INTRODUCCION AL MUESTREO

Adela Abad
Luis A. Servín

SEGUNDA EDICION

LIMUSA

INTRODUCCION AL MUESTREO

Introducción al muestreo

SEGUNDA EDICION

Adela Abad
Luis A. Servín

NORIEGA Editores

EDITORIAL LIMUSA
MEXICO • ESPAÑA • VENEZUELA • ARGENTINA
COLOMBIA • PUERTO RICO

Primera edición: 1978
Primera reimpresión C: 1981
Segunda edición: 1982
Primera reimpresión: 1984
Segunda reimpresión: 1984
Tercera reimpresión: 1987
Impreso en México
(6422)

ISBN 968 — 18 — 1542 — 4

PROLOGO

Esta obra sobre muestreo probabilístico va dirigida a los estudiantes de licenciatura que no tienen bases rigurosas de matemáticas, probabilidad y estadística. Contiene un número reducido de desarrollos matemáticos y su enfoque es de aplicaciones sobre el tema, por lo que resulta accesible a los estudiantes de Ciencias Sociales y también es adecuada para cursos semestrales en Ingeniería, donde el interés no sea del tipo matemático, sino dirigido a su utilización como una herramienta más que la Estadística Matemática pone a la disposición del investigador en las diferentes áreas del conocimiento humano. Sin embargo, como ocurre usualmente, es necesario que el lector haya cursado Estadística Elemental y que esté familiarizado con el valor medio o promedio de una serie de observaciones, de la varianza como medida de dispersión y de la distribución normal.

Este libro ha surgido de varios cursos que hemos impartido a nivel universitario a actuarios, economistas, ingenieros, matemáticos, sociólogos y a licenciados en ciencias de la comunicación. Consideramos que esta versión es susceptible de mejoría y por ello aceptaremos cualquier comentario o crítica del lector y procuraremos utilizarla para mejorar su calidad.

En los dos primeros capítulos se presentan algunos conceptos del muestreo probabilístico y se repasan algunos otros de estadística. Después aparecen secuencialmente: el muestreo aleatorio simple, tamaño de la muestra, muestreo estratificado, muestreo por conglomerados, sistemático y el submuestreo. En cada capítulo hay ejemplos resueltos que ayudan al estudiante a comprender el tema y al final de cada uno aparece una lista de ejercicios cuyas respuestas numéricas están al final del texto.

Agradecemos la intervención del Dr. Ariel Kleiman, quien nos instó a recopilar el material disperso y nos proporcionó valiosas sugerencias; también damos las gracias por su gran colaboración a los doctores Edmundo Berumen, Gabriel Vera e Ignacio Méndez y al maestro Arturo G. García. La Act. Evangelina González trabajó en la obtención de las respuestas de una buena parte de los ejercicios y la Lic. Araceli Luévano participó en diferentes actividades durante el proceso de desarrollo de este trabajo. Finalmente agradecemos a la Editorial Limusa, por su eficiente colaboración.

Adela Abad
México, D.F. Luis A. Servín

CONTENIDO

CAPITULO 1

GENERALIDADES

1.1 MOTIVACION

Consideremos al conjunto de industrias de la transformación existentes en el país en la primera quincena del mes de febrero de 1977. Algunas de ellas son muy pequeñas en el sentido de que emplean a 1 o 2 personas, otras son de tamaño regular o mediano, algunas otras son grandes y otras son muy grandes; y físicamente se encuentran dispersas en todo el territorio nacional. Nos interesa determinar el número total de personas empleadas en ellas, y sabemos de la existencia de archivos de uso público en los cuales están registradas este tipo de empresas y también aparece registrado su personal. Sin embargo, no queremos hacer uso de ellos, ya que aunque consideramos que ahí están anotadas todas las empresas que nos interesan, no nos parece razonable el número de obreros y de empleados ahí registrados. Es decir, contamos con una lista de empresas junto con sus direcciones y deseamos determinar el número total de obreros y de empleados en ellas. Si en la lista existen 90 000 empresas, ¿cómo podemos determinar este número total?

Una manera consiste en acudir telefónica o personalmente a todas y a cada una de las empresas y preguntarles y registrar su número de obreros y de empleados, es decir, hacer un censo o enumeración completa sobre todos los miles de empresas en lista. Es fácil imaginar el enorme trabajo que requeriría el desarrollo de este censo a miles de empresas esparcidas en cientos de miles de kilómetros cuadrados.

Si todas las empresas tuvieran teléfono:

¿cuántas llamadas telefónicas se requerirían?
¿cuántos teléfonos y durante cuántas horas se podrían utilizar al día?

¿cuántas personas habría disponibles para hacer esas llamadas? en promedio, ¿cuántas llamadas con éxito se podrían hacer durante el día y por teléfono?

Si no todas las empresas tuvieran teléfono:

en promedio, ¿cuántas visitas con éxito se podrían hacer al día por persona?
¿cuántos viajes habría que hacer a los diferentes lugares de la nación?
¿de cuántas personas habría que disponer para terminar el trabajo en cuatro meses?
¿cuánto habría que pagar a 50 personas por concepto de viáticos si deben viajar por todo el país durante cuatro meses?
¿cuánto habría que pagar a 60 personas por concepto de sueldos, si trabajaran continuamente durante cinco meses?

Pudiéramos listar más preguntas interesantes que surgen cuando se piensa en un censo. Sin embargo, a nuestro juicio existe una de ellas que es de gran importancia: ¿es necesario determinar exactamente el número total de obreros y de empleados?

Usualmente no es necesario, basta con una aproximación, siempre y cuando el error que se cometa no exceda a aquel que el usuario de la información está dispuesto a aceptar de antemano.

Un recurso técnico que puede ayudar satisfactoriamente en la solución de este problema, y que está justificado matemáticamente, es el uso de las técnicas de muestreo probabilístico. Estas técnicas están estructuradas de tal manera que en muchas ocasiones logran los mismos propósitos que los de un censo,* y necesitan sustancialmente menos recursos humanos, materiales y financieros. Fundamentalmente, en base al conocimiento de una fracción de la población a estudiar se derivan o se infieren conclusiones para toda ella.

Si sólo se tiene conocimiento de lo que sucede en una fracción de la población, que puede ser, digamos, de un milésimo, de un centésimo o de un décimo de ella, ¿cómo es posible que se le pueda tener confianza a los resultados obtenidos mediante su uso? El control que se tiene sobre la muestra elegida y en general sobre todo el proceso para inferir conclusiones generales, es de tipo estadístico matemático. Se puede demostrar que si la fracción de la población que es revisada se elige con determinadas reglas de manera probabilística, es decir, que no sea la persona quien elija a la muestra, sino el azar, como en el caso de un dado que al lanzarlo no se sabe de antemano qué número saldrá, entonces es estadísticamente posible

* En relación a los parámetros que estiman.

hacer afirmaciones sobre la magnitud del error cometido. En un censo, por el contrario, esto no es posible, no se puede conocer la magnitud del error, si lo hubo.

En la práctica, es fácil que se cometan errores al desarrollar un censo. Esto es debido a los miles o a los cientos de miles de cuestionarios o formas que se deben manejar; a los cientos, o a los cientos de miles de personas que se involucran en el proceso y que muchas veces, por su gran número y por las restricciones de tipo práctico, es imposible darles el entrenamiento necesario; a los grandes problemas para coordinar a un proceso muy complicado; a pérdidas o a traspapeleo de los cuestionarios o formas entre muchos lugares de almacén temporal y a la dificultad para procesar grandes volúmenes de información. En cada paso o etapa que involucra un censo, está presente la posibilidad de que los errores se cometan sin que se conozca su magnitud.

En ocasiones, cuando se desarrolla un censo, después de meses de trabajo se decide suspender toda labor referente a entrevistas sin que estén terminadas satisfactoriamente. Y al analizar los resultados obtenidos se encuentra, con que sólo se desarrollaron debidamente el 60 por ciento de ellas. ¿Y el resto? , ¿cómo podemos considerar al trabajo realizado? , ¿como un censo? , ¿como una muestra?

No es un censo desde el momento en que no se entrevistaron o no se obtuvieron las observaciones de todas las unidades; en realidad es una *muestra*, y además es una muestra no probabilística por lo cual no se sabe la magnitud del error involucrado.

Volviendo al caso del listado con 90 000 empresas y estando conscientes de que permitimos un error, claramente sería más fácil entrevistar a 600 de ellas que a las 90 000. En el caso de una muestra como ésta, ciertamente que el costo total será más reducido, ya que se requiere visitar a menos empresas, se requieren menos viajes, se requieren menos cuestionarios, se puede proporcionar al personal un entrenamiento más adecuado por ser menor su número y el proceso en general es susceptible de mayor control. Esencialmente, para planear una encuesta es necesario conocer el error* que el usuario de la información está dispuesto a aceptar o la disponibilidad de dinero y de tiempo con que se cuenta para realizar el trabajo. Lógicamente, a mayor exactitud en los resultados deseados, mayor será el costo total de la escuesta, ya que para obtener mayor exactitud es necesario incrementar el tamaño de la muestra. Estaremos en el límite cuando deseáramos terminar con un error de cero, y esto, teóricamente equivalente a un censo**.

* El error en términos de precisión (ver el apartado 4.1).

** Teóricamente, porque usualmente existen problemas que impiden obtener el valor verdadero del parámetro. Ver el apartado 1.5.

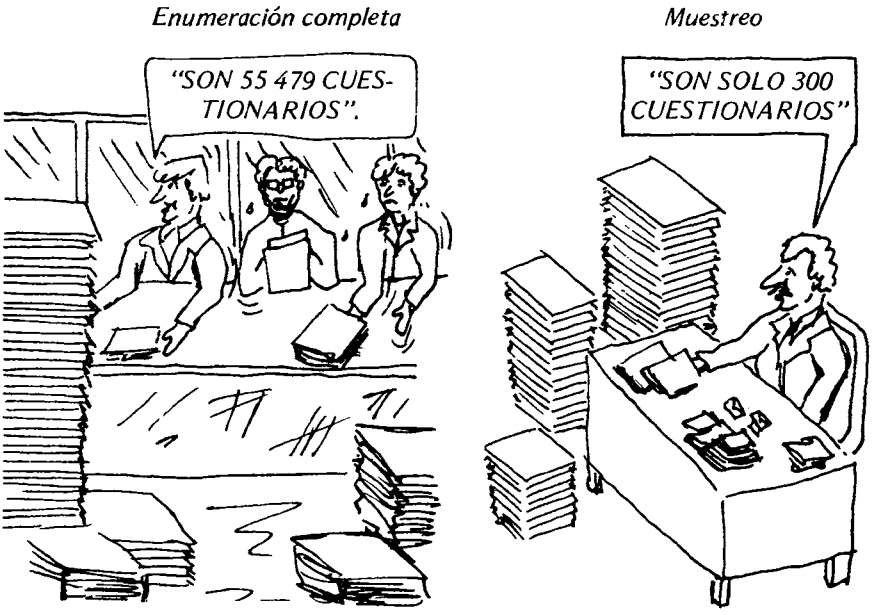


Figura 1.1

El censo o enumeración completa, es una técnica que permite averiguar o determinar el valor de parámetros que ocurren o que existen en un conjunto de elementos o unidades en consideración mediante una revisión a todos ellos; por ejemplo, en el caso de las industrias de la transformación es necesario determinar el valor del parámetro: número de obreros y de empleados en las industrias de transformación en una determinada región geográfica del país y en un tiempo o periodo de tiempo determinado. Otros ejemplos de parámetros son los siguientes: el número medio de miembros por familia en la ciudad de Guanajuato en enero de 1977, el valor total de la producción agrícola en un conjunto definido de parcelas y el estudio referido a un año en particular; el porcentaje de pólizas que no han sido pagadas en una compañía de seguros, y la relación entre el número de familias en un sector de la ciudad al número de familias en el mismo y que cuentan con seguro contra incendio.

Las técnicas de muestreo permiten *estimar* los mismos parámetros que aquellos en el caso de un censo, es decir, permiten obtenerlos aproximadamente a través de una muestra. Si esa muestra se obtiene de una manera que denominamos probabilística se le llama *muestra probabilística*, y al conjunto de esas técnicas se les denomina *técnicas de muestreo probabilístico*. De la misma manera, al conjunto de pasos que hay que seguir para llegar a una estimación

de algún parámetro poblacional se le denomina *encuesta* y ésta puede ser probabilística o no probabilística.

¿Por qué una encuesta? ; una encuesta se genera o se desarrolla con el fin de cumplir o de satisfacer un objetivo a nivel de muestreo, aquel objetivo que norma a toda la encuesta. En el caso de las industrias de la transformación éste consiste en obtener el número total de obreros y de empleados en ellas. Sin embargo, generalmente existe otro objetivo que es el del usuario de la información. Veamos el siguiente caso: un grupo de inversionistas desea establecer un centro de capacitación para los obreros y los empleados que trabajan en las industrias de transformación en determinada región del país. El grupo está seguro del futuro éxito de su empresa y para efectos de la construcción del centro de capacitación, necesita primordialmente conocer la demanda o número máximo de alumnos que podrá esperar. El grupo de inversionistas ha trabajado en este tipo de empresas durante varios años y sabe por experiencia que con sus políticas usuales de publicidad, el número máximo de alumnos que solicitarán entrenamiento durante los primeros seis semestres de funcionamiento del centro, será aproximadamente del 9 por ciento del total de obreros y de empleados en ese sector del aparato productivo del país. Así pues, se tiene la necesidad de llevar a cabo un estudio en la región para determinar ese total.

En este ejemplo se desea determinar el número total de obreros y de empleados (objetivo de la encuesta), para conocer la demanda futura al centro de capacitación en los tres años siguientes y por lo tanto, determinar la magnitud de la construcción para ese periodo de tiempo, éste es el objetivo del usuario de la información. En otras palabras, si se desarrolla una encuesta para determinar el número total de obreros y de empleados, es porque el uso futuro de este conocimiento ya está plenamente determinado. Una encuesta queda justificada mediante ambos objetivos, los cuales, pueden coincidir en algunos casos. El propósito de este libro es el de estudiar y el de comentar diferentes técnicas de muestreo probabilístico, y por ello sólo ocasionalmente se hará referencia a los objetivos propios del investigador o usuario de la información y a otras actividades como aquellas que son necesarias para localizar y entrevistar (*trabajo de campo*) a las unidades que pertenecen a la población y que fueron seleccionadas para la muestra (*unidades muestrales*), así como aquellas necesarias para preparar la información adecuadamente y sujetarla a un procesamiento.

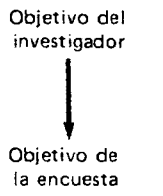


Figura 1.2 El objetivo del investigador o del usuario de la información determina la estructura de la encuesta, posiblemente no de manera única en el sentido de que las dos partes interactúan para llegar a un acuerdo, un consenso práctico que conviene a ambas partes y que toma en cuenta las limitaciones de tipo práctico.

La utilización de las técnicas de muestreo es muy amplia. Se les usa de una manera generalizada en agricultura y en ganadería; en todo tipo de industrias, en comercios y en servicios; es difícil concebir una actividad económica en un país, en la cual no se usen este tipo de técnicas. Y, naturalmente, ocurre algo similar en las diferentes áreas del conocimiento humano: actuaría, antropología, biología, contaduría, demografía, economía, ingeniería, medicina, mercadotecnia, oceanografía, psicología, publicidad, sociología, etc. Se les usa para averiguar el valor de un porcentaje, como el formado por las familias afectas a una marca de jabón; el valor de una media, como es el número medio de miembros por familia; el valor de un total; como es el valor total de la producción agrícola de trigo en un estado determinado; o para averiguar relaciones del tipo: número total de personas en una ciudad, entre el número total de personas que tienen seguro de vida en la misma ciudad. Y como se verá en el apartado 1.2, las *unidades o elementos* que conforman a la población sujeta a estudio pueden ser de naturaleza muy diversa, en lugar de hablar de personas se puede hablar de agrupaciones de ellas, se puede hablar de expedientes, de muebles, de edificios, de maquinaria industrial, de alimentos enlatados, de cajas con fruta fresca, de bodegas, de registros magnéticos, de animales o de agrupaciones de ellos, etc.

1.2 POBLACIONES

Repetidamente estaremos tratando con colecciones de objetos o de entes que se caracterizan por poseer ciertas propiedades específicas. Denominaremos conjuntos o más específicamente, *poblaciones* a esas colecciones o agrupaciones y diremos que cada una de ellas está formada por *elementos o unidades.*

De esta manera podemos tener, por ejemplo, una población de personas, una población de viviendas, una población de expedientes

de estudiantes, una población de animales, etc. Con ello queremos decir que el término población, en nuestro contexto, puede contener como unidades o elementos a personas, animales o cosas según sea el caso particular de que se trate. Las poblaciones que serán de nuestro interés están formadas por un número finito* de elementos, por ejemplo, las viviendas en la ciudad de Aguascalientes, los alumnos en la Universidad Nacional Autónoma de México o el ganado en la cuenca lechera mas grande del Estado de México, cada uno de ellos referido a una fecha específica.

En un estudio sociológico puede ser de interés estudiar atributos o actitudes de un conjunto dado de personas, digamos aquellas que en un momento en el año emigran de una región a otra con el afán de nuevas fuentes de trabajo, y de ellas se desea estudiar su estructura por sexo y edades, estado civil, estado de salud, deseos de emigración, satisfacción de necesidades, etc.

La población en estudio debe estar definida sin ambigüedad, de manera que no dé lugar a confusiones. Debe estar formada por algo que denominamos *elementos* o *unidades* (personas, hojas, reses, etc.) las cuales las consideramos contenidas en *unidades* (familias, expedientes, establos, etc.) y que se encuentran localizadas en determinado *lugar* o *región* geográfica y en un tiempo o *periodo* de tiempo dado.

También son de interés especial partes o fracciones de la población original, esto es, subconjuntos del conjunto original, a los que nos referiremos con los nombres de *subpoblaciones* o de *dominios de estudio*.

Así, en el caso de las viviendas en la ciudad de Aguascalientes, podemos definir a la subpoblación de viviendas alquiladas; o en el caso de los alumnos, podemos tener aquellos que ingresaron por primera vez el año pasado, y en el caso del ganado podemos definir la subpoblación formada por las reses que tienen un peso superior a 300 kilogramos, y así por el estilo.

1.3 CARACTERISTICAS DE LOS ELEMENTOS

Según nuestros propósitos la población *per se* no es de interés, sino que la estudiaremos por las *características*, propiedades o atributos que posea cada uno de sus elementos, y que nos interesan. Así se dice que se está desarrollando "una encuesta en familias, referente a ingresos y a tipo de alimentación". Aquí, las características a estudiar en cada familia son su ingreso y su tipo de

* Las técnicas de muestreo probabilístico se enuncian para poblaciones finitas, es decir, poblaciones que contienen o que están formadas por un número tal de elementos o de unidades que, de ser necesario, es posible contarlas y decir cuántas son.

alimentación, habiendo establecido previamente las definiciones adecuadas de ingreso y de tipo de alimentación por familia. En el ejemplo de los alumnos de la universidad, en un momento dado interesa su centro docente de procedencia, su edad, su estado civil, etc. En el caso del ganado nos puede interesar su rendimiento en litros de leche en un día específico, su edad y su tipo de alimentación.

Como ejemplos de otras características de interés, tenemos:

de una familia	núm. de miembros, zona de la ciudad en que vive, núm. de personas que trabajan, núm. de personas que estudian, grado máximo de escolaridad del jefe de la familia, etc.
de un predio agrícola	régimen de tenencia de la tierra, tipo de riego, superficie sembrada, producción por hectárea, etc.
de una industria	núm. de obreros y de empleados, producción anual, sector económico al que corresponde, núm. de profesionistas que emplea, etc.
de un asegurado de una institución de seguridad social.	estado civil, núm. de beneficiarios, grupo de salario de cotización, trayectoria que sigue para ir de su casa al trabajo, etc.
de un enfermo	sexo edad diagnóstico médico número de días de hospitalización tipo de seguro que posee
de un producto terminado	¿está defectuoso o no? núm. de fallas por unidad

de un estudiante universitario	estado civil, cantidad de dinero empleado en la compra de libros el semestre pasado. trabaja o no.

1.4 PARAMETROS POBLACIONALES

Supongamos que la población de interés está formada por las personas o miembros que conforman a cinco familias; denominémoslas por F_1, F_2, F_3, F_4 y F_5 y consideremos que éstas tienen 1, 5, 7, 2 y 5 miembros respectivamente. El número *total* de miembros en la población es la suma de ellos en cada una de las familias, por lo que la población contiene $1 + 5 + 7 + 2 + 5 = 20$ miembros en total. El número *medio* de miembros por familia es el total de ellos dividido entre el número de familias, que en este caso son cinco, obteniendo como resultado: $\frac{20}{5} = 4$ miembros por familia. Si consideramos adicionalmente el *porcentaje* de familias con más de dos miembros, intuitivamente diremos que ese porcentaje vale en este caso 60, es decir, el 60% de las familias tienen más de dos miembros.

Al total de miembros en la población, al número medio de miembros por familia y al porcentaje de familias con más de dos miembros, referidos a una fecha específica, se les denomina *parámetros poblacionales.* Y así como en el caso de la media la especificamos en términos de "miembros por familia", en otros ejemplos o en otras situaciones sus unidades de medida pueden ser diferentes, según sea la característica de interés y según la población sujeta a estudio, ya que ésta puede estar formada por otro tipo de unidades como se ilustra en el apartado 1.3.

Los parámetros poblacionales de uso más generalizado, es decir, aquellos que ocurren con más frecuencia en la práctica son cuatro: *totales, medias, proporciones* o *porcentajes** y los *cocientes* o *razones,* por ejemplo, valor total de la producción en pesos al peso total de la producción en toneladas y, por consiguiente, éstos serán los enfocados en este libro. Debemos notar que para obtener los valores de tales parámetros, es necesario revisar a todas y a cada una de las unidades o elementos en la población. Como ya vimos en el apartado 1.1, a este proceso de revisión exhaustiva se le denomina *censo* o *enumeración completa.*

* Para transformar una proporción en un porcentaje, la multiplicamos por 100.

1.5 METODO DE MEDICION

Hay algunas características que son fáciles, o relativamente fáciles, de identificar o *medir*, como en el caso del número de hijos por familia, el número de cuartos por vivienda, el número de personas que entran a una tienda en un intervalo de tiempo determinado, el número de vehículos que cruzan un puente por unidad de tiempo, el peso en kilogramos por saco de frijol; y otras características que definitivamente son difíciles de identificar o de medir, como son: ingreso por familia, el número de errores en cada documento expedido por alguna institución, el número de ocasiones en que una persona padeció gripe el año pasado, el número de abortos padecidos por persona, etc.

Esto significa que el criterio o *método* de *medición* empleado en una encuesta o en un censo particular, influye en menor o mayor grado en la determinación de un parámetro poblacional. Es claro que si somos más cuidadosos con el método de medición obtendremos un valor censal más parecido al verdadero en la población, pero conlleva un trabajo de campo mas largo, equipo y personal especializado y en general, un costo mayor. Y a la inversa, si consideramos un método de medición sencillo y expedito, tendremos rapidez en el trabajo de campo, no requeriremos entrevistadores especializados y nos costará menos; pero el error resultante será mayor que en el caso anterior.

Por ejemplo, si queremos saber si una persona está enferma o no, podemos confiar en su respuesta o someterla a una serie de preguntas, análisis y exámenes de laboratorio con el mismo fin. En el caso de una encuesta a científicos, podemos tomar como tales aquellos que nos indiquen los directivos de una institución, o seleccionarlos de acuerdo a una definición prestablecida de investigador. En ambos casos el criterio o método de medición es decisivo en nuestro estudio. Al realizar una encuesta debemos decidirnos por algún método de medición a usar y una vez que lo hemos seleccionado, es claro que solamente éste debe ser usado durante todo el trabajo de campo, ya que de esta manera habrá homogeneidad en las respuestas obtenidas y tendrá sentido el resultado alcanzado al final. Entonces, los instructivos de capacitación deben referirse al mismo concepto, la capacitación debe efectuarse sobre el mismo concepto, los entrevistadores no desvirtuarán a ese concepto durante la entrevista mediante aclaraciones individuales, subjetivas, y los resultados se enunciarán para la población muestreada y bajo el concepto y/o definiciones empleadas.

1.6 VARIABILIDAD

En una población sujeta a estudio la magnitud de la característica que es de interés, normalmente varía de unidad a unidad; así en el caso del ingreso familiar, una familia tiene 5 000 pesos al mes y la vecina de 7 000. Existen varias maneras de referirse a la *variación* de una característica y éstas se ilustran en seguida.

Supongamos que en una calle céntrica observamos a cada uno de los vehículos que pasan (automóviles particulares, taxis, autobuses, etc.) y contamos y registramos su número de ocupantes. Después de revisar decenas de vehículos encontramos que en el caso de los autos, su número de ocupantes varía con los días de la semana y con la hora del día, pero que tienen un intervalo de variación, digamos, entre 1 y 6; y que éste es menor o más pequeño que el intervalo de variación del número de pasajeros en autobuses, ya que aunque también éste presenta variaciones en el tiempo, algunos autobuses vienen casi vacíos, otros medianamente llenos y otros muy llenos. A esto nos referimos diciendo que es más variable el número de ocupantes en autobuses que en automóviles, es decir, que presenta una mayor variabilidad la característica "número de ocupantes" en autobuses que en autos; que presenta mayor dispersión, mayor *variancia*, o como también podemos decir, menor concentración. El término estadístico consagrado para este concepto es el de varianza y será ampliamente usado en todo el libro a lo largo de cada diseño de muestreo.

1.7 MARCO DE REFERENCIA MUESTRAL

Marco de referencia muestral o *marco muestral* es una manera o medio de representar e identificar a los elementos o unidades en la población. En el caso de los trabajadores de una fábrica, el marco puede estar formado por la nómina más reciente. Lo mismo puede ser válido para los empleados de alguna institución gubernamental. En algunas ocasiones a los elementos o unidades que conforman al marco les llamaremos elementos o unidades muestrales.

Existen casos en los cuales el marco queda representado por un conjunto de fotografías aéreas o de mapas, en los que se han identificado segmentos de área o manzanas de ciudades. A menudo, además de identificar a las unidades muestrales se les suele añadir algunas características de interés, como pueden ser medidas de su tamaño, es decir, atributos de ellas que nos permiten saber lo importante que es cada unidad para algún estudio específico. Así, por ejemplo, en un marco muestral de industrias, en adición al

nombre y dirección del establecimiento puede aparecer el número total de obreros y de empleados en una fecha dada y el valor de su producción anual en pesos o en millones de pesos.

Es deseable que el marco contenga a todas las unidades muestrales que son de nuestro interés, y que no incluya *unidades falsas*, o sea, elementos que son ajenos o que dejaron de pertenecer a la población. Por ejemplo, en el caso de la nómina de empleados, no queremos que aparezcan aquellos que han sido dados de baja. Tampoco queremos que contenga elementos repetidos, es decir, que aparezcan unidades muestrales registradas más de una vez. Y evidentemente desearemos que el marco sea legible. Las sorpresas derivadas al momento de la entrevista en las cuales se esperaba a un elemento muestral o una sola entrevista, y resultan ser más de una, muchas veces no pueden ser evitadas, pero hay que tenerlas presentes como una posibilidad y por lo tanto instrumentar medidas para afrontar cada eventualidad exitosamente.

Cuando la encuesta es pequeña, digamos sobre los 300 empleados y obreros en la nómina de una empresa en la ciudad de Morelia, y se refiere al sueldo que perciben, es relativamente fácil comprobar la presencia o ausencia de cada una de estas incongruencias o defectos; sin embargo, a medida que la encuesta se va haciendo más compleja, cuando el dinero y el tiempo no alcanzan, las incongruencias anteriores pueden presentarse, y lo realmente peligroso, es que el estadístico o el técnico no se percate de ellas a tiempo.

Ejemplo 1.1 Un grupo de economistas está interesado en realizar una encuesta sobre los empleados en las industrias de la fabricación de masa para tortillas en tres ciudades de México. El grupo define al número medio de empleados por establecimiento como la suma de los empleados en todos y cada uno de ellos, dividida entre el número de establecimientos.

Para desarrollar la encuesta se cuenta con un listado actualizado de los molinos en las tres ciudades que contiene la información siguiente:

MOLIENDA DE MAIZ (MOLINOS)

Nombre del establecimiento	*Dirección*
1. _____	_____
2. _____	_____
3. _____	_____
⋮	
1 300. _____	_____

En base a una muestra de 20 establecimientos elegida de alguna manera, se desea estimar el número medio de empleados por establecimiento, para el día 23 de junio de 1977.

En esta encuesta, la población sujeta a estudio está compuesta por los 1 300 molinos de masa existentes en el mes de junio de 1977 en las tres ciudades, los cuales están representados en el listado anterior.

Los elementos o unidades sujetos a estudio son cada uno de los molinos, y en el conjunto de ellos el parámetro poblacional de interés es el número medio de empleados por establecimiento. El objetivo de la encuesta es la estimación de ese parámetro, es decir, se trata de estimar una media poblacional. La característica que interesa de cada establecimiento es su número de empleados en el día 23 de junio de 1977, y el método de medición a emplear consiste en preguntar al encargado del molino por ellos y confiar plenamente en su respuesta. El marco de referencia para el estudio es el listado anterior, en el cual cada establecimiento en las ciudades está representado sin ambigüedad en algún renglón de la lista.

Ejemplo 1.2 Para el estudio de actitudes en un conjunto dado de estudiantes, se cuenta con un listado de aquéllos por grupo y por carrera que a determinada fecha se han inscrito como alumnos oficiales, siendo éstos en total 75 400. Sin embargo, por uno u otro motivo el estudio se pospone ocho meses después de lós cuales se continúa bajo las mismas condiciones iniciales y el mismo marco. Algunos problemas de marco que se presentarán son los siguientes:

i) No cobertura, ya que a la fecha de la encuesta habrá alumnos que antes no existían debido a actualizaciones tardías de los archivos escolares, inscripciones tardías y a la presencia de un nuevo semestre académico con nuevas inscripciones.

ii) Repetición de elementos provocada por el tipo de listado con que se cuenta, ya que éste es por grupo y por carrera. Pero ocurre con frecuencia que un estudiante está registrado en dos o más carreras además de los problemas de actualización de archivos derivados de los cambios de grupo que ocurren frecuentemente durante las primeras semanas de cada período académico.

iii) Presencia de elementos extraños los cuales corresponden a todas aquellas personas que en un momento dado se inscribieron, pero que, posteriormente se dieron de baja por conveniencia o por haber terminado sus estudios. Ellos, ya no son estudiantes al menos en esa escuela.

iv) Ambigüedad, posiblemente derivada a partir de homónimos.

Ejemplo 1.3 Al observar los salarios diarios de un conjunto de 10 personas se encontró lo siguiente: 175, 300, 125, 100, 100, 275, 150, 150, 200, 200.

i) El salario mínimo observado fue de 100 pesos al día, el máximo de 300, la diferencia entre el máximo y el mínimo o rango de las observaciones es de $300 - 100 = 200$.

ii) El salario medio es de: $175 + 300 + 125 + 100 + 100 + 275 + 150 + 150 + 200 + 200$ dividido entre 10, es decir, $\frac{1775}{10} = 177.5$ pesos al día.

iii) La varianza de las observaciones es de (Apartado 3.6):

$$S^2 = \frac{1}{N-1}\left(\Sigma y_i^2 - \frac{(\Sigma y_i)^2}{N}\right)$$

$$= \frac{1}{10-1}\left(175^2 + 300^2 + 125^2 + 100^2 + 100^2 + 275^2 + 150^2 + 150^2 + 200^2 + 200^2 - \frac{(1775)^2}{10}\right)$$

$$= \frac{1}{9}(356\,875 - 315\,062.5) = 4\,645.83 \text{ pesos al cuadrado.}$$

1.8 EJERCICIOS

1.1. *i.* Describa dos ejemplos de poblaciones a estudiar. *ii.* Especifique sus unidades o elementos. *iii.* Enuncie dos ejemplos de características de interés en cada una de ellas. *iv.* ¿Cuántos elementos tiene cada una de sus poblaciones? *v.* ¿Puede definir algunas subpoblaciones en ellas? , ¿cuáles?

1.2. El número de miembros y de beneficiarios asociados a un organismo público en cada una de las 20 familias de una manzana fueron los de la tabla 1.1.

Tabla 1.1

Familia	*No. de miembros*	*No. de beneficiarios*	*Familia*	*No. de miembros*	*No. de beneficiarios*
1	2	0	11	11	5
2	5	5	12	6	0
3	3	1	13	6	0
4	6	0	14	3	1
5	9	0	15	7	0
6	7	0	16	6	0
7	5	0	17	6	0
8	5	1	18	4	0
9	6	3	19	9	2
10	4	0	20	8	0

Determine:

i. El número total de miembros en las 20 familias.

ii. El número medio de miembros por familia.

iii. El porcentaje de familias con al menos un beneficiario.

iv. El cociente del total de beneficiarios al total de habitantes en la manzana.

v. El número medio de beneficiarios por familia, para las familias con al menos un beneficiario.

1.3 Enuncie una situación en la cual un censo fuera preferible a un muestreo.

1.4 En una encuesta sobre la salud de las personas en una ciudad, se parte de un mapa de ella, se dibujan las manzanas, se numeran, se efectúa una selección de manzanas, y para aquellas manzanas seleccionadas se contruye un listado de las viviendas que las conforman. Así se tiene un listado de viviendas para cada manzana en la muestra. A partir de esos listados se hace una nueva selección ahora de viviendas dentro de manzana, y para cada vivienda seleccionada, el entrevistador ocurre a ella y llena un cuestionario por cada persona o miembro de la vivienda. I) ¿Cuáles son los elementos en esta encuesta?, II) ¿Cuáles son las unidades?, III) Proporcione seis ejemplos de características a estudiar. IV) En base a sus características elegidas en (III), defina cuatro parámetros poblacionales, V) ¿Qué método de medición emplearía para cada uno de ellos?, VI) ¿Cuál fué el marco de referencia?.

CAPITULO 2

ALGUNOS CONCEPTOS DE ESTADISTICA Y DE MUESTREO

2.1 VARIABLE ALEATORIA Y DISTRIBUCION DE PROBABILIDADES

En muestreo probabilístico se elige a algunas unidades de la población de tal manera que aquellas seleccionadas lo son porque el azar así lo quiso. Por ejemplo, supongamos que tenemos a la familia número 1 y a la familia número 2, y que deseamos seleccionar a una de ellas de manera que ambas tengan la misma oportunidad de ser elegidas. Tomamos una moneda y convenimos en que si al lanzarla aparece cara fue elegida la familia número 1, y si aparece águila fue elegida la familia número 2. Lanzamos la moneda y aparece una águila, entonces decimos que la familia número 2 fue elegida *aleatoriamente* o que la eligió *el azar.*

El dispositivo que estamos usando para elegir a una familia es una moneda. Antes de lanzarla no sabemos cuál será su resultado; una vez que la lanzamos queda materializado uno de ellos, por ejemplo, aparece águila y decimos que el azar la materializó. En estas condiciones el resultado que ofrecerá la moneda es variable y desconocido. ¿Caerá *cara* o *águila*?, ¿*1* o *2*? Si el resultado del proceso de lanzarla lo simbolizamos por la variable X, los valores posibles de esta variable son dos: *cara* o *águila*, *1* o *2*. Antes de lanzarla no sabemos cuál será su resultado, no sabemos qué valor tomará X; por ello a X se le denomina *variable aleatoria.* Y decimos que la cara y el águila tienen la misma oportunidad de aparecer, y tienen la misma *probabilidad* de aparición; como resultado del lanzamiento esperamos por igual a una y a la otra. La probabilidad de que aparezca águila es la probabilidad de que X tome el valor 2 y ésta es igual a 1/2. Y de la misma manera, para el resultado cara o X igual a 1.

Cuando se conoce la probabilidad de que una variable aleatoria tome cada uno de sus valores, se dice que se conoce la frecuencia de aparición de cada uno de ellos (1/2 o el 50 por ciento de las veces para el águila en el caso de la moneda), o también se dice que se conoce la distribución de probabilidades de la variable aleatoria en consideración.*

Si en lugar de una moneda usamos un dado los resultados posibles son 6. En este caso los valores que puede tomar la variable aleatoria X son: *1, 2, 3, 4, 5* o *6*. Y nuevamente cada resultado del dado tiene la misma oportunidad o probabilidad de aparecer y ésta vale 1/6; como resultado de su lanzamiento esperamos por igual a cualquiera de las seis caras. También en este caso conocemos la frecuencia de aparición de cada valor o la distribución de probabilidades de la variable aleatoria.

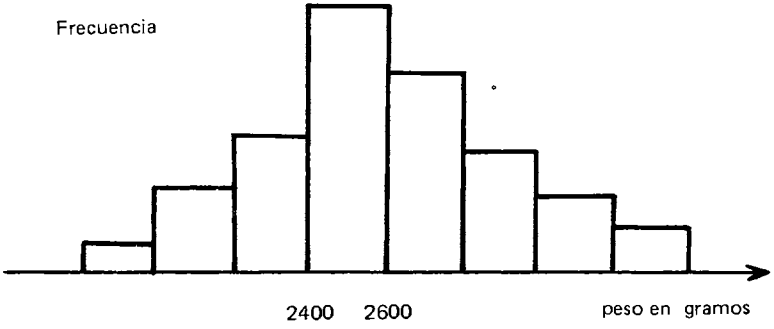


Figura 2.1. Una manera de representar la distribución de probabilidades de una variable aleatoria es mediante este diagrama denominado Histograma de frecuencias.

En la figura 2.1 aparece una manera de representación experimental de la distribución de probabilidades de la variable aleatoria X: peso de recién nacidos en un caso hipotético. El peso de recién nacidos es una variable aleatoria porque hasta que ocurre el nacimiento se puede conocer el valor de X. En ella el rectángulo mayor se puede obtener contando el número de nacimientos tales que su peso se encuentra entre 2 400 y 2 600 gramos y dividiéndolo entre el total de nacimientos en consideración; con esta fracción se dibuja un rectángulo de área equivalente a ella.

Con la moneda podemos generar dos *números aleatorios* diferentes y con el dado podemos generar a seis de ellos. Como

* Formalmente existe diferencia entre los conceptos de frecuencia y de probabilidad, pero para el alcance de este libro los consideramos como equivalentes.

veremos posteriormente, existen diversas maneras de generar una gran cantidad de números aleatorios diferentes (apartado 5.7).

Supongamos que al dado le cambiamos la numeración de sus caras, poniendo el número 1 una vez, el número 2 una vez y el número 3 cuatro veces. Al lanzarlo, el número 1 aparece con probabilidad de 1/6, el número 2 con probabilidad de 1/6 y el número 3 tiene en cada tirada cuatro oportunidades de aparecer, por lo que su probabilidad de aparición es de 4/6. Por ello en cada lanzamiento el 3 tiene más oportunidad de aparecer que el 1 y que el 2. Es razonable esperar que en un lanzamiento concreto aparezca un 3, aunque evidentemente puede darse el caso de que aparezca alguno de los otros números. Así surge el concepto de valor esperado o de esperanza matemática de una variable aleatoria; éste nos indica el resultado que bajo cierta definición (apartado 2.2) esperaríamos que ocurriera en atención a la probabilidad que tiene cada resultado posible de aparecer.

En muestreo probabilístico se elige a algunas unidades de la población de manera aleatoria o aleatoriamente, y en base a ellas derivamos o inferimos una conclusión válida para toda la población a través de una función de las observaciones (valores de atributos de las unidades elegidas en la muestra) llamada estimador, la cual por ser una función de variables aleatorias, también es una variable aleatoria. Para muestras diferentes y en igualdad de condiciones experimentales, generalmente tomará valores diferentes debido a la presencia del azar. Un estimador específico tiene una distribución de probabilidades y como veremos más adelante, para muestras diferentes tiende a tomar muchos valores relativamente próximos en magnitud a su valor esperado, tanto superiores como inferiores a él y menos valores relativamente lejanos al mismo. Su distribución de probabilidades tiende a parecerse a la silueta de una campana y con algunas condiciones y supuestos se le da el nombre de distribución normal. Ella es simétrica respecto al valor esperado del estimador (apartado 2.2). Por ello, dada una muestra, no sólo interesa el valor particular del estimador, sino también algo que sugiera la forma de la distribución, esto es, ¿qué tan achatada es la silueta? Una manera de indicar lo rápido que cae la curva (silueta de la campana) es la *desviación estándar* o *error estándar* del estimador. Su definición formal se hace a través del concepto de variancia (apartado 2.2) del cual es su raíz cuadrada. Existe otra manera de referirse a la forma de la distribución y ésta es en términos de *intervalos de confianza*. Se calcula un intervalo que aunque varía de muestra a muestra nos indica el porcentaje de veces que bajo repeticiones del experimento, ese tipo de intervalos incluirán a un valor que está en consideración, y en ocasiones este

valor coincide con el parámetro poblacional que está en estudio; por ejemplo el peso medio de los recién nacidos en el hospital A en el transcurso de un año determinado.

En realidad, los conceptos de desviación estándar de un estimador e intervalo de confianza, son maneras de referirse o de indicar lo adecuado, lo bueno o lo preciso de un estimador; un intervalo de confianza es un conjunto de valores de tal manera calculados, que uno tiene cierta confianza o cierta seguridad, de que dentro de ellos, dentro del intervalo, se encuentre el parámetro poblacional en estudio.

2.2 ESPERANZA MATEMATICA Y VARIANCIA

Sea la variable aleatoria X, la cual puede tomar los valores x_1, $x_2, \ldots, x_n$, con probabilidades $p_1, p_2, \ldots p_n$ respectivamente.*

La expresión:

$$E(X) = \sum_{i=1}^{i=n} x_i p_i \qquad 2.1$$

es la esperanza matemática o valor medio o simplemente *media* de la variable aleatoria X. De acuerdo con la fórmula, para obtenerla debemos multiplicar cada valor que toma la variable aleatoria por la probabilidad de que lo tome y sumar todos los productos. Por ejemplo, en el caso de la moneda, la variable aleatoria X toma los valores 1 y 2, cada uno de ellos son probabilidad de $\frac{1}{2}$; entonces su esperanza se calcula de la manera siguiente:

$$E(X) = 1(\tfrac{1}{2}) + 2(\tfrac{1}{2}) = \frac{3}{2}$$

Se puede demostrar que:

$$E(a) = a, \qquad a \text{ constante.}$$

$$E(a + bX) = a + bE(X), \; a \text{ y } b \text{ constantes.} \qquad 2.2$$

Es decir, la esperanza de una constante es ella misma, y la esperanza del producto de una constante por una variable aleatoria es igual a la constante multiplicada por la esperanza de la variable aleatoria.

* Para las definiciones en este apartado existen las análogas en el caso continuo.

Además se verifica también que:

$$E(X + Y) = E(X) + E(Y) \qquad 2.3$$

en la que X y Y son variables aleatorias. La expresión:

$$V(X) = E((X - E(X))^2) = \sum_{i=1}^{i=n} (x_i - E(X))^2 (p_i) \qquad 2.4$$

se llama *variancia* de la variable aleatoria X (apartado 1.6) y nos indica la variabilidad o la dispersión de la característica en cuestión. Una variable aleatoria que tenga una variancia relativamente pequeña se dice que está poco dispersa o bastante concentrada. Otra forma de calcular la variancia es según la expresión siguiente:

$$V(X) = E(X^2) - (E(X))^2 \qquad 2.5$$

A la raíz cuadrada positiva de la variancia de X se le llama desviación estándar y tiene como unidad de medida la misma que la de la variable aleatoria X.

Calculemos la variancia de la variable aleatoria X en el caso de la moneda:

$$V(X) = (1 - \frac{3}{2})^2 (\frac{1}{2}) + (2 - \frac{3}{2})^2 (\frac{1}{2}) = \frac{1}{4}$$

y su desviación estándar es $(\frac{1}{4})^{1/2} = \frac{1}{2}$

Se puede demostrar que:

$$\left.\begin{aligned} &V(a) = 0 \\ &V(a + bX) = b^2 (V(X)), \end{aligned}\right\} \quad 2.6$$

en las que a y b son constantes. La variancia de una constante es cero y la variancia del producto de una constante por una variable aleatoria es igual a la constante elevada al cuadrado y multiplicada por la variancia de la variable aleatoria.

Si X y Y son dos variables aleatorias, la variancia de su suma vale:

$$V(X + Y) = V(X) + 2COV(X, Y) + V(Y) \qquad 2.7$$

en donde COV (X, Y), representa la covariancia entre las variables aleatorias. Esta vale cero si X y Y son independientes, y se calcula según la expresión siguiente:

$$\text{COV}(X, Y) = \text{E}((X - \text{E}(X))(Y - \text{E}(Y))) \qquad 2.8$$

Se puede demostrar que también es igual a:

$$\text{COV}(X, Y) = \text{E}(XY) - \text{E}(X)\text{E}(Y) \qquad 2.9$$

Además:

$$\text{V}(X_1 + X_2 + \ldots + X_n) = \text{V}\left(\sum_{i=1}^{i=n} X_i\right)$$

$$= \sum_{i=1}^{i=n} \text{V}(X_i) + \sum_{i \neq j} \text{COV}(X_i, X_j) \qquad 2.10$$

en la que $X_1, X_2, \ldots, X_n$ son n variables aleatorias.

2.3 DISTRIBUCION NORMAL

Si en el puente sobre una carretera observamos y registramos el número de vehículos que cruzan por unidad de tiempo, se encuentra que la característica en estudio "número de vehículos por unidad de tiempo" tiene una distribución cuya forma pudiera ser la siguiente:

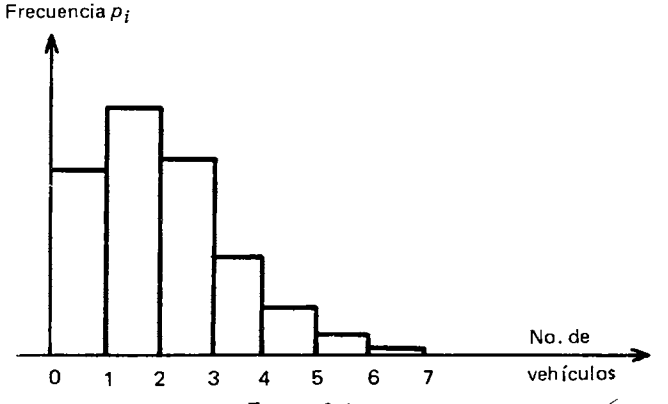


Figura 2.2

Por otra parte, al estimar el número medio de vehículos por unidad de tiempo, se encuentra que para muestras repetidas en las mismas condiciones, el estimador de ese número medio tiende a distribuirse según una distribución normal.*

Cuando se desea estimar algún parámetro poblacional siempre existen dos distribuciones en consideración, las cuales no tienen por qué ser iguales: la propia de la característica que se estudia en la población y la del estimador. Esta última tiende a distribuirse como una normal.

La *distribución normal* tiene una gran importancia en el estudio del muestreo probabilístico debido a que generalmente se supone normalidad en la distribución de los estimadores. Este supuesto es razonable a la luz de los resultados experimentales y es apoyado formalmente por el "Teorema del Límite Central". Es conocido que una distribución normal queda caraçterizada por el conocimiento de sus dos parámetros: media y variancia. Al tomar una desviación estándar a la derecha e izquierda de su media se encierra un área bajo la curva que comprende al 68% del total bajo ella. Al tomar dos desviaciones el área es del 95% y al tomar tres aumenta al 99.9%. Por lo cual, en aquellos casos en que el supuesto de normalidad es razonable y para efectos del cálculo de intervalos de confianza, se usa frecuentemente un nivel de confianza del 95%, es decir, una abscisa de 1.96 (que a veces es aproximada por un 2)./

2.4 ESTIMADORES

Las técnicas de muestreo probabilístico requieren que aquellas unidades o elementos que se elijan para la muestra, lo sean de acuerdo a procedimientos determinados; un procedimiento de éstos en particular se denomina procedimiento o *método de selección.* Una vez que la muestra ha sido seleccionada, el mecanismo para inferir o derivar una conclusión de ella a la población consiste en el llamado *método de estimación* y éste es alguna función de los valores muestrales, a la que se le denomina *estimador.* Cuando se indican los dos métodos, el de selección y el de estimación, se dice que se tiene el *diseño muestral* o diseño de muestreo correspondiente.

