Instituto Nacional de Estadística

Oposiciones al Cuerpo de Diplomados en Estadística del Estado

BOE núm. 270, de 12 de octubre de 2020, pág. 87165

Muestreo

INE

6 de julio de 2021

Índice general

Tema 1

Conceptos: población objetivo y muestreada, unidad de observación y de muestreo, marco de muestreo. Sesgo de selección y de medición. Errores de muestreo y ajenos al muestreo: ventajas del uso del muestreo frente al uso de censos. Muestreo de probabilidad: marco de referencia, el error cuadrático medio de un estimador y estimadores insesgados.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía:

S. Lohr (2010). *Sampling: Design and Analysis*. 2nd. Duxbury Press

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

1.1 Introducción

La necesidad de información estadística parece interminable en la sociedad moderna en que vivimos. En particular, se recogen datos de forma regular para satisfacer la necesidad de información sobre conjuntos específicos de elementos, llamados poblaciones finitas. Por ejemplo, nuestro objetivo puede ser obtener información sobre los hogares en una ciudad y sus comportamientos en materia de gastos, empresas en una determinada industria y sus beneficios, las personas de un país y su situación laboral, o las granjas de una región y su producción de cereales.

Una de las formas más importantes de recogida de datos en la producción estadística oficial para satisfacer estas necesidades es una encuesta muestral, es decir, una investigación parcial de la población finita a través de una encuesta. Una encuesta muestral

cuesta menos que un censo, es más rápida y puede ser, incluso, más acurada[1] que los censos.

A lo largo del siglo XX el muestreo con encuestas[2] ha evolucionado hacia un conjunto de teorías, métodos y operaciones usadas diariamente en todo el mundo.

En muchos países, se constituye legalmente un instituto nacional de estadística con el fin de proporcionar información estadística sobre la situación del país. Las encuestas son una parte importante de esta actividad. Por ejemplo, en España, el Instituto Nacional de Estadística (INE) se rige, básicamente, por la Ley 12/1989, de 9 de mayo, de la Función Estadística Pública (LFEP) [3], que regula la actividad estadística para fines estatales, que es competencia exclusiva del Estado.

Por tanto, los INEs producen regularmente estadísticas sobre características y actividades nacionales importantes, incluyendo la demografía (distribución por edad y sexo, fertilidad, mortalidad), agricultura (distribución de las cosechas), población activa (empleo), salud y condiciones de vida, industria y comercio. Gran parte de la teoría básica de muestreo se desarrolló en oficinas de estadística.

En las universidades, el muestreo es ampliamente utilizado, especialmente en sociología e investigación de la opinión pública, y también en economía, ciencias políticas y psicología. El muestreo ha crecido mucho y es un enfoque hoy día aceptado universalmente como forma de obtener información. Todos los años se dedican muchos recursos a realizar encuestas.

Los medios de comunicación proporcionan al público resultados de encuestas nuevas o periódicas. Y es ampliamente aceptado que una muestra puede proporcionar una imagen acurada de una población más grande[4]; por ejemplo, una muestra bien seleccionada de unas miles de personas puede describir una población de varios millones. Sin embargo, reunir los datos es muy costoso. Por tanto, por razones de efectividad de los costes, es imprescindible usar los mejores métodos disponibles para diseñar las muestras y para el cálculo de estimaciones, utilizar la información auxiliar, etcétera.

[1] Empleamos el término *acurado* como traducción literal de *accurate* pensando en el indicador del error cuadrático medio, por distinción al término *preciso*, cuyo uso restringimos para referirnos solo a la varianza de estimador (véase más abajo).

[2] Traducimos *survey sampling* como *muestreo con encuestas* para hacer explícitos los dos elementos fundamentales de este modo de producir información (muestras y encuestas), que distinguirlo de otros (censos, registros administrativos, nuevas fuentes de datos digitales).

[3] https://www.boe.es/buscar/doc.php?id=BOE-A-1989-10767

[4] No obstante, desde la propuesta original a finales del s. XIX de emplear muestras pasaron alrededor de 40 años hasta que empezaron a emplearse de manera rutinaria.

1.2 Conceptos: población objetivo y muestreada, unidad de observación y de muestreo, marco de muestreo

En esta sección introduciremos la terminología básica del *muestreo con encuestas* (*survey sampling*). Una encuesta hace referencia a un conjunto finito de *elementos* o *unidades de observación* llamado *población finita*. Existe una regla de enumeración que define de forma unívoca a los elementos de la población. El objetivo de una encuesta es proporcionar información sobre la población finita en cuestión o sobre subpoblaciones de especial interés, por ejemplo, 'hombres' y 'mujeres' pueden ser dos subpoblaciones de 'todas las personas'. Estas subpoblaciones se denominan *dominios de estudio* o simplemente *dominios*.

Definición 1

Las *unidades de observación*, también denominadas *unidades de análisis*, o simplemente *elementos*, son las entidades que forman la población. Entenderemos este término como los elementos de la población sobre los cuales se realiza la medición de las características de interés y los valores obtenidos son grabados.

Con una encuesta se pretende obtener información sobre *características de la población*, *parámetros*, *agregados* o *indicadores* desconocidos. Los parámetros son funciones de los valores de las *variables de estudio* (a veces también denominadas *variables objetivo*[5]. Son medidas cuantitativas desconocidas de interés, por ejemplo, los ingresos totales, los ingresos medios, la producción total, el número de desempleados, tanto para la población completa como usualmente para dominios específicos.

Definición 2

La *población objetivo* es el conjunto de elementos sobre los cuales queremos obtener información y para la que se desean calcular estimaciones de los parámetros de interés.

En la mayoría de las encuestas, el acceso y la observación de elementos individuales de la población se establece mediante un *marco de muestreo*, un recurso que identifica todos los elementos de la población objetivo con las *unidades de muestreo*. Se trata de un instrumento que permite acceder a las unidades de análisis. Una definición más formal de marco de muestreo se puede encontrar en Marco de muestreo.

Definición 3

Las *unidades de muestreo* son las unidades que se seleccionan en la muestra. Son las entidades que forman el marco muestral.

[5] *target variables*

Nótese la diferencia entre unidad de análisis y unidad de muestreo a través de este ejemplo. En una encuesta en la que quieran analizarse menores de edad, los elementos o unidades de análisis serían las personas menores de 18 años. Ahora bien, para acceder a ellas a menudo se selecciona una muestra de un marco de viviendas, por lo que las viviendas serían las unidades de muestreo. Por último, al ser menores, es posible que la información sea proporcionada por un adulto (padre/madre/tutor legal), que sería la *unidad informante*. En muchas operaciones estadísticas, los tres tipos de unidades coinciden.

Definición 4

La *población muestreda* es el conjunto de todos los elementos de la población que pueden ser extraídos en una muestra, esto es, el conjunto de unidades que forman parte del marco (las unidades de muestreo).

A partir de la población, se selecciona una *muestra*, es decir, un subconjunto de elementos. Esto se puede llevar a cabo seleccionando unidades del marco. Una muestra será una *muestra probabilística* si se obtiene utilizando un mecanismo aleatorio como se verá en la sección 1.5.

Los elementos de la muestra son *observados*, es decir, para cada elemento de la muestra, las variables de estudio son *medidas* y sus valores son *grabados*. La medición se ajusta a un *plan de medición* bien definido, especificado en términos de instrumentos de medida, una o más operaciones de medida, el orden entre estas y las condiciones bajo las cuales se llevan a cabo.

Por último, los valores de las variables grabados se usan para calcular *estimaciones (puntuales)* de los parámetros de interés de la población finita (totales, medias, medianas, proporciones, coeficientes de regresión, etc.). También se calculan estimaciones de la precisión de las estimaciones. Las estimaciones son finalmente publicadas.

En una encuesta por muestreo, la observación (medición) se limita a un subconjunto de la población. El tipo de encuestas en las que se observa/mide toda la población se llama *censo*. A continuación se muestra mediante un ejemplo el proceso general por fases de una encuesta.

Ejemplo 1. Las encuestas de población activa se llevan a cabo en muchos países. Estas encuestas tienen como objetivo responder preguntas como: ¿Cuántas personas activas hay en el país y en cada una de sus regiones? ¿Qué proporción de éstas están desempleadas? En este caso, algunos de los conceptos clave son los siguientes. *Población:* Todas las personas del país con ciertas excepciones (como menores de 16 años, personas ingresadas en instituciones). *Dominios de interés:* Grupos por edad y sexo de la población, grupos por ocupación y regiones del país. *Variables:* Cada persona, en el momento de la encuesta, se puede clasificar en (a) perteneciente a la población activa o no, y (b) empleada o no. Por tanto, hay una variable de interés que toma el valor 'uno' si la persona

pertenece a la población activa y 'cero', en caso contrario. Para medir el desempleo, se define una segunda variable de interés que toma el valor 'uno' si una persona está desempleada, 'cero', en caso contrario. Son esenciales las definiciones precisas. Si el motivo es estimar el desempleo en un mes determinado y una persona entrevistada indica que ha trabajado una semana durante ese mes, pero está desempleada el día de la entrevista, debe haber una regla precisa que indica si esa persona está desempleada o no. *Características de interés de la población:* Número de personas activas/ocupadas/-paradas/inactivas, proporción de ocupados/parados en la población activa. *Muestra:* Se selecciona una muestra de personas de la población de la manera más eficiente teniendo en cuenta los recursos existentes. *Observaciones/mediciones:* Un entrevistador visita a cada persona incluida en la muestra, le pregunta las cuestiones incluidas en un cuestionario estandarizado y graba las respuestas. *Procesamiento de datos y estimación:* Los datos grabados son depurados, es decir, se preparan para la fase de estimación; se tienen en cuenta las reglas para la falta de respuesta; se calculan las estimaciones de las características de la población. Se calculan indicadores sobre la precisión de las estimaciones. Se publican los resultados. ■

Marco de muestreo

El *marco* o *marco de muestreo* es cualquier material o recurso usado para obtener acceso a la población finita de interés. Con la ayuda del marco debe ser posible (1) identificar y seleccionar una muestra de forma que respete un diseño muestral probabilístico (véase 1.5) y (2) establecer contacto con los elementos seleccionados (por teléfono, correo, dirección postal, etc.). La siguiente definición es de (Lessler y Kalsbeek 1992):

Definición 5

Un ***marco de muestreo*** consiste en materiales, procedimientos y recursos que identifican, distinguen y permiten el acceso a los elementos de la población objetivo. Se compone de un conjunto finito de unidades al que se aplica el diseño muestral probabilístico [o no probabilístico, en sentido más general]. Incluye también información auxiliar (medidas de tamaño, información demográfica) usadas para (1) técnicas muestrales especiales, como la estratificación o la selección muestral proporcional al tamaño o (2) técnicas de estimación especiales, como la estimación de razón o de regresión.

Los marcos muestrales son listas o procedimientos para identificar todos los elementos de la población objetivo. Algunos ejemplos de marcos son los siguientes.

Ejemplo 2. El Padrón Municipal [6] es un marco que contiene información sobre todos los vecinos de los municipios de España. Este marco contiene, para cada individuo,

[6] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177011&menu=resultados&idp=1254734710990

información sobre variables como la fecha y el lugar de nacimiento, el sexo, la nacionalidad o el domicilio habitual. Faltan algunas personas, e incluye algunas que realmente no pertenecen a él, pero es un buen marco muestral. Una característica muy interesante es que proporciona acceso directo a la población de España. A menudo se usa un muestreo estratificado a partir de este marco para las encuestas dirigidas a personas llevadas a cabo por el INE. Se puede contactar fácilmente con los elementos muestreados (individuos). ■

Ejemplo 3. El Directorio Central de Empresas (DIRCE) [7] es el marco muestral usado en el INE para las encuestas a empresas. Es un marco bastante complejo y está basado en la información de varias fuentes. Por un lado, utiliza información de registros administrativos, como el *Impuesto sobre el Valor Añadido*, el *Impuesto de Sociedades* y el *Impuesto sobre la Renta de las Personas Físicas* de la Agencia Estatal de Administración Tributaria, el *Registro de Cuentas de Cotización* y el *Registro de Trabajadores Activos en Cuenta Propia* de la Seguridad Social, los movimientos del Registro Mercantil y también información de las encuestas estructurales y coyunturales de empresas realizadas por el INE. Es necesaria la actualización continua para registrar los 'nacimientos' (nuevas empresas que inician su actividad), 'muertes' (finalización de la actividad de la empresa) y cambios en la clasificación basados en el tamaño, la actividad o su ubicación geográfica. ■

Usaremos el término *muestreo directo de elementos* para denotar la selección muestral de un marco que identifica directamente a los elementos individuales de la población de interés. Es decir, las unidades del marco son objetos del mismo tipo que aquellos que queremos medir y observar. Una selección de elementos puede tener lugar directamente del marco. De forma ideal, el conjunto de elementos identificados en el marco coincide con el conjunto de elementos en la población de interés, esto es, las unidades de muestreo coinciden con las unidades que conforman la población objetivo.

Por ejemplo, si la población de interés son los individuos residentes en España, podemos llevar a cabo un muestreo directo de elementos a partir del Padrón Municipal indicado en el ejemplo 2. Aquí, la unidad muestral coincide con el elemento, que es el individuo. (Los dos conjuntos realmente no son exactamente iguales, pero las diferencias son pequeñas). El marco del ejemplo 3 se puede usar para el muestreo directo de elementos con el objetivo de estudiar la población de empresas en España; en este caso, las unidades muestrales coinciden con los elementos, que son las empresas.

Sin embargo, en muchas ocasiones en la práctica no es posible realizar un muestreo directo de elementos. Es importante realizar la siguiente distinción en relación al concepto de marco:

i. Un marco como una lista directa o identificación directa de elementos de la población objetivo.

ii. Un marco como una lista o identificación de conjuntos (más grandes o más pequeños) de elementos de la población objetivo.

[7] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736160707&menu=ultiDatos&idp=1254735576550

En el caso (i), se puede llevar a cabo un muestreo directo de elementos. En el caso (ii), el acceso a los elementos es más indirecto, concretamente, seleccionando conjuntos de elementos y observando todos o algunos de los elementos en estos conjuntos seleccionados. En muchas situaciones, el caso (ii) es la única opción, ya que no es posible encontrar o construir (sin un coste prohibitivo) una lista de elementos. El número total de elementos en la población a menudo es desconocido en el caso (ii). Por ejemplo, pensemos en la población de hogares en un gran área metropolitana. En muchas ciudades no existe nada parecido a un registro completo de hogares. Se debe considerar otras unidades muestrales distintos de los hogares. Una forma es definir unidades muestrales como viviendas y seleccionar una muestra de este tipo de unidades. Con relativa facilidad podemos entonces conseguir acceso a los hogares en (un número reducido de) las viviendas seleccionadas.

Las características básicas que un marco muestral debe tener idealmente son (Burg y col. 2019):

- el marco debe estar disponible en formato digital;
- el marco tiene por objeto representar la población objetivo tan acuradamente como sea posible;
- el marco contiene las unidades muestrales básicas correspondientes a los elementos de la población objetivo y asigna a cada unidad muestral un identificador unívoco;
- el marco incluye variables de enlace (*linking variables*), que permiten conectar las unidades muestrales básicas con registros externos;
- el marco está enriquecido con variables auxiliares, permitiendo un mejor uso (al menos con las variables de contacto);
- Si existen unidades muestrales compuestas de unidades muestrales básicas (por ejemplo, hogares a partir de personas), el enlace entre ambos tipos de unidades está incluido en el marco.

La calidad del marco puede evaluarse considerando distintos tipos de errores (Burg y col. 2019):

- Errores de cobertura debidos a unidades muestrales faltantes, erróneamente incluidas o duplicadas. Es uno de los aspectos más importantes de la calidad del marco.
- Errores de clasificación y de dominios de las unidades muestrales (p.ej. en el código de actividad económica principal de una empresa o en el municipio de residencia de una persona).
- Errores en la información de contacto de las unidades muestrales (p.ej. direcciones postales desactualizadas).

- Errores de alineación[8] en las unidades muestrales.
- Errores de unidad en las unidades muestrales compuestas (p.ej. al componer erróneamente hogares a partir de personas).

En primer lugar, la *subcobertura* (*under-coverage*), en la que determinadas partes y elementos de la población objetivo no están integrados de modo sistemático y correcto, conduce a severos problemas en el uso del marco. Por ejemplo, personas viviendo en el extranjero o personas sin hogar a menudo no están incluidos en numerosos procedimientos administrativos de registro de la población que alimentan el marco, que, por tanto, no las contiene. El impacto de la subcobertura en la estimación mediante estimadores lineales se hace presente, sobre todo, en el sesgo de las estimaciones. Puede hacerse una distinción entre la subcobertura por diseño y la subcobertura intrínseca. La primera se produce cuando se excluyen voluntariamente por parte del estadístico determinadas unidades muestrales (por ejemplo, porque son difíciles de localizar o su contacto tiene asociado costes muy altos). La subcobertura intrínseca se produce cuando las unidades muestrales no se encuentran en el marco por otras razones no voluntarias. Por su parte, la *sobrecobertura* (*over-coverage*) tiene lugar si existen unidades muestrales duplicadas, no existentes o fuera del ámbito de la población objetivo. Se distinguen igualmente dos tipos de sobrecobertura: listados duplicados[9] (*duplicate listings*) y enumeración errónea. El primer tipo corresponde a elementos de la población objetivo que están referidos al menos dos veces mediante unidades muestrales. Los duplicados afectan especialmente a la calidad de las estimaciones por su efecto a través de las variables auxiliares. Además, incrementan el coste tanto de la recogida como del procesamiento de datos. El segundo tipo hace referencia a elementos no elegibles para la población objetivo bajo análisis. Su efecto negativo, en contraposición a la subcobertura, surge en la variabilidad (varianza) de los estimadores, al reducir el tamaño muestral (al ser descartadas durante la recogida).

En segundo lugar, los errores de clasificación y de dominios equivalen a subcobertura en un dominio y sobrecobertura en otro dominio (p.ej. una empresa con un código de actividad económica principal erróneo o una persona con un municipio de residencia erróneo). Por tanto, este tipo de errores introduce información auxiliar incorrecta.

En tercer lugar, las variables de contacto en un marco desempeñan un papel fundamental para poder recoger la información en la correspondiente fase de producción. Deben estar, por tanto, monitorizadas y comprobadas periódicamente.

Por último, la relación entre las unidades muestrales es cada vez más importante. Tradicionalmente ya en las estadísticas sociales estas relaciones han sido importantes para identificar hogares y unidades compuestas de diversa naturaleza. En las encuestas

[8] Traducimos *alignment* como alineación. Por error de alineación (*alignment error* se indica la falta de concordancia entre variables respecto de una misma unidad. Por ejemplo, entre nombres de empresas y razones sociales. Este error conlleva la aparición de más errores (como el error de unidad (*unit error*)).

[9] En español también se usa el término *unidades repetidas*.

económicas, aunque tradicionalmente las unidades legales han sido objeto de estudio, los grupos y *holdings* empresariales formados por varias unidades representan más recientemente un objetivo de análisis económico de creciente importancia. En este sentido, los errores de alineamiento y de unidad deben ser detectados y corregidos en la creación y mantenimiento de los marcos.

Todos estos errores son errores *ajenos al muestreo* (*non-sampling errors*). En la sección 1.4 se explicará la diferencia entre este tipo de errores y los errores de muestreo.

1.3 Sesgo de selección y de medición

A la hora de elegir una muestra, es importante asegurarse de no incurrir en el *sesgo de selección*, que consiste en no incluir en la muestra parte de la población objetivo, situación que ocurre con frecuencia debido a que las unidades más fáciles de seleccionar no son representativas de las más difíciles de elegir. Algunos ejemplos de sesgo de selección son los siguientes:

- Usar un procedimiento de selección de la muestra tal que, sin saberlo los investigadores, dependa de alguna característica asociada con las propiedades de interés. Por ejemplo, en una encuesta para determinar la frecuencia con la que los adolescentes hablan de una determinada enfermedad con sus padres, se tomó una muestra de conveniencia de adolescentes. Los adolescentes dispuestos a hablar con los investigadores sobre la enfermedad tenían más probabilidad de estar dispuestos a hablar con sus progenitores. Los investigadores, que solo tuvieron en cuenta las cantidades de tiempo mencionadas por los adolescentes en la muestra, probablemente sobreestimaron la cantidad de comunicación existente entre los padres y los adolescentes.
- La llamada muestra de juicio, donde el encuestador usa su propio juicio para elegir las unidades muestrales. Por ejemplo, si se quiere determinar el tiempo medio que invierte una persona en un centro comercial, el encuestador podría guiarse por su juicio para seleccionar deliberadamente aquellas personas que, por su apariencia o comportamiento, parecen gastar un tiempo promedio en el centro comercial, y así confirmar su opinión.
- La existencia de errores en la especificación de la población objetivo. Por ejemplo, en encuestas relativas a las elecciones para elegir representantes políticos, supóngase que se selecciona la muestra a partir de los votantes registrados que han votado en las anteriores elecciones, esto puede conllevar a no incluir en la muestra a grupos que, no habiendo votado en anteriores comicios, sí lo hacen en los actuales.
- La no inclusión de toda la población objetivo en el marco de muestreo (el conocido problema de la *subcobertura* que ya hemos analizado antes).
- La falta de respuesta puede distorsionar en gran medida los resultados de una encuesta. Normalmente los individuos que no responden a la encuesta tienen características diferentes de los que sí responden.

- Generar la muestra a partir de voluntarios es otro caso obvio de sesgo de selección, tal y como ocurre en las encuestas que realizan, por ejemplo, programas de televisión en directo a través de llamadas telefónicas, redes sociales, etc. Este tipo de muestras pueden ser sesgadas por muchas razones: una persona realiza la encuesta múltiples veces, los receptores de la encuesta (espectadores del programa en directo) no tienen por qué representar a toda la población objetivo, quienes participan tampoco, etc.
- El reemplazo de un encuestado es otra fuente de sesgo de selección. Por ejemplo, si un encuestador tiene asignado entrevistar a una unidad familiar y no hay nadie en casa (o se trata de una zona de difícil acceso) es posible que busque una alternativa (ir a casa de un vecino). En este caso es probable que la unidad encuestada finalmente difiera de la original en varias características, lo que puede conducir a un sesgo de selección.

Además del mencionado sesgo de selección, otro concepto a tener en cuenta a la hora de diseñar una encuesta es el *sesgo de medición*. Se trata del problema de no medir correctamente las respuestas, es decir, la respuesta recogida por el encuestador no es totalmente precisa. Este sesgo sucede cuando el instrumento de medida usado presenta alguna tendencia a diferir del valor verdadero. Como ocurre con el sesgo de selección, el sesgo de medición debe tenerse en cuenta desde el principio para minimizarlo en la etapa del diseño de la encuesta.

Algunos ejemplos donde se produce sesgo de medición son:

- Las personas encuestadas no dicen la verdad. Por ejemplo, los granjeron pueden infraestimar su cosecha con el objetivo de recibir mayores ayudas económicas.
- Las personas encuestadas no entienden la pregunta.
- Los personas encuestadas a veces dan respuestas diferentes según el entrevistador que esté realizando las preguntas. Por ejemplo, para no herir la sensibilidad del entrevistador.
- En ocasiones se dan respuestas para contentar o impresionar al entrevistador.
- El entrevistador podría influir en la precisión de la respuesta. Por ejemplo, al leer de forma incorrecta las preguntas, al registrar las respuestas de forma inadecuada o bien al condicionar las respuestas del encuestado.

Aunque lo recomendable es que no exista sesgo de medición, en ocasiones es imposible. Por ejemplo, en una estudio sobre aves en un determinado territorio, los investigadores se paraban cada cierta distancia y contaban las aves que podían ver y/o escuchar en la zona. En este caso casi siempre van a subestimar el número de aves totales de la zona.

1.4 Errores de muestreo y ajenos al muestreo: ventajas del uso del muestreo frente al uso de censos

En el proceso de estimación de un parámetro de la población a través de muestreo probabilístico (véase la sección 1.5.1) se origina un error de estimación. El error de estimación se define como la desviación de la estimación respecto al verdadero valor del parámetro desconocido que se desea estimar. El error de estimación es debido a dos tipos de errores: el *error de muestreo*, originado al observar los parámetros de interés sobre un subconjunto de la población en lugar de sobre la población entera, y otro tipo de errores, los denominados errores *ajenos al muestreo*.

El *error de muestreo* del estimador se debe exclusivamente al cálculo de la estimación a partir de los datos de un subconjunto de la población (*muestra*).

Por su parte, los errores *ajenos al muestreo* son el resto de errores que se pueden producir a lo largo de la investigación estadística: deficiencias en el marco de muestreo, falta de respuesta, errores de medida y errores de procesamiento de la información. Algunos de estos errores han sido expuestos ya en las secciones anteriores. Los errores *ajenos al muestreo* pueden producirse en cualquiera de las siguientes fases de la encuesta:

i. Fase de selección de la muestra. Esta fase consiste en obtener, siguiendo el diseño muestral seleccionado para la encuesta, una muestra de elementos a partir de la utilización de un marco muestral adecuado. Los errores de estimación asociados a esta fase son el error de muestreo y las posibles deficiencias existentes en el marco (ya estudiados en Marco de muestreo), de las cuales destaca la *subcobertura* que produce serios problemas, ya que existen elementos de la población objetivo que no están presentes en el marco, por lo que no pueden ser seleccionados.

ii. Fase de recogida de datos. Esta fase consiste en implementar el plan de recogida de datos para la muestra seleccionada. Se pueden producir errores debido a la falta de respuesta y errores de medida. La falta de respuesta se da, por ejemplo, cuando existe una negativa o incapacidad para responder por parte del informante, o bien el informante no se encuentra en su domicilio en el momento de la entrevista. Las principales fuentes de los errores de medida son el entrevistador (defectos en la labor de los entrevistadores por falta de formación, interpretación o grabación incorrecta de las respuestas dadas por el informante), el informante (respuestas incorrectas de forma intencionada o no, interpretación incorrecta de las preguntas del cuestionario), el cuestionario (diseño incorrecto o inadecuado) y el modo de la entrevista (véase p.ej. Groves 1989).

iii. Fase de procesamiento de datos. En esta fase se procesa y prepara la información recogida para la fase de estimación y análisis (fase iv). Incluye la codificación de los datos, es decir, la transcripción del cuestionario a un medio adecuado para la fase iv, el proceso de depuración de los datos, mediante la implementación de técnicas de detección y corrección de errores y *outliers*, la imputación de datos faltantes y el recontacto con los informantes para clarificar cualquier tipo de información en caso de ser necesario y no suponer un coste demasiado elevado. Los errores

que pueden surgir de esta fase incluyen por ejemplo errores de transcripción, codificación y errores en los valores imputados.

En la fase de estimación y análisis, además de realizar los cálculos de las estimaciones de los parámetros poblacionales, se obtienen medidas de precisión de estas estimaciones, como por ejemplo la estimación de la varianza del estimador o del error de muestreo, $\widehat{\mathbb{V}}(\widehat{\theta})$ o $\widehat{\sigma}(\widehat{\theta})$, respectivamente.

El error de muestreo, originado como consecuencia de la observación de una muestra, se produce cuando se realiza una encuesta por muestreo. Si se observa o mide toda la población, esto es, si se realiza un *censo*, entonces no hay error de muestreo. Sin embargo, un censo no significa automáticamente 'estimación sin error'. Los errores ajenos al muestreo se producen tanto en una encuesta por muestreo como en un censo. El valor exacto del parámetro de interés solo podrá ser obtenido en casos especiales, si se realiza un censo, no hay errores de medida, no hay falta de respuesta, etc.

El error de muestreo de una encuesta puede ser cuantificado si se ha utilizado un muestreo probabilístico para la selección de la muestra (véase la sección 1.5). Las principales ventajas del uso del muestreo frente al uso de censos son las siguientes:

- El muestreo con encuestas puede proporcionar información confiable a un coste menor que un censo. Por otra parte, en ocasiones, una unidad de análisis debe ser destruida para ser observada, por lo que en este caso una muestra proporciona información confiable sobre la población, mientras que un censo la destruiría, y con ello, la necesidad de información sobre ella.
- Los datos pueden ser recogidos de forma más rápida, por lo que las estimaciones de los parámetros de interés se pueden publicar de una manera programada y en un tiempo razonable.
- Las estimaciones basadas en el muestreo con encuestas suelen ser más acuradas que aquellas basadas en un censo, ya que en una muestra se puede prestar más atención a la recogida de la información y a la calidad de los datos a través del entrenamiento del personal y el seguimiento de las personas que no han respondido a la encuesta. Sin embargo, un censo requiere de una gran organización administrativa y requiere de muchas personas dedicadas a la recogida de la información. La complejidad administrativa y la presión por obtener las estimaciones a tiempo, puede conducir a muchos errores.

1.5 Muestreo de probabilidad: marco de referencia, el error cuadrático medio de un estimador y estimadores insesgados

1.5.1 Muestreo de probabilidad: marco de referencia

Considérese una población constituida por N elementos $\{u_1, ..., u_N\}$, que se denota por $U = \{1, \ldots, N\}$, donde el tamaño de la población N es conocido. Supongamos que se desea estudiar la variable objetivo 'ingreso del hogar'. Podríamos estar interesados por

ejemplo en obtener información sobre el total de ingresos de los hogares de la población, esto es,

$$Y_U = \sum_{k \in U} y_k$$

o bien sobre el ingreso medio

$$\bar{y}_U = \frac{1}{N} \sum_{k \in U} y_k,$$

donde y_k es el ingreso del hogar k.

Para ello se seleccionará una muestra, esto es, un subconjunto de elementos de la población U seleccionados de un marco muestral mediante un esquema de *muestreo probabilístico*. S representará a la muestra aleatoria y n_S al número de elementos o cardinal de S.

El *muestreo probabilístico* o *muestreo de probabilidad* es un enfoque de la selección de muestras que satisface determinadas condiciones, las cuales, para el caso de selección directa de elementos de la población, se describen a continuación:

1. Podemos definir el conjunto de muestras Ω que se pueden obtener con el proceso de muestreo.
2. Existe una probabilidad de selección conocida $p(s)$ asociada a cada muestra posible $s \in \Omega$.
3. El procedimiento otorga a cada elemento en la población una probabilidad no nula de selección.
4. Se selecciona una muestra mediante un mecanismo aleatorio bajo el cual cada posible muestra $s \in \Omega$ recibe exactamente una probabilidad $p(s)$.

Una muestra obtenida bajo estas cuatro condiciones se llama *muestra probabilística*.

La función $p(s)$ define una distribución de probabilidad sobre Ω. Se llama *diseño de muestreo*. Una definición más rigurosa se enuncia en la definición 7.

La probabilidad a la que se refiere el punto 3 se llama *probabilidad de inclusión* del elemento.

Definición 6

La *probabilidad de inclusión*, también denominada *probabilidad de inclusión de primer orden* del elemento u_k, se define como la probabilidad de que u_k pertenezca a la muestra. Se denota por π_k y se puede calcular como:

$$\pi_k = \mathbb{P}(u_k \in S) \tag{1.1}$$

La probabilidad de inclusión de segundo orden de los elementos u_k y u_l se define como la probabilidad de que u_k y u_l pertenezcan a la muestra. Se denota por π_{kl} y se puede calcular como:

$$\pi_{kl} = \mathbb{P}(u_k, u_l \in S) \tag{1.2}$$

La aleatorización a la que se refiere el punto 4 se lleva a cabo normalmente mediante la implementación de un algoritmo muestral, esto es, un procedimiento de selección de muestras probabilísticas. Existen múltiples tipos de algoritmos muestrales, que pueden clasificarse en (i) enumerativos, (ii) de martingalas, (iii) secuenciales, (iv) por extracción individual, (v) eliminatorios y (vi) de rechazo (véase Tillé 2006, para los detalles).

Generalmente hablando, estos algoritmos consisten en la realización secuencial de experimentos aleatorizados que producen como resultado un elemento seleccionado en la muestra tras cada experimento o bien la inclusión o exclusión en la muestra de cada elemento del marco. En el ejemplo 4 se muestra un ejemplo del primer caso.

Ejemplo 4. Consideremos el siguiente procedimiento de selección de la muestra:

1. Se selecciona un elemento de los N posibles con igual probabilidad: $\frac{1}{N}$.
2. Se selecciona un segundo elemento de entre los $N-1$ restantes con igual probabilidad: $\frac{1}{N-1}$.
3. $\vdots$
n. Se selecciona un elemento de entre los $N-n+1$ restantes con igual probabilidad: $\frac{1}{N-n+1}$.

Este algoritmo secuencial es una posible forma de realizar un muestreo con probabilidades iguales y sin reemplazamiento, que produce una muestra de tamaño n. Consiste en realizar n experimentos aleatorizados (en este caso extracciones) donde el resultado es la selección de un elemento en la muestra.

■

Definición 7

Dado un mecanismo de selección de la muestra (algoritmo muestral), se define el concepto de *diseño de muestreo*, *diseño muestral* o, simplemente, *diseño* como una función $p(\cdot)$ que a cada muestra s le hace corresponder la probabilidad de que dicha muestra sea seleccionada, $p(s)$, para todo s del conjunto de posibles muestras, denotado por Ω.

En otras palabras, el diseño muestral $p(\cdot)$ es la función de probabilidad de la variable

aleatoria S, que toma valores en Ω:

$$\mathbb{P}(S = s) = p(s), \text{ para todo } s \in \Omega.$$

La función $p(\cdot)$ define una función de probabilidad sobre el espacio muestral:

i. $p(s) \geq 0$, para todo $s \in \Omega$.

ii. $\sum_{s \in \Omega} p(s) = 1$.

El diseño muestral define un conjunto de muestras posibles, donde la probabilidad de selección de cada una de ellas es estrictamente positiva, $p(s) > 0$. El resto de muestras tendrán probabilidad nula de selección y no están en Ω.

Ejemplo 5. Bajo el mecanismo de selección de la muestra definido en el Ejemplo 4, la cardinalidad del espacio muestral es $\binom{N}{n}$ y todas las muestras tienen probabilidad igual a $\frac{1}{\binom{N}{n}}$. El diseño muestral es:

$$p(s) = \frac{1}{\binom{N}{n}}, \text{ para todo } s \in \Omega.$$

Se denota como diseño del *muestreo aleatorio simple sin reemplazamiento* (véase el tema 2).

■

A partir de ahora se considerarán exclusivamente muestras probabilísticas.

1.5.2 Estimadores y sus propiedades

Sea θ un vector $j-$dimensional que representa los parámetros poblacionales de interés, esto es, $\theta = (\theta_1, \ldots, \theta_j)$. Un estimador es un estadístico, es decir, una función real de la muestra aleatoria, propuesto para producir valores (estimaciones puntuales) del parámetro poblacional θ, que tomará distintos valores en función de la muestra elegida. Así, un estimador se describe en función de los valores muestrales $\theta = \theta(S)$.

Consideremos por ejemplo que se desea estimar el total poblacional de la variable y, $\theta(y_1, \ldots, y_N) = Y_U = \sum_{k \in U} y_k$, podemos definir el estimador $\widehat{\theta}(S) = \frac{N}{n} \sum_{k \in S} y_k$, usado en un diseño de muestreo con probabilidades iguales y sin reemplazamiento.

Definición 8

Un estimador puntual del parámetro poblacional θ es una función $\widehat{\theta}$ que a cada posible muestra S le hace corresponder una estimación $\widehat{\theta}(S)$ de θ.

Es importante distinguir entre estimador y estimación. Una estimación es el valor $\widehat{\theta}(s)$ que se puede calcular para una muestra particular s del conjunto de muestras posibles Ω. Por ejemplo, en el caso del muestreo aleatorio simple sin reemplazamiento, la variable aleatoria

$$\widehat{\theta}(S) = \frac{N}{n} \sum_{k \in S} y_k$$

es un estimador del parámetro poblacional $\theta = Y_U = \sum_{k \in U} y_k$; mientras que la estimación obtenida para una muestra particular seleccionada es el valor

$$\widehat{\theta}(s) = \frac{N}{n} \sum_{k \in s} y_k.$$

Al conjunto de posibles valores que puede tomar el estimador $\widehat{\theta}$ junto a la probabilidad de que $\widehat{\theta}$ tome dichos valores bajo un diseño muestral $p(s)$ se conoce como la distribución del estimador $\widehat{\theta}$ en el muestreo.

Definición 9

Sea $\widehat{\theta}$ un estimador para θ y C el conjunto de todos los posibles valores que produce el estimador. Para cada valor $c \in C$, la probabilidad de que el estimador tome dicho valor viene dado por

$$P_C = \mathbb{P}(\widehat{\theta} = c) = \sum_{s \in \Omega_C} p(s)$$

donde Ω_C es el conjunto de muestras s para las que $\widehat{\theta}(s) = c$.

La distribución del estimador en el muestreo es el par $\{C, P_C\}$.

A continuación se definen algunas propiedades importantes de los estimadores, el concepto de *insesgadez*, *error cuadrático medio*, *error de muestreo* o *error estándar* y *error relativo* o *coeficiente de variación* del estimador.

Definición 10

Un estimador $\widehat{\theta}$ es insesgado para el parámetro θ si su sesgo es 0, esto es:

$$\mathbb{B}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta}] - \theta = 0,$$

donde la esperanza del estimador se define como

$$\mathbb{E}[\widehat{\theta}] = \sum_{s \in \Omega} p(s) \cdot \widehat{\theta}(s).$$

Definición 11

El error cuadrático medio[a] del estimador $\widehat{\theta}$ se define como

$$\mathbb{MSE}(\widehat{\theta}) = \mathbb{E}\left[\left(\widehat{\theta} - \theta\right)^2\right] = \sum_{s\in\Omega} p(s)\cdot\left(\widehat{\theta}(s) - \theta\right)^2.$$

[a] *Mean square error*.

Definición 12

La varianza del estimador se define como

$$\mathbb{V}(\widehat{\theta}) = \sum_{s\in\Omega} p(s)\cdot\left[\widehat{\theta}(s) - \mathbb{E}[\widehat{\theta}]\right]^2.$$

Se puede demostrar el siguiente resultado:

Proposición 1

$$\mathbb{MSE}(\widehat{\theta}) = \mathbb{V}(\widehat{\theta}) + \left[\mathbb{B}(\widehat{\theta})\right]^2.$$

Demostración 1

En efecto:

$$\begin{aligned}\mathbb{MSE}(\widehat{\theta}) &= \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}(\widehat{\theta}) + \mathbb{E}(\widehat{\theta}) - \theta)^2\right] = \\ &= \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2\right] + \left[\mathbb{E}[\widehat{\theta}] - \theta\right]^2 = \\ &= \mathbb{V}(\widehat{\theta}) + \left[\mathbb{B}(\widehat{\theta})\right]^2\end{aligned}$$

Ejemplo 6. Consideremos el diseño de muestreo con probabilidades iguales y sin reemplazamiento para el cual se había propuesto el estimador

$$\widehat{\theta} = \frac{N}{n}\sum_{k\in S} y_k.$$

Este estimador es insesgado para $\theta = \sum_{k\in U} y_k$:

$$\mathbb{E}[\widehat{\theta}] = \mathbb{E}\left[\frac{N}{n}\sum_{k\in U} I_k\cdot y_k\right] = \frac{N}{n}\sum_{k\in U}\mathbb{E}[I_k]\cdot y_k = \frac{N}{n}\sum_{k\in U}\pi_k\cdot y_k \underbrace{=}_{\pi_k=\frac{n}{N}} \sum_{k\in U} y_k = \theta.$$

■

El sesgo y el error cuadrático medio son medidas importantes de la calidad de un estimador. En general, entre varios estimadores posibles para estimar el parámetro θ, se escogerá aquel cuya distribución en el muestreo se concentre más estrechamente alrededor del verdadero parámetro θ. Podríamos usar entonces como criterio de selección aquel estimador que tenga el menor error cuadrático medio, ya que hay razón de peso para creer que si el error cuadrático medio del estimador es bajo, la muestra extraída producirá una estimación cercana al valor verdadero. El error cuadrático medio se puede reducir a través de la varianza del estimador o bien reduciendo el sesgo. Normalmente el investigador se encargará de tomar un estimador que sea insesgado o aproximadamente insesgado y escogerá aquel que tenga menor varianza.

