos.* El estimador directo es insesgado pero tiene varianza grande; el estimador sintético tiene menor varianza, pero puede tener un sesgo grande. Éstos se pueden combinar para formar un estimador compuesto:

$$\hat{t}_d(\text{comp}) = \alpha_d \hat{t}_d(\text{dir}) + (1 - \alpha_d)\hat{t}_d(\text{sin})$$

para $0 \le \alpha_d \le 1$. Es difícil estimar los pesos óptimos relativos α_d, pero una posible solución tiene α_d relacionado con el tamaño de muestra en el dominio d. Entonces, si se observan demasiado pocas unidades en el dominio d, α_d será cercano a cero y se confiará más en el estimador sintético.

4 *Estimadores basados en el modelo.* En un enfoque basado en el modelo se usa una superpoblación para predecir los valores del dominio d. Con frecuencia, el modelo "pide prestada la fuerza" de los datos en dominios íntimamente relacionados o incorpora información auxiliar de datos administrativos o de otras encuestas.

Los modelos mixtos, descritos en la sección 11.4, se usan con frecuencia en la estimación de áreas pequeñas. En la NAEP, si Y_{jd} es el avance en matemáticas del estudiante j en el dominio d en la población, se podría postular un modelo como

$$Y_{jd} = \beta_{0d} + (u_{jd} - \hat{u}_d)\beta_1 + \varepsilon_{jd},$$

donde $\beta_{0d} = \beta_0 + z_d\gamma_0 + \delta_{0d}$, las ε_{jd} son variables aleatorias independientes con media 0 y varianza σ^2, las δ_{0d} son variables aleatorias independientes con media 0 y varianza σ_δ^2, y ε_{jd} y δ_{0d} son independientes entre sí. El covariado en el nivel de estudiante u_{jd} (sólo usamos un covariado para mayor sencillez, pero por supuesto se podrían incluir varios covariados) podría provenir de registros administrativos; por ejemplo, la calificación del alumno en un examen de conocimientos aplicado a todos los estudiantes del estado o sus calificaciones en las clases de matemáticas. Por ejemplo, un covariado en el nivel de dominios z_d sería una evaluación del nivel socioeconómico del dominio o una variable relacionada con los métodos de la enseñanza en el dominio. El enfoque del modelo mixto permite que la estimación para el dominio d pidan prestada la fuerza de otros dominios a través del modelo para β_{0d}; suponemos una ecuación común de regresión para predecir el avance promedio en el dominio d, y todos los dominios del área de interés contribuyan a estimar los parámetros en esa ecuación de regresión. Del mismo modo, en este ejemplo, todos los estudiantes de la muestra en el área de interés contribuyen a la estimación de β_1.

La estimación indirecta (ya sea sintética, compuesta o basada en el modelo) es esencialmente un ejercicio para predecir los datos faltantes. Así, los estimadores indirectos son muy dependientes del modelo usado para predecir los datos faltantes (por ejemplo, el estimador indirecto supone que los cocientes son homogéneos a través de los dominios. De ser posible, hay que verificar empíricamente las hipótesis del modelo; un método para ello es pretender que algunos de los datos a la mano en realidad no están

disponibles y comparar el estimador indirecto con el estimador directo calculado con todos los datos.

12.4 Muestreo para eventos raros

En ocasiones quisiéramos investigar características de una población que son difíciles de hallar y que están dispersas ampliamente en la población objetivo. Por ejemplo, relativamente pocas personas son víctimas de delitos violentos en un año dado, pero tal vez desee obtener información acerca de la población de víctimas de delitos violentos. En una encuesta de epidemiología, tal vez se quiera estimar la incidencia de una enfermedad rara y asegurarse de tener los casos suficientes de la enfermedad en la muestra para analizar las diferencias de las personas sin la enfermedad.

Por supuesto, una posibilidad es extraer una muestra muy grande. Esto se hace en la NCVS, que se emplea para estimar las tasas de incidencia. Como pretendía estimar las tasas de cantidades de víctimas para muchos tipos diferentes e investigar las experiencias de las familias con el paso del tiempo, la NCVS se diseñó para ser aproximadamente autoponderada. Sin embargo, si se está interesado en las víctimas de violencia familiar, el tamaño de muestra es muy pequeño. La NCVS tendría que ser prohibitivamente cara para seguir siendo una encuesta autoponderada y proporcionar tamaños de muestra suficientes para todos los tipos de víctimas de delitos.

Se han propuesto varios métodos para poder estimar la frecuencia de la característica rara y estimar cantidades de interés para las poblaciones raras. Muchas de estas ideas se analizan en Kalton y Anderson (1986) y varios se basan en conceptos ya analizados en este libro. Describiremos en forma breve algunos de estos métodos para tener una idea general de lo disponible y dónde poder aprender más.

12.4.1 Muestreo estratificado con asignación no proporcional

En ocasiones, los estratos se pueden construir de modo que la característica rara fuese más frecuente en uno de los estratos (digamos, en el primer estrato). Entonces, una muestra estratificada en la que la fracción de muestreo sea mayor en el primer estrato puede brindar una estimación más precisa de la frecuencia de la característica rara en la población general. La fracción mayor de muestreo en el estrato 1 también aumenta el tamaño de muestra del dominio o los miembros de la población con la característica rara. La Encuesta Nacional de Salud Materno e infantil (MIHS), analizada en el ejemplo 11.1, extrae una muestra con una fracción mayor de registros de bebés con bajo peso al nacer para garantizar un tamaño de muestra adecuado.

