Técnicas básicas de Muestreo con SAS

Javier Portela García-Miguel

María Villeta López

1ª Edición (2007)

ISBN 9788496866133

ISBN 9788496866133

ISBN 9788496866133

ISBN 9788496866133

ISBN 9788496866133

ISBN 9788496866133

ISBN 9788496866133

ISBN 9788496866133

ISBN 9788496866133

ISBN 9788496866133

Prefacio

Este libro de texto presenta una exposición de la teoría básica del muestreo estadístico, destinada a su uso en todos los ámbitos de la realidad en los cuales se aplica. Su creación responde a la necesidad de plasmar nuestra experiencia como profesores de la materia en la Escuela de Estadística de la Universidad Complutense de Madrid. El libro podrá ser útil en cursos de muestreo, pero también puede servir como referencia de consulta para la aplicación práctica de los métodos presentados. El libro puede servir de referencia a cualquier profesional del campo de la Estadística, de la Investigación de Mercados, Encuestas y Sondeos de opinión, de la Medicina, Ciencias Sociales o Ingeniería.

Los conocimientos matemáticos necesarios para abordar la mayor parte del contenido del libro incluyen la familiaridad con los conceptos presentados en un curso básico de cálculo de probabilidades, así como ciertos conocimientos de inferencia estadística. El profesional o estudiante que acceda a este texto debería conocer al menos los conceptos habituales de un curso básico de estadística, como conocimientos de estadística descriptiva, las propiedades de las distribuciones binomial y normal, y resultados de inferencia tales como la construcción de intervalos de confianza.

El libro contiene 11capítulos, divididos en apartados. Los conceptos y resultados presentados están ilustrados con ejemplos numerados basados en gran parte en datos reales. Debido a la naturaleza de los métodos presentados, los ejemplos van paulatinamente incorporando conceptos ilustrados con anterioridad, de modo que su complejidad es creciente, en un esfuerzo por familiarizar al estudiante con las técnicas de manera integrada. En cada capítulo se resumen las expresiones de los estimadores en tablas de fórmulas, de interés práctico para el lector por su rápido acceso. Se incluyen numerosos ejercicios resueltos y propuestos, además de una sección explicando la utilización del SAS en obtención de muestras y estimación.

Este manual incluye un CD-rom con los programas SAS o macros necesarios para la obtención de muestras y estimación en los métodos de muestreo estudiados, y archivos de datos que servirán a las aplicaciones y ejercicios propuestos.

El texto comienza con dos temas de introducción a los conceptos básicos en muestreo. Se presenta en el capítulo segundo el método de muestreo aleatorio simple con reemplazamiento, de interés por su simplicidad matemática, y que permitirá abordar métodos de muestreo más complicados con algunos conceptos importantes ya introducidos, como la estimación de la proporción, media y total, el tratamiento del problema de estimación del error de muestreo y del cálculo del tamaño muestral.

El capítulo siguiente aborda el muestreo aleatorio simple sin reemplazamiento, considerado el método de referencia en muestreo, por lo cual es ilustrado con gran cantidad de ejemplos. En el capítulo quinto se presenta el muestreo estratificado, también de extrema importancia, pues en la práctica es muy frecuente la partición de la población para realizar el muestreo de manera independiente en cada una de las partes. El capítulo seis presenta el muestreo sistemático, que reemplaza a menudo en la realidad al muestreo aleatorio simple sin reemplazamiento por sus virtudes prácticas, además de por su justificación teórica.

El capítulo siete aborda importantes técnicas de estimación indirecta como son la estimación de razón y de regresión, que pueden mejorar en muchas circunstancias a la estimación directa habitual, si se dispone de información poblacional de una variable auxiliar relacionada

ISBN 9788496866133

con la variable de interés. El muestreo con probabilidades desiguales, presentado en el capítulo ocho, es un modo de intentar mejorar la precisión de la estimación respecto a la obtenida en el muestreo aleatorio simple, y será sobre todo útil en muestreo por conglomerados, cuyos conceptos se introducirán en el capítulo nueve. En éste se presenta el muestreo por conglomerados monoetápico, de gran aplicación práctica, para el cual se particiona la población en grupos de unidades, escogiendo aleatoriamente algunos de éstos, y examinando todas las unidades elementales dentro de cada uno de los grupos seleccionados. Finalmente, en el último capítulo del libro se presenta el muestreo por conglomerados bietápico, en el cual se escogen conglomerados aleatoriamente y dentro de éstos, muestras aleatorias de unidades elementales.

Finalmente se incluye un capítulo de síntesis de la utilización del paquete estadístico SAS en muestreo, necesario para aclarar ciertos conceptos de la utilización de este programa en la práctica real de la obtención de muestras y estimación.

Esperamos que este texto sirva de ayuda a todos aquellos profesionales y estudiantes para los cuales el muestreo es una técnica de interés, pero principalmente a los alumnos de la Escuela de Estadística de la Universidad Complutense de Madrid, a quienes lo dedicamos.

ISBN 9788496866133

ISBN 9788496866133

1 INTRODUCCIÓN AL MUESTREO

1.1 Muestreo y sus aplicaciones

El muestreo es una fuente de acceso a la realidad. Entendemos por muestra una fracción de la población representativa de ésta, de manera que pueda utilizarse para extraer conclusiones sobre la población. Los procedimientos y técnicas empleados para escoger esta muestra deben de estar orientados a que cada muestra posible no introduzca sesgos o desviaciones claras y sea suficientemente precisa. La teoría estadística del muestreo se ocupa de los métodos y técnicas para diseñar la elección de la muestra y obtener aproximaciones a ciertas características poblacionales como pueden ser proporciones, medias o totales de las variables de interés.

Frente a las muestras, existen los llamados censos, o enumeraciones completas de todas las unidades poblacionales. Estos censos resultan costosos y su duración suele ser larga. Las muestras constituyen modelos reducidos de la realidad poblacional, son más rápidas y menos costosas que los censos, y sus resultados suelen ser extrapolables al universo del que se extraen.

Veremos a continuación algunos ejemplos de la utilización del muestreo en diversos campos.

Algunas investigaciones realizadas por el INE (Instituto Nacional de Estadística) son:

- Encuesta de Población Activa (EPA), sobre actividad, ocupación y paro de la población.
- Encuesta de Presupuestos Familiares (EPF), sobre nivel de ingresos y gastos familiares.

Encuestas realizadas por empresas de sondeos de opinión incluyen:

- Sondeos de opinión, actitudes sociales.
- Encuestas electorales.

Estudios realizados en el marco de la denominada investigación de mercados:

- Estudios de audiencia en radio y televisión.

ISBN 9788496866133

- Estudio de preferencias respecto a productos a lanzar al mercado o ya existentes.

En la industria y servicios:

- Encuestas sobre la actividad industrial.
- Control de calidad, muestreo de aceptación de lotes.
- Estudios sobre tiempos de vida de maquinaria y equipos.
- Estimación de inventarios.
- Auditorías contables.

En Biología:

- Estudios de niveles de contaminación en ríos o aire.
- Estimación del número de individuos de una determinada especie animal o vegetal.
- Estudios geográficos de distribución de especies en ciertas áreas.

En Medicina:

- Estimación de prevalencia de determinadas enfermedades en la población.
- Estudios de la efectividad de vacunas.
- Estudio de diferentes causas de muerte.

En agricultura:

- Estudios de la distribución de los diferentes tipos de cultivo en un país o región.
- Estudios relativos a la producción agrícola.

Como se ve, la diversidad de campos en que se aplica o podría aplicar el muestreo es inmensa. A continuación veremos con más detalle algunas de las características que justifican el uso de muestreo o por el contrario, lo desaconsejan.

ISBN 9788496866133

1.2 Ventajas y desventajas del muestreo

Dado que el muestreo supone riesgo, es útil indicar en qué casos conviene obtener muestras, en lugar de censos o investigaciones exhaustivas:

- Cuando la población es tan grande que el censo excede las posibilidades del investigador.
- Cuando la población sea lo suficientemente uniforme para que cualquier muestra de una buena representación de la misma.
- Cuando el proceso de medición o investigación de las unidades sea destructivo.

Existen otras razones o ventajas que aporta el muestreo respecto a los censos:

- Las muestras son menos costosas.
- Mayor rapidez en la obtención de los resultados.
- Al reducir el volumen de trabajo el personal escogido es menor, puede estar más capacitado y ser sometido a entrenamiento particular. Además el proceso de toma de datos y depuración es de mayor calidad que en un censo. Así, una muestra puede conducir a resultados más exactos que una enumeración completa.

En cuanto a las desventajas o imposibilidad de utilizar el muestreo, se tiene:

- No es posible utilizar muestreo cuando se necesite información de cada uno de los elementos poblacionales.
- El muestreo exige, en comparación con los censos, menos trabajo material pero más preparación y refinamiento. Si la encuesta es compleja, los análisis estadísticos derivados de ella necesitan gran sofisticación (uso de ponderaciones, corrección de las probabilidades desiguales, etc.).
- Cuando se requiere máxima calidad no se suele utilizar el muestreo (componentes central nuclear, motores aviones, etc.).

1.3 Concepto de población, marco y muestra

Se definirán a continuación algunos conceptos fundamentales en la teoría de muestreo. Se denomina **Población** al conjunto de elementos del cual se desea obtener una información. Una **muestra** es cualquier subconjunto de la población que se utiliza para obtener resultados extrapolables al universo del que se extraen. El **Marco** es la población restringida a la información concreta que se dispone de ella (es decir, listados de la población). Se desea en principio que las unidades presentes en el marco sean todas las poblacionales, y nada más que las poblacionales, pero en la práctica los marcos pueden presentar imperfecciones como unidades poblacionales ausentes, duplicados o elementos no poblaciones que están presentes. Las muestras se obtienen

ISBN 9788496866133

a partir del marco, con lo que, hablando estrictamente, el muestreo obtiene resultados extrapolables al marco, y no a la población. A pesar de estas razones, en la práctica a menudo se asume que las imperfecciones del marco son despreciables respecto al conjunto de la población y por lo tanto no se tienen en consideración estas irregularidades.

Para ilustrar los problemas que genera la creación de un marco apropiado es suficiente pensar en una encuesta de muestreo para conocer los lugares de destino de vacaciones de las familias de una ciudad . Si los datos censales de que se dispone son de hace dos años, todas las familias que han migrado a esta ciudad estarían ausentes y aquellas que han abandonado la ciudad como residencia estarían incorrectamente presentes. Por lo tanto habría que corroborar el censo con el Padrón municipal de habitantes, recibos de impuestos o de alquileres, etc.

1.4 Etapas en un proceso muestral

Se pueden especificar las siguientes etapas:

1. Especificación de objetivos , variables de interés, medición, etc.

   En esta etapa se deben definir los conceptos que se desean medir, el instrumento de medición y la manera operativa de llevar a a cabo esta medición. Es una etapa que aunque a primera vista no ofrezca dificultades es muy influyente en el resultado final de la investigación y a menudo presenta muchos problemas. Medir algo tan sencillo como la utilización de gas o no en un hogar puede presentar circunstancias a estudiar a priori como la asociación de la factura del gas con ese hogar, presencia de contadores colectivos, instalación pero no utilización o avería, etc.

   El instrumento de medida (encuestadores y modus operandi) también puede llegar a tener enorme influencia sobre los resultados medidos. No es lo mismo realizar encuestas telefónicas o visitas no concertadas, con los problemas de no respuesta asociados, que utilizar un panel de hogares fijo o concertar la entrevista previamente con el responsable del hogar. La formación y medios de los encuestadores, la longitud de la encuesta, la tecnología empleada son factores que pueden influir en el resultado.

2. Determinación de los elementos de la población y el marco poblacional.

   Como se ha comentado, a menudo la determinación del marco de manera que se corresponda, salvo escasas excepciones, con la población, es difícil. A veces no se dispone de un único listado y es necesario un trabajo investigador especial para reunir información de diversas fuentes. Otras veces la utilización de ciertos marcos puede presentar problemas legales de violación de la intimidad. En ocasiones es necesario variar las unidades objetivo (de individuos a familias , por ejemplo) debido a las imposibilidades prácticas de elaboración de un marco (por ejemplo, localizar a una persona para entrevistarla puede ser difícil, pero es más fácil entrevistar a un miembro cualquiera de su hogar).

3. Determinación del plan de muestreo. Especificar el diseño muestral (tipo de muestreo, tamaño muestral y estimación).

   Para establecer el plan de muestreo se parte de la información estructural prefijada en el marco. Para que el diseño sea apropiado en el sentido de obtener la máxima precisión con el mínimo coste, a veces se requiere tener información previa sobre los valores aproximados

ISBN 9788496866133

que suele tomar la variable de interés o alguna relacionada con ella, en los diferentes estratos o secciones de la población. Por ello es importante disponer de información auxiliar (estudios anteriores sobre temas similares en la población) o bien se recurre a una encuesta piloto, que es un estudio de bajo coste que ayuda a identificar características poblacionales de interés para reducir costes en la encuesta final, e ilustra sobre defectos o problemas que puedan surgir en el trabajo de campo, permitiendo corregirlos antes de la encuesta definitiva.

A partir de toda esta información, se elabora el esquema matemático y práctico para la obtención de las muestras, anticipando aproximadamente el tipo de estimación que se utilizará finalmente sobre los datos, la tabulación informática, los ajustes por no respuesta, etc.

4. Recogida de datos o trabajo de campo (adiestramiento del personal, encuestas por diversos medios, ...).

   En esta etapa se procede a obtener los datos muestrales. Es necesario tener un plan de toma de datos muy claro, que no deje ninguna circunstancia al arbitrio del encuestador (cuantas visitas hay que repetir a un hogar antes de sustituirlo por otro por ausencia, por cuál se sustituirá, qué determinaciones tomar en caso de direcciones equivocadas, repeticiones de llamadas telefónicas, etc.).

5. Resumen y análisis de datos. Análisis de la falta de respuesta, estimación de errores, etc.

   En esta etapa se analizan y tabulan los datos obtenidos, depurando errores como ausencias o duplicados, y calculando las estimaciones adecuadas al diseño muestral utilizado. Se estudian modos de tratar la ausencia de respuesta y se estiman los errores de muestreo, obteniendo intervalos de confianza.

1.5 Tipos de muestreo probabilístico

En este apartado se describen brevemente los principales tipos de muestreo probabilístico y la forma de aplicarlos.

Los muestreos probabilísticos se caracterizan porque en ellos cada elemento de la población tiene una probabilidad conocida de antemano de ser seleccionado. En este tipo de muestreo está justificado el uso de la inferencia estadística, pudiéndose aproximar el nivel de error de las estimaciones. Para seleccionar el tipo de muestreo se tienen en cuenta razones de precisión, coste, cuestiones administrativas, de disposición de información, etc.

Muestreo aleatorio simple.

El m.a.s. consiste en escoger las unidades muestrales con igual probabilidad . Este tipo de muestreo puede realizarse con reposición, donde cada unidad puede ser escogida varias veces, o sin reposición. El m.a.s. se aplica fundamentalmente en poblaciones pequeñas, plenamente identificables. En el caso de poblaciones grandes, la utilización de este método presenta dificultades:

- Es difícil obtener un listado de toda la población.

ISBN 9788496866133

- Aún si fuera posible obtener un listado, la posible dispersión geográfica de la muestra obtenida daría lugar a costes demasiado elevados.

El muestreo aleatorio simple se presenta como el prototipo de muestreo por su sencillez y facilidad para calcular los errores de muestreo.

Muestreo aleatorio sistemático.

Es un método de características prácticas muy interesantes y que da, salvo raras excepciones, resultados similares o mejores que el m.a.s. Consiste en dividir un listado de la población en $k = N/n$ partes iguales , obtener aleatoriamente el denominado punto de a-rranque y seleccionar los n items correspondientes a los lugares equidistantes en la lista de k unidades, a partir de ese punto de arranque. Este método es utilizado en encuestas en lugares públicos, muestreo en agricultura, procesos de control de calidad, auditorías, etc. y suele sustituir adecuadamente al m.a.s.

Muestreo estratificado.

A menudo la población presenta ciertas divisiones más o menos evidentes en cuanto al comportamiento de la variable de interés. Con la **estratificación** o partición de la población en subpoblaciones o **estratos ,** se persiguen distintos fines:

- Dar estimaciones separadas para ciertas subpoblaciones del estudio.
- Agrupar unidades de muestreo homogéneas entre sí en el mismo estrato, con lo cual se mejorará la precisión de las estimaciones globales.
- Utilizar métodos diferentes de muestreo en los distintos estratos.

El muestreo estratificado facilita en general el trabajo de campo, pues muchas veces los estratos corresponden a criterios geográficos. Además, suele mejorar en cuanto a precisión al m.a.s. si los estratos son homogéneos internamente y diferentes entre sí respecto a la variable de interés.

Muestreo por conglomerados.

En una población localizada en un área geográfica grande, las unidades elementales (niños, familias, pacientes, etc.) pueden estar muy dispersas por el territorio geográfico y el coste de desplazamiento a todas las unidades escogidas por procedimiento de muestreo puede ser prohibitivo.

Para evitar este aumento de costes, se suele recurrir a una muestra de grupos de unidades elementales (escuelas, edificios, hospitales, etc.) llamados **conglomerados**. Cuando los conglomerados se delimitan por criterios geográficos, el muestreo se denomina muestreo por áreas. Si en cada conglomerado de la muestra se entrevista a todas las unidades elementales que lo forman, se dice que el muestreo es en **una etapa**. Si dentro de cada uno de los conglomerados de la muestra se obtiene a su vez una muestra de unidades elementales, el muestreo es **bietápico**. El procedimiento puede generalizarse a cualquier número de etapas. En cada una

ISBN 9788496866133

de éstas existe un tipo de unidades elementales llamadas sucesivamente de primera etapa, de 2ª etapa, etc. A este tipo de muestreo se le denomina **polietápico**.

El muestreo por conglomerados suele reducir mucho los costes. En principio los conglomerados deben ser heterogéneos internamente (pues cada uno de ellos tiene que representar a la población) y parecidos entre sí (pues se seleccionan aleatoriamente algunos y por lo tanto cada conglomerado concreto debe representar a los demás).

Muestreo en dos fases.

En este tipo de muestreo se selecciona una muestra inicial suficientemente grande de forma rápida, sencilla y poco costosa, a fin de que su información sirva de base para una submuestra de ésta que será aquella donde se recoja la información de la variable de interés. Por ejemplo, en una primera fase se puede seleccionar un gran número de familias, recogiendo datos de domicilio, alquiler u otras variables fáciles de conseguir. En una segunda fase se selecciona una submuestra de estas familias a la cual se aplica ya la encuesta de presupuestos familiares.

Generalmente el muestreo en dos fases se utiliza en casos en que la estimación se apoya en una variable auxiliar (ver el capítulo "Estimación Indirecta"), y la primera fase es abordada para obtener información sobre la variable auxiliar, en general poco costosa.

Muestreo de captura-recaptura.

Es utilizado, por ejemplo, en estudios de poblaciones animales. En este tipo de muestreo se extrae en primer lugar una muestra, se "marca" y devuelve a su hábitat. A continuación se extrae una segunda muestra y se observa el número de items marcados. Se utiliza esta información para calcular el tamaño poblacional. Este tipo de muestreo también se utiliza en estimaciones de prevalencia de enfermedades a través de listados de enfermos.

Muestreo inverso.

Es utilizado cuando se desea estimar la proporción de una cualidad rara. Se extraen sucesivamente unidades elementales de la población hasta obtener un número prefijado de items con la cualidad de interés. Se utilizan razonamientos probabilísticos relacionados con la distribución binomial negativa para estimar la proporción objetivo.

1.6 Muestreo no probabilístico

El muestreo no probabilístico no controla la aleatoriedad introducida en el proceso de selección de muestras. Algunos métodos no probabilísticos son:

Muestreo circunstancial o sin norma.

Es un tipo de muestreo arbitrario: las unidades muestrales se seleccionan a criterio del investigador, a menudo por motivos prácticos como cercanía o voluntariedad. También entra en este tipo de muestreo el llamado muestreo de juicio, en el que el investigador selecciona ciertas unidades que él considera subjetivamente representativas de la población. Este tipo de

ISBN 9788496866133

muestreo puede contener sesgos si esa arbitrariedad en la elección lleva a una relación directa o indirecta entre las unidades escogidas y la variable de interés.

Muestreo por cuotas.

Para evitar la rigidez de los esquemas en muestreo probabilísticos se recurre a menudo en encuestas de opinión al muestreo por cuotas, basado en el establecimiento de cuotas en la población de manera que cada cuota esté debidamente representada. Se suele pedir al entrevistador de calle que seleccione de manera aleatoria las unidades muestrales, siempre que termine rellenando sus hojas con un determinado número de individuos en cada cuota. Por ejemplo, el entrevistador debe haber entrevistado al final de la jornada 10 mujeres y 10 hombres de cada tramo de edad especificado, etc.

Este tipo de muestreo garantiza una representación de ciertos sectores minoritarios que son interesantes para el investigador y que con otro tipo de muestreo tendrían pocas posibilidades de estar en la muestra. A veces las cuotas son proyecciones proporcionales de los sectores poblacionales, obtenidos a menudo del censo. En cualquier caso, la técnica de la postestratificación en la estimación permite, por ejemplo, corregir en parte el sesgo debido al muestreo por cuotas. En general este muestreo no es costoso, agiliza y simplifica el trabajo de campo, y en sondeos de opinión refleja resultados muy aceptables. Su principal defecto es que no es probabilístico, y por lo tanto no es posible estimar de manera rigurosa el nivel de error o sesgo cometido en las estimaciones. Por ello, cuando se trata de encuestas oficiales, se utilizan más muestreos de tipo probabilístico.

ISBN 9788496866133

2 CONCEPTOS BÁSICOS EN MUESTREO

En muestreo probabilístico, la muestra $(u_1, u_2, ..., u_n)$.es seleccionada de una población mediante un proceso aleatorio que atribuye probabilidades de aparición a cada elemento poblacional. Como en general el objetivo final del muestreo es aproximar o estimar características poblacionales de la variable de interés tales como la media o proporcion poblacional, existen varios conceptos de crucial interés que es necesario abordar previamente al estudio de cualquier tipo de muestreo o estimación. En resumen, en este tema se presentan de forma sucesiva los siguientes conceptos en este orden:

- La combinatoria relativa a la selección de muestras.
- La definición y de características poblacionales a aproximar, y la construcción de características muestrales asociadas de interés.
- Recordatorio del concepto de variable aleatoria.
- Definiciones asociadas al proceso de estimación.

2.1 Combinatoria básica en muestreo

Recordaremos a continuación algunos cálculos relacionados con el muestreo en una población finita de tamaño N, donde se extrae una muestra de tamaño n, $(u_1, u_2, ..., u_n)$.

Para fijar ideas, se supone una urna con N bolas numeradas, de las cuales se extraen n. Los resultados siguientes se pueden consultar en cualquier libro básico de cálculo de probabilidades.

Propiedad 2.1 (cálculo del número de muestras diferentes bajo distintos tipos de muestreo).

a) El número de muestras diferentes que se obtienen cuando se extraen n unidades con reemplazamiento, teniendo en cuenta el orden, es N^n.

ISBN 9788496866133

b) El número de muestras diferentes que se obtienen cuando se extraen n unidades sin reemplazamiento, teniendo en cuenta el orden, es $n!\binom{N}{n}$.

c) El número de muestras diferentes que se obtienen cuando se extraen n unidades con reemplazamiento, sin tener en cuenta el orden, es $\binom{N+n-1}{n}$.

d) El número de muestras diferentes que se obtienen cuando se extraen n unidades sin reemplazamiento , sin tener en cuenta el orden, es $\binom{N}{n}$.

Ejemplo 2.1.

Supongamos una urna con $N = 3$ bolas numeradas. Fijamos $n = 2$.

1-.Hay $N^n = 3^2 = 9$ muestras diferentes en muestreo con reemplazamiento, teniendo en cuenta el orden: $(1,1)$, $(1,2)$, $(1,3)$, $(2,1)$, $(2,2)$, $(2,3)$, $(3,1)$, $(3,2)$, $(3,3)$.

2-. Hay $n!\binom{N}{n} = 2!\binom{3}{2} = 6$ muestras diferentes en muestreo sin reemplazamiento, teniendo en cuenta el orden: $(1,2)$, $(1,3)$, $(2,1)$, $(2,3)$, $(3,1)$, $(3,2)$.

3-. Hay $\binom{N+n-1}{n} = \binom{3+2-1}{2} = 6$ muestras diferentes en muestreo con reemplazamiento, sin tener en cuenta el orden: $\{1,1\}$, $\{1,2\}$, $\{1,3\}$, $\{2,2\}$, $\{2,3\}$, $\{3,3\}$.

4-. Hay $\binom{N}{n} = \binom{3}{2} = 3$ muestras diferentes en muestreo sin reemplazamiento, sin tener en cuenta el orden: $\{1,2\}, \{1,3\}, \{2,3\}$.

Estos resultados se utilizarán, en cada tipo de muestreo , para calcular las probabilidades de aparición de cada una de las distintas muestras, dando lugar a una distribución de probabilidad sobre los resultados posibles .

Ejemplo 2.2.

Partiendo de la Figura 2.1, se desea estimar el peso medio del contenido en manzanas por árbol, considerando los cuatro árboles como la población, y extrayendo una muestra de dos árboles.

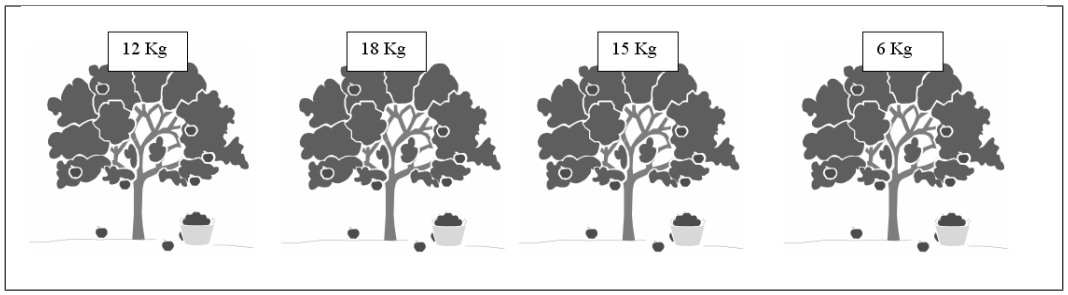


Figura 2.1. Pesos de las manzanas obtenidas en cada árbol.

Por el resultado anterior, hay exactamente:

(a) $N^n = 4^2 = 16$ muestras en muestreo con reemplazamiento, teniendo en cuenta el orden.

ISBN 9788496866133

(b) $n!\binom{N}{n} = 2!\binom{4}{2} = 12$ muestras en muestreo sin reemplazamiento, teniendo en cuenta el orden.

(c) $\binom{N+n-1}{n} = \binom{4+2-1}{2} = 10$ muestras en muestreo con reemplazamiento, sin tener en cuenta el orden.

(d) $\binom{N}{n} = \binom{4}{2} = 6$ muestras en muestreo sin reemplazamiento, sin tener en cuenta el orden.

Dependiendo del tipo de muestreo que escojamos, la configuración de muestras es diferente, como hemos visto en el ejemplo anterior. Por ejemplo, una muestra posible de las 16 del caso (a) es el primer árbol repetido dos veces, lo que da lugar a los pesos $(12, 12)$ Si se estima la media poblacional del peso en contenido en manzanas por la media muestral, se obtiene una estimación de $\overline{y}$, $\widehat{\overline{y}} = \frac{12+12}{2} = 12$. Suponiendo el caso (d), una muestra posible es, por ejemplo, la que contiene los árboles 3 y 4, lo que da lugar a los valores $\{15, 6\}$ y a la estimación $\widehat{\overline{y}} = \frac{15+6}{2} = 10.5$. En la tabla siguiente se exponen las diferentes muestras posibles y la media muestral obtenida en cada caso.

En la tabla 2.1 se pueden observar dos cuestiones de interés:

1) Según el tipo de muestreo utilizado, el conjunto de muestras posibles es diferente, y por lo tanto da lugar a diferentes valores del estimador.

2) Los valores del estimador difieren de una muestra a otra para cada tipo de muestreo. Para cada muestra obtenida en la práctica, daremos una estimación diferente, una aproximación diferente a la realidad poblacional. Como se verá en el apartado de estimación, esta variabilidad refleja el "error de muestreo": cuanto más alta, menos preciso será el estimador.

En el ejemplo, se observa que el verdadero valor de la media poblacional es $\overline{y} = \frac{12+18+15+6}{4} = 12.75$.

	Muestras posibles	**Media muestral en cada caso**
(a)	$(1,1),(1,2),(1,3),(1,4),\ (2,1),(2,2),$ $(2,3),(2,4),(3,1),(3,2),(3,3),(3,4),$ $(4,1),(4,2),(4,3),(4,4)$	$12, 15, 13.5, 9, 15, 18,$ $16.5, 12, 13.5, 16.5, 15, 10.5,$ $9, 12, 10.5, 6$
(b)	$(1,2),(1,3),(1,4),\ (2,1),(2,3),(2,4),$ $(3,1),(3,2),(3,4),(4,1),(4,2),(4,3)$	$15, 13.5, 9, 15, 16.5, 12,$ $13.5, 16.5, 10.5, 9, 12, 10.5$
(c)	$\{1,1\},\{1,2\},\{1,3\},\{1,4\},\{2,2\},$ $\{2,3\},\{2,4\},\{3,3\},\{3,4\},\{4,4\}$	$12, 15, 13.5, 9, 18,$ $16.5, 12, 15, 10.5, 6$
(d)	$\{1,2\},\{1,3\},\{1,4\},$ $\{2,3\},\{2,4\},\{3,4\}$	$15, 13.5, 9,$ $16.5, 12, 10.5$

Tabla 2.1. Muestras y medias muestrales asociadas en el ejemplo de los árboles.

Se recoge a continuación una serie de definiciones y resultados conocidos, que serán útiles en lo sucesivo.

ISBN 9788496866133

2.2 Características poblacionales y muestrales

Suponemos que y_i para $i = 1, ..., N$ son los valores de la variable y en una población de tamaño N. Análogamente cuando se dispone de una segunda variable de interés x sus valores son x_i con $i = 1, ..., N$. Se definen a continuación algunos momentos poblacionales y muestrales, que no son sino funciones de los valores poblacionales y muestrales, respectivamente.

MOMENTOS POBLACIONALES
Media poblacional: $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$
Total poblacional: $N\overline{y}$
Varianza poblacional: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2$
Cuasivarianza poblacional: $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})^2$
Coeficiente de variación poblacional: $CV(y) = \frac{\sigma}{\overline{y}}$
Cuasicovarianza poblacional: $S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})(x_i - \overline{x})$
Coeficiente de Correlación poblacional: $\rho = \frac{S_{xy}}{S_x S_y}$

Estas definiciones tienen su equivalente cuando se dispone de una muestra de tamaño n: y_i con $i = 1, ..., n$, como se verá en la siguiente tabla.

MOMENTOS MUESTRALES
Media muestral: $.\widehat{\overline{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$
Varianza muestral: $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\overline{y}})^2$
Cuasivarianza muestral: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{\overline{y}})^2$
Coeficiente de variación muestral: $\widehat{CV}(y) = \frac{\widehat{\sigma}}{\widehat{\overline{y}}}$
Cuasicovarianza muestral: $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{\overline{y}})(x_i - \widehat{\overline{x}})$
Coeficiente de Correlación muestral: $r = \frac{s_{xy}}{s_x s_y}$

ISBN 9788496866133

Si y es una variable dicotómica ($y = 1$ si la unidad poblacional tiene cierta cualidad, $y = 0$ si no la tiene), se definen las siguientes características:

MOMENTOS POBLACIONALES EN CASO DE PROPORCIONES
Proporción poblacional: $p = \frac{1}{N}\sum_{i=1}^{N} y_i$
Varianza poblacional : $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2 = p(1-p) = pq$
Cuasivarianza poblacional: $s^2 = \frac{N}{N-1}p(1-p)$

MOMENTOS MUESTRALES EN CASO DE PROPORCIONES
Proporción muestral: $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$
Varianza muestral: $\widehat{\sigma}^2 = \widehat{p}(1-\widehat{p}) = \widehat{pq}$
Cuasivarianza muestral: $s^2 = \frac{n}{n-1}\widehat{p}(1-\widehat{p})$

Veamos a continuación ciertas propiedades de interés relativas a los momentos definidos anteriormente:

Propiedad 2.2 (varianzas y covarianzas poblacionales y muestrales).

a) $\sigma^2 = \frac{N-1}{N}S^2$ y análogamente $\widehat{\sigma}^2 = \frac{n-1}{n}s^2$

b) $\sum_{i=1}^{N}(y_i - \overline{y})^2 = \sum_{i=1}^{N} y_i^2 - N\overline{y}^2$ y análogamente $\sum_{i=1}^{n}(y_i - \widehat{\overline{y}})^2 = \sum_{i=1}^{n} y_i^2 - n\widehat{\overline{y}}^2$.

c) $\sum_{i=1}^{N}(y_i - \overline{y})(x_i - \overline{x}) = \sum_{i=1}^{N} x_i y_i - N\overline{xy}$ y análogamente

$\sum_{i=1}^{n}(y_i - \widehat{\overline{y}})(x_i - \widehat{\overline{x}}) = \sum_{i=1}^{n} x_i y_i - n\widehat{\overline{x}}\widehat{\overline{y}}$.

2.3 Variables aleatorias

La muestra seleccionada proviene de un proceso aleatorio, y por lo tanto cualquier característica muestral (por ejemplo la media muestral) es una variable aleatoria que toma diferentes valores para cada una de las muestras posibles. La probabilidad de cada uno de estos valores de esta característica es la probabilidad de que la muestra asociada sea la escogida en el proceso de muestreo.

Por lo tanto, para cualquier estudio de los posibles resultados de un proceso de muestreo es necesario utilizar el concepto de variable aleatoria. Recordemos en primer lugar este concepto,

ISBN 9788496866133

necesario para desarrollar los conceptos de estimación e inferencia estadística presentes en todos los métodos de muestreo probabilístico.

1) Una variable aleatoria **discreta** Y es el resultado de un experimento probabilístico que asocia probabilidades $p_1, ..., p_M$ a los valores respectivos $Y_1, ..., Y_M$.

La esperanza de Y se define como $E[Y] = \sum_{i=1}^{M} p_i Y_i$, y la varianza de Y como

$V(Y) = E(Y - E(Y))^2$.

2) Una variable aleatoria **continua** Y con función de densidad $f(y)$ es el resultado de un experimento probabilístico que asocia probabilidad $P(a \leq Y \leq b) = \int_a^b f(y)dy$ a los valores de Y en el intervalo (a, b). La esperanza de Y se define como

$$E[Y] = \int_{-\infty}^{+\infty} yf(y)dy$$

y la varianza de Y como

$V(Y) = E(Y - E(Y))^2$.

Supongamos a continuación que Y es una variable aleatoria, con esperanza $E[Y]$ y varianza $V[Y]$. Se recuerda también la definición de covarianza entre dos variables aleatorias X e Y, como $COV(X, Y) = COV(Y, X) = E[XY] - E[X]E[Y]$. a y b son constantes arbitrarias.

Propiedad 2.3 (esperanzas y varianzas en variables aleatorias).

1) (Linealidad de la Esperanza) $E[aY + b] = aE[Y] + E[b] = aE[Y] + b$

2) $V[Y] = E[Y^2] - E[Y]^2$

3) $COV(X, a) = E[Xa] - E[X]E[a] = aE[X] - E[X]a = 0$

4) (Linealidad de la Covarianza)

$COV(aX + b, Y) = aCOV(X + b, Y) =$

$= a[COV(X, Y) + COV(b, Y)] = aCOV(X, Y)$

5) X e Y independientes$\Rightarrow COV(X, Y) = 0$

6) $V[aY] = a^2V[Y]$

7) $V[X + Y] = V[X] + V[Y] + 2COV(X, Y)$

Supongamos que $Y_1, ..., Y_n$ son observaciones independientes e idénticamente distribuidas de la variable aleatoria Y.

8) $E[\overline{Y}] = E[\frac{1}{n}\sum_{i=1}^{n} Y_i] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] = E[Y]$

ISBN 9788496866133

9) $V[\overline{Y}] = V[\frac{1}{n}\sum_{i=1}^{n} Y_i] = \frac{1}{n^2}\sum_{i=1}^{n} V[Y_i] + 2\sum_{i,j}^{n} COV(Y_i, Y_j) = \frac{1}{n}V[Y]$ pues $COV(Y_i, Y_j) = 0$ al ser Y_i e Y_j independientes.

Ejemplo 2.3.

En el Ejemplo 2.2 , supongamos que el muestreo se lleva a cabo con reemplazamiento y con probabilidades iguales, y se considera el orden en el resultado (caso (a)). Entonces la probabilidad de aparición de cada muestra es la misma: $p_i = \frac{1}{4^2} = \frac{1}{16}$ para cada una de las 16 muestras posibles. Como cada muestra lleva asociada una media muestral, se puede construir la variable aleatoria $\widehat{\overline{Y}}$ = "media muestral", que asocia probabilidad $\frac{1}{16}$ a cada uno de los valores respectivos 12, 15, 13.5, 9, 15, 18, 16.5, 12, 13.5, 16.5, 15, 10.5, 9, 12, 10.5, 6. Como hay valores repetidos, éstos se pueden agrupar (sumando las probabilidades correspondientes al mismo valor), y la tabla de probabilidades de la variable aleatoria $\widehat{\overline{Y}}$ = media muestral, queda:

$\widehat{\overline{y}}$	6	9	10.5	12	13.5	15	16.5	18
$p(\widehat{\overline{y}})$	1/16	2/16	2/16	3/16	2/16	3/16	2/16	1/16

Tabla 2.2. Probabilidades para la v.a. media muestral.

Se ha construido la tabla de valores y probabilidades asociadas al estadístico muestral $\widehat{\overline{y}}$. Análogamente se podría haber hecho para cualquier estadístico muestral de interés. Se ha visto entonces que los procesos aleatorios de selección muestral llevan a la construcción de variables aleatorias relacionadas. Por lo tanto permitirá el desarrollo de técnicas de inferencia adecuadas a los procesos de muestreo.

2.4 Estimadores

Se supone que θ es una característica poblacional de interés que deseamos estimar, por ejemplo, $\theta = \overline{y}$, y disponemos de valores de una muestra $y_1, ..., y_n$ proveniente de la población. La estimación consiste en utilizar un estadístico o función de la muestra para aproximar una característica poblacional.

Un **estimador de** θ es cualquier función de la muestra $T(y_1, ..., y_n)$ que utilizamos para estimar el parámetro poblacional θ. Un estimador es una variable aleatoria, pues depende de las observaciones $y_1, ..., y_n$, escogidas entre la población según el diseño de muestreo que hayamos elegido. Una propiedad importante en nuestro objetivo de aproximación es que el centro de gravedad de esta variable aleatoria, es decir, su esperanza matemática, coincida con lo que se desea estimar.

Así, se denomina **estimador insesgado** de θ a T si $E(T) = \theta$.

El **sesgo de un estimador** T se define como $sesgo(T) = B(T) = E(T) - \theta$.

Otra cuestión importante es la variabilidad del estimador T. Si éste es muy variable, los valores de T difieren mucho de una muestra a otra y por lo tanto la aproximación que daremos a θ

ISBN 9788496866133

depende mucho de la muestra obtenida. Se define por lo tanto la **precisión de un estimador** como el inverso de la varianza del estimador , $\frac{1}{V(T)}$.

De las dos definiciones anteriores se desprende que si un estimador es insesgado y su varianza pequeña, será un buen estimador, pues aunque cada muestra de lugar a un valor diferente de T estos valores serán muy similares entre sí y por lo tanto cercanos a su esperanza que coincide con θ . Además, si dos estimadores diferentes son insesgados el mejor de los dos será el que tenga menor varianza.

Una vez definido un estimador, es necesario dar una idea de su nivel de error respecto a lo que se quiere estimar. Para ello se utiliza su variabilidad. Hay varias maneras de expresar este error de muestreo, que veremos a continuación.

Error de muestreo de un estimador.

Es la desviación típica del estimador , $\sqrt{V(T)}$. En algunos textos se define el error de muestreo como el error de muestreo absoluto (ver definición más adelante).

Error de muestreo relativo de un estimador.

Es el coeficiente de variación del estimador, es decir, la desviación típica del estimador corregida por su esperanza: $EMR = \frac{\sqrt{V(T)}}{E(T)}$.

De gran importancia es la construcción de intervalos de confianza para la característica poblacional a partir de una estimación. El siguiente desarrollo permite construir el intervalo para estimadores aproximadamente normales e insesgados.

Error Cuadrático Medio de un estimador.

Este se define como $ECM(T) = sesgo^2(T) + V(T)$. Si el estimador es insesgado, $sesgo(T) = 0$ y por lo tanto $ECM(T) = V(T)$.

Intervalo de Confianza para una característica poblacional suponiendo estimador insesgado y normal.

Bajo la hipótesis de normalidad, el estimador T se distribuye como una normal con esperanza $E(T)$ y varianza $V(T)$, es decir, $T \equiv N(E(T), V(T))$.

Como T es insesgado, $T \equiv N(\theta, V(T))$. Así, $P(z_{\alpha/2} < \frac{T-\theta}{\sqrt{V(T)}} < -z_{\alpha/2}) = 1-\alpha$ y por lo tanto $P(T - z_{\alpha/2}\sqrt{V(T)} < \theta < T + z_{\alpha/2}\sqrt{V(T)}) = 1-\alpha$.

Usualmente se utiliza la estimación $\widehat{V}(T)$ para calcular este intervalo, pues $V(T)$ es desconocida. El intervalo de confianza aproximado para θ al $(1-\alpha)\%$ se calcula entonces como

$$(T - z_{\alpha/2}\sqrt{\widehat{V}(T)}, T + z_{\alpha/2}\sqrt{\widehat{V}(T)}).$$

Para intervalos de confianza al 95%, el valor de $z_{\alpha/2}$ es $z_{\alpha/2} = 1.96$ Para intervalos de confianza al 99%, es $z_{\alpha/2} = 2.57$ y para intervalos de confianza al 90%, es $z_{\alpha/2} = 1.64$.

ISBN 9788496866133

En general, aunque la distribución del estimador no sea exactamente normal, los intervalos de confianza creados suelen ser bastante robustos frente al fallo de la hipótesis de normalidad.

A partir de la noción de intervalo de confianza se define el error de muestreo absoluto de un estimador para cierto α prefijado.

Error de muestreo absoluto de un estimador.

Generalmente se denomina a la semianchura del intervalo de confianza suponiendo normalidad, $e = z_{\alpha/2}\sqrt{V(T)}$, error de muestreo absoluto para un nivel de confianza prefijado α.

Efecto de Diseño .

Es el cociente, para el mismo tamaño muestral, entre la varianza del estimador T bajo un cierto tipo de muestreo, y el estimador usual T' suponiendo muestreo aleatorio simple: $\dfrac{V(T)}{V_{m.a.s.}(T')}$.

Ejemplo 2.4.

Supongamos que en cada uno de los tipos de muestreo presentados en el Ejemplo 2.2 el muestreo se realiza con probabilidades iguales, de modo que la probabilidad de cada muestra es la misma. Supongamos el muestreo tipo (a), es decir, con reemplazamiento, y se tendrá en cuenta el orden. Supongamos que queremos comparar los siguientes estimadores de la media poblacional:

(i) La media muestral $\widehat{\overline{y}}$.

(ii) El mínimo muestral min que es el mínimo valor de los obtenidos en la muestra.

(iii) El máximo muestral max que es el máximo valor de los obtenidos en la muestra.

Obviamente los estimadores (ii) y (iii) van a ser malos estimadores de la media poblacional, y se presentan con el único objetivo de servir al ejemplo.

Para comparar los tres estimadores, vamos a calcular el sesgo, varianza y error cuadrático medio de cada uno de ellos. Como cada muestra tiene igual probabilidad $p = \frac{1}{16}$, se obtiene

(i) $E(\widehat{\overline{y}}) = \frac{1}{16}(12+15+13.5+9+15+18+16.5+12+13.5+16.5+15+10.5+9+12+10.5+6) = 12.75.$

El sesgo del estimador $\widehat{\overline{y}}$ es $E(T) - \theta = E(\widehat{\overline{y}}) - \overline{y} = 12.75 - 12.75 = 0$. Por lo tanto el estimador es insesgado. La varianza de $\widehat{\overline{y}}$ es $V(\widehat{\overline{y}}) = E[(\widehat{\overline{y}} - 12.75)^2] = \frac{1}{16}\sum_{i=1}^{16}(\widehat{\overline{y}}_i - 12.75)^2 = 3.137$. El error cuadrático medio de $\widehat{\overline{y}}$ es

$ECM(\widehat{\overline{y}}) = sesgo^2(\widehat{\overline{y}}) + V(\widehat{\overline{y}}) = 0 + 3.137 = 3.137.$

(ii) $E(\min) = \frac{1}{16}(12+12+12+6+12+18+15+6+12+15+15+6+6+6+6+6) = 10.3125$. El sesgo de min es $E(\min) - \overline{y} = 10.31 - 12.75 = -2.4375$ y por lo tanto el mínimo muestral, como estimador de la media poblacional, tiende a subestimar su valor.

ISBN 9788496866133

$V(\min) = \frac{1}{16}\sum_{i=1}^{16}(\min -10.31)^2 = 4.10$ y por lo tanto el error cuadrático medio será

$ECM(\min) = sesgo^2(\min) + V(\min) = (-2.4375)^2 + 4.10 = 9.74.$

(iii) $E(\max) = \frac{1}{16}(12+18+15+12+18+18+18+18+15+18+15+15+12+18+15+6) = 14.5$. El sesgo de max es $E(\max) - \overline{y} = 14.5 - 12.75 = 1.75$ y por lo tanto el máximo muestral, como estimador de la media poblacional, tiende a sobreestimar su valor.

$V(\max) = \frac{1}{16}\sum_{i=1}^{16}(\max -14.5)^2 = 4.7$ y por lo tanto el error cuadrático medio es

$ECM(\max) = sesgo^2(\max) + V(\max) = (1.75)^2 + 4.7 = 7.76.$

Una representación gráfica de los comportamientos de los tres estimadores en el orden del ejemplo se presenta en la Figura 2.2, donde cada punto negro representa el valor del estimador en cada muestra, y el valor de la verdadera media poblacional está representado por el círculo blanco.

Se observa en la Figura 2.2 cómo el estimador $\widehat{\overline{y}}$ es centrado o insesgado para la media poblacional $\overline{y} = 12.75$, mientras que los estimadores min y max no son insesgados, pues sus centros de gravedad difieren de $\overline{y} = 12.75$. Además la figura ayuda a intuir que estos dos últimos estimadores tienen mayor varianza que $\widehat{\overline{y}}$.

Por otra parte, al ser este diagrama en realidad un tipo de histograma, se observa cómo la forma de la distribución del estimador $\widehat{\overline{y}}$ corresponde a la forma de campana típica de la distribución normal (resultado asociado al Teorema Central del Límite, pues se conoce que la media muestral converge a una distribución normal). Los estimadores min y max parecen más bien distribuirse de forma asimétrica.

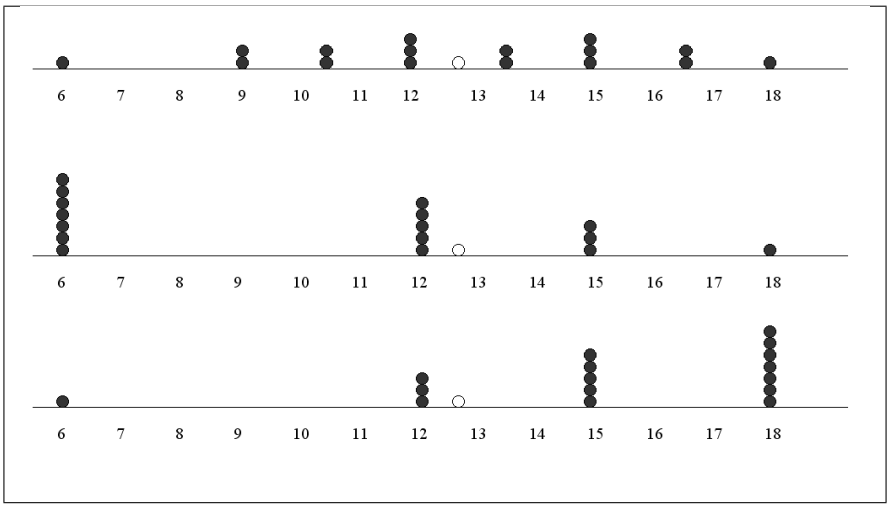


Figura 2.2. Estimaciones de $\overline{y}$ obtenidas con los estimadores $\widehat{\overline{y}}$, min y max.

Obviamente en la práctica no se pueden realizar este tipo de cálculos (obteniendo el valor del estimador para todas las muestras posibles), pues sólo se dispondrá de una muestra y no de

ISBN 9788496866133

todas las muestras posibles en la población. Pero se pueden hallar resultados teóricos que nos permitan anticipar si determinados estimadores son insesgados o no, o comparar la expresión de la varianza de diferentes estimadores, antes de obtener cualquier muestra, llegando a poder establecer estimadores apropiados para cada situación.

2.5 Sesgos no relacionados directamente con la forma del estimador

En la mayor parte de los estudios que tienen como base el muestreo, existen factores prácticos, en cierto modo externos a la forma del estimador, que hacen que la estimación tenga un sesgo importante, subestimando o sobreestimando la característica poblacional que se desea estimar. Una posible clasificación de estos sesgos es catalogarlos como **sesgos de selección** y **sesgos de medición**.

Respecto a los **sesgos de selección** , relacionados con la selección de la muestra, definición de la población y del marco, se pueden enumerar los siguientes:

- Sesgos debidos a la voluntariedad de los participantes. Es decir, si solamente se tienen en cuenta las encuestas de individuos que participan voluntariamente, puede que sean más propensos, en general, a responder de cierta manera a algunas preguntas que los participantes no encuestados por no ser voluntarios.
- Errores en la especificación de la población objetivo. Por ejemplo, en una encuesta electoral se puede entrevistar a aquellos individuos que están en una base de datos de votantes de anteriores elecciones, pudiendo incurrir en el sesgo de que en el presente existan nuevos votantes hacia otra línea política, no representados en el listado mencionado. Si además existen muchos indecisos que en la encuesta no declaran su preferencia pero en el último momento se declinan claramente por una preferencia específica, este tipo de ausencia de respuesta hace incurrir en un sesgo claro a las estimaciones.
- No incluir a toda la población objetivo, lo que se denomina subcobertura. Si se utilizan listados o censos que datan de mucho tiempo atrás, puede que gran parte de la población objetivo actual no esté representada (si por ejemplo ha habido mucha inmigración en el periodo transcurrido desde el listado utilizado).
- Sustituciones en el trabajo de campo. A veces, las personas que no suelen estar en el hogar opinan de modo muy diferente a las que suelen estar más tiempo en él, por lo que la sustitución de las personas ausentes puede llevar a un sesgo importante en ciertos casos.

Los **sesgos de medición** están relacionados con el proceso de obtener datos de las unidades elementales muestreadas. Algunos ejemplos prácticos son:

- En ocasiones las personas no dicen la verdad. Si se realiza una encuesta agrícola, las personas pueden dar respuestas sobre la producción inferiores a los datos reales con la esperanza de obtener mayores subvenciones. En encuestas sobre consumo de drogas u homosexualidad las respuestas son delicadas y a menudo hay un sesgo hacia la infraestimación .

ISBN 9788496866133

- Las personas no siempre comprenden las preguntas. Si la pregunta contiene términos poco habituales las personas pueden responder al azar, sin pedir más información, para evitar mostrar su ignorancia. La ambigüedad de ciertas preguntas también arroja problemas de sesgo.

- Las personas pueden dar diferentes respuestas a diferentes entrevistadores, o en diferentes situaciones. La forma en que el entrevistador pregunta puede inducir a la persona a responder de un modo u otro.

- En mediciones supuestamente objetivas (medicina, biología, procesos industriales), diferentes investigadores pueden adoptar ligeras variaciones que inciden en un sesgo en la medición. Asímismo, el instrumento de medición o las condiciones (clima, hora, etc.) pueden introducir sesgos en la variable respuesta.

ISBN 9788496866133

2.6 Ejercicios resueltos

Ejercicio 1.1.

Supongamos una urna con $N = 4$ bolas numeradas del 1 al 4, De ellas las tres primeras son blancas y la cuarta negra.

1) Decir cuántas muestras diferentes hay en los siguientes supuestos:

a) Se extraen dos bolas con reemplazamiento, teniendo en cuenta el orden.

b) Se extraen dos bolas sin reemplazamiento, teniendo en cuenta el orden.

c) Se extraen dos bolas con reemplazamiento, sin tener en cuenta el orden.

d) Se extraen dos bolas sin reemplazamiento, sin tener en cuenta el orden.

2) Decir cuántas muestras contienen la bola negra en cada uno de los casos anteriores, y presentar estas muestras.

3) Si se extraen 3 bolas, presentar la distribución del estimador de la media poblacional "media muestral de los números obtenidos", en el caso d) de los anteriores, suponiendo que todas las bolas tienen la misma probabilidad de ser extraídas. Hallar la esperanza y varianza de este estimador.

4) En el supuesto de extraer tres bolas, presentar la distribución del estimador de la proporción poblacional de bolas negras "proporción muestral de bolas negras en las bolas obtenidas", suponiendo que todas las bolas tienen la misma probabilidad de ser extraídas (y por lo tanto todas las muestras posibles tienen la misma probabilidad).

5) Verificar que la varianza del estimador, en el apartado 3), es igual a $\frac{N-n}{N}\frac{S^2}{n}$, donde S^2 es la cuasivarianza muestral de la población, y N es el tamaño poblacional=4.

---

1)

a) Hay $N^n = 4^2 = 16$ muestras diferentes en muestreo con reemplazamiento, teniendo en cuenta el orden.

b) Hay $n!\binom{N}{n} = 2!\binom{4}{2} = 12$ muestras diferentes en muestreo sin reemplazamiento, teniendo en cuenta el orden.

c) Hay $\binom{N+n-1}{n} = \binom{4+2-1}{2} = 10$ muestras diferentes en muestreo con reemplazamiento, sin tener en cuenta el orden.

d) Hay $\binom{N}{n} = \binom{4}{2} = 6$ muestras diferentes en muestreo sin reemplazamiento, sin tener en cuenta el orden.

2) Hay que contar las muestras que contienen dos de las tres bolas blancas en cada caso. Las demás contienen al menos una vez la bola negra.

a) Hay $3^2 = 9$ muestras diferentes teniendo en cuenta sólo las bolas blancas. Como en total había 16 muestras, habrá $16 - 9 = 7$ muestras que contienen la bola negra, que son $(1,4)$, $(4,1)$, $(2,4)$, $(4,2)$, $(3,4)$, $(4,3)$, $(4,4)$).

b) Hay $2!\binom{3}{2} = 6$ muestras diferentes teniendo en cuenta sólo las bolas blancas. Como en total había 12 muestras, habrá $12 - 6 = 6$ muestras que contienen la bola negra

ISBN 9788496866133

(que son $(1,4)$, $(4,1)$, $(2,4)$, $(4,2)$, $(3,4)$, $(4,3)$).

c) Hay $\binom{3+2-1}{2} = 6$ muestras diferentes teniendo en cuenta sólo las bolas blancas. Como en total había 20 muestras, habrá $10-6=4$ muestras que contienen la bola negra

(que son $\{1,4\},\{2,4\},\{3,4\},\{4,4\}$).

d) $\binom{3}{2} = 3$ muestras diferentes teniendo en cuenta sólo las bolas blancas. Como en total había 6 muestras, habrá $6-3=3$ muestras que contienen la bola negra (que son $\{1,4\}$, $\{2,4\}$, $\{3,4\}$).

3)

Muestra	Estimador $\widehat{\overline{y}}$	$p_{muestra}$
$\{1,2,3\}$	2	1/4
$\{1,2,4\}$	7/3	1/4
$\{1,3,4\}$	8/3	1/4
$\{2,3,4\}$	3	1/4

La esperanza es: $E(\widehat{\overline{y}}) = \frac{1}{4}(2+\frac{7}{3}+\frac{8}{3}+3) = 2.5$.

La varianza es: $V(\widehat{\overline{y}}) = \frac{1}{4}((2-2.5)^2+(\frac{7}{3}-2.5)^2+(\frac{8}{3}-2.5)^2+(3-25.)^2) = 0.1388$.

4)

Muestra	Estimador $\widehat{p}$	$p_{muestra}$
$\{1,2,3\}$	0	1/4
$\{1,2,4\}$	1/3	1/4
$\{1,3,4\}$	1/3	1/4
$\{2,3,4\}$	1/3	1/4

La esperanza es: $E(\widehat{p}) = \frac{1}{4}(0+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}) = 0.25$.

La varianza es: $V(\widehat{p}) = \frac{1}{4}((0-0.25)^2+(\frac{1}{3}-0.25)^2+(\frac{1}{3}-0.25)^2+(\frac{1}{3}-0.25)^2) = 0.0208$.

5) Como en la población es $\overline{y} = \frac{1}{4}(1+2+3+4) = 2.5$ y además $S^2 = \frac{1}{4-1}(1+2^2+3^2+4^2-4\cdot 2.5^2) = \frac{5}{3}$, tenemos que

$$\frac{N-n}{N}\frac{S^2}{n} = \frac{4-3}{4}\frac{5/3}{3} = 0.1388.$$

Ejercicio 1.2.

Se pretende realizar un proceso de muestreo sin reemplazamiento con probabilidades desiguales de $n=2$ unidades en una población de 4 donde los valores de la variable de interés y son respectivamente $y_1=2$, $y_2=3$, $y_3=5$, $y_4=1$.

ISBN 9788496866133

Se sabe que con este esquema de muestreo las probabilidades de selección de cada una de las muestras posibles está en la tabla siguiente:

Muestra	$p_{muestra}$
$\{1,2\}$	0.047
$\{1,3\}$	0.076
$\{1,4\}$	0.111
$\{2,3\}$	0.160
$\{2,4\}$	0.233
$\{3,4\}$	0.371

Estudiar el estimador de la media poblacional "media muestral" , calculando su sesgo, varianza, error cuadrático medio, error de muestreo y error de muestreo relativo .

Calculando el valor del estimador :

Muestra	$p_{muestra}$	$\widehat{\overline{y}}_{muestra}$
$\{1,2\}$	0.047	2.5
$\{1,3\}$	0.076	3.5
$\{1,4\}$	0.111	1.5
$\{2,3\}$	0.160	4
$\{2,4\}$	0.233	2
$\{3,4\}$	0.371	3

donde cada $\widehat{\overline{y}}_{muestra}$ se calcula a partir de los valores de la variable de interés en la muestra obtenida . Por ejemplo, para la muestra $\{1,2\}$, se tiene $\widehat{\overline{y}}_{muestra} = \frac{1}{2}(y_1 + y_2) = \frac{1}{2}(2+3) = 2.5$.

La esperanza del estimador es

$$E(\widehat{\overline{y}}) = \sum p_{muestra}\widehat{\overline{y}}_{muestra} = 0.047 \cdot 1.5 + ... + 0.371 \cdot 3.5 = 2.90.$$

La varianza es

$$V(\widehat{\overline{y}}) = E(\widehat{\overline{y}} - E(\widehat{\overline{y}}))^2 = \sum p_{muestra}(\widehat{\overline{y}}_{muestra} - 2.90)^2 = 0.333.$$

Como la verdadera media poblacional es $\overline{y} = 2.75$, el sesgo del estimador es $sesgo(\widehat{\overline{y}}) = E(\widehat{\overline{y}}) - \overline{y} = 2.90 - 2.75 = 0.15$.

El error cuadrático medio es

ISBN 9788496866133

$ECM(\widehat{\overline{y}}) = sesgo^2(\widehat{\overline{y}}) + V(\widehat{\overline{y}}) = 0.0225 + 0.333 = 0.3555.$

El error de muestreo se ha definido como la desviación típica del estimador. Por lo tanto es $\sqrt{V(\widehat{\overline{y}})} = 0.577$. El error de muestreo relativo de un estimador es $EMR = \dfrac{\sqrt{V(T)}}{E(T)}$. En este caso, $EMR = \dfrac{\sqrt{V(\widehat{\overline{y}})}}{E(\widehat{\overline{y}})} \simeq 0.20$.

Ejercicio 1.3.

Se dispone de la población de cinco observaciones con valores $y_1 = 1, y_2 = 4, y_3 = 7, y_4 = 2$, $y_5 = -2$. Se pretende tomar por muestreo sin reemplazamiento y probabilidades iguales, muestras de tamaño $n = 4$. Debido al método de muestreo, todas las muestras tienen la misma probabilidad.

1) Presentar el espacio muestral asociado.al experimento, sin tener en cuenta el orden en las muestras.

2) Supongamos que queremos estimar el mínimo poblacional a partir del mínimo muestral. Estudiar el estimador, hallando su varianza, sesgo y error cuadrático medio.

3) Supongamos que deseamos estimar la media poblacional a partir de la mediana muestral, definiendo la mediana muestral como la primera observación de la muestra ordenada de menor a mayor, que deja a su izquierda al menos un 50% estricto de los valores muestrales. Estudiar el estimador, hallando su varianza, sesgo y error cuadrático medio.

---

1) Al tratarse de muestreo sin reemplazamiento sin tener en cuenta el orden, hay $\binom{5}{4} = 5$ muestras posibles. Estas son:

$\{1,2,3,4\}, \{1,2,3,5\}, \{1,2,4,5\}, \{1,3,4,5\}, \{2,3,4,5\}$.

2) Construimos la tabla con los valores del estimador asociado a cada muestra:

Muestra	$p_{muestra}$	$min_{muestra}$
$\{1,2,3,4\}$	1/5	2
$\{1,2,3,5\}$	1/5	-2
$\{1,2,4,5\}$	1/5	-2
$\{1,3,4,5\}$	1/5	-2
$\{2,3,4,5\}$	1/5	-2

pues por ejemplo, la muestra $\{1,2,3,4\}$ corresponde a los valores $\{1,4,7,2\}$, cuyo mínimo es 2.

La esperanza del estimador mínimo muestral es

$$E(\widehat{\min}) = \sum p_{muestra} \min{}_{muestra} = \frac{1}{5}(2 - 2 - 2 - 2 - 2) = -1.2.$$

La varianza es:

$$V(\widehat{\min}) = E(\widehat{\min} - E(\widehat{\min}))^2 = \frac{1}{5}[(2 + 1.2)^2 + 4 \cdot (-2 + 1.2)^2] = 2.56.$$

ISBN 9788496866133

El mínimo poblacional es $\min = -2$, y por lo tanto el sesgo del estimador "mínimo muestral" es $E(\widehat{\min}) - \min = -1.2 + 2 = 0.8$.

El error cuadrático medio será

$ECM(\widehat{\min}) = sesgo^2(\widehat{\min}) + V(\widehat{\min}) = 0.8^2 + 2.56 = 3.2$.

3) Se construye la tabla igual que en el apartado anterior:

Muestra	$p_{muestra}$	$Med_{muestra}$
$\{1,2,3,4\}$	1/5	4
$\{1,2,3,5\}$	1/5	4
$\{1,2,4,5\}$	1/5	2
$\{1,3,4,5\}$	1/5	2
$\{2,3,4,5\}$	1/5	4

$E(Med) = \sum p_{muestra} Med_{muestra} = \frac{1}{5}(4+4+2+2+4) = 3.2$

La varianza es:

$V(Med) = E(Med - E(Med))^2 = \frac{1}{5}[(4-3.2)^2 + ... + (4-3.2)^2] = 0.96$.

La media poblacional es $\overline{y} = 2.4$, y por lo tanto el sesgo del estimador "mediana muestral" es $E(Med) - \overline{y} = 3.2 - 2.4 = 0.8$.

El error cuadrático medio será

$ECM(\widehat{\min}) = sesgo^2(\widehat{\min}) + V(\widehat{\min}) = 0.8^2 + 0.96 = 1.6$.

Ejercicio 1.4.

Se dispone de una urna que contiene 5 bolas, 2 con el número 1, 2 bolas con el número 2 y 1 bola con el número 3. Se extraen dos bolas sin reposición y asignando probabilidades iguales a todas las bolas de la urna. Debido al método de muestreo, todas las muestras tienen la misma probabilidad. Sea t el estadístico=suma de los números obtenidos en las dos bolas. Presentar la distribución del estadístico, dibujar el gráfico de su distribución de probabilidad y calcular su esperanza y varianza.

Para el cálculo del estadístico no interviene el orden, con lo cual hay $\binom{5}{2} = 10$ muestras posibles. Si numeramos las bolas del 1 al 5, sabiendo que las bolas 1 y 2 tienen el número 1, las 3 y 4 el número 2 y la 5 el número 3, obtenemos la siguiente tabla:

ISBN 9788496866133

Muestra	Valores	t	$p_{muestra}$
$\{1,2\}$	$\{1,1\}$	2	1/10
$\{1,3\}$	$\{1,2\}$	3	1/10
$\{1,4\}$	$\{1,2\}$	3	1/10
$\{1,5\}$	$\{1,3\}$	4	1/10
$\{2,3\}$	$\{1,2\}$	3	1/10
$\{2,4\}$	$\{1,2\}$	3	1/10
$\{2,5\}$	$\{1,3\}$	4	1/10
$\{3,4\}$	$\{2,2\}$	4	1/10
$\{3,5\}$	$\{2,3\}$	5	1/10
$\{4,5\}$	$\{2,3\}$	5	1/10

Las probabilidades de las muestras para los valores iguales de t se agregan, de manera que

$P(t=2) = \sum P(muestras \mid t=2) = 1/10$

$P(t=3) = \sum P(muestras \mid t=3) = 1/10 + 1/10 + 1/10 + 1/10 = 4/10$

$P(t=4) = \sum P(muestras \mid t=4) = 1/10 + 1/10 + 1/10 = 3/10$

$P(t=5) = \sum P(muestras \mid t=5) = 1/10 + 1/10 = 2/10$

Se puede comprobar que la suma de las probabilidades de t para sus posibles valores es 1.

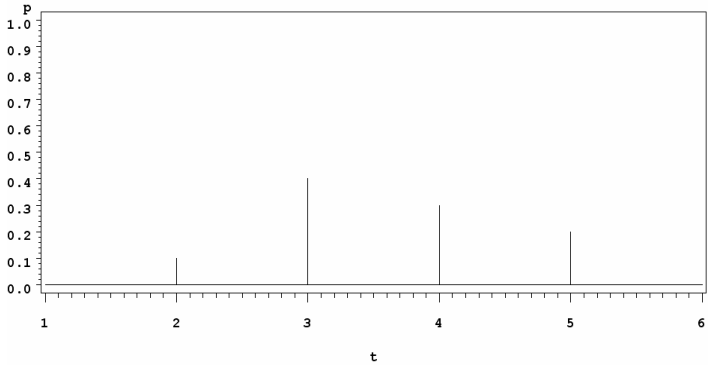


Figura 1.1. Tabla de probabilidades del estadístico t

La Figura 1.1 muestra la tabla de probabilidades. En cuanto a la esperanza y varianza:

$$E(T) = \sum p(T=t)t = \frac{1}{10} \cdot 2 + \frac{4}{10} \cdot 3 + \frac{3}{10} \cdot 4 + \frac{2}{10} \cdot 5 = 3.6$$

La varianza es:

$$V(T) = E(T - E(T))^2 = \sum p(T=t)(t-3.6)^2 = 0.728.$$

ISBN 9788496866133

Ejercicio 1.5

Se dispone de una población de 4 observaciones con las variables x e y, según aparecen en la tabla:

Observación	x	y
1	1	1
2	2	2.5
3	0	0.2
4	3	3.1

Se extraen muestras sin reemplazamiento y con probabilidades iguales de tamaño 3.

1) Presentar la distribución del estimador de la cuasicovarianza poblacional entre x e y, "cuasicovarianza muestral", calculando su sesgo.

2) Presentar la distribución del estimador de la cuasi-desviación típica de x, "cuasi-desviación típica muestral de x", calculando su sesgo.

3) Utilizar los resultados del apartado 2) para estudiar la distribución del estimador de la cuasivarianza de x, "cuasivarianza muestral de x", calculando su sesgo.

---

1) Hay $\binom{4}{3} = 4$ muestras posibles con igual probabilidad. Calcularemos el valor de la cuasicovarianza muestral para cada una de estas muestras.

Muestra $\{1, 2, 3\}$:

$\widehat{\overline{x}} = 1, \widehat{\overline{y}} = 1.23$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{\overline{y}})(x_i - \widehat{\overline{x}}) = \frac{1}{3-1}[(1-1.23)(1-1) + (2.5-1.23)(2-1) + (0.2-1.23)(0-1)] = 1.15$$

Igualmente se obtiene para las demás muestras:

Muestra $\{1, 2, 4\}$: $s_{xy} = 1.05$

Muestra $\{1, 3, 4\}$: $s_{xy} = 2.28$

Muestra $\{2, 3, 4\}$: $s_{xy} = 2.31$

Cada uno de estos cuatro valores tiene probabilidad $\frac{1}{4}$, que es la probabilidad de cada una de las muestras. La distribución de s_{xy} es por lo tanto,

ISBN 9788496866133

s_{xy}	p
1.05	0.25
1.15	0.25
2.28	0.25
2.31	0.25

La esperanza de s_{xy} es

$E(s_{xy}) = \sum 0.25(1.05 + 1.15 + 2.28 + 2.31) = 1.7$

El valor de la cuasicovarianza poblacional es

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})(x_i - \overline{x}) = 1.7$$

Por lo tanto el sesgo del estimador es cero (es un estimador insesgado con este tipo de muestreo).

2) Actuando análogamente al apartado anterior, se obtiene:

Muestra $\{1, 2, 3\} : s_x = 1$

Muestra $\{1, 2, 4\} : s_x = 1$

Muestra $\{1, 3, 4\} : s_x = 1.527$

Muestra $\{2, 3, 4\} : s_x = 1.527$

La cuasi desviación típica poblacional es S, con

$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 = 1.666$ y por lo tanto $S = 1.290944$.

La esperanza de s_x es $\frac{1}{4}(1+1+1.527+1.527) = 1.2635$. El estimador es sesgado, con sesgo aproximado -0.027444.

3) Elevando al cuadrado las cuasi desviaciones típicas obtenidas, se tiene:

Muestra $\{1, 2, 3\} : s_x^2 = 1$

Muestra $\{1, 2, 4\} : s_x^2 = 1$

Muestra $\{1, 3, 4\} : s_x^2 = 2.3317$

Muestra $\{2, 3, 4\} : s_x^2 = 2.3317$

La esperanza de s_x^2 es $\frac{1}{4}(1+1+2.3317+2.3317) = 1.666 = S^2$ y por lo tanto el estimador cuasivarianza muestral es insesgado para la cuasivarianza poblacional.

Ejercicio 1.6

En la siguiente frase:

"En un lugar de la Mancha, de cuyo nombre no quiero acordarme, vivía un hidalgo de los de lanza en astillero..." , si se escogen dos palabras al azar y sin reemplazamiento, con igual probabilidad (con lo cual todas las muestras tienen la misma probabilidad), decir:

ISBN 9788496866133

1) Cuántas muestras contienen al menos un artículo o una preposición. Calcular la probabilidad de que alguna palabra de la muestra sea un artículo o preposición.

2) Calcular la probabilidad de que alguna palabra de la muestra sea un verbo.

3) Calcular la probabilidad de que alguna palabra de la muestra sea un verbo o artículo o preposición.

---

1) Hay en total $N = 21$ palabras, de las cuales 10 son artículos o preposiciones. Por tanto hay 11 palabras que no son ni artículo ni preposición. Hay en total $\binom{21}{2} = 210$ muestras. De ellas, $\binom{11}{2} = 55$ no contienen algún artículo o preposición. Por lo tanto, hay 210-55=155 muestras que contienen al menos un artículo o preposición. Como todas las muestras tienen igual probabilidad, ésta debe ser $p_i = \frac{1}{\binom{21}{2}}$ (para que se cumpla la condición de que la suma de las probabilidades sea 1).

Por lo tanto la probabilidad de que una muestra contenga un artículo o preposición es la suma de las probabilidades de todas las muestras que contienen un artículo o preposición, es decir: $55 \cdot \frac{1}{\binom{21}{2}} = \frac{55}{210} = 0.26$.

2) Hay en la frase dos verbos, con lo cual hay $\binom{19}{2} = 171$ muestras que no contienen verbos. La probabilidad de que una muestra contenga algún verbo es por lo tanto, razonando como en el apartado anterior, $(210 - 171) \cdot \frac{1}{\binom{21}{2}} = \frac{39}{210} = 0.185$.

3) Razonando como en los dos apartados anteriores, la probabilidad es $(210 - \binom{21-12}{2}) \cdot \frac{1}{\binom{21}{2}} = 0.828$.

Ejercicio 1.7.

Tener en cuenta el orden o no no tiene efecto en la mayoría de los procesos de muestreo sin reemplazamiento en la práctica. Por ello el cálculo de probabilidades sobre las muestras y estimadores se suele realizar en general sin tener en cuenta el orden, pues la probabilidad de las muestras sólo es alterada por una constante. En muestreo con reemplazamiento, sin embargo, la probabilidad de las muestras varía de muestra a muestra si no se tiene en cuenta el orden, como se verá en el tema siguiente.

Supongamos una población de 3 unidades, numeradas del 1 al 3, en la cual se extrae una muestra de tamaño $n = 2$, asignando igual probabilidad a todas las unidades y por lo tanto a todas las muestras. Se calcula a continuación la suma de los valores obtenidos, que llamaremos s. Se pide:

1) Presentar todas las muestras en los casos siguientes:

a) Extracción sin reemplazamiento, teniendo en cuenta el orden.

b) Extracción sin reemplazamiento, sin tener en cuenta el orden.

2) Presentar la tabla de probabilidades de s en cada caso, y comparar los resultados.

---

1) a) Hay $n!\binom{N}{n} = 2!\binom{3}{2} = 6$ muestras diferentes en muestreo sin reemplazamiento, teniendo en cuenta el orden: $(1,2), (1,3), (2,1), (2,3), (3,1), (3,2)$.

b) Hay $\binom{N}{n} = \binom{3}{2} = 3$ muestras diferentes en muestreo sin reemplazamiento, sin tener en cuenta el orden: $\{1,2\}, \{1,3\}, \{2,3\}$.

2)

ISBN 9788496866133

a)

Muestra	$(1,2)$	$(1,3)$	$(2,1)$	$(2,3)$	$(3,1)$	$(3,2)$
s	3	4	3	5	4	5
p	1/6	1/6	1/6	1/6	1/6	1/6

Las probabilidades son:

s	p_s
3	2/6
4	2/6
5	2/6

b)

Muestra	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$
s	3	4	5
p	1/3	1/3	1/3

La tabla de probabilidades de s es:

s	p_s
3	1/3
4	1/3
5	1/3

La distribución de s es la misma en los dos casos.

Ejercicio 1.8

Supongamos la población con los siguientes datos:

ISBN 9788496866133

Observación	y	x
1	1	1
2	2	2.2
3	0.5	0.2
4	3	3.1

Se extraerá una muestra de tamaño $n = 2$ sin reemplazamiento y con igual probabilidad . Se desea estimar la llamada razón poblacional, definida por $R = \dfrac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$ a través de la razón muestral, definida por $\widehat{R} = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$. Hallar el sesgo del estimador para los datos de este ejercicio.

La distribucíón del estimador puede verse en la siguiente tabla:

Muestra	$p_{muestra}$	$\widehat{R}_{muestra}$
$\{1,2\}$	1/6	0.9375
$\{1,3\}$	1/6	1.25
$\{1,4\}$	1/6	0.975
$\{2,3\}$	1/6	1.041
$\{2,4\}$	1/6	0.943
$\{3,4\}$	1/6	1.06

Donde, por ejemplo, se ha calculado para la muestra $\{1,2\}$:

$$\widehat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{1+2}{1+2.2} = 0.9375.$$

Se calcula la esperanza de $\widehat{R}$:

$E(\widehat{R}) = \sum \frac{1}{6}(0.9375 + ... + 1.06) = 1.034.$

La razón poblacional es $R = \dfrac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = R = \dfrac{6.5}{6.5} = 1$, con lo cual el sesgo de $\widehat{R}$ es $E(\widehat{R}) - R = 0.034$.

ISBN 9788496866133

Ejercicio 1.9.

En el ejercicio 1.1, apartado 3), suponer la hipótesis de normalidad y construir un intervalo de confianza al 95% para la media poblacional, suponiendo que ha salido la muestra $\{1, 2, 3\}$, y suponiendo conocida la varianza del estimador, que era $.V(\widehat{\overline{y}}) = 0.1388$

En el caso de la muestra $\{1, 2, 3\}$, la media muestral es $\widehat{\overline{y}} = 2$. Al ser $z_{\alpha/2} = 1.96$ para el intervalo al 95% , el intervalo de confianza se construye como $(\widehat{\overline{y}} - \sqrt{V(\widehat{\overline{y}})}z_{\alpha/2}, \widehat{\overline{y}} + \sqrt{V(\widehat{\overline{y}})}z_{\alpha/2}) = (2 - \sqrt{0.1388} \cdot 1.96, 2 + \sqrt{0.1388} \cdot 1.96) = (1.269, 2.73)$.

Ejercicio 1.10

El procedimiento proc means del SAS calcula estadísticos descriptivos muestrales univariantes. Su sintaxis básica es:

```
proc means data=archivo;
var variables;
run;
```

Por defecto aparecen en la ventana output las medias, cuasi-desviaciones típicas, máximo y mínimo de las variables presentes en el archivo, a menos que se nombren variables en el apartado var variables, en cuyo caso aparecen los estadísticos solamente para las variables mencionadas.

Si se requieren otros estadísticos como varianza (VAR), suma (SUM) o coeficiente de variación (CV), hay que especificarlo mediante la palabra clave de la siguiente manera:

```
proc means data=archivo mean std var sum cv;
var variables;
run;
```

Para calcular covarianzas, se utiliza el procedimiento proc corr, que también calcula los coeficientes de correlación. Su sintaxis básica es:

```
proc corr data=archivo cov;
var variables;
run;
```

Por defecto presenta la matriz de (cuasi) varianzas-(cuasi)covarianzas y la matriz de correlaciones de todas las variables presentes en el archivo, a menos que se nombren variables en el apartado var variables, en cuyo caso aparecen solamente para las variables mencionadas.

Utilizar el procedimiento means y el procedimiento corr del SAS para calcular las medias, cuasi-varianzas, cuasi-desviaciones típicas y cuasi-covarianzas poblacionales y muestrales para todas las muestras del ejercicio.

ISBN 9788496866133

En primer lugar se crea un archivo de datos temporal con la información poblacional, para a continuación aplicar los procedimientos:

```
data uno;
input x y;
cards;
1 1
2 2.5
0 0.2
3 3.1
;
proc means data=uno mean var std;
run;
proc corr data=uno cov;run;
```

Las salidas en la ventana output son:

```
                                                Desviacion
Variable           Media        Varianza         estandar
-----------------------------------------------------------
x              1.5000000       1.6666667       1.2909944
y              1.7000000       1.7800000       1.3341664
-----------------------------------------------------------
```

para el procedimiento means, y

```
            Matriz de covarianza
                   x                 y
x       1.666666667       1.700000000
y       1.700000000       1.780000000
```

para el procedimiento corr, donde se observa que las varianzas de x e y están en la diagonal izda,arriba-dcha, abajo y la cuasivarianza poblacional está en la otra diagonal.

Para realizar los cálculos sobre todas las muestras, basta ir borrando cada vez una observación del archivo uno y repetir el programa. Por ejemplo, para la muestra $\{1, 2, 3\}$:

```
data uno;
input x y;
cards;
1 1
2 2.5
0 0.2
;
proc means data=uno mean var std;
run;
proc corr data=uno cov;run;
```


ISBN 9788496866133

2.7 Ejercicios propuestos

1) La siguiente tabla presenta una población de 5 hospitales denotados por A, B, C, D, y E, indicando el número de camas por hospital.

Hospital	Número de camas
A	160
B	220
C	850
D	510
E	110

a) Calcular el número medio de camas y la desviación típica del número de camas de la población de 5 hospitales.

b) ¿Cuántas muestras no ordenadas de 2 hospitales, sin reemplazamiento, pueden obtenerse?

c) enumera cada una de las posibles muestras de 2 hospitales y calcula el número medio de camas por hospital para cada muestra.

d) Suponiendo que las muestras del apartado c) son equiprobables, calcula la medi y varianza de la distribución de medias muestrales. Comparar esta media y la raíz cuadrada de esta varianza (desviación típica) con la media y desviación ípica poblacionales, respectivamente.

e) ¿Cuántas muestras (no ordenadas) de 4 hospitales pueden obtenerse de esta población? Especificarlas, junto con sus respectivas medias muestrales.

f) Calcula la esperanza y varianza de las medias muestrales del apartado e) y compáralas con la esperanza y varianza obtenidas en el apartado d).

2) Como una sección de un estudio de mercado, se eligió una manzana de 4 casas, y de ella Jermías seleccionó una muestra de 2 casas del siguiente modo. Identificó la casa numerándolas del 1 al 4, y cuando estableció la lista de muestras sin reemplazamiento ni orden de tamaño 2, lo hizo así:

$1 \rightarrow \{1, 2\}$

$1 \rightarrow \{1, 3\}$

$1 \rightarrow \{1, 4\}$

$1 \rightarrow \{2, 3\}$

$1 \rightarrow \{2, 4\}$

ISBN 9788496866133

ólvidándose, desgraciadamente, de la combinación $\{3, 4\}$. Eligió aleatoriamente un número entre 1 y 5 (resultó ser el4) que correspondió a la combinación $\{2, 3\}$, por lo que se muestrearon las casas 2 y 3. La variable de interés fue el gasto médico familiar anual, gasto que se muestra en la siguiente tabla:

Familia	Gastos
1	345
2	126
3	492
4	962

a) basándote en el procedimiento de muestreo de Jeremías, ¿cuál es la media, desviación típica y ECM del estimador del gasto médico medio familiar anual, suponiendo el estimador media muestral?

b) ¿Es insesgado este plan de muestreo?

c)¿Tiene cada casa la misma probabilidad de formar parte de la muestra? ¿Por qué?

3) En una muestra aleatoria de 80 cojinetes para cigüeñal de automóvil, 15 de ellos tienen un acabado de superficie más áspero de lo que permiten las especificaciones. Suponiendo que es apropiada la aproximación Normal a la Binomial, encontrar un intervalo de confianza al 95% para la fracción de cojinetes no conformes.

4) Por experiencias anteriores se sabe que un estimador insesgado de θ tiene una desviación típica del 1.8%. Si una muestra proporciona la estimación 12.508, determinar el intervalo de confianza para θ, con $\alpha = 0.05$, admitiendo la hipótesis de normalidad.

5) La calificaciones de 10 alumnos en un determinado ejericcio fueron: tres cincos, cuatro seises y tres sietes. Obtener todas las posibles parejas de notas, sin tener en cuenta el orden, que se pueden realizar con las calificaciones dadas, así como sus correspondientes probabilidades si las extracciones se realizan:

a) Con reposición

b) Sin reposición.

6) Dada una población con 5 elementos, se extrae una muestra de tamaño 3 mediante el siguiente procedimiento: de un urna con tres bolas marcadas con los números 1,2,3 se extraen al azar y sin reposición, dos bolas; a continuación, de otra urna con dos bolas numeradas 4 y 5, se extrae una de éstas. Establecer las muestras posibles y sus probabilidades.

7) Considérese el siguiente método de selección de una muestra de tamaño 2 en una población de 3 unidades $\{u_1, u_2, u_3\}$: se extrae una primera unidad con probabilidades iguales de selección, si ésta resulta ser u_1 se extrae la segunda entre las dos restantes con probabilidades iguales; si la primera no es u_1, la segunda se extrae de las tres que componen la población, asignando doble

ISBN 9788496866133

probabilidad a u_1 que a cada una de las otras dos. Calcula la distribución de probabilidad en el conjunto de las muestras.

8) Las calificaciones de un alumno en los tres exámenes parciales de una asignatura han sido: 1, 2, y 3.

a) Obtener la calificación media.

b) Si extraemos muestras de dos calificaciones con reposición y teniendo en cuenta el orden, enumerar todas las posibles muestras, calcular sus correspondientes probabilidades y la distribución de la media muestral, su esperanza y varianza.

c) Como el apartado b), pero las extracciones son ahora sin reposición y sin tener en cuenta el orden.

ISBN 9788496866133

3 MUESTREO ALEATORIO SIMPLE CON REEMPLAZAMIENTO (m.a.s.r.)

El muestreo aleatorio simple con reemplazamiento es matemáticamente el tipo de muestreo más sencillo. Consiste en escoger con igual probabilidad cada unidad elemental, y repetir este proceso n veces, siendo n el tamaño deseado de la muestra. Como se verá más adelante, en la práctica es más usual el muestreo aleatorio simple sin reemplazamiento, al ser de precisión mayor. Sin embargo, el tipo de muestreo que nos ocupa en este tema sirve siempre de introducción para sentar las bases de modelos de muestreo más complejos, pues es más sencillo de tratar, y su precisión representará siempre una cota inferior de interés respecto al muestreo aleatorio simple sin reemplazamiento.

Existen ocasiones prácticas en que el muestreo con reemplazamiento es más adecuado que el muestreo sin reemplazamiento, como por ejemplo estimar el peso de peces en un lago (cada vez que se pesca uno es necesario devolverlo al lago para que no muera), o estudios de medidas repetidas.

3.1 Propiedades básicas

3.1.1 Probabilidades de obtención de muestras

En muestreo aleatorio simple con reemplazamiento (m.a.s.r.) se seleccionan n unidades de una población de N unidades, con reemplazamiento e igual probabilidad. En cada extracción de las n extracciones independientes, cada unidad tiene por lo tanto probabilidad $\frac{1}{N}$ de ser escogida. La probabilidad de una muestra concreta ordenada $(u_1, u_2, ..., u_n)$ es, por tratarse de extracciones independientes, el producto de las probabilidades de cada unidad u_i, es decir, $\frac{1}{N^n}$. Por lo tanto :

$$P((u_1, u_2, ..., u_n)) = \frac{1}{N^n}$$

ISBN 9788496866133

Si lo que deseamos es calcular la probabilidad de obtener una muestra $\{u_1, ..., u_n\}$ sin tener en cuenta el orden , y donde hay Z elementos distintos, es equivalente a calcular la probabilidad de obtener una muestra de tamaño n donde aparece k_1 veces la unidad u_1, k_2 veces la unidad $u_2, ..., k_z$ veces la unidad u_z.

Habrá que contar cuántas muestras ordenadas dan lugar a las unidades $\{u_1, ..., u_n\}$. Como se admiten repeticiones, serán variaciones con repetición: $\dfrac{n!}{\prod_{i=1}^{z} k_i!}$ es el número de muestras ordenadas que dan lugar a $\{u_1, ..., u_n\}$ Obsérvese que si todas las unidades de $\{u_1, ..., u_n\}$ son distintas, $z = n$, y por lo tanto $k_i = 1$ para todo $i = 1, ..., n$, y entonces el número de muestras es el número de permutaciones $n!$.

Ahora, al ser cada muestra de las N^n posibles equiprobable, basta aplicar la regla de Laplace para obtener

$$P(\{u_1, ..., u_n\}) = \frac{\text{Casos favorables}}{\text{Casos posibles}} = \frac{n!}{N^n \prod_{i=1}^{z} k_i!}.$$

Se observa que si se tiene en cuenta el orden, todas las muestras tienen igual probabilidad (lo que no ocurre si no se tiene en cuenta el orden). Esto es de gran ayuda simplificadora en cuanto a los desarrollos teóricos, por lo cual en muestreo m.a.s.r. se suele tener en cuenta el orden en muchos de estos desarrollos.

Ejemplo 3.1.

Supongamos una urna con 4 unidades numeradas 1,2,3,4. Se toman 3 unidades según m.a.s. con reemplazamiento. Calculemos:

a) La probabilidad de obtener primero un 1, después un 2 y después otra vez un 2 .

b) La probabilidad de obtener un 1 y dos doses en nuestra muestra sin importar el orden.

c) Verificar que hay 3 muestras de las 64 posibles que contienen un uno y dos doses.

d) Si obtenemos la media muestral calculada según los números que aparecen en las bolas de la muestra, calcular la probabilidad de que esta media sea 5/3.

---

a) $P((1,2,2)) = \dfrac{1}{4^3} = \dfrac{1}{64} = 0.015625$

b) $P(\{1,2,2\}) = \dfrac{3!}{4^3 1!2!} = \dfrac{3}{64} = 0.046875$

c) $(1,2,2), (2,1,2), (2,2,1)$

d) Para ver los diferentes valores que puede tomar esa media muestral, es necesario calcularla para las $\binom{N+n-1}{n} = \binom{4+3-1}{3} = 20$ diferentes muestras donde no importa el orden (pues la media es idéntica

ISBN 9788496866133

si las muestras cambian internamente de orden). Mostramos las muestras con la media muestral calculada entre paréntesis:

$\{1,1,1\}(1); \{1,1,2\}(4/3); \{1,1,3\}(5/3); \{1,1,4\}(2);\ \{1,2,2\}(5/3);\ \{1,2,3\}\ (2);\ \{1,2,4\}\ (7/3),\ \{1,3,4\}$ (3), $\{1,3,3\}$ (7/3); $\{1,4,4\}(3);\ \{2,2,2\}(2);\ \{2,2,3\}(7/3);\ \{2,2,4\}(8/3);\ \{2,3,4\}(3); \{3,3,2\}$ (8/3); $\{3,3,3\}$ (3); $\{3,3,4\}(10/3);\ \{4,4,2\}(10/3);\ \{4,4,3\}$ (11/3); $\{4,4,4\}(4)$.

Por lo tanto las muestras que dan lugar a una media muestral de 5/3 son las que contienen dos unos y un 3 y las que contienen dos doses y un uno. Como $P(\{1,2,2\}) = \frac{3}{64}$ y $P(\{1,1,3\}) = \frac{3!}{4^3 2!1!} = \frac{3}{64}$, la probabilidad de que la media muestral sea 5/3 en muestreo con reemplazamiento en esta población es de $P(\{1,2,2\}) + P(\{1,1,3\}) = \frac{6}{64}$.

3.1.2 Probabilidades de Inclusión

Definición (probabilidad de inclusión).

La probabilidad de que aparezca una unidad determinada i , con $i = 1, ..., N$ en una muestra de tamaño n, se denomina **probabilidad de inclusión** de la unidad i y se denota por π_i. La probabilidad de inclusión de dos unidades i, j se denota por π_{ij}.

Propiedad 3.1.

Para cada $i = 1, ..., N$, la probabilidad de inclusión π_i en muestreo aleatorio simple con reemplazamiento es $\pi_i = 1 - (1 - \frac{1}{N})^n$.

Demostración.

$\pi_i = 1-$Probabilidad de que la unidad i no pertenezca a la muestra de tamaño n.

Esta última es la probabilidad de que i no salga en la primera extracción, ni en la segunda,..., ni en la enésima. Al ser extracciones independientes y en cada una de ellas la probabilidad de que i no salga es $(1 - \frac{1}{N})$ por ser muestreo con reemplazamiento, la probabilidad de que i no salga en ninguna es $(1 - \frac{1}{N})^n$ con lo que $\pi_i = 1 - (1 - \frac{1}{N})^n$.

Propiedad 3.2.

Para cada $i, j = 1, ..., N$, $i \neq j$, la probabilidad de inclusión π_{ij} en muestreo aleatorio simple con reemplazamiento es $\pi_{ij} = 1 - 2(1 - \frac{1}{N})^n + (1 - \frac{2}{N})^n$.

Demostración.

Con un razonamiento análogo a la demostración anterior, tenemos que

$\pi_{ij} = 1 - P(i$ o bien j no pertenezca a la muestra$) =$

$= 1 - [P(i \notin$ muestra$)+P(j \notin$ muestra$)$-$P(i, j \notin$ muestra$)] =$

ISBN 9788496866133

$$= 1 - [(1 - \frac{1}{N})^n + (1 - \frac{1}{N})^n - (1 - \tfrac{2}{N})^n] = 1 - 2(1 - \frac{1}{N})^n + (1 - \tfrac{2}{N})^n$$

3.1.3 Selección de una muestra aleatoria simple con reemplazamiento

Supongamos que disponemos de un listado con las unidades poblacionales numeradas del 1 al N.

Tablas de números aleatorios.

Tradicionalmente se utilizaban tablas de números aleatorios para elegir con equiprobabilidad las unidades. Aunque hay varias maneras de utilizar estas tablas, usualmente se escoge página, línea y columna de comienzo arbitrariamente. A continuación se lee el número hata tantas cifras como tiene N. Si el número es mayor que N, se descarta. Si es menor, se selecciona la unidad con ese número. El procedimiento se repite n veces para obtener una muestra de tamaño n (véase que es muestreo con reemplazamiento pues la misma unidad puede ser seleccionada varias veces). Para evitar demasiados descartes, se puede refinar el proceso utilizando múltiplos de N , es decir, tomando como puntos de corte N, $2N, 3N...$, hasta que el número mN supere el número de cifras de N. Así, si el número seleccionado de la tabla está por ejemplo entre N y $2N$, ya no se descarta, sino que se le resta N y se selecciona. En general se resta kN si el número está entre kN y $(k+1)N$.

Calculadora y ordenador.

El método anterior ya prácticamente no se usa, al disponer todos los ordenadores e incluso las calculadoras, de la generación pseudoaleatoria de una variable aleatoria uniforme $U(0,1)$. Para seleccionar un número entero i entre 1 y N con equiprobabilidad , basta generar u de una $U(0,1)$ y hacer $i = [N * u] + 1$ donde el corchete es el operador "parte entera". Para obtener una m.a.s.r. de tamaño n , se repite el procedimiento n veces. Casi todos los paquetes estadísticos, hojas de cálculo y lenguajes de programación tienen sentencias para controlar la semilla del generador pseudoaleatorio, si se desea.

3.2 Estimación en muestreo aleatorio simple con reemplazamiento

3.2.1 Estimación de la media poblacional

Las unidades poblacionales toman valores $y_1, y_2, ..., y_N$.En muestreo aleatorio simple con reemplazamiento, cada valor muestral y_i , con $i = 1, ..., n$, toma valores $y_1, y_2, ..., y_N$ con probabilidades iguales $\frac{1}{N}, ..., \frac{1}{N}$.

Por lo tanto cada posible valor muestral y_i es una variable aleatoria con Esperanza $E(y_i) = \sum_{j=1}^{N} \frac{1}{N} y_j = \overline{y}$, y Varianza $V(y_i) = \sum_{j=1}^{N} \frac{1}{N}(y_j - E(y_j))^2 = \frac{1}{N}\sum_{j=1}^{N}(y_j - \overline{y})^2 = \sigma^2$.

Teorema 3.1 (estimación de la media).

La media muestral $\widehat{\overline{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$ es un estimador insesgado de la media poblacional.

ISBN 9788496866133

Demostración.

$E(\widehat{\overline{y}}) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \frac{1}{n}\sum_{i=1}^{n} \overline{y} = \overline{y}.$

Teorema 3.2 (varianza del estimador) .

La varianza de $\widehat{\overline{y}}$ es $V(\widehat{\overline{y}}) = \frac{\sigma^2}{n}$.

Demostración.

Por independencia del proceso de muestreo (las covarianzas entre observaciones son cero), $V(\widehat{\overline{y}}) = V(\frac{1}{n}\sum_{i=1}^{n} y_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(y_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$.

En la práctica, necesitamos un estimador de $V(\overline{y})$ para poder evaluar la precisión de la estimación de la media, y construir intervalos de confianza.

Teorema 3.3 (estimación de la varianza del estimador) .

En m.a.s.r, s^2 es un estimador insesgado de σ^2 y además $\widehat{V}(\widehat{\overline{y}}) = \frac{s^2}{n}$ es un estimador insesgado de $V(\widehat{\overline{y}}) = \frac{\sigma^2}{n}$.

Demostración.

Se tendrán en cuenta los siguientes resultados:

a) $E(\widehat{\overline{y}}^2) = V(\widehat{\overline{y}}) + E(\widehat{\overline{y}})^2 = \frac{\sigma^2}{n} + \overline{y}^2$.

b) $E(y_i^2) = \sigma^2 + \overline{y}^2$.

Así, $E(s^2) = E\left(\frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 - n\widehat{\overline{y}}^2)\right) = \frac{n}{n-1}E\left(\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \widehat{\overline{y}}^2\right) =$

$$= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} E(y_i^2) - E(\widehat{\overline{y}}^2)\right) = \frac{n}{n-1}\left(\sigma^2 + \overline{y}^2 - \frac{\sigma^2}{n} - \overline{y}^2\right) = \sigma^2.$$

Por lo tanto, $E\left(\frac{s^2}{n}\right) = \frac{1}{n}E(s^2) = \frac{\sigma^2}{n}$.

Ejemplo 3.2.

Disponemos de una población de tamaño $N = 5$ y queremos estimar la media poblacional mediante m.a.s.r. y $n = 3$. Calcular un estimador insesgado de la media poblacional y dar un estimador insesgado de la varianza del estimador en los casos siguientes:

a) Suponiendo que los elementos muestrales escogidos por m.a.s.r. son $(1, 2, 2)$, con $y_1 = 4$, $y_2 = 8$.

b) Suponiendo que los elementos muestrales escogidos por m.a.s.r. son $(1, 3, 2)$, con $y_1 = 4$, $y_2 = 8$. , $y_3 = 6$.

ISBN 9788496866133

---

a) $\widehat{\overline{y}} = \frac{1}{3}(4+8+8) = \frac{20}{3}$

$$\widehat{V}(\widehat{\overline{y}}) = \frac{s^2}{n} = \frac{1}{n(n-1)}(\sum_{i=1}^{n} y_i^2 - n\widehat{\overline{y}}^2) = \frac{1}{3\cdot 2}((4^2+8^2+8^2) - 3(\tfrac{20}{3})^2) = \frac{16}{9}$$

b) $\widehat{\overline{y}} = \dfrac{1}{3}(4+6+8) = 6$

$$\widehat{V}(\widehat{\overline{y}}) = \frac{s^2}{n} = \frac{1}{n(n-1)}(\sum_{i=1}^{n} y_i^2 - n\widehat{\overline{y}}^2) = \frac{1}{3\cdot 2}((4^2+6^2+8^2) - 3\cdot 6^2) = \frac{4}{3}$$

Los resultados anteriores permiten deducir cómo estimar el Total poblacional y la proporción poblacional, en su caso.

3.2.2 Estimación del Total poblacional y proporción

Corolario 3.1 (estimación del total) .

(a) Un estimador insesgado del total poblacional.en m.a.s.r. es $N\widehat{\overline{y}}$, con varianza $V(N\widehat{\overline{y}}) = N^2\dfrac{\sigma^2}{n}$.

(b) $N^2\dfrac{s^2}{n}$ es un estimador insesgado de $V(N\widehat{\overline{y}})$.

Si y es una variable dicotómica ($y_i = 1$ si la unidad poblacional i tiene cierta cualidad, $y_i = 0$ si no la tiene), media muestral coincide con proporción muestral y media poblacional coincide con proporción poblacional. Así los siguientes resultados son inmediatos:

Corolario 3.2 (estimación de la proporción) .

(a) La proporción muestral $\widehat{p}$ es un estimador insesgado de la proporción poblacional p en m.a.s.r.

(b) $V(\widehat{p}) = \dfrac{\sigma^2}{n} = \dfrac{p(1-p)}{n} = \dfrac{pq}{n}$.

(c) $s^2 = \dfrac{n}{n-1}\widehat{p}\widehat{q}$ es un estimador insesgado de $\sigma^2 = pq$ y además $\widehat{V}(\widehat{p}) = \dfrac{\widehat{p}\widehat{q}}{n-1}$ es un estimador insesgado de $V(\widehat{p}) = \dfrac{pq}{n}$.

3.3 Tamaño de la muestra en la estimación de la media en m.a.s.r.

Se planteará ahora el problema de determinar el tamaño muestral para obtener determinada precisión o coste en muestreo aleatorio simple con reemplazamiento, en estimación de la media. Para la estimación del total o proporción, pueden aplicarse pequeñas modificaciones.

ISBN 9788496866133

Dada una población de tamaño N en la que queremos estimar $\overline{y}$ se plantea el problema de determinar el tamaño muestral n para obtener una determinada precisión Esta precisión puede ser prefijada en términos de error de muestreo, error relativo de muestreo, error de muestreo absoluto o en otros términos. Generalmente en m.a.s.r. es necesario tener una estimación previa, a partir de información sobre la población (histórica o muestras piloto), de la desviación típica poblacional σ de la característica de interés y. Llamaremos a esta estimación $\widehat{\sigma}$. Supondremos normalidad del estimador, que se cumple con n suficientemente grande, por el Teorema Central del Límite.

3.3.1 Tamaño muestral con error de muestreo prefijado

Se plantea el problema de determinar cuál es el tamaño muestral mínimo necesario para obtener un valor de error de muestreo=ϕ.en estimación de la media en m.a.s.r. Como el error de muestreo es

$$\sqrt{V(T)} = \sqrt{V(\widehat{\overline{y}})} = \frac{\sigma}{\sqrt{n}} \simeq \frac{\widehat{\sigma}}{\sqrt{n}},$$

basta hacer $\dfrac{\widehat{\sigma}}{\sqrt{n}} = \phi$ y por lo tanto $n = \dfrac{\widehat{\sigma}^2}{\phi^2}$.

3.3.2 Tamaño muestral con error de muestreo relativo prefijado

Se trata de determinar cuál es el tamaño muestral mínimo necesario para obtener un valor de error de muestreo relativo=ϕ.en estimación de la media en m.a.s.r. Se supone que se dispone de una estimación previa $\widehat{\overline{y}}$ de la media poblacional, de una muestra piloto. Como el error de muestreo relativo

$$\phi = \frac{\sqrt{V(\widehat{\overline{y}})}}{\widehat{\overline{y}}} \simeq \frac{\widehat{\sigma}/\sqrt{n}}{\widehat{\overline{y}}} \text{ y por lo tanto } n = \frac{\widehat{\sigma}^2}{\phi^2 \widehat{\overline{y}}^2}.$$

3.3.3 Tamaño muestral con error de muestreo absoluto prefijado

El problema es ahora determinar cuál es el tamaño muestral mínimo necesario para obtener un valor de error de muestreo absoluto=e.en estimación de la media en m.a.s.r. Se conoce el nivel de confianza α, y se supone normalidad, con lo que $z_{\alpha/2}$ es conocido. Como el error de muestreo absoluto

$$e = z_{\alpha/2}\sqrt{V(\widehat{\overline{y}})} \simeq z_{\alpha/2}\widehat{\sigma}/\sqrt{n} \text{ y por lo tanto } n = \frac{z_{\alpha/2}^2 \widehat{\sigma}^2}{e^2}.$$

ISBN 9788496866133

Ejemplo 3.3.

En una población de tamaño $N = 1000$ se desea estimar la media poblacional. Se dispone de información de una muestra piloto que permite estimar la varianza poblacional como $\widehat{\sigma}^2 = 7$, y la media poblacional como $\widehat{\overline{y}} = 25$. Calcular el mínimo tamaño muestral necesario para la estimación en los siguientes supuestos:

a) El error máximo de muestreo es de $\phi = 0.25$.

b) El error máximo relativo de muestreo es de $\phi = 0.05$.

c) El error máximo absoluto de muestreo es de $e = 1.5$, con un grado de confianza de $\alpha = 0.05$.

d) Con la ayuda de un programa informático, presentar los diferentes errores de muestreo obtenidos para $n = 1, 2, ..., 10$.

e) Presentar una gráfica de la evolución del error absoluto de muestreo, respecto a n.

---

a)$n = \dfrac{\widehat{\sigma}^2}{\phi^2} = \dfrac{7}{0.25^2} = 112$

b)$n = \dfrac{\widehat{\sigma}^2}{\phi^2 \widehat{\overline{y}}^2} = \dfrac{7}{0.05^2 \cdot 25^2} = 4.48 \Rightarrow n = 5$

c)$n = \dfrac{z^2_{\alpha/2}\widehat{\sigma}^2}{e^2} = \dfrac{1.96^2 \cdot 7}{1.5^2} = 11.9$

d) El recurso de utilizar un ordenador para seleccionar el tamaño muestral es a menudo más útil que recurrir a las fórmulas, pues éstas no siempre se pueden deducir tan fácilmente, y además tablas como la que se va a presentar ofrecen información de gran utilidad práctica para seleccionar el tamaño muestral, combinada posiblemente con los costes asociados al tamaño n.

La programación consiste en un bucle de n =1 hasta 10, calculando dentro del bucle los tres errores de muestreo para el n en curso. El programa sería algo así como:

```
nfin=10,sigma2=7,mediay=25,zeta=1.96

repetir desde n=1 hasta n=nfin
    e1=sqroot(sigma2/n)
    e2=e1/mediay
    e3=zeta*e1
fin_repetir
n       Error          Error              Error
        de muestreo    de muestreo        de muestreo
                       relativo           absoluto
1        2.65            0.11              5.18
2        1.87            0.07              3.67
3        1.53            0.06              2.99
4        1.32            0.05              2.59
```


ISBN 9788496866133

5	1.18	0.05	2.32
6	1.08	0.04	2.12
7	1.00	0.04	1.96
8	0.94	0.04	1.83
9	0.88	0.04	1.73
10	0.84	0.03	1.64

Tabla 3.1. Errores de muestreo para diferentes tamaños muestrales.

e) El programa anterior, llevando n hasta un número suficientemente alto, nos permite representar gráficamente el error de muestreo absoluto (semianchura del intervalo de confianza).

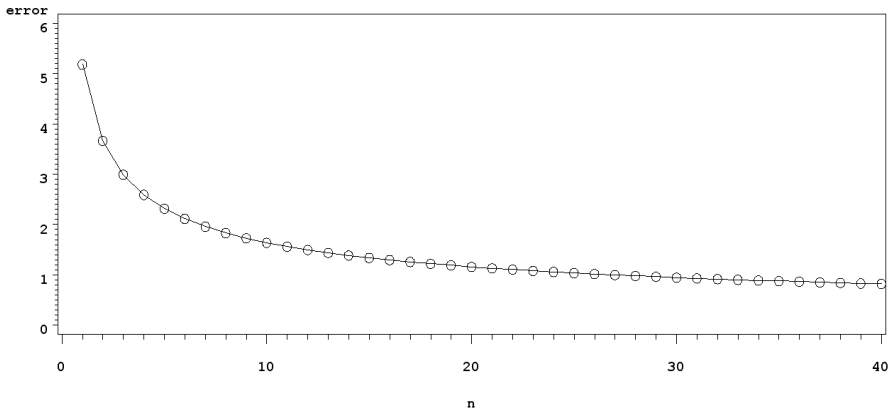


Figura 3.1. Error de muestreo absoluto en función de n.

Se observa que a partir de un cierto n el error de muestreo no desciende de manera muy significativa, con lo cual se intuye que a menudo, para estimar ciertas cantidades poblacionales, la muestra no necesita ser excesivamente grande, pues esto aumentaría mucho los costes obteniéndose una precisión muy similar.

Ejemplo 3.4.

En el Ejemplo 2.2 , supongamos que se realiza muestreo aleatorio simple con reemplazamiento y se desea calcular el error de muestreo dependiendo del número de unidades tomadas. En estos datos, $\sigma = 4.43$. Si tomamos $n = 1$ unidades, el error de muestreo será $\dfrac{\sigma}{\sqrt{n}} = \dfrac{4.43}{1} = 4.43$. Si tomamos dos unidades, el error de muestreo será $\dfrac{\sigma}{\sqrt{n}} = \dfrac{4.43}{\sqrt{2}} = 3.13$. Realizando una tabla como la del ejemplo anterior, tenemos:

n	Error de muestreo
1	4.43
2	3.13
3	2.55
4	2.21
5	1.98

ISBN 9788496866133

```
6    1.80
7    1.67
8    1.56
9    1.47
10   1.40
```

Tabla 3.2. Errores de muestreo para diferentes tamaños muestrales.

Algunas puntualizaciones son necesarias en este momento:

1) En muestreo con reemplazamiento, el tamaño de la muestra puede ser superior al de la población, pues se pueden repetir items.

2) Si se toma una muestra de tamaño $n = 1$, se puede estimar la media poblacional , pero no se puede estimar la varianza del estimador ni por lo tanto la precisión de nuestra estimación, pues s^2 no se puede calcular. Esto ocurre en general en todos los tipos de muestreo, y es particularmente importante en casos en que la población se subdivide (muestreo estratificado o por conglomerados), pues si se requieren estimaciones de errores por subdivisión será necesario un tamaño mínimo de $n = 2$ en cada subdivisión.

Ejemplo 3.5.

Daremos un ejemplo de simulación para observar la distribución de la variable aleatoria media muestral $\widehat{\overline{y}}$ en el ejemplo 2.2 de los árboles. Supongamos que se obtiene una muestra de tamaño $n = 3$ mediante muestreo con reemplazamiento y se calcula la media muestral $\widehat{\overline{y}}$. Este proceso se realiza 100 veces, y se obtiene el histograma de la Figura 3.2, que representa una aproximación a la forma de la distribución de probabilidad de $\widehat{\overline{y}}$. Según se observa, el estimador aparece como normalmente distribuido, lo que justifica el uso de los intervalos de confianza habituales para la media poblacional. Obsérvese también que la esperanza de $\widehat{\overline{y}}$ está en torno a 12.75, que es la media poblacional, pues $\widehat{\overline{y}}$ es un estimador insesgado.

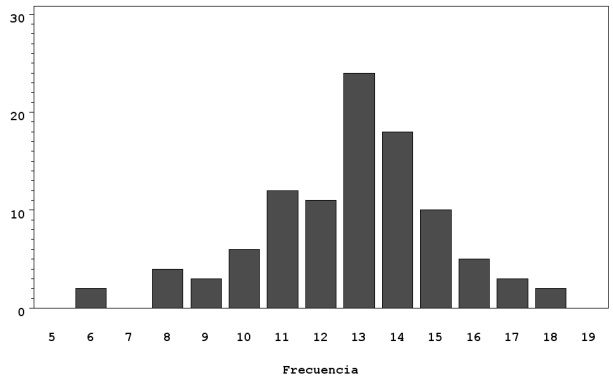


Figura 3.2. Distribución de $\widehat{\overline{y}}$ con $n = 3$ y m.a.s.r. en el ejemplo de los árboles.

ISBN 9788496866133

Ejemplo 3.6.

El proceso de determinación del tamaño muestral no debe tomarse como algo rígido, aplicando de manera automática las fórmulas descritas, sino que éstas deben ser un punto de referencia, pero tomando en cuenta también costes, problemas de trabajo de campo, medición, etc.

Veremos un ejemplo del proceso práctico empleado, teniendo en cuenta las fórmulas presentadas.

Supongamos, a modo de ejercicio teórico, que se desea estimar la longitud media de costa de las provincias costeras españolas. La población, es decir, la información de las 22 provincias costeras, se presenta en la Tabla 3.1. Asumamos que estamos interesados en que la semianchura del intervalo de confianza sea aproximadamente de 40 km. con un grado de confianza de 95%. Es decir, la muestra nos debe permitir decir que aproximadamente esa longitud media que queremos estimar está entre 40 km. arriba o 40 km abajo de nuestra estimación, con un grado de confianza del 95%.

Para ello, se debe tomar en primer lugar una muestra piloto , de bajo coste, pues hay que estimar σ para calcular el tamaño muestral a partir de la fórmula correspondiente. Se decide tomar esta muestra con tamaño de 2 provincias. Supongamos que son seleccionadas Almería y Barcelona. La media de longitud de estas dos provincias es $\frac{1}{2}(249+161) = 205$. Entonces la desviación típica estimada se puede calcular como la cuasidesviación típica muestral por ser un estimador insesgado. En este caso es $\widehat{\sigma} = s = \sqrt{\frac{1}{2-1}[(249-205)^2 + (161-205)^2]} = 62.2$. De modo que para que el error absoluto de muestreo sea $e = 40$ ha de ser $n = \dfrac{z^2_{\alpha/2}\widehat{\sigma}^2}{e^2} = \dfrac{(1.96)^2(62.2)^2}{40^2} = 9.28$.

Provincia	Km.	Provincia	Km.
Guipúzcoa	92	Almería	249
Vizcaya	154	Murcia	274
Cantabria	284	Alicante	244
Asturias	401	Valencia	135
Lugo	144	Castellón	139
Coruña (A)	956	Tarragona	278
Pontevedra	398	Barcelona	161
Huelva	122	Girona	260
Cádiz	285	Balears(Illes)	1.428
Málaga	175	Palmas(Las)	815
Granada	79	Tenerife	768

Tabla 3.3. Longitud de km. de costa en las provincias costeras españolas.

ISBN 9788496866133

Como se ha constatado en anteriores ejemplos, es más informativo obtener la tabla de errores absolutos en función de n, presentada en la Tabla 3.4.

Se decide entonces tomar una muestra de $n = 10$ provincias.para examinar sus tamaños de costa y estimar la media poblacional.

Una vez determinado el tamaño muestral a tomar se realiza el muestreo m.a.s.r. Supongamos que las 10 provincias escogidas son sucesivamente Cádiz, Granada, Cantabria, Asturias, Granada, Valencia, Tarragona, Valencia, Lugo, Almería (obsérvese que hay repeticiones pues se trata de m.a.s. con reemplazamiento).

Para esta muestra la media muestral es $\widehat{\overline{y}} = 206.8$, $s^2 = 11484$, y $s = 107$.

Al realizar un intervalo de confianza al 95% con estos datos , se observa que el error absoluto de muestreo queda estimado en $e = z_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}} = 1.96\frac{107}{\sqrt{10}} = 66.4$, muy lejos del $e = 40$ que nos sirvió inicialmente para determinar el tamaño muestral.

Evidentemente se trata de una **estimación** del error absoluto de muestreo . A pesar de ello, como el $\widehat{\sigma}$ estimado de esta muestra es más fiable que el de la muestra piloto, por estar basado en una muestra de tamaño mayor, $n = 10$ frente a una muestra de tamaño $n = 2$, hay que deducir que la semianchura del intervalo está más cerca de 73.62 que de 40, por lo que no se ha conseguido el objetivo prefijado de precisión.

Igualmente la tabla de errores de muestreo presentada anteriormente no es muy fiable, por estar basada en $\widehat{\sigma} = 62.2$. De este ejemplo se deduce que la determinación del tamaño muestral no es un problema sencillo ni que se debe tratar de manera automática. La variabilidad de la variable en cuestión, y la estimación de esta variabilidad afecta mucho a los resultados.

Como ilustración, la media poblacional de las 22 provincias es $\overline{y} = 291.56$, y la desviación típica es $\sigma = 247.70$, lejos de las estimaciones piloto ($\widehat{\sigma} = 62.2$) y muestral ($\widehat{\sigma} = s = 107$). Acudiendo a los datos, se observa que la presencia o ausencia de la provincia Islas Baleares en la muestra (con 1428 km de costa), afecta en gran medida a cualquier estimación tanto de la media como de la varianza.

n	Error de muestreo
1	122
2	86.2
3	70.4
4	61.0
5	54.5
6	49.8
7	46.1
8	43.1
9	40.6
10	38.6
11	36.8
12	35.2
13	33.8
14	32.6
15	31.5
16	30.5
17	29.6

ISBN 9788496866133

```
18   28.7
19   28.0
20   27.3
```

Tabla 3.4. Errores de muestreo para diferentes tamaños muestrales.

3.4 Tablas de fórmulas

MUESTREO ALEATORIO SIMPLE CON REEMPLAZAMIENTO			
Parámetro poblacional	$\overline{y}$	$N\overline{y}$	p
Estimador	$\widehat{\overline{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$	$N\widehat{\overline{y}}$	$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$
Varianza	$\frac{\sigma^2}{n}$	$N^2\frac{\sigma^2}{n}$	$\frac{pq}{n}$
Estimador de la varianza	$\frac{s^2}{n}$	$N^2\frac{s^2}{n}$	$\widehat{V}(\widehat{p}) = \frac{\widehat{p}\widehat{q}}{n-1}$

TAMAÑOS MUESTRALES EN m.a.s.r.		
Errores o costes prefijados	**Tamaño muestral (media)**	**Tamaño muestral (proporción)**
$\phi = \sqrt{V(\widehat{\overline{y}})}$	$\frac{\widehat{\sigma}^2}{\phi^2}$	$\frac{\widehat{p}(1-\widehat{p})}{\phi^2}$
$\phi = \frac{\sqrt{V(\widehat{\overline{y}})}}{\overline{y}}$	$\frac{\widehat{\sigma}^2}{\phi^2\widehat{\overline{y}}^2}$	$\frac{(1-\widehat{p})}{\widehat{p}\phi^2}$
$e = z_{\alpha/2}\sqrt{V(\widehat{\overline{y}})}$	$\frac{z^2_{\alpha/2}\widehat{\sigma}^2}{e^2}$	$\frac{z^2_{\alpha/2}\widehat{p}(1-\widehat{p})}{e^2}$

3.5 Obtención de muestras por m.a.s.r. con SAS

En el paquete estadístico SAS, se pueden extraer muestras por m.a.s.r. de dos modos: por programación directa y mediante la utilización del procedimiento proc surveyselect . Se verán a continuación las dos maneras:

3.5.1 Mediante programación directa

Supongamos que los datos están en el archivo SAS temporal llamado datos, que tiene N observaciones. Se supone que el archivo contiene una variable de identificación de las unidades elementales, que llamaremos en general ID.

ISBN 9788496866133

El programa de obtención de una muestra por m.a.s.r. consiste en generar un número aleatorio uniforme y entero entre 1 y N, que indicará la observación a leer. Se leerá esta observación y se guardará en el archivo que llamaremos "muestra". Este procedimiento se repetirá n veces, de manera que en el archivo llamado "muestra" habrá finalmente n observaciones obtenidas por m.a.s.r. Suponiendo $N = 50$, y $n = 10$, este programa puede ser:

```
data muestra;
do i=1 to 10;
  obse=int(ranuni(1234567)*50+1);
  set datos point=obse;
  output;
end;
stop;
run;
proc print data=muestra;run;
```

Al disponer de la variable de identificación ID, una vez que el SAS nos presenta el listado de las observaciones escogidas por m.a.s.r. mediante el procedimiento print, se podría pasar a la fase del trabajo de campo consistente en acceder a las observaciones escogidas y recolectar información en cada una de ellas para pasar al proceso de estimación.

3.5.2 Mediante el Procedimiento Surveyselect

El procedimiento surveyselect con la opción method=urs genera muestras por m.a.s. con reemplazamiento. La sintaxis básica es:

```
proc surveyselect data=poblacion out=muestra method=urs n=tama\U{f1}o
outhits seed=semilla;
run;
```

La opción outhits especifica que si una observación es elegida varias veces, aparezca repetida en el archivo muestral. Si no se indica la opción outhits, la observación aparece una vez, aunque en el archivo muestral estará también presente la variable Numberhits que indica el número de veces que ha sido seleccionada cada observación.

Seed=semilla indica la semilla de aleatorización. Si no se utiliza esta opción, la semilla utilizada es el valor en el momento del reloj del ordenador.

Asumiendo que los datos están en el archivo SAS temporal llamado datos, con $N = 50$ observaciones, y se desea una muestra de $n = 10$ observaciones obtenidas por m.a.s.r., el programa SAS que genera el archivo llamado "muestra" que contiene esas observaciones muestrales será

A continuación ya se podría utilizar el proc surveyselect para obtener muestras por m.a.s.r.:

```
proc surveyselect data=datos out=muestra method=urs n=10 outhits;
run;
```


ISBN 9788496866133

3.6 Estimación en m.a.s.r. con SAS

3.6.1 Estimación con el Procedimiento Surveymeans

Estos cálculos se pueden realizar utilizando el procedimiento surveymeans.

Estimación de la media o proporción

A continuación se utilizaría el procedimiento surveymeans con las opciones básicas. Suponiendo que la variable de interés se llama y en el archivo muestra, y nos interesa la estimación de su **media** (o **proporción** si y es una variable 0-1):

```
proc surveymeans data=muestra;
var y;
run;
```

que aporta la estimación de la media o proporción, la desviación estándar del estimador (raíz cuadrada de la varianza estimada, calculada tal y como aparece en las fórmulas de las tablas) y el intervalo de confianza al 95%.

Estimación del total

Si lo que interesa es la estimación del **total**, será necesario, en el proc surveymeans, utilizar la variable de peso con valor constante $\frac{N}{n}$ creada con el nombre SamplingWeight por el proc surveyselect. En el caso que nos ocupa, $\frac{N}{n} = \frac{50}{10} = 5$.

En el procedimiento surveymeans se debe especificar esta variable de peso (weight Samplingweight) y que lo que interesa es la estimación del total (opción sum):

```
proc surveymeans data=muestra sum;
weight samplingweight;
var y;
run;
```

obteniendo la estimación del total, desviación estándar del estimador e intervalo de confianza al 95%.

Si el archivo de muestra no proviene del proc surveyselect, se debe añadir la variable de peso al archivo. Para no añadir confusión, la llamaremos también samplingweight , dándole el valor constante $\frac{N}{n} = \frac{50}{10} = 5$:

```
data muestra;
  set muestra;
  Samplingweight=5;
run;
```


ISBN 9788496866133

A continuación ya se utiliza el proc surveymeans como se ha visto:

```
proc surveymeans data=muestra sum;
weight samplingweight;
var y;
run;
```

Otra posibilidad alternativa, más sencilla y directa de utilización, es utilizar la macro masr, creada, como todas las macros que se verán en este libro, por los autores del mismo.

3.6.2 Estimación con la macro estimasr

La macro estimasr presenta en la ventana output los estimadores y sus varianzas. Su sintaxis es la siguiente:

```
%estimasr(archivo,variable,npobla,z);
```

Donde:

archivo es el archivo de datos SAS que contiene la muestra.

variable es la variable de interés, sobre la cual se desea obtener estimaciones.

npobla es el número de observaciones poblacional N (se utiliza solamente para la estimación del total).

z es una variable indicador con valor z=1 si el archivo de muestra proviene de un surveyselect SIN la opción outhits y z=0 si no es así.

Una aplicación de esta macro con los números anteriores, suponiendo que los datos provienen de un proc surveyselectsin la opción outhits, sería:

```
%estimasr(muestra,y,50,1);
```

Recordemos que el archivo SAS con los datos muestrales se denominaba exactamente "muestra". Y si no se supone que los datos provienen de un proc surveyselect:

```
%estimasr(muestra,y,50,0);
```

En la ventana output aparecen los estimadores para la media (o proporción si e una variable 0-1) y para el total, junto con sus varianzas, desviaciones típicas (llamadas desviaciones estándar en la salida) e intervalos de confianza.

Estos intervalos calculados por el SAS pueden llevar a confusión, pues están calculados a partir de la distribución t de Student, y no de la distribución normal. Concretamente, en lugar de $z_{\alpha/2}$ se utiliza $t_{n-1,\alpha/2}$, donde este valor representa el valor que deja a su derecha $\alpha/2$ en una t

ISBN 9788496866133

de Student con $n-1$ grados de libertad. Para $n > 30$ la proximidad de la distribución Normal a la t de Student es ya muy alta.

En todo caso, si se desean calcular los percentiles de las distribuciones normal o t de Student con SAS para construir intervalos de confianza, se pueden utilizar las funciones probit y tinv. En el siguiente ejemplo se calcula el percentil $z_{0.025}$ de una normal (es decir, $\alpha = 0.05$). Como la función probit(p) calcula el punto que deja a su izquierda el valor p, hay que poner probit(1-0.025) =probit(0.975). Igualmente se aplica la función tinv(p, n) para calcular el punto $t_{n-1,\alpha/2}$. Si se desea calcular $t_{9,0.025}$, hay que poner tinv$(0.975, 9)$:

```
data;
  xnormal=probit(0.975);
  xtstudent=tinv($0.975,9);$
 put xnormal= xtstudent=;
run;
```

En la ventana LOG sale:

```
xnormal=1.9599639845 xtstudent=2.2621571628
```

con lo que se aprecia la diferencia, que lleva en realidad a intervalos más anchos (estimación más conservadora) al utilizar la t de Student. Recordemos que esta distribución fue creada para tratar problemas con muestras pequeñas, para las cuales existen estas diferencias.

En el resto de este libro se utilizará por defecto la aproximación normal con ánimo de simplificar, aún a sabiendas de que en muchos casos puedan existir aproximaciones un poco más precisas.

ISBN 9788496866133

3.7 Ejercicios resueltos

Ejercicio 2.1.

Se realiza un proceso de muestreo m.a.s.r. con tamaño $n = 8$ en una población de 100 vacas para estimar la cantidad promedio de leche que producen al mes y si han tenido cierta enfermedad. Las vacas fueron numeradas de 1 a 100 antes de tomar la muestra. Los datos recogidos son los siguientes:

Vaca	Leche	Ha tenido enfermedad	Vaca	Leche	Ha tenido enfermedad
5	40	SI	52	38	SI
23	35	NO	64	51	NO
36	50	NO	71	23	NO
36	50	NO	92	28	SI

Se pide:

a) Estimar la media de litros producidos al mes por las vacas y dar un intervalo de confianza al 95%. Estimar también el total de litros producidos por las 100 vacas y dar un intervalo de confianza al 95%.

b) Estimar la proporción de vacas que ha tenido la enfermedad y dar un intervalo de confianza al 95%.

c) Asumiendo los datos muestrales de la tabla como la única información de que disponemos, decir qué tamaño muestral debería tomarse, cuando se quiere estimar la media poblacional de l producción de leche por vaca, en cada uno de los siguientes casos:

c1) Para obtener un error de muestreo de 2.36

c2) Para obtener un error de muestreo relativo de 0.05

c3) Para obtener un error de muestreo absoluto de 3.25, asumiendo $\alpha = 0.05$.

c4) Presentar, mediante un programa en SAS, la tabla de errores de muestreo, relativo y absoluto con $\alpha = 0.05$, a medida que aumenta n, en el caso de estimación de la proporción.

---

a) Hay que recalcar que por tratarse de muestreo con reemplazamiento, pueden repetirse las observaciones. En caso de repetición hay que considerar la repetición de la observación en el cálculo de los estimadores (si no, no serían insesgados).

La estimación de la media de litros de leche producidos al mes es

$$\widehat{\overline{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i = 39.37$$

ISBN 9788496866133

La varianza estimada del estimador es

$$\widehat{V}(\widehat{\overline{y}}) = \frac{s^2}{n} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (y_i - \widehat{\overline{y}})^2 = 13.92$$

Y por lo tanto el intervalo de confianza al 95% es

$$(\widehat{\overline{y}} - 1.96\sqrt{\widehat{V}(\widehat{\overline{y}})}, \widehat{\overline{y}} + 1.96\sqrt{\widehat{V}(\widehat{\overline{y}})}) = (32.05, 46.68).$$

Para el total, basta multiplicar la estimación de la media por N y la de la varianza por N^2 :

El estimador del total es $N\widehat{\overline{y}} = 3937.5$ y su varianza estimada es $N^2 \frac{s^2}{n} = 139263$.

El intervalo de confianza es

$$(3937.5 - 1.96\sqrt{139263}, 3937.5 - 1.96\sqrt{139263}) = (3205, 4668).$$

En realidad, el intervalo de confianza se puede comprobar que se puede calcular así:

$$(N\widehat{\overline{y}} - 1.96\sqrt{N^2\widehat{V}(\widehat{\overline{y}})}, \widehat{\overline{y}} + 1.96\sqrt{N^2\widehat{V}(\widehat{\overline{y}})}) =$$

$$=(N[\widehat{\overline{y}} - 1.96\sqrt{\widehat{V}(\widehat{\overline{y}})}], N[\widehat{\overline{y}} + 1.96\sqrt{\widehat{V}(\widehat{\overline{y}})}]).$$

b) Para el caso de la proporción, se denota por $y_i = 1$ si la vaca ha tenido la enfermedad y $y_i = 0$ si no. La estimación de la proporción, cálculo de varianzas e intervalo de confianza se realizan así:

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{3}{8} = 0.375$$

y la varianza

$$\widehat{V}(\widehat{p}) = \frac{\widehat{pq}}{n-1} = 0.033.$$

El intervalo de confianza, suponiendo la aproximación Normal , es:

$$(\widehat{p} - 1.96\sqrt{\widehat{V}(\widehat{p})}, \widehat{\overline{y}} + 1.96\sqrt{\widehat{V}(\widehat{p})}) = (0.0189, 0.73).$$

c)

c1) El error de muestreo se cifra en la desviación típica del estimador, es decir: $\phi = \sqrt{V(\widehat{\overline{y}})} = 2.36$. Esto es equivalente a

$V(\widehat{\overline{y}}) = 2.36^2$ y por lo tanto, a $\frac{\sigma^2}{n} = 2.36^2$ y $n = \frac{\sigma^2}{2.36^2}$. Como no tenemos más información que la muestra obtenida, aproximaremos σ^2 por s^2, que se calcula en el ejercicio por ser

$$\frac{s^2}{8} = 13.92 \Longrightarrow s^2 = 111.36$$

y queda $\widehat{n} = \frac{s^2}{2.36^2} \simeq 20$.

También se podía haber aplicado directamente la fórmula de la tabla $n = \frac{\widehat{\sigma}^2}{\phi^2}$, con $\widehat{\sigma}^2 = s^2$.

c2) En este caso, el error de muestreo relativo es

ISBN 9788496866133

$\phi = \dfrac{\sqrt{V(\widehat{\overline{y}})}}{\overline{y}} = 0.05$ y como se ve en la tabla de fórmulas, queda $n = \dfrac{\widehat{\sigma}^2}{\phi^2 \widehat{\overline{y}}^2} = \dfrac{111.36}{0.05^2 39.37^2} = 28.7$ con lo cual habría que tomar al menos $n = 29$ para intentar garantizar ese nivel de error.

c3) El error de muestreo absoluto es $e = z_{\alpha/2}\sqrt{V(\widehat{\overline{y}})} = 3.25$ y acudiendo a la fórmula, $n = \dfrac{z^2_{\alpha/2}\widehat{\sigma}^2}{e^2} = \dfrac{1.96^2 111.36}{3.25^2} = 40.5$. Con lo cual ciframos $n = 41$ para intentar garantizar como máximo ese nivel de error.

c4) El programa podría ser así:

```
data;
p=0.375;
put 'n' @10 'Error' @20 'E.relativo' @31 'E.absoluto' /;
do n=2 to 100;
  var=p*(1-p)/(n-1);
  e1=var**0.5;
  e2=var/p;
  e3=1.96*(var**0.5);
  put n @10 e1 4.2 @20 e2 4.2 @31 e3 4.2;
end;
run;
```

La salida, obviando los datos centrales, es

```
n         Error     E.relativo E.absoluto

2         0.48      0.63       0.95
3         0.34      0.31       0.67
4         0.28      0.21       0.55
5         0.24      0.16       0.47
...........
96        0.05      0.01       0.10
97        0.05      0.01       0.10
98        0.05      0.01       0.10
99        0.05      0.01       0.10
100       0.05      0.01       0.10
```

Hay que remarcar que hay nivel de error distinto de cero aunque se tomen muestras de tamaño 100, pues es muestreo con reemplazamiento. En muestreo sin reemplazamiento, obviamente, al tomar muestras de tamaño $n = N = 100$ no habría error de muestreo pues ya se tendría toda la población.

Ejercicio 2.2.

En un estudio de pesca realizado en un lago, se pescan peces uno a uno, se mide su longitud en centímetros y especie y se devuelven al lago (vivos). El proceso se repite 10 veces, esperando una hora entre cada vez. Se obtienen los siguientes datos:

ISBN 9788496866133

Especie	Longitud		Especie	Longitud
CARPA	21		CARPA	20
TRUCHA	23		CARPA	23
CARPA	10		CARPA	23
CARPA	28		TRUCHA	15
TRUCHA	29		TRUCHA	25

a) Dar un intervalo de confianza aproximado al 95% para la proporción de truchas en el lago.

b) Dar un intervalo de confianza aproximado al 95% para la media de la longitud de las truchas en el lago.

c) Dar un intervalo de confianza aproximado al 95% para la media de la longitud de las carpas en el lago.

d) Dar un intervalo de confianza aproximado al 95% para la media de la longitud de los peces del lago.

e) Realizar los apartados anteriores con la macro estimasr en el SAS (salvo los I. de confianza).

f) ¿Es posible estimar el número de peces en el lago?

g) ¿Cuántos peces habría que muestrear aproximadamente para tener 10 truchas con una probabilidad al menos de 0.90?

h) Por razones ecológicas, se desea estimar con un grado de fiabilidad muy alto la proporción de truchas en el lago. Concretamente, se desea que la verdadera proporción esté , con un grado de confianza del 99%, en un intervalo de anchura como mucho 0.2. ¿Cuántos peces habrá que pescar?

---

a) Al devolverse los peces al lago, se está realizando muestreo con reemplazamiento. Se asume también que todos los peces tienen la misma probabilidad de ser pescados.

Denotando por $y_i = 1$ si el pez es una trucha, e $y_i = 0$ si no, la proporción estimada es:

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{4}{10} = 0.40.$$

y la varianza del estimador es:

$$\widehat{V}(\widehat{p}) = \frac{\widehat{pq}}{n-1} = \frac{0.4 \cdot 0.6}{9} = 0.02666.$$

El intervalo de confianza, suponiendo la aproximación Normal , es:

$$(\widehat{p} - 1.96\sqrt{\widehat{V}(\widehat{p})}, \widehat{\overline{y}} + 1.96\sqrt{\widehat{V}(\widehat{p})}) = (0.08, 0.72).$$

ISBN 9788496866133

b) Concretamente, el muestreo de truchas se ha realizado con reemplazamiento y con igual probabilidad (es decir, todas las truchas tienen la misma probabilidad de ser escogidas). Por tanto se pueden aplicar las fórmulas usuales a la submuestra de truchas.

Llamando y a la longitud, se obtiene:

$$\widehat{\overline{y}}_{truchas} = \frac{1}{n_{truchas}} \sum_{i=1}^{n_{truchas}} y_i = \frac{1}{4}(23 + 29 + 15 + 25) = 23$$

La varianza estimada del estimador es

$$\widehat{V}(\widehat{\overline{y}}_{truchas}) = \frac{s^2_{truchas}}{n_{truchas}} = \frac{1}{n_{truchas}(n_{truchas} - 1)} \sum_{i=1}^{n_{truchas}} (y_i - \widehat{\overline{y}}_{truchas})^2 = 8.666$$

Y por lo tanto el intervalo de confianza al 95% es $(17.23, 28.77)$.

c) Realizando las mismas operaciones, con $n_{Carpas} = 6$, queda

$\widehat{\overline{y}}_{Carpas} = 20.8333$

$\widehat{V}(\widehat{\overline{y}}_{truchas}) = 5.9611$

y el intervalo de confianza es: $(16.04, 25.62)$.

d) Para todos los peces, se utilizan todas las observaciones y $n = 10$. Tenemos:

$\widehat{\overline{y}} = 21.70$

$\widehat{V}(\widehat{\overline{y}}) = 3.267$

y el intervalo de confianza es: $(18.15, 25.24)$.

e) Para utilizar la macro se pueden crear tres archivos de datos en un solo paso data: el que contiene todos los peces, el de truchas y el de carpas:

```
data todos truchas carpas;
input especie $ long;
if especie='CARPA' then do;y=0;output carpas;end;
else do;y=1;output truchas;end;
output todos;
cards;
CARPA 21
TRUCHA23
CARPA 10
CARPA 28
TRUCHA 29
CARPA 20
CARPA 23
CARPA 23
TRUCHA 15
TRUCHA 25
;
```

La variable y es la que servirá para la proporción, y la variable long para la longitud. Para cada estimación se procede con un archivo. Al no constar el número de peces en el lago, se deja a missing su casilla en la macro:

Estimación de la proporción de truchas:

ISBN 9788496866133

```
%estimasr(todos,y,.,2);
```

Estimación de la media de la longitud de truchas:

```
%estimasr(truchas,long,.,2);
```

Estimación de de la media de la longitud de carpas:

```
%estimasr(carpas,long,.,2);
```

Estimación de de la media de la longitud de peces:

```
%estimasr(todos,long,.,2);
```

f) Existe un método de muestreo que se denomina muestreo de captura-recaptura. Habría que marcar (con una argollita o metal) los n_1 peces pescados en esta tanda, volver a pescar otro número n_2 de peces en una segunda tanda, y contar el número k de peces de la segunda tanda que estaban marcados. Entonces se estimaría el número N de peces en el lago por $\widehat{N} = \dfrac{n_1 n_2}{k}$.

g) Es un ejemplo de aplicación probabilística de los resultados obtenidos con el muestreo. Se ha estimado por $\widehat{p} = 0.40$ la proporción de truchas en el lago. Muestrear peces hasta que aparezcan 10 truchas cuando la probabilidad de que aparezca una trucha es $\widehat{p}$, es un experimento clásico modelizado por la distribución Binomial Negativa: una variable con distribución Binomial Negativa indica el número de fracasos antes del r-ésimo éxito. Denotando por número de fracasos el número de peces pescados=n y el r-ésimo éxito la décima vez que aparece una trucha, podemos obtener la probabilidad de cada n.

Se puede obtener la tabla de estas probabilidades con la función probnegb(p,n,m) del SAS, que da la probabilidad de tener menos de m fallos antes del n-ésimo éxito. Realizando un bucle variando m:

```
data;
do m=10 to 30;
 p=probnegb(0.4,10,m);
 put m=  p=;
end;
run;
```

Se obtiene que la probabilidad es mayor que 0.90 a partir de $m = 23$ peces pescados.

h) Al fijar la anchura del intervalo a 0.2, se está diciendo que la semianchura es $z_{\alpha/2}\sqrt{V(\widehat{p})} = \dfrac{0.2}{2} = 0.1$. Como en este caso $1 - \alpha = 0.99$ y $\alpha = 0.01$, $z_{\alpha/2} = 2.57$, se puede aplicar la fórmula deducida: $n = \dfrac{z^2_{\alpha/2}\widehat{p}(1-\widehat{p})}{e^2}$ con $e = 0.1$.

Como tenemos una estimación de $\widehat{p} = 0.4$ a través de la muestra, se obtiene que

$$n = \frac{2.57^2 0.24}{0.1^2} = 158 \text{ peces.}$$

ISBN 9788496866133

Ejercicio 2.3.

Se dispone de una población con 4 unidades u_1, u_2, u_3, u_4 . Se desea utilizar m.a.s.r. con $n = 3$. Responder a las siguientes preguntas:

a) Si se tiene en cuenta el orden, ¿cuántas muestras posibles hay?

b) ¿Cuál es la probabilidad de obtener la muestra (u_1, u_1, u_2) si se tiene en cuenta el orden?.

c) ¿Cuál es la probabilidad de que la muestra contenga el items u_1?¿Cuál es la probabilidad de que la muestra contenga a los items u_2 y u_3? Si el muestreo se realiza con $n = 4$, ¿cuál es la probabilidad de que las cuatro unidades estén en la muestra?

---

a) Recurriendo a las expresiones de combinatoria vistas en el tema anterior, hay $N^n = 4^3 = 64$ muestras con reemplazamiento, teniendo en cuenta el orden.

b)Si se tiene en cuenta el orden, la probabilidad de cada muestra es

$$P((u_1, u_1, u_2)) = \frac{1}{N^n} = \frac{1}{4^3} = 0.015625.$$

c) La probabilidad de que la muestra contenga el item u_1 es la probabilidad de inclusión de primer orden, igual para u_1 que para los demás, es decir, $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 1 - (1 - \frac{1}{N})^n = 0.578$.

