INSTITUTO DE ESTADÍSTICA DE ANDALUCÍA

Juan Francisco Muñoz Rosas

Aportaciones a los métodos de estimación de parámetros lineales y no lineales con información auxiliar

Tesis premiada por el Instituto de Estadística de Andalucía

Instituto de Estadística de Andalucía
CONSEJERÍA DE ECONOMÍA, INNOVACIÓN Y CIENCIA

Juan Francisco Muñoz Rosas

Aportaciones a los métodos de estimación de parámetros lineales y no lineales con información auxiliar

Instituto de Estadística de Andalucía
Pabellón de Nueva Zelanda
Leonardo Da Vinci, 21
Isla de la Cartuja
41092 Sevilla
Teléfono: 955 03 38 00
Fax: 955 03 38 16-17
www.juntadeandalucia.es/institutodeestadistica

Juan Francisco Muñoz Rosas

Aportaciones a los métodos de estimación de parámetros lineales y no lineales con información auxiliar

JUNTA DE ANDALUCIA

Instituto de Estadística de Andalucía
CONSEJERÍA DE ECONOMÍA, INNOVACIÓN Y CIENCIA

Datos catalográficos
Muñoz Rosas, Juan Francisco

Aportaciones a los métodos de estimación de parámetros lineales y no lineales con información auxiliar / autor, Juan Francisco Muñoz Rosas. -- Sevilla : Instituto de Estadística de Andalucía, 2010

112 p. ; 30 cm. + 1 disco compacto (CD-Rom). -- (Tesis)
D.L. SE. 7631-2010
ISBN 978-84-96659-83-4
Tesis premiada por el Instituto de Estadística de Andalucía

1. Estadística matemática. 2. Estimación estadística. 3. Probabilidades. 4. Muestreo. I. Instituto de Estadística de Andalucía. II. Título. III. Serie

519.2(043.2)

Directora
María del Mar Rueda García
Departamento de Estadística e Investigación Operativa
Facultad de Ciencias
UNIVERSIDAD DE GRANADA

Autor
Juan Francisco Muñoz Rosas
Licenciado en Ciencias y Técnicas Estadísticas
Departamento de Métodos Cuantitativos para la Economía y la Empresa
UNIVERSIDAD DE GRANADA

Año de Edición: 2010 Instituto de Estadística de Andalucía

Depósito Legal: SE-7631-2010
I.S.B.N.: 978-84-96659-83-4
Tirada: 300 ejemplares

Índice

Introdución

1.1. Problemas planteados

En el campo del muestreo en poblaciones finitas son numerosas las aportaciones que pueden hacerse a los métodos de estimación con información auxiliar de parámetros lineales y no lineales. Por ejemplo, en los últimos años han surgido nuevas metodologías para obtener estimadores más precisos usando información auxiliar. Estas nuevas metodologías son los estimadores de calibración (Deville y Särndal, 1992) y el método de verosimilitud empírica (Chen y Sitter, 1999). De estas metodologías, el método de verosimilitud empírica tiene un buen comportamiento asintótico y empírico, pero a causa de su reciente aparición, existen bastantes situaciones donde no ha sido analizado. En este trabajo se plantean diversos escenarios (presencia de datos faltantes, estimación de la función de distribución bajo un enfoque basado en el diseño muestral, etc) donde este método no había sido examinado, se estudian sus propiedades más importantes y se comprueba su eficiencia desde el punto de vista teórico y empírico.

Por otro lado, los métodos clásicos estudiados en muestreo de poblaciones finitas se han centrado en la estimación de parámetros lineales como la media o el total. En las últimas décadas se ha estado tratando el problema de la estimación de la función de distribución por diversos autores, pero este no es el caso de la estimación de los cuantiles, los cuales no han sido definidos ni analizados en algunas situaciones, como por ejemplo en los diseños muestrales más complejos, etc. De este modo, en este trabajo se pretende plantear y estudiar la estimación de los cuantiles en aquellas situaciones que aunque son más complejas no son las menos utilizadas, puesto que son los diseños muestrales empleados por la mayoría de los organismos y agencias estadísticas, investigaciones sociales y económicas, etc. Además, los cuantiles son muy utilizados en estos organismos por la información que recogen y para obtener medidas de gran importancia para el interés de una nación, como por ejemplo la estimación de las líneas de pobreza, proporción de bajos ingresos, etc.

Existen determinados problemas para algunos de los estimadores de cuantiles que han sido propuestos en la literatura del muestreo. En primer lugar, varios de los estimadores de la función de distribución no cumplen las propiedades de una verdadera función de distribución, mientras que existen otros estimadores que dependen estrictamente de un modelo de superpoblación. En algunas ocasiones, puede ocurrir que no exista ningún modelo que se ajuste suficientemente bien a la población en estudio, por lo que una perspectiva basada en el diseño muestral resultaría más apropiada.

En resumen, los objetivos que se persiguen en este trabajo son: (i) analizar el método de verosimilitud empírica en campos no tratados (estimación de la función de distribución desde una perspectiva basada en el diseño muestral y usando una aproximación modelo-asistida,en presencia de datos faltantes, estimación de cuantiles, etc), (ii) estudiar el comportamiento de los cuantiles en diseños más complejos (muestreo en dos ocasiones con probabilidades desiguales o con múltiples variables auxiliares, muestreo bifásico, etc).

1.2. Objetivos científicos y aportes a la teoría del muestreo

A continuación se indica cómo se distribuye el presente texto y se comenta de forma breve los principales objetivos científicos y las aportaciones a la teoría del muestreo en poblaciones finitas.

En la siguiente sección se describe el marco de trabajo general seguido a lo largo del texto y se dan algunos conceptos básicos de la teoría del muestreo en población finitas. El objetivo es familiarizarse con la notación y conceptos que van a ser usados en todo el texto.

En la teoría de muestreo en poblaciones finitas el objetivo principal de cualquier metodología es la de mejorar las estimaciones de los parámetros en estudio en el sentido de construir nuevos estimadores que, para el mismo tamaño muestral, tengan menor error de estimación, lo que implica mayor precisión en las estimaciones de los parámetros, o equivalentemente, tengan el mismo error que los ya conocidos pero con un menor tamaño muestral, lo que produce una disminución en el coste real de la realización de la encuesta. Existen dos procedimientos para intentar mejorar las precisiones de las estimaciones. Por un lado, se pueden emplear nuevas técnicas de estimación y por otro, usar métodos de muestreo más complejos que utilicen más información (muestreo en ocasiones sucesivas, etc), o que la información auxiliar sea más fiable (muestreo bifásico), etc. La primera de estas técnicas se lleva a cabo en el Capítulo 2, en donde se pone a prueba el método de verosimilitud empírica como método de estimación, mientras que la segunda técnica se aplica en el Capítulo 3 para el problema de la estimación de cuantiles.

Como se ha comentado, el método de verosimilitud empírica se desarrolla en el Capítulo 2 bajo distintos escenarios. Esta reciente metodología obtiene estimadores tan eficientes (ver Chen y Sitter, 1999, Wu, 2002, Rue-

da, Muñoz, Berger, Arcos y Martínez 2006, etc.) como los utilizados clásicamente en muestreo de poblaciones finitas, lo que lo convierte en una alternativa válida a usar en las encuestas por muestreo, puesto que si el escenario es el apropiado puede ayudar a obtener estimaciones más eficientes, reducir costes en las encuestas, etc. En la Sección 2.2 se recopilan los principales aspectos y resultados del método de verosimilitud empírica. Además, bajo esta metodología se plantean varias situaciones de un interés relevante en la teoría del muestreo, de los que destacan el problema de los datos faltantes y la estimación de la función de distribución y cuantiles.

Cuando se realiza un estudio mediante encuestas o cualquier otro procedimiento, es usual encontrarse en presencia de datos faltantes, que vienen dados por parte del entrevistado o por cualquier otra circunstancia (pérdida casual de información, errores en la etapa de manipulación de datos, etc). Ante tal problema, una técnica frecuentemente utilizada es eliminar del estudio a aquellos individuos que presentan datos faltantes en alguna de sus variables. El inconveniente principal de esta técnica es el incremento del sesgo en las estimaciones. Otra técnica habitualmente utilizada es la imputación, que presenta el inconveniente de obtener en algunas ocasiones inferencias no válidas como consecuencia de considerar los valores imputados como si éstos fueran valores verdaderos.

En la Sección 2.3 se propone un camino alternativo para el tratamiento de los datos faltantes que no necesita eliminar del estudio a ningún individuo, aprovechando toda la información que se tiene en la muestra. Este procedimiento se desarrolla bajo el método de verosimilitud empírica. Se estudian las propiedades teóricas y mediante un estudio de simulación, se contrasta la precisión de los estimadores propuestos con otros estimadores conocidos y también diseñados para el tratamiento de datos faltantes. Véase también Rueda, Muñoz, Berger, Arcos y Martínez (2006).

El problema de la estimación de la función de distribución es un tema actual y muy importante del muestreo en poblaciones finitas, por tratarse de una función que permite determinar las características más importantes de la población en estudio, proporcionando información relevante acerca del comportamiento global de la población. Sin duda, los estimadores estudiados clásicamente en la teoría del muestreo, como totales, medias, proporciones y varianzas, no ofrecen tanta información como la función de distribución. El problema de la estimación de cuantiles y de otros parámetros de tipo no funcional queda resuelto con el conocimiento de la función de distribución, puesto que éstos pueden obtenerse mediante inversión directa de la función de distribución. Además, permite obtener medidas importantes como las líneas de pobreza, proporción de bajos ingresos, etc. y son muy útiles en investigaciones de tipo social o económico. Debido a la importancia de estos parámetros en algunas investigaciones o estudios, se debe disponer de buenos métodos y técnicas para obtener las mejores estimaciones posibles.

Bajo la aproximación modelo-calibrada, Chen y Wu (2002) propusieron estimadores de la función de distribución usando el método de verosimilitud empírica. Por otro lado, estos estimadores están basados en información auxiliar a través de un único punto del conjunto de valores para los que se define la función de distribución, presentando el problema de obtener estimaciones menos precisas cuando el argumento en el que se evalúa la función de distribución se encuentra bastante alejado del punto considerado para la variable auxiliar. Por tanto, estos estimadores presentan dos inconvenientes principalmente: (i) es necesario el conocimiento y el uso de un modelo de superpoblación para los datos muestrales del estudio y (ii) se hace un uso poco eficiente de la información auxiliar.

Asumiendo el método de verosimilitud empírica, en la Sección 2.4 se propone un estimador modelo-asistido para la función de distribución basado en un uso efectivo de la información auxiliar. Este estimador será más eficiente cuanto mayor sea la correlación entre las variables auxiliares y la variable principal. Además, no resulta necesario el conocimiento de un modelo de superpoblación, puesto que el estimador propuesto no es dependiente del modelo. El uso efectivo de la información auxiliar se justifica porque el estimador propuesto está basado en tres puntos perfectamente repartidos en el recorrido de valores en donde se define la función de distribución, de modo que, independientemente del valor donde se evalúe la función de distribución, este valor estará cercano a alguno de los tres puntos, obteniendo estimaciones más precisas para la función de distribución. Esto permitirá también mejorar la calidad de la estimación de los cuantiles y de aquellos otros parámetros relacionados con éstos y que suelen obtenerse en las grandes instituciones estadísticas. Una propiedad deseable de un estimador de la función de distribución, es que éste sea por sí mismo una verdadera función de distribución. Este es otro punto importante a la hora de obtener estimadores eficientes para los cuantiles poblacionales. Notamos que el estimador propuesto también posee esta propiedad.

En el Capítulo 3 se analiza el problema de la estimación de cuantiles bajo distintos esquemas de muestreo frecuentemente usados en la práctica, varios métodos de estimación y por último, usando el método de verosimilitud empírica.

La Sección 3.2 resuelve el problema de la estimación de cuantiles en muestreo bifásico cuando las muestras en cada una de las fases son seleccionadas mediante cualquier diseño muestral, con probabilidades iguales o desiguales. Se proponen varios estimadores de tipo directo, razón y exponencial que proporcionan estimaciones óptimas para un determinado cuantil. Se analizan propiedades importantes de estos estimadores tales como la insesgadez, estimación de varianzas, etc. Como caso particular, se investiga también el muestreo bifásico aplicado a la estratificación, diseño muestral que ofrece importantes ganancias en eficiencia debido a los beneficios que produce el muestreo estratificado. Todas estas propiedades se ven desde un punto de vista teórico, aunque el análisis de los estimadores se completa con estudios empíricos llevados a cabo para los cuartiles y bajos distintos diseños muestrales con probabilidades desiguales. En términos de sesgo y de eficiencia relativa, estos estudios reflejan que los estimadores propuestos mejoran a otros estimadores diseñados en muestreo bifásico.

La mayoría de las investigaciones llevadas a cabo por los organismos nacionales de estadística son periódicas,

es decir, se repiten a intervalos regulares de tiempo. Bajo este escenario, es aplicable la metodología propuesta en la Sección 3.3 para estimar cuantiles en muestreo en dos ocasiones, lo que puede permitir obtener una mayor precisión en la etapa de estimación como se ha comprobado desde el punto de vista teórico y práctico. Esta investigación se ha llevado a cabo, por un lado, para el caso de múltiples variables auxiliares, y por otro, bajo el uso de un diseño muestral arbitrario, siendo varios los objetivos científicos y aportes a la teoría del muestreo, puesto que los métodos tradicionales de estimación en muestreo de ocasiones sucesivas se han centrado en el problema de la estimación de parámetros lineales. Para el caso de la estimación de cuantiles, la situación es bastante diferente, y sólo recientemente este campo ha sido tratado por los estudios de investigación. En cualquier caso, los estudios existentes están basados únicamente en muestreo aleatorio simple y utilizan sólo la variable de interés en la fase de estimación, o bien sólo están diseñados para una única variable auxiliar.

En la Sección 3.4 se plantea el problema de la estimación de cuantiles mediante estimadores modelo-asistidos basados en el método de verosimilitud empírica. La aplicación de estos estimadores a la estimación de algunas medidas de pobreza también se discute dentro de esta sección. Debido a la complejidad natural de los cuantiles y principalmente de las medidas de pobreza que se manejan, se propone usar la técnica bootstrap para el problema de la estimación de las varianzas de los estimadores. En los numerosos estudios empíricos llevados a cabo, puede observarse que tanto los estimadores propuestos como las estimaciones de las varianzas presentan un buen cumplimiento en términos de sesgo y eficiencia relativa.

Una valoración global de los resultados obtenidos así como las principales conclusiones de todos los estudios de este texto se resumen en el Capítulo 4.

El texto se completa con una serie de apéndices de consulta sobre varios aspectos relacionados con los estudios llevados a cabo. Así, el Apéndice A recoge las principales propiedades y características de las poblaciones finitas que han sido usadas en los estudios de simulación. Además de un breve resumen estadístico de los datos de estas poblaciones, se muestran los diagramas de dispersión de tales poblaciones.

Por último, notar que todos los estudios de simulación se han llevado a cabo mediante el lenguaje de programación R. Todos los procedimientos y funciones para obtener en R tanto los estimadores propuestos en este texto como el resto de estimadores para cada diseño muestral están disponibles en el Apéndice **??**.

Son numerosas las razones por las que se ha usado este software. En primer lugar, es un lenguaje intuitivo con una gran cantidad de argumentos estadísticos que facilitan la implementación de los estimadores propuestos. Otros programas como $Mathematica$, $Matlab$, $C++$, etc., carecen de tales procedimientos estadísticos. Por otro lado, es un paquete que destaca por su rapidez y que permite obtener el mayor número de simulaciones en menor tiempo. R es un lenguaje de programación gratuito y disponible a cualquier usuario, al contrario de otros específicos de estadística como SAS, que debido a sus altas licencias está únicamente disponible, en la mayoría de los casos, a las grandes empresas. El dispositivo gráfico que dispone R y su compatibilidad con $S-PLUS$ son otros argumentos que hacen que la mayoría de los investigadores en el campo del muestreo en poblaciones finitas prefieran el uso de este software. Sirva de ejemplo los artículos publicados en este sentido (por ejemplo Wu, 2005) así como las conferencias internacionales sobre el programa R que también se están abriendo paso, como la segunda conferencia internacional de usuarios de R que se celebró del 15 al 17 de junio de 2006 en Viena, Austria. De hecho, el gran auge que está teniendo este software hace que se estén introduciendo día a día nuevos procedimientos y paquetes estadísticos.

1.3. Notación y conceptos básicos

En esta sección se describe el marco de trabajo usual en el ámbito del muestreo de poblaciones finitas. Además, se introducen algunos conceptos básicos y la notación común que se sigue a lo largo del texto.

Se denomina *población* a un conjunto de unidades del que se desea obtener cierta información. Esta población se denota como U, es finita y contiene N elementos distintos e identificados, es decir, $U=\{1,\ldots,i,\ldots,N\}$.

En la población U es posible medir o contar en cada unidad una o varias *características* o *variables*, o clasificar sus unidades de acuerdo a ellas. A partir de estos resultados se puede llegar al conocimiento de valores como la media, el total, la proporción, función de distribución, etc., a los que se denomina *parámetros poblacionales*. La media, el total, etc., son parámetros lineales, mientras que la función de distribución, cuantiles, etc., son parámetros no lineales.

Existen dos estrategias posibles para la recopilación de datos: (i) examinar todas las unidades de la población, es decir, realizar un censo, y (ii) examinar, según unos planes establecidos con anterioridad, unas pocas unidades de la población que son representativas, es decir, obtener una muestra, y suponer que de los resultados obtenidos se infieren a las características de toda la población.

En la práctica, determinados parámetros poblacionales son desconocidos y no pueden calcularse mediante un censo. Por está razón, se recurre a una muestra para estimar estos parámetros poblacionales. Así, una muestra es un subconjunto de unidades, s, de U seleccionados de acuerdo con un diseño de muestreo específico, d, que asigna una probabilidad conocida, $p(s)$, tal que $p(s)>0$ para todo $s\in S$, donde S es el conjunto de las posibles muestras s y $\sum_{s\in S}p(s)=1$. El valor de la media, total, proporción o función de distribución obtenido a partir de la muestra se denomina *estimador* del correspondiente parámetro poblacional.

Dentro de esta población interesa estudiar ciertas características de una variable de *estudio*, *interés* o *principal* denominada y. Las variables *auxiliares* son aquellas, que sin ser objeto de estudio, son usadas para varios fines, como por ejemplo, para la selección de unidades en la mues-

tra, mejorar las estimaciones, etc. Asociado al elemento i de la muestra se conoce exactamente y sin error el valor de la característica de interés, esta cantidad se denotará como y_i. Para P variables auxiliares, el vector de variables auxiliares viene dado por $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_p, \ldots, \mathbf{x}_P)$, donde $\mathbf{x}_p = (x_{1p}, \ldots, x_{ip}, \ldots, x_{Np})^t$. Se asume que estas variables auxiliares también son conocidas para aquellos individuos seleccionados en la muestra. En algunas ocasiones, se supone que los totales o medias poblacionales de las variables auxiliares son conocidos, es decir, las cantidades $\mathbf{X} = (X_1, \ldots, X_P)$ o $\overline{\mathbf{X}} = (\overline{X}_1, \ldots, \overline{X}_P)$ son conocidas, donde $X_p = \sum_{i=1}^N x_{ip}$ y $\overline{X}_p = N^{-1} \sum_{i=1}^N x_{ip}$.

La probabilidad de inclusión de primer orden asociadas al plan de muestreo d para un individuo i, π_i, indica la probabilidad que tiene este individuo de pertenecer a la muestra s. Asimismo, π_{ij} indica la probabilidad de que ambas unidades i y j pertenezcan a la muestra s. A esta cantidad se le llama probabilidad de inclusión de segundo orden. Otras cantidades que serán usadas son los pesos básicos del diseño $d_i = \pi_i^{-1}$, $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$, etc.

De este modo, los principales parámetros poblacionales desconocidos en la práctica y que habrá que estimar son la media poblacional de la variable de interés,

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^N y_i,$$

el total poblacional,

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

la función de distribución,

$$F_y(t) = \frac{1}{N} \sum_{i=1}^N \delta(t - y_i),$$

y el cuantil para un orden β ($0 < \beta < 1$),

$$Q_y(\beta) = F_y^{-1}(\beta) = \text{ínf}\{t \mid F_y(t) \geq \beta\},$$

donde $\delta(\cdot)$ es la función indicadora que toma el valor $\delta(a) = 1$ si $a \geq 0$ y $\delta(a) = 0$ en otro caso y $F_y^{-1}(\cdot)$ denota la función inversa de $F_y(\cdot)$.

Sin ningún tipo de información auxiliar, la media poblacional de la variable de interés, $\overline{Y}$, suele estimarse mediante el estimador de tipo Hortviz-Thompson

$$\overline{y}_{HT} = \frac{1}{N} \sum_{i \in s} d_i y_i. \tag{1.1}$$

Para el caso de la estimación de la función de distribución, este estimador viene dado por

$$\widehat{F}_{HTy}(t) = \frac{1}{N} \sum_{i \in s} d_i \delta(t - y_i), \tag{1.2}$$

aunque suele usarse el estimador de tipo Hájek que es una verdadera función de distribución. Este estimador viene dado por

$$\widehat{F}_{HKy}(t) = \sum_{i \in s} d_i^* \delta(t - y_i), \tag{1.3}$$

donde $d_i^* = d_i / \sum_{j \in s} d_j$. El cuantil de orden β puede estimarse directamente mediante la inversión de este último estimador, esto es,

$$\widehat{Q}_{HKy}(\beta) = \widehat{F}_{HKy}^{-1}(\beta) = \text{ínf}\{t \mid \widehat{F}_{HKy}(t) \geq \beta\}. \tag{1.4}$$

2. El método de verosimilitud empírica

El método de verosimilitud empírica para la estimación de parámetros fue propuesto en Chen y Qin (1993), aunque fueron Chen y Sitter (1999) quienes establecieron las bases teóricas principales de este método, y partir de las cuales se han basado todos los estudios posteriores. En este capítulo se investiga esta técnica reciente en diferentes campos del muestreo en poblaciones finitas.

En la Sección 2.2 se recogen los principales aspectos de esta metodología para el caso de la estimación de la media poblacional, pueden verse las propiedades asintóticas más importantes y los diferentes tipos de estimadores basados en cada una de las perspectivas de estimación.

En cualquier estudio es usual encontrarse con el problema de datos faltantes. En la Sección 2.3 se propone usar un estimador basado en el método de verosimilitud empírica como solución al problema de la existencia de datos faltantes (véase también Rueda, Muñoz, Berger, Arcos y Martínez 2006).

La estimación de la función de distribución mediante el método de verosimilitud empírica se estudia en la Sección 2.4. Se propone usar la aproximación modelo-asistida para obtener tal estimador, y se hace un uso eficiente de la información auxiliar al estar basado el estimador en varias variables auxiliares y en varios puntos de estimación.

2.1. Introducción

En la teoría del muestreo en poblaciones finitas, el objetivo principal de un método determinado para la obtención de estimadores o de cualquier diseño muestral es el de mejorar las estimaciones de los parámetros en estudio en el sentido de construir nuevos estimadores que, para el mismo tamaño muestral, tengan menor error de estimación, lo que implica mayor precisión en las estimaciones de los parámetros, o equivalentemente, tengan el mismo error que los ya conocidos pero con un menor tamaño muestral, lo que produce una disminución en el coste real de la realización de la encuesta.

Por estas razones fundamentalmente, la metodología del muestreo en poblaciones finitas precisa de nuevas aportaciones que abaraten los costes de los estudios o investigaciones estadísticas, se mejoren las estimaciones desde el punto de vista de la eficiencia o sesgadez y se dispongan, en general, de mejores propiedades.

Es conocido que según la información que se utilize en la etapa de estimación de parámetros, se tienen dos caminos para intentar mejorar la precisión de las estimaciones: por un lado, utilizar diseños muestrales más complejos (muestreos estratificados, por conglomerados, polietápicos, adaptativos, etc.) basados únicamente en los datos de la característica de interés, y por otro lado, emplear las metodologías propias de la teoría del muestreo en poblaciones finitas basadas en el uso de información auxiliar. Esta información auxiliar, dada a través de un vector de variables auxiliares, debe estar altamente correlacionada con la característica de interés para poder obtener mayor precisión en la etapa de estimación. Estas dos alternativas se pueden combinar para perseguir el objetivo de obtener mejores estimaciones, es decir, usar diseños muestrales más complejos en métodos de estimación de parámetros que utilicen información auxiliar es una opción muy atractiva en la materia que nos ocupa (véase Hedayat y Sinha, 1991).

El método de verosimilitud empírica, que se desarrolla a largo de este capítulo, permite combinar las dos ideas anteriores y es bastante eficiente como se ha comprobado tanto desde el punto de vista teórico como empírico (véase Chen y Qin, 1993, Chen y Sitter, 1999, Zhong, 2000, Chen y Wu, 2002, Sitter y Wu, 2002, Wu, 2003, Wu, 2004a, 2004c, Rueda y Muñoz, 2005, 2006a, 2006d, etc.).

Los primeros métodos que incorporan información auxiliar en la fase de estimación son los llamados métodos indirectos de estimación, entre los que destacan los conocidos métodos de razón, diferencia y regresión. Estos estimadores no siempre garantizan que se produzca una disminución del error de muestreo respecto a los estimadores que no usan información auxiliar. Esta ganancia en precisión depende en mayor medida de la relación entre las variables auxiliares y la variable objeto de estudio, del buen uso de las hipótesis que se supongan para emplear un procedimiento u otro, y de que dichas hipótesis se ajusten en mayor o menor medida al problema real.

Los estimadores anteriores se basan únicamente en los datos muestrales, es decir, utilizan un enfoque basado en el diseño muestral. Recientemente, en muestreo se está utilizando la perspectiva basada en modelos (ver p.e. Pérez, 2002 y Sánchez-Crespo, 2002) y la nueva aproximación modelo-calibrada (Wu y Sitter, 2001). Estas aproximaciones se basan en modelos de superpoblación y son dependientes de dichos modelos. El objetivo de estos métodos es obtener estimaciones más precisas, resultados más concluyentes en la comparación de estrategias, producir estrategias óptimas, obtener propiedades asintóticas más atractivas, etc., pero cuando el esquema de trabajo está perfectamente identificado con un modelo de superpoblación. Bajo esta perspectiva cobra especial importancia el uso de variables auxiliares cuyos valores tienen que ser conocidos para todos los individuos de la población. Por tanto, para poder usar este enfoque se debe conocer el adecuado modelo de superpoblación

asociado a los datos de la población en estudio. En resumen, estas aproximaciones son más eficientes que el enfoque basado en el diseño muestral cuando el modelo de superpoblación se ajusta bien, y pueden llegar a obtener propiedades no deseables, como inferencias no válidas, cuando se usa un modelo de superpoblación erróneo. En consecuencia, para llegar a cabo estas aproximaciones, sería conveniente obtener más información: el modelo de superpoblación apropiado y todos los valores de las variables auxiliares para todos los individuos de la población. Cuando no se dan estas circunstancias, puede resultar más apropiado un método de estimación basado en el diseño muestral.

Una alternativa intermedia entre los métodos anteriores y la clásica estimación basada en diseños, es la aproximación modelo-asistida. Ésta consiste en usar un modelo de superpoblación para obtener una estimación de un determinado parámetro poblacional, y entonces, usar éste último en la etapa de estimación. Sin perdida de eficiencia, la ventaja de este estimador es que sus estimaciones no son dependientes del modelo de superpoblación, permitiendo obtener inferencias válidas independientemente de si el modelo resulta ser apropiado o no para los datos de la población de estudio. El conocido estimador de regresión generalizado (Cassel *et al.*, 1976, Särndal, 1980), los estimadores de calibración (Deville y Särndal, 1992, Théberge, 1999, Wu y Luan, 2003) y el propio estimador de verosimilitud empírica (Chen y Qin, 1993, Chen y Sitter, 1999) pueden ser categorizados como aproximaciones modelo-asistidas.

Son dos los métodos para obtener estimadores que han aparecido recientemente: los estimadores de calibración y los de verosimilitud empírica. Los primeros fueron propuestos por Deville y Särndal (1992), y desde entonces se han comprobado sus propiedades teóricas, se han obtenido numerosas modificaciones, y se ha extendido el método a diversos esquemas de muestreo, siendo todos los resultados obtenidos bastante satisfactorios.

El método de verosimilitud empírica para la estimación de parámetros es más novedoso que el método de calibración. Fue propuesto en Chen y Qin (1993) para muestreo aleatorio simple, aunque el auge y el interés de esta metodología se produce en 1999 cuando Chen y Sitter plantean el método para cualquier diseño muestral. Al igual que el método de calibración, este método permite incorporar información auxiliar de una o varias variables adicionales, y se puede plantear tanto desde una perspectiva modelo-asistida, como desde la reciente aproximación modelo-calibrada (Wu y Sitter, 2001).

Los estimadores de verosimilitud empírica para la media poblacional basados en el diseño muestral y bajo la aproximación modelo-calibrada, serán vistos en la Sección 2.2. Las principales propiedades asintóticas de estos estimadores podrán también consultarse en esta sección. Nótese que el método de verosimilitud empírica usa la aproximación modelo-asistida para determinar un determinado parámetro o variable, y posteriormente se basa en el diseño muestral para determinar los estimadores. Por simplicidad y sin pérdida de generalidad, en este caso nos referiremos como aproximación modelo-asistida o aproximación basada en el diseño muestral.

Todos los métodos generales de estimación de parámetros asumen que no existen datos faltantes en la muestra. Cuando existen observaciones perdidas en la muestra, la solución más simple es eliminar aquellos individuos con observaciones incompletas y restringir el estudio a los individuos que presentan observaciones completas para todas las variables. De este modo, con este conjunto de observaciones se puede aplicar cualquier técnica de estimación de parámetros. Una consecuencia de este método es la reducción de individuos en la muestra respecto a la muestra planificada, lo que produce mayores sesgos en las estimaciones y mayor varianza muestral. Usando el método de verosimilitud empírica, en la Sección 2.3 se proponen estimadores para el problema de datos faltantes con buenas propiedades asintóticas y empíricas. Estos estimadores aprovechan todas las observaciones muestrales, estén éstas completas o incompletas para las variables del estudio.

Otro tema de actualidad en muestreo es el problema de la estimación de la función de distribución. Los estudios se han centrado clásicamente en la estimación de parámetros poblacionales de tipo puntual, como totales, medias, proporciones y varianzas. La estimación de la función de distribución es un campo muy importante al tratarse de una función que permite determinar las características más importantes de la población en estudio, proporcionando información relevante acerca del comportamiento global de la población. Obtener buenos estimadores para tal función no es tan simple como en el caso de los estimadores puntuales. Para este problema, un buen estimador, $\widehat{F}(t)$, ha de cumplir las propiedades básicas de una verdadera función de distribución:

1. $\lim_{t\to-\infty} \widehat{F}(t) = 0 \quad ; \quad \lim_{t\to+\infty} \widehat{F}(t) = 1.$
2. $\widehat{F}(t)$ es no decreciente, es decir, $\forall\ t_1 < t_2$ se verifica $\widehat{F}(t_1) \leq \widehat{F}(t_2)$.
3. Dado $t > t^*$, $\lim_{t\to t^*}, \widehat{F}(t) = \widehat{F}(t^*)$.

Varios de los estimadores propuestos en la literatura del muestreo en poblaciones finitas no satisfacen todas estas propiedades y no son, por tanto, funciones de distribución. Por ejemplo, la función de distribución estimada mediante el método de calibración no cumple los requisitos necesarios para ser una verdadera función de distribución.

En la Sección 2.4 se propone un estimador modelo-asistido para la función de distribución basado en el diseño muestral que cumple estas propiedades y goza de una excelente ganancia en eficiencia como consecuencia de un uso efectivo de la información auxiliar. Éstas son dos ventajas importantes de este estimador propuesto basado en el método de verosimilitud empírica. En esta sección, también pueden consultarse los principales estimadores de verosimilitud pseudo empírica modelo-calibrados para la función de distribución.

En resumen, este capítulo ofrece una descripción detallada del método de verosimilitud empírica en la estimación de la media o total de la población. El objetivo de este análisis es mostrar de forma sencilla cómo se construye este estimador en distintos diseños muestrales y para los distintos enfoques existentes en muestreo, cuáles son sus propiedades más importantes y la relación que tiene con otros estimadores más conocidos. Usando este

esquema teórico, se aportan nuevas soluciones al problema de los datos faltantes y a la estimación de la función de distribución.

2.2. Estimación de la media poblacional

2.2.1. Estimadores basados en el diseño muestral

La metodología de verosimilitud empírica fue usada por Owen (1988, 1990, 1991), Molina y Skinner (1992), etc, como un método para la construcción de regiones de confianza con observaciones independientes. Owen afirmó que el estadístico de verosimilitud empírica tiene una distribución asintótica χ^2, y por tanto se puede usar para la estimación de intervalos de confianza y contraste de hipótesis. Qin y Lawless (1994, 1995) usan el método de verosimilitud empírica para la estimación puntual cuando la información se incorpora a través de la maximización de la función de verosimilitud empírica. A raíz de aquí, este método se popularizó y una gran gama de desarrollos sobre verosimilitud empírica han sido descritos en el reciente libro de Owen (2001) para distintos ámbitos.

Históricamente el uso de verosimilitud empírica fue propuesto por Hartley y Rao (1968), pero la primera aplicación formal en muestreo para poblaciones finitas del método de verosimilitud empírica se debe a Chen y Qin (1993), que lo estudiaron bajo muestreo aleatorio simple.

A continuación se detalla de forma breve la idea principal del método de verosimilitud empírica para el problema de la estimación de la media muestral de y, $\overline{Y} = N^{-1}\sum_{i=1}^{N} y_i$, y para muestreo aleatorio simple. En este caso, el estimador usual es el estimador de tipo Hortviz-Thompson, dado por

$$\overline{y}_{HT} = \frac{1}{n}\sum_{i\in s} y_i = \sum_{i\in s}\frac{1}{n}y_i. \tag{2.1}$$

En la expresión (2.1) se observa que el estimador usa n puntos y_i de la muestra con el mismo peso $(1/n)$ para estimar el parámetro. Puede ocurrir que ciertas observaciones y_i sean más determinantes que otras para el calculo del parámetro. Bajo estas circunstancias es conveniente darle a las observaciones más determinantes un mayor peso que aquellas que son menos influyentes para estimar el valor del parámetro. Esta es la idea de los estimadores de verosimilitud empírica, es decir, pretenden cambiar los pesos $1/n$ por otros pesos $\widehat{p}_i$, $i = \{1,\ldots,n\}$, con el objetivo de mejorar la estimación del parámetro. Las variables auxiliares juegan un papel importante en este método, puesto que son usadas para obtener los nuevos pesos.

Sea p_i la masa de probabilidad de y_i, con $i \in s$. El estimador máximo verosímil empírico de $\overline{Y}$ se define como

$$\overline{y}_{PE} = \sum_{i\in s}\widehat{p}_i y_i,$$

donde $\widehat{p}_i$, $i = \{1,\ldots,n\}$, maximiza la función de verosimilitud empírica, $L(\mathbf{p}) = \prod_{i\in s} p_i$ sujeta a las restricciones

$$\sum_{i\in s} p_i = 1 \quad (p_i > 0), \tag{2.2}$$

$$\sum_{i\in s} p_i \mathbf{x}_i = \overline{\mathbf{X}}. \tag{2.3}$$

La información auxiliar se incorpora en la segunda restricción. Esta expresión se justifica al asumir que los pesos que dan una estimación perfecta para $\overline{\mathbf{X}}$, deberían de dar una buena precisión en la estimación de $\overline{Y}$. Resulta razonable asumir que las estimaciones serán más eficientes a medida que y y $\mathbf{x}$ presenten una relación lineal más fuerte.

Este problema de maximización con restricciones puede resolverse mediante el método de los multiplicadores de Lagrange. Véase también, por ejemplo, Aitchison y Silvey (1958), Hall (1990) y Hall y La Scala (1990).

Los estimadores de verosimilitud empírica se pueden diseñar desde distintas perspectivas, siendo el investigador quien debe decidir el modo de aplicar el método de verosimilitud empírica. Algunos de los distintos enfoques a través de los cuales se puede diseñar esta metodología son los siguientes:

(E1). Sustitución de $L(\mathbf{p})$.

En Chen y Qin (1993) se usa la función $L(\mathbf{p})$ para obtener los estimadores de verosimilitud empírica, mientras que Chen y Sitter (1999) usaron el logaritmo de esta función a nivel poblacional, esto es, propusieron usar

$$l(\mathbf{p}) = \log\prod_{i=1}^{N} p_i = \sum_{i=1}^{N}\log(p_i).$$

Notamos que el hecho de utilizar logaritmos no produce ningún cambio en las estimaciones al tratarse la función logaritmo de una función estrictamente creciente que conserva los puntos extremos de la función original. La ventaja es una mayor facilidad para obtener estimaciones. El problema que se plantea es cómo estimar $l(\mathbf{p})$ a través de una función eficiente $\widehat{l}(\mathbf{p})$. Tomando $\log(p_i)$ como una variable de la que se pretende estimar su total, este planteamiento presenta fácil solución. Como se detalla en Chen y Sitter (1999) y para un determinado diseño muestral general, $l(\mathbf{p})$ se puede estimar a través de la denominada log-función de verosimilitud pseudo empírica, dada por:

$$\widehat{l}(\mathbf{p}) = \sum_{i\in s} d_i\log(p_i),$$

donde d_i son pesos básicos que hacen que $\widehat{l}(p)$ sea insesgada bajo el diseño para $l(\mathbf{p})$, es decir

$$E\left[\widehat{l}(\mathbf{p})\right] = E\left[\sum_{i\in s} d_i\log(p_i)\right] = \sum_{i=1}^{N}\log(p_i) = l(\mathbf{p}).$$

Este cambio en la función de verosimilitud empírica hace que esta técnica sea aplicable bajo un diseño muestral general, a diferencia del método original propuesto por Chen y Qin (1993) que está diseñado

exclusivamente para muestreo aleatorio simple. Bajo este método de muestreo, ambas perspectivas del método de verosimilitud empírica producen las mismas estimaciones.

(E2). Sustitución de la restricción $\sum_{i\in s} p_i \mathbf{x}_i = \overline{\mathbf{X}}$.

Al imponer que $\sum_{i\in s} p_i \mathbf{x}_i = \overline{\mathbf{X}}$, se están considerando valores para p_i que proporcionan estimaciones perfectas para $\overline{\mathbf{X}}$, y podemos plantearnos cómo de efectivo es el uso que se está haciendo de la información adicional a través de la condición anterior. Por este motivo, si la información auxiliar $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_P)$ es conocida, una cuestión a preguntarse sería: *¿Cual es la mejor expresión a usar en la restricción (2.3) para hallar el estimador de verosimilitud empírica?* . Para resolver esta pregunta se ha definido la cantidad $\mathbf{u}_i = u(y_i, \mathbf{x}_i)$, con $i = \{1, \ldots, N\}$, siendo $u(\cdot)$ una función conocida de y_i y de $\mathbf{x}_i$ y que verifica

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{u}_i = \mathbf{0}.$$

De este modo, $\mathbf{u}_i$ es una variable de calibración que reemplaza la expresión (2.3) por

$$\sum_{i\in s} p_i \mathbf{u}_i = \frac{1}{N}\sum_{i=1}^{N} \mathbf{u}_i = \mathbf{0}, \qquad (2.4)$$

donde $\mathbf{u}_i = \mathbf{x}_i - \overline{\mathbf{X}}$. La cuestión que surge ahora es cómo escoger $u(\cdot)$ para obtener estimadores más eficientes. En resumen, este método dispone de numerosas alternativas o soluciones dependiendo de la función $u(\cdot)$ escogida. Una elección apropiada de esta función supondrá más exactitud en las estimaciones. El uso de la aproximación modelo-calibrada es una solución óptima a este problema cuando no pueda asumirse una relación lineal entre y y $\mathbf{x}$.

(E3). Utilización de la aproximación modelo-calibrada.

En (E2) se usa una aproximación modelo-asistida, esto es, se asume una relación lineal (aunque pueden establecerse relaciones de otro tipo) para determinar unos valores $\mathbf{u}_i$ apropiados, y posteriormente, se realizan estimaciones basadas en el diseño. Si la relación entre la variable de interés y y el vector de variables auxiliares $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_P)$ puede ser descrita a través de un modelo de superpoblación con una buena bondad de ajuste, puede resultar útil el uso de estimadores modelo-calibrados (Wu y Sitter, 2001) frente a los estimadores basados en el diseño. Esta aproximación consiste en asumir un determinado modelo de superpoblación, obtener los valores estimados para la variable y mediante este modelo, y a continuación usarlos en la etapa de estimación.

En este sentido, se han propuesto varios modelos que dan lugar a los estimadores óptimos modelo-calibrados. Éstos usan el criterio de mínima esperanza bajo el modelo de superpoblación de la varianza basada en el diseño para obtener la solución óptima (véase por ejemplo los trabajos de Godambe, 1955, Godambe y Thompson, 1973 y Cassel *et al.*, 1976). Los estimadores modelo-calibrados se desarrollan con detalle en la Sección 2.2.3.

La perspectiva dada en Chen y Sitter (1999) es más apropiada como se ha comprobado en las investigaciones posteriores. Además, puede ser aplicada a cualquier diseño muestral, no estando limitada exclusivamente al muestreo aleatorio simple. De este modo, los primeros pasos antes de aplicar el método de verosimilitud empírica son:

1. Enfocar el problema bajo un modelo de población fija, es decir, basado en el diseño muestral y aplicando la aproximación modelo-asistida, o bien, asumir un modelo de superpoblación para poder aplicar el enfoque modelo-calibrado.
2. Determinar la función $u(\cdot)$ utilizada en la restricción (2.4). Para el enfoque basado en el diseño muestral se suele usar $\mathbf{u}_i = \mathbf{x}_i - \overline{\mathbf{X}}$, mientras que bajo el enfoque modelo-calibrado, la función $u(\cdot)$ es única y fácilmente deducible a partir del modelo de superpoblación.