Como ya se ha comentado, en el proceso de estimación de un parámetro de la población se originan dos tipos de errores: el error de muestreo y los errores ajenos al muestreo. El error de muestreo del estimador se debe exclusivamente al cálculo de la estimación a partir de los datos de un subconjunto de la población.

Definición 13

El *error de muestreo* o *error estándar* del estimador θ se define como la raíz cuadrada de la varianza del estimador, denominado $\sigma(\widehat{\theta})$:

$$\sigma(\widehat{\theta}) = \left[\mathbb{V}(\widehat{\theta})\right]^{1/2}.$$

Otro indicador de precisión que a menudo suele obtenerse en las encuestas es una estimación del error relativo o *coeficiente de variación* del estimador.

Definición 14

El *error relativo* o *coeficiente de variación* del estimador se define como el cociente entre el error de muestreo y el valor esperado del estimador, esto es:

$$\mathrm{CV}(\widehat{\theta}) = \frac{\sigma(\widehat{\theta})}{\mathbb{E}[\widehat{\theta}]}.$$

El estimador comúnmente usado para estimar el coeficiente de variación cuando el estimador θ es insesgado o aproximadamente insesgado es:

$$cve(\widehat{\theta}) = \frac{\widehat{\sigma}(\widehat{\theta})}{\widehat{\theta}}.$$

A continuación veremos dos estimadores importantes en el contexto del muestreo con probabilidades desiguales: el estimador de Horvitz-Thompson (HT) y el estimador de

Hansen y Hurwitz (HH). En este tipo de muestreo la probabilidad de que u_k pertenezca a la muestra, esto es, la probabilidad de inclusión de primer orden π_k, no es igual para todo $u_k \in U$. Se denotará $u_k \in S$ o bien $k \in S$ para indicar que el elemento k pertenece a la muestra aleatoria.

Definición 15

Se denomina *estimador* π o estimador de Horvitz-Thompson de Y_U a:

$$\widehat{Y}_U^{\text{HT}} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

El coeficiente $\frac{1}{\pi_k}$ que multiplica a cada elemento se denomina *factor de expansión* (también denominado *peso de muestreo* o *factor de elevación*) e incrementa la importancia de los elementos en la muestra, de tal forma que *podría decirse* que el elemento $k-$ésimo, cuando está presente en la muestra, representa no solo a dicho elemento sino a $\frac{1}{\pi_k}$ elementos de la población. Así, el factor de expansión *podría interpretarse* como el número de elementos en la población a los que representa cada unidad en la muestra.

Ejemplo 7. El estimador $\widehat{Y}_U^{\text{HT}}$ para el diseño *muestral aleatorio simple sin reemplazamiento* es, como ya se ha comentado con anterioridad,

$$\widehat{Y}_U^{\text{HT}} = \frac{N}{n} \sum_{k \in S} y_k.$$

■

Antes de proceder a obtener la expresión de la varianza del estimador y el estimador de la varianza, vamos a definir la variable aleatoria *indicador de pertenencia a la muestra* del elemento k, denotado por $I_k = I_k(S)$, para representar la pertenencia de un elemento a la muestra. I_k toma dos valores: 1 si el elemento u_k pertenece a la muestra y 0, en caso contrario, esto es,

$$I_k = I_k(S) = \begin{cases} 1, \text{ si } u_k \in S, \\ 0, \text{ si } u_k \notin S. \end{cases}$$

I_k es una variable aleatoria con distribución Bernoulli con

$$\mathbb{P}(I_k = 1) = \mathbb{P}(u_k \in S) = \pi_k$$

Comentario 1. A partir de ahora y en los temas sucesivos se ignorará la distinción tipográfica entre S, la muestra aleatoria, y s, una muestra particular seleccionada del conjunto de muestras posibles. Por simplicidad, se usará s.

■

Algunas propiedades básicas del estadístico I_k se enuncian a continuación.

Proposición 2

Dado un diseño muestral $p(s)$ se tiene, para todo $k, l = 1, \ldots, N$:

i. $\mathbb{E}[I_k] = \pi_k$;

ii. $\mathbb{V}(I_k) = \pi_k \cdot (1 - \pi_k)$;

iii. $\mathbb{C}(I_k, I_l) = \pi_{kl} - \pi_k \cdot \pi_l$,

donde $\mathbb{E}[I_k]$ y $\mathbb{V}(I_k)$ representan la esperanza y la varianza de I_k, respectivamente, y $\mathbb{C}(I_k, I_l)$, la covarianza de I_k e I_l.

Demostración 2

I_k es una variable aleatoria con distribución Bernoulli con $\mathbb{P}(I_k = 1) = \pi_k$, usando (1.1). Por tanto:

i. $\mathbb{E}[I_k] = \pi_k$.

ii. $\mathbb{V}(I_k) = \mathbb{E}[I_k^2] - [\mathbb{E}[I_k]]^2 = \pi_k \cdot (1 - \pi_k)$.

iii. $\mathbb{C}(I_k, I_l) = \mathbb{E}[I_k \cdot I_l] - \mathbb{E}[I_k] - \mathbb{E}[I_l] = \pi_{kl} - \pi_k \cdot \pi_l$, usando (1.2).

Proposición 3

Dado un diseño muestral $p(s)$ con tamaño de muestra fijo n, se tiene:

i. $\sum_{k \in U} \pi_k = n$.

ii. $\sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \pi_{kl} = n \cdot (n - 1)$.

iii. $\sum_{\substack{l \in U \\ l \neq k}} \pi_{kl} = (n - 1) \cdot \pi_k$.

Demostración 3

i. $$\sum_{k \in U} \pi_k = \sum_{k \in U} \mathbb{E}\left[I_k\right] = \mathbb{E}\left[\sum_{k \in U} I_k\right] = \mathbb{E}[n] = n.$$

ii. $$\sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \pi_{kl} = \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \mathbb{E}[I_k \cdot I_l] = \mathbb{E}\left[\sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} I_k \cdot I_l\right] = \mathbb{E}[n \cdot (n - 1)] = n \cdot (n - 1).$$

iii. $$\sum_{\substack{l \in U \\ l \neq k}} \pi_{kl} = \sum_{\substack{l \in U \\ l \neq k}} \mathbb{E}[I_k \cdot I_l] = \mathbb{E}\left[I_k \cdot \left(\sum_{l \in U} I_l - I_k\right)\right] = \mathbb{E}[I_k \cdot (n - I_k)] =$$

$$= n \cdot \mathbb{E}[I_k] - \underbrace{\mathbb{E}[I_k^2]}_{\mathbb{E}[I_k]} = (n-1) \cdot \mathbb{E}[I_k] = (n-1) \cdot \pi_k.$$

A continuación se obtiene la varianza del estimador de Horvitz-Thompson, así como el estimador insesgado de la varianza, a través del siguiente Teorema.

Teorema 4

El estimador HT

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

es insesgado para el total $Y_U = \sum_{k \in U} y_k$ y la expresión de la varianza del estimador es

$$\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}}) = \sum_{k \in U} \sum_{l \in U} \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \cdot (\pi_{kl} - \pi_k \cdot \pi_l) \tag{1.3}$$

Si $\pi_{kl} > 0$ para todo $k, l \in U$, un estimador insesgado de $\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}})$ viene dado por la siguiente expresión debida a Horvitz y Thompson 1952:

$$\widehat{\mathbb{V}}\left(\widehat{Y}_U^{\mathrm{HT}}\right) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \cdot \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \tag{1.4}$$

Demostración 4

El estimador $\widehat{Y}_U^{\mathrm{HT}}$ es insesgado. En efecto:

$$\begin{aligned} \mathbb{E}[\widehat{Y}_U^{\mathrm{HT}}] &= \mathbb{E}\left[\sum_{k \in s} \frac{y_k}{\pi_k}\right] = \mathbb{E}\left[\sum_{k \in U} \frac{y_k}{\pi_k} \cdot I_k\right] = \sum_{k \in U} \frac{y_k}{\pi_k} \cdot \mathbb{E}[I_k] = \\ &= \sum_{k \in U} \frac{y_k}{\pi_k} \cdot \pi_k = \sum_{k \in U} y_k = Y_U \end{aligned}$$

La expresión de la varianza de $\widehat{Y}_U^{\mathrm{HT}}$ es:

$$\begin{aligned} \mathbb{V}(\widehat{Y}_U^{\mathrm{HT}}) &= \mathbb{V}\left(\sum_{k \in U} \frac{y_k}{\pi_k}\right) = \mathbb{V}\left(\sum_{k \in U} \frac{y_k}{\pi_k} \cdot I_k\right) = \\ &= \sum_{k \in U} \frac{y_k^2}{\pi_k^2} \cdot \mathbb{V}(I_k) + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \mathbb{C}\left(\frac{y_k}{\pi_k} \cdot I_k, \frac{y_l}{\pi_l} \cdot I_l\right) = \\ &= \sum_{k \in U} \frac{y_k^2}{\pi_k^2} \cdot \mathbb{V}(I_k) + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \cdot \mathbb{C}\left(I_k, I_l\right) = \end{aligned}$$

$$= \sum_{k\in U}\frac{y_k^2}{\pi_k^2}\cdot\pi_k(1-\pi_k)+\sum_{k\in U}\sum_{\substack{l\in U\\ l\neq k}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}(\pi_{kl}-\pi_k\pi_l) =$$
$$= \sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.$$

En cuanto a la insesgadez de$\widehat{\mathbb{V}}\left(\widehat{Y}_U^{\mathrm{HT}}\right)$, usando el indicador de pertenencia de un elemento a la muestra y siempre que $\pi_{kl}>0$ para todo $k,l\in U$, se puede escribir

$$\widehat{\mathbb{V}}\left(\widehat{Y}_U^{\mathrm{HT}}\right)=\sum_{j\in U}\sum_{l\in U}\frac{y_k}{\pi_k}\cdot\frac{y_l}{\pi_l}\cdot\frac{\pi_{kl}-\pi_k\cdot\pi_l}{\pi_{kl}}\cdot I_k\cdot I_l$$

cuya esperanza es

$$\mathbb{E}\left[\widehat{\mathbb{V}}\left(\widehat{Y}_U^{\mathrm{HT}}\right)\right] = \mathbb{E}\left[\sum_{j\in U}\sum_{l\in U}\frac{y_k}{\pi_k}\cdot\frac{y_l}{\pi_l}\cdot\frac{\pi_{kl}-\pi_k\cdot\pi_l}{\pi_{kl}}\cdot I_k\cdot I_l\right] =$$
$$= \sum_{j\in U}\sum_{l\in U}\frac{y_k}{\pi_k}\cdot\frac{y_l}{\pi_l}\cdot\frac{\pi_{kl}-\pi_k\cdot\pi_l}{\pi_{kl}}\cdot\pi_{kl}=\mathbb{V}\left(\widehat{Y}_U^{\mathrm{HT}}\right).$$

Corolario 5

Dado un diseño muestral $p(s)$ con un tamaño de muestra fijo n, entonces la varianza del estimador HT puede ser también escrita de la siguiente forma:

$$\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}})=-\frac{1}{2}\sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\cdot\pi_l)\cdot\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2. \tag{1.5}$$

Si $\pi_{kl}>0$ para todo $k,l\in U$, un estimador insesgado de $\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}})$ viene dado por la siguiente expresión debida a Yates y Grundy 1953 y Sen 1953:

$$\widehat{\mathbb{V}}(\widehat{Y}_U^{\mathrm{HT}})=-\frac{1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl}-\pi_k\cdot\pi_k}{\pi_{kl}}\cdot\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2. \tag{1.6}$$

Demostración 5

Se puede demostrar que las expresiones (1.3) y (1.5) son equivalentes cuando el tamaño del diseño es fijo. En efecto:

$$\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}}) = -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\cdot\pi_l)\cdot\left[\left(\frac{y_k}{\pi_k}\right)^2-2\cdot\frac{y_k}{\pi_k}\cdot\frac{y_l}{\pi_l}+\left(\frac{y_l}{\pi_l}\right)^2\right] =$$

$$= \sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\cdot\pi_l)\cdot\frac{y_k}{\pi_k}\cdot\frac{y_l}{\pi_l}-\sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\cdot\pi_l)\cdot\left(\frac{y_k}{\pi_k}\right)^2$$

donde $\sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\cdot\pi_l)\cdot\left(\frac{y_k}{\pi_k}\right)^2=\sum_{k\in U}\left(\frac{y_k}{\pi_k}\right)^2\sum_{l\in U}(\pi_{kl}-\pi_k\cdot\pi_l)=0$, ya que:

$$\begin{aligned}\sum_{l\in U}(\pi_{kl}-\pi_k\cdot\pi_l) &= \sum_{l\in U}\pi_{kl}-\pi_k\sum_{l\in U}\pi_l=\sum_{l\in U}\mathbb{E}[I_k\cdot I_l]-\pi_k\sum_{l\in U}\mathbb{E}[I_l]=\\ &= \mathbb{E}\left[\sum_{l\in U}I_k\cdot I_l\right]-\pi_k\cdot\mathbb{E}\left[\sum_{l\in U}I_l\right]=\mathbb{E}[n\cdot I_k]-\pi_k\cdot\mathbb{E}[n]=\\ &= n\cdot\pi_k-n\cdot\pi_k=0\end{aligned}$$

Se obtiene así la expresión (1.3).

En cuanto a la insesgadez del estimador de la varianza, si $\pi_{kl}>0$ para todo $k,l\in U$, entonces $\widehat{\mathbb{V}}(\widehat{Y}_U^{\mathrm{HT}})$ se puede escribir como

$$\begin{aligned}\widehat{\mathbb{V}}(\widehat{Y}_U^{\mathrm{HT}}) &= -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}I_l\cdot I_l\cdot\frac{\pi_{kl}-\pi_k\cdot\pi_k}{\pi_{kl}}\cdot\left(\frac{y_k}{\pi_l}-\frac{y_l}{\pi_l}\right)^2\\ \mathbb{E}[\widehat{\mathbb{V}}(\widehat{Y}_U^{\mathrm{HT}})] &= -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}\mathbb{E}[I_l\cdot I_l]\cdot\frac{\pi_{kl}-\pi_k\cdot\pi_k}{\pi_{kl}}\cdot\left(\frac{y_k}{\pi_l}-\frac{y_l}{\pi_l}\right)^2=\\ &= -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}\pi_{kl}\cdot\frac{\pi_{kl}-\pi_k\cdot\pi_k}{\pi_{kl}}\cdot\left(\frac{y_k}{\pi_l}-\frac{y_l}{\pi_l}\right)^2=\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}})\end{aligned}$$

Como se ha comentado, las dos expresiones de la varianza del estimador (1.3) y (1.5) son idénticas cuando el diseño produce muestras de tamaño fijo. Sin embargo, los dos estimadores insesgados de las varianzas (1.4) y (1.6) no son necesariamente iguales.

Puede comprobarse que para el diseño del *muestreo aleatorio simple sin reemplazamiento*, que es un diseño que produce muestras de tamaño fijo, las dos expresiones de la varianza del estimador dado en el Ejemplo 7 y del estimador de la varianza producen los mismos resultados.

El estimador HT será usado en los temas sucesivos cuando se utilicen esquemas de muestreo donde la selección de la muestra se realiza sin reemplazamiento. Existen otros esquemas de muestreo que permiten seleccionar elementos ya extraídos previamente, denominados esquemas de muestreo con reemplazamiento.

Supongamos que la probabilidad de que un elemento de la población sea seleccionado es $p_k>0,\ \forall k=1,\ldots,N$, con $\sum_{k\in U}p_k=1$ y se realizan n extracciones con reemplazamiento. El esquema general de un proceso de muestreo con reemplazamiento consiste en

seleccionar el primer elemento de la muestra, por ejemplo k_1, y devolverlo nuevamente al conjunto de posibles elementos a seleccionar. A continuación se extrae el segundo elemento de la muestra, k_2, y se vuelve a reponer, y así sucesivamente hasta realizar n extracciones. Se puede observar que las n extracciones son independientes y las probabilidades de seleccionar una unidad son las mismas en cada extracción. Podemos escribir la muestra resultante como una *muestra ordenada* de n elementos como sigue:

$$os = (k_1, \ldots, k_n).$$

La *muestra ordenada* contiene información sobre el orden de extracción y el número de veces que cada elemento aparece en la muestra. Sin embargo, la información sobre el orden de extracción no interesa para nuestro propósito, ya que muestras con los mismos elementos pero en distinto orden las consideraremos la misma. Así, para un diseño de tamaño fijo n, el número posible de muestras distintas es una combinación con repetición de N elementos tomados de n en n.

Se define la variable aleatoria m_k como el número de veces que aparece el elemento k en la muestra con $k = 1, \ldots, N$, cuya distribución es binomial de parámetros n y p_k. La probabilidad de seleccionar una muestra viene dada por el modelo multinomial:

$$\mathbb{P}(e_1 = m_1, \ldots, e_N = m_N) = \frac{n!}{m_1! \cdot \ldots \cdot m_N!} \cdot (p_1)^{m_1} \cdot \ldots \cdot (p_N)^{m_N},$$

donde $\sum_{k \in U} m_k = n$.

La probabilidad de inclusión del elemento k, $\forall k \in U$ se puede calcular como:

$$\begin{aligned} \pi_k &= \mathbb{P}(u_k \in s) = \mathbb{P}(m_k \neq 0) = 1 - \mathbb{P}(m_k = 0) = \\ &= 1 - \binom{n}{0} \cdot p_k^0 \cdot (1 - p_k)^n = 1 - (1 - p_k)^n \end{aligned}$$

El estimador que se usará en los esquemas de muestreo con reemplazamiento se define a continuación.

Definición 16

Sea Y_U el total poblacional de la variable y el parámetro de interés. El estimador de Y_U debido a Hansen y Hurwitz 1943 es

$$\widehat{Y}_U^{\text{HH}} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_{k_i}}{p_{k_i}}$$

y lo denominaremos el *estimador HH*.

Teorema 6

El estimador HH

$$\widehat{Y}_U^{\text{HH}} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_{k_i}}{p_{k_i}}$$

es un estimador insesgado para el total $Y_U = \sum_{k\in U} y_k$ y la expresión de la varianza del estimador es

$$\mathbb{V}(\widehat{Y}_U^{\text{HH}}) = \frac{1}{n}\sum_{k\in U}\left(\frac{y_k}{p_k} - Y_U\right)^2 \cdot p_k. \tag{1.7}$$

Un estimador insesgado de $\mathbb{V}(\widehat{Y}_U^{\text{HH}})$ viene dado por

$$\widehat{\mathbb{V}}(\widehat{Y}_U^{\text{HH}}) = \frac{1}{n\cdot(n-1)}\sum_{i=1}^{n}\left(\frac{y_{k_i}}{p_{k_i}} - \widehat{Y}_U^{\text{HH}}\right)^2. \tag{1.8}$$

Demostración 6

Se define $Z_i = \frac{y_{ki}}{p_{ki}}$ como una variable aleatoria que toma los valores $Z_i = \frac{y_k}{p_k}$ si $k_i = k$ para todo $k = 1, ..., N$, es decir, si el elemento k se ha seleccionado en la extracción i-ésima. Así pues, se dispone de n variables aleatorias cuya distribución de probabilidad viene dada por

$$\mathbb{P}\left(Z_i = \frac{y_k}{p_k}\right) = \mathbb{P}(k_i = k) = p_k, \quad k \in U.$$

El estimador HH puede escribirse en función de Z_i:

$$\widehat{Y}_U^{\text{HH}} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \bar{Z}$$

Las variables aleatorias $Z_1, ..., Z_n$ son independientes e idénticamente distribuidas, ya que se realizan extracciones independientes con las mismas probabilidades de selección $(p_1, ..., p_N)$ en cada extracción.

$$\begin{aligned}
\mathbb{E}[Z_i] &= \sum_{k\in U}\frac{y_k}{p_k}\cdot\mathbb{P}\left(Z_i = \frac{y_k}{p_k}\right) = \sum_{k\in U}\frac{y_k}{p_k}\cdot p_k = Y_U \\
\mathbb{V}(Z_i) &= \mathbb{E}\left[(Z_i - \mathbb{E}[Z_i])^2\right] = \mathbb{E}\left[(Z_i - Y_U)^2\right] = \sum_{k\in U}\left(\frac{y_k}{p_k} - Y_U\right)^2\cdot p_k
\end{aligned}$$

Dado que $\widehat{Y}_U^{\text{HH}}$ es la media aritmética de n variables aleatorias independientes e

idénticamente distribuidas, se tiene:

$$\begin{aligned}\mathbb{E}\left[\widehat{Y}_U^{\mathrm{HH}}\right] &= \frac{1}{n}\sum_{i=1}^{n} Y_U = Y_U \\ \mathbb{V}\left(\widehat{Y}_U^{\mathrm{HH}}\right) &= \frac{1}{n}\sum_{k\in U}\left(\frac{y_k}{p_k} - Y_U\right)^2 \cdot p_k\end{aligned}$$

En cuanto a la insesgadez de $\widehat{\mathbb{V}}\left(\widehat{Y}_U^{\mathrm{HH}}\right)$, dado que $Z_1, ..., Z_n$ son variables aleatorias independientes e idénticamente distribuidas, se tiene que la cuasivarianza muestral

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_{k_i}}{p_{k_i}} - \widehat{Y}_U^{\mathrm{HH}}\right)^2$$

es un estimador insesgado de la varianza de Z_i, quedando demostrada así la insesgadez de $\widehat{\mathbb{V}}\left(\widehat{Y}_U^{\mathrm{HH}}\right)$.

Bibliografía

Burg, T., A. Kowarik, M. Six, G. Brancato y D. Krapavickaité (2019). *Quality Guidelines for Frames in Social Statistics*. ESSnet KOMUSO Quality in Multisource Statistics. URL: https://ec.europa.eu/eurostat/cros/system/files/qgfss-v1.51.pdf.

Groves, R.M. (1989). *Survey errors and survey costs*. New York: Wiley.

Hansen, M.H. y W.N. Hurwitz (1943). "On the theory of sampling from finite populations". En: *Ann. Math. Statist.* 14, págs. 333-362.

Horvitz, D.G. y D.J. Thompson (1952). "A generalization of sampling without replacement from a finite universe". En: *Journal of the American Statistical Association* 47, págs. 663-685.

Lessler, J.T. y W.D. Kalsbeek (1992). *Nonsampling error in surveys*. New York: Wiley.

Lohr, S. (2010). *Sampling: Design and Analysis*. 2nd. Duxbury Press.

Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.

Sen, A.R. (1953). "On the estimate of the variance in sampling with varying probabilities". En: *Journal of the Indian Society of Agricultural Statistics* 5, págs. 119-127.

Tillé, Y. (2006). *Sampling algorithms*. Springer.

Yates, F. y P.M. Grundy (1953). "Selection without replacement from within strata with probability proportional to size". En: *Journal of the Royal Statistical Society B* 15, págs. 253-261.

Tema 2

Muestreo aleatorio simple. Estimadores insesgados, varianzas y la estimación de las varianzas. Intervalos de confianza. Estimación del tamaño de la muestra. Selección sistemática. Estimación por razones.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía:

W. G. Cochran (1977). *Sampling Techniques*. 3rd. Wiley

S. Lohr (2010). *Sampling: Design and Analysis*. 2nd. Duxbury Press

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

2.1 Introducción

El muestreo aleatorio simple es un diseño de muestreo directo de elementos. La selección de la muestra bajo este tipo de diseños requiere la existencia de un marco que identifique las unidades de muestreo, que son precisamente los elementos de la población.

Considérese una población constituida por N elementos $\{u_1, ..., u_N\}$, que se denota por $U = \{1, \ldots, N\}$ y supóngase que el parámetro de interés es el total poblacional de la *variable de estudio* y.

El muestreo aleatorio simple pertenece a la categoría de diseños denominados *diseños de muestreo con probabilidades iguales*. La característica común de este tipo de diseños es que las probabilidades de inclusión de primer orden π_k son iguales, esto es $\pi_k = constante,$ para todo $k = 1, \ldots, N$. π_k representa la probabilidad de que el elemento u_k pertenezca a la muestra. Se denotará $u_k \in s$ o bien $k \in s$ para indicar que el elemento k pertenece a la muestra aleatoria.

Supongamos que se desea obtener información sobre la variable 'ingreso del hogar'. A esta variable de interés se la denomina *variable de estudio*. Podríamos estar interesados por ejemplo en obtener información sobre el total de ingresos de los hogares de la población, esto es,

$$Y_U = \sum_{k \in U} y_k$$

Existen dos formas de seleccionar una muestra aleatoria simple: *muestreo aleatorio simple sin reemplazamiento*, donde todos los elementos de la muestra son distintos, y *muestreo aleatorio simple con reemplazamiento*, donde un mismo elemento puede aparecer más de una vez en la muestra.

2.2 Muestreo aleatorio simple sin reemplazamiento

2.2.1 Definición

El muestreo aleatorio simple sin reemplazamiento (también denominado sin reposición o sin reemplazo) es un diseño de muestreo con probabilidades iguales que produce muestras de tamaño fijo. Denotamos el muestreo aleatorio simple sin reemplazamiento como diseño *srswor*.

El diseño *srswor* consiste en seleccionar un subconjunto de n elementos de la población, no repetidos, donde cada elemento tiene la misma probabilidad de pertenecer a la muestra. Téngase en cuenta que en este diseño no hay reemplazamiento de las unidades seleccionadas previamente. Por otra parte, muestras con los mismos elementos pero con un orden distinto se consideran iguales. Así, la cardinalidad del espacio muestral es $\binom{N}{n}$ y todas las muestras tienen probabilidad igual a $\frac{1}{\binom{N}{n}}$.

El diseño muestral es:

$$p(s) = \frac{1}{\binom{N}{n}}, \text{ para todo } s \in \Omega.$$

Bajo el diseño *srswor*, las probabilidades de inclusión de primer y segundo orden de todos los elementos de la población son iguales. En efecto:

$$\begin{aligned}
\pi_k &= \mathbb{P}(u_k \text{ aparezca una vez en la muestra y las } n-1 \text{ unidades} \\
&\quad \text{restantes que forman parte de la muestra no sean } u_k) = \\
&= \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(N-n)!\cdot(n-1)!}}{\frac{N!}{(N-n)!\cdot n!}} = \frac{n}{N}; \;\; k = 1, \ldots, N. \\
\pi_{kl} &= \mathbb{P}(u_k, u_l \in s) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{\frac{(N-2)!}{(N-n)!\cdot(n-2)!}}{\frac{N!}{(N-n)!\cdot n!}} = \frac{n\cdot(n-1)}{N\cdot(N-1)}; \;\; k \neq l = 1, \ldots, N.
\end{aligned}$$

A continuación se presentan varios procedimientos para implementar el diseño $srswor$ y obtener como resultado una muestra aleatoria simple sin reemplazamiento.

Ejemplo 8. Consideremos el siguiente algoritmo secuencial de selección de la muestra, que consiste en realizar n experimentos aleatorizados (en este caso extracciones) donde el resultado es la selección de un elemento en la muestra.

1. Se selecciona un elemento de los N posibles con igual probabilidad: $\frac{1}{N}$.
2. Se selecciona un segundo elemento de entre los $N-1$ restantes con igual probabilidad: $\frac{1}{N-1}$.
3. $\vdots$

n. Se selecciona un elemento de entre los $N-n+1$ restantes con igual probabilidad: $\frac{1}{N-n+1}$.

■

Sin embargo, cuando el tamaño de la población es grande suele ser más conveniente usar mecanismos donde el resultado de cada experimento sea la inclusión o exclusión en la muestra del elemento. Veamos algunos ejemplos de algoritmos de este tipo.

Ejemplo 9. Este mecanismo de selección de la muestra es debido a Fan, Muller y Rezucha 1962.

- Se consideran N realizaciones independientes de una variable aleatoria con distribución uniforme sobre el intervalo $(0,1)$: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$.
- Sea $\frac{n}{N}$ una constante fijada, la selección o no selección de la primera unidad u_1 en la muestra se determina de la siguiente forma: si $\varepsilon_1 < \frac{n}{N}$, el elemento u_1 es seleccionado en la muestra, en caso contrario no.
- Para los siguientes elementos de la población, $k = 2, \ldots, N$, el elemento es seleccionado si
$$\varepsilon_k < \frac{n - n_k}{N - k + 1}$$
donde n_k es el tamaño de la muestra hasta ese momento, es decir, el número de elementos que han resultado seleccionados en la muestra de entre los primeros $k-1$ elementos para los que ya se ha realizado el experimento.
- El procedimiento termina cuando $n_k = n$.

■

Nótese que para implementar el esquema descrito en el Ejemplo 9 el tamaño de la población N debe ser conocido. En caso de no ser conocido, es necesario realizar una pasada preliminar a lo largo del marco muestral para determinar N. En la práctica puede ocurrir que el tamaño de la población de nuestro estudio sea desconocido, por ejemplo si la población de interés está referida a un subconjunto de los elementos del marco.

Ejemplo 10. McLeod y Bellhouse 1983 proponen un método sencillo de selección de una muestra aleatoria simple de tamaño n, que no requiere del conocimiento previo del tamaño N.

- Se seleccionan los primeros n elementos en la muestra inicial: $k_1, k_2, \ldots, k_n$.
- Para cualquier elemento $k > n$, se considera una realización de una variable aleatoria con distribución uniforme sobre el intervalo $(0, 1)$: ε_k. Se calcula

$$j = 1 + [\varepsilon_k \cdot k]$$

  donde $[\cdot]$ denota la parte entera. Si $j \leq n$ entonces el elemento j de la muestra actual es reemplazado por el elemento k de la población.
- El algoritmo termina cuando se ha realizado el experimento para todos los elementos.

Por tanto, al finalizar el algoritmo k contiene el valor del tamaño poblacional N. ■

Ejemplo 11. Otro esquema para seleccionar una muestra aleatoria simple es el siguiente:

- Se consideran N realizaciones independientes de una variable aleatoria con distribución uniforme sobre el intervalo $(0, 1)$: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$.
- Los valores obtenidos se ordenan de menor a mayor

$$\varepsilon_{(k_1)} < \varepsilon_{(k_2)} < \ldots < \varepsilon_{(k_N)}$$

  donde $\varepsilon_{(k_j)}$ es el j-ésimo valor más pequeño de los N valores obtenidos.
- La muestra de tamaño n estará compuesta por los n primeros elementos que se corresponden precisamente con aquellos para los que se obtuvieron los valores ε_k más pequeños $\varepsilon_{(k_1)}, \varepsilon_{(k_2)}, \ldots, \varepsilon_{(k_n)}$.

■

El esquema de selección de muestras descrito en el Ejemplo 11 presenta la ventaja de que permite obtener de forma simultánea varias muestras aleatorias simples no solapadas, es decir, donde cualquier elemento de la población solo puede haber sido seleccionado en una muestra. La segunda muestra estaría formada por los siguientes n elementos que se corresponden con los que siguientes valores ε_k más bajos. Esta muestra no está solapada con la primera. Y así sucesivamente.

Las muestras sin solapamientos son deseables por ejemplo cuando es necesario realizar varias encuestas diferentes sobre la misma población en un corto espacio de tiempo. Esto es un beneficio ya que así se reduce la carga del informante.

2.2.2 Estimadores, varianza y estimador de la varianza

Se propone el estimador de Horvitz-Thompson como estimador del total poblacional Y_U. Veamos a continuación cuál es el estimador y la varianza bajo el diseño *srswor*. Para demostrar los resultados se hará uso de las expresiones obtenidas en el Tema 1.

Proposición 7

El estimador de Horvitz-Thompson (HT) es

$$\widehat{Y}_U^{\mathrm{HT}} = \frac{N}{n} \cdot \sum_{k \in s} y_k$$

La varianza del estimador viene dada por

$$\mathbb{V}_{srswor}\left(\widehat{Y}_U^{\mathrm{HT}}\right) = N^2 \cdot \left(\frac{1}{n} - \frac{1}{N}\right) \cdot S_{yU}^2 = N^2 \cdot \frac{1-f}{n} \cdot S_{yU}^2 \tag{2.1}$$

donde $S_{yU}^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{y}_U)^2$ es la cuasivarianza poblacional de y, $f = \frac{n}{N}$ se denomina fracción de muestreo y $1 - f$ factor de corrección para poblaciones finitas.

Demostración 7

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} \frac{y_k}{\frac{n}{N}} = \frac{N}{n} \cdot \sum_{k \in s} y_k$$

La expresión de la varianza de $\widehat{Y}_U^{\mathrm{HT}}$ es:

$$\begin{aligned}
\mathbb{V}_{srswor}(\widehat{Y}_U^{\mathrm{HT}}) &= \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \cdot \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} = \\
&= \sum_{k \in U} \frac{y_k^2}{\pi_k} \cdot (1 - \pi_k) + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \cdot (\pi_{kl} - \pi_k \pi_l) = \\
&= \sum_{k \in U} \frac{y_k^2}{\frac{n}{N}} \cdot \left(1 - \frac{n}{N}\right) + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{y_k}{\frac{n}{N}} \cdot \frac{y_l}{\frac{n}{N}} \cdot \left[\frac{n \cdot (n-1)}{N \cdot (N-1)} - \frac{n}{N} \cdot \frac{n}{N}\right] = \\
&= \frac{N-n}{n} \cdot \left[\sum_{k \in U} y_k^2 - \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{y_k \cdot y_l}{N-1}\right]
\end{aligned}$$

Teniendo en cuenta que la varianza es invariante ante un cambio de origen, se tiene:

$$\mathbb{V}_{srswor}\left(\widehat{Y}_U^{\mathrm{HT}}\right) = \frac{N-n}{n} \cdot \left[\sum_{k\in U}(y_k - \bar{y}_U)^2 - \sum_{k\in U}\sum_{\substack{l\in U\\ l\neq k}} \frac{(y_k - \bar{y}_U)\cdot(y_l - \bar{y}_U)}{N-1}\right]$$

Recordando que $\sum_{k\in U}(y_k - \bar{y}_U) = 0$, se tiene

$$\left(\sum_{k\in U}(y_k - \bar{y}_U)\right)^2 = \sum_{k\in U}(y_k - \bar{y}_U)^2 + \sum_{k\in U}\sum_{\substack{l\in U\\ l\neq k}}(y_k - \bar{y}_U)\cdot(y_l - \bar{y}_U) = 0$$

De esta forma, la expresión de $\mathbb{V}_{srswor}(\widehat{Y}_U^{\mathrm{HT}})$ queda:

$$\begin{aligned}
\mathbb{V}_{srswor}\left(\widehat{Y}_U^{\mathrm{HT}}\right) &= \frac{N-n}{n}\cdot\left[\sum_{k\in U}(y_k - \bar{y}_U)^2 + \frac{1}{N-1}\cdot\sum_{k\in U}(y_k - \bar{y}_U)^2\right] = \\
&= \frac{N-n}{n}\cdot\left(1 + \frac{1}{N-1}\right)\cdot\sum_{k\in U}(y_k - \bar{y}_U)^2 = \\
&= N^2\cdot(1-f)\cdot\frac{S_{yU}^2}{n}
\end{aligned}$$

Proposición 8

Bajo el diseño del muestreo aleatorio simple sin reemplazamiento, un estimador insesgado de la cuasivarianza poblacional es la cuasivarianza muestral:

$$S_{ys}^2 = \frac{1}{n-1}\cdot\sum_{k\in s}(y_k - \bar{y}_s)^2$$

Demostración 8

$$\begin{aligned}
\mathbb{E}\left[S_{ys}^2\right] &= \frac{1}{n-1}\cdot\mathbb{E}\left[\sum_{k\in s}\left(y_k^2 - 2\cdot y_k\cdot\bar{y}_s + \bar{y}_S^2\right)\right] = \frac{1}{n-1}\cdot\mathbb{E}\left[\sum_{k\in s}y_k^2 - n\cdot\bar{y}_s^2\right] = \\
&= \frac{n}{n-1}\cdot\mathbb{E}\left[\sum_{k\in U}\frac{y_k^2}{n}\cdot I_k - \bar{y}_s^2\right] = \frac{n}{n-1}\left[\sum_{k\in U}\frac{y_k^2}{n}\cdot\underbrace{\mathbb{E}[I_k]}_{\frac{n}{N}} - \mathbb{E}\left[\bar{y}_s^2\right]\right] = \\
&= \frac{n}{n-1}\left[\frac{1}{N}\cdot\sum_{k\in U}y_k^2 - \bar{y}_U^2 - \mathbb{V}_{srswor}\left(\bar{y}_s\right)\right] = \frac{n}{n-1}\cdot\left[\sigma_{yU}^2 - \mathbb{V}\left(\bar{y}_s\right)\right]
\end{aligned}$$

Usando (2.1) se tiene:

$$\begin{aligned}\mathbb{E}\left[S_{ys}^2\right] &= \frac{n}{n-1}\cdot\left[\sigma_{yU}^2-\frac{1-f}{n}\cdot S_{yU}^2\right]= \\ &= \frac{n}{n-1}\cdot\left[\sigma_{yU}^2-\frac{1-f}{n}\cdot S_{yU}^2\right]= \\ &\underbrace{=}_{S_{yU}^2=\frac{N}{N-1}\cdot\sigma_{yU}^2} \frac{N}{N-1}\cdot\sigma_{yU}^2=S_{yU}^2\end{aligned}$$

Corolario 9

Un estimador insesgado de la varianza del estimador de Horvitz-Thompson es

$$\widehat{\mathbb{V}}_{srswor}(\widehat{Y}_U^{\mathrm{HT}})=N^2\cdot\frac{1-f}{n}\cdot S_{ys}^2 \tag{2.2}$$

Demostración 9

Al resultado se llega de forma obvia utilizando que la cuasivarianza muestral es un estimador insesgado de la cuasivarianza poblacional.