La probabilidad de inclusión de u_2 y u_3 es igual para todo i, j, y es

$$\pi_{23} = \pi_{ij} = 1 - 2(1 - \frac{1}{N})^n + (1 - \frac{2}{N})^n = 0.28125.$$

Si el muestreo es con $n = 4$, se puede recurrir al cálculo de probabilidades para calcular la probabilidad de que todos caigan en la muestra. Si se tiene en cuenta el orden, con $n = 4$, hay exactamente $n!$ muestras con los items u_1, u_2, u_3, u_4 (es decir, todas las ordenaciones posibles de esos items). Como teniendo en cuenta el orden, con $N = 4$ y $n = 4$ hay $N^n = 4^4 = 256$ muestras, y teniendo en cuenta el orden todas tienen igual probabilidad $\frac{1}{N^n}$, la probabilidad de que en la muestra estén los items u_1, u_2, u_3, u_4 es $\frac{n!}{N^n} = \frac{4!}{256} = 0.09375$.

Ejercicio 2.4.

En una barraca de feria donde se dispara con escopeta de balines, aparecen patitos que salen al azar con igual probabilidad y pudiéndose repetir. Cada vez que sale uno el jugador, que es muy buen tirador, acierta. El número de patitos diferentes son tres. El primero vale 5 puntos, el segundo 10 y el tercero 15. El jugador sólo dispone de dos disparos.

Estudiar la distribución del estimador insesgado de los puntos promedio que hará el jugador. Realizar los cálculos en primer lugar teniendo en cuenta el orden, y seguidamente sin tenerlo en cuenta. Verificar que el estimador es insesgado para la media de los puntos.

Se presentan a continuación los valores de $\widehat{\overline{y}}$ para cada muestra y teniendo en cuenta el orden:

ISBN 9788496866133

Muestra	$\widehat{\overline{y}}$	$p_{muestra}$
$(1,1)$	5	1/9
$(1,2)$	7.5	1/9
$(1,3)$	10	1/9
$(2,1)$	7.5	1/9
$(2,2)$	10	1/9
$(2,3)$	12.5	1/9
$(3,1)$	10	1/9
$(3,2)$	12.5	1/9
$(3,3)$	15	1/9

La distribución de $\widehat{\overline{y}}$ es:

$\widehat{\overline{y}}$	$p(\widehat{\overline{y}})$
5	1/9
7.5	2/9
10	3/9
12.5	2/9
15	1/9

Si no se tiene en cuenta el orden, las probabilidades de cada muestra se calculan según la expresión $\dfrac{n!}{N^n \prod_{i=1}^{\tilde{z}} k_i!}$.

Por ejemplo, para la muestra $\{1,1\}$, se tiene $P(\{1,1\}) = \dfrac{n!}{N^n \prod_{i=1}^{\tilde{z}} k_i!} = \dfrac{2!}{3^2 2!} = \dfrac{1}{9}$ mientras que para la muestra $\{1,3\}$ es

$$P(\{1,3\}) = \frac{n!}{N^n \prod_{i=1}^{\tilde{z}} k_i!} = \frac{2!}{3^2 1!1!} = \frac{2}{9}.$$

ISBN 9788496866133

Muestra	$\widehat{\overline{y}}$	$p_{muestra}$
$\{1,1\}$	5	1/9
$\{1,2\}$	7.5	2/9
$\{1,3\}$	10	2/9
$\{2,2\}$	10	1/9
$\{2,3\}$	12.5	2/9
$\{3,3\}$	15	1/9

La distribución de $\widehat{\overline{y}}$ es:

$\widehat{\overline{y}}$	$p(\widehat{\overline{y}})$
5	1/9
7.5	2/9
10	3/9
12.5	2/9
15	1/9

que por supuesto es igual a la obtenida teniendo en cuenta el orden.

Para verificar que el estimador es insesgado para la media poblacional de los puntos, que es $\overline{y} = \frac{1}{3}(5+10+15) = 10$, hay que calcular la esperanza de la distribución del estimador. Esta es:

$$E(\widehat{\overline{y}}) = \frac{1}{9}(5 + 2 \cdot 7.5 + 3 \cdot 10 + 2 \cdot 12.5 + 15) = 10 = \overline{y}.$$

Ejercicio 2.5.

En una granja que tiene 1000 pollos se ha pesado una muestra obtenida por muestreo aleatorio simple con reemplazamiento 15 pollos. De estos, 8 tenían cierta enfermedad. Llamando y_i al peso del pollo i, se han obtenido los siguientes datos para el total de pollos muestreados:

$$\sum_{i=1}^{15} y_i = 34 \text{ y } \sum_{i=1}^{15} y_i^2 = 140.$$

Para los pollos enfermos, se tiene

$$\sum_{i=1}^{8} y_i = 15 \text{ y } \sum_{i=1}^{8} y_i^2 = 29.$$

Suponiendo correcta la aproximación normal,

a) Presentar un intervalo de confianza al 95% para la media del peso de los pollos en general.

ISBN 9788496866133

b) Presentar un intervalo de confianza al 95% para la proporción de pollos enfermos.

c) Presentar un intervalo de confianza al 95% para la media del peso de los pollos enfermos .

d) Presentar un intervalo de confianza al 95% para la media del peso de los pollos enfermos .

---

a) El estimador de la media es $\widehat{\overline{y}} = \frac{1}{15}\sum_{i=1}^{15} y_i = 2.267$. Para el estimador de la varianza del estimador, calcularemos en primer lugar s^2 :

$s^2 = \frac{1}{15-1}(\sum_{i=1}^{15} y_i^2 - 15\widehat{\overline{y}}^2) = 4.49.$

Luego $\widehat{V}(\widehat{\overline{y}}) = \frac{s^2}{n} = 0.30.$

El intervalo de confianza será $(2.267 - 1.96\sqrt{0.30}, 2.267 + 1.96\sqrt{0.30}) = (1.19, 3.34)$.

b) La proporción muestral es $\widehat{p} = \frac{8}{15} = 0.533$. Por lo tanto el estimador de la varianza del estimador es $\widehat{V}(\widehat{p}) = \frac{\widehat{pq}}{n-1} = 0.0177.$

El intervalo de confianza es $(0.533 - 1.96\sqrt{0.0177}, 0.533 + 1.96\sqrt{0.0177}) = (0.27, 0.79)$.

c) Análogamente al apartado a), se obtiene una estimación de $\widehat{\overline{y}} = \frac{1}{8}\sum_{i=1}^{8} y_i = 1.875$ y

$s^2 = \frac{1}{8-1}(\sum_{i=1}^{8} y_i^2 - 8\widehat{\overline{y}}^2) = 0.125$.Con lo que $\widehat{V}(\widehat{\overline{y}}) = \frac{s^2}{n} = 0.156$ y el intervalo de confianza es $(1.63, 2.12)$.

d) Para los pollos sanos, como $\sum_{i=1}^{15} y_i = 34$ y $\sum_{i=1}^{8} y_i = 15$, se tiene $\sum_{i=1}^{7} y_i = 34 - 15 = 19$ y además, $\sum_{i=1}^{7} y_i^2 = 140 - 29 = 121$.

Entonces, $\widehat{\overline{y}} = 2.71$, $s^2 = 11.6$, y $\widehat{V}(\widehat{\overline{y}}) = 1.65$, con lo que el intervalo de confianza sería $(0.18, 5.23)$.

Ejercicio 2.6.

El archivo SAS 'madrid' contiene los edificios a construir por municipio en la Comunidad de Madrid con licencia otorgada en 1999, excluida la capital Madrid (son 179 municipios en total).

a) Utilizar el proc surveyselect para obtener una muestra por m.a.s.r. de 40 municipios. Utilizar la opción seed=123456 en el proc surveyselect para obtener la misma muestra aleatoria que se analizará aquí. Comprobar si hay alguno repetido. Guardar el archivo con el nombre muestramadrid.

b) Realizar el mismo apartado a) sin incorporar la semilla, y observar varias veces las muestras obtenidas.

c) Utilizar la macro estimasr para estimar, utilizando el archivo muestramadrid, la media poblacional de edificios construidos por municipio. Comprobar con el proc means sobre el archivo poblacional cuál es la media poblacional real. Hacer lo mismo para estimar el total de edificios construidos en la Comunidad de Madrid (excluida la capital). Comprobar el valor poblacional.

ISBN 9788496866133

d) Utilizando la opción copiar y pegar, realizar el proceso de extraer una muestra y estimar con la macro estimasr la media. Utilizar sucesivamente las semillas 12341, 12342, 12343, 12344, 12345. Apuntar la media muestral en cada caso y observar la variabilidad del estimador.

e) Realizar el mismo proceso cambiando el tamaño muestral a $n = 110$ (se puede utilizar el mismo programa del apartado d), usando buscar y reemplazar $n = 40$ por $n = 130$). Comentar.

f) Supongamos que se desea estimar la media con un error de muestreo absoluto de 30, considerando $\alpha = 0.05$. Utilizando la información de muestramadrid como si ésta fuera una muestra piloto, calcular el valor n^* que daría lugar a ese error de muestreo como.máximo.

g) Utilizando la información poblacional, calcular el valor exacto de la varianza del estimador y el valor n^{**} real que haría falta en el caso del apartado f).

h) Extraer una muestra de tamaño n^* con la semilla 1234 y construir un intervalo de confianza para la media.

El programa para extraer la muestra es:

```
proc surveyselect data=madrid method=urs seed=123456 out=muestramadrid n=40 outhits;
run;
```

ejecutando un proc print:

```
options nocenter;
proc print data=muestramadrid noobs;
var muni edif;
run;
```

se obtiene:

muni	edif
Alcorcon	380
Aldea del Fresno	9
Aldea del Fresno	9
Ambite	8
Belmonte de Tajo	3
Belmonte de Tajo	3
Camarma de Esteruelas	0
Campo Real	43
Cervera de Buitrago	0
Collado Mediano	13
Colmenarejo	125
Fresno de Torote	0
Galapagar	184
Galapagar	184
Grinon	133
Guadalix de la Sierra	0
Guadarrama	136
Manzanares el Real	67
Montejo de la Sierra	0

ISBN 9788496866133

```
Navalafuente                    0
Navarredonda y San Mames        0
Navarredonda y San Mames        0
Navas del Rey                  16
Paracuellos de Jarama          48
Parla                         279
Pinto                         102
Rascafria                       0
San Lorenzo de El Escorial     81
San Lorenzo de El Escorial     81
San Lorenzo de El Escorial     81
San Sebastian de los Reyes    438
Santorcaz                       9
Torrejon de Ardoz              17
Torrelaguna                     0
Valdemanco                      0
Velilla de San Antonio          0
Villaconejos                   12
Villamanrique de Tajo           0
Villamantilla                   8
Villanueva de Perales           0
```

con lo que se observa las repeticiones de ciertos municipios que han caído en la muestra (el primero que aparece repetido es Aldea del Fresno).

b) Se realiza el mismo programa, poniendo otro nombre al archivo de salida y quitando la opción seed.

c) La macro estimasr se compila, y después se ejecuta:

```
%estimasr(muestramadrid,edif,178,2);
```

obteniendo:

```
************************************
ESTIMACION DE LA MEDIA O PROPORCION
************************************
                     Estadisticos

                           Error std de       Var de la
Variable           Media        la media           media
-----------------------------------------------------------
edif           61.725000       16.431364      269.989728
-----------------------------------------------------------
```

y para el total:

ISBN 9788496866133

```
**********************
ESTIMACION DEL TOTAL
**********************

                      Estadisticos

                                Desviacion          Var de
Variable            Suma          estandar         la suma
-----------------------------------------------------------
edif                10987     2924.782817         8554355
-----------------------------------------------------------
```

Los valores poblacionales se obtienen con el proc means:

```
proc means data=madrid mean sum std;
var edif;
run;
```

y son: la media poblacional $\overline{y}$ =71.16 y el total poblacional, $N\overline{y}$ =12667.

d) Realizando el análisis , se obtienen las siguientes medias: 123.35, 74.07, 57.37, 128.10, 67.57 . Con lo cual la varianza del estimador es muy alta. Hay que recordar que se trata de una variable con alta variabilidad (valga la redundancia) pues la cuasi desviación típica poblacional es 145.48, con una media de 71.16.

e) Cambiando a $n = 130$, se aprecia un cambio en la variabilidad del estimador: 80.67, 58.44, 53.60, 85.52, 77.83. hay que recordar que la varianza del estimador es $\dfrac{S^2}{n}$, que disminuye al aumentar n. Pero no se puede mitigar demasiado el hecho de que S^2 sea muy alta, como en este caso.

f) Tratando muestramadrid como una muestra piloto, se obtiene la varianza $\widehat{\sigma}^2 \simeq s^2$=10799.59 mediante un proc means. Así, sustituyendo en

$$n^* = \frac{z^2_{\alpha/2}\widehat{\sigma}^2}{e^2} = \frac{1.96^2 10799.59}{30^2} = 46.$$

g) Con la información poblacional, se sabe que $\sigma^2 = \dfrac{N-1}{N}S^2 = \dfrac{178-1}{178}\ 21166.10 = 21047$ (el SAS aporta la cuasivarianza poblacional $S^2 = 21166.1$, no la varianza).

Por lo tanto, se tiene que el verdadero valor de n para obtener un error de muestreo absoluto de $e = 30$ es

$$n^{**} = \frac{z^2_{\alpha/2}\sigma^2}{e^2} = \frac{1.96^2 21047}{30^2} = 89.83 \simeq 90.$$

Se observa entonces que la determinación del tamaño muestral a través de muestras piloto puede llevar a infravalorar la variabilidad real de la población y la necesidad de tomar muestras más grandes.

h) Se obtiene una media de 82.54 con un error estándar de 17.78 y por lo tanto la semianchura del intervalo de confianza estimado es $1.96 \cdot 17.78 = 34.84$, cercano a 30 como debería ser. El intervalo obtenido es $(47.69, 117.38)$.

ISBN 9788496866133

Ejercicio 2.7.

Tratar el siguiente fragmento de "La Vida es Sueño", de Calderón de la Barca, suponiendo que cada palabra es una observación, y realizando m.a.s.r. de tamaño $n = 2$, con el objetivo de estimar el total de artículos en el texto.

"Que es la vida? un frenesí.

Que es la vida? una ilusión,

una sombra, una ficción,

y el mayor bien es pequeño;

que toda la vida es sueño,

y los sueños, sueños son."

a) ¿Si el estimador del total de artículos fuera distribuido según una Normal, cuál sería su distribución exacta?

b) Suponer que la muestra obtenida fuera $\{Que, es\}$. ¿Cómo sería el intervalo de confianza? ¿Que problema hay?

c) Obtener 10 muestras por m.a.s.r. de tamaño $n = 2$ con el SAS. Observar y apuntar los valores que toma el estimador.

Para tratar el texto, probar con las dos técnicas siguientes:

c1) Eliminar todos los signos de puntuación y leer el texto con la opción input @@ , que lee de corrido en las líneas. Utilizar a continuación el proc surveyselect y simplemente proc print para contar el número de artículos y por lo tanto la proporción muestral, 10 veces.

c2) Crear un archivo en el paso data de la siguiente manera: poner un 1 si la palabra es un artículo y 0 si no. A continuación utilizar el proc surveyselect y la macro estimasr 10 veces.

¿Cuál de las dos técnicas es más operativa?

d) Realizar el apartado c) con c2) pero con $n = 15$. Comparar los valores del estimador obtenidos.

e) Calcular la probabilidad de que ordenando esas palabras al azar se formen los mismos versos de Calderón.

---

a) Hay $N = 33$ palabras en total, de las cuales 9 son artículos. El estimador del total, al ser insesgado, tiene esperanza igual a 9. Su varianza exacta es $N^2\dfrac{\sigma^2}{n} = N^2\dfrac{pq}{n}$, donde p es la proporción de artículos $p = \dfrac{9}{33} = 0.27$. Por lo tanto, la varianza del estimador del total $N\widehat{p}$, con $n = 2$, es $N^2\dfrac{pq}{n} = 33^2\dfrac{9}{33}\dfrac{24}{33}\dfrac{1}{2} = 108$. Como la esperanza es 9, si el estimador fuera normalmente distribuido, tendría una distribución Normal $N(\mu, \sigma^2) = N(9, 108)$.

b) En el caso de la muestra $\{Que, es\}$, la proporción muestral es $\widehat{p} = 0$. La estimación del total sería $N\widehat{p} = 0$, y la varianza estimada del estimador sería $\widehat{V}(\widehat{p}) = \dfrac{\widehat{pq}}{n-1} = 0$. El intervalo de confianza sería $(0, 0)$.

ISBN 9788496866133

Este es un problema que se da a menudo en estudios de proporciones cuando las muestras son pequeñas y además la proporción de la cualidad a estudiar no es grande . Se puede dar por lo tanto también en otros tipos cualesquiera de muestreo.

Aunque se puede aplicar una corrección en estos casos, es más adecuado recurrir al llamado "muestreo inverso": extraer unidades sin reemplazamiento de la población hasta encontrar al menos m observaciones con la cualidad de interés ($m > 1$). Al menos así se garantiza una estimación diferente de 0. El estimador de la proporción en este caso, basado en la distribución binomial negativa (como en el ejercicio 2.2), es $\widehat{p} = \dfrac{m-1}{n-1}$ donde n es el número de unidades extraídas. Obviamente en este método n no es conocido de antemano, depende del azar y por lo tanto es una variable aleatoria.

c)

c1) El programa para leer el texto, extraer la muestra y presentarla sería así:

```
data calderon;
input palabra $ @@;
cards;
Que es la vida un frenesí
Que es la vida una ilusión
una sombra una ficción
y el mayor bien es pequeño
Que toda la vida es sueño
y los sueños sueños son
;
proc surveyselect data=calderon method=urs outhits out=m n=2;
run;
proc print data=m;
```

El proc surveyselect se realiza 10 veces, apuntando el número de artículos encontrado. Para realizarlo 10 veces, se puede copiar y pegar las líneas que comienzan con el proc surveyselect 10 veces y ejecutarlo todo seguido.

Otra manera más elegante, cuando se quiere repetir un programa varias veces (en este caso ni siquiera es necesario cambiar nada en el programa) es crear una macro: las macros en SAS permiten repetir un proceso mediante bucles. Veamos en este caso su utilización:

```
%macro repecalderon;
%do i=1 %to 10;
  proc surveyselect data=calderon method=urs outhits out=m n=2 noprint;
  run;
  proc print data=m;
  run;
%end;
%mend;
```

La ejecución de la macro se realiza así:

```
%repecalderon;
```

Con lo cual se obtiene el resultado requerido. Si además no se quiere que en la ventana OUTPUT salga la información del proc surveyselect se puede utilizar en este procedimiento la opción noprint, como está escrito en la macro.

ISBN 9788496866133

En la ejecución, los valores obtenidos para el número de artículos en cada muestra han sido 1,0,1,1,0,1,0,1,0,0 (el lector habrá obtenido otros valores pues en este caso no hemos fijado la semilla y el SAS utiliza por defecto el reloj del ordenador). Las estimaciones del total en cada caso ($N\widehat{p}$) son 16.5,0,16.5,16.5,0,16.5,0,16.5,0,0.

c2) En casos como estos, es más fácil realizar una codificación previa. Normalmente se realiza un programa que verifica si la palabra está en una lista (en este caso la lista de artículos posibles) y se pone 1 si pertenece a la lista y cero si no. Haciéndolo a mano en este caso, quedaría:

```
data calderon;
input arti;
cards;
0 0 1 0 1 0
0 0 1 0 1 1
1 0 1 0
0 1 0 0 0 0
0 0 1 0 0 0
0 1 0 0 0
;
%macro repecalderon;
%do i=1 %to 10;
 proc surveyselect data=calderon method=urs outhits out=m n=2 noprint;
 run;
 %estimasr(m,arti,33,2);
%end;
%mend;

%repecalderon;
```

Donde se ha incluido dentro de la macro, la otra macro "estimasr" para estimar cada vez el total. Los resultados en este caso son parecidos.

d) Realizando el programa anterior, cambiando en el proc surveyselect a n=15, se obtiene (en nuestro caso): 2.2, 4.4, 11, 11, 2.2, 4.4, 4.4, 4.4, 2.2, 6.6.

e) Aunque no es un problema específico de muestreo, sino de cálculo de probabilidades, se presenta por su interés. Para ello hay primero que numerar las palabras de la 1 a la 33 (suponiendo el orden que aparece en los versos). El número de ordenaciones de esas palabras es 33! Pero de esas ordenaciones, hay que tener en cuenta que no es una sólo la que da lugar a los versos de Calderón. Como la palabra "es" está repetida, la ordenación de los versos 1...33 y la misma ordenación cambiando la palabra 2 por la 8 dan lugar a los mismos versos. Como la palabra "es" está repetida 4 veces, "que" 3 veces, etc. Hay exactamente 4!3!3!2!2!2!2!3! = 82944 ordenaciones que coinciden con los versos de Calderón (hemos contado las repeticiones usando el proc freq del SAS). Si todas las 33! ordenaciones tienen igual probabilidad, la probabilidad de concidir con los versos es

$$\frac{4!3!2!2!2!2!2!3!}{33!} = \frac{82944}{33!} = 9.55 \cdot 10^{-33}.$$

ISBN 9788496866133

3.8 Ejercicios propuestos

1) Sea el conjunto de la población $U = \{1, 2, 3\}$. Se realiza m.a.s.r. con $n = 2$. Encontrar la distribución conjunta de las dos componentes de la muestra y observar que las distribuciones marginales de las dos componentes son independientes e igualmente distribuidas.

2) En un área existen 10.000 viviendas. Los datos de un censo anterior hacen suponer que, aproximadamente, los 2/3 corresponden a régimen de alquiler. Se pide el tamaño muestral necesario para estimar con m.a.s.r. la proporción actual de viviendas en alquiler, con una varianza del estimador igual a 0.0016.

3) El número de viviendas de un municipio es igual a 5.200. Calcular el tamaño muestral necesario para estimar mediante m.a.s.r. el número de viviendas desocupadas, con una desviación típica del estimador igual a 40, sabiendo que una encuesta piloto ha mostrado que la proporción de viviendas desocupadas era 0.12.

4) En un proceso de control de calidad de componentes electrónicos, se debe devolver el componente al lote muy rápidamente una vez controlado, estando obligados por lo tanto a realizar muestreo con reemplazamiento. Hay 500 componentes en el lote, y tomando una muestra de 10 se obtuvieron los siguientes valores para el rendimiento de cada componente: 1,2,4,2,3,5,3,5,6,6. Además, se comprobó que los cinco primeros componentes tenían un defecto.

a) Suponiendo que la distribución del estimador es normal, presentar un intervalo de confianza al 95% para la media del rendimiento de los componentes.

b) Dar un intervalo de confianza al 95% para el rendimiento de los componentes con defecto.

c) Dar un intervalo de confianza al 95% para el rendimiento de los componentes sin defecto.

5) Se pide a 5 agricultores que elijan y midan por separado un terreno al azar en una parcela dividida que contiene 20 terrenos . Los dos primeros eligen el primer terreno de los 20 y observan que su tamaño es 500 m^2. Los tres últimos eligen los terrenos 14, 21 y 7, obteniendo tamaños de 100, 400 y 250 m^2 respectivamente.

a) Calcular un intervalo de confianza basado en la distribución normal, para el tamaño medio de los terrenos.

b) ¿Cuál era la probabilidad, a priori, de que se hubieran escogido los terrenos 1,1,14,21 y 7 por ese orden? ¿Cuál era la probabilidad a priori de escoger los mismos terrenos sin tener en cuenta el orden? ¿cuál era la probabilidad de que los terrenos 1 y 14 estuvieran en la muestra?

c) Presentar un intervalo de confianza para la proporción de terrenos mayores de 300 m^2.

6) Una población de tres unidades toma valores respectivos en y de $\{5, 6, 7\}$. Si se realiza m.a.s.r. de tamaño $n = 2$,

a) Presentar la distribución del estimador de la media poblacional $\widehat{\overline{y}}$.

b) Presentar la distribución del estimador de la varianza σ^2.

c) Calcular la esperanza y varianza del estimador de la media con m.a.s.r. y $n = 2$ directamente con la ayuda del apartado a). Calcular la esperanza y varianza del estimador de la media con m.a.s.r. y $n = 2$ si se supone conocida la información poblacional.

ISBN 9788496866133

d) Calcular el error de muestreo exacto si se toman muestras por m.a.s.r. de tamaño $n = 5$, y también con $n = 10$.

7) Se realiza un estudio de la longitud de los gatos en un centro de acogida. Hay 100 gatos en total, y se muestrean 10 por m.a.s.r. Se observa que hay 5 blancos y 5 negros. Se obtienen los datos generales:

$\sum_{i=1}^{10} y_i = 380$ y $\sum_{i=1}^{10} y_i^2 = 14440$.

Para los gatos blancos, se tiene

$\sum_{i=1}^{5} y_i = 210$ y $\sum_{i=1}^{8} y_i^2 = 8000$.

a) Suponiendo correcta la aproximación normal, presentar intervalos de confianza al 95% para la longitud media de los gatos en general, para la de los gatos blancos y para la de los gatos negros.

b) Presentar un intervalo de confianza para la proporción de gatos blancos.

c) Se desea estimar, con un grado de confianza del 95%, la proporción de gatos blancos con un error de muestreo absoluto de 0.10. ¿Cuántos gatos habría que muestrear con m.a.s.r.?.

d) Con la información muestral, ¿cuántos gatos habría que mirar para tener una probabilidad como mínimo de 0.80 de encontrarnos con 12 gatos negros? (recurrir al ejercicio resuelto 2.2).

8) En el archivo SAS esguinces está el número de altas por esguinces y luxaciones en hospitales de las provincias españolas en 1998 (hay 52).

a) Utilizar el proc surveyselect y la macro estimar en SAS para extraer una muestra por m.a.s.r. de tamaño $n = 10$ y estimar el número medio de esguinces por provincia. Dar un intervalo de confianza al 95%. Hacer lo mismo con el total de esguinces en España.

b) Estimar y dar un intervalo de confianza para la proporción de provincias que tienen más de 200 altas por esguinces.

ISBN 9788496866133

ISBN 9788496866133

4 MUESTREO ALEATORIO SIMPLE SIN REEMPLAZAMIENTO (m.a.s.)

El muestreo aleatorio simple sin reemplazamiento es considerado la referencia básica en todos los tipos de muestreo. Intuitivamente es el muestreo que aplicaría un investigador sin conocimientos de estadística, pues consiste en escoger de la población las unidades elementales, asignando igual probabilidad a cada una de ellas y sin repetir unidades. Como se vio en el tema introductorio, existe incluso una medida llamada efecto de diseño, para comparar cualquier tipo de muestreo con el muestreo aleatorio simple.

Sin embargo, su utilización práctica está restringida a poblaciones pequeñas y controladas, no muy dispersas en cuanto al tiempo necesario para ir de una unidad a otra. En poblaciones grandes (por ejemplo en un estudio realizado en la la capital de un país, considerando como unidades elementales los hogares) el coste de viajar a unidades seleccionadas por muestreo aleatorio simple es prohibitivo. En ocasiones, este tipo de muestreo no es fácil, pues la información en forma de listado no está disponible o es incompleta. En estos casos y en otros la alternativa del denominado muestreo sistemático, que se estudiará en el tema 6, es más útil.

4.1 Propiedades básicas

4.1.1 Probabilidades de obtención de muestras

En muestreo aleatorio simple sin reemplazamiento (m.a.s.) se seleccionan n unidades de una población de N unidades, sin reemplazamiento e igual probabilidad. Las extracciones son por lo tanto dependientes, y además las muestras obtenidas tienen todas sus unidades diferentes. Por simetría en el diseño y equiprobabilidad entre todas las unidades , cada una de las muestras de tamaño n tiene idéntica probabilidad. Se ha visto que el número de muestras diferentes teniendo en cuenta el orden es $n!\binom{N}{n}$. Por lo tanto, si $(u_1, u_2, ..., u_n)$ es una muestra ordenada, la probabilidad de obtener esa muestra en m.a.s. es

ISBN 9788496866133

$$P((u_1, u_2, ..., u_n)) = \frac{1}{n!\binom{N}{n}}$$

Si lo que deseamos es calcular la probabilidad de obtener una muestra $\{u_1, ..., u_n\}$ sin tener en cuenta el orden, el número de casos favorables es el número de muestras ordenadas que contienen los n elementos distintos $\{u_1, ..., u_n\}$, es decir, el número de permutaciones de los n elementos, $n!$. Así, por equiprobabilidad de las muestras se puede aplicar la regla de Laplace:

$$P(\{u_1, ..., u_n\}) = \frac{\text{Casos favorables}}{\text{Casos posibles}} = \frac{n!}{n!\binom{N}{n}} = \frac{1}{\binom{N}{n}}.$$

Por consenso, en m.a.s. no se suele tener en cuenta el orden en las muestras obtenidas, debido a que se llega a los mismos resultados en el cálculo de probabilidades sobre los estimadores, pero para aclarar conceptos se ilustran en los ejemplos las posibilidades muestrales también cuando el orden se tiene en cuenta.

Cuando el muestreo es con reemplazamiento (m.a.s.r.) sí se tiene en cuenta el orden, para evitar complicaciones en la enumeración de las muestras posibles.

Ejemplo 4.1.

Supongamos una urna con 4 unidades numeradas 1,2,3,4. Se toman 3 unidades según m.a.s. Calculemos:

a) La probabilidad de obtener primero un 1, después un 2 y después otra vez un 3 .

b) La probabilidad de obtener un 1, un 2 y un 3 en nuestra muestra sin importar el orden.

c) Verificar que hay 6 muestras de las 24 posibles que contienen un uno, un dos y un tres.

d) Si obtenemos la media muestral calculada según los números que aparecen en las bolas de la muestra, calcular la probabilidad de que esta media sea 7/3.

---

a) $P((1,2,3)) = \frac{1}{3!\binom{4}{3}} = \frac{1}{24} = 0.04166.$

b) $P(\{1,2,3\}) = \frac{1}{\binom{4}{3}} = \frac{1}{4} = 0.25.$

c) $(1,2,3), (2,1,3), (1,3,2), (2,3,1), (3,1,2), (3,2,1).$

d) Para ver los diferentes valores que puede tomar esa media muestral, es necesario calcularla para las $\binom{N}{n} = 4$ diferentes muestras donde no importa el orden (pues la media es idéntica si las muestras cambian internamente de orden). Mostramos las muestras con la media muestral calculada entre paréntesis:

$\{1,2,3\}(3); \{1,2,4\}(7/3); \{1,3,4\}(8/3); \{2,3,4\}(9/3).$

Por lo tanto las muestras que dan lugar a una media muestral de 7/3 son las que contienen un 1, un 2 y un 4. Entonces $P(\{1,2,4\}) = \frac{1}{\binom{4}{3}} = \frac{1}{4}$ es la probabilidad de que la media muestral sea 7/3 en

ISBN 9788496866133

muestreo sin reemplazamiento en esta población.

4.1.2 Probabilidades de Inclusión

Propiedad 4.1 (probabilidad de inclusión) .

Para cada $i = 1, ..., N$, la probabilidad de inclusión π_i en muestreo aleatorio simple sin reemplazamiento es $\pi_i = \dfrac{n}{N}$.

Demostración.

El número de muestras no ordenadas de tamaño n que contienen la unidad i se calcula "fijando" i en la muestra no ordenada y calculando cuántas muestras no ordenadas hay para rellenar las $n-1$ restantes celdas con las $N-1$ restantes unidades poblacionales. Estas son $\binom{N-1}{n-1}$. Por equiprobabilidad de las muestras se puede aplicar la regla de Laplace: $\pi_i = \dfrac{\text{Casos favorables}}{\text{Casos posibles}} = \dfrac{\binom{N-1}{n-1}}{\binom{N}{n}} = \dfrac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{(N)!}{(n)!(N-n)!}} = \dfrac{n}{N}$.

Propiedad 4.2 (probabilidad de inclusión) .

Para cada $i, j = 1, ..., N$, $i \neq j$, la probabilidad de inclusión π_{ij} en muestreo aleatorio simple sin reemplazamiento es $\pi_{ij} = \dfrac{n(n-1)}{N(N-1)}$.

Demostración.

Con un razonamiento análogo a la demostración anterior, tenemos que

$$\pi_{ij} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}.$$

Ejemplo 4.2.

Supongamos una urna con 4 unidades numeradas 1,2,3,4. Se toman 2 unidades según m.a.s. .

a) Calcular la probabilidad de obtener un 1 en la muestra a través de la expresión de la probabilidad de inclusión, y también por simple conteo.

b) Calcular la probabilidad de obtener un 1 y un 2 en la muestra a través de la expresión de la probabilidad de inclusión, y también por simple conteo.

---

a) $\pi_1 = \dfrac{n}{N} = \dfrac{2}{4} = \dfrac{1}{2}$.

Si tenemos en cuenta el orden hay las siguientes muestras posibles ($n!\binom{N}{n} = 12$):

ISBN 9788496866133

$(1,2),(1,3),(1,4),(2,3),(2,4),(3,4),(2,1),(3,1),(4,1),(3,2),(4,2),(4,3),$

de las cuales 6 contienen un 1: $(1,2),(1,3),(1,4),(2,1),(3,1),(4,1)$. Así, $\pi_1 = \dfrac{6}{12} = \dfrac{1}{2}$.

Si no tenemos en cuenta el orden, hay $\binom{N}{n} = 6$ muestras posibles:

$\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\}$

de las cuales 3 contienen un 1: $\{1,2\},\{1,3\},\{1,4\}$. Con lo que $\pi_1 = \dfrac{3}{6} = \dfrac{1}{2}$.

Como se ha comentado, en los cálculos de probabilidades en m.a.s por consenso no se tiene en cuenta el orden, pues se llega a los mismos resultados que teniéndolo en cuenta debido a la simetría del diseño, y es más sencillo.

b) $\pi_{12} = \dfrac{n(n-1)}{N(N-1)} = \dfrac{2(2-1)}{4(4-1)} = \dfrac{2}{12} = \dfrac{1}{6}$.

Por conteo: teniendo en cuenta el orden: 2 casos favorables de los 12 (observar muestras posibles en apartado a))=$\frac{1}{6}$. Sin tener en cuenta el orden: un caso favorable de los 6 posibles=$\frac{1}{6}$.

4.1.3 Selección de una muestra aleatoria simple sin reemplazamiento

Supongamos que disponemos de un listado con las unidades poblacionales numeradas del 1 al N.

Tablas de números aleatorios-. Se utiliza el método visto en m.a.s.r. pero si un número está ya escogido se descarta.

Calculadora y ordenador-. Tanto con calculadora como con ordenador se puede utilizar el método visto en m.a.s.r. , descartando las unidades que ya están en la muestra. Otro método más eficaz a menudo consiste en generar N observaciones $U(0,1)$, haciendo corresponder el primer número generado con la primera unidad de la población, el segundo con la segunda unidad y así sucesivamente. A cada unidad poblacional le corresponde un par (i,u), donde i es su orden inicial y u el valor de la v.a. $U(0,1)$. A continuación se ordenan los N pares por los números aleatorios generados u y se escogen las unidades que ocupan las n primeras posiciones $j = 1,...,n$ en esa nueva lista ordenada. Con ordenador, tanto en paquetes estadísticos como lenguajes de programación y hojas de cálculo, es un método muy sencillo de desarrollar.

Ejemplo 4.3.

Tenemos una población con $N = 5$ y con las unidades numeradas $1,2,3,4,5$. Supongamos que queremos una m.a.s. de tamaño $n = 3$.

1) Creamos con la calculadora u ordenador $N = 5$ números provenientes de una $U(0,1)$. Obtenemos la lista siguiente:

ISBN 9788496866133

```
i  u
1  0.529
2  0.955
3  0.583
4  0.812
5  0.584
```

2) Ordenamos la lista por u :

```
i  u
1  0.529
3  0.583
5  0.584
4  0.812
2  0.955
```

3) Tomamos los $n = 3$ primeros items de la lista: las unidades seleccionadas son aquellas con orden inicial $i = 1, 3, 5$.

4.2 Estimación en muestreo aleatorio simple sin reemplazamiento

4.2.1 Estimación de la media poblacional

Teorema 4.1 (estimación de la media).

La media muestral $\widehat{\overline{y}} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$ es un estimador insesgado de la media poblacional.

Demostración.

En primer lugar hay que ver cuál es la probabilidad de que cada unidad y_i de la muestra tome valor y_j , siendo y_j la unidad j-ésima de la población $y_1, y_2, ..., y_N$. Para ello basta ver que de las $n!\binom{N}{n}$ muestras ordenadas posibles, hay exactamente $(n-1)!\binom{N-1}{n-1}$ muestras ordenadas que tienen valor y_j en la posición i (en este caso hay que calcular las probabilidades teniendo en cuenta el orden pues i es la posición ordenada del elemento muestral). Por lo tanto, para cada valor y_j perteneciente a la población $y_1, y_2, ..., y_N$ e y_i el elemento muestral que está en la posición i, se tiene

$$P(\text{ el elemento muestral } y_i = y_j) = \frac{\text{Casos favorables}}{\text{Casos posibles}} = \frac{(n-1)!\binom{N-1}{n-1}}{n!\binom{N}{n}} = \frac{1}{N} \text{ para todo } i = 1, ..., n, \text{ y para todo } j = 1, ..., N$$

Es decir, en m.a.s. cada unidad muestral y_i toma valores de la población $y_1, y_2, ..., y_N$ con probabilidades $\dfrac{1}{N}, ..., \dfrac{1}{N}$, al igual que ocurría en m.a.s.r.

ISBN 9788496866133

Entonces, $E(y_i) = \sum_{i=1}^{N} \frac{1}{N} y_i = \overline{y}$ para todo $i = 1, ..., n$, y además

$$V(y_i) = \sum_{j=1}^{N} \frac{1}{N}(y_j - E(y_j))^2 = \frac{1}{N}\sum_{j=1}^{N}(y_j - \overline{y})^2 = \sigma^2.$$

Así, para demostrar que $\widehat{\overline{y}}$ es insesgado para $\overline{y}$ basta ver que

$$E(\widehat{\overline{y}}) = E(\frac{1}{n}\sum_{i=1}^{n} y_i) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \overline{y}.$$

Teorema 4.2 (varianza del estimador).

La varianza de $\widehat{\overline{y}}$ en m.a.s. es $V(\widehat{\overline{y}}) = \frac{N-n}{N}\frac{S^2}{n} = \frac{N-n}{N-1}\frac{\sigma^2}{n}$.

Se define $f = \frac{n}{N}$ como la **fracción de muestreo** y $(1-f)$ como el **factor de corrección por población finita (c.p.f.)**, $V(\widehat{\overline{y}}) = (1-f)\frac{S^2}{n}$.

Demostración.

$$V(\widehat{\overline{y}}) = E((\widehat{\overline{y}} - \overline{y})^2) = E\left((\frac{1}{n}\sum_{i=1}^{n} y_i - \overline{y})^2\right) = E\left(\frac{1}{n^2}(\sum_{i=1}^{n} y_i - n\overline{y})^2\right) =$$

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

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

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

$$= \frac{\sigma^2}{n} + \frac{1}{n^2}\sum_{i\neq j}^{n} E\left((y_i - \overline{y})(y_j - \overline{y})\right) = (1).$$

Ahora, $(y_i - \overline{y})(y_j - \overline{y})$ es una variable aleatoria que toma valores en todos los posibles $(y_{i'} - \overline{y})(y_{j'} - \overline{y})$ con $y_{i'}, y_{j'}$ pertenecientes a la población. Cada uno de estos valores tiene la misma probabilidad que la de que los dos elementos muestrales y_i y y_j tomen valores $y_{i'}, y_{j'}$ en una muestra de tamaño n. Razonando como en la demostración anterior, esta probabilidad es $\frac{(n-2)!\binom{N-2}{n-2}}{n!\binom{N}{n}} = \frac{1}{N(N-1)}$. Por lo tanto la esperanza es

$$E\left((y_i - \overline{y})(y_j - \overline{y})\right) = \sum_{i\neq j}^{N} \tfrac{1}{N(N-1)}(y_i - \overline{y})(y_j - \overline{y}).$$

Además, el número de sumandos en $\sum_{i\neq j}^{n}$ es exactamente $2\binom{n}{2} = n(n-1)$ por ser el número de pares i, j en una muestra de tamaño n (contando también los pares j, i). Por lo tanto,

ISBN 9788496866133

$$(1) \quad = \frac{\sigma^2}{n} + \frac{1}{n^2} n(n-1) \sum_{i \neq j}^{N} \frac{1}{N(N-1)} (y_i - \overline{y})(y_j - \overline{y}) =$$

$$= \frac{\sigma^2}{n} + \frac{(n-1)}{nN(N-1)} \sum_{i \neq j}^{N} (y_i - \overline{y})(y_j - \overline{y}) \overset{(*)}{=}$$

$$\overset{(*)}{=} \frac{\sigma^2}{n} + \frac{(n-1)}{nN(N-1)} (-N\sigma^2) = \frac{N-n}{n(N-1)} \sigma^2 = \frac{N-n}{N} \frac{S^2}{n}$$ que es lo que se quería demostrar.

La igualdad $(*)$ se obtiene de que $\sum_{i \neq j}^{N} (y_i - \overline{y})(y_j - \overline{y}) = -N\sigma^2$. En efecto, como $\sum_{i=1}^{N} (y_i - \overline{y}) = 0 \Rightarrow \left(\sum_{i=1}^{N} (y_i - \overline{y}) \right)^2 = 0 \Rightarrow$

$$\Rightarrow \left(\sum_{i=1}^{N} (y_i - \overline{y})^2 + \sum_{i \neq j}^{N} (y_i - \overline{y})(y_j - \overline{y}) \right) = 0 \Rightarrow$$

$$\Rightarrow \sum_{i \neq j}^{N} (y_i - \overline{y})(y_j - \overline{y}) = - \sum_{i=1}^{N} (y_i - \overline{y})^2 = -N\sigma^2.$$

En la práctica es necesario estimar esta varianza como referencia para estudiar la precisión del estimador de la media.

Teorema 4.3 (estimación de la varianza del estimador).

En muestreo aleatorio simple s^2 es un estimador insesgado de S^2, y por lo tanto un estimador insesgado de la varianza de $\widehat{\overline{y}}$ en m.a.s. es $\widehat{V}(\widehat{\overline{y}}) = \frac{N-n}{N} \frac{s^2}{n}$.

Demostración.

$$E(s^2) = \frac{1}{n-1} E\left(\sum_{i=1}^{n} (y_i - \widehat{\overline{y}})^2 \right) = \frac{1}{n-1} E\left(\sum_{i=1}^{n} y_i^2 - n\widehat{\overline{y}}^2 \right) =$$

$$= \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^{n} E(y_i^2) - E(\widehat{\overline{y}}^2) \right).$$

Como $E(\widehat{\overline{y}}^2) = V(\widehat{\overline{y}}) + E(\widehat{\overline{y}})^2 = \frac{N-n}{N} \frac{S^2}{n} + \overline{y}^2$, la última igualdad queda

$$E(s^2) = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^{n} E(y_i^2) - \frac{N-n}{N} \frac{S^2}{n} - \overline{y}^2 \right).$$

Ahora, como y_i^2 es una variable aleatoria que toma valores de la población $y_1^2, ..., y_N^2$ cada uno con probabilidad $\frac{1}{N}$, tenemos que

$$E(s^2) = \frac{n}{n-1} \left(\frac{1}{N} \sum_{i=1}^{N} y_i^2 - \frac{N-n}{N} \frac{S^2}{n} - \overline{y}^2 \right) =$$

$$= \frac{n}{n-1} \left(\frac{1}{N} (\sum_{i=1}^{N} y_i^2 - N\overline{y}^2) - \frac{N-n}{N} \frac{S^2}{n} \right) =$$

ISBN 9788496866133

$= \frac{n}{n-1}\left(\frac{N-1}{N}S^2 - \frac{N-n}{N}\frac{S^2}{n}\right) = S^2$ que es lo que se quería demostrar.

Ahora, como $E(\ \widehat{V}(\widehat{\overline{y}})) = \frac{N-n}{N}\frac{E(s^2)}{n} = \frac{N-n}{N}\frac{S^2}{n}$ tenemos que $\widehat{V}(\widehat{\overline{y}})$ es un estimador insesgado de $V(\widehat{\overline{y}})$.

Ejemplo 4.4.

Sea una población de 4 unidades, con valores respectivos $y_i = i$. Está claro que $\overline{y} = \frac{1}{4}(1+2+3+4) = 2.5$ y $\sigma^2 = \frac{1}{4}((1-2.5)^2 + ...) = \frac{5}{4}$. Supongamos que seleccionamos una muestra de tamaño $n = 2$. Las muestras posibles no ordenadas son $\binom{N}{n} = \binom{4}{2} = 6$. Cada una de estas muestras tiene la misma probabilidad $\frac{1}{6}$de ser escogida, y da lugar a un valor de la media muestral $\widehat{\overline{y}}$:

muestra s	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
$\widehat{\overline{y}}$	1.5	2	2.5	2.5	3	3.5

Tabla 4.1. Medias muestrales para cada muestra posible.

La media muestral es por lo tanto una variable aleatoria que toma valores 1.5, 2,..., cada uno con probabilidad $\frac{1}{6}$. Vemos que la media muestral es un estimador insesgado pues $E(\widehat{\overline{y}}) = \frac{1}{6}(1.5+2+2.5+2.5+3+3.5) = 2.5 = \overline{y}$. La varianza $V(\widehat{\overline{y}})$ se puede calcular directamente, como $V(\widehat{\overline{y}}) = \frac{1}{6}((1.5-2.5)^2 + ... + (3.5-2.5)^2) = 0.4166$, pero también a través de la expresión demostrada anteriormente, $V(\widehat{\overline{y}}) = \frac{N-n}{N}\frac{S^2}{n}$. Como el valor poblacional $S^2 = \frac{N}{N-1}\sigma^2 = \frac{4}{3}\frac{5}{4} = \frac{5}{3}$, $V(\widehat{\overline{y}}) = \frac{4-2}{4}\frac{5}{3\cdot 2} = \frac{5}{12} = 0.4166$.

Vamos a comprobar ahora que el estimador $\widehat{V}(\widehat{\overline{y}})$ es insesgado para $V(\widehat{\overline{y}})$. Para cada muestra posible de las 6, calculamos $\widehat{V}(\widehat{\overline{y}}) = \frac{N-n}{N}\frac{s^2}{n}$ obteniendo los siguientes valores: $0.125, 0.5, 1.125, 0.125, 0.5, 0.125$. Como cada una de esas muestras tiene probabilidad $\frac{1}{6}$, $\widehat{V}(\widehat{\overline{y}})$ es una variable aleatoria que toma valores $0.125, 0.5, ...$ cada uno con probabilidad $\frac{1}{6}$. Así, tenemos que $E(\widehat{V}(\widehat{\overline{y}})) = \frac{1}{6}(\ 0.125+0.5+1.125+0.125+0.5+0.125) = 0.4166$ como queríamos comprobar.

4.2.2 Estimación del Total poblacional

Corolario 4.1 (estimación del total).

(a) Un estimador insesgado del total poblacional.en m.a.s. es $N\widehat{\overline{y}}$, con varianza $V(N\widehat{\overline{y}}) = N(N-n)\frac{S^2}{n}$.

(b) $N(N-n)\frac{s^2}{n}$ es un estimador insesgado de $V(N\widehat{\overline{y}})$.

ISBN 9788496866133

4.2.3 Estimación de la Proporción poblacional

Si y es una variable dicotómica ($y_i = 1$ si la unidad poblacional i tiene cierta cualidad, $y_i = 0$ si no la tiene), media muestral coincide con proporción muestral y media poblacional coincide con proporción poblacional. Así los siguientes resultados son inmediatos.

Corolario 4.2 (estimación de la proporción).

(a) La proporción muestral $\widehat{p}$ es un estimador insesgado de la proporción poblacional p.

(b) $V(\widehat{p}) = \dfrac{N-n}{N}\dfrac{S^2}{n} = \dfrac{N-n}{N}\dfrac{Np(1-p)}{(N-1)n} = \dfrac{N-n}{N-1}\dfrac{pq}{n}$.

(c) $s^2 = \dfrac{n}{n-1}\widehat{pq}$ es un estimador insesgado de $S^2 = \dfrac{N}{N-1}pq$ y además $\widehat{V}(\widehat{p}) = \dfrac{N-n}{N}\dfrac{\widehat{pq}}{n-1}$ es un estimador insesgado de $V(\widehat{p}) = \dfrac{N-n}{N}\dfrac{pq}{n}$.

Veremos ahora una propiedad de la varianza del estimador en el caso de la proporción.

Teorema 4.4 (cotas para la varianza del estimador de la proporción).

$$V(\widehat{p}) \leq \frac{N-n}{4(N-1)n} \leq \frac{1}{4}$$

Demostración.

Como $pq \leq \dfrac{1}{4}$ pues $p \leq \frac{1}{2}$ y $q \leq \frac{1}{2}$, $V(\widehat{p}) = \dfrac{N-n}{N-1}\dfrac{pq}{n} \leq \dfrac{N-n}{4(N-1)n} \leq \dfrac{1}{4}$.

Es conveniente resaltar que la varianza del estimador es proporcional a $p(1-p) = p-p^2$, función simétrica que alcanza su máximo en $p = q = \dfrac{1}{2}$, como se ve en la Figura 4.1. La varianza del estimador en la estimación de una proporción en m.a.s., y también en m.a.s.r., será menor en los casos en los que p esté cerca de 0 o 1, y máxima en $p = \dfrac{1}{2}$. Por ello a menudo se presentan los cálculos de tamaño muestral de manera conservadora, situándonos en el peor caso, $p = q = \dfrac{1}{2}$, sin necesidad de tomar previamente una muestra piloto.

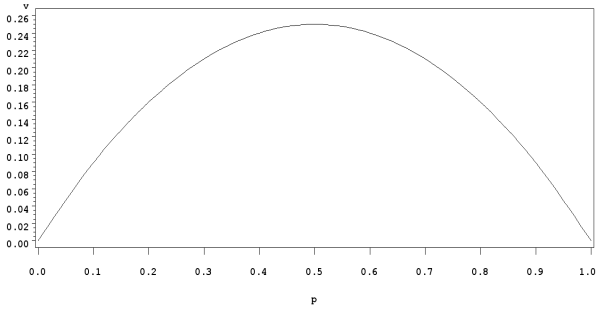


Figura 4.1. Varianza de $\widehat{p}$ en función de p.

ISBN 9788496866133

Ejemplo 4.5.

Se ha seleccionado una m.a.s. de tamaño $n = 100$ para estimar: a) la fracción de $N = 300$ estudiantes de COU que pretenden asistir a la Universidad, y b) la fracción de estudiantes que han trabajado a tiempo parcial durante su estancia en el instituto.

Sean y_i y x_i $(i = 1, ..., 100)$ las respuestas del iésimo estudiante seleccionado, dando valor $y_i = 1$ si pretende asistir a la Universidad , con $y_i = 0$ si no, y $x_i = 1$ si trabaja a tiempo parcial , con $x_i = 0$ si no. Supongamos que el número de respuestas positivas en el primer caso son 25, es decir, $\sum_{i=1}^{n} y_i = 25$, y en el segundo caso 30, o sea $\sum_{i=1}^{n} x_i = 30$.

Los estimadores respectivos de ambas fracciones (proporciones) son $\widehat{p}_1 = \frac{1}{100}25 = 0.25$ y $\widehat{p}_2 = \frac{1}{100}30 = 0.30$. Las varianzas estimadas son

$$\widehat{V}(\widehat{p}_1) = \frac{N-n}{N}\frac{\widehat{p}_1\widehat{q}_1}{n-1} = \frac{300-100}{300}\frac{0.25(0.75)}{99} = 0.00126 \text{ y}$$

$$\widehat{V}(\widehat{p}_2) = \frac{N-n}{N}\frac{\widehat{p}_2\widehat{q}_2}{n-1} = \frac{300-100}{300}\frac{0.30(0.70)}{99} = 0.00141.$$

4.3 Corrección por Población Finita . Comparación entre m.a.s.r. y m.a.s.

En todas las expresiones de las varianzas y en las varianzas estimadas de los estimadores aparece el término $\frac{N-n}{N} = \left(1 - \frac{n}{N}\right) = (1-f)$. Cuando el estimador se expresa en función de la varianza σ^2 en vez de en función de la cuasivarianza S^2, el término que aparece es $\frac{N-n}{N-1}$.Por ejemplo, en la estimación de la media poblacional, $V(\widehat{\overline{y}}) = \frac{N-n}{N}\frac{S^2}{n} = \frac{N-n}{N-1}\frac{\sigma^2}{n}$. Si la fracción de muestreo $f = \frac{n}{N}$ es pequeña, los términos $\frac{N-n}{N}$ y $\frac{N-n}{N-1}$ son cercanos a la unidad y pueden ser despreciados (eliminados) en las fórmulas de varianzas. Está claro que cuando $N \to \infty$, estos términos son 1, así que despreciar un término c.p.f. equivale a considerar la población infinita.

El efecto práctico de esta simplificación es sobreestimar la varianza del estimador. Esta sobreestimación, en m.a.s., se cifra exactamente en $(1 - \frac{N-n}{N})\frac{S^2}{n} = \frac{S^2}{N}$ unidades, con lo que conocer una estimación precisa de S^2, ayudaría a la decisión de la eliminación del término de corrección por población finita y consecuente simplificación de todas las fórmulas.

Ejemplo 4.6.

La provincia de Girona tiene $N = 209$ municipios de menos de 10.000 habitantes. Si se desea estimar la media poblacional del total de habitantes por municipio con una m.a.s. de tamaño n, y se supone que se conoce por censos anteriores la cuasivarianza de la variable y =total de habitantes, y ésta

ISBN 9788496866133

es $S^2 = 2709316$, la sobreestimación de la varianza del estimador en que se incurre al despreciar el término $\frac{N-n}{N}$ es del orden de $\frac{S^2}{N} = \frac{2709316}{209} = 12963$, lo que en términos de desviación típica del estimador es de 113 unidades (habitantes).

La semianchura del intervalo de confianza presentado será aproximadamente $z_{\alpha/2} \cdot 113$ unidades mayor que si no desprecia el término $\frac{N-n}{N}$. Para un intervalo de confianza con $\alpha = 0.05$, esto equivale a 223 unidades más ancho en cada lado. Obsérvese que esta sobreestimación no depende de n.

Si lo que se persigue es estimar la superficie media de los municipios de menos de 10.000 habitantes en Girona, $S^2 = 520.29$ y por lo tanto la sobreestimación de la varianza del estimador será solamente del orden de $\frac{520.29}{209} = 2.49$.

Si por el contrario lo que se quiere es estimar la media poblacional del total de habitantes por municipio en toda España, en los municipios de menos de 10.000 habitantes, $N = 7458$, y $S^2 = 3668872$ y entonces $\frac{S^2}{N} = 492$, lo que lleva a un aumento de anchura del intervalo al 95% de 22 unidades, en lugar de las 223 unidades que daba en el caso particular de la provincia de Girona.

De este ejemplo se deduce que el tamaño poblacional afecta al término c.p.f. y por lo tanto en la decisión de eliminarlo por simplificación, pero también influyen la variabilidad, escala de la variable y pretensiones de precisión del investigador.

En la estimación de la media en m.a.s.r., vimos que $V(\widehat{\overline{y}})_{m.a.s.r.} = \frac{\sigma^2}{n}$. En m.a.s. es $V(\widehat{\overline{y}})_{m.a.s.} = \frac{N-n}{N-1}\frac{\sigma^2}{n}$. Por lo tanto, $V(\widehat{\overline{y}})_{m.a.s.r.} > V(\widehat{\overline{y}})_{m.a.s.}$ siempre, con lo cual el muestreo aleatorio simple sin reemplazamiento es un método de muestreo más preciso que el muestreo aleatorio simple con reemplazamiento .

Si la fracción de muestreo f es pequeña, es decir, N es muy grande respecto a n, y se considera el término $\frac{N-n}{N-1}$ como 1, ambas varianzas son equivalentes con lo cual se consideran los dos métodos similares. Esto es obvio también teniendo en cuenta que si f es pequeña, la probabilidad de caer en una repetición en m.a.s.r. es muy baja y por lo tanto ambos métodos darán lugar en la práctica a muestras equiprobables similares.

Hay que remarcar también que $V(\widehat{\overline{y}})_{m.a.s.r.} - V(\widehat{\overline{y}})_{m.a.s.} = \frac{n-1}{n}\frac{S^2}{N}$ es el aumento absoluto de varianza del estimador correspondiente a utilizar m.a.s.r. en lugar de m.a.s. Es habitual también expresar las comparaciones relativas a la precisión en términos de cociente: $\frac{V(\widehat{\overline{y}})_{m.a.s.r.}}{V(\widehat{\overline{y}})_{m.a.s.}}$ expresa la precisión del m.a.s. (inverso de la varianza del estimador) dividida por la precisión del m.a.s.r. , y es

$$\frac{V(\widehat{\overline{y}})_{m.a.s.r.}}{V(\widehat{\overline{y}})_{m.a.s.}} = \frac{N-1}{N-n}$$

indicando el aumento relativo porcentual de precisión del muestreo aleatorio simple sin reemplazamiento respecto al m.a.s. con reemplazamiento. Cuanto mayor sea n, mayor va a ser la ganancia relativa en precisión del m.a.s. frente al m.a.s.r. (obsérvese que si $n = N$ en m.a.s.

ISBN 9788496866133

sin reemplazamiento, obtenemos todas las unidades poblacionales y por tanto información perfecta, mientras que si $n = N$ en m.a.s. con reemplazamiento nada asegura obtener todas las unidades, al poder repetirse las unidades muestreadas).

En general, aunque el m.a.s.r. es menos preciso que m.a.s., a veces ese tipo de muestreo reduce costes, pues las repeticiones no generan coste alguno. Por lo cual, para el mismo coste prefijado, m.a.s.r. puede obtener en algunos casos, en promedio, una precisión mayor que m.a.s.

Ejemplo 4.7.

Basándose en el ejemplo 4.6, $V(\widehat{\overline{y}})_{m.a.s.r.} - V(\widehat{\overline{y}})_{m.a.s.} = \dfrac{n-1}{n}\dfrac{S^2}{N} = \dfrac{n-1}{n}12963$. y $\dfrac{V(\widehat{\overline{y}})_{m.a.s.r.}}{V(\widehat{\overline{y}})_{m.a.s.}} = \dfrac{N-1}{N-n} = \dfrac{209-1}{209-n}$. A medida que aumenta n, la diferencia entre las varianzas del estimador aumenta, pero sólo hasta la cantidad mínima $\dfrac{S^2}{N}$, como se ve en la Figura 4.2. La precisión relativa del m.a.s. frente al m.a.s.r. aumenta también con n, como se observa en la Figura 4.3.

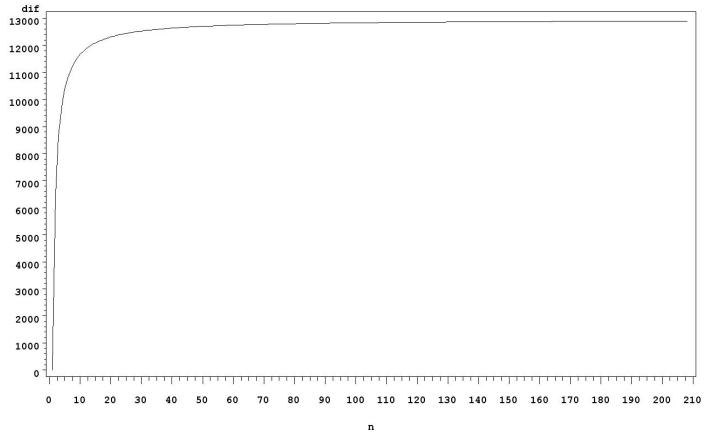


Figura 4.2. Diferencia entre varianzas $\dfrac{n-1}{n}\dfrac{S^2}{N}$

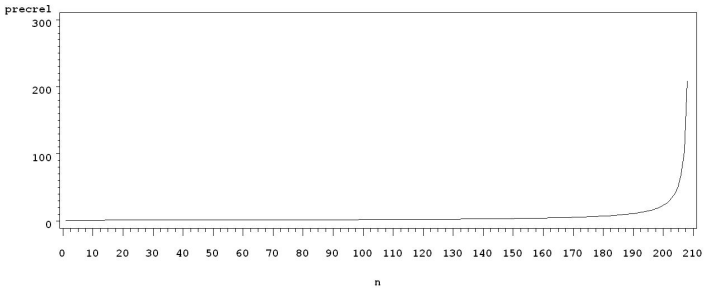


Figura 4.3. Precisión relativa del m.a.s. frente al m.a.s.r.

ISBN 9788496866133

4.4 Determinación del tamaño de la muestra en m.a.s.

Supondremos Normalidad del estimador, que se cumple con pocas restricciones, por el Teorema Central del Límite, al ser una suma de variables aleatorias idénticamente distribuidas (aunque la hipótesis de independencia no es cierta en este caso). Supondremos que $\widehat{S}^2$ e $\widehat{\overline{y}}$ son aproximaciones a S^2 e $\overline{y}$ obtenidas posiblemente a través de una muestra piloto. En el caso de estimación de proporciones, se sustituirá $\widehat{S}^2$ por $\frac{N}{N-1}\widehat{p}(1-\widehat{p})$ e $\widehat{\overline{y}}$ por $\widehat{p}$, siendo $\widehat{p}$ la proporción estimada en la muestra piloto. Eventualmente se toma directamente $\widehat{p}=0.5$ (el caso peor) para evitar tener que obtener información de una muestra piloto, y en este caso se sustituye $\widehat{S}^2$ por $\frac{N}{N-1}\widehat{pq} = \frac{N}{N-1}\frac{1}{4}$.

En todos los desarrollos siguientes se debe tener en cuenta que si N es grande, los términos de c.p.f. $\frac{N-n}{N}$ pueden ser eliminados de las expresiones dando lugar a fórmulas más sencillas.

4.4.1 Tamaño muestral con error de muestreo prefijado

Se plantea el problema de determinar cuál es el tamaño muestral mínimo necesario para obtener un valor de error de muestreo=ϕ en estimación de la media en m.a.s. Como el error de muestreo es $\phi = \sqrt{V(T)} = \sqrt{V(\widehat{\overline{y}})}$ y $V(\widehat{\overline{y}}) = \frac{N-n}{N}\frac{S^2}{n}$, basta hacer $\frac{N-n}{N}\frac{\widehat{S}^2}{n} = \phi^2$ y por lo tanto $N\widehat{S}^2 - n\widehat{S}^2 = \phi^2 Nn \Rightarrow n(N\phi^2+\widehat{S}^2) = N\widehat{S}^2 \Rightarrow n = \frac{N\widehat{S}^2}{N\phi^2+\widehat{S}^2} = \frac{1}{\frac{1}{N}+\frac{\phi^2}{\widehat{S}^2}}$.

Para el caso de las proporciones, queda:

$$n = \frac{N\widehat{S}^2}{N\phi^2+\widehat{S}^2} = \frac{N\frac{N}{N-1}\widehat{p}(1-\widehat{p})}{N\phi^2+\frac{N}{N-1}\widehat{p}(1-\widehat{p})} = \frac{N\widehat{p}(1-\widehat{p})}{\phi^2(N-1)+\widehat{p}(1-\widehat{p})}.$$

4.4.2 Tamaño muestral con error de muestreo relativo prefijado

Se trata de determinar cuál es el tamaño muestral mínimo necesario para obtener un valor de error de muestreo relativo=ϕ.en estimación de la media en m.a.s. Se supone que se dispone de una estimación previa $\widehat{\overline{y}}$ de la media poblacional, de una muestra piloto. Como el error de muestreo relativo es $\phi = \frac{\sqrt{V(\widehat{\overline{y}})}}{\widehat{\overline{y}}} \simeq \sqrt{\frac{N-n}{N}\frac{\widehat{S}^2}{n}\frac{1}{\widehat{\overline{y}}}}$ y por lo tanto, elevando al cuadrado en ambos lados de la igualdad y despejando n como en el caso anterior, $n = \frac{1}{\frac{1}{N}+\frac{(\phi\widehat{\overline{y}})^2}{\widehat{S}^2}} = \frac{N\widehat{S}^2}{\widehat{S}^2+N(\phi\widehat{\overline{y}})^2}$.

Para proporciones, queda

ISBN 9788496866133

$$n = \frac{1}{\frac{1}{N} + \frac{(\phi\widehat{\overline{y}})^2}{\widehat{S}^2}} = \frac{N(1-\widehat{p})}{(1-\widehat{p}) + (N-1)\widehat{p}\phi^2}.$$

4.4.3 Tamaño muestral con error de muestreo absoluto prefijado

El problema es ahora determinar cuál es el tamaño muestral mínimo necesario para obtener un valor de error de muestreo absoluto=e.en estimación de la media en m.a.s. Se conoce el nivel de confianza α, y se supone normalidad, con lo que $z_{\alpha/2}$ es conocido. Como el error de muestreo absoluto es $e = z_{\alpha/2}\sqrt{V(\widehat{\overline{y}})} \simeq z_{\alpha/2}\sqrt{\frac{N-n}{N}\frac{\widehat{S}^2}{n}}$ y por lo tanto $n = \frac{1}{\frac{1}{N} + \frac{e^2}{z^2_{\alpha/2}\widehat{S}^2}} = \frac{Nz^2_{\alpha/2}\widehat{S}^2}{z^2_{\alpha/2}\widehat{S}^2 + Ne^2}$.

Para proporciones, es:

$$n = \frac{Nz^2_{\alpha/2}\widehat{p}(1-\widehat{p})}{z^2_{\alpha/2}\widehat{p}(1-\widehat{p}) + (N-1)e^2}$$

4.4.4 Tamaño muestral considerando costes

En ocasiones puede desarrollarse un enfoque más práctico para estimar el tamaño de la muestra, basado en el coste de obtención de la muestra. Sean las siguientes cantidades:

z = error en la estimación muestral=valor del estimador t-verdadero valor del parámetro $\theta = (t-\theta)$.

$l(z)$ = pérdida al tomar una decisión que lleva al error z.

$f(z,n)$ =distribución de probabilidad de z, dado el tamaño muestral n.

$L(n)$= $\int l(z)f(z,n)dz$ =pérdida esperada dado n.

$C(n)$ =coste asociado al tamaño muestral n.

Una manera de escoger n es minimizando el coste más la pérdida esperada:

$\underset{n}{Min}\{F(n) = C(n) + L(n)\}$

Existen varias funciones de pérdida que se pueden aplicar. Un método habitual es fijar las funciones de pérdida y coste siguientes:

a) $l(z) = \lambda z^2$, con $\lambda = cte$. Esto implica $L(n) = \lambda E(z^2)$.

b) $C(n) = c_0 + c_1 n^\alpha$, con c_0, c_1 y α constantes independientes de n.

En el caso particular de estimación de la media en m.a.s., tenemos

$$z = \widehat{\overline{y}} - \overline{y} \Rightarrow L(n) = \lambda E\left((\widehat{\overline{y}} - \overline{y})^2\right) = \lambda V(\widehat{\overline{y}}) = \lambda\frac{N-n}{N}\frac{S^2}{n}$$

ISBN 9788496866133

La función a minimizar en n es $F(n) = C(n) + L(n) = c_0 + c_1 n^\alpha + \lambda \frac{N-n}{N} \frac{S^2}{n}$

Derivando respecto de n, $F'(n) = c_1 \alpha n^{\alpha-1} - \lambda \frac{S^2}{n^2}$. Al ser la segunda derivada positiva, igualar a cero $f'(n)$ nos dará posibles valores del n donde se minimiza $F(n)$:

$F'(n) = 0 \Rightarrow n^* = \left(\lambda \frac{S^2}{c_1 \alpha}\right)^{\frac{1}{\alpha+1}}$. Para el caso más sencillo de función de coste lineal, $\alpha = 1$, tenemos que $C(n) = c_0 + c_1 n$ y $n^* = \sqrt{\lambda \frac{S^2}{c_1}}$.

Se observa que el n óptimo aumenta con la variabilidad S^2 de y y disminuye con el coste c_1 en que se incurre por cada unidad muestral.

Ejemplo 4.8.

En una población de tamaño $N = 1000$ se desea estimar la media poblacional mediante m.a.s. Consideraremos una función de pérdida cuadrática: $l(z) = z^2$ y como coste fijo del trabajo de campo $c_0 = 10$, con un gasto extra de dos unidades monetarias por cada unidad muestral. Encontrar el tamaño muestral óptimo para minimizar el coste más la pérdida esperada, sabiendo que la cuasivarianza poblacional es $S^2 = 200$.¿Qué precisión se obtiene con el n óptimo?

Tenemos que $l(z) = (\widehat{\overline{y}} - \overline{y})^2$, es decir, $\lambda = 1$.

$C(n) = 10 + 2n$, es decir, $c_1 = 2$ y $\alpha = 1$.

Así, $n^* = \sqrt{\lambda \frac{S^2}{c_1}} = \sqrt{\frac{200}{2}} .= 10$.

La precisión de la estimación que se obtiene es $\frac{1}{V(\widehat{\overline{y}})}$, y como $V(\widehat{\overline{y}}) = \frac{N-n^*}{N} \frac{S^2}{n^*} = \frac{1000-10}{1000} \frac{S^2}{10} =$ 19.8, la precisión obtenida será de $\frac{1}{V(\widehat{\overline{y}})} = \frac{1}{19.8} = 0.05$.

4.5 Ganancia en precisión al aumentar n

Una cuestión importante desde el punto de vista intuitivo es que si N es suficientemente grande, a partir de un cierto valor de n aumentar el tamaño muestral no redunda en una ganancia significativa en precisión, y esa es una de las principales razones que hace que los censos se sustituyan a menudo por estudios basados en el muestreo.

Fijado el error de muestreo en ϕ, el valor de n que da lugar a ese error de muestreo es $n = \frac{1}{\frac{1}{N} + \frac{\phi^2}{S^2}}$.El valor de n necesario es más grande si N aumenta. Si el tamaño poblacional es grande, $1/N \simeq 0$ y $n \simeq \frac{S^2}{\phi^2}$. Es decir, a medida que N se acerca a infinito el valor del tamaño

ISBN 9788496866133

muestral para una precisión ϕ está acotado en $\dfrac{S^2}{\phi^2}$, como se ve en la Figura 4.3.

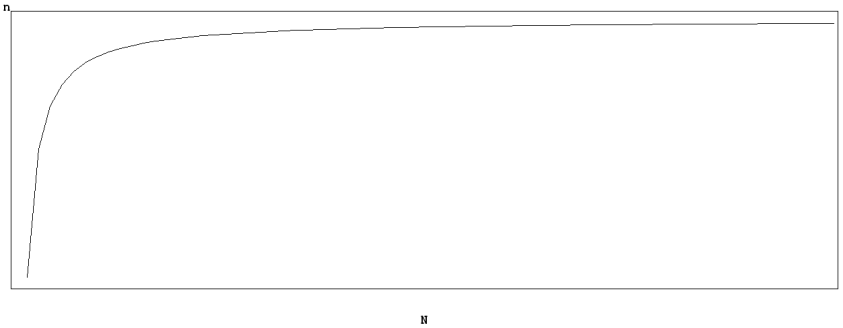


Figura 4.4. n necesario para un error ϕ cuando N aumenta.

De modo que si no hay información suficiente acerca de N, basta calcular la cota superior $\dfrac{S^2}{\phi^2}$ como el valor máximo a tomar para n. Es decir, al ser S^2 fijo, el tamaño muestral necesario para un cierto nivel de error ϕ aumenta aproximadamente proporcionalmente al inverso del cuadrado de ϕ.

Ejemplo 4.9.

A menudo la intuición traiciona a los desconocedores de las nociones básicas de muestreo: muchos pensarán que si la población es más grande, la muestra a tomar debe ser también más grande. Pero esa relación no es tan importante a partir de un tamaño poblacional N suficientemente alto, como se ve en el gráfico anterior. Por ejemplo, una muestra de tamaño $n = 100$ de una población de $N = 100.000$ unidades tiene prácticamente la misma precisión que una muestra de tamaño $n = 100$ de una población de $N = 100$ millones de unidades:

Para $N = 100.000$, $V(\widehat{\overline{y}}) = \dfrac{100.000 - 100}{100.000}\dfrac{S^2}{100} = \dfrac{S^2}{100}(0.999)$.

Para $N = 100.000.000$, $V(\widehat{\overline{y}}) = \dfrac{100.000.000 - 100}{100.000.000}\dfrac{S^2}{100} = \dfrac{S^2}{100}(0.999999)$.

ISBN 9788496866133

4.6 Tablas de fórmulas

MUESTREO ALEATORIO SIMPLE			
Parámetro poblacional	$\overline{y}$	$N\overline{y}$	p
Estimador	$\widehat{\overline{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$	$N\widehat{\overline{y}}$	$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$
Varianza	$\frac{N-n}{N}\frac{S^2}{n}$	$N(N-n)\frac{S^2}{n}$	$\frac{N-n}{N-1}\frac{pq}{n}$
Estimador de la varianza	$\frac{N-n}{N}\frac{s^2}{n}$	$N(N-n)\frac{s^2}{n}$	$\frac{N-n}{N}\frac{\widehat{p}\widehat{q}}{n-1}$

TAMAÑOS MUESTRALES EN m.a.s.		
Errores o costes prefijados	**Tamaño muestral (media)**	**Tamaño muestral (proporción)**
$\phi = \sqrt{V(\widehat{\overline{y}})}$	$n = \frac{N\widehat{S}^2}{N\phi^2 + \widehat{S}^2}$	$\frac{N\widehat{p}(1-\widehat{p})}{\phi^2(N-1) + \widehat{p}(1-\widehat{p})}$
$\phi = \frac{\sqrt{V(\widehat{\overline{y}})}}{\overline{y}}$	$n = \frac{N\widehat{S}^2}{\widehat{S}^2 + N(\phi\widehat{\overline{y}})^2}$	$\frac{N(1-\widehat{p})}{(1-\widehat{p}) + (N-1)\widehat{p}\phi^2}$
$e = z_{\alpha/2}\sqrt{V(\widehat{\overline{y}})}$	$n = \frac{Nz^2_{\alpha/2}\widehat{S}^2}{z^2_{\alpha/2}\widehat{S}^2 + Ne^2}$	$\frac{Nz^2_{\alpha/2}\widehat{p}(1-\widehat{p})}{z^2_{\alpha/2}\widehat{p}(1-\widehat{p}) + (N-1)e^2}$
$C(n) = c_0 + c_1 n^{\alpha}$	$n = \left(\lambda\frac{\widehat{S}^2}{c_1\alpha}\right)^{\frac{1}{\alpha+1}}$	

ISBN 9788496866133

4.7 Obtención de muestras por m.a.s. con SAS

Al igual que se vio al estudiar el muestreo m.a.s.r., se pueden extraer muestras por m.a.s de dos modos: por programación directa y mediante la utilización del procedimiento proc surveyselect . Se verán a continuación las dos maneras:

4.7.1 Mediante programación directa

Supongamos que los datos están en el archivo SAS temporal llamado datos, que tiene N observaciones. Se supone que el archivo contiene una variable de identificación de las unidades elementales, que llamaremos en general ID.

supongamos en lo sucesivo que $N = 50$ y $n = 10$.

El programa de obtención de una muestra por m.a.s.r. consistirá en los pasos vistos en la sección 3.3: ordenación aleatoria del archivo de datos y toma de los n primeros números del archivo. El programa sería:

```
data datos;
  set datos;
  u=ranuni(123456);
run;
proc sort data=datos;run;
data muestra;
  set datos;
  if _n_>10 then stop;
run;
```

A continuación se podría pasar a la fase del trabajo de campo consistente en acceder a las observaciones escogidas y recolectar información en cada una de ellas para pasar al proceso de estimación.

4.7.2 Mediante el Procedimiento Surveyselect

Asumiendo que los datos están en el archivo SAS temporal llamado datos, con $N = 50$ observaciones, y se desea una muestra de $n = 10$ observaciones obtenidas por m.a.s, el programa sería sencillamente:

```
proc surveyselect data=datos out=muestra method=srs n=10;
run;
```

La opción method=srs indica que el tipo de muestreo utilizado es el muestreo aleatorio simple.

ISBN 9788496866133

4.8 Estimación en m.a.s. con SAS

4.8.1 Estimación con el Procedimiento Surveymeans

Estimación de la media o proporción

En este caso se añade el valor de N, necesario para el cálculo de varianzas:

```
proc surveymeans data=muestra N=50;
var y;
run;
```

obteniendo la estimación de la media o proporción, la desviación estándar del estimador y el intervalo de confianza al 95%.

Estimación del total

Si lo que interesa es la estimación del **total**, hay que crear una nueva variable de peso, con valor $\frac{N}{n}$ que en este caso es 5:

```
data muestra;
  set muestra;
  peso=5;
run;
```

Después se procede a ejecutar el proc surveymeans con esta variable de peso:

```
proc surveymeans data=muestra N=50 sum;
weight peso;
run;
```

obteniendo la estimación del total (que aparece como "suma"), desviación estándar del estimador e intervalo de confianza al 95%.

Otra posibilidad es utilizar la macro mas, que se explica a continuación.

4.8.2 Estimación con la macro estimas

La macro mas presenta en la ventana output los estimadores y sus varianzas. Su sintaxis es la siguiente:

```
%estimas(archivo,variable,npobla);
```


ISBN 9788496866133

Donde:

archivo es el archivo de datos SAS que contiene la muestra.

variable es la variable de interés, sobre la cual se desea obtener estimaciones.

npobla es el número de observaciones poblacional N (se utiliza solamente para la estimación del total).

En la ventana OUTPUT aparecen los estimadores para la media (o proporción si e una variable 0-1) y para el total (que aparece como "suma"), junto con sus varianzas, desviaciones típicas (llamadas desviaciones estándar en la salida) e intervalos de confianza.

Una aplicación de esta macro con los números anteriores sería:

```
%estimas(muestra,y,50);
```


ISBN 9788496866133

4.9 Ejercicios resueltos

Ejercicio 3.1.

Un aficionado a los crucigramas tiene a su disposición tres crucigramas, y desea realizar dos al azar por m.a.s., para estimar el tiempo que tarda en promedio en acabar un crucigrama. En realidad, el tiempo que tardaría en hacer los crucigramas 1,2, y 3, sería, por orden: 10 minutos, 15 minutos y 20 minutos.

a) Estudiar la distribución del estimador insesgado del total de tiempo promedio que tarda en realizar un crucigrama. Comprobar que es insesgado.

b) Estudiar la distribución del estimador insesgado del total de tiempo que tardaría en realizar los tres. Comprobar que es insesgado.

c) Estudiar la distribución de s^2 como estimador insesgado de S^2.Comprobar que es insesgado.

d) Estudiar la distribución de $\dfrac{N-n}{N}\dfrac{s^2}{n}$ como estimador insesgado de la varianza del estimador de la media. Comprobar que es insesgado.

---

a) Como se ha visto, en m.a.s. basta analizar las muestras sin tener en cuenta el orden. Para cada muestra se obtendrá el valor del estimador. Por ser sin reemplazamiento y sin tener en cuenta el orden, hay $\binom{3}{2} = 3$ muestras posibles. La tabla de probabilidades es:

Muestra	$\widehat{\overline{y}}$	p
$\{1,2\}$	12.5	1/3
$\{1,3\}$	15	1/3
$\{2,3\}$	17.5	1/3

Obviamente se puede comprobar que $\widehat{\overline{y}}$ es insesgado para $\overline{y}$ pues $E(\widehat{\overline{y}}) = \dfrac{1}{3}(12.5 + 15 + 17.5) = 15 = \overline{y}$.

b) La distribución del total $N\widehat{\overline{y}}$ es 37.5,45, y 52.5 todos con probabilidad 1/3.

c) Hay que calcular s^2 para todas las muestras, con lo que es, respectivamente, s^2 =12.5, 50, 12.5 , todos los valores con probabilidad 1/3. La cuasivarianza poblacional es $S^2 = 25$, con cual se ve que $E(s^2) = \dfrac{1}{3}(12.5 + 50 + 12.5) = 25$ y por lo tanto s^2 es insesgado para S^2.

d) Para cada muestra, se obtiene el estimador de la varianza de $\widehat{\overline{y}}$: $\widehat{V}(\widehat{\overline{y}}) = \dfrac{N-n}{N}\dfrac{s^2}{n}$ = 2.08, 8.33 y 2.08 respectivamente, todos con probabilidad 1/3. La varianza real del estimador es $V(\widehat{\overline{y}}) = \dfrac{N-n}{N}\dfrac{S^2}{n} = 4.166$, con lo que se observa que $E(\widehat{V}(\widehat{\overline{y}})) = \dfrac{1}{3}(2.08 + 8.33 + 2.08) = 4.166$ y por lo tanto $\widehat{V}(\widehat{\overline{y}})$ es insesgado para $V(\widehat{\overline{y}})$.

ISBN 9788496866133

Ejercicio 3.2.

Se pretende estudiar el número de personas que acuden los miércoles a servicios administrativos del ayuntamiento en cierta comunidad. Se escogen por m.a.s. 10 ventanillas de las 100 que hay y se cuenta el número de personas que acuden a ella un miércoles. La distribución de frecuencias viene dada en la tabla siguiente, donde "Frecuencia" indica el número de ventanillas donde se ha dado ese resultado:

Número de personas	Frecuencia
70	1
65	3
92	1
87	1
45	2
30	2

a) Dar un intervalo de confianza a 95% suponiendo normalidad para la estimación del número medio de personas por ventanilla.

b) Con los datos obtenidos, dar el número de ventanillas que habría que muestrear para obtener un error de muestreo de 4 unidades. Dar el número de ventanillas necesario para obtener un error de muestreo relativo de 0.08. Dar el número de ventanillas necesario para obtener un error de muestreo absoluto de 8 suponiendo normalidad y $\alpha = 0.05$.

c) ¿En el caso del apartado a), que diferencia habría en la varianza estimada del estimador si en lugar de m.a.s. se hubiera utilizado m.a.s.r.?.

---

a) La media muestral es $\widehat{\overline{y}} = \frac{1}{10}(70 + 3 \cdot 65 + 92 + 87 + 2 \cdot 45 + 2 \cdot 30) = 59.4$.

Además, $s^2 = \frac{1}{n-1}[\sum_{i=1}^{n} y_i^2 - n\widehat{\overline{y}}^2] = 463.82$. Por lo tanto, la varianza estimada de $\widehat{\overline{y}}$ es $\widehat{V}(\widehat{\overline{y}}) = \frac{N-n}{N}\frac{s^2}{n} = 41.74$.

El intervalo de confianza para la media será $(\widehat{\overline{y}} - 1.96\sqrt{\widehat{V}(\widehat{\overline{y}})}, \widehat{\overline{y}} + 1.96\sqrt{\widehat{V}(\widehat{\overline{y}})}) = (46.7, 72)$.

b) Recurriendo a las fórmulas, se tiene que, para el error de muestreo,

$$n^* = \frac{N\widehat{S}^2}{N\phi^2 + \widehat{S}^2} = \frac{Ns^2}{N\phi^2 + s^2} = \frac{100 \cdot 463.82}{100 \cdot 4^2 + 463.82} = 22.47.$$

Con $n^* = 23$ se satisfaría ese nivel de error.

Para el error de muestreo relativo:

ISBN 9788496866133

$$n^{**} = \frac{N\widehat{S}^2}{\widehat{S}^2 + N(\phi\widehat{\overline{y}})^2} = \frac{100 \cdot 463.82}{463.82 + 100 \cdot 0.06^2 \cdot 59.4^2} = 26.7$$

Con $n^{**} = 27$ se satisfaría ese nivel de error.

En el caso del error de muestreo absoluto:

$$n^{***} = \frac{Nz^2_{\alpha/2}\widehat{S}^2}{z^2_{\alpha/2}\widehat{S}^2 + Ne^2} = \frac{100 \cdot 1.96^2 \cdot 463.82}{1.96^2 \cdot 463.82 + 100 \cdot 8^2} = 21.7$$

Con lo que $n^{***} = 22$.

c) Si se hubiera utilizado m.a.s.r. y obtenido los mismos datos, el estimador de la media poblacional sería el mismo, pero se sabe que la varianza del estimador en muestreo aleatorio simple con reemplazamiento tiene que ser mayor que la obtenida en muestreo sin reemplazamiento. En este caso, estimaríamos la varianza del estimador por $\frac{s^2}{n} = \frac{463.82}{190} = 46.82$ en lugar de por $\frac{N-n}{N}\frac{s^2}{n} = 41.74$.

Ejercicio 3.3.

En un hospital hay 230 habitaciones con 345 camas, y se realiza un m.a.s. de tamaño $n = 8$ habitaciones para estudiar la ocupación promedio en un momento dado del día. Se obtienen los datos:

Número de camas	Número de camas ocupadas
2	1
2	2
1	1
2	0
3	2
2	1
2	2
3	1
1	0
2	2

a) Estimar el número de camas ocupadas por habitación y dar un intervalo de confianza al 95% suponiendo normalidad.

b) Estimar el total de camas ocupadas en el hospital y dar un intervalo de confianza al 95%.

ISBN 9788496866133

c) Estimar la proporción de camas ocupadas en el hospital y dar un intervalo de confianza al 95%.

d) Suponiendo que las estimaciones del apartado a) son buenas aproximaciones, ¿Cuántas habitaciones habrá que examinar si se requiere un coeficiente de variación del estimador de la media de camas 0.10?

---

a) Se tiene $\widehat{\overline{y}} = 1.2$ y $s^2 = 0.622$. Por lo tanto, $\widehat{V}(\widehat{\overline{y}}) = \dfrac{N-n}{N}\dfrac{s^2}{n} = \dfrac{230-10}{230}\dfrac{0.62}{10} = 0.0595$.

El intervalo de confianza al 95% para la media de camas ocupadas por habitación en ese momento del día en el hospital será $(\widehat{\overline{y}} - 1.96\sqrt{\widehat{V}(\widehat{\overline{y}})}, \widehat{\overline{y}} + 1.96\sqrt{\widehat{V}(\widehat{\overline{y}})}) = (0.72, 1.68)$.

b) Al ser $N = 230$, el estimador del total es $N\widehat{\overline{y}} = 276$ camas ocupadas, su varianza es $N^2\ \widehat{V}(\widehat{\overline{y}}) = 3147.55$ y el intervalo de confianza es $(165.6, 386.4)$.

c) Al conocer el número de camas en total y ser este constante, la proporción de camas ocupadas es el número de camas ocupadas entre el total de camas. Se estimará por $\dfrac{N\widehat{\overline{y}}}{total\ camas} = \dfrac{276}{345} = 0.80$ (un 80% de camas ocupadas).

Para estimar la varianza de este estimador, al ser el total de camas una constante, se tiene que

$$\widehat{V}\left(\frac{N\widehat{\overline{y}}}{total\ camas}\right) = \frac{1}{total\ camas^2}\ \widehat{V}(N\widehat{\overline{y}}) = \frac{1}{345^2}\ 3147.55 = 0.02644.$$

El intervalo de confianza será $(0.80 - 1.96\sqrt{0.02644}, 0.80 + 1.96\sqrt{0.02644}) = (0.637, 0.962)$.

Hay que señalar que sería incorrecto (sería una estimación, pero en general peor) realizar la estimación hallando la proporción de camas ocupadas en cada habitación de la muestra, y después la media de estas proporciones. Como el número de camas en cada habitación es una variable aleatoria, estaríamos construyendo el estimador $\sum_{i=1}^{n} \dfrac{y_i}{x_i}$ donde y_i = número de camas ocupadas en la habitación i, y x_i =número de camas en la habitación i.

Este estimador es el estimador de los momentos de la "razón media" poblacional, que es la media de las proporciones de camas ocupadas por habitación $\sum_{i=1}^{N} \dfrac{y_i}{x_i}$. Pero no es un estimador correcto de la proporción de camas ocupadas en el hospital, que es exactamente $\dfrac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$ y se suele llamar "razón" poblacional.

Otro estimador posible, sería estimar la proporción de camas ocupadas por la razón muestral $\dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$. Este estimador puede tener menor varianza que el mostrado en la resolución $\dfrac{N\widehat{\overline{y}}}{total\ camas}$, si la correlación entre x e y es alta. El lector puede acudir a la teoría sobre estimación indirecta (estimación de razón) para más explicaciones sobre este tema.

d) En realidad se nos está pidiendo n para obtener un error de muestreo relativo de

$\phi = \dfrac{\sqrt{V(\widehat{\overline{y}})}}{\overline{y}}$ =Coeficiente de variación del estimador=0.10.

El tamaño necesario es

ISBN 9788496866133

$$n^* = \frac{N\widehat{S}^2}{\widehat{S}^2 + N(\phi\overline{\widehat{y}})^2} = \frac{230 \cdot 0.622}{0.622 + 230(0.10 \cdot 1.2)^2} = 36.36.$$

Por lo tanto haría falta examinar $n^* = 37$ camas.

Ejercicio 3.4.

Una empresa de auditoría desea estudiar la proporción de facturas no declaradas a Hacienda en un acta de contabilidad. El acta contiene 525 facturas, y se toma una m.a.s. de 20 facturas. De ellas, 8 son no declaradas.

a) Estimar la proporción de facturas no declaradas y calcular la desviación típica del estimador.

b) ¿Cuál es la fracción de muestreo? ¿Cual es el coeficiente de corrección por población finita? ¿Cuál es el error de muestreo aproximado? ¿Cual es el error de muestreo relativo aproximado? ¿Cual es el error de muestreo absoluto aproximado suponiendo $\alpha = 0.01$?

c) Se suele definir el "efecto de diseño" como el cociente entre la varianza del estimador bajo un cierto tipo de diseño, entre la varianza del estimador básico utilizando m.a.s. Calcular el efecto de diseño de utilizar m.a.s.r. en este caso.

d) Si la estimación de la proporción fuera exacta, y no utilizamos la aproximación normal, ¿cuál sería la probabilidad de, realizando un m.a.s. de 200 facturas, encontrarse con 100 o más de 100 no declaradas? (hay que utilizar la distribución hipergeométrica). Utilizando la aproximación normal, responder a la misma pregunta. Utilizar tablas o un paquete estadístico para responder a estas preguntas.

---

a) La proporción muestral es $\widehat{p} = \frac{8}{20} = 0.40$. La varianza estimada del estimador es $\widehat{V}(\widehat{p}) = \frac{N-n}{N}\frac{\widehat{p}\widehat{q}}{n-1} = \frac{525-20}{525}\frac{0.4 \cdot 0.6}{20-1} = 0.01215$. Por lo tanto la desviación típica es $\sqrt{0.01215} = 0.11$.

b)

La fracción de muestreo es $f = \frac{n}{N} = \frac{20}{1500} = 0.01333$.

El coeficiente de corrección por población finita es $\frac{N-n}{N} = 1 - f = 0.9866$.

El error de muestreo aproximado es la desviación típica estimada del estimador (es aproximado pues se trata de una estimación, no del error real) $\sqrt{\widehat{V}(\widehat{p})} = 0.11$.

El error de muestreo relativo aproximado es $\frac{\sqrt{\widehat{V}(\widehat{p})}}{\widehat{p}} = 0.275$.

El error de muestreo absoluto aproximado suponiendo $\alpha = 0.01$ es $z_{0.005}\sqrt{\widehat{V}(\widehat{p})} = 2.57\sqrt{\widehat{V}(\widehat{p})} = 0.2827$.

c) La varianza obtenida en m.a.s.r. sería la misma exceptuando el coeficiente de corrección por población finita: $\widehat{V}(\widehat{p}) = \frac{\widehat{p}\widehat{q}}{n-1} = 0.012315$. Por lo tanto el "efecto de diseño" al utilizar m.a.s.r. se estima en el cociente de las varianzas en m.a.s.r. y en m.a.s, que es $\frac{0.012315}{0.01215} = 1.013582$.

Obviamente,cuando el efecto de diseño es mayor que uno, es que el método tiene mayor varianza que el muestreo aleatorio simple y es en teoría peor.

ISBN 9788496866133

d) Supongamos entonces que $p = 0.4$ y por lo tanto hay $0.4 \cdot 1500 = 600$ facturas no declaradas, y 900 declaradas. Si se extraen 200 por m.a.s., la probabilidad de encontrarse, por ejemplo, con 100 no declaradas es, al ser todas las muestras equiprobables, el número de muestras de tamaño 200 que podrían contener 100 no declaradas, entre el número total de muestras. Esto es:

$$\frac{\binom{600}{100}\binom{900}{100}}{\binom{1500}{200}}$$

pues hay $\binom{600}{100}$ maneras de tomar sin reemplazo 100 facturas de las 600 no declaradas, y por cada una de éstas, $\binom{900}{100}$ maneras de tomar 100 facturas de las declaradas. El denominador es el número de muestras de tamaño 200 que se pueden tomar de las 1500 facturas.

Para calcular la probabilidad de que al menos haya 100 no declaradas habrían que sumar, considerando el razonamiento anterior, los términos

$$\frac{\binom{600}{100}\binom{900}{100}}{\binom{1500}{200}} + \frac{\binom{600}{101}\binom{900}{99}}{\binom{1500}{200}} + ... + \frac{\binom{600}{200}\binom{900}{0}}{\binom{1500}{200}}.$$

Se ve que el número de facturas no declaradas en la muestra de tamaño 200 sigue una distribución hipergeométrica, por lo cual sólo es necesario calcular la probabilidad con la ayuda de tablas o paquetes estadísticos. En particular, en SAS se utiliza la función probhypr(N,K,n,x) donde N es el tamaño de la población, K el de la subpoblación de items con la cualidad, n el tamaño muestral y x el número de items con la cualidad sobre el que se quiere calcular la probabilidad. La función devuelve la probabilidad de que haya, en una muestra de tamaño n, x o menos items con la cualidad. En nuestro caso se utiliza 1-p para que devuelva la probabilidad de 100 o más items.

```
data;
proba=1-probhypr(1500,600,200,100);
put proba=;
run;
```

obteniendo un resultado de *proba* =0.0008062602.

Al ser la población finita, se ha utilizado la distribución hipergeométrica. Si se considera infinita (a partir de N suficientemente grande), se utilizaría la distribución binomial, y además, a partir de un np grande, la aproximación normal.

La aproximación normal tiene en cuenta que el número de unidades con la cualidad obtenidas en una muestra de tamaño n, y sabiendo que la probabilidad de obtener la cualidad en un item es p, sigue una normal $N(\mu, \sigma^2) = N(np, np(1-p))$. Por lo tanto basta calcular la probabilidad de $X > 100$ bajo esa distribución normal.

Con el SAS, se utiliza la función probnorm(x)=$P(Z < x)$, con Z una variable Normal (0.1). Sabiendo que $P(X > 100) = P(Z > \frac{100-\mu}{\sigma})$ donde Z es una variable distribuida como Normal (0.1), hay que calcular 1-probnorm($\frac{100-\mu}{\sigma}$).

```
data;
u=(100-200*0.4)/((200*0.4*0.6)**0.5);
proba=1-pronorm(u);
put proba=;
run;
```

obteniendo proba=0.001946208.

ISBN 9788496866133

Ejercicio 3.5.

Una empresa inmobiliaria desea estimar el precio promedio de los pisos en venta de una zona. Para ello escoge por m.a.s. 15 de los 100 pisos que están en venta. Obtiene una media muestral del precio de 17.000 euros, con una cuasivarianza muestral de 625 millones. Tras esta operación, que se utilizará como muestra piloto, se desea realizar otro estudio de manera que con un grado de confianza del 95% la estimación esté dentro del 10% del precio promedio real. ¿Cuántos pisos deberá muestrear por m.a.s. en una segunda encuesta?.

Para que la estimación diste del precio real un 10% ha de ser

$P(\mid \widehat{\overline{y}} - \overline{y} \mid < 0.10\overline{y}) = 0.95.$

Suponiendo normalidad del estimador, y como la desviación típica del estimador es $\sqrt{\frac{N-n}{N}\frac{S^2}{n}}$,

se tiene que

$$P(\frac{\mid \widehat{\overline{y}} - \overline{y} \mid}{\sqrt{\frac{N-n}{N}\frac{S^2}{n}}} < \frac{0.10\overline{y}}{\sqrt{\frac{N-n}{N}\frac{S^2}{n}}}) = 0.95$$

es decir,

$$P(-\frac{0.10\overline{y}}{\sqrt{\frac{N-n}{N}\frac{S^2}{n}}} < \frac{(\widehat{\overline{y}} - \overline{y})}{\sqrt{\frac{N-n}{N}\frac{S^2}{n}}} < \frac{0.10\overline{y}}{\sqrt{\frac{N-n}{N}\frac{S^2}{n}}}) = 0.95$$

Como el estimador es insesgado, $\widehat{\overline{y}} \equiv N(\overline{y}, \sqrt{\frac{N-n}{N}\frac{S^2}{n}})$ y por lo tanto el término central es $Z \equiv N(0,1)$. Entonces,

$$P(-\frac{0.10\overline{y}}{\sqrt{\frac{N-n}{N}\frac{S^2}{n}}} < Z < \frac{0.10\overline{y}}{\sqrt{\frac{N-n}{N}\frac{S^2}{n}}}) = 0.95$$

Lo que implica que $\frac{0.10\overline{y}}{\sqrt{\frac{N-n}{N}\frac{S^2}{n}}} = z_{\alpha/2}$, con $\alpha = 0.05$.

Por lo tanto,

$\frac{0.10\overline{y}}{\sqrt{\frac{N-n}{N}\frac{S^2}{n}}} = 1.96$ y

$\frac{0.10}{1.96} = \frac{\sqrt{\frac{N-n}{N}\frac{S^2}{n}}}{\overline{y}}$ lo que es lo mismo que decir que hay que encontrar el tamaño muestral n tal que el error de muestreo relativo sea 0.10. Con la fórmula habitual, se tiene que

$$n^* = \frac{N\widehat{S}^2}{\widehat{S}^2 + N(\phi\widehat{\overline{y}})^2} = \frac{100 \cdot 625000000}{625000000 + 100 \cdot (0.10 \cdot 17000)^2} = 68.3.$$

Por lo tanto habría que muestrear $n = 69$ pisos.

ISBN 9788496866133

Ejercicio 3.6.

En un estudio agrícola en el área A se pretende estimar el número de árboles que han comenzado a dar frutos antes de determinada fecha.

a) Antes de comenzar el estudio no se tiene ninguna información, salvo el número de árboles que es $N = 100$. ¿Qué tamaño muestral se les recomendaría tomar si se pretende escoger los árboles por muestreo aleatorio simple de modo que la semianchura del intervalo de confianza al 95% sea como mucho 5?

b) Supongamos que se sabe por estudios realizados en un pueblo cercano que el verdadero porcentaje de árboles que dan frutos prematuramente está entre 0.2 y 0.35. ¿Qué tamaño muestral se recomendaría en este caso?.

c) Se ha realizado finalmente el estudio con el tamaño muestral obtenido en el apartado b) y se ha obtenido una proporción muestral de árboles que han fructificado antes de tiempo de 0.25. Establecer un intervalo de confianza para la cantidad "número de árboles que han fructificado en el área A".

---

a) En el caso de estudio de proporciones, la varianza del estimador es máxima si $p = q = 0.5$. Por lo tanto, en ausencia de información se suele tomar el valor $p = 0.5$ por defecto, porque dará lugar a un n máximo (es situarse en el caso peor).

En este caso, si $p = q = 0.5$, la semianchura del intervalo de confianza es $1.96\sqrt{N^2 \cdot \frac{N-n}{N-1}\frac{pq}{n}} = 5$ (recordemos que se trata de estimar el total). Haciendo $1.96\sqrt{\frac{N-n}{N-1}\frac{pq}{n}} = 5/N = 0.05$ queda transformada a la expresión usual de un error de muestreo absoluto para la proporción.

Despejar n equivale a aplicar la fórmula:

$$n = \frac{N z^2_{\alpha/2}\widehat{p}(1-\widehat{p})}{z^2_{\alpha/2}\widehat{p}(1-\widehat{p}) + (N-1)e^2} = \frac{100 \cdot 1.96^2 \cdot 0.5 \cdot 0.5}{1.96^2 \cdot 0.5 \cdot 0.5 + 99 \cdot (0.05)^2} = 80 \text{ árboles.}$$

b) La varianza del estimador, como función de p, es una parábola creciente desde $p = 0$ hasta $p = 0.5$ y decreciente hasta $p = 1$ con máximo en $p = 0.5$. Por lo tanto, es peor (daría una varianza más alta) el caso $p = 0.35$ que $p = 0.20$. Situándose en el caso peor $p = 0.35$, el tamaño muestral necesario, aplicando la fórmula:

$$n = \frac{N z^2_{\alpha/2}\widehat{p}(1-\widehat{p})}{z^2_{\alpha/2}\widehat{p}(1-\widehat{p}) + (N-1)e^2} = \frac{100 \cdot 1.96^2 \cdot 0.35 \cdot 0.65}{1.96^2 \cdot 0.35 \cdot 0.65 + 99 \cdot (0.05)^2} = 77.9 \simeq 78 \text{ árboles.}$$

c) La varianza estimada del estimador es $\widehat{V}(\widehat{p}) = \frac{N-n}{N}\frac{\widehat{p}\widehat{q}}{n-1} = \frac{100-78}{100}\frac{0.25 \cdot 0.75}{78-1} = 0.000535$.

el intervalo de confianza será $(0.25 - 1.96\sqrt{0.000535}, 0.25 + 1.96\sqrt{0.000535}) = (0.20, 0.29)$.

Ejercicio 3.7.

En un estudio de mercado se desea estimar la media poblacional de ingresos en la venta de cierto artículo en miles de euros en las 200 tiendas de una región, de modo que el coeficiente de variación del estimador no supere el 5%. Por anteriores estudios, se conoce que el coeficiente de variación de la variable ingresos era de 0.35.

a) ¿Cuál es el tamaño muestral que se debe tomar si se desea realizar el muestreo por m.a.s.?

ISBN 9788496866133

b) ¿Cuál es el tamaño muestral que se debe tomar si se desea realizar el muestreo por m.a.s.r.?

---

a) Se desea hallar n tal que $\phi = \dfrac{\sqrt{V(\widehat{\overline{y}})}}{\overline{y}} = 0.05$. De modo que despejando n se accede a la fórmula $n = \dfrac{N\widehat{S}^2}{\widehat{S}^2 + N(\phi\widehat{\overline{y}})^2}$.

El Coeficiente de Variación de y es

$CV(y) = \dfrac{\sigma}{\overline{y}} = \dfrac{N-1}{N}\dfrac{S_y}{\overline{y}}$, con lo que

$\dfrac{S_y}{\overline{y}} = \dfrac{N}{N-1}CV(y)$.

Dividiendo por $\widehat{\overline{y}}^2$ en numerador y denominador en la expresión de n, queda:

$$n = \frac{N\widehat{S}^2}{\widehat{S}^2 + N(\phi\widehat{\overline{y}})^2} = \frac{N(\frac{S_y}{\overline{y}})^2}{(\frac{S_y}{\overline{y}})^2 + N\phi^2} = \frac{N(\frac{N}{N-1}CV(y))^2}{(\frac{N}{N-1}CV(y))^2 + N\phi^2}.$$

Como se sabe que $CV(y)$ es aproximadamente 0.35, sustituyendo queda

$$n = \frac{200(\frac{200}{200-1}0.35)^2}{(\frac{200}{200-1}0.35)^2 + 200\cdot 0.05^2} = 14.24$$ con lo que se tomará $n = 15$.

b) En caso de muestreo con reemplazamiento, se utiliza la expresión para n:

$$n = \frac{\widehat{\sigma}^2}{\phi^2\widehat{\overline{y}}^2} = \frac{CV(y)^2}{\phi^2} = \frac{0.35^2}{0.05^2} = 49.$$

Ejercicio 3.8.

Se dispone de los siguientes datos obtenidos de un estudio realizado en 10 restaurantes escogidos por m.a.s. en un determinado municipio que tiene 130 restaurantes. Se anotó, en promedio, el tiempo en minutos que tarda en servirse el primer plato en una comida de Menú del día. También se anotó si el restaurante merecía la calificación A (más calidad) o B (calidad media), C (calidad baja).

ISBN 9788496866133

Tiempo en servir 1º plato	Calidad
5	A
10	B
20	B
5	A
3	B
8	B
4	B
13	C
7	A
6	B

a) Estimar y dar un intervalo de confianza al 95% para el número de restaurantes que tienen la calidad A.

b) Estimar el tiempo promedio en servir el primer plato en los restaurantes del municipio.

c) Si quisiéramos plantear un intervalo de confianza para el tiempo promedio en servir el primer plato en los restaurantes de calidad C, ¿qué problemas habría?.

d) Suponiendo que la estimación dada en el apartado a) fuera cierta, presentar un intervalo aproximado de confianza al 95% para el tiempo promedio en servir el primer plato en los restaurantes de calidad A.

e) ¿Cuántos restaurantes habría que muestrear por m.a.s. para tener una semianchura del intervalo de confianza al 95% de como máximo 8 restaurantes en la estimación del apartado a)?

---

a) La estimación del total es $N\widehat{p} = 130\frac{3}{10} = 39$, y $\widehat{p} = 0.30$. La varianza estimada del estimador será

$$N^2\widehat{V}(\widehat{p}) = N^2\frac{N-n}{N}\frac{\widehat{p}\widehat{q}}{n-1} = 130^2\frac{130-10}{130}\frac{0.30\cdot 0.70}{10-1} = 364.$$

Por lo tanto un intervalo de confianza suponiendo normalidad sería

$$(39 - 1.96\sqrt{364}, 39 + 1.96\sqrt{364}) = (1.6, 76.4).$$

Hay que observar que el extremo inferior se puede aumentar, pues se sabe que al menos hay 3 restaurantes tipo A (los que han aparecido en la muestra). El intervalo es muy ancho debido al pequeño tamaño muestral y a la variabilidad de la cualidad (al estar p cerca de 0.5, el caso con más variabilidad).

b) La media muestral es $\widehat{\overline{y}} = 8.1$ y la cuasivarianza $s^2 = 26.32$. La varianza del estimador será

ISBN 9788496866133

$$\widehat{V}(\widehat{\overline{y}}) = \frac{N-n}{N}\frac{s^2}{n} = \frac{130-10}{130}\frac{26.32}{10} = 2.43.$$

El intervalo de confianza al 95% asociado, suponiendo normalidad, es $(5.04, 11.15)$.

c) Al haber solamente una observación, la varianza del estimador no se puede estimar. aunque sí se puede dar una estimación del tiempo promedio en los restaurantes tipo C, que es 13 (la única observación y por lo tanto la media muestral).

d) Si se supone la estimación del apartado a) cierta, entonces el número de restaurantes tipo A es $N_A = 39$. La media muestral del tiempo promedio en los restaurantes tipo A de la muestra es $\widehat{\overline{y}} = 5.66$ y $s^2 = 1.33$. La varianza del estimador se puede estimar en

$$\widehat{V}(\widehat{\overline{y}}) = \frac{N_A - n_A}{N_A}\frac{s^2}{n_A} = \frac{39-3}{39}\frac{1.33}{3} = 0.409.$$

El intervalo de confianza al 95% aproximado será

$$(5.66 - 1.96\sqrt{0.409}, 5.66 + 1.96\sqrt{0.409}) = (4.4, 6.91).$$

e) Se transformará la fórmula de la proporción, sabiendo que, cuando se estudia la proporción y

$$e = z_{\alpha/2}\sqrt{V(\widehat{p})},$$

se tiene

$$n = \frac{N z^2_{\alpha/2}\widehat{p}(1-\widehat{p})}{z^2_{\alpha/2}\widehat{p}(1-\widehat{p}) + (N-1)e^2}$$

Pero en este caso, al estudiar el total de clase, el error absoluto de muestreo es

$$e = z_{\alpha/2}\sqrt{N^2 V(\widehat{p})} = N z_{\alpha/2}\sqrt{V(\widehat{p})},$$

con lo cual basta sustituir en la fórmula $z_{\alpha/2}$ por $N z_{\alpha/2}$. Así,

$$n = \frac{N^3 z^2_{\alpha/2}\widehat{p}(1-\widehat{p})}{N^2 z^2_{\alpha/2}\widehat{p}(1-\widehat{p}) + (N-1)e^2} = \frac{130^3 1.96^2 0.30\cdot 0.70}{130^2 1.96^2 0.30\cdot 0.70 + (130-1)8^2} = 81.$$

Ejercicio 3.9.

Se realizó un estudio sobre la afluencia de hormigas en hogares de una comunidad de 1000 viviendas. Se extrajo una m.a.s. de 10 viviendas y se contó el número de viviendas que tenían un problema con las hormigas. Resultaron salir 4 viviendas con este problema.

a) Construir un intervalo de confianza para la proporción de viviendas afectadas. Suponiendo que la muestra fuera de tamaño $n = 50$ en lugar de tamaño $n = 10$ y saliera el mismo porcentaje de viviendas afectadas en la muestra, ¿cómo varía la semianchura del intervalo de confianza?.

b) Suponiendo despreciable el término de corrección por población finita, responder ahora al apartado a).

c) Supongamos que el coste de evaluar las viviendas se cifra en una cantidad fija $c_0 = 3$ más una cantidad que depende del tamaño muestral $c_1 \cdot \sqrt{n}$, con $c_1 = 2$. Si se pretende minimizar la función "coste más 200 veces la varianza del estimador", ¿que cantidad de viviendas habría que examinar? Responder a la pregunta utilizando la fórmula habitual con $\lambda = 20$ y también programando en SAS la función coste más varianza aumentando n paulatinamente.

ISBN 9788496866133

a) $\widehat{p} = 0.4$,.con lo cual $\widehat{V}(\widehat{p}) = \dfrac{100-10}{100}\dfrac{0.4\cdot 0.6}{9} = 0.024$, con lo que $1.96\cdot\sqrt{\widehat{V}(\widehat{p})} = 0.30$ y el intervalo es $(0.096, 0.70)$.

Si la muestra fuera de tamaño $n = 100$, sería $\widehat{V}(\widehat{p}) = \dfrac{100-50}{100}\dfrac{0.4\cdot 0.6}{49} = 0.00244$ y $1.96\cdot\sqrt{\widehat{V}(\widehat{p})} = 0.0968$, mucho más pequeño.

b) En el primer caso sería $\widehat{V}(\widehat{p}) \simeq 0.0266$ y $1.96\cdot\sqrt{\widehat{V}(\widehat{p})} = 0.32$ y en el segundo, $\widehat{V}(\widehat{p}) = 0.00489$ y $1.96\cdot\sqrt{\widehat{V}(\widehat{p})} = 0.137$.

El término de corrección por población finita en el primer caso era 0.90 y en el segundo 0.50, por lo que afecta proporcionalmente más en el segundo caso la eliminación.

c) Como habitualmente se realiza en el caso de proporciones, se sustituye en la fórmula el término utilizado para medias $\widehat{S}^2$ por $\dfrac{N}{N-1}\widehat{p}(1-\widehat{p})$. Además, la función de coste es

$C(n) = c_0 + c_1\sqrt{n}$, con lo que $\alpha = 1/2$.

La función a minimizar es por lo tanto

$f(n) = c_0 + c_1\sqrt{n} + 200V(\widehat{p})$.

Utilizando la fórmula,

$$n = \left(\lambda\frac{\widehat{S}^2}{c_1\alpha}\right)^{\frac{1}{\alpha+1}} = \left(\lambda\frac{\frac{N}{N-1}\widehat{p}(1-\widehat{p})}{c_1\alpha}\right)^{\frac{1}{\alpha+1}} = \left(200\cdot\frac{\frac{100}{100-1}0.4\cdot 0.6}{2\cdot\frac{1}{2}}\right)^{\frac{2}{3}} = 13.3.$$

Habría que examinar $n = 13$ o $n = 14$ viviendas.

Con SAS, basta hacer un bucle en un paso data, calculando el valor de la función cada vez:

```
data uno;
put @5 'n' @20 'funcion' /;
do n=2 to 20;
f=n/100;
coste=3+2*(n**0.5);
var=(1-f)*0.24/n;
fun=coste+200*var;
put @5 n @20 fun;
output;
end;
run;
```

Se obtiene:

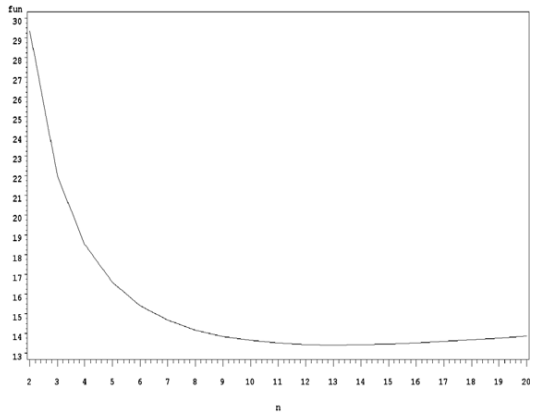
```
n               funcion
2               29.348427125
3               21.984101615
4               18.52
5               16.592135955
6               15.418979486
7               14.668645479
8               14.176854249
9               13.853333333
```


ISBN 9788496866133

```
10              13.64455532
11              13.516885944
12              13.44820323
13              13.423410243
14              13.431886202
15              13.465966692
16              13.52
17              13.589740663
18              13.671948041
19              13.764113677
20              13.86427191
```

Se observa cómo el valor de la función es mínimo para $n = 13$. Se puede hacer un gráfico para ver la función minimizada (se ha añadido la sentencia output en el bucle para guardar todas las observaciones):

```
symbol i=join;
proc gplot data=uno;plot fun*n;run;
```



Ejercicio 3.10.

En un estudio sobre hábitos de audiencia de televisión, se realizó un m.a.s. sobre la población de niños de un colegio de 1500 niños. Se obtuvieron 100 encuestas, de las cuales se obtuvo la siguiente información sobre el número de horas que veían la televisión al día (cada niño encuestado respondía con una estimación de las horas promedio que veía al día la televisión). Se obtuvieron los siguientes resultados: media muestral: 1.5 horas. Cuasivarianza muestral: 0.64.

a) Obtener un intervalo de confianza al 95% para la media poblacional de horas que pasan al día viendo la televisión los niños de ese colegio.

b) Realizar un programa SAS para presentar los errores de muestreo aproximados (normal, relativo y absoluto con $\alpha = 0.05$) aumentando n primero de 1 en uno de $n = 2$ hasta $n = 20$, luego de 100 en 100 a partir de $n = 100$. Hacer la misma tabla, suponiendo m.a.s.r. Presentar un gráfico del error de muestreo en m.a.s. y otro en m.a.s.r. a medida que aumenta n.

c) Observando la tabla de m.a.s. del apartado b), ¿a partir de qué n se tiene un error de muestreo de 0.20? Corroborarlo con las fórmulas habituales de determinación del tamaño muestral.

ISBN 9788496866133

a) Se tiene que $\widehat{\overline{y}} = 1.5$ es el estimador insesgado de la media poblacional. Además,

$$\widehat{V}(\widehat{\overline{y}}) = \frac{N-n}{N}\frac{s^2}{n} = \frac{1500-100}{1500}\frac{0.64}{100} = 0.0059733.$$

Con lo cual el intervalo de confianza será $(1.5 - 1.96\sqrt{0.0059733}, 1.5 + 1.96\sqrt{0.0059733}) = (1.348, 1.651)$.

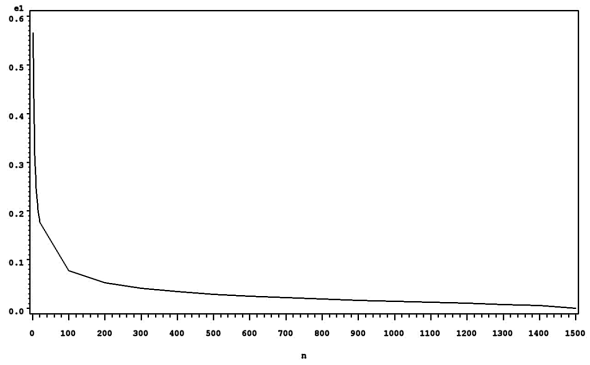
b) Se recurrirá al programa del ejercicio 2.1, cambiando los parámetros y ligeramente las fórmulas :

```
data uno;
s2=0.64;ngrande=1500;ymedia=1.5;
put 'n' @10 'Error' @25 'E.relativo' @40 'E.absoluto' /;
do n=2 to 20,100 to 1500 by 100;
  var=s2/n*(1-n/ngrande);
  e1=var**0.5;
  e2=var/1.5;
  e3=1.96*(var**0.5);
  put n @10 e1 6.4 @25 e2 6.4 @40 e3 6.4;
  output;
end;
run;
```

En la ventana LOG se observa cómo desciende el error de muestreo, que para $n = 1500$ es cero (si aparece 0 en algunas observaciones antes en el error de muestreo relativo es debido al formato de salida: se pueden cambiar los 6.4 por 10.8 para ver más decimales).

Se ha puesto la sentencia output; dentro del bucle para guardar las observaciones en el archivo uno y poder presentar el gráfico. Éste se realiza con la opción symbol i=join; para unir los puntos por líneas.

```
symbol i=join;
proc gplot data=uno;plot e1*n;
run;
```



Para m.a.s.r el programa es igual, cambiando var= var=s2/n*(1-n/ngrande); por var=s2/n;.

c) Se observa que habría que tomar como mínimo $n = 16$ para obtener un error de muestreo de 0.20. Si recurrimos a la fórmula:

$$n^* = \frac{N\widehat{S}^2}{N\phi^2 + \widehat{S}^2} = \frac{1500 \cdot 0.64}{1500 \cdot 0.20^2 + 0.64} = 15.8$$ dando lugar a un n mínimo de 16.

ISBN 9788496866133

Ejercicio 3.11.

El archivo SAS madrid, con 179 observaciones y utilizado en el ejercicio 2.7, contiene datos sobre edificios por construir en municipios en la Comunidad de Madrid en 1998. Se trata de comprobar las diferencias existentes al tomar muestras sin reemplazamiento frente a muestras con reemplazamiento.

a) En el ejercicio 2.7 se vio que la varianza poblacional era 21047. Con m.a.s.r. se vio que haría falta un tamaño $n = 90$ para obtener un error de muestreo de 30. ¿Que tamaño haría falta con m.a.s.?.

b) Con esa información poblacional, calcular el error de muestreo exacto en los siguientes casos: m.a.s. con tamaño $n = 20$ y m.a.s.r. con tamaño $n = 20$ para la estimación de la media poblacional.

c) Con $n = 40$, extraer una m.a.s. con el proc surveyselect y estimar la media y total poblacional con la macro estimas, usando la semilla 12345.

d) Se han creado para el propósito de este ejercicio, las macros variasmas y variasmasr, para observar gráficamente y comparar el comportamiento de los dos estimadores de la media, bajo mas y bajo masr. La sintaxis es:

```
%variasmas(replicas,tama);
```

y

```
%variasmasr(replicas,tama);
```

Las macros repiten, tantas veces como indica el parámetro "replicas", el proceso de extraer una muestra de tamaño "tama" y calcular el valor del estimador de la media. Todo con el archivo temporal madrid, para la media de la variable edif. Finalmente presentan un histograma (es el objetivo de la macro) de los valores obtenidos de los estimadores en las réplicas.

d1) Para observar cómo se comporta el estimador de la media bajo mas a medida que se aumenta el tamaño muestral, realizar 50 réplicas con muestras de tamaño $n = 10$, después de tamaño $n = 50$ y después de tamaño $n = 100$. Comparar visualmente los gráficos obtenidos.

d2) Para comparar m.a.s. con m.a.s.r, realizar las mismas operaciones anteriores con la macro variasmasr y comparar, en términos de varianza-rango, los gráficos obtenidos con los vistos en m.a.s.

d3) Observar que cuando la muestra es pequeña, la distribución es asimétrica (como lo es la de la variable edif) pero cuando aumenta n, se acerca más a la normalidad.

---

a) El tamaño sería $n = \dfrac{N z^2_{\alpha/2} \widehat{S}^2}{z^2_{\alpha/2} \widehat{S}^2 + N e^2}$, donde $S^2 = \dfrac{N}{N-1}\sigma^2 = 21165.2$, así:

$n = \dfrac{179 \cdot 1.96^2 \cdot 21165.2}{1.96^2 \cdot 21165.2 + 179 \cdot 30^2} \simeq 60$, considerablemente menor que el valor $n = 90$ obtenido para m.a.s.r.

ISBN 9788496866133

b) El error de muestreo es en m.a.s.r., $\sqrt{V(\widehat{\overline{y}})} = \frac{\sigma}{\sqrt{n}} = 32.43$ y en m.a.s.,

$$\sqrt{V(\widehat{\overline{y}})} = \sqrt{\frac{N-n}{N}\frac{S^2}{n}} = 30.66.$$

c) El programa sería el siguiente:

```
proc surveyselect data=madrid noprint out=muestra method=srs n=40 seed=12345;
run;
%estimas(muestra,edif,179);
```

Obteniendo:

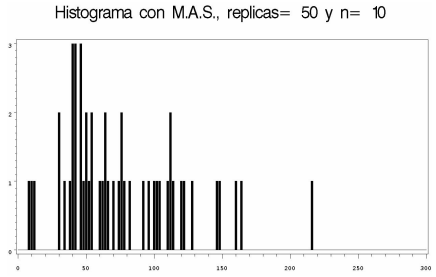
```
************************************

ESTIMACIÓN DE LA MEDIA O PROPORCIÓN

************************************
                                 Estadísticos

                           Error std de       Var de la
Variable           Media        la media           media     95% CL para la media
-------------------------------------------------------------------------------------
edif           50.300000       12.830546      164.622911    24.3477711 76.2522289
-------------------------------------------------------------------------------------

**********************

ESTIMACIÓN DEL TOTAL

**********************
                                 Estadísticos

                             Desviación          Var de
Variable             Suma        estándar         la suma       95% CL para Suma
-----------------------------------------------------------------------------------
edif          9003.700000     2296.667735         5274683    4358.25103 13649.1490
-----------------------------------------------------------------------------------
```

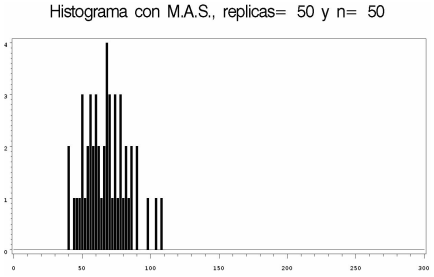
d) Se utiliza la sintaxis:

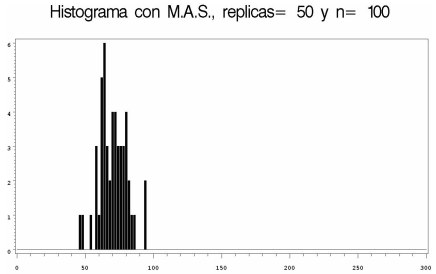
```
%variasmas(50,10);
%variasmas(50,50);
%variasmas(50,100);
```

Observando los histogramas respectivos:

ISBN 9788496866133







Los gráficos obtenidos con la macro variasmasr son similares, con mayor variabilidad en general. El programa sería:

```
%variasmasr(50,10);
%variasmasr(50,50);
%variasmasr(50,100);
```


ISBN 9788496866133

4.10 Ejercicios propuestos

1) En un estudio sobre el posible uso del muestreo para reducir el trabajo de inventario de existencias en una bodega, se realizó un recuento del valor de los artículos de cada uno de los 36 estantes de la bodega. Los valores obtenidos fueron:

29, 38, 42, 44, 45, 47, 51, 53, 53, 54, 46, 56, 58, 58, 59, 60, 60, 60, 60, 61, 61, 61, 62, 64, 65, 65, 67, 67, 68, 69, 71, 74, 77, 82, 85.

Se quiere que la estimación del valor total a partir de una muestra sea correcta salvo un error máximo de 200, con una probabilidad de error de 0.05. Un consultor sugiere que una muestra aleatoria de 12 estantes es suficiente para hacer la estimación. ¿Es correcta la afirmación?.

2) Los resultados de una encuesta realizada a una m.a.s. de 1000 personas establecen que que el 55% de los españoles considera el fútbol como su deporte favorito. Un periódico afirma: "Con una muestra de este tamaño, se puede decir con un 95% de seguridad que los resultados están dentro de más o menos 3% de lo que estarían si la población completa de adultos hubiera sido encuestada". Decir si se está de acuerdo con esta afirmación.

3) Se quiere estimar mediante m.a.s. la proporción de padres españoles que si pudieran hacer retroceder el tiempo decidirían tener hijos. Suponiendo que dicha proporción es inferior a 0.3, calcular el tamaño muestral para que el error absoluto de muestreo cometido sea inferior a 0.06 con probabilidad 0.95.

4) Una empresa de marketing está interesada en conocer la proporción de personas que se abonarían a una oferta que intentan realizar de un nuevo producto. Disponen de un fichero de 200000 direcciones, y piensan realizar un m.a.s. sin reemplazamiento.

a) ¿Qué tamaño muestral deben elegir para estimar la proporción de individuos que se abonarían con un error absoluto de muestreo del 0.5% a un nivel de confianza de 95%?.

b) Suponer que se sabe por experiencia que la proporción de abonos a este tipo e oferta suele estar generalmente entre el dos y el tres por ciento. ¿Cuál sería entonces dicho tamaño?.

5) En un estudio realizado sobre 10 pacientes escogidos por m.a.s. en un hospital con 200 pacientes se anotó el tiempo de permanencia en días en el Centro. También se preguntó al paciente su opinión sobre el Hospital.

ISBN 9788496866133

Tiempo de permanencia	Opinión
7	Buena
5	Media
6	Media
5	Media
3	Mala
8	Buena
4	Mala
2	Media
7	Buena
6	Mala

a) Estimar y dar un intervalo de confianza al 95% para el número de pacientes con opinión mala sobre el hospital.

b) Estimar el tiempo promedio de permanencia.

c) Suponiendo que la estimación dada en el apartado a) fuera cierta, presentar un intervalo aproximado de confianza al 95% para el tiempo de permanencia de los pacientes con opinión mala .

e) ¿Cuántos pacientes habría que muestrear por m.a.s. para tener una semianchura del intervalo de confianza al 95% de como máximo 8 pacientes en la estimación del apartado a)?.

6) Se pretende estudiar el número de peras por peral que hay en una huerta. Se escogen 10 perales de los 40 que hay y se obtiene la tabla de frecuencias presentada.

a) Dar un intervalo de confianza a 95% suponiendo normalidad para la estimación del número medio de peras por peral.

b) ¿En el caso del apartado a), que diferencia habría en la varianza estimada del estimador si en lugar de m.a.s. se hubiera utilizado m.a.s.r.?.

c) Suponiendo la estimación del apartado a) suficientemente precisa, decir cuál sería el error de muestreo aproximado si se seleccionaran 20 perales? ¿Y 39 perales?.

ISBN 9788496866133

Peras	Frecuencia
40	1
25	3
18	1
30	2
20	1
16	2

7) En un pueblo se desea conocer el número de familias donde algún miembro ha sido afectado por la gripe. No se tiene más información que el número de familias del pueblo, que es 500. ¿Cuántas familias habría que seleccionar por m.a.s. para asegurarse un error absoluto máximo de muestreo de 60, al nivel 95% al estimar el número de familias afectadas?.

b) Se conoce por los pueblos circundantes que la proporción de familias afectadas está entre 0.15 y 0.30. ¿Qué número de familias a muestrear se recomendaría en este caso?.

c) Se realizó el estudio con el tamaño muestral recomendado en el apartado b) y resultó una proporción de familias afectadas de 0.18. Dar un I.C. al 95% para el total de familias afectadas.

8) El archivo SAS pesos contiene datos sobre una población de 1200 personas con las variables estatura, peso, dieta y sexo. Se trata de rellenar la tabla que aparece debajo a partir de los siguientes resultados.

a) Comprobar con un proc means la verdadera proporción de personas que hacen dieta y la media poblacional de la estatura.

b) Extraer 4 muestras por m.a.s. de tamaño 40, con semillas 1234, 1235, 1236, 1237 y estimar la media de estatura y la proporción de personas que hacen algún tipo de dieta (la variable dieta está codificada como 1=hace dieta y 0=no hace dieta).

c) Realizar el apartado a) pero con tamaño muestral 400.

ISBN 9788496866133

Tamaño	Semilla	Proporción dieta	Media estatura
40	1234		
40	1235		
40	1236		
40	1237		
400	1234		
400	1235		
400	1236		
400	1237		

d) Suponiendo la última muestra de $n = 400$ suficientemente fiable, estimar el error de muestreo en la estimación de la media de estatura que se alcanzaría con $n = 1000$.

9) El archivo SAS guisa contiene datos sobre la superficie (variable super) y producción de guisantes en las 51 provincias españolas en 1998.

a) Extraer 4 muestras por m.a.s. de tamaño 10, con semillas 1234, 1235, 1236, 1237 y estimar la superficie total cultivada de guisantes en 1998.

b) Extraer 4 muestras por m.a.s.r. de tamaño 10, con semillas 1234, 1235, 1236, 1237 y estimar la superficie total cultivada de guisantes en 1998. comprobar si se repite alguna provincia.

c) Utilizar un proc means para observar la cuasivarianza de la variable superficie. Deducir cuál sería el tamaño muestral necesario exacto para tener un error de muestreo de 750 en la estimación de la media de superficie cultivada de guisante por provincia, con m.a.s.

d) Extraer una muestra aleatoria simple sin reemplazamiento con el tamaño muestral obtenido en c) y calcular la estimación del error de muestreo obtenido. Calcular también el error de muestreo exacto (puede ser algo menor que 750, debido a la corrección aplicada en el apartado c) por ser n entero). Comparar.

ISBN 9788496866133

ISBN 9788496866133

5 MUESTREO ESTRATIFICADO

5.1 Introducción y notación

Este tipo de muestreo se aplica cuando la población de tamaño N está particionada en L estratos o clases disjuntas de modo que si el tamaño de la clase o estrato h es N_h, se cumple que $\sum_{h=1}^{L} N_h = N$.

En la Figura 5.1, hay 3 estratos ($h = 1$ =ovejas blancas, $h = 2$ =ovejas negras y $h = 3$ = ovejas pintas), con tamaños respectivos $N_1 = 8$, $N_2 = 6$, $N_3 = 10$. El tamaño de la población es $N = 24$.

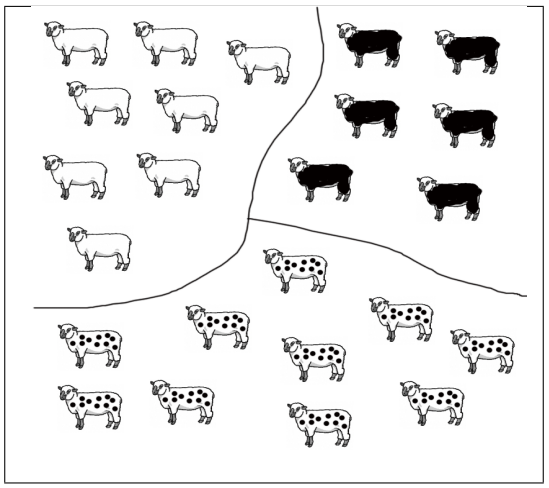


Figura 5.1. Población particionada en 3 estratos.

Se denomina **peso o ponderación** relativa al estrato h, a $W_h = \dfrac{N_h}{N}$, de manera que $\sum_{h=1}^{L} W_h = 1$. La ponderación relativa es el tamaño relativo del estrato respecto al total. En la figura, el peso de los 3 estratos es respectivamente $W_1 = 0.33, W_2 = 0.25, W_3 = 0.42$.

ISBN 9788496866133

El muestreo estratificado consiste en seleccionar aleatoriamente n_h elementos de cada estrato $h = 1, ..., L$. El tamaño final de la muestra será $n = \sum_{h=1}^{L} n_h$. La selección dentro de cada estrato es independiente del resto de estratos, de ma-nera que el muestreo estratificado es una subdivisión de un problema de muestreo en una población, en L problemas de muestreo en L poblaciones independientes. Matemáticamente, esto implica que no hay relaciones de dependencia (covarianzas) entre unidades muestreadas en un estrato y en otro diferente. Además esta división permite obtener estimaciones por separado en cada estrato.

Los motivos que llevan a realizar muestreo estratificado son tanto prácticos como teóricos. Entre ellos se pueden destacar:

- La necesidad de obtener estimaciones separadas para cada subpoblación o estrato (por ejemplo, provincias).
- Una mejor organización del trabajo de campo (por ejemplo al asignar recursos y grupos de trabajo a cada estrato geográfico).
- La posibilidad de mejorar la precisión de los estimadores globales, a través de una buena selección de estratos y de una afijación apropiada (selección de los tamaños muestrales n_h).

En la Figura 5.1, interesa realizar muestreo estratificado para estimar la cantidad media de lana que se extrae de una oveja, obteniendo también información sobre las diferencias entre los distintos de ovejas, y una mayor precisión en el estimador global si las diferencias entre los 3 tipos de ovejas en cuanto a cantidad de lana individual son importantes.

En general , para obtener un estimador con mayor precisión, conviene que los estratos sean homogéneos por dentro y diferentes entre sí, respecto a la variable de interés. Por ello la creación de estratos suele realizarse a menudo en función de una variable auxiliar relacionada con la variable de interés (por ejemplo en muestreos sobre unidades familiares de índole social, se pueden utilizar como estratos tramos de renta, pues están relacionados con las variables sociales de la encuesta).

Sea y la variable de interés. Sea y_{hi} = valor de la variable de interés en la i-ésima unidad del estrato h, para $h = 1, ..., L$, e $i = 1, ..., N_h$.

Estadísticos poblacionales.

$\overline{y}_h = \dfrac{1}{N_h} \sum_{i=1}^{N_h} y_{hi}$ = media de y en el estrato h.

$N_h \overline{y}_h = \sum_{i=1}^{N_h} y_{hi}$ = Total de y en el estrato h.

$\sigma_h^2 = \dfrac{1}{N_h} \sum_{i=1}^{N_h} (y_{hi} - \overline{y}_h)^2$ = Varianza de y en el estrato h.

ISBN 9788496866133

$S_h^2 = \dfrac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{hi} - \overline{y}_h)^2 = \dfrac{N_h}{N_h - 1}\sigma_h^2$. =Cuasivarianza de y en el estrato h.

$\overline{y} = \dfrac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h} y_{hi} = \dfrac{1}{N}\sum_{h=1}^{L} N_h\overline{y}_h = \sum_{h=1}^{L} W_h\overline{y}_h =$ Media poblacional de y.

$N\overline{y} = \sum_{h=1}^{L} N_h\overline{y}_h =$ Total poblacional.

$\sigma^2 = \dfrac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi} - \overline{y})^2$ =Varianza poblacional.

$S^2 = \dfrac{1}{N-1}\sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi} - \overline{y})^2 = \dfrac{N}{N-1}\sigma^2 =$ Cuasivarianza poblacional.

Vemos entonces que en muestreo estratificado la media poblacional puede ser expresada en términos de una suma ponderada de las medias por estratos. El total poblacional es obviamente la suma directa de los totales por estratos. En el caso de la varianza poblacional existe una descomposición útil desde el punto de vista teórico que se verá más adelante.

Estadísticos muestrales.

En cada estrato h se toma una muestra $y_{h1}, ..., y_{hn_h}$.

$\widehat{\overline{y}}_h = \dfrac{1}{n_h}\sum_{i=1}^{n_h} y_{hi} =$ media muestral de y en el estrato h.

$s_h^2 = \dfrac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \widehat{\overline{y}}_h)^2$ =Cuasivarianza muestral de y en el estrato h.

5.1.1 Descomposición de la varianza de una población estratificada

Teorema 5.1 (descomposición de la varianza).

$$\sigma^2 = \sum_{h=1}^{L} W_h\sigma_h^2 + \sum_{h=1}^{L} W_h(\overline{y}_h - \overline{y})^2$$

Demostración.

$$\sigma^2 = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi} - \overline{y})^2 = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi} - \overline{y}_h + \overline{y}_h - \overline{y})^2 =$$

$$= \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi} - \overline{y}_h)^2 + \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h}(\overline{y}_h - \overline{y})^2 + \frac{2}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi} - \overline{y}_h)(\overline{y}_h - \overline{y}) =$$

$$= \frac{1}{N}\sum_{h=1}^{L} N_h\sigma_h^2 + \frac{1}{N}\sum_{h=1}^{L} N_h(\overline{y}_h - \overline{y})^2 + \frac{2}{N}\sum_{h=1}^{L}(\overline{y}_h - \overline{y})\sum_{i=1}^{N_h}(y_{hi} - \overline{y}_h).$$

El último sumando es cero, pues para todo h,

ISBN 9788496866133

$\sum_{i=1}^{N_h}(y_{hi} - \overline{y}_h) = N_h\overline{y}_h - N_h\overline{y}_h = 0.$

Así,

$$\sigma^2 = \frac{1}{N}\sum_{h=1}^{L} N_h\sigma_h^2 + \frac{1}{N}\sum_{h=1}^{L} N_h(\overline{y}_h - \overline{y})^2 = \sum_{h=1}^{L} W_h\sigma_h^2 + \sum_{h=1}^{L} W_h(\overline{y}_h - \overline{y})^2.$$

De modo que la varianza poblacional se puede descomponer en dos sumandos: el primero relativo a la variabilidad dentro de los estratos y el segundo a la variabilidad entre estratos.

5.2 Estimación en muestreo estratificado con m.a.s. en cada estrato

5.2.1 Estimación de la media poblacional

Teorema 5.2 (estimador de la media).

Definamos con el subíndice *st* de estratificación a $\overline{y}_{st} = \sum_{h=1}^{L} W_h\widehat{\overline{y}}_h = \sum_{h=1}^{L} \frac{N_h}{N}\widehat{\overline{y}}_h$ como el estimador de la media poblacional en muestreo estratificado con m.a.s. en cada estrato. Se verifica que $\overline{y}_{st}$ es un estimador insesgado de la media poblacional $\overline{y}$.

Demostración.

$E(\overline{y}_{st}) = E\left(\sum_{h=1}^{L} W_h\widehat{\overline{y}}_h\right) = \sum_{h=1}^{L} W_h E(\widehat{\overline{y}}_h) = \sum_{h=1}^{L} W_h\overline{y}_h = \overline{y}$, donde $E(\widehat{\overline{y}}_h) = \overline{y}_h$ por tratarse cada estrato de una subpoblación independiente, donde $\widehat{\overline{y}}_h$ es el estimador usual de la media en m.a.s., y por lo tanto es insesgado.

Teorema 5.3 (varianza del estimador).

La varianza del estimador $\overline{y}_{st}$ es

$$V(\overline{y}_{st}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2}\frac{S_h^2}{n_h} = \sum_{h=1}^{L} \frac{W_h(N_h - n_h)}{N}\frac{S_h^2}{n_h}.$$

Demostración.

$V(\overline{y}_{st}) = V\left(\sum_{h=1}^{L} W_h\widehat{\overline{y}}_h\right) = \sum_{h=1}^{L} W_h^2 V\left(\widehat{\overline{y}}_h\right)$ por independencia del muestreo de unos estratos a otros.

Entonces,

$$V(\overline{y}_{st}) = \sum_{h=1}^{L} W_h^2\frac{N_h - n_h}{N_h}\frac{S_h^2}{n_h} = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2}\frac{S_h^2}{n_h}$$

por ser cada varianza $V\left(\widehat{\overline{y}}_h\right)$ la varianza del estimador $\widehat{\overline{y}}_h$ de $\overline{y}_h$ cuando se hace m.a.s. en cada estrato.

ISBN 9788496866133

Teorema 5.4 (estimación de la varianza del estimador).

Un estimador insesgado de $V(\overline{y}_{st})$ es

$$\widehat{V}(\overline{y}_{st}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2} \frac{s_h^2}{n_h}.$$

Demostración.

$$*E(\widehat{V}(\overline{y}_{st})) = E\left(\sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2} \frac{s_h^2}{n_h}\right) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2 n_h} E(s_h^2) =$$

$$= \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2} \frac{S_h^2}{n_h} = V(\overline{y}_{st}).$$

5.2.2 Estimación del total poblacional

Corolario 5.1 (estimador del total).