* Con media, la de la población y cierta variancia.

Existen muchos estimadores que pueden ser usados. Claramente debemos utilizar al mejor estimador para el problema que esté en consideración; si no sabemos cuál es el mejor, como sucede a menudo, usamos alguno que cumpla con propiedades deseables.

En los libros sobre Estadística se enuncian varias propiedades deseables de los estimadores. En este libro y en este apartado nos referiremos únicamente a dos de ellas, pero no debemos interpretarlo como si sólo nos interesaran ésas; en realidad algunos de los estimadores que se propondrán son ricos en propiedades.*Usemos la siguiente notación: un acento cincunflejo $\hat{\theta}$ sobre θ significa que se trata de un estimador, y el símbolo bajo el acento, el parámetro que se desea estimar. Entonces $\hat{\theta}$ se lee "el estimador de θ".

Sea θ un parámetro poblacional, se dice que $\hat{\theta}$ es un estimador *insesgado* de él si $E(\hat{\theta})$ es igual a θ; en caso contrario se dice que θ es sesgado y que su sesgo es igual a $E(\hat{\theta}) - \theta$ **Usualmente los estimadores son funciones del número de unidades n en la muestra. Se dice que $\hat{\theta}$ es *consistente* cuando al remplazar el tamaño n de la muestra por el tamaño N de la población se reproduce el parámetro poblacional en consideración.

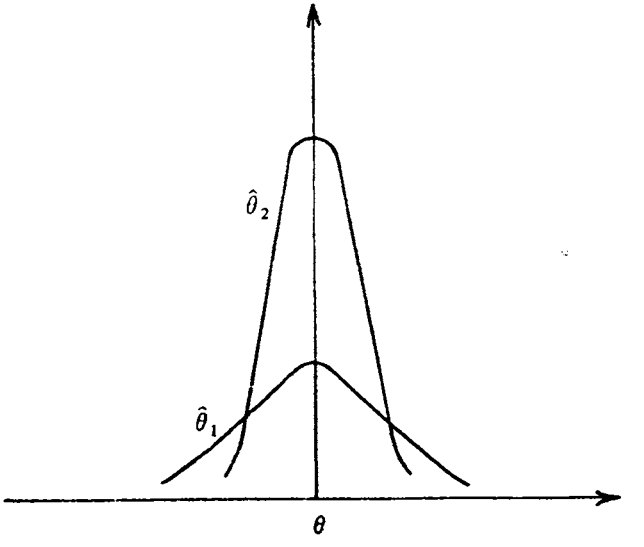


Figura 2.3

* Considerando una función de densidad asociada al estimador.

** Cuando un estimador es sesgado del parámetro poblacional en consideración es preferible hablar de su error cuadrático medio en lugar de su varíancia. (Cochran W. G. 1963. *Sampling Techniques* Apartado 1.8. J. Wiley & Sons. N. Y. Segunda edición).

La concentración o dispersión de una distribución será analizada a través de la variancia o de su raíz cuadrada, la desviación estándar. A menor variancia, menor dispersión, o sea mayor concentración; y a mayor variancia, mayor dispersión o sea menor concentración.

Supongamos que tenemos dos estimadores insesgados y consistentes $\hat{\theta}_1$ y $\hat{\theta}_2$ del parámetro poblacional θ y que éstos tienen variancias diferentes tal como se observa en la figura 2.3. $\hat{\theta}_2$ está más concentrado alrededor del valor verdadero θ, significando que hay gran probabilidad de que sea pequeña la diferencia entre la estimación $\hat{\theta}_2$ $(X_1, X_2, \ldots, X_n)$ y el valor verdadero θ; en cambio, es probable que las diferencias respectivas para $\hat{\theta}_1$ sean mayores, por lo que preferimos a $\hat{\theta}_2$, como estimador de θ.

En este libro se verán diferentes métodos de selección y de estimación. Muchas veces la elección de estos procedimientos se hace en base a la variancia muestral el estimador o a su error cuadrático medio.

2.5 POBLACION A ESTUDIAR Y POBLACION MUESTREADA

La *población a estudiar* es aquella sobre la que se desea efectuar inferencias y queda definida antes de iniciar el trabajo de campo. Generalmente ésta va sufriendo transformaciones a medida que se avanza hacia y sobre el trabajo de campo. Muchas veces esto hace necesario redefinir la población a estudiar, de manera que se tenga una población que sea alcanzable en términos prácticos.

Sin embargo, aunque ocurran redefiniciones es usual que esas poblaciones discrepen a la hora del trabajo de campo, y es necesario agregar las aclaraciones pertinentes cuando se emiten los resultados de la encuesta y sus conclusiones. Esos resultados y conclusiones sólo serán válidos para la *población muestreada;* en este sentido, el trabajo de campo se dirige a hacer coincidir las dos poblaciones.

Como ejemplo, considérese una encuesta sobre las industrias cuyo único giro es la fabricación de ropa en el Estado de Aguascalientes. Este tipo de industrias están registradas en diferentes organismos gubernamentales y en las cámaras industriales. Sin embargo, estos listados no son completos. Muchas de ellas, principalmente las de tipo familiar, no aparecen en los listados y son difíciles de localizar, por lo que es necesario redefinir la población,

por ejemplo, a aquellas industrias que aparecen registradas en el organismo encargado de la seguridad social. Sin embargo, al desarrollar la encuesta puede ocurrir que las industrias entrevistadas tengan más de un giro, por lo que no coinciden la población a estudiar y la muestreada. Esto, en algunos casos, puede ser de tal naturaleza que los resultados obtenidos no sean satisfactorios para el propósito inicial del estudio, o lo sean parcialmente.

2.6 FINALIDAD DE LA TEORIA DEL MUESTREO PROBABILISTICO

La finalidad del muestreo es proporcionar diseños muestrales, esto es, métodos de selección y de estimación, que arrojen los mejores resultados (usualmente mínima variancia) al menor costo posible. Para un costo dado podemos utilizar diferentes diseños muestrales, algunos de los cuales pueden diferir únicamente en la manera de efectuar la estimación, pero podrán compararse entre sí en términos de la variancia de su distribución en el muestreo o de su error cuadrático medio.

Ejemplo 2.1 En el capítulo anterior se presentó en el ejemplo 1.1 una encuesta referente a los empleados en los molinos productores de masa para tortillas. En él se identificó a la población sujeta a estudio, a los elementos o unidades que la componen, a la característica a observar en cada unidad, al método de medición a usar y al marco de referencia.

El grupo de economistas materializó de alguna manera una muestra consistente de veinte molinos, a los cuales identificó y marcó en el listado. Se fue a visitar a cada molino en la muestra y se les hizo la pregunta correspondiente a los encargados de cada establecimiento. Sin embargo, los entrevistadores encontraron una dificultad. Los establecimientos visitados, además de hacer la molienda del maíz y así producir la masa, se dedicaban a la fabricación de tortilla en el mismo local y, los empleados atendían indistintamente al molino y a la fabricación de tortilla por lo que no se podía hacer la distinción entre el número de empleados dedicados exclusivamente a la actividad en cuestión.

Los entrevistadores consultaron al grupo de economistas y éstos decidieron que la pregunta se hiciera sobre todos los empleados del establecimiento. Los datos obtenidos fueron los siguientes:

Molino	1	2	3	4	5	6	7	8	9	10
No. de empleados	2	2	2	3	1	2	1	1	1	1
Molino	11	12	13	14	15	16	17	18	19	20
No. de empleados	2	2	3	1	4	2	1	3	1	2

Para el procesamiento de la información se dieron las instrucciones siguientes: sume los empleados declarados en cada establecimiento en la muestra y divida la suma total entre 20; éste es el estimador usado. Posteriormente se pudo comprobar que el nombre del listado estaba equivocado, porque sus componentes eran molinos tortillerías.

Y así, como ya estaba desarrollado todo el trabajo, decidieron usar los resultados con la indicación de que la encuesta era válida para aquellos establecimientos en el listado los cuales hacían tanto masa como tortillas. De manera que la media estimada fue de 37/20 empleados por establecimiento "molino tortillería".

Es claro que la población inicial sujeta a estudio difiere de la población muestreada.

Ejemplo 2.2 Supongamos una población con 12 000 familias, de las cuales, 5 000 tienen ingresos de 200 pesos diarios, 3 000 de 300, 2 000 de 400, 1 000 de 500, 500 de 600 y 500 de 700 pesos. Su distribución es del tipo de la figura 2.4.

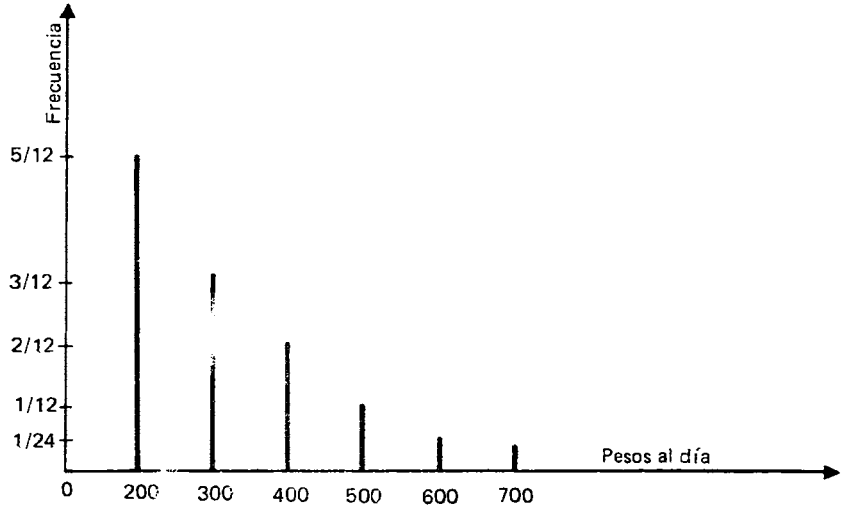


Figura 2.4

De esa población sacamos 50 muestras de tamaño 100 y estimamos a su ingreso medio mediante la media muestral. Las estimaciones obtenidas fueron tales que su histograma es el de la figura 2.5.* ¿Qué interpretación puede hacer del experimento y de la figura 2.5?

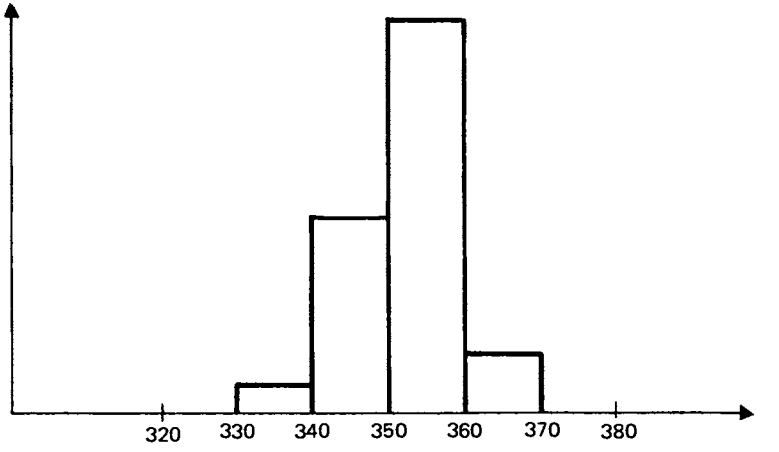


Figura 2.5

En otra etapa del experimento, el número de muestras se aumentó a 200, y con un tamaño de muestra de 250, obteniéndose como resultado la figura 2.6. Este nuevo resultado, ¿modifica o refuerza en algo a sus conclusiones anteriores?

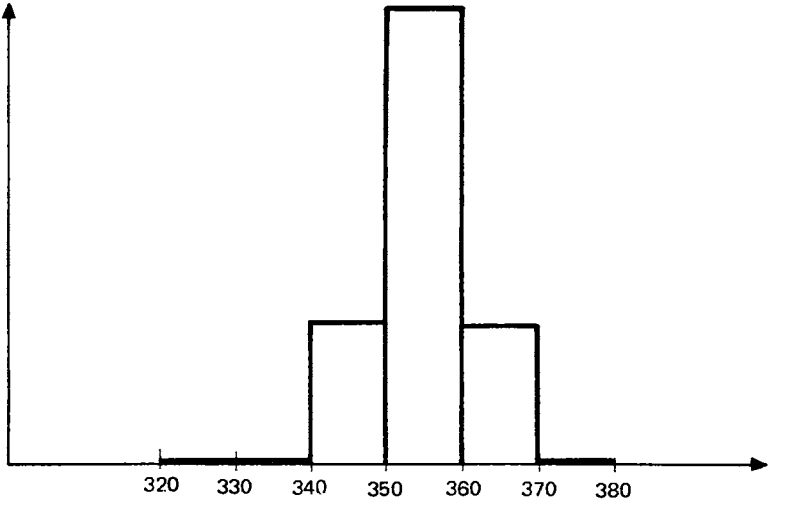


Figura 2.6

* El procedimiento empleado fue una simulación en una computadora digital.

2.7 EJERCICIOS

2.1 Véase el ejercicio *2* del capítulo *1*. Marcamos una canica por familia con el número de ésta y de sus miembros; las ponemos en una urna, las mezclamos y elegimos aleatoriamente a cinco de ellas, obteniendo los resultados siguientes:

Canica	1	2	3	4	5
No. de miembros	5	5	3	6	7

i) ¿Cómo estimaría el número medio de miembros por familia?

ii) ¿Qué estimador utilizó?

iii) ¿Cuál es su valor particular para esta muestra?

iv) ¿La muestra fue buena?

2.2 En el ejercicio anterior, y con el propósito de evaluar los resultados del muestreo, es decir, de saber qué tan buena es nuestra estimación, usamos la ecuación siguiente:

$$\left(1 - \frac{5}{20}\right)\left(\frac{1}{5}\right)\frac{\sum_{i=1}^{i=5}(y_i - \bar{y})^2}{5-1} = \left(1 - \frac{5}{20}\right)\left(\frac{1}{5}\right)\frac{1}{5-1}\left(\sum_{i=1}^{5} y_i^2 - \frac{\left(\sum_{i=1}^{5} y_i\right)^2}{5}\right)$$

llamada "estimador de la variancia";

i) ¿Cuál es su valor particular para la muestra elegida?

ii) ¿En qué unidades de medida está el resultado obtenido?

iii) A la raíz cuadrada de ella se le llama *error estándar*, ¿cuál es su valor en este caso?

iv) Haga un bosquejo de una distribución normal, con media igual al número medio de miembros por familia encontrado en el ejercicio 1.2 y con la desviación estándar calculada anteriormente.

v) Según su dibujo, ¿le parece que está muy dispersa esa distribución?, es decir, ¿su variancia es grande o pequeña?

2.3 En una encuesta de opinión desarrollada sobre los obreros de una fábrica, se encontró el porcentaje de obreros favorables a cierta regla. Este porcentaje fué de 38%. Como había información suficiente para obtener dos estimaciones, se hizo esto y se obtuvo, para la segunda, nuevamente 38%, aunque no así para sus errores estándar. En el primer caso se encontró 0.07 y en el segundo 0.05. ¿Cuál de las dos estimaciones es mejor?, dé sus razones.

CAPITULO 3

MUESTREO ALEATORIO SIMPLE

3.1 MUESTREO ALEATORIO SIMPLE

Extraer una *muestra aleatoria simple* consistente de n unidades elegidas de entre N de que consta la población, es extraerlas de manera que a todas y a cada una de las $\binom{N}{n}$ muestras posibles se les asigne y respete una probabilidad igual de ser elegidas; a cada unidad en la población se le asigna una probabilidad conocida e igual a $\frac{n}{N}$ de aparecer en la muestra. Así, en el ejemplo *1.1*, el tamaño de la población es de $N = 1\,300$ molinos y el tamaño de la muestra es de $n = 20$; el número de muestras diferentes es de:

$$\binom{1\,300}{20} = \frac{1\,300!}{20!\ (1\,300 - 20)!} = \frac{1\,300!}{20!\ 1\,280!}$$

$$= \frac{(1\,300)\ (1\,299)\ (1\,298)\ \ldots\ (1\,282)\ (1\,281)}{(20)\ (19)\ (18)\ \ldots\ (2)\ (1)}$$

y la probabilidad de que cualquier unidad aparezca en la muestra es de $\frac{20}{1\,300}$.

La definición anterior de *muestreo aleatorio simple* está en términos de muestras; su uso práctico requiere que previamente se les liste, lo que resulta casi imposible en la mayoría de los casos. Se puede probar (ejercicio 3.6) que seleccionar una muestra bajo estas condiciones, es equivalente a elegir n números aleatorios diferentes que estén comprendidos entre *1* y N; los números así elegidos forman la muestra.

El que las n unidades muestrales seleccionadas se elijan dife-

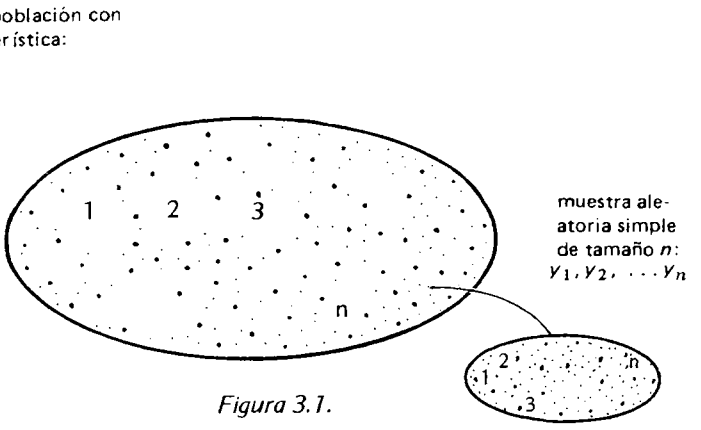


Figura 3.1.

rentes, califica a este esquema de muestreo como *sin reemplazo* ya que cada unidad elegida previamente no tiene oportunidad de ser seleccionada nuevamente en la muestra.

3.2 NOTACION

La población sujeta a estudio está formada por N unidades o elementos, los cuales poseen una característica específica de interés para nosotros, por ejemplo: el número de empleados en cada establecimiento, el número de alumnos en cada grupo escolar o el sexo de cada persona. Denotaremos el valor particular de esta característica en el primer elemento por y_1, en el segundo por y_2 y en el N-ésimo por y_N. Así, si en el primer establecimiento existen 4 empleados, en el segundo 2 y en el N-ésimo 15, diremos que el valor de la característica en el primer establecimiento es $y_1 = 4$, en el segundo $y_2 = 2$ y en el N-ésimo $y_N = 15$.

Entonces la *media poblacional* (parámetro poblacional) de esa característica será:

$$\bar{Y} = \frac{y_1 + y_2 + \ldots + y_N}{N} = \frac{\sum_{i=1}^{N} y_i}{N}$$

Y el *total poblacional,* que es otro parámetro queda dado por la suma de todos los valores de la característica en cada una de las unidades de la población:

$$Y = y_1 + y_2 + \ldots + y_N = \sum_{i=1}^{N} y_i$$

que también se puede expresar como $N\bar{Y}$.

Al extraer una muestra de tamaño n, las unidades seleccionadas pueden ser las primeras n unidades que aparecen en el listado, ya que ésta es una muestra posible, y nos referiremos a ella como y_1, y_2, y_3, . . ., y_n. Sin embargo, por lo general las unidades seleccionadas estarán dispersas en todo el listado y nos referiremos a ellas de la misma manera: y_1, y_2, . . ., y_n. Entonces, por ejemplo, y_7 en la muestra y_1, y_2, . . ., y_n será la séptima unidad elegida en la muestra y puede corresponder a la unidad 315 en el listado de la población.

3.3 TABLAS DE NUMEROS ALEATORIOS

La extracción de una muestra aleatoria se efectúa seleccionando una a una las n unidades. Como en la práctica las poblaciones no son pequeñas se usan *tablas de números aleatorios* (tabla 3.1). Estas tablas están construidas de manera que se garantiza estadísticamente la aleatoriedad de sus elementos. Una manera de obtenerlas es a través de computadoras digitales usando funciones de biblioteca ya construidas para este fin. (Ver el apartado 5.7).

Supongamos que deseamos elegir a dos números aleatorios entre 1 y 50. Para ello podemos utilizar pares de números en la tabla y éstos pueden ser adyacentes. Iniciemos con los primeros dos del segundo grupo de números (99) y avancemos hacia la derecha (56), (31), . . . El 99 no está comprendido entre 1 y 50 de manera que se descarta, el 56 igualmente, el 31 sí se elige y el 28 también; estos dos últimos números constituyen nuestra selección.

Ahora elijamos a cinco números aleatorios comprendidos entre 1 y 915. Para ello podemos usar agrupaciones de tres dígitos. Si comenzamos arriba y a la derecha de la tabla con el número 574 y continuamos hacia abajo, la muestra queda dada por:

574
399
74
195
308

Tabla 3.1

Números aleatorios

16 99 41	99 56 31	28 76 26	73 19 15	57 44 94
43 34 83	23 42 99	06 39 84	31 61 29	39 99 99
75 61 18	80 72 61	42 89 22	99 28 05	07 47 07
10 30 57	40 79 30	65 93 49	42 75 51	19 54 93
81 28 33	05 75 61	21 33 19	36 05 37	30 84 80
78 53 31	66 07 92	02 98 59	71 79 85	35 10 66
72 15 95	21 55 17	89 53 16	56 69 09	40 91 40
12 45 96	72 44 26	82 09 42	12 61 62	55 05 41
32 47 05	82 59 05	62 75 24	36 84 19	11 21 43
15 43 91	52 51 78	58 99 04	03 75 71	34 50 75
80 13 31	30 09 65	32 06 21	28 93 23	13 08 42
59 94 08	38 68 75	20 90 19	78 48 34	58 17 11
62 26 89	09 89 13	27 30 01	31 43 47	21 58 65
36 18 06	08 14 55	62 84 68	78 16 59	22 63 76
83 55 66	36 27 51	28 55 48	41 74 80	03 94 39
37 72 86	68 78 75	47 51 41	76 77 04	50 65 67
60 83 53	77 48 13	81 62 79	93 64 22	98 46 95
19 63 11	54 88 95	27 69 08	85 39 27	27 81 68
92 57 48	11 69 88	47 30 38	47 84 69	43 62 78
88 07 20	65 91 35	05 53 18	61 92 33	65 08 73

Si la muestra anterior debiera estar comprendida entre 100 y 915, entonces se descartaría el número 74 y buscaríamos uno nuevo, éste es el 351.

3.4 ESTIMACION DE MEDIAS, TOTALES, PORCENTAJES Y COCIENTES POBLACIONALES

En los apartados anteriores nos hemos referido al estudio de una característica en cada unidad en la muestra, y su valor para la unidad i-ésima lo representamos por y_i. En este apartado y para construir la tabla 3.2, se han colectado cuatro características de cada unidad en la muestra, de manera que para su representación hemos empleado cuatro letras diferentes: y_i, x_i, z_i y w_i, en otras palabras para cada unidad muestral habrá cuatro características en estudio. Con ellas será más adecuada la presentación de los estimadores fundamentales en el muestreo aleatorio simple.

En un estudio desarrollado sobre los nombres de personas

utilizados en México, se tomó una muestra aleatoria simple de 10 personas de entre 100, obteniéndose los resultados de la tabla 3.2 respecto a los atributos: sexo, número de letras, de vocales y de consonantes en sus nombres:

Tabla 3.2

Persona	*Nombre*	*No. letras* y_i	*Sexo* x_i	*No. vocales* z_i	*No. consonantes* w_i
1	Mario	5; $y_1 = 5$	M, $x_1 = 0$	$z_1 = 3$	$w_1 = 2$
2	Juan	4; $y_2 = 4$	M, $x_2 = 0$	$z_2 = 2$	$w_2 = 2$
3	Raúl	4; $y_3 = 4$	M, $x_3 = 0$	$z_3 = 2$	$w_3 = 2$
4	Gloria	6; $y_4 = 6$	F, $x_4 = 1$	$z_4 = 3$	$w_4 = 3$
5	Cosme	5; $y_5 = 5$	M, $x_5 = 0$	$z_5 = 2$	$w_5 = 3$
6	Carlos	6; $y_6 = 6$	M, $x_6 = 0$	$z_6 = 2$	$w_6 = 4$
7	Luz	3; $y_7 = 3$	F, $x_7 = 1$	$z_7 = 1$	$w_7 = 2$
8	Ernesto	7; $y_8 = 7$	M, $x_8 = 0$	$z_8 = 3$	$w_8 = 4$
9	Rosa	4; $y_9 = 4$	F, $x_9 = 1$	$z_9 = 2$	$w_9 = 2$
10	Norma	5; $y_{10} = 5$	F, $x_{10} = 1$	$z_{10} = 2$	$w_{10} = 3$
		$\sum_{i=1}^{n} y_i = 49$;	$\sum_{i=1}^{n} x_i = 4$;	$\sum_{i=1}^{n} z_i = 22$;	$\sum_{i=1}^{n} w_i = 27$

Estimación de medias: Intuitivamente, el número medio de letras por nombre será de $\frac{49}{10}$ que es igual a 4.9. En general, si extraemos una muestra aleatoria simple de tamaño n de entre una población de N elementos y deseamos estimar una media poblacional usamos como estimador a la *media muestral*, o sea:

$$\hat{\bar{Y}} = \bar{y} = \frac{\sum_{i=1}^{i=n} y_i}{n} \qquad 3.1$$

$\hat{\bar{Y}}$ se lee, "el estimador de $\bar{Y}$" por lo que la expresión anterior se lee así: "el estimador de $\bar{Y}$ es $\bar{y}$".

Estimación de totales: Y si en la misma población se desea estimar el número total de letras en sus 100 nombres, multiplicamos el número medio de letras por nombre, por el número total de nombres o personas, dando como resultado 4.9 X 100 = 490 letras. Para estimar un total, habiendo seleccionado una muestra aleatoria simple multiplicamos la media muestral $\bar{y}$ por N:

$$\hat{Y} = N\bar{y} = N\frac{\sum_{i=1}^{i=n} y_i}{n} \qquad 3.2$$

Estimación de porcentajes: Usando la misma muestra podemos estimar el porcentaje de personas de sexo femenino y ésta será igual a la fracción de mujeres en la muestra multiplicada por 100; en este caso $\frac{4}{10}100 = 40\%$. En general, para estimar el porcentaje de unidades poblacionales que tienen una determinada cualidad (sexo femenino), o pertenecen a una clase determinada, encuéntrese la fracción de ellas en la muestra y multiplíquese por 100,

$$\hat{P} = p = \frac{a}{n}100\% \qquad 3.3$$

en que $\hat{P} = p$ se lee "el estimador del porcentaje poblacional P es p"; y éste es el porcentaje en la muestra. En la ecuación 3.3, "a" es el número de unidades en la muestra que tienen o que poseen la característica de interés.

En el caso de proporciones o de porcentajes se recurre a una variable aleatoria auxiliar que toma los valores uno o cero, según que la unidad tenga o no la característica buscada. En la tabla 3.2 y para el caso de sexo se utilizó la variable auxiliar X que toma los valores $x_1, x_2, \ldots, x_{10}$ ahí anotados.

Estimación de razones o de cocientes: A partir de la tabla 3.2 también podemos estimar el número de vocales por consonante. Si a este cociente le llamamos R, su valor es $\frac{22}{27}$ En general, si deseamos estimar una razón $R = \frac{Y}{X}$ utilizamos el cociente de las medias muestrales, o el de los totales en la muestra como sigue:

$$* \hat{R} = \frac{\frac{\sum_{i=1}^{n} y_i}{n}}{\frac{\sum_{i=1}^{n} x_i}{n}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \qquad 3.4$$

* Debemos observar que con la notación de la tabla 3.2, $\hat{R}$ debería quedar definido como el cociente de $\sum_{i=1}^{i=n} z_i$ a $\sum_{i=1}^{i=n} w_i$. En la ecuación 3.4 hemos usado $\sum_{i=1}^{i=n} y_i$ a $\sum_{i=1}^{i=n} x_i$ notación que será usual para razones en el resto del libro.

3.5 CONSISTENCIA E INSESGAMIENTO DE LOS ESTIMADORES

Si en los estimadores $\bar{y}$, $N\bar{y}$, p y $\hat{R}$ propuestos en las ecuaciones 3.1 a 3.4 sustituimos n por N vemos que se reproducen los parámetros poblacionales respectivos, por ejemplo: $\bar{y} = \frac{1}{n}\sum_{i=1}^{i=n} y_i$; al sustituir n por N, tenemos $\frac{1}{N}\sum_{i=1}^{i=N} y_i$, pero esta expresión es igual a la media poblacional $\bar{Y}$; es decir, estos estimadores son consistentes (apartado 2.4). En el caso de la proporción, debemos recordar el recurso de introducir una variable aleatoria que toma los valores uno o cero. Al sumar los unos en toda la población se obtiene el total de unidades que tienen la propiedad de interés. De manera que p es un estimador consistente de P. El estimador media muestral además de ser consistente es insesgado del parámetro $\bar{Y}$, para comprobarlo debemos obtener su esperanza matemática:

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

Entonces según 2.1, 2.2 y 2. 3:

$$E(\bar{y}) = E\left(\frac{1}{n}\sum_{i=1}^{i=n} y_i\right) = \frac{1}{n} E\left(\sum_{i=1}^{i=n} y_i\right) = \frac{1}{n}\sum_{i=1}^{i=n} E(y_i)$$

$$= \frac{1}{n}\sum_{i=1}^{i=n}\left(\sum_{j=1}^{j=N} y_j \left(\begin{array}{l}\text{Probabilidad de que}\\ \text{la unidad sea elegida}\\ \text{en la } j\text{-ésima extracción}\end{array}\right)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{i=n}\left(\sum_{j=1}^{j=N} y_j \left(\begin{array}{l}\text{probabilidad de que}\\ \text{no sea elegida en las}\\ \text{primeras } j-1\\ \text{extracciones}\end{array}\right)\left(\begin{array}{l}\text{probabilidad de que}\\ \text{lo sea en la } j\text{-ésima}\end{array}\right)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{i=n}\left(\sum_{j=1}^{j=N} y_j\left(\frac{N-j+1}{N}\right)\left(\frac{1}{N-j+1}\right)\right)^{*}$$

* En el ejercicio 3.13 se pide que se obtenga el término $\frac{(N-j+1)}{N}$

$$= \frac{1}{n} \sum_{i=1}^{i=n} \left(\sum_{j=1}^{j=N} y_j \frac{1}{N} \right) = \frac{1}{nN} n \sum_{j=1}^{j=N} y_j = \bar{Y}$$

Y con respecto a $N\bar{y}$, estimador del total poblacional, usando el resultado de la ecuación 2.2 se tiene:

$$E(N\bar{y}) = N(E(\bar{y})) = N\bar{Y} = Y$$

Entonces $\bar{y}$ y $N\bar{y}$ son estimadores insesgados de $\bar{Y}$ y de Y respectivamente. Para el caso de proporciones o de porcentajes, con ayuda de la *variable auxiliar* se encuentra que el total de unidades en la población, con la característica de interés es $A = \sum_{i=1}^{i=N} y_i$; donde para esta situación y_i es la variable auxiliar, además la proporción poblacional es $P = \frac{A}{N} = \frac{1}{N} \sum_{i=1}^{i=N} y_i = \bar{Y}$ que es la media poblacional. A su vez en términos de la muestra el estimador de la proporción es $p = \frac{a}{n} = \frac{1}{n} \sum_{i=1}^{i=n} y_i = \bar{y}$; que es la media muestral de los valores y_i de la variable auxiliar, o sea que *estimar una proporción o un porcentaje es estimar una media donde la variable toma como valores 1 o 0.* Y como la media muestral es insesgada de $\bar{Y}$ se concluye que p es un estimador insesgado de P. Si en lugar de una proporción queremos estimar el número total de unidades A que pertenecen a la clase de interés, usamos como estimador a Np; por ejemplo, para obtener una estimación del número total de registros defectuosos en un archivo, multiplicamos a la proporción de este tipo de registros (defectuosos) por el total de ellos (defectuosos y no defectuosos). Este estimador también será insesgado de A.

Finalmente, para el caso de $\hat{R}$, la obtención de su esperanza matemática no es tan simple, dado que se trata de un cociente de variables aleatorias. El estimador $\hat{R}$ es sesgado de R (ver el apartado 2.4) y el sesgo disminuye al aumentar el tamaño de la muestra. Si se expande con la serie de Taylor se puede demostrar (ver el apartado 5.1 y los ejercicios 5.3 y 5.4) que el estimador $\hat{R}$ es insesgado bajo la aproximación lineal de la serie, y que el resto de los términos decrecen rápidamente para tamaños crecientes de n.

Ejemplo 3.1 Para una encuesta sobre la industria de la fabricación de nieves y helados en la ciudad de San Luis Potosí se dispone de un listado de 342 empresas con el nombre y dirección de cada una de ellas. Se tiene la idea de que la mayoría de esas empresas, las cuales son relativamente pequeñas, además de fabricar el producto lo venden al menudeo. De manera que una primera estimación a hacer, se refiere al porcentaje de empresas que fabrican y venden su producto al menudeo. Otras estimaciones que se desean obtener son las siguientes: número medio de empleados por establecimiento y el total de ellos en las 342 empresas.

Se numera consecutivamente el listado de empresas empezando con 1 y terminando con el 342 y posteriormente se elige a 15 números aleatorios diferentes entre 1 y 342. Los establecimientos en la muestra resultan ser aquellos con los números 11, 129, 50, 85, 341, 320, 294, 7, 330, 329, 265, 237, 266, 71 y 280, los cuales se marcan en el listado y se prepara una lista separada de ellos con sus nombres y direcciones. Se hacen las visitas correspondientes y se obtienen los siguientes resultados para las empresas que fabrican y venden el producto al menudeo:

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}
0	1	1	0	1	1	1	1	0	1	1	1	0	1	0

Se indican con un 1 las empresas que fabrican y venden. Entonces el porcentaje estimado de empresas que fabrican y venden al menudeo según la ecuación 3.3 es:

$$p = \frac{a}{n} 100 = \frac{10}{15} 100 = 66.7\%$$

Los resultados a la pregunta sobre el número de empleados fueron:

Establecimiento	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. de empleados	3	1	5	2	7	2	2	1	2	1	1	4	2	2	1

Usando la fórmula 3.1 obtenemos:

$$\bar{y} = \frac{\sum_{i=1}^{i=n} y_i}{n} = \frac{36}{15} = 2.4 \text{ empleados por establecimiento}$$

2.4 es el valor estimado de la media por establecimiento. Como estimador del total usamos la expresión 3.2 y obtenemos:

$$\hat{Y} = N\bar{y} = 342\ (2.4) = 820.8 \text{ empleados.}$$

De las 342 empresas, se muestreó a 15 de ellas, es decir, al $\frac{15}{342}\,100 = 4.4\%$ aproximadamente. Al cociente $\frac{n}{N} = \frac{15}{342}$ se le denomina *fracción de muestreo* y se le denota por la letra f; representa a la fracción de unidades en la población que son muestreadas.

La teoría del muestreo probabilístico permite derivar más resultados a partir de la muestra, que la *estimación puntual* (66.7% de establecimientos, 2.4 empleados por establecimiento, 820.8 empleados). Nos estamos refiriendo a resultados para evaluar lo adecuado de la encuesta y éstos quedan dados por estimaciones de variancias, las que son tratadas en el apartado siguiente. Con su ayuda podemos establecer intervalos de confianza y tener así *estimaciones por intervalos* como se les denomina generalmente. En los ejercicios se pide estimar las variancias para cada una de las estimaciones anteriores y también obtener intervalos de confianza para ellas.

3.6 VARIANCIAS DE LOS ESTIMADORES Y ESTIMADORES DE ELLAS

Se han propuesto estimadores para cada uno de los parámetros que interesan usualmente: medias, totales, porcentajes y cocientes. La aplicación de ellos nos proporciona como resultado un cierto número real; así en el apartado 3.4 obtuvimos 4.9 letras por nombre en el caso de la media. ¿Cómo nos fue en el muestreo?, ¿qué tan buena es la estimación?, ¿de qué magnitud será el error cometido?

El muestreo probabilístico permite contestar este tipo de pre-

guntas y para ello hay que calcular las variancias de los estimadores. Obtengamos la variancia de la media muestral aplicando la ecuación 2.5 y recordando que la esperanza de la media muestral $\bar{y}$ es la media poblacional $\bar{Y}$.

i) **Variancia de la media muestral:**

$$V(\bar{y}) = E((\bar{y})^2) - \bar{Y}^2$$

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

$$= \frac{1}{n^2} E\left(\sum_{i=1}^{n} y_i^2 + \sum_{i\neq j}^{n} y_i y_j\right) - \bar{Y}^2$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n} E(y_i^2) + \sum_{i\neq j}^{n} E(y_i y_j)\right) - \bar{Y}^2,$$

para obtener esta expresión se han usado los resultados 2.2 y 2.3. Recordando que $E(y_i) = (1/N)\sum_{j=1}^{N} y_j$:

$$V(\bar{y}) = \frac{1}{n^2}\left(\sum_{i=1}^{n}\frac{1}{N}\sum_{j=1}^{N} y_j^2 + \sum_{i\neq j}^{n}\sum_{k\neq l}^{N} y_k y_l \text{ (probabilidad de que } y_k \text{ se obtenga en la } i\text{-ésima extracción) (probabilidad de que } y_l \text{ se obtenga en la } j\text{-ésima extracción dado que } y_k \text{ lo fue en la } i\text{-ésima)}\right) - \bar{Y}^2$$

$$= \frac{1}{n^2}\left(\frac{n}{N}\sum_{j=1}^{N} y_j^2 + n(n-1)\sum_{k\neq l}^{N} y_k y_l \frac{1}{N}\frac{1}{N-1}\right) - \bar{Y}^2$$

$$= \frac{1}{nN}\sum_{j=1}^{N} y_j^2 + \frac{n-1}{nN(N-1)}\sum_{k\neq l}^{N} y_k y_l - \bar{Y}^2$$

Pero: $(\Sigma y_j)^2 = \Sigma y_j^2 + \sum_{i\neq j} y_i y_j$, entonces:

$$V(\bar{y}) = \frac{1}{nN}\sum_{j=1}^{N} y_j^2 - \frac{n-1}{nN(N-1)}\sum_{j=1}^{N} y_j^2 + \frac{n-1}{nN(N-1)}\left(\sum_{j=1}^{N} y_j\right)^2 - \bar{Y}^2$$

$$V(\bar{y}) = \left(\frac{1}{nN} - \frac{n-1}{nN(N-1)}\right)\sum_{j=1}^{N} y_j^2 + \frac{n-1}{nN(N-1)}(\Sigma\ y_j)^2 - \bar{Y}^2$$

$$= \frac{N-n}{nN(N-1)}\sum_{j=1}^{N} y_j^2 + \bar{Y}^2\ \frac{n-N}{n(N-1)}$$

$$= \frac{N-n}{N}\frac{1}{n}\ \frac{\sum_{j=1}^{N} y_j^2 - N\bar{Y}^2}{N-1} = \frac{N-n}{N}\frac{1}{n}\ \frac{\sum_{j=1}^{N}(y_j - \bar{Y})^2}{N-1}$$

$$= \frac{S^2}{n}\left(1 - \frac{n}{N}\right) = \frac{S^2}{n}(1-f) \qquad 3.5$$

En la ecuación 3.5, S^2 representa y se define como la variancia de las y_i en la población, o sea la *variancia poblacional* en tanto que $\frac{S^2}{n}$ es la variancia de la media muestral en poblaciones infinitas. Esta variancia aparece reducida mediante el factor $(1 - \frac{n}{N})$ por lo que se le llama *factor de corrección para población finita* y su influencia depende de la relación $\frac{n}{N}$ llamada *fracción de muestreo* o fracción muestral.

ii) Variancia de $N\bar{y}$:

Para encontrar la variancia del estimador del total poblacional hacemos lo siguiente:

$$V(N\bar{y}) = N^2 V(\bar{y}) = \frac{N^2 S^2}{n}(1-f) \qquad 3.6$$

iii) Variancia de p:

En el caso de un porcentaje sólo hay que expresar S^2 en términos de P y Q, donde Q vale $1 - P$, de la manera siguiente:

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

de donde:

$$V(p) = \frac{NPQ}{N-1}\,\frac{1-f}{n} \qquad 3.7$$

iv) Variancia de $\hat{R}$:

En el caso de razones se demuestra que la variancia del estimador $\hat{R}$ está dada por (apartado 5.1):

$$V(\hat{R}) \doteq \frac{1-f}{n\bar{X}^2}\,\frac{\sum_{i=1}^{i=N} (y_i - Rx_i)^2}{N-1} \qquad 3.8$$

Sin embargo, dada una muestra no podemos encontrar o calcular el valor particular de las variancias anteriores ya que éstas requieren conocer parámetros poblacionales que sólo se pueden obtener mediante un censo. La solución consiste en proponer estimadores de esos parámetros, basados en los resultados de la muestra.

Por ejemplo, si sobre la tabla 3.2 deseamos calcular $V(\bar{y})$ la varianza del número medio de letras por nombre, entonces, de acuerdo a la ecuación 3.5: n vale 10, N vale 100, $\frac{n}{N} = \frac{10}{100} = \frac{1}{10}$, uno de cada diez. $1 - \frac{n}{N} = \frac{9}{10}$,

$$\frac{1}{n}(1-f) = \frac{1}{10}\,\frac{9}{10} = \frac{9}{100}$$

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

En esta última expresión para el cálculo del parámetro poblacional S^2 y en el miembro de la derecha, se requiere conocer cada uno de los valores de $y_1, y_2, \ldots, y_{N-1}, y_N = y_{100}$. Pero sólo conocemos a $y_1, y_2, \ldots, y_{10}$ a través de la muestra. Entonces, si no se dispone de un censo (si dispusiéramos del censo no requeriríamos el muestreo), no puede ser calculada la varianza verdadera. Lo mismo es cierto para un total, un porcentaje y una razón.

La solución consiste en utilizar la misma muestra de que disponemos para estimar u obtener aproximadamente el valor de la varianza verdadera. Y nos referimos a ella como un estimador de la varianza verdadera. En la situación de una media o de un total estimamos aproximadamente a S^2, y en un porcentaje estimamos a P. En las líneas siguientes se proponen estimadores para cada una de las varianzas anteriores. Consecuentemente, las ecuaciones propuestas 3.9, 3.9.1, 3.10, 3.11.1 y 3.11.2 serán las que se usen en la práctica para obtener varianzas, errores estándar e intervalos confidenciales. En tanto que las ecuaciones 3.5, 3.6, 3.7 y 3.8 deben ser vistas como un paso necesario para obtener al conjunto de estimadores de varianzas (3.9, 3.9.1, 3.10, 3.11.1 y 3.11.2).

i') Estimador de $V(\bar{y})$:

Un estimador insesgado de la variancia de la media muestral es el siguiente:

$$\hat{V}(\bar{y}) = \frac{s^2}{n}(1-f), \text{ en que } s^2 = \frac{\sum_{i=1}^{i=n}(y_i - \bar{y})^2}{n-1} \qquad 3.9$$

$$\hat{V}(\bar{y}) = \frac{1-f}{n}\,\frac{1}{n-1}\left(\sum^{n} y_i^2 - \frac{(\sum^{n} y_i)^2}{n}\right)$$

Para fines operacionales, esta última ecuación es más adecuada que la original.

Veámoslo:

$$E\left((1-f)\frac{s^2}{n}\right) = \frac{1-f}{n}\,E(s^2)$$

$$= \frac{1-f}{n}\,\frac{1}{n-1}\,E\left(\sum_{i=1}^{i=n}((y_i - \bar{Y}) - (\bar{y} - \bar{Y}))^2\right)$$

$$= \frac{1-f}{n}\,\frac{1}{n-1}\left(\frac{n}{N}\sum_{i=1}^{i=N}(y_i - \bar{Y})^2 - n((\bar{y} - \bar{Y})^2)\right)$$

$$= \frac{S^2}{n}(1-f)$$

ii') Estimador de $V(N\bar{y})$:

El estimador de variancia de $N\bar{y}$ es obtenido mediante el resultado de la expresión 3.9:

$$\hat{V}(N\bar{y}) = N^2 \hat{V}(\bar{y}) = \frac{N^2 s^2}{n}(1 - f)$$

$$\hat{V}(N\bar{y}) = \frac{N^2(1-f)}{n} \frac{1}{n-1} \left(\sum^{n} y_i^2 - \frac{(\sum^{n} y_i)^2}{n}\right) \qquad 3.9.1$$

este estimador de variancia es insesgado de $V(N\bar{y})$ por ser una transformación lineal de $\hat{V}(\bar{y})$.

iii') Estimador de V(p):

Para el caso de porcentajes hay que usar la expresión de $V(\bar{y})$ en la cual y_i toma los valores 1 ó 0 y llegar a que:

$$\left.\begin{aligned} &\hat{V}(p) = \frac{N-n}{(n-1)N} pq, \text{ y a que:} \\ &\hat{V}(\hat{A}) = \frac{N(N-n)}{n-1} pq \end{aligned}\right\} \quad 3.10$$

Donde p es la proporción o porcentaje estimado mediante la media muestral o su equivalente la expresión 3.3.

Las expresiones 3.10 resultan ser estimadores insesgados de $V(p)$ y de $V(\hat{A})$ respectivamente.

iv') Estimador de $V(\hat{R})$:

iv.1. *Caso en que se conoce* $\bar{X}$

$$\hat{V}(\hat{R}) = \frac{1-f}{n\bar{X}^2} \frac{\overset{i=n}{\Sigma} (y_i - \hat{R}x_i)^2}{n-1} \qquad 3.11.1$$

$$\hat{V}(\hat{R}) = \frac{1-f}{n\bar{X}^2} \frac{1}{n-1} \left(\sum^{n} y_i^2 - 2\hat{R} \sum^{n} y_i x_i + \hat{R}^2 \sum^{n} x_i^2\right)$$

Tabla 3.3
Estimadores Aplicables a Muestreo Aleatorio Simple

Parámetro	*Estimador del Parámetro*	*Variancia del estimador*	*Estimador de la Variancia*	*Intervalos de confianza* *
Media	$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$	$V(\bar{y}) = (1-f) \frac{S^2}{n}$	$\hat{V}(\bar{y}) = (1-f) \frac{s^2}{n}$	$\bar{y} \mp t(\hat{V}(\bar{y}))^{1/2}$
Total	$N\bar{y} = \frac{N}{n} \sum_{i=1}^{n} y_i$	$V(N\bar{y}) = (1-f) \frac{N^2 S^2}{n}$	$\hat{V}(N\bar{y}) = (1-f) \frac{N^2 s^2}{n}$	$N\bar{y} \mp t(\hat{V}(N\bar{y}))^{1/2}$
Porcentaje	$p = \frac{a}{n} 100$ $\hat{A} = N \frac{a}{n}$	$V(p) = \frac{NPQ\ (1-f)}{(N-1)\ n}$ $V(\hat{A}) = \frac{N^3 PQ}{N-1} \frac{(1-f)}{n}$	$\hat{V}(p) = \frac{N-n}{(n-1)N} pq$ $\hat{V}(A) = \frac{N(N-n)}{n-1} pq$	$p \mp t(\hat{V}(p))^{1/2}$ $\hat{A} \mp t(\hat{V}(\hat{A}))^{1/2}$
Razones	$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$	$V(\hat{R}) = \frac{1-f}{n\bar{X}^2} \frac{\sum_{i=1}^{N} (y_i - Rx_i)^2}{N-1}$	$\hat{V}(\hat{R}) = \frac{1-f}{n\bar{X}^2} \frac{\sum_{i=1}^{n} (y_i - \hat{R}x_i)^2}{n-1}$ Nota: Cuando se desconoce $\bar{X}$ puede ser usado $\bar{x}$ en su lugar.	$\hat{R} \mp t(\hat{V}(\hat{R}))^{1/2}$

* Para obtener intervalos de confianza del 95% use $t = 2$.

iv.2. *Caso en que se desconoce* $\bar{X}$

$$\hat{V}(\hat{R}) = \frac{1-f}{n\bar{x}^2} \frac{\sum_{i=1}^{i=n} (y_i - \hat{R}x_i)^2}{n-1} \qquad 3.11.2$$

$$\hat{V}(\hat{R}) = \frac{1-f}{n\bar{x}^2} \frac{1}{n-1} \left(\sum^{n} y_i^2 - 2\hat{R}\sum^{n} y_i x_i + \hat{R}^2 \sum^{n} x_i^2\right)$$

en la cual $\bar{x}$ es la media muestral: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Las expresiones 3.11.1 y 3.11.2 difieren en el término $\bar{X}$, media poblacional de la característica X en un caso y la media muestral $\bar{x}$ de ella en el otro. Siempre que se conozca $\bar{X}$ deberá emplearse la expresión 3.11.1, en ocasiones eso no es posible y se le sustituye en 3.11.2 por su estimador $\bar{x}$.