Ejemplo 12. Se desea estimar el consumo mensual de folios de los trabajadores de una empresa. Para ello se decide realizar un muestreo aleatorio simple sin reemplazamiento. Se seleccionan 200 de los 1000 trabajadores que hay en la empresa, obteniéndose un consumo medio de 60 folios por trabajador, y una cuasivarianza muestral de 25. La estimación del consumo mensual de folios a partir del uso del estimador de Horvitz-Thompson es

$$\widehat{Y}_U^{\mathrm{HT}}=\frac{N}{n}\sum_{k\in s}y_k=N\cdot\bar{y}_s=1000\cdot 60=60000$$

La estimación de la varianza del estimador es

$$\widehat{\mathbb{V}}_{srswor}\left(\widehat{Y}_U^{\mathrm{HT}}\right)=N^2\cdot(1-f)\cdot\frac{S_{ys}^2}{n}=1000^2\cdot\left(1-\frac{200}{1000}\right)\cdot\frac{25}{200}=100000$$

Luego la estimación del error de muestreo vendrá dado, por definición, por la raíz cuadrada de 100000. ■

Estimación de la media poblacional

Un estimador insesgado de la media poblacional $\bar{y}_U$ se obtiene diviviendo el estimador de Horvitz-Thompson para el caso del total por el tamaño poblacional N:

$$\widehat{\bar{y}}_U^{\mathrm{HT}}=\frac{1}{N}\cdot\widehat{Y}_U^{\mathrm{HT}}=\frac{1}{n}\cdot\sum_{k\in s}y_k=\bar{y}_s$$

En efecto:

$$\mathbb{E}\left[\widehat{\bar{y}}_U^{\text{HT}}\right] = \mathbb{E}\left[\frac{1}{N}\cdot\widehat{Y}_U^{\text{HT}}\right] = \frac{1}{N}\cdot Y_U = \bar{y}_U$$

De esta forma, la varianza y el estimador de la varianza pueden obtenerse de forma sencilla:

$$\begin{aligned}\mathbb{V}_{srswor}\left(\widehat{\bar{y}}_U^{\text{HT}}\right) &= \frac{1}{N^2}\cdot\mathbb{V}\left[\widehat{Y}_U^{\text{HT}}\right] = \frac{1-f}{n}\cdot S_{yU}^2\\ \widehat{\mathbb{V}}_{srswor}\left(\widehat{\bar{y}}_U^{\text{HT}}\right) &= \frac{1-f}{n}\cdot S_{ys}^2\end{aligned}$$

2.2.3 Estimación en dominios

En la mayoría de encuestas se suele estar interesado en realizar estimaciones no solo para la población completa U, sino también para determinadas subpoblaciones, denominadas *dominios de estudio* o simplemente *dominios*. Por ejemplo, podríamos tener interés en estimar el número de personas desempleadas por grupos de edad, sexo, ocupación o bien por regiones del país. La estimación realizada en determinadas subpoblaciones se conoce como *estimación en dominios*.

Sea U_d el dominio o subpoblación de interés, se define N_d como el tamaño de U_d, esto es, el número de elementos de la población que pertenecen al dominio U_d y, por otra parte, P_d, la proporción de elementos de la población U que pertenecen a U_d.

- **Estimación del tamaño absoluto y relativo de un dominio**

  Suponiendo conocido N y desconocido N_d, se desea estimar el valor de N_d y P_d. Para ello, se define la variable z_d, que toma los valores

  $$z_{dk} = \begin{cases}1, & \text{si } k \in U_d\\ 0, & \text{si } k \notin U_d\end{cases}$$

  para todo $k \in 1, \ldots, N$.

  El total de la variable z_d es el número de elementos que pertenecen a U_d y la media de la variable z_d es la proporción de elementos que pertenecen a U_d:

  $$\begin{aligned}\sum_{k\in U} z_{dk} &= N_d\\ \bar{z}_{dU} = \sum_{k\in U}\frac{z_{dk}}{N} &= \frac{N_d}{N} = P_d\end{aligned}$$

  Sea s la muestra aleatoria seleccionada de tamaño n con $n_d = \sum_{k\in s} z_{dk}$ el número de elementos de la muestra que pertenecen al dominio y $p_d = \frac{n_d}{n}$ la proporción de elementos de la muestra que pertenecen a U_d.

A continuación se aplicarán los resultados de la Proposición 7 para obtener el estimador HT de N_d y la varianza del estimador.

$$\widehat{N}_d = N \cdot p_d$$

Tomando $Q_d = 1 - P_d$, la varianza del estimador viene dada por:

$$\mathbb{V}_{srswor}\left(\widehat{N}_d\right) = N^2 \cdot (1-f) \cdot \frac{S^2_{z_dU}}{n} = N^2 \cdot \frac{N-n}{N-1} \cdot \frac{\sigma^2_{z_dU}}{n} = N^2 \cdot \frac{N-n}{N-1} \cdot \frac{P_d \cdot Q_d}{n}$$

ya que

$$\begin{aligned}
\sigma^2_{z_dU} &= \frac{1}{N} \cdot \sum_{k \in U} (z_{dk} - P_d)^2 = \frac{1}{N} \cdot \left[\sum_{k \in U} z^2_{dk} - 2 \cdot P_d \sum_{k \in U} z_{dk} + N \cdot P_d^2\right] = \\
&= \frac{1}{N} \cdot \left[\sum_{k \in U} z_{dk} - 2 \cdot N \cdot P_d^2 + N \cdot P_d^2\right] = \frac{1}{N} \cdot \left[N \cdot P_d - N \cdot P_d^2\right] = \\
&= P_d \cdot (1 - P_d) = P_d \cdot Q_d
\end{aligned}$$

Usando (2.2), un estimador insesgado de la varianza es:

$$\widehat{\mathbb{V}}_{srswor}\left(\widehat{N}_d\right) = N^2 \cdot (1-f) \cdot \frac{S^2_{ys}}{n} = N^2 \cdot (1-f) \cdot \frac{p_d \cdot q_d}{n-1}$$

ya que $S^2_{ys} = \frac{n}{n-1} \cdot p_d \cdot q_d$.

De forma análoga, para el caso de la proporción de individuos que pertenecen al dominio, un estimador insesgado para P_d es

$$\widehat{P}_d = p_d$$

La varianza y el estimador de la varianza se pueden calcular a partir de las expresiones vistas para el estimador de N_d:

$$\begin{aligned}
\mathbb{V}_{srswor}\left(\widehat{P}_d\right) &= \frac{1}{N^2} \cdot \mathbb{V}_{srswor}\left(\widehat{N}_d\right) = \frac{N-n}{N-1} \cdot \frac{P_d \cdot Q_d}{n} \\
\widehat{\mathbb{V}}_{srswor}\left(\widehat{P}_d\right) &= \frac{1}{N^2} \cdot \widehat{\mathbb{V}}_{srswor}\left(\widehat{N}_d\right) = (1-f) \cdot \frac{p_d \cdot q_d}{n-1}
\end{aligned}$$

Comentario 2. Estos resultados son importantes ya que permiten la estimación de la proporción poblacional y el total poblacional de los elementos de la población que pertenecen a un determinado dominio, como se muestra en el ejemplo siguiente. ■

Ejemplo 13. En un territorio donde existen 1500 colegios, se desea conocer la opinión de estos acerca de un nuevo proyecto educativo que se pretende implantar. Para ello, se selecciona una muestra aleatoria simple sin reemplazamiento de 300 colegios, obteniéndose la siguiente información muestral:

A favor	En contra	En blanco
225	50	25

Sea $z_{dk} = \begin{cases} 1, & \text{si el colegio está a favor del nuevo proyecto} \\ 0, & \text{en otro caso} \end{cases}$

Una estimación de la proporción de colegios a favor del proyecto educativo, usando el estimador HT, es

$$\widehat{P}_d = p_d = \frac{225}{300} = 0{,}75$$

Mientras que la estimación del número de colegios a favor del proyecto educativo es $\widehat{N} = 1500 \cdot 0{,}75 = 1125$.

La estimación del error de muestreo es

$$\widehat{\sigma}_{srswor}\left(\widehat{P}_d\right) = \sqrt{\left(1 - \frac{300}{1500}\right) \cdot \frac{0{,}75 \cdot 0{,}25}{299}} \approx 0{,}0224.$$

■

- **Estimación del total y la media de un dominio cuando el tamaño del dominio es desconocido.**

  Supongamos que estamos interesados en estimar el total de ingresos de los hogares con familia numerosa, esto es,

  $$Y_{U_d} = \sum_{k \in U_d} y_k$$

  Se define una nueva variable, y_d, tal que

  $$y_{dk} = \begin{cases} y_k, & \text{si } k \in U_d \\ 0, & \text{otro caso} \end{cases}$$

  El total de ingresos de los hogares con familia numerosa puede expresarse en función de esta nueva variable teniendo en cuenta todos los hogares de la población y, de esta manera, poder hacer uso de los resultados vistos en la Proposición 7 y el Corolario 9.

  $$Y_{U_d} = \sum_{k \in U} y_{dk}$$

El estimador HT es

$$\widehat{Y}_{U_d}^{\mathrm{HT}} = \sum_{k\in s} \frac{y_{dk}}{\pi_k} = \frac{N}{n} \sum_{k\in s} y_{dk} = \frac{N}{n} \sum_{k\in S_d} y_k$$

donde S_d es el subconjunto de la muestra aleatoria s que pertenece al dominio U_d.

La varianza de $\widehat{Y}_{U_d}^{\mathrm{HT}}$ viene dada por

$$\begin{aligned}
\mathbb{V}_{srswor}\left(\widehat{Y}_{U_d}^{\mathrm{HT}}\right) &= \frac{N^2 \cdot (1-f)}{n} \cdot S_{y_dU}^2 = \\
&= \frac{1-f}{n} \cdot \frac{N^2}{N-1} \cdot \left[\sum_{k\in U} y_{dk}^2 - N \cdot \left(\frac{1}{N} \sum_{k\in U} y_{dk}\right)^2\right] = \\
&= \frac{1-f}{n} \cdot \frac{N^2}{N-1} \cdot \left[\sum_{k\in U_d} y_k^2 - \frac{1}{N} \cdot \left(\sum_{k\in U_d} y_k\right)^2\right]
\end{aligned}$$

Un estimador insesgado de la varianza es

$$\begin{aligned}
\widehat{\mathbb{V}}_{srswor}\left(\widehat{Y}_{U_d}^{\mathrm{HT}}\right) &= \frac{N^2 \cdot (1-f)}{n} \cdot S_{y_dS}^2 = \\
&= \frac{1-f}{n} \cdot \frac{N^2}{n-1} \cdot \left[\sum_{k\in s} y_{dk}^2 - n \cdot \left(\frac{1}{n} \sum_{k\in s} y_{dk}\right)^2\right] = \\
&= \frac{1-f}{n} \cdot \frac{N^2}{n-1} \cdot \left[\sum_{k\in S_d} y_k^2 - \frac{1}{n} \cdot \left(\sum_{k\in S_d} y_k\right)^2\right]
\end{aligned}$$

Comentario 3. En numerosas ocasiones el tamaño del dominio N_d es desconocido. Sin embargo, si N_d fuese conocido, uno normalmente preferiría usar el estimador alternativo

$$\widehat{Y}_U^{Rat} = \frac{N_d}{n_d} \sum_{k\in S_d} y_k = N_d \cdot \bar{y}_{S_d}$$

Nótese que n_d es una variable aleatoria. Las expresiones de la varianza (que a menudo es mucho menor que la del estimador HT) y estimador de la varianza de este tipo de estimadores se estudiarán en 2.7. Estimación por razones.

■

2.3 Muestreo aleatorio simple con reemplazamiento

2.3.1 Definición

Un diseño muestral con reemplazamiento permite la selección de muestras con unidades repetidas, a diferencia de los esquemas sin reemplazamiento, que producen

muestras con todos sus elementos distintos. El muestreo aleatorio simple sin reemplazamiento, también denominado muestreo aleatorio simple con reposición, es un diseño de muestreo con probabilidades iguales que produce muestras de tamaño fijo. Denotamos el muestreo aleatorio simple con reemplazamiento como diseño *srswr*.

El diseño *srswr* consiste en seleccionar de forma independiente n elementos de la población de tamaño N con una probabilidad igual de extracción de $p_k = \frac{1}{N},\ \forall k = 1, \ldots, N$. Téngase en cuenta que en este diseño hay reemplazamiento de cada elemento extraído, es decir, elementos seleccionados en la muestra pueden ser elegidos de nuevo en la siguiente extracción.

Ejemplo 14. El siguiente procedimiento es una posible implementación del diseño *srswr*. Consiste en seleccionar secuencialmente y de forma independiente elementos de la población hasta obtener un subconjunto de n unidades: $k_1, k_2, \ldots, k_n$. En cada extracción, todos los elementos de la población tienen la misma probabilidad de extracción: $p_k = \frac{1}{N}$. Cada vez que un elemento es elegido, se vuelve a reponer, de forma que en cada extracción los N elementos de la población son susceptibles de ser seleccionados. ■

Sea $os = (k_1, \ldots, k_n)$ la muestra ordenada resultante del algoritmo muestral, donde se tiene información sobre el orden de extracción y el número de veces que cada elemento aparece en la muestra. Para nuestro propósito, la información sobre el orden de extracción no es de interés, pues consideraremos como la misma muestra aquellas que contengan los mismos elementos. Por tanto, la cardinalidad del espacio muestral es una combinación con repetición de N elementos tomados de n en n.

Se define la variable aleatoria m_k como el número de veces que aparece el elemento k en la muestra, cuya distribución es binomial de parámetros n y $\frac{1}{N}$, por lo que la esperanza y la varianza son, respectivamente:

$$\begin{aligned} \mathbb{E}(m_k) &= n \cdot \frac{1}{N} \\ \mathbb{V}(m_k) &= n \cdot \frac{1}{N} \cdot \left(1 - \frac{1}{N}\right) \end{aligned}$$

La probabilidad de seleccionar una muestra viene dada por el modelo multinomial:

$$\mathbb{P}(e_1 = m_1, \ldots, e_N = m_N) = \frac{n!}{m_1! \cdot \ldots \cdot m_N!} \cdot \left(\frac{1}{N}\right)^n,$$

En el diseño *srswr* las probabilidades de inclusión de primer y segundo orden de todos los elementos de la población son iguales.

$$\begin{aligned} \pi_k &= 1 - \left(1 - \frac{1}{N}\right)^n,\ \ k = 1, \ldots, N; \\ \pi_{kl} &= 1 - 2 \cdot \left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n,\ \ k \neq l = 1, \ldots, N. \end{aligned}$$

2.3.2 Estimadores, varianza y estimador de la varianza

Se propone el estimador de Hansen y Hurwitz como estimador del total poblacional Y_U. Veamos a continuación cuál es el estimador y la varianza bajo el diseño *srswr*. Para demostrar los resultados se hará uso de las expresiones obtenidas en el Tema 2.

Proposición 10

El estimador de Hansen-Hurwitz (HH) es

$$\widehat{Y}_U^{\text{HH}} = \frac{N}{n} \cdot \sum_{i=1}^{n} y_{k_i}$$

La varianza del estimador viene dada por

$$\mathbb{V}_{srswr}\left(\widehat{Y}_U^{\text{HH}}\right) = N \cdot (N-1) \cdot \frac{S_{yU}^2}{n} = N^2 \cdot \frac{\sigma_{yU}^2}{n} \tag{2.3}$$

Demostración 10

$$\widehat{Y}_U^{\text{HH}} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_{k_i}}{p_{k_i}} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_{k_i}}{\frac{1}{N}} = \frac{N}{n} \cdot \sum_{i=1}^{n} y_{k_i}$$

En cuanto a la varianza del estimador, usando la expresión (1.7), se llega a

$$\begin{aligned}
\mathbb{V}_{srswr}\left(\widehat{Y}_U^{HH}\right) &= \frac{1}{n}\sum_{k\in U}\left(\frac{y_k}{p_k} - Y_U\right)^2 \cdot p_k = \frac{1}{n}\sum_{k\in U}\left(\frac{y_k}{\frac{1}{N}} - Y_U\right)^2 \cdot \frac{1}{N} = \\
&= \frac{1}{n}\sum_{k\in U}\frac{\left(N\cdot y_k - N\cdot \bar{y}_U\right)^2}{N} = \frac{N^2}{n}\sum_{k\in U}\frac{\left(y_k - \bar{y}_U\right)^2}{N} = \\
&= \frac{N^2}{n}\cdot\sigma_{yU}^2
\end{aligned}$$

Comentario 4. En el diseño de muestreo aleatorio simple los estimadores de Horvitz-Thompson y Hansen y Hurwitz coinciden. ■

Proposición 11

Bajo el diseño del muestreo aleatorio simple con reemplazamiento, un estimador insesgado de la varianza poblacional es la cuasivarianza muestral:

$$S_{yos}^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(y_{k_i} - \bar{y}_{os})^2$$

Demostración 11

En la demostración 8 se llegó al siguiente resultado:

$$\mathbb{E}\left[S^2_{yos}\right] = \frac{n}{n-1} \cdot \left[\sigma^2_{yU} - \mathbb{V}\left(\bar{y}_{os}\right)\right]$$

Usando (2.3) se tiene lo que se pretendía demostrar

$$\mathbb{E}\left[S^2_{yos}\right] = \frac{n}{n-1} \cdot \left[\sigma^2_{yU} - \frac{\sigma^2_{yU}}{n}\right] = \sigma^2_{yU}$$

Corolario 12

Un estimador insesgado de la varianza es

$$\widehat{\mathbb{V}}\left(\widehat{Y}^{\text{HH}}_U\right) = N^2 \cdot \frac{S^2_{yos}}{n}$$

Demostración 12

Al resultado se llega de forma obvia utilizando que la cuasivarianza muestral es un estimador insesgado de la varianza poblacional.

2.3.3 Comparación del muestreo aleatorio simple sin y con reemplazamiento

Consideremos por un lado la estrategia compuesta por el diseño *srswr* y el estimador HH y, por otra parte, la estrategia formada por el diseño *srswor* y el estimador HT. Entonces, se tiene que el cociente de varianzas

$$\frac{V_{srswr}\left(\widehat{Y}^{\text{HH}}_U\right)}{V_{srswor}\left(\widehat{Y}^{\text{HT}}_U\right)} = \frac{N \cdot (N-1) \cdot \frac{S^2_{yU}}{n}}{N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{S^2_{yU}}{n}} = \frac{N-1}{N \cdot \left(1 - \frac{n}{N}\right)} = \frac{N-1}{N-n} \geq 1, \text{ para todo } n \geq 1.$$

Se obtiene que el muestreo sin reemplazamiento es más eficiente (salvo si el tamaño de la muestra es $n = 1$).

Ejemplo 15. Volvamos al contexto del Ejemplo 12, donde se estaba interesado en estimar el consumo mensual de folios de los trabajadores de una empresa. Nos preguntamos ahora cuál sería el tamaño de muestra necesario para garantizar la misma precisión (error de muestreo) usando el diseño *srswr* que la obtenida con el diseño *srswor*, suponiendo que la cuasivarianza poblacional es 30.

Por un lado, la varianza del estimador HT es

$$\mathbb{V}_{srswor}\left(\widehat{Y}_U^{\mathrm{HT}}\right) = N^2 \cdot (1-f) \cdot \frac{S_{yU}^2}{n} = 1000^2 \cdot \left(1 - \frac{200}{1000}\right) \cdot \frac{30}{200} = 120000$$

La varianza del estimador HH es

$$\mathbb{V}_{srswr}\left(\widehat{Y}_U^{\mathrm{HH}}\right) = N^2 \cdot \frac{\sigma_{yU}^2}{n'}$$

Dado que $\sigma_{yU}^2 = \frac{N-1}{N} \cdot S_{yU}^2$, el tamaño de muestra necesario puede calcularse como sigue:

$$n' = 1000^2 \cdot \frac{\frac{999}{1000} \cdot 30}{120000} = 249{,}75$$

Como cabía esperar, se necesitará un tamaño de muestra mayor, 250.

■

2.4 Intervalos de confianza

Un intervalo de confianza para el parámetro θ es un intervalo aleatorio de la forma

$$IC_\theta(s) = [L(s), U(s)]$$

donde L y U son dos estadísticos tales que $L(s) \leq U(s)$ para cada s.

Se desea que la probabilidad de que el intervalo $IC_\theta(s)$ contenga el parámetro θ sea cercana a la unidad. Esta probabilidad se denomina *nivel de confianza* y se representa como

$$\mathbb{P}\left[IC_\theta(s) \ni \theta\right] = 1 - \alpha,$$

donde α es la probabilidad acumulada de aquellas muestras para las cuales el intervalo de confianza no incluye al valor verdadero θ. Esto es:

$$\alpha = \sum_{s \in \Omega_0} p(s),$$

donde $p(\cdot)$ representa el diseño muestral y Ω_0 el conjunto de muestras para las cuales el intervalo de confianza obtenido no incluye a θ.

Una vez extraída una muestra, s, el intervalo de confianza para el parámetro θ se puede calcular y viene dado por

$$IC_\theta(s) = \left[L(s), U(s)\right].$$

Ejemplo 16. Sea $\widehat{Y}_U^{\mathrm{HT}}$ el estimador HT para el total poblacional Y_U con distribución normal cuya esperanza es Y_U y su varianza es $\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}})$, conocida. Un intervalo de confianza para Y_U viene dado por:

$$IC_{Y_U}(s) = \left[\widehat{Y}_U^{\mathrm{HT}} \pm z_{\alpha/2} \cdot \sqrt{\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}})}\right]$$

donde $z_{\alpha/2}$ es el valor que verifica $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$ donde Z es una variable aleatoria con distribución $N(0,1)$. ■

Ejemplo 17. Sea $\widehat{Y}_U^{\mathrm{HT}}$ el estimador HT para el total Y_U con distribución normal cuya esperanza es Y_U y su varianza es $\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}})$, desconocida. Un intervalo de confianza para y viene dado por:

$$IC_{Y_U}(s) = \left[\widehat{Y}_U^{\mathrm{HT}} \pm t_{n-1;\alpha/2} \cdot \sqrt{\widehat{\mathbb{V}}(\widehat{Y}_U^{\mathrm{HT}})}\right]$$

donde $t_{n-1;\alpha/2}$ es el valor que verifica $\mathbb{P}(T > t_{n-1;\,\alpha/2}) = \alpha/2$ considerando T una variable aleatoria con distribución t de Student con $n-1$ grados de libertad. ■

En los ejemplos anteriores se ha podido obtener un intervalo de confianza con nivel de confianza $1-\alpha$. Sin embargo, normalmente resulta complicado obtener intervalos de confianza con nivel de confianza exacto $1-\alpha$. A menudo se suele calcular el siguiente intervalo de confianza para el parámetro θ:

$$IC_\theta(s) = \left[\widehat{\theta} \pm z_{\alpha/2} \cdot \sqrt{\widehat{\mathbb{V}}(\widehat{\theta})}\right] \tag{2.4}$$

El intervalo de confianza dado en (2.4) contendrá al verdadero parámetro θ para una proporción de muestras, extraídas bajo el mismo diseño, de aproximadamente $1-\alpha$, si se verifican las siguientes condiciones:

1. La distribución del estimador $\widehat{\theta}$ en el muestreo es aproximadamente normal con esperanza θ y varianza $\mathbb{V}(\widehat{\theta})$. Esta condición es equivalente a afirmar que el Teorema Central del Límite se puede aplicar a $\widehat{\theta}$.
2. Existe un estimador consistente, $\widehat{\mathbb{V}}(\widehat{\theta})$, para $\mathbb{V}(\widehat{\theta})$.

Bajo estas condiciones, si consideramos la variable aleatoria

$$\frac{\widehat{\theta}-\theta}{\sqrt{\widehat{\mathbb{V}}(\widehat{\theta})}} = \frac{\widehat{\theta}-\theta}{\sqrt{\mathbb{V}(\widehat{\theta})}} \cdot \left(\frac{\mathbb{V}(\widehat{\theta})}{\widehat{\mathbb{V}}(\widehat{\theta})}\right)^{1/2},$$

es fácil ver que es una variable aleatoria con distribución aproximadamente $N(0,1)$, por lo que puede justificarse el uso del intervalo confianza dado en (2.4). Nótese que si $\mathbb{V}(\widehat{\theta})$ es conocida se usará el intervalo de confianza

$$IC_\theta(s) = \left[\widehat{\theta} \pm z_{\alpha/2} \cdot \sqrt{\mathbb{V}(\widehat{\theta})}\right]. \tag{2.5}$$

Ejemplo 18. Continuando con el Ejemplo 13, un intervalo de confianza al 95 % aproximadamente para la proporción de colegios a favor del proyecto educativo viene dado por

$$\left[\widehat{P}_d \pm z_{\alpha/2} \cdot \widehat{\sigma}_{srswor}\left(\widehat{P}_d\right)\right] = [0{,}75 \pm 1{,}96 \cdot 0{,}0224]\,.$$

■

2.5 Estimación del tamaño de la muestra

En la planificación de toda encuesta el investigador debe especificar los objetivos y las expectativas de la misma, así como realizar un inventario de recursos disponibles en términos de presupuesto, personal, metodología estadística, tecnologías, logística y cualquier otro equipamiento. Deberá equilibrar la precisión que desea obtener en las estimaciones de los parámetros de interés y el coste de la investigación estadística para determinar el tamaño de la muestra necesario de forma que permita cubrir los objetivos y no superar los recursos disponibles.

El investigador podría especificar una precisión deseada en términos absolutos: que la diferencia en valor absoluto entre el estimador y el verdadero valor del parámetro sea como mucho e_α con una probabilidad de $1-\alpha$.

$$\mathbb{P}(|\widehat{Y}_U^{\text{HT}} - Y_U| \leq e_\alpha) = 1 - \alpha$$

Ahora necesitamos determinar una ecuación que relacione la precisión y el tamaño de muestra. Suponiendo que el estimador HT tiene distribución normal con esperanza Y_U y varianza $\mathbb{V}(\widehat{Y}_U^{\text{HT}})$ se puede obtener el siguiente intervalo de confianza:

$$IC_{Y_U}(s) = \left[\widehat{Y}_U^{\text{HT}} \pm z_{\alpha/2} \cdot \sqrt{\mathbb{V}(\widehat{Y}_U^{\text{HT}})}\right]$$

Por tanto, se debe determinar el tamaño de muestra que satisfaga

$$e_\alpha = z_{\alpha/2} \cdot \mathbb{V}(\widehat{Y}_U^{\text{HT}}) = z_{\alpha/2} \cdot \sqrt{N^2 \cdot (1-f) \cdot \frac{S_{yU}^2}{n}}$$

De aquí podemos obtener la expresión del tamaño de muestra para el caso de muestreo aleatorio simple sin reemplazamiento

$$n = \frac{z_{\alpha/2}^2 \cdot N^2 \cdot S_{yU}^2}{e_\alpha^2 + z_{\alpha/2}^2 \cdot N \cdot S_{yU}^2} \tag{2.6}$$

En el caso de muestreo aleatorio simple con reemplazamiento, se tiene

$$n = \frac{z_{\alpha/2}^2 \cdot N^2 \cdot \sigma_{yU}^2}{e_\alpha^2} \tag{2.7}$$

Si por el contrario se fija un error de muestreo e, que se define como la desviación típica del estimador, entonces el tamaño de muestra es, para el caso sin reemplazamiento,

$$n = \frac{N^2 \cdot S_{yU}^2}{e_\alpha^2 + N \cdot S_{yU}^2}$$

Alternativamente, el investigador podría considerar una precisión deseada en términos relativos: que la diferencia en relativo entre el estimador y el verdadero valor del parámetro sea como mucho $e_{r\alpha}$ con una probabilidad de $1 - \alpha$.

$$\left(\left| \frac{\widehat{Y}_U^{\mathrm{HT}} - Y_U}{Y_U} \right| \leq e_{r\alpha} \right) = 1 - \alpha$$

El tamaño de muestra para una precisión relativa fijada se obtiene sustituyendo, en las expresiones (2.6) y (2.7), e_α por $e_{r\alpha} \cdot Y_U$. Por ejemplo, en el caso con reemplazamiento se llega a

$$n = \frac{z_{\alpha/2}^2 \cdot N^2 \cdot \sigma_{yU}^2}{e_{r\alpha}^2 \cdot Y_U^2} = \frac{z_{\alpha/2}^2 \cdot [CV(Y_U)]^2}{e_{r\alpha}^2}$$

donde $CV(Y_U)$ es el coeficiente de variación poblacional, que se define como el cociente entre la desviación típica y la media. Es una medida de error relativo que tiende a ser más estable en el tiempo en comparación con la varianza.

El investigador también podría estar interesado en determinar el tamaño de muestra fijado un error relativo de muestreo e_r, que se define como el coeficiente de variación del estimador, esto es,

$$e_r = CV(\widehat{Y}_U^{\mathrm{HT}}) = \frac{\sigma(\widehat{Y}_U^{\mathrm{HT}})}{\mathbb{E}(\widehat{Y}_U^{\mathrm{HT}})} = \frac{\sigma(\widehat{Y}_U^{\mathrm{HT}})}{Y_U}. \tag{2.8}$$

Por tanto, el tamaño de muestra para el caso con reemplazamiento es

$$n = \frac{[CV(Y_U)]^2}{e_r^2}$$

Finalmente, el investigador deberá estimar las cantidades desconocidas S_{yU}^2 o $CV(Y_U)$ a través de alguna de las siguientes formas:

1. *Muestra piloto*: se trata de una muestra a pequeña escala extraída para obtener información sobre la población. Suele utilizarse como guía para el diseño de la encuesta principal, por lo que podría ser usada para estimar el tamaño de muestra.
2. Uso de información de estudios realizados con anterioridad que guarden relación con la investigación estadística en cuestión.
3. Si no se dispone de más información, se deben estimar las cantidades desconocidas. Podrían por ejemplo estimarse a partir de la suposición de una distribución hipotética de los datos de la población que el investigador considere.

Comentario 5. Si estamos interesados en estimar una proporción poblacional, por ejemplo, la proporción de colegios a favor de un nuevo proyecto educativo, también es posible tomar $P_d = \frac{1}{2}$ y estimar S^2_{yU} por $\frac{N}{N-1} \cdot \frac{1}{4}$, ya que la varianza del estimador se maximiza cuando P_d toma dicho valor. Entonces, para poblaciones grandes, $S^2_{yU} \approx \frac{1}{4}$.

■

Ejemplo 19. Sea una población constituida por 1000 hogares, de la cual se desea determinar el tamaño de muestra necesario para estimar el gasto total semanal en alimentación de los hogares con un error relativo deseado de $2\,\%$, mediante un muestreo aleatorio simple sin reemplazamiento. Se extrae una muestra piloto de 50 hogares, obteniéndose un coeficiente de variación $\frac{S_{ys}}{\bar{y}_s}$ en la muestra de $0{,}2$.

De la ecuación (2.8) se llega a la siguiente expresión

$$e_r^2 = \frac{1}{Y_U^2} \cdot \mathbb{V}\left(\widehat{Y}_U^{\mathrm{HT}}\right) = \frac{1}{Y_U^2} \cdot N^2 \cdot \frac{1-f}{n} \cdot S^2_{yU} = \left(\frac{1}{n} - \frac{1}{N}\right) \cdot \frac{S^2_{yU}}{\bar{y}_U^2}$$

De aquí se obtiene la fórmula para el tamaño de la muestra

$$n = \frac{\dfrac{S^2_{yU}}{\bar{y}_U^2}}{e_r^2 + \dfrac{1}{N} \cdot \dfrac{S^2_{yU}}{\bar{y}_U^2}}$$

Utilizando la información proporcionada por la muestra piloto, podemos estimar el tamaño de muestra:

$$n = \frac{0{,}2^2}{0{,}02^2 + \frac{1}{1000} \cdot 0{,}2^2} \approx 90{,}9$$

El tamaño de muestra que garantiza la precisión deseada es tomar $n = 91$.

■

2.6 Muestreo sistemático

El muestreo sistemático es un diseño de muestreo directo de elementos. Consiste en seleccionar aleatoriamente, y con igual probabilidad, un elemento de entre los primeros a elementos del marco poblacional. El entero positivo a se fija previamente y se denomina *intervalo muestral*. El resto de la muestra se selecciona sistemáticamente tomando cada uno de los siguientes a elementos a continuación, hasta el final de la lista. Por tanto, existen sólo a muestras posibles, cada una con una probabilidad $\frac{1}{a}$ de ser seleccionada.

El muestreo sistemático ofrece varias ventajas prácticas, en particular su simplicidad de ejecución. El hecho de solo realizar una única selección aleatoria es una gran ventaja. Es fácil, por ejemplo, para un entrevistador seleccionar una muestra sistemática mientras está en campo.

Definición 17

sea a el *intervalo muestral* fijo y sea n la parte entera de $\frac{N}{a}$, donde N es el tamaño poblacional. Entonces

$$N = na + c$$

donde el entero c verifica $0 \leq c < a$. Si $c = 0$, el tamaño muestral n será seleccionado mediante el procedimiento que presentamos a continuación. Si $c > 0$, el tamaño muestral será n o $n + 1$. La selección de la muestra se realiza de la forma siguiente:

i Se selecciona un número aleatorio entre 1 y a con igual probabilidad $\frac{1}{a}$. Denótese por r y se conoce como el *punto de arranque* o *arranque aleatorio*.

ii La muestra seleccionada está compuesta por

$$s = \{k : k = r + (j-1) \cdot a \leq N; j = 1, 2, ..., n_s\} = s_r \tag{2.9}$$

donde el tamaño muestral n_s es $n + 1$ (cuando $r \leq c$), o n (cuando $c < r \leq a$).

Como se puede observar, una vez seleccionada la primera unidad, el resto de elementos que formarán parte de la muestra viene determinado por la posición que ocupa esa unidad en el conjunto de los primeros a elementos del marco. El espacio muestral está formado por a muestras que no se solapan $\{s_1, s_2, \ldots, s_a\}$, donde todos los elementos son distintos en cada muestra. Las muestras son equiprobables, con probabilidad $\frac{1}{a}$.

El diseño del muestreo sistemático, que se denota por sys, viene dado por

$$p(s) = \begin{cases} \frac{1}{a} & \text{si } s \in \{s_1, s_2, \ldots, s_a\} \\ 0 & \text{otro caso} \end{cases}$$

Comentario 6. Las a muestras posibles respresentan una partición de la población total U en a subpoblaciones, esto es, $U = \bigcup_{r=1}^{a} s_r$.

	Muestra, s				
	s_1	$\ldots$	s_r	$\ldots$	s_a
Valores de y	y_1	$\ldots$	y_r	$\ldots$	y_a
	y_{a+1}	$\ldots$	y_{a+r}	$\ldots$	y_{2a}
	$\vdots$		$\vdots$		$\vdots$
	$y_{(n-1)a+1}$	$\ldots$	$y_{(n-1)a+r}$	$\ldots$	y_N
Total muestral	Y_{s_1}	$\ldots$	Y_{s_r}	$\ldots$	Y_{s_a}

Tabla 2.1: Muestreo sistemático en el caso $c = 0$.

Así pues, el diseño sys puede ser descrito como una selección aleatoria con igual probabilidad de una de las a subpoblaciones. En la muestra seleccionada se encuesta a todos sus miembros. La tabla 2.1 ilustra la situación para el caso en el que $c = 0$.

■

Teorema 13

Bajo el diseño sys, considerando el intervalo muestral a, el estimador de Horvitz-Thompson del total poblacional Y_U es

$$\widehat{Y}_U^{\text{HT}} = a \sum_{k \in s} y_k$$

donde s es uno de los elementos del conjunto de posibles muestras $\{s_1, \ldots, s_r, \ldots, s_a\}$, con s_r definido por (2.9).

La varianza viene dada por

$$\mathbb{V}_{sys}(\widehat{Y}_U^{\text{HT}}) = a \sum_{r=1}^{a} \left(Y_{s_r} - \frac{1}{a} \cdot Y_U \right)^2$$

donde $Y_{s_r} = \sum_{k \in s_r} y_k$.

Demostración 13

Dado que cada elemento k pertenece a una y solo una de las a muestras sistemáticas equiprobables, entonces $\pi_k = \frac{1}{a} \ \forall \in U$. Por definición, la expresión del estimador HT queda

$$\widehat{Y}_U^{\text{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k} = a \sum_{k \in s} y_k$$

Las probabilidades de inclusión de segundo orden son

$$\pi_{kl} = \begin{cases} \frac{1}{a}, & \text{si } k \text{ y } l \text{ pertenecen a la misma muestra sistemática} \\ 0, & \text{otro caso} \end{cases}$$

Por tanto, la varianza del estimador HT viene dado por

$$\begin{aligned} \mathbb{V}_{sys}(\widehat{Y}_U^{\text{HT}}) &= \sum_{k \in U} \sum_{l \in U} \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l} \cdot (\pi_{kl} - \pi_k \cdot \pi_l) \\ &= \sum_{k \in U} \sum_{l \in U} \frac{\pi_{kl}}{\pi_k \cdot \pi_l} \cdot y_k \cdot y_l - \left(\sum_{k \in U} y_k \right)^2 = \end{aligned}$$

$$= a\sum_{r=1}^{a}\left\{\sum_{k\in s_r}\sum_{l\in s_r} y_k\cdot y_l\right\} - Y_U^2 = a\sum_{r=1}^{a} Y_{s_r}^2 - Y_U^2 =$$
$$= a\sum_{r=1}^{a}\left(Y_{s_r} - \frac{1}{a}\cdot Y_U\right)^2$$

Comentario 7. Si se verifica $N = a\cdot n$, el estimador HT viene dado por

$$\widehat{Y}_U^{\mathrm{HT}} = N\sum_{k\in s_r}\frac{y_k}{n} = N\cdot \bar{y}_{s_r}$$

La varianza del estimador HT es

$$\mathbb{V}_{sys}(\widehat{Y}_U^{\mathrm{HT}}) = \frac{N^2}{a}\sum_{r=1}^{a}(\bar{y}_{s_r} - \bar{y}_U)^2 \tag{2.10}$$

■

Comentario 8. Dado que la condición $\pi_{kl} > 0$ no se verifica para todo $k \neq l$, no se debería usar el estimador de la varianza del estimador HT dado en el Tema 1. La fórmula da, en este caso, un resultado sin sentido.

■

En el diseño sys, como ya se ha comentado, si $c = 0$, entonces $N = n\cdot a$, y todas las posibles muestras tienen el mismo tamaño muestral n. Sin embargo, si $c > 0$, el tamaño muestral será $n+1$ (si $r \leq c$) o n (si $r > c$). A continuación se presenta un método para controlar el tamaño de muestra de forma que siempre sea n. Se denomina el método de *muestreo sistemático circular*.

Este método consiste en considerar el marco como si este fuera circular, es decir, el último elemento $(k = N)$ va seguido del primero $(k = 1)$, y así sucesivamente. Se selecciona un número aleatorio r entre 1 y N con igual probabilidad. Sea a el entero más cercano a $\frac{N}{n}$. Entonces la muestra está formada por los elementos k tales que, para $j = 1, .., n$,

$$k = r + (j-1)\cdot a \quad \text{si} \quad r + (j-1)\cdot a \leq N$$

o

$$k = r + (j-1)\cdot a - N \quad \text{si} \quad r + (j-1)\cdot a > N$$

De esta forma, cada muestra tendrá tamaño n, y $\pi_k = \frac{n}{N}$ para cada k.

Comentario 9. Para calcular la varianza del estimador HT cuando se usa el método circular, debemos primero calcular π_{kl} para todo $k \neq l$. Esto requiere especial atención, ya que las muestras posibles no son necesariamente disjuntas por parejas.

■

2.6.1 Eficiencia del muestreo sistemático y comparación con el muestreo aleatorio simple

La eficiencia del muestreo sistemático depende en gran medida de la ordenación particular de los N elementos del marco. Si la ordenación de los elementos es tal que las muestras sistemáticas resultantes tienen aproximadamente el mismo valor de Y_{s_r}, entonces la varianza será pequeña. Por tanto, cuanto más homogéneos sean los elementos pertenecientes a una misma muestra sistemática, menos eficiente es el muestreo sistemático. Luego para conseguir una ordenación poblacional favorable para el muestreo sistemático, deberíamos hacer un esfuerzo por conseguir una ordenación que implique un bajo grado de homogeneidad entre los elementos dentro de la misma muestra.