(a) $N\overline{y}_{st}$ es un estimador insesgado del total poblacional, con varianza

$$V(N\overline{y}_{st}) = \sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h}$$

(b) Un estimador insesgado de $V(N\overline{y}_{st})$ es $\widehat{V}(N\overline{y}_{st}) = \sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h}$

5.2.3 Estimación de la proporción poblacional

Definamos p como la proporción poblacional, p_h como la proporción en el estrato h, $\widehat{p}_h$ como la proporción muestral en el estrato h, y $p_{st} = \sum_{h=1}^{L} W_h \widehat{p}_h$ como el estimador estratificado de la proporción poblacional. Los resultados son directos, al tratarse las proporciones muestrales y poblacionales como medias muestrales y poblacionales respectivamente cuando la variable y es dicotómica.

Corolario 5.2 (estimador de la proporción).

(a) p_{st} es un estimador insesgado de la proporción poblacional, con varianza

$$V(p_{st}) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h - 1} \frac{p_h q_h}{n_h}.$$

(b) Un estimador insesgado de $V(p_{st})$ es $\widehat{V}(p_{st}) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h} \frac{\widehat{p}_h \widehat{q}_h}{n_h - 1}$.

ISBN 9788496866133

Ejemplo 5.1.

Una empresa publicitaria decide realizar una encuesta por muestreo para estimar el número medio de horas por semana que se escucha la radio en los hogares del municipio. Dicho municipio abarca 2 pueblos A y B, y un área rural. El pueblo A rodea una fábrica y la mayoría de los hogares son de obreros jóvenes con hijos en edad escolar. El pueblo B es un suburbio de una ciudad vecina, y consta de habitantes de más edad, con pocos niños en casa. Existen 155 hogares en el pueblo A, 62 en el pueblo B y 93 en el área rural. Se realiza muestreo estratificado obteniendo los datos siguientes, donde para cada estrato se refieren las horas por semana que se escucha la radio en cada hogar muestreado.

Estrato 1 (Pueblo A): 35, 28, 26, 41, 43,29,32,37,36,25,29,31,39,38,40,45,28,27,35,34.

Estrato 2 (Pueblo B): 27, 4, 49, 10, 15, 41, 25, 30.

Estrato 3 (Área rural): 8 15 21 7 14 30 20 11 12 32 34 24.

a) ¿Sería interesante estratificar la población?

b) Estimar el tiempo medio que se ve la televisión en horas por semana, para los hogares del municipio, y dar una estimación de la varianza del estimador, además de un Intervalo de confianza al 95% para la estimación suponiendo normalidad del estimador.

c) Similar al apartado b) pero referido solamente a los hogares del pueblo B

d) Comparar los resultados anteriores.

e) Estimar el número total de horas por semana que las familias del municipio dedican a ver la televisión y dar un I.C. al 95% para ese número total.

f) La empresa publicitaria quiere estimar la proporción de hogares donde se escucha un programa de entrevistas. Para ello, utilizando el anterior muestreo estratificado, recoge los siguientes datos muestrales: En el pueblo A, de los hogares muestreados 16 escuchan el programa de entrevistas . En el pueblo B, 2, y en el área rural, 3. Estimar la proporción de hogares que escuchan el programa, y dar un I.C. al 95% para esa proporción.

---

a) La partición de la población en 3 estratos facilita la tarea de selección de muestras y el trabajo de campo. Además se espera cierta homogeneidad dentro de cada estrato y posiblemente alta variabilidad (diferencias) entre los 3 estratos en cuanto al número de horas de escucha de radio . Esto favorece la precisión de un estimador global respecto a un m.a.s. no estratificado, como se verá más adelante. Por último, la estatificación permite dar estimaciones e intervalos de confianza para cada estrato por separado.

b) Vemos que el tamaño muestral n_h difiere en los 3 estratos. Obtenemos los siguientes datos muestrales:

ISBN 9788496866133

	Estrato 1	Estrato 2	Estrato 3
N_h	155	62	93
n_h	20	8	12
$\widehat{\overline{y}}_h$	33.9	25.125	19
s_h^2	35.358	232.411	87.636

Tabla 5.1. Datos muestrales.

Además, $N = 155 + 62 + 93 = 310$ y $n = n_1 + n_2 + n_3 = 40$.

El estimador de la media poblacional es

$\overline{y}_{st} = \sum_{h=1}^{L} \frac{N_h}{N} \widehat{\overline{y}}_h = \frac{N_1}{N}\widehat{\overline{y}}_1 + \frac{N_2}{N}\widehat{\overline{y}}_2 + \frac{N_3}{N}\widehat{\overline{y}}_3 = \frac{155}{310}33.9 + \frac{62}{310}25.125 + \frac{93}{310}19 = 27.7$ horas a la semana por hogar del municipio

El estimador de la varianza de $\overline{y}_{st}$ será

$$\widehat{V}(\overline{y}_{st}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2} \frac{s_h^2}{n_h} =$$

$$= \frac{155(155-20)}{310^2}\frac{35.358}{20} + \frac{62(62-8)}{310}\frac{232.411}{8} + \frac{93(93-12)}{310^2}\frac{87.636}{12} = 1.97.$$

Con lo cual un intervalo de confianza al 95% para $\overline{y}$ vendrá dado por

$$(\overline{y}_{st} - z_{\alpha/2}\sqrt{\widehat{V}(\overline{y}_{st})}, \overline{y}_{st} + z_{\alpha/2}\sqrt{\widehat{V}(\overline{y}_{st})}) = (27.7 - 1.96\sqrt{1.97}, 27.7 + 1.96\sqrt{1.97}) =$$

$$= (24.92, 30.45)$$

c) Para el pueblo B se tiene $\widehat{\overline{y}}_2 = 25.125$ con $\widehat{V}(\widehat{\overline{y}}_2) = \frac{N_2 - n_2}{N_2}\frac{s_2^2}{n_2} = \frac{62-8}{62}\frac{232.411}{8} = 25.303$.

d) La estimación de la media poblacional $\overline{y}$ es bastante buena, pero la estimación particular de la media del estrato 2 no lo es tanto, es decir, la varianza estimada del estimador toma un valor muy alto, debido a la gran variabilidad interna de ese estrato ($s_2^2 = 232.411$) y al pequeño tamaño de la muestra.

e) En el estrato 1, la proporción muestral es

$$\widehat{p}_1 = \frac{\text{número de unidades (hogares) que escuchan el programa de entrevistas}}{\text{número de unidades (hogares) muestreados}} = \frac{16}{20} = 0.8.$$

Análogamente, $\widehat{p}_2 = 0.25$ y $\widehat{p}_3 = 0.5$.

La estimación de la proporción poblacional vendrá dada por

$$p_{st} = \sum_{h=1}^{L} W_h \widehat{p}_h = \frac{155}{310}0.8 + \frac{62}{310}0.25 + \frac{93}{310}0.5 = 0.6$$

ISBN 9788496866133

La estimación de la varianza de ese estimador es

$$\widehat{V}(p_{st}) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h} \frac{\widehat{p}_h \widehat{q}_h}{n_h - 1} = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2} \frac{\widehat{p}_h \widehat{q}_h}{n_h - 1} =$$

$$\frac{155(155-20)}{310^2} \frac{0.8 \cdot 0.2}{19} + \frac{62(62-8)}{310} \frac{0.25 \cdot 0.75}{7} + \frac{93(93-12)}{310^2} \frac{0.5 \cdot 0.5}{11} = 0.0045$$

con lo cual un intervalo de confianza al 95% para p vendrá dado por

$$= (0.6 - 1.96\sqrt{0.0045}, 0.6 + 1.96\sqrt{0.0045}) = (0.46, 0.73).$$

5.3 Afijación muestral

Dado el tamaño muestral n, se denomina afijación muestral al reparto de la muestra en los diferentes estratos. Es decir, es la división de n en L números n_h, con $h = 1, ..., L$, de manera que $\sum_{h=1}^{L} n_h = n$. A continuación se verán varios tipos de afijación, suponiendo m.a.s. en cada estrato.

5.3.1 Afijación igual

Consiste en asignar el mismo tamaño muestral en cada estrato, es decir, $n_1 = n_2 = ... = n_h = ... = n_L$. Como $n = \sum_{h=1}^{L} n_h = L n_h$ pues todos los n_h son iguales, se tiene que $n_h = \frac{n}{L}$.Este tipo de afijación favorece a los estratos de menor tamaño y perjudica a los grandes, en cuanto a precisión. También se puede denominar afijación uniforme.

5.3.2 Afijación proporcional

Consiste en asignar a cada estrato h un tamaño muestral proporcional al tamaño relativo de dicho estrato, es decir, $\frac{n_h}{n} = \frac{N_h}{N}$.Así, $n_h = n\frac{N_h}{N} = nW_h$. Se verifica que con está afijación, todas las unidades de la población tienen la misma probabilidad de figurar en la muestra de n unidades. Es posiblemente la más utilizada, cuando los tamaños de los estratos varían mucho entre sí.

5.3.3 Afijación de varianza mínima

Dado un tamaño muestral total n , esta afijación consiste en asignar a cada estrato un tamaño muestral n_h de modo que la varianza $V(\overline{y}_{st})$ del estimador estratificado sea mínima. Para ello es necesario disponer de una aproximación a las cuasivarianzas por estrato S_h^2.

Se trata de un problema de optimización no lineal con una restricción de igualdad donde se puede utilizar el método de los multiplicadores de Lagrange.

ISBN 9788496866133

$Min\ V(\overline{y}_{st})$

sujeto a

$$\sum_{h=1}^{L} n_h - n = 0.$$

Definiendo el lagrangiano

$$\Phi(n_1, ..., n_h) = V(\overline{y}_{st}) + \lambda \left(\sum_{h=1}^{L} n_h - n \right) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2} \frac{S_h^2}{n_h} + \lambda \left(\sum_{h=1}^{L} n_h - n \right)$$

Derivando respecto de cada n_h,

$$\frac{\partial \Phi}{\partial n_h} = -\frac{N_h^2 S_h^2}{N^2 n_h^2} + \lambda = 0 \text{ para todo } h = 1, ..., L$$

Así, $\lambda = \frac{N_h^2 S_h^2}{N^2 n_h^2}$ con lo cual y $\sqrt{\lambda} = \frac{N_1 S_1}{N n_1} = \frac{N_2 S_2}{N n_2} = ... \frac{N_h S_h}{N n_h} = ... = \frac{N_L S_L}{N n_L}$

Como $\frac{a}{b} = \frac{c}{d} \Rightarrow \frac{a+c}{b+d} = \frac{a}{b}$, tenemos que

$$\frac{\sum_{h=1}^{L} N_h S_h}{\sum_{h=1}^{L} N n_h} = \frac{N_h S_h}{N n_h} \text{ para todo } h \text{ y, por lo tanto ,}$$

$$\frac{\sum_{h=1}^{L} N_h S_h}{N n} = \frac{N_h S_h}{N n_h} \text{ lo que implica que } n_h = n \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \text{ es la afijación de varianza mínima.}$$

Vemos que en la expresión intervienen tanto la cuasivarianza en cada estrato como el tamaño del estrato (a mayor cuasivarianza y mayor tamaño en el estrato h, mayor tamaño muestral es requerido en h). Si la cuasivarianza no varía mucho de estrato a estrato , es decir $S_h \simeq S$ para todo h, esta afijación coincide con la afijación proporcional, por lo que la afijación de varianza mínima sólo tiene utilidad cuando la diferencia de variabilidad entre estratos es alta, pues tiene el defecto de requerir las estimaciones $\widehat{S}_h$.

5.3.4 Afijación óptima con costes variables

En este tipo de afijación se considera que el coste de observación de cada unidad muestral en el estrato h es C_h. Así, el coste total de la muestra es $C = \sum_{h=1}^{L} C_h n_h$. La afijación consiste en minimizar la varianza del estimador sujeto al coste fijo C :

$Min\ V(\overline{y}_{st})$

sujeto a

ISBN 9788496866133

$$\sum_{h=1}^{L} C_h n_h - C = 0$$

Con lo cual , desarrollando como en la afijación de varianza mínima, $\frac{\partial \Phi}{\partial n_h} = -\frac{N_h^2 S_h^2}{N^2 n_h^2} + \lambda C_h = 0$ para todo $h = 1, ..., L$ y entonces,

$\sqrt{\lambda} = \frac{\frac{N_h S_h}{\sqrt{C_h}}}{N n_h}$ y por lo tanto, $n_h = n \frac{\frac{N_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{C_h}}}$.

Esta relación se puede expresar también de otra manera, pues como

$$C = \sum_{h=1}^{L} C_h n_h = n \frac{\sum_{h=1}^{L} N_h S_h \sqrt{C_h}}{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{C_h}}}$$

esto implica que $n = C \frac{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} N_h S_h \sqrt{C_h}}$ y

$$n_h = C \frac{\frac{N_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} N_h S_h \sqrt{C_h}} \text{ para todo } h = 1, ..., L.$$

Si los costes son similares en los distintos estratos, esta afijación coincide con la de varianza mínima. Por otro lado, el criterio es igualmente válido para otras funciones de coste, en cuyo caso habría que adaptar la expresión de la afijación, minimizando la varianza sujeto a la expresión correspondiente.

5.3.5 Afijación fija

Es una afijación arbitraria de los tamaños n_h en cada estrato, motivada por motivos prácticos, de coste u organización.

ISBN 9788496866133

Ejemplo 5.2.

En una comarca compuesta por tres pueblos A,B, y C se desea conocer la edad media de sus habitantes. Para ello se dispone de un presupuesto de de 1000 euros, y se supone que el coste por observación es de $C_A = C_B = 0.8$ euros, y $C_C = 1.2$ euros.

(a) Determinar el tamaño muestral en cada pueblo y el tamaño muestral en total si de una encuesta previa se ha estimado que las cuasivarianzas son $S_A^2 = 30^2, S_B^2 = 32^2, S_C^2 = 40^2$, y sabiendo que el número de habitantes en cada pueblo es $N_A = 25.000, N_B = 12.000, N_C = 2.000$. El objetivo es obtener la máxima precisión a coste fijo.

(b) Presentar también las diferentes configuraciones de la muestra, suponiendo el tamaño n derivado de los cálculos anteriores, para la afijación igual, proporcional y de varianza mínima.

(c) Realizar un programa y presentar una tabla que indique los costes y precisión para diferentes afijaciones, variando los n_h de 50 en 50 y suponiendo $n = 1200$. Expresar en un gráfico la relación entre coste y varianza para diferentes afijaciones.

(a) Hay tres estratos A,B,C. El presupuesto o coste general fijo es $C = 1000$. Hay que hallar n_A, n_B, n_C, tales que se minimice la varianza $V(\overline{y}_{st})$ sujeto a $1000 = 0.8(n_A + n_B) + 1.2n_C$. Aplicando el resultado de afijación óptima,

$$n_h = C \frac{\frac{N_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} N_h S_h \sqrt{C_h}}$$

En este caso $\sum_{h=1}^{L} N_h S_h \sqrt{C_h} = (25.000)30\sqrt{0.8}+(12.000)32\sqrt{0.8}+(2.000)40\sqrt{1.2} = 1101916.04$. Además, $\frac{N_A S_A}{\sqrt{C_A}} = \frac{25000 \cdot 30}{\sqrt{0.08}} = 838525.5$, $\frac{N_B S_B}{\sqrt{C_B}} = 429325.05$ y $\frac{N_C S_C}{\sqrt{C_C}} = 73029.6$.

Así,

$$n_A = 1000\frac{838525.5}{1101916.04} = 761,$$

$$n_B = 1000\frac{429325.05}{1101916.04} = 390,$$

$$n_A = 1000\frac{73029.6}{1101916.04} = 66.$$

Así, $n = 760 + 390 + 66 = 1217$.

(b) Con $n = 1217$ y afijación igual, se tendría aproximadamente $n_A = n_B = n_C = 406$.

Con afijación proporcional, $n_A = 1217\frac{25000}{39000} = 780$, $n_B = 1217\frac{12000}{39000} = 375$, $n_C = 1217\frac{2000}{39000} = 62$

ISBN 9788496866133

Con afijación de varianza mínima, se tiene que $\sum_{h=1}^{L} N_h S_h = 25000 \cdot 30 + 12000 \cdot 32 + 2000 \cdot 40 = 1214000$, y $n_A = n\dfrac{N_A S_A}{\sum_{h=1}^{L} N_h S_h} = 752$, $n_B = n\dfrac{N_B S_B}{\sum_{h=1}^{L} N_h S_h} = 385$, $n_C = n\dfrac{N_C S_C}{\sum_{h=1}^{L} N_h S_h} = 80$.

(c) La programación consiste en dos bucles haciendo variar n_1 y n_2, de 50 en 50, calculando n_3 y dentro del bucle la varianza del estimador y el coste. El programa sería algo así como:

```
s1=30;s2=32;s3=40
c1=0.8;c2=0.8;c3=1.2
ng1=25000;ng2=12000;ng3=2000;ng=39000
repetir desde n1=50 hasta 1150
  repetir desde n2=50 hasta 1150-n1
    n3=1200-n1-n2
    varianza=(ng1-n1)*(s1**2)/n1+(ng2-n2)*(s2**2)/n2+(ng3-n3)*(s3**2)/n3
    varianza=varianza/(ng**2)
    coste=c1*n1+c2*n2+c3*n3
  fin_repetir
fin_repetir
```

Ordenando por varianza y coste, los primeros valores de la tabla son:

nA	nB	nC	varianza	coste	
550	450	200	.000053052	1040	
600	400	200	.000053055	1040	
550	400	250	.000053192	1060	
500	450	250	.000053638	1060	
600	350	250	.000053836	1060	
500	500	200	.000053946	1040	
650	350	200	.000054043	1040	
600	450	150	.000054317	1020	***
500	400	300	.000054479	1080	
650	400	150	.000054665	1020	
550	350	300	.000054675	1080	
550	500	150	.000054763	1020	
450	500	250	.000055130	1060	
450	450	300	.000055522	1080	
450	550	200	.000055765	1040	
...					

Tabla 5.2. Varianza y coste para varios n_1, n_2 y n_3.

ISBN 9788496866133

Si se ordenan por coste y varianza, queda:

nA	nB	nC	varianza	coste
650	500	50	.000078677	980
700	450	50	.000078846	980
600	550	50	.000079104	980
750	400	50	.000079682	980
550	600	50	.000080122	980
800	350	50	.000081334	980
500	650	50	.000081776	980
850	300	50	.000084094	980
450	700	50	.000084175	980
400	750	50	.000087515	980
900	250	50	.000088513	980
350	800	50	.000092125	980
950	200	50	.000095727	980
300	850	50	.000098575	980
250	900	50	.000107909	980
...				

Tabla 5.2. Varianza y coste para varios n_1, n_2 y n_3.

Como se ve, si no se asume coste fijo ni varianza fija , no hay una afijación que minimice simultáneamente ambos parámetros. En la Figura 5.2 se observa la relación entre coste y varianza.

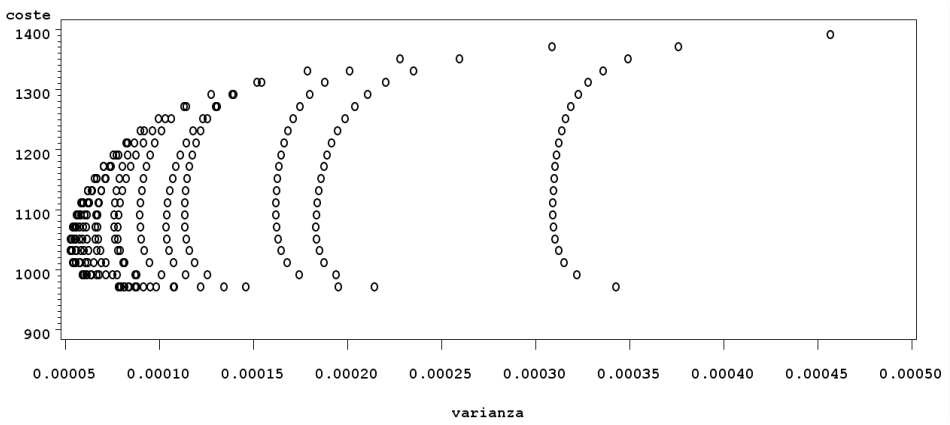


Figura 5.2. Relación entre coste y varianza para varias afijaciones.

Observando el gráfico, una solución de equilibrio sería alguna de las cercanas a la esquina inferior izquierda. Por ejemplo, la afijación que da lugar a varianza=0.000054317, y coste=1020 . Es decir, la afijación $n_A = 600$, $n_B = 450$, $n_C = 150$ (marcada con *** en la Tabla 5.2).

Pero es el investigador quien debe decidir en última instancia, de acuerdo a las tablas y gráficos, y a la importancia relativa que tienen la varianza y el coste en el trabajo que se lleva a cabo. Supongamos que decide que no puede asumir un coste mayor de 1000 ni una varianza del estimador mayor que

ISBN 9788496866133

0.00007. Entonces, es fácil truncar el gráfico empleado dibujando dos rectas correspondientes a esos puntos de corte, como aparece en la Figura 5.3, restringiéndose a la esquina izquierda y decidiéndose en este caso por la solución $n_A = 650$, $n_B = 450$, $n_C = 100$,con varianza= 0.000059433 y coste=1000.

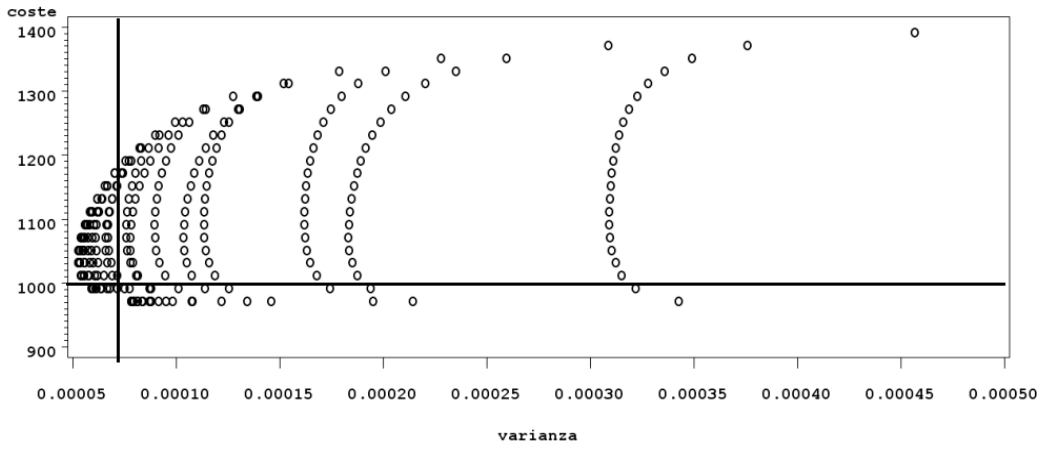


Figura 5.3. Relación entre coste y varianza con cotas superiores.

5.4 Comparaciones con m.a.s.

En esta sección se comparan la estrategia m.a.s. con las estrategias estratificadas con afijación proporcional y de varianza mínima, suponiendo el mismo tamaño muestral n y en el caso en que N_h y N sean suficientemente grandes (en ese caso S_h^2 y S^2 son próximas a σ_h^2 y σ^2). Se observa en la comparación como la estratificación puede abocar a cierta ganancia en precisión.

Teorema 5.5 (comparación entre m.a.s. y muestreo estratificado).

Suponiendo N y N_h suficientemente grandes,

$V(\overline{y}_{st}, afij.var.mín.) \lesssim V(\overline{y}_{st}, afij.prop.) \lesssim V(\widehat{\overline{y}})$.

Demostración.

En m.a.s., $V(\widehat{\overline{y}}) = \dfrac{N-n}{N}\dfrac{S^2}{n} \simeq \dfrac{\sigma^2}{n}$ por considerarse N suficientemente grande y por lo tanto $(1-f) \simeq 1$ y $S^2 \simeq \sigma^2$.

En muestreo estratificado con afijación proporcional, sustituyendo $n_h = n\dfrac{N_h}{N}$ en la fórmula de varianza, y realizando la aproximación suponiendo N y N_h grandes,

$$V(\overline{y}_{st}, afij.prop.) = \sum_{h=1}^{L} \frac{N_h(N_h-n_h)}{N^2}\frac{S_h^2}{n_h} = \sum_{h=1}^{L} \frac{N_h(N_h-n\frac{N_h}{N})}{N^2}\frac{S_h^2 N}{nN_h} =$$

ISBN 9788496866133

$$=\sum_{h=1}^{L} \frac{N_h(1-\frac{n}{N})}{N}\frac{S_h^2}{n} \simeq \sum_{h=1}^{L}\frac{N_h}{N}\frac{\sigma_h^2}{n} = \frac{1}{n}\sum_{h=1}^{L} W_h\sigma_h^2.$$

Como se ha visto en la descomposición de la varianza en una población estratificada,

$$\sigma^2 = \sum_{h=1}^{L} W_h\sigma_h^2 + \sum_{h=1}^{L} W_h(\overline{y}_h - \overline{y})^2$$

con lo que

$$\frac{1}{n}\sum_{h=1}^{L} W_h\sigma_h^2 = \frac{\sigma^2}{n} - \frac{1}{n}\sum_{h=1}^{L} W_h(\overline{y}_h - \overline{y})^2 \leq \frac{\sigma^2}{n}.$$

Por lo tanto, cuando N y N_h son suficientemente grandes, $\frac{1}{n}\sum_{h=1}^{L} W_h\sigma_h^2 \leq \frac{\sigma^2}{n}$ y así

$$V(\overline{y}_{st}, afij.prop.) \lesssim V(\widehat{\overline{y}}, m.a.s.).$$

Por otra parte, se sabe que

$$V(\overline{y}_{st}, afij.var.mín). \lesssim V(\overline{y}_{st}, afij.prop.)$$

pues en la afijación de varianza mínima, como su nombre indica es la que da lugar a la menor varianza del estimador. Así,

$$V(\overline{y}_{st}, afij.var.mín). \lesssim V(\overline{y}_{st}, afij.prop.) \lesssim V(\widehat{\overline{y}}).$$

Aunque se han supuesto N_h y N grandes, en general el muestreo estratificado con afijación proporcional o de varianza mínima produce estimaciones más precisas que el m.a.s. La ganancia en precisión es mayor, para la afijación proporcional, si los tamaños difieren mucho de unos estratos a otros. Para la afijación de varianza mínima la ganancia en precisión aumenta si tanto los tamaños de los estratos como la variabilidad de la variable de interés difieren mucho de unos estratos a otros.

Ejemplo 5.3.

Se desea estimar la media de producción media de lana por oveja en un rebaño de ovejas de razas mezcladas, como el que aparece en la Figura. 5.4. Se puede utilizar m.a.s. o muestreo estratificado con diferentes afijaciones.

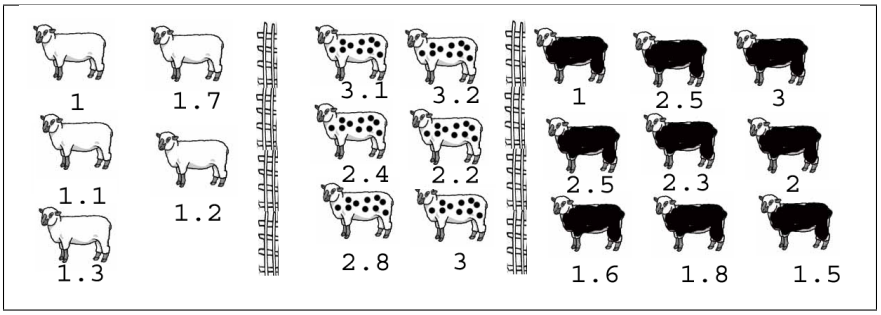


Figura 5.4. Kilogramos de lana obtenidos de cada oveja.

ISBN 9788496866133

Suponiendo $n = 6$:

a) Calcular la varianza del estimador de la media bajo muestreo estratificado con afijación proporcional.

b) Calcular la varianza del estimador bajo muestreo estratificado con afijación de varianza mínima .

c) Calcular la varianza del estimador bajo muestreo aleatorio simple.

d) Calcular la varianza del estimador bajo muestreo estratificado, con $n_1 = 3$, $n_2 = 2$, $n_3 = 1$.

e) ¿Qué problemas dan en este ejemplo las afijaciones proporcional y de varianza mínima?.

f) Presentar un gráfico de cajas de la población expuesta.

---

a) Se observa que las medias poblacionales por estrato son muy diferentes, pues $\overline{y}_1 = 1.26$, $\overline{y}_2 = 2.783$, e $\overline{y}_3 = 2.022$, con lo cual el muestreo estratificado mejoraría al muestreo aleatorio simple, si la afijación es proporcional o de varianza mínima. Nótese que la media poblacional a estimar es $\overline{y} = 2.06$.

Suponiendo las varianzas poblacionales por estrato conocidas, $S_1^2 = 0.073$, $S_2^2 = 0.1617$, $S_3^2 = 0.38$. Se cifra el tamaño muestral total en $n = 6$.

Por afijación proporcional, sería $n_1 = 6\dfrac{5}{20} = 1.5$, $n_2 = 6\dfrac{6}{20} = 1.8$, $n_3 = 6\dfrac{9}{20} = 2.7$ lo que se puede aproximar por $n_1 = 1, n_2 = 2$ y $n_3 = 3$ y además,

$$V(\overline{y}_{st}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2}\frac{S_h^2}{n_h} = \frac{5(5-1)}{20^2}\frac{0.073}{1} + \frac{6(6-2)}{20^2}\frac{0.1617}{2} + \frac{9(9-3)}{20^2}\frac{0.38}{3} = 0.017051.$$

b) Si se utiliza afijación de varianza mínima, se tiene que $\sum_{h=1}^{L} N_h S_h = 5{\cdot}0.27 + 6{\cdot}0.402 + 9{\cdot}0.6164 = 9.31$. Entonces $n_1 = 6\dfrac{5 \cdot 0.27}{9.31} = 0.87$, $n_2 = 6\dfrac{6 \cdot 0.402}{9.31} = 1.55$, $n_3 = 6\dfrac{9 \cdot 0.6164}{9.31} = 3.57$ lo que da una afijación aproximada $n_1 = 1, n_2 = 2$ y $n_3 = 3$ equivalente a la afijación proporcional y por lo tanto corresponde a la misma varianza del estimador.

c) Si se utiliza muestreo aleatorio simple, $S^2 = 0.552$ y entonces $V(\widehat{\overline{y}}) = \dfrac{20-6}{20}\dfrac{0.552}{6} = 0.0644$, mayor que la varianza obtenida en muestreo estratificado. En general, aunque N no sea grande, si la diferencia de las medias por estratos es amplia, se produce una mejora al utilizar muestreo estratificado con afijación proporcional o de varianza mínima.

d) Si se utiliza muestreo estratificado pero la afijación es arbitraria, se corre el peligro de incurrir en una estimación peor que la del m.a.s. En este ejemplo, si $n_1 = 3$, $n_2 = 2$, $n_3 = 1$. Entonces, la varianza del estimador será:

$V(\overline{y}_{st}) = \sum_{h=1}^{L} \dfrac{N_h(N_h - n_h)}{N^2}\dfrac{S_h^2}{n_h} = \dfrac{5(5-3)}{20^2}\dfrac{0.073}{3} + \dfrac{6(6-2)}{20^2}\dfrac{0.1617}{2} + \dfrac{9(9-1)}{20^2}\dfrac{0.38}{1} = 0.0738$, mayor que la varianza del estimador en m.a.s.

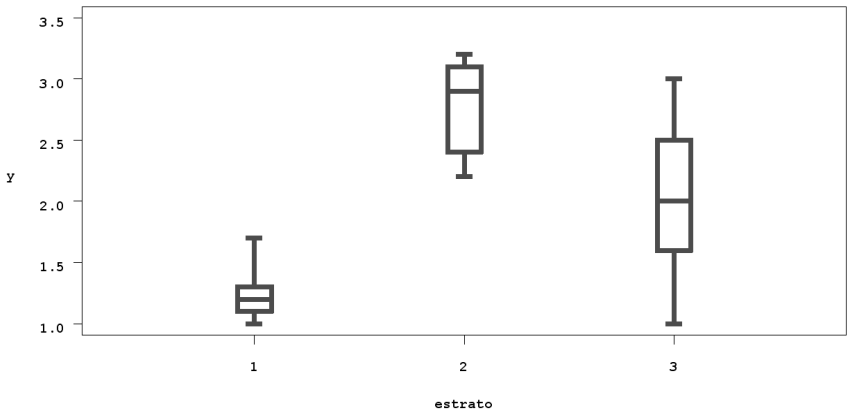
e) Ambas afijaciones dan lugar a un tamaño muestral de una unidad en un estrato, lo que impide estimar la varianza del estimador (se necesitan al menos dos unidades para calcular s_h^2 en cada estrato h) y por lo tanto aproximar en la práctica la precisión de nuestras estimaciones. Una solución práctica en este caso es tomar $n_1 = 2, n_2 = 2$ y $n_3 = 2$ (afijación igual), con lo que

ISBN 9788496866133

$V(\overline{y}_{st}) = \dfrac{5(5-2)}{20^2}\dfrac{0.073}{2} + \dfrac{6(6-2)}{20^2}\dfrac{0.1617}{2} + \dfrac{9(9-2)}{20^2}\dfrac{0.38}{2} = 0.03614$ que sigue siendo menor que la obtenida con m.a.s.

f) El gráfico de cajas permite observar y comparar la conformación de la población para los diferentes estratos. Normalmente no se dispone de toda la información poblacional, como en este ejemplo, pero igualmente se puede presentar este tipo de gráfico sobre la muestra estratificada obtenida.

En el gráfico se pueden observar las diferencias entre los 3 estratos en media de la variable de interés, lo que justifica el muestreo estratificado, y también en variabi-lidad, lo que justifica tomar un tamaño muestral en la afijación mayor en el tercer estrato y menor en el primero.



Figra 5.5. Diagrama de cajas por estrato.

Ejemplo 5.4

En este ejemplo se realizará un estudio de simulación basado en el ejemplo anterior. Se comparará la distribución del estimador usual de m.a.s. con la distribución del estimador por muestreo estratificado con afijación proporcional. Se fijará $n = 6$, y $n_1 = 1$, $n_2 = 2$, $n_3 = 3$.

El proceso de obtener una muestra estratificada con esa afijación, calculando el valor del estimador de la media $\overline{y}_{st} = \sum_{h=1}^{L} \dfrac{N_h}{N}\widehat{\overline{y}}_h$, se realizará 200 veces y se obtendrá el histograma de la variable $\overline{y}_{st}$ (recordemos que la obtención de la muestra estratificada consiste en realizar m.a.s. en cada estrato h de tamaño n_h).

Análogamente se realizará 200 veces el proceso de obtener una muestra aleatoria simple de tamaño $n = 6$ y calcular el estimador de la media $\widehat{\overline{y}}$, dibujando finalmente el histograma.

ISBN 9788496866133

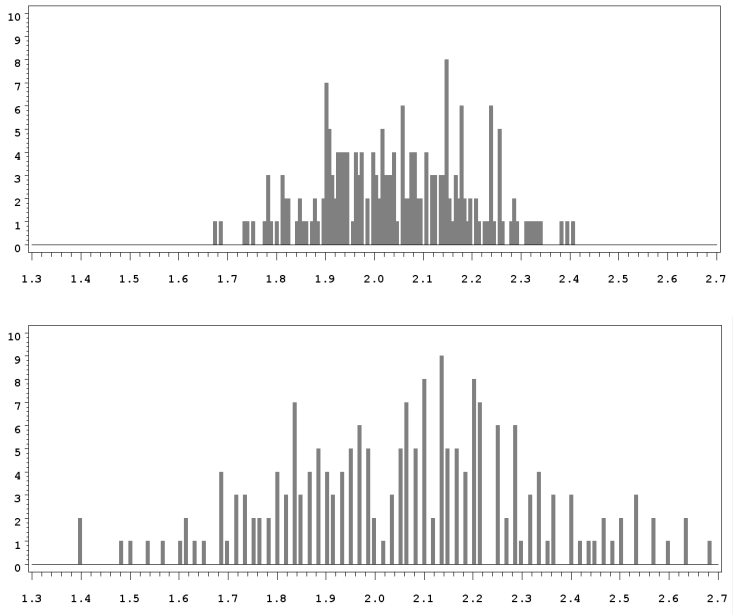


Figura 5.5. Histograma de 200 obtenciones del estimador estratificado (arriba) y del estimador por m.a.s. (abajo) sobre los datos del ejemplo.

Como se observa en el gráfico, la variabilidad del estimador media muestral con m.a.s. es mayor que en muestreo estratificado con afijación proporcional. Se comprueba también que ambos estimadores son insesgados: su centro de gravedad está en torno a la media poblacional 2.06. Además se puede comprobar que la distribución de ambos estimadores es aproximadamente normal.

5.5 Estratos como dominios de estudio

Algunos de los planteamientos para la afijación en muestreo estratificado realizados en las anteriores secciones, consistentes en minimizar la varianza del estimador global, pueden no tener sentido si los principales objetivos de la investigación son una estimación suficientemente precisa en cada estrato y/o la comparación entre estratos. Bajo estos criterios, la afijación puede ser algo diferente. Una manera habitual de actuar es exploratoria, a través de tablas y gráficos como en anteriores ejemplos, decidiendo la afijación que mejor corresponda a los intereses del investigador, en función de las precisiones obtenidas por estratos.

En el caso particular en que lo que interesa es la comparación entre estratos vía contraste de hipótesis, por ejemplo, un criterio es escoger una afijación que de una buena precisión en la estimación de las diferencia de medias entre estratos. Si hay dos estratos, el criterio será escoger la afijación que cumpla

$$Min\left\{V(\widehat{\overline{y}}_1-\widehat{\overline{y}}_2)\right\}=Min\left\{\frac{N_1-n_1}{N_1}\frac{S_1^2}{n_1}+\frac{N_2-n_2}{N_2}\frac{S_2^2}{n_2}\right\}$$

sujeto a

$n_1+n_2=n.$

ISBN 9788496866133

Derivando y utilizando el lagrangiano, queda $n_1 = n\dfrac{1}{S_1/S_2+1}$ y $n_2 = n - n_1$.

También se puede minimizar la función sujeta a un coste fijo con una cierta función de coste. Si la intención es realizar solamente un contraste de hipótesis y no interesa tanto la estimación de la diferencia $\overline{y}_1 - \overline{y}_2$, se pueden eliminar los términos de corrección por población finita. Esto puede arrojar alguna diferencia en el caso de utilizar funciones de coste.

Ejemplo 5.5

Considerando el ejemplo de las ovejas, supongamos que la población sólo contiene los estratos 2 y 3 (que llamaremos 1 y 2 respectivamente por simplificar la notación). Si se dispone de recursos para un tamaño muestral final $n = 6$, y el objetivo es comparar ambos estratos, la afijación sería

$n_1 = 6\dfrac{1}{\dfrac{0.27}{0.40}+1} = 3.58$ y $n_2 = 6 - 3.58 = 2.41$ con lo que $n_1 = 4$ y $n_2 = 2$. En términos comparativos, la afijación proporcional en este caso sería $n_1 = 6\dfrac{6}{15} = 2.4$ y $n_2 = 3.6$ con lo que $n_1 = 2$ y $n_2 = 4$. La afijación de varianza mínima sería $n_1 = 6\dfrac{6 \cdot 0.40}{3.885} = 3.70$ y $n_2 = 2.30$, que da una afijación similar a la obtenida al minimizar la varianza de la diferencia.

5.6 Construcción de los estratos

Ejemplo 5.6.

Si la población está configurada en tres estratos con valores respectivos de y : $\{1,2,3\}$, $\{1,2,3\}$, $\{1,2,3\}$, y se realiza muestreo estratificado con $n = 6$ con afijación proporcional o de varianza mínima queda la misma configuración $n_1 = n_2 = n_3 = 2$ y $V(\overline{y}_{st}) = 3\left(\dfrac{3-2}{3}\cdot\dfrac{1}{2}\right) = 0.5$. Si para el mismo tamaño muestral $n = 6$ se realiza muestreo aleatorio simple, se obtiene $V(\widehat{\overline{y}}) = \dfrac{9-6}{9}\cdot\dfrac{3}{4} = 0.25$, con lo cual en este caso el muestreo aleatorio simple mejoraría al muestreo estratificado. Es un caso extremo, pues los estratos son iguales entre sí tanto en valores como en tamaño. Si construimos los estratos en la misma población como $\{1,1,1\},\{2,2,2\},\{3,3,3\}$, obtenemos que la varianza $V(\overline{y}_{st})$ del estimador estratificado es cero, pues $S_h = 0$ para todo $h = 1,2,3$. Sin embargo, $V(\widehat{\overline{y}}) = 0.25$. Es otro caso extremo, donde ahora el muestreo estratificado sería perfecto. A partir de este ejemplo se intuye que la construcción adecuada de los estratos (homogéneos por dentro y diferentes entre sí) redunda en la precisión final en igual o mayor medida que otros factores como el tipo de estimador o el tamaño muestral.

La decisión sobre el número de estratos L depende de factores como los costes, la precisión y también de la necesidad de obtener estimaciones separadas. El número de estratos debe ser suficientemente pequeño para que al menos se obtenga $n_h \geq 2$, de cara a poder estimar la cuasivarianza en cada estrato. Esto lleva a $L \leq \dfrac{n}{2}$. Utilizar un número grande de estratos puede

ISBN 9788496866133

no redundar en una mejora significativa en la precisión en términos relativos, incluso Cochran (1990) recomienda tomar $L \leq 6$, pero a menudo factores prácticos (geográficos, administrativos) son los que deciden el tipo de estrato y el número. de éstos.

Una vez determinado el número de estratos L, queda por definir cuáles son estos estratos. Frecuentemente existen subdivisiones en la realidad que los definen (provincias, distritos, municipios, etc.). A menudo ciertos grupos especiales deben considerarse estratos (por ejemplo, una gran ciudad debe considerarse un estrato en sí misma). Cuando se realizan subdivisiones arbitrarias por motivos de organización o de estructura interna de la población, debe de tenerse en cuenta que los estratos conforman una partición de la población: deben de ser disjuntos y su unión debe ser toda la población.

Si no existen subdivisiones naturales u organizativas, una posibilidad es utilizar una variable auxiliar x para delimitar los estratos, de la que se conocen los valores para toda la población. Esta variable es interesante que esté correlacionada con la variable de interés y para justificar la estratificación.El método de la **regla de acumulación de la raíz de la frecuencia**, o regla de Dalenius-Hodges, consiste en los siguientes pasos:

1. Ordenar la población por la variable continua x.
2. Realizar una tabla de frecuencias considerando arbitrariamente un número M suficientemente grande de intervalos de valores de X , llamados I_i, $i = 1, ...M$. Denotar por f_i la frecuencia absoluta en el intervalo I_i.
3. Calcular $\sqrt{f_i}$, la raíz de la frecuencia absoluta para cada intervalo I_i, y el valor acumulado $\sqrt{f_i}^{ac}$ para cada intervalo I_i. Definir la suma $\sum_{i=1}^{M} \sqrt{f_i} = \sqrt{f_M}^{ac}$. Calcular $k = \frac{1}{L}\sqrt{f_M}^{ac}$.
4. Los puntos de corte de los estratos vienen definidos por el límite superior del intervalo cuyo valor $\sqrt{f_i}^{ac}$ esté lo más cerca posile de los múltiplos de k $(k, 2k, ..., (L-1)k)$.

Ejemplo 5.7.

Un investigador desea estimar el promedio anual de ventas de este año para 56 empresas. Se encuentran disponibles los datos de frecuencias en una clasificación del años pasado por incrementos de 50.000 euros que se presentan en la Tabla 5.3. Se trata de asignar las empresas a $L = 3$ estratos.

ISBN 9788496866133

Ingreso en miles	f_i
$100 - 150$	11
$150 - 200$	14
$200 - 250$	9
$250 - 300$	4
$300 - 350$	5
$350 - 400$	8
$400 - 450$	3
$450 - 500$	2

Tabla 5.3. Ingresos y frecuencia de empresas.

Donde f_i =número de empresas que tienen sus ingresos en el tramo i.

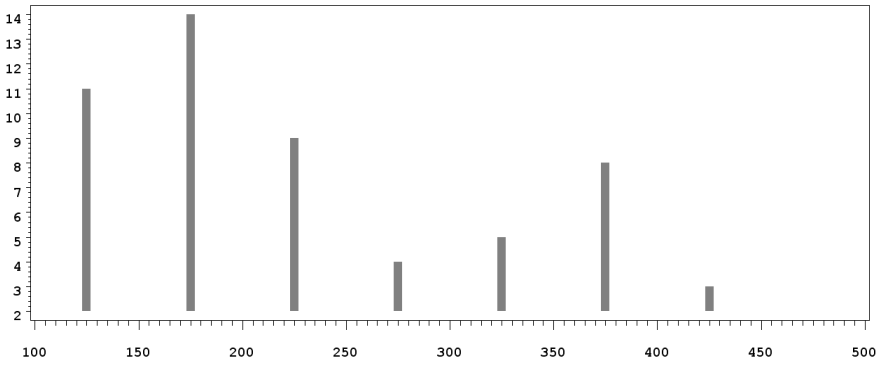


Figura 5.6. Número de empresas con ingresos en cada intervalo.

En este caso el número de intervalos de la tabla inicial de frecuencias $M = 8$ viene ya dado. Se trata de completar la tabla con los pasos 2 y 3.

ISBN 9788496866133

Ingreso en miles	f_i	$\sqrt{f_i}$	$\sqrt{f_i}^{ac}$
$100-150$	11	3.32	3.32
$150-200$	14	3.74	7.06
$200-250$	9	3.00	10.06
$250-300$	4	2.00	12.06
$300-350$	5	2.24	14.30
$350-400$	8	2.83	17.13
$400-450$	3	1.73	18.86
$450-500$	2	1.41	20.27

Tabla 5.4. Tabla de frecuencias

Así, $\sum_{i=1}^{8}\sqrt{f_i} = \sqrt{f_8}^{ac}$ =20.27. Como $k = \frac{1}{L}\sum_{i=1}^{M}\sqrt{f_i} = \frac{20.27}{3} = 6.76$, y los múltiplos de k son $(6.76, 13.52)$, observando donde caen estos valores en la columna de la raíz de frecuencia acumulada, los estratos estarán delimitados así: Estrato 1 (Ingresos de 100 a 200), Estrato 2 (Ingresos de 200 a 350), Estrato 3 (Ingresos de 350 a 500).

5.7 Tamaño de la muestra con m.a.s. por estrato

Nos centraremos en el caso en que se desea estimar la media poblacional, cuando el error absoluto de muestreo e y el nivel de confianza α están prefijados. Se desea conocer el tamaño muestral n para obtener ese nivel de error. Como la varianza del estimador depende de los tamaños n_h, hay que plantearse el problema en función de los diferentes tipos de afijación. Se suponen conocidas estimaciones piloto de S_h^2 para cada estrato h.

5.7.1 Afijación igual

Como $n_h = \dfrac{n}{L}$,

$$V(\overline{y}_{st}) = \sum_{h=1}^{L}\frac{W_h(LN_h-n)}{N}\frac{S_h^2}{n} = \sum_{h=1}^{L}LW_h^2\frac{S_h^2}{n} - \sum_{h=1}^{L}\frac{W_h}{N}S_h^2.$$

Así, como $e = z_{\alpha/2}\sqrt{V(\overline{y}_{st})}$, se tiene que

$$V(\overline{y}_{st}) = \frac{e^2}{z_{\alpha/2}^2} \Rightarrow \frac{e^2}{z_{\alpha/2}^2} = \sum_{h=1}^{L}LW_h^2\frac{S_h^2}{n} - \sum_{h=1}^{L}\frac{W_h}{N}S_h^2$$

ISBN 9788496866133

y entonces

$$n = \frac{L\sum_{h=1}^{L} W_h^2 S_h^2}{\frac{e^2}{z_{\alpha/2}^2} + \sum_{h=1}^{L} \frac{W_h}{N} S_h^2}$$

y por lo tanto

$$n_h = \frac{n}{L} = \frac{\sum_{h=1}^{L} W_h^2 S_h^2}{\frac{e^2}{z_{\alpha/2}^2} + \sum_{h=1}^{L} \frac{W_h}{N} S_h^2} = \frac{\sum_{h=1}^{L} N_h^2 S_h^2}{N^2 \frac{e^2}{z_{\alpha/2}^2} + \sum_{h=1}^{L} N_h S_h^2}.$$

5.7.2 Afijación proporcional

En este caso $n_h = nW_h$. Entonces

$$V(\overline{y}_{st}) = \sum_{h=1}^{L} \frac{W_h(N_h - nW_h)}{N} \frac{S_h^2}{nW_h} =$$

$$= \sum_{h=1}^{L} \frac{(N_h - nW_h)}{N} \frac{S_h^2}{n} = \sum_{h=1}^{L} W_h \frac{S_h^2}{n} - \sum_{h=1}^{L} \frac{W_h}{N} S_h^2 = \frac{e^2}{z_{\alpha/2}^2}$$

con lo cual

$$n = \frac{\sum_{h=1}^{L} W_h S_h^2}{\sum_{h=1}^{L} \frac{W_h}{N} S_h^2 + \frac{e^2}{z_{\alpha/2}^2}} = \frac{\sum_{h=1}^{L} N_h S_h^2}{\sum_{h=1}^{L} W_h S_h^2 + N\frac{e^2}{z_{\alpha/2}^2}}$$

y por lo tanto

$$n_h = nW_h = N_h \frac{\sum_{h=1}^{L} N_h S_h^2}{\sum_{h=1}^{L} N_h S_h^2 + \frac{e^2}{z_{\alpha/2}^2} N^2}.$$

5.7.3 Afijación de varianza mínima

Cálculos similares llevan a

$$n_h = \frac{N_h S_h (\sum_{h=1}^{L} N_h S_h)}{\sum_{h=1}^{L} N_h S_h^2 + \frac{e^2}{z_{\alpha/2}^2} N^2}.$$

ISBN 9788496866133

5.7.4 Afijación óptima con costes variables

Igualmente, se obtiene que

$$n_h = \frac{N_h S_h \sqrt{C_h}(\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{C_h}})}{\sum_{h=1}^{L} N_h S_h^2 + \frac{e^2}{z_{\alpha/2}^2} N^2}.$$

5.8 Postestratificación

A menudo no se dispone de la información para la realización de los estratos, o incluso conocidos estos, el estrato al que pertenece una unidad no se conoce a veces hasta después de realizado el muestreo. Este es el caso de características de carácter social y personal como religión, ideas políticas, etc. Las unidades solamente se pueden clasificar en estratos una vez conocidos los datos de la muestra. Supondremos N_h y por lo tanto W_h conocidos, pues pueden conocerse a partir de anteriores investigaciones u estadísticas oficiales.

En estos casos, se puede utilizar un diseño no estratificado para seleccionar la muestra de tamaño n, y una vez seleccionada, se puede a posteriori asignar las unidades de la muestra a los L estratos, y de este modo obtener mejores estimaciones. Se observa que en este caso los tamaños muestrales en cada estrato son variables aleatorias n_h , pues varían de una muestra a otra. Ello produce un aumento del error de estimación.

Supongamos ahora que la muestra se obtiene a través de m.a.s. pero una vez seleccionada, se decide estratificarla "a posteriori" y estimar la media poblacional $\overline{y}$ mediante una media post-estratificada $\overline{y}_{post.}$ Así, el tamaño muestral en el estrato h , n_h, es aleatorio antes de seleccionar la muestra y fijo una vez seleccionada. El estimador postestratificado será similar al estratificado:

$\overline{y}_{post} = \sum_{h=1}^{L} \frac{N_h}{N} \widehat{\overline{y}}_h$ donde $\widehat{\overline{y}}_h$ es la media muestral de y en el estrato h, creado a posteriori.

Este método puede llegar a ser casi tan preciso como el muestreo estratificado proporcional, siempre y cuando:

a) La muestra sea razonablemente grande.Si los estratos varían mucho en tamaño, la muestra debe ser más grande para garantizar que caen suficientes observaciones en cada estrato.

b) Las ponderaciones W_h no están exentas de errores, pues N_h suele ser una estimación, pero se supone que el nivel de error cometido en esta estimación es despreciable.

En concreto, se puede demostrar que

1) $\overline{y}_{post}$ es un estimador insesgado de $\overline{y}$.

2) $V(\overline{y}_{post}) \simeq \frac{1-f}{n} \sum_{h=1}^{L} W_h S_h^2 + \frac{1-f}{n^2} \sum_{h=1}^{L} (1 - W_h) S_h^2$

ISBN 9788496866133

En esta última expresión se observa que la varianza de $\overline{y}_{post}$ tiene dos sumandos: el primero es exactamente la varianza del estimador por muestreo estratificado con afijación proporcional. El segundo se puede considerar como la variabilidad adicional introducida por la postestratificación, siendo menor que el anterior término debido al n^2 en el denominador.

Se puede aplicar el siguiente estimador de la varianza en estimación post estratificada:

$$\widehat{V}(\overline{y}_{post}) \simeq \frac{1-f}{n}\sum_{h=1}^{L} W_h s_h^2 + \frac{1-f}{n^2}\sum_{h=1}^{L}(1-W_h)s_h^2.$$

Ejemplo 5.8.

Se observa una m.a.s. de 100 personas de una población que se supone igualmente dividida entre varones y mujeres. El objetivo es estimar el peso medio de la población. La muestra revela la siguente información:

Datos globales: $n = 100$, $\widehat{\overline{y}} = 66$ es la media muestral y por lo tanto el estimador estándar de la media del peso poblacional.

En esta muestra se observa que los datos desagregados consisten en:

Varones: $n_1 = 20$, $\widehat{\overline{y}}_1 = 90$.

Mujeres: $n_2 = 80$, $\widehat{\overline{y}}_2 = 60$.

Si se aplica la postestratificación, y considerando que hay igual número de varones y mujeres, $\frac{N_1}{N} = \frac{N_2}{N} = \frac{1}{2}$,

$$\overline{y}_{post} = \frac{N_1}{N}\widehat{\overline{y}}_1 + \frac{N_2}{N}\widehat{\overline{y}}_2 = \frac{1}{2}90 + \frac{1}{2}60 = 75$$

que es una estimación más fiel a la realidad, pues la estimación anterior $\widehat{\overline{y}} = 66$ estaba alterada debido a la descompensación en la muestra entre hombres y mujeres ($20\% - 80\%$) respecto a la distribución real ($50\% - 50\%$).

5.9 Unidades autorrepresentadas

En muchos casos prácticos existen en la población unidades que por sus características especiales deben ser excluidas del muestreo, tratadas como casos aparte, e incorporadas en la estimación final. La razón es que su valor en la característica de interés y está muy alejado del valor medio del resto de la población, y el hecho de que el azar en cualquier tipo de muestreo haga que estas unidades estén o no incluidas en la muestra genera una variabilidad muy grande en el estimador.

El método de escoger estas unidades a priori, medir su valor en y y realizar el muestreo en el resto de la población, incorporando adecuadamente al estimador el valor obtenido en las unidades autorrepresentadas, es perfectamente lícito y lleva a menudo a grandes mejoras en precisión.

ISBN 9788496866133

Una manera de ver las unidades autorrepresentadas es considerar cada una de ellas un estrato independiente, del cual se toman todas las unidades. La estimación del total consiste en sumar las estimaciones de los totales parciales en cada estrato, y la varianza total consiste en la suma ponderada de las varianzas de las estimaciones por estrato.

El siguiente resultado permite utilizar esta técnica en un caso general de muestreo.

Teorema 5.6 (unidades autorrepresentadas).

Sean $u_1, ..., u_k$ k unidades poblacionales sobre las cuales se ha medido a priori la variable de interés, obteniéndose los valores $y_1, ..., y_k$. Sean $u_{k+1,...,}u_N$ las unidades poblacionales restantes. Sea $(N-k)\widehat{\overline{y}}'$ un estimador insesgado del total $(N-k)\overline{y}'$ en la subpoblación $u_{k+1,...,}u_N$, con varianza $(N-k)^2V(\widehat{\overline{y}}')$ y estimador $(N-k)^2\widehat{V}(\widehat{\overline{y}}')$. Entonces

(a) $\widehat{N\overline{y}} = y_1 + ... + y_k + (N-k)\widehat{\overline{y}}'$ es un estimador insesgado del total $N\overline{y}$ en toda la población.

(b) $V(\widehat{N\overline{y}}) = (N-k)^2V(\widehat{\overline{y}}')$ con estimador $(N-k)^2\widehat{V}(\widehat{\overline{y}}')$.

(c) $\frac{1}{N}\widehat{N\overline{y}}$ es un estimador insesgado de la media poblacional, con varianza

$$\frac{(N-k)^2}{N^2}V(\widehat{\overline{y}}')$$

y estimador

$$\frac{(N-k)^2}{N^2}\widehat{V}(\widehat{\overline{y}}').$$

Demostración.

Al ser seleccionadas con probabilidad 1 cada una de las unidades poblacionales $u_1, ..., u_k$ no hay varianza en la estimación de su valor.

(a) $E(\widehat{N\overline{y}}) = E(y_1) + ... + E(y_k) + (N-k)E(\widehat{\overline{y}}') = y_1 + ... + y_k + (N-k)\overline{y}' = N\overline{y}$, pues $(N-k)\overline{y}'$ representa el total poblacional excluyendo las unidades $u_1, ..., u_k$.

(b)$V(\widehat{N\overline{y}}) = V(y_1) + ... + V(y_k) + (N-k)^2V(\widehat{\overline{y}}') = (N-k)^2V(\widehat{\overline{y}}')$.

(c) Es inmediato.

El caso (c) se puede aplicar también para estudios de proporciones, como se verá en un ejemplo más adelante.

Las unidades autorrepresentadas pueden darse en situaciones diversas como:

(i) Unidades cuyos valores en y son tan peculiares respecto al resto de la población que deben de ser tomadas en cuenta por separado, por razones teóricas de precisión en el estimador.

(ii) Unidades que tienen una importancia práctica especial (administrativa, de representación, etc.), y por su interés deben ser examinadas individualmente.

(iii) Unidades sobre las cuales se dispone de información concreta por anteriores estudios. Entonces se deben excluir del muestreo e incorporar su información en la estimación final como aparece en el Teorema.

ISBN 9788496866133

(iv) Un número reducido de casos que no estaban disponibles en el marco inicial de muestreo, y son tan importantes que es necesaria su medición e incorporación a la estimación para representar correctamente a la población objeto de estudio.

Ejemplo 5.9.

Se desea estimar el total de población de la provincia de Girona, tomando una muestra aleatoria de 10 municipios. En total hay 221 municipios. Pero la presencia o ausencia de Girona capital en la muestra puede dar lugar a estimaciones muy diferentes (dicho de otra manera,la varianza del estimador es muy alta debido a esa unidad elemental). Esto se ve reflejado en la Figura 5.7. Supongamos que se dispone de información mas o menos exacta del número de habitantes de Girona capital=71858.

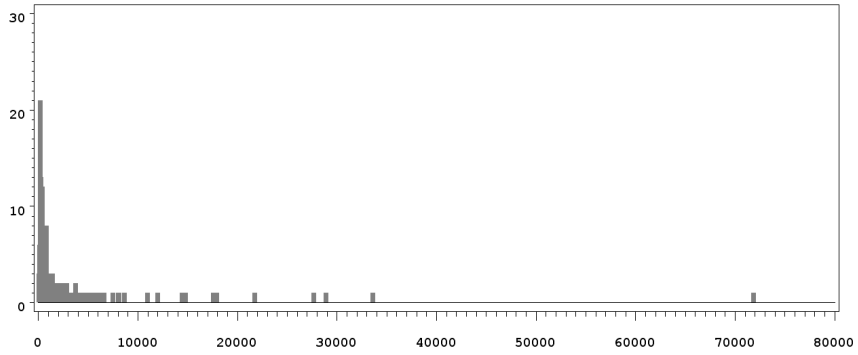


Figura 5.7. Histograma del número de habitantes por municipio en la provincia de Girona.

Por lo tanto, se excluye del muestreo la capital, y se extrae una m.a.s. de tamaño $n = 10$ de entre los $N' = N - k = 221 - 1 = 220$ municipios que quedan después de excluir la capital. Supongamos que la media muestral obtenida en esa m.a.s. es $\widehat{\overline{y}}' = 1925$. Se sabe que $(N-1)\widehat{\overline{y}}'$ es un estimador insesgado del total en la subpoblación provincia de Girona excluyendo Girona capital, y el valor para nuestra muestra es $(N-1)\widehat{\overline{y}}' = 220 \cdot 1925 = 423500$. El estimador del total para toda la provincia es $y_{Girona-capital} + (N-1)\widehat{\overline{y}}' = 71858 + 423500 = 495358$.

Se ha calculado también en la muestra $s^{2\prime} = 7189584$. Una estimación de la varianza del estimador del total para toda la población es

$$(N-1)^2 \widehat{V}(\widehat{\overline{y}}') = 220^2 \frac{220-10}{220} \frac{7189584}{10} = 3.32 \cdot 10^{10},$$

con una desviación típica de 182252.

Además, una estimación de la media de habitantes por municipio es $\frac{1}{221} 495358 = 2241$ con una varianza estimada para el estimador de

$$\frac{3.32 \cdot 10^{10}}{221^2} = 680080.$$

Suponiendo normalidad del estimador, un intervalo de confianza al 95% para la población media por municipio es $(624.6, 3857.3)$.

Si se hubiera realizado el muestreo aleatorio simple de $n = 10$ unidades incluyendo la capital, recordemos que la probabilidad de inclusión de cada unidad en m.a.s. es de $\pi_i = \frac{n}{N}$. Por lo tanto

© N.J. Portela, M. Villeta
ISBN 9788496866133

Girona capital solamente tendría una probabilidad de $\frac{10}{221} = 0.045$ de caer en la muestra, con lo que lo más probable es que no cayera en ella, y que por lo tanto el valor del estimador de la media de habitantes por municipio subestimara el verdadero valor (que en realidad es $\overline{y} = 2457$).

Ejemplo 5.10.

En un estudio de muestreo sobre la presencia de glaucoma (una enfermedad ocular que puede degenerar en ceguera) en la provincia de Segovia, muchos de los potenciales enfermos de la población objetivo ya habían sido diagnosticados en hospitales, con lo cual no era necesario tenerlos en cuenta para el muestreo. Sin embargo incorporar sus diagnósticos a la estimación realizada en el resto de la población era necesario para una estimación correcta de la prevalencia de la enfermedad en la provincia.

Para simplificar, considérese la población de la Figura 5.8, donde los diagnosticados como enfermos de glaucoma están en color negro y los sanos en color blanco.

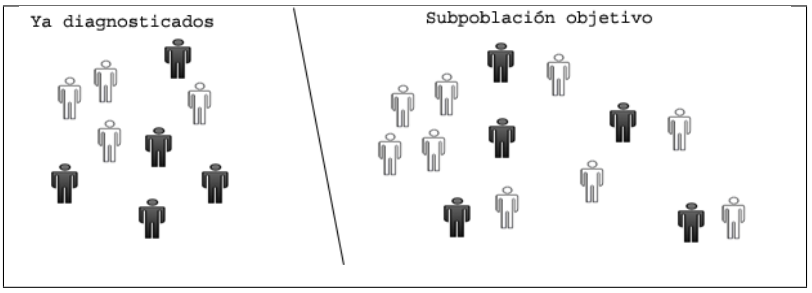


Figura 5.8. Población con un subconjunto previamente diagnosticado.

El muestreo se reducirá a la subpoblación objetivo, que es el conjunto de los no diagnosticados previamente. Después se incorporará la información de los ya diagnosticados. Supongamos que se extrae una muestra aleatoria simple sin reemplazamiento de tamaño $n = 5$ de la subpoblación objetivo y caen en la muestra 3 enfermos de glaucoma y 2 pacientes sanos. Al ser un estudio de proporciones, la variable de interés es $y = 1$ si el paciente tiene diagnóstico positivo de glaucoma e $y = 0$ si no. La proporción muestral es $\widehat{p}' = \dfrac{3}{5}$. Entonces el estimador del total en la subpoblación objetivo es $(N-k)\widehat{\overline{y}}' = (N-k)\widehat{p}' = (23-9)\dfrac{3}{5} = 8.4$. Incorporando la información conocida del resto de la población, la estimación del total de enfermos en toda la población es $5 + 8.4 = 13.4$. La varianza estimada del estimador del total será

$$(N-k)^2\widehat{V}(\widehat{p}') = 14^2\frac{14-5}{14}\frac{0.6\cdot 0.4}{5-1} = 7.56.$$

El estudio pretende estimar la prevalencia de la enfermedad en toda la población y dar una estimación de la precisión de la estimación. Como el total estimado es 13.4, la estimación de la proporción poblacional (la prevalencia de la enfermedad) será $\widehat{p} = \dfrac{13.4}{23} = 0.58$.

La varianza estimada de esta proporción será

$$\frac{(N-k)^2}{N^2}\widehat{V}(\widehat{p}') = 0.01429.$$

ISBN 9788496866133

5.10 Tablas de fórmulas

MUESTREO ESTRATIFICADO CON m.a.s. EN CADA ESTRATO			
Parámetro poblacional	$\overline{y}$	$N\overline{y}$	p
Estimador	$\overline{y}_{st} = \sum_{h=1}^{L} \frac{N_h}{N}\widehat{\overline{y}}_h$	$N\overline{y}_{st}$	$\widehat{p} = \sum_{h=1}^{L} \frac{N_h}{N}\widehat{p}_h$
Varianza	$\sum_{h=1}^{L} \frac{N_h(N_h-n_h)}{N^2}\frac{S_h^2}{n_h}$	$\sum_{h=1}^{L} N_h(N_h-n_h)\frac{S_h^2}{n_h}$	$\sum_{h=1}^{L} W_h^2\frac{N_h-n_h}{N_h-1}\frac{p_hq_h}{n_h}$
Estimador de la Varianza	$\sum_{h=1}^{L} \frac{N_h(N_h-n_h)}{N^2}\frac{s_h^2}{n_h}$	$\sum_{h=1}^{L} N_h(N_h-n_h)\frac{s_h^2}{n_h}$	$\sum_{h=1}^{L} W_h^2\frac{N_h-n_h}{N_h}\frac{\widehat{p}_h\widehat{q}_h}{n_h-1}$

MUESTREO POSTESTRATIFICADO CON m.a.s. global			
Parámetro poblacional	$\overline{y}$	$N\overline{y}$	p
Estimador	$\overline{y}_{post} = \sum_{h=1}^{L} \frac{N_h}{N}\widehat{\overline{y}}_h$	$N\overline{y}_{post}$	$\sum_{h=1}^{L} \frac{N_h}{N}\widehat{p}_h$
Varianza aprox.	$\frac{1-f}{n}\sum_{h=1}^{L} W_hS_h^2 + \frac{1-f}{n^2}\sum_{h=1}^{L}(1-W_h)S_h^2$	$NV(\overline{y}_{post})$	$V(\overline{y}_{post})$
Estimador de la Varianza	$\frac{1-f}{n}\sum_{h=1}^{L} W_hs_h^2 + \frac{1-f}{n^2}\sum_{h=1}^{L}(1-W_h)s_h^2$	$N\widehat{V}(\overline{y}_{post})$	$\widehat{V}(\overline{y}_{post})$

ISBN 9788496866133

AFIJACIONES Y TAMAÑOS MUESTRALES CON e PREFIJADO		
Afijación		Tamaño muestral para estimar la media dado e
Igual	$n_h = \dfrac{n}{L}$	$n_h = \dfrac{\sum_{h=1}^{L} N_h^2 \widehat{S}_h^2}{N^2 \dfrac{e^2}{z_{\alpha/2}^2} + \sum_{h=1}^{L} N_h \widehat{S}_h^2}$
Proporcional	$n_h = n \dfrac{N_h}{N}$	$n_h = N_h \dfrac{\sum_{h=1}^{L} N_h \widehat{S}_h^2}{\sum_{h=1}^{L} N_h \widehat{S}_h^2 + \dfrac{e^2}{z_{\alpha/2}^2} N^2}$
Varianza mínima	$n_h = n \dfrac{N_h \widehat{S}_h}{\sum_{h=1}^{L} N_h \widehat{S}_h}$	$n_h = \dfrac{N_h \widehat{S}_h (\sum_{h=1}^{L} N_h S_h)}{\sum_{h=1}^{L} N_h \widehat{S}_h^2 + \dfrac{e^2}{z_{\alpha/2}^2} N^2}$
Óptima	$n_h = n \dfrac{\dfrac{N_h \widehat{S}_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} \dfrac{N_h \widehat{S}_h}{\sqrt{C_h}}}$	$n_h = \dfrac{N_h \widehat{S}_h \sqrt{C_h} (\sum_{h=1}^{L} \dfrac{N_h \widehat{S}_h}{\sqrt{C_h}})}{\sum_{h=1}^{L} N_h \widehat{S}_h^2 + \dfrac{e^2}{z_{\alpha/2}^2} N^2}$

Notas:

Para el estudio de la estimación de la proporción p, se sustituirá en cada caso $\widehat{S}_h^2$ por

$$\frac{N_h}{N_h - 1} \widehat{p}_h (1 - \widehat{p}_h).$$

Para el estudio de la estimación del total, se sustituirá $\widehat{S}_h$ por $N \widehat{S}_h$.

Para el estudio del error de muestreo $\phi = \sqrt{V(\overline{y}_{st})}$, se sustituirá $\dfrac{e^2}{z_{\alpha/2}^2}$ por ϕ^2.

Para el estudio del error de muestreo relativo, $\phi = \dfrac{\sqrt{V(\overline{y}_{st})}}{\overline{y}_{st}}$, se sustituirá $\dfrac{e^2}{z_{\alpha/2}^2}$ por $\phi^2 \cdot \overline{y}_{st}^2$.

ISBN 9788496866133

5.11 Obtención de muestras por muestreo estratificado con SAS

Supongamos que se desea extraer muestras por muestreo estratificado con m.a.s. en cada estrato de un archivo SAS. Utilizaremos para ello el procedimiento surveyselect. Supongamos que el archivo datos contiene la variable indicativa de estrato, llamaa estrato, y toma valores 1,2,3,...indicativos del estrato.

Afijación igual

Supongamos que se desean extraer exactamente $n_h = 5$ unidades de cada estrato. Para esta afijación la sintaxis es sencilla:

```
proc surveyselect data=datos out=muestra sampsize=5;
strata estrato;
run;
```

Afijación proporcional

Es la opción por defecto que utiliza el proc surveyselect. Con afijación proporcional, $\frac{n}{N} = \frac{n_h}{N_h}$, y este valor de proporción hay que indicarlo enla opción samprate al ejecutar el procedimiento. Supongamos que en el archivo datos hay $N = 50$ observaciones en total, y que se desea afijación proporcional con $n = 10$. Entonces $\frac{n}{N} = 0.2$.

La sintaxis para obtener la muestra es la siguiente:

```
proc surveyselect data=datos out=muestra samprate=0.2;
strata estrato;
run;
```

Otros tipos de Afijación

En este caso hay que indicar en el proc surveyselect el archivo que contiene la información del tamaño muestral requerido en cada estrato (si la afijación es de varianza mínima hay que haber calculado previamente los valores n_h). El archivo debe contener la variable de estrato y la variable _nsize_, que contiene los valores n_h.

Suponiendo que tenemos en el archivo datos 3 estratos, y queremos $n_1 = 3$, $n_2 = 6$, $n_3 = 2$,se puede crear previamente el archivo (que llamaremos afijaciones) en un paso data, de la siguiente manera:

```
data afijacion;
input estrato _nsize_;
cards;
1 3
2 6
3 2
;
```


ISBN 9788496866133

Una vez creado el archivo, se utiliza en el proc surveyselect en la opción sampsize.

Para ejecutar este procedimiento correctamente en el caso de muestreo estratificado con afijación aportada en archivo, es rigurosamente necesario que el archivo poblacional (datos) esté previamente ordenado por la variable de estrato, y en el mismo orden que el archivo de afijaciones. Esto se puede realizar con el proc sort del SAS.

La sintaxis final para seleccionar la muestra quedaría:

```
proc sort data=datos;by estrato;run;
proc surveyselect data=datos out=muestra method=srs sampsize=afijacion;
strata estrato;
run;
```

5.12 Estimación en muestreo estratificado o post-estratificado con SAS

5.12.1 Estimación con la macro estimestrat

El proc surveymeans requiere diferentes sintaxis para diferentes tipos de afijación, y diferentes estimaciones (medias o totales). Para simplificar las cosas, presentamos una macro que calcula las estimaciones.

En cualquier caso, previamente a la utilización de la macro pueden tenerse los tamaños poblacionales por estrato en un archivo SAS, muy similar al archivo de afijaciones estudiado en la anterior sección. Se puede crear con un paso data. Supongamos que en la población hay 3 estratos, con tamaños poblacionales respectivos 20,10 y 20:

```
data tamas;
input estrato Nh;
cards;
1 20
2 10
3 20
;
```

Esto es opcional, pues también se pueden tener los tamaños directamente en el archivo muestral.

La sintaxis de la macro es

```
estimestrat(muestra,archivo2,vary,vartama,post,indicador,nestratos,N);
```

donde

muestra es el archivo que contiene la muestra.

ISBN 9788496866133

archivo2 es el archivo que contiene la información del tamaño de los estratos(opcional).

vary es la variable de interés.

vartama es la variable que indica el tamaño de los estratos

post

1 Si la muestra proviene de muestreo estratificado con m.a.s. en cada estrato.

2 Si la muestra proviene de un m.a.s. (post-estratificación).

indicador

1 si el tamaño de los estratos está indicado en el archivo2 en la variable vartama.

2 si el tamaño de los estratos está en la variable vartama en el archivo muestra .

nestratos es el número de estratos.

N es el tamaño poblacional.

En nuestro ejemplo, suponiendo que los datos muestrales provienen de un diseño estratificado, y están en el archivo llamado muestra5, sería

```
%estimestrat(muestra5,tamas,y,nh,1,1,3,50);
```

La macro presenta en la ventana LOG los estimadores y varianzas estimadas para la media o proporción y total, y los intervalos de confianza respectivos al 95%.

ISBN 9788496866133

5.13 Ejercicios resueltos

Ejercicio 4.1.

Se realiza un estudio sobre las notas obtenidas en matemáticas en tercer curso en las dos escuelas que hay en una pequeña población. En la primera hay 300 niños y en la segunda 600 niños. Se obtiene, utilizando afijación igual con tamaño total $n = 50$, una muestra de niños en cada escuela por m.a.s. Los datos obtenidos están en la tabla:

Escuela	$\widehat{\overline{y}}_h$	s_h^2
I	5.7	1.2
II	6.8	1.6

a) Suponiendo normalidad, plantear un intervalo de confianza al 95% para la media de la nota en matemáticas en toda la población. Plantearlo para la Escuela I.

b) Utilizando la información de esta encuesta, ¿qué tamaño muestral en total y por estrato habría que tomar con afijación proporcional para obtener un error de muestreo absoluto menor de 0.15 con $\alpha = 0.05$? ¿y con afijación de varianza mínima?

---

a) Como $N = 300 + 600 = 900$, la estimación es

$$\overline{y}_{st} = \sum_{h=1}^{L} \frac{N_h}{N} \widehat{\overline{y}}_h = \frac{300}{900} 5.7 + \frac{600}{900} 6.8 = 6.43.$$

Como el tamaño total es $n = 50$, y es afijación igual, $n_I = 25$ y $n_{II} = 25$.La varianza estimada del estimador es

$$\widehat{V}(\overline{y}_{st}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2} \frac{s_h^2}{n_h} = \frac{300(300-25)}{900^2} \frac{1.2}{25} + \frac{600(600-25)}{900^2} \frac{1.6}{25} = 0.032.$$

El intervalo de confianza al 95% para la media poblacional será

$$(6.43 - 1.96\sqrt{0.032}, 6.43 + 1.96\sqrt{0.032}) = (6.07, 6.78).$$

Para la Escuela I, como se ha realizado m.a.s. de tamaño $n_I = 25$, es

$V(\widehat{\overline{y}}_I) = \dfrac{N_I - n_I}{N_I} \dfrac{s_I^2}{n_I} = \dfrac{300 - 25}{300} \dfrac{1.2}{25} = 0.044$ y por lo tanto, el intervalo es ()

$$(5.7 - 1.96\sqrt{0.044}, 5.7 + 1.96\sqrt{0.044}) = (5.28, 6.11).$$

b) Con afijación proporcional, aproximando las cuasivarianzas por estrato por las cuasivarianzas muestrales obtenidas en la encuesta, es

$$n_I = N_I \frac{\sum_{h=1}^{L} N_h S_h^2}{\sum_{h=1}^{L} N_h S_h^2 + \dfrac{e^2}{z_{\alpha/2}^2} N^2} = 300 \frac{300 \cdot 1.2 + 600 \cdot 1.6}{300 \cdot 1.2 + 600 \cdot 1.6 + \dfrac{0.15^2}{1.96^2} 900^2} = 65.3$$

y como

ISBN 9788496866133

$n_h = n\dfrac{N_h}{N}$ para todo h, entonces $n = n_h\dfrac{N}{N_h} = 196$ y $n_{II} = n\dfrac{N_{II}}{N} = 196\dfrac{600}{900} = 130.66$. Así, se tomarán $n_I = 65$ observaciones en la escuela I y $n_{II} = 131$ observaciones en la escuela II, con un amaño total de $n = 196$.

Con afijación de varianza mínima,

$$n_h = \frac{N_h S_h(\sum_{h=1}^{L} N_h S_h)}{\sum_{h=1}^{L} N_h S_h^2 + \dfrac{e^2}{z^2_{\alpha/2}} N^2}$$

Calculando los términos:

$$\sum_{h=1}^{L} N_h S_h = 300 \cdot \sqrt{1.2} + 600 \cdot \sqrt{1.6} = 1087.58.$$

$$\sum_{h=1}^{L} N_h S_h^2 = 300 \cdot 1.2 + 600 \cdot 1.6 = 1320.$$

Así,

$$n_I = \frac{N_I S_I(\sum_{h=1}^{L} N_h S_h)}{\sum_{h=1}^{L} N_h S_h^2 + \dfrac{e^2}{z^2_{\alpha/2}} N^2} = \frac{300 \cdot \sqrt{1.2}(1087.58.)}{1320 + \dfrac{0.15^2}{1.96^2} 900^2} = 300 \cdot \sqrt{1.2} \cdot 0.179 = 59.$$

y

$n_{II} = 600 \cdot \sqrt{1.6} \cdot 0.179 = 136.$

Con lo cual $n = n_I + n_{II} = 195$.

Ejercicio 4.2.

En una encuesta sobre hábitos de dormir (número de horas promedio al días) realizada a hombres y mujeres en una empresa, con m.a.s. en cada estrato, había inicialmente 35 mujeres y 56 hombres. Se obtuvieron los siguientes datos:

Estrato	$\widehat{\overline{y}}_h$	s_h^2	n_h
Hombres	7.5	0.36	8
Mujeres	7	0.5	5

a) ¿Qué afijación se utilizó? ¿Mejoraría mucho la estimación una afijación distinta?

b) Calcular un Intervalo de Confianza al 95% bajo el supuesto de normalidad, para el promedio de horas dormidas al día en general entre los trabajadores de esa empresa.

c) ¿Qué tamaño muestral se necesitaría para estimar con un error de muestreo de 0.05 la media poblacional del número de horas que duermen en promedio los trabajadores de la empresa, utilizando afijación proporcional?.

---

a) Al ser $\dfrac{n_1}{N_1} = \dfrac{8}{56} = \dfrac{n_2}{N_2} = \dfrac{5}{35} = \dfrac{1}{7}$, se trata de afijación proporcional.

ISBN 9788496866133

El error de muestreo estimado al estimar la media es, en este caso,

$$\widehat{V}(\overline{y}_{st}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2} \frac{s_h^2}{n_h} = \frac{56(56-8)}{91^2}\frac{0.36}{8} + \frac{35(35-5)}{91^2}\frac{0.5}{5} = 0.0272.$$

Suponiendo exactas las estimaciones de las cuasivarianzas, y con el mismo tamaño muestral de $n = 13$, si se utilizara afijación de varianza mínima, los tamaños muestrales serían :

$n_1 = n\dfrac{N_1\widehat{S}_1}{\sum_{h=1}^{L} N_1\widehat{S}_1} = 13\dfrac{56 \cdot 0.6}{56 \cdot 0.6 + 35 \cdot 0.707} = 7.48$ y $n_2 = 13 - 7.48 = 5.51$ con lo cual se tomarían $n_1 = 7$ y $n_2 = 6$.

El error de muestreo aproximado sería, en este caso,

$\widehat{V}(\overline{y}_{st}) = \dfrac{56(56-8)}{91^2}\dfrac{0.36}{7} + \dfrac{35(35-5)}{91^2}\dfrac{0.5}{6} = 0.02725$ muy parecido al obtenido con afijación proporcional.

Con afijación igual en este caso sería puesto que n es impar aproximadamente un error de muestreo similar, $n_1 = 7$ y $n_2 = 6$.

b) Como

$$\overline{y}_{st} = \sum_{h=1}^{L} \frac{N_h}{N}\widehat{\overline{y}}_h = \frac{56}{91}7.5 + \frac{35}{91}7 = 7.30$$

Se tiene que el intervalo es

$$(7.30 - 1.96\sqrt{0.0272}, 7.30 + 1.96\sqrt{0.0272}) = (6.97, 7.62).$$

c) Con afijación proporcional, y $\phi = 0.05$,

$$n_h = N_h \frac{\sum_{h=1}^{L} N_h\widehat{S}_h^2}{\sum_{h=1}^{L} N_h\widehat{S}_h^2 + \phi^2 N^2}$$

Como

$$\sum_{h=1}^{L} N_h S_h^2 = 56 \cdot 0.36 + 35 \cdot 0.5 = 37.66$$

y, por lo tanto,

$$n_1 = 56\frac{37.66}{37.66 + 0.05^2 91^2} = 36.13$$

y al ser afijación proporcional, $n_2 = N_2\dfrac{n_1}{N_1} = 22.5$.

Se necesitarían aproximadamente $n_1 = 37$ y $n_2 = 23$ observaciones (n=60 en total).

Ejercicio 4.3.

Se estudia la proporción de vacas de una cierta raza en dos grandes rebaños. El primero tiene aproximadamente 100 vacas, y el segundo 200. Se toma una m.a.s. de 20 vacas del primero y de 45 del segundo. Se observan, en el primero, 8 vacas de la raza en cuestión, y en el segundo, 15.

ISBN 9788496866133

a) Dar un intervalo de confianza al 95% para la proporción de vacas de la raza en cuestión en conjunto en los dos rebaños.

b) Dar un intervalo de confianza al 95% para el total de vacas de la raza en cuestión en conjunto en los dos rebaños.

c)¿Cuál es el error de muestreo relativo aproximado al estimar la proporción? ¿y al estimar el total?

¿Qué tamaño muestral se debería tener para obtener un error de muestreo relativo de 0.05 para estimar la proporción utilizando afijación proporcional?.

---

a) $\widehat{p}_1 = \dfrac{8}{20} = 0.4$ y $\widehat{p}_2 = \dfrac{15}{45} = 0.33$. Por lo tanto,

$$\widehat{p}_{st} = \frac{100}{300}0.4 + \frac{200}{300}0.33 = 0.355$$

Los pesos respectivos de cada estrato son $\dfrac{1}{3}$ y $\dfrac{2}{3}$.

y

$$\widehat{V}(\widehat{p}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h} \frac{\widehat{p}_h \widehat{q}_h}{n_h - 1} = \frac{1}{3^2}\frac{100-20}{100}\frac{0.4\cdot 0.6}{20-1} + \frac{2^2}{3^2}\frac{200-45}{200}\frac{0.33\cdot 0.66}{45-1} = 0.00286.$$

Así, el intervalo es

$$(0.355 - 1.96\sqrt{0.00286}, 0.355 + 1.96\sqrt{0.00286}) = (0.25, 0.46).$$

b) La estimación del total es $N\widehat{p}_{st} = 106.5$, y su varianza $N^2 V(\widehat{p}_{st}) = 9585000$. El intervalo será:

$$(75, 138).$$

c) El error de muestreo relativo aproximado estimando la proporción es $\dfrac{\sqrt{V(\widehat{p}_{st})}}{\widehat{p}_{st}} = 0.15$.

Estimando el total, es $\dfrac{\sqrt{N^2 V(\widehat{p}_{st})}}{N\widehat{p}_{st}} = 0.15$ (es el mismo).

Con afijación proporcional, el tamaño muestral para cada estrato será :

$$n_h = N_h \frac{\sum_{h=1}^{L} N_h \widehat{S}_h^2}{\sum_{h=1}^{L} N_h \widehat{S}_h^2 + \dfrac{e^2}{z_{\alpha/2}^2} N^2}$$

donde $\widehat{S}_h^2$ se sustituye por $\dfrac{N_h}{N_h - 1}\widehat{p}_h(1-\widehat{p}_h)$ y $\dfrac{e^2}{z_{\alpha/2}^2}$ por $\phi \cdot \widehat{p}_{st}$.

Se calcula $\sum_{h=1}^{L} N_h \widehat{S}_h^2 = 100\dfrac{100}{100-1}0.4(1-0.4) + 200\dfrac{200}{200-1}0.33(1-0.33) = 68.91$

Para el primer estrato, es:

$$n_1 = 100\frac{68.91}{68.91 + (0.05\cdot 0.355)^2 \cdot 300^2} = 10.8.$$

y, por ser afijación proporcional, $n_2 = \dfrac{n_1}{N_1} N_2 = 21.66$.

ISBN 9788496866133

Con lo cual se tomaría $n_1 = 11$ y $n_2 = 22$, con un total de $n = 33$ observaciones.

Ejercicio 4.4.

Se ha estudiado el número de margaritas que crecen en las dos áreas que conforman un jardín. Cada área es dividida en 20 parcelas, y se seleccionan 5 parcelas al azar en cada área. Los números de margaritas obtenidos en cada una de esas parcelas son 15,40,10,25,30 en el área 1, y 40,50,35,20,25 en las parcelas de la segunda área.

a) Dar un intervalo de confianza para el total de margaritas en el jardín.

b) Realizar los cálculos en SAS con la ayuda de la macro estimestrat .

---

a) En el área 1, se tiene $\widehat{\overline{y}}_1 = 24$ y $s_1^2 = 142.5$. En el área 2, $\widehat{\overline{y}}_1 = 34$ y $s_2^2 = 142.5$.

entonces, $\overline{y}_{st} = \dfrac{20}{40}\,\widehat{\overline{y}}_1 + \dfrac{20}{40}\,\widehat{\overline{y}}_2 = 29$ y $N\overline{y}_{st} = 1160$.

$\widehat{V}(\overline{y}_{st}) = \sum_{h=1}^{L} \dfrac{N_h(N_h - n_h)}{N^2}\dfrac{s_h^2}{n_h} = \dfrac{20(20-5)}{40^2}(2\dfrac{142.5}{5}) = 10.6875$ y la varianza del total será $N^2\widehat{V}(\overline{y}_{st}) = 17100$.

El intervalo de confianza para el total será

$(1160 - 1.96\sqrt{17100}, 1160 + 1.96\sqrt{17100}) = (903.7, 1416.3)$.

b) Con SAS, se introducen los datos en un archivo:

```
data margarita;
input estrato marga tama;
cards;
1 15 20
1 40 20
1 10 20
1 25 20
1 30 20
2 40 20
2 50 20
2 35 20
2 20 20
2 25 20
;
```

La macro estimestrat se ejecuta así:

```
%estimestrat(margarita,.,marga,tama,2,2,40);
```

Obteniendo el mismo resultado.

Otra forma de crear el archivo margarita son:

ISBN 9788496866133

```
data margarita;
do i=1 to 5;estrato=1;tama=20;input marga @;output;end;
do i=1 to 5;estrato=2;tama=20;input marga @;output;end;
cards;
15 40 10 25 30 40 50 35 20 25
;
```

Ejercicio 4.5.

Se desea construir un diseño estratificado para estudiar la variable ventas mensuales de determinado artículo en una empresa que dispone de 1340 franquicias. Se dispone de información sobre las ventas de artículos parecidos, que está indicada en la tabla, agrupadas en frecuencias .

nº Ventas	nº de franquicias
0-50	100
50-100	150
100-150	200
150-200	300
200-250	210
250-300	180
300-350	100
350-400	100

Presentar cual sería la configuración en dos estratos según la regla de la frecuencia acumulada. Hacer lo mismo con 3 estratos.

Se añaden a la tabla las columnas correspondientes a la raíz de la frecuencia de la raíz de frecuencia, acumulada.

ISBN 9788496866133

nº Ventas	nº de franquicias	$\sqrt{f_i}$	$\sqrt{f_i}^{ac}$
0-50	100	10	10
50-100	150	12.24	22.24
100-150	200	14.14	36.38
150-200	300	17.32	53.7
200-250	210	14.49	68.19
250-300	180	13.41	81.6
300-350	100	10	91.6
350-400	100	10	101.6

Para dos estratos, se calcula $k = \frac{1}{L}\sum_{i=1}^{M}\sqrt{f_i} = \frac{101.6}{2} = 50.8$. El $L-1=1$ múltiplos de $k=50.8$ es 50.8 .Por lo tanto el punto de corte es el intervalo superior que contiene a 50.8 en la columna de raíz de frecuencia acumulada: el intervalo 150-200. Por lo tanto el primer estrato constará de las empresas que tengan ventas de 0 a 200 en el artículo parecido (con el que se ha construido la tabla), y el estrato 2 constará de las empresas restantes (de 200 a 400 ventas).

Para un estrato, $k = \frac{101.6}{3} = 33.8$. Los $L-1=2$ múltiplos de k son 33.8 y 67.7. Por lo tanto el primer estrato estará formado por las empresas con ventas de 0 a 150, el segundo de aquellas con ventas de 150 a 250, y el tercero con las restantes.

Ejercicio 4.6.

Se desea estimar la nota media de los alumnos en una escuela de 800 alumnos, donde hay 500 mujeres y 300 hombres. Se toma una m.a.s. de 20 alumnos y se sospecha que la nota media de las mujeres es superior a la de los hombres, por lo que se anota el sexo de los encuestados. Se obtienen los siguientes datos en la muestra: hubo 6 mujeres, con notas medias

6.5 5 7 7.5 8 5

y 14 hombres, con notas medias

4.5 5 6 8 4 6.5 7 8 5 4 5 6 5 5.5

a) Si no se tiene en cuenta el sexo, dar la estimación de la nota promedio en la escuela y el intervalo de confianza al 95%.

b) Hacer lo mismo teniendo en cuenta el sexo.

---

a) Como se trata de una m.a.s., hay que aplicar la estimación usual en m.a.s.:

$\widehat{\overline{y}} = 5.925$, con $s^2 = 1.74$. La varianza estimada del estimador es

$\widehat{V}(\widehat{\overline{y}}) = \frac{N-n}{N}\frac{s^2}{n} = \frac{800-20}{800}\frac{1.74}{20} = 0.085$. El intervalo de confianza es

ISBN 9788496866133

$(5.925 - 1.96\sqrt{0.085}, 5.925 + 1.96\sqrt{0.085}) = (5.35, 6.49)$.

b) Al no ser muestreo estratificado, pero conocer los tamaños poblacionales de los estratos, se puede recurrir a la estimación por post-estratificación.

La media muestral en mujeres es $\widehat{\overline{y}}_1 = 6.5$, y en hombres es $\widehat{\overline{y}}_1 = 5.67$.

El estimador global será

$\overline{y}_{post} = \sum_{h=1}^{L} \frac{N_h}{N} \widehat{\overline{y}}_h = \frac{500}{800}\widehat{\overline{y}}_1 + \frac{300}{800}\widehat{\overline{y}}_2 = \frac{500}{800}6.5 + \frac{300}{800}5.67 = 6.18$, distinto del estimador $\widehat{\overline{y}} = 5.925$.

El estimador de la varianza en estimación por post estratificación es

$$\widehat{V}(\overline{y}_{post}) = \frac{1-f}{n} \sum_{h=1}^{L} W_h s_h^2 + \frac{1-f}{n^2} \sum_{h=1}^{L} (1 - W_h) s_h^2$$

como $s_1^2 = 1.58$ y $s_2^2 = 1.69$, y $f = \frac{800 - 20}{800} = 0.975$

$$\widehat{V}(\overline{y}_{post}) = \frac{1-f}{n} \sum_{h=1}^{L} W_h s_h^2 + \frac{1-f}{n^2} \sum_{h=1}^{L} (1 - W_h) s_h^2 =$$

$=\frac{0.025}{20}[\frac{300}{800}1.58 + \frac{500}{800}1.69] + \frac{0.025}{20^2}[\frac{500}{800}1.58 + \frac{300}{800}1.69] = 0.0021$, sensiblemente menor que la varianza sin tener en cuenta los posibles estratos (aunque hay que tener en cuenta que en ambos casos se trata de varianzas estimadas).

El intervalo de confianza en este caso será

$(6.18 - 1.96\sqrt{0.0021}, 6.18 + 1.96\sqrt{0.0021}) = (6.09, 6.26)$.

Ejercicio 4.7

En un estudio médico en un hospital se desea saber el número de mujeres que han tenido hijos alguna vez. De las 200 pacientes ingresadas, hay 30 de las cuales ya se conoce la respuesta por tenerla registrada en los ordenadores: 18 mujeres de esas 30 han tenido hijos. En el resto de mujeres de las que no se tiene información es necesario hacer un muestreo. Se escogen por m.a.s. 20 mujeres y se cuentan 9 con hijos.

Estimar el total de mujeres con hijos en el hospital, y la proporción de mujeres con hijos. Dar un intervalo de confianza al 95% para ambas cantidades.

En la subpoblación muestreada (N' =170 mujeres), la proporción muestral es $\widehat{p}' = \frac{9}{20} = 0.45$. En esa misma subpoblación, se estima por lo tanto en $N'0.45 = 170 \cdot 0.45 = 76.5$ el total de mujeres con hijos. La estimación de este total tiene una varianza estimada de

$$170^2 \widehat{V}(\widehat{p}') = 170^2 \frac{N' - n}{N'} \frac{\widehat{p}'\widehat{q}'}{n-1} = 170^2 \frac{170 - 20}{170} \frac{0.45 \cdot 0.55}{20 - 1} = 332.17.$$

En todo el hospital, se estimará en $18 + 76.5 = 94.5$ el total de mujeres con hijos. Esta estimación tiene una varianza de 332.17, pues la cantidad 18 añadida no aporta error de muestreo. Por lo tanto el intervalo de confianza para el total de mujeres con hijos es

$(94.5 - 1.96\sqrt{332.17}, 94.5 + 1.96\sqrt{332.17}) = (58.77, 130.22)$.

ISBN 9788496866133

Para la estimación de la proporción, se parte de la estimación anterior del total: $\widehat{p} = \dfrac{94.5}{200} = 0.4725$ y la varianza de este estimador es

$$\frac{1}{200^2} 332.17 = 0.0083.$$

El intervalo de confianza para la proporción es entonces:

$$(0.4725 - 1.96\sqrt{0.0083}, 0.4725 - 1.96\sqrt{0.0083}) = (0.29, 0.65).$$

Ejercicio 4.8.

Determinar el tamaño n de la muestra que con afijación de varianza mínima produzca la misma precisión que una muestra aleatoria simple no estratificada de tamaño n' para estimar la proporción p de cierta clase en la población. Suponer en ambos casos muestreo con reemplazamiento, y aplicar después el resultado a los siguientes datos con $n' = 500$:

Estrato	W_h	p_h
I	0.3	0.5
II	0.3	0.7
III	0.4	0.2

La afijación de varianza mínima en muestreo con reemplazamiento queda:

$$n_h = n\frac{N_h\sigma_h}{\sum_{h=1}^{L} N_h\sigma_h} = n\frac{N_h\sqrt{p_h(1-p_h)}}{\sum_{h=1}^{L} N_h\sqrt{p_h(1-p_h)}} \overset{Def}{\equiv} nK_h.$$

La varianza del estimador de la proporción en muestreo estratificado con m.a.s.r. en cada estrato será:

$$V(\widehat{p}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{p_h q_h}{n_h} = \sum_{h=1}^{L} W_h^2 \frac{p_h q_h}{nK_h}$$

La varianza en m.a.s. con tamaño n' será

$$V(\widehat{p}) = \frac{p(1-p)}{n}$$

donde la proporción poblacional es $p = \sum_{h=1}^{L} W_h p_h$.

Igualando ambas expresiones y despejando n, se obtiene:

$$\sum_{h=1}^{L} W_h^2 \frac{p_h q_h}{nK_h} = \frac{p(1-p)}{n'}$$

$$\Leftrightarrow$$

$$n = \left(\sum_{h=1}^{L} W_h^2 \frac{p_h q_h}{K_h}\right)\left(\frac{p(1-p)}{n'}\right)^{-1}.$$

Para aplicar el resultado a los datos, hay que calcular primero K_h y p :

ISBN 9788496866133

$$K_h = \frac{N_h\sqrt{p_h(1-p_h)}}{\sum_{h=1}^{L} N_h\sqrt{p_h(1-p_h)}} = \frac{W_h\sqrt{p_h(1-p_h)}}{\sum_{h=1}^{L} W_h\sqrt{p_h(1-p_h)}}$$

donde se ha dividido por N en el numerador y denominador, y por lo tanto,

$$\sum_{h=1}^{L} W_h\sqrt{p_h(1-p_h)} = 0.3\sqrt{0.5(1-0.5)} + 0.3\sqrt{0.7(1-0.7)} + 0.4\sqrt{0.2(1-0.2)} = 0.15 + 0.137 + 0.16 = 0.447$$

$$K_1 = \frac{0.3\sqrt{0.5(1-0.5)}}{0.447} = 0.335$$

y análogamente $K_2 = 0.306$ y $K_3 = 0.358$.

Además,

$$p = \sum_{h=1}^{L} W_h p_h = 0.3 \cdot 0.5 + 0.3 \cdot 0.7 + 0.4 \cdot 0.2 = 0.44$$

Por lo tanto,

$$n = \left(\sum_{h=1}^{L} W_h^2 \frac{p_h q_h}{K_h}\right)\left(\frac{p(1-p)}{n'}\right)^{-1} =$$

$$= \left(0.3^2 \cdot \frac{0.25}{0.335} + 0.3^2 \cdot \frac{0.21}{0.306} + 0.4^2 \cdot \frac{0.16}{0.358}\right)\left(\frac{0.44 \cdot 0.56}{500}\right)^{-1} = 406.$$

Para obtener la misma precisión que en muestreo aleatorio simple con reeplazamiento con $n' = 500$, utilizando muestreo estratificado con afijación mínima y muestreo con reemplazamiento en cada estrato, hará falta muestrear $n = 406$ unidades. Concretamente, $nK_1 = 136$ en el primer estrato, $nK_2 = 124$ en el segundo estrato, y $nK_3 = 146$ en el tercer estrato.

Ejercicio 4.9.

En un estudio de mercado para estimar cierta media poblacional se ha cifrado el coste de evaluar cada unidad en el primer estrato en 4 unidades monetarias, mientras que el coste de evaluar cada unidad en el segundo estrato se cifra en 7 unidades. Se posee una estimación de las cuasivarianzas en cada estrato que aparece en la tabla:

Estrato	W_h	S_h^2
1	0.2	3
2	0.8	5

a) Calcular la afijación teniendo en cuenta que se ha fijado un coste máximo de $C = 200$ unidades.

b) Realizar un programa SAS para calcular el coste y varianza para combinaciones de 1 en 1 para n_1 hasta 10 y n_2 hasta 30. En el cálculo de la varianza, despreciar el término $\frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$ pues no influye a nivel comparativo.

ISBN 9788496866133

a) Se trata de construir la afijación óptima con coste fijo. Así, a partir de la expresión de n_h en función de n y de la restricción $C = \sum_{h=1}^{L} C_h n_h$ se obtiene

$$n_h = C\frac{\frac{N_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} N_h S_h \sqrt{C_h}} = C\frac{\frac{W_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} W_h S_h \sqrt{C_h}}$$

donde se ha dividido por N en numerador y denominador. Por lo tanto:

$$\sum_{h=1}^{L} W_h S_h \sqrt{C_h} = 0.2 \cdot \sqrt{3} \cdot \sqrt{4} + 0.8 \cdot \sqrt{5} \cdot \sqrt{7} = 5.4$$

y entonces:

$$n_1 = 200\frac{\frac{0.2 \cdot \sqrt{3}}{\sqrt{4}}}{5.4} = 6.4$$

y

$$n_1 = 200\frac{\frac{0.8 \cdot \sqrt{5}}{\sqrt{7}}}{5.4} = 25.$$

por lo que se toman $n_1 = 6$ y $n_2 = 25$ para no superar el coste. Por lo que $n = 31$.

b) Hay que ver en primer lugar que

$$\sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2}\frac{S_h^2}{n_h} = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h} + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2.$$

Como en el segundo término no intervienen los n_h no influye a nivel de comparación. el programa es:

```
data costes;
w1=0.2;w2=0.8;s1=3**0.5;s2=5**0.5;
c1=4;c2=7;
do n1=1 to 10;
 do n2=1 to 30;
  vari=((w1*s1)**2)/n1+((w2*s2)**2)/n2;
  coste=c1*n1+c2*n2;
  output;
 end;
end;
run;
```

Para presentar los datos, ordenamos el archivo por coste y varianza, y pedimos que sólo se presenten la observaciones con coste ente 190 y 210:

```
proc sort data=costes ;by descending coste vari;
proc print data=costes (where=(190<coste<210)) noobs;
var n1 n2 coste vari;
run;
```

Obteniendo el listado en la ventana OUTPUT:

ISBN 9788496866133

n1	n2	coste	vari
5	27	209	0.14252
10	24	208	0.14533
3	28	208	0.15429
8	25	207	0.14300
1	29	207	0.23034
6	26	206	0.14308
4	27	205	0.14852
9	24	204	0.14667
2	28	204	0.17429
7	25	203	0.14514
5	26	202	0.14708
10	23	201	0.15113
3	27	201	0.15852
8	24	200	0.14833
1	28	200	0.23429
6	25	199	0.14800
4	26	198	0.15308
9	23	197	0.15246
2	27	197	0.17852
7	24	196	0.15048
5	25	195	0.15200
10	22	194	0.15745
3	26	194	0.16308
8	23	193	0.15413
1	27	193	0.23852
6	24	192	0.15333
4	25	191	0.15800

Se observa que la solución $n_1 = 6$ y $n_2 = 25$ es la de menor varianza en el entorno de $C = 200$.

Ejercicio 4.10.

En el archivo SAS pesos se dispone de datos de una población de 1200 hombres y mujeres, de la variable peso. Se pretende realizar muestreo aleatorio simple por estratos para estimar la media, y compararlo con la precisión obtenida por m.a.s normal.

a) Puesto que se dispone de la información poblacional, calcular la varianza exacta del estimador de la media con $n = 40$ y afijación igual en los siguientes casos:

a1) Muestreo estratificado con m.a.s. en cada estrato.

a2) Muestreo estratificado con m.a.s.r. en cada estrato.

a3) Muestreo aleatorio simple.

a4) Muestreo aleatorio simple con reemplazamiento.

a5) Muestreo aleatorio simple y post estratificación.

Comentar también sin hacer los cálculos si se ganaría algo utilizando afijación de varianza mínima.

ISBN 9788496866133

b) Extraer 5 muestras por muestreo estratificado de tamaños $n_1 = n_2 = 20$, calcular el estimador de la media con la macro estimestrat y apuntar los valores del estimador. Utilizar las semillas 1234,1235,1236,1237,1238.

c) Hacer lo mismo pero con m.a.s. directo.

d) Hacer lo mismo, pero esta vez utilizando postestratificación (obtener la muestra con m.a.s. y estimar con la macro estimestrat).

---

a) Utilizando el siguiente programa:

```
proc means data=peso mean var;
var peso;
by sexo;
run;
```

se obtienen las cuasivarianzas por estrato. Utilizando el mismo programa sin "by peso;" se obtiene la cuasivarianza poblacional (y la media poblacional del peso, que es 69.89).

```
proc means data=peso mean 11var;
var peso;
run;
```

Así, llamando 1 al estrato de hombres y 2 al de mujeres, se tiene:

$S_1^2 = 114.97$, $S_2^2 = 115.93$ y $S^2 = 163.95$. La media poblacional del peso exacta es $\widehat{\overline{y}} = 69.89$.

Hay 575 hombres y 625 mujeres, con lo cual la afijación igual es muy parecida a la proporcional (la afijación proporcional con $n = 40$ sería de $n_1 = 19$ y $n_2 = 21$).

De modo que la varianza en cada caso, con $n = 40$, $n_1 = n_2 = 20$, es:

a1) Muestreo estratificado, con m.a.s. en cada estrato:

$$V(\overline{y}_{st}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2} \frac{s_h^2}{n_h} = \frac{575(575-20)}{1200^2}\frac{114.97}{20} + \frac{625(625-20)}{1200^2}\frac{115.93}{20} = 2.79.$$

a2) Muestreo estratificado, con m.a.s.r. en cada estrato:

En este caso, se suman las varianzas en cada estrato multiplicadas por el peso al cuadrado. Como las varianzas difieren de las obtenidas cuand se utiliza m.a.s. colamente en el coeficiente por población finita en cada estrato $\frac{N_h - n_h}{N_h}$, queda:

$$V(\overline{y}_{st}) = \sum_{h=1}^{L} \frac{N_h^2}{N^2}\frac{\sigma_h^2}{n_h} = \sum_{h=1}^{L} \frac{N_h^2}{N^2}\frac{N_h - 1}{N_h}\frac{S_h^2}{n_h} =$$

$\frac{575^2}{1200^2}\frac{575-1}{575}\frac{114.97}{20} + \frac{625^2}{1200^2}\frac{625-1}{625}\frac{115.93}{20} = 2.88$, ligeramente inferior (hay que darse cuenta de que cuando la fracción de muestreo f es pequeña en cada estrato, la probabilidad de repetir observación en m.a.s.r. es muy baja y por lo tanto las estimaciones suelen ser muy similares en general a las obtenidas con m.a.s.).

a3) Muestreo aleatorio simple:

$$V(\widehat{\overline{y}}) = \frac{N-n}{N}\frac{S^2}{n} = \frac{1200-40}{1200}\frac{163.95}{40} = 3.96.$$

ISBN 9788496866133

a4) Muestreo aleatorio simple con reemplazamiento:

$$V(\widehat{\overline{y}}) = \frac{\sigma^2}{n} = \frac{N-1}{N}\frac{S^2}{n} = 4.09.$$

a5) Muestreo aleatorio simple + postestratificación:

Como $f = \frac{40}{1200}$, $1 - f = 0.9666$.

$$V(\overline{y}_{post}) \simeq \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2 + \frac{1-f}{n^2}\sum_{h=1}^{L}(1-W_h)S_h^2 =$$

$$= \frac{0.966}{40}[\frac{575}{1200}114.97 + \frac{625}{1200}115.93] + \frac{0.966}{40^2}[\frac{625}{1200}114.97 + \frac{575}{1200}115.93] = 2.857.$$

Con lo cual obtenemos que es mejor estratificar (con esa afijación), pero la ganancia es pequeña respecto a utilizar m.a.s. con postestratificación .

La afijación de varianza mínima no parece tener tanto sentido, porque hay muy poca diferencia en cuanto a las varianzas y tamaños en cada estrato. La ganancia en error sería muy pequeña frente a la complicación de tener que tener estimadas las cuasivarianzas por estrato (al ser un ejercicio teórico tenemos estas cuasivarianzas, pero en la práctica no se dispondría de ellas).

b) Antes de poder ejecutar la macro, hay que añadir la variable de tamaño de los estrato. Se puede hacer de dos maneras: en un archivo aparte, o añadirlo a la muestra (en este caso mejor a la población, pues se van a tomar varias muestras). También se crea la variable estrato:

```
data peso;set pesos;estrato=sexo;
if sexo=1 then tama=575;else tama=625;
run;
```

A continuación se repite el siguiente programa 5 veces, cambiando la semilla cada vez:

```
proc surveyselect data=peso out=muestra method=srs sampsize=20 seed=1234;
strata estrato;
run;
%estimestrat(muestra,.,peso,tama,1,2,2,1200);
```

Las medias estratificadas obtenidas son:

70.85, 69.73, 68.71, 69.29, 71.16

c) Utilizando m.a.s., se realiza el siguiente programa 5 veces, cambiando la semilla cada vez:

```
proc surveyselect data=peso out=muestra method=srs sampsize=40 seed=1234;
run;
%estimas(muestra,peso,1200);
```

Obteniendo las medias muestrales:

70.67, 68.6, 71.12, 67.02, 69.02.

d) El programa básico es

ISBN 9788496866133

```
proc surveyselect data=peso out=muestra method=srs sampsize=40 seed=1234;
run;
%estimestrat(muestra,.,peso,tama,2,2,2,1200);
```

Y variando la semilla se obtienen las medias:

68.71,69.6,70.52,67.15, 69.07.

ISBN 9788496866133

5.14 Ejercicios propuestos

1) Se tomó una muestra de 2000 personas por m.a.s. pertenecientes a una población de 50980 habitantes, y se recogieron datos sobre ingresos y nivel educativo. Los resultados se resumen en la siguiente tabla:

Nivel educativo	Media muestral	Cuasivarianza muestral	Tamaño muestral
1	4.1	34.8	1585
2	13	92.2	250
3	25	174.2	105
4	38.2	320.4	60

a) Estimar el ingreso medio por individuo en esa comunidad.

b) Suponer que, a pesar de que no hay ningunalista de personas de la comunidad con cada nivel educativo, se conocen las proporciones de personas de la población pertenecientes a cada uno de los 4 niveles educativos. Éstas son respectivamente:$p_1 = 0.8033$, $p_1 = 0.1315$, $p_3 = 0.0470$, $p_1 = 0.0245$.

Contando con esta información, estimar el ingreso medio por individuo y facilitar un I.C. al 95%.

c) ¿Cuál de las dos estimaciones elegirías?

d) ¿Es más preciso el procedimiento del apartado b) que un muestreo estratificado con afijación proporcional?

2) Una población está dividida en tres estratos. Sus pesos respectivos son: 0.45, 0.35, 0.2. En una encuesta piloto se han estimado las proporciones de una cualidad de interés en cada uno de los estratos: estas estimaciones son: $p_1 = 0.48$, $p_1 = 0.39$, $p_1 = 0.58$. Se considera que los tamaños de los estratos son suficientemente grandes respecto a los tamaños muestrales.

a) Comparar la precisión de una muestra estratificada con afijaciónn proporcional de tamaño total 500 para estimar la proporción poblacional con la precisión de una m.a.s. no estratificada de tamaño 600 .

b) Opinar si la ganancia o pérdida en precisión en el apartado anterior parece significativa o no.

3) Se desea estimar un total de una población de 770 elementos mediante muestreo estratificado. Se sabe que la característica X, dada por la distribución de frecuencias siguientes, está relacionada con la varaible de interés Y:

ISBN 9788496866133

X	Frecuencias
2	100
3	80
4	90
5	200
8	90
10	30
15	90
20	50
50	10
100	20
200	10

a) Determinar 3 estratos para el estudio de dicha población.

b) Hallar la afijación de varianza mínima para una muestra de tamaño 100, suponiendo que el muestreo en cada estrato se realiza mediante m.a.s. con reposición, y que la variable de interés Y está relacionada con X según el modelo $Y_i = aX_i + b$.

4) Se tomó una m.a.s. de 52 trabajadores de una fábrica de 1400, recogiéndose datos sobre el efecto contaminante de los gases sobre los pulmones, medido en una variable continua. Los datos recogidos se organizan a continuación en una tabla, donde se apunta también en que grado de exposición a los gases trabaja habitualmente el obrero:

Exposición	Tamaño muestral	Media muestral	Cuasivarianza muestral
Alta	30	85.1	160
Media	10	76.2	110
Baja	12	43.1	100

a) Estimar el efecto contaminante medio en los trabajadores de la fábrica.

b) Suponer que se conocen, antes de analizar los datos, las proporciones de los trabajadores que están expuestos en cada una de las áreas:

ISBN 9788496866133

Alta: proporción=0.45. Media: proporción=0.25. Baja: proporción=0.30. Estimar, contando con esta información, el efecto contaminante medio, y dar un I.C. al 95%.

c) ¿Cuál de las dos estimaciones elegirías?.

5) Se van a muestrear en un pueblo familias para estimar los gastos mensuales que no son de alimentación. Se distingue entre familias de renta baja y de renta alta, habiendo en el pueblo 15000 de renta baja y 3000 de renta alta.

Se sabe por estudios anteriores que las familias de renta alta tienen aproximadamente 4 veces más gastos que las de renta baja. Además, se espera que la cuasivarianza del gasto en cada estrato sea aproximadamente proporcional a la media del mismo, con ambas proporcionalidades similares.

Explicar cómo distribuir una muestra de 800 familias en los dos estratos, suponiendo m.a.s. en cada estrato.

6) La marca de coches "Nidia" desea conocer, de los 10000 coches vendidos en 2004 a través de sus concesionarios, el número medio de ruedas cambiadas por coche desde entonces. Para ello se tomó una m.a..s. de 100 coches, recogiéndose datos sobre la cantidad de kilómetros recorridos y el número de ruedas cambiadas desde su adquisición. Los resultados obtenidos se resumen en la siguiente tabla:

Kilometraje	Tamaño muestral	Media muestral	Cuasivar. muestral
Elevado	40	8.2	0.60
Medio	40	4.3	0.41
Bajo	20	0.9	0.09

a) Estimar el número medio de ruedas cambiadas por coche desde su adquisición en 2004 en los concesionarios "Nidia".

b) Suponer ahora que, antes de analizar los datos obtenidos, se conocen las proporciones de coches objeto de estudio cuyo kilometraje se encuentra en cada uno de los tres niveles fijados para su clasificación:

Kilometraje	Elevado	Medio	Bajo
Proporción	0.3	0.5	0.2

Contando con esta información, estimar el número medio de ruedas cambiadas por cohce, y facilitar un I.C. al 95% suponiendo normalidad.

c) ¿Cuál de las dos estimaciones elegirías? Razonar la respuesta.

7) Los pesos relativos de los tres estratos de una población son:

ISBN 9788496866133

I	II	III
0.2	0.5	0.3

En una encuesta piloto se han encontrado los valores siguientes para las proporciones de los tres estartos, relativos a una cierta característica que interesa estudiar:

I	II	III
0.4	0.3	0.6

Resolver los siguientes apartados considerando que los tamaños de los estratos son suficientemente grandes con relación a los respectivos tamaños muestrales.

a) ¿Crees que una muestra estratificada con m.a.s. en cada estrato con afijación proporcional de tamaño 400 dará mayor precisión para estimar la proporción poblacional de dicha característica, que una m.a.s. de tamaño 500 sin estratificar?.

b) ¿Para qué tamaño muestral del diseño estratificado se igualan las precisiones?

c) ¿Te parece que la estratificación aporta en este caso una ganancia sustancial en precisión?

8) En el archivo SAS pesos están datos de peso, estatura, sexo y dieta de una población de 1200 personas.

a) Crear un archivo SAS llamado mujeres donde estén solamente las mujeres.

b) Comprobar con

```
proc freq data=mujeres;tables dieta;run;
```

cuántas mujeres han hecho dieta (dieta=1) y cuántas no (dieta=0).

Utilizar

```
proc means data=mujeres;run;
```

para comprobar el peso medio poblacional de las mujeres.

c) Extraer 4 muestras estratificadas de tamaño 100 del archivo mujeres, siendo los estratos formados por la variable dieta, con afijación proporcional. Utilizar las semillas 1234, 1235, 1236, 1237.

d) Utilizar la macro estimestrat para estimar la media poblacional del peso de las mujeres en cada una de las muestras anteriores.

e) Utilizar

```
proc sort data=mujeres;by dieta;
proc means data=mujeres var;run;
```


ISBN 9788496866133

para obtener la cuasivarianza poblacional por estrato de la variable peso.

f) Calcular la afijación de varianza mínima dadas las cuasivarianzas poblacionales anteriores, siempre con $n = 100$.

g) Obtener 4 muestras con las mismas semillas, con la afijación dada en el apartado f), y estimar la media poblacional con la macro estimestrat en cada caso.

h) ¿Cuál sería la varianza exacta del estimador del peso medio de las mujeres con el mismo tamaño muestral, si se utiliza m.a.s. sin estratificar?.

9) Realizar en SAS el ejercicio resuelto 4.6 con la ayuda de la macro estimestrat.

10) El archivo SAS ganado contiene las 50 provincias españolas que tienen algún tipo de ganado, con las variables Bovino, Ovino, Caprino, Porcino y Equino que indican el número de cabezas sacrificadas en 1998 respectivamente en cada tipo de ganado.

Se estudiará en este ejercicio la variable caprino.

a) Con

```
proc means data=ganado sum;var caprino;run;
```

se obtiene el total poblacional que queremos estimar a través el muestreo estratificado.

Con el paso data

```
data cabras;
set ganado;
if caprino<30000 then estrato=1;
else estrato=2;
run;
```

se crea la división de la población en dos estratos, las provincias con menos de 30.000 cabezas sacrificadas en 1998 están en el estrato 1 y las demás en el estrato 2.

Con

```
proc freq data=cabras;tables estrato;run;
```

se observa el tamaño de los dos estratos.

b) Realizar ahora muestreo estratificado de tamaño $n = 10$ con la semilla 1234, con afijación proporcional, y estimar el total de cabezas sacrificadas con la macro estimestrat.

c) Utilizar

```
proc sort data=cabras;by estrato;
proc means data=cabras var;by estrato;run;
```


ISBN 9788496866133

para calcular las cuasivarianzas por estrato. Calcular la afijación de varianza mínima para $n = 10$. ¿Qué ocurre?.

d) Considerar las 9 provincias con mayor número de cabezas de caprino como unidades autorrepresentadas, y tomar una m.a.s. de tamaño n=1 del resto de la población. Construir finalmente el estimador del total. ¿Es mejor o peor que el obtenido con afijación proporcional?.¿Qué conclusión sacas para poblaciones muy asimétricas en muestreo estratificado?.

ISBN 9788496866133

6 MUESTREO SISTEMÁTICO

6.1 Introducción

El esquema de muestreo mediante el cual la primera unidad es seleccionada aleatoriamente, pero las demás son seleccionadas automáticamente según un patrón determinado se conoce como muestreo sistemático. El muestreo sistemático es un sistema práctico y eficiente en muchas ocasiones entre las cuales se cuentan, por ejemplo:

- Muestreo geográfico, donde se recorre el área una cantidad fija de espacio cada vez, tomando unidades sucesivas y a partir de la primera unidad que ha sido seleccionada aleatoriamente. Se utiliza tanto en agricultura, ecología, como en encuestas sociales en trabajo de campo.
- Muestreo en control de calidad, donde se toman items de producción cada cierto número de items a partir del primero, que ha sido seleccionado aleatoriamente.
- Muestreo en auditorías o en informática, donde se deben revisar largos listados.
- Muestreo en encuestas en ciudades, donde se entrevista cada cierto número de peatones que pasan por un lugar

Entre las ventajas que tiene este muestreo conviene citar que es de aplicación sencilla, que en casos en que no se disponga de un listado a priori de todas las unidades de la población es una solución apropiada, que da una pauta de actuación más fácil de aplicar en el trabajo de campo, y que puede recoger cierto efecto de estratificación cuando las unidades están ordenadas respecto a una variable correlada con la variable de interés, en cuyo caso puede mejorar al m.a.s.

Para realizar muestreo sistemático para la obtención de una muestra de tamaño n en una población de tamaño N se supondrá de momento que $k = \dfrac{N}{n}$ es un número entero. Obsérvese que la población está dividida en k partes con n unidades cada una. En este caso, el procedimiento de muestreo sistemático es el siguiente:

ISBN 9788496866133

1. Se selecciona la unidad de partida i , con i entre las k primeras unidades, con igual probabilidad $\frac{1}{k}$.
2. Las restantes $n-1$ unidades de la muestra vienen determinadas automáticamente pues son aquellas que toman la posición i en cada una de las n particiones realizadas de la población. Es decir, las unidades seleccionadas son las $\{i, i+k, i+2k, i+3k, ..., i+(n-1)k\}$

Está claro que hay tantas muestras posibles, k, como unidades de partida . Es decir, a cada unidad de partida posible i le corresponde una muestra $\{i+k, i+2k, i+3k, ..., i+(n-1)k\}$. Recuérdese que, sin tener en cuenta el orden, en m.a.s. hay $\binom{N}{n}$ muestras posibles. A pesar de que en muestreo sistemático el número $k = \frac{N}{n}$ es mucho menor, el muestreo sistemático puede proveer estimadores con precisión similar al m.a.s.

6.2 Estimación en muestreo sistemático

Supongamos que $k = \frac{N}{n}$ es un número entero. La media muestral puede expresarse como $\widehat{\overline{y}} = \frac{1}{n}\sum_{j=0}^{n-1} y_{i+jk}$, donde i es el punto de arranque de la muestra sistemática escogido por m.a.s , de entre los números $i = 1, .., k$, y k es $k = \frac{N}{n}$ entero.

Teorema 6.1 (estimación de la media).

La media muestral $\widehat{\overline{y}}$ es un estimador insesgado de la media poblacional $\overline{y}$.

Demostración.

Cada muestra posible i de las k muestras tiene probabilidad $\frac{1}{k}$ de ser escogida y da lugar a una media muestral $\widehat{\overline{y}}_i$, con lo que por la definición de esperanza,

$E(\widehat{\overline{y}}) = \frac{1}{k}\sum_{i=1}^{k}\widehat{\overline{y}}_i = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{n}\sum_{j=0}^{n-1} y_{i+jk}\right) = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=0}^{n-1} y_{i+jk}$. Como $\sum_{i=1}^{k}\sum_{j=0}^{n-1} y_{i+jk}$ es la suma de todas las observaciones de la población, y $nk = N$, $E(\widehat{\overline{y}}) = \frac{1}{N}\sum_{i=1}^{N} y_i = \overline{y}$.

Teorema 6.2 (varianza del estimador).

La varianza del estimador $\widehat{\overline{y}}$ es $V(\widehat{\overline{y}}) = \frac{1}{k}\sum_{i=1}^{k}(\widehat{\overline{y}}_i - \overline{y})^2$.

Demostración.

Es directo por la definición de varianza, debido a que el los valores $\widehat{\overline{y}}_i$ son equiprobables por ser escogidos por m.a.s., y su esperanza es $\overline{y}$.

ISBN 9788496866133

Corolario 6.1 (estimación del total y proporción).

Un estimador insesgado del total poblacional es $N\widehat{\overline{y}}$, y un estimador insesgado de la proporción poblacional es la proporción muestral $\widehat{p}$.

Ejemplo 6.1.

Supongamos que sobre los siguientes datos se desea obtener una muestra sistemática de tamaño $n = 4$.

Obs	1	2	3	4	5	6	7	8	9	10	11	12
y	2	3	2	5	6	8	5	4	6	5	2	1

Tabla 6.1. Datos de ejemplo

Entonces $k = \dfrac{N}{n} = \dfrac{12}{4} = 3$. El punto de arranque varía de u_1 a u_3. Por lo tanto las muestras posibles son las observaciones $\{1, 4, 7, 10\}$, $\{2, 5, 8, 11\}$, y $\{3, 6, 9, 12\}$ que corresponden a los valores $\{2, 5, 5, 5\}$, $\{3, 6, 4, 2\}$, y $\{2, 8, 6, 1\}$ con medias muestrales respectivas 4.25, 3.75, y 4.25. Obsérvese que con diferentes tipos de muestreo el espacio de posibles muestras varía. Por ejemplo, en muestreo con reemplazamiento habría, sin tener en cuenta el orden, $\binom{12+4-1}{4} = 1365$ muestras posibles, mientras que en m.a.s. habría $\binom{12}{4} = 495$ muestras posibles.

Esto no tiene por qué redundar en que un método de muestreo sea más preciso que otro, aún con los mejores estimadores insesgados posibles. Puede ocurrir que el muestreo sistemático arroje estimaciones precisas , a pesar de que su espectro de posibles muestras sea tan reducido (pues siempre está limitado a las k muestras equivalentes a los k puntos de arranque).

En el ejemplo, se puede comprobar que efectivamente el estimador media muestral es insesgado, pues la media poblacional es $\frac{1}{12}(2 + 3 + ... + 1) = 4.0833$ y la esperanza del estimador, al ser cada una de las tres muestras equiprobables, es $\frac{1}{3}(4.25 + 3.75 + 4.25) = 4.0833$.

6.3 Estimación de la media poblacional cuando $k = \frac{N}{n}$ no es entero

Supongamos que $k = \dfrac{N}{n}$ no es un número entero. Entonces hay que decidir si se toma k como el entero superior a $\dfrac{N}{n}$ o el inferior.

ISBN 9788496866133

Ejemplo 6.2.

Sea la población formada por 5 observaciones numeradas 1,2,3,4,5:

Obs	1	2	3	4	5
y	2	3	2	5	6

Tabla 6.2. Datos de ejemplo

Supongamos que $n = 2$. Entonces $k = \dfrac{5}{2} = 2.5$. Hay dos posibilidades: tomar $k = 2$ o $k = 3$. Si se toma $k = 2$, las 2 muestras sistemáticas posibles, escogidas a partir del inicio aleatorio entre las dos primeras observaciones, son $\{1, 3, 5\}$y $\{2, 4\}$. Vemos que la segunda muestra tiene una observación menos. Si se toma $k = 3$, las 3 muestras posibles son $\{1, 4\}, \{2, 5\}$ y $\{3\}$ Vemos que la tercera muestra tiene una observación menos. Así, el tamaño de la muestra es en este caso también una variable aleatoria.

Propiedad 6.1.

La media muestral $\widehat{\overline{y}}$ no es un estimador insesgado de la media poblacional $\overline{y}$ cuando $\dfrac{N}{n}$ no es entero.

Basta ver en el ejemplo que la media poblacional es $\overline{y} = 3.6$.Ahora, cuando $k = 2$, las medias muestrales obtenidas de las dos posibles muestras son respectivamente, $\frac{10}{3}$ y $\frac{8}{2}$. Como cada muestra es equiprobable, la esperanza del estimador será $\frac{1}{2}(\frac{10}{3} + \frac{8}{2}) = 3.667$ que no es igual que $\overline{y}$.

Si $k = 3$, las medias muestrales respectivas en cada muestra son $\frac{7}{2}, \frac{9}{2}, 3$. La esperanza de la media muestral será $\frac{1}{3}(\frac{7}{2} + \frac{9}{2} + 3) = 3.333 \neq \overline{y}$.

Teorema 6.3 (estimador de la media).

Sea y_i el total calculado sobre la muestra sistemática i, considerando válidas todas las muestras (aunque tengan tamaños diferntes), y una vez escogido k para el proceso de selección, que puede haber sido el entero inferior o el superior. El estimador $\widetilde{\overline{y}}' = \dfrac{k}{N} y_i$ es un estimador insesgado de $\overline{y}$.

Demostración.

Cada muestra sistemática tiene probabilidad $\dfrac{1}{k}$ de ser escogida. Así,

$E(\widetilde{\overline{y}}_i) = E\left(\dfrac{k}{N} y_i\right) = \dfrac{1}{k} \sum_{i=1}^{k} \dfrac{k}{N} y_i = \dfrac{1}{N} \sum_{i=1}^{k} y_i = \overline{y}$. pues la última igualdad deriva del hecho de que la suma de los totales $\sum_{i=1}^{k} y_i$ sobre las muestras sistemáticas posibles coincide con la suma de toda la población.

ISBN 9788496866133

Ejemplo 6.2.

Supongamos en el ejemplo anterior $k = 2$. Entonces, la primera muestra da lugar a $y_i = 10$ y por lo tanto a $\widehat{\overline{y}}' = \frac{2}{5}10 = 4$. La segunda muestra da lugar a $y_i = 8$ y por lo tanto a $\widehat{\overline{y}}' = \frac{2}{5}8 = 3.2$. Como cada una de las 2 muestras tiene probabilidad $\frac{1}{2}$, la esperanza del estimador $\widehat{\overline{y}}'$ será $\frac{1}{2}(4 + 3.2) = 3.6 = \overline{y}$.

Ahora, si $k = 3$, la primera muestra da lugar a $y_i = 7$ y $\widehat{\overline{y}}' = \frac{3}{5}7 = 4.2$. La segunda muestra da lugar a $y_i = 9$ y $\widehat{\overline{y}}' = \frac{3}{5}9 = 5.4$. La tercera da $y_i = 2$ y $\widehat{\overline{y}}' = \frac{3}{5}2 = 1.2$. Como cada una de las 3 muestras tiene probabilidad $\frac{1}{3}$, la esperanza del estimador $\widehat{\overline{y}}'$ será $\frac{1}{3}(4.2 + 5.4 + 1.2) = 3.6 = \overline{y}$.

Normalmente debido a este tipo de complicaciones se intenta evitar que $\frac{N}{n}$ no sea entero, bien cambiando li-geramente n o bien eliminando alguna observación poblacional al azar previamente. Además, si la población es grande, el sesgo sería pequeño en caso de k no entero. En el resto del capítulo se asumirá $k = \frac{N}{n}$ entero.

6.4 Muestreo sistemático en áreas

Frecuentemente se utiliza el muestreo sistemático en un área geográfica, como por ejemplo escoger varios árboles frutales para evaluar su producción. En este sentido, es necesario utilizar el concepto bidimensional para evitar soluciones incorrectas. Si se dispone de un área cuadriculada, dividida en 9 columnas y 6 filas, y se numeran las cuadrículas del 1 al 54, y para obtener una muestra sistemática de $n = 6$ se calcula $k = \frac{N}{n} = \frac{54}{6} = 9$, puede ocurrir que el punto de arranque sea u_1 (la primera cuadrícula), dando lugar a la configuración muestral correspondiente a la Figura 6.1, que es obviamente poco atractiva en cuanto a representatividad. Las otras 8 muestras sistemáticas posibles son igualmente poco atractivas, al coincidir en la misma columna todos los árboles.

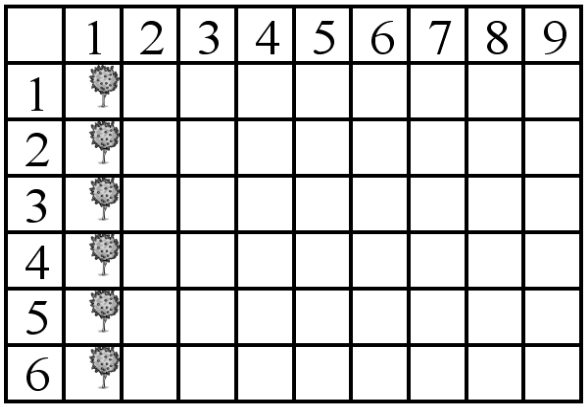


Figura 6.1. Planteamiento incorrecto del muestreo sistemático en un área.

ISBN 9788496866133

Existen varias soluciones a este problema. La más sencilla es utilizar muestreo sistemático "alineado". Supongamos que el tamaño muestral n puede descomponerse en el producto de dos enteros n_1 y n_2, de modo que $n_1 \cdot n_2 = n$. Se asume que $n_1 \geq n_2$. Supongamos también que r=número de columnas es múltiplo de n_1, y que s =número de filas es múltiplo de n_2, con $r \geq s$. Todas estas restricciones no son tan fuertes, pues en la práctica siempre se pueden adecuar las divisiones geográficas o los tamaños muestrales para cumplirlas.

Entonces el muestreo sistemático alineado consiste en

1) Calcular $k_1 = \dfrac{r}{n_1}$, $k_2 = \dfrac{s}{n_2}$.

2) Escoger el cuadro de arranque, cuyas coordenadas son (i, j), donde i está escogido entre $1, ...k_1$ con equiprobabilidad y j está escogido entre $1, ...k_2$ con equiprobabilidad.

3) La muestra sistemática correspondiente al punto de arranque es la que consta de los cuadros (i, j), $(i + k_1, j), ...(i + (n_1 - 1)k_1, j)$, $(i, j + k_2), ..., (i, j + (n_2 - 1)k_2)$.

Ejemplo 6.3.

Supongamos que en un área como la presentada en la Figura 6.1, se pretende realizar muestreo sistemático con $n = 6$. Entonces se fijan $n_1 = 3$, $n_2 = 2$. De este modo $\dfrac{r}{n_1} = \dfrac{9}{3} = 3$ y $\dfrac{s}{n_2} = 3$. Los posibles puntos de arranque son todo el cuadro marcado con X en la parte superior izquierda de la Figura.6.2. En la misma figura aparecen 3 muestras del total de las 9 muestras sistemáticas posibles.

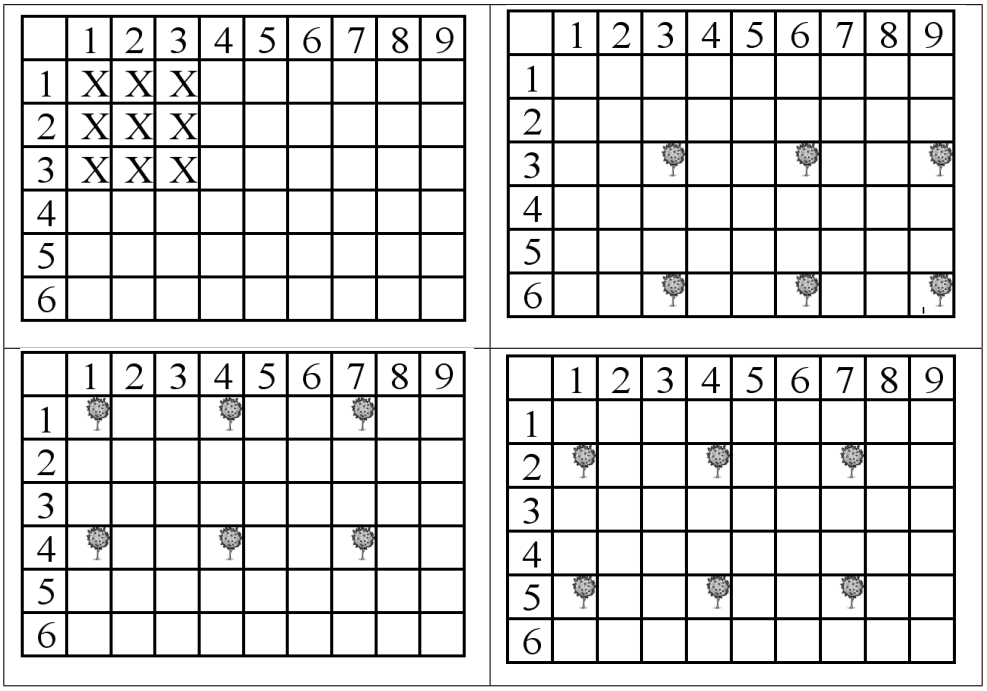


Figura 6.2. Posibles puntos de arranque y muestras sistemáticas posibles.

ISBN 9788496866133

6.5 Descomposición de la varianza en muestreo sistemático

La varianza poblacional admite una descomposición de utilidad para comprender por qué el muestreo sistemático puede mejorar al m.a.s.

Definición.

Supongamos que y_{ij} es la unidad j dentro de la muestra sistemática i. .Llamaremos σ_w^2 a la variabilidad media interna de las muestras sistemáticas. Es decir,

$$\sigma_w^2 = \frac{1}{k}\sum_{i=1}^{k}\sigma_i^2 = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n}\sum_{j=1}^{n}(y_{ij}-\widehat{\overline{y}}_i)^2 = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\widehat{\overline{y}}_i)^2.$$

Llamaremos σ_b^2 a la variabilidad entre muestras sistemáticas, $\sigma_b^2 = \frac{1}{k}\sum_{i=1}^{k}(\widehat{\overline{y}}_i-\overline{y})^2$, que coincide con la varianza del estimador de la media.

Teorema 6.4 (descomposición de la varianza).

La varianza de la población σ^2 se puede descomponer en la suma de las varianzas dentro de muestras sistemáticas σ_w^2 y entre muestras sistemáticas σ_b^2. Es decir,

$\sigma^2 = \sigma_w^2 + \sigma_b^2$.

Demostración.

En primer lugar hay que notar que $\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\widehat{\overline{y}}_i)(\widehat{\overline{y}}_i-\overline{y}) = 0$ pues

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\widehat{\overline{y}}_i)(\widehat{\overline{y}}_i-\overline{y}) =$$

$$=\sum_{i=1}^{k}\widehat{\overline{y}}_i\sum_{j=1}^{n}y_{ij} - \overline{y}\sum_{i=1}^{k}\sum_{j=1}^{n}y_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{n}\widehat{\overline{y}}_i^2 + \overline{y}\sum_{i=1}^{k}\sum_{j=1}^{n}\widehat{\overline{y}}_i =$$

$$=n\sum_{i=1}^{k}\widehat{\overline{y}}_i^2 - nk\overline{y}^2 - n\sum_{i=1}^{k}\widehat{\overline{y}}_i^2 + nk\overline{y}^2 = 0.$$

Luego como

$$N\sigma^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\overline{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\widehat{\overline{y}}_i+\widehat{\overline{y}}_i-\overline{y})^2 =$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\widehat{\overline{y}}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(\widehat{\overline{y}}_i-\overline{y})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\widehat{\overline{y}}_i)(\widehat{\overline{y}}_i-\overline{y}) =$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\widehat{\overline{y}}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(\widehat{\overline{y}}_i-\overline{y})^2 = N\sigma_w^2 + N\sigma_b^2\,,$$

se tiene que

$\sigma^2 = \sigma_w^2 + \sigma_b^2$.

ISBN 9788496866133

6.6 Comparación con m.a.s.

Utilizando la descomposición de la varianza , $\sigma^2 = \sigma_w^2 + \sigma_b^2$, la varianza del estimador de la media en el método habitual en muestreo sistemático es $V(\widehat{\overline{y}}_{sis}) = \sigma_b^2 = \sigma^2 - \sigma_w^2$ y la varianza del estimador de la media en muestreo aleatorio simple es $V(\widehat{\overline{y}}_{m.a.s.}) = \frac{N-n}{N}\frac{S^2}{n}$. Definiendo ahora la cuasivarianza interna de las muestras sistemáticas como $S_w^2 = \frac{n}{n-1}\sigma_w^2$, se tiene que $V(\widehat{\overline{y}}_{sis}) = \frac{N-1}{N}S^2 - \frac{n-1}{n}S_w^2$

Teorema 6.7 (comparación del muestreo sistemático con m.a.s.).

a) Si $S_w^2 > S^2$, $V(\widehat{\overline{y}}_{sis}) < V(\widehat{\overline{y}}_{m.a.s.})$

b) Si $S_w^2 < S^2$, $V(\widehat{\overline{y}}_{sis}) > V(\widehat{\overline{y}}_{m.a.s.})$

c) Si $S_w^2 = S^2$, $V(\widehat{\overline{y}}_{sis}) = V(\widehat{\overline{y}}_{m.a.s.})$

Demostración.

Si $S_w^2 > S^2$,entonces

$$V(\widehat{\overline{y}}_{sis}) = \frac{N-1}{N}S^2 - \frac{n-1}{n}S_w^2 < \frac{N-1}{N}S^2 - \frac{n-1}{n}S^2 =$$

$$\frac{N-n}{N}\frac{S^2}{n} = V(\widehat{\overline{y}}_{m.a.s.}).$$

Se demuestra análogamente el caso $S_w^2 < S^2$ y $S_w^2 = S^2$.