El muestreo estratificado no proporcional puede funcionar bien cuando la asignación es eficiente para todos los puntos de interés. Por ejemplo, en la MIHS, una preocupación fundamental eran los bebés con bajo peso al nacer, quienes tienen mucho más problemas de salud, pero la estratificación no proporcional puede no ser útil para todos los elementos de interés de otras encuestas. Un diseño donde las ciudades de Nueva York y San Francisco participen con exceso en la muestra es sensible a la estimación de la frecuencia de SIDA y obtención de personas afectadas, ya que se supone que estas ciudades tienen la

máxima frecuencia de la enfermedad en Estados Unidos; el diseño no sería tan eficiente para estimar la frecuencia del mal de Alzheimer, que es rara, pero no está concentrada en tales ciudades.

12.4.2 Muestreo en dos etapas

Revise las unidades de la muestra de la primera etapa para determinar si tienen o no la característica rara. Luego extraiga una submuestra de todas (o una alta fracción de muestreo) las unidades con esa característica para la muestra de la segunda etapa. Si la técnica de revisión es completamente precisa, use la muestra de la primera etapa para estimar la frecuencia de la característica rara y la muestra de la segunda etapa para estimar otras cantidades de esa población.

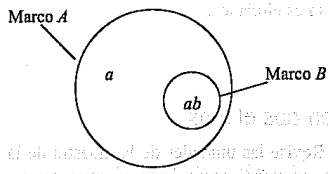
¿Qué ocurre si la técnica de revisión no es completamente precisa? Al extraer una muestra de regiones árticas en búsqueda de morsas, es probable que desde el aire no vea morsas en algunos de los sectores, debido a que los animales están debajo del hielo. Preguntar a las personas si tienen diabetes no siempre produce una respuesta precisa, pues no siempre saben si tienen la enfermedad. Como señala Deming (1977), colocar a una persona con diabetes en el estrato "sin diabetes" es más serio que colocar a una persona sin diabetes en el estrato "diabetes": si se extrae una submuestra sólo del estrato "diabetes", es probable que se descubran las personas sin diabetes colocadas erróneamente en ese estrato, mientras que el error de los diabéticos clasificados erróneamente en el estrato "sin diabetes" no se descubrirá. Una posible solución es ampliar el criterio de revisión de modo que abarque todas las unidades que podrían tener la característica rara; otra solución consiste en extraer una submuestra de ambos estratos en la segunda etapa, pero usando una fracción de muestreo mucho mayor en el estrato "probablemente con diabetes".

12.4.3 Encuestas con marcos múltiples

Aunque es probable que no tenga una lista de todos los miembros de la población rara, tal vez posea algunos marcos de muestreo incompletos que contengan un alto porcentaje de unidades con la característica rara. A veces puede combinar estos marcos incompletos, omitiendo los duplicados, para construir un marco de muestreo completo para la población. Alternativamente, se pueden seleccionar muestras independientes en los marcos y luego combinar las estimaciones de las muestras de los marcos incompletos (y tal vez de un marco completo) para obtener estimaciones globales de la población. Esta idea fue explorada por vez primera por Hartley (1962).

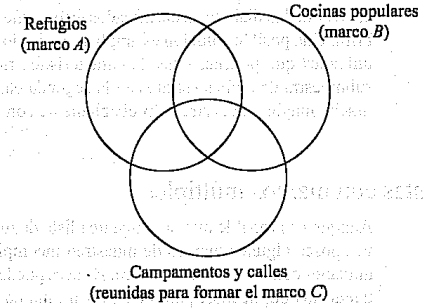
Por ejemplo, suponga que quiere estimar la frecuencia del mal Alzheimer en la población no institucionalizada. Como muchos usuarios de los centros de atención a adultos padecen el mal, sería de esperar que una muestra de tales centros proporcione un mayor porcentaje de personas con mal de Alzheimer que una encuesta general entre la población, pero no todas las personas afectadas asisten a un centro de atención a adultos. Así, se podrían tener dos marcos de muestreo: el marco A, que es el marco de muestreo para la encuesta general de la población, y el marco B, que es el marco de muestreo de los centros de atención a los adultos. Como se supone que todas las personas abarcadas en el marco B deberían estar en el marco de la encuesta general, existen dos dominios: ab, que consta de las perso-

nas en el marco A y en el marco B, y a, que consta de personas en el marco A pero no en el B.



Al realizar la encuesta, determine si cada persona de la muestra bajo el marco A está también en el marco B. Luego estime el total de la población mediante $\hat{t}_a + \hat{t}_{ab}$, donde $\hat{t}_a$ es una estimación del total en el dominio a y $\hat{t}_{ab}$ es una estimación del total en el dominio ab. Se pueden usar varias estimaciones para los totales de los dos dominios; Skinner y Rao (1996) describen algunos de ellos.

Iachan y Dennis (1993) describen el uso de marcos múltiples para obtener una muestra de la población carente de hogar en Washington, D.C. se usaron cuatro marcos: (1) los refugios para los desamparados, (2) las cocinas populares, (3) los campamentos establecidos en edificios desocupados y debajo de los puentes y (4) las calles, con muestras tomadas por manzanas. En teoría, los cuatro marcos deben capturar gran parte de la población sin hogar, aunque las personas desamparadas son móviles y algunas podrían esconderse a propósito.



La pertenencia a más de un marco se estimó preguntando a quienes respondieron la encuesta si habían estado o esperaban estar en cocinas populares, refugios o en la calle durante el periodo de muestreo (24 horas).