A continuación se definirán dos medidas del grado de homogeneidad que existe entre los elementos dentro de la misma muestra sistemática; antes veremos que la variación total, SST , puede descomponerse en la variación dentro de las muestras sistemáticas[1], SSW, y la variación entre las muestras sistemáticas[2], SSB, como en un análisis de la varianza:

$$\underbrace{\sum_{k\in U}(\bar{y}_{s_r}-\bar{y}_U)^2}_{SST}=\underbrace{\sum_{r=1}^{a}\sum_{k\in s_r}(y_k-\bar{y}_{s_r})^2}_{SSW}+\underbrace{\sum_{r=1}^{a}n\cdot(\bar{y}_{s_r}-\bar{y}_U)^2}_{SSB}$$

Consideremos por simplicidad el caso donde $N=a\cdot n$ y veamos que la varianza del estimador HT dada en (2.10) puede representarse en función de SSB:

$$\mathbb{V}_{sys}(\widehat{Y}_U^{\mathrm{HT}}) \;=\; \frac{N^2}{a}\sum_{r=1}^{a}(\bar{y}_{s_r}-\bar{y}_U)^2=\frac{N^2}{a}\cdot\frac{SSB}{n}=N\cdot SSB \tag{2.11}$$

Una disminución de la variación dentro de las muestras SSW viene acompañada de un correspondiente incremento en la variación entre las muestras SSB, por lo que, nuevamente, se puede ver que cuanto más homogéneos sean los elementos dentro de la misma muestra sistemática, menos eficiente el muestreo sistemático es.

Definición 18

El siguiente coeficiente se conoce como una medida de homogeneidad entre los elementos de una misma muestra sistemática:

$$\delta=1-\frac{N-1}{N-a}\cdot\frac{SSW}{SST} \tag{2.12}$$

Se puede demostrar que $-\frac{a-1}{N-a}\leq\delta\leq 1$. En un extremo, si $SSW=0$ entonces $\delta=1$, que se da cuando hay homogeneidad completa dentro de las muestras sistemáticas. En el otro extremo, si $SSB=0$ entonces $\delta=-\frac{a-1}{N-a}$, que sucede cuando existe heterogeneidad completa dentro de las muestras (situación ideal).

[1] en inglés *within systematic samples*
[2] en inglés *between systematic samples*

Otra posible medida de homogeneidad es usar el denominado *coeficiente de correlación intraclase*

$$\rho = 1 - \frac{n}{n-1} \cdot \frac{SSW}{SST}$$

Se puede interpretar como una medida de la correlación entre pares de elementos dentro de la misma muestra sistemática. En un extremo, $\rho = 1$ si $SSW = 0$, es decir, hay una homogeneidad completa (no hay variación) dentro de las muestras sistemáticas. En el otro extremo, $\rho = -\frac{1}{n-1}$ si $SSB = 0$, es decir, heterogeneidad completa dentro de las muestras.

La relación que existe entre δ y ρ es

$$\rho = 1 - \frac{N}{N-1} \cdot (1 - \delta)$$

En efecto:

$$\frac{SSW}{SST} = \frac{n-1}{n} \cdot (\rho - 1) \text{ y por otra parte } \frac{SSW}{SST} = \frac{N-a}{N-1} \cdot (\delta - 1) = \frac{a \cdot (n-1)}{N-1} \cdot (\delta - 1)$$

Igualando las dos expresiones se tiene el resultado:

$$\rho = 1 + \frac{a \cdot n}{N-1} \cdot (\delta - 1) = 1 - \frac{N}{N-1} \cdot (1 - \delta)$$

Utilizaremos δ como medida de homogeneidad en lugar de ρ, ya que la representación de esta última solo se verifica para clases (muestras) de igual tamaño. Sin embargo, la representación de δ permite ser aplicada tanto si las muestras son de igual tamaño como si no.

Teorema 14

Bajo el diseño sys (con $N = a \cdot n$, donde a es un entero), la varianza del estimador HT del total poblacional se puede escribir como

$$\mathbb{V}_{sys}(\widehat{Y}_U^{\text{HT}}) = N^2 \cdot \frac{S_{yU}^2}{n} \cdot [(1-f) + (n-1) \cdot \delta] \tag{2.13}$$

donde $f = \frac{n}{N} = \frac{1}{a}$ es la fracción muestral.

Demostración 14

Usando 2.11 y 2.12:

$$\delta = 1 - \frac{N-1}{N-a} \cdot \frac{SSW}{SST} = 1 - \frac{N-1}{N-a} \cdot \left(1 - \frac{SSB}{SST}\right)$$

Por tanto

$$SSB = SST \cdot \left[1 + (\delta - 1) \cdot \frac{N-a}{N-1}\right] = \frac{SST}{N-1} \cdot [(N-1) + (\delta-1) \cdot a \cdot (n-1)] =$$
$$= S_{yU}^2 \cdot [a - 1 + a \cdot (n-1) \cdot \delta] = N \cdot \frac{S_{yU}^2}{n} \cdot [(1-f) + (n-1) \cdot \delta]$$

La varianza es, como se pretendía demostrar,

$$\mathbb{V}_{sys}(\widehat{Y}_U^{\mathrm{HT}}) = N \cdot SSB = N^2 \cdot \frac{S_{yU}^2}{n} \cdot [(1-f) + (n-1) \cdot \delta]$$

La expresión de la varianza dada en (2.13) es útil porque permite comparar de forma sencilla el muestreo sistemático y el muestreo aleatorio simple. La varianza del estimador HT del total población bajo el diseño $srswor$ es

$$\mathbb{V}_{srswor}\left(\widehat{Y}_U^{\mathrm{HT}}\right) = N^2 \cdot (1-f) \cdot \frac{S_{yU}^2}{n}$$

El diseño sys es más eficiente que $srswor$ si $\delta < 0$. Para crear una situación donde se verifique esta condición, debemos (si es posible) organizar la población de forma que los y_k dentro de cada muestra sistemática muestren una considerable heterogeneidad.

Ejemplo 20. El siguiente ejemplo muestra el efecto que tienen las distintas ordenaciones de la población cuando se aplica muestreo sistemático. Supóngase que se desea obtener una muestra de tamaño n de una población formada por $N = 100$ elementos y que la variable de estudio y toma los valores $1, 2, ..., 100$. Entonces el número de muestras es $a = \frac{N}{n} = 10$. Se tiene:

$$S_{yU}^2 = \frac{N(N+1)}{12} = \frac{100 \cdot 101}{12}$$

Independientemente de la ordenación de la población, la varianza bajo el diseño $srswor$ es

$$\mathbb{V}_{srswor}(\widehat{Y}_U^{\mathrm{HT}}) = N^2 \cdot \frac{1-f}{n} \cdot S_{yU}^2 = 7{,}575 \cdot 10^5$$

Ahora se aplicará el diseño sys con dos ordenaciones distintas:

a) Supongamos la ordenación en la que $y_k = k \ \ (k = 1, ..., 100)$, es decir, una tendencia lineal perfecta en los valores y_k. La tabla 2.2 muestra las diez posibles muestras sistemáticas y Y_{s_r} es el total de la muestra r-ésima.

   En este caso, $\mathbb{V}_{sys}(\widehat{Y}_U^{\mathrm{HT}}) = 8{,}25 \cdot 10^4$, se obtiene una eficiencia mayor con el diseño sys. La medida de homogeneidad δ es $-0{,}089$, que no está lejos del valor mínimo $\delta_{min} = -0{,}1$.

	r									
	1	2	3	4	5	6	7	8	9	10
y_k	1	2	3	4	5	6	7	8	9	10
	11	12	13	14	15	16	17	18	19	20
	21	22	23	24	25	26	27	28	29	30
	31	32	33	34	35	36	37	38	39	40
	41	42	43	44	45	46	47	48	49	50
	51	52	53	54	55	56	57	58	59	60
	61	62	63	64	65	66	67	68	69	70
	71	72	73	74	75	76	77	78	79	80
	81	82	83	84	85	86	87	88	89	90
	91	92	93	94	95	96	97	98	99	100
Y_{s_r}	460	470	480	490	500	510	520	530	540	550

Tabla 2.2: Ordenación con tendencia lineal perfecta.

b) Una ordenación óptima (varianza mínima) para el muestreo sistemático viene dada en la siguiente tabla 2.3.

Puesto que todos los totales muestrales Y_{s_r} son iguales, $\mathbb{V}_{sys}(\widehat{Y}_U^{\text{HT}}) = 0$, y $\delta = \delta_{min} = -0{,}1$.

	r									
	1	2	3	4	5	6	7	8	9	10
y_k	1	2	3	4	5	6	7	8	9	10
	20	19	18	17	16	15	14	13	12	11
	21	22	23	24	25	26	27	28	29	30
	40	39	38	37	36	35	34	33	32	31
	41	42	43	44	45	46	47	48	49	50
	60	59	58	57	56	55	54	53	52	51
	61	62	63	64	65	66	67	68	69	70
	80	79	78	77	76	75	74	73	72	71
	81	82	83	84	85	86	87	88	89	90
	100	99	98	97	96	95	94	93	92	91
Y_{s_r}	505	505	505	505	505	505	505	505	505	505

Tabla 2.3: Ordenación óptima (varianza mínima).

c) Un valor grande positivo de δ se asocia con una ordenación como la de la siguiente tabla 2.4.

Obtenemos $\mathbb{V}_{sys}(\widehat{Y}_U^{\text{HT}}) = 8{,}25 \cdot 10^6$, por lo que es más eficiente el diseño $srswor$. Aquí $\delta = 0{,}989$, está cerca del valor máximo $\delta_{max} = 1$.

	r									
	1	2	3	4	5	6	7	8	9	10
y_k	1	11	21	31	41	51	61	71	81	91
	2	12	22	32	42	52	62	72	82	92
	3	13	23	33	43	53	63	73	83	93
	4	14	24	34	44	54	64	74	84	94
	5	15	25	35	45	55	65	75	85	95
	6	16	26	36	46	56	66	76	86	96
	7	17	27	37	47	57	67	77	87	97
	8	18	28	38	48	58	68	78	88	98
	9	19	29	39	49	59	69	79	89	99
	10	20	30	40	50	60	70	80	90	100
Y_{s_r}	55	155	255	355	455	555	655	755	855	955

Tabla 2.4: Ordenación con un δ grande positivo.

d) Una ordenación al azar se muestra en la tabla final 2.5. Se obtiene $\mathbb{V}_{sys}(\widehat{Y}_U^{\text{HT}}) = 7{,}1766 \cdot 10^5$, que está cerca de $\mathbb{V}_{srswor}(\widehat{Y}_U^{\text{HT}})$ y $\delta = -0{,}005$. Esta ordenación fue creada mediante una permutación aleatoria de los enteros de 1 a 100. La población se puede decir que está en un orden aleatorio, por lo que, se espera que δ sea próximo a 0.

	r									
	1	2	3	4	5	6	7	8	9	10
y_k	48	14	71	13	40	59	18	45	6	53
	38	23	11	58	70	22	24	88	77	84
	10	51	98	65	93	68	25	32	99	9
	17	26	8	78	34	87	96	39	20	54
	56	79	31	86	43	66	2	62	57	5
	73	7	80	27	60	89	76	81	85	83
	3	28	33	90	55	1	21	69	61	92
	74	37	44	94	12	72	100	30	63	97
	75	41	16	82	35	95	67	50	64	29
	49	42	15	19	46	36	47	91	52	4
Y_{s_r}	443	348	407	612	488	595	476	587	584	510

Tabla 2.5: Ordenación al azar

Este ejemplo muestra el cuidado que hay que tener cuando se usa el muestreo sistemático. El precio por la simplicidad de este diseño puede conducir a una alta pérdida de eficiencia. Por otro lado, si la población está ordenada de forma favorable al muestreo sistemático, se puede obtener una gran ganancia de precisión.

2.6.2 Estimación de la varianza

Uno de los problemas del muestreo sistemático es que no existe un método directo para la estimación de varianzas a partir de una muestra sistemática. No podemos evaluar la variabilidad muestral de la estimación puntual. Existen algunas aproximaciones que no son perfectas para tratar este problema. Una de ellas es usar un estimador sesgado de la varianza; otra consiste en modificar la selección sistemática para permitir una estimación insesgada de la varianza.

Supongamos que hay una razón de peso para creer que, en una aplicación específica, el diseño sys es por lo menos tan bueno como el diseño $srswor$ en términos de precisión. Si s_r es la muestra sistemática seleccionada, podemomos considerar como estimador de la varianza la expresión correspondiente en muestreo aleatorio simple, esto es,

$$\widehat{V} = \frac{N^2(1-f)}{n} S^2_{ys_r} \tag{2.14}$$

Se han propuesto otras alternativas a $(\widehat{V})$. En Wolter 2007 hay un capítulo entero sobre la estimación de la varianza en muestreo sistemático.

Supongamos que estamos en una situación en la que el muestreo sys es más eficiente que el $srswor$, es decir,

$$\mathbb{V}_{sys}(\widehat{Y}_U^{\text{HT}}) < \mathbb{V}_{srswor}(\widehat{Y}_U^{\text{HT}})$$

lo cual se verifica si y sólo si $\delta < 0$. Entonces se puede demostrar que el estimador de la varianza dado en (2.14) *sobreestimará* la varianza $\mathbb{V}_{sys}(\widehat{Y}_U^{\text{HT}})$. Un estimador de la varianza, $\widehat{V}$, se dice que sobreestima si su esperanza es superior a la varianza para la cual $\widehat{V}$ es usado como estimador. Esto es,

$$\mathbb{E}_{sys}(\widehat{V}) > \mathbb{V}_{sys}(\widehat{Y}_U^{\text{HT}})$$

Otra solución al problema de la estimación de la varianza es modificar el diseño sys. Por ejemplo, en lugar de usar sólo un arranque aleatorio y el intervalo muestral a, podemos usar $m > 1$ arranques aleatorios y el intervalo muestral $m \cdot a$. Esto proporciona una muestra que consiste en m muestras sistemáticas, cada una de un tamaño $\frac{n}{m}$. Asumamos por simplicidad que $\frac{n}{m}$ y $a = \frac{N}{n}$ son enteros. Una muestra aleatoria simple de m enteros se selecciona de entre 1 a ma. Sean $r_1, r_2, \ldots, r_m$ los números seleccionados. La muestra viene dada por

$$s = \left\{ k : k = r_i + (j-1) \cdot m \cdot a; i = 1, \ldots, m, j = 1, \ldots, \frac{n}{m} \right\}$$

En este caso $\pi_k = \dfrac{m}{m \cdot a} = \dfrac{n}{N}$ para cada k,y las probabilidades de inclusión de segundo orden son

$$\pi_{kl} = \begin{cases} \frac{n}{N} & \text{si } k \text{ y } l \text{ pertenecen a la misma muestra} \\ \frac{n}{N} \cdot \frac{m-1}{m \cdot a - 1} & \text{si } k \text{ y } l \text{ pertenecen a muestras distintas} \end{cases}$$

2.7 Estimación por razones

Para aplicar la estimación por razones debemos disponer de los valores de dos variables, y y x, para cada elemento de la muestra. La variable x es a menudo denominada *variable auxiliar*. Considérense los totales poblacionales de estas dos variables, esto es,

$$Y_U = \sum_{k \in U} y_k; \quad X_U = \sum_{k \in U} x_k.$$

La razón poblacional es

$$R = \frac{Y_U}{X_U} = \frac{\bar{y}_U}{\bar{x}_U}$$

Supongamos que se desea estimar el porcentaje de gasto total que los hogares destinan a la alimentación en una población de N individuos. En este caso, se pretende precisamente estimar la razón poblacional R, donde la variable de estudio y es el gasto destinado a la alimentación y la variable auxiliar x es el gasto total. Para estimar R podría seleccionarse, por ejemplo, una muestra aleatoria simple s de tamaño n y obtener la información del gasto total x_k y del gasto en alimentación y_k, para todo $k \in s$. Como veremos, en la estimación por razones, el estimador de R viene dado por

$$\widehat{R} = \frac{\sum\limits_{k \in s} y_k}{\sum\limits_{k \in s} x_k} = \frac{\bar{y}_s}{\bar{x}_s} \tag{2.15}$$

2.7.1 Razones para su uso y estimadores

A continuación se exponen algunos usos de la estimación por razones con el correspondiente estimador en cada caso:

1. Cuando se desea estimar una razón. Por ejemplo, en el caso expuesto anteriormente, cuando se pretendía estimar el porcentaje de gasto en alimentación. Como hemos comentado, el estimador de R viene dado por (2.15).

2. Cuando se desea estimar el total poblacional de y pero se desconoce el tamaño N de la población. En este caso, no se puede usar el estimador $\widehat{Y}_U^{\mathrm{HT}} = N \cdot \bar{y}_s$. Sin embargo, podría usarse una variable auxiliar x relacionada con la variable de estudio y para la cual se conozca el total poblacional, ya que se sabe que

$$N = \frac{\sum\limits_{k \in U} x_k}{\bar{x}_U}$$

Luego si $\sum_{k \in U} x_k$ es conocido, podemos estimar N a través de $\dfrac{\sum_{k \in U} x_k}{\bar{x}_s}$. Por ejemplo, en una redada de peces, supóngase que se desea estimar el total de peces con una longitud superior a 12 cm. Si se desconoce N, podría pesarse toda la redada, ya que esto se mide con facilidad, y utilizar el hecho de que la variable y,

tener una longitud mayor de 12, está relacionada con x, el peso. El estimador sería

$$\widehat{Y}_U^{\text{Rat}} = \frac{\sum_{k\in U} x_k}{\bar{x}_s} \cdot \bar{y}_s = \widehat{R} \cdot X_U \tag{2.16}$$

donde $X_U = \sum_{k\in U} x_k$.

3. Cuando se desea aumentar la precisión de las estimaciones puede usarse información auxiliar que está correlacionada con la variable de estudio, de forma que al incorporar dicha información en la estimación se logra reducir la varianza. Supóngase que se desea estimar el total poblacional de y, que puede escribirse como

$$Y_U = \frac{Y_U}{X_U} \cdot X_U = R \cdot X_U$$

Así, el estimador por razón del total de y es el dado en (2.16). Suponiendo que la correlación entre x e y, $\rho_{(x,y)}$, es alta, se ganará en eficiencia al incorporar la información de x respecto a usar el diseño $srswor$, donde $\rho_{(x,y)}$ es el coeficiente de correlación de Pearson:

$$\rho_{(x,y)} = \frac{\sum_{k\in U}(x_k - \bar{x}_U)(y_k - \bar{y}_U)}{(N-1)S_{xU}S_{yU}}$$

donde S_{yU} y S_{xU} son, respectivamente, las cuasidesviaciones típicas poblacionales de la variable de estudio y la variable auxiliar.

2.7.2 Sesgo, varianza y estimador de la varianza

Hasta ahora los estimadores que se han propuesto cuentan con la propiedad de insesgadez. Sin embargo, el estimador de la razón es sesgado, tal y como se enuncia a continuación.

Teorema 15

El estimador de la razón es sesgado, con sesgo

$$\mathbb{B}(\widehat{R}) = -\frac{\mathbb{C}(\widehat{R}, \bar{x}_s)}{\bar{x}_U} \tag{2.17}$$

Demostración 15

Consideremos la covarianza $\mathbb{C}\left(\widehat{R}, \widehat{X}_U^{\text{HT}}\right)$. Como $\widehat{R} \cdot \bar{x}_s = \bar{y}_s$, la covarianza puede

escribirse como

$$\begin{aligned}\mathbb{C}(\widehat{R},\bar{x}_s) &= \mathbb{E}(\widehat{R}\cdot\bar{x}_s) - \mathbb{E}(\widehat{R})\cdot\mathbb{E}(\bar{x}_s) = \\ &= \mathbb{E}(\bar{y}_s) - \mathbb{E}(\widehat{R})\cdot\mathbb{E}(\bar{x}_s) = \\ &= \bar{y}_U - \mathbb{E}(\widehat{R})\cdot\bar{x}_U = \\ &= -\bar{x}_U\cdot[\mathbb{E}(\widehat{R}) - R]\end{aligned}$$

ya que las medias muestrales de la variable auxiliar y la de estudio, $\bar{x}_s$ y $\bar{y}_s$, son estimadores insesgados de $\bar{x}_U$ y $\bar{y}_U$, respectivamente, bajo el diseño $srswor$. Por tanto:

$$\mathbb{B}(\widehat{R}) = \mathbb{E}(\widehat{R}) - R = -\frac{\mathbb{C}\left(\widehat{R},\bar{x}_s\right)}{\bar{x}_U}.$$

Dado que $\mathbb{C}(\widehat{R},\bar{x}_s) = \rho_{(\widehat{R},\bar{x}_s)}\cdot\sigma(\widehat{R})\cdot\sigma(\bar{x}_s)$, donde $\rho_{(\widehat{R},\bar{x}_s)}$ es el coeficiente de correlación entre el estimador de la razón y la media muestral. Entonces

$$\mathbb{B}(\widehat{R}) = -\frac{\rho_{(\widehat{R},\bar{x}_s)}\cdot\sigma(\widehat{R})\cdot\sigma(\bar{x}_s)}{\bar{x}_U} = -\rho_{(\widehat{R},\bar{x}_s)}\cdot\sigma(\widehat{R})\cdot CV(\bar{x}_s)$$

donde $CV(\bar{x}_s)$ es el coeficiente de variación de $\bar{x}_s$.

Al cociente entre el sesgo y la desviación típica del estimador se le denomina la *razón de sesgo*.

$$\frac{\mathbb{B}(\widehat{R})}{\sigma(\widehat{R})} = -\rho_{(\widehat{R},\bar{x}_s)}\cdot CV(\bar{x}_s)$$

Comentario 10. Se puede observar que la razón de sesgo es cero, y por consiguiente el sesgo, cuando $\widehat{R}$ y $\bar{x}_s$ son incorreladas. ■

Hartley y Ross 1954 fijaron una cota superior para la razón del sesgo:

$$\frac{|\mathbb{B}(\widehat{R})|}{\sigma(\widehat{R})} = |\rho_{(\widehat{R},\bar{x}_s)}|\cdot CV(\bar{x}_s) \leq CV(\bar{x}_s)$$

ya que el coeficiente de correlación no puede exceder a 1.

Dado que $CV(\bar{x}_s) = \dfrac{S_{xU}}{\bar{x}_U}\cdot\sqrt{\dfrac{1}{n}-\dfrac{1}{N}}$ en muestreo aleatorio simple sin reemplazamiento, entonces

$$\frac{|\mathbb{B}(\widehat{R})|}{\sigma(\widehat{R})} \leq \sqrt{\frac{1}{n}-\frac{1}{N}}\cdot\frac{S_{xU}}{\bar{x}_U} = \sqrt{\frac{1}{n}-\frac{1}{N}}\cdot CV(X_U) \tag{2.18}$$

donde $CV(X_U)$ es el coeficiente de variación poblacional de x.

Comentario 11. Se observa de (2.18) que la razón de sesgo se aproxima a cero cuando aumenta el tamaño muestral n (que es lo que ocurre normalmente). Cuanto menor sea el coeficiente de variación de x y mayor sea el tamaño de muestra, más pequeña será la cota y, por tanto, la razón de sesgo y el sesgo. ■

Aplicaremos ahora la *técnica de linealización de Taylor* para aproximar el estimador no lineal $\widehat{R}$ mediante un estadístico que es función lineal de $\bar{y}_s$ y $\bar{x}_s$, denotado por $\widehat{R}_0$, con el que es más sencillo trabajar. Cuando la aproximación es buena, $\widehat{R}_0$ se comportará aproximadamente como $\widehat{R}$ y podremos usar la varianza de $\widehat{R}_0$ como una aproximación de $\mathbb{V}(\widehat{R})$. Asimismo podremos obtener un estimador de la varianza del estimador de la razón.

El estimador de la razón es una función de dos variables aleatorias $\bar{y}_s$ y $\bar{x}_s$, ya que

$$\widehat{R} = \frac{\bar{y}_s}{\bar{x}_s} = f(\bar{y}_s, \bar{x}_s).$$

El desarrollo de Taylor de la función f de orden 1 alrededor del punto $(\bar{y}_U, \bar{x}_U)^t$ es:

$$\widehat{R} = \widehat{R_0} = R + a_1 \cdot (\bar{y}_s - \bar{y}_U) + a_2 \cdot (\bar{x}_s - \bar{x}_U)$$

donde a_1 y a_2 son las derivadas parciales

$$a_1 = \frac{\partial f}{\partial \bar{y}_s} = \frac{1}{\bar{x}_s} \quad ; \quad a_2 = \frac{\partial f}{\partial \bar{x}_s} = -\frac{\bar{y}_s}{(\bar{x}_s)^2}.$$

Evaluándolas en el punto $(\bar{y}_U, \bar{x}_U)^t$, obtenemos

$$\begin{aligned} \left.\frac{\partial f}{\partial \bar{y}_s}\right|_{(\bar{y}_U, \bar{x}_U)^t} &= \frac{1}{\bar{x}_U} \\ \left.\frac{\partial f}{\partial \bar{x}_s}\right|_{(\bar{y}_U, \bar{x}_U)^t} &= \frac{\bar{y}_U}{(\bar{x}_U)^2} = \frac{R}{\bar{x}_U}. \end{aligned}$$

Por tanto, podemos aproximar $\widehat{R}$ como

$$\begin{aligned} \widehat{R} &= R_0 = R + \frac{1}{\bar{x}_U} \cdot (\bar{y}_s - \bar{y}_U) - \frac{R}{\bar{x}_U} \cdot (\bar{x}_s - \bar{x}_U) = \\ &= R + \frac{1}{\bar{x}_U} \cdot (\bar{y}_s - R \cdot \bar{x}_s) \end{aligned}$$

Bajo esta aproximación, se tiene $\mathbb{E}(\widehat{R}) = \mathbb{E}(\widehat{R}_0) = R$. En otras palabras, el sesgo de $\widehat{R}$, aunque no nulo, se aproxima por cero. Se podría obtener una expresión mejorada

para el sesgo extendiendo el desarrollo de Taylor para incluir también los términos de segundo orden.

Entonces, la varianza aproximada de $\widehat{R}$ se obtiene como

$$\begin{aligned} AV(\widehat{R}) = \mathbb{V}(\widehat{R}_0) &= \mathbb{V}\left[R + \frac{1}{\bar{x}_U} \cdot (\bar{y}_s - R \cdot \bar{x}_s)\right] = \\ &= \frac{1}{(\bar{x}_U)^2} \cdot \mathbb{V}\left(\bar{y}_s - R \cdot \bar{x}_s\right) = \\ &= \frac{1}{(\bar{x}_U)^2} \cdot \frac{1-f}{n} \cdot \frac{1}{N-1} \sum_{k \in U} (y_k - R x_k)^2 = \\ &= \frac{1}{(\bar{x}_U)^2} \cdot \frac{1-f}{n} \cdot \left(S_{yU}^2 + R^2 \cdot S_{xU}^2 - 2 \cdot R \cdot S_{xyU}\right) \end{aligned} \tag{2.19}$$

donde S_{yU}^2 y S_{xU}^2 son, respectivamente, las cuasivarianzas poblacionales de la variable de estudio y la variable auxiliar. S_{xyU} es la cuasicovarianza poblacional entre las variables x e y.

Un estimador de la varianza apropiado consiste en sustituir los valores poblacionales desconocidos por sus respectivos estimadores insesgados, ya que, aunque se obtiene un estimador sesgado, el sesgo es despreciable en muestras grandes.

$$\widehat{\mathbb{V}}(\widehat{R}) = \frac{1}{(\bar{x}_U)^2} \cdot \frac{1-f}{n} \cdot \frac{1}{n-1} \sum_{k \in s} (y_k - \widehat{R} \cdot x_k)^2$$

Si $\bar{x}_U$ es desconocida, puede usarse el estimador insesgado $\bar{x}_s$.

Los resultados obtenidos se resumen y enuncian en el siguiente Teorema.

Teorema 16

Usando la linealización de Taylor, el estadístico de razón $\widehat{R} = \dfrac{\bar{y}_s}{\bar{x}_s}$ se aproxima de la siguiente forma

$$\widehat{R} = \widehat{R_0} = R + \frac{1}{\bar{x}_U} \cdot (\bar{y}_s - R \cdot \bar{x}_s) \tag{2.20}$$

El estimador $\widehat{R}$ es aproximadamente insesgado para R, con varianza aproximada

$$AV(\widehat{R}) = \frac{1}{(\bar{x}_U)^2} \cdot \frac{1-f}{n} \cdot \frac{1}{N-1} \sum_{k \in U} (y_k - R \cdot x_k)^2 \tag{2.21}$$

Un estimador de la varianza es

$$\widehat{\mathbb{V}}(\widehat{R}) = \frac{1}{(\bar{x}_U)^2} \cdot \frac{1-f}{n} \cdot \frac{1}{n-1} \sum_{k \in s} (y_k - \widehat{R} \cdot x_k)^2 \tag{2.22}$$

Demostración 16

Argumentación incluida más arriba.

Supongamos que estamos interesados en estimar el total poblacional de una variable de interés y, esto es, el total $Y_U = \sum_{k \in U} y_k$, que se puede escribir como

$$Y_U = \frac{\bar{y}_U}{\bar{x}_U} \cdot X_U = R \cdot X_U$$

Siempre y cuando X_U sea una cantidad conocida, R se puede estimar mediante $\widehat{R} = \frac{\bar{y}_s}{\bar{x}_s}$ y, por tanto, el total de y se puede estimar mediante el estimador de razón

$$\widehat{Y}_U^{\text{Rat}} = \frac{\bar{y}_s}{\bar{x}_s} \sum_{k \in U} x_k = \widehat{R} \cdot X_U$$

Ahora estamos en condiciones de obtener una expresión aproximada para la varianza del estimador de razón y un estimador de la varianza.

Corolario 17

Sea Y_U es el total desconocido de una variable de estudio y. Si X_U es el total conocido de una variable auxiliar x, entonces el estimador de razón

$$\widehat{Y}_U^{\text{Rat}} = \widehat{R} \cdot X_U$$

es aproximadamente insesgado para Y_U. Su varianza aproximada viene dada por

$$AV(\widehat{Y}_U^{\text{Rat}}) = \frac{1-f}{n} \cdot \frac{N^2}{N-1} \sum_{k \in U} (y_k - R \cdot x_k)^2 \tag{2.23}$$

Un estimador de la varianza es

$$\widehat{\mathbb{V}}(\widehat{Y}_U^{\text{Rat}}) = \frac{1-f}{n} \cdot \frac{N^2}{n-1} \sum_{k \in s} (y_k - \widehat{R} \cdot x_k)^2 \tag{2.24}$$

Demostración 17

Elemental a partir del Teorema 16 y la definición de estimador de razón del total poblacional dada en (2.16).

Comentario 12. La expresión para la varianza aproximada AV dada por (2.23) es cero si $y_k - R \cdot x_k = 0$ para todos los elementos $k \in U$. Esto no ocurrirá en la práctica, pero frecuentemente podremos conseguir un conjunto de valores $y_k - R \cdot x_k$ que, aunque no sean nulos, sean pequeños. En este caso, la varianza aproximada AV también será pequeña. Por tanto, el estimador de razón resulta ser muy preciso cuando los pares poblacionales $(y_k, x_k)^t$ se encuentren dispersos cerca de una línea recta que pasa por el origen y con una pendiente R (desconocida).

■

La expresión de la varianza aproximada dada en (2.23) también se puede escribir, por lo visto en (2.19), de la siguiente forma:

$$AV(\widehat{Y}_U^{\text{Rat}}) \;=\; X_U^2 \cdot AV(\widehat{Y}_U^{\text{Rat}}) = N^2 \cdot \frac{1-f}{n} \cdot \left(S_{yU}^2 + R^2 \cdot S_{xU}^2 - 2 \cdot R \cdot S_{xyU}\right)$$

Esta representación es útil porque permite comparar de forma sencilla la varianza del estimador por razón y el estimador de Horvitz-Thompson.

El estimador de Horvitz-Thompson bajo el diseño del muestreo aleatorio simple es

$$\mathbb{V}(\widehat{Y}_U^{\text{HT}}) = N^2 \cdot \frac{1-f}{n} \cdot S_{yU}^2$$

Por tanto, el cociente de varianzas es

$$\begin{aligned}
\frac{AV(\widehat{Y}_U^{\text{Rat}})}{\mathbb{V}(\widehat{Y}_U^{\text{HT}})} &= \frac{N^2 \cdot \frac{1-f}{n} \cdot \left(S_{yU}^2 + R^2 \cdot S_{xU}^2 - 2 \cdot R \cdot S_{xyU}\right)}{N^2 \cdot \frac{1-f}{n} \cdot S_{yU}^2} = \\
&= 1 + R^2 \cdot \frac{S_{x_U}^2}{S_{yU}^2} - 2 \cdot R \cdot \frac{S_{xyU}}{S_{yU}^2} = 1 + \frac{\bar{y}_U^2}{\bar{x}_U^2} \cdot \frac{S_{x_U}^2}{S_{yU}^2} - 2 \cdot \frac{\bar{y}_U}{\bar{x}_U} \cdot \frac{S_{xyU}}{S_{yU}^2} = \\
&= 1 + \frac{cv_x^2}{cv_y^2} - 2 \cdot \frac{\bar{y}_U}{\bar{x}_U} \cdot \rho_{(x,y)} \cdot \frac{S_{xU}}{S_{yU}} = 1 + \frac{cv_x}{cv_y} \cdot \left(\frac{cv_x}{cv_y} - 2\rho_{(x,y)}\right)
\end{aligned}$$

donde los dos coeficientes de variación son $cv_x = \dfrac{S_{xU}}{\bar{x}_U}$ y $cv_y = \dfrac{S_{yU}}{\bar{y}_U}$.

El cociente de varianzas es menor que uno cuando el coeficiente de correlación entre x e y es superior a $\dfrac{1}{2} \cdot \dfrac{cv_x}{cv_y}$. Así pues, la estimación por razón es más precisa que usar una estimación basada en el estimador de Horvitz-Thompson, que no incluye la información de la variable auxiliar, cuando

$$\rho_{(x,y)} > \frac{1}{2} \cdot \frac{cv_x}{cv_y}$$

Si se dispone de una población con coeficientes de variación de x e y aproximadamente iguales, entonces es mejor usar la estimación por razones cuando la correlación entre las variables es mayor que $\frac{1}{2}$.

Ejemplo 21. Se dispone de una muestra aleatoria simple de 50 hogares de una población constituida por 1000 hogares. Se conoce el ingreso semanal (x) y el gasto semanal en alimentación (y) de los 50 hogares. Los datos de la muestra se resumen a continuación:

$$\sum_{k\in s} y_k = 44210;\ \sum_{k\in s} x_k = 132000;$$

$$\sum_{k\in s} y_k^2 = 40630000;\ \sum_{k\in s} x_k^2 = 357920000;\ \sum_{k\in s} x_k y_k = 119560000$$

Supongamos que estamos interesados en estimar el porcentaje del ingreso que se ha destinado a la alimentación. Dado que el ingreso es variable en función del hogar seleccionado, se usará el estimador de la razón

$$\widehat{R} = \frac{\sum\limits_{k\in s} y_k}{\sum\limits_{k\in s} x_k} = \frac{44210}{132000} \approx 0{,}3349$$

El porcentaje del ingreso destinado a la alimentación estimado es de aproximadamente $33{,}49\,\%$.

La estimación del error de muestreo (desviación típica de la varianza aproximada) es

$$\widehat{\sigma}(\widehat{R}) = \sqrt{\frac{1}{(\bar{x}_s)^2}\frac{1-f}{n}\frac{1}{n-1}\left(\sum_{k\in s} y_k^2 - 2\widehat{R}\sum_{k\in s} x_k y_k + \widehat{R}^2\sum_{k\in s} x_k^2\right)} \approx 0{,}0062$$

Por tanto, el coeficiente de variación estimado es aproximadamente de $1{,}85\,\%$.

Si sabemos que el total de ingresos semanal de las familias de la población es de $2{,}5\cdot 10^6$, entonces podemos dar una estimación por razón del gasto total en alimentación, usando como variable auxiliar el ingreso semanal x:

$$\widehat{Y}_U^{Rat} = \widehat{R}\cdot X_U = \frac{44210}{132000}\cdot 2{,}5\cdot 10^6 \approx 837311$$

En este caso, dado que conocemos el total de la variable auxiliar se ha podido calcular la estimación por razón. Usaremos el total X_U para obtener la estimación del error de muestreo y del coeficiente de variación. El coeficiente de variación estimado es aproximadamente del $1{,}96\,\%$.

■

Bibliografía

Cochran, W. G. (1977). *Sampling Techniques*. 3rd. Wiley.

Fan, C.T., M.E. Muller e I. Rezucha (1962). "Development of sampling plans by using sequential (item by item) techniques and digital computers". En: *Journal of the American Statistical Association* 57, págs. 387-402.

Hartley, H.O. y A. Ross (1954). "Unbiased ratio estimators". En: *Nature* 174, págs. 270-271.

Lohr, S. (2010). *Sampling: Design and Analysis*. 2nd. Duxbury Press.

McLeod, A.I. y D.R. Bellhouse (1983). "A convenient algorithm for drawing a simple random sample". En: *Applied Statistics* 32, págs. 182-184.

Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.

Wolter, K. (2007). *Introduction to variance estimation*. 2nd. New York: Springer.

Tema 3

Muestreo estratificado y sus ventajas. Teoría del muestreo estratificado: estimadores y errores de muestreo. Asignación de las observaciones en los estratos.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía:

W. G. Cochran (1977). *Sampling Techniques*. 3rd. Wiley

S. Lohr (2010). *Sampling: Design and Analysis*. 2nd. Duxbury Press

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

3.1 Muestreo estratificado y sus ventajas

El muestreo estratificado es un diseño de muestreo probabilístico en el que se divide la población objeto de estudio en diferentes subpoblaciones no superpuestas, denominadas estratos. La muestra estratificada se obtiene tras seleccionar una muestra en cada estrato. La selección de la muestra en cada estrato se realiza de forma independiente.

El muestreo estratificado es un método poderoso y flexible que es ampliamente usado en la práctica. En las encuestas económicas (dirigidas a las empresas) es el tipo de diseño que se suele utilizar. Por ejemplo, en *Índices de Comercio al por Menor*[1] y la *Estadística Estructural de Empresas: Sectores Industrial, Comercio, Servicios*[2]. Algunas variables que se usan para realizar la partición de la población son, entre otras, la actividad económica principal en función de agrupaciones de CNAE[3], el tamaño en función del número

[1] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176900&menu=metodologia&idp=1254735576799

[2] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736143952&menu=metodologia&idp=1254735576550

[3] https://ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177032&menu=ultiDatos&idp=1254735976614

de asalariados y la Comunidad Autónoma donde se ubica la sede, información que se encuentra disponible en el marco usado para las encuestas económicas, el *Directorio Central de Empresas (DIRCE)*[4]. El muestreo estratificado también es utilizado en encuestas a hogares. Por ejemplo, en la *Encuesta de Población Activa*[5] se aplica un diseño bietápico, con estratificación de las unidades de primera etapa (las secciones censales) y sin submuestreo en las unidades secundarias (viviendas familiares principales). Las secciones censales, división territorial con fines estadísticos y administrativos, son áreas geográficas perfectamente delimitadas cuyo tamaño de población viene regulado por la Ley Orgánica del Régimen Electoral General[6].