La tabla 3.3 es un resumen de las expresiones utilizadas en este capítulo.

TABLA 3.4
Fórmulas para el cálculo de varianzas estimadas (m. a. s.)

Parámetro	*Estimador*	*Se requiere*
Media	$\frac{1-f}{n} \frac{1}{n-1} \left(\sum^{n} y_i^2 - \frac{(\sum^{n} y_i)^2}{n}\right)$	*i)* La suma de cuadradros de cada observación. *ii)* La suma de las observaciones elevada al cuadrado.
Total	$\frac{N^2(1-f)}{n} \frac{1}{n-1} \left(\sum^{n} y_i^2 - \frac{(\sum^{n} y_i)^2}{n}\right)$	Misma información que para la media.
Razón	$\frac{1-f}{n\bar{X}^2} \frac{1}{n-1} \left(\sum^{n} y_i^2 - 2\hat{R}\sum^{n} y_i x_i + \hat{R}^2 \sum x_i^2\right)$ Donde $\hat{R} = \frac{\bar{y}}{\bar{x}}$	*i)* La suma de cuadrados de cada observación de la variable en el numerador. *ii)* La suma de cuadrados de cada observación de la variable en el denominador. *iii)* La suma de productos mixtos de las observaciones de una variable por la otra.
Proporción	$\frac{a}{n}$	*i)* El número total de unidades muestrales con la característica de interés.

3.7 INTERVALOS DE CONFIANZA

Realmente es difícil indicar un tamaño de muestra mínimo a partir del cual sea razonable suponer la normalidad en la distribución de los estimadores. Esto se hace más evidente si se recuerdan los histogramas obtenidos en el ejemplo 2.2 sobre la estimación de ingresos a partir de una distribución experimental. Cuando los tamaños de muestra son del orden de millares o de centenares, realmente no se titubea acerca de la distribución. Pero en casi todas las encuestas es necesario obtener estimaciones para subpoblaciones o dominios de estudio. Por ejemplo, en una encuesta a estudiantes, se desean estimaciones por sexo, edad, carrera y período académico. Esto equivale a que el tamaño de muestra total se divida inicialmente en dos, por la consideración de dos sexos. Si partimos de $n = 5\,000$, tendríamos aproximadamente 2 500 para cada sexo. Al considerar ocho edades diferentes, digamos: −18, 18, 19, 20, 21, 22, 23, +23, tendríamos $\frac{2\,500}{8} = 313$ casos para cada edad. Si existen tres carreras y ocho períodos académicos (8 semestres) pudiéramos esperar $\frac{313}{3(8)} = 13$ estudiantes en la muestra para cada celda. Aquí estamos en problemas porque hasta esperaríamos algunas celdas vacías, además que el número 13 no es muy bueno como tamaño de muestra, posiblemente sea demasiado pequeño. En estas condiciones, no siempre se pueden prometer estimaciones para cualquier subpoblación. Para fines operacionales, se enuncia en algunos casos la regla siguiente: para aquellas celdas con al menos 25 observaciones se efectuarán las estimaciones, en caso contrario, no se efectuarán. Con esto en mente, prosigamos hacia intervalos confidenciales.

Como se dijo en el apartado 2.3, generalmente es válido suponer que los estimadores se distribuyen normalmente; esto permite calcular intervalos de confianza de manera simple. Prácticamente esto es válido si el tamaño de muestra es mayor de 30.

Para una *media* poblacional el intervalo de confianza queda dado por:

$$\{\bar{y} - t(\hat{V}(\bar{y}))^{1/2},\ \bar{y} + t(\hat{V}(\bar{y}))^{1/2}\}^* \qquad 3.12$$

* Para un tamaño de muestra pequeño, t debe ser buscado en las tablas de la distribución t con n-1 grados de libertad, ya que la variancia es desconocida.

En el caso del *total* poblacional el intervalo de confianza se construye así:

$$\{N\bar{y} - t(\hat{V}(N\bar{y}))^{1/2},\ N\bar{y} + t(\hat{V}(N\bar{y}))^{1/2}\} \qquad 3.13$$

En el caso de *porcentajes* por:

$$\{p - t(\hat{V}(p))^{1/2},\ p + t(\hat{V}(p))^{1/2}\} \qquad 3.14$$

Y en el caso del *cociente* por:

$$\{\hat{R} - t(\hat{V}(\hat{R}))^{1/2},\ \hat{R} + t(\hat{V}(\hat{R}))^{1/2}\} \qquad 3.15$$

En donde t es la abscisa en la distribución normal tal que nos deja al centro de la misma un área igual a la confianza requerida. Así para intervalos de confianza del 95%, el valor de t es de 1.96 que usualmente se redondea a 2.

En el caso de porcentajes las unidades se encuentran dentro de la clase de interés o fuera de ella y como el muestreo es sin remplazo, a medida que se van extrayendo las unidades, las probabilidades de selección van variando. Este conjunto de características: población finita, existencia de dos clases en la población; C o su complemento, y extracción de una muestra aleatoria sin remplazo permiten concluir que la distribución verdadera del estimador p es la hipergeométrica. (Ver el ejercicio 3.12.)

Ejemplo 3.2 En la encuesta del apartado 3.4 referente a nombres de personas se tiene $N = 100$, $n = 10$ y la muestra está materializada en la tabla 3.2.

a) Estime el número medio de letras por nombre y calcule una estimación del error estándar.

$$\hat{Y} = \bar{y} = \frac{\sum^{n} y_i}{n} = \frac{49}{10} = 4.9 \text{ letras por nombre.}$$

$$\hat{V}(\bar{y}) = (1 - f)\frac{\hat{S}^2}{n} = (1 - f)\frac{s^2}{n}$$

$$= \left(1 - \frac{10}{100}\right)\frac{1}{10} \quad \frac{\sum^{n}(y_i - \bar{y})^2}{n - 1}$$

$$= \frac{1}{100}\left(\sum^{n} y_i^2 - \frac{(\sum^{n} y_i)^2}{n}\right)$$

$$= \frac{1}{100}\left(253 - \frac{(49)^2}{10}\right)$$

$$= 0.129 \text{ (letras/nombre)}^2$$

y el error estándar es:

$$(0.129)^{1/2} = 0.359 \text{ letras/nombre.}$$

Si se desean intervalos de confianza del 95%:

$$\{4.9 - 2(0.359),\ 4.9 + 2(0.359)\} =$$

$$\{4.182,\ 5.618\} \text{ letras/nombre}$$

b) Estime el número total de letras en los 100 nombres y de intervalos de confianza del 90% para su estimación.

$$\hat{Y} = N\bar{y} = 4.9(100) = 490 \text{ letras.}$$

$$\{490 - 1.64(100)(0.359),\ 490 + 1.64(100)(0.359)\} =$$

$$\{431,\ 549\} \text{ letras.}$$

c) Estime el porcentaje de personas de sexo femenino y encuentre el error estándar de su estimación.

$$\hat{P} = p = \frac{a}{n}\,100 = \frac{4}{10}100 = 40\%$$

$$e.e.(p) = \left(\frac{N-n}{(n-1)N}\ pq\right)^{1/2}$$

$$= \left(\frac{100-10}{9\,(100)}\ 40(60)\right)^{1/2} = 15.5\%$$

d) Estime el número de vocales por consonante y encuentre el error estándar de su estimación.

$$\hat{R} = \frac{\Sigma z_i}{\Sigma w_i} = \frac{22}{27} = 0.81 \text{ vocales/consonante.}$$

$$\text{e.e.}(\hat{R}) = \left(\left(\frac{1-f}{n\bar{w}^2}\right)\frac{\sum_{1}^{n}(z_i - \hat{R}w_i)^2}{n-1}\right)^{1/2}$$

$$= \left(\frac{\frac{90}{100}}{(10)(2.7)^2}\ \frac{\Sigma z_i^2 - 2\hat{R}\Sigma z_i w_i + \hat{R}^2 \Sigma w_i^2}{9}\right)^{1/2}$$

$$= \left(\frac{0.09}{(2.7)^2}\ \frac{52 - 1.62(61) + (0.81)^2(79)}{9}\right)^{1/2}$$

$$= \left(\frac{0.050119}{7.29}\right)^{1/2} = 0.0829$$

$$= 0.0829 \text{ vocales por consonante.}$$

Nota. En muchos casos el factor de corrección para población finita es despreciable y se puede suprimir en las expresiones anteriores.

Ejemplo 3.3 En este ejemplo se trata de mostrar algunas de las dificultades prácticas que se encuentran en las encuestas usuales y que debido a sus repercusiones hacen razonable y conveniente el que se les dedique un poco de tiempo para así evitar sorpresas desagradables.

Un grupo de médicos desea realizar una investigación sobre personas diabéticas en el Valle de México. El grupo dispone de 10 listados diferentes de ese tipo de enfermos cada uno con su dirección. Los listados se han formado a través de 8 años de atención a ese tipo de enfermos en 10 diferentes instituciones médicas. El grupo de médicos asegura que su listado es bueno, ya que se han dado instrucciones al personal administrativo para que se preocupe por su actualización y, por ello, debe ser usado sin reservas en el estudio.

El técnico que va a diseñarles el muestreo les indica que es pertinente hacer algunas comprobaciones sobre el listado antes de decidir sobre algún método de selección, sobre todo por la delicadeza de la investigación que se desea llevar a cabo.

Al grupo de médicos este comentario del técnico le parece ridículo, falto de fundamento y consideran que redundaría en un retraso muy substancial en el estudio, en adición al costo que

requeriría. El técnico arguye que si se hace una selección y se intenta visitar a las personas en la muestra, muchas de ellas habrían cambiado de domicilio, a otras no se les encontraría en casa, otras no querrán contestar, algunas ya habrán fallecido, y aún otras no cumplirán con la definición de diabético que haya sido adoptada para el estudio. Todo esto, indica el técnico, puede hacer que un buen porcentaje de las entrevistas, digamos el 50% no se puedan llevar a cabo, y por ello, los resultados serían pobres y no tendrían la validez necesaria para emitir una conclusión adecuada.

Ante esto, los médicos deciden asignar una parte de sus recursos para determinar si las inquietudes del técnico tienen fundamento o carecen de él; así, eligen una muestra aleatoria de 10 nombres de diabéticos en cada una de las 10 listas y envían a mensajeros de los hospitales a las direcciones correspondientes. Resulta alarmante advertir que cuatro días después los mensajeros no han concluido su trabajo por lo que los médicos deciden reunir a todos ellos y analizar lo que haya ocurrido.

El primer mensajero quien fue asignado a una zona pobre de la ciudad encontró que 4 de las 10 direcciones que le asignaron eran correctas. De las 6 restantes 3 familias se habían cambiado a otro domicilio, a 2 no se pudo localizar y la restante era incorrecta ya que la dirección correspondía a una fábrica que estaba ahí desde hacia muchos años.

El segundo mensajero tuvo más éxito que el primero, ya que encontró 6 direcciones correctas en una de las cuales el enfermo ya había fallecido, a la séptima no la pudo localizar y las tres restantes no las había trabajado por falta de tiempo.

Los demás mensajeros mostraron resultados similares con excepción de dos de ellos quienes tuvieron éxito en 8 y 9 direcciones respectivamente. Después de una pequeña revisión en los listados sobre las 8 y 9 direcciones con éxito se pudo comprobar que eran de enfermos recientes atendidos en esos hospitales en los dos últimos años.

Así, por los resultados obtenidos resulta razonable la inquietud del técnico. De haberse llevado a cabo la encuesta, sin hacer las verificaciones que se han comentado, no se hubieran podido desarrollar muchas entrevistas por causas como las señaladas y esto haría difícil el procesamiento de la información como ya se verá más adelante. Además los resultados que se obtuvieran no serían del todo válidos para cualquier diabético en la lista; principalmente

porque la enfermedad evoluciona a través del tiempo y se observarían razonablemente cosas diferentes entre personas diabéticas que han contraído la enfermedad recientemente y aquellas que la tienen desde hace varios años.

Ejemplo 3.4. En una encuesta a varias etapas ha quedado seleccionada una oficina (conglomerado) con 475 empleados, y dentro de ella es necesario seleccionar a 25 empleados para una muestra aleatoria simple. El responsable de la oficina nos proporciona un listado de su personal en el cual sólo figuran los nombres. Para efectuar el sorteo y asegurar la identificación única de cada empleado podemos numerarla del 1 al 475. Usando la tabla 3.1 de números aleatorios y empezando en la esquina superior izquierda a partir del 169 (¿por qué tres columnas?) la muestra sería la siguiente: 169, 433, 103, 124, 324, 154, 361, 377, 196, 118, 57, 331, 391, 408, 286, 353, 311, 234, 215, 300, 386, 98, 81, 362, 116.

Ejemplo 3.5. Consideremos a un conjunto de edificios educacionales en los cuales existen 25 000 estudiantes, los cuales a determinada hora, todos y cada uno de ellos se encuentran en algún salón. Nuestro propósito es seleccionar estudiantes, pero no contamos con una lista de ellos. Para efectos de la selección puede usarse el hecho de que en algún momento del día todos están en los salones. Para construir nuestro marco, hacemos una lista de los 625 salones existentes y los numeramos de alguna manera. El paso siguiente consiste en seleccionar salones (*unidades primarias de muestreo*) y dentro de salones seleccionar algunos estudiantes (*unidades secundarias de muestreo*). A los estudiantes en la muestra se les aplica el cuestionario. De los 625 salones o conglomerados primarios elegimos a 25 con fracción de muestreo para las unidades primarias de $f_1 = \frac{25}{625} = \frac{1}{25}$ y con muestreo aleatorio simple. Una vez hecha la selección marcamos o identificamos a los 25 salones en la muestra y elegimos a estudiantes dentro de ellos con fracción de muestreo para las unidades secundarias de $f_2 = \frac{1}{10}$. Esto también con muestreo aleatorio simple. La probabilidad final de selección para cada estudiante es la fracción de muestreo general

$$f = f_1 \cdot f_2 = \frac{1}{25} \frac{1}{10} = \frac{1}{250}$$

Este ejemplo, envuelve al concepto de muestreo multietápico en el cual se eligen unidades grandes (salones) y dentro de ellos se efectúa un nuevo sorteo (estudiantes). Realmente el número de etapas puede ser cualquiera: ciudad, colonia, manzana, vivienda y persona. A los conglomerados más grandes se les denomina unidades primarias, a las siguientes secundarias, terciarias, etc.

3.8 EJERCICIOS

3.1 Una población consta de 1 050 unidades numeradas del 1 al 1 050. Es necesario seleccionar una muestra aleatoria simple sin remplazo de tamaño 13. ¿Cómo lo haría usted?

3.2 Suponga que las unidades en la población anterior están numeradas como se indica a continuación:

1 001	9 811
1 002	9 813
•	9 910
1 317	9 911
1 318	9 912
1 319	9 913
•	•
•	•
•	•
2 040	9 918

¿Cómo seleccionaría una muestra aleatoria simple sin remplazo de tamaño 10?

3.3 En una encuesta desarrollada sobre una población de 10 000 familias, se tomó una muestra aleatoria de 40 de ellas, de manera que la fracción de muestreo fue de $\frac{40}{10\,000} = \frac{1}{250}$ es decir se entrevistó a una familia de cada 250. El número de personas que trabajan y el número total de miembros en cada familia de la muestra aparecen en la tabla 3.5.
Estime: *a*) El número medio de personas que trabajan por familia y encuentre intervalos de confianza del 95%;
b) El total de personas que trabajan y dé una estimación del error estándar.

3.4 Usando la información del ejercicio 3.3 estime el porcentaje de familias con más de cinco miembros y encuentre el error estándar de su estimación.

3.5 Sobre el mismo ejercicio 3.3, estime el número de miembros por persona que trabajan y el error estándar de su estimación.

3.6 Sobre la definición de muestreo aleatorio simple en el apartado 3.1, muestre que si a cada muestra posible se le elige con probabilidad igual, esto es equivalente a que la muestra elegida haya sido obtenida mediante la selección de n números aleatorios diferentes entre 1 y N.

3.7 En el ejemplo 3.1 derive intervalos de confianza del 95% para el porcentaje de empresas que fabrican y venden su producto al menudeo.

3.8 En el ejemplo 3.1, estime las variancias de la media y del total de empleados y calcule intervalos de confianza del 95% para cada uno de ellos.

3.9 Un grupo asesor de una escuela técnica piensa que los planes de estudio del plantel están un poco desactualizados y, mediante una encuesta sobre los egresados de ella, piensa derivar resultados que le ayuden en su reestructuración. Las preguntas que se deben formular, van dirigidas para aquellos egresados que estén trabajando como investigadores. El listado muestra 638 nombres cada uno con su dirección, de ellos se elige una muestra aleatoria de tamaño 20. Al hacer el trabajo de campo, los entrevistadores preguntan al egresado si es o no investigador. Si responde afirmativamente le hacen la entrevista y en caso contrario no la hacen. Al devolver los cuestionarios, el grupo asesor encuentra que 18 de los egresados en la muestra se calificaron como investigadores. Y así, emite instrucciones para que en el procesamiento de la información, en el cálculo de porcentajes y medias se use como tamaño de muestra 20.

i) ¿Cree usted que está bien definida la población objetivo? Indique sus razones.
ii) En el supuesto de que la población estuviera bien definida, ¿sería correcto usar el tamaño de muestra de 20 que indica el grupo asesor?

3.10 En cada cuestionario de un conjunto de 800 provenientes de una encuesta agrícola existe un dato de un porcentaje referente a una cualidad de la parcela agrícola. Los cuestionarios no han sido procesados aún, y se desea tener alguna idea del valor de ese porcentaje en los diferentes cuestionarios. Para ello, aprovechando su numeración consecutiva se elige aleatoriamente a 60 de ellos y se estima el porcentaje teniendo éste como valor 32%. Otra persona dice que la muestra fue muy pequeña y decide aumentarla a 120, calcula el porcentaje y obtiene como valor 33.4%. Una tercera persona aumenta el tamaño de muestra hasta 250 y encuentra como valor estimado a 32.9%.

¿Qué comentarios puede usted hacer respecto a los valores obtenidos en las diferentes muestras?

Tabla 3.5

No. de familia	*No. de personas que trabajan* y_i	*No. de miembros* x_i
1	1	3
2	3	7
3	1	5
4	1	1
5	1	9
6	2	8
7	1	8
8	2	5
9	3	7
10	1	3
11	1	4
12	1	4
13	1	8
14	1	11
15	4	4
16	1	7
17	3	3
18	2	3
19	1	3
20	2	2
21	2	5
22	2	4
23	1	9
24	1	6
25	1	6
26	1	7
27	2	6
28	3	6
29	1	5
30	1	5
31	1	9
32	1	8
33	1	4
34	3	6
35	1	1
36	7	7
37	1	3
38	1	3
39	6	6
40	3	9
$\sum_{i=1}^{40} y_i = 73$	$\sum y_i x_i = 413$	$\sum_{i=1}^{i=40} x_i = 220$
$\sum y_i^2 = 207$		$\sum x_i^2 = 1436$

3.11 En una urna existen B canicas blancas y A canicas azules. Se extrae a n de ellas aleatoriamente y con reposición. Se desea determinar la probabilidad de que b canicas de entre las n extraídas ($b \leqslant n$) sean blancas. ¿Cuál es la distribución del número b de canicas blancas en cada muestra de tamaño n? , ¿cuál es la media y la variancia de esta distribución?

3.12 En la urna del ejercicio 3.11 la muestra aleatoria es extraída sin reposición. ¿Cuál es la probabilidad de que b canicas ($b \leqslant n \leqslant A + B$) sean blancas? ¿Cuál es la distribución del número b de canicas blancas en cada muestra de tamaño n? ¿Cuál es la media y la variancia de esta distribución?

3.13 En el apartado 3.5 se derivó la esperanza de la media muestral y al hacerlo se afirma que: "la probabilidad de que no sea elegida en las primeras $j-1$ *extracciones*" *es* $\frac{N-j+1}{N}$, ¿está usted de acuerdo?

3.14 Una escuela tiene 20 salones en la planta baja numerados del 1 al 20 y 16 en la planta alta numerados del 1 al 16.

i) Indique brevemente cómo numeraría o identificaría a los salones para seleccionar una muestra aleatoria simple de tamaño 5.

ii) Utilizando los números aleatorios siguientes y avanzando de arriba hacia abajo, obtenga los 5 salones en la muestra.

Números aleatorios

74	50
90	98
25	46
01	81
41	11
31	39
25	04

CAPITULO 4

DETERMINACION DEL TAMAÑO DE LA MUESTRA

4.1 PRECISION ESTADISTICA

Sabemos que el valor correcto de un parámetro poblacional se puede determinar a través de un censo, es decir, por el estudio de toda la población.* Sin embargo, en muestreo al tomar una muestra de tamaño n sólo revisamos una fracción $\frac{n}{N}$ de una población de tamaño N, y en base a ella, inferimos el valor del parámetro en la población completa. Es evidente que en estas condiciones existirá un error en la estimación. Siempre, o casi siempre, existirá un error, no estamos exentos de él; pero éste es controlable, ya que como vimos en 3.6, la variancia de la media muestral es $V(\bar{y}) = (1 - \frac{n}{N})\frac{S^2}{n}$ y ésta se encuentra en función del tamaño de la muestra. Cuando n aumenta la variancia $V(\bar{y})$ disminuye hasta alcanzar el caso límite $n = N$ en el que la variancia del estimador se anula, el muestreo se convierte en censo y el error de estimación $d = 0$.

Entonces el muestreo lleva involucrado un error que denotamos por d; éste es un determinado porcentaje del valor del parámetro poblacional en consideración, digamos: $d = (x\%)\ \bar{Y}$ ó $d = (x\%)\ Y$. El usuario de la información debe estar consciente de este error, con respecto al parámetro poblacional existe un determinado porcentaje de él que está dispuesto a aceptar como error de muestreo.

* Aunque ya vimos en 1.5 que en ocasiones el método de medición aunado a otro tipo de eventualidades no permiten llegar al valor correcto.

Sin embargo, como el azar está presente en cada muestra que se elige; el estadístico no puede asegurar para todas las muestras posibles un mismo error. Ya que como se señala en la siguiente distribución de un estimador (figura 4.1), bajo la aproximación normal,

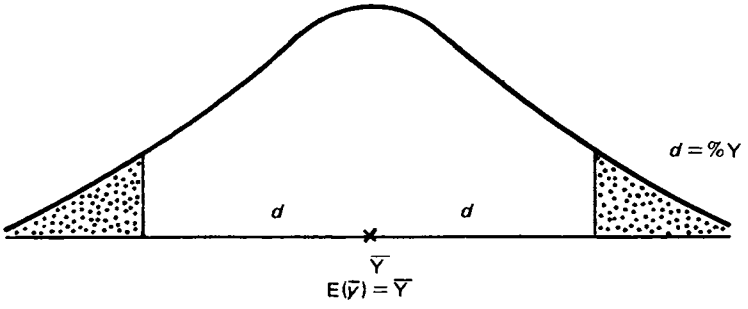


Figura 4.1

los intervalos de magnitud d a la derecha e izquierda del parámetro $\bar{Y}$, por lo general no cubren a toda la distribución; por lo cual en las áreas punteadas sigue existiendo probabilidad, es decir, oportunidad de que algunas muestras sean elegidas y arrojen un error mayor a d.

Para tomar en consideración este hecho, se introduce el concepto de *confianza*; es decir, se acepta un error $d = (x\%)\ \bar{Y}$ con una confianza del 95% en el sentido de que si se muestreara repetidas veces, en promedio, 95 de cada 100 muestras tendrían máximo un error de magnitud d, y 5 de ellas tendrían un error mayor. Y como es intuitivo el hecho de que, si en una urna existen 95 bolas verdes y 5 rojas y se revolviera y se extrajera una de ellas, es muy probable, es decir, se tiene bastante confianza (95%) de que la bola que salga sea verde (probabilidad de verde = $\frac{19}{20}$), aunque no estamos exentos de que salga roja, con probabilidad de $\frac{1}{20}$.

En muestreo, la *precisión* con que se desea una estimación queda indicada por un error d igual a x por ciento del parámetro poblacional y una confianza entre 0 y 100%. Evidentemente las confianzas más solicitadas serán, digamos del 80% en adelante.

Y esto debido al hecho de que si en la urna tenemos 99 canicas verdes y una roja, al extraer una aleatoriamente, pues, apostamos a verde. Si en lugar de tener una roja tenemos 5 ó 10 de ellas y el

complemento a 100 de color verde, sin duda seguiríamos apostando a que la canica extraída es verde, ya que en el peor de los casos hay 90 verdes contra 10 rojas. Si la proporción la movemos ahora a 80 verdes y 20 rojas, posiblemente sigamos apostando a verde. Pero si ahora ponemos 60 verdes y 40 rojas o 50 verdes y 50 rojas ya no es fácil tomar la decisión para la apuesta. En el último caso tanto las verdes como las rojas tienen probabilidad de un medio y se vuelve una lotería. Por ello decimos que las confianzas más empleadas serán del 80% en adelante, digamos.

Como la desviación estándar del estimador está dada por el cociente entre el error estipulado y el valor de la abscisa t en la distribución normal que nos deja en la parte central de la curva una área igual a la confianza especificada se verifica que

$$V = \left(\frac{d}{t}\right)^2$$

Entonces, otra manera de especificar la precisión con la que se desea una estimación es indicando una desviación estándar o una variancia deseada para nuestro estimador.

Cuadro 4.1 Maneras usuales de referirse a la precisión de un estimador o a la precisión con la que se desea una estimación.

Precisión
- error d y confianza C
- desviación estándar del estimador
- variancia del estimador
- recursos disponibles

En el cuadro 4.1 ha sido adicionada una manera de especificar la precisión del estimador y ésta es en términos de los recursos disponibles para el estudio. Tiene sentido, ya que en muchas ocasiones uno estima tamaños de muestra tales que no es posible desarrollar el estudio debido al alto costo que ello implica. Entonces el problema se plantea de la manera siguiente: dispongo de cierto

dinero y deseo desarrollar una encuesta para estimar tal (es) parámetro (s). ¿Para qué precisión me alcanza?

4.2 TAMAÑO DE LA MUESTRA PARA LA ESTIMACION DE MEDIAS

Si queremos que nuestro estimador $\bar{y}$ de la media poblacional $\bar{Y}$ tenga a lo más una variancia $V = (\frac{d}{t})^2$, se debe verificar que

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

es decir,

$$\frac{S^2}{n} - \frac{S^2}{N} \leq V,$$

y despejando n tenemos:

$$n \geq \frac{S^2}{V + \frac{S^2}{N}} = \frac{\frac{S^2}{V}}{1 + (\frac{1}{N})\frac{S^2}{V}} \qquad 4.1$$

Como deseamos tomar un tamaño de muestra no mayor que el necesario, en la expresión 4.1, usamos la igualdad. El denominador de 4.1 tiende a la unidad cuando N es grande, por lo cual usamos al numerador como primera aproximación al tamaño de la muestra. Esto es:

$$n_0 = \frac{S^2}{V} = \frac{S^2 t^2}{d^2}$$

que se suele corregir así:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \qquad 4.2$$

Las expresiones 4.2 constituyen las ecuaciones necesarias para encontrar el tamaño de la muestra cuando se desea estimar un *valor medio.* Primero se determina n_0 y posteriormente se corrige este valor con la ecuación para la n. La n así determinada es el tamaño de la muestra necesario.

Ejemplo 4.1 En un archivero hay 60 expedientes, los cuales contienen un número variable de hojas cada uno, un censo practicado en ellos muestra lo siguiente:

Tabla 4.1

Expediente	No. hojas	Expediente	No. hojas	Expediente	No. hojas
1	1	21	5	41	3
2	1	22	5	42	5
3	1	23	6	43	4
4	7	24	3	44	6
5	1	25	4	45	1
6	3	26	2	46	3
7	5	27	2	47	3
8	5	28	1	48	3
9	4	29	4	49	2
10	2	30	2	50	1
11	3	31	2	51	2
12	3	32	2	52	2
13	6	33	3	53	1
14	9	34	2	54	3
15	3	35	8	55	5
16	5	36	9	56	3
17	7	37	5	57	3
18	1	38	1	58	3
19	1	39	1	59	2
20	3	40	3	60	2

$$\Sigma\, y_i = 198, \qquad \Sigma\, y_i^2 = 900$$

El número medio de hojas por expediente es:

$$\bar{Y} = \frac{1}{N}\sum^{N} y_i = \frac{198}{60} = 3.3 \text{ hojas/expediente}$$

La variancia poblacional S^2 vale:

$$S^2 = \frac{\sum^{N} (y_i - \bar{Y})^2}{N-1} = \frac{\sum^{N} y_i^2 - N\,\bar{Y}^2}{N-1} = \frac{900 - 60(3.3)^2}{60-1} =$$

$$= 4.18 \text{ (hojas)}^2$$

Queremos encontrar el tamaño de muestra necesario para estimar el número medio de hojas por expediente con un error no superior al 20% *del valor de* $\bar{Y}$ y una confianza del 95% o sea:

$d = 20\%\ \bar{Y} = (0.2)\ (3.3) = 0.66$ hojas y $t = 1.96$. Por comodidad usaremos $t = 2$.

$$n_0 = \frac{S^2}{(\frac{d}{t})^2} = \frac{4.18}{(\frac{0.66}{2})^2} = 38.4$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{38.4}{1 + \frac{38.4}{60}} = \frac{38.4}{1 + 0.64} = 23.4$$

Claramente el número de expedientes o en general de unidades en la muestra debe ser un número entero, entonces el resultado 23.4 debe ser redondeado al entero inmediato superior para estar seguros de que el tamaño de muestra elegido sea suficiente para los propósitos del estudio, así $n = 24$. En este caso actúa fuertemente la fracción $\frac{n_0}{N}$ en el cálculo de n, la fracción de muestreo es $f = \frac{n}{N} = \frac{24}{60} = \frac{1}{2.5}$. Si N fuera igual a 6 000, el tamaño de la muestra sería $n = 39$ y la fracción de muestreo $f = \frac{39}{6\,000} = \frac{1}{154}$.

Por otra parte, si en lugar del error anterior del 20% pedimos el 5% se tiene $d = 5\%\ (3.3) = 0.165$ y para la misma confianza, obtenemos $n_0 = 614.1$ y $n = 55$. Es natural el aumento necesario en el tamaño de muestra ya que se requiere mayor precisión.

4.3 TAMAÑO DE MUESTRA PARA LA ESTIMACION DE TOTALES

Queremos que el estimador $N\bar{y}$ del total poblacional Y tenga a lo más una variancia $V = (\frac{d}{t})^2$, entonces:

$$V(N\bar{y}) = \frac{N^2 \cdot S^2}{n}\left(1 - \frac{n}{N}\right) \leqslant V = \left(\frac{d}{t}\right)^2 \qquad \text{luego:}$$

$$n_0 = \frac{N^2 \cdot S^2}{V}; \quad n = \frac{n_0}{1 + \frac{n_0}{N}} \qquad 4.3$$

De manera similar a la determinación del tamaño de muestra para medias (apartado 4.2), en la situación actual de la determinación del tamaño n de muestra para *totales*, usando las ecuaciones 4.3 calculamos n_0 y posteriormente corregimos este valor con la expresión para n.

Ejemplo 4.2 En la tabla 4.1 el total de hojas es: $Y = \Sigma\ y_i = 198$. Si deseamos estimar este parámetro con un error d no mayor al 20% de Y entonces $d = 0.2(198) = 39.6$ hojas, y considerando una confianza del 80%, $t = 1.28$, el tamaño de muestra sería:

$$n_0 = \frac{(60)^2\ (4.18)}{(\frac{39.6}{1.28})^2} = 15.72$$

$$n = \frac{15.72}{1 + \frac{15.72}{60}} = 12.46$$

Por lo que elegimos $n = 13$.

4.4 TAMAÑO DE MUESTRA PARA LA ESTIMACION DE PORCENTAJES

Análogamente a los casos presentados en 4.3 y 4.4, deseamos estimar un porcentaje P de manera que nuestro estimador de porcentajes p tenga a lo más una variancia $V = (\frac{d}{t})^2$, entonces, en base al criterio de normalidad:

$$V(p) = \frac{P\,Q}{n}\,\frac{N - n}{N - 1} \leqslant V = \left(\frac{d}{t}\right)^2$$

y usando N en vez de $N - 1$:

$$\frac{PQ}{n} - \frac{PQ}{N} \leqslant V,$$

de donde:

$$n_0 = \frac{PQ}{V};\ n = \frac{n_0}{1 + \frac{n_0}{N}} \qquad 4.4$$

En las expresiones 4.4, al determinar n_0 es necesario que tanto P como Q estén en porcentajes y similarmente dado que $V = (\frac{d}{t})^2$, el error d también debe estar en porcentajes.*

Cuadro 4.2

¿El problema es determinar el tamaño de muestra para estimar:	
	... una media? , use $\frac{S^2}{V}$
	... un total? , use $\frac{N^2 S^2}{V}$
	... una proporción? , use $\frac{PQ}{V}$

Ejemplo 4.3 En la tabla 4.1 el porcentaje de expedientes con una hoja es $P = \frac{12}{60}100 = 20\%$. Si deseamos estimar este porcentaje con un error d no mayor al 20 por ciento de P, entonces: $d = 0.20\ (20) = 4\%$ y empleando una confianza del 95% tenemos:

Cálculo de n_0 en porcentajes:

$$n_0 = \frac{20(80)}{(\frac{4}{2})^2} = 400$$

Cálculo de n_0 en por unidad:

$$n_0 = \frac{(0.20)\ (0.80)}{(\frac{0.04}{2})^2} = 400^{**}$$

$$n = \frac{400}{1 + (\frac{400}{60})} = 52.17$$

Por lo cual elegimos $n = 53$. Dado que la población es de tamaño 60, generalmente en términos prácticos esto significa que hay que censar. En la tabla 4.2 aparecen los tamaños de muestra y las fracciones muestrales que se requerirían para diferentes tamaños

* Las expresiones 4.4 también pueden ser usadas cuando P, Q y d están en por unidad.

** Para efecto de claridad, se presentan los cálculos de n_0 cuando P, Q y d están expresados en porcentajes y cuando lo están en por unidad, en este último caso P y Q toman valores entre 0 y 1.

de la población, a fin de obtener el porcentaje de error esperado, con su respectivo nivel de confianza.

Tabla 4.2

$P = 20\%$ $d = 4\%$ $t = 2$

N	n	$f = \frac{n}{N}$	*En promedio una unidad de cada:*
600	240	240/600	2.5
6 000	375	375/6 000	16
60 000	398	398/60 000	151
600 000	400	400/600 000	1 500
6 000 000	400	400/6 000 000	15 000

Debemos notar que al variar el valor del porcentaje P, el producto PQ que aparece en el numerador de la expresión 4.4 tiene una variación como se muestra en la figura 4.2. La gráfica es simétrica con respecto a $P = 50\%$ valor en el cual también alcanza el máximo y para otros valores de P el producto va disminuyendo hasta que se anula para $P = 0\%$ y para $P = 100\%$.

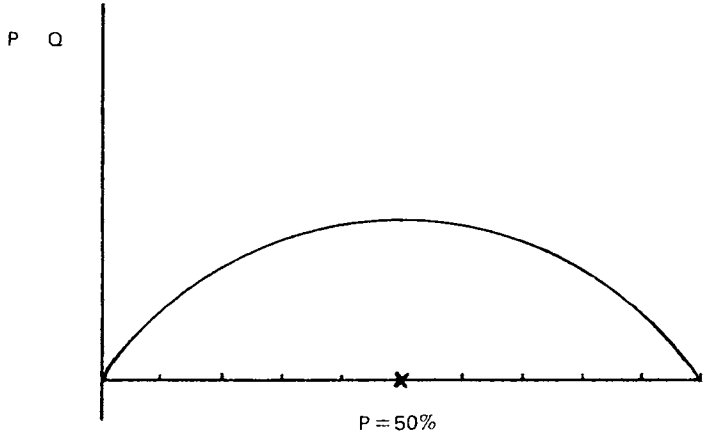


Figura 4.2

Entonces, si tenemos dos o más valores posibles de P, aquel que hace máximo el producto PQ es el valor que está más cerca de $P = 50\%$. Como veremos posteriormente, esto es importante cuan-

do se desea determinar n y sólo se tiene una idea de un intervalo en el cual se puede encontrar el porcentaje P.

4.5 ESTIMACIONES DE VARIANCIAS

Dado un problema de estimación de parámetros poblacionales mediante muestreo probabilístico, es pertinente determinar sin ambigüedad qué parámetro está en estudio: una media, un total, una proporción. Si éste no es evidente, hay que preguntar al usuario de la información cómo lo define poblacionalmente. Al hacer esto, su identificación resulta inmediata o las incongruencias de definición se hacen evidentes. Una vez que se le tiene identificado y suponiendo muestreo aleatorio simple, hay que localizar o determinar la expresión correspondiente para efectos de cálculo del tamaño de muestra.

En seguida debemos considerar la precisión con que nos piden la estimación. Hemos visto cuatro maneras de especificarla:

i) mediante un error y un nivel de confianza;
ii) mediante una desviación estándar,
iii) mediante una variancia, y
iv) mediante la especificación del presupuesto disponible.

Al tratar de remplazar los símbolos por valores en las fórmulas para obtener el tamaño de la muestra nos damos cuenta de que éstas se encuentran en términos de parámetros poblacionales (S^2, P), los cuales sólo se pueden conocer a través de un censo.

El procedimiento a seguir, es sustituirlos por valores estimados. Esto requiere de información adicional; preguntas adicionales al usuario de las estimaciones, estudios adicionales o auxiliares en la población objetivo, uso de información arrojada por encuestas anteriores y, a veces, conjeturas sobre la población.

En el caso de porcentajes, se puede preguntar al interesado (usuario), acerca del rango de valores en que cree que se encuentra el porcentaje que le interesa, pues casi siempre se tiene alguna idea de ello.

Si se puede suponer normalidad en la distribución de la característica en estudio, se puede preguntar por los valores extremos que se dan en la población y con ello estimar gruesamente el rango de variación. Como en seis desviaciones estándar está contenida prácticamente la distribución, se puede hacer $k \cdot S$ igual al rango de

variación y de ahí despejar S. En lugar de tomar $k = 6$, se suele tomar un valor más conservador como $k = 4$.

Un procedimiento más seguro y frecuentemente usado para estimar variancias, consiste en una *prueba piloto.* Se introduce en la población una muestra aleatoria pequeña y en base a ella, se estiman las variancias. En realidad, la prueba piloto puede arrojar más resultados; nos permite probar todo o casi todo el mecanismo técnico de muestreo que se diseñó y nos permite estimar tiempos y costos. Comprobamos si se pueden localizar prácticamente las unidades en la muestra como indica el método de selección; comprobamos si el método de medición es práctico y arroja la precisión necesaria; comprobamos si las formas de vaciado o cuestionario son entendibles en el lenguaje que hablan las unidades muestra, y si su redacción es la adecuada, y si en sí, es práctico en el sentido de que sea manuable y no extenso (Chevry). Y por último, da lineamientos para el entrenamiento a los encuestadores.

El éxito de una encuesta no queda asegurado por la elaboración de un método de selección y uno de estimación. El trabajo colateral, como es la elaboración de cuestionarios, manuales y entrenamiento al personal, así como la coordinación general del trabajo de campo, es vital para el buen término de la encuesta.

4.6 CUESTIONARIOS CON VARIAS PREGUNTAS Y EL CASO DE PRECISIONES POR SUBDIVISION

Generalmente, las encuestas se desarrollan para estimar varios parámetros; en consecuencia, el cuestionario está integrado por varias preguntas. En este caso, para efectos del tamaño de muestra hay que considerar dos aspectos:

i) Existe un único parámetro poblacional o una única pregunta que gobierna al tamaño de muestra y las demás preguntas se introducen sin importar la precisión con que ellas sean estimadas.

ii) Existen varias preguntas que son importantes y cada una de ellas debe ser estimada con determinada precisión; entonces se calcula el tamaño de muestra para cada una de ellas y se elige el mayor.

Claramente con este criterio, el tamaño de muestra será el apropiado para la pregunta que requiere el máximo, y el resto de

ellas lo tendrá superior al necesario, luego, se les estimará con una precisión mayor a la requerida.

Al hablar de la precisión es necesario distinguir dos casos:

i) Interesa obtener una estimación con determinada precisión a nivel poblacional.

ii) Interesa obtener una estimación para fracciones o subdivisiones de la población; son en realidad varias estimaciones sobre la misma característica. Por ejemplo, se desea estimar el sueldo medio de los empleados en una institución, por departamento, con una determinada precisión para cada uno de ellos. Hacemos una estimación del tamaño de muestra para cada subdivisión* y el tamaño de muestra global será la suma de ellos. La selección se puede hacer al menos de dos formas, en una la hacemos según muestreo aleatorio simple sobre todo el listado de empleados y, en la otra, como se verá en el capítulo 6, el esquema más apropiado es el muestreo estratificado.

Debemos notar que en el primer caso, cuando se practica la selección aleatoria sobre el listado de empleados de todos los departamentos; en promedio tenderán a ser elegidas tantas unidades de cada departamento como sea la proporción o el porcentaje de unidades que lo componen con respecto al total en la lista o en la institución. Por ejemplo, si elegimos 100 números aleatorios entre 1 y 1 000, en promedio esperaríamos que se seleccionaran $\frac{200}{1\,000}\,100 = 20$ números aleatorios entre 150 y 350.

Ejemplo 4.4 Un economista desea hacer un estudio sobre los profesores de una universidad en referencia a la cantidad de dinero por semana que cada profesor dedica a la alimentación de su familia. Para ello ocurre a los niveles administrativos correspondientes, a fin de conseguir un listado de los 2 000 maestros que trabajan en la institución. Los niveles administrativos superiores también tienen necesidad de obtener alguna información entre los maestros y se ponen de acuerdo para aprovechar esa encuesta y para introducir otras 51 preguntas adicionales, las cuales son las siguientes:

* El estudio de este tema requiere mayor detalle estadístico matemático que el pretendido en este libro; se recomienda ver el libro *"Sampling Techniques"*, 1963, W.G. Cochran, J. Wiley and Sons, N. Y. segunda edición.

1. ¿Cuánto dinero dedicó usted a la alimentación de su familia la semana pasada?
2. ¿De cuántos alumnos desearía usted que fueran sus grupos?
3. Para cada clase ¿lleva usted un único libro de texto?
4. ¿En qué sector de la ciudad vive usted?
5. ¿Cuántos idiomas habla usted aparte de su lengua materna?
6. En el semestre anterior, ¿pudo usted concluir la enseñanza de las materias a su cargo?
7. Si pudiera hacerlo, ¿cambiaría usted el programa de estudios?
8. ¿Tiene usted asignado un cubículo?
9. ¿El salón de clases que le corresponde es confortable?
10. ¿Usa automóvil para venir a la universidad?
11. ¿Qué otros empleos tiene usted aparte del de profesor de la universidad?
12. ¿Asistió usted a la XIIa. Exposición sobre Material Didáctico?
13. ¿A qué asociaciones profesionales pertenece?
14. Si la universidad aumentara los sueldos a sus maestros en 30% ¿Le convendría a usted dedicarse de tiempo completo a las labores académicas?
15. ¿Cuál es su opinión sobre el sistema de universidad abierta?
16. ¿En alguna ocasión ha tomado usted el curso que sobre lectura dinámica ofrece la universidad a sus empleados?
17. ¿Está usted empadronado?
18. Debido a la insuficiencia actual de nuestros edificios, ¿preferiría usted que se redujera el espacio de la sala de lectura de la biblioteca o el área de cubículos de profesores?
19. En promedio ¿cuánto tiempo dedica usted a la preparación de sus clases?

20. En su primer ingreso a la universidad ¿cuánto tiempo esperó para cobrar su primer pago?
21. ¿Considera usted que actualmente trabaja sobre la profesión en la que se graduó?
22. ¿En qué está especializado actualmente?
23. Si usted tuvo inasistencias durante el último semestre ¿cuál fue el principal motivo de ellas?
24. ¿Trabaja actualmente o ha trabajado en alguna ocasión en alguna industria o empresa privada?
25. ¿Cuál es su opinión sobre el sistema de enseñanza vigente actualmente en su Escuela o Facultad?
26. ¿Cuál es su opinión sobre el sistema educativo nacional?
27. En su área de enseñanza, ¿cuántos maestros cree usted que faltan a nivel nacional?
28. Se tiene la intención de organizar un congreso sobre educación superior, ¿participaría usted activamente?
29. ¿Ha viajado al extranjero en los últimos cinco años?
30. Si usted no tiene estudios de posgrado, ¿le gustaría llevar a cabo alguno en nuestra universidad?
31. ¿Ha hecho publicaciones científicas en México en alguna ocasión?
32. Usualmente, ¿en qué emplea los fines de semana?
33. ¿Acostumbra usted a leer antes de dormirse? ¿Qué tipo de literatura?
34. ¿Ha tomado algún curso sobre metodología pedagógica?
35. ¿Ha estado becado en alguna ocasión por nuestra universidad?
36. ¿Qué universidad nacional considera usted que es la de mayor prestigio?

37. ¿En qué lugar catalogaría usted a nuestra universidad comparándola con las de todo el mundo?

38. ¿En qué lugar la catalogaría usted dentro de las universidades latinoamericanas?

39. ¿Cuántas recámaras tiene su vivienda y cuántas camas hay en cada una de ellas?

40. Para el diseño y manufactura de su material didáctico, ¿le proporciona la universidad todo el material que necesita? ¿Cómo o dónde aprendió a diseñarlo?

41. ¿Ha asistido a algún curso sobre métodos para impartir clases? ¿Dónde y en qué fecha? ¿Quién se lo patrocinó?

42. Está usted satisfecho con el método que por reglamento seguimos para asignar calificaciones?

43. Usualmente ¿de qué tipo son sus exámenes? (orales, escritos, de otro tipo).

44. Existe un convenio con universidades sudamericanas para intercambio de profesores sobre diferentes áreas del conocimiento con duración aproximada de seis meses. Si lo propusiéramos a usted como candidato ¿estaría usted dispuesto a viajar para el año entrante?

45. ¿Le han satisfecho los cursos que sobre funcionamiento y uso de las computadoras electrónicas ha impartido la universidad a su personal docente? Actualmente se piensa estructurar uno más con duración de tres meses, ¿asistiría usted?

46. ¿Qué cursos de extensión profesional le gustaría tomar en el año siguiente?

47. ¿Ha usado en alguna ocasión nuestro Centro de idiomas? ¿Le parece adecuada su localización?

48. ¿Ha leído en alguna ocasión algún artículo sobre la reforma educativa en México? ¿Dónde lo leyó?

49. ¿Ha leído algún artículo sobre el origen de la estructura educativa mexicana en el siglo pasado?

50. ¿Cree usted que la educación superior en México es una manera de obtener poder y de enriquecerse? ¿Y en el mundo entero?

51. En promedio, ¿cuántos años piensa usted que nuestros egresados trabajan sobre su especialidad profesional?

52. ¿Después de cuántos años piensa usted que nuestros egresados logran un empleo adecuado a su profesión?