Estimadores bajo muestreo aleatorio simple

Una vez tenidas en cuenta estas consideraciones previas, empezaremos analizando el método de verosimilitud empírica según Chen y Qin (1993), el cual está diseñado para muestreo aleatorio simple.

Este estimador fue la primera aplicación formal del método de verosimilitud empírica en poblaciones finitas para la estimación de parámetros lineales y usando información auxiliar. Este planteamiento no se puede extender a diseños muestrales más complejos.

Según Chen y Qin (1993), el uso de verosimilitud empírica en el contexto de poblaciones finitas se puede plantear de dos formas diferentes:

1. Si todos los valores de y_i están disponibles para la población en estudio, la función de verosimilitud se define como $L^*(\mathbf{p}) = \prod_{i=1}^{N} p_i$, donde p_i es la densidad de la observación y_i. En la práctica esta situación no se va a presentar y lo más usual es que y_i sea conocida para los individuos de la muestra s. En tal caso la función de verosimilitud empírica para cualquier muestra s, con $s \subseteq S$, se define como $L(\mathbf{p}) = \prod_{i\in s} p_i$, donde se requiere que $\sum_{i=1}^{n} p_i \leq 1$. Este planteamiento fue propuesto por Jagers (1986) y es el que se sigue en varios estudios de estimación de parámetros en muestreo de poblaciones finitas mediante verosimilitud empírica (Chen y Qin, 1993, Zhong y Rao, 1996, etc).
2. Según el esquema de muestreo propuesto por Hartley y Rao (1968), los cuales consideraban que la variable de interés sólo puede tomar un número finito de valores, es decir, y_i, con $i = \{1, \ldots, I\}$. Bajo esta situación, la población media se define como:

$$\overline{Y} = \sum_{i=1}^{I} \frac{N_i}{N} y_i,$$

donde N_i es el número de unidades en la población con característica y_i. Bajo muestreo aleatorio simple de tamaño n, la verosimilitud basada en el diseño está dada por una distribución hipergeométrica multidimensional:

$$L(N_1,\ldots,N_I)=\prod_{i=1}^{I}\frac{\left(\begin{array}{c}N_i\\n_i\end{array}\right)}{\left(\begin{array}{c}N\\n\end{array}\right)},$$

donde n_i es el número de unidades en la muestra con la característica y_i. Cuando $N\to\infty$, $N_i/N\to p_i$, y $n/N\to 0$, la verosimilitud se puede aproximar por una función de verosimilitud de una distribución multinomial, a saber:

$$\frac{n!}{\prod_{i=1}^{I}n_i!}\prod_{i=1}^{I}p_i^{n_i}.$$

Utilizando el primer planteamiento propuesto por Jagers (1986), al maximizar $L(\mathbf{p})$ sin usar información auxiliar, resulta $\widehat{p}_i=1/n$ para cada $i\in s$, y el estimador de verosimilitud empírica está dado por

$$\overline{y}_{EL}=\sum_{i\in s}\widehat{p}_i y_i=\frac{1}{n}\sum_{i\in s}y_i=\overline{y}_{HT},$$

coincidiendo con el estimador directo usual para la media poblacional.

Cuando se dispone de alguna información auxiliar, ésta puede usarse en la etapa de maximización de la función de verosimilitud para obtener nuevos pesos p_i que produzcan estimaciones más eficientes para la media. Se asume que la información auxiliar disponible para la población verifica

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{u}_i=\mathbf{0},$$

donde $\mathbf{u}_i=u(y_i,\mathbf{x}_i)$ es una función conocida de y_i y de $\mathbf{x}_i$ de vectores valuados. De este modo, el nuevo problema consiste en maximizar $L(\mathbf{p})$ sujeto a las restricciones:

$$\sum_{i\in s}p_i=1\quad(p_i\geq 0),\tag{2.5}$$

$$\sum_{i\in s}p_i\mathbf{u}_i=\mathbf{0}.\tag{2.6}$$

Usando el método de los multiplicadores de Lagrange, los valores esperados para p_i, con $i\in s$, están dados por:

$$\widehat{p}_i^{*}=\frac{1}{n\,(1+\lambda^t\mathbf{u}_i)},\tag{2.7}$$

donde λ es la solución de la ecuación

$$\sum_{i\in s}\frac{\mathbf{u}_i}{1+\lambda^t\mathbf{u}_i}=\mathbf{0}.\tag{2.8}$$

El estimador de verosimilitud empírica para la media poblacional bajo muestreo aleatorio simple y usando la metodología de Chen y Qin (1993) está dado por

$$\overline{y}_{EL}=\sum_{i\in s}\widehat{p}_i^{*}y_i.\tag{2.9}$$

Asumiendo que la relación entre y y el vector $\mathbf{x}$ es lineal, la función de calibración usual viene dada por $\mathbf{u}_i=\mathbf{x}_i-\overline{\mathbf{X}}$, en cuyo caso la restricción (2.6) resulta ser

$$\sum_{i\in s}p_i\mathbf{u}_i=\sum_{i\in s}p_i(\mathbf{x}_i-\overline{\mathbf{X}})=$$

$$=\sum_{i\in s}p_i\mathbf{x}_i-\sum_{i\in s}p_i\overline{\mathbf{X}}=\sum_{i\in s}p_i\mathbf{x}_i-\overline{\mathbf{X}}=\mathbf{0}\Rightarrow$$

$$\Rightarrow\sum_{i\in s}p_i\mathbf{x}_i=\overline{\mathbf{X}},\tag{2.10}$$

que indica que las cantidades p_i dan estimaciones perfectas para $\overline{\mathbf{X}}$, y por tanto, deberían dar una buena aproximación para la media de variable de interés si la relación entre y y $\mathbf{x}$ es lineal.

Cuando $\mathbf{u}_i=\mathbf{x}_i-\overline{\mathbf{X}}$, las soluciones a las ecuaciones (2.7) y (2.8) también son obtenidas por Hartley y Rao (1968) a través de una aproximación similar. Estos autores demostraron que el estimador de regresión es asintóticamente equivalente al estimador dado en (2.9). Un resultado similar puede hacerse para el estimador de la mediana propuesto por Kuk y Mak (1989) cuando $\mathbf{u}_i=\delta(\mathbf{x}\leq M_{\mathbf{x}})-0{,}5$, siendo $M_{\mathbf{x}}$ la mediana de $\mathbf{x}$, y $\delta(\cdot)$ la función indicadora que toma el valor $\delta(a)=1$ si $a\geq 0$ y el valor 0 en otro caso.

Puede ocurrir que la ecuación (2.8) no tenga solución. Esta situación surge cuando el conjunto convexo $\{\mathbf{u}_i,i\in s\}$ no contiene al $\mathbf{0}$. Se han planteado dos soluciones para este problema:

1. Usar la verosimilitud euclídea propuesta por Owen (1991):
$$\frac{1}{2}\sum_{i\in s}(1-np_i)^2,$$
y no requerir que $0\leq p_i\leq 1$.
2. Reemplazar la restricción (2.6) por
$$\sum_{i\in s}p_i\mathbf{u}_i=\widetilde{\mathbf{u}},$$
tal que $\widetilde{\mathbf{u}}$ está dentro del conjunto convexo y tiende a $\mathbf{0}$.

En cualquier caso, cuando n es grande, la situación en la cual la ecuación (2.8) no tiene solución es poco probable.

Existen situaciones extremas en las cuales el método de verosimilitud empírica es incapaz de usar la información auxiliar, como por ejemplo, cuando $\mathbf{x}$ es dicotómica y todas las observaciones son $\mathbf{x}_i=1$. Estos casos también son poco probables en la práctica.

Estimadores bajo un diseño muestral general

El estimador del apartado anterior esta diseñado sólo para muestreo aleatorio simple, y su metodología no se puede extender a otros diseños muestrales más complejos. Chen y Sitter (1999) proponen una aproximación de verosimilitud pseudo empírica que es aplicable a cualquier diseño muestral y coincide bajo muestreo aleatorio simple con el estimador propuesto en Chen y Qin (1993).

El método de verosimilitud empírica para un diseño muestral general asume que la muestra s es seleccionada usando algún diseño muestral, $p(\cdot)$, es decir, la muestra

$s \subseteq S$ es extraída con probabilidad $p(s)$. El objetivo es maximizar la verosimilitud de la población en estudio, es decir, maximizar $L^*(\mathbf{p}) = \prod_{i=1}^{N} p_i$. Por conveniencia, y teniendo en cuenta la monotonía de la función logaritmo, se considera el objetivo de maximizar $l(\mathbf{p}) = \log L^*(\mathbf{p}) = \sum_{i=1}^{N} \log p_i$. En la práctica, solo de disponen de los valores y_i para las unidades de la muestra, pudiéndose, por tanto, utilizar únicamente las cantidades p_i para $i \in s$. Esto provoca que se necesite una estimación eficiente para $l(\mathbf{p})$. Esta estimación viene dada por la llamada función de verosimilitud pseudo empírica

$$\widehat{l}(\mathbf{p}) = \sum_{i \in s} d_i \log p_i, \tag{2.11}$$

que tiene la propiedad de ser una estimación insesgada bajo el diseño para $l(\mathbf{p})$, esto es

$$E[\widehat{l}(\mathbf{p})] = E\left[\sum_{i \in s} d_i \log p_i\right] = \sum_{i=1}^{N} \log p_i = l(\mathbf{p}),$$

donde $E[\cdot]$ denota la esperanza bajo el diseño muestral.

La información auxiliar se incorpora a través de la función de calibración $\mathbf{u}_i = u(y_i, \mathbf{x}_i)$, donde $u(\cdot)$ es una función de y_i y de $\mathbf{x}_i$ que debe satisfacer:

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{u}_i = \mathbf{0}.$$

Las cantidades $\widehat{p}_i$ necesarias para obtener el *estimador de verosimilitud pseudo empírica* ($PEMLE$) se obtienen maximizando la función dada en (2.11) sujeta a las restricciones (2.5) y (2.6).

Usando el método de los multiplicadores de Lagrange para resolver este problema, se obtiene, para $i \in s$, las cantidades

$$\widehat{p}_i = \frac{d_i^*}{1 + \lambda^t \mathbf{u}_i}, \tag{2.12}$$

donde el vector de multiplicadores de Lagrange, λ, es la solución de la expresión:

$$\sum_{i \in s} \frac{d_i^* \mathbf{u}_i}{1 + \lambda^t \mathbf{u}_i} = \mathbf{0}, \tag{2.13}$$

siendo $d_i^* = d_i / \sum_{j \in s} d_j$. El $PEMLE$ para la media poblacional se define entonces como

$$\overline{y}_{PE} = \sum_{i \in s} \widehat{p}_i y_i. \tag{2.14}$$

Se recuerda que asumiendo una relación lineal entre y y $\mathbf{x}$ se suele considerar la función de calibración $\mathbf{u}_i = \mathbf{x}_i - \overline{\mathbf{X}}$. En este caso, la restricción (2.6) puede expresarse como:

$$\sum_{i \in s} p_i \mathbf{x}_i = \overline{\mathbf{X}}.$$

En el caso de no disponer de información auxiliar, en cuyo caso se toma $\mathbf{u}_i = \mathbf{0}$, el método de verosimilitud empírica produce $\widehat{p}_i = d_i^*$, y el $PEMLE$ viene dado por

$$\overline{y}_{PE} = \sum_{i \in s} d_i^* y_i,$$

que coincide con el estimador directo para la media poblacional de tipo Hájek. En general, este estimador no coincide con el estimador directo usual de tipo Horvitz-Thompson, aunque se demuestra que disfruta de buenas propiedades respecto a este último (véase Rao, 1966, Basu, 1971 y Särndal *et al.*, 1992). Respecto al problema de la estimación de la función de distribución, el estimador de tipo Hájek disfruta de mejores propiedades, puesto que el estimador de tipo Horvitz-Thompson no cumple las propiedades para ser una verdadera función de distribución (en concreto $\lim_{t \to +\infty} \widehat{F}_{HTy}(t) \neq 1$), propiedades que si posee el estimador de tipo Hájek.

Esta propiedad para la función de distribución también se cumple para cualquier función de calibración, y no tan solo para $\mathbf{u}_i = \mathbf{0}$. Esto es, las cantidades $\widehat{p}_i$ dadas en (2.12) son estrictamente positivas y satisfacen $\sum_{i \in s} \widehat{p}_i = 1$ (como puede comprobarse en (2.5)), condiciones necesarias para estimar una verdadera función de distribución, hecho que no sucede, por ejemplo, con los estimadores de regresión generalizados ($GREG$) definidos en Cassel *et al.* (1976) y Särndal (1980) o los estimadores de calibración propuestos en Deville y Särndal (1992).

A continuación, se dan expresiones del $PEMLE$ para algunos diseños muestrales más simples y conocidos. De estos ejemplos se desprende que la aplicabilidad de esta metodología no es tan complicada y que estos estimadores están relacionados con otros estimadores tradicionales.

Ejemplo 2.1 ***Muestreo Aleatorio Simple.***

Bajo este diseño $\pi_i = n/N$, $d_i = 1/\pi_i = N/n$ y $\sum_{j \in s} d_j = N$, obteniéndose

$$d_i^* = \frac{d_i}{\sum_{j \in s} d_j} = \frac{1}{n}. \tag{2.15}$$

Si no se dispone de información auxiliar, $\mathbf{u}_i = \mathbf{0}$, $\widehat{p}_i = d_i^*$ y el $PEMLE$ para la media poblacional está dado por

$$\overline{y}_{PE} = \sum_{i \in s} \widehat{p}_i y_i = \frac{1}{n} \sum_{i \in s} y_i, \tag{2.16}$$

que coincide con el estimador usual bajo muestreo aleatorio simple ($\overline{y}_{HT}$) y con el estimador $\overline{y}_{EL}$ propuesto en Chen y Qin (1993).

Usando la información auxiliar, el $PEMLE$ viene dado por

$$\overline{y}_{PE} = \sum_{i \in s} \widehat{p}_i y_i, \tag{2.17}$$

donde

$$\widehat{p}_i = \frac{1}{n(1 + \lambda^t \boldsymbol{u}_i)}, \tag{2.18}$$

y λ es la solución de la ecuación

$$\sum_{i \in s} \frac{\boldsymbol{u}_i}{1 + \lambda^t \boldsymbol{u}_i} = \boldsymbol{0}. \tag{2.19}$$

Puede observarse que este estimador coincide, de nuevo, con el estimador $\overline{y}_{EL}$.

Ejemplo 2.2 ***Muestreo con probabilidades iguales y con reemplazamiento.***

En los métodos de muestreo con reemplazamiento se demuestra (véase Han-sen y Hurwitz, 1943) que $d_i = 1/(n\alpha_i)$, donde α_i es la probabilidad de que la

unidad i-ésima sea seleccionada. Además, al tratarse de un muestreo con probabilidades iguales se tiene que $\alpha_i = 1/N$ y por tanto $d_i = N/n$, que coincide con los pesos básicos en un muestreo aleatorio simple. En consecuencia, las expresiones (2.15), (2.16), (2.17), (2.18) y (2.19) coinciden en este diseño. La única diferencia está en la muestra, es decir, el método para seleccionarla es distinto y además aquí es posible tener unidades repetidas.

Ejemplo 2.3 ***Muestreo con probabilidades desiguales y sin reemplazamiento.***

Se tiene que $d_i = 1/\pi_i$,

$$\widehat{p}_i = \frac{d_i^*}{1+\lambda^t u_i}, \text{ donde } d_i^* = \frac{1/\pi_i}{\sum_{j\in s} 1/\pi_j},$$

y λ es solución de la ecuación (2.13). Sabido esto, el $PEMLE$ se construye según (2.14).

Bajo este muestreo existen muchos procedimientos para extraer una muestra (consúltese, por ejemplo, Chaudhuri y Vos, 1988). Todos ellos poseen expresiones que permiten calcular las cantidades π_i, necesarias para obtener el $PEMLE$. En este texto se usan los métodos de Lahiri, Midzuno y Poisson (véase Lahiri, 1951, Midzuno, 1952, Hájek, 1964, Ogus y Clark, 1971, Singh, 2003, etc). En el Apéndice **??** *pueden consultarse funciones en el lenguaje de programación R que permiten extraer muestras basadas en estos procedimientos de muestreo con probabilidades desiguales.*

Ejemplo 2.4 ***Muestreo con probabilidades desiguales y con reemplazamiento.***

Es sabido que en este caso $d_i = 1/(n\alpha_i)$, donde α_i es la probabilidad de que la unidad i-ésima sea seleccionada en cada extracción y por tanto

$$d_i^* = \frac{d_i}{\sum_{j\in s} d_j} = \frac{1/\alpha_i}{\sum_{j\in s} 1/\alpha_j}. \qquad (2.20)$$

Y así, el $PEMLE$ se construye mediante la expresión (2.14). En el caso particular de usar el tamaño de cada unidad como una variable auxiliar para la asignación de probabilidades, se tiene que $\alpha_i = M_i/M$, donde M_i es el tamaño de la unidad i, y $M = \sum_{i=1}^N M_i$. Sustituyendo este valor en la expresión (2.20), se obtiene una expresión más simple para el $PEMLE$.

Una cuestión sin resolver hasta el momento es el procedimiento a seguir para despejar λ en la expresión (2.13), donde además, se ha de verificar que las cantidades $\widehat{p}_i$ sean positivas. La resolución de este problema no es tan simple al tratarse de ecuaciones no lineales, debiéndose emplear métodos específicos para la resolución de ecuaciones no lineales, como el de bisección o el de Newton-Raphson. A continuación se describe una modificación del método de Newton-Raphson, propuesto en Chen *et al.* (2002), para el calculo del $PEMLE$ en caso de que este problema tenga una única solución y ésta exista.

Sea

$$g(\lambda) = \sum_{i\in s} \frac{d_i^* \mathbf{u}_i}{1+\lambda^t \mathbf{u}_i}.$$

Para una muestra dada, s, el conjunto de valores factibles de λ tal que $\widehat{p}_i > 0$ está dado por el conjunto convexo $A = \{\lambda : 1+\lambda^t \mathbf{u}_i > 0, i \in s\}$. El problema de maximizar la función $\widehat{l}(\mathbf{p})$, definida en (2.11), sujeta a las restricciones (2.5) y (2.6) es similar al problema de maximizar la función cóncava

$$\widetilde{l}(\lambda) = \sum_{i\in s} d_i^* \log(1+\lambda^t \mathbf{u}_i),$$

con respecto a λ sobre el conjunto convexo A, puesto que $\partial\widetilde{l}(\lambda)/\partial\lambda = g(\lambda)$. Si la única solución de $g(\lambda) = 0$ existe, ésta puede encontrarse aplicando la siguiente modificación del algoritmo de Newton-Raphson:

Algoritmo 2.1

Paso 0: *Sea $\lambda_0 = \mathbf{0}$, $k = 0$, $\gamma_0 = 1$ y $\epsilon = 10^{-8}$.*

Paso 1: *Calcular $\Delta(\lambda_k)$ donde*

$$\Delta(\lambda) = \left\{ \frac{\partial}{\partial\lambda} g^*(\lambda) \right\}^{-1} \quad ;$$

$$g^*(\lambda) = \left\{ -\sum_{i\in s} \frac{d_i^* \boldsymbol{u}_i \boldsymbol{u}_i^t}{(1+\lambda^t \boldsymbol{u}_i)^2} \right\}^{-1} \sum_{i\in s} \frac{d_i^* \boldsymbol{u}_i}{1+\lambda^t \boldsymbol{u}_i}.$$

Si $\|\Delta(\lambda_k)\| < \epsilon$, se detiene el algoritmo y la solución es λ_k. En otro caso ir al Paso 2

Paso 2: *Calcular $\delta_k = \gamma_k \Delta(\lambda_k)$. Si $1 + (\lambda_k - \delta_k)^t \boldsymbol{u}_i \leq 0$ para algún i o $\widetilde{l}(\lambda_k - \delta_k) < \widetilde{l}(\lambda_k)$, entonces tomar $\gamma_k = \gamma_k/2$ y repetir el Paso 2.*

Paso 3: *Considerar $\lambda_{k+1} = \lambda_k - \delta_k$, $k = k+1$ y $\gamma_{k+1} = (k+1)^{-1/2}$. Ir al Paso 1.*

La expresión $\|\cdot\|$ denota la norma euclídea.

La demostración de este resultado puede consultarse en Chen *et al.* (2002). Así mismo, puede comprobarse que este algoritmo es similar a la modificación del método de Newton descrito en Polyak (1987). Los cambios del paso 2 aseguran que en cada iteración el valor de λ sigue dentro del rango de A y que la función cóncava $\widetilde{l}(\lambda)$ se mueve alrededor del punto máximo. El algoritmo es simple, eficiente y la convergencia está garantizada, lo cual indica que, salvo en casos extraños, el $PEMLE$ puede siempre obtenerse.

Estimadores bajo muestreo estratificado

La metodología de verosimilitud empírica para obtener estimadores en muestreo de poblaciones finitas se extiende a diseños muestrales más complejos, como por ejemplo muestreo estratificado. Siguiendo la notación clásica del muestreo estratificado, se define la log-función de verosimilitud en muestreo estratificado como

$$l(\mathbf{p}) = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \log(p_{hi}), \qquad (2.21)$$

que puede verse como un total poblacional, cuya estimación insesgada a partir de la muestra s y bajo un diseño muestral específico está dada por

$$\widehat{l}(\mathbf{p}) = \sum_{h=1}^{L} \sum_{i\in s_h} d_{hi} \log(p_{hi}). \qquad (2.22)$$

En este caso, d_{hi} son los pesos diseñados básicos que hacen que $\widehat{l}(\mathbf{p})$, denominada log-función de verosimilitud pseudo empírica, sea insesgada bajo el diseño para $l(\mathbf{p})$. Por ejemplo, asumiendo muestreo aleatorio simple en cada estrato, se tiene $d_{hi} = N_h/n_h$.

En muestreo estratificado, el $PEMLE$ se obtiene maximizando la función (2.22) sujeta a las restricciones

$$\sum_{i\in s_h} p_{hi} = 1 \quad (p_{hi} > 0), \quad h = \{1,\ldots,L\}, \tag{2.23}$$

$$\sum_h W_h \sum_{i\in s_h} p_{hi}\mathbf{x}_{hi} = \overline{\mathbf{X}}. \tag{2.24}$$

En la restricción (2.24) se ha considerado por comodidad una relación lineal entre y y $\mathbf{x}$, aunque es posible modificar esta restricción en caso de existir o considerar oportuno asumir otro tipo de relación entre y y $\mathbf{x}$.

Una vez obtenidas todas las soluciones $\widehat{p}_{hi}$ de este problema, el $PEMLE$ bajo muestreo estratificado está dado por

$$\overline{y}_{PEst} = \sum_{h=1}^{L} W_h \sum_{i\in s_h} \widehat{p}_{hi} y_{hi}. \tag{2.25}$$

Dependiendo de si las cantidades $\overline{\mathbf{X}}_h = N_h^{-1}\sum_{i=1}^{N_h}\mathbf{x}_{hi}$ son conocidas o no, el cálculo de este estimador se puede orientar en dos caminos distintos.

En primer lugar, si las cantidades $\overline{\mathbf{X}}_h$ son conocidas para $h = \{1,\ldots,L\}$, y asumiendo una relación lineal, la restricción (2.24) puede sustituirse por la restricción

$$\sum_{i\in s_h} p_{hi}\mathbf{x}_{hi} = \overline{\mathbf{X}}_h, \quad h = \{1,\ldots,L\}, \tag{2.26}$$

y el problema que se plantea en este caso es maximizar (2.22) sujeta a las restricciones (2.23) y (2.26). Según este planteamiento, el calculo del $PEMLE$ bajo muestreo estratificado es bastante simple, esto es, se calcula el $PEMLE$ para cada estrato, $\overline{y}_{PEh}$, y el estimador final viene dado por

$$\overline{y}_{PEst} = \sum_{h=1}^{L} W_h \overline{y}_{PEh}.$$

Por otro lado, cuando $\overline{\mathbf{X}}_h$ son desconocidas para cualquier h, la restricción (2.26) no puede establecerse, y el problema de maximizar (2.22) sujeto a las restricciones (2.23) y (2.24) no es una cuestión tan simple. Incluso resulta imposible aplicar el Algoritmo 2.1 bajo muestreo estratificado debido a que la función (2.22) y la restricción (2.24) están formuladas para el conjunto de los estratos, esto es, contienen dobles sumatorias, mientras que la restricción (2.23) está formulada a nivel del estrato, es decir, contiene una sola sumatoria. Existen dos estrategias a seguir para buscar una solución óptima:

(G1). En lugar de la restricción (2.24), considerar otra restricción arbitraria para cada estrato y buscar la solución intermedia bajo esta situación. La solución final se obtiene a través del método de verosimilitud empírica.

(G2). Reemplazar las restricciones de modo que las nuevas estén todas formuladas a nivel del conjunto de los estratos, y por tanto el Algoritmo 2.1 pueda ser aplicado.

La estrategia (G1) fue seguida por Chen y Sitter (1999). El planteamiento que se propuso es el siguiente. El $PEMLE$ bajo muestreo estratificado se calcula considerando los pesos $\widehat{p}_{hi}$ obtenidos al maximizar la función (2.22) sujeta a las restricciones

$$\begin{aligned} &\textstyle\sum_h \sum_{i\in s_h} p_{hi} = 1, \\ &\textstyle\sum_h \sum_{i\in s_h} p_{hi}\mathbf{x}_{hi} = \overline{\mathbf{X}}. \end{aligned} \tag{2.27}$$

Estas restricciones surgen al incorporar la información auxiliar contenida en el tamaño de cada estrato, es decir, toda la información auxiliar usada para construir el $PEMLE$ se puede incluir en los vectores $\mathbf{u}_i = \mathbf{U}_i^* - \overline{\mathbf{U}}^*$, donde $i = \{1,\ldots,N\}$, $\mathbf{U}_i^* = (\mathbf{x}_i, \nu_{1i},\ldots,\nu_{Li})^t$, $\overline{\mathbf{U}}^* = (\overline{\mathbf{X}}, W_1,\ldots,W_L)^t$ y ν_{hi} vale 1 si $i \in h$ y 0 en otro caso. En este sentido, la información de los tamaños de los estratos se usa de forma efectiva, lo cual no ocurre ni con el estimador de regresión generalizado $(GREG)$ ni con el estimador óptimo de regresión (ORE) propuesto en Rao (1994), y esto hace que se obtengan mejores estimaciones. A su vez, bajo muestreo estratificado, el ORE es más eficiente que el $GREG$ porque usa la correlación entre y y $\mathbf{x}$. Asumiendo muestreo estratificado aleatorio, el $PEMLE$ es equivalente al ORE (y ambos mejores que el $GREG$) puesto que los pesos muestrales son constantes dentro de cada estrato e incluyen el tamaño del estrato que es equivalente a incluir la correlación. No obstante, asumiendo otro diseño muestral, por ejemplo muestreo estratificado con probabilidades proporcionales al tamaño en cada estrato, el $PEMLE$ es más eficiente que el ORE debido a que usa los tamaños de los estratos que contienen información importante que no es suministrada ni por los pesos muestrales ni por la correlación. En resumen, bajo muestreo estratificado, el $PEMLE$ gana en eficiencia respecto a otros estimadores (véase, por ejemplo, Chen y Sitter, 1999, Zhong y Rao, 1996, Zhong y Rao, 2000).

Según lo descrito, se ha de resolver el problema de maximizar (2.22) sujeta a las restricciones (2.27). Como las restricciones

$$\begin{aligned} &\textstyle\sum_{i\in s_h} p_{hi} = W_h, \quad \forall\, h = \{1,\ldots,L\}, \\ &\textstyle\sum_{i\in s_h} p_{hi}\mathbf{x}_{hi} = W_h\widetilde{\mathbf{x}}_h, \quad \forall\, h = \{1,\ldots,L\}, \end{aligned} \tag{2.28}$$

son equivalentes a las dadas en (2.27), el problema se resuelve buscando las cantidades

$$\widetilde{\mathbf{x}}_h, \quad h = \{1,\ldots,L\}, \tag{2.29}$$

tal que $\sum_h W_h\widetilde{\mathbf{x}}_h = \overline{\mathbf{X}}$ y maximizando (2.22) sujeta a las nuevas restricciones (2.28). Aplicando el método de los multiplicadores de Lagrange, la solución que se obtiene es

$$\widehat{p}_{hi} = \frac{W_h d_{hi}}{d_h + \lambda_h^t(\mathbf{x}_{hi} - \widetilde{\mathbf{x}}_h)}, \tag{2.30}$$

donde λ_h para $h = \{1,\ldots,L\}$, se obtiene de la ecuación

$$\sum_{i\in s_h} \frac{d_{hi}(\mathbf{x}_{hi} - \widetilde{\mathbf{x}}_h)}{d_h + \lambda_h^t(\mathbf{x}_{hi} - \widetilde{\mathbf{x}}_h)} = \mathbf{0}, \tag{2.31}$$

y $d_h = \sum_{i \in s_h} d_{hi}$. Sabido esto, el valor máximo para la función (2.22) es

$$\sum_h \sum_{i \in s_h} d_{hi} \log(\widehat{p}_{hi}) =$$

$$= -\sum_h \sum_{i \in s_h} d_{hi} \log\left[d_h + \lambda_h^t(\mathbf{x}_{hi} - \widetilde{\mathbf{x}}_h)\right] + \qquad (2.32)$$

$$+ \sum_h \sum_{i \in s_h} d_{hi} \left[\log(d_{hi}) + \log(W_h)\right]. \qquad (2.33)$$

Como (2.33) es constante, se puede maximizar (2.32) respecto a $\widetilde{\mathbf{x}}_h$ y bajo la condición $\sum_h W_h \widetilde{\mathbf{x}}_h = \overline{\mathbf{X}}$. Notamos que λ_h es una función que depende de $\widetilde{\mathbf{x}}_h$. Usando de nuevo el método de Lagrange, se tiene

$$l(\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_L, \mathbf{t}) =$$

$$-\sum_h \sum_{i \in s_h} d_{hi} \log\left[d_h + \lambda_h^t(\mathbf{x}_{hi} - \widetilde{\mathbf{x}}_h)\right] - \mathbf{t}^t \left(\sum_{h=1}^L W_h \widetilde{\mathbf{x}}_h - \overline{\mathbf{X}}\right).$$

Tomando derivadas respecto a $\widetilde{\mathbf{x}}_h$ e igualando al vector de ceros se obtiene

$$-\sum_{i \in s_h} \frac{d_{hi}\left[\dfrac{\partial \lambda_h^t}{\partial \widetilde{\mathbf{x}}_h}(\mathbf{x}_{hi} - \widetilde{\mathbf{x}}_h) - \lambda_h^t\right]}{d_h + \lambda_h^t(\mathbf{x}_{hi} - \widetilde{\mathbf{x}}_h)} - \mathbf{t}^t W_h = -\lambda_h^t - \mathbf{t}^t W_h = \mathbf{0},$$

y por tanto $\lambda_h^t = W_h \mathbf{t}^t$. La expresión (2.31) puede expresarse como

$$\sum_{i \in s_h} \frac{d_{hi}(\mathbf{x}_{hi} - \widetilde{\mathbf{x}}_h)}{d_h + W_h \mathbf{t}^t(\mathbf{x}_{hi} - \widetilde{\mathbf{x}}_h)} = \mathbf{0}. \qquad (2.34)$$

Debido a estos desarrollos, puede emplearse el siguiente algoritmo para la búsqueda de los pesos $\widehat{p}_{hi}$ necesarios para obtener el $PEMLE$ en muestreo estratificado.

Algoritmo 2.2

Paso 1. *Fijar un vector* $\mathbf{t}$ *y obtener las cantidades* $\widetilde{\mathbf{x}}_h$, $h = \{1, \ldots, L\}$, *soluciones de la expresión (2.34).*

Paso 2. *Si* $\sum_h W_h \widetilde{\mathbf{x}}_h = \overline{\mathbf{X}}$, *se calculan las cantidades* $\widehat{p}_{hi}$ *según (2.30), donde* $\lambda_h = W_h \mathbf{t}$. *En caso contrario, elegir otro* $\mathbf{t}$ *y volver al paso anterior.*

Una vez calculadas las cantidades $\widehat{p}_{hi}$, con $i \in s_h$ y $h = \{1, \ldots, L\}$, mediante el algoritmo anterior, el $PEMLE$ está dado por

$$\overline{y}_{PE} = \sum_{h=1}^L \sum_{i \in s_h} \widehat{p}_{hi} y_{hi}.$$

Se deben de tener en cuenta las siguientes observaciones cuando se aplica el Algoritmo 2.2:

- Las cantidades $\widetilde{\mathbf{x}}_h$ se pueden ver como funciones que dependen de $\mathbf{t}$, según la expresión (2.34).
- Se tiene que $\sum_h W_h \widetilde{\mathbf{x}}_h$ es monótona respecto $\mathbf{t}$. Esto es importante para determinar las soluciones $\widetilde{\mathbf{x}}_h$, puesto que aumentando o disminuyendo el valor $\mathbf{t}$, es posible llegar fácilmente a ellas.
- La unicidad de la solución está asegurada como consecuencia de la monotonía de $\sum_h W_h \widetilde{\mathbf{x}}_h$ respecto $\mathbf{t}$.

Este algoritmo, que también ha sido descrito en Zhong y Rao (2000), es más eficiente cuando la variable auxiliar $\mathbf{x}$ es unidimensional, puesto que en este caso puede encontrarse la solución incrementando o disminuyendo el valor de $\mathbf{t}$, el cual es unidimensional. Cuando se tiene más de una variable auxiliar, buscar la solución es un problema más complejo al tener que aumentar o disminuir un vector $\mathbf{t}$. Además, el calculo de $\widehat{p}_{hi}$ requiere resolver repetidamente sistemas no-lineales de grandes dimensiones según la expresión (2.34), y esto en la práctica es difícil de calcular. Por estas razones, se han buscado aproximaciones alternativas, que sean eficientes y fáciles de llevar a la práctica tanto si se dispone de una variable auxiliar como si son varias.

En Wu (2004b) se detalla el siguiente planteamiento que resuelve los inconvenientes anteriores y se basa en la estrategia (G2).

El objetivo que se persigue es poder aplicar el Algoritmo 2.1 de Chen *et al.* (2002). Para ello, tanto la log-función de verosimilitud pseudo empírica como las restricciones deben estar formuladas para el conjunto de los estratos, esto es, todas deben tener dobles sumatorias. Para este propósito, se tiene que reemplazar la expresión (2.23) por otra similar formulada a nivel poblacional. Sean las restricciones

$$\sum_{h=1}^L W_h \sum_{i \in s_h} p_{hi} = 1, \qquad (2.35)$$

$$\sum_{i \in s_h} p_{hi} = 1, \quad h = \{1, \ldots, L-1\}. \qquad (2.36)$$

Manteniendo al margen (2.35), se combinan (2.36) y (2.24) añadiendo en el vector de variables auxiliares $L-1$ variables indicadoras para cada estrato. Esto es, si $\mathbf{x}_{hi} = (x_{hi1}, \ldots, x_{hiP})$, se define

$$\begin{aligned}
\mathbf{z}_{1i} &= (1, 0, \ldots, 0, x_{1i1}, \ldots, x_{1iP})^t, \\
\mathbf{z}_{2i} &= (0, 1, \ldots, 0, x_{2i1}, \ldots, x_{2iP})^t, \\
&\vdots \\
\mathbf{z}_{(L-1)i} &= (0, 0, \ldots, 1, x_{(L-1)i1}, \ldots, x_{(L-1)iP})^t, \\
\mathbf{z}_{Li} &= (0, 0, \ldots, 0, x_{Li1}, \ldots, x_{LiP})^t,
\end{aligned}$$

y $\overline{\mathbf{Z}} = (W_1, \ldots, W_{L-1}, \overline{X}_1, \ldots, \overline{X}_P)^t$, siendo $(\overline{X}_1, \ldots, \overline{X}_P)^t = \overline{\mathbf{X}}$. Así, las restricciones (2.36) y (2.24) se pueden combinar mediante la restricción

$$\sum_{h=1}^L W_h \sum_{i \in s_h} p_{hi} \mathbf{z}_{hi} = \overline{\mathbf{Z}}. \qquad (2.37)$$

El problema de maximizar $\widehat{l}(p)$ sujeta a (2.23) y (2.24) es equivalente a maximizar $\widehat{l}(p)$ sujeta a (2.35) y (2.37). Usando el método de los multiplicadores de Lagrange a éste último planteamiento, se obtiene

$$\widehat{p}_{hi} = \frac{d_{hi}^*}{1 + \lambda^t \mathbf{u}_{hi}} = \mathbf{0},$$

donde

$$d_{hi}^* = \frac{d_{hi}}{W_h \sum_{h=1}^L \sum_{i \in s_h} d_{hi}} \quad , \quad \mathbf{u}_{hi} = \mathbf{z}_{hi} - \overline{\mathbf{Z}},$$

y λ es solución de

$$\sum_{h=1}^{L}\sum_{i\in s_h}\frac{d_{hi}\mathbf{u}_{hi}}{1+\lambda^t\mathbf{u}_{hi}}=\mathbf{0}. \tag{2.38}$$

En esta situación es posible aplicar el Algoritmo 2.1, estando garantizada la convergencia a la única solución, si tal solución existe.

Ejemplo 2.5 ***Estimadores bajo muestreo bifásico.***

Los estimadores comentados hasta el momento en esta sección están basados en un diseño muestral general y utilizan el vector media poblacional de las variables auxiliares para obtener las estimaciones. Cuando este vector es desconocido, ni los estimadores de verosimilitud empírica ni cualquier otro estimador basado en información auxiliar puede ser utilizado, puesto que la mayoría de éstos se construyen con ayuda de $\overline{\boldsymbol{X}}$ *para mejorar la precisión en la estimación de parámetros de la variable de interés. Véase, por ejemplo, Cochran (1977) y Särndal* et al. *(1992) para consultar los numerosos estimadores en la literatura del muestreo de poblaciones finitas que hacen uso de la información auxiliar.*

En la situación anterior, donde tan solo se conocen los datos muestrales de las variables auxiliares, es necesario estimar $\overline{\boldsymbol{X}}$ *o intentar dar una buena aproximación mediante alguna técnica o recurso. El muestreo bifásico (también denominado muestreo doble o en dos fases) permite estimar estas cantidades desconocidas y por tanto, es posible utilizar todos los métodos basados en información auxiliar.*

De este modo, en este ejemplo se resuelve el problema de la estimación de parámetros lineales en muestreo bifásico con diseños muestrales arbitrarios en cada una de las dos fases y aplicando el método de verosimilitud empírica.

En muestreo bifásico, el método de verosimilitud empírica puede ser aplicado como sigue. El $PEMLE$ *viene dado por*

$$\overline{y}_{PEb}=\sum_{i\in s}\widehat{p}_i y_i \tag{2.39}$$

donde los pesos $\widehat{p}_i$ *maximizan la log-función de verosimilitud pseudo empírica*

$$\widehat{l}(p)=\sum_{i\in s}d_i\log(p_i) \tag{2.40}$$

sujeta a las restricciones

$$\sum_{i\in s}p_i=1\quad(p_i\geq 0) \tag{2.41}$$

$$\sum_{i\in s}p_i\boldsymbol{u}_i'=\boldsymbol{0} \tag{2.42}$$

donde para todo $i\in s$, $d_i=d_i'd_{i/s'}$, *y* $\boldsymbol{u}_i'$ *es una función que depende de* y *y de los valores de* $\boldsymbol{x}$ *obtenidos en la muestra de la primera fase,* s'. *Además, esta función ha de verificar*

$$\frac{1}{n'}\sum_{i\in s'}\boldsymbol{u}_i'=\boldsymbol{0}.$$

Asumiendo relación lineal entre y *y* $\boldsymbol{x}$, *es usual considerar* $\boldsymbol{u}_i'=\boldsymbol{x}_i-\overline{\boldsymbol{x}}'$, *y la restricción (2.42) se puede expresar como*

$$\sum_{i\in s}p_i\boldsymbol{x}_i=\frac{1}{n'}\sum_{i\in s'}\boldsymbol{x}_i=\overline{\boldsymbol{x}}',$$

que viene a indicar que si los pesos que van a se estimados se ponderan sobre los datos muestrales del vector de variables auxiliares de la segunda fase, se obtendrá la cantidad $\overline{\boldsymbol{x}}'$, *es decir, la media muestral del vector de las variables auxiliares obtenida a partir de la muestra de la primera fase. De ahí la importancia de realizar un gran esfuerzo para obtener una buena estimación para* $\overline{\boldsymbol{X}}$ *con los datos de la muestra de la primera fase.*

La solución del problema planteado se resuelve por el método de los multiplicadores de Lagrange, obteniendo como solución para todo $i\in s$ *las cantidades*

$$\widehat{p}_i=\frac{d_i^*}{1+\lambda^t\boldsymbol{u}_i'},$$

donde

$$d_i^*=\frac{d_i}{\sum_{j\in s}d_j}=\frac{d_i'd_{i/s'}}{\sum_{j\in s}d_j'd_{j/s'}},$$

y λ *es el vector de multiplicadores de Lagrange que se obtiene de la ecuación*

$$\sum_{i\in s}\frac{d_i^*u_i'}{1+\lambda^t\boldsymbol{u}_i'}=\boldsymbol{0}.$$

2.2.2. Propiedades teóricas

En esta sección se describen las propiedades asintóticas más importantes de los estimadores de verosimilitud empírica basados en el diseño muestral. En primer lugar, se describen las propiedades teóricas más importantes del estimador de verosimilitud empírica propuesto en Chen y Qin (1993) bajo muestreo aleatorio simple. A continuación, se demuestra la relación que tiene el $PEMLE$ con los conocidos estimadores de regresión. Esta sección se completa con las propiedades teóricas de los estimadores de verosimilitud empírica en muestreo estratificado y su relación con otros estimadores.