12.4.4 Muestreo de redes

En una encuesta familiar, como la NCVS, cada familia proporciona información exclusivamente de los problemas ocurridos a miembros de esa familia. En una muestra de red para estudiar a las víctimas de crímenes (Czaja y Blair 1990; vea Sudman et al 1988 para el método general), cada familia de la población está vinculada a otras unidades de la población; la familia de la muestra también puede proporcionar información acerca de las unidades vinculadas a ella (llamada la *red* para esa familia). Por ejemplo, podríamos definir la red de una familia como los hermanos de los miembros adultos de la familia.

Suponga que se extrae una muestra con probabilidades iguales. A cada miembro adulto de una familia seleccionada para estar en la muestra se le pide que proporcione información acerca de los incidentes delictivos que le han ocurrido a él o a cada uno de sus hermanos. Así, se puede saber si alguien ha sido víctima de robo, si su familia es seleccionada para la muestra o bien si se ha seleccionado la familia de su hermano. La probabilidad de incluir a la persona en la muestra depende de la cantidad de familias separadas donde tiene hermanos; si tiene muchos hermanos en familias distintas, el peso asignado a la persona será menor que el de una persona que no tiene hermanos.

12.4.5 Muestreo de bola de nieve

El muestreo de bola de nieve se basa en la premisa de que los miembros de la población rara se conocen entre sí. Para extraer una muestra de bola de nieve de las personas desamparadas, se encuentran unas cuantas de estas personas. Se pide a cada una de ellas que identifique a otras sin hogar para su muestra, luego pide a las nuevas personas de su muestra que identifiquen a otras sin hogar, etcétera, hasta alcanzar un tamaño de muestra deseado. El muestreo de bola de nieve puede crear una muestra relativamente grande de una población rara, pero no es una muestra de probabilidad; hay que establecer hipótesis fuertes sobre el modelo (¡que generalmente no se cumplen!) para generalizar los resultados de una muestra de bola de nieve a la población. Sin embargo, este tipo de muestreo puede ser útil en las primeras etapas de una investigación para aprender algo acerca de la población rara.

12.4.6 Muestreo secuencial

En el muestreo secuencial, las observaciones o unidades primarias se extraen una o unas cuantas a la vez y la información de las unidades primarias extraídas con anterioridad puede usarse para modificar el diseño de muestreo para las unidades primarias seleccionadas posteriormente. En un método que surge desde Stein (1945) y Cox (1952), se extrae una muestra inicial y los resultados de esa muestra se emplean para estimar el tamaño adicional de muestra necesario para lograr una precisión deseada. Si se quiere que la muestra contenga un cierto número de miembros de la población rara, podemos usar la muestra inicial para obtener una estimación preliminar de la frecuencia, que utilizamos para estimar el tamaño necesario de la segunda muestra.

Después de recoger la segunda muestra, se combina con la inicial para obtener estimaciones de la población. Por lo general, un esquema de muestreo secuencial debe tomarse en cuenta para la estimación; por ejemplo, en el método de Cox, la varianza muestral obtenida después de combinar los datos de las muestras inicial y segunda está sesgada hacia abajo (Lohr 1990). El libro de Wetherill y Glazebrook (1986) es un buen punto de partida para lecturas adicionales relativas a los métodos secuenciales.

El muestreo por conglomerados con adaptación supone que la población rara está reunida: los caribúes aparecen en manadas, una enfermedad infecciosa está concentrada en regiones del país o las herramientas están reunidas en sitios específicos de una excavación arqueológica. Se selecciona una muestra de probabilidad inicial de unidades primarias (con frecuencia, cuadrados, en las aplicaciones a la vida salvaje). Para cada unidad primaria en la muestra original se mide una respuesta, como el número de caribúes en la unidad primaria. Si la cantidad de caribúes en la unidad i excede una cifra predeterminada, entonces se agre-

gan a la muestra las unidades vecinas. De nuevo, la naturaleza de adaptación del esquema de muestreo debe tomarse en cuenta al estimar las cantidades de la población; si usted estima la densidad de los caribúes como (número de caribúes observados)/(número de unidades primarias en la muestra), su estimación de la densidad será demasiado alta. Thomson y Seber (1996) describen varios enfoques del muestreo por conglomerados con adaptación y dan bibliografía sobre el tema.

12.4.7 Ausencia de respuesta en el muestreo de poblaciones raras

Nunca nos gusta la ausencia de respuesta, pero ésta puede tener un riesgo particular en las encuestas de poblaciones raras. Si es más probable que los miembros de la población con la característica rara no respondan en relación con los miembros sin la característica rara, las estimaciones de la frecuencia serán sesgadas. En ciertas encuestas de salud, la propia característica puede llevar a una ausencia de respuesta: una encuesta de pacientes con cáncer puede tener ausencia de respuesta debido a que la enfermedad evita que las personas contesten.

12.5 Respuesta aleatorizada

A veces, se desea realizar una encuesta con cuestiones muy delicadas, como "¿usa usted cocaína?" o "¿alguna vez ha robado algo de una tienda?" o "¿indicó un ingreso menor en su declaración de impuestos?"

En estas preguntas, es de esperar que quienes deban responder "sí" mientan. Es recomendable una forma de la pregunta que anime a responder con la verdad, sin incomodar a las personas. Horvitz *et al* (1967), en una variante de la idea original de Warner (1965), sugieren el uso de dos preguntas, la pregunta delicada y la inocua, usando un mecanismo de aleatorización (como el lanzamiento de una moneda) para determinar la pregunta que debe responder la persona. Si se usa el lanzamiento de una moneda como mecanismo de aleatorización, se podría indicar a la persona que responda a la pregunta "¿usó usted cocaína la semana pasada?" si la moneda cae en cara y "¿está el minutero de su reloj entre 0 y 30?" si la moneda cae en cruz. El entrevistador no sabe si la moneda cae en cara o cruz y por tanto no sabe qué pregunta se deba contestar. Se espera que la aleatorización y el conocimiento de que el entrevistador no sepa qué pregunta se responde anime a quienes respondan a decir la verdad si han usado cocaína la semana pasada.