Una vez que tienen estructurado este cuestionario, le piden a un estadístico que les estime el tamaño de muestra que se debe usar. Para ello, el estadístico los interroga sobre cuál es la pregunta más importante en el estudio, a lo cual responden que todas por igual. El estadístico vuelve a hacer la misma pregunta formulada de una y de otra manera y al final el grupo llega a la conclusión de que las preguntas verdaderamente importantes son las 7 primeras.

Con esta información, analiza el cuestionario, formula más preguntas y llega a las siguientes conclusiones:

Con la pregunta 1, se intenta estimar la cantidad media de dinero que los profesores dedican a la alimentación de sus familias, entonces se trata de estimar una media.

Con la pregunta 2, se intenta estimar el tamaño medio ideal de los grupos, se trata de una media.

Con la pregunta 3, se quiere estimar el porcentaje de maestros que siguen un único libro de texto, se trata de un porcentaje.

Con la pregunta 4, se desea estimar principalmente el porcentaje de sus maestros que viven en determinada zona de la ciudad, se trata de un porcentaje.

Con la pregunta 5, se desea estimar principalmente el porcentaje de maestros que sólo hablan el idioma castellano, también ésta es un porcentaje.

Con la pregunta 6, se desea estimar el porcentaje de maestros que cubrieron íntegramente sus programas de estudio en el semestre anterior, se trata de un porcentaje.

Y por fin, con la última pregunta, se desea estimar el porcentaje de maestros inconformes con sus programas de estudio, se trata de un porcentaje.

El paso siguiente, es la discusión de las precisiones con que se desea cada estimación. El economista dice que la media que a él le interesa debe encontrarse en alrededor de $ 1 000, ya que la mayoría de los maestros son casados, entre 30 y 50 años de edad y el nivel de sueldos de la universidad es bueno. Además añade que él ha hecho algunos sondeos que arrojan resultados congruentes con sus supuestos. Con estos datos el estadístico estima gruesamente una variancia S^2 de 50 000 y, como se pide un error del 5% y una confianza del 95%, según la expresión 4.2 obtiene:

$$n_0 = \frac{50\,000}{(\frac{50}{2})^2} = 80$$

de donde:

$$n = \frac{80}{1 + \frac{80}{2000}} \doteq 77$$

Similarmente, para la media de la pregunta 2 se obtiene un tamaño de muestra de 98. Al preguntar sobre el porcentaje de la pregunta 3, le indican que éste debe encontrarse entre 30 y 60% y que la desean estimar con un 5% de error y un 95% de confianza. Entonces el tamaño de muestra se obtiene con la expresión 4.4

$$n_0 = \frac{pq}{V} = \frac{50\ (50)}{(\frac{5}{2})^2} = 400$$

$$n = \frac{400}{1 + \frac{400}{2\,000}} \doteq 334$$

De la misma manera se trabaja el resto de porcentajes hasta obtener estos resultados:

Pregunta	1	2	3	4	5	6	7

Tamaño de muestra	77	98	334	300	200	200	370

Se concluye que el tamaño de muestra a usar es de 370, superior en 293 unidades al tamaño de muestra que se requeriría para obtener exclusivamente la estimación que pretendía el economista.

En este ejemplo se ha tratado de ilustrar la situación usual en muchas encuestas, acerca de que cuando se va a desarrollar una de ellas, los interesados tienden a introducir más y más preguntas en los cuestionarios con la salvaguarda de que "ya que van a visitar a tales personas, pregunten de paso tales y cuales cosas".

Sin tocar de momento los fuertes problemas a que da lugar un cuestionario extenso, se trata de ilustrar la habilidad de que debe hacer gala el estadístico para detectar cuáles preguntas son introducidas sin tener una importancia capital, y cuándo la tienen verdaderamente. El hecho es que cuando se interroga sobre la pregunta importante, los interesados contestan que todas lo son, y que de no ser así, "no se hubieran formulado". En realidad, algunas de ellas son debidas a simple curiosidad y, en otras, el uso que se les va a dar no es de tal importancia como para que ellas gobiernen a la encuesta. No es raro ver encuestas con cuestionarios formados por varias decenas de hojas. Y desde el punto de vista de la entrevista o, en general, del método que se emplee para recabar la información es deseable que sea lo más reducido posible.

4.7 EJERCICIOS

4.1 Obtenga la expresión 4.3 para el tamaño de la muestra en el caso de la estimación de totales.

4.2 Obtenga la expresión 4.4 para el tamaño de la muestra en el caso de la estimación de porcentajes.

4.3 En el ejemplo 4.1 se encontró un tamaño de muestra de 24 para estimar el número medio de hojas por expediente, con un error que no excede al 20% y a una confianza del 95%.

Usando las tablas de números aleatorios (tabla 3.1) podemos materializar una muestra. Comenzando en la esquina superior izquierda y continuando posteriormente hacia abajo tenemos la tabla 4.3:

Tabla 4.3

Expediente	*No. hojas*	*Expediente*	*No. hojas*	*Expediente*	*No. hojas*
1	5	9	5	17	3
2	4	10	2	18	6
3	2	11	1	19	2
4	3	12	2	20	1
5	2	13	2	21	5
6	3	14	1	22	3
7	2	15	1	23	5
8	9	16	1	24	3

Estime el número medio de hojas por expediente y obtenga intervalos de confianza del 95%.

4.4 En el ejemplo 4.2, para la estimación del total de hojas en los 60 expedientes con un error del 20% y una confianza del 80% se llegó a $n = 13$.

Continuando en las tablas de números aleatorios, a partir del último número usado en el ejercicio 1: *i*) Obtenga una muestra; *ii*) estime el total de hojas, y *iii*) dé una estimación del error estándar.

4.5 La producción, en un día, de tarjetas perforadas de una persona se encuentra en una gaveta; siendo el total de ellas 2 000. Se quiere estimar el porcentaje de tarjetas que tienen al menos un error, mediante una muestra aleatoria. ¿Qué tamaño de muestra es necesario si se piensa que el porcentaje está entre 68 y 80%, y se acepta un error estándar de 3%?

4.6 En un archivo de 10 000 000 de nombres de habitantes del país se desea estimar el porcentaje de ellos cuyo apellido empieza con la letra *K*. El archivo no está ordenado alfabéticamente.

Considerando que el porcentaje es de aproximadamente 0.5 por ciento se pide encontrar el tamaño de muestra requerido bajo muestreo aleatorio simple si se acepta un error estándar no mayor de 0.05%.*

4.7 Para efectos de una planeación económica en la región occidental de México, es necesario estimar de entre 10 000 establos: *a*) el número medio de vacas lecheras por establo con un error del 10% y una confianza del 95%; y *b*) el rendimiento medio de leche por establo con un error del 10% y una confianza del 95%.

Una muestra aleatoria piloto de tamaño 20 arrojó las siguientes estimaciones:

* Cuando el atributo buscado es raro, un esquema de muestreo de tipo general como es el aleatorio ya no es eficiente, requiere tamaños de muestra muy grandes y es necesario recurrir a otros métodos.

Número medio de vacas por establo igual a 40,
s^2 igual a 1 000.
Rendimiento medio de leche por establo igual
a 300 litros y s^2 igual a 1 600.

¿Qué tamaño de muestra se necesita?

4.8 Si en el ejercicio 4.7 se deseara estimar el número total de vacas lecheras en los 10 000 establos con un error de 30 000 vacas y una confianza del 95%, ¿qué tamaño de muestra se requeriría?

4.9 Una compañía manufacturera de juguetes infantiles desea introducir un nuevo tipo de caballitos, los cuales puede fabricar con uno de dos materiales distintos al mismo costo. Ambos tienen prácticamente, la misma duración y aparentemente el mismo atractivo. Se desea usar una sola clase de material y para tomar una decisión se considera conveniente realizar una pequeña encuesta sobre la zona residencial que es considerada como su mayor cliente y de la cual se tiene un listado reciente de 10 000 familias. Por lo tanto, se trata de una encuesta de opinión, en la cual se entrevistará a la madre en cada familia. ¿Puede indicar algún valor razonable para el porcentaje que se busca, y en función de él, proponer alguna precisión?

CAPITULO 5

MUESTREO ALEATORIO SIMPLE (continuación)

5.1 ESTIMACION DE RAZONES E INTERVALOS DE CONFIANZA

En el apartado 3.4 comenzamos a tratar el caso de la estimación de cocientes o razones. En este capítulo volvemos sobre este tema, ya que está asociado a una clase importante de estimadores llamados de razón; los cuales como veremos en 5.2, cuando son usados de manera apropiada, resultan ser mejores que aquellos basados en la media muestral.* Para usar a este tipo de estimadores se requiere de información adicional y se les define como cocientes de variables aleatorias. Recordemos inicialmente la estimación de razones ya introducida en 3.4.

Cuando es necesario estimar la razón o el cociente de dos parámetros poblacionales $R = \frac{Y}{X}$ en base a una muestra aleatoria simple de tamaño n, a cada unidad en la muestra se le hacen dos mediciones cuyos resultados para la unidad i-ésima simbolizamos por y_i y x_i; el estimador que usamos para ello es el presentado en la ecuación 3.4 y que repetimos en seguida:

$$\hat{R} = \frac{\sum_{n}^{n} y_i}{\sum_{n}^{n} x_i} = \frac{\frac{\sum_{n}^{n} y_i}{n}}{\frac{\sum_{n}^{n} x_i}{n}} = \frac{\bar{y}}{\bar{x}} \qquad 5.1$$

*En el cual el denominador es una constante.

Este es consistente pero sesgado, y se puede demostrar que el sesgo disminuye a medida que el tamaño de la muestra aumenta. Si se expande según la serie de Taylor y se toma de ella la aproximación lineal se encuentra que $\hat{R}$ es insesgado bajo ella, y que su variancia está dada por (ejercicios 5.3 y 5.4):

$$V(\hat{R}) \doteq \frac{1-f}{n\bar{X}^2} \frac{\sum^{N} (y_i - R\,x_i)^2}{N-1} \qquad 5.2$$

Un estimador de esta variancia es el siguiente:

$$\hat{V}(\hat{R}) \doteq \frac{1-f}{n\bar{X}^2} \frac{\sum^{n} (y_i - \hat{R}\,x_i)^2}{n-1}$$

$$\hat{V}(\hat{R}) = \frac{1-f}{n\bar{X}^2} \frac{\sum^{n} y_i^2 - 2\hat{R}\sum^{n} y_i x_i + \hat{R}^2 \sum^{n} x_i^2}{n-1}^{*} \qquad 5.3$$

Para tamaños de muestra no muy pequeños (al menos 30), consideramos apropiada la aproximación normal para el estimador $\hat{R}$ y con ella, también las expresiones anteriores para las variancias.

Ejemplo 5.1 En un estudio desarrollado al sur de la ciudad de México, en una zona formada por 70 manzanas, se listaron a las 3 000 familias que la componían, y se eligieron aleatoriamente a 40 de ellas. A cada familia en la muestra se le preguntó sobre el número de miembros y de autos que tenía. Las respuestas fueron las de la tabla 5.1 en la cual:

y_i: número de miembros
x_i: número de autos

* Si la media poblacional $\bar{X}$ es desconocida, usamos en su lugar a la media muestral $\bar{x}$.

Tabla 5.1

y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i
3	3	4	2	7	7	6	5
5	2	4	2	9	4	10	3
6	2	5	3	7	4	6	3
5	5	7	4	7	3	4	4
5	5	9	2	8	2	5	4
4	1	5	2	6	2	7	4
5	3	5	1	6	3	6	3
5	2	6	1	8	1	6	5
6	2	6	3	5	1	9	2
3	3	5	3	5	2	6	2

$\Sigma\, y_i = 236;$ $\Sigma\, x_i = 115$

$\Sigma\, y_j^2 = 1494;$ $\Sigma\, x_j^2 = 401$

$$\Sigma\, y_i \circ x_i = 685$$

Se desea estimar el número de miembros por auto en las 3 000 familias; entonces:

$$\hat{R} = \frac{236}{115} = 2.05 \text{ miembros por auto}$$

Para calcular el error estándar usamos la ecuación 5.3, y como en ella aparece $\bar{X}$, desconocida en este caso, en su lugar usamos su estimador $\bar{x}$:

$$\hat{V}(\hat{R}) = \frac{1-f}{n\bar{x}^{-2}} \frac{\Sigma\, y_i^2 - 2\hat{R}\,\Sigma\, y_i x_i + \hat{R}^2\,\Sigma\, x_i^2}{n-1} =$$

$$= \frac{1 - (\frac{40}{3\,000})}{40\,(\frac{115}{40})^2} \frac{1\,494 - 2(2.05)\,(685) + (2.05)^2\,(401)}{40-1}$$

$$= 0.028$$

Y el error estándar es $(0.028)^{1/2} = 0.167$ miembros/auto.

Considerando adecuada la aproximación normal para $\hat{R}$, los intervalos de confianza para el caso de razones o cocientes quedan dados por:

$$\left\{\hat{R} - t\,(\hat{V}(\hat{R}))^{1/2}\ ,\ \hat{R} + t\,(\hat{V}(\hat{R}))^{1/2}\right\} \qquad 5.4$$

Entonces con la expresión 5.4 podemos calcular intervalos del 95% de confianza para el cociente número de miembros por auto en el ejemplo 5.1:

$$\{2.05 - 2(0.167) \quad , \quad 2.05 + 2(0.167)\}\ ,$$

$$\{1.716\ ,\ 2.384\}$$

Ejemplo 5.2 Un comerciante que se dedica a la compra de cosechas de frijol está interesado en estimar el peso medio de piedras por saco que vienen en un lote de 1 000 de ellos, ya que ha encontrado en algunos de ellos hasta tres kilos y medio de piedras revueltas con la semilla y desea disminuir el pago por saco proporcionalmente al peso medio de los elementos extraños en cada uno de ellos. También desea estimar la relación entre el peso total de piedras en kilos al peso total de la semilla en los 1 000 sacos.

Los sacos se encuentran colocados en 25 hileras de 40 sacos cada una. Una columna del edificio, cercana a la primer fila le sirve de referencia para marcar al saco número 1. Después el comerciante adopta una manera de asignar un número consecutivo a cada uno del resto de ellos y elige a 25 aleatoriamente.

A cada saco en la muestra lo marca, lo separa del resto de ellos, lo abre y vacía su contenido en un par de cribas separando así la semilla y la materia extraña, posteriormente pesa a cada uno de ellos por separado y obtiene los resultados de la tabla 5.2:

Para estimar el peso medio de piedras por saco usamos como estimador a la media muestral:

$$\bar{y} = \frac{53}{25} = 2.12 \text{ kilogramos de piedra por saco.}$$

Una estimación del peso total de las piedras en todos los sacos la obtenemos mediante $N\bar{y}$:

Tabla 5.2

No. de saco	*Peso de las piedras en kilos*	*Peso de la semilla en kilos*
1	2	47
2	2.5	46
3	1.5	46
4	1.5	48
5	2	47
6	3	50
7	2	50
8	2.5	51
9	2	49
10	2	49
11	2	48
12	2	50
13	2	49
14	2.5	48
15	3	45
16	2.5	47
17	2	46
18	3	49
19	2	50
20	1.5	48
21	1	48
22	2	48
23	2	46
24	2.5	50
25	2	50
Totales	53	1 205

$\hat{Y} = 1\,000(2.12) = 2\,120$ kilogramos de piedra

En el ejercicio 5.6 se pide estimar intervalos de confianza del 95% para el peso medio de piedras por saco y para el total.

Por tratarse de un cociente, la relación entre el peso total de piedras al peso total de la semilla la estimamos mediante la expresión 5.1:

$$\hat{R} = \frac{53}{1205} = 0.044 \text{ kilos de piedra por kilo de semilla.}$$

Y su error estándar se pide que sea calculado en el ejercicio 5.7.

5.2 ESTIMACION DE MEDIAS Y DE TOTALES CON ESTIMADORES DE RAZON

En esta sección mantendremos fijo el método de selección que hasta ahora hemos venido empleando, a saber, muestreo aleatorio simple, pero introduciremos una modificación en el método de estimación, y los propósitos de éste serán los usuales: estimación de medias y totales.*

En el apartado 3.4 vimos cómo estimar una media y un total a través de la media muestral $\bar{y}$ y de $N \cdot \bar{y}$. En algunas ocasiones se dispone de información adicional que pueden proporcionar las mismas unidades muestrales y que puede ser usada para mejorar la estimación, es decir, para hacerla más precisa. La información adicional o auxiliar de que se habla está constituida por los valores particulares de cierta característica de las unidades de muestreo, la cual puede o no ser colectada al mismo tiempo que la característica actual en estudio; pero, para que sea valiosa a nuestros fines, debe estar correlacionada positivamente con la característica en estudio. Mientras mayor sea la correlación, más precisa será la estimación.

Por ejemplo, si se mide a niños recién nacidos, se obtiene para sus tallas $x_1, x_2, \ldots, x_N$ centímetros cada niño. Siete meses más tarde se les vuelve a medir y se obtiene $y_1, y_2, \ldots, y_N$ centímetros cada niño. En este ejemplo, la relación de magnitudes entre y_i y x_i es del tipo: y_i mayor que x_i, y el cociente $\frac{y_i}{x_i}$ para cada i, será más o menos constante, es decir, las dos variables a los cero y siete meses de edad, están fuertemente correlacionadas positivamente. Si en estas condiciones deseamos estimar la talla o longitud media de los niños a los siete meses de edad en base a una muestra aleatoria simple de tamaño n, usamos como estimador al siguiente:

$$\hat{\bar{Y}}_R = \left(\frac{\sum_{n} y_i}{\sum_{n} x_i} \right) \cdot \bar{X} \qquad 5.5$$

Donde y_i es la talla o longitud del niño i-ésimo a los siete meses de edad y x_i fue la talla o longitud del mismo niño recién nacido en la muestra, y $\bar{X}$ fue la talla o longitud media de los N

* Recordar que un porcentaje puede ser tratado como una media (apartado 3.5).

niños recién nacidos. En general este estimador es mejor que $\bar{y}$, la media muestral de las tallas o longitudes a los 7 meses de edad.

En otro caso, en varios estantes con gavetas tenemos un número variable de tarjetas, una por cada miembro de la familia: padre, madre e hijos, y este tipo de familias tienen generalmente más de dos hijos; deseamos estimar el número total de tarjetas asociadas a los hijos, éstas en cada gaveta están correlacionadas positivamente con el número total de tarjetas en ellas.* Tomamos una muestra aleatoria simple de n gavetas de entre las N que existen en los diferentes muebles; y para cada gaveta en la muestra se cuenta el número dehijos (y_i) y se registra el peso en gramos de las tarjetas en esa gaveta (x_i); y al terminar la encuesta se pesan todas las tarjetas en los estantes o gaveteros (X). En estas condiciones un estimador del número total de tarjetas de hijos es el siguiente:

$$\hat{Y}_R = \frac{\sum_{n} y_i}{\sum_{n} x_i} \cdot X \qquad 5.6$$

En el cual y_i es el número de tarjetas asociadas a los hijos y x_i es el peso en gramos de la gaveta i-ésima en la muestra y, X es el peso en gramos de todas las tarjetas en los N estantes.

Las expresiones de las variancias y de sus estimadores, correspondientes a las ecuaciones 5.5 y 5.6 son inmediatas a partir de 5.1, 5.2 y 5.3 ya que difieren en una constante; entonces la expresión siguiente:

$$V(\hat{\bar{Y}}_R) \doteq \frac{1-f}{n} \frac{\sum^{N} (y_i - R x_i)^2}{N-1} \qquad 5.7$$

es la variancia del estimador de razón de la media $\bar{Y}$, y la expresión siguiente es un estimador de esa variancia.

$$\hat{V}(\hat{\bar{Y}}_R) \doteq \frac{1-f}{n} \frac{\sum^{n} (y_i - \hat{R} x_i)^2}{n-1} \qquad 5.8$$

De manera similar, las expresiones 5.9 y 5.10 constituyen respectivamente la variancia del estimador de razón (5.6) del total Y y un estimador de esa variancia:

* Ver el ejercicio 5.11.

$$V(\hat{Y}_R) \doteq \frac{N^2\ (1-f)}{n} \frac{\sum^{N} (y_i - R\ x_i)^2}{N-1} \qquad 5.9$$

$$\hat{V}(\hat{Y}_R) \doteq \frac{N^2\ (1-f)}{n} \frac{\sum^{n} (y_i - \hat{R}\ x_i)^2}{n-1} \qquad 5.10$$

Como aplicaciones de la técnica y de las expresiones anteriores, en los ejercicios 5.1 y 5.2 se pide que para casos numéricos se completen los ejemplos aquí ya iniciados sobre recién nacidos y sobre gaveteros.

5.3 ESTIMACION DE PORCENTAJES CUANDO HAY MAS DE DOS CLASES

Consideremos el caso del ejemplo 5.1 en el cual se sortearon familias y una de las preguntas se refería a su número de miembros. Supongamos que en este ejemplo es de interés estimar el porcentaje de familias con 4, con 5, con 6, etc. miembros. En estas condiciones la población entera se fragmenta en aquellas familias compuestas por menos de cuatro, cuatro, cinco, seis, etc. miembros. Todas estas clases son disjuntas, es decir, cada familia pertenece a una y sólo a una clase, y al estimar el porcentaje de familias que se encuentran en la clase i-ésima es como si sólo hubiera dos clases: la i-ésima y su complemento a la población total, de manera que el estimador de ese porcentaje es:

$$\hat{P}_i = \frac{a_i}{n} 100 \qquad 5.11$$

Y, similarmente, si se desea estimar el porcentaje de familias con i y j miembros, es decir, el porcentaje de familias que se encuentran en la clase formada por aquellas que tienen i miembros más aquellas que tienen j miembros, consideramos que estos dos tipos o clases de familias forman una nueva clase y el resto de ellas el complemento al total, entonces:

$$\hat{P}_{i+j} = \frac{a_i + a_j}{n} 100 \qquad 5.12$$

En las expresiones 5.11 y 5.12, a_i y a_j simbolizan al número total de unidades en la muestra y que se encuentran en la clase de interés.

Para los estimadores de esta sección, los intervalos de confianza se calculan como en el apartado 3.7.

5.4 ESTIMACION DE VALORES MEDIOS Y DE PORCENTAJES EN DOMINIOS DE ESTUDIO

En muchos casos se desea hacer estimaciones para un dominio de estudio o subpoblación. El sorteo se efectúa sobre todas las unidades de la población de manera que, del tamaño de muestra inicial n, sólo una fracción n_d menor o igual a n y mayor o igual a cero cayó en el dominio de interés, entonces el estimador de la media en el dominio d–ésimo es:

$$\hat{\bar{Y}}_d = \frac{\sum_{k=1}^{n_d} y_{dk}}{n_d} \qquad 5.13$$

siempre que n_d sea mayor que cero. En esta expresión la suma se extiende sólo para aquellas unidades que están en la muestra y que pertenecen al dominio d–ésimo de interés. Puede haber tantos dominios como se quiera, pero como la selección es aleatoria, nada se puede asegurar sobre el tamaño de muestra resultante en cada dominio. Esto puede verse como una restricción de la selección aleatoria. Si uno desea estar seguro de que todos y cada uno de los dominios estén presentes en la muestra, es necesario cambiar de esquema. Veremos más adelante que en ocasiones una selección sistemática (capítulo 7) es la solución para esto, y desde luego, la estratificación juega un papel muy importante en las posibles soluciones.

Como el estimador 5.13 es un cociente de variables aleatorias, lo identificamos como uno de razón. Sin incurrir en un error grave podemos usar la ecuación siguiente como expresión de su variancia:

$$V(\hat{\bar{Y}}_d) = \frac{1 - \frac{n_d}{N_d}}{n_d} \frac{\sum_{k=1}^{N_d} (y_{dk} - \bar{Y}_d)^2}{N_d - 1} = \frac{1 - f_d}{n_d} S_d^2 \qquad 5.14$$

Un estimador de esta variancia es la expresión siguiente:

$$\hat{V}(\hat{\bar{Y}}_d) = \frac{1 - f_d}{n_d} \frac{\sum_{k=1}^{n_d} (y_{dk} - \bar{y}_d)^2}{n_d - 1} = \frac{1 - f_d}{n_d} s_d^2 \qquad 5.15$$

En 5.14 y 5.15 N_d simboliza al número total de unidades que conforman al dominio d-ésimo y $\bar{Y}_d$ y $\bar{y}_d$ son las medias poblacional y muestral respectivamente de ese dominio.

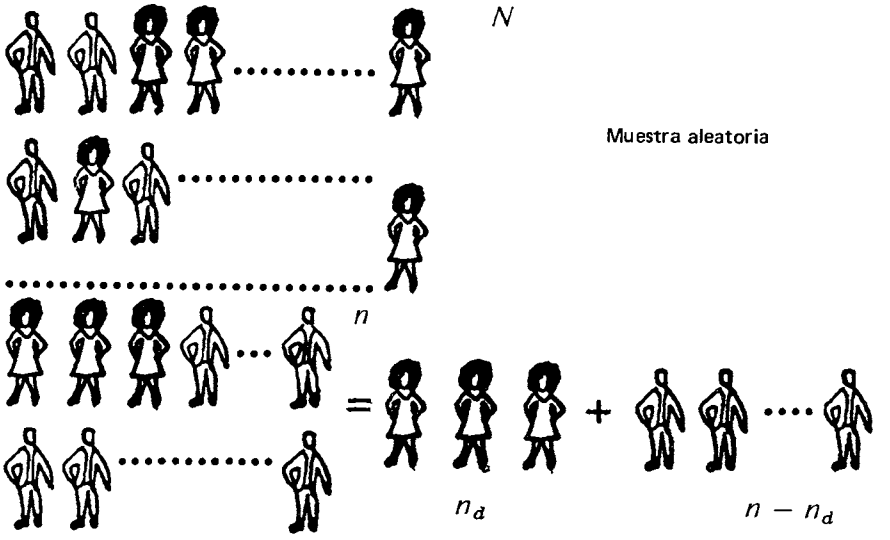


Figura 5.1 En una población formada por personas, un dominio de estudio posible es el de las mujeres; y en una muestra aleatoria de tamaño n *resultaron seleccionadas* $n_d = 3$ *mujeres con valor de su característica en estudio: número de pares de zapatos que poseen:* $y_{d1} = 7, y_{d2} = 5$ y $y_{d3} = 8$.

Usualmente el investigador está seguro de que en la población de tamaño N sujeta a estudio, puede definir alguna subpoblación en particular, pero le resulta imposible conocer al número de unidades N_d que la conforman, entonces no se conoce su tamaño; en esta situación, en la expresión 5.15 en lugar de $\frac{n_d}{N_d}$ que es igual a f_d se puede usar $\frac{n}{N}$ (Ver ejercicio 5.5).

En el caso de la estimación de proporciones en dominios de estudio cada observación vale uno o cero, según que la unidad se encuentre o no en la clase de interés. Así, si del dominio d-ésimo aparecen n_d unidades en la muestra de las cuales n_{dc} pertenecen a la clase de interés, el estimador a usar será:

$$\hat{P}_{dc} = p_{dc} = \frac{n_{dc}}{n_d} = \frac{a_d}{n_d} * \qquad 5.16$$

Y la estructura de su variancia es la misma que aquella en la expresión 5.14 excepto que y_{dk} vale uno o cero y similarmente para el estimador de su variancia en la expresión 5.15.

* Si desea que $\hat{P}_{dc}$ esté en porcentaje, multiplique a la expresión 5.16 por 100.

5.5 ESTIMACION DE TOTALES EN DOMINIOS DE ESTUDIO

Si el estimador de la media poblacional en el dominio d-ésimo es $\hat{\bar{Y}}_d$ definido como en la expresión 5.13, y se desea formar el estimador del total respectivo, éste será:

$$\left.\begin{aligned} \hat{Y}_d &= N_d \hat{\bar{Y}}_d \\ \hat{V}(\hat{Y}_d) &= N_d^2 \frac{1 - f_d}{n_d} s_d^2 \end{aligned}\right\} 5.17$$

En las expresiones 5.17 los estimadores están en términos del número total de unidades o de elementos que conforman al dominio de estudio d-ésimo; en ellas n_d, f_d y s_d^2 se definen de la misma manera que en el apartado 5.4. En el caso de que no se conozca el tamaño N_d del dominio, se pueden usar las expresiones siguientes:

$$\left.\begin{aligned} \hat{Y}_d &= \frac{N}{n} \sum_{k=1}^{n_d} y_{dk} \\ \hat{V}(\hat{Y}_d) &= \frac{N^2 \left(1 - \frac{n}{N}\right)}{n} (s_d')^2 \end{aligned}\right\} 5.18$$

en la cual:

$$(s_d')^2 = \frac{1}{n-1}\left(\sum_{k=1}^{k=n_d} y_{dk}^2 - \frac{\left(\sum_{k=1}^{k=n_d} y_{dk}\right)^2}{n}\right)$$

Ejemplo 5.3 En una zona turística mexicana localizada en las costas del Océano Pacífico se desea desarrollar un estudio; éste se refiere a los hoteles del lugar y se desea estimar su carga media turística, así como su capacidad máxima. El estudio forma parte de otro más amplio dentro del cual se piensa desarrollar una campaña de publicidad a gran escala y con ella estimar el número de cuartos faltantes bajo una demanda dada de ellos.

Para desarrollar la encuesta se cuenta con un listado que muestra los 900 hoteles de la zona en estudio y en él aparecen los nombres y direcciones de cada uno de ellos. En el estudio es de interés la estimación del número medio de turistas por hotel en una fecha dada, y esta estimación se requiere por separado para cual-

quier tipo de hotel y para aquellos que cuentan con teléfono y estacionamiento para vehículos; también interesa el total de cuartos en todos los hoteles, así como el total de cuartos restringido a aquellos hoteles que cuentan con teléfono y estacionamiento.

En base a la lista de 900 hoteles se elige aleatoriamente una muestra de 50 de ellos, se hacen las visitas correspondientes y se obtiene la información de la tabla 5.3.

Tabla 5.3

No. del hotel	*No. de turistas en la fecha dada*	*No. total de cuartos*	*¿El hotel cuenta con teléfono y estacionamiento?*
1	80	50	sí
2	50	30	sí
3	60	60	sí
4	45	20	sí
5	15	10	sí
6	20	15	no
7	25	15	sí
8	20	17	sí
9	15	13	no
10	10	20	sí
11	14	10	no
12	20	12	no
13	25	15	no
14	33	14	sí
15	80	38	sí
16	60	25	sí
17	35	20	sí
18	20	20	sí
19	23	15	sí
20	36	25	sí
21	18	10	sí
22	44	20	sí
23	40	22	sí
24	39	30	sí
25	90	50	sí
26	12	10	no
27	10	10	no
28	5	15	no
29	12	10	sí
30	13	13	no
31	10	17	sí

Tabla 5.3 (continuación)

No. del hotel	*No. de turistas en la fecha dada*	*No. total de cuartos*	*¿El hotel cuenta con teléfono y estacionamiento?*
32	20	25	sí
33	30	20	sí
34	32	20	sí
35	140	80	sí
36	85	40	sí
37	20	13	sí
38	7	9	sí
39	16	14	sí
40	19	17	sí
41	31	16	sí
42	41	23	sí
43	6	8	no
44	8	10	sí
45	10	15	sí
46	12	10	sí
47	14	10	sí
48	12	12	sí
49	18	14	sí
50	7	9	sí
Totales	1 507	1 016	

El número medio de turistas por hotel en la fecha dada y para cualquier tipo de hotel lo estimamos mediante la media muestral:

$$\hat{\bar{Y}} = \frac{1\,507}{50} = 30.14 \text{ turistas por hotel.}$$

Y la estimación de su varianza poblacional S^2 es como sigue:

$$\hat{S}^2 = \frac{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}{n-1} = \frac{80\,337 - \frac{(1\,507)^2}{50}}{50-1} = 712.57$$

Entonces la varianza de $\hat{\bar{Y}}$ es:

$$\left(1 - \frac{50}{900}\right)\frac{712.57}{50} = (1 - 0.055)14.24 = 13.46$$

La estimación del número medio de turistas por hotel en la fecha dada y para aquellos hoteles que cuentan con teléfono y estacionamiento debe tratarse como una estimación dentro de un dominio de estudio, ya que, de entre los hoteles sorteados, solamente algunos tienen la característica de interés: poseer teléfono y estacionamiento; entonces usando la expresión 5.13 tenemos:

$$\hat{\bar{Y}}_d = \frac{\sum_{k=1}^{n_d} y_{dk}}{n_d} = \frac{1\,367}{40}$$

= 34.175 turistas por hotel con teléfono y estacionamiento.

La variancia del estimador $\hat{\bar{Y}}_d$ debe ser calculada mediante la expresión 5.15. Este cálculo se solicita en el ejercicio 5.8.

El total de cuartos en todos los hoteles, lo estimamos expandiendo a la media muestral con N:

$$\hat{Y} = N\bar{y} = 900(30.14) = 27126 \text{ cuartos}$$

Para estimar el total de cuartos en aquellos hoteles que cuentan con teléfono y estacionamiento, debemos proceder como en el caso de un dominio de estudio. En este ejemplo no se dispone del número total de unidades u hoteles N_d en el dominio, por lo cual debemos usar como estimador de él la expresión 5.18, entonces:

$$\hat{Y}_d = \frac{N}{n} \sum_{k=1}^{n_d} y_{dk} = \frac{900}{50}(895) = 16\,110 \text{ cuartos}$$

Y su variancia la estimamos con:

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

pero (ecuación 5.18):

$$(s'_d)^2 = \frac{1}{n-1}\left(\sum_{k=1}^{n_d} y_{dk}^2 - \frac{\left(\sum_{k=1}^{n_d} Y_{dk}\right)^2}{n}\right)$$

$$\frac{1}{n-1}\left(\sum_{k=1}^{nd} y_{dk}^2 - \frac{\left(\sum_{k=1}^{nd} y_{dk}\right)^2}{n}\right) = \frac{1}{50-1}\left(28\,993 - \frac{(895)^2}{50}\right)$$

$$= \frac{1}{49}\left(28\,993 - 16\,020.5\right) = \frac{12\,972.5}{49} = 264.7$$

Entonces:

$$\hat{V}(\hat{Y}_d) = \frac{(900)^2\left(1 - \frac{50}{900}\right)(264.7)}{50}$$

$$= \frac{(900)^2\ (0.944)\ (264.7)}{50} = 4\,048\,004$$

Y su desviación o error estándar es de 2 012 cuartos en los hoteles con teléfono y estacionamiento. Con este error estándar podemos calcular intervalos de confianza del 95%:

$L_i = 16\,110 - 2(2\,012) = 12\,086$ cuartos

$L_s = 16\,110 + 2(2\,012) = 20\,134$ cuartos

5.6 OTROS USOS DE LOS DOMINIOS DE ESTUDIO

En los casos prácticos de encuestas es frecuente que el marco muestral no sea del todo correcto, o que, a pesar del cuidado con que se desarrolle el trabajo de campo, existan algunas unidades muestrales de las cuales no se tenga observación por causas diferentes: *i*) la persona a entrevistar se rehúsa a contestar; *ii*) el entrevistador no hizo la pregunta porque accidentalmente la brincó al desarrollar la entrevista; *iii*) no fue posible encontrar "en casa" a la persona a entrevistarse; *iv*) en el cuestionario, las respuestas aparecen borrosas; *v*) la unidad seleccionada como perteneciente a la muestra ya no pertenece a la población sujeta a estudio, porque, por ejemplo, aunque aparece en la última nómina a la fecha del trabajo de campo ya dejó su empleo, la empresa cambió de giro, etc.

En estos casos, se logra obtener satisfactoriamente sólo una parte del tamaño de muestra inicial *n*. Si el estimador propuesto inicialmente era: Al nacer la media muestral, esperando obtener una observación de todas las unidades y, sin embargo, sólo se tuvo

éxito en algunas de ellas formalmente el estimador anterior debe cambiar, una solución consiste en usar el concepto de dominios de estudio o subpoblaciones, más generalmente, las estimaciones solicitadas se refieren a subpoblaciones en cuyo caso habría que considerar, digamos, al resto de la muestra como empleado en elementos extraños. El estimador del parámetro (Media) vuelve a ser la media muestral, pero su varianza es ahora mayor (esta media muestral viene a ser un estimador de razón). Aquí existe una diferencia significativa en cuanto al uso de esquemas de muestreo autoponderados o de probabilidades iguales y los no autoponderados o de probabilidades variables. En los primeros, el estimador de la media se mantiene como la media muestral, aunque su estimador de varianza debe ser tratado como una razón de variables aleatorias, es decir, un estimador de razón. En los diseños de probabilidades variables es necesario construir el estimador respectivo; en cuyo caso, y en diseños a varias etapas, esta situación puede volverse relativamente complicada. Lo anterior hace que los diseños autoponderados sean más populares y de mayor éxito práctico, aunque posiblemente no sean los mejores desde el punto de vista teórico.* Claramente, en toda encuesta siempre existen errores pequeños y grandes; algunos de ellos son susceptibles de ser pasados por alto y otros no. En todo caso, el técnico debe usar su criterio para llegar a una decisión sobre la relevancia del error cometido. La consecuencia que se tiene al usar las expresiones correspondientes a subpoblaciones es que las precisiones de las estimaciones disminuyen, principalmente en el caso más general en el cual se desconocen los tamaños de los dominios.

5.7 RESTRICCIONES DE TIPO PRACTICO EN EL USO DEL MUESTREO ALEATORIO SIMPLE

Para llevar a cabo la selección de las unidades que conformarán a la muestra, este esquema requiere un listado de las diferentes unidades o elementos que integran a la población bajo estudio en el cual cada unidad quede identificada sin ambigüedad. Y para propósitos de identificación de las unidades, cada una de ellas debe tener asociado un número natural,** aunque el folio o numeración de las unidades no necesariamente debe ser consecutivo. En la práctica se presentan muchos casos en que la numeración de ellas está salteada (el caso de claves, por ejemplo) y no se procede a corregirla o a renumerarla al tiempo de la selección. La necesidad

* Ver Kish L. Survey Sampling. 1965. John Wiley and Sons. N. Y. y el capítulo sobre muestreo estratificado en este libro (6.3).

** O cualquier otro procedimiento que permita su identificación sin error.

de contar con este listado es una primera restricción, importante, de este esquema. Por otro lado, como la selección es aleatoria, todas las unidades tienen la misma probabilidad de ser elegidas, y la muestra tiende a dispersarse en toda la población, lo que generalmente no es conveniente para efectos del trabajo de campo ya que obliga a viajar "sin control" por toda la población.* En adición, aunque las unidades no sean muchas —digamos 100—, pero sin numeración, la selección es muy tediosa y fuertemente sujeta a errores. Por lo anterior, generalmente el uso del muestreo aleatorio simple, como esquema fundamental, queda recomendado para poblaciones relativamente pequeñas.

En la actualidad casi no existe restricción en cuanto a la disponibilidad de tablas de números aleatorios, en el sentido de que cuando una tabla preconstruida no alcanza, con ayuda de las funciones de biblioteca de las computadoras actuales es fácil generar un número muy grande de números aleatorios en el rango que se desee. Si r es un número aleatorio entre cero y uno,

$$R = a + (b - a)r$$

es un número aleatorio entre a y b, donde éstos son números naturales, incluido el cero. "a" es el extremo inferior del intervalo y "b" es el extremo superior (Naylor). Entonces, si se desean, por ejemplo 20 000 números aleatorios que estén comprendidos entre 1 y 10 000 000, con relativa facilidad se puede elaborar un pequeño programa de computadora para obtener esos números aleatorios.

5.8 FALSEAMIENTO DEL ESQUEMA DE SELECCION

Como se indicó en el capítulo 4, el diseño muestral por sí solo no constituye la encuesta. La elaboración de cuestionarios y de manuales, el entrenamiento al personal para el trabajo de campo y la coordinación de éste, así como el procesamiento y la correcta interpretación de los resultados de la encuesta son etapas que deben ser desarrolladas con el debido cuidado para evitar fracasos y malas interpretaciones.

* Cuando es necesario muestrear sobre archivos magnéticos en los cuales los registros aparecen seriados, generalmente es factible el uso del muestreo aleatorio simple, ya que sólo es necesario generar los números aleatorios con los cuales se conformará la muestra, ordenarlos y posteriormente avanzar sobre el archivo contando y detectando a los registros en la muestra.

Desde el punto de vista técnico, fallas de gabinete provocadas por una selección aleatoria deficiente u originadas durante el trabajo de campo, respecto a la localización y entrevista a las unidades seleccionadas, se reflejan en un falseamiento al esquema de selección. A algunas unidades se les asigna probabilidad de cero de aparecer en la muestra y a otras se les asigna probabilidades positivas, varias veces mayores que la necesaria. Y, desafortunadamente, en la mayoría de los casos se desconocen esas probabilidades variables introducidas a última hora. Por ejemplo, la persona *i*-ésima está en la muestra, pero no se le pudo localizar y el entrevistador optó por sustituirla por la unidad vecina. En este caso, se le asignó una probabilidad de cero a la unidad buscada inicialmente y doble o posiblemente varias veces la probabilidad de selección que debería regir a la vecina. *En los esquemas de muestreo probabilísticos no debe existir la sustitución de unidades muestrales* porque el azar materializó una muestra; si algunas unidades son sustituidas por otras, aparentemente con los mismos valores del atributo en estudio de la previamente elegida, se está falseando el esquema de selección porque a esas últimas unidades ya no las eligió el azar, y por otro lado el tamaño de muestra se está aumentando. Por ello ya no se pueden usar las bondades del estimador y del diseño en general.

Ejemplo 5.4 En un estudio sobre la disponibilidad de un artículo en las ferreterías de una población grande, se elige aleatoriamente a 30 de ellas, se hacen las visitas correspondientes y después de colectados los cuestionarios y revisadas las notas y observaciones que los enumeradores anotaron en ellos, se encontró lo siguiente: (ver la tabla 5.4)

En este ejemplo, el establecimiento número 4 en la muestra, no pertenece a la población en estudio; sin embargo, se pueden hacer las estimaciones tratando a la población como un dominio de estudio en el cual esta unidad no pertenece a la clase de interés. El establecimiento número 11 no se sabe si pertenece o no ya que pudo estar cerrado de manera ocasional y, como el enumerador no escribió ninguna otra información que aclare sobre su existencia, no se puede decidir sobre ella; una serie de revisitas puede resolver este problema.

En el caso del establecimiento número 18, si la dirección es la correcta, la unidad no pertenece a la población bajo estudio, pero,

Tabla 5.4

Ferretería	*Tiene el artículo*	*Observaciones*
1	no	
2	no	
3	no	
4		Está clausurada
5	sí	
6	no	
7	sí	
8	no	
9	no	
10	sí	
11		Está cerrada
12	sí	
13	sí	
14	no	
15	sí	
16	sí	
17	sí	
18		Hay un edificio residencial en la dirección dada.
19	si	
20	sí	
21		Como no estaba el encargado, el enumerador visitó a la siguiente ferretería en la misma calle.
22	sí	
23	no	
24		Parece que existe el artículo, pero no estaba el encargado.
25	sí	
26	sí	
27	sí	
28	sí	
29		Ya no existe
30	sí	

como en el caso de la unidad número 4 en la muestra, se puede proceder como si se tratara de un dominio de estudio.

La observación que se haya colectado al sustituir al establecimiento número 21, se debe eliminar, ya que no se permite la sustitución de unidades muestrales cualquiera que sea el caso o situación que se presente.

La situación ambigua de la unidad número 24 puede eliminarse al efectuar alguna(s) revisita(s).

Y, por último, la situación de la unidad muestral número 29, es similar a los ya presentados y discutidos en las unidades 4 y 18. Esta unidad no pertenece a la población bajo estudio.

El caso de unidades que no pertenecen a la población de interés, pero que están ocupando un lugar en el marco muestral y por lo tanto entran al sorteo, puede resolverse a través del concepto de dominios de estudio. Pero problemas como aquel que surgió en el establecimiento número 21, en el que el entrevistador hizo una sustitución, o problemas de no respuesta porque la persona se rehúsa a contestar, aunque se siguen trabajando como dominios de estudio se reflejan en un desconocimiento pleno de una parte dé la población original.

5.9 MUESTREO ALEATORIO SIMPLE CON REMPLAZO

En el apartado 3.1 se presentó la selección aleatoria sin reemplazo. Ahora mostramos algunos resultados que son válidos cuando la selección es *con remplazo.* Este tipo de selección es usada sobre todo en muestreo por conglomerados y en submuestreo (capítulos 7 y 8). Su uso hace que los desarrollos matemáticos sean más simples, y las diferencias en precisión entre una selección con remplazo y una sin él, usualmente son de escasa importancia.

Cuando se selecciona aleatoriamente y con reemplazo a n unidades, las selecciones son independientes y en cada extracción la probabilidad de que la unidad i-ésima sea elegida es $\frac{1}{N}$. El número de veces que cada unidad puede aparecer en la muestra es 0, 1, . . . n. Sea t_i el número de veces que la unidad i-ésima aparece en la muestra, entonces;

$$\hat{\bar{Y}} = \bar{y} = \frac{\sum_{i=1}^{N} t_i \quad y_i}{n}$$

t_i es una variable aleatoria, y se distribuye como una binomial (ver el ejercicio 3.11), entonces en la consideración de este nuevo esquema de selección la esperanza matemática de la media muestral es:

$$E(\bar{y}) = \frac{1}{n} \sum_{i=1}^{N} y_i \cdot E(t_i) = \frac{1}{n} \sum_{i=1}^{N} y_i \frac{n}{N}$$

$$E(\bar{y}) = \bar{Y}$$

luego, en el muestreo aleatorio simple con remplazo, la media muestral considerada como estimador de la media poblacional resulta ser insesgada de ella.

Ahora encontremos su variancia:

$$V(\bar{y}) = \frac{1}{n^2} V\left(\sum_{i=1}^{N} t_i y_i\right) =$$

$$= \frac{1}{n^2} \sum_{i=1}^{N} y_i^2 \cdot V(t_i) + 2 \sum_{i<j}^{N} y_i y_j \cdot \mathrm{Cov}(y_i, y_j)$$

$$= \frac{1}{n^2} \sum_{i=1}^{N} y_i^2 \frac{n}{N} \frac{N-1}{n} - 2 \sum_{i<j}^{N} y_i y_j \frac{n}{N^2}$$

$$= \frac{N-1}{N} \frac{S^2}{n}$$

Y se puede demostrar que un estimador insesgado de ella es:

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

5.10 EJERCICIOS

5.1 En el apartado 5.2, la población de niños a que se hace referencia es la formada por aquellos nacidos en un sanatorio determinado y en el cual se llevan estadísticas de los recién nacidos. En una semana nacieron 2 000 niños y su talla o longitud media calculada para todos ellos fue de 46 centímetros. A los siete meses de edad se elige aleatoriamente a 30 de ellos, cada niño en la muestra es medido (y_i) y posteriormente se colecta su talla o longitud inicial (x_i) a partir de sus fichas de nacimiento. Los datos son los siguientes, (ver tabla 5.5).

Estime: *a*) la talla o longitud media de los niños a los 7 meses de edad, *b*) el error estándar de su estimador y *c*) calcule intervalos de confianza del 95% para la talla de los niños mediante *i*) la media muestral, y *ii*) el estimador de razón. ¿Qué método es más preciso?

5.2 En el apartado 5.2, sobre el ejemplo de las gavetas en los 6 estantes con 240 gavetas en total, se pide estimar el número total de tarjetas pertenecientes a hijos, así como el error estándar y dar intervalos de confianza del 95%. El tamaño de la muestra fue de 20 gavetas y el peso total de las tarjetas en ellas fue de 30 kilos*; la muestra arrojó los resultados de la tabla 5.6.

* 30 kilogramos son iguales a 30 000 gramos, que son las unidades de medida para x_i en la tabla 5.6.

Tabla 5.5

y_i	x_i	y_i	x_i	y_i	x_i
52	38	70	53	52	39
62	43	71	50	56	42
73	50	55	40	57	41
57	45	59	47	60	46
68	45	71	47	58	44
54	42	58	44	74	50
53	40	72	48	48	37
51	38	74	49	52	39
63	46	63	46	57	44
70	48	53	40	70	48

Tabla 5.6

y_i:	160	180	190	240	200	150	190	190	240	220
x_i:	150	120	130	170	160	140	140	120	180	160
y_i:	200	200	220	240	180	160	190	200	130	170
x_i:	150	140	170	180	140	110	130	140	120	130

Calcule las estimaciones anteriores usando: *i*) $N\bar{y}$; *ii*) el estimador de razón.