A continuación se detallan las razones por las que el muestreo estratificado es tan popular:

1. Cuando se requieren estimaciones de precisión separadas para determinadas subpoblaciones (dominios de estudio) y la pertenencia al dominio de cada elemento de la población aparece definido en el marco muestral, entonces cada dominio de estudio puede ser tratado como un estrato separado y obtener así una muestra probabilística adecuada de cada estrato.
2. La tasa de falta de respuesta, los errores de medida y la información auxiliar pueden diferir considerablemente de una subpoblación a otra. Si ocurre esto, parece adecuado pensar que la elección de un diseño de muestreo y un estimador quizá no debería ser el mismo en todos los estratos y debería elegirse el más conveniente en cada subpolación para así mejorar la eficiencia en la estimación.
3. La conveniencia administrativa puede imponer el uso de la estratificación. Por ejemplo, si la agencia encargada de la encuesta dispone de oficinas de campo en una serie de distritos geográficos, cada una de las cuales puede supervisar la encuesta para una parte de la población. De esta forma, podría reducirse el coste global de la encuesta. En este caso, parece natural tomar cada distrito como un estrato.
4. Cuando la estratificación permite dividir la población hererogénea en subpoblaciones internamente homogéneas. Si cada estrato es homogéneo internamente, en el sentido de que las medidas de las características de estudio varían poco de una unidad a otra, se puede obtener una estimación precisa del parámetro poblacional de estudio para cualquier estrato a partir de una pequeña muestra en ese estrato. Esto producirá una ganancia en precisión en las estimaciones finales de los parámetros poblacionales de interés.

El muestreo estratificado tiene la ventaja de producir muestras más representativas de la población respecto de la variable o variables de estratificación, debido a que se asegura

[4] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736160707&menu=ultiDatos&idp=1254735576550

[5] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176918&menu=metodologia&idp=1254735976595

[6] https://www.boe.es/eli/es/lo/1985/06/19/5

que todos los elementos de cada subpoblación tienen representación en la muestra. Por tanto, la muestra se extiende sobre todos los estratos definidos. Si la estratificación se realiza de forma correcta, tal que suponga la formación de grupos internamente homogéneos y heterogéneos entre ellos en relación a las características de estudio, entonces se produce una ganancia en precisión en la estimación de dichas características para la población completa. Por otra parte, el muestreo estratificado permite aplicar distintos diseños de muestreo en cada subpoblación, así como un estimador diferente, por lo que presenta la ventaja de poder usar el más apropiado para cada estrato. Por último, si el muestreo estratificado se organiza y administra convenientemente, este tipo de diseño puede suponer la reducción del coste global de la encuesta.

El principal inconveniente del muestreo estratificado es que en el marco debe estar disponible información auxiliar para realizar la estratificación de la población para todas las unidades. Además, esta información auxiliar también debe de estar depurada y reflejar lo más fielmente la realidad, ya que, en caso contrario, la estimación se verá afectada y será necesario realizar un proceso de postestratificación[7] para poder obtener estimaciones acuradas.

Consideraciones técnicas para aplicar un muestreo estratificado

Si el estadístico encargado de la investigación decide realizar un muestreo estratificado, debe tener en cuenta las siguientes cuestiones:

i. Construcción de los estratos

   (a) El estadístico debe elegir, en caso de que sea posible, la característica o características que va a usar para dividir la población en estratos. Las características escogidas se denominan variables de estratificación. Por ejemplo, podría considerarse la edad y el sexo como variables de estratificación o bien estratificar por grupos ocupacionales.

   (b) Decidir cuántos estratos debería haber. Por ejemplo, si se usa como variable de estratificación la edad, se debe determinar el número de grupos de edad.

   (c) Determinar los límites de cada estrato a partir de las variables de estratificación elegidas. Por ejemplo, si se usan grupos de edad, se debe decidir qué intervalos de edad se usarán para configurar los estratos.

   En ocasiones la construcción de los estratos puede venir fijada por los dominios de difusión, que a su vez serán especificados por las necesidades de información internas (de otros ministerios o subdirecciones dentro del INE) y externas (Reglamentos/Directivas europeos, etc.).

ii. Elección del diseño de muestreo y el estimador dentro de cada estrato

   (a) Especificación del diseño de muestreo y el tamaño de muestra en cada estrato. A menudo se suele aplicar el mismo tipo de diseño en todos los estratos.

[7] véase la sección 7.6 del Särndal, Swensson y Wretman 1992

(b) Especificación del estimador a usar en cada estrato. Se suele aplicar también el mismo en todos los estratos.

3.2 Teoría del muestreo estratificado

3.2.1 Notación

Considérese una población constituida por N elementos $\{u_1, ..., u_N\}$, que se denota por $U = \{1, \ldots, N\}$ y supóngase que el parámetro de interés es el total poblacional de la *variable de estudio* y. Se realiza una estratificación de la población U, es decir, se realiza una partición de la población en H subpoblaciones, denominados estratos, denotados por $U_1, U_2, \ldots, U_H$. Por tanto, U_h contiene los elementos de la población que pertenecen al estrato h, $\forall h = 1, \ldots, H$.

El muestreo estratificado consiste en seleccionar, en cada estrato h, una muestra probabilística de U_h de acuerdo al diseño de muestreo establecido para ese estrato, $p_h(\cdot)$, $h = 1, \ldots, H$. El resultado de la muestra total será la unión de todas las muestras seleccionadas en cada estrato, esto es,

$$s = s_1 \cup s_2 \cup \cdots \cup s_H$$

El número de unidades seleccionadas de U_h se denomina el tamaño muestral del estrato h y se denota por n_h. Así, el tamaño muestral total, n, puede representarse como

$$n = \sum_{h=1}^{H} n_h$$

Como consecuencia de la independencia en la selección de la muestra en cada estrato, el diseño del muestreo estratificado es

$$p(s) = p_1(s_1)p_2(s_2)\cdots p_H(s_H)$$

Se denota diseño st.

El número de elementos en el estrato h se supone conocido y se denota por N_h. Dado que los estratos son una partición de la población total, se tiene:

$$N = \sum_{h=1}^{H} N_h$$

Por otra parte, el total poblacional de la variable y puede representarse como sigue:

$$Y_U = \sum_{k \in U} y_k = \sum_{h=1}^{H} Y_{U_h} = \sum_{h=1}^{H} N_h \cdot \bar{y}_{U_h}$$

donde $Y_{U_h} = \sum\limits_{k \in U_h} y_k$ es el total del estrato h y $\bar{y}_{U_h} = \frac{1}{N_h} \sum\limits_{k \in U_h} y_k$ la media del estrato h.

La media poblacional de la variable y puede representarse como

$$\bar{y}_U = \sum_{h=1}^{H} W_h \cdot \bar{y}_{U_h}$$

donde $W_h = \frac{N_h}{N}$ es el peso poblacional del estrato h.

El número de elementos de la población U que pertenecen a un determinado dominio d, es decir, que pertenecen a dicha clase, se puede expresar como

$$N_d = \sum_{h=1}^{H} \sum_{k \in U_h} z_{dk}$$

donde

$$z_{dk} = \begin{cases} 1, & \text{si } k \in U_d \\ 0, & \text{si } k \notin U_d \end{cases} \quad \text{para todo } k \in 1, \ldots, N.$$

Esto es, una variable que toma únicamente dos valores, 0 y 1, en función de si el elemento pertenece al dominio o no. Sea $N_{dh} = \sum_{k \in U_h} z_{dk}$ el número de elementos del estrato U_h que pertenecen al dominio d y $P_{dh} = \frac{1}{N_h} \sum_{k \in U_h} z_{dk} = \frac{N_{dh}}{N_h}$ la proporción de elementos de dicho estrato que pertenecen al dominio, entonces N_d se puede escribir como

$$N_d = \sum_{h=1}^{H} N_{dh} = \sum_{h=1}^{H} N_h \cdot P_{dh}$$

Por último, la proporción poblacional de los elementos de U que pertenecen a un determinado dominio d se representa como

$$P_d = \frac{N_d}{N} = \sum_{h=1}^{H} W_h \cdot P_{dh}$$

3.2.2 Estimadores y errores de muestreo

Supongamos que se considera usar el estimador de Horvitz-Thompson en todos los estratos, mientras que los diseños de muestreo aplicados sí podrían ser diferentes de una subpoblación a otra. A continuación se obtendrá el estimador de Horvitz-Thompson bajo el diseño del muestreo estratificado, así como la varianza del estimador y un estimador de dicha varianza.

Proposición 18

Bajo muestreo estratificado, el estimador de Horvitz-Thompson (HT) del total poblacional $Y_U = \sum_{k \in U} y_k$ es

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{h=1}^{H} \widehat{Y}_{U_h}^{\mathrm{HT}} \tag{3.1}$$

donde $\widehat{Y}_{U_h}^{\mathrm{HT}}$ es el estimador HT del total poblacional del estrato h, Y_{U_h}.

La varianza del estimador es

$$\mathbb{V}_{st}\left(\widehat{Y}_U^{\mathrm{HT}}\right) = \sum_{h=1}^{H} \mathbb{V}_{p_h}\left(\widehat{Y}_{U_h}^{\mathrm{HT}}\right) \tag{3.2}$$

donde $\mathbb{V}_{p_h}\left(\widehat{Y}_{U_h}^{\mathrm{HT}}\right)$ es la varianza del estimador $\widehat{Y}_{U_h}^{\mathrm{HT}}$ bajo el diseño de muestreo p_h utilizado en el estrato h. Por definición, el error de muestreo del estimador, $\sigma_{st}\left(\widehat{Y}_U^{\mathrm{HT}}\right)$, es la raíz cuadrada de la varianza del estimador.

Un estimador insesgado de la varianza viene dado por

$$\widehat{\mathbb{V}}_{st}\left(\widehat{Y}_U^{\mathrm{HT}}\right) = \sum_{h=1}^{H} \widehat{\mathbb{V}}_{p_h}\left(\widehat{Y}_{U_h}^{\mathrm{HT}}\right) \tag{3.3}$$

suponiendo que existe un estimador insesgado de $\mathbb{V}_{p_h}\left(\widehat{Y}_{U_h}^{\mathrm{HT}}\right)$, que es $\widehat{\mathbb{V}}_{p_h}\left(\widehat{Y}_{U_h}^{\mathrm{HT}}\right)$. Asimismo, un estimador del error de muestreo se calcula como la raíz cuadrada de $\widehat{\mathbb{V}}_{st}\left(\widehat{Y}_U^{\mathrm{HT}}\right)$.

Demostración 18

Por definición de muestreo estratificado, es fácil ver que para todo $k \in U_h$, se tiene

$$\pi_k = \mathbb{P}(k \in s) = \mathbb{P}(k \in s_h)$$

Por tanto, el estimador HT del total poblacional es

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{h=1}^{H} \sum_{k \in s_h} \frac{y_k}{\pi_k} = \sum_{h=1}^{H} \widehat{Y}_{U_h}^{\mathrm{HT}}$$

Dado que la muestra estratificada se obtiene, por definición, seleccionando una muestra en cada estrato de forma independiente, las variables aleatorias $\widehat{Y}_{U_h}^{\mathrm{HT}}$ son independientes y, de esta forma, se obtienen de forma sencilla los resultados (3.2) y (3.3).

Comentario 13. El muestreo estratificado permite aplicar diferentes diseños de muestreo y estimadores en los distintos estratos. Sin embargo, es habitual usar el mismo diseño y el mismo estimador en todos. Un de los diseños más populares es aplicar la selección de una muestra aleatoria simple en cada estrato. Denotaremos a este diseño por *strs* y se denomina *muestreo aleatorio estratificado*.

■

Proposición 19

Bajo el diseño del muestreo aleatorio estratificado, el estimador HT del total poblacional es

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{h=1}^{H} N_h \cdot \bar{y}_{s_h}$$

donde $\bar{y}_{s_h} = \sum_{k \in s_h} \frac{y_k}{n_h}$ es la media muestral del estrato h.

La varianza del estimador HT es

$$\mathbb{V}_{strs}\left(\widehat{Y}_U^{\mathrm{HT}}\right) = \sum_{h=1}^{H} N_h^2 \cdot \frac{1-f_h}{n_h} \cdot S_{yU_h}^2 \tag{3.4}$$

donde $f_h = \frac{n_h}{N_h}$ es la fracción de muestreo en el estrato h y

$$S_{yU_h}^2 = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \bar{y}_{U_h})^2$$

es la cuasivarianza del estrato h y $\bar{y}_{U_h} = \sum_{k \in U_h} \frac{y_k}{N_h}$ es la media de la población U_h.

Un estimador insesgado de la varianza viene dado por

$$\widehat{\mathbb{V}}_{strs}\left(\widehat{Y}_U^{\mathrm{HT}}\right) = \sum_{h=1}^{H} N_h^2 \cdot \frac{1-f_h}{n_h} \cdot S_{ys_h}^2$$

donde

$$S_{ys_h}^2 = \frac{1}{n_h - 1} \sum_{k \in s_h} (y_k - \bar{y}_{s_h})^2$$

es la cuasivarianza muestral del estrato h.

Demostración 19

Recordemos los resultados obtenidos en el Tema 2 para el diseño *srswor*:

- Estimador HT:

$$\widehat{Y}_{U_h}^{\mathrm{HT}} = N_h \sum_{k \in s_h} \frac{y_k}{n_h} = N_h \cdot \bar{y}_{s_h}$$

- Varianza:

$$\mathbb{V}_{srswor}\left(\widehat{Y}_{U_h}^{\mathrm{HT}}\right) = N_h^2 \cdot \frac{1-f_h}{n_h} \cdot S_{yU_h}^2$$

- Estimador de la varianza:

$$\widehat{\mathbb{V}}_{srswor}\left(\widehat{Y}_{U_h}^{\mathrm{HT}}\right) = N_h^2 \cdot \frac{1-f_h}{n_h} \cdot S_{ys_h}^2$$

Utilizando estos resultados en las expresiones (3.1), (3.2) y (3.3), se llega a lo que se pretendía demostrar.

Ejemplo 22. Se dispone de una población constituida por 284 municipios (Apéndice B de Särndal, Swensson y Wretman 1992) de la que se desea estimar el total de escaños del Partido Socialdemócrata en el consejo municipal (variable y). La población se ha dividido en cuatro estratos en función de la variable de estratificación número de escaños. Se realiza un muestreo estratificado donde en cada estrato se selecciona una muestra aleatoria simple independiente de municipios. Los resultados obtenidos son los siguientes:

Número de escaños	N_h	n_h	$\sum_{k\in s_h} y_k$	$\sum_{k\in s_h} y_k^2$
$31-40$	44	5	89	1647
$41-50$	168	21	441	9735
$51-70$	56	10	280	8294
71 o más	16	4	152	5794

Bajo el diseño $strs$, el estimador de HT del total de escaños del Partido Socialdemócrata es

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{h=1}^{H} N_h \cdot \bar{y}_{s_h} = 44 \cdot \frac{89}{5} + 168 \cdot \frac{441}{21} + 56 \cdot \frac{280}{10} + 16 \cdot \frac{152}{4} = 6487{,}20$$

Calculemos ahora las cuasivarianzas muestrales de cada estrato:

$$S_{ys_h}^2 = \frac{1}{n_h - 1}\left(\sum_{k\in s_h} y_k^2 - n_h \cdot \bar{y}_{s_h}^2\right)$$

$$S_{ys_1}^2 = 15{,}7; \quad S_{ys_2}^2 = 23{,}7; \quad S_{ys_3}^2 = \frac{454}{9}; \quad S_{ys_4}^2 = 6$$

La estimación de la varianza es

$$\begin{aligned}\widehat{\mathbb{V}}_{strs}\left(\widehat{Y}_U^{\mathrm{HT}}\right) &= \sum_{h=1}^{H} N_h^2 \cdot \frac{1-f_h}{n_h} \cdot S_{ys_h}^2 = 44^2 \cdot \frac{1-5/44}{5} \cdot 15{,}7 + \\ &+ 168^2 \cdot \frac{1-21/168}{21} \cdot 23{,}7 + 56^2 \cdot \frac{1-10/56}{10} \cdot \frac{454}{9} +\end{aligned}$$

$$+ \quad 16^2 \cdot \frac{1-4/16}{4} \cdot 6 \approx 46541{,}93$$

La estimación del error relativo o coeficiente de variación del estimador es aproximadamente de 3.33 %.

$$cve(\widehat{Y}_U^{\mathrm{HT}}) = \frac{\widehat{\sigma}_{strs}(\widehat{Y}_U^{\mathrm{HT}})}{\widehat{Y}_U^{\mathrm{HT}}} \approx 0{,}0333$$

■

Estimación de la media poblacional bajo el diseño strs

Un estimador insesgado de la media poblacional $\bar{y}_U$ se obtiene diviviendo el estimador de Horvitz-Thompson para el caso del total por el tamaño poblacional N:

$$\widehat{\bar{y}}_U^{\mathrm{HT}} = \frac{1}{N} \cdot \widehat{Y}_U^{\mathrm{HT}} = \frac{1}{N} \sum_{h=1}^{H} N_h \cdot \bar{y}_{s_h} = \sum_{h=1}^{H} W_h \cdot \bar{y}_{s_h}$$

La varianza y el estimador de la varianza pueden obtenerse de forma sencilla:

$$\begin{aligned} \mathbb{V}_{strs}\left(\widehat{\bar{y}}_U^{\mathrm{HT}}\right) &= \frac{1}{N^2} \cdot \mathbb{V}\left[\widehat{Y}_U^{\mathrm{HT}}\right] = \sum_{h=1}^{H} W_h^2 \cdot \frac{1-f_h}{n_h} \cdot S_{yU_h}^2 \\ \widehat{\mathbb{V}}_{strs}\left(\widehat{\bar{y}}_U^{\mathrm{HT}}\right) &= \sum_{h=1}^{H} W_h^2 \cdot \frac{1-f_h}{n_h} \cdot S_{ys_h}^2 \end{aligned}$$

Estimación de la proporción y el número de elementos que pertenecen a un dominio bajo el diseño strs

Sea $n_{dh} = \sum_{k \in s_h} z_{dk}$ el número de elementos de la muestra extraída en el estrato h que pertenecen al dominio y $p_{dh} = \frac{n_{dh}}{n_h}$ la proporción de elementos de la muestra extraída en el estrato h que pertenecen al dominio. Dado que la variable z_d es un caso particular de una característica que toma únicamente los valores 0 y 1, entonces el estimador HT es, aplicando lo estudiado,

$$\widehat{N}_d = \sum_{h=1}^{H} N_h \cdot p_{dh}$$

Tomando $Q_{dh} = 1 - P_{dh}$, la varianza del estimador viene dada por:

$$\begin{aligned} \mathbb{V}_{strs}\left(\widehat{N}_d\right) &= \sum_{h=1}^{H} N_h^2 \cdot \frac{1-f_h}{n_h} \cdot S_{z_d U_h}^2 = \sum_{h=1}^{H} N_h^2 \cdot \frac{N_h - n_h}{N_h - 1} \cdot \frac{\sigma_{z_d U_h}^2}{n_h} = \\ &= \sum_{h=1}^{H} N_h^2 \cdot \frac{N_h - n_h}{N_h - 1} \cdot \frac{P_{dh} \cdot Q_{dh}}{n_h} \end{aligned}$$

ya que

$$\sigma^2_{z_d U_h} = P_{dh} \cdot (1 - P_{dh}) = P_{dh} \cdot Q_{dh}$$

Un estimador insesgado de la varianza es:

$$\widehat{\mathbb{V}}_{strs}\left(\widehat{N}_d\right) = \sum_{h=1}^{H} N_h^2 \cdot (1 - f_h) \cdot \frac{S^2_{z_d s_h}}{n_h} = \sum_{h=1}^{H} N_h^2 \cdot (1 - f_h) \cdot \frac{p_{dh} \cdot q_{dh}}{n_h - 1}$$

ya que $S^2_{z_d s_h} = \frac{n_h}{n_h - 1} \cdot p_{dh} \cdot q_{dh}$, donde $q_{dh} = 1 - p_{dh}$

De forma análoga, para el caso de la proporción de individuos que pertenecen al dominio, un estimador insesgado para P_d es

$$\widehat{P}_d = \sum_{h=1}^{H} W_h \cdot p_{dh}$$

La varianza y el estimador de la varianza se pueden calcular a partir de las expresiones vistas para el estimador de N_d:

$$\begin{aligned}
\mathbb{V}_{strs}\left(\widehat{P}_d\right) &= \frac{1}{N^2} \cdot \mathbb{V}_{strs}\left(\widehat{N}_d\right) = \sum_{h=1}^{H} W_h^2 \cdot \frac{N_h - n_h}{N_h - 1} \cdot \frac{P_{dh} \cdot Q_{dh}}{n_h} \\
\widehat{\mathbb{V}}_{strs}\left(\widehat{P}_d\right) &= \frac{1}{N^2} \cdot \widehat{\mathbb{V}}_{strs}\left(\widehat{N}_d\right) = \sum_{h=1}^{H} W_h^2 \cdot (1 - f_h) \cdot \frac{p_{dh} \cdot q_{dh}}{n_h - 1}
\end{aligned}$$

3.3 Asignación de las observaciones en los estratos

Consideremos una población dividida en H estratos de la cual se desea estimar el total poblacional de la variable y. Se decide realizar un muestreo estratificado para el cual ya se han decidido los diseños a aplicar en cada estrato. El estimador a utilizar es el de Horvitz-Thompson.

Antes de seleccionar la muestra en cada estrato, el estadístico debe determinar cuántas unidades muestrales n_h extraerá de cada estrato. El problema de determinar n_h se conoce como el problema de la *afijación* o *asignación de la muestra*.

Definición 19

El concepto de *asignación de la muestra* se define como el reparto o distribución del tamaño de la muestra total, n, entre los distintos estratos.

3.3.1 Asignación proporcional

La asignación proporcional se define como

$$n_h = n \cdot \frac{N_h}{N}, \quad h = 1, \ldots, H \tag{3.5}$$

Esto es, el número de unidades muestrales en cada estrato es proporcional al tamaño del estrato. Por definición, la probabilidad de que un elemento de la población aparezca en la muestra s es igual a la probabilidad de que dicho elemento aparezca en la muestra s_h:

$$\pi_k = \mathbb{P}(k \in s) = \mathbb{P}(k \in s_h) = \frac{n_h}{N_h}$$

Si se utiliza la asignación proporcional, entonces $\pi_k = \frac{n}{N}$, que es equivalente a la probabilidad de que el elemento k pertenezca a una muestra aleatoria simple, $\forall k = 1, \ldots, N$. Sin embargo, una "mala" muestra, poco representativa de la población, por ejemplo con todos los elementos pertenecientes a un único estrato, no podría ocurrir en una muestra estratificada con asignación proporcional.

Asimismo, cada elemento de la muestra tiene el mismo *factor de expansión* (o *peso de muestreo*), por lo que todos los elementos representan el mismo número de unidades de la población. Cuando ocurre esto se dice que la muestra es *autoponderada*.

Definición 20

Un diseño de muestreo $p(s)$ se dice que produce muestras *autoponderadas* respecto a un estimador del parámetro poblacional θ si dicho estimador ($\widehat{\theta}$) se puede expresar en función del total muestral de la siguiente forma:

$$\widehat{\theta} = K \sum_{k \in s} y_k$$

donde K es una constante y se denomina *factor de expansión*.

El muestreo aleatorio estratificado con asignación proporcional produce muestras *autoponderadas* respecto al estimador de Horvitz-Thompson. En efecto, para el caso del total poblacional, se tiene:

$$\widehat{Y}_U^{\text{HT}} = \sum_{h=1}^{H} N_h \cdot \bar{y}_{s_h} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in s_h} y_k = \frac{N}{n} \sum_{h=1}^{H} \sum_{k \in s_h} y_k = \frac{N}{n} \sum_{k \in s} y_k$$

Como se puede observar, todas las observaciones aparecen multiplicadas por un mismo factor $\frac{N}{n}$.

La varianza bajo muestreo aleatorio estratificado con asignación proporcional es:

$$\begin{aligned}\mathbb{V}_{strs,prop}\left(\widehat{Y}_U^{\text{HT}}\right) &= \sum_{h=1}^{H} N_h^2 \cdot (1-f_h) \cdot \frac{S_{yU_h}^2}{n_h} = N^2 \sum_{h=1}^{H} \frac{W_h^2 \cdot S_{yU_h}^2}{n \cdot \frac{N_h}{N}} - N \sum_{h=1}^{H} W_h \cdot S_{yU_h}^2 = \\ &= N^2 \cdot \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} W_h \cdot S_{yU_h}^2\end{aligned}$$

Comentario 14. Obsérvese que para un tamaño de muestra n fijado, la asignación proporcional puede ser siempre calculada cuando los tamaños poblacionales de los estratos N_h sean conocidos.

■

3.3.2 Asignación óptima

Supongamos que el coste total de la encuesta puede ser expresado como

$$C = c_0 + \sum_{h=1}^{H} n_h \cdot c_h$$

donde c_0 es el coste fijo y c_h es el coste de observar un elemento del estrato h.

Definición 21

El problema de la asignación óptima de la muestra se puede formular de dos formas distintas:

1. Determinación de los tamaños muestrales n_h que minimicen la varianza del estimador V sujeto a la restricción de un coste fijo C.

$$\begin{cases} \textit{minimizar } V = \sum_{h=1}^{H} \frac{N_h^2 \cdot S_{yU_h}^2}{n_h} - \sum_{h=1}^{H} N_h \cdot S_{yU_h}^2 \\ \textit{sujeto a } C = c_0 + \sum_{h=1}^{H} n_h \cdot c_h \end{cases} \tag{3.6}$$

2. Determinación de los tamaños muestrales n_h que minimicen el coste total de la encuesta C sujeto a una precisión fijada V.

$$\begin{cases} \textit{minimizar } C = c_0 + \sum_{h=1}^{H} n_h \cdot c_h \\ \textit{sujeto a } V = \sum_{h=1}^{H} \frac{N_h^2 \cdot S_{yU_h}^2}{n_h} - \sum_{h=1}^{H} N_h \cdot S_{yU_h}^2 \end{cases} \tag{3.7}$$

Proposición 20

En el diseño de muestreo estratificado, la asignación óptima de la muestra, suponiendo la función de costes $C = c_0 + \sum_{h=1}^{H} n_h \cdot c_h$, es tal que n_h es proporcional a $\frac{N_h \cdot s_h}{\sqrt{c_h}}$, esto es:

$$n_h \propto \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}$$

Demostración 20

El problema de optimización que se plantea es el especificado en (3.6) o en (3.7). Se resuelve de forma análoga en los dos casos.

Sea $V^* = \sum_{h=1}^{H} \frac{N_h^2 \cdot S_{yU_h}^2}{n_h}$ y $C^* = C - c_0$, entonces el problema es equivalente a minimizar el producto

$$V^* \cdot C^* = \left(\sum_{h=1}^{H} \frac{N_h^2 \cdot S_{yU_h}^2}{n_h} \right) \cdot \left(\sum_{h=1}^{H} n_h \cdot c_h \right)$$

De la desigualdad de Cauchy se tiene

$$\left(\sum a_h^2 \right) \cdot \left(\sum b_h^2 \right) \geq \left(\sum a_h \cdot b_h \right)^2 .$$

La igualdad se da si y solo si $\frac{b_h}{a_h}$ es constante para cada h.

Tomando $a_h = \left(\frac{N_h^2 \cdot S_{yU_h}^2}{n_h} \right)^{1/2}$ y $b_h = (n_h \cdot c_h)^{1/2}$ se llega a

$$V^* \cdot C^* \geq \left[\sum_{h=1}^{H} N_h \cdot S_{yU_h} \cdot \sqrt{c_h} \right]^2$$

La igualdad se da cuando $\left(\frac{n_h \cdot c_h}{\frac{N_h^2 \cdot S_{yU_h}^2}{n_h}} \right)^{1/2} \equiv constante$ o, de forma equivalente, se puede decir que

$$n_h \propto \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}$$

Si se plantea el problema de **minimizar la varianza del estimador HT para un coste C fijado**, esto es, el problema especificado en (3.6), entonces dado que $n_h = \alpha \cdot \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}$

para una constante α y $C = c_0 + \sum_{h=1}^{H} n_h \cdot c_h$, se tiene:

$$C = C_0 + \sum_{h=1}^{H} n_h \cdot c_h = c_0 + \sum_{h=1}^{H} \alpha \cdot \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}} \cdot c_h = c_0 + \alpha \sum_{h=1}^{H} N_h \cdot S_{yU_h} \cdot \sqrt{c_h}$$

Por tanto

$$\alpha = \frac{C - c_0}{\sum_{h=1}^{H} N_h \cdot S_{yU_h} \cdot \sqrt{c_h}}$$

El tamaño del estrato h viende dado por la expresión

$$n_h = \frac{C - c_0}{\sum_{h=1}^{H} N_h \cdot S_{yU_h} \cdot \sqrt{c_h}} \cdot \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}, \quad h = 1, \ldots, H.$$

Usando lo anterior, la varianza mínima es

$$V_{opt} = \frac{1}{C - c_0} \cdot \left[\sum_{h=1}^{H} N_h \cdot S_{yU_h} \cdot \sqrt{c_h} \right]^2 + B$$

donde $B = -\sum_{h=1}^{H} N_h \cdot S^2_{yU_h}$.

Si se plantea el problema de **minimizar el coste C fijado para una precisión dada V**, esto es, el problema especificado en (3.7), entonces dado que $n_h = \alpha \cdot \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}$ y $V = \sum_{h=1}^{H} \frac{N_h^2 \cdot S^2_{yU_h}}{n_h} + B$, se tiene:

$$V = \sum_{h=1}^{H} \frac{N_h^2 \cdot S^2_{yU_h}}{\alpha \cdot \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}} + B = \frac{1}{\alpha} \sum_{h=1}^{H} N_h \cdot S_{yU_h} \cdot \sqrt{c_h} + B$$

Por tanto

$$\alpha = \frac{\sum_{h=1}^{H} N_h \cdot S_{yU_h} \cdot \sqrt{c_h}}{V - B}$$

El tamaño del estrato h viene dado por

$$n_h = \frac{\sum_{h=1}^{H} N_h \cdot S_{yU_h} \cdot \sqrt{c_h}}{V - B} \cdot \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}, \quad h = 1, \ldots, H.$$

El coste mínimo es

$$C_{opt} = c_0 + \frac{1}{V - B} \cdot \left[\sum_{h=1}^{H} N_h \cdot S_{yU_h} \cdot \sqrt{c_h} \right]^2$$

Comentario 15. Sea el tamaño de la muestra total un valor fijo dado, n. Dado que $n = \sum_{h=1}^{H} n_h$, es posible expresar la asignación óptima para los problemas de optimización planteados en (3.6) y (3.7) como:

$$n_h = n \cdot \frac{\frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}}{\sum_{h=1}^{H} \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}} \tag{3.8}$$

En efecto, como $n_h = \alpha \cdot \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}$, se tiene:

$$n = \sum_{h=1}^{H} \alpha \cdot \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}; \quad \alpha = \frac{n}{\sum_{h=1}^{H} \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}}; \quad n_h = \frac{n}{\sum_{h=1}^{H} \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}} \cdot \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}$$

■

Comentario 16. Se puede observar que cuanto mayor es la variación dentro de un estrato y mayor es el tamaño poblacional de dicho estrato, mayor será el número de unidades muestrales del estrato h. Asimismo, cuanto menor es el coste de observar una unidad en el estrato h, mayor será n_h.

■

Comentario 17. La asignación óptima podría producir un tamaño de muestra n_h para algún estrato mayor que el tamaño poblacional correspondiente de dicho estrato, N_h. Este problema puede ocurrir en la práctica y surge cuando la fracción de muestreo total es considerable y algunos estratos son mucho más variables que otros. A continuación se detalla el algoritmo a seguir cuando ocurre este problema para finalmente tener una asignación óptima tal que $n_h \leq N, \ \forall h = 1, \ldots, H$, para un valor de n dado.

- Sea H' el conjunto de estratos para los que se obtiene $n_h > N_h$.
- Tomar

$$\begin{aligned} n_h^{'} &= N_h, \ \forall h \in H' \\ n_h^{'} &= \left(n - \sum_{k \in H'} N_h \right) \cdot \frac{\frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}}{\sum_{h \notin H'} \frac{N_h \cdot S_{yU_h}}{\sqrt{c_h}}}, \ \forall h \notin H' \end{aligned}$$

- Si todos los estratos verifican ahora $n_h \leq N_h$, ésta será la solución óptima. En caso contrario, se repite el proceso descrito hasta que se cumpla la condición $n_h \leq N_h \ \forall h$.

■

3.3.3 Otras asignaciones de la muestra

En este apartado se van a estudiar otros tipos de afijaciones alternativas que podrían ser útiles y que generalmente producen buenos resultados. A partir de ahora supondremos iguales los costes de observar una unidad en todos los estratos, $c_h \equiv c, \ \ \forall h = 1, \ldots, H$.

Asignación de Neyman

La asignación óptima teniendo en cuenta que los costes de observar una unidad son iguales en todos los estratos viene dada por

$$n_h = n \cdot \frac{N_h \cdot S_{yU_h}}{\sum_{h=1}^{H} N_h \cdot S_{yU_h}} \tag{3.9}$$

Se denomina *asignación de Neyman*, debido a la contribución importante de Neyman 1934.

Como se puede apreciar, el cálculo de los tamaños de muestra n_h requiere que las cuasidesviaciones típicas de cada subpoblación sean conocidas. Asimismo, en el caso del problema de minimizar la varianza dado un coste fijado C, la varianza mínima solo se podrá obtener si S_{yU_h} es conocida para todo $h = 1, \ldots, H$. Normalmente en la práctica estos valores no están disponibles. Para solventar este problema se podrían usar aproximaciones cercanas a las cuasidesviaciones típicas verdaderas haciendo uso de la experiencia pasada en el caso de encuestas repetidas en el tiempo. La asignación obtenida de esta forma podría ser cercana a la óptima. Una posible alternativa es usar la asignación proporcional, ya estudiada, o bien usar información auxiliar correlada con la variable de estudio.

Asignación óptima con información auxiliar

Sea x una variable auxiliar que presenta una correlación alta con la variable de estudio y, tal que las cuasidesviaciones típicas poblacionales de x, S_{xU_h}, son conocidas. Bajo estas condiciones, un método que suele ser usado en la práctica con buenos resultados es considerar la información conocida de esta variable auxiliar x en la expresión de la asignación óptima de la siguiente forma:

$$n_h = n \cdot \frac{N_h \cdot S_{xU_h}}{\sum_{h=1}^{H} N_h \cdot S_{xU_h}}$$

Si la correlación entre x e y es perfecta, esto es, $y_k = a + b \cdot x_k, \ \ k = 1, \ldots, N$, entonces la asignación es óptima, ya que en este caso $S^2_{yU_h} = b^2 \cdot S^2_{xU_h}$. Por tanto:

$$n_h = n \cdot \frac{N_h \cdot S_{xU_h}}{\sum_{h=1}^{H} N_h \cdot S_{xU_h}} = n \cdot \frac{N_h \cdot \frac{S_{yU_h}}{b}}{\sum_{h=1}^{H} N_h \cdot \frac{S_{yU_h}}{b}} = n \cdot \frac{N_h \cdot S_{yU_h}}{\sum_{h=1}^{H} N_h \cdot S_{yU_h}},$$

que es la asignación de Neyman. Si la correlación no es perfecta pero es fuerte, la expresión de n_h suele conducir a valores cercanos a la asignación óptima.

Asignación proporcional al total de la variable y

Suponiendo que la variable y toma valores positivos, la asignación proporcional al total de la variable y se define como

$$n_h = n \cdot \frac{\sum\limits_{k \in U_h} y_k}{\sum\limits_{k \in U} y_k}$$

Para poder aplicar este tipo de afijación, los totales de la variable y en cada estrato deben ser conocidos y esto precisamente no suele ocurrir, por lo que en la práctica no podrá ser usada.

Por otra parte, se puede observar que esta asignación es óptima si el coeficiente de variación

$$cv\left(Y_{U_h}\right) = \frac{S_{yU_h}}{\bar{y}_{U_h}}$$

es constante en todos los estratos, $cv\left(Y_{U_h}\right) \equiv cv, \ \ \forall h = 1, \ldots, H.$

$$\begin{aligned} n_h &= n \cdot \frac{\sum\limits_{k \in U_h} y_k}{\sum\limits_{k \in U} y_k} = n \cdot \frac{\sum\limits_{k \in U_h} y_k}{\sum\limits_{h=1}^{H} \sum\limits_{k \in U_h} y_k} = n \cdot \frac{N_h \cdot \bar{y}_{U_h}}{\sum\limits_{h=1}^{H} N_h \cdot \bar{y}_{U_h}} = \\ &= n \cdot \frac{N_h \cdot \frac{S_{yU_h}}{cv}}{\sum\limits_{h=1}^{H} N_h \cdot \frac{S_{yU_h}}{cv}} = n \cdot \frac{N_h \cdot S_{yU_h}}{\sum\limits_{h=1}^{H} N_h \cdot S_{yU_h}} \end{aligned}$$

Asignación proporcional al total de una variable auxiliar

Sea x una variable auxiliar que toma valores positivos y que presenta una correlación alta con la variable de estudio y, tal que los totales de la variable en cada estrato $\sum_{k \in U_h} x_k$ son conocidos. Se define la asignación proporcional al total de la variable x como

$$n_h = n \cdot \frac{\sum\limits_{k \in U_h} x_k}{\sum\limits_{k \in U} x_k}$$

Este tipo de afijación ha resultado ser útil en la práctica. La justificación de su uso radica en que si x e y tienen correlación alta y el coeficiente de variación es aproximadamente el mismo en todos los estratos, entonces esta asignación no debería estar lejos de la óptima.

Ejemplo 23. En el contexto del Ejemplo 22, supongamos ahora que se dispone de la siguiente información sobre la variable de estratificación x, que denota el número de escaños en el consejo municipal:

Número de escaños	N_h	$\sum_{k\in U_h} x_k$	$\sum_{k\in U_h} x_k^2$
$31-40$	44	1518	52764
$41-50$	168	7524	339344
$51-70$	56	3198	184168
71 o más	16	1260	100016

Se desea obtener una muestra estratificada de 40 para estimar el total de escaños del Partido Socialdemócrata. Dado que no tenemos información sobre la variable de estudio, para determinar la distribución de la muestra entre los distintos estratos, se podría usar la asignación proporcional, la asignación proporcional al total de x o bien la asignación óptima con la información auxiliar de x.

Asignación proporcional:

$$n_h = n \cdot \frac{N_h}{N} \Longrightarrow n_1 = 6,\ n_2 = 24,\ n_3 = 8,\ n_4 = 2$$

Asignación proporcional al total de la variable auxiliar x:

$$n_h = n \cdot \frac{\sum\limits_{k\in U_h} x_k}{\sum\limits_{k\in U} x_k} \Longrightarrow n_1 = 5,\ n_2 = 22,\ n_3 = 9,\ n_4 = 4$$

Asignación óptima con información auxiliar (x):

$$n_h = n \cdot \frac{N_h \cdot S_{xU_h}}{\sum\limits_{h=1}^{H} N_h \cdot S_{xU_h}} \Longrightarrow n_1 = 5,\ n_2 = 21,\ n_3 = 10,\ n_4 = 4$$

■

3.3.4 Asignación en el caso de múltiples variables de estudio

Sean $y_1, y_2, \ldots, y_I$ una serie de variables de estudio con $I \geq 2$, de las que se quiere estimar el total poblacional de cada una de ellas, $\sum\limits_{k\in U} y_{ik},\ \ \forall i = 1, \ldots, I$.