Este teorema pone de manifiesto que en el muestreo sistemático interesa que la variabilidad dentro de cada muestra sistemática sea alta. Las siguientes consideraciones son importantes a la hora de aplicar este tipo de muestreo:

- Si la ordenación de los datos es aleatoria, el muestreo sistemático es igual de eficiente que el m.a.s., pues la variabilidad interna de cada muestra sistemática será similar a la poblacional.
- Si los datos están ordenados de acuerdo a la variable de interés o a otra variable fuertemente correlada con ésta, el muestreo sistemático será más eficiente que el m.a.s. Este orden hace que la variabilidad de cada muestra sistemática sea mayor en general que la variabilidad poblacional S^2.
- La variabilidad interna de la muestra sistemática será menor que la general ($S_w^2 < S^2$) y por lo tanto el muestreo sistemático peor que el m.a.s., en casos específicos donde por ejemplo existe una relación cíclica en los datos. En esos casos cíclicos, el caso peor se da cuando la constante k es múltiplo de la longitud del ciclo (en este caso S_w^2 puede llegar a valer cero). Pero aún en estos casos cíclicos puede haber valores de k que den lugar a $S_w^2 > S^2$, es decir que el muestreo sistemático pueda ser más preciso que el m.a.s. Si existe una sospecha de datos periódicos una manera de evitar efectos perniciosos sobre el estimador es utilizar muestreo sistemático replicado, como se explicará en la sección

ISBN 9788496866133

de muestras interpenetrantes, es decir, en lugar de una única muestra sistmática, tomar varias de tamaño menor.

Estas consideraciones llevan a reafirmar el muestreo sistemático como una opción práctica recomendable. Los casos en los que el muestreo sistemático es peor que el m.a.s. pueden darse por ejemplo cuando las unidades son los días del año, horas o cualquier momento del tiempo.

Ejemplo 6.4.

Supongamos los siguientes datos:

$\{1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5\}$.

La cuasivarianza poblacional es $S^2 = 2.0833$. Supongamos que queremos una muestra de tamaño $n = 5$. Dependiendo de la ordenación el muestreo sistemático puede ser mejor, igual o peor que el m.a.s.:

Ordenación cíclica: $\{1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5\}$

Ordenación aleatoria : $\{4,2,5,2,3,3,5,3,2,4,1,4,4,1,4,2,1,2,3,1,5,1,5,3,5\}$

Ordenación ascendente:$\{1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,4,4,4,4,4,5,5,5,5,5\}$

La Figura 6.3 representa las tres ordenaciones.

Como $n = 5$, $k = \dfrac{25}{5} = 5$. Las muestras sistemáticas y el valor de S_w^2 para cada ordenación son:

Ordenación cíclica: $\{1,1,1,1,1\},\{2,2,2,2,2\},\{3,3,3,3,3\},\{4,4,4,4,4\},\ \{5,5,5,5,5\}$. $S_w^2 = 0 < S^2$, pues la variabilidad interna de cada muestra sistemática es 0.

Ordenación aleatoria: $\{4,3,1,2,5\},\{2,5,4,1,1\},\{5,3,4,2,5\},\{2,2,1,3,3\}$,

$\{3,4,4,1,5\}$. $S_w^2 = 2.10 \simeq S^2$.

Ordenación ascendente: $\{1,2,3,4,5\},\{1,2,3,4,5\},\{1,2,3,4,5\},\{1,2,3,4,5\}$,

$\{1,2,3,4,5\}$. $S_w^2 = 2.5 > S^2$.

Se observa que el muestreo sistemático con ordenación asdendente mejora al m.a.s. por ser $S_w^2 > S^2$, mientras que la ordenación cíclica con $n = 5$ hace que el muestreo sistemático sea realmente impreciso (si bien es cierto que con $n = 4$, por ejemplo, ya no sería lo mismo). La ordenación aleatoria en este caso ha dado precisión similar para el muestreo sistemático y para el m.a.s. Hay que comentar que hay ordenaciones aleatorias posibles que dan una precisión menor para el muestreo sistemático respecto al m.a.s., y otras que dan precisión mayor (sólo hay que tener en cuenta que la ordenación cíclica y la ascendente son casos particulares de todas las posibles ordenaciones aleatorias). En promedio ambos métodos, sistemático y m.a.s., son equivalentes si la ordenación es aleatoria.

ISBN 9788496866133

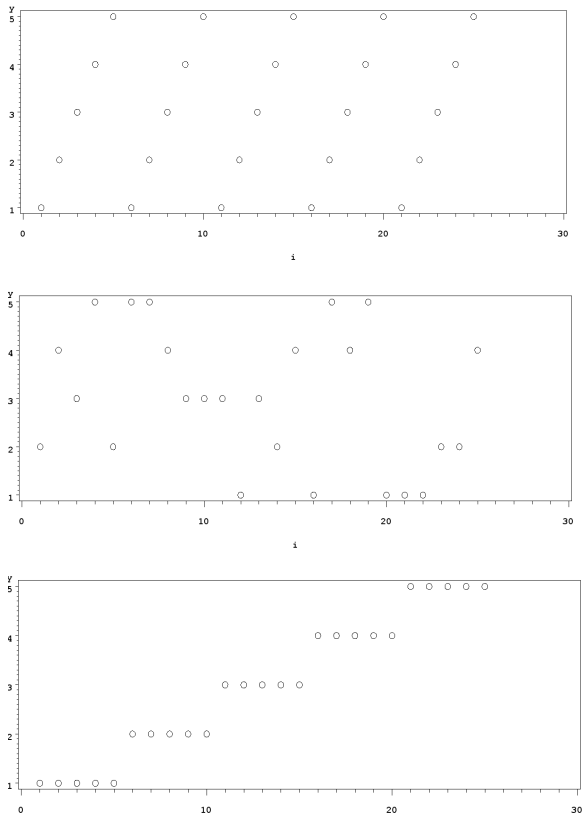


Figura 6.3. Tres ordenaciones diferentes para los mismos datos.

6.7 Estimación de la varianza: muestras interpenetrantes

Con una sola muestra sistemática no existe un estimador insesgado de la varianza del estimador. Una posibilidad si el muestreo sistemático es apropiado (igual o mejor que el m.a.s.), es utilizar el estimador de la varianza usual de m.a.s. $\widehat{V}(\widehat{\overline{y}}) = \dfrac{N-n}{N}\dfrac{s^2}{n}$ como una aproximación fiable, donde s^2 es la cuasivarianza de la muestra sistemática. Del mismo modo se puede utilizar la fórmula habitual de m.a.s. para estimar la varianza del estimador del total y proporción poblacionales.

Otra posibilidad para acceder a un estimador insesgado es realizar muestreo sistemático replicado; es decir, en lugar de tomar una muestra sistemática de tamaño n, se toman m muestras sistemáticas cada una de tamaño $n' = \dfrac{n}{m}$, suponiendo este número entero. Este método se denomina también método de las muestras interpenetrantes o replicadas. Si $\widehat{\overline{y}}_i$ es la media muestral de la muestra sistemática i, y definiendo $k' = \dfrac{N}{n'}$, entonces se dan los siguientes resultados:

Teorema 6.5 (estimación de la media con muestras interpenetrantes).

$\widehat{\overline{y}}_m = \dfrac{1}{m}\sum_{i=1}^{m}\widehat{\overline{y}}_i$ es un estimador insesgado de $\overline{y}$.

ISBN 9788496866133

Demostración.

Es directo pues $\widehat{\overline{y}}_m$ es la media muestral de m elementos $\widehat{\overline{y}}_i$, obtenidos por m.a.s., y en m.a.s. la media muestral es insesgado para la media poblacional.

Corolario 6.2 (varianza del estimador). $V(\widehat{\overline{y}}_m) = \dfrac{k'-m}{k'}\dfrac{1}{m(k'-1)}\sum_{i=1}^{k'}(\widehat{\overline{y}}_i - \overline{y})^2$.

Demostración.

Igualmente a la demostración anterior, al tomarse los m valores $\widehat{\overline{y}}_i$ por m.a.s. en una población finita de k' elementos, y al ser la cuasivarianza de $\widehat{\overline{y}}_i$ el término $\dfrac{1}{(k'-1)}\sum_{i=1}^{k'}(\widehat{\overline{y}}_i - \overline{y})^2$, aplicando la expresión de la varianza del estimador de la media en m.a.s. se tiene el resultado.

Teorema 6.6 (estimación de la varianza del estimador).

$$\widehat{V}(\widehat{\overline{y}}_m) = \frac{k'-m}{k'}\frac{1}{m(m-1)}\left(\sum_{i=1}^{m}\widehat{\overline{y}}_i^2 - m\widehat{\overline{y}}_m^2\right) = \frac{k'-m}{k'}\frac{1}{m(m-1)}\sum_{i=1}^{m}(\widehat{\overline{y}}_i - \widehat{\overline{y}}_m)^2$$

es un estimador insesgado de $V(\overline{y}_m)$.

Demostración.

Se aplica el hecho de que en m.a.s., la cuasivarianza muestral $\dfrac{1}{(m-1)}\sum_{i=1}^{m}(\widehat{\overline{y}}_i - \widehat{\overline{y}}_m)^2$ es insesgado para la cuasivarianza poblacional $\dfrac{1}{(k'-1)}\sum_{i=1}^{k'}(\widehat{\overline{y}}_i - \overline{y})^2$.

Ejemplo 6.4.

Supongamos los siguientes datos:

Obs	1	2	3	4	5	6	7	8	9	10	11	12
y	2	3	2	5	6	8	5	4	6	5	2	1

Tabla 6.3. Datos de ejemplo

Se desea obtener una muestra sistemática de tamaño $n = 6$. Para poder estimar la varianza utilizaremos el método de las muestras interpenetrantes: Tomaremos $m = 2$ muestras sistemáticas, cada una de tamaño $n' = \dfrac{n}{m} = \dfrac{6}{2} = 3$. Así, k' queda definido como $k' = \dfrac{N}{n'} = \dfrac{12}{3} = 4$.

Supongamos que después de escoger el arranque aleatorio de cada una de las dos muestras entre los $k' = 4$ primeros números, son elegidos el 1 y el 3. Entonces las dos muestras sistemáticas son $\{1, 5, 9\}$ y $\{3, 7, 11\}$ que corresponden a los valores de y $\{2, 6, 6\}$ y $\{2, 5, 2\}$. Así, $\widehat{\overline{y}}_1 = \dfrac{1}{3}(2 + 6 + 6) = 4.667$, y $\widehat{\overline{y}}_2 = \dfrac{1}{3}(2 + 5 + 2) = 3$. por lo tanto $\widehat{\overline{y}}_m = \dfrac{1}{m}\sum_{i=1}^{m}\widehat{\overline{y}}_i = \dfrac{1}{2}(4.667 + 3) = 3.8333$ y la estimación de la varianza será

ISBN 9788496866133

$$\widehat{V}(\overline{y}_m) = \frac{k'-m}{k'}\frac{1}{m(m-1)}\left(\sum_{i=1}^{m}\widehat{\overline{y}}_i - m\widehat{\overline{y}}_m^2\right) = \frac{4-2}{4}\frac{1}{2(2-1)}\left((4.667^2+3^2) - 2\cdot 3.83^2\right) = 0.36.$$

Como en este caso tenemos toda la población, podemos calcular la varianza real del estimador. Como $\overline{y} = \frac{1}{12}(2+...+1) = 4.083$ entonces se tiene que

$$V(\widehat{\overline{y}}_m^2) = \frac{k'-m}{k'}\frac{1}{m(k'-1)}\sum_{i=1}^{k'}(\widehat{\overline{y}}_i - \overline{y})^2 =$$

$$=\frac{4-2}{4}\frac{1}{2\cdot 4}[(4.667-4.083)^2 + (5.333-4.083)^2 + (3-4.083)^2 + (3.333-4.083)^2] = 0.2384.$$

6.8 Tablas de fórmulas

MUESTREO SISTEMÁTICO		
Parámetro poblacional	$\overline{y}$	$N\overline{y}$
Estimador	$\widehat{\overline{y}} = \frac{1}{n}\sum_{j=0}^{n-1} y_{i+jk}$	$N\widehat{\overline{y}}$
Varianza	$\frac{1}{k}\sum_{i=1}^{k}(\widehat{\overline{y}}_i - \overline{y})^2$	$N^2V(\widehat{\overline{y}})$
Estimador muestras interpenetrantes	$\widehat{\overline{y}}_m = \frac{1}{m}\sum_{i=1}^{m}\widehat{\overline{y}}_i$	$N^2\widehat{\overline{y}}_m$
Varianza muestras interpenetrantes	$\frac{k'-m}{k'}\frac{1}{m(k'-1)}\sum_{i=1}^{k'}(\widehat{\overline{y}}_i - \overline{y})^2$	$N^2V(\widehat{\overline{y}}_m)$
Estimador de la Varianza muestras interpenetrantes	$\frac{k'-m}{k'}\frac{1}{m(m-1)}\left(\sum_{i=1}^{m}\widehat{\overline{y}}_i^2 - m\overline{y}_m^2\right)$	$N^2V(\widehat{\overline{y}}_m)$

ISBN 9788496866133

Y para la proporción:

Parámetro poblacional	p
Estimador	$\widehat{p} = \frac{1}{n} \sum_{j=0}^{n-1} y_{i+jk}$
Varianza	$\frac{1}{k} \sum_{i=1}^{k} (\widehat{p}_i - p)^2$
Estimador muestras interpenetrantes	$\widehat{p}_m = \frac{1}{m} \sum_{i=1}^{m} \widehat{p}_i$
Varianza muestras interpenetrantes	$\frac{k' - m}{k'} \frac{1}{m(k'-1)} \sum_{i=1}^{k'} (\widehat{p}_i - p)^2$
Estimador de la Varianza muestras interpenetrantes	$\frac{k' - m}{k'} \frac{1}{m(m-1)} \left(\sum_{i=1}^{m} \widehat{p}_i^2 - m\widehat{p}_m^2 \right)$

ISBN 9788496866133

6.9 Obtención de muestras por muestreo sistemático con SAS

En este caso se puede utilizar el procedimiento surveyselect de manera sencilla. Suponiendo que se desea una muestra sitemática de tamaño $n = 10$, la sintaxis del programa es como sigue:

```
proc surveyselect data=datos out=muestra method=sys n=10;
run;
```

Si lo que se desea es obtener muestras sistemáticas por el método de las muestras interpenetrantes, se puede utilizar la opción rep=m, donde m es el número de muestras interpenetrantes a extraer. Por ejemplo, si en lugar de extraer una muestra sistemática de tamaño 10 se extraen dos de tamaño 5 cada una, se utilizaría la siguiente sintaxis:

```
proc surveyselect data=datos out=muestra method=sys n=5 rep=2;
run;
```

El archivo muestra contiene las variables presentes en el archivo datos para las observaciones muestrales, y la variable llamada Replicate que representa el índice o número de réplica o muestra interpenetrante.

6.10 Estimación en muestreo sistemático con SAS

Como se ha visto, al no disponer de un estimador insesgado de la varianza se puede optar por el estimador usual de la varianza que se utilizaría bajo muestreo aleatorio simple sin reemplazamiento. Así, el programa a utilizar podría ser la misma macro mas (puesto que el estimador es la media muestral):

```
%mas(archivo,variable,npobla);
```

donde los valores a introducir archivo, variable y npobla son los mismos que se estudiaron al ver la macro mas en el tema sobre muestreo aleatorio simple sin reemplazamiento.

Otra posibilidad, si se dispone de información proviniente del método de muestras interpenetrantes, es utilizar la siguiente macro, diseñada para construir los estimadores en ese caso.

6.10.1 Estimación con la macro estimpen

Para utilizar esta macro, el archivo de datos muestrales debe contener la variable de interés y una variable llamada **replicate** para poder utilizar también archivos de muestra que puedan provenir del proc surveyselect. Esta variable contiene el número índice de la muestra sistemática (pues se supone que se han recogido varias) a la que pertenece cada observación, siendo replicate números enteros: replicate=1,2,...

La macro estimpen se utiliza con la siguiente sintaxis:

ISBN 9788496866133

```
%estimpen(archivo,variable,m,N);
```

donde

archivo es el archivo que contiene las muestras.

variable es la variable de interés .

m es el número de muestras interpenetrantes.

N es el tamaño poblacional.

La macro presenta en la ventana LOG la estimación de la media o proporción y del total, junto con sus varianzas e intervalos de confianza.

ISBN 9788496866133

6.11 Ejercicios resueltos

Ejercicio 5.1.

En una plantación de calabacín se pretende estimar el número de calabacines con virus . El terreno se divide en 10 áreas con 4 calabacines que constan de los siguientes valores:

Área	1	2	3	4	5	6	7	8	9	10	11	12
nº calabacines con virus	1	1	1	0	2	2	3	3	2	4	3	4

Comparar los siguientes métodos de muestreo respecto a la varianza obtenida:

a) m.a.s. con $n = 4$.

b) m.a.s.r. con $n = 4$.

c) Muestreo sistemático con $n = 4$.

d) Muestreo estratificado, dividiendo la población en 4 partes por el orden en que está y tomando una observación en cada estrato.

e) Muestreo estratificado, dividiendo la población en los siguientes estratos: áreas 1,5,9, áreas 2,6,10, áreas 3,7,11, áreas 4,8,12.

---

a) Hay $N\widehat{\overline{y}} = 26$ calabacines en todo el terreno, e $\widehat{\overline{y}} = \dfrac{26}{12} = 2.166$. Además, $S^2 = 1.60$.

La varianza del estimador del total bajo m.a.s. con tamaño $n = 4$ es:

$$V(N\widehat{\overline{y}}) = N^2 \frac{N-n}{N} \frac{S^2}{n} = 12^2 \frac{12-4}{12} \frac{1.60}{4} = 38.4.$$

b) La varianza del estimador del total bajo m.a.s.r. con tamaño $n = 4$ es:

$$V(N\widehat{\overline{y}}) = N^2 \frac{\sigma^2}{n} = N^2 \frac{N-1}{N} \frac{S^2}{n} = 12^2 \frac{12-1}{12} \frac{1.60}{4} = 52.8.$$

c) Bajo muestreo sistemático, la varianza es

$$V(N\widehat{\overline{y}}_{sis}) = N^2 \frac{1}{k} \sum_{i=1}^{k} (\widehat{\overline{y}}_i - \overline{y})^2,$$

donde $\widehat{\overline{y}}_i$ es la media muestral en la muestra sistemática i.

Como $n = 4$ y $N = 12$, hay $k = \dfrac{12}{4} = 3$ muestras sistemáticas posibles, según el arranque sea de las áreas 1,2, o 3:

Muestra sistemática 1:

1,0,3,4 con media $\widehat{\overline{y}}_1 = 2$.

Muestra sistemática 2:

1,2,3,3 con media $\widehat{\overline{y}}_2 = 2.25$.

ISBN 9788496866133

Muestra sistemática 2:

1,2,2,4 con media $\widehat{\overline{y}}_3 = 2.25$.

Por ello:

$$V(N\widehat{\overline{y}}_{sis}) = N^2\frac{1}{k}\sum_{i=1}^{k}(\widehat{\overline{y}}_i - \overline{y})^2 = 12^2\frac{1}{3}[(2-2.166)^2 + (2.25-2.166)^2 + (2.25-2.166)^2\ ] = 2$$

mucho menor que la obtenida bajo m.a.s. Esto es debido a la tendencia ordenada creciente en la población.

d) Las cuasivarianzas respectivas de los 4 estratos son :

$S_1^2 = 0$, $S_2^2 = 1.33$,$S_3^2 = 0.33$,$S_4^2 = 0.33$.

La varianza del estimador del total es:

$$V(N\overline{y}_{st}) = \sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} = 3(3-1)(0 + 1.33 + 0.33 + 0.33) = 12.$$

e) Con esta configuración de estratos, se tiene:

Las cuasivarianzas respectivas de los 4 estratos son :

$S_1^2 = 0.33$, $S_2^2 = 2.33$,$S_3^2 = 1.33$,$S_4^2 = 4.33$.

La varianza del estimador del total es:

$$V(N\overline{y}_{st}) = \sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} = 3(3-1)(0.33 + 2.33 + 1.33 + 4.33) = 50.$$

Ejercicio 5.2.

En una granja se desea estudiar la comida que les sobra a los caballos una hora después de darles de comer. Están dispuestos en el establo numerados del 1 al 24 y para poder estimar la varianza se toman dos muestras sistemáticas de tamaño 4. Se observan los siguientes resultados en kilos: Muestra sistemática 1: 2,1,1,0.5. Muestra sistemática 2: 1,1,2,1.

a) Dar un intervalo de confianza al 95% para la estimación del promedio de comida que deja cada caballo.

b) Supongamos que se le dan 3 kilos inicialmente a cada caballo. Dar un intervalo de confianza para la estimación del promedio de kilos que come cada caballo.

---

a) La media obtenida en la muestra sistemática 1 es $\widehat{\overline{y}}_1 = 1.125$ y en la segunda muestra, $\widehat{\overline{y}}_2 = 1.25$.

El estimador de la media poblacional es $\widehat{\overline{y}}_m = \frac{1}{2}(\widehat{\overline{y}}_1 + \widehat{\overline{y}}_2) = 1.1875$.

Utilizando la estimación de la varianza del estimador por muestras interpenetrantes, se tiene que $k' = 6$ y $m = 2$. Así,

$$\widehat{V}(\overline{y}_m) = \frac{k'-m}{k'}\frac{1}{m(m-1)}\left(\sum_{i=1}^{m}\widehat{\overline{y}}_i^2 - m\overline{y}_m^2\right) = 0.00026.$$

El intervalo de confianza será

$$(1.1875 - 1.96\sqrt{0.00026}, 1.1875 + 1.96\sqrt{0.00026}) = (1.08, 1.28).$$

ISBN 9788496866133

b) Sea x en este caso la variable "kilos comidos" . Entonces $x = 3 - y$. Las medias sistemáticas quedan:

$\widehat{\overline{x}}_i = 3 - \widehat{\overline{y}}_i$

y entonces

$\widehat{\overline{x}}_m = 3 - \widehat{\overline{y}}_m = 1.8125$, por lo cual $V(\widehat{\overline{x}}_m) = V(\widehat{\overline{y}}_m)$ y el intervalo de confianza tiene la misma anchura y queda:

$(1.7, 1.9)$.

Ejercicio 5.3.

Se desea estimar la proporción de pisos con niños en una determinada comunidad. En ésta se numeran los pisos del 1 al 16. Los datos aparecen en la tabla:

Piso	1	2	3	4	5	6	7	8
Niños	SI	NO	NO	NO	SI	NO	NO	NO

Piso	9	10	11	12	13	14	15	16
Niños	SI	SI	SI	NO	SI	NO	SI	NO

a) Si se utiliza muestreo sistemático con $n = 4$, y el número base para el arranque aleatorio es el número 3, ¿cuál es la muestra obtenida?. Dar la estimación de la proporción poblacional con esa muestra y una aproximación a su varianza .

b) Suponer que se utiliza el método de las muestras interpenetrantes, con dos muestras de tamaño 2, y los arranques aleatorios resultan ser 4 y 5. Dar un intervalo de confianza al 95% para la proporción poblacional. Idem, para el total de niños en esa comunidad.

c) Calcular la varianza exacta del estimador suponiendo el tipo de estimación del apartado b) (dos muestras sistemáticas de tamaño $n = 2$).

---

a) Con $n = 4$, se tiene que $k = 16/4 = 4$. Si el arranque es el número 3, se toman las observaciones 3, $3 + 4$, $3 + 2 \cdot 4$, y $3 + 3 \cdot 4$, es decir, las observaciones 3,7,11 y 15.

La proporción estimada con esta muestra sistemática es $\widehat{p} = \dfrac{1}{2}$. Al no tener más que una muestra sistemática, y no existir seguramente ordenamientos cíclicos en los datos, utilizaremos la aproximación dada por el estimador de la varianza en m.a.s. Si los datos siguen un orden aleatorio con respecto a la existencia de niños, será una aproximación correcta, mientras que si existe una tendencia creciente o decreciente, será una estimación sobrevalorada de la varianza. En cualquier caso, es una estimación conservadora correcta.

$$\widehat{V}(\widehat{p}) \simeq \frac{N-n}{N}\frac{\widehat{pq}}{n-1} = \frac{16-4}{16}\frac{0.25}{4-1} = 0.0625.$$

b) En la primera muestra están las observaciones 4 y 12, y en la segunda las observaciones 5 y 13. En la primera es $\widehat{p}_1 = 0$ y en la segunda $\widehat{p}_2 = 1$. Se obtiene $\widehat{p}_m = \dfrac{1}{2}(0 + 1) = 0.5$.

La varianza estimada es, teniendo en cuenta que $k' = \dfrac{N}{n'} = \dfrac{16}{2} = 8$ y $m = 2$,

ISBN 9788496866133

$$\widehat{V}(p_m) = \frac{k'-m}{k'}\frac{1}{m(m-1)}\left(\sum_{i=1}^{m}\widehat{p}_i^2 - m\widehat{p}_m^2\right) = 0.1875.$$

Un intervalo de confianza será excesivamente ancho, pues es $1.96\sqrt{0.1875.} = 0.8$.

c) Hay 8 muestras sistemáticas, con proporciones respectivas:

1,0.5,0.5,0,1,0,0.5,0. Sabiendo que $p = \frac{7}{16} = 0.4375$, se calcula la varianza exacta:

$$V(\widehat{p}_m) = \frac{k'-m}{k'}\frac{1}{mk'}\sum_{i=1}^{k'}(\widehat{p}_i - p)^2 = \frac{8-2}{8}\frac{1}{2\cdot 8}[2(1-0.4375)^2 + 3(0.5-0.4375)^2 + 3(0-0.4375)^2] = 0.076.$$

Ejercicio 5.4.

Se desea estimar la cantidad promedio facturada en un archivo de 12 facturas. Antes de acceder a ellas se puede elaborar una ordenación por fecha o dejarlas como están, aproximadamente ordenadas por el montante. En la tabla siguiente aparecen ordenadas por fecha:

2	5	3	7	2	6	1	8	12	2	3	4

Y en esta tabla , por montante:

1	2	2	2	3	3	4	5	6	7	8	12

Si se realiza muestreo sistemático con $n = 3$, ¿cuál será la varianza del estimador en cada caso?

Hay $k = \frac{12}{3} = 4$ muestras sistemáticas. En el primer caso, sus medias respectivas son $\widehat{\overline{y}}_1 = 5.33$, $\widehat{\overline{y}}_2 = 4.33$, $\widehat{\overline{y}}_3 = 2.33$, $\widehat{\overline{y}}_4 = 6.33$.

Por lo tanto, la varianza será

$$V(\widehat{\overline{y}}) = \frac{1}{k}\sum_{i=1}^{k}(\widehat{\overline{y}}_i - \overline{y})^2 = \frac{1}{4}[(5.33-4.58)^2 + ... + (6.33-4.58)^2] = 2.1875.$$

En el segundo caso:

$\widehat{\overline{y}}_1 = 3.33$, $\widehat{\overline{y}}_2 = 4$, $\widehat{\overline{y}}_3 = 4.66$, $\widehat{\overline{y}}_4 = 6.33$.

Entonces,

$V(\widehat{\overline{y}}) = \frac{1}{k}\sum_{i=1}^{k}(\widehat{\overline{y}}_i - \overline{y})^2 = 1.24$, mucho menor, debido a la ordenación. Sería similar si la ordenación fuera descendente.

Ejercicio 5.5.

En un estudio sobre consumo eléctrico en un pueblo, éste se divide en tres zonas A, B y C. En el área A se realiza m.a.s. de 10 familias de las 100 de las que consta el área A y se obtiene un promedio de $\widehat{\overline{y}} = 20$ kilowatios al mes en esas 10 familias, con una cuasivarianza de 25.

ISBN 9788496866133

En el área B hay 200 familias, y se realiza m.a.s.r. de 20 familias, obteniendo una media muestral de 18 kilowatios, con una cuasivarianza muestral de 30.

En el área C hay 300 familias, que se ordenan geográficamente, y se ha decidido realizar muestreo sistemático replicado de dos muestras de 20 familias cada una. En la primera muestra la media obtenida en las familias muestreadas es de 15 kilowatios y en la segunda de 19 kilowatios.

a) Calcular el estimador global de la media de consumo eléctrico por familia en el pueblo y dar una estimación para su varianza.

b) Calcular el estimador del total del consumo eléctrico en el pueblo.

---

a) Se trata de muestreo estratificado, con pesos $W_1 = \dfrac{100}{600} = 0.166$, $W_2 = \dfrac{200}{600} = 0.333$ y $W_3 = \dfrac{300}{600} = 0.5$.

En el primer estrato (A), la estimación es $\widehat{\overline{y}}_1 = 20$, con varianza estimada, por ser m.a.s., de

$$\widehat{V}(\widehat{\overline{y}}_1) = \frac{N_1 - n}{N_1}\frac{s_1^2}{n} = \frac{100-10}{100}\frac{25}{10} = 2.25.$$

En el segundo estrato (B), la estimación es $\widehat{\overline{y}}_2 = 18$, con varianza estimada, por ser m.a.s.r., de

$$\widehat{V}(\widehat{\overline{y}}_2) = \frac{s_1^2}{n} = \frac{30}{20} = 1.5.$$

En el tercer estrato la estimación es $\widehat{\overline{y}}_3 = \dfrac{1}{2}(15+19) = 17$, $k' = \dfrac{N}{n'} = 15$ y la varianza se estima por

$$\widehat{V}(\widehat{\overline{y}}_3) = \frac{15-2}{15}\frac{1}{2(2-1)}\left(\sum_{i=1}^{m}\widehat{\overline{y}}_i^2 - 2\cdot 17^2\right) = 3.4.$$

El estimador global es

$$\widehat{\overline{y}}_{st} = W_1\ \widehat{\overline{y}}_1 + W_2\ \widehat{\overline{y}}_2 + W_3\ \widehat{\overline{y}}_3 = 17.83.$$

Y su varianza estimada,

$$\widehat{V}(\widehat{\overline{y}}_{st}) = W_1\ \widehat{V}(\widehat{\overline{y}}_1) + W_2\ \widehat{V}(\widehat{\overline{y}}_2) + W_3\ \widehat{V}(\widehat{\overline{y}}_3) = 2.6.$$

b) El estimador del total será $N\widehat{\overline{y}}_{st} = 600 \cdot 17.83 = 10698$ y su varianza $\widehat{V}(N\widehat{\overline{y}}_{st}) = N^2\widehat{V}(\widehat{\overline{y}}_{st}) = 936000$.

Ejercicio 5.6.

Supongamos la población 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4.

a) Extraer 2 muestras sistemáticas de tamaño $n = 4$ con SAS, con semillas 1234 y 1235 y calcular el valor del estimador.

b) Reordenar el archivo aleatoriamente con la semilla 1234 y volver a ejecutar el apartado a) .

d) Ordenar el archivo por la propia variable y repetir el apartado a).

---

a) En primer lugar se crea el archivo:

ISBN 9788496866133

```
data uno;
input y @@;
cards;
1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4
;
```

A continuación se extraen las muestras y se estima:

```
proc surveyselect data=uno method=sys n=4 out=muestra seed=1234;
run;
proc means data=muestra;run;
```

Se obtiene, para la semilla 1234, la media 1 y para 1235 la media 2.

b) Para ordenar el archivo aleatoriamente:

```
data uno;set uno;u=ranuni(1234);
proc sort data=uno;by u;
```

Luego se obtienen las muestras y se estima:

```
proc surveyselect data=uno method=sys n=4 out=muestra seed=1234;
run;
proc means data=muestra;run;
```

Obteniendo medias respectivas de 2.5 y 3.

c) Para ordenar por la variable, extraer y estimar:

```
proc sort data=uno;by y;
proc surveyselect data=uno method=sys n=4 out=muestra seed=1234;
run;
proc means data=muestra;run;
```

Obteniendo 2.5 y 2.5, que además coincide con la media poblacional.

Ejercicio 5.7.

En el archivo SAS pesos se dispone de la información del peso y estatura de 1200 personas. La varianza exacta del estimador de la media del peso por m.a.s. con $n = 40$ se calculó en el Ejercicio 4.10, y era 3.96. Si se desea comparar con muestreo sistemático con $n = 40$, por ejemplo con el método de dos muestras interpenetrantes de tamaño $n' = 20$ cada una, haría falta calcular la media muestral obtenida para las $\dfrac{1200}{20} = 600$ muestras para calcular la varianza del estimador.

En cambio, aproximaremos la varianza del estimador sistemático, tomando varias veces muestras sistemáticas de tamaño $n = 40$ y calculando el valor del estimador cada vez, y calculando

ISBN 9788496866133

finalmente su varianza (esto se puede hacer así al disponer de toda la población, pues en la práctica solo se puede estimar esa varianza a través de una muestra).

a) Ordenar aleatoriamente la población mediante una uniforme con semilla 12345, y realizar el proceso de extraer una muestra sistemática de tamaño $n = 40$ con semilla 1234 y obtener su media muestral.

La macro variasis realiza este proceso tantas veces como se indique, presentando en la ventana OUTPUT la varianza del estimador sobre todas las pruebas y un histograma del esimador. Realizarla con 500 repeticiones.

La sintaxis es

```
%variasis(archivo,vary,n,repe);
```

b) Hacer lo mismo, pero ordenando los datos por la variable estatu (estatura). Comparar.

c) Utilizar el método de las muestras interpenetrantes tomando dos muestras de tamaño $n' = 20$ y obteniendo las estimaciones y errores de muestreo con la macro estimpen.

d) Realizar muestreo estratificado con sexo como la variable de estrato y con muestreo sistemático de $n = 20$ en cada estrato y utilizar la macro estimestrat para construir intervalos de confianza para la media basados en la estimación de la varianza como si fuera m.a.s..

---

a) Para ordenar los datos aleatoriamente, se añade la variable uniforme con semilla 12345 y después se ordena por ella.

```
data pesos2;
set pesos;
u=ranuni(12345);
run;
proc sort data=pesos2;by u;run;
```

A continuación se extrae la muestra y se calcula la media muestral:

```
proc surveyselect data=pesos2 out=muestra method=sys n=40 seed=12345;
run;
proc means data=muestra;var peso;run;
```

Se obtiene una media de 72.42.

Para realizar el proceso 500 veces repetidas,

```
%variasis(pesos2,peso,40,500);
```

Obteniendo una varianza del estimador en nuestro caso (la semilla varía), de 3.30.

b) Se trata de observar si ordenar los datos por una variable auxiliar relacionada con la variable de interés puede mejorar al estimador por muestreo sistemático. Ordenando:

ISBN 9788496866133

```
proc sort data=pesos2;by estatu;run;
```

y luego ejecutando la macro:

```
%variasis(pesos2,peso,40,500);
```

Se obtiene una varianza aproximada del estimador mucho menor, de 2.50 en nuestro caso.

c) Para obtener una muestra y estimar por el método de las muestras interpenetrantes, se realiza el proc surveyselect con la opción rep:

```
proc surveyselect data=pesos2 out=muestra method=sys n=20 rep=2 seed=12345;
run;
%estimpen(muestra,peso,2,1200);
```

Obteniendo:

```
                    ******************************
                        Formulas para medias
                    ******************************

------------------------------------------------------
Media sistemática por muestras interpenetrantes=67.925
------------------------------------------------------

------------------------------------------------------------
Varianza del estimador de la media =5.175625
------------------------------------------------------------

-------------------------------------------------------------------------
Intervalo de confianza para la media: (63.466 , 72.384 )
-------------------------------------------------------------------------
```

d) Basta añadir en el proc surveyselect la opción strata y ejecutar la macro sobre el archivo de salida:

```
proc surveyselect data=pesos2 out=muestra method=sys n=20 ;
strata estrato;
run;
/*estimestrat(muestra,archivo2,vary,vartama,post,indicador,nestratos,N);*/
%estimestrat(muestra,.,peso,tama,1,2,2,1200);
```


ISBN 9788496866133

6.12 Ejercicios propuestos

1) En la auditoría realizada a una empresa de servicios informáticos, se desea comparar las facturas con el valor que aparece en los libros de contabilidad. La intención es comparar un 12% de las facturas. Supongamos las diferentes situaciones:

i) Las cuentas están ordenadas por orden cronológico, y las facturas más antiguas tienen tendencia a tener valores más pequeños que las más actuales.

ii) Las cuentas siguen un orden que puede considerarse aleatorio.

iii) Las cuentas están agrupadas por departamentos y dentro de cada uno de éstos por orden cronológico (con la tendencia mencionada en el apartado i)).

En cada una de las tres situaciones, razonar qué diseño es más apropiado: sistemático o m.a.s. para seleccionar la muestra de facturas.

2) En una huerta abandonada se dispone de las plantas tomate, lechuga, col y rábano. Se dispone de 12 parcelas. En las parcelas 1,2,3, y 12 las plantas tienen más riego natural. Se numeran los items en cada columna hacia abajo y de izquierda a derecha:

1	2	3	4	5	6	7	8	9	10	11	12
T	T	T	T	T	T	T	T	T	T	T	T
L	L	L	L	L	L	L	L	L	L	L	L
C	C	R		R	C	C	R	R	R	C	C
R	R	C		R	R	C	C			C	R
C	C			C		R					

Se duda entre tomar una muestra sistemática o bien una m.a.s. del 20%. S pide decir para cuáles de estas características la muestra sistemática será más precisa que el m.a.s.:

a) proporción de tomates.

b) proporción de coles o rábanos.

c) proporción de plantas con más riego natural.

3) Una compañía está interesada en estimar cuántos empleados están a favor de promocionar el deporte patrocinando a un equipo de atletismo. Para ello se seleccionó una muestra tomando cada décimo trabajador que salía del edificio al final de la jornada laboral un día concreto. Sa-biendo que hay 2000 empleados, y que de los encuestados 132 estaban de acuerdo en promocionar el deporte, contestar a las siguientes cuestiones:

a) Estimar el número de trabajadores que están a favor de la promoción deportiva y establecer el error de muestreo.

ISBN 9788496866133

b) Determinar el tamaño muestral necesario para estimar la proporción de trabajadores a favor de la promoción con un error de meustreo de 0.01. ¿Cómo se obtendrá la muestra en este caso?.

4) Durante un año, se comprobaron diversos procedimientos de tratamiento de cierta enfermedad en 150 personas en un hospital especializado. De la lista de pacientes sometidos a dichos procedimientos, se obtuvieron tres muestras sistemáticas cada 50 pacientes y se midió la tensión arterial, con el siguiente resultado:

Muestra	Paciente	Tensión
1	30	
	80	
	130	
2	17	
	67	
	117	
3	22	
	72	
	122	

a) Estimar la tensión arterial media de todos los pacientes y dar un I.C. para la estimación.

b) Suponiendo suficientemente precisa la muestra anterior, decir cuál sería el error de muestreo si se hubieran tomado 6 muestras replicadas en lugar de 3.

5) Una tienda con cuatro departamentos tiene las cuentas corrientes ordenadas por departamentos, con las cuentas vencidas al principio de la lista de cada departamento. Para un día particular, las cuentas aparecen de la forma siguiente:

	Departamento			
	1	2	3	4
Nº de cuentas	1-11	12-20	21-28	29-40
Cuentas vencidas	1-4	12-14	21-25	29-32

La tienda desea estimar la proporción de cuentas vencidas mediante muestreo sistemático.

a) Determinar todas las posibles muestras sistemáticas con k=10 y calcular la varianza exacta

ISBN 9788496866133

de la proporción muestral.

b) Comparar los resultados con la varianza aproximada que habría sido obtenida con una m.a.s. de tamaño 4 de esta población. ¿Qué conclusiones se pueden obtener?. Propón un diseño sistemático alternativo que mejore la situación.

6) Dada una población de tamaño N, se considera la variable $y_i = 3 + 3i$, para $i = 1, ..., N$. Se extrae una muestra sistemática con tamaño $n = \frac{N}{k}$ entero, donde k es también un número entero.

a) Calcular $\overline{y}$, y S_y^2.

b) Calcular la varianza del estimador de la media bajo muestreo sistemático y la varianza del estimador de la media bajo m.a.s.

7) Las 36 viviendas de una calle en cierto pueblo de la Costa del Sol numeradas del 1 al 36 se ordenan alfabéticamente en un archivo de acuerdo con el apellido del padre de familia. Viviendas en las cuales el padre es extranjero ocurren en los números 3, 5-7, 11-13, 15,16, 20, 21, 22, 25-28, 30-34, y 35.

Comparar la precisión de una muestra sistemática que toma una observación cada 4, con una m.a.s. del mismo tamaño para estimar la proporción de viviendas en las cuales el jefe de familia es extranjero.

8) Se dispone de la población de cuatro valores $\{1, 2, 3, 4\}$, de los cuales se tomará una muestra de tamaño $n = 2$. Calcular la varianza del estimador sistemático de la media (por simple conteo), y la varianza del estimador de la media por m.a.s. para cada una de las ordenaciones posibles de los datos. Comparar.

9) El archivo SAS ganado contiene el número de cabezas de ganado de diverso tipo sacrificadas en 1998 en 50 provincias españolas.

a) Calcular la media poblacional por provincia de cabezas sacrificadas de ganado porcino con el proc means.

b) Extraer 3 muestras sistemáticas de tamaño $n = 10$ con las semillas respectivas 1234, 1235, 1236. Estimar para cada muestra la media poblacional por provincia del número de cabezas de ganado porcino sacrificadas.

c) Ordenar el archivo por el ganado porcino y volver a realizar el apartado b) ¿Se observa una mejora en la estimación?.

d) Realizar un muestreo sistemático equivalente en tamaño al $n = 10$ utilizado anteriormente con dos muestras replicadas. Utilizar la macro estimpen para obtener el estimador y su varianza estimada.

10) Realizar el ejercicio 4, apartado a) en SAS, con la macro estimpen.

ISBN 9788496866133

7 ESTIMACIÓN INDIRECTA

En este capítulo se tratará la estimación de las características poblacionales media, total y proporción de una variable de interés y cuando se dispone de información poblacional sobre una variable auxiliar x que se supone altamente correlada con y. La variable x, al aportar información sobre y, permitirá reducir el error en la estimación. Supondremos a lo largo del tema que se dispone de una muestra aleatoria simple sin reemplazamiento de tamaño n , de manera que a cada unidad muestreada se le mide no solamente el valor y_i, sino también el valor de la variable auxiliar x_i. Las observaciones muestrales serán por lo tanto pares (x_i, y_i).

7.1 Estimadores de razón

7.1.1 Ejemplos introductorios

A modo de presentación, a continuación se muestran tres ejemplos ilustrativos de situaciones donde se utiliza la estimación de razón.

ISBN 9788496866133

Ejemplo 7.1.

Utilización del estimador de la razón para estimar una tasa o proporción.

Supongamos que se realiza un m.a.s. sobre pueblos de una determinada comarca para estudiar el alcance de una epidemia de gripe. Se obtiene una tabla como la siguiente:

Pueblo muestreado	x =nº de habitantes	y =nº de afectados de gripe
1	600	10
2	1000	31
3	780	8
...	...	...
n	...	...

Tabla 7.1. Muestra de n pueblos.

Se desea estudiar la proporción real de afectados en la población. Es decir, se desea estimar la proporción de afectados de gripe en la comarca, que es:

$$\text{Prop. de afectados} = \frac{\text{Total de afectados en la población}}{\text{Total de habitantes en la población}} =$$

$$= \frac{N \cdot \text{Media de afectados/pueblo}}{N \cdot \text{Media de habitantes/pueblo}},$$

donde N = número de pueblos. En este caso, las unidades poblacionales son los pueblos, y a cada uno de ellos se le mide la variable y =número de afectados de gripe y la variable x =número de habitantes.

La cantidad a estimar se puede entonces denotar por

$$\frac{N\overline{y}}{N\overline{x}} = \frac{\overline{y}}{\overline{x}} = R.$$

Este cociente se denomina **razón poblacional.** Una posibilidad de estimarla es utilizar el **método de los momentos**; es decir, sustituir cada momento poblacional por su momento muestral. Así, queda

$$\widehat{R} = \frac{\widehat{\overline{y}}}{\widehat{\overline{x}}}$$

que es la razón muestral o **estimador de la razón.** Otra forma de presentarlo es

$$\widehat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}.$$

ISBN 9788496866133

En el ejemplo, $\widehat{R}$ estima la proporción de afectados en la población. Es importante tener en cuenta la observación siguiente:

La razón muestral $\widehat{R} = \dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}}$ es el estimador utilizado para la proporción de afectados. Este valor es diferente del valor obtenido si calculamos la proporción de afectados $\dfrac{y_i}{x_i}$ en cada pueblo de los muestreados, y finalmente calculamos la media de estas proporciones, que será $\dfrac{1}{n}\sum_{i=1}^{n}\dfrac{y_i}{x_i} = \widehat{\overline{R}}$. Este valor se conoce como razón media muestral $\widehat{\overline{R}}$, y aunque puede utilizarse como un estimador de la proporción poblacional, su sesgo y varianza suelen ser mayores que los del estimador usual de razón $\widehat{R}$. El motivo es que en realidad lo que se está estimando con $\widehat{\overline{R}}$ es su equivalente poblacional, la razón media $\overline{R} = \dfrac{1}{N}\sum_{i=1}^{N}\dfrac{y_i}{x_i}$. En el ejemplo, $\overline{R}$ es la **proporción media por pueblo** de afectados por gripe, que no es lo mismo que la proporción poblacional pues $\overline{R} = \dfrac{1}{N}\sum_{i=1}^{N}\dfrac{y_i}{x_i} \neq \dfrac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = R.$

Ejemplo 7.2.

Utilización del estimador de razón para estimar una media.

Supongamos que se realiza un m.a.s. sobre una población de individuos a los que se mide el peso y estatura. Se obtiene la tabla de datos 7.2, y se desea estimar la media poblacional del peso de los individuos de la población. Supongamos que la media de estatura de la población $\overline{x}$ es conocida.

Individuo muestreado	y=peso	x=estatura
1	80	1.80
2	65	1.67
3	69	1.70
...	...	...
n	...	...

Tabla 7.2. Muestra de n individuos.

Un estimador posible de $\overline{y}$ es $\widehat{\overline{y}}$, ya que se ha utilizado m.a.s. Otra posibilidad es la siguiente. Como $\widehat{R}$ es un estimador de R, tenemos que $\widehat{R} \simeq R$ y por lo tanto $\widehat{R} = \dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}} \simeq \dfrac{\overline{y}}{\overline{x}}$. Así, $\overline{y} \simeq \dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}}\overline{x} = \widehat{R}\overline{x}$ y por lo tanto hemos construido un estimador de $\overline{y}$ llamado el **estimador de razón de** $\overline{y}$, denotado por

$\overline{y}_R = \widehat{R}\overline{x}$.

ISBN 9788496866133

Como en el ejemplo anterior, se podrían comparar los dos estimadores $\widehat{\overline{y}}$ e $\overline{y}_R$ en cuanto a varianzas.

Nótese que el estimador de razón requiere, adicionalmente al estimador básico $\widehat{\overline{y}}$:

a) Tomar datos de la variable auxiliar x a cada unidad muestral.

b) Conocer la media poblacional de la variable auxiliar, $\overline{x}$.

Por lo tanto, el estimador de razón puede mejorar la precisión del estimador usual $\widehat{\overline{y}}$, pero es a cambio de requerir mayor información, lo que puede redundar en costes adicionales.

Ejemplo 7.3.

Utilización del estimador de razón para estimar un total sin conocer N.

Supongamos que se dispone de un camión con N cestas de fresas, N desconocido. Deseamos conocer la cantidad de azúcar de las fresas del camión, pues está relacionada con el valor monetario de la carga. En este caso podemos conocer el peso de la carga del camión. Si definimos como unidades elementales las cestas de fresas, como variable de interés y = cantidad de azúcar de cada cesta de fresas, y como variable auxiliar x = peso de cada cesta de fresas, se tiene que $N\overline{y}$ es la cantidad a estimar, y $N\overline{x}$ es el peso total de la carga, conocido. Sea la razón poblacional $R = \dfrac{N\overline{y}}{N\overline{x}}$ y sea el estimador de la razón $\widehat{R} = \dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}}$. Como se ha visto en el ejemplo anterior, un estimador de la media de y es el estimador de razón $\overline{y}_R = \widehat{R}\overline{x}$. Por lo tanto, un estimador del total será

$N\overline{y}_R = \widehat{R}N\overline{x}$.

Al ser $N\overline{x}$ conocido, se puede utilizar el estimador denominado **estimador de razón del total**

$N\overline{y}_R = \dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}}N\overline{x}$.

Para resolver el problema, se toma una m.a.s. de n cestas , se mide la cantidad de azúcar y_i y el peso x_i de cada una de ellas , se calcula el estimador de la razón $\dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}}$ y a continuación $N\overline{y}_R$. Nótese que en este ejemplo, aún para calcular el total, no ha sido necesario conocer N . También se intuye que este estimador funcionará bien, pues la correlación entre la cantidad de azúcar y el peso de cada cesta de fresas se supone alta.

7.1.2 Definición del estimador de razón.

Supongamos que se realiza un muestreo aleatorio simple sin reemplazamiento de tamaño n sobre una población de tamaño N, y se observan los valores del par (x_i, y_i) para cada unidad muestral. Además, se supone conocida la media poblacional $\overline{x}$ de la variable auxiliar x.

ISBN 9788496866133

$R = \dfrac{N\overline{y}}{N\overline{x}} = \dfrac{\overline{y}}{\overline{x}} = \dfrac{\sum\limits_{i=1}^{N} y_i}{\sum\limits_{i=1}^{N} x_i}$ es la **razón poblacional.**

$\widehat{R} = \dfrac{N\widehat{\overline{y}}}{N\widehat{\overline{x}}} = \dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}} = \dfrac{\sum\limits_{i=1}^{n} y_i}{\sum\limits_{i=1}^{n} x_i}$es la **razón muestral** o el **estimador de la razón.**

$\overline{y}_R = \widehat{R}\overline{x} = \dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}}\overline{x}$ es el **estimador de razón de la media.**

$N\overline{y}_R = \widehat{R}N\overline{x} = \dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}}N\overline{x}$ es el **estimador de razón del total .**

Teorema 7.1 (sesgo de los estimadores de razón).

El estimador de la razón es un estimador sesgado, con sesgo $B(\widehat{R}) = \dfrac{-cov(\widehat{R}, \widehat{\overline{x}})}{\overline{x}}$.

Además el sesgo del estimador de razón de la media $\overline{y}_R$ es $B(\overline{y}_R) = -cov(\widehat{R}, \widehat{\overline{x}})$.

Demostración.

El estimador es sesgado pues $E(\widehat{R}) = E\left(\dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}}\right)$, mientras que $R = \dfrac{\overline{y}}{\overline{x}} = \dfrac{E(\widehat{\overline{y}})}{E(\widehat{\overline{x}})}$. La igualdad $E\left(\dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}}\right) = \dfrac{E(\widehat{\overline{y}})}{E(\widehat{\overline{x}})}$ no se suele cumplir , salvo casos especiales. Por lo tanto en general $E(\widehat{R}) \neq R$.

Calculemos este sesgo. Sea

$$B(\widehat{R}) = E(\widehat{R}) - R = E(\widehat{R}) - \frac{\overline{y}}{\overline{x}} = \frac{1}{\overline{x}}\left(\overline{x}E(\widehat{R}) - \overline{y}\right) = \frac{1}{\overline{x}}\left(\overline{x}E(\widehat{R}) - E(\widehat{\overline{y}})\right)$$

$$= \frac{1}{\overline{x}}\left(E(\widehat{\overline{x}})E(\widehat{R}) - E(\widehat{R}\widehat{\overline{x}})\right) = \frac{-cov(\widehat{R}, \widehat{\overline{x}})}{\overline{x}}.$$

Como consecuencia, el estimador de razón de la media $\overline{y}_R = \widehat{R}\overline{x}$ será sesgado. Concretamente, $E(\overline{y}_R) - \overline{y} = \overline{x}\left[E(\widehat{R}) - R\right] = -cov(\widehat{R}, \widehat{\overline{x}})$. Del mismo modo ocurre con el estimador de razón del total.

Teorema 7.2 (relación lineal).

Si la relación entre x e y es proporcional, el sesgo del estimador de la razón es $B(\widehat{R}) = 0$. Si la relación entre x e y es lineal, pero no proporcional (existe una constante no nula en la recta de regresión), el estimador de la razón puede ser sesgado.

ISBN 9788496866133

Demostración.

Supongamos relación proporcional entre x e y. Es decir, que $y_i = kx_i$ para toda unidad poblacional. Entonces $\widehat{R} = \dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}} = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \dfrac{k\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i} = k$ es constante y por lo tanto $cov(\widehat{R}, \widehat{\overline{x}}) = 0$, con lo que $B(\widehat{R}) = 0$.

La hipótesis del teorema es puramente teórica, pues si esa proporcionalidad se da, el estimador de la razón es perfecto: $\widehat{R} = k = R$ y también lo será el estimador de razón de la media $\overline{y}_R$. Pero el resultado es útil para mostrar que para reducir el sesgo del estimador interesa que x e y tengan una relación cercana a la proporcionalidad. Como se verá, también interesa para reducir la varianza del estimador. Obsérvese que el caso de proporcionalidad cumple como consecuencia que $\left|\rho_{xy}\right| = 1$.

Si la relación es lineal pero no proporcional, se tiene que $y_i = ax_i + b$ con $b \neq 0$ y entonces

$\widehat{R} = \dfrac{\sum_{i=1}^{n}(ax_i + b)}{\sum_{i=1}^{n} x_i} = a + \dfrac{b}{\widehat{\overline{x}}}$ y así,

$cov(\widehat{R}, \widehat{\overline{x}}) = cov(a + \dfrac{b}{\widehat{\overline{x}}}, \widehat{\overline{x}}) = b \cdot cov(\dfrac{1}{\widehat{\overline{x}}}, \widehat{\overline{x}})$ que puede ser distinto de cero, y por lo tanto $B(\widehat{R}) \neq 0$.

Teorema 7.3 (cota para el sesgo relativo).

Una cota para el sesgo relativo de $\widehat{R}$ es $\left|\dfrac{B(\widehat{R})}{\sigma_{\widehat{R}}}\right| \leq CV(\widehat{\overline{x}})$, donde $\sigma_{\widehat{R}}$ es la desviación típica de la razón muestral y $CV(\widehat{\overline{x}}) = \dfrac{\sqrt{V(\widehat{\overline{x}})}}{\overline{x}}$ es el coeficiente de variación de la media muestral de x.

Demostración.

Al ser $\left|\dfrac{B(\widehat{R})}{\sigma_{\widehat{R}}}\right| = \dfrac{\left|cov(\widehat{R}, \widehat{\overline{x}})\right|}{\sigma_{\widehat{R}}\overline{x}} = \dfrac{\left|\rho_{\widehat{R},\widehat{\overline{x}}}\right| \sigma_{\widehat{\overline{x}}}}{\overline{x}} \leq \dfrac{\sigma_{\widehat{\overline{x}}}}{\overline{x}} = CV(\widehat{\overline{x}})$, donde $\rho_{\widehat{R},\widehat{\overline{x}}}$ es el coeficiente de correlación entre $\widehat{R}$ y $\widehat{\overline{x}}$.

Como $CV(\widehat{\overline{x}}) = \dfrac{1}{\overline{x}}\sqrt{V(\widehat{\overline{x}})} = \dfrac{S_x}{\overline{x}}\sqrt{\dfrac{(1-f)}{n}}$, cuanto menor sea la variabilidad relativa de la variable auxiliar $\dfrac{S_x}{\overline{x}} = CV(x)$ y mayor el tamaño muestral n, más pequeña será la cota sobre el sesgo y como consecuencia el sesgo. A veces se dispone de gran información sobre x (por ejemplo todos sus valores poblacionales); entonces S_x y $\overline{x}$ son conocidas y se puede determinar el tamaño muestral n necesario para que el sesgo no exceda una cantidad. Usualmente se considera $\left|B(\widehat{R})\right|$ despreciable si es una fracción pequeña de su desviación típica, por ejemplo $CV(\widehat{\overline{x}})$ entre 0.1 y 0.2 (Kish, 1965) y entonces $\left|B(\widehat{R})\right| \leq 0.2\sigma_{\widehat{R}}$.

ISBN 9788496866133

Como consecuencia de los anteriores resultados, se observa que hay tres factores que pueden hacer reducir el sesgo:

a) Relación de proporcionalidad entre x e y.

b) Un bajo coeficiente de variación de la variable auxiliar x , $CV(x) = \dfrac{\sigma_x}{\overline{x}}$

c) Un tamaño muestral suficientemente alto .

Aunque no es necesario que los tres se cumplan para que el sesgo sea despreciable, son pautas a tener en cuenta.

Teorema 7.4 (aproximación del sesgo).

Una aproximación en primer orden al sesgo del estimador de la razón es

$$\widehat{B}(\widehat{R}) = \frac{-cov(\widehat{\overline{y}}, \widehat{\overline{x}}) + RV(\widehat{\overline{x}})}{\overline{x}^2}$$

Demostración.

Como $\widehat{R} - R = \dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}} - R = \dfrac{\widehat{\overline{y}} - R\widehat{\overline{x}}}{\widehat{\overline{x}}} = \dfrac{\widehat{\overline{y}} - R\widehat{\overline{x}}}{\overline{x}} \dfrac{\overline{x}}{\widehat{\overline{x}}}$

Se utilizará el desarrollo de Taylor de la función $\dfrac{1}{1+x}$, para aproximar el término $\dfrac{\overline{x}}{\widehat{\overline{x}}}$:

$$\frac{\overline{x}}{\widehat{\overline{x}}} = \frac{1}{1 + \left(\dfrac{\widehat{\overline{x}} - \overline{x}}{\overline{x}}\right)} = 1 - \frac{\widehat{\overline{x}} - \overline{x}}{\overline{x}} + \left(\frac{\widehat{\overline{x}} - \overline{x}}{\overline{x}}\right)^2 - ...$$

Así,

$$\widehat{R} - R = \frac{\widehat{\overline{y}} - R\widehat{\overline{x}}}{\overline{x}} \left[1 - \frac{\widehat{\overline{x}} - \overline{x}}{\overline{x}} + \left(\frac{\widehat{\overline{x}} - \overline{x}}{\overline{x}}\right)^2 - ...\right]$$

Tomando esperanzas y despreciando a partir del tercer término del desarrollo en serie:

$$B(\widehat{R}) = E(\widehat{R} - R) \simeq \frac{E(\widehat{\overline{y}} - R\widehat{\overline{x}})}{\overline{x}} - \frac{E\left[(\widehat{\overline{y}} - R\widehat{\overline{x}})(\widehat{\overline{x}} - \overline{x})\right]}{\overline{x}^2} =$$

$$= \frac{\overline{y} - \dfrac{\overline{y}}{\overline{x}}\overline{x}}{\overline{x}} - \frac{E\left[(\widehat{\overline{y}} - R\widehat{\overline{x}})(\widehat{\overline{x}} - \overline{x})\right]}{\overline{x}^2} =$$

$$= 0 - \frac{E\left[\widehat{\overline{y}}(\widehat{\overline{x}} - \overline{x})] - RE[\widehat{\overline{x}}(\widehat{\overline{x}} - \overline{x}))\right]}{\overline{x}^2} =$$

$$= -\frac{E(\widehat{\overline{y}}\widehat{\overline{x}}) - E(\widehat{\overline{y}})\overline{x} - R\left[E(\widehat{\overline{x}}^2) - \overline{x}E(\widehat{\overline{x}})\right]}{\overline{x}^2} =$$

ISBN 9788496866133

$$=-\frac{E(\widehat{\overline{y}\overline{x}})-E(\widehat{\overline{y}})E(\widehat{\overline{x}})-R\left[E(\widehat{\overline{x}}^2)-E(\widehat{\overline{x}})^2\right]}{\overline{x}^2}=$$

$$=\frac{-cov(\widehat{\overline{x}},\widehat{\overline{y}})+RV(\widehat{\overline{x}})}{\overline{x}^2}.$$

Ejemplo 7.4.

Se utilizarán los datos poblacionales de municipios de Girona, excluyendo la capital Girona por los motivos indicados en el Ejemplo 5.9. Supongamos que la variable auxiliar es el número de mujeres por municipio, y se desea estimar el número medio de hombres por municipio. Se considerará conocida la media poblacional de mujeres, $\overline{x} = 1074.76$. Se supone que el tamaño muestral a tomar es $n = 10$. Aunque se dispone de toda la población (y en este sentido el ejemplo es puramente teórico), para calcular el sesgo de $\widehat{R}$ habría que calcular todas las muestras posibles, equiprobables, que son $\binom{220}{10} = 5.9 \cdot 10^{16}$ y a partir de ellas, $-cov(\widehat{R},\widehat{\overline{x}})$. En lugar de ello procederemos a un estudio de simulación, obteniendo sucesivamente $K = 200$ muestras de tamaño n =10, calculando $\widehat{R}_i$, $\widehat{\overline{x}}_i$ y el estimador de razón de la media para cada una de ellas, y obteniendo también finalmente una aproximación a $-cov(\widehat{R},\widehat{\overline{x}})$ como $\frac{1}{K}\left(\sum \widehat{R}_i\widehat{\overline{x}}_i - K\overline{R}\overline{\overline{x}}\right)$, siendo $\overline{R}$ la razón media muestral y $\overline{\overline{x}}$ el promedio de las medias muestrales $\widehat{\overline{x}}_i$, y con ello una aproximación al sesgo del estimador de la media.

En las simulaciones se obtiene

$$\sum_{i=1}^{n} \widehat{R}_i\widehat{\overline{x}}_i = 212699.22 \ ,$$

$\overline{R} = 0.9936073$ y $\overline{\overline{x}} = 1070.39$,

con lo que el sesgo aproximado en la estimación de la media es $-cov(\widehat{R},\widehat{\overline{x}}) \simeq 2.69$, lo que es un sesgo muy pequeño pues la magnitud de la media a estimar es $\overline{y} = 1067.66$. La causa para que el estimador de razón tenga un sesgo tan pequeño, a pesar de un tamaño muestral $n = 10$ no muy alto, es la relación cercana a la proporcionalidad entre el número de varones por municipio (variable y) y el número de mujeres por municipio (variable x). El coeficiente de correlación poblacional entre x e y es de $\rho = 0.99933$. También se observa en la nube de puntos poblacional de la Figura 7.1 esta relación.

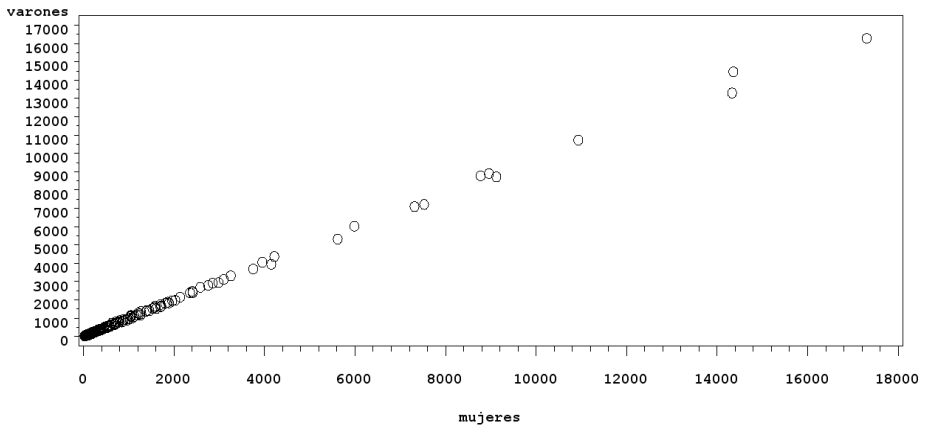
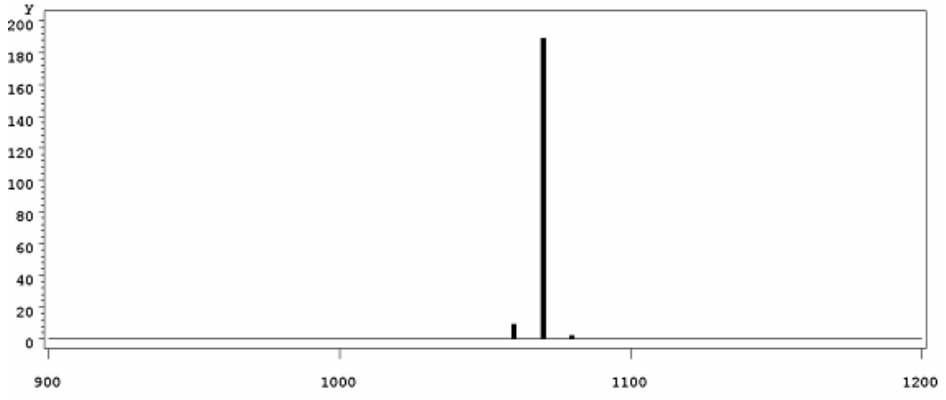


Figura 7.1. Relación entre varones y mujeres por municipio en la provincia de Girona.

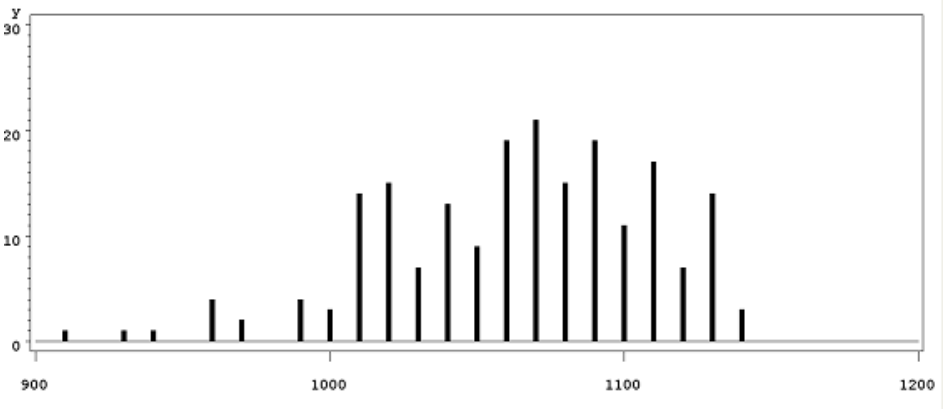
ISBN 9788496866133

La simulación realizada también nos permite estudiar el comportamiento del estimador de razón de la media comparado con el estimador usual de m.a.s., media muestral. La Figura 7.2 reproduce el histograma de los valores de los estimadores obtenidos en las 200 muestras.

Es evidente que la variabilidad del estimador de razón es mucho menor que la del estimador directo (que no tiene en cuenta la información de la variable auxiliar x). Además, como el sesgo es tan pequeño, está claro en este caso que la mejora introducida en la estimación por el uso de la variable auxiliar es muy grande.



Estimador de razón



Estimador directo

Figura 7.2. Comportamiento del estimador de razón y del estimador media muestral.

7.1.3 Varianza aproximada del estimador de la razón

Teorema 7.5 (varianza aproximada del estimador de la razón).

Una aproximación en primer orden a la varianza del estimador de la razón es

ISBN 9788496866133

$$V(\widehat{R}) \simeq \frac{1}{\overline{x}^2}\frac{N-n}{Nn}(S_y^2 + R^2 S_x^2 - 2RS_{xy}).$$

Demostración.

Utilizando el primer término del desarrollo en serie, $\widehat{R} - R \simeq \dfrac{\widehat{\overline{y}} - R\widehat{\overline{x}}}{\overline{x}}$. Es importante comentar que con esta aproximación el sesgo se está considerando nulo, pues $E\left(\dfrac{\widehat{\overline{y}} - R\widehat{\overline{x}}}{\overline{x}}\right) = 0$ como se vio en la demostración anterior. Ahora,

$$V(\widehat{R}) = V(\widehat{R} - R) \simeq \frac{V(\widehat{\overline{y}} - R\widehat{\overline{x}})}{\overline{x}^2} = \frac{1}{\overline{x}^2}\left[V(\widehat{\overline{y}}) + R^2 V(\widehat{\overline{x}}) - 2Rcov(\widehat{\overline{x}}, \widehat{\overline{y}})\right].$$

Se puede demostrar que, en m.a.s., $cov(\widehat{\overline{x}}, \widehat{\overline{y}}) = \dfrac{N-n}{N}\dfrac{S_{xy}}{n}$, donde S_{xy} es la cuasicovarianza poblacional. Las demás varianzas $V(\widehat{\overline{x}})$ y $V(\widehat{\overline{y}})$ son las usuales para m.a.s. Así,

$$V(\widehat{R}) \simeq \frac{1}{\overline{x}^2}\frac{N-n}{Nn}(S_y^2 + R^2 S_x^2 - 2RS_{xy}).$$

Aunque esta aproximación es la utilizada en la práctica, hay que tener en cuenta que aporta cierto error en la evaluación de la precisión del estimador y construcción de intervalos de confianza. Un criterio práctico para considerar válida esta aproximación es el mismo que el que se vio estudiando el sesgo, observar si

$$\widehat{CV}(\widehat{\overline{x}}) = \sqrt{\frac{(1-f)}{n}}\frac{s_x}{\widehat{\overline{x}}} \leq 0.2.$$

Corolario 7.1 (varianza aproximada del estimador de razón de la media).

Una aproximación a la varianza del estimador de razón de la media poblacional es

$$V(\overline{y}_R) \simeq \frac{N-n}{Nn}(S_y^2 + R^2 S_x^2 - 2RS_{xy}).$$

Propiedad 7.1 (estimador de la varianza del estimador de la razón).

Un estimador de los momentos de la varianza del estimador de la razón es

$$\widehat{V}(\widehat{R}) = \frac{1}{\overline{x}^2}\frac{N-n}{Nn}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}).$$

Propiedad 7.2 (estimador de la varianza del estimador).

Un estimador de los momentos de la varianza del estimador de la media es

$$\widehat{V}(\overline{y}_R) = \frac{N-n}{Nn}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}).$$

Demostración.

Se trata del estimador de los momentos de la aproximación anterior (sustituyendo los valores poblacionales S_y^2, R, S_x^2, S_{xy} por sus equivalentes muestrales respectivos). No es insesgado, pero se asume un sesgo pequeño en esta estimación en caso de ser la estimación de razón apropiada.

ISBN 9788496866133

Como la aproximación a la varianza utilizada asume sesgo aproximadamente cero del estimador de la razón y de la media, a la hora de calcular intervalos de confianza se utilizarán estos últimos estimadores de la varianza, junto con los estimadores correspondientes de razón, media o total, despreciando por lo tanto el posibles sesgo del estimador.

Se verá a continuación un resultado útil desde el punto de vista del cálculo manual o informático del estimador de la varianza.

Teorema 7.6 (expresiones alternativas para las varianzas).

$$\text{(a1)} V(\overline{y}_R) \simeq \frac{N-n}{Nn}\frac{1}{N-1}(\sum_{i=1}^{N} y_i^2 + R^2\sum_{i=1}^{N} x_i^2 - 2R\sum_{i=1}^{N} x_i y_i)$$

$$\text{(a2)} \widehat{V}(\overline{y}_R) = \frac{N-n}{Nn}\frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 + \widehat{R}^2\sum_{i=1}^{n} x_i^2 - 2\widehat{R}\sum_{i=1}^{n} x_i y_i)$$

$$\text{(b1)} V(\overline{y}_R) \simeq \frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}(y_i - Rx_i)^2$$

$$\text{(b2)}\ \widehat{V}(\overline{y}_R) = \frac{N-n}{Nn}\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{R}x_i)^2$$

Demostración.

Se demostrarán (a2) y (b2), pues la demostración de (a1) y (b1) es idéntica.

(a2)

$$\frac{N-n}{Nn}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}) =$$

$$= \frac{N-n}{Nn}\frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 - n\widehat{\overline{y}} + \widehat{R}^2\sum_{i=1}^{n} x_i^2 - n\widehat{R}^2\widehat{\overline{x}} - 2\widehat{R}\sum_{i=1}^{n} x_i y_i + 2\widehat{R}n\widehat{\overline{x}\overline{y}}) =$$

$$= \frac{N-n}{Nn}\frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 - n\widehat{\overline{y}} + \widehat{R}^2\sum_{i=1}^{n} x_i^2 - n\frac{\widehat{\overline{y}}}{\widehat{\overline{x}}}\widehat{\overline{x}} - 2\widehat{R}\sum_{i=1}^{n} x_i y_i + 2\frac{\widehat{\overline{y}}}{\widehat{\overline{x}}}n\widehat{\overline{x}\overline{y}}) =$$

$$= \frac{N-n}{Nn}\frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 + \widehat{R}^2\sum_{i=1}^{n} x_i^2 - 2\widehat{R}\sum_{i=1}^{n} x_i y_i).$$

(b2)

Por el resultado obtenido en (a2), tenemos que

$$\widehat{V}(\overline{y}_R) = \frac{N-n}{Nn}\frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 + \widehat{R}^2\sum_{i=1}^{n} x_i^2 - 2\widehat{R}\sum_{i=1}^{n} x_i y_i) =$$

$$= \frac{N-n}{Nn}\frac{1}{n-1}\sum_{i=1}^{n}(y_i^2 + \widehat{R}^2 x_i^2 - 2\widehat{R}x_i y_i) = \frac{N-n}{Nn}\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{R}x_i)^2.$$

ISBN 9788496866133

7.1.4 Comparación de la estimación de razón con la estimación directa bajo m.a.s.

Seguidamente veremos las condiciones en las cuales la estimación de razón es más eficiente que la estimación directa.

Teorema 7.7 (comparación entre estimación de razón y directa).

Suponiendo adecuada la aproximación a la varianza del estimador de razón, y asumiendo m.a.s., el estimador de razón de la media $\overline{y}_R$ tiene menor varianza que el estimador usual de m.a.s. basado en la media muestral $\widehat{\overline{y}}$, si

$$\rho_{xy} > \frac{1}{2}\frac{CV(x)}{CV(y)}.$$

Demostración.

$$V(\overline{y}_R) \simeq \frac{N-n}{Nn}(S_y^2 + R^2 S_x^2 - 2RS_{xy}) < V(\widehat{\overline{y}}) \Leftrightarrow$$

$$\Leftrightarrow \frac{N-n}{Nn}S_y^2 + R^2\frac{N-n}{Nn}S_x^2 - 2R\frac{N-n}{Nn}S_{xy} < \frac{N-n}{Nn}S_y^2$$

$$\Leftrightarrow R^2\frac{N-n}{Nn}S_x^2 - 2R\frac{N-n}{Nn}S_{xy} < 0 \Leftrightarrow RS_x^2 - 2\rho_{xy}S_xS_y < 0 \Leftrightarrow$$

$$\Leftrightarrow \frac{\overline{y}}{\overline{x}}S_x - 2\rho_{xy}S_y < 0 \Leftrightarrow \rho_{xy} > \frac{1}{2}\frac{S_x/\overline{x}}{S_y/\overline{y}} = \frac{1}{2}\frac{CV(x)}{CV(y)}.$$

Estos resultados traen las siguientes consecuencias:

a) Si las variabilidades relativas de x e y son similares, basta que $\rho_{xy} > 0.5$ para que el estimador de razón sea más eficiente que el estimador directo de m.a.s., $\widehat{\overline{y}}$.

b) La aproximación a la varianza considerada asume sesgo cero en la estimación de la razón (y por lo tanto de la media), con lo cual la comparación aproximada se extiende al error cuadrático medio de los dos estimadores, es decir, $MSE(\overline{y}_R) < MSE(\widehat{\overline{y}})$ si $\rho_{xy} > \frac{1}{2}\frac{CV(x)}{CV(y)}$.

c) El coeficiente de variación de x aparece una vez más como un factor de mejora de la estimación de razón: no sólo al disminuir este coeficiente se reduce el sesgo del estimador, sino también su varianza.

d) Si ρ_{xy} es alto, la estimación de razón $\overline{y}_R$ suele mejorar a la estimación $\widehat{\overline{y}}$. La diferencia entre las dos estimaciones es mayor cuanto mayor sea ρ_{xy}, pues

$V(\widehat{\overline{y}}) - V(\overline{y}_R) \propto 2\rho_{xy}S_y - \frac{\overline{y}}{\overline{x}}S_x$, aumenta con ρ_{xy}.

e) En la práctica se desconocen ρ_{xy} y $CV(y)$, y en general también $CV(x)$. Por lo tanto se realizan las aproximaciones de los momentos r_{xy}, $\widehat{CV}(x)$ y $\widehat{CV}(y)$ si se desea comprobar aproximadamente lo adecuado de la estimación de razón.

ISBN 9788496866133

Ejemplo 7.5.

Los datos de la Tabla 7.1 son una muestra de n =10 provincias españolas, del total de $N = 53$, a las que se midió el número de nacimientos en noviembre de 1997 (variable x) y diciembre de 1997 (variable y). Se desea estimar el total de nacimientos en Diciembre de 1997 y la media de nacimientos por provincia en Diciembre de 1997, sabiendo que la media poblacional en Noviembre fue de $\overline{x} = 536.09$.

Noviembre	1104	74	106	125	603	960	152	2947	97	465
Diciembre	759	69	122	154	644	1065	161	4090	109	457

Tabla 7.3. Nacimientos en una muestra de provincias españolas (1997).

En la muestra, $\widehat{\overline{x}} = 663.30$ y $\widehat{\overline{y}} = 763.0$, con lo que $\widehat{R} = \dfrac{763.0}{663.3} = 1.15$. El estimador de razón de la media en Diciembre será $\overline{y}_R = \widehat{R}\overline{x} = 1.15 \cdot 536.09. = 616.50$ y el estimador de razón del total de nacimientos en Diciembre será $N\overline{y}_R = 53 \cdot 616.50 = 32674.68$.

Para la estimación de varianzas, tenemos que $s_x = 886.33$ y $s_y = 1216.62$. Además $s_{xy} = 1057881.5$ y $r_{xy} = 0.98154$. Así,

$\widehat{V}(\overline{y}_R) = \dfrac{N-n}{Nn}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}) = \dfrac{53-10}{53 \cdot 10}((1216.62)^2 + (1.15 \cdot 886.33)^2 - 2 \cdot 1.15 \cdot 1057881.5). =$ 6974.7 y la desviación típica del estimador $\sqrt{\widehat{V}(\overline{y}_R)} = 83.5$.

Para el total, la desviación típica del estimador es $\sqrt{\widehat{V}(N\overline{y}_R)} = 4426.27$.

Si no utilizamos la variable auxiliar x ="nacimientos en Noviembre", el estimador de la media en Diciembre será el usual de m.a.s. , $\widehat{\overline{y}} = 763.0$ con una desviación típica aproximada de

$\sqrt{\widehat{V}(\widehat{\overline{y}})} = \sqrt{\dfrac{N-n}{Nn}s_y^2} = 346.53$, mucho mayor que 83.5.

Si estimamos los coeficientes de variación , $\widehat{CV(x)} = s_x/\widehat{\overline{x}} = 1.336$ y $\widehat{CV(y)} = s_y/\widehat{\overline{y}} = 1.59$, con lo que $\dfrac{1}{2}\dfrac{CV(x)}{CV(y)} \simeq 0.42$. Al ser $r_{xy} = 0.98154$, esta claro que la estimación de razón mejora al m.a.s., como ya se intuía al estimar las varianzas de ambos estimadores.

La Figura 7.3, nube de puntos de la muestra, pone de relieve la relación de proporcionalidad entre x e y.

ISBN 9788496866133

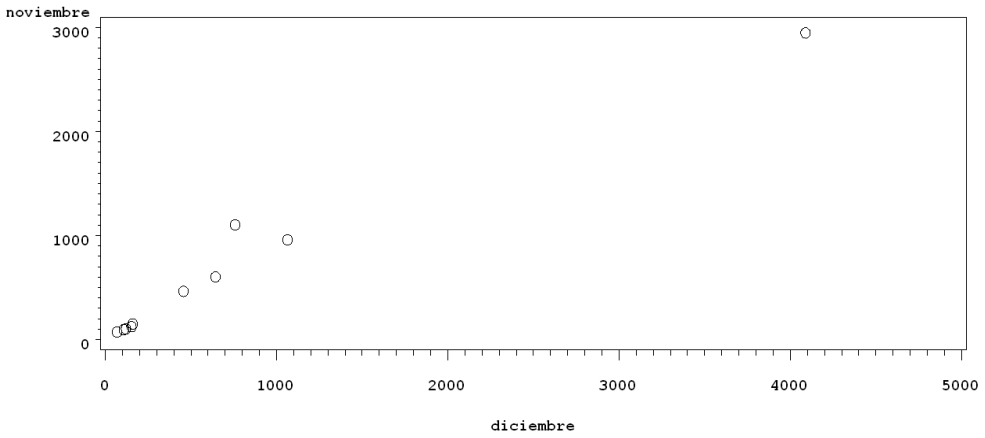


Figura 7.3. Nacimientos en Noviembre frente a nacimientos en Diciembre.

7.1.5 Estimadores de razón en muestreo estratificado

Supondremos en este apartado que la población está particionada en L estratos, y se realiza muestreo aleatorio simple sin reemplazamiento de tamaño n_h en cada estrato, obteniendo las observaciones (y_{1_h}, x_{1_h}), ... , (y_{n_h}, x_{n_h}), para cada uno de los estratos $h = 1, ..., L$. Entonces se puede abordar el problema de la construcción de un estimador global de dos maneras: el esimador separado de razón y el estimador combinado de razón.

(i) **Estimador separado de razón**

Para construir el **estimador separado de razón** de la media poblacional $\overline{y}$ se exige conocer la media poblacional de x por estrato , $\overline{x}_h$. Es obvio que esto implica conocer también la media poblacional de x, que será

$$\overline{x} = \sum_{h=1}^{L} \frac{N_h}{N} \overline{x}_h,$$

pues

$$\sum_{h=1}^{L} \frac{N_h}{N} \overline{x}_h = \sum_{h=1}^{L} \frac{N_h}{N} \frac{1}{N_h} \sum_{i=1}^{N_h} x_{i_h} = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} x_{i_h} = \overline{x}.$$

El estimador separado de la media poblacional se construye como es habitual en muestreo estratificado: ponderando los estimadores de las medias por estrato. Por tanto este estimador no supone que la verdadera razón poblacional permanezca constante al pasar de un estrato a otro.

Definición.

El estimador **separado de razón** de la media poblacional es

$$t_s = \sum_{h=1}^{L} \frac{N_h}{N} \overline{y}_{R_h} = \sum_{h=1}^{L} \frac{N_h}{N} \widehat{R}_h \overline{x}_h = \sum_{h=1}^{L} \frac{N_h}{N} \frac{\widehat{\overline{y}}_h}{\widehat{\overline{x}}_h} \overline{x}_h$$

ISBN 9788496866133

donde $\widehat{R}_h = \dfrac{\widehat{\overline{y}}_h}{\widehat{\overline{x}}_h}$ es el estimador de la razón en el estrato h, y $\widehat{\overline{x}}_h$ e $\widehat{\overline{y}}_h$ son las medias muestrales obtenidas en el estrato h.

Propiedad 7.3 (sesgo del estimador separado)

El sesgo del estimador separado de razón de la media es

$$B(t_s) = -\sum_{h=1}^{L} \frac{N_h}{N} cov(\widehat{R}_h, \widehat{\overline{x}}_h)$$

Demostración

$$E(t_s) - \overline{y} = \sum_{h=1}^{L} \frac{N_h}{N} E(\overline{y}_{R_h}) - \overline{y} = \sum_{h=1}^{L} \frac{N_h}{N} E(\overline{y}_{R_h}) - \sum_{h=1}^{L} \frac{N_h}{N} \overline{y}_h =$$

$$\sum_{h=1}^{L} \frac{N_h}{N} \left[E(\overline{y}_{R_h}) - \overline{y}_h\right] = \sum_{h=1}^{L} \frac{N_h}{N} B(\overline{y}_{R_h}) = -\sum_{h=1}^{L} \frac{N_h}{N} cov(\widehat{R}_h, \widehat{\overline{x}}_h).$$

Teorema 7.7 (varianza del estimador separado).

Una aproximación a la varianza del estimador separado de razón de la media es

$$V(t_s) \simeq \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} (S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h S_{xyh})$$

Demostración .

$$V(t_s) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 V(\overline{y}_{R_h}) \simeq \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} (S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h S_{xyh})$$

por ser independiente el muestreo en los diferentes estratos.

Corolario 7.3 (estimador de la varianza del estimador separado).

Un estimador de la varianza del estimador separado de razón de la media es

$$\widehat{V}(t_s) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} (s_{yh}^2 + \widehat{R}_h^2 s_{xh}^2 - 2\widehat{R}_h s_{xyh}).$$

Es una vez más el estimador de los momentos sobre la aproximación a la varianza. En ocasiones (véase el primer ejemplo de este capítulo) interesa por sí solo un estimador de la razón poblacional R. La estimación separada lleva a construir un estimador de este tipo. Recordemos que $\overline{y}_R = \widehat{R}\overline{x}$, con lo que $\widehat{R}_s = \dfrac{\overline{y}_R}{\overline{x}}$.

Definición.

Se denomina **estimador separado de la razón** poblacional R a $\widehat{R}_s = \dfrac{t_s}{\overline{x}}$. Este estimador tiene varianza $V(\widehat{R}_s) = \dfrac{1}{\overline{x}^2} V(t_s)$.

(ii) **Estimador combinado de razón**

Para construir el **estimador combinado de razón** de la media poblacional $\overline{y}$ se exige conocer la media poblacional de la variable auxiliar x, $\overline{x}$. El estimador combinado se construye

ISBN 9788496866133

estimando la media poblacional de $\overline{y}$ y de $\overline{x}$ como usualmente en m.a.s. estratificado, y posteriormente creando un estimador del tipo de razón a partir de esas dos estimaciones.

Definición .

El estimador **combinado de razón** de la media $\overline{y}$ es

$$t_c = \frac{\sum_{h=1}^{L} \frac{N_h}{N}\widehat{\overline{y}}_h}{\sum_{h=1}^{L} \frac{N_h}{N}\widehat{\overline{x}}_h}\overline{x} = \frac{\sum_{h=1}^{L} W_h\widehat{\overline{y}}_h}{\sum_{h=1}^{L} W_h\widehat{\overline{x}}_h}\overline{x} = \frac{\overline{y}_{st}}{\overline{x}_{st}}\overline{x}.$$

El estimador **combinado de la razón** R es $\widehat{R}_c = \dfrac{\overline{y}_{st}}{\overline{x}_{st}}$, con lo que $t_c = \widehat{R}_c\overline{x}$.

Teorema 7.8 (varianza del estimador combinado).

Una aproximación a la varianza del estimador combinado de razón de la media es

$$V(t_c) \simeq \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h}(S_{yh}^2 + R^2 S_{xh}^2 - 2RS_{xyh}).$$

Además, una aproximación a la varianza de $\widehat{R}_c$ es

$$V(\widehat{R}_c) \simeq \frac{1}{\overline{x}^2}V(t_c).$$

Demostración .

Utilizando la aproximación $\dfrac{\overline{x}}{\overline{x}_{st}} \simeq 1$, se tiene que

$$\widehat{R}_c - R = \frac{\overline{y}_{st}}{\overline{x}_{st}} - R = \frac{\overline{y}_{st} - R\overline{x}_{st}}{\overline{x}}\frac{\overline{x}}{\overline{x}_{st}} \simeq \frac{\overline{y}_{st} - R\overline{x}_{st}}{\overline{x}}$$

con lo que

$$V(\widehat{R}_c) = V(\widehat{R}_c - R) \simeq \frac{1}{\overline{x}^2}V(\overline{y}_{st} - R\overline{x}_{st}) = \frac{1}{\overline{x}^2}\left(V(\overline{y}_{st}) + R^2V(\overline{x}_{st}) - 2Rcov(\overline{x}_{st}, \overline{y}_{st})\right) =$$

$$= \frac{1}{\overline{x}^2}\sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h}(S_{yh}^2 + R^2 S_{xh}^2 - 2RS_{xyh}),$$

pues $cov(\overline{x}_{st}, \overline{y}_{st}) = cov(\sum_{h=1}^{L} W_h\widehat{\overline{x}}_h, \sum_{h=1}^{L} W_h\widehat{\overline{y}}_h) =$

$$= \sum_{h=1}^{L} cov(W_h\widehat{\overline{x}}_h, W_h\widehat{\overline{y}}_h) + \underbrace{\sum_{h\neq h'}^{L} cov(W_h\widehat{\overline{x}}_h, W_{h'}\widehat{\overline{y}}_{h'})}_{0} =$$

$$= \sum_{h=1}^{L} W_h^2 cov(\widehat{\overline{x}}_h, \widehat{\overline{y}}_h) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h}S_{xyh}.$$

El término indicado es 0 como consecuencia de la independencia del muestreo entre los estratos diferentes h y h'.

ISBN 9788496866133

Corolario 7.4 (estimador de la varianza del estimador combinado) .

Un estimador de la varianza del estimador de razón de la media poblacional es

$$\widehat{V}(t_c) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h}(s_{yh}^2 + \widehat{R}_c^2 s_{xh}^2 - 2\widehat{R}_c s_{xyh}).$$

Además, un estimador de la varianza de $\widehat{R}_c$ es

$$\widehat{V}(\widehat{R}_c) = \frac{1}{\overline{x}^2}\widehat{V}(t_c).$$

7.1.6 Comparación entre el estimador de razón separado y el combinado

En cuanto a la información disponible :

- Para construir el estimador de razón separado se necesitan conocer las medias de la variable auxiliar por estrato, $\overline{x}_h$. El estimador de razón combinado sólo necesita conocer la media poblacional $\overline{x}$.
- En compensación, el estimador de razón separado permite estimaciones de razón globales pero también separadas por estratos, mientras que el estimador combinado sólo genera estimaciones globales.

En cuanto a la precisión y sesgo en la estimación:

- Si las razones en los estratos R_h son muy similares, la estimación global separada será muy similar a la combinada.

En términos prácticos, si se dispone de información por estrato $\overline{x}_h$ y existen diferencias en términos de la razón R_h en los estratos, lo cual suele ser el caso más general, se prefiere estimación separada. En estos casos suele ocurrir que la varianza de este estimador es bastante inferior a la varianza del estimador combinado. El problema de n_h pequeños no sólo afecta a la estimación separada, también redunda en una mala estimación combinada , por lo que la comparación, si los R_h son variables, puede ser también favorable a la estimación separada.

Es decir, por lo general, solamente la falta de la información sobre las medias $\overline{x}_h$ y/o la escasa diferencia entre razones en los estratos R_h justificaría la estimación combinada frente a la separada.

Ejemplo 7.6.

Utilizando los datos de la provincia de Girona mencionados en el ejemplo 7.4, supon-gamos que se excluye la capital y se estratifica la población en municipios mayores de 1.000 habitantes y menores de 1.000 habitantes Hay entonces $N_1 = 143$ municipios en el estrato 1 ($\geq$1000 habitantes) y $N_2 = 77$ municipios en el estrato 2 ($<$1000 habitantes). Supongamos que se desea una muestra final de tamaño $n = 20$.

ISBN 9788496866133

Si no disponemos de mayor información previa (es decir, las varianzas por estrato), una opción es utilizar afijación proporcional. En este caso queda $n_1 = 20\frac{143}{220} = 13$ y $n_2 = 7$.

Se supone que se dispone de la media poblacional por estrato de la variable auxiliar número de mujeres por municipio. Estas medias son $\overline{x}_1 = 2725.86$ y $\overline{x}_1 = 185.71$.

Se procede a extraer las muestras aleatorias simples dentro de cada estrato, obteniendo los siguientes datos de las variable número de hombres y número de mujeres en los municipios muestreados:

Total hombres	1660	996	1387	1888	682	8787
Total mujeres	1723	1023	1390	1864	710	8777

Total hombres	1407	1186	3954	2925	2955	4375	8913
Total mujeres	1458	1264	4153	2860	2987	4219	8962

Tabla 7.4. Estrato 1 (municipios con más de 1000 habitantes)

Total hombres	155	63	123	260	146	66	351
Total mujeres	160	60	124	220	124	68	360

Tabla 7.5. Estrato 2 (municipios con menos de 1000 habitantes)

De las muestras obtenidas, se extraen los siguientes estadísticos:

	N_h	n_h	$\widehat{\overline{y}}_h$	$\widehat{\overline{x}}_h$	s_{xh}	s_{yh}	s_{xyh}	$\widehat{R}_h$
Estrato 1	143	13	3162.69	3183.85	2759.41	2765.39	7627487.24	0.9933
Estrato 2	77	7	166.28	159.42	103.8	104.91	10734	1.043

Tabla 7.6. Estadísticos muestrales por estrato

Con lo cual el estimador de razón de la media $\overline{y}_1$ en el estrato 1 es $\widehat{R}_1\overline{x}_1 = 2707.6$ y en el estrato 2 será $\widehat{R}_2\overline{x}_2 = 166.27$.

El estimador de razón separado de la media es:

$$t_s = \sum_{h=1}^{L} \frac{N_h}{N}\overline{y}_{R_h} = \frac{143}{220}2707.6 + \frac{77}{220}185.71 = 1824.93.$$

Con varianza estimada:

$$\widehat{V}(t_s) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h}(s_{yh}^2 + \widehat{R}_h^2 s_{xh}^2 - 2\widehat{R}_h s_{xyh}) =$$

$$= \left(\frac{143}{220}\right)^2 \frac{(143-13)}{143\cdot 13}(2765.39^2 + 0.9933^2 2759.41^2 - 2\cdot 0.9933\cdot 7627487.24) +$$

ISBN 9788496866133

$$+\left(\frac{77}{220}\right)^2 \frac{(77-7)}{77\cdot 7}(104.91^2 + 1.043^2 103.8^2 - 2\cdot 1.043\cdot 10734) =$$

$$= 214.76 + 5.34 = 220.1$$

con desviación típica del estimador $\sqrt{\widehat{V}(t_s)} = 14.8$. Se observa también que la contribución a la varianza del estimador del primer estrato es muy superior a la del segundo, por lo cual se intuye que en este caso, una mejor afijación sería la de varianza mínima, aumentando el tamaño muestral destinado al primer estrato y disminuyéndolo en el segundo estrato.

Un intervalo de confianza al 95% para la media poblacional del número de hombres por municipio basado en el estimador separado será entonces $(1824.93 \pm 1.96\cdot 14.8) = (1795.9, 1853.93)$.

En cuanto a la estimación combinada, se tiene que $\overline{y}_{st}$ =2113.94 y $\overline{x}_{st} = 2125.3$ por lo cual $\widehat{R}_c = \dfrac{2113.94}{2125.3} = 0.99465$. El valor exacto de $\overline{x}$ se puede calcular a partir de las medias poblacionales por estrato:

$$\overline{x} = W_1\overline{x}_1 + W_2\overline{x}_2 = 0.65\cdot 2725.86 + 0.35\cdot 185.71 = 2046.80.$$

De este modo,

$$t_c = \widehat{R}_c\overline{x} = 0.99465\cdot 2046.80 = 2035.85.$$

La varianza estimada será

$$\widehat{V}(t_c) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h}(s_{yh}^2 + \widehat{R}_c^2 s_{xh}^2 - 2\widehat{R}_c s_{xyh}) =$$

$$= \left(\frac{143}{220}\right)^2 \frac{(143-13)}{143\cdot 13}(2765.39^2 + 0.9946^2 2759.41^2 - 2\cdot 0.9946\cdot 7627487.24) +$$

$$+\left(\frac{77}{220}\right)^2 \frac{(77-7)}{77\cdot 7}(104.91^2 + 0.9946^2 103.8^2 - 2\cdot 0.9946\cdot 10734) =$$

$$= 210.2 + 4.937 = 215.13.$$

Aunque las diferencias son pequeñas, la varianza estimada del estimador combinado es algo menor que la del separado. Esto se da probablemente por ser las razones en los dos estratos muy similares. Se puede utilizar, en este caso, la estimación separada exclusivamente para dar estimaciones por estrato, y la combinada para la estimación conjunta.

7.2 Estimadores de regresión

7.2.1 Introducción

La estimación de razón supone una relación de proporcionalidad directa entre la variable de interés y y la variable auxiliar x. Es decir, $y_i = Rx_i$. Ocurre que a menudo que la relación entre las dos variables, aún siendo lineal, no pasa por el origen. En este caso se puede poner $y_i = a + bx_i$ con $a, b \neq 0$.

ISBN 9788496866133

En muestreo aleatorio simple , si esa relación se cumple, se obtiene que $y_i = a + bx_i$ para $i = 1, ..., n \Rightarrow \widehat{\overline{y}} = a + b\widehat{\overline{x}}$. Como además $\overline{y} = a + b\overline{x}$, restando ambas ecuaciones se obtiene $\overline{y} - \widehat{\overline{y}} = b(\overline{x} - \widehat{\overline{x}})$ con lo que $\overline{y} = \widehat{\overline{y}} + b(\overline{x} - \widehat{\overline{x}})$.

Definición .

Supongamos que se dispone de $\overline{x}$, la media poblacional de x, y se obtiene una m.a.s. de n pares (x_i, y_i). El **estimador de regresión de la media** poblacional $\overline{y}$ es

$\overline{y}_{reg} = \widehat{\overline{y}} + b(\overline{x} - \widehat{\overline{x}})$,

donde b es una constante que puede ser conocida de anteriores estudios, pero en general se estima por $\widehat{b}$, como se verá a continuación.

7.2.2 Sesgo y varianza del estimador

Teorema 7.9 (sesgo del estimador de regresión).

El sesgo del estimador de regresión es $B(\overline{y}_{reg}) = -cov(b, \widehat{\overline{x}})$. Si b es una constante supuestamente conocida, el sesgo es $B(\overline{y}_{reg}) = 0$. En general $\widehat{b}$ es una estimación a partir de la muestra, con lo que $B(\overline{y}_{reg}) = -cov(\widehat{b}, \widehat{\overline{x}})$.

Demostración .

$$B(\overline{y}_{reg}) = E(\overline{y}_{reg}) - \overline{y} = E(\widehat{\overline{y}}) + E[b(\overline{x} - \widehat{\overline{x}})] - \overline{y} =$$

$$= E(b\overline{x}) - E(b\widehat{\overline{x}}) = \overline{x}E(b) - E(b\widehat{\overline{x}}) = E(\widehat{\overline{x}})E(b) - E(b\widehat{\overline{x}}) = -cov(b, \widehat{\overline{x}}).$$

Teorema 7.10 (varianza del estimador de regresión).

Si b es una constante conocida, la varianza del estimador de regresión bajo m.a.s. es

$$V(\overline{y}_{reg}) = \frac{N-n}{Nn}(S_y^2 + b^2 S_x^2 - 2bS_{xy}).$$

Demostración .

$$V(\overline{y}_{reg}) = V(\widehat{\overline{y}}) + V[b(\overline{x} - \widehat{\overline{x}})] - 2cov(\widehat{\overline{y}}, b(\overline{x} - \widehat{\overline{x}})) =$$

$$= V(\widehat{\overline{y}}) + b^2 V[(\overline{x} - \widehat{\overline{x}})] - 2bcov(\widehat{\overline{y}}, (\overline{x} - \widehat{\overline{x}})) =$$

$$= V(\widehat{\overline{y}}) + b^2 V(\widehat{\overline{x}}) - 2bcov(\widehat{\overline{y}}, \widehat{\overline{x}}), \text{ al ser } \overline{x} \text{ y } b \text{ constantes.}$$

Así, $V(\overline{y}_{reg}) = \frac{N-n}{Nn}(S_y^2 + b^2 S_x^2 - 2bS_{xy})$.

Usualmente no disponemos del conocimiento de b, con lo que se hace necesario estimarlo. Una posibilidad es tomar el valor de b que haga mínima la varianza del estimador.

Teorema 7.11 (valor de b que minimiza la varianza del estimador de regresión).

El valor de b que hace mínima la expresión $V(\overline{y}_{reg})$ es

ISBN 9788496866133

$b^* = \dfrac{S_{xy}}{S_x^2}$.

Demostración .

Haciendo

$f(b) = \dfrac{N-n}{Nn}(S_y^2 + b^2 S_x^2 - 2bS_{xy}) \Rightarrow f'(b) = \dfrac{N-n}{Nn}(2bS_x^2 - 2S_{xy}) = 0 \Leftrightarrow b = \dfrac{S_{xy}}{S_x^2}$ y además como $f''(b) > 0$ se cumple que $b^* = \dfrac{S_{xy}}{S_x^2}$ es un mínimo.

Corolario 7.5 (estimador de regresión con b estimado).

(i) El estimador de los momentos de b^* es $\widehat{b} = \dfrac{s_{xy}}{s_x^2}$. Así, el estimador de regresión con b estimado es

$\overline{y}_{reg} = \widehat{\overline{y}} + \widehat{b}(\overline{x} - \widehat{\overline{x}})$.

(ii) El estimador de los momentos de la constante a en la recta de regresión $y = a + bx$ es $\widehat{a} = \widehat{\overline{y}} - \widehat{b}\widehat{\overline{x}}$.

Corolario 7.6 (varianza mínima del estimador de regresión).

El valor de $V(\overline{y}_{reg})$ en el caso de $b^* = \dfrac{S_{xy}}{S_x^2}$ es

$V(\overline{y}_{reg}) = \dfrac{N-n}{Nn} S_y^2(1-\rho_{xy}^2)$.

Demostración.

Sustituyendo $b^* = \dfrac{S_{xy}}{S_x^2}$ en la expresión de la varianza, queda

$$V(\overline{y}_{reg}) = \frac{N-n}{Nn}(S_y^2 + b^2 S_x^2 - 2bS_{xy}) = \frac{N-n}{Nn}(S_y^2 + (\frac{S_{xy}}{S_x^2})^2 S_x^2 - 2\frac{S_{xy}}{S_x^2} S_{xy}) =$$

$$= \frac{N-n}{Nn}(S_y^2 - \frac{S_{xy}^2}{S_x^2}) = \frac{N-n}{Nn} S_y^2(1-\rho_{xy}^2).$$

Se ve claramente que cuanto mayor sea el coeficiente de correlación entre x e y, menor será la varianza del estimador.

Corolario 7.7 (estimador de la varianza del estimador de regresión).

El estimador de los momentos de la varianza $V(\overline{y}_{reg})$ es

$\widehat{V}(\overline{y}_{reg}) = \dfrac{N-n}{Nn}(s_y^2 + \widehat{b}^2 s_x^2 - 2\widehat{b}s_{xy}) = \dfrac{N-n}{Nn} s_y^2(1-r_{xy}^2)$,

donde $\widehat{b} = \dfrac{s_{xy}}{s_x^2}$ y r_{xy} es el coeficiente de correlación muestral entre x e y, $r_{xy} = \dfrac{s_{xy}}{s_x s_y}$.

ISBN 9788496866133

7.2.3 Comparaciones con estimación directa y estimación de razón bajo m.a.s.

Se realizan a continuación comparaciones con los métodos de estimación media muestral y estimador de razón. Se asumirá que el tamaño muestral es suficientemente grande para que las aproximaciones a las varianzas utilizadas en este capítulo sean válidas.

Teorema 7.12 (Comparaciones entre m.a.s., razón y regresión).

Bajo m.a.s.,

a) $V(\widehat{\overline{y}}) \geq V(\overline{y}_{reg})$, siendo iguales estas varianzas si y sólo si $\rho_{xy} = 0$.

b) $V(\overline{y}_R) \geq V(\overline{y}_{reg})$, siendo iguales estas varianzas si y sólo si $R = \dfrac{S_{xy}}{S_x^2} = b^*$. Esta igualdad es equivalente a que la recta de regresión entre x e y , calculada con b^*, pase por el origen.

Demostración .

En m.a.s.,

$$V(\widehat{\overline{y}}) = \frac{N-n}{N}\frac{S_y^2}{n}.$$

Con el estimador de razón ,

$$V(\overline{y}_R) \simeq \frac{N-n}{Nn}(S_y^2 + R^2 S_x^2 - 2RS_{xy}).$$

Con el estimador de regresión utilizando el b de varianza mínima,

$$V(\overline{y}_{reg}) = \frac{N-n}{Nn}S_y^2(1-\rho_{xy}^2).$$

(a) Es directo, pues como $\rho_{xy}^2 \leq 1$, $\dfrac{V(\widehat{\overline{y}})}{V(\overline{y}_{reg})} = \dfrac{1}{(1-\rho_{xy}^2)} \geq 1$, con igualdad $\Leftrightarrow \rho_{xy} = 0$.

(b) Sea ahora

$$\begin{aligned} V(\overline{y}_R) - V(\overline{y}_{reg}) &= \frac{N-n}{Nn}(R^2 S_x^2 - 2RS_{xy} + \rho_{xy}^2 S_y^2) = \\ &= \frac{N-n}{Nn}(R^2 S_x^2 - 2R\rho_{xy}S_x S_y + \rho_{xy}^2 S_y^2) = \\ &= \frac{N-n}{Nn}(RS_x - \rho_{xy}S_y)^2 \geq 0. \end{aligned}$$

Además la igualdad solo se da cuando $R = \rho_{xy}\dfrac{S_y}{S_x} = \dfrac{S_{xy}}{S_x^2} = b^*$.

En este último caso, a partir de la ecuación planteada de la recta de regresión se tiene que $\overline{y} = a + b^*\overline{x}$, entonces $a = \overline{y} - b^*\overline{x} = \overline{y} - R\overline{x} = 0$ (la recta pasa por el origen).

Es necesario hacer algunas puntualizaciones:

ISBN 9788496866133

- El estimador de regresión, cuando se realiza m.a.s. en la población, tiene una varianza menor que el estimador media muestral o que el estimador de razón. Esto indica que en general, disponer de una variable auxiliar x suele llevar a una mayor eficiencia en la estimación de las características poblacionales de y.

- La estimación de regresión mejora a la estimación de razón y a la media muestral teniendo en cuenta el b^* óptimo, que es desconocido al depender de la cuasicovarianza poblacional S_{xy}. Este b se estima por $\widehat{b}$, con $\widehat{b} = \dfrac{s_{xy}}{s_x^2}$. Hay que remarcar que este estimador de b^* es sesgado, aunque si se dispone de S_x, se puede demostrar que $\widehat{b}' = \dfrac{s_{xy}}{S_x^2}$ es insesgado. Así, la supuesta mejora inducida por la estimación de regresión está sujeta a la estimación del parámetro b. Si la relación lineal entre x e y no se da, b estará cercana a cero y así, $\overline{y}_{reg} \simeq \widehat{\overline{y}}$. Pero la estimación de b en este último caso tendrá una gran varianza y sesgo, con lo que es preferible el estimador $\widehat{\overline{y}}$.

- Del mismo modo, si la relación es lineal pero la recta está cercana a pasar por el origen, o se conoce una relación de proporcionalidad directa entre x e y, es preferible la estimación de razón, al presentar un modelo más sencillo para la estimación.

- Un paso previo a la estimación de regresión debe por lo tanto incluir un estudio habitual de la correlación entre x e y, gráficos habituales de nubes de puntos y un estudio básico (ANOVA) de regresión. El contraste de hipótesis sobre la igualdad a cero del parámetro a, constante de la recta de regresión, puede ser un criterio a tener en cuenta para decidirse por estimación de razón en caso de que a no sea significativo.

Ejemplo 7.7.

Se dispone de una muestra obtenida bajo m.a.s. de 10 provincias españolas, con el número de reses sa-crificadas de ganado bovino y porcino, en 1998. Se desea estimar la media de cabezas de ganado bovino sacrificadas por provincia, considerada la población un conjunto de 50 provincias, y sabiendo que la media de cabezas de ganado porcino sacrificadas es $\overline{x} = 497229.98$.

A priori se puede pensar que el estimador de razón puede ser una buena alternativa. Pero también es cierto que sólo tenemos una idea vaga de la proporcionalidad que existe entre bovino y porcino, de manera que la relación pudiera ser lineal pero no exactamente proporcional. En general para este tipo de estudios se puede tener en cuenta datos de años anteriores o anteriores investigaciones similares, que ayudan a decidir por una estimación de razón o regresión.

Atendiendo a la muestra, esta es:

	275601	54413	59279	38210	2118	11583	19428
	3881483	139670	915660	195212	22428	150915	25301

ISBN 9788496866133

	88679	15986	16207
	328682	21558	179166

Tabla 7.7. Reses de ganado bovino y porcino sacrificadas en 1998 en España.

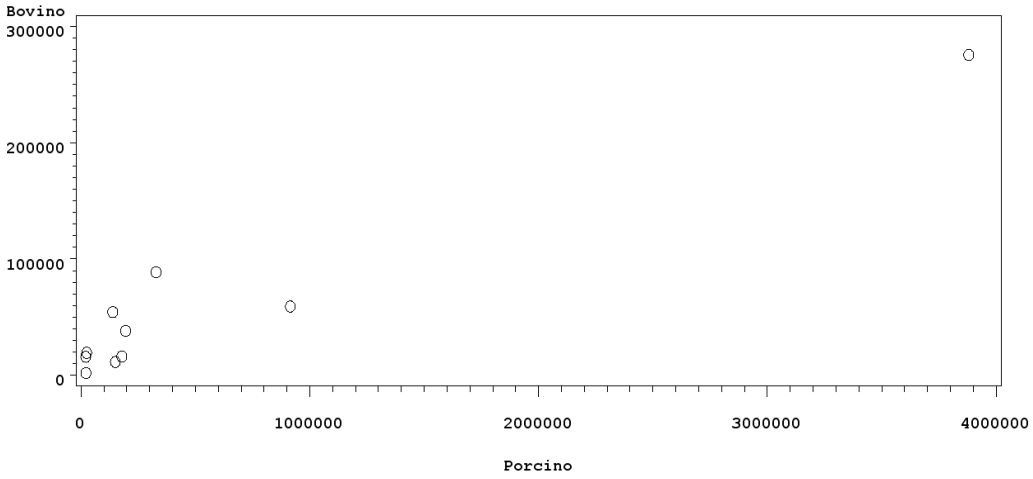


Figura 7.4. Reses sacrificadas de ganado bovino frente a porcino, 1998.

También se puede realizar un análisis habitual de regresión (tabla ANOVA) para verificar si la relación es puramente proporcional o no. El resultado siguiente es la salida del procedimiento reg del programa SAS para la muestra.

```
                        Analysis of Variance

                                  Sum of          Mean
Source                     DF     Squares        Square  F Value  Pr > F

Model                       1  54668988644  54668988644   101.58  <.0001
Error                       8   4305293948    538161744
Corrected Total             9  58974282592


Root MSE                23198    R-Square     0.9270
Dependent Mean          58150    Adj R-Sq     0.9179
Coeff Var           39.89364
```


ISBN 9788496866133

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	19691	8269.01534	2.38	0.0445
Porcino	1	0.06563	0.00651	10.08	<.0001

La constante a se puede estimar por $\widehat{a} = \widehat{\overline{y}} - \widehat{b}\widehat{\overline{x}} = 19691$.

Se observa que el contraste sobre la constante de la recta de regresión (estimada en 19691) da un p-valor de 0.0445. Con lo que no está claro si se puede adoptar estimación de regresión o de razón. Ciñéndose al simple resultado del contraste de hipótesis sobre la constante, para un nivel de significación de $\alpha = 0.05$ se adoptaría estimación de regresión y para $\alpha = 0.01$ estimación de razón.

Veamos las estimaciones directa, de regresión, y de razón, y las estimaciones de las varianzas. Los estadísticos para la muestra son:

$\widehat{\overline{y}} = 58150.40$, $\widehat{\overline{x}} = 586007.50$, $s_s = 1187547.73$, $s_y = 80948.74$, $r_{xy} = 0.96281$, $s_{xy} = 9.25 \cdot 10^{10}$.

Así, $\widehat{b} = \dfrac{s_{xy}}{s_x^2} = 0.0656$ (se puede también comprobar esto en la tabla ANOVA anterior) y entonces

$$\overline{y}_{reg} = \widehat{\overline{y}} + \widehat{b}(\overline{x} - \widehat{\overline{x}}) = 58150.40 + 0.0656(497229.98 - 586007.50) = 52326.59$$

y la varianza estimada del estimador será

$$\widehat{V}(\overline{y}_{reg}) = \frac{N-n}{Nn} s_y^2 (1 - r_{xy}^2) = \frac{50-10}{50 \cdot 10} 80948.74^2 (1 - 0.96281^2) = 38266136.26$$

y $\sqrt{\widehat{V}(\overline{y}_{reg})} = 6185.96$.

En cuanto a la estimación por razón, $\widehat{R} = \dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}} = 0.09923$ y $\overline{y}_R = \widehat{R}\overline{x} = 49340.13$.

La varianza estimada del estimador de razón es:

$$\widehat{V}(\overline{y}_R) = \frac{N-n}{Nn}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R} s_{xy}) =$$

$$= \frac{50-10}{50 \cdot 10}(80948.74^2 + 0.09923^2 1187547.73^2 - 2 \cdot 0.09923 \cdot 9.25 \cdot 10^{10}) = 166519939.7.$$

y $\sqrt{\widehat{V}(\overline{y}_R)} = 12904.26$.

Si no utilizamos la información auxiliar del ganado porcino, y estimamos la media directamente:

$\widehat{\overline{y}} = 58150.40$ con una varianza de $\widehat{V}(\widehat{\overline{y}}) = \dfrac{N-n}{Nn} s_y^2 = \dfrac{50-10}{50 \cdot 10} 80948.74^2 = 524215880.6$ y una desviación típica de $\sqrt{\widehat{V}(\widehat{\overline{y}})} = 22895.76$.

Obviamente, si tomamos como referencia las estimaciones de varianzas para comparar los tres estimadores la estimación por regresión es la más precisa, y por tanto la mejor, si suponemos el sesgo despreciable.

ISBN 9788496866133

Teniendo en cuenta el resultado del teorema de comparación , se observa que $r_{xy} = 0.96281$ (significativamente distinto de 0 para ese tamaño muestral) y además $\widehat{R} = 0.09923$ es aparentemente distinto de $\widehat{b} = 0.0656$, que junto con el p-valor relativamente significativo de la constante a de regresión induce a considerar que la recta de regresión no pasa por el origen. Esto corrobora que probablemente en este caso la estimación de regresión será la más adecuada.

7.2.4 Estimación de regresión en muestreo estratificado

Al igual que para la estimación de razón, se pueden realizar dos tipos de estimación en muestreo estratificado, con m.a.s. en cada estrato: estimación separada y combinada.

Definición .

El estimador **separado de regresión** de la media poblacional es

$\overline{y}_{regs} = \sum_{h=1}^{L} \frac{N_h}{N} \overline{y}_{reg_h}$, con

$\overline{y}_{reg_h} = \widehat{\overline{y}}_h + b_h(\overline{x}_h - \widehat{\overline{x}}_h)$,

siendo $\widehat{\overline{x}}_h$ y $\widehat{\overline{y}}_h$ las medias muestrales obtenidas en el estrato h.

Teorema 7.13 (varianza del estimador separado de regresión).

La varianza del estimador separado de regresión de la media cuando los b_h son constantes asignadas, es

$$V(\overline{y}_{regs}) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h}(S_{yh}^2 + b_h^2 S_{xh}^2 - 2b_h S_{xyh}).$$

Este resultado es directo por ser la suma de varianzas por estratos, al ser muestreo independiente en cada estrato.

Corolario 7.7 (estimador de la varianza).

Un estimador de la varianza del estimador separado de regresión de la media es

$$\widehat{V}(\overline{y}_{regs}) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h}(s_{yh}^2 + \widehat{b}_h^2 s_{xh}^2 - 2\widehat{b}_h s_{xyh}).$$

donde los $\widehat{b}_h$ son los estimadores de b_h, es decir,

$\widehat{b}_h = \frac{s_{xyh}}{s_{xh}^2}$.

Definición.

El estimador **combinado de regresión** de la media poblacional es

$\overline{y}_{regc} = \overline{y}_{st} + b(\overline{x} - \overline{x}_{st})$,

ISBN 9788496866133

donde $\overline{x}_{st}$ y $\overline{y}_{st}$ son las medias estratificadas $\overline{y}_{st} = \sum_{h=1}^{L} \frac{N_h}{N} \widehat{\overline{y}}_h$ y $\overline{x}_{st} = \sum_{h=1}^{L} \frac{N_h}{N} \widehat{\overline{x}}_h$.

Teorema 7.14 (varianza del estimador combinado de regresión).

La varianza del estimador combinado de regresión de la media considerando b fijo es

$$V(\overline{y}_{regc}) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} (S_{yh}^2 + b^2 S_{xh}^2 - 2bS_{xyh}).$$

Cuando la constante b es desconocida, el valor b a utilizar en el estimador proviene del siguiente teorema.

Teorema 7.15 (b óptimo en la varianza del estimador combinado).

El valor de b que minimiza la varianza del estimador combinado de regresión es

$$b_c = \frac{\sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} S_{xyh}}{\sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} S_{xh}^2}.$$

Demostración.

Definiendo $f(b) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} (S_{yh}^2 + b^2 S_{xh}^2 - 2bS_{xyh})$,

$f'(b) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} (2bS_{xh}^2 - 2S_{xyh})$ y $f''(b) > 0$ con lo que haciendo $f'(b) = 0$ se obtiene que

$b_c = \frac{\sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} S_{xyh}}{\sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} S_{xh}^2}$ es un mínimo.

Corolario 7.8 (forma del estimador combinado con b estimado).

Cuando b es desconocido, el estimador combinado de regresión toma la forma

$$\overline{y}_{regc} = \overline{y}_{st} + \widehat{b}_c(\overline{x} - \overline{x}_{st}),$$

donde $\widehat{b}_c$ es el estimador de los momentos de b_c, es decir

$$\widehat{b}_c = \frac{\sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} s_{xyh}}{\sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} s_{xh}^2}.$$

Corolario 7.9 (estimación de la varianza).

Cuando b es desconocido, el estimador de los momentos de la varianza del estimador de regresión combinado $\overline{y}_{regc}$ es

ISBN 9788496866133

$$\widehat{V}(\overline{y}_{regc}) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h}(s_{yh}^2 + \widehat{b}_c^2 s_{xh}^2 - 2\widehat{b}_c s_{xyh}).$$

Los mismos comentarios que se hicieron para comparar la estimación separada y combinada en estimación de razón son aplicables para la estimación de regresión.

Ejemplo 7.8.

En las Figuras 7.5 a 7.8 se explicarán gráficamente las diferencias entre las situaciones apropiadas para los distintos métodos de estimación indirecta en muestreo estratificado. Los segmentos representan la relación lineal entre x e y.

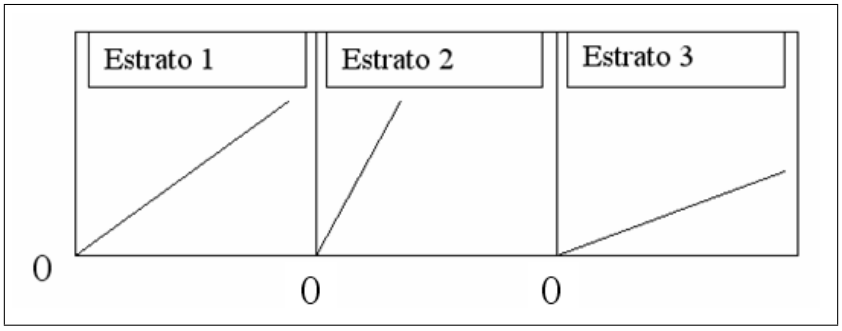


Figura 7.5. Estimador separado de razón.

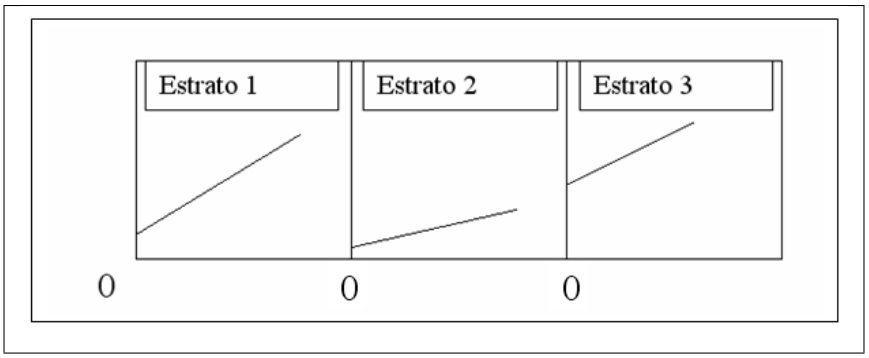


Figura 7.6. Estimador separado de regresión.

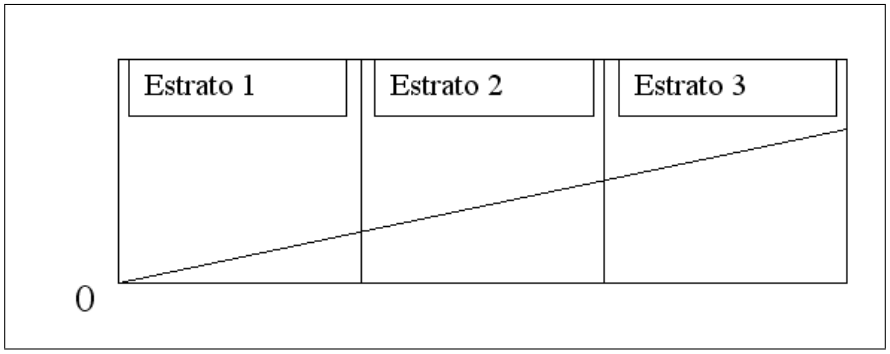


Figura 7.7. Estimador combinado de razón.

ISBN 9788496866133

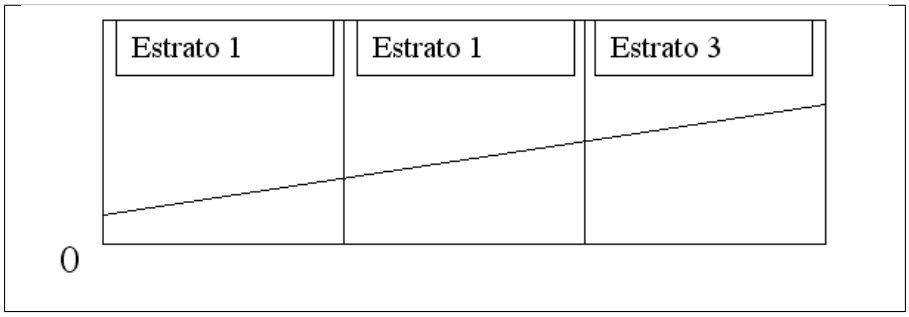


Figura 7.8. Estimador combinado de regresión.

El estimador separado de razón tiene sentido en casos en los que la razón (la pendiente de la recta ente y y x) varía de un estrato a otro.

El estimador separado de regresión tiene sentido en casos en los que la recta de regresión ente y y x varía de un estrato a otro, y esta recta no pasa por el origen en cada caso. Si alguna de las rectas por estrato pasara por el origen se utilizaría estimación de razón en ese estrato.

El estimador combinado de razón tiene sentido en casos en los que la razón (la pen-diente de la recta ente y y x) no varía mucho de un estrato a otro.

El estimador combinado de regresión tiene sentido en casos en los que la recta de regresión ente y y x no varía de un estrato a otro, y esta recta de regresión, común a todos los estratos, no pasa por el origen .

Ejemplo 7.9.

Es frecuente que las relaciones entre x e y varíen de un estrato a otro, no sólo en cuanto a la pendiente o constante de la recta de regresión, sino también a que la relación lineal puede ser importante en algún estrato y despreciable en otros. En estos casos se debe tener en cuenta la característica principal del muestreo estratificado: la independencia del muestreo entre los diferentes estratos. En general se intentará hacer en estos casos estimación separada, teniendo cuidado de escoger el mejor método en cada estrato. Veamos estas ideas a través de un ejemplo.

Se dispone de datos del crecimiento de tomates en tres regiones, de modo que las unidades elementales son las plantaciones de tomates. Tanto en la Región I como en la II y III se puede tomar como variable auxiliar para ayudar a estimar la producción total de tomates, la cosecha de calabacín del año anterior. En total hay $N_1 = 600$ plantaciones en la región I , $N_2 = 400$ plantaciones en la región II.y $N_3 = 800$ plantaciones en la región III. Además, se dispone de la cosecha total de calabacín en toneladas en cada región: $N_1\overline{x}_1 = 9120$ toneladas para la región I, 11200 toneladas para la región II, 15280 toneladas para la región III.

Se decide realizar muestreo estratificado con afijación proporcional con $n = 60$. Así, $n_1 = 20$, $n_2 = 13$, y $n_3 = 27$. Los datos muestrales obtenidos están en la tabla siguiente:

ISBN 9788496866133

	N_h	n_h	$\widehat{\overline{y}}_h$	$\widehat{\overline{x}}_h$	s_{xh}	s_{yh}	s_{xyh}	$\widehat{R}_h$	r_{xy}	$\overline{x}_h$	$\widehat{b}_h$
Región I	600	20	368.25	17.7	5.4	85.87	208.66	20.8	0.45	15.2	7.15
Región II	400	13	145	32.4	8.8	94.7	100	4.47	0.12	28	1.29
Región III	800	27	306.92	20.7	9.1	162.1	1224.3	14.82	0.83	19.1	14.78

Tabla 7.8. Estadísticos para cada región

Las estimaciones correspondientes de la media de tomates en cada estrato, por el estimador de razón, de regresión y media muestral son:

	$\overline{y}_{Rh}$	$\widehat{V}(\overline{y}_{Rh})$	$\overline{y}_{regh}$	$\widehat{V}(\overline{y}_{regh})$	$\widehat{\overline{y}}_h$	$\widehat{V}(\widehat{\overline{y}}_h)$
Región I	316.16	546.6	350.37	284.22	368.25	356.39
Región II	125.16	716.0	139.32	657.82	145	667.43
Región III	283.06	292.6	283.27	292.58	306.92	940.35

Tabla 7.9. Estimaciones para cada región

Para comparar mejor los tres métodos, observemos también los contrastes de regresión dados por la salida del paquete estadístico SAS, además del gráfico nube de puntos en cada estrato.

REGIÓN I

```
                        Analysis of Variance
                                 Sum of          Mean
Source                    DF     Squares        Square  F Value  Pr > F

Model                      1       90786         90786    15.29  0.0002
Error                     58      344342   5936.92244
Corrected Total           59      435127

Root MSE              77.05143    R-Square     0.2086
Dependent Mean       368.25000    Adj R-Sq     0.1950
Coeff Var             20.92367

                     Parameter Estimates
                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|
Intercept     1      239.75392       34.33219       6.98      <.0001
x1            1        7.25967        1.85647       3.91      0.0002
```


ISBN 9788496866133

REGIÓN II

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5511.87220	5511.87220	0.61	0.4403
Error	38	344415	9063.55534		
Corrected Total	39	349927			

Root MSE	95.20271	R-Square	0.0158
Dependent Mean	145.02500	Adj R-Sq	-0.0101
Coeff Var	65.64572		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	101.33372	58.01355	1.75	0.0888
x2	1	1.34746	1.72788	0.78	0.4403

REGIÓN III

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	474079	474079	56.51	<.0001
Error	25	209737	8389.48647		
Corrected Total	26	683816			

Root MSE	91.59414	R-Square	0.6933
Dependent Mean	306.92593	Adj R-Sq	0.6810
Coeff Var	29.84243		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.88433	44.36431	0.02	0.9843
x3	1	14.78197	1.96641	7.52	<.0001

ISBN 9788496866133

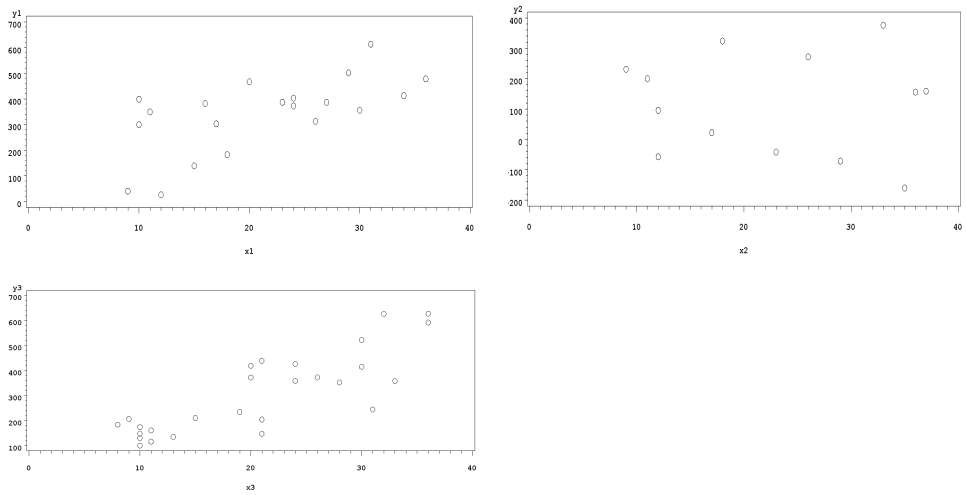


Figura 7.9. Relación entre y y x en las regiones I (arriba izda), II (arriba dcha), y III (abajo).

Por los análisis de regresión, gráficos y varianzas estimadas se intuye que en la Región I el mejor método de estimación es el de regresión. La varianza estimada es mucho menor que las otras dos, la recta parece no pasar por el origen y los dos contrastes de regresión, constante y pendiente, dan parámetros significativos.

En la Región II, no parece que se deba tener en cuenta la variable auxiliar: la varianza del estimador de regresión es muy similar a la del m.a.s., el R^2 es muy pequeño, y los contrastes de regresión no rechazan la nulidad de los dos parámeros de la recta, y se ve en el gráfico la horizontalidad de la relación entre x e y.

Por último, en la Región III la relación parece aproximadamente proporcional. La recta parece pasar por el origen, el contraste sobre la constante de regresión no rechaza la igualdad de ésta a cero, y la varianza estimada del estimador de razón es similar a la del estimador de regresión.

Habiendo decidido el mejor método de estimación en cada uno de los estratos, resta construir el estimador separado de la media o del total. En estos casos es más cómodo realizar los cálculos a partir de la estimación del total por estratos. Así, el total estimado en la Región I será, puesto que se ha decidido utilizar estimación de regresión en ese estrato,

$N_1\overline{y}_{reg1} = 600 \cdot 350.37 = 210222$ con una desviación típica de

$$\sqrt{\widehat{V}(N_1\overline{y}_{reg1})} = \sqrt{600^2 \cdot 284.22} = 10115.29.$$

Para la Región II será

$N_2\widehat{\overline{y}}_2 = 400 \cdot 145 = 58000$ con una desviación típica de

$$\sqrt{\widehat{V}(N_2\widehat{\overline{y}}_2)} = \sqrt{400^2 \cdot 667.43} = 10333.86.$$

Y el estimador del total en la Región III :

$N_3\overline{y}_{R3} = 800 \cdot 283.06 = 226448$ con una desviación típica de

$$\sqrt{\widehat{V}(N_3\overline{y}_{R3})} = \sqrt{800^2 \cdot 292.6} = 13684.4.$$

ISBN 9788496866133

Por lo tanto, un estimador del total poblacional será la suma de los tres estimadores por estrato, y su varianza será la suma de las tres varianzas:

$\widehat{N\overline{y}} = N_1\overline{y}_{reg1} + N_2\widehat{\overline{y}}_2 + N_3\overline{y}_{R3} = 494670.$

$\widehat{V}(\widehat{N\overline{y}}) = \widehat{V}(N_1\overline{y}_{reg1}) + \widehat{V}(N_2\widehat{\overline{y}}_2) + \widehat{V}(N_3\overline{y}_{R3}) = 396370557.6.$

Utilizando este resultado, la estimación de la media poblacional de producción de tomates por plantación será

$\widehat{\overline{y}} = \frac{1}{N}\widehat{N\overline{y}} = \frac{494670}{1800} = 274.8.$

$\widehat{V}(\widehat{\overline{y}}) = \frac{1}{N^2}\widehat{V}(\widehat{N\overline{y}}) = 122.34.$

Hay que recordar que este estimador construido no es insesgado pues tanto el estimador de razón como el de regresión son sesgados. Aunque se supone que el sesgo será pequeño debido a la relación lineal clara entre x e y en los estratos 1 y 3.

7.3 Tablas de fórmulas

ESTIMACIÓN DE RAZÓN bajo m.a.s.			
Parámetro poblacional	R	$\overline{y}$	$N\overline{y}$
Estimador	$\widehat{R} = \frac{\widehat{\overline{y}}}{\widehat{\overline{x}}}$	$\overline{y}_R = \widehat{R}\overline{x}$	$N\overline{y}_R$
Varianza	$V(\widehat{R}) \simeq \frac{N-n}{Nn\overline{x}^2}(S_y^2 + R^2S_x^2 - 2RS_{xy})$	$\overline{x}^2V(\widehat{R})$	$N^2V(\overline{y}_R)$
Estimador de la Varianza	$\widehat{V}(\widehat{R}) = \frac{N-n}{Nn\overline{x}^2}(s_y^2 + \widehat{R}^2s_x^2 - 2\widehat{R}s_{xy})$	$\overline{x}^2\widehat{V}(\widehat{R})$	$N^2\widehat{V}(\overline{y}_R)$
Sesgo	$B(\widehat{R}) = \frac{-cov(\widehat{R},\widehat{\overline{x}})}{\overline{x}}$	$\overline{x}B(\widehat{R})$	$N\overline{x}B(\widehat{R})$
Aproximación al Sesgo	$\frac{-cov(\widehat{\overline{y}},\widehat{\overline{x}}) + RV(\widehat{\overline{x}})}{\overline{x}^2}$	$\overline{x}\widehat{B}(\widehat{R})$	$N\overline{x}\widehat{B}(\widehat{R})$

Expresiones alternativas para las varianzas en estimación de razón.

$$V(\overline{y}_R) = \frac{N-n}{Nn}\frac{1}{N-1}(\sum_{i=1}^{N} y_i^2 + R^2\sum_{i=1}^{N} x_i^2 - 2R\sum_{i=1}^{N} x_iy_i)$$

$$\widehat{V}(\overline{y}_R) = \frac{N-n}{Nn}\frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 + \widehat{R}^2\sum_{i=1}^{n} x_i^2 - 2\widehat{R}\sum_{i=1}^{n} x_iy_i)$$

ISBN 9788496866133

$$V(\overline{y}_R) = \frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}(y_i - Rx_i)^2$$

$$\widehat{V}(\overline{y}_R) = \frac{N-n}{Nn}\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{R}x_i)^2$$

ESTIMACIÓN DE REGRESIÓN		
Parámetro poblacional	$\overline{y}$	$N\overline{y}$
Estimador	$\overline{y}_{reg} = \widehat{\overline{y}} + \widehat{b}(\overline{x} - \widehat{\overline{x}})$	$N\overline{y}_{reg}$
Varianza	$V(\overline{y}_{reg}) = \frac{N-n}{Nn}(S_y^2 + b^2S_x^2 - 2bS_{xy})$	$N^2V(\overline{y}_{reg})$
Estimador de la Varianza	$\widehat{V}(\overline{y}_{reg}) = \frac{N-n}{Nn}(s_y^2 + \widehat{b}^2s_x^2 - 2\widehat{b}s_{xy})$	$N^2\widehat{V}(\overline{y}_{reg})$

con

$$\widehat{b} = \frac{s_{xy}}{s_x^2}.$$

Expresiones alternativas para las varianzas:

$$V(\overline{y}_{reg}) = \frac{N-n}{Nn}(1-\rho_{xy}^2)S_y^2.$$

$$\widehat{V}(\overline{y}_{reg}) = \frac{N-n}{Nn}(1-r_{xy}^2)s_y^2.$$

ESTIMADOR SEPARADO DE RAZÓN			
Parámetro poblacional	$\overline{y}$	R	$N\overline{y}$
Estimador	$t_s = \sum_{h=1}^{L}\frac{N_h}{N}\overline{y}_{R_h} = \sum_{h=1}^{L}W_h\widehat{R}\overline{x}_h$	$\widehat{R}_s = \frac{t_s}{\overline{x}}$	Nt_s
Varianza aprox.	$\sum_{h=1}^{L}W_h^2\frac{(N_h-n_h)}{N_hn_h}(S_{yh}^2 + R_h^2S_{xh}^2 - 2R_hS_{xyh})$	$\frac{1}{\overline{x}^2}V(t_s)$	$N^2V(t_s)$
Estimador de la Varianza	$\sum_{h=1}^{L}W_h^2\frac{(N_h-n_h)}{N_hn_h}(s_{yh}^2 + \widehat{R}_h^2s_{xh}^2 - 2\widehat{R}_hs_{xyh})$	$\frac{1}{\overline{x}^2}\widehat{V}(t_s)$	$N^2\widehat{V}(t_s)$

ISBN 9788496866133

ESTIMADOR COMBINADO DE RAZÓN			
Parámetro poblacional	$\overline{y}$	R	$N\overline{y}$
Estimador	$t_c = \dfrac{\overline{y}_{st}}{\overline{x}_{st}}\overline{x}$	$\widehat{R}_c = \dfrac{\overline{y}_{st}}{\overline{x}_{st}}$	Nt_c
Varianza aprox.	$\sum_{h=1}^{L} W_h^2 \dfrac{(N_h - n_h)}{N_h n_h}(S_{yh}^2 + R^2 S_{xh}^2 - 2RS_{xyh})$	$\dfrac{1}{\overline{x}^2}V(t_c)$	$N^2V(t_c)$
Estimador de la Varianza	$\sum_{h=1}^{L} W_h^2 \dfrac{(N_h - n_h)}{N_h n_h}(s_{yh}^2 + \widehat{R}_c^2 s_{xh}^2 - 2\widehat{R}_c s_{xyh})$	$\dfrac{1}{\overline{x}^2}\widehat{V}(t_c)$	$N^2\widehat{V}(t_c)$

ESTIMADOR SEPARADO DE REGRESIÓN		
Parámetro poblacional	$\overline{y}$	$N\overline{y}$
Estimador	$t_{regs} = \sum_{h=1}^{L} \dfrac{N_h}{N}\overline{y}_{regh}$	Nt_{regs}
Varianza	$V(t_{regs}) = \sum_{h=1}^{L} W_h^2 \dfrac{(N_h - n_h)}{N_h n_h}(S_{yh}^2 + b_h^2 S_{xh}^2 - 2b_h S_{xyh})$	$N^2V(t_{regs})$
Estimador de la Varianza	$V(t_{regs}) = \sum_{h=1}^{L} W_h^2 \dfrac{(N_h - n_h)}{N_h n_h}(s_{yh}^2 + \widehat{b}_h^2 s_{xh}^2 - 2\widehat{b}_h s_{xyh})$	$N^2\widehat{V}(t_{regs})$

ESTIMADOR COMBINADO DE REGRESIÓN		
Parámetro poblacional	$\overline{y}$	$N\overline{y}$
Estimador	$\overline{y}_{regc} = \overline{y}_{st} + \widehat{b}_c(\overline{x} - \overline{x}_{st})$	$N\overline{y}_{regc}$
Varianza	$V(\overline{y}_{regs}) = \sum_{h=1}^{L} W_h^2 \dfrac{(N_h - n_h)}{N_h n_h}(S_{yh}^2 + b^2 S_{xh}^2 - 2bS_{xyh})$	$N^2V(\overline{y}_{regc})$
Estimador de la Varianza	$\widehat{V}(\overline{y}_{regc}) = \sum_{h=1}^{L} W_h^2 \dfrac{(N_h - n_h)}{N_h n_h}(s_{yh}^2 + \widehat{b}_c^2 s_{xh}^2 - 2\widehat{b}_c s_{xyh})$	$N^2\widehat{V}(\overline{y}_{regc})$

Donde

ISBN 9788496866133

$$\widehat{b}_c = \frac{\sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} s_{xyh}^2}{\sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} s_{xh}^2}$$

ISBN 9788496866133

7.4 Estimación de razón y regresión con SAS bajo m.a.s.

Este capítulo asume muestreo aleatorio simple, tanto sencillo como por estratos. Por lo tanto, para la obtención de muestras, aunque el objetivo sea la estimación de razón o regresión, el método utilizado es el mismo estudiado en el capítulo de muestreo aleatorio simple sin reemplazamiento.