5.3. Si $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$, y, $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$, demuestre que bajo muestreo aleatorio simple la covariancia entre ellas está dada por:

$$\mathrm{COV}(\bar{x}, \bar{y}) = \frac{1-f}{n} \frac{\sum_{i=1}^{N} (y_i - \bar{Y})(x_i - \bar{X})}{N-1}$$

5.4 Usando las definiciones y el resultado del ejercicio 5.3 y considerando la expansión por la serie de Taylor de la función $g(\bar{x}, \bar{y})$ alrededor del punto $(\bar{X}, \bar{Y})$:

$$g(\bar{x}, \bar{y}) = g(\bar{X}, \bar{Y}) + (x - \bar{X}) \left.\frac{\partial g\ (\bar{x}, \bar{y})}{\partial \bar{x}}\right|_{\substack{\bar{x}=\bar{X} \\ \bar{y}=\bar{Y}}} + (\bar{y} - \bar{Y}) \left.\frac{\partial g\ (\bar{x}, \bar{y})}{\partial \bar{y}}\right|_{\substack{\bar{x}=\bar{X} \\ \bar{y}=\bar{Y}}}$$

+ (despreciable)

Entonces:

$$\mathrm{Var}\left[g(\bar{x}, \bar{y}) \right] = \mathrm{E}\left[g(\bar{x}, \bar{y}) - g(\bar{X}, \bar{Y}) \right]^2$$

$$= \left[\frac{\partial g(\bar{x}, \bar{y})}{\partial \bar{x}}\Big/_{\substack{\bar{x}=\bar{X} \\ \bar{y}=\bar{Y}}}\right]^2 \cdot \text{Var}\,\bar{x} + \left[\frac{\partial g(\bar{x}, \bar{y})}{\partial \bar{y}}\Big/_{\substack{\bar{x}=\bar{X} \\ \bar{y}=\bar{Y}}}\right]^2 \cdot \text{Var}\,\bar{y} +$$

$$+ 2\left(\frac{\partial g(\bar{x}, \bar{y})}{\partial \bar{x}}\Big/_{\substack{\bar{x}=\bar{X} \\ \bar{y}=\bar{Y}}}\right)\left(\frac{\partial g(\bar{x}, \bar{y})}{\partial \bar{y}}\Big/_{\substack{\bar{x}=\bar{X} \\ \bar{y}=\bar{Y}}}\right) \cdot \text{Cov}\,(\bar{x}, \bar{y})$$

Si ahora, hacemos $g(\bar{x}, \bar{y}) = \frac{\bar{y}}{\bar{x}} = \hat{R}$, demuestre que

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

5.5 En el caso de dominio de estudio demuestre que con muestreo aleatorio simple $E\left(\frac{n_d}{N_d}\right) = \frac{n}{N}$

5.6 En referencia al ejemplo 5.2, calcule intervalos de confianza del 95% para el peso medio de piedras por saco, así como para el total en los 1 000 sacos.

5.7 En referencia al ejemplo 5.2, estime el error estándar del estimador del cociente entre el peso total de la piedra y el peso total de la semilla usando la aproximación normal y considerando como válida la expresión 5.3.

5.8 En referencia al ejemplo 5.3, estime la variancia del número medio de turistas por hotel para aquellos hoteles que cuentan con teléfono y estacionamiento.

5.9 En una encuesta sobre el personal de una empresa, se quiere estimar el porcentaje de empleados que se enteran regularmente de los cambios introducidos al reglamento de seguridad interno, así como el porcentaje de ellos que regularmente fuman cigarrillos. Para la encuesta se usan los listados de pago correspondientes a la última quincena. De ella se elige aleatoriamente a 40 empleados de entre un total de 5 000, obteniéndose los resultados de la tabla 5.7.

Calcule las estimaciones solicitadas, así como intervalos de confianza del 95% para el porcentaje de empleados que se enteran regularmente de los cambios introducidos al reglamento interno de seguridad.

5.10 Considere las expresiones obtenidas para las variancias de la media muestral en los esquemas de selección con y sin remplazo. ¿Cuál es la magnitud de su diferencia?

5.11 Con referencia al ejemplo de los gaveteros del apartado 5.2, ¿por qué es necesario que las familias generalmente tengan más de dos hijos?

Tabla 5.7

No. del empleado	*Conocen los cambios al reglamento*	*Fuman cigarrillos*
1	no	sí
2	no	no
3	no	sí
4	no	sí
5	no	no
6	sí	no
7	no	no
8	sí	no
9	sí	no
10	no	no
11	sí	no
12	sí	no
13	sí	no
14	sí	no
15	renunció	—
16	sí	no
17	sí	sí
18	sí	no
19	no	sí
20	sí	sí
21	sí	sí
22	sí	no
23	tiene permiso por 6 meses	
24	renunció	—
25	no	sí
26	no	no
27	sí	no
28	no	no
29	sí	no
30	sí	no
31	sí	no
32	no	no
33	no	no
34	no	sí
35	sí	sí
36	no	no
37	no	no
38	no	no
39	no	no
40	no	no

CAPITULO 6

MUESTREO ESTRATIFICADO

6.1 MUESTREO ESTRATIFICADO

En múltiples ocasiones resulta posible y conveniente partir o fraccionar a la población original en subdivisiones de tal naturaleza que ellas formen una partición. En estas condiciones cada unidad pertenece a una y sólo a una subdivisión y la unión de todas ellas conforma a la población original. En cuanto al método de selección y en parte al de estimación, a cada una de las subdivisiones se les trata de manera independiente, aunque el método de estimación las unirá en forma global. A un esquema de este tipo se le conoce como muestreo estratificado y a cada subdivisión trabajada de manera independiente se le denomina *estrato*. Todos los estratos son disjuntos y su unión es igual a la población original.

El muestreo estratificado es ampliamente usado por varios motivos:

i) desde el punto de vista del método de selección permite trabajar o estudiar a cada estrato por separado;
ii) permite derivar estimaciones por estrato o a nivel de estrato y cada una de ellas ser estudiada con la precisión solicitada;
iii) las estimaciones así derivadas resultan ser usualmente más precisas que aquellas derivadas mediante una selección aleatoria;
iv) ayuda a resolver muchos problemas de coordinación del trabajo de campo.

La población original de tamaño N es fragmentada en L estratos de los cuales el h-ésimo es de tamaño N_h, $h = 1, 2, \ldots L$. Cada unidad de la población aparece en uno y sólo en un estrato; y la selección de n_h, $h = 1, 2, \ldots L$, unidades en el estrato h-ésimo se efectúa de manera independiente de las selecciones en el resto de los estratos. La variancia de la característica en estudio en la población total está definida mediante la expresión siguiente:

$$S^2 = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y})^2}{N - 1}$$

y su variancia en el estrato h-ésimo está definida por:

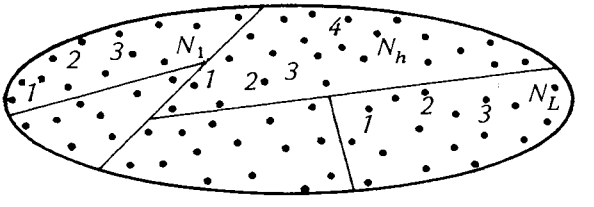
$$S_h^2 = \frac{\sum_{i}^{N_h} (y_{hi} - \bar{Y}_h)^2}{N_h - 1}$$

donde $\bar{Y}$ y $\bar{Y}_h$ simbolizan respectivamente a la media de toda la población o media general y la media del estrato h-ésimo:

$$\bar{Y}_h = \frac{\sum_{i=1}^{i=N_h} y_{hi}}{N_h}$$

$$\bar{Y} = \frac{\sum_{h=1}^{h=L} \sum_{i=1}^{i=N_h} y_{hi}}{N} = \frac{\sum_{h=1}^{h=L} \sum_{i=1}^{i=N_h} y_{hi}}{N_1 + N_2 + \ldots + N_h + \ldots + N_L}$$

La población de tamaño N es fragmentada en L estratos



De cada estrato se elige una muestra de manera independiente del resto de las selecciones.
$y_{h_1}, y_{h_2}, \ldots \ldots y_{hn_h}$.

Figura 6.1 En el estrato h*-ésimo las unidades en la muestra son las numeradas: 1, 2, . . ., n_h y el valor de su característica es $y_{h_1}, y_{h_2}, \ldots y_{hn_h}$.*

Ejemplo 6.1: Para desarrollar una encuesta sobre granjas que se dedican a la crianza de ganado bovino, se tienen 15 listados pertenecientes a 12 regiones diferentes que cuentan con este tipo de granjas. En una de las regiones su listado correspondiente aparece fraccionado en 4 partes por lo que el total de ellos es de 15.

De entre las zonas geográficas diferentes en estudio, 5 de ellas están separadas entre sí por más de 100 kilómetros, mientras que el resto de la regiones se localizan alrededor de cuatro de las cinco antes referidas y la quinta está particularmente lejana (ver la figura 6.2). Se desea estimar el total de cabezas existentes en las 12 regiones, y debido a que es de interés especial, en el estudio, el tipo de ganado con que cuenta la región lejana, resulta conveniente contar además con estimaciones de ella por separado

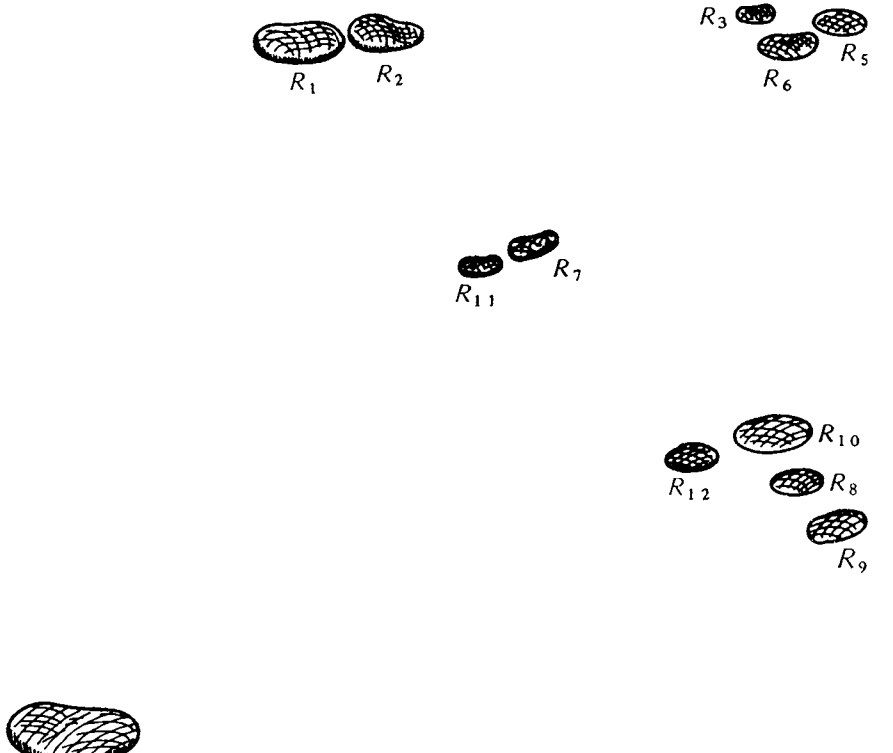


Figura 6.2.

Según los listados existentes los números totales de granjas por región son como sigue:

Región	1	2	3	4	5	6	7	8	9	10	11	12	Total
No. Total de granjas	221	75	30	400	50	90	60	50	63	77	42	25	1 183

Agrupando a las regiones que se encuentran relativamente cercanas se obtiene la tabla 6.1:

Tabla 6.1

	Regiones
I	1 y 2
II	3, 5 y 6
III	4 muy lejana
IV	7 y 11
V	8, 9, 10 y 12

De manera que desde el punto de vista geográfico una estratificación posible estaría definida por los grupos anteriores, los diferentes estratos y sus tamaños en términos del número de granjas son los de la tabla 6.2.

Tabla 6.2

Estrato h:	1	2	3	4	5	Total N
Regiones:	1 y 2	3, 5 y 6	4	7 y 11	8, 9, 10 y 12	
Tamaño N_h:	296	170	400	102	215	1183

El número total de granjas en el estudio es:

$$N = N_1 + N_2 + \ldots + N_L$$
$$N = 296 + 170 + 400 + 102 + 215$$
$$= 1\,183$$

La estratificación anterior se ha hecho según un criterio geográfico, el cual no necesariamente toma en cuenta los diferentes

tamaños de las granjas, es decir, el número de cabezas por granja; sin embargo, si no se tiene más información y en el supuesto de que no existen problemas de vías de comunicación entre ellas, ésta puede ser adecuada. Con respecto a la regionalización anterior, ¿qué crítica puede usted hacer?

6.2 ESTIMACION DE MEDIAS

Para una característica particular en estudio, su media poblacional se define de la manera siguiente:

$$\bar{Y} = \frac{\sum_h \sum_i y_{hi}}{N} = \frac{\sum_h Y_h}{N} = \frac{\sum_h N_h \bar{Y}_h}{N}$$

en la cual los subíndices en y_{hi} hacen referencia al estrato h-ésimo y a la unidad i-ésima en el estrato anterior.

Consideremos que $\bar{y}_{est}$ con la definición según la expresión 6.1, es el estimador de la media poblacional en muestreo estratificado.

$$\bar{y}_{est} = \frac{\sum_h N_h \bar{y}_h}{N} \qquad 6.1.$$

A y_{est} se le denomina *media estratificada;* para obtenerla es necesario multiplicar a cada estimador de la media de los estratos $\bar{y}_h$ por el número de unidades en el N_h, sumar los productos para todos los estratos y dividir por el número total N de unidades en la población. Si la selección dentro de cada estrato es con muestreo aleatorio simple, entonces el estimador de la media en el estrato h-ésimo $\bar{y}_h$ coincide con la media muestral (ecuación 3.1) y sólo conservamos el subíndice h en ella para evitar confusiones con las medias muestrales de los diferentes estratos. Calculemos la esperanza matemática de la media estratificada con la selección aleatoria antes referida:

$$\mathrm{E}(\bar{y}_{est}) = \mathrm{E}\left(\frac{\sum_h N_h \bar{y}_h}{N}\right)$$

$$\mathrm{E}(\bar{y}_{est}) = \frac{\sum_h N_h}{N} \mathrm{E}(\bar{y}_h) = \frac{\sum_h N_h \bar{Y}_h}{N} = \bar{Y},$$

Por lo cual decimos que cuando la selección de las unidades de la muestra en cada estrato se hace con muestreo aleatorio simple, el estimador $\bar{y}_{est}$ es insesgado de la media poblacional $\bar{Y}$.*

Suponiendo el mismo tipo de selección aleatoria en cada estrato, la variancia de la media estratificada está dada por:

$$V(\bar{y}_{est}) = V\left(\frac{\Sigma\, N_h \bar{y}_h}{N}\right)$$

$$V(\bar{y}_{est}) = \sum_h \frac{N_h^2}{N^2} V(\bar{y}_h) = \frac{\sum_h N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} \qquad 6.2$$

también la obtención de un estimador insesgado de la variancia de $\bar{y}_{est}$ es inmediato:

$$\hat{V}(\bar{y}_{est}) = \sum_h \frac{N_h^2}{N^2}(1 - f_h)\frac{s_h^2}{n_h} \qquad 6.3$$

en donde:

$$s_h^2 = \frac{\sum^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1} \quad \text{y,}$$

$$\bar{y}_h = \frac{\sum^{n_h} y_{hi}}{n_h}$$

En la expresión 6.1, $\frac{N_h}{N}$ es la ponderación que recibe $\bar{y}_h$, por lo que se denomina ponderación del estrato h-ésimo, y en 6.2 y 6.3 el cociente $f_h = \frac{n_h}{N_h}$ se le denomina fracción de muestreo del estrato h-ésimo. En este capítulo se considera que la selección dentro de cada estrato se efectúa con muestreo aleatorio simple. En la ecuación 6.3 y en la definición de s_h^2 éste aparece dividido por $n_h - 1$ por lo cual, para que exista el estimador de la variancia en el estrato h-ésimo, es necesario que el tamaño de muestra en él, n_h sea al menos de 2.

* En realidad la media estratificada será insesgada de la media poblacional siempre que el estimador de la media en cada estrato sea insesgado de su media general.

La ecuación 6.2, muestra la variancia de la media estratificada, y ésta viene siendo una suma ponderada según la fracción $(\frac{N_h}{N})^2$de las variancias del estimador de la media en cada estrato, no existiendo covariancias por la independencia de la selección de estrato a estrato. Entonces, para que una estratificación arroje variancias pequeñas, en otras palabras para que sea muy precisa, se requiere que la variabilidad dentro de cada estrato S_h^2, sea pequeña, lo cual se logra formando los estratos de manera que internamente resulten homogéneos respecto a la característica en estudio, es decir, que las unidades tiendan a parecerse respecto a la magnitud de la característica en estudio. Externamente, puede y es deseable que exista gran heterogeneidad entre estratos, ya que esta variabilidad no coopera para la variancia del estimador en la expresión 6.2. Por ejemplo, si en una ciudad se hace una encuesta para estimar el ingreso medio familiar; se pueden colocar manzanas de familias con ingreso muy bajo en un estrato, manzanas con familias de ingresos bajos en otro estrato, las de ingresos medios en otro y por último aquellas con ingresos altos y muy altos en otro estrato. Al introducir muestras independientes en cada estrato, la media estimada del estrato 1 es $\bar{y}_1 = \frac{\sum_{i=1} y_{1i}}{n_1}$, y su variancia está dada por $(1 - f_1)\ \frac{S_1^2}{n_1}$; la media estimada del estrato 2 es $\bar{y}_2 = \frac{\sum_{i=1}^{n_2} y_{2i}}{n_2}$ y con una variancia de $(1 - f_2)\ \frac{S_2^2}{n_2}$ y de la misma manera para el resto de los estratos.

Ejemplo 6.2: En una industria que elabora tapas de plástico existen 400 máquinas que fabrican ese producto. Las máquinas han sido adquiridas por la empresa según sus condiciones económicas a través de varios años y así, existen en ella 240 que son operadas manualmente y por lo tanto de bajo rendimiento; 100 semiautomáticas y 60 completamente automáticas, de alto rendimiento. Se desea estimar el número medio de tapas producidas por máquina en la primera semana de junio.

Aunque es factible practicar una muestra aleatoria simple sobre las máquinas, ya que es relativamente fácil numerarlas y hacer la selección, sabemos que este tipo de selección revolvería máquinas de pequeño y alto rendimiento, por lo que es más aconsejable

practicar una estratificación según el criterio: modo de operación de la máquina, *i. e.* manual, semiautomática y automática. Entonces formamos tres estratos de tamaños $N_1 = 240$, $N_2 = 100$ y $N_3 = 60$, y las ponderaciones de cada uno de ellos son $\frac{240}{400}$, $\frac{100}{400}$ y $\frac{60}{400}$ respectivamente. Si decidimos emplear en el primer estrato un tamaño de muestra de 12, en el segundo de 5 y en el tercero de 3, es decir, $n_1 = 12$, $n_2 = 5$ y $n_3 = 3$, el tamaño total de la muestra será de $n = n_1 + n_2 + n_3 = 20$. Entonces las fracciones de muestreo en cada estrato serán $f_1 = \frac{12}{240}$, $f_2 = \frac{5}{100}$ y $f_3 = \frac{3}{60} = \frac{1}{20}$. Supongamos que al tomar la muestra los datos son como sigue:

ESTRATO 1

Número de tapas producidas

2 600, 2 000, 1 800, 1 700, 2 400, 1 600,
1 700, 2 400, 1 100, 2 100, 2 300, 1 800.

Entonces;

$$\sum_{i=1}^{n_1} y_{1i} = 23\,500, \quad \sum_{i=1}^{n_1} y_{1i}^2 = 48\,010\,000, \quad \left(\sum_{i=1}^{n_1} y_{1i}\right)^2 = 552\,250\,000$$

ESTRATO 2

Número de tapas producidas

4 000, 5 200, 6 000, 8 300, 6 600

Entonces:

$$\Sigma y_{2i} = 30\,100, \quad \Sigma y_{2i}^2 = 191\,490\,000, \quad (\Sigma y_{2i})^2 = 906\,010\,000$$

ESTRATO 3

Número de tapas producidas

17 900, 24 000, 19 000.

Entonces:

$$\Sigma\, y_{3i} = 60\,900,\ \Sigma\, y_{3i}^2 = 1\,257\,410\,000,\ (\Sigma\, y_{3i})^2 = 3\,708\,810\,000$$

Las medias muestrales por estrato son:

$$\bar{y}_1 = \frac{23\,500}{12} = 1\,958.33,$$

$$\bar{y}_2 = \frac{30\,100}{5} = 6\,020$$

$$\bar{y}_3 = \frac{60\,900}{3} = 20\,300,$$

y son a la vez los estimadores de los rendimientos medios por máquina en cada estrato; sus errores estándar se calculan como sigue:

$$\hat{V}(\bar{y}_1) = (1 - f_1)\frac{s_1^2}{n_1}$$

$$= (1 - \frac{12}{240})\frac{1}{12}\frac{1}{12\text{-}1}\left(\Sigma y_{1i}^2 - \frac{(\Sigma y_{1i})^2}{12}\right)$$

$$= (1 - 0.05)\frac{1}{12(11)}\left(48\,010\,000 - \frac{552\,250\,000}{12}\right)$$

$$= \frac{0.95}{132}\,1\,990\,000 = 14\,322$$

luego el error estándar de $\bar{y}_1$ vale 120 aproximadamente.

$$\hat{V}(\bar{y}_2) = (1 - f_2)\frac{s_2^2}{n_2}$$

$$= (1 - \frac{5}{100})\frac{1}{5}\frac{1}{5-1}\left(\Sigma y_{2i}^2 - \frac{(\Sigma y_{2i})^2}{5}\right)$$

$$= (1 - 0.05)\frac{1}{5}\frac{1}{4}\left(191\,490\,000 - \frac{906\,010\,000}{5}\right.$$

$$= \frac{0.95}{20}\ 10\ 288\ 000 = 488\ 680$$

y la raíz cuadrada de 488 680 es el error estándar de $\bar{y}_2$, o sea 699.

$$\hat{V}(\bar{y}_3) = (1 - f_3)\frac{s_3^2}{n_3}$$

$$\hat{V}(\bar{y}_3) = \frac{0.95}{6}\left(1\ 257\ 410\ 000 - \frac{3\ 708\ 810\ 000}{3}\right)$$

$$= 3\ 347\ 167$$

y el error estándar de $\bar{y}_3$ es 1 830.

Con los resultados anteriores podemos calcular la media estratificada y ésta vale:

$$\bar{y}_{est} = \frac{N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3}{N} =$$

$$= \frac{240(1\ 958.33) + 100(6\ 020) + 60(20\ 300)}{400}$$

$$= 5\ 725 \text{ tapas/máquina.}$$

5.725 tapas por máquina es la estimación solicitada del número medio de tapas producidas por máquina, su error estándar se calcula de la manera siguiente:

$$\hat{V}(\bar{y}_{est}) = \Sigma\ \left(\frac{N_h}{N}\right)^2\ \hat{V}(\bar{y}_h)$$

$$= \left(\frac{N_1}{N}\right)^2\ \hat{V}(\bar{y}_1) + \left(\frac{N_2}{N}\right)^2\ \hat{V}(\bar{y}_2) + \left(\frac{N_3}{N}\right)^2\ \hat{V}(\bar{y}_3)$$

$$= \left(\frac{240}{400}\right)^2\ 14\ 322 + \left(\frac{100}{400}\right)^2\ 488\ 680 + \left(\frac{60}{400}\right)^2\ 3\ 347\ 167$$

$$= 0.36\ (14\ 322) + 0.0625\ (488\ 680) + 0.0225(3\ 347\ 167)$$

$$= 111\ 010$$

entonces su error estándar es de 333 tapas.

6.3 AFIJACION PROPORCIONAL

En el muestreo estratificado el tamaño total n de la muestra introducida en los L estratos, es la suma de los tamaños de las muestras en cada uno de ellos. Cuando se tiene como dato n, es necesario emplear algún criterio para *afijarla* entre los diferentes estratos, es decir, ¿qué tamaño de muestra se le debe asignar a cada estrato?

Una manera de hacerlo, es afijándolo o asignándolo según la ponderación $\frac{N_h}{N}$ de cada estrato. A los estratos más grandes se les asigna mayor tamaño de muestra, y a los más chicos menor muestra. A este criterio se le denomina *afijación proporcional* y es ampliamente usado en las encuestas de tipo práctico.

Matemáticamente este criterio queda representado por:

$$n_h = \left(\frac{N_h}{N}\right)n, \qquad h = 1, 2, \ldots, L \tag{6.4.}$$

Es decir, $n_1 = \frac{N_1}{N}n$; $n_2 = \frac{N_2}{N}n$; $n_3 = \frac{N_3}{N}n$; $\ldots$; $n_L = \frac{N_L}{N}n$

Así, en el ejemplo 6.2, el tamaño total n de la muestra vale 20 y usando afijación proporcional obtenemos como tamaño de muestra para cada uno de los estratos:

$$n_1 = (20)\left(\frac{240}{400}\right) = 12, \quad n_2 = (20)\left(\frac{100}{400}\right) = 5, \quad n_3 = (20)\left(\frac{60}{400}\right) = 3.$$

El uso de este criterio de afijación simplifica considerablemente a los estimadores. Con él, la estructura de la media estratificada se reduce a:

$$\bar{y}_{est} = \frac{\sum_h N_h \frac{\sum y_{hi}}{n_h}}{N} = \frac{\sum_h N_h \frac{N}{N_h n} \sum y_{hi}}{N}$$

$$\bar{y}_{est} = \left(\frac{1}{n}\right) \sum_h \sum_i y_{hi}, \tag{6.5.}$$

y se dice que la expresión 6.5 es un *estimador autoponderado.* Como se observa en ella, para encontrar el valor estimado sólo hay que sumar los valores obtenidos en la muestra y dividir por el tamaño de ésta.*

Con el mismo criterio de afijación proporcional, la estructura de la variancia de la media estratificada se reduce a:

$$V(\bar{y}_{est}) = (1 - \frac{n}{N})\frac{1}{n}\sum_h \frac{N_h}{N} S_h^2 = \frac{1-f}{nN}\sum_h N_h S_h^2$$

$$V(\bar{y}_{est}) = \frac{1-f}{nN} \sum_h N_h S_h^2 \qquad 6.6$$

donde $f = \frac{n}{N}$ es la fracción de muestreo general o global y el estimador de la variancia se reduce a:

$$\hat{V}(\bar{y}_{est}) = \frac{1-f}{nN} \sum_h N_h s_h^2 \qquad 6.7$$

Sin embargo, 6.1, 6.2 y 6.3 son expresiones generales válidas para cualquier afijación que se use.

6.4 ESTIMACION DE TOTALES

Si se desea estimar el valor total de una característica habiendo usado muestreo estratificado, usamos la expresión siguiente:

$$\hat{Y}_{est} = N\bar{y}_{est} \qquad 6.8$$

es decir, para obtener al estimador estratificado del total poblacional, multiplicamos a la media estratificada por el total N de unidades en la población.

En el ejercicio 6.9 se pide que se demuestre que $\hat{Y}_{est}$ es *consistente e insesgado* del total poblacional. Su variancia es inmediata a partir de 6.2 y vale:

$$V(N\bar{y}_{est}) = N^2 \hat{V}(\bar{y}_{est}), \qquad 6.9$$

y un estimador insesgado de ella es:

* Notar que este estimador sólo es válido si la muestra fue afijada proporcionalmente al tamaño relativo de los estratos.

$$\hat{V}(\hat{N}\bar{y}_{est}) = \sum_h N_h^2 \, (1 - f_h) \frac{s_h^2}{n_h} \qquad 6.10$$

Con afijación proporcional las expresiones 6.8 y 6.10 se convierten en:

$$\hat{Y}_{est} = \left(\frac{N}{n}\right) \sum_h \sum_i y_{hi} \qquad 6.11$$

$$\hat{V}(N\bar{y}_{est}) = \frac{N(1 - f)}{n} \sum_h N_h \, s_h^2 \qquad 6.12$$

Consideremos que en el ejemplo 6.2 deseamos estimar el número total de tapas fabricadas, entonces según la expresión 6.8:

$$\hat{Y}_{est} = 400(5725) = 2\,290\,000 \text{ tapas},$$

y su error estándar estimado es de:

$$400(333) = 133\,200 \text{ tapas}.$$

6.5 ESTIMACION DE PORCENTAJES

En muestreo estratificado el estimador de un porcentaje poblacional, es análogo a la media estratificada:

$$\left.\begin{aligned} p_{est} &= \left(\frac{1}{N}\right) \sum_{h=1}^{h=L} N_h p_h \\ p_h &= \frac{a_h}{n_h} 100 \end{aligned}\right\} \qquad 6.13$$

en ella p_h es el estimador del porcentaje en el estrato h-ésimo. La variancia del estimador p_{est} queda dada por la expresión siguiente:

$$V(p_{est}) = \sum_h \frac{N_h^2 \, (N_h - n_h) \, (P_h \, Q_h)}{N^2 \, (N_h - 1) \, n_h} \qquad 6.14$$

y un estimador de esta variancia es:

$$\hat{V}(p_{est}) = \sum_h \frac{N_h^2}{N^2} \frac{N_h - n_h}{(n_h - 1)N_h} \, p_h q_h \qquad 615$$

Las expresiones 6.13 a 615 se derivan sin mayores complicaciones considerando que para porcentajes la variable aleatoria y_{hi} toma los valores de *uno o cero,* por lo cual sólo hay que copiar las expresiones del apartado 3.6 en términos de muestreo estratificado.

Si el tamaño total n de la muestra se afija proporcionalmente al tamaño relativo de los estratos (afijación proporcional) las ecuaciones 6.13 y 6.15 se simplifican a la forma siguiente:

$$\left.\begin{aligned} p_{est} &= \left(\frac{1}{n}\sum a_h\right) 100 \\ \hat{V}(p_{est}) &= \frac{1-f}{N} \sum_h \frac{N_h^2}{nN_h - N} p_h q_h \end{aligned}\right\} \quad 6.16$$

En la tabla 6.1 se resumen los diferentes estimadores que resultan a nivel estrato y a nivel población; cuando se ha practicado una estratificación y dentro de cada estrato se usa muestreo aleatorio simple y la muestra total n se afija proporcionalmente al tamaño relativo $\frac{N_h}{N}$ de los estratos.

Ejemplo 6.3 Un grupo de médicos está desarrollando una encuesta en una ciudad para estimar el número total de niños en la población que tienen, o que han padecido de tosferina durante el último mes. En ella existen tres instituciones gubernamentales que ofrecen asistencia médica a los niños. Cada una cuenta con un registro de los enfermos atendidos en ellas en el mes en cuestión. Una inspección a estos listados muestra que el número total de casos registrados en las instituciones A, B y C y en el último mes fue de 20, 17 y 142 respectivamente. El resto de los niños enfermos en la ciudad son atendidos por médicos particulares, de los cuales se tienen dos listados con 450 y 800 nombres y direcciones respectivamente.

Se decide formar tres estratos. En el primero aparecen las instituciones gubernamentales, en el segundo el listado D con 450 médicos particulares y en el tercero, el listado E con los 800 restantes. La separación de médicos particulares en dos estratos fue debida únicamente a que sus listados mostraban numeraciones independientes.

Tabla 6.3

Estimadores aplicables a muestreo estratificado, con afijación proporcional y muestreo aleatorio simple en cada estrato.

	ESTRATO		*POBLACION*	
Media	$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$	$\hat{V}(\bar{y}_h) = (1 - f_h)\frac{s_h^2}{n_h}$	$\bar{y}_{est} = \frac{1}{n} \sum_{h=1}^{L} \sum_{i=1}^{n_h} y_{hi}$	$\hat{V}(\bar{y}_{est}) = \frac{1 - f}{nN} \sum_{h=1}^{L} N_h s_h^2$
Total	$N_h\bar{y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}$	$\hat{V}(N_h\bar{y}_h) = N(N\text{-}n)\frac{s_h^2}{n_h}$	$N\bar{y}_{est} = \frac{N}{n} \sum_{h=1}^{L} \sum_{i=1}^{n_h} y_{hi}$	$\hat{V}(N\bar{y}_{est}) = \frac{N(1 - f)}{n} \sum_{h=1}^{L} N_h s_h^2$
Porcentaje	$p_h = \frac{a_h}{n_h} 100$	$\hat{V}(p_h) = \frac{N_h - n_h}{N_h(n_h - 1)} p_h q_h$	$p_{est} = \left[\frac{1}{n} \sum_{h=1}^{L} a_h\right] 100$	$\hat{V}(p_{est}) = \frac{1 - f}{N} \sum_{h=1}^{L} \frac{N_h^2}{nN_h - N} p_h q_h$

El grupo de médicos que desarrolla el estudio está dispuesto a visitar a 30 médicos particulares de entre los 1 250 existentes en sus listas D y E. Así, usando afijación proporcional, los tamaños de la muestra en cada estrato son como sigue:

$$n_2 = \frac{450}{1\,250}\,30 = 10.8,$$

por lo cual tomamos $n_2 = 11$, y n_3 vale:

$$n_3 = 30 - n_2 = 19$$

Usando las listas D y E, se efectúa el sorteo, se marcan los médicos incluidos en cada muestra y al visitarlos se obtiene la información de la tabla 6.4

Tabla 6.4

Estrato No.2

Médico	1	2	3	4	5	6	7	8	9	10	11
Casos	1	1	1	5	3	1	5	2	5	1	1

Estrato No.3

Médico	1	2	3	4	5	6	7	8	9	10
Casos	7	2	1	3	1	1	1	1	1	3

Médico	11	12	13	14	15	16	17	18	19
Casos	3	1	1	2	1	1	1	1	1

Un estimador del número total de niños con tosferina en la población y en el período de tiempo en cuestión se obtiene mediante la suma del número total arrojado por el censo practicado en las listas de las instituciones gubernamentales, más las estimaciones de los totales en los estratos 2 y 3:

$$\hat{Y} = \text{Total en las instituciones gubernamentales} +$$

$$+ N_2\bar{y}_2 + N_3\bar{y}_3$$

$$= 20 + 17 + 142 + 450\left(\frac{26}{11}\right) + 800\left(\frac{33}{19}\right)$$

$$= 2\,632 \text{ niños.}$$

Y un estimador de la variancia de este total, la obtenemos mediante la suma ponderada de las variancias de cada estrato, siendo ésta de cero para el estrato 1, ya que se practicó un censo; entonces:

$$V(Y) = 0 + N_2^2\ (1 - f_2)\frac{s_2^2}{n_2} + N_3^2\ (1 - f_3)\frac{s_3^2}{n_3}$$

$$+ (450)^2\ (1 - \frac{11}{450})\ \frac{94 - \frac{(26)^2}{11}}{11\ (10)}$$

$$+ (800)^2\ (1 - \frac{19}{800})\ \frac{97 - \frac{(33)^2}{19}}{19\ (18)}$$

$$= 58\,445.35 + 72{,}499$$

$$= 130\,944 \text{ (niños)}^2$$

Y su error estándar es de 362 niños.

En seguida se presenta un ejercicio tal que su solución va evolucionando a través de varios ejercicios seriados, moviendo en cada ocasión el esquema de selección hasta llevarlo a submuestreo en el capítulo 8. Introduce algunas ideas que han aparecido en las hojas previas pero que indudablemente, requieren de mayor explicación. En particular se hace referencia al muestreo sistemático, el cual es una manera de selección tal que se elige un arranque, el cual viene siendo una unidad muestral, y a partir de ella se localiza el resto mediante saltos de longitud constante. Por ejemplificar, si en una lista de las 70 viviendas que existen en una manzana (conglomerado) se obtiene como arranque a la vivienda (elemento) número 3, o que se encuentra localizada en el tercer renglón, y se debe seleccionar a una vivienda de cada diez, entonces, las viviendas en la muestra son:

3, 17, 27, 37, 47, 57 y 67. Esto es una selección sistemática la cual termina con un tamaño de muestra de 7.

Ejemplo 6.4. Se desea hacer un estudio sobre el personal que labora en una fábrica que cuenta con edificios en 15 estados del país. El estudio se refiere a opiniones y actitudes de los empleados y obreros. En la muestra se decea tener representados a 1 de cada 30 empleados y existen en total 42 090 de ellos. Administrativamente, el personal de cada estado es independiente de la oficina central en cuanto o su nómina, de tal manera que, las listas de obreros y empleados se tienen para cada uno de ellos. La distribución del personal en cada entidad aparece en la tabla 6.5.

TABLA 6.5

Entidad	*No. de empleados*	*No. de hojas*
1. Guanajuato	19 043	635
2. Hidalgo	429	15
3. Jalisco	5 010	167
4. Michoacán	1 114	38
5. Morelos	721	25
6. Nayarit	474	16
7. Nuevo León	4 415	148
8. Oaxaca	450	15
9. Puebla	2 750	92
10. Querétaro	487	17
11. Quintana Roo	150	5
12. S. Luis P.	925	31
13. Sinaloa	2 800	94
14. Sonora	2 900	97
15. Tabasco	422	15
	42 090	1 410

Para obtener a uno de cada treinta empleados en la muestra se requiere una muestra total de $n = 1\ 403$ (¿por qué?), los cuales serán sorteados a partir de algún esquema de selección apropiado.

Ejemplo 6.5. (Continuación del ejemplo 6.4) Entonces, debemos elegir a 1 403 empleados de entre los 42 090 esparcidos en las quince entidades federativas. Disponemos de quince listados de empleados, uno por cada entidad federativa. Si deseamos obtener una selección

aleatoria simple, pues, habría necesidad de unir a esos listados de tal manera de asegurar una identificación única para cada empleado, y después efectuar la selección. Esto sería embarazoso (intente el método). Si esto fuera necesario, posiblemente sería más adecuado recurrir a una selección sistemática (capítulo 7) con un intervalo igual a 30, de tal manera que seleccionaríamos a un número aleatorio entre 1 y 30, el cual tomamos como arranque (supongamos que fue el 15) y, las instrucciones para la selección serían las siguientes:

i) Ordene los listados, digamos alfabéticamente.
ii) Encuentre el renglón número 15 del primer listado: este empleado está en la muestra.
iii) A partir del empleado número 15, vuelva a contar del 1 al 30. El empleado número 30 está en la muestra.
iv) Continúe la cuenta del 1 al 30 hasta agotar todas las listas.

La selección anterior puede ser superada mediante una estratificación en la cual cada estrato es definido como un estado. Tendríamos 15 estratos, tales que, usando afijación proporcional sus tamaños de muestra serían:

$$n_1 = \frac{19\ 043}{42\ 090}\ 1\ 403 \doteq 635$$

$$n_2 = \frac{429}{42\ 090}\ 1\ 403 \doteq 14$$

$$n_3 = \frac{5\ 010}{42\ 090}\ 1\ 403 = 167$$

$$n_4 = \frac{1\ 114}{42\ 090}\ 1\ 403 \doteq 37$$

$$n_5 = \frac{721}{42\ 090}\ 1\ 403 \doteq 24$$

$$n_6 = \frac{474}{42\ 090}\ 1\ 043 = 16$$

$$n_7 = \frac{4\ 415}{42\ 090}\ 1\ 403 = 147$$

$$n_8 = \frac{450}{42\ 090}\ 1\ 043 = 15$$

$$n_9 = \frac{2\ 750}{42\ 090}\ 1\ 403 = 92$$

$$n_{10} = \frac{487}{42\ 090}\ 1\ 403 \doteq 16$$

$$n_{11} = \frac{150}{42\ 090}\ 1\ 403 = 5$$

$$n_{12} = \frac{925}{42\ 090}\ 1\ 403 = 31$$

$$n_{13} = \frac{2\ 800}{42\ 090}\ 1\ 403 \doteq 93$$

$$n_{14} = \frac{2\ 900}{42\ 090}\ 1\ 403 \doteq 97$$

$$n_{15} = \frac{422}{42\ 090}\ 1\ 403 = 14$$

Ejemplo 6.6. (Continuación del ejemplo 6.5) Para efectuar la selección anterior (Ejemplo 6.5) en cada uno de los estratos y en el caso concreto del primero de ellos, podemos obtener una muestra aleatoria simple de tamaño 635 de entre los 19 043 empleados en Guanajuato. Esto equivale a obtener 635 números aleatorios diferentes entre 1 y 19 043 (si estuvieran numerados del 1 al 19 043). Para continuar con Hidalgo, habría que elegir a 14 números aleatorios diferentes entre 1 y 429, y así sucesivamente.

Como en el ejemplo 6.5 la selección aleatoria anterior pudo haberse efectuado mediante una selección sistemática con fracción de muestreo 1 de cada 30. (¿Realmente es la misma fracción de muestreo 1/30 para cada estrato?, ¿por qué?)

6.6. EJERCICIOS

6.1 Demuestre que en la ecuación 6.3 el estimador de la variancia de la media estratificada resulta insesgado cuando dentro de cada estrato la selección se hace con muestreo aleatorio simple.

6.2 Se desea tomar una muestra de un archivo de dos millones de nombres ordenado alfabéticamente según el primer apellido, para hacer estimaciones sobre características que no tienen relación con el orden del archivo. Cada bloque de nombres cuyo primer apellido tiene la misma letra está numerado consecutivamente empezando en 1 y terminando con el número del último, del bloque.

Con respecto a la dificultad de selección, *i*) ¿elegiría usted una muestra aleatoria? , ¿elegiría una muestra mediante una estratificación? , ¿por qué?

6.3 En el ejemplo 6.2 considere que el tamaño total de la muestra es de 120. ¿Qué tamaño de muestra le corresponde a cada estrato bajo *i*) afijación proporcional, *ii*) afijación igual? *

6.4 Si en el ejemplo 6.2 se desea estimar el número medio de tapas por máquina a nivel estrato con un error que no exceda al 10% y una confianza del 95%, ¿qué tamaño de muestra necesita para cada estrato y cuál es la muestra total? Suponga que las estimaciones de las medias por estrato y de sus variancias son las siguientes:

$$\bar{y}_1 = 1\,900, \qquad \bar{y}_2 = 6\,000, \qquad \bar{y}_3 = 20\,000$$

$$s_1^2 = 1\,800\,000, \qquad s_2^2 = 12\,000\,000, \qquad s_3^2 = 30\,000\,000$$

Tabla 6.6

Ciudad	*No. de Establecimientos*	*Establecimientos muestreados*	*Producción (kilos)*	*No. de empleados*
A	48	5	600, 800, 900, 700 1 200	2, 2, 2, 2, 3
B	127	10	900, 900, 500, 300 800, 600, 900, 800, 800, 700	2, 2, 1, 1, 2 1, 2, 2, 2, 2
C	390	10	500, 400, 700, 900 700, 1 100, 400 800, 800, 500	1, 1, 1, 2, 1 3, 2, 2, 1. 1

Estime la producción media del día en kilos por establecimiento en cada ciudad y para las tres ciudades, así como intervalos confidenciales del 95% para su estimación global.

* Es aquélla que asigna el mismo tamaño de muestra a cada estrato.

6.5 En el ejercicio 6.4 indique cuáles serían los estimadores a usar *i*) a nivel estrato, *ii*) a nivel global.

6.6 En tres ciudades de México se llevó a cabo una encuesta sobre la industria de la fabricación de tortilla, se estratificaron los establecimientos según la ciudad en la que operaban, y dentro de cada uno de los estratos se tomaron las muestras aleatorias que se indican en la tabla 6.6, registrándose su producción en kilos para un día particular y su número de empleados.

6.7 Estime la producción total en kilos de tortilla para las tres ciudades del ejercicio 6.6 y calcule intervalos de confianza del 95% para su estimación.

6.8 En el ejercicio 6.6 estime el número medio de empleados por tortillería e indique intervalos de confianza del 95%.

6.9 Demuestre que el estimador 6.8 es consistente e insesgado del total poblacional.

6.7 EL TAMAÑO DE LA MUESTRA

El problema de derivar las expresiones para obtener el tamaño de muestra cuando se ha introducido una estratificación en la población bajo estudio y se desea estimar una media, un total o un porcentaje, se desarrolla con relativa facilidad si se considera razonable o adecuado el supuesto de normalidad en la distribución de los estimadores en cada uno de los casos en consideración. Y esto es lo que se hace usualmente. En estas condiciones, la variancia del estimador del parámetro en consideración se tomará igual al cuadrado del cociente definido entre el error permitido d y el valor de la abscisa t encontrado en las tablas de la distribución normal y tal que nos deja al centro de la distribución un área igual a la confianza con que requerimos la estimación.

Tamaño de la muestra para la estimación de medias. Si la estimación del parámetro en estudio se desea a nivel global, es decir, se desea, por ejemplo, hacer una estimación del número medio de hijos por empleado, para los empleados de cierta institución contenida en treinta edificios tales que, para facilitar la selección, cada uno de ellos se define como un estrato, entonces, para el caso de una media poblacional (apartado 6.2) el estimador correspondiente es la media estratificada y utilizando la expresión 6.2 que muestra la variancia de la media estratificada, tenemos lo siguiente:

$$\left(\frac{d}{t}\right)^2 = V = \sum_h \frac{N_h^2}{N^2} \frac{S_h^2}{n_h} - \sum_h \frac{N_h}{N^2} S_h^2$$

y además:

$$n = n_1 + n_2 + \ldots + n_L \qquad 6.17$$

En ambas expresiones aparecen los tamaños de muestra correspondientes a cada estrato y deseamos resolverlas para el tamaño total n de la muestra; para ello supongamos que n se afija entre los diferentes estratos de manera proporcional luego, según esta afijación:

$$n_h = \frac{N_h}{N} n = (W_h)\, n$$

es decir, la muestra la asignamos a los estratos en proporción directa a su tamaño relativo. Usando esta afijación en la expresión 6.17 y simplificando obtenemos el resultado siguiente:

$$V = \sum_h \frac{N_h^2}{N^2} \frac{S_h^2}{(W_h)n} - \sum_h \frac{N_h}{N^2} S_h^2$$

Resolviendo para n y definiendo a n_0 como se indica, obtenemos las expresiones 6.18 para el cálculo del tamaño de la muestra cuando se desea estimar una media poblacional y la muestra se afija proporcionalmente.

$$n_0 = \frac{1}{NV} \sum_h N_h S_h^2$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \qquad 6.18$$

Una vez que se ha calculado n según estas expresiones 6.18, el tamaño de muestra para cada estrato se calcula según 6.19.

$$n_h = \left(\frac{N_h}{N}\right) n,\ h = 1, 2, \ldots, L \qquad 6.19$$

Tamaño de muestra para la estimación de totales. Para derivar las expresiones correspondientes al cálculo del tamaño de la muestra

en el caso de la estimación de un total poblacional y cuando se usa afijación proporcional, el proceso anterior se debe repetir partiendo de la nueva expresión de la variancia. Los resultados que se obtienen son las expresiones 6.20:

$$n_0 = \left(\frac{N}{V}\right) \sum_h N_h S_h^2 \quad ; \qquad n = \frac{n_0}{1 + \frac{n_0}{N}} \qquad 6.20$$

Tamaño de la muestra para la estimación de porcentajes. Las expresiones para calcular el tamaño de la muestra para la estimación de un porcentaje bajo afijación proporcional son las siguientes:

$$\left.\begin{aligned} n_0 &= \frac{1}{NV} \sum_h N_h P_h Q_h \\ n &= \frac{n_0}{1 + \frac{n_0}{N}} \end{aligned}\right\} \qquad 6.21$$

En 6.21, la variancia V es igual a $(\frac{d}{t})^2$, en la cual d está expresada en porcentaje de la misma manera que P_h y Q_h. Para una situación específica, debemos notar que si las estimaciones se desearan de tal manera que cumplieran con determinada precisión por estrato, sería necesario estimar los tamaños de muestra de manera independiente para cada uno de ellos.