El mecanismo de aleatorización puede ser cualquier cosa, pero debe tener una probabilidad conocida P de que se plantee a la persona la pregunta delicada y $1-P$ de que se plantee la pregunta inocua. Fox y Tracy (1986) describen otras formas de respuesta aleatorizada.

La clave para la respuesta aleatorizada es conocer la probabilidad de que la persona responda sí a la pregunta inocua, p_I. Queremos estimar p_S, la proporción que responde sí a la pregunta delicada.

EJEMPLO 12.8 En una implantación de la respuesta aleatorizada (Duffy y Waterton 1988), la persona que responde recibe una baraja de 50 naipes; 10 tienen la instrucción "responde 'sí'", 10 tienen la instrucción "responde 'no'", y las otras 30 contienen la pregunta delicada "¿alguna vez ha bebido más allá del límite legal inmediatamente antes de manejar un auto?" Se pide a quien responde que examine el naipe (debe darse cuenta de que algunos naipes no preguntan la cuestión delicada), que la revuelva y que seleccione una. La persona que responde no muestra la carta al entrevistador, pero se le pide que responda a la pregunta delicada con la verdad si ésta aparece en la carta, y en caso contrario que responda sí o no, según lo indicado por la carta. En este contexto,

$$P = P(\text{plantear la pregunta delicada}) = 0.6,$$

y

$$p_I = P(\text{decir sí} \mid \text{se plantea la pregunta inocua}) = 0.5. \quad \blacksquare$$

Si todos contestan la pregunta con la verdad, entonces

$$\begin{aligned}\phi &= P(\text{quien responde dice sí}) \\ &= P(\text{sí} \mid \text{se plantea la pregunta delicada})P(\text{preguntar la cuestión delicada}) \\ &\quad + P(\text{sí} \mid \text{se plantea la pregunta inocua})P(\text{preguntar la cuestión inocua}) \\ &= p_S P + p_I(1-P).\end{aligned}$$

Sea $\hat{\phi}$ la proporción de "sí" de la muestra. Como conocemos P y p_I, podemos estimar p_S como

$$\hat{p}_S = \frac{\hat{\phi} - (1-P)p_I}{P}. \tag{12.10}$$

Entonces la varianza estimada de $\hat{p}_S$ es

$$\hat{V}(\hat{p}_S) = \frac{\hat{V}(\hat{\phi})}{P^2}.$$

El "castigo" para la respuesta aleatorizada aparece en el factor $1/P^2$ en la varianza estimada. Si $P = 1/3$, por ejemplo, la varianza es nueve veces mayor de lo que sería si a cada persona de la muestra se le hubiese planteado la pregunta delicada y ella hubiera respondido con la verdad.

Usted necesita reflexionar antes de elegir P: Mientras mayor sea P, menor será la varianza de $\hat{p}_S$, pero si P es demasiado grande, las personas que responden podrían pensar que el entrevistador sabrá cuál pregunta están contestando. Algunas personas supondrán que sólo $P = 0.5$ es "justo" y que no existen otras probabilidades al elegir entre los dos elementos.

EJEMPLO 12.9 Se elige una muestra aleatoria simple de alumnos de último año de bachillerato. Cada estudiante de la muestra recibe una tarjeta con las siguientes preguntas:

Pregunta 1: ¿Has copiado alguna vez en un examen?

Pregunta 2: ¿Naciste en julio?

Sabemos, por las actas de nacimiento, que $p_I = 0.085$. Suponga que el mecanismo de aleatorización es una ruleta, con $P = 1/5$. De las 800 personas de la encuesta, 175 contestan sí a la pregunta indicada por la ruleta. Entonces, $\hat{\phi} = 175/800$. Como ésta es una muestra aleatoria simple,

$$\hat{V}(\hat{\phi}) = \frac{\hat{\phi}(1-\hat{\phi})}{n-1} = 0.0002139.$$

Así,

$$\hat{p}_S = \frac{\frac{175}{800} - \left(\frac{4}{5}\right).085}{\frac{1}{5}} = .75375,$$

y

$$\hat{V}[\hat{p}_S] = \frac{0.0002139}{\left(\frac{1}{5}\right)^2} = 0.0053. \blacksquare$$

Sin embargo, antes de usar los métodos de respuesta aleatorizada en su encuesta, haga una prueba del método con personas de la población para ver si la complicación adicional realmente aumenta el acatamiento de las condiciones y parece reducir el sesgo. Brown y Harding (1973), al comparar la respuesta aleatorizada con un cuestionario anónimo que planteaba las preguntas en forma directa, hallaron que las estimaciones del uso de drogas entre los oficiales del ejército eran mayores para el método de respuesta aleatorizada que para el cuestionario. Se supone que una estimación mayor en esta situación tiene menor sesgo. Sin embargo, no todas las pruebas de campo muestran que la respuesta aleatorizada es una mejora.