Propiedades en muestreo aleatorio simple

A continuación se estudian las propiedades asintóticas del estimador de verosimilitud empírica descrito en Chen y Qin (1993). Asumamos muestreo aleatorio simple, donde el tamaño de la muestra, n, y el tamaño de la población, N, tienden a infinito cuando un cierto índice, ν, tiende a infinito, es decir, existe una sucesión de poblaciones finitas indexadas por ν, donde $\pi_\nu=\{(x_{1_\nu},y_{1_\nu}),\ldots,(x_{N_\nu},y_{N_\nu})\}$ y el tamaño poblacional N_ν tiende a infinito. Por comodidad, se suprime el índice ν siempre que sea posible y se considera sólo una variable auxiliar. Sea

$$\sigma_x^2=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{X})^2,\quad \sigma_y^2=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{Y})^2,$$

$$\sigma_{xy}=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{X})(y_i-\overline{Y}),$$

y $\overline{x},\overline{y},s_x^2,s_y^2$ y s_{xy} sus correspondientes versiones muestrales. Se considera que la función de calibración satisface $\sum_{i=1}^{N}u_i=0$ y se tiene que

$$\sigma_u^2=\frac{1}{N-1}\sum_{i=1}^{N}u_i^2,\quad \sigma_{yu}=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{Y})u_i.$$

La media poblacional de variable de interés se estima a través del estimador $\overline{y}_{EL} = \sum_{i\in s} \widehat{p}_i y_i$. Los siguientes teoremas pueden ser definidos.

Teorema 2.1 *Suponiendo que cuando $\nu \to \infty$, el tamaño poblacional N, el tamaño muestral n, y $N - n$ tienden a infinito, y*

$$\frac{1}{N}\left\{\sum_{i=1}^{N}|u_i|^3\right\}, \quad \frac{1}{N}\left\{\sum_{i=1}^{N}|y_i|^3\right\},$$

tienen una cota superior independiente de ν, entonces se verifica

$$\frac{n^{1/2}(\overline{y}_{EL} - \overline{Y})}{\sigma_\nu} \to N(0,1),$$

donde $\sigma_\nu^2 = \left(1 - \frac{n}{N}\right)\left(\sigma_y^2 - \frac{\sigma_{yu}^2}{\sigma_u^2}\right)$.

La demostración de este resultado puede consultarse en Chen y Qin (1993). Una consecuencia importante que puede observarse de este teorema es que a mayor correlación entre u e y, mayor será la ganancia en precisión. Se demuestra que la eficiencia asintótica del método es equivalente a la del método de regresión.

En la práctica, la cantidad σ_ν^2 es desconocida, con lo que se tiene que buscar un buen estimador. Una alternativa es la estimación de σ_y^2, σ_{yu} y σ_u^2 por separado, aunque para tamaños muestrales moderados trabaja mejor el estimador jackknife para la varianza. En el siguiente teorema, debido a Chen y Qin (1993), se demuestra que el estimador jackknife es un buen estimador para σ_ν^2.

Teorema 2.2 *Bajo las mismas condiciones del Teorema 2.1, si $\overline{y}_{EL}(-j)$ es el estimador cuando la observación j-ésima es eliminada y*

$$\widehat{\sigma}_J^2 = \left(1 - \frac{n}{N}\right)(n-1)\sum_{i\in s}(\overline{y}_{EL}(-j) - \overline{y}_{EL})^2,$$

entonces,

$$\widehat{\sigma}_J^2 - \sigma_\nu^2 = o_p(1).$$

Propiedades para un diseño muestral general

En lo que sigue, se asume una sola variable auxiliar y la función de calibración $u_i = x_i - \overline{X}$. Consideremos también las siguientes condiciones

(C2.1). $u^* = max_{i\in s}|u_i| = o_p(n^{1/2})$,

(C2.2). $\frac{\sum_{i\in s} d_i u_i}{\sum_{i\in s} d_i u_i^2} = O_p(n^{-1/2})$.

El siguiente teorema, debido a Chen y Sitter (1999), puede establecerse.

Teorema 2.3 *Bajo las condiciones (C2.1) y (C2.2), el $PEMLE$ de $\overline{Y}$ cuando $\overline{X}$ es conocida, es asintóticamente equivalente al estimador de regresión generalizado $(GREG)$. Es decir,*

$$\lambda = \frac{\overline{x}_w - \overline{X}}{\sum_{i\in s} d_i^*(x_i - \overline{x}_w)^2} + o_p(n^{-1/2}),$$

y así $\overline{y}_{PE} = \overline{y}_{GREG} + o_p(n^{-1/2})$, *donde*

$$\overline{y}_{GREG} = \sum_{i\in s}\widetilde{d}_i y_i, \quad \widetilde{d}_i = d_i^*\left[1 - \frac{(x_i - \overline{x}_w)(\overline{x}_w - \overline{X})}{\sum_{i\in s} d_i^*(x_i - \overline{x}_w)^2}\right],$$

$$\overline{y}_w = \sum_{i\in s} d_i^* y_i, \quad \overline{x}_w = \sum_{i\in s} d_i^* x_i \quad \textit{y} \quad d_i^* = \frac{d_i}{\sum_{j\in s} d_j}.$$

Las condiciones (C2.1) y (C2.2) deben satisfacerse para que este teorema pueda establecerse. Sin embargo, estas condiciones no son muy restrictivas y los diseños muestrales mas conocidos las satisfacen. En Chen y Sitter (1999) se demuestra cómo estas condiciones se cumplen en tres diseños comunes, como son, el muestreo con probabilidades proporcionales al tamaño con reemplazamiento, el método de Rao-Hartley-Cochran y el muestreo por conglomerados.

Un punto importante es la estimación de la varianza del estimador $\overline{y}_{PE}$. Según el Teorema 2.3, resulta evidente que cualquier estimador de la varianza consistente para $\overline{y}_{GREG}$ será consistente para el $PEMLE$. Aunque esto es asintóticamente válido, no es atractivo usar un estimador de la varianza del $GREG$ para estimar la varianza del $PEMLE$. Una alternativa óptima es aplicar estimadores de la varianza remuestreados, tales como jackknife, bootstrap y replicaciones de muestras repetidas balanceadas (ver Shao y Wu (1989, 1992), Chen y Qin (1993) y Shao (1994)) sobre $\overline{y}_{PE}$, recalculando $\widehat{p}_i$ en cada muestra.

Propiedades en muestreo estratificado

La primera propiedad del $PEMLE$ en muestreo estratificado se basa en el Teorema 2.3.

Corolario 2.1 *Bajo las condiciones (C2.1) y (C2.2) se tiene*

$$\overline{y}_{PE} = \overline{y}_w - \frac{\sum_{h=1}^{L}\sum_{i\in s_h} d_{hi}^*(x_{hi} - \overline{x}_w)y_{hi}}{\sum_{h=1}^{L}\sum_{i\in s_h} d_{hi}^*(x_{hi} - \overline{x}_w)^2}(\overline{x}_w - \overline{X}) + o_p(n^{-1}$$

donde

$$n = \sum_{h=1}^{L} n_h \quad , \quad \overline{y}_w = \sum_{h=1}^{L}\sum_{i\in s_h} d_{hi}^* y_{hi},$$

$$\overline{x}_w = \sum_{h=1}^{L}\sum_{i\in s_h} d_{hi}^* x_{hi} \quad \textit{y} \quad d_{hi}^* = \frac{d_{hi}}{\sum_{h=1}^{L}\sum_{j\in s_h} d_{hj}}.$$

Considerando muestreo aleatorio estratificado, es decir, cuando $d_{hi} = N_h/n_h$, la expresión anterior se reduce a

$$\overline{y}_{PE} = \overline{y}_{st} - \frac{\sum_{h=1}^{L}\sum_{i\in s_h} W_h(x_{hi} - \overline{x}_{st})y_{hi}/n_h}{\sum_{h=1}^{L}\sum_{i\in s_h} W_h(x_{hi} - \overline{x}_{st})^2/n_h}(\overline{x}_{st} - \overline{X}) +$$

$$+ o_p(n^{-1/2}) = \overline{y}_{GREG} + o_p(n^{-1/2}),$$

donde

$$\overline{y}_{st} = \sum_{h=1}^{L} W_h \overline{y}_h \quad \text{y} \quad \overline{x}_{st} = \sum_{h=1}^{L} W_h \overline{x}_h.$$

Esta no es la mejor aproximación posible, puesto que se sabe que el estimador de regresión óptimo (ORE), definido en Rao (1994), funciona mejor que el $GREG$ en muestreo estratificado. Por este motivo, en Chen y Sitter (1999) se busca una mejor aproximación. En el siguiente corolario se relaciona el $PEMLE$ con el ORE bajo muestreo aleatorio estratificado. Para ello, se asume que existe una sucesión de poblaciones finitas indexadas por ν, tal que cuando $\nu \to \infty$ se verifican las condiciones

(C2.3). $0 \leq c_1 \leq \sum_{h=1}^{L} W_h \sigma_h^2 \leq c_2 \leq \infty$,

(C2.4). $\max\{n_h^{-1} W_h\} = O(n^{-1})$,

(C2.5). $N^{-1} \sum_{h=1}^{L} \sum_{i=1}^{N_h} |x_{hi}|^3 = O(1)$,

(C2.6). $N^{-1} \sum_{h=1}^{L} \sum_{i=1}^{N_h} |y_{hi}|^3 = O(1)$.

Corolario 2.2 *Bajo muestreo aleatorio estratificado y las condiciones (C2.3), (C2.4), (C2.5) y (C2.6), el $PEMLE$ de $\overline{Y}$, cuando $\overline{X}$ es conocida, es asintóticamente equivalente a $\overline{y}_{st}^{*}$, esto es, $\overline{y}_{PE} = \overline{y}_{st}^{*} + o_p(n^{-1/2})$, donde*

$$\overline{y}_{st}^{*} = \overline{y}_{st} - \frac{\sum_{h=1}^{L} W_h \sum_{i \in s_h} (x_{hi} - \widetilde{x}_h) y_{hi} / n_h}{\sum_{h=1}^{L} W_h \sum_{i \in s_h} (x_{hi} - \widetilde{x}_h)^2 / n_h} (\overline{x}_{st} - \overline{X}),$$

y las cantidades $\widetilde{x}_h$ están definidas en (2.29). Cuando L permanece finito, $\widetilde{x}_h - \overline{x}_h = O_p(n^{-1/2})$ y el estimador $\overline{y}_{PE}$ es asintóticamente equivalente al estimador lineal óptimo dado en Rao (1994).

Asumiendo otros diseños muestrales en cada estrato, las comparaciones con respecto otros estimadores son demasiado dificultosas y se ha de recurrir a la simulación para realizar las comparaciones.

En este caso, la estimación de la varianza se obtiene también a través de estimadores de la varianza remuestreados. En Chen y Sitter (1999), se demuestra que bajo muestreo aleatorio estratificado el estimador de la varianza jackknife para el $PEMLE$ es consistente.

2.2.3. Estimadores modelo-calibrados

Una de las restricciones considerada en los estimadores de verosimilitud empírica viene dada por

$$\sum_{i \in s} p_i \mathbf{u}_i = \mathbf{0}, \tag{2.43}$$

donde $\mathbf{u}_i = u(y_i, \mathbf{x}_i)$ y $u(\cdot)$ es una función conocida de y y de $\mathbf{x}$ que verifica

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{u}_i = \mathbf{0}. \tag{2.44}$$

Asumiendo una relación lineal entre la característica de interés y el vector auxiliar de variables, se utiliza frecuentemente la expresión $\mathbf{u}_i = \mathbf{x}_i - \overline{\mathbf{X}}$, y se plantea la cuestión de cómo de efectivo es el uso que se está haciendo de la información auxiliar. Si tal relación no es lineal, los estimadores de verosimilitud empírica obtenidos a partir de la expresión $\mathbf{u}_i = \mathbf{x}_i - \overline{\mathbf{X}}$ pueden resultar ineficaces y surge, por tanto, el problema de encontrar una función de calibración apropiada para los datos del estudio, es decir, que se adapte a cada situación para poder usar la información auxiliar de la mejor manera posible. Una alternativa eficiente para resolver este problema es el uso de los estimadores modelo-calibrados, los cuales están basados en modelos de superpoblación.

Recientemente, en la literatura del muestreo se están utilizando a menudo estimaciones que no están basadas en el diseño muestral, sino que dependen de un determinado modelo de superpoblación que relaciona la variable de interés a través de las variables auxiliares. Tales procedimientos son los estimadores basados en modelos y los estimadores modelo-calibrados. Con la aparición de los modelos de superpoblación la teoría de muestreo tuvo un gran empuje pues se le dotó de un instrumento muy valioso que permitió obtener resultados más concluyentes en la comparación de estrategias y eventualmente producir estrategias óptimas en varias situaciones. Ejemplos e información sobre modelos de superpoblación pueden consultarse, por ejemplo, en Godambe (1955), Godambe y Thompson (1973), Cassel *et al.* (1976), Pérez (2002) y Sánchez-Crespo (2002).

Los estimadores modelo-calibrados están propuestos en Wu y Sitter (2001), y se obtienen adaptando un modelo de superpoblación, y a continuación, usando los valores estimados mediante este modelo en la etapa de estimación. Así, se obtiene una función eficiente de calibración, y además es posible encontrar la mejor función $u(\cdot)$ en el sentido de mínima esperanza bajo un modelo de superpoblación de la varianza asintótica basada en el diseño.

Los valores $\mathbf{u}_i$ pueden expresarse como

$$\mathbf{u}_i = \mathbf{w}_i - \frac{1}{N} \sum_{i=1}^{N} \mathbf{w}_i,$$

donde $\mathbf{w}_i$ es una función conocida. Es fácil demostrar que bajo esta situación también se verifica (2.44), y por tanto, se cumplen las condiciones necesarias para aplicar la metodología de verosimilitud empírica. Operando en la restricción (2.43) se llega a la restricción alternativa

$$\sum_{i \in s} p_i \mathbf{w}_i = \frac{1}{N} \sum_{i=1}^{N} \mathbf{w}_i, \tag{2.45}$$

que es la que suele usarse en los estimadores modelo-calibrados de verosimilitud empírica. Por tanto, el problema de buscar unos valores óptimos $\mathbf{u}_i$ para obtener estimadores más eficientes, es similar al de encontrar la cantidades $\mathbf{w}_i$, para $i \in s$.

La idea de definir estimadores óptimos bajo un modelo y asumiendo el criterio de mínima esperanza bajo un modelo de superpoblación de la varianza asintótica basada en el diseño ha sido discutida por diversos autores, véase, por ejemplo, Godambe (1955), Godambe y Thompson (1973) y Cassel *et al.* (1976).

Un primer estimador modelo-calibrado surge cuando se asume el siguiente esquema asintótico. Existe una sucesión de poblaciones finitas indexadas por ν. El tamaño poblacional y el tamaño muestral para la población ν-ésima se denotan como N_ν y n_ν. Cuando $\nu \to \infty$, $N_\nu \to \infty$ y $n_\nu \to \infty$. El índice ν se suprimirá para simplificar notación. Por ejemplo, véase Isaki y Fuller (1982)

para un mayor detalle de este esquema asintótico. Por último, sea $y_1, y_2, \ldots, y_N$ una muestra aleatoria de un modelo de superpoblación ξ tal que

$$E_\xi(y_i) = \mu_i, \quad V_\xi(y_i) = \sigma_i^2, \quad i = \{1, 2, \ldots, N\}, \quad (2.46)$$

y $y_1, y_2, \ldots, y_N$ son independientes entre ellos. E_ξ y V_ξ denotan la esperanza y la varianza bajo el modelo de superpoblación.

Sea $\widetilde{y}_{C_\mathbf{w}}$ el estimador de verosimilitud pseudo empírica modelo-calibrado de $\overline{Y}$ cuando se usa $C_\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N\}$ en la restricción (2.45) y L^* un conjunto de sucesiones $C_\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N\}$ que verifican

$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}_i)^6 = O(1)$$

y

$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}_i)^2 \to \mathbf{c} \neq \mathbf{0} \text{ cuando } N \to \infty.$$

Estas condiciones sobre la sucesión $C_\mathbf{w} \in L^*$ no son muy restrictivas y se usan para facilitar las demostraciones. Asumiremos que $\{\mu_1, \ldots, \mu_N\} \in L^*$.

Se dice que un diseño muestral es regular si el diseño que resulta de un tamaño de muestra indexado tiene probabilidades de inclusión π_i y π_{ij} independientes de la característica y_i dada $\mathbf{x}_i$, y además satisface las siguientes condiciones:

(C2.7). $max_{i\in s}\left(\frac{nd_i}{N}\right) = O(1).$

(C2.8). $\frac{1}{N}\sum_{i\in s} d_i\mathbf{w}_i - \frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i = O_p(n^{-1/2})$ para cualquier sucesión de funciones $(\mathbf{w}_1, \ldots, \mathbf{w}_N) \in L^*$.

En Wu (2003) se demuestra que entre todas las clases de estimadores $\widetilde{y}_{C_\mathbf{w}}$ con $C_\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N\} \in L^*$, el valor $C_\mu = \{\mu_1, \ldots, \mu_N\}$ como variable de calibración en (2.45) minimiza $E_\xi[AV_p(\widetilde{y}_{C_\mathbf{w}})]$ bajo el modelo (2.46) y para cualquier diseño muestral regular. AV_p denota la varianza asintótica bajo el diseño. Así, el estimador de verosimilitud pseudo empírica modelo-calibrado $(MCPE)$ que presenta la propiedad arriba comentada, se construye tomando $\mathbf{w}_i = \mu_i$, o lo que es lo mismo, tomando $\mathbf{u}_i = \mu_i - N^{-1}\sum_{i=1}^{N}\mu_i$. Sustituyendo estas cantidades $\mathbf{u}_i$ en las expresiones (2.12) y (2.13) se obtiene un primer estimador de verosimilitud empírica basado en la aproximación modelo-calibrada.

Otra alternativa para construir estimadores modelo-calibrados es asumir que $y_1, y_2, \ldots, y_N$ es una muestra aleatoria de un modelo de superpoblación semiparamétrico ξ tal que

$$E_\xi(y_i|\mathbf{x}_i) = \mu_i = \mu(\mathbf{x}_i, \theta), \quad V_\xi(y_i|\mathbf{x}_i) = \nu_i^2\sigma^2, \quad i = \{1, \ldots, N\}, \quad (2.47)$$

donde $\theta = (\theta_0, \theta_1, \ldots, \theta_P)^t$ y σ^2 son parámetros poblacionales desconocidos, $\mu(\mathbf{x}, \theta)$ es una función conocida de $\mathbf{x}$ y de θ, ν_i es una función conocida de $\mathbf{x}_i$ o bien de $\mu_i = \mu(\mathbf{x}_i, \theta)$ y E_ξ y V_ξ denotan la esperanza y la varianza con respecto al modelo de superpoblación. Además, se asume que los pares $(y_1, \mathbf{x}_1); (y_2, \mathbf{x}_2); \ldots; (y_N, \mathbf{x}_N)$ son mutuamente independientes.

Este modelo es bastante general, e incluye dos casos muy importantes:

1. **El modelo de regresión lineal o no lineal**

   $$y_i = \mu(\mathbf{x}_i, \theta) + \nu_i\epsilon_i \quad i = \{1, \ldots, N\},$$

   donde ϵ_i son variables aleatorias independientes e idénticamente distribuidas, con $E_\xi(\epsilon_i) = 0$, $V_\xi(\epsilon_i) = \sigma^2$ y $\nu_i = \nu(x_i)$ una función conocida y estrictamente positiva que depende de $\mathbf{x}_i$.

2. **El modelo lineal generalizado**

   $$g(\mu_i) = \mathbf{x}_i^t\theta \quad V_\xi(y_i|\mathbf{x}_i) = \nu(\mu_i) \quad i = \{1, \ldots, N\},$$

   donde $\mu_i = E_\xi(y_i|\mathbf{x}_i)$, $g(\cdot)$ es una función de enlace y $\nu(\cdot)$ es la función varianza.

Los verdaderos parámetros del modelo son desconocidos, aunque pueden estimarse mediante cualquier método basado en el diseño. Asumiendo una aproximación basada en el modelo, la dupla $(y_i, \mathbf{x}_i)$ con $i \in s$ puede verse como una muestra independiente idénticamente distribuida del modelo de superpoblación. Los parámetros θ se pueden estimar usando procedimientos estándares. Bajo el enfoque basado en el diseño, los datos muestrales pueden no seguir la misma estructura del modelo que la población finita completa bajo un esquema muestral complejo, y θ puede carecer de sentido desde el punto de vista del diseño. En este caso, θ se reemplaza por θ_N, una estimación de θ basada en los datos de la población completa. θ_N se reemplaza entonces por $\widehat{\theta}$, una estimación basada en el diseño de los datos muestrales (véase Godambe y Thompson, 1986).

Asumiendo el modelo (2.47), el estimador de verosimilitud pseudo empírico modelo-calibrado se construye tomando $\mathbf{w}_i = \mu(\mathbf{x}_i, \widehat{\theta})$. Los valores $\mathbf{u}_i$ vienen dados por $\mathbf{u}_i = \widehat{\mu}_i - N^{-1}\sum_{i=1}^{N}\widehat{\mu}_i$, donde $\widehat{\mu}_i = \mu(\mathbf{x}_i, \widehat{\theta})$. Considerando estas cantidades en las expresiones (2.12) y (2.13) se obtiene el $MCPE$.

Al igual que ocurre bajo el primer $MCPE$ que se ha definido, en Wu (2003) se demuestra que entre todas las clases de estimadores $\widetilde{y}_{C_\mathbf{w}}$, donde $C_\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N\} \in L^*$, el valor $C_\mu = \{\mu(\mathbf{x}_1, \theta), \ldots, \mu(\mathbf{x}_N, \theta)\}$ como variable de calibración en (2.45) minimiza $E_\xi[AV_p(\widetilde{y}_{C_\mathbf{w}})]$ bajo el modelo (2.47) y para cualquier diseño muestral regular.

A continuación se resumen las observaciones más importantes sobre los estimadores de verosimilitud empírica basados en una aproximación modelo-calibrada.

1. En Wu y Sitter (2001) se demuestra que reemplazar θ por $\widehat{\theta}$ en $\mu_i = \mu(\mathbf{x}_i, \theta)$, no cambia asintóticamente el estimador resultante.
2. Con probabilidad tendiendo a uno, el $MCPE$ existe y se puede calcular usando el algoritmo 2.1 de Chen *et al.* (2002).
3. El uso efectivo de la información auxiliar depende los parámetros estimados y de la relación entre la variable respuesta y las covarianzas. Por tanto, usar la calibración sobre las variables auxiliares sin un estudio exhaustivo previo no es usualmente una buena aproximación.

4. Es sabido que para una relación lineal entre y y el vector de variables auxiliares, se toma $\mathbf{u}_i = \mathbf{x}_i - \overline{\mathbf{X}}$ para la construcción del $PEMLE$. En esta situación, el $PEMLE$ y el $MCPE$ son asintóticamente equivalentes si se considera $\widehat{\mu}_i = \mathbf{x}_i^t\widehat{\theta}$ como variable de calibración para el cálculo de la aproximación modelo-calibrada. La demostración de este resultado puede consultarse en Wu y Sitter (2001).
5. Si la relación entre y y **x** es lineal, tan sólo el conocimiento de $\overline{\mathbf{X}}$ es suficiente para obtener estimadores eficientes para la media o el total poblacional. Si dicha relación no es lineal o el parámetro de interés no es una función lineal, una información auxiliar completamente disponible y/o más datos sobre el modelo son esenciales para una estimación óptima.
6. Al igual que se ha comentado anteriormente, las cantidades $\widehat{p}_i$ son positivas. Esta propiedad no se cumple ni en los estimadores de calibración ni en calculo del $GREG$ y juega un papel muy importante en la estimación de otros parámetros de interés en el muestreo, como son la función de distribución, cuantiles, varianza y otras funciones cuadráticas.

2.2.4. Propiedades teóricas

Sea el esquema asintótico siguiente: se asume que existe una sucesión de diseños muestrales y una sucesión de poblaciones finitas indexadas por ν. El tamaño muestral n_ν y el tamaño poblacional N_ν se aproximan a infinito cuando $\nu \to \infty$.

Las condiciones siguientes son necesarias para poder aplicar el Teorema 2.4.

(C2.9). $\widehat{\theta} = \theta_N + O_p(n^{-1/2})$ y $\theta_N \to \theta$.

(C2.10). Para cada $\mathbf{x}_i$, $\dfrac{\partial\mu(\mathbf{x}_i,\mathbf{t})}{\partial\mathbf{t}}$ es continua en **t** y

$$\left|\frac{\partial\mu(\mathbf{x}_i,\mathbf{t})}{\partial\mathbf{t}}\right| \leq h(\mathbf{x}_i,\theta)$$

para **t** en un entorno de θ, y $N^{-1}\sum_{i=1}^N h(\mathbf{x}_i,\theta) = O_p(1)$.

(C2.11). Los pesos básicos muestrales, $d_i = \pi_i^{-1}$, hacen que los estimadores de Horvitz-Thompson para ciertas medias muestrales estén asintóticamente normalmente distribuidos.

(C2.12). $\mathbf{u}^* = \max_{i\in s}|\mathbf{u}_i| = o_p(n^{1/2})$, donde $\mathbf{u}_i = \mu(\mathbf{x}_i,\theta_N) - \dfrac{1}{N}\sum_{i=1}^N \mu(\mathbf{x}_i,\theta_N)$.

(C2.13). $\dfrac{\sum_{i\in s} d_i\mathbf{u}_i}{\sum_{i\in s} d_i\mathbf{u}_i^2} = O_p(n^{1/2})$.

(C2.14). $h^* = \max_{i\in s}|h_i| = o_p(n)$, siendo $h_i = h(\mathbf{x}_i,\theta_N)$.

El siguiente teorema puede establecerse.

Teorema 2.4 *Bajo el esquema asintótico descrito y las condiciones anteriores (C2.9)∼(C2.14), se tiene que*

$$\overline{y}_{MCPE} = \overline{y}_{MC} + o_p(n^{-1/2}),$$

donde $\overline{y}_{MC}$ es el estimador modelo-calibrado para la media obtenido mediante el método de calibración y cuya expresión viene dada por

$$\overline{y}_{MC} = \overline{y}_{HT} + \left\{\frac{1}{N}\sum_{i=1}^N \widehat{\mu}_i - \frac{1}{N}\sum_{i\in s} d_i\widehat{\mu}_i\right\}\widehat{B}_N,$$

con

$$\widehat{B}_N = \frac{\sum_{i\in s} d_i q_i(\widehat{\mu}_i - \overline{\mu})(y_i - \overline{y})}{\sum_{i\in s} d_i q_i(\widehat{\mu}_i - \overline{\mu})^2}, \quad \overline{y} = \frac{\sum_{i\in s} d_i q_i y_i}{\sum_{i\in s} d_i q_i}$$

y

$$\overline{\mu} = \frac{\sum_{i\in s} d_i q_i\widehat{\mu}_i}{\sum_{i\in s} d_i q_i}.$$

Las cantidades q_i son constantes positivas.

Puesto que $\overline{y}_{MCPE}$ es asintóticamente equivalente al $\overline{y}_{MC}$, las mismas expresiones de la varianza y del estimador de la varianza de $\overline{y}_{MC}$ pueden usarse para $\overline{y}_{MCPE}$. Estas varianzas asintóticas basadas en el diseño vienen dadas por

$$AV(\overline{y}_{MCPE}) = \frac{1}{N^2}\sum_{i<j}^N(\pi_i\pi_j - \pi_{ij})\left(\frac{U_i}{\pi_i} - \frac{U_j}{\pi_j}\right)^2,$$

donde π_{ij} son las probabilidades de inclusión de segundo orden, $U_i = y_i - \mu_i B_N$, $\mu_i = \mu(\mathbf{x}_i,\theta_N)$,

$$B_N = \frac{\sum_{i=1}^N q_i(\mu_i - \overline{\mu}_N)(y_i - \overline{Y})}{\sum_{i=1}^N q_i(\mu_i - \overline{\mu}_N)^2} \quad \text{y} \quad \overline{\mu}_N = \frac{1}{N}\sum_{i=1}^N \mu_i.$$

Un estimador para esta varianza viene dado por

$$\widehat{V}(\overline{y}_{MCPE}) = \frac{1}{N^2}\sum_{i<j}^n \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{u_i}{\pi_i} - \frac{u_j}{\pi_j}\right)^2,$$

donde $u_i = y_i - \widehat{\mu}_i\widehat{B}_N$.

Estas varianzas asintóticas y la demostración del teorema se pueden consultar en Wu y Sitter (2001).

Aunque estas aproximaciones son asintóticamente válidas, resulta más atractivo usar estimadores de varianzas remuestreados sobre el $MCPE$.

2.3. Tratamiento de datos faltantes

En esta sección se propone un estimador para la media poblacional cuando algunas observaciones de la variable de estudio o de las variables auxiliares están perdidas en la muestra. El nuevo estimador es válido para cualquier diseño muestral (con probabilidades iguales o desiguales) y está basado en el método de verosimilitud empírica. El estimador propuesto se compara con otros estimadores conocidos en un estudio empírico.

2.3.1. Introducción

En la práctica, es común el uso de información auxiliar poblacional en la etapa de estimación. Esta técnica tiene muchas ventajas. Por ejemplo, una adecuada información auxiliar puede producir una reducción considerable en el sesgo y el error muestral.

Cuando una o más variables auxiliares correlacionadas con la variable de estudio están disponibles, el método de calibración (Deville y Särndal, 1992) y el método de verosimilitud pseudo empírica (Chen y Qin, 1993, Chen y Sitter, 1999, Wu y Sitter, 2001, Wu, 2002) pueden usarse para estimar el total poblacional, la media poblacional, funciones de distribución y cuantiles. Ambos métodos usan información auxiliar de una o más variables auxiliares.

Generalmente, estas técnicas proporcionan estimadores que son más eficientes que los estimadores tradicionales, tales como el estimador de Horvitz y Thompson (1952) y el estimador tipo Hájek para la media (Rao, 1966, Basu, 1971, Särndal *et al.*, 1992). Sin embargo, el método de verosimilitud empírica asume respuesta completa sin valores perdidos, esto es, se asume que ninguna unidad muestral falla para proporcionar información en las variables de estudio y auxiliares.

La pérdida de información es una propiedad común en las investigaciones por muestreo. Esta pérdida de información puede ocurrir por varias razones: los individuos muestreados pueden negarse a participar en el estudio, los entrevistadores no pueden contactar con los individuos del estudio, pérdida accidental de información, etc.

En esta sección, se asume que si hay falta de respuesta, ésta es uniforme. Tratar con datos faltantes en una investigación por muestreo no es un asunto relativamente sencillo. Existen una gran variedad de métodos en el caso de existir valores perdidos en los datos muestrales.

Ante la presencia de datos faltantes, la solución más simple es eliminar las unidades con falta de respuesta y aplicar el método de verosimilitud empírica a las unidades restantes. Sin embargo, este método, el cual Rubin (1987) llamó análisis de casos completos, puede producir sesgo en las estimaciones y varianzas muestrales más grandes (ver Rubin, 1987 o Little y Rubin, 1987).

La imputación es otra técnica que puede usarse en los individuos con falta de respuesta (Little y Rubin, 1987, Rao y Toutenburg, 1995, Särndal, 1992, Chen *et al.*, 2000). La imputación consiste en sustituir los valores perdidos por un valor adecuado. Tratar los valores imputados como si estos fueran valores verdaderos y posteriormente usar el método de verosimilitud empírica puede dirigir a inferencias no válidas. Por ejemplo, la varianza puede resultar seriamente subestimada cuando la proporción de valores perdidos no es pequeña (Rao y Shao, 1992, Särndal, 1990, 1992). Además, en algunas encuestas realizadas por organismos oficiales de estadística (como por ejemplo en la Oficina de Estadística de Suecia) está prohibida la imputación como solución al problema de datos faltantes.

Otra opción es intentar mejorar la precisión de las estimaciones incluyendo los valores observados de la variable auxiliar donde la variable de estudio está perdida. Así, aunque se tenga un valor perdido para y, el valor de x es observado y utilizado en el proceso de estimación.

Los estimadores de tipo razón, diferencia o producto también asumen respuesta completa. Algunos autores han definido estimadores de tipo razón en presencia de datos faltantes. Estos estimadores solamente han sido definidos para una clase limitada de diseños muestrales. Por ejemplo, Tracy y Osahan (1994), Toutenburg y Srivastava (1998, 1999, 2000) desarrollaron estimadores de tipo razón para muestreo aleatorio simple sin reemplazamiento.

En esta sección se propone modificar el estimador de verosimilitud pseudo empírica ($PEMLE$), el cual puede obtenerse bajo cualquier diseño muestral con probabilidades iguales o desiguales. El estimador propuesto usa toda la información muestral recogida para la variable de estudio y una variable auxiliar x, esto es, el estimador propuesto es función de los valores de x para las unidades con datos y perdidos, y función de los valores de y para las unidades con valores x perdidos.

Se considera la situación en la cual existen observaciones perdidas en una de las características para algunos individuos, pero no en la otra, es decir, la pérdida de información se produce para ambas características separadamente, pero no simultáneamente. De este modo, sea p ($p \geq 0$) el número de unidades que responden a x pero no a y, es decir, asumimos que tenemos p datos perdidos para la variable y. También se tiene información auxiliar incompleta, esto es, q ($q \geq 0$) unidades muestrales responden a y pero no a x. Notamos que p y q son números enteros. Así, se tiene un conjunto de $n-p-q$ unidades ($p+q \leqslant n$) que responden a ambas variables y y x. Con este esquema, se pueden formar los tres siguientes conjuntos disjuntos de unidades muestrales

$$\begin{aligned} s_A &= \{i \in s \,|\, x_i, y_i \text{ no están perdidos}\}, \\ s_B &= \{i \in s \,|\, x_i \text{ no está perdido, } y_i \text{ está perdido}\}, \\ s_C &= \{i \in s \,|\, y_i \text{ no está perdido, } x_i \text{ está perdido}\}. \end{aligned}$$

Asumiendo muestreo aleatorio simple sin reemplazamiento, Toutenburg y Srivastava (2000) propusieron cuatro estimadores para la media poblacional de y:

$$\overline{y}_{T1} = \overline{y}^A \left[\frac{n_{pq}\overline{x}^A + p\overline{x}^B}{(n-q)\overline{x}^A} \right], \tag{2.48}$$

$$\overline{y}_{T2} = \overline{y}^A \left[\frac{(n-q)\overline{x}^A}{n_{pq}\overline{x}^A + p\overline{x}^B} \right], \tag{2.49}$$

$$\overline{y}_{T3} = \frac{(n_{pq}\overline{x}^A + p\overline{x}^B)(n_{pq}\overline{y}^A + q\overline{y}^C)}{(n-q)(n-p)\overline{x}^A}, \tag{2.50}$$

$$\overline{y}_{T4} = \left[\frac{n_{pq}\overline{y}^A + q\overline{y}^C}{n_{pq}\overline{x}^A + p\overline{x}^B} \right] \left[\frac{n-q}{n-p}\overline{x}^A \right], \tag{2.51}$$

donde $n_{pq} = n-p-q$, $\overline{y}^i$ y $\overline{x}^i$ son las medias muestrales basadas en s_i, con $i = A, B, C$.

Los estimadores $\overline{y}_{T1}$ y $\overline{y}_{T2}$ dependen de las muestras s_A y s_B, y no dependen de la muestra s_C. Sin embargo, $\overline{y}_{T3}$ y $\overline{y}_{T4}$ dependen de las muestras s_A, s_B y s_C. Toutenburg y Srivastava (2000) demostraron que ninguno de estos estimadores es uniformemente superior a otro. Una elección apropiada del estimador requiere el conocimiento de parámetros poblacionales.

Rueda y González (2004) propusieron varios estimadores que pueden usarse bajo cualquier diseño muestral en presencia de datos faltantes. Estos estimadores

están basados en métodos de tipo razón, diferencia y regresión. Por ejemplo, el estimador siguiente es asintóticamente insesgado, bajo muestreo aleatorio simple es asintóticamente normal y es mejor, en el sentido de error cuadrático medio, que el resto de estimadores propuestos.

$$\overline{y}_{Reg} = \widehat{\alpha}_{reg}\overline{y}^{A}_{HT} + (1-\widehat{\alpha}_{reg})\overline{y}^{C}_{HT} + \tag{2.52}$$

$$+\frac{\widehat{Cov}_{i\in s_A}(x,y)}{\widehat{Var}_{i\in s_A}(x)}\left[\overline{X} - \left(\widehat{\beta}_{reg}\overline{x}^{A}_{HT} + (1-\widehat{\beta}_{reg})\overline{x}^{B}_{HT}\right)\right],$$

donde $\overline{y}^{i}_{HT}$ y $\overline{x}^{i}_{HT}$ son los estimadores de Horvitz-Thompson (1952) basados en s_i ($i = A, B, C$), $\widehat{Cov}_{i\in s_A}(x,y)$ y $\widehat{Var}_{i\in s_A}(x)$ denotan los estimadores de la covarianza y varianza basados en s_A. Los valores óptimos $\widehat{\alpha}_{reg}$ y $\widehat{\beta}_{reg}$ pueden consultarse en Rueda y González (2004).

2.3.2. Estimador propuesto

A continuación se definen algunos estimadores de tipo Hájek. Las propiedades más importantes de este tipo de estimadores están descritas en Rao (1966), Basu (1971) y Särndal *et al.* (1992).

$$\bar{y}^{A}_{w} = \sum_{i\in s_A} d^{A*}_{i} y_i \quad ; \quad \bar{y}^{C}_{w} = \sum_{i\in s_C} d^{C*}_{i} y_i \quad ; \tag{2.53}$$

$$\bar{y}^{AC}_{w} = \sum_{i\in s_A\cup s_C} d^{AC*}_{i} y_i;$$

$$\bar{x}^{A}_{w} = \sum_{i\in s_A} d^{A*}_{i} x_i \quad ; \quad \bar{x}^{B}_{w} = \sum_{i\in s_B} d^{B*}_{i} x_i \quad ; \tag{2.54}$$

$$\bar{x}^{AB}_{w} = \sum_{i\in s_A\cup s_B} d^{AB*}_{i} x_i;$$

con

$$d^{A*}_{i} = \frac{d^{A}_{i}}{\sum_{j\in s_1} d^{A}_{j}} \quad , \quad d^{B*}_{i} = \frac{d^{B}_{i}}{\sum_{j\in s_B} d^{B}_{j}} \quad , \tag{2.55}$$

$$d^{C*}_{i} = \frac{d^{C}_{i}}{\sum_{j\in s_C} d^{C}_{j}},$$

$$d^{AB*}_{i} = \frac{d^{AB}_{i}}{\sum_{j\in s_A\cup s_B} d^{AB}_{j}} \quad , \tag{2.56}$$

$$d^{AC*}_{i} = \frac{d^{AC}_{i}}{\sum_{j\in s_A\cup s_C} d^{AC}_{j}},$$

$$d^{A}_{i} = 1/\pi^{A}_{i},\ d^{B}_{i} = 1/\pi^{B}_{i}, \tag{2.57}$$

$$d^{C}_{i} = 1/\pi^{C}_{i},\ d^{AB}_{i} = 1/\pi^{AB}_{i},\ d^{AC}_{i} = 1/\pi^{AC}_{i}.$$

Las cantidades π^{A}_{i}, π^{B}_{i}, π^{C}_{i}, π^{AB}_{i} y π^{AC}_{i} son, respectivamente, las probabilidades de inclusión de primer orden de las muestras s_A, s_B, s_C, $s_A \cup s_B$ y $s_A \cup s_C$.

Cuando $u_i = 0$ (sin usar información auxiliar), se obtiene $\widehat{p}_i = d^{*}_{i}$ y el estimador de verosimilitud pseudo empírico ($PEMLE$) coincide con el estimador de tipo Hájek dado por $\sum_{i\in s} d^{*}_{i} y_i$. Este estimador no usa la variable auxiliar x.

Sea el $PEMLE$ de $\bar{Y}$ dado por

$$\overline{y}^{A}_{PE} = \sum_{i\in s_A} \widehat{p}^{A}_{i} y_i,$$

donde $\widehat{p}^{A}_{i}$ maximiza $l(p^A) = \sum_{i\in s_A} d^{A}_{i} \log p^{A}_{i}$ sujeta a

$$\sum_{i\in s_A} p^{A}_{i} = 1 \quad (0 \le p^{A}_{i} \le 1), \tag{2.58}$$

$$\sum_{i\in s_A} p^{A}_{i} u_i = 0. \tag{2.59}$$

Considerando el método de multiplicadores de Lagrange, $\widehat{p}^{A}_{i}$ está dado por

$$\widehat{p}^{A}_{i} = \frac{d^{A*}_{i}}{1+\lambda^{A}u_i}, \text{ para } i \in s_A, \tag{2.60}$$

donde el vector de multiplicadores de Lagrange, λ^A, se obtiene de la ecuación

$$\sum_{i\in s_A} \frac{d^{A*}_{i} u_i}{1+\lambda^{A}u_i} = 0. \tag{2.61}$$

El estimador $\overline{y}^{A}_{PE}$ no usa la información de las muestras s_B y s_C. A continuación se define un $PEMLE$ que considera la información de s_A y s_B. Como la variable de interés contiene $n-p-q$ valores, el nuevo vector de pesos $\widehat{p}^{AB}_{i}$ debe definirse con dimensión $n-p-q$. Así, el nuevo estimador está dado por

$$\overline{y}^{AB}_{PE} = \sum_{i\in s_A} \widehat{p}^{AB}_{i} y_i,$$

donde $\widehat{p}^{AB}_{i}$ ($i \in s_A$) se obtiene como $\widehat{p}^{A}_{i}$ (el cual tiene dimensión $n-p-q$), aunque en este caso se usa el vector de multiplicadores de Lagrange λ^{AB}, el cual está basado en las muestras s_A y s_B, en la expresión (2.60). λ^{AB} se obtiene de (2.61) después de sustituir d^{A*}_{i} por d^{AB*}_{i}.

Pueden usarse otros métodos como el de imputación para obtener el $PEMLE$ basado en las muestras s_A y s_B, aunque éstos no están relacionados con el método de verosimilitud empírica.