Ejemplo 6.7 En el ejemplo 6.2 sobre la compañía manufacturadora de tapas de plástico, se desea estimar el número medio de tapas producidas por máquina con un error menor o a lo más igual a 500 tapas y a una confianza el 95%. Como el rendimiento por unidad depende del tipo de máquina empleada, resulta natural definir una estratificación según su manera de funcionamiento, y además esta estratificación es factible desde el punto de vista práctico. Así, como ya se indicó en el ejemplo 6.2 formamos tres estratos cuyos tamaños son: $N_1 = 240$, $N_2 = 100$ y $N_3 = 60$.

El tamaño de muestra necesario usando afijación proporcional, debe ser calculado con las expresiones 6.18. Como estimaciones de las variancias S_h^2 por estrato, podemos tomar las estimadas del mismo ejemplo 6.2, a saber: $s_1^2 = 180\,909$, $s_2^2 = 2\,572\,000$ y $s_3^2 = 10\,570\,000$.

Entonces:

$$n_0 = \frac{(240)(180\,909) + (100)(2\,572\,000) + (60)(10\,570\,000)}{400\,(\frac{500}{2})^2}$$

$$= 37.4$$

y al considerar la corrección por población finita tenemos:

$$n = \frac{37.4}{1 + (\frac{37.4}{400})} = 34.2$$

Tomamos $n = 35$

Los tamaños de muestra para cada estrato resultan ser de:

$$n_1 = \frac{240}{400}\,35 = 21$$

$$n_2 = \frac{100}{400}\,35 = 8.75$$

$$n_3 = \frac{60}{400}\,35 = 5.25$$

Redondeando a n_2 al entero superior obtenemos:

$$n_1 = 21, n_2 = 9, n_3 = 5$$

Si en lugar de una precisión global, se pidiera estimar la media poblacional de manera que las estimaciones por estrato tuvieran la misma precisión que la solicitada inicialmente para la global y con la cual se hicieron los cálculos anteriores, los tamaños de muestra serían:

$$n_{01} = \frac{S_1^2}{V} = \frac{180\,909}{(\frac{500}{2})^2} = 2.89$$

y también:

$$n_1 = \frac{2.89}{1 + \frac{2.89}{240}} = 2.86. \quad \text{Tomamos } n_1 = 3$$

$$n_{02} = \frac{2\,572\,000}{(\frac{500}{2})^2} = 41.15$$

$$n_2 = \frac{41.15}{1 + \frac{41.15}{100}} = 29.15. \quad \text{Tomamos } n_2 = 30$$

$$n_{03} = \frac{10\,570\,000}{(\frac{500}{2})^2} = 169.12$$

$$n_3 = \frac{169.12}{1 + \frac{169.12}{60}} = 44.29. \quad \text{Tomamos } n_3 = 45$$

El tamaño total de la muestra resulta ser $n = 3 + 30 + 45 = 78$, superior en 43 unidades al calculado (35) para la precisión global.

6.8 AFIJACION OPTIMA

Dado un tamaño de muestra total n para ser afijado entre los estratos diferentes, éste puede ser repartido entre ellos con cualquier criterio que especifiquemos. Sin embargo, sabemos que las variancias de los estimadores en cada estrato son función del tamaño de muestra asignado a cada uno de ellos, por ello es deseable que el tamaño de muestra n_h sea grande. Por otro lado, los costos en que se incurre por concepto de transportación dentro de cada estrato son variables. En algunos estratos resulta barato localizar a las unidades en la muestra y efectuar la(s) medición(es) correspondiente(s), en otros, resulta más o menos caro y aun, en algunos casos, puede ser muy costoso; en estas condiciones no es conveniente que los tamaños de muestra n_h sean grandes. Estos dos factores, a saber *precisión* y *costos* por estrato, son usados como lineamientos para afijar de manera *óptima* la muestra total n entre los estratos.

En un caso mantenemos *fija la precisión* solicitada, y nos preguntamos por los tamaños de muestra por estrato tales que arrojan un *costo total mínimo.* En otro caso, mantenemos *fijo al costo total* y determinamos los tamaños de muestra tales que *minimizan la variancia.*

Bajo *afijación óptima,* las expresiones que se obtienen para el tamaño de la muestra resultan ser usualmente muy complicadas para su aplicación práctica. Y aunque se trata de la mejor afijación, a menudo se prefiere sacrificar precisión con el propósito de manejar expresiones de afijación más simples y que, sin embargo, en la mayoría de los casos, no se alejan demasiado de la afijación óptima. Usando afijación óptima, se puede demostrar (Cochran, teorema 5.8) que si la fracción de muestreo $\frac{n_h}{N_h}$ es ignorada para cada h, la variancia de la media estratificada es menor, o a lo más igual, a la variancia de la media estratificada es menor, o a lo más igual, a la variancia de la misma bajo afijación proporcional, y que ésta a su vez es menor o igual que la variancia de la media muestral bajo una selección aleatoria. En las condiciones del teorema y si estratificamos esto significa que lo peor que nos puede ocurrir es equivalente al caso en el cual hubiéramos seleccionado aleatoriamente a las unidades sin estratificar (capítulo 3). En cualquier otra situación, con afijación proporcional obtendremos resultados mejores que si hubiéramos practicado una selección aleatoria y, evidentemente, los *mejores resultados* se obtendrán mediante el uso de la *afijación óptima.*

Como hemos visto en este capítulo, la afijación proporcional es muy simple de aplicar y en ocasiones lleva a estimadores autoponderados, como ocurrió en el apartado 6.3; además, arroja precisiones que se encuentran entre aquélla correspondiente a la afijación óptima y la correspondiente a una selección aleatoria como muestra el teorema antes referido. Por ello, esta afijación es ampliamente usada y recomendable.

Los conceptos que intervienen o que contribuyen al costo total de una encuesta son muy variados. En algunos casos el proceso para conseguir o para construir el marco muestral resulta sustancialmente costoso; por ejemplo, es necesario contar con un listado que contenga aquellas industrias de transformación que interesan para poder desarrollar una encuesta sobre ellas a nivel nacional. Parece relativamente fácil acudir a tal organismo público, en el cual

sabemos que por ley deben estar registradas las empresas y solicitar un listado de ellas. Si de esta manera pudiéramos conseguir un listado y, además, ocurriéramos a otros organismos, los cuales igualmente deben contar con los mismos listados, obtuviéramos las copias respectivas y procediéramos a compararlos, encontraríamos una divergencia entre ellos; por ejemplo, varias industrias que aparecen en un listado no están en el otro, o si aparecen, están registradas en una actividad económica diferente. Si ésta fuera la situación, podríamos tratar de construir un solo listado combinando los contenidos de cada uno de ellos, aunque en ocasiones son miles o decenas de miles de nombres y es un problema complicado el tratar de conciliarlos. De manera que obtener un marco muestral adecuado puede significar una tarea laboriosa que requiere mucho tiempo, equipo y personal.

En otras ocasiones, conseguir el equipo necesario que demanda el método de medición aprobado o entrenar adecuadamente al personal que intervendrá en las diferentes etapas, se torna una actividad crítica en el desarrollo general de la encuesta. También, en ocasiones, el llegar a un acuerdo sobre el cuestionario que deberá ser usado en definitiva, requiere de una serie de reuniones largas convocando a personal especializado y que derivan en un retraso sustancial para todo el proceso.

Los comentarios anteriores los hacemos para hacer resaltar el hecho de que en muchas ocasiones no vale la pena estar considerando refinamientos matemáticos que llevan a la obtención de tamaños de muestra óptimos o a métodos de estimación complejos y rebuscados que aunque teóricamente producen estimaciones muy precisas, las aplicaciones o usos prácticos no los requieren o que son tan complejos que las personas que deben interpretarlos se confunden provocando que los cálculos se vuelvan más largos y que todo el proceso quede más sujeto a errores.

6.9 ESTIMACION DE MEDIAS Y DE TOTALES EN SUBPOBLACIONES DE TAMAÑO CONOCIDO

En el apartado 5.4 y 5.5 fueron presentados dos tipos de estimadores para aplicarse según fuera conocido el número de unidades que conforman al dominio o que no lo fuera. Si el tamaño resulta conocido, los estimadores presentados se derivan de la media muestral en el dominio de interés d-ésimo; en el caso

contrario, no se conocen los tamaños, el estimador cambia y la penalidad o la consecuencia por su desconocimiento es un incremento en las variancias, es decir, los estimadores son menos precisos.

Cuando deseamos hacer estimaciones globales para el dominio d-ésimo y se ha practicado una estratificación en la población, es usual que los *dominios de estudio o subpoblaciones* queden *repartidos* en varios o en todos los estratos. Supongamos que en el estrato h-ésimo de tamaño N_h existen N_{hd} unidades pertenecientes al dominio d-ésimo, y que de ellas, resultan seleccionadas en la muestra n_{hd}, entonces, podemos usar como *estimador de la media en ese dominio* a la expresión siguiente:

$$\hat{\bar{Y}}_d = \frac{\sum_h N_{hd} \sum_{i=1}^{n_{hd}} \frac{y_{hdi}}{n_{hd}}}{\sum_h N_{hd}} = \frac{\sum_h N_{hd}\bar{y}_{hd}}{\sum_h N_{hd}} \qquad 6.22$$

y como estimador del total en el mismo dominio a la ecuación 6.23.

$$\hat{Y}_d = \sum_h N_{hd} \sum_{i=1}^{n_{hd}} \frac{y_{hdi}}{n_{hd}} = \sum_h N_{hd}\bar{y}_{hd} \qquad 6.23$$

Las variancias respectivas de las expresiones 6.22 y 6.23 son las siguientes:

$$V(\hat{\bar{Y}}_d) = \frac{1}{(\sum_h N_{hd})^2} \sum_h \frac{N_{hd}^2 S_{hd}^2}{n_{hd}} \left(1 - \frac{n_{hd}}{N_{hd}}\right) \qquad 6.24$$

$$V(\hat{Y}_d) = \sum_h \frac{N_{hd}^2 S_{hd}^2}{n_{hd}} \left(1 - \frac{n_{hd}}{N_{hd}}\right) \qquad 6.25$$

Y los estimadores de estas variancias son las mismas expresiones, pero en ellas S_{hd}^2 es sustituido por su estimador s_{hd}^2.

6.10 EL ESTIMADOR DE RAZON COMBINADO

En varias ocasiones resulta útil o necesario emplear estimadores de razón habiéndose hecho una estratificación. Un estimador gene-

ralmente usado y que es de esta clase, es el denominado *estimador de razón combinado.* Para su empleo se construyen estimaciones estratificadas de cada uno de los parámetros involucrados y posteriormente se *combinan* para formar el estimador correspondiente:

$$\left.\begin{aligned} \hat{Y}_{RC} &= \frac{\hat{Y}_{est}}{\hat{X}_{est}} X = \frac{\bar{y}_{est}}{\bar{x}_{est}} X \\ \hat{Y}_{RC} &= \frac{\sum_h N_h \bar{y}_h}{\sum_h N_h \bar{x}_h} X \end{aligned}\right\} \qquad 6.26$$

Como muestra la expresión 6.26 para la aplicación del estimador de razón combinado se requiere conocer el *valor total del parámetro X* el cual es usado como variable auxiliar, su conocimiento es necesario de manera global sin hacer distinción entre estratos, y esto es deseable porque generalmente ese total es conocido o es relativamente fácil obtenerlo, aunque no por estrato, sino para toda la población. Por otra parte, aunque el estimador en la expresión 6.26 está sujeto a sesgo, éste usualmente es despreciable y, generalmente, puede ser usado aun con muestras pequeñas en cada estrato.

Se puede demostrar (ejercicio 6.12) que si el tamaño de muestra n es grande, la variancia del estimador de razón combinado es de la forma siguiente:

$$V(\hat{Y}_{RC}) = \sum_h \frac{N_h^2(1 - f_h)}{n_h} \frac{1}{N_h - 1} \sum_{i=1}^{N_h} ((y_{hi} - \bar{Y}_h) - R(x_{hi} - \bar{X}_h))^2 \qquad 6.27$$

Un estimador de la variancia anterior es:

$$\hat{V}(\hat{Y}_{RC}) = \sum_h \frac{N_h^2(1 - f_h)}{n_h(n_h - 1)} \sum_{i=1}^{n_h} ((y_{hi} - \hat{R}x_{hi}) - (\bar{y}_h - \hat{R}\bar{x}_h))^2 . \qquad 6.28$$

donde $$\hat{R} = \frac{\bar{y}_{est}}{\bar{x}_{est}}$$

En la situación en que sólo exista un estrato, el estimador Y_{RC} en la expresión 6.26 coincide con el presentado en la ecuación 5.6

Ejemplo 6.8. En las capitales de varios estados del país existen gaveteros que contienen un número variable de tarjetas, una por cada miembro de la familia: padre, madre e hijos. Usando los gaveteros de los diferentes estados se desea estimar el número total de tarjetas que pertenecen a los hijos.

Como se mencionó en el apartado 5.2, el número de tarjetas en cada gaveta y que pertenecen a los hijos, está correlacionado positivamente con el número total de tarjetas en la misma gaveta, excepto en el caso de que las familias no tengan hijos o que tengan 1 o 2, en el cual al aumentar el número total de tarjetas en la gaveta no aumenta el número de tarjetas correspondientes a los hijos. Además supongamos que existe un único proveedor de esas tarjetas, y así de los suministros o abastecimientos que se han hecho a nivel nacional, conocemos el número de cajas y el peso de las tarjetas contenidas en ellas y supongamos además que no existen pérdidas y que, en general, podemos conocer el peso total de las tarjetas en todos los estados.

Si efectuamos una estratificación geográfica por estado y dentro del *h-ésimo* de ellos tomamos una muestra aleatoria de n_h gaveteros de entre los N_h, y para cada gavetero en la muestra contamos el número de tarjetas asociadas a los hijos (y_{hi}) y registramos el peso en gramos de todas las tarjetas en esa gaveta (x_{hi}), para estimar el total deseado podemos usar el estimador de razón combinado de la ecuación 6.26.

6.11 ESTIMACION DE MEDIAS Y DE TOTALES EN SUBPOBLACIONES DE TAMAÑOS DESCONOCIDOS

En el apartado 6.9 vimos el caso en el cual los tamaños de los dominios son conocidos, ahora tratamos la situación más general en la cual ellos son desconocidos. Por facilidad presentamos inicialmente el estimador del total y posteriormente el estimador de la media.

Como se vio en 5.5, para formar el estimador del total en el estrato *h-ésimo* (ecuación 5.18) a la suma de las observaciones obtenidas de las unidades en la muestra y en el dominio de interés, se le afecta del factor de expansión $\frac{N_h}{n_h}$. Entonces, el estimador del total en el dominio *d*-ésimo es:

$$\hat{Y}_d = \sum_h \hat{Y}_{hd} = \sum_h \frac{N_h}{n_h} \sum_{i=1}^{n_{hd}} y_{hdi} \qquad 6.29$$

Y usando los resultados del mismo apartado 5.5 podemos derivar sin dificultad un estimador de la variancia de $\hat{Y}_d$; éste es el siguiente:

$$\hat{V}(\hat{Y}_d) = \sum_h \frac{N_h^2 (1 - f_h)}{n_h (n_h - 1)} \left(\sum_{i=1}^{n_{hd}} y_{hdi}^2 - \frac{\left(\sum_{i=1}^{n_{hd}} y_{hdi} \right)^2}{n_h} \right) \qquad 6.30$$

ahora, la media poblacional de la característica en estudio en el dominio *d-ésimo* está definida como la suma de las observaciones en él y esta cantidad dividida entre el total de unidades en el mismo. De acuerdo a esta definición, en la ecuación 6.29, tenemos un estimador de su numerador, pero desconocemos el denominador en la misma, total de unidades en el dominio. Para este tipo de unidades cada una de ellas en la muestra tiene como observación uno o cero, según que pertenezca o no al dominio de interés. Entonces su número total en la muestra y para el estrato h-ésimo es n_{hd}, y el total de unidades estimado en el mismo estrato es:

$$\frac{N_h}{n_h} n_{hd}$$

luego, la expresión siguiente en un estimador del valor medio.

$$\hat{\bar{Y}}_d = \frac{\hat{Y}_d}{\hat{N}_d} = \frac{\sum_h \frac{N_h}{n_h} \sum_{i=1}^{n_{hd}} y_{hdi}}{\sum_h \frac{N_h}{n_h} n_{hd}} \qquad 6.31$$

Como n_{hd}, número de unidades en la muestra y favorables al dominio *d-ésimo* es una variable aleatoria, la expresión 6.31 es un estimador de razón y es posible verificar (ejercicio 6.13), que tiene la misma forma que aquélla del estimador de razón combinado en el apartado 6.10. Por ello, un estimador de su variancia es el siguiente:

$$\hat{V}(\hat{\bar{Y}}_d) \doteq \frac{1}{\left(\sum_h \frac{N_h}{n_h} n_{hd} \right)^2} \sum_h \frac{N_h^2 (1 - f_h)}{n_h (n_h - 1)}$$

$$\left[\sum_i (y_{hdi} - \bar{y}_{hd})^2 + n_{hd} \left(1 - \frac{n_{hd}}{n_h} \right) (\bar{y}_{hd} - \hat{\bar{Y}}_d)^2 \right] \qquad 6.32$$

En el caso de afijación proporcional las expresiones 6.31 y 6.32 se simplifican a las formas siguientes:

$$\hat{\bar{Y}}_d = \frac{\sum_h \sum_{i=1}^{n_{hd}} y_{hdi}}{\sum_h n_{hd}} \qquad 6.33$$

$$\hat{V}(\hat{\bar{Y}}_d) \doteq \frac{n(1-f)}{(\sum_h n_{hd})^2} \sum_h \frac{N_h}{N_h n - N} \left(\sum_i (y_{hdi} - \bar{y}_{hd})^2 + n_{hd} \left(1 - \frac{n_{hd}}{n_h}\right) \left(\bar{y}_{hd} - \hat{\bar{Y}}_d \right)^2 \right) \qquad 6.34$$

Ejemplo 6.9 En los predios ejidales de un estado de la República Mexicana se desea desarrollar un estudio sobre la utilización de equipo de tracción mecánica. Para ello se dispone de una serie de listados por municipio de ese tipo de predios, conteniendo los nombres y alguna otra información de ellos sobre su localización.

El número total de predios ejidales por municipio y para todos ellos aparece en la tabla 6.7.

Tabla 6.7

Municipio	1	2	3	4	5	6	7
No. de predios	400	175	1 320	4 200	150	3 075	800
Municipio	8	9	10	11	12	13	
No. de predios	40	392	5 190	1 730	745	815	

El estudio en la entidad federativa en cuestión se desarrolla con el propósito de construir varias matrices de insumo-producto. Para ello, la entidad federativa es dividida en cuatro zonas, y para cada una de ellas se solicita una matriz. Cada zona se define como una agrupación de municipios y sus constituciones en función de ellos y del número de predios ejidales son las de la tabla 6.8

Se desea estimar el número medio de tractores y de camiones de carga existentes por predio, incluyendo para ello a todos los predios, tengan o no tengan el vehículo. Y para ello se dispone del

dinero suficiente para llevar a cabo 200 entrevistas en todo el estado.

Dado que otras estimaciones requeridas y las cuales no apare-

Tabla 6.8

	Municipios	*No. de predios ejidales*
Zona 1	7, 4, 9 y 13	6 207
Zona 2	5 y 6	3 225
Zona 3	1, 2, 3 y 10	7 085
Zona 4	Resto del estado	2 515
Total de predios		19 032

cen en este ejemplo eran solicitadas a nivel de zona, la estratificación inmediata fue aquella en la cual se define a cada estrato como una zona. Al desarrollar las entrevistas, se encontrará que algunos predios ejidales no cuentan con el equipo buscado, otros tendrán una o dos unidades y otros más de dos. Si se tuviera previamente alguna información sobre esto, sería adecuado formar nuevas agrupaciones de predios dentro de cada zona, para así, por ejemplo, definir una nueva o nuevas estratificaciones que fueran más acertadas y aun se pudiera proporcionar las estimaciones como se requieren. Si no tenemos esta información consideramos adecuada la estratificación anterior y los tamaños de muestra por estrato usando afijación proporcional son como sigue:

$$n_1 = 6\,207\,\frac{(200)}{19\,032} = 65.23 \text{ ; tomemos } n_1 = 65^*$$

$$n_2 = 3\,225\,\frac{(200)}{19\,032} = 34$$

$$n_3 = 7\,085\,\frac{(200)}{19\,032} = 74.45 \text{ ; tomemos } n_3 = 74^*$$

$$n_4 = 300 - 65 - 34 - 74 = 127$$

* En el estudio concreto se usaron estos valores.

Se hace el sorteo sobre los listados de cada zona, se llevan a cabo las visitas y se obtiene la tabla 6.9

Tabla 6.9

Zona 1. Tractores:

1	0	0	0	1	0	0	1	1	3	1	0	0	0	1
PP	3	1	5	PP	1	1	2	0	0	1	0	0	0	0
0	0	7	1	3	1	1	1	1	2	0	3	1	1	2
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2	1	0	0	PP										

Zona 1. Camiones:

0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
PP	1	0	2	PP	0	0	0	0	0	0	0	0	0	0
0	0	3	0	0	0	0	1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	PP										

En donde PP significa propiedad privada. El resto de los datos aparecen resumidos en la tabla siguiente:

Zona	*2*	*3*	*4*
No. de tractores	8	100	4
No. de camiones	4	55	0
No. de predios privados	1	7	3

Como en las zonas aparecieron algunos predios que son de propiedad privada, y éstos no deben formar parte del estudio, éste debe hacerse con el concepto de dominios de estudio. Para el caso de tractores de acuerdo a la ecuación 6.33 tenemos:

$$\hat{Y}_d = \frac{\sum_{h}^{4} \sum_{i}^{n_{hd}} y_{hdi}}{\sum_{h} n_{hd}} = \frac{51 + 8 + 100 + 4}{62 + 33 + 67 + 124}$$

$$= \frac{163}{286} = 0.57$$

Tractores por predio ejidal tenga o no tenga el vehículo.

Para el caso de camiones:

$$\hat{\bar{Y}}_d = \frac{13 + 4 + 55 + 0}{62 + 33 + 67 + 124} = \frac{72}{286} =$$

$$= 0.25 \quad \begin{matrix} \text{Camiones por predio ejidal} \\ \text{tenga o no el vehículo.} \end{matrix}$$

Para efectos del cálculo de la variancia, debe ser usada la expresión 6.34, su cálculo lo dejamos al lector. En este ejercicio hemos empleado las expresiones más simplificadas de los estimadores que son las correspondientes a la afijación proporcional, aunque formalmente por efecto del redondeo al calcular los tamaños de muestra no lo sea, pero su efecto es despreciable.

6.12 EJERCICIOS

6.10 Obtenga las expresiones 6.20, para calcular el tamaño de la muestra en el caso de totales.

6.11 Obtenga las expresiones 6.21, para calcular el tamaño de la muestra en el caso de porcentajes.

6.12 Obtenga la expresión 6.27, de la variancia del estimador de razón combinado en muestreo estratificado.

6.13 Usando una variable auxiliar que tome como valores uno o cero según que la unidad se encuentre o no en el dominio d-ésimo, muestre que la ecuación 6.31 es un cociente de medias estratificadas, y que, por lo tanto, es de la forma del estimador de razón combinado en la ecuación 6.26.

6.14 Suponga que en el ejemplo 6.8 hay 7 estados y que el número de gaveteros y datos de la muestra son los siguientes: (tabla 6.10).

Donde

$$Q_h = \sum_{i=1}^{n_h} ((y_{hi} - \hat{R}x_{hi}) - (\bar{y}_h - \hat{R}\bar{x}_h))^2$$

Usando la media estratificada estime el número total de tarjetas asociadas a los hijos y calcule una estimación del error estándar.

6.15 En el ejercicio anterior, 6.14, y usando el estimador de razón combinado, estime el número total de tarjetas asociadas a los hijos y calcule una estimación del error estándar. El peso total de las tarjetas en todos los

gaveteros es de 725 kilogramos. Considerando a los estimadores usados en el ejercicio 6.14 y en el actual, ¿cuál es mejor?

Tabla 6.10

Estado	N_h	n_h	$\bar{y}_h$	$\bar{x}_h$*	Q_h		$(\Sigma\, y_{hi})^2$
1	400	10	710	464	87 002	5 390 000	50 410 000
2	100	6	500	334	22 500	1 680 000	9 000 000
3	200	7	829	545	20 804	4 920 000	33 640 000
4	100	6	817	547	7 151	4 110 000	24 010 000
5	150	7	786	496	2 595	4 490 000	30 250 000
6	300	8	650	402	4 916	3 640 000	27 040 000
7	200	7	815	490	5 366	4 770 000	32 490 000

6.16 Se cuenta con tres listados de establecimientos en los cuales aparecen mezclados y sin distinción alguna tortillerías y molinos-tortillerías. Estos últimos producen tanto masa como tortillas. Se hace una estratificación por listado y se toman muestras aleatorias con los resultados de la tabla 6.11.

Tabla 6.11

	No. de establecimientos	*Establecimientos en la muestra*		*No. de empleados en la muestra*	
		Molinos	*Molinos-tortillerías*	*Molinos*	*Molinos-tortillerías*
Estrato 1	48	2	3	6	5
Estrato 2	127	3	7	9	8
Estrato 3	390	4	6	9	6

Estime el número medio de empleados por tipo de establecimiento.

6.17 En el ejercicio 6.16 calcule el número de establecimientos que deben muestrearse si se desea estimar el número medio de empleados, sin distinción de establecimiento. La estimación se desea para los tres listados en conjunto con un error no mayor del 5% y una confianza del 95%. Use $d = 0.05(1.6) = 0.08$, $s_1^2 = 0.2$; $s_2^2 = 0.233$ y $s_3^2 = 0.5$ (ejercicio 6.8; notar que en este ejercicio a cualquier tipo de establecimiento se le denomina "tortillería").

* Peso medio en gramos.

6.18 Hace 10 años se estimó el número medio de familias por manzana en las 400 manzanas de un poblado. Se pensó que este número dependía del nivel socioeconómico de cada una de ellas y, así, las manzanas fueron estratificadas en dos estratos de tamaños 60 y 340 respectivamente. Se tomó una muestra aleatoria de manzanas en cada estrato y se obtuvo como resultado:

$$\bar{y}_1 = 25 \, , \bar{y}_2 = 55$$

$$s_1^2 = 50 \, , s_2^2 = 170$$

Ahora en la actualidad, se desea repetir la encuesta usando el mismo marco muestral, es decir, las mismas manzanas en los estratos previamente definidos. Pero ahora se desea que la estimación en el estrato 1, tenga un error no mayor a 3 familias por manzana, y en el 2 no mayor a 2 familias por manzana; ambos casos a una confianza del 95%. ¿Qué tamaño de muestra por estrato es necesario? Suponga que las medias actuales son 10% mayores que las de hace 10 años, y que las variancias aumentaron en 20% .

CAPITULO 7

MUESTREO POR CONGLOMERADOS Y MUESTREO SISTEMATICO

7.1 MUESTREO POR CONGLOMERADOS, PROBABILIDADES IGUALES

En los capítulos anteriores hemos tratado los casos en los cuales se obtiene una única observación de cada una de las unidades muestrales seleccionadas. Por ejemplo, supongamos que se han seleccionado familias y que preguntamos a cada una de ellas por su número de miembros, por su ingreso total en pesos o por el número de cuartos de que está formada la vivienda en que habita; en cada una de estas situaciones la familia es vista como un todo, es decir, arroja o proporciona una única respuesta con respecto a la característica en estudio y es ella, la familia, la que actúa como unidad durante el sorteo; después de él, con los datos proporcionados por cada unidad en la muestra se procede a estimar una media, un total, una razón o un porcentaje; en esta última situación de porcentajes, la unidad muestral es la que se clasifica en la clase C o en su complemento; por ejemplo, las familias están formadas por menos de tres personas (clase C o clase de interés) o por cuatro o más de ellas (complemento).

En algunos esquemas de muestreo cada una de las unidades que son sorteadas y que se eligen para la muestra no dan o no proporcionan una única observación. Por ejemplo, en el caso de las familias, se sortean unidades familiares y a cada uno de los miembros componentes de las familias en la muestra se les pregunta

sobre el cereal de su preferencia; si en una familia existen cinco personas se tendrán cinco respuestas. Con esta información se estima el porcentaje de personas que prefieren o que son afectas a cierto tipo de cereal. En este ejemplo se han elegido familias, las cuales están compuestas o formadas por personas, y no es de interés una característica propia de la familia, sino una característica que es propia de cada una de las personas en el núcleo llamado familia. La estimación que se desea hacer no se refiere a ella, sino a sus componentes; pero se emplea (a la familia) como medio de acceso para llegar a ella (a la persona). La unidad muestral que es seleccionada (familia), está compuesta por elementos (miembros o personas) y estos últimos son los que interesan, ellos poseen la característica en estudio, sin embargo, a ellos no se les sorteó directamente, sino que resultaron elegidos en virtud de estar contenidos en la unidad más grande (familia) que resultó en la muestra.

A las unidades sorteadas formadas por elementos se les denomina *conglomerados* y, así, se habla de la observación y_{ij} perteneciente al elemento *j-ésimo* en la *unidad muestral* o *conglomerado i-ésimo.* En esta notación el primer subíndice (i) denota al conglomerado en cuestión.

El muestreo por conglomerados es un esquema en el cual se eligen conglomerados de elementos y, a cada conglomerado que resulte en la muestra, se le *revisa completamente,* es decir, se censa.

Figura 7.1 En el muestreo por conglomerados es necesario censar a cada uno de ellos que se encuentre en la muestra.

Por ejemplo, si resultaron elegidas las cajas de manzanas 7, 145 y 791, se apartan esas cajas, se abren y se procede a revisar cada una de las manzanas en ellas para así tener el y_{ij} correspondiente a cada manzana en cada una de las cajas.

Como sabemos, si se desea hacer una selección aleatoria de elementos, debemos contar con un listado de ellos, numerarlos de alguna manera y hacer la selección. En la situación anterior de las personas en familias esto equivale a tener un listado de miembros de familias, lo cual para una ciudad o para un conjunto de miles de empleados es casi imposible, muy costoso y tardado tratar de construirlo.

En buena medida es por estas dificultades que para estos casos resulta deseable el muestreo por conglomerados, ya que su aplicación no requiere de la elaboración previa de un listado de elementos sino únicamente de conglomerados de ellos; así, podemos definir como conglomerados a familias, a salones de clase, a gaveteros, a vagones, a cajas de productos terminados, etc., según sea el estudio que se desarrolle.

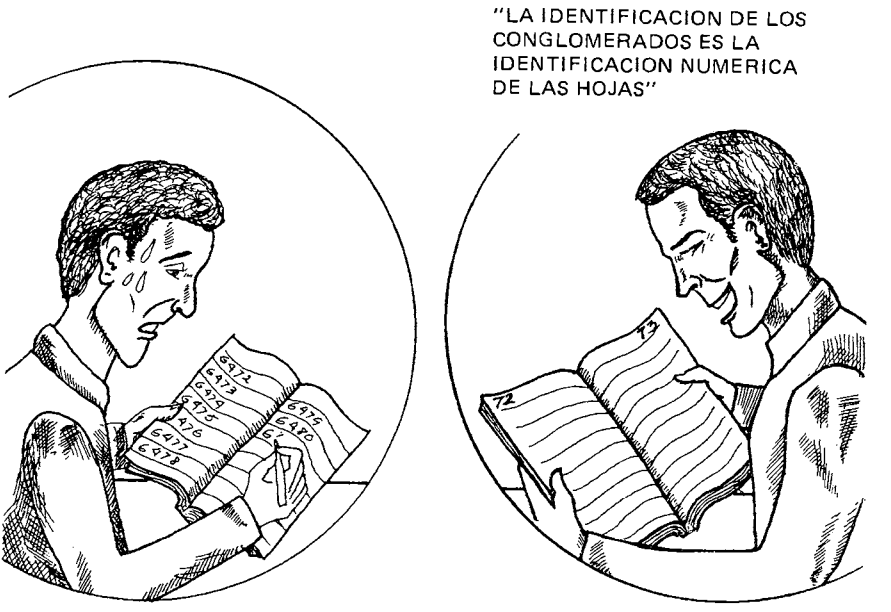


Figura 7.2 En un estudio sobre el número medio de veces que el artículo "el" aparece por renglón dentro de las páginas de un libro; resulta más fácil construir un marco de conglomerados (páginas) que construir un marco de elementos (renglones).

En una encuesta por conglomerados, por ejemplo en el caso de cajas de productos terminados, todos los conglomerados tienen el mismo número M de productos terminados o elementos debido a la uniformidad en el proceso de producción, y nos referiremos a ellos diciendo que los conglomerados son de *igual tamaño M,* o de tamaños iguales. Claramente, el caso más general es aquél en el cual los conglomerados son de *tamaños desiguales,* por ejemplo, las familias contienen un número variable de personas o elementos y los sacos de correo contienen un número variable de cartas y de sobres cada uno de ellos. Las técnicas de muestreo probabilístico cubren estas posibilidades tan frecuentes en la práctica y así, en la literatura sobre el tema se enuncian estimadores que cubren ambos casos. En el manejo práctico de ellos hay que tener presente que los conglomerados se eligen *aleatoriamente* y lo que interesa son los elementos dentro de conglomerados y todos ellos fueron o quedaron automáticamente seleccionados al elegir al conglomerado muestra y, por último, que cada uno de éstos debe ser revisado completamente.

Prácticamente, en una situación específica, la decisión sobre el tipo de esquema de muestreo que debe emplearse queda fuertemente sujeta al criterio de la persona que lo hace. En alguna situación los elementos pueden tener alguna identificación numérica consecutiva o alguna clave que sugiriera el uso de una selección aleatoria, sin embargo, por ejemplo en el caso de un archivo magnético, puede ocurrir que para poder hacer la selección sea necesario elaborar un programa de computadora electrónica que lo efectúe y de esa manera obtener los elementos en la muestra, o puede ser necesario hacer trámites de tipo administrativo para llegar a los mismos resultados, y en ocasiones estos trámites diferentes y obstáculos llegan a constituirse en verdaderos problemas que uno desea evitar. Posiblemente la solución sea cambiar de esquema de muestreo, usar algún esquema que no requiera de aquellas listas de elementos. De esta manera, por ejemplo, se eligen directamente gaveteros con expedientes que sí resultan fácilmente accesibles y el esquema de muestreo permite hacer las estimaciones a nivel de expediente. En otras situaciones aunque se cuente con un listado de productos terminados, se antoja o parece natural seleccionar a conjuntos o a conglomerados de ellos por no ser conveniente, digamos, romper demasiadas cajas o no querer hacer demasiados movimientos en el almacén o bodega.

Como se dijo anteriormente, el uso del plan de muestreo por conglomerados requiere de un censo o revisión completa en aquellas unidades seleccionadas. Mientras el conglomerado sea de un tamaño moderado o de tal naturaleza que las observaciones y_{ij} puedan obtenerse con relativa facilidad, el esquema puede aplicarse sin mayores problemas. No es difícil pensar en situaciones tales en las que el conglomerado elegido es de tamaño tal o de naturaleza tal que el tratar de introducir un censo en él resulta ser una tarea irrealizable en términos prácticos. Por ejemplo, si los conglomerados elegidos son conjuntos de viviendas de 100 manzanas y dentro de las viviendas deben ser entrevistadas las personas mayores de 30 años, se antoja como demasiado grande el conjunto que hay que censar; o si la unidad elegida como conglomerado fuera el archivero, existiendo dentro de él miles de hojas y fuera necesario calcular estimaciones a nivel de hoja; es razonable pensar que no es apropiado el esquema de muestreo por conglomerados con éstos definidos en los términos anteriores.

Cuando en una situación específica, los conglomerados por los cuales se puede optar son inmanejables mediante el presente esquema de muestreo se pueden sugerir otros planes, como son el muestreo sistemático (apartado 7.7) o el submuestreo (capítulo 8). En algunas ocasiones y bajo determinado criterio resulta factible introducir una estratificación y dentro de cada estrato trabajar con la selección adecuada. Según sea la situación particular de cada caso práctico y el criterio y los recursos (dinero, tiempo) de la persona que hace el diseño, así será el método de selección a usarse y la manera bajo la cual se harán las estimaciones a partir de los datos muestrales.

Según el nuevo esquema de muestreo de este capítulo, en el apartado siguiente se proponen dos métodos para la estimación de parámetros poblacionales, y ellos consideran que fue elegida una muestra aleatoria simple de n conglomerados; entonces los conglomerados son seleccionados con probabilidades iguales, todos tienen la misma oportunidad de ser elegidos. Posteriormente, en el apartado 7.4 se introduce otro tipo de selección y se enuncian los estimadores correspondientes sin hacer demostraciones.

7.2 ESTIMACION DE TOTALES Y DE MEDIAS

Los elementos en la población se encuentran contenidos en N conglomerados* y en ellos el i-ésimo es de tamaño Mi, tiene Mi elementos. Con esta notación el número total de elementos en toda la población resulta ser la suma de los diferentes tamaños:

$$M = \sum_{i=1}^{N} Mi,$$

de manera que el tamaño medio de los conglomerados se calcula como $M = M/N$. Mediante una selección aleatoria de n conglomerados vamos a estimar el valor medio por unidad (conglomerado) de alguna característica en estudio, usualmente nos referimos a este valor como *media por unidad*, entonces el estimador a usar es la media muestral:

$$\hat{\bar{Y}} = \bar{y} = \frac{1}{n} \sum^{n} y_i \ , \qquad 7.1.1$$

en la cual $y_i = \sum_{j=1}^{M\,i} y_{ij}$.** El estimador de la variancia resulta ser:

$$\hat{V}(\bar{y}) = \frac{1 - f}{n} \frac{\sum^{n} (y_i - \bar{y})^2}{n - 1}, \qquad 7.1.2$$

Las expresiones anteriores 7.1. (7.1.1 y 7.1.2) resultan de la aplicación directa de la teoría desarrollada en el capítulo 3. E igualmente el valor total de la característica en todas las unidades y en todos los elementos queda estimado mediante la expansión de la media muestral según N:

$$\hat{Y} = N\bar{y} = \frac{N}{n} \sum_{i=1}^{i=n} y_i \qquad 7.2.1$$

con la variancia estimada según la expresión siguiente:

* Notar que en este esquema de muestreo, N simboliza al número total de conglomerados existentes.

** Aquí, y_i simboliza al valor total en el conglomerado *i-ésimo* de la característica en estudio.

$$\hat{V}(N\bar{y}) = \frac{N^2\ (1-f)}{n}\ \frac{\sum^{n}\ (y_i - \bar{y})^2}{n-1} \qquad 7.2.2$$

$$\bar{y} = \frac{\sum^{n}\ y_i}{n}$$

Ahora, si tenemos en cuenta que el número total de elementos en la población es de M, el *estimador de la media por elemento* es el respectivo del total (7.2.1) dividido por M:

$$\hat{\bar{Y}} = \bar{\bar{y}} = \frac{\hat{Y}}{M} = \frac{1}{\bar{M}}\ \frac{1}{n}\ \sum^{n}\ y_i \qquad 7.3.1$$

$$y_i = \sum_{j=1}^{j=M\,i}\ y_{ij}$$

En 7.3.1 se han usado los símbolos $\bar{\bar{Y}}$ y $\bar{\bar{y}}$ para designar a la media poblacional por elemento y a su estimador respectivamente.* El estimador de su variancia es el siguiente:

$$\left.\begin{aligned} \hat{V}(\bar{\bar{y}}) &= \frac{1-f}{n\bar{M}^2}\ \frac{\sum^{n}\ (y_i - \bar{y})^2}{n-1} \\ \hat{V}(\bar{\bar{y}}) &= \frac{1-f}{n\bar{M}^2}\ \frac{1}{n-1}\ \left(\sum_{i=1}^{i=n}\ y_i^2 \ - \frac{1}{n}\ \left(\sum_{i=1}^{i=n}\ y_i\right)^2\right) \end{aligned}\right\} \qquad 7.3.2$$

donde: $\bar{y} = \dfrac{\sum^{i=n}\ y_i}{n}$; $\bar{M} = \dfrac{\sum^{i=N}\ M_i}{N} = \dfrac{M}{N}$

Las expresiones 7.2. y 7.3 son estimadores insesgados de los valores total y medio por elemento respectivamente. Además de estos estimadores se suele recomendar a otra pareja de ellos que vienen a ser estimadores de razón y que usualmente son más

* Esta notación de doble barra y barra simple hace la distinción entre la media por elemento $\bar{\bar{Y}}$ y la media por unidad $\bar{Y}$.

$$\bar{\bar{Y}} = \frac{1}{M}\ \sum_{i=1}^{i=N}\ Y_i \ ; \quad \bar{Y} = \frac{1}{N}\ \sum_{i=1}^{i=N}\ Y_i$$

poderosos que los insesgados. Estos estimadores usan como variable auxiliar al número de elementos de que consta cada conglomerado, y son los siguientes:

Para estimar el total poblacional:

$$\left.\begin{aligned} \hat{Y}_R &= \frac{\sum_n^n y_i}{\sum_n^n M_i} M \\ \hat{V}(\hat{Y}_R) &\doteq \frac{N^2 (1-f)}{n} \frac{\sum_n^n (y_i - \bar{\bar{y}}_R M_i)^2}{n-1} \end{aligned}\right\} \quad 7.4$$

Para estimar la media por elemento poblacional:

$$\left.\begin{aligned} \hat{\bar{\bar{Y}}}_R &= \bar{\bar{y}}_R = \frac{\hat{Y}_R}{M} = \frac{\sum_n^n y_i}{\sum_n^n M_i} \\ \hat{V}(\bar{\bar{y}}_R) &\doteq \frac{1-f}{n\bar{M}^2} \frac{\sum_n^n (y_i - \bar{\bar{y}}_R M_i)^2}{n-1} \end{aligned}\right\} \quad 7.5$$

En las expresiones 7.4 y 7.5 de las variancias, se ha colocado un signo de aproximación ($\doteq$), ya que como se dijo en el capítulo 5, estos estimadores son sesgados y para su cálculo se ha usado la parte lineal del desarrollo en serie de Taylor. En seguida aparece un ejemplo de aplicación de los estimadores enunciados por este apartado.

Ejemplo 7.1 En una compañía muy grande existen 10 000 empleados repartidos en 600 oficinas y se dispone de listados que muestran a los empleados clasificados por oficina. Se elige aleatoriamente a 20 oficinas y en cada una de ellas en la muestra se identifica a sus empleados y se les pregunta por el número de hijos menores de cuatro años que tienen. Los resultados aparecen en la tabla 7.1.

La política de la compañía es tal que no se admiten parientes cercanos como empleados, y para efectos de instalación de guarderías se desea estimar el número medio de hijos menores de cuatro años por empleado y por oficina, así como calcular una estimación

del total de niños para todos los empleados de la compañía. En este ejemplo el número medio de niños por empleado viene a ser una media por elemento, en tanto que el número medio de niños por oficina es una media por unidad, ya que la oficina desempeña el papel de conglomerado.

Tabla 7.1

Oficina	*No. de empleados*	*No. de hijos menores de 4 años*	*Oficina*	*No. de empleados*	*No. de hijos menores de 4 años*
1	15	30	11	20	30
2	18	54	12	30	30
3	12	12	13	22	42
4	15	15	14	15	30
5	10	10	15	20	40
6	20	80	16	16	24
7	15	30	17	18	45
8	16	32	18	20	40
9	18	54	19	25	25
10	18	36	20	25	75

Efectuando las sumas y sumas de cuadrados correspondientes se obtienen los resultados siguientes:

$$\sum_{i=1}^{20} M_i = 368, \sum_{i=1}^{i=20} y_i = \sum_{i=1}^{20} \sum_{j=1}^{M_i} y_{ij} = 734$$

$$\sum_{i=1}^{i=20} y_i^2 = \sum_{i=1}^{20} \left(\sum_{j=1}^{M_i} y_{ij}\right)^2 = 33\,336; \sum_{i=1}^{i=20} M_i y_i = 14\,241$$

$$\sum^{20} M_i^2 = 7\,186$$

i) Usando los estimadores insesgados: el número medio de hijos por empleado, lo podemos estimar mediante las ecuaciones 7.3:

$$\bar{\bar{y}} = \frac{600}{20\,(10\,000)} 734 = 2\,202 \text{ niños/empleado.}$$

Y despreciando f su variancia estimada es la siguiente:

$$\hat{V}(\bar{y}) = \left(\frac{600}{10\,000}\right)^2 \frac{1}{(20)19} \left[\sum^{n} y_i^2 - \frac{(\sum^{n} y_i)^2}{n}\right]$$

$$= \frac{0.0036}{380}\left(33\,336 - \frac{538\,756}{20}\right)$$

$$= 0.06 \text{ (niños/empleado)}^2,$$

el error estándar es de 0.25 niños/empleado. El número medio de hijos o niños por oficina (media por unidad) según la ecuación 7.1.1 es:

$\bar{y} = 734/20 = 36.7$ niños/oficina.

Para estimar el total de niños, cuyos padres laboran en las 600 oficinas usamos la expresión 7.2.1:

$\hat{Y} = N\bar{y} = 600(36.7) = 22\,020$ niños

y su variancia estimada queda calculada según la ecuación 7.2.2:
$(600)^2\ (6\,398.2)/((20)(19))$
o sea 6 061 452 (niños)^2, con un error estándar de 2 462 niños.

ii) Usando los estimadores de razón: según las expresiones 7.4 y 7.5, y despreciando como antes a f tenemos:

$\bar{\bar{y}}_R = 734/368 = 1.99$ niños/empleado

$$\hat{V}(\bar{\bar{y}}_R) \doteq \frac{1-f}{n\bar{M}^2} \frac{\sum^{n} y_i^2 - 2\bar{\bar{y}}_R \sum^{n} M_i y_i + \bar{\bar{y}}_R^2 \sum^{n} M_i^2}{n-1}$$

$$= \left(\frac{600}{10\,000}\right)^2 \frac{1}{20(19)} (33\,336 - 2\,(1.99)(14\,241)$$

$$+ (1.99)^2\ (7\,186)) = 0.0484 \text{ (niños/empleado)}^2,$$

y su error estándar resulta ser de 0.22 niños/empleado.

$\hat{Y}_R = 1.99\ (10\,000) = 19\,900$ niños, y su varianza es de $(0.0484)\ (100\,000\,000) = 4\,840\,000$ (niños)^2, con un error estándar de 2 200 niños.

7.3 ESTIMACION DE PORCENTAJES

La población sujeta a estudio está contenida en N conglomerados, de los cuales el *i-ésimo* es de tamaño M_i; se eligen aleatoriamente n de ellos y en base a esta muestra deseamos estimar el porcentaje de elementos que pertenecen a la clase C; por ejemplo, en cada industria (conglomerado) una parte de los empleados son hombres y otra parte son mujeres. En estas condiciones los estimadores correspondientes son los siguientes:

$$\left.\begin{aligned} \hat{P}_R = p_R &= \frac{\sum_{}^{n} a_i}{\sum_{}^{n} M_i} 100 \\ \hat{V}(p_R) &\doteq \frac{1-f}{n\bar{M}^2} \frac{\sum_{}^{n} a_i^2 - 2p_R \sum_{}^{n} a_i M_i + p_R^2 \sum_{}^{n} M_i^2}{n-1} \end{aligned}\right\} \quad 7.6\ (*)$$

En estas ecuaciones, si el término $\sum_{i=1}^{N} M_i/N$ usado para definir al tamaño medio de los conglomerados resulta desconocido, éste puede ser sustituido por su estimador:

$$\sum_{i=1}^{i=n} \frac{M_i}{n}$$

La precisión del esquema de muestreo por conglomerados depende del tamaño de ellos y de su estructura interna. Es deseable que internamente los conglomerados sean lo más *heterogéneos* posible, es decir, que haya muchos valores por arriba de la media general y otros muchos por abajo de ella. Si esto sucede, es muy posible que el *coeficiente de correlación intraconglomerado* (ejercicio 7.3) sea negativo o muy cercano a cero, y así, el esquema será más preciso que una selección aleatoria de elementos o tan preciso como ella. Sin embargo, en la práctica ocurre muchas veces que los conglomerados ya están formados, tienen una cierta estructura y no se les reconstruye a la hora de la selección, como posiblemente

* En las ecuaciones 7.6, a_i representa el número de elementos en el conglomerado *i*-ésimo que poseen la característica en estudio.

fuera deseable en un intento porque este diseño resulte más eficiente. En estas condiciones la ventaja que uno tiene en la aplicación del esquema de muestreo por conglomerados es que no requerimos de un marco de muestreo de elementos; ésta es la principal razón para su uso.