EJEMPLO 12.10 Duffy y Waterton (1988) emplearon una muestra por conglomerados en dos etapas para seleccionar a las personas que responderían a su encuesta para estimar la incidencia de varios problemas relacionados con el alcohol en Edimburgo, Escocia. Las 20 unidades primarias (distritos de votación) se seleccionaron con probabilidad proporcional al número de votantes registrados. Luego se eligieron 75 personas al azar de cada distrito seleccionado y las personas en hospitales y otras instituciones se eliminaron de la muestra. A la quinta parte de quienes respondieron se les asignó al azar para contestar preguntas directas; las otras participaron en la respuesta aleatorizada. Como ésta era una muestra por conglomerados, hay que usar las fórmulas del capítulo 6 para estimar ϕ y $V(\hat{\phi})$, con $\hat{V}(\hat{p}_S) = \hat{V}(\hat{\phi})/P^2$. Para este estudio, la tasa de respuesta fue de 81.1% para el grupo con preguntas directas y 76.5% para el grupo de respuesta aleatorizada. Las estimaciones de p_S, la proporción de personas que habían bebido más del límite legal inmediatamente antes de conducir un auto, fue de 0.469 para el grupo con preguntas directas y 0.382 para el grupo de respuesta aleatorizada (la diferencia entre estas proporciones no fue significativa desde el punto de vista estadístico). Así, en este estudio, los investigadores determinaron que la respuesta aleatorizada no aumentó la tasa de respuesta ni la incidencia estimada de la característica delicada.

Sin embargo, la respuesta aleatorizada sí aumentó la complejidad de las entrevistas. Los entrevistadores reportaron que pocas personas del grupo de respuesta aleatorizada revisaron las tarjetas antes de elegir alguna. Varias de las personas que contestaron, en particular las personas de mayor edad y con menor instrucción, tuvieron dificultades en entender el método. Además, muchos contestaron "di sí" o "di no" en lugar de "sí" o "no" cuando extrajeron una de las tarjetas con preguntas inocuas, de modo que el entrevistador supo qué tarjeta se había seleccionado. Duffy y Waterton sugieren que la habilidad del entrevistador puede ser más importante que la técnica de la encuesta para obtener respuestas verdaderas y altas tasas de respuesta. ■

12.6 Ejercicios

***1** (Requiere probabilidad.) Suponga que la muestra de la primera etapa es una muestra aleatoria simple de tamaño $n^{(1)}$ y que la submuestra de la segunda etapa es una muestra aleatoria simple de tamaño $n^{(2)}$, con $n^{(2)} < n^{(1)}$. Muestre que

$$V(\hat{t}_{y}^{(2)}) = N^2\left(1 - \frac{n^2}{N}\right)\frac{S_y^2}{n^{(2)}}$$

es la misma varianza que se obtendría al extraer de manera directa una muestra aleatoria simple de tamaño $n^{(2)}$.

***2** (Requiere probabilidad.) Para el muestreo en dos etapas con estimación por proporción (página 379), suponga que la muestra de la primera etapa es una muestra aleatoria simple de tamaño $n^{(1)}$ y que la muestra de la segunda etapa es una muestra aleatoria simple de tamaño fijo $n^{(2)}$.

a Muestre que $P(Z_i = 1) = n^{(1)}/N$, y $P(D_i = 1 \mid Z) = Z_i n^{(2)}/n^{(1)}$.

b Muestre que la varianza del estimador es

$$V(\hat{t}_{yr}^{(2)}) \approx N^2\left(1 - \frac{n^{(1)}}{N}\right)\frac{S_y^2}{n^{(1)}} + N^2\left(1 - \frac{n^{(2)}}{n^{(1)}}\right)\frac{S_d^2}{n^{(2)}}$$

donde S_d^2 es la varianza de población de las d_i y $d_i = y_i - (t_y/t_x)x_i$.

c Sea $e_i = y_i - (\hat{t}_y^{(2)}/\hat{t}_x^{(2)})x_i$ y sean s_y^2 y s_e^2 las varianzas muestrales de las y_i y las e_i de la muestra de la segunda etapa. Muestre que

$$\hat{V}(\hat{t}_{yr}^{(2)}) = N^2\left(1 - \frac{n^{(1)}}{N}\right)\frac{S_y^2}{n^{(1)}} + N^2\left(1 - \frac{n^{(2)}}{n^{(1)}}\right)\frac{s_e^2}{n^{(2)}}$$

estima $V(\hat{t}_{yr}^{(2)})$.

***3** *Estimación de la varianza en el muestreo en dos etapas para la estratificación.* Muestre que (12.4) es un estimador insesgado de $V(\hat{t}_{\text{est}}^{(2)})$ en la sección 12.1.3. SUGERENCIA: Use el resultado del capítulo 4 (página 101) de que $S^2 = \left[\sum_{h=1}^{H}(N_h - 1)S_h^2 + \sum_{h=1}^{H} N_h(\bar{y}_{hU} - \bar{y}_U)^2\right]/(N-1)$.

***4** (Requiere cálculo.) *Asignación óptima para el muestreo en dos etapas con estratificación.* Se logra más eficiencia para el muestreo en dos etapas cuando se extrae una submuestra de más observaciones en los estratos con una varianza grande, un valor grande de N_h, o un bajo costo. Rao (1973) propone hacer $m_h = v_h n_h$ para el estrato h, con v_h, $h = 1, \ldots, H$, constantes por determinar antes del muestreo.

a Sea c el costo por extraer una unidad en la muestra de la primera etapa y determinar su pertenencia a algún estrato. Sea c_h el costo por medir y para una unidad del estrato h en la segunda etapa. Suponga que el costo total será una función lineal:

$$C = cn + \sum_{h=1}^{H} c_h m_h.$$

El costo total C varía de una muestra a otra, pues los m_h sólo se determinan después de extraer la muestra de la primera etapa. Muestre que el costo esperado es

$$E[C] = cn + n\sum_{h=1}^{H} c_h v_h W_h, \qquad \textbf{(12.11)}$$

donde $W_h = N_h/N$.