Aunque $\overline{y}^{AB}_{PE}$ parece mejor estimador que $\overline{y}^{A}_{PE}$ al usar información de las muestras s_A y s_B, este estimador no resulta apropiado porque las condiciones $\sum_{i\in s_A} \widehat{p}^{AB}_{i} = 1$ y $\sum_{i\in s_A} \widehat{p}^{AB}_{i} u_i = 0$ no se cumplen. El estimador no queda bien construido y las ventajosas propiedades del método de verosimilitud empírica no se sostienen. En el estudio empírico de la Sección 2.3.4 puede confirmase esta observación.

Desafortunadamente, el estimador propuesto $\overline{y}^{A}_{PE}$ no usa información de la variable de estudio y proporcionada por la muestra s_C. Para resolver este problema, se propone una clase de estimadores que usan toda la información de la variable y incluida en las muestras s_A y s_C (véase también Rueda, Muñoz, Berger, Arcos y Martínez, 2006). Esta clase viene dada por

$$\overline{y}_{PE\alpha} = \alpha\overline{y}^{A}_{PE} + (1-\alpha)\bar{y}^{C}_{w}, \tag{2.62}$$

donde α es una constante debidamente escogida que verifica $0 < \alpha < 1$. En la Sección 2.3.3, se proponen valores apropiados para α. El estimador $\bar{y}^{C}_{w}$ está definido en (2.53).

Se observa que si $\alpha = 1$, el estimador resultante es $\overline{y}^{A}_{PE}$, y por tanto, este estimador está incluido en la clase $\overline{y}_{PE\alpha}$.

Cualquier estimador de esta clase usa toda la información disponible de las muestras s_A y s_C sin usar técnicas de imputación. Los valores de x de la muestra s_B no se usan para la estimación. No obstante, los valores de la variable y están perdidos para $i \in s_B$. Incluir esta información en la clase considerando $\overline{y}_{PE}^{AB}$ en lugar de $\overline{y}_{PE}^{A}$ empeoraría las estimaciones. En la Sección 2.3.4, un estudio de simulación muestra que los estimadores de la clase propuesta son tan eficientes como otros estimadores que usan la información de cada muestra (s_A, s_B y s_C).

2.3.3. Propiedades teóricas

En esta sección se demuestra que el estimador $\overline{y}_{PE\alpha}$ propuesto en (2.62) es asintóticamente insesgado. La varianza asintótica de $\overline{y}_{PE\alpha}$ también se deriva.

Sean las siguientes condiciones.

(C2.15). $u^{A*} = \max_{i\in s_A} |u_i| = o_p(n^{1/2})$.

(C2.16). $\dfrac{\sum_{i\in s_A} d_i^A u_i}{\sum_{i\in s_A} d_i^A u_i^2} = O_p(n^{1/2})$.

Estas condiciones fueron usadas por Chen y Sitter (1999), los cuales demuestran que varios diseños muestrales más comunes las satisfacen. Dadas estas condiciones, el siguiente resultado puede obtenerse.

Corolario 2.3 *Bajo las condiciones (C2.15) y (C2.16), se tiene que*

$$\overline{y}_{PE\alpha} = \alpha\bar{y}_{GREG}^{A} + (1-\alpha)\bar{y}_w^C + o_p(n^{-1/2}) \qquad (2.63)$$

donde

$$\bar{y}_{GREG}^{A} = \overline{y}_w^A + (\bar{X} - \overline{x}_w^A)b, \qquad (2.64)$$

con

$$b = \frac{\sum_{i\in s_A} d_i^{A*} x_i y_i - \overline{y}_w^A \overline{x}_w^A}{\sum_{i\in s_A} d_i^{A*}(x_i - \overline{x}_w^A)^2}. \qquad (2.65)$$

Demostración

Chen y Sitter (1999) demostraron que $\overline{y}_{PE}^{A}$ es asintóticamente equivalente a $\bar{y}_{GREG}^{A}$. Sabido esto, este resultado se sigue fácilmente. □

Teorema 2.5 *Bajo las condiciones (C2.15) y (C2.16), se tiene que*

$$\bar{y}_{GREG}^{A} \doteq \overline{y}_{GREG}^{A2},$$

donde

$$\overline{y}_{GREG}^{A2} = \overline{y}_w^A + (\bar{X} - \overline{x}_w^A)B, \qquad (2.66)$$

con

$$B = \frac{Cov(x,y)}{Var(x)}. \qquad (2.67)$$

Demostración

Para establecer este resultado, se asume que la población finita envuelve una sucesión de poblaciones donde n y N aumentan de modo que $n/N \to f$ cuando $n \to \infty$ y donde f es una constante.

Randles (1982) demostró que el comportamiento asintótico de algunas familias comunes de estadísticos podía establecerse aunque algunos parámetros vitales en la formulación del estadístico fuesen desconocidos. Este autor demostró que si $T_n(\widehat{\lambda})$ es una función de datos que usa el estimador $\widehat{\lambda}$, el cual también es una función de los datos que estima consistentemente el parámetro λ, entonces $T_n(\widehat{\lambda})$ y $T_n(\lambda)$ tienen la misma distribución límite y se verifica

$$\frac{\partial \mu(\gamma)}{\partial \gamma}\Big|_{\gamma=\lambda} = 0,$$

donde $\mu(\gamma) = \lim_{n\to+\infty} E_\lambda[T_n(\gamma)]$ y la esperanza es considerada cuando el verdadero parámetro es λ.

Sea $T_n(\gamma) = \bar{y}_w^A + (\bar{X} - \bar{x}_w^A)\gamma$. Notamos que $T_n(b) = \bar{y}_{GREG}^A$ ha sido establecido en (2.64). Consideremos $\mu(\gamma) = \lim_{n\to\infty} E_\gamma[T_n(\gamma)]$. Notamos que cuando $\gamma = B$, el cual está definido en (2.67), se obtiene $\mu(B) = \tilde{\bar{Y}}$ donde $\tilde{\bar{Y}} = \lim_{n\to\infty} \bar{Y}$. Puesto $\mu(\gamma)$ verifica

$$\frac{\partial \mu(\gamma)}{\partial \gamma}\Big|_{\gamma=B} = 0,$$

esto implica que $\bar{y}_{GREG}^A \doteq \bar{y}_{GREG}^{A2}$. Esto completa la demostración. □

Usando el Corolario 2.3 y el Teorema 2.5 se obtiene

$$\bar{y}_{PE\alpha} \doteq \alpha\bar{y}_{GREG}^{A2} + (1-\alpha)\bar{y}_w^C, \qquad (2.68)$$

el cual implica que $\bar{y}_{PE\alpha}$ es asintóticamente insesgado.

Teorema 2.6 *Bajo las condiciones (C2.15) y (C2.16), la varianza asintótica de $\bar{y}_{PE\alpha}$ está dada por*

$$AV(\bar{y}_{PE\alpha}) = \alpha^2\Big[V(\bar{y}_w^A) + B^2V(\bar{x}_w^A) - 2BCov(\bar{y}_w^A, \bar{x}_w^A)\Big] + \qquad (2.69)$$
$$+(1-\alpha)^2V(\bar{y}_w^C) + 2\alpha(1-\alpha)\Big[Cov(\bar{y}_w^A, \bar{y}_w^C) - BCov(\bar{x}_w^A, \bar{y}_w^C)\Big].$$

Demostración

La aproximación (2.68) implica que la varianza asintótica de $\bar{y}_{PE\alpha}$ está dada por

$$V\Big(\alpha\bar{y}_{GREG}^{A2} + (1-\alpha)\bar{y}_w^C\Big) = \qquad (2.70)$$

$$\alpha^2V(\bar{y}_{GREG}^{A2}) + (1-\alpha)^2V(\bar{y}_w^C) + 2\alpha(1-\alpha)Cov(\bar{y}_{GREG}^{A2}, \bar{y}_w^C).$$

Usando (2.66), la varianza de $\bar{y}_{GREG}^{A2}$ es

$$\begin{aligned} V(\bar{y}_{GREG}^{A2}) &= V\Big(\bar{y}_w^A + (\bar{X} - \bar{x}_w^A)B\Big) \qquad (2.71)\\ &= V\Big(\bar{y}_w^A - \bar{x}_w^A B\Big)\\ &= V(\bar{y}_w^A) + B^2V(\bar{x}_w^A) - 2BCov(\bar{y}_w^A, \bar{x}_w^A). \end{aligned}$$

El valor $Cov(\bar{y}_{GREG}^{A2}, \bar{y}_w^C)$ está dado por

$$Cov(\bar{y}_{GREG}^{A2}, \bar{y}_w^C) = Cov(\bar{y}_w^A, \bar{y}_w^C) - BCov(\bar{x}_w^A, \bar{y}_w^C). \qquad (2.72)$$

Así de (2.70), (2.71) y (2.72), la varianza asintótica de $\bar{y}_{PE\alpha}$ está dada por (2.69). El Teorema 2.6 se sigue fácilmente. □

El estimador óptimo de la clase propuesta está dado por el estimador definido en (2.62) con un valor α que minimize la varianza asintótica dada por (2.69).

La varianza asintótica (2.69) puede expresarse como

$$AV(\bar{y}_{PE\alpha}) = \alpha^2M^* + (1-\alpha)^2N^* + 2\alpha(1-\alpha)L^*,$$

donde

$$M^* = V(\bar{y}_w^A) + B^2 V(\bar{x}_w^A) - 2BCov(\bar{y}_w^A, \bar{x}_w^A) \quad (2.73)$$

$$N^* = V(\bar{y}_w^C), \quad (2.74)$$

$$L^* = Cov(\bar{y}_w^A, \bar{y}_w^C) - BCov(\bar{x}_w^A, \bar{y}_w^C). \quad (2.75)$$

El valor α_{opt} que minimiza la varianza asintótica es solución de la ecuación

$$\frac{\partial AV(\bar{y}_{PE\alpha})}{\partial \alpha}\Big|_{\alpha=\alpha_{opt}} =$$

$$2\alpha_{opt} = M^* - 2(1-\alpha_{opt})N^* + 2(1-2\alpha_{opt})L^* = 0,$$

la cual implica

$$\alpha_{opt} = \frac{N^* - L^*}{M^* + N^* - 2L^*}. \quad (2.76)$$

Sustituyendo α_{opt} en (2.69), se obtiene la varianza asintótica mínima, dada por

$$AV(\bar{y}_{PE\alpha_{opt}}) = \alpha_{opt}^2 M^* + (1-\alpha_{opt})^2 N^* + 2\alpha_{opt}(1-\alpha_{opt})L^*. \quad (2.77)$$

Desafortunadamente, el valor óptimo α_{opt} depende de parámetros poblacionales desconocidos, los cuales pueden estimarse a partir de los datos muestrales.

Bajo muestreo aleatorio simple y muestreo estratificado, $\sum_{i\in s} d_i = N$, esto es, el estimador de Horvitz-Thompson y el estimador de tipo Hájek son idénticos, y por tanto, los estimadores de las varianzas y covarianzas de las expresiones (2.73), (2.74) y (2.75) pueden obtenerse fácilmente. Una expresión analítica para (2.73), (2.74) y (2.75) bajo muestreo aleatorio simple puede encontrarse en Rueda y González (2004).

Con estas estimaciones, puede obtenerse una aproximación $\tilde{\alpha}_{opt}$ de α_{opt}. Por lo tanto, la expresión del estimador propuesto viene dada por

$$\tilde{\bar{y}}_{PE\alpha_{opt}} = \tilde{\alpha}_{opt}\bar{y}_{PE}^A + (1-\tilde{\alpha}_{opt})\bar{y}_w^C. \quad (2.78)$$

También es posible establecer la insesgadez asintótica de $\tilde{\bar{y}}_{PE\alpha_{opt}}$.

2.3.4. Propiedades empíricas

En esta sección se comparan los estimadores propuestos con otros estimadores alternativos usando un estudio empírico basado en poblaciones reales y simuladas, usadas previamente en estudios de estimadores de regresión y razón, estimación de la varianza e intervalos de confianza.

Las poblaciones naturales usadas en este estudio son la Fam1500 y Hospitals (véase Apéndice A). Se recuerda que los coeficientes de correlación están dados por $\rho_{y,x_1} = 0{,}848$ y $\rho_{y,x_2} = 0{,}546$ en la población Fam1500 y $\rho_{y,x} = 0{,}911$ en la población Hospitals.

Paralelamente a Wu y Sitter (2001), se han generado cuatro poblaciones de $N = 2000$ unidades mediante muestras independientes e idénticamente distribuidas mediante el modelo

$$y = \theta_0 + \theta_1 x + \epsilon, \quad (2.79)$$

donde $x \sim Gamma(1,1)$, $\epsilon \sim N(0,\sigma^2)$ y $\theta_0 = \theta_1 = 1$. Los coeficientes de correlación están dados por 0.6, 0.7, 0.8 y 0.9, y las poblaciones se llaman Pop06, Pop07, Pop08 y Pop09, respectivamente. Pueden consultarse más detalles de estas poblaciones en el Apéndice A.

La precisión de los estimadores propuestos se ha analizado por medio de un estudio empírico, donde para cada población se han representado tres números diferentes de valores perdidos para la variable x, p. Varios valores perdidos de y, q, se han representado en el eje de abscisas. De este modo, el comportamiento de los estimadores puede observase para relaciones fuertes y débiles entre variables y diferentes situaciones de datos perdidos.

El comportamiento de los estimadores $\overline{y}_{PE}^A$ y $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ se compara con los siguientes estimadores: (i) el estimador estándar de tipo Hájek para la media poblacional basado en las muestras s_A y s_C, es decir, $\bar{y}_w^{AC}$; (ii) $\overline{y}_{T1}$, $\overline{y}_{T2}$, $\overline{y}_{T3}$ y $\overline{y}_{T4}$, los estimadores propuestos en Toutenburg y Srivastava (2000); (iii) $\overline{y}_{PE}^{AB}$, el $PEMLE$ basado en las muestras s_A y s_B. Aunque se ha señalado que los pesos no quedan bien definidos, se usa en el estudio de simulación para observar su comportamiento; (iv) $\overline{y}_{Reg}$, el estimador propuesto en Rueda y González (2004) basado en las muestras s_A, s_B y s_C.

Para cada una de las seis poblaciones, se han generado $B = 1000$ muestras independientes bajo muestreo aleatorio simple con tamaño muestral n. A continuación, se eliminan de la muestra p elementos de la variable auxiliar y q elementos de la variable de estudio de forma aleatoria. Bajo este escenario, las submuestras s_A, s_B y s_C pueden definirse fácilmente. El cumplimiento de todos los estimadores se mide en términos de Sesgo Relativo (SR) y de Eficiencia Relativa (ER), donde

$$SR_j = \frac{1}{B}\sum_{b=1}^{B}\frac{|\overline{y}_j(b) - \overline{Y}|}{\overline{Y}} \quad ; \quad ER_j = \frac{ECM(\overline{y}_j)}{ECM(\overline{y}_w^{AC})},$$

b indica la b-ésima simulación, el Error Cuadrático Medio empírico está dado por

$$ECM(\overline{y}_j) = B^{-1}\sum_{b=1}^{B}(\overline{y}_j(b) - \overline{Y})^2,$$

y $j = 1,\ldots,8$ se refiere a los estimadores $\overline{y}_{PE}^A$, $\overline{y}_{PE}^{AB}$, $\overline{\overline{y}}_{PE\alpha_{opt}}$, $\overline{y}_{Reg}$, $\overline{y}_{T1}$, $\overline{y}_{T2}$, $\overline{y}_{T3}$ y $\overline{y}_{T4}$.

Las simulaciones se han llevado a cabo en R y los códigos se encuentran en el Apéndice **??**.

En primer lugar, se observa que el estimador $\overline{y}_{T3}$ posee una considerable ganancia en precisión respecto a los estimadores $\overline{y}_{T1}$, $\overline{y}_{T2}$ y $\overline{y}_{T4}$. Con el fin de obtener más claridad en las figuras, las líneas correspondientes a los estimadores $\overline{y}_{T1}$, $\overline{y}_{T2}$ y $\overline{y}_{T4}$ no se han incluido.

Las Figuras B.1, B.2 y B.3 representan los valores de la Eficiencia Relativa (eje de ordenadas) para los estimadores $\overline{y}_{PE}^A$, $\overline{y}_{PE}^{AB}$, $\widetilde{\overline{y}}_{PE\alpha_{opt}}$, $\overline{y}_{Reg}$ y $\overline{y}_{T3}$ bajo muestreo aleatorio simple y diferentes valores de p y q. Las líneas horizontales en el punto 1 representan la ER para $\bar{y}_w^{AC}$, el estimador estándar.

De estas figuras, se puede llegar a las siguientes conclusiones generales:

1. Si aumenta la relación entre y y x y, además, el número de datos faltantes es escaso, todos los estimadores (excepto $\overline{y}_{T3}$) obtienen mejores estimaciones con respecto al estimador estándar. Cuando ambos p y q incrementan, las estimaciones son

peores con respecto a $\overline{y}_w^{AC}$, y de ahí, que todas las líneas sean crecientes.

2. Los mejores resultados se consiguen a través del estimador $\widetilde{\overline{y}}_{PE\alpha_{opt}}$, esto es, el ECM es siempre menor que el resto de estimadores y siempre mejora considerablemente los resultados proporcionados por el estimador directo $\overline{y}_w^{AC}$.
3. El peor comportamiento lo muestra el estimador de Toutenburg y Srivastava (2000). Esto puede deberse al hecho de que este estimador no usa $\overline{X}$ como información auxiliar.

Comparando entre los estimadores basados en el método de verosimilitud empírica, se observa

1. Los estimadores $\overline{y}_{PE}^{A}$ y $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ son equivalentes cuando existe una fuerte relación entre y y x y el número de datos perdidos es pequeño. La ganancia en eficiencia de $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ con respecto a $\overline{y}_{PE}^{A}$ es mayor en el caso contrario.
2. $\overline{y}_{PE}^{AB}$ nunca es mejor que los estimadores $\overline{y}_{PE}^{A}$ o $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ en términos de eficiencia. La razón para esto es que sus pesos no están bien definidos.

Un estimador que usa la información de s_A, s_B y s_C es $\overline{y}_{Reg}$. En las poblaciones Hospitals y Fam1500 (cuando se usa x_1), $\overline{y}_{PE}^{A}$, $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ y $\overline{y}_{Reg}$ son equivalentes. En el resto de los casos, $\overline{y}_{Reg}$ nunca mejora en eficiencia a $\widetilde{\overline{y}}_{PE\alpha_{opt}}$. Aunque $\overline{y}_{Reg}$ usa información de s_A, s_B y s_C, $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ es considerablemente más eficiente cuando la correlación entre y y x es baja y aumentan los valores de p y q.

Finalmente, comparamos el estimador propuesto con el estimador estándar:

1. $\overline{y}_w^{AC}$ es únicamente más eficiente que $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ cuando la relación entre variables es débil y el número total de datos perdidos, $p+q$, es alto. En este caso, el resto de estimadores obtienen significativamente peores estimaciones. Esto ocurre, por ejemplo, en Pop06, $p=80$, $q=60$, esto es, el 70 % de la muestra son valores perdidos. En la práctica, esta situación es improbable o inaceptable. No obstante, este caso se muestra para poder revelar el comportamiento de los estimadores en situaciones extremas.
2. Como se esperaba, cuando el número de valores de x perdidos, p, incrementa, la ganancia en precisión del estimador propuesto con respecto a $\overline{y}_w^{AC}$ es menor. Equivalentemente, cuando p permanece fijo, la ganancia en precisión decrece cuando el número de valores perdidos q aumenta. Este resultado es lógico porque si p/q es pequeño, se proporciona más información por la muestra s_C en relación con la muestra s_B, y $\overline{y}_w^{AC}$ también usa la información de s_C.

Las Figuras B.4, B.5 y B.6 muestran los valores del Sesgo Relativo (SR) para todos los estimadores. Puede observarse que los valores SR están todos en un rango razonable, teniendo los estimadores $\overline{y}_{PE}^{A}$ y $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ el mejor comportamiento en términos de SR. Estas figuras presentan similares resultados que la ER, y por tanto, se puede llegar a las mismas conclusiones.

En resumen, estas simulaciones muestran como un uso apropiado de las muestras s_A y s_C por el estimador propuesto puede reducir el error de los estimadores directo, regresión, de verosimilitud pseudo empírica, etc. Por tanto, el estimador propuesto $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ es una alternativa óptima para la estimación de parámetros lineales en presencia de datos faltantes y con un buen uso de la información auxiliar.

2.4. Estimación de la función de distribución

2.4.1. Introducción

El problema de la estimación de la función de distribución es un tema actual y muy importante del muestreo en poblaciones finitas, por tratarse de una función que permite determinar las características más importantes de la población en estudio, proporcionando información relevante acerca del comportamiento global de la población. Sin duda, los estimadores estudiados clásicamente en la teoría del muestreo, como totales, medias, proporciones y varianzas, no ofrecen tanta información como la función de distribución, aunque obtener estimadores eficientes para tal función no es tan simple como en el caso de los estimadores puntuales.

La estimación de cuantiles y de otros parámetros de tipo no funcional también queda resuelto con el conocimiento de la función de distribución. Los cuantiles, por ejemplo, pueden obtenerse mediante inversión directa de la función de distribución. Además, permite obtener medidas importantes como la determinación de las líneas de pobreza, proporción de bajos ingresos, etc. y son muy útiles en investigaciones de tipo social o económico. Debido a la importancia de estos parámetros en algunas investigaciones o estudios, se debe disponer de buenos métodos y técnicas para obtener las mejores estimaciones posibles.

Recordemos que la función de distribución para una variable de interés, y, y una población finita, U, es la proporción de unidades en U para las cuales el valor de y es menor o igual que t. El problema de la estimación de la función de distribución en la presencia de información auxiliar ha recibido recientemente mucha atención debido a las importantes propiedades que posee, el interés considerable que tiene cuando, por ejemplo, y es una medida de gastos o ingresos, etc.

La función de distribución poblacional,

$$F_y(t) = \frac{1}{N}\sum_{i=1}^{N}\delta(t-y_i), \tag{2.80}$$

satisface las siguientes condiciones:

(C2.17). $\lim_{t\to-\infty} F_y(t) = 0 \quad ; \quad \lim_{t\to+\infty} F_y(t) = 1.$

(C2.18). $F_y(t)$ es monótona no-decreciente: $\forall\ t_1 < t_2$, $F_y(t_1) \leq F_y(t_2)$.

(C2.19). $F_y(t)$ es continua por la derecha: Dado $t > t^*$, $\lim_{t\to t^*} F_y(t) = F_y(t^*)$.

Varios de los estimadores propuestos en la literatura del muestreo en poblaciones finitas no satisfacen todas estas propiedades y no son, por tanto, funciones de distribución. Por ejemplo, la función de distribución estimada mediante el método de calibración no cumple los requisitos necesarios para ser una verdadera función de distribución.

Asumamos que la variable de estudio, y, está altamente asociada con un vector auxiliar de variables, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip}, \ldots, x_{iP})^t$, donde los valores $\mathbf{x}_1$,...,$\mathbf{x}_N$ son conocidos para toda la población. Como se ha comentado en varias ocasiones, en las investigaciones por muestreo es común el uso de está información poblacional auxiliar en la etapa de estimación para incrementar la precisión de los estimadores de una media o un total. Bajo este escenario, el uso de la información auxiliar ha sido extensamente estudiado, pero bastante menos ha sido el esfuerzo por aplicarlo a la estimación de la función de distribución y cuantiles poblacionales. Notamos que la aplicación de las técnicas usuales para la estimación de medias y totales en el escenario de la estimación de la función de distribución producen resultados no deseables y, en general, con una pérdida significativa en eficiencia.

Por otro lado, el número de variables auxiliares a usar en la etapa de estimación es otro punto de vista interesante en la estimación de la función distribución. Algunos de los estimadores en la literatura están construidos para una única variable auxiliar, y el uso de otras variables auxiliares resulta imposible o con un alto coste computacional. Si estas variables presentan una fuerte relación con la variable de estudio, éstas deberían incluirse en el estudio y parece razonable asumir que podrían obtenerse mejores propiedades. Estos estimadores tienen la desventaja de la pérdida de eficiencia provocada por el hecho de no poder usar esta información auxiliar multivariante. Estas consideraciones sugieren que un uso más eficiente de la información auxiliar en la etapa de estimación es posible en el problema de la estimación de la función de distribución.

Sabemos que el método de verosimilitud pseudo empírica es una técnica reciente que puede usarse para la estimación de medias o totales poblacionales (Chen y Qin, 1993, Chen y Sitter, 1999), funciones de distribución (Chen y Wu, 2002, Wu, 2003) y otros parámetros. Asumiendo este método, Chen y Wu (2002) propusieron estimadores modelo-calibrados para estimar la función de distribución. Estos estimadores están construidos por medio de restricciones que requieren el uso de un valor fijado t_0. Estos estimadores sufren una considerable pérdida de eficiencia cuando t_0 se encuentra alejado de t, el punto donde se evalúa la función de distribución. El estimador propuesto en la Sección 2.4.3 emplea el método de verosimilitud empírica y permite el uso de información auxiliar multivariante. Este estimador está basado en una aproximación modelo-asistida. Además, se usa un conjunto apropiado de puntos en las restricciones para evitar el problema de la pérdida de eficiencia.

2.4.2. Algunos estimadores de la función de distribución

En este apartado se describen los principales trabajos y enfoques relacionados con la estimación de la función de distribución poblacional. Destacamos las propiedades más importantes de estos estimadores, prestando especial interés a los estimadores modelo-calibrados de verosimilitud empírica. Estos últimos presentan bastantes similitudes con el estimador propuesto en la Sección 2.4.3, por lo que señalaremos las principales diferencias entre unos y otros. Todos los estimadores que se exponen a continuación están basados en distintas aproximaciones. Aprovecharemos la ocasión para describir los tipos de inferencias que existen recientemente en muestreo de poblaciones finitas.

En la expresión (2.80) se observa que la función de distribución puede verse como una media poblacional de la variable $z_i = \delta(t - y_i)$, y por tanto, sin utilizar ningún tipo de información auxiliar, la estimación de la función de distribución es un caso especial de la estimación de la media poblacional. Haciendo uso de esta perspectiva, los estimadores más conocidos son el de Horvitz y Thompson (1952), dado por

$$\widehat{F}_{HTy}(t) = \frac{1}{N}\sum_{i\in s} d_i\delta(t - y_i),$$

y el estimador de tipo Hájek dado por

$$\widehat{F}_{HKy}(t) = \frac{\sum_{i\in s} d_i\delta(t - y_i)}{\sum_{j\in s} d_j} = \sum_{i\in s} d_i^*\delta(t - y_i),$$

donde $d_i = 1/\pi_i$. Nótese que el estimador de Horvitz y Thompson puede usarse únicamente cuando el tamaño poblacional es conocido, mientras que el de tipo Hájek puede emplearse en ambas situaciones. Bajo cualquier diseño muestral en el cual $\sum_{i\in s} d_i = N$, puede demostrarse que $\widehat{F}_{HTy}(t) = \widehat{F}_{HKy}(t)$.

En presencia de información auxiliar, Rao *et al.* (1990) propusieron dos nuevos estimadores basados en el diseño muestral: el estimador de tipo razón dado por

$$\widehat{F}_r(t) = \frac{1}{N}\frac{\sum_{i\in s} d_i\delta(t - y_i)}{\sum_{i\in s} d_i\delta(t - \widehat{R}x_i)}\sum_{i\in U}\delta(t - \widehat{R}x_i), \qquad (2.81)$$

y el estimador diferencia dado por

$$\widehat{F}_d(t) = \frac{1}{N}\left\{\sum_{i\in s} d_i\delta(t - y_i) + \sum_{i\in U}\delta(t - \widehat{R}x_i) - \right. \qquad (2.82)$$

$$\left. - \sum_{i\in s} d_i\delta(t - \widehat{R}x_i)\right\}$$

donde

$$\widehat{R} = \frac{\sum_{i\in s} d_i y_i}{\sum_{i\in s} d_i x_i}.$$

Se observa que ambos estimadores utilizan como información auxiliar la variable $\delta(t - \widehat{R}x)$.

Al no utilizar ningún tipo de información auxiliar, los estimadores $\widehat{F}_{HTy}(t)$ y $\widehat{F}_{HKy}(t)$ son menos eficientes que $\widehat{F}_r(t)$ y $\widehat{F}_d(t)$, pero sin embargo, éstos últimos tienen el inconveniente de dar valores, por lo general, fuera del rango

$[0,1]$ y no siempre son funciones monótonas respecto a t, con lo que no cumplen las propiedades de la función de distribución. Por este motivo, son numerosos los casos en los que la inversión directa de $\widehat{F}_r(t)$ y $\widehat{F}_d(t)$ no produce buenas estimaciones para los cuantiles.

En Rao *et al.* (1990) y en Francisco y Fuller (1991) se propone transformar $\widehat{F}_d(t)$ en una función monótona antes de obtener estimaciones para los cuantiles. Estos procesos tienen básicamente dos inconvenientes: (i) no son transformaciones triviales y (ii) se desconoce la pérdida de eficiencia al realizar la transformación.

Otro estimador para la función de distribución bastante reciente es el obtenido mediante el método de calibración descrito en Deville y Särndal (1992). Al igual que los anteriores que utilizan información auxiliar tienen la propiedad no deseable de no ser una auténtica función de distribución. Esto se debe a que los pesos que se utilizan para ponderar las unidades muestrales de la variable de interés, $\delta(t-y_i)$, pueden ser negativos, y por tanto, el estimador resultante puede llegar a ser decreciente. Además se demuestra que su límite cuando $t \to +\infty$ es distinto de 1.

Por tanto, es deseable requerir que un estimador para la función de distribución sea por si mismo una verdadera función de distribución. Nótese, que una verdadera función de distribución debe satisfacer las condiciones (C2.17), (C2.18) y (C2.19).

El conocido estimador de regresión generalizado ($GREG$) (Cassel *et al.*, 1976, 1977, Särndal, 1980, Deng y Wu, 1987, Särndal *et al.*, 1989) es un estimador modelo-asistido que está basado en un modelo lineal. Más recientemente, son dos los principales métodos en la literatura que están categorizados como aproximaciones modelo-asistidas. Estos procedimientos son el de calibración (Deville y Särndal, 1992) y el de verosimilitud empírica (Chen y Qin, 1993, Chen y Sitter, 1999). Notamos que estos procedimientos no son dependientes de un modelo, aunque usan uno de ellos para construir el estimador. En otras palabras, los estimadores modelo-asistidos son aproximadamente (asintóticamente) insesgados bajo el diseño, independientemente de si el modelo es correcto o no, y son particularmente eficientes si el modelo en el que se basa es correcto. Así, la aproximación modelo-asistida proporciona inferencias válidas bajo el modelo asumido y al mismo tiempo, está protegido contra una mala especificación del modelo en el sentido de proporcionar inferencias válidas basadas en el diseño, independientemente de la relación de la variable de interés con la variable auxiliar. Un ejemplo de estimadores modelo-asistidos para la función de distribución son los estimadores $\widehat{F}_r(t)$ y $\widehat{F}_d(t)$.

Otro procedimiento para estimar parámetros lineales o no lineales en poblaciones finitas es la aproximación basada en modelos, la cual asume un modelo de superpoblación y donde los estimadores son dependientes del modelo. Chambers y Dunstan (1986) y Dorfman y Hall (1993) propusieron estimadores basados en modelos para la función de distribución. El estimador de Chambers y Dunstan presenta el inconveniente de ser inconsistente bajo el diseño. Además, se necesita llevar a cabo un cuidadoso contraste sobre el modelo antes de que estos estimadores sean usados. Todos estos métodos presentan un grado de dificultad en la computación y un pobre cumplimiento cuando el modelo especificado es incorrecto. Bajo muestreo aleatorio simple, Wang y Dorfman (1996) combinaron los estimadores de Chambers y Dunstan (1986) con estimadores de tipo diferencia basados en el diseño en un estimador híbrido, que bajo ciertas condiciones, es más eficiente que ambos estimadores. No obstante, este estimador hereda las desventajas de ambos estimadores y tiene una complicada generalización a diseños muestrales más complejos. Silva y Skinner (1995) llevaron a cabo un estudio exhaustivo de las propiedades del estimador, y destacaron algunos problemas importantes, como por ejemplo, la pérdida en eficiencia cuando este estimador se usa en la estimación de cuantiles.

Finalmente, la recientemente desarrollada aproximación modelo-calibrada (Wu y Sitter, 2001) puede también usarse en las investigaciones por muestreo. Estos estimadores se obtienen, en primer lugar, adaptando un modelo de superpoblación, y a continuación, usando los valores estimados mediante este modelo en la etapa de estimación. Por tanto, si para una población dada se conoce el modelo de superpoblación asociado o un modelo que se ajuste bastante bien a dicha población, entonces puede resultar interesante utilizar la perspectiva modelo-calibrada para la estimación de la función de distribución poblacional mediante el método de verosimilitud empírica.

Chen y Wu (2002) plantean una aproximación modelo-calibrada para obtener tres estimadores de la función de distribución usando el método de verosimilitud empírica y tres modelos de superpoblación distintos. Estos modelos son bastantes generales, e incluyen los casos más importantes usados en muestreo. Bajo los modelos que se describen, estos estimadores tienen mínima esperanza bajo el modelo de la varianza asintótica basada en el diseño entre una clase de estimadores, es decir, son óptimos dentro de esa clase. Además, estos estimadores son asintóticamente insesgados bajo el diseño si se satisface el modelo y aproximadamente insesgados bajo el modelo. Por último, los estimadores resultantes son verdaderas funciones de distribución y permiten obtener cuantiles eficientemente mediante inversión directa.

Sea un modelo de superpoblación semi-paramétrico, ξ, en el cual se supone que la relación entre y y $\mathbf{x}$ puede describirse de la forma siguiente

$$E_\xi(y_i|\mathbf{x}_i) = \mu(\mathbf{x}_i,\theta),\ V_\xi(y_i|\mathbf{x}_i) = \sigma_i^2,\ \text{con } i = \{1,\ldots,N\},$$

donde θ es un vector de parámetros de la superpoblación. Para este vector, se puede obtener un estimador basado en el diseño, $\widehat{\theta}$, utilizando métodos generales para la estimación de ecuaciones (véase por ejemplo Godambe y Thompson, 1986 y Wu y Sitter, 2001).

Dado el modelo ξ, el estimador modelo-calibrado de verosimilitud empírica ($MCPE$) para la función de distribución viene dado por

$$\widehat{F}_{MCPE}(t) = \sum_{i\in s}\widehat{p}_i\delta(t-y_i) = \sum_{i\in s}\widehat{p}_i z_i, \qquad (2.83)$$

donde los pesos $\widehat{p}_i$ maximizan la función (2.11) sujeta a las restricciones (2.5) y (2.45). La función w_i de la restricción (2.45) viene dada por

$$w_i = E_\xi(z_i|\mathbf{x}_i) = E_\xi(\delta(t_0-y_i)|\mathbf{x}_i) = P(y_i \le t_0|\mathbf{x}_i).$$

El valor t_0 en la segunda restricción se considera fijo para conseguir que el estimador $\widehat{F}_{MCPE}(t)$ sea una verdadera función de distribución. Se pueden proponer otras expresiones para w_i, pero se ha considerado $w_i = E_\xi(z_i|\mathbf{x}_i)$ porque de entre todos los posibles valores $w_i = w(\mathbf{x}_i)$, el valor $w_i = E_\xi(z_i|\mathbf{x}_i)$ minimiza la esperanza bajo el modelo de la varianza asintótica basada en el diseño muestral.

En lo que sigue, se describen tres estimadores de verosimilitud pseudo empírica modelo-calibrados distintos para la función de distribución basados en diferentes modelos de superpoblación (véase Chen y Wu, 2002). Wu (2003) proporciona resultados de optimalidad para estos estimadores.

Estimadores bajo un modelo de regresión

Un modelo de superpoblación comúnmente usado en poblaciones finitas es el modelo de regresión, que viene dado por

$$y_i = \mu(\mathbf{x}_i, \theta) + \nu_i \varepsilon_i, \quad i = \{1, \ldots, N\}, \tag{2.84}$$

donde ν_i es una función conocida de $\mathbf{x}_i$, y ε_i, con $i = \{1, \ldots, N\}$, son variables aleatorias independientes e idénticamente distribuidas con media 0 y varianza σ^2.

Para un modelo de regresión lineal se tiene que $\mu(\mathbf{x}_i, \theta) = \mathbf{x}_i^t\theta$, aunque se puede considerar cualquier otro modelo no lineal. Sea θ_N y σ_N los estimadores de θ y σ, respectivamente, basados en los datos poblacionales. Se sabe que bajo un modelo de regresión lineal con varianzas homogéneas y θ de dimensión P, $\theta_N = (\mathbf{x}^t\mathbf{x})^{-1}\mathbf{x}^t\mathbf{y}$, donde $\mathbf{x}$ es la matriz de orden $N \times P$, $\mathbf{y} = (y_1, \ldots, y_N)^t$, y

$$\sigma_N^2 = \frac{(\mathbf{y} - \mathbf{x}\theta_N)^t(\mathbf{y} - \mathbf{x}\theta_N)}{(N - P)}.$$

Bajo el modelo (2.84), las cantidades w_i en (2.45) vienen dadas por

$$\begin{aligned} w_i^* = \quad & E_\xi(z_i|\mathbf{x}_i) = P(y_i \le t_0|\mathbf{x}_i) = \\ = \quad & P(\mu(\mathbf{x}_i, \theta_N) + \nu_i\varepsilon_i \le t_0) = \\ = \quad & G\left(\frac{t_0 - \mu(\mathbf{x}_i, \theta_N)}{\nu_i}\right), \end{aligned} \tag{2.85}$$

donde $G(\cdot)$ es la función de distribución de los términos ε_i, esto es,

$$G(t) = \frac{1}{N}\sum_{i=1}^{N} \delta(t - \varepsilon_i).$$

Como el vector θ_N es desconocido, es necesario buscar una estimación eficiente para poder obtener las cantidades w_i^*. Para este propósito, también es necesario una estimación de $G(\cdot)$. Una posible estimación viene dada por los residuos estimados, $\widehat{\varepsilon}_i$, y la función $G_n(\cdot)$, donde

$$\widehat{\varepsilon}_i = \frac{y_i - \mu(\mathbf{x}_i, \widehat{\theta})}{\nu_i},$$

$$G_n(t) = \sum_{i \in s} d_i^* \delta(t - \widehat{\varepsilon}_i) = \frac{\sum_{i \in s} d_i \delta(t - \widehat{\varepsilon}_i)}{\sum_{j \in s} d_j},$$

y $\widehat{\theta}$ es la estimación basada en el diseño para θ_N. En conclusión, se llega a que las cantidades w_i de la restricción (2.45) vienen dadas por

$$w_i = G_n\left(\frac{t_0 - \mu(\mathbf{x}_i, \widehat{\theta})}{\nu_i}\right). \tag{2.86}$$

En algunas situaciones, resulta razonable asumir que los términos de error ε_i en el modelo (2.84) están normalmente distribuidos. En este caso, se llega a que

$$w_i^* = \Phi\left(\frac{t_0 - \mu(\mathbf{x}_i, \theta_N)}{\nu_i \sigma_N}\right), \tag{2.87}$$

donde $\Phi(\cdot)$ es la función de distribución de la ley de probabilidad normal estándar. Se observa que se considera θ_N y no θ en la definición de w_i^*. Esto se hace para que las cantidades w_i^* estén bien definidas sobre la población y puedan tomar todos los argumentos posibles basados en el diseño. En la práctica, se sustituye θ_N y σ_N por $\widehat{\theta}$ y $\widehat{\sigma}$ respectivamente, donde éstas últimas cantidades son las estimaciones basadas en el diseño muestral de los parámetros desconocidos del modelo. De este modo, se llega a la expresión

$$w_i = \Phi\left(\frac{t_0 - \mu(\mathbf{x}_i, \widehat{\theta})}{\nu_i \widehat{\sigma}}\right). \tag{2.88}$$

En resumen, el estimador $MCPE$ según el modelo (2.84) está dado por $\widehat{F}_{MCPE}^{(1)}(t) = \sum_{i \in s} \widehat{p}_i \delta(t - y_i)$, donde los pesos $\widehat{p}_i$ maximizan la función (2.11) sujeta a las restricciones (2.5) y (2.45). Las cantidades w_i de la segunda restricción vienen dadas por (2.86), o por los valores (2.88) en caso de existir normalidad en los errores del modelo de superpoblación.

Estimadores bajo un modelo lineal generalizado

Resulta atractivo adaptar un modelo lineal generalizado a las cantidades $w_i = E_\xi(z_i|\mathbf{x}_i) = P(y_i \le t_0|\mathbf{x}_i)$. Para ello se considera el modelo de regresión logístico

$$\log\left(\frac{w_i}{1 - w_i}\right) = \mathbf{x}_i^t\theta, \tag{2.89}$$

con función varianza $V(w) = w(1 - w)$. Bajo este modelo, el parámetro poblacional θ_N puede definirse como una solución de las ecuaciones de estimación óptimas basadas en la población, esto es, $\sum_{i=1}^{N} \mathbf{x}_i(z_i^* - w_i) = \mathbf{0}$, donde $z_i^* = \delta(t_0 - t)$. Así,

$$w_i^* = \frac{\exp(\mathbf{x}_i^t\theta_N)}{1 + \exp(\mathbf{x}_i^t\theta_N)}. \tag{2.90}$$

Un estimador basado en el diseño, $\widehat{\theta}$, para el parámetro poblacional θ_N puede obtenerse resolviendo la correspondiente versión muestral del sistema anterior, esto es, $\sum_{i \in s} d_i \mathbf{x}_i(z_i^* - w_i) = \mathbf{0}$. De este modo, un segundo $MCPE$, esta vez bajo el modelo (2.89), viene dado por $\widehat{F}_{MCPE}^{(2)}(t) = \sum_{i \in s} \widehat{p}_i \delta(t - y_i)$, donde los pesos $\widehat{p}_i$ se obtienen considerando

$$w_i = \frac{\exp(\mathbf{x}_i^t\widehat{\theta})}{1 + \exp(\mathbf{x}_i^t\widehat{\theta})}. \tag{2.91}$$

El modelo de regresión logístico da una razonable estimación en la mayoría de las estimaciones.