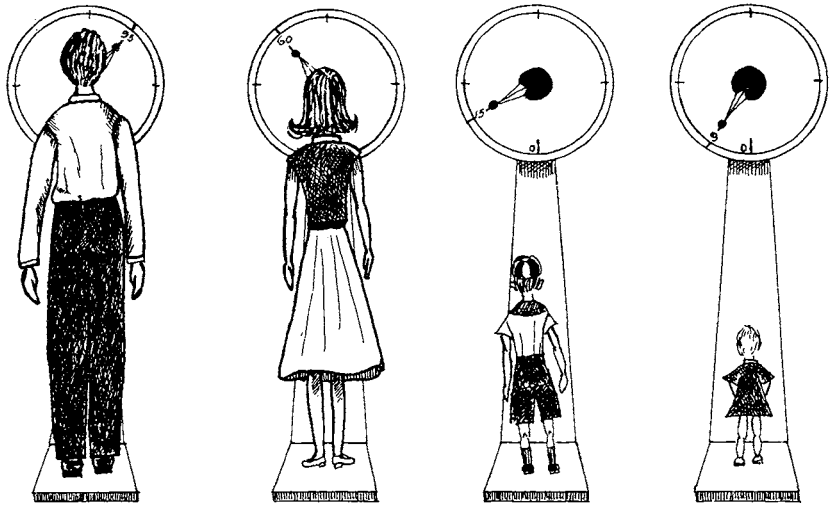


Figura 7.3 Una familia formada por niños y adultos es un conglomerado más heterogéneo que una familia formada por adultos, respecto a la característica: peso por persona en kilogramos.

7.4 SELECCION CON PROBABILIDAD PROPORCIONAL AL TAMAÑO

Cuando los conglomerados son de tamaños desiguales, es deseable asignar una mayor oportunidad de aparecer en la muestra a aquellos conglomerados que son grandes, y menor oportunidad a los que son pequeños. Al proceder de esta manera se está tomando en consideración el número de elementos que conforman a cada uno de ellos y no se les pondera a todos por igual como ocurre en los primeros apartados de este capítulo cuando se les elige de una manera aleatoria. A un tipo de selección que toma en cuenta el *"tamaño relativo"* de las unidades muestrales se le denomina *selección con probabilidad proporcional a un estimador del tamaño* (ppet). Dada la unidad muestral, por ejemplo, una oficina, algunos

estimadores de su tamaño pueden ser su número de empleados, su número de archiveros, su superficie en metros cuadrados, el número de horas que labora a la semana, etc.; en el caso de camiones con productos terminados, algunos estimadores de su tamaño pueden ser: su peso en toneladas, su número de cajas, el valor de los artículos, etc. Es claro que para un estudio específico, el estimador del tamaño de los conglomerados que se elija debe ser lo más correlacionado positivamente posible con el parámetro en estudio, ya que, por ejemplo, para referirnos a la capacidad de un salón de clase podemos hablar de su superficie, pero no tendría sentido hablar de la superficie de la mesa del maestro; para referirnos al número de cartas por saco de correo podemos hablar de su peso, en el supuesto de que éste contiene cartas únicamente.

Cuando el estimador del tamaño del conglomerado *i-ésimo* z_i coincide con el tamaño relativo según el número de elementos del conglomerado, es decir, $z_i = M_i / \sum_{i}^{N} M_i$ y que viene a ser igual a la probabilidad de selección del conglomerado *i-ésimo,* se dice que la selección es, o se efectúa con *probabilidad proporcional al tamaño del conglomerado (ppt).*

Una manera de efectuar una selección con probabilidad proporcional al tamaño de los conglomerados es la siguiente. Supongamos que hay cinco oficinas con 20, 10, 15, 25 y 20 empleados respectivamente. El total de empleados en ellas es de 90, y la probabilidad de selección que se debe asignar a la primera oficina es: $z_1 = 20/90$, a la segunda $z_2 = 10/90$, a la tercera $z_3 = 15/90$, a la cuarta $z_4 = 25/90$ y a la quinta $z_5 = 20/90$; además es cierto que $z_1 + z_2 + z_3 + z_4 + z_5 = 1$. Formemos la tabla 7.2.

Tabla 7.2

Oficina	*No. de empleados*	*No. acumulado de empleados*	*Intervalo de selección*
* 1	20	20	1 a 20
2	10	30	21 a 30
* 3	15	45	31 a 45
4	25	70	46 a 70
* 5	20	90	71 a 90

Y supongamos que debemos elegir a tres de ellas con probabilidad proporcional a su tamaño. Con ayuda de la tabla del capítulo 3 seleccionamos a un número aleatorio entre 1 y 90. Tomemos

como primer número al 16, como éste se encuentra comprendido en el intervalo de 1 a 20, resulta seleccionado el conglomerado u oficina 1. El siguiente número en la tabla es el 43, y éste se encuentra en el intervalo de 31 a 45, entonces el conglomerado 3 es elegido. El siguiente número es el 75 y de esta manera queda elegido el conglomerado 5. Supongamos que la muestra anterior debiera ser de tamaño 4, el siguiente número en la tabla es el 10, y éste se encuentra comprendido entre 1 y 20. Esto significa que el conglomerado 1 está incluido dos veces en la muestra y como cada conglomerado debe sujetarse a un censo, simplemente, en la muestra, se repetirá dos veces su observación. Cuando se muestrea con probabilidad proporcional al tamaño, o, en general, con probabilidad proporcional a un estimador del tamaño, el esquema de muestreo debe permitir el remplazo de unidades,* y así un conglomerado puede aparecer en la muestra 0, 1, 2, . . ., n veces.

7.5 ESTIMADORES A USAR CUANDO SE SELECCIONA CON ppt.

En el apartado 7.2, se enunciaron dos métodos para la estimación de medias y de totales en el caso de conglomerados que han sido seleccionados con probabilidades iguales, a saber, estimadores insesgados y estimadores de razón que resultan ser sesgados. Ahora, sin hacer las demostraciones respectivas, vamos a proponer otros estimadores insesgados para medias y para totales cuando los conglomerados se seleccionan con probabilidad proporcional a su tamaño y se permite el remplazo. Supongamos que existen N conglomerados, de los cuales el i-ésimo es de tamaño M_i y éste es seleccionado con probabilidad $z_i = M_i / \sum^{N} M_i$, proporcional a su tamaño (ppt), si la muestra debe estar formada por n conglomerados, los siguientes son estimadores insesgados de la media por elemento y del total.

Estimadores referentes a la media por elemento:

$$\hat{\bar{\bar{Y}}}_{ppt} = \bar{\bar{y}}_{ppt} = \frac{1}{n} \sum^{n} \frac{y_i}{M_i} \quad \Big| \quad \hat{V}(\bar{\bar{y}}_{ppt}) = \frac{1}{n(n-1)} \sum^{n} \left(\frac{y_i}{M_i} - \bar{\bar{y}}_{ppt}\right)^2 \qquad 7.7$$

* En este libro nos restringimos al caso de selección con remplazo, aunque existen técnicas para muestrear con ppt. que no la permiten. (Cochran, W.G. sección 9.14. 1963. Sampling Techniques. J. Wiley & Sons. N.Y. segunda edición.

Estimadores referentes al total de la característica en estudio:

$$\left.\begin{aligned} \hat{Y}_{ppt} &= M\bar{\bar{y}}_{ppt} \\ \hat{V}(\hat{Y}_{ppt}) &= M^2\,(\hat{V}(\bar{\bar{y}}_{ppt})) \end{aligned}\right\} \quad 7.8$$

7.6 EJERCICIOS

7.1 Un archivo de nombres ordenado alfabéticamente se encuentra en tarjetas sin identificación numérica y contienen cinco nombres cada una de ellas. Las tarjetas se encuentran en 100 gavetas repartidas en 10 muebles. Cada gaveta contiene entre 30 y 150 tarjetas y cada tarjeta es de medio milímetro de espesor y su peso es de medio gramo. Se desea estimar el número total de nombres en los 10 muebles *i*) ¿Qué esquema de muestreo propone usted? , indique sus razones. *ii*) ¿Cuáles son sus instrucciones para que se tome la muestra físicamente?

7.2 En referencia al ejercicio 7.1, ¿de qué signo, y de qué valor cree usted que sea la correlación entre el número de tarjetas y el número de nombres?

7.3 Si el coeficiente de correlación intraconglomerado está dado por:

$$rho = \frac{E(y_{ij} - \bar{\bar{Y}})\,(y_{ik} - \bar{\bar{Y}})}{E(y_{ij} - \bar{\bar{Y}})^2} = \frac{2\sum_i \sum_{j<k} (y_{ij} - \bar{\bar{Y}})\,(y_{ik} - \bar{\bar{Y}})}{(M-1)\,(NM-1)S^2}$$

en la cual $\bar{\bar{Y}}$ simboliza a la media general por elemento y S^2 a la variancia total entre los elementos: ¿qué se requiere para que *rho* sea: *i*) ¿positivo? , *ii*) ¿negativo? Si *rho* resulta positivo, el muestreo por conglomerados es menos preciso que una selección aleatoria de elementos equivalente según la relación siguiente:

$$V(\bar{\bar{y}}) \doteq \frac{1-f}{nM} S^2\,(1 + (M-1)\,rho)$$

7.4 En adición a la pregunta a los empleados sobre su número de hijos menores de cuatro años, en el ejemplo 7.1, se les preguntó si usarían o no los servicios de la guardería. Los resultados obtenidos fueron los de la tabla 7.3.

Tabla 7.3

Oficina	*No. de empleados*	*No. de empleados que respondieron "sí"*
1	15	3
2	18	10
3	12	10
4	15	12
5	10	10
6	20	8
7	15	10
8	16	13
9	18	11
10	18	9
11	20	15
12	30	30
13	22	17
14	15	9
15	20	13
16	16	10
17	18	14
18	20	8
19	25	12
20	25	15

Estime: *i*) el porcentaje y el número total de empleados que respondieron afirmativamente; *ii*) calcule una estimación de los errores estándar.

7.5 En el ejercicio 7.3 aparece una relación que muestra la variancia del estimador de la media muestral por elemento en términos del coeficiente de correlación intraconglomerado. ¿Cuál es el valor mínimo que puede tomar el coeficiente de correlación intraconglomerado según esta relación?

7.6 Un comerciante que se dedica a la compra de jitomate recibe un lote de 5 000 cajas de ese producto agrícola. El comerciante está seguro de que el fruto no está podrido, pero sí está consciente de que la estibación de las cajas y su transportación no fueron las adecuadas. Por lo anterior, decide elegir aleatoriamente a 30 cajas dentro del lote completo y ordena que para cada caja en la muestra se cuente el número total de jitomates y el número parcial de ellos que están reventados. Los resultados aparecen en la tabla 7.4.

Estime el porcentaje de jitomates reventados en el lote completo y obtenga intervalos de confianza del 95% para él.

Tabla 7.4

Caja	*No. total de frutos*	*No. de frutos reventados*
1	200	0
2	232	0
3	210	3
4	305	1
5	244	25
6	290	3
7	185	100
8	227	11
9	261	0
10	298	0
11	250	1
12	273	43
13	209	17
14	260	3
15	240	1
16	210	55
17	190	40
18	229	7
19	255	2
20	232	0
21	288	10
22	240	0
23	227	4
24	239	14
25	248	11
26	225	1
27	220	3
28	220	1
29	229	1
30	240	98

7.7 En una escuela existen 300 grupos diferentes de alumnos. El número de alumnos en cada uno de ellos varía entre 6 y 55. Un sociólogo desea llevar a cabo una encuesta para estimar el porcentaje de alumnos que tienen más de 10 hermanos, así como también al porcentaje de ellos que usan piloncillo en lugar de azúcar al tomar su café o té. El sociólogo está consciente de que una muestra aleatoria de grupos asignaría igual oportunidad de aparecer en la muestra a cada uno de ellos, así éstos fueran pequeños o grandes. Y dado que en su listado aparecen las claves de los grupos y el total de alumnos inscritos, decide elegir a 20 de ellos con probabilidad proporcional al número de alumnos por grupo. Para efectos

de la selección prepara su tabla, lleva a cabo ésta y al término de las entrevistas obtiene la tabla 7.5.

Tabla 7.5

Grupo	*No. de alumnos*	*No. de alumnos con más de 10 hermanos*	*No. de alumnos que usan piloncillo*
1	48	9	12
2	50	30	19
3	50	21	3
4	50	14	7
5	27	19	8
6	35	15	7
7	9	9	1
8	39	23	9
9	17	11	3
10	17	11	3
11	11	7	7
12	13	11	1
13	50	10	1
14	20	3	3
15	29	19	9
16	6	1	4
17	14	3	4
18	14	3	4
19	41	9	7
20	40	14	6

Calcule las estimaciones puntuales solicitadas y calcule estimaciones de los errores estándar.

7.7 LA MUESTRA SISTEMATICA

En la práctica es deseable contar con un esquema de muestreo tal que permita una selección que sea relativamente fácil y rápida de las unidades muestrales; esto es reforzado por el hecho de que un buen porcentaje de los errores cometidos en las encuestas se derivan de fallas cometidas durante la selección. Por ejemplo, al emprender una encuesta sobre listas de estudiantes, se les pide a los entrevistadores que en aquellos grupos de alumnos seleccionados en la muestra se elija a 5 de ellos aleatoriamente. Para ello es necesario adicionar una tabla de números aleatorios y dar las instrucciones y el entrenamiento necesario para que los entrevistadores ob-

tengan la muestra. Sin embargo, puede ocurrir que el personal no aplique o no siga las instrucciones como debiera y con ello se cambia repetidas veces el criterio de selección, ya que alguna parte del personal lo sigue o lo interpreta de una manera y otras de manera diferente.

En ocasiones, el técnico que elabora el diseño de muestreo no se llega a enterar de estos cambios introducidos durante la selección, o si se entera, ocurre fuera del tiempo oportuno y ya no puede aplicar un remedio adecuado o lo aplica parcialmente.

Supongamos que de una lista de 100 personas es necesario obtener una muestra consistente de 4 de ellas; una selección aleatoria requiere que se encuentre a 4 números aleatorios comprendidos entre 1 y 100 y después de ello hay que localizarlos en la lista. Otra manera de hacerlo y que es más simple cuando se cumplen razonablemente ciertas hipótesis (apartado 7.9), consiste en encontrar a un número aleatorio entre 1 y 25; si, por ejemplo, este número resultó ser el 18, las cuatro unidades que conformarán a la muestra son las siguientes: 18, 18 + 25, 18 + 50 y la 18 + 75. En este tipo de selección se procede a encontrar aleatoriamente a la primera unidad en la muestra para determinar al resto de ellas se avanza a brincos constantes a través de la lista, de esta manera las $n - 1$ unidades restantes quedan automáticamente seleccionadas.

A este tipo de muestra (esquema de muestreo) se le denomina *muestra sistemática* (muestreo sistemático); como cualquier otro esquema puede ser usado para toda la selección y, en lugar de muestreo aleatorio simple se puede usar muestreo sistemático, o en lugar de una selección aleatoria de conglomerados a éstos se les elige sistemáticamente, o como en el caso más evidente de una estratificación: dentro de un estrato se puede emplear una muestra sistemática, en otro una muestra aleatoria, en otro una muestra por conglomerados según el criterio del técnico y según lo requieran las condiciones particulares de cada caso en la práctica.

Principalmente, la aplicación del muestreo sistemático se torna más conveniente en aquellos casos en que la selección no puede hacerse en el gabinete, sino que por uno u otro problema debe ser desarrollada durante el trabajo de campo, lejos del personal especializado que pudiera detectar oportunamente procedimientos equivocados por una interpretación errónea de las instrucciones contenidas en los manuales o porque en muchas ocasiones se sobrestima al personal de campo, la serie de pláticas y ejercicios que debieran

conformar su entrenamiento se ve reducida a una única explicación rápida y con ella se espera falsamente que los entrevistadores se desempeñen de manera satisfactoria.

Las unidades que van a ser elegidas sistemáticamente pueden ser conglomerados o elementos, ya sea en un estrato o en una población entera. O de manera análoga, habiendo elegido a varias unidades primarias en la muestra, las cuales pudieran ser ciudades, se desea obtener una selección sistemática de manzanas dentro de ciudades para continuar con otra selección sistemática de viviendas dentro de las manzanas seleccionadas y eventualmente terminar con una selección sistemática de 1 de cada 2 adultos responsables dentro de vivienda.

Al extraer una muestra sistemática, pueden considerarse varios casos:

i) Extraerla con una fracción de muestreo preestablecida f. En esta situación el intervalo de la muestra sistemática es el inverso de f y, frecuentemente no se sabe el tamaño de muestra con el cual se va a terminar, y el objetivo es simplemente mantener fija una determinada fracción de muestreo. Consideremos por ejemplo que han sido elegidos tres salones de clase a los cuales se debe entrar con fracciones de muestreo de 1/9, 1/6.1 y 1/6. No sabemos cuántos alumnos hay en cada salón y simplemente entramos con las fracciones anteriores y al final sabemos cuántos y cuáles están en la muestra. El denominador de la fracción de muestreo puede ser o no ser entero ($f = 1/9, f = 1/6.1$).

ii) Extraer exactamente n elementos para la muestra. Aquí es necesario calcular el intervalo de la muestra sistemática $k = N/n$ el cual puede resultar entero o fraccionario.

iii) Extraer un tamaño de muestra de aproximadamente n, es decir, uno puede terminar con n, $n - 1$ ó $n + 1$, es indiferente. En esta situación se parte de n y N para calcular el intervalo de la muestra sistemática $k = N/n$, el cual, de salir fraccionario puede ser redondeado al entero inmediato superior o inferior de manera que el método de selección es simple: Obtenga el arranque el cual es un número aleatorio entre 1 y k, la muestra está dada por: $r, r + k, r + 2k, \ldots$ Por ejemplo si $N = 21$, $n = 7$, $k = 3$ y $r = 2$. Entonces la muestra es: 2, 5, 8, 11, 14, 17 y 21.

Los métodos alternativos para las situaciones cuando k es fraccionario son: intervalos fraccionarios y la muestra cíclica. En intervalos

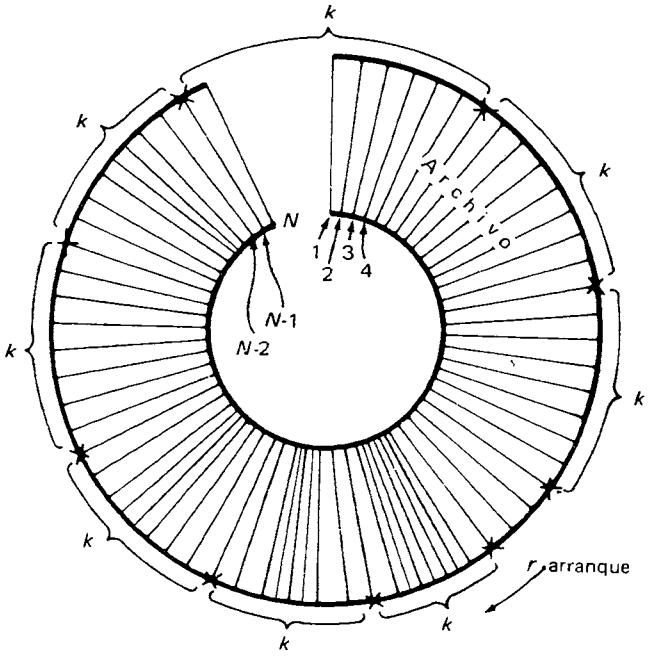


Figura 7.4 Para la aplicación de la muestra sistemática cíclica se considera que la cola del archivo se une al principio del mismo de manera que al efectuar los saltos de longitud k *y al llegar al término de la lista se continúa el conteo con el inicio del mismo hasta completar* n

fraccionarios se determina el valor de k y se cuenta el número de decimales que tiene (por ejemplo 7.02). Ahora se suprime el punto decimal y se considera a un kNUEVO el cual es el original sin el punto decimal. Se extrae a un número aleatorio entre 1 y kNUEVO al cual se le suma consecutivamente kNUEVO. Finalmente se suprimen tantas cifras a la derecha como decimales existan en el k original. Los elementos resultantes forman la muestra. Por ejemplo si $N = 70$ y $n = 8$; $k = \frac{70}{8} = 8.75$. Entonces kNUEVO = 875, tomamos $1 \leqslant r \leqslant 875$ y supongamos que es el 400. Entonces:

	400
1	275
2	150
3	025
3	900
4	775
5	650
6	525

La muestra está formada por, 4, 12, 21, 30, 39, 47, 56 y 65.

Continuando con el ejemplo anterior en el cual k resultó ser de 8.75, y usando una muestra cíclica; se elige a un número aleatorio entre 1 y N, este es el arranque, luego se redondea k al entero inmediato superior o inferior y se le suma consecutivamente al arranque hasta obtener un tamaño de muestra exactamente n. Es necesario considerar que el archivo es cíclico en el sentido de que su parte final se une a su inicio para poder continuar la cuenta como ilustra la figura 7.4. Supongamos que $1 \leqslant r \leqslant N$, y r, resulta ser 42. Considerando a $k = 8$ tenemos:

42	
50	
58	
66	
74	4
	12
	20
	28

Las situaciones descritas anteriormente aparecen en el diagrama 7.1 en el cual la primer pregunta es sobre n (parte superior del diagrama). Aquí es necesario recordar que cuando se entra con una fracción de muestreo fija, usualmente no se conoce n ~~por lo cual~~ continuaríamos a la derecha con n no fijo. El resto del diagrama es autoexplicativo.

Con estos métodos en mente, podemos continuar con el estimador de la media respectivo teniendo en mente que los métodos anteriores son tales que nos mantienen una probabilidad de selección constante e igual a $f = \frac{n}{N}$ para cada unidad o elemento en las unidades de muestreo de la etapa que se esté trabajando.*

* Notar que cuando se desea un tamaño de muestra de aproximadamente n y para facilitar o agilizar la selección, k es redondeado al entero inmediato superior o inferior, realmente al k original se le redefine moviendo ligeramente la f original, pero todas las unidades tendrán finalmente una probabilidad f redefinida de ser seleccionadas.

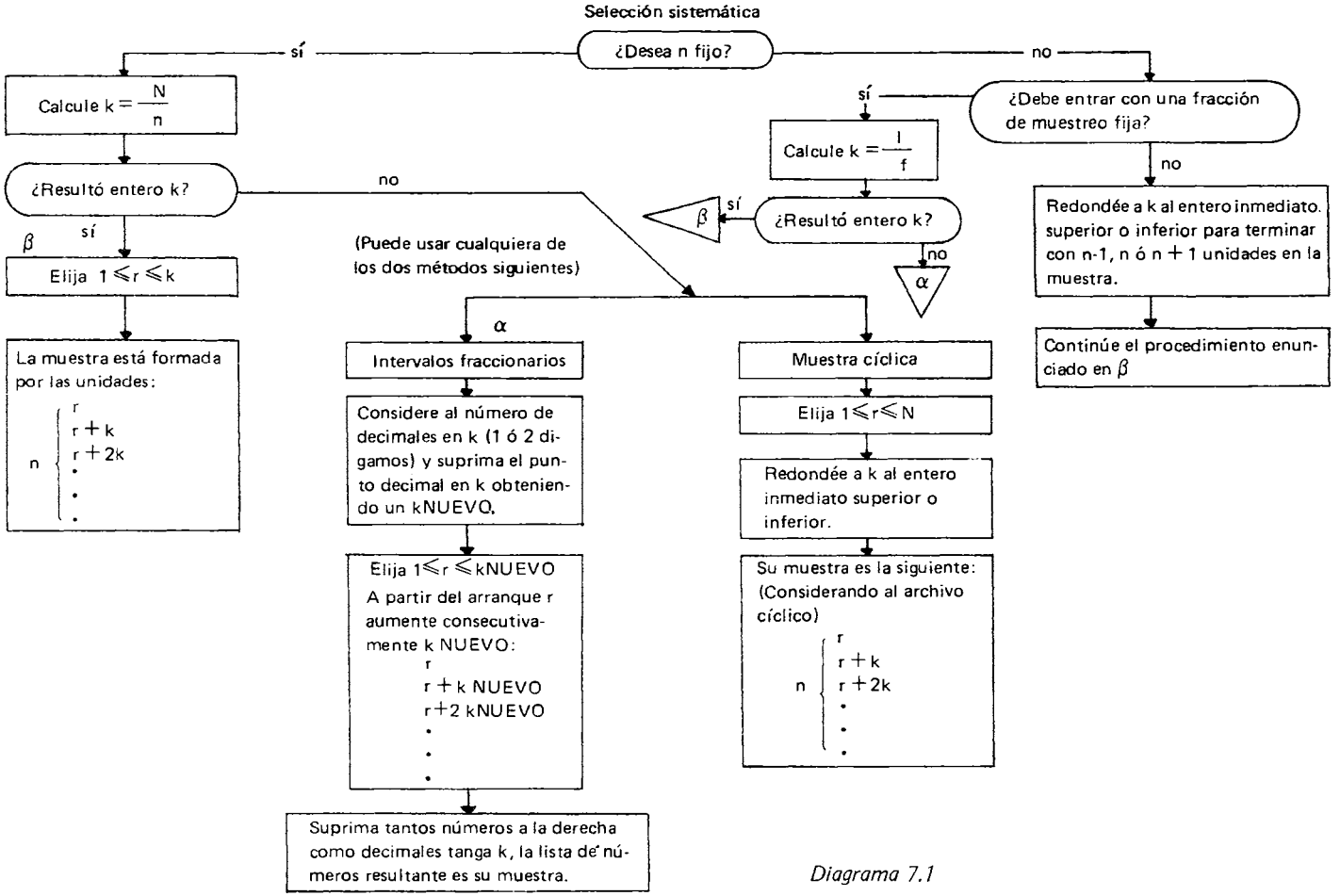

Selección sistemática
¿Desea n fijo?
sí
no
Calcule k = N/n
¿Resultó entero k?
no
β
sí
Elija 1 ≤ r ≤ k
La muestra está formada por las unidades:
n: r, r + k, r + 2k, ., ., .
(Puede usar cualquiera de los dos métodos siguientes)
α
Intervalos fraccionarios
Considere al número de decimales en k (1 ó 2 digamos) y suprima el punto decimal en k obteniendo un kNUEVO.
Elija 1 ≤ r ≤ kNUEVO
A partir del arranque r aumente consecutivamente k NUEVO:
r
r + k NUEVO
r+2 kNUEVO
.
.
.
Suprima tantos números a la derecha como decimales tanga k, la lista de números resultante es su muestra.
Muestra cíclica
Elija 1 ≤ r ≤ N
Redondée a k al entero inmediato superior o inferior.
Su muestra es la siguiente: (Considerando al archivo cíclico)
n: r, r + k, r + 2k, ., ., .
¿Debe entrar con una fracción de muestreo fija?
sí
Calcule k = l/f
¿Resultó entero k?
β
sí
no
α
no
Redondée a k al entero inmediato. superior o inferior para terminar con n-1, n ó n + 1 unidades en la muestra.
Continúe el procedimiento enunciado en β

Diagrama 7.1

En estas condiciones, un estimador insesgado de la media poblacional es la denominada *media sistemática,* la cual se define de la manera siguiente:

$$\bar{y}_{sist} = \frac{\sum_{i=n} y_i}{n}$$

En este esquema de muestreo al estimador de la media poblacional $\bar{Y}$, se le denomina media sistemática y la manera de calcularla es la misma que aquélla para la media muestral, es decir, suma de observaciones entre el tamaño de la muestra.

La media sistemática es insesgada de la media poblacional ya que el número de muestras posibles es k, y la probabilidad de que cada una de ellas sea elegida es $\frac{1}{k}$, entonces:

$E(\bar{y}_{sist})$ = Suma sobre las muestras diferentes de los productos $\frac{1}{k}\bar{y}_{sist}$

= Suma sobre las muestras diferentes de los productos $\frac{1}{k}\frac{\sum_{n} y_i}{n}$

$= \frac{1}{kn}$ (Suma sobre las muestras diferentes de $(y_1 + y_2 + \ldots + y_n)$)

$= \frac{1}{N}(y_1 + y_2 + y_3 + \ldots + y_N)$

$= \bar{Y}$ *

Y su varianza está dada por:

$$V(\bar{y}_{sist}) = \frac{1}{k}\sum_{i=1}^{i=k}(\bar{Y}_{sist(i)} - \bar{Y})^2 \qquad 7.10$$

* Ya que hay sólo k muestras posibles y cada una de ellas con elementos diferentes.

Ejemplo 7.2 Para desarrollar una encuesta de opinión sobre la localización de dos fraccionamientos industriales, se cuenta con un listado de industrias pequeñas de la transformación, ordenado por actividad económica; es decir, aparece un grupo de fábricas de hielo, luego un grupo de herrerías, otro más de fabricación de calzado y así sucesivamente. Por la localización geográfica de los fraccionamientos se supone que la opinión de cada una de las empresas es razonablemente independiente de su actividad económica.

En el listado el número total de establecimientos es de 3 800 y se desea introducir una única muestra aleatoria o sistemática sobre todo el listado, con la condición de que todos los tipos de empresas queden representados en ella. Una selección aleatoria se espera que se esparza por todo el listado, sin embargo, existe la posibilidad de que algunos tipos de empresas no aparezcan en la muestra, sobre todo de aquellos grupos pequeños.

Por lo anterior se decide emplear una muestra sistemática que tenga un intervalo de longitud igual al número de establecimientos en el grupo más pequeño de empresas.

El listado de empresas está como sigue:

Fábricas de hielo:

1. _____
2. _____
•
•
35. _____

Herrerías:

1. _____
2. _____
•
•
70. _____

Fabricación de calzado:

1. _____
2. _____
•
•
40. _____

Fabricación de muebles de madera:

1. _____
2. _____
•
•

90. _____

Fabricación de artículos para el hogar:

1. _____
2. _____
•
•
50. _____

Fabricación de ropa:

1. _____
2. _____
•
•
20. _____

Talleres automotrices:

1. _____
2. _____
•
•
190. _____
Etc.

Los grupos siguientes a talleres automotrices tienen un número de empresas que exceden a los 20 establecimientos, por lo que el intervalo de la muestra sistemática es de 20. Un número aleatorio elegido entre 1 y 20 resulta ser el 7. De manera que las empresas que conforman a la muestra son: 7, 27, 47, 67, etc., hasta terminar el listado. El tamaño de muestra resulta ser de $n = N/k = 3\,800/20$

= 190 establecimientos. Al desarrollar las entrevistas se encontró que 110 de ellas son favorables al fraccionamiento A.

En estas condiciones el porcentaje estimado de empresas que se inclinan por el fraccionamiento A es de:

$$\frac{\Sigma y_i}{n}100 = \frac{110}{190}100 = 57.89\%$$

58% aproximadamente; en la ecuación anterior y_i ha tomado los valores de *uno* o *cero*. La variancia del estimador del porcentaje lo estimamos según la expresión 3.10 (ver el apartado 7.9) y al hacer los cálculos obtenemos:

$$\frac{3\,800 - 190}{189(3\,800)}(57.89)(42.11) = 12.25$$

al extraer su raíz cuadrada se obtiene el valor de 3.5% que es su error estándar.

7.8 EQUIVALENCIAS DE LA MUESTRA SISTEMATICA

En este esquema de muestreo, sólo el arranque resulta ser aleatorio, es decir, la primera unidad elegida es seleccionada de manera aleatoria y todas las demás unidades hasta completar n no son seleccionadas por el azar, ya que una vez que se eligió el arranque éstas quedan plenamente determinadas. En estas condiciones, al hacer una comparación de ella con la selección por conglomerados, se encuentra que la selección sistemática equivale a elegir o seleccionar un conglomerado aleatoriamente, el cual está formado por n elementos: el que sirvió como arranque para la muestra (r entre 1 y k) y el resto de unidades hasta completar n. Como a partir de r, los saltos son de orden k, los diferentes conglomerados que pueden ser formados y elegidos en el supuesto de que $N = 42$ y $n = 7$, son los que aparecen en la tabla 7.6.

Debido a esta equivalencia y al hecho de que el tamaño de muestra resulta ser de 1, desde el punto de vista de conglomerados concluimos que la variancia de los estimadores no puede ser estimada formalmente,* ya que el denominador de las expresiones correspondientes se anula (ver las fórmulas 7.1 a 7.5). También, los comentarios sobre la precisión de la muestra sistemática son los

* Decimos que no puede ser estimada formalmente, porque bajo determinados supuestos podemos enunciar un estimador de ella como se verá en el apartado 7.9.

Tabla 7.6*

Conglomerados

Elementos en la muestra	1	2	·	·	·	6
1	1	2	3	4	5	6
2	7	8	9	10	11	12
·	13	14	15	16	17	18
·	19	20	21	22	23	24
·	25	26	27	28	29	30
6	31	32	33	34	35	36
$n = 7$	37	38	39	40	41	42

$$k = \frac{N}{n} = \frac{42}{7} = 6$$

mismos que aquéllos aplicables al muestreo por conglomerados: los conglomerados deben ser heterogéneos.

Ahora, intentemos comparar a la muestra sistematica con una muestra estratificada, para ello veamos la tabla 7.7.

Tabla 7.7

$k=6$

Estratos	1	2	·	·	·	6
1	1	2	3	4	5	6
2	7	8	9	10	11	12
3	13	14	15	16	17	18
4	19	20	21	22	23	24
5	25	26	27	28	29	30
6	31	32	33	34	35	36
$n = 7$	37	38	39	40	41	42

Si suponemos que existen n estratos y cada uno de ellos formado por k elementos, al elegir un número entre 1 y k (por ejemplo, el 2) da la apariencia de haberse elegido a una unidad de cada uno de los estratos; aunque en esta situación, sólo en uno de ellos se eligió la unidad aleatoriamente y en los estratos restantes quedó determinada a partir de la primera selección y esto resulta en una divergencia respecto al esquema formal de estratificación, ya que como hemos visto anteriormente, éste requiere muestras independientes de estrato a estrato.

* Debemos notar que para una población específica (N = 42) y para un tamaño dado de muestra ($n = 7$) cada unidad de la población sólo puede aparecer en una sola muestra.

7.9 SU USO

Este esquema de muestreo es ampliamente usado, principalmente en los casos de archivos de tarjetas, expedientes y hojas, así como en archivos magnéticos. Su uso generalizado se debe a su facilidad de aplicación, ya que sólo es necesario seleccionar el arranque y de ahí en adelante avanzar a brincos constantes, con ayuda, por ejemplo, de una regla o usando el folio de los documentos. En muchas ocasiones, aunque $k = N/n$ no resulte entero, uno supone que lo fue con el fin de que el arranque quede localizado al principio del archivo. Cuando k no resulta ser entero, las muestras diferentes no reciben probabilidades iguales de selección y el estimador, la media sistemática, pasa de insesgado a sesgado. Sin embargo, generalmente su efecto es despreciable y puede ser preferible suponer entero a k para evitar tomar de esta manera una muestra cíclica o intervalos fraccionarios como ya se indicó en el apartado 7.7, la cual aunque restituye el insesgamiento, son relativamente más complicados de practicar. Sin embargo, debe ser mencionado que en muchos esquemas de muestreo, la fracción de muestreo está fija, en cuyo caso, si k es fraccionario, pues, es necesario usar intervalos fraccionarios por ejemplo.

Aunque la aplicación del muestreo sistemático es muy simple, es necesario que el técnico cuente con alguna información sobre la población en la que lo va a aplicar, para así evitar sorpresas: cuando se introdujo el muestreo aleatorio simple, se dijo que la muestra debe ser aleatoria y esto se logra mediante el uso de las tablas de números aleatorios, en otras palabras, el mecanismo de aleatorización de la población son las tablas. En contraposición, en muestreo sistemático sólo una unidad es elegida aleatoriamente y las demás quedan forzadas a pertenecer a la muestra y aún en muchos casos se acostumbra fijar hasta el arranque. Así, se dice, por ejemplo: tómese una muestra sistemática con arranque al centro del intervalo. Si en estas condiciones de selección se quiere usar el supuesto de aleatorización para estimar la variancia, como en el capítulo 3, la única esperanza que uno tiene es que la población en sí se encuentre aleatorizada u "ordenada" aleatoriamente respecto a la característica buscada, como en el caso de la urna que se agita y después, por una perforación, sale una canica. En esta situación el estimador de la variancia de la media sistemática viene a ser la expresión siguiente:

$$\hat{V}(\bar{y}_{sist}) = \frac{1-f}{n} \frac{\sum^{n} (y_i - \bar{y}_{sist})^2}{n-1}$$

El supuesto de *aleatorización "natural"* en la población es el frecuentemente usado cuando se aplica este tipo de esquema. Por ello, en la práctica, antes de decidirse por una muestra sistemática se requiere analizar un poco la población para estar seguros de que tal suposición es cierta razonablemente.

Lo anterior no significa que el muestreo sistemático esté restringido a este caso; hay muchos otros en que se puede aplicar y lo hace mejor que una selección aleatoria, pero su análisis se vuelve el de una serie de casos particulares, cada uno con sus propios estimadores.*

Ejemplo 7.3 En un archivo de 10 000 tarjetas perforadas se desea hacer una estimación del porcentaje de tarjetas que tienen al menos un error de perforación. El archivo se presta para la introducción de una muestra sistemática y es pertinente la pregunta siguiente: los errores de perforación existentes en el archivo, ¿tienen alguna relación con el orden de las tarjetas? Suponiendo que la perforista u operadora trabajó normalmente, sólo existirán errores debidos a distracciones accidentales o a cansancio. Si no estuvo trabajando a marchas forzadas, el error debido a cansancio será mínimo y se puede despreciar. Así, podemos suponer aleatorización en las tarjetas e introducir la muestra sistemática.

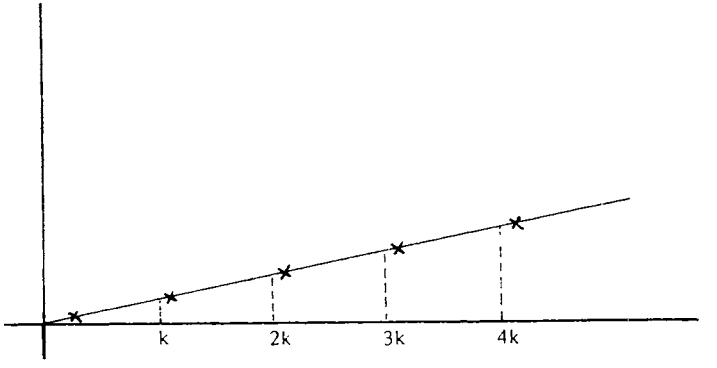


Figura 7.5.

* Sobre este tema, el lector puede consultar el libro de L. Kish.

Si el cansancio no fuera despreciable porque las 10 000 tarjetas hayan sido perforadas de manera continua, habría una tendencia en las tarjetas con error, a crecer a medida que se avanza a lo largo del archivo. Si la tendencia fuera lineal (realmente no lo es), como lo muestra la figura 7.5, sería poco afortunada una muestra sistemática al principio o al final del intervalo. Una manera de mejorarla sería tomando el arranque al centro del intervalo.

Ejemplo 7.4. (Continuación del ejemplo 6.6) Supongamos que las listas del personal en cada estado aparecen mecanografiadas en hojas tamaño carta y que el número de hojas por delegación es el registrado en la tabla 6.5. Ahí se ha supuesto que en promedio cada hoja tiene 30 nombres, debido a ello, y pensando en un esquema de muestreo por conglomerados, se requieren 47 hojas en la muestra (¿por qué?). En estas condiciones el tamaño de muestra resultante no tiene por qué coincidir con el deseado, pero se parecerá a él. Esto ocurre en la práctica y de hecho constituye una manera de diseño, es decir, se diseña para terminar con un tamaño de muestra esperado igual a algún valor en particular y realmente, al final se terminará con una cantidad mayor o menor que la deseada. Generalmente esto ocurre cuando uno fija la fracción de muestreo general para obtener estimadores autoponderados o como también se dice, diseños con igual probabilidad, aunque también existen métodos para controlar esta variación (Ver el libro de Kish. Capítulo 7).

Las 47 hojas en la muestra pueden ser seleccionadas con $f = \frac{47}{1\,410}$ o uno de cada treinta y esto puede ser hecho ya sea con muestreo aleatorio simple o mediante una selección sistemática como en el ejemplo 6.5.

Ejemplo 7.5. (Continuación del ejemplo 7.4) Como continuación del ejemplo anterior y avanzando en la complejidad del diseño, consideramos ahora un submuestreo en el cual, la unidad primaria es la hoja con los nombres y la unidad secundaria es el nombre en sí. Nuestro objetivo es terminar con $n = 1\,403$ nombres. Aunque pudiéramos proponer un esquema tal que procurara terminar con exactamente 1 403, nuestro objetivo actual es proponer esquemas autoponderados en los cuales pudiera fluctuar el tamaño de muestra final, pero que mantendría las probabilidades de selección constantes y por lo tanto la sencillez en los estimadores. En el ejemplo anterior 7.4 al seleccionar 47 hojas en la muestra podemos decir que las probabilidades de selección fueron las siguientes: $f_1 = \frac{47}{1\,410}$; $f_2 = \frac{1}{1}$.

Es decir, la fracción de muestreo general fue $f = f_1 \cdot f_2 = \frac{47}{1\,410}\,\frac{1}{1} =$ $= \frac{47}{1\,410} = \frac{1}{30}$ y requerimos un censo en las primarias seleccionadas.

Si $f_1 = \frac{94}{1\,410}$ y, $f_2 = \frac{1}{2}$ entonces, $f = \frac{94}{1\,410}\,\frac{1}{2} = \frac{1}{30}$.

Aquí estamos solicitando 94 hojas en la muestra y dentro de cada una de ellas entramos con fracción de muestreo 1 de cada 2. El número de primarias en la muestra se ha duplicado. Otra opción es usar 188 primarias en la muestra y dentro de cada una de ellas seleccionar a uno de cada cuatro nombres, es decir, $f = \frac{188}{1\,410}\,\frac{1}{4} = \frac{1}{30}$. A medida que aumentamos el número de primarias en la muestra reducimos el número de nombres en la muestra dentro de cada hoja. Si las listas de nombres siguen al orden de adscripción del personal a cada uno de los departamentos u oficinas en la empresa, lo que estamos haciendo al aumentar el número de primarias en la muestra, es aumentar al número de departamentos u oficinas en ella, aumentando, por así decirlo, la dispersión de la muestra y por lo tanto, evitando que se centralice, introduciendo tamaños de muestra relativamente pequeños en cada oficina o departamento respaldados en que usualmente ofrecen una correlación intraclase positiva (pocas observaciones nos dicen lo mismo que muchas observaciones), y el precio que se está pagando por ello, es tener que recorrer más edificios o ciudades (digamos) buscando a las distintas oficinas seleccionadas.

7.10 EJERCICIOS

7.8 En el ejercicio 7.1 sobre el archivo de nombres ordenado alfabéticamente, *i*) ¿se puede suponer un "orden" aleatorio sobre los primeros apellidos? , *ii*) ¿Por qué? , *iii*) ¿Y sobre las longitudes de los primeros apellidos?

7.9 Muestre que cuando $k = \frac{N}{n}$ no es entero, el uso de la muestra cíclica, hace que la media sistemática sea insesgada de la media poblacional.

7.10 Suponga que se cuenta con una lista de 100 alimentos que usualmente se consumen en una población y que en ella también aparecen los consumos promedios en gramos de alimento por persona. Un grupo de dietistas está interesado en estimar el total de alimento en gramos consumido por persona sin tener en cuenta el tipo de alimento.* La lista está

* El ejercicio es académico, ya que el total de alimento en gramos es la suma de las 100 cantidades, la cual puede ser obtenida sin dificultad.

ordenada según el consumo por alimento de mayor a menor. Alguien sugiere que se introduzca una muestra sistemática de intervalo 10 con arranque en el primer alimento y se estime como si fuera muestreo aleatorio simple. Se toma la muestra con los resultados siguientes:

Alimento	*Gramos por persona*
Maíz	100
Trigo	70
Azúcar	60
Frijol	40
Naranja	30
Arroz	15
Papas	12
Garbanzo	6
Pescados	2
Melón	1

Otra persona sugiere que se numeren los alimentos y que se tome una muestra aleatoria de tamaño 10; y una tercera persona indica que se censen los cinco primeros alimentos y se muestre aleatoriamente el resto. ¿Tiene razón la primera persona que propuso la muestra sistemática? Comente cada propuesta. Un censo practicado en la lista tiene como resultado 2 500 gramos.

7.11 En los hospitales de una institución de seguridad social es llevada una forma especial de sumarización, para registrar seis datos de cada persona hospitalizada. Una vez que el paciente ha salido del hospital, esta forma especial es separada de su expediente clínico y almacenada por separado en lotes correspondientes a un año. Las formas se van almacenando diariamente según su orden de llegadas y al término de 1975 se han colectado 20 000 de ellas.

Usando una muestra sistemática de intervalo 400 se extrae la muestra siguiente (ver la tabla 7.8).

Estime el periodo medio de estancia por paciente (para pacientes que ingresan y egresan el mismo día se considera un día de estancia), y encuentre intervalos de confianza del 95% para el mismo.

Tabla 7.8

Hoja No.	*Fecha de ingreso*	*Fecha de egreso*	*Tipo de servicio*	*Diagnóstico de mayor relevancia*	*Motivo del egreso*
1	10/I	10/I/74	04	1	S
2	9/I	14/I/74	07	3	S
3	20/I	21/I/74	23	7	S
4	14/I	25/I/74	12	2	T
5	28/I	28/I/74	07	3	M
6	30/I	30/I/74	14	1	S
7	27/I	1/II/74	19	3	S

Hoja No.	*Fecha de ingreso*	*Fecha de egreso*	*Tipo de servicio*	*Diagnóstico de mayor relevancia*	*Motivo del ingreso*
8	3/II	4/II	33	4	S
9	9/II	9/II	22	2	T
10	10/II	15/II	07	2	T
11	3/II	19/II	11	1	S
12	2/II	23/II	19	3	M
13	25/II	28/II	14	4	M
14	3/III	3/III	25	4	M
15	10/III	10/III	07	4	S
16	15/III	19/III	36	4	S
17	23/III	24/III	22	2	S
18	5/IV	5/IV	04	4	S
19	19/III	11/IV	07	1	S
20	12/IV	13/IV	03	1	T
21	22/IV	22/IV	14	1	T
22	30/IV	2/V	35	1	M
23	9/V	9/V	29	1	S
24	19/V	19/V	18	3	S
25	5/VI	5/VI	06	5	M
26	1/VI	13/VI	20	2	T
27	17/VI	17/VI	36	5	S
28	1/VII	1/VII	14	4	S
29	7/VII	9/VII	04	4	S
30	10/VII	14/VII	10	6	S
31	20/VII	25/VII	18	3	S
32	7/VIII	7/VIII	04	5	T
33	15/VIII	15/VIII	04	5	S
34	19/VIII	20/VIII	35	4	T
35	1/IX	1/IX	19	4	S
36	5/IX	5/IX	08	4	S
37	7/IX	11/IX	04	2	S
38	30/VIII	17/IX	07	1	S
39	5/X	8/X	30	7	S
40	18/X	18/X	19	2	S
41	27/X	27/X	13	3	S
42	3/XI	4/XI	08	1	S
43	10/XI	10/XI	24	3	S
44	15/XI	16/XI	13	4	S
45	23/XI	23/XI	04	4	S
46	2/XII	2/XII	04	5	S
47	6/XII	8/XII	04	3	M
48	15/XII	17/XII	14	4	M
49	25/XII	25/XII	01	1	M
50	27/XII	27/XII	22	4	M

7.12 Sobre el ejercicio 7.11 y de entre las personas curadas (motivo de egreso S), estime el porcentaje de ellas que fueron atendidas tanto en el servicio 04 como en el 07.

