TECNICAS DE MUESTREO

WILLIAM G. COCHRAN
Professor of Statistics, Emeritus-Harvard University

COMPAÑIA EDITORIAL CONTINENTAL, S. A. MEXICO

DISTRIBUIDORES:

ESPAÑA–ARGENTINA–CHILE–VENEZUELA–COLOMBIA–PERU

Bolivia — Brasil — Costa Rica — Dominicana — Ecuador — El Salvador — Estados Unidos
Guatemala — Honduras — Nicaragua — Panamá — Paraguay — Portugal — Puerto Rico
Uruguay

Indice de Materias

Cap.		Pág.
	Prefacio	7
1	**INTRODUCCION**	19
	1.1 Las Ventajas del Método de Muestreo	19
	1.2 Algunos Usos de las Encuestas por Muestreo	21
	1.3 Etapas Principales en una Encuesta por Muestreo	24
	1.4 El Papel de la Teoría del Muestreo	28
	1.5 Muestreo Probabilista	29
	1.6 Alternativas al Muestreo Probabilista	30
	1.7 Uso de la Distribución Normal	32
	1.8 El Sesgo y sus Efectos	34
	1.9 El Error Cuadrático Medio	37
	Ejercicios	38
2	**MUESTREO ALEATORIO SIMPLE**	41
	2.1 Muestreo Aleatorio Simple	41
	2.2 Selección de una Muestra Aleatoria Simple	42
	2.3 Definiciones y Notación	44
	2.4 Propiedades de las Estimaciones	45
	2.5 Varianzas de las Estimaciones	47
	2.6 La Corrección por Población Finita	49
	2.7 Estimación del Error Estándar a Partir de una Muestra	50
	2.8 Límites de Confianza	52
	2.9 Un Método Alternativo de Prueba	53
	2.10 Muestreo Aleatorio con Restitución	55
	2.11 Estimación de una Razón	56
	2.12 Estimaciones de Medias en Subpoblaciones	60
	2.13 Estimaciones de Totales en Subpoblaciones	62
	2.14 Comparación Entre las Medias de los Dominios	65
	2.15 Validez de la Aproximación Normal	66
	2.16 Estimadores Lineales de la Media de Población	72
	Ejercicios	73

Cap.		Pág.
3	**MUESTREO PARA PROPORCIONES Y PORCENTAJES**	**79**
3.1	Características Cualitativas	79
3.2	Varianzas de las Estimaciones Muestrales	80
3.3	El Efecto de P en los Errores Estándar	83
3.4	La Distribución Binomial	84
3.5	La Distribución Hipergeométrica	85
3.6	Límites de Confianza	87
3.7	Clasificación en más de dos Clases	90
3.8	Límites de Confianza Cuando Existen más de dos Clases	91
3.9	La Distribución Condicional de p	92
3.10	Proporciones y Totales Sobre Subpoblaciones	93
3.11	Comparaciones Entre Dominios Diferentes	94
3.12	Estimación de Proporciones en el Muestreo por Conglomerados	95
	Ejercicios	100
4	**LA ESTIMACION DEL TAMAÑO DE LA MUESTRA**	**103**
4.1	Un Ejemplo Hipotético	103
4.2	Análisis del Problema	105
4.3	La Especificación de la Precisión	106
4.4	La Fórmula para n al Hacer un Muestreo para Determinar una Proporción	107
4.5	Atributos Raros-Muestreo Inverso	108
4.6	La Fórmula para n con Datos Continuos	109
4.7	Estimaciones Anticipadas de Varianzas de Población	110
4.8	Tamaño de la Muestra con más de una Característica	114
4.9	Tamaño de la Muestra Cuando las Estimaciones se Quieren para Subdivisiones de la Población	115
4.10	El Tamaño de la Muestra en Problemas de Decisión	117
4.11	El Efecto del Diseño *(Deff)*	119
	Ejercicios	120
5	**MUESTREO ALEATORIO ESTRATIFICADO**	**125**
5.1	Descripción	125
5.2	Notación	126
5.3	Propiedades de las Estimaciones	127
5.4	La Varianza Estimada y Límites de Confianza	132

Cap.		Pág.
5.5	La Asignación Optima	133
5.6	Precisiones Relativas del Muestreo Aleatorio Estratificado y del Muestreo Aleatorio Simple	136
5.7	¿En qué Casos Produce la Estratificación Considerables Ganancias de Precisión?	138
5.8	Asignación que Requiere más del 100% del Muestreo	142
5.9	Estimación del Tamaño de la Muestra con Datos Continuos	143
5.10	Muestreo Estratificado para Proporciones	145
5.11	Ganancias en Precisión en el Muestreo Estratificado para Proporciones	147
5.12	Estimación del Tamaño de Muestra con Proporciones	149
	Ejercicios	149
5A	**OTROS ASPECTOS DEL MUESTREO ESTRATIFICADO**	**155**
5A.1	Efectos de las Desviaciones a Partir de la Asignación Optima	155
5A.2	Efectos de Errores en los Tamaños de los Estratos	157
5A.3	El Problema de la Asignación con más de una Característica	160
5A.4	Otros Métodos de Asignación con más de un Atributo	161
5A.5	Estratificación en dos Direcciones, con Muestras Pequeñas	165
5A.6	Selección Controlada	167
5A.7	La Construcción de los Estratos	169
5A.8	Número de Estratos	174
5A.9	Estratificación Después de la Selección de la Muestra (Postestratificación)	177
5A.10	Muestreo por Cuota	178
5A.11	Estimación a Partir de una Muestra de la Ganancia Debida a la Estratificación	179
5A.12	Estimación de la Varianza con una Unidad por Estrato	181
5A.13	Estratos como Dominios de Estudio	184
5A.14	Estimación de Totales y Medias Sobre Subpoblaciones	186

Cap.		Pág.
	5A.15 Muestreo a Partir de dos Marcos	189
	Ejercicios	190
6	**ESTIMADORES DE RAZON**	**195**
	6.1 Métodos de Estimación	195
	6.2 El Estimador de Razón	196
	6.3 Varianza Aproximada de la Estimación de Razón	198
	6.4 Estimación de la Varianza a Partir de una Muestra	201
	6.5 Límites de Confianza	202
	6.6 Comparación de la Estimación de Razón con la Media por Unidad	203
	6.7 Condiciones Bajo las Cuales el Estimador de Razón es un Estimador Insesgado Lineal Optimo	204
	6.8 Sesgo de la Estimación de Razón	207
	6.9 Exactitud de las Fórmulas para la Varianza Estimada	209
	6.10 Estimaciones de Razón en Muestreo Aleatorio Estratificado	211
	6.11 La Estimación de Razón Combinada	212
	6.12 Comparación de las Estimaciones Separadas y Combinada	214
	6.13 Cálculo Abreviado de la Varianza Estimada	217
	6.14 Asignación Optima con una Estimación de Razón	220
	6.15 Estimaciones Insesgadas del Tipo de Razón	222
	6.16 Comparación de los Métodos	225
	6.17 Estimación Mejorada de la Varianza	227
	6.18 Comparación de dos Razones	229
	6.19 Razón de dos Razones	232
	6.20 Estimaciones de Razón Multivariantes	234
	6.21 Estimadores Producto	235
	Ejercicios	236
7	**ESTIMADORES DE REGRESION**	**239**
	7.1 La Estimación de Regresión Lineal	239
	7.2 Estimaciones de Regresión con b Preasignada	241
	7.3 Estimaciones de Regresión Cuando b se Calcula a Partir de la Muestra	243
	7.4 Estimación de Muestra de la Varianza	245
	7.5 Comparación en Muestras Grandes con la Estimación de Razón y la Media por Unidad	246

Cap.		Pág.
	7.6 Exactitud de las Fórmulas de Muestras Grandes Para $V(\bar{y}_{lr})$ y $v(\bar{y}_{lr})$	247
	7.7 Sesgo de la Estimación de Regresión Lineal	249
	7.8 El Estimador de Regresión Lineal en un Modelo de Regresión Lineal	250
	7.9 Estimaciones de Regresión en Muestreo Estratificado	251
	7.10 Coeficientes de Regresión Estimados de la Muestra	252
	7.11 Comparación de los dos Tipos de Estimaciones de Regresión	254
	Ejercicios	255
8	**MUESTREO SISTEMATICO**	**257**
	8.1 Descripción	257
	8.2 Relación con el Muestreo Conglomerado	259
	8.3 Varianza de la Media Estimada	260
	8.4 Comparación del Muestreo Sistemático con el Muestreo Aleatorio Estratificado	265
	8.5 Poblaciones en Orden Aleatorio	266
	8.6 Poblaciones con Tendencia Lineal	268
	8.7 Métodos para Poblaciones que Presentan Tendencias Lineales	269
	8.8 Poblaciones con Variación Periódica	271
	8.9 Poblaciones Autocorrelacionadas	273
	8.10 Poblaciones Naturales	276
	8.11 Estimación de la Varianza a Partir de una sola Muestra	278
	8.12 Muestreo Sistemático Estratificado	282
	8.13 Muestreo Sistemático en dos Dimensiones	283
	8.14 Resumen	285
	Ejercicios	287
9	**MUESTREO POR CONGLOMERADOS, DE UNA ETAPA: CONGLOMERADOS DEL MISMO TAMAÑO**	**289**
	9.1 Motivos del Muestreo por Conglomerados	289
	9.2 Una Regla Simple	290
	9.3 Comparaciones de Precisión, Hechas a Partir de Datos de Encuestas	294
	9.4 Varianza en Términos de la Correlación Dentro de Conglomerados	298
	9.5 Funciones de la Varianza	300

Cap.		Pág.
	9.6 Una Función de Costos	302
	9.7 Muestreo Conglomerado para Proporciones	304
	Ejercicios	306
9A	**MUESTREO CONGLOMERADO DE UNA ETAPA: CONGLOMERADOS DE TAMAÑOS DESIGUALES**	**309**
	9A.1 Unidades Conglomerados de Tamaños Desiguales	309
	9A.2 Muestreo con Probabilidad Proporcional al Tamaño	311
	9A.3 Selección con Probabilidades Desiguales y con Restitución	312
	9A.4 La Medida Optima del Tamaño	316
	9A.5 Exactitudes Relativas de las Tres Técnicas	316
	9A.6 Muestreo con Probabilidades Desiguales sin Restitución	320
	9A.7 El Estimador de Horvitz-Thompson	321
	9A.8 El Método de Brewer	323
	9A.9 El Método de Murthy	326
	9A.10 Métodos Relacionados con el Muestreo Sistemático	328
	9A.11 El Método de Rao, Hartley, Cochran	329
	9A.12 Comparaciones Numéricas	331
	9A.13 Estimaciones Estratificadas y de Razón	334
	Ejercicios	336
10	**SUBMUESTREO CON UNIDADES DE TAMAÑOS IGUALES**	**339**
	10.1 Muestreo en dos Etapas	339
	10.2 Determinación de Medias y Varianzas en Muestreo de dos Etapas	340
	10.3 Varianza de la Media Estimada en Muestreo de dos Etapas	342
	10.4 Estimación Muestral de la Varianza	343
	10.5 La Estimación de Proporciones	345
	10.6 Fracciones Optimas de Muestreo y Submuestreo	346
	10.7 Estimación de m_{opt} de un Reconocimiento Piloto	349
	10.8 Muestreo de Tres Etapas	352
	10.9 Muestreo Estratificado de las Unidades	355
	10.10 Asignación Optima con Muestreo Estratificado	356
	Ejercicios	357

Cap.		Pág.
11	**EL SUBMUESTREO CON UNIDADES DE DIFERENTES TAMAÑOS**	**359**
	11.1 Introducción	359
	11.2 Métodos de Muestreo Cuando $n = 1$	361
	11.3 Muestreo con Probabilidad Proporcional al Tamaño Estimado	365
	11.4 Resumen de los Métodos para $n = 1$	367
	11.5 Métodos de Muestreo Cuando $n > 1$	368
	11.6 Dos Resultados Utiles	368
	11.7 Unidades Seleccionadas con Probabilidades Iguales: Estimador Insesgado	371
	11.8 Unidades Seleccionadas con Probabilidades Iguales: Estimación de la Razón al Tamaño	372
	11.9 Unidades Seleccionadas con Probabilidades Iguales y con Restitución: Estimador Insesgado	375
	11.10 Selección de las Unidades sin Restitución	377
	11.11 Comparación de los Métodos	379
	11.12 Razones a Otra Variable	380
	11.13 Elección de Fracciones de Muestreo y Submuestreo. Probabilidades Iguales	382
	11.14 Probabilidades de Selección Optima y Tasas de Muestreo y Submuestreo	384
	11.15 Muestreo Estratificado. Estimadores Insesgados	386
	11.16 Muestreo Estratificado. Estimaciones de Razón	387
	11.17 Estimadores no Lineales en Encuestas Complejas	388
	11.18 Desarrollo en Series de Taylor	389
	11.19 Réplicas Repetidas Equilibradas	390
	11.20 El Método Jackknife	392
	11.21 Comparación de los Tres Métodos	393
	Ejercicios	395
12	**MUESTREO DOBLE**	**399**
	12.1 Descripción de la Técnica	399
	12.2 Muestreo Doble para Estratificación	400
	12.3 Asignación Optima	403
	12.4 Varianza Estimada en Muestreo Doble para Estratificación	406
	12.5 Muestreo Doble para Comparaciones Analíticas	408
	12.6 Estimadores de Regresión	412
	12.7 Asignación Optima y Comparación con el Muestreo Simple	414

Cap.		Pág.
	12.8 Varianza Estimada en Muestreo Doble para Regresión	416
	12.9 Estimadores de Razón	417
	12.10 Muestreo Repetido de la Misma Población	418
	12.11 Muestreo en dos Ocasiones	420
	12.12 Muestreo en más de dos Ocasiones	423
	12.13 Simplificaciones y Adelantos Ulteriores	426
	Ejercicios	431
13	**FUENTES DE ERROR EN LAS ENCUESTAS**	435
	13.1 Introducción	435
	13.2 Efectos de la no-Respuesta	436
	13.3 Tipos de no-Respuesta	440
	13.4 Revisitas	442
	13.5 Un Modelo Matemático de los Efectos de las Revisitas	444
	13.6 Fracción Optima de Muestreo Entre los que no Responden	448
	13.7 Ajustes por Sesgo sin Revisitas	452
	13.8 Un Modelo Matemático para Errores de Medición	455
	13.9 Efectos del Sesgo Constante	458
	13.10 Efectos de Errores no Correlacionados Dentro de la Muestra	460
	13.11 Efectos de la Correlación Intramuestral Entre Errores de Medición	462
	13.12 Resumen de los Efectos de Errores de Medición	464
	13.13 El Estudio de los Errores de Medición	464
	13.14 Mediciones Repetidas de Submuestras	466
	13.15 Submuestras Interpenetrantes	469
	13.16 Combinación de la Interpenetración y la Medición Repetida	472
	13.17 Preguntas Delicadas: Respuestas Aleatorizadas	473
	13.18 La Segunda Pregunta no Relacionada	475
	13.19 Resumen	478
	Ejercicios	479
	BIBLIOGRAFÍA	483
	RESPUESTAS A LOS EJERCICIOS	497
	INDICE ALFABÉTICO	505

CAPITULO 1

Introducción

1.1 LAS VENTAJAS DEL METODO DE MUESTREO

Nuestros conocimientos, actitudes y acciones están basados en gran parte sobre muestras. Esto es verdad real en la vida cotidiana como en la investigación científica. Así, la opinión que tiene una persona de una institución que realiza miles de transacciones diarias, a menudo se basa en uno o dos contactos ocurridos en el transcurso de muchos años. Es clásico (y cómico) el personaje que después de pasar diez días en un país extranjero, escribe un libro para enseñar a sus habitantes cómo revitalizar su industria, reformar su sistema político, equilibrar su presupuesto y mejorar la comida de sus hoteles. Pero en realidad, la diferencia que existe entre este personaje y el estudioso de ciencias políticas que vive veinte años en ese país dedicado a estudiarlo, es que el primero basa sus conclusiones en una muestra mucho más pequeña y es menos consciente de su gran ignorancia. Tanto en la ciencia como en los asuntos humanos careceremos de los recursos necesarios para estudiar más de un fragmento de los fenómenos que pueden ampliar nuestros conocimientos.

En este libro se presenta la teoría en que se basan los buenos métodos de muestreo. En la mayoría de las aplicaciones para las que se edificó esta teoría, el conjunto del que se desea obtener información es finito y bien delimitado como sería por ejemplo el conjunto de habitantes de una ciudad, las máquinas de una fábrica o los peces de un lago. En ciertos casos, puede parecer factible el obtener la información por medio de una enumeración completa o censo del conjunto. Los administradores acostumbrados a tratar con censos inicialmente desconfiaban de las muestras y se resistían a utilizarlas. Aunque esta actitud ha desaparecido, conviene revisar las ventajas del muestreo en comparación con la enumeración completa.

Costo reducido

Si los datos se obtienen únicamente de una pequeña fracción del total, los gastos son menores que los que se realizarían si se llevara a cabo un censo completo. En poblaciones muy grandes se pueden obtener resultados lo suficientemente exactos cuando se analizan muestras que representan sólo una pequeña fracción de la población. En los Estados Unidos las investigaciones periódicas más importantes las lleva a cabo el Gobierno, que usa muestras de alrededor de 105 000 personas o sea aproximadamente, una persona por cada 1 240. Las encuestas realizadas para obtener información en conexión con las ventas o campañas de publicidad en la investigación de mercados, pueden emplear muestras de sólo unos cuantos miles de observaciones.

Mayor rapidez

Por la misma razón, los datos pueden ser recolectados y resumidos más rápidamente con una muestra que con una enumeración completa. Esta es una consideración vital cuando se necesita la información con urgencia.

Más posibilidades

Para obtener la información en ciertos tipos de encuestas, se utilizan los servicios de personal altamente calificado o equipo muy especializado de disponibilidad limitada. Por lo tanto, en estos casos el censo completo es impracticable y como alternativa a la obtención de datos por muestreo, sólo existe la de no obtenerlos. De ahí que las encuestas basadas en el muestreo tienen más posibilidades y flexibilidad respecto a la información que puede obtenerse. Por otra parte, se desea una información exacta de muchas subdivisiones de la población, el tamaño de la muestra requerida es en ocasiones tan grande que la mejor opción es la enumeración completa.

Mayor exactitud

Debido a que al reducir el volumen de trabajo se puede emplear personal más capacitado y someterlo a un entrenamiento intensivo y debido también a que en estas condiciones será factible la supervisión cuidadosa del trabajo de campo y del procesamiento de los resultados, una muestra puede producir resultados más exactos que la enumeración completa.

1.2 ALGUNOS USOS DE LAS ENCUESTAS POR MUESTREO

Para quien observa el desarrollo del muestreo y sus aplicaciones en los últimos 25 años, lo más sorprendente es el rápido aumento del número y tipos de encuestas realizadas por muestreo. La Oficina de Estadística de las Naciones Unidas publica ocasionalmente un informe sobre las Encuestas por Muestreo de actualidad ("Sample Surveys of Current Interest") que llevan a cabo los países miembros. El informe de 1968 incluye una lista de encuestas procedentes de 46 países. Muchas de ellas contienen información de gran interés sobre la planeación nacional en áreas tales como la producción agrícola y la utilización de la tierra, el desempleo y la fuerza de trabajo disponible, la producción industrial, los precios al mayoreo y menudeo, las condiciones de salud de la población, y los ingresos y gastos familiares. Pero también existen estudios más especializados, por ejemplo, los arreglos para vacaciones anuales (Australia), las causas de divorcio (Hungría), la deuda e inversiones rurales (India), el consumo doméstico de agua (Israel), los radioescuchas (Malasia), los gastos de vacaciones (Holanda), la estructura de edades en el ganado vacuno (Checoslovaquia), y los empleos vacantes (E. U. A.).

El muestreo ha venido a representar un papel muy importante en los censos nacionales que se llevan a cabo cada 10 años. En los Estados Unidos, se incluyó en el censo de 1940 una muestra del 5%. Que hace preguntas extra sobre la ocupación, tamaño de la familia, fertilidad, etc., a todas aquellas personas cuyos nombres estaban en 2 de los 40 renglones de cada página de la lista. El uso del muestreo se aumentó grandemente en 1950. De una muestra del 20% (un nombre cada 5 renglones) se obtuvo información sobre el ingreso, la escolaridad, la migración y el servicio en las fuerzas armadas. Al tomar una de cada 6 personas en la muestra del 20%, se creó otra muestra del 3⅓% para obtener información sobre el matrimonio y la fertilidad. Una serie de preguntas, sobre la condición y edad de las viviendas, se dividió en 5 grupos, cada grupo se contestó en una de cada cinco casas. También se empleó el muestreo para aumentar la rapidez de la publicación de los resultados. Las tabulaciones preliminares para muchos atributos importantes, obtenidas por muestreo, se publicaron un año y medio antes que los reportes finales.

Este proceso se continuó en los censos de 1960 y 1970. Salvo cierta información básica que se necesitaba obtener de cada persona por razones constitucionales o legales, todo el censo se cambió a

una base de muestreo. Este cambio acompañado de un aumento en la mecanización utilizada para manejar la información, dio por resultado una publicación mucho más rápida y ahorros sustanciales.

Además de utilizar las muestras continuadas para los censos, las oficinas gubernamentales las emplean para obtener información de actualidad requerida. En los E. U. A. se llevan a cabo, por ejemplo, la encuesta de población (Current Population Survey), que proporciona datos mensuales sobre la magnitud y estructura de la fuerza de trabajo y del número de desempleados, la Encuesta Nacional de Salud, y la serie de muestras requeridas para el cálculo del índice mensual de precios al consumidor.

En más pequeña escala, los gobiernos locales —ciudad, estado y condado— tienen uso cada vez más de encuestas por muestreo para obtener la información requerida para la planeación y solución de problemas urgentes. En la planeación y en la solución de problemas urgentes. En la mayoría de las ciudades de los E. U. A. existen agencias dedicadas a planear y dirigir encuestas por muestreo para sus clientes.

La investigación de mercados depende básicamente del método de muestreo. Hay un escrutinio ininterrumpido de las audiencias de diferentes programas de radio y televisión, así como del número de lectores de diversas publicaciones periódicas (incluyendo la publicidad que en ellas aparece). Los fabricantes y distribuidores quieren saber la reacción del público a nuevos productos, empaques o presentaciones, así como sus quejas, y las razones por las que prefieren un producto y no otro.

Los negocios y las industrias utilizan el muestreo para aumentar la eficiencia de sus operaciones internas. Las importantes áreas de control de calidad y de muestreo de aceptación quedan fuera del alcance de este libro. Pero obviamente, las decisiones que se tomen respecto al nivel o cambio de calidad, o bien, la aceptación o rechazo de los lotes de producción, estarán bien fundadas sólo si los resultados obtenidos de la muestra son válidos para todo el lote (con un margen razonable de tolerancia). El muestreo de los registros de transacciones comerciales (cuentas, nóminas, existencias y personal), usualmente mucho más fácil que el muestreo de personas, puede proporcionar información de utilidad en forma rápida y económica. También, por medio del muestreo se pueden obtener ahorros en la estimación de los inventarios, en los estudios de la condición y el tiempo de vida de la maquinaria y el equipo, en la inspección de la exactitud y rapidez del trabajo del empleado de oficina, al investigar cómo el personal clave distribuye su tiempo entre diversas tareas y, más generalmente, en el nuevo campo conocido como

investigación de operaciones. Los libros de Deming (1960) y Slonim (1960) contienen muchos ejemplos interesantes que muestran la gama de aplicaciones del método de muestreo en el comercio.

Los sondeos de opinión, de actitudes y electorales a los que se debe en gran parte el primer contacto del muestreo con el gran público, siguen siendo un rasgo popular en los periódicos. En el campo de la contabilidad y la auditoría, en donde durante años se ha usado el muestreo, se palpa un nuevo interés por adaptar los desarrollos modernos a los problemas particulares de dicho campo. Por ejemplo Neter (1972) describe cómo ahorran dinero las compañías aéreas y ferroviarias al utilizar muestras de registros para separar los ingresos procedentes del servicio de carga, de los que provienen del servicio de pasajeros. El papel del muestreo en los litigios ha sido objeto de acaloradas discusiones. Gallup (1972) ha recalcado la gran contribución que las encuestas por muestreo pueden aportar al proceso de información gubernamental, al determinar rápidamente la opinión de la gente respecto a los nuevos programas de gobierno y al demostrar su importancia como fuentes de información en las ciencias sociales.

A grandes rasgos, las encuestas por muestreo pueden clasificarse en dos categorías principales — *descriptivas* y *analíticas*. En una encuesta descriptiva el único objetivo es obtener cierta información respecto a grandes grupos: por ejemplo, el número de hombres, mujeres y niños que acostumbran ver cierto programa de televisión. En una encuesta analítica se hacen comparaciones entre varios subgrupos de una población, para averiguar si existen ciertas diferencias entre ellos y formular o verificar hipótesis sobre sus causas. La encuesta de fertilidad en Indianápolis, por ejemplo, fue un intento para determinar hasta dónde los matrimonios planean el número y espaciamiento de sus hijos, las actitudes de ambos cónyuges a ese respecto, las razones de dichas actitudes y el éxito que logran en su planeación familia (Kiser y Whelpton, 1953).

La distinción entre encuestas analíticas y descriptivas no es perfectamente clara. Muchas encuestas proporcionan datos que sirven para ambos propósitos (el descriptivo y el analítico). Junto con el aumento del número de encuestas descriptivas se ha notado un aumento del número de encuestas realizadas con fines analíticos, particularmente, las relacionadas con el estudio del comportamiento humano y la salud. Podemos mencionar entre otras las encuestas sobre el estado de los dientes de los niños antes y después de la fluorización del agua, sobre las tasas de mortalidad y las causas de muerte de fumadores en los diferentes niveles de consumo de tabaco

y finalmente, la gigantesca encuesta sobre la efectividad de la vacuna Salk. El estudio que hizo Coleman (1966) sobre la igualdad en las oportunidades de educación, realizado en una muestra nacional de escuelas, contenía numerosos análisis de regresión que estimaban las contribuciones relativas de las características de la escuela, el medio familiar, y la actitud del niño frente a los diferentes resultados en los exámenes.

1.3 ETAPAS PRINCIPALES EN UNA ENCUESTA POR MUESTREO

Como introducción a una discusión sobre el papel que desempeña la teoría en una encuesta por muestreo, es útil describir brevemente los pasos involucrados en la planeación y ejecución de una encuesta. Las encuestas varían considerablemente en su complejidad. Es una tarea fácil el tomar una muestra de 5 000 tarjetas cuidadosamente arregladas y numeradas en un archivo. Pero la situación es otra si se desea tomar una muestra de los residentes de una región donde se usa un medio de transporte fluvial a través de la selva, donde no hay mapas, donde se hablan 15 dialectos diferentes y los habitantes desconfían de todo extranjero inquisitivo. Problemas que son desconcertantes en una encuesta pueden ser triviales o inexistentes en otra.

Los pasos principales en una encuesta están agrupados más o menos arbitrariamente bajo 11 encabezados y se citan a continuación.

Objetivos de la encuesta

Una exposición clara de los objetivos es lo más útil. Sin esto, es fácil olvidarlos en una encuesta compleja al preocuparse por los detalles de la planeación y por lo tanto tomar decisiones que varían de los objetivos.

Población bajo muestreo

La palabra *población* se emplea para denominar el conjunto del que se elige la muestra. La definición de población puede no presentar problema, por ejemplo, cuando se muestrea un grupo de focos de luz eléctrica a fin de estimar su tiempo de vida promedio. Por otra parte, en el muestreo de una población de propiedades agrícolas, se deben fijar las reglas para definir lo que es por ejemplo, un rancho o una hacienda, y surgen casos dudosos al tratar de hacer una delimitación exacta. Estas reglas deben ser aplicables a la práctica: el enumerador debe ser capaz de decidir en el campo, sin demasiados titubeos, si un caso dudoso pertenece o no a la población.

La población que se muestrea (la población *muestreada*) debe coincidir con la población sobre la cual se desea información (la población *objetivo*). En ocasiones, por razones de factibilidad, o simple conveniencia, la población muestreada es más restringida que la población objetivo. De ser así, debe recordarse que las conclusiones extraídas de la muestra son aplicables a la población muestreada, y habrá que recurrir a otras fuentes de información para decidir hasta qué grado se aplican estas conclusiones a la población objetivo. Toda información que se obtenga respecto a las diferencias entre ambas poblaciones será de utilidad.

Los datos recogidos

Es conveniente cerciorarse que todos los datos son pertinentes a la encuesta y que no se omiten datos esenciales. Particularmente en presencia de poblaciones humanas, existe la tendencia a hacer un número excesivo de preguntas que no se analizan posteriormente. Un cuestionario demasiado largo produce una baja general de la calidad de las respuestas, tanto a las preguntas importantes como a las otras.

Grado de precisión deseado

Los resultados de una encuesta por muestreo están siempre sujetos a cierta incertidumbre porque sólo se mide una parte de la población, y por los errores en las mediciones realizadas. Esta falta de certeza se puede reducir al tomar muestras más grandes y emplear mejores dispositivos de medición. Pero esto suele costar tiempo y dinero, en consecuencia, la especificación del grado de precisión deseado, es un paso importante en la preparación de la encuesta. Este paso es responsabilidad de la persona que va a utilizar los datos y puede presentar dificultades, porque los administradores no están acostumbrados a pensar en términos de la magnitud del error tolerable en las estimaciones, compatible con una buena decisión. El estadístico puede ayudar en esta etapa.

Métodos de medición

Puede existir la posibilidad de escoger el método de medición y el método de inspección de la población. Los datos acerca del estado de salud de una persona se pueden obtener de sus declaraciones,

o de un examen médico. La encuesta puede emplear un cuestionario autoadministrado, o un proceso de entrevistas en las que los entrevistadores simplemente leen un cuestionario prescrito, o bien, un proceso en el que se permite mucha libertad en la forma y el orden de las preguntas. La inspección puede ser por correo, por visitas personales, por teléfono o por una combinación de los tres medios. Se ha estudiado mucho sobre los métodos de entrevista y sus problemas (véase, por ejemplo, Hyman, 1954 y Payne, (1951).

Una parte importante del trabajo preliminar es la construcción de las formas de registro donde entrarán las preguntas y las respuestas. En cuestionarios sencillos a veces es posible precodificar las respuestas, es decir, colocarlas de tal modo que puedan transferirse rutinariamente a un equipo mecánico. De hecho para la construcción de buenas formas de registro se necesita prever la estructura de las tablas de resúmenes finales que se utilizan para obtener las conclusiones.

El marco

Antes de seleccionar la muestra, la población debe ser dividida en partes llamadas *unidades de muestreo* o *unidades*. Estas deben cubrir la totalidad de la población y no traslaparse en el sentido de que todo elemento de la población pertenezca a una y solamente a una unidad. Algunas veces, la unidad apropiada es obvia como en el caso de la población de los focos donde la unidad es el foco. En otras ocasiones, existe la posibilidad de escoger lo que será la unidad de muestreo. En el muestreo de los residentes de una ciudad, la unidad puede ser una persona, los miembros de una familia, o las personas que viven en una manzana. En el muestreo de una cosecha agrícola la unidad puede ser un lote, una granja, o una área de terreno cuya forma y dimensiones quedan a nuestra discreción.

Frecuentemente, la construcción de esta lista de unidades de muestreo, llamada *marco*, es uno de los principales problemas prácticos. Con base en amargas experiencias, los técnicos han adquirido una actitud crítica frente a las listas que han sido recolectadas en forma rutinaria y con algún propósito específico. Aunque se asegure lo contrario estas listas suelen ser incompletas, parcialmente ilegibles, o contienen duplicaciones de magnitud desconocida. Será difícil encontrar un buen marco cuando la población es especializada, como por ejemplo, la población de editores de libros o de criadores de pavos. Jessen (1955) presenta un método interesante para construir un marco a partir de las ramas de un árbol frutal.

Selección de la muestra

Existe, actualmente, una gran variedad de planes para seleccionar una muestra. Por cada plan considerado, se pueden hacer, *grosso modo*, estimaciones del tamaño de la muestra, partiendo de un conocimiento del nivel de precisión deseado. Los costos relativos y el tiempo empleado para cada plan también se comparan antes de tomar una decisión.

La encuesta piloto

Es de gran utilidad probar el cuestionario y los métodos de campo en pequeña escala. Esto casi siempre da por resultado mejoras al cuestionario y puede evitar otros problemas que serían serios a mayor escala, por ejemplo, que el costo fuera mucho mayor que el esperado.

Organización del trabajo de campo

En encuestas extensas se encuentran muchos problemas de orden administrativo. El personal debe recibir un entrenamiento sobre el propósito de la encuesta y los métodos de medición que se emplearán. Además, se debe supervisar adecuadamente su trabajo. Un procedimiento de verificación *anticipado* sobre la calidad de las respuestas es de incalculable valor. Se deben hacer planes para manejar las no-respuestas, es decir, la falla del enumerador para obtener la información de ciertas unidades de la muestra.

Resumen y análisis de los datos

El primer paso después de realizar la encuesta es el editar los cuestionarios obtenidos, con la esperanza de corregir errores o cuando menos desechar los datos que obviamente están equivocados. Habrá necesidad de tomar ciertas decisiones respecto al procedimiento de cálculo en los casos de omisión de respuestas de quienes responden o de eliminación de datos en el proceso de edición. Después se realizarán los cálculos que conduzcan a las estimaciones. Puede haber diferentes métodos de estimación para los mismos datos.

Una práctica aconsejable en la presentación de los datos es informar la magnitud esperada de error en las estimaciones más importantes. Una de las ventajas del muestreo de probabilidad es que se pueden hacer tales enunciados (de error esperado) aunque deberán ser muy calificados si la cantidad de no-respuestas es importante.

Información conseguida para encuestas futuras

Cuanta más información de una población se tenga inicialmente, más fácil será el diseño de una muestra que proporcione estimaciones exactas. Toda muestra obtenida es una guía potencial de futuros muestreos, por los datos que revela sobre las medias, las desviaciones estándar, y la naturaleza de la variabilidad de las medidas principales, así como sobre los costos de obtención de datos. La práctica de muestreo avanzará más rápidamente si se prevé lo necesario para reunir y registrar este tipo de información.

Hay otro aspecto importante en el que una muestra completa facilita la obtención de otras posteriores Las cosas nunca resultan como se planearon para la obtención de una muestra compleja. Un muestreador hábil aprende a reconocer los errores de ejecución y a evitar que se repitan.

1.4 EL PAPEL DE LA TEORIA DEL MUESTREO

El objeto de encontrar los pasos para una encuesta por muestreo es recalcar que el muestreo es un negocio práctico y exige muchas y diversas habilidades. En algunos pasos, como por ejemplo en la definición de la población, en la determinación de los datos a recoger y de los métodos de medición, y en la organización del trabajo de campo, poco o nada tiene que ver la teoría del muestreo. Aunque estos asuntos no se discutirán en el resto del libro, hay que tener presente su importancia. El muestreo requiere atención en todas las fases de la actividad: un trabajo mediocre en una de ellas puede arruinar toda la encuesta.

El propósito de la teoría del muestreo es que éste sea más eficiente. Su objetivo es desarrollar métodos de selección de muestras y de estimación, que proporcionen, al menor costo posible, estimaciones con la suficiente exactitud para nuestros propósitos. Este principio de exactitud específica a costo mínimo aparece una y otra vez en la presentación de la teoría.

Para aplicar este principio, debemos ser capaces de predecir en cualquier método de muestreo que se considere, la precisión y el costo esperados. Respecto a la precisión, no podremos predecir cuál será el error de una estimación en una situación específica, porque esto implicaría el conocimiento del verdadero valor de la población. En lugar de ello, la precisión de un procedimiento de muestreo se juzga al examinar la distribución de frecuencia generada para las estimaciones, suponiendo que el proceso de muestreo se aplica varias veces a la misma población. Desde luego, ésta es la técnica estándar con la que se juzga la precisión en la teoría estadística.

Una simplificación adicional que podemos hacer, consiste en suponer, lo que es razonable en la práctica si se trata de muestras de tamaño común, que las estimaciones de muestra tienen una distribución aproximadamente normal. Con una estimación distribuida en forma normal se conoce la distribución de frecuencias, si son conocidas la media y la desviación estándar o la varianza. Una parte considerable de la teoría del muestreo se ocupa de encontrar fórmulas para estas medias y varianzas.

Hay dos diferencias entre la teoría estándar de encuestas por muestreo y la teoría clásica del muestreo como aparece en los libros de estadística matemática. En la teoría clásica, las mediciones hechas sobre las unidades de muestreo de la población suele suponerse que siguen una distribución de frecuencia de forma matemática conocida, como sería la distribución normal, cuyos parámetros, media y varianza, por ejemplo, se estimarían a partir de los datos de las muestras. Por otro lado, en la teoría de las encuestas por muestreo, se supone que sólo se dispone de una información muy limitada sobre dicha distribución, y sobre todo, no se supone conocida su forma matemática, así que el enfoque se puede describir como independiente de un modelo o de una distribución de frecuencia. Esta es una actitud natural para encuestas muy grandes en las que se efectúan numerosas mediciones diferentes de las unidades que siguen diversas distribuciones de frecuencia. Para las encuestas en las que sólo se realizan pocas mediciones en cada unidad, el estudio de sus distribuciones de frecuencia puede justificar la hipótesis de que son de forma matemática conocida, lo que permite la aplicación de la teoría clásica.

Otra diferencia es que las poblaciones en una encuesta tienen un número finito de unidades. Los resultados son ligeramente más complicados cuando el muestreo es de una población finita y no de una infinita. Por razones prácticas, a menudo se ignoran estas diferencias en los resultados para poblaciones finitas e infinitas. Los casos en que no se ignoren, los indicaremos.

1.5 MUESTREO PROBABILISTA

Los procedimientos de muestreo considerados en este libro comparten las siguientes propiedades matemáticas.

1. Podemos definir el conjunto de muestras distintas $S_1, S_2, \ldots, S_v$, que el procedimiento es capaz de elegir si se aplica a una población específica. Esto significa que podemos decir con precisión cuáles son las unidades de muestreo que pertenecen a S_1, S_2, etc.

Supongamos, por ejemplo, que la población consta de seis unidades, numeradas de 1 a 6. Un procedimiento común para elegir una muestra de tamaño 2 ofrece tres posibilidades $S_1 \sim (1, 4)$; $S_2 \sim (2.5)$; $S_3 \sim (3, 6)$. Nótese que no se incluyen todas las posibles muestras de tamaño 2.

2. Cada muestra posible S_i tiene asignada una probabilidad de selección π_i.

3. Se selecciona una de las S_i por un proceso aleatorio, en el que cada S_i tiene una probabilidad π_i de ser elegida. En el ejemplo anterior, podríamos asignar la misma probabilidad a cada muestra. Posteriormente, la selección se podría realizar al elegir un número aleatorio entre 1 y 3. Si el número es j, se toma la muestra S_j.

4. El método para calcular la estimación a partir de la muestra debe ser definido y debe conducir a una estimación única para cualquier muestra específica. Podemos decir, por ejemplo, que la estimación es el promedio de las mediciones correspondientes a las unidades individuales de la muestra.

Para todo procedimiento de muestreo que satisfaga estas condiciones, podemos calcular la distribución de frecuencia de las estimaciones que genera el proceso, si se aplica repetidamente a la misma población. Sabemos la frecuencia con que se elige cualquier muestra S_i, y sabemos cómo calcular la estimación a partir de los datos de S_i. Por lo tanto, es claro que se puede desarrollar una teoría de muestreo para cada procedimiento de este tipo, aunque los detalles del desarrollo puedan ser intrincados. Un método de esta clase se conoce con el nombre de *muestreo probabilista*.

En la práctica, rara vez se extrae una muestra de probabilidad dando las S_i y los números π_i como se esbozó anteriormente. Es un trabajo muy laborioso para una gran población, en la cual un procedimiento de muestreo puede producir billones de muestras posibles. Por lo general, la extracción se hace al especificar probabilidades de inclusión en la muestra para las unidades individuales y extraer unidades, una a la vez, o en grupos, hasta constituir la muestra del tamaño y tipo deseado. Desde el punto de vista teórico, basta saber que si quisiéramos, podríamos especificar las subcolecciones S_i y las π_i, siempre y cuando tuviéramos tiempo ilimitado para hacerlo.

1.6 ALTERNATIVAS AL MUESTREO PROBABILISTA

A continuación veremos algunos tipos comunes de muestreo no probabilistas.

1. La muestra es una parte de la población fácilmente accesible. Una muestra de carbón en un vagón abierto se puede tomar a 15 o 20 cm de la parte superior.

2. La muestra se selecciona a la ventura. Al tomar diez conejos de una jaula en un laboratorio, el investigador puede sacar aquellos que alcance con la mano, sin una planeación consciente.

3. Con una población pequeña pero heterogénea, el investigador inspecciona la totalidad de ésta y selecciona una pequeña muestra de unidades "típicas", es decir, unidades que a su parecer están cercanas al promedio de la población. Este método algunas veces es llamado de *juicio* o de *selección intencional*.

4. La muestra consta esencialmente de voluntarios, en estudios en los cuales el proceso de medición es desagradable o penoso para la persona que está siendo investigada.

En condiciones adecuadas cualquiera de estos métodos puede dar resultados útiles. Sin embargo, no son los indicados para el desarrollo de una teoría de muestreo, ya que no involucran ningún elemento aleatorio en el procedimiento de selección. Casi la única manera de examinar qué tan bueno puede ser uno de los métodos es encontrar una situación en la cual los resultados sean conocidos, ya sea para la población total o para una muestra basada en probabilidades y posteriormente hacer una comparación. Pero aun así, si un método resulta adecuado en la comparación, puede ser inadecuado al variar las condiciones.

En relación con lo anterior (muestreos probabilistas y no probabilistas) señalamos que algunos de los primeros usos del muestreo que hicieron los gobiernos a nivel urbano o nacional, desde 1850, tenían como finalidad reducir los costos de las estimaciones de los resultados de un censo. Para los tributos más importantes del censo, se calculaban los totales de la ciudad o el país, aprovechando completamente la información obtenida del censo. Para las mediciones restantes se tomaba una muestra que variaba entre el 15 y el 25% de los formularios devueltos después del censo, con objeto de aligerar el trabajo de estimación de los totales de la ciudad o el país para dichas mediciones. Se usaron dos métodos diferentes para la selección de la muestra. El primero, llamado de *selección aleatoria* es una aplicación del muestreo de probabilidad según el cual cada unidad de la población (cada formulario devuelto después del censo) tiene la misma oportunidad de ser incluido en la muestra. Para este método se constató que, como se dijo anteriormente, con ayuda de la teoría del muestreo y la distribución normal es posible predecir, en forma aproximada y con los datos de la muestra, la magnitud del error esperado en las estimaciones hechas

a partir de la muestra. Más aún, en los ítems más importantes para los que se disponía de los datos completos que suministró el censo, es posible, hasta cierto grado verificar la exactitud de las predicciones.

El segundo método es la *selección a propósito*. Esta no se definía en detalle, pero presentaba dos rasgos comunes. La unidad de muestreo consistía en grupos de formularios, a menudo relativamente grandes. Por ejemplo, en el censo italiano de 1921, el país tenía 8 354 comunas, agrupadas en 214 distritos. Al extraer una muestra del 14%, los estadísticos italianos Gini y Galvani seleccionaron a propósito 29 distritos, en lugar de 1 250 comunas. En segundo lugar, los 29 distritos se eligieron de tal modo que la muestra proporcionara estimaciones exactas para 7 importantes variables de control, cuyos resultados eran conocidos para todo el país, y con esto se esperaba que la muestra diera buenas estimaciones de las otras variables altamente correlacionadas con las de control.

En la década de los 20, el Instituto Estadístico Internacional nombró una comisión para estudiar las ventajas y desventajas de los dos métodos. El informe de Jensen (1926) pareció favorecer el método de selección a propósito. Sin embargo, este método se abandonó relativamente pronto como técnica de muestreo para obtener estimaciones nacionales en las encuestas que medían muchas variables, ya que carecía de la flexibilidad que ofrecían los métodos de muestreo probabilista desarrollados posteriormente, no podía predecir a partir de la muestra la exactitud esperada en las estimaciones y usaba unidades de muestreo demasiado grandes. Gini y Galvani concluyeron que el método de muestreo denominado muestreo aleatorio estratificado (Cap. 5) que utiliza la comuna como unidad de muestreo, habría dado mejores resultados que su método.

1.7 USO DE LA DISTRIBUCION NORMAL

En ocasiones es útil usar la palabra **estimador** para designar la regla por la cual se calcula alguna característica μ de la población, a partir de los resultados de la muestra, y la palabra *estimación* para el valor obtenido de una muestra específica. Un estimador $\hat{\mu}$ de μ dado por un plan de muestreo se llama insesgado si el valor medio de $\hat{\mu}$, tomado sobre todas las muestras posibles proporcionadas por el plan, es igual a μ. En la notación de la Sec. 1.5, puede escribirse esta condición

$$E(\hat{\mu}) = \sum_{i=1}^{v} \pi_i \hat{\mu}_i = \mu$$

donde $\hat{\mu}_i$ es la estimación dada por la i-ésima muestra. El símbolo E, que significa "valor esperado de", es de uso común.

Como se mencionó en la Sec. 1.4, las muestras en algunas encuestas son tan grandes que siguen una distribución aproximadamente normal. Además, con el muestreo probabilista, tenemos fórmulas que dan la media y la varianza de las estimaciones. Supongamos que hemos tomado una muestra por un proceso que se sabe que proporciona un estimador insesgado y que hemos calculado la estimación de muestra $\hat{\mu}$ y su desviación estándar $\sigma_{\hat{\mu}}$ (que suele llamarse su error estándar). ¿Qué tan buena es esta estimación? No podemos conocer el valor exacto del error de estimación $(\hat{\mu} - \mu)$ pero por las propiedades de la curva normal, puede esperarse que

0.32 (en 1 caso de 3) el error absoluto $|\hat{\mu} - \mu|$ excede $\sigma_{\hat{\mu}}$
0.05 (en 1 caso de 20) el error absoluto $|\hat{\mu} - \mu|$ excede $1.96\sigma_{\hat{\mu}} \doteq 2\sigma_{\hat{\mu}}$
0.01 (en 1 caso de 100) el error absoluto $|\hat{\mu} - \mu|$ excede $2.58\sigma_{\hat{\mu}}$

Por ejemplo, si una muestra probabilista de los registros de acumuladores de uso rutinario en una fábrica muestra una vida promedio $\hat{\mu} = 394$ días, con error estándar de $\sigma_{\hat{\mu}} = 4.6$ días; se espera que en 99 de 100 casos la vida promedio en la población de baterías se sitúe entre

$$\hat{\mu} = 394 - (2.58)(4.6) = 382 \text{ días}$$

y

$$\hat{\mu}_U = 394 + (2.58)(4.6) = 406 \text{ días}$$

Los límites, 382 días y 406 días, se llaman *límites de confianza* inferior y superior. Con una sola estimación de una encuesta, la afirmación "μ está entre 382 y 406 días" no es correcta con certeza. La magnitud "99% de confianza" implica que si se usara el mismo plan de muestreo muchas veces en una población, haciendo un enunciado de confianza a partir de cada muestra, el 99% (aprox.) serían correctas y 1% estarían equivocados. Cuando el muestreo ha sido introducido en una operación en la que previamente se haya usado un censo completo, una demostración de esta propiedad algunas veces se hace sacando muestras repetidas del tipo propuesto de una población para la cual exista un registro completo, de tal manera que μ sea conocido (véase, por ejemplo, Trueblood y Cyert, 1957). La verificación práctica de que aproximadamente la proporción establecida de aseveraciones es correcta, es útil para educar y tranquilizar a los administradores respecto a la naturaleza del muestreo. De modo semejante, cuando se toma una sola muestra de cada una de una serie de poblaciones diferentes, cerca de un 95% de las aseveraciones de confianza al 95% son correctas.

La discusión precedente presupone que $\sigma_{\hat{\mu}}$, según se calcula de la muestra, es conocido exactamente. En realidad, tanto $\sigma_{\hat{\mu}}$, como $\hat{\mu}$, están sujetos a error de muestreo. Con una variable normalmente distribuida, se usan las tablas de distribución t de Student, en lugar de las tablas normales para calcular los límites de confianza para μ cuando la muestra es pequeña. El reemplazo de la tabla normal por la tabla de t casi no hace diferencia si el número de grados de libertad en $\sigma_{\hat{\mu}}$ excede a 50. Con ciertos tipos de muestreo estratificado y con el método de muestreo repetido (Sec. 11.19) los grados de libertad son pequeños y la tabla t es necesaria.

1.8 EL SESGO Y SUS EFECTOS

En la teoría de las encuestas por muestreo es necesario considerar los estimadores sesgados por 2 razones:

1. En algunos de los problemas más comunes, particularmente en la estimación de razones, se encuentra que los estimadores convenientes y apropiados son sesgados.

2. Aun con los estimadores que son insesgados en muestreo probabilista, los errores de medición y las no-respuestas pueden producir sesgos en los números que se calculan a partir de los datos. Esto sucede, por ejemplo, cuando casi todas las personas que se niegan a ser entrevistadas se oponen a cierto gasto de los fondos públicos, en tanto que aquellas que han sido entrevistadas se dividen igualmente en pro y en contra.

Para examinar el efecto del sesgo, supóngase que la estimación $\hat{\mu}$ está normalmente distribuida alrededor de una media m que está a una distancia B del verdadero valor μ de la población, como se ve en la Fig. 1.1. La magnitud del sesgo es $B = m - \mu$. Supongamos que desconocemos la existencia del sesgo. Calculamos la desviación estándar σ de la distribución de frecuencias del estimador —esto será, desde luego, la desviación estándar respecto a la media m de la distribución, no respecto a la verdadera media μ—. Estamos usando σ en lugar de $\sigma_{\hat{\mu}}$. Como una afirmación acerca de la exactitud de la estimación, declaramos que la probabilidad es de 0.05 de que el estimador $\hat{\mu}$ esté en error por más de 1.96σ.

Consideraremos ahora cómo la presencia del sesgo distorsiona esta probabilidad. Para hacer esto calculamos la verdadera probabilidad de que la estimación está en error por más de 1.96σ, en donde el error es medido a partir de la verdadera media μ. Ambos lados de la distribución deben ser examinados separadamente. Para el ex-

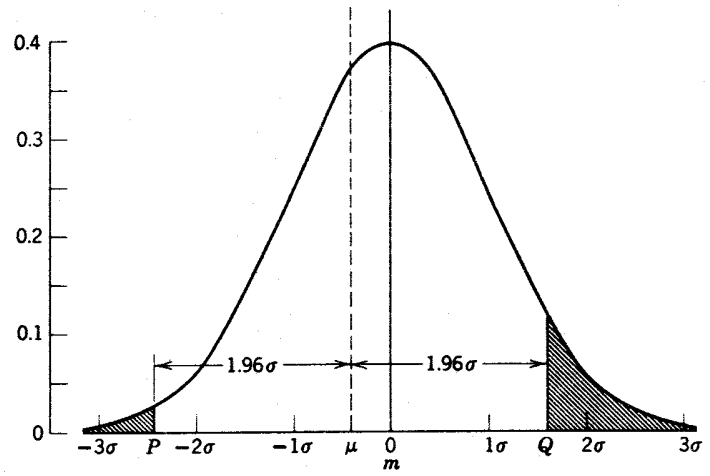


FIG. 1-1. Efecto del sesgo en los errores de estimación

tremo derecho, la probabilidad de un error de más de $+ 1.96\sigma$ es el área sombreada a partir de Q en la Fig. 1-1. Esta área está dada por

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{\mu+1.96\sigma}^{\infty} e^{-(\hat{\mu}-m)^2/2\sigma^2}\,d\hat{\mu}$$

Sea $\hat{\mu} - m = \sigma t$. El límite inferior del intervalo de integración para t es

$$\frac{\mu-m}{\sigma}+1.96=1.96-\frac{B}{\sigma}$$

Por lo que el área es:

$$\frac{1}{\sqrt{2\pi}}\int_{1.96-(B/\sigma)}^{\infty} e^{-t^2/2}\,dt$$

De manera semejante, el extremo izquierdo, es decir, el área sombreada a la izquierda a partir de P, tiene una área de:

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-1.96-(B/\sigma)} e^{-t^2/2}\,dt$$

Al observar las integrales, está claro que la perturbación depende únicamente de la razón del sesgo a la desviación estándar. Los resultados se muestran en la Tabla 1.1.

En el caso de considerar la probabilidad total de cometer un error de más de 1.96σ, el sesgo tiene poco efecto, siempre y cuando

TABLA 1.1. EFECTO DE UN SESGO B EN LA PROBABILIDAD DE COMETER UN ERROR MAYOR QUE 1.96σ

	Probabilidad de error		
B/σ	$< -1.96\sigma$	$>1.96\sigma$	Total
0.02	0.0238	0.0262	0.0500
0.04	0.0228	0.0274	0.0502
0.06	0.0217	0.0287	0.0504
0.08	0.0207	0.0301	0.0508
0.10	0.0197	0.0314	0.0511
0.20	0.0154	0.0392	0.0546
0.40	0.0091	0.0594	0.0685
0.60	0.0052	0.0869	0.0921
0.80	0.0029	0.1230	0.1259
1.00	0.0015	0.1685	0.1700
1.50	0.0003	0.3228	0.3231

sea menor que una décima de la desviación estándar. En este punto, la probabilidad total es de 0.0511 en lugar de ser 0.05 que pensamos que es. Conforme aumenta más este sesgo, la perturbación se vuelve más seria. Para $B = \sigma$, la probabilidad de error es de 0.17, más de 3 veces el valor supuesto.

Los dos lados de la distribución se afectan de una manera diferente. Con un sesgo positivo, como en el ejemplo, la probabilidad de una sobreestimación más allá de 1.96σ disminuye con rapidez del valor supuesto de 0.025 para volverse despreciable cuando $B = \sigma$. La probabilidad de la sobreestimación correspondiente se eleva en forma constante. En la mayoría de las aplicaciones el error total es de interés primordial, pero en ocasiones estamos particularmente interesados en los errores cometidos en una sola dirección.

Como una regla de trabajo, el efecto del sesgo en la exactitud de un estimador es despreciable si el sesgo es menor que un décimo de la desviación estándar de la estimación. Si tenemos un método de estimación que produce sesgo para el cual $B/\sigma < 0.1$, en donde B es el valor absoluto del sesgo, se puede argumentar que el sesgo no es una desventaja del método. Aun con $B/\sigma = 0.2$, es modesta la perturbación en la probabilidad del error.

En el uso de estos resultados se deben distinguir las dos fuentes de sesgo mencionadas al principio de esta sección. Con sesgos que surgen en la estimación de proporciones, un límite superior a la proporción B/σ se puede determinar matemáticamente. Si la muestra es lo bastante grande, podemos tener la confianza de que B/σ no excederá a 0.1. Por otro lado, con sesgos causados por errores de medición o de no-respuesta, generalmente es imposible encontrar el límite superior para B/σ que es pequeño. Este enfadoso problema se discute en el Cap. 13.

1.9 EL ERROR CUADRATICO MEDIO

A fin de comparar un estimador sesgado, con un estimador insesgado, o bien, dos estimadores que tienen cantidades diferentes de sesgo, un criterio útil es el del error cuadrático medio (ECM) del estimador, medido a partir del valor de la población que se está estimando. Formalmente,

$$\begin{aligned}\text{ECM}(\hat{\mu}) &= E(\hat{\mu}-\mu)^2 = E[(\hat{\mu}-m)+(m-\mu)]^2 \\ &= E(\hat{\mu}-m)^2+2(m-\mu)E(\hat{\mu}-m)+(m-\mu)^2 \\ &= (\text{variación de } \hat{\mu}) + (\text{sesgo})^2\end{aligned}$$

El doble producto desaparece, ya que $E(\hat{\mu} - m) = 0$.

El uso del ECM, como un criterio de la exactitud de un estimador, equivale a considerar dos estimadores que tienen el mismo ECM como equivalentes. Esto no es estrictamente correcto porque las distribuciones de frecuencia de los errores $(\hat{\mu} - \mu)$, de tamaños diferentes no serán las mismas para los dos estimadores si tienen cantidades diferentes de sesgo. Sin embargo, ha sido demostrado por Hansen, Hurwitz y Madow (1953) que si B/σ es aproximadamente menor un medio, las dos distribuciones de frecuencia son casi idénticas en relación con los errores *absolutos* $|\hat{\mu} - \mu|$ de magnitudes diferentes. La Tabla 1.2 ilustra este resultado.

TABLA 1.2. PROBABILIDAD DE UN ERROR ABSOLUTO $\geq 1\sqrt{\text{ECM}}$, $1.96\sqrt{\text{ECM}}$ Y $2.576\sqrt{\text{ECM}}$

	Probabilidad		
B/σ	$1\sqrt{\text{ECM}}$	$1.96\sqrt{\text{ECM}}$	$2.576\sqrt{\text{ECM}}$
0	0.317	0.0500	0.0100
0.2	0.317	0.0499	0.0100
0.4	0.319	0.0495	0.0095
0.6	0.324	0.0479	0.0083

Aun si $B/\sigma = 0.6$, los cambios en las probabilidades al compararlas con los valores obtenidos para $B/\sigma = 0$ son ligeros.

Debido a la dificultad de poder asegurar que ningún sesgo insospechado se introduce en las estimaciones, generalmente hablaremos de la *precisión* de un estimador más que de su *exactitud*. La exactitud se refiere a la magnitud de las desviaciones respecto a la media verdadera μ mientras que precisión se refiere a la magnitud de las desviaciones respecto a la media m obtenida por la aplicación repetida del procedimiento de muestreo.

EJERCICIOS

1.1. Suponga que usa el muestreo para estimar el número total de palabras de un libro que contiene ilustraciones.

(a) ¿Hay algún problema para la definición de la población? (b) ¿Cuáles son los pros y los contras de: (1) la página, (2) el renglón, como unidad de muestreo?

1.2. Se va a tomar una muestra de una lista de nombres que están en tarjetas (un nombre por tarjeta) numeradas consecutivamente, las cuales se encuentran en un archivo. Cada nombre tendrá la misma oportunidad de ser incluido en la muestra. ¿Qué problemas surgen de las siguientes situaciones comunes? (a) Algunos de los nombres no pertenecen a la población-objeto, a pesar de que este hecho no puede ser verificado para ningún nombre hasta que no se seleccione. (b) Algunos nombres aparecen en más de una tarjeta. Todas las tarjetas con el mismo nombre llevan números consecutivos y, por lo tanto, aparecen juntas en el archivo. (c) Algunos nombres aparecen en más de una tarjeta, pero las que llevan el mismo nombre pueden estar colocadas en cualquier lugar dentro del archivo.

1.3. El problema para encontrar un marco completo que permita la obtención de una muestra es frecuentemente un obstáculo. ¿Qué clase de marcos pudieran ser convenientes para las siguientes encuestas? ¿Tienen los marcos alguna deficiencia seria? (a) Una encuesta de tiendas que venden petacas en una gran ciudad. (b) Una encuesta de artículos que se dejan en los trenes subterráneos o autobuses. (c) Una encuesta de las personas mordidas por víboras durante el último año.

1.4. Un directorio de la ciudad, de hace cuatro años, enlista las direcciones en orden a lo largo de cada calle, y da los nombres de las personas que viven en cada dirección. Para una encuesta que se lleva a cabo actualmente por medio de entrevistas a la gente de la ciudad, ¿cuáles son las deficiencias de este marco? ¿Pueden ser remediadas por los entrevistadores durante el desarrollo de la encuesta? Al usar el directorio, ¿sacaría usted una lista de direcciones (domicilios) o una lista de personas?

1.5. En una estimación para muestreo del valor real de objetos pequeños en el inventario de una gran firma, el valor real y el registrado en los libros se obtuvieron para cada objeto de la muestra. Para la muestra total, la razón del valor real al registrado fue de 1.021; esta estimación está distribuida aproximadamente de una manera normal con un error estándar de 0.0082. Si el valor del inventario según los libros de registro es de \$80 000, calcule los límites de confianza del 95% para el valor real.

1.6. Frecuentemente los datos se deben tratar como una muestra, a pesar de que a primera vista parece que constituyen una enumeración completa. El propietario de un lote de estacionamiento encuentra que el negocio disminuye los domingos por la mañana. Después de 26 domingos de operación, su percepción promedio por domingo es exactamente \$10.00. El error estándar de esta cifra, calculado a partir de las variaciones de semana a semana, es de \$1.20. Los costos del cuidador son de \$7.00 cada domingo. El propietario quiere mantener abierto el lote los domingos por la mañana siempre y cuando su utilidad esperada fuera de \$5.00. ¿Cuál es la probabilidad de que la utilidad a largo plazo sea de, al menos, \$5.00? ¿Qué suposición se debe hacer para responder a esta pregunta?

1.7. En la Tabla 1.2, ¿qué pasaría a la probabilidad de excederse en $1\sqrt{\text{ECM}}$, $1.96\sqrt{\text{ECM}}$ y $2.576\sqrt{\text{ECM}}$ cuando B/σ tiende a ser infinita, es decir, cuando el ECM es debido enteramente a sesgo? ¿Concuerdan sus resultados con la dirección de los cambios anotados en la Tabla 1.2 conforme B/σ varíe de 0 a 0.6?

1.8. Cuando es necesario comparar dos estimaciones que tienen diferentes distribuciones de frecuencia de errores $(\hat{\mu} - \mu)$, ocasionalmente es posible, en problemas especializados, calcular el costo o la pérdida que resultaría de un error $(\hat{\mu} - \mu)$ de cualquier tamaño dado. Se prefiere la estimación que da la pérdida esperada más pequeña, siendo el resto de las condiciones iguales. Demuestre que si la pérdida es una función cuadrática $\lambda(\hat{\mu} - \mu)^2$ del error, debemos escoger la estimación con el error cuadrático medio menor.

CAPITULO 2

Muestreo Aleatorio Simple

2.1 MUESTREO ALEATORIO SIMPLE

El *muestreo aleatorio simple* es un método de selección de n unidades en un conjunto de N de tal modo que cada una de las ${}_NC_n$ muestras distintas tengan la misma oportunidad de ser elegidas. En la práctica, un muestreo aleatorio se realiza unidad por unidad. Se numeran las unidades de 1 a N. Posteriormente se extrae una serie de n números aleatorios entre 1 y N, ya sea utilizando una tabla de números aleatorios o mediante un programa de computación que produce una tabla semejante. En cada extracción, el proceso debe otorgar la misma oportunidad de selección a todos y cada uno de los números que *no hayan salido*. Las unidades que llevan estos n números constituyen la muestra.

Fácilmente se verifica que todas las ${}_NC_n$ muestras distintas, tienen la misma oportunidad de ser extraídas por este método. Considérese una muestra determinada, es decir, una colección de n unidades especificadas. En la primera extracción, la probabilidad de que se seleccione una de estas n unidades es n/N. En la segunda, la probabilidad que se extraiga una de las restantes $(n-1)$ unidades especificadas es $(n-1)/(N-1)$, y así sucesivamente. Por lo tanto, la probabilidad de que se extraigan las n unidades especificadas es

$$\frac{n}{N}\cdot\frac{(n-1)}{(N-1)}\cdot\frac{(n-2)}{(N-2)}\cdots\frac{1}{(N-n+1)}=\frac{n!(N-n)!}{(N)!}=\frac{1}{{}_NC_n} \qquad (2.1)$$

Como en todas las extracciones subsecuentes se descarta un número extraído, este método también se llama muestreo aleatorio *sin restitución*. El muestreo aleatorio *con restitución* es perfectamente

factible: en cada extracción todos los N miembros de la población reciben la misma oportunidad de extracción, sin que importe el número de veces que se extrajeron antes. Las fórmulas de varianzas y varianzas estimadas de las estimaciones realizadas a partir de la muestra son a menudo más simples cuando el muestreo es con restitución que en el caso contrario. Por esta razón, se utiliza el muestreo con restitución en los planes de muestreo más complicados, aunque a primera vista parece inútil tener dos o más veces la misma unidad dentro de la muestra.

2.2 SELECCION DE UNA MUESTRA ALEATORIA SIMPLE

Las tablas de números aleatorios son tablas de los dígitos 0, 1, 2, . . . 9, donde cada dígito tiene la misma probabilidad de ser seleccionado en cada extracción. Entre las tablas más grandes están las que publicó la Rand Corporation (1955) con un millón de dígitos y las de Kendall y Smith (1938) con 100 000 dígitos. En los textos estándar de estadística se pueden encontrar muchas tablas. La Tabla 2.1 muestra 1 000 dígitos aleatorios y para ilustración, de Snedecor y Cochran (1967).

Cuando se usan estas tablas para seleccionar una muestra aleatoria simple, el primer paso es la enumeración de las unidades de la población de 1 a N. Si el primer dígito de N es un número entre 5 y 9, el siguiente método de selección es el adecuado. Supóngase $N = 528$ y queremos $n = 10$. Tómense tres columnas de la Tabla 2.1, digamos la 25, 26 y 27. Recórrase hacia abajo cada columna seleccionando los 10 primeros números *distintos*, entre 001 y 528. Estos son 36, 509, 364, 417, 348, 127, 149, 186, 290 y 162. Para los dos últimos números saltamos a las columnas 30 a 32. En selecciones repetidas es aconsejable cambiar el punto de partida en la tabla.

La desventaja de este método es que los números de tres dígitos 000 y todos los del 529 a 999 no se usan aunque al saltarse números no se pierde mucho tiempo. Cuando el primer dígito de N es menor que 5, algunos pueden preferir este método si n es pequeño y se dispone de una tabla de números aleatorios bastante grande.

Con $N = 128$, por ejemplo, un segundo método con menos rechazos y de más fácil aplicación es el siguiente. En una serie de números de tres dígitos se substrae 200 de todos los números que hay entre 201 y 400, se substrae 400 de todos los números entre 401 y 600, 600 de todos los números entre 601 y 800, 800 de todos los

TABLA 2.1. UN MILLAR DE DÍGITOS ALEATORIOS

	00–04	05–09	10–14	15–19	20–24	25–29	30–34	35–39	40–44	45–49
00	54463	22662	65905	70639	79365	67382	29085	69831	47058	08186
01	15389	85205	18850	39226	42249	90669	96325	23248	60933	26927
02	85941	40756	82414	02015	13858	78030	16269	65978	01385	15345
03	61149	69440	11286	88218	58925	03638	52862	62733	33451	77455
04	05219	81619	10651	67079	92511	59888	84502	72095	83463	75577
05	41417	98326	87719	92294	46614	50948	64886	20002	97365	30976
06	28357	94070	20652	35774	16249	75019	21145	05217	47286	76305
07	17783	00015	10806	83091	91530	36466	39981	62481	49177	75779
08	40950	84820	29881	85966	62800	70326	84740	62660	77379	90279
09	82995	64157	66164	41180	10089	41757	78258	96488	88629	37231
10	96754	17676	55659	44105	47361	34833	86679	23930	53249	27083
11	34357	88040	53364	71726	45690	66334	60332	22554	90600	71113
12	06318	37403	49927	57715	50423	67372	63116	48888	21505	80182
13	62111	52820	07243	79931	89292	84767	85693	73947	22278	11551
14	47534	09243	67879	00544	23410	12740	02540	54440	32949	13491
15	98614	75993	84460	62846	59844	14922	48730	73443	48167	34770
16	24856	03648	44898	09351	98795	18644	39765	71058	90368	44104
17	96887	12479	80621	66223	86085	78285	02432	53342	42846	94771
18	90801	21472	42815	77408	37390	76766	52615	32141	30268	18106
19	55165	77312	83666	36028	28420	70219	81369	41943	47366	41067

números entre 801 y 999 y desde luego 000 de todos los números entre 000 y 200. Todos los residuos mayores que 129 y los números 000, 200, etc., se desechan. Al utilizar las columnas 05 a 07 de la Tabla 2.1 obtenemos 26, 52, 7, 94, 16, 48, 41, 80, 128 y 92, la extracción que requiere 15 números de tres dígitos para $n = 10$. En esta muestra la tasa de rechazo $5/15 = 33\%$ es cercana a la probabilidad de rechazo $72/200 = 36\%$, para este método. Al usar este método con número N como 384, nótese que uno substrae 400 de un número entre 401 y 800, pero automáticamente desecha todos los números mayores a 800. La substracción de 800 de los números entre 801 y 999 daría una mayor probabilidad de aceptación de residuos entre 001 y 199 que de los residuos entre 200 y 384.

Hay otros método de muestreo que se prefieren al muestreo aleatorio simple por razones de conveniencia o de mayor precisión. El muestreo aleatorio simple es el más adecuado para una introducción a la teoría del muestreo.

2.3 DEFINICIONES Y NOTACION

En una encuesta por muestreo elegimos ciertas propiedades que intentamos medir y registrar para cada unidad que venga dentro de la muestra. Estas propiedades de las unidades se llaman *características* o simplemente, *atributos*.

Los valores obtenidos para cualquier atributo específico en las N unidades que comprenden la población, se representan por y_1, y_2, ..., y_N. Los valores correspondientes para las unidades en muestreo se denotan por y_1, y_2, ..., y_n, o, si queremos referirnos a un miembro de la muestra en general, por y_i $(i = 1, 2, \ldots, n)$. Nótese que la muestra no consistirá de las *primeras* n unidades en la población, excepto en el caso, usualmente raro, en que se seleccionen precisamente estas unidades. Si este punto no se olvida, de acuerdo a mi experiencia, no habrá confusión.

Las letras *mayúsculas* se refieren a las características de la *población* y las *minúsculas* a las de la *muestra*. Para los totales y las medias tenemos las siguientes definiciones:

	Población	Muestra
Total:	$Y = \sum^{N} y_i = y_1 + y_2 + \cdots + y_N$	$\sum^{n} y_i = y_1 + y_2 + \cdots + y_n$
Media:	$\bar{Y} = \frac{y_1 + y_2 + \cdots + y_N}{N} = \frac{\sum^{N} y_i}{N}$	$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{\sum^{n} y_i}{n}$

Aun cuando el muestreo se realiza con muchos propósitos, el interés se centra, con frecuencia, en cuatro características de la población.

1. Media = $\bar{Y}$ (por ejemplo, el número promedio de niños por escuela).

2. Total = Y (por ejemplo, el número total de hectáreas de trigo en una región).

3. La proporción de dos totales o medias $R = Y/X = \bar{Y}/\bar{X}$ (por ejemplo, la razón del activo neto al activo total en un grupo de familias).

4. La proporción de unidades que caen dentro de alguna clase definida (por ejemplo, proporción de personas con dientes postizos).

En este capítulo se discute la estimación de las primeras tres cantidades.

El símbolo $\hat{}$ identifica una estimación muestral de una característica de la población. En este capítulo sólo se consideran los estimadores más simples:

	Estimadores
Media de la población $\bar{Y}$	$\hat{\bar{Y}} = \bar{y}$ = media muestral
Total de la población Y	$\hat{Y} = N\bar{y} = N\sum^{n} y_i/n$
Razón de la población R	$\hat{R} = \bar{y}/\bar{x} = \sum^{n} y_i \Big/ \sum^{n} x_i$

En $\hat{Y}$ al factor N/n que multiplica al total muestral, algunas veces se llama el factor de *expansión, elevación* o *inflación*. Su inverso n/N, la razón del tamaño de la muestra respecto a la población, se llama *fracción de muestreo* y se identifica con la letra f.

2.4 PROPIEDADES DE LAS ESTIMACIONES

La precisión de cualquier estimación basada en una muestra depende del método por el cual se calcula la estimación a partir de los datos de la muestra, y del plan de muestreo. Para ahorrar espacio nos hemos referido algunas veces a "la precisión de la media muestral" o "la precisión del muestreo aleatorio simple", sin mencionar específicamente el otro factor fundamental. Esto sólo ha ocurrido en los casos donde es claro, en el contexto, cuál es el factor que falta. Al estudiar cualquier fórmula que se presenta, el lector debe asegurarse que conoce el método específico de muestreo y el método de estimación para el cual se ha establecido la fórmula.

En este libro, un método de estimación se llama *consistente* si la estimación se vuelve exactamente igual al valor de la población cuando $n = N$, o sea, cuando la muestra está constituida por la población total. Para el muestreo aleatorio simple, es obvio que $\bar{y}$ y $N\bar{y}$ son estimaciones consistentes de la media de la población y del total, respectivamente. La consistencia es una propiedad deseable de los estimadores. Por otro lado, un estimador inconsistente no necesariamente es inútil, ya que puede dar una precisión satisfactoria cuando n, comparado con N, es pequeño. Es posible que su utilidad esté confinada a esta situación Hansen, Hurwitz y Madow (1953) y Murthy (1967) dan otra definición de la consistencia semejante a la de la estadística clásica. Un estimador es consistente, si la probabilidad de que sea erróneo en una cantidad dada, o en una mayor

tiende a cero al crecer la muestra. Un enunciado exacto de esta definición requiere cierto cuidado para planes de encuesta complejos.

Como hemos visto, un método de estimación es *insesgado* si el valor promedio de la estimación, que se toma sobre todas las muestras posibles de un tamaño n dado, es exactamente igual al valor verdadero de la población. Si el método va a ser insesgado, este resultado debe ser cierto para cualquier población de valores finitos y_i y para cualquier n. Para investigar si $\bar{y}$ es insesgado con el muestreo aleatorio simple, calculamos el valor de $\bar{y}$ para todas las ${}_NC_n$ muestras y buscamos el promedio de las estimaciones. Con el símbolo E se indica este promedio sobre todas las muestras posibles.

Teorema 2.1. La media muestral $\bar{y}$ es un estimador insesgado de $\bar{Y}$.

Prueba: Por definición,

$$E\bar{y} = \frac{\sum \bar{y}}{{}_NC_n} = \frac{\sum (y_1 + y_2 + \cdots + y_n)}{n[N!/n!(N-n)!]} \tag{2.2}$$

En donde la suma se extiende sobre todas las ${}_NC_n$ muestras. Para evaluar esta suma, buscamos en cuántas muestras aparece un valor específico cualquiera, y_i. Como están disponibles otras $(N-1)$ unidades para el resto de la muestra y otros $(n-1)$ lugares para completar la muestra, el número de muestras que contienen y_i es

$${}_{N-1}C_{n-1} = \frac{(N-1)!}{(n-1)!(N-n)!} \tag{2.3}$$

De donde:

$$\sum (y_1 + y_2 + \cdots + y_n) = \frac{(N-1)!}{(n-1)!(N-n)!}(y_1 + y_2 + \cdots + y_N)$$

Sustituyendo en (2.2) tenemos:

$$E\bar{y} = \frac{(N-1)!}{(n-1)!(N-n)!}\,\frac{n!(N-n)!}{nN!}(y_1 + y_2 + \cdots + y_N)$$

$$= \frac{(y_1 + y_2 + \cdots + y_N)}{N} = \bar{Y} \tag{2.4}$$

Corolario. $\hat{Y} = N\bar{y}$ es un estimador insesgado del total de la población Y.

Una prueba menos complicada del Teorema 2.1 es como sigue: como cada unidad aparece en el mismo número de muestras, está claro que

$E(y_1 + y_2 + \cdots + y_n)$ debe ser un múltiplo de $y_1 + y_2 + \cdots + y_N$ (2.5)

El multiplicador debe ser n/N, ya que la expresión de la izquierda tiene n términos y la de la derecha tiene N términos. Esto conduce al resultado.

2.5 VARIANZAS DE LAS ESTIMACIONES

La varianza de los valores y_i en una población finita usualmente se define como

$$\sigma^2 = \frac{\sum_1^N (y_i - \bar{Y})^2}{N} \tag{2.6}$$

Por motivos de notación, los resultados se presentan en términos de una expresión ligeramente diferente, en donde el divisor $(N-1)$ se usa en lugar de N. Entonces tenemos:

$$S^2 = \frac{\sum_1^N (y_i - \bar{Y})^2}{N-1} \tag{2.7}$$

Esta convención la usan quienes enfocan la teoría del muestreo por medio del análisis de la varianza. Su ventaja es que la mayoría de los resultados toman una forma ligeramente más simple. Siempre y cuando se use consistentemente la misma notación, todos los resultados son equivalentes en ambas notaciones. Consideremos ahora la varianza de $\bar{y}$. Por esto queremos decir $E(\bar{y} - \bar{Y})^2$ tomada sobre todas las ${}_NC_n$ muestras.

Teorema 2.2. La varianza de la media $\bar{y}$ de una muestra aleatoria simple es

$$V(\bar{y}) = E(\bar{y} - \bar{Y})^2 = \frac{S^2}{n}\,\frac{(N-n)}{N} = \frac{S^2}{n}(1-f) \tag{2.8}$$

en donde $f = n/N$ es la fracción de muestreo.

Prueba:

$$n(\bar{y} - \bar{Y}) = (y_1 - \bar{Y}) + (y_2 - \bar{Y}) + \cdots + (y_n - \bar{Y}) \tag{2.9}$$

Por el argumento de simetría usado en la relación (2.5), se tiene que

$$E[(y_1 - \bar{Y})^2 + \cdots + (y_n - \bar{Y})^2] = \frac{n}{N}[(y_1 - \bar{Y})^2 + \cdots + (y_N - \bar{Y})^2] \tag{2.10}$$

y también que

$$E[(y_1-\bar{Y})(y_2-\bar{Y})+(y_1-\bar{Y})(y_3-\bar{Y})+\cdots+(y_{n-1}-\bar{Y})(y_n-\bar{Y})]$$
$$=\frac{n(n-1)}{N(N-1)}[(y_1-\bar{Y})(y_2-\bar{Y})+(y_1-\bar{Y})(y_3-\bar{Y})$$
$$+\cdots+(y_{N-1}-\bar{Y})(y_N-\bar{Y})] \qquad (2.11)$$

En (2.11) la suma de los productos se extiende sobre todas las parejas de unidades en la muestra y en la población, respectivamente. La suma del lado izquierdo contiene $n(n-1)/2$ términos y la suma de la derecha $N(N-1)/2$ términos.

Ahora, elevando al cuadrado (2.9) y promediando sobre todas las muestras aleatorias simples. Usando (2.10) y (2.11), obtenemos

$$n^2E(\bar{y}-\bar{Y})^2=\frac{n}{N}\Big\{(y_1-\bar{Y})^2+\cdots+(y_N-\bar{Y})^2$$
$$+\frac{2(n-1)}{N-1}[(y_1-\bar{Y})(y_2-\bar{Y})+\cdots+(y_{N-1}-\bar{Y})(y_N-\bar{Y})]\Big\}$$

Completando el cuadrado en el término que incluye el doble producto, tenemos

$$n^2E(\bar{y}-\bar{Y})^2=\frac{n}{N}\Big\{\Big(1-\frac{n-1}{N-1}\Big)[(y_1-\bar{Y})^2+\cdots+(y_N-\bar{Y})^2]$$
$$+\frac{n-1}{N-1}[(y_1-\bar{Y})+\cdots+(y_N-\bar{Y})]^2\Big\}$$

El segundo término dentro de la llave desaparece, ya que la suma de y_i es igual a $N\bar{Y}$. La división entre n^2 da

$$V(\bar{y})=E(\bar{y}-\bar{Y})^2=\frac{N-n}{nN(N-1)}\sum_{i=1}^{N}(y_i-\bar{Y})^2=\frac{S^2}{n}\frac{(N-n)}{N}$$

Corolario 1. El error estándar de $\bar{y}$ es

$$\sigma_{\bar{y}}=\frac{S}{\sqrt{n}}\sqrt{(N-n)/N}=\frac{S}{\sqrt{n}}\sqrt{1-f} \qquad (2.12)$$

Corolario 2. La varianza de $\hat{Y}=N\bar{y}$, como un estimador del total de la población Y, es

$$V(\hat{Y})=E(\hat{Y}-Y)^2=\frac{N^2S^2}{n}\frac{(N-n)}{N}=\frac{N^2S^2}{n}(1-f) \qquad (2.13)$$

Corolario 3. El error estándar de $\hat{Y}$ es

$$\sigma_{\hat{Y}}=\frac{NS}{\sqrt{n}}\sqrt{(N-n)/N}=\frac{NS}{\sqrt{n}}\sqrt{1-f} \qquad (2.14)$$

2.6 LA CORRECCION POR POBLACION FINITA

Para una muestra al azar de tamaño n de una población infinita, se sabe que la varianza de la media es σ^2/n. El único cambio en este resultado, cuando la población es finita, es la introducción del factor $(N-n)/N$. Los factores $(N-n)/N$ para la varianza y $\sqrt{(N-n)/N}$ para el error estándar se denominan *correcciones debidas a población finita* o *corrección por finitud* (cpf). Algunos autores que presentan los resultados en términos de σ dan estas correcciones con un divisor $(N-1)$ en lugar de N. Siempre y cuando la fracción n/N sea pequeña, estos factores toman valores cercanos a la unidad y el tamaño de la población como tal no tiene un efecto directo en el error estándar de la media de la muestra. Por ejemplo, si S es la misma en dos poblaciones, una muestra de 500 de una población de 200 000 da una estimación de la media de la población casi tan preciso como una muestra de 500 de una población de 10 000. Las personas que no están familiarizadas con el muestreo, con frecuencia encuentran este resultado muy difícil de creer y, desde luego es asombroso. Para ellas es obvio que, si la información se obtuvo únicamente con una pequeña fracción de la población, la media de la muestra no puede ser exacta. Es conveniente que el lector considere por qué este punto de vista es erróneo.

En la práctica, el cpf se puede ignorar siempre y cuando la fracción de muestreo no exceda un 5% y para muchos propósitos, aun si es tan alto como el 10%. El efecto de ignorar la corrección es sobreestimar el error estándar de la estimación $\bar{y}$.

El siguiente teorema, que es una extensión del Teorema 2.2, no se necesita para la discusión de este capítulo, pero se prueba aquí para referencia posterior.

Teorema 2.3. Si y_i, x_i son dos variables definidas en toda unidad de la población, y $\bar{y}$, $\bar{x}$ son las medias derivadas de una muestra aleatoria simple tamaño n, entonces su *covarianza* es:

$$E(\bar{y}-\bar{Y})(\bar{x}-\bar{X})=\frac{N-n}{nN}\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X}) \qquad (2.15)$$

Este teorema se reduce al Teorema 2.2 si las variables y_i, x_i son iguales en cada unidad.

Prueba. Aplicando el Teorema 2.2 a la variable $u_i = y_i + x_i$. La media de la población de u_i es $\bar{U} = \bar{Y} + \bar{X}$, y por el Teorema 2.2, tenemos:

$$E(\bar{u} - \bar{U})^2 = \frac{N-n}{nN} \frac{1}{N-1} \sum_{i=1}^{N} (u_i - \bar{U})^2$$

o sea,

$$E[(\bar{y} - \bar{Y}) + (\bar{x} - \bar{X})]^2 = \frac{N-1}{nN} \frac{1}{N-1} \sum_{i=1}^{N} [(y_i - \bar{Y}) + (x_i - \bar{X})]^2 \qquad (2.16)$$

Desarrollando los términos cuadráticos en ambos lados. Por el Teorema 2.2,

$$E(\bar{y} - \bar{Y})^2 = \frac{N-n}{nN} \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2$$

con una relación similar para $E(\bar{x} - \bar{X})^2$. Por lo tanto, estos dos términos se cancelan en ambos lados de la Ec. (2.16). El resultado del teorema (Ec. 2.15) se deriva de los dobles productos.

2.7. ESTIMACION DEL ERROR ESTANDAR A PARTIR DE UNA MUESTRA

Las fórmulas para los errores estándar de la estimación de la media y total de la población se usan principalmente para tres propósitos: (1) para comparar la precisión obtenida por el muestreo aleatorio simple con otros métodos de muestreo, (2) para estimar el tamaño de la muestra que se necesita en una encuesta que esté siendo planeada, (3) para estimar la precisión realmente obtenida en una encuesta que se haya terminado. Las fórmulas involucran a S^2, la varianza de la población. Esta no será conocida en la práctica pero puede ser estimada de los datos de la muestra. El resultado importante se establece en el Teorema 2.4.

Teorema 2.4. En una muestra aleatoria simple,

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

es una estimación insesgada de

$$S^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N-1}$$

Prueba. Podemos escribir:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} [(y_i - \bar{Y}) - (\bar{y} - \bar{Y})]^2 \qquad (2.17)$$

$$= \frac{1}{n-1} \left[\sum_{i=1}^{n} (y_i - \bar{Y})^2 - n(\bar{y} - \bar{Y})^2 \right] \qquad (2.18)$$

Ahora, tomando el promedio sobre todas las muestras simples al azar de tamaño n. Por el argumento de simetría usado en el Teorema 2.2,

$$E\left[\sum_{i=1}^{n} (y_i - \bar{Y})^2 \right] = \frac{n}{N} \sum_{i=1}^{N} (y_i - \bar{Y})^2 = \frac{n(N-1)}{N} S^2$$

por definición de S^2. Además, por el Teorema 2.2,

$$E[n(\bar{y} - \bar{Y})^2] = \frac{N-n}{N} S^2$$

Por lo tanto,

$$E(s^2) = \frac{S^2}{(n-1)N} [n(N-1) - (N-n)] = S^2 \qquad (2.19)$$

Corolario. Las estimaciones insesgadas de las varianzas de $\bar{y}$ y $\hat{Y} = N\bar{y}$ son

$$v(\bar{y}) = s_{\bar{y}}^2 = \frac{s^2}{n} \left(\frac{N-n}{N} \right) = \frac{s^2}{n} (1-f) \qquad (2.20)$$

$$v(\hat{Y}) = s_{\hat{Y}}^2 = \frac{N^2 s^2}{n} \left(\frac{N-n}{N} \right) = \frac{N^2 s^2}{n} (1-f) \qquad (2.21)$$

Para los errores estándar tomamos

$$s_{\bar{y}} = \frac{s}{\sqrt{n}} \sqrt{1-f}, \qquad s_{\hat{Y}} = \frac{Ns}{\sqrt{n}} \sqrt{1-f} \qquad (2.22)$$

Estas estimaciones son ligeramente sesgadas: para la mayoría de las aplicaciones, este sesgo no tiene importancia.

El lector debe fijarse en los símbolos empleados para representar las varianzas verdaderas y estimadas de las respectivas estimaciones. Por lo tanto, para $\bar{y}$ escribimos

varianza verdadera: $V(\bar{y}) = \sigma_{\bar{y}}^2$

varianza estimada: $v(\bar{y}) = s_{\bar{y}}^2$

2.8 LIMITES DE CONFIANZA

Generalmente se presupone que las estimaciones $\bar{y}$ y $\hat{Y}$ se distribuyen en forma normal alrededor del valor correspondiente de la población. Las razones para esta suposición y sus limitaciones se consideran en la Sec. 2.15. Si la suposición es verdadera, los límites de confianza superior e inferior para la media y total de la población son como sigue:

Media:

$$\hat{\bar{Y}}_L = \bar{y} - \frac{ts}{\sqrt{n}}\sqrt{1-f}, \qquad \hat{\bar{Y}}_U = \bar{y} + \frac{ts}{\sqrt{n}}\sqrt{1-f} \tag{2.23}$$

Total:

$$\hat{Y}_L = N\bar{y} - \frac{tNs}{\sqrt{n}}\sqrt{1-f}, \qquad \hat{Y}_U = N\bar{y} + \frac{tNs}{\sqrt{n}}\sqrt{1-f} \tag{2.24}$$

El símbolo t es el valor del desvío normal correspondiente a la probabilidad de confianza deseada. Los valores más comunes son:

Probabilidad de confianza (%)	50	80	90	95	99
	0.67	1.28	1.64	1.96	2.58

Si el tamaño de la muestra es menor que 50, los puntos de porcentaje se pueden tomar de la tabla t de Student con $(n-1)$ grados de libertad, siendo estos, los grados de libertad en la varianza estimada s^2. La distribución t se ajusta exactamente sólo si las observaciones de y_i están normalmente distribuidas y N es infinito. Discrepancias moderadas de la normalidad no afectan grandemente. Para muestras pequeñas con distribuciones muy asimétricas se necesitan métodos especiales.

Ejemplo. Se recabaron firmas para una petición en 676 hojas. Cada hoja tenía espacio suficiente para 42 firmas pero en muchas de las hojas se recabó un número menor. Se contó el número de firmas por hoja en una muestra al azar de 50 hojas (muestra del 7% aproximadamente); los resultados se muestran en la Tabla 2.2.

Estimar el número total de firmas para la petición y los límites de confianza al 80%.

La unidad de muestreo es una hoja y las observaciones, y_i, son el número de firmas por hoja. Como aproximadamente la mitad de las hojas tenían el número máximo de firmas (42), los datos se presentan como una distribución de frecuencia. Nótese que la distribución original difiere bastante de la normal, estando la frecuencia mayor en el extremo superior. No obstante, hay una razón para creer, basados en la experiencia, que las medias de muestras de 50 observaciones se distribuyen aproximadamente conforme a la normal.

Encontramos

$$n = \sum f_i = 50, \qquad y = \sum f_i y_i = 1471, \qquad \sum f_i y_i^2 = 54\,497$$

Por lo tanto, la estimación del número total de firmas es

$$\hat{Y} = N\bar{y} = \frac{(676)(1471)}{50} = 19\,888$$

La varianza s^2 de la muestra es

$$s^2 = \frac{1}{n-1}[\sum f_i(y_i - \bar{y})^2] = \frac{1}{n-1}\left[\sum f_i y_i^2 - \frac{(\sum f_i y_i)^2}{\sum f_i}\right]$$

$$= \frac{1}{49}\left[54\,497 - \frac{(1471)^2}{50}\right] = 229.0$$

Utilizando la expresión (2.22) los límites de confianza al 80% son

$$19\,888 \pm \frac{tNs}{\sqrt{n}}\sqrt{1-f} = 19\,888 \pm \frac{(1.28)(676)(15.13)\sqrt{1-0.0740}}{\sqrt{50}}$$

Lo anterior da como resultado 18 107 y 21 669 para los límites al 80%. Una numeración completa demostró que había 21 045 firmas.

TABLA 2.2

TABLA 2.2. RESULTADOS PARA UNA MUESTRA DE 50 HOJAS DE PETICION y_i = NUMERO DE FIRMAS: f_i = FRECUENCIA

y_i	42	41	36	32	29	27	23	19	16	15	
f_i	23	4	1	1	1	2	1	1	2	2	
y_i	14	11	10	9	7	6	5	4	3		Total
f_i	1	1	1	1	1	3	2	1	1		50

2.9 UN METODO ALTERNATIVO DE PRUEBA

Cornfield (1944) sugirió un método para comprobación de los resultados obtenidos de un muestreo aleatorio simple sin restitución, que nos permite el uso de resultados estándar de la teoría de poblaciones infinitas. Sea a_i una variable aleatoria que toma el valor de 1 si la i-ésima unidad está en la muestra y el valor de 0 si no lo está. La media $\bar{y}$ de la muestra se puede escribir

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{N} a_i y_i \tag{2.25}$$

donde la suma se extiende sobre todas las unidades N en la población. En esta expresión, las a_i son variables aleatorias y las y_i son un conjunto de números fijos.

Claramente,

$$\Pr(a_i = 1) = \frac{n}{N}, \qquad \Pr(a_i = 0) = 1 - \frac{n}{N}$$

Por lo tanto, a_i se distribuye como una variable binomial unitaria con $P = n/N$. Por lo que

$$E(a_i) = P = \frac{n}{N}, \qquad V(a_i) = PQ = \frac{n}{N}\left(1 - \frac{n}{N}\right) \tag{2.26}$$

Para encontrar $V(\bar{y})$ necesitamos también la covarianza de a_i y a_j. El producto $a_i\, a_j$ es 1 si las unidades i-ésima y j-ésima están en la muestra y es 0 de otra manera. La probabilidad de que dos unidades específicas estén ambas en la muestra es $n(n-1)/N(N-1)$. Por lo tanto,

$$\begin{aligned}\operatorname{Cov}(a_i a_j) &= E(a_i a_j) - E(a_i)E(a_j) \\ &= \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = -\frac{n}{N(N-1)}\left(1 - \frac{n}{N}\right)\end{aligned} \tag{2.27}$$

Usando este enfoque para encontrar $V(\bar{y})$, tenemos, de la expresión (2.25),

$$V(\bar{y}) = \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 V(a_i) + 2\sum_{i<j}^{N} y_i y_j \operatorname{Cov}(a_i a_j)\right] \tag{2.28}$$

$$= \frac{1-f}{nN}\left(\sum y_i^2 - \frac{2}{N-1}\sum y_i y_j\right) \tag{2.29}$$

usando (2.26) y (2.27). Completando el cuadrado en el término del doble producto, tenemos

$$V(\bar{y}) = \frac{1-f}{nN}\left(\frac{N}{N-1}\sum y_i^2 - \frac{1}{N-1}Y^2\right) \tag{2.30}$$

$$= \frac{1-f}{n(N-1)}\sum (y_i - \bar{Y})^2 = \frac{(1-f)S^2}{n} \tag{2.31}$$

Este método ofrece un mecanismo para probar fácilmente los Teoremas 2.3 y 2.4. Se puede usar para encontrar momentos de orden superior en la distribución de $\bar{y}$, aunque para este propósito un método dado por Tukey (1950), con un desarrollo adicional por Wishart (1952), es más poderoso.

2.10 MUESTREO ALEATORIO CON RESTITUCION

Un enfoque semejante se aplica cuando el muestreo es con restitución. En este caso la unidad i-ésima puede aparecer 0, 1, 2, . . . , n veces en la muestra. Sea t_i el número de veces que la i-ésima unidad aparece en la muestra, entonces:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{N} t_i y_i \tag{2.32}$$

Como la probabilidad de que la i-ésima unidad esté presente en la muestra es $1/N$ para cada unidad muestreada, la variable t_i se distribuye como un número binomial de éxitos en n experimentos con $p = 1/N$. Por lo que:

$$E(t_i) = \frac{n}{N}, \qquad V(t_i) = n\left(\frac{1}{N}\right)\left(1 - \frac{1}{N}\right) \tag{2.33}$$

Conjuntamente, las variables t_i siguen una distribución multinomial, por lo tanto:

$$\operatorname{Cov}(t_i t_j) = -\frac{n}{N^2} \tag{2.34}$$

Utilizando las Ecs. (2.32), (2.33) y (2.34), tenemos para el muestreo con restitución,

$$V(\bar{y}) = \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 \frac{n(N-1)}{N^2} - 2\sum_{i<j}^{N} y_i y_j \frac{n}{N^2}\right] \tag{2.35}$$

$$= \frac{1}{nN}\sum_{i=1}^{N} (y_i - \bar{Y})^2 = \frac{\sigma^2}{n} = \frac{N-1}{N}\frac{S^2}{n} \tag{2.36}$$

En consecuencia, $V(\bar{y})$ en un muestreo sin restitución, es solamente $(N-n)/(N-1)$ veces su valor en el muestreo con restitución. Si en vez de $\bar{y}$ la media $\bar{y}_d$ de las unidades diferentes o distintas de la muestra se usa como estimación cuando el muestreo sea con restitución, Murthy (1967) ha mostrado que el término guía en la varianza promedio de $\bar{y}_d$ es $(1 - f/2)S^2/n$, con base en el trabajo de Basu (1968) y de Des Raj y Khamis (1958). En algunas aplicaciones, el costo de la medición de unidades distintas de la muestra puede predominar, así que el costo de la muestra es proporcional al número de unidades distintas. En esta situación, Seth y J. N. K. Rao (1964) mostraron que para un costo promedio dado, $V(\bar{y})$ en el muestreo sin restitución es menor que $V(\bar{y}_d)$ en el muestreo con restitución. También probaron que es válido el resultado más general en el que si $\bar{y}_{d'} = f(\nu)\bar{y}_d/Ef(\nu)$, donde ν es el número

de unidades distintas de la muestra y $f(\nu)$ es una función de ν, entonces $V(\bar{y}) < V(\bar{y}_{d'})$ si $S^2 < N\bar{Y}^2$, condición que satisfacen casi todas las poblaciones encontradas en las encuestas por muestreo.

2.11 ESTIMACION DE UNA RAZON

Frecuentemente la cantidad que va a ser estimada a partir de una muestra aleatoria simple es la razón de dos variables, ambas variando de unidad a unidad. En una encuesta de familias podríamos obtener el número promedio de trajes por hombre adulto, el promedio de gastos en cosméticos por mujer adulta y el promedio de número de horas por semana que la televisión es vista por niños de edades entre 10 y 15 años. Con el objeto de estimar la primera de estas características, registraremos para la i-ésima familia ($i = 1, 2, \ldots, n$) el número de hombres adultos x_i que viven en la i-ésima unidad habitacional y el número total de trajes y_i que poseen estas personas. El parámetro de la población a ser estimado es la razón.

$$R = \frac{\text{Número total de trajes}}{\text{Número total de hombres adultos}} = \frac{\sum_1^N y_i}{\sum_1^N x_i} \tag{2.37}$$

la estimación correspondiente de muestreo es

$$\hat{R} = \frac{\sum_1^n y_i}{\sum_1^n x_i} = \frac{\bar{y}}{\bar{x}} \tag{2.38}$$

Ejemplos de este tipo ocurren frecuentemente cuando la unidad de muestreo (la familia) comprende un grupo o conglomerado de elementos (hombres adultos) y nuestro interés está en la media de la población *por elemento*. Las razones también aparecen en otras aplicaciones, por ejemplo, la razón de préstamos para propósitos de construcción a los préstamos totales de un banco o la razón de hectáreas de trigo en relación a las hectáreas totales en una propiedad.

La distribución derivada del muestreo de $\hat{R}$ es más complicada que la de $\bar{y}$ porque ambos, el numerador $\bar{y}$ y el denominador $\bar{x}$ varían de muestra a muestra. En muestras pequeñas la distribución de $\hat{R}$ es asimétrica y $\hat{R}$ por lo común es una estimación ligeramente sesgada de R. En muestras grandes, la distribución de $\hat{R}$ tiende a la

normalidad y el sesgo se vuelve despreciable. Los siguientes resultados aproximados servirán para la mayoría de las aplicaciones; la distribución de $\hat{R}$ se estudia con mayor detalle en el Cap. 6.

Teorema 2.5. Si las variables y_i, x_i son medidas en cada unidad de una muestra aleatoria simple de tamaño n, supuestamente grande, el (ECM) y la varianza de $\hat{R} = \bar{y}/\bar{x}$ son, aproximadamente,

$$\text{ECM}(\hat{R}) \doteq V(\hat{R}) \doteq \frac{1-f}{n\bar{X}^2} \frac{\sum_{i=1}^N (y_i - Rx_i)^2}{N-1} \tag{2.39}$$

en donde $R = \bar{Y}/\bar{X}$ es la razón de las medias de la población y $f = n/N$

Prueba:

$$\hat{R} - R = \frac{\bar{y}}{\bar{x}} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}} \tag{2.40}$$

Si n es grande, $\bar{x}$ no debe diferir mucho de $\bar{X}$. Para anular teniendo resuelta la distribución de la razón de dos aleatorias variables $(\bar{y} - R\bar{x})$ y $\bar{x}$, la aproximación consiste en reemplazar $\bar{x}$ por $\bar{X}$ en el denominador de (2.40). Esto da

$$\hat{R} - R \doteq \frac{\bar{y} - R\bar{x}}{\bar{X}} \tag{2.41}$$

Ahora, el promedio sobre todas las muestras aleatorias simples de tamaño n es

$$E(\hat{R} - R) \doteq \frac{E(\bar{y} - R\bar{x})}{\bar{X}} = \frac{\bar{Y} - R\bar{X}}{\bar{X}} = 0 \tag{2.42}$$

puesto que $R = \bar{Y}/\bar{X}$. Esto demuestra que para el orden de aproximación usado aquí $\hat{R}$ es un estimador insesgado de R.

Utilizando la Ec. (2.41) también obtenemos el resultado

$$\text{ECM}(\hat{R}) = E(\hat{R} - R)^2 \doteq \frac{1}{\bar{X}^2} E(\bar{y} - R\bar{x})^2 \tag{2.43}$$

La cantidad $\bar{y} - R\bar{x}$ es la media de la muestra de la variable $d_i = y_i - Rx_i$, cuya media de la población $\bar{D} = \bar{Y} - R\bar{X} = 0$. Por lo que podemos encontrar $V(\hat{R})$ aplicando el teorema 2.2 para la varianza de la media de una muestra aleatoria simple utilizando la variable d_i y dividiendo por X^2. Esto da

$$V(\hat{R}) = \frac{1}{\bar{X}^2} E(\bar{y} - R\bar{x})^2 = \frac{1}{\bar{X}^2} \frac{S_d^2}{n}(1-f) \tag{2.44}$$

$$= \frac{1-f}{n\bar{X}^2} \frac{\sum_{i=1}^{N} (d_i - \bar{D})^2}{(N-1)} = \frac{1-f}{n\bar{X}^2} \frac{\sum_{i=1}^{N} (y_i - Rx_i)^2}{N-1} \tag{2.45}$$

Lo anterior completa la prueba.

Vale la pena enfatizar la manera como se probó el Teorema 2.5. En el Teorema 2.2 se demostró que la fórmula para la varianza de la media muestral $\bar{y}$ da la fórmula para la varianza aproximada de la proporción $\bar{y}/\bar{x}$, si la variable y_i se reemplaza por la variable $(y_i - Rx_i)/\bar{X}$. Este mismo resultado, o su extensión natural, es cierto también en situaciones más complejas y se usa más adelante y con frecuencia en este libro.

Como una estimación muestral de

$$\frac{\sum_{i=1}^{N} (y_i - Rx_i)^2}{N-1}$$

se acostumbra tomar

$$\frac{\sum_{i=1}^{n} (y_i - \hat{R}x_i)^2}{n-1}$$

Se puede demostrar que esta estimación tiene un sesgo del orden de $1/n$.

Para el error estándar estimado de $\hat{R}$, se tiene:

$$s(\hat{R}) = \frac{\sqrt{1-f}}{\sqrt{n}\bar{X}} \sqrt{\frac{\sum (y_i - \hat{R}x_i)^2}{n-1}} \tag{2.46}$$

Si $\bar{X}$ no es conocida, la afirmación $\bar{x}$ de la muestra se sustituye en el denominador.

Una manera de calcular $s(\hat{R})$ consiste en expresarla como:

$$s(\hat{R}) = \frac{\sqrt{1-f}}{\sqrt{n}\bar{X}} \sqrt{\frac{\sum y_i^2 - 2\hat{R}\sum y_i x_i + \hat{R}^2 \sum x_i^2}{n-1}} \tag{2.47}$$

Ejemplo. La Tabla 2.3 muestra el número de personas (x_1), el ingreso semanal de la familia (x_2) y los gastos semanales en comida (y), en una muestra aleatoria simple de 33 familias de bajos ingresos. Como la muestra es pequeña, los datos sólo se proponen para ilustrar los cálculos.

De la muestra se estima (a) el promedio de gastos semanales en alimentos por familia, (b) el gasto semanal en alimentos por persona y (c) el porcentaje de ingreso que se gaste en alimento. Calcule los errores estándar de estas estimaciones.

Gasto semanal en alimentos por familia. Es la media muestral común y corriente:

$$\bar{y} = \frac{907.2}{33} = \$27.49$$

Por el Teorema 2.2 (ignorando el cpf), su error estándar es:

$$s_{\bar{y}} = \frac{1}{\sqrt{n}} \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \frac{1}{\sqrt{n(n-1)}} \sqrt{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}$$

$$= \frac{1}{\sqrt{(33)(32)}} \sqrt{28\,224 - (907.2)^2/33} = \$1.76$$

(La suma de cuadrados sin corregir: 28 224 se muestra en la parte inferior de la Tabla 2.3.)

TABLA 2.3. TAMAÑO, INGRESO SEMANAL Y COSTO DE ALIMENTOS DE 33 FAMILIAS

Núm. de familia	Tamaño x_1	Ingreso x_2	Costo de alimentos y	Núm. de familia	Tamaño x_1	Ingreso x_2	Costo de alimentos y
1	2	62	14.3	18	4	83	36.0
2	3	62	20.8	19	2	85	20.6
3	3	87	22.7	20	4	73	27.7
4	5	65	30.5	21	2	66	25.9
5	4	58	41.2	22	5	58	23.3
6	7	92	28.2	23	3	77	39.8
7	2	88	24.2	24	4	69	16.8
8	4	79	30.0	25	7	65	37.8
9	2	83	24.2	26	3	77	34.8
10	5	62	44.4	27	3	69	28.7
11	3	63	13.4	28	6	95	63.0
12	6	62	19.8	29	2	77	19.5
13	4	60	29.4	30	2	69	21.6
14	4	75	27.1	31	6	69	18.2
15	2	90	22.2	32	4	67	20.1
16	5	75	37.7	33	2	63	20.7
17	3	69	22.6				
				Total	123	2394	907.2

$\sum x_1^2 = 533$, $\sum x_2^2 = 177\,254$, $\sum y^2 = 28\,224$

$\sum x_1 y = 3595.5$, $\sum x_2 y = 66\,678$

Gasto semanal en alimento por persona. Como el tamaño de la familia varía, la estimación es una razón de dos variables,

$$\hat{R}_1 = \frac{\sum y}{\sum x_1} = \frac{907.2}{123} = \$7.38 \text{ por persona}$$

Las sumas de cuadrados y de productos necesarios para calcular $S(\hat{R})$ según la Ec. (2.47) se encuentran en la parte inferior de la Tabla (2.3). Necesitamos, además,

$$2\hat{R}_1 = 14.7512, \qquad \hat{R}_1^2 = 54.3996, \qquad \bar{x}_1 = 3.7273$$

Los decimales extra en $\hat{R}_1$, $2\hat{R}_1$, $\hat{R}_1^2$ se incluyen por exactitud. Entonces, utilizando la Ec. (2.47)

$$s(\hat{R}_1) = \frac{1}{\sqrt{33}(3.7273)}\sqrt{\frac{(28\,224) - (14.7512)(3595.5) + (54.3996)(533)}{32}}$$
$$= \$0.534$$

Porcentaje del ingreso gastado en alimentos. Esto nuevamente es una razón de dos variables

$$\hat{R}_2 = 100\frac{\sum y}{\sum x_2} = \frac{(100)(907.2)}{2394} = 37.9\%$$

Utilizando la Ec. (2.47) el lector puede verificar que el error estándar es 2.38%.

2.12 ESTIMACIONES DE MEDIAS EN SUBPOBLACIONES

En muchas encuestas, la población se subdivide en clases dentro de las cuales se realizan estimaciones diferentes. En una encuesta de familias, se puede desear hacer estimaciones por separado para familias con 0, 1, 2, . . . niños, para propietarios de casa e inquilinos o para familias en grupos de ocupación diferente. El término *dominios de estudio* ha sido dado a estas subpoblaciones por la subcomisión de muestreo de las Naciones Unidas (1950).

En la situación más simple, cada unidad en la población cae dentro de uno de los dominios. Consideremos que el j-ésimo dominio contiene N_j unidades y que n_j es el número de unidades, en una muestra aleatoria simple de tamaño n, que pertenecen a este dominio. Si $y_{jk}(k = 1, 2, \ldots, n_j)$ son las medidas u observaciones obtenidas en estas unidades, la media de población $\bar{Y}_j$ es estimada en el j-ésimo dominio por

$$\bar{y}_j = \sum_{k=1}^{n_j} \frac{y_{jk}}{n_j} \tag{2.48}$$

A primera vista, $\bar{y}_j$ parece ser una estimación de razón, descrita en la Sec. 2.11, puesto que, a pesar de que n es fija, n_j variará de una muestra de tamaño n a otra. Se puede evitar la complicación de usar una estimación de razón al considerar la distribución de $\bar{y}_j$ sobre muestras en donde tanto n como n_j son constantes. Sea $n_j > 0$.

En la totalidad de muestras con n y n_j constantes, la probabilidad de sacar un conjunto específico de n_j unidades de las N_j unidades en el dominio j está dada por

$$\frac{{}_{N-N_j}C_{n-n_j}}{{}_{N-N_j}C_{n-n_j} \cdot {}_{N_j}C_{n_j}} = \frac{1}{{}_{N_j}C_{n_j}}$$

Como cada conjunto específico de n_j unidades del j-ésimo dominio puede aparecer con todas las selecciones de $(n - n_j)$ unidades obtenidas de las $(N - n_j)$ que no están en el dominio j, el numerador de la expresión antes indicada es el número de muestras que contienen un conjunto específico de n_j y el denominador es el número total de muestras. Por lo tanto, los Teoremas 2.1, 2.2 y 2.4 se pueden aplicar usando las observaciones y_{jk}, si ponemos n_j en lugar de n y N_j por N.

Del Teorema 2.1: $\bar{y}_j$ es una estimación insesgada de $\bar{Y}_j$ (2.49)

Del Teorema 2.2: el error estándar $\bar{y}_j$ es $\frac{S_j}{\sqrt{n_j}}\sqrt{1-(n_j/N_j)}$ (2.50)

en donde

$$S_j^2 = \sum_{k=1}^{N_j} \frac{(y_{jk} - \bar{Y}_j)^2}{N_j - 1} \tag{2.51}$$

Del Teorema 2.4: una estimación del error estándar de $\bar{y}_j$ es

$$\frac{s_j}{\sqrt{n_j}}\sqrt{1-(n_j/N_j)} \tag{2.52}$$

en donde

$$s_j^2 = \sum_{k=1}^{n_j} \frac{(y_{jk} - \bar{y}_j)^2}{n_j - 1} \tag{2.53}$$

Si no se conoce el valor de N_j, se puede usar la fracción n/N en lugar de n_j/N_j, para calcular el cpf. (Con el muestreo aleatorio simple, n_j/N_j es una estimación insesgada de n/N.)

2.13 ESTIMACIONES DE TOTALES EN SUBPOBLACIONES

En una lista de facturas recuperables de una compañía, en donde ciertas facturas han sido pagadas y otras no, podríamos estimar por muestreo la cantidad total, en dólares, de facturas no pagadas. Si N_j (el número de facturas no pagadas en la población) es conocido, no hay problema. La estimación muestral es $N_j\bar{y}_j$ y su error estándar condicional es N_j veces la expresión (2.50).

Alternativamente, si la cantidad total recuperable es conocida, se puede usar una estimación de razón. Con la muestra se puede obtener una estimación de la razón (cantidad total de facturas no pagadas)/(cantidad total de las facturas). Esto se multiplica por la cantidad total conocida de facturas recuperables en la lista.

Si no se conocen N_j ni el total recuperable, estas estimaciones no se pueden obtener. En su lugar, multiplicamos el total muestral de las y, de unidades que están en el j-ésimo dominio, por el factor de expansión N/n. Esto da la estimación

$$\hat{Y}_j = \frac{N}{n}\sum_{k=1}^{n_j} y_{jk} \tag{2.54}$$

Demostraremos que $\hat{Y}_j$ es insesgada y obtendremos su error estándar considerando muestras repetidas de tamaño n. El recurso de mantener n_j fijo, al igual que n, no ayuda en este problema.

En la presentación de la prueba regresamos a la notación original, en donde y_i representa la medida de la i-ésima unidad en la población. Definiendo una nueva variable y'_i para cada unidad en la población, en donde

$$y_i' = \begin{cases} y_i \text{ si la unidad está en el } j\text{-ésimo dominio} \\ 0 \text{ si la unidad se encuentra fuera del } j\text{-ésimo dominio} \end{cases}$$

El total de la población de la y_i' es

$$\sum_{i=1}^{N} y_i' = \sum_{j\text{-ésimo dom.}} y_i = Y_j \tag{2.55}$$

En una muestra aleatoria simple de tamaño n, $y'_i = y_i$ para cada una de las n_j unidades que caen en el j-ésimo dominio; $y_i' = 0$ para cada una de las $n - n_j$ unidades restantes. Si $\bar{y}'$ es la medida ordinaria de la muestra de las y'_i, la cantidad

$$N\bar{y}' = \frac{N}{n}\sum_{i=1}^{n} y_i' = \frac{N}{n}\sum_{k=1}^{n_j} y_{jk} = \hat{Y}_j \tag{2.56}$$

Este resultado demuestra que la estimación $\hat{Y}_j$ como se define en la Ec. (2.54) es N veces la media de la muestra de las y'_i.

En muestras repetidas de tamaño n podemos aplicar claramente los Teoremas 2.1, 2.2 y 2.4 a la variable y_i'. Esto demuestra que $\hat{Y}_j$ es una estimación insesgada de $\hat{Y}_j$ con un error estándar.

$$\sigma(\hat{Y}_j) = \frac{NS'}{\sqrt{n}}\sqrt{1-(n/N)} \tag{2.57}$$

en donde S' es la desviación estándar de las y_i' en la población. A fin de calcular S', visualizamos a la población como consistente de los N_j valores y_i que están en el j-ésimo dominio y de $N - N_j$ valores 0. De esta manera,

$$S'^2 = \frac{1}{N-1}\left(\sum_{j\text{-ésimo dom.}} y_i^2 - \frac{Y_j^2}{N}\right) \tag{2.58}$$

Aplicando el Teorema 2.4, una estimación del error estándar de $\hat{Y}_j$ derivado de una muestra es

$$s(\hat{Y}_j) = \frac{Ns'}{\sqrt{n}}\sqrt{1-(n/N)} \tag{2.59}$$

Al calcular s', se le da un valor 0 a cualquier unidad que no está en el j-ésimo dominio. Algunos estudiantes parecen tener una objeción psicológica para hacer esto, pero el método es bueno.

Los métodos presentados en esta sección y la sección precedente, también se aplican a encuestas en donde el marco usado contiene unidades que no pertenecen a la población definida. Un ejemplo ilustra esta situación.

Ejemplo. De una lista de 2 422 gastos familiares menores, se sacó una muestra aleatoria simple de 180 observaciones a fin de estimar el gasto total de operación de la casa. Ciertos tipos de gastos (ropa y mantenimiento del automóvil) no se consideraron pertinentes. De las 180 observaciones muestrales, 152 fueron pertinentes. La suma y la suma de cuadrados sin corregir de las cantidades pertinentes (en dólares) fueron como sigue:

$$\sum y_i' = 343.5, \qquad \sum y_i'^2 = 1491.38$$

El gasto total de operación familiar estimado y el error estándar de la estimación son:

$$\hat{Y}_j = \frac{N}{n}\sum_{i=1}^{n} y_i' = \frac{(2422)(343.5)}{180} = \$4622$$

De la Ec. (2.59).

$$s(\hat{Y}_j) = \frac{Ns'}{\sqrt{n}}\sqrt{1-(n/N)}$$

Al calcular s' consideramos nuestra muestra de 180 observaciones con 28 ceros. Por lo tanto,

$$s'^2 = \frac{1}{(179)}\left[\sum y_i'^2 - \frac{(\sum y_i')^2}{180}\right]$$

$$= \frac{1}{(179)}\left[1491.38 - \frac{(343.5)^2}{180}\right] = 4.670$$

Finalmente,

$$s_{\hat{Y}_j} = (2422)\sqrt{\frac{4.670}{180}\left(1 - \frac{180}{2422}\right)} = \$375$$

La estimación no es precisa, su coeficiente de variación, 375/4622, es aproximadamente 8%.

En este ejemplo, los gastos de mantenimiento del automóvil y los gastos de ropa fueron considerados como no pertinentes y, por lo tanto, con valor 0 en la muestra. En algunas aplicaciones se conoce por adelantado que ciertas unidades en la población no contribuyen en nada al total que está siendo estimado. Por ejemplo, en una inspección de tiendas donde se trata de estimar la venta total de petacas, algunas tiendas no venden petacas; en estudios de explotaciones agrícolas, ciertas unidades, en muestreo de áreas, no contienen explotaciones agrícolas en absoluto. Algunas veces es posible, con un poco de esfuerzo, identificar y contar las unidades que no contribuyen en nada, así en nuestra notación $(N - N_j)$, y por tanto, N_j es conocido.

Por consiguiente, vale la pena examinar en cuánto se reduce $V(\hat{Y}_j)$ cuando N_j es conocido. Si N_j no es conocido, entonces, la Ec. (2.57) queda

$$V(\hat{Y}_j) = \frac{N^2 S'^2}{n}\left(1 - \frac{n}{N}\right)$$

Si $\bar{Y}_j$ y S_j son la media y la desviación estándar en el dominio de interés (es decir, obtenidas con las unidades que no son cero) el lector puede verificar que

$$(N-1)S'^2 = (N_j - 1)S_j^2 + N_j\bar{Y}_j^2\left(1 - \frac{N_j}{N}\right) \tag{2.60}$$

Como los términos $1/N_j$ y $1/N$ son casi siempre despreciables,

$$S'^2 \doteq P_j S_j^2 + P_j Q_j \bar{Y}_j^2 \tag{2.61}$$

en donde $P_j = N_j/N$ y $Q_j = 1 - P_j$. Esto da

$$V(\hat{Y}_j) \doteq \frac{N^2}{n}(P_j S_j^2 + P_j Q_j \bar{Y}_j^2)\left(1 - \frac{n}{N}\right) \tag{2.62}$$

Si se identifican las unidades no nulas, sacamos de ellas una muestra de tamaño n_j. La estimación del total del dominio es $N_j\bar{y}_j$ con varianza

$$V(N_j\bar{y}_j) = \frac{N_j^2}{n_j}S_j^2\left(1 - \frac{n_j}{N_j}\right) = \frac{N^2}{n_j}P_j^2 S_j^2\left(1 - \frac{n_j}{N_j}\right) \tag{2.63}$$

Las varianzas comparables son (2.62) y (2.63). En la Ec. (2.62) el número promedio de unidades diferentes de 0 en la muestra de tamaño n es nP_j. Si tomamos $n_j = nP_j$ en (2.63), de tal manera que el número de unidades diferentes de 0 que va a ser medido es más o menos el mismo con ambos métodos, (2.63) se vuelve

$$V(N_j\bar{y}_j) = \frac{N^2}{n}P_j S_j^2\left(1 - \frac{n}{N}\right) \tag{2.64}$$

La razón de varianzas (2.64) a (2.62) es

$$\frac{V(N_j \text{ conocido})}{V(N_j \text{ no conocido})} = \frac{S_j^2}{S_j^2 + Q_j\bar{Y}_j^2} = \frac{C_j^2}{C_j^2 + Q_j} \tag{2.65}$$

en donde $C_j = S_j/\bar{Y}_j$, es el coeficiente de variación entre las unidades diferentes de 0. Como podría esperarse, la reducción en la varianza debida a un conocimiento de N_j es mayor cuando la proporción de unidades, con valor 0 es grande y cuando y_j varía relativamente poco entre las unidades con valores diferentes de 0. Para un estudio más amplio de este problema, véase Jessen y Houseman (1944).

2.14 COMPARACION ENTRE LAS MEDIAS DE LOS DOMINIOS

Supongamos que $\bar{y}_j$, $\bar{y}_k$ son los promedios de la muestra en el j-ésimo y k-ésimo dominios de un grupo de ellos, en los que se han clasificado las unidades de una muestra aleatoria simple. La varianza de sus diferencias es

$$V(\bar{y}_j - \bar{y}_k) = V(\bar{y}_j) + V(\bar{y}_k) \tag{2.66}$$

Esta fórmula se aplica también a la diferencia entre dos razones $\hat{R}_j$ y $\hat{R}_k$.

Debe observarse que rara vez es de interés científico preguntar si $\bar{Y}_j = \bar{Y}_k$ porque estas medias no serán exactamente iguales en una población finita, excepto por casualidad, aun cuando los datos en ambos dominios fueran sacados al azar de la misma población infinita. En lugar de esto, probamos la hipótesis nula de que los dos dominios se sacaron de poblaciones *infinitas* con la misma media. Por consiguiente, omitimos el cpf cuando calculamos $V(\bar{y}_j)$ y $V(\bar{y}_k)$, usando la fórmula

$$V(\bar{y}_j - \bar{y}_k) = \frac{S_j^2}{n_j} + \frac{S_k^2}{n_k} \tag{2.67}$$

Se obtiene una ecuación semejante a (2.67) para pruebas de significación, si se formula la pregunta: ¿Podrían las muestras de los dos dominios haber sido tomadas al azar de la misma población *finita*?

Con esta hipótesis nula puede probarse (Ej. 2.16) que

$$V(\bar{y}_j - \bar{y}_k) = S_{jk}^2\left(\frac{1}{n_j} + \frac{1}{n_k}\right)$$

donde S_{jk}^2 es la varianza de la población finita que consiste de los dominios combinados.

2.15 VALIDEZ DE LA APROXIMACION NORMAL

La confianza de que la aproximación normal es adecuada en la mayoría de situaciones prácticas proviene de diferentes fuentes. Se han hecho muchos estudios en la teoría de probabilidades sobre la distribución de las medias de muestras aleatorias. Se ha probado que para cualquier población que tiene una desviación estándar finita, la distribución de la media muestral tiende a la normalidad conforme n aumenta (véase, por ejemplo, Feller, 1957). Este trabajo se refiere a poblaciones infinitas.

Para el muestreo sin restitución de poblaciones finitas, Hájek (1960) ha dado las condiciones necesarias y suficientes en las que la distribución de la media muestral tiende a la normalidad, con base en los trabajos de Erdös y Rényi (1959) y Madow (1948). Hájek supone una secuencia de valores n_ν, N_ν que tienden a infinito de tal manera que $(N_\nu - n_\nu)$ también tiende a infinito. Las medidas de la ν-ésima población se identifican por $y_{\nu i}$, $(i = 1, 2, \ldots,$

$N_\nu)$. Para esta población, sea $S_{\nu\tau}$ el conjunto de unidades en la población para la cual

$$|y_{\nu i} - \bar{Y}_\nu| > \tau\sqrt{n_\nu(1-f_\nu)}S_\nu$$

en donde $\bar{Y}_\nu$, S_ν, f_ν, son la media de la población, desviación estándar y cpf, y τ es un número > 0. Entonces, la condición tipo Lindeberg

$$\lim_{\nu\to\infty} \frac{\sum_{S_{\nu\tau}} (y_{\nu i} - \bar{Y}_\nu)^2}{(N_\nu - 1)S_\nu^2} = 0$$

es necesaria y suficiente para asegurar que $\bar{y}_\nu$ tienda a la normalidad con la media y la varianza dadas en los Teoremas 2.1 y 2.2.

Este conocimiento deja algo que desear. No es fácil contestar la pregunta directa: "¿Qué tan grande debe ser n en esta población para que la aproximación normal sea suficientemente exacta?". Las distribuciones que no son normales varían mucho, tanto en la naturaleza como en el grado de su discrepancia de la normalidad. Las distribuciones de muchos tipos de empresas (tiendas, plantas avícolas, ciudades) exhiben una marcada asimetría positiva con pocas unidades grandes y muchas unidades pequeñas. La misma clase de asimetría se observa en ciertas poblaciones biológicas (por ejemplo, el número de ratas o moscas por cuadra en una ciudad).

Como una ilustración de una distribución positivamente asimétrica, en la Fig. 2.1 se muestra la distribución de frecuencia del número de habitantes en 196 ciudades grandes de los Estados Unidos en 1920. (Las cuatro ciudades más grandes, Nueva York, Chicago, Filadelfia y Detroit, fueron omitidas. Su inclusión hubiera extendido la escala horizontal a más de cinco veces el tamaño que se nota en la Fig. 2.1 y hubiera, desde luego, acentuado enormemente la asimetría.) En la Fig. 2.2 se muestra la distribución de frecuencia del número total de residentes en cada una de las 200 muestras aleatorias simples con $n = 49$ sacadas de esta población. La distribución de los totales de la muestra, e igualmente de las medias, es más similar a una curva normal, pero aún muestra algo de asimetría positiva.

Del estudio de las distribuciones teóricas asimétricas y de los resultados de los experimentos de muestreo en poblaciones asimétricas reales, se pueden hacer algunas aseveraciones respecto a lo

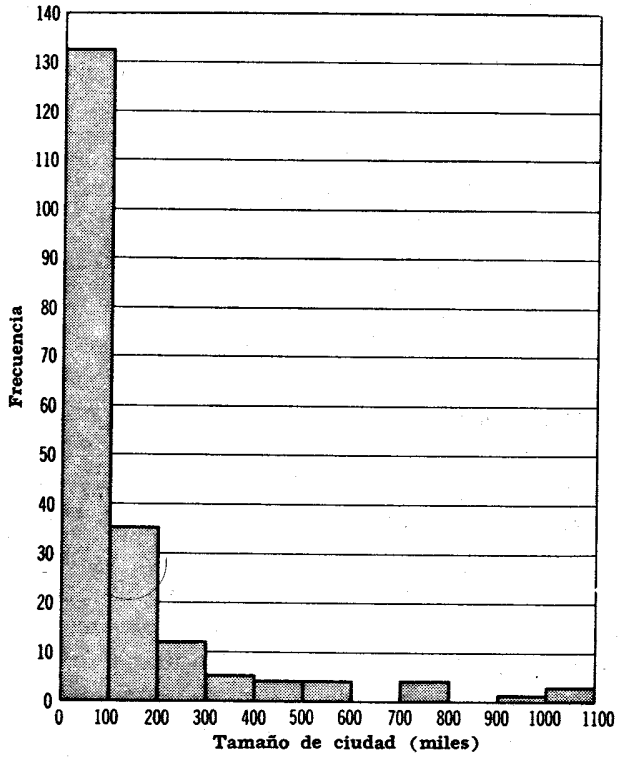


FIG. 2.1. Distribución de frecuencias de tamaños de 196 ciudades de los Estados Unidos en 1920

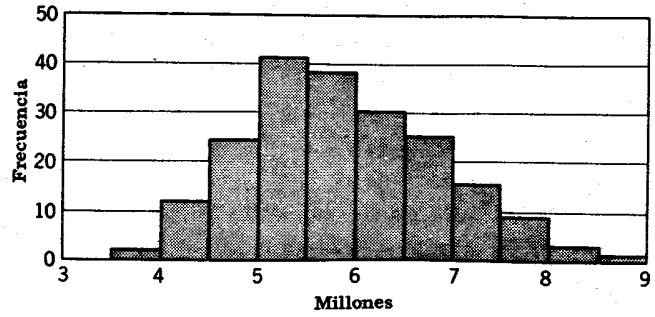


FIG. 2.2. Distribución de frecuencias de los totales de 200 muestras aleatorias simples con $n = 49$

que sucede usualmente a los límites y probabilidades de confianza cuando se muestrean dos poblaciones positivamente asimétricas como sigue:

1. La frecuencia con la cual la aseveración

$$\bar{y} - 1.96 s_{\bar{y}} < \bar{Y} < \bar{y} + 1.96 s_{\bar{y}}$$

está equivocada, es generalmente más alta que el 5%.

2. La frecuencia con la cual

$$\bar{Y} > \bar{y} + 1.96 s_{\bar{y}}$$

es *mayor* que 2.5%.

3. La frecuencia con la cual

$$\bar{Y} < \bar{y} - 1.96 s_{\bar{y}}$$

es *menor* que 2.5%.

Como una ilustración, consideremos una variable y que la podemos suponer distribuida binomialmente, de tal forma que la distribución exacta de $\bar{y}$ se puede obtener de las tablas binomiales. La variable y toma sólo dos valores —el valor h con probabilidad P y el valor 0 con probabilidad Q—. La media de la población es $\bar{Y} = Ph$. Una muestra aleatoria simple de tamaño n contiene a unidades con valor h y $n - a$ unidades que tienen el valor 0. En la muestra,

$$\sum y = ah, \qquad \bar{y} = \frac{ah}{n}$$

$$(n-1)s^2 = \sum y^2 - n\bar{y}^2 = ah^2 - \frac{a^2h^2}{n}$$

$$s_{\bar{y}}^2 = \frac{s^2}{n} = \frac{h^2}{n^2}\frac{a(n-a)}{n-1}$$

Por lo tanto, los límites normales de confianza al 95% para $\bar{Y}$ son

$$\bar{y} \pm 1.96 s_{\bar{y}} = \frac{h}{n}\left[a \pm 1.96\sqrt{\frac{a(n-a)}{n-1}}\right] \tag{2.68}$$

Sea $n = 400$, $P = 0.1$. Entonces $\bar{Y} = 0.1h$. Al probar encontramos que si $a = 29$ en la Ec. (2.68) el límite de confianza superior es $39.18h/400 = 0.098h$, mientras que $a = 30$ da $40.34h/400 = 0.101h$. De aquí que cualquier valor de $a \leq 29$ da un límite de confianza superior muy bajo. En forma similar, encontramos que si $a \geq 54$ el límite inferior es muy alto.

La variable a sigue una distribución binomial con $n = 400$, $P = 0.1$. Las tablas (Harvard Computation Laboratory, 1955) muestran que

Pr (límite superior establecido muy bajo) = Pr $(a \leq 29)$ = 0.0357
Pr (límite inferior establecido muy alto) = Pr $(a \geq 54)$ = 0.0217

Pr (postulado de confianza equivocado) = 0.0574

La probabilidad total de estar equivocado no está muy lejos de 0.05. En más del 60% de los postulados equivocados, la verdadera media es mayor que el límite superior establecido.

No hay una regla general segura para establecer qué tan grande debe ser n para poder usar la aproximación normal en el cálculo del límite de confianza. Para poblaciones en las cuales la desviación principal de normalidad consiste de una asimetría marcadamente positiva, una regla burda que ha sido útil en ocasiones es

$$n > 25G_1^2 \tag{2.69}$$

donde G_1 es la medida de asimetría de Fisher (Fisher, 1932).

$$G_1 = \frac{E(y_i - \bar{Y})^3}{\sigma^3} = \frac{1}{N\sigma^3}\sum_{i=1}^{N}(y_i - \bar{Y})^3 \tag{2.70}$$

Esta regla está diseñada para que los enunciados probabilistas al 95% de confianza no estén equivocados más del 6% de las veces. Lo anterior se obtiene matemáticamente, suponiendo que cualquier perturbación debida a los momentos mayores al tercero de la distribución de $\bar{y}$ es despreciable. La regla sólo intenta controlar la frecuencia *total* de postulados equivocados, ignorando la dirección del error de estimación.

Al calcular G_1, o una estimación, para una población específica, podemos obtener una idea aproximada del tamaño de muestra necesario para la aplicación de la aproximación normal en el cálculo de límites de confianza. El resultado se debe verificar experimentalmente siempre que sea posible.

Ejemplo. La información en la Tabla 2.4 muestra el número de acres dedicados a cultivos en 556 propiedades en el Condado de Seneca, Nueva York. La información proviene de una serie de estudios realizados por West (1951), quien obtuvo muestras repetidas de tamaño 100 de esta población y examinó las distribuciones de frecuencias de $\bar{y}$, s, y las t de Student, para varias características de interés en encuestas de administración de ranchos.

El cálculo de G_1 se muestra al pie de la Tabla. Los cálculos se hacen utilizando la escala codificada, y, puesto que G_1 es un número puro, no hay necesidad de regresar a la escala original. Nótese que el intervalo de la primera clase era ligeramente diferente de los otros.

Puesto que $G_1 = 1.9$, tomamos como una n mínima

$$n = (25)(1.9)^2 = 90$$

Para muestras de tamaño 100, West encontró en esta característica (acres en cultivos) que ni la distribución de $\bar{y}$ ni la de t de Student difieren significativamente de las correspondientes distribuciones normales teóricas.

TABLA 2.4. DISTRIBUCIÓN DE FRECUENCIAS DE ACRES CULTIVADOS EN 556 RANCHOS

Intervalos de clase (acres)	Escala codificada y_i	Frecuencia f_i	$f_i y_i$	$f_i y_i^2$	$f_i y_i^3$
0–29	−0.9	47	−42.3	38.1	−34.3
30–63	0	143	0	0	0
64–97	1	154	154	154	154
98–131	2	82	164	328	656
132–165	3	62	186	558	1 674
166–199	4	33	132	528	2 112
200–233	5	13	65	325	1 625
234–267	6	6	36	216	1 296
268–301	7	4	28	196	1 372
302–335	8	6	48	384	3 072
336–369	9	2	18	162	1 458
370–403	10	0	0	0	0
404–437	11	2	22	242	2 662
438–471	12	0	0	0	0
472–505	13	2	26	338	4 394
Totales		556	836.7	3 469.1	20 440.7

$$E(y_i) = \bar{Y} = \frac{836.7}{556} = 1.50486$$

$$E(y_i^2) = \frac{3469.1}{556} = 6.23939$$

$$E(y_i^3) = \frac{20\,440.7}{556} = 36.76385$$

$$\sigma^2 = E(y_i^2) - \bar{Y}^2 = 3.97479$$

$$\kappa_3 = E(y_i - \bar{Y})^3 = E(y_i^3) - 3E(y_i^2)\bar{Y} + 2\bar{Y}^3$$
$$= 15.411$$

$$G_1 = \frac{\kappa_3}{\sigma^3} = \frac{15.411}{7.925} = 1.9$$

Una buena práctica de muestreo tiende a hacer la aproximación normal más válida. La falla en la aproximación normal ocurre prin-

cipalmente cuando la población contiene algunos individuos extremos que dominan el promedio de muestra cuando están presentes. Sin embargo, estos extremos también tienen un efecto más serio al aumentar la varianza de la muestra y disminuir la precisión. Por lo tanto, se aconseja segregarlos y hacer planes separados para su control, tal vez enumerándolos completamente si no son muy numerosos. Esta eliminación de los individuos extremos de la población reduce la asimetría y mejora la aproximación normal. Esta técnica es un ejemplo del muestreo estratificado que se discutirá en el Cap. 5.

2.16 ESTIMADORES LINEALES DE LA MEDIA DE POBLACION

En muestreo aleatorio simple nos preguntamos si la media muestral $\bar{y}$ es el mejor estimador de $\bar{Y}$. Esto naturalmente ha dado lugar a un buen número de trabajos. La respuesta depende del conjunto de competidores permitidos para $\bar{y}$, y de la definición de "el mejor". Para el muestreo, las unidades de la población suelen numerarse en alguna forma de 1 a N. Estos números suelen llamarse las *etiquetas* que se ponen a las unidades.

Los primeros resultados probados por Horvitz y Thompson (1952) para estimadores lineales en muestreo aleatorio simple son los siguientes. Si cualquier y_i recibe siempre el mismo peso w_i siempre que se extraiga la unidad rotulada con i, la media muestral $\bar{y}$ es el *único* estimador insesgado de $\bar{Y}$ de forma $\sum^{n} w_i y_i$. Dado que cada unidad aparece en una fracción n/N de muestras aleatorias simples, $E\left(\sum^{n} w_i y_i\right) = n\left(\sum^{N} w_i y_i\right)\Big/N \equiv \bar{Y}$ sólo si cada $w_i = 1/n$. Si, por otro lado, el peso dependiera sólo del *orden* con que se extrae la unidad para la muestra, entonces $\bar{y}$ tiene varianza mínima entre los estimadores lineales insesgados de la forma $\sum_{d}^{n} w_d y_{(d)}$, donde $y_{(d)}$ es el valor y en la unidad que se saca en la extracción d-ésima.

Una clase más amplia de competidores es el conjunto $\sum^{n} w_{is} y_i$, donde w_{is} puede depender de las otras unidades que caen en la muestra, así como de i. Godambe (1955) demostró que en esta clase no existe estimación insesgada de $\bar{Y}$ con varianza mínima para todas las poblaciones.

Hartley y J. N. K. Rao (1968-1969) han desarrollado otras propiedades de $\bar{y}$, al igual que Royall (1968) y C. R. Rao (1971). En cualquier población finita habrá cuando mucho $T \leq N$ valores distintos de los y_i. En la muestra, sean n_t valores iguales a y_t donde $\sum n_t = n$. Hartley y Rao (1968) demuestran que $\bar{y}$ tiene varianza mínima entre los estimadores insesgados de $\bar{Y}$ que sólo son funciones de n_t y y_t. Para muestreo aleatorio con restitución, prueban que la media de los valores *distintos* en la muestra es el estimador de máxima verosimilitud de $\bar{Y}$, aunque no tiene la varianza mínima en todas las poblaciones.

C. R. Rao (1971), siguiendo los trabajos de Kempthorne (1969), consideró los estimadores insesgados $\hat{\bar{Y}} = \sum w_{is} y_i$ discutidos por Godambe. Para representar el caso donde las etiquetas l no dan información respecto a los valores y_l, C. R. Rao calculó el promedio de $V(\hat{\bar{Y}})$ sobre todas las $N!$ permutaciones de las etiquetas asignadas a los valores y demostró que sobre estas permutaciones, $\bar{y}$ tiene varianza promedio mínima. Royall (1970b) dio un resultado más general.

El trabajo de Godambe (1955) estimuló numerosas investigaciones en el diseño de muestras y la estimación incluyendo temas como los criterios para juzgar los estimadores, el papel de la máxima verosimilitud, el uso de la información auxiliar que las etiquetas pueden contener respecto a los y_i, los estimadores bayesianos, y los métodos de estimación cuando se pueden hacer suposiciones sobre la distribución de frecuencia de los y_i. La influencia de este trabajo en la práctica del muestreo ha sido hasta ahora limitada, pero debe aumentar constantemente. En ocasiones nos referiremos a dicho trabajo. Para una revisión de estos trabajos, véase J. N. K. Rao* (1975a) y Smith (1976).

EJERCICIOS

2.1. En una población con $N = 6$ los valores de y_i son 8, 3, 1, 11, 4 y 7. Calcular la media de muestra $\bar{y}$ para todas las posibles muestras aleatorias simples de tamaño 2. Verificar que $\bar{y}$ es una estimación insesgada de $\bar{Y}$ y que su varianza es la expresada en el Teorema 2.2.

2.2. Para la misma población, calcular s^2 para todas las muestras aleatorias simples de tamaño 3 y verificar que $E(s^2) = S^2$.

* En lo sucesivo al mencionar el apellido Rao, nos referiremos exclusivamente a J. N. K. Rao a menos que especifique lo contrario.

2.3. Si las muestras aleatorias de tamaño 2 se sacan con restitución de la población, demuestre, buscando todas las muestras posibles, que $V(\bar{y})$ satisface la ecuación

$$V(\bar{y}) = \frac{\sigma^2}{n} = \frac{S^2}{n}\frac{(N-1)}{N}$$

2.4. Una muestra aleatoria simple de 30 familias se obtuvo de una área de la ciudad que contiene 14 848 familias. El número de personas por familia en la muestra obtenida fue como sigue:

5, 6, 3, 3, 2, 3, 3, 3, 4, 4, 3, 2, 7, 4, 3, 5, 4, 4, 3, 3, 4, 3, 3, 1, 2, 4, 3, 4, 2, 4

Estime el número total de gente en el área y calcule la probabilidad de que esta estimación esté dentro del $\pm 10\%$ del valor verdadero.

2.5. En un estudio sobre el posible uso del muestreo para reducir el trabajo de inventario de existencias en una bodega, se hizo un conteo del valor de los artículos de cada uno de los 36 estantes en la bodega. Los valores aproximados fueron:

29, 38, 42, 44, 45, 47, 51, 53, 53, 54, 56, 56, 56, 58, 58, 59, 60, 60,
60, 60, 61, 61, 61, 62, 64, 65, 65, 67, 67, 68, 69, 71, 74, 77, 82, 85.

La estimación del valor total a partir de una muestra debe ser correcta módulo un error máximo de $200.00, excepto para una posibilidad en veinte. Un consultor sugiere que una muestra aleatoria simple de 12 estantes es suficiente para hacer la estimación. ¿Está usted de acuerdo?

$$\Sigma y = 2138, \qquad \Sigma y^2 = 131\,682$$

2.6. Después de tomar la muestra en la Tabla 2.2 (Pág. 53), se contó el número de hojas llenas (con 42 firmas cada una) y se encontró que eran 326. Use esta información para obtener una mejor estimación del número total de firmas y encuentre el error estándar de su estimación.

2.7. De una lista de 468 academias con cursos de 2 años se sacó una muestra aleatoria simple de 100. La muestra contenía 54 instalaciones públicas y 46 privadas. Los datos para el número de estudiantes (y) y el número de profesores (x) se muestran a continuación.

	n	$\Sigma(y)$	$\Sigma(x)$
Pública	54	31 281	2 024
Privada	46	13 707	1 075
	$\Sigma(y^2)$	$\Sigma(yx)$	$\Sigma(x^2)$
Pública	29 881 219	1 729 349	111 090
Privada	6 366 785	431 041	33 119

(a) Para cada tipo de instalación en la población, estimar la razón (número de estudiantes/número de profesores). (b) Calcular los errores estándar de las estimaciones. (c) Para las instalaciones públicas, encontrar los límites de confianza al 90% para la razón de estudiante/maestro en la población total.

2.8. En el ejemplo anterior pruebe al nivel del 5% si la razón de estudiante/maestro es significativamente diferente en ambos tipos de instalaciones.

2.9. Para las instalaciones públicas, estimar el número total de maestros: (a) a sabiendas que el número total de instalaciones públicas en la población es 251, (b) sin conocer esta cifra. En cada caso calcule el error estándar de su estimación.

2.10. La Tabla siguiente muestra el número de residentes en cada una de las 197 ciudades de Estados Unidos que tenían más de 50 000 habitantes en 1940. Calcular el error estándar del número total de habitantes estimado en las 197 ciudades utilizando los siguientes métodos de muestreo: (a) una muestra aleatoria simple de tamaño 50, (b) una muestra que incluya las 5 ciudades más grandes y una muestra aleatoria simple de tamaño 45 para las 192 ciudades restantes, (c) una muestra que incluya las 9 ciudades más grandes y una muestra aleatoria simple de tamaño 41 para las ciudades restantes.

2.11. Calcular el coeficiente de asimetría G_1 para la población original y para la población restante después de eliminar: (a) las 5 ciudades más grandes, (b) las 9 ciudades más grandes.

DISTRIBUCIÓN DE FRECUENCIA DE LOS TAMAÑOS DE LAS CIUDADES

Clase del tamaño (miles)	f	Clase del tamaño (miles)	f	Clase del tamaño (miles)	f
50–100	105	550–600	2	....	....
100–150	36	600–650	1	1500–1550	1
150–200	13	650–700	2	....	....
200–250	6	700–750	0	1600–1650	1
250–300	7	750–800	1	....	....
300–350	8	800–850	1	1900–1950	1
350–400	4	850–900	2	....	....
400–450	1	900–950	0	3350–3400	1
450–500	3	950–1000	0	....	....
500–550	0	1000–1050	0	7450–7500	1

Las lagunas en los intervalos se indican con . . .

2.12. Se va a realizar una pequeña encuesta para comparar los propietarios de casas con los arrendatarios. En la población, cerca del 75% son propietarios y 25% son arrendatarios. Para una característica, se piensa que la varianza es aproximadamente de 15, tanto para propietarios como para arrendatarios. El error estándar de la diferencia entre las dos medias de los dominios no debe exceder a 1. ¿Qué tan grande debe ser la muestra (a) si los propietarios y los arrendatarios se pueden identificar antes de obtener la muestra, ? (b) si no es así, (Una respuesta aproximada será suficiente en (b); una solución exacta requiere de tablas binomiales.)

2.13. Una muestra aleatoria simple de tamaño 3 se obtiene de una población de tamaño N con restitución. Demuestre que las probabilidades de que la muestra contenga unidades *diferentes* de 1, 2 y 3 (por ejemplo, *aaa, aab, abc,* respectivamente) son

$$P_1=\frac{1}{N^2}, \qquad P_2=\frac{3(N-1)}{N^2}, \qquad P_3=\frac{(N-1)(N-2)}{N^2}$$

Como una estimación de $\bar{Y}$ se toma y', la media no ponderada sobre las diferentes unidades en la muestra. Demostrar que la varianza promedio de $\bar{y}'$ es

$$V(\bar{y}')=\frac{(2N-1)(N-1)S^2}{6N^2}\doteq(1-f/2)S^2/3$$

Una manera de hacer esto es demostrar que

$$V(\bar{y}')=S^2\left(\frac{N-1}{N}P_1+\frac{N-2}{2N}P_2+\frac{N-3}{3N}P_3\right)$$

Por lo tanto, demostrar que $V(\bar{y}') < V(\bar{y})$, donde $\bar{y}$ es la media aritmética de las n observaciones en la muestra. El resultado $V(\bar{y}') < V(\bar{y})$ para cualquier $n > 2$ fue probado por Raj y Khamis (1958).

2.14. Dos dentistas A y B, hicieron una encuesta para investigar el estado de los dientes de 200 niños en una colonia. El Dr. A seleccionó una muestra aleatoria simple de 20 niños y contó el número de dientes cariados de cada niño, con los siguientes resultados:

Número de dientes cariados, por niño	0	1	2	3	4	5	6	7	8	9	10
Número de niños	8	4	2	2	1	1	0	0	0	1	1

El Dr. B, usando las mismas técnicas dentales examinó a los 200 niños y sólo registró aquellos que no tenían los dientes cariados. Encontró que 60 niños no tenían dientes dañados.

Estimar el número total de dientes cariados en los niños de la colonia, (a) usando sólo los resultados de A, (b) usando los resultados de A y B, (c) ¿Son insesgadas las estimaciones? (d) ¿Qué estimación espera que sea más precisa?

2.15. Una compañía intenta entrevistar una muestra aleatoria simple de empleados asociados con ella por más de cinco años. La compañía tiene \$1 000 para este estudio y cada entrevista cuesta \$10. No hay una lista separada de empleados con más de 5 años de servicio, pero se puede compilar una lista de los expedientes a un costo de \$200. La compañía puede: (a) compilar la lista y entrevistar a las personas incluidas en una muestra aleatoria simple sacada de los empleados elegidos o (b) sacar una muestra aleatoria simple de todos los empleados, y entrevistar sólo a los elegibles. El costo de rechazar aquellos empleados incluidos en la muestra que no son elegibles es despreciable.

Demostrar que para estimar un total sobre la población de empleados elegibles, el plan (a) da una varianza más pequeña que el plan (b) sólo si $C < 2\sqrt{Q_j}$, donde C es el coeficiente de variación de la característica bajo estudio entre los empleados elegibles y Q_j es la proporción de empleados no elegibles de la compañía. Ignore la cpf.

2.16. Se toma una muestra aleatoria simple de tamaño $n = n_1 + n_2$ con media $\bar{y}$ en una población finita y también una submuestra aleatoria simple de tamaño n_1 de la misma población y con media $\bar{y}_1$. Demostrar que (a) $V(\bar{y}_1 - \bar{y}_2) = S^2[(1/n_1) + (1/n_2)]$ donde $\bar{y}_2$ es la media de las restantes n_2 unidades de la muestra, (b) $V(\bar{y}_1 - \bar{y}) = S^2[(1/n_1) - (1/n)]$, (c) Cov $\{\bar{y}, \bar{y}_1 - \bar{y}\} = 0$. El muestreo repetido implica extracciones repetidas tanto de la muestra como de la submuestra.

2.17. El número de muestras aleatorias simples distintas de tamaño n es desde luego $N!/n!(N-n)!$. Es de interés encontrar conjuntos más pequeños de muestras de tamaño n que tengan las mismas propiedades que el conjunto de muestras aleatorias simples. Un conjunto tal es el de los *diseños por bloques incompletos equilibrados* (*bib*). Estas son muestras de n distintas unidades tomadas entre N, tales que (i) cada unidad aparece en el mismo número (r) de muestras, (ii) cada par de unidades aparece simultáneamente en λ muestras.

Verifíquese que $\lambda = r(n-1)/(N-1)$ y que el número de muestras distintas en el conjunto es rN/n. Sobre el conjunto de las muestras *bib*, demuéstrese en la notación usual que si $\bar{y}$ es la media de una muestra (a) $V(\bar{y}) = (1-f)S^2/n$ y (b) $v(\bar{y}) = (1-f)\sum^{n}(y_i - \bar{y})^2/n(n-1)$ es una estimación insesgada de $V(\bar{y})$.

Nota. No existe un método general para encontrar el más pequeño r para el cual pueda construirse un *bib*. En ocasiones, el más pequeño r conocido proporciona $N!/n!(N-n)!$ muestras, lo que nos lleva a las muestras aleatorias simples. Pero para $N = 91$, $n = 10$ el más pequeño conjunto *bib* tiene 91 muestras contra más de 6 trillones (E.U.) de muestras aleatorias simples. Avadhani y Sukhatme (1973) han mostrado cómo se pueden usar los diseños *bib* para tratar de reducir los gastos de viaje entre unidades de muestreo.

2.18. La siguiente es una ilustración debida a Royall (1968), el hecho de que en el muestreo aleatorio simple la media de muestra $\bar{y}$ no presenta varianza uniformemente mínima en la clase de los estimadores de la forma $\sum^{n} w_{is}y_i$ considerada por Godambe (1955) donde el peso w_{is} puede depender de las otras unidades que caigan en la muestra. Para $N = 3$, $n = 2$ considerar el estimador

$$\hat{\bar{Y}}_{12}=\tfrac{1}{2}y_1+\tfrac{1}{2}y_2; \qquad \hat{\bar{Y}}_{13}=\tfrac{1}{2}y_1+\tfrac{2}{3}y_3; \qquad \hat{\bar{Y}}_{23}=\tfrac{1}{2}y_2+\tfrac{1}{3}y_3$$

donde $\hat{\bar{Y}}_{ij}$ es el estimador para la muestra que tiene las unidades (i, j). Demostrar que en los resultados de Royall $\hat{\bar{Y}}_{ij}$ es insesgado y $V(\hat{\bar{Y}}_{ij}) < V(\bar{y})$ si $y_3(3y_2 - 3y_1 - y_3) > 0$. La ilustración se toma de un ejemplo anterior de Roy y Chakravarti (1960).

2.19. Este ejercicio es otro ejemplo de estimadores construidos con base en ciertos rasgos particulares de las poblaciones. Después de decidir la toma de una muestra aleatoria simple se dieron cuenta de que y_1 sería inacostumbradamente bajo y y_{N_e} inacostumbradamente alto. Para esta situación, Särndal (1972) examinó el siguiente estimador insesgado de $\bar{Y}$.

$$\hat{\bar{Y}}_S = \bar{y} + c \quad \text{si la muestra contiene } y_1 \text{ pero no } y_N$$
$$= \bar{y} - c \quad \text{si la muestra contiene } y_N \text{ pero no } y_1$$
$$= \bar{y} \quad \text{para todas las otras muestras.}$$

donde c es una constante. Demuestre el resultado de Särndal en el que $\bar{Y}_S$ es insesgado con

$$V(\hat{\bar{Y}}_S) = (1-f)\left[\frac{S^2}{n} - \frac{2c}{(N-1)}(y_N - y_1 - nc)\right]$$

de manera que $V(\hat{\bar{Y}}_S) < V(\bar{y})$ si $0 < c < (y_N - y_1)/n$.

2.20. Para una población con $N = 8$ y valores $y_i = 1, 4, 5, 5, 6, 6, 8, 13$, demostrar que con $n = 4$, $V(\bar{y}) = 1.5$, en tanto que $V(\hat{\bar{Y}}_S) = 0.214$ cuando $c = 1.5$ (su mejor valor), y 0.357 cuando $c = 1$ o 2.

Con base en la información del Ejercicio 2.19, otro plan de muestreo consiste en incluir tanto y_1 como y_8 en cada muestra, y extraer una muestra aleatoria simple de tamaño 2 de $y_2 \ldots, y_7$ con media $\bar{y}_2$. La estimación de $\bar{Y}$ es

$$\hat{\bar{Y}}_{st} = (y_1 + 6\bar{y}_2 + y_8)/8$$

Demostrar que $\hat{\bar{Y}}_{st}$ es insesgado, con varianza $9V(\bar{y}_2)/16$. Para esta población demostrar que $V(\hat{\bar{Y}}_{st}) = 0.350$. Este estimador es un ejemplo de muestreo estratificado (Cap. 5) con tres estratos: y_1 ; $y_2 \ldots y_7$ y y_8.

CAPITULO 3

Muestreo para Proporciones y Porcentajes

3.1 CARACTERISTICAS CUALITATIVAS

En algunas ocasiones deseamos estimar el número total, la proporción, o el porcentaje de unidades en la población, que poseen alguna característica o atributo, o que caen dentro de alguna clase definida. Muchos de los resultados, que regularmente se publican, derivados de censos y encuestas, son de esta forma, por ejemplo, el número de personas sin empleo, el porcentaje de la población originaria de un lugar. La clasificación puede ser introducida en forma directa en el cuestionario, en forma de preguntas que se contestan con un "sí" o un "no". En otros casos las medidas originales son más o menos continuas y la clasificación se introduce al tabular los resultados. Por consiguiente, podemos registrar la edad de los entrevistados con aproximación al año y publicar el porcentaje de la población que tiene 60 o más años cumplidos.

Notación. Suponemos que todas y cada una de las unidades en la población caen dentro de una, de dos posibles clases C y C'. La notación es como sigue:

Número de unidades en C en		Proporción de unidades en C en	
Población	Muestra	Población	Muestra
A	a	$P = A/N$	$p = a/n$

La estimación muestral de P es p, y la estimación muestral de A es Np o Na/n. En el trabajo estadístico, la distribución *binomial* se aplica con frecuencia a estimaciones como a y p. Como se verá, la

distribución correcta en poblaciones finitas es la *hipergeométrica*, aun cuando la binomial es una aproximación generalmente satisfactoria.

3.2 VARIANZAS DE LAS ESTIMACIONES MUESTRALES

Por medio de un artificio simple se pueden aplicar los teoremas presentados en el Cap. 2 a esta situación. Para cualquier unidad en la muestra o población, se define y_i como 1 si la unidad está en C, y como 0 si la unidad está en C'. Para esta población de valores y_i, está claro que

$$Y = \sum_1^N y_i = A \tag{3.1}$$

$$\bar{Y} = \frac{\sum_1^N y_i}{N} = \frac{A}{N} = P \tag{3.2}$$

En la misma forma, para la muestra.

$$\bar{y} = \frac{\sum_1^n y_i}{n} = \frac{a}{n} = p \tag{3.3}$$

De modo que, el problema de estimar A y P es similar a la estimación del total y la media de una población en la cual, todos los valores y_i son 1 o 0. Para usar los teoremas del Cap. 2, debemos primero expresar S^2 y s^2 en términos de P y p. Nótese que

$$\sum_1^N y_i^2 = A = NP, \qquad \sum_1^n y_i^2 = a = np$$

Por lo tanto,

$$S^2 = \frac{\sum_1^N (y_i - \bar{Y})^2}{N-1} = \frac{\sum_1^N y_i^2 - N\bar{Y}^2}{N-1}$$

$$= \frac{1}{N-1}(NP - NP^2) = \frac{N}{N-1}PQ \tag{3.4}$$

donde $Q = 1 - P$. En forma similar,

$$s^2 = \frac{\sum_1^n (y_i - \bar{y})^2}{n-1} = \frac{n}{n-1}pq \tag{3.5}$$

Al aplicar los Teoremas 2.1, 2.2 y 2.4 a esta población se obtienen los siguientes resultados para un muestreo aleatorio simple de las unidades clasificadas.

Teorema 3.1. La proporción en la muestra $p = a/n$ es una estimación insesgada de la proporción en la población $P = A/N$.

Teorema 3.2. La varianza de p es

$$V(p) = E(p-P)^2 = \frac{S^2}{n}\left(\frac{N-n}{N}\right) = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right) \tag{3.6}$$

Al usar (3.4).

Corolario 1. Si p y P son los *porcentajes* de la muestra y de la población respectivamente, que caen en la clase C, (3.6) continúa siendo la varianza de p.