b Con v_h fijo,

$$V\left(\bar{y}_{\text{est}}^{(2)}\right) = S^2\left(\frac{1}{n}-\frac{1}{N}\right)+\frac{1}{n}\sum_{h=1}^{H} W_h S_h^2\left(\frac{1}{v_h}-1\right).$$

Muestre que $V(\bar{y}_{\text{est}}^{(2)})$ se minimiza, sujeto a la restricción en (12.11), cuando

$$v_h = \sqrt{\frac{cS_h^2}{c_h\left(S^2-\sum_{j=1}^{H} W_j S_j^2\right)}}.$$

SUGERENCIA: Use multiplicadores de Lagrange. O bien, use (12.11) para expresar n como una función del costo esperado y de los demás valores, sustituya esta expresión en vez de n en $V(\bar{y}_{\text{est}}^{(2)})$ y luego calcule las derivadas parciales.

c Para un costo esperado dado C^*, determine el valor de n.

Se han propuesto otras formas de asignación óptima; vea otros métodos y algoritmos en Treder y Sedransk (1993).

5 Use los resultados del ejercicio 4 para determinar una asignación óptima para una encuesta de seguimiento similar a la del ejemplo 12.1. Suponga que los costos relativos son $c = 1$ y $c_h = 20$, para $h = 1, 2, 3$. Use los datos del ejemplo 12.1 para estimar cantidades como W_h y S_h^2. ¿En qué difiere su asignación de la utilizada? ¿De la asignación proporcional?

6 Observe que en (12.6), $\hat{N} = n_1/\hat{p}$, donde $\hat{p}$ es la proporción muestral de individuos marcados de la segunda muestra. Use la linealización para determinar una estimación de $V(\hat{N})$.

7 Con frecuencia, la distribución de $\hat{N}$ en (12.6) no es aproximadamente normal. Sin embargo, la distribución de $\hat{p} = m/n_2$, es frecuentemente cercana a la normalidad, y es fácil construir los intervalos de confianza para p. Para los datos del ejemplo 12.5, determine un intervalo de confianza de 95% para $\hat{p}$. ¿Cómo puede utilizar ese intervalo y obtener un intervalo de confianza para N? ¿Cuál es la relación del intervalo de confianza resultante con los otros ya calculados? ¿Es simétrico el intervalo con respecto de $\hat{N}$?

***8** (Requiere probabilidad.) En un lago con N peces, n_1 de ellos marcados, la probabilidad de obtener m peces recapturados y $n_2 - m$ no capturados anteriormente en una muestra aleatoria simple de tamaño n_2 es

$$\mathcal{L}(N \mid n_1, n_2) = \frac{\binom{n_1}{m}\binom{N-n_1}{n_2-m}}{\binom{N}{n_2}}.$$

La estimación de máxima verosimilitud $\hat{N}$ de N es el valor que maximiza $\mathcal{L}(N)$; es el valor que hace al valor observado de m parecer más probable si conocemos n_1 y n_2. Determine la estimación de máxima verosimilitud de N. SUGERENCIA: ¿Cuándo ocurre que $\mathcal{L}(N) \geq \mathcal{L}(N-1)$?

***9** (Requiere estadística matemática.) *Estimación de máxima verosimilitud de N en muestras grandes*. Suponga que n_1 de los N peces de un lago están marcados. Se extrae entonces una muestra aleatoria simple de n_2 peces, y n de ellos están marcados. Suponga que N, n_1 y n_2 son todos "grandes". Entonces, la probabilidad de que m de los peces de la muestra estén marcados es aproximadamente

$$\mathcal{L}(N) = \binom{n^2}{m}\left(\frac{n_1}{N}\right)^m\left(1-\frac{n_1}{N}\right)^{n_2-m}.$$

a Muestre que $\hat{N} = n_1 n_2/m$ es la estimación de máxima verosimilitud de N.

b Use la teoría de máxima verosimilitud para mostrar que la varianza asintótica de $\hat{N}$ es aproximadamente $N^2(N-n_1)/(n_1 n_2)$.

***10** (Requiere cálculo.) Suponga que el costo por atrapar un pez es el mismo para cada pez en la primera y segunda muestras y que usted tiene los recursos suficientes para atrapar un total de $n_1 + n_2 = C$. Si conocemos N y C y $C < N$, ¿cuáles deben ser los valores de n_1 y n_2 para minimizar la varianza en el ejercicio 9(b)?

***11** (Requiere probabilidad.)

a Para la estimación de Chapman $\overline{N}$ en (12.7), sea X la variable aleatoria que denota la cantidad de individuos marcados en la segunda muestra. ¿Cuál es la distribución de probabilidad de X?

b Muestre que $E[\tilde{N}] = N$ si $n_2 \geq N - n_1$.

12 Los investigadores del Departamento de Recursos Naturales de Wisconsin (1993) usaron la captura y recaptura para estimar la cantidad de martas en el área de estudio Monico en Wisconsin.

a En el primer estudio, se capturaron siete martas entre el 11 de agosto de 1981 y el 31 de enero de 1982. Se capturaron 12 martas entre el 1 y el 19 de febrero de 1982; de esas 12, cuatro también habían sido capturadas en la primera muestra. Dé una estimación de la cantidad total de martas en el área, junto con un intervalo de confianza de 95% de la estimación.

b En el segundo estudio, se capturaron 16 martas entre el 28 de septiembre y el 31 de octubre de 1982, y 19 martas entre el 1 y el 17 de noviembre de 1982; 11 de las 19 martas de la segunda muestra ya habían sido capturadas en la primera muestra. Dé una estimación de la cantidad total de martas en el área, junto con un intervalo de confianza de 95% para su estimación.

c ¿Qué hipótesis está estableciendo para calcular estas estimaciones? ¿Qué significan estas hipótesis en términos del comportamiento de las martas y la posibilidad de atraparlas?