7.4.1 Estimación de razón y regresión en muestreo aleatorio simple sin estratos

Utilizaremos para ello la macro razreg. La sintaxis es la siguiente:

```
%estimrazreg (archivo,variabley,variablex,mediax,ngrande,n);
```

donde

archivo es el archivo que contiene los datos muestrales.

variabley es el nombre de la variable y.

variablex es el nombre de la variable x .

mediax es la media poblacional de x (se puede aproximar por la media muestral si no se dispone de esa información).

ngrande es el tamaño poblacional N.

n es el tamaño muestral.

Si por ejemplo se desea estimar por razón o regresión la razón, media o total de la variable y llamada estatura, que está junto a la variable x llamada peso en el archivo muestral llamado muestra2, con $n = 100$ observaciones, que proviene de una población con $N = 1000$ individuos y la media poblacional de la variable peso es conocida e igual a 71, la sintaxis sería:

```
%estimrazreg (muestra2,estatura,peso,71,1000,100);
```

Esta macro presenta en la ventana LOG los estimadores de razón y regresión de la media y total, con sus varianzas e intervalos de confianza al 95%, así como el estimador de la razón y su varianza e intervalo de confianza al 95%.

7.4.2 Estimación de razón y regresión en muestreo aleatorio simple estratificado

Para ello se utilizará la macro estimrazestrat. La sintaxis es la siguiente:

```
estimrazestrat (archivo1,archivo2,variabley,variablex,varestrato,
vartama,varmedx,mediax,indicador,ngrande);
```


ISBN 9788496866133

archivo1 es el archivo que contiene los datos muestrales.

archivo2 es el archivo que contiene los datos de cada estrato (opcional).

variabley es el nombre de la variable y.

variablex es el nombre de la variable x .

varestrato es la variable que indica el estrato.

vartama es la variable que contiene el tamaño por estrato Nh.

varmedx es la variable que contiene la media de x por estrato.

mediax es la media poblacional de x (se puede aproximar por la media muestral si no se dispone de esa información).

indicador

1 si es el archivo1 muestral el que contiene los datos de cada estrato.

2 si es el archivo2 el que contiene los datos de cada estrato.

ngrande es el tamaño poblacional N.

Observaciones:

- Si existe el archivo2 debe de tomar la forma (los nombres de las variables pueden variar):

```
estrato Nh  medxh
1       300 6.5
2       400 7.8
....
```

- Si no existe el archivo2, en todas las observaciones del archivo1 deben estar presentes las variables de tamaño de estrato y de medias de x, y son constantes por estrato.

- Si no se dispone de la media poblacional de x, medx, puede sustituirse por la media muestral por estrato, o bien se deja como ausente (misssing) y la estimación separada no es realizada, existiendo sin embargo la estimación combinada.

Supongamos que se desean estimar medias o totales por razón o regresión, separadas o combinadas. El archivo muestral se denomina datos1 y el archivo con la información por estratos datestrat. La variable estatura es la variable de interés, mientras que la variable auxiliar es la variable peso. La media poblacional del peso es conocida e igual a 71. La población tiene tamaño $N = 1000$. La variable de tamaño de estrato se llama nh, la variable estratos ciudad y la variable de media de x por estratos medx. La sintaxis sería:

```
estimrazestrat (datos1,datestrat,estatura,peso,estrato,
nh,medx,71,2,1000);
```


ISBN 9788496866133

Esta macro presenta en la ventana LOG los estimadores de razón y regresión separados y combinados de la media y total, con sus varianzas e intervalos de confianza al 95%, así como el estimador separado y combinado de la razón y su varianza e intervalo de confianza al 95%.

ISBN 9788496866133

7.5 Ejercicios resueltos

Ejercicio 6.1.

En una empresa se dispone de datos de facturación de cierto producto en las 100 sucursales de esta empresa, del año 2001. El total de facturación ese año fue de 200 millones.

El año 2002 sólo se puede disponer de una muestra tomada por m.a.s. de 5 sucursales, arrojando los siguientes datos (en millones):

2001	1	1.5	3	2	4
2002	0.75	2	3.5	2.5	5

a) Dar un intervalo de confianza al 95% para la media de facturación del año 2002, utilizando estimación de razón

b) Dar un intervalo de confianza al 95% utilizando estimación de expansión (la usual por m.a.s.).

c) Dar un intervalo de confianza al 95% utilizando estimación de regresión.

---

a) Se llamará y a los datos en 2002 y x a los datos en 2001. La media poblacional en 2001 es $\overline{x} = \dfrac{200}{100} = 2$.

El estimador de la razón es $\widehat{R} = \dfrac{13.75}{11.5} = 1.19$.

El estimador de razón de la media es $\overline{y}_R = \widehat{R}\overline{x} = 1.19. \cdot 2 = 2.39$.

Para estimar su varianza, se calcularán s_y^2, s_x^2 y s_{xy}. Estos son:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{\overline{y}})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\widehat{\overline{y}}^2\right) = 2.56$$

s_x^2 se calcula análogamente y es $s_x^2 = 1.45$.

Además, $s_{xy} = \dfrac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\widehat{\overline{xy}}\right) = 1.906$.

Entonces,

$$\widehat{V}(\overline{y}_R) = \frac{N-n}{Nn}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}) = \frac{200-5}{200\cdot 5}(2.56 + 1.19^2 \cdot 1.45 - 2\cdot 1.19 \cdot 1.906) = 0.015.$$

El intervalo de confianza, asumiendo normalidad del estimador, es:

$(2.39 - 1.96\sqrt{0.015}, 2.39 + 1.96\sqrt{0.015}) = (2.15, 2.63)$.

b) El estimador es la media muestral, $\widehat{\overline{y}} = \dfrac{13.75}{5} = 2.75$.

Su varianza estimada es

$$\widehat{V}(\widehat{\overline{y}}) = \frac{N-n}{N}\frac{s_y^2}{n} = \frac{100-5}{100}\frac{2.56}{5} = 0.48.$$

ISBN 9788496866133

Se ve que es mucho menor, al ser el coeficiente de correlación entre x e y alto (el coeficiente de correlación muestral es r=0.988).

y el intervalo de confianza

$(2.75 - 1.96\sqrt{0.48}, 2.75 - 1.96\sqrt{0.48}) = (1.38, 4.11)$.

c) El coeficiente de regresión es $\widehat{b} = \dfrac{s_{xy}}{s_x^2} = 1.31$.

El estimador de regresión será

$$\overline{y}_{reg} = \widehat{\overline{y}} + \frac{s_{xy}}{s_x^2}(\overline{x} - \widehat{\overline{x}}) = 2.75 + 1.31 \cdot (2 - 2.3) = 2.35.$$

La varianza estimada del estimador de regresión será

$$\widehat{V}(\overline{y}_{reg}) = \frac{N-n}{Nn}(s_y^2 + \widehat{b}^2 s_x^2 - 2\widehat{b}s_{xy}) = \frac{100-5}{100 \cdot 5}(2.56 + 1.31^2 \cdot 1.45 - 2 \cdot 1.31 \cdot 1.906) = 0.010.$$

El intervalo de confianza será $(2.15, 2.56)$.

Ejercicio 6.2

En un estudio agrario se pretende estimar el número de plantas de tomate en un terreno. El terreno está dividido en 120 parcelas de distinto tamaño, y se sabe que en total hay 3000 m^2. Se extrae una muestra por muestreo sistemático de 10 parcelas, y se registra para cada parcela el tamaño del terreno en m^2 y el número de plantas de tomate que contiene. Se obtiene la siguiente tabla:

Tamaño	10	15	30	50	20	45	15	10	5	55
Plantas	150	250	400	600	300	500	100	100	45	650

a) Estimar por razón el total de plantas de tomate. Dar una estimación de la varianza del estimador.

b) Realizar los mismos cálculos con el estimador por m.a.s.

Al suponerse que el muestreo sistemático en este caso al menos tiene una varianza tan pequeña como el m.a.s., se utilizarán las fórmulas de m.a.s., sabiendo que el error de muestreo puede ser incluso menor.

La media muestral del tamaño es $\widehat{\overline{x}} = 25.5$, y de las plantas, $\widehat{\overline{y}} = 309.5$. También $s_y^2 = 47958$, $s_x^2 = 335.83$ y $s_{xy} = 3922.5$.

El estimador de la razón del total es:

$$N\overline{y}_R = \frac{\widehat{\overline{y}}}{\widehat{\overline{x}}} N\overline{x} = \frac{309.5}{25.5} 3000 = 36.412$$

y $\widehat{R} = 12.13$

La varianza estimada de este estimador es

$$\widehat{V}(N\overline{y}_R) = N^2 \frac{N-n}{Nn}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}).$$

ISBN 9788496866133

Así,

$$\widehat{V}(N\overline{y}_R) \simeq 120\frac{120-10}{10}(47958 + 12.13^2 \cdot 335.83 - 2 \cdot 12.13 \cdot 3922.5) = 2.922.337$$

b) El estimador del total será $N\widehat{\overline{y}} = 37140$, y su varianza estimada $\widehat{V}(N\widehat{\overline{y}}) = N^2\frac{N-n}{N}\frac{s_y^2}{n} = 120(120-10)\frac{47958}{10} = 63.304.560$.

Ejercicio 6.3

Se quiere estimar el porcentaje de personas en un edificio que tienen móvil. El edificio consta de 60 viviendas, de las cuales se escogen 10 por m.a.s. Se sabe que en total hay 210 personas en el edificio. En las viviendas muestreadas, se obtienen los datos siguientes:

nº de personas	3	5	6	1	7	5	2	3	6	7
nº personas con móvil	2	3	3	1	4	3	1	1	4	5

a) Estimar por razón la proporción de personas con móvil y dar un intervalo de confianza al 95% suponiendo normalidad.

b) Estimar, utilizando m.a.s., la proporción de personas con móvil. Comparar este estimador con el anterior.

---

a) Si llamamos x a la variable número de personas en la familia, e y a la variable número de personas con móvil, la proporción poblacional de personas con móvil es $R = \frac{\sum_{i=1}^N y_i}{\sum_{i=1}^N x_i}$. Por lo tanto la razón muestral $\widehat{R}$ será el estimador de esta proporción:

$$\widehat{R} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = 0.6.$$

Además , $\overline{x} = \frac{210}{60} = 3.5$, y se calculan también $s_y^2 = 2.011$, $s_x^2 = 4.5$, y $s_{xy} = 2.83$, con lo que la varianza del estimador de la razón es

$$\widehat{V}(\widehat{R}) = \frac{N-n}{Nn\overline{x}^2}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}) = \frac{60-10}{60 \cdot 10 \cdot 3.5^2}(2.011 + 0.6^2 \cdot 4.5 - 2 \cdot 0.6 \cdot 2.83) = 0.0015.$$

El intervalo de confianza, asumiendo normalidad del estimador, es:

$$(0.6 - 1.96\sqrt{0.0015}, 0.6 - 1.96\sqrt{0.0015}) = (0.52, 0.67).$$

b) Sin ayuda de la variable auxiliar, se estima el total de personas con móvil, y como el número de personas total es una constante conocida, se estima la proporción por el cociente entre ambas cantidades.

El total de personas con móvil se estima por $N\widehat{\overline{y}} = 60 \cdot 2.70 = 162$.

La varianza estimada de este estimador es

$$\widehat{V}(N\widehat{\overline{y}}) = N^2\frac{N-n}{N}\frac{s_y^2}{n} = 60(60-10) \cdot \frac{2.011}{10} = 603.3.$$

La proporción estimada de personas con móvil es $\frac{N\widehat{\overline{y}}}{210} = \frac{162}{210} = 0.77$, y este estimador tiene varianza estimada

ISBN 9788496866133

$\widehat{V}(\frac{N\widehat{\overline{y}}}{210}) = \frac{1}{210^2}\widehat{V}(N\widehat{\overline{y}}) = \frac{1}{210^2}603.3 = 0.01368.$

Parece que la precisión aumenta, utilizando la información de la variable auxiliar (estimación de razón) . El coeficiente de correlación muestral entre x e y es en este caso $r = \frac{s_{xy}}{s_x s_y} = 0.94$, y el cociente entre los coeficientes de variación estimados es

$\frac{\widehat{CV}(x)}{\widehat{CV}(y)} = \frac{s_x/\widehat{\overline{x}}}{s_y/\widehat{\overline{y}}} = 0.90$, con lo que $r > \frac{1}{2}\frac{\widehat{CV}(x)}{\widehat{CV}(y)} = 0.45.$

Ejercicio 6.4

En un laboratorio se desea estimar el volumen del estómago de cierta especie de rata. Hay ratas machos y hembras. Se realiza m.a.s. de tamaño $n = 5$ de los 30 machos que hay y $n = 6$ de las 40 hembras que hay. Se dispone de la información del peso de las ratas. El peso total de las ratas macho es de 7500 gramos, y el de las ratas hembra de 12000 gramos. Los datos muestrales obtenidos fueron, para las ratas macho:

Volumen estómago	1.75	1	2.5	1.5	2
Peso	200	150	400	150	250

Para las ratas hembra:

Volumen estómago	2.25	3	2.5	1.5	1.8
Peso	250	300	350	100	200

a) Estimar por razón y regresión el volumen medio del estómago de las ratas macho y de las ratas hembra, por separado y calcular la varianza de los estimadores.

b) Estimar el volumen medio del estómago de las ratas por estimación de razón separada y combinada.

---

a)

ESTIMACIÓN DE RAZÓN

Se llamará y al volumen del estómago y x al peso.

Para las ratas macho, la media poblacional del peso es $\overline{x} = \frac{7500}{30} = 250$. En la muestra, se tiene $\widehat{\overline{x}} = 230, \widehat{\overline{y}} = 1.75$, $s_y^2 = 0.31$, $s_x^2 = 10750$ y $s_{xy} = 53.12$.

La estimación por razón de la media del volumen de estómago será $\overline{y}_R = \widehat{R}\overline{x} = \frac{1.75}{230}250 = 1.90$, con $\widehat{R} = 0.0076$.

La varianza de la estimación será:

$\widehat{V}(\overline{y}_R) = \frac{N-n}{Nn}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}) = \frac{30-5}{30\cdot 5}(0.31 + 0.0076^2 \cdot 10750 - 2 \cdot 0.0076 \cdot 53.12) = 0.021.$

En el caso de las ratas hembra, se tiene que la media poblacional del peso es $\overline{x} = \frac{12000}{40} = 300$. En la muestra, se tiene $\widehat{\overline{x}} = 240, \widehat{\overline{y}} = 2.21$, $s_y^2 = 0.34$, $s_x^2 = 9250$ y $s_{xy} = 48.87$.

ISBN 9788496866133

La estimación por razón de la media del volumen de estómago será $\overline{y}_R = \widehat{R}\overline{x} = \frac{2.21}{240}300 = 2.76$, con $\widehat{R} = 0.009$.

La varianza de la estimación será:

$$\widehat{V}(\overline{y}_R) = \frac{N-n}{Nn}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}) = 0.04.$$

ESTIMACIÓN DE REGRESIÓN

En el caso de las ratas macho, $\widehat{b} = \frac{s_{xy}}{s_x^2} = \frac{53.12}{10750} = 0.00494$.

El estimador de la media es

$$\overline{y}_{reg} = \widehat{\overline{y}} + \widehat{b}(\overline{x} - \widehat{\overline{x}}) = 1.75 + 0.00494(250 - 230) = 1.848.$$

Y su varianza estimada es

$$\widehat{V}(\overline{y}_{reg}) = \frac{N-n}{Nn}(s_y^2 + \widehat{b}^2 s_x^2 - 2\widehat{b}s_{xy}) = 0.008.$$

En el caso de las ratas hembra, $\widehat{b} = \frac{s_{xy}}{s_x^2} = \frac{48.87}{9250} = 0.0052$.

El estimador de la media es

$$\overline{y}_{reg} = \widehat{\overline{y}} + \widehat{b}(\overline{x} - \widehat{\overline{x}}) = 2.52$$

Y su varianza estimada es

$$\widehat{V}(\overline{y}_{reg}) = \frac{N-n}{Nn}(s_y^2 + \widehat{b}^2 s_x^2 - 2\widehat{b}s_{xy}) = 0.015.$$

b) El estimador separado de razón es la suma ponderada de los estimadores de razón en cada estrato:

$$t_s = \sum_{h=1}^{L} \frac{N_h}{N}\overline{y}_{R_h} = \frac{30}{70}1.90 + \frac{40}{70}2.76 = 2.39.$$

Su varianza será la suma de las varianzas con las ponderaciones al cuadrado:

$$\widehat{V}(t_s) = \sum_{h=1}^{L} (\frac{N_h}{N})^2 \widehat{V}(\overline{y}_{R_h}) = (\frac{30}{70})^2 0.021 + (\frac{40}{70})^2 0.04 = 0.0169.$$

El estimador separado de regresión se construye de manera similar:

$$t_{regs} = \sum_{h=1}^{L} \frac{N_h}{N}\overline{y}_{regh} = 2.23$$

con

$$\widehat{V}(t_{regs}) = \sum_{h=1}^{L} (\frac{N_h}{N})^2 \widehat{V}(\overline{y}_{reg}) = 0.006.$$

Para los estimadores combinados, es necesario calcular las estimaciones estratificadas:

$$\overline{y}_{st} = \frac{30}{70}1.75 + \frac{40}{70}2.21 = 2.01$$

y

$$\overline{x}_{st} = \frac{30}{70}230 + \frac{40}{70}240 = 235.7.$$

ISBN 9788496866133

Además, $\overline{x} = \frac{30}{70}250 + \frac{40}{70}300 = 278.57$.

Así, el estimador combinado de razón de la media es

$$t_c = \frac{\overline{y}_{st}}{\overline{x}_{st}}\overline{x} = \frac{2.01}{235.7}278.57 = 2.37.$$

Su varianza es

$$\widehat{V}(t_c) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h}(s_{yh}^2 + \widehat{R}_c^2 s_{xh}^2 - 2\widehat{R}_c s_{xyh})$$

donde $\widehat{R}_c = \frac{\overline{y}_{st}}{\overline{x}_{st}} = 0.00852$.

Calculando queda:

$$\widehat{V}(t_c) = 0.016.$$

Para el estimador combinado de regresión, se tiene que

$$\overline{y}_{regc} = \overline{y}_{st} + \widehat{b}_c(\overline{x} - \overline{x}_{st})$$

donde

$$\widehat{b}_c = \frac{\sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} s_{xyh}}{\sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h} s_{xh}^2} = 0.005.$$

Así,

$$\overline{y}_{regc} = \overline{y}_{st} + \widehat{b}_c(\overline{x} - \overline{x}_{st}) = 2.01 + 0.005 \cdot (278.57 - 235.7) = 2.23.$$

y

$$\widehat{V}(\overline{y}_{regc}) = \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h}(s_{yh}^2 + \widehat{b}_c^2 s_{xh}^2 - 2\widehat{b}_c s_{xyh}) = 0.006.$$

Ejercicio 6.5

Se desea estimar la proporción de items defectuosos en un proceso de control de calidad donde el número de items varía por lote. Se sabe que los items están repartidos en 100 lotes, de los cuales se escogen 10 por m.a.s., obteniendo los siguientes datos:

Lote	1	2	3	4	5	6	7	8	9	10
nº de items	25	35	40	10	15	45	20	30	10	50
nº de items defectuosos	3	4	5	2	2	5	2	2	1	6

Estimar la proporción de items defectuosos y dar un intervalo de confianza al 95%, despreciando el coeficiente de correción por población finita.

La proporción de items defectuosos se estima por la razón muestral $\widehat{R} = \frac{\widehat{\overline{y}}}{\widehat{\overline{x}}} = \frac{3.2}{25.5} = 0.125$.

ISBN 9788496866133

Además, $s_y^2 = 2.84$, $s_x^2 = 335.83$, y $s_{xy} = 19.33$.

Si se desprecia el coeficiente de corrección por población finita, la varianza es:

$$\widehat{V}(\widehat{R}) = \frac{N-n}{Nn\overline{x}^2}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}) \simeq \frac{1}{n\overline{x}^2}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}).$$

Al no disponer de la cantidad de items en total, hay que aproximar $\overline{x}^2$ por $\widehat{\overline{x}}^2$, con lo que queda

$$\widehat{V}(\widehat{R}) \simeq \frac{1}{10 \cdot (25.5)^2}(2.84 + 0.125^2 \cdot 335.83 - 2 \cdot 0.125 \cdot 19.33) = 0.0005.$$

El intervalo de confianza al 95% para la proporción de interés será:

$$(0.125 - 1.96\sqrt{0.0005}, 0.125 + 1.96\sqrt{0.0005}) = (0.08, 0.168).$$

Ejercicio 6.6

Se dispone de una población con 5 observaciones de las variables x e y, como aparecen en la tabla.

Observación	1	2	3	4	5
x	1	3	4	6	5
y	0.5	2	5	7	4

Si se realiza m.a.s. de $n = 2$, Responder a las preguntas siguientes:

a) Presentar la distribución del estimador de razón de la media tomando x como variable auxiliar. Calcular su esperanza y su varianza.

b) Calcular el sesgo del estimador de razón de la media poblacional de y directamente y también con la fórmula $B(\overline{y}_R) = -cov(\widehat{R}, \widehat{\overline{x}})$.

*

---

a) Hay $\binom{5}{2} = 10$ muestras posibles, cada una con igual probabilidad $\frac{1}{10}$. Se sabe que $\overline{x} = 3.8$. Para cada una de ellas, se calcula la razón $\widehat{R}$ y el estimador de razón de la media $\overline{y}_R$.

Por ejemplo, para la muestra $\{1, 2\}$ $\widehat{R} = \frac{0.5+2}{1+3} = 0.625$ y $\overline{y}_R = \widehat{R}\overline{x} = 0.625 \cdot 3.8 = 2.375$.

ISBN 9788496866133

$Muestra$	$\widehat{R}$	$\overline{y}_R$	p	$\widehat{\overline{x}}$
$\{1,2\}$	0.625	2.375	$\frac{1}{10}$	2
$\{1,3\}$	1.1	4.18	$\frac{1}{10}$	2.5
$\{1,4\}$	1.07	4.07	$\frac{1}{10}$	3.5
$\{1,5\}$	0.75	2.85	$\frac{1}{10}$	3
$\{2,3\}$	1	3.8	$\frac{1}{10}$	3.5
$\{2,4\}$	1	3.8	$\frac{1}{10}$	4.5
$\{2,5\}$	0.75	2.85	$\frac{1}{10}$	4
$\{3,4\}$	1.2	4.56	$\frac{1}{10}$	5
$\{3,5\}$	1	3.8	$\frac{1}{10}$	4.5
$\{4,5\}$	1	3.8	$\frac{1}{10}$	5.5

La distribución del estimador es entonces:

$\overline{y}_R$	p
2.375	$\frac{1}{10}$
2.85	$\frac{2}{10}$
3.8	$\frac{4}{10}$
4.07	$\frac{1}{10}$
4.18	$\frac{1}{10}$
4.56	$\frac{1}{10}$

La esperanza de $\overline{y}_R$ es $E(\overline{y}_R) = \frac{1}{10}(2.375 + 2 \cdot 2.85 + 4 \cdot 3.8 + 4.07 + 4.18 + 4.56) = 3.6$

La varianza de $\overline{y}_R$ es

$V(\overline{y}_R) = \frac{1}{10}((2.375 - 3.6)^2 + 2 \cdot (2.85 - 3.6)^2 + 4 \cdot (3.8 - 3.6)^2 +$

$(4.07 - 3.6)^2 + (4.18 - 3.6)^2 + (4.56 - 3.6)^2) = 0.426.$

b) Directamente, como $\overline{y} = 3.7$, el sesgo es $B(\overline{y}_R) = E(\overline{y}_R) - \overline{y} = 3.6 - 3.7 = -0.1$.

Para calcularlo como $B(\overline{y}_R) = -cov(\widehat{R}, \widehat{\overline{x}})$, para calcular esta covarianza se ve que

$E(\widehat{R}) = \frac{1}{\overline{x}} E(\overline{y}_R) = 0.947$ y $E(\widehat{\overline{x}}) = \overline{x} = 3.8$ por ser m.a.s. Así

$cov(\widehat{R}, \widehat{\overline{x}}) = E[(\widehat{R} - E(\widehat{R}))((\widehat{\overline{x}} - E(\widehat{\overline{x}}))] =$

ISBN 9788496866133

$$=\frac{1}{10}[(0.625-0.947)(2-3.8)+...+(1-0.947)(5.5-3.8)]=0.1.$$

Con lo cual

$B(\overline{y}_R)=-cov(\widehat{R},\widehat{\overline{x}})=-0.1$ como se había calculado directamente.

Ejercicio 6.7

En un pueblo A perteneciente a cierta mancomunidad se realiza un muestreo para determinar ciertas características de la zona. Concretamente interesa estudiar la proporción de mujeres enfermas de gripe y el número total de hombres. Se sabe que en el pueblo, que consta de 50 manzanas, hay en total unas 2000 mujeres. Se escogen por muestreo aleatorio simple 6 manzanas y se cuentan las mujeres, hombres, y número de mujeres enfermas.

Manzana	Mujeres enfermas	Hombres	Mujeres
1	2	37	30
2	1	26	20
3	6	99	100
4	2	36	25
5	3	56	50
6	4	71	65

a) Estimar la proporción de mujeres enfermas en el pueblo y dar un intervalo de confianza al 95% para la estimación suponiendo normalidad del estimador.

b) Estimar el número total de hombres en el pueblo comparando el estimador básico de muestreo aleatorio simple con los dos estimadores usuales de estimación indirecta (presentar un I.C. al 95% para cada método). Explicar las diferencias entre los tres métodos de estimación utilizando medidas de precisión, descriptivas y gráficos.

c) Se dispone de datos tomados del conjunto de pequeños pueblos restantes de la mancomunidad a la que pertenece el pueblo A. En total hay en esos pueblos unas 30 manzanas, con un total de 1500 mujeres. Se toma una muestra de 5 manzanas y se observan, en esos datos, los siguientes estadísticos: media muestral de mujeres enfermas por manzana: 1. Cuasivarianza muestral de mujeres enfermas por manzana:3. media muestral de mujeres por manzana: 55. Cuasivarianza muestral de mujeres por manzana: 900. Cuasicovarianza entre mujeres y mujeres enfermas=21. Estimar por dos métodos de estimación indirecta la proporción de mujeres enfermas en toda la mancomunidad (incluyendo el pueblo A), calcular la varianza estimada de los dos estimadores y explicar los resultados. Decidir cual sería , de los dos, el mejor estimador.

---

a) Definiendo y como la variable número de mujeres enfermas y x como el número de mujeres, la proporción se estimará por la razón muestral, $\widehat{R}=\dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}}=\dfrac{3}{48.33}=0.062$.

ISBN 9788496866133

Además, $s_y^2 = 3.2$, $s_x^2 = 926.6$, y $s_{xy} = 54$, con lo que $r_{xy} = 0.99$ y está justificado el uso del estimador de la razón (otra manera de estimar esa proporción era, como se vio estudiando m.a.s., estimar el total de enfermas por $N\widehat{\overline{y}}$ y dividirlo por el total de mujeres, que es 2000. Pero en este caso es más adecuado utilizar el estimador de la razón).

La varianza estimada del estimador de la razón es, teniendo en cuenta también que la media poblacional del número de mujrese por manzana es $\overline{x} = \dfrac{2000}{50} = 40$:

$$\widehat{V}(\widehat{R}) = \frac{N-n}{Nn\overline{x}^2}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}) = \frac{50-6}{50\cdot 6\cdot 40^2}(3.2 + 0.062^2\cdot 926.6 - 2\cdot 0.062\cdot 54) = 0.000006.$$

El intervalo de confianza al 95% es

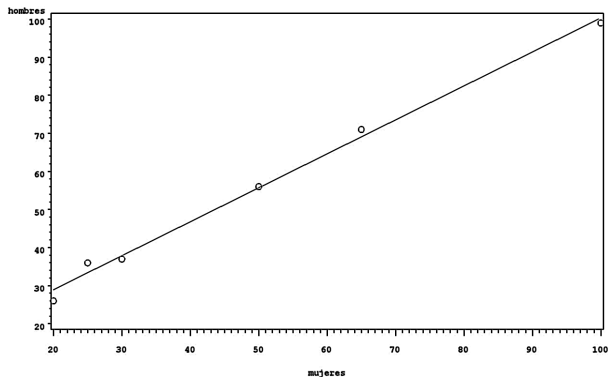
$$(0.062 - 1.96\sqrt{0.000006}, 0.062 - 1.96\sqrt{0.000006}) = (0.057, 0.066).$$

b) Definiendo y como el número de hombres y x como el número de mujeres, la estimación de expansión del total de hombres es $N\widehat{\overline{y}} = 50\cdot 54.16 = 2708.33$.

Para comparar esta estimación con la estimación de razón, teniendo en cuenta que $s_y^2 = 742.96$, $s_x^2 = 926.6$, y $s_{xy} = 827.33$, entonces se tiene que $r_{xy} = \dfrac{s_{xy}}{s_x s_y} = 0.997$ y por lo tanto, al ser tan alto el coeficiente de correlación está justificada la estimación de razón o regresión. Se puede de todos modos calcular $\dfrac{\widehat{CV}(x)}{\widehat{CV}(y)} = \dfrac{s_x/\widehat{\overline{x}}}{s_y/\widehat{\overline{y}}} = \dfrac{0.503}{0.63} = 0.80$ y por lo tanto $r_{xy} > \dfrac{1}{2}\dfrac{\widehat{CV}(x)}{\widehat{CV}(y)} = 0.40$ lo que justifica la no utilización del estimador de expansión en este caso.

La estimación de razón del total es $N\overline{y}_R = N\dfrac{\widehat{\overline{y}}}{\widehat{\overline{x}}}\overline{x} = 2241.38$.

Por lo que respecta a escoger entre estimación de razón o regresión, el coeficiente de regresión es $\widehat{b} = \dfrac{s_{xy}}{s_x^2} = 0.89$ y el estimador de la razón es $\widehat{R} = 1.12$ (son suficientemente diferentes para que las varianzas de los estimadores, que sólo difieren en ese término, sean también diferentes, siempre a favor de la estimación de regresión). Por otra parte, la constante estimada de la recta de regresión es $\widehat{a} = \widehat{\overline{y}} - \widehat{b}\widehat{\overline{x}} = 11.14$ lo que parece suficientemente grande en magnitud. El contraste de hipótesis en un modelo de regresión sobre la constante a rechaza la hipótesis de diferencia de cero. Se presenta la nube de puntos de la muestra en la figura siguiente, donde se aprecia que no se podría considerar cero la constante de regresión:



Nube de hombres por mujeres

Como $\overline{y}_{reg} = \widehat{\overline{y}} + \widehat{b}(\overline{x} - \widehat{\overline{x}}) = 46.72$, La estimación de regresión del total es $N\overline{y}_{reg} = 2336$.

ISBN 9788496866133

c) Llamando, como en el apartado a), y a las mujeres enfermas e x a las mujeres, se tiene que $s_x^2 = 900$, $s_y^2 = 3$, $s_{xy} = 21$, $\overline{x} = 50$, $\widehat{\overline{x}} = 55$, e $\widehat{\overline{y}} = 1$. Si se utiliza el tipo de estimación del apartado a), la estimación de la proporción en ese conjunto de pequeños pueblos se hará a través de la razón muestral $\widehat{R} = 0.018$. La varianza de este estimador es

$$\widehat{V}(\widehat{R}) = \frac{N-n}{Nn\overline{x}^2}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}) = 0.000169.$$

Para obtener una estimación conjunta de los dos estratos (el pueblo A y los restantes pueblos) se puede optar por estimación separada o combinada. Para el caso de la estimación separada, se tiene que el estimador de la razón es

$$R_s = \frac{t_s}{\overline{x}} = \frac{1}{\overline{x}} \sum_{h=1}^{L} \frac{N_h}{N} \widehat{R}_h \overline{x}_h = \frac{1}{43.75}\left(\frac{50}{80}0.062 \cdot 40 + \frac{30}{80}0.018 \cdot 50\right) = 0.043.$$

pues $\overline{x} = \dfrac{3500}{80} = 43.75$.

Y su varianza estimada será

$$\widehat{V}\left(\frac{1}{\overline{x}} \sum_{h=1}^{L} \frac{N_h}{N} \widehat{R}_h \overline{x}_h\right) = \sum_{h=1}^{L} (\frac{N_h}{N}\frac{\overline{x}_h}{\overline{x}})^2 \widehat{V}(\widehat{R}_h) = (\frac{50}{80}\frac{40}{43.75})^2 0.000006 + (\frac{30}{80}\frac{50}{43.75})^2 0.000169 = 0.000026.$$

Si se opta por estimación combinada, el estimador es

$$\widehat{R}_c = \frac{\overline{y}_{st}}{\overline{x}_{st}} = \frac{\frac{50}{80}3 + \frac{30}{80}1}{\frac{50}{80}48.33 + \frac{30}{80}55} = 0.0442.$$

Y su varianza estimada,

$$V(\widehat{R}_c) = \frac{1}{\overline{x}^2} \sum_{h=1}^{L} W_h^2 \frac{(N_h - n_h)}{N_h n_h}(s_{yh}^2 + \widehat{R}_c^2 s_{xh}^2 - 2\widehat{R}_c s_{xyh}) =$$

$$= \frac{1}{43.75^2}(\frac{5}{8})^2 \frac{(50-6)}{50 \cdot 6}(3.2 + 0.044^2 \cdot 926.6 - 2 \cdot 0.044 \cdot 54) +$$

$$+ \frac{1}{43.75^2}(\frac{3}{8})^2 \frac{(40-5)}{40 \cdot 5}(3 + 0.044^2 \cdot 900 - 2 \cdot 0.044 \cdot 21) = 0.0000444.$$

En general, si hay diferencias en las razones por estrato y hay suficientes observaciones por estrato la estimación separada está más justificada que la estimación combinada, como es en este caso.

Ejercicio 6.8

El año 2000 se realizó un examen de literatura a los estudiantes de dos clases de 40 alumnos cada una, obteniendo una media de 6.5. Al año siguiente se les hizo otro examen de literatura, además de otro de matemáticas, sobre los conocimientos adquiridos en ese curso académico. Corregidos 10 exámenes seleccionados aleatoriamente entre los 80 alumnos y cotejando el resultado del año 2001 con el obtenido por los mismos alumnos en el año 2000, se obtuvo la siguiente tabla:

2000 (literatura)	5	4	6	7	7	5.5	4	8	3	6	9
2001 (literatura)	4	3.3	4.8	5.6	5.7	4.4	3.2	6.6	2.5	4.8	7.2
2001 (matemáticas)	4.7	4.0	5.5	6.3	6.3	5.3	4.4	7.3	3.4	5.7	7.9

ISBN 9788496866133

Se trata de realizar las siguientes estimaciones sabiendo que solamente se conoce el resultado de los 80 alumnos del año 2000 en literatura.

a) La mejor estimación posible de la media de 2001 del examen de literatura .

b) La mejor estimación posible de la media de 2001 del examen de matemáticas.

---

a) Al haberse realizado muestreo aleatorio simple, se puede optar por estimación de expansión, de razón o regresión. La variable auxiliar es obviamente la nota obtenida en 2000 en literatura, de la que se conoce la media poblacional $\overline{x} = 6.5$.

Se ha visto que si el coeficiente de correlación cumple $\rho_{xy} > \frac{1}{2}\frac{CV(x)}{CV(y)}$ la estimación de razón supera a la de expansión en términos de varianza. Se puede comprobar también que la varianza estimada será además menor en estimación de razón que en estimación de expansión si $r_{xy} > \frac{1}{2}\frac{\widehat{CV(x)}}{\widehat{CV(y)}}$ donde $\widehat{CV}(x) = s_x/\widehat{\overline{x}}$ y $\widehat{CV}(y) = s_y/\widehat{\overline{y}}$.

En nuestro caso, se tiene que $\widehat{\overline{x}} = 5.86$, $\widehat{\overline{y}} = 4.73$, $s_x^2 = 3.30$, $s_y^2 = 2.13$, $s_{xy} = 2.65$ y por lo tanto $r_{xy} = \frac{s_{xy}}{s_x s_y} = 0.998$, $\widehat{CV}(x) = s_x/\widehat{\overline{x}} = 0.31$ y $\widehat{CV}(y) = s_x/\widehat{\overline{y}} = 0.306$ y entonces $\frac{1}{2}\frac{\widehat{CV(x)}}{\widehat{CV(y)}} = 0.5$ con lo cual es evidente que al menos la estimación de razón será superior a la de expansión.

Falta comparar la estimación de razón con la obtenida por regresión. Se preferirá razón cuando la recta de regresión pase aproximadamente por el origen. En este caso, con estimación de regresión se tiene que $\widehat{b} = \frac{s_{xy}}{s_x^2} = 0.80$.La estimación de la constante en la recta de regresión es $a = \widehat{\overline{y}} - \widehat{b\overline{x}} = 0.042$ que es pequeño en magnitud respecto a $\widehat{b\overline{x}}$. El contraste de hipótesis sobre la significatividad de este coeficiente da que este no es es significativamente distinto de cero

Además,

$$\widehat{V}(\overline{y}_R) = \frac{N-n}{Nn}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}) = 0.00042$$

y

$$\widehat{V}(\overline{y}_{reg}) = \frac{N-n}{Nn}(s_y^2 + \widehat{b}^2 s_x^2 - 2\widehat{b}s_{xy}) = 0.00041,$$ similar al de razón,

con lo cual se preferirá el método de razón al de regresión y también frente a la estimación de expansión. Las tres estimaciones son $\widehat{\overline{y}} = 4.73$, $\overline{y}_R = 5.25$, y $\overline{y}_{reg} = 5.24$.

b) En este caso, $\widehat{\overline{y}} = 5.52$, $s_y^2 = 1.88$, $s_{xy} = 2.48$ y por lo tanto $r_{xy} = \frac{s_{xy}}{s_x s_y} = 0.995$ y $\widehat{CV}(y) = s_x/\widehat{\overline{y}} = 0.248$ y entonces $\frac{1}{2}\frac{\widehat{CV(x)}}{\widehat{CV(y)}} = 0.62$ con lo cual la estimación de razón será superior en precisión a la de expansión.

La estimación de la constante en la recta de regresión es $\widehat{a} = \widehat{\overline{y}} - \widehat{b\overline{x}} = 0.832$. El contraste de hipótesis sobre la significatividad de este coeficiente da como resultado que es significativamente distinto de cero.

Además,

$$\widehat{V}(\overline{y}_R) = \frac{N-n}{Nn}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}) = 0.012$$

y

$$\widehat{V}(\overline{y}_{reg}) = \frac{N-n}{Nn}(s_y^2 + \widehat{b}^2 s_x^2 - 2\widehat{b}s_{xy}) = 0.0016,$$

ISBN 9788496866133

que es 7 veces menor que la varianza del estimador de razón, con lo cual se preferirá el método de regresión y también frente a la estimación de expansión. Las tres estimaciones son $\widehat{\overline{y}} = 5.52$, $\overline{y}_R = 6.12$, y $\overline{y}_{reg} = 6.00$.

Ejercicio 6.9

Se dispone de la siguiente muestra obtenida mediante m.a.s. en un lote del peso en gramos y longitud en mm de ciertos tornillos obtenidos en un proceso de fabricación. Se sabe que el peso total de los 100 tornillos en el lote es de 1.5 kg.

Peso	12	14	15	13	16
longitud	15	17	18	14	19

a) Suponiendo la muestra suficientemente precisa y estimación de razón, decir cuántos tornillos habrá que muestrear para obtener un error de muestreo de 0.20 en la estimación de la longitud media.

b) Realizar el apartado a suponiendo estimación de regresión.

c) Realizar el apartado a) suponiendo m.a.s.

---

a) La varianza del estimador de razón de la media es:

$$V(\overline{y}_R) = \frac{N-n}{Nn}(S_y^2 + R^2 S_x^2 - 2RS_{xy}).$$

Si se supone la muestra suficientemente precisa, las estimaciones de las cuasivarianzas y cuasi covarianza son:

$S_y^2 \simeq s_y^2 = 4.3$

$S_x^2 \simeq s_x^2 = 2.5$

$S_{xy} \simeq s_{xy} = 3.$

Y la estimación de la razón es $\widehat{R} = 1.18$.

Para que el error de muestreo sea 0.20, ha de ser $\sqrt{V(\overline{y}_R)} = 0.2$ y por lo tanto

$0.04 = V(\overline{y}_R) = \frac{N-n}{Nn}(S_y^2 + R^2 S_x^2 - 2RS_{xy})$ y entonces,

$$0.04 = \frac{N-n}{Nn}(S_y^2 + R^2 S_x^2 - 2RS_{xy}) \simeq \frac{100-n}{100n}(4.3 + 1.18^2 \cdot 2.5 - 2 \cdot 1.18 \cdot 3)$$

y por lo tanto,

$$\frac{100-n}{100n} \cdot 0.701 = 0.04.$$

Así,

$$n(100 \cdot 0.04 + 0.701) = 100 \cdot 0.701 \text{ y } n = \frac{100 \cdot 0.701}{(100 \cdot 0.04 + 0.701)} = 14.9.$$

Por lo tanto habría que muestrear 15 tornillos.

ISBN 9788496866133

b) Como $\widehat{b} = \frac{s_{xy}}{s_x^2} = 1.2$, y la varianza del estimador de regresión es

$$V(\overline{y}_{reg}) = \frac{N-n}{Nn}(S_y^2 + b^2 S_x^2 - 2bS_{xy}),$$

se tiene que

$$0.04 = \frac{100-n}{100n}(4.3 + 1.2^2 \cdot 2.5 - 2 \cdot 1.2 \cdot 3) = \frac{100-n}{100n}(0.7)\ .$$

y por lo tanto se necesitará el mismo tamaño muestral $n = 14$.

c) Suponiendo m.a.s., se tiene que

$0.04 = \frac{100-n}{100n} S_y^2 \simeq \frac{100-n}{100n} 4.3$ y entonces

$$n = \frac{100 \cdot 4.3}{100 \cdot 0.04 + 4.3} = 51.8.$$

Bajo m.a.s., habría que seleccionar 52 tornillos para obtener ese nivel de precisión.

Ejercicio 6.10

En una ciudad interesa estudiar la proporción de perros machos de una determinada raza y el número total de perros hembra. Se sabe que en la ciudad hay aproximadamente 30 perreras, y hay en total unos 1000 perros machos. Se escogen por muestreo aleatorio simple 5 perreras y se cuentan los perros macho, los perros hembra, y número de perros macho de la raza en cuestión.

Perrera	Perros Macho Raza	Perros macho	Perros hembra
1	4	41	60
2	2	50	40
3	12	180	120
4	4	65	50
5	6	100	100

Nota: Si se define por

a= Perros Macho raza , b= Perros Macho, y c= Perros Hembra,

se tiene que

$$\sum_{i=1}^{5} a_i^2 = 216;\ \sum_{i=1}^{5} b_i^2 = 50806;\ \sum_{i=1}^{5} c_i^2 = 32100;\ \sum_{i=1}^{5} a_i b_i = 3284;\ \sum_{i=1}^{5} a_i c_i = 2560;\ \sum_{i=1}^{5} b_i c_i = 39310.$$

a) Estimar la proporción de perros machos de la determinada raza sobre los perros macho y dar un intervalo de confianza al 95% para la estimación suponiendo normalidad del estimador.

ISBN 9788496866133

b) Estimar el número total de perros hembra en el pueblo con un estimador de estimación indirecta . Comparar su varianza con la obtenida con el estimador usual de muestreo aleatorio simple.

---

a) El número de perros macho en cada perrera es una variable aleatoria, por lo que se utilizará el estimador de razón para estimar la proporción de perros macho de esa raza (suponiendo, como se comprobará, que la correlación entre perros macho de raza y perros macho es alta).

Definiendo por $y =$ Perros macho de raza, y $x =$ Perros macho, se tiene que el estimador de la proporción estará definido por

$$\widehat{R} = \frac{\widehat{\overline{y}}}{\widehat{\overline{x}}} = \frac{5.6}{87.20} = 0.064.$$

Para calcular la varianza, es necesario calcular s_y^2, s_x^2 y s_{xy}.

$$s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\widehat{\overline{y}}^2\right) = \frac{1}{5-1}\left(216 - 5 \cdot 5.6^2\right) = 14.8.$$

$$s_x^2 = \frac{1}{5-1}\left(50806 - 5 \cdot 87.2^2\right) = 3196.7.$$

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\widehat{\overline{x}}\widehat{\overline{y}}\right) = \frac{1}{5-1}\left(3284 - 5 \cdot 87.2 \cdot 5.6\right) = 210.6.$$

También se conoce la media poblacional $\overline{x} = \dfrac{1000}{30} = 33.33$.

La varianza estimada del estimador es entonces:

$$\widehat{V}(\widehat{R}) = \frac{N-n}{Nn\overline{x}^2}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}) = \frac{30-5}{30 \cdot 5 \cdot 33.33^2}(14.8 + 0.064 \cdot 3196.7 - 2 \cdot 0.064 \cdot 210.6) = 0.00014.$$

El intervalo de confianza al 95% para esa proporción será

$$(0.064 - 1.96\sqrt{0.00014}, 0.064 + 1.96\sqrt{0.00014}) = (0.041, 0.087).$$

b) Se utilizará como variable auxiliar la variable perros macho, pues es de la que se dispone de la información sobre la media poblacional.

Se comenzará por realizar estimación de regresión, descartándola a favor de la estimación de razón si hay razones suficientes para suponer que la recta pasa por el origen.

En este caso, utilizando expresiones similares a las del apartado a), y llamando $x =$Perros macho e $y =$ Perros hembra, se obtiene:

$s_y^2 = 1180$

$s_x^2 = 3196.7$

$s_{xy} = 1761.5.$

El estimador de la constante de regresión es $\widehat{b} = \dfrac{s_{xy}}{s_x^2} = \dfrac{1761.5}{3196.7} = 0.55.$

El estimador de regresión del total es $N\overline{y}_{reg}$, con

$$\overline{y}_{reg} = \widehat{\overline{y}} + \widehat{b}(\overline{x} - \widehat{\overline{x}}) = 74 + 0.55(33.33 - 87.20) = 44.3.$$

Así, $N\overline{y}_{reg} = 30 \cdot 44.3 = 1329.$

ISBN 9788496866133

Por otra parte, $\widehat{a} = \widehat{\overline{y}} - \widehat{b}\widehat{\overline{x}} = 26.04$ que parece grande en magnitud (un contraste de hipótesis sobre el parámetro en un modelo de regresión clásico da significativamente distinto de 0). El estimador de la razón $\widehat{R} = 0.84$, bastante diferente de $\widehat{b}$ en magnitud, con lo cual la varianza de ambos estimadores diferirá bastante (ambas varianzas se diferencian en que el término R está en lugar del término b y todos los demás términos son similares). La diferencia siempre es a favor del estimador de regresión, pues se demuestra que tiene menor varianza que el de razón.

Por estos motivos se prefiere en este caso la estimación de regresión a la de razón. Realizar el gráfico de la nube de puntos siempre es posible, además, como un medio gráfico de observar que la recta de regresión no pasa por el origen.

La varianza estimada del estimador es $N^2\widehat{V}(\overline{y}_{reg})$, con

$$\widehat{V}(\overline{y}_{reg}) = \frac{N-n}{Nn}(s_y^2 + \widehat{b}^2 s_x^2 - 2\widehat{b}s_{xy}) = \frac{30-5}{30\cdot 5}(1180 + 0.55^2 \cdot 3196.7 - 2\cdot 0.55 \cdot 1761.5) = 34.89.$$

y

$N^2\widehat{V}(\overline{y}_{reg}) = 31401.$

El intervalo de confianza al 95% para el total de perros hembra es

$(1329 - 1.96\sqrt{31401}, 1329 + 1.96\sqrt{31401}) = (981.7, 1676.3).$

Ejercicio 6.11

El archivo SAS guisa contiene datos de 51 provincias españolas con las variables superficie dedicada al cultivo del guisante en 1998, en Hectáreas, (super) y la producción de guisantes en toneladas métricas (produ).

Se utilizará como variable de interés la producción de guisantes y como variable auxiliar la superficie. Se desea estudiar la estimación de la media de producción de guisante por provincia.

a) Se desea comprobar la precisión global del estimador de razón y de regresión con el de expansión, bajo m.a.s. Calcular, con la ayuda de los procedimientos del SAS, las varianzas exactas de cada uno de los tres estimadores, asumiendo $n = 10$.

b) Utilizando el archivo poblacional, dibujar la nube de puntos de la producción con la superficie.

c) Extraer 5 muestras de tamaño $n = 10$ con las semillas 1234, 1235, 1236, 1237, 1238 y comprobar el valor de los estimadores de razón, regresión y m.a.s. (utilizar la macro estimrazreg y el proc means para la media muestral). Se supone conocida la media poblacional de la superficie, 1131.

---

a) Utilizando el proc corr del SAS, se obtiene la matriz de varianzas y covarianzas, así como las medias y coeficiente de correlación entre las dos variables:

```
proc corr data=guisa cov;
var produ super;
run;
```


ISBN 9788496866133

Así, $S_y^2 = 12004900.12$, $S_x^2 = 6910396.79$, y $S_{xy} = 8959207.48$. Además, $\overline{y} = 1454$ y $\overline{x} = 1131$. El coeficiente de correlación poblacional es $\rho_{xy} = 0.98$. La razón poblacional es $R = \dfrac{\overline{y}}{\overline{x}} = 1.285$.

La varianza exacta del estimador de expansión, con $n = 10$, es:

$$V(\widehat{\overline{y}}) = \frac{N-n}{Nn} S_y^2 = \frac{51-10}{51 \cdot 10} 12004900.12 = 965100.$$

La varianza exacta del estimador de razón es:

$$V(\overline{y}_R) = \frac{N-n}{Nn}(S_y^2 + R^2 S_x^2 - 2RS_{xy}) = \frac{51-10}{51 \cdot 10}(12004900.12 + 1.285^2 \cdot 6910396.79 - 2 \cdot 1.285 \cdot 8959207.48) = 30534.6$$

En el caso del estimador de regresión, se tiene que $b = \dfrac{S_{xy}}{S_x^2} = 1.29$.

Así,

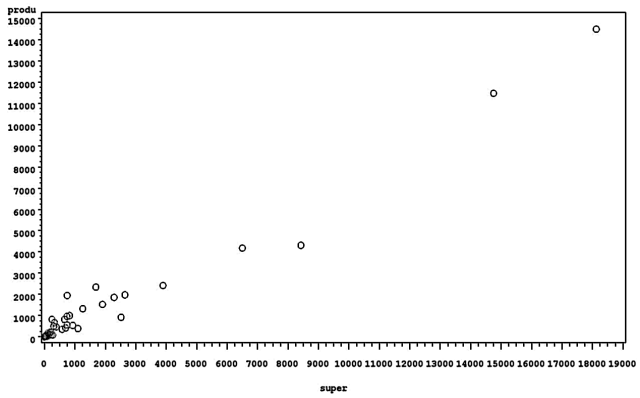
$$V(\overline{y}_{reg}) = \frac{N-n}{Nn}(S_y^2 + b^2 S_x^2 - 2bS_{xy}) = \frac{51-10}{51 \cdot 10}(12004900.12 + 1.29^2 \cdot 6910396.79 - 2 \cdot 1.29 \cdot 8959207.48) \simeq 30534$$

con lo cual en este caso la estimación de regresión no es necesaria, al no apreciarse diferencias sustanciales con la de razón.

b) El programa utilizado es

```
symbol v=circle i=none c=black;
proc gplot data=guisa;plot produ*super;run;
```

dando lugar al gráfico:



Producción por superficie

c) El programa básico es

```
proc surveyselect data=guisa method=srs out=muestra seed=1234 sampsize=10 noprint;
run;
proc means data=muestra;var produ;run;
%estimrazreg(muestra,produ,super,1454,51,10);
```


ISBN 9788496866133

Variando la semilla para las cinco muestras, se obtienen los siguientes resultados:

$\widehat{\overline{y}}$	$\overline{y}_R$	$\overline{y}_{reg}$
1048	980	996
3251	1230	1396
2153	1223	1274
466	1328	1157
788	1163	1079

donde se observa la mayor variabilidad del estimador de expansión, respecto a los de razón y regresión.

Ejercicio 6.12

El archivo SAS parcelas contiene datos sobre la producción de frutas y hortalizas en 200 parcelas, situadas ambas en dos tipos de terreno diferente de 100 parcelas cada terreno. Se desea estudiar la estimación de la producción media de hortalizas conociendo la producción media de frutas en las 200 parcelas.

a) Calcular con el proc means la media poblacional de producción de hortalizas y frutas por parcela en conjunto, y la media de frutas por estrato.

b) Extraer una m.a.s. estratificada con semilla 12345 de 10 parcelas en cada terreno y realizar la estimación separada y combinada por razón y regresión de la producción media de hortalizas por parcela, utilizando la macro estimrazestrat.

c) Realizar el mismo proceso con muestras de tamaño 50 en cada terreno y la misma semilla.

d) A la vista de la muestra de tamaño $n = 10$, estudiar si es mejor realizar estimación separada con razón en un estrato y regresión en el otro, utilizando la macro estimrazreg para cada estrato. Aplicar el caso a la estimación en la muestra obtenida de tamaño 10 por estrato.

---

a) Realizando el proc means, y llamando x a la producción de fruta e y a la de hortalizas:

```
proc means data=parcelas;run;
```

se obtiene $\overline{x} = 24.09$ e $\overline{y} = 23.82$.

Para obtener los datos por estrato, se ejecuta

```
proc means data=parcelas;by estrato;run;
```

dando $\overline{x}_1 = 26.44$ e $\overline{x}_2 = 21.75$.

b) Para poder utilizar posteriormente la macro, se añade, en el archivo poblacional, la variable de tamaño (el tamaño es 100 en cada uno de los dos estratos) y la media de x a cada estrato:

ISBN 9788496866133

```
data parcelas;set parcelas;if estrato=1 then mediax=26.44;else mediax=21.75;tama=100;
```

Se extrae la muestra y se utiliza la macro:

```
proc surveyselect data=parcelas method=srs n=10 out=muestra seed=12345;
strata estrato;
run;
%estimrazestrat(muestra,.,hortalizas,frutas,estrato,tama,mediax,24.09,1,200);
```

Obteniendo los siguientes resultados:

Estimación	Estimador	Varianza estimada
Razón separada	24.14	0.50
Razón combinada	24.18	0.59
Regresión separada	23.91	0.10
Regresión combinada	24.21	0.38

c) Con $n = 50$ y la misma semilla, se obtiene la siguiente tabla:

Estimación	Estimador	Varianza estimada
Razón separada	23.94	0.05
Razón combinada	23.85	0.01
Regresión separada	23.95	0.06
Regresión combinada	23.95	0.04

Se observa como aumenta la precisión al aumentar el tamaño muestral (recordemos que el valor real de $\overline{y}$ es 23.82).

d) Para saber si en algún estrato es conveniente no utilizar estimación de regresión (pues la recta pasa por el origen y en ese caso conviene mejor utilizar el modelo más sencillo de razón), veamos con detalle los datos obtenidos por estrato con la muestra de tamaño 10:

```
proc corr data=muestra cov;by estrato;
var frutas hortalizas;
run;
```

Se obtiene para el primer estrato:

$\widehat{\overline{x}} = 27.7, \widehat{\overline{y}} = 28.8, s_x^2 = 94.6$, $s_y^2 = 103.5$, $s_{xy} = 98.8$, y $r_{xy} = 0.99$ .Entonces,

$$\widehat{b} = \frac{s_{xy}}{s_x^2} = 1.04,$$

$$\widehat{a} = \widehat{\overline{y}} - \widehat{b}\widehat{\overline{x}} = -0.13$$

ISBN 9788496866133

y

$$\widehat{R} = \frac{\widehat{\overline{y}}}{\widehat{\overline{x}}} = 1.039.$$

con lo cual en el primer estrato, al ser $\widehat{b}$ y $\widehat{R}$ muy parecidos, y $\widehat{a}$ despreciable en términos de la magnitud del valor de $\widehat{\overline{y}}$, la recta de regresión pasa aproximadamente por el origen y es más adecuado utilizar estimación de razón.

En el segundo estrato:

$\widehat{\overline{x}} = 20.7, \widehat{\overline{y}} = 19.8, s_x^2 = 95.7$, $s_y^2 = 30.8$, $s_{xy} = 50.48$, y $r_{xy} = 0.92$.

Entonces,

$$\widehat{b} = \frac{s_{xy}}{s_x^2} = 0.52 \ ,$$

$$\widehat{a} = \widehat{\overline{y}} - \widehat{b}\widehat{\overline{x}} = 8.8$$

y

$$\widehat{R} = \frac{\widehat{\overline{y}}}{\widehat{\overline{x}}} = 0.95.$$

Al ser $\widehat{b}$ y $\widehat{R}$ my diferentes, y $\widehat{a}$ claramente no ignorable en términos de la magnitud de $\widehat{\overline{y}}$, en este segundo estrato es conveniente realizar estimación de regresión.

Para construir el estimador conjunto, se realiza calcula el estimador de razón y su varianza en el primer estrato, y el estimador de regresión y varianza en el segundo, para después unirlos con los pesos respectivos.

La estimación se realiza con la macro estimrazreg, creando antes un archivo para cada estrato:

```
data estrato1 estrato2;
set parcelas;
if estrato=1 then output estrato1;
else output estrato2;
run;
%estimrazreg(estrato1,hortalizas,frutas,26.44,100,10);
%estimrazreg(estrato2,hortalizas,frutas,21.75,100,10);
```

De la primera macro se obtiene $\overline{y}_R = 27.26$ con $\widehat{V}(\overline{y}_R) = 0.11$ y de la segunda, $\overline{y}_{reg} = 20.39$ con $\widehat{V}(\overline{y}_{reg}) = 0.36$.

El estimador separado se construye mediante

$$\widehat{\overline{y}}^* = \frac{N_1}{N}\overline{y}_R + \frac{N_2}{N}\overline{y}_{reg} = 0.5 \cdot 27.26 + 0.5 \cdot 20.39 = 23.82.$$

Su varianza estimada es

$$\widehat{V}(\widehat{\overline{y}}^*) = (\frac{N_1}{N})^2\widehat{V}(\overline{y}_R) + (\frac{N_2}{N})^2\widehat{V}(\overline{y}_{reg}) = 0.1175.$$

Se observa cómo en este caso, la estimación da prácticamente el valor real de la media poblacional, que aunque ha sido casualidad, es explicado también por haber "afinado" en la construcción del estimador.

ISBN 9788496866133

7.6 Ejercicios propuestos

1) De una población de 40 hogares se obtiene una m.a.s. de 4 hogares, que proporcionan los siguientes valores anuales en euros:

Gastos en alimentación	Gastos totales
6000	12000
5900	10600
3400	7400
7000	18000

a) Estimar el porcentaje de gasto en alimentación

b) ¿Se puede decir que el sesgo es despreciable?

2) Una compañía de seguros desea estimar el gasto medio por asegurado en que se incurre en los asegurados que han tenido algún percance. Para ello se escogen por m.a.s. 8 asegurados. Se sabe que hay 600 asegurados que han tenido algún incidente, y que el número de percances entre esta subpoblación de asegurados asciende a 900. Se obtiene, para cada asegurado muestreado, el gasto total que le ha originado a la compañía y el número de percances que ha tenido:

Asegurado	1	2	3	4	5	6	7	8
nº incidentes	1	2	1	3	1	2	3	1
gasto total	120	350	80	700	90	240	850	80

a) Basándose en estos datos, que método de muestreo (describir el estimador) y qué tamaño muestral recomendarías para obtener un error máximo de muestreo de 60 para estimar la media de gasto por asegurado?.

b) ¿Si se desea estimar el gasto por incidente, qué tamaño muestral mínimo se ha de tomar para obtener un error de muestreo de 50?

3) En una ciudad que contiene 15000 viviendas se ha tomado una m.a.s. de 600 viviendas. En cada una de ellas se ha observado el número de personas P_i y el número de habitaciones H_i obteniéndose los siguientes resultados:

$$\sum_{i=1}^{600} P_i = 2946,\ \sum_{i=1}^{600} H_i = 2150,\ \sum_{i=1}^{600} P_i^2 = 18694,\ \sum_{i=1}^{600} H_i^2 = 10997,\ \sum_{i=1}^{600} P_i H_i = 12800.$$

a) Estimar el número medio de personas por habitación, utilizando el estimador de razón.

b) ¿Es influyente el sesgo en la estimación?

ISBN 9788496866133

c) Realizar la estimación por intervalos al 95% de confianza comentando la precisión obtenida.

d) Suponiendo conocido el número total de personas en la ciudad, se desea averiguar el total de habitaciones en la misma mediante el método indirecto basado en la razón. ¿Resulta apropiada dicha estimación frente a la basada en la media por unidad?.

4) El rectorado de la Universidad Complutense de Madrid ha decidido hacer un estudio para saber cuantas horas de baja han necesitado sus profesores durante el curso escolar. Los profesores de la UCM pueden elegir entre usar la Seguridad Social (seguro A) y un seguro privado (seguro B). Se sabe que el número de profesores que utilizan el seguro A es 1000, mientras que los que utilizan el seguro B son 1500. Los profesores del primer grupo acumularon un total de 16300 horas de baja en el año anterior, mientras que los del segundo sólo 12800. Se extrae una muestra estratificada de 20 profesores obteniéndose los siguientes resultados:

Seguro A			Seguro B		
Profesor	X	Y	Profesor	X	Y
1	12	13	1	10	8
2	24	25	2	8	0
3	15	15	3	0	4
4	30	32	4	14	6
5	32	36	5	12	10
6	26	24	6	6	0
7	10	12	7	4	2
8	15	16	8	0	4
9	0	2	9	8	4
10	14	12	10	10	8

donde X = Número de horas perdidas el año anterior e Y = Número de horas perdidas este año.

Se pide:

a) Estimar el número medio de horas perdidas por los afiliados a la Seguridad Social y su error de muestreo mediante estimación de razón. Hacer lo mismo para la compañía B.

b) Estimar el número medio de horas perdidas en la población total mediante el estimador de razón separado, combinado, regresión separada y regresión combinada. Comparar resultados.

ISBN 9788496866133

5) Un investigador desea estimar el efecto de la creación de una fábrica en un entorno rural. Para ello escoge un terreno no cultivado significativo en el entorno de la fábrica, lo divide en 40 parcelas, escoge 10 por muestreo sistemático con arranque aleatorio, y cuenta el número de especies vegetales diferentes en cada parcela antes de que la fábrica comience a funcionar en el año 2000. Tres años después de ese comienzo, vuelve a contar en cada parcela el número de especies diferentes. Sabiendo que en las 40 parcelas la suma de las especies diferentes contadas en cada parcela es 440 (que no es lo mismo que el número de especies "diferentes" en todas las parcelas, pues las contadas por parcela están repetidas en unas parcelas y otras),

a) Estimar la disminución en porcentaje del número de especies vegetales en el entorno de la fábrica supuestamente debida a su presencia.

b) ¿Es relevante el sesgo del estimador?.

c) Estimar adecuadamente el número promedio por parcela de especies en la actualidad.

d) ¿Mejora el estimador del apartado anterior al estimador de expansión (media muestral)?.

Datos:

Parcela muestreada	A=n^o especies en 2000	B=n^o especies en 2003
1	6	2
2	15	6
3	10	4
...	...	...
10	7	2

Con $\sum A = 153$, $\sum B = 59$, $\sum A^2 = 2763$, $\sum B^2 = 417$ y $\sum AB = 1067$.

6) En un estudio realizado en el Parque Nacional de Ordesa, se desea investigar la edad media de los pinos de un pinar. La edad exacta se puede conocer cortando el tronco y contando los anillos, pero esto lleva más tiempo y es poco ecológico. Como el diámetro del tronco está relacionado con el número de anillos, se mide el diámetro de los 500 árboles del pinar y se obtiene que el diámetro promedio es de 50 centímetros. A continuación se seleccionan 10 pinos por muestreo sistemático con arranque aleatorio y se cortan, midiendo su edad en meses a partir de los anillos. Se obtienen los siguientes datos, donde D es el diámetro y E es la edad:

$$\sum_{i=1}^{10} E_i = 2291, \sum_{i=1}^{10} D_i = 521, \sum_{i=1}^{10} E_i^2 = 545075, \sum_{i=1}^{10} D_i^2 = 28543, \sum_{i=1}^{10} D_i E_i = 124611.$$

Suponiendo que el orden de los árboles en el pinar no tiene relación con su edad, calcular tres diferentes estimadores de la media de edad de los árboles del pinar, y compararlos, estimando sus varianzas y justificando bien la respuesta.

7) Para un estudio sobre dietas en un pueblo se escogen 8 hombres por m.a.s. y se mide su peso

ISBN 9788496866133

antes de aplicar el tratamiento. Tres meses después, se vuelve a medir el peso de esos mismos hombres, obteniendo la tabla:

Peso antes de la dieta	85	85	81	87	74	74	76	78
Peso después de la dieta	70	72	60	69	57	56	54	62

Se considera que el peso medio de los 500 hombres del pueblo es 76 kilogramos, dato obtenido de un estudio médico de gran precisión en la comarca a la que pertenece el pueblo.

a) Estimar la reducción promedio en peso de los hombres tras la dieta, si se aplicara el tratamiento a cualquier hombre del pueblo.

b) Estimar el hipotético peso promedio de los hombres del pueblo después del tratamiento si se aplicara la dieta a todos. Comparar en esta estimación los dos estimadores indirectos.

8) Un agricultor desea estimar el número medio de olivos por cada hectárea en un cierto terreno que posee. Conoce la producción total de aceituna en años anteriores, que es 12 toneladas métricas (1 Tm=1000 kilos).

El agricultor selecciona por muestreo sistemático 6 hectáreas de las 25 que posee, cuenta el número de olivos en cada una de ellas y cuando llega la recolecta marca las cajas que provienen de esas hectáreas, conociendo la producción de cada una de ellas.

Obtiene los siguientes datos:

Hectárea	1	2	3	4	5	6
nº de olivos	193	193	180	201	147	147
Producción (en kilos)	908	925	782	924	648	648

a) Estimar el número medio de olivos por hectárea. Comparar los dos estimadores indirectos, estimando sus varianzas.

b) ¿Cual sería el tamaño muestral necesario para estimar esa media con un error de muestreo de 2 unidades?.

9) Utilizar la macro estimrazreg del SAS para resolver el ejercicio propuesto número 7.

10) Utilizar la macro estimrazreg del SAS para resolver el ejercicio propuesto número 8.

11) Utilizar la macro estimrazestrat del SAS para resolver el ejercicio resuelto número 6.4.

12) Utilizar la macro estimrazreg del SAS para resolver el ejercicio resuelto número 6.8.

13) El archivo SAS activos contiene la población activa en miles de personas en las 52 provincias españolas en 1996 y 1997. Supongamos que se dispone de la información de 1996 y se desea utilizar m.a.s. de varias provincias en 1997, para utilizar estimación de razón o regresión para la media de activos por provincia. Sabiendo que la media poblacional en 1996 es 306.46, se pide:

ISBN 9788496866133

a) Extraer 5 m.a.s. de tamaño 10 con las semillas 1234, 1235, 1236, 1237, 1238. Con la macro estimrazreg, estimar por razón y regresión la media de activos por provincia. Ir apuntando los resultados en la tabla que aparece más abajo. Estimar también esa media con el estimador de expansión.

b) Se crean 4 regiones diferenciadas: 1=Norte, 2=Sur, 3=Castilla, 4=Levante, en la variable región, con número de provincias respectivas 14, 14, 15, y 9. Utilizar

```
proc sort data=activos;by region;
proc means data=activos;var acti96;by region;run;
```

para obtener las medias de la variable auxiliar por estrato. Utilizar la afijación 2, 3, 3, 2 para extraer una m.a.s. estratificada de tamaño $n = 10$. Hacerlo 5 veces, con las mismas semillas que en a).

Con la macro estimrazestrat, calcular cada vez los estimadores separados y combinados de razón y regresión de la media. Apuntar los resultados en la tabla.

Con la macro estimestrat, calcular cada vez el estimador de expansión estratificado. Apuntar los resultados en la tabla.

c) Conclusiones.

Muestra	$\overline{y}$	$\overline{y}_R$	$\overline{y}_{reg}$	$\overline{y}_{st}$	t_c	t_s	t_{regs}	$\overline{y}_{regc}$
1								
2								
3								
4								
5								

ISBN 9788496866133

8 MUESTREO CON PROBABILIDADES DESIGUALES

En este capítulo se tratará de un método de muestreo alternativo al muestreo aleatorio simple o muestreo sistemático, donde las probabilidades de aparecer en la muestra eran las mismas para cada unidad poblacional .

En muestreo con probabilidades desiguales, cada unidad i tiene asociada una probabilidad inicial p_i de ser escogida, frecuentemente calculada a partir de una variable auxiliar x_i (nótese que se requiere conocer a priori el valor de x para toda la población).

Este método de muestreo se utiliza sobre todo cuando las unidades consideradas son conglomerados (conjuntos de unidades), y puede ser bastante eficiente en caso de una relación de proporcionalidad directa entre x e y (o dicho de otro modo, entre las probabilidades de selección p_i y los valores y_i).

Respecto a la utilización de una variable auxiliar x, hay que observar que los métodos de probabilidades desiguales exigen poder conocer todos los valores x_i de la población antes de llevar a cabo el muestreo para construir las p_i, mientras que otros métodos como el muestreo aleatorio simple combinado con estimación de razón o regresión solamente exigían el conocimiento de los valores x_i muestrales y de $\overline{x}$.

8.1 Muestreo con probabilidades desiguales con reemplazamiento

En este apartado se estudiará el caso en que cada unidad poblacional $i = 1, ..., N$ lleva asociada una probabilidad p_i de ser escogida, y se realiza la selección sucesiva e independiente de n unidades, de modo que la probabilidad de selección permanece constante en cada extracción, al ser con reemplazamiento.

Las probabilidades p_i pueden ser:

ISBN 9788496866133

a) Arbitrarias (cumpliendo $\sum_{i=1}^{N} p_i = 1$).

b) Construidas a partir de una variable auxiliar x conocida para toda la población, de manera que $p_i = \dfrac{x_i}{\sum_{i=1}^{N} x_i}$. Como a menudo la variable x suele estar relacionada con la magnitud o "tamaño" de la unidad, este método de muestreo se denomina también **muestreo proporcional al tamaño con reemplazamiento** (pptr) aunque x no indique forzosamente tamaño .

Ejemplo 8.1.

Supongamos que se quiere realizar un muestreo pptr de $n = 3$ viviendas de las $N = 6$ que conforman un edificio. Se desea que la probabilidad asociada a cada vivienda sea proporcional al número de habitantes de la vivienda x.

Vivienda	**nº habitantes**	**probabilidad de selección** $p_i = \frac{x_i}{\sum x_i}$
1	8	8/30
2	6	6/30
3	3	3/30
4	5	5/30
5	4	4/30
6	4	4/30

Tabla 8.1. Probabilidades de selección

Para obtener las tres viviendas de la muestra, se sortea repetidamente un número entre 1 y 6, con probabilidades respectivas $p_1 = \frac{8}{30}, ..., p_6 = \frac{4}{30}$. Al ser muestreo con reemplazamiento, las viviendas pueden repetirse en la muestra.

Usualmente se utilizan los métodos de probabilidades desiguales cuando las unidades son grandes conglomerados o conjuntos de subunidades de interés. La variable au-xiliar x a partir de la cual se construyen las probabilidades suele estar correlacionada con y. Algunos ejemplos son:

ISBN 9788496866133

unidad	**variable de estudio** y	**variable auxiliar** x
pueblos	población	población censo anterior
pueblos	ingresos	declaración renta anterior
terrenos	área de cierto cultivo	área geográfica total
fábricas	producción	nº trabajadores

Tabla 8.2. Ejemplos de muestreo ppt

8.1.1 Estimación en muestreo pptr

En caso de muestreo pptr, es necesario construir diferentes estimadores de los que son habituales en m.a.s., al haber alterado la equiprobabilidad de cada unidad poblacional.

Teorema 8.1 (estimador insesgado del total).

Sea p_i la probabilidad de seleccionar la unidad i en la muestra. Sea una muestra obtenida mediante muestreo proporcional al tamaño con reemplazamiento. Entonces

$$t_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i},$$

denominado **estimador de Hansen-Hurwitz** del total poblacional, es un estimador insesgado de $N\overline{y}$.

Demostración.

$$E(t_{HH}) = \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{y_i}{p_i}\right).$$

Ahora, cada $\dfrac{y_i}{p_i}$ muestral es una variable aleatoria que toma valores posibles $\dfrac{y_1}{p_1}, ..., \dfrac{y_N}{p_N}$, siendo las probabilidades respectivas de estos valores $p_1, ..., p_N$. Así, por la definición de esperanza, $E\left(\dfrac{y_i}{p_i}\right) = \sum_{j=1}^{N} p_j\left(\dfrac{y_j}{p_j}\right) = \sum_{j=1}^{N} y_j = N\overline{y}$ para todo i, y entonces

$$E(t_{HH}) = \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{y_i}{p_i}\right) = N\overline{y}.$$

Corolario 8.1.

Si $p_i = \dfrac{1}{N}$ el muestreo pptr es equivalente al muestreo aleatorio simple con reemplazamiento, pues consiste en escoger con reemplazamiento entre N unidades con probabilidades iguales. Además, en este caso el estimador de Hansen-Hurwitz coincide con el estimador usual en m.a.s.r, es decir, $t_{HH} = N\overline{y}_s$.

ISBN 9788496866133

Teorema 8.2 (varianza del estimador).

$$V(t_{HH}) = \frac{1}{n}\left(\sum_{j=1}^{N}\frac{y_j^2}{p_j} - (N\overline{y})^2\right).$$

Demostración.

$V(t_{HH}) = \frac{1}{n^2}\sum_{i=1}^{n} V\left(\frac{y_i}{p_i}\right)$ por ser muestreo con reemplazamiento (extracciones independientes).

Por otra parte,

$$V\left(\frac{y_i}{p_i}\right) = E\left[\left(\frac{y_i}{p_i}\right)^2\right] - \left[E\left(\frac{y_i}{p_i}\right)\right]^2.$$

Cada $\left(\frac{y_i}{p_i}\right)^2$ muestral es una variable aleatoria que toma valores posibles $(\frac{y_1}{p_1})^2$, ... , $(\frac{y_N}{p_N})^2$, siendo las probabilidades respectivas de estos $p_1, ..., p_N$. Así,

$V\left(\frac{y_i}{p_i}\right) = \sum_{j=1}^{N} p_j\frac{y_j^2}{p_j^2} - (N\overline{y})^2 = \sum_{j=1}^{N}\frac{y_j^2}{p_j} - (N\overline{y})^2$ para todo i, y por lo tanto

$$\frac{1}{n^2}\sum_{i=1}^{n} V\left(\frac{y_i}{p_i}\right) = \frac{1}{n}V\left(\frac{y_i}{p_i}\right) = \frac{1}{n}\left(\sum_{j=1}^{N}\frac{y_j^2}{p_j} - (N\overline{y})^2\right).$$

Teorema 8.3 (estimador de la varianza).

Un estimador insesgado de $V(t_{HH})$ es

$$\widehat{V}(t_{HH}) = \frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\frac{y_i^2}{p_i^2} - nt_{HH}^2\right).$$

Demostración.

$$E[\widehat{V}(t_{HH})] = \frac{1}{(n-1)}\left[\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{y_i^2}{p_i^2}\right) - E(t_{HH}^2)\right] =$$

$$= \frac{1}{(n-1)}\left[\sum_{j=1}^{N}\frac{y_j^2}{p_j} - (V(t_{HH}) + (E(t_{HH}))^2)\right] =$$

$$= \frac{1}{(n-1)}\left[\sum_{j=1}^{N}\frac{y_j^2}{p_j} - \frac{1}{n}\sum_{j=1}^{N}\frac{y_j^2}{p_j} + \frac{1}{n}(N\overline{y})^2 - (N\overline{y})^2\right] =$$

$$= \frac{1}{(n-1)}\left[\left(\frac{n-1}{n}\right)\sum_{j=1}^{N}\frac{y_j^2}{p_j} - \left(\frac{n-1}{n}\right)(N\overline{y})^2\right] =$$

$$= \frac{1}{n}\left(\sum_{j=1}^{N}\frac{y_j^2}{p_j} - (N\overline{y})^2\right) = V(t_{HH}).$$

Otras formas de la varianza y del estimador de la varianza vienen expresadas a continuación.

ISBN 9788496866133

Teorema 8.4 (expresiones alternativas para las varianzas).

(a) $V(t_{HH}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{y_i}{p_i} - N\overline{y}\right)^2 p_i.$

(b) $\widehat{V}(t_{HH}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - t_{HH}\right)^2.$

Demostración.

$$\text{(a) } \frac{1}{n}\sum_{i=1}^{N}\left(\frac{y_i}{p_i} - N\overline{y}\right)^2 p_i = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{y_i^2}{p_i^2} - 2N\overline{y}\frac{y_i}{p_i} + (N\overline{y})^2\right) p_j =$$

$$= \frac{1}{n}\left[\sum_{i=1}^{N}\frac{y_i^2}{p_i} - 2N\overline{y}\sum_{i=1}^{N} y_i + (N\overline{y})^2\sum_{i=1}^{N} p_i\right] =$$

$$= \frac{1}{n}\left[\sum_{i=1}^{N}\frac{y_i^2}{p_i} - (N\overline{y})^2\right] = V(t_{HH}),$$

pues se ha utilizado que $\sum_{i=1}^{N} p_i = 1$ y $\sum_{i=1}^{N} y_i = N\overline{y}$.

$$\text{(b) } \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - t_{HH}\right)^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i^2}{p_i^2} - 2\frac{y_i}{p_i}t_{HH} + t_{HH}^2\right) =$$

$$= \frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\frac{y_i^2}{p_i^2} - 2t_{HH}\sum_{i=1}^{n}\frac{y_i}{p_i} + \sum_{i=1}^{n} t_{HH}^2\right) =$$

$$= \frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\frac{y_i^2}{p_i^2} - nt_{HH}^2\right) = \widehat{V}(t_{HH}).$$

Se ha utilizado en el desarrollo que $\sum_{i=1}^{n}\frac{y_i}{p_i} = nt_{HH}$.

Para estimar la media poblacional basta con corregir las expresiones obtenidas para la estimación del total.

Corolario 8.2 (estimación de la media y proporción).

1. $\frac{1}{N}t_{HH}$ es un estimador insesgado de la media poblacional $\overline{y}$, con varianza $\frac{1}{N^2}V(t_{HH})$
2. Un estimador insesgado de la varianza del estimador de la media es $\frac{1}{N^2}\widehat{V}(t_{HH})$.
3. Supongamos que la variable y toma valores 0 ó 1 y se desea estimar la proporción p de valores 1 en la población. Un estimador insesgado de esta proporción es $\frac{1}{N}t_{HH}$, donde $t_{HH} = \frac{1}{n}\sum_{j=1}^{n}\frac{y_j}{p_j}$, siendo y_j el total en el conglomerado i, que en este caso es el número de valores "1" en el conglomerado i. Este estimador tiene varianza $\frac{1}{N^2}V(t_{HH})$ y un estimador insesgado de esta varianza es $\frac{1}{N^2}\widehat{V}(t_{HH})$.

ISBN 9788496866133

El siguiente teorema y corolario justifican por qué la variable auxiliar x debe estar altamente correlada con y para obtener buenas estimaciones, y además la relación debe de ser aproximadamente de proporcionalidad (la recta de regresión pasa por el origen).

Teorema 8.5 (probabilidades proporcionales a y).

La varianza $V(t_{HH})$ es mínima y además $V(t_{HH}) = 0$ si $p_i = \dfrac{y_i}{\sum_{i=1}^{N} y_i}$, es decir si las probabilidades de selección p_i son proporcionales a los y_i.

Demostración.

Se trata de resolver el problema de minimización:

$\underset{p_1,\ldots,p_N}{Min}\ V(t_{HH})$, sujeto a $\sum_{i=1}^{N} p_i = 1$.

La función de Lagrange es

$$\Phi = \frac{1}{n}\left(\sum_{i=1}^{N}\frac{y_i^2}{p_i} - (N\overline{y})^2\right) + \lambda(\sum_{i=1}^{N} p_i - 1)$$

Derivando respecto de p_i e igualando a cero , queda

$-\frac{1}{n}\frac{y_i^2}{p_i^2} + \lambda = 0$ para todo i, y así, $p_i = \frac{y_i}{\sqrt{n\lambda}}$. Como además $\sum_{i=1}^{N} p_i = 1$ esto implica que $\sum_{i=1}^{N}\frac{y_i}{\sqrt{n\lambda}} = 1$ y por lo tanto $\lambda = \frac{1}{n}(N\overline{y})^2$ y sustituyendo en la primera expresión ,

$$-\frac{1}{n}\frac{y_i^2}{p_i^2} + \frac{1}{n}(N\overline{y})^2 = 0 \Rightarrow p_i = \frac{y_i}{N\overline{y}} = \frac{y_i}{\sum_{i=1}^{N} y_i}.$$

La varianza es mínima en esos valores de p_i y además, en ese caso,

$$\frac{1}{n}\left(\sum_{i=1}^{N}\frac{y_i^2}{p_i} - (N\overline{y})^2\right) = \frac{1}{n}\left(\sum_{i=1}^{N}\frac{y_i^2 N\overline{y}}{y_i} - (N\overline{y})^2\right) =$$

$$= \frac{1}{n}\left(N\overline{y}\sum_{i=1}^{N} y_i - (N\overline{y})^2\right) = 0.$$

Los valores poblacionales y_i se desconocen en la práctica, pero el teorema induce a escoger como variable auxiliar x aquella de las disponibles que esté más altamente correlacionada con y.

Corolario 8.3 (probabilidades proporcionales a una variable auxiliar).

Supongamos que las probabilidades de selección se escogen proporcionales a la variable auxiliar x. Si la relación entre x e y es lineal y la recta pasa por el origen, $V(t_{HH}) = 0$.

Demostración.

Si $y = bx$, entonces

ISBN 9788496866133

$$p_i = \frac{x}{\sum_{i=1}^{N} x_i} = \frac{\frac{1}{b} y_i}{\frac{1}{b} \sum_{i=1}^{N} y_i} = \frac{y_i}{\sum_{i=1}^{N} y_i}$$

y por lo tanto $V(t_{HH}) = 0$.

Ejemplo 8.2.

El siguiente ejemplo es puramente teórico, y pretende mostrar el funcionamiento del estimador presentado. Se supone una población de 4 unidades A,B,C,D, donde se plantea una variable de interés y y una variable de tamaño x. Supongamos que se realiza muestreo pptr con $n = 2$.

	A	B	C	D
y	3	4	5	6
x	2	4	6	8

Tabla 8.3. Datos de ejemplo

Por lo tanto las p_i serán, respectivamente, $\frac{2}{20}, \frac{4}{20}, \frac{6}{20}, \frac{8}{20}$, es decir $\{0.1, 0.2, 0.3, 0.4\}$. El total poblacional, que consideraremos la cantidad a estimar, es $\sum_{i=1}^{N} y_i = 18$.

Las muestras posibles teniendo en cuenta el orden son 16. Cada una de estas muestras (i, j) de tamaño 2 tiene una probabilidad de $p_i p_j$ de ser escogida, debido a la independencia en el muestreo. Se puede construir entonces la tabla de probabilidades para cada valor posible del estimador. Por ejemplo, Para la muestra (A, B) la probabilidad de ser escogida es $0.1 \cdot 0.2 = 0.02$, y el valor del estimador del total es

$$t_{HH} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i} = \frac{1}{2}\left(\frac{3}{0.1} + \frac{4}{0.2}\right) = 25.$$

La tabla 8.1 ofrece todos los valores de t_{HH}, así como sus respectivas probabilidades de selección.

Un histograma directo de los valores del estimador sería incorrecto en este caso, al no ser estos valores equiprobables. Lo correcto es representar cada valor del estimador con una barra proporcional en altura a la probabilidad de ese valor. La Figura 8.2 refleja la función de probabilidad del estimador. Se ve como la aproximación normal sería incorrecta.

Se puede comprobar que el estimador es insesgado, como se demostró, pues

$$E(t_{HH}) = \sum_{\substack{muestras \\ posibles}} t_{HH_i} \cdot p(muestra_i) = 30 \cdot 0.01 + 25 \cdot 0.02 + ... + 15 \cdot 0.16 = 18.$$

La varianza del estimador se puede calcular con la definición básica de varianza, puesto que se dispone de la tabla de probabilidades y valores, o bien recurriendo a la expresión ya demostrada

$$V(t_{HH}) = \frac{1}{n}\left(\sum_{j=1}^{N} \frac{y_j^2}{p_j} - (N\overline{y})^2\right) = 9.667.$$

ISBN 9788496866133

Observando la tabla 8.1, se ve que las muestras más probables son (D, D), que arroja una estimación de 15 para el total, y (D, C) y (C, D), con estimaciones de 15.83.

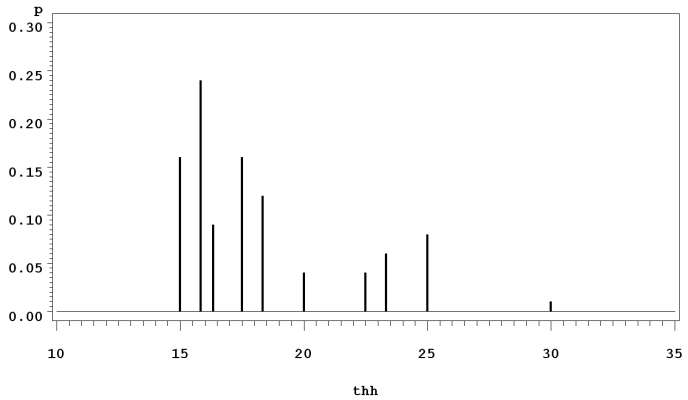


Figura 8.1. Función de probabilidad de t_{HH}.

$Muestra$	t_{HH}	$p(muestra)$
(A, A)	30.00	0.01
(A, B)	25.00	0.02
(A, C)	23.33	0.03
(A, D)	22.50	0.04
(B, A)	25.00	0.02
(B, B)	20.00	0.04
(B, C)	18.33	0.06
(B, D)	17.50	0.08
(C, A)	23.33	0.03
(C, B)	18.33	0.06
(C, C)	16.33	0.09
(C, D)	15.83	0.12
(D, A)	22.50	0.04
(D, B)	17.50	0.08
(D, C)	15.83	0.12
(D, D)	15.00	0.16

Tabla 8.4. Muestras posibles, valor del estimador y probabilidades.

ISBN 9788496866133

Ejemplo 8.3.

Un estudio pretende estimar el número de recetas médicas emitidas en total una región. Para ello se realiza muestreo con probabilidad proporcional al tamaño de $n = 10$ pueblos de los $N = 300$ que componen la región. La variable de tamaño utilizada es x =número de habitantes del pueblo.

Tras realizar el sorteo, caen en la muestra ciertos pueblos, con probabilidades respectivas de aparición $\{0.003,\ 0.01,\ 0.008,\ 0.02,\ 0.006,\ 0.07,\ 0.004,\ 0.002,\ 0.01,\ 0.003\}$ y los valores del número de recetas médicas obtenidos en esos pueblos son, respectivamente,

$\{40, 100, 75, 210, 65, 610, 38, 20, 90, 45\}$.

El estimador de Hansen-Hurwitz del total poblacional será, por lo tanto,

$$t_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} = \frac{1}{n}(\frac{40}{0.003} + \frac{100}{0.01} + \frac{75}{0.008} + ... + \frac{45}{0.003}) = \frac{1}{n}\cdot 106255.95 = 10626$$

y el estimador de la media de recetas expedidas por pueblo será $\frac{1}{N}t_{HH} = 35.41$.

El resultado para la estimación de la media puede parecer extraño, habida cuenta de que observando los valores muestrales de y, sólo una observación supera el valor 35.41. La razón para ese valor del estimador es que, en general, las probabilidades de que las observaciones presentes en esta muestra fueran elegidas son altas, (compárese, por ejemplo, con las probabilidades de cada observación si éstas fueran idénticas para todos los pueblos: estas serían $p_i = \frac{1}{N} = 0.0033$ para todo i). El hecho de que estas probabilidades sean relativamente altas hace disminuir el valor del estimador pue las p_i están en el denominador.

Esto no es un defecto, pues si el estimador no tuviera esta propiedad de corrección no sería insesgado. Es razonable suponer que al caer en la muestra pueblos proporcionalmente más grandes (pues sus p_i son grandes), el resto de pueblos serán en general más pequeños, y esto justifica intuitivamente el valor pequeño del estimador.

Respecto a la proporcionalidad relativa entre la probabilidad de selección y la variable de interés, se puede observar que, para la muestra obtenida, r_{py} =0.998. Pero como se ha visto, es deseable no solamente una relación lineal entre p_i e y, sino que la relación sea de proporcionalidad (constante de la recta de regresión aproximadamente igual a cero). La Figura 8.2 expresa la relación entre p_i e y en la muestra obtenida, que parece de proporcionalidad casi exacta. El estimador parece por lo tanto bastante fiable. La estimación de la varianza del estimador será, además,

$$\widehat{V}(t_{HH}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - t_{HH}\right)^2 =$$

$$= \frac{1}{10\cdot 9}\left[\left(\frac{40}{0.003} - 10626\right)^2 + \left(\frac{100}{0.01} - 10626\right)^2 + ... + \left(\frac{45}{0.003} - 10626\right)^2\right] =$$

$$= 404839.43$$

y su desviación típica $\sqrt{\widehat{V}(t_{HH})} = 636.27$.

ISBN 9788496866133

Suponiendo normalidad del estimador, el intervalo de confianza al 95% para el total de recetas emitidas en la región será

$$(10626 \pm 1.96 \cdot 636.27) = (9378.5, 11872.68).$$

El intervalo de confianza al 95% para la media de recetas por pueblo emitidas será

$$(35.41 \pm 1.96 \cdot \frac{636.27}{300}) = (31.25, 39.56).$$

Obsérvese que no es directo, a priori, comparar la estimación de la precisión con muestreo pptr con la precisión estimada obtenida con m.a.s., pues ésta última es desconocida al haber sido el muestreo diferente. Aunque existen métodos para estimar S^2 (y por lo tanto la varianza del estimador media muestral) a partir de una muestra obtenida por muestreo pptr, no se expondran aquí.

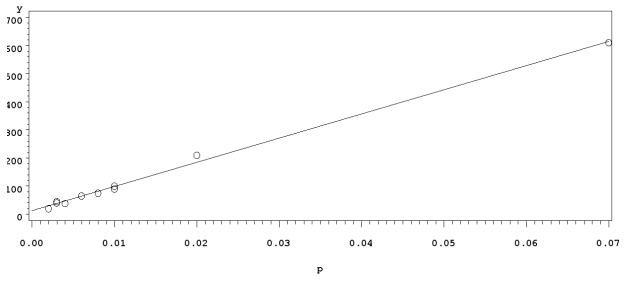


Figura 8.2. Relación entre p_i e y.

8.2 Métodos de selección de la muestra

En caso de muestreo pptr, existen varios métodos para escoger las unidades poblacionales con pro-babilidades proporcionales a x_i.

a) **Método acumulativo**

1) Se construye la tabla siguiente:

ID	V. auxiliar	Total acumulado
1	x_1	$T_1 = x_1$
2	x_2	$T_2 = x_1 + x_2$
.	.	.
.	.	.
N	x_N	$T_N = x_1 + x_2 + ... + x_N$

Tabla 8.5. Método acumulativo

ISBN 9788496866133

2) Se selecciona un número aleatorio uniforme $R \in (0, T_N)$.

3) Considerando $T_0 = 0$, se selecciona la observación i tal que $T_{i-1} < R \leq T_i$.

Los pasos 2-3 se repiten n veces para obtener la muestra de tamaño n.

El método es similar a calcular primero $p_i = \dfrac{x}{\sum_{i=1}^{N} x_i}$, calcular a continuación los totales acumulados sobre los p_i, y seleccionar el número $R \in U(0,1)$.

Para realizar el método con calculadora u ordenador, se calcula $R = T_N \cdot u$, donde u proviene de una $U(0,1)$.

Hay que hacer notar que este sistema requiere calcular los totales acumulados, lo cual puede hacerse largo. Además, si existe un coste relacionado con el conocimiento de cada x_i, el método es costoso pues requiere conocer todos los x_i a priori. El siguiente método evita tener que conocer a priori todos los x_i.

b) **Método de Lahiri**

Antes de comenzar la selección, se determina un número x_0 tal que $x_0 \geq \max(x_1, ..., x_n)$.

1) Se escoge aleatoriamente y con equiprobabilidad un número entero $i \in (1, N)$.

2) Se escoge aleatoriamente y con equiprobabilidad un número $R \in (0, x_0)$.

3) Si $R \leq x_i$, entonces la observación i es seleccionada. Si no, es rechazada (se vuelve a repetir el proceso).

Los pasos 1-2-3 se repiten n veces para obtener la muestra de tamaño n. Hay que observar que el método puede ser largo e ineficiente si x_0 es mucho mayor que el máximo, pues habrá muchas unidades rechazadas.

Para escoger el número entero i con calculadora u ordenador, basta generar u de una $U(0,1)$ y hacer $i = [N * u] + 1$ donde el corchete indica la parte entera.

Ejemplo 8.3.

Se utilizarán los datos del Ejemplo 8.2. Supongamos que queremos extraer una muestra con muestreo pptr de tamaño n.

	A	B	C	D
y_i	3	4	5	6
x_i	2	4	6	8
p_i	0.1	0.2	0.3	0.4

Tabla 8.6. Datos de ejemplo

ISBN 9788496866133

Método acumulativo

1) La tabla es:

ID	$V.auxiliar$	$Total\ acumulado$
A	2	$T_1 = 2$
B	4	$T_2 = 6$
C	6	$T_3 = 12$
D	8	$T_4 = 20$

Tabla 8.7. Totales acumulados

2) Se escoge un número Uniforme $R \in U(0, T_N)$. Utilizando la calculadora, se obtiene por ejemplo 17.41.

3) Como el valor 17.41 está entre 12 y 20, se selecciona la observación D.

Los pasos 2 y 3 se realizarán n veces, obteniendo una muestra de tamaño n.

Método de Lahiri

Supongamos que se pone $x_0 = 10$.

1) Se escoge aleatoriamente y con equiprobabilidad un número entero $i \in (1, 4)$. Por ejemplo, utilizando la calculadora se obtiene $i = 2$.

2) Se escoge aleatoriamente y con equiprobabilidad un número entero $R \in (0, 10)$. Por ejemplo, utilizando la calculadora se obtiene $R = 6$.

3) Si $R \leq x_i$, entonces la observación i es seleccionada. Si no, es rechazada (se vuelve a repetir el proceso). Como en nuestro caso $6 > x_2 = 4$, se repite el proceso:

1) Obtenemos con la calculadora $i = 4$.

2) Obtenemos con la calculadora $R = 7$.

3) Como $7 \leq x_4 = 8$, se acepta la observación $4 = D$.

El método a partir del paso 1) se realizará n veces, obteniendo una muestra de tamaño n.

8.3 Muestreo con probabilidades desiguales sin reemplazamiento

En este tipo de muestreo, que se notará por ppt (muestreo proporcional al tamaño), cada unidad poblacional i tiene probabilidad p_i inicial de ser seleccionada, pero una vez que está en la

ISBN 9788496866133

muestra, no puede volver a ser escogida. Como ocurre habitualmente , la varianza del estimador será menor que en muestreo con reemplazamiento, pero existen cuestiones relacionadas con el esquema de selección y estimación de varianzas que complican a menudo su utilización.

8.3.1 Estimación en muestreo ppt

En este tipo de muestreo es necesario contar con el cálculo de las probabilidades de inclusión π_i de cada unidad poblacional. Veremos en primer lugar dos resultados que serán útiles.

Teorema 8.6.

Sea π_i la probabilidad de inclusión de la unidad poblacional i en la muestra, en muestreo sin reemplazamiento (las probabilidades pueden ser desiguales). Entonces $\sum_{i=1}^{N} \pi_i = n$.

Demostración.

Definamos la variable aleatoria $e_i = \begin{cases} 1 \text{ si la unidad } i \text{ está en la muestra} \\ 0 \text{ si la unidad } i \text{ no está en la muestra} \end{cases}$.

Entonces $E(e_i) = \pi_i$. Además, en muestreo sin reemplazamiento, $\sum_{i=1}^{N} e_i = n$ pues al no haber repeticiones, se cuenta $e_i = 1$ por cada unidad muestral de las n que pertenecen, y 0 por cada unidad que no pertenece. Tomando esperanzas en ambos lados, $E\left(\sum_{i=1}^{N} e_i\right) = n$ y por lo tanto $\sum_{i=1}^{N} E(e_i) = \sum_{i=1}^{N} \pi_i = n$.

Teorema 8.7.

Sea π_{ij} la probabilidad de inclusión de las unidades poblacionales i y j en la muestra, en muestreo sin reemplazamiento . Entonces dado $i \in \{1, ..., N\}$, $\sum_{j \neq i}^{N} \pi_{ij} = (n-1)\pi_i$.

Demostración.

Se tiene que

$$\sum_{j \neq i}^{N} \pi_{ij} = \sum_{j \neq i}^{N} P(j \in \text{muestra} |\ i \in \text{muestra}) P(i \in \text{muestra}) =$$

$$\sum_{j \neq i}^{N} P(j \in \text{muestra} |\ i \in \text{muestra})\pi_i = \pi_i \sum_{j \neq i}^{N} P(j \in \text{muestra} |\ i \in \text{muestra}).$$

Ahora, esta suma de probabilidades es la correspondiente a la selección sin reemplazamiento de $(n-1)$ unidades de todas las $(N-1)$ restantes (pues i ya está en la muestra). Así, $\sum_{j \neq i}^{N} P(j \in \text{muestra} |\ i \in \text{muestra}) = (n-1)$ y por lo tanto $\sum_{j \neq i}^{N} \pi_{ij} = (n-1)\pi_i$.

ISBN 9788496866133

Ejemplo 8.4.

Supóngase que se quiere extraer una muestra de tamaño $n = 2$ con probabilidades proporcionales al tamaño y sin reemplazamiento, de la población del Ejemplo 8.3.

	A	B	C	D
y_i	3	4	5	6
p_i	0.1	0.2	0.3	0.4

Tabla 8.8. Datos de ejemplo

El método básico de muestreo ppt sin reemplazamiento consiste en mantener las probabilidades de extraer cada unidad proporcionales a las probabilidades iniciales p_i, en cada paso. Es decir, si en la primera extracción ha salido la unidad i, en la segunda la probabilidad de cada unidad j es $\frac{p_j}{1-p_i}$. Si en la 1ª y 2ª extracción han salido las unidades i y j, en la tercera la probabilidad de que salga k es $\frac{p_k}{1-p_i-p_j}$, y así sucesivamente.

En este esquema la probabilidad de cada muestra, ordenada, se calcula según el ejemplo que sigue:

$P(A,B) = P($ sale A en la primera extracción y sale B en la segunda)=

=P(sale A en la primera extracción)$\cdot P$(sale B en la segunda| ha salido A en la primera)=

$$=p_A \cdot \frac{p_B}{1-p_A} = 0.1 \cdot \frac{0.2}{1-0.1} = 0.02222.$$

Hay que remarcar que no existe simetría en el sentido de que $P(B,A) = p_B \cdot \frac{p_A}{1-p_B} = 0.025$.

Así, tenemos la tabla de probabilidades 8.9.

ISBN 9788496866133

$Muestra$	$p(muestra)$
(A, B)	0.0222
(A, C)	0.0333
(A, D)	0.0444
(B, A)	0.0250
(B, C)	0.0750
(B, D)	0.10
(C, A)	0.0428
(C, B)	0.0857
(C, D)	0.1714
(D, A)	0.0666
(D, B)	0.1333
(D, C)	0.20

Tabla 8.9. Probabilidades para cada muestra

Con lo cual

$$\pi_A = P(A,B) + P(A,C) + P(A,D) + P(B,A) + P(C,A) + P(D,A) =$$

$$= 0.0222 + 0.0333 + 0.0444 + 0.0250 + 0.0428 + 0.0666 = 0.234.$$

Del mismo modo, se calcula $\pi_B = 0.4412$, $\pi_C = 0.6082$ y $\pi_D = 0.7157$.

Se puede verificar entonces el resultado $\sum_{i=1}^{N} \pi_i = n$ pues

$$\sum_{i=1}^{N} \pi_i = \pi_A + \pi_B + \pi_C + \pi_D = 2.$$

Utilizando que las probabilidades de inclusión de segundo orden se pueden también calcular, pues por ejemplo $\pi_{AB} = P(A,B) + P(B,A) = 0.0472$, se puede comprobar que $\sum_{j \neq i}^{N} \pi_{ij} = (n-1)\pi_i$ pues

$$\sum_{j \neq A}^{N} \pi_{Aj} = \pi_{AB} + \pi_{AC} + \pi_{AD} = 0.0472 + 0.0761 + 0.111 = 0.234 = (n-1)\pi_A.$$

Igualmente se puede comprobar para $\sum_{j \neq B}^{N} \pi_{Bj}$, $\sum_{j \neq C}^{N} \pi_{Cj}$ y $\sum_{j \neq D}^{N} \pi_{Dj}$.

ISBN 9788496866133

El siguiente teorema define el estimador insesgado más utilizado en este tipo de muestreo.

Teorema 8.8 (estimador insesgado del total).

Sea una muestra $y_1, ..., y_n$ obtenida mediante muestreo ppt, donde las probabilidades de inclusión respectivas de las unidades muestrales son π_i. Entonces el **estimador de Horvitz-Thompson** del total poblacional $t_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$ es un estimador insesgado de $N\overline{y}$.

Demostración.

Definamos la variable aleatoria $e_i = \begin{cases} 1 \text{ si la unidad } i \text{ está en la muestra} \\ 0 \text{ si la unidad } i \text{ no está en la muestra} \end{cases}$.

Entonces como π_i es la probabilidad de que la unidad i esté en la muestra, $e_i \equiv B(1, \pi_i)$ es una Bernoulli. Así, $E(e_i) = \pi_i$ y $V(e_i) = \pi_i(1 - \pi_i)$. Se puede reescribir el estimador de la siguiente manera:

$t_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{N} \frac{y_i}{\pi_i} e_i$ y por lo tanto,

$$E(t_{HT}) = E\left(\sum_{i=1}^{N} \frac{y_i}{\pi_i} e_i\right) = \sum_{i=1}^{N} \frac{y_i}{\pi_i} E(e_i) = \sum_{i=1}^{N} y_i = N\overline{y}.$$

La probabilidad de inclusión en m.a.s. es $\pi_i = \frac{n}{N}$, con lo que el siguiente corolario es directo.

Corolario 8.3.

Cuando se trata de muestreo aleatorio simple sin reemplazamiento, el estimador t_{HT} es tal que $t_{HT} = N\widehat{\overline{y}}$.

Teorema 8.9 (varianza del estimador).

La varianza del estimador de Horvitz-Thompson es

$$V(t_{HT}) = \sum_{i=1}^{N} \frac{1 - \pi_i}{\pi_i} y_i^2 + \sum_{i,j,i \neq j}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j$$

o, de otro modo,

$$V(t_{HT}) = \sum_{i=1}^{N} \frac{1 - \pi_i}{\pi_i} y_i^2 + 2 \sum_{i<j}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j.$$

Demostración.

Utilizando la definición de la variable aleatoria e_i, se tiene también que , para $i \neq j$,

$e_i e_j = \begin{cases} 1 \text{ si las dos unidades } i \text{ y } j \text{ están en la muestra} \\ 0 \text{ en cualquier otro caso} \end{cases}$

ISBN 9788496866133

La probabilidad de que $e_i e_j = 1$ es la probabilidad de inclusión de ambas unidades, π_{ij}, y además, $E(e_i e_j) = 1 \times \pi_{ij} = \pi_{ij}$. Además, se puede ver que

$$e_i^2 = \begin{cases} 1 \text{ si la unidad } i \text{ está en la muestra} \\ \\ 0 \text{ si la unidad } i \text{ no está en la muestra} \end{cases}$$

La probabilidad de que $e_i^2 = 1$ es π_i. Además, $E(e_i^2) = 1 \times \pi_i = \pi_i$.

Ahora,

$$V(t_{HT}) = V\left(\sum_{i=1}^{N} \frac{y_i}{\pi_i} e_i\right) = \sum_{i=1}^{N} V\left(\frac{y_i}{\pi_i} e_i\right) + \sum_{i,j,i\neq j}^{N} cov(\frac{y_i}{\pi_i} e_i, \frac{y_j}{\pi_j} e_j)$$

pues al tratarse de muestreo sin reemplazamiento existe una relación de dependencia entre las unidades extraídas y la covarianza es no nula. Entonces,

$$V(t_{HT}) = \sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2} V(e_i) + \sum_{i\neq j}^{N} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j} cov(e_i, e_j) =$$

$$= \sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2}\pi_i(1-\pi_i) + \sum_{i,j,i\neq j}^{N} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}[E(e_i e_j) - E(e_i)E(e_j)] =$$

$$= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_{i,j,i\neq j}^{N} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}[\pi_{ij} - \pi_i\pi_j] =$$

$$= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_{i,j,i\neq j}^{N} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} y_i y_j.$$

Teorema 8.10 (otra forma para la varianza del estimador).

La varianza del estimador de Horvitz-Thompson puede también expresarse mediante la forma de Yates-Grundy:

$$V(t_{HT})_{YG} = \frac{1}{2} \sum_{i,j,i\neq j}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

o

$$V(t_{HT})_{YG} = \sum_{i<j}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

Demostración.

$$V(t_{HT})_{YG} = \frac{1}{2} \sum_{i,j,i\neq j}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 =$$

$$= \frac{1}{2} \sum_{i,j,i\neq j}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i^2}{\pi_i^2} + \frac{y_j^2}{\pi_j^2} - 2\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\right) =$$

$$= \frac{1}{2} \sum_{i,j,i\neq j}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i^2}{\pi_i^2} + \frac{y_j^2}{\pi_j^2}\right) - \sum_{i,j,i\neq j}^{N} (\pi_i\pi_j - \pi_{ij})\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} =$$

ISBN 9788496866133

$$= \frac{1}{2}\sum_{i,j,i\neq j}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i^2}{\pi_i^2}\right) + \frac{1}{2}\sum_{i\neq j}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_j^2}{\pi_j^2}\right) -$$

$$- \sum_{i,j,i\neq j}^{N}(\pi_i\pi_j - \pi_{ij})\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} =$$

$$= \sum_{i,j,i\neq j}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i^2}{\pi_i^2}\right) - \sum_{i,j,i\neq j}^{N}(\pi_i\pi_j - \pi_{ij})\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

pues el primer término y el segundo dan el mismo valor al sumar i o j en todo N. Sustituyendo el sumatorio $\sum_{i,j,i\neq j}^{N}$ por su versión extendida $\sum_{i=1}^{N}\sum_{j\neq i}^{N}$ y desarrollando :

$$V(t_{HT})_{YG} = \sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi_i\pi_j\left(\frac{y_i^2}{\pi_i^2}\right) - \sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi_{ij}\left(\frac{y_i^2}{\pi_i^2}\right) - \sum_{i,j,i\neq j}^{N}(\pi_i\pi_j - \pi_{ij})\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} =$$

$$= \sum_{i=1}^{N}\pi_i\left(\frac{y_i^2}{\pi_i^2}\right)\sum_{j\neq i}^{N}\pi_j - \sum_{i=1}^{N}\left(\frac{y_i^2}{\pi_i^2}\right)\sum_{j\neq i}^{N}\pi_{ij} - \sum_{i,j,i\neq j}^{N}(\pi_i\pi_j - \pi_{ij})\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}.$$

Como $\sum_{j\neq i}^{N}\pi_j = n - \pi_i$ por ser $\sum_{j=1}^{N}\pi_j = n$, y además se sabe que $\sum_{j\neq i}^{N}\pi_{ij} = (n-1)\pi_i$, se tiene que:

$$V(t_{HT})_{YG} = \sum_{i=1}^{N}\pi_i\left(\frac{y_i^2}{\pi_i^2}\right)(n-\pi_i) - \sum_{i=1}^{N}\left(\frac{y_i^2}{\pi_i^2}\right)(n-1)\pi_i - \sum_{i,j,i\neq j}^{N}(\pi_i\pi_j - \pi_{ij})\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} =$$

$$= \sum_{i=1}^{N}\left(\frac{y_i^2}{\pi_i}\right)[(n-\pi_i) - (n-1)] - \sum_{i,j,i\neq j}^{N}(\pi_i\pi_j - \pi_{ij})\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} =$$

$$= \sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}y_i^2 + \sum_{i,j,i\neq j}^{N}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}y_iy_j.$$

El siguiente resultado permite estimar la varianza del estimador de Horvitz-Thompson.

Teorema 8.11 (estimador de la varianza).

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

$$\widehat{V}(t_{HT}) = \sum_{i=1}^{n}\frac{(1-\pi_i)}{\pi_i^2}y_i^2 + \sum_{i\neq j}^{n}\frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}}\frac{y_iy_j}{\pi_i\pi_j}.$$

o, de otro modo,

$$\widehat{V}(t_{HT}) = \sum_{i=1}^{n}\frac{(1-\pi_i)}{\pi_i^2}y_i^2 + 2\sum_{i<j}^{n}\frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}}\frac{y_iy_j}{\pi_i\pi_j}.$$

Demostración.

Definiremos e_i como en las demostraciones anteriores, y

$$e_{ij} = \begin{cases} 1 \text{ si las dos unidades } i \text{ y } j \text{ están en la muestra} \\ 0 \text{ en cualquier otro caso} \end{cases}$$

ISBN 9788496866133

de manera que $E(e_{ij}) = \pi_{ij}$.

Así, el primer sumando queda

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

y el segundo

$$\sum_{i\neq j}^{n} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j} = \sum_{i\neq j}^{N} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j} e_{ij}.$$

Por lo que

$$E(\widehat{V}(t_{HT})) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 E(e_i) + \sum_{i\neq j}^{N} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j} E(e_{ij}) =$$

$$= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_{i,j,i\neq j}^{N} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j} y_i y_j = V(t_{HT}).$$

Teorema 8.12 (otro estimador de la varianza).

Un estimador insesgado de la varianza del estimador de Horvitz-Thompson en la forma de Yates-Grundy, es:

$$\widehat{V}(t_{HT})_{YG} = \frac{1}{2} \sum_{i,j,i\neq j}^{n} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 = \sum_{i<j}^{n} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

Demostración.

Al igual que en la demostración anterior,

$$\frac{1}{2} \sum_{i,j,i\neq j}^{n} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 = \frac{1}{2} \sum_{i,j,i\neq j}^{N} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 e_{ij}$$

y por lo tanto

$$E(\widehat{V}(t_{HT})_{YG}) = \frac{1}{2} \sum_{i,j,i\neq j}^{N} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 E(e_{ij}) =$$

$$= \frac{1}{2} \sum_{i,j,i\neq j}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 = V(t_{HT})_{YG}.$$

Aunque sean estimadores insesgados de la misma cantidad $V(t_{HT})$, pueden tomar valores diferentes en cada muestra concreta, pues la propiedad de ser insesgado afecta a la esperanza o centro de gravedad de la distribución del estimador, no a cada valor particular. Además, aunque sean insesgados puede ocasionalmente salir en la estimación, un valor negativo. También estos resultados son ilustrativos de que pueden existir (y ser igualmente útiles) diferentes estimadores insesgados para un mismo parámetro.

ISBN 9788496866133

Ejemplo 8.5.

En este ejemplo teórico se estudiarán el comportamiento del estimador del total y de los estimadores de varianza para los datos del Ejemplo 8.4.

Calculando el estimador y el estimador de la varianza para cada muestra obtenida, resulta por ejemplo, para la muestra (A, B),

$t_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \frac{3}{0.234} + \frac{4}{0.4412} = 21.89$ y el estimador de la varianza de t_{HT} quedará:

$$\widehat{V}(t_{HT}) = \sum_{i=1}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + 2 \sum_{i<j}^{n} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j} =$$

$$= \frac{(1-0.234)}{0.234^2} 3^2 + \frac{(1-0.4412)}{0.4412^2} 4^2 + 2 \frac{(0.0472 - 0.234 \cdot 0.4412)}{0.0472} \frac{3}{0.234} \frac{4}{0.4412} = -104.17.$$

Al ser negativo, es incorrecto como estimador de la varianza, con lo cual será necesario comprobar el estimador de la varianza en la forma de Yates-Grundy .Este es:

$$\widehat{V}(t_{HT})_{YG} = \sum_{i<j}^{n} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 =$$

$$= \frac{(0.234 \cdot 0.4412 - 0.0472)}{0.0472} \left(\frac{3}{0.234} - \frac{4}{0.4412} \right)^2 = 16.74.$$

Se puede comprobar que el estimador t_{HT} es insesgado, pues

$$E(t_{HT}) = 0.0222 \cdot 21.89 + ... + 0.2 \cdot 16.60 = 18.$$

La varianza del estimador t_{HT} se puede calcular pues es

$$V(t_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + 2 \sum_{i<j}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j \simeq 3.42.$$

ISBN 9788496866133

$Muestra$	$p(muestra)$	t_{HT}	$\widehat{V}(t_{HT})$	$\widehat{V}(t_{HT})_{YG}$
(A,B)	0.0222	21.89	-104	16.74
(A,C)	0.0333	21.04	-31	18.41
(A,D)	0.0444	21.20	36.52	10.02
(B,A)	0.0250	21.89	-104	16.74
(B,C)	0.0750	17.29	-27.4	0.478
(B,D)	0.10	17.45	12.18	0.165
(C,A)	0.0428	21.04	-31.0	18.41
(C,B)	0.0857	17.29	-27.4	0.478
(C,D)	0.1714	16.60	22.75	0.005
(D,A)	0.0666	21.20	36.52	10.02
(D,B)	0.1333	17.45	12.18	0.165
(D,C)	0.20	16.60	22.75	0.005

Tabla 8.10 Estimaciones obtenidas para cada muestra posible.

Hay que tener en cuenta los siguientes comentarios respecto a los resultados obtenidos:

1. La varianza del estimador en el caso de muestreo sin reemplazamiento es mucho menor que en muestreo con reemplazamiento, donde era, para los mismos datos y configuración de probabilidades, 9.667 frente al 3.42 obtenido en muestreo ppt.

2. Los estimadores de varianza parecen poco estables. El estimador $\widehat{V}(t_{HT})$ da un valor negativo para 6 de las 12 muestras posibles, mientras que los dos estimadores tienen una variabilidad enorme, siendo muy difícil en la práctica, establecer con fiabilidad la precisión de las estimaciones a partir de la muestra.

3. Sin embargo si para los mismos datos se utiliza m.a.s. para estimar el total con una muestra de tamaño $n = 2$, se obtiene $V(N\widehat{\overline{y}}) = N(N-n)S^2 = 4(4-2)\cdot 1.25 = 10$ mayor que la obtenida en muestreo pptr con esas probabilidades (9.66) , y mucho mayor que en muestreo ppt (3.42). Lo que explica que con una adecuada selección de las probabilidades, se utilicen los métodos ppt y pptr en la práctica, sin preocuparse tanto de los problemas presentados en la estimación de varianzas.

4. Se puede comprobar gráficamente que la variabilidad del estimador es menor utilizando muestreo sin reemplazamiento, que con reemplazamiento, observando la gráfica de barras de la distribución de probabilidad del estimador,.como aparece en la Figura 8.3.

ISBN 9788496866133

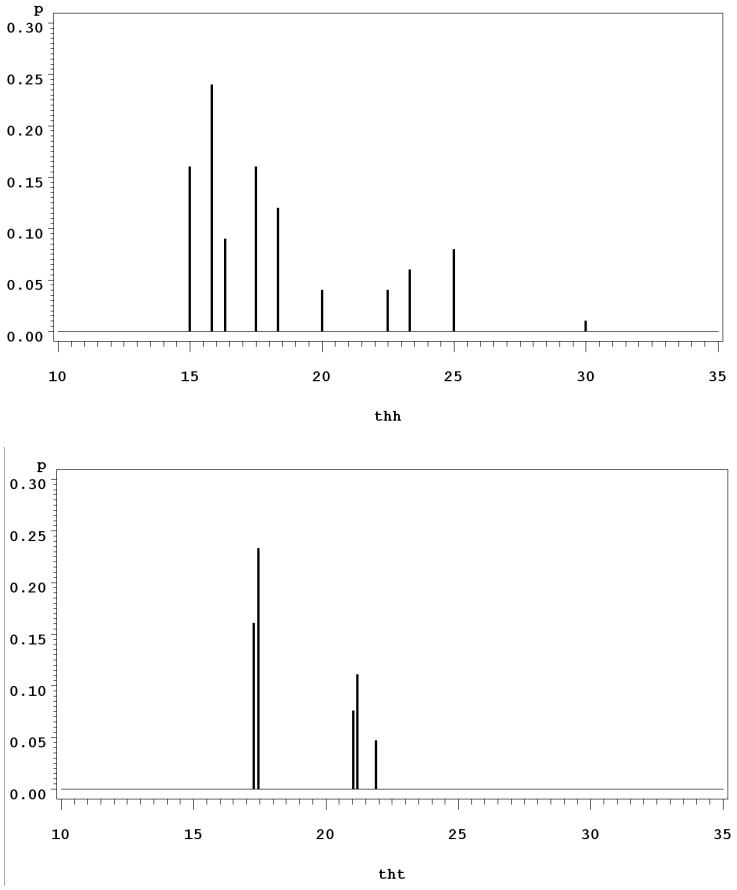


Figura 8.3. Distribuciones del estimador t_{HH} (muestreo pptr, arriba) y del estimador t_{HT} (muestreo ppt, abajo).

8.3.2 Selección de las probabilidades de inclusión

De forma comparable a como ocurría en muestreo con probabilidades desiguales y con reemplazamiento, en muestreo ppt el estimador tendrá más precisión cuánto más cerca estén las probabilidades de inclusión π_i de una relación proporcional a y_i.

Teorema 8.13 (probabilidades de inclusión proporcionales a y).

Si π_i es proporcional a y_i para todo $i = 1, ..., N$, es decir, $\pi_i = ky_i$, entonces $V(t_{HT}) = 0$.

Demostración.

Si $\pi_i = ky_i$, entonces$\dfrac{y_i}{\pi_i} = \dfrac{1}{k}$, constante para todo i . Sustituyendo este término en la expresión de la varianza en la forma de Yates-Grundy,

$$V(t_{HT}) = V(t_{HT})_{YG} = \frac{1}{2} \sum_{i,j,i\neq j}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 = 0.$$

Este resultado nos indica que las probabilidades de inclusión deben estar cerca de la proporcionalidad con y para obtener un estimador eficiente.

ISBN 9788496866133

Ejemplo 8.6.

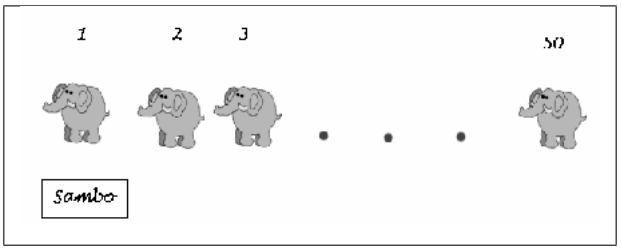


Figura 8.4. 50 elefantes de circo numerados para muestreo ppt.

Una crítica a los métodos de probabilidades desiguales aparece en un famoso ejemplo en (Basu, 1971). En un circo hay 50 elefantes y se necesita conocer el peso total de todos los elefantes por motivos de desplazamiento. Pesarlos a todos es muy difícil, pero el propietario del circo conoce el valor de hace tres años del peso w de un elefante llamado Sambo, y por lo tanto razona que el peso total de todos los elefantes debe ser aproximadamente $50 \cdot w$. Un estadístico, que a la sazón trabaja como payaso en el circo, propone utilizar muestreo con probabilidades desiguales con una muestra de tamaño $n = 1$, asignando las siguientes probabilidades de selección, que para $n = 1$ coinciden con las de inclusión:

$\pi_{Sambo} = p_{Sambo} = \dfrac{99}{100}$ para Sambo.

$\pi_i = p_i = \dfrac{1}{4900}$ para cada uno de los 49 elefantes restantes.

Así, la probabilidad de tener que pesar el elefante muestreado es muy pequeña, de $\dfrac{1}{4900}$, pues para Sambo se utilizaría el dato histórico w sin necesidad de pesarlo.

Con este planteamiento, si el elefante escogido es Sambo, el peso estimado de todos los elefantes según el estimador de Horvitz-Thompson es

$$t_{HT} = \frac{y_i}{\pi_i} = \frac{100}{99} w$$

es decir, se estima el peso de todos los elefantes juntos como si pesaran aproximadamente tanto como solamente Sambo.

Si el elefante escogido es otro cualquiera, por ejemplo Jumbo, con peso W, el peso estimado de todos los elefantes juntos es

$$t_{HT} = \frac{y_i}{\pi_i} = 4900W$$

lo cual también es inverosímil, pues se estima el peso de los 50 elefantes como 4900 veces el peso de uno de ellos.

Es obvio que el ejemplo no ataca las bondades teóricas de los métodos ppt, pero alerta sobre su utilización indebida. Aunque el ejemplo estaba planteado para alertar sobre los peligros de utilizar los métodos ppt en muestras pequeñas, hay que señalar lo siguiente:

- En el ejemplo, no hay manera de saber si las probabilidades de inclusión planteadas están correladas con la variable de interés, el peso. Para ser así, los elefantes diferentes de Sambo

ISBN 9788496866133

deberían pesar aproximadamente lo mismo, y además aproximadamente 4900 veces menos que Sambo, lo que es imposible. Al no ser aproximadamente proporcionales las π_i con las y_i, el método de muestreo ppt (o pptr, pues coinciden para $n = 1$) no es adecuado.

- Una muestra de tamaño $n = 1$ no puede representar bien una población heterogénea.

Supongamos el siguiente ejemplo alternativo, donde se representan cinco elefantes con su dieta.

Nombre	Sambo	Niko	Jumbo	Koku	Papo
Peso	5000	500	500	500	500
Dieta	4800	475	520	450	650

Tabla 8.11. Dieta y peso de 5 elefantes

Aunque la heterogeneidad de la población es grande, pues Sambo pesa 10 veces más que cualquiera de sus congéneres, si se utiliza muestreo proporcional a la dieta las estimaciones, aún para $n = 1$, son verosímiles (el verdadero peso total poblacional es 7000):

Nombre	Sambo	Niko	Jumbo	Koku	Papo
Peso	5000	500	500	500	500
$\pi_i = p_i$	0.701	0.069	0.076	0.066	0.088
t_{HT}	7130.2	7205.3	6581.7	7605.6	5704.2

Tabla 8.12. Probabilidades de inclusión y estimador

La precisión de las estimaciones se debe a que la relación entre π_i e y_i es aproximadamente proporcional. El coeficiente de correlación es $\rho_{\pi y} = 0.999$ y la recta de regresión pasa aproximadamente por el origen.

Como se ha visto, las probabilidades de inclusión deben estar cerca de la proporcionalidad con y para obtener un buen estimador. Pero lo único que se puede controlar de manera sencilla es la selección de una variable auxiliar x, de relación cercana a la proporcionalidad con y, que lleve a probabilidades de selección **en primera extracción** p_i aproximadamente proporcionales a y. Como se verá en el siguiente resultado, la proporcionalidad $p_i \propto y_i$ no implica forzosamente la proporcionalidad $\pi_i \propto y_i$.

Teorema 8.14 (muestreo ppt).

Supongamos que $n = 2$. Se realiza muestreo con probabilidades p_i en la primera extracción y en la segunda, excluyendo la primera unidad extraída, se extrae la segunda unidad con probabilidad proporcional respecto de la suma de las probabilidades de las unidades que restan (esto se podría llamar el **método básico** de muestreo ppt). Entonces,

ISBN 9788496866133

a) $\pi_i = p_i(1 + \sum_{j \neq i}^{N} \frac{p_j}{1-p_j})$.

b) $\pi_{ij} = p_i p_j (\frac{1}{1-p_i} + \frac{1}{1-p_j})$.

Demostración.

a) La probabilidad de inclusión de la unidad i será

$\pi_i = P$(sale i en 1ª extracción)$+P$(No sale i en 1ª extracción, sale i en 2ª extracción)=

$$=p_i + \sum_{j \neq i}^{N} P(\text{sale } j \text{ en 1ª extracción, } i \text{ en segunda})=$$

$$= p_i + \sum_{j \neq i}^{N} P(i \text{ en segunda } | \text{sale } j \text{ en 1ª extracción}) P(\text{sale } j \text{ en 1ª extracción}) =$$

$$= p_i + \sum_{j \neq i}^{N} \frac{p_i}{1-p_j} p_j = p_i(1 + \sum_{j \neq i}^{N} \frac{p_j}{1-p_j}).$$

Se ha utilizado el hecho de que $P(i$ en segunda $|$sale j en 1ª extracción$)=\frac{p_i}{1-p_j}$ por ser $1-p_j$ la suma de las probabilidades restantes habiendo salido j en 1ª extracción.

b) $\pi_{ij} = P(i$ en 1ª extracción$)P(j$ en 2ª$|$ i en 1ª $)+$

$$+P(j \text{ en 1ª extracción}) P(i \text{ en 2ª} | \ j \text{ en 1ª }) =$$

$$= p_i \frac{p_j}{1-p_i} + p_j \frac{p_i}{1-p_j}.$$

Para $n > 2$ se puede razonar del mismo modo y desarrollar las expresiones para π_i y π_{ij}. Lo más importante, es que, como se ve, **con el método básico p_i no es proporcional a π_i** .

Se han desarrollado métodos de muestreo con probabilidades desiguales y respetando el hecho de ser muestreo sin reemplazamiento, pero que cumplan la relación $\pi_i \propto p_i$.Por lo tanto, con estos métodos, si la variable auxiliar x tiene relación aproximada de proporcionalidad a y, y se construyen $p_i \propto x_i \propto y_i$, obtendremos aproximadamente $\pi_i \propto y_i$ y por lo tanto una gran precisión en el estimador. Se verán a continuación algunos de estos métodos.

Los principales objetivos de estos métodos de selección son :

a) Proporcionalidad $\pi_i \propto p_i$.

b) Sencillez de implementación del método .

c) Buenas propiedades en cuanto a la estimación (expresión sencilla para π_i).

d) Buenas propiedades en cuanto a la estimación de la varianza (expresión sencilla para π_{ij}).

ISBN 9788496866133

i) **Método de Brewer**

Presentaremos el caso en que $n = 2$, aunque existe una extensión del método para $n > 2$ que permite calcular las probabilidades de inclusión de forma recursiva. El procedimiento es:

1) La primera unidad se selecciona con probabilidad $p_i \frac{(1-p_i)}{(1-2p_i)} \left(\sum_{i=1}^{N} p_i \frac{(1-p_i)}{(1-2p_i)} \right)^{-1}$.

2) Sea la primera unidad seleccionada la unidad j. La segunda unidad se selecciona sin reemplazamiento, con probabilidad $\frac{p_i}{1-p_j}$.

El método exige $p_i < 0.5$ para todas las unidades. Para el caso de una muestra de tamaño n, ha de ser $p_i < \frac{1}{n}$ para todas las unidades de la población.

El método de Brewer verifica, para $n = 2$:

a) $\pi_i = 2p_i$.

b) $\pi_{ij} = 2p_i p_j \left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j} \right) \left(1 + \sum_{k=1}^{N} \frac{p_k}{1-2p_k} \right)^{-1}$.

ii) Método de Sampford

El método se puede utilizar para cualquier tamaño n.

1) La primera unidad se selecciona con probabilidad p_i.

2) La segunda unidad se selecciona con reemplazamiento, con probabilidad

$$\frac{p_i}{1-np_i} \left(\sum_{k=1}^{N} \frac{p_k}{1-np_k} \right)^{-1}.$$

3) Si la segunda unidad es similar a alguna que ya está en la muestra, se descarta. El paso 2 se repite haste tener n unidades distintas en la muestra.

Al igual que el método de Brewer, para poder aplicar el método ha de ser $p_i < \frac{1}{n}$ para todo i.

Existe un modo de calcular los π_i y los π_{ij}. En el caso de $n = 2$, se cumplen las mismas relaciones que en el método de Brewer:

a) $\pi_i = 2p_i$.

b) $\pi_{ij} = 2p_i p_j \left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j} \right) \left(1 + \sum_{k=1}^{N} \frac{p_k}{1-2p_k} \right)^{-1}$.

Además, el método de Sampford cumple $\pi_i \pi_j - \pi_{ij} \geq 0$, con lo que se asegura la positividad del estimador de la varianza del estimador de Horvitz-Thompson en la forma de Yates-Grundy.

iii) Método de Hanurav

Es un método para $n = 2$, aunque existe una extensión para $n > 2$ (método Hanurav-Vijayan). El procedimiento es:

ISBN 9788496866133

1) Se reordena la población y redefinen los p_i de manera que $p_1 \leq p_2 ... \leq p_N$.

2) Sea $\phi = 2(1-p_N)(p_N - p_{N-1})/(1-p_N-p_{N-1})$. Se realiza un sorteo de Bernoulli con probabilidad de éxito ϕ y de fracaso $1-\phi$.

3) Si el resultado es éxito, se selecciona la unidad N y la segunda unidad es seleccionada de entre $1, ..., N-1$ con probabilidades respectivas $p_1, ..., p_{N-1}$.

4) Si el resultado es fracaso, sean las probabilidades $\phi_i^* = p_i/(1-p_N-p_{N-1})$ para $i = 1, ..., N-1$ y $\phi_N^* = (p_N - \phi/2)/(1-\phi)$. Se seleccionan con reemplazamiento dos unidades con estas probabilidades. Si no coinciden, se aceptan. Si hay coincidencia, se rechazan y se realiza muestreo con reemplazamiento con probabilidades proporcionales a $\phi_i^{*^2}$. Si vuelven a coincidir, se repite el proceso desde el principio.

El método exige $p_i < 0.5$ para todas las unidades. Para el caso de una muestra de tamaño n, ha de ser $p_i < \dfrac{1}{n}$ para todas las unidades de la población. Existe un modo de calcular las π_i y las π_{ij}, y se cumple que $\pi_i \pi_j - \pi_{ij} \geq 0$.

En estos métodos, el estudio de casos con $n = 2$ está justificado desde el punto de vista en que las unidades a seleccionar puedan ser en realidad conglomerados de unidades de posteriores etapas, de modo que a menudo se elige una pequeña cantidad.

En general, los esquemas existentes requieren $p_i < \dfrac{1}{n}$ para todas las unidades de la población. Esto con frecuencia no se cumple, así que desde el punto de vista práctico se adoptan soluciones como:

- Estratificar previamente por grupos relacionados con la variable x. Así, estos estratos serán homogéneos respecto a x y por lo tanto en cada grupo las probabilidades p_i no diferirán excesivamente, llevando probablemente a $p_i < \dfrac{1}{n}$ para todas las unidades dentro de cada estrato. Aunque a veces la estratificación, al reducir el número de unidades, puede tener una consecuencia opuesta a la deseada: que surjan unidades con $p_i > \dfrac{1}{n}$.
- Como caso particular del anterior, se pueden considerar unidades autorepresentadas : si hay pocas unidades con $p_i > \dfrac{1}{n}$, puede forzarse a que entren en la muestra como si fueran estratos independientes, reduciendo el tamaño muestral n para el resto de la muestra. Los cálculos de estimadores para el total y varianzas son sencillos teniendo en cuenta la independencia en el muestreo sobre los estratos.
- El método básico de muestreo, que no tiene la restricción $p_i < \dfrac{1}{n}$, no aboca a una probabilidad de inclusión π_i proporcional a p_i, pero si la relación entre x e y es aproximadamente de proporcionalidad, puede dar lugar en todo caso a buenas estimaciones.
- Utilizar otros tipos de muestreo que aunque no son estrictamente sin reemplazamiento, dan lugar a mejores estimaciones que el muestreo con reemplazamiento (método de Narain , método de muestreo secuencial con mínimo reemplazamiento de Chromy), etc.

ISBN 9788496866133

- Otra posibilidad es variar las probabilidades de selección. Es decir, truncar superiormente por una constante los valores de la variable auxiliar x, de modo que se crea esta segunda variable x' como variable para las probabilidades de selección. Siempre que este truncamiento no afecte a que la correlación entre x' e y siga siendo alta, da lugar igualmente a buenas estimaciones.

Ejemplo 8.7.

Si se utilizan los datos de los $N = 209$ municipios de la provincia de Girona, y se considera como variable auxiliar el número de mujeres por municipio, y la variable de interés el número de hombres, se puede ver en la Tabla 8.3 que la probabilidades iniciales de extracción arrojan diferencias que pueden hacer inviables los métodos presentados anteriormente.

Los datos de la tabla 8.3 están ordenados de mayor a menor respecto al número de mujeres. Si se desea utilizar muestreo ppt con alguno de los métodos descritos, ha de ser $p_i < \frac{1}{n}$ para todas las observaciones. El caso peor es Girona capital, donde $p_i = 0.136$, y como ha de ser $0.136 < \frac{1}{n}$, entonces $n < \frac{1}{0.136} = 7.33$. Es decir, para poder utilizar alguno de los métodos de selección ppt que tienen supuestamente buenas propiedades, el tamaño máximo de la muestra habría de ser $n = 7$.

Municipio	Mujeres	p_i
Girona	37343	0.13639
Figueres	17308	0.06322
Blanes	14363	0.05246
Olot	14334	0.05235
Salt	10938	0.03995
...	...	...
Meranges	30	0.000109
La Vajol	33	0.000120
Sales de Llierca	27	0.000098

Tabla 8.13. Municipios de la provincia de Girona.

Para solucionar este problema, en primer lugar hay que remarcar que Girona capital tiene el doble de mujeres que el siguiente municipio en la lista. A partir de este segundo municipio, el número de mujeres no tiene cortes tan bruscos.

ISBN 9788496866133

La posibilidad más habitual es, como se hizo en el ejemplo 5.9, considerar Girona como una unidad autorrepresentada (obviamente en la práctica hay que tener en cuenta los costes asociados a esta decisión).

Las probabilidades de extracción (proporcionales al nº de mujeres sin Girona) se reconstruyen sin tener en cuenta Girona, y el primer municipio de la lista pasa a ser Figueres, con su nueva probabilidad de primera extracción $p_i = 0.0732$. En este momento, la restricción para poder utilizar los métodos descritos es $n < \dfrac{1}{0.0732} = 13.6$. Si se requiere un tamaño muestral más grande, se puede realizar la extracción con el método básico de muestreo sin reemplazamiento, o bien estratificar para intentar corregir $p_i < \dfrac{1}{n}$ en cada estrato, o bien recurrir a más unidades autorrepresentadas.

En cuanto a la estimación del total de hombres en la provincia de Girona, si se realiza muestreo ppt en la subpoblación que no contiene a Girona capital con cualquiera de los métodos expuestos, y t_{HT} es el estimador del total dentro de esa subpoblación, el estimador del total final será $t_{HT} + y_{Girona}$, donde y_{Girona} es el valor de y en Girona capital. La estimación de la varianza será $\widehat{V}(t_{HT})$ o $\widehat{V}(t_{HT})_{YG}$, pues al añadir el valor constante de la capital la varianza del estimador no cambia.

Ejemplo 8.8

Se realizará un estudio de simulación para comprobar el comportamiento del estimador de Horvitz-Thompson en la población descrita en el ejemplo anterior. En este estudio se toman $k = 100$ muestras de tamaño $n = 10$ cada una por muestreo proporcional al número de mujeres en cada municipio, sin reemplazamiento. El objetivo es estimar el número medio de hombres por municipio.

Se utilizará la exclusión de Girona capital, como se ha indicado en el anterior ejemplo, incorporándola después al valor del estimador. Así, el estimador del número medio de hombres final será $\frac{1}{N}(t_{HT} + y_{Girona})$, donde t_{HT} estima el total de hombres en la subpoblación excluida la capital Girona.

El paquete estadístico SAS utiliza el método de Hanurav por defecto. Este será el método de extracción aplicado.

Se ha realizado el mismo ejercicio de simulación utilizando los tipos de muestreo y estimación siguientes (en todos ellos se procede a la corrección por la unidad autorrepresentada): muestreo pptr con estimador de Hansen-Hurwitz, muestreo aleatorio simple con estimación de razón, y m.a.s. con estimación directa.

La tabla 8.4 presenta la media y desviación típica del valor del estimador obtenido sobre las 100 muestras, para cada uno de los métodos de estimación/muestreo.

Método	ppt	pptr	m.a.s.(estimador $\widehat{\overline{y}}$)	m.a.s.(estimador $\overline{y}_R$)
Media del Estimador	1219.10	1221	1322.17	1226
D. típica del Estimador	18	19.31	726	27

Tabla 8.14. Resumen de estimaciones bajo diferentes esquemas.

Se observa cómo el muestreo con probabilidades proporcionales al tamaño mejora incluso al m.a.s. con estimación de razón, en cuanto a la varianza del estimador. Si no se utiliza la información del número

ISBN 9788496866133

de mujeres por municipio (variable auxiliar), y se utiliza m.a.s. con el estimador $\widehat{\overline{y}}$ la estimación es mucho peor, con una desviación típica de 726.

La Figura 8.5 presenta los histogramas de los valores de los estimadores en las 100 muestras para los cuatro métodos empleados.

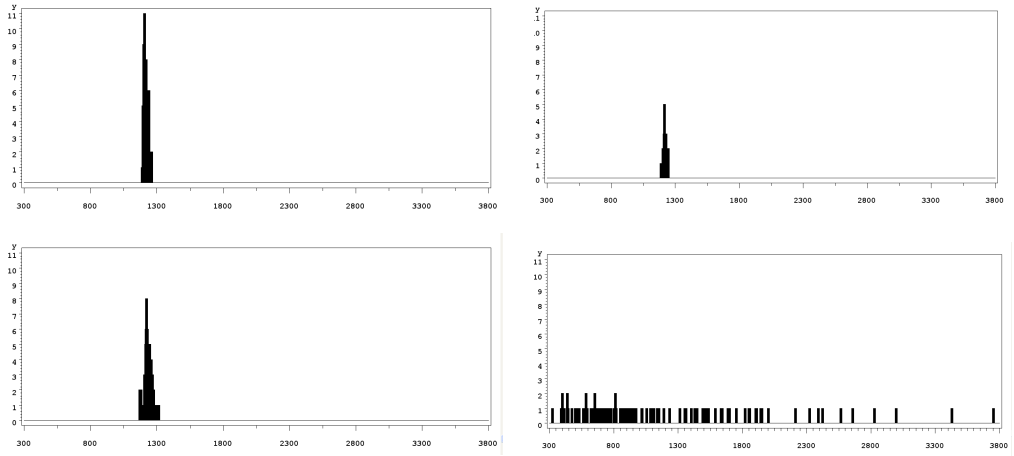


Figura 8.5. Histograma del valor de los estimadores en los métodos ppt (arriba, izda.), pptr (arriba, dcha.), m.a.s. con estimación de razón (abajo, izda.), m.a.s. con estimador media muestral (abajo, dcha.).

Se observa la gran variabilidad debida a no utilizar variable auxiliar en m.a.s., y la ganancia en precisión de los muestreos ppt respecto a m.a.s. , aún con estimación de razón. La comparación entre muestreo ppt y pptr es más sutil, pero existe (es más preciso, en general, muestreo sin reemplazamiento que con reemplazamiento para el mismo tamaño y método empleado).

Hay que tener en cuenta que diferentes tipos de muestreo y/o estimación requieren diferente información adicional. Aquellos que dan mejor resultado suelen requerir más información, que puede estar o no asociada a costes importantes.

En muestreo proporcional al tamaño sin reemplazamiento, se requiere conocer los valores de la variable auxiliar para todos los elementos de la población, previamente al muestreo.