Corolario 2. La varianza de $\hat{A} = Np$, la estimación del número total de unidades en la clase C, es

$$V(\hat{A}) = \frac{N^2PQ}{n}\left(\frac{N-n}{N-1}\right) \tag{3.7}$$

Teorema 3.3. Una estimación insesgada de la varianza de p, derivada de una muestra, es

$$v(p) = s_p^2 = \frac{N-n}{(n-1)N}pq \tag{3.8}$$

Prueba. En el corolario del Teorema 2.4 se demostró que para una variable continua y_i una estimación insesgada de la varianza de la media muestral $\bar{y}$ es

$$v(\bar{y}) = \frac{s^2}{n}\frac{(N-n)}{N} \tag{3.9}$$

Para proporciones, P toma el lugar de $\bar{y}$, y en (3.5) se demostró que

$$s^2 = \frac{n}{n-1}pq \tag{3.10}$$

Por lo tanto:

$$v(p) = s_p^2 = \frac{N-n}{(n-1)N}pq \tag{3.11}$$

De lo anterior se tiene que si N es muy grande en relación a n, de tal manera que el factor de corrección por finitud es muy pequeño, una estimación insesgada de la varianza de p es:

$$\frac{pq}{n-1}$$

Este resultado puede parecer singular a algunos lectores, puesto que la expresión pq/n casi siempre se usa en la práctica como una estimación de la varianza. El hecho es que pq/n es, aun en poblaciones infinitas, sesgada.

Corolario. Una estimación insesgada de la varianza de $\hat{A} = Np$, el número total estimado de unidades en la clase C en la población, es

$$v(\hat{A}) = s_{Np}{}^2 = \frac{N(N-n)}{n-1}pq \tag{3.12}$$

Ejemplo. De una lista de 3 042 nombres y direcciones, una muestra aleatoria simple de 200 nombres mostró, al revisarla, 38 direcciones equivocadas. Estime el número total de direcciones que se necesita corregir en la lista y encuentre el error estándar de esta estimación. Tenemos

$$N = 3042, \qquad n = 200, \qquad a = 38, \qquad p = 0.19$$

El número total estimado de direcciones equivocadas es

$$\hat{A} = Np = (3042)(0.19) = 578$$

$$s_{\hat{A}} = \sqrt{[(3042)(2842)(0.19)(0.81)/199]} = \sqrt{6686} = 81.8$$

Puesto que la fracción de muestreo es menor del 7%, el factor de corrección por finitud (cpf) no afecta. Para quitar el factor, se sustituye el término $N - n$ por N. Si además, sustituimos $n - 1$ por n, tenemos la fórmula más simple

$$s_{Np} = N\sqrt{pq/n} = (3042)\sqrt{(0.19)(0.81)/200} = 84.4$$

Este resultado concuerda muy bien con el anterior de 81.8.

Las fórmulas anteriores de varianza y varianza estimada de p, son ciertas sólo si las *unidades* son clasificadas dentro de C o C', de tal manera que p es la razón del número de unidades en C en la muestra, al número total de unidades en la muestra. En muchas encuestas cada unidad está compuesta por un grupo de elementos y son éstos los que se clasifican. Como ejemplos tenemos:

Unidad de muestreo	Elementos
Familia	Miembros de la familia
Restaurante	Empleados
Canasta de huevos	Huevos
Arbol de durazno	Duraznos

Si se saca una muestra aleatoria simple de unidades para estimar la producción P de *elementos* en la población que pertenecen a la clase C, las fórmulas anteriores no se aplican. Los métodos adecuados aparecen en la Sec. 3.12.

3.3 EL EFECTO DE P EN LOS ERRORES ESTANDAR

La Ec. (3.6) muestra cómo la varianza del porcentaje estimado cambia con P, para valores de n y N fijos. Si se ignora la (cpf), tenemos

$$V(p) = \frac{PQ}{n}$$

La función PQ y su raíz cuadrada se muestran en la Tabla 3.1. Estas funciones se pueden considerar como la varianza y la desviación estándar, respectivamente, para una muestra de tamaño 1.

TABLA 3.1. VALORES DE PQ Y $\sqrt{PQ}$

P = Porcentaje de la población en la clase C

P	0	10	20	30	40	50	60	70	80	90	100
PQ	0	900	1600	2100	2400	2500	2400	2100	1600	900	0
$\sqrt{PQ}$	0	30	40	46	49	50	49	46	40	30	0

Las funciones alcanzan sus valores más grandes cuando la población está igualmente dividida entre las dos clases, y dichas funciones son simétricas respecto a este punto. El error estándar de p cambia relativamente poco cuando P está entre el 30% y el 70%. Con el valor máximo de $\sqrt{PQ}$, 50, se necesita una muestra de tamaño 100 para reducir el error estándar de la estimación al 5%. Para alcanzar un error estándar de 1% se requiere un tamaño de muestra de 2 500.

Este enfoque no es el apropiado cuando el interés está en el número total de unidades en la población que se encuentran en la clase C. En este caso es más natural preguntar: ¿es posible que la estimación sea correcta módulo un error, digamos, del 7% del valor verdadero? Por lo tanto, tendemos a pensar en el error estándar, como una fracción o porcentaje del valor verdadero, NP. La fracción es

$$\frac{\sigma_{Np}}{NP} = \frac{N\sqrt{PQ}}{\sqrt{n}NP}\sqrt{\frac{N-n}{N-1}} = \frac{1}{\sqrt{n}}\sqrt{\frac{Q}{P}}\sqrt{\frac{N-n}{N-1}} \tag{3.13}$$

Esta cantidad se llama normalmente el *coeficiente de variación* de la estimación. Si la (cpf) se ignora, el coeficiente es $\sqrt{Q/nP}$. La proporción $\sqrt{Q/P}$, que se puede considerar como el coeficiente de variación para una muestra de tamaño 1, se exhibe en la Tabla 3.2.

TABLA 3.2. VALORES DE $\sqrt{Q/P}$ PARA DIFERENTES VALORES DE P

P = Porcentaje de la población en la clase C

P	0	0.1	0.5	1	5	10	20
$\sqrt{Q/P}$	∞	31.6	14.1	9.9	4.4	3.0	2.0
P	30	40	50	60	70	80	90
$\sqrt{Q/P}$	1.5	1.2	1.0	0.8	0.7	0.5	0.3

Para un tamaño fijo de muestra, el coeficiente de variación del total estimado en la clase C disminuye consistentemente conforme aumenta el verdadero porcentaje en C. El coeficiente es alto cuando P es menor que el 5%. Se necesitan muestras muy grandes para obtener estimaciones precisas del número total de individuos que poseen cualquier atributo que sea raro en la población. Para $P = 1\%$, debemos tener $\sqrt{n} = 99$ con el fin de reducir el coeficiente de variación de la estimación a 0.1 o 10%. Esto da un tamaño de muestra de 9 801. El muestreo aleatorio simple, o cualquier método de muestreo que se adapte para propósitos generales, resulta costoso en la estimación del número total de unidades que posean un atributo raro y se encuentren en número reducido en la población.

3.4 LA DISTRIBUCION BINOMIAL

Puesto que la población es de un tipo particularmente simple, en donde la y_i toma los valores 1 o 0, podemos encontrar la distribución real de frecuencias de la estimación p y no sólo su media y varianza.

La población contiene A unidades que están en la clase C y $N - A$ unidades en C', en donde $P = A/N$. Si la primera unidad que se analiza sucede que pertenece a C, quedarán en la población $A - 1$ unidades en C y $N - A$ en C'. Por lo tanto, la proporción de unidades en C, después de obtener la primera unidad, cambia ligeramente a $(A - 1)/(N - 1)$. Alternativamente, si la primera unidad analizada se encuentra en C', la proporción en C cambia a $A/(N - 1)$. En el muestreo sin restitución, la proporción cambia de esta manera a lo largo del proceso de obtención de unidades. En la presente sección, estas variaciones se ignoran, es decir, se presupone que P es constante. Esto equivale a suponer que A y $N - A$ son grandes con respecto al tamaño de muestra n, o que el muestreo es con restitución.

Con esta suposición, el proceso de obtención de individuos para formar la muestra consiste de una serie de n ensayos, en cada uno de los cuales, la probabilidad de que la unidad que se analiza pertenezca a C es P. Esta situación da lugar a la distribución de frecuencia conocida como binomial para el número de unidades en C en la muestra. La probabilidad de que la muestra contenga a unidades de C es:

$$\Pr(a) = \frac{n!}{a!(n-a)!} P^a Q^{n-a} \tag{3.14}$$

De esta expresión podemos tabular la distribución de frecuencia de a, de $p = a/n$, o del total estimado Np.

Existen 3 conjuntos de tablas satisfactorias. Todas ellas dan a P en intervalos de 0.01. La amplitud para n es como sigue:

U.S. Bureau of Standards (1950).

$n = 1(1)49$ (*es decir, varía de* 1 *a* 49 *con intervalos de* 1).

Romig (1952): $n = 50(5)100$.

Harvard Computation Laboratory (1955):

$n = 1(1)50(2)100(10)200(20)500(50)1\,000$

3.5 LA DISTRIBUCION HIPERGEOMETRICA

La distribución de p puede encontrarse sin suponer que la población es grande en relación a la muestra. Los números de unidades en las dos clases C y C' en la población, son A y A', respectivamente. Calcularemos la probabilidad de que los números correspondientes en la muestra sean a y a' donde

$$a + a' = n, \qquad A + A' = N$$

En el muestreo aleatorio simple cada una de las $\binom{N}{n}$ selecciones diferentes de n unidades obtenidas de N tienen igual probabilidad de obtenerse. Para encontrar la probabilidad deseada, contamos el número de muestras que contienen exactamente a unidades pertenecientes a C y a' de C'. El número de selecciones diferentes de a unidades entre las A que están en C es $\binom{A}{a}$ mientras que el nú-

mero de selecciones diferentes de a' entre las A' es $\binom{A'}{a'}$. Cada una de las selecciones del primer tipo se puede combinar con cualquiera del segundo para dar una muestra diferente del tipo requerido. El número total de muestras de los tipos requeridos es, por lo tanto,

$$\binom{A}{a}\cdot\binom{A'}{a'}$$

De aquí que, si se obtiene una muestra aleatoria simple de tamaño n la probabilidad de que sea del tipo requerido es

$$\Pr(a, a'|A, A') = \binom{A}{a}\cdot\binom{A'}{a'}\Big/\binom{N}{n} \tag{3.15}$$

Esta es la distribución de frecuencias de a o np, de la cual, la de p es inmediatamente derivable. La distribución se llama *hipergeométrica.*

Para propósitos de cálculo la probabilidad hipergeométrica (3.15) puede escribirse como sigue:

$$\frac{n!}{a!(n-a)!}\cdot\frac{A(A-1)\ldots(A-a+1)(A')(A'-1)\ldots(A'-a'+1)}{N(N-1)\ldots(N-n+1)} \tag{3.16}$$

Ejemplo. Una familia de 8 miembros está formada por 3 hombres y 5 mujeres. Encontrar la distribución de frecuencias del número de hombres en una muestra aleatoria simple de tamaño 4. En este caso

$$A=3;\quad A'=5,\qquad N=8;\qquad n=4$$

De la Ec. (3.16) la distribución del número de hombres, a, es la siguiente:

a	Probabilidad
0	$\frac{4!}{0!4!}\cdot\frac{5.4.3.2}{8.7.6.5}=\frac{1}{14}$
1	$\frac{4!}{1!3!}\cdot\frac{3.5.4.3}{8.7.6.5}=\frac{6}{14}$
2	$\frac{4!}{2!2!}\cdot\frac{3.2.5.4}{8.7.6.5}=\frac{6}{14}$
3	$\frac{4!}{3!1!}\cdot\frac{3.2.1.5}{8.7.6.5}=\frac{1}{14}$
4	Imposible $=0$

El lector puede verificar que el número medio de hombres $\frac{3}{2}$ y que la varianza es $\frac{15}{28}$. Estos resultados concuerdan con las fórmulas establecidas previamente en la Sec. 3.2, las cuales dan

$$E(np)=nP=\frac{nA}{N}=\frac{(4)(3)}{8}=\frac{3}{2}$$

$$V(np)=nPQ\frac{N-n}{N-1}=4\cdot\frac{3}{8}\cdot\frac{5}{8}\cdot\frac{4}{7}=\frac{15}{28}$$

3.6 LIMITES DE CONFIANZA

Primero discutiremos el significado de los límites de confianza para el caso de características cualitativas. En la muestra, a individuos de los n pertenecen a la clase C. Supóngase que las inferencias se van a hacer respecto al número A en la población que pertenece a la clase C. Para el límite de confianza superior respecto a A, calculamos un valor $\hat{A}_U$ tal, que para éste, la probabilidad de obtener a o menos individuos pertenecientes a C en la muestra es una pequeña cantidad α_U, por ejemplo, 0.025. Formalmente, $\hat{A}_U$ satisface la ecuación

$$\sum_{j=0}^{a}\Pr(j, n-j|\hat{A}_U, N-\hat{A}_U)=\alpha_U U \tag{3.17}$$

donde Pr es la probabilidad para la distribución hipergeométrica, según se definió en (3.15).

Cuando α_U se escoge por adelantado (3.17) requiere en general un valor no entero para satisfacerla, mientras que conceptualmente $\hat{A}_U$ debería ser un número entero. En la práctica elegimos $\hat{A}_U$ como el valor entero más pequeño de A tal que el lado izquierdo de (3.17) es menor que, o igual a, α_U. En la misma forma el límite de confianza inferior $\hat{A}_L$ es el valor entero más grande tal que

$$\sum_{j=a}^{n}\Pr(j, n-j|\hat{A}_L, N-\hat{A}_L)\leq\alpha_L \tag{3.18}$$

Los límites de confianza para P se encuentran entonces al tomar $\hat{P}_U=\hat{A}_U/N$, $\hat{P}_L=\hat{A}_L/N$.

Existen numerosos métodos disponibles para calcular los límites de confianza.

Métodos exactos

Chung y DeLury (1950) presentan gráficas con los límites de confianza para P al 90, 95 y 99% con N = 500, 2 500 y 10 000. Los valores para tamaños de población intermedia se obtienen por interpolación. Lieberman y Owen (1961) presentan tablas para térmi-

nos individuales y acumulativos de la distribución hipergeométrica, pero N se extiende solamente hasta 100.

La aproximación normal

De la varianza estimada de p obtenida de la Ec. (3.8), se deriva que una forma de la aproximación normal para obtener los límites de confianza para P es

$$p \pm \left[t\sqrt{1-f}\sqrt{pq/(n-1)} + \frac{1}{2n} \right] \qquad (3.19)$$

donde $f = n/N$ y t es el desvío normal correspondiente a la probabilidad de confianza. El uso del término más familiar $\sqrt{pq/n}$ raras veces produce diferencias apreciables. El último término de la derecha es una corrección por continuidad. Esto sólo da lugar a una ligera mejora en la aproximación. Sin embargo, sin la corrección, la aproximación normal usualmente da un intervalo de confianza muy estrecho.

El error en la aproximación normal depende de todas las cantidades n, p, N, α_U y α_L. La cantidad para la cual el error es más sensible es np o más específicamente, el número observado en la clase *más pequeña*. La Tabla 3.3 presenta las reglas de trabajo para decidir cuándo se puede usar la aproximación normal (3.19).

TABLA 3.3. VALORES MÁS PEQUEÑOS DE np PARA USO DE LA APROXIMACIÓN NORMAL

p	np = Núm. observado en la clase *más pequeña*	n = Tamaño de muestra
0.5	15	30
0.4	20	50
0.3	24	80
0.2	40	200
0.1	60	600
0.05	70	1400
~0*	80	∞

* Esto significa que p es extremadamente pequeña, de tal manera que np sigue la distribución de Poisson.

Las reglas en la Tabla 3.3 están construidas de tal manera que con los límites de confianza al 95% la verdadera frecuencia con la que los límites no incluyen a P no es más grande que 5.5%. Además, la probabilidad de que el límite superior sea inferior a P está entre 2.5 y 3.5% y la probabilidad de que el límite inferior exceda a P está entre 2.5 y 1.5%.

Ejemplo 1. En una muestra aleatoria simple de tamaño 100, obtenida de una población de tamaño 500, existen 37 unidades en la clase C. Encontrar los límites de confianza al 95% para la proporción y para el número total en la clase C en la población. En este ejemplo:

$$n = 100, \qquad N = 500, \qquad p = 0.37$$

El ejemplo es del tipo para el cual se recomienda la aproximación normal. El error estándar estimado de p es

$$\sqrt{(1-f)pq/(n-1)} = \sqrt{(0.8)(0.37)(0.63)/99} = 0.0434$$

La corrección por continuidad, $1/2n$, es igual a 0.005. Por lo tanto, los límites al 95% para P se estiman como

$$0.37 \pm (1.96 \times 0.0434 + 0.005) = 0.37 \pm 0.090$$

$$\hat{P}_L = 0.280, \qquad \hat{P}_U = 0.460$$

Los límites según las gráficas de Chung y DeLury son 0.285 y 0.462, respectivamente.

Para encontrar los límites para el número total en la clase C en la población, multiplicamos por N y obtenemos 140 y 230, respectivamente.

Aproximaciones binomiales

Cuando la aproximación normal no es aplicable, los límites para P pueden encontrarse utilizando las tablas binomiales (Sec. 3.4) y ajustarlas, si es necesario, para tomar en consideración la cpf. La Tabla VIII1 en las *tablas estadísticas* de Fisher y Yates (1957) da los límites de confianza binomiales para P para cualquier valor de n y es una alternativa útil para las tablas binomiales ordinarias. El Ej. 2 muestra cómo se calcula la aproximación binomial.

Ejemplo 2. Para otro atributo en la muestra del Ej. 1, nueve de las 100 unidades pertenecen a la clase C. De la tabla de Romig para $n = 100$, los límites de confianza al 95% para P son 0.041 y 0.165. (Las tablas de Fisher-Yates dan 0.042 y 0.164.) Si f, la fracción de muestreo, es menor que el 5%, los límites que se encuentran por este procedimiento son lo suficientemente aproximados para la mayoría de los propósitos. En este ejemplo, $f = 0.2$ y es necesario un ajuste.

Para aplicar el ajuste, reducimos el intervalo entre p y cada uno de los límites por el factor $\sqrt{1-f} = \sqrt{0.8} = 0.894$. Los límites ajustados son entonces:

$$\hat{P}_L = 0.090 - (0.894)(0.090 - 0.041) = 0.046$$

$$\hat{P}_U = 0.090 + (0.894)(0.165 - 0.090) = 0.157$$

Los límites que se obtienen de las gráficas de Chung y DeLury son 0.045 y 0.157, respectivamente.

Ejemplo 3. En los registros de auditoría en los cuales se necesita una tasa muy pequeña de error, el límite de confianza superior para A es de particular importancia. Supongamos que se verifican 200 de 1 000 registros y que se acepta el conjunto de 1 000 si no se encuentran errores. Se han creado tablas especiales que dan los límites de confianza superiores para el número de errores en un grupo de observaciones. Una buena aproximación resulta de las siguientes relaciones. La probabilidad de no encontrar errores en n cuando A errores están presentes en N es, de la distribución hipergeométrica,

$$\frac{(N-A)(N-A-1)\dots(N-A-n+1)}{N(N-1)\dots(N-n+1)} \doteq \left(\frac{N-A-u}{N-u}\right)^n$$

donde $u = (n-1)/2$. Por ejemplo, con $n = 200$, $A = 10$, $N = 1\,000$, la aproximación da $(890.5/900.5)^{200}$, la cual, por logaritmos resulta ser 0.107. Por lo tanto, $A = 10$ (una tasa de error del 1%) es aproximadamente el límite de confianza superior al 90% para el número de errores en el grupo.

3.7 CLASIFICACION EN MAS DE DOS CLASES

Frecuentemente, en la presentación de resultados, se clasifican las unidades en más de dos clases. Así, una muestra de una población humana se puede arreglar en 15 grupos correspondientes a intervalos de 5 años de edad. Aun cuando una pregunta se supone que puede contestarse por un simple "sí" o "no", el resultado realmente obtenido puede caer en cualquiera de las siguientes 4 clases: "sí", "no", "no sé", y "sin respuesta". La extensión de la teoría a estos casos la ilustra una situación en la que existen 3 clases.

Supongamos que el número que cae en la i-ésima clase es A_i en la población y a_i en la muestra, donde

$$N = \sum A_i, \qquad n = \sum a_i, \qquad P_i = \frac{A_i}{N}, \qquad p_i = \frac{a_i}{n}$$

Cuando la muestra de tamaño n es pequeña en relación al total de A_i, las probabilidades P_i se pueden considerar efectivamente constantes a lo largo de la obtención de la muestra. La probabilidad de obtener la muestra observada está dada por la expresión *multinomial*.

$$\Pr(a_i) = \frac{n!}{a_1!a_2!a_3!} P_1^{a_1} P_2^{a_2} P_3^{a_3} \tag{3.20}$$

Esta es una extensión correcta de la distribución binomial y es una buena aproximación cuando la fracción de muestreo es pequeña.

La expresión correcta de la probabilidad de obtener la muestra observada es

$$\Pr(a_i|A_i) = \binom{A_1}{a_1}\binom{A_2}{a_2}\binom{A_3}{a_3}\Big/\binom{N}{n} \tag{3.21}$$

Esta expresión es la extensión natural de (3.15), Sec. 3.5, para la distribución hipergeométrica. El numerador es el número de muestras distintas de tamaño n que se pueden formar con a_1 unidades en la clase 1, a_2 en la clase 2 y a_3 en la clase 3.

3.8 LIMITES DE CONFIANZA CUANDO EXISTEN MAS DE DOS CLASES

Deben distinguirse dos casos diferentes.

Caso 1. Calculamos

$$p = \frac{\text{Número en cualquier clase en la muestra}}{n} = \frac{a_1}{n}$$

o

$$p = \frac{\text{Número total en un grupo de clases}}{n} = \frac{a_1 + a_2 + a_3}{n}$$

En cualquiera de estas situaciones, a pesar de que la clasificación original contiene más de dos clases, p misma se obtiene de una subdivisión de las n unidades en solamente dos clases. La teoría antes presentada se aplica a este caso. Los límites de confianza se calculan como se describió en la Sec. 3.6.

Caso 2. Algunas veces se omiten ciertas clases, al calcular p de una subdivisión en dos partes de las clases restantes. Por ejemplo, podríamos omitir las personas que no saben, o no dieron, una respuesta y considerar la proporción del número de respuestas "sí" respecto al número de respuestas "sí" "no". Las proporciones que estructuralmente son de este tipo, con frecuencia son de interés en encuestas por muestreo. El denominador de una proporción tal no es n sino un número n' más pequeño.

Aunque n' varía de muestra a muestra, pueden seguir usándose los resultados previos, al considerar la distribución condicional de p y las muestras en las que tanto n como n' son fijas. Este recurso ya se empleó en la Sec. 2.12. Supongamos que

$$p = \frac{a_1}{a_1 + a_2}, \qquad n' = a_1 + a_2, \qquad n = a_1 + a_2 + a_3$$

por lo que a_3 es el número en la muestra que cae en las clases que, por el momento, no estamos interesados. Entonces, como se muestra en la siguiente sección, la distribución condicional de a_1 y

a_2 es la distribución hipergeométrica obtenida cuando la muestra es de tamaño n' y el tamaño de la población es $N' = A_1 + A_2$. Por lo tanto, utilizando la Ec. (3.19), la aproximación normal para los límites de confianza condicionales de $P = A_1/(A_1 + A_2)$ es

$$p \pm \left[t\sqrt{\left(1-\frac{n'}{N'}\right)\frac{pq}{(n'-1)}+\frac{1}{2n'}} \right] \tag{3.22}$$

Si el valor de N' no es conocido, n/N se puede sustituir por n'/N' en el término cpf de la Ec. (3.22).

3.9 LA DISTRIBUCION CONDICIONAL DE p

Para encontrar esta distribución, restringimos nuestra atención a muestras de tamaño n en las que $n' = a_1 + a_2$ cae en las clases 1 y 2. El número de muestras distintas de este tipo es

$$\binom{N'}{n'}\binom{N-N'}{n-n'} = \binom{A_1+A_2}{a_1+a_2}\binom{A_3}{a_3} \tag{3.23}$$

Dentro de estas muestras, el número que tiene a_1 en la clase 1 y a_2 en la clase 2 ya ha sido dado como el numerador en (3.21), Sec. 3.7. Al dividir este numerador por (3.23), tenemos

$$\Pr(a_1|A_1, A_2, n, n') = \binom{A_1}{a_1}\binom{A_2}{a_2}\Big/\binom{A_1+A_2}{a_1+a_2} \tag{3.24}$$

Esta es una distribución hipergeométrica ordinaria para una muestra de tamaño n' de una población de tamaño $N' = A_1 + A_2$.

Ejemplo. Considere una población que consiste en 5 unidades *b*, *c*, *d*, *e*, *f*, que pertenecen a tres clases.

Clase	A_i	Unidades indicadas por
1	1	*b*
2	2	*c, d*
3	2	*e, f*

Con muestras aleatorias de tamaño 3, deseamos estimar $P = A_1/(A_1 + A_2)$, o en este caso 1/3. Entonces, $N = 5$ y $N' = 3$.

Hay diez posibles muestras de tamaño 3, todas ellas con iguales probabilidades iniciales. Estas se agrupan de acuerdo con el valor de n'.

$n' = 1$

Muestra	a_1	a_2	p	Probabilidad condicional	$(p-P)$
bef	1	0	1	$\frac{1}{3}$	$\frac{2}{3}$
cef o *def*	0	1	0	$\frac{2}{3}$	$-\frac{1}{3}$

Si las muestras se especifican por los valores a_1, a_2, sólo son posibles dos tipos: $a_1 = 1$, $a_2 = 0$; $a_1 = 0$, $a_2 = 1$. Sus probabilidades condicionales, 1/3 y 2/3, respectivamente, concuerdan con la expresión general (3.24). Además,

$$E(p) = \tfrac{1}{3}$$

$$\sigma_p^2 = \left(\frac{1}{3}\right)\left(\frac{4}{9}\right)+\left(\frac{2}{3}\right)\left(\frac{1}{9}\right) = \frac{6}{27} = \frac{2}{9}$$

La estimación p es insesgada y su varianza concuerda con la fórmula general

$$\sigma_p^2 = \left(\frac{N'-n'}{N'-1}\right)\frac{PQ}{n'} = \left(\frac{3-1}{3-1}\right)\left(\frac{1}{3}\right)\left(\frac{2}{3}\right) = \frac{2}{9}$$

Para $n' = 2$ hay seis posibles muestras que sólo dan 2 grupos de valores de a_1, a_2.

$n' = 2$

Muestra	a_1	a_2	p	Probabilidad condicional	$(p-P)$
bce, bcf, bde, o *bdf*	1	1	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{6}$
cde o *cdf*	0	2	0	$\frac{1}{3}$	$-\frac{1}{3}$

La estimación es nuevamente insesgada y su varianza es

$$\sigma_p^2 = \left(\frac{2}{3}\right)\left(\frac{1}{36}\right)+\left(\frac{1}{3}\right)\left(\frac{1}{9}\right) = \frac{1}{18}$$

lo cual se puede verificar utilizando la fórmula general. Nótese que la varianza es solamente ¼ parte de la obtenida cuando $n' = 1$. Con un enfoque condicional, la varianza cambia con la configuración de la muestra que se obtuvo.

Para $n' = 3$, existe una sola muestra posible, *bcd*. Esto da la fracción de población correcta, ⅓. La varianza condicional de p es 0, como está indicado por la fórmula general, la cual se reduce a 0 cuando $N' = n'$.

3.10 PROPORCIONES Y TOTALES SOBRE SUBPOBLACIONES

Si se van a hacer estimaciones separadas para cada una de un número de subpoblaciones o dominios de estudio, a las que se desti-

nan las unidades en la muestra, es posible aplicar los resultados en las Secs. 3.8 y 3.9. Los datos de la muestra se pueden presentar como sigue:

Clase	Dominio 1		Dominio 2		...	Dominio k		Total
	C	C'	C	C'	...	C	C'	
Núm. de unidades	a_1	a_1'	a_2	a_2'	...	a_k	a_k'	n

De las n unidades, se encuentra que $(a_1 + a_1')$ caen en el dominio 1 y de éstas, a_1 pertenece a la clase C. La proporción que pertenece a la clase C en el dominio 1 se estima por $p_1 = a_1/(a_1 + a_1')$. La distribución de frecuencia y los límites de confianza para p_1 se discutieron en el Caso 2 de las Secs. 3.8 y 3.9.

Para estimar el número *total* de unidades A_1 pertenecientes a la clase C en el dominio 1, hay dos posibilidades. Si N_1, el número total de unidades en el dominio 1 en la población, es conocido, podemos usar la estimación condicional.

$$\hat{A}_1 = N_1 p_1 = \frac{N_1 a_1}{a_1 + a_1'} \tag{3.25}$$

Su error estándar se calcula como sigue:

$$s(\hat{A}_1) = N_1\sqrt{1-(n_1/N_1)}\sqrt{p_1 q_1/(n_1-1)} \tag{3.26}$$

donde $n_1 = a_1 + a_1'$.

Si N_1 no es conocido, la estimación es

$$\hat{A}_1' = \frac{Na_1}{n} \tag{3.27}$$

con un error estándar estimado

$$s(\hat{A}_1') = N\sqrt{1-(n/N)}\sqrt{pq/(n-1)} \tag{3.28}$$

donde $p = a_1/n$.

3.11 COMPARACIONES ENTRE DOMINIOS DIFERENTES

Puesto que las proporciones son estimadas independientemente en los diferentes dominios, las comparaciones entre dichas proporciones se hacen por métodos estándar elementales. Por ejemplo, para probar si la proporción $p_1 = a_1/(a_1 + a_1')$ difiere significativamente de la proporción $p_2 = a_2/(a_2 + a_2')$, formamos una tabla 2×2 común y corriente,

	Dominio 1	Dominio 2
C	a_1	a_2
C'	a_1'	a_2'
Total	n_1	n_1'

La prueba de χ^2 ordinaria (Fisher, 1958) o la aproximación normal a la distribución de $(p_1 - p_2)$ es apropiada. En forma semejante, comparaciones entre proporciones para más de dos dominios se hacen utilizando los métodos correspondientes a la tabla de contingencia $2 \times k$.

En ocasiones se desea probar si a_1 difiere significativamente de a_2; por ejemplo, si el *número* de republicanos que están a favor de una proposición es mayor que el *número* de demócratas también a favor. Con base en la hipótesis nula de que estos dos números son iguales en la población, el total $n' = a_1 + a_2$ en las dos clases en cuestión, debe dividirse con igual probabilidad entre ellas. Consecuentemente, podemos considerar a a_1 como un número binomial de éxitos en n' ensayos, con probabilidad de éxito = ½ en la hipótesis nula. Se puede verificar que el desvío normal (corregido por continuidad) es

$$\frac{2(|a_1 - \frac{1}{2}n'| - \frac{1}{2})}{\sqrt{n'}}$$

3.12 ESTIMACION DE PROPORCIONES EN EL MUESTREO POR CONGLOMERADOS

Como se mencionó en la Sec. 3.2, los métodos precedentes no son válidos si cada unidad es un conglomerado de elementos y estamos estimando la proporción de estos que pertenecen a la clase C.

Si cada unidad contiene el mismo número de elementos m, sea $p_i = a_i/m$ la proporción de elementos en la i-ésima unidad perteneciente a la clase C. La proporción de elementos que pertenecen a C en la muestra es

$$p = \frac{\sum a_i}{nm} = \frac{1}{n}\sum^{n} p_i$$

es decir, la estimación p es la media no ponderada de las cantidades p_i. En consecuencia, si y_i se reemplaza por p_i, las fórmulas del Cap. 2 se pueden aplicar directamente para dar la varianza verdadera y estimada de p.

$$V(p)=\frac{1-f}{n}\frac{\sum^{N}(p_i-P)^2}{N-1} \tag{3.29}$$

Una estimación muestral insesgada de esta varianza es

$$v(p)=\frac{1-f}{n}\frac{\sum^{n}(p_i-p)^2}{n-1} \tag{3.30}$$

Ejemplo 1. Un grupo de 61 pacientes leprosos fue tratado con una droga durante 48 semanas. Para medir el efecto de la droga en el bacilo de la lepra, se probó bacteriológicamente la presencia del bacilo en 6 lugares diferentes del cuerpo de cada paciente. De las 366 observaciones, 153, o sea el 41.8% fueron negativas. ¿Cuál es el error estándar de este porcentaje?

Este ejemplo proviene de un experimento controlado, no de una encuesta, pero ilustra qué tan errónea puede ser la fórmula binomial. Utilizando la fórmula binomial tenemos $n = 366$ y

$$\text{e.e.}\,(p)=\sqrt{pq/(n-1)}=\sqrt{(41.8)(58.2)/365}=2.58\%$$

Cada paciente es un conglomerado con $m = 6$ elementos (lugares). Para encontrar el error estándar con la fórmula correcta, necesitamos la distribución de frecuencia de los 61 valores de p_i. Es más conveniente tabular la distribución de y_i, el número de lugares negativos por paciente. Con p_i expresado en porcentaje, $p_i = 100y_i/6$. De la distribución presentada en la Tabla 3.4 encontramos $\sum fy^2 = 669$ y

$$\text{e.e.}\,(\bar{y})=\sqrt{\frac{\sum f_i(y_i-\bar{y})^2}{n(n-1)}}=\sqrt{\frac{669-[(153)^2/61]}{(61)(60)}}=0.279$$

Por lo tanto,

$$\text{e.e.}\,(p)=\frac{100}{6}\text{e.e.}\,(\bar{y})=4.65\%$$

Esta cifra es aproximadamente 1.8 veces el valor dado por la fórmula binomial. La fórmula binomial requiere la suposición de que los resultados obtenidos en los lugares diferentes en el mismo paciente sean independientes, aun cuando en realidad presentan una fuerte correlación positiva. La última línea en la Tabla 3.4 muestra los números esperados de pacientes con 0, 1, 2, ... lugares negativos, calculados con la binomial $(0.58 + 0.42)^6$. Nótese el marcado exceso de frecuencias f de pacientes con 0,5 y 6 lugares negativos.

TABLA 3.4. NÚMERO DE LUGARES NEGATIVOS POR PACIENTE

$y_i = 6p_i/100$	0	1	2	3	4	5	6	Total
f	17	11	4	4	7	14	4	61
fy_i	0	11	8	12	28	70	24	153
f_{exp}	2.3	10.1	18.3	17.6	9.6	2.8	0.3	61.0

Si el tamaño de un conglomerado no es constante, sea m_i, el número de elementos en el i-ésimo conglomerado y sea $p_i = a_i/m_i$. La proporción de unidades pertenecientes a la clase C en la muestra es

$$p=\frac{\sum^{n} a_i}{\sum^{n} m_i} \tag{3.31}$$

Estructuralmente, ésta es una estimación de razón típica discutida en la Sec. 2.11 y después en el Cap. 6. Es ligeramente sesgada aun cuando el sesgo rara vez tiene importancia práctica.

Si ponemos a_i por y_i y m_i por x_i en la Ec. (2.39), la varianza aproximada de p es

$$V(p)\doteq\frac{1-f}{n\bar{M}^2}\frac{\sum(a_i-Pm_i)^2}{N-1} \tag{3.32}$$

donde P es la proporción de elementos en C en la población y $\bar{M} = \sum^{N} m_i/N$ es el número promedio de elementos por conglomerado. Una expresión alternativa es

$$V(p)\doteq\frac{1-f}{n}\sum^{N}\left(\frac{m_i}{\bar{M}}\right)^2\frac{(p_i-P)^2}{N-1} \tag{3.33}$$

Esta forma muestra que la varianza aproximada involucra una suma ponderada de cuadrados de desviaciones de p_i respecto al valor de la población P.

Para la varianza estimada tenemos

$$v(p)=\frac{1-f}{n\bar{m}^2}\frac{\sum a_i^2-2p\sum a_im_i+p^2\sum m_i^2}{n-1} \tag{3.34}$$

donde $\bar{m} = \sum m_i/n$ es el número promedio de elementos por conglomerado en la muestra.

Ejemplo 2. Una muestra aleatoria simple de 30 familias se obtuvo de un censo hecho en 1947 en los Distritos 6 y 7 del Eastern Health District de Baltimore. La población contiene aproximadamente 15 000 familias. En la Tabla 3.5 las personas en cada familia se clasifican, (a) de acuerdo a si han consultado un médico en los últimos 12 meses, (b) de acuerdo a su sexo.

Nuestro propósito es comparar la fórmula de razón con la binomial que es inapropiada. Considerando primero la proporción de personas que han consultado un doctor. Con la fórmula binomial, tendremos:

$$n=104, \qquad p=\frac{30}{104}=0.2885$$

TABLA 3.5. INFORMACIÓN DE UNA MUESTRA ALEATORIA SIMPLE DE 30 FAMILIAS

Familia número	Núm. de personas m_i	Número de		Médico visto en el último año	
		Hombres a_i	Mujeres	Sí a_i	No
1	5	1	4	5	0
2	6	3	3	0	6
3	3	1	2	2	1
4	3	1	2	3	0
5	2	1	1	0	2
6	3	1	2	0	3
7	3	1	2	0	3
8	3	1	2	0	3
9	4	2	2	0	4
10	4	3	1	0	4
11	3	2	1	0	3
12	2	1	1	0	2
13	7	3	4	0	7
14	4	3	1	4	0
15	3	2	1	1	2
16	5	3	2	2	3
17	4	3	1	0	4
18	4	3	1	0	4
19	3	2	1	1	2
20	3	1	2	3	0
21	4	1	3	2	2
22	3	2	1	0	3
23	3	2	1	0	3
24	1	0	1	0	1
25	2	1	1	2	0
26	4	3	1	2	2
27	3	1	2	0	3
28	4	2	2	2	2
29	2	1	1	0	2
30	4	2	2	1	3
Totales	104	53	51	30	74

Luego

$$v_{bin}(p)=\frac{pq}{n}=\frac{(0.2885)(0.7115)}{104}=0.00197$$

Con la fórmula de razón, notamos que hay 30 conglomerados y tomamos

$n=$ 30

$m_i=$ número total en la i-ésima familia

$a_i=$ número de personas en la i-ésima familia que ha visto al doctor

$p=$ 0.2885, como antes

$$m=\frac{104}{30}=3.4667$$

$\sum a_i^2=86; \sum m_i^2=404; \sum a_i m_i=113$

La cpf se puede ignorar. Por lo tanto (utilizando la Ec. 3.34),

$$v(p)=\frac{(86)-2(0.2885)(113)+(0.2885)^2(404)}{(30)(29)(3.4667)^2}=0.00520$$

La varianza dada por el método de razón, 0.00520, es mucho más grande que la obtenida con la fórmula binomial, 0.00197. Por diversas razones, las familias difieren en la frecuencia en que sus miembros han consultado al médico. Para la muestra completa, la proporción de quienes lo consultaron es solamente un poco más de uno en cuatro, pero hay varias familias en las que cada miembro de ellas ha visto al médico. Se obtienen resultados similares para cualquier característica en la cual los miembros de la misma familia tienden a actuar en la misma forma.

En la estimación de la proporción de hombres en la población, los resultados son diferentes. Por el mismo método de cálculo, encontramos

Fórmula binomial: $v(p)=0.00240$
Fórmula de razón: $v(p)=0.00114$

En este caso, la fórmula binomial *sobreestima* la varianza. La razón por la que ocurre esto es interesante. La mayoría de las familias se forman como resultado de un matrimonio; por lo tanto, contienen cuando menos un hombre y una mujer. En consecuencia, la proporción de hombres por familia se separa menos de ½ de lo que se podría esperar con la fórmula binomial. Ninguna de las 30 familias, excepto una con un solo miembro, está compuesta totalmente por hombres o por mujeres. Si la distribución binomial fuese aplicable, con una verdadera P de aproximadamente ½, las familias con todos los miembros de un mismo sexo constituirían la cuarta parte de la familia de tamaño 3 y ⅛ de las familias de tamaño 4. Esta propiedad de la distribución del sexo ha sido discutida por Hansen y Hurwitz (1942). Kish (1957) ha dado otras ilustraciones del error cometido con el uso impropio de la fórmula binomial en investigaciones sociológicas.

EJERCICIOS

3.1. En una población con $N = 6$, $A = 4$ y $A' = 2$, calcular el valor de a para todas las posibles muestras aleatorias simples de tamaño 3. Verificar los teoremas dados para la media y varianza de $p = a/n$. Verificar que

$$\frac{N-n}{(n-1)N}pq$$

es una estimación insesgada de la varianza de p.

3.2. En una muestra aleatoria simple de 200 obtenida de una población de 2 000 colegios, 120 de éstos estuvieron a favor de una propuesta, 57 se opusieron y 23 se abstuvieron de opinar. Estimar los límites de confianza al 95% para el número de colegios en la población, que favorecieron la propuesta.

3.3. En los resultados de la muestra anterior, ¿se obtiene una evidencia contundente de que la mayoría de los colegios en la población favorecieron la propuesta?

3.4. Una población con $N = 7$ consiste de los elementos B_1, C_1, C_2, C_3, D_1, D_2 y D_3. Se toma una muestra aleatoria simple de tamaño 4 con el fin de estimar la proporción de C'_s respecto a $C'_s + D'_s$. Calcule las distribuciones condicionales de esta proporción, p, y verifique la fórmula de su varianza condicional.

3.5. En el ejercicio anterior, ¿cuál es la probabilidad de que una muestra de tamaño 4 contenga B_1? Encuentre la varianza promedio de p en el Ejerc. 3.4 en todas las muestras aleatorias simples de tamaño 4. Es 0.0393 contrariamente a 0.025 con $N = 6$, $n = 4$ y B_1 ausente.

3.6. Se eligió una muestra aleatoria simple de 290 familias de una área de la ciudad conteniendo 14 828 familias. A cada familia se le preguntó si la casa era suya o rentada y también si tenían el uso exclusivo de un baño interior. Los resultados fueron como sigue:

	Propia		Rentada		Total
Uso exclusivo de un baño	Sí	No	Sí	No	
	141	6	109	34	290

(a) para las familias que rentan, estimar el porcentaje en el área de familias que cuentan con un baño interior de uso exclusivo y dar el error estándar de su estimación; (b) estimar el número total de familias que rentan casa en el área y que no tienen un baño interior para uso exclusivo y dar el error estándar de esta estimación.

3.7. Si en el Ej. 3.6, el número total de familias que rentan casa en el área de la ciudad es de 7 526, construya una nueva estimación del número de arrendatarios que no cuentan con la facilidad de un baño exclusivo y del error estándar de esta estimación.

3.8. Para estimar el número total de unidades en la clase C en el dominio 1 (Sec. 3.10), la estimación $\hat{A}_1 = N_1 p_1$ se recomendó si se conocía N_1, contra $\hat{A}_1' = Na_1/n$ si N_1 no era conocido. Ignorando la cpf, demostrar que en muestras grandes, la razón de la varianza de $\hat{A}_1$ respecto a la de $\hat{A}_1'$ es, aproximadamente, $Q_1/(Q_1 + P_1\pi)$, donde π es la proporción de la población que no

se encuentra en el dominio 1, y P_1, como en la Sec. 3.10, es la proporción de unidades en el dominio 1 que pertenecen a la clase C. Establezca las condiciones en las que el conocimiento de N_1 da lugar a reducciones grandes de la varianza.

3.9. En una muestra aleatoria simple de tamaño 5 obtenida de una población de tamaño 30, ninguna unidad en la muestra pertenecía a la clase C. Utilizando la distribución hipergeométrica, encuentre el límite superior respecto al número A de unidades pertenecientes a la clase C en la población, correspondiente a una probabilidad de confianza del 95% sobre un lado de la distribución. Encuentre también la aproximación respecto a A_U obtenida al calcular el límite binomial superior al 95% P_U y al acortar el intervalo como se describió en la Sec. 3.6. Pruebe también el método descrito en la Pág. 90, Ej. 3.

3.10. Un servicio médico estudiantil tiene un registro del número total de estudiantes elegibles N y del número total de visitas Y, que hicieron los estudiantes durante un año. Algunos no hicieron visitas. El servicio médico desea estimar el número promedio de visitas Y/N_1 realizadas por los N_1 estudiantes que hicieron, por lo menos, una visita, pero no se conoce el valor de N_1. Se toma una muestra aleatoria simple de n estudiantes elegibles. En ella, n_1 estudiantes de los n seleccionados hicieron por lo menos una visita y el número total de visitas fue y. Ignore la cpf en esta pregunta, (a) demuestre que y/n_1 es una estimación insesgada de Y/N_1 y que su varianza condicional es S^2/n_1, donde S^2 es la varianza del número de visitas entre los estudiantes que por lo menos hicieron una. (b) Un segundo método para estimar Y/N_1 es usar $\hat{N}_1 = Nn_1/n$ como una estimación de N_1 y, por lo tanto, Yn/Nn_1 como una estimación de Y/N_1. Demuestre que esta estimación es sesgada y que la razón del sesgo respecto al valor verdadero Y/N_1 es aproximadamente $(N - N_1)/nN_1$. Encuentre una expresión aproximada para la varianza de la estimación Y_n/Nn_1 y demuestre que la estimación en (a) tiene una varianza mayor si

$$S^2 > \frac{(N-N_1)n_1}{N_1 n}\left(\frac{Y}{N_1}\right)^2$$

Nota. Si p es una estimación binomial de P, basado en n ensayos, entonces aproximadamente,

$$E\left(\frac{1}{p}\right) = \frac{1}{P} + \frac{Q}{nP^2}, \qquad V\left(\frac{1}{p}\right) = \frac{Q}{nP^3}$$

3.11. ¿Cuál de las dos estimaciones previas parece más precisa en las circunstancias siguientes? $N = 2\,004$, $Y = 3\,011$. La muestra con $n = 100$ demostró que 73 estudiantes hicieron cuando menos una visita. El número total de visitas fue de 152 y la varianza estimada s^2 fue 1.55.

3.12. Una muestra aleatoria simple de n unidades conglomeradas, cada una con m elementos, se obtiene de una población en la cual la proporción de elementos en la clase C es P. Conforme varía la correlación intraconglomerada, ¿cuáles son los posibles valores más altos y más bajos de la verdadera varianza de p (la estimación muestral de P) y cómo se comparan con la varianza binomial? Ignore la cpf.

3.13. En la muestra de 30 familias de la Tabla 3.5, la información presentada a continuación se refiere al número de visitas al dentista realizadas

en el último año. Estimar la varianza de la proporción de personas que vieron a un dentista y comparar ésta con la estimación de la varianza binomial.

3.14. En el muestreo de una característica no común, un método consiste en seguir sacando una muestra aleatoria simple hasta que se encuentren m unidades que poseen esta característica no común (Haldane, 1945) en donde m se elige por adelantado. Si se ignora la cpf, probar que la probabilidad de que la muestra total requerida sea de tamaño n es

$$\frac{(n-1)!}{(m-1)!(n-m)!}P^mQ^{n-m} \qquad (n \geq m)$$

donde P es la frecuencia de la característica rara. Encontrar el tamaño promedio de la muestra total y demostrar que si $m > 1$, $p = (m-1)/(n-1)$ es una estimación insesgada de P. (Para una mayor discusión véase Finney, 1949, y Sandelius, 1951, quienes consideran un plan en el cual el muestreo continúa hasta que se encuentren m individuos o el tamaño total de la muestra alcance un límite n_0 predeterminado.) Véase también la Sec. 4.5.

Núm. de personas	Dentista visto		Núm. de personas	Dentista visto	
	Sí	No		Sí	No
5	1	4	5	1	4
6	0	6	4	4	0
3	1	2	4	1	3
3	2	1	3	1	2
2	0	2	3	0	3
3	0	3	4	1	3
3	1	2	3	0	3
3	1	2	3	1	2
4	1	3	1	0	1
4	0	4	2	0	2
3	1	2	4	0	4
2	0	2	3	1	2
7	2	5	4	1	3
4	1	3	2	0	2
3	0	3	4	0	4

CAPITULO 4

La Estimación del Tamaño de la Muestra

4.1 UN EJEMPLO HIPOTETICO

Al planear una encuesta por muestreo, siempre se alcanza una etapa en donde hay que tomar una decisión respecto al tamaño de la muestra. Esta decisión es importante. Una muestra demasiado grande implica un despilfarro de recursos y una muy pequeña disminuye la utilidad de los resultados. La decisión no siempre puede tomarse satisfactoriamente; a menudo no disponemos de la suficiente información para saber si el tamaño de la muestra seleccionada, es el óptimo. La teoría del muestreo proporciona un marco dentro del cual se puede pensar inteligentemente respecto a este problema.

A continuación veremos un ejemplo hipotético en el que se resaltan los pasos necesarios para llegar a una solución. Un antropólogo está preparando un estudio sobre los habitantes de cierta isla. Entre otras cosas, desea estimar el porcentaje de habitantes que pertenecen al grupo sanguíneo O. Le han asegurado la cooperación necesaria para extraer una muestra aleatoria simple. ¿Qué tan grande debe ser la muestra?

Este asunto no se puede discutir sin antes contestar a otra pregunta. ¿Qué tan exactamente desea el antropólogo conocer el porcentaje de personas del grupo sanguíneo O? A lo anterior, nuestro amigo contesta que estará satisfecho si el porcentaje no contiene un error mayor del $\pm 5\%$, o sea que si la muestra indica que el 43% de la población es del grupo sanguíneo O, el porcentaje para toda la isla se encontrará con certeza entre el 38 y el 48%.

Para evitar malentendidos, es aconsejable aclarar al antropólogo que no le podemos asegurar una exactitud dentro de un 5%, a menos que se analice el tipo sanguíneo de todos los habitantes. Por muy grande que se tome n, existe la posibilidad de una muestra desafortunada que presente un error mayor al 5% deseado. El antropólogo contesta fríamente que lo sabe, que acepta el riesgo de una posibilidad en veinte de obtener una muestra poco afortunada y que todo lo que pide es un valor de n y no una clase de estadística.

Ahora estamos en posición de estimar en términos generales el valor de n. Para simplificar las cosas se ignora la cpf y el porcentaje de muestra p se supone normalmente distribuido. Si estas hipótesis son razonables, puede verificarse cuando se conoce el valor n inicial.

En términos técnicos, p debe encontrarse en el intervalo $(P \pm 5)$, excepto para un caso en 20. Dado que p se supone normalmente distribuida alrededor de P, se encontrará en el intervalo $(P \pm 2\sigma_p)$, salvo para una posibilidad en 20. Además,

$$\sigma_p \doteq \sqrt{PQ/n}$$

De modo que podemos escribir

$$2\sqrt{PQ/n} = 5 \quad \text{o} \quad n = \frac{4PQ}{25}$$

En este momento aparece una dificultad común a todos los problemas para la estimación del tamaño de la muestra. Se obtuvo una fórmula para n, pero n depende de una propiedad de la población sujeta al muestreo. En este caso, la propiedad es la cantidad P que desearíamos medir. Por lo tanto, preguntamos al antropólogo si nos puede dar una idea del valor que se espera de P. El contesta que con base en datos obtenidos previamente con otros grupos étnicos, y de acuerdo a sus especulaciones sobre la historia racial de la isla, le sorprendería que P estuviera fuera del intervalo del 30 al 60%.

Esta información será suficiente para proporcionar una respuesta útil. Para cualquier valor de P entre 30 y 60, el producto PQ está entre 2 100 y un máximo de 2 500 con $P = 50$. El valor correspondiente de n está entre 336 y 400. Para no correr riesgos se toma 400 como estimación inicial de n.

Ahora pueden reexaminarse las hipótesis hechas en este análisis. Con $n = 400$ y p entre 30 y 60, la distribución de p debería estar cerca de la normal. Si se requiere o no la cpf, dependen del número de habitantes de la isla. Si la población excede a 8 000, la fracción de muestreo es menor al 5% y no se necesita ajuste para la cpf.

El método para aplicar el reajuste en caso necesario, se discutirá en la Sec. 4.4.

4.2 ANALISIS DEL PROBLEMA

Los pasos principales involucrados en la elección del tamaño de la muestra son los siguientes:

1. Debe existir algún enunciado respecto a lo que se espera de la muestra. Este puede darse en términos de límites de error deseados, como en el ejemplo anterior, o bien en términos de alguna decisión o acción que debe tomarse una vez que se conocen los resultados de la muestra. La responsabilidad de este enunciado es primordialmente de las personas que van a usar los resultados de la encuesta, aunque con frecuencia, dichas personas necesitan de una guía para expresar sus deseos en términos numéricos.
2. Se debe encontrar una ecuación que relacione n con la precisión deseada de la muestra. La ecuación variará según el contenido del enunciado de precisión y el tipo de muestreo propuesto. Una de las ventajas del muestreo probabilista es que permite la elaboración de esta ecuación.
3. Esta ecuación tendrá como parámetros ciertas propiedades desconocidas de la población, que deben estimarse para obtener resultados específicos.
4. Con frecuencia sucede que los datos estipulan para ciertas subdivisiones mayores de la población y que los límites de error deseados se establecen para cada subdivisión. De ser así, se hace un cálculo separado para el valor n en cada subdivisión y el n total se encuentra por adición.
5. Generalmente se mide más de un atributo o característica en una encuesta por muestreo: en ocasiones, el número de atributos es grande. Si se estipula un grado de precisión para cada atributo, los cálculos conducirán a una serie de valores conflictivos de n uno para cada atributo. Por lo tanto, debe encontrarse un método para reconciliar estos valores.
6. Finalmente, debe apreciarse el valor elegido de n, para que sea consistente con los recursos de muestreo disponibles. Esto exige una estimación del costo, trabajo, tiempo y materiales que se necesitan para obtener la muestra del tamaño propuesto. En ocasiones es claro que n debe reducirse drásticamente, y entonces es necesario tomar una decisión difícil, que es la de proceder con una muestra mucho más pequeña, lo que reduce la precisión, o bien, abandonar los esfuerzos hasta contar con mayores recursos.

En secciones subsecuentes se examinarán en detalle estas dificultades.

4.3 LA ESPECIFICACION DE LA PRECISION

La precisión deseada se puede establecer, al definir la cantidad de error tolerable en las estimaciones muestrales. Esta cantidad se determina mejor a la luz de los usos a que se destinan los resultados de la muestra. En ocasiones, es difícil saber qué tanto error *debería* tolerarse, particularmente cuando los resultados se destinan a varios fines. Supongamos que preguntamos al antropólogo por qué desea que el porcentaje del conjunto de personas en el grupo sanguíneo O tenga una corrección de un 5% en lugar de 4 o 6%. Su respuesta podría ser que los datos sobre el grupo sanguíneo se destinan principalmente a la clasificación racial y sospecha que los isleños pertenecen a un grupo racial con un valor de P aproximado al 35%, o a un tipo racial con un valor de P cercano al 50%. Un límite de error de 5% en la estimación le parece lo suficientemente pequeño para permitir la clasificación en uno de estos dos tipos. Sin embargo, se opondría en forma violenta a límites de error entre 4 o 6%.

De modo que el límite de error de 5% que escogió el antropólogo, es en cierta medida arbitrario. A este respecto, el ejemplo muestra cómo se suelen decidir los límites de error. De hecho el antropólogo, estaba más seguro de lo que quería, en comparación con muchos científicos y administradores. Cuando por primera vez se pregunta a estas personas el grado de precisión deseado, a menudo confiesan que nunca han considerado el asunto y no tienen idea de la respuesta. Me he percatado, gracias a la experiencia, que después de discutir el asunto, son capaces de indicar, al menos aproximadamente, el tamaño del límite de error que les parece razonable.

En muchas situaciones prácticas no podemos ir más adelante. Parte de la dificultad es que no se sabe lo suficiente sobre las consecuencias que producirían errores de diferentes magnitudes en las decisiones prácticas que se toman con base en los resultados de la encuesta. Aun cuando estas consecuencias se conozcan, los resultados de muchas encuestas importantes, los utilizan diferentes personas con diversos propósitos, y algunos de ellos no se han previsto al planear la encuesta. Por lo tanto, es de esperarse que exista un cierto elemento de adivinanza en la especificación de la precisión para un futuro más o menos inmediato.

Si la muestra se toma con un propósito bien definido (como sería una decisión entre "sí" o "no", o bien, sobre la cantidad de dinero que se va a gastar en alguna empresa), entonces, la precisión requerida se puede enunciar usualmente de una manera más específica, en términos de las consecuencias de los errores de decisión. Un enfoque general a los problemas de esta índole se presenta en la Sec. 4.10, que ofrece un punto de partida lógico para la solución.

4.4 LA FORMULA PARA n AL HACER UN MUESTREO PARA DETERMINAR UNA PROPORCION

Las unidades se clasifican en dos clases, C y C'. Se ha convenido en algún margen de error d de la proporción estimada p de las unidades en la clase C, y existe un pequeño riesgo α, que estamos dispuestos a correr, de que el error real supere a d; es decir, queremos que

$$\Pr(|p-P| \geq d) = \alpha$$

Se supone un muestreo aleatorio simple y p se toma con distribución normal. A partir del Teorema 3.2, Sec. 3.2,

$$\sigma_p = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{PQ}{n}}$$

Por lo tanto, la fórmula que relaciona n con el grado de precisión deseado es

$$d = t\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{PQ}{n}}$$

donde t es la abscisa de la curva normal que corta un área de α en las colas de la distribución. Al resolver para n encontramos

$$n = \frac{\dfrac{t^2PQ}{d^2}}{1+\dfrac{1}{N}\left(\dfrac{t^2PQ}{d^2}-1\right)} \tag{4.1}$$

Para fines prácticos, se sustituye una estimación anticipada p de P en esta fórmula. Si N es grande, una primera aproximación es

$$n_0 = \frac{t^2pq}{d^2} = \frac{pq}{V} \tag{4.2}$$

donde

$V = \dfrac{pq}{n_0}$ = varianza deseada de la proporción de muestra

En la práctica, primero calculamos n_0. Si n_0/N es despreciable, una aproximación satisfactoria a n de (4.1) será n_0. De no ser así, es claro, al comparar con 4.1 y 4.2 que n se obtiene como

$$n = \frac{n_0}{1+(n_0-1)/N} \doteq \frac{n_0}{1+(n_0/N)} \tag{4.3}$$

Ejemplo. En el ejemplo hipotético de los grupos sanguíneos teníamos

$$d=0.05, \quad p=0.5, \quad \alpha=0.05, \quad t=2$$

Luego

$$n_0 = \frac{(4)(0.5)(0.5)}{(0.0025)} = 400$$

Supongamos que sólo hay 3 200 isleños. La cpf será necesaria y encontramos

$$n = \frac{n_0}{1+(n_0-1)/N} = \frac{400}{1+\frac{399}{3200}} = 356$$

La fórmula para n_0 también se aplica si d, p y q se expresan como porcentajes en lugar de proporciones. Dado que el producto pq crece al tender p a ½ o 50%, una estimación conservadora de n es la que se obtiene al seleccionar como p el valor más cercano a ½ en el intervalo en que se espera esté p. Si por ejemplo, p parece estar entre 5 y 9%, tomamos 9% para la estimación de n.

Algunas veces, particularmente al estimar el número total NP de unidades en la clase C, deseamos controlar en NP el error *relativo* r en lugar del error absoluto, así, por ejemplo, podríamos desear estimar NP con un error no mayor al 10%, o sea, queremos que

$$\Pr\left(\frac{|Np-NP|}{NP} \geq r\right) = \Pr\left(|p-P| \geq rP\right) = \alpha$$

Para esta especificación, sustituimos rP por rp en vez de d en las Fórmulas (4.1) y (4.2). A partir de (4.2) obtenemos

$$n_0 = \frac{t^2pq}{r^2p^2} = \frac{t^2}{r^2}\frac{q}{p} \tag{4.2}'$$

La Fórmula (4.3) permanece inalterada.

4.5 ATRIBUTOS RAROS—MUESTREO INVERSO

Al estimar n a partir de las Fórmulas (4.1), (4.2) y (4.2)′, el muestreador introduce su mejor estimación anticipada de la proporción de población P. Si se sabe que P está entre el 30 y el 70%, como en el ejemplo de la Sec. 4.1, una estimación exacta de P no es crucial. Pero con un atributo raro (por ejemplo, $P \leq 10\%$), el valor

necesario de n, para un error relativo r específico es 11 veces más grande cuando $P = 1\%$ que cuando $P = 10\%$. En esta situación (un valor pequeño de P pero no muy bien conocido de antemano), el método de Haldane (1945) del muestreo continuo hasta que m de los atributos raros se hayan encontrado en la muestra, tiene una gran ventaja. El método suele llamarse *muestreo inverso*.

Si n es el tamaño de la muestra en el que aparece el m-ésimo atributo raro, $(m > 1)$, una estimación insesgada de P es $p = (m-1)/(n-1)$. Para N muy grande, P pequeña y $m \geq 10$, una buena aproximación de $V(p)$ es $mP^2Q/(m-1)^2$. Luego, $cv(p) \doteq (mQ)^{1/2}/(m-1) < \sqrt{m}/(m-1)$, que será un límite superior bastante ceñido si P es pequeño. De modo que al fijar m de antemano, podemos controlar el valor de $\text{cv}(p)$ sin conocer anticipadamente P. El valor $m = 27$ da $\text{cv}(p) < 20\%$, pero el valor $m = 102$ se necesita para $\text{cv}(p) < 10\%$. El valor de n con este método es una variable aleatoria, pero será grande si P es pequeña.

4.6 LA FORMULA PARA n CON DATOS CONTINUOS

Generalmente se desea controlar el error *relativo* r en la estimación del total o la media de la población. Con una muestra aleatoria simple de media $\bar{y}$, queremos que

$$\Pr\left(\left|\frac{\bar{y}-\bar{Y}}{\bar{Y}}\right| \geq r\right) = \Pr\left(\left|\frac{N\bar{y}-N\bar{Y}}{N\bar{Y}}\right| \geq r\right) = \Pr\left(|\bar{y}-\bar{Y}| \geq r\bar{Y}\right) = \alpha$$

donde α es una pequeña probabilidad. Suponemos que $\bar{y}$ sigue una distribución normal: del Teorema 2.2. Corolario 1, sabemos que el error estándar es

$$\sigma_{\bar{y}} = \sqrt{\frac{N-n}{N}}\frac{S}{\sqrt{n}}$$

Así que

$$r\bar{Y} = t\sigma_{\bar{y}} = t\sqrt{\frac{N-n}{N}}\frac{S}{\sqrt{n}} \tag{4.4}$$

Al resolver para n

$$n = \left(\frac{tS}{r\bar{Y}}\right)^2 \Big/ \left[1+\frac{1}{N}\left(\frac{tS}{r\bar{Y}}\right)^2\right]$$

Obsérvese que la característica de población de la que depende n es su coeficiente de variación $S/\bar{Y}$. Esto a menudo es más estable y fácil de conjeturar de antemano que el valor mismo de S.

Como primera aproximación tomamos

$$n_0 = \left(\frac{tS}{r\bar{Y}}\right)^2 = \frac{1}{C}\left(\frac{S}{\bar{Y}}\right)^2 \tag{4.5}$$

al sustituir una estimación anticipada de $(S/\bar{Y})$. La cantidad C es el $(\text{cv})^2$ deseado de la estimación muestral.

Si n_0/N es apreciable, calculamos n como en (4.3)

$$n = \frac{n_0}{1+(n_0/N)} \tag{4.3}$$

Si en lugar del error relativo r queremos controlar el error absoluto d en $\bar{y}$, tomamos $n_0 = t^2S^2/d^2 = S^2/V$, donde V es la varianza deseada de $\bar{Y}$.

Ejemplo. En viveros que producen arbolitos para venta es aconsejable estimar, al final del invierno o al principio de la primavera, el número de arbolitos en buen estado que probablemente se dispondrán puesto que con base en esto se determina la demanda y aceptación de las órdenes. Johnson (1943) llevó a cabo un estudio de los métodos de muestreo para la estimación del número total de plántulas. La información que se indica a continuación se obtuvo de una almáciga de arces plateados de un pie de ancho y 430 pies de largo. La unidad de muestreo fue un pie de longitud de la almáciga, de tal manera que $N = 430$. Mediante una enumeración completa de la almáciga se encontró que $\bar{Y} = 19$, $S^2 = 85.6$, siendo éstos los verdaderos valores de la población.

Con un muestreo aleatorio simple, ¿cuántas unidades se deben tomar para estimar $\bar{Y}$ dentro del 10% con probabilidades de 1 en 20? De la Ec. (4.5) obtenemos

$$n_0 = \frac{t^2S^2}{r^2\bar{Y}^2} = \frac{(4)(85.6)}{(1.9)^2} = 95$$

Puesto que n_0/N no es despreciable, tomamos

$$n = \frac{95}{1+\frac{95}{430}} = 78$$

Casi el 20% de la almáciga se debe contar para alcanzar la precisión deseada.

Las fórmulas para n dadas aquí se aplican sólo al muestreo aleatorio simple en el que se usa la media muestral como estimación de $\bar{Y}$. Las fórmulas apropiadas para otros métodos de muestreo y estimación se presentan al discutir estas técnicas.

4.7 ESTIMACIONES ANTICIPADAS DE VARIANZAS DE POBLACION

El ejemplo del vivero no es típico en cuanto a que la varianza de población S^2 se conocía. En la práctica, hay cuatro caminos para

estimar las varianzas de población en determinaciones de tamaños de muestra: (1) Al tomar la muestra en dos pasos, siendo el primero una muestra aleatoria simple de tamaño n_1 de donde se estima s_1^2 o p_1 de S^2 o P y se obtiene el valor requerido de n; (2) por los resultados de una encuesta piloto; (3) por muestreo previo de la misma población o una semejante; y (4) al conjeturar respecto a la estructura de la población y ayudarse con algunos resultados matemáticos.

El método 1 proporciona las estimaciones más confiables de S^2 o P, pero no se usa frecuentemente porque retarda la consumación de la encuesta. Cuando el método es factible, Cox (1952) con base en un trabajo de Stein (1945) muestra cómo calcular n a partir de A_1^2 o p_1 de modo que la estimación final de $\bar{y}$ o p tenga una varianza preasignada V, un límite de error preasignado d, o un coeficiente de variación preasignado. La primera muestra se supone lo suficientemente grande para despreciar los términos de orden $1/n_1^2$. A continuación se mencionan algunos resultados.

Los resultados dados aquí suponen $n_1 \leq n$, el tamaño de la muestra final. De no ser así, véase Cox (1952).

Estimación de $\bar{Y}$ dado cv $= \sqrt{C}$

Los resultados suponen una distribución normal de y_i. Si s_1^2 es la varianza estimada a partir de la primera muestra, se toman unidades adicionales para que el tamaño de la muestra final sea

$$n = \frac{s_1^2}{C\bar{y}_1^2}\left(1+8C+\frac{s_1^2}{n_1\bar{y}_1^2}+\frac{2}{n_1}\right) \tag{4.6}$$

La media $\bar{y}$ de la muestra final está ligeramente sesgada. Tómese $\hat{\bar{Y}} = \bar{y}(1-2C)$.

Estimación de $\bar{Y}$ con Varianza V

Se toman unidades adicionales hasta que el tamaño de la muestra total sea

$$n = \frac{s_1^2}{V}\left(1+\frac{2}{n_1}\right) \tag{4.7}$$

Si S se conociera exactamente, el tamaño requerido de la muestra sería S^2/V. El efecto de no conocer S es incrementar el tamaño promedio por un factor $(1 + 2/n_1)$.

Estimación de P con Varianza V

Sea p_1 la estimación de P a partir de la primera muestra. El tamaño combinado de las dos primeras muestras debería ser

$$n = \frac{p_1 q_1}{V} + \frac{3 - 8 p_1 q_1}{p_1 q_1} + \frac{1 - 3 p_1 q_1}{V n_1} \tag{4.8}$$

El primer término de la derecha es el tamaño requerido si se sabe que P es igual a p_1. Con este método, la estimación binomial ordinaria p hecha a partir de la muestra completa, de tamaño n, es ligeramente sesgada. Para corregir el sesgo, tómese

$$\hat{P} = p + \frac{V(1-2p)}{pq}$$

Estimación de P con cv $= \sqrt{C}$ dado

Tómese

$$n = \frac{q_1}{C p_1} + \frac{3}{p_1 q_1} + \frac{1}{C p_1 n_1} \tag{4.9}$$

La estimación es $\hat{P} = p - Cp/q$. En todos los resultados anteriores se ignora la cpf.

Ejemplo. Un muestreador desea estimar P con un coeficiente de variación de 0.1 (10%) y conjetura que P se encontrará entre 5 y 20%. Este intervalo es demasiado amplio para dar una buena estimación inicial del n requerido. Como el cv de P es $\sqrt{Q/nP}$, fácilmente se verifica que $n = 400$ es adecuado para $P = 20\%$, pero $n = 1900$ se requerirá si P es sólo 5%.

De acuerdo con esto, el muestreador toma una muestra inicial con $n_1 = 396$ y encuentra $p_1 = 0.101$. Dado que $\sqrt{C} = 0.1$, $C = 0.01$. La Ec. 4.9 conduce a

$$n = \frac{(0.899)}{(0.01)(0.101)} + \frac{3}{(0.0908)} + \frac{1}{(0.01)(40)} = 926$$

La muestra combinada da $np = 88$; $p = 88/926 = 0.0950$. La corrección por sesgo, Cp/q, viene a ser 0.0011, lo que da una estimación final de 0.094 o 9.4%.

En el segundo método, una pequeña encuesta piloto sirve para varios fines, especialmente si se duda de la posibilidad de la encuesta principal. Si la encuesta piloto es en sí una encuesta aleatoria simple, se aplican los métodos precedentes. Pero con frecuencia, el trabajo piloto se limita a una parte de la población que se puede manejar convenientemente, o que revelará la magnitud de ciertos problemas. Se debe tener en cuenta la opción selectiva de la encuesta piloto al usar sus resultados para estimar S^2 o P. Por ejemplo, es una práctica común la de confinar el trabajo piloto a algunos

conglomerados de unidades. Así, el valor s^2 calculado, mide principalmente la variación dentro de un conglomerado y puede ser una subestimación del valor pertinente S^2. La relación entre las variaciones intra e interconglomerados se discutirá en el Cap. 9. El mismo problema surge en el muestreo por conglomerados para proporciones, en donde la fórmula pq/n puede subestimar el efecto de la variación entre conglomerados. Cornfield (1951) ilustra adecuadamente la estimación del tamaño de la muestra en muestreo por conglomerados para proporciones.

El tercer método —uso de los resultados de encuestas previas —, señala el valor que tiene el poner a disposición, o al menos mantener accesible, cualquier dato sobre la desviaciones estándar obtenidas en encuestas previas. Infortunadamente, el costo de computación de las desviaciones estándar en encuestas complejas es elevado, aun con máquinas electrónicas, y es frecuente que sólo aquellas desviaciones estándar requeridas para dar una vaga idea de la precisión de las estimaciones principales, se computen y registren. Si se encuentran datos anteriores apropiados, el valor de S^2 puede requerir un ajuste, para actualizarlo. Con datos asimétricos en los que $\bar{Y}$ cambia con el tiempo, frecuentemente se encuentra que S^2 varía de acuerdo a una razón comprendida entre $k\bar{Y}$ y $k\bar{Y}^2$ donde k es una constante. Por lo tanto, si se cree que $\bar{Y}$ ha aumentado en 10% en el intervalo de tiempo transcurrido desde la encuesta previa, podríamos incrementar nuestra estimación inicial de S^2 entre un 10 y un 20%.

Por último, en ocasiones es posible hacer una estimación útil de S^2 a partir de una información relativamente escasa respecto a la naturaleza de la población. En estudios previos sobre el número de gusanos en suelos, se utilizó un instrumento para tomar una muestra de ($9 \times 9 \times 5$ plg) de la capa superficial. Para estimar n, el muestreador necesitaba conocer la desviación estándar del número de gusanos encontrados en una extracción con ayuda del instrumento. Si los gusanos se distribuyeran al azar en la capa superficial, el número encontrado en un pequeño volumen se apegaría a una distribución de Poisson para la cual $S^2 = \bar{Y}$. Como los gusanos pueden tender a congregarse se decidió suponer $S^2 = 1.2\bar{Y}$, donde el factor 1.2 es un factor arbitrario de seguridad. Aunque se desconocía $\bar{Y}$, sus valores de importancia económica respecto al daño en la cosecha se pudieron delinear. Estas dos informaciones hicieron posible la determinación de los tamaños de muestra que resultaron satisfactorios.

Deming (1960) muestra cómo algunas distribuciones matemáticas simples son útiles en la estimación de S^2, a partir de cierto conocimiento del intervalo donde se encuentre y de una idea general de la forma de la distribución. Si la distribución es como una binomial, con una proporción p de las observaciones en un extremo del intervalo y una proporción q en el otro, $S^2 = pqh^2$ donde h es la longitud del intervalo cubierto por la distribución. Cuando $p = q = \frac{1}{2}$ el valor de $S^2 = 0.25h^2$ es el máximo posible para una longitud h dada. Otras relaciones útiles son $S^2 = 0.083h^2$ para una distribución rectangular, $S^2 = 0.056h^2$ para una distribución que tiene forma de triángulo rectángulo y $S^2 = 0.042h^2$ para un triángulo isósceles.

Estas relaciones no son de mucha ayuda si h es grande o poco conocida. Sin embargo, si h es grande, es conveniente estratificar la población (Cap. 5) de modo que dentro de un estrato, el intervalo se reduzca considerablemente. Es usual que dentro de un estrato, la forma también se simplifique (más cercana a rectangular). En consecuencia, estas relaciones son efectivas en la predicción de S^2, y, por lo tanto de n, dentro de estratos individuales.

4.8 TAMAÑO DE LA MUESTRA CON MAS DE UNA CARACTERISTICA

En la mayoría de las encuestas se recoge información sobre más de una característica. Un método para determinar el tamaño de la muestra es la especificación de los márgenes de error para las características que se consideran vitales en la encuesta. Una estimación inicial del tamaño requerido de la muestra se hace separadamente para cada una de estas características importantes.

Al completar la estimación de n para cada característica, se concreta la situación. Puede ser que los n requeridos estén bastante próximos. Si el valor más grande de n está dentro de los límites del presupuesto, se toma este valor. Pero con frecuencia, hay suficiente variación entre los n y, por lo tanto, no se selecciona el más grande, ya sea por consideraciones presupuestales o porque este valor dará una precisión global sustancialmente más elevada que la considerada en un principio. En este caso, el estándar de precisión deseado puede relajarse un poco para algunas características, lo que permite utilizar un valor más pequeño de n.

Hay casos en que los n requeridos, para diferentes características, son tan discordantes que algunos se deben abandonar en la investigación; con los recursos disponibles, la precisión esperada para estas características es totalmente inadecuada. La dificultad

puede no deberse exclusivamente al tamaño de la muestra. Algunas características requerirán tipos diferentes de muestreo en comparación con otras. Con poblaciones que se muestran repetidamente es útil reunir información respecto a las características que se pueden combinar económicamente en una encuesta general y las que requieren métodos especiales. Como ejemplo, en la Tabla 4.1 presentamos una clasificación de características en 4 tipos, sugerida por la experiencia en encuestas agrícolas regionales. En esta clasificación, una encuesta general quiere decir una en la cual las unidades están más o menos uniformemente distribuidas en alguna región, como sucede, por ejemplo con una muestra aleatoria simple.

TABLA 4.1. UN EJEMPLO DE TIPOS DIFERENTES DE CARACTERÍSTICAS EN ENCUESTAS REGIONALES

Tipo	Descripción de las Características	Tipo de Muestreo Requerido
1	Bien distribuido a través de la región y ocurriendo con frecuencia razonable en todas partes.	Una encuesta general con baja razón de muestreo.
2	Bien distribuida a través de la región pero con baja frecuencia.	Una encuesta general, pero con una razón de muestreo más elevada.
3	Ocurre con una frecuencia razonable en la mayoría de las partes de la región, pero con una distribución más esporádica, estando ausente en algunas partes y altamente concentrada en otras.	Para obtener mejores resultados se hace una muestra estratificada con diferentes intensidades. en partes diferentes de la región (Cap. 5). Puede a veces incluirse en una encuesta general con muestreo suplementario.
4	Distribución muy esporádica o concentrada en una pequeña parte de la región.	No adecuada para una encuesta general. Requiere una muestra especialmente construida para su distribución.

4.9 TAMAÑO DE LA MUESTRA CUANDO LAS ESTIMACIONES SE QUIEREN PARA SUBDIVISIONES DE LA POBLACION

Se planea a menudo para presentar estimaciones no sólo de la población como un todo, sino también para ciertas subdivisiones. Si éstas pueden identificarse por adelantado, como es el caso de las regiones geográficas se hace un cálculo separado de n para cada región. Supóngase que la media de cada subdivisión se debe esti-

mar con una varianza específica V. Para la i-ésima subdivisión, tenemos $n_i = S_i^2/V$, por lo tanto, el tamaño de la muestra total $n = \sum S_i^2/V$. Las S_i^2 individuales serán en promedio más pequeñas que S^2, la varianza de la población, pero a menudo sólo un poco más pequeñas. De modo que si existen k subdivisiones, $n \doteq kS^2/V$, en tanto que si sólo se desea obtener la estimación para la población como un todo, se toma $n = S^2/V$.

Por lo tanto, si queremos estimaciones con varianza V para cada una de las k subdivisiones, el tamaño de la muestra puede aproximarse a k veces el n necesario para una estimación global de la misma precisión. Este aspecto tiende a ignorarse en los cálculos del tamaño de la muestra, y le ocurre especialmente a las personas sin experiencia en métodos de encuesta.

Si las subdivisiones representan clasificaciones por variables tales como la edad, el sexo, el ingreso, los años de educación, etc., la subdivisión a la que una persona pertenece no se conoce sino hasta después de haber tomado la muestra. Incluso se pueden hacer estimaciones anticipadas del tamaño de la muestra, si se conocen las proporciones π_i de las unidades que pertenecen a las diferentes subdivisiones. Si se selecciona una muestra aleatoria simple de tamaño n, el tamaño esperado de la muestra a partir de la i-ésima subdivisión es $n\pi_i$. La varianza promedio de la media a partir de esta subdivisión es

$$V(\bar{y}_i) = E\left(\frac{S_i^2}{n_i}\right) \doteq \frac{S_i^2}{n\pi_i} \tag{4.10}$$

si $n\pi_i$ es grande. Por lo tanto, requerimos $n \doteq S_i^2/\pi_i V$ para hacer $V(\bar{y}_i) = V$. Si esto es válido para cada subdivisión,

$$n \doteq \text{máx}\left(\frac{S_i^2}{\pi_i V}\right) \tag{4.11}$$

Si las subdivisiones están hechas en clases como la edad o el ingreso, S_i^2/π_i puede ser menor que S^2 para las clases centrales, pero puede ser grande para una clase extrema con π_i pequeña. En este caso, tendremos que incrementar el valor de V en esta subdivisión, o bien, encontrar alguna manera de identificar de antemano unidades en esta subdivisión para que se puedan muestrear a una mayor tasa. En ocasiones, el método del muestreo doble (Cap. 12) es útil para este propósito.

Las exigencias en el tamaño de la muestra son aún mayores en estudios analíticos, donde las especificaciones son

$$V(\bar{y}_i - \bar{y}_j) \leq V \tag{4.12}$$

para cada par de subdivisiones (dominios). En este caso

$$n \doteq \text{máx}_{i,j} \frac{1}{V}\left(\frac{S_i^2}{\pi_i} + \frac{S_j^2}{\pi_j}\right) \tag{4.13}$$

Si las S_i^2 no son muy diferentes de S^2, n será $2kS^2/V$ cuando los k dominios son del mismo tamaño, y aún mayores en el otro caso. El efecto de los términos de cpf, despreciados en esta discusión, es la reducción en cierta medida de los n requeridos.

4.10 EL TAMAÑO DE LA MUESTRA EN PROBLEMAS DE DECISION

En ocasiones, puede desarrollarse un enfoque más lógico para determinar el tamaño de la muestra, cuando se va a tomar una decisión práctica, basada en los resultados de la muestra. Se puede presuponer que la decisión estará más sólidamente fundamentada, si la estimación muestral tiene un error, pequeño, en lugar de uno elevado. Podemos calcular en términos monetarios la pérdida $l(z)$ que ocurrirá al tomar una decisión que presente un error correspondiente de z en la estimación. Aunque el valor real de z no se puede predecir de antemano, la teoría del muestreo nos permite encontrar la distribución de frecuencia $f(z, n)$ de z, que para un método de muestreo específico, dependerá del tamaño de la muestra n. Por lo tanto, la *pérdida esperada* para un tamaño de muestra dado es

$$L(n) = \int l(z) f(z, n)\, dz \tag{4.14}$$

El propósito de tomar la muestra es disminuir esta pérdida. Si $C(n)$ es el costo de una muestra de tamaño n es razonable elegir n para minimizar.

$$C(n) + L(n) \tag{4.15}$$

puesto que éste es el costo total involucrado al tomar la muestra y las decisiones a partir de sus resultados. La elección de n determina tanto el tamaño óptimo de la muestra, como el grado de precisión más ventajoso.

Alternativamente puede presentarse el mismo enfoque en términos de la *ganancia* monetaria que ocurre al tener la información muestral en lugar de la pérdida derivada de los errores en dicha información. Si se usa la ganancia monetaria, construimos una ganancia esperada $G(n)$ a partir de un tamaño de muestra n, donde $G(n)$ es cero si no se toma muestra alguna. *Maximizamos*

$$G(n) - C(n)$$

En esta forma, el principio es equivalente a la regla de la economía clásica de la maximización de la utilidad.

La aplicación más simple se presenta cuando la función pérdida $l(z)$, es λz^2, donde λ es una constante. Lo que da como resultado que

$$L(n) = \lambda E(z^2) \tag{4.16}$$

Por ejemplo, si $\hat{\bar{Y}}$ es la estimación muestral de $\bar{Y}$, y $z = \hat{\bar{Y}} - \bar{Y}$,

$$L(n) = \lambda V(\hat{\bar{Y}}) = \frac{\lambda S^2}{n} - \frac{\lambda S^2}{N} \tag{4.17}$$

si se usa muestreo aleatorio simple.

El tipo más simple de función costo para la muestra es

$$C(n) = c_0 + c_1 n \tag{4.18}$$

donde c_o es el costo fijo. Al derivar, el valor de n que minimiza costo más pérdida es

$$n = \sqrt{\lambda S^2 / c_1} \tag{4.19}$$

Yates (1960) da una forma más general de este resultado. El mismo análisis se aplica a cualquier método de muestreo y de estimación en el que la varianza de la estimación es inversamente proporcional a n y el costo es una función lineal de n.

Blythe (1945) describe la aplicación de este principio a la estimación del volumen de madera en un lote con fines de venta (ver Ejercicio 4.11). Nordin (1944) discute el tamaño óptimo de la muestra para estimar las ventas potenciales de un mercado al que desea entrar un fabricante. Si las ventas pueden pronosticarse exactamente, la cantidad de equipo fijo y de producción por unidad de tiempo se podrá asignar de tal manera que maximice la utilidad esperada por el fabricante. Grundy y otros (1954-1956) consideran el tamaño óptimo de una segunda muestra, cuando se conocen los resultados de una muestra inicial.

Este enfoque ha sido desarrollado sustancialmente en trabajos sobre la teoría de decisión estadística. Las generalizaciones incluyen la sustitución de la utilidad por el valor monetario como una escala en la que se miden costos o pérdidas, el uso explícito de información anterior subjetiva respecto a parámetros desconocidos, al expresar esta información como distribuciones de probabilidad a *priori* de los parámetros desconocidos, y la investigación de diferentes tipos de funciones de costo y pérdida, y de datos tanto cualitativos

como cuantitativos. Se puede ver una relación bastante completa del método en el trabajo de Raiffa y Schlaifer (1961). Aunque todavía no es evidente la frecuencia con que podrán llevarse a una solución completa los problemas de decisión con ayuda de este método, puede afirmarse que es valioso ya que estimula las reflexiones respecto a los factores importantes de una buena decisión. Una área que parece adaptarse convenientemente a las aplicaciones es la del muestreo de los lotes de artículos en un proceso de producción en masa, para decidir si se acepta o se rechaza el lote con base en su calidad estimada. Sittig (1951) considera la economía de la determinación del tamaño de muestra, al tomar en cuenta costos de inspección y de aceptación de artículos defectuosos, así como costos ligados a la existencia de artículos reglamentarios en lotes rechazados.

4.11 EL EFECTO DEL DISEÑO (*Deff*)

Con los planes de muestreo más complejos que se describen posteriormente, una cantidad útil es el llamado *efecto del diseño* (*deff*) del plan de muestreo (Kish, 1965). Este autor describe dicha cantidad como la razón de la varianza de la estimación obtenida a partir de la muestra más compleja a la varianza de la estimación obtenida a partir de una muestra aleatoria simple del mismo número de unidades. El efecto del diseño tiene dos usos primordiales —en la estimación del tamaño de muestra y en la apreciación de la eficiencia de planes más complejos—. Por ejemplo, al estimar la proporción de personas que presentan cierto atributo, a menudo es conveniente usar la casa en vez de la persona, como unidad de muestreo. Como se notó en el Cap. 3, la fórmula PQ/n no se puede usar en estos planes. Para estimar la proporción de los que han visto a un médico (Sec. 3.12) una muestra aleatoria simple de casas, dio $v(p) = 0.00520$ contra $pq/n = 0.00197$ para una muestra aleatoria simple del mismo tamaño constituida por personas. Una estimación del *deff* para esta muestra por conglomerados y esta variable es $520/197 = 2.6$. Cuando las fracciones de muestreo son pequeñas, podemos, por lo tanto, estimar el tamaño de la muestra al calcular el n (número de personas) requerido con una muestra aleatoria simple de personas y multiplicar por 2.6. Al observar las razones *deff* de esta manera para las variantes importantes con un plan complejo, podemos usar las fórmulas simples de este capítulo para estimar el

tamaño de la muestra con el plan complejo y también juzgar si dicho plan es ventajoso en eficiencia, en relación a su costo y complejidad.

Para estimar el *deff* a partir de los resultados de una muestra compleja, se requiere algo de álgebra. Necesitamos mostrar en qué forma estos resultados proporcionan, de ser posible, estimaciones insesgadas de S^2 o de PQ. Se dan ejemplos de estos cálculos para muestreo aleatorio estratificado en la Sec. 5A.11 y para muestreo por conglomerados en la Sec. 9.3, cuando los conglomerados son del mismo tamaño.

EJERCICIOS

4.1. En un distrito en donde hay 4 000 casas, el porcentaje de propietarios va a ser estimado con una muestra no mayor al 2% y el porcentaje de propietarios de 2 autos con una de no mayor al 1%. (Las cifras 2 y 1% son los valores absolutos, y no los cv.) El porcentaje verdadero de propietarios se piensa que está entre 45 y 65% y el porcentaje de propietarios de dos autos entre 5 y 10%. ¿Qué tan grande debe ser una muestra para satisfacer las dos finalidades?

4.2. En la población de 676 hojas de solicitud (Tabla 2.2 Pág. 53) ¿qué tan grande debe ser la muestra si se va a estimar el número total de firmas con un margen de error de 1 000 con probabilidad de 95%? Suponga que el valor de s^2 dado en la Pág. 53 es el valor de S^2 en la población.

4.3. Se va a realizar una encuesta para determinar la existencia de enfermedades comunes en una población grande. Para cualquier enfermedad que afecte cuando menos al 1% de los individuos en la población, se desea estimar el número total de casos, con un coeficiente de variación de no más del 20%. (*a*) ¿Qué tamaño de muestra aleatoria simple es necesario al suponer que la presencia de la enfermedad puede ser reconocida sin equivocación? (*b*) ¿Qué tamaño de muestra se necesita si se desea obtener el número total de casos separadamente para hombres y mujeres, con la misma precisión?

4.4. En un muestreo de insectos se va a estimar el número de gusanos por acre con un límite de error del 30% al nivel de probabilidad del 95%, en cualquier campo en donde la cantidad de gusanos exceda a 200 000 por acre en la capa superior de 5 plg del suelo. La barrena para muestrear mide $9 \times 9 \times 5$ plg de alto. Suponiendo que el número de gusanos en una muestra sigue una distribución ligeramente más variable que la de Poisson, tomamos $S^2 = 1.2\bar{Y}$. ¿Qué tamaño de muestra aleatoria simple es necesaria? (Un acre = 43 560 pies^2.)

4.5. Los siguientes coeficientes de variación se obtuvieron en una encuesta de predios agrícolas en Iowa, la unidad fue una área de una milla cuadrada (información de R. J. Jessen):

Característica	Coeficiente de variación estimado (%)
Acres en predios agrícolas	38
Acres de maíz	39
Acres de avena	44
Número de trabajadores de la familia	100
Número de trabajadores contratados	110
Número de desempleados	317

Se planea una encuesta para estimar características de área con un coeficiente de variación de 2.5% y número de trabajadores (excluyendo a los desempleados) con un coeficiente de variación del 5%. Con una muestra aleatoria simple, ¿cuántas unidades se necesitan? ¿Qué tan bien se supone que esta muestra estime el número de desempleados?

4.6. Por muestreo experimental, el valor medio de una variable aleatoria se va a estimar con una varianza $V = 0.0005$. Los valores de la variable aleatoria para las primeras 20 muestras se proporcionan a continuación. ¿Cuántas muestras adicionales se necesitan?

No. de muestra	Valor de la variable aleatoria	No. de muestra	Valor de la variable aleatoria
1	0.0725	11	0.0712
2	0.0755	12	0.0748
3	0.0759	13	0.0878
4	0.0739	14	0.0710
5	0.0732	15	0.0754
6	0.0843	16	0.0712
7	0.0727	17	0.0757
8	0.0769	18	0.0737
9	0.0730	19	0.0704
10	0.0727	20	0.0723

4.7. Se diseña una encuesta familiar para estimar la proporción de familias que poseen ciertos atributos. Para las principales características de interés, el valor de P se espera que oscile entre el 30 y el 70%. Con una muestra aleatoria simple, ¿qué tan grande debe ser el valor de n para estimar los siguientes promedios con un error estándar que no exceda de 3%? (*a*) la media P global. (*b*) las medias *individuales* P_i para las clases de ingresos —menores de \$5 000; \$5 000 a \$10 000; mayores a \$10 000. ($i = 1, 2, 3$)—. (*c*) las diferencias entre medias ($P_i - P_j$) para cada una de las parejas de clases en (*b*). Dar una respuesta separada para (*a*), (*b*) y (*c*). Las estadísticas de ingresos indican que las proporciones de familias con ingresos en las tres clases anteriores son 50, 38 y 12%.

4.8. Los colegios con programas de estudio de cuatro años, en los Estados Unidos se dividieron en clases de 4 tamaños diferentes de acuerdo con sus inscripciones de 1952-1953. Las desviaciones estándar dentro de cada clase se muestran a continuación.

	Clase			
	1	2	3	4
Número de estudiantes	<1000	1000–3000	3000–10 000	más de 10 000
S_i	236	625	2008	10 023

Si conoce los límites de clase pero no los valores de S_i, ¿qué tan bien puede estimar los valores de los S_i usando cifras matemáticas simples (Sec. 4.7)? Ningún colegio tiene menos de 200 estudiantes y el más grande tiene cerca de 50 000 estudiantes.

4.9. Con una función cuadrática de pérdida y una función lineal de costos como en la Sec. 4.10, S^2 se reduce a S'^2 mediante un plan de muestreo más eficiente, c_0, c_1 y λ permanecen sin cambio. Si n', V' denotan un nuevo tamaño óptimo de muestra y $V(\hat{\bar{Y}})$, demostrar que $n' < n$ y que $V' < V$.

4.10. Si la función de pérdida debida a un error en $\bar{y}$ es $\lambda\,|\bar{y} - \bar{Y}|$ y si el costo $C = c_0 + c_1 n$, demuestre que con un muestreo aleatorio simple, ignorando la cpf, el valor más económico de n es

$$\left(\frac{\lambda S}{c_1\sqrt{2\pi}}\right)^{2/3}$$

4.11. (Adaptado de Blythe, 1945.) El precio de venta de un lote de madera en pie es UW, en donde U es el precio por unidad de volumen y W es el volumen de la madera en el lote. Se cuenta el número N de troncos en el lote, y mediante una muestra aleatoria simple de n troncos, se estima el volumen promedio por tronco. Se hace una estimación pagada por el vendedor y ésta es provisionalmente aceptada por el comprador. Después, el comprador determina el volumen exacto comprado, y el vendedor le reembolsa, en caso de que el comprador haya pagado más de lo que le fue entregado. Si el comprador ha pagado menos de lo recibido, no menciona el hecho.

Construya la función de pérdida del vendedor, Al suponer que el costo de la medición de n troncos es cn, encuentre el valor óptimo de n. La desviación estándar del volumen por tronco se puede indicar por S y la cpf se puede ignorar.

4.12. (a) La presencia o la ausencia de dos características se debe medir en cada unidad de una muestra aleatoria simple de una población grande. Si P_1, P_2 son los porcentajes de unidades de la población que tienen las características 1 y 2, un cliente desea estimar $(P_1 - P_2)$ con un error estándar que no exceda dos puntos por ciento. ¿Cuál es el tamaño de la muestra, que debe sugerirse si el cliente piensa que tanto P_1 como P_2 están entre el 40 y el 60%, y que las características se distribuyen independientemente en las unidades?

(b) Suponga que en (a) el cliente opina que las características están positivamente correlacionadas, pero no conoce la correlación, si se sugiere una muestra inicial de 200, con los siguientes resultados.

Características		
1	2	Número de unidades
Sí	Sí	72
Sí	No	44
No	Sí	14
No	No	70
		200

¿Qué tamaño de muestra se recomienda para estimar $(P_1 - P_2)$ con un error estándar $\leq 2\%$?

4.13. (a) Suponga que se está estimando la razón de sexos, que es cercana a la unidad, y que se pueden muestrear domicilios donde viven 4 personas, padre, madre y dos hijos. Al ignorar la pequeña proporción de familias con gemelos idénticos, encontrar el factor *deff* para una muestra aleatoria simple de n casas contra una muestra de las $4n$ personas. (b) Las familias donde hubiese gemelos idénticos, ¿elevarían o reducirían el factor *deff*?

CAPITULO 5

Muestreo Aleatorio Estratificado

5.1 DESCRIPCION

En el muestreo estratificado, la población de N unidades se divide primero en subpoblaciones de $N_1, N_2, \ldots, N_L$ unidades, respectivamente. Estas subpoblaciones, no se traslapan y en su conjunto comprenden a toda la población, por lo tanto,

$$N_1+N_2+\cdots+N_L=N$$

Las subpoblaciones se denominan *estratos*. Para obtener todo el beneficio de la estratificación, los valores de los N_h deben ser conocidos. Una vez determinados los estratos, se extrae una muestra de cada uno, las extracciones deben hacerse independientemente en los diferentes estratos. Los tamaños de muestras dentro de los estratos se denotan con $n_1, n_2, \ldots, n_L$, respectivamente.

Si se toma una muestra aleatoria simple en cada estrato, el procedimiento total se describe como un *muestreo aleatorio estratificado*.

La estratificación es una técnica común. Para esto hay muchas razones; las principales son las siguientes.

1. Si los datos deseados deben tener una precisión conocida en algunas subdivisiones de la población, es aconsejable tratar cada subdivisión como una "población" por derecho propio.

2. Por conveniencia administrativa, puede ser necesario el uso de la estratificación; así, por ejemplo, la agencia que realiza una encuesta, podría tener sucursales en el campo, cada una de las cuales supervisaría la encuesta de una parte de la población.

3. Los problemas de muestreo pueden tener marcadas diferencias en diversas partes de la población. Con poblaciones humanas, las personas que viven en instituciones (como hoteles, hospitales,

cárceles) se colocan en un estrato diferente de las que viven en casas ordinarias, ya que otro método de muestreo es el apropiado para cada una de estas dos situaciones. Al muestrear negocios, podríamos tener una lista de las grandes firmas que se deben colocar en un estrato diferente. Quizá algún tipo de muestreo por áreas debe usarse para las firmas más pequeñas.

4. La estratificación puede dar lugar a una ganancia en la precisión de las estimaciones de características de la población total. Quizá sea posible dividir una población heterogénea en subpoblaciones, en las que cada una sea internamente homogénea. Esto es lo que sugiere el nombre de *estratos*, con su implicación de una división en capas. Si cada estrato es homogéneo, en cuanto a que las medidas varíen ligeramente de una unidad a otra, una estimación precisa de cualquier media de estrato se puede obtener a partir de una pequeña muestra en dicho estrato. Y posteriormente podrán combinarse estas estimaciones en una estimación precisa para toda la población.

La teoría del muestreo estratificado se ocupa de las propiedades de las estimaciones de una muestra estratificada y de la mejor elección para los tamaños de muestras n_h que deben dar la precisión máxima. En este desarrollo se presupone que ya se construyeron los estratos. Los problemas de construcción de los estratos y de la determinación del número de estratos se verán posteriormente (Sec. 5A.7).

5.2 NOTACION

El subíndice h denota el estrato, e i la unidad dentro del estrato. La notación es una extensión natural de la usada anteriormente. Todos los símbolos siguientes se refieren al estrato h.

N_h	número total de unidades
n_h	número de unidades en la muestra
y_{hi}	valor obtenido para la i-ésima unidad
$W_h = \dfrac{N_h}{N}$	ponderación del estrato
$f_h = \dfrac{n_h}{N_h}$	fracción de muestreo en el estrato

$$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h} \qquad \text{media verdadera}$$

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} \qquad \text{media de muestra}$$

$$S_h^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2}{N_h - 1} \qquad \text{varianza verdadera}$$

Nótese que el divisor en la varianza es $(N_h - 1)$.

5.3 PROPIEDADES DE LAS ESTIMACIONES

Para la media de población por unidad, la estimación usada en muestreo estratificado es $\bar{y}_{st}$ (st significa *estratificado*), donde

$$\bar{y}_{st} = \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{N} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{5.1}$$

donde $N = N_1 + N_2 + \cdots + N_L$.

La estimación $\bar{y}_{st}$ en general no es la misma que la media muestral. La media muestral $\bar{y}$ puede escribirse como

$$\bar{y} = \frac{\sum_{h=1}^{L} n_h \bar{y}_h}{n} \tag{5.2}$$

La diferencia es que en $\bar{y}_{st}$ las estimaciones a partir de estratos individuales reciben sus ponderaciones correctas N_h/N. Es evidente que $\bar{y}$ coincide con $\bar{y}_{st}$ cuando en cada estrato

$$\frac{n_h}{n} = \frac{N_h}{N} \quad \text{o} \quad \frac{n_h}{N_h} = \frac{n}{N} \quad \text{o} \quad f_h = f$$

Esto significa que la fracción de muestreo es la misma en todos los estratos. Esta estratificación se describe como estratificación con asignación *proporcional* de los números n_h y da lugar a una muestra autoponderada. Si se hacen numerosas estimaciones, este tipo de muestra ahorrará tiempo.

Las propiedades principales de la estimación $\bar{y}_{st}$ se bosquejan en los siguientes teoremas. Los dos primeros, se aplican al mues-

treo estratificado en general y no se restringen al muestreo aleatorio estratificado; es decir, que la muestra en un estrato cualquiera no es necesariamente una muestra aleatoria simple.

Teorema 5.1. Si en cada estrato la estimación muestral $\bar{y}_h$ es insesgada, entonces y_{st} es una estimación insesgada de la media de población $\bar{Y}$.

Prueba.

$$E(\bar{y}_{st}) = E\sum_{h=1}^{L} W_h \bar{y}_h = \sum_{h=1}^{L} W_h \bar{Y}_h$$

puesto que las estimaciones son insesgadas en los estratos individuales. Pero la media de población $\bar{Y}$ puede escribirse

$$\bar{Y} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h} y_{hi}}{N} = \frac{\sum_{h=1}^{L} N_h \bar{Y}_h}{N} = \sum_{h=1}^{L} W_h \bar{Y}_h$$

lo que completa la demostración.

Teorema 5.2. Si las muestras se extraen independientemente en los diferentes estratos,

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) \tag{5.3}$$

donde $V(\bar{y}_h)$ es la varianza de $\bar{y}_h$ sobre muestras repetidas del estrato h.

Prueba. Dado que

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{5.4}$$

$\bar{y}_{st}$ es una función lineal de los $\bar{y}_h$ con ponderaciones fijas W_h. Por lo tanto, podemos citar el resultado en estadísticas para la varianza de una función lineal.

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) + 2\sum_{h=1}^{L}\sum_{j>h}^{L} W_h W_j \operatorname{Cov}(\bar{y}_h \bar{y}_j) \tag{5.5}$$

Pero como las muestras se extraen independientemente en los diferentes estratos, se anulan todos los términos de covarianza. Esto da el resultado (5.3).

Para resumir los Teoremas 5.1 y 5.2: si $\bar{y}_h$ es una estimación insesgada de $\bar{Y}_h$ en cada estrato, y la selección de muestras es independiente en los diferentes estratos, entonces $\bar{y}_{st}$ es una estimación insesgada de $\bar{Y}$ con varianza $\sum W_h^2 V(\bar{y}_h)$.

Lo importante con respecto a este resultado es que la varianza de $\bar{y}_{st}$ sólo depende de las varianzas de las estimaciones de las medias $\bar{Y}_h$ de los estratos individuales. Si fuese posible dividir una población altamente variable en estratos de un modo tal que todos los atributos tuviesen el mismo valor dentro de un estrato, podríamos estimar $\bar{Y}$ sin error alguno. La Ec. 5.4 muestra que el uso de las ponderaciones de estratos correctos N_h/N al hacer la estimación y_{st} es el que conduce a este resultado.

Teorema 5.3. Para el muestreo aleatorio estratificado, la varianza de la estimación $\bar{y}_{st}$ es

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1-f_h) \tag{5.6}$$

Prueba. Puesto que $\bar{y}_h$ es una estimación insesgada de Y_h puede aplicarse el Teorema 5.2. Además, por el Teorema 2.2, aplicado a un estrato individual

$$V(\bar{y}_h) = \frac{S_h^2}{n_h}\frac{N_n - n_h}{N_h}$$

Por sustitución en el resultado del Teorema 5.2, obtenemos

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 V(\bar{y}_h) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} = \sum W_h^2 \frac{S_h^2}{n_h}(1-f_h)$$

En los siguientes corolarios se dan algunos casos particulares de esta fórmula.

Corolario 1. Si las fracciones de muestreo n_h/N_h son despreciables en todos los estratos,

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum \frac{N_h^2 S_h^2}{n_h} = \sum \frac{W_h^2 S_h^2}{n_h} \tag{5.7}$$

Esta es la fórmula apropiada cuando las correcciones por población finita se pueden ignorar.

Corolario 2. Con asignación proporcional, sustituimos

$$n_h = \frac{nN_h}{N}$$

en (5.6). La varianza se reduce a

$$V(\bar{y}_{st}) = \sum \frac{N_h}{N}\frac{S_h^2}{n}\left(\frac{N-n}{N}\right) = \frac{1-f}{n}\sum W_h S_h^2 \tag{5.8}$$

Corolario 3. Si el muestreo es proporcional y las varianzas en todos los estratos tienen el mismo valor, S_w^2, obtenemos el siguiente resultado

$$V(\bar{y}_{st}) = \frac{S_w^2}{n}\left(\frac{N-n}{N}\right) \tag{5.9}$$

Teorema 5.4. Si $\hat{Y}_{st} = N\bar{y}_{st}$ es la estimación del total de la población Y, entonces

$$V(\hat{Y}_{st}) = \sum N_h(N_h - n_h)\frac{S_h^2}{n_h} \tag{5.10}$$

Lo anterior es inmediato, considerando el Teorema 5.3.

Ejemplo. La Tabla 5.1 muestra el número de habitantes en 1920 y 1930, en miles, en 64 grandes ciudades de los Estados Unidos. Los datos fueron obtenidos tomando entre la 5a. y la 68a. ciudades de los Estados Unidos del total ordenado respecto al número total de habitantes en 1920. Las ciudades se colocaron en dos estratos, el primero contiene las 16 ciudades más grandes y el segundo las 48 restantes.

TABLA 5.1. TAMAÑOS DE 64 CIUDADES (EN MILES) EN 1920 Y 1930

Tamaño en 1920 (x_{hi})				Tamaño en 1930 (y_{hi})			
Estrato				Estrato			
$h = 1$	2			1	2		
797	314	172	121	900	364	209	113
773	298	172	120	822	317	183	115
748	296	163	119	781	328	163	123
734	258	162	118	805	302	253	154
588	256	161	118	670	288	232	140
577	243	159	116	1238	291	260	119
507	238	153	116	573	253	201	130
507	237	144	113	634	291	147	127
457	235	138	113	578	308	292	100
438	235	138	110	487	272	164	107
415	216	138	110	442	284	143	114
401	208	138	108	451	255	169	111
387	201	136	106	459	270	139	163
381	192	132	104	464	214	170	116
324	180	130	101	400	195	150	122
315	179	126	100	366	260	143	134

Nota: Las ciudades están clasificadas en el mismo orden en ambos años.

Totales y sumas de cuadrados

	1920		1930	
Estrato	$\sum(x_{hi})$	$\sum(x_{hi}^2)$	$\sum(y_{hi})$	$\sum(y_{hi}^2)$
1	8 349	4 756 619	10 070	7 145 450
2	7 941	1 474 871	9 498	2 141 720

El número total de habitantes del total de 64 ciudades en 1930 se va a estimar de una muestra de tamaño 24. Encontrar el error estándar del error estimado por: (1) una muestra aleatoria simple, (2) una muestra aleatoria estratificada con asignación proporcional, (3) una muestra aleatoria estratificada con 12 unidades obtenidas de cada estrato.

Esta población se asemeja a las poblaciones de muchos tipos de negocios en que algunas unidades —las ciudades grandes— contribuyen sustancialmente al total y presentan una variabilidad mucho mayor que las restantes.

Los totales por estrato y las sumas de cuadrados se muestran abajo de la Tabla 5.1. En este ejemplo sólo se utiliza la información de 1930: la información de 1920 aparece en un ejemplo posterior.

Para la población completa en 1930, encontramos

$$Y = 19\,568 \qquad S^2 = 52\,448$$

Las tres estimaciones de Y se representan por $\hat{Y}_{ran}$, $\hat{Y}_{prop}$, y $\hat{Y}_{igual}$.

1. Con muestreo aleatorio simple

$$V(\hat{Y}_{ran}) = \frac{N^2S^2}{n}\frac{N-n}{N} = \frac{(64)^2(52\,448)}{24}\left(\frac{40}{64}\right) = 5\,594\,453$$

del Teorema 2.2, Corolario 2. El error estándar es

$$\sigma(\hat{Y}_{ran}) = 2365$$

2. Las varianzas para los estratos individuales son

$$S_1^2 = 53\,843, \qquad S_2^2 = 5581$$

Nótese que el estrato con las ciudades más grandes tiene una varianza aproximadamente 10 veces mayor que la del otro estrato.

Con asignación proporcional, tenemos $n_1 = 6$, $n_2 = 18$. De la Ec. (5.7) multiplicando por N^2, tenemos

$$V(\hat{Y}_{prop}) = \frac{N-n}{n}\sum N_hS_h^2$$

$$= \tfrac{40}{24}[(16)(53\,843)+(48)(5581)] = 1\,882\,293$$

$$\sigma(\hat{Y}_{prop}) = 1372$$

3. Para $n_1 = n_2 = 12$ utilizamos la fórmula general (5.9):

$$V(\hat{Y}_{igual}) = \sum N_h(N_h - n_h)\frac{S_h^2}{n_h}$$

$$= \frac{(16)(4)(53\,843)}{12} + \frac{(48)(36)(5581)}{12} = 1\,090\,827$$

$$\sigma(\hat{Y}_{igual}) = 1044$$

En este ejemplo las muestras de igual tamaño en ambos estratos son más precisas que la asignación proporcional. Ambas son superiores al muestreo aleatorio simple.

5.4 LA VARIANZA ESTIMADA Y LIMITES DE CONFIANZA

Si se toma una muestra aleatoria simple dentro de cada estrato, una estimación insesgada de S_h^2 (del Teorema 2.4) es

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 \tag{5.11}$$

Por lo tanto, obtenemos el siguiente.

Teorema 5.5. Con muestreo aleatorio estratificado, una estimación insesgada de la varianza de $\bar{y}_{st}$ es

$$v(\bar{y}_{st}) = s^2(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h} \tag{5.12}$$

Una forma alternativa para propósitos de cálculo es

$$s^2(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 s_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h s_h^2}{N} \tag{5.13}$$

El segundo término del lado derecho de la expresión anterior representa la reducción debida a la cpf.

Para calcular esta estimación, debe haber cuando menos dos unidades provenientes de todos y cada uno de los estratos. La estimación de la varianza cuando la estratificación se lleva al punto en el cual sólo se escoge una unidad por estrato, se discute en la Sec. 5A.12.

Las fórmulas para los límites de confianza son:

Media de la población: $\bar{y}_{st} \pm ts(\bar{y}_{st})$ (5.14)

Total de la población: $N\bar{y}_{st} \pm tNs(\bar{y}_{st})$ (5.15)

Estas fórmulas suponen que $\bar{y}_{st}$ está normalmente distribuida y que $s(\bar{y}_{st})$ está bien determinada, de modo que el multiplicador t pueda encontrarse en las tablas de la distribución normal.

Si sólo se dan algunos grados de libertad para cada estrato, el procedimiento usual para tomar en cuenta el error de muestreo inherente a una cantidad como $s(\bar{y}_{st})$ consiste en leer el valor t en las tablas de la t de Student, en lugar de la tabla normal. La distribución de $s(\bar{y}_{st})$ es en general demasiado compleja, para permitir una aplicación estricta de este método. El siguiente es un método aproximado de asignación de un número efectivo de grados de libertad a $s(\bar{y}_{st})$ (Satterthwaite, 1946).

Podemos escribir

$$s^2(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} g_h s_h^2, \quad \text{donde } g_h = \frac{N_h(N_h - n_h)}{n_h}$$

El número efectivo de grados de libertad n_e es

$$n_e = \frac{\left(\sum g_h s_h^2\right)^2}{\sum \frac{g_h^2 s_h^4}{n_h - 1}} \tag{5.16}$$

El valor de n_e siempre está entre el más pequeño de los valores $(n_h - 1)$ y su suma. La aproximación toma en cuenta el hecho de que S_h^2 puede variar de un estrato al otro y supone que y_{hi} sea normal, pues depende del resultado de que la varianza de s_h^2 sea $2\sigma_h^4/(n_h - 1)$. Si la distribución de y_{hi} tiene una curtosis positiva, la varianza de s_h^2 será mayor que esto y la Fórmula 5.16 sobreestima los grados efectivos de libertad.

5.5 LA ASIGNACION OPTIMA

En el muestreo estratificado, los valores de los tamaños de muestras n_h en los respectivos estratos, los elige quien hace el muestreo. Se pueden seleccionar para minimizar $V(\bar{y}_{st})$ para un costo específico de tomar la muestra, o bien, para minimizar el costo para un valor específico de $V(\bar{y}_{st})$.

La función de costo más simple es de la forma

$$\text{costo} = C = c_0 + \sum c_h n_h \tag{5.17}$$

Dentro de cualquier estrato, el costo es proporcional al tamaño de la muestra, pero el costo por unidad c_h puede variar entre estratos. El término c_0 representa un costo fijo.

Esta función de costo es apropiada cuando la parte principal del costo es la de tomar las medidas en cada unidad. Si los costos de viaje entre las unidades son importantes, los estudios empíricos y matemáticos sugieren que los costos de viaje se representen mejor por medio de la expresión $\Sigma t_h \sqrt{n_h}$ donde t_h es el costo de viaje por unidad (Beardwood y otros, 1959). Aquí sólo se considera la función de costo lineal (5.17).

Teorema 5.6. En el muestreo aleatorio estratificado, con una función de costo lineal de la forma (5.17), la varianza de la media estimada $\bar{y}_{st}$ es un mínimo para un costo específico C y el costo es un mínimo para una varianza específica $V(\bar{y}_{st})$ donde n_h es proporcional a $W_hS_h/\sqrt{c_h}$.

Prueba. Tenemos

$$C = c_0 + \sum_{h=1}^{L} c_h n_h \tag{5.17}$$

$$V = V(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}(1-f_h) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h} \tag{5.18}$$

Nuestros problemas son ya sea (1) elegir los n_h para minimizar V con C especificada, o bien, (2) elegir los n_h para minimizar C con V especificada. Lo que ocurre es que a excepción de los pasos finales, estos problemas tienen la misma solución. Al elegir los n_h para minimizar V con C fija o C con V fija, ambos son equivalentes a minimizar el producto,

$$V'C' = \left(V + \sum \frac{W_h^2 S_h^2}{N_h}\right)(C - c_0)$$

$$= \left(\sum \frac{W_h^2 S_h^2}{n_h}\right)\left(\sum c_h n_h\right) \tag{5.19}$$

Stuart (1954) observó que (5.19) se puede minimizar "limpiamente" usando la desigualdad de Cauchy-Schwarz. Si a_h, b_h son dos conjuntos de L números positivos, esta desigualdad proviene de la identidad

$$\left(\sum a_h^2\right)\left(\sum b_h^2\right) - \left(\sum a_h b_h\right)^2 = \sum_{i}\sum_{j>i} (a_i b_j - a_j b_i)^2 \tag{5.20}$$

De (5.20) vemos que

$$\left(\sum a_h^2\right)\left(\sum b_h^2\right) \geq \left(\sum a_h b_h\right)^2 \tag{5.21}$$

y la igualdad ocurrirá si y solamente si b_h/a_h es constante para todo h. En (5.19) tómese

$$a_h = \frac{W_h S_h}{\sqrt{n_h}}, \qquad b_h = \sqrt{c_h n_h}, \qquad a_h b_h = W_h S_h \sqrt{c_h}$$

La desigualdad (5.21) da

$$V'C' = \left(\sum \frac{W_h^2 S_h^2}{n_h}\right)\left(\sum c_h n_h\right) = \left(\sum a_h^2\right)\left(\sum b_h^2\right) \geq \left(\sum W_h S_h \sqrt{c_h}\right)^2$$

Por lo tanto, ninguna elección de los n_h puede hacer $V'C'$ menor que

$$\left(\sum W_h S_h \sqrt{c_h}\right)^2$$

El valor mínimo ocurre cuando

$$\frac{b_h}{a_h} = \frac{n_h \sqrt{c_h}}{W_h S_h} = \text{constante} \tag{5.22}$$

como se afirma en el teorema.

En términos del tamaño total de muestra n_h en un estrato, tenemos

$$\frac{n_h}{n} \equiv \frac{W_h S_h/\sqrt{c_h}}{\sum (W_h S_h/\sqrt{c_h})} = \frac{N_h S_h/\sqrt{c_h}}{\sum (N_h S_h/\sqrt{c_h})} \tag{5.23}$$

Este teorema da lugar a las siguientes reglas de conducta. En un estrato dado, se toma una muestra más grande si

1. El estrato es más grande.
2. El estrato es más variable internamente.
3. El muestreo es más barato en el estrato.

Se necesita un paso más para completar la asignación. La Ec. (5.23) da el n_h en términos de n, pero todavía no conocemos qué valor tiene n. La solución depende de si la muestra se escoge para satisfacer un costo total especificado C o para dar una varianza especificada V para y_{st}. Si el *costo* es fijo, se sustituyen los valores óptimos de n_h en la función de costo (5.17) y se resuelven para n. Esto da

$$n = \frac{(C - c_0)\sum (N_h S_h/\sqrt{c_h})}{\sum (N_h S_h \sqrt{c_h})} \tag{5.24}$$

Si V es fija, se sustituye la n_h óptima en la fórmula para $V(\bar{y}_{st})$. Encontramos

$$n = \frac{\left(\sum W_h S_h \sqrt{c_h}\right)\sum W_h S_h/\sqrt{c_h}}{V + (1/N)\sum W_h S_h^2} \tag{5.25}$$

donde $W_h = N_h/N$.

Un caso especial importante surge si $c_h = c$, esto es, si el costo por unidad es el mismo en todos los estratos. El costo se convierte en $C = c_0 + cn$ y la asignación óptima para un costo fijo se reduce a la asignación óptima para un tamaño de muestra fijo. El resultado de este caso especial es como sigue.

Teorema 5.7. En el muestreo aleatorio estratificado la $V(\bar{y}_{st})$ se minimiza para un tamaño de muestra total fijo n si

$$n_h = n\frac{W_h S_h}{\sum W_h S_h} = n\frac{N_h S_h}{\sum N_h S_h} \tag{5.26}$$

Esta asignación se llama a veces *asignación de Neyman*, por Neyman (1934) cuya prueba dio prominencia al resultado. Más tarde se descubrió una prueba anterior de Tschuprow (1923).

Una fórmula para la varianza mínima con n fija se obtiene al sustituir el valor de n en (5.26) en la fórmula general para $V(\bar{y}_{st})$. El resultado es

$$V_{mín}(\bar{y}_{st}) = \frac{\left(\sum W_h S_h\right)^2}{n} - \frac{\sum W_h S_h^2}{N} \tag{5.27}$$

El segundo término de la derecha representa la cpf.

5.6 PRECISIONES RELATIVAS DEL MUESTREO ALEATORIO ESTRATIFICADO Y DEL MUESTREO ALEATORIO SIMPLE

Si se usa la estratificación de manera inteligente, casi siempre da como resultado una varianza más pequeña para la media estimada o para el total, que la dada por un muestreo aleatorio simple comparable. Sin embargo, no es verdad que *cualquier* muestreo aleatorio estratificado dé una varianza menor que un muestreo aleatorio simple. Si los valores de los n_h están lejos del óptimo, el muestreo estratificado puede tener una varianza más alta. De hecho, aun la estratificación con asignación óptima para un tamaño de muestra total fijo puede dar una varianza más alta, aunque esto es más bien una curiosidad académica y no algo que sucede en la práctica.

En esta sección se comparan el muestreo aleatorio simple y el muestreo aleatorio estratificado con asignación proporcional y óptima. Esta comparación demuestra cómo se logra una ganancia gracias a la estratificación.

Las varianzas de las *medias* estimadas se denotan con V_{ran}, V_{prop} y $V_{ópt}$ respectivamente.

Teorema 5.8. Si se ignoran los términos en $1/N_h$, relativamente a la unidad,

$$V_{ópt} \leq V_{prop} \leq V_{ran} \tag{5.28}$$

donde la asignación óptima es para n fijo, o sea, con $n_h \propto N_h S_h$.

Prueba.

$$V_{ran} = (1-f)\frac{S^2}{n} \tag{5.29}$$

$$V_{prop} = \frac{(1-f)}{n}\sum W_h S_h^2 = \frac{\sum W_h S_h^2}{n} - \frac{\sum W_h S_h^2}{N} \tag{5.30}$$

[de la Ec. (5.8), Sec. 5.3.]

$$V_{ópt} = \frac{\left(\sum W_h S_h\right)^2}{n} - \frac{\sum W_h S_h^2}{N} \tag{5.31}$$

[de la Ec. (5.27), Sec. 5.5.]

Tenemos, usando la identidad algebraica estándar para el análisis de la varianza de una población estratificada

$$\begin{aligned}(N-1)S^2 &= \sum_h \sum_i (y_{hi} - \bar{Y})^2 \\ &= \sum_h \sum_i (y_{hi} - \bar{Y}_h)^2 + \sum_h N_h(\bar{Y}_h - \bar{Y})^2 \\ &= \sum_h (N_h - 1)S_h^2 + \sum_h N_h(\bar{Y}_h - \bar{Y})^2 \end{aligned} \tag{5.32}$$

Si los términos $1/N_h$ son despreciables y, por tanto, también lo es $1/N$, (5.32) obtenemos

$$S^2 = \sum W_h S_h^2 + \sum W_h(\bar{Y}_h - \bar{Y})^2 \tag{5.33}$$

Por lo tanto

$$V_{ran} = (1-f)\frac{S^2}{n} = \frac{(1-f)}{n}\sum W_h S_h^2 + \frac{(1-f)}{n}\sum W_h(\bar{Y}_h - \bar{Y})^2 \tag{5.34}$$

$$= V_{prop} + \frac{(1-f)}{n}\sum W_h(\bar{Y}_h - \bar{Y})^2 \tag{5.35}$$

por la definición de $V_{ópt}$ debemos tener $V_{prop} \geq V_{ópt}$. Por (5.30) y (5.31) su diferencia es

$$\begin{aligned} V_{prop} - V_{ópt} &= \frac{1}{n}\left[\sum W_h S_h^2 - \left(\sum W_h S_h\right)^2\right] \\ &= \frac{1}{n}\left[\sum W_h(S_h - \bar{S})^2\right] \end{aligned} \tag{5.36}$$

donde $\bar{S} = \sum W_h S_h$ es una media ponderada de las S_h.

A partir de (5.35) y (5.36) con los términos en $1/N_h$ despreciables,

$$V_{ran} = V_{ópt} + \frac{1}{n}\sum W_h(S_h - \bar{S})^2 + \frac{(1-f)}{n}\sum W_h(\bar{Y}_h - \bar{Y})^2 \qquad (5.37)$$

Para resumir, en la Ec. (5.37) hay dos componentes de la disminución en la varianza cuando pasamos del muestreo aleatorio simple a la asignación óptima. La primera componente (término a la extrema derecha) proviene de la eliminación de las diferencias entre las medias de estratos, la segunda (término medio a la derecha) proviene de la eliminación del efecto de las diferencias entre las desviaciones estándar de estratos. La segunda componente representa la diferencia en varianza entre la asignación óptima y la proporcional.

Si los términos en $1/N_h$ no son despreciables, el sustituir por S^2 de (5.32) se obtiene

$$V_{ran} = V_{prop} + \frac{(1-f)}{n(N-1)}\left[\sum N_h(\bar{Y}_h - \bar{Y})^2 - \frac{1}{N}\sum (N-N_h)S_h^2\right] \qquad (5.38)$$

en lugar de (5.35).

Por lo tanto, la estratificación proporcional da una varianza más elevada que el muestreo aleatorio simple si

$$\sum N_h(\bar{Y}_h - \bar{Y})^2 < \frac{1}{N}\sum (N-N_h)S_h^2 \qquad (5.39)$$

Matemáticamente, esto puede ocurrir. Supóngase que las S_h^2 son iguales a Sw^2 de modo que la asignación proporcional es óptima en el sentido de Neyman. Entonces (5.39), viene a ser

$$\sum N_h(\bar{Y}_h - \bar{Y})^2 < (L-1)S_w^2$$

o

$$\frac{\sum N_h(\bar{Y}_h - \bar{Y})^2}{L-1} < S_w^2 \qquad (5.40)$$

Los lectores familiarizados con el análisis de la varianza, reconocerán esta relación, que implica que el cuadro de la media entre estratos es más pequeña que dentro de los estratos, es decir, que la razón F es menor que 1.

5.7 ¿EN QUE CASOS PRODUCE LA ESTRATIFICACION CONSIDERABLES GANANCIAS DE PRECISION?

La variable ideal para la estratificación, es el valor de y en sí, o sea, la cantidad que se va a medir en la encuesta. Si pudiésemos

estratificar por los valores de y, no habría traslape entre los estratos y la varianza dentro de ellos, sería mucho menor que la varianza global, particularmente si hay muchos estratos. Esta situación se ilustra mediante el ejemplo de la Sec. 5.3, Pág. 130. La población consistía en los tamaños (números de habitantes) de 64 ciudades en 1930, estratificadas por tamaños. Aunque sólo había dos estratos la estratificación proporcional reducía la estimación muestral ($\hat{Y}$) de 2365 a 1372. Al estratificar con $n_1 = n_2 = 12$, que es el óptimo en la asignación de Neyman, se obtenía una nueva reducción a 1044

En la práctica desde luego, no podemos estratificar por valores de y. Pero algunas aplicaciones importantes se acercan a esta situación, y dan lugar a una mejora considerable en la precisión, si se satisfacen las siguientes condiciones.

1. La población consta de instituciones que varían mucho en tamaño.
2. Las principales variables a medir están íntimamente relacionadas con los tamaños de las instituciones.
3. Se cuenta con una buena medida de tamaño para establecer los estratos.

Como ejemplos tenemos los negocios de un tipo específico, por ejemplo, tiendas de alimentos (en las encuestas que tratan sobre el volumen de los negocios o el número de empleados), escuelas (en encuestas relacionadas con el número de alumnos), hospitales (en estudios sobre la cantidad de pacientes), y formas de impuestos (para características altamente correlacionadas con el ingreso susceptible de impuesto). En los Estados Unidos, las granjas varían mucho en tamaño, al medirlas según su extensión total o su ingreso bruto, pero las características comunes, como son la producción de ciertas cosechas o tipos de ganado, con frecuencia muestran solamente una correlación moderada con respecto al tamaño de la granja, por lo tanto, no se puede ganar mucho al estratificar los tamaños de las granjas.

Si el tamaño de la institución permanece estable a través del tiempo, o al menos en periodos cortos, generalmente, su mejor medida práctica es el tamaño de la institución tomado en un censo reciente. El ejemplo de la Sec. 5.3, ilustra la situación en la que se disponía de buenos datos previos. La Tabla 5.2 muestra las S_h y los óptimos $n_h \propto N_h S_h$ resultantes, cuando la asignación se hace de datos de 1920 y de 1930, respectivamente.

Los datos de 1920 indican un n_1 de 11.56 contra un óptimo "verdadero" de 12.21 para los datos de 1930. Se redondea a enteros,

TABLA 5.2. CÁLCULO DE LA ASIGNACIÓN OPTIMA

Estrato	N_h	1920 Datos S_h	1920 Datos N_hS_h	1920 Datos n_h	1930 Datos S_h	1930 Datos N_hS_h	1930 Datos n_h
1	16	163.30	2612.80	11.56	232.04	3712.64	12.21
2	48	58.55	2810.40	12.44	74.71	3586.08	11.79
Totales	64		5423.20	24.00		7298.72	24.00

ambos conjuntos de datos dan la misma asignación —un tamaño de muestra de 12 para cada estrato.

Nótese que la fracción óptima de muestreo es 75% en el estrato 1, pero sólo 25% en el estrato 2. Con frecuencia se encuentra que debido a la gran variabilidad del estrato que consiste en las instituciones grandes, la fórmula requiere un muestreo del 100% en dicho estrato. De hecho, la asignación puede requerir más del 100% de muestreo (ver Sec. 5.8). Nótese también que las S_h son menores en 1920 que en 1930. Los datos de 1920 dan una impresión demasiado optimista de la precisión que se obtendrá de una encuesta en 1930. Como se mencionó en la Sec. 4.7, la posibilidad de un cambio en los niveles de las S_h siempre debería considerarse al usar datos anteriores, aunque una consideración del cambio debe contener algo de adivinanza.

La estratificación geográfica en la que los estratos son áreas compactas, como municipios o colonias de una ciudad, es común —a menudo por conveniencia administrativa o porque se quieren datos separados para cada estrato— y generalmente viene acompañada con un incremento en la precisión porque operan muchos factores para lograr que las personas que viven o las cosechas que se cultivan en una misma área muestren semejanzas en sus principales características. Sin embargo, las ganancias debidas a la estratificación geográfica, en general son modestas así, por ejemplo, la Tabla 5.3 muestra datos publicados por Jessen (1942) y Jessen y Houseman (1944) sobre la efectividad de la estratificación geográfica para cierto número de atributos económicos típicos de las granjas.

Se representan cuatro tamaños de estratos —la ciudad, el municipio, el área típica de granja y el estado—. Para dar una idea de los tamaños relativos de los estratos, hay alrededor de 1 600 poblaciones, 100 municipios y 5 áreas en Iowa.

En la tabla se toma la precisión de un método de estratificación como inversamente proporcional al valor de $V(\bar{y}_{st})$ que da el método.

TABLA 5.3. PRECISIÓN RELATIVA DE LAS DIFERENTES CLASES DE ESTRATIFICACIÓN GEOGRÁFICA EN PORCENTAJE

Estado	No. de atributos	Estrato: Población	Estrato: Municipio	Estrato: Area agrícola	Estrato: Estado
Iowa, 1938	18	115	100	96	91
Iowa, 1939	19	121	100	97	91
Florida, 1942					
Area de cítricos	14	144	100	..	
Area de hortalizas	15	111	100	..	
California, 1942	17	113	100	97	

Por lo tanto, la precisión relativa del método 1 respecto al 2 es la razón $V_2(\bar{y}_{st})/V_1(\bar{y}_{st})$, expresada como porcentaje. Los datos mostrados son promedios sobre los números de atributos de la segunda columna. El municipio se toma como estándar en cada caso. Como se indicó, las ganancias en precisión son moderadas. En Iowa, el uso de 1 600 estratos (poblaciones) comparado con la no estratificación (el estado) aumenta la precisión en 30% más o menos; es decir, reduce la varianza en un 25% aproximadamente.

En lo que respecta a la estratificación proporcional, en contra de la estratificación óptima, hay dos situaciones en que la estratificación óptima gana gallardamente. El primero es el caso ya discutido en donde la población consiste en instituciones grandes y pequeñas estratificadas por alguna medida del tamaño. Las varianzas S_h^2 son usualmente mucho mayores para las instituciones grandes que para las pequeñas, lo que hace ineficiente la estratificación proporcional. La segunda situación se encuentra en encuestas donde algunos estratos son mucho más costosos de muestrear que otros. La influencia de los factores $\sqrt{c_h}$ puede hacer que la asignación resulte mediocre.

Al planear una asignación en donde el n_h estimado no difiere mucho de la proporcionalidad, vale la pena estimar qué tanto más grandes se convierten $V(\bar{y}_{st})$ o $V(\hat{Y}_{st})$ si se usa la asignación proporcional. Los óptimos en el problema de asignación son más bien bajos (ver la Sec. 5A.2) y el incremento de la varianza puede resultar sorprendentemente pequeño. Además, la superioridad del óptimo calculado aquí a partir de los valores estimados de los S_h es siempre exagerado por los errores en los S_h estimados. La sencillez y autoponderación de la asignación proporcional valen probablemente de un 10 a un 20% de incremento en la varianza.

5.8 ASIGNACION QUE REQUIERE MAS DEL 100% DEL MUESTREO

Como se mencionó en la Sec. 5.7, la fórmula para el óptimo puede producir en cierto estrato un n_h que sea mayor que el correspondiente N_h. Considérese el ejemplo de los tamaños de las ciudades en la Sec. 5.3. Una muestra de 24 ciudades distribuidas entre los estratos, requería 12 ciudades de 16 en el primero y 12 de 48 en el segundo. Si la muestra hubiese sido de tamaño 48, la asignación hubiera requerido 24 ciudades de 16 en el primer estrato. Lo más que se puede hacer es tomar todas las ciudades en el estrato, y dejar 32 ciudades para el segundo estrato en lugar de las 24 postuladas por la fórmula. Este problema sólo surge cuando la fracción global de muestreo es importante, y algunos estratos son mucho más variables que otros. Esto ocurre en la práctica con mucha frecuencia.

Si la estimación original da $n_1 > N_1$ cuando hay más de dos estratos, la asignación óptima revisada es:

$$\tilde{n}_1 = N_1; \qquad \tilde{n}_h = (n - N_1)\frac{W_h S_h}{\sum_2^L W_h S_h}, \qquad (h \geq 2) \tag{5.41}$$

siempre que $\tilde{n}_h \leq N_h$ para $h \geq 2$. Si ocurriese que $\tilde{n}_2 > N_2$, cambiaríamos la asignación a

$$\tilde{n}_1 = N_1; \qquad \tilde{n}_2 = N_2; \qquad \tilde{n}_h = (n - N_1 - N_2)\frac{W_h S_h}{\sum_3^L W_h S_h}, \qquad (h \geq 3) \tag{5.41}'$$

Siempre y cuando $\tilde{n}_h \leq N_h$ para $h \geq 3$. Continuamos con este proceso hasta que cada $\tilde{n}_h \leq N_h$. La asignación resultante puede demostrarse que es óptima para un valor de n dado, como era de esperarse.

Hay que tener cuidado al usar la fórmula correcta para $V(\bar{y}_{st})$. La fórmula general (5.6) de la Sec. 5.3 es correcta si se sustituyen los $\tilde{n}_h$ dados por la asignación óptima *revisada*. La Fórmula (5.27) para $V_{mín}(\bar{y}_{st})$,

$$V_{mín}(\bar{y}_{st}) = \frac{(\sum W_h S_h)^2}{n} - \frac{\sum W_h S_h^2}{N} \tag{5.27}$$

no vale ya. Si $\sum'$ denota suma en los estratos para los cuales $\tilde{n}_h \leq N_h$, una fórmula alternativa correcta es

$$V_{mín}(\bar{y}_{st}) = \frac{(\sum' W_h S_h)^2}{n'} - \frac{\sum' W_h S_h^2}{N} \tag{5.42}$$

donde n' es el tamaño total revisado de la muestra en estos estratos.

5.9 ESTIMACION DEL TAMAÑO DE LA MUESTRA CON DATOS CONTINUOS

Las fórmulas para la determinación del valor de n en una asignación óptima estimada son las que se dieron en la Sec. 5.5. Esta sección presenta las fórmulas para cualquier asignación con algunos casos especiales de utilidad. Se supone que la estimación se ha especificado con varianza V. Si en lugar de ello, se especifica el margen de error d (Sec. 4.4), $V = (d/t)^2$ donde t es el desvío correspondiente a la probabilidad que podemos conceder de que el error exceda el margen deseado.

Estimación de la Media de Población $\bar{Y}$

Sea s_h la estimación de S_h y sea $n_h = w_h n$ donde los w_h han sido elegidos. En estos términos, la $V(\bar{y}_{st})$ anticipada (por el Teorema 5.3, Sec. 5.3.) es

$$V = \frac{1}{n}\sum \frac{W_h^2 s_h^2}{w_h} - \frac{1}{N}\sum W_h s_h^2 \tag{5.43}$$

con $W_h = N_h/N$. Esto da como fórmula general para n

$$n = \frac{\sum \frac{W_h^2 s_h^2}{w_h}}{V + \frac{1}{N}\sum W_h s_h^2} \tag{5.44}$$

si se ignora la cpf tenemos, como primera aproximación

$$n_0 = \frac{1}{V}\sum \frac{W_h^2 s_h^2}{w_h} \tag{5.45}$$

Si n_0/N no es despreciable, podemos calcular n como

$$n = \frac{n_0}{1 + \frac{1}{NV}\sum W_h s_h^2} \tag{5.46}$$

En casos particulares, las fórmulas toman varias formas que pueden ser convenientes para el cálculo. Algunas son:

Asignación óptima supuesta (para n fijo): $w_h \propto W_h s_h$.

$$n = \frac{(\sum W_h s_h)^2}{V + \frac{1}{N}\sum W_h s_h^2} \tag{5.47}$$

Asignación proporcional: $w_h = W_h = N_h/N$.

$$n_o = \frac{\sum W_h s_h^2}{V}, \qquad n = \frac{n_0}{1+\frac{n_0}{N}} \tag{5.48}$$

Estimación del Total de Población

Si V es la $V(\hat{Y}_{st})$ deseada, las fórmulas principales son

En general:

$$n = \frac{\sum \frac{N_h^2 s_h^2}{w_h}}{V + \sum N_h s_h^2} \tag{5.49}$$

Optimo supuesto (para n fijo):

$$n = \frac{(\sum N_h s_h)^2}{V + \sum N_h s_h^2} \tag{5.50}$$

Proporcional:

$$n_0 = \frac{N}{V}\sum N_h s_h^2, \qquad n = \frac{n_0}{1+\frac{n_0}{N}} \tag{5.51}$$

Ejemplo. Este ejemplo proviene de un artículo de Cornell (1947) que describe una muestra de universidades y colegios de los Estados Unidos tomada en 1946 por la oficina de educación con objeto de estimar las inscripciones para el año académico 1946-1947. La ilustración es para la población de 196 colegios de maestros y escuelas normales. Se les dispuso en siete estratos, de los cuales se ignorará uno pequeño. Los cinco primeros estratos se construyeron por tamaño de la institución: el sexto sólo contenía colegios para mujeres. Se calcularon las estimaciones s_h de las S_h a partir de los resultados obtenidos para el año académico 1943-1944. Se usó una estratificación "óptima" basada en estos valores s_h.

El objetivo era un coeficiente de variación de 5% en el total de inscripciones estimado. En 1943 el número total de inscripciones para este grupo de colegios fue de 56 472. De modo que el error estándar deseado es

$$(0.05)(56\,472) = 2824$$

y, por lo tanto, la varianza deseada es

$$V = (2824)^2 = 7\,974\,976$$

Se puede objetar que el número de inscripciones será mayor en 1946 que en 1943 y que la asignación debería hacerse para este incremento. En realidad el cálculo supone solamente que el cv por colegio es igual en 1943 que en 1946, —lo que parece razonable.

TABLA 5.4. DATOS PARA LA ESTIMACIÓN DEL TAMAÑO DE MUESTRA

Estrato	N_h	s_h	$N_h s_h$	n_h
1	13	325	4 225	9
2	18	190	3 420	7
3	26	189	4 914	10
4	42	82	3 444	7
5	73	86	6 278	13
6	24	190	4 560	10
Totales	196		26 841	56

La Tabla 5.4 muestra los valores de N_h, s_h y $N_h s_h$, que se conocían antes de determinar n.

La fórmula apropiada para n es (5.50) que se aplica a una asignación "óptima" para estimar un total. Con sólo 196 unidades en esta población no es probable que la cpf sea despreciable. Sin embargo, para propósitos de ilustración, una primera aproximación que ignora la cpf será la que se busque. Esto es:

$$n_0 = \frac{(\sum N_h s_h)^2}{V} = \frac{(26\,841)^2}{7\,974\,976} = 90.34$$

Obviamente es necesario un ajuste. Para la n correcta en (5.50), tenemos

$$n = \frac{n_0}{1+\frac{1}{V}\sum N_h s_h^2} = \frac{90.34}{1+\frac{4\,640\,387}{7\,974\,976}} = 57.1$$

Se escogió un tamaño de muestra de 56 unidades.* La n_h para estratos individuales aparece en la columna del lado derecho de la Tabla 5.4.

5.10 MUESTREO ESTRATIFICADO PARA PROPORCIONES

Si queremos estimar la proporción de unidades en la población que pertenecen a una clase definida C, la estratificación ideal se obtiene si colocamos en el primer estrato toda unidad que pertenezca a C y en el segundo toda unidad que no pertenezca a C. Si esto falla, tratamos de construir estratos tales que la proporción en la clase C varíe tanto como sea posible de estrato a estrato.

Sea

$$P_h = \frac{A_h}{N_h}, \qquad p_h = \frac{a_h}{n_h}$$

la proporción de unidades en C en el h-ésimo estrato y en la muestra de ese estrato, respectivamente. Para la proporción en la pobla-

* Los resultados aritméticos difieren ligeramente de los obtenidos por Cornell (1947).

ción total, la estimación apropiada en el muestreo aleatorio estratificado es

$$p_{st} = \sum \frac{N_h p_h}{N} \tag{5.52}$$

Teorema 5.9. Con el muestreo aleatorio estratificado, la varianza de p_{st} es

$$V(p_{st}) = \frac{1}{N^2} \sum \frac{N_h^2 (N_h - n_h)}{N_h - 1} \frac{P_h Q_h}{n_h} \tag{5.53}$$

Prueba. Este es un caso particular del teorema general para la varianza de la media estimada. Del Teorema 5.3

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum N_h (N_h - n_h) \frac{S_h^2}{n_h} \tag{5.54}$$

Sea y_{hi} una variable que toma el valor de uno cuando la unidad está en C y 0 en otro caso. En la Sec. 3.2, Ec. 3.4, se demostró que para esta variable

$$S_h^2 = \frac{N_h}{N_h - 1} P_h Q_h \tag{5.55}$$

Esto da el resultado.

Nota: Prácticamente en todas las aplicaciones, aun si la cpf no es despreciable, los términos en $1/N_h$ serán despreciables y se puede usar la siguiente fórmula ligeramente más simple.

$$V(p_{st}) = \frac{1}{N^2} \sum N_h (N_h - n_h) \frac{P_h Q_h}{n_h} = \sum \frac{W_h^2 P_h Q_h}{n_h} (1 - f_h) \tag{5.56}$$

Corolario 1. Cuando pueda ignorarse la cpf

$$V(p_{st}) = \sum W_h^2 \frac{P_h Q_h}{n_h} \tag{5.57}$$

Corolario 2. Con asignación proporcional

$$V(p_{st}) = \frac{N - n}{N} \frac{1}{nN} \sum \frac{N_h^2 P_h Q_h}{N_h - 1} \tag{5.58}$$

$$\doteq \frac{1 - f}{n} \sum W_h P_h Q_h \tag{5.59}$$

Para la estimación muestral de la varianza, se sustituye $p_h q_h / (n_h - 1)$ por la incógnita $P_h Q_h / n_h$ en cualquiera de las fórmulas indicadas anteriormente.

La mejor elección de n_h para minimizar $V(p_{st})$ se desprende de la teoría general en la Sec. 5.5.

Varianza mínima para un tamaño de muestra total fijo.

$$n_h \propto N_h \sqrt{N_h/(N_h - 1)} \sqrt{P_h Q_h} \doteq N_h \sqrt{P_h Q_h}$$

Por lo que

$$n_h \doteq n \frac{N_h \sqrt{P_h Q_h}}{\sum N_h \sqrt{P_h Q_h}} \tag{5.60}$$

Varianza mínima con un costo fijo, donde costo $= c_0 + \sum C_h n_h$.

$$n_h \doteq n \frac{N_h \sqrt{P_h Q_h / c_h}}{\sum N_h \sqrt{P_h Q_h / c_h}} \tag{5.61}$$

El valor de n se encuentra como en la Sec. 5.5.

5.11 GANANCIAS EN PRECISION EN EL MUESTREO ESTRATIFICADO PARA PROPORCIONES

Si los costos por unidad son los mismos en todos los estratos, dos reglas de trabajo útiles son: (*a*) la ganancia en precisión del muestreo aleatorio estratificado sobre el muestreo aleatorio simple es pequeña o modesta a menos que la P_h varíe mucho de estrato a estrato; (*b*) la asignación óptima con una n fija gana poco sobre la asignación proporcional si todas las P_h caen entre 0.1 y 0.9.

Para ilustrar el primer resultado, en la Tabla 5.5 se muestra la comparación del muestreo aleatorio estratificado (asignación proporcional) con el muestreo aleatorio simple en 3 estratos de igual tamaño ($W_h = \frac{1}{3}$). Están incluidos 4 casos, el primero con $P_h = 0.4$, 0.5 y 0.6 en los tres estratos y el último (y más extremo) con $P_h =$ 0.1, 0.5 y 0.9. Las siguientes dos columnas muestran las varianzas

TABLA 5.5. PRECISIÓN RELATIVA DEL MUESTREO ESTRATIFICADO Y EL MUESTREO ALEATORIO SIMPLE

	Simple	Estratificado	
P_h	$nV(p)/(1-f) = PQ$	$nV(p_{st})/(1-f) = \frac{1}{3}\sum P_h Q_h$	Precisión relativa (%)
0.4, 0.5, 0.6	2500	2433	103
0.3, 0.5, 0.7	2500	2233	112
0.2, 0.5, 0.8	2500	1900	132
0.1, 0.5, 0.9	2500	1433	174

de la proporción estimada, multiplicada por $n/(1-f)$ y la última columna está formada por las precisiones relativas del muestreo estratificado respecto al muestreo aleatorio simple. La ganancia en precisión es grande solamente en los últimos dos casos.

Para comparar la asignación proporcional con la asignación óptima con n fijo, se encontrará que si se ignora la cpf

$$V_{\acute{o}pt}=\frac{(\sum W_h\sqrt{P_hQ_h})^2}{n}, \qquad V_{prop}=\frac{\sum W_hP_hQ_h}{n} \tag{5.62}$$

Por lo tanto, la precisión relativa de la asignación proporcional respecto a la asignación óptima es,

$$\frac{V_{\acute{o}pt}}{V_{prop}}=\frac{(\sum W_h\sqrt{P_hQ_h})^2}{\sum W_hP_hQ_h} \tag{5.63}$$

Si todas las P_h se encuentran entre los dos valores P_0 y $(1-P_0)$, estamos interesados en el valor más pequeño que tomará la precisión relativa. Por simplicidad, consideremos dos estratos de igual tamaño ($W_1 = W_2$). La precisión relativa mínima se obtiene cuando $P_1 = \frac{1}{2}$ y $P_2 = P_0$. Entonces, la precisión relativa es

$$\frac{V_{\acute{o}pt}}{V_{prop}}=\frac{(0.5+\sqrt{P_0Q_0})^2}{2(0.25+P_0Q_0)} \tag{5.64}$$

Algunos valores de esta función se presentan en la Tabla 5.6. Aun con $P_0 = 0.1$ o tan alto como 0.9, la precisión relativa es 94%. En la mayoría de los casos, la simplicidad y la característica de autoponderación de la estratificación proporcional hace más que compensar la ligera pérdida en precisión.

TABLA 5.6. PRECISIÓN RELATIVA DE LA ASIGNACIÓN PROPORCIONAL RESPECTO A LA ASIGNACIÓN OPTIMA

P_0	0.4 o 0.6	0.3 o 0.7	0.2 o 0.8	0.1 o 0.9	0.05 o 0.95
RP(%)	100.0	99.8	98.8	94.1	86.6

Deben notarse las limitaciones de este ejemplo. No toma en consideración los costos diferenciales debidos al muestreo en los diferentes estratos. En algunas encuestas las P_h son muy pequeñas, pero varían de, digamos, 0.001 a 0.05 en diferentes estratos. Aquí habrá una ganancia más sustancial respecto a la estratificación óptima.

5.12 ESTIMACION DEL TAMAÑO DE MUESTRA CON PROPORCIONES

Las fórmulas pueden deducirse de las más generales en la Sec. 5.9. Sea V la varianza deseada en la estimación de la proporción P en la población total. Las fórmulas para los dos tipos principales de asignación son las siguientes.

Proporcional:

$$n_0=\frac{\sum W_hp_hq_h}{V}, \qquad n=\frac{n_0}{1+\frac{n_0}{N}} \tag{5.65}$$

Optimo supuesto:

$$n_0=\frac{(\sum W_h\sqrt{p_hq_h})^2}{V}, \qquad n=\frac{n_0}{1+\frac{1}{NV}\sum W_hp_hq_h} \tag{5.66}$$

donde n_0 es la primera aproximación, la cual ignora la cpf, y n es el valor corregido tomando en consideración las cpf. En el desarrollo de estas fórmulas, los factores $N_h/(N_h-1)$ se han considerado como la unidad.

Estos resultados son aplicables a la estimación de una *proporción*. Si es preferible pensar en términos porcentuales, las mismas fórmulas se aplican si p_h, q_h, V, etc., se expresan como porcentajes. Para la estimación del número total de unidades en la población que pertenecen a la clase C, es decir, de NP, todas las varianzas se multiplican por N^2.

EJERCICIOS

5.1. En una población con $N = 6$ y $L = 2$ los valores de y_{hi} son 0, 1, 2 en el estrato 1 y 4, 6, 11 en el estrato 2. Se va a tomar una muestra con $n = 4$. (*a*) Demostrar que las n_h óptimas bajo la asignación de Neyman, aproximando a números enteros, son $n_h = 1$ en el estrato 1 y $n_h = 3$ en el estrato 2. (*b*) Calcular la estimación $\bar{y}_{st}$ para todas las muestras posibles que pueden derivarse bajo la asignación óptima y bajo la asignación proporcional. Verificar que las estimaciones son insesgadas. Por lo tanto, encontrar directamente $V_{\acute{o}pt}(\bar{y}_{st})$ y $V_{prop}(\bar{y}_{st})$. (*c*) Verificar que $V_{\acute{o}pt}(\bar{y}_{st})$ concuerda con la fórmula dada en la Ec. 5.6 y que $V_{prop}(\bar{y}_{st})$ concuerda con la fórmula dada en la Ec. 5.8, Pág. 129. (*d*) El uso de la fórmula 5.27, Pág. 136, para calcular $V_{\acute{o}pt}(\bar{y}_{st})$ es ligeramente incorrecto porque no considera el hecho de que las n_h fueron aproximadas a enteros. ¿Qué tan bien concuerda este resultado con el valor corregido?

5.2. Las familias de un pueblo se van a muestrear para estimar la cantidad promedio de bienes por familia que se pueden convertir en dinero en efectivo rápidamente. Las familias se estratifican en un estrato de renta alta y otro de renta baja. Se piensa que una casa en el estrato de renta alta tiene cerca de 9 veces más bienes que los existentes en una casa en el estrato de renta baja, y se espera que S_h sea proporcional a la raíz cuadrada de la media del estrato.

Existen 4 000 familias en el estrato de renta alta y 20 000 familias en el estrato de renta baja. (*a*) ¿Cómo distribuiría una muestra de 1 000 familias entre los dos estratos? (*b*) Si el objetivo es estimar la diferencia entre los bienes por familia en ambos estratos, ¿cómo debe estar distribuida la muestra?

5.3. La información que aparece a continuación, representa la estratificación de todas las propiedades agrícolas en un condado, clasificadas por tamaño promedio de acres de maíz por propiedad en cada estrato. Para una muestra de 100 propiedades, calcule los tamaños de muestras en cada estrato bajo (*a*) asignación proporcional,[9] (*b*) asignación óptima. Compare las precisiones de estos métodos con la del muestreo aleatorio simple.

Tamaño de la propiedad (acres)	Número de propiedades N_h	Promedio de acres de maíz $\bar{Y}_h$	Desviación estándar S_h
0–40	394	5.4	8.3
41–80	461	16.3	13.3
81–120	391	24.3	15.1
121–160	334	34.5	19.8
161–200	169	42.1	24.5
201–240	113	50.1	26.0
241–	148	63.8	35.2
Total o media	2010	26.3	

5.4. Verifique el resultado establecido en la fórmula 5.38, Sec. 5.6:

$$V_{ran} = V_{prop} + \frac{(1-f)}{n(N-1)}\left[\sum N_h(\bar{Y}_h - \bar{Y})^2 - \frac{1}{N}\sum (N-N_h)S_h^2\right]$$

5.5 Un muestreador tiene dos estratos con tamaños relativos W_1, W_2. El cree que S_1, S_2 pueden ser consideradas iguales pero piensa que c_2 puede estar entre 2 veces c_1 y 4 veces c_1. Preferiría usar la asignación proporcional pero no quiere incurrir en un aumento sustancial en la varianza comparada con la asignación óptima. Para un costo dado $C = c_1n_1 + c_2n_2$, ignorando la cpf, demostrar que

$$\frac{V_{prop}(\bar{y}_{st})}{V_{opt}(\bar{y}_{st})} = \frac{W_1c_1 + W_2c_2}{(W_1\sqrt{c_1} + W_2\sqrt{c_2})^2}$$

Si $W_1 = W_2$, calcule el aumento relativo en la varianza al usar la asignación proporcional cuando $c_2/c_1 = 2, 4$.

5.6. Un muestreador propone tomar una muestra aleatoria estratificada. Espera que sus costos de campo serán de la forma $\sum c_h n_h$. Sus estimaciones adelantadas de las cantidades relevantes para los dos estratos son como sigue:

Estrato	W_h	S_h	C_h
1	0.4	10	\$4
2	0.6	20	\$9

(*a*) Encuentre los valores de n_1/n y n_2/n que minimizan los costos de campo totales para un valor dado de $V(\bar{y}_{st})$. (*b*) Encontrar el tamaño de muestra requerido bajo esta asignación óptima, para hacer $V(\bar{y}_{st}) = 1$. Ignore la cpf. (*c*) ¿Cuál será el costo total de campo?

5.7. Después de haber tomado la muestra en el Ejercicio 5.6, el muestreador encuentra que sus costos de campo fueron, en realidad, de \$2.00 por unidad en el estrato 1 y de \$12.00 en el estrato 2. (*a*) ¿Qué tanto más alto es el costo de campo en comparación con lo que se había previsto? (*b*) Si hubiesen conocido los costos correctos de campo por adelantado, ¿podría haber obtenido la $V(\bar{y}_{st}) = 1$ para el costo de campo estimado originalmente en el Ejercicio 5.6? (*Nota.* La desigualdad Cauchy-Schwarz, en la Pág. 134, con $V' = 1$, da la respuesta a esta pregunta sin encontrar la nueva asignación.)

5.8. En una estratificación con 2 estratos, los valores de W_h y S_h son como sigue:

Estrato	W_h	S_h
1	0.8	2
2	0.2	4

Calcule los tamaños de muestra n_1, n_2 necesarios en ambos estratos para satisfacer las siguientes condiciones. Cada caso requiere un cálculo separado. (Ignore la cpf.) (*a*) El error estándar de estimación de la media de la población $\bar{y}_{st}$ debe ser 0.1 y el tamaño de muestra total $n = n_1 + n_2$ debe ser minimizado. (*b*) El error estándar de la media estimada de cada estrato debe ser 0.1. (*c*) El error estándar de la diferencia entre 2 medias estimadas de los estratos debe ser 0.1, nuevamente minimizando el tamaño de muestra total.

5.9. Con dos estratos, a un muestrador le gustaría tener $n_1 = n_2$ por conveniencia administrativa, en lugar de usar los valores dados por la asignación de Neyman. Si $V(\bar{y}_{st})$, $V_{opt}(\bar{y}_{st})$ representan las varianzas dadas por la asignación $n_1 = n_2$ y la asignación de Neyman, respectivamente, demuestre que el aumento fraccionario en la varianza es

$$\frac{V(\bar{y}_{st}) - V_{opt}(\bar{y}_{st})}{V_{opt}(\bar{y}_{st})} = \left(\frac{r-1}{r+1}\right)^2$$

en donde $r = n_1/n_2$ de acuerdo con la asignación de Neyman. Para los estratos en el Ejercicio 5.8, caso *a*, ¿cuál sería el aumento fraccionario de la varianza al usar $n_1 = n_2$ en lugar de la asignación óptima?

5.10. Si la función de costos es de la forma $C = c_0 + \sum t_h \sqrt{n_h}$, en donde c_0 y la t_h son números conocidos, demuestre que con el fin de minimizar $V(\bar{y}_{st})$ para un costo total fijo n_h debe ser proporcional a

$$\left(\frac{W_h^2 S_h^2}{t_h}\right)^{2/3}$$

Encontrar los n_h para una muestra de tamaño 1 000 bajo las siguientes condiciones:

Estrato	W_h	S_h	t_h
1	0.4	4	1
2	0.3	5	2
3	0.2	6	4

5.11. Si $V_{prop}(\bar{y}_{st})$ es la varianza de la media estimada de una muestra aleatoria estratificada de tamaño n con asignación proporcional y $V(\bar{y})$ es la varianza de la media de una muestra aleatoria simple de tamaño n, demuestre que la razón

$$\frac{V_{prop}(\bar{y}_{st})}{V(\bar{y})}$$

no depende del tamaño de muestra, en tanto que la razón

$$\frac{V_{mín}(\bar{y}_{st})}{V_{prop}(\bar{y}_{st})}$$

disminuye conforme n aumenta. (Esto implica que la asignación óptima para n fijo se vuelve más efectiva en relación a la asignación proporcional conforme n aumenta.) (Utilice las fórmulas 5.8 y 5.27.)

5.12. Compare los valores obtenidos para $V(p_{st})$ bajo la asignación proporcional y la asignación óptima para un tamaño de muestra fijo en las dos poblaciones siguientes. Cada estrato es de igual tamaño. La cpf se puede ignorar.

Población 1		Población 2	
Estrato	P_h	Estrato	P_h
1	0.1	1	0.01
2	0.5	2	0.05
3	0.9	3	0.10

¿Qué resultado general se ilustra con estas dos poblaciones?