13 Suponga que el lago tiene N peces, y que n_1 de ellos están marcados. Entonces se extrae una muestra de tamaño n_1 del lago. Se eligen tres valores de N, n_1 y n_2. Aproximamos la distribución de $\hat{N}$ extrayendo 1000 muestras distintas de tamaño n_2 de la población de N unidades y trazando un histograma de las $\hat{N}$ resultantes de las distintas muestras. Repita esto para otros valores de N, n_1 y n_2. ¿Cuándo parece que el histograma tiene aproximadamente una distribución normal?

[Una versión alternativa de este problema consiste en calcular la distribución de probabilidad de $\hat{N}$ para distintos valores de N, n_1 y n_2 usando la distribución hipergeométrica dada en el ejercicio 8. Tal vez desee usar la fórmula de Stirling (véase Durrett 1994, 156) para aproximar los factoriales.]

14 Utilice el método de captura y recaptura con dos muestras para estimar la cantidad total de granos de palomitas de maíz o frijoles secos en un paquete o para estimar la cantidad total de monedas en un frasco. Describa en forma completa lo que hizo y de la estimación del tamaño de la población junto con un intervalo de confianza de 95% para N. ¿Cómo eligió los tamaños de las dos muestras?

15 Repita el ejercicio 14, usando tres muestras y modelos loglineales. ¿Sería de esperar que el modelo de completa independencia se ajuste bien a este caso? ¿Realmente ocurre esto?

16 Domingo-Salvany et al (1995) también usaron la captura y recaptura en el estudio de la sala de emergencia, dividiendo la lista en cuatro muestras de acuerdo con el trimestre. Los siguientes datos son parte de la primera tabla de su artículo.

	TR1 sí TR2 sí	TR1 sí TR2 no	TR1 no TR2 sí	TR1 no TR2 no
TR3 sí, TR4 sí	29	35	35	96
TR3 sí, TR4 no	48	58	80	400
TR3 no, TR4 sí	25	77	50	376
TR3 no, TR4 no	97	357	312	?

Ajuste modelos loglineales a estos datos. ¿Cuál modelo cree que sea mejor? Use su modelo para estimar la cantidad de personas en la celda faltante y construya un intervalo de confianza de 95% para su estimación.

17 Cochi *et al* (1989) registraron los datos relativos al síndrome de rubéola congénita mediante dos fuentes. El Registro Nacional del Síndrome de Rubéola Congénita (NCRSR) obtuvo los datos a través de informes voluntarios de los departamentos de salud estatales y locales. El programa de inspección de defectos al nacer (BDMP) obtuvo datos de los registros de descarga de hospitales en un subconjunto de los mismos. A continuación de muestran los datos de 1970 a 1985 de ambos sistemas:

Año	NCRSR	BDMP	Ambos	Año	NCRSR	BDMP	Ambos
1970	45	15	2	1978	18	9	2
1971	23	3	0	1979	39	11	2
1972	20	6	2	1980	12	4	1
1973	22	13	3	1981	4	0	0
1974	12	6	1	1982	11	2	0
1975	22	9	1	1983	3	0	0
1976	15	7	2	1984	3	0	0
1977	13	8	3	1985	1	0	0

a Los autores afirman que el NCRSR y el BDMP son fuentes independientes de información. ¿Cree que esto sea plausible? ¿Qué ocurre con las otras hipótesis del método de captura y recaptura?

b Use la estimación de Chapman (12.7) para determinar $\hat{N}$ para cada año.

c Ahora reúna los datos de todos los años y estime la cantidad total de casos del síndrome de rubéola congénita entre 1970 y 1985. ¿Cuál es la relación de su estimación obtenida al reunir todos los datos con la suma de las estimaciones de la parte (b)? ¿Cuál cree que sea más confiable?

d ¿Existe alguna evidencia del declive del síndrome? Proporcione un análisis estadístico para justificar su respuesta.

18 Frank (1978) informa acerca del siguiente experimento para estimar la cantidad de carpas en un tanque. Las dos primeras muestras usaron una trampa para carpas para atrapar los peces, mientras que la tercera utilizó una red. Las carpas atrapadas en la primera muestra fueron marcadas en la aleta caudal y las atrapadas en la segunda muestra fueron marcadas en la aleta pectoral izquierda.

¿Muestra 1?	¿Muestra 2?	¿Muestra 3?	Cantidad de peces
S	S	S	17
S	N	S	28
N	S	S	52
N	N	S	234
S	S	N	80
S	N	N	223
N	S	N	400

¿Qué modelo loglineal proporciona el mejor ajuste a estos datos? Usando este modelo, estime la cantidad total de peces y proporcione un intervalo de confianza de 95% para su estimación.

19 En el experimento del ejercicio 18, ¿qué significa en términos del comportamiento de los peces el hecho de una interacción entre la presencia en la primera muestra y la presencia en la segunda? ¿Y entre la presencia en la primera muestra y la presencia en la tercera?