La mejor asignación para una variable no será en general la mejor para otra variable distinta. Por tanto, algúna solución debe ser adoptada en una encuesta donde existen numerosas características de estudio para obtener la mejor afijación. El primer paso será reducir el conjunto de variables consideradas para la obtención de la afijación a

un número relativamente pequeño de características que se consideran más importantes.

Una posible forma de asignación es la sugerida por Yates 1960, método que se describe a continuación. Considerando el diseño $strs$, el estimador del total de la variable y_i viene dado por

$$\widehat{Y}_{iU}^{\mathrm{HT}} = \sum_{h=1}^{H} N_h \cdot \bar{y}_{is_h}$$

donde $\bar{y}_{is_h}$ es la media de la variable y_i en la muestra s_h extraída del estrato h. Obsérvese que la expresión de la varianza dada en (3.4) puede escribirse de la siguiente forma para el estimador del total de y_i:

$$\mathbb{V}_i = \mathbb{V}_{strs}\left(\widehat{Y}_{iU}^{\mathrm{HT}}\right) = B_i + \sum_{h=1}^{H} \frac{A_{ih}}{n_h}$$

con

$$A_{ih} = N_h^2 \cdot S_{iU_h}^2; \quad B_i = -\sum_{h=1}^{H} N_h \cdot S_{iU_h}^2$$

donde $S_{iU_h}^2$ es la cuasivarianza poblacional de la variable y_i en el estrato U_h.

Consideremos la siguiente combinación lineal de las varianzas $\mathbb{V}_i$ dada por

$$\mathbb{V}_{lin} = \sum_{i=1}^{I} \omega_i \cdot \mathbb{V}_i = \sum_{i=1}^{I} \omega_i \cdot B_i + \sum_{h=1}^{H} \sum_{i=1}^{I} \frac{\omega_i \cdot A_{ih}}{n_h}$$

donde ω_i es el peso que representa la importancia de la variable y_i.

El método consiste en minimizar la expresión $\mathbb{V}_{lin}$ sujeto a la función de costes

$$C = c_0 + \sum_{h=1}^{H} n_h \cdot c_h,$$

o bien minimizar el coste sujeto a una varianza $\mathbb{V}_{lin}$ fijada. Este procedimiento tiene la ventaja de simplificar el problema reduciéndolo al caso unidimiensional.

Sea $V^* = \sum_{h=1}^{H} \sum_{i=1}^{I} \frac{\omega_i \cdot A_{ih}}{n_h}$ y $C^* = C - c_0$, entonces el problema planteado es equivalente a minimizar el producto

$$V^* \cdot C^* = \left(\sum_{h=1}^{H} \sum_{i=1}^{I} \frac{\omega_i \cdot A_{ih}}{n_h}\right) \cdot \left(\sum_{h=1}^{H} n_h \cdot c_h\right)$$

De la desigualdad de Cauchy se tiene

$$\left(\sum a_h^2\right) \cdot \left(\sum b_h^2\right) \geq \left(\sum a_h \cdot b_h\right)^2.$$

donde $a_h = \left(\sum_{i=1}^{I} \frac{\omega_i \cdot A_{ih}}{n_h} \right)^{1/2}$ y $b_h = (n_h \cdot c_h)^{1/2}$.

La igualdad se da cuando

$$\frac{b_h}{a_h} = \frac{(n_h \cdot c_h)^{1/2}}{\left(\sum_{i=1}^{I} \frac{\omega_i \cdot A_{ih}}{n_h} \right)^{1/2}} \equiv constante$$

o, de forma equivalente, se puede decir que, bajo el diseño $strs$,

$$n_h \propto \frac{1}{\sqrt{c_h}} \cdot \left(\sum_{i=1}^{I} \omega_i \cdot A_{ih} \right)^{1/2} = \frac{N_h}{\sqrt{c_h}} \cdot \left(\sum_{i=1}^{I} \omega_i \cdot S^2_{iU_h} \right)^{1/2}$$

La principal debilidad de este procedimiento es la posible arbitrariedad en la elección de los pesos ω_i de cada variable. Para más información, véase (Rao 1979).

Otra forma de abordar el problema de asignación de la muestra consiste en minimizar el coste dado por la función $C = c_0 + \sum_{h=1}^{H} n_h \cdot c_h$ bajo las siguientes restricciones:

- Se especifican unas tolerancias para la varianza correspondiente a cada variable.

  $$\mathbb{V}_i \leq \mathbb{V}_{i0}, \quad i = 1, \ldots, I.$$

  donde $\mathbb{V}_{i0}$ es la varianza deseada para la variable i
- $n_h \leq Nh, \quad h = 1, \ldots, H.$
- $n_h \geq 1, \quad h = 1, \ldots, H$ (requerida en caso de querer calcular la media de un estrato) o bien $n_h \geq 2, \quad h = 1$ (requerida en caso de querer calcular la cuasivarianza en un estrato).

Este problema de optimización puede ser descrito como un problema de programación matemática convexa. En (Danielsson 1975) se demuestra que este problema tiene una solución que puede ser enunciada analíticamente aunque en una forma compleja.

3.4 Comparación de precisión del estimador de Horvitz-Thompson en muestreo aleatorio estratificado según el tipo de asignación y el muestreo aleatorio simple

En primer lugar se va a realizar la comparación de la precisión del estimador HT bajo el diseño del muestreo aleatorio estratificado considerando la asignación óptima con respecto a la asignación proporcional, para comprobar que el estimador es siempre igual o más preciso cuando se usa la asignación óptima:

$$\mathbb{V}_{strs,opt}\left(\widehat{Y}_U^{\mathrm{HT}}\right) \leq \mathbb{V}_{strs,prop}\left(\widehat{Y}_U^{\mathrm{HT}}\right) \tag{3.10}$$

donde $\mathbb{V}_{strs,opt}\left(\widehat{Y}_U^{\mathrm{HT}}\right)$ es la varianza bajo muestreo aleatorio estratificado con asignación óptima y $\mathbb{V}_{strs,prop}\left(\widehat{Y}_U^{\mathrm{HT}}\right)$ con asignación proporcional.

La varianza del estimador HT bajo muestreo aleatorio estratificado con asignación óptima es, usando la asignación de Neyman dada en (3.9):

$$\begin{aligned}\mathbb{V}_{strs,opt}\left(\widehat{Y}_U^{\mathrm{HT}}\right) &= \sum_{h=1}^{H} N_h^2 \cdot (1-f_h) \cdot \frac{S_{yU_h}^2}{n_h} = \\ &= N^2 \sum_{h=1}^{H} \frac{W_h^2 \cdot S_{yU_h}^2}{n \cdot \frac{W_h \cdot S_{yU_h}}{\sum_{h=1}^{H} W_h \cdot S_{yU_h}}} - N \sum_{h=1}^{H} W_h \cdot S_{yU_h}^2 = \\ &= \frac{N^2}{n} \cdot \left(\sum_{h=1}^{H} W_h \cdot S_{yU_h}\right)^2 - N \sum_{h=1}^{H} W_h \cdot S_{yU_h}^2\end{aligned}$$

La varianza bajo muestreo aleatorio estratificado con asignación proporcional es, usando la asignación dada en (3.5):

$$\begin{aligned}\mathbb{V}_{strs,prop}\left(\widehat{Y}_U^{\mathrm{HT}}\right) &= \sum_{h=1}^{H} N_h^2 \cdot (1-f_h) \cdot \frac{S_{yU_h}^2}{n_h} = N^2 \sum_{h=1}^{H} \frac{W_h^2 \cdot S_{yU_h}^2}{n \cdot \frac{N_h}{N}} - N \sum_{h=1}^{H} W_h \cdot S_{yU_h}^2 = \\ &= N^2 \cdot \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} W_h \cdot S_{yU_h}^2\end{aligned}$$

Restando las dos expresiones, se obtiene lo que se pretendía demostrar:

$$\begin{aligned}\mathbb{V}_{strs,prop}\left(\widehat{Y}_U^{\mathrm{HT}}\right) - \mathbb{V}_{strs,opt}\left(\widehat{Y}_U^{\mathrm{HT}}\right) &= \frac{N^2}{n} \sum_{h=1}^{H} W_h \cdot S_{yU_h}^2 - \frac{N^2}{n} \cdot \left(\sum_{h=1}^{H} W_h \cdot S_{yU_h}\right)^2 = \\ &= \frac{N^2}{n} \cdot \left[\sum_{h=1}^{H} W_h \cdot S_{yU_h}^2 - \left(\sum_{h=1}^{H} W_h \cdot S_{yU_h}\right)^2\right] = \\ &= \frac{N^2}{n} \sum_{h=1}^{H} W_h \cdot \left(S_{yU_h} - \bar{S}_y\right)^2 \geq 0\end{aligned}$$

donde $\bar{S}_y = \sum_{h=1}^{H} W_h \cdot S_{yU_h}$ es la media de las cuasidesviaciones típicas de los estratos.

Comentario 18. Obsérvese que la igualdad de precisiones se obtiene cuando las cuasidesviaciones típicas S_{yU_h} son iguales en todos los estratos. La ganancia en precisión de la asignación óptima respecto de la proporcional será mayor cuanta más variación exista en las cuasivarianzas de los estratos.

■

Comparemos la precisión obtenida con el estimador HT bajo el diseño del muestreo aleatorio estratificado con asignación proporcional con respecto al diseño del muestreo aleatorio simple.

De la descomposición del análisis de la varianza, tenemos que la variación total, SST, se puede descomponer como la suma de la variación interestrato, SSB, y la variación intraestrato, SSW.

$$(N-1)\cdot S_{yU}^2 = \sum_{k\in s}(y_k-\bar{y}_U)^2 = \sum_{h=1}^{H} N_h\cdot(\bar{y}_{U_h}-\bar{y}_U)^2 + \sum_{h=1}^{H}(N_h-1)\cdot S_{yU_h}^2 \qquad (3.11)$$

De forma equivalente, podemos representarlo como

$$SST = SSB + SSW$$

En efecto:

$$\begin{aligned} SST &= \sum_{h=1}^{H}\sum_{k\in U_h}(y_k-\bar{y}_U)^2 = \sum_{h=1}^{H}\sum_{k\in U_h}(y_k-\bar{y}_{U_h}+\bar{y}_{U_h}-\bar{y}_U)^2 = \\ &= \sum_{h=1}^{H}\sum_{k\in U_h}(\bar{y}_{U_h}-\bar{y}_U)^2 + \sum_{h=1}^{H}\sum_{k\in U_h}(y_k-\bar{y}_{U_h})^2 + 2\sum_{h=1}^{H}\sum_{k\in U_h}(y_k-\bar{y}_{U_h})\cdot(\bar{y}_{U_h}-\bar{y}_U) = \\ &= \underbrace{\sum_{h=1}^{H} N_h\cdot(\bar{y}_{U_h}-\bar{y}_U)^2}_{SSB} + \underbrace{\sum_{h=1}^{H}(N_h-1)\cdot S_{yU_h}^2}_{SSW} \end{aligned}$$

ya que el último sumando es 0.

Recordemos que la varianza del estimador HT bajo el diseño $srswor$ es

$$\begin{aligned} \mathbb{V}_{srswor}(\widehat{Y}_U^{\mathrm{HT}}) &= N^2\cdot\left(\frac{1}{n}-\frac{1}{N}\right)\cdot S_{yU}^2 = \\ &= \frac{N^2}{N-1}\cdot\left(\frac{1}{n}-\frac{1}{N}\right)\cdot\left[\sum_{h=1}^{H} N_h\cdot(\bar{y}_{U_h}-\bar{y}_U)^2 + \sum_{h=1}^{H}(N_h-1)\cdot S_{yU_h}^2\right] = \\ &= \frac{N^3}{N-1}\cdot\left(\frac{1}{n}-\frac{1}{N}\right)\cdot\left[\sum_{h=1}^{H} W_h\cdot(\bar{y}_{U_h}-\bar{y}_U)^2 + \sum_{h=1}^{H}\left(W_h-\frac{1}{N}\right)\cdot S_{yU_h}^2\right] \end{aligned}$$

donde se ha usado (3.11).

Por tanto:

$$\mathbb{V}_{srswor}\left(\widehat{Y}_U^{\mathrm{HT}}\right) - \mathbb{V}_{strs}\left(\widehat{Y}_U^{\mathrm{HT}}\right) =$$

$$= \frac{N^3}{N-1}\left(\frac{1}{n}-\frac{1}{N}\right)\left[\sum_{h=1}^{H} W_h(\bar{y}_{U_h}-\bar{y}_U)^2+\sum_{h=1}^{H}\left(W_h-\frac{1}{N}\right)S^2_{yU_h}-\frac{N-1}{N}\sum_{h=1}^{H}W_hS^2_{y_h}\right]=$$
$$= \frac{N^3}{N-1}\left(\frac{1}{n}-\frac{1}{N}\right)\left[\sum_{h=1}^{H} W_h(\bar{y}_{U_h}-\bar{y}_U)^2+\sum_{h=1}^{H}\left(W_h-\frac{1}{N}\right)S^2_{yU_h}-\frac{N-1}{N}\sum_{h=1}^{H}W_hS^2_{y_h}\right]=$$
$$= \frac{N^3}{N-1}\left(\frac{1}{n}-\frac{1}{N}\right)\left[\sum_{h=1}^{H} W_h(\bar{y}_{U_h}-\bar{y}_U)^2-\frac{1}{N}\sum_{h=1}^{H}(1-W_h)S^2_{yU_h}\right]$$

Comentario 19. Del resultado anterior, se puede afirmar que es teóricamente posible que el muestreo aleatorio estratificado con asignación proporcional conduzca a una varianza ligeramente mayor que el diseño $srswor$, en el caso de que todos los estratos presenten medias $\bar{y}_{Uh}$ iguales o aproximadamente iguales. Sin embargo, en la mayoría de los casos esta igualdad no se da y normalmente $\sum_{h=1}^{H} W_h \cdot (\bar{y}_{U_h}-\bar{y}_U)^2$ excederá considerablemente el valor de $\frac{1}{N}\sum_{h=1}^{H}(1-W_h)\cdot S^2_{yU_h}$. ■

Comentario 20. Si la variación interestrato representa una proporción significativa de la variación total, entonces las diferencias entre las medias de los estratos son una de las principales razones de la variación de y. En tal caso, el diseño $strs$ con asignación proporcional produce una varianza sustancialmente menor. Así pues, cuanto más difieran las medias de los estratos más ganancia en precisión se obtendrá con la estratificación respecto a aplicar muestreo aleatorio simple.

■

Ejemplo 24. En el contexto del Ejemplo 22 y el Ejemplo 23, supongamos que se conoce el total del número de escaños en los consejos municipales, para cada estrato. Por consiguiente, se dispone del total poblacional de la variable de estudio y. La información viene recogida en la siguiente tabla:

Número de escaños	N_h	$\sum_{k\in U_h} x_k$	$\sum_{k\in U_h} x_k^2$
$31-40$	44	1518	52764
$41-50$	168	7524	339344
$51-70$	56	3198	184168
71 o más	16	1260	100016

Calculemos la varianza para las tres asignaciones obtenidas en el Ejemplo 23 y comparemos el resultado respecto a aplicar el diseño del muestreo aleatorio simple usando el estimador HT.

Las cuasivarianzas poblacionales de cada estrato son:

$$S^2_{yU_h} = \frac{1}{N_h-1}\cdot\left(\sum_{k\in U_h} y_k^2 - N_h\cdot\bar{y}^2_{U_h}\right)$$

$$S^2_{yU_1} \approx 18{,}48; \quad S^2_{yU_2} \approx 24{,}55; \quad S^2_{yU_3} \approx 34{,}61; \quad S^2_{yU_4} \approx 22{,}93$$

La varianza del estimador usando la asignación proporcional es

$$\mathbb{V}_{strs,prop}\left(\widehat{Y}_U^{\mathrm{HT}}\right) = \sum_{h=1}^{H} N_h^2 \cdot \frac{1-f_h}{n_h} \cdot S^2_{yU_h} \approx 44092{,}60$$

La varianza es ligeramente menor aplicando muestreo estratificado con asignación proporcional, ya que la varianza obtenida con asignación proporcional al total de x y la varianza aplicando asignación óptima con información auxiliar es, respectivamente:

$$\mathbb{V}_{strs,prop_x} \approx 44934{,}73; \quad \mathbb{V}_{strs,opt_x} \approx 45228{,}54$$

Para calcular la varianza del estimador de expansión bajo muestreo aleatorio simple, debemos obtener en primer lugar la cuasivarianza poblacional de y.

$$S^2_{yU} = \frac{1}{N-1} \cdot \left[\sum_{h=1}^{H} N_h \cdot (\bar{y}_{U_h} - \bar{y}_U)^2 + \sum_{h=1}^{H} (N_h - 1) \cdot S^2_{yU_h}\right] \approx 52{,}56$$

donde se ha usado $\bar{y}_{U_1} = \frac{189}{11}$, $\bar{y}_{U_2} = \frac{3383}{168}$, $\bar{y}_{U_3} = \frac{1545}{56}$, $\bar{y}_{U_4} = 38{,}5625$.

Se sabe que la variable de estratificación x tiene correlación positiva alta con la variable de estudio y (es aproximadamente 0.76), por lo que cabe esperar que la varianza obtenida bajo muestreo aleatorio simple con tamaño $n = 40$ y usando el estimador de expansión HT sea mucho mayor que usando muestreo estratificado con las asignaciones que tienen en cuenta la información de x. Asimismo, dado que SSB representa un porcentaje significativo de la variación total (aproximadamente $52\,\%$), se podría esperar una varianza sustancialmente menor que con el diseño $srswor$.

En efecto, la varianza del diseño $srswor$ es

$$\mathbb{V}(\widehat{Y}_U^{\mathrm{HT}}) = N^2 \cdot \frac{1-f}{n} \cdot S^2_{yU} \approx 91058{,}80$$

Se observa una ganancia importante con la estratificación, tal y como cabía esperar.

■

Bibliografía

Cochran, W. G. (1977). *Sampling Techniques*. 3rd. Wiley.

Danielsson, S. (1975). "Optimal allokering vid vissa klasser av urvalsförfaranden". Tesis doct. Department of Mathematics, University of Linköping, Sweden.

Lohr, S. (2010). *Sampling: Design and Analysis*. 2nd. Duxbury Press.

Neyman, J. (1934). "On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection". En: *Journal of the Royal Statistical Society* 97, págs. 558-625.

Rao, J. N. K. (1979). "Optimization in the design of sample surveys". En: *J. S. Rustagi (ed.), Optimizing Methods in Statistics: Proceedings of an International Conference*. New York: Academic Press, págs. 419-434.

Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.

Yates, F. (1960). *Sampling Methods for Censuses and Surveys*. 3rd. London: Charles Griffin y Co.

Tema 4

Muestreo por conglomerados con probabilidades idénticas en una etapa. Muestreo por conglomerados con probabilidades idénticas en dos etapas. Muestreo con probabilidades diferentes en una etapa con reemplazo. Muestro con probabilidades diferentes en dos etapas con reemplazo. Muestreo con probabilidades diferentes sin reemplazo.

Este tema está elaborado como una adaptación casi literal en español de la siguiente bibliografía:

C.-E. Särndal, B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer

S. Lohr (2010). *Sampling: Design and Analysis*. 2nd. Duxbury Press

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

4.1 Introducción

Los diseños considerados hasta el momento asumen que es posible el muestreo directo de los elementos. Es decir, los elementos de la población se pueden usar como elementos muestrales en una única etapa de muestreo. Sin embargo, en muchas encuestas de tamaño mediano y grande, el muestreo directo de elementos no se usa por una o dos de las siguientes razones:

i. No existe un marco muestral que identifique a todos y cada uno de los elementos de la población, y la producción de tal marco puede ser difícil, cara o imposible.

   Por ejemplo, no se puede construir un listado con todos los clientes de una tienda, pero sí podemos construir una lista de todos los individuos de una ciudad para los cuales sólo existe una lista de viviendas; sin embargo, la construcción de esta lista consumirá mucho tiempo y será muy cara.

ii. Los elementos de la población están dispersos sobre un área muy grande, en cuyo caso el muestreo directo de elementos dará lugar a una muestra demasiado dispersa. Por tanto, el coste del trabajo de campo sería prohibitivo debido al alto

coste de los viajes si son necesarias entrevistas personales. También puede ser dificultosa la supervisión del trabajo de campo, lo que puede dar lugar a una tasa de falta de respuesta muy grande y a errores de muestreo graves.

Si la población objetivo son los alumnos de un determinado curso, por ejemplo cuarto de ESO, en España, es más barato tomar una muestra de colegios y entrevistar a todos los alumnos de cuarto de los colegios seleccionados que entrevistar a toda una muestra aleatoria simple de alumnos de cuarto de ESO. Con una muestra aleatoria simple de alumnos es posible que en algún caso hubiera que viajar hasta algún colegio para entrevistar a un único alumno.

El *muestreo por conglomerados* es un diseño de muestreo no directo de elementos en el que la población finita está agrupada en subpoblaciones llamadas *conglomerados*. Consiste en seleccionar un subconjunto de conglomerados donde todos los elementos de la población del conglomerado seleccionado son encuestados. El muestreo por conglomerados también se llama *muestreo por conglomerados en una única etapa* o *monoetápico*. Por contra, en el *muestreo de conglomerados con submuestreo* la muestra de elementos se obtiene como resultado de dos etapas de muestreo:

i. Los elementos de la población se agrupan en primer lugar en subpoblaciones disjuntas, llamadas *unidades muestrales primarias* (PSUs del inglés *primary sampling units*). Se obtiene la probabilidad de selección de cada PSU (*muestreo de primera etapa*).

ii. Para cada PSU de primera etapa, se decide el tipo de unidad de muestreo a usar en el muestreo de la segunda etapa. Estas *unidades muestrales de segunda etapa* (SSUs del inglés *secondary sampling units*) pueden ser elementos o conglomerados de elementos. Se obtiene la probabilidad muestral de las SSUs para cada PSU en el muestreo de primera etapa. Cuando las SSUs son conglomerados, se encuesta a cada unidad en las SSUs seleccionados en el caso de utilizar solo dos etapas; en caso contrario, se sigue muestreando.

Comentario 21. Cuando cada SSU es un elemento, usamos el término muestreo bietápico; cuando cada SSU es un conglomerado de elemento, usamos el término muestreo de conglomerados con submuestreo o muestreo multietápico. ■

Ejemplo 25. Supóngase que quieren estimarse el número de *tablets* electrónicas por residente en un país. No existe un listado de todos estos dispositivos vendidos y el muestreo directo sobre personas conduciría a una muestra dispersa cuya recogida de datos superaría nuestras restricciones presupuestarias. Por tanto, no puede plantearse un muestreo directo sobre las personas. En su lugar, puede considerarse un listado de secciones censales y entrevistar a todos los residentes de cada sección censal seleccionada. Se trata de un muestreo por conglomerados monoetápico donde las unidades muestrales primarias (las PSUs) son las secciones censales y las unidades de análisis o elementos son las personas residentes. No hay SSUs. ■

Ejemplo 26. La Encuesta sobre Uso de Drogas en Enseñanzas Secundarias en España, ESTUDES [1], es una operación estadística que tiene por objetivo recabar información de valor para diseñar y evaluar políticas dirigidas a prevenir el consumo de drogas y otras adicciones y los problemas derivados del mismo. Para ello se realiza una encuesta a estudiantes de 14 a 18 años matriculados en la ESO, Bachillerato, Ciclos de Formación Profesional Básica y Ciclos Formativos de Grado Medio de Formación Profesional.

Se realiza un muestreo por conglomerados bietápico, en el que, en primera etapa, se seleccionan aleatoriamente centros educativos (unidades de primera etapa) y, en segundo lugar, aulas (unidades de segunda etapa), cumplimentando el cuestionario a todos los alumnos presentes en las mismas.

En este caso los centros educativos son las PSUs y las aulas son las SSUs. Los alumnos son los elementos. ■

Definición 22

El *muestreo multietápico* consiste en un muestreo en tres o más etapas. Hay una jerarquía de unidades muestrales: las unidades muestrales de primera etapa, las unidades muestrales secundarias dentro de las PSUs, las unidades muestrales terciarias dentro de las SSUs, y así sucesivamente. Las unidades muestrales en la última etapa de muestreo se llaman *unidades de muestreo últimas* y las de aquellas en la etapa anterior a la última se llaman *unidades de muestreo penúltimas*.

Comentario 22. El término muestreo multietápico de elementos se aplica cuando las unidades de muestreo últimas son elementos. En el muestreo multietápico de conglomerados, las unidades de muestreo últimas son conglomerados de elementos. ■

Ejemplo 27. La Encuesta Europea de Salud en España[2] es una operación estadística dirigida al conjunto de personas de 15 y más años que reside en viviendas familiares en todo el territorio nacional. Su objetivo principal es obtener datos sobre el estado de salud, la utilización de los servicios sanitarios y los factores determinantes de salud, de manera armonizada y comparable a nivel europeo.

Para ello utiliza un muestreo trietápico. Las unidades de primera etapa son las secciones censales. Las unidades de segunda etapa son las viviendas familiares principales, investigándose a todos los hogares que tienen su residencia habitual en las mismas. Dentro de cada hogar se selecciona a un adulto (15 o más años).

Esto es un ejemplo de muestreo trietápico en el que las secciones censales son las PSUs, las viviendas son las SSUs y el adulto que rellena el cuestionario es la unidad muestral

[1] https://pnsd.sanidad.gob.es/profesionales/sistemasInformacion/sistemaInformacion/encuestas_ESTUDES.htm

[2] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176784&menu=metodologia&idp=1254735573175

terciaria o unidad muestral final. ■

Asimismo, hasta ahora se han analizado exclusivamente diseños de muestreo donde las probabilidades de selección de las unidades de muestreo son iguales. Este tipo de diseños proporcionan esquemas que, frecuentemente, son fáciles de diseñar y explicar. Sin embargo, no siempre es posible aplicarlos y no son tan eficientes como los diseños con probabilidades desiguales.

En este tema se estudiarán diseños de muestreo con probabilidades diferentes sin reemplazamiento y con reemplazamiento, por lo que se aplicarán los estimadores de Horvitz-Thompson y Hansen y Hurwitz, que se definieron en el tema 1. A lo largo del tema se considerará que la población está agrupada en subpoblaciones. Se estudiarán los estimadores, las varianzas y los estimadores de las varianzas en diseños monoetápicos y bietápicos cuando las unidades muestrales de primera y segunda etapa son seleccionadas con probabilidades diferentes. Los resultados obtenidos se aplicarán al caso particular en el que la selección de las unidades muestrales sea con probabilidades idénticas en una y en dos etapas.

4.2 Muestreo con probabilidades diferentes sin reemplazo

En esta sección se estudiarán diseños de muestreo en los que las unidades de primera etapa son seleccionadas mediante un esquema sin reemplazo, por lo que se usará el estimador de Horvitz-Thompson (HT) definido en el tema 1, así como las expresiones de la varianza del estimador y el estimador de la varianza obtenidas en el Teorema 4.

Existen diversos algoritmos de selección de la muestra sin reemplazamiento, por ejemplo podría utilizarse el muestreo sistemático, que es de un uso muy extendido por su facilidad de implementación, aunque tiene el inconveniente de que muchas de las probabilidades de inclusión de segundo orden son cero. Por otra parte, en el tema 2 se estudiaron algunos algoritmos muestrales para la selección de una muestra aleatoria simple sin reemplazo. También existen mecanismos de selección de la muestra con probabilidades diferentes. Por ejemplo, podría pensarse en seleccionar unidades primarias con probabilidades proporcionales al tamaño (entendido como una medida del tamaño de la unidad primaria, que podría ser por ejemplo el número de elementos). La mayoría de los algoritmos de selección con probabilidades diferentes son complicados de implementar. Generalmente los cálculos de las probabilidades de inclusión de segundo orden se vuelven rápidamente engorrosas a medida que el tamaño de la muestra se incrementa. En (Brewer y Hanif 1983) se pueden consultar diversos métodos para seleccionar muestras con probabilidades diferentes sin reemplazo.

A continuación se presenta un método para la selección de una única unidad primaria ($n_I = 1$), con probabilidades proporcionales al tamaño, que denominaremos *método del total acumulado*:

Sea T_i el tamaño acumulado hasta la unidad primaria i, donde $T_0 = 0$.

1. Se calcula $T_i = T_{i-1} + N_i$, donde N_i es el tamaño de la unidad primaria i.
2. Se considera una realización de una uniforme sobre el intervalo $(0, 1)$: ε. La unidad i es seleccionada si
$$T_{i-1} < \varepsilon \cdot T_{N_I} \leq T_i$$
donde N_I es el número de unidades primarias en las que se agrupa la población.

El índice I se usa para identificar entidades asociadas con las unidades de primera etapa. Se puede comprobar que se tiene efectivamente un diseño con probabilidades proporcionales al tamaño:

$$\pi_{Ii} = \mathbb{P}(T_{i-1} < \varepsilon \cdot T_{N_I} \leq T_i) = \frac{T_i - T_{i-1}}{T_{N_I}} = \frac{N_i}{\sum_{i=1}^{N_I} N_i}$$

donde $T_{N_I} = \sum_{i=1}^{N_I} N_i$ es el tamaño de la población.

Ejemplo 28. Supóngase que una población está agrupada en 6 subpoblaciones para la cual se dispone de la información del número de individuos en cada una de ellas $N_1 = 15, N_2 = 10, N_3 = 30, N_4 = 20, N_5 = 15$ y $N_6 = 10$. Calculemos el total acumulado T_i:

i	N_i	T_i
1	15	15
2	10	25
3	30	55
4	20	75
5	15	90
6	10	100

Se obtiene una realización de una uniforme sobre el intervalo $(0, 1)$, ε. Suponiendo que $25 < \varepsilon \cdot 100 \leq 55$, entonces la unidad primaria seleccionada es $i = 3$. ■

4.2.1 Muestreo por conglomerados en una etapa

Considérese que la población finita $U = \{1, .., k, ..., N\}$ se divide en N_I subpoblaciones, llamadas conglomerados, y denotadas por $U_1, ..., U_i, ..., U_{N_I}$. El conjunto de conglomerados se representa simbólicamente por

$$U_I = \{1, ..., i, ..., N_I\}.$$

El índice I se utilizará para identificar entidades asociadas con los conglomerados, que son las unidades primarias. La razón para usar I en lugar de C, que sería más natural, es que el primero facilita la transición al muestreo de conglomerados con submuestreo, en que I se referirá a la primera etapa, II a la segunda etapa y si hubiese más etapas se podrían seguir usando estos subíndices.

El número de elementos de la población en el i-ésimo conglomerado U_i se denota por N_i. La partición de U se expresa mediante las ecuaciones

$$U = \bigcup_{i \in U_I} U_i \quad \text{y} \quad N = \sum_{i \in U_I} N_i.$$

De momento, consideraremos que solo se realiza una etapa de muestreo, es decir, todos los elementos de las unidades muestrales seleccionadas serán observados.

Definición 23

El *muestreo por conglomerados en una única etapa* (o simplemente *muestreo por conglomerados*) se define ahora de la siguiente forma:

i. Una muestra probabilística s_I de conglomerados se selecciona a partir de U_I de acuerdo con el diseño $p_I(\cdot)$. El tamaño de s_I se denota por n_I, para un diseño de tamaño fijo, o por n_{s_I} para un diseño de tamaño variable.

ii. Cada elemento poblacional de los conglomerados seleccionados es observado (entrevistado).

Aquí $p_I(\cdot)$ puede ser cualquiera de los diseños convencionales, es decir, muestreo aleatorio simple sin reemplazamiento, muestreo sistemático, muestreo estratificado, etcétera. Seguiremos usando s como símbolo del conjunto de elementos que se observan. Es decir,

$$s = \bigcup_{i \in s_I} U_i.$$

El tamaño de s es

$$n_s = \sum_{i \in s_I} N_i.$$

Señalamos que incluso si $p_I(\cdot)$ es un diseño de tamaño fijo, el número de elementos observados n_s en general no será fijo, porque los tamaños de los conglomerados N_i pueden variar.

Las probabilidades de inclusión de la unidad de muestreo de primer y segundo orden inducidas por el diseño $p_I(\cdot)$ son

$$\pi_{Ii} = \sum_{s_I \ni i} p_I(s_I)$$

y para dos conglomerados i y j

$$\pi_{Iij} = \sum_{s_I \ni i,j} p_I(s_I).$$

Se verifica que $\pi_{Iii} = \pi_{Ii}$. Volvamos a las probabilidades de inclusión del elemento. Como la muestra s contiene todos los elementos en los conglomerados seleccionados, para cada k en U_i, tenemos

$$\pi_k = \mathbb{P}(s \ni k) = \mathbb{P}(s_I \ni i) = \pi_{Ii}. \tag{4.1}$$

Las probabilidades de inclusión de segundo orden vienen dadas por

$$\pi_{kl} = \mathbb{P}(s \ni k, l) = \mathbb{P}(s_I \ni i) = \pi_{Ii} \tag{4.2}$$

si tanto k como l pertenecen al mismo conglomerado U_i, y

$$\pi_{kl} = \mathbb{P}(s \ni k, l) = \mathbb{P}(s_I \ni i, j) = \pi_{Iij} \tag{4.3}$$

si k y l pertenecen a distintos conglomerados U_i y U_j. Cabe señalar que $\pi_{kk} = \pi_k$. Es conveniente introducir la notación simplificada

$$Y_{U_i} = \sum_{k \in U_i} y_k$$

para el total de la unidad primaria i-ésima. El total poblacional a estimar se puede expresar entonces como

$$Y_U = \sum_{k \in U} y_k = \sum_{i \in U_I} Y_{U_i}.$$

Sea $\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$. Tenemos entonces el siguiente resultado:

Teorema 21

En el muestreo por conglomerados, el estimador de Horvitz-Thompson del total poblacional $Y_U = \sum_{k \in U} y_k$ se puede escribir como

$$\widehat{Y}_U^{\text{HT}} = \sum_{i \in s_I} \frac{Y_{U_i}}{\pi_{Ii}}. \tag{4.4}$$

La varianza viene dada por

$$\mathbb{V}[\widehat{Y}_U^{\text{HT}}] = \sum_{i \in U_I} \sum_{j \in U_I} \Delta_{Iij} \frac{Y_{U_i}}{\pi_{Ii}} \frac{Y_{U_j}}{\pi_{Ij}}. \tag{4.5}$$

Un estimador insesgado de la varianza es

$$\widehat{\mathbb{V}}[\widehat{Y}_U^{\text{HT}}] = \sum_{i \in s_I} \sum_{j \in s_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \frac{Y_{U_i}}{\pi_{Ii}} \frac{Y_{U_j}}{\pi_{Ij}}. \tag{4.6}$$

Demostración 21

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{k\in s} \frac{y_k}{\pi_k} = \sum_{i\in s_I} \sum_{k\in U_i} \frac{y_k}{\pi_k}$$

usamos la ecuación (4.1) y obtenemos

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{i\in s_I} \frac{\left(\sum_{k\in U_i} y_k\right)}{\pi_{Ii}} = \sum_{i\in s_I} \frac{Y_{U_i}}{\pi_{Ii}}.$$

Los otros resultados se obtienen inmediatamente a partir de los resultados estudiados en el tema 1 sobre el estimador de Horvitz-Thompson, teniendo en cuenta que $\sum_{i\in s_I} \frac{Y_{U_i}}{\pi_{Ii}}$ es el estimador de Horvitz-Thompson de $\sum_{i\in U_I} Y_{U_i}$. Las probabilidades de primer orden adecuadas son las probabilidades de inclusión de las unidades primarias.

Si $p_I(\cdot)$ es un diseño muestral de tamaño fijo, la varianza $\mathbb{V}[\widehat{Y}_U^{\mathrm{HT}}]$ en el Teorema 21 también se puede expresar como

$$\mathbb{V}[\widehat{Y}_U^{\mathrm{HT}}] = -\frac{1}{2} \sum_{i\in U_I} \sum_{j\in U_I} \Delta_{Iij} \left(\frac{Y_{U_i}}{\pi_{Ii}} - \frac{Y_{U_j}}{\pi_{Ij}}\right)^2 \tag{4.7}$$

con el estimador insesgado de la varianza

$$\widehat{\mathbb{V}}(\widehat{Y}_U^{\mathrm{HT}}) = -\frac{1}{2} \sum_{i\in s_I} \sum_{j\in s_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \left(\frac{Y_{U_i}}{\pi_{Ii}} - \frac{Y_{U_j}}{\pi_{Ij}}\right)^2. \tag{4.8}$$

El Teorema 21 conduce a varias conclusiones interesantes sobre la eficiencia del muestreo por conglomerados. Asumimos que se usa el estimador de Horvitz-Thompson y que $p_I(\cdot)$ es un diseño de tamaño fijo, por eso se utilizan las ecuaciones (4.7) y (4.8).

a. A partir de (4.7), podemos ver que si todos los $\check{Y}_{U_i} = \frac{Y_{U_i}}{\pi_{Ii}}$ son iguales, entonces $\mathbb{V}[\widehat{Y}_U^{\mathrm{HT}}] = 0$. Por tanto, si podemos elegir π_{Ii} aproximadamente proporcional a los totales de los conglomerados Y_{U_i}, entonces el muestreo por conglomerados será altamente eficiente.

b. Si los tamaños de los conglomerados N_i son conocidos en la fase de planificación, se puede elegir un diseño con $\pi_{Ii} \propto N_i$. Como $Y_{U_i} = N_i \bar{y}_{U_i} = \sum_{k\in U_i} y_k$, es una buena elección si hay poca variación entre las medias de los conglomerados $\bar{y}_{U_i}$. Si todos los $\bar{y}_{U_i}$ son iguales, tendríamos, en efecto, $\mathbb{V}[\widehat{Y}_U^{\mathrm{HT}}] = 0$.

c. Un diseño de muestreo por conglomerados con probabilidades iguales (es decir, uno en el que todos los π_{Ii} son iguales) a menudo es una mala opción cuando los conglomerados tienen distinto tamaño. Para que un diseño así sea eficiente,

debemos tener $\bar{y}_{U_i}$ aproximadamente proporcionales a N_i^{-1}. Esto, sin embargo, raramente se da en la práctica.

Comentario 23. El muestreo por conglomerados y el muestreo estratificado.
Los conglomerados recuerdan a los estratos, pero sólo de manera superficial: un conglomerado, al igual que un estrato, es una agrupación de los elementos de una población (individuos de una ciudad, empresas de una comunidad autónoma, establecimientos industriales de una provincia). Sin embargo, el proceso de construcción y selección es un poco distinto en estos dos métodos.