5.13. Demuestre que en la estimación de proporciones los resultados correspondientes al Teorema 5.8 son como sigue:

$$V_{ran} = V_{prop} + \frac{(1-f)}{n} \sum W_h (P_h - P)^2$$

$$V_{prop} = V_{opt} + \frac{\sum W_h (\sqrt{P_h Q_h} - \sqrt{\overline{P_h Q_h}})^2}{n}$$

en donde

$$\sqrt{\overline{P_h Q_h}} = \sum W_h \sqrt{P_h Q_h}.$$

5.14. En una compañía, el 62% de los empleados son hombres experimentados o sin experiencia, el 31% son mujeres oficinistas y el 7% son supervisores. De una muestra de 400 empleados la compañía desea estimar la proporción que utiliza ciertos servicios de diversión. Las conjeturas son de que los servicios son utilizados por el 40 o 50% de los hombres, 20 o 30% de las mujeres y 5 o 10% de los supervisores. (*a*) ¿Cómo asignaría la muestra dentro de estos tres grupos? (*b*) Si las verdaderas proporciones de los usuarios fueran 48, 21 y 4%, respectivamente, ¿cuál sería el error estándar de la proporción p_{st} estimada? (*c*) ¿Cuál sería el error estándar de p derivado de una muestra aleatoria simple con $n = 400$?

5.15. La Fórmula (5.27) para la varianza mínima de $\bar{y}_{st}$ bajo la asignación de Neyman es la siguiente.

$$V_{mín}(\bar{y}_{st}) = \frac{(\sum W_h S_h)^2}{n} - \frac{\sum W_h S_h^2}{N}$$

Un estudiante comenta: "Dado que $\sum W_h S_h^2 > (\sum W_h S_h)^2$ a menos que todas las S_h sean iguales, la fórmula debe estar equivocada porque cuando n se aproxima a N dará un valor negativo para $V(\bar{y}_{st})$". ¿Quién tiene la razón, el estudiante o la fórmula?

5.16. Por la Fórmula (5.26) para la asignación de Neyman, la fracción de muestreo en el estrato h es $f_h = n_h/N_h = nS_h/N \sum W_j S_j$. Las situaciones en que la fórmula requiere más del 100% del muestreo en un estrato ($f_h > 1$) es de esperarse que sean aquellas en las que la fracción global del muestreo n/N sea importante y uno de los estratos presente una variabilidad excepcionalmente alta: El siguiente es un ejemplo para una población pequeña con $N = 100$, $n = 40$.

Estrato	N_h	S_h	Optimo n_h
1	60	2	15
2	30	4	15
3	10	15	10
	100		40

(*a*) Verificar que los óptimos n_h son como se muestran en la columna de la derecha. (*b*) calcular $V(\bar{y}_{st})$ por la Fórmula (5.6) y por la Fórmula (5.42) y mostrar que ambas dan $V(\bar{y}_{st}) = 0.12$.

CAPITULO 5 A

Otros Aspectos del Muestreo Estratificado

5A.1 EFECTOS DE LAS DESVIACIONES A PARTIR DE LA ASIGNACION OPTIMA

En este capítulo se discuten varios temas especiales del uso práctico del muestreo estratificado. Las Secs. 5A.1 a 5A.8, 5A.10 y 5A.15 tratan sobre problemas que pueden surgir en la planeación de una muestra; las secciones restantes tratan sobre las técnicas del análisis de los resultados. La presente sección considera la pérdida de precisión al no lograr la asignación óptima de la muestra.

Supóngase que se pretende usar la asignación óptima para un n dado. El tamaño de la muestra n'_h en el estrato h debería ser

$$n_h' = \frac{n(W_h S_h)}{\sum W_h S_h} \tag{5A.1}$$

A partir de la Ec. (5.27) la varianza mínima es

$$V_{mín}(\bar{y}_{st}) = \frac{1}{n}(\sum W_h S_h)^2 - \frac{1}{N}\sum W_h S_h^2 \tag{5A.2}$$

En la práctica, como no se conocen las S_h, sólo podemos aproximar esta asignación. Si $\hat{n}_h$ es el tamaño usado en el estrato h de la muestra, la varianza que de hecho se alcanza, a partir de la Ec. (5.6) es

$$V(\bar{y}_{st}) = \sum \frac{W_h^2 S_h^2}{\hat{n}_h} - \frac{1}{N}\sum W_h S_h^2 \tag{5A.3}$$

El aumento en varianza que origina la asignación imperfecta es

$$V(\bar{y}_{st}) - V_{mín}(\bar{y}_{st}) = \sum \frac{W_h^2 S_h^2}{\hat{n}_h} - \frac{1}{n}(\sum W_h S_h)^2 \tag{5A.4}$$

En el primer término de la derecha, sustitúyase $W_h S_h$ por su expresión en términos de n_h' a partir de (5A.1). Esto da el siguiente resultado interesante

$$V(\bar{y}_{st}) - V_{mín}(\bar{y}_{st}) = \frac{(\sum W_h S_h)^2}{n^2}\left(\sum \frac{n_h'^2}{\hat{n}_h} - n\right)$$

$$= \frac{(\sum W_h S_h)^2}{n^2} \sum \frac{(\hat{n}_h - n_h')^2}{\hat{n}_h} \qquad (5A.5)$$

Regresando a la Ec. (5A.2), si la cpf (último término a la derecha) es despreciable, tenemos que

$$\frac{V_{mín}(\bar{y}_{st})}{n} = \frac{(\sum W_h S_h)^2}{n^2} \qquad (5A.6)$$

Por lo tanto, el incremento proporcional en la varianza que resulta a partir de desviaciones de la asignación óptima es

$$\frac{V(\bar{y}_{st}) - V_{mín}(\bar{y}_{st})}{V_{mín}(\bar{y}_{st})} = \frac{1}{n} \sum_{h=1}^{L} \frac{(\hat{n}_h - n_h')^2}{\hat{n}_h} \qquad (5A.7)$$

donde $\hat{n}_h$ es el tamaño muestral real y n_h' el tamaño muestral óptimo en el estrato h. La cpf no es despreciable, entonces el signo de igualdal en (5A.7) se convierte en signo de desigualdad $\geq$.

Sea $g_h = |\hat{n}_h - n_h'|/\hat{n}_h$ la diferencia absoluta en los tamaños de la muestra en el estrato h expresada como una fracción del tamaño real n_h. Entonces (5A.7) viene a ser

$$\frac{V - V_{mín}}{V_{mín}} = \sum_{h=1}^{L} \frac{\hat{n}_h}{n} g_h^2 \qquad (5A.8)$$

una media ponderada de los g_h^2. Un límite superior conservador para $(V - V_{mín})/V_{mín}$ es, por tanto, g^2 donde g es la diferencia proporcional más grande en cualquier estrato. De modo que si $g = 0.2$ o 20%, el incremento proporcional en la varianza no podrá exceder $(0.2)^2$ o 4%. Si $g = 30\%$ el incremento proporcional en la varianza es cuando más 9%. En este sentido puede decirse que el óptimo es bajo.

Además, (5A.8) sugiere que el límite superior g^2 con frecuencia estimará en exceso el incremento proporcional en la varianza en una cantidad sustancial. La Tabla 5A.1 da un ejemplo con tres estratos para $n = 340$. La asignación óptima requiere tamaños de muestra de 200, 100 y 40 en tanto que los tamaños reales usados son 150, 120 y 70.

Como el valor de g es 0.43 (estrato 3), la regla aproximada da 18% como el incremento proporcional en la varianza. De la columna

TABLA 5A.1. EFECTOS DE LAS DESVIACIONES DE LA ASIGNACIÓN OPTIMA

Estrato	n_h' (ópt.)	$\hat{n}_h$ (real)	$\frac{\lvert\hat{n}_h - n_h'\rvert}{\hat{n}_h}$	$\frac{(\hat{n}_h - n_h')^2}{\hat{n}_h}$
1	200	150	0.33	16.7
2	100	120	0.17	3.3
3	40	70	0.43	12.9
Total	340	340	—	32.9

de la derecha, sabemos que el incremento real es de 32.9/340 = 9.7%.

Evans (1971) examinó la misma cuestión en términos de los efectos de los errores en las S_h estimadas y desarrolló una regla aproximada que muestra si un óptimo estimado es de esperarse que sea más preciso que una asignación proporcional. Este autor supone que el coeficiente de variación de la S_h estimada es el mismo en todos los estratos. Este supuesto es apropiado cuando las S_h fueron estimadas de una muestra preliminar del mismo tamaño en cada estrato. Evans nos dice cómo calcular el tamaño de una muestra preliminar requerida para hacer una asignación "óptima" en promedio mejor que la asignación proporcional. Sukhatme mostró anteriormente (1953) que una muestra inicial pequeña da usualmente una probabilidad elevada que la asignación "óptima" será superior a la proporcional. Ver Sukhatme y Sukhatme (1970), Pág. 87.

5A.2 EFECTOS DE ERRORES EN LOS TAMAÑOS DE LOS ESTRATOS

Para un tipo de estratificación deseada, los totales de estratos N_h pueden no conocerse con exactitud, al ser obtenidos de datos de censos anticuados. En vez de las verdaderas proporciones de estratos W_h, tenemos estimaciones w_h. La estimación de $\bar{Y}$ de la muestra es $\sum w_h \bar{y}_h$.

En términos generales, las consecuencias de usar ponderaciones que están en error son las siguientes:

1. La estimación de la muestra tiene *sesgo*. Debido al sesgo, medimos la exactitud de la estimación mediante su error cuadrático medio respecto a $\bar{Y}$, en lugar de medirla por su varianza respecto a su propia media (véase la Sec. 1.9).

2. El sesgo permanece constante conforme aumenta el tamaño de la muestra. Consecuentemente, siempre se alcanza un tamaño de

muestra para el cual la estimación es menos precisa que el muestreo aleatorio simple y se pierde toda ganancia en precisión debido a los estratos.

3. La estimación usual $s(\bar{y}_{st})$ subestima el verdadero error de $\bar{y}_{st}$, ya que no contiene la contribución del sesgo al error.

Para justificar estas declaraciones, nótese que en muestreo repetido, el valor medio de la estimación es $\sum w_h \bar{Y}_h$. El sesgo alcanza, por lo tanto, una cantidad igual a

$$\sum (w_h - W_h)\bar{Y}_h$$

Esto es independiente del tamaño de muestra. Para encontrar el error cuadrático medio (ECM) de la estimación, es fácil verificar que el término de la varianza está dado por la fórmula usual, con w_h en lugar de W_h. Por lo tanto,

$$\mathrm{ECM}(\bar{y}_{st}) = \sum \frac{w_h^2 S_h^2}{n_h}(1-f_h) + \left[\sum (w_h - W_h)\bar{Y}_h\right]^2 \qquad (5A.9)$$

Esta expresión fue dada por Stephan (1941). Finalmente, la fórmula usual para $s^2(\bar{y}_{st})$ es claramente una estimación sin sesgo del primer término en (5A.9) pero no toma en cuenta el segundo término.

Ejemplo. Este ilustra la pérdida en precisión debida a ponderaciones incorrectas cuando la estratificación es: (a) poco efectiva, (b) altamente efectiva. Considérese una población grande con $S^2 = 1$, divisible en dos estratos con $W_1 = 0.9$, $W_2 = 0.1$. Supongamos $S_1 = S_2 = S_h$. Entonces, despreciando los términos en $1/N_h$,

$$\begin{aligned} S^2 &\doteq \sum W_h S_h^2 + \sum W_h(\bar{Y}_h - \bar{Y})^2 \\ &= S_h^2 + W_1 W_2(\bar{Y}_1 - \bar{Y}_2)^2 \end{aligned} \qquad (5A.10)$$

esto es,

$$1 = S_h^2 + 0.09(\bar{Y}_1 - \bar{Y}_2)^2$$

En (a) tome $\bar{Y}_1 - \bar{Y}_2 = 1$. Entonces $S_h^2 = 0.91$ y la estratificación proporcional reduce la varianza en 9%, comparada con muestreo aleatorio simple.

En (b) tome $\bar{Y}_1 - \bar{Y}_2 = 3$, dando $S_h^2 = 0.19$, una reducción en varianza de más de 80%.

Con dos estratos y ponderaciones incorrectas el sesgo puede escribirse

$$(w_1 - W_1)(\bar{Y}_1 - \bar{Y}_2)$$

ya que $(w_1 - W_1) = -(w_2 - W_2)$. Suponga que las ponderaciones estimadas son $w_1 = 0.92$ y $w_2 = 0.08$. El sesgo alcanza $(0.02)(1) = 0.02$ en (a) y 0.06 en (b). Por lo tanto, tenemos los siguientes ECM comparables para una muestra de tamaño n:

Muestreo aleatorio simple: $V(\bar{y}) = \dfrac{1}{n}$

Muestreo aleatorio estratificado:

$$(a)\text{: } \mathrm{ECM}(y_{st}) = \frac{0.91}{n} + 0.0004$$

$$(b)\text{: } \mathrm{ECM}(y_{st}) = \frac{0.19}{n} + 0.0036$$

Como se ve en la Tabla 5A.2, con relación a (a), el muestreo aleatorio simple comienza a ganar con $n = 300$. Hay, sin embargo, poco qué escoger entre los dos métodos hasta $n = 1\,000$.

En (b), con más en juego, la estratificación es superior hasta $n = 200$, aunque la mayor parte de la ganancia potencial ya se ha perdido con este tamaño de muestra. Más allá de $n = 300$ la estratificación se vuelve marcadamente inferior al muestreo aleatorio simple. Una estimación exacta de las W_h es en particular importante cuando la estratificación es altamente efectiva o cuando el tamaño de muestra es grande.

En algunas encuestas puede tomarse una muestra preliminar grande de tamaño n' para estimar W_h. Esta técnica, conocida como *muestreo doble*, o *muestreo en dos fases*, tiene numerosas aplicaciones y se discute en el Cap. 12. Se mostrará que con muestreo doble el error cuadrático medio de $\bar{y}_{st}$ es aproximadamente

$$\frac{\sum W_h S_h^2}{n} + \frac{\sum W_h(\bar{Y}_h - \bar{Y})^2}{n'} \qquad (5A.11)$$

Comparando este ECM con S^2/n, como lo da la Ec. (5A.10), vemos que la mayor parte de la ganancia de la estratificación se retiene, siempre que n' sea mucho más grande que n. Para ponerlo en forma más general, un grupo de ponderaciones estimadas preserva la mayor parte de la ganancia potencial con estratificación, si las

TABLA 5A.2. VALORES COMPARABLES DEL ECM($\bar{y}$)

		Aleatorio estratificado	
n	Aleatorio simple	(a)	(b)
50	0.0200	0.0186	0.0074
100	0.0100	0.0095	0.0055
200	0.0050	0.0049	0.0045
300	0.0033	0.0034	0.0042
400	0.0025	0.0027	0.0041
1000	0.0010	0.0013	0.0038

ponderaciones son más exactamente estimadas de lo que lo serían con una muestra aleatoria de tamaño n.

5A.3 EL PROBLEMA DE LA ASIGNACION CON MAS DE UNA CARACTERISTICA

Dado que generalmente, la mejor asignación para una característica no será la mejor para otra, debe llegarse a algún compromiso en un reconocimiento con características numerosas. El primer paso es reducir las características consideradas en la asignación a un número relativamente pequeño donde se consideran las más importantes. Si se dispone de buenos datos anteriores, podemos entonces calcular la asignación óptima para cada característica separadamente y ver hasta qué punto existe desacuerdo. En una encuesta de un tipo especializado las correlaciones entre las características pueden ser altas y las asignaciones pueden diferir relativamente poco.

Ejemplo. Los datos dados por Jessen (1942) ilustran una encuesta de este tipo con fincas. El Estado de Iowa se dividió en cinco regiones geográficas, cada una designada de acuerdo con su principal actividad agrícola. Supóngase que estas regiones serán usadas como estratos en una encuesta de explotaciones lecheras. Las tres características de mayor interés son el número de vacas ordeñadas por día, el número de galones de leche por día y las entradas totales anuales en efectivo por productos lácteos. De un reconocimiento hecho en 1938, las desviaciones estándar estimadas s_h dentro de estratos se muestran en la Tabla 5A.3. En la Tabla 5A.4 se dan las asignaciones óptimas de Neyman, basadas en estas s_h para las características individuales en una muestra de 1 000 fincas.

Las asignaciones óptimas individuales difieren una de la otra sólo moderadamente. Con una excepción, las tres se desvían en la misma dirección en comparación con la asignación proporcional. Así, en el primer estrato, la asignación proporcional sugiere 197 fincas y las asignaciones individuales conducen a números entre 236 y 258. El promedio de los tamaños de muestra

TABLA 5A.3. DESVIACIONES ESTÁNDAR DENTRO DE ESTRATOS

Estrato	$W_h = \frac{N_h}{N}$	s_h Vacas ordeñadas	s_h Galones de leche	Entradas por productos lácteos ($)
Lechería noreste	0.197	4.6	11.7	332
Existencias de grano	0.191	3.4	9.8	357
Ganado occidental	0.219	3.3	7.0	246
Pastizal del sur	0.184	2.8	6.5	173
Ganado oriental	0.208	3.7	9.8	279

TABLA 5A.4. TAMAÑOS DE MUESTRA DENTRO DE ESTRATOS ($n = 1\,000$)

	Asignación				
		Optima para			
Estrato	Proporcional	Vacas	Galones	Entradas	Promedio m_h
Lechería noreste	197	254	258	236	250
Existencias de grano	191	182	209	246	212
Ganado occidental	219	203	171	194	189
Pastizal del sur	184	145	134	115	131
Ganado oriental	208	216	228	209	218

TABLA 5A.5. VARIANZAS ESPERADAS DE LAS MEDIAS ESTIMADAS

Tipo de asignación	Vacas	Galones	Entradas
Optima	0.0127	0.0800	76.9
Compromiso	0.0128	0.0802	77.6
Proporcional	0.0131	0.0837	80.9

óptimos para las tres características, mostrado en la columna de la derecha, provee una asignación de compromiso satisfactoria.

La Tabla 5A.5 muestra las varianzas de muestreo esperadas de y_{st}, como las dan las asignaciones óptimas individuales, la de compromiso y la proporcional. Las fórmulas son:

$$v_{opt} = \frac{(\sum W_h s_h)^2}{n}, \qquad v_{comp} = \sum \frac{(W_h S_h)^2}{m_h}, \qquad v_{prop} = \frac{\sum W_h s_h^2}{n}$$

La asignación de compromiso da resultados casi tan precisos como cuando fuese posible unas asignaciones óptimas separadas para cada ítem. Más notable aún es el hecho de que la asignación proporcional es tan sólo un poco menos precisa que la de compromiso o los óptimos individuales. Además la Tabla 5A.5 sobreestima la precisión de las asignaciones de óptimas y de compromiso porque éstos fueron a partir de varianzas estimadas. Este resultado es otra ilustración de que el óptimo es plano, lo que se mencionó en la Sec. 5A.1.

5A.4 OTROS METODOS DE ASIGNACION CON MAS DE UN ATRIBUTO

Una asignación alternativa de compromiso es la que sugiere Chatterjee (1967), en ella se eligen los n_h que minimizan el promedio de los incrementos proporcionales en varianza a partir de

(5A.7) al tomarlos sobre las variables. Si j denota una variable, esto nos lleva a elegir

$$n_h = n\sqrt{\sum_j n_{jh}'^2}\Big/\sum_h \sqrt{\sum_j n_{jh}'^2} \qquad (5A.12)$$

donde n'_{jh} es el tamaño óptimo de la muestra en el estrato h para la variable j. Para los datos de la Tabla 5A.4, donde los óptimos individuales sólo difieren poco, los n_h de Chatterjee varían a partir del promedio m_h de la Tabla 5A.4 en cuando mucho en una unidad en cualquier estrato.

En algunas encuestas, las asignaciones óptimas para variables individuales difieren tanto, que no hay una solución de compromiso obvia. Se necesita algún principio para determinar la asignación a usar y Yates (1960) sugirió dos de ellos, que veremos a continuación.

El primero se aplica a encuestas con un objetivo específico en donde la pérdida debida a un error de una magnitud dada en una estimación, se puede medir en términos de dinero o utilidad, como se vio en la Sec. 4.10. Con k variables y funciones cuadráticas de pérdida, es razonable expresar la pérdida esperada total, como una función lineal de las varianzas de las medias o totales estimadas de población. Para las medias

$$L = \sum_j^k a_j V(\bar{y}_{jst}) = \sum_j^k a_j \sum_h^L W_h^2 S_{jh}^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \qquad (5A.13)$$

donde S^2_{jh} es la varianza de la j-ésima variable en el estrato h. El intercambio del orden de la suma da

$$L = \sum_h \frac{W_h^2}{n_h}\left(\sum_j a_j S_{jh}^2\right) - \frac{1}{N}\sum W_h\left(\sum_j a_j S_{jh}^2\right) \qquad (5A.14)$$

Con una función lineal para los costos de muestreo, tenemos

$$C = c_0 + \sum c_h n_h \qquad (5A.15)$$

Al minimizar el producto de $(C - c_0)$ y el primer término en L (el término que depende de n_h) llegamos a,

$$n_h \propto \frac{W_h}{\sqrt{c_h}}\sqrt{\sum_j a_j S_{jh}^2} \qquad (5A.16)$$

por la desigualdad de Cauchy-Schwarz. La constante de proporcionalidad se encuentra al satisfacer la restricción dada para L o C. Por

ejemplo supóngase que el valor de L se ha especificado y que el término cpf puede ignorarse. Tenemos

$$n_h = \frac{n(W_h A_h/\sqrt{c_h})}{\sum (W_h A_h/\sqrt{c_h})} \qquad (5A.17)$$

donde $A_h = \sqrt{\sum_j a_j S_{jh}^2}$. El tamaño de la muestra total es, por (5A.14)

$$n = \frac{1}{L}\left(\sum_h \frac{W_h A_h}{\sqrt{c_h}}\right)\left(\sum_h W_h A_h \sqrt{c_h}\right) \qquad (5A.18)$$

En la segunda sugerencia se especifica la varianza deseada de V_j para *cada* variable. Para medias de población esto significa que

$$\sum_{h=1}^{L} \frac{W_h^2 S_{jh}^2}{n_h} - \sum_{h=1}^{L} \frac{W_h S_{jh}^2}{N} \leq V_j \qquad (j = 1, 2, \ldots, k) \qquad (5A.19)$$

Se usan signos de desigualdad porque la asignación más económica puede proporcionar varianzas más pequeñas que las V_j para algunas características.

En esta sugerencia, el costo C [Ec. (5A.15)] se minimiza de acuerdo a las tolerancias V_j y las condiciones $0 \leq n_h \leq N_h$. El problema es de programación no lineal. Hartley y Hocking (1963), Chatterjee (1966), Zukhovitsky y Avdeyeva (1966), Huddleston y otros (1970) han dado logaritmos para su solución. Anteriormente Dalenius (1957) dio una solución gráfica ingeniosa, mientras que Yates (1960) y Kokan (1963) desarrollaron métodos de aproximación sucesiva ilustrados en la segunda edición de este libro.

Un primer paso útil es encontrar la asignación óptima separadamente para cada variable así como el costo que satisface su tolerancia. Tómese la variable, digamos y_1 para la cual el costo C_1 es el más elevado, y examínese si los valores óptimos n_h para y_1 satisfacen todas las otras $(k - 1)$ tolerancias. De ser así, se usa esta asignación y el problema está resuelto, ya que ninguna otra asignación, satisfará la tolerancia V_1 para y_1 a un costo tan bajo como C_1.

Al estudiar una serie de ejemplos en un problema relacionado, Booth y Sedransk (1969) han señalado que, a falta de un programa de cálculo, se puede obtener una buena solución al segundo problema de Yates, al resolver el primer problema que es más fácil. Debemos especificar que L en (5A.13) tendrá el valor $V^* = \sum a_j V_j$ donde las V_j son las tolerancias individuales deseadas y las a_j se hacen inversamente proporcionales a las V_j. Así, con dos variables $a_1 = V_2/(V_1 + V_2)$, $a_2 = V_1/(V_1 + V_2)$ y

$$V^* = \frac{2V_1V_2}{(V_1 + V_2)} \qquad (5A.20)$$

Ejemplo. (Para cuatro estratos, dos variables.) Los datos y la aplicación aproximada del método se demuestran en las columnas (1) a (6) de la Tabla 5A.6. El problema es encontrar el n más pequeño para que

$$V(\bar{y}_{1st}) \leqq 0.04, \qquad V(\bar{y}_{2st}) \leq 0.01$$

TABLA 5A.6. DATOS ARTIFICIALES PARA CUATRO ESTRATOS, DOS VARIABLES

Columna	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Estrato	W_h	S_{1h}^2	S_{2h}^2	$A_h^2 = \sum_j a_j S_{jh}^2$	$W_h A_h$	n_h	$\tilde{n}_h$
1	0.4	25	1	5.8	0.963	206	194
2	0.3	25	4	8.2	0.859	183	180
3	0.2	25	16	17.8	0.844	180	187
4	0.1	25	64	56.2	0.750	160	171
				Totales	3.416	729	732

Al encontrar la asignación óptima para cada variable en forma independiente, fácilmente se verifica que $n = 625$ es el requerido para satisfacer la primera restricción, y $n = 676$ para satisfacer la segunda. Sin embargo, $n = 676$ con su asignación no satisface la primera restricción al obtenerse 0.0589 en lugar de 0.04 para V_1. Una solución de iteración para satisfacer ambas restricciones (que presentamos en la segunda edición) dio $\tilde{n} = 732$, con el $\tilde{n}_h$ que aparece en la columna (7) de la Tabla 5A.6.

Para usar el método de Booth y Sedransk con $V_1 = 0.04$ y $V_2 = 0.01$, especificamos

$$L = 0.2\,V(\bar{y}_{1st}) + 0.8\,V(\bar{y}_{2st}) = \frac{2(0.04)(0.01)}{(0.05)} = 0.016$$

A partir de (5A.18) con $c_h = 1$, tenemos

$$n = \frac{\left(\sum W_h A_h\right)^2}{L} = \frac{(3.416)^2}{0.016} = 729$$

Al usar la columna (5) de la Tabla 5A.6. De (5A.17), la columna (5) también conduce a los valores n_h que aparecen en la columna (6) de la Tabla 5A.6. Como se ve en las columnas (6) y (7), las dos soluciones n_h y $\tilde{n}_h$ concuerdan satisfactoriamente.

Como lo hacen notar Booth y Sedransk $n \leq \tilde{n}$ en todos los problemas de este tipo, ya que n satisface la restricción $L - V^*$, pero no es necesario que satisfaga la restricción sobre cada variable, en tanto que la asignación $\tilde{n}$ satisface L, al igual que las restricciones individuales.

5A.5 ESTRATIFICACION EN DOS DIRECCIONES, CON MUESTRAS PEQUEÑAS

Supongamos que hay dos criterios de estratificación, digamos por R filas y C columnas, lo que da RC celdas. Si $n \geq RC$ cada celda se puede representar en la muestra. Un problema surge cuando $n < RC$ y desearíamos que la muestra nos diera la representación proporcional de cada criterio de estratificación. En un método simple desarrollado por Bryant, Hartley y Jessen (1960), la técnica sólo requiere que n exceda al más grande de R y C.

Para ilustrar este método, supongamos que se estratificó una pequeña población de 165 escuelas, según el tamaño de la localidad, en cinco clases y, según el costo promedio por alumno, en cuatro clases. Los números de escuelas m_{ij} y las proporciones de escuelas $P_{ij} = m_{ij}/165$ en cada una de las 20 celdas se muestran en la Tabla 5A.7.

TABLA 5A.7. NÚMERO Y PROPORCIÓN DE ESCUELAS EN CADA CELDA

Tamaño de la ciudad		Gasto por alumno A	B	C	D	Totales		$n_{i.}$
I	m_{1j}	15	21	17	9	$m_{1.}$	62	
	P_{1j}	0.091	0.127	0.103	0.055	$P_{1.}$	0.376	4
II	m_{2j}	10	8	13	7	$m_{2.}$	38	
	P_{2j}	0.061	0.049	0.079	0.042	$P_{2.}$	0.231	2
III	m_{3j}	6	9	5	8	$m_{3.}$	28	
	P_{3j}	0.036	0.055	0.030	0.049	$P_{3.}$	0.170	2
IV	m_{4j}	4	3	6	6	$m_{4.}$	19	
	P_{4j}	0.024	0.018	0.036	0.036	$P_{4.}$	0.114	1
V	m_{5j}	3	2	5	8	$m_{5.}$	18	
	P_{5j}	0.018	0.012	0.030	0.049	$P_{5.}$	0.109	1
Totales	$m_{.j}$	38	43	46	38		165	
	$P_{.j}$	0.230	0.261	0.278	0.231		1.000	
	$n_{.j}$	2	3	3	2			

El objetivo de dar a cada escuela una oportunidad igual de selección, en tanto que se da a cada clase marginal su representación proporcional. En esta ilustración $n = 10$. Calcule los números $n_{i.} = nP_{i.}$ y $n_{.j} = nP_{.j}$, donde estos productos se redondean al entero más próximo (con un ajuste menor posterior, si se necesita, de tal manera que los $n_{i.}$ y los $n_{.j}$ ambas sumen n). Estos números se muestran en la Tabla 5A.7.

El siguiente paso es extraer $n = 10$ celdas con probabilidad $n_{i.}n_{.j}/n^2$ para la ij-ésima celda. Esto se hace al construir un cuadrado $n \times n$ (Tabla 5A.8). En la hilera 1 se extrae una columna al azar. En la hilera 2 se extrae al azar una de las columnas restantes, y así sucesivamente. Al final, cada hilera y cada columna contienen una unidad. (Esta extracción se hace más rápidamente mediante una permutación al azar de los números 1 a 10.) Los resultados de una extracción se indican con $\times$ en la Tabla 5A.8.

TABLA 5A.8. CUADRADO 10×10 PARA EXTRAER LA MUESTRA

		Columna									
Hilera		1	2	3	4	5	6	7	8	9	10
		A		*B*			*C*			*D*	
1	I	×									
2					×						
3			×								
4								×			
5	II						×				
6									×		
7	III			×							
8										×	
9	IV					×					
10	V										×

Nótese que las columnas 1 y 2 se asignan al estrato marginal A ya que $n_1 = 2$. Similarmente, las hileras 1 hasta 4 se asignan al estado marginal I, ya que $n_1 = 4$ y así las demás. Esto completa la asignación de la muestra a las 20 celdas. La asignación aparece en una forma más compacta en la Tabla 5A.9. Dos escuelas se extraen al azar de

TABLA 5A.9. ASIGNACIÓN DE LA MUESTRA A LAS 20 CELDAS

	A	*B*	*C*	*D*	Total
I	2	1	1	0	4
II	0	0	2	0	2
III	0	1	0	1	2
IV	0	1	0	0	1
V	0	0	0	1	1
Total	2	3	3	2	10

entre las 15 escuelas en la celda I*A* y así las demás. La probabilidad de que se extraiga una escuela en la hilera i, columna j es proporcional a $n_{i.}n_{.j}/P_{ij}$. Así, las probabilidades no son iguales, aunque lo serán aproximadamente si $P_{ij} \doteq n_{i.}n_{.j}/n^2$.

Una estimación sin sesgo de la media por escuela es

$$\bar{y}_U = \frac{1}{n}\sum \frac{n^2 P_{ij}}{n_{i.}n_{.j}} y_{ij}$$

donde y_{ij} es el *total* de la muestra en la ij-ésima celda. Sin embargo, si $P_{ij} \doteq n_{i.}n_{.j}/n^2$, la media de la muestra $\bar{y}$ es probablemente preferible, ya que su sesgo debería ser despreciable. Una estimación de la varianza de la muestra está disponible tanto para la estimación sesgada como para la no sesgada, puesto que n es cuando menos dos veces el más grande de R y C y cuando menos se extraen dos unidades en cada hilera y en cada columna.

Si en algunas celdas P_{ij} *difiere* marcadamente de $n_{i.}n_{.j}/n^2$, un paso extra conserva las probabilidades de selección de las escuelas casi constantes. Después de calcular las $n_{i.}$ y $n_{.j}$, examine las cantidades $D_{ij} = nP_{ij} - n_{i.}n_{.j}/n$, después de redondearlas a enteros. Si en cualquier celda, D_{ij} es un entero positivo, automáticamente asigne D_{ij} unidades a dicha celda. Reduzca n, los n_i, los n_j en las cantidades requeridas par esta asignación fija y siga con la asignación restante como antes.

5A.6 SELECCION CONTROLADA

Goodman y Kish (1950) nos dan otra técnica para este problema con muestras pequeñas, llamada *selección controlada*. Existe una ilustración simple debida a Hess, Riedel y Fitzpatrick (1966) que aplicaron el método a un muestreo de hospitales, donde se ilustra la idea básica. La estratificación principal es por el tamaño del hos-

pital (dos estratos). También es deseable la representación de dos tipos de propietarios de hospital, pero sólo se va a extraer una unidad (hospital) de cada estrato principal. Al numerar las unidades dentro de los estratos se usa una prima (′) que indica un tipo de propiedad, la ausencia de este símbolo indica el otro tipo de propiedad. La Tabla 5A.10 (lado izquierdo) muestra las unidades en ambos estratos.

TABLA 5A.10. ORDENAMIENTO DE LAS UNIDADES DENTRO DE LOS ESTRATOS PARA SELECCIÓN CONTROLADA

Estrato del orden original		Estrato del orden revisado	
I Gran hospital	II Pequeño hospital	Gran hospital	Pequeño hospital
1	1	1	3′
2	2	2	4′
3′	3′	3′	5′
4′	4′	4′	1
	5′		2

Si se extrae la unidad 1 o 2 del estrato I, desearíamos extraer las unidades 3′, 4′ o 5′ del estrato II, de modo que ambos tipos de estratificación se encuentren presentes con $n = 2$. De manera semejante, 3′ o 4′ (estrato I) es la unidad deseada con 1 o 2 (estrato II). La selección controlada hace tan alta como es matemáticamente posible la probabilidad de estas combinaciones deseadas, al conservar la misma probabilidad de selección dentro de los estratos y, por tanto, las estimaciones insesgadas por medio de las fórmulas usuales para el muestreo estratificado. El propósito es una mayor exactitud para un n dado o bien, un ahorro en costos de campo.

Con el muestreo aleatorio estratificado, la probabilidad de una combinación deseada es $(.5)(.6) + (.5)(.4) = .5$. Esta probabilidad puede incrementarse a .9 por dos cambios simples en la selección de la muestra. Se reacomodan las unidades del estrato II, de modo que las combinaciones deseadas (3′, 4′, 5′) con 1 y 2 en el estrato 1 aparezcan primero, como se ve a la derecha de la Tabla 5A.10. Posteriormente se extrae un número aleatorio r entre 1 y 100 y se usa para seleccionar las unidades de *ambos* estratos. En el estrato I, $1 \leq r \leq 25$ selecciona la unidad 2, etc., de modo que cada unidad recibe un cuarto de probabilidad de ser elegida. De manera semejante, en el estrato II, la condición $1 \leq r \leq 20$, selecciona 3′, la condición $21 \leq r \leq 40$ selecciona 4′, y así sucesivamente. De modo que $1 \leq r \leq 20$ selecciona (1,3′), $20 \leq r \leq 25$ selecciona (1,4′) y así sucesivamente. Las selecciones conjuntas y sus probabilidades son como sigue.

Par:	(1, 3′)	(1, 4′)	(2, 4′)	(2, 5′)	(3′, 5′)	(3′, 1)	(4′, 1)	(4′, 2)
Probabilidad	.20	.05	.15	.10	.10	.15	.05	.20

La única combinación no deseada es (3′, 5′). Por tanto, la probabilidad total de las combinaciones deseadas es .90.

Como el muestreo no es independiente en los dos estratos, las fórmulas para $V(\bar{y}_{st})$ y $v(\bar{y}_{st})$ no se aplican. Hess, Riedel y Fitzpatrick (1976), dan fórmulas aproximadas. En esta monografía, también se da un algoritmo para la aplicación de la selección controlada en problemas con más estratos, con un n mayor, y con controles más complejos.

Para otro método que usa diseños por bloques incomplejos balanceados, véase Avadhani y Sukhatme (1973).

5A.7 LA CONSTRUCCION DE LOS ESTRATOS

Este asunto presenta varios problemas. ¿Cuál es la mejor característica para la construcción de los estratos? ¿Cómo se determinan los límites entre estratos? ¿Cuántos estratos debería haber? Para una sola característica o variable y, la mejor característica es, por supuesto, la distribución de frecuencia de y. La siguiente mejor es probablemente la distribución de frecuencia de alguna otra cantidad altamente correlacionada con y. Dado el número de estratos, las ecuaciones para determinar los mejores límites entre ellos bajo asignación proporcional y de Neyman, han sido obtenidas por Dalenius (1957), otros investigadores han encontrado algunos métodos de aproximación más rápidos. Consideraremos la asignación de Neyman, puesto que generalmente es superior a la asignación proporcional en poblaciones donde se gana más al estratificar.

Se supone primero que los estratos se establecen al usar el valor en sí de y.

Sean y_0, y_L el valor menor y mayor de y en la población. El problema es encontrar límites intermedios entre estratos $y_1, y_2, \ldots, y_{L-1}$ tales que

$$V(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \qquad (5A.21)$$

sea un mínimo. Si se ignora la cpf, es suficiente con minimizar $\sum W_h S_h$. Dado que y_h aparece en esta suma solamente en los términos $W_h S_h$ y $W_{h+1} S_{h+1}$, tenemos

$$\frac{\partial}{\partial y_h}(\sum W_h S_h) = \frac{\partial}{\partial y_h}(W_h S_h) + \frac{\partial}{\partial y_h}(W_{h+1} S_{h+1})$$

Ahora, si $f(y)$ es la función de frecuencia de y,

$$W_h = \int_{y_{h-1}}^{y_h} f(t)\, dt, \qquad \frac{\partial W_h}{\partial y_h} = f(y_h) \tag{5A.22}$$

Además,

$$W_h S_h^2 = \int_{y_{h-1}}^{y_h} t^2 f(t)\, dt - \frac{\left[\int_{y_{h-1}}^{y_h} t f(t)\, dt\right]^2}{\int_{y_{h-1}}^{y_h} f(t)\, dt} \tag{5A.23}$$

Al derivar (5A.23) obtenemos

$$S_h^2 \frac{\partial W_h}{\partial y_h} + 2 W_h S_h \frac{\partial S_h}{\partial y_h} = y_h^2 f(y_h) - 2 y_h \mu_h f(y_h) + \mu_h^2 f(y_h) \tag{5A.24}$$

donde μ_h es la media de y en el estrato h. Súmese $S_h^2 \partial W_h / \partial y_h$ al lado izquierdo y la cantidad igual $S_h^2 f(y_h)$ al lado derecho. Esto da, al dividir por $2S_h$

$$\frac{\partial (W_h S_h)}{\partial y_h} = S_h \frac{\partial W_h}{\partial y_h} + W_h \frac{\partial S_h}{\partial y_h} = \frac{1}{2} f(y_h) \frac{(y_h - \mu_h)^2 + S_h^2}{S_h} \tag{5A.25}$$

Similarmente encontramos

$$\frac{\partial (W_{h+1} S_{h+1})}{\partial y_h} = -\frac{1}{2} f(y_h) \frac{(y_h - \mu_{h+1})^2 + S_{h+1}^2}{S_{h+1}} \tag{5A.26}$$

Por lo tanto, las ecuaciones de cálculo para y_h son

$$\frac{(y_h - \mu_h)^2 + S_h^2}{S_h} = \frac{(y_h - \mu_{h+1})^2 + S_{h+1}^2}{S_{h+1}} \quad (h = 1, 2, \ldots, L-1) \tag{5A.27}$$

Desafortunadamente, estas ecuaciones están mal adaptadas para cálculos prácticos, ya que tanto μ_h como S_h dependen de y_h. Un método rápido aproximado, debido a Dalenius y Hodges (1959) se presenta para minimizar $\sum W_h S_h$. Sea

$$Z(y) = \int_{y_0}^{y} \sqrt{f(t)}\, dt \tag{5A.28}$$

Si los estratos son numerosos y estrechos, $f(y)$ debería ser aproximadamente constante (rectangular) dentro de un estrato dado. Por lo tanto.

$$W_h = \int_{y_{h-1}}^{y_h} f(t)\, dt \doteq f_h(y_h - y_{h-1}) \tag{5A.29}$$

$$S_h \doteq \frac{1}{\sqrt{12}}(y_h - y_{h-1}) \tag{5A.30}$$

$$Z_h - Z_{h-1} = \int_{y_{h-1}}^{y_h} \sqrt{f(t)}\, dt \doteq \sqrt{f_h}(y_h - y_{h-1}) \tag{5A.31}$$

donde f_h es el valor "constante" de $f(y)$ en el estrato h. Al sustituir estas aproximaciones, encontramos

$$\sqrt{12} \sum_{h=1}^{L} W_h S_h \doteq \sum_{h=1}^{L} f_h (y_h - y_{h-1})^2 \doteq \sum_{h=1}^{L} (Z_h - Z_{h-1})^2 \tag{5A.32}$$

Ya que $(Z_L - Z_0)$ es fija, es fácil verificar que la suma en la derecha se minimiza al hacer $(Z_h - Z_{h-1})$ constante.

Dada $f(y)$, la regla es formar la cumulativa de $\sqrt{f(y)}$ y escoger las y_h de tal manera que éstas formen intervalos iguales en la escala cum $\sqrt{f(y)}$. La Tabla 5A.11 ilustra el uso de esta regla.

Ejemplo. Los datos muestran la distribución de frecuencia del porcentaje de préstamos bancarios dedicados a préstamos industriales en una

TABLA 5A.11. CÁLCULO DE LOS LÍMITES DE ESTRATO CON LA REGLA CUM $\sqrt{f(y)}$

Préstamos industriales / Préstamos totales %	$f(y)$	Cum $\sqrt{f(y)}$	Préstamos industriales / Préstamos totales %	$f(y)$	Cum $\sqrt{f(y)}$
0–5	3464	58.9	50–55	126	340.3
5–10	2516	109.1	55–60	107	350.6
10–15	2157	155.5	60–65	82	359.7
15–20	1581	195.3	65–70	50	366.8
20–25	1142	229.1	70–75	39	373.0
25–30	746	256.4	75–80	25	378.0
30–35	512	279.0	80–85	16	382.0
35–40	376	298.4	85–90	19	386.4
40–45	265	314.7	90–95	2	387.8
45–50	207	329.1	95–100	3	389.5

población de 13 435 bancos de los Estadose Unidos de Norteamérica (McEvoy, 1956). La distribución es oblicua, con su moda en el extremo inferior. En la columna cum $\sqrt{f}$, $58.9 = \sqrt{3\,436}$, $109.1 = \sqrt{3\,464} + \sqrt{2\,516}$, y así sucesivamente.

Supóngase que se quieren cinco estratos. Dado que el total de cum $\sqrt{f}$ es 389.5, los puntos de división deberían estar a 77.9, 155.8, 233.7 y 311.6 en esta escala. Los puntos disponibles más cercanos son los siguientes:

	Estrato				
	1	2	3	4	5
Límites	0–5%	5–15%	15–25%	25–45%	45–100%
Intervalo en cum $\sqrt{f}$	58.9	96.6	73.6	85.6	74.8

Los dos primeros intervalos, 58.9 y 96.6, son bastante desiguales, pero no se pueden mejorar sin una subdivisión más fina de las clases originales.

Si los intervalos de clase en la distribución original de y son de longitudes desiguales, se necesita un ligero cambio. Cuando el intervalo cambia de uno de longitud d a uno de longitud ud el valor de $\sqrt{f}$ para el segundo intervalo se multiplica por $\sqrt{u}$ al formar cum $\sqrt{f}$.

Otro método propuesto por Sethi (1963) consiste en encontrar los límites dados por las ecuaciones de cálculo (5A.27) para una distribución continua estándar, parecida a la población de estudio. Para las distribuciones normales y varias distribuciones χ^2 Sethi tabuló los límites óptimos para asignaciones de Neyman, iguales y proporcionales para $L \leq 6$. Si una de estas distribuciones parece aproximarse a la población de estudio, los límites se pueden encontrar en las tablas de Sethi.

Otros dos métodos aproximados requieren algo de prueba y error. A partir de las relaciones (5A.32) la regla de Dalenius-Hodges es más o menos equivalente a hacer W_hS_h constante, como lo conjeturó antes Dalenius y Gurney (1951). Una regla semejante es la de Ekman (1959), en la que $W_h(y_h - y_{h-1})$ se hace igual a una constante.

Al comparar ciertas poblaciones teóricas y ocho poblaciones de estudio, Cochran (1961) encontró que la regla cum $\sqrt{f}$ y la regla de Ekman tenían resultados consistentemente buenos (no intentó el método de Sethi). En un estudio sobre capacidad (camas) de hospitales en los Estados Unidos, cuya distribución se asemeja a una distribución χ^2 con un grado de libertad, Hess, Sethi y Balakrishnan

(1966) encontraron que el método de Ekman es ligeramente superior al método cum $\sqrt{f}$ y al de Sethi para $L > 2$, también Murthy (1967) obtiene buenos resultados con el método de Ekman.

Las relaciones (5A. 32) tienen una consecuencia interesante. Si W_hS_h es constante, la asignación de Neyman, da un tamaño de muestra *constante* $n_h = n/L$ en todos los estratos. Para los métodos aproximados, las comparaciones realizadas sugieren que la regla simple $n_h = n/L$ es satisfactoria.

Hasta aquí, hemos supuesto en forma poco realista que la estratificación se puede basar en los valores de y. En la práctica, se utiliza alguna otra variable x (tal vez el valor de y en un censo reciente). Dalenius (1957) desarrolla ecuaciones para límites de x que minimizan $\sum W_hS_{yh}$, siempre y cuando se tenga cierto conocimiento de la regresión de y sobre x. Si esta regresión es no lineal, estos límites pueden diferir considerablemente de los óptimos, cuando x es la variable que se mide. Sin embargo, las ecuaciones significan que si la regresión de y sobre x es lineal, y la correlación entre y y x es alta, dentro de todos los estratos, los dos conjuntos de límites deberían ser casi los mismos. Sea

$$y = \alpha + \beta x + e$$

donde $E(e) = 0$ para todo x y e, las x no son correlacionadas. La varianza de e dentro del estrato h es S_{eh}^2. Entonces, los límites de x que minimizan $\bar{V}(y_{st})$ satisfacen las ecuaciones (Dalenius 1957).

$$\frac{\beta^2[(x_h-\mu_{xh})^2+S_{xh}^2]+2S_{eh}^2}{\beta S_{xh}\sqrt{1+S_{eh}^2/\beta^2 S_{xh}^2}} = \frac{\beta^2[(x_h-\mu_{x,h+1})^2+S_{x,h+1}^2]+2S_{e,h+1}^2}{\beta S_{x,h+1}\sqrt{1+S_{e,h+1}^2/\beta^2 S_{x,h+1}^2}}$$

Si $S_{eh}^2/\beta^2 S_{xh}^2$ es pequeño para todo h, estas ecuaciones se reducen a la forma (5A.27) que da los limites óptimos para x. Pero $S_{eh}^2/\beta^2 S_{xh}^2 = (1-\rho_h^2)/\rho_h^2$ donde ρ_h es la correlación entre y y x dentro del estrato h.

Aunque es necesario continuar la investigación, este resultado sugiere que la regla de cum $\sqrt{f}$ aplicada a x debería dar una estratificación eficiente para otra variable y que tiene una regresión lineal de x con correlación elevada. Algunos resultados numéricos por Cochran (1961) corroboran esta conjetura. Además, si las ρ_h son apenas moderadas, como sucederá cuando aumenta el número de estratos, la falla, al usar los límites óptimos de x, debería tener un efecto menor en y.

La discusión anterior es desde luego, principalmente pertinente al muestreo de instituciones estratificadas por alguna medida de tamaño. La situación es diferente cuando un grupo de variables está

estrechamente relacionado con y_1 y otro grupo está relacionado con y_2 con una distribución de frecuencia marcadamente distinta. Una posibilidad es buscar límites de estratos de compromiso que satisfagan las tolerancias deseables en $V(\bar{y}_{1st})$ y $V(\bar{y}_{2st})$, siguiendo el enfoque general dado en la Sec. 5A.4, pero no se han desarrollado los métodos de cálculo.

En la estratificación geográfica el problema es menos tratable matemáticamente porque existen muchas maneras diferentes de formar los límites de estratos. El procedimiento usual es estimar algunas variables de alta correlación con las características principales de la encuesta y usar una combinación de sentido común y prueba y error para construir buenos límites para estas variables seleccionadas. Dado que las ganancias en precisión con la estratificación es probable que sean modestas, no vale la pena hacer un gran esfuerzo para mejorar los límites. Las bases de la estratificación para ítems económicos han sido discutidas por Stephan (1941) y Hagood y Bernert (1945) y para ítems agrarios por King y McCarty (1941).

5A.8 NUMERO DE ESTRATOS

Las dos preguntas pertinentes a una decisión sobre el número de estratos L son (a) ¿en qué proporción decrece la varianza de $\bar{y}_{st}$ conforme aumenta L? (b) ¿Cómo se afecta el costo del reconocimiento al aumentar L?

Con respecto a: (a), supóngase primero que los estratos se construyen con los valores de y. Para tomar el caso más simple sea rectangular la distribución de y en el intervalo $(a, a + d)$. Entonces, S_y^2, antes de la estratificación, es $d^2/12$, de tal manera que con una muestra aleatoria simple de tamaño n, $V(\bar{y}) = d^2/12n$. Si se crean L estratos de igual tamaño, la varianza dentro de cualquier estrato es $S_{yh}^2 = d^2/12L^2$. Por lo tanto, para una muestra estratificada, con $W_h = 1/L$ y $n_h = n/L$,

$$V(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_{yh}\right)^2 = \frac{1}{n}\left(\sum_{h=1}^{L} \frac{1}{L}\cdot\frac{d}{\sqrt{12}L}\right)^2 = \frac{d^2}{12nL^2} = \frac{V(\bar{y})}{L^2} \tag{5A.33}$$

Entonces, con una distribución rectangular la varianza de $\bar{y}_{st}$ decrece inversamente con el *cuadrado* del número de estratos. Es notable que esta relación se mantiene, aproximadamente, cuando se estratifican distribuciones oblicuas con extensión finita mediante la selección de límites óptimos para la asignación de Neyman. En ocho distribuciones de datos del tipo que puede que ocurra en la práctica, Cochran (1961) encontró que los valores promedio de

$V(\bar{y}_{st})/V(\bar{y})$ fueron 0.232, 0.098 y 0.053 para $L = 2, 3, 4$, en comparación con 0.250, 0.111 y 0.062 para la distribución rectangular

Estos resultados, que sugieren que la multiplicación de estratos es provechosa, presentan un cuadro engañador respecto a lo que pasa cuando alguna otra variable x se usa para construir los estratos Si $\phi(x) = E(y \mid x)$ es la regresión de y respecto a x, podemos escribir

$$y = \phi(x) + e \tag{5A.34}$$

donde ϕ y e no están correlacionados. Por lo tanto

$$S_y^2 = S_\phi^2 + S_e^2 \tag{5A.35}$$

De acuerdo con los resultados anteriores, la creación de L estratos para x puede reducir S_ϕ^2 a S_ϕ^2/L^2 si $\phi(x)$ es lineal o en un grado menor si $\phi(x)$ no es lineal. Pero el término S_e^2 no se reduce por la estratificación según x. Conforme L aumenta, tarde o temprano se alcanza un valor en el cual domina el término S_e^2. Aumentos posteriores en L producirán solamente una reducción trivial proporcional en $V(\bar{y}_{st})$.

Cuan pronto se alcance el punto de retornos decrecientes depende de un número de factores —particularmente los tamaños relativos de S_e^2 y S_ϕ^2 y la naturaleza de $\phi(x)$—. Sólo unos pocos ejemplos con datos actuales se encuentran en la literatura. Para suplementarlos, se usa aquí un procedimiento teórico simple. Supóngase que la selección óptima de límites de estrato por medio de x, con muestras de igual tamaño n/L en cada estrato, reduce $V(\bar{x}_{st})$ en una relación proporcional a $1/L^2$. Entonces

$$V(\bar{x}_{st}) = \frac{L}{n}\sum_{h=1}^{L} W_h^2 S_{xh}^2 = \frac{S_x^2}{nL^2} \tag{5A.36}$$

Supóngase también que la regresión de y en x es lineal, esto es,

$$y = \alpha + \beta x + e \tag{5A.37}$$

donde S_e^2 es constante. Entonces,

$$V(\bar{y}_{st}) = \frac{L}{n}\sum_{h=1}^{L} W_h^2 S_{yh}^2 = \frac{L\beta^2}{n}\sum_{h=1}^{L} W_h^2 S_{xh}^2 + \frac{LS_e^2}{n}\sum_{h=1}^{L} W_h^2 \tag{5A.38}$$

Para cualquier grupo de L estratos, $\sum W_h^2 \geq \frac{1}{L}$. Al usar (5A.36) tenemos

$$V(\bar{y}_{st}) \geq \frac{1}{n}\left(\frac{\beta^2 S_x^2}{L^2} + S_e^2\right) = \frac{S_y^2}{n}\left[\frac{\rho^2}{L^2} + (1-\rho^2)\right] \tag{5A.39}$$

donde ρ es la correlación entre y y x en la población por estratificar.

Con este modelo, la Tabla 5A.12 muestra $V(\bar{y}_{st})/V(\bar{y})$ para $\rho = 0.99$, 0.95, 0.90 y 0.85 y $L = 2$ a 6, suponiendo que la relación (5A.39) es una igualdad. Las columnas del lado derecho de la tabla dan $V(\bar{y}_{st})/V(\bar{y})$ para tres conjuntos de datos actuales, descritos abajo de la tabla en donde x es el valor de y en alguna ocasión anterior.

Los resultados para el modelo de regresión indican que a menos que ρ exceda 0.95, deberá esperarse poca reducción de la varianza más allá de $L = 6$. Los conjuntos de datos 2 y 3 apoyan esta conclusión, aunque un nuevo incremento de L puede ser provechoso en el caso del registro en colegios (grupo). En dos comparaciones sobre datos de encuesta, Hess, Sethi y Balakrishnan (1966) encontraron que $V(\bar{y}_{st})$ decrecía más rápidamente con L de lo que(5A.39) predice, lo que sugiere que el modelo (5A.37) es una sobresimplificación.

Para completar este análisis, requerimos una función de costo que muestre cómo éste depende de L. Dalenius (1957) sugiere la relación $C = LC_s + nC_n$. La razón de costo C_s/C_n variará con el tipo de encuesta. Un aumento en el número de estratos implica trabajo extra en la planificación y extracción de la muestra y aumenta el número de ponderaciones usadas en el cálculo de los estimados, a menos que los mismos sean autoponderados. En algunas encuestas ca-

TABLA 5A.12. $V(\bar{y}_{st})/V(\bar{y})$ COMO UNA FUNCIÓN DE L PARA EL MODELO DE REGRESIÓN LINEAL Y ALGUNOS DATOS REALES

	Modelo de regresión lineal				Datos, grupo		
	$\rho =$						
L	0.99	0.95	0.90	0.85	1	2	3
2	0.265	0.323	0.392	0.458	0.197	0.295	0.500
3	0.129	0.198	0.280	0.358	0.108	0.178	0.375
4	0.081	0.154	0.241	0.323	0.075	0.142	0.244
5	0.059	0.134	0.222	0.306	0.065	0.105	0.241
6	0.047	0.123	0.212	0.298	0.050	0.104	0.212
∞	0.020	0.098	0.190	0.277	—	—	—

Tipos de Datos

Grupo	Datos	x	y	Fuente
1	Inscripciones a colegios	1952	1958	Cochran (1961)
2	Tamaño de las ciudades	1940	1950	Cochran (1961)
3	Ingresos familiares	1929	1933	Dalenius y Gurney (1951)

si no se requiere cambio alguno en la organización del trabajo de campo; en otras, se establece una unidad de campo separada en cada estrato. Cualquiera que sea la forma de la función de costo, los resultados en la Tabla 5A.12 sugieren que si un aumento en L más allá de 6 necesita cualquier reducción en n para mantener el costo constante, el aumento es rara vez provechoso.

La discusión en esta sección está confinada a encuestas en las cuales se van a hacer estimaciones globales. Si también se desean estimaciones para subdivisiones de la población, es más fuerte el argumento en favor de un número mayor de estratos.

5A.9 ESTRATIFICACION DESPUES DE LA SELECCION DE LA MUESTRA (POSTESTRATIFICACION)

Con algunas variables adecuadas para la estratificación, el estrato al que pertenece una unidad no se conoce sino hasta después de recoger los datos. Ejemplos comunes son las características personales como la edad, el sexo, estatura, etc., y el nivel de educación. Los tamaños de los estratos N_h se pueden obtener de manera bastante exacta a partir de las estadísticas oficiales, pero las unidades se pueden clasificar en estratos, solamente después de conocer los datos de la muestra. Aquí suponemos que W_h y los N_h son conocidos.

Un procedimiento consiste en tomar una muestra aleatoria simple de tamaño n y clasificar las unidades. En lugar de la media muestral $\bar{y}$ usamos la estimación $\bar{y}_W = \sum W_h \bar{y}_h$, donde $\bar{y}_h$ es la media de las unidades de la muestra que caen en el estrato h y $W_h = N_h/N$. Este método es casi tan preciso como el muestreo estratificado proporcional siempre y cuando: (a) la muestra sea razonablemente grande, digamos mayor que 20 en cada estrato, y (b) los efectos de los errores en las ponderaciones W_h puedan ignorarse (ver la Sec. 5A.2).

Para demostrarlo, sea m_h el número de unidades de la muestra que se encuentran en el estrato h donde m_h variará de una muestra a otra. Para muestras en donde m_h son números fijos que exceden todos a cero,

$$V(\bar{y}_W) = \sum \frac{W_h^2 S_h^2}{m_h} - \frac{1}{N} \sum W_h S_h^2 \qquad (5A.40)$$

El valor promedio de $V(\bar{y}_w)$ en muestras repetidas de tamaño n debe calcularse ahora. Esto requiere un poco de cuidado porque uno o más de los m_h podrían ser nulos. Si esto ocurriera, habría que com-

binar dos o más estratos antes de hacer la estimación, y se lograría una estimación menos precisa. Al incrementar n, la probabilidad de que cualquier m_h sea nulo se hace tan pequeña, que la contribución a la varianza de esta fuente es despreciable.

En el caso en que m_h sea nulo, se ignora, Stephan (1945) ha demostrado que para los términos de orden n^{-2}

$$E\left(\frac{1}{m_h}\right)=\frac{1}{nW_h}+\frac{1-W_h}{n^2W_h^2} \qquad (5A.41)$$

Por tanto,

$$E[V(\bar{y}_w)]=\frac{1-f}{n}\sum W_hS_h^2+\frac{1}{n^2}\sum(1-W_h)S_h^2 \qquad (5A.42)$$

El primer término es el valor de $V(\bar{y}_{st})$ para estratificación proporcional. El segundo, representa el incremento en varianza que surge ya que los m_h no se distribuyen proporcionalmente. Pero

$$\frac{1}{n^2}\sum(1-W_h)S_h^2=\frac{1}{n}\left(\frac{L}{n}\right)\bar{S}_h^2-\frac{1}{n^2}\sum W_hS_h^2=\frac{1}{n\bar{n}_h}\bar{S}_h^2-\frac{1}{n^2}\sum W_hS_h^2 \qquad (5A.43)$$

donde $\bar{S}_h^2$ es el promedio de las S_h^2, y $\bar{n}_h = n/L$ es el número promedio de unidades por estrato. Así que, si las S_h^2 no difieren mucho, el incremento es aproximadamente de $(L-1)/L\bar{n}_h$ veces la varianza para estratificación proporcional, al ignorar la cpf. El incremento será pequeño si $\bar{n}_h$ es razonablemente grande.

Este método también puede aplicarse a una muestra ya estratificada por otro factor, por ejemplo, en cinco regiones geográficas, a condición de que los W_h se conozcan separadamente en cada región. Esta estratificación doble se emplea mucho en las encuestas nacionales de los Estados Unidos: véase Bean (1970) para una descripción de las fórmulas de estimación de la encuesta de salud por entrevistas, realizada por el National Center for Health Statistics.

5A.10 MUESTREO POR CUOTA

En otro método usado para sondeos de opinión y en las investigaciones de mercados, los n_h requeridos en cada estrato se calculan por adelantado, de modo que la estratificación sea proporcional. El enumerador tiene instrucciones de seguir muestreando hasta obtener la cuota necesaria en cada estrato. Las variables más comunes para estratificación son el área geográfica, la edad, el sexo, la raza y alguna medida del nivel económico. Si el enumerador elige personas al azar, dentro de las áreas geográficas, y asignase a cada una su estrato apropiado, el método coincidiría con el muestreo aleatorio estratificado. Se requiere una cantidad considerable de trabajo de campo para completar las cuotas, porque en las últimas etapas, la mayoría de las personas examinadas, corresponden a cuotas ya saturadas.

Para hacer más expédito el llenado de las cuotas, se permite cierta flexibilidad al encuestador, en lo que respecta a las personas o domicilios por incluirse. El grado de flexibilidad en general es variable, pero en general el muestreo por cuota puede describirse como muestreo estratificado con una selección más o menos no aleatoria dentro de los estratos. Por esta razón, las fórmulas de error de muestreo no podrán aplicarse con confianza a los resultados de muestras por cuota. Cierto número de comparaciones entre los resultados de muestras por cuota y por probabilidad han sido resumidos por Stephan y McCarthy (1958), quienes dan una crítica excelente sobre el comportamiento de ambos tipos de reconocimiento. El método de cuota es probable que produzca muestras que contienen sesgo en características tales como: ingresos, educación y ocupación, aunque a menudo concuerda bien con las muestras por probabilidad en cuestiones de opinión y actitud.

5A.11 ESTIMACION A PARTIR DE UNA MUESTRA DE LA GANANCIA DEBIDA A LA ESTRATIFICACION

Cuando se tomó una muestra aleatoria estratificada, puede ser de interés como una guía en la conducción de futuras encuestas, estimar la ganancia en precisión en relación con muestreo aleatorio simple.

Los datos disponibles de la muestra son los valores de N_h, n_h, $\bar{y}_h$ y s_h^2. De la Sec. 5.4, la varianza estimada de la media ponderada a partir de una muestra estratificada es, por la Fórmula (5.13),

$$v(\bar{y}_{st})=\sum\frac{W_h^2s_h^2}{n_h}-\sum\frac{W_hs_h^2}{N}$$

El problema es comparar esta varianza con una estimación de la varianza de la media que se hubiera obtenido de una muestra aleatoria simple. Un procedimiento usado algunas veces calcula la desviación cuadrática media de la media de la muestra,

$$s^2=\frac{\sum(y_{hi}-\bar{y})^2}{n-1}$$

donde se ignoran los estratos. Esta se toma como estimación de s^2, de tal modo que $\hat{V}_{ran} = (N-n)s^2/Nn'$ para la media de una muestra aleatoria simple. Este método funciona bastante bien si la asignación es proporcional, ya que una muestra aleatoria simple está distribuida en sí misma, aproximadamente en proporción, entre los estratos. Pero, si se ha adoptado una asignación alejada de la proporcional, la muestra tomada realmente no se parece a una muestra aleatoria simple, y este s^2 puede ser un estimador mediocre. Damos un procedimiento general cuya prueba se debe a J. N. K. Rao (1962).

Teorema 5A.1. Dados los resultados de una muestra aleatoria estratificada, un estimador insesgado de V_{ran}, varianza de la media de una muestra aleatoria simple de la misma población, es

$$V_{ran} = \frac{(N-n)}{n(N-1)}\left[\frac{1}{N}\sum_h^L \frac{N_h}{n_h}\sum_j^{n_h} y_{hj}^2 - \bar{y}_{st}^2 + v(\bar{y}_{st})\right] \tag{5A.44}$$

donde $v(\bar{y}_{st})$ es el estimador insesgado usual de $V(\bar{y}_{st})$.

Prueba.

$$V_{ran} = \frac{(N-n)}{nN}S^2 = \frac{(N-n)}{n(N-1)}\left[\frac{1}{N}\sum_h^L \sum_j^{N_h} y_{hj}^2 - \bar{Y}^2\right] \tag{5A.45}$$

Ahora bien

$$\frac{1}{N}E\left(\sum_h^L \frac{N_h}{n_h}\sum_j^{n_h} y_{hj}^2\right) = \frac{1}{N}\sum_h^L \sum_j^{N_h} y_{hj}^2 \tag{5A.46}$$

También como $v(\bar{y}_{st})$ y $\bar{y}_{st}$ son estimadores insesgados de $V(\bar{y}_{st})$ y $\bar{Y}$, respectivamente,

$$Ev(\bar{y}_{st}) = V(\bar{y}_{st}) = E(\bar{y}_{st}^2) - \bar{Y}^2 \tag{5A.47}$$

y luego un estimador insesgado de $\bar{Y}^2$ en (5A.45) es

$$\bar{y}_{st}^2 - v(\bar{y}_{st}) \tag{5A.48}$$

De (5A.46) y (5A.48) se infiere que un estimador muestral insesgado de V_{ran} en (5A.45) es

$$V_{ran} = \frac{(N-n)}{n(N-1)}\left[\frac{1}{N}\sum_h^L \frac{N_h}{n_h}\sum_j^{n_h} y_{hj}^2 - \bar{y}_{st}^2 + v(\bar{y}_{st})\right] \tag{5A.44}$$

Así se demuestra el Teorema 5A.1.

Con asignación proporcional, $(N_h/n_h = N/n)$, los dos primeros términos dentro de los corchetes cuadrados vienen a ser $(1/n)$ veces la suma de cuadrados dentro de muestra $= (n-1)s^2/n$. La Fórmula (5A.44) se reduce entonces a

$$v_{ran} = \frac{(N-n)}{n(N-1)}\left[\frac{(n-1)}{n}s^2 + v(\bar{y}_{st})\right] \tag{5A.49}$$

Si n es grande, $(n-1) \doteq n$, $(N-1) \doteq N$, y el término en $v(\bar{y}_{st})$ es de orden $1/n$ relativamente al término en s^2 en (5A.49).

Luego,

$$v_{ran} \doteq \frac{(N-n)}{nN}s^2 \tag{5A.50}$$

para asignación proporcional. En el caso general la simplificación correspondiente (n grande) es

$$v_{ran} \doteq \frac{(N-n)}{nN}\left[\frac{1}{N}\sum \frac{N_h}{n_h}\sum^{n_h} y_{hj}^2 - \bar{y}_{st}^2\right] \tag{5A.51}$$

Ejemplo. Los cálculos se ilustran a partir de los tres primeros estratos en la muestra de los colegios para maestros (Sec. 5.9). Los datos de la Tabla 5A.13 son para la muestra posterior de 1946. Las medias representan inscripciones por colegio en miles. Los valores s_h^2 son un poco mayores que en la segunda edición, debido a una corrección.

TABLA 5A.13. DATOS BÁSICOS DE UNA MUESTRA ESTRATIFICADA DE COLEGIOS DE MAESTROS

Estrato	N_h	n_h	$\bar{y}_h$	s_h^2	$\frac{N_h}{n_h}\left(\sum y_{hj}^2\right)$
1	13	9	2.200	1.8173	83.920
2	18	7	1.638	0.0735	49.429
3	26	10	0.992	0.0859	27.596
	—	—			—
	57	26			160.945

Si n es pequeña usamos la Fórmula (5A.44). Encontramos $\bar{y}_{st} = 1.4715$. Los valores de las y_h, para la muestra no se reportaron, pero el número en la columnas de la derecha puede obtenerse de las precedentes de la Tabla 5A.13. Las fórmulas dan

$$v(\bar{y}_{st}) = \frac{1}{N^2}\sum \frac{N_h(N_h-n_h)}{n_h}s_h^2 = 0.00497$$

$$v_{ran} = \frac{31}{(26)(56)}\left[\frac{160.945}{57} - (1.4715)^2 + 0.00497\right] = 0.01412$$

La estratificación parece haber reducido la varianza a (aprox.) un tercio del valor para una muestra aleatoria simple, siendo el factor por efecto de diseño, estimado, (Sec. 4.11), igual a $0.00497/0.01412 = 0.35$.

5A.12 ESTIMACION DE LA VARIANZA CON UNA UNIDAD POR ESTRATO

Si la población es altamente variable y se conocen numerosos criterios efectivos por estratificación ésta se puede llevar a cabo hasta el punto en que la muestra contenga sólo una unidad por es-

trato. En este caso, las fórmulas dadas anteriormente para estimar $V(\hat{Y}_{st})$ y $V(\bar{y}_{st})$ no podrán usarse. Con L par, se puede intentar una estimación al agrupar los estratos en pares que de antemano se cree que tienen aproximadamente los mismos totales de estrato. La asignación en pares deberá hacerse antes de ver los resultados de la muestra por razones que veremos en lo sucesivo.

Sean y_{j1}, y_{j2} las observaciones de la muestra, en un par típico, donde j recorre los valores de 1 a $L/2$. Sea $\hat{Y}_{j1} = N_{j1}y_{j1}$, $\hat{Y}_{j2} = N_{j2}y_{j2}$ los totales estimados de estratos. Ahora

$$\hat{Y}_{j1} - \hat{Y}_{j2} = (Y_{j1} - Y_{j2}) + (\hat{Y}_{j1} - Y_{j1}) - (\hat{Y}_{j2} - Y_{j2}) \tag{5A.52}$$

Luego, al promediar en todas las muestras de este par

$$E(\hat{Y}_{j1} - \hat{Y}_{j2})^2 = (Y_{j1} - Y_{j2})^2 + N_{j1}(N_{j1} - 1)S_{j1}^2 + N_{j2}(N_{j2} - 1)S_{j2}^2 \tag{5A.53}$$

Para $V(\hat{Y}_{st})$ considérese la estimación

$$v_1(\hat{Y}_{st}) = \sum_{j=1}^{L/2} (\hat{Y}_{j1} - \hat{Y}_{j2})^2 \tag{5A.54}$$

Por (5A.53) el valor esperado de esta cantidad es

$$Ev_1(\hat{Y}_{st}) = \sum_{h=1}^{L} N_h(N_h - 1)S_h^2 + \sum_{j=1}^{L/2} (Y_{j1} - Y_{j2})^2 \tag{5A.55}$$

El primer término de la derecha es la varianza correcta (por el Teorema 5.4 con $n_h = 1$). El segundo, representa un sesgo positivo cuyo tamaño depende del éxito alcanzado al seleccionar pares de estratos cuyos verdaderos totales difieren un poco. La forma de la estimación (5A.54) nos dice que la construcción de pares al hacer que los totales estimados de *muestra* difieran lo menos posible, puede conducir a una seria subestimación. La técnica se llama método de los "estratos contraídos".

Con L impar, cuando menos un grupo debe ser de tamaño diferente de 2. La extensión de la estimación (5A.54) a G grupos de cualesquier tamaños elegidos $L_j \geq 2$ es

$$v_1(\hat{Y}_{st}) = \sum_{j=1}^{G} \frac{L_j}{L_j - 1} \sum_{k=1}^{L_j} (\hat{Y}_{jk} - \hat{Y}_j/L_j)^2 \tag{5A.56}$$

donde $\hat{Y}_j$ es el total estimado para el grupo j. Para $L_j = 2$, cuando $\hat{Y}_j = \hat{Y}_{j1} + \hat{Y}_{j2}$, esta forma concuerda con (5A.54). Como ocurre con (5A.54), la esperanza de esta $v_1(\hat{Y}_{st})$ da la varianza correcta $V(\hat{Y}_{st})$ más un sesgo positivo que sustituye Y_{jk} y Y_j por $\hat{Y}_{jk}$ y $\hat{Y}_j$ en (5A.56)

Cuando se conoce una variable auxiliar A_h para cada estrato, que pronostica el total de estrato Y_k, Hansen, Hurwitz y Madow (1953) sugirieron el estimador de varianza alternativo

$$v_2(\hat{Y}_{st}) = \sum_{j=1}^{G} \frac{L_j}{L_j - 1} \sum_{k=1}^{L_j} (\hat{Y}_{jk} - A_{jk}\hat{Y}_j/A_j)^2 \tag{5A.57}$$

Si A_k es un buen pronóstico, el sesgo positivo del término v_2 que proviene de las desviaciones $(\hat{Y}_{jk} - A_{jk}\hat{Y}_j/A_j)^2$ es de esperarse que sea menor al término correspondiente en v_1, aunque difiriera de v_1, v_2 también da una estimación sesgada del término en la S_h^2 que aparece en $V(\hat{Y}_{st})$. Hartley, Rao y G. Kiefer (1969) encontraron que v_2 es menos sesgado que v_1 en dos de tres poblaciones, con poca diferencia en la tercera.

Estos autores desarrollaron un método que no involucra la contracción de estratos. Este método usa una o más variables auxiliares x_{1h}, x_{2h} y así sucesivamente, sobre las cuales las verdaderas medias de estratos $\bar{Y}_h$ se cree que tienen una regresión lineal. Si y_h es el valor muestral en el estrato h, el método usa las desviaciones

$$d_h = y_h - \bar{y} - \sum_i b_i(x_{ih} - \bar{x}_i) \tag{5A.58}$$

La matriz de varianza y covarianza de las d_h puede expresarse como una función lineal de las σ_{h2} más ciertos términos de sesgo. Al invertir esta relación, se obtienen estimaciones σ_h^2 donde $\sigma_h^2 = (N_h - 1)S_h^2/N_h$, lo queda

$$v_3(\hat{Y}_{st}) = \sum_{h=1}^{L} N_h^2 \hat{\sigma}_h^2 \tag{5A.59}$$

El método parece prometedor y se extiende a estimaciones de razón, pero los autores previenen que se necesitan métodos adicionales de comparación con "estratos contraídos".

Al utilizar otro enfoque, Fuller (1970) desarrolló un método de construcción de estratos que proporciona una estimación muestral insesgada de $V(\hat{Y}_{st})$ con una unidad por estrato. Por sencillez, supongamos que $N/n = N/L = k$ (un entero). Elíjase un número aleatorio r entre 1 y k. El primer estrato consiste en las unidades numeradas de $(r + 1)$ hasta $(r + k)$, el segundo, de las unidades numeradas de $(r + k + 1)$ hasta $(r + 2k)$ y así sucesivamente, el último estrato (L-ésimo, n-ésimo) de las unidades numeradas de $r + (n - 1)k + 1$ hasta $N = nk$ y aquellas de 1 a r. A primera vista, este último estrato parece ser una elección mediocre. Pero como Fuller lo hace notar, el método da resultados en la estratificación geográfica con unidades de área. Aquí, la estratificación suele basarse en la noción de que las unidades cercanas entre sí tienden a ser semejantes. Al numerar las unidades como serpentina, podemos tener y_N cerca de y_1 de modo que el estrato que contiene y_N y y_1 también puede ser internamente homogéneo. La estimación $v(\hat{Y}_{st})$ es una suma ponderada de las diferencias $(y_h - y_{h+1})^2$.

El método circular sería menos efectivo para una población con tendencia creciente de y_1 a y_N en donde el estrato que incluye y_1 y

y_N tendría una considerable variabilidad interna. Para esta situación, Fuller da un segundo plan, un poco más complejo, que debería proporcionar una buena precisión con tendencia creciente y también proporciona una estimación insesgada de $V(\hat{Y}_{st})$.

5A.13 ESTRATOS COMO DOMINIOS DE ESTUDIO

En esta sección veremos las encuestas cuyo propósito primordial es el de comparar entre diferentes estratos, supuestamente identificables de antemano. Las reglas para asignar los tamaños de muestra para los estratos, son diferentes a las que se aplican cuando el objetivo es estimar globales de población. Si sólo hay dos estratos, podríamos elegir n_1, n_2 para minimizar la varianza de la diferencia $(\bar{y}_1 - \bar{y}_2)$ entre las medias estimadas de estratos. Al omitir la cpf por razones que vimos en la Sec. 2.14, tendremos

$$V(\bar{y}_1 - \bar{y}_2) = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \tag{5A.60}$$

Con una función costo lineal

$$C = c_0 + c_1 n_1 + c_2 n_2 \tag{5A.61}$$

V se minimiza cuando

$$n_1 = \frac{\frac{nS_1}{\sqrt{c_1}}}{S_1/\sqrt{c_1} + S_2/\sqrt{c_2}}, \qquad n_2 = \frac{\frac{nS_2}{\sqrt{c_2}}}{S_1/\sqrt{c_1} + S_2/\sqrt{c_2}} \tag{5A.62}$$

Con L estratos, $L > 2$, la asignación óptima depende de la precisión deseada en las diferentes comparaciones. Por ejemplo, el costo podría minimizar siempre y cuando el conjunto de $L(L-1)/2$ haga que $V(\bar{y}_h - \bar{y}_i) \leq V_{hi}$, donde los valores de V_{hi} se eligen según la precisión que se considere necesaria para una comparación satisfactoria de los estratos h e i.

Con frecuencia hay un método de asignación, más simple pero adecuado, especialmente si las S_h y c_h no difieren mucho. Un enfoque es el de minimizar la varianza *promedio* de la diferencia entre todos los $L(L-1)/2$ pares de estratos, o sea minimizar

$$\bar{V} = \frac{2}{L}\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + \cdots + \frac{S_L^2}{n_L}\right) \tag{5A.63}$$

$\bar{V}$ se minimiza, para C fija, por medio de la regla de (5A.62)

$$n_h \propto \frac{S_h}{\sqrt{c_h}} \tag{5A.64}$$

Esta regla puede dar como resultado una comparación más precisa en algunos pares de estratos y en otros una comparación menos precisa de los apropiado. Como alternativa tenemos la selección de n_h de modo que la desviación estándar de la diferencia sea la misma, digamos $\sqrt{V}$ para cada par de estratos. Esto equivale a hacer $S_h^2/n_h = V/2$ para cada estrato. Para un costo fijo, este método proporciona menos precisión global que el primero. Se puede verificar que las dos asignaciones óptimas dan

$$\bar{V} = \frac{2(\sum S_h\sqrt{c_h})^2}{L(C-c_0)}, \qquad V = \frac{2(\sum S_h^2 c_h)}{(C-c_0)} \tag{5A.65}$$

Se infiere de la desigualdad de Cauchy-Schwarz que V es siempre mayor que $\bar{V}$ excepto cuando $S_h\sqrt{c_h}$ = constante. Si V es bastante mayor que $\bar{V}$ en ocasiones se puede encontrar una asignación de compromiso, con un poco de prueba y error, que dará una varianza promedio cercana a $\bar{V}$ y conserve $V(\bar{y}_h - \bar{y}_i)$ razonablemente constante.

Si el objetivo es obtener estimaciones para cada estrato, así como estimaciones globales para toda la población, al planear una encuesta podríamos especificar las siguientes condiciones

$$V(\bar{y}_h) = \frac{S_h^2}{n_h}(1-f_h) \leq V_h, \qquad V(\bar{y}_{st}) = \sum \frac{N_h^2 S_h^2}{n_h}(1-f_h) \leq V$$

Se incluyen ahora los términos cpf porque el propósito es especificar la precisión con la que deben estimarse las medias en la población finita. Las condiciones sobre las $V(\bar{y}_h)$ determinan límites inferiores a los valores de los n_h. Si estos límites inferiores satisfacen la condición sobre $V(\bar{y}_{st})$, se ha resuelto el problema de asignación. Cuando la condición impuesta a $V(\bar{y}_{st})$ no se satisface, tenemos un método gráfico elaborado por Dalenius (1957).

Cuando los $L = 2^k$ estratos representan todas las combinaciones de k factores en dos niveles, surgen problemas más complejos y el objetivo es estimar los efectos promedio de los factores. Si el estrato o celda al que pertenece cualquier miembro de la población, se conoce de antemano (antes del muestreo) podrá sacarse del estrato h una muestra de tamaño deseado. n_h Para 2, 3 o 4 factores, Sedransk (1967) ha dado métodos para determinar los n_h que minimizan el costo en diferentes especificaciones respecto a las varianzas de los efectos promedio estimados de los factores, y la potencia deseada de una prueba para interacciones.

5A.14 ESTIMACION DE TOTALES Y MEDIAS SOBRE SUBPOBLACIONES

Con frecuencia sucede que las subpoblaciones o dominios de estudio estén representados en todos los estratos. Si la estratificación es geográfica, por ejemplo, pueden desearse estimaciones separadas de toda la población, para hombres y mujeres, para diferentes grupos de edades, para usuarios y no usuarios de equis pasta dental y cosas semejantes. El problema presenta algunas complicaciones. Las fórmulas básicas las dio Yates (1953), y Durbin (1958) y Hartley (1959) aportaron discusiones y pruebas. En las Secs. 2.10 y 2.11 se vieron los métodos aplicables a un solo estrato.

La notación siguiente se aplica a las unidades del estrato h que se encuentran en el dominio j.

Notación.

Número de unidades: N_{hj}, $\sum_j N_{hj} = N_h$

Número en la muestra: n_{hj}, $\sum_j n_{hj} = n_h$

Medición en la unidad individual: y_{hij}

Media muestral: $\bar{y}_{hj} = \sum_{i=1}^{n_{hj}} \frac{y_{hij}}{n_{hj}}$

Media de dominio: $\bar{Y}_{hj} = \sum_{i=1}^{N_{hj}} \frac{y_{hij}}{N_{hj}}$

El total y la media de la población para el dominio j sobre todos los estratos, es respectivamente:

$$Y_j = \sum_h N_{hj} \bar{Y}_{hj}, \qquad \bar{Y}_j = \frac{Y_j}{N_j}$$

donde $N_j = \sum_h N_{hj}$.

Las cosas se complican porque los n_{hj} son variables aleatorias. Si los N_{hj} se conociesen, el problema sería simple. Como estimaciones de Y_j y $\bar{Y}_j$, podríamos usar

$$\hat{Y}_h{}' = \sum_h N_{hj} \bar{y}_{hj}, \qquad \hat{\bar{Y}}_j{}' = \frac{\hat{Y}_j{}'}{N_j}$$

Por el método descrito en la Sec. 2.12 la fórmula ordinaria para $V(\bar{y}_{hj})$ sigue siendo válida, siempre que todos los $n_{hj} > 0$. Así

$$V(\hat{Y}_j{}') = \sum_h \frac{N_{hj}^2 S_{hj}^2}{n_{hj}} \left(1 - \frac{n_{hj}}{N_{hj}}\right) \tag{5A.66}$$

donde las S^2_{hj} es la varianza entre unidades en el dominio j, dentro del estrato h. Pero en la práctica rara vez se conocen los N_{hj}.

Estimación de Totales de Dominios

A falta de los N_{hj} cada total de estrato del dominio se estima como en la Sec. 2.13. Estos totales se suman para obtener un total de dominio estimado, o sea

$$\hat{Y}_j = \sum_h \frac{N_h}{n_h} \sum_i^{n_{hj}} y_{hij} \tag{5A.67}$$

La varianza verdadera y estimada de $\hat{Y}_j$ se encuentra por medio del artificio usado en la Sec. 2.13. Se introduce una variable y'_{hi} que es igual a y_{hi} para todas las unidades del dominio j e igual a cero para todas las otras unidades de la población. Como se demostró en la Sec. 2.13, esto da para la varianza estimada

$$v(\hat{Y}_j) = \sum_h \frac{N_h^2}{n_h(n_h-1)} (1-f_h) \left[\sum_i^{n_{hj}} y_{hij}^2 - \frac{(\sum y_{hij})^2}{n_h} \right] \tag{5A.68}$$

Estimación de Medias de Dominio

Para estimar la media de dominio Y_j/N_j se requiere una estimación muestral de N_j. La siguiente es una estimación insesgada

$$\hat{N}_j = \sum_h \frac{N_h}{n_h} n_{hj} \tag{5A.69}$$

De modo que tomamos

$$\hat{\bar{Y}}_j = \frac{\hat{Y}_j}{\hat{N}_j} = \frac{\sum_h (N_h/n_h) \sum_i y_{hij}}{\sum_h (N_h/n_h) n_{hj}} \tag{5A.70}$$

Con estratificación proporcional, $\hat{\bar{Y}}_j$ se reduce a la media de la muestra ordinaria de las unidades que caen en el dominio j. En general esta estimación se conoce como una *estimación* de razón combinada, que discutiremos más tarde en la Sec. 6.11. Para demostrarlo, se introduce otra variable muda x'_{hi} que es igual a 1 para cada unidad en el dominio j y cero para todas las otras unidades, donde ahora, i va de 1 a N_h. Claramente,

$$\bar{x}_h{}' = \frac{\sum_i^{n_h} x_{hi}{}'}{n_h} = \frac{n_{hj}}{n_h}, \qquad \bar{y}_h{}' = \frac{\sum_i^{n_h} y_{hi}}{n_h} = \frac{\sum_i^{n_{hj}} y_{hij}}{n_h} = \frac{n_{hj}}{n_h} \bar{y}_{hj} \tag{5A.71}$$

De modo que la media de dominio estimada puede escribirse

$$\hat{\bar{Y}}_j = \frac{\sum_h (N_h/n_h)\sum_i y_{hij}}{\sum_h (N_h/n_h)n_{hj}} = \frac{\sum_h N_h \bar{y}_h'}{\sum_h N_h \bar{x}_h'} = \frac{\bar{y}_{st}}{\bar{x}'_{st}} \qquad (5A.72)$$

Esta es la fórmula para la estimación combinada de razón para las dos variables y_{hi}' y x_{hi}'. En la Sec. 6.11 vemos que la varianza estimada puede expresarse aproximadamente como

$$v(\hat{\bar{Y}}_j) \doteq \frac{1}{\hat{N}_j^2}\sum_h \frac{N_h^2(1-f_h)}{n_h(n_h-1)}\sum_i^{n_h}[y_{hi}' - \hat{\bar{Y}}_j x_{hi}' - (\bar{y}_h' - \hat{\bar{Y}}_j \bar{x}_h')]^2 \qquad (5A.73)$$

La segunda suma puede escribirse

$$\sum_i^{n_h}(y_{hi}' - \hat{\bar{Y}}_j x_{hi}')^2 - n_h(\bar{y}_h' - \hat{\bar{Y}}_j \bar{x}_h')^2 = \sum_i^{n_{hj}}(y_{hij} - \hat{\bar{Y}}_j)^2 - \frac{n_{hj}^2}{n_h}(\bar{y}_{hj} - \hat{\bar{Y}}_j)^2 \qquad (5A.74)$$

Al usarlo (5A.71). Además, el primer término de (5A.74) puede expresarse alternativamente como

$$\sum_i^{n_{hj}}(y_{hij} - \bar{y}_{hj})^2 + n_{hj}(\bar{y}_{hj} - \hat{\bar{Y}}_j)^2$$

Insertando estos resultados en (5A.73) finalmente se tiene para la varianza estimada

$$v(\hat{\bar{Y}}_j) \doteq \frac{1}{\hat{N}_j^2}\sum_h \frac{N_h^2(1-f_h)}{n_h(n_h-1)}\left[\sum_i (y_{hij} - \bar{y}_{hj})^2 + n_{hj}\left(1 - \frac{n_{hj}}{n_h}\right)(\bar{y}_{hj} - \hat{\bar{Y}}_j)^2\right] \qquad (5A.75)$$

El término a la derecha representa una contribución entre estratos a la varianza. Las diferencias entre las medias de estratos no se eliminan totalmente de la varianza de la media estimada de cualquier subpoblación. La contribución entre estratos es pequeña si los términos $1 - n_{hj}/n_h$ son pequeños, es decir si la subpoblación es casi tan grande como la población total.

Como ha señalado Durbin (1958), (5A.75) se aplica también a las medias estimadas para toda la población, si la muestra fuera incompleta por razones tales como: no-respuestas, siempre y cuando, se use la estimación $\hat{\bar{Y}}_j$. En este caso $\hat{\bar{Y}}_j$ se interpreta como la media estimada para la parte de la población que daría una respuesta con los métodos empleados de recolección de datos. Pero existe una complicación adicional ya que la parte no-respuesta de la población suele tener una media diferente de la parte que sí responde. Por lo tanto, $\hat{\bar{Y}}_j$ es una estimación sesgada de la media de toda la población y esta contribución sesgada no aparece en (5A.75).

5A.15 MUESTREO A PARTIR DE DOS MARCOS

Uno de los primeros ejemplos del uso de una muestra tomada de una lista B de grandes negocios, en combinación con una muestra de un marco A por áreas, que cubre la población completa, es la encuesta por muestreo en 1949 de las tiendas al menudeo, que tomó la oficina de censos, y que describen Hansen, Hurwitz y Madow (1953) (Pág. 516 y siguientes). Los objetivos del uso combinado de una lista marco incompleta y un marco por áreas completo, eran ganar exactitud y ahorrar dinero. Los grandes negocios en ocasiones se pueden muestrear a bajo costo gracias a una combinación de correo y teléfono: además, al ser negocios grandes también tienen la varianza más grande para las variables y que se miden. Al colocarlos en un estrato separado con asignación óptima (o con muestreo del 100% si esto parece ser lo más cercano al óptimo) se producen incrementos importantes de exactitud. En la encuesta de las tiendas al menudeo, los negocios de la lista marco que caían en la muestra de áreas se identificaron y retiraron del área de muestreo, para que la población muestreada cayera en dos estratos distintos. Este proceso llamado *deduplicación* lo llevó a cabo el supervisor de campo antes del muestreo. La deduplicación se complica a veces y presenta errores. Algunas de las dificultades prácticas las discuten Hansen, Hurwitz y Jabine (1963) al igual que discuten la manera como se pueden utilizar las listas incompletas.

La identificación de los miembros de la muestra a partir de la muestra del marco A que pertenecen a la lista marco B, en ocasiones requiere que se midan las variables y para estos miembros. Entonces el muestreador tiene tres muestras a su disposición en las que se ha medido y —una muestra del estrato a (la parte de A que no pertenece a B) y dos muestras independientes del estrato B. Una de ellas se denota con el subíndice ab y es la muestra obtenida de A que se identifica como perteneciente a B, y la otra se obtiene por muestreo directo del marco B. Con el marco A completo y muestreo aleatorio simple de ambos marcos, Hartley (1962) propuso que ambas muestras B se usaran en la estimación postestratificada.

$$\hat{Y} = N_a\bar{y}_a + N_{ab}(p\bar{y}_{ab} + q\bar{y}_B) \qquad (5A.76)$$

donde $\bar{y}_a, \bar{y}_{ab}, \bar{y}_B$ denotan las medias de muestra respectivas. Los factores de ponderación p y q para las dos muestras que pertenecen al marco B, con $p + q = 1$ se eligen para minimizar $V(\hat{Y})$ con una función costo de la forma

$$C = c_A n_A + c_B n_B \qquad (5A.77)$$

En (5A.76) los tamaños de estratos $N_{ab} = N_B$ y $N_a = (N_A - N_B)$ serán conocidos.

Con $S_B^2 > S_a^2$ y $c_B < c_A$, Hartley demostró que este método puede dar grandes reducciones en $V(\hat{Y})$ al compararlas con el muestreo de solamente el marco A, aunque la muestra del marco A esté postestratificada en los dos estratos a y $B = ab$.

El problema es más difícil si el marco A también es incompleto, requiriéndose entonces dos marcos A y B con alguna duplicación para cubrir totalmente a la población. Para postestratificación hay tres estratos distintos: a (unidades en A solamente); ab (unidades en A y B a la vez); y b (unidades en B solamente). Los tres estratos no podrán muestrearse en forma directa y habrá que sacar muestras de tamaños n_A, n_B de los marcos A y B. Además, los tamaños N_a, N_{ab}, N_b de los estratos por lo general no se conocen. Para muestreo aleatorio simple, a partir de dos marcos A y B, Hartley (1962) sugirió la estimación

$$\hat{Y} = \frac{N_A}{n_A} y_a + p\frac{N_A}{n_A} y_{ab} + q\frac{N_B}{n_B} y_{ba} + \frac{N_B}{n_B} y_b \qquad (5A.78)$$

donde, al igual que antes, $p + q = 1$ y los y son *totales* de muestra en los estratos. Los subíndices ab y ba denotan las muestras en el estrato duplicado que se encontraron n_A y n_B. Hartley determinó p y q para minimizar $V(\hat{Y})$ con costo fijo. Lund (1968) y Fuller y Burmeister (1972) dieron mejores estimaciones que las de Hartley (5A.78) esencialmente por utilizar mejores estimaciones de N_a, N_{ab}, N_b que las empleadas por Hartley en su estimación Fuller y Burmeister (1972) también trataron el caso en que el marco A es de áreas con submuestreo de unidades de área. Hartley (1974) da un método general al muestreo de dos marcos, aplicable a cualquier diseño de muestras en ambos marcos.

EJERCICIOS

5A.1. Al planear una encuesta de ventas en cierto tipo de tienda, con $n = 550$, se dispone de buenas estimaciones de S_h, provenientes de una encuesta anterior, en dos de los tres estratos. El tercer estrato consiste de tiendas nuevas y tiendas que no tuvieron ventas en la encuesta anterior, de tal manera que se tiene que conjeturar un valor para S_3. Si S_3, es actualmente 10, calcule $V(\bar{y}_{st})$ según una asignación estimada de Neyman cuando S_3 se conjetura como (a) 5, (b) 20. Muestre que en ambos casos el aumento proporcional en varianza sobre el óptimo verdadero es ligeramente mayor al 2%.

Estrato	W_h	Verdadera S_h	Estimada S_h (a)	(b)
1	0.3	30	30	30
2	0.6	20	20	20
3	0.1	10	5	20

5A.2. Muestre que si todas las S_h, excepto S_L, se estiman correctamente y S_L se estima como $\hat{S}_L = S_L(1 + \lambda)$, el aumento proporcional en $V_{\acute{o}pt}(\bar{y}_{st})$, al usar $\hat{S}_L$ en lugar de la verdadera S_L para la asignación de Neyman, es

$$\frac{\lambda^2 n'_L(n - n'_L)}{(1+\lambda)n^2}$$

donde n'_L es el tamaño de muestra en el estrato L bajo la verdadera asignación de Neyman. Verifique que esta fórmula concuerda con los resultados del ejercicio 5A.1. (La concordancia no es exacta debido al redondeo de las n_h a enteros.) De aquí, muestre que una subestimación de 50% de S_L tiene el mismo efecto que una sobreestimación del 100%.

5A.3. Si hay dos estratos y si ϕ es la razón n_1/n_2 real al n_1/n_2 óptimo de Neyman demuestre que cualesquiera que sean los valores de N_1, N_2, S_1 y S_2, la razón $V_{min}(\bar{y}_{st})/V(\bar{y}_{ts})$ nunca es menor que $4\phi(1 + \phi)^2$, cuando las cpf pueden ignorarse.

5A.4. Los resultados de una muestra altamente simple con $n = 1\,000$ pueden clasificarse en tres "estratos", con $\bar{y}_h = 10.2$, 12.6 y 17.1; $s_h^2 = 10.82$ (la misma en cada estrato) y $s^2 = 17.66$. Las ponderaciones estimadas para los estratos son $w_h = 0.5$, 0.3 y 0.2, respectivamente. Se sabe que estas ponderaciones son inexactas, pero se piensa que todas son correctas dentro de un 5%, de tal manera que en el peor de los casos son o $W_h = 0.525$, 0.285 y 0.190 o $W_h = 0.475$, 0.315 y 0.210. Mediante los métodos de la Sec. 5A.2, ¿recomendaría usted la estratificación? (Donde se necesite, suponga que $\bar{y}_h = \bar{Y}_h$ y $s_h^2 = S_h^2$.)

5A.5. En una muestra aleatoria estratificada con dos variables, el objetivo es satisfacer las especificaciones

$$V(\bar{y}_{1st}) \le V_1; \qquad V(\bar{y}_{2st}) \le V_2$$

para el costo mínimo $C = \sum c_h n_h$. La cpf puede ignorarse. (a) Demostrar el resultado de Chatterjee (1972) en el que una asignación de compromiso será necesaria si

$$\frac{\sum W_h S_{2h}\sqrt{c_h}}{\sum W_h (S_{1h}^2/S_{2h})\sqrt{c_h}} \le \frac{V_2}{V_1} \le \frac{\sum W_h (S_{2h}^2/S_{1h})\sqrt{c_h}}{\sum W_h S_{1h}\sqrt{c_h}}.$$

(b) Si V_2/V_1 es igual o excede al límite superior, la asignación óptima para y_1 satisface ambas tolerancias con un resultado correspondiente cercano al límite inferior.

5A.6. Una encuesta con tres estratos se ha planeado para estimar el porcentaje de familias que tienen cuentas de ahorros en el banco y la cantidad promedio invertida por familia. Los siguientes son estimaciones de los porcentajes P_h y las S_h intraestratos para la cantidad invertida.

Estrato	W_h	$P_h(\%)$	$S_h(\$)$
1	0.6	20	90
2	0.3	40	180
3	0.1	70	520

Calcular los tamaños más pequeños de muestra y los n_h que satisfagan ambos requisitos: (a) el porcentaje de familias debe estimarse con desviación estándar igual a 2 y la cantidad promedio invertida con error estándar igual a \$5. (b) El porcentaje de familias debe estimarse con un error estándar de 1.5 y la cantidad invertida promedio con un error estándar de \$5.

La parte (b) requiere una asignación de compromiso, ya sea por medio de un programa de cálculo o bien, por el método de la segunda edición (Pág 164). La asignación $n_h = 371\,344\,315$, con $n = 1\,030$ satisface las dos tolerancias. Demostrar que el método de Booth-Sedransk, (Sec. 5A.4) da $n_h/n =$ 0.431, 0.326, 0.243. Esta asignación requeriría $n = 1\,073$ para satisfacer ambas tolerancias.

5A.7. La tabla de la parte superior de la Pág. 192, es de la distribución de frecuencia de una población de 911 tamaños de ciudades, para ciudades entre 10 000 y 60 000, dispuestas en clases de 2 000. Para abreviar los cálculos, se da una y' codificada y los valores de $\sqrt{f}$, cum $\sqrt{f}$, cum f, fy', y $\sum fy'^2$. Aplicar la regla de Dalenius-Hodges para crear dos estratos con asignación óptima en el sentido de Neyman. Encontrar los valores de W_h y S_h para cada uno de los estratos. Verificar (a) que los tamaños óptimos de muestra son casi los mismos en ambos estratos y (b) al encontrar S^2 para toda la población, que

$$\frac{V(\bar{y})}{V_{opt}(\bar{y}_{st})} \doteq 4.8$$

5A.8. La distribución en triángulo rectángulo $f(y) = 2(1 - y)$, $0 < y < 1$, se divide en dos estratos en el punto a. (a) Mostrar que

$$W_1 = a(2-a), \qquad W_2 = (1-a)^2$$

$$S_1^2 = \frac{a^2(6-6a+a^2)}{18(2-a)^2}, \qquad S_2^2 = \frac{(1-a)^2}{18}$$

(b) Muestre que bajo la regla cum. $\sqrt{f}$ la mejor elección de a es $1 - 1/\sqrt[3]{4} = 0.37$ y que con este límite, la óptima n_1/n_2 es cerca de 27/25 y $V(\bar{y}_{st})$ es cerca del 27% del valor dado por muestreo aleatorio simple.

5A.9. En los Ejercicios 5A.7 y 5A.8, demostrar que la regla de Ekman $W_h(y_h - y_{h-1}) =$ constante concuerda muy bien con la regla cum $\sqrt{f}$ al determinar los límites de estratos.

	f	y'	$\sqrt{f}$	Cum f	Cum $\sqrt{f}$	fy'
10 –	205	0	14.3	205	14.3	0
12 –	135	1	11.6	340	25.9	135
14 –	106	2	10.3	446	36.2	212
16 –	82	3	9.1	528	45.3	246
18 –	61	4	7.8	589	53.1	244
20 –	42	5	6.5	631	59.6	210
22 –	32	6	5.7	663	65.3	192
24 –	30	7	5.5	693	70.8	210
26 –	27	8	5.2	720	76.0	216
28 –	18	9	4.2	738	80.2	162
30 –	22	10	4.7	760	84.9	220
32 –	21	11	4.6	781	89.5	231
34 –	19	12	4.4	800	93.9	228
36 –	16	13	4.0	816	97.9	208
38 –	14	14	3.7	830	101.6	196
40 –	17	15	4.1	847	105.7	255
42 –	9	16	3.0	856	108.7	144
44 –	8	17	2.8	864	111.5	136
46 –	11	18	3.3	875	114.8	198
48 –	9	19	3.0	884	117.8	171
50 –	7	20	2.6	891	120.4	140
52 –	4	21	2.0	895	122.4	84
54 –	5	22	2.2	900	124.6	110
56 –	5	23	2.2	905	126.8	115
58 –	6	24	2.4	911	129.2	144
Totales	911		129.2			4 407

$$\sum fy'^2 = 50\,395$$

5A.10. Se dispone de una suma de \$5 000 para una muestra estratificada. En la notación de la Sec. 5A.8, la función de costo se piensa que es aproximadamente, $C = 200L + 10n$ y

$$V(\bar{y}_{st}) \doteq \frac{S^2}{n}\left[\frac{\rho^2}{L^2} + (1-\rho^2)\right]$$

donde ρ es la correlación entre la variable empleada para construir los estratos y la variable a medir en la encuesta. Calcule la L óptima para $\rho = 0.95$, 0.9 y 0.8. ¿Cuál es un buen número de compromiso para los estratos a usar para los tres valores de ρ?

5A.11. Los siguientes datos están tomados de una muestra estratificada de vendedores de llantas en marzo de 1945 (Deming y Simmons, 1946). Los vendedores se asignaron a los estratos según el número de llantas nuevas en existencia durante un censo previo. Las medias de muestra $\bar{y}_h$ son el número medio de llantas nuevas por vendedor (a) estimar la ganancia en precisión debida a la estratificación. (b) comparar este resultado con lo que se gana al usar la asignación proporcional.

Límites de estratos	N_h	W_h	$\bar{y}_h$	s_h^2	n_h
1–9	19 850	0.8032	4.1	34.8	3 000
10–19	3 250	0.1315	13.0	92.2	600
20–29	1 007	0.0407	25.0	174.2	340
30–39	606	0.0245	38.2	320.4	230
Totales	24 713	0.9999			4 170

5A.12. Una población tiene dos estratos de tamaños realtivos $W_1 = 0.8$, $W_2 = 0.2$ y varianzas dentro de estratos $S_1^2 = 100$, $S_1^2 = 400$. Se toma una muestra aleatoria estratificada que satisface las siguientes condiciones: (i) las medias de cada estrato deben estimarse con varianza ≤ 1; (ii) $V(\bar{y}_{st}) \leq 0.5$. Al ignorar la cpf encontrar los valores de n_1, n_2 que satisfacen las tres condiciones para el mínimo $n = n_1 + n_2$.

Sugerencia. Nótese que $-\partial V(\bar{y}_{st})/\partial n_1 > -\partial V(\bar{y}_{st})/\partial n_2$ si $n_1 < 2n_2$. Fuller (1966) ha discutido varios métodos para tratar este problema.

5A.13. En un ejemplo de Nordbotten (1956) que estudió Kokan (1963) se planea una encuesta para estimar el empleo total Y_1 y el valor de la producción Y_2 en fábricas de muebles. Cuando las fábricas se estratifican por tamaños, los N_h y las S_{ih}^2 estimadas burdamente son como sigue:

Estrato	N_h	S_{1h}^2	S_{2h}^2
Grande	600	200	500 000
Pequeña	1 000	10	4 000
	1 600		

La condición de que las estimaciones de Y_1 y Y_2 no estén equivocadas en más del 6% ($P = 0.95$) equivale a las tolerancias

$$V_1 = V(\bar{y}_{1st}) \leq 0.0351 \text{: } V_2 = V(\bar{y}_{2st}) \leq 56.25$$

Demostrar que la asignación óptima para y_1 con $n_1 = 450$, $n_2 = 167$, satisface las dos tolerancias. Obsérvese que en este problema no puede ignorarse la cpf.

5A.14. En muestreo aleatorio estratificado con una unidad por estrato, se supone que los estratos pueden agruparse en pares con $N_{j1} = N_{j2} = N_j$ ($j = 1, 2, \ldots, L/2$). Un método alternativo de muestreo saca dos unidades al azar de cada par de estratos. Demostrar que para este método

$$V(\hat{Y}_{st2}) = \frac{(N_j - 1)}{(2N_j - 1)} \sum_{j=1}^{L/2} \left[2N_j(N_j - 1) \sum_{i=1}^{2} S_{ji}^2 + (Y_{j1} - Y_{j2})^2 \right]$$

Por lo tanto, demostrar que el valor esperado de la estimación $v_1(\hat{Y}_{st})$ de estratos contraídos en la Fórmula (5A.54), Sec. 5A.12, sobreestima $V(\hat{Y}_{st2})$, la varianza que se aplicaría si se usaran estratos dos veces más grandes.

CAPITULO 6

Estimadores de Razón

6.1 METODOS DE ESTIMACION

Uno de los rasgos de la estadística teórica es la creación de una vasta teoría que discute los métodos de obtención de buenas estimaciones a partir de los datos. En el desarrollo de la teoría, específicamente para encuestas de muestreo, se han utilizado poco estos conocimientos, por dos causas principales. Primero, en las encuestas que contienen un gran número de atributos, es una gran ventaja, aunque se disponga de máquinas computadoras, el poder utilizar procedimientos de estimación que requieran poco más que simples sumas, en tanto que los métodos superiores de estimación de la estadística teórica, como son la máxima verosimilitud, podrían necesitar una serie de aproximaciones sucesivas antes de encontrar la estimación. En segundo lugar, como se dijo en la Sec. 1.4, son diferentes las actitudes tomadas en estas dos divisiones de investigación. La mayoría de los metodos de investigación de la estadística teórica suponen que se conoce la forma funcional de la distribución de frecuencia que sigue a los datos de la muestra, y el método de estimación está cuidadosamente engranado de acuerdo a este tipo de distribución. En la teoría de encuestas por muestreo se ha preferido hacer, cuando más, algunos supuestos respecto a esta distribución de frecuencia (que es muy asimétrica, o que es bastante simétrica) y dejar fuera de la discusión su forma funcional específica. Esta preferencia conduce a la utilización de métodos simples de estimación que dan buenos resultados en cierta gama de tipos de distribuciones de frecuencia. Esta actitud resulta razonable para tratar encuestas en las que el tipo de distribución puede variar de un atributo a otro, y cuando no deseamos detenernos a examinarlas todas, antes de decidir cómo hacer cada estimación.

En consecuencia, actualmente, las técnicas de estimación para el trabajo de encuestas por muestreo son de alcances restringidos. Ahora consideraremos dos técnicas, el método de razón en este capítulo y el método de regresión lineal en el Cap. 7.

6.2 EL ESTIMADOR DE RAZON

En el método de razón se obtiene una variable auxiliar x_i, correlacionada con y_i, para cada unidad de la muestra. El total de población X de las x_i debe conocerse. En la práctica, x_i suele ser el valor de y_i en alguna ocasión anterior cuando se tomó un censo completo. La finalidad de este método es obtener un incremento de precisión, al utilizar la correlación entre y_i y x_i. Por el momento, suponemos muestreo aleatorio simple.

La estimación de razón de Y, el total de la población para los y_i, es

$$\hat{Y}_R = \frac{y}{x}X = \frac{\bar{y}}{\bar{x}}X \tag{6.1}$$

donde y, x son los totales de muestra de y_i y x_i respectivamente.

Si x_i es el valor de y_i en una ocasión anterior, el método de la razón utiliza la muestra para estimar el cambio relativo Y/X ocurrido desde la ocasión anterior. El cambio relativo estimado y/x se multiplica por el total de población X conocido anteriormente, para proporcionar una estimación del total de la población actual. Si la razón y_i/x_i es casi la misma en todas las unidades del muestreo, los valores de y/x varían poco entre las muestras y la estimación de razón es altamente precisa. En otra aplicación, x_i puede ser el número total de acres de una finca y y_i el número total de acres dedicados a cierto cultivo. La estimación de razón será exitosa en este caso si todos los agricultores dedican más o menos el mismo porcentaje de su propiedad a dicho cultivo.

Si la cantidad que se ha de estimar es $\bar{Y}$, el valor de la media de la población de y_i, la estimación de razón es

$$\hat{\bar{Y}}_R = \frac{y}{x}\bar{X}$$

A menudo se deseará estimar una razón y no un total o una media, por ejemplo, la razón de acres de maíz al de acres de trigo, la razón de los gastos debidos a la fuerza de trabajo, a los gastos totales, o la razón de activos corrientes a la de activos totales. La estimación muestral es $\hat{R} = y/x$. En este caso no es necesario conocer X. El uso de estimaciones de razón para este propósito ya se discutió en la Sec. 2.11 y (con muestreo por conglomerados para proporciones) en la Sec. 3.12.

Ejemplo. La Tabla 6.1 muestra el número de habitantes (en miles) en cada una de las 49 ciudades extraídas en una muestra aleatoria simple de la población de 196 grandes ciudades discutidas en la Sec. 2.15. El problema es estimar el número total de habitantes en las 196 ciudades en 1930. Se supone que se conoce el verdadero total para 1920, X, su valor es 22 919.

El ejemplo es conveniente para una estimación de razón. La mayoría de las ciudades en la muestra presentan un aumento en tamaño de 1920 a 1930 del orden de 20%. De los datos de la muestra tenemos

$$y = \sum y_i = 6\,262, \qquad x = \sum x_i = 5054$$

Consecuentemente la estimación de razón del total de 1930 para todas las 196 ciudades es

$$\hat{Y}_R = \frac{y}{x}X = \frac{6\,262}{5\,054}(22\,919) = 28\,397$$

La estimación correspondiente basada en la media de la muestra por ciudad es

$$\hat{Y} = N\bar{y} = \frac{(196)(6262)}{49} = 25\,048$$

El total correcto en 1930 es 29 351.

TABLA 6.1. TAMAÑOS DE 49 GRANDES CIUDADES DE ESTADOS UNIDOS (en miles) EN 1920 (x_i) Y 1930 (y_i)

x_i	y_i	x_i	y_i	x_i	y_i
76	80	2	50	243	291
138	143	507	634	87	105
67	67	179	260	30	111
29	50	121	113	71	79
381	464	50	64	256	288
23	48	44	58	43	61
37	63	77	89	25	57
120	115	64	63	94	85
61	69	64	77	43	50
387	459	56	142	298	317
93	104	40	60	36	46
172	183	40	64	161	232
78	106	38	52	74	93
66	86	136	139	45	53
60	57	116	130	36	54
46	65	46	53	50	58
				48	75

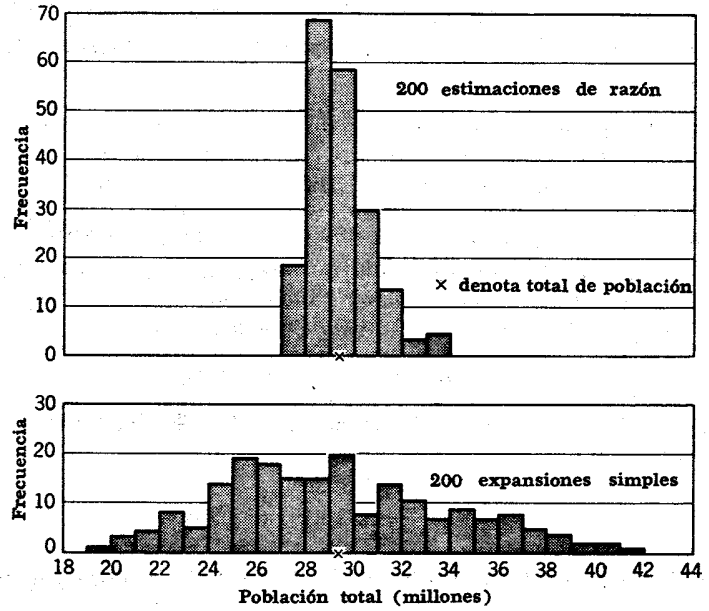


FIG. 6.1. Comparación experimental de la estimación de razón con la estimación basada en la media de la muestra

La Fig. 6.1 muestra la estimación de razón y la estimación basada en la media de la muestra por ciudad para cada una de 200 muestras aleatorias simples de tamaño 49, extraídas de esta población. Es evidente un aumento sustancial en precisión con el método de razón.

6.3 VARIANZA APROXIMADA DE LA ESTIMACION DE RAZON

La distribución de la estimación de razón ha sido siempre un problema porque tanto y como x varían de muestra a muestra. Los resultados teóricos conocidos no corresponden a lo que quisiéramos saber para aplicaciones prácticas. Los resultados principales se exponen primero sin prueba.

La estimación de razón es consistente (esto es obvio). Tiene sesgo, excepto para algunos tipos especiales de poblaciones, aunque el sesgo es despreciable en muestras grandes. La distribución límite de la estimación de razón, conforme n se hace muy grande, es normal, sujeta a algunas restricciones menores sobre el tipo de población de la cual estamos muestreando. En muestras de tamaño moderado la distribución presenta una tendencia hacia asimetría positiva en las clases de poblaciones para las cuales se usa el método

más a menudo. No poseemos fórmulas exactas para el sesgo y la varianza de muestreo de la estimación sino sólo aproximaciones que son válidas en muestras grandes.

Estos resultados equivalen a decir que no hay dificultad si la muestra es lo suficientemente grande para que (a) la razón sea casi normalmente distribuida y (b) sea válida la fórmula de la muestra grande para su varianza. Como una regla de operación, los resultados de la muestra grande pueden usarse si el tamaño de la muestra excede a 30 y es también lo suficientemente grande de tal manera que los coeficientes de variación de $\bar{x}$ y $\bar{y}$ sean ambos menores que el 10%.

Teorema 6.1. Las estimaciones de razón del total Y de la población, la media $\bar{Y}$ de la población y la razón Y/X de la población, son, respectivamente

$$\hat{Y}_R = \frac{\bar{y}}{\bar{x}}X, \qquad \hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}}\bar{X}, \qquad \hat{R} = \frac{\bar{y}}{\bar{x}}$$

En una muestra aleatoria simple de tamaño n (n grande)

$$V(\hat{Y}_R) \doteq \frac{N^2(1-f)}{n}\left[\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1}\right] \tag{6.2}$$

$$V(\hat{\bar{Y}}_R) \doteq \frac{1-f}{n}\left[\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1}\right] \tag{6.3}$$

$$V(\hat{R}) \doteq \frac{1-f}{n\bar{X}^2}\left[\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1}\right] \tag{6.4}$$

donde $f = n/N$ es la fracción de muestreo. El método que se usa en el Teorema 2.5 muestra que (6.2), (6.3) y (6.4) también son aproximaciones a los errores cuadrados medios del estimador en estas fórmulas.

El argumento que conduce al resultado aproximado (6.4) se dio en el teorema 2.5. Dado que $\hat{\bar{Y}}_R = \bar{X}\hat{R}$, $\hat{Y}_R = N\bar{X}\hat{R}$, los otros dos resultados vienen como consecuencia inmediata.

Corolario 1. Hay varias formas alternativas del resultado. Dado que $\bar{Y} = R\bar{X}$, podemos escribir

$$V(\hat{Y}_R)=\frac{N^2(1-f)}{n(N-1)}\sum_{i=1}^{N}[(y_i-\bar{Y})-R(x_i-\bar{X})]^2$$
$$=\frac{N^2(1-f)}{n(N-1)}[\sum(y_i-\bar{Y})^2+R^2\sum(x_i-\bar{X})^2$$
$$-2R\sum(y_i-\bar{Y})(x_i-\bar{X})]$$

El coeficiente de correlación ρ entre y_i y x_i en la población finita está definido por la ecuación

$$\rho=\frac{E(y_i-\bar{Y})(x_i-\bar{X})}{\sqrt{E(y_i-\bar{Y})^2E(x_i-\bar{X})^2}}=\frac{\sum^{N}(y_i-\bar{Y})(x_i-\bar{X})}{(N-1)S_yS_x}$$

Esto conduce al resultado

$$V(\hat{Y}_R)=\frac{N^2(1-f)}{n}(S_y^2+R^2S_x^2-2R_\rho S_yS_x) \tag{6.5}$$

Una forma equivalente es

$$V(\hat{Y}_R)=(1-f)\frac{Y^2}{n}\left(\frac{S_y^2}{\bar{Y}^2}+\frac{S_x^2}{\bar{X}^2}-\frac{2S_{yx}}{\bar{Y}\bar{X}}\right) \tag{6.6}$$

donde $S_{yx}=\rho S_yS_x$ es la covarianza entre y_i y x_i. Esta relación también puede escribirse como

$$V(\hat{Y}_R)=(1-f)\frac{N^2}{n}(C_{yy}+C_{xx}-2C_{yx}) \tag{6.7}$$

donde C_{yy}, C_{xx} son los cuadrados de los coeficientes de variación (cv) de y_i y x_i, respectivamente, y C_{yx} es la covarianza relativa.

Corolario 2. Dado que $\hat{Y}_R$, $\hat{\bar{Y}}_R$ y $\hat{R}$ difieren solamente por multiplicadores conocidos, el coeficiente de variación (por ejemplo, el error estándar dividido entre la cantidad que se está estimando) es el mismo para las tres estimaciones. De (6.7) el cuadrado de este cv es

$$(\text{cv})^2=\frac{V(\hat{Y}_R)}{Y^2}=\frac{1-f}{n}(C_{yy}+C_{xx}-2C_{yx}) \tag{6.8}$$

La cantidad $(\text{cv})^2$ se ha llamado la *varianza relativa* en Hansen y otros (1953). Su uso evita la repetición de las fórmulas de varianza para cantidades relacionadas como la media o el total de población estimados.

6.4 ESTIMACION DE LA VARIANZA A PARTIR DE UNA MUESTRA

De la Ec. (6.2)

$$V(\hat{Y}_R)\doteq\frac{N^2(1-f)}{n}\frac{\sum^{N}(y_i-Rx_i)^2}{N-1}$$

Como ya se mencionó en la Sec. 2.11, tomamos

$$\frac{\sum^{n}(y_i-\hat{R}x_i)^2}{(n-1)}$$

como estimación muestral de la varianza de población. Esta estimación presenta un sesgo del orden de $1/n$.

Para la varianza estimada, $v(\hat{Y}_R)$, esto da

$$v(\hat{Y}_R)=\frac{N^2(1-f)}{n(n-1)}\sum_{i=1}^{n}(y_i-\hat{R}x_i)^2 \tag{6.9}$$

Este resultado puede expresarse de muchas maneras diferentes. Por ejemplo,

$$v(\hat{Y}_R)=\frac{N^2(1-f)}{n(n-1)}(\sum y_i^2+\hat{R}^2\sum x_i^2-2\hat{R}\sum y_ix_i) \tag{6.10}$$
$$=\frac{N^2(1-f)}{n}(s_y^2+\hat{R}^2s_x^2-2\hat{R}s_{yx}) \tag{6.11}$$

donde $s_{yx}=\sum(y_i-\bar{y})(x_i-\bar{x})/(n-l)$ es la covarianza de muestra entre y_i y x_i.

Existen dos fórmulas alternativas para la estimación de muestra de la varianza. Dado que $\hat{Y}_R=N\bar{X}\hat{R}$, una forma para $\hat{R}$ es

$$v_1(\hat{R})=\frac{(1-f)}{n\bar{X}^2}(s_y^2+\hat{R}^2s_x^2-2\hat{R}s_{yx}) \tag{6.12}$$

Sin embargo, dado que $\hat{R}=\bar{y}/\bar{x}$, la cantidad $\bar{X}$ no es necesariamente conocida y a veces tampoco se conoce al estimar R. Esto sugiere la forma alternativa

$$v_2(\hat{R})=\frac{(1-f)}{n\bar{x}^2}(s_y^2+\hat{R}^2s_x^2-2\hat{R}s_{yx}) \tag{6.13}$$

Esta forma también podría usarse para $v(\hat{Y}_R)$, al tomar $v_2(\hat{Y}_R)=X^2v_2(\hat{R})$.

Esto plantea la siguiente pregunta: Si se conoce $\bar{X}$, ¿es preferible v_1 a v_2? La respuesta no la sabemos claramente aún. P.S.R.S.

Rao y J.N.K. Rao (1971) estudiaron analíticamente los sesgos en v_1 y v_2 en una población infinita por medio del modelo de regresión lineal

$$y_i = \alpha + \beta x_i + e_i$$

con

$$E(e_i|x_i) = 0,\ V(e_i|x_i) = \delta x_i^t,\ E(e_i e_j|x_i x_j) = 0,$$

donde $0 \leq t \leq 2$ y x_i tiene un rango de distribución $ax^{h-1}\ e^{-x}$. Se estudió el intervalo $0 \leq t \leq 2$ porque en las aplicaciones, la varianza residual de y_i se cree que aumenta con x_i a tasas diferentes, en diversas poblaciones. Ellos encontraron que v_2 es menos sesgada para $0 \leq t\ 3/2$, pero también, menos estable para $t = 0$ o $t = 1$.

6.5 LIMITES DE CONFIANZA

Si la muestra es lo suficientemente grande, de modo que se aplique la aproximación normal, se pueden obtener límites de confianza para Y y R.

$$Y: \hat{Y}_R \pm z\sqrt{v(\hat{Y}_R)} \tag{6.14}$$

$$R: \hat{R} \pm z\sqrt{v(\hat{R})} \tag{6.15}$$

donde z es el desvío normal correspondiente a la probabilidad de confianza elegida.

En la Sec. 6.3 se sugirió que la aproximación normal se aplica razonablemente bien si el tamaño de la muestra es al menos de 30 y si es lo suficientemente grande para que los cv de $\bar{y}$ y $\bar{x}$ sean ambos menores que 0.1. Cuando no se reúnen estas condiciones, la fórmula para $v(\hat{R})$ tiende a dar valores que son demasiado bajos y puede volverse notoria la asimetría positiva en la distribución de $\hat{R}$.

Un método alternativo para calcular límites de confianza se ha usado en ensayos biológicos (Fieller, 1932; Paulson, 1942). El método requiere que $\bar{y}$ y $\bar{x}$ sigan una distribución normal bivariable, de tal forma que $(\bar{y} - R\bar{x})$ esté normalmente distribuida. Resulta entonces que en una muestra aleatoria simple la cantidad

$$\frac{\bar{y} - R\bar{x}}{\sqrt{[(N-n)/Nn]}\sqrt{s_y^2 + R^2 s_x^2 - 2Rs_{yx}}} \tag{6.16}$$

está casi normalmente distribuida con media cero y desviación estándar unitaria.

No se conoce el valor de R, pero cualquier considerado valor de R que haga este desvío normal lo suficientemente grande puede considerarse como descartado por los datos de la muestra. Consecuentemente, los límites de confianza para R se encuentran al hacer (6.16) igual a $\pm z$ y despejar R de la ecuación cuadrática resultante. Los límites de confianza son aproximados, porque, si tratamos de comprobarlos al muestrear repetidamente una población fija con R conocida, resultan algunos valores de $\bar{y}$ y $\bar{x}$ para los cuales las dos raíces de la ecuación son imaginarias. Tales casos se vuelven raros si los cv de $\bar{y}$ y $\bar{x}$ son menores de 0.3.

Después de algunas transformaciones, las dos raíces pueden expresarse como

$$R = \hat{R}\frac{(1 - z^2 c_{\bar{y}\bar{x}}) \pm z\sqrt{(c_{\bar{y}\bar{y}} + c_{\bar{x}\bar{x}} - 2c_{\bar{y}\bar{x}}) - z^2(c_{\bar{y}\bar{y}}c_{\bar{x}\bar{x}} - c_{\bar{y}\bar{x}}^2)}}{1 - z^2 c_{\bar{x}\bar{x}}} \tag{6.17}$$

donde

$$c_{\bar{y}\bar{y}} = \frac{N-n}{Nn}\frac{s_y^2}{\bar{y}^2}$$

es el cuadrado del cv estimado de $\bar{y}$, con definiciones análogas de $c_{\bar{y}\bar{x}}$ y $c_{\bar{x}\bar{x}}$. Si $z^2 c_{\bar{y}\bar{y}}$, $z^2 c_{\bar{x}\bar{x}}$ y $z^2 c_{\bar{y}\bar{x}}$ son todas relativamente pequeñas comparadas con 1, los límites se reducen a

$$R = \hat{R} \pm z\sqrt{c_{\bar{y}\bar{y}} + c_{\bar{x}\bar{x}} - zc_{\bar{y}\bar{x}}}$$

Esta expresión es la misma que la aproximación normal (6.15). Aun con normalidad bivariada, los límites de Fieller han sido criticados por no ser lo suficientemente conservadores. Wilkinson, James y Venables (1975) explican la dificultad y dan un método alternativo.

6.6 COMPARACION DE LA ESTIMACION DE RAZON CON LA MEDIA POR UNIDAD

El tipo de estimación de Y que se estudió en capítulos anteriores es $N\bar{y}$, donde $\bar{y}$ es la media por unidad para la muestra (en muestreo aleatorio simple) o una media ponderada por unidad (en muestreo aleatorio estratificado). Estimaciones de esta clase son llamadas estimaciones basadas en la *media por unidad* o estimaciones obtenidas por *expansión simple*.

Teorema 6.2. En muestras grandes, con muestreo aleatorio simple, la estimación de razón $\hat{Y}_R$ tiene una varianza menor que la estimación $\hat{Y} = N\bar{y}$ obtenida por expansión simple, si

$$\rho > \frac{1}{2}\left(\frac{S_x}{\bar{X}}\right)\Big/\left(\frac{S_y}{\bar{Y}}\right) = \frac{\text{coeficiente de variación } x_i}{2(\text{coeficiente de variación de } y_i)}$$

Prueba. Para $\hat{Y}$ tenemos

$$V(\hat{Y}) = \frac{N^2(1-f)}{n} S_y^2$$

Para la estimación de razón tenemos de (6.5)

$$V(\hat{Y}_R) = \frac{N^2(1-f)}{n}(S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x)$$

Por lo tanto, la estimación de razón tiene una varianza más pequeña si

$$S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x < S_y^2$$

Si $R = \bar{Y}/\bar{X}$ es positivo, la condición se convierte en

$$\rho > \frac{RS_x}{2S_y} = \frac{1}{2}\left(\frac{S_x}{\bar{X}}\right)\Big/\left(\frac{S_y}{\bar{Y}}\right) = \frac{1}{2}\,\frac{\text{cv}(x)}{\text{cv}(y)} \qquad (6.18)$$

6.7 CONDICIONES BAJO LAS CUALES EL ESTIMADOR DE RAZON ES UN ESTIMADOR INSESGADO LINEAL OPTIMO

Un resultado bien conocido de la teoría de la regresión indica el tipo de población en el cual la estimación de razón puede considerarse la mejor dentro de una amplia clase de estimaciones. El resultado fue demostrado primero para poblaciones infinitas. Brewer (1963*b*) y Royall (1970*a*) ampliaron los resultados a poblaciones finitas. El resultado es válido si se satisfacen dos condiciones.

1. La relación entre y_i y x_i es una recta que pasa por el origen.
2. La varianza de y_i alrededor de esta recta es proporcional a x_i.

Se definirá ahora "un estimador óptimo insesgado lineal". Considérense todos los estimadores $\hat{Y}$ de Y como funciones lineales de los valores de muestra y_i, es decir, son de la forma

$$l_1 y_1 + l_2 y_2 + \cdots + l_n y_n$$

donde los l no dependen de y_i aunque pueden ser funciones de los x_i. La elección de los números l se restringe a aquellos que proporcionan una estimación insesgada de Y. El estimador con la varianza más pequeña se llama el mejor estimador lineal insesgado (BLUE = best linear unbiased estimator).

Formalmente, Brewer y Royall, suponen que los N valores de población $(y_i\ x_i)$ son una muestra aleatoria de una suprapoblación en la que

$$y_i = \beta x_i + \varepsilon_i \qquad (6.19)$$

donde los ε_i son independientes de los x_i y $x_i > 0$. En arreglos en donde x_i es fijo, ε_i tiene media cero y varianza λx_i. Los x_i $(i = 1, 2, \ldots, N)$ son conocidos.

En la teoría de la aleatorización utilizada en este libro hasta ahora, el total de la población finita Y se ha considerado como una cantidad fija. Por otro lado, según el modelo (6.19) tenemos que $Y = \beta X + \sum^N \varepsilon_i$ es una variable aleatoria. Al definir un estimador insesgado de acuerdo a este modelo, Brewer y Royall, utilizan un concepto de insesgado que difiere del de la teoría. Ellos consideran que un estimador $\hat{Y}$ es insesgado si $E(\hat{Y}) = E(Y)$ en selecciones repetidas de la población finita y la muestra en el modelo. Un estimador tal puede llamarse *insesgado según modelo.*

Teorema 6.3. En el modelo (6.19) el estimador de razón $\hat{Y}_R = X\bar{y}/\bar{x}$ es un estimador insesgado lineal óptimo para cualquier muestra aleatoria o no, elegida exclusivamente de acuerdo con los valores de x_i.

Prueba. Como $E(\varepsilon_i|x_i) = 0$ en muestreo repetido, de (6.19) tenemos que

$$Y = \beta X + \sum^N \varepsilon_i : E(Y) = \beta X \qquad (6.20)$$

Además, con el modelo (6.19) cualquier estimador lineal $\hat{Y}$ es de la forma

$$\hat{Y} = \sum^n l_i y_i = \beta \sum^n l_i x_i + \sum^n l_i \varepsilon_i \qquad (6.21)$$

Si conservamos fijos los n valores de muestra x_i al repetir el muestreo en el modelo (6.19),

$$E(\hat{Y}) = \beta \sum^n l_i x_i : V(\hat{Y}) = \lambda \sum^n l_i^2 x_i \qquad (6.22)$$

De (6.20) y (6.22), $\hat{Y}$ es claramente insesgado según modelo si $\sum^n l_i x_i \equiv X$. Al minimizar $V(\hat{Y})$ en estas condiciones por un multiplicador de Lagrange se obtiene

$$2l_i\, x_i = cx_i : l_i = \text{constante} = X/n\bar{x} \qquad (6.23)$$

La constante del ser el valor $X/n\bar{x}$ para satisfacer la condición del insesgado $l\sum x_i = X$. Por tanto el estimador BLUE, $\hat{Y}$ es $n\bar{y}X/n\bar{x} = X\bar{y}/\bar{x} = \hat{Y}_R$, el estimador de razón usual. Esto termina la prueba.

Además, de (6.20) y (6.21) con $l = X/n\bar{x}$,

$$\hat{Y}_R - Y = \sum^{n} l_i\varepsilon_i - \sum^{N} \varepsilon_i = (X/n\bar{x})(\sum^{n} \varepsilon_i) - \sum^{N} \varepsilon_i \tag{6.24}$$

$$= \frac{(X-n\bar{x})}{n\bar{x}} \sum^{n} \varepsilon_i - \sum^{N-n} \varepsilon_i \tag{6.25}$$

donde $\sum^{N-n}$ denota la suma sobre los $(N-n)$ valores de población ajenos a la muestra. Por tanto

$$V(\hat{Y}_R) = \frac{\lambda(X-n\bar{x})^2(n\bar{x})}{(n\bar{x})^2} + \lambda(X-n\bar{x}) = \frac{\lambda(X-n\bar{x})X}{n\bar{x}} \tag{6.26}$$

Fácilmente se ve que un estimador insesgado según modelo de λ para esta muestra es

$$\hat{\lambda} = \sum^{n} \frac{1}{x_i}(y_i - \hat{R}x_i)^2/(n-1) \tag{6.27}$$

donde $\hat{R} = \bar{y}/\bar{x}$, como de costumbre. Este valor se puede sustituir en (6.26) para dar una estimación muestral insesgada según modelo de $V(\hat{Y}_R)$.

La importancia práctica de estos resultados es que sugieren las condiciones en las que el estimador de razón es superior no sólo a $\bar{y}$ sino que además, es el mejor de toda una clase de estimadores. Cuando tratamos de decidir la clase de estimación a usarse, puede ser útil una gráfica en donde se marquen los valores de y_i contra los de x_i. Si esta gráfica muestra una relación rectilínea que pase por el origen y si la varianza de los puntos y_i alrededor de la línea parecen ser más o menos proporcionales a x_i, será difícil obtener algo mejor que el estimador de razón.

En ocasiones, la varianza de los y_i en arreglos donde x_i es fijo, *no* es proporcional a x_i. Si esta varianza residual es de la forma $\lambda v(x_i)$, donde se conoce $v(x_i)$, Brewer y Royall demostraron que el estimador BLUE se convierte en

$$\hat{Y} = X\frac{\sum^{n} w_i y_i x_i}{\sum^{n} w_i x_i^2} \tag{6.28}$$

donde $w_i = 1/v(x_i)$. En una muestra de población tomada en Grecia, Jessen y otros (1947) juzgaron que la varianza residual aumentaba

más o menos como x_i^2. Esto sugiere una regresión ponderada con $w_i = 1/x_i^2$ que da

$$\hat{Y} = \frac{X\left(\sum^{n} w_i y_i x_i\right)}{\sum^{n} (w_i x_i^2)} = \frac{X}{n}\sum^{n}\left(\frac{y_i}{x_i}\right) \tag{6.29}$$

Para una población y n dadas, $V(\hat{Y}_R)$ en (6.26) está claramente minimizada, para cada $x_i > 0$, cuando la muestra consiste de los n *más grandes* x_i de la población. En 16 pequeñas poblaciones naturales del tipo al que se han aplicado las estimaciones de razón, Royall (1970) encontró que para muestras con $n = 2$ hasta 12, la selección de los n más grandes x_i generalmente incrementaba la exactitud de $\hat{Y}_R$.

En resumen los resultados de Brewer-Royall muestran que la suposición de un cierto tipo de modelo, conduce a un estimador de razón insesgado y a las fórmulas para $V(\hat{Y}_R)$ y $v(\hat{Y}_R)$ que resultan simples y exactas para $n > 1$. Los resultados podrían usarse en la práctica en los casos donde el examen de los pares y, x a partir de los datos disponibles sugieren que el modelo es razonablemente correcto. Las fórmulas de la varianza (6.26) y (6.27) parecen ser sensibles a la inexactitud en este modelo, aunque esto requiere más estudio.

En trabajos posteriores de Royall y Herson (1973) se discute el tipo de distribución muestral requerido con respecto a los x_i, para que $\hat{Y}_R$ se mantenga insesgado cuando hay una regresión polinomial de y_i respecto a x_i.

6.8 SESGO DE LA ESTIMACION DE RAZON

En general, la estimación de razón tiene un sesgo de orden $1/n$. Como el error estándar de la estimación es del orden $1/\sqrt{n}$, la cantidad (sesgo/error estándar) también es del orden $1/\sqrt{n}$ y se vuelve despreciable al crecer considerablemente n. En la práctica, esta cantidad es de poca importancia para muestras de tamaño moderado. Sin embargo, su valor en pequeñas muestras es interesante en el muestreo estratificado con muchos estratos, en donde quisiéramos calcular y examinar las estimaciones de razón en estratos individuales con pequeñas muestras en los estratos. A continuación, presentamos dos resultados útiles respecto al sesgo.

El primero proporciona el término guía del sesgo, cuando se le desarrolla en una serie de Taylor.

$$\hat{R}-R=\frac{\bar{y}}{\bar{x}}-R=\frac{\bar{y}-R\bar{x}}{\bar{x}}$$

escriba

$$\frac{1}{\bar{x}}=\frac{1}{\bar{X}+(\bar{x}-\bar{X})}=\frac{1}{\bar{X}}\left(1+\frac{\bar{x}-\bar{X}}{\bar{X}}\right)^{-1}\doteq\frac{1}{\bar{X}}\left(1-\frac{\bar{x}-\bar{X}}{\bar{X}}\right) \tag{6.30}$$

De modo que,

$$\hat{R}-R\doteq\frac{\bar{y}-R\bar{x}}{\bar{X}}\left(1-\frac{\bar{x}-\bar{X}}{\bar{X}}\right) \tag{6.31}$$

Ahora

$$E(\bar{y}-R\bar{x})=\bar{Y}-R\bar{X}=0$$

por lo tanto, el término guía del sesgo proviene del segundo término entre corchetes. Además,

$$E\bar{y}(\bar{x}-\bar{X})=E(\bar{y}-\bar{Y})(\bar{x}-\bar{X})=\frac{1-f}{n}\rho S_y S_x \tag{6.32}$$

por el Teorema 2.3 (Pág. 49) y la definición de ρ. También,

$$E\bar{x}(\bar{x}-\bar{X})=E(\bar{x}-\bar{X})^2=\frac{1-f}{n}S_x^2$$

En consecuencia, el término guía del sesgo es

$$E(\hat{R}-R)\doteq\frac{1-f}{n\bar{X}^2}(RS_x^2-\rho S_y S_x) \tag{6.33}$$

$$\doteq\frac{1-f}{n}(C_{xx}-C_{yx})R \tag{6.34}$$

Para una justificación rigurosa de (6.34) véase David y Sukhatme (1974).
Ahora, el término guía en $V(\hat{R})$ es

$$V(\hat{R})=\frac{(1-f)}{n\bar{X}^2}(S_y^2+R^2S_x^2-2R\rho S_y S_x) \tag{6.35}$$

a partir de (6.5) al sustituir $\hat{R}=\hat{Y}_R/N\bar{X}$.

De (6.33) y (6.35) el término guía de la cantidad (sesgo/error estándar), que es el mismo para $\hat{R}$, $\hat{Y}_R$, y $\hat{\bar{Y}}_R$, puede expresarse como

$$\frac{(\text{sesgo})}{\text{e.e.}}=\text{cv}\,(\bar{x})\frac{(RS_x-\rho S_y)}{(R^2S_x^2-2R\rho S_y S_x+S_y^2)^{1/2}} \tag{6.36}$$

donde cv $(\bar{x})=\sqrt{1-f}S_x/\sqrt{n}\bar{X}$. Al sustituir estimaciones muestrales de los términos de (6.36), Kish, Namboodiri y Pillai (1962) calcularon los valores (sesgo/error estándar) para muchos atributos de varios estudios nacionales y más localizados. En los estudios nacionales, casi todos los valores (sesgo/error estándar) eran <0.03, y casi los únicos valores >0.10 en sus estudios provienen de un solo estrato con $n_h = 6$ pequeños hospitales.

El segundo resultado de Hartley y Ross (1954), da una respuesta exacta para el sesgo y un límite superior de la relación sesgo a error estándar. Consideremos la covarianza en muestras aleatorias simples de tamaño n, de las cantidades $\hat{R}$ y $\bar{x}$. Tenemos

$$\text{cov}\,(\hat{R},\bar{x})=E\left(\frac{\bar{y}}{\bar{x}}\cdot\bar{x}\right)-E(\hat{R})\,E(\bar{x}) \tag{6.37}$$

$$=\bar{Y}-\bar{X}E(\hat{R}) \tag{6.38}$$

Así que

$$E(\hat{R})=\frac{\bar{Y}}{\bar{X}}-\frac{1}{\bar{X}}\text{cov}\,(\hat{R},\bar{x})=R-\frac{1}{\bar{X}}\text{cov}\,(\hat{R},\bar{x}) \tag{6.39}$$

De modo que el sesgo en $\hat{R}$ es —cov $(\hat{R},\bar{x})/\bar{X}$. A diferencia de la aproximación de Taylor al sesgo (6.33), esta expresión es exacta. además,

$$|\text{sesgo en }\hat{R}|=\frac{|\rho_{\hat{R},\bar{x}}\sigma_{\hat{R}}\sigma_{\bar{x}}|}{\bar{X}}$$

$$\leq\frac{\sigma_{\hat{R}}\sigma_{\bar{x}}}{\bar{X}}$$

puesto que $\hat{R}$ y $\bar{x}$ no pueden tener una correlación mayor que 1. Por lo tanto,

$$\frac{|\text{sesgo en }\hat{R}|}{\sigma_{\hat{R}}}\leq\frac{\sigma_{\bar{x}}}{\bar{X}}=\text{cv de }\bar{x} \tag{6.40}$$

Desde luego, se aplica la misma cuota al sesgo en $\hat{Y}_R$ y $\hat{\bar{Y}}_R$. De modo que si el cv de $\bar{x}$ es menor que 0.1 se puede considerar despreciable el sesgo al compararlo con el error estándar.

6.9 EXACTITUD DE LAS FORMULAS PARA LA VARIANZA Y LA VARIANZA ESTIMADA

Con pequeñas muestras, digamos $n < 30$ y C_{xx} grande, se ha sospechado durante mucho tiempo que las fórmulas de muestras grandes dadas para $V(\hat{R})$ y $v(\hat{R})$ son subestimaciones. Por una expansión en serie de Taylor, Sukhatme (1954) expresó el error en

$V(\hat{R})$ en términos de los momentos bivariados de y y x. Desafortunadamente, el resultado es demasiado complicado para conducir a un método práctico en las aplicaciones.

Si y y x siguen una distribución normal bivariada, los resultados de Sukhatme se simplifican en forma considerable. Sea

$$V_1 = (C_{yy} + C_{xx} - 2C_{yx})/n$$

una notación para la primera aproximación a la varianza relativa de $\hat{R}$, al ignorar la cpf. Para los términos de orden $1/n^2$,

$$E\left(\frac{\hat{R}-R}{R}\right)^2 \doteq V_1\left(1+\frac{3C_{xx}}{n}+\frac{6C_{xx}}{n}\cdot\frac{\rho^2 C_{yy}+C_{xx}-2C_{yx}}{C_{yy}+C_{xx}-2C_{yx}}\right) \tag{6.41}$$

Como el término entre paréntesis de la derecha es menor que $6C_{xx}/n$, esto da

$$E\left(\frac{\hat{R}-R}{R}\right)^2 < V_1\left(1+\frac{9C_{xx}}{n}\right) \tag{6.42}$$

para los términos de orden $1/n^2$. Ahora bien, C_{xx}/n es el cuadrado del coeficiente de variación $\bar{x}$. De modo que si n es suficientemente grande para que el cv de $\bar{x}$ sea menor que 0.1, el uso de V_1 no debería subestimar en más del 9%. En la práctica, el multiplicador 9 de (6.42) aparece indebidamente alto comparado con (6.41), por ejemplo, si $C_{xx} = C_{yy}$, (6.41) se reduce a

$$E\left(\frac{\hat{R}-R}{R}\right)^2 \doteq V_1\left[1+\frac{C_{xx}}{n}(6-3\rho)\right] \tag{6.43}$$

Como ρ es casi siempre positivo en las aplicaciones del método de razón, resultará más representativo un multiplicador entre 3 y 6. Sin embargo, los efectos de la no normalidad en y y x también entran en el término de orden $1/n^2$.

A partir de un estudio con el método de Montecarlo efectuado por Rao (1968) en pequeñas poblaciones naturales, se mencionarán algunos resultados ilustrativos de los sesgos de las fórmulas para grandes muestras para $V(\hat{R})$ y $v_1(\hat{R})$ y para pequeñas muestras aleatorias de ocho poblaciones, cada una con $N > 30$. Las poblaciones están descritas en Rao (1969). Las fórmulas para $V(\hat{R})$ y $v_1(\hat{R})$, dadas en (6.4) y (6.12) se apreciarán como estimadores del verdadero $ECM(\hat{R})$

Para una población dada, las cantidades 100 $[ECM(\hat{R}) - V(\hat{R})]/ECM(\hat{R})$ y 100 $[ECM(\hat{R}) - Ev_1(\hat{R})]/ECM(\hat{R})$ son el porcentaje de subestimación del verdadero ECM $(\hat{R})$. Los promedios de estos porcentajes en las ocho poblaciones son las que aparecen en la Tabla 6.2.

TABLA 6.2. PORCENTAJE PROMEDIO DE LA SUBESTIMACIÓN DE ECM $(\hat{R})$

Estimador	n = 4	6	8	12
$V(\hat{R})$ en (6.4)	14	14	14	12
$v_1(\hat{R})$ en (6.12)	31	23	21	18

En estos datos, el porcentaje de subestimación en $V(\hat{R})$ declina escasamente al crecer n. Esto se explica en parte por la circunstancia de que en una población $V(\hat{R})$ dio una *sobreestimación* que declinó con n. Las indicaciones a partir de estos resultados son que los sesgos en $v_1(\hat{R})$ son mucho más serios en muestras pequeñas que los que hay en $\hat{R}$ misma, y son insatisfactorios cuando menos hasta $n = 12$. Para $n = 4$, Koop (1968) encontró subestimaciones en $v_1(\hat{R})$ que promediaban 25% en tres poblaciones con $N = 20$.

Un estimador alternativo de $v_1(\hat{R})$ que parece prometedor en cuanto a la reducción del sesgo se discutirá en la Sec. 6.17.

6.10 ESTIMACIONES DE RAZON EN MUESTREO ALEATORIO ESTRATIFICADO

Hay dos maneras de hacer una estimación de razón del total de población Y. Una de ellas es una estimación de razón *separada* del total de cada estrato y la suma de estos totales. Si y_h, x_h son los totales de muestra en el estrato h-ésimo y X_h es el total de estrato de los x_{hi}, esta estimación $\hat{Y}_{Rs}$ (s por *separada*) es

$$\hat{Y}_{Rs} = \sum_h \frac{y_h}{x_h} X_h = \sum_h \frac{\bar{y}_h}{\bar{x}_h} X_h \tag{6.44}$$

No se supone que la verdadera razón permanece constante al pasar de un estrato al otro. La estimación requiere del conocimiento de los totales separados X_h.

Teorema 6.4. Si una muestra aleatoria simple se saca en cada estrato y los tamaños de la muestra son grandes en todos los estratos,

$$V(\hat{Y}_{Rs}) = \sum_h \frac{N_h^2(1-f_h)}{n_h}(S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h\rho_h S_{yh}S_{xh}) \tag{6.45}$$

donde $R_h = Y_h/X_h$ es la razón verdadera en el estrato h, y ρ_h se define como antes en cada estrato.

Prueba. Al aplicar la Fórmula (6.2) Sec. 6.3 para una muestra aleatoria simple, se obtiene en el estrato h,

$$V(\hat{Y}_{Rh}) = \frac{N_h^2(1-f_h)}{n_h}(S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h\rho_h S_{yh}S_{xh}) \tag{6.46}$$

Como $\hat{Y}_{Rs} = \sum_h \hat{Y}_{Rh}$ y el muestreo es independiente en cada estrato, $V(\hat{Y}_{Rs}) = \sum V(\hat{Y}_{Rh})$ y se obtiene el resultado (6.45).

Esta fórmula es válida sólo si la muestra en cada estrato es lo bastante grande para que la fórmula de la varianza se aplique a cada estrato. Hay que recordar esta limitación en las aplicaciones prácticas.

Además, cuando los n_h son pequeños y el número de estratos L es grande, el sesgo en $\hat{Y}_{Rs}$ puede no ser despreciable en relación a su error estándar, como lo sugiere a grandes rasgos el siguiente argumento.

En un solo estrato hemos visto que

$$\frac{|\text{sesgo en } \hat{Y}_{Rh}|}{\sigma(\hat{Y}_{Rh})} \leq \text{cv de } \bar{x}_h$$

Si el sesgo tiene el mismo signo en todos los estratos, lo que puede ocurrir, el sesgo en $\hat{Y}_{Rs}$ será aproximadamente L veces el de $\hat{Y}_{Rh}$. Pero el error estándar de $\hat{Y}_{Rs}$ es tan sólo del orden de $\sqrt{L}$ veces el de $\hat{Y}_{Rh}$. Luego la razón

$$\frac{|\text{sesgo en } \hat{Y}_{Rs}|}{\sigma(\hat{Y}_{Rs})}$$

es de orden

$$\sqrt{L}(\text{cv de } \bar{x}_h)$$

Por ejemplo, con 50 estratos y el cv de $\bar{x}_h$ aproximadamente 0.1 en cada estrato, el sesgo en $\hat{Y}_{Rs}$ podría ser tan alto como 0.7 veces su error es estándar. La contribución del sesgo al error cuadrático medio de $\hat{Y}_{Rs}$ sería entonces como de un tercio.

Aunque en la práctica, el sesgo es mucho más pequeño que su cota superior, hay que tener presente el peligro del sesgo con estimación de razón separada si $\sqrt{L}$ (cv de $\bar{x}_h$) excede, digamos, a 0.3.

6.11 LA ESTIMACION DE RAZON COMBINADA

Puede derivarse una estimación alternativa a partir de una sola razón combinada (Hansen, Hurwitz y Gurney, 1946). De los datos de muestras calculamos

$$\hat{Y}_{st} = \sum_h N_h \bar{y}_h, \qquad \hat{X}_{st} = \sum_h N_h \bar{x}_h. \tag{6.47}$$

Estas son las estimaciones estándar de los totales de población Y y X respectivamente, obtenidas de una muestra estratificada. La estimación combinada, $\hat{Y}_{Rc}$ (c indica *combinado*) es

$$\hat{Y}_{Rc} = \frac{\hat{Y}_{st}}{\hat{X}_{st}}X = \frac{\bar{y}_{st}}{\bar{x}_{st}}X \tag{6.48}$$

donde $\bar{y}_{st} = \hat{Y}_{st}/N$, $\bar{x}_{st} = \hat{X}_{st}/N$ son las medias de población estimadas de una muestra estratificada.

La estimación $\hat{Y}_{Rc}$ no requiere un conocimiento de las X_h, sino solamente de X.

La estimación combinada está menos sujeta al riesgo de sesgo que la estimación separada. Usando el procedimiento de Hartley y Ross en la Sec. 6.8, tenemos, al escribir $\hat{R}_c = \bar{y}_{st}/\bar{x}_{st}$,

$$\begin{aligned}\text{cov}(\hat{R}_c, \bar{x}_{st}) &= E\left(\frac{\bar{y}_{st}}{\bar{x}_{st}}\cdot\bar{x}_{st}\right) - E(\hat{R}_c)E(\bar{x}_{st}) \\ &= \bar{Y} - \bar{X}E(\hat{R}_c)\end{aligned} \tag{6.49}$$

Por lo tanto,

$$E(\hat{R}_c) = R - \frac{1}{\bar{X}}\text{cov}(\hat{R}_c, \bar{x}_{st})$$

$$\frac{|\text{sesgo en } \hat{R}_c|}{\sigma_{\hat{R}_c}} = \frac{|\rho_{\hat{R}_c,\bar{x}_{st}}\cdot\sigma_{\bar{x}_{st}}|}{\bar{X}} \leq \text{cv de } \bar{x}_{st}. \tag{6.50}$$

Así, los sesgos en $\hat{R}_c$, y $\hat{Y}_{Rc}$ son despreciables en relación a sus errores estándar siempre que el cv de $\bar{x}_{st}$ sea menor que 0.1.

Teorema 6.5. Si el tamaño total de la muestra n es grande,

$$V(\hat{Y}_{Rc}) = \sum_h \frac{N_h^2(1-f_h)}{n_h}(S_{yh}^2 + R^2 S_{xh}^2 - 2R\rho_h S_{yh}S_{xh}) \tag{6.51}$$

Prueba. Esta sigue el mismo argumento que el Teorema 2.5. En el caso presente la ecuación clave es

$$(\hat{Y}_{Rc} - Y) = \frac{N\bar{X}}{\bar{x}_{st}}(\bar{y}_{st} - R\bar{x}_{st}) \doteq N(\bar{y}_{st} - R\bar{x}_{st}) \tag{6.52}$$

Ahora, considérese la variable auxiliar $u_{hi} = y_{hi} - Rx_{hi}$. El lado derecho de (6.52) es $N\bar{u}_{st}$, donde $\bar{u}_{st}$ es la media ponderada de la variable u_{hi} en una muestra estratificada. Además, la media de población $\bar{U}$ de u_{hi} es cero, ya que $R = \bar{Y}/\bar{X}$.

Por lo tanto, podemos aplicar el Teorema 5.3 a $\bar{u}_{st}$ para la varianza de la media estimada de una muestra aleatoria estratificada. Esto da

$$V(\hat{Y}_{Rc}) = N^2 V(\bar{u}_{st}) = \sum_h \frac{N_h(N_h - n_h)}{n_h} S_{uh}^2 \tag{6.53}$$

donde

$$S_{uh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (u_{hi} - \bar{U}_h)^2$$

$$= \frac{1}{N_h - 1} \sum_{i=1}^{N_h} [(y_{hi} - \bar{Y}_h) - R(x_{hi} - \bar{X}_h)]^2$$

Cuando se expande la ecuación cuadrática, se obtiene el resultado (6.51).

A partir de las Ecs. (6.45) y (6.51) es interesante notar que las varianzas aproximadas de $\hat{Y}_{Rs}$ y $\hat{Y}_{Rc}$ tomen la misma forma general, siendo la diferencia que las razones de población R_h en los estratos individuales en (6.45) se reemplazan por R en (6.51).

6.12 COMPARACION DE LAS ESTIMACIONES SEPARADA Y COMBINADA

Podemos escribir

$$V(\hat{Y}_{Rc}) - V(\hat{Y}_{Rs})$$

$$= \sum_h \frac{N_h^2(1-f_h)}{n_h} [(R^2 - R_h^2) S_{xh}^2 - 2(R - R_h)\rho_h S_{yh} S_{xh}]$$

$$= \sum \frac{N_h^2(1-f_h)}{n_h} [(R - R_h)^2 S_{xh}^2 + 2(R_h - R)(\rho_h S_{yh} S_{xh} - R_h S_{xh}^2)]$$

En situaciones en las cuales es apropiada la estimación de razón, el último término de la derecha es usualmente pequeño. (Desaparece si dentro de cada estrato la relación entre y_{hi} y x_{hi} es una línea recta que pasa por el origen.) Entonces, a menos que R_h sea constante entre estratos, el uso de una estimación de razón separada en cada estrato, la muestra en cada estrato es suficientemente grande de modo que la fórmula aproximada para $V(\hat{Y}_{Rs})$ es válida y el sesgo acumulado que puede afectar $\hat{Y}_{Rs}$ (Sec. 6.10) puede ignorarse. Con sólo una pequeña muestra en cada estrato, se debe recomendar la estimación combinada, a menos que exista una buena evidencia empírica de lo contrario.

Para las estimaciones muestrales de estas varianzas sustituimos en los lugares apropiados las estimaciones de las muestras de R_h y R. Los cuadrados medios de muestra S_{yh}^2 y S_{xh}^2 se sustituyen por las varianzas correspondientes, y la covarianza de la muestra se sustituye por el término $\rho_h S_{yh} S_{xh}$. El cuadrado medio y la covarianza de la muestra se deben calcular separadamente para cada estrato.

Ejemplo. Los datos provienen de un censo de todas las fincas en el Condado de Jefferson, Iowa. En este ejemplo y_{hi} representa acres de maíz y x_{hi} acres de la finca. La población se divide en *dos* estratos, el primero contiene fincas hasta de 160 acres. Se considera una muestra de 100 fincas. Cuando se usa muestreo estratificado, supondremos que se toman 70 fincas del estrato 1 y 30 del estrato 2, siendo esta distribución casi óptima. Los datos se muestran en la Tabla 6.3. Las últimas tres cantidades, Q_h, V_h' y V_h'', son cantidades auxiliares a usar en los cálculos, las dos últimas se definirán más tarde.

TABLA 6.3. DATOS DEL CONDADO DE JEFFERSON, IOWA

Estratos	Tamaño (acres de la finca)	N_h	S_{yh}^2	S_{yxh}	S_{xh}^2	R_h
1	0–160	1580	312	494	2055	0.2350
2	Más de 160	430	922	858	7357	0.2109
Para pob. completa		2010	620	1453	7619	0.2242

Estratos	$\bar{Y}_h$	$\bar{X}_h$	n_h	$Q_h = W_h^2/n_h$	V_h'	V_h''
1	19.40	82.56	70	0.008828	193	194
2	51.63	244.85	30	0.001525	887	907
Para pob. completa	26.30	117.28	100			

Consideramos cinco métodos para estimar la media de población de acres de maíz por finca. Las cpf se ignoran.