Estimadores bajo valores pseudo estimados de un modelo semi-paramétrico

La variable $z_i = \delta(t - y_i)$ toma solamente valores 0 ó 1, pero los valores estimados w_i construidos bajo los modelos (2.84) y (2.89) están siempre entre 0 y 1. También es posible utilizar los llamados valores pseudo estimados $w_i = \delta(t_0 - \widehat{y}_i)$, los cuales también son variables dicotómicas y donde $\widehat{y}_i$ son valores estimados para y_i.

Bajo un modelo semi-paramétrico, $E_\xi(y_i|\mathbf{x}_i) = \mu_i$ y $V_\xi(y_i|\mathbf{x}_i) = \upsilon(\mu_i)$, donde $\mu_i = \mu(\mathbf{x}_i, \theta)$ y $\upsilon(\cdot)$ es una función varianza. Los valores estimados $\widehat{y}_i$ están dados por $\mu(\mathbf{x}_i, \widehat{\theta})$. Sea $h(\cdot)$ una conocida función de enlace tal que $h(\mu_i) = \mathbf{x}_i\theta$. $\widehat{\theta}$ es el estimador máximo verosímil que se obtiene del sistema de ecuaciones

$$\sum_{i\in s} \frac{d_i \mathbf{x}_i(y_i - \mu_i)}{\upsilon(\mu_i)h'(\mu_i)} = \mathbf{0},$$

donde $h'(u) = \partial h(u)/\partial u$. θ_N es la solución a

$$\sum_{i=1}^{N} \frac{\mathbf{x}_i(y_i - \mu_i)}{\upsilon(\mu_i)h'(\mu_i)} = \mathbf{0}.$$

Por tanto, el estimador viene dado por $\widehat{F}^{(3)}_{MCPE}(t) = \sum_{i\in s} \widehat{p}_i \delta(t - y_i)$, donde los pesos $\widehat{p}_i$ se obtienen usando los valores pseudo estimados

$$w_i = \delta(t_0 - \widehat{y}_i). \tag{2.92}$$

En la práctica se usan estas cantidades debido a que los valores

$$w_i^* = \delta(t_0 - \mu(\mathbf{x}_i, \theta_N)), \tag{2.93}$$

son desconocidos.

Bajo un modelo lineal simple con una única variable auxiliar, $\mu(\mathbf{x}, \theta) = \theta_0 + \theta_1 x_i$, y

$$\frac{1}{N}\sum_{i=1}^{N} w_i = \frac{1}{N}\sum_{i=1}^{N} \delta(t_0 - (\theta_0 + \theta_1 x_i)) = F_x\left(\frac{t_0 - \theta_0}{\theta_1}\right),$$

donde $F_x(t)$ es la función de distribución de la variable x. La restricción (2.45) se resume a

$$\sum_{i\in s} p_i \delta(t_0 - (\widehat{\theta}_0 + \widehat{\theta}_1 x_i)) = F_x\left(\frac{t_0 - \widehat{\theta}_0}{\widehat{\theta}_1}\right),$$

con lo que solamente se debe conocer la distribución de frecuencias de x para obtener $\widehat{F}^{(3)}_{MCPE}(t)$.

Notamos que puede usarse cualquier modelo de superpoblación. Si el modelo de superpoblación asociado a la población en estudio es otro distinto a cualquiera de estos tres, el planteamiento para el cálculo del estimador de verosimilitud pseudo empírica modelo-calibrado es similar a lo comentado. Bastaría con obtener las cantidades w_i óptimas bajo el modelo de superpoblación asociado.

La elección del valor t_0 es un aspecto importante, puesto que los estimadores son más precisos para estimar $F_y(t)$ cuando t está en las cercanías del punto t_0. En consecuencia, ningún w_i con un valor fijo t_0 puede ser uniformemente óptimo para $F_y(t)$ en todos los valores de t. El problema de encontrar un valor óptimo t_0 no se discute en Chen y Wu (2002). De hecho, sus correspondientes estudios empíricos usan cuantiles poblacionales de la variable y para obtener el valor t_0. Esta elección no puede realizarse en la práctica debido que los cuantiles poblacionales de la variable de estudio son desconocidos. En resumen, estos estimadores presentan dos inconvenientes principalmente: (i) es necesario el conocimiento de un modelo de superpoblación para los datos muestrales del estudio y (ii) se hace un uso poco eficiente de la información auxiliar, puesto que sería posible definir los estimadores usando más de un punto t_0, utilizando de este modo más información auxiliar, lo que conlleva esperar estimaciones más precisas. Estos problemas puede solventarse en gran medida mediante la metodología propuesta en la Sección 2.4.3, donde se usa un vector $\mathbf{t}_0$ para obtener estimaciones más eficientes para cualquier t.

El estimador que se propone para la función de distribución usa una aproximación modelo-asistida y el método de verosimilitud empírica. Con el objetivo de que este estimador sea más eficiente para cualquier t, éste usa un vector $\mathbf{t}_0$ basado en los cuartiles poblacionales de una pseudo-variable que es conocida en la práctica. Además, este estimador es una verdadera función de distribución y goza de una excelente ganancia en eficiencia como consecuencia de un uso efectivo de la información auxiliar. Éstas son dos de las ventajas más importantes del estimador propuesto.

2.4.3. Estimador propuesto modelo-asistido

En esta sección se propone usar la aproximación modelo-asistida basada en el método de verosimilitud empírica para construir un estimador de la función de distribución poblacional. La información auxiliar multivariante puede incorporarse en la etapa de estimación y se hace un uso efectivo de la información auxiliar. Este estimador basado en el diseño muestral es una auténtica función de distribución que disfruta de varias propiedades importantes.

Para construir el nuevo estimador para $F_y(t)$, se modifican los pesos del estimador $\widehat{F}_{HKy}(t)$, es decir d_i^*, por unos nuevos pesos $\widehat{p}_i$. Este conjunto de pesos se determina por medio de una aproximación modelo-asistida y usando las técnicas de verosimilitud empírica (Sección 2.2).

Se considera la estimación modelo-asistida porque esta aproximación proporciona un esquema de trabajo conveniente en el cual se pueden desarrollar estimadores muy precisos. A través de un modelo de superpoblación se construyen estimadores basados en la muestra que mejoran la precisión de las estimaciones cuando el modelo es correcto, pero que también mantiene propiedades importantes, tales como consistencia y una varianza estimable, cuando el modelo es incorrecto.

Se considera el usual modelo de regresión dado por

$$y_i = \beta^t \mathbf{x}_i + v_i\varepsilon_i, \quad i = 1, \ldots, N, \tag{2.94}$$

donde v_i es una función conocida de x_i y los valores ε_i son variables aleatorias independientes e idénticamente distribuidas con media 0 y varianza σ^2.

En la práctica, los valores del vector β son desconocidos. Mediante la teoría de regresión, puede deducirse que

el estimador de mínimos cuadrados de β (Särndal *et al.*, 1992)

$$\mathbf{B} = \left(\sum_{i \in U} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\sigma^2} \right)^{-1} \cdot \sum_{i \in U} \frac{\mathbf{x}_i y_i}{\sigma^2} \tag{2.95}$$

es el mejor estimador insesgado lineal de β bajo el modelo (2.94). **B** es un parámetro poblacional desconocido, pero puede estimarse usando los datos muestrales y aplicando el principio de estimación de las probabilidades de inclusión, esto es

$$\widehat{\beta} = \left(\sum_{i \in s} \frac{d_i \mathbf{x}_i \mathbf{x}_i^t}{\sigma^2} \right)^{-1} \cdot \sum_{i \in s} \frac{d_i \mathbf{x}_i y_i}{\sigma^2}. \tag{2.96}$$

El estimador propuesto modelo-asistido basado en el método de verosimilitud empírica se obtiene definiendo la pseudo-variable g, donde $g_i = \widehat{\beta}^t \mathbf{x}_i$, para $i \in s$. Esta variable puede considerarse como una predicción para y_i bajo el anterior modelo lineal.

Sean $t_{g25} = Q_g(0{,}25)$, $t_{g50} = Q_g(0{,}5)$ y $t_{g75} = Q_g(0{,}75)$ los cuartiles poblacionales de la variable g, donde $Q_g(\alpha) = \text{ínf}\{t \mid F_g(t) \geq \alpha\} = F_g^{-1}(\alpha)$. Bajo nuestro marco de trabajo, estas cantidades están disponibles, puesto que asumimos que la información auxiliar poblacional es conocida. El estimador de verosimilitud pseudo empírica modelo-asistido para la función de distribución se define como $\widehat{F}_{MA}(t) = \sum_{i \in s} \widehat{p}_i \delta(t - y_i)$, donde los nuevos pesos $\widehat{p}_i$ se obtienen maximizando $\widehat{l}(\mathbf{p})$ sujeta a las siguientes condiciones

$$\sum_{i \in s} p_i = 1, \quad (p_i > 0), \tag{2.97}$$

$$\sum_{i \in s} p_i \delta(t_{g25} - g_i) = \frac{1}{N} \sum_{i=1}^{N} \delta(t_{g25} - g_i) = F_g(t_{g25}) = 0{,}25, \tag{2.98}$$

$$\sum_{i \in s} p_i \delta(t_{g50} - g_i) = \frac{1}{N} \sum_{i=1}^{N} \delta(t_{g50} - g_i) = F_g(t_{g50}) = 0{,}5, \tag{2.99}$$

$$\sum_{i \in s} p_i \delta(t_{g75} - g_i) = \frac{1}{N} \sum_{i=1}^{N} \delta(t_{g75} - g_i) = F_g(t_{g75}) = 0{,}75. \tag{2.100}$$

Nótese que la idea de usar $\delta(t - g_i)$, para algún t, como una variable de calibración para construir restricciones tales como (2.98), (2.99) y (2.100) fue discutida, en primer lugar, por Wu y Sitter (2001) y posteriormente elaborada en Chen y Wu (2002).

Existen dos aspectos importantes relacionados con este o cualquier otro procedimiento de estimación. Éstos son la eficiencia y la consistencia. La eficiencia se refiere al cumplimiento del estimador en términos de sesgo y error cuadrático medio. En la Sección 2.4.5, se realiza una comparación de la eficiencia de $\widehat{F}_{MA}(t)$ con respecto otros estimadores conocidos. Las restricciones (2.98), (2.99) y (2.100) son requerimientos de consistencia altamente usados y son impuestos en la práctica porque resulta razonable pensar que los pesos que dan estimaciones perfectas para las variables auxiliares, deberían también dar una buena estimación para la variable de estudio.

La elección de t_{g25}, t_{g50} y t_{g75} en (2.98), (2.99) y (2.100) se realiza por varias razones. En primer lugar, esto está altamente relacionado con la existencia de la solución del método de verosimilitud empírica. Si se usaran más de tres valores t_0, esto es, un mayor número de restricciones, se podría llegar a problemas de existencia de solución (véase la Sección 2.4.4 para un mayor detalle). Estos puntos están también especificados por motivos de eficiencia. Si se usa un único punto t_0, $\widehat{F}_{MA}(t)$ será más eficiente para t en las proximidades de t_0. Para varios valores de t_0, es razonable asumir que si éstos están perfectamente distribuidos dentro del posible rango de valores de t, entonces, $\widehat{F}_{MA}(t)$ será más eficiente. Los valores t_{g25}, t_{g50} y t_{g75} exhiben una buena distribución y por tanto, $\widehat{F}_{MA}(t)$ será más preciso cuando t se encuentre en los alrededores de los cuartiles poblacionales de la variable g. Esto afecta a un alto rango de valores de la variable de estudio.

$\widehat{F}_{MA}(t)$ será, especialmente, más eficiente cuando t es igual a uno de los valores t_{g25}, t_{g50} ó t_{g75}. Esto implica que no hay una elección óptima de valores para todo t. Por otro lado, para t igual a t_{g25}, t_{g50} y t_{g75} y si el modelo (2.94) se ajusta perfectamente a la población de estudio, esto es, $y_i = \beta^t \mathbf{x}_i = g_i$, $i = 1, \ldots, N$, entonces $\delta(t - g_i) = \delta(t - y_i)$ y $\widehat{F}_{MA}(t)$ se reduce al valor exacto de $F_y(t)$. Es de esperar, que en el caso de una información auxiliar fuertemente relacionada con la variable de estudio, la correlación entre y_i y g_i será mayor, y consecuentemente, $\widehat{F}_{MA}(t)$ cumplirá mejor en el sentido de obtener estimaciones más precisas para $F_y(t)$.

Denotando por $\mathbf{t}_g = (t_{g25}, t_{g50}, t_{g75})^t$,

$$\delta(\mathbf{t}_g - g_i) = (\delta(t_{g25} - g_i), \delta(t_{g50} - g_i), \delta(t_{g75} - g_i))^t$$

y $\mathbf{K} = (0{,}25, 0{,}50, 0{,}75)^t$, las restricciones (2.98), (2.99) y (2.100) pueden expresarse por

$$\sum_{i \in s} p_i \delta(\mathbf{t}_g - g_i) = \mathbf{K} \tag{2.101}$$

o también como

$$\sum_{i \in s} p_i \mathbf{u}_i = \mathbf{0}, \tag{2.102}$$

donde $\mathbf{u}_i = \delta(\mathbf{t}_g - g_i) - \mathbf{K}$.

Mediante el conocido método de multiplicadores de Lagrange, puede demostrarse que la solución del problema de maximización sujeto a las condiciones (2.97) y (2.102) está dado por

$$\widehat{p}_i = \frac{d_i^*}{1 + \lambda^t \mathbf{u}_i}, \tag{2.103}$$

donde el multiplicador de Lagrange λ, cuya dimensión es tres, se obtiene de la ecuación

$$h(\lambda) = \sum_{i \in s} \frac{d_i^* \mathbf{u}_i}{1 + \lambda^t \mathbf{u}_i} = \mathbf{0}. \tag{2.104}$$

Puede demostrarse que, con probabilidad tendiendo a uno cuando el tamaño muestral va a infinito, existe una única solución a $h(\lambda) = \mathbf{0}$. Si tal solución existe, ésta puede encontrarse, por ejemplo, con el Algoritmo 2.1, el cual tiene garantizada la convergencia a la solución.

2.4.4. Propiedades teóricas

Un estimador modelo-asistido para la función de distribución se ha definido en la Sección 2.4.3. A continuación estudiamos varias propiedades de este estimador, las cuales pueden ser importantes en la práctica. En concreto, se estudia la existencia del estimador, se demuestra que $\widehat{F}_{MA}(t)$ es una verdadera función de distribución, se obtiene otra propiedad relacionada con la eficiencia del estimador propuesto y se establecen algunos resultados asintóticos.

Existencia del estimador

Existen dos aspectos computacionales por los cuales el estimador $\widehat{F}_{MA}(t)$ no pueda existir: (i) en la obtención del vector $\widehat{\beta}$ y (ii) para encontrar la solución a $h(\lambda) = \mathbf{0}$ en (2.104).

En el punto (i), $\widehat{\beta}$ siempre existe cuando se aplica información auxiliar univariante. En otro caso, $(\sum_{i\in s} d_i \mathbf{x}_i \mathbf{x}_i^t)^{-1}$ no puede calcularse si no es de rango completo. Esta situación es poco probable cuando $n \geq P$.

Respecto a la segunda cuestión, se ha comentado que puede emplearse el Algoritmo 2.1.

Para el caso de la estimación de la media poblacional, la variable $\mathbf{u}_i$ que usualmente se toma es $\mathbf{u}_i = \mathbf{x}_i - \overline{\mathbf{X}}$ (Chen y Sitter, 1999), la cual está también justificada por un modelo lineal. Bajo esta situación y usando el Algoritmo 2.1, $h(\lambda) = \mathbf{0}$ falla para proporcionar la solución si: (i) el vector de medias $\overline{\mathbf{X}}$ no es un punto interior del conjunto convexo formado por $\{\mathbf{x}_i, i \in s\}$, ó (ii) la matriz $\sum_{i\in s} d_i \mathbf{u}_i \mathbf{u}_i^t$ no es de rango completo.

En (i), el estimador de verosimilitud pseudo empírica no existe. Para el caso de estimar la media poblacional, esto ocurre con una probabilidad tendiendo a cero cuando el tamaño muestral tiende a infinito. En el escenario de la estimación de la función de distribución, la situación es bastante diferente. En particular, para el procedimiento propuesto, el vector $\mathbf{K}$ es siempre un punto interior del conjunto formado por $\{\delta(\mathbf{t}_g - g_i), i \in s\}$, puesto que los componentes de este vector son 0 ó 1, mientras que los componentes de $\mathbf{K}$ toman valores dentro de $[0, 1]$. Notamos que los componentes del vector $\delta(\mathbf{t}_g - g_i)$ no pueden ser todos 0 ó 1 para $i \in s$, salvo en situaciones extremas.

Sea $\mathbf{t}_0 = (t_{0(1)}, \ldots, t_{0(h)}, \ldots, t_{0(H)})^t$ otro vector diferente de $\mathbf{t}_g$ con similar o diferente dimensión y que puede usarse en restricciones como la dada por (2.101). Respecto al punto (ii), decir que resulta necesario una cuidadosa elección del vector $\mathbf{t}_0$ para evitar o eliminar el problema de multicolinealidad. En lo que sigue, se justifica la elección $\mathbf{t}_g = (t_{g25}, t_{g50}, t_{g75})^t$. En primer lugar, si se toman valores de $t_{0(h)}$ con dos ellos muy cercanos, entonces, resulta más probable que surga el problema de la multicolinealidad. Si se usan valores extremos de t_0 (o muy elevados o demasiados pequeños), la variable indicadora $\delta(\mathbf{t}_0 - g_i)$ podría tener todos sus elementos iguales a cero o a uno para $i \in s$, y por tanto, el método de verosimilitud empírica no tendría solución. Teniendo estas consideraciones en cuenta, la elección $\mathbf{t}_g = (t_{g25}, t_{g50}, t_{g75})^t$ resulta apropiada, puesto que cada punto está alejado del resto y además, estos puntos no se encuentran cercanos a los valores extremos de la variable g, evitando que la variable indicadora $\delta(\mathbf{t}_g - g_i)$ pueda contener valores que sean todos iguales a cero o a uno para $i \in s$. Bajo este planteamiento, el problema de la multicolinealidad es improbable. Notamos que este problema decrece conforme aumenta el tamaño muestral. Por ejemplo, no se ha observado problemas de multicolinealidad para el estimador $\widehat{F}_{MA}(t)$ en los estudios de simulación de la Sección 2.4.5, mientras que cuando se usa un vector $\mathbf{t}_0$ con dimensión mayor de 5, nos encontramos problemas de multicolinealidad para tamaños muestrales mayores de 50.

Como se comentó en la Sección 2.4.3, la elección $\mathbf{t}_g = (t_{g25}, t_{g50}, t_{g75})^t$ está también especificada por motivos de eficiencia. Además, el estimador $\widehat{F}_{MA}(t)$ es fácilmente computable debido a que el vector $\mathbf{t}_g$ es de dimension igual a 3 y por tanto, el sistema (2.104) presenta un número pequeño de ecuaciones.

$\widehat{F}_{MA}(t)$ es una auténtica función de distribución

La siguiente cuestión es comprobar si el estimador propuesto es una verdadera función de distribución. Para determinar esto, debemos verificar si se satisfacen, para $\widehat{F}_{MA}(t)$, las condiciones (C2.17), (C2.18) y (C2.19) de la Sección 2.4.1.

Resultado 2.1 *El estimador $\widehat{F}_{MA}(t)$ es una verdadera función de distribución.*

Demostración

Resulta fácil demostrar que la condición (C2.17) siempre se satisface si los pesos $\widehat{p}_i$, para $i = 1, \ldots, n$, son independientes de t:

$$\lim_{t\to-\infty} \widehat{F}_{MA}(t) = \lim_{t\to-\infty} \sum_{i\in s} \widehat{p}_i \delta(t - y_i) =$$

$$= \sum_{i\in s} \widehat{p}_i \lim_{t\to-\infty} \delta(t - y_i) = \sum_{i\in s} \widehat{p}_i 0 = 0.$$

$$\lim_{t\to+\infty} \widehat{F}_{MA}(t) = \lim_{t\to+\infty} \sum_{i\in s} \widehat{p}_i \delta(t - y_i) =$$

$$= \sum_{i\in s} \widehat{p}_i \lim_{t\to+\infty} \delta(t - y_i) = \sum_{i\in s} \widehat{p}_i = 1.$$

Por otro lado, $\widehat{F}_{MA}(t)$ es una función continua por la derecha y monótona no decreciente para unos pesos $\widehat{p}_i$ que sean independientes de t:

- Sea $t_1 < t_2$, entonces $\delta(t_1 - y_i) \leq \delta(t_2 - y_i)$ para $i \in s$ y $\widehat{F}_{MA}(t_1) = \sum_{i\in s} \widehat{p}_i \delta(t_1 - y_i) \leq \sum_{i\in s} \widehat{p}_i \delta(t_2 - y_i) = \widehat{F}_{MA}(t_2)$, puesto que $\widehat{p}_i$ son los mismos valores positivos para t_1 y t_2.
- Sea $t > t^*$, $\lim_{t\to t^*} \widehat{F}_{MA}(t) = \lim_{t\to t^*} \sum_{i\in s} \widehat{p}_i \delta(t - y_i) = \sum_{i\in s} \widehat{p}_i \lim_{t\to t^*} \delta(t - y_i) = \sum_{i\in s} \widehat{p}_i \delta(t^* - y_i) = \widehat{F}_{MA}(t^*)$.

Por tanto, las condiciones (C2.17), (C2.18) y (C2.19) se satisfacen para $\widehat{F}_{MA}(t)$ si el mismo conjunto de valores $\widehat{p}_i$ son usados para cada argumento t. Como $\widehat{F}_{MA}(t)$ asume un vector fijo $\mathbf{t}_g$, entonces, $\widehat{F}_{MA}(t)$ es una verdadera función de distribución. □

$\widehat{F}_{MA}(t)$ es igual a $F_y(t)$ cuando $x_i = y_i$

En las investigaciones por muestreo que incorporan muestreo sucesivo, la variable auxiliar es la misma que la variable principal, pero medida en un periodo anterior. En este caso, la información auxiliar incluye valores poblacionales de la variable x, los cuales pueden estar próximos a los valores de y. En tal situación, resulta razonable esperar que un estimador de $F_y(t)$ debería de aproximarse a $F_y(t)$ a medida que x se aproxima a y. Esta propiedad no se satisface para el estimador estándar, puesto que éste no hace uso de la información auxiliar.

Si $y_i = x_i$, puede verse que $\widehat{\beta} = \mathbf{1}$, $g_i = y_i$ y segunda restricción planteada para el estimador $\widehat{F}_{MA}(t)$ está dada por $\sum_{i\in s} p_i\delta(\mathbf{t}_g - y_i) = F_y(\mathbf{t}_g)$. Así, $\widehat{F}_{MA}(t) = \sum_{i\in s}\widehat{p}_i\delta(t-y_i)$ es exactamente igual a $F_y(t)$ si t coincide con uno de los valores de vector $\mathbf{t}_g$. Si esto no sucede, la igualdad, en general, no se cumple, aunque se esperan que las desviaciones sean pequeñas si el argumento t está próximo a un componente de $\mathbf{t}_g$.

Comportamiento asintótico

El siguiente paso es establecer el comportamiento asintótico del estimador $\widehat{F}_{MA}(t)$. Lamentablemente, este estimador usa los vectores $\mathbf{t}_g$ y $\widehat{\beta}$, que son dependientes de la muestra, lo que dificulta la obtención del comportamiento asintótico de este estimador. No obstante, es posible obtener algunos resultados para el estimador $\widehat{F}_{MA1}(t)$ que es muy similar al estimador propuesto aunque menos eficiente al utilizar menos información auxiliar. Este estimador se obtiene equivalentemente al estimador propuesto, con la diferencia de que los pesos $\widehat{p}_i$ están basados en las restricciones (2.97) y

$$\sum_{i\in s} p_i\delta(t_0 - g_i) = \frac{1}{N}\sum_{i=1}^{N}\delta(t_0 - g_i) = F_g(t_0), \quad (2.105)$$

para un valor cualquiera t_0 especificado.

Nota 2.1 *En caso de haber establecido propiedades asintóticas como la equivalencia con otros estimadores o la determinación de la varianza del estimador $\widehat{F}_{MA}(t)$, estas expresiones serían solamente válidas para muestras de gran tamaño y por tanto, serían poco útiles en la práctica. Habitualmente, la replicación de algún tipo, como Bootstrap, Jackknife o replicación mediante muestras balanceadas (Shao y Tu, 1995), es una alternativa que se usa en la etapa de estimación de la varianza, particularmente para la estimación de varianzas de funciones de distribución que son especialmente dificultosas. Tales procedimientos son fáciles de computar (Dalgleish, 1995) y además, han demostrado un buen cumplimiento para el método de verosimilitud empírica (Chen y Sitter, 1999) y para la estimación de la función de distribución (Lombardía* et al.*, 2003, Lombardía* et al.*, 2004).*

Teorema 2.7 *Cuando el vector $\widehat{\beta}$ se reemplaza por el parámetro $\boldsymbol{B}$ dado en (2.95), el correspondiente estimador de verosimilitud pseudo empírica modelo-asistido, $\widehat{F}^B_{MA1}(t)$, cuando se usa el punto $t_0 = t$, es asintóticamente equivalente a un estimador de tipo regresión generalizado:*

$$\widehat{F}^B_{MA1}(t) = \widehat{F}_{HKy}(t) + (F_b(t) - \widehat{F}_b(t))\widehat{D} + o_p(n^{-1/2}),$$

donde $\widehat{D} = \frac{\widehat{\sigma}_{z,w}}{\widehat{\sigma}^2_w} =$

$$= \frac{\sum_{i\in s} d_i^*[\delta(t-y_i) - \widehat{F}_{HKy}(t)][\delta(t-b_i) - \widehat{F}_b(t)]}{\sum_{i\in s} d_i^*[\delta(t-b_i) - \widehat{F}_b(t)]^2},$$

$b_i = \boldsymbol{B}^t\boldsymbol{x}_i$, $F_b(t)$ es la función de distribución de la variable b y $\widehat{F}_b(t)$ denota el estimador de tipo Hájek para la función de distribución de b en el punto t. z y w denotan las variables $\delta(t-y)$ y $\delta(t-b)$, respectivamente. Por tanto, $\widehat{F}_{MA1}(t)$ es asintóticamente insesgado bajo el diseño y tiene la misma varianza asintótica que el estimador de tipo regresión generalizado.

Demostración

Para demostrar este teorema, asumimos que la población finita está envuelta en una sucesión de poblaciones donde n y N aumentan de tal forma que $(n/N) \to f$ cuando $n \to \infty$. Además, se considera la variable de calibración $\delta(t-b_i)$ en (2.105) para construir $\widehat{F}_{MA1}(t)$. Sea $u_i = \delta(t-b_i) - F_b(t)$. Puesto que $|u_i| \leq 1$, las condiciones (C2.1) y (C2.2) del Teorema 2.3 se satisfacen y por tanto

$$\lambda = \frac{\sum_{i\in s} d_i^* u_i}{\sum_{i\in s} d_i^* u_i^2} + o_p(n^{-1/2}),$$

y $\widehat{p}_i = d_i^*(1-\lambda u_i) + o_p(n^{-1/2})$. Así:

$$\widehat{F}^B_{MA1}(t) = \sum_{i\in s}\widehat{p}_i\delta(t-y_i) + o_p(n^{-1/2}) =$$

$$\sum_{i\in s} d_i^*\left[1 - \frac{\left(\widehat{F}_b(t) - F_b(t)\right)u_i}{\sum_{i\in s} d_i^* u_i^2}\right]\delta(t-y_i) + o_p(n^{-1/2}) =$$

$$\sum_{i\in s} d_i^*\delta(t-y_i) - \frac{\widehat{F}_b(t) - F_b(t)}{\sum_{i\in s} d_i^* u_i^2}\sum_{i\in s} d_i^* u_i\delta(t-y_i) + o_p(n^{-1/2}) =$$

$$\widehat{F}_{HKy}(t) + (F_b(t) - \widehat{F}_b(t))\frac{\sum_{i\in s} d_i^* u_i\delta(t-y_i)}{\sum_{i\in s} d_i^* u_i^2} + o_p(n^{-1/2}) =$$

$$\widehat{F}_{HKy}(t) + (F_b(t) - \widehat{F}_b(t))\widehat{D} + o_p(n^{-1/2}).$$

□

El resultado anterior es válido cuando se usa el parámetro poblacional **B**. El siguiente resultado garantiza que el Teorema 2.7 también se cumple cuando usamos el parámetro muestral $\widehat{\beta}$, el usado por el estimador $\widehat{F}_{MA1}(t)$.

Teorema 2.8 *Los estimadores $\widehat{F}_{MA1}(t)$ y $\widehat{F}^B_{MA1}(t)$ tienen la misma distribución límite.*

Demostración

Denotemos los estimadores modelo-asistidos de verosimilitud pseudo empírica por $\widehat{F}_{MA1}(t) = T_n(\widehat{\beta})$ y $\widehat{F}^B_{MA1}(t) = T_n(\mathbf{B})$. La expresión $T_n(\widehat{\beta})$ depende del estimator $\widehat{\beta}$, es cual es función de los datos muestrales y estima consistentemente el vector de parámetros β. Reemplazando el estimador $\widehat{\beta}$ en $T_n(\cdot)$ por γ y denotándolo por $T_n(\gamma)$, es posible encontrar la distribución límite de la media de esta expresión cuando el valor actual del parámetro

es β: $\mu(\gamma) = \lim_{n\to\infty} E_\beta[T_n(\gamma)] = \widetilde{F}_y(t)$, donde $\widetilde{F}_y(t)$ es el valor límite de $F_y(t)$ cuando $N \to \infty$. Por tanto

$$\frac{\partial \mu(\gamma)}{\partial \gamma}\Big|_{\gamma=\beta} = \left(\frac{\partial \mu(\gamma)}{\partial \gamma_1}\Big|_{\gamma=\beta}, \frac{\partial \mu(\gamma)}{\partial \gamma_2}\Big|_{\gamma=\beta}, \ldots, \frac{\partial \mu(\gamma)}{\partial \gamma_P}\Big|_{\gamma=\beta} \right)$$

$$= (0, 0, \ldots, 0).$$

Randles (1982) demostró que bajo esta condición, la distribución límite de $T_n(\hat{\beta})$ $(= \widehat{F}_{MA1}(t))$ y $T_n(\mathbf{B})$ $(= \widehat{F}^B_{MA1}(t))$ son idénticas. □

Teorema 2.9 *El comportamiento asintótico del estimador* $\widehat{F}^{D1}_y(t) = \widehat{F}_{HKy}(t) + (F_b(t) - \widehat{F}_b(t))\widehat{D}$ *es el mismo del estimador* $\widehat{F}^{D2}_y(t) = \widehat{F}_{HKy}(t) + (F_b(t) - \widehat{F}_b(t))D$*, con*

$$D = \frac{\sigma_{z,w}}{\sigma^2_w} = \frac{\sum_{i\in U} d^*_i[\delta(t-y_i) - F_y(t)][\delta(t-b_i) - F_b(t)]}{\sum_{i\in U} d^*_i[\delta(t-b_i) - F_b(t)]^2}.$$

Consecuentemente, $\widehat{F}^B_{MA1}(t)$ *es asintóticamente normal y asintóticamente insesgado bajo el diseño. Su correspondiente varianza asintótica está dada por*

$$AV(\widehat{F}^B_{MA1}(t)) = \sum_{i\in U}\sum_{l\in U} \Delta_{il}(d^*_i E_i)(d^*_l E_l), \qquad (2.106)$$

donde $\Delta_{il} = \pi_{il} - \pi_i\pi_l$ *y* $E_i = \delta(t-y_i) - \delta(t-b_i)D$*.*

Demostración

$\widehat{F}^{D1}_y(t)$ puede expresarse como sigue:

$$\widehat{F}^{D1}_y(t) = \widehat{F}_{HKy}(t) + (F_b(t) - \widehat{F}_b(t))\widehat{D} =$$

$$= \widehat{F}_{HKy}(t) + (F_b(t) - \widehat{F}_b(t))(\widehat{D} - D + D) =$$

$$= \widehat{F}_{HKy}(t) + (F_b(t) - \widehat{F}_b(t))D + (F_b(t) - \widehat{F}_b(t))(\widehat{D} - D) =$$

$$= \widehat{F}^{D2}_y(t) + (F_b(t) - \widehat{F}_b(t))(\widehat{D} - D).$$

$\widehat{F}_b(t)$ y $\widehat{D}$ son asintóticamente insesgados bajo el diseño para $F_b(t)$ y D, respectivamente, y por tanto el producto $(F_b(t) - \widehat{F}_b(t))(\widehat{D} - D)$ será de menor orden que $\widehat{F}_b(t)$. Consecuentemente, el término $(F_b(t) - \widehat{F}_b(t))(\widehat{D} - D)$ tiene menor orden que $\widehat{F}_{HKy}(t) + (F_b(t) - \widehat{F}_b(t))D$. Entonces, $\widehat{F}^{D1}_y(t)$ es asintóticamente insesgado y puesto que los estimadores $\widehat{F}_{HKy}(t)$ y $\widehat{F}_b(t)$ son asintóticamente normales, el estimador $\widehat{F}^{D1}_y(t)$ es asintóticamente normal.

La varianza asintótica de $\widehat{F}^{D1}_y(t)$ coincide con la varianza del estadístico $\widehat{F}^{D2}_y(t)$, la cual está dada por

$$V\left(\widehat{F}_{HKy}(t) + (F_b(t) - \widehat{F}_b(t))D\right) =$$

$$= V\left(\widehat{F}_{HKy}(t) + F_b(t)D - \widehat{F}_b(t)D\right) =$$

$$= V\left(\widehat{F}_{HKy}(t) - \widehat{F}_b(t)D\right),$$

puesto que $F_b(t)D$ es un término constante. Ahora

$$\widehat{F}_{HKy}(t) - \widehat{F}_b(t)D = \sum_{i\in s} d^*_i\delta(t-y_i) - \sum_{i\in s} d^*_i\delta(t-b_i)D =$$

$$= \sum_{i\in s} d^*_i\left[\delta(t-y_i) - \delta(t-b_i)D\right] = \sum_{i\in s} d^*_i E_i,$$

con $E_i = \delta(t-y_i) - \delta(t-b_i)D$.

Así, la varianza asintótica de $\widehat{F}^B_{MA1}(t)$ está dada por

$$AV(\widehat{F}^B_{MA1}(t)) = V(\widehat{F}_{HKy}(t) - \widehat{F}_b(t)D) =$$

$$= V\left(\sum_{i\in s} d^*_i E_i\right) = \sum_{i\in U}\sum_{l\in U} \Delta_{il}(d^*_i E_i)(d^*_l E_l).$$

□

Considerando el Teorema 2.8, el resultado anterior también sostiene para $\widehat{F}_{MA1}(t)$ en lugar de $\widehat{F}^B_{MA1}(t)$. Por tanto, asumiendo el estimador $\widehat{F}_{MA1}(t)$, la varianza (2.106) puede estimarse por

$$\widehat{V}(\widehat{F}_{MA1}(t)) = \sum_{i\in s}\sum_{j\in s} \frac{\Delta_{ij}}{\pi_{ij}}(\widehat{p}_i e_i)(\widehat{p}_j e_j),$$

donde $e_i = \delta(t-y_i) - \delta(t-g_i)\widehat{G}$, con $\widehat{G} = \frac{\widehat{\sigma}_{z,v}}{\widehat{\sigma}^2_v} =$

$$= \frac{\sum_{i\in s} d^*_i[\delta(t-y_i) - \widehat{F}_{HKy}(t)][\delta(t-g_i) - \widehat{F}_g(t)]}{\sum_{i\in s} d^*_i[\delta(t-g_i) - \widehat{F}_g(t)]^2},$$

y donde v denota a la variable $\delta(t-g)$.

Nota 2.2 *Algunos autores, tal como Rao* et al. *(1990), usan la pseudo-variable* $g_i = \widehat{R}^t\mathbf{x}_i$*, para* $i = 1, \ldots, N$*, para construir estimadores modelo-asistidos para la función de distribución, donde* $\widehat{R} = (\sum_{i\in s} d_i x_i)^{-1}(\sum_{i\in s} d_i y_i)$*. El problema de esta pseudo-variable es que únicamente puede usarse para una variable auxiliar. Bajo tal situación,* $\widehat{R}$ *ó* $\widehat{\beta}$ *pueden usarse.*

Nota 2.3 *El estimador* $\widehat{F}_{MA}(t)$ *es computacionalmente simple y no depende de parámetros desconocidos, puesto que el vector* $\mathbf{t}_g$ *puede calcularse fácilmente a través de* $\mathbf{x}$*, el cual asumimos es conocido. Cuando esta información no está disponible, el muestreo bifásico es una técnica apropiada para poder aplicar el estimador propuesto. Este muestreo consiste en tomar una primera muestra más grande, donde se recogen los datos de la variable auxiliar. Esto servirá como información auxiliar completa en una segunda muestra más pequeña.*

2.4.5. Propiedades empíricas

Las principales propiedades del estimador $\widehat{F}_{MA}(t)$ han sido establecidas en la Sección 2.4.4. El siguiente paso es analizar la precisión de este estimador por medio de un estudio empírico. Por tanto, en esta sección se llevan a cabo estudios de simulación para investigar el cumplimiento muestral de varios estimadores de la función de distribución existentes en la literatura del muestreo en poblaciones finitas.

Para realizar estos estudios se han usado dos poblaciones simuladas generadas bajo una relación lineal entre y y $\mathbf{x}$, y una población natural, en la cual no se sostiene una relación de este tipo.

Las poblaciones simuladas, de tamaño $N = 1000$, se han generado mediante el modelo

$$y_i = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \epsilon_i, \qquad (2.107)$$

donde las variables x_{1i} y x_{2i} se han generado de distribuciones Gamma y las cantidades ϵ_i son variables aleatorias

independientes e idénticamente distribuidas con distribución Normal de parámetros 0 y σ^2. El valor de σ^2 se escoge de modo que el coeficiente de correlación entre y_i y $\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i}$ es 0.98 para la primera población (Pob098) y 0.80 para la segunda población (Pob080). Como población natural se emplea la población Murthy, la cual presenta un comportamiento exponencial en sus datos. En el Apéndice A están disponibles las propiedades más importantes de estas poblaciones así como sus respectivos diagramas de dispersión.

La precisión del estimador propuesto $\widehat{F}_{MA}(t)$ es comparada con los siguientes estimadores: el estimador convencional $\widehat{F}_{HTy}(t)$, el estimador de Chambers y Dunstan (1986) $\widehat{F}_{CD}(t)$, los estimadores propuestos en Rao *et al.* (1990), esto es $\widehat{F}_r(t)$, $\widehat{F}_d(t)$ y $\widehat{F}_{RKM}(t)$, y por último, el primer estimador $MCPE$ propuesto en Chen y Wu (2002), el cual denotamos como $\widehat{F}^{(1)}_{MC}(t)$.

Notamos que el modelo (2.107) fue también usado por Chen y Wu (2002), teniendo el estimador $\widehat{F}^{(1)}_{MC}(t)$ el mejor cumplimiento en la mayoría de los casos. En este estudio, también se usa el estimador $\widehat{F}_{MA}(t)$ cuando se considera un valor t_0 en las restricciones. Este estimador se denota como $\widehat{F}_{MA1}(t)$. Esto nos permitirá comprobar la ganancia de precisión de usar un vector en las restricciones en lugar de usar un único valor. Así, el mismo punto $t_0 = Q_g(0{,}5)$ es usado por los estimadores $\widehat{F}^{(1)}_{MC}(t)$ y $\widehat{F}_{MA1}(t)$ para cada t, puesto que esto es necesario para obtener una auténtica función de distribución.

Se llevan a cabo dos estudios de simulación. Por un lado, se evalúan los estimadores en los puntos $t = Q_y(0{,}25)$, $t = Q_y(0{,}50)$ y $t = Q_y(0{,}75)$. Con el fin de revelar el comportamiento medio de los distintos estimadores en diferentes valores de t, se realiza otro estudio de simulación para los argumentos $t = Q_y(0{,}1), Q_y(0{,}2), \ldots, Q_y(0{,}9)$. Éste último nos permitirá observar el comportamiento del estimador $\widehat{F}_{MA}(t)$ cuando se usan valores de t alejados de $\mathbf{t}_g = (t_{g25}, t_{g50}, t_{g75})^t$.

Primera simulación

Esta primera simulación consiste en tomar una muestra aleatoria simple de las anteriores poblaciones y estimar la función de distribución en los puntos $t = Q_y(0{,}25)$, $t = Q_y(0{,}50)$ y $t = Q_y(0{,}75)$ mediante los distintos estimadores. Este proceso se repite $B = 1000$ veces para diferentes tamaños muestrales. A continuación, el cumplimiento de todos los estimadores se compara en términos de Sesgo Relativo (SR) y de Eficiencia Relativa (ER), con

$$SR(t) = \frac{1}{B}\sum_{b=1}^{B} \frac{\widehat{F}(t)_b - F_y(t)}{F_y(t)};\ ER(t) = \frac{ECM[\widehat{F}(t)]}{ECM[\widehat{F}_{HTy}(t)]}, \tag{2.108}$$

donde b expresa la b-ésima simulación, $\widehat{F}(t)$ es un estimador cualquiera de la función de distribución, $ECM[\widehat{F}(t)] = B^{-1}\sum_{b=1}^{B}[\widehat{F}(t)_b - F_y(t)]^2$ es el Error Cuadrático Medio empírico para $\widehat{F}(t)$, y $ECM[\widehat{F}_{HTy}(t)]$ se define de modo similar para el estimador estándar. Notamos que valores de ER menores de 1 indican que el estimador $\widehat{F}(t)$ es mejor que $\widehat{F}_{HTy}(t)$ en términos de error cuadrático medio.

Las funciones que permiten llevar a cabo este estudio pueden consultarse en el Apéndice **??**. La función de R usada para encontrar la solución de la ecuación $h(\lambda) = \mathbf{0}$ puede también verse en Wu (2005).