Mientras que, por lo general, la estratificación aumenta la precisión en relación con el muestreo aleatorio simple, el muestreo por conglomerados, con frecuencia, la disminuye. Los miembros de un mismo conglomerado tienden a ser más similares entre sí que los elementos seleccionados al azar de entre toda la población: los alumnos de cuarto de ESO de un mismo colegio tienden a tener un nivel de vida parecido; los peces de un mismo lago tienden a presentar concentraciones similares de mercurio. Por lo general, estas analogías surgen debido a ciertos factores subyacentes que podrían medirse o no. Por tanto, si extraemos una muestra con dos alumnos de un mismo colegio puede que no consigamos tanta información acerca de los alumnos de cuarto de ESO como la que obtendríamos al extraer una muestra de dos alumnos de colegios distintos.

El muestreo por conglomerados se utiliza en la práctica (sobre todo en las encuestas sociales) debido a que es más económico y conveniente obtener muestras por conglomerados que al azar entre la población de personas. Casi todas las encuestas sociales (a hogares, a individuos, a viviendas) utilizan el muestreo por conglomerados debido al ahorro de costes. ■

4.2.2 Muestreo por conglomerados con probabilidades idénticas en una etapa

El muestreo por conglomerados con probabilidades idénticas en una etapa consiste en seleccionar una muestra aleatoria simple sin reemplazamiento s_I de n_I unidades de muestreo (conglomerados) del total de N_I conglomerados existentes en U_I, de forma que todos los elementos de los conglomerados seleccionados son observados. Denotaremos este diseño por $1st$[3].

Se sigue del Teorema 21 y de los resultados sobre el diseño muestral aleatorio simple sin reemplazamiento que el estimador de Horvitz-Thompson del total poblacional viene dado por

$$\widehat{Y}_U^{\text{HT}} = N_I \bar{y}_{s_I}, \tag{4.9}$$

donde $\bar{y}_{s_I} = \dfrac{1}{n_I} \sum_{i \in s_I} Y_{U_i}$ es la media de los totales de conglomerados en s_I. La varianza se puede escribir como

$$\mathbb{V}_{1st}(\widehat{Y}_U^{\text{HT}}) = N_I^2 \frac{1 - f_I}{n_I} S^2_{YU_I}, \tag{4.10}$$

[3]Del inglés *simple cluster random sampling without replacement*.

donde $f_I = \frac{n_I}{N_I}$ es la fracción de muestreo de conglomerados y

$$S^2_{YU_I} = \frac{1}{N_I - 1} \sum_{i \in U_I} (Y_{U_i} - \bar{y}_{U_I})^2, \tag{4.11}$$

con $\bar{y}_{U_I} = \sum_{i \in U_I} \frac{Y_{U_i}}{N_I}$. El estimador insesgado de la varianza es

$$\widehat{\mathbb{V}}_{1st}(\widehat{Y}_U^{\text{HT}}) = N_I^2 \frac{1 - f_I}{n_I} S^2_{Ys_I}, \tag{4.12}$$

donde

$$S^2_{Ys_I} = \frac{1}{n_I - 1} \sum_{i \in s_I} (Y_{U_i} - \bar{y}_{s_I})^2.$$

De la descomposición del análisis de la varianza, tenemos que la variación total, SST, se puede descomponer en este caso como la parte de variabilidad que existe entre los conglomerados, SSB, y la parte de variabilidad dentro de los conglomerados, SSW:

$$\underbrace{\sum_{i \in U_I} \sum_{k \in U_i} (y_k - \bar{y}_U)^2}_{SST} = \underbrace{\sum_{i \in U_I} N_i \cdot (\bar{y}_{U_i} - \bar{y}_U)^2}_{SSB} + \underbrace{\sum_{i \in U_I} \sum_{k \in U_i} (y_k - \bar{y}_{U_i})^2}_{SSW}$$

O alternativamente, esto se puede escribir

$$(N - 1) \cdot S^2_{yU} = (N_I - 1) \cdot S^2_{yB} + (N - N_I) \cdot S^2_{yW}$$

donde $S^2_{yB} = \frac{1}{N_I - 1} \sum_{i \in U_I} N_i \cdot (\bar{y}_{U_i} - \bar{y}_U)^2$ y $S^2_{yW} = \frac{1}{N - N_I} \sum_{i \in U_I} \sum_{k \in U_i} (y_k - \bar{y}_{U_i})^2$

Comentario 24. Dentro del muestreo por conglomerados, se puede disponer de conglomerados con el mismo número de elementos, que denotaremos por $N_i = M$ para todo i, o bien conglomerados con distintos tamaños. Generalmente los conglomerados de individuos en la práctica no se ajustan al primer caso (conglomerados de tamaños iguales), pero podría aparecer en el muestreo agrícola. Es el caso más sencillo. Se usan los resultados del muestreo aleatorio simple con los totales por conglomerado como observaciones. El inconveniente es que casi siempre el muestreo por conglomerados con igual tamaño proporciona una menor precisión para los estimadores que en el caso de una muestra aleatoria simple con el mismo número de elementos. ■

Conglomerados del mismo tamaño

En muestreo por conglomerados del mismo tamaño M podemos estimar la media poblacional $\bar{y}_U$, dividiendo el estimador del total entre el número de elementos, con lo que obtenemos:

$$\widehat{\bar{y}}_U^{\text{HT}} = \frac{\widehat{Y}_U^{\text{HT}}}{N_I M}. \tag{4.13}$$

La varianza viene dada por

$$\mathbb{V}_{1st}\left(\widehat{\bar{y}}_U^{\text{HT}}\right) = \frac{1-f_I}{n_I}\frac{S^2_{YU_I}}{M^2}. \tag{4.14}$$

Un estimador insesgado de la varianza es

$$\widehat{\mathbb{V}}\left(\widehat{\bar{y}}_U^{\text{HT}}\right) = \frac{1-f_I}{n_I}\frac{S^2_{Ys_I}}{M^2}. \tag{4.15}$$

Ejemplo 29. Un estudiante quiere estimar las calificaciones promedio de sus compañeros de residencia. En vez de obtener una lista de todos los alumnos que están en su residencia y obtener una muestra aleatoria simple, observa que dicha residencia consta de 100 dormitorios, cada uno con cuatro estudiantes. Elige cinco dormitorios al azar y pregunta a cada persona por sus calificaciones. Los resultados son los siguientes:

Número de persona	Dormitorio (Conglomerado) 1	2	3	4	5
1	3.08	2.36	2.00	3.00	2.68
2	2.60	3.04	2.56	2.88	1.92
3	3.44	3.28	2.52	3.44	3.28
4	3.04	2.68	1.88	3.64	3.20
Total	12.16	11.36	8.96	12.96	11.08

Las unidades primarias son los dormitorios, de modo que $N_I = 100$, $n_I = 5$ y $M = 4$. La estimación del total de la población (la suma estimada de todas las calificaciones de todos los alumnos que pertenecen a la residencia, cantidad sin sentido en este ejemplo, pero útil para demostrar el procedimiento) es:

$$\widehat{Y}_U^{\text{HT}} = \frac{100}{5}(12{,}16 + 11{,}36 + 8{,}96 + 12{,}96 + 11{,}08) = 1130{,}4$$

y

$$S^2_{Ys_I} = \frac{1}{5-1}[(12{,}16 - 11{,}304)^2 + ... + (11{,}08 - 11{,}304)^2] = 2{,}256.$$

En este ejemplo, $S^2_{Ys_I}$ es solo la (cuasi)varianza muestral de los totales de los 5 dormitorios. Por tanto, usando (4.13) y (4.15) obtenemos

$$\widehat{\bar{y}}_U^{\text{HT}} = \frac{1130{,}4}{400} = 2{,}826$$

y

$$\widehat{\mathbb{V}}\left(\widehat{\bar{y}}_U^{\text{HT}}\right) = \frac{1-\frac{5}{100}}{5}\frac{2{,}256}{4^2} = 0{,}027.$$

Obsérvese que en estos cálculos sólo se utiliza los valores de la última fila 'Total' de la tabla de datos, las calificaciones individuales se emplean sólo para calcular el total de cada dormitorio. ■

Analizamos ahora más detenidamente el diseño $1st$ y lo comparamos con una selección directa aleatoria simple de elementos. Para ello será útil trabajar con el *coeficiente de correlación intraconglomerados*.

Definición 24

El coeficiente de correlación intraconglomerados poblacional ρ (Lohr 2010) se define como el coeficiente de correlación para las $N_I M(M-1)$ parejas de valores (y_k, y_l) de un mismo conglomerado, con $k \neq l$:

$$\rho = \frac{\sum_{i \in U_I} \left[\sum_{k \in U_i} \sum_{\substack{l \in U_i \\ l \neq k}} (y_k - \bar{y}_U)(y_l - \bar{y}_U) \right]}{(N_I M - 1)(M-1) S_{yU}^2}. \tag{4.16}$$

El coeficiente ρ se puede interpretar como una medida de homogeneidad dentro de los conglomerados y se puede escribir de forma alternativa como

$$\rho = 1 - \frac{M}{M-1} \cdot \frac{SSW}{SST} \tag{4.17}$$

Se puede demostrar que

$$-\frac{1}{M-1} \leq \rho \leq 1.$$

En un extremo, si $SSW = 0$ entonces $\rho = 1$, que se da cuando los conglomerados son perfectamente homogéneos. En el otro extremo, si $SSB = 0$ entonces $\rho = -\frac{1}{M-1}$, que sucede cuando existe heterogeneidad completa dentro de los conglomerados (situación ideal).

Se puede observar que la variabilidad del estimador de HT para el total poblacional de y depende completamente de la parte de la variabilidad que existe entre los conglomerados, ya que

$$\begin{aligned} S_{Y U_I}^2 &= \frac{1}{N_I - 1} \cdot \sum_{i \in U_I} (Y_{U_i} - \bar{y}_{U_I})^2 = \frac{1}{N_I - 1} \cdot \sum_{i \in U_I} \left(M \cdot \bar{y}_{U_i} - M \cdot \frac{\bar{y}_{U_I}}{M} \right)^2 = \\ &= \frac{M^2}{N_I - 1} \cdot \sum_{i \in U_I} (\bar{y}_{U_i} - \bar{y}_U)^2 = M \cdot S_{yB}^2 \end{aligned}$$

Por tanto,

$$\mathbb{V}_{1st}\left(\widehat{Y}_U^{HT}\right) = N_I^2 \cdot \frac{1-f_I}{n_I} \cdot M \cdot S_{yB}^2$$

Mientras que si se hubiera obtenido una muestra aleatoria simple de $n_I \cdot M$ elementos, entonces:

$$\mathbb{V}_{srswor} = \mathbb{V}_{srswor}\left(\widehat{Y}_U^{HT}\right) = N_I^2 \cdot M^2 \cdot \frac{(1-f_I)}{n_I \cdot M} \cdot S_{yU}^2 = N_I^2 \cdot \frac{(1-f_I)}{n_I} \cdot M \cdot S_{yU}^2$$

Si $S_{yB}^2 > S_{yU}^2$, entonces el muestreo por conglomerados es menos eficiente que el muestreo aleatorio simple.

Por otra parte, la expresión de la varianza del estimador dada en (4.10) también puede escribirse en función de ρ usando (4.17), ya que se puede demostrar:

$$S_{yB}^2 = \frac{N_I \cdot M - 1}{M \cdot (N_I - 1)} \cdot S_{yU}^2 \cdot [1 + (M-1) \cdot \rho]$$

Luego,

$$\mathbb{V}_{1st} = \mathbb{V}_{1st}\left(\widehat{Y}_U^{HT}\right) = N_I^2 \cdot \frac{(1-f_I)}{n_I} \cdot M \cdot \frac{N_I \cdot M - 1}{M \cdot (N_I - 1)} \cdot S_{yU}^2 \cdot [1 - (M-1) \cdot \rho]$$

Así tenemos:

$$\frac{\mathbb{V}_{1st}}{\mathbb{V}_{srswor}} = \frac{S_{yB}^2}{S_{yU}^2} = \frac{N_I \cdot M - 1}{M \cdot (N_I - 1)} \cdot [1 + (M-1) \cdot \rho]$$

Si el número de unidades de primera etapa N_I que pertenecen a la población es tan grande que $N_I M - 1 \approx M \cdot (N_I - 1)$ entonces el cociente de varianzas es aproximadamente $1 + (M-1) \cdot \rho$. Por tanto, cuando ρ es positivo, el muestreo por conglomerados es menos eficiente que el muestreo aleatorio simple. Sin embargo, si ρ es negativo el muestreo por conglomerados es más eficiente.

Comentario 25. Como se ha comentado, ρ proporciona una medida de homogeneidad dentro de los conglomerados. Para que el muestreo por conglomerados sea más eficiente que un muestreo aleatorio simple, nos interesa que las subpoblaciones no contengan elementos similares en su interior. Sin embargo, en los conglomerados que aparecen de manera natural en la población, normalmente ρ es positivo, pues los elementos de un mismo conglomerado tienden a ser más parecidos que elementos elegidos al azar entre la población. ρ es negativo cuando los elementos de un mismo conglomerado están más dispersos de lo que se dispersaría un grupo elegido al azar. Esto puede ocurrir por ejemplo en algunas muestras sistemáticas o en conglomerados artificiales. ■

Comentario 26. Se puede observar que el muestreo sistemático estudiado en el tema 2 se corresponde formalmente con un diseño muestral $1st$ con $n_I = 1$, donde los N_I conglomerados se corresponden con las a posibles muestras sistemáticas. Además, el muestreo sistemático con m arranques aleatorios (véase la sección 2.6.2) se puede

considerar como un diseño muestral $1st$ con $n_I = m$ y $N_I = ma$. En este caso, la ecuación (4.12) proporciona un estimador insesgado de la varianza para el estimador de Horvitz-Thompson del total poblacional. ■

Conglomerados de distinto tamaño

En las encuestas sociales es raro que los conglomerados tengan todos el mismo tamaño. La diferencia entre conglomerados con el mismo o distinto tamaño es que es probable que la variación entre los totales de los conglomerados Y_{U_i} individuales sea grande cuando los conglomerados tengan distinto tamaño. Así, esperaríamos que Y_{U_i} sea grande cuando el tamaño del conglomerado N_i fuese grande y que sea pequeño cuando N_i también lo sea. Con frecuencia, $\widehat{\mathbb{V}}[\widehat{Y}_U^{\mathrm{HT}}]$ es mayor en una muestra por conglomerados cuando las PSUs tienen distintos tamaños que cuando todas las PSUs tienen el mismo número de SSUs.

En muestreo por conglomerados de distinto tamaño no podremos usar el coeficiente de correlación intraconglomerados ρ, ya que solo está definido para conglomerados del mismo tamaño. En su lugar se usará la misma medida de homogeneidad estudiada para medir el grado de homogeneidad dentro de las muestras sistemáticas en el tema 2:

$$\delta = 1 - \frac{N-1}{N-N_I} \cdot \frac{SSW}{SST} = 1 - \frac{S_{yW}^2}{S_{yU}^2}$$

Se puede demostrar que

$$-\frac{N_I - 1}{N - N_I} \leq \delta \leq 1.$$

En un extremo, si $SSW = 0$ entonces $\delta = 1$, que se da cuando hay homogeneidad completa dentro de los conglomerados, esto es, cuando la variación es cero dentro de cada conglomerado. En el otro extremo, si $SSB = 0$, esto es, las medias de los conglomerados son iguales, entonces $\delta = -\frac{N_I - 1}{N - N_I}$.

Comentario 27. Un valor pequeño de δ significa que los elementos en el mismo conglomerado son diferentes con respecto a la variable de estudio, es decir, tienen un bajo grado de homogeneidad. Un valor grande de δ significa que los elementos en el mismo conglomerado son similares, es decir, tienen un alto grado de homogeneidad. ■

Comentario 28. El lector familiarizado con el análisis de regresión identificará δ como el coeficiente de determinación ajustado por los grados de libertad, que a menudo se denota por R_{adj}^2, cuando se ajusta la regresión lineal de y sobre N_I variables dummy (indicando la pertenencia al conglomerado) de la población entera de N datos. ■

Denotemos por $\bar{N} = \frac{N}{N_I}$ el número medio de elementos por conglomerados, sea

$$K_I = \frac{N_I^2(1 - f_I)}{n_I},$$

y sea

$$Cov = \frac{1}{N_I - 1} \sum_{i \in U_I} (N_i - \bar{N}) N_i \bar{y}_{U_i}^2 \tag{4.18}$$

la covarianza entre N_i y $N_i \bar{y}_{U_i}^2$. Se verifica fácilmente que

$$S_{YU_I}^2 = \bar{N} S_{yU}^2 \left(1 + \frac{N - N_I}{N_I - 1}\delta\right) + Cov \tag{4.19}$$

Si introducimos esta expresión en (4.10), que denominamos V_{1st}, nos da

$$V_{1st} = \left(1 + \frac{N - N_I}{N_I - 1}\delta\right) \bar{N} K_I S_{yU}^2 + K_I Cov. \tag{4.20}$$

El número esperado de elementos observados bajo $1st$, con n_I conglomerados seleccionados a partir de N_I, es

$$\mathbb{E}_{1st}(n_s) = n_I \bar{N} = n.$$

Para obtener una comparación justa, consideremos el muestreo aleatorio simple sin reemplazamiento directo, con el tamaño muestral (fijo) $n = n_I \bar{N}$. El estimador de Horvitz-Thompson de Y_U es entonces $N\bar{y}_s$, y la varianza es

$$V_{srswor} = \mathbb{V}_{srswor}(N\bar{y}_s) = \bar{N} K_I S_{yU}^2.$$

Por tanto, una tercera expresión para V_{1st} es

$$V_{1st} = \left(1 + \frac{N - N_I}{N_I - 1}\delta\right) V_{srswor} + K_I Cov, \tag{4.21}$$

a través de la cual obtenemos que el cociente de varianzas es:

$$\frac{V_{1st}}{V_{srswor}} = 1 + \frac{N - N_I}{N_I - 1}\delta + \frac{Cov}{\bar{N} S_{yU}^2}. \tag{4.22}$$

Estas expresiones dan lugar a algunas conclusiones interesantes sobre la eficiencia del muestreo $1st$.

Caso 1

Supongamos que todos los tamaños de los conglomerados N_i son iguales, es decir, $N_i = \bar{N}$ para cada i. En este caso, $Cov = 0$, y obtenemos

$$\frac{V_{1st}}{V_{srswor}} = 1 + \frac{N - N_I}{N_I - 1}\delta \approx 1 + (\bar{N} - 1)\delta. \tag{4.23}$$

Esto muestra que $V_{1st} < V_{srswor}$ si y solo si $\delta < 0$, es decir, si y solo si hay una variación suficientemente grande dentro del conglomerado. Sin embargo, muchos conglomerados

con los que trabajamos en la práctica están formados por elementos 'próximos', y, como estos elementos tienen a asemejarse entre sí más o menos, es probable que $\delta > 0$.

En consecuencia, V_{1st} es mayor que V_{srswor}, a menudo considerablemente mayor. Por ejemplo, incluso si δ es positivo pero muy cercano a cero, digamos por ejemplo que $\delta = 0{,}08$, tenemos, con un tamaño medio de conglomerado de $\bar{N} = 300$,

$$\frac{V_{1st}}{V_{srswor}} \approx 25.$$

Esto muestra una gran pérdida de eficiencia debida al muestreo por conglomerado, porque el tamaño medio del conglomerado es bastante grande en este caso.

Caso 2

Supongamos que los conglomerados varían en tamaño. Si la correlación entre N_i y $A_i = N_i \bar{y}_{U_i}^2$ es positiva, como suele ocurrir a menudo, la varianza aumenta debido a que la selección de conglomerados puede ser peor que en el Caso 1, ya que el segundo término en la fórmula (4.21) puede ser grande.

Para destacar el efecto de la variación en el tamaño de los conglomerados, consideremos el caso extremo de homogeneidad mínima, es decir, $\delta = \delta_{min} = -\frac{N_I - 1}{N - N_I}$. En este caso, todos los $\bar{y}_{U_i}$ son iguales a $\bar{y}_U$ y (4.21) se puede escribir como

$$V_{1st} = \bar{y}_U^2 K_I S_{NU_I}^2 \tag{4.24}$$

que será mayor si la varianza del tamaño de conglomerado

$$S_{NU_I}^2 = \frac{1}{N_I - 1} \sum_{i \in U_I} (N_i - \bar{N})^2$$

es grande.

En este caso

$$\frac{V_{1st}}{V_{srswor}} = \bar{N} \left(\frac{cv_n}{cv_y} \right),$$

donde los dos coeficientes de variación son $cv_n = \dfrac{S_{NU_i}}{\bar{N}}$ y $cv_y = \dfrac{S_{yU}}{\bar{y}_U}$. El cociente $\dfrac{V_{1st}}{V_{srswor}}$ puede ser mayor que la unidad, especialmente si $\bar{N}$ es grande.

Esta discusión muestra que la estrategia $(1st, \widehat{Y}_U^{\mathrm{HT}})$ es probable que sea ineficiente en muchas situaciones, especialmente si los conglomerados son homogéneos y/o de distintos tamaños. Sin embargo, desde un punto de vista del coste/eficiencia, la estrategia $(1st, \widehat{Y}_U^{\mathrm{HT}})$ puede tener ventajas, ya que a menudo es más barato encuestar conglomerados de elementos que encuestar a una muestra dispersa geográficamente que se puede obtener a partir de una selección aleatoria simple de elementos.

Sin embargo, la eficiencia del muestreo por conglomerados se puede mejorar cuando se dispone de información auxiliar. La elección de la estrategia, entonces, depende de la información disponible.

Un caso sencillo surge cuando una medida aproximada u_i del tamaño Y_{U_i} está disponible para cada conglomerado $i = 1, ..., N_I$. Si u_i es aproximadamente proporcional a Y_{U_i} se puede reducir la varianza del estimador de Horvitz-Thompson usando muestreo por conglomerados con probabilidades de inclusión $\pi_{Ii} \propto u_i$. Una alternativa es usar muestreo por conglomerados estratificado con estratos de conglomerados formados de forma que la variación de u_i sea pequeña en cada estrato. Otra alternativa es usar el estimador de razón considerando los tamaños de los conglomerados como variable auxiliar.

Comentario 29. En el caso de conglomerados de distinto tamaño, el estimador del total poblacional no plantea problemas, pero sí el de la media poblacional. El estimador insesgado HT de la media poblacional es

$$\widehat{\bar{y}}_U^{\text{HT}} = \frac{\widehat{Y}_U^{\text{HT}}}{N} = \frac{N_I \bar{y}_{s_I}}{N}$$

Para poder obtener una estimación necesitamos conocer N, pero con frecuencia sólo es posible conocer el tamaño de los conglomerados que están en la muestra y no del resto, ya que el muestreo por conglomerados se utiliza cuando no se dispone de marco poblacional o es muy difícil o costoso obtenerlo. Este método de construcción no puede aplicarse y debe construirse también un estimador para N, lo que conduce al uso del estimador de razón

$$\widehat{R} = \widehat{\bar{y}}_U^{Rat} = \frac{\sum\limits_{i \in s_I} Y_{U_i}}{\sum\limits_{i \in s_I} N_i}$$

que utiliza como información auxiliar los tamaños de los conglomerados. ■

Uso del estimador de razón cuando los conglomerados son de distinto tamaño

En el caso de muestreo por conglomerados con probabilidades idénticas cuando se dispone de conglomerados de distinto tamaño, se hará uso del estimador de razón estudiado en el tema 2 para obtener una estimación de la media poblacional y del error de muestreo.

Ya se ha comentado que el estimador de razón puede ser una alternativa al estimador de Horvitz-Thompson para mejorar la eficiencia del muestreo por conglomerados con probabilidades idénticas cuando los conglomerados son de distinto tamaño.

Teorema 22

En un muestreo monoetápico de elementos, un estimador aproximadamente insesgado para la media poblacional viene dado por

$$\widehat{\bar{y}}_U^{\text{Rat}} = \widehat{R} = \frac{\sum\limits_{i \in s_I} Y_{U_i}}{\sum\limits_{i \in s_I} N_i}, \tag{4.25}$$

La varianza aproximada de $\widehat{\bar{y}}_U^{\text{Rat}}$ es

$$\mathbb{V}\left[\widehat{\bar{y}}_U^{\text{Rat}}\right] \approx \frac{1}{\bar{N}^2} \cdot \frac{1-f_I}{n_I} \cdot \frac{1}{N_I - 1} \sum_{i \in U_I} (Y_{U_i} - R \cdot N_i)^2 \tag{4.26}$$

donde $\bar{N} = \frac{1}{N_I} \sum_{i \in U_I} N_i = \frac{N}{N_I}$

El estimador aproximadamente insesgado de la varianza es

$$\widehat{\mathbb{V}}\left[\widehat{\bar{y}}_U^{\text{Rat}}\right] \approx \frac{1}{\left(\frac{1}{n_I} \sum_{i \in s_I} N_i\right)^2} \cdot \frac{1-f_I}{n_I} \cdot \frac{1}{n_I - 1} \sum_{i \in s_I} (Y_{U_i} - \widehat{R} \cdot N_i)^2 \tag{4.27}$$

Demostración 22

La demostración es una aplicación directa del estimador de razón (véase el Teorema 16).

Comentario 30. Nótese que la razón poblacional es la media poblacional de la variable y.

$$R = \frac{\sum_{i \in U_I} Y_{U_i}}{\sum_{i \in U_I} N_i} = \bar{y}_U$$

■

También podría usarse el estimador de razón para estimar el total poblacional

$$\widehat{Y}_U^{\text{Rat}} = N \cdot \widehat{R} \tag{4.28}$$

Sin embargo, obsérvese que el estimador requiere conocer el número total de elementos que existen en la población, mientras que para el estimador HT no.

4.2.3 Muestreo por conglomerados en dos etapas

En el muestreo por conglomerados monoetápico sucede en general que la varianza del estimador HT es mayor que en el caso de un muestreo aleatorio simple sin repetición. Esto se debe a (i) la tendencia habitual de que los elementos de un mismo conglomerado son parecidos entre ellos[4] y (ii) la variación en el tamaño de los conglomerados. La varianza del estimador HT bajo muestreo de conglomerados monoetápico siempre se puede reducir seleccionando más y más conglomerados. Sin embargo, esto conlleva un mayor coste de recogida de datos por ser una muestra mayor resultando a menudo inadmisible debido a las restricciones presupuestarias. Además, al obtener una muestra

[4]Por ejemplo, porque las personas que viven en un mismo área tienen características similares.

de todos los elementos del mismo conglomerado, repetimos parcialmente la misma información en lugar de conseguir información nueva y esto implica una menor precisión para las estimaciones de las variables objetivo de la población.

Para mantener bajo control el coste y, al mismo tiempo, aumentar el número de conglomerados seleccionados, podemos extraer una submuestra en los conglomerados seleccionados en lugar de entrevistar a todas las unidades de los conglomerados seleccionados. En tal caso, a continuación debemos estimar el total poblacional de cada conglomerado Y_{U_i} a partir de las submuestras. Si la variación dentro del conglomerado es pequeña, las estimaciones $\widehat{Y}_{U_i}^{\text{HT}}$ tendrán una varianza pequeña, incluso para tamaños de submuestras relativamente moderados. En tal caso, merece la pena usar muestreo bietápico en lugar de muestreo unietápico.

Definición 25. Muestreo bietápico

Un diseño muestral de elementos bietápico se define mediante dos etapas de muestreo:

Primera etapa: Se selecciona una muestra s_I de PSUs de U_I ($s_I \subset U_I$) de acuerdo con el diseño $p_I(\cdot)$.
Segunda etapa: Para cada $i \in s_I$, se selecciona una muestra de s_i elementos de U_i ($s_i \subset U_i$) de acuerdo con un diseño $p_i(\cdot|s_I)$ **invariante** e **independiente**.

Comentario 31. La invarianza de los diseños de segunda etapa $p_i(\cdot|s_I)$ quiere decir que, para cada $i \in U_I$ y cada $s_I \subset U_I$, se cumple

$$p_i(\cdot|s_I) = p_i(\cdot).$$

En otras palabras, cada vez que un conglomerado U_i es seleccionado en la primera etapa, debe emplearse el mismo diseño de submuestreo $p_i(\cdot)$. Por ejemplo, se puede establecer que se realice un muestreo aleatorio simple de tamaño muestral n_i cada vez que el conglomerado U_i es seleccionado. ■

Comentario 32. La independencia de los diseños de segunda etapa $p_i(\cdot|s_I)$ quiere decir que, para cada cada $s_I \subset U_I$, se cumple

$$\mathbb{P}\left(\bigcup_{i\in s_I} s_i|s_I\right) = \prod_{i\in s_i} \mathbb{P}\left(s_i|s_I\right).$$

En otras palabras, el submuestreo en cada conglomerado se efectúa independientemente del submuestreo en el resto de conglomerados. ■

La muestra de elementos resultante, denotada por s, está compuesta por $n_I = |s_I|$ submuestras $s = \bigcup_{i\in s_I} s_i$ y, por tanto, el tamaño muestral final será la suma de los

tamaños muestrales de cada submuestra $n = \sum_{i \in s_I} n_i$.

Debemos fijar también la notación para las probabilidades de inclusión asociadas al muestreo de elementos bietápico. Para el diseño de primera etapa $p_I(\cdot)$ denotaremos las probabilidades de inclusión como π_{Ii} y, π_{Iij}. Denotaremos también $\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$, con $\Delta_{Iii} = \pi_{Ii}(1 - \pi_{Ii})$, y $\check{\Delta}_{Iij} = \frac{\Delta_{Iij}}{\pi_{Iij}}$.

De modo similar, para el diseño de segunda etapa $p_i(\cdot)$, usaremos la notación $\pi_{k|i}$ y $\pi_{kl|i}$. Las cantidades Δ son $\Delta_{kl|i} = \pi_{kl|i} - \pi_{k|i}\pi_{l|i}$, con $\Delta_{kk|i} = \pi_{k|i}(1 - \pi_{k|i})$ y, finalmente, $\check{\Delta}_{kl|i} = \frac{\Delta_{kl|i}}{\pi_{kl|i}}$.

Estimadores, varianza y estimador de la varianza

Para obtener el estimador HT, su varianza y el estimador HT de esta varianza podemos usar el resultado general sobre la construcción del estimador HT con las probabilidades de inclusión π_k y π_{kl} particulares del muestreo bietápico. Se sigue de las propiedades de invarianza y de independencia que las probabilidades de inclusión de los elementos son

$$\pi_k = \pi_{Ii}\pi_{k|i} \quad \text{si } k \in U_i \tag{4.29a}$$

y

$$\pi_{kl} = \begin{cases} \pi_{Ii}\pi_{k|i} & \text{si } k = l \in U_i, \\ \pi_{Ii}\pi_{kl|i} & \text{si } k, l \in U_i, k \neq l, \\ \pi_{Iij}\pi_{k|i}\pi_{l|j} & \text{si } k \in U_i \text{ y } l \in U_j. \end{cases} \tag{4.29b}$$

Teorema 23

El estimador HT, su varianza y el estimador HT de esta varianza para un muestreo de elementos bietápico están dados por

i. $\widehat{Y}_U^{\text{HT}} = \sum_{k \in U} \frac{y_k}{\pi_k}$,

ii. $\mathbb{V}\left[\widehat{Y}_U^{\text{HT}}\right] = \sum_{k,l \in U} (\pi_{kl} - \pi_k\pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$,

iii. $\widehat{\mathbb{V}}^{\text{HT}}\left[\widehat{Y}_U^{\text{HT}}\right] = \sum_{k,l \in U} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$,

donde las probabilidades de inclusión están dados por (4.29).

Demostración 23

Es una aplicación directa del Teorema de Horvitz-Thompson. ■

Aunque este resultado nos permite computar directamente las estimaciones derivadas

de este diseño muestral, el Teorema 23 no nos permite comprender cómo influye cada etapa en la construcción de las estimaciones. Para ello, debemos recurrir a un resultado general de la teoría de probabilidad (véase p.ej. Grimmet y Stirzaker 2004, pág. 69). Sean X, Y variables aleatorias. Se cumplen:

- $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X|Y]]$.
- $\mathbb{V}[X] = \mathbb{V}_Y[\mathbb{E}[X|Y]] + \mathbb{E}_Y[\mathbb{V}[X|Y]]$.

En el muestreo bietápico, condicionamos sobre el suceso de que la muestra s_I se seleccione en la primera fase. Sea $\check{y}_{k|i} = \frac{y_k}{\pi_{k|i}}$ y sea[5]

$$\widehat{Y}^{\text{HT}}_{U_i|i} = \sum_{k \in s_i} \check{y}_{k|i} \tag{4.30}$$

el estimador HT con respecto a la segunda etapa del total poblacional de la PSU $Y_{U_i} = \sum_{k \in U_i} y_k$.

En caso de submuestrear U_i repetidamente de acuerdo con el diseño $p_i(\cdot)$, $\widehat{Y}^{\text{HT}}_{U_i|i}$ es un estimador insesgado de Y_{U_i}, esto es, es insesgado condicionalmente a seleccionar s_i en la primera etapa. La varianza con respecto a la segunda etapa es

$$V_i \equiv \mathbb{V}\left[\widehat{Y}^{\text{HT}}_{U_i|i}\right] = \sum_{k \in U_i} \sum_{l \in U_i} \Delta_{kl|i}\ \check{y}_{k|i}\ \check{y}_{l|i} \tag{4.31}$$

cuyo estimador insesgado es

$$\widehat{V}_i \equiv \widehat{\mathbb{V}}^{\text{HT}}\left[\widehat{Y}^{\text{HT}}_{U_i|i}\right] = \sum_{k \in s_i} \sum_{l \in s_i} \check{\Delta}_{kl|i}\ \check{y}_{k|i}\ \check{y}_{l|i}. \tag{4.32}$$

Como sucede con el muestreo directo de elementos, se pueden emplear fórmulas alternativas para diseños de tamaño muestral fijo. Si el diseño $p_i(\cdot)$ es de tamaño fijo, V_i también se puede escribir como

$$V_i = -\frac{1}{2} \sum_{k \in U_i} \sum_{l \in U_i} \Delta_{kl|i}\ (\check{y}_{k|i} - \check{y}_{l|i})^2 \tag{4.33}$$

cuyo estimador insesgado viene dado por

$$\widehat{V}_i = -\frac{1}{2} \sum_{k \in s_i} \sum_{l \in s_i} \check{\Delta}_{kl|i}\ (\check{y}_{k|i} - \check{y}_{l|i})^2. \tag{4.34}$$

Para las estimaciones HT en el muestreo bietápico, deben combinarse las aportaciones de ambas etapas.

[5] Adviértase la diferencia entre $\widehat{Y}^{\text{HT}}_{U_i}$ y $\widehat{Y}^{\text{HT}}_{U_i|i}$. El primero es el estimador HT del total poblacional de la variable y en el conglomerado U_i, que se construye con las probabilidades de inclusión π_k, $k \in s_i$. El segundo es el estimador HT del total poblacional de la variable y en el conglomerado U_i, condicionado a que se ha escogido el conglomerado U_i en la primera etapa, por tanto, se construye con las probabilidades inclusión $\pi_{k|i}$.