En muestreo proporcional al tamaño con reemplazamiento se requiere la misma información, pero en promedio es menos costoso que el método anterior al poderse repetir unidades en el muestreo, y ahorrarnos la medición en estas unidades a partir de la segunda vez que son elegidas. El m.a.s. con estimación de razón requiere conocer la media poblacional de la variable auxiliar, y no forzosamente los valores de esta variable para toda la población, pero sí los valores de esta variable para las unidades que caen en la muestra.

Finalmente, el m.a.s. sin utilizar la variable auxiliar en la estimación no requiere de información adicional.

ISBN 9788496866133

8.4 Tablas de fórmulas

PROBABILIDADES DESIGUALES CON REEMPLAZAMIENTO			
Parámetro poblacional	$N\overline{y}$	$\overline{y}$	p
Estimador	$t_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$	$\frac{t_{HH}}{N}$	$\frac{t_{HH}}{N}$
Varianza	$\frac{1}{n}\left(\sum_{i=1}^{N}\frac{y_i^2}{p_i} - (N\overline{y})^2\right)$	$\frac{V(t_{HH})}{N^2}$	$\frac{V(t_{HH})}{N^2}$
Estimador de la Varianza	$\frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\frac{y_i^2}{p_i^2} - nt_{HH}^2\right)$	$\frac{\widehat{V}(t_{HH})}{N^2}$	$\frac{\widehat{V}(t_{HH})}{N^2}$

PROBABILIDADES DESIGUALES SIN REEMPLAZAMIENTO			
Parámetro poblacional	$N\overline{y}$	$\overline{y}$	p
Estimador	$t_{HT} = \sum_{i=1}^{n}\frac{y_i}{\pi_i}$	$\frac{t_{HT}}{N}$	$\frac{t_{HT}}{N}$
Varianza	$\sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}y_i^2 + 2\sum_{i<j}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}y_iy_j$	$\frac{V(t_{HT})}{N^2}$	$\frac{V(t_{HT})}{N^2}$
Varianza (Yates-Grundy)	$\sum_{i<j}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$	$\frac{V(t_{HT})_{YG}}{N^2}$	$\frac{V(t_{HT})_{YG}}{N^2}$
Estimador de la Varianza	$\sum_{i=1}^{n}\frac{(1-\pi_i)}{\pi_i^2}y_i^2 + 2\sum_{i<j}^{n}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\frac{y_iy_j}{\pi_i\pi_j}$	$\frac{\widehat{V}(t_{HT})}{N^2}$	$\frac{\widehat{V}(t_{HT})}{N^2}$
Estimador de la Varianza (Yates-Grundy)	$\sum_{i<j}^{n}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$	$\frac{\widehat{V}(t_{HT})_{YG}}{N^2}$	$\frac{\widehat{V}(t_{HT})_{YG}}{N^2}$

ISBN 9788496866133

8.5 Obtención de muestras con probabilidades desiguales con SAS

8.5.1 Muestreo ppt con reemplazamiento

La utilización del procedimiento surveyselect es sencilla. Supongamos que la variable auxiliar se denomina x. La sintaxis para obtener el archivo de muestra de tamaño $n = 10$ del archivo poblacional datos, con probabilidades proporcionales a x_i es:

```
proc surveyselect data=datos out=muestra method=pps_wr n=10 outhits;
size x;
run;
```

8.5.2 Muestreo ppt sin reemplazamiento

En este caso, debido a las razones expuestas en la teoría, es conveniente utilizar algún método alternativo al básico. El proc surveyselect del SAS utiliza por defecto el método de Hanurav-Vijayan, y tiene la virtud de aportar, dentro del mismo archivo muestral, las probabilidades de inclusión de primer y segundo orden, necesarias para el cálculo de estimadores.

Por ello se utilizará este procedimiento. Si la variable auxiliar es x, y se desea una muestra de tamaño $n = 3$ obtenida por muestreo ppt del archivo poblacional datos1, la sintaxis es :

```
proc surveyselect data=datos out=muestra method=pps n=3 jtprobs;
size x;
run;
```

El archivo muestral contiene, además de la información presente en el archivo datos, las probabilidades de inclusión (que se han solicitado a través de la opción jtprobs) . Presentamos abajo una parte de este archivo para aclarar cómo están dispuestas estas probabilidades.

Las π_i están representadas en la variable llamada SelectionProb. Las π_{1j} en la variable JtProb_1, para $j = 1, 2, 3$ (por lo tanto, en el ejemplo, $\pi_{13} = 0.00656$), las π_{2j} en la variable JtProb_2, etc. Es obvio que al ser muestreo sin reemplazamiento la probabilidad de seleccionar dos veces la misma unidad es 0 (es decir, $\pi_{11} = \pi_{22} = \pi_{33} = 0$).

x	Selection Prob	JtProb_1	JtProb_2	JtProb_3
1.76810	0.03232	0	.001479431	.003079121
3.71252	0.06786	.001479431	0	.006563197
7.72682	0.14123	.003079121	.006563197	0

Una cuestión importante al realizar muestreo ppt con el proc surveyselect (y en realidad con la mayoría de los métodos creados a tal efecto) es que las probabilidades de selección inicial p_i deben ser, para todas las observaciones, tales que $p_i < \frac{1}{n}$. Esto representa en muchos casos un

ISBN 9788496866133

impedimento, que se puede solucionar utilizando el método de selección proporcional al tamaño con mínimo reemplazamiento, o método de Chromy (method=pps_seq) y utilizar estimaciones tipo Hansen Hurwitz (es decir, como si fuera muestreo con rempazamiento) , pues las varianzas estimadas siempre serán una aproximación conservadora.

8.6 Estimación en muestreo con probabilidades desiguales con SAS

8.6.1 Muestreo ppt con reemplazamiento

Se utilizará la macro estimpptr para realizar las estimaciones.

En esta macro, se tiene la alternativa de aportar las probabilidades de selección de dos modos:

(1) En el mismo archivo muestral, como una variable llamada pi.

(2) (opcional) Aportando un archivo que contenga en la variable pi las probabilidades de selección p_i para cada unidad poblacional. En este caso la variable de identificación de cada observación debe tomar los valores enteros desde 1 hasta N. Esta variable de identificación también debe estar presente en el archivo muestral. Es decir, este archivo "poblacional" debe tomar la forma:

```
ID pi    ....
1  0.032
2  0.012
3  0.456
4  0.222
...
```

y en el archivo muestral también debe estar presente la misma variable de identificación, para las observaciones que estén en él.

La sintaxis de la macro estimpptr es

```
estimpptr(archivo1,archivo2,variabley,id,ngrande,n,indicador);
```

donde

archivo1 es el archivo que contiene la muestra.

archivo2 es el archivo que contiene las probabilidades para todos los elementos de la población (opcional).

variabley es la variable de interés.

id es la variable de identificación

ngrande es el tamaño poblacional N.

ISBN 9788496866133

n es el tamaño muestral.

indicador estipula el modo en que aparecen las probabilidades de selección pi:

1 como la variable pi en el archivo de muestra.

2 como la variable pi en el archivo2.

Supongamos que se desea obtener estimaciones para los datos muestrales presentes en el archivo muestra1, con $n = 10$ observaciones. Asumamos que este archivo contiene ya las probabilidades p_i (indicador=1). La población tiene tamaño $N = 100$ y la variable de interés se llama z en el archivo de la muestra. La variable de identificación no necesita aparecer. La sintaxis correspondiente es:

```
%estimpptr(muestra1,.,z,.,100,10,1);
```

Se ha puesto el símbolo missing (".") en lugar del archivo2 pues no se dispone de este archivo (en realidad cuando el indicador es 1 la macro no usa el archivo2, con lo cual se puede poner cualquier nombre).

Supongamos que, en este mismo ejemplo, el archivo muestral no contiene las probabilidades p_i, pero éstas están en el archivo poblacional llamado datos1 junto con la variable de identificación codigo. La sintaxis correspondiente en este caso es:

```
%estimpptr(muestra1,datos1,z,codigo,100,10,2);
```

La macro estimpptr presenta en la ventana LOG los estimadores de Hansen Hurwitz de la media y total, y sus varianzas e intervalos de confianza al 95%.

8.6.2 Muestreo ppt sin reemplazamiento

Se utilizará la macro estimppt para realizar las estimaciones.

Para utilizar esta macro se solicita que el archivo muestral contenga las probabilidades de inclusión de primer y segundo orden, tal y como aparecen en el archivo de salida del procedimiento surveyselect (con lo cual se supone que los datos muestrales provienen de la selección realizada a través del proc surveyselect, aunque se puede crear un archivo similar con un paso data si se dispone de las probabilidades de inclusión).

Se asume por lo tanto que los datos muestrales tienen la forma descrita en el apartado de selección ppt presentado anteriormente.

La sintaxis de la macro estimppt es la siguiente:

```
%estimppt(archivo1,variabley,n,ngrande);
```


ISBN 9788496866133

donde

archivo1 es el archivo que contiene la muestra. Debe contener los datos muestrales y las probabilidades de inclusión

de primer y segundo orden, tal y como aparecen en la salida de un proc surveyselect. En general es el archivo proveniente de un proc surveyselect.

variabley es la variable de interés

ngrande es el tamaño poblacional N

n es el tamaño muestral

Como ejemplo, asumiendo que el archivo muestral mues1 proviene del proc surveyselect con el método pps, la variable de interés es la variable alfa, el tamaño muestral es $n = 10$ y el tamaño poblacional es $n = 100$, la sintaxis sería:

```
%estimppt(mues1,alfa,10,100);
```

La macro estimppt presenta en la ventana LOG los estimadores de Horwitz Thompson de media y total, junto con sus varianzas estimadas (la usual de Horwitz-Thompson y la de Yates-Grundy) y los intervalos de confianza al 95% asociados a esas varianzas.

ISBN 9788496866133

8.7 Ejercicios resueltos

Ejercicio 7.1

Una población tiene 3 unidades con valores en 3, 2 y 7 en la variable y. Los valores respectivos de la variable x para esas observaciones son 6, 5 y 13. Suponiendo $n = 2$:

a) En muestreo pptr proporcional a la variable x, presentar la distribución del estimador del total poblacional. Comprobar que es insesgado.

b) Hacer lo mismo suponiendo muestreo ppt (con el método básico).

---

a) Las probabilidades respectivas de las tres observaciones son $p_1 = \dfrac{x_1}{\sum x_i} = \dfrac{6}{24} = 0.25$, $p_2 = 0.208$ y $p_3 = 0.542$.

Para cada muestra (i, j), al ser muestreo con reemplazamiento, la probabilidad es $p_i p_j$.

Las muestras posibles (al ser muestreo con reemplazamiento se tiene en cuenta el orden), sus probabilidades y el valor del estimador del total son:

Muestra	p	t_{HH}
$(1,1)$	0.062	12
$(1,2)$	0.052	10.8
$(1,3)$	0.135	12.4
$(2,1)$	0.052	10.8
$(2,2)$	0.043	9.6
$(2,3)$	0.112	11.2
$(3,1)$	0.135	12.4
$(3,2)$	0.112	11.2
$(3,3)$	0.293	12.9

donde por ejemplo, para la muestra $(1,2)$ se han calculado las probabilidades así:

$P((1,2)) = p_1 p_2 = 0.25 \cdot 0.208 = 0.052$.

Y el valor del estimador es

$$t_{HH} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i} = \frac{1}{2}\left(\frac{3}{0.25} + \frac{2}{0.208}\right) = 10.8.$$

La distribución del estimador es por lo tanto:

ISBN 9788496866133

t_{HH}	$p(t_{HH})$
9.6	0.043
10.8	0.105
11.2	0.225
12	0.062
12.4	0.271
12.9	0.293

La esperanza del estimador es

$E(t_{HH}) = 0.043 \cdot 9.6 + ... + 0.293 \cdot 12.9 = 12 = N\overline{y}$, con lo cual t_{HH} es insesgado como estimador del total.

b) Como $n = 2$, se calculan las probabilidades de inclusión así:

$\pi_i = p_i(1 + \sum_{j \neq i}^{N} \frac{p_j}{1 - p_j})$.

$\pi_{ij} = p_i p_j(\frac{1}{1 - p_i} + \frac{1}{1 - p_j})$.

Calcularemos en primer lugar cada $\frac{p_j}{1 - p_j}$ y como cada $\frac{1}{1 - p_j} = \frac{p_j}{1 - p_j} + 1$:

$j = 1 : \frac{0.25}{1 - 0.25} = 0.333;\ \frac{1}{1 - 0.25} = 1.333.$

$j = 2 : \frac{0.208}{1 - 0.208} = 0.263;\ \frac{1}{1 - 0.208} = 1.26.$

$j = 3 : \frac{0.542}{1 - 0.542} = 1.183;\ \frac{1}{1 - 0.542} = 2.18.$

Así,

$\pi_1 = 0.25(1 + 0.263 + 1.183) = 0.6115.$

$\pi_2 = 0.208(1 + 0.333 + 1.183) = 0.5233.$

$\pi_3 = 0.542(1 + 0.263 + 0.333) = 0.865.$

y

$\pi_{12} = 0.25 \cdot 0.208(1.333 + 1.26) = 0.135.$

$\pi_{13} = 0.25 \cdot 0.542(1.333 + 2.18) = 0.476.$

$\pi_{23} = 0.208 \cdot 0.542(1.26 + 2.18) = 0.389.$

Se presentan en la tabla las muestras (sin tener en cuenta el orden, pues es muestreo sin reemplazamiento), con sus probabilidades y el estimador del total.

ISBN 9788496866133

El estimador del total de Horvitz Thompson es $t_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$. Por ejemplo, para la muestra $\{1,2\}$ es $t_{HT} = \frac{3}{0.6115} + \frac{2}{0.5233} = 8.727$.

Muestra	π_{ij}	t_{HH}
$\{1,2\}$	0.135	8.727
$\{1,3\}$	0.476	13
$\{2,3\}$	0.389	11.9

Se comprueba que el estimador es insesgado, pues

$E(t_{HH}) = 0.135 \cdot 8.727 + 0.476 \cdot 13 + 0.389 \cdot 11.9 = 12.$

Ejercicio 7.2

Se realiza un muestreo estratificado para determinar una estimación para el número de plantas con un virus en una finca. Esta se divide en parcelas de la parte Norte y de la parte Sur. En la parte Norte hay 4 parcelas, con un total de 200 plantas, de las cuales se escogen dos con reemplazamiento , asignando probabilidades de elección proporcionales el número de plantas de cada parcela. En la primera parcela muestreada hay 21 plantas enfermas, y 45 plantas. En la segunda hay 30 plantas enfermas y 58 plantas.

En la parte Sur hay 3 parcelas , con un total de 145 plantas, de las cuales se escogen 2 con probabilidades proporcionales al número de plantas, pero esta vez sin reemplazamiento. En la primera parcela muestreada hay 17 plantas enfermas, y 45 plantas. En la segunda hay 19 plantas enfermas y 47 plantas.

Suponiendo normalidad del estimador,

a) Dar un I.C. al 95% para el número total de plantas enfermas en la finca.

b) Dar un I.C. al 95% para la proporción de plantas enfermas en la finca.

---

a) Se trata de muestreo estratificado, por lo cual se realizarán las estimaciones por separado para cada estrato (norte y Sur), para construir finalmente un estimador global.

Estrato Norte:

Se trata de muestreo pptr. La probabilidad de la primera parcel escogida es $p_1 = \frac{45}{200} = 0.225$, pues es proporcional al número de plantas. La probabilidad de la segunda es $p_2 = \frac{58}{200} = 0.29$.

El estimador del total de plantas enfermas en ese estrato es $t_{HH} = \frac{1}{2}\left(\frac{21}{0.225} + \frac{30}{0.29}\right) = 98.39$.

La varianza estimada de ese estimador es

$$\widehat{V}(t_{HH}) = \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} \frac{y_i^2}{p_i^2} - n t_{HH}^2\right) = \frac{1}{2(2-1)}\left(\frac{21^2}{0.225^2} + \frac{30^2}{0.29^2} - 2 \cdot 98.39^2\right) = 25.73.$$

ISBN 9788496866133

Estrato Sur:

En este caso es muestreo ppt. Para calcular los estimadores y varianzas es necesario calcular las probabilidades de inclusión. Suponiendo que se ha realizado el muestreo mediante el método básico, se tiene que

$$\pi_i = p_i(1 + \sum_{j\neq i}^{N} \frac{p_j}{1-p_j})$$

y

$$\pi_{ij} = p_i p_j(\frac{1}{1-p_i} + \frac{1}{1-p_j}).$$

Las probabilidades iniciales son $p_1 = \frac{45}{145} = 0.31$ y $p_2 = \frac{47}{145} = 0.324$.Por lo tanto $p_3 = \frac{53}{145} = 0.366$.

Así,

$$\pi_1 = p_1(1 + \frac{p_2}{1-p_2} + \frac{p_3}{1-p_3}) = 0.637$$

y

$$\pi_2 = p_2(1 + \frac{p_1}{1-p_1} + \frac{p_3}{1-p_3}) = 0.656.$$

y

$$\pi_{12} = p_1 p_2(\frac{1}{1-p_1} + \frac{1}{1-p_2}) = 0.294.$$

El estimador de Horvitz-Thompson del total en este estrato es

$$t_{HT} = \frac{17}{0.637} + \frac{19}{0.656} = 55.65.$$

y su varianza estimada:

$$\widehat{V}(t_{HT}) = \sum_{i=1}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + 2\sum_{i<j}^{n} \frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j} =$$

$$= \frac{(1-0.637)}{0.637^2} 17^2 + \frac{(1-0.656)}{0.656^2} 19^2 + 2\frac{(0.294 - 0.637 \cdot 0.656)}{0.294} \frac{17}{0.637} \cdot \frac{19}{0.656} = -104.25.$$

Al ser negativo se utilizará el estimador de la varianza en la forma de Yates-Grundy:

$$\widehat{V}(t_{HT})_{YG} = \sum_{i<j}^{n} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 = \frac{(0.637 \cdot 0.656 - 0.294)}{0.294} \left(\frac{17}{0.637} - \frac{19}{0.656}\right)^2 = 2.18.$$

El estimador global se construirá sumando las estimaciones del total en cada estrato:

$$\widehat{N\overline{y}} = 98.39 + 55.65 = 154.04.$$

Su varianza será:

$$V(\widehat{N\overline{y}}) = \widehat{V}(t_{HH}) + \widehat{V}(t_{HT}) = 25.73 + 2.18 = 27.91.$$

El intervalo de confianza al 95% para el total de plantas enfermas en la finca es

$$(154.04 - 1.96\sqrt{27.91}, 154.04 + 1.96\sqrt{27.91}) = (143.6, 164.4).$$

b) Basta aplicar que la proporción estimada será $\frac{\widehat{N\overline{y}}}{345} = 0.446$, y su varianza $\widehat{V}(\frac{\widehat{N\overline{y}}}{345}) = \frac{1}{345^2} V(\widehat{N\overline{y}}) = 0.000234$.

ISBN 9788496866133

El intervalo de confianza es

$(0.446 - 1.96\sqrt{0.000234}, 0.446 - 1.96\sqrt{0.000234}) = (0.416, 0.476)$.

Ejercicio 7.3

Se desea estimar el número medio de alumnos por aula que asisten a clase habitualmente en cierta Facultad. Hay 6 clases con diferente número de alumnos matriculados. Se escogen 3 clases con reposición y con probabilidad proporcional al número de alumnos matriculados. Si los números respectivos de alumnos matriculados son 57,63,80,44,60,62, se pide:

a) Utilizando una tabla de números aleatorios, se obtienen los números 120,4,230. ¿A qué clases corresponden si utilizamos el método acumulativo de selección ?

b) Suponiendo que el número de alumnos que asisten habitualmente a clase en el orden de las tres clases mencionadas es respectivamente de 52,38 y 28, estimar el número total de alumnos que asisten a clase habitualmente en la Facultad y estimar la varianza del estimador.

c) Estimar el número de alumnos que asisten a clase habitualmente, por aula, en la Facultad y estimar la varianza del estimador..

---

a) Se construye la tabla de totales acumulados:

Observación	x_i	T_i
1	57	57
2	63	120
3	80	200
4	44	244
5	60	304
6	62	366

Los números 120, 4 y 230, éstos pertenecen a los intervalos respectivos correspondientes a las observaciones 2,1 y 4 (los intervalos son cerrados por la derecha).

b) Para las observaciones 2, 1 y 4, se tiene $p_2 = \dfrac{63}{366} = 0.172$, $p_1 = \dfrac{57}{366} = 0.155$, y $p_4 = \dfrac{44}{366} = 0.12$.

El total se estima a través del estimador de Hansen Hurwitz:

$$t_{HH} = \frac{1}{3}\left(\frac{52}{0.172} + \frac{38}{0.155} + \frac{28}{0.12}\right) = 260.$$

La varianza se estima por

$$\widehat{V}(t_{HH}) = \frac{1}{3(3-1)}\left(\left(\frac{52}{0.172}\right)^2 + \left(\frac{38}{0.155}\right)^2 + \left(\frac{28}{0.12}\right)^2 - 3 \cdot 260^2\right) = 524.8.$$

c) Se trata de estimar la media, pues las unidades elementales son las aulas en este ejercicio, por lo tanto

ISBN 9788496866133

$\widehat{\overline{y}} = \frac{1}{N} t_{HH} = 43.33$ y

$\widehat{V}(\frac{1}{N} t_{HH}) = \frac{1}{N^2} \widehat{V}(t_{HH}) = 14.57.$

Ejercicio 7.4

Se dispone de una población con 3 observaciones y se compararán ciertos procesos de muestreo y estimación con $n = 2$.

y	x
2	3
3	5
4	6

a) Calcular la varianza exacta del estimador del total si se utiliza muestreo pptr con variable auxiliar x.

b) Calcular la varianza exacta del estimador del total bajo m.a.s. de tamaño $n = 2$.

c) Calcular el error cuadrático medio exacto del estimador bajo m.a.s. con estimación de razón, suponiendo conocida $\overline{x}$.

---

a) Las probabilidades respectivas de cada observación son $p_1 = \frac{3}{14} = 0.214$, $p_2 = \frac{5}{14} = 0.357$, $p_3 = \frac{6}{14} = 0.429$.

La varianza del estimador de Hansen Hurwitz es:

$$V(t_{HH}) = \frac{1}{n} \left(\sum_{i=1}^{L} \frac{y_i^2}{p_i} - (N\overline{y})^2 \right) = \frac{1}{2} \left(\frac{2^2}{0.214} + \frac{3^2}{0.357} + \frac{4^2}{0.429} - 9^2 \right) = 0.10.$$

b) La cuasivarianza poblacional es $S_y^2 = 1$, y entonces $V(\widehat{\overline{y}}) = \frac{3-2}{3 \cdot 2} \cdot 1 = 0.166$.

c) La estimación de razón del total es $N\overline{y}_R = N \frac{\widehat{\overline{y}}}{\widehat{\overline{x}}} \overline{x} = 14 \frac{\widehat{\overline{y}}}{\widehat{\overline{x}}}$. Se calculará el valor del estimador para cada muestra posible, para calcular su esperanza y sesgo, y su varianza.

Para la muestra $\{1, 2\}$, se tiene $N\overline{y}_R = 14 \frac{2.5}{4} = 8.75$; para $\{1, 3\}$, $N\overline{y}_R = 9.333$; para $\{2, 3\}$, $N\overline{y}_R = 8.909$.

La esperanza del estimador es $E(N\overline{y}_R) = \frac{1}{3}(8.75 + 9.333 + 8.909) = 8.9974$.

El sesgo es por lo tanto $N\overline{y} - E(N\overline{y}_R) = 0.0026$.

La varianza del estimador es

$$V(N\overline{y}_R) = \frac{1}{3}((8.75 - 8.9974)^2 + (9.333 - 8.9974)^2 + (8.909 - 8.9974)^2) = 0.06055.$$

El error cuadrático medio es $ECM(N\overline{y}_R) = V(N\overline{y}_R) + sesgo^2(N\overline{y}_R) = 0.06055 + 0.0026^2 \simeq 0.06055$.

ISBN 9788496866133

Ejercicio 7.5

De las 25 granjas de un pueblo se selecciona una muestra aleatoria de tamaño n mediante el siguiente procedimiento: la primera extracción se realiza con probabilidades proporcionales al tamaño de las granjas y las $n-1$ restantes con probabilidades iguales, siendo todas las extracciones sin reposición.

a) Calcular la probabilidad de que la granja i aparezca en la muestra.

b) Calcular una estimación insesgada del dinero que entre todas las granjas del pueblo dedican anualmente a comprar piensos, con los siguientes datos: $n = 3$, tamaño total=100 y

Granja muestral	1	2	3
Dinero que emplea al año	61	30	50
Tamaño	5	3	6

c) Establecer la desviación típica de la estimación del apartado b).

---

a) Hay que calcular las probabilidades de inclusión con ese método de muestreo.

Llamando x_i al tamaño de la granja i, la primera extracción se realiza con probabilidades respectivas $p_i = \frac{x_i}{\sum_{i=1}^{N} x_i} = \frac{x_i}{N\overline{x}}$. La probabilidad de que la granja i no salga en la 1ª extracción es por lo tanto $1 - \frac{x_i}{N\overline{x}}$.

La probabilidad de que la granja i esté en la muestra es la probabilidad de que salga en primera extracción más la probabilidad de que salga en alguna de las restantes:

$\pi_i = P(i \in 1^{\text{a}} \text{ extracción}) +$

$P(i \notin 1^{\text{a}} \text{ extracción y sale en la } 2^{\text{a}}) +$

$+P(i \notin 1^{\text{a}} \text{ extracción}, i \notin 2^{\text{a}} \text{ extracción y sale en la } 3^{\text{a}}) +$

...

$+P(i \notin 1^{\text{a}} \text{ extracción ni en ninguna de las restantes y sale en la última}) =$

$$= \frac{x_i}{N\overline{x}} + (1 - \frac{x_i}{N\overline{x}})(\frac{1}{N-1} + \frac{N-2}{N-1}\frac{1}{N-2} + ... + \frac{1}{N-1}).$$

Se observa que todos los sumandos se simplifican y queda $\frac{1}{N-1}$ en cada sumando. Como hay $n-1$ sumandos, queda:

$$\pi_i = \frac{x_i}{N\overline{x}} + (1 - \frac{x_i}{N\overline{x}})(\frac{n-1}{N-1}).$$

Otra manera de calcular el segundo término es la siguiente: Una vez extraída la primera granja, las $n-1$ restantes se extraen con probabilidades iguales y sin reemplazamiento de las $N-1$ que quedan, por lo cual se trata de muestreo aleatorio simple sin reemplazamiento de $n-1$ en una población de $N-1$, y se sabe que la probabilidad de inclusión de cada observación i en esta submuestra es, por ser m.a.s., $\pi_i^* = \frac{n-1}{N-1}$.

M.J. Portela, M. Villeta
ISBN 9788496866133

b) Se utilizará el estimador de Horvitz-Thompson, por ser muestreo ppt sin reemplazamiento.

Las probabilidades de inclusión son:

$$\pi_1 = \frac{5}{100} + (1 - \frac{5}{100})\frac{2}{24} = 0.12917.$$

$$\pi_2 = \frac{3}{100} + (1 - \frac{3}{100})\frac{2}{24} = 0.111083.$$

$$\pi_3 = \frac{6}{100} + (1 - \frac{6}{100})\frac{2}{24} = 0.13833.$$

Luego:

$$t_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \frac{61}{0.129} + \frac{30}{0.111} + \frac{50}{0.138} \simeq 1104.38$$

c) Para estimar la varianza, hay que calcular las probabilidades de inclusión de segundo orden. Se realizará un cálculo de probabilidades similar al obtenido anteriormente, dependiendo de lo ocurrido en cada extracción:

$$\pi_{ij} = P(i \in 1^{\text{a}}, j \in 2^{\text{a}} \text{ o } 3^{\text{a}}) + P(j \in 1^{\text{a}}, i \in 2^{\text{a}} \text{ o } 3^{\text{a}}) + P(i, j \in 2^{\text{a}} \text{ o } 3^{\text{a}}) =$$

$$= \frac{x_i}{N\overline{x}}(\frac{2}{N-1}) + \frac{x_j}{N\overline{x}}(\frac{2}{N-1}) + (1 - \frac{x_i}{N\overline{x}} - \frac{x_j}{N\overline{x}})(\frac{1}{N-1}\frac{1}{N-2}2).$$

En el último sumando se ha calculado la probabilidad de que i pertenezca a la 2^{a} extracción y j a la 3^{a} +la probabilidad de que j pertenezca a la 2^{a} extracción e i a la 3^{a} . Se puede utilizar también que la probabilidad de inclusión de dos observaciones en m.a.s. es $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$. En este caso, $N^* = N - 1$ y queda $\pi^*_{ij} = \frac{2(2-1)}{(N-1)(N-2)} = \frac{1}{N-1}\frac{1}{N-2}2$.

π_{ij} se puede simplificar factorizando y queda:

$$\pi_{ij} = \frac{2}{N\overline{x}(N-1)(N-2)}[(N-3)(x_i + x_j) + N\overline{x}].$$

De este modo, queda:

$$\pi_{12} = \frac{2}{100(24)(23)}[(25-3)(5+3) + 100] = 0.01$$

y análogamente,

$$\pi_{13} = 0.01239$$

y

$$\pi_{23} = 0.010797.$$

Ahora se estima la varianza a través del estimador usual de Horvitz-Thompson:

$$\widehat{V}(t_{HT}) = \sum_{i=1}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + 2\sum_{i<j}^{n} \frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j} \simeq 28468.$$

Por lo tanto, la desviación típica del estimador es

$$\sqrt{\widehat{V}(t_{HT})} = 168.7.$$

ISBN 9788496866133

Ejercicio 7.6

Se realiza un estudio para estimar, en un terreno de 10 hectáreas, qué superficie en total está cubierta por arbolado. El terreno está dividido en 20 parcelas de distinto tamaño, y se seleccionan de entre éstas 5 con reemplazamiento y probabilidad proporcional al tamaño de la parcela. Se obtiene que la proporción de terreno cubierto por arbolado en cada parcela de las parcelas muestreadas, es respectivamente 0.2, 0.3, 0.25, 0.5, y 0.8. Calcular un intervalo de confianza al 95% para la superficie total en el terreno cubierta por árboles.

Se ha realizado muestreo con probabilidades proporcionales al tamaño de las parcelas. Si se denomina por x_i al tamaño de la parcela i en hectáreas, la probabilidad de selección para esa parcela es $p_i = \dfrac{x_i}{\sum x_i} = \dfrac{x_i}{10}$.

El estimador de Hansen Hurwitz del total de terreno cubierto por arbolado es

$$t_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} = \frac{10}{5}\sum_{i=1}^{n}\frac{y_i}{x_i}$$

donde y_i representa el terreno cubierto por arbolado en la parcela i.

Los datos que se dan en el problema son las proporciones de terreno cubierto por arbolado en cada parcela, o sea,

$$\frac{\text{Terreno cubierto por arbolado en parcela } i}{\text{Total de terreno en parcela } i} = \frac{y_i}{x_i}.$$

Así que :

$$t_{HH} = \frac{10}{5}\sum_{i=1}^{n}\frac{y_i}{x_i} = \frac{10}{5}(0.2 + 0.3 + 0.25 + 0.5 + 0.8) = 4.10 \text{ hectáreas.}$$

La varianza estimada de t_{HH} es:

$$\widehat{V}(t_{HH}) = \frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\frac{y_i^2}{p_i^2} - nt_{HH}^2\right) = \frac{1}{n(n-1)}\left(\sum_{i=1}^{n}10^2(\frac{y_i}{x_i})^2 - nt_{HH}^2\right) =$$

$$= \frac{1}{5(4-1)}\left(10^2(0.2^2 + 0.3^2 + 0.25^2 + 0.5^2 + 0.8^2) - 5\cdot 4.10^2\right) = 1.6133$$

por lo cual un intervalo de confianza al 95% será

$$(4.10 - 1.96\sqrt{1,6133}, 4.10 + 1.96\sqrt{1,6133}) = (1.61, 6.59).$$

Ejercicio 7.7

Utilizar en SAS las macros estimpptr y estimppt para calcular los estimadores y sus varianzas en cada estrato del ejercicio 7.2.

Para estimación en caso de muestreo con reemplazamiento (estrato Norte) se utilizará la macro estimpptr, creando primero un archivo con la información muestral:

```
data uno;
input id y pi;
cards;
```


ISBN 9788496866133

```
1 21 0.225
2 30 0.29
;
%estimpptr(uno,.,y,id,4,2,1);
```

Obteniendo los resultados del ejercicio 7.2, es decir: $t_{HH} = 98.49$ y $\widehat{V}(t_{HH}) = 25.57$.

En el caso del estrato Sur, donde se ha realizado muestreo sin reemplazamiento, se utiliza la macro estimppt. En el archivo creado hay que incorporar las probabilidades de inclusión:

```
data dos;
input y selectionprob jtprob_1 jtprob_2;
cards;
17 0.637 0 0.294
19 0.656 0.294 0
;
%estimppt(dos,y,2,3);
```

Obteniendo los mismos resultados que en el ejercicio 7.2.

Ejercicio 7.8

En el archivo SAS guisa, donde la variable super indica la superficie dedicada al cultivo del guisante y la variable produ la producción de guisante en provincias españolas en 1998, se realizará un estudio para comprobar la eficacia de los estimadores pptr y ppt, estimando la producción total de guisante utilizando muestreo con probabilidad proporcional al tamaño (en este caso, la superficie dedicada al cultivo).

a) Crear en el archivo guisa la variable pi, que indica la probabilidad de ser escogida para cada provincia. Observar el listado del archivo, ordenado por esa probabilidad. ¿Cuáles son las provincias con mayor probabilidad de ser escogidas? ¿Y las que menos?.

b) Extraer 5 muestras con probabilidad proporcional a la superficie con reemplazamiento, de tamaño $n = 10$, con las semillas 1234, 1235, 1236, 1237, 1238. Utilizar la macro estimpptr para estimar el total y observar los valores del estimador. Observar también los estimadores de la varianza.

c) Realizar el mismo proceso con muestreo ppt . Utilizar la macro estimppt.

d) Realizar muestreo aleatorio simple 5 veces con las mismas semillas , observando el valor de la media muestral. Calcular en estos casos también el estimador de razón y regresión mediante la macro estimrazreg. Comparar con los resultados anteriores.

---

a) Para crear la variable de probabilidad se realiza un proc means para obtener el total de superficie. A continuación se crea la variable pi en un paso data, junto con la variable de identificación necesaria en la ejecución de la macro. La ordenación se realiza con el proc sort y el listado con el proc print.

```
proc means data=guisa sum;run;
data guisa;set guisa;pi=super/74173;id=_n_;
proc sort data=guisa;by pi;
proc print data=guisa;run;
```


ISBN 9788496866133

En cuanto a superficie de producción, las provincias con menor superficie y por lo tanto con menor pi son Asturias, Cantabria y varias más (todas con superficie 0). Las que más superficie dedican son Ciudad Real y Valladolid. Por lo tanto serán aquellas con mayor probabilidad de salir en la muestra.

b) El programa base es:

```
proc surveyselect data=guisa out=muestra n=10 method=pps_wr seed=1234 outhits;
size pi;
run;
%estimpptr(muestra,.,produ,id,51,10,1);
```

En esta primera muestra se obtiene, por ejemplo, Álava 3 veces y Badajoz 2, (estaban en la lista ordenada en las posiciones 37 y 45, con lo que es normal que hayan sido escogidas).

El estimador en esta muestra es 103.625. Para las muestras restantes, con las diferentes semillas, es 56395, 62945, 75885, 51549.

c) Al intentar obtener muestras proporcionales a la variable super con el proc surveyselect, con tamaño $n = 10$, se obtiene un mensaje de error, pues hay provincias con $p_i > \frac{1}{n} = 0.10$. Una manera de tratar esto es seleccionar obligatoriamente la unidad de p_i más alto y calcular las nuevas p_i con las $n - 1 = 9$ unidades restantes. Si existen unidades con $p_i > \frac{1}{n-1}$ se repite el proceso, seleccionando obligatoriamente la de valor p_i más alto, comprobando si existen en el nuevo archivo unidades con $p_i > \frac{1}{n-2}$ y así sucesivamente. Este problema suele ocurrir a menudo en poblaciones donde la variable es asimétrica.

Con SAS, una manera de hacerlo es ir eliminando las observaciones seleccionadas del archivo y recalcular p_i cada vez (utilizando un proc means para calcular la suma de los valores de la variable super y haciendo pi=super/suma en un paso data). Se comprueba entonces en cada caso, con un proc print, si existen observaciones con $p_i > \frac{1}{n-k}$ en cada caso, donde se han eliminado k del archivo.

Otra manera más rápida es ir eliminando una a una las observaciones con mayor valor en super en un paso data y ejecutando cada vez el proc surveyselect reduciendo n en una unidad cada vez. En el momento que no hay mensaje de error, es que todas las observaciones que quedan cumplen $p_i < \frac{1}{n-k}$.

En nuestro caso, se seleccionan obligatoriamente y por orden sucesivo, según el esquema presentado, las provincias de Ciudad Real, Valladolid, Sevilla y Palencia (las que mayor superficie dedican al cultivo del guisante). Finalmente el tamaño muestral restante es $10 - 4 = 6$. Todas las provincias restantes cumplen $p_i < \frac{1}{6}$. El paso data para eliminar esas observaciones es:

```
data guisa2;set guisa;if super>6000 then delete;run;
```

Entonces se realizará muestreo ppt con $n = 6$ en esas provincias restantes, considerando fijas en la muestra las provincias mencionadas, y el estimador del total se construirá con el esquema habitual cuando hay unidades autorrepresentadas. Es decir, como las 4 provincias autorrepresentadas suman 34467 en producción de guisante, el estimador del total será $t_{HT} + 34467$ y su varianza estimada $\widehat{V}(t_{HT})$ pues las cuatro provincias autorrepresentadas no añaden variabilidad al estimador por estar seleccionadas de manera no aleatoria.

El proceso se hace como en el apartado b) , repetidamente 5 veces con las semillas usuales. En la sintaxis del proc surveyselect se indica el archivo modificado guisa2, el tamaño $n = 6$ y la opción jtprobs para que el archivo de muestra contenga las probabilidades de inclusión, necesarias para las estimaciones con la macro estimppt. En muestreo ppt no hace falta tener calculada la variable pi (la macro estimppt no la va a utilizar, sino que utilizará las probabilidades de inclusión en su lugar), por lo que en el apartado size se indica directamente la variable super. En la macro, hay que señalar que el tamaño poblacional es el nuevo tras haber eliminado las 4 provincias, es decir $N' = 47$:

ISBN 9788496866133

```
proc surveyselect data=guisa2 out=muestra n=6 method=pps seed=1234 jtprobs;
size super;
run;
%estimppt(muestra,produ,6,47);
```

Con las 5 semillas respectivas , se obtienen los estimadores en la subpoblación muestreada, de 20720, 15364, 30311, 20783 y 17256.

Los estimadores finales respectivos son 20720+34467 = 55187, 15364+34467 = 49831, 30311+34467 = 64778, 20783+34467 = 55250, y 17256+34467 = 51273. Como se observa, se trata de estimaciones muy cercanas al verdadero valor poblacional de la producción total, que es 57694.

d) Se calcula la media poblacional de superficie, que es 1454.37, para utilizarla en la estimación indirecta. El programa para realizar la estimación de expansión es el proc means y para la estimación indirecta la macro estimrazreg:

```
proc surveyselect data=guisa out=muestra n=10 method=srs seed=1234 ;
run;
proc means data=muestra;var produ;run;
%estimrazreg (muestra,produ,super,1454.37,51,10);
```

Finalmente, se presentan en la siguiente tabla los valores obtenidos para la estimación del total con las 5 semillas, en cada método realizado.

Semilla	$N\widehat{\overline{y}}$	$N\overline{y}_R$	$N\overline{y}_{reg}$	t_{HH}	t_{HT}
1234	53448	50017	50825	103625	55187
1235	165846	62751	71226	56395	49831
1236	109803	62378	64974	62945	64778
1237	23796	67754	59015	75885	55250
1238	40203	59360	55053	51549	51273

Se observa como el método más preciso es muestreo ppt sin reemplazamiento (aunque ha habido que cambiar las condiciones de muestreo, utilizando unidades autorrepresentadas). La estimación de razón y regresión son bastante buenas. Éstas también tienen la ventaja de que no es necesario a priori conocer los valores de todas las provincias en la variable super, solamente su media poblacional. El estimador de expansión es menos preciso debida a la gran variabilidad S_y^2 de la variable de interés .

Ejercicio 7.9

Supongamos que tenemos la siguiente población de 4 unidades numeradas 1,2,3,4 con valores en $y = \{2, 3, 4, 5\}$, y se desea obtener una muestra de tamaño $n = 3$, con probabilidades de inclusión de primer orden, en ese caso, de $\pi_1 = 0.6$, $\pi_2 = 0.9$, $\pi_3 = 0.6$, $\pi_4 = 0.9$, y de segundo orden de:

$\pi_{12} = 0.5$

$\pi_{13} = 0.2$

ISBN 9788496866133

$\pi_{14} = 0.5$

$\pi_{23} = 0.5$

$\pi_{24} = 0.8$

$\pi_{34} = 0.5$

Utilizar la macro estimppt del SAS para obtener estimaciones para el total con la muestra 1,2,3 y con la muestra 1,3,4.

Para ello hay que crear el archivo donde aparecen los valores de y y las probabilidades de inclusión para los elementos muestrales.

```
data uno;
input y selectionprob jtprob_1-jtprob_3;
cards;
 2  0.6   0 0.5   0.2
 3  0.9   0.5 0   0.5
 4  0.6   0.2 0.5 0
;
```

a continuación se ejecuta la macro estimppt:

```
%estimppt(uno,y,3,4);
```

y se obtiene una estimación de 13.33. con una varianza de Yates-Grundy de 0.61.

Para la muestra 1,3,4 es similar:

```
data uno;
input y selectionprob jtprob_1-jtprob_3;
cards;
 2  0.6   0 0.2  0.5
 4  0.6   0.2 0  0.5
 5  0.9   0.5 0.5 0
;
```

Y la macro estimppt:

```
%estimppt(uno,y,3,4);
```

Obteniendo una estimación de 15.55 con una varianza de Yates-Grundy de 9.38.

Ejercicio 7.10

En una ciudad con 8 escuelas elementales un vendedor de ordenadores pretende estimar el número de alumnos zurdos. Utiliza el número de alumnos matriculados en las escuelas para seleccionar con probabilidad proporcional al tamaño, 3 escuelas. La información sobre las escuelas muestreadas y las probabilidades de selección aparece en la tabla siguiente:

ISBN 9788496866133

Escuela	nº de alumnos zurdos	π_i
1	10	0.45
2	4	0.40
3	2	0.50

Las probabilidades de inclusión de segundo orden π_{ij} son:

	j		
i	1	2	3
1	0	0.20	0.20
2	0.20	0	0.20
3	0.20	0.20	0

Utilizar la macro estimppt en el SAS para calcular la varianza y los intervalos de confianza al 95% para el total de alumnos zurdos, basados en la varianza estimada por Horvitz-Thompson y por Yates-Grundy.

Para ejecutar la macro es necesario introducir los datos en un archivo creando las variables selectionprob (π_i) y jtprob_1 hasta jtprob_3 (π_{ij}). La creación del archivo se produce en el paso data:

```
data dos;
input escuela zurdos selectionprob jtprob_1-jtprob_3;
cards;
1 10 0.45 0 0.2 0.2
2 4  0.40 0.2 0 0.2
3 2  0.50 0.2 0.2 0
;
```

a continuación se ejecuta la macro estimppt:

```
%estimppt(dos,zurdos,3,8);
```

Las varianzas estimadas son 361.82 con el estimador de Horvitz-Thompson y 26.56 con el estimador de Yates-Grundy.

Un intervalo basado en el estimador de la varianza de H-T: (-1.06, 73.50), que se puede dejar en (0, 73.50), y un intervalo basado en la estimación de la varianza de Yates-Grundy, de (26.11 , 46.32) .

En la práctica estas diferencias suelen ocurrir y es difícil orientarse por uno de los dos estimadores. Como referencia, siempre se puede tener en cuenta la suposición de que se ha realizado muestreo ppt y calcular la varianza del estimador de Hansen-Hurwitz en ese caso, pues siempre servirá de aproximación como cota inferior

ISBN 9788496866133

. En este caso, se asume $\pi_i = np_i$ en el método de selección utilizado por el vendedor de ordenadores, y por lo tanto $p_i = \dfrac{\pi_i}{n}$. Se estimará con la macro estimpptr la varianza del estimador de Hansen-hurwitz, en este supuesto:

```
data dos;
set dos;
pi=selectionprob/3;
run;
%estimpptr(dos,.,zurdos,escuela,8,3,1);
```

La varianza estimada es 258.71, más cercana a la que arrojaba el estimador de la varianza de Horvitz-Thompson.

ISBN 9788496866133

8.8 Ejercicios propuestos

1) Una cadena de supermercados posee tiendas en 9 ciudades españolas. El director de la compañía piensa que sería interesante crear guarderías anexas a los supermercados para facilitar a sus empleados el cuidado de sus hijos. Por ello desea estimar el número medio de hijos de empleados menores de 5 años por tienda. Para conseguirlo, el director decide muestrear con reposición 4 ciudades con probabilidad proporcional al número de tiendas. Si los respectivos números de tiendas en las 9 ciudades son:

10, 25, 8, 31, 40, 16, 5, 23, 17

Se pide:

a) Mostrar cómo seleccionar la muestra con ayuda de una tabla de números aleatorios (respetar el orden en que se presentan las tiendas que hay por ciudad).

b) Si los números aleatorios seleccionados son 122, 31, 37 y 100, ¿qué ciudades constituyen la muestra?.

c) Suponiendo que el número total de hijos de empleados menores de 5 años obtenidos en las 4 ciudades muestrales son (según el orden de selección) 90, 153, 78, 205 respectivamente, estimar el número medio de hijos de empleados menores de 5 años por tienda, de toda la compañía, y dar un I.C. al 95% suponiendo normalidad.

2) Se realiza un estudio para estimar, en un colegio de 300 alumnos en 10 clases de 3º, cuántos alumnos utilizan móvil. Se seleccionan 5 clases con reemplazamiento y probabilidad proporcional al número de alumnos. Se obtiene que la proporción de alumnos que utilizan móvil en las 5 clases muestreadas es respectivamente 0.05, 0.1, 0.2, 0.2, y 0.06. Calcular un intervalo de confianza al 95% para el número de alumnos que utilizan móvil.

3) De las 30 jaulas de un gallinero se selecciona una muestra aleatoria de tamaño 3 mediante el siguiente procedimiento: la primera extracción se realiza con probabilidades proporcionales al número de gallinas de las jaulas y las 2 restantes con probabilidades iguales, siendo todas las extracciones sin reposición.

Calcular una estimación insesgada de los huevos puestos al día entre todas las jaulas del gallinero , sabiendo que hay en total 200 gallinas, con los siguientes datos:

Jaula	1	2	3
Huevos	5	3	2
Tamaño	6	5	4

Establecer también la desviación típica de la estimación.

4) Una población tiene 3 unidades con valores en 5, 2 y 8 en la variable y. Los valores respectivos de la variable x para esas observaciones son 7, 3, y 10. Suponiendo $n = 2$:

a) En muestreo pptr proporcional a la variable x, presentar la distribución del estimador del total poblacional. Comprobar que es insesgado.

ISBN 9788496866133

b) Hacer lo mismo suponiendo muestreo ppt (con el método básico).

5) En una población con 3 observaciones se desea estimar el total con $n = 2$.

y	x
1	2
4	6
6	13

a) Calcular la varianza exacta del estimador del total si se utiliza muestreo pptr con variable auxiliar x.

b) Calcular la varianza exacta del estimador del total bajo m.a.s. de tamaño $n = 2$.

c) Calcular el error cuadrático medio exacto del estimador bajo m.a.s. con estimación de razón, suponiendo conocida $\overline{x}$.

d) Calcular la varianza exacta del estimador del total si se utiliza muestreo ppt con variable auxiliar x.

6) En un proceso de control de calidad realizado en dos fábricas similares, se realiza un muestreo estratificado para determinar una estimación para el número de lotes defectuosos. En el proceso de la primera fábrica hay 10 lotes, con un total de 200 items. Se escogen dos lotes con reemplazamiento , asignando probabilidades de elección proporcionales al número de items. En el primer lote muestreado hay 4 items defectuosos, y 20 items en total. En el segundo lote hay 35 items defectuosos y 60 items. en la segunda fábrica hay 3 lotes a examinar, con un total de 150 items. Se escogen 2 lotes con probabilidades proporcionales al número de items sin reemplazamiento. En el primer lote muestreado hay 60 items, de los cuales 30 son defectuosos. En la segunda fábrica hay 10 items defectuosos y 30 items.

Suponiendo normalidad del estimador,

a) Dar un I.C. al 95% para el número total de items defectuosos en conjunto en las dos fábricas.

b) Dar un I.C. al 95% para la proporción de items defectuosos en conjunto en las dos fábricas.

7) Resolver el ejercicio propuesto 1), c) en SAS, con la ayuda de la macro estimpptr.

8) El archivo SAS aprobados contiene los alumnos matriculados en Selectividad en Junio en las 42 Universidades españolas en 1996 (variable matri) , y el número de aprobados (variable junio). Se trata de estimar el total de aprobados utilizando la variable matri como variable auxiliar. Se irá rellenando la tabla que aparece más abajo a medida que se realicen las estimaciones.

a) Calcular el verdadero total de alumnos aprobados en Junio con

```
proc means data=aprobados sum;var junio;run;
```


ISBN 9788496866133

b) Extraer 5 muestras con probabilidad proporcional a los alumnos matriculados con reemplazamiento, de tamaño $n = 6$, con las semillas 1234, 1235, 1236, 1237, 1238. Utilizar la macro estimpptr para estimar el total y observar los valores del estimador. Observar también los estimadores de la varianza.

c) Realizar el mismo proceso con muestreo ppt . Utilizar la macro estimppt.

d) Realizar muestreo aleatorio simple 5 veces con las mismas semillas , observando el valor de la media muestral. Calcular en estos casos también el estimador de razón y regresión mediante la macro estimrazreg. Comparar con los resultados anteriores: ¿Qué estimadores parecen tener mayor varianza y cuáles menor?.

Muestra	$N\overline{y}$	$N\overline{y}_R$	$N\overline{y}_{reg}$	t_{HH}	t_{HT}
1					
2					
3					
4					
5					

e) Supongamos que se seleccionan forzosamente las Universidades más grandes, UCM, País Vasco, Santiago, Autónoma de Madrid y Barcelona. La Universidad restante se extrae por m.a.s. del resto. Realizarlo sucesivamente 5 veces con las semillas anteriores y calcular el valor del estimador final. ¿Que tal es la precisión de este último estimador?.

f) Como se pueden conocer las cuasivarianzas exactas de las variables, calcular la varianza exacta del estimador del apartado e), la de t_{HH} bajo muestreo pptr, la del estimador de expansión bajo m.a.s. y la del estimador de razón del total bajo m.a.s.. ¿Cuál es menor?.

ISBN 9788496866133

ISBN 9788496866133

9 MUESTREO POR CONGLOMERADOS EN UNA ETAPA

En este capítulo se estudiará el tipo de muestreo que selecciona aleatoriamente conjuntos de unidades llamados conglomerados, de una partición de la población en estos grupos. Al ser muestreo en una etapa, se considera que en cada conglomerado de los seleccionados se examinarán todas las unidades pertenecientes a él.

9.1 Introducción

Se comenzará con la definición de conglomerado (cluster) de unidades.

Definición.

Un **conglomerado** es un conjunto de unidades, que a su vez es una clase o parte de una partición de la población.

Como esta definición puede ser equívoca respecto a los estratos, que también son particiones de la población, hay que añadir que los conglomerados se crean con el propósito de seleccionar una muestra de entre ellos por algún método de muestreo, dejando otros sin examinar. La diferencia respecto al muestreo estratificado es que en éste último se seleccionan unidades dentro de todos y cada uno de los estratos. El muestreo estratificado, para ser eficiente, requería cierta diferencia entre los estratos respecto a la variable de estudio, y cierta homogeneidad interna. En el muestreo por conglomerados, como se van a selecionar sólo algunos, conviene que cada uno de ellos esté en cierto modo representado por los demás (homogeneidad entre conglomerados), y, además, que cada uno de ellos sea en sí mismo una buena representación de la población (heterogeneidad intra conglomerados).

La Figura 9.1 representa una partición típica del muestreo por conglomerados (en este caso, los conglomerados son los corralitos): estos son parecidos entre sí, y la variabilidad interna de cada uno de ellos representa en cierto modo la variabilidad de la población. Compárese con la partición en estratos de la Figura 4.1. para comprender la diferencia.

ISBN 9788496866133

Nótese también en la Figura 9.1 que los conglomerados no tienen por qué tener el mismo tamaño (número de unidades elementales), aunque a menudo se construyen de tamaños iguales por simplificar los problemas de muestreo y estimación.

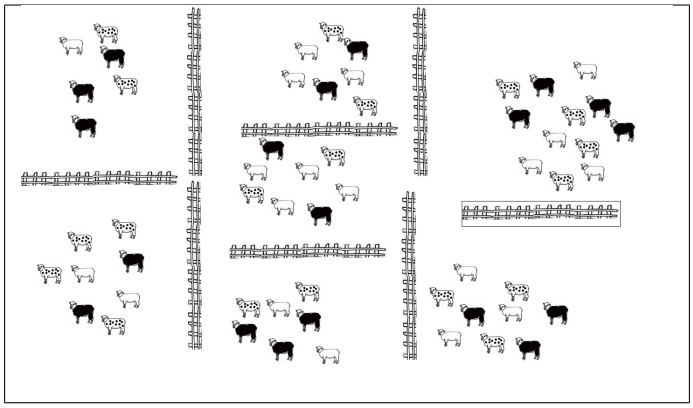


Figura 9.1. Partición de la población en 7 conglomerados.

El muestreo por conglomerados presenta algunas ventajas que veremos a continuación:

- Suele ser menos costoso en tiempo o recursos que otros tipos de muestreo, pues la disposición de las unidades en grupo facilita las tareas administrativas de localización, desplazamiento y toma de datos.
- A menudo no se dispone de información exhaustiva de todas las unidades de la población, o es costoso conseguirla. En muestreo por conglomerados sólo se necesita recabar información sobre los conglomerados seleccionados, disminuyendo los costes .
- Frecuentemente los conglomerados existen ya como unidades administrativas, facilitando la tarea de particionar la población.

Por otro lado, y como principal inconveniente del método, aunque teóricamente se construyan conglomerados homogéneos entre ellos, en la práctica suele ocurrir que no lo son suficientemente, disminuyendo así la precisión del estimador. Véase en los ejemplos de la tabla 9.2 que puede existir variabilidad alta entre los conglomerados respecto a la variable de interés. Así, en general, el muestreo por conglomerados suele ser menos preciso que el muestreo aleatorio simple para el mismo tamaño muestral. A pesar de esto, para el mismo coste, el muestreo por conglomerados suele ser más preciso que otros métodos, incluido el m.a.s., debido a que permite tomar muestras de mayor tamaño por un precio similar, al reducirse mucho los gastos del trabajo de campo.

Respecto a la organización del método, un conglomerado puede a su vez, ser dividido internamente en subgrupos, y cada uno de estos subgrupos formar una partición interna de unidades. Si existe este tipo de jerarquía, se utiliza la siguiente terminología: se denominan a los primeros conglomerados **unidades de primera etapa**, a los subgrupos **unidades de segunda etapa**, etc. El método de muestreo en varias etapas consiste en seleccionar por muestreo algunos de los primeros grupos, y a continuación , y dentro de los grupos ya seleccionados, seleccionar por muestreo algunos de los subgrupos, y así sucesivamente hasta llegar a las unidades elementales. En este capítulo se tratará del muestreo por conglomerados monoetápico.

ISBN 9788496866133

Definición.

El método de muestreo por conglomerados **monoetápico** o en una etapa consiste en seleccionar por algún método de muestreo una muestra de conglomerados, y dentro de cada uno de los seleccionados examinar todas las unidades elementales.

Conglomerados	**unidades elementales**	**variable de interés** y
familias	individuos	gasto mensual
edificios	hogares	consumo eléctrico
granjas	gallinas	presencia de enfermedad
parcelas de terreno	árboles frutales	producción
mancomunidades	municipios	consumo de gas
hospitales	pacientes internos	tiempo internado
escuelas	alumnos	nota final de curso

Tabla 9.1. Ejemplos de conglomerados en muestreo monoetápico.

Hay que remarcar que al examinar todas las unidades elementales dentro de cada conglomerado de los seleccionados, es como si estuviéramos tratando simplemente con "grandes" unidades elementales, pues no hay ningún azar asociado al resultado dentro del conglomerado una vez escogido. Es como si cada conglomerado llevase asociada una característica de interés (que puede ser su total, proporción o media) y basta aplicar los resultados asociados a la técnica de selección o estimación escogida (m.a.s.r., m.a.s., ppt, etc.) sobre esa característica de interés como si ésta fuera la variable habitual y. Así, los resultados teóricos en el fondo suelen corresponderse directamente con los ya vistos anteriormente, variando simplemente la notación. Sin embargo, hay nuevos conceptos relativos a la homogeneidad "intra" y "entre" conglomerados y a la comparación con m.a.s., que no se han estudiado hasta ahora.

Notación

En este diseño de muestreo, se supone que la población está particionada en L conglomerados, de tamaños respectivos N_i, $i = 1, ..., L$.

El **tamaño medio de los conglomerados** es $\overline{N} = \dfrac{1}{L}\sum_{i=1}^{L} N_i = \dfrac{N}{L}$.

La **media del conglomerado** i es $\overline{y}_i = \dfrac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$.

El **total del conglomerado** i es $y_i \overset{def}{=} N_i\overline{y}_i = \sum_{j=1}^{N_i} y_{ij}$.

ISBN 9788496866133

La **media poblacional** es $\overline{y} = \dfrac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{N_i} y_{ij} = \dfrac{1}{N}\sum_{i=1}^{L} N_i\overline{y}_i$.

La **cuasivarianza del del conglomerado** i es

$$S_i^2 = \frac{1}{N_i-1}\sum_{j=1}^{N_i}(y_{ij}-\overline{y}_i)^2 = \frac{N_i}{N_i-1}\sigma_i^2.$$

En estudio de proporciones, $S_i^2 = \dfrac{N_i}{N_i-1}p_i(1-p_i)$.

La **cuasivarianza poblacional** es $S^2 = \dfrac{1}{N-1}\sum_{i=1}^{L}\sum_{j=1}^{N_i}(y_{ij}-\overline{y})^2 = \dfrac{N}{N-1}\sigma^2$.

En estudio de proporciones, $S^2 = \dfrac{N}{N-1}p(1-p)$.

La **cuasivarianza intra-conglomerados** es $S_w^2 = \dfrac{1}{L(\overline{N}-1)}\sum_{i=1}^{L}\sum_{j=1}^{N_i}(y_{ij}-\overline{y}_i)^2$, que es un promedio de la cuasivarianza interna de cada conglomerado.

En estudio de proporciones, $S_w^2 = \dfrac{1}{L}\sum_{i=1}^{L}\dfrac{N_i}{N_i-1}p_i(1-p_i)$.

La **cuasivarianza entre conglomerados** es $S_b^2 = \dfrac{1}{(L-1)}\sum_{i=1}^{L} N_i(\overline{y}_i-\overline{y})^2$. Si todos los conglomerados tienen el mismo tamaño, $S_b^2 = \dfrac{\overline{N}}{(L-1)}\sum_{i=1}^{L}(\overline{y}_i-\overline{y})^2$.

En estudio de proporciones,

$$S_b^2 = \frac{1}{(L-1)}\sum_{i=1}^{L} N_i(p_i-p)^2.$$

9.2 Conglomerados de igual tamaño

En este apartado se estudiará el caso en que los conglomerados pueden considerarse de igual tamaño $\overline{N}$. En la notación anterior, esto equivale a sustituir cada N_i por $\overline{N}$. Si la variación en tamaño de los conglomerados es pequeña, aproximadamente entre 5% y 10%, se pueden utilizar los resultados de este apartado, pues la contribución del tamaño de los conglomerados a la varianza del estimador suele ser despreciable. Durante el desarrollo de este apartado se va a asumir m.a.s. en la selección de los conglomerados.

9.2.1 Análisis de la varianza en muestreo por conglomerados

Veremos en primer lugar una descomposición de la varianza poblacional en términos de las varianzas intra-conglomerados y entre conglomerados.

ISBN 9788496866133

Teorema 9.1 (descomposición de la varianza).

Se verifica que

$$(N-1)S^2 = L(\overline{N}-1)S_w^2 + (L-1)S_b^2.$$

Demostración.

Se desarrollará la demostración para el caso general en que N_i puede variar, pues $N_i = \overline{N}$ para todo i (conglomerados del mismo tamaño) es un caso particular.

$$(N-1)S^2 = \sum_{i=1}^{L}\sum_{j=1}^{N_i}(y_{ij}-\overline{y})^2 = \sum_{i=1}^{L}\sum_{j=1}^{N_i}(y_{ij}-\overline{y}_i+\overline{y}_i-\overline{y})^2 =$$

$$= \sum_{i=1}^{L}\sum_{j=1}^{N_i}(y_{ij}-\overline{y}_i)^2 + \sum_{i=1}^{L}\sum_{j=1}^{N_i}(\overline{y}_i-\overline{y})^2 + 2\sum_{i=1}^{L}\sum_{j=1}^{N_i}(y_{ij}-\overline{y}_i)(\overline{y}_i-\overline{y}).$$

Como

$$\sum_{i=1}^{L}\sum_{j=1}^{N_i}(y_{ij}-\overline{y}_i)(\overline{y}_i-\overline{y}) = \sum_{i=1}^{L}\sum_{j=1}^{N_i} y_{ij}\overline{y}_i - \sum_{i=1}^{L}\sum_{j=1}^{N_i} y_{ij}\overline{y} - \sum_{i=1}^{L}\sum_{j=1}^{N_i}\overline{y}_i^2 + \sum_{i=1}^{L}\sum_{j=1}^{N_i}\overline{y}_i\overline{y} =$$

$$= \sum_{i=1}^{L} N_i\overline{y}_i^2 - \sum_{i=1}^{L} N_i\overline{y}_i\overline{y} - \sum_{i=1}^{L}\sum_{j=1}^{N_i}\overline{y}_i^2 + \sum_{i=1}^{L} N_i\overline{y}_i\overline{y} = 0,$$

tenemos que

$$(N-1)S^2 = \sum_{i=1}^{L}\sum_{j=1}^{N_i}(y_{ij}-\overline{y}_i)^2 + \sum_{i=1}^{L}\sum_{j=1}^{N_i}(\overline{y}_i-\overline{y})^2 =$$

$$= \sum_{i=1}^{L}\sum_{j=1}^{N_i}(y_{ij}-\overline{y}_i)^2 + \sum_{i=1}^{L} N_i(\overline{y}_i-\overline{y})^2 = L\overline{(N}-1)S_w^2 + (L-1)S_b^2.$$

Así, la variabilidad total se puede descomponer en variabilidad entre conglomerados y variabilidad intraconglomerados. Como se ha comentado, en muestreo por conglomerados interesa que la descomposición sea favorable a una variabilidad "intra" alta frente a una variabilidad "entre" baja.

La siguiente definición permite aproximar de otro modo el concepto de variabilidad o heterogeneidad intra-conglomerados.

Definición.

Supongamos que los tamaños de los conglomerados son iguales, $\overline{N}$. El **coeficiente de correlación intra-conglomerados** es

$$\delta = \frac{\sum_{i=1}^{L}\sum_{j\neq k=1}^{\overline{N}}(y_{ij}-\overline{y})(y_{ik}-\overline{y})}{(\overline{N}-1)(L\overline{N}-1)S^2}.$$

Este coeficiente refleja el grado de homogeneidad interna de los grupos, por lo cual a mayor δ, peor será la configuración de conglomerados, y a menor δ, (más heterogeneidad interna),

ISBN 9788496866133

mejor será esa configuración. Se puede demostrar que el valor mínimo de δ (correspondiente a la mejor configuración de conglomerados) es $\delta = -\frac{1}{\overline{N}-1}$.

Ejemplo 9.1

Supongamos que tenemos una población de 9 elementos, con valores $\{1, 2, 3, 1, 2, 3, 1, 2, 3\}$. Entonces la media poblacional es $\overline{y} = 2$ y la cuasivarianza poblacional es $S^2 = 0.75$.

Si escogemos la siguiente configuración de conglomerados:

$\{1, 2, 3\}$, $\{1, 2, 3\}$, $\{1, 2, 3\}$

entonces

$$\delta = \frac{\sum_{i=1}^{L}\sum_{j\neq k=1}^{\overline{N}}(y_{ij}-\overline{y})(y_{ik}-\overline{y})}{(\overline{N}-1)(L\overline{N}-1)S^2} =$$

$$= \frac{3[(1-2)(2-2)+(1-2)(3-2)+(2-2)(1-2)+(2-2)(3-2)]}{(3-1)(3\cdot 3-1)0.75} +$$

$$+ \frac{3[(3-2)(1-2)+(3-2)(2-2)]}{(3-1)(3\cdot 3-1)0.75} =$$

$$= \frac{3\cdot -2}{(3-1)(3\cdot 3-1)0.75} = -\frac{1}{2}$$

Además, $-\frac{1}{\overline{N}-1} = -\frac{1}{3-1} = -\frac{1}{2}$ con lo cual la configuración escogida es la mejor posible.

Si, por el contrario, se utiliza la configuración:

$\{1, 1, 1\}$, $\{2, 2, 2\}$, $\{3, 3, 3\}$

entonces

$$\delta = \frac{6(1-2)(1-2)+6(3-2)(3-2)+6(2-2)(2-2)}{(3-1)(3\cdot 3-1)0.75} =$$

$$= \frac{12}{(3-1)(3\cdot 3-1)0.75} = 1.$$

que es positiva, y, como se verá más adelante, lleva a que el muestreo por conglomerados, con esta configuración, sea inferior al m.a.s. para el mismo tamaño muestral.

9.2.2 Estimación de la media

Veremos a continuación un estimador de la media en el caso de conglomerados de igual tamaño.

ISBN 9788496866133

Teorema 9.2 (estimación de la media).

Supongamos que se realiza un m.a.s. de n conglomerados entre L, siendo éstos de igual tamaño $\overline{N}$. En cada conglomerado i de los n muestreados se examinan todas las unidades y por lo tanto se obtiene la media poblacional en el conglomerado $\overline{y}_i$. Entonces

$$\overline{y}_c = \frac{1}{n}\sum_{i=1}^{n}\overline{y}_i$$

es un estimador insesgado de $\overline{y}$.

Demostración.

$E(\overline{y}_c) = \frac{1}{n}\sum_{i=1}^{n}E(\overline{y}_i)$. Ahora, por el mismo razonamiento que se hizo al estudiar m.a.s., cada $\overline{y}_i$ es una variable aleatoria que toma valores $\overline{y}_1, ..., \overline{y}_L$ cada uno con probabilidad $\frac{1}{L}$. entonces para todo i, es

$$E(\overline{y}_i) = \frac{1}{L}\sum_{j=1}^{L}\overline{y}_j$$

y por lo tanto

$$E(\overline{y}_c) = \frac{1}{n}\sum_{i=1}^{n}E(\overline{y}_i) = E(\overline{y}_c) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{L}\sum_{i=1}^{L}\overline{y}_i = \frac{1}{L}\sum_{i=1}^{L}\overline{y}_i = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{\overline{N}}\sum_{j=1}^{\overline{N}}y_{ij} = \overline{y}.$$

Teorema 9.3 (varianza del estimador).

La varianza de $\overline{y}_c$ es

$$V(\overline{y}_c) = \frac{L-n}{L}\frac{1}{n(L-1)}\sum_{i=1}^{L}(\overline{y}_i - \overline{y})^2 = \frac{L-n}{L\overline{N}n}S_b^2.$$

Demostración.

Es otra consecuencia de la utilización del muestreo aleatorio simple sobre los grupos. Basta redefinir el conjunto de conglomerados como una población de L unidades, cada una de las cuales toma valor $\overline{y}_i$, y donde se realiza m.a.s. de tamaño n.

El estimador definido $\overline{y}_c$ coincide con el estimador usual media muestral en m.a.s., pues $\overline{y}_c = \frac{1}{n}\sum_{i=1}^{n}\overline{y}_i$, y por lo tanto su varianza será

$$V(\overline{y}_c) = \frac{L-n}{L}\frac{S'^2}{n}$$

donde en este caso S'^2 es la cuasivarianza de la variable $\overline{y}_i$, que es $\frac{1}{L-1}\sum_{i=1}^{L}(\overline{y}_i - \overline{y})^2$ al ser $\frac{1}{L}\sum_{i=1}^{L}\overline{y}_i = \overline{y}$. Como

$$S_b^2 = \frac{\overline{N}}{L-1}\sum_{i=1}^{L}(\overline{y}_i - \overline{y})^2 \ ,$$

ISBN 9788496866133

entonces

$$S'^2 = \frac{S_b^2}{\overline{N}}$$

y se obtiene el resultado.

La siguiente expresión relaciona esta varianza con el coeficiente de correlación intra-conglomerados.

Teorema 9.4 (otra forma para la varianza del estimador).

$$V(\overline{y}_c) = \frac{(L-n)N\sigma^2}{L(L-1)n\overline{N}^2}(1+(\overline{N}-1)\delta).$$

Demostración.

En primer lugar se puede expresar δ en función de σ^2. Por ser

$$S^2 = \frac{N}{N-1}\sigma^2 = \frac{L\overline{N}}{L\overline{N}-1}\sigma^2,$$

entonces

$$\delta = \frac{\sum_{i=1}^{L}\sum_{j\neq k=1}^{\overline{N}}(y_{ij}-\overline{y})(y_{ik}-\overline{y})}{L\overline{N}(\overline{N}-1)\sigma^2}$$

Ahora,

$$V(\overline{y}_c) = \frac{L-n}{Ln(L-1)}\sum_{i=1}^{L}(\overline{y}_i-\overline{y})^2 = \frac{L-n}{Ln(L-1)}\sum_{i=1}^{L}\left(\frac{1}{\overline{N}}\sum_{j=1}^{\overline{N}}y_{ij} - \frac{1}{\overline{N}}\sum_{j=1}^{\overline{N}}\overline{y}\right)^2 =$$

$$= \frac{L-n}{Ln(L-1)\overline{N}^2}\sum_{i=1}^{L}\left[\sum_{j=1}^{\overline{N}}(y_{ij}-\overline{y})\right]^2 =$$

$$= \frac{L-n}{Ln(L-1)\overline{N}^2}\sum_{i=1}^{L}\left[\sum_{j=1}^{\overline{N}}(y_{ij}-\overline{y})^2 + \sum_{j\neq k=1}^{\overline{N}}(y_{ij}-\overline{y})(y_{ik}-\overline{y})\right] =$$

$$= \frac{L-n}{Ln(L-1)\overline{N}^2}\left[N\sigma^2 + N(\overline{N}-1)\sigma^2\delta\right] = \frac{(L-n)N\sigma^2}{L(L-1)n\overline{N}^2}(1+(\overline{N}-1)\delta).$$

Corolario 9.1 (forma aproximada para la varianza del estimador).

Una aproximación a $V(\overline{y}_c)$ es

$$V(\overline{y}_c) \simeq \frac{L-n}{L}\frac{S^2}{n\overline{N}}(1+(\overline{N}-1)\delta).$$

Demostración.

$$V(\overline{y}_c) = \frac{L-n}{Ln\overline{N}}\frac{N\sigma^2}{(L-1)\overline{N}}(1+(\overline{N}-1)\delta).$$

ISBN 9788496866133

Sólo hay que demostrar que

$$\frac{N\sigma^2}{(L-1)\overline{N}} \simeq S^2.$$

Utilizando que $S^2 = \frac{N}{N-1}\sigma^2$,

$$\frac{N\sigma^2}{(L-1)\overline{N}} = \frac{(N-1)S^2}{(L-1)\overline{N}} = \frac{(N-1)}{(N-\overline{N})}S^2.$$

El término $\frac{(N-1)}{(N-\overline{N})} = \frac{1}{1-\frac{\overline{N}}{N}} - \frac{1/N}{1-\frac{\overline{N}}{N}} \simeq 1$ pues se considera $\frac{\overline{N}}{N}$ despreciable. Entonces se tiene que $\frac{N\sigma^2}{(L-1)\overline{N}} \simeq S^2$ y por lo tanto

$$V(\overline{y}_c) \simeq \frac{L-n}{L}\frac{S^2}{n\overline{N}}(1+(\overline{N}-1)\delta).$$

9.2.3 Estimación de varianzas

Para estudiar la precisión del estimador en la práctica, es necesario construir estimadores de su varianza.

El siguiente resultado es un paso previo para estimar el coeficiente de correlación intra conglomerados.

Teorema 9.5 (aproximaciones a δ).

Suponiendo N y L suficientemente grandes, se tiene que:

(a) $\delta \simeq \frac{S_b^2 - S^2}{(\overline{N}-1)S^2}$.

(b) $\delta \simeq 1 - \frac{\overline{N}}{\overline{N}-1}\frac{L(\overline{N}-1)S_w^2}{(N-1)S^2}$.

Demostración.

(a) Como

$$V(\overline{y}_c) \simeq \frac{L-n}{L}\frac{S^2}{n\overline{N}}(1+(\overline{N}-1)\delta)$$

y además

$$V(\overline{y}_c) = \frac{L-n}{L\overline{N}n}S_b^2,$$

igualando ambos términos se tiene que

ISBN 9788496866133

$$\delta \simeq \frac{S_b^2 - S^2}{(\overline{N}-1)S^2} \ .$$

(b) Utilizando que

$$L(\overline{N}-1)S_w^2 = (N-1)S^2 - (L-1)S_b^2$$

por la propiedad de descomposición de la varianza, se tiene que

$$1 - \frac{\overline{N}}{\overline{N}-1}\frac{L(\overline{N}-1)S_w^2}{(N-1)S^2} = \frac{(\overline{N}-1)(N-1)S^2 - \overline{N}(N-1)S^2 + \overline{N}(L-1)S_b^2}{(\overline{N}-1)(N-1)S^2} =$$

$$= \frac{-(N-1)S^2 + \overline{N}(L-1)S_b^2}{(\overline{N}-1)(N-1)S^2} = \frac{\frac{\overline{N}(L-1)}{N-1}S_b^2 - S^2}{(\overline{N}-1)S^2}.$$

Ahora,

$\frac{\overline{N}(L-1)}{N-1} = \frac{N-\overline{N}}{N-1} = \frac{N}{N-1} - \frac{\overline{N}}{N-1} \simeq 1$ pues el segundo término se considera aproximadamente cero suponiendo $\overline{N}$ despreciable frente a N. Así,

$$1 - \frac{\overline{N}}{\overline{N}-1}\frac{L(\overline{N}-1)S_w^2}{(N-1)S^2} \simeq \frac{S_b^2 - S^2}{(\overline{N}-1)S^2} \simeq \delta.$$

El siguiente resultado permite estimar el coeficiente de correlación intra conglomerados, lo que será útil en el sentido de diagnosticar si la configuración de conglomerados permitirá obtener una buena precisión en las estimaciones.

Corolario 9.2 (estimación de δ).

Un estimador de los momentos de δ es

$$\widehat{\delta} = 1 - \frac{\overline{N}}{\overline{N}-1}\frac{L(\overline{N}-1)s_w^2}{(N-1)s^2}$$

donde

$$s_w^2 = \frac{1}{n(\overline{N}-1)}\sum_{i=1}^{n}\sum_{j=1}^{\overline{N}}(y_{ij} - \overline{y}_i)^2$$

y

$$s^2 = \frac{1}{n\overline{N}-1}\sum_{i=1}^{n}\sum_{j=1}^{\overline{N}}(y_{ij} - \overline{y}_c)^2.$$

En muestreo por conglomerados no se cumple que s^2 sea un estimador insesgado de S^2. El siguiente resultado permite estimar la cuasivarianza poblacional S^2 y la varianza del estimador.

Teorema 9.6 (estimación de varianzas poblacionales).

a) Un estimador insesgado de S_w^2 es s_w^2.

ISBN 9788496866133

b) Un estimador insesgado de S_b^2 es $s_b^2 = \dfrac{\overline{N}}{(n-1)} \sum_{i=1}^{n} (\overline{y}_i - \overline{y}_c)^2$.

c) Un estimador insesgado de S^2 es $\widehat{S}^2 = \dfrac{L(\overline{N}-1)s_w^2 + (L-1)s_b^2}{(N-1)}$.

Demostración.

a) $E(s_w^2) = \dfrac{1}{n} \sum_{i=1}^{n} E\left[\dfrac{1}{\overline{N}-1} \sum_{j=1}^{\overline{N}} (y_{ij} - \overline{y}_i)^2 \right]$.

El término $S_i^2 \overset{def.}{=} \dfrac{1}{\overline{N}-1} \sum_{j=1}^{\overline{N}} (y_{ij} - \overline{y}_i)^2$ es la cuasivarianza poblacional interna en el conglomerado i. Como se eligen n conglomerados por m.a.s. de entre L, es una variable aleatoria que toma valores $S_1^2, ..., S_L^2$ con probabilidades iguales $\dfrac{1}{L}$. Su esperanza será entonces $E(S_i^2) = \dfrac{1}{L} \sum_{j=1}^{L} S_j^2$. Así,

$$E(s_w^2) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{L} \sum_{j=1}^{L} S_j^2 = \frac{1}{L} \sum_{j=1}^{L} S_j^2 = \frac{1}{L(\overline{N}-1)} \sum_{i=1}^{L} \sum_{j=1}^{\overline{N}} (y_{ij} - \overline{y}_i)^2 = S_w^2.$$

b) $E(s_b^2) = \overline{N} E\left[\dfrac{1}{(n-1)} \sum_{i=1}^{n} (\overline{y}_i - \overline{y}_c)^2 \right]$. Considerando cada conglomerado como una unidad de una población de L unidades, de la cual se toma una m.a.s. de n unidades, y se denota por $\overline{y}_i$ la característica de interés de esa unidad, se observa que $\dfrac{1}{(n-1)} \sum_{i=1}^{n} (\overline{y}_i - \overline{y}_c)^2$ no es más que la cuasivarianza muestral $s_{\overline{y}_i}^2$ de la característica $\overline{y}_i$, y por ser m.a.s. $E(s_{\overline{y}_i}^2) = S_{\overline{y}_i}^2 = \dfrac{1}{(L-1)} \sum_{i=1}^{L} (\overline{y}_i - \overline{y}_c)^2$. Así,

$$E(s_b^2) = \overline{N} E(s_{\overline{y}_i}^2) = \frac{\overline{N}}{(L-1)} \sum_{i=1}^{L} (\overline{y}_i - \overline{y}_c)^2 = S_b^2.$$

c) Al ser insesgados s_w^2 y s_b^2, y por la propiedad de descomposición de la varianza, se tiene que

$$E\left[L(\overline{N}-1)s_w^2 + (L-1)s_b^2. \right] = L(\overline{N}-1)S_w^2 + (L-1)S_b^2 = (N-1)S^2.$$

Entonces,

$$E(\widehat{S}^2) = E\left[\frac{L(\overline{N}-1)s_w^2 + (L-1)s_b^2}{(N-1)} \right] = S^2.$$

Corolario 9.3 (otro estimador de δ).

Otro estimador de los momentos de δ es

$$\widehat{\delta} = \frac{s_b^2 - \widehat{S}^2}{(\overline{N}-1)\widehat{S}^2}.$$

El siguiente importante corolario ofrece la posibilidad de estimar la varianza del estimador de manera insesgada.

ISBN 9788496866133

Corolario 9.4 (estimador de la varianza del estimador).

Un estimador insesgado de la varianza

$$V(\overline{y}_c) = \frac{L-n}{L\overline{N}n} S_b^2.$$

es

$$\widehat{V}(\overline{y}_c) = \frac{L-n}{L\overline{N}n} s_b^2 = \frac{(1-f_1)}{n(n-1)} \sum_{i=1}^{n} (\overline{y}_i - \overline{y}_c)^2$$

Ejemplo 9.2

Un agricultor desea estimar la producción media de sus perales. Divide su terreno en 20 grupos de árboles de 6 árboles cada conjunto, y escoge aleatoriamente 5 grupos, examinando todos los árboles dentro de cada grupo. Los datos obtenidos en kilogramos de producción, en cada uno de los árboles de los grupos muestreados, son los siguientes:

Grupo\Árbol	1	2	3	4	5	6
1	18	10	15	20	14	8
2	9	13	10	11	7	17
3	12	15	14	6	8	9
4	10	12	13	7	6	8
5	12	16	13	12	10	7

Tabla 9.2. Datos de producción de perales.

De este modo, se obtienen las medias en cada conglomerado, que son $\overline{y}_1 = 15.4$, $\overline{y}_2 = 10$, $\overline{y}_3 = 11$, $\overline{y}_4 = 9.6$, $\overline{y}_5 = 12.6$.

Así, la estimación de la producción media por peral es

$$\overline{y}_c = \frac{1}{n} \sum_{i=1}^{n} \overline{y}_i = \frac{1}{5}(15.4 + 10 + 11 + 9.6 + 12.6) = 11.72.$$

La estimación de la varianza del estimador será

$$\widehat{V}(\overline{y}_c) = \frac{L-n}{Ln} \frac{1}{(n-1)} \sum_{i=1}^{n} (\overline{y}_i - \overline{y}_c)^2 = \frac{20-5}{20 \cdot 5} \frac{1}{(5-1)} \sum_{i=1}^{5} (\overline{y}_i - \overline{y}_c)^2 = 0.8358.$$

Si se desea profundizar más en la estructura de conglomerados, se puede estimar el coeficiente de homogeneidad δ. Para ello es necesario calcular

$$s_w^2 = \frac{1}{n(\overline{N}-1)} \sum_{i=1}^{n} \sum_{j=1}^{\overline{N}} (y_{ij} - \overline{y}_i)^2 = 1.62$$

y

ISBN 9788496866133

$$s^2 = \frac{1}{n\overline{N}-1}\sum_{i=1}^{n}\sum_{j=1}^{\overline{N}}(y_{ij}-\overline{y}_c)^2 = 12.79.$$

Así, tenemos que

$$\widehat{\delta} = 1 - \frac{\overline{N}}{\overline{N}-1}\frac{L(\overline{N}-1)s_w^2}{(N-1)s^2} = 1 - \frac{6}{6-1}\frac{20(6-1)1.62}{(120-1)12.79} = 0.87.$$

Una manera gráfica de observar las diferencias entre conglomerados, (perjudiciales en cuanto a la precisión del estimador) es utilizar un diagrama de cajas:

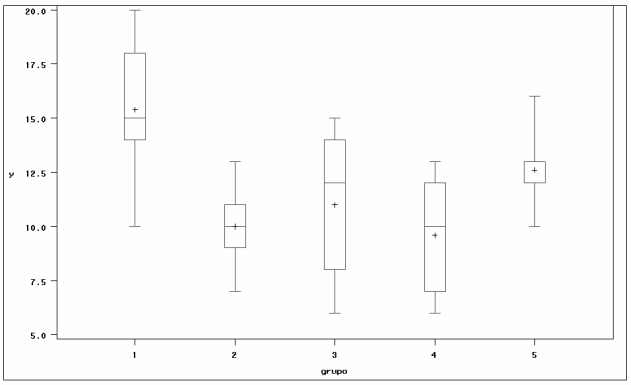


Figura 9.2. Diagrama de cajas por conglomerado.

En la Figura 9.2 se aprecia la variabilidad entre conglomerados, que puede afectar a la precisión de la estimación. Como se verá a continuación, cuando δ es positivo el muestreo por conglomerados es más impreciso que el m.a.s. para el mismo tamaño muestral, aunque como frecuentemente ocurre, utilizando conglomerados existe una ganancia respecto a la reducción en trabajo de campo: el agricultor sabe que solamente tiene que desplazarse a 5 grupos para realizar la toma de datos, cosa que no ocurriría con el m.a.s., más complejo de implementar. Una alternativa apropiada en este problema práctico concreto sería utilizar muestreo sistemático.

9.2.4 Estimación del total y proporción

En el caso de conglomerados del mismo tamaño, se aplican directamente los resultados anteriores.

Propiedad 9.1 (estimación del total y proporción).

En m.a.s. monoetápico de conglomerados de tamaños iguales:

a) Un estimador insesgado del total es $N\overline{y}_c = \frac{N}{n}\sum_{i=1}^{n}\overline{y}_i$, con varianza $V(N\overline{y}_c) = N^2V(\overline{y}_c)$ y estimador insesgado de la varianza $N^2\widehat{V}(\overline{y}_c)$.

b) Supongamos que la variable y toma valores 0 ó 1 y se desea estimar la proporción p de valores 1 en la población. Un estimador insesgado de esta proporción p es $\overline{y}_c = \widehat{p}_c = \frac{1}{n}\sum_{i=1}^{n}p_i$,

ISBN 9788496866133

donde p_i es la proporción en el conglomerado i. En este caso, $S_b^2 = \dfrac{\overline{N}}{(L-1)}\sum_{i=1}^{L}(p_i - p)^2$ y la varianza es $V(\widehat{p}_c) = V(\overline{y}_c) = \dfrac{L-n}{L\overline{N}n}S_b^2$. El estimador de la varianza será $\widehat{V}(\widehat{p}_c) = \dfrac{L-n}{L\overline{N}n}s_b^2$, donde $s_b^2 = \dfrac{\overline{N}}{(n-1)}\sum_{i=1}^{n}(p_i - \widehat{p}_c)^2$.

Ejemplo 9.3

Se desea estimar el número de franquicias de una determinada empresa que han cumplido sus objetivos anuales, y la proporción de franquicias que los cumplen. Para ello se agrupan las franquicias en 25 regiones geográficas con 10 franquicias cada una, y se obtiene una m.a.s. de 5 regiones, examinando todas las franquicias dentro de cada una de estas regiones .

Los resultados obtenidos son 3 franquicias que han cumplido los objetivos en la primera región analizada, y sucesivamente 4,3,2,3 en las regiones 2,3,4, y 5 respectivamente.

Suponiendo la creación de la variable y, con valor 1 si la franquicia en cuestión ha cumplido los objetivos y con valor 0 si no, se puede estimar la proporción poblacional a través de $\widehat{p}_c$:

$$\widehat{p}_c = \frac{1}{n}\sum_{i=1}^{n} p_i = \frac{1}{5}(\frac{3}{10} + \frac{4}{10} + \frac{3}{10} + \frac{2}{10} + \frac{3}{10}) = 0.3.$$

Para calcular la varianza estimada del estimador , se tiene que

$$s_b^2 = \frac{\overline{N}}{(n-1)}\sum_{i=1}^{n}(p_i - \widehat{p}_c)^2 = 0.05$$

y por lo tanto

$$\widehat{V}(\widehat{p}_c) = \frac{L-n}{L\overline{N}n}s_b^2 = \frac{25-5}{25\cdot 10\cdot 5}0.05 = 0.0008.$$

Si se quiere estimar el total, se utiliza el resultado anterior, pues

$N\overline{y}_c = N\widehat{p}_c = 250 \cdot 0.3 = 75$

y su varianza estimada será

$N^2\widehat{V}(\widehat{p}_c) = 50.$

9.2.5 Comparación con m.a.s.

En este apartado se comparará la precisión del muestreo por conglomerados monoetápico con m.a.s., con conglomerados de igual tamaño, con el m.a.s., suponiendo el mismo tamaño muestral=número de unidades elementales muestradas. Además se estudiará cómo buscar el tamaño muestral adecuado en algunos casos.

Teorema 9.7 (comparación con m.a.s.).

Suponiendo el mismo número de unidades elementales muestradas, se verifica que

ISBN 9788496866133

$$\frac{V(\overline{y}_c)}{V(\overline{y}_s)} \simeq 1 + (\overline{N} - 1)\delta.$$

Demostración.

Muestrear n conglomerados de tamaño $\overline{N}$ equivale a obtener $n\overline{N}$ unidades elementales, con lo que el m.a.s. equivalente debe ser de $n\overline{N}$ unidades. Así $V(\overline{y}_s) = \dfrac{N - n\overline{N}}{N}\dfrac{S^2}{n\overline{N}}$ y como $V(\overline{y}_c) \simeq \dfrac{L-n}{L}\dfrac{S^2}{n\overline{N}}(1 + (\overline{N} - 1)\delta) = \dfrac{N - n\overline{N}}{N}\dfrac{S^2}{n\overline{N}}(1 + (\overline{N} - 1)\delta)$, entonces

$$\frac{V(\overline{y}_c)}{V(\overline{y}_s)} \simeq 1 + (\overline{N} - 1)\delta.$$

Corolario 9.5 (comparación en función de δ).

Asumiendo L suficientemente grande como para hacer válida la expresión aproximada de $V(\overline{y}_c)$, entonces:

a) Si $\delta = -\dfrac{1}{\overline{N} - 1}$, entonces $V(\overline{y}_c) = 0 \leq V(\overline{y}_s)$.

b) Si $-\dfrac{1}{\overline{N} - 1} \leq \delta \leq 0$, entonces $V(\overline{y}_c) \leq V(\overline{y}_s)$.

c) Si $\delta = 0$, entonces $V(\overline{y}_c) = V(\overline{y}_s)$.

d) Si $\delta > 0$, entonces $V(\overline{y}_c) > V(\overline{y}_s)$.

Como consecuencia, si la homogeneidad interna de los conglomerados es alta (δ alto), el muestreo por conglomerados será menos preciso que el m.a.s., algo natural. En la práctica, se utiliza un estimador de los momentos del coeficiente de correlación intra-conglomerados.

Ejemplo 9.4

En el ejemplo 9.2, se puede estimar la relación entre realizar muestreo por conglomerados y m.a.s. a través de la estimación de δ : como $\widehat{\delta} = 0.87$, se tiene que $1 + (\overline{N} - 1)\delta = 1 + 5 \cdot 0.87 = 5.35$, con lo cual la varianza del estimador realizando muestreo por conglomerados será aproximadamente 5 veces más grande que la obtenida realizando m.a.s., para el mismo tamaño muestral.

9.2.6 Estudio del tamaño muestral

De cara a estudiar el tamaño muestral para una precisión o coste prefijados, se comenzará con una comparación con m.a.s. derivada de los resultados anteriores.

Corolario 9.6 (comparación con m.a.s. en términos de tamaño muestral).

El tamaño muestral (en unidades elementales) necesario en muestreo por conglomerados monoetápico de tamaños iguales, para obtener la misma precisión que en un m.a.s. con tamaño muestral $n_{m.a.s.}$, es

ISBN 9788496866133

$n_c \simeq n_{m.a.s.}(1 + (\overline{N} - 1)\delta)$.

Demostración.

Se utilizará que $\dfrac{L-n}{L} = \dfrac{N-n\overline{N}}{N}$, y sea $n\overline{N} = n_c$ el tamaño muestral final en muestreo monoetápico.

Al ser las dos varianzas

$$V(\overline{y}_c) \simeq \frac{N-n_c}{N}\frac{S^2}{n_c}(1 + (\overline{N} - 1)\delta)$$

y

$$V(\overline{y}_s) = \frac{N-n_{m.a.s.}}{N}\frac{S^2}{n_{m.a.s.}}$$

entonces, para tener la misma precisión,ha de ser

$$\frac{N-n_c}{N}\frac{S^2}{n_c}(1 + (\overline{N} - 1)\delta) \simeq \frac{N-n_{m.a.s.}}{N}\frac{S^2}{n_{m.a.s.}}.$$

Despreciando los términos de corrección por población finita,

$$\frac{1}{n_c}(1 + (\overline{N} - 1)\delta) \simeq \frac{1}{n_{m.a.s.}}$$

y por lo tanto

$n_c \simeq n_{m.a.s.}(1 + (\overline{N} - 1)\delta)$.

La expresión $\dfrac{V(\overline{y}_c)}{V(\overline{y}_s)} \simeq (1 + (\overline{N} - 1)\delta)$ es denominada "efecto de diseño". El coeficiente δ suele decrecer cuando el tamaño medio de los conglomerados $\overline{N}$ crece , pues se va haciendo mayor la heterogeneidad interna, pero en términos relativos el factor $(1 + (\overline{N} - 1)\delta)$ suele crecer con $\overline{N}$. En todo caso, para un estudio previo es necesario hallar un equilibrio entre tamaño muestral, coste, precisión y tamaño de los conglomerados.

Normalmente la comparación directa con m.a.s. en términos de precisión no tiene interés práctico, pues el muestreo por conglomerados tiene mucho menor coste que el m.a.s. Además, en grandes poblaciones el muestreo aleatorio simple es frecuentemente inutilizable por motivos prácticos, de coste y de falta de información. Veremos a continuación cómo obtener una aproximación al tamaño muestral óptimo en muestreo monoetápico para una precisión y/o un coste dados.

Tamaño muestral para un error de muestreo dado ϕ.

Recordemos que n_c es el número de unidades elementales obtenidas en muestreo por conglomerados monoetápico, de n conglomerados seleccionados mediante m.a.s.

El error de muestreo ϕ es la desviación típica del estimador. En este caso $\phi = \sqrt{V(\overline{y}_c)}$ y entonces $\phi^2 = V(\overline{y}_c)$. Así,tomando la expresión aproximada de $V(\overline{y}_c)$:

$\dfrac{N-n_c}{N}\dfrac{S^2}{n_c}(1 + (\overline{N} - 1)\delta) = \phi^2$ y despejando, se obtiene

ISBN 9788496866133

$$n_c = \frac{N(1+(\overline{N}-1)\delta)S^2}{N\phi^2+(1+(\overline{N}-1)\delta)S^2}.$$

Como $n_c = n\overline{N} = n\dfrac{N}{L}$, siendo n el número de conglomerados seleccionados, se obtiene

$$n = L\frac{(1+(\overline{N}-1)\delta)S^2}{N\phi^2+(1+(\overline{N}-1)\delta)S^2} = L\frac{(1+(\frac{N}{L}-1)\delta)S^2}{N\phi^2+(1+(\frac{N}{L}-1)\delta)S^2}.$$

Si el número de conglomerados L es conocido, n está determinado por esta expresión, siempre que se tenga una buena aproximación a S^2. Si se pueden adoptar diferentes configuraciones de conglomerados de igual tamaño, existen varias soluciones que se pueden presentar en una tabla, asociando a cada L un n que arroje la precisión requerida, y asociando probablemente el coste. En este último caso, además, cada diferente construcción de conglomerados lleva a un diferente δ. A menudo esta información no está estimada para todas las posibilidades de configuraciones de conglomerados, con lo que se suele prefijar el número de conglomerados L en un valor que permite a través de información anterior, tener estimada δ (y además, como es habitual, S^2).

Si se desprecia el término de corrección por población finita, queda

$$n_c \simeq \frac{S^2(1+(\overline{N}-1)\delta)}{\phi^2}$$

con lo que

$$n \simeq \frac{1}{\overline{N}}\frac{S^2(1+(\overline{N}-1)\delta)}{\phi^2}.$$

Las mismas consideraciones respecto a L son válidas en esta última aproximación.

Tamaño muestral para un coste dado C.

En este caso es necesario expresar C en función de los diferentes costes que surgen en este tipo de muestreo. Una función de coste sencilla es

$$C = c_0 + nc_1 + n\overline{N}c_2,$$

donde c_0 es un coste fijo , c_1 es el coste asociado a cada conglomerado (viaje, requerimiento de información, etc.) y c_2 es el coste asociado a cada unidad elemental (toma de datos, viaje dentro del conglomerado, costes administrativos, etc.) dentro de los conglomerados. Es directo que

$$n = \frac{C-c_0}{c1+\overline{N}c_2}$$

por lo que si el tamaño de los conglomerados está prefijado, fijar el coste lleva a un valor de n concreto.

ISBN 9788496866133

Ejemplo 9.5

Supongamos otra vez el Ejemplo 9.2, asumiendo que una buena estimación de S^2 es $S^2 = 7$. En primer lugar, haremos un análisis a través de medios informáticos del error de muestreo obtenido para diferentes tamaños muestrales n_c. Recordemos que había 120 árboles en total. Supongamos que prefijamos el número de conglomerados en $L = 20$ de $\overline{N} = 6$ árboles cada conglomerado, y que la estimación de δ en el ejemplo 9.2 es correcta, es decir, $\widehat{\delta} = 0.87$.

El error de muestreo al cuadrado es aproximadamente

$$\phi^2 = \frac{N - n_c}{N}\frac{S^2}{n_c}(1 + (\overline{N} - 1)\delta) \simeq \frac{120 - n_c}{120}\frac{7}{n_c}(1 + (6 - 1)0.87) =$$

$$= \frac{120 - n_c}{120}\frac{37.45}{n_c} = \frac{120 - n\overline{N}}{120}\frac{37.45}{n\overline{N}} = \frac{120 - n \cdot 6}{120}\frac{37.45}{6n}.$$

Realizando un sencillo bucle de programación, se obtiene el error al cuadrado ϕ^2 para diferentes valores de $n = \frac{n_c}{\overline{N}}$. Supongamos además ciertos costes $c_0 = 3, c_1 = 2$ y $c_2 = 3$, asumiendo una función de coste lineal. Se presentará en la tabla también el valor del coste $C = c_0 + nc_1 + n\overline{N}c_2$ para cada valor de n.

n	error^2	Coste
1	5.93	23
2	2.81	43
3	1.77	63
4	1.25	83
5	0.94	103
6	0.73	123
7	0.58	143
8	0.47	163
9	0.38	183
10	0.31	203
11	0.26	223
12	0.21	243
13	0.17	263
14	0.13	283
15	0.10	303
16	0.08	323
17	0.06	343
18	0.03	363
19	0.02	383
20	0.00	403

Obsérvese que si se toman los 20 conglomerados, al ser muestreo monoetápico se está examinando toda la población y por lo tanto el error de muestreo es cero.

Si el error de muestreo está prefijado, se puede obtener a partir de la tabla el n mínimo para obtenerlo, o bien a través de la fórmula del Teorema. Por ejemplo, supongamos que queremos un error máximo de $\phi = \phi^2 = 1$. Entonces, ha de ser:

$$n = L\frac{(1 + (\overline{N} - 1)\delta)S^2}{N\phi^2 + (1 + (\overline{N} - 1)\delta)S^2} = 20\frac{37.45}{120 + 37.45} = 4.7,$$

ISBN 9788496866133

es decir n ha de ser mayor que 4, que como se ve en la tabla es el punto de corte a partir del cual el error es inferior o igual a $\phi = \phi^2 = 1$.

Otra posibilidad para la función de coste es utilizar que el coste de viaje es aproximadamente proporcional a la raíz cuadrada del número de conglomerados. Para ilustrar esta idea, considérese la figura 9.3, donde los conglomerados se reparten a la misma distancia en un área rectangular A.

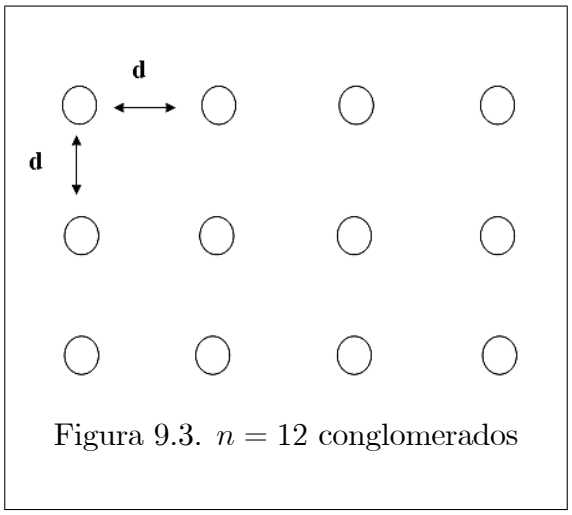


Figura 9.3. $n = 12$ conglomerados

Si la distancia horizontal o vertical entre dos conglomerados es d , y n_1 es el número de conglomerados en una fila y n_2 el número de conglomerados en una columna, la longitud del rectángulo es $d(n_1 - 1)$ y la altura será $d(n_2 - 1)$. Además, el número total de conglomerados es $n = n_1 n_2$. En el ejemplo de la figura, $n_1 = 4$, $n_2 = 3$, $n = n_1 n_2 = 12$.

El área total es

$$A = d(n_1 - 1) \times d(n_2 - 1) = d^2(n_1 - 1)(n_2 - 1).$$

Si el número de conglomerados por columna o fila es más o menos grande, se puede aproximar $A \simeq d^2 n_1 n_2$, con lo que $A \simeq d^2 n$ y la distancia entre dos conglomerados es $d \simeq \sqrt{\frac{A}{n}}$.

Si se trata de recorrer todos los conglomerados por el camino más corto, la distancia recorrida será aproximadamente $d(n-1)$, que es el número de caminos entre los conglomerados dos a dos (en el ejemplo, $d \times 11$). Aproximando $d(n-1) \simeq dn$, y como $d \simeq \sqrt{\frac{A}{n}}$, entonces la distancia recorrida en total será

$$\sqrt{\frac{A}{n}} \times n = \sqrt{nA}.$$

Por lo tanto, como el coste relativo al viaje entre conglomerados es aproximadamente proporcional a la distancia recorrida, será proporcional a la raíz cuadrada del número de conglomerados. Esto se suele representar en la función de coste como un término adicional $c_0\sqrt{n}$:

$$C = c_0\sqrt{n} + nc_1 + n\overline{N}c_2.$$

ISBN 9788496866133

El valor de c_0 tiene en cuenta las consideraciones realizadas (el área A) y el coste de cada viaje. En el valor de c_1 están excluidos ahora los costes por viaje, sólo cuentan costes relativos a la adquisición de información o añadidos extra por cada conglomerado. La constante c_2 es, como anteriormente, relativa al coste de cada toma de datos y adquisición de información previa en cada unidad elemental. Hay que tener en cuenta que a menudo estos costes se consideran también en términos de horas/hombre.

En esta función de coste, se puede despejar n, haciendo $n' = \sqrt{n}$ y resolviendo la ecuación de segundo grado en n'. Si el tamaño medio de los conglomerados $\overline{N}$ (y el número de éstos , L) no está prefijado, la optimización de la varianza sujeta al coste fijo C se puede realizar también por métodos numéricos, arrojando ésta valores óptimos de $\overline{N}$. Otra posibilidad más sencilla y directa es programar la expresión del coste y de la varianza, y dar valores a $\overline{N}$, observando los valores de n y de la varianza obtenida, presentando la tabla y/o gráficos para la toma de decisiones sobre el número óptimo de conglomerados.

Para la elección del tamaño muestral, también es posible plantear el problema como minimización del coste sujeto a una precisión dada. Los desarrollos son del mismo tipo.

9.3 Conglomerados de tamaño desigual

Se estudiará a continuación el caso en que los conglomerados varían en tamaño. Si la variación en tamaño es muy grande, el hecho de que caigan unos conglomerados u otros en la muestra hará variar mucho el valor del estimador, o, dicho de otra manera, el efecto de la variación del tamaño de los conglomerados sobre la varianza del estimador puede llegar a ser importante. Hay varias maneras de intentar paliar este efecto cuando se sospecha que puede ser grave:

- Estratificación de los conglomerados por tamaño. Se realiza estimación separada en estratos formados por conglomerados de tamaño similar. Requiere conocer a priori los tamaños de todos los conglomerados, lo que en la práctica puede ser difícil, pero pueden aproximarse estos tamaños con variables auxiliares, si es solamente para la formación de estratos.

- Estimación de razón a tamaño. Es uno de los métodos más eficaces si no se puede estratificar. Normalmente el total por conglomerado está relacionado de manera proporcional con el tamaño, así que la estimación de razón está justificada.

- Estimación con probabilidades desiguales, asignando probabilidades mayores a los conglomerados de mayor tamaño. Requiere conocer a priori el tamaño de todos los conglomerados o una variable auxiliar muy correlacionada con el tamaño. Este método permite evitar el riesgo práctico que existe en muestreo aleatorio simple de que los conglomerados grandes (más importantes en términos relativos para el investigador) queden fuera de la muestra, o que puedan estar en la muestra conglomerados muy pequeños con escasa representatividad.

En el caso de conglomerados de tamaño desigual, se define como hasta ahora $\overline{N} = \dfrac{1}{L}\sum_{i=1}^{L} N_i = \dfrac{N}{L}$ el tamaño medio de los conglomerados.

ISBN 9788496866133

9.3.1 Estimación de la media

En este apartado se procederá a estudiar dos distintos tipos de estimación bajo m.a.s. de los conglomerados: estimación insesgada y estimación de razón a tamaño.

Teorema 9.8 (estimador insesgado de la media).

(a) Bajo muestreo aleatorio simple sin reemplazamiento monoetápico de conglomerados, un estimador insesgado de la media poblacional $\overline{y}$ es

$$\widehat{\overline{y}} = \frac{1}{\overline{N}n}\sum_{i=1}^{n} N_i\overline{y}_i = \frac{1}{\overline{N}n}\sum_{i=1}^{n} y_i,$$

donde y_i representa el total en el conglomerado i.

(b) La varianza de $\widehat{\overline{y}}$ es

$$V(\widehat{\overline{y}}\,) = \frac{1}{n\overline{N}^2}\frac{L-n}{L}\frac{1}{L-1}\sum_{i=1}^{L}(y_i-\overline{y}_t)^2,$$

donde $\overline{y}_t$ representa el total medio por conglomerado, es decir, $\overline{y}_t = \frac{1}{L}\sum_{i=1}^{L} y_i$.

(c) Un estimador insesgado de esta varianza es

$\widehat{V}(\widehat{\overline{y}}) = \frac{1}{n\overline{N}^2}\frac{L-n}{L}\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\widehat{\overline{y}}_t)^2$, donde $\widehat{\overline{y}}_t$ es el total medio por conglomerado estimado, $\widehat{\overline{y}}_t = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{N}\widehat{\overline{y}}$.

Demostración.

(a) $E(\widehat{\overline{y}}) = \frac{1}{\overline{N}n}\sum_{i=1}^{n} E(y_i)$. Como y_i es una variable aleatoria que toma valores $y_1,..,y_L$ con probabilidades iguales $\frac{1}{L}$, pues se trata de m.a.s. sobre los conglome-rados, se obtiene que $E(y_i) = \frac{1}{L}\sum_{j=1}^{L} y_j$. Entonces,

$$E(\widehat{\overline{y}}) = \frac{1}{\overline{N}n}\sum_{i=1}^{n}\frac{1}{L}\sum_{j=1}^{L} y_j = \frac{1}{\overline{N}}\frac{1}{L}\sum_{j=1}^{L} y_j = \frac{1}{N}\sum_{j=1}^{L} y_j = \frac{1}{N}N\overline{y} = \overline{y}.$$

(b) $V(\widehat{\overline{y}}) = \frac{1}{\overline{N}^2}V(\frac{1}{n}\sum_{i=1}^{n} y_i)$. Pero $V(\frac{1}{n}\sum_{i=1}^{n} y_i)$ es la varianza usual de la media muestral en m.a.s., considerando los conglomerados como las unidades, L como el tamaño de la población y n el de la muestra. Entonces,

$$V(\widehat{\overline{y}}) = \frac{1}{\overline{N}^2}\frac{L-n}{L}\frac{1}{n}\frac{1}{L-1}\sum_{i=1}^{L}(y_i-\overline{y}_t)^2$$

por ser la varianza de la media muestral en m.a.s. la cuasivarianza poblacional (en este caso la cuasiva-rianza de la variable aleatoria y_i), dividida por el tamaño muestral n y multiplicada por el coeficiente de corrección por población finita, $(1-f) = \frac{L-n}{L}$.

ISBN 9788496866133

(c) Con un razonamiento similar al anterior,

$$E(\widehat{V}(\widehat{\overline{y}})) = \frac{1}{n\overline{N}^2}\frac{L-n}{L}E\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\widehat{\overline{y}}_t)^2\right].$$

Se trata de la esperanza de la cuasivarianza muestral en m.a.s., con lo que será la cuasivarianza poblacional, y así,

$$E(\widehat{V}(\widehat{\overline{y}})) = \frac{1}{n\overline{N}^2}\frac{L-n}{L}\frac{1}{L-1}\sum_{i=1}^{L}(y_i-\overline{y}_t)^2 = V(\widehat{\overline{y}}).$$

El estimador anterior suele tener una varianza muy alta en el caso en que la variación en tamaño es grande. Este estimador insesgado tiene sentido sólo si el tamaño muestral de conglomerados n es relativamente grande .

Por lo tanto, conviene utilizar otro tipo de estimador en caso de tamaños muy variables. Un estimador muy utilizado es el estimador de **razón a tamaño**, considerado en el siguiente resultado.

Teorema 9.9 (Estimador de razón a tamaño).

Bajo m.a.s. de conglomerados de distinto tamaño,

(a) Un estimador de la media, basado en la técnica de estimación de razón, es

$$\widehat{\overline{y}}_R = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} N_i}.$$

(b) La varianza aproximada de este estimador es

$$V(\widehat{\overline{y}}_R) \simeq \frac{1}{\overline{N}^2}\frac{L-n}{Ln}\frac{1}{L-1}\sum_{i=1}^{L}N_i^2(\overline{y}_i-\overline{y})^2.$$

(c) Un estimador de los momentos de esta varianza es

$$\widehat{V}(\widehat{\overline{y}}_R) = \frac{1}{\overline{N}^2}\frac{L-n}{Ln}\frac{1}{n-1}\sum_{i=1}^{n}N_i^2(\overline{y}_i-\widehat{\overline{y}}_R)^2.$$

En este último caso, si no hay información sobre $\overline{N}$, éste valor se puede sustituir por $\widehat{\overline{N}} = \frac{1}{n}\sum_{i=1}^{n}N_i$.

Demostración.

(a) En primer lugar, definiendo los conglomerados como las unidades, L como el tamaño de la población y n el de la muestra aleatoria simple, y_i como la variable de interés a medir en cada unidad (conglomerado) y N_i como la variable auxiliar, tenemos que la razón poblacional

$$R = \frac{\sum_{i=1}^{L} y_i}{\sum_{i=1}^{L} N_i} = \frac{1}{N}\sum_{i=1}^{L} y_i = \overline{y},$$

ISBN 9788496866133

con lo que estimar la razón poblacional es equivalente a estimar la media poblacional. Así,

$\widehat{\overline{y}}_R = \widehat{R} = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} N_i}$ será un estimador (sesgado) de la media poblacional.

(b) Como se ha visto, la varianza aproximada del estimador de la razón se puede poner utilizando la forma (b1) indicada en el Teorema del capítulo de estimación indirecta. Utilizando la notación adecuada (L = tamaño población, N_i variable auxiliar), en este caso el $\overline{y}_R$ definido en el capítulo mencionado es $\overline{y}_R = \widehat{R}\overline{N}$ y por lo tanto $\widehat{R} = \dfrac{\overline{y}_R}{\overline{N}}$ y entonces,

$$V(\widehat{\overline{y}}_R) = V(\widehat{R}) = V(\frac{\overline{y}_R}{\overline{N}}) = \frac{1}{\overline{N}^2} V(\overline{y}_R) = \frac{1}{\overline{N}^2} \frac{L-n}{Ln} \frac{1}{L-1} \sum_{i=1}^{L} (y_i - RN_i)^2$$

El término

$$\sum_{i=1}^{L} (y_i - RN_i)^2 = \sum_{i=1}^{L} (y_i - \frac{\sum_{i=1}^{L} y_i}{\sum_{i=1}^{L} N_i} N_i)^2 = \sum_{i=1}^{L} (y_i - N_i\overline{y})^2 = \sum_{i=1}^{L} N_i^2 (\overline{y}_i - \overline{y})^2$$

con lo que queda

$$V(\widehat{\overline{y}}_R) = \frac{1}{\overline{N}^2} \frac{L-n}{Ln} \frac{1}{L-1} \sum_{i=1}^{L} N_i^2 (\overline{y}_i - \overline{y})^2.$$

(c) Es una consecuencia directa de lo anterior, y se demuestra igual, utilizando (b2) del capítulo de estimación indirecta.

Ejemplo 9.6

Una empresa desea conocer el número promedio de litros de leche consumidos al mes por familia en una comunidad de 3000 hogares, agrupados de manera natural en 150 edificios de viviendas. Se seleccionan por muestreo aleatorio simple 5 edificios y en cada uno de estos se entrevista a todas las familias. Se obtienen los datos siguientes:

ISBN 9788496866133

Edificio	Litros consumidos en total en todos los hogares	nº de hogares
1	225	15
2	420	20
3	510	30
4	60	10
5	160	20

Tabla 9.3. Consumo de leche en hogares

Se desea estimar la media poblacional de los litros por unidad elemental (familia). Obviamente en este caso el estimador tipo razón puede mejorar al estimador insesgado, al ser el tamaño de conglomerados muestreado pequeño y estar previsiblemente muy relacionado el consumo total de leche con el tamaño del conglomerado (a más hogares, más consumo). Veamos en todo caso los valores obtenidos por los dos estimadores.

Nótese que en este ejemplo se tiene $n = 5$, $N = 3000$, $L = 150$ y $\overline{N} = \dfrac{3000}{150} = 20$.

El estimador insesgado es

$$\widehat{\overline{y}} = \frac{1}{\overline{N}n}\sum_{i=1}^{n} y_i = \frac{1}{20\cdot 5}(225+420+510+60+160) = 13.75$$

Dado que el total medio estimado por conglomerado es

$$\widehat{\overline{y}}_t = \frac{1}{5}(225+420+510+60+160) = 20\cdot\widehat{\overline{y}} = 275,$$

la varianza estimada del estimador de la media será

$$\widehat{V}(\widehat{\overline{y}}) = \frac{1}{n\overline{N}^2}\frac{L-n}{L}\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\widehat{\overline{y}}_t)^2 =$$

$$=\frac{1}{5\cdot 20^2}\frac{150-5}{150}\frac{1}{5-1}\sum_{i=1}^{n}(y_i-\widehat{\overline{y}}_t)^2 = 16.7.$$

El estimador de razón a tamaño será:

$$\widehat{\overline{y}}_R = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} N_i} = \frac{(225+420+510+60+160)}{(15+20+30+10+20)} = 14.47.$$

y su varianza estimada:

$$\widehat{V}(\widehat{\overline{y}}_R) = \frac{1}{\overline{N}^2}\frac{L-n}{Ln}\frac{1}{n-1}\sum_{i=1}^{n} N_i^2(\overline{y}_i-\widehat{\overline{y}}_R)^2 =$$

ISBN 9788496866133

$$= \frac{1}{20^2}\frac{150-5}{150\cdot 5}\frac{1}{5-1}[15^2(15-14.47)^2+...+20^2(8-14.47)^2] = 5.65.$$

Aparentemente el estimador de razón a tamaño tiene mayor precisión que el estimador insesgado (recordemos que no hay certeza pues tanto $\widehat{V}(\widehat{\overline{y}})$ como $\widehat{V}(\widehat{\overline{y}}_R)$ son estimaciones).

Una manera de corroborar esto es calcular el coeficiente de correlación muestral entre y_i y N_i. Este es, en este caso, $r = 0.86$, con lo cual al ser relativamente alto, se intuye que es correcto en este caso utilizar el estimador de razón en lugar del insesgado, a pesar del sesgo.

9.3.2 Estimación del total y proporción

Se aplicarán los resultados vistos anteriormente al caso de estimación de totales y proporciones.

Propiedad 9.2 (estimación del total y proporción).

En muestreo aleatorio simple sin reemplazamiento monoetápico de conglomerados de tamaños desiguales,

a) Un estimador del total es

$$N\widehat{\overline{y}}_R = N\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} N_i},$$

con varianza $V(N\widehat{\overline{y}}_R) = N^2V(\widehat{\overline{y}}_R)$ y estimador de la varianza $N^2\widehat{V}(\widehat{\overline{y}}_R)$.

b) Supongamos que la variable y toma valores 0 ó 1 y se desea estimar la proporción p de valores 1en la población. Un estimador de esta proporción p es $\widehat{\overline{y}}_R = \widehat{p}_R = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} N_i}$. La varianza de este estimador es

$$V(\widehat{\overline{y}}_R) = V(\widehat{p}_R) = \frac{1}{\overline{N}^2}\frac{L-n}{Ln}\frac{1}{L-1}\sum_{i=1}^{L} N_i^2(p_i-p)^2$$

donde p_i es la proporción en el conglomerado i. El estimador de la varianza será

$$\widehat{V}(\widehat{p}_R) = \frac{1}{\overline{N}^2}\frac{L-n}{Ln}\frac{1}{n-1}\sum_{i=1}^{n} N_i^2(p_i-\widehat{p}_R)^2.$$

Ejemplo 9.7

Supongamos que se desea estimar por muestreo por conglomerados monoetápico, la población de hombres en España en 1998, tomando como conglomerados las provincias y unidades elementales los municipios. En cada provincia seleccionada se realiza un estudio completo por municipio para calcular la población de hombres en esa provincia. Se sabe que hay 8.098 municipios en toda España.

ISBN 9788496866133

Supongamos que de las 52 provincias se plantea escoger $n = 10$. Obsérvese el principal problema del muestreo por conglomerados, y es que la variabilidad entre conglomerados es muy alta al haber provincias mucho más pobladas que otras. Evidentemente esto se puede corregir utilizando estimaciones tipo razón, o como veremos más adelante, probabilidades desiguales, asignando probabilidades mayores a las provincias con mayor valor en una variable auxiliar relacionada con la población de hombres.

Una cuestión interesante en este caso es que el estimador de razón a tamaño no tiene a priori por qué mejorar al estimador insesgado, pues el que una provincia tenga más unidades elementales, que en este caso son los municipios, no tiene por qué redundar en mayor valor de la variable de interés . Por ejemplo, en provincias con ciudades grandes como Madrid o Barcelona no es necesario que haya muchos municipios para obtener un valor muy grande en la variable de interés.

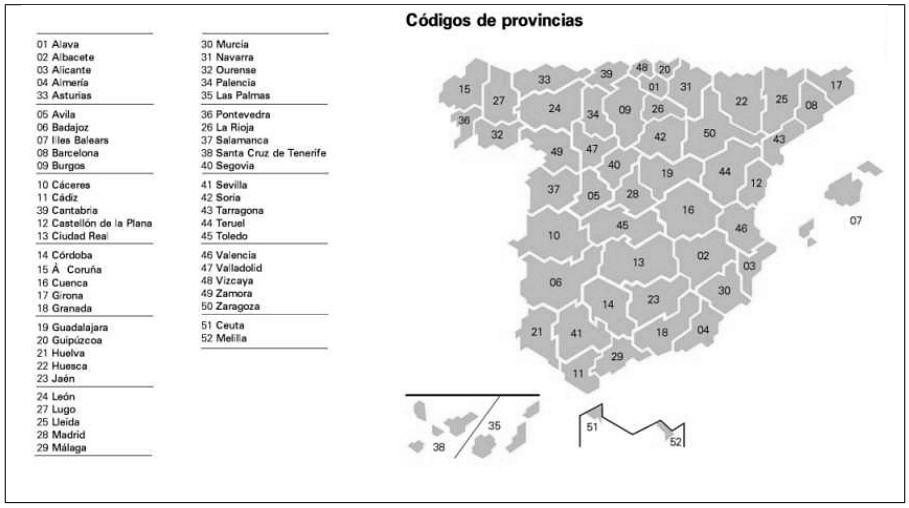


Figura 9.4. 52 Provincias españolas

Las provincias obtenidas por m.a.s., con el valor de la población de hombres y el número de municipios se presentan a continuación:

ISBN 9788496866133

Provincia	Pob. hombres	nº municipios
A Coruña	528.571	94
Almería	225.388	102
Badajoz	321.189	163
Tarragona	268.596	183
Barcelona	2.265.359	310
Murcia	514.527	45
Valencia	1.033.149	265
Cantabria	257.805	102
Málaga	570.326	100
Segovia	73.618	208

Tabla 9.4. Provincias muestreadas

En este ejemplo, $n = 10$, $N = 8.098$, $L = 52$, y $\overline{N} = \dfrac{N}{L} = 155.73$.

El estimador insesgado de la media de hombres por municipio será

$$\widehat{\overline{y}} = \frac{1}{\overline{N}n} \sum_{i=1}^{n} y_i = 3890.4$$

y por lo tanto el estimador del total de la población de hombres en España es $N\widehat{\overline{y}} = 31.504.501$.

Dado que $\widehat{\overline{y}}_t = \overline{N}\widehat{\overline{y}} = 605852$, la varianza estimada del estimador de la media es

$$\widehat{V}(\widehat{\overline{y}}) = \frac{1}{n\overline{N}^2} \frac{L-n}{L} \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \widehat{\overline{y}}_t)^2 = 1367721.4$$

y por lo tanto, la varianza del estimador del total es $8098^2 \widehat{V}(\widehat{\overline{y}}) = 8.96 * 10^{13}$

Si se utiliza el estimador de razón a tamaño,

$$\widehat{\overline{y}}_R = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} N_i} = 3854$$ y entonces el estimador del total es $N \cdot 3854 = 31.209.692$

y

$$\widehat{V}(\widehat{\overline{y}}_R) = \frac{1}{\overline{N}^2} \frac{L-n}{Ln} \frac{1}{n-1} \sum_{i=1}^{n} N_i^2 (\overline{y}_i - \widehat{\overline{y}}_R)^2 = 808910.5$$

y por lo tanto la varianza estimada del estimador del total es $8098^2 \widehat{V}(\widehat{\overline{y}}_R) = 5.30 * 10^{13}$

ISBN 9788496866133

La varianza en estimación de razón a tamaño parece bastante menor que la del estimador insesgado. El coeficiente de correlación muestral entre la población de hombres y el número de municipios es $r = 0.65$, con lo cual al final sí parece que el estimador de razón también mejora al insesgado, al igual que en el ejemplo anterior.

Se sabe por otra parte, que el número de hombres en España en 1998 es 19.488.465, con lo cual con ambos estimadores estaríamos sobreestimando ese valor. Utilizando probabilidades desiguales se puede corregir ese problema, como se verá más adelante.

9.3.3 Comparaciones entre los dos estimadores

Se han visto dos alternativas para estimar las características poblacionales en caso de que los conglomerados sean de distinto tamaño. Aunque en general el método de razón tiende a tener mejores resultados, conviene tener en cuenta las siguientes consideraciones:

1. Para construir el estimador insesgado $\widehat{\overline{y}}$ sólo es necesario conocer el tamaño poblacional N. En el estimador de razón $\widehat{\overline{y}}_R$ es necesario conocer los tamaños de cada uno de los conglomerados muestreados.
2. Para estimar la varianza de $\widehat{\overline{y}}_R$, si no se dispone del tamaño poblacional N (y por lo tanto no se conoce el tamaño medio $\overline{N} = \dfrac{N}{L}$), se puede estimar $\overline{N}$ por la media muestral $\widehat{\overline{N}} = \dfrac{1}{n}\sum_{i=1}^{n} N_i$.
3. La estimación de razón tiene sesgo, y tanto la varianza expuesta como la estimación de la va-rianza son aproximaciones. Pero en la práctica el sesgo es pequeño, por ser los N_i aproximadamente proporcionales a los y_i, y la estimación de varianza suele ser una buena aproximación.
4. Es inmediato comprobar que si los tamaños de los conglomerados son iguales, $N_i = \overline{N}$ para todo i, entonces coinciden los tres estimadores vistos $\widehat{\overline{y}} = \widehat{\overline{y}}_R = \overline{y}_c$.

9.3.4 Muestreo monoetápico con probabilidades desiguales y reemplazamiento

Otra manera de evitar el efecto perjudicial que puede tener sobre la estimación la diferencia de tamaños entre conglomerados es asignar probabilidades proporcionales a una variable auxiliar x relacionada con la variable de interés, o proporcionales al tamaño si se sabe que éste está relacionado con y. Los resultados son equivalentes a los estudiados en muestreo con probabilidades desiguales, pues cada conglomerado i puede tomarse como una gran unidad elemental, con característica de interés su total y_i.

Teorema 9.10 (estimador del total y varianzas).

Supongamos que se seleccionan n conglomerados de los L posibles, según un esquema de muestreo con probabilidades respectivas p_i y reemplazamiento. Dentro de cada conglomerado se examinan todas las unidades elementales.

ISBN 9788496866133

(a) Un estimador insesgado para el total $N\overline{y}$ es $t_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$ donde y_i es el total en el conglomerado i.

(b) La varianza de t_{HH} es $V(t_{HH}) = \frac{1}{n}\left(\sum_{i=1}^{L}\frac{y_i^2}{p_i} - (N\overline{y})^2\right)$.

(c) Un estimador insesgado de $V(t_{HH})$ es $\widehat{V}(t_{HH}) = \frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\frac{y_i^2}{p_i^2} - nt_{HH}^2\right)$

(d) Un estimador insesgado para la media poblacional es $\frac{1}{N}t_{HH}$, con varianza $\frac{1}{N^2}V(t_{HH})$ y estimador insesgado de la varianza $\frac{1}{N^2}\widehat{V}(t_{HH})$.

(e) Si la variable y es una variable con valores $0, 1$ y se trata de estimar la proporción p de unos en la población, se aplica la misma expresión $\frac{1}{N}t_{HH}$ que en apartado (d) y las varianzas de acuerdo con esto. Hay que notar que en este caso t_{HH}, que es un estimador del total, estimaría el número de valores "1" en la población.

Demostración.

Basta definir la población como una población de L unidades cada una con un valor de interés y_i, y donde se muestrean por muestreo con probabilidades desiguales y con reemplazamiento n conglomerados. Como se verifica que $\sum_{i=1}^{L} y_i = N\overline{y}$, el estimador habitual de Hansen-Hurwitz del total del valor de interés y_i en esos L conglomerados, será a su vez un estimador del total de la población $N\overline{y}$. Las varianzas y estimadores de ellas son las habituales de la estimación de Hansen-Hurwitz.

Cuando se realiza muestreo aleatorio simple sobre los conglomerados este puede hacerse con reemplazamiento. Se sabe que este tipo de muestreo es un caso particular de muestreo pptr asumiendo las probabilidades iguales. El siguiente resultado expone este caso particular.

Teorema 9.11 (estimación en m.a.s.r.).

Supongamos que se seleccionan n conglomerados de los L posibles, según un esquema de muestreo aleatorio simple con reemplazamiento. Dentro de cada conglomerado se examinan todas las unidades elementales.

(a) Un estimador insesgado para el total $N\overline{y}$ es $t_{HH} = \frac{L}{n}\sum_{i=1}^{n} y_i = L\widehat{\overline{y}}_t$.

(b) La varianza de t_{HH} es $V(t_{HH}) = \frac{1}{n}\left(L\sum_{i=1}^{L} y_i^2 - (N\overline{y})^2\right)$.

(c) Un estimador insesgado de $V(t_{HH})$ es $\widehat{V}(t_{HH}) = \frac{L^2}{n(n-1)}\left(\sum_{i=1}^{n} y_i^2 - n\widehat{\overline{y}}_t^2\right)$

(d) Los resultados para media y proporción se aplican al igual que en los apartados (d) y (e) del teorema anterior.

ISBN 9788496866133

Demostración.

El muestreo aleatorio simple con reemplazamiento es equivalente a asignar probabilidades $p_i = \frac{1}{L}$ para todo i. Basta sustituir este valor en las expresiones habituales de la estimación de Hansen-Hurwitz, en el Teorema anterior.

El siguiente resultado es otro caso particular, y es útil cuando disponemos de información directa sobre el tamaño de los conglomerados, pues en este caso suele convenir asignar probabilidades proporcionales al tamaño.

Teorema 9.12 (estimación con probabilidades proporcionales al tamaño).

Supongamos que se asignan a cada conglomerado probabilidades proporcionales a su tamaño, es decir $p_i = \frac{N_i}{N}$. Entonces

(a) Un estimador insesgado para el total $N\overline{y}$ es $t_{HH} = N\frac{1}{n}\sum_{i=1}^{n}\overline{y}_i$

(b) La varianza de t_{HH} es $V(t_{HH}) = \frac{N}{n}\sum_{j=1}^{L} N_i\left(\overline{y}_i - \overline{y}\right)^2$.

(c) Un estimador insesgado de $V(t_{HH})$ es $\widehat{V}(t_{HH}) = \frac{N^2}{n(n-1)}\sum_{i=1}^{n}(\overline{y}_i - \frac{t_{HH}}{N})^2$.

Demostración.

(a) Es inmediato, sustituyendo $p_i = \frac{N_i}{N}$ en la expresión de t_{HH}.

(b) Se sustituye $p_i = \frac{N_i}{N}$ en la expresión alternativa de la varianza

$$V(t_{HH}) = \frac{1}{n}\sum_{j=1}^{L}\left(\frac{y_j}{p_j} - N\overline{y}\right)^2 p_j,$$

y utilizando que y_i =total en el conglomerado $= N_i\overline{y}_i$.

(c) Se sustituye $p_i = \frac{N_i}{N}$ en la expresión alternativa de la estimación de la varianza

$$\widehat{V}(t_{HH}) = \frac{1}{n(n-1)}\sum_{j=1}^{n}\left(\frac{y_j}{p_j} - t_{HH}\right)^2.$$

Ejemplo 9.8

En el ejemplo 9.7 se comentaba que la diferencia entre conglomerados podría ser muy grande, pues la presencia o ausencia de provincias como Madrid o Barcelona en la muestra final de conglomerados es determinante para grandes variaciones en la estimación final. Una posibilidad es asignar probabilidades proporcionales a la población total en cada provincia, suponiendo que se dispone de esta información. Así, las probabilidades iniciales de selección quedarían:

ISBN 9788496866133

Provincia	p_i
Madrid	0.124
Barcelona	0.115
...	...
Melilla	0.00153

Tabla 9.5. Probabilidades de selección

Se ha realizado la selección de 10 provincias, con reemplazamiento y probabilidades p_i proporcionales al número de habitantes en la provincia. Al ser muestreo con reemplazamiento, algunas provincias han sido seleccionadas más de una vez. En la siguiente tabla quedan los resultados:

Provincia	Pob. hombres	p_i	nº de veces seleccionada
Madrid	2.444.919	0.124	2
Barcelona	2.261.746	0.115	2
Valencia	1.060.156	0.054	1
Vizcaya	553.761	0.028	1
Murcia	551.343	0.027	1
Granada	391.867	0.020	1
Lleida	178.152	0.009	1
Segovia	73.410	0.003	1

Tabla 9.5. Provincias muestreadas y probabilidades de selección

El estimador del total es

$$t_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} = \frac{1}{10}(2\cdot\frac{2444919}{0.124} + ... + \frac{73410}{0.003}) \simeq 20.245.671$$

Y su varianza estimada es:

$$\widehat{V}(t_{HH}) = \frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\frac{y_i^2}{p_i^2} - nt_{HH}^2\right) = 225860525988 = 2.25 * 10^{11}.$$

Aunque se trata de una estimación de la varianza, nos da una idea de que su magnitud es mucho menor que cuando se utilizaba m.a.s., sin información auxiliar, y estimaciones insesgadas o de razón

ISBN 9788496866133

a tamaño. Por supuesto, en este caso la correlación entre la población de hombres y p_i es muy alta al estar p_i construida a partir de la población total. Recordemos que el número real de hombres en España en 1998 es 19.488.465, con lo cual la estimación en este caso sería bastante correcta.

Se puede observar en la Figura 9.3 la relación de proporcionalidad entre el número de hombres en cada conglomerado (provincia).y p_i.

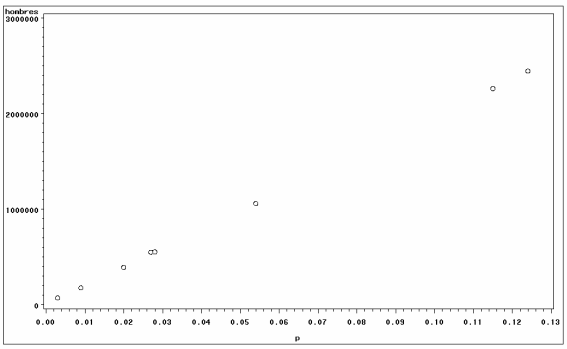


Figura 9.5. Relación entre y_i y p_i.

Si suponemos normalidad del estimador, un intervalo de confianza al 95% basado en el estimador de Hansen Hurwitz será

$(20.245.671 \pm 1.96 \cdot 475247.8) = (19.314.185, 21.177.156)$.

Hay que observar en este ejemplo que el estimador de Hansen Hurwitz no utiliza el número de municipios en cada provincia para la estimación. Y es que de alguna manera esa información está implícita en los totales por provincia y_i.

9.3.5 Muestreo monoetápico con probabilidades desiguales y sin reemplazamiento

Igualmente que en el caso anterior, todos los resultados vistos al estudiar muestreo con probabilidades desiguales sin reemplazamiento son aplicables aquí, simplemente considerando L como el tamaño poblacional y el total por conglomerado y_i como la característica de interés. El siguiente teorema es válido para el total, pero las habituales correcciones lo hacen válido para estimación de media y proporción.

Teorema 9.13 (estimación en muestreo sin reemplazamiento).

Supongamos que se seleccionan n conglomerados de los L posibles, según un esquema de muestreo con probabilidades respectivas de inclusión π_i y π_{ij} y sin reemplazamiento. Dentro de cada conglomerado se examinan todas las unidades elementales.

(a) $t_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$ es un estimador insesgado de $N\overline{y}$.

(b) $V(t_{HT}) = \sum_{i=1}^{L} \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_{i,j,i\neq j}^{L} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j} y_i y_j$

ISBN 9788496866133

(c) $V(t_{HT})_{YG} = \frac{1}{2}\sum_{i,j,i\neq j}^{L}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 = \sum_{i<j}^{L}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$

(d) $\widehat{V}(t_{HT}) = \sum_{i=1}^{n}\frac{(1-\pi_i)}{\pi_i^2}y_i^2 + \sum_{i\neq j}^{n}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\frac{y_iy_j}{\pi_i\pi_j}$

(e) $\widehat{V}(t_{HT})_{YG} = \frac{1}{2}\sum_{i,j,i\neq j}^{n}\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 = \sum_{i<j}^{n}\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$

Donde $V(t_{HT})_{YG}$ es la expresión de la varianza en la forma de Yates-Grundy y se sabe que $V(t_{HT})_{YG} = V(t_{HT})$.

Demostración.

Al igual que en el caso de muestreo con reemplazamiento, se define la población como una población de L unidades una con característica y_i, y donde se escogen, por muestreo con probabilidades desiguales y sin reemplazamiento, n conglomerados. Como se verifica que $\sum_{i=1}^{L} y_i = N\overline{y}$, el estimador habitual de Horvitz-Thompson del total de la característica y_i en esos L conglomerados, será a su vez un estimador del total de la población $N\overline{y}$.

9.3.6 Tamaño de la muestra

Si los conglomerados son de tamaño desigual y estamos utilizando muestreo monoetápico, el tamaño muestral final será $n' = \sum_{i=1}^{n} N_i = n\widehat{\overline{N}}$ que es una variable aleatoria pues depende de qué conglomerados caigan en la muestra. Si los conglomerados ya están configurados, y $\overline{N} = \frac{1}{L}\sum_{i=1}^{L} N_i$ está prefijado y es conocido, el número de conglomerados a muestrear n correspondiente a un coste fijo esperado

$C = c_0 + nc_1 + n\overline{N}c_2$ es $n = \dfrac{C - c_0}{c_1 + \overline{N}c_2}$.

Si la función de coste esperado es $C = c_0\sqrt{n} + nc_1 + n\overline{N}c_2$, es

$$n = \left[\frac{\sqrt{c_0^2 + 4C(c_1 + \overline{N}c_2)} - c_0}{2(c_1 + \overline{N}c_2)}\right]^2.$$

Se trata de funciones de coste esperado, pues el término $n\overline{N}c_2$ es la esperanza del coste por unidad elemental, $E(n'c_2) = E(n\widehat{\overline{N}}c_2) = n\overline{N}c_2$.

Si $\overline{N}$ no está prefijado, y se puede elegir entre diferentes configuraciones de conglomerados , entonces se puede realizar la optimización respecto a las varianzas sujeto a coste fijo, o, más fácil y directo, recurrir a la enumeración informática de las dife-rentes posibilidades para la configuración de conglomerados y cálculo aproximado de la varianza y coste asociados a diferentes tamaños n.