1. Muestra aleatoria simple: estimación de la media por finca.

$$V_1 = \frac{S_y^2}{n} = \frac{620}{100} = 6.20$$

2. Muestra aleatoria simple: estimación de razón.

$$V_2 - \frac{1}{n}(S_y^2 + R^2 S_x^2 - 2RS_{yx})$$

$$= \frac{1}{100}[620 + (0.2242)^2(7619) - 2(0.2242)(1453)]$$

$$= 3.51$$

3. Muestra aleatoria estratificada: estimación de la media por finca.

$$V_3 = \sum \frac{W_h^2}{n_h} S_{yh}^2 = \sum Q_h S_{yh}^2 = 4.16$$

4. Muestra aleatoria estratificada: estimación de razón usando una razón separada en cada estrato.

$$V_4 = \sum Q_h (S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h S_{yxh}) = \sum Q_h V_h' = 3.06$$

5. Muestreo aleatorio estratificado: estimación de razón usando una razón combinada.

$$V_5 = \sum Q_h (S_{yh}^2 + R^2 S_{xh}^2 - 2RS_{yxh}) = \sum Q_h V_h'' = 3.10$$

Las precisiones relativas de los diferentes métodos pueden resumirse como sigue:

Método de muestreo	Método de estimación	Precisión relativa
1. Aleatorio simple	Media por finca	100
2. Aleatorio simple	Razón	177
3. Aleatorio estratificado	Media por finca	149
4. Aleatorio estratificado	Razón separada	203
5. Aleatorio estratificado	Razón combinada	200

Los resultados muestran un punto interesante de amplia aplicación. La estratificación por tamaño de las fincas cumple el mismo propósito general que una estimación de razón en la cual, el denominador es el tamaño de la finca. Ambos planes disminuyen el efecto de las variaciones en tamaño de las fincas sobre el error de muestreo de la media estimada de acres de maíz por finca. Por ejemplo, la ganancia en precisión de una estimación de razón es 77% cuando se usa muestreo aleatorio simple, pero es solamente 36% (203 contra 149) cuando se usa muestreo estratificado.

En el diseño de encuestas, puede haber una elección entre introducir un factor dentro de la estratificación o utilizarlo en el método de estimación. La mejor decisión depende de las circunstancias. Los puntos relevantes son: (*a*) algunos factores, por ejemplo, localización geográfica, que se pueden introducir más fácilmente en la estratificación que en el método de estimación; (*b*) el asunto depende de la naturaleza de la relación entre y_i y x_i. Todos los métodos simples de estimación trabajan en forma más efectiva con una relación lineal. Con una relación compleja o discontinua, la estratificación puede ser más efectiva, ya que, si hay suficientes estratos, la estratificación eliminará los efectos de casi cualquier clase de relación entre y_i y x_i. (*c*) Si algunas variables importantes son aproximadamente proporcionales a x_i, pero otras son aproximadamente proporcionales a otra variable z_i, es mejor usar x_i y z_i como denominadores en las estimaciones de razón que estratificar de acuerdo a una de ellas.

6.13 CALCULO ABREVIADO DE LA VARIANZA ESTIMADA

Si $n_h = 2$ en todos los estratos, Keyfitz (1957) proporcionó métodos simples para el cálculo de aproximaciones a las varianzas estimadas de $\hat{Y}_{RC}$ o $\hat{R}_C$ o más generalmente, de funciones de una o más variables de la forma $\hat{Y}_{st}$. Para $\hat{R}_c$ tenemos

$$\hat{R}_c = \frac{\hat{Y}_{st}}{\hat{X}_{st}} = \frac{\sum_h \hat{Y}_h}{\sum_h \hat{X}_h} = \frac{\sum_h \frac{N_h}{2}(y_{h1} + y_{h2})}{\sum_h \frac{N_h}{2}(x_{h1} + x_{h2})} \tag{6.54}$$

El método de Keyfitz utiliza la identidad para $n_h = 2$,

$$2s_{yh}^2 = 2\sum_{i=1}^{2} (y_{hi} - \bar{y}_h)^2 \equiv (y_{h1} - y_{h2})^2 = (dy_h)^2 \tag{6.55}$$

donde $dy_h = (y_{hi} - y_{h2})$. Por lo tanto

$$v(\hat{Y}_h) = \left(\frac{N_h}{2}\right)^2 2(1-f_h)s_{yh}^2 = (1-f_h)(y_{h1}' - y_{h2}')^2 = (1-f_h)(dy_h')^2 \tag{6.56}$$

donde $y'_{hi} = N_h\, y_{hi}/2$. De igual forma, para la estimación muestral de la covarianza,

$$\operatorname{cov}(\hat{Y}_h \hat{X}_h) = (1-f_h)(dy_h')(dx_h') \tag{6.57}$$

Ahora

$$\hat{R}_c - R = \frac{\hat{Y}_{st}}{\hat{X}_{st}} - R \doteq \frac{\hat{Y}_{st} - R\hat{X}_{st}}{X} = \frac{Y}{X}\left(\frac{\hat{Y}_{st}}{Y} - \frac{\hat{X}_{st}}{X}\right) \tag{6.58}$$

Como el muestreo es independiente en estratos diferentes

$$v(\hat{R}_c) = \left(\frac{Y}{X}\right)^2 \sum_h (1-f_h)\left[\left(\frac{dy_h'}{Y}\right)^2 + \left(\frac{dx_h'}{X}\right)^2 - 2\left(\frac{dy_h'}{Y}\right)\left(\frac{dx_h'}{X}\right)\right] \tag{6.59}$$

$$= \left(\frac{Y}{X}\right)^2 \sum_h (1-f_h)\left(\frac{dy_h'}{Y} - \frac{dx_h'}{X}\right)^2 \tag{6.60}$$

Keyfitz (1957) ha extendido este método a los estimadores postestratificados, al muestreo de varias etapas y a la obtención de varianzas de las diferencias de las estimaciones, a partir de encuestas sucesivas en muestras periódicas. Woodruff (1971) da un enfoque

general para los estimadores no lineales, las probabilidades de selección desiguales y las muestras de tamaño n_h en los estratos. Como ilustración del método de Woodruff, considérese una función $f(\hat{\mathbf{Y}})$ donde $\hat{\mathbf{Y}}$ representa el vector, o conjunto, de m variables $\hat{Y}_j = \sum_h \hat{Y}_{jh}$. Con muestreo aleatorio simple en el estrato h, las $\hat{Y}_{jh}$ son de la forma $(N_h/n_h)\sum y_{jhi}$, y la suma se extiende a todas las n_h unidades de muestra del estrato h. Por la aproximación de Taylor,

$$f(\hat{\mathbf{Y}}) - f(\mathbf{Y}) \doteq \sum_j^m \frac{\partial f}{\partial Y_j}(\hat{Y}_j - Y_j) = \sum_j \sum_h \frac{\partial f}{\partial Y_j}(\hat{Y}_{jh} - Y_{jh}) \tag{6.61}$$

El artificio consiste en invertir el orden de la suma y escribir

$$f(\hat{\mathbf{Y}}) - f(\mathbf{Y}) \doteq \sum_h \sum_j \frac{\partial f}{\partial Y_j}(\hat{Y}_{jh} - Y_{jh}) = \sum_h (\hat{U}_h - U_h) \tag{6.62}$$

donde

$$\hat{U}_h = \sum_j \frac{\partial f}{\partial Y_j}\hat{Y}_{jh} = \frac{N_h}{n_h}\sum_i^{n_h}\left(\sum_j \frac{\partial f}{\partial Y_j} y_{jhi}\right) = \sum_i^{n_h} u_{hi} \tag{6.63}$$

Al evaluar las $(\partial f/\partial Y_j)$ en las estimaciones $\hat{Y}_j$, los números u_{hi} se calculan a partir de (6.63) para cada unidad de muestra en el estrato h. Con muestreo aleatorio simple en los estratos, se aplica la fórmula usual para la varianza estimada de la suma $\hat{U}_h$. Por lo tanto, de (6.62), una estimación de muestra aproximada de la varianza de $f(\hat{\mathbf{Y}})$ es

$$v[f(\hat{\mathbf{Y}})] \doteq \sum_h \frac{(N_h - n_h)n_h}{N_h} \frac{\sum_i^{n_h}(u_{hi} - \bar{u}_h)^2}{(n_h - 1)} \tag{6.64}$$

La ventaja de este método es que las covarianzas de las $\hat{Y}_{jh}$ no necesitan calcularse.

Para ilustrar los u_{hi} para el estimador $\hat{R}_c = \hat{Y}/\hat{X}$, es útil escribir $\hat{R}_c = \hat{Y}_1/\hat{Y}_2$, así que

$$f(\mathbf{Y}) = \frac{Y_1}{Y_2}: \quad \frac{\partial f}{\partial Y_2} = \frac{1}{Y_2}: \quad \frac{\partial f}{\partial Y_2} = -\frac{Y_1}{Y_2^2} \tag{6.65}$$

Por lo tanto, a partir de (6.63) con $n_h = 2$,

$$u_{hi} = \frac{N_h}{n_h}\left(\frac{y_{1hi}}{\hat{Y}_2} - \frac{\hat{Y}_1}{\hat{Y}_2^2}y_{2hi}\right) = \frac{N_h}{n_h}\frac{(y_{1hi} - \hat{R}_c y_{2hi})}{\hat{Y}_2} \tag{6.66}$$

en términos de los y_{hi}, x_{hi} originales

$$u_{hi} = \frac{N_h}{n_h}\frac{(y_{hi} - \hat{R}_c x_{hi})}{\hat{X}} \tag{6.67}$$

Su varianza estimada se aproxima como en (6.64).

El método de Keyfitz se usa, por ejemplo, en la encuesta de salud por medio de entrevistas del National Center for Health Statistics. La unidad de muestra es del tipo conglomerado, un condado o grupo de condados vecinos. Cada unidad se divide en segmentos de aproximadamente 9 domicilios, se eligen un promedio de 13 segmentos de una unidad muestreada. Las variables y_{h1}, y_{h2} podrían ser los números de personas que padecen una enfermedad específica y x_{h1}, x_{h2} los números totales de personas en las muestras de las dos unidades del estrato h.

Además de la estratificación geográfica inicial de los condados, se postestratifican las personas de la muestra por edad, sexo y color. Así, en lugar del estimador $\hat{Y}_{Rc} = X\hat{R}_c$ del número total de enfermos, el estimador es

$$\hat{Y}_{PS} = \sum_a X_a \hat{R}_a = \sum_a X_a \frac{\hat{Y}_a}{\hat{X}_a} \tag{6.68}$$

donde a representa una clase de edad-sexo y color y X_a es el total de población conocido en esta clase. Aquí tenemos una función de *dos* conjuntos de variables aleatorias $-\hat{Y}_a$ y $\hat{X}_a$. Además, por muchas razones y_{ahi} y $y_{a'hi}$ pueden estar correlacionadas para dos clases diferentes a y a' en una unidad conglomerada; así, por ejemplo, para una enfermedad infecciosa, el número de casos puede ser elevado para todas las clases de la unidad. Al aplicar (6.63) e ignorar la cpf, tenemos con $n_h = 2$,

$$v(\hat{Y}_{PS}) = \sum_h \left[\sum_a X_a\left(\frac{dy_{ah}'}{\hat{X}_a} - \frac{\hat{Y}_a dx_{ah}'}{\hat{X}_a^2}\right)\right]^2 \tag{6.69}$$

$$\doteq \sum_h \left[\sum_a X_a\left(\frac{\hat{Y}_a}{\hat{X}_a}\right)\left(\frac{dy_{ah}'}{\hat{Y}_a} - \frac{dx_{ah}'}{\hat{X}_a}\right)\right]^2 \tag{6.70}$$

donde $dy_{ah}' = N_h(y_{ah1} - y_{ah2})/2$.

En esa aplicación, el método de Keyfitz también trata otras complejidades de la encuesta como son: la selección de unidades primarias con probabilidades desiguales, los ajustes para no respuestas y el uso del método de estratos contraídos. Además, como el tiempo de cálculo, aun con este método simple, sólo permite cálculos de varianza para un número limitado de atributos, se dan gráficas de la relación entre la desviación estándar de $(\hat{Y})/\hat{Y}$ y $\hat{Y}$ para diferentes tipos de atributo lo que ayuda en la predicción de la d.e. $(\hat{Y})$ para atributos de los que aún no se ha calculado. Bean (1970) presenta claramente estos métodos y resultados.

6.14 ASIGNACION OPTIMA CON UNA ESTIMACION DE RAZON

La asignación óptima de los n_h puede ser diferente con una estimación de razón que con una media por unidad. Considérese primero la variante $\hat{Y}_{Rs}$. Del Teorema 6.4, su varianza es

$$V(\hat{Y}_{Rs}) = \sum_h \frac{N_h(N_h-n_h)}{n_h}(S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h\rho_h S_{yh}S_{xh})$$

$$= \sum_h \frac{N_h(N_h-n_h)}{n_h} S_{dh}^2, \quad \text{con } S_{dh}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h} d_{hi}^2 \qquad (6.71)$$

donde $d_{hi} = y_{hi} - R_h x_{hi}$ es la desviación de y_{hi} desde $R_h x_{hi}$. Mediante los métodos dados en el Cap. 5 para encontrar la distribución óptima, se deduce que (6.71) es mínima respecto a un costo total de la forma $\Sigma c_h n_h$, cuando

$$n_h \propto \frac{N_h S_{dh}}{\sqrt{c_h}}$$

Se recordará que para una varianza mínima con una media por unidad se escoge n_h proporcional a $N_h S_{yh}/\sqrt{c_h}$.

En el planeamiento de una muestra, la distribución con una estimación de razón puede ser un poco complicada, porque parece difícil especular sobre los valores probables de S_{dh}. Hay dos reglas útiles. Con una población en la cual la estimación de razón es la mejor estimación lineal sin sesgo, S_{dh} será aproximadamente proporcional a $\sqrt{\bar{X}_h}$ (según el Teorema 6.3). En este caso, los n_h deberían ser proporcionales a $N_h\sqrt{\bar{X}_h}/\sqrt{c_h}$. Algunas veces, la varianza de d_{hi} puede ser más aproximadamente proporcional a $\bar{X}_h^2$. Esto conduce a la distribución de n_h proporcional a $N_h\bar{X}_h/\sqrt{c_h}$, o sea, al total de x_{hi} por estrato, dividido entre la raíz cuadrada del costo por unidad. Un ejemplo de este tipo lo discute Hansen, Hurwitz y Gurney (1946) con una muestra diseñada para estimar las ventas de tiendas al menudeo.

Si se usa la estimación, $\hat{Y}_{Rc}$ se aplica mismo el razonamiento general.

Ejemplo. Los diferentes métodos de asignación pueden compararse con los datos recolectados en una lista completa de 257 huertos comerciales de duraznos en Carolina del Norte, en junio de 1946 (Finkner, 1950). El propósito fue determinar el procedimiento de muestreo más eficiente para estimar la producción comercial de duraznos en esta área. Se obtuvo información sobre el número de árboles de durazno y la producción total estimada en cada huerto. La alta correlación entre estas dos variables sugirió el uso de una estimación de razón. Se omitió un huerto muy grande.

Para esta ilustración, se divide el área geográficamente en tres estratos. El número de árboles (duraznos) en una huerta se denota por x_{hi} y la producción estimada en bushels por y_{hi}. Se considera solamente la primera estimación de razón $\hat{Y}_{Rs}$ (basada en una razón separada en cada estrato), ya que el principio es el mismo para ambos tipos de estimaciones de razón estratificadas.

Se comparan cuatro métodos de asignación. (*a*) n_h proporcional a N_h, (*b*) n_h proporcional a $N_h S_{yh}$, (*c*) n_h proporcional a $N_h\sqrt{\bar{X}_h}$, y (*d*) n_h proporcional a $N_h\bar{X}_h = X_h$. El tamaño de muestra es 100. Los datos para estas comparaciones se resumen en la Tabla 6.4.

TABLA 6.4. DATOS DE LA ENCUESTA DE PRODUCCIÓN DE DURAZNOS EN CAROLINA DEL NORTE

Estratos	S_{xh}^2	S_{yxh}	S_{yh}^2	S_{xh}	S_{yh}	$\bar{X}_h$	$\bar{Y}_h$	R_h	S_{dh}^2
1	5186	6462	8699	72.01	93.27	53.80	69.48	1.29133	658
2	2367	3100	4614	48.65	67.93	31.07	43.64	1.40475	573
3	4877	4817	7311	69.83	85.51	56.97	66.39	1.16547	2706
Pob.	3898	4434	6409	62.43	80.06	44.45	56.47	1.27053	1433

Estratos	N_h	(*a*)	$N_h S_{yh}$	(*b*)	$\sqrt{\bar{X}_h}$	$N_h\sqrt{\bar{X}_h}$	(*c*)	$N_h\bar{X}_h$	(*d*)
1	47	18	4384	22	7.33	344.5	20	2529	22
2	118	46	8016	40	5.57	657.3	39	3666	32
3	91	36	7781	38	7.55	687.1	41	5184	46
Pob.	256	100	20181	100	20.45	1688.9	100	11379	100

La parte superior de la tabla muestra los datos básicos. El método empleado para calcular las cuatro varianzas fue encontrar primero los n_h para cada tipo de asignación. Estos valores aparecen en las columnas encabezadas (*a*) hasta (*d*) en la parte inferior del cuadro. Así, con asignación (*a*), $n_h = nN_h/N$, de tal manera que en el primer estrato.

$$n_1 = \frac{(100)(47)}{256} = 18$$

Cuando se han obtenido los n_h, se encuentra la $V(\hat{Y}_{Rs})$ correspondiente sustituyendo en la fórmula

$$V(\hat{Y}_{Rs}) = \sum_h \frac{N_h(N_h-n_h)}{n_h} S_{dh}^2$$

donde

$$S_{dh}^2 = S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h S_{yxh}$$

Las cantidades S_{dh}^2 son las mismas para todas las cuatro asignaciones y se dan en el extremo derecho de la mitad superior de la Tabla 6.4.

En la Tabla 6.5 se muestran las varianzas y precisiones relativas.

TABLA 6.5. COMPARACIÓN DE CUATRO MÉTODOS DE ASIGNACIÓN

Método de asignación: n_h proporcional a	Varianza: Estrato 1	Estrato 2	Estrato 3	Total	Precisión relativa
1. N_h	49 824	105 833	376 215	531 872	100
2. $N_h S_{yh}$	35 144	131 847	343 446	510 437	104
3. $N_h\sqrt{\bar{X}_h}$	41 750	136 964	300 312	479 026	111
4. $N_h\bar{X}_h$	35 144	181 710	240 888	457 742	116

No hay mucho que escoger entre las diferentes asignaciones, como se esperaría, ya que los n_h no difieren grandemente en los cuatro métodos. El método 4, en el cual la asignación es proporcional al número total de árboles en el estrato, parece ser ligeramente superior a los otros.

6.15 ESTIMACIONES INSESGADAS DEL TIPO DE RAZON

Como hemos señalado, las estimaciones insesgadas o menos sesgadas que $\hat{R}$ o $\hat{Y}_R$ del tipo de razón pueden ser útiles en encuestas de numerosos estratos y pequeñas muestras de cada estrato, si parece apropiada la estimación de razón separada. Discutiremos rápidamente tres métodos que proporcionan estimaciones insesgadas y tres métodos que eliminan el término de orden $1/n$ [ver (6.33)].

En la comparación de estos métodos son importantes las siguientes preguntas: (a) ¿El ECM se compara favorablemente con el de la estimación ordinaria de razón? (b) ¿Proporciona el método una estimación muestral satisfactoria de la varianza? Como hemos visto, ésta es una dificultad con $\hat{R}$. Comenzamos por describir los métodos. Los del tipo insesgado requieren el conocimiento de $\bar{X}$.

Métodos Insesgados

Una estimación, debida a Hartley y Ross (1954), puede derivarse partiendo de la media $\bar{r}$ de las razones y_i/x_i y corrigiéndola por sesgo.

$$\bar{r}=\frac{1}{n}\sum^{n} r_i=\frac{1}{n}\sum^{n}\frac{y_i}{x_i}$$

Ahora

$$\frac{1}{N}\sum_{i=1}^{N} r_i(x_i-\bar{X})=\frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{x_i}\cdot x_i-\left(\frac{1}{N}\sum_{i=1}^{N} r_i\right)\bar{X}$$
$$=\bar{Y}-\bar{X}E(r_i)=\bar{X}[R-E(r_i)] \qquad (6.72)$$

Pero en muestreo aleatorio simple $E(\bar{r}) = E(r_i)$. Por lo tanto,

$$\text{sesgo en } \bar{r}=E(\bar{r})-R=-\frac{1}{\bar{X}N}\sum_{i=1}^{N} r_i(x_i-\bar{X}) \qquad (6.73)$$

Por el Teorema 2.3, una estimación insesgada de muestra de

$$\frac{1}{N-1}\sum_{i=1}^{N} r_i(x_i-\bar{X})$$

es

$$\frac{1}{n-1}\sum_{i=1}^{n} r_i(x_i-\bar{x})=\frac{n}{n-1}(\bar{y}-\bar{r}\bar{x})$$

Al sustituir en (6.73), la estimación $\bar{r}$, corregida por sesgo, se vuelve

$$\hat{R}_{HR}=\bar{r}+\frac{n(N-1)}{(n-1)N\bar{X}}(\bar{y}-\bar{r}\bar{x}) \qquad (6.74)$$

La estimación insesgada correspondiente del total de población $\hat{Y}$ es

$$\hat{R}_{HR}X=\bar{r}X+\frac{n(N-1)}{n-1}(\bar{y}-\bar{r}\bar{x}) \qquad (6.75)$$

Por argumentos semejantes, se deriva otra estimación insesgada (Mickey, 1959) a partir de las n razones $\hat{R}_i$ que se obtienen al *retirar* cada unidad en turno de la muestra, de modo que $\hat{R}_i=\sum y/\sum x$ sobre los restantes $(n-1)$ miembros. Si $\hat{\bar{R}}_{-}$ denota la media de las $\hat{R}_i$, la estimación de Mickey es

$$\hat{R}_M=\hat{\bar{R}}_{-}+\frac{n(N-n+1)}{N\bar{X}}(\bar{y}-\hat{\bar{R}}_{-}\bar{x}). \qquad (6.76)$$

En un tercer método, Lahiri (1951) demostró que la estimación ordinaria de razón $\hat{R}$ es insesgada si se saca la muestra con probabilidad proporcional a $\sum x_i$. Tal vez, el método más simple para hacer esto (Midzuno, 1951) es el de extraer el primer miembro de la muestra con probabilidad proporcional a x_i. Los restantes $(n-1)$ miembros de la muestra se sacan con probabilidad iguales. Es fácil probar (Ejercicio 6.10) que con este método, la probabilidad de sacar una muestra específica es proporcional a $\sum x_i$, y que $\hat{R}=\sum^{n} y_i/\sum^{n} x_i$ es insesgada para este método de selección de muestra.

Métodos con un sesgo del orden $1/n^2$

Estos métodos consisten en hacer un ajuste a $\hat{R}$. El primero, debido a Quenouille (1956), se aplica a una vasta clase de proble-

mas estadísticos en los que la estimación propuesta tiene un sesgo de orden $1/n$. Se ha llamado *jackknife* porque se trata de un instrumento de múltiples usos. La utilidad de este método para estimaciones de razón la señaló Durbin (1959).

Al ignorar la cpf por el momento, el sesgo de estimaciones como $\hat{R}$ se puede desarrollar en una serie del tipo

$$E(\hat{R}) = R + \frac{b_1}{n} + \frac{b_2}{n^2} + \cdots \tag{6.77}$$

Si $n = mg$, dividamos la muestra al azar en g grupos de tamaño m. A partir de (6.77)

$$E(g\hat{R}) = gR + \frac{b_1}{m} + \frac{b_2}{gm^2} + \cdots \tag{6.78}$$

Ahora, sea $\hat{R}_j$ la razón ordinaria $\sum y / \sum x$, calculada de la muestra, después de *omitir* el grupo j-ésimo. Como $\hat{R}_j$; se obtiene de una muestra aleatoria simple de tamaño $m(g-1)$, tenemos

$$E(\hat{R}_j) = R + \frac{b_1}{(g-1)m} + \frac{b_2}{(g-1)^2 m^2} + \cdots \tag{6.79}$$

Por lo tanto

$$E[(g-1)\hat{R}_j] = (g-1)R + \frac{b_1}{m} + \frac{b_2}{(g-1)m^2} + \cdots \tag{6.80}$$

Al sustraer de (6.78) se obtiene, hasta el orden n^{-2}.

$$E[g\hat{R} - (g-1)\hat{R}_j] = R - \frac{b_2}{g(g-1)m^2} = R - \frac{b_2}{n^2}\frac{g}{(g-1)}$$

Ahora el sesgo es de orden $1/n^2$. Podemos construir g estimaciones de este tipo, una para cada grupo. El estimador de Quenouille (el jackknife) es el promedio de estas g estimaciones, o sea

$$\hat{R}_Q = g\hat{R} - (g-1)\hat{R}_- \tag{6.81}$$

donde $\hat{R}_-$ es el promedio de las g cantidades $\hat{R}_j$. Como lo muestra Quenouille, la varianza de $\hat{R}_Q$ difiere de la de $\hat{R}$ por términos de orden $1/n^2$. Cualquier incremento en la varianza debido a este ajuste por sesgo, debería, por lo tanto, ser despreciable en muestras moderadamente grandes. La elección $m = 1$, $g = n$ parece la mejor con el método Jackknife en muestras pequeñas.

Si no puede ignorarse la cpf, el término guía en el sesgo de $\hat{R}$, como en (6.33) es de la forma $b_1(1-f)/n$. Puede demostrarse (Ejercicio 6.10) que para quitar ambos términos en $1/n$ y en $1/N$ necesitamos

$$\hat{R}_Q = w\hat{R} - (w-1)\hat{R}_- \tag{6.82}$$

donde $w = g[1 - (n-m)/N]$, o con $m = 1$, $g = n$, $w = n[1 - (n-1)/N]$.

El estimador de Beale (1962) es

$$\hat{R}_B = \frac{\bar{y} + [(1-f)/n](s_{yx}/\bar{x})}{\bar{x} + [(1-f)/n](s_x^2/\bar{x})} = \hat{R}\cdot\frac{1 + [(1-f)/n]c_{yx}}{1 + [(1-f)/n]c_{xx}} \tag{6.83}$$

en tanto que el de Tin (1965) es la cantidad estrechamente relacionada

$$\hat{R}_T = \hat{R}\left[1 - \frac{(1-f)}{n}\left(\frac{s_x^2}{\bar{x}^2} - \frac{s_{yx}}{\bar{y}\bar{x}}\right)\right] = \hat{R}\left[1 - \frac{(1-f)}{n}(c_{xx} - c_{yx})\right] \tag{6.84}$$

donde $s_{yx} = \sum^n (y_i - \bar{y})(x_i - \bar{x})/(n-1)$, $s_x^2 = \sum^n (x_i - \bar{x})^2/(n-1)$, de modo que c_{yx}, c_{xx} son la covarianza relativa y la varianza relativa de la muestra de x.

La estructura de $\hat{R}_T$ se puede ver al notar que por (6.34), el término guía en el valor esperado de $\hat{R}$ puede escribirse

$$R\left[1 + \frac{(1-f)}{n}(C_{xx} - C_{yx})\right]$$

Evidentemente, $\hat{R}_T$ es el ajuste de $\hat{R}$ por una estimación de muestra de la corrección del sesgo requerida. $\hat{R}_B$ y $\hat{R}_T$ son idénticas a los términos de orden $1/n$, y en general se comportan en forma similar.

6.16 COMPARACION DE LOS METODOS

Ejemplo. La población artificial de la Tabla 6.6 contiene tres estratos con $N_h = 4$, $n_h = 2$ en cada estrato. La población se construyó deliberadamente de modo que $(a)R_h$ varíe marcadamente de un estrato al otro, lo que

TABLA 6.6. UNA PEQUEÑA POBLACIÓN ARTIFICIAL

	Estrato I		Estrato II		Estrato III	
	y	x	y	x	y	x
	2	2	2	1	3	1
	3	4	5	4	7	3
	4	6	9	8	9	4
	11	20	24	23	25	12
Totales	20	32	40	36	44	20
R_h	0.625		1.111		2.200	

favorece una estimación de razón *separada* $\hat{Y}_{Rs}$, y (*b*) $\hat{R}_h$ sobreestima en cada estrato, con la amenaza de un sesgo importante cumulativo en $\hat{Y}_{Rs}$. Se compararon los siguientes métodos de estimación del *total* de población Y.

Expansión simple: $\sum N_h \bar{y}_h$

Razón combinada: $(\bar{y}_{st}/\bar{x}_{st})X$

Razón separada: $\sum (\bar{y}_h/\bar{x}_h)X_h = \sum \hat{R}_h X_h$

Las estimaciones restantes, la *separada* de Hartley y Ross, la de Lahiri y la de Quenouille, los métodos de Beale y de Tin, tienen la misma forma que la estimación de razón separada, salvo que $\hat{R}_{(HR)h}$, $\hat{R}_{(_)h}$, etc., reemplazan $\hat{R}_h$. (Para $n_h = 2$, los métodos de Hartley-Ross y de Mickey, son idénticos). Hay $6^3 = 216$ muestras posibles. Los resultados son exactos salvo errores de redondeo. Por su ayuda en algunos de los cálculos estoy en deuda con el doctor Joseph Sedransk.

Los resultados de la Tabla 6.7 exhiben algunos aspectos interesantes. Para la estimación de razón combinada, la contribución del cuadrado del sesgo al error cuadrático medio es trivial, a pesar de las condiciones extremas, pero esta estimación es mediocre en lo que respecta la varianza, por la amplia variación de los R_h. Al juzgar por el ECM, la estimación de razón separada es mucho más exacta que la estimación combinada, pero está fuertemente sesgada. De los métodos insesgados, el de Hartley-Ross muestra una varianza relativamente alta, como se ha encontrado para $n_h = 2$ en algunos estudios sobre poblaciones naturales. El método de Lahiri se comporta particularmente bien. Esta población se adapta al método de Lahiri porque una unidad en cada estrato tiene valores excepcionalmente altos de y_i así como de x_i, con la consecuencia de que esta unidad tiene una probabilidad elevada de salir y las muestras que contienen esta unidad dan buenas estimaciones de R_h.

Los métodos de Quenouille, Beale y Tin, produjeron disminuciones importantes del sesgo, en comparación con la estimación de razón separada, y todos tuvieron menores ECM, de modo que en este ejemplo, lograron sus propósitos principales.

TABLA 6.7. RESULTADOS PARA DIFERENTES ESTIMACIONES DE Y

Método	Varianza	(Sesgo)²	ECM
Expansión simple	820.3	0.0	820.3
Razón combinada	262.8	6.5	269.3
Razón separada	35.9	24.1	60.0
Razón separada de Hartley-Ross	153.6	0.0	153.6
Separada de Lahiri	19.6	0.0	19.6
Separada de Quenouille	42.9	1.1	44.0
Separada de Beale	28.9	8.0	36.9
Separada de Tin	28.6	5.7	34.3

El estudio de J. N. K. Rao (1969) de poblaciones naturales, que se mencionó en la Sec. 6.9, comparaba los métodos de Quenouille, Beale y Tin, para $n = 2, 4, 6, 8$ en 15 de dichas poblaciones. Para

$n = 2$, la prueba más severa, las medias y cuartiles superiores de las cantidades $|\text{sesgo}|/\sqrt{\text{ECM}}$ fueron: $\hat{R}_Q$, 3%, 7%; $\hat{R}_B$, 8%, 12%; $\hat{R}_T$, 8%, 19%; en contraposición a 15%, 20% para $\hat{R}$. Los métodos más complejos parecen ser útiles en lo que respecta al sesgo en estas muestras muy pequeñas.

El mismo estudio comparaba los ECM de cinco de las estimaciones de esta sección con el de $\hat{R}$. (Se omitió el método de Lahiri, porque el estudio estaba confinado a las muestras aleatorias simples.)

Para cada método se calculó la razón 100 ECM ($\hat{R}_Q$) ECM ($\hat{R}$), y así sucesivamente para cada población.

Para $n = 4$, las estimaciones de Quenouille y Mickey, fueron un poco inferiores a $\hat{R}$ en estas poblaciones, pero para $n \geq 6$ todos los métodos tenían un promedio de ECM muy cercano a los de $\hat{R}$. Para una muestra muy pequeña en una *sola* población, este estudio sugiere que estos métodos más complejos no presentan ventaja material en exactitud sobre $\hat{R}$. Pero el hecho de que reducen el sesgo con un incremento pequeño o nulo del ECM en un estrato, debería darles ventaja en una estimación de razón separada con numerosos estratos, que tienen muestras pequeñas.

En modelo de regresión lineal, comparaciones de los ECM de estos métodos para *n*, efectuadas por P.S.R.S. Rao y J.N.K. Rao (1971), Hutchinson (1971) y J.N.K. Rao y Kuzik (1974) dieron resultados que en general concuerdan con los de poblaciones naturales.

6.17 ESTIMACION MEJORADA DE LA VARIANZA

Un estimador que vale la pena considerar para muestras pequeñas o moderadas es el que sugirió Tukey (1958) para la estimación de jackknife (Quenouille) $\hat{R}_Q$. Con *g* grupos, $n = mg$, y *f* despreciable, $\hat{R}_Q$ es el promedio de las *g* cantidades.

$$\hat{R}'_j = g\hat{R} - (g-1)\hat{R}_j$$

donde $\hat{R}_j = \sum y / \sum x$, se calcula después de omitir el grupo *j*. Si las $\hat{R}'_j$ pudieran considerarse como *g* estimaciones independientes de *R*, entonces, con muestreo aleatorio simple, la siguiente sería una estimación insesgada de $V(\hat{R}_Q)$

$$v(\hat{R}_Q) = \frac{(1-f)}{g} \frac{\sum^{g} (\hat{R}'_j - \hat{R}_Q)^2}{(g-1)} \tag{6.85}$$

Dado que $\hat{R}'_j - \hat{R}_Q = -(g-1)(\hat{R}_j - \hat{R}_{_})$, (6.85) se calcula más fácilmente como

$$v(\hat{R}_O)=\frac{(1-f)(g-1)}{g}\sum^{g}(\hat{R}_j-\hat{R}_-)^2 \qquad (6.86)$$

donde $\hat{R}_-$ es la media de las g cantidades $\hat{R}_j$.

Por supuesto, las $\hat{R}_j'$ o $\hat{R}$ para diferentes j no son independientes y la Fórmula (6.86) es una aproximación. Hasta aquí, las propiedades analíticas de $v(\hat{R}_Q)$ se han establecido sólo para muestras grandes. Arvesen (1969) mostró que para una amplia clase de estimaciones que incluyen $\hat{R}_Q$ y que son simétricas en los elementos de la muestra [estimaciones que se conocen como las U-estadísticas de Hoeffding (1948)], la Fórmula $v(\hat{R}_O)$ se hace insesgada cuando se fija g o para $g = n$ al crecer n lo suficiente.

Del estudio de Rao (1969) de ocho poblaciones naturales, se reportaron en la Sec. 6.9 para $n = 4, 6, 8, 12$, porcentajes de subestimaciones por promedio de muestras pequeñas en la $v_1(\hat{R})$ estándar, como estimación del verdadero ECM($\hat{R}$). Los correspondientes porcentajes de sesgo promedio en $v(\hat{R}_Q)$ se muestran con fines comparativos en la Tabla 6.8; éstos son los promedios de los ocho números: $100\,[v(\hat{R}_Q) - \text{ECM}(\hat{R}_Q)]/\text{ECM}(\hat{R}_Q)$.

TABLA 6.8. PORCENTAJE DE SESGO PROMEDIO EN ESTIMACIONES DE VARIANZA

Promedio de	$n=$ 4	6	8	12
$100[v(\hat{R}_O)-\text{ECM}(\hat{R}_O)]/\text{ECM}(\hat{R}_O)$	+11%	+10%	+6%	+1%
$100[v(\hat{R}_O)-\text{ECM}(\hat{R})]/\text{ECM}(\hat{R})$	+11%	+10%	+6%	+1%
$100[v_1(\hat{R})-\text{ECM}(\hat{R})]/\text{ECM}(\hat{R})$	−31%	−23%	−21%	−18%

La Tabla 6.8 también da los promedios de las cantidades $100\,[v(\hat{R}_Q) - \text{ECM}(\hat{R})]/\text{ECM}(\hat{R})$. Estos promedios son de interés para el investigador que usa $\hat{R}$, pero está dispuesto a reemplazar $v_1(\hat{R})$ por $v(\hat{R}_Q)$ como estimador de $V(\hat{R})$ si parece menos sesgado. En vista de los sesgos en $\hat{R}$ y $\hat{R}_Q$, las comparaciones más apropiadas de $V(\hat{R}_Q)$ se hacen con los ECM.

En estas poblaciones, $v(\hat{R}_Q)$ es una ligera sobreestimación de ECM ($\hat{R}_Q$) y ECM($\hat{R}$), en tanto que $v_1(\hat{R})$ tiene sesgos negativos sustanciales en estas muestras pequeñas.

La estabilidad de $v(\hat{R}_Q)$ con relación a la de $v_1(\hat{R})$ juzgada por los cuadrados de los coeficientes de variación de estas estimaciones de varianza, resultó bastante pobre en estas muestras. En estudios de Rao y Beegle (1968) de $v(\hat{R}_Q)$ y $v_1(\hat{R})$ en un modelo de regresión lineal de y respecto a x en una población infinita con x normal, sin embargo, $v(\hat{R}_Q)$ y $v_1(\hat{R})$ resultaron más o menos igualmente estables para $n = 4$ hasta $n = 12$.

Con una estimación de razón separada y numerosos estratos, estos resultados sugieren que $\sum X_h^2 v(\hat{R}_{Oh})$ es superior a $\sum X_h^2 v_1(\hat{R}_h)$ como estimador de $V(\hat{Y}_{Rs})$. Puede esperarse que el primero tenga menos sesgo y ambos sean adecuadamente estables. Pero con pocos estratos, estas afirmaciones son cuestionables antes de tener información suficiente a nuevas comparaciones.

6.18 COMPARACION DE DOS RAZONES

En muestreos analíticos frecuentemente es necesario estimar la diferencia $\hat{R} - \hat{R}'$ entre dos razones y calcular el error estándar de $\hat{R} - \hat{R}'$. Las fórmulas que se dan aquí, son para la varianza *estimada* de $\hat{R} - \hat{R}'$, ya que éstas son las que se requieren más frecuentemente. Las cpf se omiten por las razones dadas en la Sec. 2.14.

Primero se supone muestreo aleatorio simple. Se pueden distinguir tres casos.

Las Dos Razones son Independientes

Esto ocurre cuando las unidades se clasifican en dos clases distintas y queremos comparar las razones estimadas separadamente en ambas clases. Por ejemplo, en un estudio sobre gastos domésticos una muestra aleatoria simple de familias se puede subdividir en casas propias y alquiladas, para comparar las proporciones de gasto en el mantenimiento de la casa en las dos clases. Si las razones estimadas se denotan por $\hat{R} = \bar{y}/\bar{x}$, $\hat{R}' = \bar{y}'/\bar{x}'$, entonces

$$v(\hat{R}-\hat{R})=v(\hat{R})+v(\hat{R}') \qquad (6.87)$$

Las Dos Razones Tienen el Mismo Denominador

Cuando la unidad es un conglomerado de familias, puede quererse comparar la proporción de adultos del sexo masculino que usan rasuradoras eléctricas con la proporción de quienes usan navajas. En cualquier unidad, y = número de hombres adultos que usan rasuradoras eléctricas, y' = número de hombres adultos que usan navajas y x = número total de hombres adultos.

$$\hat{R}-\hat{R}'=\frac{\bar{y}-\bar{y}'}{\bar{x}}$$

Si $d_i = y_i - y_i'$, la varianza estimada de $\hat{R} - \hat{R}'$ puede calcularse como

$$v(\hat{R}-\hat{R}') \doteq \frac{1}{n(n-1)\bar{x}^2} \sum_{i=1}^{n} [d_i - (\hat{R}-\hat{R}')x_i]^2 \qquad (6.88)$$

Las Dos Razones Tienen Denominadores Diferentes pero Pueden Estar Correlacionadas

Un ejemplo es la comparación de la proporción de hombres que fuman, con la proporción de mujeres que fuman, en una encuesta en la cual la unidad es un conglomerado de casas. Matemáticamente, éste es el caso más general.

$$v(\hat{R}-\hat{R}') = v(\hat{R}) + v(\hat{R}') - 2\,\mathrm{cov}\,(\hat{R}\hat{R}') \qquad (6.89)$$

El único término que no es familiar es cov $(\hat{R}\hat{R}')$. Al escribir, en la forma usual

$$\hat{R} - R \doteq \frac{\bar{y} - R\bar{x}}{\bar{X}} \qquad \hat{R}' - R' \doteq \frac{\bar{y}' - R'\bar{x}'}{\bar{X}'}$$

tenemos

$$\mathrm{cov}\,(\hat{R}\hat{R}') \doteq \frac{1}{n\bar{X}\bar{X}'}\,\mathrm{cov}\,(y_i - Rx_i)(y_i' - R'x_i')$$

Una estimación de la muestra puede calcularse como sigue:

$$\mathrm{cov}\,(\hat{R}\hat{R}') \doteq \frac{1}{n(n-1)\bar{x}\bar{x}'} \sum^{n} (y_i y_i' - \hat{R}y_i'x_i - \hat{R}'y_i x_i' + \hat{R}\hat{R}'x_i x_i') \qquad (6.90)$$

Ejemplo. El experimento de campo de la vacuna Salk contra la polio se condujo entre niños en los primeros tres grados en todas las escuelas de cierto número de condados. Los condados no fueron seleccionados al azar, ya que se favoreció a aquellos con una historia de polio previa, pero para esta ilustración se supondrá que fueron una muestra aleatoria de alguna población.

Los niños cuyos padres no les dieron permiso para participar en la prueba, se llamaron el grupo de los "no inoculados" y, desde luego, no recibieron inyecciones. La mitad de los niños que sí tenían permiso, recibieron tres inyecciones de un líquido inerte y se llamaron el grupo "placebo". De los datos en la Tabla 6.9, compare las frecuencias $\hat{R}$, $\hat{R}'$ de polio con parálisis en los grupos "no inoculado" y "placebo". Para reducir la cantidad de datos, la comparación se restringe a 34 condados, cada uno con más de 4 000 niños en ambos grupos combinados.

En estos datos cualquier variación en el grado de ataque de polio de condado a condado produciría una correlación positiva entre $\hat{R}$ y $\hat{R}'$. Las siguientes cantidades se derivan de los totales.

Placebo: $\hat{R} = \dfrac{88}{167.4} = 0.525687, \qquad \bar{x} = \dfrac{167.4}{34} = 4.9235$

TABLA 6.9. NÚMERO DE NIÑOS (x, x') Y CASOS DE PARALÍTICOS (y, y') POR CONDADO

x*	x'	y†	y'	x	x'	y	y'
4.1	2.4	0	0	13.8	25.6	3	3
3.5	8.0	1	6	10.5	8.1	2	0
4.1	6.1	7	2	21.6	25.9	10	7
2.6	4.6	2	1	3.5	6.7	2	2
2.4	1.5	2	1	6.8	7.3	3	8
2.2	1.9	0	0	2.3	3.7	0	1
1.1	4.0	1	1	2.6	2.9	2	0
1.6	4.0	1	2	6.0	11.1	3	1
5.7	7.8	1	4	11.0	14.8	7	11
3.3	11.0	3	7	19.4	42.5	11	14
1.0	3.8	0	1	6.8	13.7	6	2
2.0	5.2	1	0	1.2	4.0	3	1
8.3	19.0	4	4	5.4	9.3	11	6
1.0	3.7	1	5	1.7	2.6	0	2
1.1	4.2	0	1	2.1	2.3	0	0
2.3	6.8	1	2	1.5	2.6	0	0
1.9	3.5	0	2	3.0	4.0	0	2
			Totales	167.4	284.6	88	99

* x, x' = números de niños "placebo" y "no inoculados" (en miles).
† y, y' = números de casos de polio con parálisis en los grupos placebo y no inoculados.

No inoculado: $\hat{R}' = \dfrac{99}{284.6} = 0.347857, \qquad \bar{x}' = \dfrac{284.6}{34} = 8.3706$

Para $v(\hat{R})$, $v(\hat{R}')$ y cov $(\hat{R}\hat{R}')$, se requieren todas las sumas de cuadrados y productos sin corregir entre las cuatro variantes.

$$v(\hat{R}) = \frac{1}{n(n-1)\bar{x}^2}(\textstyle\sum y^2 - 2\hat{R}\sum yx + \hat{R}^2 \sum x^2)$$

$$= \frac{1}{(34)(33)(4.9235)^2}[(564) - (1.05137)(822.2) + (0.27635)(1661.92)]$$

$$= 0.00584$$

Similarmente, encontramos $v(\hat{R}') = 0.00240$.

$$\mathrm{cov}\,(\hat{R}\hat{R}') = \frac{1}{n(n-1)\bar{x}\bar{x}'}(\textstyle\sum yy' - \hat{R}\sum y'x - \hat{R}'\sum yx' + \hat{R}\hat{R}'\sum xx')$$

$$= \frac{(497) - (0.52569)(844.6) - (0.34786)(1397.4) + (0.52569)(0.34786)(2690.8)}{(34)(33)(4.9235)(8.3706)}$$

$$= 0.00127$$

Por lo tanto,

$$\text{e.e.}(\hat{R}-\hat{R}')=\sqrt{0.00584+0.00240-0.00254}=0.0754$$

Dado que $\hat{R}-\hat{R}'=0.1778$, la diferencia alcanza importancia al nivel de 5% (la distribución de $\hat{R}-\hat{R}'$ puede ser un tanto asimétrica para este tamaño de muestra). Una posible explicación es que los niños no inoculados pueden haber tenido más protección natural contra la polio que los niños placebo.

Puede darse el mismo problema en muestras estratificadas en las cuales los dominios del estudio atraviesan los estratos. Si $\hat{R}_h, \hat{R}_h'$ parecen variar de estrato a estrato, la comparación se basará probablemente en un examen de los valores de $\hat{R}_h-\hat{R}_h'$ en los estratos individuales. Al encontrar los errores estándar de $\hat{R}_h-\hat{R}_h'$, es posible determinar si estas diferencias varían de estrato a estrato y, si no, calcular una diferencia global eficiente.

Si las $\hat{R}_h, \hat{R}_h'$ no exhiben una variación real de estrato a estrato, puede ser suficiente con comparar las estimaciones combinadas $\hat{R}_c$ y $\hat{R}_c'$. Como anteriormente

$$v(\hat{R}_c-\hat{R}_c')=v(\hat{R}_c)+v(\hat{R}_c')-2\,\text{cov}\,(\hat{R}_c\hat{R}_c') \tag{6.91}$$

donde, al poner $d_{hi}=(y_{hi}-\bar{y}_h)-\hat{R}_c(x_{hi}-\bar{x}_h)$,

$$v(\hat{R}_c)=\frac{1}{\bar{x}_{st}^{\,2}}\sum_h\frac{N_h^{\,2}}{n_h(n_h-1)}\sum_i d_{hi}^{\,2} \tag{6.92}$$

$$\text{cov}\,(\hat{R}_c\hat{R}_c')=\frac{1}{\bar{x}_{st}\bar{x}_{st}'}\sum_h\frac{N_h^{\,2}}{n_h(n_h-1)}\sum_i d_{hi}d_{hi}' \tag{6.93}$$

Una discusión más prolija sobre la comparación de razones, que incluye fórmulas abreviadas de cálculo cuando las muestras las permiten, ha sido dada por Kish y Hess (1959*b*).

6.19 RAZON DE DOS RAZONES

En algunas aplicaciones deseamos estimar la razón R/R' de dos razones. Así, en el ejemplo precedente, podríamos interesarnos en la razón R/R' de las tasas de niños paralíticos que recibieron un "placebo" y de los niños no inoculados, o la razón de las proporciones de hombres y mujeres que trabajan, tomados de una muestra conglomerada. Si los datos (y, x) están a la disposición en la misma muestra para dos periodos de tiempo, la cantidad podría ser la razón de gastos semanales alimenticios por cada casa en los dos tiempos diferentes.

Con una muestra aleatoria simple (es decir de conglomerados) la estimación muestral de R/R' es $\hat{R}/\hat{R}'=(\bar{y}/\bar{x})/(\bar{y}'/\bar{x}')$, llamada en

ocasiones *estimación de razón doble*. Como para la razón simple, el término guía del sesgo de $\hat{R}/\hat{R}'$ es del orden de $1/n$, pero más complejo que el de $\hat{R}$ o $\hat{R}'$. Podemos escribir

$$\hat{R}/\hat{R}'=(R/R')(1+\delta\bar{y})(1+\delta\bar{x}')/(1+\delta\bar{x})(1+\delta\bar{y}') \tag{6.94}$$

donde $\delta\,\bar{y}$ denota $(\bar{y}-\bar{Y})/\bar{Y}$, y así sucesivamente. Cuando se desarrolla esta expresión, hay seis términos cuadráticos de orden $1/n$ que aparecen en el sesgo de $\hat{R}/\hat{R}'$. Rao y Pereira (1968) dan una expresión exacta para este sesgo.

La fórmula (6.8) para la varianza relativa $V(\hat{R})/R^2=C_{\hat{R}\hat{R}}$ de una razón, se puede escribir

$$C_{\hat{R}\hat{R}}=C_{\bar{y}\bar{y}}+C_{\bar{x}\bar{x}}+2C_{\bar{y}\bar{x}}$$

De aquí que el término guía en $V(\hat{R}/\hat{R}')$ sea

$$V(\hat{R}/\hat{R}')\doteq\left(\frac{R}{R'}\right)^2(C_{\hat{R}\hat{R}}+C_{\hat{R}'\hat{R}'}-2C_{\hat{R}\hat{R}'}) \tag{6.95}$$

donde

$$C_{\hat{R}\hat{R}'}=C_{\bar{y}\bar{y}'}+C_{\bar{x}\bar{x}'}-C_{\bar{y}\bar{x}'}-C_{\bar{y}'\bar{x}} \tag{6.96}$$

para la estimación de muestra correspondiente $v(\hat{R}/\hat{R}')$ sustituimos estimaciones muestrales de los términos en (6.95).

Ejemplo. Para la razón de las tasas de placebo y casos no inoculados $\hat{R}/\hat{R}=0.52569/0.34786=1.511$. Estimar el error estándar de esta razón. Los cálculos del ejemplo precedente, proporcionan

$$\hat{C}_{\hat{R}\hat{R}}=\frac{0.00584}{(0.5257)^2}=0.0211;\qquad \hat{C}_{\hat{R}'\hat{R}'}=\frac{0.00240}{(0.3479)^2}=0.0198;$$

$$\hat{C}_{\hat{R}\hat{R}'}=\frac{0.00127}{(0.5257)(0.3479)}=0.00694$$

$$v(\hat{R}/\hat{R}')=(1.511)^2(0.0211+0.0198)-0.0139)=0.0617$$

$$\text{e.e.}\,(\hat{R}/\hat{R}')=0.248$$

La estimación de razón doble se ha utilizado ocasionalmente en lugar de $\hat{Y}_R=\hat{R}X$, para estimar un *total* de población Y, como lo sugirió Keyfitz (Yates, 1960). Supóngase que $\hat{R}'=(\bar{y}'/\bar{x}')$ se conoce para la *misma* muestra en un periodo previo y que también se conoce $R'=Y'/\bar{X}'$. Si $\hat{R}'/R'$ se encontró que es, digamos un poco superior a >1, podríamos argüir intuitivamente que $\hat{R}$ también es de esperarse que dé una sobreestimación de R que debería ajustarse hacia abajo, al dividirla entre la razón $\hat{R}'/R'$. Esto conduce a la estimación de razón doble $\hat{Y}_{DR}$.

$$\hat{Y}_{DR}=\frac{R'}{\hat{R}'}(\hat{R}X)=\frac{\hat{R}}{\hat{R}'}(R'X) \tag{6.97}$$

Como la varianza relativa de $\hat{Y}_R$ es $C_{\hat{R}\hat{R}}$, en tanto que la de $\hat{Y}_{DR}$ es

$$C_{\hat{R}\hat{R}} + C_{\hat{R}'\hat{R}'} - 2C_{\hat{R}\hat{R}'} \tag{6.98}$$

la razón doble dará una estimación más precisa en muestras grandes si la correlación entre $\hat{R}$ y $\hat{R}'$ es bastante alta.

6.20 ESTIMACIONES DE RAZON MULTIVARIANTES

Olkin (1958) ha extendido la estimación de razón a la situación en la cual p variables auxiliares $x(x_1, x_2, \ldots x_p)$ están disponibles. La estimación propuesta para el total de la población, digamos $\hat{Y}_{MR}$ para la *razón multivariable*, es

$$\hat{Y}_{MR} = W_1\frac{\bar{y}}{\bar{x}_1}X_1 + W_2\frac{\bar{y}}{\bar{x}_2}X_2 + \cdots + w_p\frac{\bar{y}}{\bar{x}_p}X_p$$
$$= W_1\hat{Y}_{R_1} + W_2\hat{Y}_{R_2} + \cdots + W_p\hat{Y}_{R_p}$$

donde las W_i son ponderaciones por determinarse para maximizar la precisión de $\hat{Y}_{MR}$, respecto a $\Sigma W_i = 1$. Este tipo de estimación parece conveniente cuando la regresión de y en $x_1, x_2, \ldots x_p$ es lineal y pasa por el origen. Deben conocerse los totales de población X_i.

El método se describe para dos variables x, ya que esta sería la aplicación más frecuente. Tenemos

$$\hat{Y}_{MR} - Y = W_1(\hat{Y}_{R_1} - Y) + W_2(\hat{Y}_{R_2} - Y)$$

Por lo tanto, al suponer que el sesgo es despreciable,

$$V(\hat{Y}_{MR}) = W_1^2V(\hat{Y}_{R_1}) + 2W_1W_2\,\mathrm{cov}\,(\hat{Y}_{R_1}\hat{Y}_{R_2}) + W_2^2V(\hat{Y}_{R_2})$$
$$= W_1^2V_{11} + 2W_1W_2V_{12} + W_2^2V_{22} \tag{6.99}$$

donde $V_{11} = V(\hat{Y}_{R_1})$, etc. Los valores de W_1, W_2, los cuales minimizan la varianza, respecto a $W_1 + W_2 = 1$, se encuentra que son

$$W_1 = \frac{V_{22} - V_{12}}{V_{11} + V_{22} - 2V_{12}}, \qquad W_2 = \frac{V_{11} - V_{12}}{V_{11} + V_{22} - 2V_{12}}$$

y la varianza mínima es

$$V_{mín}(\hat{Y}_{MR}) = \frac{V_{11}V_{22} - V_{12}^2}{V_{11} + V_{22} - 2V_{12}} \tag{6.100}$$

Con p variante, es necesario calcular la inversa V^{ij} de la matriz V_{ij}. Entonces, la óptima $W_i = \Sigma_i/\Sigma$, donde Σ_i es la suma de los elementos en la i-ésima columna de V^{ij} y Σ es la suma de todos los elementos p^2 de V^{ij}. La varianza mínima es $1/\Sigma$.

En la práctica. las ponderaciones se determinan a partir de varianzas y covarianzas v_{ij} estimadas. De (6.7) en la Sec. 6.3,

$$v_{11} = \frac{(1-f)\hat{Y}^2}{n}(c_{yy} + c_{11} - 2c_{y1})$$

$$v_{22} = \frac{(1-f)\hat{Y}^2}{n}(c_{yy} + c_{22} - 2c_{y2})$$

donde $c_{yy} = s_y^2/\bar{y}^2$, etc. La covarianza puede expresarse como

$$v_{12} = \frac{(1-f)\hat{Y}^2}{n}(c_{yy} + c_{12} - c_{y1} - c_{y2})$$

Un método conveniente de cálculo es obtener primero la matriz

$$C = \begin{pmatrix} c_{yy} & c_{y1} & c_{y2} \\ c_{y1} & c_{11} & c_{12} \\ c_{y2} & c_{12} & c_{22} \end{pmatrix}$$

Si $v_{ij}' = nv_{ij}/(1-f)\hat{Y}^2$, la matriz v_{ij}' se obtiene fácilmente al tomar contrastes diagonales en C, esto es,

$$v_{11}' = c_{yy} + c_{11} - c_{y1} - c_{y1}$$

$$v_{12}' = c_{yy} + c_{12} - c_{y1} - c_{y2} \qquad \text{etc.}$$

El factor $(1-f)\hat{Y}^2/n$ no se necesita para calcular w_i, pero debe ser insertado cuando se calcula la varianza mínima. Entonces

$$v_{mín}(\hat{Y}_{MR}) = \frac{(1-f)\hat{Y}^2}{n}\,\frac{(v_{11}'v_{22}' - v_{12}'^2)}{(v_{11}' + v_{22}' - 2v_{12}')}. \tag{6.101}$$

En vista de la cantidad de cálculos involucrados, esta estimación probablemente se restringirá a encuestas pequeñas de alcance especializado. El método es capaz de dar un marcado aumento en precisión sobre $\hat{Y}_{R_1}$ o $\hat{Y}_{R_2}$ solas.

6.21 ESTIMADORES PRODUCTO

Si una variable auxiliar x tiene correlación negativa con y donde x y y son variables que sólo toman valores positivos, un análogo natural del estimador de razón es el estimador producto, para el cual

$$\hat{\bar{Y}}_p = \frac{\bar{x}}{\bar{X}}\bar{y}; \qquad \hat{Y}_p = N\frac{\bar{x}}{\bar{X}}\bar{y} \tag{6.102}$$

Por la expansión usual en serie de Taylor, el análogo de (6.8) para el estimador producto en una muestra aleatoria simple grande, es

$$(cv)^2 = \frac{(1-f)}{n}(C_{yy} + C_{xx} + 2C_{yx}) \qquad (6.103)$$

donde $(cv)^2$ es el cuadrado del coeficiente de variación de $\hat{\bar{Y}}_p$ o de $\hat{Y}_p$. P.S.R.S. Rao y Mudholkar (1967) han extendido el estimador de razón multivariante de Olkin a una combinación ponderada de estimadores de razón (para x_i positivamente correlacionados con y) y estimadores producto (para x_i negativamente correlacionados con y).

EJERCICIOS

6.1. Una encuesta piloto de 21 hogares dio los siguientes datos para número de miembros (x), hijos (y_1), automóviles (y_2) y receptores de TV (y_3).

x	y_1	y_2	y_3	x	y_1	y_2	y_3	x	y_1	y_2	y_3
5	3	1	1	2	0	0	0	6	3	2	0
2	0	1	1	3	1	1	1	4	2	1	1
4	1	2	0	2	0	1	0	4	2	1	1
4	2	1	1	6	4	2	1	3	1	0	1
6	4	1	1	3	1	0	0	2	0	1	1
3	1	1	2	4	2	1	1	4	2	1	1
5	3	1	1	5	3	1	1	3	1	1	1

Suponiendo que se conoce el total de población X, ¿recomendaría usted que se usaran estimaciones de razón en lugar de expansiones simples para estimar el total de hijos, automóviles y receptores de TV?

6.2. En un campo de cebada se pesaron el grano, y_i, y el grano más la paja, x_i, en cada una de un gran número de unidades de muestreo localizadas al azar por todo el campo. También se pesó la cosecha total (grano más paja) del campo completo. Se obtuvieron los siguientes datos: $c_{yy} = 1.13$, $c_{yx} = 0.78$, $c_{xx} = 1.11$. Calcule la ganancia en precisión obtenida al estimar el rendimiento en grano del campo, de la razón grano a cosecha total en lugar de usar el rendimiento medio de grano por unidad.

Se necesitan 20 min para cortar, trillar y pesar el grano en cada unidad, 2 min para pesar la paja en cada unidad y 2 horas para recolectar y pesar toda la cosecha del campo. ¿Cuántas unidades se deben tomar por campo para que la estimación de razón sea más económico que la media por unidad?

6.3. Para los datos en el Tabla 6.1, $\hat{Y}_R = 28{,}367$ y $c_{\bar{y}\bar{y}} = 0.0142068$, $c_{\bar{y}\bar{x}} = 0.0146541$, $c_{\bar{x}\bar{x}} = 0.0156830$. Calcule los límites cuadráticos de confianza al 95% para Y y compárelos con los límites encontrados mediante la aproximación normal.

6.4. Los valores de y y x se miden para cada unidad en una muestra aleatoria simple de una población. Si se conoce $\bar{X}$, la media de población de x, ¿cuál de los procedimientos siguientes recomienda, para estimar $\bar{Y}/\bar{X}$? (a) Usar siempre $\bar{y}/\bar{X}$. (b) Usar algunas veces $\bar{y}/\bar{X}$ y otras veces $\bar{y}/\bar{x}$. (c) Siempre usar $\bar{y}/\bar{x}$. Dé sus razones.

6.5. Los datos siguientes son de una pequeña población artificial con $N = 8$ y dos estratos de mismo tamaño.

Estrato 1		Estrato 2	
x_{1i}	y_{1i}	x_{2i}	y_{2i}
2	0	10	7
5	3	18	15
9	7	21	10
15	10	25	16

Para una muestra aleatoria estratificada en la cual $n_1 = n_2 = 2$, compare los ECM de $\hat{Y}_{Rs}$ y $\hat{Y}_{Rc}$ encontrando los resultados para todas las muestras posibles. ¿Hasta qué punto es la diferencia en el ECM debida a sesgos en las estimaciones?

6.6. En el Ejercicio 6.5 calcule la varianza dada usando el método de Lahiri de selección de la muestra dentro de cada estrato y una estimación de razón separada.

6.7. Cuarenta y cinco estados de los Estados Unidos (excluyendo los 5 más grandes) fueron colocados en 9 estratos con cinco estados cada uno, los estados dentro del mismo estrato tienen aproximadamente la misma razón de población de 1950 a 1940. Una muestra aleatoria estratificada con $n_h = 2$ dio los siguientes resultados para la población en 1960 (y) y 1950 (x), en millones.

	Estrato								
	1	2	3	4	5	6	7	8	9
y_{h1}	0.23	0.63	0.97	2.54	4.67	4.32	4.56	1.79	2.18
x_{h1}	0.13	0.50	0.91	2.01	3.93	3.96	4.06	1.91	1.90
y_{h2}	4.95	2.85	0.61	6.07	3.96	1.41	3.57	1.86	1.75
x_{h2}	2.78	2.38	0.53	4.84	3.44	1.33	3.29	2.01	1.32

Dado que el total de población en 1950, X, es 97.94, estime la población en 1960 mediante la estimación de razón combinada. Encuentre el error estándar de su estimación mediante el método abreviado de Keyfitz (Sec. 6.13). El total correcto para 1960 fue 114.99. ¿Coincide su estimación con esta cifra dentro de los errores de muestreo?

6.8. En el ejemplo de una razón bivariante dado por Olkin, se tomó una muestra de 50 ciudades de una población de 200 ciudades. Las variables y, x_1, x_2 son número de habitantes por ciudad en 1950, 1940 y 1930, respectivamente. Para la población, $\bar{Y} = 1699$, $\bar{X}_1 = 1482$, $\bar{X}_2 = 1420$ (en cientos) y, para la muestra $\bar{y} = 1896$, $\bar{x}_1 = 1693$, $\bar{x}_2 = 1643$. La matriz C según se define en la Sec. 6.20 es

	y	x_1	x_2
y	1.213	1.241	1.256
x_1	1.241	1.302	1.335
x_2	1.256	1.335	1.381

Estime $\bar{Y}$ mediante (a) la media de la muestra, (b) la razón del número de habitantes de 1950 a 1940 y (c) la estimación bivariante. Calcule el error estándar estimado de cada estimación.

6.9. Pruebe que con el método de Midzuno de selección de muestras (Sec. 6.15) la probabilidad de que se extraiga cualquier muestra específica es

$$\frac{(n-1)!(N-n)!}{(N-1)!}\frac{\sum^{n}(x_i)}{X}$$

6.10. En poblaciones pequeñas, el término guía del sesgo de $\hat{R}$ en muestras aleatorias simples de tamaño n es de la forma

$$E(\hat{R}-R)=\frac{b_1(1-f)}{n}=\frac{b_1}{n}-\frac{b_1}{N}$$

donde b_1 no depende de n, N. Si $n = mg$ y la muestra se divide al azar en g grupos de tamaño m, sea $\hat{R}_j=\sum y/\sum x$ tomados entre los $(n-m)$ miembros restantes de la muestra cuando se omite de ésta el grupo j. Demuestre que en el sesgo de la estimación

$$w\hat{R}-(w-1)\hat{R}_j$$

ambos términos en b_1 se anulan si $w=g[1-(n-m)/N]$.

CAPITULO 7

Estimadores de Regresión

7.1 LA ESTIMACION DE REGRESION LINEAL

Al igual que la estimación de razón de regresión lineal se ha diseñado para aumentar la precisión con el uso de una variable auxiliar x_i correlacionada con y_i. Cuando se examine la relación entre y_i y x_i, podrá encontrarse que aunque se trata de una relación aproximadamente lineal, la línea no pasa por el origen. Esto sugiere una estimación basada sobre la regresión lineal de y_i respecto a x_i y no sobre la razón de ambas variables.

Supongamos que cada y_i y x_i se obtienen para todas y cada una de las unidades de la muestra y que se conoce la media de población $\bar{X}$ de los valores x_i. La estimación de regresión lineal de $\bar{Y}$, media de población de los y_i es

$$\bar{y}_{lr}=\bar{y}+b(\bar{X}-\bar{x}) \tag{7.1}$$

donde el subíndice *lr* denota *regresión lineal* y b es una estimación del cambio en y por incremento unitario de x. El fundamento de esta estimación es que si $\bar{x}$ está por debajo del promedio, también deberíamos esperar que $\bar{y}$ lo estuviera por una cantidad $b(\bar{X}-\bar{x})$ debido a la regresión de y_i sobre x_i. Para una estimación del total de población Y, tomamos $\hat{Y}_{lr}=N\bar{y}_{lr}$.

Watson (1937) utilizó una regresión del área foliar respecto al peso foliar, para estimar el área promedio de las hojas de una planta. El procedimiento consistía en pesar todas las hojas de la planta. Se determinó el área y el peso de cada hoja, para una pequeña muestra de hojas. Posteriormente se ajustó la media muestral del área por medio de la regresión respecto al peso de las hojas. La cuestión es desde luego que es fácil pesar una hoja, pero determinar su área es laborioso.

Este ejemplo ilustra una situación general en que la estimación de regresión es útil. Supóngase que podemos estimar rápidamente el x_i de cierta característica para cada unidad y que también podemos, por un método más costoso, determinar el valor correcto de y_i de la característica para una muestra aleatoria simple de las unidades. Un experto en ratas podría efectuar una rápida estimación visual del número de ratas en cada manzana de una parte de la ciudad y después por medio de trampas determinar el número real en cada una de las manzanas de una muestra aleatoria simple. En otra aplicación que describe Yates (1960), se hizo una estimación visual del volumen de madera en cada lote de una población de lotes de un décimo de acre, y después se midió el volumen real para una muestra de los lotes. La estimación de regresión

$$\bar{y}+b(\bar{X}-\bar{x})$$

ajusta la media de muestra de las mediciones reales por la regresión de las mediciones reales efectuadas respecto a las estimaciones rápidas, que no necesariamente carecerán de sesgo. Si $x_i - y_i = D$, de modo que la estimación rápida sea perfecta, excepto por un sesgo constante D, entonces con $b = 1$ la estimación de regresión viene a ser

$$\bar{y}+(\bar{X}-\bar{x})=\bar{X}+(\bar{y}-\bar{x})$$

= (media de población de la estimación rápida) + (ajuste por sesgo)

Si no se supone un modelo de regresión lineal, nuestro conocimiento de las propiedades de la estimación de regresión tiene los mismos alcances que el de la estimación de razón. La estimación de regresión es consistente, en el sentido trivial de que cuando la muestra comprende a toda la población, $\bar{x}=\bar{X}$, y la estimación de regresión se reduce a $\bar{Y}$. Como se mostrará, la estimación suele estar sesgada, pero la razón del sesgo al error estándar se reduce cuando la muestra es grande. Estamos frente a una fórmula para la varianza de la estimación en el caso de una muestra grande, pero será necesaria más información respecto a la distribución de la estimación en muestras pequeñas y respecto al valor de n que se necesita para utilizar prácticamente los resultados de grandes muestras.

Por una elección adecuada de b, la estimación de regresión incluye, como casos particulares, la media por unidad y la estimación de razón. Es obvio que si b se toma como igual a cero, $\bar{y}_{lr}$ se reduce a $\bar{y}$. Si $b=\bar{y}/\bar{x}$,

$$\bar{y}_{lr}=\bar{y}+\frac{\bar{y}}{\bar{x}}(\bar{X}-\bar{x})=\frac{\bar{y}}{\bar{x}}\bar{X}=\hat{\bar{Y}}_R \tag{7.2}$$

7.2 ESTIMACIONES DE REGRESION CON b PREASIGNADA

Aunque en la mayoría de las aplicaciones se estima b a partir de los resultados de la muestra, en ocasiones es razonable elegir previamente el valor de b. En encuestas repetidas, cálculos anteriores pueden mostrar que los valores de muestra de b permanecen razonablemente constantes; o si x es el valor de y en un censo reciente, un conocimiento general de la población podría sugerir que b no se aleja mucho de la unidad, de modo que se elige $b = 1$. Como la teoría del muestreo de estimaciones de regresión con b preasignada es simple e instructiva, consideraremos este caso para empezar.

Teorema 7.1. En el muestreo aleatorio simple, donde b_0 es una constante asignada, la estimación de regresión lineal

$$\bar{y}_{lr}=\bar{y}+b_0(\bar{X}-\bar{x})$$

es insesgada, con varianza

$$V(\bar{y}_{lr})=\frac{1-f}{n}\frac{\sum_{i=1}^{N}[(y_i-\bar{Y})-b_0(x_i-\bar{X})]^2}{N-1} \tag{7.3}$$

$$=\frac{1-f}{n}(S_y^2-2b_0S_{yx}+b_0^2S_x^2) \tag{7.4}$$

Nótese que no se necesitó supuesto alguno sobre la relación entre y y x en la población finita

Prueba. Dado que b_0 es constante en muestreo repetido

$$E(\bar{y}_{lr})=E(\bar{y})+b_0E(\bar{x}-\bar{X})=\bar{Y} \tag{7.5}$$

por el Teorema 2.1. Además, $\bar{y}_{lr}$ es la media de muestra de las cantidades $y_i - b_0(x_i - \bar{X})$, cuya media de población es $\bar{Y}$. Así pues, por el Teorema 2.2,

$$V(\bar{y}_{lr})=\frac{1-f}{n}\cdot\frac{\sum_{i=1}^{N}[(y_i-\bar{Y})-b_0(x_i-\bar{X})]^2}{N-1} \tag{7.6}$$

$$=\frac{1-f}{n}(S_y^2-2b_0S_{yx}+b_0^2S_x^2) \tag{7.7}$$

Corolario. Una estimación de muestra insesgada de $V(\bar{y}_{lr})$ es

$$v(\bar{y}_{lr})=\frac{1-f}{n}\frac{\sum_{i=1}^{n}[(y_i-\bar{y})-b_0(x_i-\bar{x})]^2}{n-1} \tag{7.8}$$

$$=\frac{1-f}{n}(s_y^2-2b_0s_{yx}+b_0^2s_x^2) \tag{7.9}$$

Esto resulta de inmediato al aplicar el Teorema 2.4 a la variable auxiliar $y_i - b_0(x_i - \bar{X})$.

En este momento podemos preguntar: ¿Cuál es el mejor valor de b_0? La respuesta la da el Teorema 7.2.

Teorema 7.2. El valor de b_0 que minimiza $V(\bar{y}_{lr})$ es

$$b_0 = B = \frac{S_{yx}}{S_x^2} = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})(x_i - \bar{X})}{\sum_{i=1}^{N} (x_i - \bar{X})^2} \tag{7.10}$$

que podría llamarse el coeficiente de regresión lineal de y respecto a x en la población finita. Notar que B no depende de las propiedades de ninguna muestra extraída, por lo cual, teóricamente podría preasignarse. La varianza mínima resultante es

$$V_{mín}(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2 (1-\rho^2) \tag{7.11}$$

donde ρ es el coeficiente de correlación de la población entre y y x.

Prueba. En la expresión (7.4) para $V(\bar{y}_{lr})$, hagamos

$$b_0 = B + d = \frac{S_{yx}}{S_x^2} + d \tag{7.12}$$

Esto da

$$V(\bar{y}_{lr}) = \frac{1-f}{n}\left[S_y^2 - 2S_{yx}\left(\frac{S_{yx}}{S_x^2} + d\right) + S_x^2\left(\frac{S_{yx}^2}{S_x^4} + 2d\frac{S_{yx}}{S_x^2} + d^2\right)\right]$$

$$= \frac{1-f}{n}\left[\left(S_y^2 - \frac{S_{yx}^2}{S_x^2}\right) + d^2 S_x^2\right] \tag{7.13}$$

Es claro que ésta se minimiza cuando $d = 0$. Como $\rho^2 = S_{yx}^2/S_y^2 S_x^2$,

$$V_{mín}(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2 (1-\rho^2) \tag{7.14}$$

Puede usarse el mismo análisis para ver qué tanto se aleja b_0 de B sin incurrir en una pérdida importante de precisión. De (7.13) y (7.14),

$$V(\bar{y}_{lr}) = \frac{1-f}{n}[S_y^2(1-\rho^2) + (b_0 - B)^2 S_x^2] \tag{7.15}$$

$$= V_{mín}(\bar{y}_{lr})\left[1 + \frac{(b_0 - B)^2 S_x^2}{S_y^2(1-\rho^2)}\right] \tag{7.16}$$

Dado que $BS_x = \rho S_y$, esto se puede escribir

$$V((\bar{y}_{lr}) = V_{mín}(\bar{y}_{lr})\left[1 + \left(\frac{b_0}{B} - 1\right)^2 \frac{\rho^2}{(1-\rho^2)}\right] \tag{7.17}$$

Por lo tanto, si el incremento proporcional en la varianza debe ser menor que α, tendremos

$$\left|\frac{b_0}{B} - 1\right| < \sqrt{\alpha(1-\rho^2)/\rho^2} \tag{7.18}$$

Por ejemplo, si $\rho = 0.7$, el incremento en la varianza es inferior al 10%, ($\alpha = 0.1$), siempre que

$$\left|\frac{b_0}{B} - 1\right| < \sqrt{(0.1)(0.51)/(0.49)} = 0.32$$

La expresión (7.18) nos dice que para asegurar un incremento proporcional *pequeño* en la varianza, b_0/B deberá estar cerca de 1, si ρ es muy elevado; pero puede alejarse bastante de 1, si ρ es solamente moderado.

7.3 ESTIMACIONES DE REGRESION CUANDO b SE CALCULA A PARTIR DE LA MUESTRA

El Teorema 7.2 sugiere que si b debe calcularse a partir de la muestra, la ya conocida estimación por mínimos cuadrados de B será una estimación efectiva, esto es

$$b = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{7.19}$$

La teoría de la regresión lineal desempeña un papel prominente en la metodología estadística. Los resultados estándar de esta teoría no son enteramente adecuados a las encuestas por muestreo, porque se necesita suponer que la regresión de población de y respecto a x sea lineal, que la varianza residual de y alrededor de la línea de regresión sea constante y que la población sea infinita. Si los dos primeros supuestos son demasiado incorrectos, probablemente no se use una estimación de regresión lineal. Pero en encuestas en donde la regresión de y respecto a x se cree que es aproximadamente lineal, es útil poder usar $\bar{y}_{lr}$ sin necesidad de suponer linealidad exacta o varianza residual constante.

En consecuencia, presentamos un método que no supone ninguna relación específica entre y_i y x_i. Como en la teoría análoga para la estimación de razón, sólo se obtienen resultados de muestras grandes.

Con b como en (7.19), el estimador de regresión lineal de $\bar{Y}$ en muestras aleatorias simples es

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}) = \bar{y} - b(\bar{x} - \bar{X}) \tag{7.20}$$

En la Sec. 7.7 veremos que el estimador $\bar{y}_{lr}$, al igual que $\bar{y}_R$, tiene un sesgo del orden de $1/n$. Al encontrar el error de muestreo de $\bar{y}_{lr}$, se reemplaza el valor de b de la muestra, en (7.20) por el coeficiente de regresión de la población B, en (7.10). En el Teorema 7.3, el error cometido en esta aproximación se verá que es del orden $1/\sqrt{n}$ relativo a los términos retenidos. En primer lugar examinaremos la relación entre b y B.

Se introduce la variable auxiliar e_i definida por la relación

$$e_i = y_i - \bar{Y} - B(x_i - \bar{X}) \tag{7.21}$$

Dos propiedades de los e_i son que $\sum_N e_i = 0$ y

$$\sum_N e_i(x_i - \bar{X}) = \sum_N (y_i - \bar{Y})(x_i - \bar{X}) - B\sum_N (x_i - \bar{X})^2 = 0 \tag{7.22}$$

Por definición de B. Ahora

$$\begin{aligned} b &= \sum_n y_i(x_i - \bar{x}) / \sum_n (x_i - \bar{x})^2 \\ &= \sum_n [\bar{Y} + B(x_i - \bar{X}) + e_i](x_i - \bar{x}) / \sum (x_i - \bar{x})^2 \\ &= B + \sum_n e_i(x_i - \bar{x}) / \sum_n (x_i - \bar{x})^2 \end{aligned} \tag{7.23}$$

Un resultado necesario en el Teorema 7.3 es que $(b - B)$ sea del orden de $1/\sqrt{n}$. Por el Teorema 2.3, $\sum_n e_i(x_i - \bar{x})/(n-1)$ es una estimación insesgada de $\sum_N e_i(x_i - \bar{X})/(N-1)$ que por (7.22) es cero. Así $\sum_n e_i(x_i - \bar{x})/(n-1)$ se distribuye alrededor de una media cero en muestras repetidas. Como el error estándar de una covarianza de muestra se sabe que es de orden $1/\sqrt{n}$, $\sum_n e_i(x_i - \bar{x})/(n-1)$ es de orden $1/\sqrt{n}$. Pero $\sum (x_i - \bar{x})^2/(n-1) = s_x^2$ es de orden unitario. Por lo tanto, $(b - B)$, que a partir de (7.23) es la razón de estas dos cantidades, es de orden $1/\sqrt{n}$.

Teorema 7.3. Si b es la estimación por mínimos cuadrados de B y

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}) \tag{7.24}$$

entonces, en muestras aleatorias simples de tamaño n, con n grande,

$$V(\bar{y}_{lr}) \doteq \frac{(1-f)}{n} S_y^2 (1 - \rho^2) \tag{7.25}$$

donde $\rho = S_{yx}/S_y S_x$ es la correlación de población entre y y x.

Prueba. El error de muestreo de $\bar{y}_{lr}$ surge de la cantidad

$$\bar{y}_{lr} - \bar{Y} = \bar{y} - \bar{Y} + b(\bar{X} - \bar{x}) \tag{7.26}$$

Como aproximación, reemplacemos $\bar{y}_{lr}$ por

$$\tilde{y}_{lr} = \bar{y} + B(\bar{X} - \bar{x}) \tag{7.27}$$

donde B es el coeficiente de regresión lineal en la población de y respecto a x. El error en que incurre esta aproximación es $(B - b)(\bar{X} - \bar{x})$. Esta cantidad es de orden $1/n$ en una muestra aleatoria simple de tamaño n, ya que tanto $(b - B)$ como $(\bar{x} - \bar{X})$ son de orden $1/\sqrt{n}$. Pero el error de muestreo en $\tilde{y}_{lr}$ es del orden $1/\sqrt{n}$ dado que es el error en la media muestral de la variable auxiliar $(y_i - Bx_i)$. Por lo tanto, el término guía en $E(\bar{y}_{lr} - \bar{Y})^2$ es $V(\tilde{y}_{lr})$. Por (7.11) en muestras grandes,

$$E(\bar{y}_{lr} - \bar{Y})^2 \doteq V(\tilde{y}_{lr}) = \frac{(1-f)}{n} S_y^2 (1 - \rho^2) \tag{7.28}$$

7.4 ESTIMACION DE MUESTRA DE LA VARIANZA

Como estimación de muestra de $V(\bar{y}_{lr})$, válida para muestras grandes, podemos usar

$$v(\bar{y}_{lr}) = \frac{1-f}{n(n-2)} \sum_{i=1}^{n} [(y_i - \bar{y}) - b(x_i - \bar{x})]^2 \tag{7.29}$$

$$= \frac{1-f}{n(n-2)} \left\{ \sum (y_i - \bar{y})^2 - \frac{[\sum (y_i - \bar{y})(x_i - \bar{x})]^2}{\sum (x_i - \bar{x})^2} \right\} \tag{7.30}$$

esta última es la fórmula de cálculo abreviada usual. Se deduce como a continuación.

En el Teorema 7.3, Ec. (7.28), teníamos, dado que $S_y^2(1 - \rho^2) = S_e^2$,

$$V(\bar{y}_{lr}) \doteq \frac{(1-f)}{n} S_e^2$$

Del Teorema 2.4 tenemos como estimación insesgada de S_e^2

$$s_e^2 = \frac{1}{n-1} \sum_{i=1}^{n} (e_i - \bar{e})^2$$

Ahora bien, la Ec. (7.21) se infiere que

$$e_i - \bar{e} = (y_i - \bar{y}) - B(x_i - \bar{x}) = [(y_i - \bar{y}) - b(x_i - \bar{x})] + (b - B)(x_i - \bar{x}) \quad (7.31)$$

El segundo término de la derecha, de orden $1/\sqrt{n}$ puede despreciarse en relación al primero que es de orden unitario. De modo que en muestras grandes podemos usar

$$\frac{1}{n-1} \sum_{i=1}^{n} [(y_i - \bar{y}) - b(x_i - \bar{x})]^2 \quad (7.32)$$

como estimación de S_e^2. El divisor $(n - 2)$ en lugar de $(n - 1)$ se sugirió en (7.29) y (7.30), porque se usa en la teoría estándar de la regresión y sabemos que da una estimación insesgada de S_e^2, si la población es infinita y la regresión lineal.

7.5 COMPARACION EN MUESTRAS GRANDES CON LA ESTIMACION DE RAZON Y LA MEDIA POR UNIDAD

Para estas comparaciones, el tamaño de la muestra n deberá ser lo suficientemente grande de modo que las fórmulas aproximadas para las varianzas de las estimaciones de razón y regresión sean válidas. Las tres varianzas comparables para la media de población estimada $\bar{Y}$ son

$$V(\bar{y}_{lr}) = \frac{N-n}{Nn} S_y^2 (1 - \rho^2) \qquad \text{(regresión)}$$

$$V(\bar{y}_R) = \frac{N-n}{Nn} (S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x) \qquad \text{(razón)}$$

$$V(\bar{y}) = \frac{N-n}{Nn} S_y^2 \qquad \text{(media por unidad)}$$

Es claro que la varianza de la estimación de regresión es más pequeña que la de la media por unidad, a menos que $\rho = 0$, y en este caso las dos varianzas son iguales.

La varianza de la estimación de regresión es menor que la de la estimada de razón si

$$-\rho^2 S_y^2 < R^2 S_x^2 - 2R\rho S_y S_x \quad (7.33)$$

Esto es equivalente a las desigualdades

$$(\rho S_y - R S_x)^2 > 0 \qquad \text{o} \qquad (B - R)^2 > 0 \quad (7.34)$$

De este modo, la estimación de regresión es más precisa que la estimación de razón a menos que $B = R$. Esto ocurre cuando la relación entre y_i y x_i es una línea recta que pasa por el origen.

Ejemplo. La precisión de las estimaciones de regresión, razón y media por unidad de una muestra aleatoria simple se pueden comparar usando los datos reunidos en la lista completa de huertos de durazno descrita en la Pág. 220. En este ejemplo, y_i es la producción estimada de duraznos en una huerta, y x_i el número de árboles de durazno en la huerta. Compararemos las estimaciones de la producción total de los 256 huertos, a partir de una muestra de 100 huertos. Es dudoso que la muestra sea lo bastante grande para que las fórmulas de la varianza sean completamente válidas, ya que los cv de $\bar{y}$ y $\bar{x}$ son ambos algo más altos de 10%, pero el ejemplo servirá para ilustrar los cálculos. Los datos básicos son los siguientes:

$$S_y^2 = 6\,409 \qquad S_{yx} = 4\,434 \qquad S_x^2 = 3\,898$$

$$R = 1.270 \qquad \rho = 0.887 \qquad n = 100 \qquad N = 256$$

$$V(\hat{Y}_{lr}) = \frac{N(N-n)}{n} S_y^2 (1 - \rho^2)$$

$$= \frac{(256)(156)}{100} (6\,409)(1 - 0.787) = 545\,000$$

$$V(\hat{Y}_R) = \frac{N(N-n)}{n} (S_y^2 + R^2 S_x^2 - 2RS_{yx})$$

$$= \frac{(256)(156)}{100} [6\,409 + (1.613)(3\,898) - 2(1.270)(4\,434)]$$

$$= 573\,000$$

$$V(\hat{Y}) = \frac{N(N-n)}{n} S_y^2 = 2\,559\,000$$

Hay poco que escoger entre las estimaciones de regresión y de razón, como puede esperarse de la naturaleza de las variables. Ambas técnicas son muy superiores a la media por unidad.

El resultado precedente sobre la superioridad de la estimación de regresión lineal es sólo para muestras grandes. En pequeñas muestras de poblaciones naturales, los resultados de la estimación de regresión son desilusionantes. En ocho poblaciones naturales del tipo en el que se ha usado la estimación de razón, Rao (1969), con estudio apoyado en Montecarlo, encontró que el promedio de las razones ECM $(\hat{Y}_{lr})$/ECM $(\hat{Y}_R)$ era de 1.15 para $n = 12$, 1.36 para $n = 8$ y 1.51 para $n = 6$. Estas bajas eficiencias de $\hat{Y}_{lr}$ no se debían a sesgos más grandes en las estimaciones de regresión, porque las correspondientes razones de varianzas eran casi las mismas.

7.6 EXACTITUD DE LAS FORMULAS DE MUESTRAS GRANDES PARA $V(\bar{y}_{lr})$ y $v(\bar{y}_{lr})$

No se tienen resultados analíticos generales sobre la exactitud de las fórmulas aproximadas (7.25) para $V(\bar{y}_{lr})$ y (7.29) para $v(\bar{y}_{lr})$

en muestras pequeñas o moderadas. Los estimadores aproximados en (7.25) y (7.29) son

$$V(\bar{y}_{lr}) = \frac{(1-f)}{n} S_y^2 (1-\rho^2) \tag{7.25}$$

$$v(\bar{y}_{lr}) = \frac{1-f}{n(n-2)} \sum_i^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2 \tag{7.29}$$

Supongamos que los y_i para $i = 1, 2, \ldots N$ son una muestra aleatoria de una población infinita en el modelo

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{7.35}$$

donde, para x_i fijo, las ε_i están distribuidas independientemente con media nula, varianza $\sigma_\varepsilon^2 = \sigma_y^2(1 - \rho^2)$; con este modelo, Cochran (1942) dio el resultado para los términos de orden $1/n^2$.

$$EV(\bar{y}_{lr}) = \frac{(1-f)}{n} \sigma_y^2 (1-\rho^2) \left(1 + \frac{1}{n-3} + \frac{2G_1^2}{n^2}\right) \tag{7.36}$$

donde $G_1 = k_{3x}/\sigma_x^3$ es la medida de Fisher de la asimetría relativa de la distribución de x. Dado que $S_y^2(1 - \rho^2)$ en (7.25) es una estimación insesgada de $\sigma_y^2(1 - \rho^2)$ en este modelo, (7.36) sugiere que con x simétricamente distribuida, el porcentaje de subestimación mediante $V(\bar{y}_{lr})$ es $100/(n - 2)$ con este modelo.

De estudios con el método de Montecarlo, en ocho poblaciones naturales pequeñas, (Rao, 1968), la Tabla 7.1 muestra para $n = 6$, 8, 12, el porcentaje de subestimación promedio de la varianza de $\bar{y}_{lr}$ por las aproximaciones de $V(\bar{y}_{lr})$ y $v(\bar{y}_{lr})$.

Para x_i simétrica, la Fórmula (7.36) sugiere porcentaje de subestimaciones, mediante $V(\bar{y}_{lr})$ iguales a 25, 17 y 10% para $n = 6$, 8, 12. Los porcentajes para $V(\bar{y}_{lr})$ en la Tabla 7.1 son sustancialmente más elevados por cantidades que difícilmente se deben a la simetría de x en estas poblaciones, sino más bien se atribuyen a las deficiencias en el modelo lineal. Las subestimaciones son aún mayores para las estimaciones de muestra de la varianza $v(\bar{y}_{lr})$. Además, al comparar con la Tabla 6.2, Pág. 211. que se aplica a la estimación de razón en las mismas ocho poblaciones, se ve que el porcentaje de las subestimaciones en V y v para $\bar{y}_{lr}$ son, cuando menos, el doble de las de $\bar{y}_R$ en muestras del mismo tamaño.

TABLA 7.1. PORCENTAJE DE SUBESTIMACIÓN PROMEDIO DE LA VARIANZA DE $\bar{y}_{lr}$

Estimador	n = 6	8	12
$V(\bar{y}_{lr})$ en (7.25)	38	34	28
$v(\bar{y}_{lr})$ en (7.29)	48	42	33

7.7 SESGO DE LA ESTIMACION DE REGRESION LINEAL

El estimador $\bar{y}_{lr}$ tiene un sesgo de $1/n$ en muestreo aleatorio simple. Tenemos

$$E(\bar{y}_{lr}) = \bar{Y} - Eb(\bar{x} - \bar{X}) \tag{7.37}$$

de modo que como expresión del sesgo, tenemos $-Eb(\bar{x} - \bar{X}) = -\text{cov}(b, \bar{x})$. El término guía del sesgo, resulta ser

$$\frac{-(1-f)}{n} \frac{Ee_i(x_i - \bar{X})^2}{S_x^2} \tag{7.38}$$

Este término representa una contribución de la componente *cuadrática* de la regresión de $\bar{y}$ respecto a $\bar{x}$. De modo que si una gráfica de muestra de y_i contra x_i aparece aproximadamente lineal, no existe riesgo de un sesgo importante en $\bar{y}_{lr}$.

Para mostrar la validez de (7.38), habrá que efectuar algunos desarrollos algebraicos. Por (7.23), Pág. 244,

$$b = B + \frac{\sum^n e_i(x_i - \bar{x})}{\sum^n (x_i - \bar{x})^2} \tag{7.39}$$

Reemplacemos $\sum^n (x_i - \bar{x})^2$ por su término guía, nS_x^2 y escribamos

$$\sum^n e_i(x_i - \bar{x}) = \sum^n e_i(x_i - \bar{X}) + n\bar{e}(\bar{X} - \bar{x}) \tag{7.40}$$

Por lo tanto, el término guía del sesgo $-Eb(\bar{x} - \bar{X})$ de $\bar{y}_{lr}$ es el promedio de

$$\frac{-\sum^n e_i(x_i - \bar{X})(\bar{x} - \bar{X})}{nS_x^2} + \frac{\bar{e}(\bar{x} - \bar{X})^2}{S_x^2} \tag{7.41}$$

Sea $u_i = e_i(x_i - \bar{X})$. Por (7.22) su media de población $\bar{U} = 0$. Por lo tanto, el valor promedio del primer término en (7.41) puede escribirse

$$\frac{-E(\bar{u} - \bar{U})(\bar{x} - \bar{X})}{S_x^2} = -\frac{(1-f)}{n} \frac{E(u_i - \bar{U})(x_i - \bar{X})}{S_x^2} \tag{7.42}$$

por el Teorema (2.3) (Pág. 199) para el valor promedio de una covarianza de muestra en muestreo aleatorio simple. Esto a su vez es igual a (7.38), o sea

$$-\frac{(1-f)}{n}\frac{Ee_i(x_i-\bar{X})^2}{S_x^2}$$

En el segundo término de (7.41), $\bar{e}$ es $O(1/\sqrt{n})$ y $(\bar{x}-\bar{X})^2$ es $O(1/n)$, de modo que este término es de un orden menor que (7.38). En consecuencia, (7.38) es el término guía del sesgo de $\bar{y}_{lr}$.

7.8 EL ESTIMADOR DE REGRESION LINEAL EN UN MODELO DE REGRESION LINEAL

Supongamos que los valores de la población finita $y_i (i = 1, 2, \ldots N)$ se extraen aleatoriamente de una superpoblación infinita en la que

$$y = \alpha + \beta x + \varepsilon \tag{7.43}$$

donde los ε son independientes, con media 0 y varianza σ_ε^2 para x fijo. Al sustituir directamente a partir del modelo encontramos que

$$b = \frac{\sum_1^n y_i(x_i-\bar{x})}{\sum_1^n (x_i-\bar{x})^2} = \beta + \frac{\sum_1^n \varepsilon_i(x_i-\bar{x})}{\sum_1^n (x_i-\bar{x})^2} \tag{7.44}$$

$$\bar{y}_{lr} - \bar{Y} = (\bar{\varepsilon}_n - \bar{\varepsilon}_N) + (\bar{X}-\bar{x})\frac{\sum_1^n \varepsilon_i(x_i-\bar{x})}{\sum_1^n (x_i-\bar{x})^2} \tag{7.45}$$

donde $\bar{\varepsilon}_n$ y $\bar{\varepsilon}_N$ son las medias sobre la muestra y sobre la población finita. De (7.45) vemos que en este modelo, $E(\bar{y}_{lr} - \bar{Y}) = 0$, de modo que $\bar{y}_{lr}$ es insesgada según modelo, para cualquier tamaño de muestra.

En lo que respecta a la varianza, de (7.45) vemos que para un conjunto de x dado,

$$V(\bar{y}_{lr}) = E(\bar{y}_{lr} - \bar{Y})^2 = \sigma_\varepsilon^2\left[\left(\frac{1}{n}-\frac{1}{N}\right) + \frac{(\bar{X}-\bar{x})^2}{\sum_1^n (x_i-\bar{x})^2}\right] \tag{7.46}$$

Este resultado es válido para todo $n > 1$ y cualquier muestra seleccionada únicamente por valores de x. Este método y su generalización al caso de varianzas residuales fueron dados por Royall (1970). En este modelo, un plan de muestreo determinado que tuviera éxito en hacer $\bar{x} = \bar{X}$ minimizará $V(\bar{y}_{lr})$ para un n dado.

También, para cualquier muestra elegida exclusivamente según los valores de x_i, el estimador de mínimos cuadrados usual

$$s_\varepsilon^2 = \sum_1^n [(y_i-\bar{y}) - b(x_i-\bar{x})]^2/(n-2) \tag{7.47}$$

es un estimador insesgado según modelo de σ_ε^2 para $n > 2$.

Por lo tanto, en problemas en los que se aplica este modelo, se pueden establecer resultados simples y exactos para la varianza y la media de $\bar{y}_{lr}$ válidos para cualquier muestra de tamaño mayor a 2 y que sólo requieran una selección de muestra, según los valores de x_i; el elemento aleatorio lo proporciona gratis la distribución de los ε supuesta en el modelo.

7.9 ESTIMACIONES DE REGRESION EN MUESTREO ESTRATIFICADO

Al igual que para la estimación de razón, se pueden hacer dos tipos de estimación de regresión en el muestreo aleatorio estratificado. En la primera estimación $\bar{y}_{lrs}$ (s significa separado) se calcula una estimación de la regresión separada para cada media de estrato, o sea,

$$\bar{y}_{lrh} = \bar{y}_h + b_h(\bar{X}_h - \bar{x}_h) \tag{7.48}$$

Entonces, con $W_h = N_h/N$,

$$\bar{y}_{lrs} = \sum_h W_h \bar{y}_{lrh} \tag{7.49}$$

Esta estimación es apropiada cuando se piensa que los verdaderos coeficientes de regresión B_h varían entre los estratos.

La segunda estimación de regresión $\bar{y}_{lrc}$ (c significa combinado) es apropiada cuando se ha supuesto que los B_h son pequeños en todos los estratos. Para calcular $\bar{y}_{lrc}$, encontramos primero

$$\bar{y}_{st} = \sum_h W_h \bar{y}_h \qquad \bar{x}_{st} = \sum_h W_h \bar{x}_h$$

Entonces

$$\bar{y}_{lrc} = \bar{y}_{st} + b(\bar{X} - \bar{x}_{st}) \tag{7.50}$$

Las dos estimaciones se considerarán inicialmente en el caso en que los b_h y b se eligen de antemano, pues sus propiedades son excepcionalmente simples en esta situación. De la Sec. 7.2, se sabe que $\bar{y}_{lrh}$ es una estimación insesgada de $\bar{Y}_h$, de modo que $\bar{y}_{lrs}$ es una estimación insesgada de $\bar{Y}$. Como el muestreo es independiente en estratos diferentes, vemos por el Teorema 7.1 que

$$V((\bar{y}_{lrs}) = \sum_h \frac{W_h^2(1-f_h)}{n_h}(S_{yh}^2 - 2b_h S_{yxh} + b_h^2 S_{xh}^2) \tag{7.51}$$

El Teorema 7.2 muestra que $V(\bar{y}_{lrs})$ es mínima cuando $b_h = B_h$, el verdadero coeficiente de regresión en el estrato h. El valor mínimo de la varianza se puede escribir

$$V_{mín}(\bar{y}_{lrs}) = \sum_h \frac{W_h^2(1-f_h)}{n_h}\left(S_{yh}^2 - \frac{S_{yxh}^2}{S_{xh}^2}\right) \tag{7.52}$$

Volviendo la estimación combinada con b preasignado, (7.50) muestra que y_{lrc} es también una estimación insesgada de $\bar{Y}$. Dado que $\bar{y}_{lrc}$ es la estimación usual de una muestra estratificada para la variable $y_{hi} + b(\bar{X} - x_{hi})$, podemos aplicar el Teorema 5.3 a esta variable, lo que da como resultado

$$V(\bar{y}_{lrc}) = \sum_h \frac{W_h^2(1-f_h)}{n_h}(S_{yh}^2 - 2bS_{yxh} + b^2 S_{xh}^2) \tag{7.53}$$

El valor de b que hace mínima esta varianza es

$$B_c = \sum_h \frac{W_h^2(1-f_h)S_{yxh}}{n_h} \Big/ \sum_h \frac{W_h^2(1-f_h)S_{xh}^2}{n_h} \tag{7.54}$$

La cantidad B_c es una media ponderada de los coeficientes de regresión de los estratos $B_h = S_{yxh}/S_{xh}^2$. Si escribimos

$$a_h = \frac{W_h^2(1-f_h)}{n_h} S_{xh}^2 \text{ ; entonces } B_c = \Sigma a_h B_h / \Sigma a_h.$$

De (7.52) y (7.53), con B_c en lugar de b, encontramos

$$V_{min}(\bar{y}_{lrc}) - V_{mín}(\bar{y}_{lrs}) = \sum a_h B_h^2 - (\sum a_h) V_c^2$$
$$= \sum a_h (B_h - B_c)^2 \tag{7.55}$$

Este resultado muestra que con las elecciones óptimas, la estimación separada tiene una varianza menor que la estimación combinada, a menos que B_h sea el mismo en todos los estratos.

Estas elecciones óptimas serían, desde luego, un conocimiento avanzado de los valores de S_{yxh} y S_{xh}^2.

7.10 COEFICIENTES DE REGRESION ESTIMADOS DE LA MUESTRA

El análisis precedente es útil al indicar el tipo de estimaciones b_h y b de la muestra que pueden ser eficientes cuando se buscan estimaciones de regresión. Con la estimación separada, el análisis sugiere que tomemos

$$b_h = \frac{\sum_i (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_i (x_{hi} - \bar{x}_h)^2} \tag{7.56}$$

la estimación de mínimos cuadrados dentro de estrato de B_h.

Al aplicar el Teorema 7.3 a cada estrato, tenemos

$$V(\bar{y}_{lrs}) = \sum_h \frac{W_h^2(1-f_h)}{n_h} S_{yh}^2(1-\rho_h^2) \tag{7.57}$$

siempre que el tamaño de muestra n_h sea grande en todos los estratos. Para obtener una estimación de varianza de la muestra, se sustituye

$$s_{y\cdot xh}^2 = \frac{1}{n_h - 2}\left[\sum_i (y_{hi} - \bar{y}_h)^2 - b_h^2 \sum_i (x_{hi} - \bar{x}_h)^2\right] \tag{7.58}$$

en lugar de $S_{yh}^2(1 - \rho_h^2)$ en (7.57).

La estimación $\bar{y}_{lrs}$ presenta la misma dificultad que la correspondiente estimación de razón, o sea que la razón del sesgo al error estándar puede llegar a ser apreciable. De la Sec. 7.7 resulta que las estimaciones de regresión $\bar{y}_{lrh}$ en los estratos individuales pueden tener sesgos de orden $1/n_h$ y los sesgos pueden ser del mismo signo en todos los estratos, de tal manera que el sesgo general en $\bar{y}_{lrs}$ también puede ser de orden $1/n_h$. Dado que el término principal en el sesgo viene de la regresión cuadrática de y_{hi} en x_{hi}, como se mostró en la Sec. 7.7, este peligro es más agudo cuando la relación entre las variables se aproxima más bien al tipo cuadrático que al lineal.

Con la estimación combinada vimos que la varianza es mínima cuando $b = B_c$, según se define en (7.54). Esto sugiere que tomemos

$$b_c = \sum_h \frac{W_h^2(1-f_h)}{n_h(n_h-1)} \sum_i (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h) \Big/ \sum_h \frac{W_h^2(1-f_h)}{n_h(n_h-1)} \sum_i (x_{hi} - \bar{x}_h)^2$$

como una estimación de B_c. Si la estratificación es proporcional y si podemos reemplazar $(n_h - 1)$ en b_c por n_h, b_c se reduce a la conocida estimación combinada de mínimos cuadrados

$$b_c' = \sum_h \sum_i (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h) \Big/ \sum_h \sum_i (x_{hi} - \bar{x}_h)^2$$

En ciertas circunstancias se prefieren otras estimaciones que b_c o b_c'. Por ejemplo, si los verdaderos coeficientes de regresión B_h son los mismos en todos los estratos, pero las varianzas residuales alrededor de la línea de regresión difieren sustancialmente de un estrato al otro, puede ser más precisa una media ponderada diferente

de los b_h, ponderada inversamente con la varianza estimada. Sin embargo, la ganancia en precisión por lo que respecta a $\bar{y}_{lrc}$ es posible que sea pequeña.

Dado que

$$\bar{y}_{lrc} - \bar{Y} = \bar{y}_{st} - \bar{Y} + b_c(\bar{X} - \bar{x}_{st})$$
$$= [\bar{y}_{st} - \bar{Y} + B_c(\bar{X} - \bar{x}_{st})] + (b_c - B_c)(\bar{X} - \bar{x}_{st}) \qquad (7.59)$$

resulta que si los errores de muestreo de b_c son despreciables

$$V(\bar{y}_{lrc}) = \sum_h \frac{W_h^2(1-f_h)}{n_h}(S_{yh}^2 - 2B_cS_{yxh} + B_c^2S_{xh}^2) \qquad (7.60)$$

Como estimación de $V(\bar{y}_{lrc})$, podemos tomar

$$v(\bar{y}_{lrc}) = \sum_h \frac{W_h^2(1-f_h)}{n_h(n_h-1)} \sum_i [(y_{hi} - \bar{y}_h) - b_c(x_{hc} - \bar{x}_h)]^2 \qquad (7.61)$$

7.11 COMPARACION DE LOS DOS TIPOS DE ESTIMACIONES DE REGRESION

No se pueden dar reglas absolutas para decidir si la estimación separada o combinada es mejor en una situación específica: se requiere ejercer buen juicio para hacer la elección. Los defectos de la estimación separada son que está más expuesta a sesgos cuando las muestras son pequeñas dentro de los estratos individuales y que su varianza tiene una mayor contribución debida a los errores de muestreo en los coeficientes de regresión. El defecto de la estimación combinada es que su variancia es inflada si los coeficientes de regresión de la población difieren de estrato a estrato.

Si tenemos confianza en que las regresiones son lineales y si B_h parece ser el mismo en todos los estratos, hasta donde pueda juzgarse, debe preferirse la estimación combinada. Si las regresiones parecen lineales (en tal forma que el peligro de sesgo parece pequeño) pero B_h parece variar de estrato a estrato, es recomendable la estimación separada. Si hay alguna curvilinealidad en las regresiones cuando se usa una estimación de regresión lineal, la estimación combinada es probablemente más segura a menos que las muestras sean grandes en todos los estratos.

Mickey (1959) y Williams (1963) han desarrollado estimadores del tipo de regresión que son insesgados, pero aún no se han probado lo suficiente. Rao (1969) encontró que el estimador de Mickey generalmente es inferior a los estimadores de regresión estándar, y a los estimadores de razón en poblaciones naturales. Una versión del estimador llamado de jackknifed también puede construirse en esta situación. Con $n = mg$ sea $\bar{y}'_{(lr)j}$ la estimación de regresión estándar, calculada a partir de la muestra en donde se omitió el grupo j, $(j = 1, 2, \ldots g)$. Entonces, la forma jackknifed es

$$\bar{y}_{(lr)Q} = g\bar{y}_{lr} - (g-1)\left(\sum^{g} \bar{y}'_{(lr)j}\right)\Big/ g \qquad (7.62)$$

EJERCICIOS

7.1. Un agricultor experimentado hace una estimación ocular del peso x_i de los duraznos en cada árbol en un huerto de $N = 200$ árboles. Encuentra un peso total de $X = 11\,600$ *lb*. Los duraznos en una muestra aleatoria simple de 10 árboles se cosechan y pesan, con los siguientes resultados:

		Número de árbol										
		1	2	3	4	5	6	7	8	9	10	Total
Peso real	y_i	61	42	50	58	67	45	39	57	71	53	543
Peso estimado	x_i	59	47	52	60	67	48	44	58	76	58	569

Como una estimación del peso total real Y, tomamos

$$\hat{Y} = N[\bar{X} + (\bar{y} - \bar{x})]$$

Calcule la estimación y encuentre su error estándar.

7.2. ¿Daría la estimación de regresión lineal, con el b de mínimos cuadrados de la muestra, una estimación más precisa en 7.1?

7.3. De los datos de la muestra en la Tabla 6.1 (Pág. 197) calcule la estimación de regresión del número total de habitantes en 1930 en las 196 ciudades grandes. Encuentre el error estándar aproximado de esta estimación y compare su precisión con la de la estimación de razón.

7.4. En el Ejercicio 7.3 encuentre el número total estimado de habitantes y su error estándar si se toma b como 1.

7.5. En la siguiente población con $N = 5$, verifique (a) que la regresión de y en x es lineal y (b) que la estimación de regresión lineal no tiene sesgo en muestras aleatorias con $n = 3$. Los pares (y, x) son (3, 0), (5, 0). (8, 2), (8, 3), (12, 3).

7.6. Una medición burda x, hecha en cada unidad, está relacionada a la verdadera medida y en la unidad por la ecuación

$$x = y + e + d$$

donde d es un sesgo constante y e es un error de medición, no correlacionado con y, el cual tiene media cero y varianza S_e^2 en la población, que se supone infinita. En muestras aleatorias simples de tamaño n, compare las varianzas de (a) la estimación de "diferencia" $[\bar{y} + (X - \bar{x})]$ de la media $\bar{Y}$ y (b) la estimación de regresión lineal, usando el valor de b que da la varianza mínima. (Las varianzas pueden depender de S_y^2.)

7.7. Al desarrollar todos los casos posibles compare los ECM de las estimaciones de regresión separada y combinada del total Y de la siguiente población, cuando se extraen muestras aleatorias simples de tamaño 2 de cada estrato. Para cada estimación, ¿cuánto es el sesgo de contribución para ECM?

Estrato 1		Estrato 2	
x_{1i}	y_{1i}	x_{2i}	y_{2i}
4	0	5	7
6	3	6	12
7	5	8	13

Use las estimaciones de mínimos cuadrados ordinarios de los B, b_h y b_c en la Pág. 253.

7.8. En la población del Ejercicio 7.7 mostrar que si el B óptimo preasignado pudiese usarse en cada caso, $V(\hat{Y}_{lrs}) = 4.39$, $V(\hat{Y}_{lrc}) = 4.43$, $= 4.43$ siendo ambas estimaciones insesgadas.

7.9. Por el mismo método, comparar el ECM de las estimaciones de *razón* separada y combinada en la población del Ejercicio 7.7. Dado que la razón Y/X es $8/17 = 0.47$ en el estrato 1, y $32/19 = 1.68$ en el estrato 2, la teoría de muestras grandes sugeriría que $\hat{Y}_{Rs}$ fuese mayor que $\hat{Y}_{Rc}$. Sin embargo, se encontrará que en estas muestras muy pequeñas, $\hat{Y}_{Rc}$ tiene el menor ECM. Su superioridad no se debe a que tenga un sesgo menor, ni a que $\hat{Y}_{Rs}$ o $\hat{Y}_{Rc}$ estén materialmente sesgadas. Otra discordancia con respecto a la teoría de las muestras grandes es, como se verá, que $\hat{Y}_{Rs}$ y $\hat{Y}_{Rc}$ tienen ECM menores que las estimaciones de regresión correspondientes.

CAPITULO 8

Muestreo Sistemático

8.1 DESCRIPCION

A primera vista, este método de muestreo es muy diferente al muestreo aleatorio simple. Supongamos que las N unidades de la población se numeran de 1 a N en cierto orden. Para elegir una muestra de n unidades, tomamos una unidad al azar entre las k primeras y luego tomamos las subsecuentes a intervalos de k. Así, por ejemplo, si k es 15 y la primera unidad que se extrae es la número 13, entonces las subsecuentes se numeran 28, 43, 58, etc. La selección de la primera unidad determina toda la muestra, que se denomina muestra de *todas* las k-ésimas unidades.

A continuación veremos las ventajas aparentes de este método respecto al muestreo aleatorio simple.

1. Es más fácil sacar una muestra y a menudo, más fácil hacerlo sin cometer errores. Esta es una ventaja particular cuando la extracción se hace en el área. Aunque la extracción se haga en una oficina, este método puede ahorrar mucho tiempo. Por ejemplo, si las unidades se describen en tarjetas del mismo tamaño y se colocan en un archivero, se pueden extraer las tarjetas que están separadas por una pulgada a lo largo del archivero. Esta operación es rápida mientras que el muestreo aleatorio simple será laborioso. Claro está que el método descrito anteriormente se aparta un tanto de la regla de "cada k-ésima".

2. Intuitivamente, el muestreo sistemático parece ser más preciso que el aleatorio simple. En efecto, estratifica la población en n estratos, que consisten de las primeras k unidades, las segundas k unidades, etc. Por lo tanto, podemos esperar que la muestra sistemática sea tan preciso como la muestra aleatoria estratificada co-

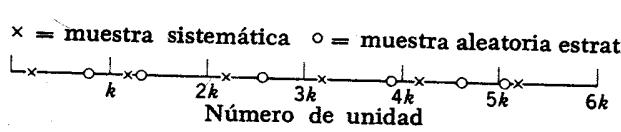


FIG. 8.1. Muestreo aleatorio estratificado y sistemático

rrespondiente con *una* unidad por estrato. La diferencia es que con la muestra sistemática, las unidades ocurren en la misma posición relativa del estrato, mientras que con el muestreo aleatorio estratificado, la posición dentro del estrato se determina separadamente por aleatorización dentro de cada estrato (véase la Fig. 8.1). La muestra sistemática se reparte más uniformemente sobre la población, y este hecho, algunas veces ha dado al muestreo sistemático una precisión considerablemente mayor que la del muestreo aleatorio estratificado.

Una modificación al muestreo sistemático consiste en elegir cada unidad en el centro o cerca del centro de cada estrato. O sea que en lugar de empezar la sucesión con un número aleatorio elegido entre 1 y k, tomamos el número inicial como $(k+1)/2$ si k es impar y como $k/2$ o $(k+2)/2$ si k es par (Madow, 1953). Este procedimiento lleva el propósito del muestreo sistemático a su conclusión lógica. Si y_i puede considerarse como una función continua de una variable continua i, hay razones para esperar que esta muestra localizada al centro será más precisa que una localizada aleatoriamente. Algunas investigaciones en poblaciones naturales apoyan esta opinión, aunque las muestras localizadas al centro tienden a comportarse erráticamente. Aquí nos ocuparemos en forma exclusiva de muestras en donde existe algún elemento aleatorio.

Como en general, N no es un múltiplo entero de k, las diferentes muestras sistemáticas de la misma población finita podrán variar su tamaño en una unidad. Así, con $N = 23$, $k = 5$, los números de unidades en las cinco muestras sistemáticas aparecen en la Tabla 8.1. Las tres primeras muestras tienen $n = 5$ y las últimas dos $n = 4$. Este hecho introduce una perturbación en la teoría del muestreo sistemático, que probablemente sea despreciable si n excede a 50, y se le ignorará para simplificar la presentación de la teoría. No es de esperarse que sea grande aun cuando n sea pequeña.

Otro método, sugerido por Lahiri en 1952 (véase Murthy, 1967, Pág. 139) proporciona un tamaño constante de muestra y una media de muestra insesgada. Considérense las N unidades dispuestas alrededor de un círculo, y sea k el entero más cercano a N/n. Se-

TABLA 8.1. LAS POSIBLES MUESTRAS SISTEMÁTICAS PARA $N = 23$, $k = 5$

Número de la muestra sistemática

I	II	III	IV	V
1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23		

lecciónese un número al azar entre 1 y N y tómese cada k-ésima unidad a partir de ahí y siguiendo el círculo, hasta alcanzar las n unidades deseadas. Supongamos que se desea $n = 5$ con $N = 23$. Entonces $k = 5$. Si el número aleatorio es 19, tomamos las unidades 19, 1, 6, 11, 16. Con este método es fácil verificar que cada unidad tiene la misma probabilidad de selección. Si se desean $n = 4$ unidades con $N = 23$, tomamos $k = 6$.

8.2 RELACION CON EL MUESTREO CONGLOMERADO

Existe otra manera de considerar el muestreo sistemático. Con $N = nk$ las k muestras sistemáticas posibles aparecen en las columnas de la Tabla 8.2. En él es evidente que la población se ha dividido en k grandes unidades de muestra, cada una de las cuales cuenta con n unidades originales. La operación de elegir una muestra sistemática aleatoriamente localizada, es sólo la de elegir *una* de estas grandes unidades de muestreo al azar. Por lo tanto, el muestreo sistemático viene a ser la elección de *una sola* unidad de muestreo compleja, que constituye la muestra total. Una muestra sistemática

TABLA 8.2. COMPOSICIÓN DE LAS k MUESTRAS SISTEMÁTICAS

Número de muestra

	1	2	···	i	···	k
	y_1	y_2		y_i		y_k
	y_{k+1}	y_{k+2}		y_{k+i}		y_{2k}
	···	···		···		···
	$y_{(n-1)k+1}$	$y_{(n-1)k+2}$		$y_{(n-1)k+i}$		y_{nk}
Medias	$\bar{y}_1$	$\bar{y}_2$		$\bar{y}_i$		$\bar{y}_k$

es una muestra aleatoria simple de una unidad conglomerada, tomada en una población de k unidades conglomeradas.

8.3 VARIANZA DE LA MEDIA ESTIMADA

Se han desarrollado numerosas fórmulas para la varianza de $\bar{y}_{sy}$, media de una muestra sistemática. Las tres que se dan a continuación se aplican a cualquier clase de muestreo conglomerado en donde éstos contienen n elementos y la muestra consiste en un conglomerado. En estas fórmulas suponemos $N = nk$.

Si $N = nk$, es fácil verificar que $\bar{y}_{sy}$ es una estimación insesgada de $\bar{Y}$ para una muestra sistemática aleatoriamente localizada.

En el análisis que sigue, el símbolo y_{ij} denota el j-ésimo miembro de la i-ésima muestra sistemática, por lo tanto, $j = 1, 2, \ldots n$, $i = 1, 2, \ldots k$. La media de la i-ésima muestra se denota con $\bar{y}_{i.}$.

Teorema 8.1. La varianza de la media de una muestra sistemática es

$$V(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy} \tag{8.1}$$

donde

$$S^2_{wsy} = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2$$

es la varianza entre las unidades que se encuentran dentro de la misma muestra sistemática. El denominador de esta varianza, $k(n-1)$, se construye por medio de las reglas usuales en el análisis de la varianza: cada una de las k muestras contribuye con $(n-1)$ grados de libertad a la suma de cuadrados del numerador.

Prueba. Por la identidad usual del análisis de la varianza

$$(N-1)S^2 = \sum_i \sum_j (y_{ij} - \bar{Y})^2$$

$$= n \sum_i (\bar{y}_{i.} - \bar{Y})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$$

Pero la varianza de $\bar{y}_{sy}$ es por definición

$$V(\bar{y}_{sy}) = \frac{1}{k} \sum_{i=1}^{k} (\bar{y}_{i.} - \bar{Y})^2$$

de aquí

$$(N-1)S^2 = nkV(\bar{y}_{sy}) + k(n-1)S^2_{wsy} \tag{8.2}$$

El resultado prosigue.

Corolario. La media de una muestra sistemática es más precisa que la media de una muestra aleatoria simple si y solamente si

$$S^2_{wsy} > S^2$$

Prueba. Si $\bar{y}$ es la media de una muestra aleatoria simple de tamaño n,

$$V(\bar{y}) = \frac{N-n}{N} \frac{S^2}{n}$$

De (8.1), $V(\bar{y}_{sy}) < V(\bar{y})$ si y solamente si

$$\frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy} < \frac{N-n}{N} \frac{S^2}{n} \tag{8.3}$$

esto es, si

$$k(n-1)S^2_{wsy} > \left(N - 1 - \frac{N-n}{n}\right) S^2 = k(n-1)S^2 \tag{8.4}$$

Este resultado importante que en general se aplica al muestreo conglomerado, enuncia que el muestreo sistemático es más preciso que el muestreo aleatorio simple si la varianza dentro de las muestras sistemáticas es *mayor* que la varianza de la población total. El muestreo sistemático es preciso cuando las unidades dentro de una misma muestra son heterogéneas y es impreciso cuando son homogéneas. El resultado es intuitivamente obvio. Si hay poca variación dentro de una muestra sistemática en relación con la de la población, las unidades sucesivas en la muestra repiten más o menos la misma información.

Otra forma para la varianza se da en el Teorema 8.2.

Teorema 8.2.

$$V(\bar{y}_{sy}) = \frac{S^2}{n}\left(\frac{N-1}{N}\right)[1 + (n-1)\rho_w] \tag{8.5}$$

donde ρ_w es el coeficiente de correlación entre pares de unidades que están en la misma muestra sistemática. Se define como

$$\rho_w = \frac{E(y_{ij} - \bar{Y})(y_{iu} - \bar{Y})}{E(y_{ij} - \bar{Y})^2} \tag{8.6}$$

donde el numerador es promediado sobre todos los $kn(n-1)/2$ pares distintos y el denominador sobre todos los N valores de y_{ij}. Como el denominador es $(N-1)S^2/N$, esto da

$$\rho_w = \frac{2}{(n-1)(N-1)S^2} \sum_{i=1}^{k} \sum_{j<u} (y_{ij} - \bar{Y})(y_{iu} - \bar{Y}) \tag{8.7}$$

Prueba

$$n^2kV(\bar{y}_{sy}) = n^2 \sum_{i=1}^{k} (\bar{y}_{i.} - \bar{Y})^2$$

$$= \sum_{i=1}^{k} [(y_{i1} - \bar{Y}) + (y_{i2} - \bar{Y}) + \cdots + (y_{in} - \bar{Y})]^2$$

Los términos al cuadrado constituyen la suma total de cuadrados de las desviaciones de $\bar{Y}$, es decir, $(N-1)S^2$. Esto da

$$n^2kV(\bar{y}_{sy}) = (N-1)S^2 + 2\sum_{i}\sum_{j<u}(y_{ij} - \bar{Y}(y_{iu} - \bar{Y}) \tag{8.8}$$

$$= (N-1)S^2 + (n-1)(N-1)S^2\rho_w \tag{8.9}$$

Por lo tanto,

$$V(\bar{y}_{sy}) = \frac{S^2}{n}\left(\frac{N-1}{N}\right)[1 + (n-1)\rho_w] \tag{8.10}$$

Este resultado muestra que la correlación positiva entre unidades en la misma muestra, infla la varianza de la media de la muestra. Aun una pequeña correlación positiva puede tener un gran efecto, debido al multiplicador $(n-1)$.

Los dos teoremas anteriores expresan $V(\bar{y}_{sy})$ en términos de S^2, por lo tanto, la relacionan con la varianza de una muestra aleatoria simple. Existe un teorema similar al 8.2 que expresa $V(\bar{y}_{sy})$ en términos de la varianza de una muestra aleatoria estratificada, en la cual, los estratos se componen de las primeras k unidades, las segundas k, etc. En nuestra notación, el suscrito j en y_{ij} denota el estrato. La media del estrato se escribe $\bar{y}_{.j}$.

Teorema 8.3.

$$V(\bar{y}_{sy}) = \frac{S^2_{wst}}{n}\left(\frac{N-n}{N}\right)[1 + (n-1)\rho_{wst}] \tag{8.11}$$

donde

$$S^2_{wst} = \frac{1}{n(k-1)} \sum_{j=1}^{n}\sum_{i=1}^{k} (y_{ij} - \bar{y}_{.j})^2 \tag{8.12}$$

Esta es la varianza entre unidades comprendidas en el mismo estrato. El divisor $n(k-1)$ se usa porque cada uno de los n estratos contribuyeron $(k-1)$ grados de libertad. Además,

$$\rho_{wst} = \frac{E(y_{ij} - \bar{y}_{.j})(y_{iu} - \bar{y}_{.u})}{E(y_{ij} - \bar{y}_{.j})^2} \tag{8.13}$$

Esta cantidad es la correlación entre las desviaciones de los pares de características de las medias de los estratos que están en la misma muestra sistemática.

$$\rho_{wst} = \frac{2}{n(n-1)(k-1)} \sum_{i=1}^{k}\sum_{j<u} \frac{(y_{ij} - \bar{y}_{.j})(y_{iu} - \bar{y}_{.u})}{S^2_{wst}} \tag{8.14}$$

La prueba es similar a la del Teorema 8.2.

Corolario. Una muestra sistemática tiene la misma precisión que la correspondiente muestra aleatoria estratificada con una unidad por estrato, si $\rho_{wst} = 0$. Esto es así porque para este tipo de muestra, $V(\bar{y}_{st})$ es (Teorema 5.3, Corolario 3)

$$V(\bar{y}_{st}) = \left(\frac{N-n}{N}\right)\frac{S^2_{wst}}{n} \tag{8.15}$$

Otras fórmulas para $V(\bar{y}_{sy})$, apropiadas para una población autocorrelacionada, fueron presentadas por W. G. y L. H. Madow (1944), quienes hicieron el primer estudio teórico sobre la precisión del muestreo sistemático.

Ejemplo. Los datos en la Tabla 8.3 son para una pequeña población artificial que exhibe una clara y constante tendencia ascendente. Tenemos $N = 40, k, = 10, n = 4$. Cada columna representa una muestra sistemática y las hileras son los estratos. El ejemplo ilustra la situación en la cual la correlación "dentro de estratos" es positiva. Por ejemplo, en la primera muestra cada uno de los cuatro números 0, 6, 18 y 26 están abajo de la media del estrato al que pertenecen. Esto es consistentemente cierto, con unas pocas excepciones, en las primeras cinco muestras sistemáticas. En las últimas cinco muestras, las desviaciones de la media del estrato, son en su mayoría positivas. En consecuencia, los términos de los productos cruzados en ρ_{wst} son predominantemente positivos. Del Teorema 8.3 esperamos que el muestreo sistemático sea menos preciso que el muestreo aleatorio estratificado con una unidad por estrato.

TABLA 8.3. DATOS DE 10 MUESTRAS SISTEMÁTICAS con $n = 4$, $N = kn = 40$

Estratos	Número de muestras sistemáticas 1	2	3	4	5	6	7	8	9	10	Medias de estratos
I	0	1	1	2	5	4	7	7	8	6	4.1
II	6	8	9	10	13	12	15	16	16	17	12.2
III	18	19	20	20	24	23	25	28	29	27	23.3
IV	26	30	31	31	33	32	35	37	38	38	33.1
Totales	50	58	61	63	75	71	82	88	91	88	72.7

TABLA 8.4. ANÁLISIS DE VARIANZA

	g.l.	s.c.	c.m
Entre hileras (estratos)	3	4828.3	
Dentro de estratos	36	485.5	$13.49 = S^2_{wst}$
totales	39	5313.8	$136.25 = S^2$

La varianza $V(\bar{y}_{sy})$ se encuentra directamente a partir de los totales de las muestras sistemáticas como

$$V(\bar{y}_{sy}) = V_{sy} = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{Y})^2 = \frac{1}{n^2 k}\sum_{i=1}^{k}(n\bar{y}_{i.}-n\bar{Y})^2$$

$$= \frac{1}{160}\left[(50)^2+(58)^2+\cdots+(88)^2-\frac{(727)^2}{10}\right] = 11.63$$

Para el muestreo aleatorio estratificado necesitamos un análisis de varianza de la población "entre hileras" y "dentro de hileras". Este se presenta en la Tabla 8.4. Por lo tanto, las varianzas de las medias estimadas a partir de muestras aleatorias simples y estratificadas será

$$V_{ran} = \left(\frac{N-n}{N}\right)\frac{S^2}{n} = \frac{9}{10}\cdot\frac{136.25}{4} = 30.66$$

$$V_{st} = \left(\frac{N-n}{N}\right)\frac{S^2_{wst}}{n} = \frac{9}{10}\cdot\frac{13.49}{4} = 3.04$$

Tanto el muestreo aleatorio estratificado como el sistemático son mucho más efectivos que el muestreo aleatorio simple, pero como se anticipó, el sistemático tiene menos precisión que el aleatorio estratificado.

La Tabla 8.5 muestra los mismos datos, con el orden de las observaciones invertido en los estratos segundo y cuarto. Esto hace que ρ_{wst} sea negativa, porque logra que la mayoría de los productos

TABLA 8.5. DATOS DE LA TABLA 8.3 CON EL ORDEN INVERTIDO EN LOS ESTRATOS II Y IV

Estratos	Números de la muestra sistemática 1	2	3	4	5	6	7	8	9	10	Medias de estratos
I	0	1	1	2	5	4	7	7	8	6	4.1
II	17	16	16	15	12	13	10	9	8	6	12.2
III	18	19	20	20	24	23	25	28	29	27	23.3
IV	38	38	37	35	32	33	31	31	30	26	33.1
Totales	73	74	74	72	73	73	73	75	75	65	72.7

cruzados entre las desviaciones desde las medias de estratos sean negativos para pares de observaciones que están en la misma muestra sistemática. Por ejemplo, en la primera muestra sistemática, las desviaciones a partir de las medias de estratos son ahora -4.1, $+4.8$, -5.3, $+4.9$. De los seis productos de pares de desviaciones, cuatro son negativos. A grandes rasgos, la misma situación se aplica en cada muestra sistemática.

Este cambio no afecta V_{ran} y V_{st}. Con muestreo sistemático se obtiene un incremento dramático en la precisión, como se ve al comparar los totales de muestra sistemática de la Tabla 8.5 con los de la Tabla 8.3. Ahora tenemos

$$V_{sy} = \frac{1}{160}\left[(73)^2+(74)^2+\cdots+(65)^2-\frac{(727)^2}{10}\right] = 0.46$$

En ocasiones es posible explotar este resultado al numerar las unidades para crear correlaciones negativas dentro de los estratos. Para ello se requiere un conocimiento aproximado de las tendencias dentro de la población. Sin embargo, como veremos más adelante, la situación de la Tabla 8.5 es una en la que resulta difícil obtener una buena estimación del error estándar de $\bar{y}_{sy}$ a partir de la muestra.

8.4 COMPARACION DEL MUESTREO SISTEMATICO CON EL MUESTREO ALEATORIO ESTRATIFICADO

El éxito del muestreo sistemático con relación al muestreo aleatorio simple o aleatorio estratificado, depende mucho de las propiedades de la población. En algunas poblaciones, el muestreo sistemático es extremadamente preciso y en otras resulta menos preciso que el muestreo aleatorio simple. Para algunas poblaciones y algunos valores de n, $V(\bar{y}_{sy})$ aun puede *incrementarse* al tomar una muestra grande, lo que constituye una desviación sorprendente del buen comportamiento. Por lo tanto, es difícil dar un consejo general respecto a las situaciones donde se aconseja el muestreo sistemático. Es necesario conocer algo sobre la estructura de la población para usarlo de manera efectiva.

Se han seguido dos lineamientos en la investigación de este problema. Uno de ellos es comparar los diferentes tipos de muestreo en poblaciones artificiales, en donde y_i es una función simple de i. El otro es hacer las comparaciones para poblaciones naturales. Algunos de los resultados principales los presentaremos en las secciones subsecuentes.

8.5 POBLACIONES EN ORDEN ALEATORIO

El muestreo sistemático se usa en algunos casos, por su conveniencia, en poblaciones en las cuales la numeración de las unidades es efectivamente aleatoria. Esto es así en el muestreo a partir de un archivo arreglado alfabéticamente por apellidos, si la característica que se mide no tiene relación con el apellido del individuo. No habrá entonces tendencia o estratificación en y_i conforme procedemos a lo largo del archivo ni tampoco correlación entre valores vecinos.

En esta situación, se esperaría que el muestreo sistemático fuera esencialmente equivalente al muestreo aleatorio simple y que tuviera la misma varianza. Para cualquier población finita simple, con valores dados de n y k, esto no es exactamente cierto, porque V_{sy}, que se basa en k grados de libertad, es más bien errática cuando k es pequeña y puede resultar mayor o menor que V_{ran}. Hay dos resultados que muestran, que en promedio, las dos varianzas son iguales.

Teorema 8.4. Considérense todas las $N!$ poblaciones finitas las cuales están formadas por las $N!$ permutaciones de cualquier conjunto de números $y_1, y_2, \ldots, y_N$. Entonces el promedio sobre estas poblaciones finitas,

$$E(V_{sy}) = V_{ran} \tag{8.16}$$

Nótese que V_{ran} es la misma para todas las permutaciones.

Este resultado, probado por W. G. y L. H. Madow (1944), muestra que, si el orden de las características en una población finita específica puede considerarse como sacado al azar de entre las $N!$ permutaciones, el muestreo sistemático es, en promedio, equivalente al muestreo aleatorio simple.

El segundo tratamiento es considerar la población finita como tomada al azar de una superpoblación infinita que tiene ciertas propiedades. El resultado que se prueba no se aplica a ninguna población finita sola (es decir, a ningún conjunto específico de valores $y_1, y_2, \ldots, y_N$) sino al promedio de todas las poblaciones finitas que pueden sacarse de la población infinita.

El símbolo $\mathscr{E}$ denota los promedios sobre todas las poblaciones finitas que se pueden sacar de esta superpoblación.

Teorema 8.5. Si las varianzas $y_i (i = 1, 2, \ldots, N)$ se sacan al azar de una superpoblación en la cual

$$\mathscr{E}y_i = \mu, \qquad \mathscr{E}(y_i-\mu)(y_j-\mu) = 0 \quad (i \neq j), \qquad \mathscr{E}(y_i-\mu)^2 = \sigma_i^2$$

entonces

$$\mathscr{E}V_{sy} = \mathscr{E}V_{ran}$$

Las condiciones cruciales son que todos los y_i tienen la misma media μ, esto es, no hay tendencia y no existe correlación lineal entre los valores de y_i y y_j en dos puntos diferentes. La varianza σ_i^2 puede cambiar de punto a punto en la serie.

Prueba. Para cualquier población finita específica

$$V_{ran} = \frac{N-n}{Nn} \frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N-1}$$

ahora

$$\sum_{i=1}^{N} (y_i - \bar{Y})^2 = \sum_{i=1}^{N} [(y_i - \mu) - (\bar{Y} - \mu)]^2$$

$$= \sum_{i=1}^{N} (y_i - \mu)^2 - N(\bar{Y} - \mu)^2$$

Como y_i y y_j, no tienen correlación $(i \neq j)$,

$$\mathscr{E}(\bar{Y} - \mu)^2 = \frac{1}{N^2} \sum_{i=1}^{N} \sigma_i^2 \tag{8.17}$$

Por lo tanto,

$$\mathscr{E}V_{ran} = \frac{N-n}{Nn(N-1)} \left(\sum_{i=1}^{N} \sigma_i^2 - N \frac{\sum \sigma_i^2}{N^2} \right) \tag{8.18}$$

El resultado es

$$\mathscr{E}V_{ran} = \frac{N-n}{N^2 n} \sum_{i=1}^{N} \sigma_i^2 \tag{8.19}$$

Volviendo a V_{sy}, dejemos que $\bar{y}_u$ denote la media de la u-ésima muestra sistemática. Para cualquier población finita específica,

$$V_{sy} = \frac{1}{k} \sum_{u=1}^{k} (\bar{y}_u - \bar{Y})^2 \tag{8.20}$$

$$= \frac{1}{k} \left[\sum_{u=1}^{k} (\bar{y}_u - \mu)^2 - k(\bar{Y} - \mu)^2 \right] \tag{8.21}$$

Por el teorema para la varianza de la media de una muestra no correlacionada de una población infinita,

$$\mathcal{E}V_{sy} = \frac{1}{k}\left(\frac{\sum_{i=1}^{N}\sigma_i^2}{n^2} - \frac{k\sum_{i=1}^{N}\sigma_i^2}{N^2}\right) \tag{8.22}$$

$$= \frac{N-n}{N^2 n}\sum_{i=1}^{N}\sigma_i^2 = \mathcal{E}V_{ran} \tag{8.23}$$

8.6 POBLACIONES CON TENDENCIA LINEAL

Si la población consiste únicamente de una tendencia lineal, com se ilustra en la Fig. 8.2, es muy fácil adivinar la naturaleza de lc resultados. De la Fig. 8.2 parece como si V_{sy} y V_{st} (con una unida por estrato) fueran ambas menores que V_{ran}. Además, V_{sy} será mayc que V_{st}, porque si la muestra sistemática es muy baja en un estratc es muy baja en todos los estratos, mientras que la muestra aleatori estratificada provee una oportunidad para que los errores dentro d los estratos se cancelen.

Para examinar los resultados matemáticamente podemos supone que $y_i = i$. Tenemos

$$\sum_{i=1}^{N} i = \frac{N(N+1)}{2}, \qquad \sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$$

La varianza de la población, S^2, está dada por

$$S^2 = \frac{1}{N-1}(\sum y_i^2 - N\bar{Y}^2)$$

$$= \frac{1}{N-1}\left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4}\right] = \frac{N(N+1)}{12} \tag{8.24}$$

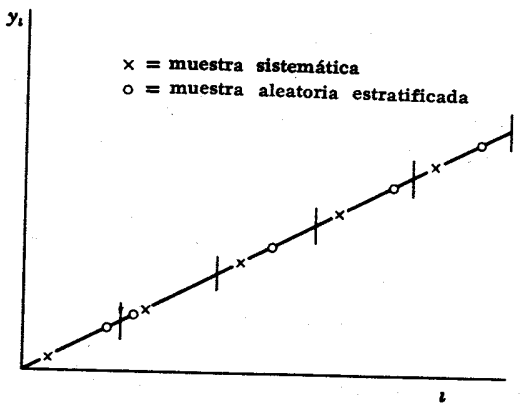


FIG. 8.2. Muestreo sistemático de una población con tendencia lineal

De aquí, la varianza de la media de una muestra aleatoria simple es

$$V_{ran} = \frac{N-n}{N}\cdot\frac{S^2}{n} = \frac{n(k-1)}{N}\cdot\frac{N(N+1)}{12n} = \frac{(k-1)(N+1)}{12} \tag{8.25}$$

Para encontrar la varianza dentro de estratos, S_w^2, sólo necesitamos reemplazar N por k en (8.24). Esto da

$$V_{st} = \frac{N-n}{N}\cdot\frac{S_w^2}{n} = \frac{n(k-1)}{nk}\cdot\frac{k(k+1)}{12n} = \frac{(k^2-1)}{12n} \tag{8.26}$$

Para muestreo sistemático, la media de la segunda muestra excede la de la primera en 1; la media de la tercera excede la de la segunda en 1, y así sucesivamente. Entonces, las medias $\bar{y}_u$ pueden reemplazarse por los números 1, 2, . . . , k. Así, por una aplicación posterior de (8.24),

$$\sum_{u=1}^{k}(\bar{y}_u - \bar{Y})^2 = \frac{k(k^2-1)}{12}$$

Esto da

$$V_{sy} = \frac{1}{k}\sum(\bar{y}_u - \bar{Y})^2 = \frac{k^2-1}{12} \tag{8.27}$$

A partir de las Fórmulas (8.25), (8.26) y (8.27) deducimos, como ya se anticipó,

$$V_{st} = \frac{k^2-1}{12n} \leq V_{sy} = \frac{k^2-1}{12} \leq V_{ran} = \frac{(k-1)(N+1)}{12} \tag{8.28}$$

La igualdad solamente se da cuando $n = 1$. Por lo tanto, para eliminar el efecto de una tendencia lineal, sospechada o insospechada, el muestreo sistemático resulta mucho más efectivo que el muestreo aleatorio simple, pero menos efectivo que el aleatorio estratificado.

8.7 METODOS PARA POBLACIONES QUE PRESENTAN TENDENCIAS LINEALES

El comportamiento del muestreo sistemático en presencia de una tendencia lineal podrá mejorarse de varias maneras. Una de ellas es utilizar una muestra centralmente localizada. Otra es la de cambiar la estimación de una media no ponderada a una media ponderada en donde todos los miembros internos de la muestra tendrán peso la unidad (antes de dividir por n) pero pesos diferentes asig-

nados al primero y último miembros. Si el número aleatorio sacado entre 1 y k es i, estos pesos son

$$1 \pm \frac{n(2i-k-1)}{2(n-1)k} \tag{8.29}$$

Se usa el signo + para el primer miembro, y el signo − para el último. Para todo i, los dos pesos obviamente suman 2. Se puede verificar que si la población consiste en una tendencia lineal y $N = nk$, la media de muestra ponderada da la media de población correcta. El efecto de estas *correcciones extremas* lo examinó Yates (1948) y se deben a él.

Bellhouse y Rao (1975) extendieron las correcciones de Yates al caso $N \neq nk$, cuando el muestreo sistemático se obtiene por el método circular de Lahiri (Sec. 8.1), que garantiza n constante. Como antes, los pesos que difieren de 1 se aplican a los números de muestra primero y último *en el orden secuencial original de la población*. Por ejemplo, si el número aleatorio inicial para extraer la muestra es 19 con $N = 23, n = 5$, las unidades 19, 1, 6, 11, 16 constituyen la muestra, el primero y último miembro son y_1 y y_{19}. Surgen dos casos.

Caso 1. i pequeño para el que $i + (n-1)k \leq N$ de modo que las n unidades se obtienen sin pasar sobre y_N. Los pesos para el primero (+) y último (−) miembros son

$$1 \pm \frac{n[2i+(n-1)k-(N+1)]}{2(n-1)k} \tag{8.30}$$

Caso 2. $i + (n-1)k > N$. Sea n_2 el número de unidades de la muestra que se obtienen al pasar sobre y_N. Así, con $i = 19, n_2 = 4$. Los pesos para el primero (+) y el último (−) miembros son

$$1 \pm \frac{n}{2(N-k)}\left[2i+(n-1)k-(N+1)-2n_2\frac{N}{n}\right] \tag{8.31}$$

En ambos casos, los miembros internos de la muestra reciben el peso 1 en el total de muestra. Con $N = 23$, $n = k = 5$, $i = 19$, $n_2 = 4$ los pesos primero y último son $1 \pm (-7/18)$. De modo que y_1 recibe un peso de 11/18 mientras que y_{19} recibe 25/18.

Dos métodos alternativos intentan cambiar el método de selección de la muestra de modo que la *media* de muestra no la afecte una tendencia lineal. Con $N = nk$ y n *par*, un método que sugiere Sethi (1965) divide la población en $n/2$ estratos de tamaño $2k$, al elegir dos unidades equidistantes del extremo de cada estrato. Si el número aleatorio inicial es i los $n/2$ pares de unidades son los numerados

$$[i+2jk,\ 2(j+1)k-i+1], \qquad j=0,1,2,\ldots \tfrac{1}{2}n-1 \tag{8.32}$$

Esta selección elimina el efecto de una tendencia lineal en cualquier estrato de $2k$ unidades, aunque la pendiente lineal varíe de un estrato a otro. Murthy (1967) denominó este método *muestreo sistemático equilibrado.*

El método *modificado* de Singh y otros (1968) elige pares de unidades equidistantes de los extremos de la población. Con n *par*, los $n/2$ pares equidistantes que empiezan con la unidad $i(i = 1, 2, \ldots k)$ son

$$[i+jk,\ (N-jk)-i+1], \qquad j=0,1,2,\ldots \tfrac{1}{2}n-1 \tag{8.33}$$

Cuando n *es impar* en estos métodos, j llega a $\frac{1}{2}(n-1)-1$ en (8.32) y (8.33). El método equilibrado (8.32) suma el miembro restante de la muestra cerca del extremo en $[i+(n-1)k]$; el método modificado lo hace cerca del medio centro en $[i+\frac{1}{2}(n-1)k]$. El efecto de una tendencia lineal no se elimina completamente en $\bar{y}$ para n impar.

Bellhouse y Rao (1975) han efectuado comparaciones de la actuación de estos dos métodos con las correcciones de Yates y con el muestreo sistemático ordinario en modelos de suprapoblaciones que representan tendencias lineales y parabólicas, y variación periódica y autocorrelacionada, también en algunas pequeñas poblaciones naturales las han efectuado estos autores y Singh (comunicación personal). En general, los tres métodos (Yates, equilibrado, modificado) se comportan similarmente y resultan superiores al muestreo sistemático ordinario en presencia de una tendencia lineal o parabólica.

La población de la Tabla 8.3 por ejemplo, es un caso donde estos métodos se comportarían muy bien. El muestreo sistemático ordinario, dio $V_{sy} = 11.63$. Para los otros métodos se tienen varianzas comparables ($n = 4$, $k = 10$): Yates, 1.29; Sethi (equilibrado) 0.46; Singh (modificado) 0.34. El método equilibrado es el que se tiene en la Tabla 8.5 al hacer la inversión en los estratos II y IV de la Tabla 8.3.

8.8 POBLACIONES CON VARIACION PERIODICA

Si la población consiste en una tendencia periódica, como sería, por ejemplo, una curva sinusoidal simple, la efectividad del muestreo sistemático depende del valor de k. Esto se puede ver gráficamente en la Fig. 8.3. En esta representación, la altura de la curva es la observación y_i. Los puntos de muestra A representan el caso menos favorable al muestreo sistemático, este caso se da siempre

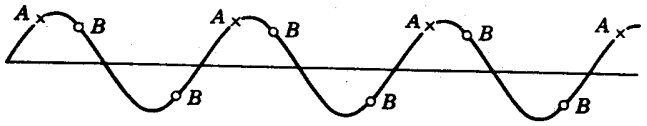


FIG. 8.3. Variación periódica

que k es igual al periodo de la curva sinusoidal o es un múltiplo entero del periodo. Todas las observaciones dentro del muestreo sistemático son exactamente iguales, de modo que la muestra no ofrece más precisión que una sola observación tomada al azar de la población.

El caso más favorable (muestra B) se da cuando k es un múltiplo impar del semiperiodo. Cada muestra sistemática tiene una media exactamente igual a la media verdadera de población dado que las desviaciones sucesivas arriba y abajo de la línea media se cancelan. La varianza muestral de la media es por lo tanto nula. Entre estos dos casos, la muestra tiene varios grados de efectividad que dependen de la relación entre k y la longitud de onda.

Las poblaciones que exhiben una curva sinusoidal exacta no se espera que se den en la práctica. Las poblaciones con una tendencia periódica definida sí son bastante comunes. Ejemplos de ellas son el flujo de tráfico que pasa por un punto dado de una carretera las 24 horas del día y las ventas de un almacén en un periodo de siete días. Para estimar un promedio sobre un cierto periodo de tiempo, no es aconsejable tomar la observación de la muestra todos los días a las 4 p.m., o cada martes; en lugar de esto, la estrategia apropiada sería la de espaciar la muestra a todo lo largo de la curva periódica, asegurándose que cada día de la semana recibe la misma representación, en el caso de las ventas.

Algunas poblaciones tienen una clase de efecto periódico menos obvio. Las nóminas semanales en un pequeño sector de una fábrica pueden siempre listar a los trabajadores en el mismo orden y contener de 19 a 23 nombres cada semana. Una muestra sistemática de uno de 20 nombres durante un periodo de semanas consta principalmente de los registros de un trabajador, o los registros de dos o tres trabajadores. De modo semejante, una muestra sistemática de nombre de un registro civil, podría contener demasiados nombres de cabezas de familia o demasiados niños. Si hay tiempo suficiente para estudiar la estructura periódica, usualmente se puede diseñar una muestra sistemática para capitalizar sobre esta estructura. De otro modo, es preferible una muestra aleatoria simple o estratificada, cuando se sospecha, pero no se conoce con seguridad, la existencia de un efecto periódico.

En algunas poblaciones naturales puede darse una variación casi periódica, difícil de anticipar. L. H. Madow (1946) encontró evidencia en este sentido en un lote de arbolillos en una población más bien pequeña ($N = 420$). Finney (1950) discutió un fenómeno semejante respecto al volumen de madera por franja de terreno en el bosque Dehra Dun, aunque reexaminando los datos Milne (1959) sugirió que la periodicidad aparente podía ser resultado del proceso de medición. El efecto de la casi periodicidad es que el muestreo sistemático tendrá un comportamiento mediocre en ciertos valores de n, y un buen comportamiento en otros. No se sabe si este efecto es frecuente. Matérn (1960) cita ejemplos en los que las fuerzas naturales, como serían las mareas, podrían producir variación periódica espacial, pero en su opinión no se han encontrado casos lo suficientemente claros en las encuestas forestales.

8.9 POBLACIONES AUTOCORRELACIONADAS

Con muchas poblaciones naturales hay razón para esperar que dos observaciones y_i, y_j serán más semejantes cuando i y j sean cercanos en la serie, que cuando ocupen posiciones distantes entre sí. Esto ocurrirá siempre que haya fuerzas naturales que induzcan un cambio lento al recorrer la serie. En un modelo matemático para este efecto podemos suponer que y_i y y_j están positivamente correlacionadas, siendo su correlación función solamente de su separación, $i - j$, que disminuye al crecer la distancia. Aunque este modelo está sobresimplificado, puede representar uno de los rasgos sobresalientes de muchas poblaciones naturales.

Para investigar si este modelo se aplica a una población, podemos calcular el conjunto de correlaciones ρ_u para pares de atributos que distan u unidades, y graficar esta correlación respecto a u. Esta curva, o la función que representa, se llama un *correlograma*. Aunque el modelo sea válido, el correlograma, no será una función suave para una población finita cualquiera, debido a las irregularidades que causa la finitud de la población. En una comparación de muestreo aleatorio estratificado con sistemático para este modelo, estas irregularidades dificultan la deducción de resultados para una población finita cualquiera. La comparación puede hacerse sobre el promedio de toda una serie de poblaciones finitas tomadas al azar de una suprapoblación infinita a la que se aplica el modelo. Esta técnica ya se usó en el Teorema 8.5 y en las Secs. 6.7 y 7.8.

Por lo tanto, suponemos que las observaciones y_i $(i = 1, 2, \ldots, N)$ se sacan de una suprapoblación en donde

$$\mathscr{E}(y_i) = \mu, \quad \mathscr{E}(y_i - \mu)^2 = \sigma^2, \quad \mathscr{E}(y_i - \mu)(y_{i+u} - \mu) = \rho_u \sigma^2 \tag{8.34}$$

con

$$\rho_u \geq \rho_v \geq 0, \text{ siempre que } u < v$$

El trazo de un conjunto de y_i a partir de esta suprapoblación crea una sola población finita de tamaño N.

La varianza promedio para el muestreo sistemático se denota con

$$\mathscr{E}V_{sy} = \mathscr{E}E(\bar{y}_{sy} - \bar{Y})^2$$

Para esta clase de poblaciones es fácil mostrar que el muestreo aleatorio estratificado supera al aleatorio simple, pero no se puede establecer un resultado general respecto al sistemático. Dentro de la clase hay suprapoblaciones en donde el muestreo sistemático supera al aleatorio estratificado, pero también hay suprapoblaciones en donde el muestreo sistemático es inferior al aleatorio simple para algunos valores de k.

Puede obtenerse un teorema general al suponer además, que el correlograma es cóncavo hacia arriba.

Teorema 8.6. Si además de las condiciones (8.34) tenemos

$$\delta_u^2 = \rho_{u+1} + \rho_{u-1} - 2\rho_u \geq 0 \quad [u = 2, 3, \ldots, (kn-2)] \tag{8.35}$$

entonces

$$\mathscr{E}V_{sy} \leq \mathscr{E}V_{st} \leq \mathscr{E}V_{ran} \tag{8.36}$$

para cualquier tamaño de muestra. Además, a menos que $\delta_u^2 = 0$, $u = 2, 3, \ldots, (kn - 2)$,

$$\mathscr{E}V_{sy} < \mathscr{E}V_{st} \tag{8.37}$$

Cochran (1946) ha demostrado este teorema.

Un bosquejo del argumento para $n = 2$ ilustra el papel desempeñado por la condición "cóncava hacia arriba". En la muestra sistemática, los miembros del par distan siempre k unidades. Por esto,

$$\mathscr{E}V(\bar{y}_{sy}) = \tfrac{1}{4}(\sigma^2 + \sigma^2 + 2\rho_k\sigma^2) = \tfrac{1}{2}\sigma^2(1 + \rho_k) \tag{8.38}$$

Con la muestra estratificada, hay k posiciones posibles para la unidad tomada de cada estrato, al hacer k^2 combinaciones de posiciones. Los números de combinaciones 1, 2, ..., $(2k - 1)$ unidades distantes son como sigue

Distancia	1	2 ... $(k-1)$	k	$(k+1)$ ... $(2k-1)$	Total
Número	1	2 ... $(k-1)$	k	$(k-1)$ ... 1	k^2

De aquí, el valor promedio de $V(\bar{y}_{st})$, tomado sobre las k^2 combinaciones, puede escribirse

$$\mathscr{E}V(\bar{y}_{st}) = \frac{\sigma^2}{2k^2}\left[\sum_{u=1}^{k-1} u(2 + \rho_u + \rho_{2k-u}) + k(1 + \rho_k)\right] \tag{8.39}$$

Del mismo modo, $\mathscr{E}V(\bar{y}_{sy})$ puede expresarse como

$$\mathscr{E}V(\bar{y}_{sy}) = \frac{\sigma^2}{2k^2}\left[\sum_{u=1}^{k-1} u(2 + 2\rho_k) + k(1 + \rho_k)\right] \tag{8.40}$$

De aquí,

$$\mathscr{E}V(\bar{y}_{st}) - \mathscr{E}V(\bar{y}_{sy}) = \frac{\sigma^2}{2k^2}\left[\sum_{u}^{k-1} u(\rho_u + \rho_{2k-u} - 2\rho_k)\right] \tag{8.41}$$

Pero si

$$\rho_{u+1} + \rho_{u-1} \geq 2\rho_u \qquad (u = 2, 3, \ldots)$$

es fácil demostrar que todos los términos dentro del paréntesis son positivos. Esto completa la prueba. Brevemente, la separación *promedio* es k, tanto para la muestra sistemática como para la estratificada, pero debido a la concavidad, la muestra estratificada pierde más en precisión, cuando la distancia es menor que k, de lo que gana, cuando la distancia excede a k.

Quenouille (1949) ha demostrado que las desigualdades en el Teorema 8.6 permanecen válidas cuando se relajan dos de las condiciones de tal manera que

$$\mathscr{E}(y_i) = \mu_i, \qquad \mathscr{E}(y_i - \mu_i)^2 = \sigma_i^2 \tag{8.42}$$

En este caso cada una de las tres varianzas promedio se aumenta en la misma cantidad.

En lo que se refiere a aplicaciones prácticas, varios autores han propuesto correlogramas cóncavos hacia arriba, como modelos para poblaciones naturales específicas. La función $\rho_u = \tanh(u^{-3/5})$ la sugirió Fisher y Mackenzie (1922) para la correlación entre lluvia semanal en dos estaciones meteorológicas las cuales distan una distancia u, la función $\rho_u = e^{-\lambda u}$ por Osborne (1942) y Matérn (1947) para reconocimientos de bosques y uso de la tierra, y la función $\rho_u = (l - u)/l$ por Wold (1938) para ciertos tipos de series de tiempo económicas.

8.9 POBLACIONES NATURALES

Se realizaron investigaciones en una variedad de poblaciones naturales. Los datos se describen en la Tabla 8.6. Los primeros tres estudios se hicieron a partir de mapas. En el primero, la población finita consiste de 288 altitudes a distancias sucesivas de 0.1 millas en una región ondulada. En los siguientes dos, los datos son fracciones de las longitudes de líneas trazadas sobre un mapa en el que se encuentra un cierto tipo de cobertura (por ejemplo, hierba). Estos ejemplos se pueden considerar como muy cercanos a la variación continua en el sentido matemático.

TABLA 8.6. POBLACIONES NATURALES USADAS EN ESTUDIOS DE MUESTREO SISTEMÁTICO

Referencia	N	Tipo de Datos
Yates (1948) tabla 13	288	Altitudes leídas a intervalos de 0.1 millas en un mapa de reconocimiento militar.
Osborne (1942)	*	Porcentaje de área en (a) tierra cultivada, (b) matorral, (c) hierba, (d) bosques, en líneas paralelas trazadas en un mapa.
Osborne (1942)	*	Porcentaje de área en abeto Douglas en líneas paralelas trazadas en un mapa.
Yates (1948)	192	Temperatura del suelo (12 plg bajo la hierba) durante 192 días consecutivos.
Yates (1948)	192	Temperatura del suelo (4 plg bajo suelo desnudo) durante 192 días.
Yates (1948)	192	Temperatura del aire durante 192 días.
Yates (1948)	96	Rendimientos de 96 hileras de papas.
Finney (1948)	160	Volumen de madera negociable por franja de 3 cadenas de ancho y de longitud variable (Bosque Monte Stuart).
Finney (1948)	288	Volumen de madera virgen por franja de 2.5 cadenas de ancho y 80 cadenas de largo (Bosque Monte Black).
Finney (1950)	292	Volumen de madera por franja de 2 cadenas de ancho y largo variable (Bosque Dehra Dun).
Johnson (1943)	400†	Número de plántulas por pie de era en 4 almácigos de árboles de madera dura.
Johnson (1943)	400†	Número de plántulas por pie de era en 3 almácigos de coníferas.
Johnson (1943)	400†	Número de plántulas por pie de era en 6 almácigos de coníferas trasplantadas.

* Teóricamente, N es infinito, si se pueden concebir líneas infinitamente delgadas.
† Aproximadamente. El número varió de era a era.

Los tres estudios siguientes se basaron en las temperaturas durante 192 días consecutivos: (a) 12 plg bajo suelo, (b) 4 plg bajo suelo, (c) en el aire. Este trío representa una gradación en dirección de mayor influencia de los cambios erráticos de día a día en el tiempo, en comparación con la lenta influencia de las estaciones.