Teorema 24

En el muestreo de elementos bietápico, el estimador HT del total poblacional $Y_U = \sum_{k\in U} y_k$ viene dado por

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{i\in s_I} \frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}} = \sum_{i\in s_I} \frac{1}{\pi_{Ii}} \sum_{k\in s_i} \frac{y_k}{\pi_{k|i}} \tag{4.35}$$

La varianza de $\widehat{Y}_U^{\mathrm{HT}}$ se puede escribir como la suma de dos componentes,

$$V_{2st} \equiv \mathbb{V}\left[\widehat{Y}_U^{\mathrm{HT}}\right] = V_{PSU} + V_{SSU} \tag{4.36}$$

donde

$$V_{PSU} = \sum_{i\in U_I} \sum_{j\in U_I} \Delta_{Iij} \frac{Y_{U_i}}{\pi_{Ii}} \frac{Y_{U_j}}{\pi_{Ij}}, \tag{4.37a}$$

$$V_{SSU} = \sum_{i\in U_I} \frac{V_i}{\pi_{Ii}} \equiv \sum_{i\in U_I} \frac{1}{\pi_{Ii}} \sum_{k\in U_i} \sum_{l\in U_i} \Delta_{kl|i}\, \breve{y}_{k|i}\, \breve{y}_{l|i} \tag{4.37b}$$

El estimador HT de la varianza V_{2st} se construye estimando cada componente por separado de modo insesgado:

$$\begin{aligned} \widehat{V}_{PSU} &= \sum_{i\in s_I} \sum_{j\in s_I} \breve{\Delta}_{Iij} \frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}} \frac{\widehat{Y}_{U_j|j}^{\mathrm{HT}}}{\pi_{Ij}} - \sum_{i\in s_I} \frac{1}{\pi_{Ii}} \left(\frac{1}{\pi_{Ii}} - 1\right) \widehat{V}_i \\ &\equiv \sum_{i\in s_I} \sum_{j\in s_I} \breve{\Delta}_{Iij} \frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}} \frac{\widehat{Y}_{U_j|j}^{\mathrm{HT}}}{\pi_{Ij}} \\ &\quad - \sum_{i\in s_I} \frac{1}{\pi_{Ii}} \left(\frac{1}{\pi_{Ii}} - 1\right) \sum_{k\in s_i} \sum_{l\in s_i} \breve{\Delta}_{kl|i}\, \breve{y}_{k|i}\, \breve{y}_{l|i}, \end{aligned} \tag{4.38a}$$

$$\widehat{V}_{SSU} = \sum_{i\in s_I} \frac{\widehat{V}_i}{\pi_{Ii}^2} = \sum_{k\in s_i} \frac{1}{\pi_{Ii}^2} \sum_{k\in s_i} \sum_{l\in s_i} \breve{\Delta}_{kl|i}\, \breve{y}_{k|i}\, \breve{y}_{l|i}. \tag{4.38b}$$

El estimador HT de $\widehat{Y}_U^{\mathrm{HT}}$ está dado por:

$$\widehat{V}_{2st} \equiv \widehat{\mathbb{V}}^{\mathrm{HT}}\left[\widehat{Y}_U^{\mathrm{HT}}\right] = \widehat{V}_{PSU} + \widehat{V}_{SSU} = \sum_{i\in s_I} \sum_{j\in s_I} \breve{\Delta}_{Iij} \frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}} \frac{\widehat{Y}_{U_j|j}^{\mathrm{HT}}}{\pi_{Ij}} + \sum_{i\in s_I} \frac{\widehat{V}_i}{\pi_{Ii}}. \tag{4.39}$$

Comentario 33. Para construir la demostración del Teorema 24 simplificamos la notación, permitiendo así su generalización al caso multietápico:

$$\begin{aligned}
\mathbb{E}_I\mathbb{E}_{II}\left[\widehat{Y}_U^{\mathrm{HT}}\right] &= \mathbb{E}_{p_I}\left[\mathbb{E}\left[\widehat{Y}_U^{\mathrm{HT}}|s_I\right]\right], \\
\mathbb{V}_I\mathbb{E}_{II}\left[\widehat{Y}_U^{\mathrm{HT}}\right] &= \mathbb{V}_{p_I}\left[\mathbb{E}\left[\widehat{Y}_U^{\mathrm{HT}}|s_I\right]\right], \\
\mathbb{E}_I\mathbb{V}_{II}\left[\widehat{Y}_U^{\mathrm{HT}}\right] &= \mathbb{E}_{p_I}\left[\mathbb{V}\left[\widehat{Y}_U^{\mathrm{HT}}|s_I\right]\right].
\end{aligned}$$

Es decir, el subíndice I indica la esperanza o la varianza con respecto al diseño $p_I(\cdot)$ usado en la primera fase y II indica la esperanza condicionada o la varianza condicionada con respecto al conjunto de diseños $p_i(\cdot)$, $i \in s_I$, usado en la segunda fase, dado s_I. ■

Demostración 24

La ecuación (4.35) equivale al estimador de Horvitz-Thompson toda vez que reconocemos las probabilidades de inclusión de primer orden de cada unidad $k \in U$ de la población (véase la ecuación (4.29a)):

$$\begin{aligned}
\widehat{Y}_U^{\mathrm{HT}} &= \sum_{k\in s}\frac{y_k}{\pi_k} = \sum_{i\in s_I}\sum_{k\in s_i}\frac{y_k}{\pi_{Ii}\pi_{k|i}} \\
&= \sum_{i\in s_I}\frac{\left(\sum_{k\in s_i}\frac{y_k}{\pi_{k|i}}\right)}{\pi_{Ii}} = \sum_{i\in s_I}\frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}}.
\end{aligned}$$

Sin embargo, también es conveniente saber cómo influye cada etapa en la insesgadez del estimador:

$$\begin{aligned}
\mathbb{E}_I\mathbb{E}_{II}\left[\widehat{Y}_U^{\mathrm{HT}}\right] &= \mathbb{E}_{p_I}\left[\mathbb{E}\left[\widehat{Y}_U^{\mathrm{HT}}|s_I\right]\right] \\
&= \mathbb{E}_{p_I}\left[\mathbb{E}\left[\sum_{i\in s_I}\frac{1}{\pi_{Ii}}\sum_{k\in s_i}\frac{y_k}{\pi_{k|i}}|s_I\right]\right] \\
&= \mathbb{E}_{p_I}\left[\sum_{i\in s_I}\frac{1}{\pi_{Ii}}\mathbb{E}\left[\sum_{k\in s_i}\frac{y_k}{\pi_{k|i}}|s_I\right]\right] \\
&= \mathbb{E}_{p_I}\left[\sum_{i\in s_I}\frac{1}{\pi_{Ii}}\left[\sum_{k\in U_i}y_k\right]\right] = \mathbb{E}_{p_I}\left[\sum_{i\in s_I}\frac{Y_{U_i}}{\pi_{Ii}}\right] \\
&= \sum_{i\in U_I}Y_{U_i} \\
&= \sum_{k\in U}y_k = Y_U
\end{aligned}$$

Para la varianza, usamos los momentos condicionales calculados gracias a las propiedades de invarianza e independencia:

$$\mathbb{E}_{II}\left[\widehat{Y}_U^{\mathrm{HT}}\right] = \mathbb{E}\left[\widehat{Y}_U^{\mathrm{HT}}|s_I\right] = \sum_{i\in s_I}\mathbb{E}_{p_i}\left[\frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}}|s_I\right] \underbrace{=}_{\text{invarianza}} \sum_{i\in s_I}\mathbb{E}_{p_i}\left[\frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}}\right] = \sum_{i\in s_I}\frac{Y_{U_i}}{\pi_{Ii}}, \tag{4.40a}$$

$$\mathbb{V}_{II}\left[\widehat{Y}_U^{\mathrm{HT}}\right] = \mathbb{V}\left[\widehat{Y}_U^{\mathrm{HT}}|s_I\right] \underbrace{=}_{\text{independencia}} \sum_{i\in s_I}\mathbb{V}\left[\frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}}|s_I\right] \underbrace{=}_{\text{invarianza}} \sum_{i\in s_I}\mathbb{V}_{p_i}\left[\frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}}\right] = \sum_{i\in s_I}\frac{V_i}{\pi_{Ii}^2}. \tag{4.40b}$$

De este modo, la varianza se escribe inmediatamente como

$$\begin{aligned} V_{2st} &= \mathbb{V}\left[\widehat{Y}_U^{\mathrm{HT}}\right] \\ &= \mathbb{V}_I\mathbb{E}_{II}\left[\widehat{Y}_U^{\mathrm{HT}}\right] + \mathbb{E}_I\mathbb{V}_{II}\left[\widehat{Y}_U^{\mathrm{HT}}\right] \\ &= \mathbb{V}_I\left[\sum_{i\in s_I}\frac{Y_{U_i}}{\pi_{Ii}}\right] + \mathbb{E}_I\left[\sum_{i\in s_I}\frac{V_i}{\pi_{Ii}^2}\right] \\ &= \sum_{i\in U_I}\sum_{j\in U_I}\Delta_{Iij}\frac{Y_{U_i}}{\pi_{Ii}}\frac{Y_{U_j}}{\pi_{Ij}} + \sum_{i\in U_I}\frac{V_i}{\pi_{Ii}} \end{aligned}$$

que demuestra la expresión de la varianza (4.36).

Ahora debemos demostrar las expresiones para la estimación insesgada de la varianza. Procedemos nuevamente analizando cada componente. Por la propiedad de independencia y la definición de V_i, se cumple

$$\mathbb{E}_{II}\left[\widehat{Y}_{U_i|i}^{\mathrm{HT}}\,\widehat{Y}_{U_j|j}^{\mathrm{HT}}\right] = \begin{cases} Y_{U_i}^2 + V_i, & \text{si } i = j, \\ Y_{U_i}Y_{U_j}, & \text{si } i \neq j. \end{cases}$$

Ahora tenemos

$$\begin{aligned} \mathbb{E}\left[\sum_{i\in s_I}\sum_{j\in s_I}\breve{\Delta}_{Iij}\frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}}\frac{\widehat{Y}_{U_j|j}^{\mathrm{HT}}}{\pi_{Ij}}\right] &= \mathbb{E}_I\left[\sum_{i\in s_I}\sum_{j\in s_I}\breve{\Delta}_{Iij}\frac{\mathbb{E}_{II}\left[\widehat{Y}_{U_i|i}^{\mathrm{HT}}\,\widehat{Y}_{U_j|j}^{\mathrm{HT}}\right]}{\pi_{Ii}\pi_{Ij}}\right] \\ &= \mathbb{E}_I\left[\sum_{i\in s_I}\sum_{j\in s_I}\breve{\Delta}_{Iij}\frac{Y_{U_i}}{\pi_{Ii}}\frac{Y_{U_j}}{\pi_{Ij}}\right] + \mathbb{E}_I\left[\sum_{i\in s_I}\breve{\Delta}_{Iii}\frac{V_i}{\pi_{Ii}^2}\right] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\in U_I}\sum_{j\in U_I}\Delta_{Iij}\frac{Y_{U_i}}{\pi_{Ii}}\frac{Y_{U_j}}{\pi_{Ij}} + \sum_{i\in U_I}\frac{\pi_{Ii}-\pi_{Ii}^2}{\pi_{Ii}^2}V_i \\
&= V_{PSU} + \sum_{i\in U_I}\left(\frac{1}{\pi_{Ii}}-1\right)V_i \qquad (4.41a)
\end{aligned}
$$

y

$$
\begin{aligned}
\mathbb{E}\left[-\sum_{i\in s_I}\frac{1}{\pi_{Ii}}\left(\frac{1}{\pi_{Ii}}-1\right)\widehat{V}_i\right] &= -\mathbb{E}_I\left[\sum_{i\in s_I}\frac{1}{\pi_{Ii}}\left(\frac{1}{\pi_{Ii}}-1\right)\mathbb{E}_{II}\left[\widehat{V}_i\right]\right] \\
&= -\mathbb{E}_I\left[\sum_{i\in s_I}\frac{1}{\pi_{Ii}}\left(\frac{1}{\pi_{Ii}}-1\right)V_i\right] \\
&= -\sum_{i\in U_I}\left(\frac{1}{\pi_{Ii}}-1\right)V_i \qquad (4.41b)
\end{aligned}
$$

La suma de (4.41a) y (4.41b) implica que $\mathbb{E}\left[\widehat{V}_{PSU}\right] = V_{PSU}$.

A continuación demostramos que $\widehat{V}_{SSU}$ es insesgado para V_{SSU}:

$$
\begin{aligned}
\mathbb{E}\left[\widehat{V}_{SSU}\right] &= \mathbb{E}_I\left[\sum_{i\in s_I}\frac{\mathbb{E}_{II}\left[\widehat{V}_i\right]}{\pi_{Ii}^2}\right] \\
&= \mathbb{E}_I\left[\sum_{i\in s_I}\frac{V_i}{\pi_{Ii}^2}\right] = \sum_{i\in U_I}\frac{V_i}{\pi_{Ii}} = V_{SSU}.
\end{aligned}
$$

Recopilando tenemos que

$$
\mathbb{E}\left[\widehat{\mathbb{V}}^{\mathrm{HT}}\left[\widehat{Y}_U^{\mathrm{HT}}\right]\right] = \mathbb{E}\left[\widehat{V}_{PSU}+\widehat{V}_{SSU}\right] = V_{PSU}+V_{SSU} = \mathbb{V}\left[\widehat{Y}_U^{\mathrm{HT}}\right]
$$

La demostración queda completada.

Comentario 34. En la práctica muchas encuestas con diseño bietápico hacen uso de los llamados *diseños autoponderados*[6]. Un diseño bietápico se dice que es autoponderado si cada unidad de segunda etapa de la muestra representa al mismo número de unidades secundarias en la población. En ese caso, se dice que la muestra es *autoponderada*. Un diseño que proporciona muestras autoponderadas es el siguiente. Sea u_i una medida del tamaño de la unidad primaria i-ésima. En primera etapa se escoge un diseño muestral proporcional al tamaño de cada unidad primaria de modo que $\pi_{Ii} \propto u_i$ y, en segunda etapa, se escoge un muestreo aleatorio simple con fracción de muestreo en cada conglomerado dada por $\frac{n_i}{N_i} = \frac{1}{u_i}$. Si en la primera etapa se emplea un diseño de

[6] *Self-weighting design.*

tamaño muestral fijo n_I, entonces

$$\pi_{Ii} = \frac{u_i}{\sum\limits_{i \in U_I} u_i} \cdot n_I.$$

Las probabilidades de inclusión de los elementos de la población se reducen entonces a

$$\pi_k = \pi_{Ii}\pi_{k|i} = \frac{u_i}{\sum\limits_{i \in U_I} u_i} \cdot n_I \cdot \frac{n_i}{N_i} = \frac{n_I}{\sum\limits_{i \in U_I} u_i}.$$

El estimador HT se simplifica notablemente:

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k} = \frac{\sum\limits_{i \in U_I} u_i}{n_I} \sum_{k \in s} y_k = \frac{\sum\limits_{i \in U_I} u_i}{n_I} Y_s,$$

esto es, el estimador es proporcional al total muestral de la variable a estimar. Aparte de la simplificación computacional, este tipo de diseños permite un mayor control del trabajo de campo, pues si $\frac{N_i}{u_i} \approx K$, siendo K una constante para todo i, entonces $n_i \approx K$ y el número de entrevistas a realizar en cada conglomerado es básicamente el mismo. Esto permite un mayor control de la logística de la fase de recogida de datos (por ejemplo, asignando un entrevistador por conglomerado de modo que todos tengan la misma carga de trabajo). ■

4.2.4 Muestreo por conglomerados con probabilidades idénticas en dos etapas

Particularicemos ahora los resultados obtenidos para el caso de aplicar muestreo por conglomerados con probabilidades idénticas en dos etapas, es decir, usar el muestreo aleatorio simple sin reemplazamiento en ambas etapas. En la primera etapa se toma una muestra s_I de n_I conglomerados a partir de los N_I existentes y, a continuación, para cada conglomerado $i \in s_I$, se selecciona una muestra de tamaño n_i sobre los N_i elementos del conglomerado. Aplicando el Teorema 24 obtenemos el estimador HT del total poblacional Y_U que puede escribirse como

$$\widehat{Y}_U^{\mathrm{HT}} = \sum_{i \in s_I} \frac{\widehat{Y}_{U_i|i}^{\mathrm{HT}}}{\pi_{Ii}} = \frac{N_I}{n_I} \sum_{i \in s_I} N_i \bar{y}_{s_i}. \tag{4.42}$$

dado que $\pi_{Ii} = \frac{n_I}{N_I}$ para todo $i \in U_I$.

La varianza es

$$V_{2st} = N_I^2 \frac{1 - f_I}{n_I} S_{Y U_I}^2 + \frac{N_I}{n_I} \sum_{i \in U_I} N_i^2 \frac{1 - f_i}{n_i} S_{y U_i}^2 \tag{4.43}$$

donde

- $f_I = \frac{n_I}{N_I}$ es la fracción de muestreo en primera etapa;
- $f_i = \frac{n_i}{N_i}$ es la fracción de muestreo en segunda etapa;
- $S^2_{YU_I} = \frac{1}{N_I - 1} \sum_{i \in U_I} (Y_{U_i} - \bar{y}_{U_I})^2$ es la cuasivarianza en U_I de los totales poblacionales Y_{U_i} de los conglomerados i, con $\bar{y}_{U_I} = \sum_{i \in U_I} \frac{Y_{U_i}}{N_I}$;
- $S^2_{yU_i} = \frac{1}{N_i - 1} \sum_{k \in U_i} (y_k - \bar{y}_{U_i})^2$ es la cuasivarianza de la variable objetivo y en el conglomerado U_i, con $\bar{y}_{U_i} = \sum_{k \in U_i} \frac{y_k}{N_i}$.

El estimador insesgado de la varianza es

$$\widehat{V}_{2st} = N_I^2 \frac{1 - f_I}{n_I} S^2_{\widehat{Y} s_I} + \frac{N_I}{n_I} \sum_{i \in s_I} N_i^2 \frac{1 - f_i}{n_i} S^2_{y s_i} \tag{4.44}$$

donde

- $S^2_{\widehat{Y} s_I} = \frac{1}{n_I - 1} \sum_{i \in s_I} \left[\widehat{Y}^{\text{HT}}_{U_i|i} - \left(\frac{1}{n_I} \sum_{i \in s_I} \widehat{Y}^{\text{HT}}_{U_i|i} \right) \right]^2$ es la cuasivarianza en s_I de los totales estimados $\widehat{Y}^{\text{HT}}_{U_i|i} = N_i \bar{y}_{s_i}$ de los conglomerados i;
- $S^2_{y s_i} = \frac{1}{n_i - 1} \sum_{k \in s_i} (y_k - \bar{y}_{s_i})^2$ es la cuasivarianza de la variable objetivo y en el conglomerado U_i, con $\bar{y}_{s_i} = \frac{1}{n_i} \sum_{k \in s_i} y_k$.

Comentario 35. Se puede observar que las expresiones del estimador, la varianza y el estimador de la varianza para el muestreo estratificado obtenidas en el tema 3 son un caso particular de las fórmulas obtenidas para el muestreo por conglomerados con probabilidades idénticas en dos etapas con $n_I = N_I$. ■

Ejemplo 30. Se dispone de una población constituida por 284 municipios (Apéndice B de Särndal, Swensson y Wretman 1992) de la que se desea estimar el total de escaños en el consejo municipal (variable y). Supóngase que la población está agrupada en $N_I = 50$ conglomerados. Se realiza un muestreo por conglomerados en dos etapas sin reemplazamiento para estimar el total de la variable y.

i	N_i	y_k
19	5	41, 49, 49
45	8	49, 49, 45
47	5	31, 31, 35
50	9	39, 41, 61
31	7	49, 51, 33

En la primera etapa, se extrae una muestra aleatoria simple s_I de tamaño $n_I = 5$. Para cada conglomerado en s_I, se realiza un muestreo aleatorio simple de tamaño $n_i = 3$ de los N_i elementos del i-ésimo conglomerado. Se obtienen los siguientes datos:

Para calcular la estimación a partir del estimador HT y su varianza estimada asociada, primero obtenemos $\widehat{Y}^{\text{HT}}_{U_i|i} = N_i \bar{y}_{s_i}$ y $S^2_{ys_i}$ para cada $i \in s_I$:

i	$\widehat{Y}^{\text{HT}}_{U_i\|i}$	$S^2_{ys_i}$
19	$\frac{695}{3}$	$\frac{64}{3}$
45	$\frac{1144}{3}$	$\frac{16}{3}$
47	$\frac{485}{3}$	$\frac{16}{3}$
50	423	148
31	$\frac{931}{3}$	$\frac{292}{3}$

Usando la ecuación (4.42) llegamos a:

$$\widehat{Y}^{\text{HT}}_U = \frac{50}{5} \cdot \left(\frac{695}{3} + \frac{1144}{3} + \frac{485}{3} + 423 + \frac{931}{3} \right) = 15080$$

y

$$S^2_{\widehat{Y}s_I} = \frac{1}{n_I - 1} \sum_{i \in s_I} \left[\widehat{Y}^{\text{HT}}_{U_i|i} - \left(\frac{1}{n_I} \sum_{i \in s_I} \widehat{Y}^{\text{HT}}_{U_i|i} \right) \right]^2 \approx 11410{,}9$$

Finalmente, la fórmula (4.44) nos lleva a la siguiente estimación de la varianza:

$$\begin{aligned} \widehat{V}_{2st}(\widehat{t}_{pi}) = & \frac{50 \cdot 45}{5} \cdot 11410{,}9 + 10 \cdot \left(\frac{5 \cdot 2}{3} \cdot \frac{64}{3} + \frac{8 \cdot 5}{3} \cdot \frac{16}{3} + \right. \\ & \left. \frac{5 \cdot 2}{3} \cdot \frac{16}{3} + \frac{9 \cdot 6}{3} \cdot 148 + \frac{7 \cdot 4}{3} \cdot \frac{292}{3} \right) \approx 5183960{,}7 \end{aligned}$$

El coeficiente de variación estimado es aproximadamente del 15 %. ■

Comentario 36. El muestreo por conglomerados con probabilidades idénticas en dos etapas produce muestras autoponderadas si $\frac{N_i}{n_i}$ es constante ($N_i = c \cdot n_i$). En efecto:

$$\widehat{Y}^{\text{HT}}_U = \frac{N_I}{n_I} \sum_{i \in s_I} N_i \bar{y}_{s_i} = \frac{N_I}{n_I} \sum_{i \in s_I} \frac{N_i}{n_i} \sum_{k \in s_i} y_k = \frac{N_I}{n_I} c \sum_{k \in s} y_k.$$

■

Uso del estimador de razón cuando los conglomerados son de distinto tamaño

De la misma forma que en muestreo por conglomerados monoetápico, podemos hacer uso del estimador de razón para estimar la media poblacional en un diseño bietápico cuando los tamaños de los conglomerados son distintos y no hay conocimiento del número de elementos de la población N.

Asimismo, puede ser una alternativa al estimador HT, ya que el primer componente de la varianza dada en (4.43) puede ser muy grande cuando los tamaños de los conglomerados son demasiado variables, aunque las medias de los conglomerados sean casi constantes.

Aplicando los resultados estudiados sobre el estimador de razón al caso de los estimadores $\widehat{Y}_U^{\text{HT}}$ y $\widehat{N}_U^{\text{HT}}$ con probabilidades idénticas en primera y segunda etapa, tenemos que, considerando conocidos únicamente el tamaño de los conglomerados N_i para todo $i \in s_I$, un estimador aproximadamente insesgado para la media poblacional en un muestreo bietápico de elementos viene dado por

$$\widehat{\bar{y}}_U^{\text{Rat}} = \frac{\sum_{i \in s_I} N_i \cdot \bar{y}_{U_i}}{\sum_{i \in s_I} N_i}$$

La fórmula de la varianza aproximada del estimador es

$$\mathbb{V}(\widehat{\bar{y}}_U^{\text{Rat}}) \approx \frac{1}{\bar{N}^2} \left[\frac{1 - f_I}{n_I} \frac{\sum_{i \in U_I} (N_i \bar{y}_{U_i} - R N_i)^2}{N_I - 1} + \frac{1}{n_I N_I} \sum_{i \in U_I} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{yU_i}^2}{n_i} \right]$$

El estimador aproximadamente insesgado de la varianza es

$$\widehat{\mathbb{V}}(\widehat{\bar{y}}_U^{\text{Rat}}) \approx \frac{1}{\left(\frac{1}{n_I} \sum_{i \in s_I} N_i\right)^2} \left[\frac{1 - f_I}{n_I} \frac{\sum_{i \in s_I} \left(N_i \bar{y}_{s_i} - \widehat{R} N_i\right)^2}{n_I - 1} + \frac{1}{n_I N_I} \sum_{i \in s_I} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{ys_i}^2}{n_i} \right]$$

4.3 Muestreo con probabilidades diferentes con reemplazo

En esta sección se estudiarán diseños de muestreo en los que las unidades muestrales de primera etapa son seleccionadas mediante un esquema con reemplazamiento. En caso de muestreo en dos etapas, las unidades secundarias se seleccionan mediante un esquema sin reemplazamiento.

Veamos algunos algoritmos de selección de la muestra con reemplazamiento para seleccionar las unidades muestrales de primera etapa. En el tema 2 se estudió un algoritmo

muestral para la selección de una muestra aleatoria simple directa con reemplazo. También existen mecanismos de selección de la muestra con probabilidades diferentes. Un esquema sencillo para seleccionar una muestra de tamaño $n_I = 1$ es usar el *método del total acumulado*. Cuando el tamaño de muestra es $n_I > 1$, un mecanismo de selección de la muestra bajo muestreo con probabilidades proporcionales al tamaño con reemplazo consiste en repetir, de forma independiente, n_I veces el *método del total acumulado*.

A continuación se describe el método de selección sistemática. Este método es muy popular por su simplicidad. Se trata de una generalización del muestreo sistemático estudiado en el tema 2. El algoritmo es el siguiente:

- Se obtienen los tamaños acumulados $T_i = T_{i-1} + N_i$, con $T_0 = 0$.
- Se determina el intervalo muestral a, donde a es un entero positivo. Sea n_I la parte entera de $\frac{T_{N_I}}{a}$, entonces podemos expresar T_{N_I} como:

  $$T_{N_I} = n_I \cdot a + c$$

  donde $0 \leq c < a$.
- Se selecciona un número aleatorio (arranque aleatorio) entre 1 y a con igual probabilidad $\frac{1}{a}$: r.
- La muestra seleccionada es

  $$s_r = \{k : \ T_{i-1} < r + (j-1) \cdot a \leq T_i, \ \text{ para algún } j = 1, 2, \ldots, n_I\}$$

  donde n_I es el tamaño de muestra, que será n_I si $c < r \leq a$ o bien $n_I + 1$ si $r \leq c$.

Ejemplo 31. Continuando con el Ejemplo 28, supongamos ahora que se desea seleccionar una muestra de tamaño $n_I = 4$ unidades primarias mediante el método generalizado de selección sistemática.

Dado que $n_I = 4$, el intervalo muestral a es 25 y $c = 0$:

$$100 = T_{N_I} = n_I \cdot a = 4 \cdot 25$$

Si el arranque aleatorio seleccionado entre 1 y 25 es 4, entonces los conglomerados correspondientes a $4, 29, 54$ y 79 son aquellos que se seleccionan para la muestra. Estos grupos son $1, 3, 3$ y 5, respectivamente.

Obsérvese que este método no proporciona una verdadera muestra aleatoria con reemplazamiento, ya que los grupos 1 y 2 no pueden aparecer ambos en la muestra y es imposible que aparezca alguno de ellos más de una vez. Por otra parte, con probabilidad igual a 1 el grupo 3, 4, 5 o 6 aparecerá en la muestra. A pesar de ello, en muchas ocasiones este método resulta sencillo de implantar respecto a otros que proporcionan una muestra aleatoria. ■

Los diseños de muestreo con reemplazo suelen ser menos eficientes que los diseños sin reemplazo. Sin embargo, suelen utilizarse ya que facilitan la elección y análisis de la muestra, permiten rebajar la carga computacional relacionada con la estimación de la varianza.

4.3.1 Muestreo con probabilidades diferentes en una etapa con reemplazo

Considérese una muestra ordenada $\{i_1, \ldots, i_\nu, \ldots, i_{n_I}\}$ de conglomerados siguiendo un esquema de muestreo con reemplazamiento tal que, en cada extracción, la probabilidad de seleccionar el conglomerado i es p_i, para todo $i \in \{1, \ldots, N_I\}$ donde N_I es el número total de conglomerados existentes en la población. El tamaño de la muestra seleccionada, s_I, se denota por n_I. El total poblacional en el conglomerado U_{i_ν} se denota por $Y_{U_{i_\nu}}$.

Teorema 25

En un muestreo monoetápico con reemplazamiento, un estimador insesgado para el total poblacional Y_U viene dado por

$$\widehat{Y}_U^{HH} = \frac{1}{n_I} \sum_{\nu=1}^{n_I} \frac{Y_{U_{i_\nu}}}{p_{i_\nu}} \tag{4.45}$$

La varianza de $\widehat{Y}_U^{HH}$ está dada por

$$\mathbb{V}\left[\widehat{Y}_U^{HH}\right] = \frac{1}{n_I} \sum_{i \in U_I} p_i \left(\frac{Y_{U_i}}{p_i} - Y_U\right)^2 \tag{4.46}$$

Un estimador insesgado para esta varianza está dado por

$$\widehat{\mathbb{V}}\left[\widehat{Y}_U^{HH}\right] = \frac{1}{n_I(n_I - 1)} \sum_{\nu=1}^{n_I} \left[\frac{Y_{U_{i_\nu}}}{p_{i_\nu}} - \widehat{Y}_U^{HH}\right]^2 \tag{4.47}$$

Demostración 25

Los resultados pueden obtenerse inmediatamente a partir del Teorema 6. ■

Ejemplo 32. Una universidad cuenta con 15 grupos que asisten a una asignatura. Hay $N = 647$ estudiantes en total y cada uno de los grupos cuenta con N_i estudiantes, con N_i conocidos para todo $i = 1, \ldots, 15$.

Se extrae una muestra de unidades primarias con reemplazamiento y con probabilidades

proporcionales a los tamaños: $\{12, 14, 14, 5, 1\}$. Por tanto:

$$p_i = \frac{N_i}{\sum\limits_{i \in U_I} N_i} = \frac{N_i}{N}$$

Se desea estimar el total de horas que los estudiantes invirtieron en estudiar la asignatura. La respuesta Y_{U_i} es el total de horas que todos los estudiantes del grupo i destinaron a su estudio. Los resultados obtenidos son los siguientes:

Grupo	N_i	p_i	Y_{U_i}	$\frac{Y_{U_i}}{p_i}$
12	24	$\frac{24}{647}$	75	2021.875
14	100	$\frac{100}{647}$	203	1313.410
14	100	$\frac{100}{647}$	203	1313.410
5	76	$\frac{76}{647}$	191	1626.013
1	44	$\frac{44}{647}$	168	2470.364

Una estimación del total poblacional del número de horas invertidas en el estudio es

$$\widehat{Y}_U^{\text{HH}} = \frac{2021{,}875 + 1313{,}410 + 1313{,}410 + 1626{,}013 + 2470{,}364}{5} = 1749{,}014$$

Una estimación de la varianza es

$$\widehat{\mathbb{V}}\left(\widehat{Y}_U^{\text{HH}}\right) = \frac{1}{5} \cdot \frac{(2021{,}875 - 1749{,}014)^2 + ... + (2470{,}364 - 1749{,}014)^2}{4} = 49470{,}656$$

De manera que la estimación del error de muestreo es

$$\widehat{\sigma}\left(\widehat{Y}_U^{\text{HH}}\right) = 222{,}42$$

La cantidad promedio de tiempo que un estudiante estudió la asignatura es

$$\widehat{\overline{y}}_U^{\text{HH}} = \frac{1749{,}014}{647} = 2{,}70$$

Luego la estimación del error de muestreo es

$$\sigma\left(\widehat{\bar{y}}_U^{\mathrm{HH}}\right) = \sqrt{\frac{49470{,}656}{647^2}} = 0{,}34$$

■

Particularicemos ahora para el caso en que los conglomerados son seleccionados mediante muestreo aleatorio simple con reemplazamiento. El estimador HH del total poblacional Y_U es

$$\widehat{Y}_U^{\mathrm{HH}} = \frac{N_I}{n_I}\sum_{\nu=1}^{n_I} Y_{U_{\nu_i}}$$

La varianza se puede escribir como

$$\mathbb{V}(\widehat{Y}_U^{\mathrm{HH}}) = \frac{N_I^2}{n_I}\cdot\sigma^2_{YU_I}$$

donde

$$\sigma^2_{YU_I} = \frac{1}{N_I}\sum_{i\in U_I}(Y_{U_i} - \bar{y}_{U_I})^2$$

con $\bar{y}_{U_I} = \sum_{i\in U_I}\frac{Y_{U_i}}{N_I}$. El estimador insesgado de la varianza es

$$\widehat{\mathbb{V}}(\widehat{Y}_U^{\mathrm{HH}}) = \frac{N_I^2}{n_I}\cdot S^2_{Ys_I}$$

donde $S^2_{Ys_I} = \frac{1}{n_I-1}\sum_{i\in s_I}(Y_{U_i} - \bar{y}_{s_I})^2$ con $\bar{y}_{s_I} = \sum_{i\in s_I}\frac{Y_{U_i}}{n_I}$.

Nótese que s_I representa la muestra ordenada de conglomerados seleccionados.

4.3.2 Muestreo con probabilidades diferentes en dos etapas con reemplazo

Consideremos el siguiente tipo de diseño muestral:

i. En la primera etapa de muestreo, se selecciona una muestra ordenada de conglomerados (PSUs)
$$os_I = \{i_1, \ldots, i_\nu, \ldots, i_{n_I}\}$$
siguiendo un esquema de muestreo con reemplazamiento tal que, en cada extracción, la probabilidad de seleccionar el conglomerado i es p_i, para todo $i \in \{1, \ldots, N_I\}$.

ii. En la segunda etapa, se selecciona una submuestra de n_i elementos del conglomerado i. Se satisfacen las mismas propiedades de invarianza e independencia ya descritas con anterioridad (véase los Comentarios 31 y 32). El diseño elegido para la submuestra es un esquema sin reemplazamiento. Con frecuencia se suele elegir el muestreo aleatorio simple sin reemplazamiento o el muestreo sistemático.

iii. Si un conglomerado (PSU) es seleccionado más de una vez, se submuestrea independientemente tantas veces como haya sido seleccionado.

Denotemos por $\widehat{Y}^{\text{HT}}_{U_{i_\nu}|os_I}$ el estimador HT del total poblacional $Y_{U_{i_\nu}}$ en el conglomerado U_{i_ν}; por $V_{i_\nu} = \mathbb{V}\left[\widehat{Y}^{\text{HT}}_{U_{i_\nu}|os_I}|os_I\right]$ la varianza del estimador $\widehat{Y}^{\text{HT}}_{U_{i_\nu}|os_I}$ condicionada al conglomerado seleccionado en primera etapa. Entonces, podemos demostrar el siguiente resultado.

Teorema 26

En un muestreo bietápico bajo las condiciones i, ii y iii anteriores, un estimador insesgado para el total poblacional Y_U está dado por

$$\widehat{Y}^{HH}_U = \frac{1}{n_I}\sum_{\nu=1}^{n_I}\frac{\widehat{Y}^{\text{HT}}_{U_{i_\nu}|os_I}}{p_{i_\nu}}. \tag{4.48}$$

La varianza de $\widehat{Y}^{HH}_U$ está dada por

$$\mathbb{V}\left[\widehat{Y}^{HH}_U\right] = \frac{1}{n_I}\sum_{i=1}^{N_I} p_i\left(\frac{Y_{U_i}}{p_i} - Y_U\right)^2 + \frac{1}{n_I}\sum_{i=1}^{N_I}\frac{V_i}{p_i}. \tag{4.49}$$

Un estimador insesgado de la varianza viene dado por

$$\widehat{\mathbb{V}}\left[\widehat{Y}^{HH}_U\right] = \frac{1}{n_I(n_I-1)}\sum_{\nu=1}^{n_I}\left[\frac{\widehat{Y}^{\text{HT}}_{U_{i_\nu}|os_I}}{p_{i_\nu}} - \widehat{Y}^{HH}_U\right]^2. \tag{4.50}$$

Demostración 26

Al tratarse de una muestra ordenada en primera etapa podemos definir las variables aleatorias $Z_\nu = \frac{Y_{i_\nu}}{p_{i_\nu}}$ y $\widehat{Z}_\nu = \frac{\widehat{Y}^{\text{HT}}_{U_{i_\nu}}}{p_{i_\nu}}$. Puesto que son variables aleatorias independientes e idénticamente distribuidas con función de masa de probabilidad p_i, se cumplen

$$\mathbb{E}\left[\widehat{Z}_\nu\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{Z}_\nu|os_I\right]\right] = \mathbb{E}\left[Z_\nu\right] = Y_U$$

y

$$\begin{aligned}\mathbb{V}\left[\widehat{Z}_\nu\right] &= \mathbb{V}\left[\mathbb{E}\left[\widehat{Z}_\nu|os_I\right]\right] + \mathbb{E}\left[\mathbb{V}\left[\widehat{Z}_\nu|os_I\right]\right] \\ &= \mathbb{V}\left[Z_\nu\right] + \mathbb{E}\left[\frac{V_{i_\nu}}{p^2_{i_\nu}}\right]\end{aligned}$$

$$= \sum_{i=1}^{N_I} p_i \left(\frac{Y_{U_i}}{p_i} - Y_U \right)^2 + \sum_{i=1}^{N_I} \frac{V_i}{p_i}.$$

Como $\widehat{Y}_U^{HH}$ es la media de n_I variables aleatorias $\widehat{Z}_\nu$ independientes distribuidas idénticamente, el Teorema se sigue de resultados genéricos de la media de este tipo de variables aleatorias.

Particularizando al caso de muestreo con reemplazamiento con probabilidades desiguales en primera etapa y muestreo aleatorio simple sin reemplazamiento en segunda etapa, se tiene que el estimador es

$$\widehat{Y}_U^{\mathrm{HH}} = \frac{1}{n_I} \cdot \sum_{\nu=1}^{n_I} \frac{N_{i_\nu} \cdot \bar{y}_{s_{i_\nu}}}{p_{i_\nu}}$$

donde $\bar{y}_{s_{i_\nu}} = \frac{1}{n_i} \sum_{k \in s_{i_\nu}} y_k$ representa la media muestral de los elementos seleccionados en el conglomerado i_ν.

La varianza del estimador viene dada por

$$\mathbb{V}\left(\widehat{Y}_U^{\mathrm{HH}}\right) = \frac{1}{n_I} \cdot \sum_{i=1}^{N_I} p_i \cdot \left(\frac{Y_{U_i}}{p_i} - Y_U \right)^2 + \frac{1}{n_I} \cdot \sum_{i=1}^{N_I} \frac{N_i^2}{p_i} \cdot \frac{1 - f_i}{n_i} \cdot S_{y_{U_i}}^2 \tag{4.51}$$

Un estimador insesgado de la varianza es

$$\widehat{\mathbb{V}}\left(\widehat{Y}_U^{\mathrm{HH}}\right) = \frac{1}{n_I \cdot (n_I - 1)} \cdot \sum_{\nu=1}^{n_I} \left[\frac{N_{i_\nu} \cdot \bar{y}_{S_{i_\nu}}}{p_{i_\nu}} - \widehat{Y}_U^{\mathrm{HH}} \right]^2$$

En el caso de que se realice un muestreo aleatorio simple con reemplazo de las unidades de primera etapa, las fórmulas del estimador y el estimador de la varianza se obtienen tomando $p_i = \frac{1}{N_I} \ \forall i \in U_I$ y la varianza dada en (4.51) queda:

$$\mathbb{V}\left(\widehat{Y}_U^{\mathrm{HH}}\right) = \frac{N_I^2}{n_I} \sigma_{YU_I}^2 + \frac{N_I}{n_I} \sum_{i \in U_I} N_i^2 \frac{1 - f_i}{n_i} S_{yU_i}^2$$

Ejemplo 33. En el contexto del Ejemplo 32, supóngase que en lugar de encuestar a todos los estudiantes de los conglomerados seleccionados, se extrae una submuestra de cinco estudiantes de cada grupo seleccionado, mediante un muestreo aleatorio simple sin reemplazamiento. Los resultados obtenidos son los siguientes:

Grupo	N_i	p_i	y_k	$\bar{y}_{s_i}$	$\widehat{Y}^{\mathrm{HT}}_{U_{i_\nu}\mid os_I}$	$\frac{\widehat{Y}^{\mathrm{HT}}_{U_{i_\nu}\mid os_I}}{p_i}$
12	24	0,0371	2; 3; 2,5; 3; 1,5	2,4	57,6	1552,8
14	100	0,1546	2,5; 2; 3; 0; 0,5	1,6	160,0	1035,2
14	100	0,1546	3; 0,5; 1,5; 2; 3	2,0	200,0	1294,0
5	100	0,1175	1; 2,5; 3; 5; 2,5	2,8	212,8	1811,6
1	76	0,0680	4; 4,5; 3; 2; 5	3,7	162,8	2393,9

y_k indica la cantidad de horas que el alumno k invirtió en el estudio. Obsérvese que el conglomerado 14 se ha extraído dos veces, pero, sin embargo, el conjunto de estudiantes seleccionados en segunda etapa es distinto, ya que el submuestreo en cada conglomerado se efectúa independientemente del submuestreo en el resto de conglomerados.

Recordemos que las unidades de primera etapa (los grupos) fueron seleccionados mediante un esquema de muestreo con reemplazamiento y probabilidades proporcionales al número de estudiantes en cada grupo, esto es,

$$p_i = \frac{N_i}{N}$$

La estimación basada en el estimador HH es

$$\begin{aligned}\widehat{Y}_U^{HH} &= \frac{1}{n_I}\sum_{\nu=1}^{n_I}\frac{\widehat{Y}^{\mathrm{HT}}_{U_{i_\nu}}}{p_{i_\nu}} = \frac{1}{n_I}\sum_{\nu=1}^{n_I}\frac{N_{i_\nu}\cdot\bar{y}_{s_{i_\nu}}}{\frac{N_{i_\nu}}{N}} = \frac{N}{n_I}\sum_{\nu=1}^{n_I}\bar{y}_{s_{i_\nu}} = \\ &= \frac{647}{5}\cdot[2{,}4+1{,}6+2+2{,}8+3{,}7] = 1617{,}5.\end{aligned}$$

La estimación de la varianza es:

$$\begin{aligned}\widehat{\mathbb{V}}\left[\widehat{Y}_U^{HH}\right] &= \frac{1}{n_I(n_I-1)}\sum_{\nu=1}^{n_I}\left[\frac{\widehat{Y}^{\mathrm{HT}}_{U_{i_\nu}\mid os_I}}{p_{i_\nu}} - \widehat{Y}_U^{HH}\right]^2 = \\ &= \frac{1}{5\cdot 4}\left[(1552{,}8-1617{,}5)^2+\ldots+(2393{,}9-1617{,}5)^2\right] = \\ &= 54419{,}17\end{aligned}$$

Por tanto, la estimación del error de muestreo es aproximadamente $233{,}28$.

■

Bibliografía

Brewer, K.R.W. y M. Hanif (1983). *Sampling with unequal probabilities*. Springer.

Grimmet, G.R. y D.R. Stirzaker (2004). *Probability and random processes*. 3rd. Oxford Science Publications.

Lohr, S. (2010). *Sampling: Design and Analysis*. 2nd. Duxbury Press.

Särndal, C.-E., B. Swensson y J.H. Wretman (1992). *Model assisted survey sampling*. New York: Springer.