20 Egeland et al (1995) usaron el método de captura y recaptura para estimar la cantidad total de casos de síndrome fetal por alcoholismo entre los nativos de Alaska nacidos entre 1982 y 1989. Se usaron dos fuentes de casos: 13 identificados por médicos particulares y 45 identificados por el Servicio de Salud para los Indígenas (IHS). Ocho casos estaban en ambas listas.

a Estime la cantidad total de casos de síndrome fetal por alcoholismo. Dé un intervalo de confianza de 95% para su estimación, usando la prueba ji cuadrada invertida o la técnica de bootstrap.

b La estimación de captura y recaptura se basa en la hipótesis de que las dos fuentes de datos son independientes; es decir, un niño en la lista IHS tiene la misma probabilidad de aparecer en la lista de los médicos particulares que un niño que no está en la lista IHS. ¿Cree que esta hipótesis sea válida en este caso? ¿Por qué? ¿Qué consejo daría a los investigadores si les preocupa la independencia?

c Suponga que es menos probable que los niños atendidos por médicos particulares sean vistos por el IHS. ¿Es probable que $\hat{N}$ subestime o sobreestime la cantidad de niños con síndrome fetal por alcoholismo? Justifique su respuesta.

21 Una universidad desea estimar la proporción de sus estudiantes que han usado cocaína. Los estudiantes se clasificaron en tres grupos: licenciatura, posgrado y escuela profesional (es decir, medicina o derecho) y se seleccionaron en la muestra al azar dentro de los grupos. Como había cierta preocupación por el hecho de que los estudiantes no estuviesen dispuestos a revelar su uso de la cocaína a un funcionario universitario, se empleó el siguiente método. En una caja se colocaron y revolvieron bien 30 bolas rojas, 16 azules y cuatro blancas. Luego se pidió a cada estudiante que extrajese una bola de la caja. Si la bola extraída era roja, la persona contestaba la pregunta 1 y en caso contrario contestaba la pregunta 2.

Pregunta 1: ¿Alguna vez ha usado cocaína?

Pregunta 2: ¿Es blanca la bola que extrajo?

Los resultados son los siguientes:

Grupo	Licenciatura	Posgrado	Profesional
Cantidad total de estudiantes en el grupo	8972	1548	860
Cantidad de estudiantes en la muestra	900	150	80
Cantidad de los que contestaron sí	123	27	27

Suponiendo que todas las respuestas fueron verdaderas, estime la proporción de estudiantes que han usado cocaína y reporte el error estándar de su estimación. Compare este error estándar con el error que sería de esperar si se preguntase a los estudiantes de la muestra la pregunta directa y si la contestaran con la verdad.

Ahora suponga que todas las personas que respondieron lo hicieron con la verdad con el método de respuesta aleatorizada, pero que 25% de quienes utilizaron cocaína negaron el hecho al cuestionárseles en forma directa. ¿Cuál método da una estimación de la proporción global de estudiantes que han usado cocaína con el menor error cuadrático medio?

22 Kuk (1990) propone el siguiente método de respuesta aleatorizada. Se pide a quien responde que genere dos variables binarias independientes X_1 y X_2 con $P(X_1 = 1) = \theta_1$ y $P(X_2 = 1) = \theta_2$. Conocemos las probabilidades θ_1 y θ_2. Ahora pedimos a quien responde que diga el valor de X_1 si la persona está en la clase sensible y X_2 en caso contrario. Suponga que la verdadera proporción de personas en la clase sensible es p_S.

a ¿Cuál es la probabilidad de que quien responda reporte 1?

b Use su respuesta de la parte (a) para dar una estimación $\hat{p}_S$ de p_S. ¿Qué condiciones deben satisfacer θ_1 y θ_2?

c ¿Cuál es el valor de $V(\hat{p}_S)$ si se extrae una muestra aleatoria simple?

A

Conceptos de probabilidad utilizados en muestreo

No tengo muchos recuerdos de ese día, excepto la rapidez de Johnson: al destacar el doctor Beattie que había visto a los carruajes 1 y 1000 (el primero y el último), como si fuera algo notable, Johnson replicó: "¿y eso qué? Hay la misma posibilidad de ver esos que cualesquier otros dos". Evidentemente tenía razón, aunque ver los dos extremos (cada uno de los cuales es, en cierto grado, más llamativo que los demás) puede llamar más la atención que ver cualquier otro par.

—James Boswell, *The Life of Samuel Johnson*

La esencia del muestreo de probabilidad es que podemos calcular la probabilidad con la que se seleccionará como muestra cualquier subconjunto de observaciones de la población. La mayor parte de las demostraciones de los resultados de la teoría de aleatorización usados en este libro dependen de conceptos de probabilidad. En este apéndice presentamos un breve repaso de algunas de las ideas básicas utilizadas. Para obtener más detalles y para la deducción de las demostraciones, el lector puede consultar una bibliografía más amplia sobre probabilidad, como Ross (1998) o Durrett (1994).

Como todo el trabajo de la teoría de aleatorización se ocupa de variables aleatorias discretas, en esta sección sólo daremos los resultados para estas variables. Usamos los resultados de las secciones B.1 a B.3 en los capítulos 2 a 4, y los resultados de la sección B.4 en los capítulos 5 y 6.

A.1 Probabilidad

Suponga que realiza un experimento en el cual usted puede describir todos los resultados que se podrían dar, pero sin saber exactamente cuál de ellos ocurrirá. Usted puede lanzar una moneda, extraer un naipe de una baraja o elegir tres nombres de un sombrero que contiene 20 opciones. Se asignan probabilidades a los distintos resultados y a conjuntos formados por resultados (llamados **eventos**), de acuerdo con la posibilidad de que ocurran tales eventos. Sea Ω el **espacio muestra**, la lista de todos los resultados posibles, para el lanzamiento de una moneda, $\Omega = \{\text{cara, cruz}\}$. Las probabilidades en los espacios muestra finitos