Los estudios restantes se refieren a los rendimientos de plantas o árboles en secuencias que yacen a lo largo de una línea. En el estudio sobre las papas, que es típico del grupo, la población finita consiste en los rendimientos totales de 96 hileras en el campo. Como no se ha hecho aquí una revisión exhaustiva de la literatura, pueden encontrarse más datos.

En algunos de los estudios V_{sy} se compara con la varianza V_{st2} para una muestra aleatoria estratificada con estratos de tamaño $2k$ y dos unidades por estrato. Esta comparación es interesante porque de los datos de la muestra se puede obtener un estimador insesgado de V_{st2}. Esto no es posible para V_{st1} (con estratos de tamaño k y una unidad por estrato) o para V_{sy}. Otros autores reportan comparaciones entre V_{sy} con V_{st1} y V_{st2}. La mayoría de estas fuentes no presentan comparaciones con V_{ran} en una forma fácilmente utilizable, pero parece que en general V_{st2} dio ganancias en precisión sobre V_{ran}.

En las publicaciones de Yates y Finney se han dado comparaciones para una gama de valores de n y de k dentro de cada población finita. En esos casos, los datos en la Tabla 8.7 son las medias geométricas de las razones de varianzas para los valores individuales de k. Los otros autores hicieron cálculos solamente para un valor de k por población, pero pueden dar datos para diferentes características o para varias poblaciones del mismo tipo natural. Aquí, de nuevo, se tomaron las medias geométricas de las razones de varianzas.

Aunque los datos son limitados en extensión, los resultados son impresionantes. En los estudios que permiten comparación con V_{st1} el muestreo sistemático presenta una ganancia consistente en precisión la cual, aunque modesta, es buena. La mediana de las razones V_{st1}/V_{sy} es 1.4. Las ganancias en comparación con V_{st2} son sustanciales al ser la razón mediana 1.9.

La tendencia interna de los resultados concuerda con lo esperado, aunque no debería hacerse mucho caso de este hecho, en vista del pequeño número de estudios. Las ganancias son mayores para los tipos de datos en los cuales podríamos suponer que la variación sería cercana a la continua. El descenso en V_{st1}/V_{sy} de la temperatura del suelo a la del aire podría también anticiparse con baje en este mismo punto de vista. En las últimas tres características

TABLA 8.7. PRECISIÓN RELATIVA DE MUESTREO ALEATORIO ESTRATIFICADO Y SISTEMÁTICO

Datos	Intervalo de k	Precisión relativa de sistemático a estratificado	
		V_{st1}/V_{sy}	V_{st2}/V_{sy}
Altitudes	2–20	2.99	5.68
Porcentaje de área (4 tipos de cubierta)		—	4.42
Porcentaje de área (abeto Douglas)		—	1.83
Temperatura del suelo (12 plg)	2–24	2.42	4.23
Temperatura del suelo (4 plg)	4–24	1.45	2.07
Temperatura del aire	4–24	1.26	1.65
Papas	3–16	1.37	1.90
Volumen de madera (Monte Stuart)	2–32	1.07	1.35
Volumen de madera (Monte Black)	2–24	1.19	1.44
Volumen de madera (Dehra Dun)	2–32	1.39	1.89
Plántulas de madera dura	14	—	1.89
Plántulas de coníferas	14–24	—	2.22
Coníferas trasplantadas	12–22	—	0.93

(datos de los almácigos forestales), el único que no muestra ganancia alguna es el de las coníferas trasplantadas, las cuales son más viejas y más uniformes que las plántulas.

8.11 ESTIMACION DE LA VARIANZA A PARTIR DE UNA SOLA MUESTRA

A partir de los resultados de una muestra aleatoria simple con $n > 1$, podemos calcular una estimación insesgada de la varianza de la media de la muestra, siendo esta estimación insesgada *cualquiera que sea la forma de la población*. Dado que una muestra sistemática puede considerarse como una muestra simple aleatoria con $n = 1$, esta útil propiedad no prevalece en el caso de una muestra sistemática. Como ilustración considerese el ejemplo de la "curva sinusoidal". Sea

$$y_i = m + a \operatorname{sen} \frac{\pi i}{2}$$

donde $k = 4$ e $i = 1, 2, \ldots, 4n$. Las observaciones sucesivas en la población son

$$(m+a), m, (m-a), m, (m+a), m, (m-a), m, \ldots$$

Si $i = 1$ se escoge como el primer miembro, *todos* los miembros de la muestra sistemática tienen el valor $(m + a)$. Para cada una de las otras tres posibles elecciones de i, todos los miembros tienen los valores m, $(m - a)$, o m, respectivamente. Entonces, a partir de una *sola* muestra no tenemos manera de estimar el valor de a pero el valor real de la varianza de muestreo de la media de la muestra sistemática es $a^2/2$. La ilustración muestra que es imposible construir una varianza estimada que no tenga sesgo si hay variación periódica.

Estos resultados no significan que nada pueda hacerse. Excluyendo el caso de la variación periódica, podemos saber bastante sobre la estructura de la población como para poder desarrollar un modelo matemático que represente adecuadamente el tipo de variación presente. Podremos entonces desarrollar una fórmula para la varianza estimada que es aproximadamente sin sesgo para este modelo, aunque pudiera ser muy sesgada para otros modelos. La decisión de usar uno de estos modelos queda a juicio del muestreador.

A continuación se presentan algunos modelos simples junto con sus varianzas estimadas correspondientes. No se ofrecen pruebas.

Los modelos más simples se aplican a poblaciones en las cuales y_i se compone de una tendencia más una componente "aleatoria". Entonces

$$y_i = \mu_i + e_i$$

donde μ_i es alguna función de i. Para la componente aleatoria, suponemos que hay una superpoblación en la cual

$$\mathscr{E}(e_i) = 0, \qquad \mathscr{E}(e_i^2) = \sigma_i^2, \qquad \mathscr{E}(e_i e_j) = 0 \qquad (i \neq j)$$

Una fórmula S_{sy}^2 propuesta para la varianza estimada se denomina insesgada si

$$\mathscr{E}E(s_{sy}^2) = \mathscr{E}V_{sy}$$

esto es, si es sin sesgo para todas las poblaciones finitas que se pueden extraer de la superpoblación.

Población en Orden "Aleatorio"

$$\mu_i = \text{constante} \qquad (i = 1, 2, \ldots, N)$$

$$s_{sy1}^2 = \frac{N-n}{Nn} \frac{\sum (y_i - \bar{y}_{sy})^2}{n-1} \tag{8.43}$$

Este caso es aplicable cuando estamos seguros de que el orden es esencialmente aleatorio con respecto a las características que se están midiendo. La fórmula para la varianza es la misma que para una muestra aleatoria simple y no tiene sesgo si el modelo es correcto.

Solamente Efectos de la Estratificación

$$\mu_i \text{ constante} \quad (rk+1 \le i \le rk+k)$$

$$s^2_{sy2} = \frac{N-n}{Nn} \frac{\sum (y_i - y_{i+k})^2}{2(n-1)} \tag{8.44}$$

En este caso, la media es constante dentro de cada estrato de k unidades. La estimación s^2_{sy2}, que se basa en la diferencia sucesiva del cuadrado medio, no in sesgada. Contiene una contribución indeseable de la diferencia entre las μ en estratos vecinos y el primero y último estrato tienen muy poca influencia en la estimación de la componente aleatoria de la varianza. Con una muestra razonablemente grande, esta estimación sería en general demasiado alta suponiendo que el modelo es correcto.

Tendencia Lineal

$$\mu_i = \mu + \beta i$$

$$s^2_{sy3} = \frac{N-n}{N} \frac{n'}{n^2} \frac{\sum (y_i - 2y_{i+k} + y_{i+2k})^2}{6(n-2)} \qquad (1 \le i \le n-2) \tag{8.45}$$

La estimación se basa en términos cuadráticos sucesivos en la secuencia y_i. La suma de cuadrados contiene $(n-2)$ términos. Con una tendencia lineal hemos visto (Sec. 8.7) que la misma puede ser eliminada al usar correcciones terminales. El término n'/n^2 es la suma de cuadrados de las ponderaciones en $\bar{y}_{wsy}$. A menos que n sea pequeña, n'/n^2 se puede reemplazar por el factor usual $1/n$. Debido a que los estratos en los extremos reciben ponderaciones demasiado pequeñas, la estimación contiene sesgo a menos que σ_i^2 sea constante, pero será satisfactoria si n es grande y el modelo es correcto.

Si está presente una variación continua de tipo más complejo, la fórmula precedente puede arrojar resultados pobres. En la Tabla 8.8 la segunda y tercera fórmulas se aplican a seis eras de almácigo forestal (Johnson, 1943). La fórmula cuadrática es ligeramente mejor que la basada en diferencias sucesivas, pero ambas producen serias sobreestimaciones.

Fórmulas desarrolladas a partir de suposiciones simples acerca de la naturaleza del correlograma las han discutido Osborne (1942),

TABLA 8.8. VARIANZAS DE LAS MEDIAS DE MUESTRAS DE NÚMEROS DE PLÁNTULAS (DATOS DE JOHNSON)

	Era	Real V_{sy}	s^2_{sy2}	s^2_{sy3}
Arce plateado	1	0.91	2.8	2.5
	2	0.74	3.6	2.9
Olmo americano	1	4.8	28.4	12.6
	2	15.5	22.6	18.6
Pinabete blanco	1	5.5	17.2	11.2
	2	2.0	11.6	6.4
Pino blanco	1	8.2	21.0	21.9

Cochran (1946) y Matérn (1947). Si las observaciones sucesivas en muestras sistemáticas se indican por y_1', y_2' y así sucesivamente, Yates (1949) ha sugerido un estimador basado en diferencias d_u de las cuales la primera es

$$d_1 = (\tfrac{1}{2}y_1' + y_3' + y_5' + y_7' + \tfrac{1}{2}y_9') - (y_2' + y_4' + y_6' + y_8') \tag{8.46}$$

La diferencia siguiente, d_2, puede empezar con y_9', y así sucesivamente. Para la varianza estimada de $\bar{y}_{sy}$ tomamos

$$\hat{V}(\bar{y}_{sy}) = \frac{N-n}{Nn} \sum_{u=1}^{g} \frac{d_u^2}{7.5g} \tag{8.47}$$

El factor 7.5 es la suma de los cuadrados de los coeficientes en cualquier d_u, y g es el número de diferencias que da la muestra ($g \doteq n/9$). En las poblaciones naturales examinadas por Yates, una fórmula de este tipo resultó superior a s^2_{sy2} basada en diferencias sucesivas, pero todavía sobreestimada $V(\bar{y}_{sy})$.

Resumiendo, hay escasez de fórmulas para la varianza estimada, pero todas son de aplicabilidad limitada.

Con $N = nk$, supongamos que n es divisible por un entero m, (digamos 10). El método siguiente utiliza parcialmente muestreo sistemático y proporciona una estimación muestral insesgada de $V(\bar{y}_{sy})$ basada en $(m-1)$ grados de libertad. Extráigase una muestra aleatoria simple de tamaño m a partir de las unidades numeradas de 1 a mk. Para cada unidad de esta muestra, tómese también cada unidad en una posición (mk-ésima) posterior. De hecho, este método divide la población en mk conglomerados, cada uno de tamaño $N/mk = n/m$, y elige m conglomerados al azar, de modo que tenemos una muestra aleatoria simple de m conglomerados. Por ejemplo, supongamos que queremos una muestra del 20% de una población con $N = 2\,400$ de modo que $n = 480$ $k = 5$, tómese una muestra aleatoria simple de tamaño $m = 10$ entre las unidades numeradas de 1 a $mk = 50$, y tómese cada quincuagésima unidad de ahí en

adelante. Entonces tenemos 10 muestras conglomeradas, cada una de tamaño 2 400/50 = 48.

Gautschi (1957) ha examinado la exactitud de este método en las estructuras de población consideradas en este capítulo. Como podría anticiparse, la exactitud está entre la del muestreo aleatorio simple y la del muestreo sistemático con $m = 1$.

8.12 MUESTREO SISTEMATICO ESTRATIFICADO

Hemos visto que si las unidades se ordenan apropiadamente, el muestreo sistemático proporciona una clase de estratificación con fracciones de muestreo iguales. Si estratificamos según otro criterio, podríamos sacar una muestra sistemática separada dentro de cada estrato con puntos iniciales determinados independientemente. Esto es recomendable si se quieren estimaciones separadas para cada estrato o si se usan fracciones de muestreo desiguales. Este método es más preciso que el muestreo aleatorio estratificado si el muestreo sistemático dentro de los estratos es más preciso que el aleatorio simple dentro de los estratos. Si $\bar{y}_{syh}$ es la media de la muestra sistemática en el estrato h, la estimación $\bar{Y}$ de la media de población y su varianza son

$$\bar{y}_{stsy} = \sum W_h \bar{y}_{syh}, \qquad V(\bar{y}_{stsy}) = \sum W_h^2 V(\bar{y}_{syh})$$

Con tan sólo unos pocos estratos, el problema de encontrar una estimación de muestra de esta cantidad equivale al problema ya discutido de encontrar una estimación muestral satisfactoria de $V(\bar{y}_{syh})$ en cada estrato.

Cuando los estratos son más numerosos puede ser preferible una estimación basada en el método de estratos contraídos (Sec. 5A.12) De los resultados de esa sección se sigue que la estimación

$$v(\bar{y}_{stsy}) = \sum{}' W_h'^2 (\bar{y}_{syh} - \bar{y}_{syj})^2 \tag{8.48}$$

donde la suma se extiende sobre los pares de estratos, es en promedio una sobreestimación, aunque haya variación periódica dentro de los estratos.

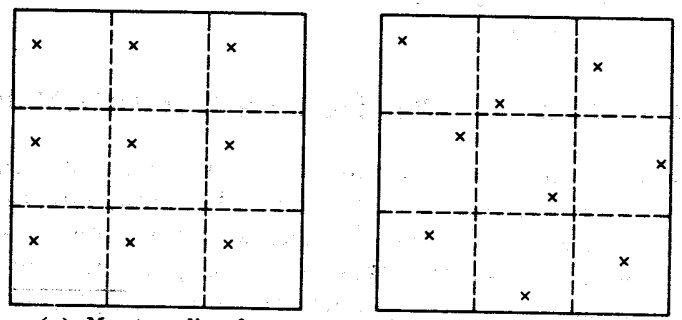
Se puede obtener una estimación insesgada de la varianza del error si se sacan dos muestras sistemáticas con diferente punto inicial aleatorio e intervalo $2k$ dentro de cada estrato, al proporcionar cada estrato un grado de libertad. Se perderá algo de precisión si el muestreo sistemático resulta efectivo. Si hay muchos estratos, podrá usarse una muestra sistemática en la mayoría de ellos al sacar dos muestras tales en una submuestra aleatoria de estratos con el propósito de estimar el error.

8.13 MUESTREO SISTEMATICO EN DOS DIMENSIONES

Al muestrear una área, la extensión más simple del muestreo sistemático unidimensional es el método de la "rejilla cuadrada" que aparece en la Fig. 8.4*a*. La muestra queda completamente determinada al elegir un par de números aleatorios para fijar las coordenadas de la unidad superior izquierda. El comportamiento de la rejilla cuadrada se ha estudiado en poblaciones teóricas y en poblaciones naturales. Metérn (1960) investigó el mejor tipo de muestra cuando la correlación entre cualesquiera dos puntos del área es una función cóncava hacia arriba y monótona decreciente de su separación d. Para correlogramas tales como $e^{-\lambda d}$ la rejilla da buenos resultados, y es superior al muestreo aleatorio simple o estratificado con una unidad por estrato, aunque Matérn da razones para esperar que lo mejor en esta situación es una red triangular en la que los puntos se encuentran en vértices de triángulos equiláteros.

En 14 experimentos de uniformidad agrícola Haynes (1948) encontró que la rejilla cuadrada tenía aproximadamente la misma precisión que el muestreo aleatorio simple en dos dimensiones. Milne (1959) examinó en 50 ensayos de uniformidad la rejilla cuadrada *central*, en la que el punto está en el centro del cuadrado. Dio mejores resultados que el muestreo aleatorio simple y tal vez un poco mejores que el muestreo aleatorio estratificado, aunque esta diferencia no fuese estadísticamente significativa. Estos resultados sugieren que cuando menos para datos de este tipo, los efectos de autocorrelación son débiles. Para estimar el área cubierta por bosque o agua sobre un mapa, Matérn encontró en dos ejemplos que la rejilla cuadrada es mejor que el método aleatorio.

La Fig. 8.4*b* exhibe una muestra sistemática alternativa llamada muestra *no alineada*. Las coordenadas de la unidad superior izquierda se seleccionan primero, mediante un par de números aleatorios. Otros dos números aleatorios determinarán las coordenadas *horizontales* de las dos unidades restantes en la primera *columna* de los estratos. Se requieren otros dos para fijar las coordenadas *verticales* de las unidades restantes en la primera *fila* de estratos. El intervalo constante k (igual a los lados de los cuadrados) fija entonces las posiciones de todos los puntos. Las investigaciones de Quenouille (1949) y Das (1950) para correlogramas simples de dos dimensiones indican que el diseño no alineado a menudo es superior a la rejilla cuadrada y al muestreo aleatorio estratificado.



(*a*) Muestra alineada (*b*) Muestra no alineada

FIG. 8.4. Dos tipos de muestra sistemática bidimensional

Nueva evidencia de la superioridad de una muestra no alineada se obtiene de la experiencia en un diseño experimental en el que se encontró que el cuadrado latino es un método preciso para disponer tratamientos en un campo rectangular. El cuadrado latino de 5×5 de la Fig. 8.5*a* puede verse como una división del campo en cinco muestras sistemáticas, una para cada letra. Existe evidencia de que este cuadrado particular llamado el cuadrado latino del "movimiento del rey" es un tanto más preciso que un cuadrado elegido al azar, de 5×5, probablemente porque no hay alineación en las diagonales, ni en las filas, ni en las columnas.

El principio del cuadrado latino lo utilizó Homeyer y Black (1946) al muestrear campos de avena rectangulares. Cada campo contenía 21 lotes. Las tres muestras sistemáticas posibles se denotan con *A*, *B* y *C*, respectivamente, en la Fig. 8.5*b*. Este arreglo con una de las letras elegidas al azar en cada campo dio un incremento en precisión aproximado al 25%, sobre el muestreo aleatorio estratificado con dos filas como estratos. El arreglo no satisface exactamente

A	*B*	*C*	*D*	*E*
D	*E*	*A*	*B*	*C*
B	*C*	*D*	*E*	*A*
E	*A*	*B*	*C*	*D*
C	*D*	*E*	*A*	*B*

(*a*) Cuadrado latino de "movimiento del rey"

A	*B*	*C*
B	*C*	*A*
C	*A*	*B*
A	*B*	*C*
B	*C*	*A*
C	*A*	*B*
A	*B*	*C*

(*b*) Diseño sistemático para un campo rectangular 3×7

FIG. 8.5. Dos diseños sistemáticos basados en el cuadrado latino

la propiedad del cuadrado latino, pues cada letra aparece tres veces en una columna y dos veces en las otras columnas, pero se acerca a esta propiedad tanto como es posible.

Yates (1960) que denomina estos arreglos *muestreo reticular*, discute su uso en muestreo de dos y tres dimensiones. En tres dimensiones, cada fila, columna y nivel vertical, puede representarse en la muestra al elegir p unidades de entre las p^3 de la población. Con p^2 unidades en la muestra, cada una de las p^2 combinaciones de niveles, filas y columnas, de filas y alturas verticales, y de columnas y alturas verticales, se podrá representar. Patterson (1954) ha investigado los arreglos que dan una estimación insesgada del error.

8.14 RESUMEN

Las muestras sistemáticas son de extracción y ejecución conveniente. En la mayoría de los estudios que hemos reportado en este capítulo, sobre poblaciones artificiales y poblaciones naturales pudieron compararse favorablemente en lo que a la precisión respecta con las muestras aleatorias estratificadas. Sus desventajas son que pueden dar una precisión mediocre cuando se presente una periodicidad insospechada, y que no se conoce aún un método confiable para la estimación de $V(\bar{y}_{sy})$ a partir de los datos de muestra.

A la luz de estos resultados podemos recomendar el muestreo sistemático en las situaciones siguientes.

1. Cuando el ordenamiento de la población es esencialmente aleatorio o contiene a lo más una estratificación débil. Aquí, el muestreo sistemático se usa por conveniencia y se esperan pocas ganancias de precisión. Se cuenta con estimaciones muestrales del error que son razonablemente insesgadas (Sec. 8.11).

2. Cuando se emplea una estratificación con numerosos estratos y se toma una muestra sistemática independiente en cada estrato. Los efectos de las periodicidades ocultas tienen tendencia a neutralizarse en esta situación y puede obtenerse una estimación del error que se sabe es una sobreestimación (Sec. 8.12). En forma alternativa, podemos usar la mitad de los estratos y tomar dos muestras sistemáticas, con inicios aleatorios independientes en cada estrato. Este método da una estimación insesgada del error.

3. Para unidades conglomeradas de submuestreo (Cap. 10). En este caso puede obtenerse una estimación casi insesgada o insesgada del error de muestreo en la mayoría de las situaciones prácticas. Este es un uso bastante común del muestreo sistemático.

Número de plántulas

Pies 1–20	21–40	41–60	61–80	81–100	101–120	121–140	141–160	161–180	181–200	Totales muestra sistemática
1	2	3	4	5	6	7	8	9	10	
8	20	26	34	31	24	18	16	36	10	223
6	19	26	21	23	19	13	12	8	35	182
6	25	10	27	41	28	7	8	29	7	188
23	11	41	25	18	18	9	10	33	9	197
25	31	30	32	15	29	11	12	14	12	211
16	26	55	43	21	24	20	20	13	7	245
28	29	34	33	8	33	16	17	18	6	222
21	19	56	45	22	37	9	12	20	14	255
22	17	39	23	11	32	14	7	13	12	190
18	28	41	27	3	26	15	17	24	15	214
26	16	27	37	4	36	20	21	29	18	234
28	9	20	14	5	20	21	26	18	4	165
11	22	25	14	11	43	15	16	16	4	177
16	26	39	24	9	27	14	18	20	9	202
7	17	24	18	25	20	13	11	6	8	149
22	39	25	17	16	21	9	19	15	8	191
44	21	18	14	13	18	25	27	4	9	193
26	14	44	38	22	19	17	29	8	10	227
31	40	55	36	18	24	7	31	8	5	255
26	30	39	29	9	30	30	29	10	3	235
Totales de los estratos 410	459	674	551	325	528	303	358	342	205	4155

4. Para muestrear poblaciones que tienen una variación de tipo continuo a condición de que no se requiera regularmente una estimación del error de muestreo. Si se hace una serie de encuestas de ese tipo, será suficiente una verificación ocasional de los errores de muestreo. Yates (1948) ha mostrado cómo puede hacerse esto, al tomar observaciones adicionales.

EJERCICIOS

8.1. Los datos en la Tabla (Pág. 286) son los números de plántulas por cada pie de era, en una era de 200 pies de largo.

Encuentre la varianza de la media de una muestra sistemática de cada vigésimo pie. Compare ésta con las varianzas para (a) una muestra aleatoria simple, (b) una muestra aleatoria estratificada con dos unidades por estrato, (c) una muestra aleatoria estratificada con una unidad por estrato. Todas las muestras tienen $n = 10$. [$\Sigma(y_i - \bar{Y})^2 = 23\,601$.]

8.2. Una población de 360 viviendas (numeradas de 1 a 360) en Baltimore se ordena alfabéticamente en un archivo de acuerdo con el apellido del jefe de familia. Viviendas en las cuales el jefe no es blanco ocurren en los números siguientes: 28, 31-33, 36-41, 44, 45, 47, 55, 56, 58, 68, 69, 82, 83, 85, 86, 89-94, 98, 99, 101, 107-110, 114, 154, 156, 178, 223, 224, 296, 298-300, 302-304, 306-323, 325-331, 333, 335-339, 341, 342. (Las viviendas de no blancos muestran algún "agrupamiento" debido a la asociación entre apellido y color.)

Compare la precisión de una muestra sistemática 1-en-8 con una muestra aleatoria simple del mismo tamaño para estimar la proporción de viviendas en las cuales el jefe de familia no es blanco.

8.3. Un vecindario tiene tres comunidades compactas, que consisten respectivamente, de personas de descendencia anglosajona, polaca e italiana. Hay un directorio actualizado. En él, las personas de una casa están enlistadas en el siguiente orden: marido, esposa, hijos (por edades), otros. Las casas son enlistadas en orden a lo largo de las calles. El número promedio de personas por casa es de cinco.

La elección es entre una muestra sistemática de cada quinta persona en el directorio y una muestra aleatoria simple del 20%. ¿Para cuáles de las siguientes variables espera usted que la muestra sistemática sea más precisa? (a) Proporción de personas de descendencia polaca. (b) proporción de varones. (c) proporción de hijos. Dé sus razones.

8.4. En un directorio de 13 casas en una calle, las personas están enlistadas como sigue: M = varón adulto, F = mujer adulta, m = hijo varón, f = hija.

Vivienda

1	2	3	4	5	6	7	8	9	10	11	12	13
M	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*
F	*F*	*F*	*F*	*F*	*F*	*F*	*F*	*F*	*F*	*F*	*F*	*F*
f	*f*	*m*		*m*	*f*	*f*	*m*	*m*	*m*	*f*	*f*	
m	*m*	*f*		*m*	*m*	*f*	*f*		*f*	*m*		
f	*f*			*f*		*m*						

Compare las varianzas dadas por una muestra sistemática de una de cada cinco personas y una muestra aleatoria simple del 20% para la estimación de (*a*) la proporción de varones, (*b*) la proporción de hijos, (*c*) la proporción de personas en hogares de profesionales (viviendas 1, 2, 3, 12 y 13 se describen como profesionales). ¿Respaldan estos resultados sus respuestas del Ejercicio 8.3? Para la muestra sistemática numere cada columna hacia abajo y luego vaya hacia la parte superior de la siguiente columna.

8.5. En el Ejercicio 8.1 podemos estimar $V(\bar{y}_{sy})$ (*a*) al considerar cada muestra sistemática como una muestra aleatoria simple, (*b*) al pretender que cada muestra sistemática 1-en-20 está compuesta de dos muestras sistemáticas 1-en-40 con inicios al azar separados. Para cada método compare el promedio de las varianzas estimadas con la varianza real de $\bar{y}_{sy}$.

8.6. En una población que consiste de una tendencia lineal (Sec. 8.6) demuestre que una muestra sistemática es menos precisa que una muestra aleatoria estratificada con estratos de tamaño $2k$ y dos unidades por estrato si $n > (4k + 2)/(k + 1)$.

8.7. Una población bidimensional con una tendencia lineal se puede representar por la relación

$$y_{ij} = i + j \qquad (i, j = 1, 2, ,\ldots, nk)$$

donde y_{ij} es el valor de la característica en la i-ésima hilera y la j-ésima columna. La población contiene $N^2 = n^2k^2$ unidades.

Una muestra sistemática en red cuadrada se selecciona al sacar al azar dos coordenadas iniciales independientes i_0, j_0, cada una entre 1 y k. La muestra, de tamaño n^2, contiene todas las unidades cuyas coordenadas son de la forma

$$i_0 + \gamma k, j_0 + \delta k$$

donde γ, δ son dos enteros cualesquiera entre 0 y $(n - 1)$, inclusive. Muestre que la media de esta muestra tiene la misma precisión que la media de una muestra aleatoria simple de tamaño n^2.

8.8. Si la comparación en el Ejercicio 8.7 se hiciera para una población tridimensional con tendencia lineal, ¿qué resultado esperaría usted?

8.9. En una población con $y_i = i^2 (i = 1, 2, \ldots, 16)$ comparar los valores de $E(\hat{\bar{Y}} - \bar{Y})^2$ dados por cada k-ésimo muestreo sistemático y por los métodos de Yates, Sethi y Singh, y otros, para $n = 4$, $k = 4$, $N = 16$.

CAPITULO 9

Muestreo por Conglomerados, de una Etapa: Conglomerados del Mismo Tamaño

9.1 MOTIVOS DEL MUESTREO POR CONGLOMERADOS

En los capítulos precedentes, se han hecho numerosas referencias a encuestas en las que la unidad consiste en un grupo o *conglomerado* de unidades más pequeñas que hemos llamado *elementos* o *subunidades*. Hay dos razones principales para la muy generalizada aplicación del muestreo por conglomerados. Aúnque la primera intención sea la de usar los elementos como unidades de muestreo, se ha encontrado que para muchas encuestas no se tiene una lista confiable de los elementos de la población y además sería demasiado costoso formular dicha lista. En muchos países no existen listas completas y actualizadas de la gente, las viviendas, o las granjas en grandes regiones geográficas. Sin embargo, a partir de los mapas de la región, se puede dividirla en unidades de área, como serían las manzanas de una ciudad y los terrenos con fronteras identificables rápidamente en las zonas rurales. En los Estados Unidos suelen elegirse estos conglomerados porque resuelven el problema de la construcción de una lista de unidades de muestreo.

Aunque se cuente con una lista de viviendas, puede ocurrir que por consideraciones económicas se recomiende una unidad conglomerada mayor. Para un tamaño de muestra dado, una unidad de muestreo pequeña suele ser más precisa que una unidad grande. Así, por ejemplo, una muestra aleatoria simple de 600 viviendas

cubre una ciudad más uniformemente que una muestra de 20 manzanas con un promedio de 30 viviendas por manzana. Pero se tendrán mayores costos de campo para localizar 600 viviendas y desplazarse entre ellas que para localizar 20 manzanas y visitar todas las casas de dichas manzanas. Al cotejar costo contra precisión, la unidad mayor puede resultar más conveniente.

Se puedé seleccionar racionalmente entre los dos tipos o tamaños de unidades mediante el conocido principio de elegir la unidad que da la varianza más pequeña para un costo dado, o el menor costo para una varianza prefijada. Como en muchas decisiones prácticas, puede haber factores imponderables: un tipo de unidad puede ofrecer alguna conveniencia especial o desventaja que es difícil incluir en el cálculo de los costos. Al muestrear un cultivo en crecimiento, algunas experiencias han sugerido que una unidad pequeña puede dar estimaciones sesgadas debido a la incertidumbre entre las fronteras exactas de la unidad. Homeyer y Black (1946) encontraron que las unidades de 2×2 pies, daban un rendimiento de avena superior en 8% a las unidades de 3×3 pies, posiblemente porque los muestreadores tendían a situar en los casos dudosos las plantas de la frontera dentro de la unidad. Sukhatme (1947) cita resultados semejantes para cultivos de trigo y arroz.

En este capítulo se trata el caso en el que cada unidad conglomerada contiene el mismo número de elementos o subunidades.

9.2 UNA REGLA SIMPLE

Cuando el problema es comparar unos cuantos tipos de unidad o tamaños específicos, es útil el resultado siguiente:

Teorema 9.1. Esto se aplica al muestreo aleatorio simple en el que la cpf es despreciable. La cantidad por estimar es el total de población. Para la unidad de tipo u-ésimo sean

M_u = tamaño relativo de la unidad

S_u^2 = varianza entre los totales de unidades

C_u = costo relativo para medir una unidad.

Entonces, el costo relativo para varianza especificada o la varianza relativa para costo especificado $\propto C_u S_u^2/M_u^2$.

Prueba. Supóngase que V es la varianza especificada del total estimado de población. Para el tipo u-ésimo de unidad la estimación es $N_u\bar{y}_u$ con varianza

$$V=\frac{N_u^2 S_u^2}{n_u}: \qquad n_u=\frac{N_u^2 S_u^2}{V} \tag{9.1}$$

El costo de tomar n_u unidades $C_u n_u = C_u N_u^2 S_u^2/V$. Como $N_u M_u =$ constante para los diferentes tipos de unidad, el costo es proporcional a $C_u S_u^2/M_u^2$. Por otro lado, si se especifica el costo C, $n_u = C/C_u$ y en (9.1), $V \propto C_u S_u^2/M_u^2$. Esto termina la demostración.

Corolario 1. Si definimos la *precisión neta relativa* de una unidad, como inversamente proporcional a la varianza obtenida para costo fijo, el Teorema 9.1 se puede enunciar como

$$\text{precisión neta relativa} \propto \frac{M_u^2}{C_u S_u^2} \tag{9.2}$$

Corolario 2. En el análisis de la varianza, las varianzas por unidades de diferentes tamaños suelen calcularse en lo que se llama una base común, que usualmente es la aplicable a la unidad más pequeña. Para poner las varianzas sobre una base común, la varianza S_u^2 entre los totales de unidades de tamaño M_u se divide por M_u. Sean

$S_u'^2 = \dfrac{S_u^2}{M_u}$ = varianza entre totales de unidades (sobre una base común)

$C_u' \propto \dfrac{C_u}{M_u}$ = costo relativo para tomar un determinado volumen de muestra

Entonces, el Teorema 9.1 y el Corolario 1 se pueden enunciar como sigue:

$$\text{costo relativo para la misma precisión} \propto \frac{C_u S_u^2}{M_u^2} \propto C_u' S_u'^2 \tag{9.3}$$

$$\text{precisión neta relativa} \propto \frac{1}{C_u' S_u'^2} \tag{9.4}$$

Si las diferencias en los costos para tomar la muestra se ignoran, o sea si C_u' es constante, la precisión neta relativa con la u-ésima unidad $\propto 1/S_u'^2$. Los factores *deff* de Kish para las unidades diferentes (Sec. 4.11) son, por tanto, proporcionales a las $S_u'^2 = s_u^2/M_u$.

Ejemplo. Los datos de Johnson (1941) para una era de plántulas de pino blanco proveen un ejemplo simple. La era tenía seis hileras de 434 pies de largo cada una. Hay muchas maneras de dividir la era en unidades de muestreo. Los datos para cuatro tipos de unidades se muestran en la Tabla 9.1. Dado que la era se contó completamente, los datos son correctos en lo que se refiere a valores de población.

TABLA 9.1. DATOS PARA CUATRO TIPOS DE UNIDADES DE MUESTREO

Datos preliminares	Tipo de unidad: 1 pie hilera	2 pies hilera	1 pie era	2 pies era
M_u = tamaño relativo de la unidad	1	2	6	12
N_u = número de unidades en la población	2604	1302	434	217
S_u^2 = varianza de la pobl. por unidad	2.537	6.746	23.094	68.558
Número de pies de hilera que se pueden contar en 15 min.	44	62	78	108

Las unidades fueron

un pie de una sola hilera
dos pies de una sola hilera
un pie del ancho de la era
dos pies del ancho de la era

Con las primeras dos unidades se supuso muestreo estratificado por hileras, de tal manera que S_u^2 representa las varianzas dentro de hileras. Para las dos últimas unidades se supuso muestreo aleatorio simple.

Ya que el costo principal es el de localizar y contar las unidades, los costos se estimaron mediante un estudio de tiempo (última hilera de la Tabla 9.1). Con las unidades mayores, se puede contar una muestra más grande en 15 min, por gastarse menos tiempo para ir de una unidad a la otra.

La cantidad a estimar es el número total de plántulas por era. En la notación del Teorema 9.1, la Tabla 9.1 da los valores de M_u y S_u^2. Los valores relativos de C_u, expresados como el tiempo requerido para contar una unidad, son como sigue:

	1 pie hilera	2 pies hilera	1 pie era	2 pies era
C_u (en tiempos de 15 minutos)	$\frac{1}{44}$	$\frac{2}{62}$	$\frac{6}{78}$	$\frac{12}{108}$

Según el Teorema 9.1, Corolario 1, las precisiones netas relativas se calculan en la Tabla 9.2.

La última línea de la Tabla 9.2 da las precisiones relativas cuando se toma la de la unidad más pequeña como 100. La unidad de 1 pie de era parece ser la mejor.

Las varianzas entre unidades, expresadas en una base común. También es conveniente observar. Los valores de $S_u'^2 = S_u^2/M_u$, aplicables a un solo pie de hilera son, respectivamente, 2.537, 3.373, 3.849, 5.713. Note que estas varianzas aumentan consistentemente al aumentar el tamaño de la unidad. Este resultado frecuentemente (aunque puede haber excepciones). Dado que la precisión relativa neta $\propto 1/C_u'S_u'^2$, el costo de tomar una muestra dada

TABLA 9.2. PRECISIONES NETAS RELATIVAS DE LAS CUATRO UNIDADES

	1 pie de hilera	2 pies de hilera	1 pie de era	2 pies de era
$\frac{M_u^2}{C_uS_u^2}$	$\frac{44}{2.537} = 17.34$	$\frac{(4)(62)}{(2)(6.746)} = 18.38$	$\frac{(36)(78)}{(6)(23.094)} = 20.27$	$\frac{(144)(108)}{(12)(68.558)} = 18.90$
	100	106	117	109

debe decrecer con las unidades más grandes si es que las mismas prueban ser económicas.

El Teorema 9.1 y sus corolarios siguen siendo válidos para muestreo estratificado con asignación proporcional si todos los estratos son del mismo tamaño y si S_u^2, $S_u'^2$ representan varianzas promedio dentro de los estratos. Esto es así, en las condiciones enunciadas porque la varianza del total de población estimado ignorando la cpf es $N^2S_u^2/n$ y, por lo tanto, toma la misma forma que en el muestreo aleatorio simple. El Teorema 9.1 no es válido para tipos de muestreo más complicados.

Los resultados anteriores se dan para ilustrar el procedimiento general. *Las comparaciones entre unidades deben hacerse siempre para la clase de muestreo que se usa en la práctica o bien, si esto aún no se ha decidido, para las clases de muestreo que se consideren.* Los cambios en el método de muestreo o de estimación alterarán las precisiones netas relativas de las diferentes unidades. Aun con un método fijo de muestreo y de estimación, las precisiones netas relativas varían con el tamaño de la muestra si el costo no es una función lineal del tamaño, o si el tamaño es tal que no se puedan ignorar las cpf.

Generalmente hay más de una característica a considerar. Un posible enfoque es fijar el costo total y encontrar las precisiones netas relativas para cada tipo de unidad y cada característica. A menos que un tipo sea uniformemente superior, se necesitará tomar una decisión de compromiso que dé la ponderación principal a las características más importantes.

En vista de los numerosos factores que influyen los resultados, el estudio de un tamaño óptimo de unidad en una encuesta muy extensa es una gran tarea. Jessen (1942) describe un buen ejemplo de muestreo agrícola. Un resumen de sus resultados es el que aparece en la Tabla 9.3. En él se comparan cuatro tamaños de unidades, que son un cuarto de sección, una media sección, una sección, y un bloque que consiste en dos secciones contiguas. Las secciones en una área de una milla cuadrada, contienen en promedio un poco

TABLA 9.3. ERRORES ESTÁNDAR ESTIMADOS (%) PARA CUATRO TAMAÑOS DE UNIDAD, CON MUESTREO ALEATORIO SIMPLE

Items	$S/4$	$S/2$	S	$2S$	Unidades mejores
Número de cerdos	5.0	4.9	5.3	6.2	$S/2$
Número de caballos	3.4	3.3	3.6	4.2	$S/2$
Número de borregos	17.4	15.7	14.9	14.3	$2S$
Número de pollos	3.0	3.0	3.3	3.8	$S/4$, $S/2$
Número de huevos producidos ayer	5.7	5.2	4.9	4.7	$2S$
Número de ganado	4.7	4.6	4.8	5.5	$S/2$
Número de vacas ordeñadas	3.7	3.6	3.8	4.4	$S/2$
Número de galones de leche	4.4	4.2	4.4	4.9	$S/2$
Ingresos por productos lecheros	5.5	5.2	5.4	6.0	$S/2$
Número de acres de las granjas	2.9	2.8	3.0	3.5	$S/2$
Número de acres de maíz	3.7	3.5	3.8	4.4	$S/2$
Número de acres de avena	4.6	4.8	5.6	7.0	$S/4$
Rendimiento de maíz	1.6	1.7	2.0	2.5	$S/4$
Rendimiento de avena	1.6	1.5	1.6	1.8	$S/2$
Gastos por alimentos comerciales	12.6	13.6	16.7	21.8	$S/4$
Gastos totales, operador	7.8	8.1	9.6	12.0	$S/4$
Ingresos totales, operador	6.2	6.5	7.7	9.8	$S/4$
Ingreso neto líquido, operador	6.8	6.9	7.8	9.5	$S/4$

menos que cuatro granjas. En esta comparación se especificaron el costo de campo total $1 000, la extensión del cuestionario (60 min para llenarlo) y el costo de viaje [5 cts. (USA) por milla], porque las precisiones netas relativas cambian al alterar cualquiera de estas variables. Los costos están al nivel de 1939.

Los datos de la Tabla son los errores estándar relativos (en porcentaje) de las medias estimadas por granja para 18 ítems. No hay mejor unidad para todos los artículos. La media sección y cuarto de sección, sin embargo, son superiores a las unidades más grandes para todas las características, excepto dos de ellas, al haber poco que elegir entre las secciones medias y cuartas. La media sección probablemente sería preferible porque el problema de identificar con exactitud las fronteras es más fácil.

9.3 COMPARACIONES DE PRECISION, HECHAS A PARTIR DE DATOS DE ENCUESTAS

En el ejemplo de la era de plántulas, las varianzas para los diferentes tipos de unidad se obtuvieron a partir de un recuento total

de la población. Sin embargo, a excepción de poblaciones pequeñas, raramente se puede efectuar un reconocimiento sólo con propósitos comparativos. La información sobre la unidad óptima usualmente se obtiene como un ingenioso subproducto de una encuesta cuyo propósito principal es obtener estimaciones.

Suponga que en una encuesta, cada unidad se puede dividir en M unidades menores. En lugar de registrar solamente los totales para cada unidad "grande" en la muestra, registramos los datos por separado para cada una de las M unidades menores. Entonces, puede hacerse una comparación entre las precisiones de las unidades grandes y las pequeñas. Primero se supondrá una muestra aleatoria simple de tamaño n.

El análisis de la varianza en la Tabla 9.4 se puede calcular a partir de la muestra.

La varianza estimada de una unidad grande (con base en una unidad menor) es S_b^2. Podría pensarse que una estimación apropiada de la varianza de una unidad menor sería el cuadrado medio entre todas las unidades menores en la muestra, esto es,

$$s^2 = \frac{(n-1)s_b^2 + n(M-1)s_w^2}{nM-1} \tag{9.5}$$

Esta estimación, aunque a menudo satisfactoria, está ligeramente sesgada porque la muestra no es una muestra aleatoria simple de unidades menores, dado que éstas se muestran en grupos contiguos de M unidades.

Una estimación insesgada se obtiene a partir de la muestra, al construir un análisis de la varianza como en la Tabla 9.5 para la población total, la cual contiene N unidades mayores y NM unidades menores.

TABLA 9.4. ANÁLISIS DE LA VARIANZA DE LOS DATOS DE LA MUESTRA (CON BASE EN UNA UNIDAD PEQUEÑA)

	g.l.	c.m.
Entre unidades grandes	$(n-1)$	s_b^2
Entre unidades pequeñas dentro de grandes	$n(M-1)$	s_w^2
Entre unidades pequeñas en la muestra	$nM-1$	$s^2 = \dfrac{(n-1)s_b^2 + n(M-1)s_w^2}{nM-1}$

TABLA 9.5. ANÁLISIS DE LA VARIANZA PARA LA POBLACIÓN TOTAL (CON BASE EN UNA UNIDAD PEQUEÑA)

	g.l.	c.m.
Entre unidades grandes	$N-1$	S_b^2
Entre unidades pequeñas dentro de unidades grandes	$N(M-1)$	S_w^2
Entre unidades pequeñas en la población	$NM-1$	$S^2=\dfrac{(N-1)S_b^2+N(M-1)S_w^2}{NM-1}$

De acuerdo con su definición, la varianza de la población entre unidades pequeñas es dada por la última línea de la Tabla, esto es,

$$S^2=\frac{(N-1)S_b^2+N(M-1)S_w^2}{NM-1} \tag{9.6}$$

Con el muestreo aleatorio simple, S_b^2 en la Tabla 9.4, es una estimación insesgada de S_b^2 (sacado de la Sec. 2.3). Puede mostrarse fácilmente que S_w^2 es una estimación insesgada de S_w^2. De esta manera, una estimación insesgada de la varianza S^2 entre todas las unidades pequeñas en la población es

$$\hat{S}^2=\frac{(N-1)s_b^2+N(M-1)s_w^2}{NM-1} \tag{9.7}$$

Claramente, esta expresión es casi la misma que la expresión más simple

$$\hat{S}^2\doteq\frac{s_b^2+(M-1)s_w^2}{M} \tag{9.8}$$

Si $n>50$, (9.5) para s^2 también se reduce a (9.8), de tal manera que s^2 es una aproximación satisfactoria a S^2 para $n>50$.

Las dos estimaciones s_b^2 (para la unidad grande) y S^2 (para la unidad pequeña) están en una base común y pueden sustituirse en el Teorema 9.1., Corolario 2.

Si la muestra es grande, las unidades menores pueden medirse dentro de una submuestra aleatoria de las unidades mayores (digamos 100 de cada 600). Alternativamente, podrían medirse dos unidades pequeñas tomadas al azar de cada unidad mayor. Simultáneamente pueden investigarse unidades pequeñas de más de un tamaño, siempre que se tomen datos que provean una estimación insesgada de S_w^2 para cada unidad pequeña.

Con muestreo estratificado, las varianzas para las unidades grandes y pequeñas se pueden estimar separadamente en cada estrato al usar estos métodos y luego sustituirse en la fórmula apropiada para la varianza de la estimación a partir de una muestra estratificada.

Ejemplo. Los datos son de una muestra de fincas tomada en Carolina del Norte en 1942 para estimar los empleos rurales (Finkner, Morgan, y Monroe, 1943). El método de extracción de la muestra fue localizar puntos al azar en el mapa y escoger como unidades de muestreo las tres fincas más cercanas a cada punto. El método no se recomienda ya que una finca grande tiene una probabilidad mayor de inclusión en una muestra que una finca pequeña y una finca aislada tiene una probabilidad mayor que otra en una área densamente explotada. Cualesquiera efectos de este sesgo serán ignorados.

A partir de los datos de las muestras correspondientes a fincas individuales, el grupo de tres fincas se puede comparar con la finca individual como unidad de muestreo. La característica escogida es el número de trabajadores asalariados. La muestra se estratificó, siendo el estrato un grupo de territorios similares en la densidad de su población de fincas y la razón tierra de pastoreo a tierra de labranza. Dado que la fracción de muestreo fue de 1.9%, puede ignorarse la cpf.

$$V(\hat{Y}_{st})=\sum_h\frac{N_h^2S_h^2}{n_h}$$

El procedimiento correcto es calcular $N_h^2S_h^2/n_h$ separadamente dentro de cada estrato para los dos tipos de unidad, usando un análisis de varianza y la expresión (9.8). Usaremos un procedimiento más simple como una aproximación.

Los estratos contenían en general entre 300 y 450 fincas y ya sea dos o tres unidades de tres fincas se tomaron dentro de cada estrato, para hacer el muestreo aproximadamente proporcional. Asumiendo proporcionalidad, esto es, $n_h/N_h=n/N$, podemos escribir

$$V(\hat{Y}_{st})=\frac{N}{n}\sum N_hS_h^2\doteq\frac{N^2}{n}\bar{S}_h^2$$

Y si suponemos además, que las S_h^2 no varían grandemente entre estratos de tal manera que se pueden reemplazar por su promedio $\bar{S}_h^2$.

Las estimaciones de $\bar{S}_h^2$ se obtienen del análisis de varianza en la Tabla 9.6, el cual se realizó con base en una sola finca.

TABLA 9.6. ANÁLISIS DE VARIANZA DE LA MUESTRA (NÚMERO DE TRABAJADORES PAGADOS CON BASE EN UNA SOLA FINCA)

	g.l.	c.m.
Entre unidades dentro de estratos	825	6.218
Entre fincas dentro de unidades	2768	2.918
Entre fincas dentro de estratos	3593	3.676

Para el grupo de tres fincas, el cuadrado medio $\bar{s}_{h3}^2 = 6.218$ sirve como estimación de $\bar{S}_h^2$ basado en una sola finca. Para las fincas individuales, usando (9.8) tenemos

$$\hat{\bar{S}}^2 = \frac{6.218 + 2(2.918)}{3} = 4.018$$

Por el Teorema 9.1, Corolario 2, las dos cifras 6.218 para el grupo de tres fincas y 4.018 para la finca individual, indican las varianzas relativas obtenidas para un tamaño total de muestra fijo. El grupo de fincas da cerca de dos tercios de la precisión de una sola finca. Consideraciones de costos probablemente harían que el resultado fuera más favorable para la unidad de tres fincas.

9.4 VARIANZA EN TERMINOS DE LA CORRELACION DENTRO DE CONGLOMERADOS

Las fórmulas para las varianzas a menudo se expresan en términos del coeficiente de correlación, ρ, entre elementos del mismo conglomerado. Este procedimiento ya se usó para el muestreo sistemático (Sec. 8.3).

Sea y_{ij} el valor observado para el j-ésimo elemento de la i-ésima unidad y sea y_i el total de la unidad. En muestreo conglomerado no hay porque distinguir entre las dos clases de promedios: la media por unidad $\bar{Y} = \Sigma y_i/N$ y la media por elemento $\bar{\bar{Y}} = \Sigma y_i/NM = \bar{Y}/M$. La varianza entre elementos es

$$S^2 = \frac{\sum_{i,j} (y_{ij} - \bar{\bar{Y}})^2}{NM-1}.$$

El coeficiente de correlación intraconglomerados ρ se definió (Sec. 8.3) como

$$\rho = \frac{E(y_{ij} - \bar{\bar{Y}})(y_{ik} - \bar{\bar{Y}})}{E(y_{ij} - \bar{\bar{Y}})^2} = \frac{2\sum_i \sum_{j<k} (y_{ij} - \bar{\bar{Y}})(y_{ik} - \bar{\bar{Y}})}{(M-1)(NM-1)S^2} \tag{9.9}$$

El número de los términos (productos cruzados) en el numerador E es $NM(M-1)/2$ y en el denominador E es $(NM-1)S^2/NM$.

Teorema 9.2. Una muestra aleatoria simple de n conglomerados, cada uno con M elementos, se extrae de entre los N conglomerados de la población. Entonces, la media de muestra por elemento $\bar{y}$ es una estimación insesgada de $\bar{\bar{Y}}$ con varianza

$$V(\bar{y}) = \frac{1-f}{n} \cdot \frac{NM-1}{M^2(N-1)} S^2[1+(M-1)\rho]$$

$$\doteq \frac{1-f}{nM} S^2[1+(M-1)\rho] \tag{9.10}$$

donde ρ es el coeficiente de correlación dentro del conglomerado.

Prueba. Supongamos que y_i denota el total para el i-ésimo conglomerado y $\bar{y} = \sum_{1}^{n} y_i/n$. Por los Teoremas 2.1 y 2.2, $\bar{y}$ es una estimación insesgada de $\bar{Y}$ con varianza,

$$V(\bar{y}) = \frac{(1-f)}{n} \frac{\sum (y_i - \bar{Y})^2}{N-1}$$

Pero $\bar{y} = M\bar{\bar{y}}$ y $\bar{Y} = M\bar{\bar{Y}}$. Por lo tanto, $\bar{\bar{y}}$ es una estimación insesgada de $\bar{\bar{Y}}$ con varianza

$$V(\bar{\bar{y}}) = \frac{1-f}{nM^2} \frac{\sum (y_i - \bar{Y})^2}{N-1} \tag{9.11}$$

Pero

$$(y_i - \bar{Y}) = (y_{i1} - \bar{\bar{Y}}) + (y_{i2} - \bar{\bar{Y}}) + \cdots + (y_{iM} - \bar{\bar{Y}})$$

Elevando al cuadrado y sumando sobre los N conglomerados,

$$\sum_{i}^{N} (y_i - \bar{Y})^2 = \sum_{i}^{N} \sum_{j}^{M} (y_{ij} - \bar{\bar{Y}})^2 + 2\sum_{i}^{N} \sum_{j<k}^{M} (y_{ij} - \bar{\bar{Y}})(y_{ik} - \bar{\bar{Y}})$$

$$= (NM-1)S^2 + (M-1)(NM-1)\rho S^2$$

$$= (NM-1)S^2[1+(M-1)\rho] \tag{9.12}$$

al usar la definición de ρ en (9.9). Al sustituir en (9.11) para $V(\bar{\bar{y}})$, esto da

$$V(\bar{\bar{y}}) = \frac{1-f}{n} \cdot \frac{NM-1}{M^2(N-1)} S^2[1+(M-1)\rho]$$

Esto completa la prueba.

Si se toma una muestra aleatoria simple de nM elementos, la fórmula para $V(\bar{\bar{y}})$ es la misma que en (9.10), excepto por el término dentro del paréntesis cuadrado. El factor

$$1+(M-1)\rho$$

muestra la magnitud del cambio de la varianza al usar un conglomerado en lugar de un elemento como unidad de muestreo. Este factor es, por tanto, el efecto del diseño de Kish para conglomerados de tamaño M (Sec. 4.11). Si $\rho > 0$, el conglomerado es menos preciso para un cierto volumen de muestra. Si $\rho < 0$, como a veces ocurre, el conglómerado es más preciso. El Teorema (9.2) es una simple extensión del Teorema 8.2. (Pág. 261.)

Puede darse una expresión alternativa para ρ. Sea S_b^2 la varianza entre totales de conglomerados, con base en una sola unidad. Entonces

$$\sum (y_i - \bar{Y})^2 = (N-1)MS_b^2$$

La Ec. 9.12 puede escribirse como

$$(N-1)MS_b^2 = (NM-1)S^2[1+(M-1)\rho]$$

de tal manera que

$$\rho = \frac{(N-1)MS_b^2 - (NM-1)S^2}{(NM-1)(M-1)S^2} \doteq \frac{S_b^2 - S^2}{(M-1)S^2} \tag{9.13}$$

cuando los términos en $1/N$ puedan ignorarse.

Es conveniente notar el valor del cuadrado medio intraconglomerado

$$S_w^2 = \sum_i^N \sum_j^M (y_{ij} - \bar{y}_i)^2/N(M-1) \tag{9.14}$$

Por el análisis simple de la varianza

$$(NM-1)S^2 = \sum_i^N (y_i - \bar{Y})^2/M + \sum_i^N \sum_j^M (y_{ij} - \bar{y}_i)^2$$

$$= \frac{(NM-1)}{M} S^2[1+(M-1)\rho] + N(M-1)S_w^2$$

por (9.12). Luego,

$$S_w^2 = \frac{NM-1}{NM} S^2(1-\rho) \doteq S^2(1-\rho) \tag{9.15}$$

Una buena discusión sobre los valores numéricos de ρ para diferentes características y diferentes tamaños de conglomerados la dan Hansen, Hurwitz y Madow (1953), quienes consideran ρ como una "medida de homogeneidad" del conglomerado.

9.5 FUNCIONES DE LA VARIANZA

En algunos tipos de encuestas, como son: el muestreo de suelos, el corte de cosechas y las encuestas de granjas que utilizan una unidad de muestreo de área, el tamaño de la unidad conglomerada puede admitir una variación casi continua. En la búsqueda de la mejor unidad, el problema no consiste en elegir entre dos o tres tamaños específicos que se han probado, sino encontrar el valor óptimo de

M vista como una variable continua. Este problema requiere un método para predecir la varianza S_b^2 entre unidades en la población, como función de M. Por el análisis de la varianza, S_b^2 puede verse que es (a) la varianza S^2 entre todos los elementos de la población y (b) la varianza S_w^2 entre elementos que están en la misma unidad. Nuestro método es el de predecir S_w^2 y S^2 y encontrar S_b^2 por el análisis de la varianza.

Los datos de muestra producen estimaciones de S^2 y S_w^2 para el tamaño de unidad realmente usado. Como S^2 es la varianza entre elementos, no la afectará el tamaño de la unidad. Sin embargo, S_w^2 sí se verá afectada. Podría esperarse que aumentara al aumentar el tamaño de la unidad. Si las unidades grandes a examinar poco difieren en tamaño de la unidad realmente usada, una primera aproximación será considerar S_w^2 constante, al usar la estimación que proporcionan los datos de muestra. Una investigación realizada por McVay (1947) sugiere que esta aproximación a menudo puede ser satisfactoria.

Como mejor aproximación, Jessen (1942); Mahalanobis (1944) y Hendricks (1944) intentaron desarrollar una ley general para predecir como cambia S_w^2 con respecto al tamaño de la unidad. En muchas encuestas agrícolas, S_w^2 pareció estar relacionada con M por la fórmula empírica

$$S_w^2 = AM^g \qquad (g>0) \tag{9.16}$$

donde A y g son constantes no dependientes de M. En esta fórmula, S_w^2 aumenta constantemente al aumentar M. Generalmente g es pequeña. Podría esperarse una curva de este tipo cuando hay fuerzas que ejercen influencias semejantes en elementos cercanos. El clima, el tipo de suelo, la topografía, y el anexo a los mercados tienden a asemejar las características de las explotaciones agrícolas.

En teoría, la fórmula es objetable, porque hace que S_w^2 aumente ilimitadamente al aumentar M. Si suponemos, como parece razonable, que no hay relación entre elementos muy distantes, una fórmula en la que S_w^2 se acerca a una cota superior con una M grande será más apropiada. Sin embargo, una fórmula que dé un buen ajuste sobre la gama de valores de M que se investiga será suficiente.

Si esta fórmula ajusta bien log S_w^2 aparecería como una recta al graficarla contra log M. Se necesitarán los valores de S_w^2, para cuando menos dos valores de M, con el fin de estirar las constantes log A y g. Son cuando menos necesarios tres valores de M para la apreciación de la linealidad del ajuste.

En el análisis de la varianza de la Tabla 9.5 encontramos que

$$S_b^2 = \frac{(NM-1)S^2 - N(M-1)S_w^2}{N-1}$$

$$= \frac{(NM-1)S^2 - N(M-1)AM^g}{N-1}$$

$$\doteq MS^2 - (M-1)AM^g \tag{9.17}$$

Hendricks (1944) ha señalado que la población completa podría considerarse como una sola gran unidad de muestreo que contuviera NM elementos. Si (9.16) es válida, entonces $S^2 = A(NM)^g$. La ventaja de este artificio es que los valores de A y g ahora pueden estimarse de los datos para una encuesta en la que sólo se ha usado un valor de M. Las dos ecuaciones que conducen a las estimaciones son

$$\log S_w^2 = \log A + g \log M \tag{9.18}$$

$$\log S^2 = \log A + g \log (NM) \tag{9.19}$$

La fórmula para S_b^2 viene a ser, a partir de (9.17)

$$S_b^2 \doteq AM^g[MN^g - (M-1)] \tag{9.20}$$

Este método no da una verificación de la corrección de (9.16), que podría ser bastante válida para pequeños valores de M, pero errática para un valor tan alto como NM.

La fórmula 9.16 se presenta como ejemplo de la metodología y no como regla general. Quien se enfrente a un problema similar, deberá construir y probar cualquier tipo de fórmula que parezca la más apropiada para su material. En algunos casos, $\log S_b^2$ podría ser una función lineal de M, como lo sugirió Fairfield Smith (1938) para datos de rendimiento.

9.6 UNA FUNCION DE COSTOS

En una encuesta muy vasta, la naturaleza de los costos de campo desempeña un papel importante para determinar la unidad óptima. Como ilustración del papel que juegan los factores de costo, vamos a describir una función de costo desarrollada por Jessen (1942) para encuestas agrícolas en donde las grandes unidades son conglomerados de granjas vecinas.

Se distinguen dos componentes de la función de costo de campo. La componente c_1Mn comprende los costos que directamente varían con el número total de elementos (granjas). Por lo tanto, c_1 contiene

el costo de la entrevista y el costo de desplazamiento entre las granjas del conglomerado.

La segunda componente, $c_2\sqrt{n}$, mide el costo de desplazamiento entre conglomerados. Se ha mostrado, por medio de pruebas sobre un mapa para una población fija, que este costo varía aproximadamente como la raíz cuadrada del número de conglomerados. Por lo tanto, el costo total del campo será

$$C = c_1Mn + c_2\sqrt{n} \tag{9.21}$$

al suponer un muestreo aleatorio simple e ignorar la cpf, la varianza de la media por elemento $\bar{y}$ es S_b^2/nM. De (9.17), esto es igual a

$$V(\bar{y}) = \frac{S^2 - (M-1)AM^{g-1}}{n} \tag{9.22}$$

Para determinar el tamaño óptimo de unidad, encontramos M e incidentalmente n, que minimiza V para un C fijo. La solución general es complicada, aunque su aplicación a un ejemplo numérico no presenta gran dificultad.

Mediante alguna manipulación podremos obtener la ecuación que da la M óptima. Primero se resuelve la ecuación de costo (9.21) como una cuadrática en $\sqrt{n}$. Esto da

$$\frac{2c_1M\sqrt{n}}{c_2} = \left(1 + \frac{4Cc_1M}{c_2^2}\right)^{1/2} - 1 \tag{9.23}$$

La ecuación a minimizar es

$$C + \lambda V = c_1Mn + c_2\sqrt{n} + \lambda V$$

Al diferenciar y notar que $\partial V/\partial n = -V/n$, obtenemos las ecuaciones:

n:
$$c_1M + \tfrac{1}{2}c_2n^{-1/2} = -\frac{\lambda\,\partial V}{\partial n} = \frac{\lambda V}{n} \tag{9.24}$$

M:
$$c_1n = -\frac{\lambda\,\partial V}{\partial M} \tag{9.25}$$

Al dividir (9.25) por (9.24) se elimina λ. Esto conduce a

$$\frac{n}{V}\frac{\partial V}{\partial M} = -\frac{c_1n}{c_1M + \tfrac{1}{2}c_2n^{-1/2}}$$

o

$$\frac{M}{V}\frac{\partial V}{\partial M} = -\frac{1}{1 + c_2/2c_1M\sqrt{n}} \tag{9.26}$$

Si sustituimos por $\sqrt{n}$ de (9.23), obtenemos, después de alguna simplificación,

$$\frac{M}{V}\frac{\partial V}{\partial M}=\left(1+\frac{4Cc_1M}{c_2^2}\right)^{-1/2}-1 \tag{9.27}$$

Al desarrollar completamente el lado izquierdo de esta ecuación y cambiar de signo en ambos lados, encontramos

$$\frac{AM^{g-1}[gM-(g-1)]}{S^2-(M-1)AM^{g-1}}=1-\left(1+\frac{4Cc_1M}{c_2^2}\right)^{-1/2} \tag{9.28}$$

Esta ecuación da la M óptima. El lado izquierdo no involucra ninguno de los factores de costo, y depende solamente de la forma de la función de la varianza. Puede verse que ambos lados son funciones crecientes de M dentro de la región de interés, para $g > 0$, $M \geq 1$. Supongamos que se encontró la solución para los valores específicos de C, c_1 y c_2, y queremos examinar el efecto de un aumento en c_1 sobre la solución. El lado izquierdo no depende de c_1, pero el lado derecho aumenta conforme c_1 aumenta. Consecuentemente, el valor óptimo de M disminuirán porque el término C_1M está a la derecha. Una disminución en c_2 produce un efecto similar.

Ahora, c_1 aumenta si la duración de la entrevista aumenta, mientras que c_2 disminuye si el viajar se hace más barato o si las fincas en una área dada aumentan en densidad. Estos hechos conducen a la conclusión de que el tamaño óptimo de la unidad se hace *más pequeño* cuando

aumenta la duración de las entrevistas
viajar se vuelve más barato
los elementos (fincas) aumentan en densidad
la cantidad total de dinero usada (C) aumenta

Esta conclusión es una consecuencia del tipo de función de costo y requerirá se le reexamine con otra función diferente. La misma ilustra el hecho de que la unidad óptima no es una característica fija de la población, sino que depende también del tipo de reconocimiento y de los niveles de precios y salarios.

Hansen, Hurwitz y Madow (1953) dan un excelente análisis de la construcción de funciones de costo para encuestas que involucran muestreo por conglomerados.

9.7 MUESTREO CONGLOMERADO PARA PROPORCIONES

Las mismas técnicas se aplican al muestreo conglomerado para proporciones. Supongamos que M elementos en cualquier conglomerado se pueden clasificar en dos clases y que $p_i = a_i/M$ es la proporción en la clase C en el i-ésimo conglomerado. Se toma una simple muestra aleatoria simple de n conglomerados y se usa el promedio p de las p_i observadas en la muestra como la estimación de la proporción P en la población.

Se recordará (Sec. 3.12) que no podemos usar la teoría binomial para encontrar $V(p)$ sino que debemos aplicar la fórmula para las variables continuas a las p_i. Esto da

$$V(p)=\frac{N-n}{Nn}\frac{\sum_{i=1}^{N}(p_i-P)^2}{N-1}\doteq\frac{N-n}{N^2n}\sum(p_i-P)^2 \tag{9.29}$$

Alternativamente, si tomamos una muestra aleatoria simple de nM elementos, la varianza de p se obtiene por teoría binomial (Teorema 3.2) como

$$V_{bin}(p)=\frac{(NM-nM)}{NM-1}\frac{PQ}{nM}\doteq\frac{N-n}{N}\frac{PQ}{nM} \tag{9.30}$$

si N es grande. Consecuentemente, el factor (efecto del diseño),

$$\frac{V(p)}{V_{bin}(p)}\doteq\frac{M\sum(p_i-P)^2}{NPQ}\qquad (N \text{ grande}) \tag{9.31}$$

muestra el cambio relativo en la varianza debido al uso de conglomerados. Los valores numéricos de este factor son útiles en la obtención de estimaciones preliminares del tamaño de la muestra con muestreo de conglomerado. El tamaño de muestra requerido se estima primero mediante la fórmula binomial y luego se multiplica por el factor para indicar el tamaño que será necesario con muestreo conglomerado. Para una ilustración, véase Cornfield (1951).

Si los tamaños M_i de los conglomerados son variables, la estimación $p = \Sigma a_i/\Sigma M_i$ es una estimación de razón. Su varianza la da, aproximadamente, la fórmula (Sec. 3.12)

$$V(p)\doteq\frac{N-n}{Nn\bar{M}^2}\frac{\sum_{i=1}^{N}M_i^2(p_i-P)^2}{N-1} \tag{9.32}$$

donde $\bar{M} = \Sigma M_i/N$ es el tamaño promedio del conglomerado.

Si este ejemplo se compara con una muestra aleatoria simple de $n\bar{M}$ elementos, encontramos, como una generalización de (9.31)

$$\frac{V(p)}{V_{bin}(p)}\doteq\frac{\sum M_i^2(p_i-P)^2}{N\bar{M}PQ} \tag{9.33}$$

Como con las variantes continuas, la relación entre el tamaño del conglomerado y la varianza entre conglomerados se puede in-

vestigar, ya sea expresando el factor en (9.31) y 9.33) como una función de $\bar{M}$, o buscando una relación entre la varianza dentro del conglomerado y $\bar{M}$. Si asignamos el valor 1 a cualquier unidad que cae en la clase C, y 0 a cualquier otra unidad, la ecuación fundamental del análisis de la varianza para M es

$$NMP(1-P)=M\sum(p_i-P)^2+M\sum p_i(1-p_i) \tag{9.34}$$

suma de cuadrados total = suma de cuadrados entre conglomerados + s.c. dentro de conglomerados

A partir de esta relación se puede calcular el cuadrado medio dentro de conglomerados y se le puede graficar como una función de M. McVay (1947) describe el uso de este análisis en la investigación del tamaño óptimo del conglomerado.

EJERCICIOS

9.1. Para los datos de la Tabla 9.1, comparar las precisiones netas relativas de los 4 tipos de unidad, cuando el propósito es el de estimar el número total de plántulas en la era con un error estándar de 200 plántulas (nótese que se involucra la cpf).

9.2. Para los datos de la Tabla 3.5, estimar la precisión relativa de la unidad familiar comparada con el individuo para estimar la razón de los sexos y la proporción de personas que hayan visto al doctor en los doce meses precedentes, suponiendo muestreo aleatorio simple.

9.3. Una población que consiste en 2 500 elementos, se divide en 10 estratos, conteniendo cada uno cincuenta grandes unidades compuestas de cinco elementos. El análisis de la variancia de la población para cierta característica es el siguiente, con base en los elementos.

	g.l.	c.m.
Entre estratos	9	30.6
Entre grandes unidades dentro de estratos	490	3.0
Entre elementos dentro de las grandes unidades	2 000	1.6

Ignorando la cpf. ¿considera usted que la precisión relativa de la unidad grande a la pequeña es mayor con muestreo aleatorio simple, que con muestreo aleatorio estratificado (asignación proporcional)?

9.4. Una población que contiene $L\bar{N}M$ elementos, se divide en L estratos, cada uno con $\bar{N}$ grandes unidades, y cada una de éstas con M unidades pequeñas. Las siguientes cantidades provienen del análisis de la variancia de la población, basándose en elementos.

S_1^2 = cuadrado medio entre estratos

S_2^2 = cuadrado medio entre grandes unidades dentro de estratos

S_3^2 = cuadrado medio entre elementos dentro de estratos

Si $\bar{N}$ es elevada y se ignora la cpf, mostrar que la precisión relativa de la unidad grande a la pequeña (el elemento) mejora gracias a la estratificación, si

$$\frac{(M-1)}{S_1^2}<\frac{M}{S_2^2}-\frac{1}{S_3^2}$$

9.5. En una encuesta rural, donde la unidad de muestreo es un conglomerado de M granjas, el costo de tomar una muestra de n unidades es

$$C=4tMn+60\sqrt{n}$$

donde t es el tiempo en horas gastado para obtener las respuestas de un granjero. Si se gastan \$2 000.00 en encuesta, los valores de n para $M = 1, 5, 10$; $t = \frac{1}{2}, 2$ dan lo siguiente.

	M		
	1	5	10
$t=\frac{1}{2}$ h	400	131	74
$t=2$ h	156	40	21

Verificar dos de estos valores para asegurarse de que ha entendido el uso de la fórmula.

La varianza de la media de muestra (ignorando la cpf) es

$$\frac{S^2}{Mn}[1+(M-1)\rho]$$

Si $\rho = 0.1$ para todo M entre 1 y 10, ¿cuál es el tamaño de unidad más preciso para (*a*) $t = \frac{1}{2}$ h, (*b*) $t = 2$ h? ¿Cómo explica la diferencia en los resultados?

9.6. Si hubiera \$5 000.00 para la encuesta, ¿esperaría que el tamaño óptimo de la unidad creciera o decreciera (relativamente a la que corresponde a \$2 000.00)? Dé sus razones. Puede, si lo desea, encontrar el tamaño óptimo para verificar su argumento.

CAPITULO 9A

Muestreo Conglomerado de una Etapa: Conglomerados de Tamaños Desiguales

9A.1 UNIDADES CONGLOMERADOS DE TAMAÑOS DESIGUALES

En la mayor parte de las aplicaciones, las unidades conglomerados (como son municipios, ciudades, manzanas de una ciudad) contienen números diferentes de elementos o subunidades (como son regiones geográficas, viviendas, personas). Este capítulo trata de algunos de los numerosos métodos de selección de muestra y estimación que se han producido para el caso de unidades conglomerados de tamaños desiguales. Sea M_i el número de elementos de la i-ésima unidad. Para estimar el *total* de población Y de los y_{ij} ya tenemos dos métodos conocidos.

Muestreo aleatorio simple de conglomerados: estimación insesgada

Como antes, sea

$$y_i = \sum_{j=1}^{M_i} y_{ij} = M_i \bar{y}_i$$

el *total* de elementos para la i-ésima unidad conglomerado. Dada una muestra aleatoria simple de n de las N unidades de la población, una estimación insesgada de Y es (por el Teorema 2.1, Corolario)

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i \tag{9A.1}$$

Por el Teorema 2.2, su varianza es

$$V(\hat{Y}) = \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N-1} \tag{9A.2}$$

donde $\bar{Y} = Y/N$ es la media de población por conglomerado.

Suele descubrirse que la estimación $\hat{Y}$ es de una precisión mediocre. Esto ocurre cuando las $\bar{y}_i$ (medias por *elemento*) varían poco entre las unidades y los M_i varían considerablemente. En este caso, las $y_i = M_i\bar{y}_i$ también varían considerablemente entre las unidades y la varianza (9A.2) es grande.

Muestreo aleatorio simple de conglomerados: estimación de razón a tamaño

Sea

$$M_0 = \sum_{i=1}^{N} M_i = \text{número total de elementos de la población.}$$

Si los M_i y, por tanto, M_0 se conocen, una alternativa es la estimación de razón en donde M_i se toma como la variable auxiliar x_i.

$$\hat{Y}_R = M_0 \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i} = M_0 \text{ (media muestral por elemento)}$$

En la notación de la estimación de razón, la razón de población $R = Y/X = Y/M_0 = \bar{Y}$, media de población por elemento. Por el Teorema 6.1, al suponer que el número de conglomerados de la muestra es grande,

$$V(\hat{Y}_R) \doteq \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^{N} (y_i - M_i\bar{Y})^2}{N-1} \tag{9A.3}$$

$$\doteq \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^{N} M_i^2(\bar{y}_i - \bar{Y})^2}{N-1} \tag{9A.4}$$

Como demuestra (9A.4), la varianza de $\hat{Y}_R$ depende de la variabilidad entre las medias por elemento y a menudo se encuentra que es mucho más pequeña que $V(\hat{Y})$.

Nótese que $\hat{Y}_R$ requiere un conocimiento del total M_0 de todos los M_i, en tanto que $\hat{Y}$ no lo exige. Lo contrario es verdadero cuando estimamos la media de población por elemento. En este caso las estimaciones correspondiente son

$$\hat{\bar{Y}} = \frac{\hat{Y}}{M_0} = \frac{N}{nM_0} \sum^{n} y_i, \quad \hat{\bar{Y}}_R = \frac{\hat{Y}_R}{M_0} = \frac{\sum^{n} y_i}{\sum^{n} M_i} = \text{media muestral por elemento}$$

Por lo tanto, $\hat{\bar{Y}}_R$ exige solamente el conocimiento de los M_i que caen en la muestra seleccionada.

9A.2 MUESTREO CON PROBABILIDAD PROPORCIONAL AL TAMAÑO

Si todos los M_i son conocidos, Hansen y Hurwitz (1943), han desarrollado otra técnica en donde se seleccionan las unidades con probabilidades proporcionales a sus tamaños M_i. Un método de elegir una sola muestra es el que ilustra la pequeña población siguiente de $N = 7$ unidades.

Unidad	Tamaño M_i	$\sum M_i$	Gama asignada
1	3	3	1–3
2	1	4	4
3	11	15	5–15
4	6	21	16–21
5	4	25	22–25
6	2	27	26–27
7	3	30	28–30

Se forman las sumas cumulativas de los M_i. Para elegir una unidad se saca un número aleatorio entre 1 y $M_0 = 30$. Supongamos que sea el 19. En la suma, el número 19 cae en la unidad 4 que cubre la gama de números de 16 a 21 inclusive. Con este método de extracción, la probabilidad de seleccionar cualquier unidad distinta, es proporcional a su tamaño.

Este método de selección de una unidad resulta conveniente cuando N es tan sólo moderado, o en el muestreo estratificado, cuando los N_h son pequeños o moderados, pero la acumulación de los M_i puede consumir tiempo considerable para N grande (como sería $N = 20\,000$). Para este caso, Lahiri (1951) ha dado un método alternativo que evita la acumulación. Sea $M_{máx}$ el mayor de los M_i. Extrái-

gase un número aleatorio entre 1 y N y supóngase que es i. Ahora, extráigase otro número aleatorio m entre 1 y $M_{máx}$. Si m es menor, o igual a M_i, se selecciona la i-ésima unidad. En caso contrario, se ensaya con otro par de números aleatorios. Naturalmente, este método involucra el menor número de rechazos cuando los M_i no difieren mucho entre sí.

Ahora consideremos $n > 1$; supongamos por el momento que el muestreo es *con restitución*. Para seleccionar una segunda unidad por el método cumulativo, extráigase un nuevo número aleatorio entre 1 y 30. Sin embargo, a diferencia del muestreo sin restitución, no prohibimos la selección de la unidad 4 una segunda vez. Con esta regla, las probabilidades de selección permanecen proporcionales a los tamaños en cada extracción. Una ventaja de la selección sin restitución es que las fórmulas para las varianzas verdadera y estimada de las estimaciones son simples.

Por otro lado (Sec. 9A.6), en muestreo sin restitución, el conservar las probabilidades de selección proporcionales a los tamaños elegidos es más difícil, y tarde o temprano se hace imposible al aumentar n. Esto puede verse en el caso extremo (aunque poco práctico) de $n = 7$ del ejemplo anterior. Si la selección fuera sin restitución, existe la certeza de que cada unidad sería elegida sin consideración a los tamaños originales M_i. Pero para muestreo estratificado, en donde los N_h son pequeños, se ha investigado profusamente (Sec. 9A.6 y siguientes) para desarrollar métodos prácticos de muestreo con probabilidades desiguales sin restitución.

9A.3 SELECCION CON PROBABILIDADES DESIGUALES Y CON RESTITUCION

Sea M_i/M_0 la probablidad con que se ha elegido la i-ésima unidad y con restitución, donde $M_0 = \sum M_i$. Ahora mostraremos que la siguiente es una estimación insesgada del total de población Y.

$$\hat{Y} = \frac{M_0}{n}(\bar{y}_1 + \bar{y}_2 + \cdots + \bar{y}_n)$$

$$= M_0 \text{ (media de las medias por elemento de las unidades)} \qquad (9A.5)$$

Para su comparación con métodos subsecuentes, denotaremos, esta estimación por medio de $\hat{Y}_{pps}$. Además

$$V(\hat{Y}_{pps}) = \frac{M_0}{n}\sum_{i=1}^{N} M_i(\bar{y}_i - \bar{\bar{Y}})^2 \qquad (9A.6)$$

de modo que la varianza de $\hat{Y}_{pps}$, como la de $\hat{Y}_R$, depende de la variabilidad de las *medias por elemento* de la unidad.

En algunas aplicaciones, los tamaños M_i se conocen sólo aproximadamente. En otros, el "tamaño" no es el número de elementos de la unidad sino una medida de su extensión que se piensa altamente correlacionada con el total de unidad y_i. Por ejemplo, el "tamaño" de un hospital podría medirse por el número total de camas, o por el número promedio de camas ocupadas en un cierto periodo de tiempo. De manera semejante, es posible dar varias medidas del "tamaño" de un restaurante, un banco o un distrito agrícola. Por consiguiente, consideraremos una medida de tamaño M_i' y una probabilidad correspondiente de selección $z_i = M_i'/M_0'$, donde $M_0' = \sum^{N} M_i'$. En lo que respecta los resultados teóricos los z_i pueden ser cualquier conjunto de números positivos de suma 1 sobre la población. Se mostrará que

$$\hat{Y}_{ppz} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{z_i} \qquad (9A.7)$$

es una estimación insesgada de Y con varianza

$$V(\hat{Y}_{ppz}) = \frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{y_i}{z_i} - Y\right)^2 \qquad (9A.8)$$

Las pruebas utilizan un método introducido en la Sec. 2.10. Sea t_i el número de veces que aparece la i-ésima unidad en una muestra específica de tamaño n donde t_i puede tomar cualquiera de los valores 0, 1, 2, . . . , n. Consideremos la distribución de frecuencia conjunta de los t_i para todas las N unidades de la población.

El método de extraer la muestra equivale al problema estándar de probabilidad en el cual se colocan n bolas en N cajas, y la probabilidad de que una bola determinada vaya a la caja i-ésima es igual a z_i en cada colocación. En consecuencia, la distribución conjunta de los t_i es la expresión multinomial

$$\frac{n!}{t_1!t_2!\ldots t_N!} z_1^{t_1} z_2^{t_2} \ldots z_N^{t_N}$$

Para la multinomial, son bien conocidas las siguientes propiedades de la distribución de los números t_i

$$E(t_i) = nz_i, \qquad V(t_i) = nz_i(1 - z_i), \qquad \text{Cov}\,(t_i t_j) = -nz_i z_j \qquad (9A.9)$$

Teorema 9A.1. Si una muestra de n unidades se extrae con las probabilidades z_i y con restitución, entonces

$$\hat{Y}_{ppz} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{z_i} \qquad (9A.10)$$

es una estimación insesgada de Y con varianza

$$V(\hat{Y}_{ppz})=\frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{y_i}{z_i}-Y\right)^2 \tag{9A.11}$$

Prueba. Podemos escribir

$$\hat{Y}_{ppz}=\frac{1}{n}\left(t_1\frac{y_1}{z_1}+t_2\frac{y_2}{z_2}+\cdots+t_N\frac{y_N}{z_N}\right)=\frac{1}{n}\sum_{i=1}^{N} t_i\frac{y_i}{z_i}$$

donde la suma se extiende a *todas* las unidades de la población. En muestreo repetido, las t_i son variables aleatorias, en tanto que los y_i y los z_i son un conjunto de números fijos. Por lo tanto, ya que $E(t_i) = nz_i$ por (9A.9),

$$E(\hat{Y}_{ppz})=\frac{1}{n}\sum_{i=1}^{N}(nz_i)\frac{y_i}{z_i}=\sum_{i=1}^{n} y_i = Y$$

de modo que $\hat{Y}_{ppz}$ es insesgada. También

$$V(\hat{Y}_{ppz})=\frac{1}{n^2}\left[\sum_{i=1}^{N}\left(\frac{y_i}{z_i}\right)^2 V(t_i)+2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{y_i}{z_i}\frac{y_j}{z_j}\operatorname{Cov}(t_i t_j)\right] \tag{9A.12}$$

$$=\frac{1}{n}\left[\sum_{i=1}^{N}\left(\frac{y_i}{z_i}\right)^2 z_i(1-z_i)-2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{y_i}{z_i}\frac{y_j}{z_j}z_i z_j\right] \tag{9A.13}$$

$$=\frac{1}{n}\left(\sum_{i=1}^{N}\frac{y_i^2}{z_i}-Y^2\right)=\frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{y_i}{z_i}-Y\right)^2 \tag{9A.11}$$

dado que $\sum z_i = 1$. Esto completa la prueba.

Al tomar $z_i = M_i/M_0$ en el Teorema 9A.1 se obtienen los correspondientes resultados para muestreo con probabilidad proporcional al tamaño.

Puede darse otra expresión para $V(\hat{Y}_{ppz})$. A partir de (9A.13),

$$nV(\hat{Y}_{ppz}) = \sum_{i=1}^{N}\frac{y_i^2(1-z_i)}{z_i}-2\sum_{i=1}^{N}\sum_{j>i}^{N} y_i y_j \tag{9A.14}$$

Como $(1 - z_i)$ es igual a la suma de todos los z_i restantes de la población, el coeficiente de y_i^2/z_i en (9A.14) contiene un término z_j para cada $j \neq i$. De manera semejante, el coeficiente de y_j^2/z_j contiene un término z_i. Por lo tanto,

$$V(\hat{Y}_{ppz})=\frac{1}{n}\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\frac{y_i^2 z_j}{z_i}+\frac{y_j^2 z_i}{z_j}-2y_i y_j\right)$$

$$=\frac{1}{n}\sum_{i=1}^{N}\sum_{j>i}^{N} z_i z_j\left(\frac{y_i}{z_i}-\frac{y_j}{z_j}\right)^2 \tag{9A.15}$$

Teorema 9A.2. Si se extrae una muestra de n unidades, con probabilidad proporcional a z_i con restitución, para cualquier $n > 1$, la siguiente es una estimación muestral insesgada de $V(\hat{Y}_{ppz})$,

$$v(\hat{Y}_{ppz})=\sum_{i=1}^{n}\left(\frac{y_i}{z_i}-\hat{Y}_{ppz}\right)^2\Big/ n(n-1) \tag{9A.16}$$

Prueba. Por la identidad algebraica usual

$$\sum_{i=1}^{n}\left(\frac{y_i}{z_i}-\hat{Y}_{ppz}\right)^2=\sum_{i=1}^{n}\left(\frac{y_i}{z_i}-Y\right)^2-n(\hat{Y}_{ppz}-Y)^2 \tag{9A.17}$$

de modo que por (9A.16)

$$E[n(n-1)v(\hat{Y}_{ppz})]=E\sum_{i=1}^{n}\left(\frac{y_i}{z_i}-Y\right)^2-nV(\hat{Y}_{ppz}) \tag{9A.18}$$

por la definición de $V(\hat{Y}_{ppz})$. Al introducir las variables t_i, obtenemos

$$\begin{aligned} n(n-1)E[v(\hat{Y}_{ppz})] &= E\sum_{i=1}^{N} t_i\left(\frac{y_i}{z_i}-Y\right)^2-nV(\hat{Y}_{ppz}) \\ &= n\sum_{i=1}^{n} z_i\left(\frac{y_i}{z_i}-Y\right)^2-nV(\hat{Y}_{ppz}) \\ &= n(n-1)V(\hat{Y}_{ppz}) \end{aligned} \tag{9A.19}$$

al usar (9A.11) en el Teorema 9A.1. Esto termina la prueba.

Ya que $\hat{Y}_{ppz}$ es la media de los n valores y_i/z_i, la Fórmula (9A.16) toma una forma muy simple.

Cuando la selección es estrictamente proporcional al tamaño (esto es $z_i = M_i/M_0$) los Teoremas 9A.1 y 9A.2, se expresan como sigue.

Teorema 9A.3. Si se extrae una muestra de n unidades con probabilidades al tamaño $z_i = M_i/M_0$ y con restitución

$$\hat{Y}_{pps}=\frac{M_0}{n}\sum_{i=1}^{n}\left(\frac{y_i}{M_i}\right)=\frac{M_0}{n}\sum_{i=1}^{n}(\bar{y}_i)=M_0\bar{y} \tag{9A.20}$$

donde $\bar{y}$ es la media no ponderada de las medias de unidades, es una estimación insesgada de Y con varianza

$$V(\hat{Y}_{pps})=\frac{M_0}{n}\sum_{i=1}^{N} M_i(\bar{y}_i-\bar{Y})^2 \tag{9A.21}$$

Estos resultados provienen del Teorema 9A.1, dado que $\bar{y}_i = y_i/M_i$ y $\bar{Y} = Y/M_0$.

Teorema 9A.4. En las condiciones del Teorema 9A.3, una estimación muestral insesgada de $V(\hat{Y}_{pps})$ es

$$v(\hat{Y}_{pps}) = M_0^2 \sum_{i=1}^{n} (\bar{y}_i - \bar{\bar{y}})^2/n(n-1) \qquad (9A.22)$$

El resultado se obtiene por sustitución de $z_i = M_i/M_0$ en (9A.16), dado que $\bar{y}_i = y_i/M_i$ y otra vez $\hat{Y}_{ppz} = M_0\bar{\bar{y}}$.

9A.4 LA MEDIDA OPTIMA DEL TAMAÑO

En los casos donde la medida del tamaño M_i' es una estimación de lo grande de la unidad, la siguiente es una pregunta de interés para nosotros: ¿cuál es la medida del tamaño que minimiza la varianza de $\hat{Y}_{ppz}$?

Ahora bien, $$V(\hat{Y}_{ppz}) = \frac{1}{n}\sum^{N} z_i\left(\frac{y_i}{z_i} - Y\right)^2 = \frac{1}{n}\left(\sum^{N}\frac{y_i^2}{z_i} - Y^2\right)$$

Esta expresión se anula si $z_i \propto y_i$: es decir, si $z_i = y_i/Y$. Si los y_i son todos positivos, el conjunto de z_i es un conjunto aceptable de probabilidades. De modo que las mejores medidas del tamaño son números proporcionales a los totales de atributos y_i para las unidades.

Este resultado no es de una aplicación práctica directa; si los y_i se conocen de antemano para toda la población, sería innecesaria la muestra. El resultado sugiere que si los y_i son relativamente estables en el transcurso del tiempo, los valores previos más recientemente disponibles de y_i podrían ser las mejores medidas del tamaño para este atributo. En la práctica, claro está, debe usarse una *sola* medida del tamaño para todos los atributos al seleccionar la muestra. Si se puede elegir entre diferentes medidas del tamaño, podemos esperar que la mejor sea la más cercana a la proporcionalidad respecto a los totales de unidades de los atributos principales.

9A.5 EXACTITUDES RELATIVAS DE LAS TRES TECNICAS

En esta sección comparamos las exactitudes de las tres técnicas precedentes para estimar el total de población, con unidades que son conglomerados de tamaños desiguales (suponiendo que se conocen los M_i si la técnica los requiere).

1. Selección: probabilidades iguales. Estimación: $\hat{Y}_u$.
2. Selección: probabilidades iguales. Estimación: $\hat{Y}_R$.
3. Selección: probabilidad $\propto$ tamaño. Estimación; $\hat{Y}_{pps}$.

No existe una regla simple para decidir cuál es la más exacta. El resultado depende de la relación entre $\bar{y}_i$ y M_i y de la varianza de $\bar{y}_i$ como función de M_i. La situación favorable a las estimaciones de razón y *pps* es aquella en que $\bar{y}_i$ no esté relacionada con M_i. La situación favorable a $\hat{Y}_u$ es aquella en que el *total* de unidad y_i no esté relacionado con M_i.

Para guiarse en la elección, pueden expresar las varianzas de las tres estimaciones en una forma comparable. Suponemos que $(N-1) \doteq N$ y escribimos $E(y_i - \bar{Y})^2 = \sum^{N}(y_i - \bar{Y})^2/N$. También suponemos que el sesgo $\hat{Y}_R$ es despreciable.

Para $\hat{Y}_u$ tenemos, a partir de (9A.2),

$$nV(\hat{Y}_u) = N^2(1-f)E(y_i - \bar{Y})^2 = (1-f)E(Ny_i - Y)^2 \qquad (9A.23)$$

Para $\hat{Y}_R$, y a partir de (9A.4),

$$nV(\hat{Y}_R) = N^2(1-f)EM_i^2(\bar{y}_i - \bar{Y})^2 = (1-f)E\left(\frac{M_i}{\bar{M}}\right)^2(M_0\bar{y}_i - Y)^2 \qquad (9A.24)$$

donde $\bar{M} = \sum M_i/N = M_0/N$.

De (9A.21), para $\hat{Y}_{pps}$,

$$nV(\hat{Y}_{pps}) = NM_0EM_i(\bar{y}_i - \bar{Y})^2 = M_0^2E\left(\frac{M_i}{\bar{M}}\right)(\bar{y}_i - \bar{Y})^2$$

$$= E\left(\frac{M_i}{\bar{M}}\right)(M_0\bar{y}_i - Y)^2 \qquad (9A.25)$$

De (9A.23), (9A.24), (9A.25) concluimos que $V(\hat{Y}_u)$ depende de la exactitud de las cantidades $Ny_i = NM_i\bar{y}_i$ como estimaciones de Y, en tanto que $V(\hat{Y}_R)$ y $V(\hat{Y}_{pps})$ dependen de la exactitud de las cantidades $M_0\bar{y}_i = M_0y_i/M_i$ como estimaciones de Y. Si $\bar{y}_i$ no está relacionada con M_i, esperamos que $M_0\bar{y}_i$ sea más exacta que $NM_i\bar{y}_i$, y la situación inversa, si y_i no está relacionada con M_i.

En cuanto a $\hat{Y}_R$ y $\hat{Y}_{pps}$, observamos que, por (9A.24) y (9A.25), la expresión $V(\hat{Y}_R)$ otorga un peso relativamente mayor a las unidades grandes que $V(\hat{Y}_{pps})$. Nótese también que $\hat{Y}_u$ y $\hat{Y}_R$ se benefician del término cpf, que puede ser sustancial en estratos pequeños (como, por ejemplo, con $n_h = 2$, $N_h = 10$). Este aspecto ha estimulado el desarrollo de la selección con probabilidades desiguales sin restitución. Desde luego, la Fórmula (9A.24) para $\hat{Y}_R$, sólo es válida en muestras grandes.

Se han realizado nuevas comparaciones entre los métodos a partir de un modelo de población infinta por Cochran, segunda Ed.,

Des Raj (1954-1958), Yates (1960), Zarcovic (1960) y Foreman y Brewer (1971). La mayoría de los autores suponen que la población finita es una muestra aleatoria de una suprapoblación infinita en la que

$$y_i = \alpha + \beta M_i + e_i; \qquad E(e_i|M_i) = 0 \tag{9A.26}$$

lo que se espera aproxime la relación vigente en muchas encuestas. Es necesario hacer algún supuesto respecto a la varianza de e_i en conglomerados de un tamaño dado.

De (9.12) en la Sec. 9.4 obtenemos, al dividir por $(N-1) \doteq N$,

$$V(e_i) = V(y_i|M_i) \doteq M_i S^2[M_i\rho + (1-\rho)]$$

Como una aproximación esto sugiere que $V(e_i) = cM_i^g$, donde $1 < g < 2$ en la mayoría de las aplicaciones. Del modelo (9A.26) se obtiene

$$\bar{Y} = \alpha + \beta\bar{M} + \bar{e}_N: \qquad \bar{\bar{Y}} = \frac{\alpha}{\bar{M}} + \beta + \frac{\bar{e}_N}{\bar{M}} \tag{9A.27}$$

Suponemos que aquí $\bar{e}_N$; esto se reduce a ignorar la cpf.

Se infiere que,

$$\frac{nV(\hat{Y}_u)}{N^2} = E(y_i - \bar{Y})^2 = E[\beta(M_i - \bar{M}) + e_i]^2$$

$$= \beta^2 V(M_i) + cE(M_i^g) \tag{9A.28}$$

$$\frac{nV(\hat{Y}_R)}{N^2} \doteq EM_i^2(\bar{y}_i - \bar{\bar{Y}})^2 = E(y_i - M_i\bar{\bar{Y}})^2$$

$$= E[\alpha(1 - M_i/\bar{M}) + e_i]^2 = \frac{\alpha^2 V(M_i)}{\bar{M}^2} + cE(M_i^g) \tag{9A.29}$$

$$\frac{nV(\hat{Y}_{pps})}{N^2} = \bar{M}EM_i(\bar{y}_i - \bar{\bar{Y}})^2 = \bar{M}EM_i\left[\alpha\left(\frac{1}{M_i} - \frac{1}{\bar{M}}\right) + \frac{e_i}{M_i}\right]^2$$

$$= \alpha^2 E\left[\frac{(M_i - \bar{M})^2}{M_i\bar{M}}\right] + v\bar{M}E(M_i^{g-1}) \doteq \frac{\alpha^2 V(M_i)}{\bar{M}^2} + c\bar{M}E(M_i^{g-1}) \tag{9A.30}$$

Por lo tanto, las varianzas aproximadamente comparables en este modelo son

$$\frac{nV_u}{N^2} = \beta^2 V(M_i) + cE(M_i^g) \tag{9A.31}$$

$$\frac{nV_R}{N^2} \doteq \alpha^2\frac{V(M_i)}{\bar{M}^2} + cE(M_i^g) \tag{9A.32}$$

$$\frac{nV_{pps}}{N^2} \doteq \frac{\alpha^2 V(M_i)}{\bar{M}^2} + c\bar{M}E(M_i^{g-1}) \tag{9A.33}$$

Considérese primero $\alpha = 0$, caso en el que $\bar{y}_i$ no está relacionada con M_i. Es claro que

$$V_R < V_u$$

Si β (que en este caso se convierte en $\bar{\bar{Y}}$) es grande, la superioridad de la estimación de razón puede ser considerable. Con $\alpha = 0$,

$$V_{pps} < V_u \quad \text{para} \quad g \geq 1$$

Estas desigualdades son válidas porque la covarianza de M_i y M_i^{g-1} es positiva si $g > 1$ y nula si $g = 1$, de modo que con $g \geq 1$.

$$E(M_i^g) = E[(M_i)(M_i^{g-1})] \geq \bar{M}E(M_i^{g-1})$$

Si $\beta = 0$ y $\alpha \neq 0$, por lo tanto, el *total* de unidad y_i no está relacionado con M_i, $\hat{Y}_u$ siempre es superior a $\hat{Y}_R$ y también a $\hat{Y}_{pps}$ salvo, posiblemente, en el caso poco verosímil de que $g = 2$ con $\beta = 0$.

Cuando ni α ni β se anulan, los comportamientos relativos de $\hat{Y}_u$ $\hat{Y}_R$, $\hat{Y}_{pps}$ dependen de las magnitudes relativas de $|\alpha|$ y $|\beta|$. Por ejemplo, $\hat{Y}_R$ supera $\hat{Y}_u$ si $\beta^2 > \alpha^2/\bar{M}^2$, como señalaron Foreman y Brewer (1971).

En lo que respecta a V_R contra V_{pps}, los coeficientes de α^2 son aproximadamente los mismos en V_R y V_{pps}. De los términos en c, $V_{pps} < V_R$ en el caso $g > 1$ que es de esperarse se aplique en casi todas las situaciones que tienen β diferente de 0, en tanto que $V_{pps} > V_R$ si $g < 1$.

Los resultados de este modelo concuerdan con las conclusiones antes sugeridas en esta sección. Si $\bar{y}_i$ no presenta o sólo presenta una tendencia moderada al aumentar M_i, el método de razón con probabilidades iguales y el método *pps* (probabilidades proporcionales al tamaño) son más precisos que la estimación insesgada con probabilidades iguales, y de hecho, pueden ser mucho más precisos. $\hat{Y}_u$ es superior si el total de unidad y_i no se encuentra relacionado con M_i. Hay menos amplitud de elección entre las estimaciones de la razón y las de *pps* como esperamos que g se encuentre entre 1 y 2, la estimación *pps* suele ser más precisa y sus resultados no se restringen a muestras grandes.

Las estimaciones $\hat{Y}_u$ y $\hat{Y}_R$ se benefician con el término cpf, cuando éste es apreciable. Anticipándonos a trabajos posteriores (Sec. 9A.12), la evidencia sugiere que los mejores métodos *pps* sin restitución se benefician aproximadamente con la misma cantidad del término cpf, como lo que beneficia al muestreo equiprobabilista.

9A.6 MUESTREO CON PROBABILIDADES DESIGUALES SIN RESTITUCION

Una parte importante de este trabajo se realizó para extensas encuestas, en las que las unidades conglomeradas se estratificaron según otro principio (por ejemplo, situación geográfica) en un número sustancial de estratos relativamente pequeños, al haberse extraído, tan sólo un pequeño número de unidades conglomerados de cada estrato. El caso $n_h = 2$ que proporciona un grado de libertad de cada estrato para estimaciones de errores de muestreo, es de particular interés.

Supongamos que se van a extraer dos unidades de un estrato. La primera se extrae con probabilidades z_j, proporcionales a alguna medida del tamaño. Sea la i-ésima unidad la que se selecciona. Si seguimos el método más natural en la segunda extracción, se selecciona una de las unidades restantes con probabilidades asignadas $z_j/(1 - z_i)$. De modo que la probabilidad total π_i de elegir la i-ésima unidad, ya sea en la primera o en la segunda extracción es

$$\pi_i = z_i + \sum_{j\neq i}^{N} \frac{z_j z_i}{(1-z_j)} = z_i\left(1 + \sum_{j\neq i}^{N} \frac{z_j}{1-z_j}\right) \tag{9A.34}$$

$$= z_i\left(1 + A - \frac{z_i}{1-z_i}\right) \tag{9A.35}$$

donde $A = \Sigma z_j/(1 - z_j)$ tomada sobre todas las N unidades.

Supongamos que $\pi_i = 2z_i$, al quedar proporcionales las probabilidades relativas de selección respecto a la medida de tamaño z_i. El estimador simple que introdujeron Horvitz y Thompson (1952),

$$\hat{Y}_{HT} = \sum_{i}^{2} \frac{y_i}{\pi_i} = \frac{1}{2}\sum_{i}^{2} \frac{y_i}{z_i}$$

tendrá entonces varianza nula si los z_i son proporcionales a los y_i, dado que $z_i = y_i/Y$ y cada unidad seleccionada dará la estimación correcta $y_i/z_i = Y$. Sin embargo, en (9A.35), las cantidades $z_i' = \pi_i/2$ están siempre más cerca de la igualdad que los z_i originales, a causa del segundo factor en (9A.35). En el ejemplo que dieron Yates y Grundy (1953), con $N = 4$, $n = 2$, $z_i = 0.1$, 0.2, 0.3 y 0.4, se obtuvo que los z_i' valen 0.1173, 0.2206, 0.3042 y 0.3579.

Se han usado tres métodos para tratar con esta distorsión de las probabilidades de selección que se piensan usar. Uno de ellos consiste en retener el método natural de selección de muestra precedente, pero con cambios apropiados en el método de estimación del total de población. Otro es utilizar un método diferente al de selección de muestra, que conserva $\pi_i = nz_i$ sujeto a ciertas restricciones en los valores de nz_i. Otro más, consiste en aceptar cierta distorsión de las probabilidades, si esto presenta ventajas compensatorias (por ejemplo, una mayor sencillez o generalidad). En la Sec. 9A.8 y sigs., se dan ejemplos de estos tres métodos.

Primero, presentamos la estimación general del total de población, más conocida, para muestreo con probabilidades desiguales sin restitución.

9A.7 EL ESTIMADOR DE HORVITZ-THOMPSON

Se selecciona una muestra de n unidades sin restitución por algún método. Sea

π_i = probabilidad de que la i-ésima unidad esté en la muestra.

π_{ij} = probabilidad de que la i-ésima y j-ésima unidades estén ambas en la muestra.

Son válidas las relaciones siguientes:

$$\sum_{i}^{N} \pi_i = n, \qquad \sum_{j\neq i}^{N} \pi_{ij} = (n-1)\pi_i, \qquad \sum_{i}^{N}\sum_{j>i} \pi_{ij} = \tfrac{1}{2}n(n-1) \tag{9A.36}$$

Para establecer la segunda relación llamemos $P(s)$ a la probabilidad de una muestra que consiste en n unidades *especificadas*. Entonces $\pi_{ij} = \Sigma P(s)$ sobre todas las muestras que contienen las unidades i-ésima y j-ésima, y $\pi_i = \Sigma P(s)$ sobre todas las muestras que contienen la i-ésima unidad. Cuando tomamos $\Sigma\pi_{ij}$ para $j \neq i$, cada $P(s)$, para una muestra que contenga la i-ésima unidad se cuenta $(n - 1)$ veces en la suma, puesto que hay otros $(n - 1)$ valores de j en la muestra. Esto prueba la segunda relación. La tercera relación se deduce de la segunda.

El estimador de Horvitz-Thompson (1952) del total de población es

$$\hat{Y}_{HT} = \sum_{i}^{n} \frac{y_i}{\pi_i} \tag{9A.37}$$

donde y_i es la medición para la i-ésima unidad.

Teorema 9A.5. Si $\pi_i > 0$, $(i = 1, 2, \ldots, N)$,

$$\hat{Y}_{HT} = \sum_{i}^{n} \frac{y_i}{\pi_i}$$

es un estimador insesgado de Y, con varianza

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 + 2 \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i y_j \qquad (9A.38)$$

donde π_{ij} es la probabilidad de que estén en la muestra las unidades i-ésima y j-ésima.

Prueba. Sea t_i $(i = 1, 2, \ldots, N)$ una variable aleatoria que toma el valor 1 si se extrae la i-ésima unidad y el valor cero en el otro caso. Entonces, t_i sigue una distribución binomial para una muestra de tamaño 1 con probabilidad π_i. Por lo tanto,

$$E(t_i) = \pi_i, \qquad V(t_i) = \pi_i(1-\pi_i) \qquad (9A.39)$$

El valor de Cov $(t_i t_j)$ también se necesita. Dado que $t_i t_j$ es 1, sólo si las dos unidades aparecen en la muestra,

$$\text{Cov}\,(t_i t_j) = E(t_i t_j) - E(t_i)E(t_j) = \pi_{ij} - \pi_i \pi_j \qquad (9A.40)$$

Luego, al considerar y_i fija y las t_i como variables aleatorias,

$$E(\hat{Y}_{HT}) = E\left(\sum_{i=1}^{N} \frac{t_i y_i}{\pi_i}\right) = \sum_{i=1}^{N} y_i = Y$$

$$V(\hat{Y}_{HT}) = \sum_{i}^{N} \left(\frac{y_i}{\pi_i}\right)^2 V(t_i) + 2 \sum_{i}^{N} \sum_{j>i}^{N} \frac{y_i}{\pi_j} \frac{y_j}{\pi_j} \text{Cov}\,(t_i t_j)$$

$$= \sum_{i}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 + 2 \sum_{i}^{N} \sum_{j>i}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i y_j \qquad (9A.41)$$

Esto termina la prueba.

Esta varianza se puede expresar en otra forma, al usar las dos primeras relaciones de (9A.36). Se tiene

$$\sum_{j \neq i} (\pi_{ij} - \pi_i \pi_j) = (n-1)\pi_i - \pi_i(n - \pi_i) = -\pi_i(1-\pi_i)$$

Al sustituir $(1 - \pi_i)$ en el primer término de (9A.41).

$$\sum_{i}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 = \sum_{i}^{N} \sum_{j \neq i}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i}\right)^2 = \sum_{i}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) \left[\left(\frac{y_i}{\pi_i}\right)^2 + \left(\frac{y_j}{\pi_j}\right)^2\right]$$

De modo que

$$V(\hat{Y}_{HT}) = \sum_{i}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) \left[\left(\frac{y_i}{\pi_i}\right)^2 + \left(\frac{y_j}{\pi_j}\right)^2 - 2 \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}\right]$$

$$= \sum_{i}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \qquad (9A.42)$$

Corolario. A partir de (9A.41), al usar el método t_i, se ve que un estimador $V(\hat{Y}_{HT})$ insesgado de muestra es

$$v_1(\hat{Y}_{HT}) = \sum_{i}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + 2 \sum_{i}^{n} \sum_{j>i}^{n} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j \pi_{ij}} y_i y_j \qquad (9A.43)$$

a condición de que ninguno de los π_{ij} de la población sea nulo.

Yates y Grundy (1953) así como Sen (1953) han dado un estimador de muestra diferente. De (9A.42), este estimador es

$$v_2(\hat{Y}_{HT}) = \sum_{i}^{n} \sum_{j>i}^{n} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \qquad (9A.44)$$

con la misma restricción sobre las π_{ij}.

Como los términos $(\pi_i \pi_j - \pi_{ij})$ suelen variar considerablemente y a veces son negativos, v_1 y v_2 tienden a ser cantidades inestables. Ambos estimadores pueden tomar valores negativos para ciertos métodos de selección de muestra. Rao y Singh (1973) compararon los coeficientes de variación de v_1 y v_2 en muestras de $n = 2$, a partir de 34 pequeñas poblaciones naturales que encontraron en textos y artículos sobre encuestas por muestreo, al utilizar el método de selección de muestra de Brewer (Sec. 9A.8) para el cual $\pi_i = 2z_i$ como se deseaba. El estimador v_2 si era considerablemente más estable y siempre mayor que 0 para este método, mientras que v_1 tomaba frecuentemente valores negativos.

Ahora veremos los métodos de selección de muestra. La bibliografía por Brewer y Hanif (1969) menciona más de 30, de los cuales describiremos 4. La mayoría de los métodos se complican uniformemente al crecer n por arriba de 2, aunque algunos se extienden con bastante facilidad. En la presentación nos fijaremos principalmente en el caso $n = 2$, por su sencillez y porque es la situación común en encuestas nacionales que usan numerosos estratos pequeños con $n_h = 2$.

9A.8 EL METODO DE BREWER

Para $n = 2$, este método de selección de muestra conserva $\pi_i = 2z_i$ y utiliza el estimador de Horvitz-Thompson

$$\hat{Y}_{HT}=\frac{y_i}{\pi_i}+\frac{y_j}{\pi_j}=\frac{1}{2}\left(\frac{y_i}{z_i}+\frac{y_j}{z_j}\right)$$

Con diferentes enfoques los métodos producidos por Brewer (1963), Rao (1965) y Durbin (1967) dieron todos los mismos valores de π_i y π_{ij}. Suponemos que cada $z_i < 1/2$.

Brewer extrae la primera unidad con lo que él llama probabilidades *proporcionales* revisadas a $z_i(1-z_i)/(1-2z_i)$, y la segunda unidad con probabilidades $z_i/(1-z_j)$ donde j es la unidad extraída primeramente. El divisor necesario para convertir las $z_i(1-z_i)/(1-2z_i)$ en probabilidades reales es su suma

$$D=\sum_{i=1}^{N}\frac{z_i(1-z_i)}{1-2z_i}=\frac{1}{2}\sum_{i=1}^{N}\frac{z_i(2-2z_i)}{1-2z_i}=\frac{1}{2}\left(1+\sum_{i}^{N}\frac{z_i}{1-2z_i}\right) \qquad (9A.45)$$

Con $n = 2$ la probabilidad de que se extraiga la i-ésima unidad, es la suma de las probabilidades de ser extraída en primero y segundo lugar. Así

$$\pi_i=\frac{z_i(1-z_i)}{D(1-2z_i)}+\frac{1}{D}\sum_{j\neq i}^{N}\frac{z_j(1-z_j)}{(1-2z_j)}\frac{z_i}{(1-z_j)}$$

$$=\frac{z_i}{D}\left[\frac{(1-2z_i)+z_i}{1-2z_i}+\sum_{j\neq i}^{N}\frac{z_j}{1-2z_j}\right]=\frac{z_i}{D}\left(1+\sum_{j=1}^{N}\frac{z_j}{1-2z_j}\right)=2z_i \qquad (9A.46)$$

al escribir (9A.45) por D. De manera semejante,

$$\pi_{ij}=\frac{z_iz_j}{D}\left(\frac{1}{1-2z_i}+\frac{1}{1-2z_j}\right)=\frac{2z_iz_j}{D}\frac{(1-z_i-z_j)}{(1-2z_i)(1-2z_j)} \qquad (9A.47)$$

Como este método utiliza la estimación *HT* (Horvitz-Thompson), el Teorema 9A.5 y su corolario proporcionan fórmulas para la varianza y varianza estimada de $\hat{Y}_B$.

El método tiene dos propiedades deseables. Brewer (1963) mostró que su varianza siempre es menor que la de la estimación $\hat{Y}_{ppz}$ en muestreo sin restitución. En segundo lugar, con un poco de álgebra, Rao (1965) muestra que $(\pi_i\pi_j-\pi_{ij})>0$ para todo $i\neq j$, por lo tanto, la estimación de Yates-Grundy para v_2 de la varianza es siempre positiva.

El método de Durbin (1967) extrae la primera unidad (i) con probabilidad z_i. Si la unidad i se extrajo primeramente, la probabilidad de que la unidad j se extraiga en segundo lugar se hace *proporcional* a

$$z_j\left[\frac{1}{(1-2z_i)}+\frac{1}{(1-2z_j)}\right] \qquad (9A.48)$$

En este caso, el divisor de las proporciones es

$$\sum_{j\neq i}^{N}z_j\left[\frac{1}{(1-2z_i)}+\frac{1}{(1-2z_j)}\right]=\frac{(1-z_i)}{1-2z_i}+\sum_{j\neq i}^{N}\frac{z_j}{(1-2z_j)}=1+\sum_{j=1}^{N}\frac{z_j}{1-2z_j} \qquad (9A.49)$$

Por lo tanto, el divisor es igual a $2D$ en el método de Brewer (9A.45). La probabilidad de que se extraiga en *primer* lugar la unidad i-ésima y en *segundo* la j-ésima es, por tanto,

$$P(i)P(j|i)=\frac{z_iz_j}{2D}\left[\frac{1}{(1-2z_i)}+\frac{1}{(1-2z_j)}\right] \qquad (9A.50)$$

Por simetría, esto iguala $P(j)P(i|j)$, de modo que las π_{ij} de Durbin son las mismas que las de Brewer en (9A.47).

Sampford (1967) ha extendido este método a muestras de tamaño n, siempre que $nz_i < 1$ para todas las unidades de la población. Con este método de selección de muestra, la probabilidad de que la muestra consista por ejemplo en las unidades 1, 2, . . ., n es una extensión natural de (9A.47), y es proporcional a

$$\left(1-\sum_{i=1}^{n}z_i\right)\prod_{i=1}^{n}z_i\Big/\prod_{i=1}^{n}(1-nz_i) \qquad (9A.51)$$

Para este método, puede mostrarse que $\pi_i = nz_i$. Se da una fórmula para π_{ij} con ciertas instrucciones para su cálculo por computadora. El estimador *HT* para Y se usa de modo que hay fórmulas para su varianza y varianza estimada. El estimador de Yates-Grundy v_2 en (9A.44) es siempre positivo. Sampford ha sugerido varios métodos para extraer la muestra de modo que se satisfaga (9A.51). Uno de ellos es la extracción de la primera unidad con probabilidades de z_i, y todas las unidades subsecuentes con probabilidades proporcionales a $z_j/(1-nz_j)$ *con restitución*. Si se obtiene una muestra con n unidades *distintas*, se acepta. Pero se desecha cualquier intento de obtención de muestra cuando por segunda vez aparece una unidad. Este método conduce a (9A.51). Para dar una idea de su rapidez, se presenta una fórmula del número esperado de intentos necesarios para obtener una muestra.

Para $n = 2$ el método de Durbin (1967) de extracción de la muestra a diferencia del método de Brewer, tiene la propiedad de que la probabilidad incondicional de extraer la unidad i es z_i tanto en la primera como en la segunda extracción. En muestreo de varias etapas para encuestas repetidas a intervalos regulares, Felligi (1963) señaló que es necesario o aconsejable dejar unidades fuera y reemplazarlas ocasionalmente según un procedimiento regular llamado esquema de rotación, debido a lo poco conveniente que son los interrogatorios prolongados de las mismas personas. Este

autor produjo un método de selección de unidades sucesivas que también posee la propiedad de Durbin. Su método, basado en cálculos repetidos, es semejante al de Brewer-Durbin, pero tiene valores ligeramente diferentes de las π_{ij}.

9A.9 EL METODO DE MURTHY

Este método utiliza la primera técnica de selección sugerida (Sec. 9A.6) al extraer las unidades sucesivas con probabilidades z_i, $z_j/(1-z_i)$, $z_k/(1-z_i-z_j)$ y así sucesivamente. El estimador de Murthy (1957) se apoya en trabajos anteriores debidos a Des Raj (1956*a*), que produjo estimaciones insesgadas de una manera ingeniosa, al basarse en el orden específico en que se extraen las n unidades de la muestra. Murthy mostró que correspondiendo a cualquier estimación ordenada de esta clase, podemos construir una estimación no ordenada que también es insesgada y presenta una varianza más baja.

Su estimador propuesto es

$$\hat{Y}_M = \frac{\sum_i^n P(s|i) y_i}{P(s)} \tag{9A.52}$$

donde

$P(s|i)$ = probabilidad condicional de obtener el conjunto de unidades extraído, dado que se sacó la i-ésima unidad en *primer* lugar

$P(s)$ = probabilidad incondicional de obtener el conjunto de unidades extraído.

Ahora probamos que la estimación $\hat{Y}_M$ es insesgada. Para cualquier unidad i de la población, $\Sigma P(s|i) = 1$, suma tomada sobre *todas* las muestras en las que se tiene la i-ésima unidad en primer lugar. Para mostrar esto en el caso $n = 2$, donde j es la otra unidad de la muestra

$$P(s|i) = \frac{z_j}{1-z_i}; \qquad \sum P(s|i) = \frac{\sum_{j\neq i}^N z_j}{1-z_i} = 1$$

Para $n = 3$ con j y k siendo las unidades segunda y tercera,

$$\sum P(s|i) = \sum_{j\neq i}^N \sum_{k\neq i,j}^N z_j z_k/(1-z_i)(1-z_i-z_j) = \sum_{j\neq i}^N z_j/(1-z_i) = 1$$

etc. para $n > 3$, luego, cuando sumamos $\Sigma P(s)\,\hat{Y}_M$ sobre todas las muestras de tamaño n, el coeficiente de y_i en la suma es 1, de modo que

$$E(\hat{Y}_M) = \sum P(s)\hat{Y}_M = \sum_i^N y_i = Y \tag{9A.53}$$

Murthy (1957) ha dado expresiones generales para $V(\hat{Y}_M)$ y $v(\hat{Y}_M)$ para cualquier n.

Cuando $n = 2$, la muestra que consiste en las unidades i y j,

$$P(s|i) = \frac{z_j}{1-z_i}; \qquad P(s|j) = \frac{z_i}{1-z_j} \tag{9A.54}$$

$$P(s) = \pi_{ij} = z_i P(s|i) + z_j P(s|j) = z_i z_j (2-z_i-z_j)/(1-z_i)(1-z_j) \tag{9A.55}$$

Por lo tanto, la estimación es, al usar (9A.52), (9A.54) y (9A.55),

$$\hat{Y}_M = \frac{1}{2-z_i-z_j}\left[(1-z_j)\frac{y_i}{z_i} + (1-z_i)\frac{y_j}{z_j}\right] \tag{9A.56}$$

Por definición $V(\hat{Y}_M)$ es (para $n = 2$)

$$\sum_s P(s)\hat{Y}_M^2 - Y^2 = \sum_i^N \sum_{j>i}^N \frac{z_i z_j (2-z_i-z_j)}{(1-z_i)(1-z_j)} \hat{Y}_M^2 - Y^2 \tag{9A.57}$$

Después de sustituir $\hat{Y}_M^2$ según (9A.56) y Y^2 y rearreglando términos, encontramos

$$V(\hat{Y}_M) = \sum_i^N \sum_{j>i}^N \frac{z_i z_j (1-z_i-z_j)}{2-z_i-z_j}\left(\frac{y_i}{z_i} - \frac{y_j}{z_j}\right)^2 \tag{9A.58}$$

Al comparar con $V(\hat{Y}_{ppz})$ para muestreos con restitución, Fórmula (9A.15), claramente $V(\hat{Y}_M) < V(\hat{Y}_{ppz})$.

Dado que para cualquier $v(\hat{Y}_M)$ propuesta su promedio es $\Sigma P(s)\, v(\hat{Y}_M)$, la siguiente es una estimación muestral insesgada de la varianza para $n = 2$

$$v(\hat{Y}_M) = \frac{(1-z_i)(1-z_j)(1-z_i-z_j)}{(2-z_i-z_j)^2}\left(\frac{y_i}{z_i} - \frac{y_j}{z_j}\right)^2 \tag{9A.59}$$

que siempre es positiva.

La fórmula correspondiente para muestras de tamaño n es

$$v(\hat{Y}_M) = \frac{1}{[P(s)]^2}\left\{\sum_i^n \sum_{j>i}^n [P(s)P(s|i,j) - P(s|i)P(s|j)] z_i z_j \left(\frac{y_i}{z_i} - \frac{y_j}{z_j}\right)^2\right\} \tag{9A.60}$$

donde $P(s|ij)$ es la probabilidad condicional de obtener la muestra, dado que se eligieron las unidades i y j (en cualquier orden) en

las dos primeras extracciones. Bayless (1968) ha dado un programa para computadora que calcula las $P(s|i)$ y las $P(s|ij)$.

9A.10 METODOS RELACIONADOS CON EL MUESTREO SISTEMATICO

En un método sugerido por Madow (1949), que como señala Murthy (1967) se utilizó en encuestas en la India, tenemos una extensión del muestreo sistemático. Sus ventajas son que la muestra fácilmente se extrae para cualquier n, el método conserva $\pi_i = nz_i$, y proporciona una estimación insesgada de Y. Además, la persona que construye la muestra puede en ocasiones poseer un conocimiento anticipado, que le permita arreglar aproximadamente las unidades en un orden para el cual el muestreo sistemático da buenos resultados. Una desventaja del método sistemático es, como de costumbre, la ausencia de una $v(\hat{Y}_{sys})$ insesgada.

Se puede extraer una muestra de tamaño n, al usar los z_i o las medidas de tamaño M_i' de las cuales se derivaron los $z_i = M_i'/\Sigma M_i' = M_i'/M_0'$. Si se usan los M_i', se forma una columna de totales cumulativos T_i de las cantidades nM_i'. De esta columna, dentro del intervalo 1 a nM_0', se asigna un intervalo de tamaño nM_i' a la unidad i. Para la selección de muestra, se extrae un número aleatorio r entre 1 y M_0' y se seleccionan las n unidades para las que el intervalo asignado incluye los números $r,\ r + M_0', \ldots, r + (n-1)M_0'$.

Este método se ilustra con los datos de la Sec. 9A.2 y con $N = 7$. Estamos suponiendo $n = 3$. Al formar los T_i y los intervalos asignados, se saca un número aleatorio entre 1 y $M_0 = 30$, digamos $r = 17$. Las unidades elegidas son las que tienen como intervalos asignados los que incluyen los números 17, 47, 77, es decir, las unidades 3, 4, 6.

Unidad	Tamaño M_i'	$T_i = 3\,\Sigma\, M_i'$	Intervalo Asignado	Unidades Seleccionadas
1	3	9	1–9	
2	1	12	10–12	
3	11	45	13–45	17 (unidad 3)
4	6	63	46–63	47 (unidad 4)
5	4	75	64–75	
6	2	81	76–81	77 (unidad 6)
7	3	90	82–90	
	—	—		
$M_0' =$	30			

Si $nz_i < 1$ (es decir, $nM_i' \leq M_0'$) para todo i, cualquier unidad tiene probabilidad nz_i de selección, y ninguna unidad se selecciona dos veces o más. Por ejemplo, la unidad 5 es elegida si $4 \leq r \leq 15$, lo que da probabilidad $12/30 = 3z_5$. Si $nz_i > 1$ para una o más unidades, tales unidades se pueden seleccionar más de una vez en la muestra, pero la frecuencia *promedio* de selección es $\pi_i = nz_i$. Esto ocurre en el ejemplo para la unidad 3, ya que $3M_3' = 33 > 30 = M_0'$. La unidad 3 se elige *una vez* si $1 \leq r \leq 12$ o $16 \leq r \leq 30$ (en todas las 27 elecciones) y *dos veces* si $13 \leq r \leq 15$ (3 elecciones). La frecuencia promedio de selección es $(1 \times 27 + 2 \times 3)/30 = 33/30 = 3z_3$.

Se deduce que

$$\hat{Y}_{sys} = \sum_i^n \frac{y_i}{\pi_i} = \frac{1}{n}\sum_i^n \frac{y_i}{z_i} \tag{9A.61}$$

es una estimación insesgada de Y.

Hartley y Rao (1962) examinaron este método con las unidades en orden aleatorio. Con la restricción $nz_i < 1$, (todo i), obtuvieron expresiones aproximadas para $V(\hat{Y}_{sys})$ y $v(\hat{Y}_{sys})$.

9A.11 EL METODO DE RAO, HARTLEY, COCHRAN

Para una muestra de tamaño n, este método comienza al constituir n grupos aleatorios de unidades, se debe elegir *una* unidad de cada grupo. Los números de unidades $N_1, N_2, \ldots, N_n$ en los grupos respectivos se pueden elegir de antemano: veremos que esto es ventajoso para hacer los N_g lo más iguales posible. Si Z_g es la medida total del tamaño para el grupo g, la i-ésima unidad del grupo recibe una probabilidad de selección z_i/Z_g. La estimación del total de población, Rao, Hartley y Cochran (1962) es

$$\hat{Y}_{RHC} = \sum_{g=1}^{n} Z_g \frac{y_g}{z_g} = \sum_{g=1}^{n} \hat{Y}_g \tag{9A.62}$$

donde y_g, z_g se refieren a la unidad extraída del grupo g.

Como los Z_g no serán iguales, este método no conserva las probabilidades de selección proporcionales a los tamaños y hay evidencia parcial de que este estimador padece de una pequeña imprecisión. Sus ventajas estriban en su sencillez y generalidad.

Al desarrollar $V(\hat{Y}_{RHC})$ promediamos en dos etapas. La etapa 1, es la aleatorización en grupos y la etapa 2, la selección de una unidad dentro de cada grupo. Para cualquier división específica en

grupos, $\hat{Y}_g$ en (9A.62) es un estimador insesgado del total de grupo Y_g, y por lo tanto, $E_2(\hat{Y}_{RHC}) = Y$. En el Cap. 10 se demuestra una fórmula bien conocida, para encontrar una varianza sobre dos etapas de muestreo, que es

$$V(\hat{Y}_{RHC}) = E_1[V_2(\hat{Y}_{RHC})] + V_1[E_2(\hat{Y}_{RHC})] \tag{10.2}$$

Como $E_2(\hat{Y}_{RHC}) = Y$ y es constante, tiene varianza nula, y el segundo término en (10.2) se anula en esta aplicación.

Para $V_2(\hat{Y}_{RHC})$ podemos usar, dentro de un grupo, la fórmula de la varianza para muestreo con restitución, dado que sólo se selecciona una unidad en cada grupo. Dentro de un grupo, las probabilidades de selección son z_i/Z_g. Por (9A.15), de la Sec. 9A.3, obtenemos para cualquier división específica (con $n_g = 1$),

$$V(\hat{Y}_g) = \sum_{i<j}^{N_g} z_i z_j \left(\frac{y_i}{z_i} - \frac{y_j}{z_j}\right)^2 \tag{9A.63}$$

tomada sobre pares de unidades del grupo g. Sobre el conjunto de subdivisiones aleatorias en grupos, la probabilidad de que cualquier par de unidades caiga en el grupo g es $N_g(N_g - 1)/N(N - 1)$. Por lo tanto, el valor promedio de $V_2(\hat{Y}_g)$ sobre la aleatorización es

$$E_1 V_2(\hat{Y}_g) = \frac{N_g(N_g-1)}{N(N-1)} \sum_{i}^{N} \sum_{j>i}^{N} z_i z_j \left(\frac{y_i}{z_i} - \frac{y_j}{z_j}\right)^2 = \frac{N_g(N_g-1)}{N(N-1)} \left(\sum_{i=1}^{N} \frac{y_i^2}{z_i} - Y^2\right) \tag{9A.64}$$

a partir de (9A.63). Como $\hat{Y}_{RHC} = \sum \hat{Y}_g$,

$$E_1[V_2(\hat{Y}_{RHC})] = \frac{\left(\sum_{g=1}^{N} N_g^2 - N\right)}{N(N-1)} \left(\sum_{1}^{N} \frac{y_i^2}{z_i} - Y^2\right) = \frac{n\left(\sum_{g=1}^{N} N_g^2 - N\right)}{N(N-1)} V(\hat{Y}_{ppz}) \tag{9A.65}$$

En esta forma, $V(\hat{Y}_{RHC})$ es simplemente un múltiple de la varianza de la estimación $\hat{Y}_{ppz}$ en muestreo con restitución. Si N/n es entero, la elección $N_g = N/n$ minimiza el multiplicador. En este caso

$$V(\hat{Y}_{RHC}) = \left(1 - \frac{n-1}{N-1}\right) V(\hat{Y}_{ppz}) \tag{9A.66}$$

Si $N = nR + k$, con R entero y $k < n$, la mejor elección consiste en formar k grupos de tamaño $(R + 1)$, siendo los restantes $(n - k)$ del tamaño R. Esto da

$$V(\hat{Y}_{RHC}) = \left[1 - \frac{n-1}{N-1} + \frac{k(n-k)}{N(N-1)}\right] V(\hat{Y}_{ppz}) \tag{9A.67}$$

Si N/n es entero, (9A.66) da $V(\hat{Y}_{RHC})/V(\hat{Y}_{ppz}) = (N - n)/(N - 1)$, y se obtiene la misma razón que en el muestreo aleatorio simple.

Un estimador insesgado de la varianza es

$$v(\hat{Y}_{RHC}) = \frac{\left(\sum_{g=1}^{n} N_g^2 - N\right)}{\left(N^2 - \sum_{g=1}^{n} N_g^2\right)} \sum_{g=1}^{n} Z_g \left(\frac{y_g}{z_g} - \hat{Y}_{RHC}\right)^2 \tag{9A.68}$$

con $N = nR + k$, y k grupos de tamaño $(R + 1)$, la Fórmula (9A.68) viene a ser

$$v(\hat{Y}_{RHC}) = \frac{N^2 + k(n-k) - Nn}{N^2(n-1) - k(n-k)} \sum_{g}^{n} Z_g \left(\frac{y_g}{z_g} - \hat{Y}_{RHC}\right)^2 \tag{9A.69}$$

9A.12 COMPARACIONES NUMERICAS

En la literatura se encuentran comparaciones de los comportamientos de algunos de los métodos, particularmente por Rao y Bayless (1969, 1970), en tres situaciones: (*a*) en poblaciones artificiales pequeñas [por ejemplo, las poblaciones con $N = 4$, $n = 2$ construidas por Yates y Grundy (1953)], (*b*) en el modelo de regresión lineal usado en la Sec. 9A.5, y (*c*) en 20 poblaciones naturales.

Aquí se comparan siete métodos en tres poblaciones artificiales con $N = 5$, $n = 2$. Los tamaños relativos z_i de las unidades son los mismos en las tres poblaciones *A*, *B* y *C* (Tabla 9A.1). En *A*, la *media por elemento* que es proporcional a y_i/z_i no se encuentra correlacionada con z_i. En *B*, la media por elemento se eleva al aumentar los tamaños. En *C*, los *totales* de unidad tienen poca relación con los tamaños.

Los planes se comparan como sigue.

1. El muestreo aleatorio simple de las unidades (probabilidad igual). Estimación: $\hat{Y}_{SRS} = N\bar{y}$.

TABLA 9A.1. TRES PEQUEÑAS POBLACIONES ARTIFICIALES

Tamaños relativos	(z_i)	0.1	0.1	0.2	0.3	0.3
Población *A*	y_i	0.3	0.5	0.8	0.9	1.5
	y_i/z_i	3	5	4	3	5
Población *B*	y_i	0.3	0.3	0.8	1.5	1.5
	y_i/z_i	3	3	4	5	5
Población *C*	y_i	0.7	0.6	0.4	0.9	0.6
	y_i/z_i	7	6	2	3	2

2. Muestreo aleatorio simple de las unidades. Estimación: razón al tamaño, $\hat{Y}_R = \sum y_i / \sum z_i$.
3. Muestreo con probabilidad proporcional al tamaño con restitución. Estimación: $\hat{Y}_{ppz} = (1/n)\sum (y_i/z_i)$.
4. Método de Brewer, muestreo sin restitución. Estimación: $\hat{Y}_B = (1/n)\sum (y_i/z_i)$.
5. Método de Murthy. Estimación: $\hat{Y}_M = \left[(1-z_j)\frac{y_i}{z_i} + (1-z_i)\frac{y_j}{z_j}\right] \Big/ (2-z_i-z_j)$.
6. El método Rao, Hartley y Cochran, con un grupo de tres unidades, y uno de dos unidades. Estimación: $\hat{Y}_{RHC} = \sum Z_g (y_g/z_g)$.
7. El método sistemático de Madow, con las unidades dispuestas en orden creciente de y_i/z_i. Estimación: $\hat{Y}_{sys} = (1/n)\sum (y_i/z_i)$.

Para darnos una idea de lo que se puede esperar de los comportamientos relativos de los métodos de Brewer, Madow y *RHC*, contra el muestreo aleatorio simple con probabilidades iguales, nótese que las varianzas relativas de los métodos más nuevos respecto a $V(\hat{Y}_{SRS})$ son aproximadamente los coeficientes relativos de variación de las cantidades y_i/z_i y y_i. Las razones $cv(y_i/z_i)/cv(y_i)$ son 0.19 en la población *A*, 0.11 en *B* y 6.7 en *C*. De modo que los métodos de probabilidad desigual deberían tener una eficiencia relativa aproximada de 5 en *A*, 9 en *B*, pero sólo 0.15 en *C*. Esta comparación es un poco injusta respecto a los métodos de probabilidad desigual, que ponderan los errores cuadrados en y_i/z_i por las medidas de tamaño. Los *ECM* de $\hat{Y}_R$ deberían estar no muy alejados de las varianzas de los métodos de probabilidad desigual.

En la práctica, al elegir un método, las consideraciones más importantes se refieren a la facilidad de estación de la muestra, la sencillez de la estimación, la exactitud de la estimación, y la existencia de un estimador de la varianza de la estimación. Con $n = 2$, todos los métodos son simples. Por lo que respecta a la exactitud, el método sistemático sólo se muestra en su caso más favorable y la información que dispone el encuestador rara vez le permitirá usarlo. Todos los métodos, a excepción $\hat{Y}_{sys}$ y $\hat{Y}_R$, proporcionan estimaciones muestrales insesgadas de la varianza de error. La Tabla 9A.2 presenta las varianzas (con el ECM para $\hat{Y}_R$).

En selección equiprobabilista, la razón de $V(\hat{Y})$ en el muestreo sin restitución a la varianza $V(\hat{Y})$ en muestreo con restitución es $(N-n)/(N-1) = 0.75$ para estas poblaciones. Para $\hat{Y}_{RHC}$, la razón $V(\hat{Y}_{RHC})/V(\hat{Y}_{ppz})$, donde $\hat{Y}_{ppz}$ denota muestreo con restitución es 0.8 en la Tabla 9A.2. Para los métodos con probabilidades desiguales, la razón varía entre poblaciones. El promedio de las tres razones en *A*, *B*, *C* es 0.74 para el método de Brewer y 0.73 para el método de Murthy, más o menos la misma razón que para los métodos equiprobabilistas. Los resultados para $n = 3$, 4 sobre pequeñas poblaciones naturales debidas a Bayless y Rao (1970) en general concuerdan con $(N-n)/(N-1)$.

TABLA 9A.2. VARIANZAS DE TOTALES ESTIMADOS DE POBLACION

Estimación	Población A	B	C
$\hat{Y}_{SRS}$	1.575	2.715	0.248
$\hat{Y}_R$ (*ECM*)	0.344	0.351	1.421
$\hat{Y}_{ppz}$	0.400	0.320	1.480
$\hat{Y}_B$	0.246	0.248	1.251
$\hat{Y}_M$	0.267	0.237	1.130
$\hat{Y}_{RHC}$	0.320	0.256	1.184
$\hat{Y}_{sys}$	0.150	0.140	0.760

En las poblaciones *A* y *B*, otros métodos son mucho más exactos que $\hat{Y}_{SRS}$ con muestreo aleatorio simple. La estimación de la razón al tamaño con muestreo aleatorio simple, tiene un comportamiento semejante a los de los métodos de probabilidades desiguales, aunque no tan bueno. En este último, no hay mucho que elegir entre los métodos de Brewer, Murthy o *RHC*. El método sistemático en su caso más favorable da muy buenos resultados.

En la población *C*, donde los totales de unidades poco tienen que ver con los tamaños, el muestreo aleatorio simple con probabilidades iguales es el mejor. Como se señaló al final de la Sec. 9A.13, esta superioridad es probablemente debida al estimador $N\bar{y}$, y no al muestreo aleatorio simple.

Rao y Bayless (1969) compararon 10 métodos no equiprobabilistas en 20 poblaciones naturales que se encuentran en diversos libros y artículos sobre el muestreo con *N* entre 9 y 35. Se limitaron a los métodos (*a*) que se sabe presentan varianzas más pequeñas que $\hat{Y}_{ppz}$ y (*b*) que proporcionan un estimador de varianza positivo insesgado.

Entre los métodos aquí presentados, compararon las eficiencias de $\hat{Y}_M$, $\hat{Y}_{RHC}$, y $\hat{Y}_{ppz}$ con la de $\hat{Y}_B$. Para $n = 2$, casi no había elección posible entre los tres métodos sin restitución, sin embargo, $\hat{Y}_M$ aventaja un poco y supera a $\hat{Y}_{RHC}$, siempre que los dos métodos difieren en pre-

cisión. También hemos notado (Sec. 9A.7) que el estimador de varianza puede ser inestable para métodos que usan la estimación de Horvitz-Thompson. Los resultados de Rao-Bayless comparan los coeficientes de variación de $v(\hat{Y}_M)$, $v(\hat{Y}_{RHC})$, y $v(\hat{Y}_B)$ en la forma de Yates-Grundy, con el de $v(\hat{Y}_{ppz})$, como medidas de las estabilidades de los estimadores de varianza. Con relación a $v(\hat{Y}_B)$ como 100%, las eficiencias medianas, de los otros estimadores de varianza son: $v(\hat{Y}_{RHC}) = 109\%$, $v(\hat{Y}_M) = 104\%$, $v(\hat{Y}_{ppz}) = 97\%$, y los tres métodos presentan algunas ganancias individuales, considerables.

Bayless y Rao (1970) dan comparaciones semejantes para $n = 3$ (14 poblaciones) y $n = 4$ (10 poblaciones). Se usó la extensión de Sampford del método de Brewer. Para n muy superior a 2, los métodos de Sampford y Murthy requieren cálculos de las probabilidades en computadora. Las varianzas de $\hat{Y}_M$ y $\hat{Y}_S$ concordaban estrechamente en casi todas las poblaciones, con $\hat{Y}_{RHC}$ un poco superada, ya que su eficiencia mediana relativa respecto a $\hat{Y}_S$ caía al 92% para $n = 4$. Por otro lado, en la estabilidad de los estimadores de varianza la superioridad de $\hat{Y}_{RHC}$ y $\hat{Y}_M$ aumentaba con eficiencias medianas relativas de 118% ($n = 3$) y 129% ($n = 4$) para $v(\hat{Y}_{RHC})$ y 110% ($n = 3$) y 120% ($n = 4$) para $v(\hat{Y}_M)$ con relación al 100% correspondiente a $v(\hat{Y}_S)$.

Rao y Bayless también compararon las eficiencias de $\hat{Y}$ y $v(\hat{Y})$ para algunos de los estimadores en el modelo de regresión lineal de la Sec. 9A.5 con $\alpha = 0$. En tanto que los resultados comparativos dependían de la potencia g, la tendencia general era semejante a la que se presenta en poblaciones naturales.

9A.13 ESTIMACIONES ESTRATIFICADAS Y DE RAZON

Las fórmulas precedentes se han presentado como se aplicarían a un solo estrato, aunque la concentración en valores pequeños de n implica estratificación previa, con base en otro principio (como la localización geográfica, la estratificación urbana-rural). La extensión de las fórmulas a esta estratificación se hace como de costumbre. Para cualquier método, si n_h, z_{hi}, π_{hi}, y_{hi}, etc., se refieren al estrato h,

$$\hat{Y} = \sum_h^L \hat{Y}_h : V(\hat{Y}) = \sum_h^L V(\hat{Y}_h) : v(\hat{Y}) = \sum v(\hat{Y}_h) \qquad (9A.70)$$

Las estimaciones de razón entran en los cálculos cuando la variable de interés es una razón (como mujeres desempleadas/mujeres aptas para el trabajo), o cuando se utiliza para incrementar la precisión. En muestreo con probabilidades desiguales, debe hacerse una

sola elección de los z_i o z_{hi} para todas las variables que requieren estimaciones. Así, por ejemplo, los z_i o z_{hi} pueden ser proporcionales a las ventas recientes *totales* de un tipo de negocio, cuando la encuesta debe estimar las ventas del día de ciertas clases de artículos, lo mismo que ventas totales del día. Para algunas clases de artículos, las ventas pueden no ser ni aproximadamente proporcionales a los z_i. En tales casos, el uso de la conocida "razón a la misma variable en la última ocasión" puede producir aumentos sustanciales de la precisión.

Con muestreo de probabilidades desiguales, el cambio en las fórmulas a las de los estimadores de razón se efectúa fácilmente. En una población no estratificada se reemplaza $\hat{Y}$ para cualquier método por $X\hat{Y}/\hat{X}$. Así, por ejemplo, con el estimador HT usamos $X(\sum^n y_i/\pi_i)/\sum^n (x_i/\pi_i)$ para la forma de razón, en lugar de $\sum (y_i/\pi_i)$. Para las aproximaciones estándar al ECM y ECM estimado de una estimación de razón, reemplácese y_i por $d_i = (y_i - Rx_i)$ en $V(\hat{Y})$ y por $d_i' = (y_i - \hat{R}x_i)$ en $v(\hat{Y})$.

Así, por ejemplo, con la forma de razón del estimador de Horvitz-Thompson, la aproximación a v_2, la varianza estimada en la forma de Yates-Grundy, es a partir de (9A.44), Pág. 323,

$$v_2[\hat{Y}_{HT(R)}] \doteq \sum_i^N \sum_{j>i}^N \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{d_i'}{\pi_i} - \frac{d_j'}{\pi_j}\right)^2 \qquad (9A.71)$$

En una población estratificada puede esperarse que con n_h pequeño, la estimación de razón combinada (Sec. 6.11) será la utilizada [esto es, $\hat{Y} = X(\sum \hat{Y}_h)/(\sum \hat{X}_h)$]. Para fórmulas de varianza aproximada, se reemplaza y_{hi} por $d_{hi} = y_{hi} - Rx_{hi}$ en V, y por $d_{hi}' = y_{hi} - \hat{R}x_{hi}$ en v.

Cuando las y_i para un atributo no están relacionadas a los z_i y no hay una estimación de razón adecuada factible para un atributo semejante, Rao (1966) investigó estimadores alternativos. Estos se obtienen de cualesquier estimadores de probabilidades desiguales (Horvitz-Thompson, Madow, Rao-Hartley-Cochran, etc.) al reemplazar y_i/z_i por Ny_i, siempre que y_i aparezca en el estimador. Entonces, por el estimador de Horvitz-Thompson, la forma alternativa es

$$\hat{Y}_{HT}^* = \frac{N}{n} \sum_i^n y_i \qquad (9A.72)$$

en tanto que para el método *RHC*

$$\hat{Y}_{RHC}^* = N \sum_g^n Z_g y_g \qquad (9A.73)$$

Los estimadores son sesgados, pero la intuición sugiere que si las y_i no tienen relación con las z_i los sesgos serían relativamente

pequeños. Por el mismo método que el usado al encontrar $V(\hat{Y}_{HT})$ en (9A.42), obtenemos

$$V[\hat{Y}^*_{HT}] = \frac{N^2}{n^2} \sum_{i}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij})(y_i - y_j)^2 \qquad (9A.74)$$

De modo que, $V[\hat{Y}^*_{HT}]$ depende de la cantidad de variación en los *totales* de unidades, así como con $V(\hat{Y}_{SRS})$. En la población C, Tabla 9A.2, este método dio 0.266 y 0.280 para el ECM de las formas alternativas de $\hat{Y}_{HT}$ y $\hat{Y}_M$, y estas formas se comportan casi tan bien como $\hat{Y}_{SRS}$. En ambos métodos, el término (Sesgo)^2 era aproximadamente del 4% del ECM.

EJERCICIOS

9A.1. Horvitz y Thompson (1952) dan los siguientes datos para estimaciones visuales M_i del número de familias y para los números reales y_i en 20 manzanas de la ciudad de Ames, Iowa. Para asistir en los cálculos, también se dan los valores de $\bar{y}_i$ y de $\bar{y}_i^2/M_i$. Se escoge una muestra de $n = 1$ manzana. Calcule las varianzas del número total de familias Y, obtenido mediante (a) la estimación insesgada en muestreo con iguales probabilidades, (b) la estimación de la razón en muestreo con iguales probabilidades, (c) muestreo con probabilidad proporcional a M_i. (Para la estimación de razón calcule el verdadero error cuadrático medio, no la fórmula aproximada.)

M_i	y_i	$\bar{y}_i$	y_i^2/M_i	M_i	y_i	$\bar{y}_i$	y_i^2/M_i
9	9	1.0000	9.000	19	19	1.0000	19.000
9	13	1.4444	18.778	21	25	1.1905	29.762
12	12	1.0000	12.000	23	27	1.1739	31.696
12	12	1.0000	12.000	24	21	0.8750	18.375
12	14	1.1667	16.333	24	35	1.4583	51.042
14	17	1.2143	20.643	25	22	0.8800	19.360
14	15	1.0714	16.071	26	25	0.9615	24.038
17	20	1.1765	23.529	27	27	1.0000	27.000
18	19	1.0556	20.056	30	47	1.5667	73.633
18	18	1.0000	18.000	40	37	0.9250	34.225

¿Concuerdan los resultados con la discusión en la Sec. 9A.5?

9A.2. Hay que enviar un cuestionario a una muestra de escuelas secundarias para encontrar cuáles proveen ciertas facilidades, como por ejemplo, un curso de ruso o una alberca. Si M_i es el número de estudiantes en la i-ésima escuela, la cantidad que hay que estimar para una facilidad dada es la proporción P de estudiantes que están en escuelas que tienen dicha facilidad, esto es,

$$P = \frac{\sum_{w} M_i}{\sum_{i=1}^{N} M_i}$$

donde $\sum_{w}$ es una suma sobre aquellas escuelas *con* la facilidad.

Una muestra de n escuelas se extrae con probabilidad proporcional a M_i con estimación. Para una facilidad, a de entre n escuelas se encuentra que la poseen. (a) Muestre que $\hat{P} = a/n$ es una estimación insesgada de P y que su varianza verdadera es $P(1-P)/n$. (*Sugerencia.* En el corolario al Teorema 9.4 deje que $y_i = M_i$ si la escuela tiene la facilidad e igual a 0 de otra manera.) (b) Muestre que una estimación insesgada de $V(\hat{P})$ es $v(\hat{P}) = \hat{P}(1-\hat{P})/(n-1)$.

9A.3. Las unidades grandes en una población se arreglan en un número finito de clases por tamaño: todas las unidades en la clase h contienen M_h unidades pequeñas. (a) ¿En qué condiciones el muestreo con *pps* da, en promedio, la misma distribución en clases por tamaño en la muestra que estratificación por tamaño de la unidad, con distribución óptima para un tamaño fijo de muestra? (b) Si la varianza entre unidades grandes en la clase h es kM_h, donde k es una constante para todas las clases, ¿cuál sistema de probabilidades de selección de las unidades da una muestra en la que los tamaños tienen aproximadamente la misma distribución que una muestra aleatoria estratificada con distribución óptima para un tamaño fijo de muestra?

9A.4. Para una población con $N = 3$, $z_i = \frac{1}{2}, \frac{1}{3}, \frac{1}{6}$ y $y_i = 7, 5, 2$, se extraen dos unidades sin restitución, la primera con probabilidad proporcional a z_i y la segunda con probabilidad proporcional a los otros tamaños. (a) Verificar que $\pi_1 = \frac{51}{60}$, $\pi_2 = \frac{44}{60}$, $\pi_3 = \frac{25}{60}$ y que $\pi_{12} = \frac{35}{60}$, $\pi_{13} = \frac{16}{60}$, $\pi_{23} = \frac{9}{60}$. (b) Para este método de selección de la muestra, comparar las varianzas de $\hat{Y}_{HT}$ y $\hat{Y}_M$ y también, compararlas con la varianza de $\hat{Y}_{RHC}$ al usar su método de selección de muestra. Se pueden construir las tres estimaciones posibles, o bien usar las fórmulas de la varianza. (c) Mostrar que la razón de $V(\hat{Y}_M)$ a $V(\hat{Y}_{ppz})$ en un muestreo con restitución, se acerca al valor $\frac{1}{2}$ que se aplica a muestreo equiprobabilista.

9A.5. Para la población del Ejercicio 9A.4, una segunda variable tenía los valores $y_{2i} = 8, 5, 9$, no todos estrechamente relacionados a los z_i, de modo que con el método de muestreo del Ejercicio 9A.4, $\hat{Y}_{HT}$ y $\hat{Y}$ se esperaría que se comportasen mediocremente para esta variable. Para el estimador de Rao, $\hat{Y}^*_{HT} = 1.5(y_{2i} + y_{2j})$, comparar su ECM con la varianza de $\hat{Y}_{SRS} = 1.5(y_{2i} + y_{2j})$ en muestreo equiprobabilista. ¿Cuál es la contribución del sesgo al ECM?

9A.6. Para el método de Brewer con $n = 2$ en la Sec. 9A.8 se ha mostrado que

$$\pi_{ij} = \frac{4z_i z_j (1 - z_i - z_j)}{(1-2z_i)(1-2z_j)} \Big/ \left(1 + \sum_{i}^{N} \frac{z_i}{1-2z_i}\right)$$

(a) Mostrar que si cada $z_i < \frac{1}{2}$,

$$0 < \pi_{ij} < 4z_i z_j \qquad (\text{cada } i \neq j)$$

(b) Mostrar que este resultado hace positivo al estimador de varianza de Yates-Grundy, para este método.

pequeños. Por el mismo método que el usado al encontrar $V(\hat{Y}_{HT})$ en (9A.42), obtenemos

$$V[\hat{Y}^*_{HT}] = \frac{N^2}{n^2} \sum_{i}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij})(y_i - y_j)^2 \qquad (9A.74)$$

De modo que, $V[\hat{Y}^*_{HT}]$ depende de la cantidad de variación en los *totales* de unidades, así como con $V(\hat{Y}_{SRS})$. En la población C, Tabla 9A.2, este método dio 0.266 y 0.280 para el ECM de las formas alternativas de $\hat{Y}_{HT}$ y $\hat{Y}_M$, y estas formas se comportan casi tan bien como $\hat{Y}_{SRS}$. En ambos métodos, el término (Sesgo)^2 era aproximadamente del 4% del ECM.

EJERCICIOS

9A.1. Horvitz y Thompson (1952) dan los siguientes datos para estimaciones visuales M_i del número de familias y para los números reales y_i en 20 manzanas de la ciudad de Ames, Iowa. Para asistir en los cálculos, también se dan los valores de $\bar{y}_i$ y de $\bar{y}_i^2/M_i$. Se escoge una muestra de $n = 1$ manzana. Calcule las varianzas del número total de familias Y, obtenido mediante (*a*) la estimación insesgada en muestreo con iguales probabilidades, (*b*) la estimación de la razón en muestreo con iguales probabilidades, (*c*) muestreo con probabilidad proporcional a M_i. (Para la estimación de razón calcule el verdadero error cuadrático medio, no la fórmula aproximada.)

M_i	y_i	$\bar{y}_i$	y_i^2/M_i	M_i	y_i	$\bar{y}_i$	y_i^2/M_i
9	9	1.0000	9.000	19	19	1.0000	19.000
9	13	1.4444	18.778	21	25	1.1905	29.762
12	12	1.0000	12.000	23	27	1.1739	31.696
12	12	1.0000	12.000	24	21	0.8750	18.375
12	14	1.1667	16.333	24	35	1.4583	51.042
14	17	1.2143	20.643	25	22	0.8800	19.360
14	15	1.0714	16.071	26	25	0.9615	24.038
17	20	1.1765	23.529	27	27	1.0000	27.000
18	19	1.0556	20.056	30	47	1.5667	73.633
18	18	1.0000	18.000	40	37	0.9250	34.225

¿Concuerdan los resultados con la discusión en la Sec. 9A.5?

9A.2. Hay que enviar un cuestionario a una muestra de escuelas secundarias para encontrar cuáles proveen ciertas facilidades, como por ejemplo, un curso de ruso o una alberca. Si M_i es el número de estudiantes en la i-ésima escuela, la cantidad que hay que estimar para una facilidad dada es la proporción P de estudiantes que están en escuelas que tienen dicha facilidad, esto es,

$$P = \frac{\sum_{w} M_i}{\sum_{i=1}^{N} M_i}$$

donde $\sum_{w}$ es una suma sobre aquellas escuelas *con* la facilidad.

Una muestra de n escuelas se extrae con probabilidad proporcional a M_i con estimación. Para una facilidad, a de entre n escuelas se encuentra que la poseen. (*a*) Muestre que $\hat{P} = a/n$ es una estimación insesgada de P y que su varianza verdadera es $P(1-P)/n$. (*Sugerencia.* En el corolario al Teorema 9.4 deje que $y_i = M_i$ si la escuela tiene la facilidad e igual a 0 de otra manera.) (*b*) Muestre que una estimación insesgada de $V(\hat{P})$ es $v(\hat{P}) = \hat{P}(1-\hat{P})/(n-1)$.

9A.3. Las unidades grandes en una población se arreglan en un número finito de clases por tamaño: todas las unidades en la clase h contienen M_h unidades pequeñas. (*a*) ¿En qué condiciones el muestreo con *pps* da, en promedio, la misma distribución en clases por tamaño en la muestra que estratificación por tamaño de la unidad, con distribución óptima para un tamaño fijo de muestra? (*b*) Si la varianza entre unidades grandes en la clase h es kM_h, donde k es una constante para todas las clases, ¿cuál sistema de probabilidades de selección de las unidades da una muestra en la que los tamaños tienen aproximadamente la misma distribución que una muestra aleatoria estratificada con distribución óptima para un tamaño fijo de muestra?

9A.4. Para una población con $N = 3$, $z_i = \frac{1}{2}, \frac{1}{3}, \frac{1}{6}$ y $y_i = 7, 5, 2$, se extraen dos unidades sin restitución, la primera con probabilidad proporcional a z_i y la segunda con probabilidad proporcional a los otros tamaños. (*a*) Verificar que $\pi_1 = \frac{51}{60}$, $\pi_2 = \frac{44}{60}$, $\pi_3 = \frac{25}{60}$ y que $\pi_{12} = \frac{35}{60}$, $\pi_{13} = \frac{16}{60}$, $\pi_{23} = \frac{9}{60}$. (*b*) Para este método de selección de la muestra, comparar las varianzas de $\hat{Y}_{HT}$ y $\hat{Y}_M$ y también, compararlas con la varianza de $\hat{Y}_{RHC}$ al usar su método de selección de muestra. Se pueden construir las tres estimaciones posibles, o bien usar las fórmulas de la varianza. (*c*) Mostrar que la razón de $V(\hat{Y}_M)$ a $V(\hat{Y}_{ppz})$ en un muestreo con restitución, se acerca al valor $\frac{1}{2}$ que se aplica a muestreo equiprobabilista.

9A.5. Para la población del Ejercicio 9A.4, una segunda variable tenía los valores $y_{2i} = 8, 5, 9$, no todos estrechamente relacionados a los z_i, de modo que con el método de muestreo del Ejercicio 9A.4, $\hat{Y}_{HT}$ y $\hat{Y}$ se esperaría que se comportasen mediocremente para esta variable. Para el estimador de Rao, $\hat{Y}^*_{HT} = 1.5(y_{2i} + y_{2j})$, comparar su ECM con la varianza de $\hat{Y}_{SRS} = 1.5(y_{2i} + y_{2j})$ en muestreo equiprobabilista. ¿Cuál es la contribución del sesgo al ECM?

9A.6. Para el método de Brewer con $n = 2$ en la Sec. 9A.8 se ha mostrado que

$$\pi_{ij} = \frac{4z_i z_j (1 - z_i - z_j)}{(1-2z_i)(1-2z_j)} \Big/ \left(1 + \sum_{i}^{N} \frac{z_i}{1-2z_i}\right)$$

(*a*) Mostrar que si cada $z_i < \frac{1}{2}$,

$$0 < \pi_{ij} < 4z_i z_j \qquad (\text{cada } i \neq j)$$

(*b*) Mostrar que este resultado hace positivo al estimador de varianza de Yates-Grundy, para este método.

Sugerencia. Para (a) es suficiente mostrar que

$$(1-z_i-z_j)=(1-2z_i)(1-2z_j)\left[1+\frac{z_i}{1-2z_i}+\frac{z_j}{1-2z_j}\right]$$

9A.7. (a) Para el método de Durbin, con $n = 2$ verificar directamente que la probabilidad de que se extraigan la j-ésima unidad en segundo lugar es z_j como se dijo en la Pág. 325.

(b) Con $N = 4$, $z_i = 0.1, 0.2, 0.3, 0.4$, calcular la probabilidad de que con el método de Brewer se extraiga primeramente la unidad 1, y después la probabilidad de extraerla en segundo lugar. Verificar que ambas probabilidades suman $0.2 = 2z_i$.