7.13. En una muestra sistemática de tamaño 200 viviendas y durante su trabajo de campo se encontró que 10 de ellas no eran en realidad viviendas, sino pequeñas industrias. Sin embargo, todas las estimaciones a efectuarse se refieren a viviendas. Para efectos de estimación de medias y de totales, ¿qué ecuaciones usaría usted? Y, ¿qué valor (es) de tamaño (s) de muestra usaría? ¿Por qué?

7.14. Una escuela tiene 20 salones en la planta baja numerados del 1 al 20 y 16 en la parte alta numerados del 1 al 16.

i) Indique brevemente, cómo numeraría o identificaría a los salones para seleccionar sistemáticamente a 5 de ellos.

ii) Utilizando los siguientes números aleatorios y avanzando de arriba hacia abajo, obtenga los cinco salones en la muestra (anote su arranque, el intervalo de la muestra, la muestra y el método usado).

Números aleatorios

74
90
25
01
41
37
25

CAPITULO 8

SUBMUESTREO

8.1 VENTAJAS Y DESVENTAJAS DE ESQUEMAS

En los capítulos 3 y 4 se desarrolló el muestreo aleatorio simple y, posteriormente, fueron comentadas varias razones por las cuales en muchas situaciones prácticas este esquema de muestreo resulta inoperante, a saber: por la necesidad de contar con un listado de unidades muestrales, el cual en muchas ocasiones es imposible de tener o de tratar de formar, por los errores a que está expuesta una selección aleatoria y que se reflejan en un falseamiento al esquema de muestreo y por la gran variabilidad que se tiene presente cuando se hace un único sorteo entre todas las unidades de la población, asignando a cada una de ellas una probabilidad igual de ser elegidas. Además, es frecuente que se soliciten estimaciones para subdivisiones de la población, lo cual contribuye a reforzar la necesidad de contar con otros esquemas de muestreo.

El muestreo estratificado permite abatir la variabilidad del estimador mediante el uso de estratos definidos adecuadamente y, de esta manera, la media estratificada puede ser más eficiente que la media muestral. Además este esquema permite hacer estimaciones separadas para cada estrato mediante el control del tamaño de muestra en cada uno de ellos y la independencia de la selección, y esto en adición a otras ventajas de tipo administrativo que proporciona durante el trabajo de campo.

El muestreo por conglomerados evita en buena medida la necesidad de contar con un listado de unidades o elementos, requiriendo únicamente un listado de conglomerados. Como usualmente éstos ya están formados de manera natural, generalmente no se espera que este esquema sea más eficiente que una selección aleatoria. Y, como hemos visto, se requiere que aquellos conglomerados que conforman a la muestra sean revisados completamente. Naturalmente la necesidad de la revisión exhaustiva va creando muchos problemas cuando los conglomerados son relativamente grandes (un gavetero, una manzana de viviendas), va haciendo el trabajo más tardado, más difícil y más costoso.

Por último, el muestreo sistemático, es un esquema cuya aplicación es muy simple, a prueba de errores; aunque, por otro lado, requiere de cuidados sobre la población sujeta a estudio, por las características que ella tiene y que influyen en el resultado final de una muestra sistemática. También pueden elaborarse diseños muestrales que no usen un único esquema, sino combinaciones de los existentes; por ejemplo, se pueden estratificar edificios y definir a cada uno de ellos como un estrato, entonces dentro de ellos elegir oficinas o muebles mediante una muestra sistemática. En cuanto a los métodos de estimación, por ejemplo en el caso de conglomerados, ya hemos visto que se dispone de varias maneras de combinar los datos de la muestra para producir o derivar una estimación. Estudiamos estimadores insesgados y estimadores de razón y vimos que cada uno de ellos requiere de determinada información para su cálculo, en algunos casos sólo se puede aplicar determinado estimador por carecer de información que requerirían otros que son más precisos. Esta situación práctica de contar con un tipo de información y no con otra es realmente importante, se dispone de alguna información y se carece de otra; por lo cual, es muy valioso el hecho de disponer de formas diversas para hacer las estimaciones. En ocasiones, sin tener en cuenta sustancialmente el que el estimador usado sea menos preciso que otro. Un criterio muy usado en las situaciones prácticas es aquel bajo el cual, se estructuran métodos de selección y de estimación que sean enteramente simples (diseños con igual probabilidad de selección para cada elemento en la población y con estimadores autoponderados, en cuyo caso la media muestral es el estimador de la media general), sacrificando así precisión estadística a favor de la eliminación o de la disminución de errores y malos entendidos tanto durante la selección como durante el procesamiento de la información.

Existe otro esquema de muestreo, el cual es sustancialmente más poderoso que los ya vistos anteriormente. Poderoso, usualmente, no en el sentido de que sea más preciso para iguales tamaños de muestra, sino en el sentido práctico: se puede introducir donde otros esquemas no tienen cabida prácticamente. Este es, el *submuestreo*, y más generalmente, el muestreo multietápico en el cual cuando sólo existen dos etapas se denomina submuestreo.

8.2. SUBMUESTREO

Consideremos una encuesta desarrollada sobre documentos los cuales están contenidos en gavetas dentro de estantes o muebles. Usualmente en cada gaveta, el número de documentos (tarjetas, expedientes, recetas, fichas, etc.) es de decenas o de cientos, y si el número de muebles es grande, por ejemplo 100, y se usara al mueble como conglomerado, habría que censar a aquellos que resultaran elegidos en la muestra (capítulo 7). en estas condiciones, el trabajo de campo sería muy embarazoso, muy largo, casi imposible. En este problema, usando el submuestreo a cada mueble que ya está en la muestra se le vuelve a muestrear; es decir, por ejemplo, las gavetas de cada mueble en la muestra se numeran y se obtiene una submuestra aleatoria de gavetas dentro de mueble, pueden ser elegidas dos gavetas dentro de cada uno de ellos y a las gavetas que así resultaran elegidas se les censa; éste es el proceso de selección en el submuestreo. A partir de los datos de la muestra proporcionados por las dos gavetas elegidas se hace una primera estimación para cada mueble en ella y posteriormente las estimaciones de los diferentes muebles se combinarán para producir o derivar una estimación global; así se desarrolla el proceso de estimación.

En este esquema y en este ejemplo, lo primero que se elige son los muebles, a ellos se les denomina unidades o conglomerados de primera etapa o *unidades primarias;* a las gavetas dentro del mueble se les elige posteriormente y se les denomina unidades o conglomerados de segunda etapa o *unidades secundarias.* Y como en el caso del capítulo anterior, todas las unidades primarias pueden tener el mismo número de unidades secundarias, es decir, ser del mismo tamaño $M_i = M$ o puede ser de tamaños diferentes M_i, este es el caso más general.

En una encuesta particular, la respuesta a la pregunta ¿cómo definir a cada unidad primaria? depende de cada situación específica y de la persona que elabora el diseño. A una persona le puede resultar conveniente definir a cada unidad primaria como cada gavetero, en tanto que a otra le parece más adecuado definirla como la gaveta. Es la misma pregunta que nos planteamos en el capítulo 7: ¿cómo definir al conglomerado? Puede haber varias respuestas.

De manera un tanto similar a las respuestas sobre el tamaño de los conglomerados, también los métodos de selección disponibles son variados. Se puede elegir a las unidades primarias con probabilidad igual o con probabilidades desiguales, con probabilidad proporcional al tamaño del conglomerado o proporcional a un estimador de él. También esto resulta válido para la selección secundaria, aunque en muchas ocasiones para estas unidades se prefiere una selección con probabilidad igual. Y, por último, también se dispone de varias maneras para calcular una estimación, esto es, existen varios métodos de estimación. Naturalmente la flexibilidad de que se dispone para definir a los conglomerados de primera etapa y los diferentes métodos de selección y de estimación permiten amoldar la teoría disponible a las diferentes condiciones que se dan en la práctica. Desde luego, desde el punto de vista de la precisión estadística, algunos estimadores resultan ser mejores que otros, sus propiedades son variadas: insesgados, sesgados, fuertemente sujetos a sesgo, menos sujetos a sesgo, consistentes, etc

8.3 NOTACION

La población está compuesta de elementos, los cuales deben ser repartidos o asignados a conglomerados y de tal manera que cada elemento aparezca en uno y sólo en uno de los N conglomerados de primera etapa y no es necesario que el número de elementos en todas y cada una de las unidades primarias sea constante. Suponemos que el conglomerado primario i-ésimo está formado por M_i elementos, en estas condiciones, el número total de ellos en la población es $\sum_{i=1}^{N} M_i$, al que nos referiremos con la letra M, $M = \sum^{N} M_i$. Entonces, tiene sentido hablar del tamaño medio de los conglomerados $\overline{M} = \frac{M}{N}$, o *tamaño medio de las unidades primarias.*

Tomaremos una muestra de n conglomerados de entre los N existentes y dentro de la *i-ésima primaria* en la muestra eligiremos una submuestra de m_i *unidades secundarias* o elementos y denotaremos al valor de la característica en estudio correspondiente al *elemento j-ésimo en la primaria i-ésima* por y_{ij}; en estas condiciones el valor del total y el valor de la media muestrales en ella son como sigue:

$$y_i = \sum^{m_i} y_{ij} \quad ; \quad \bar{y}_i = y_i/m_i$$

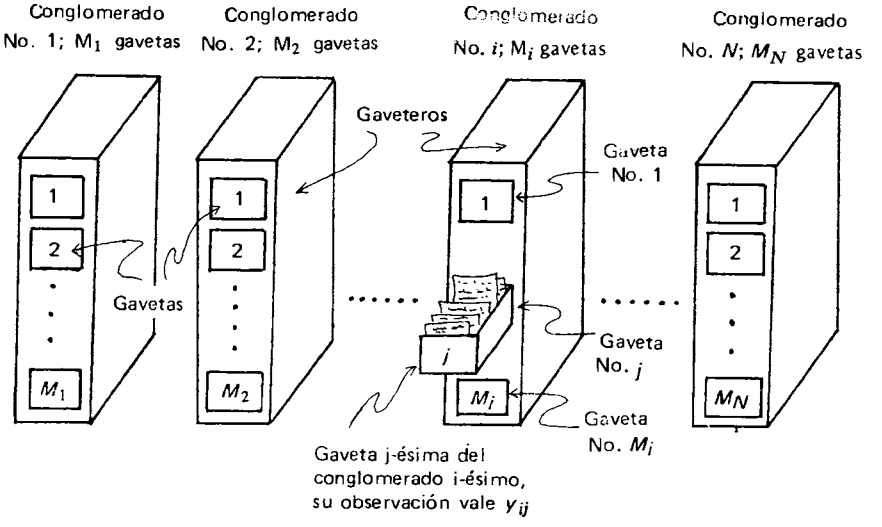


Figura 8.1. En un muestreo por conglomerados a dos etapas existen N unidades primarias o conglomerados de primera etapa (en este ejemplo, el gavetero), y cada uno de ellos está formado por un cierto número de unidades secundarias o de segunda etapa (en este ejemplo, la gaveta).

El valor de la *media por unidad primaria* o valor de la media por conglomerado en la población viene siendo el valor total de la característica en los N conglomerados y dividido entre N:

$$\bar{Y} = \frac{\sum^{N} \sum^{M_i} y_{ij}}{N} = \frac{\sum^{N} y_i}{N}$$

en tanto que el valor de la *media por elemento* en la población está dado por el valor total general dividido entre el número de elementos que existen:

$$\bar{\bar{Y}} = \frac{\sum_{N}^{N} \sum_{M_i}^{M_i} y_{ij}}{\sum_{N}^{N} M_i} = \frac{\sum_{N}^{N} y_i}{\sum_{N}^{N} M_i} = \frac{\sum^{i=N} y_i}{M} \quad *$$

y el valor del total poblacional es:

$$Y = \sum_{N}^{N} \sum_{M_i}^{M_i} y_{ij} = \sum_{N}^{N} y_i.$$

8.4 ESTIMADORES INSESGADOS DE LA MEDIA POR ELEMENTO Y DEL TOTAL

Tenemos 1, 2, 3, . . ., i, . . ., N conglomerados o unidades de primera etapa cuyos tamaños son $M_1, M_2, \ldots, M_i, \ldots, M_N$. De ellos tomamos una muestra aleatoria de tamaño n y dentro de cada conglomerado así elegido tomamos una submuestra aleatoria consistente de $m_1, m_2, \ldots, m_i, \ldots, m_n$ elementos o unidades secundarias respectivamente. En estas condiciones la expresión 8.1 constituye un estimador insesgado del parámetro poblacional media por elemento:

$$\hat{\bar{\bar{Y}}} = \bar{\bar{y}} = \frac{1}{n\bar{M}} \sum_{n}^{n} M_i \bar{y}_i \qquad 8.1$$

es decir, para estimar una media por elemento, habiendo hecho una selección aleatoria en la primera y segunda etapas del submuestreo, debemos calcular las medias muestrales de cada conglomerado en la muestra $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_i, \ldots, \bar{y}_n$; multiplicar a cada una de ellas por el tamaño del conglomerado respectivo $M_1, M_2, \ldots, M_i, \ldots, M_n$; hacer la suma de estos productos y al resultado afectarlo por el factor $\frac{1}{n\bar{M}}$, el recíproco de n veces el tamaño medio de los conglomerados. Se puede demostrar que la expresión 8.2 es un estimador insesgado de la variancia de ese estimador 8.1:

* Debemos notar que en submuestreo (ver también la nota en el apartado 7.2), los valores poblacionales pueden definirse a nivel de conglomerado o unidad primaria y a nivel de elemento o unidad secundaria. Para referirnos a la media por unidad primaria ponemos una barra ($\bar{Y}$) en tanto que para referirnos a la media por unidad secundaria o elemento ponemos dos barras ($\bar{\bar{Y}}$).

$$\hat{V}(\bar{\bar{y}}) = \frac{1 - f_1}{n\bar{M}^2} s_1^2 + \frac{1}{nN\bar{M}^2} \sum_{i}^{n} M_i^2 (1 - f_{2i}) \frac{s_{2i}^2}{m_i} \qquad 8.2$$

En esta expresión:

$$f_1 = \frac{n}{N},$$

$$f_{21} = \frac{m_1}{M_1}$$

$$f_{22} = \frac{m_2}{M_2}$$

.

.

.

$$f_{2i} = \frac{m_i}{M_i}$$

.

.

.

$$f_{2n} = \frac{m_n}{M_n}$$

y además:

$$s_1^2 = \frac{\sum_{i=1}^{i=n} (M_i\bar{y}_i - \bar{M}\bar{\bar{y}})^2}{n - 1}$$

$$s_{21}^2 = \frac{\sum_{j=1}^{j=m_1} (y_{1j} - \bar{y}_1)^2}{m_1 - 1}$$

$$s_{22}^2 = \frac{\sum_{j=1}^{j=m_2} (y_{2j} - \bar{y}_2)^2}{m_2 - 1}$$

$$s_{2\,i}^2 = \frac{\sum_{j=1}^{j=m_i} (y_{ij} - \bar{y}_i)^2}{m_i - 1}$$

$$s_{2\,n}^2 = \frac{\sum_{j=1}^{j=m_n} (y_{nj}^2 - \bar{y}_n)^2}{m_n - 1}$$

A f_1 se le denomina fracción de muestreo de las unidades primarias, en tanto que a $f_{2\,1}$, fracción de muestreo de las unidades secundarias en la primera primaria en la muestra y en general $f_{2\,i}$ es la fracción de muestreo de las unidades secundarias en la primaria i-ésima en la muestra.

s_1^2 representa el estimador de S_1^2 la variancia entre primarias, en tanto que $s_{2\,i}^2$ estima a la variancia $S_{2\,i}^2$ entre las secundarias dentro de la primaria *i-ésima.* Usualmente al primer término en la expresión 8.2 es el que más contribuye a la variancia de $\bar{y}$.

A partir de 8.1 y de 8.2 se puede formar fácilmente un estimador insesgado del valor total de la característica en estudio sobre todos los elementos, para ello sólo es necesario multiplicar al estimador de la media por elemento por el total de elementos en la población y para formar al estimador de la variancia respectivo, debemos multiplicar a la expresión 8.2 por el total de elementos elevado al cuadrado, entonces:

$$\hat{Y} = M\bar{\bar{y}} \quad , \quad \hat{V}(\hat{Y}) = M^2\,\hat{V}(\bar{\bar{y}}) \qquad 8.3$$

Ejemplo 8.1 En un estado de la república, los 9 000 comercios registrados en un organismo público se encuentran concentrados en 20 zonas diferentes de la capital y en 10 poblaciones cercanas a ella. A nivel estatal se desea estimar el número medio de empleados

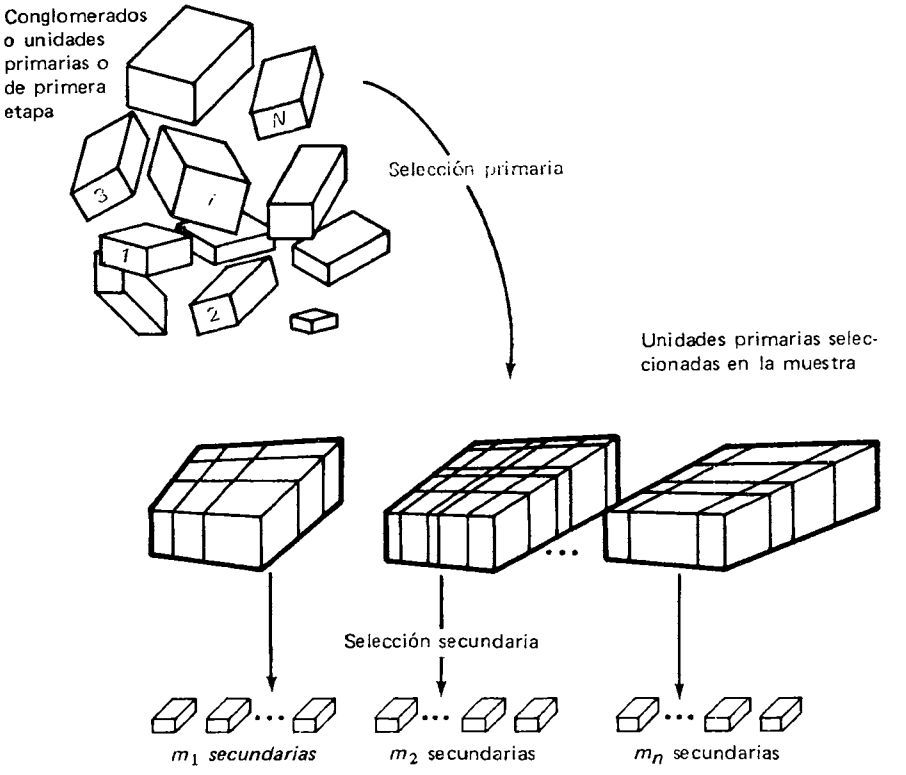


Figura 8.2

por comercio y para tal fin se considera un esquema de muestreo a dos etapas; definiendo como unidad primaria a cada una de las zonas que conglomeran a los comercios, así como a cada una de las 10 poblaciones aledañas, en estas condiciones el número total de conglomerados primarios es $N = 30$. Como unidades secundarias se considera a los comercios dentro de cada primaria. Se dispone de listados de comercios por unidad primaria y se elige a seis de ellas aleatoriamente, es decir, se eligen seis números aleatorios diferentes entre 1 y 30 y aquellas unidades asociadas a estos números elegidos quedan en la muestra. Usando los listados de comercios por unidad primaria y para aquellos en la muestra, se elige de manera aleatoria aproximadamente al 5% de ellos en cada listado. Después de hacer las selecciones primaria y secundarias se desarrolla el trabajo de campo y se obtiene la información que muestra la tabla 8.1.

Tabla 8.1

Primaria	*1*	*2*	*3*	*4*	*5*	*6*
No. de comercios	400	200	650	300	100	350
No. de comercios submuestreados	20	10	33	15	5	18
$y_i = \sum^{m_i} y_{ij}$	480	90	785	114	23	137
$M_i\bar{y}_i$	9600	1800	15462	2280	460	2663
f_{2i}	0.05	0.05	0.051	0.05	0.05	0.051
$M_i^2(1 - f_{2i})$	152000	38000	400952	85500	9500	116252
s_{2i}^2	512	36	1329	62	3	77
$(\bar{y}_i - \bar{\bar{y}}_R)^2$	61.93	50.83	58.52	72.76	132.94	72.59

Calculemos $\bar{\bar{y}}$ según la expresión 8.1:

$$\bar{\bar{y}} = \left(\frac{1}{n\bar{M}}\right) \sum^{n} M_i\bar{y}_i = \frac{30}{6(9\,000)} (32\,265)$$

$$= 17.92 \text{ empleados/comercio}$$

Esta es la estimación del número medio de empleados por comercio (media por elemento), y la variancia del estimador se calcula como sigue (ver la ecuación 8.2):

$$s_1^2 = \frac{\sum_{i=1}^{i=6} (M_i\bar{y}_i - \bar{M}\bar{\bar{y}})^2}{n-1} = 34\,693\,997$$

$$\sum_{i=1}^{i=6} M_i^2 (1 - f_{2i}) \frac{s_{2i}^2}{m_i} = 21\ 031\ 830$$

Entonces:

$$\hat{V}(\bar{y}) = \frac{1 - \frac{6}{30}}{6(300)^2} (34\ 693\ 997) + \frac{21\ 031\ 830}{6(30)(300)^2}$$

= 51.39 + 1.29 = 52.68 (empleados/comercio)^2 y su error estándar es de 7.26 empleados/comercio.

Para este número medio calculemos intervalos de confianza del 95%:

$L_i = 17.92 - 2(7.26) = 3.4$ empleados/comercio,

$L_s = 17.92 + 2(7.26) = 32.44$ empleados/comercio.

8.5 ESTIMADORES DE RAZON DE LA MEDIA POR ELEMENTO Y DEL TOTAL

En este apartado veremos otro método de estimación para los parámetros usuales, que se basa en estimadores de razón. El método de selección es el mismo que aquel del apartado anterior, a saber: de entre N conglomerados primarios, se elige aleatoriamente a n de ellos y de éstos se toma una submuestra también de manera aleatoria, tal que del i-ésimo en la muestra se elige a m_i unidades secundarias. Entonces, un estimador de razón de la media por elemento es el siguiente:

$$\hat{\bar{Y}}_R = \bar{y}_R = \frac{\sum_{n}^{n} M_i \bar{y}_i}{\sum_{n}^{n} M_i} \qquad 8.4$$

En palabras, calcúlese las medias muestrales $\bar{y}_1, \bar{y}_2, \bar{y}_1, \bar{y}_2, \ldots \bar{y}_i, \ldots, \bar{y}_n$ de cada conglomerado primario, multiplíquese a cada una

de ellas por su tamaño respectivo, súmesense estas cantidades y divídanse entre el número de elementos contenidos en los conglomerados en la muestra.

La ecuación 8.4 tiene como estimador de su variancia a la expresión siguiente:

$$\hat{V}(\bar{\bar{y}}_R) = \frac{1 - f_1}{n\bar{M}^2} s_1^2 + \frac{1}{nN\bar{M}^2} \sum^{n} M_i^2 (1 - f_{2i}) \frac{s_{2i}^2}{m_i} \qquad 8.5$$

Donde: $s_1^2 = \dfrac{\sum^{n} M_i^2 (\bar{y}_i - \bar{\bar{y}}_R)^2}{n - 1}$, y

$$s_{2i}^2 = \frac{\sum^{m_i} (y_{ij} - \bar{y}_i)^2}{m_i - 1}$$

Partiendo de las expresiones anteriores para la media por elemento, podemos formar de manera inmediata el estimador del valor total de la característica en estudio con sólo multiplicar a 8.4 por el número total de elementos:

$$\hat{Y}_R = \bar{\bar{y}}_R \sum^{N} M_i \quad , \quad \hat{V}(\hat{Y}_R) = (\sum^{N} M_i)^2 \hat{V}(\bar{\bar{y}}_R) \qquad 8.6$$

Si se desea estimar un porcentaje, tanto en 8.1 como en 8.4 en lugar de $\bar{y}_i$ debe aparecer el estimador respectivo del porcentaje poblacional en la primaria i-ésima, digamos p_i.

Ejemplo 8.2 Usando los datos del ejemplo 8.1 anterior podemos volver a calcular las estimaciones ahí solicitadas, pero usando los estimadores de razón propuestos en este apartado:

Según la expresión 8.4

$$\bar{\bar{y}}_R = \frac{\sum^{n} M_i \bar{y}_i}{\sum^{n} M_i} = \frac{32\,265}{2\,000} = 16.13 \text{ empleados/comercio.}$$

Ahora usando la expresión 8.5:

$$s_1^2 = \frac{\sum^{n} M_i^2 (\bar{y}_i - \bar{\bar{y}}_R)^2}{n - 1} = \frac{1}{5}(53\,481\,945)$$

$$= 10\,696\,389;$$

el segundo término de la ecuación 8.5 es igual al análogo del ejemplo 8.1, entonces:

$$\hat{V}(\bar{y}_R) = \frac{1 - \frac{6}{30}}{6(300)^2}(10\,696\,389) + 1.29 = 15.85 + 1.29$$

$$= 17.14 \text{ (empleados/comercio)}^2$$

con un error estándar de 4.14. Habiendo encontrado el error estándar podemos calcular un intervalo de confianza del 95%; los límites inferior y superior de él son como sigue:

$$L_i = 16.13 - 2(4.14) = 7.85 \text{ empleados/comercio.}$$

$$L_s = 16.13 + 2(4.14) = 24.41 \text{ empleados/comercio.}$$

Al comparar los límites superior e inferior obtenidos mediante el estimador insesgado y el de razón, observamos que este último resultó mas preciso, sus intervalos del 95% de confianza son más cerrados:

[3.4, 32.44] contra [7.85, 24.41] ,

esto ocurre con frecuencia.

Ejemplo 8.3 En una entidad federativa, cada uno de los agricultores se encuentran registrados en una de diez listas diferentes que corresponden a otras tantas agrupaciones de parcelas. Para efectos de estimar el número medio de hijos por agricultor se elige a cada una de las listas como unidad primaria y a cada agricultor dentro de ellas como unidad secundaria. Con ayuda de una tabla de números aleatorios se elige a tres números entre 1 y 10 y resultan seleccionados los conglomerados 2, 3 y 8 (ver la tabla 8.2).

De esta manera el tamaño medio de los conglomerados resulta ser de:

$$\bar{M} = \frac{14\,278}{10} \doteq 1\,428$$

Para submuestrear dentro de cada conglomerado primario, se elige como fracción de muestreo un centésimo; es decir, se elige a una

Tabla 8.2

Conglomerado primario	1	2	3	4	5	6	7	8	9	10	Suma
No. de agricultores por unidad primaria M_i	78	1000	400	2200	700	1400	500	2000	3000	3000	14278
Unidades primarias en la muestra		*	*					*			

de cada 100 unidades secundarias. Las observaciones correspondientes y otros cálculos necesarios aparecen en la tabla 8.3.

a) Usando el estimador insesgado de la expresión 8.1 tenemos:

$$\bar{\bar{y}} = \frac{1}{n\bar{M}} \sum_{i=1}^{i=n} M_i \bar{y}_i = \frac{22\,100}{3(1\,428)} = \frac{22\,100}{4\,284}$$

$$= 5.16 \text{ hijos por agricultor}$$

5.16 es el número medio estimado de hijos por agricultor. Para la estimación de su variancia, según la expresión 8.2 tenemos:

$$s_1^2 = \frac{\sum^{i=3} (M_i \bar{y}_i - \bar{M}\bar{\bar{y}})^2}{n-1} = \frac{1}{3-1}(39\,046\,672) = 19\,523\,336$$

$$\sum_{i=1}^{i=3} M_i^2 (1 - f_{2i}) \frac{s_{2i}^2}{m_i} = 2\,771\,208$$

Por lo cual $\hat{V}(\bar{\bar{y}})$ vale:

$$\hat{V}(\bar{\bar{y}}) = \frac{1 - \frac{3}{10}}{3(1\,428)^2}(19\,523\,336) + \frac{2\,771\,208}{3(10)(1\,428)^2}$$

$$= 2.23 + 0.0453 = 2.28$$

Tabla 8.3

Primaria	*1*	*2*	*3*	
No. de agricultores por unidad primaria en la muestra M_i	1 000	400	2 000	$\sum_{i=1}^{i=3} M_i = 3\,400$
Fracción de muestreo en las secundarias f_{2i}	0.01	0.01	0.01	
No. de agricultores submuestreados m_i	10	4	20	
No. de hijos por agricultor en la submuestra y_{ij}	6,2,11,8, 6,7,6,5, 8,10.	9,9,10, 4.	1,3,3,1,9,14, 7,10,4,5,7,7, 6,5,6,8,4,6,7,7.	
No. de hijos en la muestra por unidad primaria $y_i = \sum_{j=1}^{j=m_i} y_{ij}$	69	32	120	
No. medio estimado de hijos por unidad primaria en la muestra $\bar{y}_i = \frac{y_i}{m_i}$	6.9	8	6	
No. total estimado de hijos por unidad primaria en la muestra $M_i\bar{y}_i$	6 900	3 200	12 000	$\sum_{i=1}^{i=3} M_i\bar{y}_i = 22\,100$
$(1 - f_{2i})$	0.99	0.99	0.99	
$M_i^2 (1 - f_{2i})$	990 000	158 400	3 960 000	
s_{2i}^2	6.54	7.33	9.26	
$(\bar{y}_i - \bar{\bar{y}}_R)^2$	0.16	2.25	0.25	

Entonces el error estándar vale 1.51 hijos por agricultor.

Los límites inferior y superior de un intervalo de confianza del 95% para el número medio de hijos por agricultor valen:

$$L_i = 5.16 - 2(1.51) = 2.14 \text{ hijos por agricultor}$$

$$L_s = 5.16 + 2(1.51) = 8.18 \text{ hijos por agricultor.}$$

b) Ahora volvemos a calcular la estimación anterior usando el estimador sesgado de la expresión 8.4.

$$\hat{\bar{Y}}_R = \bar{\bar{y}}_R = \frac{\sum_{i=1}^{i=n} M_i \bar{y}_i}{\sum_{i=1}^{i=n} M_i} = \frac{22\,100}{3\,400} = 6.5 \text{ hijos por agricultor.}$$

La estimación de su variancia la hacemos con la expresión 8.5 para lo cual es necesario calcular s_1^2:

$$s_1^2 = \frac{\sum_{i=1}^{i=3} M_i^2 (\bar{y}_i - \bar{\bar{y}}_R)^2}{n-1}$$

$$= \frac{(1\,000)^2 (6.9 - 6.5)^2 + (400)^2 (8 - 6.5)^2 + (2\,000)^2 (6 - 6.5)^2}{3 - 1}$$

$$= \frac{1\,520\,000}{2} = 760\,000$$

Entonces, como los segundos términos de 8.2 y 8.5 son iguales, tenemos:

$$\hat{V}(\bar{\bar{y}}_R) = \frac{1 - \frac{3}{10}}{3(1\,428)^2} (760\,000) + 0.0453$$

$$= 0.087 + 0.0453 = 0.132$$

y su error estándar vale 0.364 hijos por agricultor; por lo cual los intervalos del 95% de confianza para esta media son:

$L_i = 6.5 - 2(0.364) = 5.772$ hijos por agricultor,

$L_s = 6.5 + 2(0.364) = 7.228$ hijos por agricultor.

Mediante la comparación de los errores estándar o de los intervalos de confianza calculados mediante los estimadores insesgados y de razón concluimos que este último es más preciso.

Ejemplo 8.4. Continuando con el tipo de diseños autoponderados empleados en los ejemplos 7.4 y 7.5, podemos mostrar o elaborar otras propuestas de diseños para los ejemplos 8.1 y 8.3 los cuales fueron presentados con diseños de probabilidades variables y/o con tamaños de muestra fijos. Supongamos que deseamos terminar con 100 comercios en la muestra, es decir, $f = \frac{100}{9000} = \frac{1}{90}$. Una propuesta con igual probabilidad es:

$$f = f_1 \cdot f_2 = \frac{1}{18}\ \frac{1}{5} = \frac{1}{90}$$

Es decir, entrar a la selección de primarias con fracción de muestreo 1 de cada 18, y dentro de ellas, seleccionar a personas con fracción 1 de cada 5. Otra propuesta es $f = f_1 \cdot f_2 = \frac{1}{9}\ \frac{1}{10} = \frac{1}{90}$.

En ambos casos no se asegura el terminar con 100 personas exactamente. Si deseamos 50 agricultores en la muestra y tenemos 10 primarias, las siguientes son dos maneras de seleccionarlos:

$$i)\quad f = \frac{1}{10}\ \frac{1}{28.556} = \frac{1}{285.56}$$

$$ii)\quad f = \frac{5}{10}\ \frac{1}{142.78} = \frac{1}{285.56}$$

El punto de partida empleado ha sido el de contar con el número de elementos que se desean en la muestra así como de su total en la población. Este último casi nunca se tiene con exactitud, pero basta con una aproximación. Con esos dos datos obtenemos la fracción

de muestreo general f y procedemos a repartirla entre las dos o más etapas de muestreo necesarias. Al hacerlo tenemos explícitas las fracciones de muestreo para cada una de la etapas. Si las unidades primarias seleccionadas resultan ser de las más grandes, el número de secundarias en la muestra excederá al esperado. Si resultan ser de las más chicas, ocurrirá lo contrario. En general la muestra quedará repartida según el número relativo de unidades primarias grandes y pequeñas ya que su probabilidad de selección es de $f_1 = \frac{n}{N}$. Entonces, como se deduce, el método elige unidades primarias grandes y chicas dándoles a cada una de ellas la misma oportunidad de ser elegidas lo cual no es lo más adecuado, ya que un salón de clase en Sociología puede tener 80 alumnos, mientras que en Matemáticas tener 12 y se antoja darles más importancia a 80 alumnos que a 12.

Para dar mayor oportunidad de selección a las unidades más grandes y menor a las más chicas, se les puede seleccionar con probabilidad proporcional a su tamaño relativo (7.4). De esta manera se toma en cuenta el tamaño de cada primaria, pero, las probabilidades de selección de cada una de ellas serán casi siempre distintas y necesariamente requerirán que se les compense al definir las fracciones de muestreo de las secundarias en aras de mantener $f = f_1 \cdot f_2$ como una constante.

8.6 EXTENSION A MUESTREO ESTRATIFICADO

En la práctica es muy usual emplear algún criterio de estratificación y así dividir a la población en subpoblaciones independientes al menos desde el punto de vista de la selección, esto permite además efectuar estimaciones por estrato, o por efectos de precisión es deseable contar con estimadores cuando las unidades primarias se han estratificado en L estratos. La notación es tal que el h-ésimo contiene $M_h = \sum_{i=1}^{i=N_h} M_{hi}$ elementos, el número de unidades primarias en él es N_h, y de ellas se elige a n_h. Aquí usamos la misma notación anterior con la adición de un subíndice que nos indica el estrato al que se hace referencia.

El estimador estratificado del *valor medio por elemento* correspondiente a la utilización de la expresión 8.1 dentro de cada estrato (estimador insesgado en cada estrato) es:

$$\bar{y}_{est} = \frac{\sum_{h=1}^{h=L} M_h \bar{y}_h}{\sum_{h=1}^{h=L} M_h} \qquad 8.7$$

y un estimador de su variancia es el siguiente:

$$\hat{V}(\bar{y}_{est}) = \sum_{h} \left[\frac{M_h}{\sum_{h=1}^{L} M_h} \right]^2 \hat{V}(\bar{y}_h) \qquad 8.8$$

en 8.7 y 8.8, $\bar{y}_h$ es el estimador 8.1 valuado con la muestra desarrollada en el estrato h -ésimo y $\hat{V}(y_h)$ su estimador de variancia correspondiente.

8.7 COMENTARIOS SOBRE LOS TAMAÑOS

Como se señaló en los apartados 7.3 y 8.1, en muchas ocasiones los elementos poblacionales se encuentran en conglomerados que pudiéramos denominar naturales y que el estadístico aprovecha para el desarrollo de sus encuestas. Por ejemplo, las personas se encuentran contenidas en familias, las familias en manzanas habitacionales, las manzanas en zonas o distritos y las zonas o distritos en ciudades. En una situación dada puede no ser conveniente fraccionar o dividir a una manzana en dos conglomerados indicando límites físicos de donde empieza y donde termina cada uno de ellos; es decir, usualmente cuando la identificación física se vuelve difícil, no es conveniente fraccionar o dividir a conglomerados naturales, pero sí es posible hacer combinaciones de ellos, por ejemplo, juntar a dos o más manzanas cercanas o a dos o más oficinas, etc. De esta manera el tamaño de las unidades primarias lo podemos hacer pequeño, regular o grande.

Desde el punto de vista de la precisión, en un esquema de submuestreo interesa que las unidades primarias sean altamente heterogéneas respecto a la(s) característica(s) en estudio (ejercicio 7.3). Una manera de lograr esta heterogeneidad consiste en definir a la unidad primaria de manera que resulte grande relativamente, para así usar el criterio que es válido en muchas ocasiones y que establece que a mayor lejanía, mayor diferencia o asemejanza entre unidades, familias pobres y familias ricas. Si ocurre esto, pueden

elegirse para la muestra a pocas unidades primarias. Si por el contrario, los conglomerados primarios resultan ser muy homogéneos, se hace necesario aumentar la fracción de muestreo de las primarias o, en otras palabras, aumentar el tamaño de muestra de ellas, para percatarse y tomar en consideración la alta variabilidad entre primarias. En esta situación, dentro de las unidades primarias muestrales se elegirán pocas unidades secundarias, ya que los elementos en la misma primaria tenderán a parecerse.

8.8 EJERCICIOS

8.1 En el ejemplo 8.1 sobre los comercios en un estado se definió a la unidad primaria como cada zona y como cada una de las diez poblaciones cercanas a la capital. ¿Qué comentarios puede hacer respecto a la variabilidad dentro de primarias? En un muestreo a dos etapas, en el cual interesa estimar la media por comercio a nivel estatal, y si los comercios aparecieran en listas por población sin más información que su nombre y su dirección, ¿cómo definiría usted a la unidad de primera etapa? Indique sus razones.

8.2 Si en el ejercicio 8.1 pudiera emplear muestreo estratificado, ¿cómo definiría a los estratos? Indique los pros y los contras de su definición.

8.3 Suponga que en el ejercicio 8.1 se definieron dos estratos; en uno se encuentran todos los comercios de la capital y en el otro el resto de comercios ¿Tendría alguna ventaja esta estratificación? La definición de unidades primarias se mantiene como en el ejemplo 8.1. De las zonas comerciales de la capital se elige aleatoriamente a 4 de ellas, siendo éstas las primarias 1, 3, 4 y 6 de la tabla 8.1 y las dos restantes son del otro estrato (2 y 5). Si en la capital existen 7 000 comercios, estime el número medio de empleados por comercio para cada uno de los estratos, así como a nivel estatal, e indique intervalos del 95% para su estimación.

8.4 Los estimadores 8.1 y 8.4 se vuelven autoponderados cuando $f_{2i} = f_2$ una constante; en este caso, encuentre la estructura de ellos y de sus respectivos estimadores de variancia.

8.5 Suponga que todos los conglomerados son de tamaño igual M, y que la fracción de muestreo de las secundarias es constante, $f_2 = m/M$. ¿Qué estructura adquieren 8.1 y 8.2?

8.6 Un organismo público desea llevar a cabo una encuesta de opinión sobre sus 130 000 empleados. Ellos se encuentran repartidos en 30 delegaciones regionales dispersas en los estados de la República Mexicana incluyendo a la capital del país. La nómina es elaborada de manera independiente en la capital y en cuatro delegaciones diferentes. Una vez que ésta ha sido elaborada se reparte a los diferentes estados y de esa manera se les paga a los empleados.

Cada una de las delegaciones que elaboran las nóminas regionales cubren a 7, 9, 4 y 9 delegaciones foráneas respectivamente. Quince días después de que la nómina ha sido pagada, se envía una copia de ella a las oficinas centrales en la capital.

Las estimaciones deseadas son de porcentajes definidos sobre los empleados, y se desean obtener para los empleados en la capital de la república y para todo el país, incluyendo a la capital. Suponiendo que se tiene acceso a los listados de empleados en las diferentes etapas del pago de la nómina, enuncie tres esquemas de muestreo diferentes que pudieran ser empleados para esta encuesta e identifique lo necesario en cada esquema.

8.7 Indique los estimadores que deberían ser empleados en cada uno de los esquemas de muestreo propuestos para el caso del ejercicio anterior, 8.6.

8.8 En una escuela con 30 salones se desea hacer una selección sistemática de 2 salones y dentro de cada salón, elegir pupitres con fracción de muestreo 1 de cada 20. Los salones y el número de pupitres por salón son como sigue:

Salón	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. de pupitres	20	25	20	27	26	25	48	30	25	40	21	20	27	35	38

Salón	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
No. de pupitres	30	40	43	48	40	21	21	25	27	29	35	40	42	42	45

i) Usando los números aleatorios siguientes y avanzando de arriba hacia abajo obtenga la selección deseada de salones y anótelos

39
48
30
07

ii) Dentro de los salones antes seleccionados y utilizando los siguientes números aleatorios, haga las selecciones de pupitres dentro de salón en la muestra con fracción de muestreo de 1 en 20. Anote el método usado y los pupitres que están en la muestra en cada salón.

Números aleatorios para el:

Primer salón de la muestra	*Segundo salón de la muestra*
75	09
19	48
15	07
30	16
03	

8.9 En una farmacia existen 50 muebles (a manera de libreros) de seis tableros cada uno de ellos. En cada mueble y sobre los tableros están los medicamentos que ahí se expenden. Se desea estimar el total de dinero invertido en los medicamentos y para esto se obtiene una selección sistemática de cinco muebles y de cada uno de ellos en la muestra se hace una selección sistemática de dos tableros en cada mueble después de lo cual, se determina el valor de la mercancía en cada tablero en la muestra con los resultados siguientes:

Mueble	*No. de tableros en cada mueble*	*No. de tableros en la muestra*	*Valor de la mercancía en cada tablero*
1	6	2	1000 1000
2	6	2	2000, 1000
3	6	2	1000, 2000
4	6	2	3000, 2000
5	6	2	3000, 1000

i) Estime el valor total de la mercancía en la farmacia.
ii) Encuentre intervalos del 95% para el total de la mercancía.

BIBLIOGRAFIA

Azorín Poch. 1969 *Curso de muestreo y aplicaciones.* Aguilar, España.

Babbie E. R. 1973. *Survey Research Methods.* Wadsworth Publishing Company, Inc. Belmont, California.

Chevry G. R. 1967 *Práctica de las encuestas estadísticas.* Ariel. España.

Cochran W. G. 1963 *Sampling techniques* John Wiley & Sons N. Y. Segunda edición.

Kish L. 1965 *Survey sampling.* John Wiley & Sons, N. Y.

Naylor T. H. Balintfy J. L. Burdich D. S. Chuk. 1977 *Experimentos de Simulación en Computadoras con Modelos de Sistemas Económicos* Editorial Limusa, S. A. México, D. F.

Raj Des. 1968. *Sampling theory.* McGraw-Hill. N. Y.

Sudman S. 1976. *Applied Sampling.* Academic Press. N. Y.

RESPUESTAS A LOS EJERCICIOS

1.2 *i*) 118, *ii*) 5.9, *iii*) $\frac{7}{20}$ 100 = 35%, *iv*) $\frac{18}{118}$ = 0.153, *v*) $\frac{18}{7}$ = 2.57

2.2 *i*) 0.33, *ii*) (miembros)^2, *iii*) 0.574

3.3 *a*) $\frac{73}{40}$ = 1.825, [1.391, 2.259] considerando $t = 2$

b) 10 000 (1,825) = 18 250, error estándar = 2 170

3.4 $\frac{20}{40}$ 100 = 50%, error estándar = 7.99%

3.5 $\hat{R} = \frac{220}{73}$ = 3.01, error estándar = 0.398

3.7 [42.07, 91.33]

3.8 $\hat{V}(\bar{y})$ = 0.189, [1.53, 3.27]

$\hat{V}(N\bar{y})$ = 22 106.2, [523.44, 1 118.2]

4.3 $\bar{y} = \frac{73}{24}$ = 3.04, [2.42, 3.66]

4.5 Tamaño de la muestra igual a 216

4.6 Tamaño de la muestra igual a 19 861

4.7 Tamaño de la muestra igual a 244

4.8 Tamaño de la muestra igual a 426

5.1 *a)* $\frac{1\,833}{1\,329}\,46 = 63.44$, *b)* $\frac{1\,970(383.74)}{2\,000(30)(29)} = 0.434$, entonces el error estándar vale 0.674

c.i) $L_i = 58.148$, $L_s = 64.052$

c.ii) $L_i = 62.122$, $L_s = 64.758$;

es más preciso el estimador de razón ya que su intervalo de confianza del 95% de confianza es más cerrado que aquel de la media muestral.

5.2 *i)* $N\bar{y} = 46\,200$ con un error estándar de 1 544.8;

$L_i = 43\,110.4 \qquad L_s = 49\,289.6$

ii) 40 104.2 tarjetas; error estándar igual a 981.4

$L_i = 38\,141.4 \qquad L_s = 42\,067$

5.6 Para el peso medio [1.929, 2.311]

Para el total [1 928.6, 2 311.4]

5.7 Error estándar igual a 0.001986

6.3 $n = 120$, *i)* $n_1 = \frac{240}{400}\,120 = 72$, $n_2 = 30$, $n_3 = 18$

ii) $n_1 = n_2 = n_3 = \frac{120}{3} = 40$

6.4 $n_1 = 109$, $n_2 = 58$, $n_3 = 20$; $n = 109 + 58 + 20 = 187$

6.6 Ciudad A 840, ciudad B 720, ciudad C 680.
La producción media en las tres ciudades es de 702.58 kilogramos al día.
Los intervalos de confianza del 95% son [598.5, 806.65]

6.7 396 957.7 kilogramos, [338 152.5, 455 762.9]

6.8 1.60 empleados por tortillería, [1.287, 1.913]

6.14 1 057 400 y el error estándar es de 37 567.06

6.15 1 134 549.4 y el error estándar es de 13 898.23.
En este ejercicio es mejor el estimador de razón combinado ya que su error estándar es menor, 13 898.23 vs. 37 567.06

6.16 Para los molinos $\frac{522.9}{213.3} = 2.452$,

para los molinos-tortillerías $\frac{383.6}{351.7} = 1.091$

6.17 Usando afijación proporcional se obtiene $n = 178$

6.18 $n_1 = 19, n_2 = 128$

7.4 *i*) 64.95%, 6 495. (*ii*) $ee\ (p_R) = 0.0509$; $ee(\hat{A}) = 508.58$

7.6 6.34%, $L_i = 2.168$, $L_s = 10.512$

7.7 45.78% y 23.87%

7.11 4.08, [2.50, 5.66]

7.12 34.38%

INDICE ALFABETICO

ESTA OBRA SE TERMINO DE IMPRIMIR EL DIA
27 DE ENERO DE 1987, EN LOS TALLERES DE
IMPRESIONES EDITORIALES, S. A.
LAGO CHALCO 230, COL. ANAHUAC
MEXICO, D. F.

LA EDICION CONSTA DE 1,000 EJEMPLARES
Y SOBRANTES PARA REPOSICION

518

Esta obra es un libro de texto para el curso de muestreo probabilístico que se imparte a estudiantes de licenciatura de las carreras de Ciencias Sociales, Economía, Administración, Actuaría y Ciencias de la Comunicación. En el libro no se hace énfasis en el rigor matemático, sino en las aplicaciones de las técnicas de muestreo probabilístico en las áreas sociales y económico-administrativas.

La exposición es clara y sencilla y se presentan en secuencia temas tales como conceptos de muestreo y estadística, muestreo aleatorio simple, tamaño de la muestra, muestreo estratificado, muestreo por conglomerados, muestreo sitemático y submuestreo. Además, en todos los capítulos hay ejemplos resueltos para facilitarle al estudiante el tema expuesto y al terminar cada capítulo hay ejercicios cuyas respuestas están al final del libro.