Las Figuras B.7 y B.8 muestran la ER para las tres poblaciones cuando se evalúan en los cuartiles poblacionales de la variable de interés. En los casos donde un estimador cumpla peor que el estimador estándar, su correspondiente línea estará omitida. Los valores absolutos de las cantidades SR para $\widehat{F}_{MA}(t)$ están todas dentro de un rango razonable y son todos menores del 1 %. Esto sostiene para el resto de estimadores en la mayoría de los casos. De este modo, estos valores no se muestran.

De las Figuras B.7 y B.8 se pueden obtener las siguientes conclusiones:

1. $\widehat{F}_{MA}(t)$ es considerablemente más preciso que el resto de estimadores en $t = Q_y(0{,}25)$ y $t = Q_y(0{,}75)$, y exhibe la más baja ER en estos casos. Cuando se estima la mediana de la variable de interés, la situación es diferente, es decir, otros estimadores presentan un similar comportamiento a $\widehat{F}_{MA}(t)$. Por ejemplo, uno de estos estimadores es $\widehat{F}^{(1)}_{MC}(t)$ en las poblaciones Pop098 y Pop080. Este estimador muestra una mayor ER en los puntos $t = Q_y(0{,}25)$ y $t = Q_y(0{,}75)$ debido a que t_0 está alejado de t. El conocimiento del modelo correcto maximiza la eficiencia de $\widehat{F}^{(1)}_{MC}(t)$, pero solamente cuando t está próximo a t_0.
2. En los casos donde hay una fuerte información auxiliar (Pop098), la ganancia de usar $\widehat{F}_{CD}(t)$, $\widehat{F}^{(1)}_{MC}(t)$, $\widehat{F}_{MA}(t)$ y $\widehat{F}_{MA1}(t)$ puede ser substancial comparada con el estimador estándar.
3. La débil linealidad en la población Murthy afecta especialmente a $\widehat{F}^{(1)}_{MC}(t)$ y $\widehat{F}_{CD}(t)$, los cuales son más eficientes cuando los datos se rigen por un modelo lineal (Pop098 y Pop080).
4. $\widehat{F}_{CD}(t)$ es menos eficiente que el estimador estándar de tipo Horvitz-Thompson cuando la función de distribución se estima en los puntos $t = Q_y(0{,}25)$ y $t = Q_y(0{,}75)$. Este estimador es bastante preciso cuando t está próximo a $Q_y(0{,}5)$, aunque llega a ser considerablemente menos eficiente cuando t está alejado de $Q_y(0{,}5)$.
5. $\widehat{F}_{MA1}(t)$ es siempre menos preciso que $\widehat{F}_{MA}(t)$. Esto revela la ganancia de usar el vector $\mathbf{t}_g$ en lugar de un valor t_0. En cualquier caso, $\widehat{F}_{MA1}(t)$ tiene un buen comportamiento y es siempre más eficiente que el estimador estándar.
6. En términos de ER, el estimador más eficiente para $F_y(t)$ se obtiene por $\widehat{F}_{MA}(Q_y(0{,}75))$ en la población Murthy. En este caso, los estimadores modelo-calibrados y basados en modelos no tienen un buen comportamiento. Esto puede deberse a que no existe una buena linealidad y a que t está alejado de t_0.
7. Los estimadores $\widehat{F}_r(t)$ y $\widehat{F}_d(t)$ son siempre considerablemente menos eficientes que $\widehat{F}_{MA}(t)$.

Segunda simulación

La simulación anterior se ha realizado en los puntos $t = Q_y(0{,}25)$, $t = Q_y(0{,}50)$ y $t = Q_y(0{,}75)$. Puede observarse que el orden de estos cuantiles coincide con el orden de los cuantiles del vector $\mathbf{t}_g$. Es esperable que $\widehat{F}_{MA}(t)$ cumpla bien en esta situación. Por este motivo, usaremos otro estudio de simulación para medir la precisión de los distintos estimadores en los puntos $t = Q_y(0{,}1), Q_y(0{,}2), \ldots, Q_y(0{,}9)$.

En este caso, el cumplimiento de los estimadores es medido mediante el Sesgo Relativo Medio (SRM) y la Eficiencia Relativa Media (ERM), dados respectivamente por

$$SRM = \frac{1}{9}\sum_{q=1}^{9} |SR(t_q)| \quad ; \quad ERM = \sqrt{\frac{1}{9}\sum_{q=1}^{9} ER(t_q)},$$

donde $SR(t)$ y $ER(t)$ están definidos en (2.108) y t_q es el q-ésimo decil para la variable de estudio.

Consideramos también una medida global del cumplimiento de los estimadores a través de los 9 cuantiles para cada muestra obtenida de las $B = 1000$ simulaciones. Esta medida es la Desviación Absoluta Máxima (DAM) que está dada por: DAM(b) $= \max_q |\widehat{F}(t_q)_b - F(t_q)|$, para $b = 1, \ldots, B$. Notamos que las medidas SRM, ERM y DAM han sido también usadas en Silva y Skinner (1995).

La Figura B.9 muestra los valores SRM, en tanto por ciento, para las tres poblaciones. Puede observarse que todos los estimadores exhiben valores SRM menores del 1 % para las poblaciones Pob098 y Murthy. Asumiendo una relación más débil (Pob080), el estimador de tipo razón presenta el peor comportamiento (su SRM ronda el 1.4 %). En la mayoría de los casos, puede observarse que los valores SRM son decrecientes según el tamaño muestral y que el estimador $\widehat{F}_{MA}(t)$ presenta el menor sesgo.

Los valores ERM para las tres poblaciones están mostrados en la Figura B.10. Estos resultados revelan que hay una razonable ganancia de eficiencia al usar $\widehat{F}_{MA}(t)$ con respecto a otros estimadores. $\widehat{F}^{(1)}_{MC}(t)$ muestra el segundo mejor comportamiento en las poblaciones Pob098 y Pob080, las cuales están basadas en un modelo lineal. A pesar de esta relación lineal entre y y $\mathbf{x}$, la pérdida de eficiencia de $\widehat{F}^{(1)}_{MC}(t)$ comparada con $\widehat{F}_{MA}(t)$ se debe al hecho de que el estimador $\widehat{F}^{(1)}_{MC}(t)$ usa un único valor fijo $t_0 = 0{,}5$, y éste es menos preciso cuando t está alejado de t_0. En términos de ERM, $\widehat{F}_{CD}(t)$ muestra el peor comportamiento de todos los estimadores considerados. $\widehat{F}_{CD}(t)$ es más preciso cuando t está cercano a $Q_y(0{,}5)$, aunque este estimador sufre una considerable pérdida de eficiencia en cuantiles extremos (de bajo o alto orden).

La Figura B.11 muestra los diagramas de cajas con bigotes de las distribuciones de los valores DAM obtenidos para las tres poblaciones. Se han tomado muestras de tamaño 100 para las poblaciones Pob098 y Pob080 y muestras de tamaño 50 para la población Murthy. Estos diagramas confirman el análisis anterior: $\widehat{F}_{CD}(t)$ presenta la máxima desviación absoluta mientras que $\widehat{F}_{MA}(t)$ muestra el mejor comportamiento en todos los casos.

En todos los estudios (ER, SR, SRM, ERM y DAM), el estimador propuesto, $\widehat{F}_{MA}(t)$, proporciona una buena mejoría sobre $\widehat{F}_{MA1}(t)$, el cual usa un único punto t_0. Esto confirma la ganancia en eficiencia al usar el vector $\mathbf{t}_g$, especialmente cuando t está alejado de t_0.

3. Aportaciones a la estimación de cuantiles

3.1. Introducción

El problema de la estimación de la totales y medias poblacionales en presencia de variables auxiliares ha sido extensamente discutido en la literatura del muestreo de poblaciones finitas. Para el problema de la estimación de la mediana y otros cuantiles, la situación es bastante diferente y tan sólo en la actualidad este problema está siendo discutido, debido en parte, al creciente interés de este tipo de medidas. Notamos que los distintos estimadores y métodos propuestos para la media y el total de una variable no tienen una extensión obvia al problema de la estimación de cuantiles.

Un ejemplo del uso de cuantiles y otras medidas relacionadas en muestreo de poblaciones finitas es el siguiente. Frecuentemente, los organismos nacionales de estadística y otras agencias se encuentran con variables, tales como ingresos, gastos, etc., que presentan distribuciones con una alta asimetría. Bajo estas circunstancias, la mediana resulta más apropiada que la media poblacional. De este modo, asumiendo datos de Encuestas Continuas de Presupuestos Familiares, los gobiernos de diferentes países obtienen numerosas medidas de pobreza, tal como la proporción de bajos ingresos, que dependen directamente de determinados cuantiles. Un ejemplo de este tipo de medidas viene dado por Eurostat (2000), en donde se define que un salario es bajo si éste está por debajo del 60 % del salario mediano mensual, es decir, el cuantil de orden $\beta = 0{,}5$ se emplea en Eurostat. A nivel nacional, el Instituto Nacional de Estadística y sus correspondientes organismos autónomos, definen una medida similar para determinar el índice de pobreza, aunque en este caso la variable principal es el gasto producido en los hogares españoles. Otros estudios de tipo económico también usan cuantiles para estudiar la relación entre gastos en alimentación de los hogares y los correspondientes ingresos, análisis de salarios y gastos, impacto de varias características demográficas, calidad en la escuela, análisis de demanda, etc. Una extensa bibliografía sobre estas y otras aplicaciones en estudios de tipo económico puede consultarse en Koenker y Hallock (2001).

Al igual que para el caso de la estimación de parámetros lineales como medias o totales, las estimaciones serán más eficientes si se incorpora información auxiliar, altamente correlacionada con la variable de interés, en la etapa de estimación. En la estimación de cuantiles, existen dos grandes métodos que incorporan la información auxiliar de forma eficiente:

M1. Estimación de cuantiles indirectos: consiste en construir estimadores de tipo razón, diferencia o regresión, tal como se construyen para la media o el total. Ejemplos de este tipo de estimación pueden verse en Kuk y Mak (1989), Arcos, Rueda y Muñoz (2006), Rueda, *et al.* (1998, 2003, 2004), etc. Notamos que para formular la mayoría de estos estimadores, se requiere conocer los cuantiles poblacionales de las variables auxiliares, o bien otro tipo de parámetro poblacional.

M2. Estimación a través de la función de distribución: La técnica habitual en muestreo de poblaciones finitas es invertir la función de distribución para obtener la estimación de un determinado cuantil. Se requiere, por tanto, usar eficientemente la información auxiliar en la etapa de estimación de la función de distribución. El inconveniente de esta técnica es que el estimador de la función de distribución debe ser una verdadera función de distribución para estimar cuantiles con mayor precisión. Aunque este hecho resulta imprescindible, existen varios estimadores en la literatura que no cumplen tal propiedad. Chambers y Dunstan (1986) fueron de los primeros investigadores en utilizar información auxiliar para construir estimadores de la función de distribución, y posteriormente invertir esta función para obtener cuantiles. Otras importantes referencias son Rao *et al.* (1990), Wang y Dorfman (1996), Dorfman y Hall (1993), Kuo (1988), Silva y Skinner (1995).

Notamos que durante el desarrollo de este capítulo se tratarán exclusivamente con estimadores derivados del método $M2$, el cual es más usado por su calidad de estimación y eficiencia.

Los primeros trabajos relacionados con el problema de la estimación de parámetros de posición, como la mediana y los cuantiles se deben a Woodruff (1952) donde se construyen intervalos de confianza bajo muestreo aleatorio simple. Posteriormente, Hill (1968) utiliza un enfoque bayesiano para la construcción de sus estimadores, mientras que Sendransk y Meyer (1978) se basan en un enfoque puramente probabilístico de distribución de estadísticos ordenados para muestreo aleatorio simple y estratificado. Pero los estimadores más eficientes y con mejores propiedades se desarrollan posteriormente bajo aproximaciones modelo-asistidas, basadas en el modelo y modelo-calibradas. También se han propuesto estimadores de cuantiles mediante intervalos de confianza basados en estimadores de razón, regresión y diferencia y usando información auxiliar multivariante (Rueda, Arcos y Artés, 1997, 1998, Rueda y Arcos, 2001, Rueda y Arcos, 2002a, Rueda y Arcos, 2002b).

En la literatura, los estimadores de cuantiles más

conocidos son los siguientes. En primer lugar, citamos el estimador de Chambers y Dunstan (1986) para la función de distribución, el cual está basado en un modelo de superpoblación. La inversión directa de esta función puede usarse para la obtención de cuantiles. Siguiendo esta técnica, Rao *et al.* (1990) propusieron estimadores de tipo razón y diferencia usando una aproximación basada en el diseño. Kuk y Mak (1989) propusieron dos estimadores para los cuales solamente es necesario conocer a nivel poblacional el valor de la mediana de una variable auxiliar. Más recientemente, Rueda *et al.* (1998) y Rueda y Arcos (2001) propusieron intervalos de confianza para los cuantiles basados en estimadores de tipo razón y diferencia de la función de distribución. En Rueda *et al.* (2003, 2004) se plantea la estimación de cuantiles mediante estimadores de tipo diferencia usando cuantiles poblacionales del mismo orden de la variable auxiliar. La estimación de cuantiles usando técnicas recientes de estimación también ha sido investigada. Por ejemplo, Chen y Wu (2002) proponen la estimación de cuantiles usando la aproximación modelo-calibrada.

Existe otro gran número de estimadores de cuantiles propuestos para distintos diseños muestrales. Los estimadores más importantes se irán citando a lo largo del presente capítulo, en el cual se trata el problema de la estimación de cuantiles desde distintos enfoques. Por un lado, se desarrollan nuevos estimadores en diseños muestrales más complejos, y por otro, se proponen estimadores asumiendo el reciente método de verosimilitud empírica.

Para formular la mayoría de los estimadores de cuantiles, ya sean a través del método $M1$ o del método $M2$, es necesario conocer los valores poblacionales de las variables auxiliares, aunque esto es poco usual en la práctica. La solución a este problema se trata en la Sección 3.2 mediante el uso del muestreo bifásico, en el cual la información auxiliar poblacional puede estimarse usando la muestra de la primera fase. Por tanto, en esta sección se proponen estimadores de cuantiles en muestreo bifásico y asumiendo que las unidades muestrales se extraen mediante métodos de muestreo con probabilidades desiguales en cada una de las dos fases. La eficiencia de estos estimadores puede mejorarse si se usa un muestreo estratificado en la primera fase. Asumiendo este último diseño muestral, denominado muestreo bifásico aplicado a la estratificación, se comprueba que los estimadores propuestos pueden llegar a ser más precisos con respecto a otros existentes en la literatura.

Por otro lado, en la Sección 3.3 se plantean nuevos estimadores de cuantiles bajo muestreo en ocasiones sucesivas. En primer lugar se definen estimadores de cuantiles basados en múltiples variables auxiliares. La introducción de tal información proporciona un marco de estimación apropiado que permite obtener estimadores más precisos. A continuación, también se proponen estimadores de cuantiles basados en muestras seleccionadas mediante muestreos probabilísticos con probabilidades desiguales (por ejemplo, con unidades proporcionales al tamaño). Notamos que éste es el caso de los organismos nacionales y agencias de estadística que realizan encuestas continuas a lo largo del tiempo. El comportamiento de todos los estimadores propuestos se analiza desde el punto de vista teórico (mediante aproximaciones asintóticas), y desde una perspectiva empírica (analizando los resultados obtenidos a partir de una serie de poblaciones).

Para cerrar este capítulo, en la Sección 3.4 se proponen estimadores para cuantiles asumiendo el método de verosimilitud empírica, expuesto con detalle en el capítulo anterior. Los estimadores propuestos usan de manera eficiente la información auxiliar, lo que se traduce en una mejoría de la precisión. Esta precisión de los estimadores propuestos se ha evaluado para el cálculo de algunas medidas de pobreza oficiales, las cuales dependen de forma directa de cuantiles. Este estudio se ha llevado a cabo asumiendo distintos estimadores de cuantiles. Los resultados obtenidos reflejan que los estimadores propuestos proporcionan estimaciones más precisas para las medidas de pobreza involucradas en tal estudio.

3.2. Estimadores bajo muestreo bifásico

En esta sección se resuelve el problema de la estimación de cuantiles bajo muestreo en dos fases o muestreo bifásico con diseños muestrales arbitrarios en cada una de las dos fases. Se proponen varios estimadores de tipo directo, razón y exponencial que proporcionan estimaciones óptimas para un determinado cuantil. Se analizan propiedades importantes de estos estimadores, tales como la insesgadez, estimación de varianzas, etc. Como caso particular, se investiga también el muestreo bifásico aplicado a la estratificación, diseño muestral que ofrece importantes ganancias en eficiencia debido a los beneficios que produce el muestreo estratificado. Todas estas propiedades se ven desde un punto de vista teórico, aunque el análisis de los estimadores se completa con un estudio empírico llevado a cabo para los cuartiles y bajos distintos diseños muestrales con probabilidades desiguales. Este estudio refleja que los estimadores propuestos mejoran a otros estimadores conocidos en términos de sesgo y eficiencia relativa.

Notamos que la mayor ventaja al usar muestreo bifásico es una alta ganancia en precisión sin un sustancial incremento en costes. De hecho, este diseño muestral se usa frecuentemente en numerosas encuestas por razones de coste y eficiencia.

3.2.1. Introducción

Para el problema de la estimación de un determinado parámetro en muestreo de poblaciones finitas, la información auxiliar juega un papel muy importante en la precisión de los estimadores. La mayoría de los estimadores basados en información auxiliar se basan en el conocimiento a nivel poblacional de las variables auxiliares. En la práctica, esta cantidad no tiene porque ser conocida. De hecho, son muy poco frecuentes las encuestas que disponen de esta información, por lo que resulta imposible obtener estos estimadores basados en información auxiliar. Una alternativa es estimar los parámetros poblacionales que usan los estimadores, aunque esto conlleva a importantes errores en la etapa de la estimación de la varianza (véase Berger, Muñoz y Rancourt, 2006). Bajo esta situación, el uso

de un muestreo bifásico es la técnica más apropiada para resolver este problema.

Por tanto, el muestreo bifásico es una herramienta útil para aquellas investigaciones en las cuales no existe conocimiento previo de las variables auxiliares a nivel poblacional. Otro punto a favor del muestreo bifásico es la creación de un esquema importante de información que permite la selección probabilística de sub-muestras. Para una mayor profundización sobre el muestreo bifásico en la estimación de medias o totales puede consultarse, por ejemplo, Prasad y Thach (2001), Särndal *et al.* (1992), Fernández y Mayor (1994) y Artés y García (2002).

En lo que respecta a la estimación de cuantiles en muestreo bifásico, los primeros autores en realizar investigaciones en este sentido fueron Singh *et al.* (2001), Singh (2003) y Allen *et al.* (2002) para el problema de la estimación de la mediana poblacional. Estos trabajos fueron desarrollados exclusivamente para muestreo aleatorio simple. Con el fin de completar estos estudios, en esta sección se proponen numerosos estimadores para un determinado cuantil cuando se lleva a cabo un muestreo bifásico con diseños muestrales arbitrarios en cada una de las dos fases.

A continuación se describe brevemente en que consiste un muestreo bifásico. Suponemos que tenemos una población U compuesta por N unidades de la que se extrae en una primera fase una muestra, s', de tamaño, n', bastante grande y de bajo costo, según cierto criterio muestral, d_1, tal que $p_{d1}(s')$ será la probabilidad de que s' sea seleccionada y donde las correspondientes probabilidades de inclusión de primer y segundo orden se denotan, respectivamente, como $\pi_i^{'}$ y $\pi_{ij}^{'}$ para i y $j \in U$. En esta muestra, una o varias variables auxiliares pueden ser recogidas fácilmente, es decir, dicha muestra permite obtener la información auxiliar necesaria para todo el proceso. Dada s', una segunda muestra s de tamaño n es seleccionada en la segunda fase mediante un diseño d_2, tal que $p(s/s')$ es la probabilidad condicional de escoger s. Las probabilidades de inclusión bajo este diseño se denotan como $\pi_{i/s'}$ y $\pi_{ij/s'}$. Notamos que $\Delta_{ij}^{'} = \pi_{ij}^{'} - \pi_i^{'}\pi_j^{'}$ y $\Delta_{ij}^{s'} = \pi_{ij/s'} - \pi_{i/s'}\pi_{j/s'}$.

3.2.2. Estimadores propuestos

Sin usar ningún tipo de información auxiliar, el candidato natural para estimar el cuantil β es $\widehat{Q}_y(\beta) = \text{ínf}\{t \mid \widehat{F}_{HTy}(t) \geq \beta\} = \widehat{F}_{HTy}^{-1}(\beta)$, donde

$$\widehat{F}_{HTy}(t) = \frac{1}{N}\sum_{i\in s}\frac{\delta(t-y_i)}{\pi_i}$$

es el estimador de tipo Horvitz y Thompson (1952) de $F_y(t)$, y las probabilidades de inclusión están dadas por $\pi_i = \sum_{s'\ni i} p_{d1}(s')\pi_{i/s'}$.

Como puede observarse, para determinar π_i se deben conocer las probabilidades $\pi_{i/s'}$ para cada s', las cuales no se conocen generalmente porque $\pi_{i/s'}$ pueden depender del diseño de la primera fase, por ejemplo si la muestra de la segunda fase es diseñada mediante un muestreo proporcional a una variable auxiliar.

Notamos que el estimador de tipo Horvitz-Thompson para la media poblacional tampoco puede obtenerse en la práctica bajo este muestreo. Por esta razón, Särndal *et al.* (1992) propusieron el uso de estimadores π^*. Usando esta idea, se definen las cantidades $\pi_i^* = \pi_i^{'}\pi_{i/s'}$ y $\pi_{ij}^* = \pi_{ij}^{'}\pi_{ij/s'}$, que permiten definir el π^*-estimador de la función de distribución como

$$\widehat{F}_{HTy}^*(t) = \frac{1}{N}\sum_{i\in s}\frac{\delta(t-y_i)}{\pi_i^*}, \tag{3.1}$$

y así, el estimador directo propuesto para un cuantil β esta dado por

$$\widehat{Q}_y^*(\beta) = \widehat{F}_{HTy}^{*-1}(\beta). \tag{3.2}$$

Notamos que $\widehat{Q}_y^*(\beta)$ no coincide generalmente con el estimador $\widehat{Q}_y(\beta)$ excepto en casos excepcionales, aunque la principal ventaja del estimador directo propuesto sobre el estándar comentado es su aplicabilidad para cualesquiera que sean los diseños muestrales usados en cada fase.

El estimador (3.2) se ha definido sin usar ninguna información auxiliar. Si esta información está disponible, el uso de estimadores indirectos nos puede ayudar a obtener estimaciones más precisas para los cuantiles en muestreo bifásico. De este modo, el siguiente paso es definir una clase de estimadores que usen información auxiliar. En primer lugar mostraremos los principales antecedentes relacionados con el tema que nos ocupa.

Asumiendo muestreo aleatorio simple y que la mediana de la variable x es conocida, Kuk y Mak (1989) propusieron el siguiente estimador de tipo razón para la mediana

$$\widehat{Q}_y^r(0{,}5) = \widehat{Q}_y(0{,}5)\frac{Q_x(0{,}5)}{\widehat{Q}_x(0{,}5)}.$$

Además, Kuk y Mak (1989) propusieron otros estimadores de cuantiles bajo muestreo aleatorio simple llamados estimadores de posición y de estratificación, pero la extensión de cualquiera de ellos a otros diseños muestrales más complejos no ha sido posible.

Rueda *et al.* (2003, 2004) propusieron, para cualquier diseño muestral d y para cualquier β, métodos de diferencia y exponenciales para estimar un cuantil β. Singh *et al.* (2001) sugirieron estimadores de tipo razón, regresión, posición y estratificación de la mediana cuando la muestra es seleccionada en dos fases y usando muestreo aleatorio simple en cada una de ellas. Bajo muestreo bifásico y muestreo aleatorio simple en cada fase, Allen *et al.* (2002) propusieron dos clases de estimadores para la mediana poblacional. Estos estimadores usan la información proporcionada por dos variables auxiliares, x y z, donde se asume que la mediana de z es conocida.

A continuación se presenta una clase de estimadores para cuantiles poblacionales cuando las muestras en ambas fases son seleccionadas mediante un esquema de muestreo arbitrario:

$$\widehat{Q}_y^{\mathcal{H}}(\beta) = H(\widehat{Q}_y^*(\beta), t^*), \tag{3.3}$$

donde $t^* = \widehat{Q}_x^*(\beta)/\widehat{Q}_x^{'}(\beta)$, y $\widehat{Q}_x^{'}(\beta)$ es el estimador de $Q_x(\beta)$ basado en la muestra de la primera fase, esto es, $\widehat{Q}_x^{'}(\beta) = \text{ínf}\{t \mid \widehat{F}_{HTx}^{'-1}(t) \geq \beta\}$, donde

$$\widehat{F}_{HTx}^{'}(t) = \frac{1}{N}\sum_{i\in s'}\frac{\delta(t-x_i)}{\pi_i^{'}}.$$

La función H satisface las siguientes condiciones:

(C3.1). Asume valores en un subconjunto convexo cerrado $\mathcal{C} \subset \mathbb{R}_2$, el cual contiene el punto $(Q_y(\beta), 1)$.

(C3.2). H es una función continua en $\mathcal{C}$, tal que $H(Q_y(\beta), 1) = Q_y(\beta)$.

(C3.3). Las primeras y segundas derivadas parciales de H existen y son continuas en $\mathcal{C}$, con

$$H_{10}(Q_y(\beta), 1) = \left.\frac{\partial H(q, t^*)}{\partial q}\right|_{(q,t^*)=(Q_y(\beta),1)} = 1.$$

Un caso particular dentro de la clase general de estimadores $\mathcal{H}$ es el estimador tipo razón, dado por:

$$\widehat{Q}^*_{yr}(\beta) = \widehat{Q}^*_y(\beta)\frac{\widehat{Q}'_x(\beta)}{\widehat{Q}^*_x(\beta)},$$

y el cual se corresponde con la elección $H(q, t^*) = q/t^*$.

Otro estimador para el cuantil β, llamado el estimador exponencial, está dado por:

$$\widehat{Q}^*_{ye}(\beta) = \widehat{Q}^*_y(\beta)\left(\frac{\widehat{Q}'_x(\beta)}{\widehat{Q}^*_x(\beta)}\right)^{\alpha},$$

siendo α una constante fija. Este estimador también se encuentra dentro de la clase $\mathcal{H}$, puesto que se corresponde con la elección $H(q, t^*) = q/(t^*)^{\alpha}$. Notamos que estos estimadores se han definido en Rueda, Arcos, Muñoz y Singh (2006).

Nota 3.1 *Si $\alpha = 0$, entonces $\widehat{Q}^*_{ye}(\beta) = \widehat{Q}^*_y(\beta)$, esto es, $\widehat{Q}^*_{ye}(\beta)$ coincide con el estimador π^*. Por otro lado, si $\alpha = 1$, entonces $\widehat{Q}^*_{ye}(\beta) = \widehat{Q}^*_{yr}(\beta)$. Por último, puede comprobarse que si $\alpha = -1$, entonces $\widehat{Q}^*_{ye}(\beta) = \widehat{Q}^*_{yp}(\beta)$, el cual puede definirse como un estimador producto.*

Nota 3.2 *Bajo muestreo aleatorio simple en cada fase y $\beta = 0{,}5$, los estimadores propuestos $\widehat{Q}^*_{yr}(\beta)$ y $\widehat{Q}^*_{ye}(\beta)$ se corresponden, respectivamente, con los estimadores $\widehat{M}^{(a)}_y$ y $\widehat{M}^{(b)}_y$ propuestos por Singh* et al. *(2001).*

3.2.3. Propiedades teóricas

En este apartado se estudian las principales propiedades del estimador $\widehat{Q}^*_y(\beta)$ y de los estimadores basados en la clase $\mathcal{H}$. Debido a que estos estimadores no son funciones continuas, serán necesarias aproximaciones lineales.

Teorema 3.1 *El estimador $\widehat{Q}^*_y(\beta)$ es asintóticamente insesgado para $Q_y(\beta)$*

Demostración

En primer lugar, el estimador $\widehat{Q}^*_y(\beta)$ puede expresarse asintóticamente como una función lineal de la función de distribución estimada y evaluada en el punto $Q_y(\beta)$ mediante la representación de Bahadur (véase, por ejemplo, Bahadur, 1966, Chambers y Dunstan, 1986, Kuk y Mak, 1989, Chen y Chen, 2000, Chen y Wu, 2002, etc):

$$\widehat{Q}^*_y(\beta) - Q_y(\beta) = \frac{1}{f_y(Q_y(\beta))}(\beta - \widehat{F}^*_{HTy}(Q_y(\beta))) + O(n^{-1/2}), \tag{3.4}$$

donde $f_y(\cdot)$ denota la derivada del valor límite de $F_y(\cdot)$ cuando $N \longrightarrow \infty$.

Además, es sabido que el estimador $\widehat{F}^*_{HTy}(t)$ es insesgado de $F(t)$. En consecuencia, se tiene que

$$E(\beta - \widehat{F}^*_{HTy}(Q_y(\beta))) = 0$$

y basándose en la ecuación (3.4), puede verse fácilmente que $E(\widehat{Q}^*_y(\beta)) = Q_y(\beta) + O(n^{-1/2})$, esto es, el estimador $\widehat{Q}^*_y(\beta)$ es asintóticamente insesgado de $Q_y(\beta)$. □

Teorema 3.2 *La varianza asintótica del estimador $\widehat{Q}^*_y(\beta)$ está dada por $AV(\widehat{Q}^*_y(\beta)) =$*

$$= \frac{1}{N^2}\frac{1}{f^2_y(Q_y(\beta))}\left[\sum_{i,j\in U}(\pi'_{ij} - \pi'_i\pi'_j)\frac{\delta(Q_y(\beta) - y_i)}{\pi'_i}\frac{\delta(Q_y(\beta)}{\pi'_j}\right.$$

$$\left. + E_{d1}\left(\sum_{i,j\in s'}(\pi_{ij/s'} - \pi_{i/s'}\pi_{j/s'})\frac{\delta(Q_y(\beta) - y_i)}{\pi^*_i}\frac{\delta(Q_y(\beta) - y}{\pi^*_j}\right.\right.$$

Demostración

De la expresión (3.4) se deduce que

$$AV(\widehat{Q}^*_y(\beta)) = \frac{1}{f^2_y(Q_y(\beta))}V\left(\widehat{F}^*_{HTy}(Q_y(\beta))\right),$$

donde $V\left(\widehat{F}^*_{HTy}(Q_y(\beta))\right) =$

$$= V_{d1}E\left(\widehat{F}^*_{HTy}(Q_y(\beta))|s'\right) + E_{d1}V\left(\widehat{F}^*_{HTy}(Q_y(\beta))|s'\right)$$

refleja la variación debida a cada una de las fases de muestreo.

Por otro lado, el error de estimación total del estimador π^* dado por (3.1), cuando se evalúa en el punto $Q_y(\beta)$, puede expresarse como suma de dos componentes

$$\widehat{F}^*_{HTy}(Q_y(\beta)) - F_y(Q_y(\beta)) =$$

$$= \left(\widehat{F}'_{HTy}(Q_y(\beta)) - F_y(Q_y(\beta))\right) +$$

$$+ \left(\widehat{F}^*_{HTy}(Q_y(\beta)) - \widehat{F}'_{HTy}(Q_y(\beta))\right) = Q_{s'} + R_s,$$

donde $Q_{s'}$ es el error debido a la primera fase del muestreo y R_s es el error debido a la segunda fase. Usando esta descomposición, se tiene que

$$V_{d1}E\left(\widehat{F}^*_{HTy}(Q_y(\beta))|s'\right) = V_{d1}(Q_{s'}) =$$

$$= \frac{1}{N^2}\sum_{i,j\in U}(\pi'_{ij} - \pi'_i\pi'_j)\frac{\delta(Q_y(\beta) - y_i)}{\pi'_i}\frac{\delta(Q_y(\beta) - y_j)}{\pi'_j}$$

y

$$E_{d1}V\left(\widehat{F}^*_{HTy}(Q_y(\beta))|s'\right) = E_{d1}V(R_s|s') = \frac{1}{N^2}\times$$

$$\times E_{d1}\left(\sum_{i,j\in s'}(\pi_{ij/s'} - \pi_{i/s'}\pi_{j/s'})\frac{\delta(Q_y(\beta) - y_i)}{\pi^*_i}\frac{\delta(Q_y(\beta) - y}{\pi^*_j}\right.$$

□

Corolario 3.1 *Un estimador insesgado de* $AV(\widehat{Q}_y^*(\beta))$ *está dado por*

$$\widehat{V}(\widehat{Q}_y^*(\beta)) = \frac{1}{N^2}\frac{1}{f_y^2(Q_y(\beta))}\times$$

$$\times\left(\sum_{i,j\in s}\frac{\pi_{ij}^{'}-\pi_i^{'}\pi_j^{'}}{\pi_{ij}^*}\frac{\delta(\widehat{Q}_y^*(\beta)-y_i)}{\pi_i^{'}}\frac{\delta(\widehat{Q}_y^*(\beta)-y_j)}{\pi_j^{'}}+\right.$$

$$\left.+\sum_{i,j\in s}\frac{\pi_{ij/s'}-\pi_{i/s'}\pi_{j/s'}}{\pi_{ij/s'}}\frac{\delta(\widehat{Q}_y^*(\beta)-y_i)}{\pi_i^*}\frac{\delta(\widehat{Q}_y^*(\beta)-y_j)}{\pi_j^*}\right).$$

En la práctica, la cantidad $f_y(Q_y(\beta))$ es desconocida. Un valor aproximado de $f_y(Q_y(\beta))$ puede obtenerse aplicando métodos estándares tal como el kernel (Silverman, 1986). Notamos que algunos de estos métodos para la estimación de densidades han sido usados, por ejemplo, en Kuk y Mak (1989) y Arcos *et al.* (2005).

El estimador de la varianza anterior no depende de esperanzas relacionadas con el diseño de la primera fase, haciendo posible su cálculo en la práctica.

Teorema 3.3 *Cualquier estimador dentro de la clase* $\mathcal{H}$ *es asintóticamente insesgado para* $Q_y(\beta)$.

Demostración

Para obtener este resultado nos basaremos en las siguientes aproximaciones lineales:

$$\widehat{Q}_y^*(\beta)-Q_y(\beta) = \frac{1}{f_y(Q_y(\beta))}(\beta-\widehat{F}_{HTy}^*(Q_y(\beta)))+O(n^{-1/2}),$$

$$\widehat{Q}_x^*(\beta)-Q_x(\beta) = \frac{1}{f_x(Q_x(\beta))}(\beta-\widehat{F}_{HTx}^*(Q_x(\beta)))+O(n^{-1/2}),$$

$$\widehat{Q}_x^{'}(\beta)-Q_x(\beta) = \frac{1}{f_x(Q_x(\beta))}(\beta-\widehat{F}_{HTx}^{'}(Q_x(\beta)))+O(n'^{-1/2}),$$

y usando la expansión de la serie de Taylor de primer orden para H sobre el punto $(Q_y(\beta),1)$:

$$\begin{aligned}\widehat{Q}_y^{\mathcal{H}}(\beta) &= H((Q_y(\beta),1))+\\ &+\left(\widehat{Q}_y^*(\beta)-Q_y(\beta)\right)H_{10}(Q_y(\beta),1)+\\ &+(t-1)H_{01}(Q_y(\beta),1)+O(n^{-1}),\end{aligned}\qquad(3.5)$$

donde H_{10} y H_{01} denotan las derivadas parciales de primer orden de H con respecto a q y t, respectivamente. Como $\widehat{F}_{HTy}^*(t)$ y $\widehat{F}_{HTx}^*(t)$ son estimadores insesgados de $F_y(t)$ y $F_x(t)$, respectivamente, puede observarse que cualquier estimador en $\mathcal{H}$ será asintóticamente insesgado para $Q_y(\beta)$. □

Para obtener las expresiones asintóticas de las varianzas, consideraremos la expansión de la serie de Taylor dada en (3.5), que da lugar a la expresión:

$$\widehat{Q}_y^{\mathcal{H}}(\beta)-Q_y(\beta) =$$

$$=\left(\widehat{Q}_y^*(\beta)-Q_y(\beta)\right)+\frac{\widehat{Q}_x^*(\beta)}{\widehat{Q}_x^{'}(\beta)}H_{01}(Q_y(\beta),1)+O(n^{-1}).$$

Desarrollando se obtiene

$$\widehat{Q}_y^{\mathcal{H}}(\beta)-Q_y(\beta)\simeq Q_y(\beta)e_0+\qquad(3.6)$$

$$+(e_1-e_2)H_{01}(Q_y(\beta),1)-e_2(e_1-e_2)H_{01}(Q_y(\beta),1),$$

donde:

$$e_0=\frac{\widehat{Q}_y^*(\beta)}{Q_y(\beta)}-1,\quad e_1=\frac{\widehat{Q}_x^*(\beta)}{Q_x(\beta)}-1\quad\text{y}\quad e_2=\frac{\widehat{Q}_x^{'}(\beta)}{Q_x(\beta)}-1.$$

Introduciendo varianzas en (3.6) y bajo una aproximación de primer orden, se llega a la expresión:

$$\begin{aligned}V(\widehat{Q}_y^{\mathcal{H}}(\beta)) &= Q_y(\beta)^2V(e_0)+H_{01}(Q_y(\beta),1)^2V(e_1-e_2)+\\ &+2H_{01}(Q_y(\beta),1)Cov(e_0,e_1-e_2).\end{aligned}$$

Por otro lado, bajo muestreo bifásico:

$$V(\widehat{Q}_y^{\mathcal{H}}(\beta))=E_{d1}V(\widehat{Q}_y^{\mathcal{H}}(\beta)/s^{'})+V_{d1}E(\widehat{Q}_y^{\mathcal{H}}(\beta)/s^{'})$$

refleja la variación debida a cada una de las dos fases de muestreo. Usando las propiedades conocidas del estimador de Horvitz-Thompson y su varianza, se obtiene

$$V_{d1}E(\widehat{Q}_y^{\mathcal{H}}(\beta)/s^{'})=\frac{1}{N^2}\frac{1}{f_y^2(Q_y(\beta))}\times$$

$$\times\left(\sum_{i,j\in U}\Delta_{ij}^{'}\frac{\delta(Q_y(\beta)-y_i)}{\pi_i^{'}}\frac{\delta(Q_y(\beta)-y_j)}{\pi_j^{'}}\right)$$

y

$$E_{d1}V(\widehat{Q}_y^{\mathcal{H}}(\beta)/s^{'})=$$

$$=E_{d1}\left(\frac{1}{N^2}\frac{1}{f_y^2(Q_y(\beta))}\sum_{i,j\in s'}\Delta_{ij}^{s'}\frac{\delta(Q_y(\beta)-y_i)}{\pi_i^*}\frac{\delta(Q_y(\beta)-y}{\pi_j^*}\right.$$

$$+\frac{H_{01}^2(Q_y(\beta),1)}{Q_x^2(\beta)}\frac{1}{N^2}\frac{1}{f_x^2(Q_x(\beta))}\times$$

$$\times\sum_{i,j\in s'}\Delta_{ij}^{s'}\frac{\delta(Q_x(\beta)-x_i)}{\pi_i^*}\frac{\delta(Q_x(\beta)-x_j)}{\pi_j^*}+$$

$$+2\frac{H_{01}(Q_y(\beta),1)}{Q_x(\beta)}\frac{1}{N^2}\frac{1}{f_y(Q_y(\beta))f_x(Q_x(\beta))}\times$$

$$\left.\times\sum_{i,j\in s'}\Delta_{ij}^{s'}\frac{\delta(Q_y(\beta)-y_i)}{\pi_i^*}\frac{\delta(Q_x(\beta)-x_j)}{\pi_j^*},\right.$$

donde $\Delta_{ij}^{'}=\pi_{ij}^{'}-\pi_i^{'}\pi_j^{'}$ y $\Delta_{ij}^{s'}=\pi_{ij/s'}-\pi_{i/s'}\pi_{j/s'}$. Esta expresión no puede obtenerse en la práctica, así que para ello

$$\sum_{i,j\in U}\Delta_{ij}^{'}\frac{\delta(Q_y(\beta)-y_i)}{\pi_i^{'}}\frac{\delta(Q_y(\beta)-y_j)}{\pi_j^{'}}$$

se estima por

$$\sum_{i,j\in s}\frac{\Delta_{ij}^{'}}{\pi_{ij}^*}\frac{\delta(\widehat{Q}_y^*(\beta)-y_i)}{\pi_i^{'}}\frac{\delta(\widehat{Q}_y^*(\beta)-y_j)}{\pi_j^{'}},$$

y

$$E_{d1}\left(\sum_{i,j\in s'}\Delta_{ij}^{s'}\frac{\delta(Q_y(\beta)-y_i)}{\pi_i^*}\frac{\delta(Q_y(\beta)-y_j)}{\pi_j^*}\right)$$

por

$$\sum_{i,j\in s}\frac{\Delta_{ij}^{s'}}{\pi_{ij/s'}}\frac{\delta(\widehat{Q}_y^*(\beta)-y_i)}{\pi_i^*}\frac{\delta(\widehat{Q}_y^*(\beta)-y_j)}{\pi_j^*}.$$

Las funciones $f_x(Q_x(\beta))$ y $f_y(Q_y(\beta))$ pueden calcularse siguiendo Silverman (1986).

Las varianzas asintóticas de los estimadores de tipo razón, producto y exponencial se derivan a partir de

$H(q,t) = q/t$, $H(q,t) = q * t$ y $H(q,t) = q/t^{\alpha}$, respectivamente.

Una vez que la clase y sus propiedades principales han sido definidas, el siguiente paso es obtener el estimador óptimo en la clase $\widehat{Q}^{*}_{ye}(\beta)$. La idea de optimalidad se define en el sentido de minimizar la varianza asintótica de estos estimadores.