9A.8. En el método sistemático de Madow se puede elegir una unidad más de una vez en la muestra si $nz_i > 1$ (es decir, $nM_i' > M_0'$). Mostrar (como se dijo en la Pág. 329) que para tales unidades, la frecuencia promedio de selección es nz_i, de modo que el estimador de Horvitz-Thompson de Y sigue insesgado para $nz_i > 1$.

CAPITULO 10

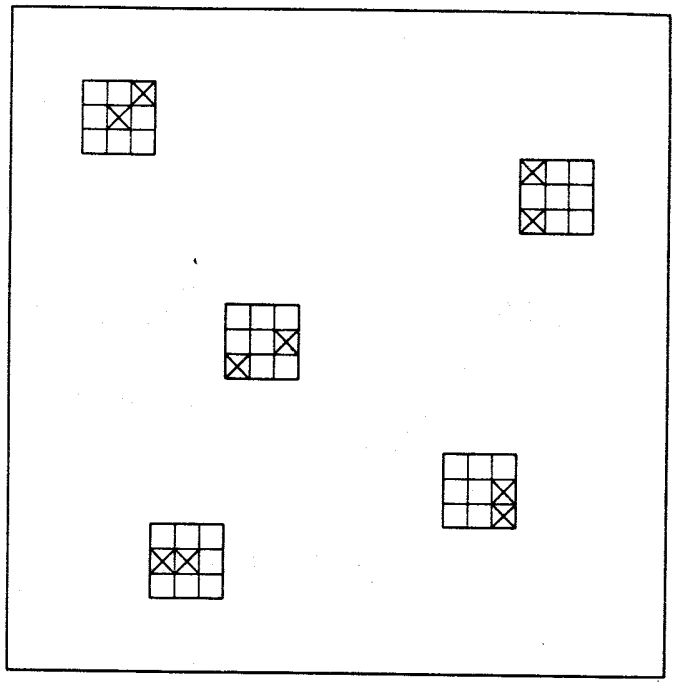
Submuestreo con Unidades de Tamaños Iguales

10.1 MUESTREO EN DOS ETAPAS

Supongamos que cada unidad de la población se puede dividir en cierto número de unidades más pequeñas, o subunidades. Se ha tomando una muestra de n unidades. Si las subunidades contenidas en una unidad seleccionada dan resultados semejantes, no parece económico medirlas todas. Una práctica acostumbrada consiste en seleccionar y medir una muestra de subunidades de alguna unidad elegida. Esta técnica se llama *submuestreo*, dado que la unidad no se mide completamente, sino que a su vez es objeto de un muestreo. Mahalanobis le ha dado el nombre de *muestreo en dos etapas*, porque la muestra se obtiene en dos pasos. Lo primero que se hace es seleccionar una muestra de unidades, que a menudo se denominan *unidades primarias*, y después se selecciona una muestra de unidades de la segunda etapa o subunidades de cada una de las unidades primarias elegidas.

El submuestreo tiene una gran variedad de aplicaciones que rebasan en mucho los alcances inmediatos de las encuestas por muestreo. Cuando un proceso involucra pruebas químicas, físicas o biológicas, realizables en una pequeña cantidad de material, comúnmente se obtendrá ésta como submuestra de una cantidad más considerable que a su vez es una muestra.

En el presente capítulo, vamos a considerar el caso más simple en el que cada unidad contiene el mismo número M de subunidades, de las cuales se elegirán m al submuestrear cualquier unidad. En la Fig. 10.1 exhibimos una representación esquemática de una muestra en dos etapas, para la que $M = 9$ y $m = 2$.



⊠ denota un elemento en la muestra

FIG. 10.1. Representación esquemática de un muestreo en dos etapas ($N = 81, n = 5, M = 9, m = 2$)

La principal ventaja del muestreo en dos etapas es su flexibilidad con respecto al muestreo de una etapa. Se reduce al muestreo de una etapa en el caso $m = M$, pero, a menos que ésta sea la mejor elección de m, tenemos la posibilidad de tomar un valor más pequeño que resulta más eficiente. Como de costumbre, lo que se trata es de alcanzar un cierto equilibrio entre precisión estadística y costo. Cuando las subunidades de la misma unidad concuerdan estrechamente, la precisión sugiere un valor pequeño de m. Por otro lado, en ocasiones es casi tan económico medir toda la unidad que submuestrearla como, por ejemplo, cuando la unidad es una vivienda y una sola persona puede proporcionar datos exactos respecto a todos los ocupantes de la misma.

10.2 DETERMINACION DE MEDIAS Y VARIANZAS EN MUESTREO DE DOS ETAPAS

En el muestreo de dos etapas, el plan de muestreo da primero un método de elección de n unidades. En seguida, para cada unidad

seleccionada, se da un método para seleccionar de un número específico de subunidades de la misma. Al encontrar la media y la varianza de una estimación hay que tomar promedios sobre todas las muestras que puede generar este proceso de dos etapas. Una manera de calcular este promedio, consiste primero en promediar la estimación sobre todas las selecciones de la segunda etapa, que pueden extraerse de un conjunto *fijo* de n unidades elegido por el plan. Posteriormente, se promedian sobre todas las posibles selecciones de n unidades por medio del plan. Para una estimación $\hat{\theta}$, este método se expresa como

$$E(\hat{\theta}) = E_1[E_2(\hat{\theta})] \tag{10.1}$$

donde E denota el valor esperado o promedio sobre todas las muestras, E_2 denota el promedio sobre todas las posibles selecciones, en la segunda etapa, de un conjunto fijo de unidades, y E_1 denota el promedio sobre todas las selecciones de la primera etapa.

Para $V(\hat{\theta})$ este método da el siguiente resultado, fácil de recordar.

$$V(\hat{\theta}) = V_1[E_2(\hat{\theta})] + E_1[V_2(\hat{\theta})] \tag{10.2}$$

donde $V_2(\hat{\theta})$ es la varianza sobre todas las posibles selecciones de submuestra para un conjunto dado de unidades. Para demostrarlo, sea $\theta = E(\hat{\theta})$ (donde θ no es necesariamente la cantidad que $\hat{\theta}$ va a estimar, puesto que $\hat{\theta}$ puede ser sesgada). Por definición

$$V(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = E_1E_2(\hat{\theta} - \theta)^2 \tag{10.3}$$

Pero

$$E_2(\hat{\theta} - \theta)^2 = E_2(\hat{\theta}^2) - 2\theta E_2(\hat{\theta}) + \theta^2 \tag{10.4}$$

$$= [E_2(\hat{\theta})]^2 + V_2(\hat{\theta}) - 2\theta E_2(\hat{\theta}) + \theta^2 \tag{10.5}$$

Ahora se promedia sobre las selecciones de la primera etapa. Dado que $E_1E_2(\hat{\theta}) = \theta$,

$$V(\hat{\theta}) = E_1[E_2(\hat{\theta})]^2 - \theta^2 + E_1[V_2(\hat{\theta})] \tag{10.6}$$

$$= V_1[E_2(\hat{\theta})] + E_1[V_2(\hat{\theta})] \tag{10.7}$$

La Fórmula (10.7) admite una extensión natural a tres o más etapas. Para muestreo de tres etapas

$$V(\hat{\theta}) = V_1[E_2(E_3(\hat{\theta}))] + E_1\{V_2[E_3(\hat{\theta})]\} + E_1\{E_2[V_3(\hat{\theta})]\} \tag{10.7'}$$

10.3 VARIANZA DE LA MEDIA ESTIMADA EN MUESTREO DE DOS ETAPAS

Se usará la notación siguiente.

y_{ij} = valor obtenido para la j-ésima subunidad en la i-ésima unidad primaria.

$\bar{y}_i = \sum_{j=1}^{m} \frac{y_{ij}}{m}$ media muestral por subunidad en la i-ésima unidad primaria.

$\bar{\bar{y}} = \sum_{i=1}^{n} \frac{\bar{y}_i}{n}$ = media global de muestra por subunidades.

$S_1^2 = \frac{\sum_{i=1}^{N} (\bar{Y}_i - \bar{\bar{Y}})^2}{N-1}$ = varianza entre medias de unidades primarias.

$S_2^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} (y_{ij} - \bar{Y}_i)^2}{N(M-1)}$ = varianza entre subunidades dentro de unidades primarias.

Hay que observar que Y_i denota el total sobre las subunidades de la i-ésima unidad (que en los Cap. 9 y 9A se denota con y_i).

Teorema 10.1. Si las n unidades y m subunidades de cada unidad elegida, se seleccionan por muestreo aleatorio simple, $\bar{\bar{y}}$ es la estimación insesgada de $\bar{\bar{Y}}$ con varianza

$$V(\bar{\bar{y}}) = \left(\frac{N-n}{N}\right)\frac{S_1^2}{n} + \left(\frac{M-m}{M}\right)\frac{S_2^2}{mn} \tag{10.8}$$

Prueba. Con muestreo aleatorio simple en ambas etapas,

$$E(\bar{\bar{y}}) = E_1[E_2(\bar{\bar{y}})] = E_1\left(\frac{1}{n}\sum^{n} \bar{Y}_i\right) = \left(\frac{1}{N}\sum^{N} \bar{Y}_i\right) = \bar{\bar{Y}} \tag{10.9}$$

Para $V(\bar{\bar{y}})$, usamos la Fórmula (10.2).

$$V(\bar{\bar{y}}) = V_1[E_2(\bar{\bar{y}})] + E_1[V_2(\bar{\bar{y}})] \tag{10.10}$$

Dado que $E_2(\bar{\bar{y}}) = \sum^{n} \bar{Y}_i/n$, el primer término de la derecha es la varianza de la media por subunidad para un muestreo aleatorio simple de una etapa, de n unidades. De modo que, por el Teorema 2.2,

$$V_1[E_2(\bar{\bar{y}})] = \left(\frac{N-n}{N}\right)\frac{S_1^2}{n} \tag{10.11}$$

Además, con $\bar{\bar{y}} = \sum^{n} \bar{y}_i/n$ y muestreo aleatorio simple en la segunda etapa,

$$V_2(\bar{\bar{y}}) = \frac{(M-m)}{Mn^2}\frac{\sum^{n} S_{2i}^2}{m} \tag{10.12}$$

donde $S_{2i}^2 = \sum_{j}^{M} (y_{ij} - \bar{Y}_i)^2/(M-1)$ es la varianza entre subunidades para la i-ésima unidad primaria. Cuando promediamos sobre las muestras de primera etapa, $\sum_{i}^{n} S_{2i}^2/n$ promedia a $\sum^{N} S_{2i}^2/N = S_2^2$.

Por lo tanto,

$$E_1[V_2(\bar{\bar{y}})] = \left(\frac{M-m}{M}\right)\frac{S_2^2}{mn} \tag{10.13}$$

El teorema resulta de la Fórmula (10.10), al sumar (10.11) y (10.13).

Si $f_1 = n/N$ y $f_2 = m/M$ son las fracciones del muestreo en la primera y segunda etapas, otra forma del resultado es

$$V(\bar{\bar{y}}) = \frac{1-f_1}{n}S_1^2 + \frac{1-f_2}{mn}S_2^2 \tag{10.14}$$

10.4 ESTIMACION MUESTRAL DE LA VARIANZA

Teorema 10.2. En las condiciones del Teorema 10.1, tenemos como estimación insesgada de $V(\bar{\bar{y}})$

$$v(\bar{\bar{y}}) = \frac{1-f_1}{n}s_1^2 + \frac{f_1(1-f_2)}{mn}s_2^2 \tag{10.15}$$

donde $f_1 = n/N$, $f_2 = m/M$, y

$$s_1^2 = \frac{\sum^{n} (\bar{y}_i - \bar{\bar{y}})^2}{n-1} \qquad s_2^2 = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2}{n(m-1)} \tag{10.16}$$

Prueba.

$$(n-1)s_1^2 = \sum^{n} (\bar{y}_i - \bar{\bar{y}})^2 = \sum^{n} \bar{y}_i^2 - n\bar{\bar{y}}^2 \tag{10.17}$$

Por lo tanto

$$(n-1)E_2(s_1^2) = \sum^{n} \bar{Y}_i^2 + \sum^{n} \frac{(1-f_2)}{m}S_{2i}^2 - n\bar{Y}_n^2 - \frac{1}{n}\sum^{n} \frac{(1-f_2)}{m}S_{2i}^2 \tag{10.18}$$

donde $\bar{Y}_n = \sum^n Y_i/n$. El último término de la derecha es válido, porque el submuestreo es independiente en diferentes unidades y $\bar{y} = \sum^n \bar{y}_i/n$. En consecuencia,

$$(n-1)E_2(s_1^2) = \sum^n (\bar{Y}_i - \bar{Y}_n)^2 + \frac{(n-1)(1-f_2)}{nm} \sum^n \frac{S_{2i}^2}{n} \qquad (10.19)$$

Al multiplicar por $(1 - f_1)/n(n - 1)$ y promediar sobre la primera etapa de muestreo aleatorio simple,

$$E\frac{(1-f_1)}{n} s_1^2 = \frac{(1-f_1)}{n} S_1^2 + \frac{(1-f_1)(1-f_2)}{mn} S_2^2 \qquad (10.20)$$

Al comparar con (10.14) para $V(\bar{y})$, nótese que el término en S_2^2 está indefecto por la cantidad $f_1(1 - f_2)S_2^2/mn$. Dado que E_1E_2 $(s_2^2) = S_2^2$, por lo tanto, una estimación insesgada de $V(\bar{y})$ es

$$v(\bar{y}) = \frac{1-f_1}{n} s_1^2 + \frac{f_1(1-f_2)}{mn} s_2^2 \qquad (10.21)$$

Corolario. Un resultado utilizado más tarde es que a partir de (10.20)

$$E(s_1^2) = S_1^2 + \frac{(1-f_2)}{m} S_2^2 = S_1^2 + \frac{S_2^2}{m} - \frac{S_2^2}{M} \qquad (10.22)$$

Resulta que una estimación insesgada de S_1^2 es

$$s_1^2 - \frac{s_2^2(1-f_2)}{m}$$

Notas al Teorema 10.2. Si $M = m$ o sea $f_2 = 1$, la Fórmula (10.15) resulta apropiada para el muestreo aleatorio simple de las unidades. Si $n = N$, la fórmula es la del muestreo aleatorio estratificado proporcional, puesto que las unidades primarias pueden considerarse entonces como estratos, todos ellos muestreados. En relación con esto, el muestreo en dos etapas es una clase de estratificación incompleta, con las unidades como estratos.

Cuando puede despreciarse $f_1 = n/N$, obtenemos el simple resultado,

$$v(\bar{y}) = \frac{s_1^2}{n} = \frac{\sum^n (\bar{y}_i - \bar{y})^2}{n(n-1)} \qquad (10.23)$$

Por lo tanto, la varianza estimada podrá calcularse con sólo conocer las medias de unidades. Este resultado es útil cuando el submuestreo es sistemático, porque en este caso no podemos calcular una estimación insesgada de S_2^2. Pero (10.23) es aún aplicable, siempre

que n/N sea pequeña. Si n/N no es pequeña, (10.23) sobreestima por la cantidad $f_1S_1^2/n$, como se ve a partir de (10.20) y (10.14).

10.5 LA ESTIMACION DE PROPORCIONES

Si las subunidades se clasifican en dos clases y estimamos la proporción de ellas que pertenece a la primera clase, se pueden aplicar las fórmulas precedentes por el artificio acostumbrado de definir y_{ij} como 1, si la subunidad correspondiente está en esta clase, y como cero en el caso contrario. Sea $p_i = a_i/m$ la proporción que cae dentro de la primera clase en la submuestra de la i-ésima unidad. Las dos varianzas estimadas s_1^2 y s_2^2 requeridas para el Teorema 10.2 son como sigue:

$$s_1^2 = \frac{\sum_{i=1}^n (p_i - \bar{p})^2}{n-1}$$

$$s_2^2 = \frac{m}{n(m-1)} \sum_{i=1}^n p_i q_i$$

donde $\bar{p} = \Sigma p_i/n$. En consecuencia, por el Teorema 10.2,

$$v(\bar{p}) = \frac{1-f_1}{n(n-1)} \sum_i^n (p_i - \bar{p})^2 + \frac{f_1(1-f_2)}{n^2(m-1)} \sum_i^n p_i q_i \qquad (10.24)$$

Ejemplo. En un estudio sobre enfermedades de las plantas, éstas fueron plantadas en 160 pequeñas parcelas que contienen nueve plantas cada una. Se escogió una muestra al azar de 40 parcelas y se examinaron tres plantas en cada parcela muestreada para ver si estaban enfermas. Se encontró que 22 parcelas no tuvieron plantas enfermas (de las tres), 11 tuvieron una, cuatro tuvieron dos y tres tuvieron tres. Estime la proporción de plantas enfermas y su e.e. El símbolo ϕ denota las frecuencias 22, 11, 4, 3.

Tenemos $N = 160$, $M = 9$, $n = 40$, $m = 3$. Para encontrar s_1^2 y s_2^2, es conveniente trabajar primero con los números de plantas enfermas $(3p_i)$ y los números de plantas sanas $(3q_i)$. Los cálculos se plantean como sigue:

	Frecuencia				
$3p_i$	ϕ	$9p_iq_i$	$9\phi p_iq_i$	$3\phi p_i$	$9\phi p_i^2$
0	22	0	0	0	0
1	11	2	22	11	11
2	4	2	8	8	16
3	3	0	0	9	27
	40		30	28	54

$$\bar{p}=\frac{3\sum\phi p_i}{3\sum\phi}=\frac{28}{120}=0.233$$

$$\sum\phi(p_i-\bar{p})^2=\frac{1}{(9)}\left(54-\frac{(28)^2}{40}\right)=3.822$$

$$\sum\phi p_i q_i=\frac{30}{9}=3.333$$

Por lo tanto, de la fórmula inmediatamente anterior a este ejemplo,

$$v(\bar{p})=\frac{(3)(3.822)}{(4)(40)(39)}+\frac{(2)(3.333)}{(4)(3)(1600)(2)}=0.00201$$

La proporción enferma es 0.233 con e.e. 0.045. La fórmula aproximada $s_1/\sqrt{n}$, de (10.23), da 0.049, una estimación razonablemente buena considerando que $f_1 = 1/4$.

10.6 FRACCIONES OPTIMAS DE MUESTREO Y SUBMUESTREO

Estas dependen del tipo de función de costos. Si los gastos de viaje entre unidades son de poca importancia, una forma de la función que ha demostrado ser útil, es

$$C=c_1 n+c_2 nm$$

La primera componente del costo, $c_1 n$ es proporcional al número de unidades primarias de la muestra, la segunda, $c_2 nm$ es proporcional al número total de unidades o elementos de la segunda etapa. Del Teorema 10.1 se ve que $V(\bar{y})$ puede escribirse

$$V(\bar{y})=\frac{1}{n}\left(S_1^2-\frac{S_2^2}{M}\right)+\frac{1}{mn}S_2^2-\frac{1}{N}S_1^2 \qquad (10.25)$$

El último término de la derecha no depende de la elección de n y m. Al minimizar V para C fijo, o C para V fija, equivale a minimizar el producto

$$\left(V+\frac{1}{N}S_1^2\right)C=\left[\left(S_1^2-\frac{S_2^2}{M}\right)+\frac{S_2^2}{m}\right](c_1+c_2 m)$$

Por la desigualdad de Cauchy-Schwarz

$$m_{opt}=\frac{S_2}{\sqrt{S_1^2-S_2^2/M}}\sqrt{c_1/c_2} \qquad (10.26)$$

siempre que $S_1^2 > S_2^2/M$. Se redondea m_{opt} al entero más cercano. Si m es un entero tal que $m < m_{opt} < m+1$, la siguiente es una regla más conveniente para m_{opt} pequeña, (Cameron, 1951): redondear hacia arriba si $m^2_{opt} > m(m+1)$, de otra manera redondear hacia abajo. Si $m_{opt} > M$ o si $S_1^2 < S_2^2/M$, tómese $m = M$, al usar muestreo de una etapa. (El producto $(V + S_1^2/N)C$ es una función estrictamente decreciente de m cuando $S_1^2 < S_2^2/M$.)

Para encontrar el valor de n, se resuelve la ecuación de costo o la ecuación de varianza, según la que se haya preasignado.

En la mayoría de las situaciones prácticas, el óptimo es relativamente bajo. Un error de unas cuantas unidades en la elección de m produce sólo una pequeña pérdida de precisión, como lo ilustra el ejemplo siguiente. Escríbase

$$S_u^2=S_1^2-\frac{S_2^2}{M} \qquad (10.27)$$

Ejemplo. Sea

$$c_1=10c_2, \qquad S_2=1.3S_u$$

entonces

$$m_{opt}=1.3\sqrt{10}=4.1$$

Consideraremos fijo el costo total y veremos cómo cambia la varianza de $\bar{y}$ al cambiar m. Se supone que N es elevado. De (10.14)

$$V(\bar{y})=\frac{S_u^2}{n}+\frac{S_2^2}{nm}$$

$$=\left(S_u^2+\frac{S_2^2}{m}\right)\frac{c_1+mc_2}{C}$$

al eliminar n por medio de la ecuación de costo. Esto da

$$V(\bar{y})=\frac{S_u^2 c_2}{C}\left(1+\frac{S_2^2}{mS_u^2}\right)\left(\frac{c_1}{c_2}+m\right)=\frac{S_u^2 c_2}{C}\left(1+\frac{1.69}{m}\right)(10+m)$$

Al omitir el factor constante, la varianza relativa se puede calcular para diferentes valores de m. La Tabla 10.1 muestra estas varianzas y las precisiones relativas (tomar como estándar la precisión máxima para $m = 4$).

Para cualquier valor de m entre 2 y 9, la pérdida de precisión relativa a la óptima es inferior al 12%.

TABLA 10.1. VARIANZAS RELATIVAS Y PRECISIONES PARA VALORES DIFERENTES DE m

$m=$	1	2	3	4	5	6	7	8	9	10
Varianza relativa	29.59	22.14	20.32	19.92	20.07	20.51	21.10	21.80	22.56	23.38
Precisión relativa	0.67	0.90	0.98	1.00	0.99	0.97	0.94	0.91	0.88	0.85

En la práctica, la elección de m requiere las estimaciones de c_1/c_2 y S_2/S_1 o en forma equivalente S_2/S_u. Debido a lo achatado de la óptima, no es necesario obtener estas razones con gran precisión. Si c_1/c_2 se conoce razonablemente bien y un valor de m, digamos m_0, se ha seleccionado, una tabla muy útil (Brooks, 1955) muestra la gama de valores de S_2^2/S_u^2 dentro de la cual m_0 da una precisión de al menos un 90% de la óptima.

La tabla se obtuvo como sigue. Para un costo dado, al suponer N grande, la precisión relativa de m_0 a m_{opt} se encuentra que es

$$\frac{V(\bar{y}|m_{opt})}{V(\bar{y}|m_0)} = \frac{(S_u\sqrt{c_1}+S_2\sqrt{c_2})^2}{S_u^2c_1+S_2^2c_2+m_0c_2S_u^2+c_1S_2^2/m_0} \tag{10.28}$$

El conjunto de valores de S_2/S_u para los cuales esta expresión excede algún nivel asignado L, son aquellos que se encuentran entre las dos raíces

$$\frac{S_2}{S_u} = \frac{\gamma \pm \sqrt{L(1-L)}(\sqrt{m_0}+\gamma^2/\sqrt{m_0})}{(L\gamma^2/m_0)-(1-L)} \tag{10.29}$$

donde $\gamma^2 = c_1/c_2$.

La Tabla 10.2, adaptada de Brooks (1955), muestra los límites superior e inferior de S_2^2/S_u^2 para $L = 0.9$. El amplio intervalo entre los límites inferior y superior es sorprendente en casi todos los casos. Nótese que la gama de valores de m_0 cambia en diferentes partes de la Tabla.

Si tenemos una idea aproximada de los valores de S_2^2/S_u^2 para las características principales en un reconocimiento, la Tabla 10.2 puede usarse para seleccionar un valor de m_0. Nótese que si ρ es la correlación entre elementos en la misma unidad primaria, según se definió en la Sec. 9.4, la razón S_2^2/S_u^2 es casi igual a $(1-\rho)/\rho$. Un valor de S_2^2/S_u^2 tan bajo como 1 corresponde a $\rho = 0.5$. Este sería un grado de correlación intraunidad inusitadamente alto. En forma similar, $\rho = 0.1$ da $S_2^2/S_u^2 = 9$, mientras $\rho = 0.01$ da $S_2^2/S_u^2 = 99$.

Ejemplo. Suponga que c_1/c_2 es casi 1 y que S_2^2/S_u^2 se espera entre 5 y 100 para las características principales. Las columnas $c_1/c_2 = 1$ dan $m_0 = 4$ como una elección satisfactoria, ya que ésta cubre las razones de 4 hasta más de 100 (realmente hasta 196). Con $c_1/c_2 = 16$ y la misma gama deseada, la tabla sugiere un valor de m_0 poco más o menos entre 15 y 20. Cálculos ulteriores a partir de (10.29) muestran que $m_0 = 18$ es mejor. Esta cubre la gama de 5.2 a 84, no tan amplia como se deseaba.

Cuando el costo de viajar entre las unidades primarias es sustancial, una función de costo más exacta puede ser

$$C = c_1n + c_t\sqrt{n} + c_2nm \tag{10.30}$$

TABLA 10.2. LÍMITES PARA S_2^2/S_u^2 DENTRO DE LOS CUALES m_0 DA AL MENOS 90% DE LA PRECISIÓN MÁXIMA

$c_1/c_2 =$	$\frac{1}{2}$		1		$c_1/c_2 =$	2		4	
m_0	L	U	L	U	m_0	L	U	L	U
1	0.0	11	0.0	4	2	0.5	8	0.2	4
2	2.0	98	1.1	22	3	1.2	21	0.5	8
3	4.1	>*	2.4	72	4	2.2	44	1.0	16
4	6.6	>	4.0	>	5	3.3	82	1.6	27
5	9.5	>	5.9	>	6	4.7	>	2.4	42
6	13	>	8.1	>	7	6.3	>	3.3	61
7	16	>	11	>	8	8.0	>	4.3	87
8	20	>	13	>	9	10	>	5.4	>
$c_1/c_2 =$	8		16		$c_1/c_2 =$	32		64	
m_0	L	U	L	U	m_0	L	U	L	U
6	1.0	17	0.3	8	5	0.1	3	0.0	2
7	1.5	24	0.5	11	10	0.4	12	0.1	7
8	2.0	32	0.7	15	15	1.2	26	0.3	14
9	2.6	42	1.0	19	20	2.7	46	0.7	24
10	3.3	53	1.3	23	25	4.5	74	1.5	37
15	7.6	>	3.5	55	30	6.9	>	2.5	52
20	13.3	>	6.6	>	35	9.7	>	3.7	71
25	20.4	>	10.5	>	40	13	>	5.2	93

* > denota "> 100."

ya que los costos de viaje tienden a ser proporcionales a $\sqrt{n}$. Si se ha especificado un valor deseado de $V(\bar{y})$, pares de valores de (n, m) que dan esta varianza se calculan fácilmente a partir de (10.25) para $V(\bar{y})$. Los costos para combinaciones diferentes se calculan entonces a partir de (10.30) y se encuentra la combinación que dé el menor costo. Cuando el costo se ha fijado con anterioridad, Hansen, Hurwitz y Madow (1953) dan un método para determinar la combinación (n, m) que minimiza la varianza y una tabla que facilita una elección rápida. Nótese que su n es nuestra m, y viceversa.

10.7 ESTIMACION DE m_{opt} DE UN RECONOCIMIENTO PILOTO

En ocasiones las estimaciones de S_2^2 y S_1^2 se obtienen de un reconocimiento piloto en el cual se escogen n' unidades prima-

rias, con m' elementos tomados de cada unidad. Esta sección trata de la elección de n' y m'. Si s_1^2 es la varianza entre medias de las unidades y s_2^2 es la varianza entre elementos dentro de las unidades, según se definió en la Sec. 10.4, (10.22) da

$$E(s_1^2) = \left(S_1^2 - \frac{S_2^2}{M}\right) + \frac{S_2^2}{m'} = S_u^2 + \frac{S_2^2}{m'} \tag{10.31}$$

Para la función simple de costo $c_1 n + c_2 nm$, teníamos

$$m_{\acute{o}pt} = \frac{S_2}{\sqrt{S_1^2 - S_2^2/M}} \sqrt{c_1/c_2}$$

Como una estimación de $m_{\acute{o}pt}$ a partir de este reconocimiento piloto, (10.31) sugiere que tomemos

$$\hat{m}_{\acute{o}pt} = \frac{s_2}{\sqrt{s_1^2 - s_2^2/m'}} \sqrt{c_1/c_2} = \frac{\sqrt{m'}}{\sqrt{(m' s_1^2/s_2^2) - 1}} \sqrt{c_1/c_2} \tag{10.32}$$

La estimación $\hat{m}_{\acute{o}pt}$ está sujeta a un error de muestreo que depende del error de muestreo de la razón s_1^2/s_2^2. Del análisis de la varianza se sabe que $m' s_1^2/s_2^2$ está distribuida como

$$F\left(1 + m' \frac{S_u^2}{S_2^2}\right)$$

donde F tiene $(n' - 1)$ y $n'(m' - 1)$ grados de libertad, siempre que las y_{ij} estén normalmente distribuidas. Este resultado conduce a la distribución de muestra de $m_{\acute{o}pt}$ para valores dados de n' y m', esto es,

$$\hat{m}_{\acute{o}pt} = \frac{\sqrt{m' c_1/c_2}}{\sqrt{F\left(1 + \frac{m' S_u^2}{S_2^2}\right) - 1}} \tag{10.33}$$

Ejemplo. Para el ejemplo en la Sec. 10.6, en el cual

$$c_1 = 10 c_2, \qquad S_2 = 1.3 S_u, \qquad m_{\acute{o}pt} = 1.3\sqrt{10} = 4.1$$

considérese qué tan bien se estima $m_{\acute{o}pt}$ a partir de una muestra piloto con $n' = 10$ y $m' = 4$. De (10.33)

$$\hat{m}_{\acute{o}pt} = \frac{6.324}{\sqrt{F[1 + (4/1.69)] - 1}} = \frac{6.324}{\sqrt{3.367F - 1}}$$

donde F tiene 9 y 30 gl. Para encontrar los límites entre los cuales $\hat{m}_{\acute{o}pt}$ se situará el 80% del tiempo, tenemos, considerando los valores de significación al 10% para F en un solo extremo,

$$F_{.10}(9, 30) = 1.8490, \qquad F_{.90}(9, 30) = 1/F_{.10}(30, 9) = 1/2.2547 = 0.4435$$

Sustituyendo estos valores de F se obtiene

límite inferior $\hat{m}_{\acute{o}pt} = 2.8$; límite superior $\hat{m}_{\acute{o}pt} = 9.0$

Como se muestra en la Tabla 10.1, cualquier m en esta gama da un grado de precisión cercano al óptimo. Así con $n' = 10$, $m' = 4$, hay 8 oportunidades de 10 que la pérdida de precisión sea pequeña con datos normales.

Los límites del 80% y 95% para $n' = 5$, 10, 20 y $m' = 4$ son los que aparecen en la Tabla 10.3. Con $n' = 20$ casi hay seguridad de que estimar $m_{\acute{o}pt}$ con precisión cercana al óptimo. Esto no es así con $n' = 5$.

TABLA 10.3. LÍMITES SUPERIORES E INFERIORES PARA $\hat{m}_{\acute{o}pt}$

n'	80%	90%
5	2.5, ∞	1.8, ∞
10	2.8, 9.0	2.3, ∞
20	3.1, 6.4	2.7, 9.1

Si la razón c_1/c_2 es la misma en la encuesta piloto que en la encuesta principal el costo de la encuesta piloto será proporcional a $c_1 n' + c_2 n' m'$. Brooks (1955) da una tabla de valores de (n', m') en la encuesta piloto más económica que proporciona una precisión relativa de 90% en la estimación de $m_{\acute{o}pt}$. La Tabla 10.4 es parte de dicha tabla.

TABLA 10.4. DISEÑOS DE MUESTRAS PILOTO QUE TIENEN UNA PRECISIÓN RELATIVA ESPERADA DE 90%

c_1/c_2	≤ 1		2		4		8		16		32		64	
S_2^2/S_u^2	n'	m'	n'	m'	n'	m'	n'	m'	n'	m'	n'	m'	n'	m'
1	7	3	6	4	6	5	5	6	5	7	4	10	4	12
2	8	5	7	7	6	9	6	9	5	13	5	14	4	20
4	9	9	8	11	8	12	7	14	7	15	5	25	5	27
8	10	14	10	15	9	17	9	18	8	22	6	32	5	44
16	10	25	10	27	10	27	10	28	8	37	7	46	6	60
32	10	46	10	47	10	48	10	49	9	58	8	69	6	102
64	10	92	10	93	10	96	10	100	10	104	8	137	7	169

Los cálculos suponen N, M elevadas: los diseños son conservadores si los términos de cpf se toman en cuenta. Nótese que no más de 10 unidades primarias son las requeridas y que los diseños no presentan sensibilidad, relativamente, respecto a la razón c_1/c_2.

10.8 MUESTREO DE TRES ETAPAS

El proceso de submuestreo se puede llevar a una tercera etapa al muestrear las subunidades (elementos) en lugar de enumerarlas completamente. Por ejemplo, en reconocimientos para estimar la producción de cosechas en la India. (Sukhatme, 1947), la aldea es una unidad de muestreo conveniente. Dentro de una aldea, solamente se seleccionan los campos cultivados con la cosecha en cuestión, así que el campo es la subunidad. Cuando se selecciona un campo, solamente ciertas partes de él son recolectadas para la determinación del rendimiento por acre; por lo tanto, se muestrea la subunidad en sí misma. Si están involucrados análisis físicos o químicos de la cosecha, puede usarse un submuestreo adicional, ya que estas determinaciones se hacen a menudo en una parte de la muestra recolectada en un campo.

Los resultados son una extensión directa de los del muestreo en dos etapas y se dan brevemente. La población contiene N unidades de primera etapa, cada una con M unidades de segunda etapa, cada una de las cuales tiene K unidades de tercera etapa. Los números correspondientes a la muestra son n, m y k, respectivamente. Sea y_{iju} el valor obtenido para la u-ésima unidad de tercera etapa en la j-ésima unidad de segunda etapa sacada de la i-ésima unidad primaria. Las medias de población relevantes por unidad de tercera etapa son como sigue:

$$\bar{Y}_{ij}=\frac{\sum_{u}^{K} y_{iju}}{K}, \qquad \bar{\bar{Y}}_i=\frac{\sum_{j}^{M}\sum_{u}^{K} y_{iju}}{MK}, \qquad \bar{\bar{\bar{Y}}}=\frac{\sum_{i}^{N}\sum_{j}^{M}\sum_{u}^{K} y_{iju}}{NMK}$$

Se requieren las siguientes varianzas de población:

$$S_1^2=\frac{\sum_{i}^{N}(\bar{\bar{Y}}_i-\bar{\bar{\bar{Y}}})^2}{N-1}$$

$$S_2^2=\frac{\sum_{i}^{N}\sum_{j}^{M}(\bar{Y}_{ij}-\bar{\bar{Y}}_i)^2}{N(M-1)}$$

$$S_3^2=\frac{\sum_{i}^{N}\sum_{j}^{M}\sum_{u}^{K}(y_{ijk}-\bar{Y}_{ij})^2}{NM(K-1)}$$

Teorema 10.3. Si se usa muestreo aleatorio simple en las tres etapas, la media de la muestra por unidad de tercera etapa $\bar{\bar{\bar{y}}}$ es estimación insesgada de $\bar{\bar{\bar{Y}}}$ con varianza

$$V(\bar{\bar{\bar{y}}})=\frac{1-f_1}{n}S_1^2+\frac{1-f_2}{nm}S_2^2+\frac{1-f_3}{nmk}S_3^2 \tag{10.34}$$

donde $f_1 = n/N$, $f_2 = m/M$, $f_3 = k/K$ son las fracciones de muestreo en las tres etapas.

Prueba. Solamente se indican los pasos principales. Escríbase

$$\bar{\bar{\bar{y}}}-\bar{\bar{\bar{Y}}}=(\bar{\bar{\bar{y}}}-\bar{\bar{\bar{Y}}}_{nm})+(\bar{\bar{\bar{Y}}}_{nm}-\bar{\bar{\bar{Y}}}_n)+(\bar{\bar{\bar{Y}}}_n-\bar{\bar{\bar{Y}}}) \tag{10.35}$$

donde $\bar{\bar{\bar{Y}}}_{nm}$ es la media de población de las nm unidades de segunda etapa que fueron seleccionadas y $\bar{\bar{\bar{Y}}}_n$ es la media de población de las n unidades primarias que fueron seleccionadas. Cuando elevamos al cuadrado y sacamos el promedio, los términos con productos cruzados desaparecen. Las contribuciones de los términos elevados al cuadrado resultan ser como sigue:

$$E(\bar{\bar{\bar{y}}}-\bar{\bar{\bar{Y}}}_{nm})^2=\frac{1-f_3}{nmk}S_3^2$$

$$E(\bar{\bar{\bar{Y}}}_{nm}-\bar{\bar{\bar{Y}}}_n)^2=\frac{1-f_2}{nm}S_2^2$$

$$E(\bar{\bar{\bar{Y}}}_n-\bar{\bar{\bar{Y}}})^2=\frac{1-f_1}{n}S_1^2$$

Al sumar estos términos, se obtiene el teorema.

Teorema 10.4. Una estimación insesgada de $V(\bar{\bar{\bar{y}}})$ a partir de la muestra es

$$v(\bar{\bar{\bar{y}}})=\frac{1-f_1}{n}s_1^2+\frac{f_1(1-f_2)}{nm}s_2^2+\frac{f_1f_2(1-f_3)}{nmk}s_3^2 \tag{10.36}$$

donde s_1^2, s_2^2, s_3^2 son en la muestra los análogos de S_1^2, S_2^2, S_3^2, respectivamente.

Prueba. Esto se puede probar mediante los métodos de la Sec. 10.4 o alternativamente, al demostrar que

$$E(s_1^2)=S_1^2+\frac{1-f_2}{m}S_2^2+\frac{1-f_3}{mk}S_3^2 \tag{10.37}$$

$$E(s_2^2)=S_2^2+\frac{1-f_3}{k}S_3^2$$

y $E(s_3^2) = S_3^2$. Para obtener el primer resultado, sea $\bar{y}_{iK}$ la media sobre todas las m unidades de segunda etapa en la i-ésima unidad primaria, al suponer que todos los K elementos fueron enumerados durante la tercera etapa. Sea $\bar{\bar{y}}_K$ la media de los n valores $\bar{y}_{iK}$. Entonces, de (10.22) para el muestreo en dos etapas, se tiene que

$$E\left[\frac{\sum^{n}(\bar{y}_{iK}-\bar{\bar{y}}_K)^2}{n-1}\right]=S_1^2=\frac{1-f_2}{m}S_2^2 \tag{10.38}$$

Ahora, si $\bar{y}_i$ es la media de la muestra para la i-ésima unidad primaria, escríbase

$$(\bar{y}_i-\bar{\bar{y}})=(\bar{y}_{iK}-\bar{\bar{y}}_K)+[(\bar{y}_i-\bar{y}_{iK})-(\bar{\bar{y}}-\bar{\bar{y}}_K)] \tag{10.39}$$

Al promediar primero sobre muestras en las cuales las unidades de primera y segunda etapa son fijadas, se puede mostrar fácilmente que

$$\frac{1}{(n-1)}E\sum^{n}[(\bar{y}_i-\bar{y}_{iK})-(\bar{\bar{y}}-\bar{\bar{y}}_K)]^2=\frac{(1-f_3)S_3^2}{mk} \tag{10.40}$$

y que la suma de términos con productos cruzados (10.39) en nada contribuye. Esto establece el resultado para $E(s_1^2)$. El correspondiente a $E(s_2^2)$ se encuentra en forma similar. Por lo tanto

$$\begin{aligned}E[v(\bar{\bar{y}})]&=\frac{1-f_1}{n}\left(S_1^2+\frac{1-f_2}{m}S_2^2+\frac{1-f_3}{mk}S_3^2\right)\\&\quad+\frac{f_1(1-f_2)}{nm}\left(S_2^2+\frac{1-f_3}{k}S_3^2\right)+\frac{f_1f_2(1-f_3)}{nmk}S_3^2\\&=\frac{1-f_1}{n}S_1^2+\frac{1-f_2}{nm}S_2^2+\frac{1-f_3}{nmk}S_3^2=V(y)\end{aligned} \tag{10.41}$$

Lo mismo que con el muestreo en dos etapas, es claro a partir de (10.36) que si f_1 es despreciable $v(\bar{\bar{y}})$ se reduce a

$$v(\bar{\bar{y}})=\frac{s_1^2}{n}=\frac{\sum^{n}(\bar{y}_i-\bar{\bar{y}})^2}{n(n-1)} \tag{10.42}$$

Esta es una estimación conservadora si f_1 no es despreciable.

Con una función de costo de la forma

$$C=c_1n+c_2nm+c_3nmk \tag{10.43}$$

los valores óptimos de k y m son

$$k_{opt}=\frac{S_3}{\sqrt{S_2^2-S_3^2/K}}\sqrt{c_2/c_3},\qquad m_{opt}=\frac{\sqrt{S_2^2-S_3^2/K}}{\sqrt{S_1^2-S_2^2/M}}\sqrt{c_1/c_2} \tag{10.44}$$

La extensión de los resultados en esta sección, a etapas adicionales de muestreo, debería ser clara a partir de la estructura de las fórmulas.

10.9 MUESTREO ESTRATIFICADO DE LAS UNIDADES

El submuestreo puede combinarse con cualquier tipo de muestreo de las unidades primarias. El submuestreo en sí mismo puede emplear estratificación o muestreo sistemático. Las fórmulas para las varianzas para estas modificaciones se pueden construir a partir de las fórmulas para los métodos más simples.

Los resultados que se presentan son para muestreo estratificado de las unidades primarias en una muestra de dos etapas. Los tamaños de las unidades primarias se suponen constantes para un estrato dado pero pueden variar de estrato a estrato. Esta situación ocurre cuando las unidades primarias se estratifican por tamaño, de tal manera que los tamaños dentro de un estrato se vuelven constantes o casi.

El h-ésimo estrato N_h unidades primarias, cada una con M_h unidades de la segunda etapa; los números correspondientes en la muestra son n_h y m_h. La media de población estimada por unidad de la segunda etapa es

$$\bar{\bar{y}}_{st}=\frac{\sum_h N_hM_h\bar{\bar{y}}_h}{\sum_h N_hM_h}=\sum_h W_h\bar{\bar{y}}_h \tag{10.45}$$

donde $W_h = N_hM_h/\Sigma N_hM_h$ es el tamaño relativo del estrato en términos de *unidades de la segunda etapa* y $\bar{\bar{y}}_h$ es la media de la muestra en el estrato. Al aplicar el Teorema 10.1 dentro de cada estrato, tenemos

$$V(\bar{\bar{y}}_{st})=\sum_h W_h^2\left(\frac{1-f_{1h}}{n_h}S_{1h}^2+\frac{1-f_{2h}}{n_hm_n}S_{2h}^2\right) \tag{10.46}$$

donde $f_{1h} = n_h/N_h$, $f_{2h} = m_h/M_h$.

A partir del Teorema 10.2, una estimación insesgada de la muestra es

$$v(\bar{\bar{y}}_{st})=\sum_h W_h^2\left[\frac{1-f_{1h}}{n_h}s_{1h}^2+\frac{f_{1h}(1-f_{2h})}{n_hm_h}s_{2h}^2\right] \tag{10.47}$$

Las varianzas correspondientes para el total de población estimado se obtienen multiplicando las Fórmulas (10.46) y (10.47) por $(\Sigma N_hM_h)^2$.

10.10 ASIGNACION OPTIMA CON MUESTREO ESTRATIFICADO

Esto trata sobre la mejor elección de los n_h y los m_h. Si los costos de viaje entre unidades no son el factor principal, el costo se puede representar adecuadamente por la fórmula

$$C=\sum_h c_{1h}n_h+\sum_h c_{2h}n_h m_h \tag{10.48}$$

A partir de (10.46), la varianza puede escribirse como

$$V(\bar{y}_{st})=\sum_h W_h{}^2\left[\frac{1}{n_h}\left(S_{1h}{}^2-\frac{S_{2h}{}^2}{M_h}\right)+\frac{1}{n_h m_h}S_{2h}{}^2-\frac{1}{N_h}S_{1h}{}^2\right]$$

La cantidad

$$V(\bar{y}_{st})+\lambda\left(\sum_h c_{1h}n_h+\sum_h c_{2h}n_h m_h-C\right)$$

donde λ es un multiplicador de Lagrange, es una función de las variables n_h y $(n_h m_h)$. Por lo tanto, para minimizar V para C fija, o viceversa, tenemos

$$n_h\sqrt{\lambda}=\frac{W_h}{\sqrt{c_{1h}}}\sqrt{S_{1h}{}^2-S_{2h}{}^2/M_h} \tag{10.49}$$

$$n_h m_h\sqrt{\lambda}=\frac{W_h S_{2h}}{\sqrt{c_{2h}}} \tag{10.50}$$

Esto da

$$m_h=\frac{S_{2h}}{\sqrt{S_{1h}{}^2-S_{2h}{}^2/M_h}}\sqrt{c_{1h}/c_{2h}} \tag{10.51}$$

La fórmula para m_h óptimo es exactamente la misma que en el muestreo no estratificado [(10.26) en la Sec. 10.6].

De (10.49), dado que $W_h \propto N_h M_h$

$$n_h \propto \frac{N_h M_h S_{uh}}{\sqrt{c_{1h}}} \qquad \text{donde } S_{uh}{}^2=S_{1h}{}^2-\frac{S_{2h}{}^2}{M_h} \tag{10.52}$$

Ya que son convenientes estimaciones autoponderantes, consideraremos en qué circunstancias la asignación óptima conduce a una estimación autoponderante. De (10.45) se desprende que $\bar{y}_{st}$ es autoponderante si $n_h m_h/N_h M_h = f_0 =$ constante, ya que en este caso

$$\bar{\bar{y}}_{st}=\frac{\sum_h (N_h M_h/n_h m_h)\sum_i^{n_h}\sum_j^{m_h} y_{hij}}{\sum N_h M_h}=\frac{\sum_h\sum_i\sum_j y_{hij}}{f_0\sum N_h M_h}$$

$$=\frac{\sum_h\sum_i\sum_j y_{hij}}{\sum n_h m_h}=\bar{\bar{y}}$$

La condición es, como puede esperarse, que la fracción de muestreo completa f_0 sea la misma en todos los estratos.

De (10.50), la asignación óptima da

$$f_{0h}=\frac{n_h m_h}{N_h M_h}\propto\frac{S_{2h}}{\sqrt{c_{2h}}} \tag{10.53}$$

Frecuentemente c_{2h}, el costo por unidad de la segunda etapa, será aproximadamente el mismo en unidades primarias grandes y pequeñas; pero S_{2h} puede ser mayor en las unidades grandes que en las pequeñas. Sin embargo, dado que el óptimo es achatado, una muestra autoponderante será a menudo tan precisa como el óptimo. Nótese que este resultado se mantiene aun si el muestreo óptimo de las unidades primarias está lejos de ser proporcional.

EJERCICIOS

10.1. Un grupo de 20 000 registros se guarda en 400 gaveteros de archivo, con 50 registros cada uno. En una muestra en dos etapas, se extraen cinco registros al azar de cada uno de 80 gaveteros seleccionados al azar. Para una característica, las estimaciones de la varianza fueron $s_1{}^2 = 362$, $s_2{}^2 = 805$, según se definen en la Sec. 10.4. (*a*) Calcule el error estándar de la media por registro de esta muestra. (*b*) Compare éste con el error estándar dado por la fórmula aproximada (10.23) en la Sec. 10.4.

10.2. De los resultados de una muestra piloto en dos etapas, en la cual se escogieron m' subunidades de cada una de n' unidades, es útil estimar el valor de $V(\bar{y})$ que estaría dado por una muestra subsecuente que contiene m subunidades de cada una de n unidades. Muestre que una estimación insesgada de $V(\bar{y})$ es

$$\hat{V}(\bar{y})=\left(\frac{N-n}{N}\right)\frac{s_1{}^2}{n}+\frac{s_2{}^2}{mn}\left(1-\frac{m}{m'}+\frac{mn}{m'N}-\frac{mn}{MN}\right)$$

donde $s_1{}^2$ y $s_2{}^2$ se calculan a partir de una muestra preliminar. *Indicación.* Use el Teorema 10.1 y el resultado (10.22):

$$E(s_1{}^2)=S_1{}^2-\frac{S_2{}^2}{M}+\frac{S_2{}^2}{m}$$

10.3. Al muestrear campos de trigo en Kansas, con el campo como unidad primaria, King y McCarty (1941) reportaron los siguientes cuadrados medios para rendimiento en bushels por acre: $s_1{}^2 = 165$, $s_2{}^2 = 66$. Se tomaron dos submuestras por campo. Para una muestra de n campos, compare las varianzas de la media de la muestra dada por (*a*) la muestra tal y cual fue tomada, (*b*) cuatro submuestras por campo de n campos. (*c*) Al cosechar enteramente n campos.

N y M se pueden suponer grandes y constantes. En (c) suponga que la cosecha completa equivale a muestreo en una sola etapa (es decir, a tener $m = M$).

10.4. En la misma encuesta, con dos submuestras por campo, los cuadrados medios para el porcentaje de proteína fueron $s_1^2 = 7.73$, $s_2^2 = 1.43$. ¿Cuántos campos se requieren para estimar el rendimiento medio con aproximación de $\pm$ 1 bushel y el porcentaje medio de proteína con aproximación de $\pm\frac{1}{4}$ salvo en un caso de 20? Realice los cálculos suponiendo (a) que se toman dos submuestras por campo en la encuesta principal, (b) suponiendo una cosecha completa de un campo en la encuesta principal.

10.5. Para los datos sobre rendimientos de trigo en el Ej. 10.3, ¿cuál es el valor de c_1/c_2 en una función lineal de costo si la m óptima estimada es 2?

10.6. Si m/M y n/N son ambas pequeñas y la función de costo es lineal, muestre que $m = 2$ da un valor de $V(\bar{y})$ menor que $m = 1$ si

$$\frac{c_1}{c_2} > 2\frac{S_1^2}{S_2^2}$$

10.7. Un gran almacén maneja alrededor de 20 000 cuentas recibidas por mes. Una muestra del 2% ($m = 400$) se verificó cada mes durante 2 años ($n = 24$). Los números de cuentas erróneas que se encontraron por mes (de entre 400) fueron (en orden de magnitud) 0, 0, 1, 1, 2, 4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9, 9, 10, 10, 13, 14, 17, siendo errático el patrón de tiempo. A partir de los resultados en la Sec. 10.5, calcule s_1^2 y s_2^2. De aquí, calcule el error estándar de $\bar{p}$, como una estimación del porcentaje de cuentas erróneas por año, que se obtendría de la verificación de (a) 1 200 cuentas tomadas al azar, de un solo mes. (b) 300 cuentas de cada uno de cuatro meses al azar, (c) 100 cuentas cada mes. *Indicación.* Use la fórmula en el Ejercicio 10.2 con $m' = 400$ u obtenga estimaciones insesgadas de S_1^2 y S_2^2 y use el Teorema 10.1.

10.8. Al planear una encuesta en dos etapas, se esperaba que c_1/c_2 sería alrededor de 4 y que S_a^2/S_u^2 estaría entre 5 y 50. (a) ¿Cuál valor de m escogería de la Tabla 10.2? (b) Suponga que al terminar la encuesta se encontró que c_1/c_2 fue cercana a 8 y S_2^2/S_u^2 a 25. Calcule la precisión relativa dada por su m con la dada por la m óptima. (c) Haga los mismos cálculos para $c_1/c_2 = 4$, $S_2^2/S_u^2 = 100$.

10.9. Si ρ es el coeficiente de correlación entre unidades de la segunda etapa en la misma unidad primaria, pruebe que

$$\frac{1-\rho}{\rho} = \frac{S_2^2}{[(N-1)/N]S_1^2 - S_2^2/M} \doteq \frac{S_2^2}{S_u^2}$$

(Esto establece un resultado empleado en la Sec. 10.6.)

10.10. Muestre que si $S_u^2 > 0$ en la notación de la Sec. 10.6, una muestra aleatoria simple de n unidades primarias, con 1 elemento tomado por unidad, es más precisa que una muestra al azar de n elementos ($n > 1$, $M > 1$). Muestre que la precisión de los dos métodos es igual si n/N es despreciable. ¿Esperaría esto intuitivamente?

CAPITULO 11

El Submuestreo con Unidades de Diferentes Tamaños

11.1 INTRODUCCION

Al muestrear una población extensa, suelen encontrarse unidades primarias que varían en tamaño. Además, por consideraciones de costos se suele recomendar el uso del muestseo de varias etapas, de modo que los problemas discutidos en este capítulo ocurren con frecuencia. Si los tamaños no varían considerablemente, un método es el de estratificar por tamaños de unidades primarias, de modo que las unidades dentro de un estrato sean iguales o casi. Las fórmulas de la Sec. 10.9 pueden, entonces, ser una aproximación adecuada. Pero frecuentemente seguirán existiendo diferencias sustanciales de tamaño dentro de los estratos y en ocasiones será aconsejable basar la estratificación en otras variables. Gray y Corlett (1950) en una revisión de las encuestas sociales en Gran Bretaña, que son muestras a nivel nacional, con distritos como unidades primarias, señalan que el tamaño primero fue incluido como una de las variables para la estratificación, pero que al llegar a conocer mejor las características de la población, pareció más conveniente considerar otro factor.

Para adquirir un buen conocimiento práctico del muestreo de varias etapas, será necesario un esfuerzo concentrado cuando las unidades son de tamaños variables, porque la técnica es flexible a este respecto. Se pueden elegir las unidades con probabilidades iguales o bien, con probabilidades proporcionales al tamaño, o a alguna estimación del tamaño. Se han construido varias reglas para determinar las fracciones de muestreo y submuestreo y se dispone de varios métodos de estimación. Las ventajas de los diferentes métodos dependen de la naturaleza de la población, de los costos de campo, y de los datos adicionales de que disponemos.

La primera parte de este capítulo se dedica a la descripción de los métodos principales en uso. Empezaremos con una población que consta de un solo estrato. La extensión al muestreo estratificado es posible, como en los capítulos precedentes, al sumar las fórmulas de varianza apropiadas sobre los estratos. Para simplificar las cosas, suponemos primero que sólo se eligió una unidad primaria, es decir, que $n = 1$. Este caso no es tan impráctico como parece a primera vista, porque cuando hay un gran número de estratos, se puede alcanzar una precisión satisfactoria en la estimación aunque $n_h = 1$. En las encuestas mensuales del U.S. Census Bureau para estimar los números de personas empleadas, la unidad primaria es un municipio* o un grupo de municipios vecinos. Esta es una unidad grande, pero tiene ventajas administrativas que disminuyen los costos. Como los municipios están lejos de la uniformidad en cuanto a sus características, la estratificación se extiende hasta el punto en que sólo se selecciona uno de cada estrato. En consecuencia, la teoría que se va a discutir sólo se aplica a un solo estrato en este plan de muestreo.

Como en capítulos anteriores, las cantidades por estimar pueden ser el total de población Y, la media de población (usualmente la media por subunidad $\bar{Y}$), o la razón de dos variables.

Notación. La observación para la j-ésima subunidad perteneciente a la i-ésima unidad, se denota con y_{ij}. Los siguientes símbolos son referentes a la i-ésima unidad.

	Población	Muestra
Número de subunidades	M_i	m_i
Media por subunidad	$\bar{Y}_i$	$\bar{y}_i$
Total	$Y_i = M_i\bar{Y}_i$	$y_i = m_i\bar{y}_i$

Los siguientes símbolos se refieren a la totalidad de la población o de la muestra.

	Población	Muestra
Número de subunidades	$M_0 = \sum^N M_i$	$\sum^n m_i$
Total	$Y = \sum^N Y_i$	$\sum^n y_i$
Media por subunidad	$\bar{Y} = Y/M_0$	$\bar{y} = \sum y_i / \sum m_i$
Media por unidad primaria	$\bar{Y} = Y/N$	$\bar{y} = \sum y_i / n$

* Condado, en el original. (N. del T.)

11.2 METODOS DE MUESTREO CUANDO $n = 1$

Supongamos que la i-ésima unidad se seleccionó conteniendo M_i subunidades, de las cuales m_i se escogen al azar. Consideramos tres métodos de estimación de $\bar{Y}$, la media por subunidad, o *unidad de la segunda etapa*, como suele llamarse.

1. Unidades elegidas con probabilidad igual

$$\text{Estimación} = \bar{y}_I = \bar{y}_i.$$

La estimación es la media de muestra por subunidad. Es sesgada porque en muestreo repetido de la misma unidad, el promedio de $\bar{y}_i$ es $\bar{Y}_i$, y como cada unidad tiene la misma oportunidad de selección, el promedio de $\bar{Y}_i$ es

$$\frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i = \bar{Y}_a \quad \text{(digamos)}$$

Pero la media de población es

$$\bar{Y} = \frac{\sum_{i=1}^{N} M_i\bar{Y}_i}{M}, \qquad \text{donde } M = \sum_{i=1}^{N} M_i$$

Por tanto, el sesgo es igual a $(\bar{Y}_a - \bar{Y})$. Dado que el método es sesgado, vamos a calcular el error cuadrático medio (ECM) respecto a $\bar{Y}$. Escribamos

$$\bar{y}_i - \bar{Y} = (\bar{y}_i - \bar{Y}_i) + (\bar{Y}_i - \bar{Y}_a) + (\bar{Y}_a - \bar{Y})$$

Al elevar al cuadrado y tomar la esperanza matemática sobre todas las muestras posibles; todas las contribuciones de términos cruzados se anulan. Las esperanzas de los términos cuadrados se obtienen fácilmente por los métodos dados en el Cap. 10. Encontramos

$$\text{ECM}(\bar{y}_I) = \frac{1}{N}\sum_{i=1}^{N}\frac{(M_i - m_i)}{M_i}\frac{S_{2i}^2}{m_i} + \frac{1}{N}\sum_{i=1}^{N}(\bar{Y}_i - \bar{Y}_a)^2 + (\bar{Y}_a + \bar{Y})^2 \qquad (11.1)$$

donde

$$S_{2i}^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y}_i)^2$$

es la varianza entre subunidades de la i-ésima unidad.

El ECM de $\bar{y}_I$ contiene tres componentes: uno surge de la variación entre unidades, otro de la variación entre verdaderas medias de las unidades, y otro más del sesgo.

Los valores de los m_i no se han especificado. La elección más común consiste en tomar todos los tamaños m_i iguales, o tomar m_i proporcional a M_i, es decir, submuestrear una proporción fija de cualquier unidad seleccionada. La elección de los m_i sólo afecta la primera de las tres componentes de la varianza, a saber, la componente que proviene de la variación dentro de las unidades.

II. Unidades elegidas con igual probabilidad

$$\text{Estimación} = \bar{y}_{\text{II}} = \frac{NM_i\bar{y}_i}{M_0}$$

Esta estimación es *insesgada*. Dado que $\bar{y}_i$ es una estimación insesgada de $\bar{Y}_i$, el producto $M_i\bar{y}_i$ es una estimación insesgada del total de unidad Y_i. Por tanto, $NM_i\bar{y}_i$ es una estimación insesgada del total de población Y, al dividir por M_0, el número total de subunidades en la población, obtenemos una estimación insesgada de $\bar{Y}$.

Para encontrar $V(\bar{y}_{\text{II}})$ que desde luego es igual a su ECM, tenemos

$$\bar{y}_{\text{II}} - \bar{Y} = \frac{NM_i\bar{y}_i}{M_0} - \bar{Y}$$

$$= \frac{NM_i}{M_0}(\bar{y}_i - \bar{Y}_i) + \left(\frac{NM_i}{M_0}\bar{Y}_i - \bar{Y}\right)$$

Ahora $M_i\bar{Y}_i = Y_i$, el total para la unidad, y $\bar{Y} = N\bar{Y}/M_0$, donde $\bar{Y}$ es la media de población por unidad. Esto da

$$\bar{y}_{\text{II}} - \bar{Y} = \frac{NM_i}{M_0}(\bar{y}_i - \bar{Y}_i) + \frac{N}{M_0}(Y_i - \bar{Y})$$

Por lo tanto,

$$V(\bar{y}_{\text{II}}) = \frac{N}{M_0{}^2}\sum_{i=1}^{N} M_i(M_i - m_i)\frac{S_{2i}{}^2}{m_i} + \frac{N}{M_0{}^2}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \qquad (11.2)$$

La componente entre unidades de esta varianza (segundo término a la derecha) representa la variación entre los *totales* de unidades Y_i. Esta componente está afectada tanto por las variaciones en las M_i de unidad a unidad, como por las variaciones en las medias $\bar{Y}_i$ por elemento. Si las unidades varían considerablemente en tamaño, esta componente es grande aunque las medias por elemento $\bar{Y}_i$ son casi constantes de unidad a unidad. Con frecuencia esta componente es tan grande que $\bar{y}_{\text{II}}$ tienen un ECM mucho más alto que la estimación con sesgo $\bar{y}_{\text{I}}$. Entonces, ni el Método I ni el Método II son enteramente satisfactorios.

III. Unidades escogidas con probabilidad proporcional al tamaño

$$\text{Estimación} = \bar{y}_{\text{III}} = \bar{y}_i = \text{media de la muestra}$$

Esta técnica se debe a Hansen y Hurwitz (1943), y provee una media de la muestra que no tiene sesgo y no está sujeta a la inflación de la varianza en el Método II.

En muestreo repetido, la i-ésima unidad aparece con frecuencia relativa M_i/M_0. Por lo tanto,

$$E(\bar{y}_{\text{III}}) = \sum_{i=1}^{N}\frac{M_i}{M_0}\bar{Y}_i = \bar{Y}$$

Además,

$$\bar{y}_{\text{III}} - \bar{Y} = (\bar{y}_{\text{III}} - \bar{Y}_i) + (\bar{Y}_i - \bar{Y})$$

Promediemos primero sobre muestras en las cuales se selecciona la i-ésima unidad.

$$E_i(\bar{y}_{\text{III}} - \bar{Y})^2 = \left(\frac{M_i - m_i}{M_i}\right)\frac{S_{2i}{}^2}{m_i} + (\bar{Y}_i - \bar{Y})^2$$

Ahora promediemos sobre todas las selecciones posibles de la unidad. Dado que la i-ésima unidad se selecciona con frecuencia relativa M_i/M_0,

$$V(\bar{y}_{\text{III}}) = \frac{1}{M_0}\left[\sum_{i=1}^{N}(M_i - m_i)\frac{S_{2i}{}^2}{m_i} + \sum_{i=1}^{N}M_i(\bar{Y}_i - \bar{Y})^2\right] \qquad (11.3)$$

Nótese que, como en el Método I, la componente entre-unidades resulta de diferencias entre las medias $\bar{Y}_i$ por elemento en las unidades sucesivas. Si estas medias por elemento son casi iguales, esta componente es pequeña.

Ejemplo. Apliquemos estos resultados a una población pequeña, artificialmente construida. Los datos se presentan en la Tabla 11.1. Hay tres unidades con 2, 4 y 6 elementos, respectivamente. Se pueden verificar las cifras dadas para Y_i, $S_{2i}{}^2$, y $\bar{Y}_i$. La media $\bar{Y}$ de la población es $33/12$, o 2.75. La media sin

TABLA 11.1. POBLACIÓN ARTIFICIAL CON UNIDADES DE TAMAÑOS DESIGUALES

Unidad	y_{ij}	M_i	Y_i	$S_{2i}{}^2$	$\bar{Y}_i$	$\bar{Y}_i - \bar{\bar{Y}}$
1	0, 1	2	1	0.500	0.5	−2.25
2	1, 2, 2, 3	4	8	0.667	2.0	−0.75
3	3, 3, 4, 4, 5, 5	6	24	0.800	4.0	+1.25
	Totales	12	33			

ponderar de las $\bar{Y}_i$ es 2.167 = $\bar{Y}_a$, de tal manera que el sesgo en el Método I es — 0.583. Su cuadrado, la contribución al ECM, es 0.340.

Una unidad se selecciona y dos elementos muestreadcs de ella. Consideraremos cuatro métodos, dos de los cuales son variantes del Método I.

Método Ia.
Selección: unidad con probabilidad igual, $m_i = 2$.
Estimación: $\bar{y}_i$ (con sesgo).

Método Ib.
Selección: unidad con probabilidad igual $m_i = \frac{1}{2}M_i$.
Estimación: $\bar{y}_i$ (con sesgo).

Método II.
Selección: unidad con probabilidad igual, $m_i = 2$.
Estimación: $NM_i\bar{y}_i/M_0$ (insesgada).

Método III.
Selección: unidad con probabilidad M_i/M_0, $m_i = 2$.
Estimación: $\bar{y}_i$ (insesgada).

El Método *Ib* (submuestreo proporcional) no garantiza un tamaño de 2 (éste puede ser 1, 2 o 3), pero el tamaño promedio de la muestra es 2.

Al aplicar las fórmulas para el error de muestreo (11.1), (11.2) y (11.3) obtenemos los resultados de la Tabla 11.2.

TABLA 11.2. ECM DE LAS ESTIMACIONES MUESTRALES DE $\bar{Y}$

Método	Contribución al ECM de: Dentro de unidades	Entre unidades	Sesgo	ECM Total
I*a*	0.144	2.056	0.340	2.540
I*b*	0.183	2.056	0.340	2.579
II	0.256	5.792	0.000	6.048
III	0.189	1.813	0.000	2.002

Aunque el ejemplo es artificial, los resultados son característicos de los que se encuentran en comparaciones hechas en muchas poblaciones. El Método III da el ECM más pequeño porque tiene la contribución más pequeña de la variación entre unidades. El Método II, aunque sin sesgo, es muy inferior. El Método I*a* (tamaño igual de la submuestra) es ligeramente mejor que el Método *Ib* (submuestreo proporcional).

También se han hecho algunas comparaciones entre estos métodos en poblaciones reales. Para seis características (total de trabajadores, total de trabajadores agrícolas, total de trabajadores no agrícolas, estimados separadamente para hombres y mujeres), Hansen y Hurwitz (1943) encontraron que el Método III produjo grandes

reducciones en la contribución de la variación entre unidades, en comparación con el Método sin sesgo II, y reducciones que promediaron 30% en comparación con el Método I. (Ellos supusieron como despreciable la contribución de la variación dentro de unidades.) En la estimación de características típicas de instalaciones agrícolas para el Estado de Carolina del Norte, Jebe (1952) reportó reducciones del orden del 15% en la varianza total en comparación con métodos del tipo I. En ambos estudios la unidad primaria fue un condado.

11.3 MUESTREO CON PROBABILIDAD PROPORCIONAL AL TAMAÑO ESTIMADO

Como se mencionó en el Cap. 9, los tamaños M_i de las unidades apenas se conocen aproximadamente con base en datos anteriores y en otras encuestas pueden tenerse varias otras medidas posibles disponibles del tamaño de una unidad. Sea z_i la probabilidad o tamaño relativo asignado a la i-ésima unidad, donde z_i son cualquier conjunto de números positivos que suman la unidad. Todavía suponemos $n = 1$.

Método IV Una estimación insesgada de $\bar{Y}$ es

$$\bar{y}_{IV} = \frac{M_i\bar{y}_i}{z_iM_0} \tag{11.4}$$

Esto resulta porque, en muestreo repetido, la i-ésima unidad aparece con frecuencia relativa z_i, de tal manera que

$$E(\bar{y}_{IV}) = \sum_{i=1}^{N} z_i\left(\frac{M_i\bar{y}_i}{z_iM_0}\right) = \sum_{i=1}^{N}\frac{M_i\bar{y}_i}{M_0} = \bar{Y}$$

La varianza de $\bar{y}_{IV}$ se obtiene de la manera usual. Escríbase

$$\bar{y}_{IV} - \bar{Y} = \frac{M_i\bar{y}_i}{z_iM_0} - \bar{Y} \qquad \text{[por (11.4)]}$$

$$= \frac{1}{M_0}\left[\frac{M_i}{z_i}(\bar{y}_i - \bar{Y}_i) + \left(\frac{M_i}{z_i}\bar{Y}_i - M_0\bar{Y}\right)\right]$$

En la varianza, cada cuadrado recibe una ponderación z_i. Por lo tanto

$$V(\bar{y}_{IV}) = \frac{1}{M_0^2}\left[\sum_{i=1}^{N}\frac{M_i(M_i - m_i)}{z_i}\frac{S_{2i}^2}{m_i} + \sum_{i=1}^{N} z_i\left(\frac{M_i\bar{Y}_i}{z_i} - M_0\bar{Y}\right)^2\right] \tag{11.5}$$

Si $z_i = M_i/M_0$, (11.5) se reduce a (11.3) para $V(\bar{y}_{III})$. Si $z_i = 1/N$ (probabilidades iniciales iguales), (11.5) se reduce a (11.2) para

la varianza de la estimación insesgada cuando las probabilidades son iguales.

A menos que $z_i = M_i/M_0$, la componente entre unidades en (11.5) es afectada en cierto grado por las variaciones en los tamaños M_i tanto como por las variaciones en las medias $\bar{Y}_i$ por elemento.

TABLA 11.3. CÁLCULO DE $V(\bar{y}_{IV})$

Unidad	M_i	M_i/M_0	z_i	m_i	$\frac{M_i(M_i-m_i)}{z_i m_i}$	S_{2i}^2	Y_i	$\frac{Y_i}{z_i}$	$\frac{Y_i}{z_i} - Y$
1	2	0.17	0.2	2	0	0.500	1	5	−28
2	4	0.33	0.4	2	10	0.667	8	20	−13
3	6	0.50	0.4	2	30	0.800	24	60	+27

Ejemplo. La Tabla 11.3 muestra los cálculos para encontrar $V(\bar{y}_{IV})$ en la población artificial de la Tabla 11.1. Las z_i se han tomado como 0.2, 0.4 y 0.4, y las $m_i = 2$.

De (11.5), la varianza resulta ser como sigue:

$$\text{contribución dentro-de-unidades} = \sum \frac{M_i(M_i-m_i)S_{2i}^2}{z_i m_i}/M_0^2 = 0.213$$

$$\text{contribución entre-unidades} = \sum z_i\left(\frac{Y_i}{z_i} - Y\right)^2/M_0^2 = 3.583$$

Una comparación con la Tabla 11.2 muestra que el Método IV tiene una varianza más baja que el Método sin sesgo II, en el cual la unidad primaria se escoge con probabilidades iguales, pero el Método IV es decididamente inferior al Método I o al Método III. En este ejemplo, el Método IV paga un precio muy alto para obtener una estimación insesgada.

Consecuentemente, es natural considerar si la media de la muestra (como en el Método I) sería mejor que la estimación adoptada en el Método IV.

V. Unidades escogidas con probabilidad proporcional a tamaño estimado

$$\text{Estimación} = \bar{y}_V = \bar{y}_i = \text{media de la muestra}$$

La estimación es sesgada ya que, por ejemplo,

$$E(\bar{y}_i) = \sum z_i \bar{Y}_i = \bar{Y}_z$$

Si las z_i son buenas estimaciones, $\bar{Y}_z$ está cercana a la media correcta $\bar{Y} = \sum M_i \bar{Y}_i/M_0$ y el sesgo es pequeño.

Si escribimos

$$\bar{y}_V - \bar{\bar{Y}} = (\bar{y}_i - \bar{Y}_i) + (\bar{Y}_i - \bar{Y}_z) + (\bar{Y}_z - \bar{\bar{Y}})$$

las tres componentes del ECM resultan ser como sigue:

$$\text{ECM}(\bar{y}_V) = \sum_{i=1}^{N} \frac{z_i(M_i-m_i)}{M_i}\frac{S_{2i}^2}{m_i} + \sum_{i=1}^{N} z_i(\bar{Y}_i - \bar{Y}_z)^2 + (\bar{Y}_z - \bar{\bar{Y}})^2 \qquad (11.6)$$

Ejemplo. Si los valores de z_i y m_i se escogen como en la Tabla 11.3, se puede verificar que las componentes de la varianza de $\bar{y}_V$ son como se muestra en la Tabla 11.4.

TABLA 11.4. CONTRIBUCIONES AL ECM EN EL MÉTODO V

Dentro unidades	Entre unidades	Sesgo	ECM Total
0.173	1.800	0.062	2.035

Este es superior a todos los métodos, excepto el Método III (pps) y es casi tan bueno como el Método III.

11.4 RESUMEN DE LOS METODOS PARA $n = 1$

Los cinco métodos para estimar la media $\bar{\bar{Y}}$ por elemento y sus ECM en el ejemplo numérico se resumen en la Tabla 11.5.

TABLA 11.5. MÉTODOS DE MUESTREO EN DOS ETAPAS ($n = 1$)

Método	Probabilidades en la selección de unidades	Estimación de $\bar{\bar{Y}}$	Condición de sesgo	ECM en ejemplo
I	Iguales	$\bar{y}_i$	Con	Ia: 2.541 Ib: 2.579
II	Iguales	$\frac{NM_i\bar{y}_i}{M}$	Sin	6.048
III	$\frac{M_i}{M_0} \propto$ Tamaño	$\bar{y}_i$	Sin	2.002
IV	$z_i \propto$ Tamaño estimado	$\frac{M_i\bar{y}_i}{z_i M_0}$	Sin	3.796
V	$z_i \propto$ Tamaño estimado	$\bar{y}_i$	Con	2.035

11.5 METODOS DE MUESTREO CUANDO $n > 1$

Para $n > 1$ los principales métodos de muestreo son extensiones naturales, tanto de los Métodos I a IV precedentes como de los métodos discutidos en el Cap. 9A para muestreo de una etapa con unidades que son conglomerados de tamaños desiguales. Las cantidades más a menudo estimadas son los totales de población, las medias de población por subunidad y las medias o proporciones que tienen la estructura de estimaciones de razón.

En muchas aplicaciones, la cantidad M_0, número total de subunidades de la población, es desconocida, y sólo se conocen los M_i para las unidades primarias de la muestra, que se extraen dado que pueden contarse por medio de una lista. Por lo tanto, vale la pena observar que los Métodos II y IV y sus extensiones a $n > 1$, no requieren el conocimiento de M_0 para la estimación del *total* de población. El método I y sus extensiones no requieren que se conozca M_0 para estimar la media por subunidad. El Método III de probabilidad proporcional al tamaño, requiere el conocimiento de M_0.

Las estimaciones se denominan *autoponderantes* cuando simplemente son un múltiplo del total sobre todas las subunidades de la muestra. En vista de la conveniencia práctica de las estimaciones autoponderantes en encuestas a gran escala, debe observarse la condición en la cual cada estimación se vuelve autoponderante. Para las estimaciones insesgadas, la condición será, como veremos, que cada unidad de la segunda etapa en la población, o más generalmente, cada unidad de la última etapa del muestreo, tenga la misma oportunidad de extracción.

11.6 DOS RESULTADOS UTILES

Al investigar varianzas y estimaciones de muestra son útiles los dos siguientes resultados. Consisten en mostrar cómo se extienden los resultados del muestreo de una etapa a resultados de muestreo de dos etapas o muestreo de múltiples etapas. Primero los probó Durbin (1953) para estimadores que son sumas de estimadores para las unidades primarias de la muestra y posteriormente, Des Raj (1966) los extendió a funciones lineales de los estimadores de las unidades primarias, y nuevamente, Rao (1975*b*) los extendió a casos más complejos. Seguiremos el estudio de Des Raj.

Las unidades primarias se eligen sin restitución con probabilidades iguales o desiguales. En la i-ésima unidad primaria, sea $\hat{Y}_i$ un estimador insesgado del total de unidad $\hat{Y}_i$, con varianza de

segunda etapa σ_{2i}^2. El submuestreo es independiente en las diferentes unidades primarias. Se considera un estimador insesgado del total de población Y de la forma.

$$\hat{Y} = \sum_{i=1}^{n} w_{is} \hat{Y}_i \tag{11.7}$$

donde el peso w_{is} se conoce para cada muestra s, y puede depender de las otras unidades primarias contenidas en la muestra, así como de la unidad i.

Adoptaremos el artificio usado en la Sec. 2.9 al dejar w_{is}', como una variable aleatoria, que es igual a w_{is} si la unidad i aparece en la muestra, y cero en el caso contrario. Así obtenemos

$$\hat{Y} = \sum_{i=1}^{N} w_{is}' \hat{Y}_i \tag{11.8}$$

Ahora bien

$$E(\hat{Y}) = E_1 E_2(\hat{Y}) = E_1\left(\sum_{i=1}^{N} w_{is}' Y_i\right) = Y \tag{11.9}$$

si y sólo si $E_1(w_{is}') = 1$ para cada i.

Teorema 11.1.

$$V(\hat{Y}) = V\left(\sum_{i=1}^{n} w_{is} \hat{Y}_i\right) = V\left(\sum_{i=1}^{n} w_{is} Y_i\right) + \sum_{i=1}^{N} E_1(w_{is}'^2)\sigma_{2i}^2 \tag{11.10}$$

Prueba. Por la Fórmula (10.2),

$$\begin{aligned} V(\hat{Y}) &= V_1[E_2(\hat{Y})] + E_1[V_2(\hat{Y})] \\ &= V\left(\sum_{i=1}^{n} w_{is} Y_i\right) + E_1\left[\sum_{i=1}^{N} w_{is}'^2 V_2(\hat{Y}_i)\right] \end{aligned} \tag{11.11}$$

puesto que la covarianza de segunda etapa entre $\hat{Y}_i$ y $\hat{Y}_j$ $(i \neq j)$ es cero, porque el submuestreo es independiente. Por lo tanto,

$$V(\hat{Y}) = V\left(\sum_{i=1}^{n} w_{is} Y_i\right) + \sum_{i=1}^{N} [E_1(w_{is}'^2)\sigma_{2i}^2] \tag{11.12}$$

Con esto se completa la prueba.

El primer término en (11.12) es la varianza del estimador correspondiente, cuando cada unidad primaria de la muestra se enumera totalmente y se obtiene a partir de resultados del muestreo en una etapa.

Ejemplo. Para el análogo de dos etapas del estimador de Horvitz-Thompson, $\hat{Y}_{HT} = \sum^{n} \hat{Y}_i/\pi_i$, el peso $w_{is}' = 1/\pi_i$ si la unidad i está en la muestra y

cero en el caso contrario. De modo que $E_1(w_{is}'^2) = \pi_i/\pi_i^2 = 1/\pi_i$, donde π_i es la probabilidad de extraer la unidad i-ésima. Además, si m_i subunidades se han extraído de entre las M_i por muestreo aleatorio simple, siempre que se extraiga la unidad i-ésima,

$$\sigma_{2i}^2 = V_2(\hat{Y}_i) = \frac{M_i(M_i - m_i)}{m_i} S_{2i}^2 \tag{11.13}$$

Así pues, por el Teorema 11.1 y usando la Fórmula (9A.42) como análogo de una etapa de $V(\hat{Y}_{HT})$, tenemos

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 + \sum_{i=1}^{N}\frac{M_i(M_i - m_i)S_{2i}^2}{m_i\pi_i} \tag{11.14}$$

Teorema 11.2. Supóngase que tenemos una estimación insesgada $\hat{\sigma}_{2i}^2$ de la varianza de segunda etapa de σ_{2i}^2 de $\hat{Y}_i$, y una estimación muestral insesgada de $V(\sum^n w_{is}Y_i) = V(\sum^N w_{is}'Y_i)$ del muestreo de una etapa. Esta última será una forma cuadrática como

$$v\left(\sum^n w_{is}Y_i\right) = \sum_i^n a_{is}Y_i^2 + 2\sum_i^n\sum_{j>i}^n b_{ijs}Y_iY_j \tag{11.15}$$

Entonces, un estimador muestral insesgado de $V(\sum^n w_{is}\hat{Y}_i)$ es

$$v\left(\sum^n w_{is}\hat{Y}_i\right) = \left(\sum_i^n a_{is}\hat{Y}_i^2 + 2\sum_i^n\sum_{j>i}^n b_{ijs}\hat{Y}_i\hat{Y}_j\right) + \sum^n w_{is}\hat{\sigma}_{2i}^2 \tag{11.16}$$

Por lo tanto, la regla para construir el estimador muestral de $V(\sum^n w_{is}\hat{Y}_i)$ es: en un estimador insesgado de $V(\sum^n w_{is}Y_i)$ de la primera etapa del muestreo se inserta $\hat{Y}_i$ en donde aparece Y_i. A esto se suma el término $\sum^n(w_{is}\hat{\sigma}_{2i}^2)$, donde $\sum^n w_{is}\hat{Y}_i = \hat{Y}$, y $\hat{\sigma}_{2i}^2$ es un estimador insesgado de $V_2(\hat{Y}_i)$.

Prueba.

$$V\left(\sum_i^N w_{is}'Y_i\right) = \sum_i^N Y_i^2 V(w_{is}') + 2\sum_i^N\sum_{j>i}^N Y_iY_j \operatorname{cov}(w_{is}'w_{js}') \tag{11.17}$$

Intróduzcase la variable aleatoria, a_{is}', donde $a_{is}' = a_{is}$ si la unidad i pertenece a la muestra y $a_{is}' = 0$ en el caso contrario. De manera semejante, sea $b_{ijs}' = b_{ijs}$, si las unidades i y j pertenecen a la muestra y $b_{ijs}' = 0$ en el caso contrario. A partir de (11.15) para muestreo de una etapa

$$v\left(\sum^N w_{is}'Y_i\right) = \sum_i^N a_{is}'Y_i^2 + 2\sum_i^N\sum_{j>i}^N b_{ijs}'Y_iY_j \tag{11.18}$$

Si esto debe ser insesgado, al comparar (11.18) con (11.17) se ve que debemos tener $E_1(a_{is}') = V(w_{is}')$. Ahora, para nuestro estimador de la varianza (11.16)

$$E_1E_2\left(\sum_i^N a_{is}'\hat{Y}_i^2 + 2\sum_i^N\sum_{j>i}^N b_{ijs}'\hat{Y}_i\hat{Y}_j\right) + E_1E_2\left(\sum^N w_{is}'\hat{\sigma}_{2i}'^2\right)$$

$$= E_1\left(\sum_i^N a_{is}'Y_i^2 + 2\sum_i^N\sum_{j>i}^N b_{ijs}'Y_iY_j\right) + \sum_i^N[V(w_{is}') + E^2(w_{is}')]\sigma_{2i}^2 \tag{11.19}$$

En (11.19) hemos usado los resultados en los que $E(a_{is}') = V(w_{is}')$ y en los que para todo i, $E_1(w_{is}') = 1 = E_1^2(w_{is}')$. A continuación

$$E\left[v\left(\sum^n w_{is}\hat{Y}\right)\right] = V\left(\sum^n w_{is}Y_i\right) + \sum_i^N E_1(w_{is}'^2)\sigma_{2i}^2 = V\left(\sum_i^n w_{is}\hat{Y}_i\right) \tag{11.20}$$

por (11.12), y así concluye la prueba.

Consideramos ahora algunos estimadores específicos del total de población Y. En encuestas muy extensas la selección de las unidades primarias con probabilidades desiguales ha llegado a ser la técnica más usada. Las Secs. 11.9 (muestreo con restitución) y 11.10 (muestreo sin restitución) tratan sobre estos dos métodos; la Sec. 11.13 presenta el estimador de razón en muestreo *ppz*. Otras Secs. (11.7) y (11.8) dan los métodos correspondientes, cuando las unidades primarias se eligen con probabilidades iguales.

11.7 UNIDADES SELECCIONADAS CON PROBABILIDADES IGUALES: ESTIMADOR INSESGADO

Salvo mención de lo contrario, las m_i subunidades de muestra que pertenecen a la i-ésima unidad se tomarán con muestreo aleatorio simple. El estimador insesgado del total de población es

$$\hat{Y}_u = \frac{N}{n}\sum^n M_i\bar{y}_i = \frac{N}{n}\sum^n Y_i \tag{11.21}$$

Para aplicar el Teorema 11.1, obsérvese que

$$w_{is} = \frac{N}{n}; \qquad E(w_{is}') = \frac{n}{N}\frac{N}{n} = 1; \qquad E(w_{is}'^2) = \frac{n}{N}\frac{N^2}{n^2} = \frac{N}{n}$$

en consecuencia, por el Teorema 11.1 y resultados anteriores

$$V(\hat{Y}_u) = \frac{N^2}{n}(1 - f_1)\frac{\sum(Y_i - \bar{Y})^2}{(N-1)} + \frac{N}{n}\sum^N\frac{M_i^2(1 - f_{2i})S_{2i}^2}{m_i} \tag{11.22}$$

donde $f_{2i} = m_i/M_i$. El estimador se hace autoponderante si f_{2i} es constante, ($= f_2$, digamos). Entonces tenemos

$$\hat{Y}_u = \frac{N}{nf_2}\sum_i^m\sum_j^{m_i} y_{ij} \tag{11.23}$$

La cantidad nf_2/N es, desde luego, la probabilidad de que se extraiga cualquier unidad de segunda etapa.

Para una estimación muestral insesgada de la varianza, el Teorema 11.2 proporciona, a partir de (11.16),

$$v(\hat{Y}_u)=\frac{N^2(1-f_1)}{n}\frac{\sum_i^n(\hat{Y}_i-\hat{\bar{Y}}_u)^2}{n-1}+\frac{N}{n}\sum_i^n\frac{M_i^2(1-f_{2i})s_{2i}^2}{m_i} \tag{11.24}$$

11.8 UNIDADES SELECCIONADAS CON PROBABILIDADES IGUALES: ESTIMACION DE LA RAZON AL TAMAÑO

El estimador del total de población Y es

$$\hat{Y}_R=M_0\frac{\sum^n M_i\bar{y}_i}{\sum^n M_i}=M_0\frac{\sum^n \hat{Y}_i}{\sum^n M_i} \tag{11.25}$$

Este es un estimador de razón característico ya que tanto el denominador como el numerador varían de una muestra a otra. Se utiliza principalmente para estimar medias por subunidad, en donde no se requiere el conocimiento de M_0 y en donde este estimador es una extensión del método I con $n = 1$.

Para encontrar su ECM aproximado, escríbase

$$\hat{Y}_R-Y=M_0\frac{\sum^n M_i\bar{y}_i}{\sum^n M_i}-Y\doteq\frac{N}{n}\sum^n M_i(\bar{y}_i-\bar{\bar{Y}}) \tag{11.26}$$

donde $\bar{\bar{Y}}=Y/M_0$.

Dado que $\hat{Y}_u=(N/n)\sum^n M_i\bar{y}_i$, podemos obtener el ECM$(\hat{Y}_R)$ aproximado, a partir de la Fórmula (11.22) para $V(\hat{Y}_u)$, al sustituir $M_i(\bar{y}_i-\bar{\bar{Y}})$ en vez de $M_i\bar{y}_i$ o más generalmente, $(y_{ij}-\bar{\bar{Y}})$ para y_{ij}. Ahora (11.22) es

$$V(\hat{Y}_u)=\frac{N^2}{n}(1-f_1)\frac{\sum^N(Y_i-\bar{\bar{Y}})^2}{N-1}+\frac{N}{n}\sum^n\frac{M_i^2(1-f_{2i})S_{2i}^2}{m_i}$$

En la sustitución, $Y_i=M_i\bar{Y}_i$ se convierten en $M_i(\bar{Y}_i-\bar{\bar{Y}})$ en tanto que $\bar{\bar{Y}}=\sum^N y_i/N$ se reemplaza por 0. También, $S_{2i}^2=\sum(y_{ij}-\bar{Y}_i)^2/(M_i-1)$ permanece inalterado cuando $(y_{ij}-\bar{\bar{Y}})$ reemplaza a y_{ij}. Esto nos da el resultado

$$\text{ECM}\,(\hat{Y}_R)\doteq\frac{N^2}{n}(1-f_1)\frac{\sum^N M_i^2(\bar{Y}_i-\bar{\bar{Y}})^2}{N-1}+\frac{N}{n}\sum^N\frac{M_i^2(1-f_{2i})S_{2i}^2}{m_i} \tag{11.27}$$

Así como con $\hat{Y}_u$, este estimador se hace autoponderante si

$$f_{2i}=\frac{m_i}{M_i}=\text{constante}\,f_2=\frac{\bar{m}}{\bar{M}}=\frac{N\bar{m}}{M_0}$$

En este caso, la contribución dentro de unidades se puede expresar más simplemente al poner $m_i=N\bar{m}M_i/M_0$, lo que da

$$\text{ECM}(\hat{Y}_R)\doteq\frac{N^2}{n}(1-f_1)\frac{\sum^N M_i^2(\bar{Y}_i-\bar{\bar{Y}})^2}{(N-1)}+\frac{M_0^2(1-f_2)}{n\bar{m}}\sum^N\left(\frac{M_i}{M_0}\right)S_{2i}^2 \tag{11.28}$$

Puede observarse el parecido con la fórmula correspondiente en la que las unidades primarias son del mismo tamaño. De (10.14) Sec. 10.3, tenemos, al multiplicar por M_0^2 ya que $\hat{Y}=M_0\bar{\bar{y}}$,

$$V(\hat{Y})=\frac{M_0^2(1-f_1)}{n}\frac{\sum^N(\bar{Y}_i-\bar{\bar{Y}})^2}{(N-1)}+\frac{M_0^2(1-f_2)}{nm}\sum^N\left(\frac{1}{N}\right)S_{2i}^2 \tag{11.29}$$

La diferencia es que en (11.28) las contribuciones de los ECM de las unidades primarias son ponderadas, pues las unidades más grandes reciben mayor peso.

Por el Teorema 11.2, una estimación muestral aproximada del ECM$(\hat{Y}_R)$ en 11.27 está dada por

$$v(\hat{Y}_R)=\frac{N^2}{n}(1-f_1)\frac{\sum^n M_i^2(\bar{y}_i-\hat{\bar{Y}}_R)^2}{n-1}+\frac{N}{n}\sum^n\frac{M_i^2(1-f_{2i})s_{2i}^2}{m_i} \tag{11.30}$$

donde $\hat{\bar{Y}}_R=\sum^n M_i\bar{y}_i/\sum^n M_i=\hat{Y}_R/M_0$.

Cuando este estimador se utiliza para estimar la media de población por subunidad, tenemos $\hat{\bar{Y}}_R=\sum^n\hat{Y}_i/\sum^n M_i$ y $v(\hat{\bar{Y}}_R)=v(\hat{Y}_R)/M_0^2$. Cuando M_0 se desconoce, sustituimos $\hat{M}_0=N(\sum^n M_i/n)$ en vez de M_0 en el cálculo de $v(\hat{\bar{Y}}_R)$.

Cuando se seleccionan unidades primarias con probabilidades iguales, una estimación alternativa de la media de población por subunidad (otra extensión del método I) es

$$\frac{1}{n}(\bar{y}_1+\bar{y}_2+\cdots+\bar{y}_n)$$

Esta estimación es autoponderante si m_i = constante, como en el ejemplo que sigue. Cuando m_i y $\bar{Y}_i$ están no correlacionadas, esta estimación puede ser satisfactoria, pero es susceptible de un sesgo que no se anula ni aun con n grande.

Ejemplo. Del volumen, *American Men of Science*, se tomaron 20 páginas al azar. De cada página se registraron las edades de dos científicos, también

elegidos aleatoriamente. El número total de biografías por página, varía en general entre 14 y 21. Estimar la edad promedio y su error estándar de los datos de la Tabla 11.6, usando el estimador de razón.

De la columna extrema derecha,

$$\hat{\bar{Y}}_R = \frac{\sum M_i \bar{y}_i}{\sum M_i} = \frac{17\,121.5}{359} = 47.7 \text{ años}$$

TABLA 11.6. EDADES DE 40 CIENTÍFICOS EN *American Men of Science* ($n = 20$, $m = 2$)

Unidad No.	M_i	Edades y_{i1}	Edades y_{i2}	Total y_i	$M_i\bar{y}_i$
1	15	47	30	77	577.5
2	19	38	51	89	845.5
3	19	43	45	88	836.0
4	16	55	41	96	768.0
5	16	59	45	104	832.0
6	19	39	38	77	731.5
7	18	43	43	86	774.0
8	18	49	51	100	900.0
9	18	45	35	80	720.0
10	18	46	59	105	945.0
11	20	71	64	135	1 350.0
12	18	35	46	81	729.0
13	19	61	54	115	1 092.5
14	19	45	87	132	1 254.0
15	18	31	38	69	621.0
16	16	64	39	103	824.0
17	16	63	47	110	880.0
18	19	36	33	69	655.5
19	19	61	39	100	950.0
20	19	54	34	88	836.0
Totales	359			1 904	17 121.5

Como n/N puede ignorarse, tenemos de (11.30), al dividir por M_0^2,

$$v(\hat{\bar{Y}}_R) \doteq \frac{\sum^{n} M_i^2(\bar{y}_i - \hat{\bar{Y}}_R)^2}{n\bar{M}^2(n-1)}$$

El numerador se calcula como

$$\sum(M_i\bar{y}_i)^2 - 2\hat{\bar{Y}}_R \sum (M_i\bar{y}_i)M_i + \hat{\bar{Y}}_R{}^2 \sum M_i^2$$
$$= 15\,375\,020 - (95.3844)(309\,747.5) + (2274.55)(6481) = 571\,300$$

Como $\bar{M}_n = 359/20$, estimado de la muestra, se obtiene,

$$v(\hat{\bar{Y}}_R) = \frac{(20)(571\,300)}{(19)(359)^2} = 4.67$$

$$s(\hat{\bar{Y}}_R) = 2.16 \text{ años}$$

11.9 UNIDADES SELECCIONADAS CON PROBABILIDADES DESIGUALES Y CON RESTITUCION: ESTIMADOR INSESGADO

Las unidades primarias se seleccionan con probabilidades proporcionales a z_i *con restitución*. Los resultados para $z_i = M_i/M_0$ (probabilidad proporcional al tamaño) siguen como un caso especial.

La submuestra de m_i subunidades de la i-ésima unidad se supone extraída aleatoriamente *sin restitución*. Si la i-ésima unidad se selecciona más de una vez, se supone que en cada selección se restituye la totalidad de la submuestra, al hacer una nueva extracción independiente de m_i unidades sin restitución de la unidad completa.

Una estimación insesgada del total de población es

$$\hat{Y}_{ppz} = \frac{1}{n}\sum^{n} \frac{M_i\bar{y}_i}{z_i} = \frac{1}{n}\sum^{n} \frac{\hat{Y}_i}{z_i} \tag{11.31}$$

Para $n = 1$, se mostró en la Sec. 11.3 que este estimador, $M_0\bar{y}_{\text{IV}} = \hat{Y}_{\text{IV}}$, es insesgado. Su varianza se obtiene de la Fórmula (11.5) al multiplicar por M_0^2 como

$$V(\hat{Y}_{\text{IV}}) = \sum_{i=1}^{N} z_i\left(\frac{Y_i}{z_i} - Y\right)^2 + \sum_{i=1}^{N} \frac{M_i(M_i - m_i)S_{2i}{}^2}{z_i m_i} \tag{11.32}$$

Con este método de muestreo, el estimador $\hat{Y}_{ppz}$ es la media de n estimaciones independientes de la forma $\hat{Y}_{\text{IV}}$. En consecuencia, a partir de la teoría clásica del muestreo se sabe que $\hat{Y}_{ppz}$ es insesgado y

$$V(\hat{Y}_{ppz}) = \frac{1}{n}V(\hat{Y}_{\text{IV}}) = \frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{Y_i}{z_i} - Y\right)^2 + \frac{1}{n}\sum_{i=1}^{N} \frac{M_i^2(1-f_{2i})S_{2i}{}^2}{m_i z_i} \tag{11.33}$$

Además, dadas las n estimaciones independientes $\hat{Y}_{\text{IV}} = Y_i/z_i$, un estimador muestral insesgado de $V(\hat{Y}_{\text{IV}})$ es, desde luego,

$$v(\hat{Y}_{\text{IV}}) = \frac{\sum_{i=1}^{n}\left(\frac{\hat{Y}_i}{z_i} - \hat{Y}_{ppz}\right)^2}{(n-1)} \tag{11.34}$$

En consecuencia, un estimador muestral insesgado de $V(\hat{Y}_{ppz})$ es la expresión muy simple

$$v(\hat{Y}_{ppz}) = \frac{\sum_{i=1}^{n} \left(\frac{\hat{Y}_i}{z_i} - \hat{Y}_{ppz} \right)^2}{n(n-1)} \tag{11.35}$$

Estos resultados también son válidos para el muestreo de múltiples etapas, siempre que $\hat{Y}_i$ sea un estimador insesgado de Y_i, y que el submuestreo sea independiente cuando se extraiga una unidad primaria.

Para descubrir en qué ocasiones $\hat{Y}_{ppz}$ se hace autoponderante en el muestreo de dos etapas, escribimos

$$\hat{Y}_{ppz} = \frac{1}{n} \sum_{i}^{n} \frac{M_i}{z_i m_i} \sum_{j}^{m_i} y_{ij} \tag{11.36}$$

Por lo tanto, la condición es

$$\frac{n z_i m_i}{M_i} = \text{constante} = f_0 \tag{11.37}$$

Esta expresión es la probabilidad de extraer cualquier unidad específica de la segunda etapa. A partir de (11.37), $m_i/M_i = f_0/nz_i$. Si la probabilidad f_0 se elige de antemano, se podrá decir al trabajador de campo qué fracción de la segunda etapa m_i/M_i debe tomar de cualquier unidad primaria elegida. Por ejemplo, supongamos $f_0 = 1/50 = 0.02$ y $n = 60$ unidades primarias elegidas. Si $z_i = 0.0026$ para una unidad seleccionada, debemos tener $m_i/M_i = 0.02/(60)(0.0026)$, o 1 en 7.8.

Con una muestra autoponderante, el estimador de varianza (11.35) toma la forma más simple

$$v(\hat{Y}_{ppz}) = \frac{n}{(n-1) f_0^2} \sum_{i}^{n} (y_i - \bar{y})^2 \tag{11.38}$$

donde $y_i = \sum_j y_{ij}$ es el total de muestra en la unidad i-ésima. La sencillez de las varianzas estimadas (11.35) y (11.38) es un rasgo atractivo del muestreo con restitución.

Con este método, hay otras maneras de extraer las submuestras. Si la unidad i-ésima se selecciona t_i veces, una variante del método es la extracción de una *sola* submuestra de tamaño $t_i m_i$ sin restitución, a condición de que $M_i > m_i t_i$. Sukhatme (1954) ha mostrado que con este método, $V(\hat{Y}_{ppz})$ se reduce en $(n-1)\sum^{N} M_i S_{2i}^2/n$. Otro método es el de extraer una sola submuestra de tamaño m_i sin importar cuántas veces se selecciona la unidad i-ésima. La estimación $M_i\bar{y}_i/z_i$ de esta unidad recibe el peso t_i (número de veces que se ha

extraído la unidad) con cualquiera de los dos métodos. El efecto es un incremento de $V(\hat{Y}_{ppz})$ en la cantidad

$$\left(\frac{n-1}{n}\right) \sum^{N} \frac{M_i^2 (1-f_{2i}) S_{2i}^2}{m_i}$$

Para el mismo costo, rara vez puede esperarse que las diferencias en precisión entre estos métodos sean sustanciales.

Si $z_i = M_i/M_0$ la estimación insesgada (11.31) se reduce a

$$\hat{Y}_{pps} = \frac{M_0}{n} \sum_{i}^{n} \bar{y}_i \tag{11.39}$$

Es claro que esta estimación se hace autoponderante, cuando $m_i = m$, para que $\hat{Y}_{pps} = M_0 \bar{y}$.

Un estimador muestral insesgado es

$$v(\hat{Y}_{pps}) = \frac{M_0^2}{n(n-1)} \sum_{i}^{n} \left(\bar{y}_i - \frac{\hat{Y}_{pps}}{M_0} \right)^2 \tag{11.40}$$

11.10 SELECCION DE LAS UNIDADES SIN RESTITUCION

Para cualquiera de los métodos "sin restitución" estudiados en el Cap. 9A, se obtienen las fórmulas para la varianza, y varianza estimada en muestreo de dos etapas, a partir de los Teoremas 11.1, 11.2. Con los métodos de Brewer o Durbin,

$$\hat{Y}_B = \sum^{n} \frac{M_i \bar{y}_i}{\pi_i} = \frac{1}{n} \sum^{n} \frac{M_i \bar{y}_i}{z_i} = \frac{1}{n} \sum^{n} \frac{\hat{Y}_i}{z_i} \tag{11.41}$$

con varianza

$$V(\hat{Y}_B) = \sum_{i}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 + \sum_{i}^{N} \frac{M_i^2 (1-f_{2i}) S_{2i}^2}{m_i \pi_i} \tag{11.42}$$

Para el estimador correspondiente $\hat{Y}_{ppz}$ en muestreo con restitución, tenemos, a partir de (11.33), con $\pi_i = nz_i$,

$$V(\hat{Y}_{ppz}) = \frac{1}{n} \sum^{N} z_i \left(\frac{Y_i}{z_i} - Y \right)^2 + \sum^{N} \frac{M_i^2 (1-f_{2i}) S_{2i}^2}{m_i \pi_i} \tag{11.43}$$

Puesto que las contribuciones "dentro de unidades" a la varianza son las mismas en (11.33) y (11.43), cualquier ganancia relativa en la precisión al seleccionar unidades sin restitución, se diluye en las muestras de dos etapas por la contribución intraunidades a la varianza. Por ejemplo, en un muestreo de varias etapas estratificado de $n = 44$ provincias, entre $N = 147$ provincias, Des Raj (1964) encontró que la razón $V(\hat{Y}_{WOR})/V(\hat{Y}_{WR})$ promediaba alrededor de 0.79

sobre siete atributos para la componente entre unidades, pero el promedio de la razón sobre ambas componentes era 0.92.

Con $n = 2$, una estimación muestral insesgada de $V(\hat{Y}_B)$ es

$$v(\hat{Y}_B) = (\pi_1\pi_2\pi_{12}^{-1} - 1)\left(\frac{M_1\bar{y}_1}{\pi_1} - \frac{M_2\bar{y}_2}{\pi_2}\right)^2 + \sum_{i=1}^{2} \frac{M_i^2(1-f_{2i})s_{2i}^2}{m_i\pi_i} \quad (11.44)$$

donde los subíndices 1, 2 denotan las unidades elegidas.

La extensión al estimador de Sampford para $n > 2$ es directa. El estimador alternativo de *RHC*

$$\hat{Y}_{RHC} = \sum_{g}^{n} \frac{Z_g M_g \bar{y}_g}{z_g} \quad (11.45)$$

donde $Z_g = \sum z_i$ sobre el grupo g-ésimo y $M_g, \bar{y}_g$, y z_g se refieren a la unidad extraída del grupo. Para una muestra de n unidades, un estimador insesgado de $V(\hat{Y}_{RHC})$ es

$$v(\hat{Y}_{RHC}) = \frac{\left(\sum^{n} N_g^2 - N\right)}{(N^2 - \sum N_g^2)} \sum_{g}^{n} Z_g\left(\frac{M_g\bar{y}_g}{z_g} - \hat{Y}_{RHC}\right)^2 + \sum_{g}^{n} Z_g \frac{M_g^2}{z_g} \frac{(1-f_{2g})s_{2g}^2}{m_g} \quad (11.46)$$

Las Fórmulas (11.44) y (11.46) requieren cálculo separado de las contribuciones a la varianza estimada entre unidades y dentro de unidades. En encuestas extensas con numerosos estratos y atributos, la complicación de tales fórmulas hace que la estimación de las varianzas se convierta en un trabajo que requiere más tiempo de computadora del que posiblemente se le puede dedicar en la práctica. Recordemos la forma más simple de la fórmula de la varianza (11.35) cuando se extraen las unidades con restitución; esto es

$$v(\hat{Y}_{ppz}) = \frac{1}{n(n-1)} \sum^{n} \left(\frac{\hat{Y}_i}{z_i} - \hat{Y}_{ppz}\right)^2$$

donde $y_i'' = M_i\bar{y}_i/z_i$. Este resultado hace más atractivo el muestreo con restitución, si no conduce a una gran pérdida de precisión.

Entre los métodos sin restitución, Platek y Singh (1972), al planear la reorganización de la encuesta sobre la fuerza del trabajo canadiense para estratos en los que $n_h = 6$, o un múltiplo de 6, señala las ventajas de $\hat{Y}_{RHC}$ y de la versión de Hartley-Rao del muestreo sistemático (Sec. 9A.10) en cuanto a sencillez y estimación de la varianza.

Para el caso de numerosos estratos pequeños con $n_h = 2$, Durbin (1967) ha producido métodos que dan una estimación de varianza simple si las unidades se seleccionan sin restitución. Este autor usa $\hat{Y}_{HT}$ e investiga métodos para los cuales $\pi_i = 2z_i$. Dentro de un estrato, la estrategia consiste en usar métodos de selección que hacen $(\pi_i\pi_j\pi_{ij}^{-1} - 1)$ igual a 1 o 0. Estas elecciones dan como resultado la reducción de $v(\hat{Y}_{HT})$ en (11.44) al término

$$\left(\frac{\hat{Y}_1}{\pi_1} - \frac{\hat{Y}_2}{\pi_2}\right)^2$$

o al segundo término de (11.44). Brewer y Hanif (1969) mejoraron y extendieron los métodos a estimadores diferentes de $\hat{Y}_{HT}$.

11.11 COMPARACION DE LOS METODOS

En muestreo de una etapa (Sec. 9A.12), se comparó el muestreo con probabilidad igual y con la estimación insesgada $\hat{Y}_u$ con las estimaciones de razón correspondientes y con varios métodos de muestreo de probabilidades desiguales. Dicho brevemente, la razón de la varianza de $\hat{Y}_u$ a la de otro método que muestreaba sin restitución, era igual a la razón del coeficiente de variación del total de unidad Y_i a la de Y_i/z_i. En lo que respecta al muestreo con probabilidades desiguales con y sin restitución, la razón $V(\hat{Y}_{SR})/V(\hat{Y}_{CR})$, era más o menos igual a $(N - n)/(N - 1)$, al igual que en la selección con probabilidades iguales.

En muestreo de dos y más etapas, el efecto neto de la contribución a la varianza dentro de unidades es diluir las diferencias creadas por las contribuciones entre unidades. Con los métodos de submuestreo usados en la práctica, la contribución dentro de unidades suele ser aproximadamente la misma para los diferentes métodos. El efecto neto es que las precisiones relativas globales de diferentes métodos se acercan a la igualdad. En la práctica, la componente dentro de unidades suele ser, cuando menos, tan grande como la componente entre unidades.

Aun si contamos con computadoras poderosas, las formas autoponderantes de las estimaciones son convenientes y muy utilizadas. El uso de un plan autoponderante debería acarrear tan sólo una pérdida de precisión de poca importancia, porque la elección de los m_i sólo afecta la contribución a la varianza dentro de unidades. La forma autoponderante requiere $m_i \propto M_i/z_i$, en tanto que el m_i que minimiza V_2 para un número total esperado de subunidades es $m_i \propto M_iS_{2i}/z_i$. Estas dos asignaciones difieren poco, a menos que S_{2i} varíe dentro de una gama importante de valores.

En las encuestas de muchos atributos, también observamos (Sec. 9A.13) que las z_i usadas para selección de muestra podían no ser

una buena elección respecto a algunos de los atributos, cuando Y_i tuviera poca relación con z_i. Si se puede encontrar una variable auxiliar x_i tal que y_i/x_i sea razonablemente constante, una posibilidad es la de cambiar a la estimación de razón (Sec. 11.13). Si no se dispone de tal x_i, la estimación alternativa que sugiere Rao (1966) es de la forma

$$\hat{Y}^*_{ppz} = \frac{N}{n}\sum_i^N M_i \bar{y}_i \tag{11.47}$$

La evidencia preliminar mostrada por Rao, sugiere que si Y_i y z_i no están relacionadas, $\hat{Y}^*_{ppz}$ se puede comportar casi tan bien como lo haría $\hat{Y}_u$ si el muestreo fuese con probabilidades iguales (como se ilustra en la Sec. 9A.13). Si se eligen las unidades primarias con restitución, una estimación muestral insesgada de $V(\hat{Y}^*_{ppz})$ es

$$v(\hat{Y}^*_{ppz}) = \frac{N^2}{n(n-1)}\sum^n (M_i\bar{y}_i - \hat{\bar{Y}}^*_{ppz})^2 \tag{11.48}$$

donde $\hat{\bar{Y}}^*_{ppz} = \hat{Y}^*_{ppz}/N$. Esta expresión no incluye la contribución del sesgo en esta estimación de su ECM, que se espera pequeña si Y_i y z_i no están relacionadas.

11.12 RAZONES A OTRA VARIABLE

En muestreo de dos etapas, la cantidad que se debe estimar es a menudo una razón Y/X. Esto ocurre por dos razones diferentes. Como ya dijimos antes, si x es el valor de y en un censo reciente, la razón y/x puede ser relativamente estable. Una estimación del total o media de población de y, basados en esta razón, puede ser más precisa que las estimaciones consideradas hasta aquí en este capítulo.

Las estimaciones de razón de este tipo también se encuentran al estimar proporciones o medias sobre partes de la población. En una encuesta urbana con unidad primaria, donde las unidades primarias son las manzanas de una ciudad, un ejemplo de una proporción de este tipo es

$$\frac{\text{número de empleados de sexo masculino de más de 16 años de edad}}{\text{número total de habitantes de sexo masculino, mayores de 16 años}}$$

Si $y_{ij} = 1$ para cualquier empleado de sexo masculino, mayor de 16 y $y_{ij} = 0$ en el caso contrario, y $x_{ij} = 1$ para cualquier habitante del sexo masculino mayor de 14 años y $x_{ij} = 0$ en el caso contrario, la proporción de población es Y/X. Otros ejemplos de este tipo de encuesta son el ingreso promedio de familias suscritas a una cierta revista, o la cantidad promedio de dinero para gastos menores que tiene un adolescente.

Con cualquiera de los métodos precedentes de selección de unidades (probabilidades iguales o desiguales, con o sin restitución), la fórmula aproximada estándar para el ECM de la varianza de $\hat{Y}_R$ y $\hat{R}$ se obtiene fácilmente de la fórmula para $V(\hat{Y})$ para el mismo método de selección de muestra por la técnica que hemos usado repetidamente.

Para obtener $V(\hat{Y}_R)$, se reemplaza y_{ij} por $d_{ij} = y_{ij} - Rx_{ij}$ en $V(\hat{Y})$ para el método de muestreo usado. Para $V(\hat{R})$, se divide también por X^2.

Para el ECM estimado o varianza estimada $v(\hat{Y}_R)$, se reemplaza y_{ij} por $d_{ij}' = y_{ij} - \hat{R}x_{ij}$ en $v(\hat{Y})$. Para $v_2(\hat{R})$, también se divide por $\hat{X}^2$. Esto resulta por el método usado en el Teorema 2.5. Para

$$\hat{Y}_R - Y = X\frac{\hat{Y}}{\hat{X}} - Y = \frac{X}{\hat{X}}(\hat{Y} - R\hat{X}) \doteq (\hat{Y} - R\hat{X}) = \hat{D} \tag{11.49}$$

donde $d_{ij} = y_{ij} - Rx_{ij}$. Dado que $E(\hat{D}) = 0$ con cualquier plan de muestreo en el que $\hat{Y}$ y $\hat{X}$ son insesgadas, encontramos $E(\hat{Y}_R - Y)^2$ al tomar la fórmula para $V(\hat{Y})$ para este plan y al sustituir $d_{ij} = y_{ij} - Rx_{ij}$ en lugar de y_{ij}.

Por ejemplo, consideremos $\hat{Y}_u = (N/n)\sum^n M_i\bar{y}_i = (N/n)\sum^n \hat{Y}_i$ en muestreo con probabilidades iguales. De la Fórmula (11.22), Sec. 11.7,

$$V(\hat{Y}_u) = \frac{N^2}{n}(1-f_1)\frac{\sum(Y_i - \bar{Y})^2}{N-1} + \frac{N}{n}\sum^N \frac{M_i^2(1-f_{2i})S_{2i}{}^2}{m_i} \tag{11.50}$$

Por lo tanto, con muestreo de probabilidades iguales $\hat{Y}_R = X\hat{Y}_u/\hat{X}_u$,

$$V(\hat{Y}_R) \doteq \frac{N^2}{n}(1-f_1)\frac{\sum^N(Y_i - RX_i)^2}{N-1} + \frac{N}{n}\sum^N \frac{M_i^2(1-f_{2i})S_{d2i}^2}{m_i} \tag{11.51}$$

donde

$$S_{d2i}^2 = \frac{1}{M_i - 1}\sum_{i=1}^{M_i}[(y_{ij} - Rx_{ij}) - (\bar{Y}_i - R\bar{X}_i)]^2$$

Las condiciones en las cuales $\hat{Y}_R$ se reducen a un múltiplo de la razón de totales de muestra $\Sigma\Sigma y_{ij}/\Sigma\Sigma x_{ij}$ siempre son aquellas en las que la $\hat{Y}$ correspondiente se hace autoponderante, en este caso $f_{2i} = m_i/M_i$ = constante.

Para la varianza estimada, la sustitución de $d_{ij}' = y_{ij} - \hat{R}x_{ij}$ para y_{ij} en (11.24) para $v(\hat{Y}_u)$ da

$$v(\hat{Y}_R) \doteq \frac{N^2(1-f_1)}{n}\frac{\sum(\hat{Y}_i - \hat{R}\hat{X}_i)^2}{n-1} + \frac{N}{n}\sum^{n}\frac{M_i^2(1-f_{2i})s_{d'2i}^2}{m_i} \qquad (11.52)$$

De modo semejante, con selección *ppz* de probabilidades proporcionales al tamaño z con restitución teníamos

$$\hat{Y}_{ppz} = \frac{1}{n}\sum^{n} M_i\bar{y}_i/z_i$$

(de (11.33) en la Sec. 11.9)

$$V(\hat{Y}_{ppz}) = \frac{1}{n}\sum^{N} z_i\left(\frac{y_i}{z_i} - Y\right)^2 + \frac{1}{n}\sum^{N}\frac{M_i^2(1-f_{2i})S_{2i}^2}{z_i m_i} \qquad (11.53)$$

Por lo tanto, para $\hat{Y}_R = X\hat{Y}/\hat{X}$ en muestreo sin restitución,

$$V(\hat{Y}_R) \doteq \frac{1}{n}\sum^{N}\frac{1}{z_i}(Y_i - RX_i)^2 + \frac{1}{n}\sum^{N}\frac{M_i^2(1-f_{2i})S_{d2i}^2}{z_i m_i} \qquad (11.54)$$

De (11.35), una estimación de muestra $v(\hat{Y}_R)$ ligeramente sesgada es

$$v(\hat{Y}_R) = \frac{1}{n(n-1)}\sum^{n}\left(\frac{\hat{Y}_i - \hat{R}\hat{X}_i}{z_i}\right)^2 \qquad (11.55)$$

Con el método de Brewer del muestreo sin restitución $\hat{Y}_{RB} = X\hat{Y}_B/\hat{X}_B$, donde $\hat{Y}_B = \sum^{n} M_i\bar{y}_i/\pi_i$. Para $n = 2$, las Fórmulas (11.42) y (11.44) dan, para la estimación de razón,

$$V(\hat{Y}_{RB}) = \sum_{i}^{N}\sum_{j>i}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{D_i}{\pi_i} - \frac{D_j}{\pi_j}\right)^2 + \sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})S_{d2i}^2}{m_i\pi_i} \qquad (11.56)$$

donde $D_i = Y_i - RX_i$ y las unidades i y j son las dos seleccionadas.

$$v(\hat{Y}_{RB}) = (\pi_i\pi_j\pi_{ij}^{-1} - 1)\left(\frac{\hat{D}_i'}{\pi_i} - \frac{\hat{D}_j'}{\pi_j}\right)^2 + \sum_{i=1}^{2}\frac{M_i^2(1-f_{2i})s_{d'2i}^2}{m_i\pi_i} \qquad (11.57)$$

11.13 ELECCION DE FRACCIONES DE MUESTREO Y SUBMUESTREO. PROBABILIDADES IGUALES

Primero se discute el problema para la estimación razón a tamaño, cuando las unidades se toman con probabilidades iguales. La fracción de submuestreo $f_2 = m_i/M_i$ se supone constante, así que la estimación es la media de muestra por subunidad.

La más simple de las funciones de costos contiene tres términos.

c_u = costo fijo por unidad primaria
c_2 = costo por subunidad
c_l = costo de la lista por subunidades en una unidad seleccionada

El tercer término se incluye porque generalmente se debe hacer una lista de los elementos de cualquier unidad seleccionada y se debe verificar el número de ellos para extraer una submuestra. Por lo tanto,

$$\text{costo} = c_u n + c_2\sum^{n} m_i + c_l\sum^{n} M_i$$

Esta fórmula no es útil en la forma que aparece, pues el costo depende del conjunto particular de unidades elegidas. En lugar de esto, se considera el costo promedio sobre n unidades, que es igual a

$$E(C) = c_u n + c_2 n\bar{m} + c_l n\bar{M} = (c_u + c_l\bar{M})n + c_2 n\bar{m} = c_1 n + c_2 n\bar{m} \qquad (11.58)$$

donde c_1 incluye ahora el costo promedio de hacer la lista de una unidad.

Determinamos n y $\bar{m} = f_2\bar{M}$ para minimizar $V(\bar{y})$ para un costo dado o *viceversa*. De (11.28) Sec. 11.8, al dividir por $M_0^2 = (\bar{M}N)^2$,

$$\text{ECM}(\bar{y}) \doteq \frac{1-f_1}{n}\frac{\sum^{N} M_i^2(\bar{Y}_i - \bar{\bar{Y}})^2}{\bar{M}^2(N-1)} + \frac{1-f_2}{n\bar{m}}\sum^{N}\frac{M_i}{M_0}S_{2i}^2$$

Escríbase

$$S_b^2 = \frac{\sum^{N} M_i^2(\bar{Y}_i - \bar{\bar{Y}})^2}{\bar{M}^2(N-1)}$$

Esta es una varianza ponderada entre medias por elemento de unidades. Es análoga a la varianza S_1^2 de la Sec. 10.3 y se reduce a S_1^2 si todos los M_i son iguales. También podemos escribir

$$S_2^2 = \sum^{N}\frac{M_i}{M_0}S_{2i}^2$$

Esta es una media ponderada de las varianzas intraunidades. Se reduce a S_2^2 de la Sec. 10.3 si todos los M_i son iguales.

En esta notación, dado que $f_2 = \bar{m}/\bar{M}$,

$$\text{ECM}(\bar{y}) = \frac{1}{n}\left(S_b^2 - \frac{S_2^2}{\bar{M}}\right) + \frac{1}{n\bar{m}}S_2^2 - \frac{1}{N}S_b^2 \qquad (11.59)$$

Al aplicar la desigualdad de Cauchy-Schwarz, como de costumbre, a (11.58) y (11.59) obtenemos

$$\bar{m}_{\acute{o}pt} \doteq \frac{S_2}{\sqrt{S_b^2 - S_2^2/\bar{M}}}\sqrt{\frac{c_1}{c_2}} \tag{11.60}$$

Los métodos dados en la Sec. 10.8 para usar el conocimiento que se tiene respecto a las razones S_2/S_b y c_1/c_2 para guiar la selección de $m_{\acute{o}pt}$ se aplican aquí. La estimación insesgada, cuando las unidades se extraen con probabilidades iguales, puede manejarse en forma parecida.

La siguiente sección presenta un análisis más general de este problema.

11.14 PROBABILIDADES DE SELECCION OPTIMA Y TASAS DE MUESTREO Y SUBMUESTREO

Un análisis antiguo, pero importante, de Hansen y Hurwitz (1949) determina simultáneamente las z_i óptimas como funciones de los M_i y las fracciones óptimas de muestreo y submuestreo. Las unidades se seleccionan con restitución. El análisis se presenta para $\hat{Y}_R$ en la forma autoponderante, de modo que $m_i = f_0 M_i/nz_i = f_0 M_i/\pi_i$.

Como en la Sec. 11.13, la función de costos es

$$C = c_u n + c_2 \sum^n m_i + c_l \sum^n M_i$$

Aunque se conocen los M_i se incluye un costo de lista para aquellas unidades que caen en la muestra, porque esta lista será necesaria para proporcionar un marco al submuestreo.

El costo promedio de muestrear n unidades es

$$E(C) = c_u n + C_2 f_0 M_0 + c_l \sum^N \pi_i M_i \tag{11.61}$$

Por (11.54) enla Sec. 11.12,

$$V(\hat{Y}_R) = \frac{1}{n}\left[\sum^N \frac{1}{z_i}(Y_i - RX_i)^2 + \frac{M_i(M_i - m_i)}{z_i m_i} S_{d2i}^2\right] \tag{11.62}$$

Dado que $d_{ij} = y_{ij} - Rx_{ij}$, podemos escribir $(Y_i - RX_i) \doteq M_i \bar{D}_i$. Al observar que $\pi_i = nz_i$, $M_i/nz_i m_i = 1/f_0$, y combinar los términos primero y tercero en (11.62), obtenemos

$$V = V(\hat{Y}_R) = \sum^N \left[\frac{M_i^2}{\pi_i}\left(\bar{D}_i^2 - \frac{S_{d2i}^2}{M_i}\right) + \frac{M_i}{f_0} S_{d2i}^2\right] \tag{11.63}$$

El problema está en elegir n, f_0 y las $\pi_i = nz_i$ para minimizar V, sujeta a un costo promedio fijo y a la restricción

$$\sum^N z_i = 1, \qquad \sum^N \pi_i = n$$

Tómese λ y μ como multiplicadores de Lagrange, y minimícese.

$$V + \lambda\left[c_u n + c_2 f_0 M_0 + c_l \sum^N \pi_i M_i - E(C)\right] + \mu\left(n - \sum^N \pi_i\right) \tag{11.64}$$

Por derivación se tienen

$$n: \quad \lambda c_u + \mu = 0; \qquad \mu = -\lambda c_u \tag{11.65}$$

$$\pi_i: \quad \frac{-M_i^2}{\pi_i^2}\left(\bar{D}_i^2 - \frac{S_{d2i}^2}{M_i}\right) + \lambda c_l M_i - \mu = 0 \tag{11.66}$$

Las relaciones (11.65) y (11.66) conducen a

$$z_i = \frac{\pi_i}{n} \propto M_i\left(\bar{D}_i^2 - \frac{S_{d2i}^2}{M_i}\right)^{1/2} \Big/ (c_u + c_l M_i)^{1/2} \tag{11.67}$$

Como los valores individuales de $(\bar{D}_i^2 - S_{d2i}^2/M_i)$ no se conocen, podemos considerar la manera como el valor promedio puede depender del tamaño de la unidad M_i, al usar el siguiente argumento más o menos burdo. Supóngase que se dividiera una población en unidades de tamaño M. Como $E(\bar{D}_i) = 0$, la fórmula (9.10), Pág. 298 da

$$E(\bar{D}_i^2) = V(\bar{D}_i) \doteq \frac{S_d^2}{M}[1 + (M-1)\rho_M]$$

donde S_d^2 es la varianza de población de los d_{ij} y ρ_M es la correlación intraunidad para unidades de tamaño M. También de (9.15), Pág. 300,

$$E(S_{d2i}^2) \doteq S_d^2(1 - \rho_M)$$

por lo tanto,

$$E\left(\bar{D}_i^2 - \frac{S_{d2i}^2}{M_i}\right) \doteq \frac{S_d^2}{M}[1 + (M-1)\rho_M - (1-\rho_M)] = \rho_M S_d^2$$

De (11.67) esto nos da como una aproximación

$$z_i \propto \frac{M_i\sqrt{\rho_{M_i}}}{\sqrt{c_u + c_l M_i}} \tag{11.68}$$

con ρ positivo, se puede esperar que ρ_{Mi} decrezca al aumentar M_i, porque las subunidades muy distantes entre sí están menos sujetas a influencias comunes, pero esta disminución puede ser solamente ligera para $\sqrt{\rho_{M_i}}$. Las deducciones a partir de (11.68) son las siguientes:

1. El costo de poner en lista $c_l M_i$ no es importante, $z_i \propto M_i$ (es decir, selección *pps*) es la mejor si $\sqrt{\rho_{Mi}}$ varía poco sobre la gama

de valores de tamaños en la población. Si $\sqrt{\rho_{M_i}}$ decrece notablemente, las probabilidades óptimas están entre $z_i \propto M_i$ y $z_i \propto \sqrt{M_i}$.

2. Si los costos de lista predominan, z_i caería entre $\sqrt{M_i}$ y una constante (probabilidades iguales).

3. Si los costos de lista y los fijos son del mismo orden de magnitud, $z_i \propto \sqrt{M_i}$ puede ser un buen compromiso.

La derivación de (11.64) con respecto a la fracción de muestreo global f_0 da

$$f_0^2 = \frac{\sum^N M_i S_{d2i}^2}{\lambda c_2 M_0} \tag{11.69}$$

El valor de λ se encuentra en términos de las π conocidas, al sumar (11.66) sobre todas las unidades. Este paso lleva al resultado

$$f_0 = \left[\frac{(c_u + c_l \bar{M}) \sum^N (M_i/M_0) S_{d2i}^2}{c_2 \sum \frac{M_i^2}{\pi_i^2} \left(\bar{D}_i^2 - \frac{S_{d2i}^2}{M_i} \right)} \right]^{1/2} \tag{11.70}$$

Al comparar con (11.60), se verá que f_0 tiene la misma estructura que en el muestreo con probabilidades iguales, si recordamos que en (11.60), $f_0 = n\bar{m}_{opt}/M_0$ y $c_1 = c_u + c_l\bar{M}$.

El óptimo n se encuentra de la ecuación de costo promedio (11.61).

11.15 MUESTREO ESTRATIFICADO. ESTIMADORES INSESGADOS

Para los métodos insesgados ($\hat{Y}_u$, $\hat{Y}_{ppz}$, $\hat{Y}_B$, etc.) la extensión al muestreo estratificado es directa. El subíndice h denota el estrato.

El total estimado de población es $\hat{Y}_{st} = \sum \hat{Y}_h$, con

$$V(\hat{Y}_{st}) = \sum_h^L V(\hat{Y}_h), \qquad v(\hat{Y}_{st}) = \sum_h^L v(\hat{Y}_h) \tag{11.71}$$

Estas varianzas se pueden obtener de las fórmulas ya dadas.

Podemos observar las condiciones en las que las estimaciones se hacen autoponderantes. Para $\hat{Y}_{ppz}$ (Sec. 11.9) o $\hat{Y}_B$ (Sec. 11.10),

$$\hat{Y}_{st} = \sum_h^L \frac{1}{n_h} \sum_i^{n_h} \frac{M_{hi} y_{hi}}{m_{hi} z_{hi}} \tag{11.72}$$

donde y_{hi} es el *total* sobre las m_{hi} subunidades de la i-ésima unidad en el estrato h. Estas estimaciones se ve que son autoponderantes

dentro de los estratos, si la probabilidad $f_{0h} = n_h z_{hi} m_{hi}/M_{hi}$ de selección de una subunidad en el estrato es constante dentro del estrato. En este caso, la estimación viene a ser

$$\hat{Y}_{st} = \sum_h^L \frac{1}{f_{0h}} \sum_i^{n_h} y_{hi} \tag{11.73}$$

que es completamente autoponderante si f_{0h} es la misma en todos los estratos como intuitivamente podría preverse.

Si las unidades dentro de un estrato dado son del mismo tamaño (esto es $M_{hi} = M_h$), en la Sec. 10.10 se mostró que la asignación muestral que conducía a una estimación completamente autoponderante se acerca al óptimo a condición de que $S_{2h}/\sqrt{c_{2h}}$ sea razonablemente constante. Cuando los M_{hi} varían dentro de los estratos, un resultado correspondiente para los estimadores como $\hat{Y}_{ppz}$ y $\hat{Y}_B$ es el siguiente. Supongamos que la función de costos es lineal, como en las Sec. 11.13 y 11.14, y que el estimador es autoponderante dentro de los estratos (es decir, $f_{0h} = n_h z_{hi} m_{hi}/M_{hi}$). Entonces, el óptimo f_{0h} para un costo esperado dado, se puede mostrar que es de la forma

$$f_{0h} \propto \frac{1}{\sqrt{c_{2h}}} \sqrt{\sum_i (M_{hi}/M_{0h}) S_{2hi}^2} \tag{11.74}$$

donde $M_{0h} = \Sigma M_{hi}$. Así, la elección $f_{0h} = f_0$, que hace estos estimadores completamente autoponderantes, será cercana al óptimo en cuanto a precisión, a menos que S_{2hi}^2 o bien, los c_{2h} varíen mucho entre estratos, en especial porque esta elección sólo afecta la contribución de segunda etapa a la varianza.

11.16 MUESTREO ESTRATIFICADO. ESTIMACIONES DE RAZON

Las fórmulas para la estimación separada $\hat{Y}_{Rs}$ resultan de las de la Sec. 11.12 para un solo estrato, al suponer una muestra grande en cada estrato y muestreo independiente en diferentes estratos.

Para la estimación combinada, $\hat{Y}_{Rc} = X\hat{Y}_{st}/\hat{X}_{st}$,

$$\hat{Y}_{Rc} - Y = X \sum^L \hat{Y}_h / \sum^L \hat{X}_h - Y$$

$$= \frac{X}{\hat{X}_{st}} \sum^L (\hat{Y}_h - R\hat{X}_h) \doteq \sum^L (\hat{Y}_h - R\hat{X}_h)$$

Por lo tanto, sustituimos $d_{hij} = y_{hij} - Rx_{hij}$ en lugar de y_{hij} para obtener las fórmulas aproximadas de $V(\hat{Y}_{Rc})$ a partir de las corres-

pondientes para $V(\hat{Y}_{st})$ por el plan de muestreo usado. Para $v(\hat{Y}_{Rc})$ sustituimos $d'_{hij} = y_{hij} - \hat{R}_c x_{hij}$ en $v(\hat{Y}_{st})$.

Por ejemplo, con probabilidades desiguales con restitución (Sec. 11.9) la Fórmula (11.33) lleva a

$$V(\hat{Y}_{Rc}) \doteq \sum_h^L \frac{1}{n_h} \sum_i^{N_h} \left[z_{hi}\left(\frac{D_{hi}}{z_{hi}} - D_h\right)^2 + \frac{M_{hi}^2(1-f_{2hi})S_{d2hi}^2}{z_{hi}m_{hi}} \right] \qquad (11.75)$$

donde

$$D_{hi} = Y_{hi} - RX_{hi}$$

$$S_{d2hi}^2 = \frac{1}{(M_{hi}-1)} \sum_j^{M_{hi}} [(y_{hij} - Rx_{hij}) - (\bar{Y}_{hi} - R\bar{X}_{hi})]^2$$

Para la varianza estimada, la Fórmula (11.35) da, después de la sustitución,

$$v(\hat{Y}_{Rc}) \doteq \frac{\sum_h^L (D_{hi}' - \bar{D}_h')^2}{n_h(n_h-1)} \qquad (11.76)$$

donde

$$D_{hi}' = \frac{M_{hi}\bar{d}_{hi}'}{z_{hi}}, \qquad \bar{D}_h' = \frac{\sum_i^{n_h} D_{hi}'}{n_h}, \qquad \bar{d}_{hi}' = \bar{y}_{hi} - \hat{R}_c\bar{x}_{hi}$$

11.17 ESTIMADORES NO LINEALES EN ENCUESTAS COMPLEJAS

Además de la estimación de totales, de medias, razones y diferencias entre éstos, el análisis de los datos de la encuesta puede comprender estimaciones de una estructura matemática más compleja (como serían los coeficientes de correlación simple y parcial, las medianas y otros porcentiles). Los objetivos de los análisis pueden incluir la determinación de un intervalo de confianza a la cantidad estimada, o bien, la realización de una prueba de significación.

Dada una muestra aleatoria simple de una población infinita, la teoría estadística ha producido toda una variedad de métodos para lograr estos objetivos, métodos exactos de muestras pequeñas basados sobre algunos supuestos respecto a la naturaleza de las distribuciones que siguen las observaciones y métodos aproximados de grandes muestras que requieren menos hipótesis. Por otro lado, con los tipos más complejos de muestras estudiados en este libro, hemos podido dar métodos de cálculo de estimadores insesgados de las

varianzas de estimadores insesgados *lineales*, como $\hat{Y}_{st}$, $\hat{Y}_{ppz}$, o $\hat{Y}_{HT}$. Al suponer que la muestra es lo bastante grande para que estos estimadores tengan distribuciones cercanas a la normalidad, las tablas normales darán los límites de confianza y las pruebas de significación. Quedan muchos problemas pendientes para estimadores no lineales en encuestas complejas.

Se han introducido tres métodos apropiados para la estimación de los errores estándar de estimadores no lineales. Se describirán para muestras aleatorias estratificadas con $n_h = 2$ en todos los estratos, caso en el que se han estudiado más profusamente. Todos los métodos dan las estimaciones insesgadas usuales de la varianza, cuando el estimador es lineal. La muestra puede ser de varias etapas con unidades primarias extraídas con probabilidades iguales o con probabilidades desiguales con restitución.

11.18 DESARROLLO EN SERIES DE TAYLOR

Este es el método que dio las fórmulas aproximadas para las varianzas estimadas de $\hat{R}$, y $\hat{R}_c$, $\hat{R}_s$ en muestreo estratificado. La función que se va a estimar, $f(Y_1, Y_2, \ldots, Y_k) = f(\mathbf{Y})$ se expresa como una función de los totales de población de ciertas variables, al ser la estimación $f(\hat{\mathbf{Y}})$ la función correspondiente de las estimaciones insesgadas de muestra de $\hat{\mathbf{Y}}$. Para j variables con $n_h = 2$, cuando las unidades se eligen por el método *ppz* con restitución, el total de población y su estimador lineal son

$$Y_j = \sum_h^L \sum_i^{N_h} Y_{jhi}; \qquad \hat{Y}_j = \sum_h^L \sum_i^2 \frac{\hat{Y}_{jhi}}{2z_{hi}} = \sum_h^L (y'_{jh1} + y'_{jh2}) \qquad (11.77)$$

donde $y'_{jhi} = \hat{Y}_{jhi}/2z_{hi}$ y $\hat{Y}_{jhi}$ es una estimación muestral insesgada de Y_{jhi} a partir de la submuestra en esta unidad.

De (11.35) una estimación insesgada de $V(\hat{Y}_j)$ viene a ser

$$v(\hat{Y}_j) = \sum_h \sum_i^2 2(y'_{jhi} - \bar{y}'_{jhi})^2 \equiv \sum_h (y'_{jh1} - y'_{jh2})^2 \qquad (11.78)$$

De modo que, por la extensión de Woodruff (1971) en la Sec. (6.13) del método abreviado de Keyfitz para $n_h = 2$, una aproximación en serie de Taylor a la varianza de $f(\hat{\mathbf{Y}})$ es

$$v[f(\hat{\mathbf{Y}})] \doteq \sum_h \left[\sum_j^k \left(\frac{\partial f}{\partial \hat{Y}_j} \right) (y'_{jh1} - y'_{jh2}) \right]^2 \qquad (11.79)$$

La expresión de una función no lineal de las variables medidas en la forma $f(\mathbf{Y})$ puede requerir algún trabajo y precauciones que no

son posibles con algunas funciones de interés. Consideremos un ejemplo simple, el del muestreo aleatorio simple de una etapa con unidades primarias de tamaños iguales, donde debe ser estimada la correlación entre los totales de unidades de dos variables U_1 y U_2. El valor de población $f(\mathbf{Y})$ es

$$\rho = \frac{\sum_h^L \sum_i^N U_{1hi}U_{2hi} - \frac{\left(\sum_h \sum_i U_{1hi}\right)\left(\sum_h \sum_i U_{2hi}\right)}{N}}{\left[\sum_h \sum_i U_{1hi}^2 - \frac{\left(\sum_h \sum_i U_{1hi}\right)^2}{N}\right]^{1/2} \left[\sum_h \sum_i U_{2hi}^2 - \frac{\left(\sum_h \sum_i U_{2hi}\right)^2}{N}\right]^{1/2}}$$

En términos de nuestras variables Y_{jhi}, ésta es una función de cinco de estas variables

$$Y_{1hi} = U_{1hi\ \ 2hi}; \qquad Y_{2hi} = U_{1hi}^2; \qquad Y_{3hi} = U_{2hi}^2$$
$$Y_{4hi} = U_{1hi}; \qquad Y_{5hi} = U_{2hi}$$

con $y'_{hi} = \hat{U}_{1hi}\hat{U}_{2hi}/2z_{hi}$, $i = 1, 2$, $z_{hi} = 1/N_h$ en este caso.

11.19 REPLICAS REPETIDAS EQUILIBRADAS

Inicialmente, consideremos este método para una función lineal de una sola variable. Sea $f(Y) = Y$, con su estimación $f(\hat{Y}) = \hat{Y} = \sum_h (y_{h1}' + y_{h2}')$, donde $y_{hi}' = \hat{Y}_{hi}/2z_{hi}$. Una estimación insesgada de la varianza $v(\hat{Y})$ es, de (11.78)

$$v(\hat{Y}) = \sum_h (y_{h1}' - y_{h2}')^2 \tag{11.80}$$

Tómese una media muestra H, al elegir una unidad de cada estrato. La estimación de $f(Y) = Y$ de esta media muestra es $f(H) = 2\sum_h y_{hi}'$, donde hi es la unidad elegida en el estrato h. Por lo tanto, si $f(S)$ denota la estimación de $f(Y)$ hecha a partir de toda la muestra,

$$f(H) - f(S) = 2\sum_h y_{hi}' - \sum_h (y_{h1}' + y_{h2}') = \sum_h \pm (y_{h1}' - y_{h2}') \tag{11.81}$$

y los signos dependen de si se elige la unidad 1 o 2 en el estrato h.

Al comparar (11.81) con (11.80), se ve que para cualquiera de las 2^L posibles selecciones, $[f(H) - f(S)]^2$ contiene los términos cuadrados correctos $(y_{h1}' - y_{h2}')^2$ en $v(\hat{Y})$ en cada estrato. Sin embargo, para cada media muestra específica, cada par de estratos contribuye con un término cruzado que lleva un signo $\pm$ para

$[f(H) - f(S)]^2$. McCarthy (1966) observó que se puede encontrar un conjunto de medias muestras equilibradas en las que, para cada par de estratos, la mitad de los signos correspondientes a los productos cruzados del conjunto son + y la otra mitad −. Si este conjunto contiene g diferentes medias muestras, los términos de producto cruzado no deseados se cancelan al sumar sobre el conjunto, lo que da

$$\frac{1}{g}\sum_i^g [f(H_i) - f(S)]^2 \equiv \sum_h^L (y_{h1}' - y_{h2}')^2 = v(\hat{Y}) \tag{11.82}$$

Si C_i es la media muestra complementaria de H_i, tenemos como segunda y tercera formas

$$v(\hat{Y}) \equiv \frac{1}{g}\sum_i^g [f(C_i) - f(S)]^2 \equiv \frac{1}{4g}\sum_i^g [f(H_i) - f(C_i)]^2 \tag{11.83}$$

Un cuarto estimador es el promedio de los dos primeros anteriores.

Los conjuntos equilibrados de este tipo fueron propuestos anteriormente por Plackett y Burman (1946) para usarlos en un diseño experimental para el que g es cualquier múltiplo de 4. Para L estratos, el conjunto equilibrado más pequeño tiene g iguales al más pequeño múltiplo de 4 superior a L. Con $L = 5$ estratos, $g = 8$. La Tabla 11.7 muestra un conjunto equilibrado para $L = 5$, el signo + denota aquí la unidad 1, y el − denota la unidad 2 en un estrato.

TABLA 11.7 MEDIAS MUESTRAS BALANCEADAS PARA $L = 5$ ESTRATOS

	Media Muestra							
Estrato	1	2	3	4	5	6	7	8
1	+	+	+	−	+	−	−	−
2	−	+	+	+	−	+	−	−
3	−	−	+	+	+	−	+	−
4	+	−	−	+	+	+	−	−
5	−	+	−	−	+	+	+	−

Cuando este método de réplica repetida equilibrada (BRR) se aplica a una función no lineal $f(\mathbf{Y})$ de varias variables, los cuatro estimadores

$$\frac{1}{g}\sum_i^g [f(H_i) - f(S)]^2; \qquad \frac{1}{g}\sum_i^g [f(C_i) - f(S)]^2; \qquad \frac{1}{4g}\sum_i^g [f(H_i) - f(C_i)]^2$$

y el promedio de los dos primeros, todos difieren hasta cierto punto. También $f(S) = f(\hat{\mathbf{Y}})$ es sesgado y las estimaciones de varianza no

incluyen la contribución correcta del sesgo al ECM$[f(\hat{\mathbf{Y}})]$. Sin embargo, hay que observar que si se extraen m muestras independientes, S_i, cada una con una complementaria H_i, C_i, la cantidad

$$\frac{1}{2m}\sum_{i}^{m}[f(H_i)-f(C_i)]^2 \tag{11.84}$$

es una estimación insesgada de $V[f(H)]$. Por lo tanto, dicho burdamente, el método *BRR* equivale a suponer para $f(S)$, que: (a) su sesgo es despreciable, (b) $V[f(S)] \doteq (1/2)V[f(H)]$, y (c) el cálculo de $v[f(S)]$ a partir de *BRR* para una *sola* S estratificada concuerda aceptablemente con su cálculo a partir de muestras independientes S_i para considerarla útil.

El método de las réplicas repetidas, se ha usado extensamente, tanto para la selección de muestras y la estimación de la varianza por el U.S. Census Bureau en la década de los 50, como por Deming (1956-1960). La idea proviene de una técnica de Mahalanobis de las submuestras interpenetrantes (Mahalanobis, 1944). McCarthy (1969) pasa revista a las propiedades del método *BRR* con una discusión posterior de Kish y Frankel (1974).

11.20 EL METODO JACKNIFE

Para $\hat{R}$ en muestras aleatorias simples, se describió este método en la Sec. 6.17. Fue propuesto como un medio de estimación de $V(\hat{R}_Q)$, pero también pudo proponerse como estimación de $V(\hat{R})$. Omítase la unidad j-ésima de la muestra y calcúlese $\hat{R}_j$, la estimación de razón para el resto de la muestra. Al ignorar la cpf, se puede obtener un estimador aproximado de $V(\hat{R})$ a partir de (6.86) con $g = n$. Si suponemos $\hat{R} \doteq \hat{R}_{\cdot}$, la media de las $\hat{R}_j$, este estimador se convierte en

$$v(\hat{R}) \doteq \frac{(n-1)}{n}\sum_{i}^{n}(\hat{R}_j-\hat{R})^2 = \frac{(n-1)}{n}[f(S_j)-f(S)]^2 \tag{11.85}$$

la expresión de la extrema derecha indica cómo se puede aplicar el método a otras estimaciones no lineales.

Para un estimador lineal, como $\bar{y}$, es fácil verificar que la Fórmula (11.85) se reduce a la usual $v(\bar{y}) = \sum(y_i-\bar{y})^2/n(n-1)$.

Al extender este método a las muestras estratificadas con $n_h = 2$, Frankel (1971) sugiere que se omita una unidad al azar del estrato h para $h = 1, 2, \ldots, L$ en turno y se calcule $f(S_h)$ de la muestra

restante de tamaño $(2L - 1)$. Una forma de la estimación Jacknife de $V[f(S)]$ es entonces

$$v[f(S)] = \sum_{h}^{L}[f(S_h)-f(S)]^2 \tag{11.86}$$

Para un estimador lineal $f(\hat{\mathbf{Y}})$, este estimador de varianza se reduce al estimador insesgado usual en muestreo *ppz* con restitución.

Al igual que con *BRR*, existen cuatro versiones análogas del estimador Jacknife de $v[f(S)]$.

11.21 COMPARACION DE LOS TRES METODOS

Los métodos de Taylor, *BRR* y *J* han recibido una justificación teórica rigurosa muy limitada. Para evaluar su comportamiento con diferentes estimadores y tipos de encuestas, se han utilizado principalmente como base los estudios de Montecarlo. Un estudio tal lo describe Frankel (1971) y Kish y Frankel (1974).

La muestra era de una etapa con unidades conglomeradas de tamaños ligeramente diferentes, que tienen 14.1 viviendas en promedio. La población de $N = 3\,240$ unidades o 45 737 viviendas se dividió en turno en 6, 12 y 30 estratos del mismo tamaño con $n_h = 2$ en cada estrato. Se usó el muestreo aleatorio simple dentro de los estratos. Los datos se tomaron de la encuesta de población Current Population Survey (U.S. Census Bureau). Las ocho variables originales para una casa eran: el número de personas, el número de ellas con menos de 18 años de edad, el número de ellas en la fuerza de trabajo y el ingreso total, y para la cabeza de familia, el ingreso, la edad, el sexo y los años de educación. Se calcularon varias medias (razones), diferencias de medias y coeficientes de correlación simple, parcial y múltiple (tres variables independientes). En números de unidades primarias, las muestras eran muy pequeñas, 12, 24 y 60, con alrededor de 170, 340 y 850 como números de casas.

Al calcular los estimadores de varianza, se compararon las cuatro formas del método *BRR* y el método *J*. El método de Taylor no se usó para regresiones múltiples y parciales, por la dificultad en expresar las derivadas parciales $\partial f/\partial Y_j$ en una forma manipulable. Puesto que el método de muestreo era la estratificación proporcional sin restitución, el término cpf $(1 - f)$ se incluyó en todas las estimaciones de varianza con $f = n/N$, aunque se necesitaron poco con estas muestras pequeñas.

Salvo los coeficientes de correlación múltiple, los estimadores $f(\mathbf{Y})$ presentaban sesgos de relativa poca importancia para los tres tamaños de muestras, al ser las razones |sesgo|/error estándar inferiores a < 0.1, para razones, diferencias de razones y coeficientes de regresión simple, y alrededor de 0.2 para los coeficientes de correlación simple y parcial. Cuando las $v(\hat{\mathbf{Y}})$ aproximadas se consideran como estimadores de ECM($\hat{\mathbf{Y}}$), los tres métodos Taylor, *BRR* y *J* se comportaban bien para razones y diferencias de razones con el promedio de sesgo en $v(\hat{\mathbf{Y}})|/\text{ECM}(\hat{\mathbf{Y}})$ inferior a 5%. Para coeficientes de correlación simples los estimadores de varianza *BRR* presentaban sesgos sustancialmente más bajos que los estimadores *T* o *J*. Lo contrario era cierto para coeficientes de regresión simple. Con *BRR* y *J* el mejor de los cuatro métodos $v(\hat{Y})$ era el promedio de las estimaciones *H* y *C* (esto es, con *BRR*).

$$\frac{1}{2g}\left\{\sum_{i}^{g}[f(H_i)-f(S)]^2+\sum_{i}^{g}[f(C_i)-f(S)]^2\right\} \qquad (11.87)$$

En lo que respecta a los enunciados de intervalos de confianza y las pruebas de significación, lo importante era saber qué tan bien concordaban las colas de las distribuciones de probabilidad de la variable $[f(\hat{\mathbf{Y}}) - f(\mathbf{Y})]$/e.s. $[f(\hat{\mathbf{Y}})]$ de la *t* de Student con *L g.l.* Para valores elegidos de *t*, Frankel (1971) da las estimaciones de Montecarlo de las frecuencias para la variable relacionada $[f(\hat{\mathbf{Y}}) - Ef(\hat{\mathbf{Y}})]$/e.s.$[f(\hat{\mathbf{Y}})]$. Algunos resultados para 6 y 30 estratos se muestran en la Tabla 11.8 para los valores de *t* que presenta Frankel cerca de los niveles en ambas colas del 5 y 10%. La versión (11.87) de los métodos *BRR* y *J* es la que se muestra.

En resumen, el método *BRR* tiene un comportamiento consistentemente mejor en la Tabla 11.8. Salvo para las *R* múltiples, se puede considerar adecuado para el uso práctico desde el punto de vista de que en el análisis de datos, un valor de la cola del 5% en la Tabla representa un valor real entre el 3 y el 8%. El método *J* es un poco mejor que el método de Taylor. Salvo *BRR* con razones y regresiones simples, todos los métodos dan frecuencias reales en la cola, más elevadas que las de las tablas *t*, por lo tanto, las probabilidades de confianza son altas. Un aspecto que intriga es que, para los coeficientes de correlación, el aumento del tamaño de muestra de 12 a 60 no produjo una mejoría correspondiente en la aproximación de las frecuencias reales a las colas de la distribución *t*.

TABLA 11.8 PROMEDIO DE PROBABILIDADES EN LAS COLAS DE LAS DISTRIBUCIONES PARA $[f(\hat{\mathbf{Y}}) - Ef(\hat{\mathbf{Y}})]$/e.s.$[f(\hat{\mathbf{Y}})]$ COMPARADOS CON LOS DE LA *t* DE STUDENT

	6 estratos					
	$P(t) = .042$, $t = 2.576$			$P(t) = .098$, $t = 1.960$		
	BRR	*J*	Taylor	*BRR*	*J*	Taylor
Razones	.044	.049	.052	.096	.106	.112
[*a] *b*	.034	.048	.058	.085	.117	.127
[b] *r*	.052	.069	.084	.114	.137	.163
r parciales	.043	.063	—	.092	.132	—
R múltiples	.065	.088	—	.105	.160	—

	30 estratos					
	$P(t) = .059$, $t = 1.960$			$P(t) = .110$, $t = 1.645$		
	BRR	*J*	Taylor	*BRR*	*J*	Taylor
Razones	.056	.057	.057	.109	.111	.112
b	.062	.067	.068	.110	.116	.116
r	.089	.098	.102	.138	.153	.164
r parciales	.103	.121	—	.156	.181	—
R múltiples	.175	.207	—	.265	.297	—

[*a] *b* = coeficiente de regresión simple.
[b] *r* = coeficiente de correlación simple.

Este estudio abre un vasto campo de investigación de los métodos, con diferentes planes de encuesta y diferentes tipos de estimador $f(\mathbf{Y})$.

En un estudio de Montecarlo y una muestra más compleja (muestreo de dos etapas con probabilidades proporcionales al tamaño, con restitución incluyendo estratificación y postestratificación) Bean (1975) comparó los métodos de Taylor y *BRR* para estimadores del tipo de razón. Ambos métodos dieron estimaciones satisfactorias de la varianza y probabilidades de confianza bilaterales satisfactorias, calculadas a partir de la distribución normal, sin embargo, aún existe bastante asimetría y por lo tanto, no se puede confiar en los intervalos de confianza unilaterales.

EJERCICIOS

11.1. Obteniendo las estimaciones para todas las muestras posibles que se pueden extraer de la población artificial de la Tabla 11.1, por los Métodos I*a*, I*b*, II y III, verificar el ECM total dado en la Tabla 11.2.

11.2. Para los Métodos II (probabilidades iguales, estimación insesgada) y III (selección pps) volver a calcular las variancias de $\hat{\bar{Y}}$ para el ejemplo en la Tabla 11.1, cuando $m_i = 1$. Mostrar que la precisión del Método III relativamente al Método II es más baja para $m_i = 1$ que para $m_i = 2$. ¿Cuál es el resultado general que ilustra esto?

11.3. Para la población de la Tabla 11.1, si los tamaños estimados z_i son 0.1, 0.3 y 0.6 con $m_i = 2$, mostrar que la estimación insesgada (Método IV) da una varianza más pequeña que el muestreo *pps*. ¿Cómo se explica este resultado?

11.4. Los elementos de la población que tienen tres unidades primarias, se clasifican en dos clases. Los tamaños de unidades M_i y las proporciones P_i de elementos que pertenecen a la primera clase son como sigue:

$$M_1 = 100, \quad M_2 = 200, \quad M_3 = 300, \qquad P_1 = 0.40, \quad P_2 = 0.45, \quad P_3 = 0.35$$

Para una muestra de 50 elementos de una unidad primaria, comparar el ECM de los métodos I*a*, II y III para estimación de la proporción de elementos de la primera clase en la población. (En las fórmulas de la variancia de la Sec. 11.2, S_i^2 es aproximadamente P_iQ_i.)

11.5. Una muestra de n unidades primarias es seleccionada con probabilidades iguales, de cada unidad elegida se toma una fracción constante de subunidades, f_2. Si a_i de las m_i subunidades en la unidad i-ésima caen en la clase C, muéstrese que la estimación razón al tamaño (Sec. 11.8) de la proporción de población en la clase C es $\bar{p} = \Sigma a_i / \Sigma m_i$. De la Fórmula 11.36, mostrar que la siguiente es una estimación de ECM($\bar{p}$) es

$$v(\bar{p}) = \frac{1-f_1}{n\bar{M}^2}\frac{\sum^n M_i^2(p_i-\bar{p})^2}{n-1} + \frac{f_1(1-f_2)}{n^2\bar{m}\bar{M}}\sum^n \frac{M_i m_i}{m_1 - 1} p_i q_i$$

donde $p_i = a_i/m_i$.

11.6. Una firma que tiene 36 fábricas decide verificar la condición de cierto tipo de equipo del cual hay $M_0 = 25\,012$ piezas en uso. Una muestra aleatoria de 12 fábricas es la que se toma y después se verifica una submuestra del 10% en cada una de las fábricas seleccionadas. Los números de piezas examinados (m_i) y los números encontrados de piezas con alguna deterioración (a_i) son como sigue:

Fábrica	m_i	a_i	$p_i = \frac{a_i}{m_i}$	Fábrica	m_i	a_i	$p_i = \frac{a_i}{m_i}$
1	65	8	0.123	7	85	18	0.212
2	82	21	0.256	8	73	11	0.151
3	52	4	0.077	9	50	7	0.140
4	91	12	0.132	10	76	9	0.118
5	62	1	0.016	11	64	20	0.312
6	69	3	0.043	12	50	2	0.040

Estimar el porcentaje y número total de piezas defectuosas en uso y estimar sus errores estándar.

Nota. Ya que $M_i/\bar{M} \doteq m_i/\bar{m}_s$, la componente entre unidades de $v(\bar{p})$ puede calcularse como

$$\frac{1-f_1}{n\bar{m}^2(n-1)}\left(\sum a_i^2 - 2\bar{p}\sum a_i m_i + \bar{p}^2 \sum m_i^2\right)$$

y dado que las m_i son bastante grandes, la componente dentro de unidades se calcula como

$$\frac{f_1(1-f_2)}{(n\bar{m})^2}\sum a_i q_i$$

11.7. Si se seleccionan unidades primarias con probabilidades iguales y f_2 es constante, mostrar que en la notación del Ejercicio 11.5 la estimación insesgada de una proporción de población es $p = N\Sigma a_i / nM_0 f_2$ y que si los términos en $1/m_i$ pueden despreciarse, su varianza se puede calcular como

$$v(p) = \frac{1-f_1}{n(n-1)\bar{m}^2}\sum^n (a_i - \bar{a})^2 + \frac{f_1(1-f_2)}{(n\bar{m})^2}\sum^n a_i q_i$$

calcular p y su error estándar para los datos del Ejercicio 11.6.

11.8. Una muestra de n unidades primarias se elige con probabilidades proporcionales a los tamaños estimados z_i (con restitución) y con una fracción constante esperada global de muestreo, f_0. Mostrar que las estimaciones insesgadas y de razón al tamaño del total de población son respectivamente T/f_0 y $TM_o/\sum^n m_i$, donde T es el total de muestra. (Se sigue que si M_0 es desconocida, puede usarse la estimación insesgada, pero no la de razón al tamaño. Para estimar la media de población por subunidad, se invierte la situación.)