ISBN 9788496866133

9.4 Tablas de fórmulas

TAMAÑOS DE CONGLOMERADOS IGUALES (m.a.s.)			
Parámetro poblacional	$\overline{y}$	$N\overline{y}$	p
Estimador	$\overline{y}_c = \frac{1}{n}\sum_{i=1}^{n}\overline{y}_i$	$N\overline{y}_c$	$\widehat{p}_c = \frac{1}{n}\sum_{i=1}^{n} p_i$
Varianza	$\frac{(1-f_1)}{n(L-1)}\sum_{i=1}^{L}(\overline{y}_i-\overline{y})^2$	$N^2V(\overline{y}_c)$	$\frac{(1-f_1)}{n(L-1)}\sum_{i=1}^{L}(p_i-p)^2$
Estimador de la Varianza	$\frac{(1-f_1)}{n(n-1)}\sum_{i=1}^{n}(\overline{y}_i-\overline{y}_c)^2$	$N^2\widehat{V}(\overline{y}_c)$	$\frac{(1-f_1)}{n(n-1)}\sum_{i=1}^{n}(p_i-\widehat{p}_c)^2$

Otras expresiones para la varianza del estimador son:

$$\boxed{V(\overline{y}_c) = \frac{(L-n)N\sigma^2}{L(L-1)n\overline{N}^2}(1+(\overline{N}-1)\delta)}$$

$$\boxed{V(\overline{y}_c) = \frac{L-n}{L\overline{N}n}S_b^2}$$

y

$$\boxed{V(\overline{y}_c) \simeq \frac{L-n}{L}\frac{S^2}{n\overline{N}}(1+(\overline{N}-1)\delta)}$$

Otra expresión para la varianza estimada es:

$$\boxed{\widehat{V}(\overline{y}_c) = \frac{L-n}{L\overline{N}n}s_b^2}$$

con

$$s_b^2 = \frac{\overline{N}}{(n-1)}\sum_{i=1}^{n}(\overline{y}_i-\overline{y}_c)^2.$$

Un estimador de δ es

$$\boxed{\widehat{\delta} = 1 - \frac{\overline{N}}{\overline{N}-1}\frac{L(\overline{N}-1)s_w^2}{(N-1)s^2}}$$

donde

$$s_w^2 = \frac{1}{n(\overline{N}-1)}\sum_{i=1}^{n}\sum_{j=1}^{\overline{N}}(y_{ij}-\overline{y}_i)^2$$

y

ISBN 9788496866133

$$s^2 = \frac{1}{n\overline{N}-1}\sum_{i=1}^{n}\sum_{j=1}^{\overline{N}}(y_{ij}-\overline{y}_c)^2.$$

Otro estimador de δ es

$$\boxed{\widehat{\delta} = \frac{s_b^2 - \widehat{S}^2}{(\overline{N}-1)\widehat{S}^2}}$$

donde

$$\widehat{S}^2 = \frac{L(\overline{N}-1)s_w^2 + (L-1)s_b^2}{(N-1)}.$$

TAMAÑOS DESIGUALES: ESTIMACIÓN INSESGADA (m.a.s.)			
Parámetro poblacional	$\overline{y}$	$N\overline{y}$	p
Estimador	$\widehat{\overline{y}} = \frac{1}{\overline{N}n}\sum_{i=1}^{n} y_i$	$N\widehat{\overline{y}}$	$\frac{1}{\overline{N}n}\sum_{i=1}^{n} N_i p_i$
Varianza	$\frac{(1-f_1)}{n\overline{N}^2(L-1)}\sum_{i=1}^{L}(y_i-\overline{y}_t)^2$	$N^2V(\widehat{\overline{y}})$	$V(\widehat{\overline{y}})$
Estimador de la Varianza	$\frac{(1-f_1)}{n\overline{N}^2(n-1)}\sum_{i=1}^{n}(y_i-\widehat{\overline{y}}_t)^2$	$N^2\widehat{V}(\widehat{\overline{y}})$	$\widehat{V}(\widehat{\overline{y}})$

donde

$$\overline{y}_t = \frac{1}{L}\sum_{i=1}^{L} y_i$$

e

$$\widehat{\overline{y}}_t = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

ISBN 9788496866133

TAMAÑOS DESIGUALES: ESTIMACIÓN DE RAZÓN A TAMAÑO (m.a.s.)			
Parámetro poblacional	$\overline{y}$	$N\overline{y}$	p
Estimador	$\widehat{\overline{y}}_R = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} N_i}$	$N\widehat{\overline{y}}_R$	$\widehat{p}_R = \dfrac{\sum_{i=1}^{n} N_i p_i}{\sum_{i=1}^{n} N_i}$
Varianza	$\dfrac{(1-f_1)}{\overline{N}^2 n(L-1)} \sum_{i=1}^{L} N_i^2(\overline{y}_i - \overline{y})^2$	$N^2 V(\widehat{\overline{y}}_R)$	$\dfrac{(1-f_1)}{\overline{N}^2 n(L-1)} \sum_{i=1}^{L} N_i^2(p_i - p)^2$
Estimador de la Varianza	$\dfrac{(1-f_1)}{\overline{N}^2 n(n-1)} \sum_{i=1}^{n} N_i^2(\overline{y}_i - \widehat{\overline{y}}_R)^2$	$N^2 \widehat{V}(\widehat{\overline{y}}_R)$	$\dfrac{(1-f_1)}{\overline{N}^2 n(n-1)} \sum_{i=1}^{n} N_i^2(p_i - \widehat{p}_R)^2$

TAMAÑOS DESIGUALES, PROBABILIDADES DESIGUALES, CON REEMPLAZAMIENTO			
Parámetro poblacional	$N\overline{y}$	$\overline{y}$	p
Estimador	$t_{HH} = \dfrac{1}{n} \sum_{i=1}^{n} \dfrac{y_i}{p_i}$	$\dfrac{t_{HH}}{N}$	$\dfrac{t_{HH}}{N}$
Varianza	$\dfrac{1}{n}\left(\sum_{i=1}^{L} \dfrac{y_i^2}{p_i} - (N\overline{y})^2\right)$	$\dfrac{V(t_{HH})}{N^2}$	$\dfrac{V(t_{HH})}{N^2}$
Estimador de la Varianza	$\dfrac{1}{n(n-1)}\left(\sum_{i=1}^{n} \dfrac{y_i^2}{p_i^2} - n t_{HH}^2\right)$	$\dfrac{\widehat{V}(t_{HH})}{N^2}$	$\dfrac{\widehat{V}(t_{HH})}{N^2}$

ISBN 9788496866133

TAMAÑOS DESIGUALES, PROBABILIDADES DESIGUALES SIN REEMPLAZAMIENTO			
Parámetro poblacional	$N\overline{y}$	$\overline{y}$	p
Estimador	$t_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$	$\frac{t_{HT}}{N}$	$\frac{t_{HT}}{N}$
Varianza	$\sum_{i=1}^{L} \frac{1-\pi_i}{\pi_i} y_i^2 + 2 \sum_{i<j}^{L} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j} y_i y_j$	$\frac{V(t_{HT})}{N^2}$	$\frac{V(t_{HT})}{N^2}$
Varianza (Yates-Grundy)	$\sum_{i<j}^{L} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$	$\frac{V(t_{HT})_{YG}}{N^2}$	$\frac{V(t_{HT})_{YG}}{N^2}$
Estimador de la Varianza	$\sum_{i=1}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + 2 \sum_{i<j}^{n} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}} \frac{y_i y_j}{\pi_i\pi_j}$	$\frac{\widehat{V}(t_{HT})}{N^2}$	$\frac{\widehat{V}(t_{HT})}{N^2}$
Estimador de la Varianza (Yates-Grundy)	$\sum_{i<j}^{n} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$	$\frac{\widehat{V}(t_{HT})_{YG}}{N^2}$	$\frac{\widehat{V}(t_{HT})_{YG}}{N^2}$

ISBN 9788496866133

9.5 Obtención de muestras en muestreo monoetápico de conglomerados con SAS

9.5.1 Muestreo aleatorio simple de conglomerados

Para extraer una muestra de conglomerados en muestreo monoetápico se utilizará la macro extramono. La sintaxis es la siguiente:

```
%extramono(archivo1,archivo2,codif,varconglo,nconglo,n,semilla);
```

donde

archivo1 es el archivo que contiene la información poblacional.

archivo2 es el archivo que contendrá la muestra.

codif es un archivo de salida que contiene la recodificación de los conglomerados.

varconglo es la variable que indica el conglomerado en el archivo poblacional.

nconglo es el número de conglomerados diferentes.

n es el tamaño muestral.

semilla es la semilla de aleatorización para obtener la muestra. Si se desea variar de semilla cada vez, sin tener un control sobre ella, se puede poner el valor cero, de modo que se utilizará el reloj del ordenador como semilla.

La variable de conglomerado varconglo puede ser numérica o alfanumérica.

Si por ejemplo se dispone del archivo con la información poblacional data1 y la variable que indica los conglomerados se llama provincia, y se desea una muestra en el archivo muestra1, de 5 conglomerados de los 55 existentes , y se fija la semilla en el número 44444444, y el archivo de codificación se llamará codif, la sintaxis es:

```
%extramono(data1,muestra1,codif,provincia,55,5,44444444);
```

La macro extramono obtiene el archivo muestral en muestreo por conglomerados monoetápico. Este archivo contiene todas las observaciones de cada uno de los conglomerados seleccionados.

A veces el investigador realiza el proceso de escoger los conglomerados de un listado de conglomerados, sin tener la información completa de los items que forman cada conglomerado. En este caso, supongamos que tiene un archivo con la variable de identificación de conglomerado y una observación por conglomerado, y desea escoger n conglomerados por m.a.s. Entonces no es necesaria la macro extramono, basta con utilizar el proc surveyselect para escoger por m.a.s. los n conglomerados. La macro extramono está orientada a archivos que contienen la identificación de los conglomerados pero también la de todos los items en cada conglomerado.

ISBN 9788496866133

9.5.2 Muestreo de conglomerados con probabilidades desiguales, con o sin reemplazamiento

En este caso se utilizará la macro extramonoppt. La información sobre la variable auxiliar base para las probabilidades de selección puede estar incluida en el archivo de información poblacional (en cuyo caso tiene que ser constante dentro de cada conglomerado) o bien en un archivo aparte, donde aparece la variable de índice de conglomerado varconglo, y la variable auxiliar. Recordemos que como variable auxiliar es posible aportar directamente las probabilidades p_i pues no afectará al programa y es correcto igualmente.

La sintaxis es :

```
%extramonoppt(archivo1,archivo2,archivo3,inclusion,
codif,varconglo,variablex,nconglo,reemplazo,indicador,n,semilla);
```

donde

archivo1 es el archivo que contiene la información poblacional.

archivo2 es el archivo con la información de la variable auxiliar o pi por conglomerado(opcional).

archivo3 es el archivo de salida que contendrá la muestra.

inclusion es un archivo de salida que contiene las probabilidades de inclusión de los conglomerados muestrales (en caso de muestreo sin remplazamiento).

codif es un archivo de salida que contendrá la recodificación de los conglomerados y las probabilidades p_i para cada uno.

varconglo es la variable que indica el conglomerado en el archivo poblacional.

variablex es la variable auxiliar para las probabilidades de selección (puede ser p_i).

nconglo es el número de conglomerados diferentes .

reeemplazo:

1 Si se desea muestreo ppt CON reemplazamiento

2 Si se desea muestreo ppt SIN reemplazamiento

indicador:

1 Si la variablex está presente en el archivo poblacional, constante por conglomerado.

2 Si la variablex está en el archivo2, que contiene además la variable varconglo.

n es el tamaño muestral.

semilla es la semilla para la extracción de muestras.

La macro extramonoppt crea por lo tanto el archivo de muestra y el archivo de codificación de los conglomerados. El archivo de muestra contiene todas las observaciones de cada uno de los

ISBN 9788496866133

conglomerados seleccionados. Si alguno de éstos ha caído varias veces en la muestra (cuando se trata de muestreo con reemplazamiento), todas sus observaciones aparecen repetidas. La variable numberhits indica cuántas veces ha caído ese conglomerado en la muestra.

El archivo de codificación contiene la variable conglo, que son números enteros consecutivos que van de 1 al número de conglomerados, y la variable original (nombrada varconglo en la macro) que indica en el archivo poblacional el código del conglomerado. A cada número de conglo va asociado un código de varconglo.

Si se ha solicitado muestreo ppt sin reemplazamiento, el archivo de **inclusion** tiene la siguiente forma:

conglo	Unit	Selection Prob	JtProb_ 1	JtProb_ 2	JtProb_ 3	JtProb_ 4	JtProb_ 5
5	1	0.16984	0.00000	0.09418	0.09590	0.09946	0.11350
8	2	0.67845	0.09418	0.00000	0.44234	0.45414	0.49943
2	3	0.68850	0.09590	0.44234	0.00000	0.46419	0.50948
10	4	0.70595	0.09946	0.45414	0.46419	0.00000	0.52693
3	5	0.76783	0.11350	0.49943	0.50948	0.52693	0.00000

Conglo es la variable índice de conglomerado creada por la macro (también está en el archivo la varconglo original).

Debido al proceso de selección del surveyselect, la numeración de las probabilidaddes cambia. La variable Unit está por lo tanto referida a cada conglomerado, de manera a poder identificar las probabilidades de segundo orden.

SelectionProb es la variable que indica π_i para cada conglomerado: si se mantiene la numeración de la variable conglo, es $\pi_5 = 0.16, \pi_8 = 0.67, \pi_2 = 0.68$, etc.

Las columnas Jtprob_1 a Jtprob_5 indican las probabilidades de segundo orden. Estas se refieren a la codificación expresada por la variable Unit. Es decir, dentro de esa codificación $\pi_{23} = \pi_{32} = 0.44$. Si traducimos esa codificación a la numeración de conglo (ver columna conglo y columna Unit), sería $\pi_{82} = \pi_{28} = 0.44$.

Hemos supuesto que el archivo tiene esta ordenación para su utilización en la macro, pues la numeración correlativa (de 1 a 5) de los items que han caído en la muestra será más sencilla de tratar para el cálculo de estimaciones.

Como se comentó en la explicación de la macro extramono, si no se dispone dentro del archivo de la identificación de los items dentro de cada conglomerado, sino solamente de la identificación de cada conglomerado con una observación por conglomerado, se puede prescindir de esta macro y utilizar directamente el proc surveyselect con la opción pps_wr (con reemplazo) o bien pps (sin reemplazo) para extraer los conglomerados.

ISBN 9788496866133

9.6 Estimación en muestreo monoetápico de conglomerados con SAS

9.6.1 Los conglomerados han sido seleccionados por muestreo aleatorio simple

Se utilizará la macro estimono. La sintaxis es la siguiente:

```
estimono(archivo1,archivo2,varconglo,variablex,nconglo,n,ngrande);
```

donde

archivo1 es el archivo que contiene la muestra.

variabley es la variable de interés .

varconglo es la variable que indica el conglomerado en el archivo1.

nconglo es el número de conglomerados diferentes .

n es el tamaño muestral (número de conglomerados muestreado).

ngrande es el tamaño poblacional N.

La macro mono calcula y presenta en la ventana LOG estimadores, varianzas e intervalos de confianza al 95% de medias y totales bajo los siguientes supuestos:

- Estimación insesgada suponiendo tamaños iguales.

- Estimación insesgada suponiendo tamaños desiguales.

- Estimación de razón a tamaño.

9.6.2 Los conglomerados han sido seleccionados por muestreo ppt con reemplazamiento

Se utilizará la macro monopptr. La sintaxis es la siguiente:

```
estimonopptr(archi1,archi2,varconglo,variabley,indicador,nconglo,
n,ngrande);
```

donde:

archi1 es el archivo que contiene la información muestral.

archi2 es el archivo con la información de pi por conglomerado (opcional).

varconglo es la variable que indica el conglomerado en el archivo poblacional y archivo2.

variabley es la variable de interés.

indicador

ISBN 9788496866133

1 Si la variable de probabilidades pi está presente en el archivo poblacional, constante por conglomerado.

2 Si pi está en el archivo2, que contiene además la variable varconglo.

nconglo es el número de conglomerados diferentes .

n es el tamaño muestral.

ngrande es el tamaño poblacional .

Observaciones importantes para su correcta utilización:

- La macro monopptr utiliza la salida de la macro estimpptr, por lo cual ésta última ha de haber sido compilada anteriormente.

- Si un conglomerado ha sido tomado k veces repetidas (se trata de muestreo con reemplazamiento), sus observaciones deben aparecer también repetidas en el archivo muestral, y estas repeticiones estar numeradas en una variable llamada rep que va de 1 a k.

- El archivo muestral debe contener una variable llamada pi que indica la probabilidad de selección del conglomerado. Si el archivo muestral ha sido creado con la macro extramonoppt la variable de probabilidades pi y la variable de repeticiones k están ya creadas en el archivo muestral.

La macro monopptr calcula y presenta en la ventana LOG los estimadores, varianzas e intervalos de confianza al 95% de medias y totales suponiendo muestreo pptr proporcional a las probabilidades p_i presentes en la variable pi.

Supongamos que el archivo muestra2 ha sido generado a través de la macro extramonoppt donde se extrajeron con probabilidades p_i cinco conglomerados de los $L = 50$ existentes en una población de $N = 1000$ elementos. La variable de interés se llama ingresos, y la variable de conglomerados es municipio. La macro monopptr se ejecutaría así:

```
estimonopptr(muestra2,.,municipio,ingresos,1,50,5,1000);
```

En segundo lugar, supóngase que el archivo muestra2 no se generó con la macro extramono sino que fue aportada desde el exterior. La variable de codificación de conglomerados está presente en ese archivo muestra2 y también en el archivo codigos, que contiene la variable de conglomerados y la variable de probabilidades pi, obtenida de una variable auxiliar utilizada para la selección de conglomerados con probabilidades proporcionales a ella.

Entonces la sintaxis es:

```
estimonopptr(muestra2,codigos,municipio,ingresos,2,50,5,1000);
```

9.6.3 Los conglomerados han sido seleccionados por muestreo sin reemplazamiento

Se utilizará la macro estimonoppt. La sintaxis es :

ISBN 9788496866133

```
estimonoppt(archi1,inclusion,variabley,nconglo,n,ngrande);
```

donde

archi1 es el archivo que contiene la muestra.

inclusion es el archivo que contiene las probabilidades de inclusión (proviene de la macro extramonoppt).

variabley es la variable de interés .

nconglo es el número de conglomerados diferentes .

n es el tamaño muestral (número de conglomerados muestreado).

ngrande es el tamaño poblacional N.

Para ejecutar esta macro es necesario haber compilado previamente la macro estimppt, y normalmente se supone que la muestra proviene de una ejecución de la macro extramonoppt.

Si la muestra no proviene de la macro extramonoppt, deben cumplirse las condiciones:

- Existe una variable de conglomerados numerada de 1 a nconglo(L) y que se llama conglo, tanto en el archivo archi1 como en el archivo inclusion (como ambos son archivos muestrales no tienen por qué estar presentes todos los valores 1,...,L de conglo, sólo los conglomerados seleccionados).
- El archivo de inclusion contiene la variable conglo, la variable SelectionProb (las π_i) y las variables JtProb_1,...,JtProb_n (las π_{ij}).

Éstas últimas están numeradas de tal forma que la probabilidad π_{ij} está en la observación i y en la variable JtProb_j del archivo inclusion.

La macro monoppt calcula y presenta en la ventana LOG los estimadores, varianzas e intervalos de confianza al 95% de medias y totales suponiendo muestreo ppt con las probabilidades presentadas en el archivo inclusion. Si alguna estimación de varianza es negativa el intervalo de confianza no es presentado.

Supongamos que se ha obtenido el archivo de muestra mues5 y el archivo de probabilidades de inclusión inclu3 a través de la macro extramonoppt. La variable de interés se llama renta, y se han muestreado 6 conglomerados de los 40 existentes, en una población con 2000 unidades elementales. La sintaxis es:

```
estimonoppt(mues5,inclu3,renta,40,6,2000);
```


ISBN 9788496866133

9.7 Ejercicios resueltos

Ejercicio 8.1

En un proceso de control de calidad para piezas de repuesto de coches se dispone de 5 grandes lotes compuestos cada uno de 10 cajas de piezas. El número de piezas por caja puede variar. Se seleccionan por m.a.s. 2 lotes, y en cada uno de éstos se examinan todas las cajas, contando el número de piezas defectuosas y, de entre éstas, el número que requiere reparación forzosamente. En la tabla el número 3/8 indica que en el lote 1, en la caja 1 se encontraron 8 piezas defectuosas, de las cuales 3 necesitaban reparación.

Lote/Caja	1	2	3	4	5	6	7	8	9	10
1	3/8	2/6	5/7	3/6	2/4	0/3	2/5	3/7	4/8	2/4
2	2/4	3/5	0/2	1/2	2/3	0/0	1/2	3/5	4/6	5/7

a) Estimar el número total de piezas que necesitan reparación en los 5 lotes y dar un I.C. al 95%.

b) Estimar el número medio por lote de piezas que necesitan reparación y dar un I.C. al 95%.

c) Estimar el número medio por caja de piezas defectuosas y dar un I.C. al 95%.

d) Estimar la proporción de piezas defectuosas que necesitan reparación sobre todas las defectuosas y dar un I.C. al 95%.

En este ejercicio se pondrá de relieve que las estimaciones en muestreo por conglomerados monoetápico se pueden reducir en general, a las ya estudiadas en temas anteriores (estimaciones bajo m.a.s., en este ejemplo). Solamente es necesario concretar bien qué es lo que se considera unidad elemental: basta definir los conglomerados como unidades elementales pues dentro de cada uno de ellos los valores totales y/o medios son fijos al ser muestreo monoetápico y tomarse todas las unidades dentro de cada conglomerado.

En todo caso, para evitar confusiones, se considerarán las dos maneras de solucionar el problema (una es definiendo los lotes como unidades elementales y asociar a cada lote el total obtenido sobre las cajas, y la otra , definiendo los lotes como conglomerados y las cajas como unidades elementales).

a) Considerando, como se ha comentado, los lotes como unidades elementales, se obtiene que al lote 1 están asociadas en total 3+...+2=26 piezas que necesitan reparación y 8+...+4=58 piezas defectuosas en total. En el lote 2 hay respectivamente 21 piezas que necesitan reparación y 36 defectuosas en total.

Llamando y_i =número de piezas que necesitan reparación en el lote i, como se ha realizado m.a.s. de 2 lotes de los 5 lotes, la estimación del total es $N\widehat{\overline{y}} = 5 \cdot \dfrac{26+21}{2} = 117.5$.

La varianza estimada del estimador es, por ser m.a.s.,

$$\widehat{V}(N\widehat{\overline{y}}) = N^2 \frac{N-n}{N} \frac{s_y^2}{n} = 5^2 \frac{5-2}{5} \frac{12.5}{2} = 93.75$$

con lo cual el I.C. vendrá dado por:

$$(117.5 - 1.96\sqrt{93.75}, 117.5 + 1.96\sqrt{93.75}) = (98.52, 136.47).$$

ISBN 9788496866133

Si se desean aplicar las fórmulas vistas en conglomerados, el resultado es el mismo. Se define como conglomerado el lote y como unidad elemental la caja: estamos en el caso de tamaños iguales, pues cada lote tiene 10 cajas. Así, $L = 5$ y $N =$ número de unidades elementales=50. El tamaño medio del conglomerado es $\overline{N} = \dfrac{N}{L} = 10$. La estimación insesgada del total es :

$$N\overline{y}_c = N\frac{1}{n}\sum_{i=1}^{n}\overline{y}_i = \frac{N}{n}\sum_{i=1}^{n}\frac{y_i}{10}$$

donde y_i es el total por lote= número de piezas que necesitan reparación en el lote i. Así,

$$N\overline{y}_c = \frac{50}{2}\sum_{i=1}^{n}\frac{y_i}{10} = 5\cdot\frac{26+21}{2} = 117.5.$$

Para la estimación de la varianza,

$$\widehat{V}(N\overline{y}_c) = N^2\frac{(1-f_1)}{n}\frac{1}{(n-1)}\sum_{i=1}^{n}(\overline{y}_i-\overline{y}_c)^2 =$$

$$=50^2\frac{(1-2/5)}{2}\frac{1}{(n-1)}\sum_{i=1}^{n}(\overline{y}_i-\overline{y}_c)^2.$$

El término $\dfrac{1}{(n-1)}\sum_{i=1}^{n}(\overline{y}_i-\overline{y}_c)^2$ es la cuasivarianza de los valores $\dfrac{26}{10}$ y $\dfrac{21}{10}$, medias por conglomerado (la estimación de la media poblacional por caja es $\overline{y}_c = \dfrac{1}{2}(\dfrac{26}{10}+\dfrac{21}{10}) = 2.35$).

Entonces $\dfrac{1}{(n-1)}\sum_{i=1}^{n}(\overline{y}_i-\overline{y}_c)^2$ =0.125 y por lo tanto $\widehat{V}(N\overline{y}_c) = 93.75$, como se vio con el otro planteamiento.

b) Habiendo estimado el total en la población, y como hay 5 lotes, la estimación es $\dfrac{1}{5}117.5. = 23.5$ y la varianza de este estimador es $\dfrac{1}{5^2}\cdot 93.75 = 3.75$. El intervalo de confianza es $(23.5 - 1.96\sqrt{3.75}, 23.5 + 1.96\sqrt{3.75}) = (19,7, 27.29)$.

c)

Si se hace considerando los conglomerados como unidades elementales, el número de piezas defectuosas en los lotes muestreados es respectivamente de 58 y 36. Entonces, el estimador del total de piezas defectuosas es $N\widehat{\overline{y}} = 5\cdot\dfrac{58+36}{2} = 235$. Como es $s_y^2 = 242$, La varianza de este estimador es $\widehat{V}(N\widehat{\overline{y}}) = 5^2\dfrac{5-2}{5}\dfrac{242}{2} = 1815$.

Si se realiza a partir de la definición de conglomerados, $N\overline{y}_c = \dfrac{50}{2}\sum_{i=1}^{n}\dfrac{y_i}{10} = 5\cdot\dfrac{58+36}{2} = 235$, y $\widehat{V}(N\overline{y}_c) = 50^2\dfrac{(1-2/5)}{2}\dfrac{1}{(n-1)}\sum_{i=1}^{n}(\overline{y}_i-\overline{y}_c)^2 = 1815$, igualmente.

Como en total hay 50 cajas, el número medio de piezas defectuosas estimado por caja es $\dfrac{235}{50} = 4.7$ y la varianza de este estimador es

$$\frac{1}{50^2}1815 = 0.726.$$

El intervalo de confianza es $(3.84, 5.55)$.

c) Tanto el número de piezas defectuosas como el número que necesita reparación son cantidades aleatorias. Para estimar la proporción hay entonces que utilizar el estimador de la razón.

En este caso lo adecuado es considerar los lotes como unidades elementales. Se define x_i como el número total de piezas defectuosas en el lote i e y_i como el número de piezas que necesitan reparación en el lote i. El estimador de la proporción en todos los lotes es el estimador de la razón

ISBN 9788496866133

$$\widehat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{26+21}{58+36} = 0.50.$$

Calculando $s_y^2 = 12.5$, $s_x^2 = 242$ y $s_{xy} = 55$, la varianza estimada de este estimador es

$$\widehat{V}(\widehat{R}) = \frac{N-n}{Nn\overline{x}^2}(s_y^2 + \widehat{R}^2 s_x^2 - 2\widehat{R}s_{xy}) = \frac{5-2}{5\cdot 2\cdot \overline{x}^2}(12.5 + 0.25\cdot 242 - 2\cdot 0.5\cdot 55).$$

La media $\overline{x}$ se desconoce, pero se estimará por la media muestral a efectos del cálculo en la varianza:$\widehat{\overline{x}} = \frac{58+36}{2} = 47$.

Así, $\widehat{V}(\widehat{R}) = 0.00244$.

Ejercicio 8.2

Se dispone de una población con 8 observaciones:

Observación	1	2	3	4	5	6	7	8
y_i	1	2	0	1	0	2	0	0

Se desea comparar la configuración de dos conglomerados siguiente:

$\{1,2,3,4\}$ y $\{5,6,7,8\}$, con la configuración $\{1,2,5,8\}$ y $\{3,4,6,7\}$.

Decir qué configuración es mejor calculando el coeficiente de correlación intraconglomerados, y calcular también S_w^2, S_b^2 en cada caso.

Para la primera configuración, se tiene que

$$\delta = \frac{\sum_{i=1}^{L}\sum_{j\neq k=1}^{\overline{N}} (y_{ij}-\overline{y})(y_{ik}-\overline{y})}{(\overline{N}-1)(L\overline{N}-1)S^2}$$

En el primer conglomerado, es

$$\sum_{j\neq k=1}^{\overline{N}} (y_{ij}-\overline{y})(y_{ik}-\overline{y}) =$$

$$= (1-1)(2-1) + (1-1)(0-1) + (1-1)(1-1) +$$

$$+(2-1)(1-1) + (2-1)(0-1) + (2-1)(1-1) +$$

$$+(0-1)(1-1) + (0-1)(2-1) + (0-1)(1-1) +$$

$$+(1-1)(1-1) + (1-1)(2-1) + (1-1)(0-1) =$$

$$= -2.$$

En el segundo, se calcula igualmente y es:

ISBN 9788496866133

$\sum_{j\neq k=1}^{\overline{N}} (y_{ij}-\overline{y})(y_{ik}-\overline{y}) = -3.$

Entonces, como $S^2 = 0.785$,

$$\delta = \frac{-2-3}{(4-1)(8-1)0.785} = -0.303.$$

Se trata de una buena configuración, pues el mínimo valor posible de δ está muy cercano pues es $\delta = -\frac{1}{\overline{N}-1} = -0.333$. Como se sabe, para valores negativos de δ, es más preciso realizar muestreo por conglomerados que m.a.s.

En esta configuración, se tiene que

$$S_w^2 = \frac{1}{L(\overline{N}-1)} \sum_{i=1}^{L} \sum_{j=1}^{N_i} (y_{ij}-\overline{y}_i)^2 = \frac{1}{2}(\frac{2}{3}+1) = 0.833$$

y

$$S_b^2 = \frac{\overline{N}}{(L-1)} \sum_{i=1}^{L} (\overline{y}_i - \overline{y})^2 = 0.5$$

Se observa que en la descomposición de la varianza, el término relativo a la varianza dentro de conglomerados es muy alto respecto al término relativo a la varianza entre conglomerados:

$(N-1)S^2 = L(\overline{N}-1)S_w^2 + (L-1)S_b^2 = 5 + 0.5.$

Para la segunda configuración, el sumatorio del numerador es -2.75 tanto para el primer conglomerado como para el segundo. Así:

$$\delta = \frac{2 \cdot -2.75}{(4-1)(8-1)0.785} = -0.33$$

que es la mejor configuración posible. En este caso,

$$S_w^2 = \frac{1}{2}(0.9166 + 0.9166) = 0.9166$$

y

$S_b^2 = 0$

pues las medias en cada conglomerado de esta configuración son iguales. Así, los conglomerados son lo más parecidos posible entre sí y variables internamente.

Ejercicio 8.3

En un estudio realizado en una comunidad de vecinos, se escogen 2 edificios por m.a.s., de entre los 6 edificios que contiene la comunidad. En los edificios muestreados se examinan todas las viviendas, con la particularidad de que algunas están vacías permanentemente y se cuenta como si no existieran. En las viviendas no vacías examinadas, se cuenta si hay perros o no. Se obtienen los siguientes datos: Primer edificio muestreado: hay 15 viviendas no vacías, en las cuales hay 3 con perros. Segundo edificio: hay 20 viviendas no vacías, de las cuales 5 tienen perro.

Estimar la proporción de viviendas habitadas que tienen perro y dar un I.C. al 95%.

ISBN 9788496866133

Se trata de un estudio por conglomerados de distinto tamaño, pues no se consideran las unidades (viviendas) que están vacías. En este caso existen dos estimaciones posibles: la estimación insesgada y la de razón a tamaño. como a priori no se conoce N, es decir, el número de viviendas no vacías, se recurrirá al estimador de razón a tamaño.

Se tiene:

$$\widehat{p}_R = \frac{\sum_{i=1}^{n} N_i p_i}{\sum_{i=1}^{n} N_i} = \widehat{p}_R = \frac{3+5}{15+20} = 0.228.$$

La varianza de este estimador se estima por:

$$\widehat{V}(\widehat{p}_R) = \frac{(1-f_1)}{\overline{N}^2 n(n-1)} \sum_{i=1}^{n} N_i^2 (p_i - \widehat{p}_R)^2 =$$

$$= \frac{(1-2/6)}{17.5^2 2(2-1)} [15^2(0.2-0.228)^2 + 20^2(0.25-0.228)^2] = 0.0004,$$

donde se ha estimado $\overline{N}$ por el valor medio del tamaño de los conglomerados de la muestra, es decir $\widehat{\overline{N}} = \frac{15+20}{2} = 17.5$.

El intervalo de confianza correspondiente es $(0.188, 0.267)$.

Ejercicio 8.4

En un colegio se desea estimar la estatura media de los niños de tercer curso. Hay 7 clases de diferentes tamaños, con un total de 220 alumnos, y se seleccionan 3 clases por m.a.s. En la primera clase muestreada hay 20 alumnos, y se obtiene que la media de estatura es de 140 cm., en la segunda hay 35 alumnos, con una media de estatura de 155 cm., y en la tercera hay 28 alumnos con una media de 135. Dar una estimación de la media de estatura en el colegio para los niños de tercer curso y estimar la varianza del estimador.

Al ser las clases de tamaño distinto, se puede utilizar estimación insesgada o de razón a tamaño. Se calcularán los dos estimadores.

Por estimación insesgada, se tiene que

$$\widehat{\overline{y}} = \frac{1}{\overline{N} n} \sum_{i=1}^{n} y_i = \frac{1}{\overline{N} n} \sum_{i=1}^{n} N_i \overline{y}_i.$$

Como $\overline{N} = \frac{N}{L} = \frac{220}{7} = 31.42$, es

$$\widehat{\overline{y}} = \frac{1}{31.42 \cdot 3}(20 \cdot 140 + 35 \cdot 155 + 28 \cdot 135) = 127.32.$$

Puede llamar la atención que el valor obtenido sea menor a las medias observadas en cada una de las clases muestreadas. El "mecanismo" del estimador se puede explicar obsrvando que consiste en calcular una estimación $\frac{1}{n}\sum_{i=1}^{n} y_i$ del total medio por clase, y después dividir esta cantidad entre el número de alumnos medio por clase $\overline{N}$. Si el número de alumnos por clase varía mucho, la primera estimación puede ser imprecisa y dar lugar a resultados extraños como el observado, donde las clases variaban mucho en tamaño respecto a la media.

El total medio estimado por clase es $\widehat{\overline{y}}_t = \frac{1}{n}\sum_{i=1}^{n} y_i = 4001.67$. La varianza estimada de este estimador es:

ISBN 9788496866133

$$\widehat{V}(\widehat{\overline{y}}) = \frac{(1-f_1)}{n\overline{\overline{N}}^2(n-1)} \sum_{i=1}^{n}(y_i - \widehat{\overline{y}}_t)^2 =$$

$$=\frac{(1-3/7)}{31.42^2 \cdot 3(3-1)}[(20 \cdot 140 - 4001.67)^2 + (35 \cdot 155 - 4001.67)^2 + (28 \cdot 135 - 4001.67)^2] = 339.3.$$

Si se realiza estimación de razón a tamaño, entonces:

$$\widehat{\overline{y}}_R = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} N_i} = \frac{20 \cdot 140 + 35 \cdot 155 + 28 \cdot 135}{20 + 35 + 28} = 144.63.$$

y

$$\widehat{V}(\widehat{\overline{y}}_R) = \frac{(1-f_1)}{\overline{\overline{N}}^2 n(n-1)} \sum_{i=1}^{n} N_i^2(\overline{y}_i - \widehat{\overline{y}}_R)^2 =$$

$$=\frac{(1-3/7)}{31.42^2 \cdot 3(3-1)}[20^2(140 - 144.63)^2 + 35^2(155 - 144.63)^2 + 28^2(135 - 144.63)^2] = 20.53.$$

Se observa la gran diferencia en la precisión estimada de ambos estimadores. La razón para explicarlo es que la correlación entre el total por conglomerado y_i y el tamaño del conglomerado N_i es muy alta: $r_{yN} = 0.98$. Es conocido que es mejor un estimador sesgado (el de razón a tamaño lo es) pero con pequeña varianza, que un estimador insesgado con varianza alta.

Ejercicio 8.5

En una auditoría se realiza una inspección de los gastos de electricidad de una gran empresa. Se escogen 4 meses por m.a.s. y en cada uno de esos meses se examinan todas las facturas debidas a electricidad, que es un número fijo de 70 facturas. Se observan los gastos medios por factura en euros respectivamente en cada uno de esos 4 meses: 2500 (con una cuasivarianza de 60.000), 4500 (con una cuasivarianza de 70.000), 3800 (con una cuasivarianza de 50.000), 4000 (con una cuasivarianza de 50.000).

a) Teniendo como objetivo estimar el gasto anual de la empresa en electricidad, estimar el coeficiente de correlación intraconglomerados y concluir si el método de muestreo es correcto o hubiera sido "mejor" seleccionar 4x70=280 facturas por m.a.s. de las 12x70=840 existentes, a pesar del trabajo extra que esto implicaría.

b) Dar un intervalo de confianza al 95% para los gastos totales en electricidad.

---

a) El estimador de δ que se utilizará, al no tener los datos completos de cada conglomerado, es

$$\widehat{\delta} = \frac{s_b^2 - \widehat{S}^2}{(\overline{N}-1)\widehat{S}^2}$$

donde

$$\overline{y}_c = \frac{1}{4}(2500 + 4500 + 3800 + 4000) = 3700,$$

$$s_b^2 = \frac{\overline{N}}{(n-1)} \sum_{i=1}^{n}(\overline{y}_i - \overline{y}_c)^2 = \frac{70}{(4-1)}[(2500 - 3700)^2 + ... + (4000 - 3700)^2] = 50866666.67,$$

ISBN 9788496866133

$$s_w^2 = \frac{1}{n(\overline{N}-1)} \sum_{i=1}^{n} \sum_{j=1}^{\overline{N}} (y_{ij} - \overline{y}_i)^2 = \frac{1}{4}(60000 + 70000 + 50000 + 50000) = 57500,$$

y

$$\widehat{S}^2 = \frac{L(\overline{N}-1)s_w^2 + (L-1)s_b^2}{(N-1)} = \frac{12(70-1)57500 + (12-1)50866666.67}{(840-1)} = 723651.$$

Así,

$$\widehat{\delta} = \frac{s_b^2 - \widehat{S}^2}{(\overline{N}-1)\widehat{S}^2} = \frac{50866666.67 - 723651}{(70-1)723651} = 1.004$$

con lo cual al ser mayor que cero, es preferible el m.a.s. al muestreo por conglomerados en este caso (la variabilidad entre conglomerados es muy alta).

b) El estimador del total es $N\overline{y}_c = 840 \cdot 3700 = 3108000$.

La varianza estimada de este estimador es

$$\widehat{V}(N\overline{y}_c) = N^2 \frac{(1-f_1)}{n(n-1)} \sum_{i=1}^{n} (\overline{y}_i - \overline{y}_c)^2 = 2.84 \cdot 10^{10}.$$

El intervalo de confianza para el total será $(2777199, 3438800)$.

Ejercicio 8.6

Se realiza un estudio de muestreo para estimar la nota media en selectividad en Institutos de cierta ciudad. Estos pueden ser clasificados en tres tipos: Privados, Públicos y Privados subvencionados.

En todos los institutos privados hay 60 clases de último curso de Bachillerato, con un número aproximado de 1500 alumnos en total. Se extraen 2 clases de entre todos los institutos con probabilidades proporcionales al tamaño con reposición, y en cada clase se examina a todos los alumnos. Los datos obtenidos son: tamaños de las clases muestreadas: {25,30} y media de los alumnos obtenida en esas clases: {5.2,6.4}.

En los 10 institutos públicos, que congregan alrededor de 6000 alumnos de último curso de Bachillerato, el número de alumnos de cada instituto varía considerablemente. Se seleccionan por m.a.s. 3 clases de las 240 existentes en todos los institutos , y se obtiene la nota en selectividad de cada uno de los alumnos de esas clases. La media obtenida en selectividad en cada una de esas tres clases es {5.5,6.2,7.1} y el número de alumnos, {28, 35, 20}.

De los institutos privados subvencionados, se extrae una muestra sin reemplazamiento y proporcional al número de alumnos de dos institutos de los 3 que hay en la ciudad, y en cada uno de ellos se obtienen datos de todos los alumnos, sabiendo que en total en los tres institutos hay 150 alumnos en selectividad. Se obtiene que en el primer instituto hay 50 alumnos en selectividad y la nota media obtenida es 5.5. En el segundo, hay 40 alumnos y la nota media obtenida es de 6.5.

Estimar la nota media en selectividad en la ciudad y estimar la varianza del estimador.

Se trata de un problema de muestreo estratificado. En cada estrato se ha realizado un tipo diferente de estimación. Se procederá a una estimación del total en cada estrato, para obtener finalmente una estimación

ISBN 9788496866133

del total como la suma de las estimaciones en los tres estratos, y corregir para obtener una estimación de la media.

Estrato Centros Privados:

Se ha realizado muestreo de conglomerados monoetápico con probabilidades proporcionales al tamaño y reemplazamiento, donde los conglomerados son las clases y las unidades elementales los alumnos.

Para las dos clases muestreadas, se tiene: $p_1 = \dfrac{25}{1500} = 0.0166$ y $p_2 = \dfrac{30}{1500} = 0.02$. Además, $y_1 = 5.2 \cdot 25 = 130$ e $y_2 = 6.4 \cdot 30 = 192$.

El estimador del total en ese estrato es

$$t_{HH} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i} = \frac{1}{2}(\frac{130}{0.0166} + \frac{192}{0.02}) = 8700.$$

Y el estimador de la media será $\dfrac{8700}{1500} = 5.8$.

La varianza estimada del estimador del total es

$$\widehat{V}(t_{HH}) = \frac{1}{n(n-1)} \left(\sum_{i=1}^{n} \frac{y_i^2}{p_i^2} - n t_{HH}^2 \right) = \frac{1}{2(2-1)} \left((\frac{130}{0.0166})^2 + (\frac{192}{0.02})^2 - 2 \cdot 8700^2 \right) = 810000.$$

Estrato Centros Públicos:

Se ha realizado m.a.s. y los tamaños de los conglomerados (las clases) son desiguales. Al ser muy diferentes estos tamaños, se utilizará el estimador de razón a tamaño:

$$\widehat{\overline{y}}_R = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} N_i} = \frac{5.5 \cdot 28 + 6.2 \cdot 30 + 7.1 \cdot 20}{28 + 30 + 20} = 6.179$$

y por lo tanto $N\widehat{\overline{y}}_R = 37076.92$.

La varianza estimada del total es

$$\widehat{V}(N\widehat{\overline{y}}_R) = \frac{N^2(1-f_1)}{\overline{N}^2 n(n-1)} \sum_{i=1}^{n} N_i^2 (\overline{y}_i - \widehat{\overline{y}}_R)^2 =$$

$$= \frac{6000^2(1-3/240)}{(\frac{6000}{240})^2 \cdot 3(3-1)} (28^2(5.5-6.179)^2 + 30^2(6.2-6.179)^2 + 20^2(7.1-6.179)^2) = 6646896.7$$

Estrato Centros Privados subvencionados:

En este estrato los conglomerados son los centros. Hay 3 y se escogen con probabilidades proporcionales al tamaño sin reemplazamiento. Para las estimaciones es necesario calcular las probabilidades de inclusión. Llamando 1 al primer centro muestreado y 2 al segundo, se tiene:

$p_1 = \dfrac{50}{150} = \dfrac{1}{3}$, $p_2 = \dfrac{40}{150} = 0.2666$, y $p_3 = \dfrac{60}{150} = 0.4$.

Entonces,

$$\pi_1 = \frac{1}{3}(1 + \frac{0.266}{1-0.266} + \frac{0.4}{1-0.4}) = 0.676,$$

$$\pi_2 = 0.266(1 + \frac{0.333}{1-0.333} + \frac{0.4}{1-0.4}) = 0.577,$$

y

ISBN 9788496866133

$$\pi_{12} = \frac{1}{3}0.266(\frac{1}{1-0.333} + \frac{1}{1-0.266}) = 0.254.$$

El estimador del total en los 3 centros Privados subvencionados es

$$t_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = (\frac{5.5 \cdot 50}{0.676} + \frac{6.5 \cdot 40}{0.577}) = 857.41.$$

El estimador de la media en esos centros sería $\frac{857.41}{150} = 5.71$.

El estimador de la varianza del total es:

$$\widehat{V}(t_{HT}) = \frac{(1-0.676)}{0.676^2}(5.5 \cdot 50)^2 + \frac{(1-0.577)}{0.577^2}(6.5 \cdot 40)^2 +$$

$$+2\frac{(0.254 - 0.676 \cdot 0.577)}{0.254}\frac{5.5 \cdot 50}{0.676}\frac{6.5 \cdot 40}{0.577} = -56866.$$

Da un valor negativo, por lo cual habrá que calcular la varianza en la forma de Yates-Grundy:

$$\widehat{V}(t_{HT})_{YG} = \frac{(0.676 \cdot 0.577 - 0.254)}{0.254}\left(\frac{5.5 \cdot 50}{0.676} - \frac{6.5 \cdot 40}{0.577}\right)^2 = 1027.67.$$

Finalmente, para construir un estimador global de la media, se construye primero el estimador global el total, que es la suma de los tres estimadores de los totales, y su varianza estimada será la suma de las varianzas estimadas:

$N\widehat{\overline{y}}^* = 8700 + 37076.92 + 857.41 = 46634.33.$

y

$\widehat{V}(N\widehat{\overline{y}}^*) = 810000 + 6646896.7 + 1027.67 = 7457924.37$

Puesto que en la población hay 1500+6000+150=7650, el estimador de la media será finalmente $\widehat{\overline{y}}^* = \frac{46634.33.}{7650} =$ 6.096 y su varianza estimada será $\widehat{V}(\widehat{\overline{y}}^*) = \frac{7457924.37}{7650^2} = 0.1274$.

Un intervalo de confianza al 95% para la media poblacional basado en estas estimaciones es $(5.39, 6.79)$.

Ejercicio 8.7

Se desea estudiar el tiempo que las obras realizadas en una ciudad provocan contaminación acústica por encima de un determinado nivel los días de semana. Para ello se divide la ciudad en 10 zonas, y de estas zonas se escogen 3 por m.a.s. En cada zona seleccionada se cuentan las obras y en cada una de ellas se elabora un recuento de las horas semanales que esas determinadas obras hacen ruido por encima del nivel fijado.

Se obtiene que en la primera zona de las seleccionadas, hay 5 obras, con horas semanales de ruido respectivas 17,20,5,15,10. En la segunda zona hay 3 obras, con horas 21,25, 6. En la tercera zona hay 4 obras, con horas 6, 6, 6, 8 .

a) Estimar el total de las horas de ruido semanales que se dan en todas las zonas y estimar la varianza del estimador.

b) Si la normativa dijera que en una obra no se puede llegar a más de 12 horas semanales de ruido por encima del nivel prefijado, estimar la proporción de obras que incumplen esa normativa y dar un intervalo de confianza al 95% para esa proporción, suponiendo normalidad del estimador.

ISBN 9788496866133

a) Se trata de un muestreo por conglomerados monoetápico, donde los conglomerados son las zonas y las unidades elementales las obras. Los tamaños son desiguales, y se puede optar tanto por el estimador insesgado como por el de razón a tamaño, pero en este caso se desconoce el número total de obras y por lo tanto se utilizará el de razón a tamaño. En la primera zona el total de horas es $y_1 = 67$, en la segunda $y_2 = 52$ y en la tercera $y_3 = 26$.

Se tiene:

$$\widehat{\overline{y}}_R = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} N_i} = \frac{67+52+26}{5+3+4} = 12.083.$$

La estimación del total tiene una complicación y es que se desconoce el número total de obras. Éste se puede estimar, sin embargo, por $\widehat{N} = L\widehat{\overline{N}}$, donde

$$\widehat{\overline{N}} = \frac{1}{n}\sum_{i=1}^{n} N_i = \frac{1}{3}(5+3+4) = 4$$

es la estimación muestral del número medio de obras por zona. Así, $\widehat{N} = 10 \cdot 4 = 40$ obras, y entonces $\widehat{N}\widehat{\overline{y}}_R = 483.32$.

La estimación de la varianza del estimador es :

$$\widehat{V}(\widehat{N}\widehat{\overline{y}}_R) = \widehat{N}^2 \frac{(1-f_1)}{\overline{N}^2 n(n-1)} \sum_{i=1}^{n} N_i^2 (\overline{y}_i - \widehat{\overline{y}}_R)^2$$

donde también hay que estimar $\overline{N}^2$ por $\widehat{\overline{N}}^2$. Así,

$$\widehat{V}(\widehat{N}\widehat{\overline{y}}_R) = 4^2 \cdot 10^2 \frac{(1-0.3)}{4^2 \cdot 3(3-1)} [5^2(13.4-12.083)^2 + 3^2(17.33-12.083)^2 + 4^2(6.5-12.083)^2] = 9215.$$

b) Se utiliza también el estimador de razón a tamaño. Se crea la variable $y_{ij} = 1$ si el número de horas de ruido en la obra j de la zona i excede el valor 12, e $y_{ij} = 0$ si no. La proporción estimada es:

$$\widehat{p}_R = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} N_i} = \frac{3+2+0}{5+3+4} = 0.416.$$

La varianza del estimador es:

$$\widehat{V}(\widehat{p}_R) = \frac{(1-0.3)}{4^2 3(3-1)} [5^2(0.6-0.416)^2 + 3^2(0.66-0.416)^2 + 4^2(0-0.416)^2] = 0.0304.$$

El intervalo de confianza será

$$(0.416 - 1.96\sqrt{0.0304}, 0.416 + 1.96\sqrt{0.0304}) = (0.073, 0.76).$$

Ejercicio 8.8

Se trata de estimar la media en muestreo monoetápico con $n = 2$ en una población que contiene 3 conglomerados, con los tamaños y las medias poblacionales indicadas en la tabla inferior:

ISBN 9788496866133

N_i	$\overline{y}_i$	y_i
2000	30	60000
500	60	30000
20	10	200

a) Calcular la correlación entre los N_i y las medias $\overline{y}_i$. Calcular la correlación entre los N_i y los totales y_i.

b) A la vista de lo obtenido, para estimar la media, ¿es conveniente utilizar m.a.s. con estimación insesgada, m.a.s. con estimación de razón a tamaño o bien muestreo ppt con probabilidades proporcionales al tamaño?

c) Calcular el valor de los tres estimadores de la media indicados en el apartado anterior para todas las muestras, con sus probabilidades respectivas. Resumir en una tabla, y calcular a continuación la varianza y sesgo de cada uno de los tres estimadores. Concluir.

---

a) La correlación poblacional entre N_i y las medias $\overline{y}_i$ es $\dfrac{S_{N_i\overline{y}_i}}{S_{N_i}S_{\overline{y}_i}} = \dfrac{3100}{1033 \cdot 25.16} = 0.119$. La correlación entre N_i y los totales y_i es $\dfrac{S_{N_i y_i}}{S_{N_i}S_{y_i}} = 0.959$.

b) Lo que importa en la estimación, aunque se trate de la estimación de la media, es la correlación entre la variable auxiliar y los totales por conglomerado. Esto se ve en la expresión de los estimadores: el de razón a tamaño es $\widehat{\overline{y}}_R = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} N_i}$, lo que implica asumir alta correlación entre N_i e y_i. El estimador de Horvitz Thompson para la media es $\dfrac{1}{N}t_{HT} = \dfrac{1}{N}\sum_{i=1}^{n}\dfrac{y_i}{\pi_i}$, con lo que se asume del mismo modo alta correlación entre el total y_i y las probabilidades π_i.

La razón es que aunque la media en cada conglomerado no esté relacionada con el tamaño, la contribución de la media de cada conglomerado a la media poblacional sí lo está, pues en la construcción de la media global, contribuyen en mayor medida los conglomerados grandes que los pequeños:

En el ejercicio, la media poblacional es:

$$\frac{1}{2520}(30 \cdot 60000 + 60 \cdot 30000 + 10 \cdot 20) = 35.79.$$

Se observa que la media poblacional está "más cerca" de la media de los conglomerados grandes que de las medias de los pequeños.

Por ello si el tamaño de los conglomerados es muy variable, en general es muy poco recomendable utilizar m.a.s. con estimación insesgada (a no ser que la fracción de muestreo sea muy alta). O bien se utiliza m.a.s. con estimadores de razón o bien muestreo ppt (o pptr).

Como en este caso la correlación entre los tamaños y los totales por conglomerado es alta, la estimación de razón con m.a.s. y la estimación por muestreo ppt serán preferibles al m.a.s. con estimación insesgada.

Si la relación entre N_i e y_i es prácticamente de proporcionalidad , y hay grandes diferencias de tamaño, el muestreo ppt suele ser preferible al m.a.s. con estimación de razón, como se verá en este caso.

ISBN 9788496866133

c) Se construirán las probabilidades en caso de muestreo ppt:

$p_1 = \dfrac{2000}{2520} = 0.7936.$

$p_2 = 0.1984.$

$p_3 = 0.007936.$

$\pi_{12} = p_1 \left(\dfrac{p_2}{1-p_1} + \dfrac{p_2}{1-p_2} \right) = 0.95957.$

$\pi_{13} = 0.0386.$

$\pi_{23} = 0.00355.$

Así,

$\pi_1 = \pi_{12} + \pi_{13} = 0.99645$, $\pi_2 = \pi_{12} + \pi_{23} = 0.96312$ y $\pi_3 = \pi_{13} + \pi_{23} = 0.04041$.

El estimador de la media poblacional de Horvitz Thompson para la muestra $1, 2$ se calcula como

$t_{HT(1,2)} = \dfrac{1}{2520} \left(\dfrac{60000}{0.99645} + \dfrac{30000}{0.96312} \right) = 36.24.$

Análogamente, se tiene $t_{HT(1,3)} = 25.85$ y $t_{HT(2,3)} = 14.32$.

Cada una de estos valores del estimador tiene probabilidad respectiva π_{12}, π_{13}, π_{23}.

Para el estimador de razón a tamaño con la muestra 1,2, se tiene:

$\widehat{\overline{y}}_{R(1,2)} = \dfrac{90000}{2500} = 36.$

Igualmente se calculan los resultados que aparecen en la tabla. Como es bajo m.a.s., cada uno de estos valores tiene probabilidad 1/3.

La estimación de la media por m.a.s. y estimación insesgada en la muestra 1,2 es:

$\widehat{\overline{y}}_{(1,2)} = \dfrac{1}{840 \cdot 2} (60000 + 30000) = 53.57.$

El resto de estimaciones aparece en la tabla:

	m.a.s.			ppt	
Muestra	$\widehat{\overline{y}}_R$	$\widehat{\overline{y}}$	$p(muestra)$	$\dfrac{1}{N} t_{HT}$	$p(muestra)$
$(1, 2)$	36	53.5	1/3	36.2	0.95957
$(1, 3)$	29.8	35.8	1/3	25.8	0.03686
$(2, 3)$	58	17.9	1/3	14.3	0.00355

Se sabe que los estimadores $\widehat{\overline{y}}$ y $\dfrac{1}{N} t_{HT}$ son insesgados (comprobarlo). El estimador de razón a tamaño tiene sesgo $E(\widehat{\overline{y}}_R) - 35.79$, donde

$E(\widehat{\overline{y}}_R) = \dfrac{1}{3}(36 + 29.8 + 58) = 41.26$.Así que el sesgo de $\widehat{\overline{y}}_R$ es 5.47.

La varianza de $\widehat{\overline{y}}_R$ es $V(\widehat{\overline{y}}_R) = \dfrac{1}{3}((36 - 41.26)^2 + (29.8 - 41.26)^2 + (58 - 41.26)^2) = 146.4.$

ISBN 9788496866133

La varianza de $\widehat{\overline{y}}$ es $V(\widehat{\overline{y}}) = \frac{1}{3}((53.5 - 35.79)^2 + (35.8 - 35.79)^2 + (17.9 - 35.79)^2) = 211.2$.

La varianza de $\frac{1}{N}t_{HT}$ es $V(\frac{1}{N}t_{HT}) = 0.95957(36.2-35.79)^2+0.03686(25.8-35.79)^2+0.00355(14.3-35.79)^2) = 5.47$.

El error cuadrático medio de $\widehat{\overline{y}}_R$ es $5.47^2 + 146.4 = 176.3$. El de los otros dos estimadores coincide con su varianza, por ser insesgados.

La conclusión es que el muestreo ppt es muy superior en este caso al m.a.s., aún con estimación de razón a tamaño.

Como se ha comentado, el motivo es que suele dar mejores resultados en términos de precisión asegurarse de que caen en la muestra con mayor probabilidad los conglomerados más grandes y por lo tanto más representativos de la población. El hecho de que el m.a.s. otorgue la misma probabilidad al conglomerado de 20 unidades que al de 2000 tiene en este caso una gran repercusión, aumentando mucho la varianza del estimador. Si las medias dentro de cada conglomerado fueran iguales o similares, serían aproximadamente equivalentes unos métodos a otros.

Ejercicio 8.9

En una huerta se realiza un estudio para comprobar cuántos tomates tiene cada mata en promedio. Se divide la huerta en 10 parcelas y escoge 3 de éstas por m.a.s. En ellas se cuenta el número de matas y el número de tomates en cada mata. Se obtienen los siguientes resultados:

Parcela\Mata	1	2	3	4	5
1	10	6	4	5	7
2	5	6	8		
3	10	5	6		

a) Estimar el número medio de tomates por mata y calcular la varianza del estimador.

b) Si se desea obtener un error de muestreo de 0.10, ¿cuántas parcelas se tendrá que examinar?.

---

a) Se utilizará el estimador de razón a tamaño:

$$\widehat{\overline{y}}_R = \frac{32 + 19 + 21}{5 + 3 + 3} = 6.54.$$

Para la estimación de la varianza, se estimará previamente $\overline{N}$ por $\widehat{\overline{N}} = \frac{5 + 3 + 3}{3} = 3.67$.

$$\widehat{V}(\widehat{\overline{y}}_R) = \frac{(1 - 0.3)}{3.67^2 3(3 - 1)}[5^2(6.4 - 6.54)^2 + 3^2(6.33 - 6.54)^2 + 3^2(7 - 6.54)^2] = 0.024.$$

b) Para determinar el tamaño muestral, se aproxima en la varianza exacta el término

$$\frac{1}{(L-1)} \sum_{i=1}^{L} N_i^2(\overline{y}_i - \overline{y})^2$$

por el término obtenido en la muestra

ISBN 9788496866133

$$\frac{1}{(n-1)}\sum_{i=1}^{n} N_i^2(\overline{y}_i - \overline{y})^2 = 1.385.$$

Así,

$V(\widehat{\overline{y}}_R) \simeq \dfrac{N-n}{\overline{N}^2 Nn} 1.385$ y entonces,

$$0.01 = \frac{N-n}{\overline{N}^2 Nn} 1.385.$$

Aproximando N por $\widehat{N} = L\widehat{\overline{N}} = 36.7$, se obtiene que

$$n = \frac{36.7 \cdot 1.385}{0.01(3.67^2 \cdot 36.7) + 1.385} = 8.03$$

Se necesitaría seleccionar $n = 9$ parcelas.

Ejercicio 8.10

En un polideportivo, se dispone de 10 grupos de natación supuestamente de similar nivel y con 5 alumnos cada uno. Se desea estimar el tiempo medio que tardan los nadadores de ese nivel en cubrir una cierta distancia a la mayor velocidad que puedan. Se escogen por m.a.s. dos grupos y en cada uno de ellos se hace la prueba a todos los formantes del grupo. Se obtiene que en el primer grupo la media es de 80 segundos, con una cuasivarianza de 100, y en el segundo grupo la media es 70 con una cuasivarianza de 65.

a) Estimar la varianza entre grupos y la varianza entre grupos. Decir si los grupos están bien formados o son my diferentes. Estimar el coeficiente de correlación entre conglomerados. ¿Cuál sería el valor mínimo que podría alcanzar este coeficiente? Estimar la media poblacional y estimar su varianza.

b) Suponiendo las estimaciones anteriores suficientemente precisas, si se desea estimar la media con un coeficiente de variación del estimador de 0.04, y sin tener en cuenta el coeficiente de corrección por población finita ¿cuántos grupos habría que seleccionar? ¿y si se tiene en cuenta el c.p.f.?

c) Si se desea estimar la media con un error de muestreo absoluto de 7 segundos, ¿cuántos grupos habría que seleccionar?.

---

a) La varianza entre grupos es

$$s_b^2 = \frac{\overline{N}}{(n-1)}\sum_{i=1}^{n}(\overline{y}_i - \overline{y}_c)^2 = \frac{5}{(2-1)}[(80-75)^2 + (70-75)^2] = 250.$$

La varianza intra grupos es

$$s_w^2 = \frac{1}{n(\overline{N}-1)}\sum_{i=1}^{n}\sum_{j=1}^{\overline{N}}(y_{ij} - \overline{y}_i)^2 = \frac{1}{2}(100+65) = 82.5.$$

Parece que se trata de una mala configuración de grupos, pues son muy diferentes entre sí respecto a su variabilidad interna. Como consecuencia, $\widehat{\delta}$ será posiblemente positivo.

Para calcular $\widehat{\delta}$, hay que calcular primero

ISBN 9788496866133

$$\widehat{S}^2 = \frac{L(\overline{N}-1)s_w^2 + (L-1)s_b^2}{(N-1)} = \frac{10(5-1)82.5 + (10-1)250}{(50-1)} = 113.26.$$

Así

$$\widehat{\delta} = \frac{s_b^2 - \widehat{S}^2}{(\overline{N}-1)\widehat{S}^2} = \frac{250 - 113.26}{(5-1)113.26} = 0.30.$$

El valor mínimo que podría tener el coeficiente de correlación intraconglomerados es $\delta = -\dfrac{1}{(\overline{N}-1)} = -0.25$.

La media se estima por el estimador insesgado:

$$\overline{y}_c = \frac{1}{n}\sum_{i=1}^{n}\overline{y}_i = \frac{80+70}{2} = 75.$$

La varianza estimada de $\overline{y}_c$ se puede poner como:

$$\widehat{V}(\overline{y}_c) = \frac{(1-f_1)}{\overline{N}n}s_b^2 = \frac{(1-0.2)}{5\cdot 2}250 = 20.$$

b) Al ser las estimaciones suficientemente precisas, s_b^2 es una estimación insesgada de S_b^2 y al ser el coeficiente de variación del estimador

$$\frac{\sqrt{V(\overline{y}_c)}}{E(\overline{y}_c)} = 0.04,$$

entonces

$$V(\overline{y}_c) = \frac{(1-f_1)}{\overline{N}n}S_b^2 = 0.02^2E^2(\overline{y}_c).$$

Sustituyendo la esperanza de $\overline{y}_c$ por el valor de $\overline{y}_c$ obtenido en la muestra, y S_b^2 por s_b^2, se tiene que

$$\frac{(1-f_1)}{5n}250 \simeq 0.04^2\cdot 75^2.$$

Si se prescinde del término $1-f_1$ es $\dfrac{1}{5n}250 \simeq 0.04^2\cdot 75^2$ y despejando n, se obtiene que $n = 5.5$ por lo que se tomaría $n = 6$.

Si no se prescinde del término:

$\dfrac{(10-n)}{10}\dfrac{250}{5n} \simeq 0.04^2\cdot 75^2$ y entonces:

$$n = \frac{10\cdot 250}{0.04^2\cdot 75^2\cdot 10\cdot 5 + 250} = 3.57 \text{ y se tomaría } n = 4.$$

c) Si $1.96\sqrt{V(\overline{y}_c)} = 7$, entonces $V(\overline{y}_c) = \dfrac{7^2}{1.96^2}$ y

$$\frac{(10-n)}{10}\frac{250}{5n} \simeq \frac{7^2}{1.96^2}$$

por lo cual

$$n = \frac{10\cdot 250}{\dfrac{7^2}{1.96^2}\cdot 10\cdot 5 + 250} = 2.8 \text{ y se tomaría } n = 3.$$

ISBN 9788496866133

Ejercicio 8.11

Realizar el ejercicio 8.7, apartado a) con la ayuda de la macro estimono del SAS.

Para ello se crea el archivo muestral:

```
data uno;input y conglo;
cards;
17 1
20 1
5 1
15 1
10 1
21 2
25 2
6 2
6 3
6 3
6 3
8 3
;
```

A continuación se ejecuta la macro estimono, donde se introduce la estimación $\widehat{N} = 40$ como el tamaño poblacional:

```
%estimono(uno,y,zona,10,3,40);
```

Obteniendo los datos de la estimación de razón a tamaño del total y su varianza (sale 9218 en lugar del valor 9215 obtenido en la resolución del ejercicio 8.7 debido a más precisión en los decimales).

Ejercicio 8.12

En una ciudad se desea estimar el número de cafeterías donde no se permite fumar. Se decide dividir la ciudad en 10 divisiones geográficas, y se escogen 3 por m.a.s. en cada una de éstas se cuenta el número de cafeterías y el número en el que no se permite fumar. Los datos obtenidos son los siguientes: División 1: 20 cafeterías, de las cuales en 4 no se permite fumar. División 2: 30 cafeterías, de las cuales en 8 no se permite fumar. División 3: 35 cafeterías, de las cuales en 9 no se permite fumar.

a) Estimar la proporción de cafeterías donde no se puede fumar en la ciudad, con la ayuda de la macro estimono.

b) Estimar la proporción de cafeterías donde no se puede fumar en la ciudad, con la ayuda de la macro estimrazreg.

Si se desea utilizar la macro estimono, es necesario tener en el archivo una observación por unidad elemental, de modo que se crea el archivo artificialmente, con la variable $y_{ij} = 1$ si en la cafetería no se permite fumar, e $y_{ij} = 0$ si se permite:

ISBN 9788496866133

```
data cafe;
do i=1 to 20;division=1;if i<=4 then y=1;else y=0;output;end;
do i=1 to 30;division=2;if i<=8 then y=1;else y=0;output;end;
do i=1 to 35;division=3;if i<=9 then y=1;else y=0;output;end;
run;
```

El archivo tendrá entonces 85 observaciones, y en cada conglomerado (división) respectivo habrá 20, 30 y 35, donde las primeras 4, 8 y 9 observaciones corresponden a las cafeterías donde no se puede fumar ($y_{ij} = 1$) y el resto de observaciones a aquellas en las que se puede.

Para obtener estimaciones del total al ejecutar la macro, hay que estimar previamente N (el número de cafeterías en la ciudad), pues no es conocido. Así, $\widehat{N} = L\widehat{\overline{N}} = 10 \cdot \dfrac{20+30+35}{3} = 283.33$.

A continuación, se ejecuta la macro:

```
%estimono(cafe,y,division,10,3,283.33);
```

Donde se observa el estimador de razón a tamaño, que es $\widehat{p} = 0.247$, con una varianza estimada de $\widehat{V}(\widehat{p}) = 0.00019$ y un I.C. al 95% de $(0.219, 0.274)$.

b) La manera de proceder sería crear el archivo con una observación por conglomerado y los totales en la variable y_i, y aplicar la macro estimrazreg, con N_i como variable auxiliar, y utilizando la estimación muestral de $\overline{N} = 28.33$ para el cálculo de la varianza:

```
data caf;
input tama y;
cards;
20 4
30 8
35 9
;
%estimrazreg(caf,y,tama,28.33,10,3);
```

El resultado es el mismo, tanto en el valor del estimador como en su varianza teniendo en cuenta que la proporción a estimar es la razón. Se vuelve a comprobar que los problemas de estimación en muestreo por conglomerados monoetápico se reducen en general a problemas más sencillos, considerando los conglomerados como unidades elementales.

Ejercicio 8.13

Un pasto de ganado de 2000 m^2 está dividido en 6 parcelas, de las cuales se escogen 2 con probabilidades proporcionales al tamaño. En cada uno de éstos se cuentan todas las vacas y se comprueba su edad por la placa de identificación que llevan. Los datos son:

Primera parcela seleccionada: mide 400 m^2, y hay 6 vacas, con edades en años 1, 1.5, 2, 1.7, 3, 1. Segunda parcela seleccionada: mide 500 m^2 y hay 8 vacas, con edades 2, 2, 3, 1, 1.5, 1, 3.

Utilizar la macro estimonopptr para estimar la edad media de las vacas del pasto, la varianza del estimador y un I.C. al 95%.

En primer lugar se crean los datos, calculando antes la probabilidad de selección de los conglomerados muestrales. Como se trata de probabilidades proporcionales al tamaño, $p_1 = \dfrac{400}{2000} = 0.2$ y $p_2 = \dfrac{500}{2000} = 0.25$

ISBN 9788496866133

```
data vacas;
input edad parcela ;
if parcela=1 then pi=0.2;
if parcela=2 then pi=0.25;
cards;
1   1
1.5 1
2   1
1.7 1
3   1
1   1
2   2
2   2
3   2
1   2
1.5 2
1   2
3   2
;
```

A continuación se ejecuta la macro, estimando antes $\widehat{N} = L\widehat{\overline{N}} = 42$:

```
%estimonopptr(vacas,.,parcela,edad,1,6,2,42);
```

Se obtiene que la edad media de las vacas se estima por 1.25 años, con una varianza estimada de 0.0012 y un intervalo de confianza de (1.18 , 1.32).

Ejercicio 8.14

Se realiza una encuesta para estimar el gasto en comida diaria de las familias en una comunidad. Hay 10 bloques, con diferente número de viviendas. En total hay 260 viviendas. El archivo SAS comunidad contiene los datos.

a) Calcular, con un proc means y a modo informativo, la media de gasto poblacional por familia.

b) Utilizar muestreo aleatorio simple de 3 bloques con la macro extramono y semilla 1234, y para estimar la media del gasto y obtener un I. C. utilizar la macro estimono. Realizar el mismo proceso seleccionando 6 bloques y comparar los resultados.

c) Utilizar muestreo proporcional al tamaño de los bloques (la variable nviv indica el número de viviendas), con reemplazamiento, con la macro extramonoppt y semilla 1234, con tamaño 3 y después 6, y realizar las estimaciones, a través de la macro estimonopptr.

d) Realizar el mismo proceso que en el apartado c), pero esta vez sin reemplazamiento con las macros extramonoppt y estimonoppt. Decir también cuál es el valor de la probabilidad de inclusión de segundo orden los dos conglomerados con más bajo código (variable bloque) que han entrado en la muestra, y también sus probabilidades de inclusión de primer orden.

---

a) Con el programa

```
proc means data=comunidad;var gasto;run;
```


ISBN 9788496866133

se obtiene que el gasto medio por familia en comida es 636.46 euros.

b) Se ejecutan las macros:

```
%extramono(comunidad,muestra1,codif,bloque,10,3,1234);
%estimono(muestra1,gasto,conglo,10,3,260);
```

El estimador de razón a tamaño (son tamaños desiguales), con la muestra de 3 bloques, es 632.26. Su varianza estimada es 245.46.

Si se toman 6 bloques, se obtiene una estimación de la media de 630.92 y de la varianza del estimador de 96.3.

c) La sintaxis para la macro extramonoppt es:

```
%extramonoppt(comunidad,.,muestra2,.,codif,bloque,nviv,10,1,1,3,1234);
```

Vemos con un proc print data=muestra2 que han sido seleccionados los bloques 2,3 y 4. Con la macro estimonopptr se obtienen las estimaciones:

```
%estimonopptr(muestra2,.,conglo,gasto,1,10,3,260);
```

Obteniendo un estimador de la media de 643.75 con una varianza estimada de 225.48 con la muestra de 3 bloques. Con la muestra de 6, se obtienen respectivamente, 651.71 y 80.92. Hay que señalar que en la muestra de 6 bloques aparecen repetidos dos veces en la muestra los bloques 2 y 3, y aparecen una sola vez los bloques 1 y 4.

d) Ejecutando la macro :

```
%extramonoppt(comunidad,.,muestra3,inclu,codif,bloque,nviv,10,2,1,3,1234);
```

Y presentando los archivos muestra3 e inclu con un proc print en la ventana OUTPUT, se observa que han sido elegidos los bloques 2, 10 y 4. Las probabilidades de inclusión se ven en el archivo inclu. Las de primer orden π_i de los bloques 2 y 4 son respectivamente 0.276 y 0.242. La probabilidad de inclusión de segundo orden de esos bloques es $\pi_{24} = 0.049$ (hay que ver que 2 y 4 corresponden a los valores de la variable Unit respectivos 3 y 1, y entonces buscar en la matriz de probabilidades el elemento (1,3) ó (3,1)).

Para obtener las estimaciones, se ejecuta la macro:

```
%estimonoppt(muestra3,inclu,gasto,10,3,260);
```

que da lugar, para la muestra de tamaño 3, a una media estimada de 623 con varianzas estimadas de 2138 (Horvitz–Thompson) y 546 (Yates-Grundy).

Para la muestra de tamaño 6, la media estimada es de 645 con varianzas estimadas 64 deYates-Grundy y sale negativa bajo el estimador de la varianza de Horvitz–Thompson.

Ejercicio 8.15

Supongamos que en determinada población dividida en 50 conglomerados se ha estimado con certa precisión el coeficiente de correlación intraconglomerados $\delta \simeq -0.05$, y la cuasivarianza poblacional $S^2 \simeq 20$. Si los tamaños de los conglomerados son iguales, con $\overline{N} = 10$

ISBN 9788496866133

a) Calcular cuántos conglomerados hay que seleccionar por m.a.s. para que el error de muestreo al estimar la media sea de 0.25.

b) Suponiendo que la función de coste es

$C = c_0\sqrt{n} + nc_1 + n\overline{N}c_2,$

donde $c_0 = 7$, $c_1 = 0.10$, y $c_2 = 0.05$,

presentar una tabla creada con programación con SAS para que para cada tamaño muestral de $n = 2$ hasta $n = 30$, presente el error de muestreo aproximado al que se llegaría y el coste asociado a ese tamaño muestral.

---

a) La varianza del estimador de la media es aproximadamente

$$V(\overline{y}_c) \simeq \frac{L-n}{L}\frac{S^2}{n\overline{N}}(1+(\overline{N}-1)\delta).$$

Así que se pide n tal que

$$0.25^2 \simeq \frac{50-n}{50}\frac{20}{10n}(1+(10-1)(-0.05)).$$

Despejando, se obtiene:

$$n = 50\frac{(1+(10-1)(-0.05))}{50\cdot 10\cdot 0.25^2+(1+(10-1)(-0.05))} = 13.01.$$

Se necesitarían extraer $n = 14$ conglomerados.

b) Un programa podría ser el siguiente:

```
data uno;
delta=-0.05;
k=2*(1+9*delta);
put 'n' @5 'error' @15 'coste';
do n=2 to 30;
  f=n/50;
  var=(1-f)*k/n;
  error=var**0.5;
  coste=7*n**0.5+0.1*n+0.05*10*n;
  put n @5 error 5.3 @15 coste 4.2;
  output;
end;
run;
```

Obteniendo una tabla como:

```
n   error     coste

2   0.727     11.1
3   0.587     13.9
4   0.503     16.4
..................
12  0.264     31.4
```


ISBN 9788496866133

```
13  0.250     33.0
14  0.238     34.6
..................
27  0.137     52.6
28  0.131     53.8
29  0.126     55.1
30  0.121     56.3
```

donde se aprecia que a partir de n=14 se cumple que el error es menor que 0.25 (el valor del error para $n = 13$ es mayor que 0.25, pero simplemente no aparecen todos los decimales en la tabla).

ISBN 9788496866133

9.8 Ejercicios propuestos

1) En una población dividida en conglomerados de tamaño 25 se obtiene una muestra de 10 conglomerados por m.a.s. De experiencias anteriores parece razonable asumir la relación $S_b^2 = 0.32 \cdot S^2$. Si se obtiene una varianza para el estimador de la media igual a 0.01, se pide, prescindiendo del factor $1 - f$:

a) Valor del coeficiente de correlación intraconglomerados. Interpretación.

b) Valores de S_b^2 y S^2.

c) ¿Cuántas unidades elementales sería necesario obtener utilizando m.a.s. para conseguir igual precisión que en el muestreo por conglomerados?.

2) El propietario de un periódico local quiere saber cuántas familias han leído alguna vez su periódico. En su pueblo hay 1500 viviendas, alojadas en 35 edificios. Por motivos de coste, se decide seleccionar una m.a.s. de 3 edificios y en cada uno de ellos se entrevista a todas las familias.Se obtienen los siguientes datos: en el primer edificio había 36 familias, de las cuales 12 habían leído alguna vez el periódico. En el segundo edificio, había 45 familias, de las cuales 28 habían leído alguna vez el periódico, y en el tercer edificio muestreado había 26 familias de las cuales 13 habían accedido alguna vez al periódico.

a) Estimar la proporción de familias que han leído alguna vez el periódico en el pueblo, y dar un I.C. al 95% para la cantidad estimada.

b) Suponiendo suficientemente precisa la estimación anterior, dar el error de muestreo obtenido si se muestrearan 5, 10 y 20 edificios respectivamente.

3)

En un museo se dispone de 100 lotes de piezas sin clasificar. Se sabe que en total hay 1600 piezas. Se desea tener una aproximación al periodo del que datan en promedio las piezas. Se escogen 4 lotes por m.a.s. y se encuentra que en el primer lote hay 15 piezas. Estas se fechan con ayuda de expertos, que datan las piezas en promedio del año 1210, con una cuasi desviación típica de 108. En este y el resto de lotes se obtiene lo que aparece en la tabla.

Lote	Media	Cuasi d.típica	nº piezas
1	1210	108	15
2	1310	110	20
3	1260	120	30
4	1340	130	10

a) Estimar la varianza entre lotes y la varianza entre lotes. Decir si los grupos están bien formados o son muy diferentes. Estimar el coeficiente de correlación entre conglomerados. ¿Cuál sería el valor mínimo que podría alcanzar este coeficiente? Estimar la media poblacional del año de las piezas y dar un intervalo de confianza al 95% .

ISBN 9788496866133

b) Suponiendo la estimación anterior suficientemente precisa, decir en cuánto se aumentaría la precisión si se tomaran 20 lotes en lugar de 4.

4) Se realiza un control de calidad sobre las piezas de un pedido de 800 lotes con 30 piezas cada uno. Se extrae una muestra sin reposición de 25 lotes, dentro de la cual 12 lotes no tienen piezas defectuosas, 10 lotes tienen una pieza defectuosa y tres lotes tienen 2 piezas defectuosas.

a) Estimar el número total de piezas defectuosas en el pedido, el error de muestreo y el error de muestreo relativo, así como el error de muestreo absoluto con un grado de confianza de 95%.

b) Resolver el apartado a) suponiendo que el muestreo es con reposición.

5) En una ganadería las vacas están distribuidas en grandes dehesas separadas por alambradas. Se realiza un estudio para comprobar qué proporción de vacas hembra hay en promedio. Hay 15 dehesas y escoge 3 de éstas por m.a.s. En ellas se cuenta el número de vacas y toros. Se obtienen los siguientes resultados:

Dehesa\Vaca	Vacas	Toros
1	30	8
2	10	12
3	12	12

a) Plantear el problema como muestreo monoetápico de conglomerados. Estimar la proporción de vacas y estimar la varianza del estimador, por estimación insesgada y por razón a tamaño.

b) Resolver el problema planteándolo como un problema de m.a.s.+estimación de la razón. Estimar la varianza del estimador y comparar.

6) Se trata de estimar la media en muestreo monoetápico con $n = 2$ en una población que contiene 3 conglomerados, con los tamaños y las medias poblacionales indicadas en la tabla inferior:

N_i	$\overline{y}_i$
3000	25
1000	60
50	10

a) Calcular la correlación entre los N_i y las medias $\overline{y}_i$. Calcular la correlación entre los N_i y los totales y_i.

b) A la vista de lo obtenido, para estimar la media, ¿es conveniente utilizar m.a.s. con estimación insesgada, m.a.s. con estimación de razón a tamaño o bien muestreo ppt con probabilidades proporcionales al tamaño?

ISBN 9788496866133

c) Calcular el valor de los tres estimadores de la media indicados en el apartado anterior para todas las muestras, con sus probabilidades respectivas. Resumir en una tabla, y calcular a continuación la varianza y sesgo de cada uno de los tres estimadores. Concluir.

7) Supongamos que el número total de personas mayores de 65 años y el número de personas mayores de 65 años que requieren los servicios de un asistente social en 2 urbanizaciones muestreadas de un conjunto de 5 urbanizaciones, son los expuestos en la siguiente tabla:

ISBN 9788496866133

	Urbanización A		Urbanización B	
	>65	>65,asist	>65	>65,asist
1	2	1	1	1
2	1	0	1	0
3	2	0	1	1
4	1	1	3	1
5	1	0	2	0
6	1	0	1	1
7	2	1	3	0
8	1	1	1	0
9	3	1	1	0
10	1	1	3	2
11	1	0	2	1
12	2	1	3	0
13	1	0	1	1
14	3	1	1	0
15	1	0	2	1
16	3	2	2	1
17	1	0	1	0
18	2	0	2	0
19	1	1	2	1
20	3	0	1	1

Estimar las características poblacionales siguientes, facilitando intervalos de confianza al 95%:

a) Número medio de personas por urbanización mayores de 65 años que requieren los servicios

ISBN 9788496866133

de un asistente social-

b) Número de personas mayores de 65 años que requieren los servicios de un asistente social.

c) Número medio de personas por vivienda mayores de 65 años que requieren los servicios de un asistente social.

d) Número medio de personas por urbanización mayores de 65 años.

e) Proporción de personas mayor de 65 años que requieren los servicios de un asistente social.

8) Realizar el ejercicio propuesto 7) en SAS, con la ayuda de la macro estimono.

9) Realizar el ejercicio resuelto 8.13 en SAS, con la ayuda de la macro estimono.

10) El archivo SAS instituto contiene las notas medias de los alumnos de diferentes clases de un instituto.

a) Calcular, con un proc means y a modo informativo, la nota media de los alumnos del instituto.

b) Utilizar muestreo aleatorio simple de 4 aulas con la macro extramono y semilla 1234, y para estimar la nota media y obtener un I. C. utilizar la macro estimono. Realizar el mismo proceso seleccionando 7 aulas y comparar los resultados.

c) Utilizar muestreo proporcional al tamaño de las aulas con reemplazamiento,(el tamaño está indicado con la variable numero) , con la macro extramonoppt y semilla 1234, con tamaño 4 y después 7, y realizar las estimaciones con la macro estimonopptr.

d) Realizar el mismo proceso que en el apartado c), pero esta vez sin reemplazamiento con las macros extramonoppt y estimonoppt. Decir también cuál es el valor de la probabilidad de inclusión de segundo orden los dos conglomerados con más bajo código (variable bloque) que han entrado en la muestra, y también sus probabilidades de inclusión de primer orden.

e) Para comprobar la variabilidad entre conglomerados, ejecutar el siguiente programa con la última muestra obtenida:

```
proc boxplot data=muestra;plot nota*aula;run;
```

11) El archivo SAS supermat contiene un estudio realizado sobre 4810 alumnos en 91 Centros de Enseñanaza Secundaria de Madrid, sobre aptitudes de todo tipo. La variable centro indica el código de centro, y la variable tamacentro el número de alumnos encuestados en ese centro.

Para el resto del ejercicio, consideraremos el archivo como la población de interés. Los tests realizados están en escala de 1 a 100.

a) Calcular la nota poblacional media de la variable orto (test de ortografía).

b) Extraer 5 centros por muestreo proporcional a la variable tamacentro, con reemplazamiento, con la macro extramonoppt, con la semilla 1234. Calcular un I.C. al 95% para la nota media en el test ortografía con la macro estimonopptr.

c) Realizar el apartado anterior pero en muestreo sin reemplazamiento, con las macros extramonoppt y estimonoppt.

ISBN 9788496866133

ISBN 9788496866133

10 MUESTREO BIETÁPICO DE CONGLOMERADOS

Un conglomerado frecuentemente contiene demasiados elementos para obtener una medición de cada uno de ellos, o estos son tan homogéneos que la medición de sólo unos cuantos proporciona información completa sobre el conglomerado. Cuando ocurre cualquiera de las dos situaciones, el investigador puede seleccionar una muestra aleatoria de conglomerados y dentro de los seleccionados, escoger a su vez una muestra aleatoria de unidades elementales. El resultado es una muestra por conglomerados en dos etapas.

Por lo tanto, en el muestreo en dos etapas o bietápico se seleccionan primero n conglomerados o unidades primarias. Después se selecciona un número específico de subunidades o unidades secundarias o finales en cada uno de los conglomerados extraídos. Al igual que en el capítulo anterior, se considerará en primer lugar el caso más simple, en el que cada unidad primaria o conglomerado contiene el mismo número $\overline{N}$ de unidades secundarias, de las cuales se seleccionarán m en cada uno de los conglomerados escogidos.

La principal ventaja del muestreo en dos etapas es su flexibilidad respecto al muestreo en una etapa. Como de costumbre, se trata de alcanzar un cierto equilibrio entre precisión estadística y coste. Pueden muestrearse pocos conglomerados y muchos elementos en cada uno, o muchos conglomerados y pocos elementos en cada conglomerado. Los conglomerados grandes tienden a contener elementos heterogéneos, y en consecuencia se requiere una muestra grande de cada uno para lograr estimaciones precisas de los parámetros de la población. En contraste, los conglomerados pequeños frecuentemente contienen elementos relativamente homogéneos, en cuyo caso puede obtenerse información precisa sobre las características de un conglomerado seleccionando una muestra pequeña de cada uno.

Otra ventaja es que el coste respecto al muestreo monoetápico es menor, al no tomarse todas las unidades dentro de los conglomerados escogidos. Además, el problema existente en muestreo monoetápico con tamaños desiguales, relativo a que el tamaño muestral final es una variable aleatoria, no existe aquí.

ISBN 9788496866133

Entre las desventajas de este método de muestreo cabe citar que, al extraer una muestra dentro de cada conglomerado, aparece una nueva fuente de variabilidad. La precisión de los estimadores se verá afectada, pues las varianzas serán similares a las obtenidas en muestreo monoetápico, pero con un factor añadido debido a la variabilidad de la submuestra dentro de cada conglomerado. Además, los desarrollos teóricos son más complejos, al tener en cuenta esa segunda fuente de variabilidad.

10.1 Marco probabilístico

Cuando se seleccionan aleatoriamente conglomerados y dentro de cada uno de ellos unidades elementales, existen dos fuentes de variabilidad implicadas en la construcción de un estimador:

1. La distribución que asigna probabilidades de aparición, en la muestra de tamaño n, a los L conglomerados.
2. Las distribuciones de probabilidad que asignan probabilidades de aparición, en la muestra de tamaño m_i, a las unidades elementales que están dentro del conglomerado i. Estas distribuciones están condicionadas a que haya salido escogido el conglomerado i en la muestra inicial de conglomerados.

Por lo tanto, a la hora de calcular esperanzas y varianzas de los estimadores, funciones matemáticas de la muestra, habrá que tener en cuenta este esquema probabilístico. Los siguientes resultados permiten acometer la tarea de cálculo de espe-ranzas y varianzas de estimadores en el resto de este capítulo.

Teorema 10.1 (esperanza y varianza en muestreo bietápico-Teorema de Madow).

Sea $\widehat{\theta}$ el valor del estimador obtenido en muestreo bietápico. $\widehat{\theta}$ es una función de las unidades elementales obtenidas.

Sea E_1 **la esperanza en 1º etapa,** es decir, el promedio calculado sobre todas las posibles selecciones de conglomerados i en primera etapa. E_1 está relacionada con la distribución mencionada en el punto 1. Sea V_1 la **varianza en primera etapa**.

Sea E_2^i **la esperanza en 2º etapa** condicionada al conglomerado i. Es decir, el promedio del valor del estimador calculado sobre todas las posibles muestras en segunda etapa en el conglomerado i. E_2^i se refiere a la distribución mencionada en el punto 2 para un determinado conglomerado i.Sea V_2^i la **varianza en segunda etapa**.

Entonces,

(a) $E(\widehat{\theta}) = E_1[E_2^i(\widehat{\theta})]$

(b) $V(\widehat{\theta}) = V_1[E_2^i(\widehat{\theta})] + E_1[V_2^i(\widehat{\theta})]$

El apartado (b) se conoce como el Teorema de Madow.

ISBN 9788496866133

Demostración.

(a) Es una consecuencia del conocido resultado en teoría estadística

$E(X) = E_Y[E_X(X \mid Y)]$.

$$E_1[E_2^i(\widehat{\theta})] = \sum_{i=1}^{L} p_i E_2^i(\widehat{\theta}) = \sum_{i=1}^{L} p_i \sum_{s\in\Omega_i}^{\Omega_i} p_{s/i}\widehat{\theta}_s,$$

donde $\widehat{\theta}_s$ es el valor del estimador obtenido habiendo escogido la muestra s en el conglomerado i, y $p_{s/i}$ es la probabilidad condicionada de obtener la muestra s en el conglomerado i.

Así, como $p_i p_{s/i} = p_{i,s}$ es la probabilidad conjunta p(conglomerado i, muestra s), se tiene que

$$\sum_{i=1}^{L} p_i \sum_{s\in\Omega_i}^{\Omega_i} p_{s/i}\widehat{\theta}_s = \sum_{i=1}^{L} \sum_{s\in\Omega_i}^{\Omega_i} p_i p_{s/i}\widehat{\theta}_s = \sum_{i=1}^{L} \sum_{s\in\Omega_i}^{\Omega_i} p_{i,s}\widehat{\theta}_s = \sum_{i=1}^{L} \widehat{\theta}_s \sum_{s\in\Omega_i}^{\Omega_i} p_{i,s} = \sum_{i=1}^{L} \widehat{\theta}_s p_i = E(\widehat{\theta}).$$

(b) $V(\widehat{\theta}) = E(\widehat{\theta} - E(\widehat{\theta}))^2 = E_1[E_2^i((\widehat{\theta} - E(\widehat{\theta}))^2)]$.

Hay que tener en cuenta que $E(\widehat{\theta})$ es una constante. El término interior se puede desarrollar:

$$E_2^i((\widehat{\theta} - E(\widehat{\theta}))^2) = E_2^i(\widehat{\theta}^2) - 2E(\widehat{\theta})E_2^i(\widehat{\theta}) + E^2(\widehat{\theta}) = (E_2^i(\widehat{\theta}))^2 + V_2(\widehat{\theta}) - 2E(\widehat{\theta})E_2^i(\widehat{\theta}) + E^2(\widehat{\theta}).$$

Así,

$$V(\widehat{\theta}) = E_1[(E_2^i(\widehat{\theta}))^2] + E_1[V_2(\widehat{\theta})] - 2E(\widehat{\theta})E_1[E_2^i(\widehat{\theta})] + E^2(\widehat{\theta}) =$$

$$= E_1[(E_2^i(\widehat{\theta}))^2] + E_1[V_2(\widehat{\theta})] - 2E^2(\widehat{\theta})^2 + E^2(\widehat{\theta}) = E_1[(E_2^i(\widehat{\theta}))^2] + E_1[V_2(\widehat{\theta})] - E^2(\widehat{\theta}) =$$

$$E_1[(E_2^i(\widehat{\theta}))^2] + E_1[V_2(\widehat{\theta})] - (E_1[E_2^i(\widehat{\theta})])^2 = V_1[E_2^i(\widehat{\theta})] + E_1[V_2(\widehat{\theta})].$$

Pues se ha utilizado que $V_1[E_2^i(\widehat{\theta})] = E_1[(E_2^i(\widehat{\theta}))^2] - (E_1[E_2^i(\widehat{\theta})])^2$.

De modo que para calcular esperanzas de estimadores y estudiar si son insesgados, se utilizará el resultado (a), y para calcular varianzas, el apartado (b).

10.2 Conglomerados de igual tamaño

Supongamos que todos los conglomerados o unidades de primera etapa tienen el mismo tamaño $\overline{N}$, que se seleccionan n de entre los L que existen, y que de cada uno de esos conglomerados se eligen m unidades elementales de las $\overline{N}$ que existen. El tamaño poblacional es $N = L\overline{N}$. Se denotan por $f_1 = \dfrac{n}{L}$ y $f_1 = \dfrac{m}{\overline{N}}$ las fracciones de muestreo en primera etapa y segunda etapa, respectivamente.

ISBN 9788496866133

10.2.1 Notación

Características poblacionales

y_{ij} = valor de la j-ésima subunidad en la i-ésima unidad primaria.

$\overline{y} = \dfrac{1}{L\overline{\overline{N}}} \sum_{i=1}^{L} \sum_{j=1}^{\overline{N}} y_{ij} = \dfrac{1}{L} \sum_{i=1}^{L} \overline{y}_i$ = Media poblacional.

$\overline{y}_i = \dfrac{1}{\overline{N}} \sum_{j=1}^{\overline{N}} y_{ij}$ =Media en el conglomerado i.

y_i = Total en el conglomerado i.

$S_{2i}^2 = \dfrac{1}{\overline{N}-1} \sum_{j=1}^{\overline{N}} (y_{ij} - \overline{y}_i)^2$ = Cuasivarianza en el conglomerado i. Se denota $\sigma_{2i}^2 = \dfrac{\overline{N}-1}{\overline{N}} S_{2i}^2$.

$S_2^2 = \dfrac{1}{L(\overline{N}-1)} \sum_{i=1}^{L} \sum_{j=1}^{\overline{N}} (y_{ij} - \overline{y}_i)^2 = \dfrac{1}{L} \sum_{i=1}^{L} S_{2i}^2$ =Cuasivarianza en segunda etapa , o promedio de las cuasivarianzas en los conglomerados.

Se denota $\sigma_2^2 = \dfrac{1}{L} \sum_{i=1}^{L} \sigma_{2i}^2$.

(Nota: $S_2^2 = S_w^2$, siendo S_w^2 la cuasivarianza intra-conglomerados presentada en muestreo monoetápico)

$S_1^2 = \dfrac{1}{L-1} \sum_{i=1}^{L} (\overline{y}_i - \overline{y})^2$ = Cuasivarianza en primera etapa, o cuasivarianza entre medias de unidades primarias.

(Nota: $S_1^2 = \dfrac{1}{\overline{\overline{N}}} S_b^2$, siendo S_b^2 la cuasivarianza entre conglomerados presentada en muestreo monoetápico)

Características muestrales

$\widehat{\overline{y}}_i = \dfrac{1}{m} \sum_{i=1}^{m} y_{ij}$ = Media muestral en el conglomerado i.

$\overline{\overline{y}} = \dfrac{1}{n} \sum_{i=1}^{n} \widehat{\overline{y}}_i$ =Media muestral de las medias obtenidas en los conglomerados seleccionados.

$s_{2i}^2 = \dfrac{1}{m-1} \sum_{j=1}^{m} (y_{ij} - \widehat{\overline{y}}_i)^2$ = Cuasivarianza muestral en el conglomerado i.

$s_2^2 = \dfrac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \widehat{\overline{y}}_i)^2 = \dfrac{1}{n} \sum_{i=1}^{n} s_{2i}^2$ =Cuasivarianza muestral en segunda etapa , o promedio de las cuasivarianzas en los conglomerados.

ISBN 9788496866133

$s_1^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(\widehat{\overline{y}}_i - \overline{\overline{y}})^2$ = Cuasivarianza muestral en primera etapa, o cuasivarianza entre medias de unidades primarias.

10.2.2 Estimador insesgado con m.a.s. en ambas etapas

Se verá a continuación un estimador insesgado de la media poblacional, suponiendo muestreo aleatorio simple en ambas etapas.

Teorema 10.2 (estimador insesgado de la media y varianza).

(a) $\overline{\overline{y}} = \dfrac{1}{n}\sum_{i=1}^{n}\widehat{\overline{y}}_i$ es un estimador insesgado de la media poblacional.

(b) La varianza de $\overline{\overline{y}}$ es

$$V(\overline{\overline{y}}) = \frac{L-n}{L}\frac{S_1^2}{n} + \frac{\overline{N}-m}{\overline{N}}\frac{S_2^2}{nm}$$

o, en función de las fracciones de muestreo,

$$V(\overline{\overline{y}}) = (1-f_1)\frac{S_1^2}{n} + (1-f_2)\frac{S_2^2}{nm}.$$

Demostración.

(a) Se utilizará que en m.a.s. , la esperanza de la media muestral es la media poblacional.

$$E(\overline{\overline{y}}) = E_1[E_2^i(\overline{\overline{y}})] = E_1[E_2^i(\frac{1}{n}\sum_{i=1}^{n}\widehat{\overline{y}}_i)] = E_1\left[\frac{1}{n}\sum_{i=1}^{n}E_2^i(\widehat{\overline{y}}_i)\right] = E_1\left[\frac{1}{n}\sum_{i=1}^{n}\overline{y}_i\right] =$$

$$= \frac{1}{L}\sum_{i=1}^{L}\overline{y}_i = \overline{y}.$$

(b) Se utilizará el Teorema de Madow:

$$V(\widehat{\theta}) = V_1[E_2^i(\widehat{\theta})] + E_1[V_2^i(\widehat{\theta})]$$

donde en este caso $\widehat{\theta} = \overline{\overline{y}} = \dfrac{1}{n}\sum_{i=1}^{n}\widehat{\overline{y}}_i$.

En primer lugar,

$$V_1[E_2^i(\widehat{\theta})] = V_1\left[E_2^i\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\overline{y}}_i\right)\right] = V_1\left[\frac{1}{n}\sum_{i=1}^{n}\overline{y}_i\right].$$

Como se trata de muestreo aleatorio simple, y se seleccionan n unidades de entre L, considerando $\overline{y}_i$ la característica de interés, la varianza de la media muestral $\dfrac{1}{n}\sum_{i=1}^{n}\overline{y}_i$ es $V_1\left[\dfrac{1}{n}\sum_{i=1}^{n}\overline{y}_i\right] = \dfrac{L-n}{L}\dfrac{S_{\overline{y}_i}^2}{n}$. La cuasivarianza poblacional de $\overline{y}_i$ denotada por $S_{\overline{y}_i}^2$ es la cuasivarianza en primera etapa, $S_{\overline{y}_i}^2 = S_1^2$, y entonces $V_1[E_2^i(\widehat{\theta})] = \dfrac{L-n}{L}\dfrac{S_1^2}{n}$.

ISBN 9788496866133

A continuación se desarrolla el segundo sumando en el Teorema de Madow:

$$E_1[V_2^i(\widehat{\theta})] = E_1\left[V_2^i\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\overline{y}}_i\right)\right] \overset{(*)}{=} E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}V_2^i(\widehat{\overline{y}}_i)\right] \overset{(**)}{=} E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\frac{\overline{N}-m}{\overline{N}}\frac{S_{2i}^2}{m}\right] =$$

$$= \frac{1}{n}\frac{\overline{N}-m}{\overline{N}}\frac{1}{m}E_1\left(\frac{1}{n}\sum_{i=1}^{n}S_{2i}^2\right) \overset{(***)}{=} \frac{\overline{N}-m}{\overline{N}}\frac{1}{nm}\frac{1}{L}\sum_{i=1}^{L}S_{2i}^2 = \frac{\overline{N}-m}{\overline{N}}\frac{S_2^2}{nm}.$$

La igualdad $(*)$, donde las covarianzas no aparecen, proviene del hecho de la independencia del muestreo dentro de cada conglomerado, de unos conglomerados a otros.

La igualdad $(**)$ se da por tratarse de la varianza de la media muestral en muestreo aleatorio simple en cada conglomerado, es decir, muestreo de m unidades sobre una población de $\overline{N}$ unidades.

La igualdad $(***)$ se da por ser $E_1(\frac{1}{n}\sum_{i=1}^{n}S_{2i}^2)$ la esperanza de la media muestral de la característica S_{2i}^2, que en muestreo aleatorio simple es la media poblacional $\frac{1}{L}\sum_{i=1}^{L}S_{2i}^2$.

Así,

$$V_1[E_2^i(\widehat{\theta})] + E_1[V_2^i(\widehat{\theta})] = \frac{L-n}{L}\frac{S_1^2}{n} + \frac{\overline{N}-m}{\overline{N}}\frac{S_2^2}{nm}.$$

La expresión de la varianza del estimador es muy ilustrativa de lo que ocurre en muestreo bietápico: La varianza es la suma de dos fuentes de variabilidad: entre conglomerados $(\frac{L-n}{L}\frac{S_1^2}{n})$ e intra-conglomerados $(\frac{\overline{N}-m}{\overline{N}}\frac{S_2^2}{nm})$. Si la configuración de conglomerados es correcta, los conglomerados son parecidos entre sí y S_1^2 es pequeña, y por lo tanto n no importa que sea pequeña. Además, los grupos son heterogéneos internamente, y entonces S_2^2 es grande, con lo cual m debe ser relativamente grande para compensar.

Pero normalmente la configuración puede no ser óptima, y estudios piloto de variabilidad , añadidos posiblemente a una función de coste , deben ser utilizados para determinar cómo distribuir la muestra de tamaño aproximado $n\prime = (nm)$, en dos etapas. Es decir, estudiar si se tomarán muchos conglomerados y pocas unidades dentro de cada uno de ellos, o por el contrario pocos conglomerados y una muestra de tamaño alto m dentro de cada conglomerado.

Hay que remarcar además que la varianza del estimador de la media en muestreo monoetápico con tamaños iguales era

$$V(\overline{y}_c) = \frac{L-n}{L}\frac{1}{n(L-1)}\sum_{i=1}^{L}(\overline{y}_i-\overline{y})^2 = \frac{L-n}{L}\frac{S_1^2}{n},$$

con lo cual se observa que la varianza obtenida en muestreo bietápico es la misma que en muestreo monoetápico, más el factor añadido $\frac{\overline{N}-m}{\overline{N}}\frac{S_2^2}{nm}$, debido a tomar muestras, y no toda la población, dentro de cada conglomerado escogido. La coincidencia del primer término de las varianzas de muestreo bietápico con la varianza completa obtenida en muestreo monoetápico se

ISBN 9788496866133

dará, en general, en la mayoría de los diferentes marcos de muestreo que se verán (y usualmente también se da en los estimadores de las varianzas).

El siguiente resultado permite estimar la varianza del estimador de manera insesgada.

Teorema 10.3 (estimador de la varianza). Un estimador insesgado de $V(\overline{\overline{y}})$ es

$$\widehat{V}(\overline{\overline{y}}) = \frac{L-n}{L}\frac{s_1^2}{n} + \frac{\overline{N}-m}{\overline{N}}\frac{s_2^2}{mL} = (1-f_1)\frac{s_1^2}{n} + (1-f_2)\frac{s_2^2}{mL}.$$

Demostración.

$$E(\widehat{V}(\overline{\overline{y}})) = \frac{L-n}{Ln}E(s_1^2) + \frac{\overline{N}-m}{\overline{N}mL}E(s_2^2).$$

Calcularemos por separado $E(s_1^2)$ y $E(s_2^2)$:

En primer lugar, se sabe que $s_1^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(\widehat{\overline{y}}_i - \overline{\overline{y}})^2 = \dfrac{1}{n-1}\left[\sum_{i=1}^{n}\widehat{\overline{y}}_i^2 - n\overline{\overline{y}}^2\right]$, con lo que

$$E(s_1^2) = E_1\left[E_2^i\left[s_1^2\right]\right] = E_1\left[\frac{1}{n-1}E_2^i\left[\sum_{i=1}^{n}\widehat{\overline{y}}_i^2 - n\overline{\overline{y}}^2\right]\right].$$

Ahora,

$$E_2^i\left[\sum_{i=1}^{n}\widehat{\overline{y}}_i^2 - n\overline{\overline{y}}^2\right] = \sum_{i=1}^{n}E_2^i(\widehat{\overline{y}}_i^2) - nE_2^i(\overline{\overline{y}}^2) = \sum_{i=1}^{n}E_2^i(\widehat{\overline{y}}_i^2) - n\left[(E_2^i(\overline{\overline{y}}))^2 + V_2^i(\overline{\overline{y}})\right] =$$

$$= \sum_{i=1}^{n}((E_2^i(\widehat{\overline{y}}_i))^2 + \sum_{i=1}^{n}V_2^i(\widehat{\overline{y}}_i) - n\left[(E_2^i(\overline{\overline{y}}))^2 + V_2^i(\overline{\overline{y}})\right] =$$

$$= \sum_{i=1}^{n}\overline{y}_i^2 + \sum_{i=1}^{n}(1-f_2)\frac{S_{2i}^2}{nm} - n(\frac{1}{n}\sum_{i=1}^{n}\overline{y}_i)^2 + n\frac{1}{n^2}\sum_{i=1}^{n}(1-f_2)\frac{S_{2i}^2}{nm} \overset{(*)}{=}$$

$$= \sum_{i=1}^{n}(\overline{y}_i - \frac{1}{n}\sum_{i=1}^{n}\overline{y}_i)^2 + (n-1)(1-f_2)\sum_{i=1}^{n}\frac{S_{2i}^2}{nm}.$$

Se ha utilizado en $(*)$ que

$$\sum_{i=1}^{n}(\overline{y}_i - \frac{1}{n}\sum_{i=1}^{n}\overline{y}_i)^2 = \sum_{i=1}^{n}\overline{y}_i^2 - n(\frac{1}{n}\sum_{i=1}^{n}\overline{y}_i)^2$$

debido a la conocida propiedad asociada a la varianza muestral, $\sum_{i=1}^{n}(x_i-\overline{x})^2 = \sum_{i=1}^{n}x_i^2 - n\overline{x}^2$.

Así,

$$E_2^i(s_1^2) = \frac{1}{n-1}E_2^i\left[\sum_{i=1}^{n}\widehat{\overline{y}}_i^2 - n\overline{\overline{y}}^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}(\overline{y}_i - \frac{1}{n}\sum_{i=1}^{n}\overline{y}_i)^2 + (1-f_2)\sum_{i=1}^{n}\frac{S_{2i}^2}{nm}.$$

Tomando la esperanza en primera etapa, como el primer término es una cuasivarianza muestral y su esperanza bajo m.a.s. es la cuasivarianza poblacional, se tiene que

ISBN 9788496866133

$$E(s_1^2) = E_1(E_2^i(s_1^2)) = \frac{1}{L-1}\sum_{i=1}^{L}(\overline{y}_i - \overline{y})^2 + (1-f_2)\frac{1}{m}E.\left[\frac{1}{n}\sum_{i=1}^{n} S_{2i}^2\right] = S_1^2 + (1-f_2)\frac{1}{m}S_2^2.$$

En segundo lugar, calcularemos $E(s_2^2)$:

$$E(s_2^2) = E_1(E_2^i(s_2^2)) = E_1\left[E_2^i(\frac{1}{n}\sum_{i=1}^{n} s_{2i}^2)\right] = E_1\left[\frac{1}{n}\sum_{i=1}^{n} E_2^i(s_{2i}^2)\right] = E_1\left[\frac{1}{n}\sum_{i=1}^{n} S_{2i}^2\right] =$$

$$= \frac{1}{L}\sum_{i=1}^{L} S_{2i}^2 = S_2^2.$$

Así,

$$E(\widehat{V}(\overline{\overline{y}})) = \frac{L-n}{Ln}E(s_1^2) + \frac{\overline{N}-m}{\overline{N}mL}E(s_2^2) = \frac{(1-f_1)}{n}\left[S_1^2 + (1-f_2)\frac{1}{m}S_2^2\right] + \frac{(1-f_2)}{mL}S_2^2 =$$

$$= \frac{(1-f_1)}{n}S_1^2 + (1-f_1)(1-f_2)\frac{1}{nm}S_2^2 + \frac{(1-f_2)}{mL}S_2^2 =$$

$$= \frac{(1-f_1)}{n}S_1^2 + (1-f_2)\frac{1}{m}S_2^2\left[\frac{(1-f_1)}{n} + \frac{1}{L}\right].$$

Como

$$\frac{(1-f_1)}{n} + \frac{1}{L} = \frac{L-n}{Ln} + \frac{1}{L} = \frac{1}{n},$$

se tiene finalmente que

$$E(\widehat{V}(\overline{\overline{y}})) = (1-f_1)\frac{S_1^2}{n} + (1-f_2)\frac{S_2^2}{nm} = V(\overline{\overline{y}}).$$

Ejemplo 10.1

En un proceso de control de calidad industrial, se dispone de componentes electróncios agrupados en lotes. Estos componentes se evalúan obteniendo un valor que indica un cierto tipo de rendimiento. Cada lote consta de 20 componentes, y hay 140 lotes de producción. La idea es estimar el rendimiento medio de los componentes en toda la producción.

Se realiza muestreo por conglomerados bietápico, seleccionando 6 lotes, y dentro de cada uno 5 componentes. Los resultados se indican en la tabla 10.1.

En este ejemplo, se tiene $L = 140$, $\overline{N} = 20$, $N = 2800$, $n = 6$ y $m = 5$. El estimador insesgado de la media poblacional es

$$\overline{\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\overline{y}}_i = \frac{1}{6}(12 + 11.6 + 10.4 + 9.8 + 11.8 + 12.4) = 11.33.$$

ISBN 9788496866133

Lote	Rendimientos en los componentes seleccionados
1	12,10,15,6,17
2	9,8,16,13,12
3	7,5,18,12,10
4	6,10,10,15,8
5	13,14,9,15,8
6	12,10,6,18,16

Tabla 10.1. Rendimientos de componentes

Para estimar la varianza, se calculan

$$s_2^2 = \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \widehat{\overline{y}}_i)^2 = \frac{1}{6} \sum_{i=1}^{n} \frac{1}{5-1} \sum_{j=1}^{m} (y_{ij} - \widehat{\overline{y}}_i)^2 = \frac{1}{6}(18.5+10.3+25.3+11.2+9.7+22.8) = 16.3.$$

y

$$s_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\widehat{\overline{y}}_i - \overline{\overline{y}})^2 = 1.018.$$

Entonces

$$\widehat{V}(\overline{\overline{y}}) = \frac{L-n}{L} \frac{s_1^2}{n} + \frac{\overline{N}-m}{\overline{N}} \frac{s_2^2}{mL} = \frac{140-6}{140} \frac{1.018}{6} + \frac{20-5}{20} \frac{16.3}{5 \cdot 140} = 0.179.$$

Para observar la variabilidad entre conglomerados y entre conglomerados se puede recurrir a un diagrama de cajas en la Figura 10.1.

Un intervalo de confianza al 95% , suponiendo normalidad del estimador, quedaría en

$$(11.33 \pm 1.96 \cdot 0.179) = (10.97, 11.68).$$

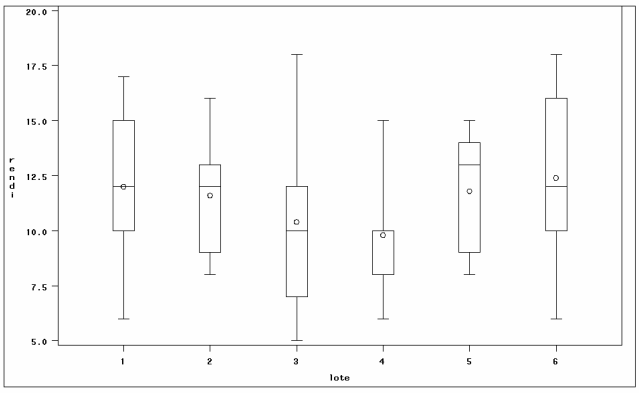


Figura 10.1. Diagrama de cajas del rendimiento por lote.

ISBN 9788496866133

10.2.3 Estimador del total y la proporción

Para estimar el total y la proporción, se aplica el resultado anterior.

Corolario 10.1 (estimador del total y proporción).

En caso de muestreo bietápico con m.a.s. en ambas etapas,

(a) Un estimador insesgado del total poblacional es $N\overline{y}$, con varianza $N^2V(\overline{y})$ y estimador insesgado de la varianza $N^2\widehat{V}(\overline{y})$.

(b) Supongamos que la variable y toma valores 0 ó 1 y se desea estimar la proporción p de valores 1 en la población. Un estimador insesgado de esta proporción p es $\widehat{\widehat{p}} = \dfrac{1}{n}\sum_{i=1}^{n}\widehat{p}_i$ donde $\widehat{p}_i$ es la proporción muestral en el conglomerado i.

(c) La varianza de $\widehat{\widehat{p}}$ es

$$V(\widehat{p}) = \frac{L-n}{L}\frac{S_1^2}{n} + \frac{\overline{N}-m}{\overline{N}}\frac{S_2^2}{nm}$$

donde, en este caso,

$$S_1^2 = \frac{1}{L-1}\sum_{i=1}^{L}(p_i-p)^2 \text{ y } S_2^2 = \frac{1}{L(\overline{N}-1)}\sum_{i=1}^{L}\sum_{j=1}^{\overline{N}}(y_{ij}-p_i)^2 = \frac{1}{L}\sum_{i=1}^{L}\frac{\overline{N}}{\overline{N}-1}p_i(1-p_i).$$

(d) Un estimador insesgado de $V(\widehat{p})$ es

$$\widehat{V}(\widehat{p}) = \frac{L-n}{L}\frac{s_1^2}{n} + \frac{\overline{N}-m}{\overline{N}}\frac{s_2^2}{mL}$$

donde, en este caso,

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{p}_i - \widehat{\widehat{p}})^2,$$

$$s_{2i}^2 = \frac{m}{m-1}\widehat{p}_i(1-\widehat{p}_i),$$

y

$$s_2^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{m}{m-1}\widehat{p}_i(1-\widehat{p}_i).$$

En (c) y (d) se ha aplicado el resultado ya conocido de que

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{y})^2 = \frac{N}{N-1}p(1-p)$$

cuando la variable y es una variable 0,1 y p es la proporción de valores "1".

ISBN 9788496866133

10.2.4 Tamaño muestral y distribución de la muestra en las etapas

Precisión prefijada

En muestreo bietápico, si se pide la distribución de la muestra (es decir, el reparto de la muestra entre conglomerados y unidades dentro de cada conglomerado), para obtener una determinada precisión o varianza del estimador, existen varias soluciones pues la ecuación

$$V(\overline{\overline{y}}) = \frac{L-n}{L}\frac{S_1^2}{n} + \frac{\overline{N}-m}{\overline{N}}\frac{S_2^2}{nm} = K$$

puede satisfacerse de manera aproximada para varias combinaciones de n y m.

Una manera práctica de afrontar el problema sería presentar un listado con esas posibles soluciones y el coste asociado a cada una de ellas, según una o varias funciones de coste. En esa tabla o listado se pueden descartar aquellas soluciones exageradamente carentes de equilibrio a juicio del investigador.

Otro planteamiento es prefijar el coste y obtener aquella solución que ofrece varianza mínima. También se puede minimizar el coste sujeto a cierta varianza prefijada, dando lugar al mismo tipo de desarrollos.

Coste prefijado

(1) Supongamos que la función de coste es $C = c_0 + nc_1 + nmc_2$. Con un desarrollo análogo al estudiado en la sección dedicada al tamaño muestral en muestreo monoetápico, se obtiene que $n^* = \dfrac{C - c_0}{c1 + mc_2}$ si suponemos fijos C y m.

(2) En caso en que la función de coste es $C = c_0 + \sqrt{n}c_1 + nmc_2$ o cualquier otra función compleja, pueden utilizarse técnicas de optimización, pero al ser n y m valores discretos, es más sencillo programar el cálculo del valor de la varianza y coste para una lista de n y m y centrarse en un intervalo de valores adecuados de C y V (por ejemplo, descartando la enumeración en la tabla de valores de C y V, fuera de los intervalos mencionados).

Ejemplo 10.2

Supongamos que en el ejemplo 10.1 de los lotes se tiene una buena estimación de $S_1^2 = 1.1$ y $S_2^2 = 15.8$. Supongamos que el tiempo y recursos de que se dispone obliga a fijar el número de componentes muestreadas en $nm = 30$ componentes. Como n y m deben de ser enteros, se puede calcular el error aproximado para diversas combinaciones de n y m tales que $nm \simeq 30$. Se realiza un programa informático asignando diversos valores a n y m y calculando $V(\overline{\overline{y}}) = \dfrac{L-n}{L}\dfrac{S_1^2}{n} + \dfrac{\overline{N}-m}{\overline{N}}\dfrac{S_2^2}{nm}$ para cada combinación. Nos restringimos en la salida presentada a aquellos valores de n y m tales que $nm \simeq 30$. Los primeros datos (ordenados por la varianza aproximada del estimador) son:

```
 n     m     n.m       Var

17     2     34    0.47508
11     3     33    0.49911
```


ISBN 9788496866133

```
16     2     32    0.50527
 8     4     32    0.52464
30     1     30    0.52914
15     2     30    0.53948
29     1     29    0.54766
```

Se observa que la mejor combinación tal que $nm = 30$ corresponde a tomar 30 conglomerados y una observación dentro de cada uno de ellos. Esto llevaría a no poder estimar s_{2i}^2 y por lo tanto a no poder utilizar el estimador de la varianza insesgado estudiado. Una solución alternativa es tomar 15 conglomerados con 2 observaciones cada uno (el aumento de la varianza del estimador respecto a la anterior solución es aproximadamente solo de 0.01).

10.2.5 Estimadores con m.a.s. en primera etapa y m.a.s.r. en segunda etapa

Los desarrollos anteriores se pueden modificar ligeramente para obtener los estimadores y varianzas en el caso en que en la segunda etapa la selección se haga con reemplazamiento.

Teorema 10.4 (estimación de la media y proporción).

En caso de muestreo por conglomerados bietápico con m.a.s. en primera etapa y m.a.s.r. en segunda etapa,

(a) Un estimador insesgado de la media es $\overline{\overline{y}} = \dfrac{1}{n}\sum_{i=1}^{n}\widehat{\overline{y}}_i$, con varianza

$$V(\overline{\overline{y}})' = \frac{L-n}{L}\frac{S_1^2}{n} + \frac{\sigma_2^2}{nm}.$$

(b) $\widehat{V}(\overline{\overline{y}})' = \dfrac{L-n}{L}\dfrac{s_1^2}{n} + \dfrac{s_2^2}{mL}$ es un estimador insesgado de $V(\overline{\overline{y}})'$.

(c) Un estimador insesgado de la proporción p es $\widehat{\widehat{p}} = \dfrac{1}{n}\sum_{i=1}^{n}\widehat{p}_i$ con varianza

$$V(\widehat{\widehat{p}})' = \frac{L-n}{L}\frac{S_1^2}{n} + \frac{\sigma_2^2}{nm}.$$

donde,

$$S_1^2 = \frac{1}{L-1}\sum_{i=1}^{L}(p_i - p)^2 \text{ y } \sigma_2^2 = \frac{1}{L}\sum_{i=1}^{L}p_i(1-p_i).$$

(d) Un estimador insesgado de $V(\widehat{\widehat{p}})'$ es

$$\widehat{V}(\widehat{\widehat{p}})' = \frac{L-n}{L}\frac{s_1^2}{n} + \frac{s_2^2}{mL}$$

donde

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{p}_i - \widehat{\widehat{p}})^2$$

ISBN 9788496866133

y

$$s_2^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{m}{m-1}\widehat{p}_i(1-\widehat{p}_i).$$

Demostración.

Recurriremos a los teoremas similares desarrollados para el caso de m.a.s. en ambas etapas.

(a) La demostración de que $\overline{\overline{y}}$ es insesgado es similar que para el caso de m.a.s. en ambas etapas.

La expresión de la varianza de $V(\overline{\overline{y}})$ es ligeramente distinta:

Al ser muestreo con reemplazamiento en segunda etapa, en el paso $(**)$ del teorema que calcula $V(\overline{\overline{y}})$ para m.a.s. en ambas etapas, hay que expresar $V_2^i(\widehat{\overline{y}}_i) = \dfrac{\sigma_{2i}^2}{m}$ en lugar de $V_2^i(\widehat{\overline{y}}_i) = \dfrac{\overline{N}-m}{\overline{N}}\dfrac{S_{2i}^2}{m}$. El resto se desarrolla de igual manera.

(b) Recurriendo al teorema que calcula $E(\widehat{V}(\overline{\overline{y}}))$ en m.a.s. en ambas etapas, basta hacer $V_2^i(\widehat{\overline{y}}_i) = \dfrac{\sigma_{2i}^2}{m}$ y $E_2^i(s_{2i}^2) = \sigma_{2i}^2$, y todos los desarrollos restantes son similares. Se obtiene que

$$E(s_1^2) = S_1^2 + \frac{\sigma_2^2}{nm} \text{ y } E(s_2^2) = \sigma_2^2,$$

y por lo tanto

$$E(\widehat{V}'(\overline{\overline{y}})) = E\left[\frac{L-n}{L}\frac{s_1^2}{n} + \frac{s_2^2}{mL}\right] = \frac{L-n}{L}\frac{S_1^2}{n} + \frac{\sigma_2^2}{nm}.$$

(c) y (d) Son consecuencia de lo inmediatamente anterior.

10.3 Conglomerados de distinto tamaño

Al muestrear una población extensa, suelen encontrarse unidades primarias que varían en tamaño. Además, por consideraciones de costes se suele recomendar el uso del muestreo en varias etapas, de modo que los problemas que se discuten en esta sección suceden con frecuencia. Así ocurre, por ejemplo, al considerar árboles como unidades primarias; en ese caso las unidades secundarias como los frutos no sumarían un número constante, igual en todos los árboles. Si los conglomerados son manzanas o edificios en una ciudad, también el número de hogares puede variar.

Para abordar este problema, se recurrirá a las técnicas observadas en muestreo por conglomerados monoetápico. Una posibilidad es la estratificación, agrupando en un mismo estrato conglomerados del mismo tamaño. Otra, utilizar técnicas como la estimación de razón, que pretende corregir una alta variabilidad del estimador en caso de tamaños muy desiguales. Por último, utilizar muestreo con probabilidades aproximadamente proporcionales al tamaño también ayuda a reducir la varianza del estimador, y permite evitar circunstancias molestas para

ISBN 9788496866133

el investigador que pueden ocurrir en m.a.s., donde en la muestra pueden aparecer con probabilidad no despreciable muchos conglomerados de pequeño tamaño, que pueden en conjunto ser poco representativos de la población.

Al tener los conglomerados tamaños desiguales está justificada una posible variación del tamaño de la muestra escogida dentro de cada conglomerado. En cada conglomerado de los escogidos en primera etapa se extraerá una muestra de tamaño m_i.

Se denotará, debido a los tamaños desiguales,

$$S_{2i}^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (y_{ij} - \overline{y}_i)^2 = \frac{N_i}{N_i - 1} \sigma_{2i}^2 = \text{Cuasivarianza en el conglomerado } i.$$

El esquema de este tema será presentar, en primer lugar, un estimador insesgado y a continuación el estimador de razón a tamaño alternativo. Seguidamente se estudiarán los casos de muestreo con probabilidades desiguales en primera etapa, con reemplazamiento y sin reemplazamiento, y respectivamente y como casos particulares de éstos , m.a.s.r. y m.a.s.

10.3.1 Estimación insesgada en caso de m.a.s. en primera y segunda etapa

Se presentará ahora un método de estimación insesgada en caso de m.a.s. en primera y segunda etapa, y tamaños desiguales.

Teorema 10.5 (estimador insesgado del total y varianzas).

Supongamos m.a.s. en ambas etapas, y tamaños de conglomerados desiguales.

(a) Un estimador insesgado para el total poblacional es

$$\widehat{N\overline{y}} = \frac{L}{n} \sum_{i=1}^{n} N_i \widehat{\overline{y}}_i \ .$$

(b) $$V(\widehat{N\overline{y}}) = \frac{L^2(1 - f_1)}{n(L-1)} \sum_{i=1}^{L} (y_i - \overline{N\overline{y}})^2 + \frac{L}{n} \sum_{i=1}^{L} \frac{N_i^2 (1 - f_{2i}) S_{2i}^2}{m_i},$$

donde $f_1 = \dfrac{n}{L}$ es la fracción de muestreo en primera etapa y $f_{2i} = \dfrac{m_i}{N_i}$ es la fracción de muestreo en segunda etapa.

(c) Un estimador insesgado de $V(\widehat{N\overline{y}})$ es

$$\widehat{V}(\widehat{N\overline{y}}) = \frac{L^2(1 - f_1)}{n(n-1)} \sum_{i=1}^{n} (N_i \widehat{\overline{y}}_i - \frac{1}{n} \sum_{i=1}^{n} N_i \widehat{\overline{y}}_i)^2 + \frac{L}{n} \sum_{i=1}^{n} \frac{N_i^2 (1 - f_{2i}) s_{2i}^2}{m_i}.$$

ISBN 9788496866133

Demostración.

Se utilizarán los resultados vistos cuando los tamaños de los conglomerados son iguales.

(a)

$$E(\widehat{N\overline{y}}) = E_1[E_2^i(\widehat{N\overline{y}})] = E_1[E_2^i(\frac{L}{n}\sum_{i=1}^{n} N_i\widehat{\overline{y}}_i)] = E_1\left[\frac{L}{n}\sum_{i=1}^{n} E_2^i(N_i\widehat{\overline{y}}_i)\right] =$$

$$=L \cdot E_1\left[\frac{1}{n}\sum_{i=1}^{n} N_i\overline{y}_i\right] = \sum_{i=1}^{L} N_i\overline{y}_i = N\overline{y}.$$

Hay que tener en cuenta en este caso que un estimador del total basado en el estimador visto en el caso de conglomerados de tamaño igual:

$\widehat{N\overline{y}}' = \frac{N}{n}\sum_{i=1}^{n}\widehat{\overline{y}}_i$ sería un **estimador sesgado**, pues

$$E_1[E_2^i(\frac{N}{n}\sum_{i=1}^{n}\widehat{\overline{y}}_i)] = NE_1[(\frac{1}{n}\sum_{i=1}^{n}\overline{y}_i)] =$$

$\frac{N}{L}\sum_{i=1}^{L}\overline{y}_i$ que es distinto de $\sum_{i=1}^{L} N_i\overline{y}_i$ en general.

Por lo tanto, en caso de tamaños desiguales, tanto para estimar el total como la media o proporción, es conveniente utilizar el estimador presentado en este teorema u otros estimadores que se verán a lo largo del capítulo.

(b) Basándose en la demostración del caso de conglomerados con tamaños iguales, basta realizar un desarrollo similar para ver que

$$V_1[E_2^i(\widehat{N\overline{y}})] = V_1\left[E_2^i\left(\frac{L}{n}\sum_{i=1}^{n} N_i\widehat{\overline{y}}_i\right)\right] = V_1\left[\frac{L}{n}\sum_{i=1}^{n} N_i\overline{y}_i\right] =$$

$$= (1-f_1)\frac{L^2}{n}\frac{1}{L-1}\sum_{i=1}^{L}(y_i - \frac{1}{L}\sum_{i=1}^{L} y_i)^2 = \frac{L^2(1-f_1)}{n(L-1)}\sum_{i=1}^{L}(y_i - \frac{N}{L}\overline{y})^2 =$$

$$=\frac{L^2(1-f_1)}{n(L-1)}\sum_{i=1}^{L}(y_i - \overline{N\overline{y}})^2.$$

Además,

$$E_1[V_2^i(\widehat{N\overline{y}})] = E_1\left[V_2^i\left(\frac{L}{n}\sum_{i=1}^{n} N_i\widehat{\overline{y}}_i\right)\right] = E_1\left[\frac{L^2}{n^2}\sum_{i=1}^{n} N_i^2 V_2^i(\widehat{\overline{y}}_i)\right] =$$

$$=E_1\left[\frac{L^2}{n^2}\sum_{i=1}^{n} N_i^2(1-f_{2i})\frac{S_{2i}^2}{m_i}\right] =$$

$$=\frac{L^2}{nL}\sum_{i=1}^{L} N_i^2(1-f_{2i})\frac{S_{2i}^2}{m_i} = \frac{L}{n}\sum_{i=1}^{L} N_i^2(1-f_{2i})\frac{S_{2i}^2}{m_i}.$$

Así,

ISBN 9788496866133

$$V(\widehat{N\overline{y}}) = V_1[E_2^i(\widehat{N\overline{y}})] + E_1[V_2^i(\widehat{N\overline{y}})] = \frac{L^2(1-f_1)}{n(L-1)}\sum_{i=1}^{L}(y_i - \overline{N}\overline{y})^2 + \frac{L}{n}\sum_{i=1}^{L}\frac{N_i^2(1-f_{2i})S_{2i}^2}{m_i}.$$

(c) Sea $E_1E_2^i\left[\dfrac{1}{n-1}\sum_{i=1}^{n}(N_i\widehat{\overline{y}}_i - \dfrac{1}{n}\sum_{i=1}^{n}N_i\widehat{\overline{y}}_i)^2\right]$.

Denotando $\widehat{y}_i = N_i\widehat{\overline{y}}_i$, y $\overline{\widehat{y}} = \dfrac{1}{n}\sum_{i=1}^{n}N_i\widehat{\overline{y}}_i$, queda

$$E_1\left[\frac{1}{n-1}E_2^i\left[\sum_{i=1}^{n}\widehat{y}_i^2 - n\overline{\widehat{y}}^2\right]\right].$$

Ahora,

$$E_2^i\left[\sum_{i=1}^{n}\widehat{y}_i^2 - n\overline{\widehat{y}}^2\right] = \sum_{i=1}^{n}E_2^i(\widehat{y}_i^2) - nE_2^i(\overline{\widehat{y}}^2) =$$

$$= \sum_{i=1}^{n}E_2^i(\widehat{y}_i)^2 + \sum_{i=1}^{n}V_2^i(\widehat{y}_i) - n\left[(E_2^i(\overline{\widehat{y}}))^2 + V_2^i(\overline{\widehat{y}})\right] =$$

$$= \sum_{i=1}^{n}y_i^2 + \sum_{i=1}^{n}N_i^2(1-f_{2i})\frac{S_{2i}^2}{m_i} - n\left[\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)^2 + \frac{1}{n^2}\sum_{i=1}^{n}N_i^2(1-f_{2i})\frac{S_{2i}^2}{m_i}\right] =$$

$$= \sum_{i=1}^{n}y_i^2 - n\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)^2 + \frac{n-1}{n}\sum_{i=1}^{n}N_i^2(1-f_{2i})\frac{S_{2i}^2}{m_i}.$$

Por lo tanto,

$$E_1\left[\frac{1}{n-1}E_2^i\left[\sum_{i=1}^{n}\widehat{y}_i^2 - n\overline{\widehat{y}}^2\right]\right] =$$

$$= E_1\left[\frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^2 - n\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)^2\right] + \frac{1}{n}\sum_{i=1}^{n}(1-f_{2i})\frac{S_{2i}^2}{m_i}\right] =$$

$$= \frac{1}{L-1}\sum_{i=1}^{L}(y_i - \overline{N}\overline{y})^2 + \frac{1}{L}\sum_{i=1}^{L}N_i^2(1-f_{2i})\frac{S_{2i}^2}{m_i}.$$

Así, la esperanza del primer término en el estimador de la varianza queda:

$$E\left[\frac{L^2(1-f_1)}{n}\frac{1}{n-1}\sum_{i=1}^{n}(N_i\widehat{\overline{y}}_i - \frac{1}{n}\sum_{i=1}^{n}N_i\widehat{\overline{y}}_i)^2\right] =$$

$$= \frac{L^2(1-f_1)}{n(L-1)}\sum_{i=1}^{L}(y_i - \overline{N}\overline{y})^2 + \frac{L^2(1-f_1)}{Ln}\sum_{i=1}^{L}N_i^2(1-f_{2i})\frac{S_{2i}^2}{m_i}.$$

Por otra parte, en el segundo término:

$$E_1\left[\frac{L}{n}\sum_{i=1}^{n}E_2^i\left[\frac{N_i^2(1-f_{2i})s_{2i}^2}{m_i}\right]\right] = E_1\left[\frac{L}{n}\sum_{i=1}^{n}\frac{N_i^2(1-f_{2i})E_2^i(s_{2i}^2)}{m_i}\right] =$$

$$= E_1\left[L\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{N_i^2(1-f_{2i})S_{2i}^2}{m_i}\right] = \sum_{i=1}^{L}\frac{N_i^2(1-f_{2i})S_{2i}^2}{m_i}.$$

ISBN 9788496866133

Así que

$$E_1\left[E_2^i\left[\widehat{V}(\widehat{N\overline{y}})\right]\right] =$$

$$=\frac{L^2(1-f_1)}{n(L-1)}\sum_{i=1}^{L}(y_i-\overline{N\overline{y}})^2+\frac{L^2(1-f_1)}{nL}\sum_{i=1}^{L}N_i^2(1-f_{2i})\frac{S_{2i}}{m_i}+\sum_{i=1}^{L}\frac{N_i^2(1-f_{2i})S_{2i}^2}{m_i}.$$

Los dos últimos sumandos se pueden simplificar. Fijándose en sus coeficientes:

$$\frac{L^2(1-f_1)}{nL}+1=\frac{L(1-f_1)+n}{n}=\frac{L(\frac{L-n}{L})+n}{n}=\frac{L}{n}.$$

De modo que:

$$E_1\left[E_2^i\left[\widehat{V}(\widehat{N\overline{y}})\right]\right]=\frac{L^2(1-f_1)}{n(L-1)}\sum_{i=1}^{L}(y_i-\overline{N\overline{y}})^2+\frac{L}{n}\sum_{i=1}^{L}N_i^2(1-f_{2i})\frac{S_{2i}}{m_i}$$

que es lo que se quería demostrar.

10.3.2 Estimación de razón a tamaño en caso de m.a.s. en primera y segunda etapa

Como se ha visto en temas anteriores, una manera de evitar los problemas de representación y alta varianza del estimador con estimación insesgada cuando los conglomerados son de tamaños desiguales, es utilizar el estimador de razón a tamaño. En muestreo por conglomerados bietápico tiene tanto interés como tenía en muestreo monoetápico. También es cierto que en la segunda etapa del muestreo bietápico siempre se puede tomar un tamaño muestral m mayor en los conglomerados más grandes, pero nada garantiza, bajo m.a.s. de las unidades de primera etapa, que los conglomerados más grandes y representativos caigan en la muestra, al estar asignándoles probabilidades iguales a todos. La estimación de razón corrige estas posibilidades que surgen en m.a.s. en primera etapa, a cambio de introducir un sesgo en la estimación que suele ser pequeño en la práctica.

Teorema 10.6 (estimador de razón a tamaño).

Se define el estimador de razón a tamaño del total poblacional $N\overline{y}$ como

$$N\widehat{\overline{y}}_R=N\frac{\sum_{i=1}^{n}N_i\widehat{\overline{y}}_i}{\sum_{i=1}^{n}N_i}.$$

Entonces

(a) Una expresión aproximada de la varianza de $N\widehat{\overline{y}}_R$ es

$$V(N\widehat{\overline{y}}_R)=\frac{L^2(1-f_1)}{n}\sum_{i=1}^{L}\frac{N_i^2(\overline{y}_i-\overline{y})^2}{L-1}+\frac{L}{n}\sum_{i=1}^{L}\frac{N_i^2(1-f_{2i})S_{2i}^2}{m_i}$$

(b) Un estimador de los momentos de $V(\widehat{\overline{y}}_R)$ es

ISBN 9788496866133

$$\widehat{V}(N\widehat{\overline{y}}_R) = \frac{L^2(1-f_1)}{n}\sum_{i=1}^{n}\frac{N_i^2(\widehat{\overline{y}}_i - \widehat{\overline{y}}_R)^2}{(n-1)} + \frac{L}{n}\sum_{i=1}^{n}\frac{N_i^2(1-f_{2i})s_{2i}^2}{m_i}.$$

Demostración.

(a)

A partir del Teorema de Madow:

$$V_1[E_2^i(N\widehat{\overline{y}}_R)] = V_1\left[N\cdot\frac{\sum_{i=1}^{n}N_iE_2^i(\widehat{\overline{y}}_i)}{\sum_{i=1}^{n}N_i}\right] = N^2V_1\left[\frac{\sum_{i=1}^{n}N_i\overline{y}_i}{\sum_{i=1}^{n}N_i}\right] \simeq$$

$$\simeq N^2\frac{1}{\overline{N}^2}\frac{L-n}{Ln}\frac{1}{L-1}\sum_{i=1}^{L}N_i^2(\overline{y}_i-\overline{y})^2 = = \frac{L^2(1-f_1)}{n}\sum_{i=1}^{L}\frac{N_i^2(\overline{y}_i-\overline{y})^2}{L-1}$$

donde se ha utilizado el resultado visto en estimación de razón a tamaño en muestreo por conglomerados monoetápico.

Además,

$$E_1[V_2^i(N\widehat{\overline{y}}_R)] = N^2E_1[V_2^i(\widehat{\overline{y}}_R)] = N^2E_1\left[\frac{\sum_{i=1}^{n}N_i^2V_2^i(\widehat{\overline{y}}_i)}{(\sum_{i=1}^{n}N_i)^2}\right] =$$

$$= N^2E_1\left[\frac{\sum_{i=1}^{n}N_i^2\frac{(1-f_{2i})S_{2i}^2}{m_i}}{(\sum_{i=1}^{n}N_i)^2}\right] = \frac{1}{n^2}N^2E_1\left[\frac{\sum_{i=1}^{n}N_i^2\frac{(1-f_{2i})S_{2i}^2}{m_i}}{\widehat{\overline{N}}^2}\right] \overset{(*)}{\simeq}$$

$$\overset{(*)}{\simeq} \frac{L^2}{n}E_1\left[\frac{1}{n}\sum_{i=1}^{n}N_i^2\frac{(1-f_{2i})S_{2i}^2}{m_i}\right] = \frac{L}{n}\sum_{i=1}^{L}N_i^2\frac{(1-f_{2i})S_{2i}^2}{m_i}.$$

Donde en $(*)$ se ha utilizado la notación $\widehat{\overline{N}} = \frac{1}{n}\sum_{i=1}^{n}N$ y se aproximado la esperanza, extrayendo el término muestral $\widehat{\overline{N}}^2$ en el denominador fuera de la esperanza, y además se ha aproximado $N^2 \simeq L^2\widehat{\overline{N}}^2$.

Sumando los dos términos $V_1[E_2^i(N\widehat{\overline{y}}_R)]$ y $E_1[V_2^i(N\widehat{\overline{y}}_R)]$, se obtiene el resultado.

(b) No es necesario demostrarlo, pues se han sustituido los momentos poblacionales

$$\sum_{i=1}^{L}\frac{N_i^2(\overline{y}_i-\overline{y})^2}{L-1} \text{ y } \sum_{i=1}^{L}\frac{N_i^2(1-f_{2i})S_{2i}^2}{m_i}$$

por sus respectivos muestrales,

$$\sum_{i=1}^{n}\frac{N_i^2(\widehat{\overline{y}}_i-\widehat{\overline{y}}_R)^2}{(n-1)} \text{ y } \sum_{i=1}^{n}\frac{N_i^2(1-f_{2i})s_{2i}^2}{m_i}.$$

ISBN 9788496866133

Ejemplo 10.3

Supongamos el mismo objetivo que en el ejemplo 9.7, es decir, estimar la población de hombres en España en 1998. Se desea utilizar muestreo por conglomerados bietápico. Se escogerán 5 provincias por m.a.s.

Se seleccionarán por m.a.s., dentro de cada provincia (conglomerado) escogida en primera etapa, una cantidad m_i de municipios. Para otorgar más importancia a las provincias con más municipios, se tomará $m_i = 0.1 \cdot N_i$, es decir, la fracción de muestreo en segunda etapa será $f_{2i} = \dfrac{m_i}{N_i} = 0.1$ constante. Hay que percatarse de que utilizar una fracción de muestreo constante en lugar de m_i constante para todo i, implica no conocer a priori el tamaño muestral final, pues éste depende de los conglomerados escogidos.

Tenemos por lo tanto, que $L = 52$, $N = 8098$ y $n = 5$. Tras el proceso de muestreo, se obtienen los datos presentados en la siguiente tabla:

Provincia	nº municipios=N_i	m_i	$\widehat{\overline{y}}_i$	s_{2i}^2
La Rioja	174	17	3831.59	212239434
Asturias	78	8	2140.14	2136467.14
Madrid	179	18	6722.22	401646298
Sevilla	105	10	5046.40	45697537.60
Cáceres	219	22	1738.80	5554809.73

Tabla 10.2. Características muestrales

La estimación insesgada del total será

$$\widehat{N\overline{y}} = \frac{L}{n}\sum_{i=1}^{n} N_i\widehat{\overline{y}}_i = 30.654.771.$$

Y la varianza estimada del estimador:

$$\widehat{V}(\widehat{N\overline{y}}) = \frac{L^2(1-f_1)}{n(n-1)}\sum_{i=1}^{n}(N_i\widehat{\overline{y}}_i - \frac{1}{n}\sum_{i=1}^{n} N_i\widehat{\overline{y}}_i)^2 + \frac{L}{n}\sum_{i=1}^{n}\frac{N_i^2(1-f_{2i})s_{2i}^2}{m_i} =$$

$$= 7.4 \cdot 10^{13} + 1.08 \cdot 10^{13} = 8.48 \cdot 10^{13}.$$

El estimador de razón a tamaño del total es

$$N\widehat{\overline{y}}_R = N\frac{\sum_{i=1}^{n} N_i\widehat{\overline{y}}_i}{\sum_{i=1}^{n} N_i} = 8098\frac{2947574.16}{755} = 31.615.172.$$

Y la varianza estimada del estimador:

$$\widehat{V}(N\widehat{\overline{y}}_R) = \frac{L^2(1-f_1)}{n}\sum_{i=1}^{n}\frac{N_i^2(\widehat{\overline{y}}_i - \widehat{\overline{y}}_R)^2}{(n-1)} + \frac{L}{n}\sum_{i=1}^{n}\frac{N_i^2(1-f_{2i})s_{2i}^2}{m_i} =$$

ISBN 9788496866133

$= 6.23 \cdot 10^{13} + 1.08 \cdot 10^{13} = 7.31 \cdot 10^{13}$.

Ambos estimadores tienen una alta varianza, en comparación con el ejemplo 9.8, donde se utilizaba muestreo proporcional a una variable auxiliar . También hay que tener en cuenta que el tamaño de la muestra final era mucho mayor al ser muestreo monoetápico en el ejemplo mencionado, y que se escogían 10 conglomerados en vez de 5 como en este ejemplo.

10.3.3 Estimación en muestreo con probabilidades desiguales y reemplazamiento en primera etapa y m.a.s. o m.a.s.r. en segunda etapa

En muestreo bietápico no es infrecuente que la primera etapa se haga con reemplazamiento, pues se reducen mucho los costes en casos de estudios en poblaciones grandes. Los estimadores son por otra parte más sencillos que en muestreo sin reemplazamiento. Además, el hecho de utilizar probabilidades proporcionales al tamaño puede ayudar en general a mejorar la precisión del estimador.

A continuación se presenta un estimador insesgado del tipo Hansen-Hurwitz en el caso de que se seleccionen los conglomerados con probabilidades respectivas p_i y reemplazamiento, y m.a.s. o m.a.s.r. en segunda etapa.

Teorema 10.7 (estimador insesgado del total y varianzas).