El valor óptimo de α está dado por

$$\alpha_{opt} = \frac{Q_x(\beta)}{Q_y(\beta)} \frac{Cov(\widehat{Q}_y(\beta), \widehat{Q}_x(\beta)) - Cov(\widehat{Q}_y(\beta), \widehat{Q}'_x(\beta))}{V(\widehat{Q}_x(\beta)) + \widehat{Q}'_x(\beta) - 2Cov(\widehat{Q}_x(\beta), \widehat{Q}'_x(\beta))}.$$

Usando las propiedades de muestreo bifásico, se obtiene:

$$\alpha_{opt} = \frac{Q_x(\beta)}{Q_y(\beta)} \frac{f_x(Q_x(\beta))}{f_y(Q_y(\beta))} \times$$

$$\times \frac{E_{d1}\left(\sum_{i,j \in s'} \Delta^{s'}_{ij} \frac{\delta(Q_y(\beta) - y_i)}{\pi^*_i} \frac{\delta(Q_x(\beta) - x_j)}{\pi^*_j}\right)}{E_{d1}\left(\sum_{i,j \in s'} \Delta^{s'}_{ij} \frac{\delta(Q_x(\beta) - x_i)}{\pi^*_i} \frac{\delta(Q_x(\beta) - x_j)}{\pi^*_j}\right)},$$

y el estimador óptimo está dado por

$$\widehat{Q}^{\alpha_{opt}}_y(\beta) = \widehat{Q}^{*}_y(\beta) \left(\frac{\widehat{Q}'_x(\beta)}{\widehat{Q}^{*}_x(\beta)}\right)^{\alpha_{opt}}.$$

Puede verse que:

$$V(\widehat{Q}^{\mathcal{H}}_y(\beta)) \geq V(\widehat{Q}^{\alpha_{opt}}_y(\beta)) = V(\widehat{Q}_y(\beta)) - K_1 =$$

$$= V(\widehat{Q}_y(\beta)) - \frac{(Cov(\widehat{Q}_y(\beta), \widehat{Q}_x(\beta)) - Cov(\widehat{Q}_y(\beta), \widehat{Q}'_x(\beta)))^2}{V(\widehat{Q}_x(\beta)) + \widehat{Q}'_x(\beta) - 2Cov(\widehat{Q}_x(\beta), \widehat{Q}'_x(\beta))},$$

esto es, el valor más bajo de la varianza de $\widehat{Q}^{\mathcal{H}}_y(\beta)$ está dado por el estimador exponencial con $\alpha = \alpha_{opt}$.

La ecuación anterior demuestra que el estimador propuesto $\widehat{Q}^{\alpha_{opt}}_y(\beta)$ es siempre más eficiente que el estimador más simple $\widehat{Q}_y(\beta)$. Puede observarse que K_1 es la cantidad que se reduce de varianza cuando se usa el estimador exponencial con el valor óptimo de α en lugar de usar el estimador $\widehat{Q}_y(\beta)$.

En la práctica, el valor de α es desconocido. Sin embargo, los datos muestrales podrán usarse para obtener un estimador para este parámetro. Un posible estimador para el valor óptimo de α está dado por

$$\widehat{\alpha} = \frac{\widehat{Q}^{*}_x(\beta)}{\widehat{Q}^{*}_y(\beta)} \frac{f_x(Q_x(\beta))}{f_y(Q_y(\beta))} \times \tag{3.7}$$

$$\times \frac{\sum_{i,j \in s} \frac{\Delta^{s'}_{ij}}{\pi_{ij/s'}} \frac{\delta(\widehat{Q}^{*}_y(\beta) - y_i)}{\pi^*_i} \frac{\delta(Q_x(\beta) - x_j)}{\pi^*_j}}{\sum_{i,j \in s} \frac{\Delta^{s'}_{ij}}{\pi_{ij/s'}} \frac{\delta(Q_x(\beta) - x_i)}{\pi^*_i} \frac{\delta(Q_x(\beta) - x_j)}{\pi^*_j}}.$$

De este modo, se puede definir un estimador óptimo para el cuantil β como:

$$\widehat{Q}^{\widehat{\alpha}}_y(\beta) = \widehat{Q}^{*}_y(\beta) \left(\frac{\widehat{Q}'_x(\beta)}{\widehat{Q}^{*}_x(\beta)}\right)^{\widehat{\alpha}}.$$

Siguiendo el procedimiento discutido en Allen *et al.* (2002), puede demostrarse que $E(\widehat{Q}^{\widehat{\alpha}}_y(\beta)) = Q_y(\beta) + o(n^{-1})$ y al primer grado de aproximación, $V(\widehat{Q}^{\widehat{\alpha}}_y(\beta)) \cong V(\widehat{Q}^{\alpha_{opt}}_y(\beta))$, esto es, los estimadores $\widehat{Q}^{\widehat{\alpha}}_y(\beta)$ y $\widehat{Q}^{\alpha_{opt}}_y(\beta)$ son asintóticamente equivalentes.

3.2.4. Propiedades empíricas

Se han propuesto varios estimadores para cuantiles en muestreo bifásico cuando las muestras en ambas fases se seleccionan con probabilidades desiguales. A continuación se lleva a cabo un estudio de simulación con el objetivo de observar el comportamiento de estos estimadores y destacar el más eficiente entre ellos. En este estudio se han considerado las poblaciones Fam1500 y Counties (véase Apéndice A).

Se han generado 1000 muestras independientes bajo diferentes métodos de muestreo en cada fase. El tamaño muestral en la primera fase, n', se ha fijado en 150, mientras que el tamaño de la muestra de la segunda fase, n, varia entre 10 y 100. Los casos considerados son los siguientes:

1. $(Mas.Midzuno)$: Las muestras en la primera fase han sido seleccionadas mediante muestreo aleatorio simple de tamaño n', mientras que las muestras de la segunda fase se han tomado mediante el método de Midzuno (véase Singh, 2003, pg. 390). Las probabilidades de inclusión en este caso vienen dadas por:

$$\pi'_i = \frac{n'}{N}, \quad \pi_{i/s'} = \frac{n'-n}{n'-1} \frac{x_i}{\sum_{j \in s'} x_j} + \frac{n-1}{n'-1} \quad \rightarrow$$

$$\rightarrow \quad \pi^*_i = \pi'_i \pi_{i/s'}.$$

2. $(Mas.Poisson)$: En la primera fase se usa muestreo aleatorio simple de tamaño n', y las muestras de la segunda fase son seleccionadas mediante el método de Poisson (véase Singh, 2003, pg. 499), de modo que las probabilidades de inclusión están dadas por:

$$\pi'_i = \frac{n'}{N}, \quad \pi_{i/s'} = n \frac{x_i}{\sum_{j \in s'} x_j} \quad \rightarrow \quad \pi^*_i = \pi'_i \pi_{i/s'}.$$

El cumplimiento de los estimadores propuestos en muestreo bifásico para un determinado cuantil se evalúa para los tres cuartiles, $\beta = 0{,}25, 0{,}50, 0{,}75$, en términos de Sesgo Relativo (%) (SR) y Eficiencia Relativa (ER) mediante aproximaciones Monte Carlo derivadas de $B = 1000$ muestras independientes. Estas medidas vienen dadas por:

$$SR_i = 100 \times \frac{1}{B} \sum_{b=1}^{B} \frac{\widehat{Q}^{i}_y(\beta)_b - Q_y(\beta)}{Q_y(\beta)}; ER_i = \frac{ECM[\widehat{Q}^{i}_y(\beta)]}{ECM[\widehat{Q}^{*}_y(\beta)]},$$

donde b indica la b-ésima simulación y $\widehat{Q}^{i}_y(\beta)$ denota el i-ésimo estimador propuesto, con

- $\widehat{Q}^{1}_y(\beta) = \widehat{Q}^{*}_y(\beta) \frac{\widehat{Q}'_x(\beta)}{\widehat{Q}^{*}_x(\beta)}$,

- $\widehat{Q}_y^2(\beta) = \widehat{Q}_y^*(\beta) \left(\dfrac{\widehat{Q}'_x(\beta)}{\widehat{Q}_x^*(\beta)} \right)^{\widehat{\alpha}}$, donde $\widehat{\alpha}$ está dado en (3.7),
- $\widehat{Q}_y^3(\beta) = \widehat{Q}_y^*(\beta) \left(\dfrac{\widehat{Q}'_x(\beta)}{\widehat{Q}_x^*(\beta)} \right)^{\alpha_{opt}}$.

$ECM[\widehat{Q}_y^i(\beta)] = B^{-1} \sum_{b=1}^{B} [\widehat{Q}_y^i(\beta)_b - Q_y(\beta)]^2$ es el Error Cuadrático Medio empírico y $ECM[\widehat{Q}_y^*(\beta)]$ se define análogamente para $\widehat{Q}_y^*(\beta)$, el estimador directo definido en (3.2). Se recuerda que este estimador no usa información auxiliar.

Las Figuras B.12,. . .,B.15 representan la eficiencia relativa para los estimadores $\widehat{Q}_y^1(\beta)$, $\widehat{Q}_y^2(\beta)$ y $\widehat{Q}_y^3(\beta)$ en las diferentes poblaciones y bajo los diseños muestrales $Mas.Midzuno$ y $Mas.Poisson$. Estas figuras muestran el comportamiento de los estimadores cuando aumenta el tamaño muestral en la segunda fase, mientras que el tamaño muestral de la primera fase permanece constante.

Cuando existe alta correlación lineal entre y y la variable auxiliar, todos los estimadores son más eficientes que el estimador $\widehat{Q}_y^*(\beta)$, mostrado con líneas horizontales. La ganancia en eficiencia relativa decrece cuando aumenta el tamaño muestral de la segunda fase. Este resultado resulta lógico porque si el tamaño muestral en la segunda fase es pequeño, entonces la muestra tendrá menos información de la variable y, y el estimador $\widehat{Q}_y^*(\beta)$ presentará mayor grado de error, mientras que los estimadores de tipo razón y exponencial son más eficientes porque usan más información. Cuando n incrementa, $\widehat{Q}_y^*(\beta)$ obtiene mejores estimaciones y más cercanas a las estimaciones de los estimadores de tipo razón y exponencial.

$\widehat{Q}_y^3(\beta)$ es el estimador más eficiente en la mayoría de los casos. Este resultado era deseable puesto que este estimador es asintóticamente óptimo en la clase (3.3). Sin embargo, el estimador $\widehat{Q}_y^2(\beta)$ presenta valores bastantes similares y no depende de valores desconocidos. Se observa que $\widehat{Q}_y^1(\beta)$ es el estimador menos eficiente de entre los estimadores propuestos. Cuando la relación lineal entre las variable es más débil, $\widehat{Q}_y^1(\beta)$ es incluso menos eficiente que el estimador directo, mientras que $\widehat{Q}_y^2(\beta)$ y $\widehat{Q}_y^3(\beta)$ continúan teniendo un buen comportamiento. En resumen, el uso del estimador exponencial mejora las estimaciones, especialmente si la relación lineal entre las variables es débil.

Por otro lado, el método de Poisson produce resultados más eficientes en el sentido de ER que el método de Midzuno y con respecto al estimador $\widehat{Q}_y^*(\beta)$. Esto se debe a que el estimador directo presenta estimaciones muy dispersas bajo el método de Poisson causadas por la heterogeneidad de las probabilidades de inclusión.

Los estimadores propuestos son casi equivalentes en la población Counties porque los coeficientes de correlación lineal están más cercanos a 1. De hecho, la ER de los estimadores propuestos en esta población es mejor que la ER en la población Fam1500.

El estudio del sesgo es otro aspecto importante, particularmente para estimadores de tipo razón, que puede probar la existencia de sub-estimaciones o sobre-estimaciones en los estimadores. Los valores SR en la población Fam1500 están todos dentro de un rango razonable, teniendo el estimador $\widehat{Q}_y^*(\beta)$ el mayor valor en torno al 3 %, como puede verse en la Figura B.16. Los valores de SR para la población Counties cuando x_1 se usa como variable auxiliar y x_2 para asignar probabilidades están mostrados en la Figura B.17. El estimador $\widehat{Q}_y^*(\beta)$ obtiene claramente sobre-estimación, especialmente cuando el tamaño muestral en la segunda fase es pequeño y bajo el diseño muestral $Mas.Poisson$. El valor absoluto de los valores SR para los estimadores propuestos son menores de 7 % para el diseño $Mas.Midzuno$ y menores de 13 % para el diseño $Mas.Poisson$, excepto en muestras pequeñas para el estimador $\widehat{Q}_y^2(\beta)$, el cual no supera el 25 %. En resumen, el estudio de los valores SR revela que los estimadores propuestos presentan un menor sesgo que el estimador directo.

3.2.5. Aplicación al muestreo estratificado

Es sabido que el muestreo estratificado es una potente técnica que proporciona resultados eficientes cuando la población está adecuadamente estratificada y las variables auxiliares y principal presentan una alta correlación. Sin embargo, el muestreo bifásico es la herramienta más apropiada cuando la información auxiliar poblacional no está disponible, que es lo que ocurre en la mayoría de los casos. Estas dos técnicas pueden combinarse en el llamado muestreo bifásico aplicado a la estratificación. Asumiendo este diseño muestral, en esta sección se define un estimador para la función de distribución y se estudian sus principales propiedades. Este estimador se usará para construir nuevos estimadores de cuantiles, y aplicando la relación entre ambos parámetros, será posible también determinar la expresión asintótica de la varianza del estimador propuesto. La estimación de la varianza es un aspecto muy importante con un alto número de aplicaciones, tal como la construcción de intervalos de confianza, obtención del tamaño muestral óptimo, etc. Por esta razón, tanto el estimador propuesto como su varianza se analizan mediante un estudio de simulación. Los resultados de este estudio reflejan algunas útiles ganancias en eficiencia del estimador propuesto y de su varianza sobre otros estimadores.

La única diferencia de este método de muestreo con respecto al expuesto en la Sección 3.2.2, es el uso adicional del muestreo estratificado. Bajo determinadas condiciones, esta técnica es particularmente eficiente, siendo frecuentemente utilizada en la práctica por diferentes razones: (i) administrativas, cuando el marco de trabajo está dividido en varios distritos geográficos, (ii) importante ganancia en eficiencia sobre diseños muestrales no estratificados, etc.

En resumen, el muestreo bifásico aplicado a la estratificación combina las principales ventajas del muestreo bifásico y muestreo estratificado. Esta técnica consiste en tomar una primera gran muestra de la población en estudio según un diseño muestral determinado. En esta muestra, se observa una variable auxiliar, la cual se usa para estratificar dicha muestra en H estratos. De cada estrato, se selecciona una muestra y se observa la variable de interés.

A continuación se describe el muestreo bifásico apli-

cado a la estratificación y el estimador natural para estimar la función de distribución. Además, se propone un estimador para la función de distribución basado en estimadores π^*.

La notación seguida para el muestreo bifásico aplicado a la estratificación es la siguiente. Una primera muestra s' de tamaño n' es diseñada según el diseño muestral d_1, de modo que $p_{d1}(s')$ es la probabilidad de que s' sea seleccionada y donde las correspondientes probabilidades de inclusión de primer y segundo orden se denotan como $\pi_i^{'}$ y $\pi_{ij}^{'}$, para $i,j \in U$. Para los elementos en s', se recoge la información de una variable auxiliar, x. Esta variable se usa para dividir s' en H pre-especificados estratos denotados como $s_h^{'}$, $(h=1,\ldots,H)$, con $n_h^{'}$ elementos en el estrato h. De este modo, de $s_h^{'}$ se puede seleccionar una muestra s_h de tamaño n_h mediante un diseño $p_h(/s')$. La muestra final será $s=\bigcup_{h=1}^{H} s_h$. La probabilidades de inclusión para las unidades de la segunda fase se denotan como $\pi_{i/s'}$ y $\pi_{ij/s'}$, para $i,j \in s'$. Notamos que $\Delta_{ij}^{'}=\pi_{ij}^{'}-\pi_i^{'}\pi_j^{'}$ y $\Delta_{ij}^{s'}=\pi_{ij/s'}-\pi_{i/s'}\pi_{j/s'}$.

El primer paso para estimar un determinado cuantil es obtener un buen estimador para la función de distribución con propiedades deseables. El candidato natural (estimador de tipo Horvitz y Thompson) para estimar la función de distribución bajo la técnica de muestreo en estudio es:

$$\widehat{F}_{st}(t)=\frac{1}{N}\sum_{h=1}^{H}\sum_{i\in s_h}\frac{\delta(t-y_i)}{\pi_i},$$

donde las probabilidades de inclusión están dadas por $\pi_i=\sum_{s'\ni i}p_{d1}(s')\pi_{i/s'}$. Este estimador no puede obtenerse siempre en la práctica debido a que las probabilidades $\pi_{i/s'}$, para cada s', deben de conocerse para poder determinar π_i. Esto no es siempre posible porque $\pi_{i/s'}$ puede depender del resultado de la primera fase (por ejemplo si la muestra de la segunda fase se selecciona mediante un muestreo proporcional a una variable auxiliar).

En la práctica, el uso del estimador de tipo Horvitz-Thompson no resulta posible ni para el problema de la estimación de la media poblacional. Por esta razón, Särndal *et al.* (1992) propusieron el uso de π^*-estimadores. Usando este idea, se introducen las cantidades $\pi_i^*=\pi_i^{'}\pi_{i/s'}$ y $\pi_{ij}^*=\pi_{ij}^{'}\pi_{ij/s'}$ para definir el π^*-estimador de la función de distribución como

$$\widehat{F}_{st}^*(t)=\frac{1}{N}\sum_{h=1}^{H}\sum_{i\in s_h}\frac{\delta(t-y_i)}{\pi_i^*}. \qquad (3.8)$$

La calidad de un estimador de la función de distribución puede medirse a través de diversas propiedades deseables (véase Chambers *et al.* , 1992). A continuación se analizan algunas de las más importantes para el estimador dado por (3.8).

Simplicidad

El cálculo de un estimador de la función de distribución, $\widehat{F}_y(t)$, será particularmente simple si

$$\widehat{F}_y(t)=\frac{1}{N}\sum_{i\in s}w_i\delta(t-y_i),$$

donde los pesos w_i dependen sólo de la etiqueta i. Esto es particularmente deseable para investigaciones con múltiples características. Puede comprobarse fácilmente que el $\widehat{F}_{st}^*(t)$ posee esta propiedad.

Unicidad en la definición

El estimador propuesto es un estimador basado en el diseño muestral, el cual no depende de la elección de un modelo. Además se ha asumido que los estratos están pre-especificados. De este modo, la expresión para $\widehat{F}_{st}^*(t)$ es única.

Sesgo

Una medida importante de la calidad de un estimador es la insesgadez. Särndal *et al.* (1992) establecieron que, para el caso de estimar el total poblacional, el π^*-estimador es insesgado. Este resultado puede extenderse fácilmente al problema de la estimación de la función de distribución, esto es, asumiendo que $z_i=\delta(t-y_i)$ es la variable de interés, el estimador (3.8) puede verse como un problema de estimación de la media poblacional de la variable z_i.

Disponibilidad de la varianza

Siguiendo la demostración del Teorema 3.2, puede comprobarse fácilmente que la varianza de $\widehat{F}_{st}^*(t)$ está dada por

$$V(\widehat{F}_{st}^*(t))=\frac{1}{N^2}\left[\sum_{i,j\in U}\Delta_{ij}^{'}\frac{\delta(t-y_i)}{\pi_i^{'}}\frac{\delta(t-y_j)}{\pi_j^{'}}+\right.$$

$$\left.+E_{d1}\left(\sum_{h=1}^{H}\sum_{i,j\in s_h^{'}}\Delta_{ij}^{s'}\frac{\delta(t-y_i)}{\pi_i^*}\frac{\delta(t-y_j)}{\pi_j^*}\right)\right]. \qquad (3.9)$$

De este modo, un estimador insesgado de esta varianza viene dado por:

$$\widehat{V}(\widehat{F}_{st}^*(t))=\frac{1}{N^2}\left(\sum_{i,j\in s}\frac{\Delta_{ij}^{'}}{\pi_{ij}^*}\frac{\delta(t-y_i)}{\pi_i^{'}}\frac{\delta(t-y_j)}{\pi_j^{'}}+\right.$$

$$\left.+\sum_{h=1}^{H}\sum_{i,j\in s_h}\frac{\Delta_{ij}^{s'}}{\pi_{ij/s'}}\frac{\delta(t-y_i)}{\pi_i^*}\frac{\delta(t-y_j)}{\pi_j^*}\right), \qquad (3.10)$$

puesto que cada componente de (3.10) es insesgado de su correspondiente componente en la ecuación (3.9).

$\widehat{F}_{st}^*(t)$ es una verdadera función de distribución

En primer lugar, notamos que varios de los estimadores propuestos en la literatura no son verdaderas funciones de distribución. Por ejemplo, ninguno de los conocidos estimadores de tipo razón y diferencia propuestos por Rao *et al.* (1990) es una función de distribución en general (véase Kuk, 1993, Mukhopadhyay, 2000).

Las condiciones (C2.18) y (C2.19) siempre se satisfacen para $\widehat{F}_{st}^*(t)$ y el valor límite de $\widehat{F}_{st}^*(t)$ es también igual a 0. En general, $\lim_{t\to+\infty}\widehat{F}_{st}^*(t)$ no es igual a

1, aunque esto se verifica para algunos diseños muestrales tal como muestreo aleatorio simple. En la Sección 3.2.7 se analiza $\lim_{t\to+\infty} \widehat{F}^*_{st}(t)$ para algunos diseños muestrales mediante un estudio de simulación. Los resultados obtenidos para la población Fam1500 sostienen que este valor está bastante próximo a 1. En resumen, el estimador $\widehat{F}^*_{st}(t)$ mantiene todas las condiciones para ser una verdadera función de distribución, excepto en $\lim_{t\to+\infty} \widehat{F}^*_{st}(t) = 1$, la cual se verifica para algunos diseños muestrales y está bastante próximo a 1 en otros.

La mayoría de los estimadores de cuantiles se obtiene mediante la inversión de la función de distribución. Asumiendo muestreo bifásico, Singh *et al.* (2001) propusieron el siguiente estimador:

$$\widehat{F}_{SJT}(t) = \frac{n_x^{'}\widetilde{F}^*_{YA}(t)}{n'} + \frac{(n'-n_x^{'})\widetilde{F}^*_{YB}(t)}{n'},$$

donde $n_x^{'}$ es el número de unidades en la primera muestra con $x \leq \widehat{Q}_x^{'}(0{,}5)$ y $\widetilde{F}^*_{YA}(t)$ y $\widetilde{F}^*_{YB}(t)$ denotando la proporción de unidades en la muestra de la segunda fase para las cuales $x \leq \widehat{Q}_x^{'}(0{,}5)$ y $x > \widehat{Q}_x^{'}(0{,}5)$, respectivamente, que tiene valores de y menores o iguales que t. $\widehat{Q}_x^{'}(0{,}5)$ es el estimador de tipo Horvitz-Thompson para $Q_x(0{,}5)$ basado en la primera muestra. De este modo, se definió el siguiente estimador para la mediana

$$\widehat{Q}_{SJT}(0{,}5) = \widehat{F}^{-1}_{SJT}(0{,}5) = \text{ínf}\{t|\widehat{F}_{SJT}(t) \geq 0{,}5\} \quad (3.11)$$

Siguiendo esta técnica, el cuantil de orden β puede estimarse a partir de $\widehat{F}^*_{st}(t)$ como

$$\widehat{Q}^*_{st}(\beta) = \widehat{F}^{*-1}_{st}(\beta) = \text{ínf}\{t|\widehat{F}^*_{st}(t) \geq \beta\}. \quad (3.12)$$

3.2.6. Propiedades teóricas

A continuación se estudian las propiedades del estimador $\widehat{Q}^*_{st}(\beta)$. Para ello, se necesita una aproximación lineal debido a que $\widehat{Q}^*_{st}(\beta)$ no es una función continua.

Teorema 3.4 *El estimador $\widehat{Q}^*_{st}(\beta)$ es asintóticamente insesgado para $Q_y(\beta)$.*

Demostración

El estimador $\widehat{Q}^*_{st}(\beta)$ puede expresarse asintóticamente como una función lineal de la función de distribución estimada evaluada en el cuantil $Q_y(\beta)$ mediante la representación de Bahadur (véase Chambers y Dunstan, 1986):

$$\widehat{Q}^*_{st}(\beta) - Q_y(\beta) = \frac{1}{f_y(Q_y(\beta))}(\beta - \widehat{F}^*_{st}(Q_y(\beta))) + O(n^{-1/2}), \quad (3.13)$$

donde $f_y(\cdot)$ denota la derivada del valor límite de $F_y(\cdot)$ cuando $N \longrightarrow \infty$. Como $\widehat{F}^*_{st}(t)$ es un estimador insesgado de $F(t)$, se tiene que $E(\beta - \widehat{F}^*_{st}(Q_y(\beta))) = 0$ y considerando la expresión (3.13), puede comprobarse fácilmente que

$$E(\widehat{Q}^*_{st}(\beta)) = Q_y(\beta) + O(n^{-1/2}).$$

□

Asumiendo la insesgadez del estimador $\widehat{Q}^*_{st}(\beta)$ y la expresión (3.13), es posible determinar fácilmente la varianza de dicho estimador al primer grado de aproximación. Esta varianza queda establecida en el siguiente corolario.

Teorema 3.5 *La varianza asintótica del estimador $\widehat{Q}^*_{st}(\beta)$ viene dada por*

$$AV(\widehat{Q}^*_{st}(\beta)) = \frac{1}{N^2}\frac{1}{f_y^2(Q_y(\beta))} \times$$

$$\times \left[\sum_{i,j\in U} \Delta_{ij}^{'} \frac{\delta(\widehat{Q}^*_{st}(\beta) - y_i)}{\pi_i^{'}} \frac{\delta(\widehat{Q}^*_{st}(\beta) - y_j)}{\pi_j^{'}} + \right.$$

$$\left. + E_{d1}\left(\sum_{h=1}^{H} \sum_{i,j\in s_h^{'}} \Delta_{ij}^{s'} \frac{\delta(\widehat{Q}^*_{st}(\beta) - y_i)}{\pi_i^*} \frac{\delta(\widehat{Q}^*_{st}(\beta) - y_j)}{\pi_j^*} \right) \right].$$

Demostración

De la expresión (3.13) se deduce que

$$AV(\widehat{Q}^*_{st}(\beta)) = \frac{1}{f_y^2(Q_y(\beta))} V\left(\widehat{F}^*_{st}(Q_y(\beta))\right),$$

donde $V\left(\widehat{F}^*_{st}(Q_y(\beta))\right)$ está dada en (3.9). □

Un estimador insesgado para esta varianza viene dado por:

$$\widehat{V}(\widehat{Q}^*_{st}(\beta)) = \frac{1}{N^2}\frac{1}{f_y^2(Q_y(\beta))} \times$$

$$\times \left(\sum_{i,j\in s} \frac{\Delta_{ij}^{'}}{\pi_{ij}^*} \frac{\delta(\widehat{Q}^*_{st}(\beta) - y_i)}{\pi_i^{'}} \frac{\delta(\widehat{Q}^*_{st}(\beta) - y_j)}{\pi_j^{'}} + \right.$$

$$\left. + \sum_{h=1}^{H} \sum_{i,j\in s_h} \frac{\Delta_{ij}^{s'}}{\pi_{ij/s'}} \frac{\delta(\widehat{Q}^*_{st}(\beta) - y_i)}{\pi_i^*} \frac{\delta(\widehat{Q}^*_{st}(\beta) - y_j)}{\pi_j^*} \right). \quad (3.14)$$

Este estimador para la varianza del estimador propuesto presenta una forma explícita, lo que permite que pueda obtenerse siempre en la práctica, es decir, la expresión (3.14) no depende del valor esperado sobre el diseño de la primera fase, haciendo posible los cálculos directos.

Una vez que la varianza del estimador ha sido determinada, intervalos de confianza y otras importantes aplicaciones derivadas de la varianza podrán también obtenerse.

En el siguiente ejemplo se determina las expresiones del estimador propuesto $\widehat{Q}^*_{st}(\beta)$ y de su correspondiente varianza estimada para el caso de selección de unidades mediante muestreo aleatorio simple.

Ejemplo 3.1 *Asumiendo muestreo aleatorio simple en cada fase, el π^*-estimador viene dado por*

$$\widehat{Q}^*_{st}(\beta) = \text{ínf}\{t \mid \sum_{h=1}^{H} \frac{n_h^{'}}{n'} \sum_{i\in s_h} \frac{\delta(t - y_i)}{n_h} \geq \beta\},$$

y el estimador de su varianza puede obtenerse de (3.14) después de sustituir las probabilidades $\pi_{i/s'}$, π_i^*, $\pi_{ij/s'}$ y π_{ij}^* por

$$\pi_{i/s'} = \frac{n_h}{n_h^{'}} \quad ; \quad \pi_i^* = \frac{n'}{N}\frac{n_h}{n_h^{'}}, \quad \textit{para } i \in s_h,$$

$$\pi_{ij/s'} = \left\{ \begin{array}{l} \dfrac{n_h(n_h-1)}{n_h^{'}(n_h^{'}-1)} \text{ si } i,j \in s_h^{'} \\ \dfrac{n_h n_l}{n_h^{'} n_l^{'}} \text{ si } i \in s_h^{'} \text{ y } j \in s_l^{'} \end{array} \right\}$$

$$\pi_{ij}^{*} = \left\{ \begin{array}{l} \dfrac{n_h(n_h-1)}{n_h^{'}(n_h^{'}-1)} \dfrac{n'(n'-1)}{N(N-1)} \text{ si } i,j \in s_h^{'} \\ \dfrac{n_h n_l}{n_h^{'} n_l^{'}} \dfrac{n'(n^{'}-1)}{N(N-1)} \text{ si } i \in s_h^{'} \text{ y } j \in s_l^{'} \end{array} \right\}$$

3.2.7. Propiedades empíricas

Asumiendo muestreo bifásico aplicado a la estratificación, se ha propuesto un estimador para un determinado cuantil poblacional, mientras que su correspondiente varianza asintótica ha sido establecida. La insesgadez del estimador de cuantiles también ha sido discutida. El siguiente paso será analizar, mediante un estudio de simulación, éstas y otras medidas importantes de calidad para los dos estimadores propuestos. Los resultados se compararan sobre otros estimadores conocidos en la literatura del muestreo en poblaciones finitas.

En este estudio se usa la población Fam1500 (véase Apéndice A), donde recordamos que las correlaciones entre la variable principal y las auxiliares vienen dadas por $\rho_{y,x_1} = 0{,}848$ y $\rho_{y,x_2} = 0{,}546$.

En primer lugar, analizaremos $\lim_{t\to\infty} F_{st}^{*}(t)$ para poder comprobar cómo de cercano se encuentra de la unidad. Recordamos que $F_{st}^{*}(t)$ será una verdadera función de distribución si este valor es igual a 1. Se ha considerado muestreo aleatorio simple (S), el método de Midzuno (M) y el método de Poisson (P). Las diferentes combinaciones de diseños muestrales se van a denotar como d_{ij}, para $i,j = \{S,M,P\}$, donde i y j van a expresar los diseños muestrales usados en la primera y segunda fase, respectivamente. Este estudio se ha llevado a cabo usando aproximaciones Monte Carlo derivadas de 1000 muestras independientes, para $\beta = 0{,}5$, $n' = 150$ y 300 y varios valores de n.

Para cada diseño muestral, las Tablas 3.1 y 3.2 muestran la esperanza empírica de $\lim_{t\to\infty} \widehat{F}_{st}^{*}(t)$ basada en 1000 muestras de la población Fam1500. Puede observarse que todos los resultados están cercanos a 1, obteniéndose mejores resultados cuando la muestra de la segunda fase es mayor. Como esperábamos, asumiendo muestreo aleatorio simple en cada una de las fases, siempre se obtiene que $\lim_{t\to\infty} \widehat{F}_{st}^{*}(t) = 1$. Esto también ocurre en la mayoría de los casos cuando se considera el método de Poisson en alguna de las dos fases. En general, la variable x_1 (para correlaciones altas) obtiene mejores resultados que la variable x_2.

El siguiente paso es comparar el comportamiento del estimador propuesto para cuantiles y de su varianza con respecto a otros estimadores. En este estudio, se ha incluido el estimador (3.11) y su correspondiente estimador de la varianza propuesto en Singh *et al.* (2001). La ganancia en eficiencia sobre muestreo no estratificado puede contrastarse si comparamos el estimador propuesto con el estimador basado en la segunda fase, sin considerar estratos en la primera fase. Este estimador será denotado como $\widehat{Q}_y^{*}(\beta)$ y lo usaremos como el estimador base en las comparaciones.

Tabla 3.1: Esperanza empírica de $\lim_{t\to\infty} \widehat{F}^*_{st}(t)$ para varios diseños muestrales y considerando la variable x_1.

n'	n	d_{SS}	d_{SM}	d_{SP}	d_{MS}	d_{MM}	d_{MP}	d_{PS}	d_{PM}	d_{PP}
150	30	1.000	1.010	1.000	1.001	1.011	1.000	1.000	1.000	1.000
	50	1.000	1.005	1.000	1.001	1.006	1.000	1.000	1.000	0.999
	70	1.000	1.003	1.000	1.001	1.004	1.000	1.000	1.000	1.000
	90	1.000	1.002	1.000	1.001	1.002	1.000	0.999	1.000	1.000
300	60	1.000	1.005	1.000	1.000	1.005	1.000	0.999	1.000	1.000
	100	1.000	1.003	1.000	1.000	1.003	1.000	1.000	1.000	1.000
	140	1.000	1.001	1.000	1.000	1.002	1.000	1.000	1.000	1.000
	180	1.000	1.001	1.000	1.000	1.001	1.000	1.000	1.000	1.000

Tabla 3.2: Esperanza empírica de $\lim_{t\to\infty} \widehat{F}^*_{st}(t)$ para varios diseños muestrales y considerando la variable x_2.

n'	n	d_{SS}	d_{SM}	d_{SP}	d_{MS}	d_{MM}	d_{MP}	d_{PS}	d_{PM}	d_{PP}
150	30	1.000	1.011	1.002	1.001	1.011	0.998	1.001	1.002	1.002
	50	1.000	1.005	1.002	1.001	1.006	1.001	1.000	1.001	0.999
	70	1.000	1.003	0.999	1.001	1.004	0.999	1.000	1.000	0.999
	90	1.000	1.002	1.000	1.001	1.002	0.999	1.000	1.001	0.999
300	60	1.000	1.005	1.000	1.000	1.005	0.999	1.000	1.000	0.999
	100	1.000	1.003	1.000	1.000	1.003	1.000	1.000	1.000	0.999
	140	1.000	1.001	1.000	1.000	1.002	1.000	0.999	1.000	0.999
	180	1.000	1.001	1.000	1.000	1.001	1.000	1.000	1.000	1.000

Tabla 3.3: Medidas de eficiencia y precisión para los estimadores de cuantiles y sus varianzas asumiendo el diseño muestral d_{SM} y la variable x_1. $\beta = 0{,}5$ y $n' = 150$.

	ER				SR (%)				$RECMR$ (%)			
n	30	50	70	90	30	50	70	90	30	50	70	90
$\widehat{Q}^*_{st}$	0.59	0.69	0.59	0.68	-0.1	-0.1	-0.1	0.0	2.7	2.2	1.7	1.5
$\widehat{Q}^*_{y}$	1.00	1.00	1.00	1.00	0.2	-0.1	0.0	0.0	3.5	2.6	2.2	1.9
$\widehat{Q}_{SJT}$	0.64	0.66	0.67	0.74	-0.2	-0.1	-0.1	0.0	2.8	2.1	1.8	1.6
$\widehat{V}(\widehat{Q}^*_{st})$	0.32	0.42	0.42	0.26	-5.2	9.2	13.2	7.4	15.8	12.7	14.9	8.6
$\widehat{V}(\widehat{Q}^*_{y})$	1.00	1.00	1.00	1.00	-16.6	-13.5	-13.5	-11.3	16.6	13.5	13.5	11.3
$\widehat{V}(\widehat{Q}_{SJT})$	1.11	2.18	2.37	2.29	27.4	30.1	31.1	23.2	27.4	30.1	31.1	23.2

Tabla 3.4: Medidas de eficiencia y precisión para los estimadores de cuantiles y sus varianzas asumiendo el diseño muestral d_{SM} y la variable x_1. $\beta = 0{,}5$ y $n' = 300$.

	ER				SR (%)				$RECMR$ (%)			
n	60	100	140	180	60	100	140	180	60	100	140	180
$\widehat{Q}^*_{st}$	0.55	0.61	0.73	0.76	-0.1	0.0	-0.1	-0.1	1.8	1.4	1.3	1.1
$\widehat{Q}^*_{y}$	1.00	1.00	1.00	1.00	0.1	0.1	0.0	-0.1	2.5	1.8	1.5	1.3
$\widehat{Q}_{SJT}$	0.58	0.62	0.73	0.80	0.0	0.0	0.0	-0.1	1.9	1.4	1.3	1.1
$\widehat{V}(\widehat{Q}^*_{st})$	0.10	0.09	0.33	0.13	-4.8	-4.1	-9.9	-4.2	11.7	8.0	10.7	5.0
$\widehat{V}(\widehat{Q}^*_{y})$	1.00	1.00	1.00	1.00	-20.2	-16.2	-13.4	-10.4	20.2	16.2	13.4	10.4
$\widehat{V}(\widehat{Q}_{SJT})$	1.18	2.10	1.68	2.38	37.7	37.6	23.7	20.2	37.7	37.6	23.7	20.2

Tabla 3.5: Medidas de eficiencia y precisión para los estimadores de cuantiles y sus varianzas asumiendo el diseño muestral d_{SM} y la variable x_2. $\beta = 0{,}5$ y $n' = 150$.

	ER				SR (%)				$RECMR$ (%)			
n	30	50	70	90	30	50	70	90	30	50	70	90
$\widehat{Q}^*_{st}$	0.59	0.60	0.72	0.77	-0.1	0.0	0.1	-0.1	2.7	2.1	1.8	1.7
$\widehat{Q}^*_{y}$	1.00	1.00	1.00	1.00	0.2	0.1	0.0	-0.1	3.5	2.7	2.1	1.9
$\widehat{Q}_{SJT}$	0.78	0.84	0.90	0.94	-0.1	0.0	0.0	-0.1	3.1	2.5	2.0	1.9
$\widehat{V}(\widehat{Q}^*_{st})$	0.27	0.12	0.28	0.24	-8.1	-1.8	-2.1	-8.6	17.5	10.4	6.7	9.5
$\widehat{V}(\widehat{Q}^*_{y})$	1.00	1.00	1.00	1.00	-19.8	-18.3	-9.0	-14.9	19.8	18.3	9.0	14.9
$\widehat{V}(\widehat{Q}_{SJT})$	0.01	0.01	0.18	0.13	0.9	-1.7	4.2	-5.7	0.9	1.8	4.2	5.7

Tabla 3.6: Medidas de eficiencia y precisión para los estimadores de cuantiles y sus varianzas asumiendo el diseño muestral d_{SM} y la variable x_2. $\beta = 0{,}5$ y $n' = 300$.

	ER				SR (%)				$RECMR$ (%)			
n	60	100	140	180	60	100	140	180	60	100	140	180
$\widehat{Q}^*_{st}$	0.57	0.57	0.66	0.73	-0.1	0.0	-0.1	0.0	1.8	1.4	1.2	1.1
$\widehat{Q}^*_{y}$	1.00	1.00	1.00	1.00	0.0	-0.1	-0.1	-0.1	2.4	1.8	1.5	1.3
$\widehat{Q}_{SJT}$	0.80	0.84	0.89	0.90	-0.1	-0.1	-0.1	0.0	2.1	1.7	1.4	1.2
$\widehat{V}(\widehat{Q}^*_{st})$	0.29	0.09	0.06	0.08	0.7	3.1	-3.2	-4.8	12.0	8.4	5.8	5.7
$\widehat{V}(\widehat{Q}^*_{y})$	1.00	1.00	1.00	1.00	-12.8	-17.0	-15.5	-14.5	12.8	17.0	15.5	14.5
$\widehat{V}(\widehat{Q}_{SJT})$	0.42	0.03	0.01	0.13	10.3	3.3	2.0	5.9	10.3	3.3	2.1	5.9

La precisión de todos los estimadores de cuantiles y sus respectivas varianzas se miden para $\beta = 0{,}5$ mediante el Sesgo Relativo (SR), la Eficiencia Relativa (ER) y la Raíz cuadrada del Error Cuadrático Medio Relativo ($RECMR$). Para un cuantil, $\widehat{Q}_y(\beta)$, están medidas están dadas por

$$SR[\widehat{Q}_y(\beta)] = \frac{E[\widehat{Q}_y(\beta)] - Q_y(\beta)}{Q_y(\beta)},$$

$$ER[\widehat{Q}_y(\beta)] = \frac{ECM[\widehat{Q}_y(\beta)]}{ECM[\widehat{Q}^*_y(\beta)]},$$

$$RECMR[\widehat{Q}_y(\beta)] = \frac{\left(ECM[\widehat{Q}_y(\beta)]\right)^{1/2}}{Q_y(\beta)},$$

y para el estimador de la varianza de un cuantil, $\widehat{V}(\widehat{Q}_y(\beta))$, las medidas son

$$SR[\widehat{V}(\widehat{Q}_y(\beta))] = \frac{E[\widehat{V}(\widehat{Q}_y(\beta))] - V[Q_y(\beta)]}{V[Q_y(\beta)]},$$

$$ER[\widehat{V}(\widehat{Q}_y(\beta))] = \frac{ECM[\widehat{V}(\widehat{Q}_y(\beta))]}{ECM[\widehat{V}(\widehat{Q}^*_y(\beta))]},$$

$$RECMR[\widehat{V}(\widehat{Q}_y(\beta))] = \frac{\left(ECM[\widehat{V}(\widehat{Q}_y(\beta))]\right)^{1/2}}{V[Q_y(\beta)]},$$

donde $E[\cdot]$, $ECM[\cdot]$ y $V[\cdot]$ denotan las Esperanzas, Errores Cuadráticos Medios y Varianzas empíricas basadas en 1000 muestras. Notamos que valores de $ER[\widehat{Q}_y(\beta)]$ y $ER[\widehat{V}(\widehat{Q}_y(\beta))]$ menores de 1 indican que $\widehat{Q}_y(\beta)$ y $\widehat{V}(\widehat{Q}_y(\beta))$ son más precisos que $\widehat{Q}^*_y(\beta)$ y $\widehat{V}(\widehat{Q}^*_y(\beta))$, respectivamente. También se ha calculado la Cobertura de los intervalos de confianza al 95 % (asumiendo distribución normal) y la longitud media de los intervalos basados en 1000 muestras.