11.9. En un estudio de sobrepoblación en una gran ciudad un estrato contenía 100 manzanas de las cuales se tomaron 10 con probabilidades proporcionales al tamaño estimado (con restitución). Se usó como fracción de muestreo esperado global $f_0 = 2\%$. Estimar el número total y el número promedio de personas por habitación, y sus errores estándar a partir de los datos que siguen.

Manzana	1	2	3	4	5	6	7	8	9	10
Cuartos	60	52	58	56	62	51	72	48	71	58
Personas	115	80	82	93	105	109	130	93	109	95

11.10. Para el método de Durbin (Sec. 11.10) en que se simplifica la estimación de varianza en muestreo con probabilidades proporcionales a z sin restitución, un método simple de selección de muestra debido esencialmente a Kish (1965) es el siguiente. El subíndice h que denota el estrato se omite y el número de unidades primarias se supone que es par.

Se arreglan las unidades en orden de z_i creciente y se marcan por pares. El método es exacto solamente si $z_i = z_j$ para miembros del mismo par; esto es lo que se supondrá aquí. Se seleccionan dos unidades *ppz con restitución*. Si se sacan dos unidades diferentes, se aceptan. Si se saca la misma unidad dos veces, se hace que la muestra consista en los dos miembros del par al que pertenece esta unidad. Muéstrese que para este método: (*a*) $\pi_i = 2z_i$, (*b*)

para unidades que no están en el mismo par, $\pi_{ij} = 2z_i z_j = \pi_i \pi_j / 2$, de modo que $\pi_i \pi_j \pi_{ij}^{-1} - 1 = 1$, y (c) para unidades en el mismo par $\pi_{ij} = 4z_i z_j = \pi_i \pi_j$, de modo que $\pi_i \pi_j \pi_{ij}^{-1} - 1 = 0$.

11.11. En la Sec. 11.9, la Fórmula (11.33) para $V(\hat{Y}_{ppz})$ en muestreo con restitución se probó con base en que siempre que se seleccione la i-ésima unidad, se extraiga de toda la muestra una submuestra aleatoria simple independiente de tamaño m_i. Demostrar los siguientes resultados para dos planes alternativos.

(a) Cuando se selecciona la unidad i-ésima t_i veces, se saca una submuestra aleatoria simple de tamaño $m_i t_i$ de la misma (supóngase $m_i t_i \leq M_i$). En este plan, $V(\hat{Y}_{ppz})$ en (11.41) se reduce en la cantidad $(n-1)\sum^{N} M_i S_{2i}^2/n$ (Sukhatme, 1954).

(b) Cuando se selecciona la i-ésima unidad t_i veces, se toma una muestra aleatoria simple de tamaño m_i. Entonces $V(\hat{Y}_{ppz})$ en (11.41) aumenta en la cantidad

$$\frac{(n-1)}{n}\sum^{N} M_i^2(1-f_{2i})S_{2i}^2/m_i$$

tanto en (a) como en (b), $\hat{Y}_{ppz} = \sum^{N} t_i M_i \bar{y}_i / nz_i$, al recibir la unidad i-ésima asignada el peso t_i.

CAPITULO 12

Muestreo Doble

12.1 DESCRIPCION DE LA TECNICA

Como hemos visto, varias técnicas de muestreo dependen de si se tiene información anticipada respecto a una variable auxiliar x_i. Las estimaciones de razón y regresión requieren que se conozca la media de población $\bar{X}$. Si se desea estratificar la población respecto a los valores de x_i, debe conocerse su distribución de frecuencia.

Cuando falta esta información, es relativamente económico tomar una gran muestra preliminar, sólo para medir x_i. El propósito de esta muestra es obtener una buena estimación de $\bar{X}$ o de la distribución de frecuencia de x_i. En una encuesta que tiene el propósito de hacer estimaciones para otra variable y_i, puede ser conveniente dedicar parte de los recursos a esta muestra preliminar, aunque esto significa que deberá disminuir el tamaño de la muestra en la encuesta principal para y_i. Esta técnica se conoce como *muestreo doble* o *muestreo de dos fases*. Como lo implica la discusión, la técnica es provechosa sólo si lo que se gana en precisión a partir de estimaciones de razón o regresión, o bien, de la estratificación, hace más que compensar la pérdida en precisión debida a la reducción del tamaño de la muestra principal.

El muestreo doble puede ser apropiado cuando la información respecto a x_i está contenida en tarjetas no tabuladas. Así, por ejemplo, en encuestas de la población civil alemana en 1945, la muestra en cualquier ciudad se extraía generalmente de las listas de registro de racionamiento. Además de la estratificación geográfica dentro de la ciudad, para la que ya se tenían los datos, se propuso la estratificación por sexo y edad. Como la muestra debía extraerse apresuradamente y como las listas se necesitaban constantemente, no era factible la tabulación de la distribución completa de edades y sexos.

Pero una muestra sistemática, moderadamente grande, sí podía extraerse de prisa. Cada persona seleccionada se clasificaba en la clase apropiada de edad-sexo. A partir de estos datos, se seleccionó una lista mucho más pequeña de personas para ser entrevistadas.

12.2 MUESTREO DOBLE PARA ESTRATIFICACION

La teoría fue dada inicialmente por Neyman (1938). La población debe estratificarse en L clases (estratos). La primera muestra es una muestra aleatoria simple de tamaño n'.
Sea

$W_h = N_h/N$ = proporción de la población que cae en el estrato h.

$w_h = n_h'/n'$ = proporción de la primera muestra que cae en el estrato h.
Entonces, w_h es una estimación insesgada de W_h.

La segunda muestra es una muestra aleatoria estratificada de tamaño n en donde las y_{hi} son las variables a medir: se toman n_h unidades del estrato h. Usualmente, la segunda muestra en el estrato h es una submuestra aleatoria de los n_h' elementos en el estrato. El objeto de la primera muestra es la estimación de los pesos de esrtatos; el de la segunda muestra es la estimación de las medias de estratos $\bar{Y}_h$.

La media de población $\bar{Y} = \sum W_h \bar{Y}_h$. Como estimación usamos

$$\bar{y}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h \tag{12.1}$$

El problema es el de elegir n' y los n_h para minimizar $V(\bar{y}_{st})$ a un costo dado.

Debemos entonces verificar si la varianza mínima es menor que la que puede obtenerse por una muestra aleatoria simple, en donde se mide y_i sola. Al presentar la teoría, suponemos que los n_h son una submuestra aleatoria de los n_h'. Por lo tanto, $n_h = v_h n_h'$ donde $0 < v_h \leq 1$ y los v_h se eligen por adelantado. El muestreo repetido implica una nueva extracción de la primera y la segunda muestra para que las w_h, n_h y $\bar{y}_h$ sean todas variables aleatorias. Así, la cuestión es un problema de estratificación en el cual no se conocen exactamente los tamaños de estratos (Sec. 5A.2).

Por sencillez se harán dos aproximaciones. El primer tamaño de muestra n' se supone lo suficientemente grande para que $w_h > 0$. En segundo lugar, cuando lleguemos a la discusión de la estrategia óptima, cada v_h óptima, como la de la fórmula, se supone ≤ 1.

Teorema 12.1. La estimación $\bar{y}_{st}$ es insesgada.

Prueba. Primero se promedia sobre las muestras donde los w_h son fijos. Dado que $\bar{y}_h$ es la media de una muestra aleatoria simple del estrato, $E(\bar{y}_h) = \bar{Y}_h$. Además, cuando promediamos sobre selecciones diferentes de la primera muestra, $E(w_h) = W_h$, puesto que la primera muestra es en sí misma una muestra aleatoria simple. En estas condiciones

$$E(\bar{y}_{st}) = E[E(\sum w_h \bar{y}_h | w_h)] = E(\sum w_h \bar{Y}_h) = \sum W_h \bar{Y}_h = \bar{Y} \tag{12.2}$$

Teorema 12.2. Si la primera muestra es aleatoria y de tamaño n', la segunda muestra es una submuestra aleatoria de la primera, de tamaño $n_h = v_h n_h'$, donde $0 < v_h \leq 1$ y los v_h son fijos

$$V(\bar{y}_{st}) = S^2\left(\frac{1}{n'} - \frac{1}{N}\right) + \sum_h^L \frac{W_h S_h^2}{n'}\left(\frac{1}{v_h} - 1\right) \tag{12.3}$$

donde S^2 es la varianza de población.

Prueba. La prueba se obtiene fácilmente con el siguiente artificio. Supóngase que las y_{hi} se midieron en todas las unidades n_h' de la primera muestra en el estrato h y no sólo sobre una submuestra aleatoria de n_h. Entonces, como $w_h = n_h'/n'$,

$$\sum_h^L w_h \bar{y}_h' = \bar{y}'$$

es la media de una muestra aleatoria simple de tamaño n' de la población. En esta forma, el promediar sobre selecciones repetidas de la muestra de tamaño n',

$$V\left(\sum_h^L w_h \bar{y}_h'\right) = S^2\left(\frac{1}{n'} - \frac{1}{N}\right) \tag{12.4}$$

pero

$$\bar{y}_{st} = \sum_h^L w_h \bar{y}_h = \sum_h^L w_h \bar{y}_h' + \sum_h^L w_h (\bar{y}_h - \bar{y}_h') \tag{12.5}$$

Sea el subíndice 2 el que se refiere a un promedio global sobre todas las submuestras aleatorias de n_h unidades que pueden extraerse de n_h' unidades dadas. Es claro que $E_2(\bar{y}_h) = \bar{y}_h'$. Los resultados siguientes se infieren inmediatamente (ver Ejercicio 2.16):

$$\begin{gathered} \operatorname{cov}[\bar{y}_h', (\bar{y}_h - \bar{y}_h')] = 0: \\ \operatorname{cov}(\bar{y}_h', \bar{y}_h) = V(\bar{y}_h'): \quad V(\bar{y}_h - \bar{y}_h') = V(\bar{y}_h) - V(\bar{y}_h') \end{gathered} \tag{12.6}$$

luego, para w_h fijo,

$$V_2\left[\sum w_h(\bar{y}_h-\bar{y}_h')\right]=\Big]=\sum w_h^2 S_h^2\left(\frac{1}{n_h}-\frac{1}{n_h'}\right)=\sum\frac{w_h S_h^2}{n'}\left(\frac{1}{\nu_h}-1\right) \quad (12.7)$$

puesto que $n_h=\nu_h n_h'=\nu_h w_h n'$.

Al promediar sobre la distribución de los w_h obtenida por selecciones repetidas de la primera muestra, tenemos, de (12.4) (12.5) y (12.7),

$$V(\bar{y}_{st})=S^2\left(\frac{1}{n'}-\frac{1}{N}\right)+\sum_h^L\frac{W_h S_h^2}{n'}\left(\frac{1}{\nu_h}-1\right) \quad (12.8)$$

Corolario 1. El resultado se puede expresar en varias formas diferentes. Por el análisis de la varianza,

$$(N-1)S^2=\sum(N_h-1)S_h^2+\sum N_h(\bar{Y}_h-\bar{Y})^2 \quad (12.9)$$

Por lo tanto, si $g'=(N-n')/(N-1)$, al multiplicar por $g'/n'N$ se obtiene

$$\frac{(N-n')S^2}{n'N}=S^2\left(\frac{1}{n'}-\frac{1}{N}\right)=\frac{g'}{n'}\sum(W_h-N^{-1})S_h^2+\frac{g'}{n'}\sum W_h(\bar{Y}_h-\bar{Y})^2 \quad (12.10)$$

De (12.3) esto da

$$V(\bar{y}_{st})=\sum_h^L\frac{W_h S_h^2}{n'}\left(\frac{1}{\nu_h}-1\right)+\frac{g'}{n'}\sum_h^L(W_h-N^{-1})S_h^2+\frac{g'}{n'}\sum_h^L W_h(\bar{Y}_h-\bar{Y})^2 \quad (12.11)$$

Además, por la definición de $g'=(N-n')/(N-1)$, resulta que

$$-\frac{1}{n'}+\frac{g'}{n'}=-\frac{1}{N}+\frac{g'}{n'N} \quad (12.12)$$

En consecuencia (12.11), los términos segundo y tercero en $\sum W_h S_h^2$, que tienen coeficientes $-1/n'$ y g'/n' se pueden escribir alternativamente, lo que da

$$V(\bar{y}_{st})=\sum_h W_h S_h^2\left(\frac{1}{n'\nu_h}-\frac{1}{N}\right)+\frac{g'}{n'N}\sum_h(W_h-1)S_h^2+\frac{g'}{n'}\sum W_h(\bar{Y}_h-\bar{Y})^2 \quad (12.13)$$

Si la clasificación no es costosa, puede no ser razonable suponer n'/N despreciable, dado que una proporción importante de la población se puede clasificar. Sin embargo, para la mayoría de las aplicaciones, podrá ignorarse el término $g'/n'N$ en (12.13). Y en este caso, (12.13) se simplifica a

$$V(\bar{y}_{st})\doteq\sum_h^L W_h S_h^2\left(\frac{1}{n'\nu_h}-\frac{1}{N}\right)+\frac{g'}{n'}\sum_h^L W_h(\bar{Y}_h-\bar{Y})^2 \quad (12.14)$$

Los resultados en el Teorema (12.2) fueron dados por Rao (1973).

Corolario 2. Si lo que se estima es una proporción en la segunda muestra, las expresiones; $S^2=NPQ/(N-1)$ y

$$S_h^2=\frac{N_h P_h Q_h}{(N_h-1)}; \qquad (\bar{Y}_h-\bar{Y})^2=(P_h-P)^2$$

se sustituyen en (12.3), (12.11) y (12.13). Tenemos en (12.14), para $n'\nu_h/N$ también despreciable

$$V(p_{st})\doteq\sum_h^L\frac{W_h P_h Q_h}{n'\nu_h}+\frac{g'}{n'}\sum_h^L W_h(P_h-P)^2 \quad (12.15)$$

Corolario 3. Los resultados para el caso en el cual la segunda muestra se toma *independientemente* de la primera, de modo que los n_h no dependan de los n_h' (excepto en lo que respecta el supuesto $n_h\le n_h'$), se dieron en la segunda edición Pág. 401 para n_h/N_h despreciable, el término guía de la varianza tiene la misma estructura que (12.14), y es

$$V(\bar{y}_{st})\doteq\sum_h^L\frac{W_h^2 S_h^2}{n_h}+\frac{g'}{n'}\sum_h^L W_h(\bar{Y}_h-\bar{Y})^2 \quad (12.16)$$

Robson (1952) y Robson y King (1953) escribieron artículos en donde extienden la teoría de la estratificación al muestreo de dos etapas y la aplican a la estimación de lectura de revistas.

12.3 ASIGNACION OPTIMA

El objetivo es elegir los n' y ν_h para minimizar $V(\bar{y}_{st})$ a un costo específico. Sea c' el costo de la clasificación por unidad y c_h el costo de medir una unidad en el estrato h. Para una muestra específica

$$C=c'n'+\sum_h c_h n_h \quad (12.17)$$

Dado que los n_h son variables aleatorias, minimizamos el costo esperado para n' y ν_h elegidas

$$E(C)=C^*=c'n'+n'\sum c_h\nu_h W_h \quad (12.18)$$

Para $V=V(\bar{y}_{st})$, la Fórmula (12.3) conduce a

$$n'(V+S^2/N)=(S^2-\sum W_h S_h^2)+\sum_h\frac{W_h S_h^2}{\nu_h} \quad (12.19)$$

El producto $C^*(V+S^2/N)$ no involucra n'. Al aplicar la desigualdad de Cauchy-Schwarz a este producto, se muestra que el producto se minimiza, si para cada h,

$$\frac{\nu_h^2 c_h}{S_h^2} = \frac{c'}{(S^2 - \Sigma W_h S_h^2)} \tag{12.20}$$

Esto da

$$\nu_h = S_h[c'/c_h(S^2 - \Sigma W_h S_h^2)]^{1/2} \tag{12.21}$$

El valor de n' se obtiene de la ecuación de costo esperado (12.18).

Por sustitución del óptimo ν_h en la fórmula para $C^*(V + S^2/N)$, la varianza mínima se encuentra que es

$$V_{min}(\bar{y}_{st}) = \frac{1}{C^*}[\Sigma W_h S_h \sqrt{c_h} + (S^2 - \Sigma W_h S_h^2)^{1/2}\sqrt{c'}]^2 - \frac{S^2}{N} \tag{12.22}$$

El uso de la Fórmula (12.21) para asignación óptima exige, en la práctica, más conocimientos de la población del que se puede esperar que tenga el encuestador. Pero afortunadamente, los errores debidos a conjeturas tienen aspectos de compensación. Así, si los ν_h de (12.21) son demasiado elevados, los n' en (12.18) serán demasiado bajos y los pesos de estratos no estarán determinados tan convenientemente como deberían. Sin embargo, en compensación parcial, las medias de estrato $\bar{y}_h$ tendrán una determinación más precisa que en la solución óptima. Cuando casi no se conocen anticipadamente los pesos de estratos W_h, Srinath (1971) y Rao (1973) sugieren un método de selección de los n_h ligeramente diferente, que se piensa es más resistente a los efectos de conjeturas mediocres de los W_h.

El caso que presenta el problema más fácil de asignación es aquel en el cual c_h y S_h son constantes. Entonces (12.21), viene a ser

$$\nu_h = \nu = \left[\left(\frac{c'}{c}\right)\frac{S_w^2}{(S^2 - S_w^2)}\right]^{1/2} = \left[\left(\frac{c'}{c}\right)\frac{1}{(\phi - 1)}\right]^{1/2} \tag{12.23}$$

donde $\phi = S^2/S_w^2$, es la eficiencia relativa de la estratificación proporcional al muestreo aleatorio simple. Por lo tanto, si conjeturásemos que la estratificación reduciría $V(\bar{y})$ en la mitad, para que $\phi = 2$, entonces $\nu_h = (c'/c)^{1/2}$.

Ejemplo. Este ejemplo no sale de un problema de muestreo doble, pero ilustra algunos aspectos de la solución. Usamos los datos del condado de Jefferson (Pág. 404). Al estimar los acres dedicados al cultivo del maíz, supóngase que podemos o bien tomar una muestra aleatoria simple de granjas, o dedicar algunos recursos a la clasificación de las granjas en dos estratos, según el tamaño de la granja. Los datos pertinentes de la población son para los acres de cultivo de maíz.

Estratos	W_h	S_h^2	S_h	$\bar{Y}_h$
1	0.786	312	17.7	19.404
2	0.214	922	30.4	51.626
Población		620		26.300

Supóngase que $C^* = 100$, $c_h = 1 = c$ y que S^2/N puede ignorarse. Esto implica que si no se usa el muestreo doble, podemos permitirnos la toma de una muestra de $n = 100$ granjas, lo que da $V(\bar{y}) = 6.20$. Sea c' el costo de la clasificación por granja en el estrato 1 ($\leq$ 160 acres) y en el estrato 2 ($>$ 160 acres). Considérense las preguntas:

1. ¿Para qué valores de c'/c trae un aumento en la precisión el muestreo doble?
2. ¿Cuál es el plan de muestreo doble óptimo si $c' = c/100$ y cuál la $V(\bar{y}_{st})$ resultante?
3. En el Prob. 2, ¿cómo cambia el plan y el valor de $V(\bar{y}_{st})$ si los ν_h se conjeturan como el doble de las fracciones óptimas?

1. De los datos de población $\Sigma W_h S_h = 20.4$, $(S^2 - \Sigma W_h S_h^2) = 177$. En consecuencia por (12.22)

$$V_{min}(\bar{y}_{st}) = 0.01\,(20.4 + 13.3\sqrt{c'})^2$$

Si ésta debe ser menor que 6.20 para muestreo aleatorio simple, $c' < 0.11$; esto es, $c'/c < 1/9$.

2. Si $c'/c = 1/100$, entonces, con el plan óptimo, (12.22) da

$$V_{min}(\bar{y}_{st}) = 0.01(20.4 + 1.3)^2 = 4.71$$

Nótese que si la clasificación por tamaño de la granja no cuesta nada, tendríamos

$$V_{min}(\bar{y}_{st}) = 0.01(20.4)^2 = 4.16$$

En lo que respecta a los detalles del plan, con $c'/c = 1/100$, de (12.21), obtenemos

$$\nu_h = S_h/133; \qquad \nu_1 = 0.133; \qquad \nu_2 = 0.229$$

Como $\Sigma W_h \nu_h = 0.1535$. de (12.18) encontramos que $n' = 612$. A su vez, esto nos da los valores esperados de 64, 30 para n_1, n_2. De modo que casi todo el dinero se gastará en la medición: sólo un 6% en la clasificación.

3. Si conjeturamos $\nu_1 = 0.266$, $\nu_2 = 0.458$, entonces $\Sigma W_h \nu_h = 0.307$ y de (12.18), $n' = 315$, lo que conduce a valores esperados $n_1 = 66$, $n_2 = 31$. De (12.3), $V(\bar{y}_{st})$ para este plan se encontrará igual a 4.85, sólo un aumento del 3% sobre el óptimo 4.71 del Prob. 2.

12.4 VARIANZA ESTIMADA EN MUESTREO DOBLE PARA ESTRATIFICACION

Si tanto $1/n'$ como $1/N$ se pueden ignorar respecto a 1 (por ejemplo, < 0.02), una estimación de muestra casi insesgada de $V(\bar{y}_{st})$ en (12.14) es simplemente la copia muestral de esta fórmula.

$$v(\bar{y}_{st}) = \sum_h^L w_h s_h^2 \left(\frac{1}{n' v_h} - \frac{1}{N}\right) + \frac{g'}{n'} \sum_h^L w_h (\bar{y}_h - \bar{y}_{st})^2 \tag{12.24}$$

$$= \sum_h^L \frac{w_h^2 s_h^2}{n_h} - \sum_h^L \frac{w_h s_h^2}{N} + \frac{g'}{n'} \sum_h^L w_h (\bar{y}_h - \bar{y}_{st})^2 \tag{12.24'}$$

donde $g' = (N - n')/(N - 1)$. Esta fórmula será suficiente para casi todas las aplicaciones. Si $1/N$ y $1/n'$ no pueden ignorarse, se necesitará una expresión algebraica más complicada.

Teorema 12.3. Una estimación muestral insesgada de $V(\bar{y}_{st})$ en muestreo doble es

$$v(\bar{y}_{st}) = \frac{n'(N-1)}{(n'-1)N}\left[\sum_h w_h s_h^2 \left(\frac{1}{n' v_h} - \frac{1}{N}\right) + \frac{g'}{n'} \sum_h s_h^2 \left(\frac{w_h}{N} - \frac{1}{n' v_h}\right) + \frac{g'}{n'} \sum_h w_h (\bar{y}_h - \bar{y}_{st})^2\right] \tag{12.25}$$

Prueba. De (12.13) la forma general de la varianza que se va a estimar es

$$V(\bar{y}_{st}) = \sum_h W_h S_h^2 \left(\frac{1}{n' v_h} - \frac{1}{N}\right) + \frac{g'}{n'N} \sum_h (W_h - 1) S_h^2 + \frac{g'}{n'} \sum_h W_h (\bar{Y}_h - \bar{Y})^2 \tag{12.26}$$

Al promediar primero para n' fijo y w_h fijo, y luego sobre las variaciones de los w_h, el promedio de $w_h s_h^2$ en (12.25) es $W_h S_h^2$, en tanto que el de s_h^2 es S_h^2. Estos resultados se usarán después de la Ec. (12.31).

En el último término de (12.25)

$$\sum w_h (\bar{y}_h - \bar{y}_{st})^2 = \sum w_h \bar{y}_h^2 - \bar{y}_{st}^2 \tag{12.27}$$

Al promediar primero respecto a w_h fijo,

$$E\left(\sum w_h \bar{y}_h^2\right) = \sum w_h \bar{Y}_h^2 + \sum w_h S_h^2 \left(\frac{1}{v_h w_h n'} - \frac{1}{w_h N}\right) \tag{12.28}$$

Además,

$$E_w E\left(\sum w_h \bar{y}_h^2\right) = \sum W_h \bar{Y}_h^2 + \sum S_h^2 \left(\frac{1}{v_h n'} - \frac{1}{N}\right) \tag{12.29}$$

También,

$$E(\bar{y}_{st}^2) = \bar{Y}^2 + V(\bar{y}_{st}) \tag{12.30}$$

Al sustraer (12.30) de (12.29) y multiplicar g'/n' se tiene

$$\frac{g'}{n'} E \sum w_h (\bar{y}_h - \bar{y}_{st})^2 = \frac{g'}{n'}\left[\sum W_h (\bar{Y}_h - \bar{Y})^2 + \sum S_h^2 \left(\frac{1}{v_h n'} - \frac{1}{N}\right) - V(\bar{y}_{st})\right] \tag{12.31}$$

Al sustituir (12.31) para encontrar $(n'-1)NEv(\bar{y}_{st})/n'(N-1)$ a partir de (12.25). Se tiene

$$\frac{(n'-1)N}{n'(N-1)} Ev(\bar{y}_{st}) = \left(1 - \frac{g'}{n'}\right) V(\bar{y}_{st}) = \frac{(n'-1)N}{n'(N-1)} V(\bar{y}_{st})$$

Esto demuestra el resultado. Obsérvese que los dos términos medios en (12.25) son del orden $1/n'N$ y $1/n'^2 v_h$ y pueden ignorarse con relación a los términos que se conservan, si $1/N$ y $1/n'$ son despreciables. Esto justifica la forma más simple (12.24).

Rao (1973) dio el resultado (12.25) en términos de los n_h y n_h' como sigue.

$$v(\bar{y}_{st}) = \frac{N-1}{N} \sum_h \left(\frac{n_h' - 1}{n' - 1} - \frac{n_h - 1}{N - 1}\right) \frac{w_h s_h^2}{n_h} + \frac{(N-n')}{N(n'-1)} \sum_h w_h (\bar{y}_h - \bar{y}_{st})^2 \tag{12.32}$$

Corolario. Para usar (12.24) en la estimación de una proporción, sustitúyase p_h por $\bar{y}_h$ y $n_h p_h q_h/(n_h - 1)$ por s_h^2.

Ejemplo. En una muestra aleatoria simple de 374 casas de un distrito, 292 estaban ocupadas por familias de personas de raza blanca y 82 por otras. Una muestra de aproximadamente una de cuatro casas dio los siguientes datos respecto a propiedad.

	Rentas	De su propiedad	Total
Blancas	31	43	74
No blancas	4	14	18

Estimar la proporción de casas rentadas en el área de la que se sacó la muestra y su error estándar.

Si el primer estrato consta de las casas ocupadas por blancos.

$$w_1 = \frac{292}{374} = 0.78, \qquad w_2 = \frac{82}{374} = 0.22$$

$$p_1 = \frac{43}{74} = 0.58, \qquad P_2 = \frac{14}{18} = 0.78$$

$$p_{st} = w_1 p_1 + w_2 p_2 = 0.624$$

$$n' = 374, \qquad n_1 = 74, \qquad n_2 = 18$$

De hecho se encuentra que sólo el primer término de (12.24) es de importancia. Luego

$$v(p_{st}) = \sum_h \frac{n_h}{(n_h-1)} \frac{w_h p_h q_h}{\nu_h n'} = \sum_h \frac{w_h{}^2 p_h q_h}{(n_h-1)} = \frac{(0.78)^2(0.2436)}{73} + \frac{(0.22)^2(0.1716)}{17}$$

$$= 0.00252$$

$$s(p_{st}) = 0.050$$

La proporción estimada de casas rentadas es 0.62 ± 0.050.

12.5 MUESTREO DOBLE PARA COMPARACIONES ANALITICAS

En la Sec. 5A.13 nos hemos ocupado de la determinación de tamaños de muestras en subgrupos de la población para ciertas comparaciones analíticas entre L medias de subgrupos. Si el objetivo es obtener varianzas de las diferencias entre las medias para cada par de subgrupos todas iguales a V, se observó que una asignación simple de compromiso, a menudo adecuada, consiste en especificar o minimizar tan sólo la varianza *promedio* de las $L(L-1)/2$ diferencias.

$$\frac{2}{L} \sum_i^L \frac{S_i^2}{n_i} = V \tag{12.33}$$

En otra aplicación, Booth y Sedransk (1969), los subgrupos formaban una clasificación factorial 2×2 y el problema era la estimación de los efectos renglón y columna con la misma precisión. En este caso, la varianza por especificar o minimizar tomaba la forma

$$\frac{1}{2} \sum_i^2 \sum_j^2 (\theta_i{}^2 + \theta_j{}^2) \frac{S_{ij}{}^2}{n_{ij}} = V \tag{12.34}$$

donde los θ_i y θ_j son pesos que el investigador elige, y los símbolos (ij) denotan los 4 subgrupos.

En la discusión de la Sec. 5A.13 los subgrupos se podían muestrear directamente. El problema viene a ser un problema de muestreo doble cuando los miembros de los subgrupos no se pueden identificar de antemano, pero las unidades pueden clasificarse, a un costo relativamente bajo, en los subgrupos de una muestra aleatoria simple inicial de tamaño n'. Si n'_i de las muestras iniciales caen en el subgrupo i, los n_i para la muestra final que se mide para proporcionar las comparaciones deseadas se extraen de los n_i'. La función de costos más simple es

$$C = c'n' + c \sum n_i = c'n' + cn \tag{12.35}$$

donde suponemos $c' \ll c$.

Para n_i fijo, ambas relaciones (12.33) y (12.34) son de la forma

$$\sum_i^L \frac{a_i^2}{n_i} = V \tag{12.36}$$

donde suponemos que $S_i{}^2$, $S_{ij}{}^2$ y por tanto los $a_i{}^2$ se conocen de antemano.

Cuando el objetivo es minimizar V para C dado, Sedransk (1965) y Booth y Sedransk (1969) intentan valores diferentes de n' en turno. Para n' dado, el valor de n se conoce por (12.35). Para n dado, los n_i que minimizan V son $n_i = na_i/(\sum a_i)$. Puede surgir alguna dificultad al usar estos n_i, porque los n_i' que proporciona la muestra inicial son variables aleatorias. En algunos subgrupos, podemos encontrar $n_i' < na_i/(\sum a_i)$ para que los n_i que minimizan, excedan los n_i' disponibles. Las reglas de asignación que sirven para manejar esta situación se ilustrarán para $L = 3$ grupos, según Sedransk (1965). Numérense los subgrupos en orden creciente de $n_i'(\sum a_i)/na_i$, y denótese con $\sum_1 a_i$ la suma sobre todas las clases, salvo la primera. Tómese

$$\begin{aligned}
&n_1 = \frac{na_1}{(\sum a_i)} && \text{si} && n_1' \geq \frac{na_1}{(\sum a_i)} \\
&n_1 = n_1' && \text{si} && n_1' < \frac{na_1}{(\sum a_i)} \\
&n_2 = \frac{(n-n_1)a_2}{(\sum_1 a_i)} && \text{si} && n_2' \geq \frac{(n-n_1)a_2}{(\sum_1 a_i)} \\
&n_2 = n_2' && \text{si} && n_2' < \frac{(n-n_1)a_2}{(\sum_1 a_i)} \\
&n_3 = n - n_1 - n_2
\end{aligned} \tag{12.37}$$

donde $\sum_1 a_i = \sum a_i - a_1$.

Estas reglas no son completas, pero cubrirán casi todos los n' que puede esperarse cercanos a los óptimos. El principio consiste en no alejarse de la asignación $n_i \propto a_i$. Para más detalles consúltese Booth y Sedransk (1969).

Ejemplo. Un ejemplo en donde el muestreo doble opera convenientemente es el siguiente: $C = 2\,000$, $c' = 1$, $c = 10$. Los tres subgrupos son de tamaños relativos $W_1 = 0.05$, $W_2 = 0.25$, $W_3 = 0.70$, $a_i{}^2 = S_i{}^2 = 10$ $(i = 1, 2, 3)$. Considérese primero muestreo sencillo. Como el costo de selección, clasificación y medida de una unidad de muestreo es de 11 unidades monetarias, podemos permitir $n = 182$ con muestreo sencillo.

La asignación óptima requeriría n_i iguales, pero en promedio, los valores de n_i de muestreo sencillo con $n = 182$ son 9.1, 45.5, 127.4. Al suponer $E(1/n_i) = 1/E(n_i)$ la V promedio de muestreo sencillo es aproximadamente

$$E(V) \doteq 10\left(\frac{1}{9.1}+\frac{1}{45.5}+\frac{1}{127.4}\right) = 1.40$$

Con muestreo doble, los cálculos para los diferentes n' muestran que el n' óptimo está cerca de 620, lo que da $n = 138$. Con $n = 138$ el n_i óptimo es 46 en cada clase. Sin embargo, $n' = 620$ proporciona tan sólo 31 elementos esperados, en la clase 1. Por lo tanto, en promedio, la regla de asignación (12.37) para muestreo doble da $n_1 = E(n'_1) = (0.05)(620) = 31$, $n_2 = n_3 = 53.5$, lo que conduce a

$$E(V) \doteq 10\left(\frac{1}{31}+\frac{2}{53.5}\right) = 0.70$$

una reducción de 50% de $E(V)$ con muestreo que no es doble.

Rao (1973) trata estos problemas por el método usado con muestreo estratificado: este método especifica la fracción ν_i de los n_i' en el subgrupo i-ésimo que se van a medir. Una ventaja es que el óptimo n' se puede determinar analíticamente. Con $C = c'n' + cn$, al igual que antes, el costo esperado es

$$C^* = c'n' + cn' \sum \nu_i W_i \tag{12.38}$$

donde W_i es el tamaño relativo del subgrupo i. Al suponer, como antes, que $E(1/n_i) \doteq 1/E(n_i)$, el promedio V es

$$E(V) = \sum_i^L \frac{a_i^2}{n' W_i \nu_i} \tag{12.39}$$

Por la desigualdad de Cauchy-Schwarz, la fracción ν_i óptima para n' fijo está dada por

$$n' W_i \nu_i = \frac{a_i}{(\sum a_i)} \frac{(C^* - n'c')}{c} \tag{12.40}$$

siempre que $\nu_i \leq 1$. La sustitución de $n'W_i\nu_i$ a partir de (12.40) en (12.39) da, para la $E(V)$ mínima

$$E(V) = \frac{c(\sum a_i)^2}{C^* - n'c'} \tag{12.41}$$

Como $E(V)$ en (12.41) decrece al decrecer n', el n' óptimo disminuirá cuando menos hasta alcanzar el valor m_2, digamos, en el que una de las ν_i, digamos ν_1, se hace 1. En este punto, (12.40) y (12.41) ya no se aplican. De (12.40), ν_1 es 1 cuando

$$n' W_1 = \frac{a_1(C^* - n'c')}{c(\sum a_i)} \tag{12.42}$$

Al resolver (12.42) para m_1, el valor de n' en el que $\nu_1 = 1$, se obtiene

$$m_1 = \frac{C^*}{[c' + cW_1(\sum a_i)/a_1]} \tag{12.43}$$

Para examinar valores de n' menores que m_1, hágase $\nu_1 = 1$ y úsese la desigualdad de Cauchy-Schwarz para obtener las restantes ν_i. Encontramos, para $i > 1$,

$$n' W_i \nu_i = \frac{a_i}{(\sum_1 a_i)} \frac{C^* - n'(c' + cW_1)}{c} \tag{12.44}$$

donde $\sum_1 a_i = \sum a_i - a_1$. La $E(V)$ mínima resultante es ahora

$$E(V) = \frac{a_1^2}{n' W_1} + \frac{c(\sum_1 a_i)^2}{C^* - n'(c' + cW_1)} \tag{12.45}$$

De esta $E(V)$ se anula la derivada $dE(V)/dn'$ en

$$n' = C^* \Big/ \left\{(c' + cW_1) + \frac{(\sum_1 a_i)}{a_1}[cW_1(c' + cW_1)]^{1/2}\right\} \tag{12.46}$$

El valor m_2 donde ν_1 y ν_2 son ambas 1, para que (12.45) deje de ser válido es

$$m_2 = C^* \Big/ \left[(c' + cW_1) + \frac{cW_2(\sum_1 a_i)}{a_2}\right] \tag{12.47}$$

Por lo tanto, las expresiones (12.45) y (12.46) sólo se aplican para los valores $m_1 \geq n' \geq m_2$. Si $dE(V)/dn'$ no se anula para $n' \geq m_2$, debemos poner ν_1, ν_2 y así sucesivamente = 1 en turno hasta encontrar el punto de vuelta de $E(V)$. Sin embargo en muchas situaciones en las que es económico el muestreo doble, el punto de vuelta de $E(V)$ ocurre para $m_1 \geq n' \geq m_2$.

Ejemplo. Para el ejemplo desarrollado en esta sección, $C^* = 2\,000$, $c' = 1$, $c = 10$, $a_i^2 = S_i^2 = 10$, $W_i = 0.05$, 0.25. 0.70. De (12.43) y (12.47) tenemos

$$m_1 = \frac{2000}{[1 + (10)(0.05)(3)} = 800$$

$$m_2 = \frac{2000}{[1.5 + (10)(.25)2]} = 308$$

Además, $dE(V)/dn'$ en la Fórmula (12.46) se anula en

$$n' = \frac{2000}{\{1.5 + (2)[(0.5)(1.5)]^{1/2}\}} = 619$$

Como este valor está entre 800 y 308, proporciona el mínimo requerido. La Fórmula (12.44) da $\nu_2 = .346$, $\nu_3 = .124$. Numéricamente, esta solución es, en esencia, la que se encontró por el método de Sedransk con $n' = 620$ y valores semejantes de n_i en los tres subgrupos.

Ambos métodos se extienden fácilmente al caso de costos diferenciales de medición en subgrupos diferentes. También se han dado sugerencias (Rao, 1973) para el caso en donde $E(1/n_i)$ es sustancialmente mayor que $1/E(n_i)$.

12.6 ESTIMADORES DE REGRESION

En algunas aplicaciones del muestreo doble, la variable auxiliar x_i se ha usado para obtener una estimación de regresión de $\bar{Y}$. En la primera (grande) muestra de tamaño n', medimos tan sólo x_i; en la segunda, una submuestra aleatoria de tamaño $n = \nu n' = n'/k$ donde la fracción ν se elige de antemano, medimos tanto x_i como y_i. La estimación de $\bar{Y}$ es

$$\bar{y}_{lr} = \bar{y} + b(\bar{x}' - \bar{x}) \tag{12.48}$$

donde $\bar{x}', \bar{x}$ son las medias de las x_i en las muestras primera y segunda, y b es el coeficiente de regresión de mínimos cuadrados de y_i respecto a x_i, calculado de la segunda muestra.

Si no se hace un supuesto respecto a la presencia de una regresión lineal en la población, $\bar{y}_{lr}$ será sesgada. Tal como en el muestreo de una etapa (Cap. 7). Puede darse una aproximación a $V(\bar{y}_{lr})$ al suponer muestreo aleatorio y $1/n$ y $1/n$ despreciables respecto a 1.

$$V(\bar{y}_{lr}) \doteq \frac{S_y^2(1-\rho^2)}{n} + \frac{\rho^2 S_y^2}{n'} - \frac{S_y^2}{N} \tag{12.49}$$

Prueba. Para encontrar el error de muestreo de $\bar{y}_{lr}$ en muestreo aleatorio simple, habíamos mostrado que si b en $\bar{y}_{lr}$ se reemplazaba por el coeficiente de regresión de población finita $B = S_{yx}/S_x^2$, el error en la aproximación era de orden $1/\sqrt{n}$ con relación al de $\bar{y}_{lr}$. El mismo método se aplica aquí. Por lo tanto, examinamos la varianza de la aproximación.

$$\tilde{y}_{lr} = \bar{y} + B(\bar{x}' - \bar{x})$$

Los subíndices 1, 2 denotarán variaciones sobre las fases primera y segunda del muestreo. Sea $u_i = y_i - Bx_i$. En la segunda fase se considera la muestra grande como una población finita. Entonces, como la pequeña muestra se extrajo al azar de la grande,

$$E_2(\tilde{y}_{lr}) = \bar{y}': \qquad V_2(\tilde{y}_{lr}) = \left(\frac{1}{n} - \frac{1}{n'}\right) s_u'^2$$

donde $s_u'^2$ es la varianza de u dentro de la muestra grande. Así, se infiere que

$$V(\bar{y}_{lr}) \doteq V(\tilde{y}_{lr}) = V_1(\bar{y}') + E_1\left(\frac{1}{n} - \frac{1}{n'}\right) s_u'^2 \tag{12.50}$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) S_y^2(1-\rho^2) \tag{12.51}$$

ya que $s_u'^2$ es una estimación insesgada de $S_u^2 = S_y^2(1-\rho^2)$. Por lo tanto

$$V(\bar{y}_{lr}) \doteq \frac{S_y^2(1-\rho^2)}{n} + \frac{\rho^2 S_y^2}{n'} - \frac{S_y^2}{N} \tag{12.49}$$

Esto completa la prueba.

Como en la Sec. 7.8 se supone, con base en Royall (1970), que la población finita es en sí misma una muestra aleatoria de una suprapoblación infinita, en donde es válido un modelo de regresión lineal. Entonces, $\bar{y}_{lr}$ se hace insesgada según modelo y pueden obtenerse resultados exactos de muestras pequeñas para su varianza. Sea el modelo de regresión en la suprapoblación

$$y = \alpha + \beta x + \varepsilon \tag{12.52}$$

donde, para x dados, los ε son independientes, con medias 0 y la varianza $\sigma_y^2(1-\rho^2)$, donde σ_y y ρ son ahora parámetros de la suprapoblación.

Al sustituir para $\bar{y}$, $\bar{Y}$, y b a partir de (12.52), cálculos algebraicos directos proporcionan

$$\bar{y}_{lr} - \bar{Y} = \bar{\varepsilon}_n - \bar{\varepsilon}_N + \beta(\bar{x}' - \bar{X}) + \frac{(\bar{x}' - \bar{x})\sum_i^n \varepsilon_i(x_i - \bar{x})}{\sum_i^n (x_i - \bar{x})^2} \tag{12.53}$$

Al promediar sobre la distribución de los ε, resulta de (12.53) que $\bar{y}_{lr}$ es insesgada según modelo para x fijas en la población finita y las dos muestras.
Además, de (12.53),

$$E[(\bar{y}_{lr} - \bar{Y})^2 | x] = \sigma_y^2(1-\rho^2)\left(\frac{1}{n} - \frac{1}{N}\right) + \beta^2(\bar{x}' - \bar{X})^2 + \sigma_y^2(1-\rho^2)\frac{(\bar{x}' - \bar{x})^2}{\sum (x_i - \bar{x})^2} \tag{12.54}$$

El último término a la derecha de (12.54) proviene del error de muestreo de b y es de orden $1/n$, con relación a los dos primeros términos de la derecha. Al promediar los dos primeros términos de la derecha sobre la distribución de las x creada por selecciones alea-

torias repetidas de la población finita y de las dos muestras, se obtiene

$$EV(\bar{y}_{lr}) \doteq \sigma_y^2(1-\rho^2)\left(\frac{1}{n}-\frac{1}{N}\right)+\rho^2\sigma_y^2\left(\frac{1}{n'}-\frac{1}{N}\right) \tag{12.55}$$

$$=\frac{\sigma_y^2(1-\rho^2)}{n}+\frac{\rho^2\sigma_y^2}{n'}-\frac{\sigma_y^2}{N} \tag{12.56}$$

Esta expresión tiene la misma forma que (12.49), excepto en que en (12.56) σ_y^2 y ρ se refieren a la suprapoblación.

El muestreo doble con regresión ha sido extendido por Khan y Tripathi (1967) al caso donde p de las variables auxiliares x se miden en la segunda muestra, al estimar $\bar{Y}$ por la regresión lineal múltiple de y en estas variables. Al ser la segunda muestra una submuestra aleatoria de la primera, y con normalidad multivariada supuesta para y y las x, la extensión de (12.54) para $p > 1$ da para la varianza promedio

$$V(\bar{y}_{lr})=\frac{S_y^2(1-R^2)}{n}\left[1+\frac{(n'-n)}{n'}\frac{p}{(n-p-2)}\right]+\frac{R^2S_y^2}{n'}-\frac{S_y^2}{N} \tag{12.57}$$

donde R es el coeficiente de correlación múltiple entre y y las x.

El caso en que la segunda muestra se extrae independientemente de la primera fue considerado por Chameli Bose (1943).

12.7 ASIGNACION OPTIMA Y COMPARACION CON EL MUESTREO SIMPLE

De la fórmula de la varianza (12.49), donde se supone $1/n$ despreciable, el muestreo doble con estimación de regresión puede compararse con una sola muestra aleatoria simple sin ajuste de regresión. Tenemos

$$V+\frac{S_y^2}{N}=\frac{S_y^2(1-\rho^2)}{n}+\frac{\rho^2S_y^2}{n'}: \qquad C=cn+c'n' \tag{12.58}$$

Por la desigualdad de Cauchy-Schwarz, el producto VC se minimiza cuando

$$\frac{cn^2}{S_y^2(1-\rho^2)}=\frac{c'n'^2}{\rho^2S_y^2} \quad \text{así que} \quad \frac{n}{n'}=\left[\frac{c'}{c}\frac{(1-\rho^2)}{\rho^2}\right]^{1/2} \tag{12.59}$$

Al sustituir en VC se tiene

$$(VC)_{min}=S_y^2(\sqrt{c(1-\rho^2)}+\sqrt{c'\rho^2})^2-\frac{CS_y^2}{N} \tag{12.60}$$

Así, para un costo especificado C,

$$V_{min}=S_y^2\frac{(\sqrt{c(1-\rho^2)}+\sqrt{c'\rho^2})^2}{C}-\frac{S_y^2}{N} \tag{12.61}$$

Si se dedican todos los recursos a una sola muestra sin ajuste de regresión, esta muestra tiene tamaño C/c y la varianza de su media es

$$V(\bar{y})=\frac{cS_y^2}{C}-\frac{S_y^2}{N} \tag{12.62}$$

Por lo tanto, el uso óptimo del muestreo doble dará una varianza menor si

$$c>(\sqrt{c(1-\rho^2)}+\sqrt{c'\rho^2})^2 \tag{12.63}$$

Esta desigualdad puede expresarse de dos maneras.

$$\frac{c}{c'}>\frac{(1+\sqrt{1-\rho^2})^2}{\rho^2} \tag{12.64}$$

o

$$\rho^2>\frac{4(c/c')}{(1+c/c')^2} \tag{12.65}$$

Las Ecs. (12.64) y (12.65) dan los intervalos críticos de c/c' para ρ dada, y de ρ para c/c' dada, que redundan en un muestreo doble ventajoso.

La Fig. 12.1 grafica los valores de la razón c/c' (en una escala logarítmica) contra ρ. La curva I es la relación cuando los muestreos sencillo y doble son de la misma precisión; la curva II es válida cuando $V_{opt} = 0.8\ V(\bar{y})$, es decir, cuando el muestreo doble da un incremento de precisión del 25%, y la curva III se refiere a un incremento de precisión del 50%. Por ejemplo, cuando $\rho = 0.8$, el muestreo doble es igual al sencillo en precisión si c/c' es 4, da un incremento de precisión de 25% de c/c' es más o menos 7½, y un incremento de 50% si el valor de c/c' es aproximadamente de 13.

Para usos prácticos, las curvas sobreestiman las ganancias alcanzadas gracias al muestreo doble, porque los mejores valores de n y n' deben ser estimados de datos previos o conjeturados. Hay cierto margen para los errores en las estimaciones que deberían fijarse antes de decidir adoptar el muestreo doble.

Para cualquier ρ, hay un límite superior a la ganancia en precisión por muestreo doble. Esto ocurre cuando la información sobre $\bar{x}'$ se obtiene gratuitamente ($c' = 0$). El límite superior a la precisión relativa es $1/(1-\rho^2)$.

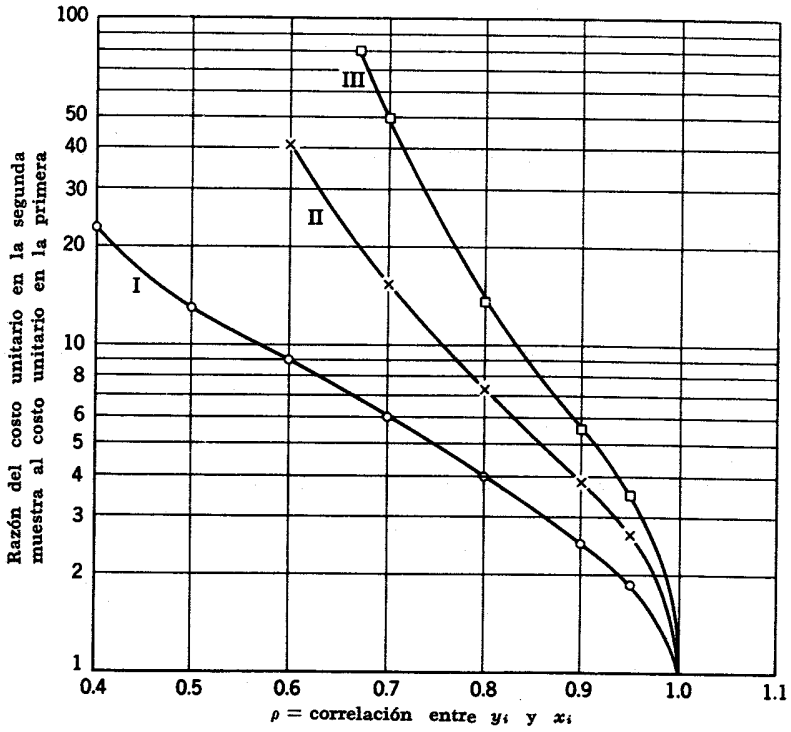


FIG. 12.1. Relación entre c/c' y ρ para tres valores fijos de la precisión relativa de muestreo doble y sencillo

Curva I: los muestreos doble y sencillo son de igual precisión.
Curva II: en muestreo doble da un incremento de precisión del 25%.
Curva III: muestreo doble da un incremento de precisión del 50%.

12.8 VARIANZA ESTIMADA EN MUESTREO DOBLE PARA REGRESION

Si los términos en $1/n$ se pueden ignorar, la $V(\bar{y}_{lr})$ está dada por (12.49):

$$V(\bar{y}_{lr}) \doteq \frac{S_y^2(1-\rho^2)}{n} + \frac{\rho^2 S_y^2}{n'} - \frac{S_y^2}{N}$$

Con un modelo de regresión lineal, la cantidad

$$s_{y.x}^2 = \frac{1}{n-2}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2 - b^2\sum_{i=1}^{n}(x_i-\bar{x})^2\right] \tag{12.66}$$

es una estimación insesgada de $S_y^2(1-\rho^2)$. Como

$$s_y^2 = \frac{\sum(y_i-\bar{y})^2}{n-1}$$

es una estimación insesgada de S_y^2, se infiere que

$$s_y^2 - s_{y.x}^2$$

Es una estimación insesgada de $\rho^2 S_y^2$.

Por lo tanto, una estimación muestral de $V(\bar{y}_{lr})$ es

$$v(\bar{y}_{lr}) = \frac{s_{y.x}^2}{n} + \frac{s_y^2 - s_{y.x}^2}{n'} - \frac{s_y^2}{N} \tag{12.67}$$

Si la segunda muestra es pequeña y los términos en $1/n$ no se pueden despreciar en relación a 1, una estimación de varianza sugerida para muestras aleatorias simples, a partir de (12.54) es

$$v(\bar{y}_{lr}) = s_{y.x}^2\left[\frac{1}{n} + \frac{(\bar{x}'-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right] + \frac{s_y^2 - s_{y.x}^2}{n'} - \frac{s_y^2}{N} \tag{12.68}$$

Esta es una híbrida de la varianza condicional y la varianza promedio.

12.9 ESTIMADORES DE RAZON

Si la primera muestra se usa para obtener $\bar{x}'$ como una estimación de $\bar{X}$ en una estimación de razón $\bar{Y}$, el estimador de $\bar{Y}$ es

$$\bar{y}_R = \frac{\bar{y}}{\bar{x}}\bar{x}' \tag{12.69}$$

Para encontrar la varianza aproximada, escríbase

$$\begin{aligned}\bar{y}_R - \bar{Y} &= \frac{\bar{y}}{\bar{x}}\bar{x}' - \bar{Y} \\ &= \left(\frac{\bar{y}}{\bar{x}}\bar{X} - \bar{Y}\right) + \frac{\bar{y}}{\bar{x}}(\bar{x}'-\bar{X}) \\ &= \frac{\bar{X}}{\bar{x}}(\bar{y}-R\bar{x}) + \frac{\bar{y}}{\bar{x}}(\bar{x}'-\bar{X})\end{aligned}$$

La primera componente es el error de la estimación ordinaria de razón (Sec. 2.11). Para obtener la varianza apropiada del error en la Sec. 2.11, reemplazábamos el factor $\bar{X}/\bar{x}$ por la unidad en este término. Para el mismo orden de aproximación, reemplazamos el factor $\bar{y}/\bar{x}$ en la segunda componente por la razón de población $R = \bar{Y}/\bar{X}$. De modo que

$$\bar{y}_R - \bar{Y} \doteq (\bar{y}-R\bar{x}) + R(\bar{x}'-\bar{X}) \tag{12.70}$$

Si la segunda muestra es una submuestra aleatoria de la primera

$$E_2(\bar{y}_R - \bar{Y}) \doteq \bar{y}' - \bar{Y}: \qquad V_2(\bar{y}_R - \bar{Y}) \doteq \left(\frac{1}{n} - \frac{1}{n'}\right) s_d'^2 \tag{12.71}$$

donde $s_d'^2$ es la varianza dentro de la segunda muestra de la variable $d_i = y_i - Rx_i$. Al promediar ahora sobre selecciones aleatorias repetidas de la primera muestra

$$V(\bar{y}_R) = V_1 E_2 + E_1 V_2$$
$$\doteq \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)(S_y^2 - 2RS_{yx} + R^2 S_x^2) \tag{12.72}$$

como $s_d'^2$ es una estimación insesgada de $S_d^2 = S_y^2 - 2RS_{yx} + R^2 S_x^2$.

Al separar los términos en $1/n$ y $1/n'$ obtenemos

$$V(\bar{y}_R) \doteq \frac{S_y^2 - 2RS_{yx} + R^2 S_x^2}{n} + \frac{2RS_{yx} - R^2 S_x^2}{n'} - \frac{S_y^2}{N} \tag{12.72'}$$

12.10 MUESTREO REPETIDO DE LA MISMA POBLACION

La práctica de confiar en muestras para la recolección de series importantes de datos publicados a intervalos regulares, se ha hecho común. Esto se debe en parte a que se ha comprendido que con una población dinámica, un censo a intervalos no frecuentes es de poca utilidad. Una información muy precisa sobre las características de una población en julio de 1960 y julio de 1970, puede no ser de mucha ayuda en la planeación que requiere un conocimiento de la población en 1976. Pueden resultar útiles una serie de pequeñas muestras tomadas a intervalos de un año o menos.

Cuando se muestrea repetidamente la misma población (salvo los cambios debidos al paso del tiempo) el muestreador está en una posición ideal para hacer estimaciones realistas, tanto de los costos como de las varianzas, para aplicar las técnicas, que conducen a la eficiencia óptima del muestreo. Una pregunta importante es, ¿con cuánta frecuencia y de qué manera debe cambiar la muestra con el transcurso del tiempo? Son muchas las consideraciones que afectan esta decisión. La gente puede estar poco dispuesta a dar el mismo tipo de información una y otra vez. Los entrevistados pueden estar influidos por la información que reciben en las entrevistas y esto puede hacerlos cada vez menos representativos con el paso del tiempo. Sin embargo, en ocasiones la cooperación es mejor en una segunda entrevista que en la primera y cuando la información es de índole técnica o confidencial, la segunda visita puede producir datos más exactos que la primera.

En el resto de este capítulo se considera el asunto del reemplazo de la muestra y el asunto relacionado con hacer estimaciones a partir de una serie repetida de muestras. Este tema de estudio es apropiado al contenido de este capítulo porque se pueden utilizar las técnicas del muestreo doble.

Con los datos de una serie de muestras, hay tres clases de cantidades que podríamos estimar.

1. El cambio en $\bar{Y}$ entre una ocasión y la siguiente.
2. El valor promedio de $\bar{Y}$ sobre todas las ocasiones.
3. El valor promedio de $\bar{Y}$ para la última ocasión.

En la mayoría de las encuestas, se centra el interés sobre el promedio actual (3), particularmente si se puede esperar que las características de la población cambien rápidamente con el tiempo. Por otro lado, con una población que sufre cambios lentos, un promedio anual (2) sobre 12 muestras mensuales o cuatro muestras trimestrales sería adecuado para los usos principales. Esta sería la situación en un estudio sobre la permanencia de ciertas enfermedades crónicas en un gran lapso de tiempo. Tratándose de una enfermedad cuya permanencia presenta una marcada variación estacional, los datos corrientes son de mayor interés, pero también son útiles los promedios anuales para comparaciones entre diferentes regiones y años. Las estimaciones del cambio (1) se desean principalmente al intentar estudiar los aspectos de las fuerzas que se sabe actúan sobre la población. Por ejemplo, si se acepta una moción que supuestamente estimula la construcción de viviendas, es interesante saber si la tasa de construcción de casas nuevas se ha incrementado en el año siguiente (al tener presente que este incremento no se debe enteramente à la nueva ley).

Supóngase que estamos en libertad de alterar o retener la composición de la muestra, y que su tamaño total debe ser el mismo en todas las ocasiones. Si deseamos maximizar la precisión, se pueden hacer los siguientes enunciados respecto a las políticas de reemplazo:

1. Para estimar el cambio, lo mejor es conservar la misma muestra en todas las ocasiones.
2. Para estimar el promedio sobre todas las ocasiones, lo mejor es sacar una nueva muestra en cada ocasión.
3. Para estimaciones actuales, se obtiene la misma precisión, ya sea conservando la misma muestra o cambiándola en cada ocasión. El reemplazo de una parte de la muestra en cada ocasión puede ser mejor que estas alternativas.

Los enunciados 1 y 2 son válidos porque casi siempre habrá una correlación *positiva* entre mediciones de la misma unidad en dos ocasiones sucesivas. El cambio estimado en una unidad tiene la varianza $S_1^2+S_2^2-2\rho S_1S_2$, donde los subíndices denotan las dos ocasiones. Si el cambio se estima de dos unidades *diferentes*, la varianza es $S_1^2+S_2^2$. Al estimar la media global para las dos ocasiones, la varianza es $(S_1^2+S_2^2+2\rho S_1S_2)/4$ si se conserva la misma unidad y $(S_1^2+S_2^2)/4$ si se elige una nueva unidad.

El enunciado 3, menos obvio, se investiga en las secciones siguientes.

12.11 MUESTREO EN DOS OCASIONES

Supóngase que las muestras son del mismo tamaño n en ambas ocasiones y que las estimaciones actuales son de interés primario. La política del reemplazo ha sido examinada por Jessen (1942). Por simplicidad, suponemos que se usa muestreo aleatorio simple.

La media de la primera muestra tiene varianza S_1^2/n, sin que haya información previa que utilizar. Al seleccionar la segunda muestra, se retienen m unidades de las de la primera muestra (m para apareadas). Las u unidades restantes (u para no apareadas) se descartan y reemplazan por una nueva selección de unidades no seleccionadas previamente.

Notación.

$\bar{y}_{hu}$ = media de la porción sin aparear en la ocasión h
$\bar{y}_{hm}$ = media de la porción apareada en la ocasión h
$\bar{y}_h$ = media de la muestra completa en la ocasión h

Las porciones apareadas y las sin aparear de la segunda muestra proveen las estimaciones independientes y_{2u}', y_{2m}' de $\bar{Y}_2$, como se muestra en la Tabla 12.1. En la porción apareada usamos una estimación de regresión de muestreo doble, donde la muestra "grande"

TABLA 12.1. ESTIMACIONES DE LAS PORCIONES APAREADA Y NO APAREADA

	Estimación	Varianza
No apareada:	$\bar{y}_{2u}'=\bar{y}_{2u}$	$\frac{S_2^2}{u}=\frac{1}{W_{2u}}$
Apareada:	$\bar{y}_{2m}'=\bar{y}_{2m}+b(\bar{y}_1-\bar{y}_{1m})$	$\frac{S_2^2(1-\rho^2)}{m}+\rho^2\frac{S_2^2}{n}=\frac{1}{W_{2m}}$

es la primera muestra y la variable auxiliar x_i es el valor de y_i en la primera ocasión. La varianza de $\bar{y}_{2m}'$ viene de (12.49), Págs. 412 y 413: nótese que las muestras m y n corresponden a n y n', respectivamente, en (12.49). La cpf es despreciable.

La mejor estimación combinada de $\bar{Y}_2$ se encuentra ponderando las dos estimaciones independientes inversamente a sus varianzas. Si W_{2u}, W_{2m} son las varianzas inversas, esta estimación es

$$\bar{y}_2'=\phi_2\bar{y}_{2u}'+(1-\phi_2)\bar{y}_{2m}' \tag{12.73}$$

donde

$$\phi_2=\frac{W_{2u}}{W_{2u}+W_{2m}}$$

Según la teoría de los mínimos cuadrados, la varianza de y_2' es

$$V(\bar{y}_2')=\frac{1}{W_{2u}+W_{2m}}$$

De la Tabla 12.1, esto resulta ser después de simplificar

$$V(\bar{y}_2')=\frac{S_2^2(n-u\rho^2)}{n^2-u^2\rho^2} \tag{12.74}$$

Nótese que si $u=0$ (apareamiento completo) o si $u=n$ (no apareamiento) esta varianza tiene el mismo valor, S_2^2/n.

El valor óptimo de u se encuentra al minimizar (12.79) con respecto a variación en u. Esto da

$$\frac{u}{n}=\frac{1}{1+\sqrt{1-\rho^2}},\qquad \frac{m}{n}=\frac{\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} \tag{12.75}$$

Cuando la u óptima se sustituye en (12.74), la varianza mínima resulta ser

$$V_{\text{ópt}}(\bar{y}_2')=\frac{S_2^2}{2n}(1+\sqrt{1-\rho^2}) \tag{12.76}$$

La Tabla 12.2 muestra, para una serie de valores de ρ, el porcentaje óptimo que se debe aparear y la ganancia relativa en precisión comparada con no apareamiento. El mejor porcentaje a aparear nunca excede 50% y disminuye constantemente conforme ρ aumenta. Cuando $\rho=1$, la fórmula sugiere $m=0$, la cual cae fuera de la gama de nuestras suposiciones, ya que m se ha supuesto razonablemente grande. El procedimiento correcto en este caso es tomar $m=2$. Las dos unidades apareadas son suficientes para determinar exactamente la línea de regresión.

TABLA 12.2. PORCENTAJE OPTIMO APAREADO

ρ	% óptimo apareado	% de ganancia en precisión	% de ganancia con $\frac{m}{n}=\frac{1}{3}$	% de ganancia con $\frac{m}{n}=\frac{1}{4}$
0.5	46	7	7	6
0.6	44	11	11	9
0.7	42	17	17	15
0.8	38	25	25	23
0.9	30	39	39	39
0.95	24	52	50	52
1.0	0	100	67	75

La ganancia en precisión más grande que se puede obtener es 100% cuando $\rho = 1$. Si ρ no es alto, las ganancias son modestas.

Aunque el porcentaje óptimo para aparear varía con ρ, en la práctica, solamente se puede usar un solo porcentaje para todas las características de una encuesta. En la Tabla 12.2, la columna de la derecha muestra los porcentajes de ganancia en precisión cuando se aparean un tercio y un cuarto de las unidades. Ambas son buenos compromisos, excepto para las características en las cuales ρ excede 0.95.

Kulldorff (1963) da una detallada discusión de este modelo; él considera el caso en el que en la segunda ocasión el costo de medición de una unidad apareada (es decir, una antes medida) puede diferir de 1 de la medición de una nueva unidad no apareada y no supone tamaños iguales de muestra en las dos ocasiones. Por lo tanto, aparte de los costos fijos, su costo en la segunda ocasión es

$$C_2 = mc_m + uc_u: \qquad \frac{C_2}{c_u} = m\delta + u \tag{12.77}$$

donde $\delta = c_m/c_u$. Si los tamaños de muestras son los mismos en ambas ocasiones de manera que $m + u = n$, la proporción no apareada óptima en la segunda ocasión se obtiene minimizando

$$\frac{VC_2}{c_u S_2^2} = [n\delta + u(1-\delta)]\frac{(n-u\rho^2)}{(n^2-u^2\rho^2)} = [\delta + \mu(1-\delta)]\frac{(1-\mu\rho^2)}{(1-\mu^2\rho^2)} \tag{12.78}$$

donde $\mu = u/n$ y V viene de (12.74). Si $\delta < 1$, siendo más económico aparear, la proporción apareada óptima es desde luego, mayor que los valores de la Tabla 12.2. También trata el caso en el que los costos deben ser los mismos en ambas ocasiones.

En algunas aplicaciones los datos para la ocasión 1 proporcionan varias variedades auxiliares correlacionadas con y_2, una de las cuales será usualmente y_1. Por ejemplo, al estimar la caza de aves acuáticas y_2 por cada cazador en Ontario de 1968 a 1969, Sen (1973a) encontró que la caza por cazador y el número de días de cacería en 1967 y 1968 estaban correlacionados con y_2. En este artículo amplió el análisis anterior al caso en que $\bar{y}_{2m}$ se ajusta por su regresión lineal múltiple respecto a las variables auxiliares y donde las muestras en las dos ocasiones son de tamaños desiguales. Con grandes muestras de tamaño igual el único cambio en (12.76) para $V_{opt}(\bar{y}_2')$ es sustituir ρ^2 por el cuadrado R^2 del coeficiente de correlación múltiple entre y_2 y las variables auxiliares (suponiendo normalidad multivariada). La teoría correspondiente para el caso en que $\bar{y}_{2m}$ se ajusta por la estimación de razón multivariada se da en Sen (1972) para tamaños de muestras iguales y en Sen (1973b) para tamaños diferentes.

12.12 MUESTREO EN MAS DE DOS OCASIONES

El problema general de reemplazo lo ha estudiado Yates (1960) y Patterson (1950), en relación tanto con estimaciones actuales como con estimaciones de cambio. Cuando hay más de dos ocasiones, aumentan las oportunidades para un uso flexible de los datos. En una ocasión h podemos tener partes de la muestra que se aparean con la ocasión $h-1$, partes que se aparean con ocasiones $h-1$ y $h-2$, etc. En un intento por mejorar la estimación actual, podemos tratar con una regresión múltiple que comprende todos los apareamientos en ocasiones anteriores. También se puede *revisar* la estimación actual para la ocasión $h-1$ después de conocer los datos para la ocasión h. En la estimación revisada podría utilizarse la regresión de la ocasión $h-1$, tanto en la ocasión $h-2$ como en la ocasión h, al suponer que estuvieron disponibles las apropiadas porciones apareadas de la muestra.

Esta sección contiene una introducción a la materia. La atención se restringirá a estimaciones actuales en las que sólo se usa la regresión en la muestra inmediatamente anterior. Esto da como resultado una pequeña pérdida de precisión, pero dado que usualmente la correlación ρ disminuye conforme aumenta el intervalo de tiempo entre las ocasiones, la pérdida en precisión rara vez será significativa. La varianza S^2 y el coeficiente de correlación ρ entre valores de características en la misma unidad en dos ocasiones sucesivas se suponen constantes todo el tiempo.

TABLA 12.3. ESTIMACIONES DE $\bar{Y}_h$ EN LA h-ESIMA OCASIÓN

	Estimación	Varianza
No apareado:	$\bar{y}_{hu}' = \bar{y}_{hu}$	$\frac{S^2}{u} = \frac{1}{W_{hu}}$
Apareado:	$\bar{y}_{hm}' = \bar{y}_{hm} + b(\bar{y}'_{h-1} - \bar{y}_{h-1,m})$	$\frac{S^2(1-\rho^2)}{m} + \rho^2 V(\bar{y}'_{h-1}) = \frac{1}{W_{hm}}$

En la h-ésima ocasión, sean m_h y u_h los números de unidades apareadas y no apareadas, respectivamente, con la $(h-1)$-ésima ocasión. Las dos estimaciones de $\bar{Y}_h$ que pueden obtenerse se dan en la Tabla 12.3. El único cambio en procedimiento en la segunda ocasión (Tabla 12.1) es que en el ajuste de la estimación de la porción apareada por la regresión, usamos la estimación mejorada $\bar{y}'_{h-1}$ en lugar de la media de la muestra $\bar{y}_{h-1}$.

La varianza de la estimación apareada $\bar{y}_{hm}'$ en la Tabla 12.3 se deriva de (12.49) que está en Págs. 412 y 413. Nótese que (a) nuestra m corresponde a la n en (12.49) y (b) el término $\rho^2 S_y^2/n'$ en (12.49) que es igual a $B^2 V(\bar{x}')$, se reemplaza por $\rho^2 V(\bar{y}'_{h-1})$, como $B = \rho$ cuando S es constante en ocasiones sucesivas y $\bar{y}'_{h-1}$ corresponde a $\bar{x}'$ en el análisis previo.

Ahora examinamos la precisión obtenida si en toda ocasión se usan las óptimas m_h y u_h y las ponderaciones óptimas. Se encontrará que la óptima m_h/n aumenta constantemente en ocasiones sucesivas, y se aproximan rápidamente a un valor límite de ½.

Al ponderar inversamente con la varianza, la mejor estimación de $\bar{y}_h$ es

$$\bar{y}_h' = \phi_h \bar{y}_{hu}' + (1-\phi_h)\bar{y}_{hm}' \tag{12.79}$$

donde $\phi_h = W_{hu}/(W_{hu} + W_{hm})$. Esto da

$$V(\bar{y}_h') = \frac{1}{W_{hu} + W_{hm}} = \frac{g_h S^2}{n}$$

donde g_h denota la razón de la varianza en la ocasión h a la de la primera ocasión. Al sustituir W_{hu}, W_{hm} de la Tabla 12.3, tenemos

$$\frac{S^2}{V(\bar{y}_h')} = \frac{n}{g_h} = S^2(W_{hu} + W_{hm}) = u_h + \frac{1}{\frac{(1-\rho^2)}{m_h} + \frac{\rho^2 g_{h-1}}{n}} \tag{12.80}$$

Ahora escogemos m_h y u_h para maximizar esta cantidad y, por lo tanto, minimizar $V(\bar{y}_h')$. Al escribir $u_h = n - m_h$ y diferenciar el lado derecho de (12.80) con respecto a m_h obtenemos

$$\frac{1-\rho^2}{m_h^2} = \left(\frac{1-\rho^2}{m_h} + \frac{\rho^2 g_{h-1}}{n}\right)^2$$

Esto da, al resolver para la $\hat{m}_h$, óptima, digamos,

$$\frac{\hat{m}_h}{n} = \frac{\sqrt{1-\rho^2}}{g_{h-1}(1+\sqrt{1-\rho^2})} \tag{12.81}$$

Cuando se sustituye este valor en (12.80), la relación se convierte después de alguna manipulación algebraica en

$$\frac{1}{g_h} = 1 + \frac{1-\sqrt{1-\rho^2}}{g_{h-1}(1+\sqrt{1-\rho^2})} \tag{12.82}$$

Esta relación puede escribirse

$$r_h = 1 + b r_{h-1}$$

donde $r_h = 1/g_h$ y $r_1 = 1/g_1 = 1$. El uso repetido de esta relación de recurrencia da

$$\frac{1}{g_h} = r_h = 1 + b + b^2 + \cdots + b^{h-1} = \frac{1-b^h}{1-b}$$

donde, de (12.82), $b = (1-\sqrt{1-\rho^2})/(1+\sqrt{1-\rho^2})$. Dado que $0 < b < 1$, el factor límite de la varianza g_∞ es

$$g_\infty = 1 - b = \frac{2\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} \tag{12.83}$$

Por lo tanto, la varianza de $\bar{y}_h'$ tiende hacia

$$V(\bar{y}_\infty') = \frac{S^2}{n}\left(\frac{2\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}}\right) \tag{12.84}$$

Finalmente, el valor límite de $\hat{m}_h$ se obtiene de (12.81) como

$$\frac{\hat{m}_\infty}{n} = \frac{\sqrt{1-\rho^2}}{g_\infty(1+\sqrt{1-\rho^2})} = \frac{1}{2}$$

independientemente del valor de ρ.

La Tabla 12.4 muestra el porcentaje óptimo a aparear —$100\hat{m}_h/n$, según se obtiene de (12.81)— y las varianzas resultantes para ρ = 0.7, 0.8, 0.9 y 0.95 y para una serie de valores de h.

TABLA 12.4. PORCENTAJE OPTIMO APAREADO Y VARIANZAS

	% apareado $100\hat{m}_h/n$				$g_h = nV(\bar{y}_h')/S^2$			
	$\rho =$				$\rho =$			
h	0.7	0.8	0.9	0.95	0.7	0.8	0.9	0.95
2	42	38	30	24	0.857	0.800	0.718	0.656
3	49	47	42	36	0.837	0.762	0.646	0.556
4	50	49	47	43	0.834	0.753	0.622	0.515
5	50	50	49	46	0.833	0.751	0.613	0.495
∞	50	50	50	50	0.833	0.750	0.607	0.476

Para la cuarta ocasión, el porcentaje óptimo apareado es cercano a 50 para todos los valores de ρ mostrados, aunque una cantidad menor de apareamiento se indica para la segunda y tercera ocasiones. Las reducciones en varianza, esto es $(1 - g_h)$, son modestas si ρ es menor que 0.8.

12.13 SIMPLIFICACIONES Y ADELANTOS ULTERIORES

En su aplicación práctica, el análisis precedente puede necesitar una modificación. Hemos supuesto que todas las políticas de reemplazo cuestan lo mismo y son igualmente factibles. Con poblaciones humanas, los costos de campo probablemente sean más bajos si se retienen las mismas unidades para un cierto número de ocasiones. Si las estimaciones del cambio en el total o en la media de la población son de interés, este factor también apunta hacia el apareamiento de más de la mitad de las unidades de una ocasión a la siguiente.

También es conveniente conservar constantes las ponderaciones y las proporciones apareadas, en lugar de cambiarlas en cada ocasión. Consecuentemente, investigaremos las varianzas de $\bar{y}_h'$ y de la estimación del cambio $(\bar{y}_h' - \bar{y}_{h-1}')$ cuando m, u y ϕ se mantienen constantes. Continuamos escribiendo $V(\bar{y}_h') = g_h S^2/n$, aunque el valor actual de g_h será diferente del de la sección precedente.

La estimación es ahora

$$\bar{y}_h' = \phi \bar{y}_{hu}' + (1-\phi)\bar{y}_{hm}'$$

Al sustituir las expresiones para las dos varianzas (de la Tabla 12.3), tenemos

$$V(\bar{y}_h') = \frac{g_h S^2}{n} = \phi^2 V(\bar{y}_{hu}') + (1-\phi)^2 V(\bar{y}_{hm}')$$

$$S^2\left[\frac{\phi^2}{u} + \frac{(1-\phi)^2(1-\rho^2)}{m}\right] + \frac{S^2\rho^2(1-\phi)^2 g_{h-1}}{n}$$

Por lo tanto,

$$g_h = \left[\frac{\phi^2}{\mu} + \frac{(1-\phi)^2(1-\rho^2)}{\lambda}\right] + \rho^2(1-\phi)^2 g_{h-1} \tag{12.85}$$

donde $\mu = u/n$, $\lambda = m/n$. Escríbase esta relación como

$$g_h = a + b g_{h-1}$$

Por aplicación repetida, tenemos, dado que $g_1 = 1$,

$$g_h = \frac{a(1-b^{h-1})}{1-b} + b^{h-1}$$

Como $b = \rho^2(1 - \phi)^2$ es menor que 1, el valor límite es

$$g_\infty = \frac{a}{1-b} = \frac{\lambda\phi^2 + \mu(1-\phi)^2(1-\rho^2)}{\lambda\mu[1-\rho^2(1-\phi)^2]} \tag{12.86}$$

El valor de la ponderación ϕ que minimiza la varianza límite puede encontrarse al diferenciar (12.86). Esto conduce a una ecuación cuadrática cuya raíz apropiada es

$$\phi_{\text{ópt}} = \frac{\sqrt{1-\rho^2}\left[\sqrt{1-\rho^2+4\lambda\mu\rho^2} - \sqrt{1-\rho^2}\right]}{2\lambda\rho^2}$$

En la práctica, no se conocerá el valor exacto de ρ y diferirá de característica a característica. En general, se puede escoger un valor simple de transacción. Claramente, $\phi_{\text{ópt}}$ será menor que $\mu = u/n$, ya que la parte apareada de la muestra da una precisión más alta por unidad que la parte no apareada. Por ejemplo, con $\mu = 0.25$, esto es, ¼ de la muestra sin aparear, $\phi_{\text{ópt}}$ resulta ser 0.216, 0.198 y 0.164 para $\rho = 0.7$, 0.8, 0.9. La elección de $\phi = 0.2$ sería adecuada para esta gama de valores de ρ.

Para la estimación de cambio, tenemos

$$V(\bar{y}_h' - \bar{y}_{h-1}') = V(\bar{y}_h') + V(\bar{y}_{h-1}') - 2\,\text{cov}\,(\bar{y}_h'\bar{y}_{h-1}') \tag{12.87}$$

Para encontrar el término de covarianza, nótese que si y_{hi}, $y_{h-1,i}$ son los valores para la i-ésima unidad en el grupo apareado en las ocasiones h y $(h - 1)$, nuestro modelo es

$$y_{hi} = \bar{Y}_h + \rho(y_{h-1,i} - \bar{Y}_{h-1}) + e_{hi}$$

donde los e_{hi} son independientes de las y. A partir de este modelo se encuentra, por sustitución, que

$$\bar{y}_{hm}' = \bar{y}_{hm} + \rho(\bar{y}_{h-1}' - \bar{y}_{h-1,m}) = \bar{Y}_h + \rho(\bar{y}_{h-1}' - \bar{Y}_{h-1}) + \bar{e}_{hm}$$

Por lo tanto, la covarianza de $\bar{y}_{hm}'$ y $\bar{y}_{h-1}'$ es $\rho V(\bar{y}_{h-1}')$. Pero

$$\text{cov}\,(\bar{y}_h'\bar{y}_{h-1}') = \text{cov}\,\{[\phi\bar{y}_{hu} + (1-\phi)\bar{y}_{hm}']\bar{y}_{h-1}'\} = \rho(1-\phi)V(\bar{y}_{h-1}')$$

TABLA 12.5. GANANCIAS PORCIENTO EN EFICIENCIA PARA LA ESTIMACIÓN $\bar{y}_h'$ ACTUAL Y LA ESTIMACIÓN DEL CAMBIO $(\bar{y}_h'-\bar{y}_{h-1}')$. PROPORCIONES APAREADAS: ½ Y ¾

Estimación actual: ganancia porciento $= 100(1-g_h)/g_h$

h	$\rho=0.7$ ½	$\rho=0.7$ ¾	$\rho=0.8$ ½	$\rho=0.8$ ¾	$\rho=0.9$ ½	$\rho=0.9$ ¾	$\rho=0.95$ ½	$\rho=0.95$ ¾
2	14	10	22	14	33	19	41	22
3	16	14	30	20	52	32	67	39
4	17	15	32	24	59	40	79	52
∞	17	15	33	26	62	50	89	74
RG^a	15	10	27	18	56	40	—	—
h	Estimación del cambio: ganancia porciento $= 100(2-g_h')/g_h$							
2	106	153	156	233	245	399	326	565
3	113	160	170	245	277	415	365	588
4	115	160	174	251	285	424	388	603
∞	115	163	178	251	292	440	397	624
RG	101	163	166	269	381	605	—	—

[a]Descrito en la Pág. 430.

dado que $\bar{y}_{hu}$ es independiente de $\bar{y}_{h-1}'$. De (12.87), esto da

$$V(\bar{y}_h'-\bar{y}_{h-1}')=\frac{S^2}{n}\{g_h+g_{h-1}[1-2\rho(1-\phi)]\} \tag{12.88}$$

De (12.85) y (12.87), las varianzas de $\bar{y}_h'$ y $(\bar{y}_h'-\bar{y}_{h-1}')$ pueden calcularse para cualesquiera valores de m, ϕ, y ρ. La Tabla 12.5 muestra las ganancias porciento en eficiencia que resultan para estas estimaciones en relación a las estimaciones obtenidas a partir de muestras independientes en cada ocasión. Las proporciones apareadas son $\lambda = m/n = ½$ y ¾. El peso ϕ se tomó como 0.35 para $\lambda = ½$ y 0.2 para $\lambda = ¾$.

No es sorprendente que las características más importantes de la Tabla 12.5 son las ganancias elevadas en eficiencia para las estimaciones del cambio cuando ρ es al menos 0.7. Además, el aumento de la proporción apareada de ½ a ¾, da lugar a ganancias sustanciales en eficiencia de las estimaciones de cambio a costa de pérdidas menores en eficiencia para las estimaciones actuales. Los resultados sugieren, que retener la proporción ⅔, ¾ o ⅘ de una ocasión a otra,

puede ser una buena práctica si las estimaciones actuales y las de cambio son importantes.

La comparación de las ganancias en eficiencia para la estimación actual, $\bar{y}_h'$ en la Tabla 12.5, con las ganancias óptimas a partir de la Tabla 12.4 sugiere que después de la segunda ocasión poca precisión es la que se pierde usando peso constante y una proporción fija de elementos apareados, a menos que $\rho \geq 0.95$.

Si ρ excede 0.8, el coeficiente de regresión $b = \rho$ puede sustituirse con tan sólo una pequeña pérdida de precisión resultante. Esto da una estimación $\bar{y}_h''$ de la forma

$$\bar{y}_h''=\phi\bar{y}_{hu}+(1-\phi)(\bar{y}_{h-1}''+\bar{y}_{hm}-\bar{y}_{h-1,m}) \tag{12.89}$$

En la encuesta de población actual hecha mensualmente en los Estados Unidos por la Oficina del Censo una cuarta parte de las unidades de segunda etapa se reemplaza cada mes, de modo que un domicilio particular permanece en la muestra durante cuatro meses consecutivos. Ese domicilio se omite en la muestra por ocho meses; pero luego se reincluye por cuatro meses, incrementando así ligeramente la precisión de las comparaciones entre años sucesivos.

La estimación compuesta usada en esta encuesta es de una forma relacionada con (12.89) pero ligeramente diferente.

$$\bar{y}_h''=(1-K)\bar{y}_h+K(\bar{y}_{h-1}''+\bar{y}_{hm}-\bar{y}_{h-1,m}) \tag{12.90}$$

donde K es un factor de ponderación constante. La diferencia es que $\bar{y}_h$, la estimación actual para toda la muestra, toma el lugar de $\bar{y}_{hu}$ en (12.89). Las cantidades $\bar{y}_{hm}$, $\bar{y}_{h-1,m}$, $\bar{y}_h$ en (12.90) son estimaciones de razón de un tipo bastante complejo.

La varianza de $\bar{y}_h''$ (debida a Bershad) está dada en Hansen, Hurwitz y Madow (1953); véase también el apéndice en Hansen y otros (1955). El estimador del cambio de un mes a otro es

$$\bar{d}_h=\bar{y}_h''-\bar{y}_{h-1}''$$

Como las unidades primarias permanecen inalteradas, sólo se afectan las componentes dentro de unidades de $V(\bar{y}_h'')$ y $V(\bar{d}_h)$ por la política de rotación de muestra.

Rao y Graham (1964) han examinado el comportamiento de las estimaciones compuestas en políticas de rotación de este tipo, en las que un entrevistado permanece en la muestra durante r meses, y luego se le saca de la muestra durante m meses. Ellos usaron como modelos un correlograma exponencial y un correlograma lineal en el tiempo, que desciende a cero. Graham (1973) ha estudiado un correlograma más complejo, requerido cuando hay una elevada correlación entre los meses h y $(h-12)$.

Sus ganancias en eficiencia para $r = 2$, 4 y $m = \infty$ corresponden a los resultados para $\lambda = \frac{1}{2}, \frac{3}{4}$ en la Tabla 12.5. Para un correlograma exponencial en el que la correlación entre resultados de la misma unidad en las ocasiones h, $h - i$ es ρ^i, las líneas rotuladas *RG* de la Tabla 12.5 muestran para $\rho = 0.7, 0.8, 0.9$ sus porcentajes ganancias en eficiencia. Para cada ρ utilizan en (12.90) el K óptimo para las estimaciones actuales al tender $h \to \infty$. Como se muestra en la Tabla 12.5 también encuentran que las ganancias de eficiencia con las estimaciones compuestas son mucho más grandes al estimar el cambio que al estimar el nivel actual.

En un contexto más general, Scott y Smith (1974) han discutido el papel que desempeñan los métodos de series de tiempo en las estimaciones hechas en encuestas repetidas de varios tipos.

En otra política de rotación se extrae una nueva muestra en cada ocasión, sin aparear. Con muestreo semanal o mensual, este plan es apropiado, cuando son primordiales las estimaciones anuales (y en menor medida las semestrales o trimestrales) como, por ejemplo, en una encuesta sanitaria con énfasis en las enfermedades crónicas. Si el cuestionario obtiene para cada unidad los resultados del mes anterior y los del mes en curso, podemos considerar las estimaciones compuestas de la forma

$$\bar{y}_h' = \bar{y}_h + \phi_h(\bar{y}'_{h-1} - \bar{y}_{h-1,h}) \tag{12.91}$$

donde $\bar{y}_h$ = estimación de datos actuales en la muestra actual.

$\bar{y}_{h-1,h}$ = estimación hecha sobre datos del mes anterior en la muestra actual.

$\bar{y}'_{h-1}$ = estimación compuesta para el mes anterior.

La teoría es discutida por Hansen, Hurwitz y Madow (1953), y Woodruff (1959), que la aplican a una encuesta de ventas al menudeo, y por Eckler (1955). En la encuesta de ventas al menudeo, la estimación compuesta involucra una estimación de razón, que es de la forma

$$\bar{y}_h'' = (1 - W)\bar{y}_h + W\left(\frac{\bar{y}_h}{\bar{y}_{h-1,h}}\right)\bar{y}''_{h-1}$$

donde W es un factor ponderante. Como las correlaciones de mes-a-mes son muy elevadas al promediar cerca de 0.98, las ganancias en precisión serán sustanciales. Un mes más tarde se calcula una estimación compuesta realizada para el mes h, que usa los resultados para el mes h de la nueva muestra tomada en el mes $(h + 1)$.

Con este método es esencial que los datos obtenidos del mes precedente sean exactos en la muestra actual. Esto podría no ser así, cuando los datos dependen de la memoria no registrada del entrevistado, sin embargo, el método puede tener éxito si los datos son del tipo que el entrevistado registra cuidadosamente por rutina.

EJERCICIOS

12.1. Se destinan \$3 000 a una encuesta para estimar una proporción. La encuesta principal costará \$10 por unidad de muestreo. Se dispone de información en registros, a un costo de \$0.25 por unidad de muestreo, que permite la clasificación de las unidades en dos estratos de tamaños casi iguales. Si la proporción verdadera es 0.2 en el estrato 1 y 0.8 en el estrato 2, estime n y n' óptimas y el valor resultante de $V(p_{st})$. ¿Produce el muestreo doble alguna ganancia en precisión sobre el muestreo simple? (Las razones n'/N, n_h/N_h se pueden ignorar.)

12.2. Para las W_h, P_h en el ejercicio 12.1. encuentre las razones de costo $c_n/c_{n'}$ para las cuales el muestreo doble es más económico que el simple.

12.3. Una población contiene L estratos de igual tamaño. Si V_{ran} denota la varianza de la media de una muestra aleatoria simple y V_{st}, V_{ds} son las varianzas correspondientes para muestreo aleatorio estratificado con asignación proporcional y para muestreo doble con estratificación, demuestre que, aproximadamente,

$$nV_{ran} = \bar{S}_h^2 + \frac{\sum_h (\bar{Y}_h - \bar{Y})^2}{L}$$

$$nV_{st} = \bar{S}_h^2$$

$$nV_{ds} = \bar{S}_h^2 + \frac{n}{n'}\frac{\sum_h (\bar{Y}_h - \bar{Y})^2}{L}$$

donde $\bar{S}_h^2$ es la varianza promedio dentro de estratos. (N y n' se pueden considerar ambas como grandes en relación con L y las n_h en muestreo doble pueden suponerse iguales a n/L.)

Por lo tanto, si $(\mathrm{RP})_{st}$ denota la precisión relativa de la muestra estratificada con la muestra aleatoria simple, con una definición correspondiente para $(\mathrm{RP})_{ds}$ demuestre que

$$(\mathrm{RP})_{ds} = \frac{(\mathrm{RP})_{st}}{1 + (n/n')[(\mathrm{RP})_{st} - 1]}$$

Para $(\mathrm{RP})_{st} = 2$. trace $(\mathrm{RP})_{ds}$ contra n/n'. ¿Qué tan pequeña debe ser esta razón para que $(\mathrm{RP})_{ds} = 1.9$?

12.4. Si $\rho = 0.8$ en muestreo doble para regresión, ¿qué tan grande debe ser n' con relación a n, si la pérdida en precisión debida a errores de muestreo en la media de la muestra grande se quiere menor al 10%?

12.5. En una aplicación de muestreo doble para regresión, la muestra pequeña fue de tamaño 87 y la grande de tamaño 300. Los siguientes cálculos se aplican a la muestra pequeña:

$$\sum (y_i - \bar{y})^2 = 17\,283, \qquad \sum (y_i - \bar{y})(x_i - \bar{x}) = 5114, \qquad \sum (x_i - \bar{x})^2 = 3248$$

Calcule el error estándar de la estimación de la regresión de $\bar{Y}$.

12.6. Para $\rho = 0.95$, verifique los datos que aparecen en la Tabla 12.4 para el porcentaje óptimo que tendría que ser apareado y para la ganancia en precisión en relación con no apareamiento. Calcule los correspondientes porcentajes de ganancias en precisión si se retiene una tercera parte de las unidades de la primera a la segunda ocasión y la mitad de las unidades se retiene en cada ocasión subsiguiente.

12.7. En muestreo aleatorio simple en dos ocasiones, suponga que la estimación en la segunda ocasión es, en la notación de la Sec. 12.11,

$$\bar{y}_2'' = (1-\phi)(\bar{y}_1 + \bar{y}_{2m} - \bar{y}_{1m}) + \phi\bar{y}_{2u}$$

(*a*) Al ignorar la cpf, muestre que

$$V(\bar{y}_2'') = \frac{S^2}{n}\left\{(1-\phi)^2\frac{[1+\mu(1-2\rho)]}{\lambda} + \frac{\phi^2}{\mu}\right\}$$

donde $\lambda = m/n$, $\mu = u/n$. (*b*) Para ρ, λ, μ, dados encuentre el valor de ϕ que minimiza $V(\bar{y}_2'')$. Muestre que si ρ excede $\frac{1}{2}$ la mejor ponderación ϕ yace entre μ y $\mu/(1+\mu)$.

12.8. Para $\mu = \frac{1}{4}$, $\mu = \frac{1}{2}$, $\rho = 0.8$ y $\rho = 0.9$, compare $V(\bar{y}_2'')$ en el ejercicio anterior con la varianza de la estimación compuesta óptima de la regresión $\bar{y}_2'$, según se da en la Ec. 12.74. (En $\bar{y}_2''$ tome $\phi = 0.2$ cuando $\mu = \frac{1}{4}$ y $\phi = 0.4$ cuando $\mu = \frac{1}{2}$.) Verifique que para estos valores de ρ la estimación $\bar{y}_2''$ es casi tan precisa como $\bar{y}_2'$ para ambas $\mu = \frac{1}{4}$ y $\mu = \frac{1}{2}$.

12.9. Una muestra independiente de tamaño n se toma cada mes. De la muestra tomada en cualquier mes se toman datos para el mes actual y para el anterior. Se forma una estimación compuesta $\bar{y}_h'$ como en (12.91), Sec. 12.13.

$$\bar{y}_h' = \bar{y}_h + \phi_h(\bar{y}'_{h-1} - \bar{y}_{h-1,h})$$

El modelo es

$$y_{hi} = \bar{Y}_h + \rho(y_{h-1,i} - \bar{Y}_{h-1}) + e_{hi}$$

donde e_{hi} es independiente de las y y tiene varianza $(1-\rho_2)$. Muestre que

(*a*) $\bar{y}_h' - \bar{Y}_h = \bar{e}_h + \phi_h(\bar{y}'_{h-1} - \bar{Y}_{h-1}) + (\rho - \phi_h)(\bar{y}_{h-1,h} - \bar{Y}_{h-1})$

(*b*) Si $V(\bar{y}_h') = g_h S^2/n$, donde S^2 es constante en todas las ocasiones,

$$g_h = (1-\rho^2) + \phi_h^2 g_{h-1} + (\rho - \phi_h)^2$$

(*c*) El óptimo $\phi_h = \rho/(1+g_{h-1})$ y el óptimo resultante g_h es

$$g_h = 1 - \frac{\rho^2}{1+g_{h-1}}$$

(*d*) El g_h limitante es $g_\infty = \sqrt{1-\rho^2}$. Estos resultados los dio Eckler (1955).

12.10. Si $E_b = V(\bar{y})/V(\bar{y}_b)$ y $E_R = V(\bar{y})/V(\bar{y}_R)$ son las eficiencias relativas de las estimaciones de regresión lineal y de razón de la media de una muestra aleatoria simple, demostrar que para $\bar{y}_b$ y para $\bar{y}_R$ la eficiencia relativa correspondiente a $\bar{y}$ en muestreo doble con selección óptima de n/n' es (ignórese $1/N$),

$$E_{ds} = E\Big/\left(1+\sqrt{\frac{c'}{c}}\sqrt{E-1}\right)^2$$

Por lo tanto, debe observarse que con cualquiera de estos estimadores, el muestreo doble no será altamente efectivo a menos que c'/c sea pequeño (como $< 1/10$). Por ejemplo, con $c'/c = 1/10$, $E = 6$ da $E_{ds} = 2.1$.

12.11. En muestreo en dos ocasiones, supóngase que $S_1 = S_2 = S$ y que las muestras son grandes, de modo que los coeficientes de regresión de y_{2i} respecto a y_{1i} y de y_{1i} respecto a y_{2i} en la parte apareada de las muestras en las dos ocasiones, son ambos efectivamente iguales a ρ. La estimación $\bar{y}_2'$ en la Sec. 12.10 y una estimación análoga $\bar{y}_1'$ se construyen usando la regresión de y_{1i} respecto a y_{2i}. Demostrar que

(*i*) $$V(\bar{y}_2' - \bar{y}_1') = \frac{2S^2(1-\rho)}{(n-u\rho)}$$

(*ii*) $$V(\bar{y}_2' + \bar{y}_1') = \frac{2S^2(1+\rho)}{(n+u\rho)}$$

[Una manera de hacerlo, es expresar $(\bar{y}_2' \pm \bar{y}_1')$ como funciones lineales de $(\bar{y}_{2m} \pm \bar{y}_{1m})$ y $(\bar{y}_{2u} \pm \bar{y}_{1u})$, que no están correlacionadas].

Nótese que, como lo sugiere la intuición, (i) se minimiza cuando $u = 0$ mientras que (ii) se minimiza cuando $u = n$.

12.12. El caso más favorable para la aplicación del método en el Ejercicio 12.1 ocurre cuando, en el estrato 1, la verdadera proproción es 0. Al estimar el número total de unidades Y_1 en la población que tienen un atributo costoso de medir, esto ocurre cuando hay un segundo atributo, barato de medir tal que sólo unidades que tengan el segundo atributo pueden poseer el primero. En una muestra aleatoria simple de tamaño n', cuéntese el número m' que tienen el atributo 2. Extráigase una submuestra de tamaño $\nu m' = m'/k$ de éstas, y cuéntese el número r que tiene el atributo 1.

(*a*) De los Teoremas 12.1 y 12.2 mostrar que $\hat{Y}_1 = Nkr/n$ es una estimación insesgada de Y_1 con varianza

$$V(\hat{Y}_1) = \frac{N^2P_1}{n'}\left[Q_1\left(1-\frac{n'}{N}\right) + \frac{P_2(k-1)}{P_1+P_2}\right]$$

donde P_1, (P_1+P_2) son las proporciones de población que tienen los atributos 1, 2. Supóngase que $1/N(P_1+P_2)$ se puede ignorar.

(*b*) Con $c = 10$, $c' = 1$, el investigador conjetura que $P_1 = 0.25$. $P_2 \doteq 0.15$. ¿Es provechoso el muestreo doble en este caso?

CAPITULO 13

Fuentes de Error en las Encuestas

13.1 INTRODUCCION

La teoría presentada en capítulos precedentes supone todo el tiempo que se usa alguna clase de muestreo probabilístico y que la observación y_i en la i-ésima unidad es el valor correcto para esa unidad. El error de la estimación emana solamente de la variación aleatoria de muestreo que está presente cuando se miden n de las unidades en lugar de la población completa de N unidades.

Estas suposiciones son razonables para los tipos más simples de muestreo, en los cuales los medios de medición son exactos y la calidad del trabajo es alta. En encuestas complejas, particularmente cuando intervienen difíciles problemas de medición, los supuestos pueden estar muy lejos de la verdad. Pueden presentarse tres fuentes adicionales de error que son las siguientes:

1. Fracaso en medir algunas de las unidades en la muestra escogida. Esto puede ocurrir por omisión, o, con poblaciones humanas, debido al fracaso en la localización de algunos individuos o su renuencia a contestar las preguntas una vez que se les localiza.
2. Errores de medición en una unidad. El medio de medición puede estar sesgado o ser impreciso. Con poblaciones humanas los entrevistados pueden no poseer información exacta o dar contestaciones sesgadas.
3. Errores introducidos en la edición, codificación y tabulación de los resultados.

Estas fuentes de error necesitan una modificación de la teoría estándar de muestreo. Los principales objetivos de tal modificación son proveer una guía respecto a la distribución de recursos entre la reducción de errores aleatorios de muestreo y la reducción de los

otros errores, así como desarrollar métodos para calcular errores estándar y límites de confianza que permanezcan válidos aun cuando estén presentes los otros errores.

13.2 EFECTOS DE LA NO-RESPUESTA

Usaremos el término *no-respuesta* para referirnos a la falla en la medición de algunas de las unidades en la muestra seleccionada. En el estudio de la no-respuesta es conveniente pensar en la población como dividida en dos "estratos", el primero consistente de todas las unidades para las cuales serían obtenidas las mediciones si las unidades llegaran a caer en la muestra y el segundo, de las unidades para las cuales no se obtendrían mediciones. Las composiciones de los dos estratos dependen íntimamente de los métodos que se usen para localizar las unidades y obtener los datos. Una encuesta en la cual se hacen al menos tres llamadas, si fuera necesario, en cada casa y en la que un supervisor con poderes excepcionales de persuasión va a visitar a todas las personas que se rehúsan a dar datos, tendrá un estrato de "no-respuesta" mucho menor que una en la cual solamente se hace un solo intento por casa.

Esta división en dos estratos distintos es, desde luego, una sobresimplificación. La casualidad juega su parte en la determinación de si una unidad se encuentra y se mide en un número dado de intentos. En una especificación más completa del problema, nosotros asignaríamos a cada unidad una probabilidad que representara la oportunidad que tiene de ser medida por un método de campo dado, si cayera en la muestra.

La muestra no provee información acerca del estrato 2 de no-respuestas. Esto no importaría si se pudiera suponer que las características del estrato 2 son las mismas que las del 1. Sin embargo, dondequiera que se han hecho comprobaciones se ha encontrado que las unidades en el estrato "no-respuesta" difieren de las unidades que son mensurables. Una ilustración aparece en la Tabla 13.1. Los datos provienen de un muestreo experimental de huertos de frutales en Carolina del Norte en 1946. Tres envíos sucesivos por correo del mismo cuestionario se hicieron a los productores. Para una de las preguntas —número de árboles frutales— había datos completos para la población (Finkner, 1950).

La declinación constante en el número de árboles frutales por productor en las respuestas sucesivas es evidente, siendo estos números 456 para las respuestas al primer envío, 382 en el segundo envío, 340 en el tercero y 290 para los que no respondieron a las

TABLA 13.1. RESPUESTAS A TRES SOLICITUDES EN UNA ENCUESTA POR CORREO

	Número de productores	% de la población	Número promedio de árboles por productor
Respuesta al primer envío	300	10	456
Respuesta al segundo envío	543	17	382
Respuesta al tercer envío	434	14	340
Personas que no responden después de 3 envíos	1839	59	290
Población total	3116	100	329

tres cartas. La respuesta total fue pobre, más de la mitad de la población no dio la información aun después de los tres intentos.

Consideremos ahora los efectos de la no-respuesta sobre la estimación de muestra. Sean N_1, N_2 los números de unidades en los dos estratos y $W_1 = N_1/N$, $W_2 = N_2/N$, de tal manera que W_2 es la proporción de no-respuesta en la población. Supongamos que se toma una muestra aleatoria simple de la población. Una vez que se completa el trabajo de campo, tenemos datos de una muestra aleatoria simple del estrato 1 pero no tenemos datos del estrato 2. Por lo tanto, la cantidad de sesgo en la media de la muestra es

$$E(\bar{y}_1) - \bar{Y} = \bar{Y}_1 - \bar{Y} = \bar{Y}_1 - (W_1\bar{Y}_1 + W_2\bar{Y}_2)$$
$$= W_2(\bar{Y}_1 - \bar{Y}_2) \qquad (13.1)$$

La cantidad de sesgo es el producto de la proporción de no-respuesta y la diferencia entre las medias de los dos estratos. Dado que la muestra no provee información acerca de $\bar{Y}_2$, el tamaño del sesgo es desconocido a menos que se puedan establecer límites para $\bar{Y}_2$ a partir de alguna fuente que no sean los datos de la muestra. Con una variable continua, los únicos límites que se pueden asignar con certeza a menudo son tan amplios que tienen poca utilidad.

Consecuentemente, con datos continuos, cualquier proporción apreciable de no-respuesta usualmente hace imposible la asignación de límites de confianza útiles a $\bar{Y}$ a partir de los resultados de la muestra. Quedamos en la posición de tener que confiar en alguna conjetura acerca del tamaño del sesgo, sin datos que apoyen semejante conjetura.

En el muestreo para porporciones, la situación es un poco más fácil, ya que la proporción desconocida P_2 en el estrato 2 debe estar entre 0 y 1. Si W_2 es conocida, estos límites de P_2 nos capacitan para la construcción de límites de confianza para la proporción P

de la población. Supóngase que se toma una muestra aleatoria simple de n unidades y que se obtienen mediciones para n_1 de las unidades en la muestra. Suponiendo n_1 suficientemente grande, los límites de confianza al 95% para P_1 están dados por

$$p_1 \pm 2\sqrt{p_1 q_1 / n_1}$$

donde p_1 es la proporción de la muestra y la cpf se ignora.

Cuando tratamos de derivar una declaración de confianza acerca de P, estaremos en terrenos seguros si suponemos $P_2 = 0$ en la búsqueda de $\hat{P}_L$ y $P_2 = 1$ cuando buscamos $\hat{P}_U$. Entonces podemos tomar, para los límites al 95%,

$$\hat{P}_L = W_1(p_1 - 2\sqrt{p_1 q_1 / n_1}) + W_2(0) \tag{13.2}$$

$$\hat{P}_U = W_1(p_1 + 2\sqrt{p_1 q_1 / n_1}) + W_2(1) \tag{13.3}$$

Es fácil verificar que estos límites son conservadores, esto es, que

$$Pr(\hat{P}_L \leq P \leq \hat{P}_U) > 0.95$$

Los límites se pueden estrechar un poco mediante un argumento más cuidadoso (Cochran, Mosteller y Tukey, 1954, Pág. 280), dado que P_2 no puede ser 0 y 1 simultáneamente, como antes se supuso.

Los límites son demasiado amplios a menos que W_2 sea muy pequeña. La Tabla 13.2 muestra los límites promedio para una muestra de tamaño $n = 1\,000$ y una serie de valores de W_2 y p_1. Como los límites en (13.2) y (13.3) dependen del valor de n_1 (número de entrevistados en la muestra), hemos tomado $n_1 = nW_1$, su valor promedio, para los cálculos en la Tabla 13.2.

El rápido aumento en la amplitud del intervalo de confianza con una W_2 creciente es evidente. Es interesante examinar cuáles valores de n serían necesarios para dar las mismas amplitudes de intervalo de confianza si W_2 fuera cero. Esto se hace fácilmente cuando p_1 es 50%. Para $W_2 = 5\%$, la Tabla 13.2 muestra que la amplitud media del intervalo de confianza es 5.6. El tamaño de

TABLA 13.2. LÍMITES DE CONFIANZA AL 95% PARA P (%) CUANDO $n = 1\,000$

% No-respuesta $100W_2$	Porcentaje de muestra, $100\,p_1$			
	5	10	20	50
0	(3.6, 6.4)	(8.1, 11.9)	(17.5, 22.5)	(46.7, 53.2)
5	(3.4, 11.1)	(7.6, 16.3)	(16.5, 26.5)	(44.4, 55.6)
10	(3.2, 15.8)	(7.2, 20.8)	(15.6, 30.4)	(42.0, 58.0)
15	(3.0, 20.5)	(6.8, 25.2)	(14.7, 34.3)	(39.6, 60.4)
20	(2.8, 25.2)	(6.3, 29.7)	(13.7, 38.3)	(37.2, 62.8)

muestra equivalente n_e, al suponer que no hay no-respuesta, se encuentra de la ecuación

$$5.6 = 2\sqrt{(50)(50)/n_e}$$

$$n_e = 320$$

Para $W_2 = 10$, 15 y 20%, los valores de n_e son 155, 90 y 60 respectivamente. Es evidente la conveniencia de dedicar una porción sustancial de los recursos a la reducción de la no-respuesta.

Si la tasa de no-respuesta de la población es desconocida, como sucederá de costumbre pueden calcularse límites conservadores de confianza de los datos de muestra por un método que sugirió un estudiante. Al calcular el límite inferior, supone que todos los no-respondientes de la muestra hubiesen dado una respuesta negativa. Al calcular el límite superior, suponer que todos los que rehusaron responder hubiesen dado una respuesta positiva. Por ejemplo, suponer $n = 1\,000$, $n_1 = 800$ y $p_1 = 10\%$, de modo que 80 miembros de la muestra dan una respuesta positiva y la tasa muestral de no-respuesta es 20%. Entonces en porcentajes,

$$\hat{P}_L = 8 - 2\sqrt{(8)(92)/1000} = 6.3\%$$

$$\hat{P}_U = 28 + 2\sqrt{(28)(72)/1000} = 30.8\%$$

Los límites son un poco más amplios que para $p_1 = 10\%$. $W_2 = 20\%$, en la Tabla 13.2.

Si se conoce W_2 de experiencia previa en el tipo particular de encuesta, Birnbaum y Sirken (1950a, 1950b) dan un método para encontrar el tamaño n de muestra que garantiza con riesgo α, un error absoluto en la proporción de muestra menor que una cantidad especificada d. No se supone conocimiento anticipado de P_1, P_2, o P. Si no hubiera no-respuesta tomaríamos (según la Sec. 4.4),

$$n = t_\alpha^2 PQ / d^2 \tag{13.4}$$

donde t_α es el desvío normal correspondiente al riesgo α de que el error exceda d. Sin información previa sobre P, tomaríamos $P = 0.5$ como el caso menos favorable, lo que da

$$n = \frac{t_\alpha^2}{4d^2} \tag{13.5}$$

Al tomar la combinación menos favorable del sesgo $W_2(P_1 - P_2)$ y el valor de P_1, Birnbaum y Sirken muestran que un valor de n que aún garantiza un error menor que d, con riesgo α, es

$$n \doteq \frac{t_\alpha^2}{4d(d - W_2)W_1} - 1 \tag{13.6}$$

TABLA 13.3. VALOR MÍNIMO DE n PARA LÍMITE DADO DE ERROR d, CON RIESGO $\alpha = 0.05$

% No-respuesta $100W_2$	d(%) 20	15	10	5
0	24	43	96	384
2	27	50	122	653
4	31	60	166	2000
6	36	75	255	....
8	43	99	521	....
10	53	142	...	....
15	112	...	...	....

Nótese que no hay valor de n que satisfaga si $W_2 > d$. Si $W_2 = 0$, esta ecuación se reduce a (13.5) aparte del término -1, el cual viene de una aproximación en el análisis. Algunos valores de n dados por el método de Birnbaum y Sirken se muestran en la Tabla 13.3.

Esta tabla cuenta la misma triste historia que la Tabla 13.2. Si nos contentamos con una estimación imprecisa ($d = 20$), cantidades de no-respuesta hasta de 10% se pueden manejar duplicando el tamaño de la muestra. Sin embargo, cualquier porcentaje considerable de no-respuesta hace imposible o muy costoso alcanzar una precisión altamente garantizada mediante el aumento del tamaño de la muestra entre los entrevistados.

13.3 TIPOS DE NO-RESPUESTA

En las secciones subsiguientes se describen algunos métodos para el manejo del problema de la no-respuesta. Una clasificación aproximada de los tipos de no-respuesta es la siguiente.

1. *No-cubrimiento.* Falla en la localización o la visita a algunas unidades de la muestra.

Este es un problema con unidades de muestreo que son áreas, en las cuales el entrevistador debe encontrar y enlistar todas las viviendas (de acuerdo con alguna definición) en una manzana de la ciudad. También surge del uso de listas incompletas. Algunas veces el mal tiempo o las escasas facilidades de transporte hacen imposible alcanzar ciertas unidades durante el periodo de la encuesta.

2. *Los no-en-casa.* Este grupo contiene personas que residen en el lugar pero que se encuentran temporalmente fuera de casa. Las familias en las cuales ambos padres trabajan y las familias sin niños

son más difíciles de alcanzar, que las familias con niños muy pequeños o con personas de edad confinadas en la casa.

3. *Incapaz de contestar.* El entrevistado puede no tener la información requerida para ciertas preguntas o puede mostrarse poco inclinado a darla. Una redacción hábil del cuestionario y su prueba previa son una salvaguardia.

4. *Los "hueso duro".* Las personas que inexorablemente rechazan ser entrevistados, que están incapacitadas o que están fuera de casa durante todo el tiempo disponible para el trabajo de campo, son las que constituyen este sector. Esto representa una fuente de sesgo que persiste, sin importar cuánto esfuerzo se ponga en la perfección de las listas.

La identificación y medida del no-cubrimiento son difíciles. Con muestreo de áreas, un método es revisitar las unidades primarias, al hacer una lista cuidadosa que sirve como comprobación. Las comparaciones de los recuentos de los números de personas o habitaciones con los de otra encuesta indican a veces que algunas se han omitido. Cuando el marco principal es un directorio de direcciones por calles, el mismo se puede completar con muestra de área, cuyo propósito es muestrear parte de la ciudad (p. ej., edificios nuevos) no cubierta adecuadamente por el directorio, y en partes donde el directorio parece exacto, buscar direcciones no incluidas en el mismo. Las encuestas que emplean estos métodos, con una discusión del problema de no-cubrimiento, son descritas por Kish y Hess (1958) y Woolsey (1956).

Con respecto a los no-en-casa, el problema es más fácil en encuestas en las cuales *cualquier* adulto en la casa es capaz de contestar las preguntas, que en aquellas en las que se entrevista un solo adulto escogido al azar. Usualmente es preferible un solo adulto si la encuesta se refiere a individuos en la cual una persona no puede responder con exactitud por otra o si se espera que las correlaciones intrahogareñas sean altas, de tal manera que la medida de más de una persona por familia es antieconómica. A este respecto, un método útil de seleccionar una sola persona por cada casa fue desarrollado por Kish (1949). El plan original fue diseñado para casas que pudieran contener hasta seis personas elegibles, pero el procedimiento puede adaptarse a familias mayores o menores.

El entrevistador enlista en el inventario a las personas elegibles en la familia y luego las numera: primero los hombres en orden decreciente de edad, luego las mujeres en el mismo orden. Cada inventario tiene impreso *uno* de los grupos de instrucciones en la Tabla 13.4.

TABLA 13.4. INSTRUCCIONES PARA LA SELECCIÓN DE UN SOLO ENTREVISTADO

Frecuencia relativa de uso	Tabla número	Si el número de adultos en la familia es					
		1	2	3	4	5	6
		Seleccione el adulto numerado					
1/6	A	1	1	1	1	1	1
1/12	B1	1	1	1	1	2	2
1/12	B2	1	1	1	2	2	2
1/6	C	1	1	2	2	3	3
1/6	D	1	2	2	3	4	4
1/12	E1	1	2	3	3	3	5
1/12	E2	1	2	3	4	5	5
1/6	F	1	2	3	4	5	6

Cada persona elegible en una familia de un tamaño dado tiene igual probabilidad de ser seleccionada, excepto que los adultos 3 y 5 en familias de tamaño 5 están ligeramente sobrerrepresentados. Dado que los entrevistados masculinos están concentrados en las Tablas A, B y C, el entrevistador puede dedicar las visitas vespertinas a las familias así designadas.

13.4 REVISITAS

Una técnica estándar es especificar el número de revisitas, o un número mínimo, que deben hacerse en cualquier unidad antes de abandonarla como "contacto imposible". Stephan y McCarthy (1958) presentan los datos de un cierto número de encuestas sobre el porcentaje de la muestra total obtenido en cada visita. Los resultados promedio se muestran en la Tabla 13.5

TABLA 13.5. NÚMERO DE VISITAS REQUERIDAS PARA ENTREVISTAS COMPLETAS

Entrevistado	% de la muestra con el cual se estableció contacto en la Primera visita	Segunda visita	Tercera o posterior visita	Porcentaje de No-respuesta	Total
Cualquier adulto*	70	17	8	5	100
Adulto al azar	37	32	23	8	100

* Dos encuestas en las cuales el entrevistado fue un ama de casa y un operario dé una finca, respectivamente, han sido incluidas en el grupo "cualquier adulto".

En encuestas en las cuales cualquier adulto de la familia podría contestar las preguntas, la primera visita obtuvo cerca del 70% de la muestra y las dos primeras visitas, 87%. El costo creciente del muestreo cuando hay que entrevistar a un adulto escogido al azar es evidente, al producir la primera visita apenas 37% de las entrevistas requeridas. El marcado éxito de la segunda visita refleja el trabajo previo del entrevistador, al investigar con anterioridad cuándo el entrevistado estaría en casa y disponible.

Muy poco se ha publicado sobre los costos relativos de las visitas posteriores a la primera. Debe esperarse que estas visitas sean más costosas por entrevista completa, ya que las casas están más esparcidas en el área asignada al entrevistador y dado que sus ocupantes son presuntamente personas que pasan más que una cantidad promedio de su tiempo fuera de casa. Basado en experiencias británicas, Durbin (1954) sugiere que las visitas ulteriores pueden ser menos costosas de lo que podría anticiparse. Las cifras siguientes muestran costos relativos estimados por entrevista completa (por ejemplo: dinero gastado en i-ésimas visitas dividido por el número de entrevistas obtenidas) para cada visita hasta la quinta, en un estudio especial relatado por Durbin y Stuart (1954).

TABLA 13.6. COSTOS RELATIVOS POR NUEVA ENTREVISTA COMPLETA EN LA i-ÉSIMA VISITA

Visita	1	2	3	4	5
Costo relativo	100	112	127	151	250

La estimación de estos costos requiere sumo cuidado. Si el entrevistado deseado no está en casa durante la primera visita, el entrevistador puede gastar algún tiempo inquiriendo cuándo estará en casa y concertando una cita tentativa. Al costear ese tiempo se debe asignar a la segunda visita y no a una primera que no tuvo éxito.

Una medida más útil es el costo promedio por entrevista completa *sobre todas las entrevistas obtenidas hasta la i-ésima visita.* Estas cifras dan los costos relativos de obtener n entrevistas completas cuando insistimos en i visitas, antes que el entrevistador se dé por vencido. Para calcular estas cifras, debemos saber cuántas entrevistas se obtienen en cada visita. En la Tabla 13.7 se hacen estos cálculos con base en dos tipos de suposiciones. El primero simula encuestas en las que cualquier adulto puede contestar las preguntas, el segundo tipo aquellas encuestas que demandan un adulto al azar. Los datos sobre los números de entrevistas obtenidas se tomaron de la Tabla 13.5.

TABLA 13.7. COSTOS RELATIVOS POR ENTREVISTA COMPLETA HASTA LA i-ÉSIMA VISITA

Entrevistado = Cualquier adulto

		i-ésima visita		Hasta i-ésima visita			Adulto "al azar"	
Visita	Costo relativo	No. de ents.*	Costo de ents.	No. total de ents.	Costo total	Costo por ent.	No. de ents.	Costo por ent.
1	100	$0.70n_o$	$70n_o$	$0.70n_o$	$70n_o$	100	$0.37n_o$	100
2	112	$0.17n_o$	$19.04n_o$	$0.87n_o$	$89.04n_o$	102	$0.32n_o$	106
3	127	$0.07n_o$	$8.89n_o$	$0.94n_o$	$97.93n_o$	104	$0.16n_o$	110
4	151	$0.04n_o$	$6.04n_o$	$0.98n_o$	$103.97n_o$	106	$0.09n_o$	114
5	250	$0.02n_o$	$5.00n_o$	$1.00n_o$	$108.97n_o$	109	$0.06n_o$	122

* Entrevistas.

Los detalles de los cálculos se muestran solamente para la primera suposición, siendo el método exactamente el mismo para la segunda. El símbolo n_o denota el tamaño *original* de la muestra.

El insistir hasta por tres visitas cuesta solamente 4% más *por entrevista completa* que las visitas únicas, si cualquier adulto es un prospecto satisfactorio; y solamente 10% más si debe entrevistarse a un adulto al azar. No se sabe qué tan típicos sean estos resultados, pero el método da estimaciones realistas del costo de la insistencia en revisitas si se han recolectado datos sobre el costo necesario y el tamaño de la muestra. También interviene el factor tiempo: las revisitas retrasan los resultados finales.

13.5 UN MODELO MATEMATICO DE LOS EFECTOS DE LAS REVISITAS

Deming (1953) desarrolló un modelo matemático útil y flexible para el examen más detallado de las consecuencias de diferentes políticas sobre revisitas. La población se divide en r clases, de acuerdo con la probabilidad de que el entrevistado se encuentre en casa. Sea

w_{ij} = probabilidad de que se encuentre un entrevistado en la j-ésima clase antes de la i-ésima visita.

p_j = proporción de la población que cae en la j-ésima clase.

μ_j = media de la característica para la j-ésima clase.

σ_j^2 = varianza de la característica para la j-ésima clase.

Por simplicidad, suponemos $w_{ij} > 0$ para todas las clases, aunque el método se adapta fácilmente para incluir personas imposibles de alcanzar. Si $\bar{y}_{ij}$ es la media para aquellos en la clase j que fueron alcanzados durante o antes de la i-ésima visita, también se supone que $E(\bar{y}_{ij}) = \mu_j$.

La media verdadera de la población para la característica es

$$\bar{\mu} = \sum_j p_j \mu_j \tag{13.7}$$

Considérese la composición de la muestra después de i visitas. Las personas en la muestra pueden clasificarse en $(r + 1)$ clases como sigue: en la primera clase y entrevistadas; en la segunda clase y entrevistadas; y así por el estilo. La $(r + 1)$-ésima clase consiste de todos aquellos aún no entrevistados después de i visitas. Si se ignora la cpf, los números que caen en estas $(r + 1)$ clases están distribuidos de acuerdo con la multinomial

$$[w_{i1}p_1 + w_{i2}p_2 + \cdots + w_{ir}p_r + (1 - \sum w_{ij}p_j)]^{n_o}$$

donde n_0 es el tamaño inicial de la muestra.

Resulta entonces que el número n_i de quienes han sido entrevistados en el curso de i visitas está binomialmente distribuido con el número de intentos = n_o y probabilidad de éxito $\Sigma w_{ij}p_j$. Por lo tanto

$$E(n_i) = \text{número esperado de entrevistas en } i \text{ visitas} = n_o \sum_j^r w_{ij}p_j \tag{13.8}$$

Para n_i fijos, los números de entrevistas n_{ij} obtenidas ($j = 1, 2, \ldots, r$) siguen una multinomial con probabilidades $w_{ij}p_j/\Sigma w_{ij}p_j$. Resulta que

$$E(n_{ij}|n_i) = \frac{n_i w_{ij} p_j}{\sum w_{ij} p_j}$$

Por lo tanto, si $\bar{y}_i$ es la media de la muestra obtenida después de i visitas,

$$E(\bar{y}_i|n_i) = E\left(\frac{\sum n_{ij}\bar{y}_{ij}}{n_i}\right) = \frac{\sum n_i w_{ij} p_j \mu_j}{n_i \sum w_{ij} p_j} = \frac{\sum w_{ij} p_j \mu_j}{\sum w_{ij} p_j} = \bar{\mu}_i \tag{13.9}$$

Dado que este resultado no depende de n_i, la media incondicional de $\bar{y}_i$ es también $\bar{\mu}_i$. El sesgo en la estimación $\bar{y}$ es, por lo tanto, $(\bar{\mu}_i - \bar{\mu})$.

La varianza condicional de $\bar{y}_i$ para n_i dado, se encuentra en forma similar que es

$$V(\bar{y}_i|n_i) = \frac{\sum_j^r w_{ij} p_j [\sigma_j^2 + (\mu_j - \bar{\mu}_i)^2]}{n_i \sum_j^r w_{ij} p_j} \tag{13.10}$$

La varianza incondicional, al ignorar los términos de orden $1/n_i^2$, está dada aproximadamente al reemplazar n_i en (13.10) por su valor esperado de (13.8).

Finalmente, el error cuadrático medio de la estimación obtenido después de i visitas es

$$\text{ECM}(\bar{y}_i|i) = V(\bar{y}_i|i) + (\bar{\mu}_i - \bar{\mu})^2 \tag{13.11}$$

Debe considerarse también el costo de hacer i visitas. El número esperado de entrevistas *nuevas* obtenidas en la k-ésima visita es $\sum (w_{kj} - w_{k-1,j}) p_j$. Por lo tanto, si c_k es el costo por entrevista completa en la k-ésima visita, el costo esperado de hacer i visitas es $n_0 C(i)$, donde

$$C(i) = c_1 \sum w_{1j} p_j + c_2 \sum (w_{2j} - w_{1j}) p_j + \cdots + c_i \sum (w_{ij} - w_{i-1,j}) p_j$$

Ejemplo. Una población con tres clases aparece en la Tabla 13.8. Se intenta que las p_j y w_{ij} representen encuestas en las cuales se entrevista a un adulto al azar. Durante la primera visita las probabilidades w_{ij} de obtener una entrevista se toman como 0.6, 0.3 y 0.1 en las tres clases. Durante la segunda y subsiguientes visitas, las probabilidades condicionales de entrevistar una persona omitida anteriormente son 0.9, 0.5 y 0.2. Estas cifras se hicieron mayores que las probabilidades correspondientes a la primera visita para representar el efecto de una investigación inteligente por parte del entrevistador.

El dato que se está estimando es un porcentaje binomial cercano a 50%. Se consideran dos grupos de μ_j (I, II). Por simplicidad, las varianzas dentro-de-clase $\sigma_j^2 = \mu_j(100 - \mu_j)$ se tomaron todas como 2 500. Los costos relativos por entrevista completa en visitas sucesivas son los dados en la Tabla 13.6.

La Tabla 13.9 muestra (a) el número total esperado de entrevistas obtenidas para un total de i visitas, (b) el costo promedio de estas visitas por entrevista, y (c) el sesgo $(\bar{\mu}_i - \bar{\mu})$ en la estimación $\bar{y}$ bajo las suposiciones I y II acerca de las μ_j.

TABLA 13.8. CARACTERÍSTICAS DE LAS TRES CLASES

	Clase 1	Clase 2	Clase 3
p_j	0.45	0.50	0.05
w_{ij}	$0.6 + (0.4)[1 - (0.1)^{i-1}]$	$0.3 + (0.7)[1 - (0.5)^{i-1}]$	$0.1 + (0.9)[1 - (0.8)^{i-1}]$
I μ_j	55	50	45
II μ_j	60	50	40

TABLA 13.9. NÚMERO DE ENTREVISTAS, COSTOS POR ENTREVISTA Y SESGOS

Número requerido de visitas	Número de entrevistas obtenidas	Costo promedio por entrevista	Sesgo I	Sesgo II
1	$0.425n_o$	100	+1.118	+2.235
2	$0.771n_o$	105	+0.711	+1.421
3	$0.882n_o$	108	+0.421	+0.842
4	$0.933n_o$	110	+0.266	+0.532
5	$0.960n_o$	114	+0.180	+0.360

En II, por ejemplo, la media verdadera de la población $\bar{\mu}$ es 54%. La media $\bar{\mu}_1$ obtenida de la primera visita es 56.235%, lo que da un sesgo de + 2.235% mostrado en la tabla. Una política que requiere tres visitas reduce este sesgo a + 0.842%.

Los valores del ECM ($\bar{y}$) obtenidos de un gasto dado de dinero se compararon para las diferentes políticas de revisita. En las primeras comparaciones la cantidad de dinero es suficiente para tomar $n_0 = 500$ si solamente se hace una visita. De la Tabla 13.9 el número esperado de entrevistas obtenidas durante la primera visita es $E(n_1) = (500)\ (0.425) = 212.5$. Si se hacen dos visitas, este número esperado debe ser reducido a $E(n_2) = 212.5/1.05 = 202.4$ para mantener el mismo costo, y en forma similar para 3, 4 y 5 revisitas. Estos valores de $E(n_i)$ se sustituyeron en la Ec. (13.10) para dar $V(\bar{y})$ y, por lo tanto, ECM ($\bar{y}$).

La Tabla 13.10 presenta los ECM que resultan para tres importes de desembolso, correspondientes a $n_0 = 500$, 1 000, 2 000 para una sola visita. Cuando $n_0 = 500$, también se dan los valores del ECM ($\bar{y}$) para la situación "no-sesgo" en la cual toda $\mu_j = 50$. Esta columna muestra el efecto de las revisitas cuando las mismas son innecesarias, ya que no resulta sesgo alguno de reducir la encuesta a una sola visita.

TABLA 13.10. VALORES DEL ECM ($\bar{y}$) PARA DIFERENTES POLÍTICAS DE REVISITA QUE CUESTAN LA MISMA SUMA

Número requerido de visitas	$n_o = 500$ (sólo para primeras visitas)			$n_o = 1000$		$n_o = 2000$	
	No sesgo	I*	II*	I	II	I	II
1	**11.8**	13.0	16.9	7.1	10.9	4.2	8.0
2	12.4	12.9	14.6	6.7	8.3	3.6	5.2
3	12.7	**12.9**	13.6	**6.5**	7.1	3.4	3.9
4	13.0	13.1	**13.4**	6.6	**6.9**	**3.3**	3.6
5	13.5	13.5	13.8	6.8	6.9	3.4	**3.5**

* Estas representan poblaciones con la más pequeña (I) y la más grande (II) cantidades de sesgo, según se definen en la Tabla 13.8.

Las políticas que dan los ECM más bajos se muestran en letra negrita. Considérese primero el tamaño de la muestra más pequeña, $n_0 = 500$. Si las revisitas no son necesarias, una política que reclama tanto como cuatro revisitas da por resultado sólo un modesto aumento en el ECM. En I, que comprende la menor cantidad de sesgo, las diferentes políticas producen aproximadamente la misma exactitud, aunque tres es el óptimo. En II, tres a cinco revisitas son satisfactorias, una sola visita da un ECM cercano al 25% sobre el mínimo.

Para los tamaños de muestra mayores el número óptimo de revisitas aumenta a cuatro o cinco y el uso de una sola visita resulta en pérdidas más sustanciales de exactitud.

Esta, desde luego, es solamente una ilustración. La importancia del método es que, conforme la información se acumula con relación a costos y sesgos relativos, se puede plantear una política económica para cualquier tipo de encuesta.

13.6 FRACCION OPTIMA DE MUESTREO ENTRE LOS QUE NO RESPONDEN

Después de hacer el primer intento para llegar a las personas de la muestra, otro método debido a Hansen y Hurwitz (1946), consiste en tomar una submuestra aleatoria de las personas no entrevistadas y hacer un gran esfuerzo para entrevistar a todos en la submuestra. Esta técnica se desarrolló primero para encuestas donde el intento inicial se hizo por correo, una submuestra de las personas que no contestaron el cuestionario completo se trató de alcanzar por el método más costoso que es la entrevista personal.

Este método se puede ver como una aplicación de la técnica del muestreo doble para estratificación, que se presentó en la Sec. 12.2. Se empieza por tomar una muestra aleatoria simple de n' unidades. Sea n_1' el número de unidades de la muestra que proporcionan los datos buscados, y n_2' el número en el grupo de no respuestas. Mediante esfuerzos más intensos, los datos se obtienen más tarde de una submuestra aleatoria $n_2 = v_2 n^{2\prime}$ entre los n_2'. Hansen y Hurwitz usan la notación $v_2 = 1/k$ de modo que $n_2 = n_2'/k$.

En el contexto de la Sec. 12.2 hay dos estratos. El estrato 1 consiste en aquellos que no responden al primer intento con una muestra medida de tamaño $n_1 = n_1'$, de modo que $v_1 = 1$. El estrato 2 son aquellos que responderían al segundo intento con $n_2 = n_1'/k$.

El costo de tomar la muestra es

$$c_0 n' + c_1 n_1' + \frac{c_2 n_2'}{k} \tag{13.12}$$

donde c son los costos por unidad: c_0 es el costo de hacer el primer intento, c_1 es el costo de procesar los resultados de este primer intento y c_2 es el costo de obtener y procesar los datos en el segundo estrato. Si W_1, W_2 son las proporciones de población de ambos estratos, el costo esperado es

$$C = c_0 n' + c_1 W_1 n' + \frac{c_2 W_2 n'}{k} \tag{13.13}$$

Como estimación de $\bar{Y}$, tomamos

$$\bar{y}' = w_1 \bar{y}_1 + w_2 \bar{y}_2 = \frac{(n_1' \bar{y}_1 + n_2' \bar{y}_2)}{n'} \tag{13.14}$$

donde $\bar{y}_1$, $\bar{y}_2$ son las medias de las muestras de tamaños $n_1 = n_1'$ y $n_2 = n_2'/k$. Por el Teorema 12.1, la estimación $\bar{y}'$ será insesgada, si las respuestas se obtienen de todas las submuestras aleatorias seleccionadas de tamaño $n_2 = n_2'/k$.

Por la Fórmula (12.3), la varianza de $\bar{y}'$ es

$$\begin{aligned} V(\bar{y}') &= \left(\frac{1}{n'} - \frac{1}{N}\right) S^2 + \frac{W_2 S_2^{\,2}}{n'} \left(\frac{1}{\nu_2} - 1\right) \\ &= \left(\frac{1}{n'} - \frac{1}{N}\right) S^2 + \frac{(k-1) W_2 S_2^{\,2}}{n'} \end{aligned} \tag{13.15}$$

Entonces, las cantidades n' y k se eligen para minimizar el producto $C(V + S^2/N)$. De (13.15) y (13.13) tenemos

$$V + S^2/N = \frac{(S^2 - W_2 S_2^{\,2})}{n'} + \frac{k W_2 S_2^{\,2}}{n'} \tag{13.16}$$

$$C = (c_0 + c_1 W_1) n' + \frac{c_2 W_2 n'}{k} \tag{13.17}$$

Por la desigualdad de Cauchy, el óptimo k es

$$k_{\text{opt}} = \sqrt{\frac{c_2 (S^2 - W_2 S_2^{\,2})}{S_2^{\,2} (c_0 + c_1 W_1)}} \tag{13.18}$$

El tamaño inicial de muestra n' se puede elegir ya sea para minimizar C con V especificada o V con C especificado, al resolver para n' de (13.16) o (13.17). Si V es especificada,

$$n'_{\text{opt}} = \frac{N[S^2 + (k-1) W_2 S_2^{\,2}]}{(NV + S^2)} \tag{13.19}$$

donde V es el valor especificado para la varianza de la media de población estimada.

Las soluciones exigen un conocimiento de W_2: a menudo, este puede estimarse de experiencias previas. Además de S^2, cuyo valor

debe estimarse de antemano en cualquier problema de "tamaño de muestra", las soluciones también involucran S_2^2 que es la varianza en el estrato de no respuestas. El valor de S_2^2 puede ser más difícil de predecir; probablemente no sea el mismo que S^2. Por ejemplo, en encuestas por correo en casi todos los tipos de empresa económica, los que responden tienden a ser los grandes operadores, con varianzas mayores entre unidades que los que no responden.

Si W_2 no se conoce bien, como una aproximación satisfactoria se busca el valor de $n'_{\acute{o}pt}$ de (13.18) y (13.19) provisionalmente para una cierta gama de valores supuestos de W_2 entre 0 y un límite superior seguro. El $n'_{\acute{o}pt}$ máximo de esta serie se toma como tamaño inicial de muestra n'. Cuando se reciben las respuestas a la encuesta por correo, se conoce el valor de n_2'. Al buscar el valor de k usado con este método, utilizamos la varianza $v_c(\bar{y}')$ condicional a los valores conocidos de n_2 y n'. Esta se puede obtener de (12.4) y (12.7) como

$$V_c(\bar{y}') = S^2\left(\frac{1}{n'} - \frac{1}{N}\right) + \frac{(k-1)n_2' S_2^2}{n'^2} \tag{13.20}$$

La Ec. (13.20) se resuelve para encontrar el k que da la varianza condicional deseada. Generalmente, el costo para este método es un poco más elevado que el costo óptimo para W_2 conocido.

Ejemplo. Este ejemplo es una condensación de la publicación por Hansen y Hurwitz (1946). La primera muestra se toma por correo y se espera que la proporción de respuestas sea 50%. La precisión deseada es aquella que daría una muestra aleatoria simple de tamaño 1 000 si no hubiera no-respuesta. El costo de enviar los cuestionarios por correo es 10 centavos y el costo de procesar el cuestionario completo es 40 centavos. Llevar a cabo una entrevista personal cuesta $4.10.

¿Cuántos cuestionarios deberían enviarse y qué porcentaje de los que no responden deberían ser entrevistados?

En términos de la función de costo (13.12) los costos por unidad en dólares son los siguientes:

c_0 = costo del primer intento = 0.1
c_1 = costo de procesar datos por respuesta = 0.4
c_2 = costo de obtener y procesar los datos por cada no-respuesta = 4.5

Los n y k óptimos se pueden obtener de (13.17) y (13.18). Si las varianzas S^2 y S_2^2 se suponen iguales y N se supone que es grande, entonces

$$k_{\acute{o}pt} = \sqrt{\frac{c_2(1-W_2)}{c_0 + c_1 W_1}} = \sqrt{\frac{(4.5)(0.5)}{0.1 + (0.4)(0.5)}} = \sqrt{7.5} = 2.739$$

$$n'_{\acute{o}pt} = \frac{S^2[1 + (k-1)W_2]}{V} = 1000\{1 + (1.739)(0.5)\}$$

$$= 1870$$

Nótese que hemos puesto $S^2/V = 1\,000$, o $V = S^2/1\,000$, ya que esta es la varianza que tendría la media de la muestra si se tomara una muestra de 1 000 y se obtuviera una respuesta completa.

Consecuentemente, deberían enviarse por correo 1 870 cuestionarios. De los 935 que no son devueltos, entrevistamos una muestra al azar de 935/2.739 o sea 341. El costo es $2 095.

Como lo ha señalado Durbin (1954), el submuestreo probablemente no muestre un beneficio marcado a menos que c_2 sea grande en relación con $(c_0 + c_1 W_1)$. Las dos cantidades son comparables, dado que $(c_0 + c_1 W_1)$ es el costo esperado por unidad de hacer el primer intento y procesar los resultados y c_2 tiene el mismo significado para el segundo intento. De las ecuaciones puede mostrarse que la razón del costo de obtener una V prescrita con $k = 1$ (sin submuestreo) al costo mínimo para k óptimo es

$$\frac{S^2(c_0 + c_1 W_1 + c_2 W_2)}{[\sqrt{(S^2 - W_2 S_2^2)(c_0 + c_1 W_1)} + \sqrt{c_2} W_2 S_2]^2} = \frac{c_0 + c_1 W_1 + c_2 W_2}{[\sqrt{W_1(c_o + c_1 W_1)} + \sqrt{c_2} W_2]^2}$$

si S^2 y S_2^2 son aproximadamente iguales. Si r es la razón de c_2 a $(c_0 + c_1 W_1)$ la razón de costo se convierte en

$$\frac{1 + rW_2}{(\sqrt{W_1} + \sqrt{r} W_2)^2}$$

Por ejemplo, la razón de costo es 1.029 para $W_1 = 0.5$ si $r = 4$, 1.146 si $r = 10$ y 1.228 si $r = 16$. Pero si S^2 es bastante más grande que S_2^2 hay más que ganar del submuestreo.

Con muestreo estratificado, los valores óptimos de los n'_h y los k_h en los estratos individuales son más complejos. Una buena aproximación consiste primero en estimar, por los métodos de las Secs. 5.5 y 5.9, los tamaños de muestra n_{oh} que se requerirían en los estratos si no hubiera no respuesta. Ahora, de (13.19) si $W_2 = 0$, tenemos

$$n_o = \frac{NS^2}{NV + S^2} \tag{13.21}$$

Luego (13.19) se puede escribir como

$$n'_{\acute{o}pt} = n_o\left[1 + \frac{(k-1)W_2 S_2^2}{S^2}\right] \tag{13.22}$$

Esta ecuación aplicada separadamente a cada estrato, da una aproximación al óptimo n_h'. Los valores de k_h se encuentran al aplicar (13.18) en cada estrato.

Estas técnicas se pueden usar con estimaciones de razón o regresión. Con la estimación de razón, las cantidades S^2 y S_2^2 se reem-

plazan por S_d^2 y S_{2d}^2, donde $d_i = y_i - Rx_i$. Con una estimación de regresión, S^2 se convierten en $S^2(1-\rho^2)$ y S_2^2 en $S_2^2(1-\rho^2)$.

13.7 AJUSTES POR SESGO SIN REVISITAS

Un método ingenioso para disminuir los sesgos presentes en los resultados de la primera visita lo sugirió Hartley (1946) y lo desarrollaron Politz y Simmons (1949-1950) y Simmons (1954). Supongamos que todas las visitas se hicieron por la noche y en días laborables. Al entrevistado se le pregunta si estaba en casa, en el momento de la entrevista, en cada una de las cinco noches precedentes de la semana. Si afirma que estaba en casa t noches de las cinco, se toma la razón $(t+1)/6$ como estimación de la frecuencia π con que está en casa durante las horas de entrevista.

Los resultados de la primera visita se ponen en seis grupos según el valor de t (0, 1, 2, 3, 4, 5). En el grupo t, sea n_t el número de entrevistas obtenidas y $\bar{y}_t$ la media de atributos. La estimación de Politz-Simmons de la media de población μ es

$$\bar{y}_{PS} = \frac{\sum_{t=0}^{5} 6n_t\bar{y}_t/(t+1)}{\sum_{t=0}^{5} 6n_t/(t+1)} \tag{13.23}$$

En este enfoque se reconoce que los resultados de la primera visita están indebidamente ponderados con personas que están en casa la mayor parte del tiempo. Dado que en promedio, una persona que está en casa una proporción π del tiempo, tiene una oportunidad relativa π de aparecer en la muestra, su respuesta debería recibir una ponderación $1/\pi$. La cantidad $6/(t+1)$ se usa como una estimación de $1/\pi$. Entonces $\bar{y}_{ps}$ es menos sesgada que la media de la muestra $\bar{y}$ de la primera visita, pero su varianza es mayor porque se reemplaza una media sin ponderar por una media con ponderaciones estimadas.

En la presentación de la media y la varianza de $\bar{y}_{ps}$, hacemos uso de la notación de la Sec. 13.5. La población se divide en clases, las personas en la j-ésima clase están en su casa una fracción πj del tiempo. Nótese que el t-ésimo grupo (p. ej., personas en casa t de las cinco noches precedentes) contendrá personas de varias clases. Sean n_{jt}, $\bar{y}_{jt}$ el número y la media del dato para aquellos en la clase j y el grupo t. Entonces, $\bar{y}_{ps}$ puede escribirse en la forma siguiente:

$$\bar{y}_{PS} = \frac{\sum\sum 6n_{jt}\bar{y}_{jt}/(t+1)}{\sum\sum 6n_{jt}/(t+1)} = \frac{N}{D} \quad \text{(digamos)} \tag{13.24}$$

Esta es una estimación de razón. En muestras grandes, su media es aproximadamente $E(N)/E(D)$.

Si n_o es el tamaño inicial de la muestra (respuestas más no-en-casa) y n_j es el número de entrevistados de la clase j, se hacen las siguientes suposiciones:

(i) $\dfrac{n_j}{n_o}$ es una estimación binomial de $p_j\pi_j$

(ii) $E(n_{jt}|n_j) = n_j \dfrac{5!}{t!(5-t)!}\pi_j^t(1-\pi_j)^{5-t}$

(iii) $E(\bar{y}_{jt}) = \mu_j$, para cualquier j y t

La suposición (ii) es cuestionable. Sin dar una explicación detallada, la misma supone que las personas responden correctamente si estaban o no en casa.

Para una j dada, al usar la suposición (ii),

$$E\sum_{t=1}^{5} n_{jt}\left(\frac{6}{t+1}\right) = n_j \sum_{t=1}^{5}\left(\frac{6}{t+1}\right)\frac{5!}{t!(5-t)!}\pi_j^t(1-\pi_j)^{5-t} \tag{13.25}$$

$$= \frac{n_j}{\pi_j}[1-(1-\pi_j)^6] \tag{13.26}$$

Por lo tanto,

$$E(D) = \sum_{j=1}^{r}\frac{E(n_j)}{\pi_j}[1-(1-\pi_j)^6] = n_o\sum_{j=1}^{r} p_j[1-(1-\pi_j)^6] \tag{13.27}$$

al usar la suposición (i). Además, dado que $E(\bar{y}_{jt}) = \mu j$ para cualquier j y t, esto da el resultado

$$E(\bar{y}_{PS}) = \bar{\mu}_{PS} \doteq \frac{\sum_{j=1}^{r} p_j\mu_j[1-(1-\pi_j)^6]}{\sum_{j=1}^{r} p_j[1-(1-\pi_j)^6]} \tag{13.28}$$

Dado que la media verdadera $\bar{\mu} = \sum p_j\mu_j$, algún sesgo permanece en $\bar{y}_{ps}$. En cierto sentido, esta estimación tiene el mismo sesgo que $\bar{y}_6$, la media de la muestra dada por el método de las revisitas, con el requisito de que se hagan cuando menos seis visitas si es necesario. Se mostró en la Sec. 13.5, Ec. 13.9, que el método de las revisitas con un total de i visitas, da una estimación insesgada de $\bar{\mu}_{ji} = \sum w_{ij}p_j\mu_j / \sum w_{ij}p_j$, donde w_{ij} es la probabilidad de que una persona en la clase j que cae en la muestra sea entrevistada. Ahora

$w_{1j} = \pi_j$. Si en visitas subsecuentes la probabilidad de encontrar en casa a una persona a quien no se alcanzó previamente permanece como π_j, entonces

$$w_{ij} = [1-(1-\pi_j)^i]$$

de tal modo que $\bar{\mu}_{PS} = \bar{\mu}_6$. Sin embargo, con el método de las revisitas, la probabilidad de una entrevista en una visita posterior puede ser mayor que π_j, como resultado de la información obtenida por el entrevistador durante la primera visita o las anteriores. En este caso, el método de las revisitas tiene menos sesgo después de seis visitas.

La varianza de $\bar{y}_{PS}$ es más bien complicada. Con la aproximación acostumbrada para una estimación de razón, puede expresarse, de acuerdo con Deming (1953), como

$$V(\bar{y}_{PS}) \doteq \frac{1}{n_o U}\{\sum \pi_j p_j B_j[\sigma_j^2 + (\mu_j - \bar{\mu}_{PS})^2] + (n_o - 1)\sum (\pi_j p_j)^2 (B_j - A_j^2)(\mu_j - \bar{\mu}_{PS})^2\}$$

de donde

$$U = 1 - \sum p_j (1-\pi_j)^6$$

$$A_j = \frac{1}{\pi_j}[1-(1-\pi_j)^6]$$

$$B_j = \sum_{t=0}^{5} \left[\frac{6}{(1+t)}\right]^2 \frac{5!}{t!(5-t)!} \pi_j^t (1-\pi_j)^{5-t}$$

Aunque esta expresión es difícil de evaluar sin aplicarla a poblaciones específicas, se pueden hacer dos comentarios. Si las μ_j no difieren grandemente, esto es, si el sesgo de las primeras visitas es moderado, el término dominante es el primero:

$$\frac{1}{n_o U}\sum \pi_j p_j B_j \sigma_j^2$$

Esta expresión tiende a ser 25 a 35% más alta que la varianza de la media sin ponderar de las primeras visitas. También, $V(\bar{y}_{PS})$ contiene un término que no decrece conforme n_o aumenta y llega a ser importante en muestras muy grandes.

Para resumir, las comparaciones hechas en poblaciones simuladas por Deming (1953), Durbin (1954) y el autor, sugieren que este método muestra la mejor ventaja en relación a las revisitas, cuando los sesgos provenientes de las visitas iniciales son sustanciales y la muestra es grande. Sin embargo, las reducciones en ECM para el mismo desembolso son pequeñas, a no ser que las revisitas

cuesten sustancialmente más de lo que aquí se afirma. La técnica Politz-Simmons tiene la ventaja de economizar tiempo. Los errores e imperfecciones en los valores de t, no considerados en el análisis, son una desventaja. El método puede aplicarse, según lo sugiere Simmons (1954), junto con varias revisitas.

Se han propuesto varios otros métodos para disminuir el sesgo causado por "no-en-casa". El de Bartholomew (1961) se aplica a una encuesta con dos visitas. Este supone que para aquellos que no están en casa durante la primera visita, el entrevistador puede hacer, por medio de una cuidadosa indagación, que la probabilidad de encontrarlos en la segunda visita sea aproximadamente igual. Si esto es así, las n_2 personas entrevistadas durante la segunda visita son una submuestra al azar de las $(n_0 - n_1)$ personas omitidas en la primera visita. Por lo tanto, $[n_1\bar{y}_1 + (n_o - n_1)\bar{y}_2]/n_o$ es una estimación insesgada de la media de la muestra original propuesta. Este método dio buenos resultados en ciertas encuestas en Inglaterra a las cuales lo aplicó Bartholomew. En encuestas repetidas Kish y Hess (1959) sugieren que las no-respuestas de encuestas recientes pueden servir para reemplazar las no-respuestas en una encuesta actual. Donde quiera que el sesgo de las primeras visitas muestre un patrón sistemático, como en la Tabla 13.1, Hendricks (1949) ha delineado métodos de extrapolación para estimar los resultados promedio que hubieran sido dados por los que no respondieron.

13.8 UN MODELO MATEMATICO PARA ERRORES DE MEDICION

Teóricamente, podemos imaginar que es posible un gran número de repeticiones independientes de la medida en la i-ésima unidad. Sea $y_{i\alpha}$ el valor obtenido en la α-ésima repetición. Entonces

$$y_{i\alpha} = \mu_i + e_{i\alpha} \tag{13.29}$$

donde μ_i = valor correcto

$e_{i\alpha}$ = error de medida.

La idea de un "valor correcto" requiere un poco de discusión. Con algunas características el concepto es simple y concreto. Por ejemplo, en un inventario hecho por muestreo, el valor correcto puede ser el número de bandas para ventilador que están en un estante a las doce del día de un día específico. En algunos casos, el valor correcto puede definirse operacionalmente. La presión sanguínea diastólica correcta de una persona en un momento específico, se

puede definir como el valor obtenido cuando se mide mediante cierto instrumento estándar bajo condiciones cuidadosamente prescritas. Sin embargo, podemos ver que nuestro instrumento estándar está sujeto a errores de medición y podemos esperar que con el tiempo se desarrollará un instrumento más preciso. Con otras características, por ejemplo, algún aspecto de la actitud de un empleado con respecto a su patrón, o de habilidad de una persona para enfrentarse a sus problemas diarios, nadie puede reclamar que tiene un método satisfactorio de medir el "valor correcto". No obstante, el concepto es útil aun en tales casos.

En mediciones repetidas de la misma unidad, los errores $e_{i\alpha}$ seguirán una distribución de frecuencia. Para la i-ésima unidad, supongamos que $e_{i\alpha}$ tenga media β_i y varianza σ_i^2. El término β_i representa un sesgo en la medición. Las magnitudes de β_i y σ_i^2 dependerán, desde luego, de la naturaleza de la característica medida y del instrumento de medición. Pueden depender también de muchos otros factores. Con poblaciones humanas, el clima político y económico prevalentes y la cantidad y tipo de publicidad que la encuesta haya recibido con anterioridad, pueden influir las respuestas al cuestionario.

El siguiente paso es considerar cómo cambian los errores de medición cuando nos movemos de una unidad a la otra. Pueden surgir varias complicaciones.

Para el componente β_i del sesgo, puede haber uno constante, digamos, $E(\beta_i) = \beta$, que afecta a todas las unidades de la población. También habrá otra componente $(\beta_i - \beta)$ que sigue una distribución de frecuencia sobre la población. Esta componente puede estar correlacionada con el valor correcto μ_i; por ejemplo, el aparato de medición puede subestimar consistentemente los valores altos de μ_i y sobrestimar los bajos.

Puede haber una correlación entre los valores de $e_{i\alpha}$ en diferentes unidades en la misma muestra. El ejemplo más sencillo es el "sesgo del entrevistador". Algunas veces se encuentran grandes diferencias entre los valores de $y_{i\alpha}$ obtenidos por diferentes entrevistadores que están muestreando partes comparables de la misma población [véase Lineau (1941), Mahalanobis (1946) y Barr (1957)].

Un efecto similar apareció cuando se cortan muestras de una cosecha por diferentes equipos y cuando se hacen análisis químicos o biológicos en laboratorios diferentes. El factor humano no es la única causa de correlaciones entre unidades que se miden más o menos al mismo tiempo. El tiempo afecta muchos procesos de medición, algunos hacen uso de materias primas cuya calidad varía de lote a lote. En la estimación del precio de venta actual de casas

TABLA 13.11. COMPONENTES DEL ERROR DE MEDICIÓN EN LA i-ÉSIMA UNIDAD

Símbolo	Naturaleza del componente
β	Sesgo constante sobre todas las unidades.
$\beta_i - \beta$	Componente variable de sesgo, que sigue alguna distribución de frecuencia con media cero conforme i varía y puede estar correlacionada con el valor correcto μ_i.
$d_{i\alpha} = e_{i\alpha} - \beta_i$	Componente fluctuante del error, que sigue alguna distribución de frecuencia con media cero y varianza σ_i^2 conforme α varía para i fija.

construidas hace algunos años, Hansen, Hurwitz y Bershad (1961) apuntan que si algunas casas en la muestra se han vendido recientemente, sus precios establecen un nivel de guía al entrevistador y al dueño en la asignación de precios a las casas que no se han vendido durante muchos años. De hecho, el precio promedio registrado para la muestra puede depender del orden en el cual aparezcan en la muestra las casas vendidas recientemente.

Para manejar estas correlaciones intramuestra en sus términos más generales, se requiere un modelo más complejo que el que se presentó aquí. En particular, las notaciones para $e_{i\alpha}$ y β_i tendrían que indicar que sus valores pueden depender de otras unidades presentes en la muestra. Sin embargo, los tipos de correlación que se cree son los más comunes, en la práctica se pueden representar por el presente modelo o por extensiones simples del mismo.

Los componentes del error de medición se resumen en la Tabla 13.11.

Hemos notado además que los valores de β_i y $d_{i\alpha}$ en diferentes unidades en la misma muestra pueden estar correlacionados uno con el otro, cuando $d_{i\alpha} = e_{i\alpha} - \beta_i$.

Hansen et al. (1951), Sukhatme y Seth (1952), Hansen, Hurwitz y Bershad (1961), han desarrollado modelos semejantes al que precede así como Hansen, Hurwitz y Pritzker (1965). Fellegi (1964) ha dado un modelo más extenso que incluye numerosas correlaciones cruzadas que puedan existir entre diferentes fuentes de errores de medición en las situaciones prácticas.

En sus artículos de 1965 y 1961 Hansen y otros., expresaron nuestro modelo en terminología un poco diferente incluida la adición de un subíndice G a todas las variables como recordatorio de que los errores y sus magnitudes en general dependen de las condiciones generales (G) de la encuesta. Ellos expresan sus resultados en términos de $d_{i\alpha}$ y de

$$\mu_i' = E(y_{i\alpha}|i) = \mu_i + \beta_i \tag{13.30}$$

La cantidad μ_i' (que denotan Y_i o P_i con una proporción) es conceptualmente el promedio obtenido de numerosas repeticiones del proceso de medición de la unidad i. Así es que

$$d_{i\alpha} = e_{i\alpha} - \beta_i = y_{i\alpha} - \mu_i' \tag{13.31}$$

Llaman a $d_{i\alpha}$, *la desviación de la respuesta* en la unidad i, en tanto que nosotros la hemos llamado la componente fluctuante del error de medición. Por lo tanto,

$$y_{i\alpha} - \mu = d_{i\alpha} + (\mu_i' - \mu') + (\mu' - \mu) \tag{13.32}$$

donde μ' es la media de *población* de los μ_i'.

Promediando sobre la muestra tenemos,

$$\bar{y}_\alpha - \mu = \bar{d}_\alpha + (\bar{\mu}' - \mu') + (\mu' - \mu) \tag{13.33}$$

Esto conduce a la fórmula

$$\text{ECM}(\bar{y}_\alpha) = V(\bar{d}_\alpha) + V(\bar{\mu}') + (\mu' - \mu)^2 + 2\,\text{cov}\,(\bar{d}_\alpha, \bar{\mu}') \tag{13.34}$$

Para la media muestral $\bar{y}_\alpha$, los términos de la derecha se llaman, respectivamente, la variancia de respuesta, la variancia de muestreo, el cuadrado del sesgo global, y el doble de la covarianza entre el sesgo de respuesta promedio en la muestra y el error de muestreo. Como $E(d_{i\alpha}|i) = 0$ en nuestro modelo, esta covarianza se anula por repeticiones del proceso de medición en el mismo conjunto de unidades de muestra en el mismo orden. Podría no anularse bajo repeticiones sobre muestras diferentes o en órdenes diferentes si los $d_{i\alpha}$ se ven afectados por otras unidades en la muestra. Aunque se va a ignorar este término por sencillez, Fellegi (1964) ha mostrado en un estudio canadiense que podría sesgar materialmente algunos métodos de estimación de las componentes de la varianza de respuesta $V(\bar{d}_\alpha)$.

Usando el método de Cornfield como se describe en la Sec. 2.9 Koch (1973) ha mostrado una descomposición general del ECM del estimador en encuestas multivariadas por muestreo con aplicaciones a las medias de subclases.

13.9 EFECTOS DEL SESGO CONSTANTE

Supóngase que las mediciones y_i en todas las unidades están sujetas a un sesgo constante β cuya magnitud se desconoce. Entonces,

la media $\bar{y}$ de una muestra aleatoria simple también está sujeta al sesgo β. En la varianza estimada del error que asignamos a la media de la muestra, el sesgo se cancela, ya que esta estimación se deriva de una suma de cuadrados de los términos $(y_i - \bar{y})^2$. Consecuentemente, el cálculo usual de los límites de confianza para $\bar{Y}$ a partir de los datos de la muestra ignora el sesgo. Los mismos resultados se aplican al muestreo aleatorio estratificado.