(a) Un estimador insesgado del total $N\overline{y}$ en el caso de muestreo con probabilidades desiguales y reemplazamiento en primera etapa y m.a.s. o m.a.s.r. en segunda etapa es

$$\widehat{N\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{p_i}$$

(b) Si el muestreo en segunda etapa es m.a.s., la varianza de $\widehat{N\overline{y}}$ es

$$V(\widehat{N\overline{y}}) = \frac{1}{n}\left(\sum_{i=1}^{L}\frac{y_i^2}{p_i} - (N\overline{y})^2\right) + \frac{1}{n}\sum_{i=1}^{L}\frac{N_i - m_i}{N_i}\frac{N_i^2}{p_i m_i}S_{2i}^2.$$

donde $y_i = N_i\overline{y}_i$ es el total en el conglomerado i.

(c) Si el muestreo en segunda etapa es m.a.s.r., la varianza de $\widehat{N\overline{y}}$ es

$$V(\widehat{N\overline{y}}) = \frac{1}{n}\left(\sum_{i=1}^{L}\frac{y_i^2}{p_i} - (N\overline{y})^2\right) + \frac{1}{n}\sum_{i=1}^{L}\frac{N_i^2}{p_i m_i}\sigma_{2i}^2.$$

donde $y_i = N_i\overline{y}_i$ es el total en el conglomerado i.

Demostración.

(a) $E(\widehat{N\overline{y}}) = E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{p_i}\right] = E_1\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{N_i}{p_i}E_2^i(\widehat{\overline{y}}_i)\right]\right] = E_1\left[\frac{1}{n}\sum_{i=1}^{n}\frac{N_i\overline{y}_i}{p_i}\right] = N\overline{y}$

pues la última igualdad deriva del hecho de que se trata del estimador de Hansen-Hurwitz del total, que es insesgado .

ISBN 9788496866133

(b) Se utilizará el Teorema de Madow:

$V(\widehat{\theta}) = V_1[E_2^i(\widehat{\theta})] + E_1[V_2^i(\widehat{\theta})]$

con $\widehat{\theta} = \widehat{N\overline{y}}$.

En primer lugar,

$E_2^i(\widehat{N\overline{y}}) = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i\overline{y}_i}{p_i}$ como se ha visto en la demostración de (a). Entonces,

$$V_1[E_2^i(\widehat{N\overline{y}})] = V_1[\frac{1}{n}\sum_{i=1}^{n}\frac{N_i\overline{y}_i}{p_i}] = \frac{1}{n}\left(\sum_{i=1}^{L}\frac{y_i^2}{p_i} - (N\overline{y})^2\right)$$

pues es la varianza del estimador del total de Hansen-Hurwitz.

Para desarrollar el segundo término en el Teorema de Madow:

$$V_2^i(\widehat{N\overline{y}}) = V_2^i\left(\frac{1}{n}\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{p_i}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{N_i^2}{p_i^2}V_2^i(\widehat{\overline{y}}_i) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{N_i^2}{p_i^2}\frac{N_i - m_i}{N_i}\frac{S_{2i}^2}{m_i}$$

por ser m.a.s. en segunda etapa y por lo tanto $V_2^i(\widehat{\overline{y}}_i)$ la varianza de la media muestral en m.a.s. Así,

$$E_1[V_2^i(\widehat{N\overline{y}})] = \frac{1}{n^2}\sum_{i=1}^{n}E_1\left[\frac{N_i^2}{p_i^2}\frac{N_i - m_i}{N_i}\frac{S_{2i}^2}{m_i}\right].$$

Ahora, como la primera etapa consiste en muestreo con probabilidades p_i y reemplazo, entonces

$$\frac{N_i^2}{p_i^2}\frac{N_i - m_i}{N_i}\frac{S_{2i}^2}{m_i}$$

es una variable aleatoria que toma valores

$$\frac{N_1^2}{p_1^2}\frac{N_1 - m_1}{N_1}\frac{S_{21}^2}{m_1}, ..., \frac{N_L^2}{p_L^2}\frac{N_L - m_L}{N_L}\frac{S_{2L}^2}{m_L}$$

con probabilidades respectivas $p_1, ..., p_L$.

Por la definición de esperanza,

$$E_1[V_2^i(\widehat{N\overline{y}})] = \frac{1}{n^2}\sum_{i=1}^{n}E_1\left[\frac{N_i^2}{p_i^2}\frac{N_i - m_i}{N_i}\frac{S_{2i}^2}{m_i}\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i=1}^{L}\frac{N_i^2}{p_i^2}\frac{N_i - m_i}{N_i}\frac{S_{2i}^2}{m_i}p_i =$$

$$\frac{1}{n}\sum_{i=1}^{L}\frac{N_i - m_i}{N_i}\frac{N_i^2}{p_i m_i}S_{2i}^2$$

y, por lo tanto,

$$V(\widehat{N\overline{y}}) = V_1[E_2^i(\widehat{N\overline{y}})] + E_1[V_2^i(\widehat{N\overline{y}})] = \frac{1}{n}\left(\sum_{i=1}^{L}\frac{y_i^2}{p_i} - (N\overline{y})^2\right) + \frac{1}{n}\sum_{i=1}^{L}\frac{N_i - m_i}{N_i}\frac{N_i^2}{p_i m_i}S_{2i}^2 .$$

ISBN 9788496866133

(c) Basta realizar el desarrollo anterior, pero con $V_2^i(\widehat{\overline{y}}_i) = \dfrac{\sigma_{2i}^2}{m_i}$ por ser la varianza de la media muestral en muestreo con reemplazamiento. Por lo tanto, en la expresión final resulta $\dfrac{\sigma_{2i}^2}{m_i}$ en lugar de $\dfrac{N_i - m_i}{N_i}\dfrac{S_{2i}^2}{m_i}$.

El siguiente método de estimación de varianzas se debe a Hansen, Hurwitz y Madow (1958) y se denomina **método de los conglomerados últimos** para la estimación de varianzas. Se denomina conglomerado último al conjunto de las unidades elementales que forman parte de cada unidad de primera etapa. Es un método muy práctico y bastante utilizado, pues permite estimar las varianzas del estimador sin necesidad de conocer las varianzas intra-conglomerados o entre conglomerados de las etapas previas.

Su principal virtud es la simplicidad en la estimación de la varianza. Además, como se ha comentado ya, la estimación de la varianza suponiendo muestreo con reemplazo en primera etapa siempre servirá como una cota superior, conservadora, para el mismo diseño de muestreo, pero realizando muestreo sin reemplazamiento. Así que en la práctica no es infrecuente que se realice muestreo sin reemplazamiento, y, para evitar las complicaciones derivadas de las estimaciones usuales (ver estimadores de Horvitz-Thompson), se utilicen las estimaciones por el método de los conglomerados últimos.

Teorema 10.8 (método de los conglomerados últimos).

Supongamos que en muestreo por conglomerados,

(1) Se seleccionan n conglomerados en muestreo con reemplazamiento, con probabilidades iguales o desiguales, en primera etapa.

(2) Se construye un estimador insesgado $\widehat{\theta}_i$ del total poblacional $N\overline{y}$, basado solamente en la información del conglomerado i, para $i = 1, ..., L$.

Entonces,

(a) El estimador $\widehat{\theta} = \dfrac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_i$ es un estimador insesgado del total poblacional.

(b) Un estimador insesgado de la varianza de $\widehat{\theta}$ es

$$\widehat{V}(\widehat{\theta}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\widehat{\theta}_i - \widehat{\theta})^2.$$

Demostración.

(a) Al ser insesgado $\widehat{\theta}_i$ para $\theta = N\overline{y}$, se tiene que $E(\widehat{\theta}_i) = E_1\left[E_2^i(\widehat{\theta}_i)\right] = \theta$. Entonces,

$$E(\widehat{\theta}) = E\left[\frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_i\right] = E_1\left[E_2^i\left[\frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_i\right]\right] = \left[\frac{1}{n}\sum_{i=1}^{n}\left[E_1E_2^i(\widehat{\theta}_i)\right]\right] = \frac{1}{n}\sum_{i=1}^{n}N\overline{y} = N\overline{y}.$$

ISBN 9788496866133

(b)

Se verán seguidamente algunas relaciones de interés para la demostración.

Hay que notar en primer lugar que $E(\widehat{\theta}_i) = E(\widehat{\theta}_j)$ y además, $V(\widehat{\theta}_i) = V(\widehat{\theta}_j)$ para todos los $i, j = 1, ..., L$ pues si p_i es la probabilidad de obtener el conglomerado i-ésimo en primera etapa,

$$V(\widehat{\theta}_i) = E(\widehat{\theta}_i^2) - \theta^2 = E_1\left[E_2^i(\widehat{\theta}_i^2)\right] - \theta^2 = \sum_{i=1}^{L} p_i E_2^i(\widehat{\theta}_i^2) - \theta^2 =$$

$$= \sum_{j=1}^{L} p_j E_2^j(\widehat{\theta}_j^2) - \theta^2 = V(\widehat{\theta}_j).$$

Por otra parte, al ser insesgados $\widehat{\theta}_i$ y $\widehat{\theta}$ para $\theta = N\overline{y}$, se tiene que

$E(\widehat{\theta}) = E(\widehat{\theta}_i) = E_1\left[E_2^i(\widehat{\theta}_i)\right] = \theta$, y que $V(\widehat{\theta}) = E(\widehat{\theta} - \theta)^2$.

Además,

$$V(\widehat{\theta}) = V\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_i\right) = \frac{1}{n^2}\left[\sum_{i=1}^{n} V(\widehat{\theta}_i) + \sum_{i\neq j}^{n} cov(\widehat{\theta}_i, \widehat{\theta}_j)\right].$$

Pero la covarianza $cov(\widehat{\theta}_i, \widehat{\theta}_j) = 0$:

$$cov(\widehat{\theta}_i, \widehat{\theta}_j) = E(\widehat{\theta}_i\widehat{\theta}_j) - E(\widehat{\theta}_i)E(\widehat{\theta}_j) = E_1\left[E_2^{i,j}(\widehat{\theta}_i\widehat{\theta}_j)\right] - \theta^2 \overset{(*)}{=}$$

$$= E_1\left[E_2^i(\widehat{\theta}_i)E_2^j(\widehat{\theta}_j)\right] - \theta^2 \overset{(**)}{=} E_1\left[E_2^i(\widehat{\theta}_i)\right]E_1\left[E_2^j(\widehat{\theta}_j)\right] - \theta^2 = \theta^2 - \theta^2 = 0.$$

La igualdad $(*)$ se da por ser la distribución de muestreo en segunda etapa independiente de unos conglomerados a otros, pues se refiere solamente a muestras intra-conglomerados.

La igualdad $(**)$ es cierta pues el muestreo **en primera etapa** es **muestreo con reemplazamiento** y por lo tanto hay independencia en el muestreo de primera etapa entre los conglomerados escogidos i y j.

En caso de **muestreo sin reemplazamiento en primera etapa**, no sería cierta esta última igualdad $(**)$, porque habría dependencia entre los conglomerados escogidos. Por consiguiente el resultado del Teorema de los conglomerados últimos no sería cierto en ese caso.

Debido a que $cov(\widehat{\theta}_i, \widehat{\theta}_j) = 0$ para $i \neq j$, se tiene que $V(\widehat{\theta}) = \frac{1}{n^2}\sum_{i=1}^{n} V(\widehat{\theta}_i) = \frac{1}{n}V(\widehat{\theta}_i)$ para todo i, por ser todos los $V(\widehat{\theta}_i)$ iguales.

Ahora, para demostrar que $\widehat{V}(\widehat{\theta})$ es insesgado, calculemos su esperanza, utilizando que $E(\widehat{\theta}) = E(\widehat{\theta}_i) = \theta$ para todo i:

$$E(\widehat{V}(\widehat{\theta})) = \frac{1}{n(n-1)}\sum_{i=1}^{n} E[(\widehat{\theta}_i - \widehat{\theta})^2] = \frac{1}{n(n-1)}\sum_{i=1}^{n} E[(\widehat{\theta}_i - \theta + \theta - \widehat{\theta})^2] =$$

$$= \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} E[(\widehat{\theta}_i - \theta)^2] + 2\sum_{i=1}^{n} E[(\widehat{\theta}_i - \theta)(\theta - \widehat{\theta})] + nE(\widehat{\theta} - \theta)^2\right) =$$

ISBN 9788496866133

$$=\frac{1}{n(n-1)}\left(\sum_{i=1}^{n}E(\widehat{\theta}_i^2)-\sum_{i=1}^{n}2\theta E(\widehat{\theta}_i)+n\theta^2+\sum_{i=1}^{n}2\theta E(\widehat{\theta}_i)\right.$$

$$\left.-2\sum_{i=1}^{n}E(\widehat{\theta}_i\widehat{\theta})-2n\theta^2+2n\theta^2+nE(\widehat{\theta}-\theta)^2\right)=$$

$$=\frac{1}{n(n-1)}\left(\sum_{i=1}^{n}E(\widehat{\theta}_i^2)+n\theta^2-2\sum_{i=1}^{n}E(\widehat{\theta}_i\widehat{\theta})+nE(\widehat{\theta}-\theta)^2\right)=$$

$$=\frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\left[V(\widehat{\theta}_i)+E(\widehat{\theta}_i)^2\right]+n\theta^2-2\sum_{i=1}^{n}E(\widehat{\theta}_i\widehat{\theta})+nV(\widehat{\theta})\right)=$$

$$=\frac{1}{(n-1)}\frac{1}{n}\left(\sum_{i=1}^{n}V(\widehat{\theta}_i)+n\theta^2+n\theta^2-2\sum_{i=1}^{n}E(\widehat{\theta}_i\widehat{\theta})+nV(\widehat{\theta})\right)=$$

$$=\frac{1}{n-1}\left(V(\widehat{\theta}_i)+2\theta^2-\frac{2}{n}\sum_{i=1}^{n}E(\widehat{\theta}_i\widehat{\theta})+V(\widehat{\theta})\right)=$$

$$=\frac{1}{n-1}\left(nV(\widehat{\theta})+\frac{2}{n}\sum_{i=1}^{n}\left[E(\widehat{\theta}_i)E(\widehat{\theta})-E(\widehat{\theta}_i\widehat{\theta})\right]+V(\widehat{\theta})\right)$$

donde en la última igualdad se ha sustituido θ^2 por $\frac{1}{n}\sum_{i=1}^{n}\theta^2=\frac{1}{n}\sum_{i=1}^{n}E(\widehat{\theta}_i)E(\widehat{\theta})$ por ser $E(\widehat{\theta})=E(\widehat{\theta}_i)=\theta$.

Entonces , se tiene que

$$E(\widehat{V}(\widehat{\theta}))=\frac{1}{n-1}\left((n+1)V(\widehat{\theta})-\frac{2}{n}\sum_{i=1}^{n}cov(\widehat{\theta}_i,\widehat{\theta})\right).$$

Desarrollando la covarianza,

$$cov(\widehat{\theta}_i,\widehat{\theta})=cov(\widehat{\theta}_i,\frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_i)=\frac{1}{n}cov(\widehat{\theta}_i,\sum_{i=1}^{n}\widehat{\theta}_i)=\frac{1}{n}\left[cov(\widehat{\theta}_i,\widehat{\theta}_i)+\sum_{i\neq j}^{n}cov(\widehat{\theta}_i,\widehat{\theta}_j)\right]=$$

$$=\frac{1}{n}\left[V(\widehat{\theta}_i)+0\right]=\frac{1}{n}V(\widehat{\theta}_i).$$

Así,

$$E(\widehat{V}(\widehat{\theta}))=\frac{1}{n-1}\left((n+1)V(\widehat{\theta})-\frac{2}{n}\sum_{i=1}^{n}\frac{1}{n}V(\widehat{\theta}_i)\right)=\frac{1}{n-1}\left((n+1)V(\widehat{\theta})-2V(\widehat{\theta})\right)=$$

$$=\frac{1}{n-1}((n-1)V(\widehat{\theta}))=V(\widehat{\theta}).$$

Hay que remarcar que en esta última parte de la demostración no ha hecho falta utilizar el desarrollo $E_1\left[E_2^i(.)\right]$, pues la demostración se basa en propiedades básicas de esperanzas, varianzas y covarianzas, que son ciertas independientemente de la forma del estimador. Por ello, independientemente del tipo de muestreo que se realice en segunda etapa, siempre que se pueda construir un estimador insesgado del total poblacional a partir de la información de un solo conglomerado, y con tal de realizar muestreo con reemplazamiento en primera etapa,

ISBN 9788496866133

sea con probabilidades iguales o desiguales, se podrá aplicar este método. Si el estimador es sesgado, como sería el caso de estimación de razón en segunda etapa, pero el sesgo se puede considerar despreciable, se puede utilizar también esta aproximación a la varainza.

También hay que decir que la estimación de la varianza sólo es posible si $n \geq 2$, como suele ocurrir en general. En muestreo por conglomerados el caso en que se selecciona un sólo conglomerado existe a veces en la práctica, sobre todo cuando hay estratificación previa de los conglomerados.

Corolario 10.2 (estimación de la varianza del estimador del total).

Supongamos el caso de muestreo con reemplazamiento con probabilidades desiguales en primera etapa y m.a.s. o bien m.a.s.r. en segunda etapa.

Un estimador insesgado de la varianza del estimador del total $\widehat{N\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{p_i}$ es

$$\widehat{V}(\widehat{N\overline{y}}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{N_i\widehat{\overline{y}}_i}{p_i} - \widehat{N\overline{y}})^2.$$

Demostración.

Basta aplicar el método de los conglomerados últimos. En efecto, un estimador insesgado del total $N\overline{y}$ basado solamente en el conglomerado i es $\frac{N_i\widehat{\overline{y}}_i}{p_i}$, pues

$$E\left[\frac{N_i\widehat{\overline{y}}_i}{p_i}\right] = E_1\left[\frac{N_i}{p_i}E_2^i(\widehat{\overline{y}}_i)\right] = E_1\left[\frac{N_i\overline{y}_i}{p_i}\right]$$

Cada $\frac{N_i\overline{y}_i}{p_i}$ en primera etapa es una variable aleatoria que toma valores $\frac{N_1\overline{y}_1}{p_1}, ..., \frac{N_L\overline{y}_L}{p_L}$ con probabilidades $p_1, ..., p_L$ y por lo tanto

$$E_1\left[\frac{N_i\overline{y}_i}{p_i}\right] = \sum_{j=1}^{L} p_j\frac{N_j\overline{y}_j}{p_j} = \sum_{j=1}^{L} N_j\overline{y}_j = N\overline{y}.$$

Así, en el método de los conglomerados últimos $\theta = N\overline{y}$, $\widehat{\theta}_i = \frac{N_i\widehat{\overline{y}}_i}{p_i}$, y $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_i = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{p_i} = \widehat{N\overline{y}}$. Por lo tanto un estimador de la varianza de $\widehat{N\overline{y}}$ es

$$\widehat{V}(\widehat{N\overline{y}}) = \widehat{V}(\widehat{\theta}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\widehat{\theta}_i - \widehat{\theta})^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{N_i\widehat{\overline{y}}_i}{p_i} - \widehat{N\overline{y}})^2.$$

El resultado es válido independientemente de que en segunda etapa la estimación de $\widehat{\overline{y}}_i$ sea a través de m.a.s. o m.a.s.r, pues lo que importa es que sea insesgado y con reemplazamiento en primera etapa para aplicar el método de conglomerados últimos.

ISBN 9788496866133

Corolario 10.3 (Estimación de la media y proporción).

Supongamos que el muestreo es con reemplazamiento y probabilidades p_i en primera etapa, y m.a.s. o m.a.s.r. en segunda etapa.

(a) Un estimador insesgado de la media poblacional es $\frac{1}{N}\widehat{N\overline{y}} = \frac{1}{Nn}\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{p_i}$, con varianza $\frac{1}{N^2}V(\widehat{N\overline{y}})$ y estimador insesgado de la varianza $\frac{1}{N^2}\widehat{V}(\widehat{N\overline{y}})$.

(b) Supongamos que la variable y toma valores 0 ó 1 y se desea estimar la proporción p de valores 1 en la población.

(b1) Un estimador insesgado de esta proporción p es $\widehat{p} = \frac{1}{Nn}\sum_{i=1}^{n}\frac{N_i\widehat{p}_i}{p_i}$ donde $\widehat{p}_i$ es la proporción muestral en el conglomerado i.

(b2) La varianza de $\widehat{p}$ es, en caso de m.a.s. en segunda etapa,

$$V(\widehat{p}) = \frac{1}{N^2n}\left(\sum_{i=1}^{L}\frac{(N_i\widehat{p}_i)^2}{p_i} - (N\widehat{p})^2\right) + \frac{1}{N^2n}\sum_{i=1}^{L}\frac{N_i-m_i}{N_i}\frac{N_i^2}{p_im_i}\frac{N_i}{N_i-1}p_i(1-p_i).$$

(b3) La varianza de $\widehat{p}$ es, en caso de m.a.s.r. en segunda etapa,

$$V(\widehat{p}) = \frac{1}{N^2n}\left(\sum_{i=1}^{L}\frac{(N_i\widehat{p}_i)^2}{p_i} - (N\widehat{p})^2\right) + \frac{1}{N^2n}\sum_{i=1}^{L}\frac{N_i^2}{p_im_i}p_i(1-p_i).$$

(b4) Un estimador insesgado de $V(\widehat{p})$ es

$$V(\widehat{p}) = \frac{1}{N^2n(n-1)}\sum_{i=1}^{n}(\frac{N_i\widehat{p}_i}{p_i} - N\widehat{p})^2.$$

Veamos a continuación algunos casos particulares de interés , según los valores de las probabilidades p_i.

Corolario 10.4 (casos particulares para p_i).

(a) Si el muestreo en primera etapa es con reemplazamiento y **probabilidades iguales** $p_i = \frac{1}{L}$**,** es decir, diseño **m.a.s.r.**, entonces,

(a1) Un estimador insesgado del total es $\widehat{N\overline{y}} = \frac{L}{n}\sum_{i=1}^{n}N_i\widehat{\overline{y}}_i$. Si el muestreo en segunda etapa es m.a.s., la varianza de $\widehat{N\overline{y}}$ es

$$V(\widehat{N\overline{y}}) = \frac{1}{n}\left(L\sum_{i=1}^{L}y_i^2 - (N\overline{y})^2\right) + \frac{L}{n}\sum_{i=1}^{L}\frac{N_i-m_i}{N_i}\frac{N_i^2}{m_i}S_{2i}^2.$$

(a2) Si el muestreo en segunda etapa es m.a.s. o m.a.s.r., o $\widehat{\overline{y}}_i$ es un estimador insesgado de $\overline{y}_i$, un estimador de la varianza $V(\widehat{N\overline{y}})$ es

$$\widehat{V}(\widehat{N\overline{y}}) = \frac{L^2}{n(n-1)}\sum_{i=1}^{n}(N_i\widehat{\overline{y}}_i - \frac{\widehat{N\overline{y}}}{L})^2.$$

ISBN 9788496866133

(b) Si el muestreo en primera etapa es con reemplazamiento y **probabilidades proporcionales al tamaño** del conglomerado, es decir $p_i = \dfrac{N_i}{N}$,

(b1) Un estimador insesgado del total es $\widehat{N\overline{y}} = \dfrac{N}{n}\sum_{i=1}^{n}\widehat{\overline{y}}_i$. Si el muestreo en segunda etapa es m.a.s., la varianza de $\widehat{N\overline{y}}$ es

$$V(\widehat{N\overline{y}}) = \frac{N}{n}\left(\sum_{i=1}^{L}\frac{{y_i}^2}{N_i} - N\overline{y}^2\right) + \frac{N}{n}\sum_{i=1}^{L}\frac{N_i - m_i}{N_i^2}\frac{N_i^2}{m_i}S_{2i}^2.$$

(b2) Si el muestreo en segunda etapa es m.a.s. o m.a.s.r., o $\widehat{\overline{y}}_i$ es un estimador insesgado de $\overline{y}_i$, un estimador de la varianza $V(\widehat{N\overline{y}})$ es

$$\widehat{V}(\widehat{N\overline{y}}) = \frac{N^2}{n(n-1)}\sum_{i=1}^{n}(\widehat{\overline{y}}_i - \frac{\widehat{N\overline{y}}}{N})^2 = \frac{N^2}{n(n-1)}\sum_{i=1}^{n}(\widehat{\overline{y}}_i - \widehat{\overline{y}})^2.$$

Ejemplo 10.4

Continuando con el ejemplo 10.3, se aborda el problema de muestreo a través de la asignación de probabilidades proporcionales a la población total de cada provincia, para luego escoger una m.a.s. de municipios dentro de cada provincia.

Se escogerán $n = 5$ provincias por este método, fijando la fracción de muestreo dentro de cada provincia en $f_{2i} = \dfrac{m_i}{N_i} = 0.1$ como en el ejemplo 10.3. Los resultados se presentan a continuación.

Provincia	p_i	N_i	m_i	$\widehat{\overline{y}}_i$	s_{2i}^2
Madrid	0.124	179	18	6722.22	401646298
Barcelona	0.115	310	31	3780.03	42056589.30
Valencia	0.054	265	26	2450.42	37896783.45
Valencia	0.054	265	26	3108.92	43310476.55
Valencia	0.054	265	26	3789.12	59520126.99

Tabla 10.3. Características muestrales

Se observa que la provincia Valencia es escogida 3 veces (en cada una de ellas se extrae una m.a.s. diferente de municipios).

El estimador del total es

$$\widehat{N\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{p_i} = \frac{1}{5}\left[\frac{179\cdot 6722}{0.124} + ... + \frac{265\cdot 3789}{0.054}\right] = 13.154.039.$$

La varianza estimada de este estimador es

ISBN 9788496866133

$$\widehat{V}(\widehat{N\overline{y}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\frac{N_i \widehat{\overline{y}}_i}{p_i} - \widehat{N\overline{y}})^2 = 2.16 \cdot 10^{14}.$$

En comparación con el ejemplo 10.3, podría parecer que la varianza es mucho mayor (al menos la estimación de la varianza es mayor). Sin embargo, el método de los conglomerados últimos para la estimación de la varianza es conveniente, en general, cuando n es suficientemente grande, pues aunque es un estimador insesgado de la varianza puede no ser preciso para muestras pequeñas de conglomerados.

En nuestro caso, utilizar probabilidades proporcionales al tamaño probablemente mejora al estimador pues el coeficiente de correlación muestral entre p_i y el total estimado por conglomerado $N_i\widehat{\overline{y}}_i$ es 0.84. Además, en este ejemplo conocemos el valor real del total poblacional, 19.488.465, más cerca de los 13 millones de este ejemplo que de los 31 millones obtenidos en el ejemplo 10.3.

Es otro ejemplo más de que para comparar la precisión de dos estimadores o tipos de muestreo hay que tener cuidado al comparar sus varianzas estimadas, pues si estas estimaciones son imprecisas pueden llevar a conclusiones erróneas. Por ello es mejor decidirse por los estimadores o técnicas de muestreo basándose en las ideas teóricas de carácter general, en lugar de dejar al resultado empírico (es decir, estimación de la varianza) decidir por nosotros.

10.3.4 Estimación en muestreo con probabilidades desiguales y sin reemplazamiento en primera etapa y m.a.s. en segunda etapa

En general el muestreo sin reemplazamiento con probabilidades desiguales tiene menor varianza que el muestreo con reemplazamiento del mismo tipo, pero las estimaciones de varianzas suelen ser complicadas, y hará falta por lo general calcular las probabilidades de inclusión. Una posibilidad es utilizar las fórmulas de muestreo con reemplazamiento para estimar las varianzas, pues siempre serán una buena aproximación conservadora de la precisión del estudio.

En todo caso, a continuación se estudiarán los estimadores clásicos para el caso de muestreo sin reemplazamiento con probabilidades desiguales.

Teorema 10.9 (estimación del total y varianzas).

Supongamos muestreo sin reemplazamiento con probabilidades desiguales en primera etapa, y m.a.s. en segunda etapa.

(a) Un estimador insesgado para el total poblacional es

$$\widehat{N\overline{y}} = \sum_{i=1}^{n} \frac{N_i \widehat{\overline{y}}_i}{\pi_i}$$

(b) La varianza de $\widehat{N\overline{y}}$ es

$$V(\widehat{N\overline{y}}) = \sum_{i=1}^{L} \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_{i \neq j}^{L} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j + \sum_{i=1}^{L} \frac{N_i^2}{\pi_i} (1 - f_{2i}) \frac{S_{2i}^2}{m_i}.$$

(c) Un estimador insesgado de $V(\widehat{N\overline{y}})$ es

ISBN 9788496866133

$$\widehat{V}(\widehat{N\overline{y}}) = \sum_{i=1}^{n} \frac{(N_i\widehat{\overline{y}}_i)^2}{\pi_i^2}(1-\pi_i) + \sum_{i\neq j}^{n} \frac{N_i\widehat{\overline{y}}_i N_j\widehat{\overline{y}}_j}{\pi_i\pi_j}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}} + \sum_{i=1}^{n}\frac{N_i^2}{\pi_i}(1-f_{2i})\frac{s_{2i}^2}{m_i}.$$

Demostración.

(a) $E\left[\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{\pi_i}\right] = E_1\left[\sum_{i=1}^{n}\frac{N_iE_2^i(\widehat{\overline{y}}_i)}{\pi_i}\right] = E_1\left[\sum_{i=1}^{n}\frac{N_i\overline{y}_i}{\pi_i}\right] = N\overline{y}$ por ser insesgado el estimador de Horvitz-Thompson.

(b) Se aplicará el Teorema de Madow:

$V_1[E_2^i(\widehat{N\overline{y}})] = V_1\left[\sum_{i=1}^{n}\frac{N_i\overline{y}_i}{\pi_i}\right] = \sum_{i=1}^{L}\frac{1-\pi_i}{\pi_i}y_i^2 + \sum_{,i\neq j}^{L}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}y_iy_j$. Además,

$$E_1[V_2^i(\widehat{N\overline{y}})] = E_1\left[V_2^i\left(\sum_{i=1}^{n}\frac{N_i\overline{y}_i}{\pi_i}\right)\right] = E_1\left[\left(\sum_{i=1}^{n}\frac{N_i^2V_2^i(\overline{y}_i)}{\pi_i^2}\right)\right] =$$

$$= E_1\left[\sum_{i=1}^{n}\frac{N_i^2}{\pi_i^2}(1-f_{2i})\frac{s_{2i}^2}{m_i}\right].$$

Definiendo la variable aleatoria $e_i = \begin{cases} 1 \text{ si el conglomerado } i \text{ está en la muestra} \\ 0 \text{ si el conglomerado } i \text{ no está en la muestra} \end{cases}$

que tiene esperanza $E(e_i) = \pi_i$, se tiene que

$$E_1\left[\sum_{i=1}^{n}\frac{N_i^2}{\pi_i^2}(1-f_{2i})\frac{s_{2i}^2}{m_i}\right] = E_1\left[\sum_{i=1}^{L}\frac{N_i^2}{\pi_i^2}(1-f_{2i})\frac{s_{2i}^2}{m_i}e_i\right] =$$

$$= \sum_{i=1}^{L}\frac{N_i^2}{\pi_i^2}(1-f_{2i})\frac{s_{2i}^2}{m_i}E(e_i) = \sum_{i=1}^{L}\frac{N_i^2}{\pi_i}(1-f_{2i})\frac{s_{2i}^2}{m_i}.$$

Uniendo $V_1[E_2^i(\widehat{N\overline{y}})] + E_1[V_2^i(\widehat{N\overline{y}})]$ se tiene el resultado (b) del Teorema .

(c) Se procederá por partes, aplicando la esperanza en cada uno de los tres sumandos:

$$E_1E_2^i\left[\sum_{i=1}^{n}\frac{(N_i\widehat{\overline{y}}_i)^2}{\pi_i^2}(1-\pi_i)\right] = E_1\left[\sum_{i=1}^{n}\frac{N_i^2E_2^i((\widehat{\overline{y}}_i)^2)}{\pi_i^2}(1-\pi_i)\right] =$$

$$= E_1\left[\sum_{i=1}^{n}\frac{N_i^2\left[V_2^i(\widehat{\overline{y}}_i) + (E_2^i(\widehat{\overline{y}}_i))^2\right]}{\pi_i^2}(1-\pi_i)\right] =$$

$$= E_1\left[\sum_{i=1}^{n}\frac{N_i^2\left[V_2^i(\widehat{\overline{y}}_i) + (E_2^i(\widehat{\overline{y}}_i))^2\right]}{\pi_i^2}(1-\pi_i)\right] =$$

$$= E_1\left[\sum_{i=1}^{n}\frac{N_i^2V_2^i(\widehat{\overline{y}}_i)}{\pi_i^2}(1-\pi_i)\right] + E_1\left[\sum_{i=1}^{n}\frac{y_i^2}{\pi_i^2}(1-\pi_i)\right] =$$

ISBN 9788496866133

$$= \sum_{i=1}^{L} \frac{N_i^2}{\pi_i}(1-\pi_i)(1-f_{2i})\frac{s_{2i}^2}{m_i} + \sum_{i=1}^{L} \frac{1-\pi_i}{\pi_i} y_i^2,$$

donde se ha utilizado en la última igualdad la variable auxiliar e_i, en ambos sumandos.

Por otra parte,

$$E_1 E_2^i \left[\sum_{i \neq j}^{n} \frac{N_i \widehat{\overline{y}}_i N_j \widehat{\overline{y}}_j}{\pi_i \pi_j} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \right] = E_1 \left[\sum_{i \neq j}^{n} \frac{y_i y_j}{\pi_i \pi_j} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \right] = \sum_{i \neq j}^{L} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j$$

donde en la última igualdad se ha utilizado la variable $e_i = 1$ si i y j están en la muestra, y 0 si no. Se verifica que $E(e_i) = \pi_{ij}$.

Por último,

$$E_1 E_2^i \left[\sum_{i=1}^{n} \frac{N_i^2}{\pi_i}(1-f_{2i})\frac{s_{2i}^2}{m_i} \right] = E_1 \left[\sum_{i=1}^{n} \frac{N_i^2}{\pi_i}(1-f_{2i})\frac{S_{2i}^2}{m_i} \right] = \sum_{i=1}^{L} N_i^2 (1-f_{2i})\frac{S_{2i}^2}{m_i}.$$

Sumando los tres términos, se tiene que

$$E(\widehat{V}(\widehat{N\overline{y}})) =$$

$$= \sum_{i=1}^{L} \frac{N_i^2}{\pi_i}(1-\pi_i)(1-f_{2i})\frac{s_{2i}^2}{m_i} + \sum_{i=1}^{L} \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_{i \neq j}^{L} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j + \sum_{i=1}^{L} N_i^2 (1-f_{2i})\frac{S_{2i}^2}{m_i} =$$

$$= \sum_{i=1}^{L} \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_{i \neq j}^{L} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j + \sum_{i=1}^{L} \frac{N_i^2}{\pi_i}(1-f_{2i})\frac{S_{2i}^2}{m_i} = V(\widehat{N\overline{y}}).$$

Ejemplo 10.5

Un investigador desea obtener una estimación del salario medio de pacientes ingresados por hospital debido a cierta enfermedad, en los hospitales de una ciudad. Para ello decide obtener una muestra de $n = 3$ hospitales de los $L = 20$ existentes, sin reemplazamiento, con probabilidades iniciales de selección p_i proporcionales al número de camas total en cada hospital. Se sabe que hay en los hospitales de toda la ciudad, aproximadamente $N = 320$ enfermos de esa enfermedad. Dentro de cada hospital de los muestreados, se selecciona por m.a.s. un 20% de los pacientes con dicha enfermedad , y se les pregunta el salario mensual aproximado. Los datos obtenidos, junto con las probabilidades de inclusión , son los siguientes:

							π_{ij}		
Hospital	nº camas	π_i	N_i	m_i	$\widehat{\overline{y}}_i$	s_{2i}^2	1	2	3
1	102	0.13	15	3	826	97344	0	0.019	0.023
2	110	0.14	11	2	1236	65536	0.019	0	0.027
3	126	0.16	15	3	1006	88804	0.023	0.027	0

Tabla 10.4. Características muestrales y probabilidades de inclusión

ISBN 9788496866133

En este ejemplo, $f_{2i} = 0.20$. El estimador insesgado del salario medio de los pacientes ingresados por la enfermedad en cuestión es

$$\frac{1}{N}\widehat{N\overline{y}} = \frac{1}{N}\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{\pi_i} = \frac{1}{320}(\frac{15\cdot 286}{0.13} + \frac{11\cdot 1236}{0.14} + \frac{15\cdot 1006}{0.16}) = 701.33.$$

La varianza estimada de este estimador se calcula a través de los tres sumandos:

$$\frac{1}{N^2}\sum_{i=1}^{n}\frac{(N_i\widehat{\overline{y}}_i)^2}{\pi_i^2}(1-\pi_i) = 161425$$

$$\frac{1}{N^2}\sum_{i\neq j}^{n}\frac{N_i\widehat{\overline{y}}_i N_j\widehat{\overline{y}}_j}{\pi_i\pi_j}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}} = 2(1317 + 2907.2 + 15238.6) = 38925.7$$

$$\frac{1}{N^2}\sum_{i=1}^{n}\frac{N_i^2}{\pi_i}(1-f_{2i})\frac{s_{2i}^2}{m_i} = 985.2$$

Así, la varianza del estimador de la media queda

$$\frac{1}{N^2}\widehat{V}(\widehat{N\overline{y}})) = 161425 + 38925.7 + 985.2 = 201335.9.$$

10.3.5 Muestras autoponderadas

Se verá a continuación un concepto de particular interés en muestreo bietápico, pues la complejidad de este tipo de diseño, que viene con frecuencia añadido a una estratificación previa, exige la posibilidad de simplificaciones que den lugar a estimadores correctos pero a la vez sencillos.

Definición.

Muestras autoponderadas. Una muestra genérica obtenida por un proceso de muestreo concreto se denomina **muestra autoponderada** con respecto a un estimador $\widehat{\theta}$ si $\widehat{\theta}$ es una función de la suma sobre las unidades elementales , es decir,$\widehat{\theta} = K\sum_{i=1}\sum_{j=1}...y_{ij...}$. El factor K se denomina factor de extrapolación. Hay que remarcar que las muestras autoponderadas para un cierto estimador $\widehat{\theta}$ no tienen por qué serlo si se utiliza otro estimador $\widehat{\theta}'$. También se suele definir la **estimación autoponderada** como aquella referida a un diseño de muestreo que da lugar a muestras autoponderadas respecto a un estimador concreto $\widehat{\theta}$.

En la mayor parte de los diseños, si el estimador lineal $\widehat{\theta}$ es insesgado, el hecho de las muestras sean autoponderadas respecto a $\widehat{\theta}$ equivale a que todas las unidades de última etapa tienen la misma probabilidad de pertenecer a la muestra. La razón es que en los estimadores lineales del total del tipo

$$\sum_{i=1}\sum_{j=1}...w_{ij..}y_{ij...}$$

los pesos $w_{ij...}$ representan el inverso de la probabilidad de que la observación correspondiente $y_{ij...}$ pertenezca a la muestra. Si los pesos son constantes, ello significa que todas las unidades tienen la misma probabilidad de ser escogidas.

ISBN 9788496866133

Por ello a menudo existe una definición alternativa de muestra autoponderada que aparece en algunos textos, como las muestras provenientes de un diseño que asigna igual probabilidad a las unidades de última etapa.

La construcción de un diseño de muestreo que da lugar a muestras autoponderadas es muy importante desde varios puntos de vista:

1. El cálculo de los estimadores es sencillo, pues son funciones de la suma de las observaciones muestrales. Para diseños complejos esto tiene cierta importancia, pues evita errores de concepto y construcción.

2. Cuando en un diseño de muestreo todas las unidades elementales tienen la misma probabilidad de pertenecer a la muestra, cada una de ellas es representativa de la población en el mismo modo. Estadísticos básicos como media, mediana, histogramas y percentiles sobre la muestra estiman las cantidades correspondientes de la población. Además, las muestras autoponderadas con frecuencia tienen una varianza menor y las estadísticas sobre la muestra son más robustas.

Sin embargo, a menudo , aunque se trate de muestras autoponderadas, éstas pueden no provenir de un diseño sencillo m.a.s. y por lo tanto no son muestras aleatorias simples (recuérdese el caso particular de muestreo con probabilidades desiguales). Por lo tanto la estimación de los errores de muestreo (varianza del estimador) no se simplifica, y hay que recurrir igual a las fórmulas de varianza para cada caso. Esto hace que técnicas como intervalos de confianza o contrastes de hipótesis aplicadas de manera estándar sobre las muestras autoponderadas, sean incorrectas (si bien es cierto que también lo serían, bajo el mismo tipo de diseño, en caso de muestras no autoponderadas).

Casos particulares

Se asume muestreo estratificado con m.a.s. en cada estrato, donde los tamaños de los estratos N_h son distintos.

(a) Supongamos muestreo estratificado con m.a.s. en cada estrato y afijación proporcional, es decir, $n_h = n\dfrac{N_h}{N}$. Un estimador insesgado de la media es

$\overline{y}_{st} = \sum_{h=1}^{L} \dfrac{N_h}{N}\widehat{\overline{y}}_h = \sum_{h=1}^{L} \dfrac{N_h}{N}\dfrac{1}{n_h}\sum_{j=1}^{n_h} y_{hj} = \sum_{h=1}^{L} \dfrac{N_h}{N}\dfrac{1}{n}\dfrac{N}{N_h}\sum_{j=1}^{n_h} y_{hj} = \dfrac{1}{n}\sum_{h=1}^{L}\sum_{j=1}^{n_h} y_{hj}$ es función de la suma sobre las unidades elementales y por lo tanto las muestras son autoponderadas.

También se puede remarcar que en este caso, la probabilidad de seleccionar cada unidad hj de la población es igual. Veamos: en el estrato h se seleccionan n_h unidades. Como se trata de m.a.s., cada unidad del estrato h es seleccionada con probabilidad $\dfrac{n_h}{N_h} = \dfrac{n}{N}$ =constante, igual para todo h.

Por otra parte, se observa que el estimador del total es

ISBN 9788496866133

$\overline{y}_{st} = \sum_{h=1}^{L} \sum_{j=1}^{n_h} \frac{N}{n} y_{hj}$, es decir, se trata de una suma de los valores muestrales y_{hj} ponderada por el peso $w_{hj} = \frac{N}{n}$, que es el inverso de la probabilidad de que cada observación sea seleccionada.

(b) Supongamos muestreo estratificado con afijación igual, es decir, $n_h = \frac{n}{L}$. Entonces

$\overline{y}_{st} = \sum_{h=1}^{L} \frac{N_h}{N} \frac{1}{n_h} \sum_{j=1}^{n_h} y_{hj} = \sum_{h=1}^{L} \frac{N_h}{N} \frac{L}{n} \sum_{j=1}^{n_h} y_{hj} = \frac{L}{nN} \sum_{h=1}^{L} N_h \sum_{j=1}^{n_h} y_{hj}$ no es suma directa sobre las unidades elementales. Por lo tanto las muestras no son autoponderadas (=la estimación no es autoponderada).

Se observa que las unidades poblacionales tienen distinta probabilidad de pertenecer a la muestra: las que pertenecen al conglomerado h tienen probabilidad $\frac{n_h}{N_h} = \frac{n}{LN_h}$ de pertenecer a la muestra. Cada una de las observaciones en los estratos con N_h más grande tiene menos probabilidad de pertenecer a la muestra que aquellas pertenecientes a estratos con tamaño N'_h más pequeño, pues $\frac{n}{LN_h} < \frac{n}{LN'_h}$.

Se observa que el estimador del total es

$N\overline{y}_{st} = \sum_{h=1}^{L} \sum_{j=1}^{n_h} w_{hj} y_{hj}$, donde $w_{hj} = \frac{N_h L}{n}$ es el inverso de la probabilidad de que la unidad hj sea seleccionada.

El siguiente resultado presenta el tipo de afijaciones que dan lugar a muestras autoponderadas en muestreo por conglomerados.

Teorema 10.11 (muestras autoponderadas en muestreo por conglomerados).

(a) En muestreo por conglomerados monoetápico,

(a1) Cuando los tamaños son iguales, bajo m.a.s., $\overline{\overline{y}}$ produce estimaciones autoponderadas.

(a2) Con los tamaños son desiguales, bajo m.a.s., $\widehat{\overline{y}}$ produce estimaciones autoponderadas.

(a3) Con tamaños desiguales, bajo muestreo pptr, t_{HH} da estimaciones autoponderadas si $p_i =$ constante$=\frac{1}{L}$ (m.a.s.r. de los conglomerados).

(a4) Con tamaños desiguales, bajo muestreo ppt, t_{HT} da estimaciones autoponderadas si $\pi_i =$ constante (m.a.s. de los conglomerados).

(b) En muestreo por conglomerados bietápico,

(b1) Cuando los tamaños son iguales, bajo m.a.s. o m.a.s.r. en primera etapa, $\overline{\overline{y}}$ produce estimaciones autoponderadas si $m =$ constante.

(b2) Con tamaños son desiguales, bajo m.a.s. en ambas etapas, el estimador insesgado $\widehat{N\overline{y}}$ produce estimaciones autoponderadas si $\frac{N_i}{m_i} =$ constante.

ISBN 9788496866133

(b3) Con tamaños desiguales, bajo muestreo pptr en primera etapa, el estimador $\widehat{N\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{p_i}$ da estimaciones autoponderadas si $\frac{N_i}{p_i m_i}$ =constante.

(b4) Con tamaños desiguales, bajo muestreo ppt, el estimador $\widehat{N\overline{y}} = \sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{\pi_i}$da estimaciones autoponderadas si $\frac{N_i}{\pi_i m_i}$ =constante.

Demostración.

(a)

(a1) $\overline{\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\overline{N}}\sum_{j=1}^{\overline{N}} y_{ij} = \frac{1}{n\overline{N}}\sum_{i=1}^{n}\sum_{j=1}^{\overline{N}} y_{ij}$ así que es función de la suma sobre las unidades elementales.

(a2) $\widehat{\overline{y}} = \frac{1}{n\overline{N}}\sum_{i=1}^{n} N_i\frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij} = \frac{1}{n\overline{N}}\sum_{i=1}^{n}\sum_{j=1}^{\overline{N}} y_{ij}$.

(a3) $t_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}\sum_{j=1}^{N_i} y_{ij}$. Así que si $\frac{1}{p_i}$ = constante, t_{HH} da estimaciones autoponderadas. Para que p_i sea constante, ha de ser $p_i = \frac{1}{L}$. (Es el caso de m.a.s.r. de conglomerados).

(a4) $t_{HT} = \sum_{i=1}^{n}\frac{y_i}{\pi_i} = \sum_{i=1}^{n}\frac{1}{\pi_i}\sum_{j=1}^{N_i} y_{ij}$. Si π_i = constante= $\frac{n}{L}$ la estimación es autoponderada (es el caso de m.a.s. de conglomerados).

(b)

(b1) $\overline{\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$ es función de la suma si m_i es constante.

(b2) $\widehat{N\overline{y}} = \frac{L}{n}\sum_{i=1}^{n} N_i\widehat{\overline{y}}_i = \frac{L}{n}\sum_{i=1}^{n}\frac{N_i}{m_i}\sum_{j=1}^{m_i} y_{ij}$ es función de la suma si $\frac{N_i}{m_i}$ es constante.

(b3) $\widehat{N\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{p_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i}{p_i m_i}\sum_{j=1}^{m_i} y_{ij}$ es función de la suma si $\frac{N_i}{m_i p_i}$ es constante.

(b4) $\widehat{N\overline{y}} = \sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{\pi_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i}{\pi_i m_i}\sum_{j=1}^{m_i} y_{ij}$ es función de la suma si $\frac{N_i}{m_i\pi_i}$ es constante.

10.3.6 Correcciones para muestras no autoponderadas

Como se ha comentado, el hecho de que una muestra sea autoponderada simplifica en gran medida los cálculos. La mayoría de los programas comerciales o de investigación sobre estimación en encuestas por muestreo se basan en la utilización de los pesos muestrales para la realización

ISBN 9788496866133

de estimaciones. Los pesos en el estimador del total son iguales al inverso de la probabilidad de inclusión de la observación. El aportar estos pesos con cada observación evita dar mayor información sobre el proceso de muestreo realizado, pues en los pesos está prácticamente toda la información necesaria para la estimación.

Si la muestra es autoponderada, los pesos son constantes y se simplifican los cálculos, evitando errores en el proceso de tabulación. Sin embargo, en la práctica ocurre frecuentemente que la muestra no es autoponderada, muchas veces porque motivos prácticos impiden un muestreo con las afijaciones adecuadas.

Como solución simplificadora, pueden utilizarse técnicas en el proceso de tratamiento de datos sobre muestras que no son en origen autoponderadas, para poder "corregir" éstas y adoptar posteriormente los métodos estándar de estimación para muestras autoponderadas. Obviamente estos métodos empeoran la calidad de la estimación, bien en sesgo o en varianza, pero a cambio de obtener la sencillez de tratamiento de los datos necesaria en muchos casos.

Algunas posibilidades son:

1. Reemplazar los pesos de cada observación por la media de los pesos calculada sobre toda la muestra. Esto es equivalente a considerar la muestra como autoponderada aunque no lo sea. Esto da lugar a estimadores sesgados. Si la covarianza entre los pesos y los valores muestrales de la variable de interés es pequeña, este sesgo suele ser pequeño.

2. Redondeo de los pesos a múltiplos de 10, 100, etc. a conveniencia, de modo que el redondeo lleve a un valor aproximadamente constante de los pesos. Este método también origina un sesgo, menos controlable que el anterior.

3. Extraer una submuestra con reemplazamiento de tamaño n' $(< n)$ de la muestra original y con probabilidades proporcionales a los pesos originales. Entonces, si el valor obtenido de la variable de interés en cada observación de la submuestra es y_k, se tiene que

$$(N\widehat{\overline{y}})' = \frac{1}{n'} \sum_{j=1}^{n'} \left(\sum w_{i\ldots} \right) y_k$$

es un estimador insesgado del total, siendo $\sum w_{i\ldots}$ la suma de los pesos originales sobre toda la muestra, y un estimador insesgado de la varianza de $(N\widehat{\overline{y}})'$ es

$$\widehat{V}((N\widehat{\overline{y}})') = \frac{1}{n'(n'-1)} \sum_{j=1}^{n'} \left((N\widehat{\overline{y}})' - \left(\sum w_{i\ldots} \right) y_k \right)^2 .$$

Este método es insesgado, pero se aumenta la varianza de la estimación respecto a la estimación habitual.

10.4 Tablas de fórmulas

En las fórmulas se tendrán en cuenta las expresiones siguientes de las varianzas poblacionales y muestrales, en el caso en que el objetivo sea estimar la media o la proporción.

ISBN 9788496866133

Varianzas en estimación de la media o total	**Varianzas en estimación de la proporción**
$S_1^2 = \frac{1}{L-1}\sum_{i=1}^{L}(\overline{y}_i - \overline{y})^2$	$S_1^2 = \frac{1}{L-1}\sum_{i=1}^{L}(p_i - p)^2$
$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{\overline{y}}_i - \widehat{\overline{y}})^2$	$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{p}_i - \widehat{\widehat{p}})^2$
$S_2^2 = \frac{1}{L(\overline{N}-1)}\sum_{i=1}^{L}\sum_{j=1}^{\overline{N}}(y_{ij} - \overline{y}_i)^2$	$S_2^2 = \frac{1}{L}\sum_{i=1}^{L}\frac{\overline{N}}{\overline{N}-1}p_i(1-p_i)$
$s_2^2 = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \widehat{\overline{y}}_i)^2$	$s_2^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{m}{m-1}\widehat{p}_i(1-\widehat{p}_i)$
$S_{2i}^2 = \frac{1}{(\overline{N}-1)}\sum_{j=1}^{\overline{N}}(y_{ij} - \overline{y}_i)^2$	$S_{2i}^2 = \frac{\overline{N}}{\overline{N}-1}p_i(1-p_i)$
$s_{2i}^2 = \frac{1}{(m-1)}\sum_{j=1}^{m}(y_{ij} - \widehat{\overline{y}}_i)^2$	$s_{2i}^2 = \frac{m}{m-1}\widehat{p}_i(1-\widehat{p}_i)$

TAMAÑOS IGUALES, M.A.S. EN AMBAS ETAPAS

Parámetro poblacional	$\overline{y}$	$N\overline{y}$	p
Estimador	$\widehat{\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\overline{y}}_i$	$N\widehat{\overline{y}}$	$\widehat{\widehat{p}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{p}_i$
Varianza	$(1-f_1)\frac{S_1^2}{n} + (1-f_2)\frac{S_2^2}{nm}$	$N^2V(\widehat{\overline{y}})$	$V(\widehat{\widehat{p}})$
Estimador de la Varianza	$(1-f_1)\frac{s_1^2}{n} + (1-f_2)\frac{s_2^2}{mL}$	$N^2\widehat{V}(\widehat{\overline{y}})$	$\widehat{V}(\widehat{\widehat{p}})$

TAMAÑOS IGUALES, m.a.s. EN PRIMERA ETAPA, M.A.S.R. EN SEGUNDA ETAPA

Parámetro poblacional	$\overline{y}$	$N\overline{y}$	p
Estimador	$\widehat{\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\overline{y}}_i$	$N\widehat{\overline{y}}$	$\widehat{\widehat{p}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{p}_i$
Varianza	$(1-f_1)\frac{S_1^2}{n} + \frac{\sigma_2^2}{nm}$	$N^2V(\widehat{\overline{y}})$	$V(\widehat{\widehat{p}})$
Estimador de la Varianza	$(1-f_1)\frac{s_1^2}{n} + \frac{s_2^2}{mL}$	$N^2\widehat{V}(\widehat{\overline{y}})$	$\widehat{V}(\widehat{\widehat{p}})$

ISBN 9788496866133

TAMAÑOS DESIGUALES, m.a.s. EN AMBAS ETAPAS **ESTIMACIÓN INSESGADA**	
Parámetro poblacional	$N\overline{y}$
Estimador	$\widehat{N\overline{y}} = \frac{L}{n}\sum_{i=1}^{n} N_i\widehat{\overline{y}}_i$
Varianza	$\frac{L^2(1-f_1)}{n(L-1)}\sum_{i=1}^{L}(y_i - \overline{N}\overline{y})^2 + \frac{L}{n}\sum_{i=1}^{L}\frac{N_i^2(1-f_{2i})S_{2i}^2}{m_i}$
Estimador de la Varianza	$\frac{L^2(1-f_1)}{n(n-1)}\sum_{i=1}^{n}(N_i\widehat{\overline{y}}_i - \frac{1}{n}\sum_{i=1}^{n}N_i\widehat{\overline{y}}_i)^2 + \frac{L}{n}\sum_{i=1}^{n}\frac{N_i^2(1-f_{2i})s_{2i}^2}{m_i}$

Para la estimación de la media y proporción :

Parámetro poblacional	$\overline{y}$	p
Estimador	$\frac{\widehat{N\overline{y}}}{N}$	$\widehat{p} = \frac{L}{Nn}\sum_{i=1}^{n} N_i\widehat{p}_i$
Varianza	$\frac{V(\widehat{\overline{y}})}{N^2}$	$\frac{V(\widehat{p})}{N^2}$
Estimador de la Varianza	$\frac{\widehat{V}(\widehat{\overline{y}})}{N^2}$	$\frac{\widehat{V}(\widehat{p})}{N^2}$

ISBN 9788496866133

TAMAÑOS DESIGUALES, m.a.s. EN AMBAS ETAPAS **ESTIMACIÓN DE RAZÓN A TAMAÑO**	
Parámetro poblacional	$\overline{y}$
Estimador	$\widehat{\overline{y}}_R = \dfrac{\sum_{i=1}^{n} N_i \widehat{\overline{y}}_i}{\sum_{i=1}^{n} N_i}$
Varianza	$\dfrac{(1-f_1)}{\overline{N}^2 n} \sum_{i=1}^{L} \dfrac{N_i^2(\overline{y}_i - \overline{y})^2}{L-1} + \dfrac{1}{L\overline{N}^2 n} \sum_{i=1}^{L} \dfrac{N_i^2(1-f_{2i})S_{2i}^2}{m_i}$
Estimador de la Varianza	$\dfrac{(1-f_1)}{\overline{N}^2 n} \sum_{i=1}^{n} \dfrac{N_i^2(\widehat{\overline{y}}_i - \widehat{\overline{y}}_R)^2}{(n-1)} + \dfrac{1}{L\overline{N}^2 n} \sum_{i=1}^{n} \dfrac{N_i^2(1-f_{2i})s_{2i}^2}{m_i}$

En el último caso, para la estimación del total y proporción :

Parámetro poblacional	$N\overline{y}$	p
Estimador	$N\widehat{\overline{y}}_R$	$\widehat{\overline{y}}_R$
Varianza	$N^2 V(\widehat{\overline{y}}_R)$	$V(\widehat{\overline{y}}_R)$
Estimador de la Varianza	$N^2 \widehat{V}(\widehat{\overline{y}}_R)$	$\widehat{V}(\widehat{\overline{y}}_R)$

ISBN 9788496866133

TAMAÑOS DESIGUALES, PROBABILIDADES DESIGUALES Y REEMPLAZAMIENTO EN 1ª ETAPA, m.a.s. EN 2ª ETAPA	
Parámetro poblacional	$N\overline{y}$
Estimador	$\widehat{N\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{p_i}$
Varianza	$\frac{1}{n}\left(\sum_{i=1}^{L}\frac{y_i^2}{p_i} - (N\overline{y})^2\right) + \frac{1}{n}\sum_{i=1}^{L}\frac{N_i - m_i}{N_i}\frac{N_i^2}{p_i m_i}S_{2i}^2.$
Estimador de la Varianza	$\frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{N_i\widehat{\overline{y}}_i}{p_i} - \widehat{N\overline{y}})^2$

TAMAÑOS DESIGUALES, PROBABILIDADES DESIGUALES Y REEMPLAZAMIENTO EN 1ª ETAPA, m.a.s.r. EN 2ª ETAPA	
Parámetro poblacional	$N\overline{y}$
Estimador	$\widehat{N\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{p_i}$
Varianza	$V(\widehat{N\overline{y}}) = \frac{1}{n}\left(\sum_{i=1}^{L}\frac{y_i^2}{p_i} - (N\overline{y})^2\right) + \frac{1}{n}\sum_{i=1}^{L}\frac{N_i^2\sigma_{2i}^2}{p_i m_i}$
Estimador de la Varianza	$\frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{N_i\widehat{\overline{y}}_i}{p_i} - \widehat{N\overline{y}})^2$

ISBN 9788496866133

TAMAÑOS DESIGUALES, PROBABILIDADES DESIGUALES Y MUESTREO SIN REEMPLAZAMIENTO EN 1ª ETAPA, m.a.s. EN 2ª ETAPA	
Parámetro poblacional	$N\overline{y}$
Estimador	$\widehat{N\overline{y}} = \sum_{i=1}^{n} \frac{N_i \widehat{\overline{y}}_i}{\pi_i}$
Varianza	$V(\widehat{N\overline{y}}) = \sum_{i=1}^{L} \frac{1-\pi_i}{\pi_i} y_i^2 + 2\sum_{i<j}^{L} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j} y_i y_j + \sum_{i=1}^{L} \frac{N_i^2}{\pi_i}(1-f_{2i})\frac{S_{2i}^2}{m_i}$
Estimador de la Varianza	$\widehat{V}(\widehat{N\overline{y}}) = \sum_{i=1}^{n} \frac{(1-\pi_i)}{\pi_i^2}(N_i\widehat{\overline{y}}_i)^2 + 2\sum_{i<j}^{n} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}} \frac{N_i\widehat{\overline{y}}_i N_j\widehat{\overline{y}}_j}{\pi_i\pi_j} + + \sum_{i=1}^{n} \frac{N_i^2}{\pi_i}(1-f_{2i})\frac{s_{2i}^2}{m_i}$
Estimador de la Varianza (Y-G)	$\widehat{V}(\widehat{N\overline{y}}) = \sum_{i<j}^{n} \frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}} \left(\frac{N_i\widehat{\overline{y}}_i}{\pi_i} - \frac{N_j\widehat{\overline{y}}_j}{\pi_j}\right)^2 + + \sum_{i=1}^{n} \frac{N_i^2}{\pi_i}(1-f_{2i})\frac{s_{2i}^2}{m_i}$

En los últimos casos, para la estimación de la media y proporción se corrigen ligeramente las expresiones anteriores:

Parámetro poblacional	$\overline{y}$	p
Estimador	$\frac{\widehat{N\overline{y}}}{N}$	$\frac{\widehat{N\overline{y}}}{N}$
Varianza	$\frac{V(\widehat{\overline{y}})}{N^2}$	$\frac{V(\widehat{\overline{y}})}{N^2}$
Estimador de la Varianza	$\frac{\widehat{V}(\widehat{\overline{y}})}{N^2}$	$\frac{\widehat{V}(\widehat{\overline{y}})}{N^2}$

ISBN 9788496866133

10.5 Obtención de muestras por muestreo bietápico con SAS

10.5.1 Muestreo aleatorio simple de conglomerados, con m.a.s. en segunda etapa

Se utilizará en este caso la macro extrabimas, con sintaxis:

```
extrabimas(archivo1,archivo2,archivo3,codif,varconglo,nconglo,n,n2,
frac2,semilla1,semilla2);
```

donde

archivo1 es el archivo que contiene la información poblacional.

archivo2 es el archivo que contendrá la muestra final.

archivo3 es el archivo que contiene la muestra de conglomerados (todas las observaciones de cada conglomerado escogido).

codif es un archivo de salida que contiene la recodificación de los conglomerados.

varconglo es la variable que indica el conglomerado en el archivo poblacional

nconglo es el número de conglomerados diferentes.

n es el número de conglomerados muestreados (primera etapa: n).

n2 es el número de observaciones m a muestrear en cada conglomerado (afijación igual).

frac2 es la fracción de muestreo f_2 en segunda etapa (afijación proporcional) .

semilla1 semilla de aleatorización para obtener la muestra de conglomerados.

semilla2 semilla de aleatorización para obtener las muestras en segunda etapa.

Observaciones:

- Para poder ejecutar la macro extrabimas es necesario haber compilado previamente la macro extramono.

- Si se desea utilizar fracción de muestreo en segunda etapa $f_{2i} = f_2$ constante (variable frac2) es necesario poner missing (".") en el lugar de la variable n2.

- Es necesario que haya suficientes observaciones en cada conglomerado para que los tamaños n2 puedan ser alcanzados.

La macro extrabimas obtiene el archivo muestral en muestreo por conglomerados bietápico, con muestreo aleatorio simple en ambas etapas. El archivo muestral archivo2 contiene las observaciones seleccionadas de cada uno de los conglomerados seleccionados. El archivo muestral archivo3 contiene todas las observaciones de los conglomerados seleccionados.

Por ejemplo, si se dispone del archivo con la información poblacional data1 y la variable que indica los conglomerados se llama provincia, y se desea una muestra en el archivo muestra1, de 5

ISBN 9788496866133

conglomerados de los 55 existentes , y se fija la semilla en primera etapa en el número 4444444, y la semilla en segunda etapa 12345, se requiere muestrear el mismo número de observaciones en cada conglomerado escogido, $m = 4$, el archivo que contiene todas las observaciones de los conglomerados escogidos en primera etapa se llamará primera, y el archivo de codificación se llamará codif, la sintaxis es:

```
extrabimas(data1,muestra1,primera,codif,provincia,55,5,4,.,4444444,12345);
```

Si se desea el mismo programa pero muestreando un 20% de las observaciones dentro de cada conglomerado escogido, será:

```
extrabimas(data1,muestra1,primera,codif,provincia,55,5,.,0.20,4444444,12345);
```

10.5.2 Muestreo aleatorio simple de conglomerados, con m.a.s.r. en segunda etapa

Se utilizará la macro extrabimasr, con la misma sintaxis que la anterior:

```
extrabimasr(archivo1,archivo2,archivo3,codif,varconglo,nconglo,n,n2,
frac2,semilla1,semilla2);
```

10.5.3 Muestreo pptr o m.a.s.r. de conglomerados, con m.a.s. en segunda etapa

Se utilizará en este caso la macro extrabipptr, con sintaxis:

```
extrabipptr(archivo1,archivo2,archivo3,archivo4,codif,
varconglo,varx,nconglo,n,n2,frac2,masr,indicador,semilla1,semilla2);
```

donde

archivo1 es el archivo que contiene la información poblacional.

archivo2 es el archivo con la información de la variable auxiliar o pi por conglomerado(opcional).

archivo3 es el archivo de salida que contiene la muestra de conglomerados (todas las observaciones de cada conglomerado escogido).

archivo4 es el archivo de salida que contendrá la muestra final.

codif es un archivo de salida que contiene la recodificación de los conglomerados.

varconglo es la variable que indica el conglomerado en el archivo poblacional.

varx es la variable auxiliar.

nconglo es el número de conglomerados diferentes.

n es el número de conglomerados muestreados (primera etapa).

ISBN 9788496866133

n2 es el número de observaciones a muestrear en cada conglomerado (afijación igual).

frac2 es la fracción de muestreo en segunda etapa (afijación proporcional).

masr

1 Si se desea masr en primera etapa.

2 Si se desea pptr en primera etapa.

indicador

1 Si la variablex está presente en el archivo poblacional, constante por conglomerado.

2 Si la variablex está en el archivo2, que contiene además la variable varconglo.

semilla1 semilla de aleatorización para obtener la muestra de conglomerados.

semilla2 semilla de aleatorización para obtener las muestras en segunda etapa.

Observaciones:

- Para poder ejecutar la macro extrabipptr es necesario haber compilado previamente la macro extramonopptr.

- Si se desea utilizar fracción de muestreo en segunda etapa $f_{2i} = f_2$ constante (variable frac2) es necesario poner missing (".") en el lugar de la variable n2.

- Es necesario que haya suficientes observaciones en cada conglomerado para que los tamaños n2 puedan ser alcanzados.

- Si se selecciona la opción masr, no se puede seleccionar la opción indicador=2 (no es necesaria la variable varx).

- Si la fracción de muestreo en segunda etapa es muy pequeña, al menos se obtiene una observación por conglomerado.

La macro extrabipptr obtiene el archivo muestral en muestreo por conglomerados bietápico, con muestreo proporcional a la varx con reemplazamiento en primera etapa y m.a.s. en segunda etapa. El archivo muestral archivo4 contiene las observaciones seleccionadas de cada uno de los conglomerados seleccionados. El archivo3 contiene todas las observaciones de los conglomerados seleccionados.

Por ejemplo, si se dispone del archivo con la información poblacional data1 y la variable que indica los conglomerados se llama provincia, y se desea una muestra en el archivo muestra1, de 5 conglomerados de los 55 existentes , y la primera etapa se hará proporcional a la variable llamada pobla, y presente en el archivo data1, y se requiere muestrear el mismo número de observaciones en cada conglomerado escogido, $m = 4$, el archivo que contiene todas las observaciones de los conglomerados escogidos en primera etapa se llamará primera, y el archivo de codificación se llamará codif, la sintaxis es:

```
%extrabipptr(data1,.,primera,muestra1,codif,provincia,pobla,
55,5,4,.,2,1);
```


ISBN 9788496866133

10.5.4 Muestreo ppt de conglomerados, con m.a.s. en segunda etapa

Se utilizará en este caso la macro extrabippt, con sintaxis:

```
extrabippt(archivo1,archivo2,archivo3,archivo4,inclusion,codif,
varconglo,varx,nconglo,n,n2,frac2,indicador,semilla1,semilla2);
```

donde:

archivo1 es el archivo que contiene la información poblacional.

archivo2 es el archivo con la información de la variable auxiliar o pi por conglomerado(opcional).

archivo3 es el archivo que contiene la muestra de conglomerados (todas las observaciones de cada conglomerado escogido).

archivo4 es el archivo que contendrá la muestra final.

inclusion es el archivo de salida que contiene las probabilidades de inclusión para los conglomerados muestrales.

codif es un archivo de salida que contiene la recodificación de los conglomerados.

varconglo es la variable que indica el conglomerado en el archivo poblacional.

varx es la variable auxiliar.

nconglo es el número de conglomerados diferentes.

n es el número de conglomerados muestreados (primera etapa).

n2 es el número de observaciones a muestrear en cada conglomerado (afijación igual).

frac2 es la fracción de muestreo en segunda etapa (afijación proporcional).

indicador

1 Si la variablex está presente en el archivo poblacional, constante por conglomerado.

2 Si la variablex está en el archivo2, que contiene además la variable varconglo.

semilla1 semilla de aleatorización para obtener la muestra de conglomerados.

semilla2 semilla de aleatorización para obtener las muestras en segunda etapa.

Observaciones:

- Para poder ejecutar la macro extrabippt es necesario haber compilado previamente la macro extramonoppt.

- Es necesario que todos los conglomerados sean tales que $p_i < 1/n$.

ISBN 9788496866133

- Si se desea utilizar fracción de muestreo en segunda etapa $f_{2i} = f_2$ constante (variable frac2) es necesario poner missing (".") en el lugar de la variable n2.

- Es necesario que haya suficientes observaciones en cada conglomerado para que los tamaños n2 puedan ser alcanzados.

- Si la fracción de muestreo en segunda etapa es muy pequeña, al menos se obtiene una observación por conglomerado.

La macro extrabippt obtiene el archivo muestral en muestreo por conglomerados bietápico, con muestreo proporcional a la varx en primera etapa y sin reemplazamiento, y m.a.s. en segunda etapa. El archivo muestral archivo4 contiene las observaciones seleccionadas de cada uno de los conglomerados seleccionados. El archivo3 contiene todas las observaciones de los conglomerados seleccionados. El archivo de inclusión es importante, pues contiene las probabilidades de inclusión de los conglomerados muestreados, de utilidad para las posteriores estimaciones.

Por ejemplo, si se dispone del archivo con la información poblacional data1 y la variable que indica los conglomerados se llama provincia, y se desea una muestra en el archivo muestra1, de 5 conglomerados de los 55 existentes , y la primera etapa se hará proporcional a la variable llamada pobla, y presente en el archivo data1, y se requiere muestrear el mismo número de observaciones en cada conglomerado escogido, $m = 4$, el archivo que contiene todas las observaciones de los conglomerados escogidos en primera etapa se llamará primera, y el archivo de codificación se llamará codif, y el de inclusion inclu, la sintaxis es:

```
%extrabippt(data1,.,primera,muestra1,inclu,codif,provincia,
pobla,10,5,4,.,1);
```

10.6 Estimación en muestreo bietápico con SAS

10.6.1 Muestreo aleatorio simple de conglomerados, con m.a.s. o m.a.s.r. en segunda etapa

Se utiliza la macro estimbimas:

```
estimbimas(muestra,vary,varconglo,vartamacong,reemplazo,nconglo,
n,ngrande);
```

donde:

muestra es el archivo que contiene la muestra.

vary es la variable de interés.

varconglo es la variable que indica el conglomerado en el archivo muestra.

vartamacong es la variable que indica el tamaño de cada conglomerado en el archivo muestra.

reemplazo

ISBN 9788496866133

1 si es m.a.s.r. en segunda etapa

2 si es m.a.s. en segunda etapa

nconglo es el número de conglomerados diferentes.

n es el tamaño muestral (número de conglomerados muestreado).

ngrande es el tamaño poblacional N.

La macro estimbimas calcula estimadores, varianzas e intervalos de confianza de medias y totales en caso de muestreo por conglomerados bietápico con mas en ambas etapas,

- suponiendo tamaños iguales.

- suponiendo tamaños desiguales, por estimación insesgada.

- suponiendo tamaños desiguales, por estimación de razón a tamaño.

La macro indica en un mensaje si los tamaños son iguales, en cuyo caso no se utiliza la estimación de razón a tamaño, sino solamente la insesgada, o desiguales, en cuyo caso se presentan los dos tipos de estimación.

10.6.2 Muestreo pptr de conglomerados , con m.a.s. o m.a.s.r. en segunda etapa

Se utiliza la macro estimbipptr:

```
estimbipptr(muestra,archivo2,vary,varconglo,vartamacong,indicador,
nconglo,n,ngrande);
```

donde:

muestra es el archivo que contiene la muestra.

archivo2 archivo que contiene las probabilidades pi junto con la variable varconglo (opcional).

vary es la variable de interés .

varconglo es la variable que indica el conglomerado en el archivo muestra.

vartamacong es la variable que indica el tamaño de cada conglomerado en el archivo muestra.

indicador

1 Si pi está presente en el archivo poblacional, constante por conglomerado.

2 Si pi está en el archivo2, que contiene además la variable varconglo.

nconglo es el número de conglomerados diferentes .

n es el tamaño muestral (número de conglomerados muestreado).

ngrande es el tamaño poblacional N.

ISBN 9788496866133

La macro estimbipptr calcula estimadores, varianzas e intervalos de confianza de medias y totales en caso de muestreo por conglomerados bietápico con muestreo pptr en primera etapa, y mas o bien masr en segunda etapa.

10.6.3 Muestreo ppt de conglomerados , con m.a.s. en segunda etapa

Se utiliza la macro estimbipptr:

```
estimbippt(muestra,inclusion,vary,varconglo,vartamacong,nconglo,
n,ngrande);
```

donde:

muestra es el archivo que contiene la muestra

inclusion archivo que contiene las probabilidades de inclusion, la variable varconglo y la variable unit (ver especificaciones abajo)

vary es la variable de interés

varconglo es la variable que indica el conglomerado en el archivo muestra

vartamacong es la variable que indica el tamaño de cada conglomerado en el archivo muestra

nconglo es el número de conglomerados diferentes

n es el tamaño muestral (número de conglomerados muestreado)

ngrande es el tamaño poblacional N

La macro estimbippt calcula estimadores, varianzas e intervalos de confianza de medias y totales en caso de muestreo por conglomerados bietápico con muestreo ppt en primera etapa, y mas en segunda etapa.

Algunas cuestiones a tener en cuenta:

- El archivo inclusión tiene que tener ciertas características: debe tomar la forma similar a un archivo de salida del proc surveyselect con el método pps. es decir, debe contener las variables jtprob_1...jtprob_n, la variable selectionprob (prob de inclusión de primer orden), la variable de conglomerado y la variable unit (índice). Un ejemplo de este archivo es:

Obs	conglo	Unit	Selection Prob	JtProb_1	JtProb_2	JtProb_3
1	10	1	0.17641	0.000000	0.023385	0.050031
2	4	2	0.18811	0.023385	0.000000	0.053614
3	7	3	0.40247	0.050031	0.053614	0.000000

ISBN 9788496866133

donde la probabilidad de inclusión del conglomerado numerado como 4 y correspondiente a unit=2 es 0.188, y la probabilidad de inclusión conjunta de los conglomerados 10 y 7 (con índices unit=1 y unit=3, por lo tanto la probabilidad está en la fila 1 y columna 3) es 0.050031.

- Si el archivo de inclusión proviene de la macro extrabippt ya tiene la estructura que hace falta.

ISBN 9788496866133

10.7 Ejercicios resueltos

Ejercicio 9.1

Se desea estudiar la media de salarios anuales de los empleados en miles de euros en establecimientos de un banco que sólo tiene sucursales en dos ciudades . En la primera ciudad hay 15 sucursales, cada una con aproximadamente 20 empleados. Se seleccionan dos sucursales con reposición, y con probabilidades arbitrarias en función de la ubicación estratégica de cada sucursal (se supone que los salarios son más altos si la ubicación es mejor). En cada sucursal de las escogidas se escogen por m.a.s. 3 empleados. Los datos obtenidos en esta ciudad, donde $\widehat{\overline{y}}_i$ representa la media de los salarios de los empleados seleccionados en miles, son:

Sucursal	p_i	$\widehat{\overline{y}}_i$
1	0.10	17
2	0.05	15

En la segunda ciudad se escogen de las 25 sucursales que existen, 2 sucursales por m.a.s.. En cada sucursal se muestrean por m.a.s. , 2 empleados. Hay gran diferencia entre el número de empleados de cada sucursal, por lo que se desea utilizar un estimador de razón. Los datos obtenidos son:

Sucursal	N_i	$\widehat{\overline{y}}_i$	s_{2i}^2
1	25	18	10
2	15	20	8

Se pide:

a) Dar una estimación e I.C. al 95% de la media de salario anual por empleado en cada ciudad.

b) Dar una estimación e I.C. al 95% de la media de salario anual por empleado en la población de sucursales del banco.

---

a) En la primera ciudad, se trata de un diseño de muestreo por conglomerados en dos etapas, con muestreo pptr en primera etapa y m.a.s. en la segunda. El estimador del total es por lo tanto:

$$\widehat{N\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{p_i} = \frac{1}{2}(\frac{20\cdot 17}{0.10} + \frac{20\cdot 15}{0.05}) = 4700$$

y por lo tanto, como $N = 15\cdot 20 = 300$ empleados, el estimador de la media es $\widehat{\overline{y}} = \dfrac{4700}{300} = 15.67$.

La varianza de este estimador es:

$$\widehat{V}(\widehat{\overline{y}}) = \frac{1}{N^2}\frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{N_i\widehat{\overline{y}}_i}{p_i} - \widehat{N\overline{y}})^2 =$$

ISBN 9788496866133

$$=\frac{1}{300^2}\frac{1}{2(2-1)}[(\frac{20\cdot 17}{0.10}-4700)^2+(\frac{20\cdot 15}{0.05}-4700)^2]=18.77.$$

El intervalo de confianza al 95% asociado a estas estimaciones es $(7.17, 24.16)$.

En la segunda ciudad, los tamaños son desiguales y se requiere el estimador de razón a tamaño. Este es:

$$\widehat{\overline{y}}_R=\frac{\sum_{i=1}^{n}N_i\widehat{\overline{y}}_i}{\sum_{i=1}^{n}N_i}=\frac{25\cdot 18+15\cdot 20}{25+20}=16.67.$$

Con varianza estimada:

$$V(\widehat{\overline{y}}_R)=\frac{(1-f_1)}{\overline{N}^2 n}\sum_{i=1}^{n}\frac{N_i^2(\widehat{\overline{y}}_i-\widehat{\overline{y}}_R)^2}{(n-1)}+\frac{1}{L\overline{N}^2 n}\sum_{i=1}^{n}\frac{N_i^2(1-f_{2i})s_{2i}^2}{m_i}=$$

$$=\frac{(1-2/25)}{22.5^2 2}[\frac{25^2(18-16.67)^2}{(2-1)}+\frac{15^2(20-16.67)^2}{(2-1)}]+$$

$$+\frac{1}{25\cdot 22.5^2 2}[\frac{25^2(1-2/18)10}{2}+\frac{15^2(1-2/20)8}{2}]=3.41.$$

Donde se ha aproximado $\overline{N}$ por su estimador muestral $\widehat{\overline{N}}=\frac{45}{2}=22.5$. Una estimación del total de empleados sería $\widehat{N}=L\widehat{\overline{N}}=562.5$.

b) Para todas las sucursales, se trata de agregar las estimaciones para los diferentes estratos.

Es más cómodo realizar la estimación general para el total y después corregir para la media:

$$\widehat{N\overline{y}}^*=[4700+562.5\cdot 16.67]=14076.8.$$

y

$$V(\widehat{N\overline{y}}^*)=300^2\cdot 18.77+562.5^2\cdot 3.41=2768245.3.$$

Así, el estimador global de la media es

$$\widehat{\overline{y}}^*=\frac{14076.8}{300+562.5}=16.32$$

y su varianza estimada:

$$V(\widehat{\overline{y}}^*)=\frac{2768245.3}{(300+562.5)^2}=3.72.$$

El intervalo de confianza será (9.02, 23.6).

Ejercicio 9.2

Se desea investigar la proporción de plantas de tomates afectadas por un tipo de hongo. En el invernadero hay 300 plantas divididas en 3 grandes bloques. Se extraen dos bloques sin reemplazamiento, con probabilidades proporcionales al tamaño en la 1ª y 2ª extracción. Dentro de cada bloque se escogen por m.a.s. 4 plantas. Se obtienen los datos:

ISBN 9788496866133

Bloque	nº de plantas afectadas en la muestra	N_i
1	3	140
2	2	100

Obtener un I.C. al 95% para la proporción de plantas afectadas en el invernadero, utilizando el estimador de Yates-Grundy de la varianza.

Es necesario calcular las probabilidades de inclusión de primer y segundo orden.

Como $p_1 = \frac{140}{300} = 0.466$, $p_2 = \frac{1}{3}$, $p_3 = \frac{60}{300} = 0.2$,

$$\pi_1 = 0.466(1 + \frac{1/3}{1-1/3} + \frac{0.2}{1-0.2}) = 0.816.$$

$$\pi_2 = \frac{1}{3}(1 + \frac{0.466}{1-0.466} + \frac{0.2}{1-0.2}) = 0.708.$$

y

$$\pi_{12} = 0.466 \cdot \frac{1}{3}(\frac{1}{1-0.466} + \frac{1}{1-1/3}) = 0.525.$$

Entonces, siendo $y_{ij} = 1$ si la planta está afectada e $y_{ij} = 0$ si no,

$$\widehat{N\overline{y}} = \sum_{i=1}^{n} \frac{N_i \widehat{\overline{y}}_i}{\pi_i} = \frac{140 \cdot \frac{3}{4}}{0.816} + \frac{100 \cdot \frac{2}{4}}{0.708} = 199.29$$

y

$$\widehat{p} = \widehat{\overline{y}} = \frac{199.29}{300} = 0.66.$$

La cuasivarianza en segunda etapa s_{2i}^2 se puede expresar, en el caso de proporciones, como

$$s_{2i}^2 = \frac{m}{m-1}\widehat{p}_i(1-\widehat{p}_i).$$

En nuestro caso es

$$s_{21}^2 = \frac{4}{4-1}\frac{3}{4}(1-\frac{3}{4}) = 0.25$$

y

$$s_{21}^2 = \frac{4}{4-1}\frac{2}{4}(1-\frac{2}{4}) = 0.33.$$

La varianza del estimador es, utilizando la de Yates-Grundy:

$$300^2 \widehat{V}(\widehat{p}) = \frac{(0.816 \cdot 0.708 - 0.525)}{0.525}\left(\frac{105}{0.816} - \frac{50}{0.708}\right)^2 +$$

$$\frac{140^2}{0.816}(1-4/140)\frac{0.25}{4} + \frac{100^2}{0.708}(1-4/100)\frac{0.33}{4} = 2926.77 \text{ con lo que}$$

$$\widehat{V}(\widehat{p}) = \frac{2926.77}{300^2} = 0.0325.$$

ISBN 9788496866133

El intervalo de confianza será $(0.31, 1)$.

Ejercicio 9.3

Se va a medir el gasto en medicinas al mes de los habitantes de cierta provincia de 700000 habitantes. Dentro de ésta se seleccionan 2 municipios con probabilidades proporcionales al tamaño y con reposición . Dentro de cada uno de los municipios se seleccionan por m.a.s. una de cada 10 farmacias . Por último se pide al dependiente seleccionar por m.a.s. la factura anónima de todos los clientes que han realizado gasto en su farmacia al mes sobre su gasto en medicinas. Los datos recogidos son los siguientes:

Primer municipio (100000 habitantes): Se examinan 3 farmacias, en las cuales se obtienen las siguientes cifras en euros: 10000 en la primera, 15000 en la segunda y 10000 en la tercera.

Segundo municipio (80000 habitantes): Se examinan 2 farmacias, en las cuales se obtienen las siguientes cifras en euros: 12000 en la primera y 30000 en la segunda.

a) Estimar el gasto total en medicinas en ese mes, en la provincia. Estimar la varianza del estimador.

b) Decir si la estimación es autoponderada.

---

a) En cada municipio se selecciona un 10% de farmacias. Así, en el primer municipio hay $N_1 = \dfrac{3}{0.10} = 30$ farmacias, y en el segundo hay $N_2 = \dfrac{2}{0.10} = 20$ farmacias.

Entonces, $N_1\widehat{\overline{y}}_1 = 30 \cdot \dfrac{10000 + 15000 + 10000}{3} = 350000$ y $N_2\widehat{\overline{y}}_2 = 20 \cdot 21000 = 420000$.

Las probabilidades de selección de los dos municipios son respectivamente, $p_1 = \dfrac{100000}{700000} = 0.143$ y $p_2 = \dfrac{80000}{700000} = 0.114$.

El estimador del total es por lo tanto,

$$\widehat{N\overline{y}} = \frac{1}{2}[\frac{350000}{0.143} + \frac{420000}{0.114}] = 3062500.$$

Su varianza estimada es:

$$\widehat{V}(\widehat{N\overline{y}}) = \frac{1}{2(2-1)}[(\frac{350000}{0.143} - 3065881.48)^2 + (\frac{420000}{0.114} - 3065881.48)^2] = 3.7 \cdot 10^{11}.$$

b) Se trata de muestreo pptr en primera etapa y m.a.s. en segunda. El estimador es:

$$\widehat{N\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i\widehat{\overline{y}}_i}{p_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i}{p_i m_i}\sum_{i=1}^{m_i} y_{ij}.$$

La estimación será autoponderada si $\dfrac{N_i}{p_i m_i} = cte.$

Como se ha visto, $f_2 = 0.10 = \dfrac{m_i}{N_i}$ es constante, pero p_i no lo es, por lo que $\dfrac{N_i}{p_i m_i}$ no es constante a menos que todos los municipios tengan la misma probabilidad p_i de ser escogidos. Por ejemplo, en los datos obtenidos, se ve que $\dfrac{N_1}{p_1 m_1} = 69.9$ y $\dfrac{N_1}{p_1 m_1} = 87.7$.

ISBN 9788496866133

La estimación por lo tanto, no es autoponderada.

Ejercicio 9.4

En un estudio en una plantación de lechugas realizada en un invernadero, se divide ésta en 40 secciones de 10 plantas. Se desea estimar el peso promedio de las lechugas, y se escogen por m.a.s. 4 secciones, y dentro de cada una de ellas se eligen 2 lechugas también por m.a.s. Se obtienen los siguientes pesos:

Sección	Pesos
1	0.8, 0.5
2	0.75, 1
3	0.4, 0.35
4	0.6, 0.7

a) Estimar el peso medio de las lechugas y dar un estimador de la varianza del estimador.

b) Realizar el apartado a) suponiendo que el muestreo de lechugas es por m.a.s.r. dentro de cada sección.

c) Realizar el apartado a) suponiendo que el número de secciones puede considerarse infinito.

d) Decir si la estimación del apartado a) es autoponderada. ¿Cuál es la probabilidad de cada lechuga de aparecer en la muestra?.

---

a) Para los diferentes apartados, es conveniente calcular antes las medias y cuasivarianzas muestrales dentro de cada sección:

$\widehat{\overline{y}}_1 = 0.65$, $s_{21}^2 = 0.045$

$\widehat{\overline{y}}_2 = 0.875$, $s_{22}^2 = 0.031$

$\widehat{\overline{y}}_3 = 0.375$, $s_{23}^2 = 0.00125$

$\widehat{\overline{y}}_4 = 0.65$, $s_{24}^2 = 0.005$

Al ser los tamaños iguales, el estimador insesgado es

$$\overline{\overline{y}} = \frac{1}{4}(0.65 + 0.875 + 0.375 + 0.65) = 0.6375.$$

Para calcular la varianza estimada se tiene que:

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{\overline{y}}_i - \overline{\overline{y}})^2 = 0.0418.$$

y

$$s_2^2 = \frac{1}{n}\sum_{i=1}^{n} s_{2i}^2 = 0.0205$$

ISBN 9788496866133

El estimador de la varianza de $\overline{\overline{y}}$ es:

$$\widehat{V}(\overline{\overline{y}}) = (1-f_1)\frac{s_1^2}{n} + (1-f_2)\frac{s_2^2}{mL} = (1-4/40)\frac{0.0418}{4} + (1-2/10)\frac{0.0205}{2\cdot 40} = 0.00962.$$

b) En caso de m.a.s.r. en segunda etapa el estimador es el mismo, pero la varianza se estima sin tener en cuenta el coeficiente de corrección por población finita en la segunda etapa:

$$\widehat{V}(\overline{\overline{y}}) = (1-4/40)\frac{0.0418}{4} + \frac{0.0205}{2\cdot 40} = 0.00967.$$

c) Si el número de secciones es infinito, en la estimación de la varianza se puede prescindir del coeficiente de corrección por población finita en primera etapa $(1-f_1)$:

$$\widehat{V}(\overline{\overline{y}}) = \frac{0.0418}{4} + (1-2/10)\frac{0.0205}{2\cdot 40} = 0.010655.$$

d) El estimador es

$$\overline{\overline{y}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\overline{y}}_i = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}.$$

Será estimación autoponderada si m_i es constante, cosa que es cierta pues $m_i = m = 2$ lechugas. Así,

$$\overline{\overline{y}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ij}.$$

La probabilidad de aparecer en la muestra para cada lechuga es $\dfrac{nm}{N} = 0.02$:

Ésta se puede calcular como la probabilidad de que sea escogida la sección a la que pertenece, $\dfrac{n}{L}$ (como se vio en teoría sobre m.a.s.) multiplicada por la probabilidad de que, supuesta la sección escogida, sea elegida la lechuga en cuestión, $\dfrac{m}{\overline{N}}$. Como $N = L\overline{N}$, se tiene el resultado.

Ejercicio 9.5

En un colegio se quiere estimar la proporción de niños que han tenido la rubeola. En el colegio hay 280 alumnos repartidos en 10 clases de alumnos de la misma edad, y se seleccionan 5 por m.a.s. Dentro de cada una de éstas clases se seleccionan por muestreo sistemático 12 alumnos y se les pregunta si han tenido la rubeola.

clase	nº total de alumnos	nº con rubeola en la muestra
1	40	2
2	20	3
3	35	4
4	25	2
5	25	1

a) Estimar de manera insesgada la proporción y el total de niños que han tenido la rubeola en el colegio.

ISBN 9788496866133

b) Dar un I.C. al 95% para cada una de éstas cantidades.

c) Decir si la estimación de la proporción es autoponderada y calcular cuál es la probabilidad de pertenecer a la muestra del alumno j de la clase i .

---

a) El muestreo sistemático en la segunda etapa se considera equivalente al m.a.s., pues la ordenación de los alumnos en la clase es aleatoria respecto a la variable de interés (haber tenido rubeola o no).

En consecuencia se tratará como un problema de muestreo bietápico de conglomerados con m.a.s. en ambas etapas.

El estimador de la proporción, puesto que se nos pide estimación insesgada y lo tamaños de los conglomerados son desiguales, será

$$\widehat{p} = \frac{L}{Nn}\sum_{i=1}^{n} N_i\widehat{p}_i = \frac{10}{280\cdot 5}(40\frac{2}{12} + ... + 25\frac{1}{12}) = 0.2113.$$

El estimador del total será $N\widehat{p} = 59$.

b) Calculando

$s_{2i}^2 = \frac{m}{m-1}\widehat{p}_i(1-\widehat{p}_i)$, es $s_{21}^2 = 1.81$, $s_{22}^2 = 0.20$, $s_{23}^2 = 024$, $s_{24}^2 = 1.81$, $s_{25}^2 = 0.0833$.

Su varianza estimada es:

$$\widehat{V}(\widehat{p}) = \frac{L^2(1-f_1)}{n(n-1)}\sum_{i=1}^{n}(N_i\widehat{p}_i - \frac{1}{n}\sum_{i=1}^{n} N_i\widehat{p}_i)^2 + \frac{L}{n}\sum_{i=1}^{n}\frac{N_i^2(1-f_{2i})s_{2i}^2}{m_i} =$$

$$=\frac{10^2(1-5/10)}{5(5-1)}[(6.67-5.91)^2 + ... + (2.08-5.91)^2]+$$

$$+\frac{10}{5}\left(\frac{40^2(1-12/40)1.81}{12} + ... + \frac{25^2(1-12/25)0.0833}{12}\right) = 0.00267.$$

Y para el total, $\widehat{V}(N\widehat{p}) = 280^2 0.00267 = 209.54$.

Los intervalos de confianza respectivos son, para la proporción, $(0.11, 0.31)$ y $(30.8, 87.5)$.

c) Al ser el estimador de la proporción

$$\widehat{p} = \frac{L}{Nn}\sum_{i=1}^{n} N_i\widehat{p}_i = \frac{L}{Nn}\sum_{i=1}^{n} N_i\frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij} = \frac{L}{Nnm}\sum_{i=1}^{n} N_i\sum_{j=1}^{m_i} y_{ij},$$

pues $m_i = 12$ constante para todo i, donde se ha denotado por $y_{ij} = 1$ si el alumno ha tenido rubeola, $y_{ij} = 0$ si no. Sería estimación autoponderada si N_i fuera constante, algo que no es cierto.

La probabilidad de que un alumno ij pertenezca a la muestra es la probabilidad de que sea seleccionada la clase a la que pertenece, que es $\frac{5}{10}$, multiplicada por la probabilidad de que sea seleccionado el alumno dada que ha sido seleccionada su clase i. Esta última probabilidad es $\frac{m}{N_i}$. Por lo tanto la probabilidad de que un alumno concreto ij sea escogido es $\frac{5}{10}\frac{12}{N_i} = \frac{6}{N_i}$. Los alumnos pertenecientes a clases más amplias tienen menos probabilidad de ser escogidos.

ISBN 9788496866133

Ejercicio 9.6

Se va a estudiar el seguimiento que una huelga va a tener entre los trabajadores de telefónica en Extremadura. Se seleccionan con probabilidades iguales y reemplazamiento oficinas dentro de cada una de las provincias. Dentro de cada oficina se eligen por m.a.s. un 20% de trabajadores y se les pregunta si van a secundar la huelga.

Se obtienen los siguientes datos:

	A	B	C
Cáceres	10 (420)	8, 6, 10	6, 5, 8
Badajoz	8 (360)	15, 6, 5	10, 3, 3

donde las columnas representan:

A= Número total de oficinas y total de trabajadores (entre paréntesis) en la provincia.

B= Número de trabajadores muestreados en cada una de las oficinas seleccionadas.

C= Número de trabajadores que secundarán la huelga, de entre los muestreados, en cada una de las oficinas.

Estimar la proporción y el total de trabajadores que irán a la huelga en Extremadura, estimando la varianza de ambos estimadores.

Se trata de muestreo estratificado, pues se trata independientemente cada provincia (estratos). Se mostrará cómo se hacen los cálculos para la provincia de Cáceres.

En ésta, se realiza muestreo bietápico con m.a.s.r. en primera etapa y m.a.s. en segunda etapa. El tamaño de los conglomerados (oficinas) es desigual, pues al muestrearse 8 trabajadores en la primera oficina, y ser éste valor un 20% del total de trabajadores en esa oficina, hay $N_1 = \dfrac{8}{0.2} = 40$ trabajadores en total en ella. De modo similar, $N_2 = 30$ y $N_3 = 50$. Hay $L = 10$ oficinas en total, y $N = 420$.

Al ser m.a.s.r. en primera etapa, se utilizan las fórmulas de muestreo bietápico con probabilidades p_i y reemplazamiento, con $p_i = \dfrac{1}{L} = 0.1$ para todo i (probabilidades iguales). Así, como $1/p_i = 10$, se tiene:

$$\widehat{Np}_{Cáceres} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i\widehat{p}_i}{p_i} = \frac{10}{3}\left(40\frac{6}{8} + 30\frac{5}{6} + 50\frac{8}{10}\right) = 316.66$$

y

$$\widehat{V}(\widehat{Np}_{Cáceres}) = \frac{1}{3(3-1)}\left((300-316.67)^2 + (250-316.67)^2 + (400-316.67)^2\right) = 1944.44.$$

Para la proporción, será:

$$\widehat{p}_{Cáceres} = \frac{\widehat{Np}_{Cáceres}}{N} = \frac{316.67}{420} = 0.753$$

y

ISBN 9788496866133

$\widehat{V}(\widehat{p}_{Cáceres}) = \dfrac{1944.44}{420^2} = 0.011.$

En la provincia de Badajoz, los cálculos son similares, llegando a:

$\widehat{Np}_{Badajoz} = 213.33$ y $\widehat{V}(\widehat{Np}_{Badajoz}) = 8711.11$, con

$\widehat{p}_{Badajoz} = 0.592$ y $\widehat{V}(\widehat{p}_{Badajoz}) = 0.067.$

En conjunto, para toda Extremadura, se tiene que el total estimado es:

$\widehat{Np}_{Extremadura} = \widehat{Np}_{Cáceres} + \widehat{Np}_{Badajoz} = 530.$

y su varianza será:

$\widehat{V}(\widehat{Np}_{Extremadura}) = 10655.55.$

Para la proporción, es:

$\widehat{p}_{Extremadura} = \dfrac{\widehat{Np}_{Extremadura}}{780} = 0.679$

y

$\widehat{V}(\widehat{p}_{Extremadura}) = \dfrac{\widehat{V}(\widehat{Np}_{Extremadura})}{780^2} = 0.0175.$

Ejercicio 9.7

Realizar los cálculos del ejercicio 9.4 , apartado a) con la ayuda del SAS y la macro estimbimas.

En primer lugar se crea el archivo de datos:

```
data lechuga;
input seccion y;
tama=10;
cards;
1 0.8
1 0.5
2 0.75
2 1
3 0.4
3 0.35
4 0.6
4 0.7
;
```

En la macro estimbimas se especifican los parámetros del ejemplo:

```
%estimbimas(lechuga,y,seccion,tama,1,40,4,400);
```

Obteniendo los mismos resultados que el ejercicio 9.4 para el estimador y su varianza.

ISBN 9788496866133

Ejercicio 9.8

Realizar los cálculos del ejercicio 9.3 , apartado a) con la ayuda del SAS y la macro estimbipptr.

En primer lugar se crea un archivo, incorporando la información calculada de las probabilidades y de los tamaños de los conglomerados (las probabilidades se dejan como fracción para obtener mayor precisión en los cálculos):

```
data municipios;
input muni gasto;
if muni=1 then do;tama=30;pi=1/7;end;
if muni=2 then do;tama=20;pi=0.8/7;end;
cards;
1 10000
1 15000
1 10000
2 12000
2 30000
;
```

A continuación se ejecuta la macro. Puesto que el número de unidades elementales (farmacias) es desconocido en el problema, se pone como missing (en el último lugar de los parámetros de la macro). La macro calculará la estimación del total, pero no podrá calcular la media.

```
%estimbipptr(municipios,.,gasto,muni,tama,1,8,2,.);
```

Se obtienen los mismos resultados de ejercicio 9.3 para el estimador del total y su varianza.

Ejercicio 9.9

Realizar los cálculos del ejercicio 9.2 , apartado a) con la ayuda del SAS y la macro estimbippt.

Para poder ejecutar la macro, hay que presentar el archivo con una observación por unidad elemental, por lo cual se crea así:

```
data dos;
input conglo y tama;
cards;
1 1 140
1 1 140
1 1 140
1 0 140
2 1 100
2 0 100
2 0 100
;
```

A continuación es necesario crear un archivo con las probabilidades de inclusión, calculadas en el ejercicio:

ISBN 9788496866133

```
data uno;
input unit selectionprob jtprob_1 jtprob_2;
conglo=unit;
cards;
1 0.816 0 0.525
2 0.708 0.525 0
;
```

Se ejecuta la macro, indicando el archivo de datos y el de las probabilidades de inclusión:

```
%estimbippt(dos,uno,y,conglo,tama,2,300);
```

Obteniendo el estimador y la varianza de Yates-Grundy (además de la de Horvitz-Thompson).

Ejercicio 9.10

Se dispone del archivo SAS comunidad, con el que se trabajó en el ejercicio 8.13. Se trata de una encuesta para estimar el gasto en comida mensual de las familias en una comunidad. Hay 10 bloques, con diferente número de viviendas, con un total de 260 viviendas.

Se desea realizar muestreo por conglomerados bietápico, con diferentes esquemas de muestreo:

a) Utilizar la macro extrabimas para extraer una m.a.s. de 3 bloques, y dentro de cada uno de los bloques una m.a.s. de 12 viviendas. Utilizar las semillas 1234 y 1234 en 1ª y 2ª etapa. Utilizar la macro estimbimas para estimar el gasto medio en comida en toda la comunidad, mediante el estimador de razón a tamaño.

Repetir el ejercicio con la semilla 1235 en 1ª y 2ª etapa y con la semilla 1236.

b) Realizar el apartado anterior, pero con m.a.s.r. en segunda etapa. Utilizar la macro extrabimasr para la extracción.

c) Realizar muestreo pptr de 3 bloques con la macro extrabipptr, y dentro de cada uno de éstos una m.a.s. de 12 viviendas. Realizar estimaciones con la macro estimbipptr. Utilizar las mismas semillas para la extracción.

d) Realizar muestreo ppt de 3 bloques con la macro extrabippt, y dentro de cada uno de éstos una m.a.s. de 12 viviendas. Realizar estimaciones con la macro estimbippt. Utilizar las mismas semillas para la extracción.

e) Realizar m.a.s. de 36 viviendas y estimar la media. Repetir el proceso con las mismas semillas indicadas.

f) Resumir todos los resultados anteriores en una tabla comparativa.

---

a) Los programas de extracción y estimación son los siguientes:

```
%extrabimas(comunidad,muestra,info1,info2,bloque,10,3,12,.,1234,1234);
%estimbimas(muestra,gasto,bloque,nviv,2,10,3,260);
```


ISBN 9788496866133

obteniendo los resultados de la tabla, cambiando la semilla cada vez.

b) En la macro estimbimas hay que indicar que hay reemplazamiento en la segunda etapa, con el número 1 en lugar del 2 en la quinta posición. Los programas son :

```
%extrabimasr(comunidad,muestra,info1,info2,bloque,10,3,12,.,1234,1234);
%estimbimas(muestra,gasto,bloque,nviv,1,10,3,260);
```

c) Se ejecutan los programas:

```
%extrabipptr(comunidad,.,info1,muestra,codif,bloque,nviv,10,3,12,.,2,1,1234,1234);
%estimbipptr(muestra,.,gasto,bloque,nviv,1,10,3,260);
```

d) Se ejecutan los programas:

```
%extrabippt(comunidad,.,info1,muestra,inclusion,codif,bloque,nviv,10,3,12,.,1,1234,1234);
%estimbippt(muestra,inclusion,gasto,conglo,nviv,3,260);
```

e) Con el proc surveyselect y el proc means se extrae la muestra y estima la media, respectivamente:

```
proc surveyselect data=comunidad out=muestra n=36 method=srs seed=1234;
run;
proc means data=muestra;var gasto;run;
```

f) La tabla que queda es:

Semilla	m.a.s. 1^a,2^a	m.a.s. 1^a, m.a.s.r. 2^a	pptr 1^a, m.a.s. 2^a	ppt 1^a, m.a.s. 2^a	m.a.s.
1234	615.5	607.5	631.9	588.6	617.5
1235	651.6	590.3	610.2	602.5	689.4
1236	639.1	666.4	652.2	632.7	639.1

Las medias respectivas de cada una de las columnas son 635.4, 621.4, 631.4, 607.9, 649.8.

Recordando que la verdadera media era 636.4, se observa que no hay mucha diferencia en las estimaciones, según los métodos empleados. Esto es debido a una cierta homogeneidad entre conglomerados (son muy parecidos entre sí). Hay que advertir que si la precisión del muestreo por conglomerados es similar a la del m.a.s., como ocurre en este caso, en general es preferible el muestreo por conglomerados por motivos prácticos de la recogida de datos.

ISBN 9788496866133

10.8 Ejercicios propuestos

1) Una empresa de investigación de mercados ideó un plan de muestreo para estimar las ventas semanales de una determinada marca de cereales en un área geográfica concreta. La empresa decidió muestrear ciudades dentro del área y supermercados dentro de las ciudades. La medición de interés fue el número de cajas vendidas del cereal en una semana específica. Se muestrearon 5 ciudades de entre las 20 del área, obteniéndose los resultados de la tabla adjunta.

Ciudad	Supermercados	Supermercados muestreados	Media muestral	Cuasivarianza muestral
1	45	9	102	20
2	36	7	90	16
3	20	4	76	22
4	18	4	94	26
5	28	6	120	12

a) Estimar las ventas promedio de todos los supermercados del área para la semana específica y facilitar el error de muestreo . ¿Es insesgado el estimador utilizado?

b) ¿Se tiene suficiente información para estimar el número total de cajas de cereales vendidas en todos los supermercados del área durante la semana? Si es así, estimar dicho total y hallar el error de muestreo.

c) Si se llevara a cabo nuevamente un estudio semejante al descrito, ¿recomendarías que las ciudades se muestrearan con probabilidades proporcionales al número de supermercados? ¿Por qué?.

2) En una determinada zona geográfica se desea saber a si las mujeres que trabajan en las PYMES (pequeñas y medianas empresas) de menos de 50 empleados les interesaría trabajar a tiempo parcial. Para obtener esta información, primero se estratifican las empresas por sectores:

Sector A: Agricultura, ganadería y alimentación.

Sector B: Servicios.

Sector C: Industria y construcción.

De dichos sectores se conoce el número de empresas, y el número de trabajadoras por empresa:

ISBN 9788496866133

	Sector A	Sector B	Sector C
Nº de empresas	30	40	30
Nº empleadas por empresa	2,4,5,6...	10	4

Se decide muestrear en cada sector del siguiente modo:

Sector A: muestreo con reposición y probabilidad proporcional al número de empleadas, de 3 empresas.

Sector B: m.a.s. de 4 empresas

Sector C: m.a.s. de 3 empresas.

A todas las trabajadoras de las empresas muestreadas se les preguntó si querían trabajar a tiempo parcial. El número de mujeres que contestó afirmativamente en cada sector fue el siguiente:

Sector A:

Nº de empleadas	5	6	3
Responden afirmativamente	5	5	3

Sector B: 2,5, 3 y 1 respectivamente.

Sector C: 2,2, y 2 respectivamente.

Sabiendo que el número total de trabajadoras en el Sector A es 150, estimar la proporción total de empleadas que desean trabajar a tiempo parcial (dentro de la población objeto de estudio) y facilitar la desviación típica del estimador utilizado.

3) De una edición ilustrada de la obra de Julio Verne "La vuelta al mundo en 80 días" se desea conocer la proporción de renglones que tienen al menos un acento ortográfico. Para ello, primero se clasifican las páginas del libro dependiendo de si poseen ilustraciones o encabezamientos o son finales de capítulos (páginas que denominaremos "incompletas"), o si por el contrario carecen de ilustraciones y espacios en blanco (páginas completas).

El número de páginas y de renglones por página, atendiendo a la clasificación establecida, es el siguiente:

	Páginas incompletas	Páginas completas
Nº de páginas	50	200
Nº de renglones por página	2,3,6,7,...	40

Para realizar el estudio se seleccionaron, de entre las páginas incompletas, por muestreo con reemplazamiento y con probabilidades proporcionales al número de renglones, 2 páginas incom-

ISBN 9788496866133

pletas. Dentro de cada una de ellas se seleccionaron por m.a.s. un 50% de los renglones que contenían. el número de renglones por página y la proporción de renglones con al menos un acento ortográfico descubiertos en ambas páginas se resumen en la siguiente tabla:

Nº de líneas por página	20	28
Proporción de renglones acentuados	0.8	0.5

Por otro lado, dentro del grupo de las páginas completas se seleccionaron por m.a.s. 4 páginas, y se observó la existencia o no de acentos ortográficos en 10 renglones de cada una de ellas (seleccionados también mediante m.a.s.). El número de líneas analizadas con al menos un acento ortográfico fue 9, 5, 8 y 4 respectivamente.

Sabiendo que el número total de líneas de las páginas incompletas es 1000, estimar la proporción de renglones del libro que tienen al menos un acento ortográfico, y facilitar un intervalo confidencial al 95%, suponiendo la hipótesis de normalidad.

4) Una gran urbanización de la Costa del Sol está constituida por casas familiares y apartamentos. Ambos tipos de viviendas se concentran en bloques. Los 30 bloques de casas familiares contienen 40 viviendas cada uno, mientras que los 20 bloques de apartamentos varían de tamaño (8 apartamentos, 4, 6, 12,...) , conteniendo entre todos ellos a 200 apartamentos.

Se desea conocer la proporción de viviendas de extranjeros en la urbanización. Para ello, se seleccionaron de entre los bloques de casas familiares 4 bloques, y en cada uno de ellos 10 viviendas (m.a.s. en ambos casos). Las proporciones observadas de viviendas de extranjeros fueron 0.4, 0.3, 0.7 y 0.2 respectivamente.

Por otro lado, dentro del conjunto de los bloques de apartamentos se seleccionaron por muestreo con reemplazamiento y probabilidades proporcionales a los tamaños 3 bloques. De cada uno de ellos se tomó por m.a.s. una fracción muestral del 50% de los apartamentos. El número de apartamentos por bloque y de viviendas de extranjeros encontrados se resumen en la tabla siguiente:

Nº de apartamentos por bloque	16	8	6
Nº de apartamentos extranjeros observados	6	3	3

Estimar la proporción de viviendas de extranjeros en la urbanización, y facilitar un intervalo confidencial al 95%.

5) Se desea conocer la proporción de hogares de una determinada Comunidad Autónoma en los que hay más de un coche. Para ello se estratifican las seccciones censales de la Comunidad en zona rural y zona urbana. Las 25 secciones de la zona rural varían de tamaño (80 hogares, 108 hogares, 95 hogares...) conteniendo entre todas ellas 2000 hogares, mientras que las 20 secciones de la zona urbana contienen 500 hogares cada una.

De entre las secciones censales de la zona rural, se seleccionaron por muestreo con reemplazamiento y probabilidades proporcionales a los tamaños 4 secciones. De cada una de ellas se tomó

ISBN 9788496866133

una fracción muestral del 20% de los hogares por m.a.s. El número de hogares por sección y de hogares observados con más de un coche se recogen en la siguiente tabla:

Nº de hogares por sección censal	135	80	45	100
Nº de hogares observados con más de un coche	3	4	2	4

Por otro lado, se seleccionaron 3 secciones censales de la zona urbana, y de cada una de ellas el 10% de los hogares (m.a.s. en ambos casos). El número de hogares observados con más de un coche fue 25, 40 y 10 respectivamente.

Estimar la proporción de hogares de la Comunidad con más de un coche, y facilitar un I.C. al 95%.

6) Se desea estimar la proporción de alumnos que practican algún deporte en una determinada zona. Para ello consideraremos como población a estudiar los alumnos del Instituto de Enseñanza Secundaria (I.E.S.) de la zona. En este curso 200/2001 se han matriculado 1000 alumnos y han sido agrupados en 50 aulas de diferente tamaño, debido a la gran variedad de optativas que ofrece la LOGSE. Realizamos un muestreo bietápico con probabilidades iguales y sin reposición en ambas etapas. En la primera etapa se obtienen 5 aulas con 6, 10, 8, 20 y 60 alumnos respectivamente. De estos alumnos, en segunda etapa, seleccionamos sólo algunos con una fracción de muestreo $f_{2i} = 4/M_i$ siendo M_i el número de alumnos matriculados en el aula iésima. A la muestra de alumnos en cada aula seleccionada se les pregunta si realizan algún deporte obteniéndose que en la primera aula hay una respuesta afirmativa, 3 en la 2º, 2 para las aulas 3 y 4, y 3 alumnos practican algún deporte en la muestra del aula 5.

Con estos datos se pide un estimador insesgado para la proporción de alumnos que practican algún deporte y su error de muestreo.

7) Para calcular los ingresos medios por vivienda en una población formada por dos barrios, uno dedicado a la industria y otro a la agricultura, se decide realizar un muestreo en cada uno de ellos. En el barrio industrial se selecciona una muestra de bloques de tamaño 3 sin reposición y con probabilidades iguales. Los tres bloques de viviendas muestreados corresponden a un 5% del total de bloques en ese barrio. Estos bloques están formados por 50 viviendas cada uno. Para cada bloque seleccionamos, también sin reposición, una submuestra de tamaño 5. En cada vivienda de la muestra se obtienen los ingresos medios. Los datos obtenidos se muestran en la siguiente tabla:

Bloque	Ingresos medios en los hogares muestreados
1	16, 19, 22, 18, 25
2	20, 19, 23, 18, 25
3	16, 15, 20, 18, 21

Por otra parte, el barrio agrícola está formado por bloques deviviendas de diversos tamaños. Seleccionamos cuatro de ellos con reposición y probabilidades proporcionales al tamaño. En

ISBN 9788496866133

cada bloque seleccionado se realiza un submuestreo tomando cinco hogares en cada uno de ellso. Los resultados obtenidos se muestran en la tabla:

Bloque	Tamaño	Ingresos medios en los hogares muestreados
1	40	12, 15, 18, 13, 17
2	60	19, 17, 20, 21, 13
3	45	10, 14, 6, 14, 16
4	55	7, 11, 15, 7, 15

Con los datos obtenidos dar una estimación de los ingresos medios por vivienda en cada barrio así como los ingresos medios en la población. Obtener los errores de muestreo en ambos casos.

8) Se desea muestrear tres hospitales con reemplazamiento de entre los seis que existen en una ciudad, con el fin de estimar la proporción de pacientes que permanecen en el centro más de dos días consecutivos. Como los hospitales varían de tamaño, éstos serán muestreados con probabilidades proporcionales al número de pacientes.

Los datos muestrales son:

Hospital	Nº de pacientes muestreados	Nº de los que permanecen más de dos días
2	43	25
4	28	15
6	19	8

Estimar la proporción de pacientes con permanencia superior a dos días consecutivos para los seis hospitales y establecer un límite para el error de estimación con un nivel de confianza de 95%.

9) Resolver el ejercicio 9.6 con la ayuda de la macro estimbipptr (nota: crear un archivo con la variable conglomerado y la variable pi que es igual para conglomerado, igual a $\frac{1}{L}$).

10) El archivo SAS supermat contiene un estudio realizado sobre 4810 alumnos en 91 Centros de Enseñanaza Secundaria de Madrid, sobre aptitudes de todo tipo. La variable centro indica el código de centro, y la variable tamacentro el número de alumnos encuestados en ese centro.

Para el resto del ejercicio, consideraremos el archivo como la población de interés. Los tests realizados están en escala de 1 a 100.

a) Calcular la nota poblacional media de la variable orto (test de ortografía).

ISBN 9788496866133

b) Extraer 5 centros por muestreo proporcional a la variable tamacentro, con reemplazamiento, y dentro de cada uno de ellos 10 alumnos, con la macro extrabipptr . Con la macro estimbipptr.calcular el valor del estimador y su varianza. Realizar el proceso con las semillas 1234, 1235, 1236, 1237, 1238. Rellenar con el estimador la parte correspondiente de la tabla inferior.

c) Realizar el mismo proceso sin reemplazamiento, con las macros extrabippt y estimbippt.

d) Realizar el mismo proceso, con m.a.s. en ambas etapas, con las macro extrabimas y estimbimas.

Muestra	m.a.s., m.a.s.	pptr, m.a.s	ppt, m.a.s.
1			
2			
3			
4			
5			

ISBN 9788496866133

11 UTILIZACIÓN DEL SAS EN MUESTREO

Al margen de la utilización del programa SAS en cada una de las técnicas, presentada en los capítulos correspondientes, se verá a continuación una síntesis de la utilización del SAS en muestreo, así como ciertas consideraciones de carácter general.

11.1 El procedimiento Surveyselect

El procedimiento Surveyselect está orientado a la extracción de muestras aleatorias de un archivo de datos SAS. Su sintaxis básica es la siguiente:

proc surveyselect data=*poblacion* out=*muestra* method=*método* n=*tamaño muestral* seed=*semilla*;

size *variable-tamaño*;

strata *estratos;*

run;

poblacion es el archivo SAS que contiene la población .

muestra es el archivo SAS de salida que contiene la muestra.

tamaño muestral es el tamaño de la muestra. Puede también aportarse en su lugar la opción rate=*fraccion* donde se especifica la fracción de muestreo.

semilla es la semilla de aleatorización para obtener la muestra. Si no se indica esta opción, el proc surveyselect utiliza la hora interna del ordenador.

ISBN 9788496866133

variable-tamaño es la variable auxiliar, para métodos de muestreo proporcionales a los valores de esa variable (la sentencia size es opcional, no se utiliza para m.a.s. por ejemplo).

estratos son las variables que indican los estratos (puede haber más de una). La sentencia strata es opcional, para muestreo estratificado. El método indicado en method=*método* se aplica a todos los estratos.

método es el método de muestreo utilizado. Se pueden nombrar los siguientes:

srs es la opción por defecto; es muestreo aleatorio simple sin reemplazamiento (m.a.s.).

urs muestreo aleatorio simple con reemplazamiento (m.a.s.r.).

pps_wr muestreo con reemplazamiento, con probabilidades proporcionales a la variable auxiliar.

pps muestreo sin reemplazamiento, con probabilidades proporcionales a la variable auxiliar, con el método de Hanurav.

sys muestreo sistemático.

El proc surveyselect no obtiene directamente muestras en muestreo por conglomerados en varias etapas, por lo cual hemos programado las macros extramono, extrabippt, extrabimas y extrabimasr para suplir esta carencia.

Algunos comentarios respecto al muestreo estratificado y ppt son necesarios a continuación.

11.1.1 Muestreo estratificado

En el capítulo de muestreo estratificado está expuesta la sintaxis del proc surveyselect para cada tipo de afijación. Si se desea realizar muestreo estratificado con diferente método de extracción en cada estrato, es necesario realizar el siguiente proceso:

1) Segmentar el archivo en varios, con un archivo diferente para cada estrato. Esto se puede realizar con el paso data:

```
data archivo1 archivo2...archivoL;
if estrato=1 then output archivo1;
...
if estrato=L then output archivoL;
```

2) Realizar la extracción de muestras por separado para cada archivo.

3) Unir todos los archivos con un paso data:

```
data union;set muestra1 muestra2....muestraL;
```

Para aplicar la diferente estimación en cada estrato hay que conservar también los archivos por separado.

ISBN 9788496866133

11.1.2 Muestreo ppt

En el muestreo ppt el proc surveyselect utiliza el método de Hanurav por defecto. Es aconsejable utilizar la opción jtprobs (ver el capítulo dedicado a muestreo ppt) para conservar las probabilidades de inclusión de los elementos muestrales.

Aunque se pueden especificar los métodos pps_Brewer y pps_Sampford, están restringidos solamente a dos observaciones por estrato, con lo cual no los hemos tratado aquí.

La restricción $p_i < \frac{1}{n}$ puede ser una molestia con bastante frecuencia. Remitimos al capítulo sobre muestreo ppt para ver las maneras prácticas de abordarlo, a través de unidades autorrepresentadas, estratificación u otros métodos.

11.2 El procedimiento Surveymeans

Es el procedimiento del SAS para realizar estimaciones sobre muestras extraídas en diversos tipos de muestreo. Su sintaxis básica es:

proc surveymeans data=*archivo N=total*;

var *variables*;

strata *estratos*;

cluster *conglomerados*;

weight *variable*;

run;

Donde *total* es el tamaño poblacional, que se indica a efectos de cálculo del coeficiente de corrección por población finita, y *estratos* y *conglomerados* indican respectivamente las variables de identificación de los estratos y conglomerados (si los hay).

El proc surveymeans asume que la variable indicada en la opción weight contiene, para cada observación, su peso en el procedimiento de muestreo. Si no se indica la opción weight, los pesos se toman como constantes para todas las observaciones.

Aunque el procedimiento surveymeans realiza los cálculos sobre una cierta variedad de métodos de muestreo, su algoritmo básico para la estimación de varianzas es el método de linearización por expansión de Taylor, que ocasionalmente puede dar lugar a estimaciones relativamente diferentes a las presentadas en este curso básico. Además, en caso de muestreo por conglomerados, se asume reemplazamiento en primera etapa, considerando solamente la variabilidad en primera etapa para el cálculo de varianzas.

Por estas razónes se han realizado las macros presentes en este trabajo. Existen también otras macros para realizar estimaciones sobre encuestas por muestreo en SAS (el conjunto de macros CLAN, realizado por Statistics Sweden, y el conjunto de macros GSE, realizado por Statistics Canada), pero no son de acceso gratuito.

ISBN 9788496866133

Estimación en muestras autoponderadas

La estimación por defecto de la media por el proc surveymeans es adecuada solamente en muestras autoponderadas, y esto suele ser fuente de errores de los usuarios. Se explica a continuación el funcionamiento del proc surveymeans en cuanto a estimación de la media.

El funcionamiento del algoritmo de estimación del proc Surveymeans consiste en calcular de manera general, como estimador de la media, una media ponderada por los pesos aportados por el usuario. Por ejemplo, en muestreo estratificado por conglomerados bietápico:

$$\widehat{\overline{y}}_{Surveymeans} = \frac{1}{w...} \sum_{h=1}^{L} \sum_{j=1}^{n_h} \sum_{k=1}^{m_{hj}} w_{ijk} y_{ijk}$$

donde $w...$ es la suma de los pesos sobre los elementos muestrales:

$$w... = \sum_{h=1}^{L} \sum_{j=1}^{n_h} \sum_{k=1}^{m_{hj}} w_{ijk}$$

y L, n_h, m_h son estratos, conglomerados y unidades respectivamente.

Hay que señalar que en este estimador, el denominador $w...$ depende de la muestra, y por lo tanto, es difícil que el estimador sea insesgado de la media poblacional si la muestra no es autoponderada. Se obtiene que, en general :

$$E(\widehat{\overline{y}}_{Surveymeans}) = E\left(\frac{1}{w...} \sum_{h=1}^{L} \sum_{j=1}^{n_h} \sum_{k=1}^{m_{hj}} w_{ijk} y_{ijk} \right) \neq \frac{1}{N} \sum_{h=1}^{L} \sum_{j=1}^{N_h} \sum_{k=1}^{M_{hj}} y_{ijk}$$

donde tanto $\dfrac{1}{w...}$ como $w_{ijk} y_{ijk}$ son variables aleatorias. La igualdad no se suele cumplir si los pesos no son iguales para todas las observaciones, pues $w...$ es la suma de los pesos de las observaciones muestrales.

Si la muestra es autoponderada, entonces los pesos en el estimador insesgado habitual son constantes, $w_{ijk} = W$ para todas las observaciones, y en el estimador del proc surveymeans los pesos se cancelan y se obtiene con el proc surveymeans el estimador insesgado de la media:

$$\widehat{\overline{y}}_{Surveymeans} = \frac{1}{N_{ijk}} \sum_{h=1}^{L} \sum_{j=1}^{n_h} \sum_{k=1}^{m_h} y_{ijk}$$

donde hemos denotado por N_{ijk} el total de observaciones de la muestra.

En el caso de muestras autoponderadas, además, al ser los pesos constantes no es necesario poner la opción weight en el proc surveymeans, pues el estimador es igual si se consideran todos los pesos como iguales.

Si por ejemplo se desea estimación sobre muestreo estratificado con dos estratos, con pesos $W_1 = \dfrac{10}{30}$ y $W_2 = \dfrac{20}{30}$, con m.a.s. de tamaño $n_1 = 2$ y $n_2 = 4$ respectivamente , los pesos relativos para cada observación son iguales, $\dfrac{10}{30 \cdot 2} = \dfrac{20}{30 \cdot 4}$. Así que no es necesario indicar los

ISBN 9788496866133

pesos en la opción weight por ser la muestra autoponderada. Se obtendrá entonces el estimador insesgado habitual, y su varianza estimada, que no coincide en todos los decimales con el estimador habitual de la varianza por utilizar el proc surveymeans la técnica de aproximación por linearización de Taylor.

En el ejemplo básico que viene en el SAS con el proc surveymeans, de muestreo por conglomerados monoetápico y estratificado (ver Ayuda y ejemplos del proc surveymeans) , se trata de una muestra autoponderada, por ser los conglomerados escogidos con afijación proporcional, y por lo tanto tampoco es necesario introducir los pesos, tal y como están indicados en el ejemplo. El programa da lugar al mismo resultado si no se utiliza la opción weight.

Como ejemplo de un caso simple que no puede ser resuelto directamente con un estimador insesgado por el proc surveymeans en su opción de estimación de media por defecto, por tratarse de estimación no autoponderada, si se ha realizado muestreo pptr en una población de N observaciones, para el cálculo del estimador de Hansen Hurwitz de la media

$$\widehat{\overline{y}} = \frac{1}{N} t_{HH} = \frac{1}{Nn} \sum_{i=1}^{n} \frac{y_i}{p_i} = \sum_{i=1}^{n} \frac{1}{Nnp_i} y_i \ ,$$

el peso relativo de cada observación i sería $\dfrac{1}{Nnp_i}$. Pero si se utiliza como peso $\dfrac{1}{Nnp_i}$, o cualquier otro peso proporcional a éste, pues dan todos el mismo resultado, el cálculo por defecto por el proc surveymeans del estimador de la media da

$$\widehat{\overline{y}} = \frac{1}{w.} \sum_{i=1}^{n} \frac{1}{Nnp_i} y_i, \text{ pero}$$

$$w. = \sum_{i=1}^{n} \frac{1}{Nnp_i} \neq 1$$

en general, pues $w.$ depende de las observaciones escogidas, y por lo tanto no se obtiene el resultado deseado, $\widehat{\overline{y}} = \sum_{i=1}^{n} \dfrac{1}{Nnp_i} y_i$.

Al calcular la media el proc surveymeans está por lo tanto orientado a la estimación sobre muestras autoponderadas, comunes en la práctica sobre todo cuando se realizan encuestas complejas con estratificación y varias etapas. Si la muestra no es autoponderada, siempre se puede recurrir a alguno de los métodos de corrección presentados en la sección 9.2.5. o bien utilizar la opción explicada a continuación.

Estimación en muestras no autoponderadas

Utilizando los pesos adecuados, se puede obtener el estimador correcto de la media en muestras no autoponderadas, a través de la estimación del total en el proc surveymeans . En el ejemplo anterior los pesos para la estimación del total serían los $\dfrac{1}{np_i}$ para cada observación muestral, y estarían en el archivo muestral en la variable llamada pesos, por ejemplo.

Se utilizará la opción sum, requiriendo una suma ponderada de los valores muestrales, para obtener una estimación del total:

ISBN 9788496866133

```
proc surveymeans data=archivo sum;
weight pesos;
run;
```

Con la opción sum el procedimiento surveymeans construye el estimador del total

$$\widehat{N\overline{y}}_{Surveymeans} = \sum_{h=1}^{L}\sum_{j=1}^{n_h}\sum_{k=1}^{m_{hj}} w_{ijk}y_{ijk}$$

que en el ejemplo será igual a

$\widehat{N\overline{y}} = \sum_{i=1}^{n} \frac{1}{np_i} y_i$ obteniendo el resultado habitual del estimador de Hansen Hurwitz para el total.

Si se desea la estimación de la media, basta obtener la estimación del total y dividirla por N. También puede utilizarse el peso $\frac{1}{Np_i}$ en el procedimiento, obteniendo el estimador correcto de la media.

Debido a que en este ejemplo se trata de muestreo con reemplazamiento, el valor de la varianza estimada por el proc surveymeans es igual al obtenido mediante la fórmula del estimador insesgado de la varianza estudiado para el estimador de Hansen Hurwitz.

Si en lugar de muestreo con probabilidades desiguales con reemplazamiento se utiliza muestreo con probabilidades desiguales sin reemplazamiento, los pesos a considerar en el proc surveymeans con la opción sum son, para cada observación muestral, $\frac{1}{\pi_i}$ para la estimación del total y $\frac{1}{N\pi_i}$ para la estimación de la media. Al no tener en cuenta las probabilidades de inclusión de segundo orden π_{ij}, el proc surveymeans obtiene un estimador de la varianza, basado en la linearización por Taylor, diferente a los estimadores clásicos de Horvitz-Thompson y Yates-Grundy.

Para otros métodos de muestreo (con o sin estratos, con o sin reemplazamiento, con o sin conglomerados, probabilidades iguales o desiguales, etc.), se construirán adecuadamente los pesos de muestreo asignándolos a una variable, y se definirán en la opción weight del proc surveymeans, ejecutado con la opción sum.

Por ejemplo, si se trata de muestreo por conglomerados bietápico con selección en primera etapa con probabilidades desiguales, los pesos para la estimación de la media serán $\frac{N_i}{Nn\pi_i m_i}$ donde N_i es el tamaño del conglomerado i, m_i es el tamaño muestral en segunda etapa en el conglomerado i y π_i es la probabilidad de inclusión del conglomerado i.

Hay que tener en cuenta que en general, considerando la estimación de la varianza, aún en casos de muestras autoponderadas, la varianza calculada para estos métodos sencillos de muestreo no tiene por qué coincidir exactamente con la estudiada en la teoría, por basarse en una aproximación, que es útil para métodos complejos de muestreo pero no es necesaria para métodos más básicos.

ISBN 9788496866133

En lo que respecta al proc surveymeans en la estimación en técnicas básicas como m.a.s.r. y m.a.s. , se han tratado en los capítulos relacionados. Una cuestión que puede ser de importancia es que el proc surveymeans utiliza la distribución t de Student en la construcción de intervalos de confianza, lo que puede ser más preciso que la aproximación Normal para muestras pequeñas .

11.3 Utilización de las macros

Las macros presentadas en este libro ha sido realizadas enteramente por los autores de éste. Su propósito es principalmente pedagógico, aunque se han utilizado en la práctica sobre encuestas y trabajos prácticos de muestreo y estimación, con buenos resultados.

Las macros de extracción de muestras comienzan por el sufijo *extra* y las macros dedicadas a estimación por el sufijo *estim*.

La utilización de las macros requiere haberlas compilado previamente. Para este propósito, se presentan todas conjuntamente en el archivo todomacros.sas para ser compiladas a la vez, además de también estar dispuestas por separado cada una en un archivo.

Respecto a su utilización, conviene tener en cuenta algunas cuestiones técnicas en caso de tener problemas al ejecutarlas:

- En general al principio de cada macro hay una línea del tipo

```
options nodate nocenter nonumber nonotes;
```

La opción nonotes elimina los comentarios realizados por el SAS en la ventana LOG sobre los procesos. La razón es que las macros implican a veces cierta cantidad de operaciones que llenan la pantalla LOG de comentarios. Si existe algún problema en la ejecución de la macro, para poder ver todos los comentarios, se puede cambiar en la macro la palabra nonotes por notes, o borrar la línea.

- Los valores missing pueden estropear el cálculo de estimaciones en muchos casos, por lo cual el archivo muestral debe ser depurado consecuentemente.
- Raramente puede ocurrir que el nombre de alguna variable de los archivos poblacionales o muestrales aportados por el usuario coincidan con los nombres de variables utilizados en la macro. No es normal que esto origine errores, pero en caso de duda se pueden cambiar los nombres de las variables en los archivos utilizados.
- Los parámetros de las macros son posicionales. Eventualmente, si el usuario lo desea, se pueden cambiar a parámetros con un igual en la definición de la macro:

```
%macro estimasr(archivo=poblacion,variable=y,npobla=numero,z=control);
```


ISBN 9788496866133

Y en la ejecución de la macro, se nombran también con un igual los parámetros

```
%estimasr(archivo=uno,variable=gasto,npobla=100,z=1);
```

Esta manera de nombrar los parámetros en la macro es preferida por algunos usuarios.

- El SAS considera las mayúsculas y minúsculas como diferentes en el nombre de las variables. En algunos casos esto puede dar problemas al no reconocer las macros el nombre de las variables, si no se indican como son. Es necesario cerciorarse por lo tanto de que al nombrar la variable en la ejecución de la macro el nombre de la variable aparece como es, con sus mayúsculas y minúsculas. Si es necesario, se puede ejecutar el procedimiento contents previamente para comprobar el nombre de las variables:

```
proc contents data=archivo;run;
```

11.4 Estimaciones sobre varias variables a la vez

Una posibilidad importante es la de poder realizar estimaciones a la vez sobre varias variables, algo que es una necesidad muy frecuente en la práctica. En el proc surveymeans esta opción es fácilmente accesible, al poder poner en el apartado var la lista de variables separada por espacios en blanco.

En el caso de la utilización de las macros, es necesario ejecutar una macro de repetición. Para cada macro y en cada circunstancia (tamaño poblacional, etc.) se crea la siguiente macro de repetición, que está también presente en el archivo todomacros.sas.

ISBN 9788496866133

```
/*****************************************************************/
/* PROGRAMA BASE PARA REPETIR CUALQUIER MACRO DE
ESTIMACIÓN CON VARIAS VARIABLES */
 /*****************************************************************/

%Macro Repetir( VarList ) ;
%Local Stop;
%Let Stop =
%Eval(%SysFunc(CountC(&VarList , %Str( ))) + (%Length(&VarList) >0 ));
%Do I = 1 %To &Stop ;
  /*%Put %Scan( &VarList , &I , %Str( ) ) ;*/
  %let variab=%Scan( &VarList , &I , %Str( ) ) ;

/*****************************************************************
     AQUí DEBAJO SE PONE LA MACRO CON SUS PARÁMETROS.
     EN LA CASILLA DE LA VARIABLE DE INTERÉS ES NECESARIO
     PONER LA PALABRA &variab
/*****************************************************************/
  %estimasr(muestra,&variab,50,1);
/*****************************************************************/
%End ;

%Mend repetir;
```

Para cada macro y circunstancia, simplemente se cambia en el texto presentado anteriormente la línea

```
%estimasr(muestra,&variab,50,1);
```

por la ejecución de la macro que interese. Por ejemplo, por

```
%estimonopptr(muestra,dos,con,&variab,1,10,5,850);
```

donde se pone la palabra clave &variab en el lugar de la variable de interés.

Finalmente, para ejecutar la repetición de las estimaciones de la macro sobre varias variables, basta ejecutar la macro repetir sobre una lista de variables (que se supone están presentes en el archivo de muestra):

```
%repetir(gasto ingresos edad);
```

11.5 Listado de las macros

Para la extracción de muestras en m.a.s.r., m.a.s., muestreo estratificado (con m.a.s., m.a.s.r., sistemático, pptr o ppt en cada estrato), muestreo sistemático, muestreo pptr y muestreo ppt

ISBN 9788496866133

se utilizará el proc surveyselect. Para estimación de medias, totales y proporciones en m.a.s. se utilizará el proc surveymeans.

Para el resto de procedimientos, se utilizarán las macros.

Estimación en muestreo aleatorio simple con reemplazamiento:

```
%estimasr(archivo,variable,npobla,z);
```

Estimación en muestreo aleatorio simple sin reemplazamiento:

```
%estimas(archivo,variable,npobla);
```

Estimación en muestreo estratificado:

```
%estimestrat(muestra,archivo2,vary,vartama,post,indicador,nestratos,N);
```

Estimación en muestreo sistemático:

```
% estimpen(archivo,variable,m,N);
```

Estimación de razón y regresión bajo m.a.s.:

```
%estimrazreg (archivo,variabley,variablex,mediax,ngrande,n);
%estimrazestrat (archivo1,archivo2,variabley,variablex,varestrato,
vartama,varmedx,mediax,indicador,ngrande);
```

Estimación en muestreo con probabilidades desiguales:

```
%estimpptr(archivo1,archivo2,variabley,id,ngrande,n,indicador);
%estimppt(archivo1,variabley,n,ngrande);
```

Extracción de muestras en muestreo por conglomerados monoetápico:

```
%extramono(archivo1,archivo2,codif,varconglo,nconglo,n,semilla);
%extramonoppt(archivo1,archivo2,archivo3,inclusion,codif,
varconglo,variablex,nconglo,reemplazo,indicador,n,semilla);
%extramonopptr(archivo1,archivo2,archivo3,codif,varconglo,
variablex,nconglo,indicador,n);
```

Estimación en muestreo por conglomerados monoetápico:

ISBN 9788496866133

```
%estimono(archivo1,variabley,varconglo,nconglo,n,ngrande);
%estimonoppt(archi1,inclusion,variabley,nconglo,n,ngran);
%estimonopptr(archi1,archi2,varconglo,variabley,indicador,
nconglo,n,ngran);
```

Extracción de muestras en muestreo por conglomerados bietápico:

```
%extrabimas(archivo1,archivo2,archivo3,codif,varconglo,nconglo,
n,n2,frac2,semilla1,semilla2);
%extrabimasr(archivo1,archivo2,archivo3,codif,varconglo,nconglo,
n,n2,frac2,semilla1,semilla2);
%extrabipptr(archivo1,archivo2,archivo3,archivo4,codif,varconglo,
varx,nconglo,n,n2,frac2,masr,
indicador,semilla1,semilla2);
%extrabippt(archivo1,archivo2,archivo3,archivo4,inclusion,codif,
varconglo,varx,nconglo,n,n2,frac2,indicador,semilla1,semilla2);
```

Estimación en muestreo por conglomerados bietápico:

```
%estimbimas(muestra,vary,varconglo,vartamacong,reemplazo,
nconglo,n,ngrande);
%estimbipptr(muestra,archivo2,vary,varconglo,vartamacong,indicador,
nconglo,n,ngrande);
%estimbippt(muestra,inclusion,vary,varconglo,vartamacong,nconglo,
n,ngrande);
```

11.6 Utilización del SAS en la práctica del muestreo

A lo largo del presente trabajo se han visto numerosos ejemplos pedagógicos en los que se dispone de toda la población en un archivo, del cual se extraen muestras y se calculan los valores de los estimadores. Estos casos, como aquellos en los que se realizan simulaciones, tienen por objetivo ilustrar sobre el comportamiento de los diferentes estimadores estudiados.

En la práctica, y antes del muestreo, no se dispone de la variable y para ningún elemento poblacional. Normalmente se dispone de un listado con el código o identificación de cada una de las unidades elementales (con la información de estratos, conglomerados y variables auxiliares si existen). Los pasos para la realización del muestreo y posterior estimación en la práctica son:

1. Lo primero que hay que hacer, si se quiere utilizar el SAS para escoger qué unidades elementales y/o de cualquier etapa hay que seleccionar, con estratos o no, es introducir los datos de identificación en un archivo SAS. Esto puede hacerse, si se dispone de esa información en un archivo de texto, Excel o cualquier formato de base de datos, con los procedimientos del SAS de importación de archivos.

ISBN 9788496866133

2. Una vez creado este archivo, se procede a seleccionar la muestra, con el procedimiento surveyselect o alguna de las macros. La muestra consistirá en una porción del archivo "poblacional", con sus mismas variables y por lo tanto con la variable de identificación de las unidades.

3. Seguidamente se procederá al trabajo de campo, recogiendo la información de cada una de las unidades cuya identificación esté en el archivo SAS muestral.

4. Una vez recogida esta información en forma de valores de variables para cada observación, se incorpora al archivo muestral SAS (si ha sido recogida en otro formato, se procede a su importación como archivo SAS).

5. Por último, se procede a la estimación de medias, proporción y totales utilizando el proc surveymeans o bien la macro correspondiente, sobre el archivo muestral que contiene la información recogida.

ISBN 9788496866133

Bibliografía

Azorín, F. Sánchez Crespo, L.J. (1986) *Métodos y Aplicaciones del muestreo.* Alianza Universidad.

Cochran, W.C. (1971) *Técnicas de muestreo.* Ed. Cecsa.

Hansen, M.H. , Hurwitz, W., Madow, W.G. (1953) *Sample survey methods and theory.* Wiley. New York.

Kish, L. (1965) *Survey Sampling.* Wiley. New York.

Lehtonen, R., Pahkinen, E. (1994) *Practical Methods for Design and Analysis of Complex Surveys.* John Wiley and Sons.

Lohr, S.L. (1999) *Muestreo: diseño y análisis.* I. Thompson Editores.

Mirás, J. (1985) *Elementos de muestreo para poblaciones finitas.* INE. Madrid.

Pérez, C. (1999) *Técnicas de muestreo estadístico. Teoría, práctica y aplicaciones informáticas.* RA-MA.

Raj, D. (1979) *La estructura de las encuestas por muestreo.* Fondo de Cultura Económica.

Scheaffer, R.L., Mendenhall, W. Ott, L. (1987) *Elementos de muestreo.* Grupo editorial Iberoamericano.

Singh, S. (2003) *Advanced Sampling Theory with Applications.* Kluwer Academic Publishers.

Singh, R., Singh, N. (1996) *Elements of Survey Sampling.* Kluwer A. Publishers.

Som, R. (1996) *Practical Sampling Techniques.* M. Dekker, Inc.

ISBN 9788496866133