La situación es esencialmente la misma con las estimaciones de razón y de regresión. Considérese la estimación de regresión

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x})$$

donde tanto las y_i como las x_i pueden estar sujetas a sesgos constantes β_y y β_x, respectivamente. Dado que la estimación de mínimos cuadrados b permanece inmutable y dado que el sesgo β_x se cancela en el término $(\bar{X} - \bar{x})$, se obtiene que $\bar{y}_{lr}$ está sujeta a un sesgo β_y. Es fácil verificar que la estimación de $V(\bar{y}_{lr})$ de la muestra no contiene ninguna contribución debida al sesgo.

$$\bar{y}_R = \frac{\bar{y}}{\bar{x}}\bar{X}$$

el sesgo es también β_y para una primera aproximación, ya que en muestras grandes $E(\bar{X}/\bar{x})$ es aproximadamente 1 aun si las x_i están sujetas a un sesgo constante. En muestras grandes, la estimación de varianza a partir de la muestra

$$v(\bar{y}_R) = \frac{(N-n)}{Nn}\,\frac{\sum (y_i - \hat{R}x_i)^2}{n-1} \tag{13.35}$$

estará casi tan libre de sesgo como una estimación de

$$E(\bar{y}_R - \bar{Y})^2$$

esto es, como estimación de la varianza alrededor de la media *sesgada* $\bar{Y}$.

Para resumir, un sesgo constante pasa desapercibido por los datos de la muestra. Como hemos visto (Sec. 1.8), las probabilidades del 95% de confianza casi no se afectan si la razón de β_y al error estándar de la media estimada es menor que 0.1, pero a medida que la razón aumenta más allá de este valor, el cálculo de los límites de confianza se vuelve engañoso. Estimaciones de *cambio* entre un periodo de tiempo u otro, entre un estrato y otro, permanecen sin sesgo, siempre que éste se mantenga constante.

13.10 EFECTOS DE ERRORES NO CORRELACIONADOS DENTRO DE LA MUESTRA

En esta sección se dan algunos resultados respecto a $V(\bar{y}_\alpha)$ y $v(\bar{y}_\alpha)$ para muestreo aleatorio simple en la situación más simple en la que los errores de medición no están correlacionados dentro de la muestra. Esta situación se puede aplicar a encuestas tomadas de archivos, a encuestas sobre cuestionarios que llena el demandado (como el caso de las encuestas por correo) en las que los miembros de una misma muestra no se consultan entre sí, y a algunas encuestas de poblaciones inanimadas en donde la medición es objetiva. El supuesto de no correlación no puede hacerse a la ligera: la correlación intramuestreal puede aparecer, por ejemplo, en la entrevista, la edición, la codificación y la transferencia de los datos a la máquina computadora, si es la misma persona la que manipula un cierto número de elementos de la muestra.

A partir de (13.34) obtenemos, para la varianza,

$$V(\bar{y}_\alpha) = V(\bar{d}_\alpha) + V(\bar{\mu}') \tag{13.36}$$

Al encontrar las varianzas, primero promediamos sobre repeticiones del proceso de medición sobre una muestra específica y después sobre diferentes muestras aleatorias simples. Ahora, $V(d_{i\alpha}) = \sigma_i^2$ y los errores no están correlacionados en unidades diferentes dentro de la muestra. Por lo tanto, para el muestreo aleatorio simple, si S denota una muestra específica

$$E(\bar{d}_\alpha^2 | S) = \frac{1}{n^2}\sum_i^n \sigma_i^2 \tag{13.37}$$

Cuando promediamos subsecuentemente sobre todas las muestras aleatorias simples, tenemos

$$V(\bar{y}_\alpha) = \frac{1}{nN}\sum_i^N \sigma_i^2 + \frac{1-f}{n}\frac{\sum_i^N (\mu_i' - \mu)^2}{N-1} \tag{13.38}$$

$$= \frac{1}{n}\sigma_d^2 + \frac{(1-f)}{n} S_{\mu'}^2 \tag{13.39}$$

donde σ_d^2 denota el promedio de población de las varianzas de los errores de medición. En la terminología de Hansen y otros, σ_d^2/n es la varianza de respuesta de la estimación de muestra de $\bar{y}_\alpha$.

Con errores no correlacionados, el mismo modelo se puede aplicar al estimar una proporción de población P (Hansen, Hurwitz y Bershad, 1961). Para cualquier unidad, sea el valor correcto μ_i igual a uno si la unidad está en la clase C y cero en el caso contrario. Si

ocurren errores de medición, esto implica que a veces las unidades se clasifiquen incorrectamente. Para la unidad i-ésima, el valor registrado $y_{i\alpha}$ es en ocasiones 1 y otras 0. Sea P_i la notación para la proporción de medidas de la unidad i-ésima para la que $y_{i\alpha}$ es igual a 1. Entonces, para un i dado, $y_{i\alpha}$ es una variable binomial en medición repetida con media $\mu_i' = P_i$ mientras que la varianza de $d_{i\alpha}$ es P_iQ_i. En esta forma, si $p_\alpha = \bar{y}_\alpha$ es la estimación de muestra (13.38), viene a ser

$$V(p_\alpha) = V(\bar{y}_\alpha) = \frac{1}{nN}\sum^N P_iQ_i + \frac{(1-f)}{n}\frac{\sum^N (P_i - P)^2}{N-1} \tag{13.40}$$

$$< \frac{1}{n(N-1)}\left(\sum^N P_i - \sum^N P_i^2 + \sum^N P_i^2 - NP^2\right)$$

$$= \frac{N}{n(N-1)}PQ \tag{13.41}$$

donde

$$P = \frac{\sum^N P_i}{N}$$

Como lo muestra (13.41), la suma de la varianza de respuesta y la varianza de muestreo tiene un límite superior $NPQ/n(N-1)$. Este límite superior es también la varianza de la media de muestra de las mediciones *correctas* sobre las unidades, si la cpf se ignora y no hay sesgo global. Ya que la medición correcta μ_i en cualquier unidad es entonces una variable binomial con media P, de modo que

$$\frac{S_{\mu'}^2}{n} = \frac{N}{n(N-1)}PQ \geq V(p_\alpha) \tag{13.42}$$

Este resultado más bien inesperado es válido porque (1) la varianza de muestreo que entra en (13.39) y (13.40) es la de $\mu_i' = (\mu_i + \beta_i)$ y (2) en la estimación de una proporción μ_i y β_i son siempre *negativamente* correlacionadas. Cuando $\mu_i = 1, \beta_i \leq 0$, pues $P_i = (\mu_i + \beta_i) \leq 1$, y de manera semejante cuando $\mu_i = 0, \beta_i \geq 0$. De modo que el término

$$\frac{S_{\mu'}^2}{n}$$

en (13.39) es siempre menor

$$\frac{S_\mu^2}{n}$$

por una cantidad aproximada de la varianza de respuesta de $\bar{y}_\alpha$.

Con errores no correlacionados, un resultado útil es que la fórmula usual para $v(\bar{y}_\alpha)$ en muestreo aleatorio simple permanece sin sesgo si puede ignorarse la cpf. Del Teorema 2.4 corolario, esta fórmula desarrollada al suponer que *no* hay errores de medición es

$$v(\bar{y}_\alpha)=\frac{1-f}{n}s^2=\frac{1-f}{n}\frac{\sum^{n}(y_{i\alpha}-\bar{y}_\alpha)^2}{(n-1)} \tag{13.43}$$

Ahora bien,

$$y_{i\alpha}-\bar{y}_\alpha=(d_{i\alpha}-\bar{d}_\alpha)+(\mu_i'-\bar{\mu}') \tag{13.44}$$

Al elevar al cuadrado y promediar primero sobre mediciones repetidas y después sobre selecciones de muestra repetidas, obtenemos

$$Ev(\bar{y}_\alpha)=\frac{1-f}{n}{\sigma_d}^2+\frac{1-f}{n}{S_{\mu'}}^2 \tag{13.45}$$

en tanto que, de (13.39)

$$V(\bar{y}_\alpha)=\frac{1}{n}{\sigma_d}^2+\frac{(1-f)}{n}{S_{\mu'}}^2 \tag{13.46}$$

Por lo tanto, $Ev(\bar{y}_\alpha)=V(\bar{y}_\alpha)$ si puede ignorarse f.

De la misma manera, las fórmulas en capítulos precedentes para las estimaciones de muestra de varianzas de error de muestreo puede mostrarse que siguen siendo válidas en muestreo estratificado y de varias etapas, como lo hacen las fórmulas de grandes muestras aplicables a las estimaciones de razón y regresión, siempre y cuando los errores de medición en $y_{i\alpha}$ y $x_{i\alpha}$ no estén correlacionados dentro de la muestra y las cpf sean despreciables.

13.11 EFECTOS DE LA CORRELACION INTRAMUESTRAL ENTRE ERRORES DE MEDICION

Supóngase que para algunos o todos los valores de $d_{i\alpha}$ para unidades en la misma muestra, algunos de estos valores, o todos, están correlacionados. El término $d\alpha^2$ puede escribirse

$$\bar{d}_\alpha^{\,2}=\frac{1}{n^2}\left(\sum_i^n {d_{i\alpha}}^2+2\sum_{i<j}^{n} d_{i\alpha}d_{j\alpha}\right) \tag{13.47}$$

Por lo tanto, al promediar sobre mediciones repetidas y muestras aleatorias simples

$$V(\bar{d}_\alpha)=E(\bar{d}_\alpha^{\,2})=\frac{1}{n}{\sigma_d}^2+\frac{2n(n-1)}{2n^2}E(d_{i\alpha}\,d_{j\alpha}) \tag{13.48}$$

donde los productos se promediaron sobre todos los pares de unidades en la misma muestra. Por analogía, con el muestreo conglomerado, el coeficiente de correlación intramuestral promedio p_w puede definirse por la ecuación

$$E(d_{i\alpha}\,d_{j\alpha})=\rho_w{\sigma_d}^2 \tag{13.49}$$

Esto nos da, a partir de (13.48),

$$V(\bar{d}_\alpha)=\frac{{\sigma_d}^2}{n}[1+(n-1)\rho_w] \tag{13.50}$$

Hansen, Hurwitz y Bershad (1961) han llamado $V(\bar{d}\alpha)\alpha$ la varianza *total* de respuesta, ya que afecta la media muestral. Su componente σ_d^2/n es llamada la varianza *simple* de respuesta, en tanto que el término $(n-1)p_w\,\sigma_d^2/n$ es la *componente correlacionada* de la varianza total de respuesta.

A partir de (13.34) obtenemos, al suponer cov $(\bar{d}_\alpha, \bar{\mu}')=0$,

$$V(\bar{y}_\alpha)=\frac{{\sigma_d}^2}{n}[1+(n-1)\rho_w]+\frac{1-f}{n}{S_{\mu'}}^2 \tag{13.51}$$

El valor promedio de la $v(\bar{y}\alpha)$ usual en (13.43) se encuentra de la misma manera y es

$$Ev(\bar{y}_\alpha)=\frac{(1-f)}{n}[{\sigma_d}^2(1-\rho_w)+{S_{\mu'}}^2] \tag{13.52}$$

Como puede esperarse que ρ_w sea positivo para muchos tipos de error de medición, la fórmula estándar para $v(\bar{y}_\alpha)$ suele ser una *subestimación* en este caso y hace que la estimación de muestra parezca más precisa de lo que es. Esto es verdad aun cuando la cpf sea despreciable.

Tal vez el ejemplo más frecuente de correlación intramuestral entre errores de medición es la antes mencionada correlación dentro del entrevistador, particularmente al tratarse de preguntas que involucran opiniones o criterios. Supongamos que $n = mk$, y que cada uno de los k entrevistadores obtenga datos de m entrevistados. Si podemos suponer que no hay correlación entre los errores de medición de diferentes entrevistadores,

$$V(\bar{d}_\alpha)=\frac{{\sigma_d}^2}{n}[1+(m-1)\rho_w] \tag{13.53}$$

donde ρ_w es el coeficiente promedio de correlación dentro del entrevistador.

Aun un valor pequeño de ρ_w puede dar una contribución importante a $V(\bar{y}_\alpha)$, porque es multiplicado por el tamaño m (más o me-

nos) de la asignación del entrevistador. También puede haber alguna correlación entre los errores de mediciones para diferentes entrevistadores (como, por ejemplo, cuando han sido entrenados o dirigidos por el mismo supervisor).

Este modelo sólo representa el tipo más simple de correlación intramuestral. Con muestreo estratificado, por ejemplo, un codificador puede procesar resultados de varios estratos y por un mal entendido en las instrucciones puede introducir errores correlacionados que se extienden por sobre los estratos. El modelo matemático se puede adaptar para aplicarlo a situaciones de este tipo.

13.12 RESUMEN DE LOS EFECTOS DE ERRORES DE MEDICION

En términos del modelo, la media $\bar{y}$ de una muestra aleatoria simple sería insesgada, con varianza S_μ^2/n (ignorando cpf), si todas las mediciones fueran perfectamente exactas. Como resultado de los tipos de errores de medición aquí discutidos, la media puede estar sujeta a un sesgo de cantidad β, y su error cuadrático medio es

$$\text{ECM}(\bar{y}_\alpha) = \frac{1}{n}\{S_{\mu'}^2 + \sigma_d^2[1+(n-1)\rho_w]\} + \beta^2 \qquad (13.54)$$

donde $\mu_i' = \mu_i + \beta_i$.

La Fórmula 13.54 contiene dos términos, $S_{\mu'}^2/n$ y $\sigma_d^2\ (1-\rho_w)/n$, que disminuyen con $1/n$. Los otros dos términos, $\rho_w\,\sigma_d^2$ y β^2, a primera vista parecen independientes de n. Pero esto puede ser una sobresimplificación. Cualquier cambio material en el tamaño de la muestra puede requerir un cambio de los métodos de campo de medición y esto puede afectar a ρ_w y β^2. Sin embargo, estos dos términos deberían cambiar con relativa lentitud, si acaso, de acuerdo con n. Ya que en muestras grandes el ECM parece estar dominado por estos dos términos; la varianza ordinaria de muestreo se vuelve no importante y es engañosa como guía para la exactitud real de los resultados.

13.13 EL ESTUDIO DE LOS ERRORES DE MEDICION

En años recientes se ha investigado profusamente el estudio de los errores de medición en las prácticas de muestreo. Los objetivos son descubrir los componentes que contribuyen principalmente al ECM y encontrar las maneras de disminuir estas contribuciones. Algunos de los métodos principales se describen en ésta y en las secciones subsiguientes. En la actualidad está perfectamente claro

que el progreso será lento y costoso. Una razón es que, como ya se mencionó, los errores de medición dependen íntimamente tanto de las características como de los procedimientos de medición. Los resultados acerca de los errores de medición hallados para una encuesta rara vez pueden suponerse aplicables a otras.

Idealmente, el mejor método para estudiar los errores de medición es obtener los valores correctos de μ_i. En la práctica, esta vía de acceso está limitada a características para las cuales existe un método factible para encontrar μ_i y por problemas de ejecución y gasto. Algunos ejemplos son dados por Belloc (1954), quien comparó los datos sobre hospitalización declarados en entrevistas a domicilio, con los registros del hospital para el individuo, y por Gray (1955), quien comparó las declaraciones de los empleados respecto a sus ausencias por enfermedad con los registros de la oficina de personal. Las comprobaciones de este tipo, algunas veces llamadas "registro-verificación", son posibles con características tales como edad, ocupación, número de años de escuela y precio pagado por automóvil. Una dificultad es que algunas veces los registros no contienen un reflejo exacto de la persona entrevistada.

Al faltar un método para determinar el valor correcto, una alternativa es volver a medir, mediante un método independiente que se considera más exacto. Kish y Lansing (1954) contrataron evaluadores profesionales para estimar los precios de venta de casas que ya habían sido declarados por su propietarios. En investigaciones estadísticas sobre enfermedades, las contestaciones de los entrevistados se han comparado, sea con los registros de su médico o con los resultados de un examen médico completo [Sagen, Dunham y Simmons (1959), Trussell y Elinson (1959)]. Los resultados de tales comparaciones no son fáciles de interpretar en términos del modelo, dado que el instrumento superior está en sí mismo sujeto a errores de medición, pero las comparaciones indicarán, cuando menos, las características para las cuales el instrumento de rutina concuerda bien con el instrumento superior y aquellas con las cuales no concuerda.

En encuestas por entrevistas a viviendas, un método con un propósito semejante consiste en volver a entrevistar una submuestra de los entrevistados con entrevistadores más expertos y con un cuestionario más detallado y exhaustivo. Después de la reentrevista, el experto discutirá con el entrevistado las discrepancias entre las respuestas originales y las repetidas, con objeto de determinar la respuesta más exacta y la razón de la discrepancia. Así puede ganarse mucha información útil. Al presentar algunos métodos para la medición de errores de respuesta, Madow (1965) discutió el uso

del muestreo doble con un estimador de diferencia de la forma $[\bar{y}'-(\bar{y}-\hat{\bar{\mu}})]$ como medio para reducir el sesgo de las respuestas. Aquí $\hat{\bar{\mu}}$ es la media de las mediciones no sesgadas o menos sesgadas hechas sobre la submuestra, en tanto que $\bar{y}'$ y $\bar{y}$ son medias de las mediciones originales sobre la muestra y submuestra.

Ocasionalmente, son factibles las comparaciones generales entre los resultados de dos reconocimientos diferentes. Para un cierto número de características, los resultados del "Censo de los Estados Unidos" se pueden comparar con los dados por el "Reconocimiento de la Población Actual" que se tomaron al mismo tiempo. Dado que el "Reconocimiento" se considera más exacto, particularmente para características difíciles de medir, se pueden obtener estimaciones imprecisas del sesgo de medición β en los datos del "Censo" [Hansen, Hurwitz y Bershad (1961)]. Stephan y McCarthy (1958), describen un cierto número de comparaciones entre resultados de muestras de cuota y muestras por probabilidad.

En las secciones siguientes consideraremos algunos métodos diseñados para producir estimaciones cuantitativas de las componentes de la varianza total (varianza de respuesta más varianza de muestreo) de una estimación muestral.

13.14 MEDICIONES REPETIDAS DE SUBMUESTRAS

En estudios recientes se ha centrado el interés en la estimación: (a) de la varianza *total* de respuesta y los tamaños relativos de la varianza simple de respuesta y la componente correlacionada como contribuciones a ella, y (b) de los tamaños relativos de la varianza total de respuesta y de la varianza de muestreo. Para (a) un método usado comúnmente es seleccionar una submuestra de los agentes de medición (entrevistadores, codificadores, etc.) y volver a medir sus asignaciones por medio de otros agentes con, supuestamente, la misma habilidad. Supongamos que mk es el tamaño de la submuestra, donde m es el tamaño de la asignación de un agente y k el número de agentes elegidos en la muestra original.

Sean y_{i1}, y_{i2} las dos mediciones sobre la unidad i-ésima de la asignación de un agente. Si (13.31) es válida,

$$y_{i\alpha} = d_{i\alpha} + \mu_i' \tag{13.55}$$

Por lo tanto, al promediar sobre la asignación,

$$\frac{E\sum_i^m (y_{i1}-y_{i2})^2}{2m} = \frac{\sigma_{d1}^2+\sigma_{d2}^2}{2} - \operatorname{cov}(d_{i1}\, d_{i2}) \tag{13.56}$$

Esta expresión estima σ_{d1}^2, que es la varianza simple de respuesta para el agente 1 en la encuesta, si se cumplen dos condiciones (A): no hay correlación entre errores de respuesta d_{i1}, d_{i2}, en la misma unidad; y $\sigma_{d1}^2 = \sigma_{d2}^2$. La Ec. (13.56) se aplica a un solo par de agentes y se promedia sobre la submuestra de k pares de agentes.

Las condiciones A pueden presentarse cuando la medición es una *codificación* y los agentes son codificadores de habilidades semejantes y están entrenados por supervisores diferentes, sin que un codificador pueda ver el trabajo de otro. Al hacer la entrevista, se puede esperar una cov (d_{i1} d_{i2}) positiva porque algunos entrevistados repiten una respuesta incorrecta de memoria. En este caso, $\sum^m (y_{i1}-y_{i2})^2/2m$ subestima σ_{d1}^2. También subestima σ_{d1}^2 si el segundo agente es más hábil que el primero como lo muestran para una medición (0,1), Hansen, Hurwitz y Pritzker (1965). Aún más, se ha encontrado que $\sum (y_{i1}-y_{i2})^2/2m$ declina si el segundo entrevistador obtiene las respuestas que obtiene el primero, aun cuando se le indique que *no* las vea hasta terminar la n entrevista repetida (Koons, 1973). Estas complicaciones ilustran lo difícil del estudio realista de los errores de medición.

Para la varianza *total* de respuesta, la estimación pertinente a partir de (13.55) para un solo par de agentes es $(\bar{y}_{.1}-\bar{y}_{.2})^2/2$, donde $\bar{y}_{.1}, \bar{y}_{.2}$ son las medias de las primeras y de las segundas m mediciones. En las condiciones (A).

$$\frac{E(\bar{y}_{.1}-\bar{y}_{.2})^2}{2} = \frac{\sigma_d^2}{m}[1+(m-1)\rho_w] \tag{13.57}$$

donde ρ_w es la correlación entre errores de respuesta en unidades diferentes por el mismo agente. La Ec. (13.57) proporciona sólo un grado de libertad, pero está promediada sobre los k pares de agentes. Con una estimación de σ_d^2 a partir de (13.56), podemos estimar los tamaños relativos de la varianza de respuesta simple y la componente correlacionada. En encuestas por entrevistas, la componente correlacionada suele encontrarse que es mucho mayor que la varianza simple de respuesta, excepto para atributos básicos como son la edad, sexo y estado civil (Fellegi, 1964). En parte, por esta razón el Censo de los Estados Unidos de 1970 utilizó ampliamente la autoenumeración por correo (Hansen y Waksberg, 1970).

Si (13.55) es aplicable, también podemos estudiar la razón de la varianza simple de respuesta a la varianza de muestreo que se aplica a la muestra en una asignación de un agente. Pues, en las condiciones A,

$$E\sum_\alpha^2\sum_i^m (y_{i\alpha}-\bar{y}_{.\alpha})^2/2(m-1) = \sigma_d^2(1-\rho_w) + S_{\mu'}^2 \tag{13.58}$$

Pritzker y Hanson (1962) han llamado a la razón

$$I=\frac{\sigma_d^2}{(\sigma_d^2+S_{\mu'}^2)} \qquad (13.58)'$$

el *índice de inconsistencia,* que es análogo a la cantidad $(1-\phi)$, donde ϕ es el coeficiente de confiabilidad usado al estudiar errores de medición en psicología. Si ρ_w es despreciable, I se puede estimar a partir de la razón de (13.56) a (13.58).

Con una variable (0,1), las Ec. (13.40) y (13.41) de la Sec. (13.10), al tener $n = 1$, muestran que para una sola medición de una sola unidad, la suma de la varianza simple de respuesta y la varianza de muestreo es $V(y_{1\alpha}) \doteq PQ$, donde $P=\sum P_i/N$, al suponer que puede despreciarse f. El conjunto de valores por pares (y_{i1}, y_{i2}) obtenido por ambos agentes se resume en la siguiente tabla 2×2 de frecuencias.

Número de respuestas

		Segundo agente 1	0	Total
Primer agente	1	a	b	$a+b$
	0	c	d	$c+d$
		$a+c$	$b+d$	m

Así, b es el número de unidades sobre las que el primer agente registra 1, y el segundo agente registra 0. En las condiciones A

$$\hat{\sigma}_d^2=\frac{\sum_i^m (y_{i1}-y_{i2})^2}{2m}=(b+c)/2m \qquad (13.59)$$

Como estimación de I a partir de la submuestra, una elección es

$$\hat{I}=\frac{\hat{\sigma}_d^2}{\hat{P}\hat{Q}}=\frac{m(b+c)}{(a+b)(c+d)+(a+c)(b+d)} \qquad (13.60)$$

que promedia las estimaciones PQ a partir de los dos conjuntos de mediciones. Las estimaciones del índice han sido publicadas tanto para el Censo como para la encuesta por muestreo por Pritzker y Hanson (1962), Fellegi (1964) y Koons (1973). Estas estimaciones son útiles al comparar la poca confiabilidad relativa de la medición para diferentes atributos en censos sucesivos, o entre métodos diferentes de medición, al proporcionar así una evaluación de los nuevos métodos de medición.

La interpretación de estas comparaciones, a menudo se ve oscurecida por las dudas que existen sobre la validez de las condiciones A. Estimaciones más complejas basadas en supuestos más realistas son las que da Fellegi (1964).

13.15 SUBMUESTRAS INTERPENETRANTES

Esta técnica de particular utilidad en el estudio de errores correlacionados fue propuesta por Mahalanobis (1946). Para presentarla en los términos más simples, se divide una muestra aleatoria de n unidades, al azar, en k submuestras, donde cada submuestra contiene $m = n/k$ unidades. El trabajo de campo y el procesamiento de la muestra se planean para que no exista correlación entre los errores de medición de 2 unidades en submuestras *diferentes.* Así, por ejemplo, supongamos que la correlación de la que nos ocupamos se origina solamente en sesgos de los entrevistadores. Si cada uno de los k entrevistadores se asigna a una submuestra diferente y si no hay correlación entre errores de medición para entrevistadores diferentes, tenemos un ejemplo de la técnica.

Con el mismo modelo matemático es conveniente rotular las unidades con subíndices dobles. Sea

$$y_{ij\alpha}=\mu'_{ij}+d_{ij\alpha} \qquad (13.61)$$

donde i denota la submuestra (entrevistador) y j el miembro dentro de la submuestra. Se ignora la cpf.

Dado que la submuestra i-ésima es una submuestra aleatoria, ella misma es una muestra aleatoria simple de tamaño m. De modo que, por (13.51), la varianza de su media es

$$V(\bar{y}_{i\alpha})=\frac{1}{m}\{S_{\mu'}^2+\sigma_d^2[1+(m-1)\rho_w]\} \qquad (13.62)$$

donde ρ_w es la correlación entre los $d_{ij\alpha}$ obtenidos por el mismo entrevistador.

Como los errores son independientes en las diferentes submuestras,

$$V(\bar{y}_\alpha)=\frac{1}{k}V(\bar{y}_{i\alpha})=\frac{1}{n}\{S_{\mu'}^2+\sigma_d^2[1+(m-1)\rho_w]\} \qquad (13.63)$$

De los resultados de la muestra se puede calcular un análisis de la varianza de las km observaciones en las componentes "Entre entrevistadores (submuestras)" con $(k-1)$ grados de libertad y "Dentro de entrevistadores" con $k(m-1)$ grados de libertad. Es fácil verificar que los valores esperados de los cuadrados medios prome-

TABLA 13.12

ESPERANZAS DE LOS CUADRADOS MEDIOS (CON BASE EN UNA UNIDAD)

	gl	cm	$E(\text{cm})$
Entre entrevistadores (submuestras)	$k-1$	$s_b^2=\dfrac{m\sum(\bar{y}_{i\alpha}-\bar{y}_\alpha)^2}{k-1}$	$S_{\mu'}^2+\sigma_d^2[1+(m-1)\rho_w]$
Dentro de entrevistadores	$k(m-1)$	$s_w^2=\dfrac{\sum\sum(y_{ij\alpha}-\bar{y}_{i\alpha})^2}{k(m-1)}$	$S_{\mu'}^2+\sigma_d^2(1-\rho_w)$

diados sobre selecciones de los entrevistadores y de las muestras aleatorias y submuestras aleatorias vienen a ser como se muestra en la Tabla 13.12.

La Tabla 13.12 contiene 2 resultados importantes. Al comparar con (13.63) vemos que sb^2/n es una estimación insesgada de $V(\bar{y}_\alpha)$, si se ignora la cpf. Así es como las submuestras interpenetrantes proporcionan una estimación de $V(\bar{y}_\alpha)$ que toma en cuenta apropiadamente la varianza simple de respuesta y la componente correlacionada.

El análisis también permite estimar la componente correlacionada, ya que

$$\frac{E(s_b^2-s_w^2)}{m}=\rho_w\sigma_d^2 \tag{13.64}$$

En consecuencia, al comparar $(m-1)\ (s_b^2-s_w^2)/m$ con s_b^2, se estima la cantidad relativa con que la componente correlacionada de la varianza de respuesta contribuye a la varianza total de $\bar{y}_\alpha$. Con mediciones en donde la componente correlacionada es mucho mayor que la varianza simple de respuesta, la razón $(m-1)\ (s_b^2-s_w^2)/ms_b^2$ se ha usado en forma alternativa como medida de la contribución relativa de la varianza total de respuesta a la varianza total de $\bar{y}_\alpha$. Tepping y Boland (1972) presentan estimaciones de esta razón para atributos que aparecen en el "Reconocimiento de la Población Actual".

Cuando el método de interpenetración se aplica en una muestra de varias etapas que cubre una extensa área geográfica, la práctica más común consiste en tener pares de entrevistadores que miden las submuestras interpenetrantes extraídas de los conglomerados *más pequeños* de entre las etapas sucesivas. De esta manera, el número

de unidades últimas en la asignación de un entrevistador se conserva en su nivel acostumbrado en la encuesta, aunque el entrevistador tenga que viajar de una área dos veces mayor que la usual. Para un conglomerado, la Tabla 13.12 proporciona un grado de libertad entre entrevistadores y $2(m-1)$ g.l. dentro de entrevistadores: las medias cuadradas correspondientes se promedian sobre los conglomerados elegidos para el estudio. Desde luego, la varianza de muestreo que se mide es sólo aquella dentro de la última etapa de la conglomeración. Como recordatorio de este hecho, el término de la varianza del muestreo en $E(cm)$ a veces se escribe $S_{\mu'^2}\ (1-\rho_s)$ en lugar de $S_{\mu'^2}$, donde ρ_s es la correlación intraconglomerados entre los errores de muestreo para subunidades diferentes.

En esta forma se usó el método de interpenetración en el estudio Response Variance Study 1 del U.S. Census Bureau (1968), diseñado para estimar las componentes correlacionadas de las varianzas totales de respuesta de atributos en el censo 1960. Las áreas en este estudio eran conglomerados compactos de viviendas, los conglomerados se dispersaron sobre todos los EE.UU. En un conglomerado de muestra cualquiera se formaron dos submuestras interpenetrantes y se asignó cada submuestra a un entrevistador diferente.

En la mitad de los conglomerados, los dos entrevistadores tenían *diferentes* jefes de equipo. En una mitad se supuso, como parece razonable, que los errores de respuesta de los dos entrevistadores no estaban correlacionados. Así $(s_b^2-s_w^2)/m$ estima $\rho_w\sigma_d^2$, donde s_b^2 es ahora el cuadrado medio promedio entre entrevistadores en el mismo conglomerado. En la otra mitad de los conglomerados, los dos entrevistadores tenían el *mismo* jefe de equipo. Aquí, el objetivo era medir hasta donde "el efecto jefe de equipo" inducía una covarianza entre $\bar{d}_{.1}$ y $\bar{d}_{.2}$ para los dos entrevistadores en un conglomerado. De ser así, s_b^2 en la Tabla 13.12, ahora estima

$$S_{\mu'}^2(1-\rho_s)+\sigma_d^2[1+(m-1)\rho_w]-mE\operatorname{cov}(\bar{d}_{.1}\bar{d}_{.2})$$

y $(s_b^2-s_w^2)/m$ estima

$$\rho_w\sigma_d^2-E(\operatorname{cov}\bar{d}_{.1}\bar{d}_{.2}) \tag{13.65}$$

Al comparar los dos conjuntos de valores de $(s_b^2-s_w^2)/m$ se ve que existe un efecto "jefe de equipo". Dado que las diferencias entre estimaciones de la varianza como s_b^2 y s_w^2 son inestables, se necesita un gran número de conglomerados en las dos mitades de este tipo de estudio para medir el efecto "jefe de equipo" con alguna precisión.

La técnica de interpenetración se extiende al muestreo estratificado y al de varias etapas. Si el interés primario está en obtener una estimación insesgada de $V(\bar{y}_\alpha)$ que tome en cuenta adecuadamente

los efectos de los errores de medición, lo único que se necesita es que la muestra consista en un número de submuestras de la misma estructura, en las que tengamos la seguridad de que los errores de medición son independientes en submuestras diferentes. Estrictamente, esto requiere que los diferentes equipos de entrevista, los supervisores y los procesadores de datos sean diferentes en las diversas submuestras. Si $\bar{y}_{i\alpha}$ es la media de la i-ésima submuestra, la cantidad $\sum(\bar{y}_{i\alpha}-\bar{y}_{\alpha})^2/k(k-1)$ es una estimación insesgada de $V(\bar{y}_{\alpha})$, con $(k-1)$ grados de libertad. Este resultado es válido porque las submuestras se pueden considerar como una unidad de muestreo compleja, en la que la muestra es en efecto, una muestra aleatoria simple de dichas unidades complejas, con errores de medición no correlacionados entre las diferentes unidades complejas. En consecuencia, los resultados de la Sec. (13.10) son aplicables.

Existen numerosas aplicaciones de este método, algunas veces llamado *muestreo replicado*, que describe Deming (1960), quien lo ha usado extensamente. Para otras discusiones de sus ventajas, puede verse a Jones (1955) y Koop (1960). Los costos de viaje de los entrevistadores aumentan por la interpenetración, pero este efecto puede mitigarse si la muestra se estratifica en áreas compactas. Así, por ejemplo, cada estrato podría contener dos muestras aleatorias asignadas a un entrevistador diferente. Cada entrevistador deberá viajar por todo el estrato y no sólo la mitad de él. Cada estrato proporciona un grado de libertad para la estimación de $V(\bar{y}_{\alpha})$.

13.16 COMBINACION DE LA INTERPENETRACION Y LA MEDICION REPETIDA

Como hemos visto (Sec. 13.14), la medición repetida de las unidades en la asignación de un agente por otro de calidad semejante, proporciona estimaciones de la varianza simple de respuesta y de la varianza total de respuesta, si se aplican las condiciones A, aunque éstas pueden subestimar en encuestas por entrevistas si los errores de los entrevistados en ambas ocasiones están correlacionados positivamente. El esquema de interpenetración (Sec. 13.15) proporciona interpretaciones de la componente correlacionada de la varianza de respuesta y su contribución a la varianza total, varianza de respuesta más varianza de muestreo.

Es mucho más lo que puede aprenderse con una combinación ingeniosa de interpenetración y repetición como la usa Fellegi (1964) en un estudio de errores de respuesta en el censo de población canadiense 1961. El estudio se llevó a cabo en 134 áreas de enumeración, cada una de las cuales contenía alrededor de 150 viviendas lo que

era la asignación de un entrevistador. Las áreas de enumeración contiguas se agruparon en 67 pares. Se asignaron dos entrevistadores a cada par, para que cada uno de ellos entrevistara una mitad aleatoria de las viviendas en el par. De esta manera, cada enumerador tenía una carga de trabajo regular, pero desparramada sobre un área doble. Entonces, las asignaciones de los dos entrevistadores se intercambiaron para obtener así la combinación deseada de interpenetración y medición repetida.

Si S_1, S_2 denotan las dos submuestras interpenetrantes e I_1, I_2, los dos entrevistadores, al comparar $(I_1\,S_1)$ con $(I_2\,S_1)$ o $(I_1\,S_2)$, con $(I_2\,S_2)$ se tiene el análisis de "medición repetida", en tanto que al comparar (I_1S_1) con (I_2S_2) o (I_1S_2) con (I_2S_1), se tiene el análisis de "interpenetración". Estas comparaciones conducen a estimaciones de la varianza simple de respuesta, la componente correlacionada, la varianza total de respuestas y el índice de inconsistencia. La varianza de muestreo involucrada es la que existe entre viviendas dentro de pares de áreas de enumeración. Un análisis más extenso de Fellegi estima también la covarianza entre desviaciones de muestreo y de respuesta para el mismo entrevistador. Este término se despreció en el modelo aquí presentado, pero Fellegi muestra que en las estimaciones de $\rho\sigma_d^2$, puede crear sesgos considerables.

Bailar y Dalenius (1969) han hecho una buena exposición de las ventajas y desventajas de las diferentes variantes del método repetido y del método de interpenetración. Hansen y Waksberg (1970) revisan el trabajo de investigación del U.S. Census Bureau sobre errores de medición en lo que respecta al censo y a algunas de las más importantes encuestas por muestreo realizadas por la misma oficina. Los cambios resultantes en el censo 1970 incluían un uso más extendido de la autoenumeración (al omitir sesgos de entrevistadores) en un censo por correo, una mayor utilización del muestreo en contraposición al censo completo y una preselección por computadora de la muestra para evitar algunos sesgos que se habían detectado en la selección de la muestra de la última etapa que hizo el entrevistador. Se encontraron errores graves en datos sobre la ocupación, la industria y la calidad de las viviendas, así como problemas de revisión en datos sobre gastos, que se seguían estudiando.

13.17 PREGUNTAS DELICADAS: RESPUESTAS ALEATORIZADAS

Una situación que se puede esperar conduzca a negativas o a respuestas evasivas por parte de los entrevistados, ocurre cuando

una pregunta de la muestra es delicada o muy personal (como, por ejemplo, ¿suele el entrevistador robar objetos de las tiendas o usar drogas?). Considérese primero la estimación de una proporción binomial —la proporción π_A de entrevistados que pertenecen a cierta clase A o han cometido cierto acto—. Al utilizar ingeniosamente un artificio de aleatorización, Warner (1965) mostró que se puede estimar esta proporción sin que el entrevistado revele su posición personal respecto a la pregunta. El objeto es ayudar a que se den respuestas veraces y se conserve plenamente lo confidencial del asunto.

El artificio de aleatorización, como es la flecha giratoria o una urna con bolas blancas y rojas, selecciona uno de dos enunciados o preguntas, cada uno de los cuales requiere una respuesta "sí" o "no" por parte del entrevistado. El entrevistador *no* sabe cuál es la respuesta dada por el entrevistado, pero sí sabe cuáles son las probabilidades relativas P y $(1 - P)$ con las que se han presentado los dos enunciados. El éxito del método depende de que el entrevistado esté convencido de que su participación no revelará su posición personal con respecto al asunto que se considera delicado o muy personal.

En la proposición original de Warner los dos enunciados son:

"Soy miembro de la clase A." (Presentada con probabilidad P)
"No soy miembro de la clase A."

Con una muestra aleatoria de n entrevistados, el entrevistador registra una estimación binomial $\hat{\phi} = m/n$ de la proporción ϕ de respuestas "afirmativas". Si las preguntas se contestan verazmente, el cociente entre ϕ y π_A en la población es

$$\phi = P\pi_A + (1-P)(1-\pi_A) = (2P-1)\pi_A + (1-P) \tag{13.66}$$

Con P conocida, esta relación sugiere la estimación

$$\hat{\pi}_{AW} = \frac{[\hat{\phi} - (1-P)]}{(2P-1)} \qquad \left(P \neq \frac{1}{2}\right) \tag{13.67}$$

Esta estimación resulta ser la estimación de mayor veracidad de π_A. (El subíndice W es por "Warner".) La estimación es insesgada, con varianza

$$V(\hat{\pi}_{AW}) = \frac{\phi(1-\phi)}{n(2P-1)^2} \tag{13.68}$$

Al escribir $(1 - \phi)$ en la forma

$$(1-\phi) = (2P-1)(1-\pi_A) + (1-P) \tag{13.69}$$

fácilmente encontramos

$$V(\hat{\pi}_{AW}) = \frac{\pi_A(1-\pi_A)}{n} + \frac{P(1-P)}{n(2P-1)^2} \tag{13.70}$$

El primer término en $V(\hat{\pi}_{AW})$ es la varianza que tendría $V(\hat{\pi}_A)$ si los n entrevistados contestaron verazmente a una pregunta directa sobre su pertenencia a la clase A. Salvo cuando π_A esté cerca de $\frac{1}{2}$ y $P > 0.85$ el segundo término es mayor que el primero, a menudo mucho mayor. Por lo tanto, en general, el método es muy impreciso. Esto podría esperarse, ya que el entrevistador no sabe si una respuesta "afirmativa" implica una pertenencia a la clase A o a su opuesta. Sin embargo, como lo mostró Warner, su método puede dar un ECM menor de lo que daría una pregunta directa delicada, si esta última produjera numerosos rechazos o respuestas falsas.

13.18 LA SEGUNDA PREGUNTA NO RELACIONADA

Como alternativa al método de Warner, Simmons sugirió (Horvitz, Shah y Simmons, 1967) que la cooperación de los entrevistados podría mejorar si el segundo enunciado no fuese delicado y no tuviese relación con el primero. Por ejemplo,

"Yo nací en el mes de mayo".

El primer enunciado no se ha alterado. Si todos responden verazmente, la proporción de población de respuestas "afirmativas" es ahora

$$\phi = P\pi_A + (1-P)\pi_U \tag{13.71}$$

donde π_U es la proporción en la población muestreada de los que nacieron en mayo. Si se conoce π_U la estimación obvia y de mayor verosimilitud de π_A es

$$\hat{\pi}_{AU} = \frac{[\hat{\phi} - (1-P)\pi_U]}{P} \tag{13.72}$$

con varianza

$$V(\hat{\pi}_{AU}) = \frac{\phi(1-\phi)}{nP^2} \tag{13.73}$$

Morton (Greenberg y otros., 1969, Pág. 532) ha sugerido cómo puede obtenerse siempre el caso π_U *conocida*. Una urna contiene bolas azules, blancas y rojas en proporciones conocidas P_1, P_2, P_3. Al sacar una bola roja se produce un enunciado delicado. Al sacar una bola

azul o blanca se produce el enunciado: "El color de esta bola es blanco". Así $\pi_U = P_2/(P_2 + P_3)$.

Dowling y Shachtman (1975) han mostrado que $V(\hat{\pi}_{AU}) < V(\hat{\pi}_{AW})$ para todas las π_A, π_U siempre que P exceda ⅓. (La varianza de $\hat{\pi}_{AW}$ es simétrica alrededor de $P = ½$, pero la de $\hat{\pi}_{AU}$ no lo es, y una P pequeña da lugar a pocas respuestas a la pregunta delicada con este método.)

Si fuese necesario estimar π_A y π_U podemos tener dos muestras aleatorias de tamaños n_1, n_2, con proporciones diferentes P_1, P_2 a la pregunta delicada. Al denotar ϕ_1, ϕ_2 las proporciones de respuestas "sí" en las poblaciones definidas por las selecciones P_1 y P_2,

$$\phi_1 = P_1\pi_A + (1-P_1)\pi_U \tag{13.74}$$

$$\phi_2 = P_2\pi_A + (1-P_2)\pi_U \tag{13.75}$$

Estas relaciones sugieren la estimación

$$\hat{\pi}_{AU} = \frac{[\hat{\phi}_1(1-P_2) - \hat{\phi}(1-P_1)]}{(P_1-P_2)} \tag{13.76}$$

con

$$V(\hat{\pi}_{AU}) = \frac{1}{(P_1-P_2)^2}\left[\frac{\phi_1(1-\phi_1)(1-P_2)^2}{n_1} + \frac{\phi_2(1-\phi_2)(1-P_1)^2}{n_2}\right] \tag{13.77}$$

Si $P_1 > ½$. Greenberg y otros, (1969) mostraron que esta varianza se minimiza cuando $P_2 = 0$, o sea, cuando a todos los miembros de la segunda muestra (n_2) ha hecho la pregunta π_U no relacionada. Moors (1971) recomendó este procedimiento, pero Greenberg y otros (1969) sugieren $P_1 + P_2 = 1$ como regla de trabajo, cuando la elección $P_2 = 0$ pudiera debilitar la cooperación de los entrevistados. Cuando, por ejemplo, $P_1 = 0.8$ en promedio el 80% de la muestra 1 y el 20% de la muestra 2 se verían confrontados con la pregunta delicada.

Con las óptimas n_1, n_2 para dada $n = n_1 + n_2$, la desigualdad de Cauchy-Schwarz muestra que la varianza mínima resultante de $\hat{\pi}_{AU}$ es

$$V_{min}(\hat{\pi}_{AU}) = \frac{1}{n(P_1-P_2)^2}[(1-P_2)\sqrt{\phi_1(1-\phi_1)} + (1-P_1)\sqrt{\phi_2(1-\phi_2)}]^2 \tag{13.78}$$

La razón n_1/n_2 minimizante es

$$\frac{n_1}{n_2} = \frac{(1-P_2)}{(1-P_1)}\sqrt{\frac{\phi_1(1-\phi_1)}{\phi_2(1-\phi_2)}} \tag{13.79}$$

Esta elección requiere estimaciones anticipadas de π_A y π_U, pero el óptimo es bastante plano. Greenberg y otros (1969) hacen algunas recomendaciones respecto a las elecciones de P_1, P_2, n_1 y n_2. Para

preservar la confidencialidad, es útil tener π_U aproximadamente igual a π_A.

Se han estudiado muchas variantes de estos métodos (como sería el uso de preguntas no relacionadas) así como los sesgos producidos en $\hat{\pi}_{AW}$ y $\hat{\pi}_{AU}$ si una fracción de los entrevistados contesta falsamente a la pregunta. Greenberg y otros (1971) han aplicado la técnica de dos muestras para estimar la media μ_A a una variable sensible discreta o continua, por métodos análogos a los que conducen a las Ecs. (13.74) y (13.75). La pregunta no relacionada estima la media μ_U para una variable no sensible. Los subgrupos aleatorios de n_i sujetos ($i = 1, 2$) reciben la pregunta delicada con probabilidad P_i, y la no delicada con probabilidad $(1 - P_i)$. La variable registrada z_i para un sujeto, sigue por tanto una mezcla de dos distribuciones en las proporciones P_i, $(1 - P_i)$, una distribución con media μ_A, varianza σ_A^2, y la otra con media μ_U, varianza σ_U^2. Por lo tanto,

$$E(z_i) = P_i\mu_A + (1-P_i)\mu_U \tag{13.80}$$

$$V(z_i) = P_i\sigma_A^2 + (1-P_i)\sigma_U^2 + P_i(1-P_i)(\mu_A - \mu_U)^2 \tag{13.81}$$

Análogamente a (13.76), la estimación de μ_A es

$$\hat{\mu}_{AU} = \frac{[(1-P_2)\bar{z}_1 - (1-P_1)\bar{z}_2]}{(P_1-P_2)} \tag{13.82}$$

Para una eficiencia máxima se requieren las condiciones $\mu_U = \mu_A$, $\sigma_U^2 = 0$, en tanto que para preservar el anonimato lo mejor es $\mu_U = \mu_A$, $\sigma_U^2 = \sigma_A^2$.

Warner (1971) ha construido un contexto teórico para una gran clase de modelos de respuestas aleatorizados. Como lo sugieren (13.74) y (13.75) el truco consiste en estimar ciertas funciones lineales de las π o μ delicadas y no relacionadas con tantas ecuaciones como parámetros a estimar.

El método se ha aplicado a la obtención de estimaciones de las proporciones de nacimientos ilegítimos, de abortos inducidos, de adictos a la heroína, de personas conectadas con el crimen organizado, y del ingreso medio y número de abortos, como aplicaciones continuas-discretas. Como el método ha llamado mucho la atención, se puede esperar que aparezcan más aplicaciones. Horvitz, Greenberg y Abernathy (1975) han hecho una excelente revisión.

Ultimamente se ha discutido el grado de privacía que los entrevistados tienen en las diferentes versiones de las entrevistas aleatorizadas. En algunas versiones el entrevistador puede conjeturar la posición de algunos entrevistados sobre la pregunta delicada con una probabilidad bastante alta de corrección, —lo que es un aspecto indeseable para este método.

13.19 RESUMEN

Con relación a sus efectos en las fórmulas dadas en capítulos anteriores, los errores que no son de muestreo se pueden clasificar como sigue:

1. Con cubrimiento incompleto y no-respuesta, la consecuencia más importante es que las estimaciones pueden contener sesgo, porque la parte no alcanzada de la población puede diferir de la parte muestreada. Hay ahora amplia evidencia de que estos sesgos varían considerablemente de característica a característica y de encuesta a encuesta, siendo algunas veces despreciables y otras considerables. Una segunda consecuencia es, desde luego, que las varianzas de las estimaciones aumenta porque la muestra obtenida realmente es más pequeña que la muestra propuesta. De este factor puede hacerse cargo, al menos en parte, la selección del tamaño de la muestra propuesta.

2. Los errores de medición que son independientes de unidad a unidad dentro de la muestra y promedian cero sobre toda la población, se toman en cuenta adecuadamente en las fórmulas usuales para el cálculo de los errores estándar de las estimaciones, siempre que los términos cpf sean despreciables. Tales errores disminuyen la precisión de las estimaciones y es conveniente encontrar si esta disminución es significativa.

3. Si los errores de medición en diferentes unidades de la muestra están correlacionados, las fórmulas usuales para los errores estándar contienen sesgo. Es probable que los errores estándar sean muy pequeños, dado que las correlaciones en su mayoría son positivas en la práctica. Este tipo de perturbación se pasa por alto fácilmente e incluso puede pasar inadvertida.

4. Un sesgo constante que afecta a todas las unidades por igual es el más difícil de advertir. Ninguna manipulación de los datos de la muestra revelará dicho sesgo.

Como se ha indicado en este capítulo, el estudio de estos problemas es lento y difícil. No obstante, se ha comenzado bien. Se demostró mucho ingenio para procurar técnicas para la evaluación y control de los errores que no son de muestreo. Aunque este es un campo en el cual es difícil hacer amplias generalizaciones, gradualmente debería acumularse la información acerca de la naturaleza y magnitudes de los errores de medición en los diferentes tipos de investigaciones. También es necesario saber más sobre lo que se puede lograr con un buen adiestramiento y la supervisión de los entrevistadores, con pruebas preliminares, con mecanismos para controlar la calidad del trabajo de campo y con una evaluación postencuesta de los éxitos y puntos débiles en la operación.

En ciertas hipótesis, una segunda medición de una submuestra de unidades por otro entrevistador de habilidad semejante proporcionará estimaciones de la varianza simple de respuesta y su razón a la variancia total de respuesta, así como una estimación aproximada de su razón y la varianza de muestreo o la que se sujeta esta submuestra. Un segundo método de —submuestras interpenetrantes— de estimaciones de la varianza total (varianza de muestreo más varianza de respuesta) y una estimación de la componente correlacionada de la varianza de respuesta. La combinación de interpenetración y medición repetida es particularmente provechosa.

EJERCICIOS

13.1. Supóngase que, mediante métodos de campo de diferentes intensidades, es posible hacer que el estrato "respuesta" consista de 60, 80, 90 o 95% de la población total. Para un porcentaje a estimar, las medias verdaderas del estrato "respuesta" son: estrato 60%, 40.7; estrato 80%, 43.5; estrato 90%, 44.8; estrato 95%, 45.4; 5% restante, 59.0. (*a*) Para un método que muestre solamente el estrato 60%, demuestre que la raíz cuadrada del error cuadrático medio del porcentaje estimado para la población total es

$$\sqrt{(2414/n)+28.94}$$

donde n es el número de cuestionarios completos obtenidos. (*b*) Demuestre que la raíz cuadrada del error cuadrático medio de 5% no se puede lograr con un método con 60% de respuesta, pero sí con algo más de 100 cuestionarios completos para métodos que tienen una respuesta de 80% o mejor. (*c*) Si se prescribe una raíz cuadrada del error cuadrático medio de 2%, ¿cuáles métodos pueden lograrlo y cuáles son los tamaños de muestra necesarios?

13.2. En 13.1 (*c*) supóngase que cuesta \$5 por cuestionario completo para el método de campo que tiene un 90% de respuesta. El obtener un cuestionario completo del *siguiente* 5% de la población cuesta \$20. Para una raíz cuadrada del error cuadrático medio de 2%, ¿es más barato usar el método con una proporción de respuesta de 90% o el de una proporción de respuesta de 95%?

13.3. Una población consiste en dos estratos de igual tamaño. La probabilidad de encontrar el entrevistado en casa y anuente a ser entrevistado durante cualquier visita es 0.9 para las personas en el estrato 1 y 0.4 para las personas en el estrato 2. (*a*) En la notación de la Sec. 13.5 demuestre que

$$w_{i1} = 1-(0.1)^i, \qquad w_{i2} = 1-(0.6)^i$$

(*b*) Si el tamaño original de la muestra es n_0, calcule el número total esperado de entrevistas obtenidas para 1, 2, 3, 4 y 5 visitas. (*c*) Si los costos relativos por entrevista a la i-ésima visita son 100, 120, 150, 200 y 300 para $i = 1, 2, 3, 4, 5$, respectivamente, calcule el costo promedio por entrevista para todas

las entrevistas obtenidas hasta la i-ésima visita. (d) El dinero disponible para la encuesta es suficiente para pagar las primeras 300 visitas completas. Si la política es insistir en i visitas, ¿cuáles son los números esperados totales de entrevistas completas que se pueden obtener por la misma cantidad de dinero cuando $i = 1, 2, 3, 4, 5$?

13.4. En el Ejercicio 13.3 las personas en el estrato 1 tienen una media de 40% para algún porcentaje binomial que se está estimando, y las personas en el estrato 2 tienen una media de 60%. (a) Calcule el sesgo en la media de la muestra para $i = 1, 2, 3, 4, 5$ visitas. (b) Calcule la varianza de las medias de la muestra para la situación de costo en la parte (d) del Ejercicio 13.3. (Para economizar cálculos, la varianza puede tomarse como $2\,600/n_i$, donde n_i es el número total esperado de entrevistas obtenidas.) (c) ¿Cuál política da el ECM más bajo?

13.5. En la Sec. 13.6 (submuestreo de quienes no responden) verificar la fórmula (Pág. 451) para la razón del costo esperado de obtener una V especificada sin submuestreo al costo mínimo esperado,

$$\text{Razón} = \frac{F(c_0+c_1W_1+c_2W_2)}{[\sqrt{(F-W_2)(c_0+c_1W_1)}+W_2\sqrt{c_2}]^2}$$

donde $F = S^2/S_2^2$. Sea $c_0 = c_1 = 1$, $c_2 = 16$. (a) Si $F = 1$, demuestre que la razón de costo tiene un máximo de 1.25 cuando $W_2 = 0.2$ o 0.25. (b) Si $F = 1.5$, demuestre que el máximo es 1.41 para $W_2 = 0.3$ o 0.35.

13.6. En una encuesta de aves de corral y cerdos mantenidos en patios y ciertas instalaciones pequeñas (Gray, 1957), una encuesta por correo con varios recordatorios, se continuó con entrevistas de una submuestra de las personas que no contestaron. Por un juicio adelantado, se escogió $k = 2$ (es decir, una submuestra del 50%). Para una característica importante se dispuso de los siguientes datos después de la encuesta, en la notación del Ejercicio 13.5.

$$\frac{c_1}{c_0} \doteq 0.15, \qquad \frac{c_2}{c_0} \doteq 9.5, \qquad W_1 \doteq 0.8, \qquad S^2 \doteq S_2^2$$

Al encontrar VC para $k = 2$ y para una k óptima, determine si $k = 2$ fue una buena elección.

13.7. En una encuesta por el método Politz-Simmons se encontraron en casa durante la primera visita 390 personas que respondieron, de una muestra inicial de 660. Los números de los que declararon haber estado en casa durante las 0, 1, ..., 5 de las cinco noches precedentes y el número de los que contestaron sí a una pregunta en la encuesta fueron como sigue:

	0/5	1/5	2/5	3/5	4/5	5/5
Número	14	35	55	74	94	118
Respuestas sí	4	13	20	30	42	156

Calcule la estimación Politz-Simmons de la proporción de contestaciones afirmativas en la población y compárelo con la estimación binomial simple.

13.8. Una población con $N = 6$ contiene tres unidades para las cuales la respuesta correcta a una pregunta es sí, y tres unidades para las cuales es no. Debido a errores de medición, la probabilidad de obtener una respuesta "sí" en una unidad sí, es 0.9; y la probabilidad de obtener una respuesta "no" en una unidad no, es también 0.9. (a) Al obtener la distribución de todas las respuestas posibles para muestras de tamaño 2, demuestre que las probabilidades de que la muestra de 0, 1, 2 respuestas "sí" son 0.218, 0.564, y 0.218. (b) Demuestre que la varianza de la proporción estimada de contestaciones "sí" es 0.1090. Verifique los resultados (13.40) y (13.41) en la Sec. 13.10. (c) ¿Cuál sería la varianza de la proporción estimada de "sí", si no hubiera errores de medición?

13.9. En una parte de la Encuesta de Trabajo en Bengala 1942 (Mahalanobis, 1946) se tomó una muestra al azar de cerca de 175 familias en cada uno de tres estratos. La muestra en cada estrato se dividió en cinco submuestras al azar, asignando cada una a un entrevistador diferente. Los cinco entrevistadores trabajaron en todos los cinco estratos. Para gastos de alimentos, la parte relevante del análisis de varianza (con base en una *sola familia*) es como sigue:

	gl	cm	E(cm)
Entre entrevistadores	4	22.3	$\sigma_{\mu'}^2+\sigma_d^2+35\sigma_{IS}^2+105\sigma_I^2$
Entrevistas × estratos	8	9.6	$\sigma_{\mu'}^2+\sigma_d^2+35\sigma_{IS}^2$
Dentro de submuestras	510	9.9	$\sigma_{\mu'}^2+\sigma_d^2$

Si g_i, w_{hi}, representan sesgos del entrevistador i, el modelo para una sola familia es

$$y_{hij\alpha} = \bar{\mu}_h + g_i + w_{hi} + (\mu'_{hij} - \bar{\mu}_h) + d_{hij\alpha}$$

Varianzas: $\qquad \sigma_I^2 \quad \sigma_{IS}^2 \quad \sigma_{\mu'}^2 \qquad \sigma_d^2$

Verifique las expresiones dadas para $E(cm)$ y estime la proporción de la varianza total de la media que puede ser adscrita a sesgos del enumerador.

13.10. Considérese un acto ilegal que haya sido cometido por el 10% de la población ($\pi_A = 0.1$). Si todos los entrevistados son veraces, compárese la $V(\hat{\pi}_A)$ para $n = 500$ dada por (a) una pregunta delicada directa, (b) el método de Warner con $P = 0.8$, (c) el método de la pregunta no relacionada con $\pi_u = 0.2$ conocida, (d) el método de dos muestras y pregunta no relacionada con π_u de hecho igual a 0.2, pero desconocida cuando $P_1 = 0.8$, $P_2 = 0$ como lo recomendó Moors, (e) el mismo método cuando $P_1 = 0.8$, $P_2 = 1 - P_1$. Supóngase que se puede usar la n_1/n_2 óptima en (d) y (e). Se han evitado algunos decimales al calcular para los métodos $V(100\hat{\pi}_A) = 10^4 V(\hat{\pi}_A)$.

13.11. En el Ejercicio 13.10, supóngase que todos los entrevistados fueron veraces con cualquiera de los métodos de respuesta aleatorizada (b), (c) (d) o (e), pero que a una pregunta directa y delicada con el método (a) algunos entrevistados que cometieron el acto lo niegan. Para $n = 500$, ¿cuál de los métodos de respuesta aleatorizada da un ECM ($\hat{\pi}_A$) menor que el método (a) si (1) 15%, (2) 20%, (3) 25% de los que cometieron el acto lo niegan con el método (a)?

Bibliografía

Armitage, P. (1947). A comparison of stratified with unrestricted random sampling from a finite population. *Biometrika*, **34**, 273-280.

Arvesen, J. N. (1969). Jackknifing U-Statistics. *Ann. Math. Stat.*, **40**, 2076-2100.

Avadhani, M. S. y Sukhatme, B. V. (1973). Controlled sampling with equal probabilities and without replacement. *Int. Stat. Rev.*, **41**, 175-183.

Bailar, B. A. (1975). The effects of rotation group bias on estimates from panel surveys. *Jour. Amer. Stat. Assoc.*, **70**, 23-30.

Bailar, B. A., y Dalenius, T. (1969). Estimating response variance components of the U.S. Bureau of the Census Survey Model. *Sankhya*, **B31**, 341-360.

Barr, A. (1957). Differences between experienced interviewers. *App. Stat.*, **6**, 180-188.

Bartholomew, D. J. (1961). A method of allowing for "not-at-home" bias in sample surveys, *App. Stat.*, **10**, 52-59.

Bartlett, M. S. (1949). Fitting a straight line when both variables are subject to error. *Biometrics*, **5**, 207-212.

Basu, D. (1958). On sampling with and without replacement. *Sankhya*, **20** 287-294.

Bayless, D. L. (1968). Variance estimation in sampling from finite populations. Ph. D. Thesis, Texas A & M University.

Bayless, D. L., y Rao, J. N. K. (1970). An empirical study of stabilities of estimators and variance estimators in unequal probability sampling (n = 3 o 4). *Jour. Amer. Stat. Assoc.*, **65**, 1645-1667.

Beale, E. M. L. (1962). Some uses of computers in operational research. *Industrielle Organisation*, **31**, 51-2.

Bean, J. A. (1970). Estimation and sampling variance in the health interview survey. National Center for Health Statistics, Washington, D. C., Series 2, 38.

Bean, J. A. (1975). Distribution and properties of variance estimators for complex multistage probability samples. National Center for Health Statistics, Washington, D. C., Series 2, 65.

Beardwood, J., Halton, J. H., y Hammersley, J. M. (1959). The shortest path through many points, *Proc. Cambridge Phil. Soc.*, **55**, 299-327.

Bellhouse, D. R. y Rao, J. N. K. (1975). Systematic sampling in the presence of a trend. *Biometrika*, **62**, 694-697.

Belloc, N. B. (1954). Validation of morbidity survey data by comparison with hospital records. *Jour. Amer. Stat. Assoc.*, **49**, 832-846.

Birnbaum, Z. W., y Sirken, M. G. (1950a). Bias due to nonavailability in sampling surveys. *Jour. Amer. Stat. Assoc.*, **45**, 98-111.

Birnbaum, Z. W., y Sirken, M. G. (1950b). On the total error due to noninterview and to random sampling. *Int. Jour. Opinion and Attitude Res.*, **4**, 179-191.

Blythe, R. H. (1945). The economics of sample size applied to the scaling of saw-logs. *Biom. Bull.*, **1**, 67-70.

Booth, G., y Sedransk, J. (1969). Planning some two-factor comparative surveys. *Jour. Amer. Stat. Assoc.*, **64**, 560-573.

Bose, Chameli (1943). Note on the sampling error in the method of double sampling. *Sankhya*, **6**, 330.

Brewer, K. W. R. (1963a). A model of systematic sampling with unequal probabilities. *Australian Jour. Stat.*, **5**, 5-13.

Brewer, K. W. R. (1963b). Radio estimation in finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Jour. Stat.*, **5**, 93-105.

Brewer, K. W. R., y Hanif, M. (1969). Sampling without replacement and probability of inclusion proportional to size. I Methods using the Horvitz and Thompson estimator. II Methods using special estimators. Inédito.

Brewer, K. W. R., y Hanif, M. (1970). Durbin's new multistage variance estimator. *Jour. Roy. Stat. Soc.*, **B32**, 302-311.

Brooks, S. (1955). The estimation of an optimum subsampling number. *Jour. Amer. Stat. Assoc.*, **50**, 398-415.

Bryant, E. C., Hartley, H. O., y Jessen, R. J. (1960). Design and estimation in two-way stratification. *Jour. Amer. Stat. Assoc.*, **55**, 105-124.

Buckland, W. R. (1951). A review of the literature of systematic sampling. *Jour. Roy. Stat. Soc.*, **B13**, 208-215.

Burstein, H. (1975). Finite population correction for binomial confidence limits. *Jour. Amer. Stat. Assoc.*, **70**, 67-69.

Cameron, J. M. (1951). Use of variance components in preparing schedules for the sampling of baled wool. *Biomerics*, **7**, 83-96.

Chatterjee, S. (1966). A programming algorithm and its statistical applications. O.N.R. Rep. Téc. 1, Department of Statistics, Harvard University, Cambridge.

Chatterjee, S. (1967). A note on optimum stratification. *Skand. Akt.*, **50**, 40-44.

Chatterjee, S. (1968). Multivariate stratified surveys. *Jour. Amer. Stat. Assoc.*, **63**, 530-534.

Chatterjee, S. (1972). A study of optimum allocation in multivariate stratified surveys. *Skand. Akt.*, **55**, 73-80.

Chung, J. H., y DeLury, D. B. (1950). *Confidence Limits for the Hypergeometric Distribution.* University of Toronto Press, Toronto, Canadá.

Cochran, W. G. (1942). Sampling theory when the sampling units are of unequal sizes. *Jour. Amer. Stat. Assoc.*, **37**, 199-212.

Cochran, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann. Math. Stat.*, **17**, 164-177.

Cochran, W. G. (1961). Comparison of methods for determining stratum boundaries. *Bull. Int. Stat. Inst.*, **38**, 2, 345-358.

Cochran, W. G., Mosteller, F., y Tukey, J. W. (1954). *Statistical Problems of the Kinsey Report. American Statistical Association, Washington, D.C.*, Pág. 280.

Coleman, J. S. (1966). *Equality of Educational Opportunity*, Oficina de Prensa de los E.U.A., Washington, D.C.

Cornell, F. G. (1947). A stratified random sample of a small finite population. *Jour. Amer. Stat. Assoc.*, **42**, 523-532.

Cornfield, J. (1944). On samples from finite populations. *Jour. Amer. Stat. Assoc.*, **39**, 236-239.

Cornfield, J. (1951). The determination of sample size. *Amer. Jour. Pub. Health*, **41**, 654-661.

Cox, D. R. (1952). Estimation by double sampling. *Biometrika*, **39**, 217-227.

Dalenius, T. (1957). *Sampling in Sweden.* Contributions to the methods and theories of sample survey practice. Almqvist and Wicksell, Stockholm.

Dalenius, T., y Gurney, M. (1951). The problem of optimum stratification. II. *Skand. Akt.*, **34**, 133-148.

Dalenius, T., y Hodges, J. L., Jr. (1959). Minimum variance stratification. *Jour. Amer. Stat. Assoc.*, **54**, 88-101.

Das, A. C. (1950). Two-dimensional systematic sampling and the associated stratified and random sampling. *Sankhya*, **10**, 95-108.

David, F. N., y Neyman, J. (1938). Extension of the Markoff theorem of least squares. *Stat. Res. Mem.*, **2**, 105.

David, I. P., y Sukhatme, B. V. (1974). On the bias and mean square error of the ratio estimator. *Jour. Amer. Stat. Assoc.*, **69**, 464-466.

Deming, W. E. (1953). On a probability mechanism to attain an economic balance between the resultant error of non-response and the bias of non-response. *Jour. Amer. Stat. Assoc.*, **48**, 743-772.

Deming, W. E. (1956). On simplifications of sampling design through replication with equal probabilities and without stages. *Jour. Amer. Stat. Assoc.*, **51**, 24-53.

Deming, W. E. (1960). *Sample Design in Business Research.* John Wiley and Sons, New York.

Deming, W. E., y Simmons, W. R. (1946). On the design of a sample for dealer inventories. *Jour. Amer. Stat. Assoc.*, **41**, 16-33.

Des Raj (1954). On sampling with probabilities proportional to size. *Ganita*, **5**, 175-182.

Des Raj (1956a). Some estimators in sampling with varying probabilities without replacement. *Jour. Amer. Stat. Assoc.*, **51**, 269-284.

Des Raj (1956b). A note on the determination of optimum probabilities in sampling without replacement. *Sankhya*, **17**, 197-200.

Des Raj (1958). On the relative accuracy of some sampling techniques. *Jour. Amer. Stat. Assoc.*, **53**, 98-101.

Des Raj (1964). The use of systematic sampling with probability proportional to size in a large-scale survey. *Jour. Amer. Stat. Assoc.*, **59**, 251-255.

Des Raj (1966). Some remarks on a simple procedure of sampling without replacement. *Jour. Amer. Stat. Assoc.*, **61**, 391-396.

Des Raj, y Khamis, S. H. (1958). Some remarks on sampling with replacement. *Ann. Math. Stat.*, **29**, 550-557.

Dowling, T. A., y Shachtman, R. H. (1975). On the relative efficiency of randomized response models. *Jour. Amer. Stat. Assoc.*, **70**, 84-87.

Durbin, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. *Jour. Roy. Stat. Soc.*, **B15**, 262-269.

Durbin, J. (1954). Non-response and call-backs in surveys. *Bull. Int. Stat. Inst.*, **34**, 72-86.

Durbin, J. (1958). Sampling theory for estimates based on fewer individuals than the number selected. *Bull. Int. Stat. Inst.*, **36**, 3, 113-119.

Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, **46**, 477-480.

Durbin, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *App. Stat.*, **16**, 152-164.

Durbin, J., y Stuart, A. (1954). Callbacks and clustering in sample surveys: an experimental study. *Jour. Roy. Stat. Soc.*, **A117**, 387-428.

Eckler, A. R. (1955). Rotation sampling. *Ann. Math. Stat.*, **26**. 664-685.

Ekman, G. (1959). An approximation useful in univariate stratification. *Ann. Math. Stat.*, **30**, 219-229.

Erdös, P., y Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Pub. Math. Inst. Hungarian Acad. Sci.*, **4**, 49-57.

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations. *Jour. Roy. Stat. Soc.*, **B31**, 195-233.

Evans, W. D. (1951). On stratification and optimum allocations. *Jour. Amer. Stat. Assoc.*, **46**, 95-104.

Fellegi, I. (1963). Sampling with varying probabilities without replacement: rotating and non-rotating samples. *Jour. Amer. Stat. Assoc.*, **58**, 183–201.

Fellegi, I. (1964). Response variance and its estimation. *Jour. Amer. Stat. Asosc.*, **59**, 1016-1041.

Feller, W. (1957). *An Introduction to Probability Theory and Its Applications*, John Wiley and Sons, New York, segunda edición.

Fieller, E. C. (1932). The distribution of the index in a normal bivariate population. *Biometrika*, **24**, 428-440.

Finkner, A. L. (1950). Methods of sampling for estimating commercial peach production in North Carolina. *North Carolina Agr. Exp. Stat. Tech. Bull.*, **91**.

Finkner, A. L., Morgan, J. J., y Monroe, R. J. (1943). Methods of estimating farm employment from sample data in North Carolina. *N. C. Agr. Exp. Sta. Tech. Bull.*, **75**.

Finney, D. J. (1948). Random and systematic sampling in timber surveys. *Forestry*, **22**, 1–36.

Finney, D. J. (1949). On a method of estimating frequencies. *Biometrika*, **36**, 233–234.

Finney, D. J. (1950). An example of periodic variation in forest sampling. *Forestry*, **23**, 96–111.

Fisher, R. A. (1958). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, decimotercera edición, sección 21. Cuarta edición (1932).

Fisher. R. A., y Mackenzie, W. A. (1922). The correlation of weekly rainfall. *Quart. Jour. Roy. Met. Soc.*, **48**, 234–245.

Fisher, R. A., y Yates, F. (1957). *Statistical Tables for Biological, Agricultural and Medical Research*. Oliver y Boyd, Edinburgh, quinta edición.

Foreman, E. K., y Brewer, K. W. R. (1971). The efficient use of supplementary information in standard sampling procedures. *Jour. Roy. Stat. Soc.* **B33**, 391–400.

Frankel, M. R. (1971). Inference from survey samples. Institute for Social Research, Ann Arbor, Mich.

Fuller, W. A. (1966). Estimation employing post strata. *Jour. Amer. Stat. Assoc.*, **61**, 1172–1183.

Fuller, W. A. (1970). Sampling with random stratum boundaries. *Jour. Roy. Stat. Soc.*, **B32**, 209–226.

Fuller, W. A., y Burmeister, L. F. (1972). Estimators for samples selected from two overlapping frames. *Proc. Soc. Stat. Sect. Amer. Stat. Assoc.*, 245–249.

Gallup, G. (1972). Opinion polling in a democracy. *Statistics, a Guide to the Unknown*. J. M. Tanur y Cols. (Eds.), Holden-Day, Inc., San Francisco, 146–152.

Gautschi, W. (1957). Some remarks on systematic sampling. *Ann. Math. Stat.*, **28**, 385–394.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *Jour. Roy. Stat. Soc.*, **B17**, 269–278.

Goodman, L. A., y Hartley, H. O. (1958). The precision of unbiased ratio-type estimators. *Jour. Amer. Stat. Assoc.*, **53**, 491–508.

Goodman, R., y Kish, L. (1950). Controlled selection–a technique in probability sampling. *Jour. Amer. Stat. Assoc.*, **45**, 350–372.

Graham, J. E. (1973). Composite estimation in two cycle rotation sampling designs. *Comm. in Stat.* **1**, 419–431.

Gray, P. G. (1955). The memory factor in social surveys. *Jour. Amer. Stat. Assoc.*, **50**, 344–363.

Gray, P. G. (1957). A sample survey with both a postal and an interview stage. *App. Stat.* **6**, 139–153.

Gray, P. G., y Corlett, T. (1950). Sampling for the social survey. *Jour. Roy. Stat. Soc.*, **A113**, 150–206.

Greenberg, B. G., y cols. (1969). The unrelated question randomized response model: Theoretical framework. *Jour. Amer. Stat. Assoc.*, **64**, 520–539.

Greenberg, B. G. y cols. (1971). Application of the randomized response technique in obtaining quantitative data. *Jour. Amer. Stat. Assoc.*, **66**, 243–250.

Grundy, P. M., Healy, M. J. R., y Rees, D. H. (1954). Decision between two alternatives–how many experiments? *Biometrics*, **10**, 317–323.

Grundy, P. M., Healy, M. J. R., y Rees. D. H. (1956). Economic choice of the amount of experimentation. *Jour. Roy. Stat. Soc.*, **B18**, 32–55.

Hagood, M. J., y Bernert, E. H. (1955). Component indexes as a basis for stratification. *Joux. Amer. Stat. Assoc.*, **40**, 330–341.

Hájek, J. (1958). Some contributions to the theory of probability sampling. *Bull. Int. Stat. Inst.*, **36**, 3, 127–134.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Pub. Math. Inst. Hungarian Acad. Sci.*, **5**, 361-374.

Haldane, J. B. S. (1945). On a method of estimating frequencies. *Biometrika*, **33**, 222–225.

Hansen, M. H., y cols. (1951). Response errors in surveys. *Jour. Amer. Stat. Assoc.*, **46**, 147–190.

Hansen, M. H., y Hurwitz, W. N. (1942). Relative efficiencies of various sampling units in population inquiries. *Jour. Amer. Stat. Assoc.*, **37**, 89–94.

Hansen, M. H., y Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *Ann. Math. Stat.*, **14**, 333–362.

Hansen, M. H., y Hurwitz, W. N. (1946). The problem of nonresponse in sample surveys. *Jour. Amer. Stat. Assoc.*, **41**, 517–529.

Hansen, M. H., y Hurwitz, W. N. (1949). On the determination of the optimum probabilities in sampling. *Ann. Math. Stat.*, **20**, 426–432.

Hansen, M. H., Hurwitz, W. N., y Bershad, M. (1961). Measurement errors in censuses and surveys. *Bull. Int. Stat. Inst.*, **38**, 2, 359–374.

Hansen, M. H., Hurwitz, W. N., y Gurney, M. (1946). Problems and methods of the sample survey of business. *Jour. Amer. Stat. Assoc.*, **41**, 173–189.

Hansen, M. H., Hurwitz, W. N., y Jabine, T. B. (1963). The use of imperfect lists for probability sampling at the U.S. Bureau of the Census. *Bull. Int. Stat. Inst.*, **40**, 1, 497–517.

Hansen, M. H., Hurwitz, W. N., y Madow. W. G. (1953). *Sample Survey Methods and Theory.* John Wiley and Sons, New York, Vols. I y II.

Hansen, M. H., Hurwitz, W. N., Nisselson, H., y Steinberg, J. (1955). The redesign of the census current population survey. *Jour. Amer. Stat. Assoc.*, **50**, 701–719.

Hansen, M. H., Hurwitz, W. N., y Pritzker, L. (1965). The estimation and interpretation of gross differences and the simple response variance. *Contributions to statistics presented to Professor P. C. Mahalanobis.* Pergamon Press, Oxford, and Statistical Publishing Society, Calcutta, 111–136.

Hansen, M. H., y Waksberg, J. (1970). Research on non-sampling errors in censuses and surveys. *Rev. Int. Stat. Inst.*, **38**, 318–332.

Hanson, R. H., y Marks, E. S. (1958). Influence of the interviewer on the accuracy of survey results. *Jour. Amer. Stat. Assoc.*, **53**, 635–655.

Hartley, H. O. (1946). Discussion of paper by F. Yates. *Jour. Roy. Stat. Soc.*, **109**, 37.

Hartley, H. O. (1959). *Analytic Studies of Survey Data.* Istituto di Statistica, Rome, volume in honor of Corrado Gini.

Hartley, H. O. (1962). Multiple frame surveys. *Proc. Soc. Stat. Sect. Amer. Stat. Assoc.*, 203–206.

Hartley, H. O. (1974). Multiple frame methodology and selected applications. *Sankhya*, **C36**, 99–118.

Hartley, H. O., y Hocking, R. (1963). Convexing programming by tangential approximation. *Management Science*, **9**, 600–612.

Hartley, H. O., y Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *Ann. Math. Stat.*, **33**, 350–374.

Hartley, H. O., y Rao, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika*, **55**, 547–557.

Hartley, H. O., y Rao, J. N. K. (1969). A new estimation theory for sample surveys, II. In *New Developments in Survey Sampling*, N. L. Johnson and H. Smith (Eds.), Wiley-Interscience, New York, 147-169.

Hartley, H. O., Rao, J. N. K., y Kiefer, G. (1969). Variance estimation with one unit per stratum. *Jour. Amer. Stat. Assoc.*, **64**, 841–851.

Hartley, H. O., y Ross, A. (1954). Unbiased ratio estimates. *Nature*, **174**, 270–271.

Harvard Computation Laboratory (1955). *Tables of the Cumulative Binomial Probability Distribution.* Harvard University Press, Cambridge, Mass.

Haynes, J. D. (1948). An empirical investigation of sampling methods for an area. M. S. thesis, University of North Carolina.

Hendricks, W. A. (1944). The relative efficiencies of groups of farms as sampling units. *Jour. Amer. Stat. Assoc.*, **39**, 367–376.

Hendricks, W. A. (1949). Adjustment for bias by non-response in mailed surveys. *Agr. Econ. Res.*, **1**, 52–56.

Hendricks, W. A. (1956). *The Mathematical Theory of Sampling.* Scarecrow Press, New Brunswick, N.J.

Hess, I., Riedel, D. C., y Fitzpatrick, T. B. (1976). *Probability Sampling of Hospitals and Patients.* University of Michigan, Ann Arbor. Mich., segunda edición.

Hess, I., Sethi, V. K., y Balakrishnan, T. R. (1966). Stratification: A practical investigation. *Jour. Amer. Stat. Assoc.*, **61**, 74–90.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.*, **19**, 293–325.

Homeyer, P. G., y Black, C. A. (1946). Sampling replicated field experiments on oats for yield determinations. *Proc. Soil Sci. Soc. America*, **11**, 341–344.

Horvitz, D. G. (1952). Sampling and field procedures of the Pittsburgh morbidity survey. *Pub. Health Reports*, **67**, 1003–1012.

Horvitz, D. G., Greenberg, B. G., y Abernathy, J. R. (1975). Recent developments in randomized response designs. *A survey of statistical design and linear models.* J. N. Srivastava (Ed.), American Elsevier Publishing Co., New York, 271–285.

Horvitz, D. G., Shah, B.V., y Simmons, W. R. (1967). The unrelated randomized response model. *Proc. Soc. Stat. Sect. Amer. Stat. Assoc.*, 65-72.

Horvitz, D. G., y Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Jour. Amer. Stat. Assoc.*, **47**, 663–685.

Huddleston, H. F., Claypool, P. L., y Hocking, R. R. (1970). Optimum sample allocation to strata using convex programming. *App. Stat.* **19**, 273–278.

Hutchinson, M. C. (1971). A Monte Carlo comparison of some ratio estimators. *Biometrika*, **58**, 313–321.

Hyman, H. H. (1954). *Interviewing in Social Research.* University of Chicago Press, Chicago, Ill.

James, A. T., Wilkinson, G. N., y Venables, W. N. (1975). Interval estimates for a ratio of means. *Sankkya* (en prensa).

Jebe, E. H. (1952). Estimation for sub-sampling designs employing the county as a primary sampling unit. *Jour. Amer. Stat. Assoc.*, **47**, 49–70.

Jensen, A. (1926). Report on the representative method in statistics. *Bull. Int. Stat. Inst.*, **22**, 359–377.

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

Jessen, R. J. (1955). Determining the fruit count on a tree by randomized branch sampling. *Biometrics*, **11**, 99–109.

Jessen, R. J., y cols. (1947). On a population sample for Greece. *Jour. Amer. Stat. Assoc.*, **42**, 357–384.

Jessen, R. J. y Houseman, E. E. (1944). Statistical investigations of farm sample surveys taken in Iowa, Florida and California. *Iowa Agr. Exp. Sta. Res. Bull.*, 329.

Johnson, F. A. (1941). A statistical study of sampling methods for tree nursery inventories. M. S. thesis, Iowa State College.

Johnson, F. A. (1943). A statistical study of sampling methods for tree nursery inventories. *Jour. Forestry*, **41**, 674–689.

Jones, H. W. (1955). Investigating the properties of a sample mean by employing random subsample means. *Jour. Amer. Stat. Assoc.*, **51**, 54–83.

Kempthorne, O. (1969). Some remarks on inference in finite sampling. *New Developments in Survey Sampling*, N. L. Johnson and H. Smith, Jr. (Eds.), John Wiley & Sons. New York, 671–695.

Kendall, M. G., y Smith, B. B. (1938) Randomness and random sampling numbers. *Jour. Roy. Stat. Soc.*, **101**, 147–166.

Keyfitz, N. (1957). Estimates of sampling variance where two units are selected from each stratum. *Jour. Amer. Stat. Assoc.*, **52**, 503–510.

Khan, S., y Tripathi, T. P. (1967). The use of multivariate auxiliary information in double-sampling. *J. Ind. Stat. Assoc.*, **5**, 42–48.

King, A. J., y McCarty, D. E (1941). Application of sampling to agricultural statistics whit emphasis on stratified samples. *Jour. Marketing*, abril, 462–474.

Kiser, C. V., y Whelpton, P. K. (1953). Resume of the Indianapolis study of social and psychological factors affecting fertility. *Population Studies*, **7**, 95–110.

Kish, L. (1949). A procedure for objetive respondent selection within the household. *Jour. Amer. Stat. Assoc.*, **44**, 380–387.

Kish, L. (1957). Confidence limits for clustered samples. *Amer. Soc. Rev.*, **22**, 154–165.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons, New York.

Kish, L. y Frankel, M. R. (1974). Inference from complex samples. *Jour. Roy. Stat. Soc.*, **B36**, 1–37.

Kish, L., y Hess, I. (1958). On noncoverage of sample dwellings. *Jour. Amer. Stat. Assoc.*, **53**, 509–524.

Kish, L., y Hess, I. (1959a). A "replacement" procedure for reducing the bias of nonresponse. *Amer. Statistician*, **13**, 4, 17–19.

Kish, L., y Hess, I. (1959b). On variances of ratios and their differences in multistage samples. *Jour. Amer. Stat. Assoc.*, **54**, 416–446.

Kish, L., y Lansing, J. B. (1954). Response errors in estimating the value of homes. *Jour. Amer. Stat. Assoc.*, **49**, 520–538.

Kish, L., Namboodiri, N. K., y Pillai, R. K. (1962). The ratio bias in surveys. *Jour. Amer. Stat. Assoc.*, **57**, 863–876.

Koch, G. (1973). An alternative approach to multivariate response error models for sample survey data with applications to estimators involving subclass means. *Jour. Amer. Stat. Ass.*, **68**, 906–913.

Kohan, A. R. (1963). Optimum allocation in multivariate surveys. *Jour. Roy. Stat. Soc.*, **A126**, 557–565.

Koons, D. A. (1973). Quality control and measurement of nonsampling error in the Health Interview Survey. *Nat. Center for Health Stat.*, Series 2, 54.

Koop, J. C. (1960). On theoretical questions underlying the technique of replicated or interpenetrating samples. *Proc. Soc. Stat. Sect. Amer. Stat. Assoc.*, 196–205.

Koop, J. C. (1968). An exercise in ratio estimation. *Amer. Statistician*. **22**, 1, 29–30.

Kulldorff, G. (1963). Some problems of optimum allocation for sampling on two occasions. *Rev. Int. Stat. Inst.*, **31**, 24–57.

Lahiri, D. B. (1951). A method for sample selection providing unbiased ratio estimates. *Bull. Int. Stat. Inst.*, **33**, 2, 133–140.

Lieberman. G. J., y Owen, D. B. (1961). *Tables of the Hypergeometric Probability Distribution*. Stanford University Press, Stanford, Calif.

Lienau, C. C. (1941). Selection, training and performance of the National Health Survey field staff. *Amer. Jour. Hygiene*, **34**, 110–132.

Lund, R. E. (1968). Estimators in multiple frame surveys. *Proc. Soc. Sci. Sect. Amer. Stat. Assoc.*, 282–288.

McCarthy, P. J. (1966). Replication: An approach to the analysis of data from complex surveys. National Center for Health Statistics, Washington, D. C., Series, 2, 14.

McCarthy, P. J. (1969). Pseudo-replication: Half-samples. *Rev. Int. Stat. Inst.*, **37**, 239–264.

McVay, F. E. (1947). Sampling methods applied to estimating numbers of commercial orchards in a commercial peach area. *Jour. Amer. Stat. Assoc.*, **42**, 533–540.

Madow, L. H. (1946). Systematic sampling and its relation to other sampling designs. *Jour. Amer. Stat. Assoc.*, **41**, 207–214.

Madow, L. H. (1950). On the use of the county as a primary sampling unit for state estimates. *Jour. Amer. Stat. Assoc.*, **45**, 30–47.

Madow, W. G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Ann. Math. Stat.*, **19**, 535–545.

Madow, W. G. (1949). On the theory of systematic sampling, II. *Ann Math. Stat.*, **20**, 333–354.

Madow, W. G. (1953). On the theory of systematic sampling, III. *Ann. Math. Stat.*, **24**, 101–106.

Madow, W. G.. y Madow, L. H. (1944). On the theory of systematic sampling. *Ann. Math. Stat.*, **15**, 1–24.

Madow, W. G. (1965). On some aspects of response error measurement. *Proc. Soc. Stat. Soc. Amer. Stat. Assoc.*, 182–192.

Mahalanobis, P. C. (1944). On large-scale sample surveys. *Phil. Trans. Roy. Soc. London*, **B231**, 329–451.

Mahalanobis, P. C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Jour. Roy. Stat. Soc.*, **109**, 325–370.

Matérn, B. (1947). Methods of estimating the accuracy of line and sample plot surveys. *Medd. fr. Statens Skogsforsknings Institut.*, **36**, 1–138.

Matérn, B. (1960). Spatial variation. *Medd. fr. Statens Skogsforsknings Institut.*, **49**, 5, 1–144.

Mickey, M. R. (1959). Some finite population unbiased ratio and regression estimators. *Jour. Amer. Stat. Assoc.*, **54**, 594–612.

Midzuno, H. (1951). On the sampling system with probability proportionate to sum of sizes. *Ann. Inst. Stat. Math.*, **2**, 99–108.

Milne, A. (1959). The centric systematic area sample treated as a random sample. *Biometrics*, **15**, 270–297.

Moors, J. J. A. (1971). Optimization of the unrelated question randomized response model. *Jour. Amer. Stat. Assoc.*, **66**, 627–629.

Murthy, M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhya*, **18**, 379–390.

Murthy, M. N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcuta, India.

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Jour. Ind. Soc. Agric. Stat.*, **3**, 169–174.

National Bureau of Standards (1950). *Tables of the Binomial Probability Distribution*. Oficina de Prensa del Gobierno de los E.U.A., Washington, D.C.

Neter, J. (1972). How a accountants save money by sampling. *Statistics, A Guide to the Unknown.* J. M. Tanur y cols. (Eds.), Holden-Day, Inc., San Francisco, 203–211.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Jour. Roy. Stat. Soc.*, **97**, 558–606.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Jour. Amer. Stat. Assoc.*, **33**, 101–116.

Nordbotten, S. (1956). Allocation in stratified sampling by means of linear programming. *Skand. Akt. Tidskr.*, **39**, 1–6.

Nordin, J. A. (1944). Determining sample size. *Jour. Amer. Stat. Assoc.*, **39**, 497–506.

Olkin, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika*, **45**, 154–165.

Osborne, J. G. (1942). Sampling errors of systematic and random surveys of cover-type areas. *Jour. Amer. Stat. Assoc.*, **37**, 256–264.

Patterson, H. D. (1950). Sampling on successive occasions with partial replacement of units. *Jour. Roy. Stat. Soc.*, **B12**, 241–255.

Patterson, H. D. (1954). The errors of lattice sampling. *Jour. Roy. Stat. Soc.*, **B16**, 140–149.

Paulson, E. (1942). A note on the estimation of some mean values for a bivariate distribution. *Ann. Math. Stat.*, **13**, 440–444.

Payne, S. L. (1951). *The Art of Asking Questions.* Princeton University Press, Princeton, N.J.

Plackett, R. L., y Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, **33**, 305–325.

Platek, R., y Singh, M. P. (1972). Some aspects of redesign of the Canadian Labor Force Survey. *Proc. Soc. Stat. Sect. Amer. Stat. Assoc.*, 397–402.

Politz, A. N., y Simmons, W. R. (1949, 1950). An attempt to get the "not at homes" into the sample without callbacks. *Jour. Amer. Stat. Assoc.*, **44**, 9–31, y **45**, 136–137.

Pritzker, L., y Hanson, R. (1962). Measurement errors in the 1960 Census of Population. *Proc. Soc. Stat. Sect. Amer. Stat. Assoc.*, 80–89.

Quenouille, M. H. (1949). Problems in plane sampling. *Ann. Math. Stat.*, **20**, 355–375.

Quenouille, H. H. (1956). Notes on bias in estimation. *Biometrika*. **43**, 353-360.

Raiffa, H., y Schlaifer, R. (1961). *Applied Statistical Decision Theory.* Harvard Business School, Cambridge, Mass.

Rand Corporation (1955). *A Million Random Digits.* Free Press, Glencoe, Ill.

Rao, C. R. (1971). Some aspects of statistical inference in problems of sampling from finite populations. *Foundations of Statistical Inference.* V. P. Godambe and D. A. Sprott (Eds.)., Holt, Rinehart, and Winston, Toronto, Canadá 177–202.

Rao, J. N. K. (1962). On the estimation of the relative efficiency of sampling procedures. *Ann. Inst. Stat. Math.*, **14**, 143–150.

Rao, J. N. K. (1965). On two simple schemes of unequal probability sampling without replacement. *Jour. Ind. Stat. Assoc.*, **3**, 173–180.

Rao, J. N. K. (1966). Alternative estimators in *pps* sampling for multiple characteristics. *Sankhya*, **A23**, 47–60.

Rao, J. N. K. (1968). Some small sample results in ratio and regression estimation. *Jour. Ind. Stat. Assoc.*, **6**, 160–168.

Rao, J. N. K. (1969). Ratio and regression estimators. *New Developments in Survey Sampling*, N. L. Johnson y H. Smith, Jr. (Eds.) John Wiley & Sons, New York, 213-234.

Rao, J. N. K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, **60**, 125-133.

Rao, J. N. K. (1975a). On the foundations of survey sampling. In *A Survey of Statistical Design and Linear Models*, J. N. Srivastava (Ed.), American Elsevier Publishing Co., New York, 489–505.

Rao, J. N. K. (1975b). Unbiased variance estimation for multistage designs. *Sankhya* (en prensa).

Rao, J. N. K., y Bayless, D. L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. *Jour. Amer. Stat. Assoc.*, **64**, 540-559.

Rao, J. N. K., y Beegle, L. D. (1967). A Monte Carlo study of some ratio estimators. *Sankhya*, **B29**, 47-56.

Rao, J. N. K., y Graham, J. E. (1964). Rotation designs for sampling on repeated occasions. *Jour. Amer. Stat. Assoc.*, **59**, 492-509.

Rao, J. N. K., y Kuzik, R. A. (1974). Sampling errors in ratio estimation. *Indian Jour. Stat.* **36**, C, 43-58.

Rao, J. N. K., Hartley, H. O., y Cochran, W. G. (1962). A simple procedure of unequal probability sampling without replacement. *Jour. Roy. Stat. Soc.* **B24**, 482-491.

Rao, J. N. K., y Pereira, N. P. (1968). On double ratio estimators. *Sankhya*, **A30**, 83-90.

Rao, J. N. K., y Singh, M.P. (1973). On the choice of estimator in survey sampling. *Australian Jour. Stat.*, **15**, 95-104.

Rao, P. S. R. S., y Mudholkar, G. S. (1967). Generalized multivariate estimations for the mean of finite populations. *Jour. Amer. Stat. Assoc.*, **62**, 1008-1012.

Rao, P. S. R. S., y Rao, J. N. K. (1971). Small sample results for ratio estimators. *Biometrika*, **58**, 625-630.

Robson, D. S. (1952). Multiple sampling of attributes. *Jour. Amer. Stat. Assoc.*, **47**, 203-215.

Robson, D. S. (1957). Applications of multivariate polykays to the theory of unbiased ratio type estimation. *Jour. Amer. Stat. Assoc.*, **52**, 511-522.

Robson, D. S., y King, A. J. (1953). Double sampling and the Curtis impact survey. *Cornell Univ. Agr. Exp. Sta. Mem.*, 231.

Romig, H. G. (1952). *50-100 Binomial Tables.* John Wiley & Sons, New York.

Roy, J., y Chakravarti, I. M. (1960). Estimating the mean of a finite population. *Ann. Math. Stat.*, **31**, 392-398.

Royall, R. M. (1968). An old approach to finite population sampling theory. *Jour. Amer. Stat. Assoc.*, **63**, 1269-1279.

Royall, R. M. (1970a). On finite population sampling theory under certain linear regression models. *Biometrika*, **57**, 377-387.

Royall, R. M. (1970b). Finite population sampling–on labels in estimation., *Ann. Math. Stat.*, **41**, 1774-1779.

Royall, R. M. (1971). Linear regression models in finite population sampling theory. *Foundations of Statistical Inference*, V. P. Godambe, y D. A. Sprott (Eds.). Holt, Rinehart & Winston, Toronto, Canadá, 259-279.

Royall, R. M., y Herson, J. (1973). Robust estimation in finite populations, I. *Jour, Amer. Stat. Assoc.*, **68**, 880-889.

Sagen, O. K., Dunham, R. E., y Simmons, W. R. (1959). Health statistics from record sources and household interviews compared. *Proc. Soc. Stat. Sect. Amer. Stat. Assoc.*, 6-15.

Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, **54**, 499-513.

Sandelius, M. (1951). Truncated inverse binomial sampling. *Skandinavisk Aktuarietidskrift*, **34**, 41-44.

Särndal, C. E. (1972). Sample survey theory vs. general statistical theory: Estimation of the population mean, *Rev. Int. Stat. Inst.*, **40**, 1-12.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, **2**, 110-114.

Scott, A. J., y Smith, T. M. F. (1974). Analysis of repeated surveys using time series methods. *Jour. Amer. Stat. Assoc.*, **69**, 674-678.

Sedransk, J. (1965). A double sampling scheme for analytical surveys. *Jour. Amer. Stat. Assoc.*, **60**, 985-1004.

Sedransk, J. (1967). Designing some multi-factor analytical studies. *Jour. Amer. Stat. Assoc.*, **62**, 1121-1139.

Sen, A. R. (1953). On the estimate of variance in sampling with varying probabilities. *Jour. Ind. Soc. Agric. Stat.*, **5**, 119-127.

Sen, A. R. (1972). Successive sampling with p $(p \geq 1)$ auxiliary variables. *Ann. Math. Stat.*, **43**, 2031-2034.

Sen, A. R. (1973a). Theory and application of sampling on repeated occasions with several auxiliary variables. *Biometrics*, **29**, 383-385.

Sen, A. R. (1973b). Some theory of sampling on successive occasions. *Australian Jour. Stat.*, **15**, 105-110.

Seth, G. R., y Rao, J. N. K. (1964). On the comparison between simple random sampling with and without replacement. *Sankhya*, **A26**, 85-86.

Sethi, V. K. (1963). A note on optimum stratification for estimating the population means. *Australian Jour. Stat.*, **5**, 20-33.

Sethi, V. K. (1965). On optimum pairing of units. *Sankhya*, **B27**, 315-320.

Simmons, W. R. (1954). A plan to account for "not-at-homes" by combining weighting and callbacks. *Jour. of Marketing*, **11**, 42-53.

Singh, D., Jindal, K. K., y Garg, J. N. (1968). On modified systematic sampling. *Biometrika*, **55**, 541-546.

Sittig, J. (1951). The economic choice of sampling system in acceptance sampling. *Bull. Int. Stat. Inst.*, **33**, V, 51-84.

Slonim, M. J. (1960). *Sampling in a Nutshell.* Simon & Schuster, New York.

Smith, H. F. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *Jour. Agric. Sci.*, **28**, 1-23.

Smith, T. M. F. (1976). The foundations of survey sampling: A Review. *Jour. Roy. Stat. Soc.*, **A139**, 183-204.

Snedecor, G. W., y Cochran, W. G. (1967). *Statistical Methods.* Iowa State University Press, Ames, Iowa, sexta edición.

Srinath, K. P. (1971). Multiphase sampling in nonresponse problems. *Jour. Amer. Stat. Assoc.*, **16**, 583-586.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Stat.*, **16**, 243-258.

Stephan, F. F. (1941). Stratification in representative sampling. *Jour. Marketing*, **6**, 38-46.

Stephan, F. F. (1945). The expected value and variance of the reciprocal and other negative powers of a positive Bernoulli variate. *Ann. Math. Stat.*, **16**, 50-61.

Stephan, F., y McCarthy, P. J. (1958). *Sampling Opinions.* John Wiley and Sons, New York, Pág. 243.

Stuart, A. (1954). A simple presentation of optimum sampling results. *Jour. Roy. Stat. Soc.*, **B16**, 239-241.

Sukhatme, P. V. (1935). Contribution to the theory of the representative method. *Supp. Jour. Roy. Stat. Soc.*, **2**, 253-268.

Sukhatme, P. V. (1947). The problem of plot size in large-scale yield surveys. *Jour. Amer. Stat. Assoc.*, **42**, 297-310.

Sukhatme, P. V. (1954). *Sampling Theory of Surveys, With Applications.* Iowa State College Press, Ames, Iowa.

Sukhatme, P. V., y Seth, G. R. (1952). Non-sampling errors in surveys. *Jour. Ind. Soc. Agr. Stat.*, **4**, 5-41.

Sukhatme, P. V., y Sukhatme, B. V. (1970). *Sampling Theory of Surveys With Applications.* Food and Agriculture Organization, Rome, segunda edición.

Tepping, B. J., y Boland, K. L. (1972). Response variance in the Current Population Survey. U.S. Bureau of the Census Working Paper Núm. 36, Oficina de Prensa del Gobierno de los E.U.A., Washington, D.C.

Tin, M. (1965). Comparison of some ratio estimators. *Jour. Amer. Stat. Assoc.*, **60**, 294-307.

Trueblood, R. M., y Cyert, R. M. (1957). *Sampling Techniques in Accounting,* Prentice-Hall, Englewood Cliffs, N.J.

Trussell, R. E., y Elinson, J. (1959). *Chronic Illness in a Large City.* Harvard University Press, Cambridge, Mass., Págs. 339-370.

Tschuprow, A. A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, **2**, 461-493, 646-683.

Tukey, J. W. (1950). Some sampling simplified. *Jour. Amer. Stat. Assoc.*, **45**, 501-519.

Tukey. J. W. (1958). Bias and confidence in not-quite large samples. *Ann. Math. Stat.*, **29**, 614.

U. N. Statistical Office (1950). *The preparation of sample survey reports.* Stat. Papers Series C, Núm. 1.

U. N. Statistical Office (1960). *Sample Surveys of Current Interest.* Octavo informe.

U. S. Bureau of the Census. (1968). *Evaluation and Research Program of the U.S. Census of Population and Housting, 1960: Effects of Interviews and Crew Leaders.* Series ER 60, Núm. 7, Washington, D.C.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Jour. Amer. Stat. Assoc.*, **60**, 63-69.

Warner, S. L. (1971). The linear randomized response model. *Jour. Amer. Stat. Assoc.*, **66**, 884-888.

Watson, D. J. (1937). The estimation of leaf areas. *Jour. Agr. Sci.*, **27**, 474.
West, Q. M. (1951). *The Results of Applying a Simple Random Sampling Process to Farm Management Data.* Agricultural Experiment Station, Cornell University.
Williams, W. H. (1963). The precision of some unbiased regression estimators. *Biometrika*, **17**, 267-274.
Wishart, J. (1952). Moment-coefficients of the *k*-statistics in samples from a finite population. *Biometrika*, **39**, 1-13.
Wold, H. O. A. (1954). *A Study in the Analysis of Stationary Time Series.* Almqvist and Wicksell, Stockholm, segunda edición.
Woodruff, R. S. (1959). The use of rotating samples in the Census Bureau's Monthly Surveys. *Proc. Soc. Stat. Sect. Amer. Stat. Assoc.*, 130-138.
Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Jour. Amer. Stat. Assoc.*, **66**, 411-414.
Woolsey, T. D. (1956). Sampling methods for a small household survey. *Pub. Health Monographs*, Núm. 40.

Yates, F. (1948). Systematic sampling. *Phil. Trans. Roy. Soc. London*, **A241**, 345-377.
Yates, F. (1960). *Sampling Methods for Censuses and Surveys.* Charles Griffin and Co., London, tercera edición.
Yates, F., y Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Jour. Roy. Stat. Soc.*, **B15**, 253-261.

Zarkovic, S. S. (1960). On the efficiency of sampling with various probabilities and the selection of units with replacement. *Metrika*, **3**, 53-60.
Zukhovitsky, S. I., y Avdeyeva, L. I. (1966). *Linear and Convex Programming.* W. B. Saunders, Philadelphia.

Respuestas a los Ejercicios

1.1. (*a*) Como ejemplos de problemas de definición, se tienen las decisiones de contar las letras en el prólogo o en el índice y de cómo tratar a los símbolos matemáticos como "palabras". Sin embargo, en un libro como éste, con muchos símbolos matemáticos, parece poco probable que se quiera contar las palabras, ya sea incluyendo u omitiendo los símbolos. (*b*) (1) Las páginas constituyen un marco conveniente. Una desventaja de una página como unidad de muestreo en la que contamos todas las palabras en una página de muestra es que con numerosas ilustraciones, el número de palabras por página puede ser muy variable, debido a las páginas incompletas. Quizá valdría la pena primero poner en lista todas las páginas incompletas y formar dos subpoblaciones o estratos de páginas incompletas y uno de páginas completas, muestreadas separadamente por el método de muestreo estratificado descrito en el Cap. 5. (2) Un problema con el renglón es que el obtener una lista de renglones para que se puedan muestrear directamente los renglones, consumiría mucho tiempo. También hay renglones incompletos al final de los párrafos. Sin embargo, como las palabras por renglón son más o menos estables en números, la solución podría ser el usar un muestreo en dos etapas (Cap. 11) en el que se extraiga primero una muestra de páginas, y después se cuente el número de renglones en cada página seleccionada y se extraiga una submuestra de renglones en estas páginas.

1.2. La cuestión supone que primero debemos extraer una muestra de tarjetas con probabilidad igual. (*a*) Si los nombres de muestra que no están en la población objeto se descartan, el único problema es que el tamaño de la muestra de nombres a partir de la población objeto será generalmente más pequeña que el número de tarjetas y además, será una variable aleatoria que depende de las tarjetas elegidas. (*b*) El problema es que los nombres que aparecen en varias tarjetas tienen probabilidades de selección más elevadas. Una manera de tratar esto es contar el número de tarjetas en las que aparece un nombre seleccionado y usar este número al hacer la estimación por métodos apropiados a la selección con probabilidades desiguales (Cap. 9A). Otra manera que da a cada nombre la misma oportunidad, pero que puede involucrar muchos rechazos, es la de retener una tarjeta sólo si es la primera del conjunto en el que aparece el nombre. (*c*) Tal como en (*b*) los nombres se seleccionan con probabilidades desiguales. No se conoce un método fácil de dar a cada nombre la misma oportunidad. Si el número de tarjetas en las que aparece cada nombre se registró en alguna parte, se puede utilizar un método de probabilidad igual como en (*b*).

1.3. Como sugerencias tenemos: (*a*) un directorio reciente de almacenes y de tiendas de equipajes, (*b*) las oficinas de artículos perdidos que hay en las compañías de trenes subterráneos y de autobuses, y (*c*) hospitales y mé-

dicos privados en el área geográfica donde se presentan mordeduras de serpiente, además de cualquier organización de salud pública donde se reportan las mordeduras, son obligatorios. Las debilidades de los tres marcos son posiblemente el estado incompleto más los costos elevados en (*c*), si las mordeduras de serpientes son raras y no se reportan a un centro. (*d*) Una lista de casas suele usarse como marco para seleccionar una muestra de familias. Aunque habrá cierto estado incompleto (familias con las que no es posible entrar en contacto) el problema mayor puede ser el de errores de medición.

1.4. Un problema es el estado incompleto debido a la nueva construcción. En una muestra de direcciones, los nuevos inquilinos usualmente pueden ser tratados por el entrevistador que verifica para cualquier dirección de muestra si hay nuevos domicilios entre esta dirección y la siguiente en el directorio y de ser así, incluirá estos nuevos domicilios en la muestra. Podrían no mencionarse las áreas completas de nuevas construcciones en el directorio, lo que requiere el desarrollo de un marco separado. Se prefiere sacar una muestra de direcciones a extraer una lista de personas, porque las direcciones son más permanentes. Sin embargo, por razones de gastos de viaje, la unidad de muestreo podría ser una manzana de la ciudad de la cual se seleccionaría una submuestra de viviendas.

1.5. \$80 390 y \$82 970.

1.6. La probabilidad de confianza es aproximadamente de 0.054 (se encontró a partir de $t = -1.67$ con 25 grados de libertad). Esto supone que los recibos futuros siguen la misma distribución de frecuencia que la muestra de 26 recibos y que esta distribución es normal.

1.7. Cuando el ECM se debe enteramente al sesgo, la estimación siempre es incorrecta por una magnitud de 1 $\sqrt{\text{ECM}}$. La probabilidad de un error mayor que 1 $\sqrt{\text{ECM}}$ es, por tanto, la unidad, y la probabilidad de un error mayor que 1.96 $\sqrt{\text{ECM}}$ o mayor que 2.576 $\sqrt{\text{ECM}}$ es cero.

2.4. $\hat{Y} = 51\,473$. Probabilidad cerca de 0.9.

2.5. Sí. $\sigma(\hat{Y})$ es 98.4.

2.6. $\hat{Y} = 20\,238$, $s(\hat{Y}) = 849$.

2.7. (*a*) Público: $\hat{R} = 15.46$. Privado: $\hat{R} = 12.75$. (*b*) Público: $s(\hat{R}) = 0.761$. Privado: $s(\hat{R}) = 0.727$. Para la cpf tomamos $f = 100/468$. (*c*) $14.2 < R < 16.7$.

2.8. Dif./e.e.$_{\text{dif}} = 2.71/1.186 = 2.28$. P cerca de 0.023. Note que no se usa la cpf en el cálculo de e.e.$_{\text{dif}}$.

2.9. (*a*) 9 408, e.e. = 780; (*b*) 9 472, e.e. = 1 104.

2.10. E.e. (en 1 000'es) = (*a*) 14 800; (*b*) 3 900; (*c*) 3 140.

2.11. 9.2 (*a*) 2.7; (*b*) 2.4.

2.12. (*a*) $n = 60$, con 30 de cada grupo; (*b*) $n = 80$ será suficiente si el número de propietarios en la muestra está entre 20 y 60. Con $n = 80$, la probabilidad de que esto pase es 0.54 (de las tablás binomiales). Con $n = 100$, cuando el número de propietarios de la muestra esté entre 19 y 81 será suficiente, esta probabilidad es 0.94.

2.14. (*a*) 420; (*b*) 490; (*c*) ambos son insesgados; (*d*) estimación (*b*).

3.2. 1 066, 1 334 dado por la aproximación normal, Ec. 3.19.

3.3. Casi concluyente.

3.6. (*a*) $76.2 \pm 3.6\%$; (*b*) $1\,738 \pm 280$ familias.

3.7. $1\,789 \pm 268$ familias.

3.8. Como un resultado exacto,

$$\frac{V(\hat{A}_1)}{V(\hat{A}_1')} = \frac{N_1^2 n Q_1}{N^2 n_1 (1-\pi)(Q_1 + P_1\pi)}$$

Ahora $N_1 = N(1-\pi)$ y en muestras grandes $n_1 \doteq n(1-\pi)$. Estas sustituciones dan el resultado apuntado. Para que $V(\hat{A}_1)V(\hat{A}_1')$ sea pequeña, debemos tener $\pi(1-Q_1)/Q_1$ grande. Esto significa que Q_1 debe ser pequeña: en otras palabras, la proporción del grupo 1 que está en la clase C debe ser grande. Para una Q_1 dada, π debe ser grande.

3.9. Todos dan $A_U = 13$. Por la hipergeométrica, la probabilidad de no unidades en C en la muestra es 0.0601 para $A_U = 12$ y 0.0434 para $A_U = 13$. Por la binomial, $P_U = 0.4507$ y $\sqrt{1 - fP_U} = 0.4114$, dando $A_U = 12.3$. El Ejercicio 3 de la Pág. 90 da 0.061 para $A_U = 12$ y 0.044 para $A_U = 13$.

3.11. La estimación (*b*) parece más precisa.

3.12. El valor más alto es PQ/n en comparación con PQ/mn por la fórmula binomial. Esto sucede cuando todo conglomerado consiste enteramente de unos o enteramente de ceros. El valor más bajo puede ser cero si todo conglomerado da la misma proporción P. (Esto es posible solamente para ciertos valores de P y m.)

3.13. La varianza es 0.00184 por el método de la razón y 0.00160 por la fórmula binomial.

3.14. Tamaño promedio de la muestra = m/P.

4.1. 735 casas. Se necesita este tamaño de muestra para casas con dos coches si $P = 10\%$.

4.2. Aproximadamente 260 hojas.

4.3. (*a*) 2 475; (*b*) 4 950.

4.4. $n = 21$ (tomando $t = 2$).

4.5. $n = 484$. Para el número de sin empleo, el cv sería aproximadamente de 15%.

4.6. 62 más.

4.7. (*a*) $n = 278$; (*b*) $n = 2\,315$; (*c*) $n = 3\,046$.

4.8. Si se supone una distribución rectangular dentro de cada clase, tomamos $S^2 = 0.083h^2$ o $S = 0.29h$. Esto da estimaciones de 230, 580, 2 030 y

11 600 en las cuatro clases. Si se usa la distribución de triángulo rectángulo en la cuarta clase, tomamos $S = 0.24h$, lo que da 9 600 para esta clase.

4.11. $n_{\bar{y}pt} = \left(\dfrac{UNS}{2c\sqrt{2\pi}}\right)^{2/3}$.

5.1. (a) La asignación de Neyman da $n_1 = 0.87$, $n_2 = 13.13$. (b) Hay tres posibles estimaciones bajo asignación óptima y nueve bajo asignación proporcional. $V_{\acute{o}pt}(\bar{y}_{st}) = \frac{1}{6} = 0.167$: $V_{prop}(\bar{y}_{st}) = \frac{7}{12} = 0.583$. ($d$) Fórmula 5.27 da $V_{\acute{o}pt}(\bar{y}_{st}) = 0.159$.

5.2. (a) $n_1 = 375$, $n_2 = 625$; (b) $n_1 = 750$, $n_2 = 250$.

5.3. $RP = 181\%$ para asignación proporcional y 214% para asignación óptima.

5.5. Los aumentos relativos máximos ocurren cuando $W_1 = W_2$ e igual 0.029 para $c_2/c_1 = 2$ y 0.111 para $c_2/c_1 = 4$.

5.6. (a) $n_1/n = \frac{1}{3}$, $n_2/n = \frac{2}{3}$; (b) $n = 264$, $n_1 = 88$, $n_2 = 176$; (c) \$1 936.

5.7. (a) \$2 288 contra \$1 936. (b) No. El costo de campo mínimo para reducir V a 1 es \$2 230.

5.8. (a) $n_1 = 384$, $n_2 = 192$; (b) $n_1 = 400$, $n_2 = 1\,600$; (c) $n_1 = 1\,200$, $n_2 = 2\,400$.

5.9. Aumento fraccionario = $\frac{1}{9}$.

5.10. $n_1 = 541$, $n_2 = 313$, $n_3 = 146$.

5.12. En población 1, $V_{prop} = 0.143/n$; $V_{\acute{o}pt} = 0.134/n$. En población 2, $V_{prop} = 0.0491/n$, $V_{\acute{o}pt} = 0.0423/n$. La reducción en varianza de asignación óptima es cerca de 6% en población 1 contra 14% en población 2.

5.14. (a) Si conjeturamos $P_1 = 45\%$, $P_2 = 25\%$, $P_3 = 7.5\%$ como un convenio, esto da $n_1 = 268$, $n_2 = 116$, $n_3 = 16$; (b) e.e. = 0.0225; (c) e.e. = 0.0241.

5.15. Al aproximarse n a N, ¿alcanza una etapa en la que la fórmula estándar $n_h \propto N_h S_h$ para la asignación óptima de Neyman, no es ya aplicable, porque ello requeriría $n_h > N_h$ al menos en un estrato. Como se observó en la Sec. 5.8, la fórmula (5.27) deja de ser válida. El estudiante está en el error si afirma que (5.27) es errónea siempre; la fórmula tiene un área limitada de aplicaciones, que sin embargo, las cubre casi todas.

5A.4. No. En cada uno de los peores casos $[\Sigma(w_h - W_h)\bar{Y}_h]^2$ es $(0.105)^2 = 0.0110$. Entonces, con estratificación, ECM ($\bar{y}_{st}$), según lo da la fórmula (5A.6) es $0.0108 + 0.0110 = 0.0218$. Con muestreo aleatorio simple $V(\bar{y}) = 0.0177$.

5A.6. (a) $n = 1024$. La asignación óptima para la segunda variable (cantidad promedio invertida) satisface ambos requerimientos.

5A.7. $W_1 = 0.728$, $W_2 = 0.272$, $S_1 = 1.806$, $S_2 = 4.698$ (en la escala codificada). (a) Los tamaños óptimos de muestra son $n_1 = 0.507n$, $n_2 = 0.493n$. (b) $V(\bar{y}) = 31.95/n$, $V_{\acute{o}pt}(\bar{y}_{st}) = 6.72/n$.

5A.8. (b) $\int_0^a \sqrt{f(y)}\,dy = \int_0^a \sqrt{2(1-y)}\,dt = 2\sqrt{2}[1-(1-a)^{3/2}]/3$. Por lo tanto, queremos $[1 - (1 - a)^{3/2}] = \frac{1}{2}$.

5A.9. En el Ejercicio 5A.7, $W_1 = 0.728$, como en la regla de Dalenius-Hodge, se acerca lo más a la regla de Ekman. En el Ejercicio 5A.8, la regla de Ekman, da $a = (3 - \sqrt{5})/2 = 0.38$.

5A.10. Las óptimas son $L = 7$ para $p = 0.95$, $L = 5$ para $p = 0.9$ y $L = 4$ para $p = 0.8$. Ya sea $L = 5$ o $L = 6$ son un buen compromiso.

5A.11. (a) La ganancia en precisión es cerca de 110%. (b) La ganancia de estratificación proporcional sobre muestreo aleatorio simple es cerca de 90%.

5A.12. Aumentar n_1 como se sugiere, dejando $n_2 = 400$. Requerimos $n_1 = 140$, dando $n = 540$.

6.1. Para la estimación de razón $V(\hat{Y}_R) = N^2(1-f)S_d^2/n$ y para la expansión simple $V(\hat{Y}) = N^2(1 - f)S^2_y/n$, donde $d = (y - Rx)$. Para la muestra de 21 familias las estimaciones de S_d^2 y S_y^2 son las siguientes. Número de hijos, $s_d^2 = 0.49$, $s_y^2 = 1.61$; número de coches, $s_d^2 = 0.41$, $s_y^2 = 0.39$; número de receptores de TV, $s_d^2 = 0.51$, $s_y^2 = 0.45$. La estimación de razón parece ser superior para los hijos.

6.2. Ganancia = 66%. Cuando menos 11 unidades por el método de razón.

6.3. Límites cuadráticos (27,100, 29,870); límites normales (27,030, 29,700).

6.4. Aplique el Teorema 6.3 a la estimación de $R = \bar{Y}/\bar{X}$. Con muestras grandes, use $\bar{y}/\bar{X}$ si $r \leq$ (cv de x)/2(cv de y) y use $\bar{y}/\bar{x}$ de otra manera, donde r y los cv son estimaciones de muestras.

6.5. Los ECM son 46.5 para la estimación de razón separada y 40.6 para la estimación de razón combinada. En ambos casos la contribución del sesgo al ECM es despreciable.

6.6. Para el método de Lahiri, $V(\hat{Y}_{Rs}) = 40.1$.

6.7. Total de población estimado = 116.21 millones. La varianza relativa es 0.00111, de tal manera que el e.e. es (0.0333) (116.21) = 3.87 millones. La estimación está dentro de 1 e.e. del total verdadero.

6.8. Las estimaciones son (a) 1 896, (b) 1 660, (c) 1 689. En (c) encontramos $w_1 = 2.38$, $w_2 = -1.38$. Los e.e. estimados son (a) 256, (b) 36.9, (c) 18.6. Para la estimación estándar en (b) se usó la fórmula

e.e. = $\hat{\bar{Y}}_R\sqrt{(1-f)(c_{yy} + c_{11} - 2c_{y1})/n}$, donde $\hat{\bar{Y}}_R$ es la estimación de razón de $\bar{Y}$, esto es, 1 660. Para el e.e. en (c) se usó $\hat{\bar{Y}}_{MR} = 1\,689$.

7.1. Estimación = 11 080; e.e. = 152 (incluyendo la cpf).

7.2. No, dado que b es muy cercano a 1.

7.3. $\bar{Y}_{lr} = 28\,177 \pm 570$. La precisión relativa es 113%.

7.4. 27 751 ± 694.

7.6. Para la estimación de la diferencia, $V(\bar{y}) = S_e^2/n$; para la estimación de la regresión lineal, $V(\bar{y}_{lr}) = S_e^2 S_y^2/n(S_e^2 + S_y^2)$. La estimación de la regresión tiene la varianza menor, pero su superioridad no tiene importancia si S_e^2/S_y^2 es pequeña.

7.7. $\text{ECM}(\hat{Y}_{lrs}) = 34.5$, $\text{Sesgo}^2 = 9.7$; $\text{ECM}(\hat{Y}_{lrc}) = 11.9$, $\text{Sesgo}^2 = 1.2$.

7.9. $\text{ECM}(\hat{Y}_{Rs}) = 8.9$, $\text{ECM}(\hat{Y}_{Rc}) = 6.7$.

8.1. Las varianzas son 8.19 (sistemático), 11.27 (aleatorio simple), 8.25 (estratificado, 2), 7.46 (estratificado, 1).

8.2. $V_{sys} = 0.00141$, $V_{ran} = 0.00340$.

8.3. La muestra sistemática debería ser mejor para la proporción de personas de ascendencia polaca, dado que esta variable exhibe estratificación geográfica. Es probablemente inferior para la proporción de hijos porque el intervalo de muestreo, 1 en 5, coincide con el tamaño promedio de una familia. Lo mismo es cierto, aunque en menor medida, para la proporción de hombres.

8.4. Las varianzas son las siguientes. Hombres. $V_{srs} = 0.0204$, $V_{sys} = 0.0216$; hijos, $V_{srs} = 0.0204$, $V_{sys} = 0.0776$; profesionales, $V_{srs} = 0.0192$, $V_{sys} = 0.0016$.

8.5. Varianza real = 8.19. Método (a) da 11.29. Para el método (b) la varianza estimada de una sola muestra es $(1 - f)\ (\bar{y}_{i1} - \bar{y}_{i2})^2/4$, donde $\bar{y}_{i1}, \bar{y}_{i2}$ son las medias de las dos mitades. El promedio es 3.24. La subestimación seria es inesperada.

8.7. Ambas varianzas son $(k^2 - 1)/6$.

8.8. El muestreo aleatorio simple es mejor, a menos que $n = 1$ o $k = 1$.

8.9. Cada k-ésima unidad, $V(\hat{Y}) = 362.2$; Yates, $\text{ECM}(\hat{Y}) = 7.3$; Sethi, $V(\hat{Y}) = 21.0$; Singh, y cols., $V(\hat{Y}) = 81.0$. Los dos últimos estimadores son insesgados en este ejemplo.

9.1. Los costos relativos de usar los cuatro tipos de unidad son 100, 90.1, 79.7 y 77.8 (tomando la primera unidad como estándar).

9.2. La precisión relativa de la familia es 211% para la proporción de sexos y 38% para la proporción que ha visitado a un médico.

9.3. La precisión relativa de la unidad grande es 0.566 con muestreo aleatorio simple y 0.625 con muestreo aleatorio estratificado.

9.5. (a) $M = 5$; (b) $M = 1$.

9.6. La M óptima debería disminuir porque el costo de viaje, el cual varía con $\sqrt{n}$, se convierte relativamente en menos importante conforme n aumenta.

9A.1. (a) 34 242; (b) 5 534; (c) 6 493.

9A.3. (a) Si la d.e. entre unidades grandes en la clase h α M_h. (b) Si la probabilidad α $\sqrt{M_h}$.

9A.4. (b) $V(\hat{Y}_{HT}) = 1.75$, $V(\hat{Y}_M) = 0.27$, $V(\hat{Y}_{RHC}) = 0.33$. En este problema y_i/z_i varía muy poco, pero $\hat{Y}_{HT}$ da resultados mediocres (relativamente) porque con el método de selección de muestra $\pi_i \neq 2z_i$. El estimador de Murthy se construye para este método de selección de muestra teniendo $V(\hat{Y}_M) = 0$ si y_i/z_i es constante, como lo muestra (9A.60). (c) $V(\hat{Y}_M)/V(\hat{Y}_{ppz}) = 0.54$ en tanto que $(N - n)/(N - 1)' = \frac{1}{2}$.

9A.5. $\text{ECM}(\hat{Y}^*_{HT(A)}) = 7.06$, $(\text{Sesgo})^2/\text{ECM} = 0.065$, $V(\hat{Y}_{SRS}) = 6.5$

10.1. (a) 2.00; (b) 2.13.

10.3. (a) $165/n$; $(b)148.5/n$; (c) $132/n$.

10.4. (a) $n = 660$ campos; (b) $n = 530$ campos. La proteína requiere menos campos que rendimiento.

10.5. $c_1/c_2 = 8$.

10.7. (a) 0.93%; (b) 0.51%; (c) 0.36%.

10.8. (a) Cualquiera $m_o = 7$ o $m_o = 8$ es apropiado; (b) 89% para $m_o = 7$ y 93% para $m_o = 8$; (c) 86% para $m_o = 7$ y 89% para $m_o = 8$.

11.2. La precisión relativa de III a II baja de 3.02 a 2.75. Si dos planes de muestreo difieren principalmente en su contribución entre-unidades a la varianza, la precisión relativa del plan superior disminuirá en general conforme la razón de la varianza dentro-de-unidades con los aumentos de la varianza total.

11.3. La explicación es, dicho brevemente, que con estos datos las Y_i/z_i son más estables que las Y_i/M_i. Si tomamos $z_i = \frac{1}{33}, \frac{8}{33}$, y $\frac{24}{33}$, la contribución entre-unidades en el método IV desaparecería.

11.4. Varianza total: 0.00504 (Ia), 0.02358 (II), 0.00554 (III).

11.6. Porcentaje estimado 14.2 ± 2.16. Número estimado 3 540 ± 540.

11.7. Porcentaje estimado 13.9 ± 2.49.

11.9. (a) Total de habitaciones 29 400; total de personas 50 550; personas por habitación 1.72; (b) e.e.: total de personas 2 440; personas por habitación 0.066.

12.1. $n = 267$, $n' = 1\,320$ o $n = 268$, $n' = 1\,280$. $V(p_{st})$ con asignación óptima es 6.67 cuando p_{st} es en %. Con muestreo simple, $V(p) = 8.33$.

12.2. $c_n/c_{n'} > 9$.

12.3. $n/n' = 1/19$.

12.4. $n' > 16n$.

12.5. Por la fórmula (12.67), e.e. = 1.25, al ignorar $1/N$.

12.6. Las ganancias en porcentaje de la segunda a la sexta ocasión son 50, 75, 91, 100 y 105, respectivamente.

12.8. Los valores de $nV(\bar{y}_2'')/S^2$ y $nV(\bar{y}_2')/S^2$ son los siguientes: $\mu = \frac{1}{4}$, $\rho = 0.8$: 0.885, 0.875; $\mu = \frac{1}{4}$, $\rho = 0.9$: 0.843, 0.840; $\mu = \frac{1}{2}$, $\rho = 0.8$: 0.824, 0.810; $\mu = \frac{1}{2}$, $\rho = 0.9$: 0.752, 0.746.

7.4. 27 751 ± 694.

7.6. Para la estimación de la diferencia, $V(\bar{y}) = S_e^2/n$; para la estimación de la regresión lineal, $V(\bar{y}_{lr}) = S_e^2 S_y^2/n(S_e^2 + S_y^2)$. La estimación de la regresión tiene la varianza menor, pero su superioridad no tiene importancia si S_e^2/S_y^2 es pequeña.

7.7. $\text{ECM}(\hat{Y}_{lrs}) = 34.5$, $\text{Sesgo}^2 = 9.7$; $\text{ECM}(\hat{Y}_{lrc}) = 11.9$, $\text{Sesgo}^2 = 1.2$.

7.9. $\text{ECM}(\hat{Y}_{Rs}) = 8.9$, $\text{ECM}(\hat{Y}_{Rc}) = 6.7$.

8.1. Las varianzas son 8.19 (sistemático), 11.27 (aleatorio simple), 8.25 (estratificado, 2), 7.46 (estratificado, 1).

8.2. $V_{sys} = 0.00141$, $V_{ran} = 0.00340$.

8.3. La muestra sistemática debería ser mejor para la proporción de personas de ascendencia polaca, dado que esta variable exhibe estratificación geográfica. Es probablemente inferior para la proporción de hijos porque el intervalo de muestreo, 1 en 5, coincide con el tamaño promedio de una familia. Lo mismo es cierto, aunque en menor medida, para la proporción de hombres.

8.4. Las varianzas son las siguientes. Hombres. $V_{srs} = 0.0204$, $V_{sys} = 0.0216$; hijos, $V_{srs} = 0.0204$, $V_{sys} = 0.0776$; profesionales, $V_{srs} = 0.0192$, $V_{sys} = 0.0016$.

8.5. Varianza real = 8.19. Método (a) da 11.29. Para el método (b) la varianza estimada de una sola muestra es $(1 - f)\ (\bar{y}_{i1} - \bar{y}_{i2})^2/4$, donde $\bar{y}_{i1}, \bar{y}_{i2}$ son las medias de las dos mitades. El promedio es 3.24. La subestimación seria es inesperada.

8.7. Ambas varianzas son $(k^2 - 1)/6$.

8.8. El muestreo aleatorio simple es mejor, a menos que $n = 1$ o $k = 1$.

8.9. Cada k-ésima unidad, $V(\hat{Y}) = 362.2$; Yates, $\text{ECM}(\hat{Y}) = 7.3$; Sethi, $V(\hat{Y}) = 21.0$; Singh, y cols., $V(\hat{Y}) = 81.0$. Los dos últimos estimadores son insesgados en este ejemplo.

9.1. Los costos relativos de usar los cuatro tipos de unidad son 100, 90.1, 79.7 y 77.8 (tomando la primera unidad como estándar).

9.2. La precisión relativa de la familia es 211% para la proporción de sexos y 38% para la proporción que ha visitado a un médico.

9.3. La precisión relativa de la unidad grande es 0.566 con muestreo aleatorio simple y 0.625 con muestreo aleatorio estratificado.

9.5. (a) $M = 5$; (b) $M = 1$.

9.6. La M óptima debería disminuir porque el costo de viaje, el cual varía con $\sqrt{n}$, se convierte relativamente en menos importante conforme n aumenta.

9A.1. (a) 34 242; (b) 5 534; (c) 6 493.

9A.3. (a) Si la d.e. entre unidades grandes en la clase h α M_h. (b) Si la probabilidad α $\sqrt{M_h}$.

9A.4. (b) $V(\hat{Y}_{HT}) = 1.75$, $V(\hat{Y}_M) = 0.27$, $V(\hat{Y}_{RHC}) = 0.33$. En este problema y_i/z_i varía muy poco, pero $\hat{Y}_{HT}$ da resultados mediocres (relativamente) porque con el método de selección de muestra $\pi_i \neq 2z_i$. El estimador de Murthy se construye para este método de selección de muestra teniendo $V(\hat{Y}_M) = 0$ si y_i/z_i es constante, como lo muestra (9A.60). (c) $V(\hat{Y}_M)/V(\hat{Y}_{ppz}) = 0.54$ en tanto que $(N - n)/(N - 1)' = \frac{1}{2}$.

9A.5. $\text{ECM}(\hat{Y}^*_{HT(A)}) = 7.06$, $(\text{Sesgo})^2/\text{ECM} = 0.065$, $V(\hat{Y}_{SRS}) = 6.5$

10.1. (a) 2.00; (b) 2.13.

10.3. (a) $165/n$; $(b)148.5/n$; (c) $132/n$.

10.4. (a) $n = 660$ campos; (b) $n = 530$ campos. La proteína requiere menos campos que rendimiento.

10.5. $c_1/c_2 = 8$.

10.7. (a) 0.93%; (b) 0.51%; (c) 0.36%.

10.8. (a) Cualquiera $m_o = 7$ o $m_o = 8$ es apropiado; (b) 89% para $m_o = 7$ y 93% para $m_o = 8$; (c) 86% para $m_o = 7$ y 89% para $m_o = 8$.

11.2. La precisión relativa de III a II baja de 3.02 a 2.75. Si dos planes de muestreo difieren principalmente en su contribución entre-unidades a la varianza, la precisión relativa del plan superior disminuirá en general conforme la razón de la varianza dentro-de-unidades con los aumentos de la varianza total.

11.3. La explicación es, dicho brevemente, que con estos datos las Y_i/z_i son más estables que las Y_i/M_i. Si tomamos $z_i = \frac{1}{33}, \frac{8}{33}$, y $\frac{24}{33}$, la contribución entre-unidades en el método IV desaparecería.

11.4. Varianza total: 0.00504 (Ia), 0.02358 (II), 0.00554 (III).

11.6. Porcentaje estimado 14.2 ± 2.16. Número estimado 3 540 ± 540.

11.7. Porcentaje estimado 13.9 ± 2.49.

11.9. (a) Total de habitaciones 29 400; total de personas 50 550; personas por habitación 1.72; (b) e.e.: total de personas 2 440; personas por habitación 0.066.

12.1. $n = 267$, $n' = 1\,320$ o $n = 268$, $n' = 1\,280$. $V(p_{st})$ con asignación óptima es 6.67 cuando p_{st} es en %. Con muestreo simple, $V(p) = 8.33$.

12.2. $c_n/c_{n'} > 9$.

12.3. $n/n' = 1/19$.

12.4. $n' > 16n$.

12.5. Por la fórmula (12.67), e.e. = 1.25, al ignorar $1/N$.

12.6. Las ganancias en porcentaje de la segunda a la sexta ocasión son 50, 75, 91, 100 y 105, respectivamente.

12.8. Los valores de $nV(\bar{y}_2'')/S^2$ y $nV(\bar{y}_2')/S^2$ son los siguientes: $\mu = \frac{1}{4}$, $\rho = 0.8$: 0.885, 0.875; $\mu = \frac{1}{4}$, $\rho = 0.9$: 0.843, 0.840; $\mu = \frac{1}{2}$, $\rho = 0.8$: 0.824, 0.810; $\mu = \frac{1}{2}$, $\rho = 0.9$: 0.752, 0.746.

12.12. (a) Sea $y_i = 1$ para toda unidad que tiene el primer atributo y $y_i = 0$ en cero contrario y sea el estrato 1 el estrato en que cada unidad tiene el segundo atributo. En la rotación del Teorema 12.1, 12.2, con $1/N$ despreciable, $S^2 = P_1Q_1$ y $S_1^2 = P_1P_2/(P_1 + P_2)^2$. Los resultados de ($a$) se siguen de los teoremas; (b) si C^* es el costo esperado, al muestreo doble con $v_1 = 1/2$, $k = 2$ (el óptimo) de $V(\hat{Y}_{st}) = N^2(0.844)/C^*$, en tanto que una muestra aleatoria simple da $V(\hat{Y}) = N^2(1.875)/C^*$, considerable.

13.1. (c) 90% de respuesta con 1 047 cuestionarios completos a 95% de respuesta con 701 cuestionarios completos.

13.2. El método con 90% de respuesta cuesta \$5 235. Con 95% de respuesta cuesta \$5.7895 por cuestionario completo o \$4 058 de costo total.

13.3. (b) $0.65n_o$, $0.815n_o$, $0.8915n_o$, $0.9351n_o$, $0.9611n_o$; (c) 100, 104, 108, 112, 117; (d) 300, 288, 277, 267, 256.

13.4. (a) Sesgo (en %) $= -3.85, -2.15, -1.21, -0.69, -0.40$; ($b$) las varianzas son 8.67, 9.03, 9.39, 9.74, 10.16; (c) cuatro visitas.

13.6. Sí. VC para $k = 2$ es solamente cerca de 2% sobre el mínimo VC.

13.7. Estimación Politz-Simmons, 39.7%; binomial, 42.3%.

13.8. (c) Varianza = 0.1.

13.9. Si el error de medición de cada enumerador fuese independiente de familia a familia, la varianza de la media de la muestra sería $(\sigma_\mu^2+\sigma_d^2+\sigma_{I_s}^2+\sigma_I^2)/525$ en lugar de $(\sigma_\mu^2+\sigma_d^2+35\sigma_{I_s}^2+105\sigma_I^2)/525$. Los sesgos del enumerador contribuyen a cerca del 55% de la varianza total.

13.10. $10^4V(\hat{\pi}_A) = V(100\hat{\pi}_A) =$ (a) 1.80 (Directo); (b) 10.69 (Warner); (c) 3.30 (π_u conocida); (d) 5.12 (Moors); (e) 6.30 ($P_2 = 1 - P_1$).

13.11. (a) $10^4\text{ECM}(\hat{\pi}_A) = 3.81$; $\hat{\pi}_{AU}$ es superior si π_U es conocida. (b) $10^4\text{ECM}(\hat{\pi}_A) = 5.47$; la probabilidad de dos preguntas $\hat{\pi}_{AU}$ es también superior si $P_2 = 0$. (c) $10^4\text{ECM}(\hat{\pi}_A) = 7.64$. Todos los métodos son superiores excepto el método original de Warner.

Indice Alfabético

— A —

Abreviaciones en la computación de la variancia de la estimación de razón, 221
Aleatorio, muestreo *ver* Muestreo aleatorio simple
Alistamiento de las unidades primarias, efecto del costo sobre la probabilidad óptima de selección, 386
Apareamiento en muestreo repetido de la misma población, 441-442
Aproximación de Satterwhaitte a número de grados de libertad, 133
Asignación:
  de Neyman, 136
    los mejores límites de estratos para, 169
    proporcional en muestreo estratificado, 127
    autoponderación muestral, 127
    comparación con asignación óptima, 136, 141, 147
    comparación con muestreo aleatorio simple, 136, 147
    comparación con postestratificación, 177
    consejos para su uso, 141, 148
    en la estimación de proporciones, 147
    fórmulas de la variancia, 128-132
Asignación óptima en muestreo estratificado:
  comparación con asignación proporcional, 136, 147
  comparación con muestreo aleatorio simple, 136, 147
  con estimaciones de razón, 220
  con más de una característica, 160-165
  con muestreo doble, 403
  de dos etapas, 346
  efecto de errores de:
    en S_h, 155-157
    en tamaños de muestra, 157
  efecto de las desviaciones desde la óptima (asignación), 155
  en muestreo para proporciones, 146
  estimación de datos previos, 139, 173
  para costo total fijo, 133-135
  para tamaño fijo de muestra, 135-136
  requiriendo un muestreo de más del 100%, 142
Asimetría:
  coeficiente de asimetría, 70, 248
  efecto de la estratificación sobre la, 72
  efecto sobre los límites de confianza, 67
Atributos *ver* Proporciones

— C —

Cambio, estimaciones del, 414, 426-428
Características
  cualitativas, 79 *ver también* Proporciones, estimación de
  raras, método de Haldane, 109
Censo
  de los E.U.A., uso del muestreo, 21-22
  uso del muestreo en, 21
Coeficiente
  de correlación:
    con una muestra sistemática, 261
    en poblaciones finitas, 200
    intraconglomerados, 264
  de regresión en población finita, 242
  de variación, 83
Consistencia
  definición, 45
  de la media de muestra aleatoria simple, 45
  de una estimación de razón, 198
  de una estimación de regresión, 241
Corrección por continuidad, 88, 95
Correcciones
  finales, 270
  por población finita, (cpf), 49
    en muestreo aleatorio estratificado, 128
    en muestreo de dos etapas, 342
    regla para ignorarla, 49
Correlación intraconglomerado, 262
Correlograma, 273
Costos de viaje, fórmula, 133
Covariancia de medias de muestra, 50
Cuadrado latino del movimiento de un caballo del ajedrez, 284
Cuadrático error medio, definición, 37
  justificación del uso del, 37
  relación con la varianza y el sesgo, 38

— D —

Desigualdad de Cauchy Schwarz, 134
Desviación
  de la respuesta, 458
  estándar en población finita, 47
Distribución
  binomial, 89
    límites de confianza, 90
    para proporciones de muestra, 85
    referencias a las tablas, 84
    uso erróneo en muestreo por conglomerados, 95-99
  condicional, de proporciones, 92
  hipergeométrica, 85
    como distribución condicional, 92
    límites de confianza, 87
    referencia a tablas y gráficas, 87
  normal, 29
    como aproximación a la hipergeométrica, 88
    como distribución límite de medias de muestra, 65-66
    su validez con medias de datos continuos, 66-72
    utilización en encuestas, 32
Doble muestreo, 339
  estimaciones de razón, 417-418
Dominios de estudio (subpoblaciones), 60
  comparaciones entre medias, 65
  comparaciones entre proporciones y totales, 94
  dominios que cruzan varios estratos, 186-188
  estimaciones de medias y totales (variable continua), 60-61
  estimaciones de proporciones y totales (la variable 0-1), 93-94
  estratos como dominios, 184
  tamaño requerido de la muestra, 115-117
Dos etapas, muestreo de, *ver* Muestreo doble

— E —

E_1, E_2, promedios sobre las etapas primera y segunda, 341
Efecto del diseño (Deff), 45
Elementos, 56
El método Jacknife, 222
  para estimar la varianza de una razón, 227
  para la varianza de una estimación de una función no lineal, 392

Encuesta piloto
  su uso al estimar tamaño de muestra, 111
  uso al estimar fracciones óptimas de muestreo y submuestreo, 349-351
Encuestas
  descriptivas, 23
  por correo, 189, 437, 448
Entrevistador, sesgo de, 456, 476
  modelos matemáticos, 455-458
E promedio sobre todas las muestras posibles, 33
Errores de medición, 455
  efectos de errores no correlacionados dentro de la muestra, 460
  efectos de la correlación entre errores, 462
  efectos del sesgo constante, 458
  modelo matemático, 455-458
  resumen de los efectos, 464
  técnicas para estudio de, 464-473
  utilización de muestras interpenetrantes, 469-472
Errores en encuestas, tipos de, 321
Error estándar
  (aproximado) de estimadores no lineales, 388
    comparación de los métodos, 393-395
    método de la serie de Taylor, 389
    método Jacknife, 392
    replicaciones repetidas equilibradas, 390
  de la desviación estándar de una muestra, 70
  de la diferencia entre dos razones, 229
  de la diferencia entre medias de dominios, 65
  de la estimación de razón, 58, 198-202
    en muestreo estratificado, 211-214
  de la estimación de regresión, 241, 245, 246
  de la media:
    de muestra aleatoria simple, 48, 50-51
    de muestra sistemática, 259-265
    de muestreo estratificado, 127-136
    en muestreo conglomerado, 298
    en muestreo de tres etapas, 352-355
  de la regresión en muestreo estratificado, 251-254
  del total de dominio, 61
    en muestreo estratificado, 187
  del total de población estimado de una muestra aleatoria simple, 49
  del total en una población que posee cierto atributo, 82
  de media de dominio, 59-60
    en muestreo estratificado, 189-192
  de proporción:
    en muestreo aleatorio simple, 81-84
    en muestreo conglomerado, 95-99
    en muestreo de dos etapas, 344
    en muestreo estratificado, 145-147
    sobre un dominio, 93-94
  de una razón en muestreo de dos etapas, 380
Estimación
  autoponderante, 127
    en muestreo de dos etapas, 371-372, 373, 376, 387
  compuesta en muestreo repetido, 429
  de razón combinada, 212
    al estimar medias de dominios, 187
    asignación óptima para, 220
    comparada con la estimación de razón separada, 214
    en muestreo estratificado de dos etapas, 387
    límite superior al sesgo relativo, 213
    multivariada, 234
    varianza, 213
    varianza estimada, 214
  de regresión combinada, 253
    comparada con la estimación de regresión separada, 254
    varianza, 253
    varianza estimada, 254

insesgada, definición, 33
simple, 216
Estimaciones
de varianzas de población:
efectos de errores en S_h sobre la precisión del muestreo estratificado, 155
para determinar el tamaño de muestra, 111
visuales, 239
Estimador
de razón, 56, 195
ajustes para disminuir el sesgo, 222
asignación óptima en muestreo estratificado, 220
como caso especial de estimación de regresión, 241
comparado con estimación de regresión, 246
comparado con media por unidad, 203
comparado con muestreo estratificado, 216
condiciones óptimas para, 207
condiciones para que sea insesgado, 204
consistencia, 198
cota superior a sesgo relativo, 209
en muestreo aleatorio estratificado, 211
en muestreo conglomerado, 56, 97
en muestreo doble, 416
en muestreo estratificado de dos etapas, 387
error estándar para comparación de dos razones, 229
estimación de Hartley Ross, 222
estimación de la varianza por método jackknife, 224
estimaciones insesgadas de tipo razón, 222
exactitud de variancia aproximada, 209
límites de confianza, 202
método de Lahiri, 226
multivariado, 234
sesgo de, 207-209
varianza en muestra grande, 55, 198
varianza estimada, 58, 201, 227
ver también: Estimación de razón combinada; Estimación separada de razón
de regresión, 239
comparado con estimador de razón, 246
comparado con media por unidad, 246
condiciones para que sea insesgado, 250
con pendiente preasignada, 241
efecto del error en la pendiente, 242
en muestreo aleatorio estratificado, 252, 253
en muestreo doble, 412
en muestreo repetido de la misma población, 420-431
exactitud de la varianza de muestra grande, 247
producto, 235
sesgo, 249
utilizaciones del, 239
varianza en muestras grandes, 245
varianza estimada, 245
ver también estimación combinada de regresión; Estimación separada de regresión
de regresión lineal, *ver* Estimador de regresión
Estimados de Horvitz-Thompson, 321-323
Estratificación, 125
doble, 165
efecto del número de estratos en la precisión, 174
efecto en la normalidad de la variable, 72
geográfica, 140
construcción de los estratos, 174
ganancias en precisión, 140-141
la mejor variable para una, 139
Estratos, 89
construcción de, 169-174
número óptimo de, 174-177
Etiquetas (o rótulos), 72
Expansión, factor de, 45

— F —

Factor
de expansión, 45
de inflación, 45
Fracción de muestreo, 45
primera etapa, 342
segunda etapa, 343
Fronteras de estratos, regla para su determinación, 169-173
Función
de costo:
para el número de estratos, 176
al determinar la fracción óptima de muestreo para no respondientes, 448
al determinar la fracción óptima de submuestreo, 346, 382
al determinar probabilidades óptimas de selección de unidades primarias, 380-388
en la determinación del tamaño de muestra, 117
en la determinación del tamaño óptimo de la unidad conglomerada, 302
en muestreo aleatorio estratificado, 133
en muestreo de dos marcos, 189
en muestreo doble:
con estratificación, 404
para comparaciones analíticas, 408
para estimaciones de regresión, 415
de la varianza, 300
de pérdidas, 117, 162

— G —

Grados de libertad
número efectivo en muestreo estratificado de, 133

— H —

Hueso duro en no-respuesta, 440

— I —

Indice de inconsistencia, 468
Items (*ver también* Características y atributos), 44

— K —

Keyfitz, método abreviado para estimación de varianza, 217-219

— L —

La mejor estimación insesgada, 204
Límites
cuadráticos de confianza para estimación de razón, 202
de confianza, 33
condicionales, 91
definición para atributos, 87
efecto de la no respuesta sobre, 436-440
efecto del sesgo en, 36
en muestreo aleatorio estratificado, 132
para estimaciones de razón, 202
para medias en muestreo aleatorio simple, 52
para proporciones o porcentajes, 87-90
validez de la aproximación normal, 66-72

— M —

Marco, 26
estimaciones cuando el marco tiene unidades fuera de la población objeto, 63-64
muestreo de dos marcos, 189-190
Media de la muestra, propiedades óptimas de la, 72-73
Mediciones repetidas, 466, 472
Medida de homogeneidad, 307
Método
de la serie de Taylor, 389
de respuesta aleatorizada, 473-474
original de Warner, 474
otras variantes, 477
segunda pregunta no relacionada, 475
Métodos
de estratos contraídos, 182, 282
de Politz y Simmons para tratar la no respuesta, 452-455
Miembro aleatorio en un domicilio, método de selección, 442

Muestra
de rejilla cuadrada, 283
inversa (método de Haldane), 108
por áreas, 125
sistemática alineada, 283
sistemática no alineada, 283
Muestras
interpenetrantes, 469-473
sistemáticas centralmente localizadas, 257
Muestreada, población, 25
Muestreo
aleatorio estratificado, 125
asignación de Neyman, 136
asignación óptima para costo fijo, 133-135
comparado con el muestreo aleatorio simple, 136-138
comparado con el muestreo sistemático, 262, 278
comparado con la estimación de razón, 216
con estimaciones de razón, 211
con estimaciones de regresión, 251
construcción de los estratos, 169-174
con una unidad por estrato, 181
estimación de la ganancia en precisión, 179
estimación de $\overline{y}_{st}$, 127-128
estimaciones para dominios de estudio, 186-188
tamaño de la muestra, 143, 149
tipo de población que resulta en ganancias considerables, 138
varianza estimada de P_{st}, 146
varianza estimada de $\overline{y}_{st}$, 132
aleatorio simple, 41
distribución de la proporción muestral, 84-86
estimador lineal óptimo de media de población, 72
límites de confianza:
para media de muestra, 52
para proporción de muestra, 52
método de extracción, 42
para clasificación en más de dos clases, 90
tamaño de muestra:
para medias, 109
para proporciones, 107
varianza estimada, 254
de media muestral, 50
de proporción muestral, 81
conglomerado con submuestreo, *ver* Muestreo de dos etapas
con restitución, 42, 55
de aceptación, 22
de dos etapas (unidades de diferentes tamaños), 359
comparación de los métodos, 379
estimador de razón, 380
fracciones óptimas de muestreo y submuestreo, 382-386
métodos con una unidad por estrato, 361-367
unidades elegidas:
con probabilidades desiguales con restitución, 375-377
con probabilidades desiguales sin restitución, 377-379
con probabilidades iguales, 361-362, 371-375
de dos etapas (unidades de mismo tamaño)
fracciones óptimas de muestreo y submuestreo, 346
muestreo estratificado de las unidades primarias, 355
sus ventajas, 339
tabla para la selección de tamaño óptimo de submuestra, 349
uso de encuesta piloto, 349
varianza
de media estimada, 342
de proporciones estimadas, 345
ver Muestreo doble
de tres etapas, 352
fracciones óptimas de muestreo y submuestreo, 355
varianza de la media por unidad de, 352-354
doble con estimaciones de regresión, 408
comparación con el muestreo aleatorio simple, 414
tamaños óptimos de muestras, 415
varianza, 412-413
varianza estimada, 416
doble con estratificación, 400
comparación con el muestreo aleatorio simple, 404
estimación de proporciones, 403
tamaños óptimos de muestras, 403
varianza, 400
varianza estimada, 406
por conglomerados (de una etapa):
comparación de $\hat{Y}_u$, $\hat{Y}_R$, $\hat{Y}_{pps}$, 316-320
estimación de razón $\hat{Y}_R$, 310
selección equiprobabilista, insesgada estimador $\hat{Y}_u$, 309
conglomerados de tamaños diferentes, 309
conglomerados de un mismo tamaño, 289
comparado con muestreo aleatorio simple, 299
estimación de proporciones por, 95-96, 304
forma y tamaño óptimo de conglomerado, 290-298, 304
la varianza como función del tamaño del conglomerado, 300
varianza de media por elemento, 298
estimador de Horvitz-Thompson, 321-323
en muestreo estratificado, 334
estimadores de razón, 334
probabilidades de selección óptima, 316
selección
métodos relacionados con el muestreo sistemático, 328
sin restitución, 320-321
selección con probabilidades desiguales con restitución (WR), Y_{pps}, 312-316
selección sin restitución, método de Brewer, 323
comparación de los métodos, 333
método de Murthy, 326
método de Rao, Hartley, Cochran, 329
por cuota, 178
probabilista, definición y propiedades, 29-30
alternativas al, 30-31
repetido de una población, 418
estimación compuesta, 429
estimaciones del cambio, 419, 427
estimaciones en curso, 420-431
políticas de rotación, 354
porcentaje óptimo de apareados, 422, 425
utilización de los que no responden de encuestas previas, 456
reticulado, 285
sin restitución, 41
sistemático, 257
comparado con muestreo aleatorio simple, 261-275
comparado con muestreo estratificado, 261-278
correcciones extremas, 270
efecto de la variación periódica, 271
en dos dimensiones, 283
en muestreo conglomerado de una etapa, 328
en muestreo de dos etapas, 344
en poblaciones
con orden "aleatorio", 265
con tendencia lineal, 268-271
en poblaciones autocorrelacionadas, 273
en poblaciones naturales, 276
estimación de la varianza, 278-282
método de extracción, 258
muestreo sistemático estratificado, 282
recomendaciones para su uso, 285
relación con el muestreo conglomerado, 259
variancia de la media, 260-265

—N—

No
cobertura, 441
normalidad:

efecto de la estratificación sobre la, 72
efecto sobre los límites de confianza, 67
frecuentemente encontrada en la práctica del muestreo, 67
respuesta, 436
efecto de revisitas, 442-448
efecto en límites de confianza, 437-440
efecto sobre la varianza en muestreo estratificado, 188
fracción óptima de muestreo entre no respondientes, 448-452
método de Politz y Simmons, 452- 455
razones para, 441
sesgo producido por, 436-438
sesgado según modelo, 204
Notación
para errores de medición, 455-458
para estimaciones de razón, 196
para muestreo aleatorio simple, 44
para muestreo en dos etapas, 342
para muestreo estratificado, 126
para proporciones, 79, 93
para varianzas de estimaciones, 52
Números aleatorios, 42

— O —

Objeto de población, 24

— P —

Pasos a seguir en una encuesta por muestreo, 24
Población, 24
muestreada, 25
objeto, 25
Poblaciones autocorrelacionadas, 273
Porcentaje óptimo apareado en muestreo repetido, 422, 425
Porcentajes, estimación de, *ver* Proporciones, estimación de
Postestratificación, 177
comparada con asignación proporcional, 177
Ppt (en muestreo de dos etapas), 363, 377
comparada con probabilidades iguales, 367-379
selección sin restitución, 377
Precisión en contraste con exactitud, 38
relativa, 136
método de cálculo, 141
Preguntas delicadas, 473-474
método de respuesta aleatorizada, 372-477
Probabilidad proporcional al tamaño ppt, (muestreo de una etapa), 311
comparada con probabilidades iguales, 316-320
método de extracción de muestra, 310
selección sin restitución, 320
Proporciones, estimación de, 79
con más de dos clases, 90, 91
efecto de la P de población en la precisión, 83
efecto de no respuestas, 437
en muestreo aleatorio estratificado, 145-150
en muestreo aleatorio simple, 79, 99
en muestreo conglomerado, 95-99, 304
en muestreo de dos etapas (unidades de mismo tamaño), 344
en muestreo de dos etapas (unidades de tamaños desiguales), 380
en muestreo doble, 403
tamaño de muestra para, 105
Prueba preliminar, 27

— R —

Razón
de dos estimadores de razón, 232
separada, estimación de, 211
asignación óptima, 220
comparada con la estimación combinada, 214
confiabilidad con sesgo, 212
varianza, 212
varianza estimada, 214
Reconocimientos analíticos, 23 *ver también* Dominios de estudio
Reentrevistas en el estudio de errores de medición de revisitas, 465, 466

Regla de $\sqrt{f}$ cumulativa, 171
Regresión separada, estimación de, 252
comparada con la estimación combinada, 254
confiabilidad con sesgo, 253
varianza, 253
varianza estimada, 253
Relativa, precisión, 136
método de cálculo, 141
Replicaciones repetidas equilibradas, 390
Revisitas, 442
costo relativo, 443-444
efectos en razón de respuestas, 442
modelo matemático, 444
política óptima, 447-448
reentrevistas en el estudio de errores de medición, 465-466

— S —

Selección
a propósito, 32
controlada, 165
Sesgo
debido a errores en pesos de estratos, 157
de estimador de razón, 207
de estimador de regresión, 249
definición, 34
del entrevistador, 457
en pequeñas unidades de área, 289
por no respuesta, 436-438
uso deliberado de procedimientos sesgados, 34-37
Subpoblación, *ver* Dominios de estudio
Subunidades (elementos), 289
Suprapoblación, 204

— T —

Tamaño
de la muestra para límites especificados de error:
análisis del problema, 105
con atributos raros, 108
con datos continuos, 109
con más de una característica, 114
con proporciones, 107
en muestreo aleatorio estratificado, 143, 148
método de Cox para muestreo de dos etapas, 111-112
minimizando pérdidas, costo más pérdida por errores, 117
para aproximación normal a límites de confianza de proporciones, 69-72, 88
para aproximación normal a límites de confianza para datos continuos, 69-72
para comparaciones entre dominios, 117
para medias sobre dominios, 115
óptimo de la submuestra:
unidades primarias de mismo tamaño, 346
unidades primarias de tamaños diferentes, 382-386
Teoría, en encuestas por muestreo, 28

— U —

Unidad de muestreo, definición, 26
Unidades primarias de muestreo, 339
Usos de las encuestas por muestreo, 21-23
en censos de cada diez años, 21-22
en la investigación del mercado, 21
en sondeos de opinión, 23
para los negocios y la industria, 22
por países miembros de la O.N.U., 21

— V —

"Valor correcto", 455
Variación periódica, efecto en el muestreo sistemático, 271
Varianza
de estimaciones muestrales, *ver* Error estándar
definición de S^2 y σ^2, 46, 49
de población, estimaciones anticipadas, 110-114
de respuesta, 458
componente correlacionada, 463
simple, 463
total, 463
simple de respuesta, 463
total de respuesta, 463
Ventajas del muestreo, 19-20
Verificaciones de registros, 465

NOTAS