Asumiendo muestreo aleatorio simple para obtener la muestra de la primera fase y el método de Midzuno para obtener la segunda muestra, en las Tablas 3.3 y 3.4 pueden observarse los resultados de las distintas medidas de precisión para los estimadores y asumiendo la variable x_1. En este caso (para una alta correlación), tanto el estimador propuesto como su correspondiente varianza son mas precisos, en términos de ER, que sus competidores. Los valores absolutos de las medidas SR, para todos los cuantiles, son siempre menores de $0{,}2\,\%$. Respecto a las varianzas, se observa que $\widehat{V}(\widehat{Q}^*_y)$ presenta subestimación, mientras que $\widehat{V}(\widehat{Q}_{SJT})$ claramente arrastra una seria sobreestimación. Los estimadores propuestos también presentan la mejor precisión en términos de $RECMR$.

A continuación se analiza la precisión de los estimadores usando una menor correlación entre la variable principal y auxiliar. Para ello, observamos las Tablas 3.5 y 3.6. El estimador propuesto para estimar cuantiles es más preciso que el resto en términos de ER. Respecto a la estimación de varianzas, $\widehat{V}(\widehat{Q}_{SJT})$ parece tener el mejor comportamiento, aunque esto sólo ocurre para una escasa correlación entre las variables (situación no deseada en la práctica) y para el caso de varianzas. Conclusiones similares pueden obtenerse a partir del sesgo y del error cuadrático medio. Como resulta razonable, éstas últimas medidas mejoran para cada estimador a medida que se aumenta el tamaño de la muestra de cualquiera de las dos fases.

Por último, se analiza la cobertura y la longitud media de los intervalos de confianza de cada estimador. Estas medidas vienen dadas por las Tablas 3.7 y 3.8 para la variable x_1 y las Tablas 3.9 y 3.10 para la variable x_2. En todos los casos se observa que el estimador propuesto tiene la menor longitud media empírica para el intervalo de confianza. Para altas correlaciones, la cobertura del estimador propuesto es mejor que la del resto de estimadores, puesto que se obtienen valores más próximos al 95 %. Para bajas correlaciones, la cobertura del estimador propuesto se ve ligeramente superada por la cobertura de $\widehat{Q}_{SJT}$, aunque éste último estimador tiene el inconveniente de presentar intervalos de confianza mucho más amplios. Todas estas propiedades teóricas y empíricas bajo muestreo bifásico aplicado a la estratificación pueden también consultarse en Rueda, Arcos, Muñoz y Singh (2006) y Rueda y Muñoz (2006c).

Tabla 3.7: Cobertura y Longitud Media de Intervalos de Confianza de los distintos estimadores bajo el diseño d_{SM} y asumiendo la variable x_1. $\beta = 0{,}5$ y $n' = 150$.

	Cobertura (%)				Longitud Media			
n	30	50	70	90	30	50	70	90
$\widehat{Q}^*_{st}$	94.1	93.4	96.6	95.3	828	656	566	512
$\widehat{Q}^*_{y}$	92.2	92.5	92.8	93.9	1010	772	646	564
$\widehat{Q}_{SJT}$	96.9	97.3	97.4	96.8	998	771	650	571

Tabla 3.8: Cobertura y Longitud Media de Intervalos de Confianza de los distintos estimadores bajo el diseño d_{SM} y asumiendo la variable x_1. $\beta = 0{,}5$ y $n' = 300$.

	Cobertura (%)				Longitud Media			
n	60	100	140	180	60	100	140	180
$\widehat{Q}^*_{st}$	94.4	93.9	93.7	93.2	568	447	385	347
$\widehat{Q}^*_{y}$	92.1	93.1	93.0	93.1	701	534	444	385
$\widehat{Q}_{SJT}$	96.8	98.1	96.9	97.0	703	541	454	398

Tabla 3.9: Cobertura y Longitud Media de Intervalos de Confianza de los distintos estimadores bajo el diseño d_{SM} y asumiendo la variable x_2. $\beta = 0{,}5$ y $n' = 150$.

	Cobertura (%)				Longitud Media			
n	30	50	70	90	30	50	70	90
$\widehat{Q}^*_{st}$	93.7	94.0	94.7	93.8	830	655	567	512
$\widehat{Q}^*_{y}$	90.7	93.5	94.1	92.8	1010	772	646	565
$\widehat{Q}_{SJT}$	93.8	94.7	95.4	94.5	1001	775	654	576

Tabla 3.10: Cobertura y Longitud Media de Intervalos de Confianza de los distintos estimadores bajo el diseño d_{SM} y asumiendo la variable x_2. $\beta = 0{,}5$ y $n' = 300$.

	Cobertura (%)				Longitud Media			
n	60	100	140	180	60	100	140	180
$\widehat{Q}^*_{st}$	94.8	95.7	94.8	92.4	568	447	385	347
$\widehat{Q}^*_{y}$	92.7	92.8	92.6	92.4	701	534	444	385
$\widehat{Q}_{SJT}$	96.3	95.1	94.8	94.7	707	541	461	406

3.3. Estimadores bajo muestreo en dos ocasiones sucesivas

El muestreo en ocasiones sucesivas es una técnica muy conocida que puede emplearse en las investigaciones longitudinales para estimar determinados parámetros poblacionales y medidas de diferencia o cambio de una variable objeto de estudio. En esta sección se discute la estimación de cuantiles en la ocasión más reciente bajo un muestreo en dos ocasiones sucesivas. Este estudio se realiza, por un lado, haciendo un uso más efectivo de la información auxiliar, es decir, considerando varias variables auxiliares en la etapa de estimación. Por otro lado, también se obtienen estimadores basados en muestreos con probabilidades de selección de unidades desiguales. Se estudian las propiedades más importantes y se deducen las expresiones de las varianzas. Como es habitual, se mide la precisión de los estimadores propuestos en estudios de simulación basados en varias poblaciones.

3.3.1. Introducción

En numerosas investigaciones por muestreo, una misma población puede ser muestreada repetidamente y la misma variable de estudio es medida en cada ocasión, de modo que se sigue el desarrollo de ésta sobre el tiempo. Por ejemplo, las encuestas de presupuestos familiares son llevadas a cabo periódicamente para estimar el número de empleados, las encuestas de opinión se llevan a cabo a intervalos regulares de tiempo para medir las preferencias de los votantes, etc. En estos casos, el uso de la teoría de un esquema de muestreo sucesivo puede ser una alternativa atractiva para mejorar las estimaciones de nivel en un punto en el tiempo, el cambio entre dos puntos, etc. (véase por ejemplo Cochran, 1977).

El muestreo en ocasiones sucesivas ha sido extensamente usado en las ciencias sociales y aplicadas para estimar medidas de nivel, cambios de un parámetro lineal tal como la media o el total (véase, por ejemplo, Särndal *et al.*, 1992), estimación de la varianza de este cambio (Berger, 2004), etc. Otros ejemplos del uso de encuestas longitudinales pueden consultarse en Ruspini (1999) para el análisis en el cambio social, Solga (2001) para el estudio de movilidad laboral, etc.

Asumiendo muestreo en dos ocasiones sucesivas, la teoría desarrollada por Jessen (1942) y Patterson (1950) proporciona el estimador óptimo de la media poblacional en la segunda ocasión, combinando dos estimadores distintos de esta media. Por un lado, se usa un estimador de tipo regresión basado en la muestra solapada de la muestra, considerando que la variable auxiliar es el valor de la variable principal en la primera ocasión. Por ultimo, se considera un estimador simple de la media basado en una muestra aleatoria de la porción no solapada de la segunda ocasión. El muestreo en ocasiones sucesivas también ha sido discutido en Narain (1953), Adhvaryu (1978), Eckler (1955), Gordon (1983), Arnab y Okafor (1992), Sen (1972, 1973), Singh y Srivastava (1973), Sen *et al.* (1975), Singh *et al.* (1992) y Singh (2003), el cual proporciona una extensa bibliografía sobre este tópico. En todos los estudios anteriores, el parámetro considerado para su estimación es la media poblacional.

Recientemente, Martínez *et al.* (2005) propusieron una metodología de estimación de cuantiles bajo muestreo en ocasiones sucesivas usando el valor de la variable principal en una ocasión anterior como variable auxiliar. Este estudio fue desarrollado bajo muestreo aleatorio simple y asumiendo que sobre la ocasión más reciente se toma una submuestra a partir de las unidades previamente seleccionadas, y que ciertas de estas unidades son reemplazadas por otras nuevas unidades seleccionadas independientemente de la muestra solapada.

Asumiendo un muestreo en dos ocasiones sucesivas, se propone un estimador para un cuantil de orden β que emplea una información auxiliar multivariante. El diseño muestral usado en cada fase es el muestreo aleatorio simple. Por otro lado, también se propone un estimador de cuantiles cuando las correspondientes muestras son seleccionadas mediante diseños muestrales arbitrarios en cada una de las dos fases que consta este esquema de muestreo. En este caso, se usará un estimador de tipo razón en la porción de muestra solapada para proporcionar el estimador óptimo de un cuantil. Para ello, se pondera las estimaciones inversamente a sus varianzas. Las propiedades del estimador propuesto se estudian bajo aproximaciones basadas en muestras de gran tamaño. El comportamiento de estos nuevos estimadores también se estudiarán bajo los datos de una población real.

La notación habitual a seguir en muestreo en ocasiones sucesivas es la siguiente. Consideramos que estamos haciendo un seguimiento continuo de la población U, de tamaño N, sobre dos, o más, periodos de tiempo con valores y_i en el periodo u ocasión más reciente. Se asume que una muestra de tamaño n' está diseñada en la ocasión anterior. En la ocasión reciente, una submuestra (llamada muestra solapada) de tamaño m es diseñada de las n' unidades seleccionadas previamente, y $u = n - m$ unidades son reemplazadas por nuevas unidades seleccionadas de la población restante. $\chi = m/n$ será la fracción de solapamiento.

En muestreo con dos ocasiones sucesivas, el estimador habitual para la estimación de cuantiles se construye como sigue. En primer lugar se estima la función de distribución a partir de la muestra s obtenida en la ocasión más reciente. Este estimador viene dado por $\widehat{F}_{yn}(t) = n^{-1}\sum_{i\in s}\delta(t-y_i)$, el cual coincide con el estimador de tipo Horvitz-Thompson bajo muestreo aleatorio simple. A continuación se estima el cuantil de orden β a partir de esta función de distribución, es decir:

$$\widehat{Q}_{yn}(\beta) = \widehat{F}_{yn}^{-1}(\beta) = \text{ínf}\left\{t : \widehat{F}_{yn}(t) \geq \beta\right\}. \qquad (3.15)$$

3.3.2. Generalización a múltiples variables auxiliares

La estimación de cuantiles bajo un muestreo con dos ocasiones sucesivas con extracción de muestras mediante muestreo aleatorio simple ha sido discutida en Martínez *et al.* (2005). Este estudio está basado en una única variable auxiliar, es decir, el uso de un número mayor de variables auxiliares no es posible. El objetivo que se persigue en la presente sección es por tanto el estudio de la

estimación de cuantiles bajo este esquema de muestreo y para un vector multivariante de variables auxiliares. En las Secciones 3.3.3 y 3.3.4 se analizan las propiedades teóricas y empíricas de este nuevo estimador. Como se ha comentado, todos estos estudios están diseñados para el clásico muestreo aleatorio simple. En la práctica el uso de técnicas de muestreo más complejas, como por ejemplo la extracción de unidades con probabilidades proporcionales al tamaño, puede producir estimaciones más eficientes. A partir de la Sección 3.3.5 se plantea el problema de la estimación de cuantiles bajo muestreo con dos ocasiones sucesivas y para un diseño muestral arbitrario.

Asumiendo muestreo aleatorio simple, en este apartado se define una clase de estimadores que pueden obtenerse a partir de un vector multivariante de variables auxiliares. En concreto, esta clase está formada por un estimador de tipo razón construido a partir de todas las variables auxiliares disponibles en las muestras que están solapadas y por un estimador de la variable de interés en la muestra no solapada de la ocasión más reciente. El estimador óptimo en el sentido de minimizar la varianza de esta clase será también obtenido.

En la presente sección y en 3.3.3 y 3.3.4 asumiremos que en la primera ocasión se dispone de P variables auxiliares, denotadas por $x_1, \ldots, x_P$. La información proporcionada por estas variables nos permitirá obtener un estimador de tipo razón multivariante a partir de las muestras solapadas. Por otro lado, también será posible obtener otro estimador para un determinado cuantil de la variable principal a partir de la muestra no solapada. La clase de estimadores propuesta en esta sección está formada por estos dos nuevos estimadores, los cuales se definen a continuación.

De modo similar a como se ha definido (3.15) y usando los datos de la muestra de la primera ocasión, pueden definirse los estimadores $\widehat{Q}_{xi}(\beta)$, para $i = 1, \ldots, P$. Análogamente, $\widehat{Q}_{xim}(\beta)$ y $\widehat{Q}_{ym}(\beta)$ denotarán los cuantiles muestrales de orden β de la muestra solapada para las variables auxiliares y principal, mientras que $\widehat{Q}_{yu}(\beta)$ denota el cuantil muestral basado en la muestra no solapada de la ocasión más reciente.

Siguiendo a Olkin (1958), se propone el siguiente estimador de tipo razón multivariante de $Q_y(\beta)$ basado en la parte solapada:

$$\widehat{Q}_{ym}^{MR}(\beta) = \sum_{1\leq i\leq P} w_i \frac{\widehat{Q}_{ym}(\beta)}{\widehat{Q}_{xim}(\beta)} \widehat{Q}_{xi}(\beta) = \sum_{1\leq i\leq P} w_i \widehat{Q}_{yrim}(\beta). \tag{3.16}$$

Los pesos w_i (verificando $\sum_{1\leq i\leq P} w_i = 1$) se obtienen de modo que maximizan la precisión del estimador $\widehat{Q}_{ym}^{MR}(\beta)$. Se usa el criterio de mínima varianza para obtener estas cantidades. Sabido esto, la varianza de este estimador viene dada por

$$V(\widehat{Q}_{ym}^{MR}(\beta)) = \sum_{1\leq i\leq P} w_i^2 V(\widehat{Q}_{yrim}(\beta)) +$$

$$+2\sum_{i<j} w_i w_j Cov(\widehat{Q}_{yrim}(\beta), \widehat{Q}_{yrjm}(\beta)).$$

Esta última ecuación puede escribirse como $V(\widehat{Q}_{ym}^{MR}(\beta)) = w'Bw$, donde $w = (w_1, \ldots, w_P)'$, $B = (b_{ij})$ y $b_{ij} = Cov(\widehat{Q}_{yrim}(\beta), \widehat{Q}_{yrjm}(\beta))$ para $i, j = 1, \ldots, P$. Para obtener el valor extremo usaremos la desigualdad de Cauchy-Schwarz, y puesto que B es semidefinida positiva, se obtiene que el valor óptimo w está dado por

$$w_{opt} = \frac{B^{-1}e}{e'B^{-1}e},$$

donde $e = (1, \ldots, 1)'$. Por tanto, la mínima varianza obtenida a partir de w_{opt} será

$$V_{min}(\widehat{Q}_{ym}^{MR}(\beta)) = \frac{1}{e'B^{-1}e}.$$

Asumiendo muestreo en dos ocasiones sucesivas, se propone el siguiente estimador compuesto que combina el anterior estimador de tipo razón múltiple basado en la muestra solapada con el estimador de la muestra no solapada:

$$\widehat{Q}_y(\beta) = W\widehat{Q}_{ymopt}^{MR}(\beta) + (1-W)\widehat{Q}_{yu}(\beta), \tag{3.17}$$

donde $\widehat{Q}_{ymopt}^{MR}(\beta)$ está dado por el estimador $\widehat{Q}_{ym}^{MR}(\beta)$ cuando se considera el valor óptimo de w, esto es w_{opt}, mientras que W es una constante que satisface $0 < W < 1$ y que es escogida de modo que el estimador $\widehat{Q}_y(\beta)$ presente la mínima varianza dentro la clase anterior. Un simple cálculo demuestra que

$$W_{opt} = \frac{V(\widehat{Q}_{yu}(\beta))}{V(\widehat{Q}_{yu}(\beta)) + V(\widehat{Q}_{ymopt}^{MR}(\beta))}. \tag{3.18}$$

En resumen, el estimador propuesto que presenta las propiedades óptimas en términos de mínima varianza está dado por

$$\widehat{Q}_{yopt}(\beta) = W_{opt}\widehat{Q}_{ymopt}^{MR}(\beta) + (1-W_{opt})\widehat{Q}_{yu}(\beta), \tag{3.19}$$

y su varianza viene dada por

$$V(\widehat{Q}_{yopt}(\beta)) = W_{opt}^2 V(\widehat{Q}_{ymopt}^{MR}(\beta)) + (1-W_{opt})^2 V(\widehat{Q}_{yu}(\beta)), \tag{3.20}$$

la cual puede también escribirse como

$$V(\widehat{Q}_{yopt}(\beta)) = \frac{V(\widehat{Q}_{yu}(\beta))V(\widehat{Q}_{ymopt}^{MR}(\beta))}{V(\widehat{Q}_{yu}(\beta)) + V(\widehat{Q}_{ymopt}^{MR}(\beta))}. \tag{3.21}$$

3.3.3. Propiedades teóricas

El siguiente paso en el estudio del estimador propuesto $\widehat{Q}_{yopt}(\beta)$ es la determinación de sus propiedades más importantes, además de la propiedad de mínima varianza ya comentada. En concreto se establece la normalidad de dicho estimador y su correspondiente varianza exacta.

Los resultados obtenidos se derivan asumiendo las siguientes condiciones:

(C3.4). Asumimos que s' es una muestra aleatoria simple de U, lo cual implica que la muestra complementaria s'^c es también una muestra aleatoria simple de U. Finalmente, asumiremos que s_m es una muestra aleatoria simple de s' y s_u es otra muestra aleatoria simple de s'^c. Bajo estas condiciones, las probabilidades de inclusión vienen dadas por: $\pi'_i = \frac{n'}{N}$, $\pi'_{ij} = \frac{n'}{N}\frac{n'-1}{N-1}$, $\pi_{i/s'} = \frac{m}{n'}$, $\pi_{ij/s'} = \frac{m(m-1)}{n'(n'-1)}$, $\pi_{i/s'^c} = \frac{u}{N-n'}$, $\pi_{ij/s'^c} = \frac{u(u-1)}{(N-n')(N-n'-1)}$.

(C3.5). Suponemos que la población finita está envuelta en una sucesión de poblaciones $\{U_\nu\}$, donde n_ν y N_ν aumentan de modo que $(n_\nu/N_\nu) \to f$ cuando $n_\nu \to \infty$.

(C3.6). Se asume que cuando $N_\nu \to \infty$ la distribución bivariante formada por (x, y) puede aproximarse por una distribución continua con densidades marginales $f_x(\cdot)$ y $f_y(\cdot)$ para x e y respectivamente, siendo $f_x(Q_x(\beta))$ y $f_y(Q_y(\beta))$ positivas.

Teorema 3.6 *El estimador de razón multivariante $\widehat{Q}_{ym}^{MR}(\beta)$ dado por (3.16) y la clase propuesta de estimadores $\widehat{Q}_y(\beta)$ dada por (3.17) son asintóticamente normales.*

Demostración

En primer lugar, los cuantiles muestrales $\widehat{Q}_{yu}(\beta)$, $\widehat{Q}_{ym}(\beta)$, $\widehat{Q}_{xi}(\beta)$ y $\widehat{Q}_{xim}(\beta)$ son asintóticamente normales como se demostró en Gross (1980).

Sean las siguientes funciones de este estimador

$$H_1(\widehat{Q}_{ym}(\beta), \widehat{Q}_{x1}(\beta), ..., \widehat{Q}_{xP}(\beta), \widehat{Q}_{x1m}(\beta), ..., \widehat{Q}_{xPm}(\beta)) = \sum_{1\le i\le P} w_i \frac{\widehat{Q}_{ym}(\beta)}{\widehat{Q}_{xim}(\beta)} \widehat{Q}_{xi}(\beta).$$

H_1 es una función continua con derivadas parciales de primer y segundo orden continuas en un entorno de $(Q_y, Q_{x1}, \ldots, Q_{xP})$. Bajo esta situación y usando los resultados de Cramer (1946), $\widehat{Q}_{ym}^{MR}(\beta)$ es asintóticamente normal.

La normalidad asintótica de la clase propuesta de estimadores se deriva fácilmente como consecuencia de la expresión lineal de la clase. □

La normalidad asintótica del estimador $\widehat{Q}_{yopt}(\beta)$ también se deriva al pertenecer este estimador a la clase (3.17).

La linealidad de la clase de estimadores también nos permitirá computar sus varianzas. Para ello, será necesario conocer las varianzas del estimador de razón multivariante basado en la muestra solapada y el estimador que solamente envuelve a la muestra no solapada, $\widehat{Q}_{yu}(\beta)$, como puede verse en (3.20) y (3.21).

Gross (1980) demostró que una expresión asintótica para la varianza del estimador $\widehat{Q}_{yu}(\beta)$ está dada por

$$V(\widehat{Q}_{yu}(\beta)) = \frac{N-u}{N}\beta(1-\beta)(u)^{-1}\{f_y(Q_y(\beta))\}^{-2}. \tag{3.22}$$

Teorema 3.7 *La varianza de $V(\widehat{Q}_{yrim}(\beta))$, con $i = 1, \ldots, P$, y la covarianza entre $\widehat{Q}_{yrim}(\beta)$ y $\widehat{Q}_{yrjm}(\beta)$, con $i, j = 1, \ldots, P$ vienen dadas por*

$$V(\widehat{Q}_{yrim}(\beta)) = \frac{\beta(1-\beta)}{f_y(Q_y(\beta))^2}\left[\left(\frac{1}{m}-\frac{1}{N}\right)+\left(\frac{1}{m}-\frac{1}{n'}\right)\times R_i\frac{f_y(Q_y(\beta))}{f_{xi}(Q_{xi}(\beta))}\left\{R_i\frac{f_y(Q_y(\beta))}{f_{xi}(Q_{xi}(\beta))}+2\left(1-\frac{P_{11}(y,x_i)}{\beta(1-\beta)}\right)\right\}\right], \tag{3.23}$$

y

$$Cov(\widehat{Q}_{yrim}(\beta), \widehat{Q}_{yrjm}(\beta)) = \frac{\beta(1-\beta)}{f_y(Q_y(\beta))^2}\left[\left(\frac{1}{m}-\frac{1}{N}\right)+\left(\frac{1}{n'}-\frac{1}{m}\right)R_i\frac{f_y(Q_y(\beta))}{f_{xi}(Q_{xi}(\beta))}\left(\frac{P_{11}(y,x_i)}{\beta(1-\beta)}-1\right)+\left(\frac{1}{n'}-\frac{1}{m}\right)R_j\frac{f_y(Q_y(\beta))}{f_{xj}(Q_{xj}(\beta))}\left(\frac{P_{11}(y,x_j)}{\beta(1-\beta)}-1\right)-\left(\frac{1}{n'}-\frac{1}{m}\right)R_iR_j\frac{f_y^2(Q_y(\beta))}{f_{xi}(Q_{xi}(\beta))f_{xj}(Q_{xj}(\beta))}\times\left(\frac{P_{11}(x_i,x_j)}{\beta(1-\beta)}-1\right)\right], \tag{3.24}$$

donde $P_{11}(y, x_i)$ denota la proporción de valores en la población para los cuales $y \le Q_y(\beta)$ y $x_i \le Q_{xi}(\beta)$, y $R_i = Q_y(\beta)/Q_{xi}(\beta)$.

Demostración

El estimador $\widehat{Q}_{yrim}(\beta)$ puede expresarse como

$$\widehat{Q}_{yrim}(\beta) = Q_y(\beta)(1+e_0)(1+e_{2i})(1-e_{1i}+e_{1i}^2+\ldots), \tag{3.25}$$

donde $e_0 = \dfrac{\widehat{Q}_{ym}(\beta)}{Q_y(\beta)} - 1$, $e_{1i} = \dfrac{\widehat{Q}_{xim}(\beta)}{Q_{xi}(\beta)} - 1$ y $e_{2i} = \dfrac{\widehat{Q}_{xi}(\beta)}{Q_{xi}(\beta)} - 1$, $i = 1, \ldots, P$.

Considerando la expansión de serie de Taylor se obtiene la expresión

$$(\widehat{Q}_{yrim}(\beta) - Q_y(\beta))(\widehat{Q}_{yrjm}(\beta) - Q_y(\beta)) \cong Q_y(\beta)^2\times (e_0+e_{2i}-e_{1i}+e_{1i}^2-e_{1i}e_{2i}-e_{1i}e_0+e_0e_{2i}+\ldots)\times (e_0+e_{2j}-e_{1j}+e_{1j}^2-e_{1j}e_{2j}-e_{1j}e_0+e_0e_{2j}+\ldots).$$

La expresión asintótica de la covarianza de los estimadores $\widehat{Q}_{yrim}(\beta)$ y $\widehat{Q}_{yrim}(\beta)$ se obtiene tomando esperanzas (se han considerado solamente términos de orden uno). Las esperanzas de las variables e_i pueden derivarse de Singh (2003):

$$\mathtt{E}[e_0^2] = \frac{N-m}{Nm}\beta(1-\beta)(Q_y(\beta)f_y(Q_y(\beta)))^{-2},$$

$$\mathtt{E}[e_{1i}^2] = \frac{N-m}{Nm}\beta(1-\beta)(Q_{xi}(\beta)f_{xi}(Q_{xi}(\beta)))^{-2},$$

$$\mathtt{E}[e_{2i}^2] = \mathtt{E}[e_{1i}e_{2i}] = \frac{N-n'}{Nn'}\beta(1-\beta)(Q_{xi}(\beta)f_{xi}(Q_{xi}(\beta)))^{-2},$$

$$\mathtt{E}[e_0e_{1i}] = \frac{N-m}{Nm}(P_{11}(y,x_i)-\beta(1-\beta))\times (Q_{xi}(\beta)Q_y(\beta)f_{xi}(Q_{xi}(\beta))f_y(Q_y(\beta)))^{-1},$$

$$\mathtt{E}[e_0e_{2i}] = \frac{N-n'}{Nn'}(P_{11}(y,x_i)-\beta(1-\beta))\times (Q_{xi}(\beta)Q_y(\beta)f_{xi}(Q_{xi}(\beta))f_y(Q_y(\beta)))^{-1},$$

$$\mathtt{E}[e_{1j}e_{2i}] = \mathtt{E}[e_{2j}e_{2i}] = \frac{N-n'}{Nn'}(P_{11}(x_j,x_i)-\beta(1-\beta))\times \times (Q_{xj}(\beta)f_{xj}(Q_{xj}(\beta))Q_{xi}(\beta)f_{xi}(Q_{xi}(\beta)))^{-1},$$

$$\mathtt{E}[e_{1j}e_{1i}] = \frac{N-m}{Nm}(P_{11}(x_j,x_i)-\beta(1-\beta))\times \times (Q_{xj}(\beta)f_{xj}(Q_{xj}(\beta))Q_{xi}(\beta)f_{xi}(Q_{xi}(\beta)))^{-1}.$$

Sustituyendo estos valores y operando adecuadamente, se obtiene la expresión dada en (3.24). □

Por tanto, usando las expresiones (3.22) (3.23) y (3.24), la matriz B, la varianza del estimador propuesto dado en (3.20) o (3.21) y el valor W_{opt} definido en (3.18) quedan determinadas.

3.3.4. Propiedades empíricas

En la Sección 3.3.2 se ha definido un estimador óptimo dentro de la clase (3.17). La normalidad y la varianza asintótica de este estimador se ha establecido en la Sección 3.3.3. El siguiente paso en este estudio es comprobar la exactitud de este estimador. En este apartado, la eficiencia del estimador propuesto y su varianza serán analizadas. En primer lugar, se analiza la ganancia en eficiencia de la varianza asintótica del estimador $\widehat{Q}_{yopt}(\beta)$ con la varianza de $\widehat{Q}_{yn}(\beta)$, el estimador estándar basado en la ocasión más reciente y el cual está dado en (3.15). A continuación, el comportamiento de estos estimadores serán contrastados en una situación real mediante un estudio empírico.

En ambos estudios se usaran dos poblaciones naturales: la población Counties y la población Turismos (véase Apéndice A). La población turismos resulta interesante en este caso porque dispone de cuatro variables auxiliares. Se pueden comparar los varios estimadores usando un número distinto de variables auxiliares, de modo que pueda observarse la evolución de la ganancia en precision al aumentar el número de variables auxiliares usadas en la etapa de estimación.

Comparaciones teóricas

El primer estudio consiste en comparar la varianza del estimador óptimo propuesto dado en (3.21) con la varianza del estimador frecuentemente usado, $\widehat{Q}_{yn}(\beta)$. Este estudio nos permitirá conocer el comportamiento de las varianzas teóricas de los estimadores. Gross (1980) comprobó que una expresión asintótica para la varianza del estimador $\widehat{Q}_{yn}(\beta)$ está dada por

$$V(\widehat{Q}_{yn}(\beta)) = \frac{N-n}{N}\beta(1-\beta)(n)^{-1}\{f_y(Q_y(\beta))\}^{-2}.$$

En las Figuras B.18 y B.19, las varianzas teóricas de los estimadores $\widehat{Q}_{yopt}(\beta)$ y $\widehat{Q}_{yn}(\beta)$ son comparadas por medio de sus cocientes, esto es, las figuras muestran los Ratios Teóricos $RT = V(\widehat{Q}_{yopt}(\beta))/V(\widehat{Q}_{yn}(\beta))$. En este estudio, se representan diferentes valores de m en el eje de abscisas y el estimador propuesto se ha obtenido para cada valor de P ($P = 1, 2$ en la población Counties y $P = 1, 2, 3, 4$ en la población Turismos). Las líneas horizontales muestran los RT para el estimador $\widehat{Q}_{yn}(\beta)$. Notamos que valores de RT por debajo de 1 indican que $V(\widehat{Q}_{yopt}(\beta))$ es menor que $V(\widehat{Q}_{yn}(\beta))$, y por tanto el estimador propuesto es más eficiente.

De estas comparaciones teóricas, se pueden destacar la siguientes conclusiones:

1. Para ambas poblaciones, el estimador propuesto parece tener uniformemente menor varianza que el estimador estándar, $\widehat{Q}_{yn}(\beta)$, y a su vez menor varianza que el estimador propuesto cuando éste utiliza una única variable auxiliar.
2. Las mejores propiedades se obtienen cuando se usan todas las variables auxiliares.
3. Cuando los tamaños muestrales en ambas ocasiones son iguales, la fracción de solapamiento óptima está entre 0.2 y 0.4. Una fracción de solapamiento más alta resulta apropiada cuando el tamaño muestral en la ocasión reciente es menor que el tamaño muestral de la primera ocasión.
4. En ambas poblaciones, los ratios más bajos se obtienen cuando los tamaños muestrales son $n' = 75$ y $n = 25$, en cuyo caso los RT, para valores grandes de χ, son aproximadamente iguales a 0.4, esto es, la varianza asintótica del estimador propuesto presenta una mejoría del $60\,\%$ con respecto a la varianza asintótica del estimador estándar.

Estudio empírico

El siguiente paso consiste en llevar a cabo un estudio de simulación con el fin de revelar la ganancia en eficiencia de $\widehat{Q}_{yopt}(\beta)$ con respecto a $\widehat{Q}_{yn}(\beta)$ en una situación real. De nuevo, las poblaciones Counties y Turismos serán usadas. Este estudio también muestra el comportamiento de $\widehat{Q}_{yopt}(\beta)$ cuando este estimador usa un número diferente de variables auxiliares.

Se generan $B = 1000$ muestras independientes bajo muestreo con dos ocasiones sucesivas. Todas las muestras (solapadas y no solapadas) se obtienen bajo muestreo aleatorio simple. El cumplimiento de estos estimadores se evalúa para el cuantil de orden $\beta = 0{,}5$ en términos de Sesgo Relativo (SR) y Eficiencia Relativa (ER), con

$$SR = \frac{1}{B}\sum_{b=1}^{B}\frac{\widehat{Q}_{yopt}(\beta)_b - Q_y(\beta)}{Q_y(\beta)}; ER = \frac{ECM[\widehat{Q}_{yopt}(\beta)]}{ECM[\widehat{Q}_{yn}(\beta)]},$$

donde b indica la b-ésima simulación, el Error Cuadrático Medio empírico está dado por

$$ECM[\widehat{Q}_{yopt}(\beta)] = \frac{1}{B}\sum_{b=1}^{B}[\widehat{Q}_{yopt}(\beta)_b - Q_y(\beta)]^2,$$

y donde $ECM[\widehat{Q}_{yn}(\beta)]$ se define de modo similar para $\widehat{Q}_{yn}(\beta)$. Por tanto, el comportamiento empírico del estimador propuesto se compara con el estimador estándar mediante diferentes valores de P.

Las generaciones aleatorias, cálculos y obtención de estimadores se han obtenido mediante el programa R. Los detalles de la programación están disponibles en el Apéndice **??**.

Las Figuras B.20 y B.21 representan la ER obtenida en el estudio de simulación. En la Figuras B.22 y B.23 se muestra la evolución de los valores óptimos W_{opt} con respecto a la fracción de solapamiento. Los valores SR están todos dentro de un rango razonable y por tanto se han omitido.

De las Figuras B.20, B.21, B.22 y B.23 se pueden hacer las siguientes observaciones:

1. Los resultados confirman un buen comportamiento por parte del estimador óptimo propuesto en comparación con el estimador estándar, y a su vez con respecto al estimador óptimo simple, es decir, el estimador propuesto óptimo basado en una única variable auxiliar.
2. Este estudio también nos muestra que se obtienen estimaciones más precisas cuando se usa un mayor número de variables auxiliares.

3. Cuando los tamaños muestrales en ambas ocasiones son iguales, la fracción de solapamiento óptima está entre $0{,}2$ y $0{,}4$. En otro caso, no puede observarse una fracción de solapamiento óptima.
4. Los valores W_{opt} son crecientes con respecto a la fracción de solapamiento. Este resultado era predecible puesto que a medida que aumenta el tamaño muestral de la parte solapada con respecto al tamaño de la muestra no solapada, el estimador de razón multivariante debería tener un mayor peso dentro del estimador propuesto. En todos los casos, los valores más altos de W_{opt} se obtienen cuando se usan todas las variables auxiliares en la etapa de estimación. Este resultado demuestra que se obtienen estimaciones más precisas cuando se usan todas las variables auxiliares: de la expresión (3.18) puede observarse que W_{opt} es mayor si $V(\widehat{Q}_{ymopt}(\beta))$ tiene valores más pequeños, y bajo esta situación, el estimador óptimo propuesto obtiene estimaciones más precisas.
5. Cuando el tamaño muestral en la segunda ocasión es menor que el tamaño en la primera ocasión, se obtiene una mayor ganancia en precisión, y esta ganancia aumenta a medida que crece la diferencia entre los tamaños muestrales. Este resultado es razonable porque si n es pequeño en relación con n', entonces, la primera muestra proporcionará mayor información, y el estimador de razón múltiple basado en la muestra solapada presentará también un menor grado de error.

En Rueda, Muñoz y Arcos (2006) pueden consultarse más detalles sobre la estimación de cuantiles en muestreo con dos ocasiones sucesivas y para un vector multivariante de variables auxiliares.

3.3.5. Muestreo con probabilidades desiguales

Asumiendo muestreo en dos ocasiones sucesivas y diseños muestrales arbitrarios para la selección de las distintas muestras que requieren ser seleccionadas bajo este esquema, Särndal *et al.* (1992) demostraron que el estimador de tipo Horvitz-Thompson de una media no puede siempre usarse en la práctica debido a que el estimador requiere el cálculo de las probabilidades de inclusión π_i, y esto no es posible para las unidades de la muestra s_u o para las unidades de la muestra s_m.

Los distintos esquemas de muestreo que pueden plantearse bajo un muestreo en dos ocasiones sucesivas y sus correspondientes probabilidades de inclusión son los que se detallan a continuación. La muestra de la primera fase s' con tamaño n' está diseñada según un diseño muestral d_1, tal que $p_{d1}(s')$ es la probabilidad de que s' sea escogida. Las correspondientes probabilidades de inclusión de primer y segundo orden vienen dadas por $\pi_i^{'}$, $\pi_{ij}^{'}$, para $i,j \in U$. Dada s', en la segunda ocasión, una muestra solapada s_m con tamaño m, es diseñada según un diseño d_2, tal que $p_m(s_m/s')$ es la probabilidad condicional de escoger s_m. Las probabilidades de inclusión bajo este diseño se denotan como $\pi_{i/s'}$ y $\pi_{ij/s'}$. La muestra no solapada s_u es por tanto seleccionada de $U - s' = s'^c$ según el diseño d_3, tal que $p_u(s_u/s'^c)$ es la probabilidad condicional de escoger s_u. Las probabilidades de inclusión bajo este diseño se denotarán como π_{i/s'^c} y π_{ij/s'^c}.

Además, en esta sección y en las dos siguientes asumiremos que se dispone de una única variable auxiliar, x, que serán los valores de la variable principal que toman los individuos en el primer periodo u ocasión. También puede considerase que x es una variable auxiliar altamente correlacionada con la variable principal, aunque en la práctica esto no es lo habitual.

A continuación se define un estimador compuesto basado en estimadores π^* (véase Särndal *et al.*, 1992, p.347) y que combina un estimador construido en la muestra solapada con otro estimador basado en la muestra no solapada.

Así, usando la muestra no solapada, s_u, es posible obtener el siguiente estimador para la función de distribución

$$\widehat{F}_{yu}(t) = \frac{1}{N}\sum_{i\in s_u}\frac{\delta(t-y_i)}{\pi_i'^c\pi_{i/s'^c}},$$

el cual es un estimador π^*. El correspondiente estimador para el cuantil de orden β viene por tanto dado por

$$\widehat{Q}_{yu}(\beta) = \text{ínf}\{t : \widehat{F}_{yu}(t) \geq \beta\}. \tag{3.26}$$

A partir de la muestra solapada pueden construirse los siguientes estimadores de la función de distribución

$$\widehat{F}_{ym}(t) = \frac{1}{N}\sum_{i\in s_m}\frac{\delta(t-y_i)}{\pi_i'\pi_{i/s'}}, \tag{3.27}$$

$$\widehat{F}_{xm}(t) = \frac{1}{N}\sum_{i\in s_m}\frac{\delta(t-x_i)}{\pi_i'\pi_{i/s'}}, \tag{3.28}$$

los cuales son estimadores π^* basados en la segunda y primera ocasión respectivamente. Usando también la muestra de la primera fase, es posible construir un estimador de tipo Horvitz-Thompson para la variable auxiliar como sigue

$$\widehat{F}_{x}(t) = \frac{1}{N}\sum_{i\in s'}\frac{\delta(t-x_i)}{\pi_i'}. \tag{3.29}$$

Usando los estimadores dados en (3.27), (3.28) y (3.29) y basándonos en la muestra solapada y en la muestra de la primera fase, se propone el siguiente estimador de tipo razón

$$\widehat{Q}^r_{ym}(\beta) = \widehat{Q}_{ym}(\beta)\frac{\widehat{Q}_x(\beta)}{\widehat{Q}_{xm}(\beta)}, \tag{3.30}$$

donde

$$\widehat{Q}_{ym}(\beta) = \text{ínf}\{t : \widehat{F}_{ym}(t) \geq \beta\}, \tag{3.31}$$

$$\widehat{Q}_{xm}(\beta) = \text{ínf}\{t : \widehat{F}_{xm}(t) \geq \beta\}, \tag{3.32}$$

$$\widehat{Q}_{x}(\beta) = \text{ínf}\{t : \widehat{F}_{x}(t) \geq \beta\}. \tag{3.33}$$

Siguiendo a Jessen (1942), se propone el estimador compuesto $\widehat{Q}^R_y(\beta)$ para $Q_y(\beta)$ como combinación lineal del estimador (3.26) y el estimador (3.30). Este estimador viene dado por

$$\widehat{Q}^R_y(\beta) = w\widehat{Q}^r_{ym}(\beta) + (1-w)\widehat{Q}_{yu}(\beta), \tag{3.34}$$

donde w es un peso constante y no negativo. El siguiente paso será determinar w de modo que se minimice la varianza del estimador compuesto $\widehat{Q}_y^R(\beta)$.

Teorema 3.8 *La varianza mínima del estimador* $\widehat{Q}_y^R(\beta)$ *viene dada por*

$$V_{min}(\widehat{Q}_y^R(\beta)) = \frac{V_1V_2 - C^2}{V_1 + V_2 - 2C}.$$

Demostración

La varianza de $\widehat{Q}_y^R(\beta)$ viene dada por

$$V(\widehat{Q}_y^R(\beta)) = w^2V(\widehat{Q}_{ym}^r(\beta)) + (1-w)^2V(\widehat{Q}_{yu}(\beta))$$
$$+2w(1-w)Cov(\widehat{Q}_{yu}(\beta), \widehat{Q}_{ym}^r(\beta)) =$$
$$= w^2V_1 + (1-w)^2V_2 + 2w(1-w)C =$$
$$(V_1 + V_2 - 2C)\{w - \frac{V_2 - C}{V_1 + V_2 - 2C}\}^2 + \frac{V_1V_2 - C^2}{V_1 + V_2 - 2C} \geq$$
$$\frac{V_1V_2 - C^2}{V_1 + V_2 - 2C} = V_{min}(\widehat{Q}_y^R(\beta)),$$

puesto que $V_1 + V_2 - 2C > 0$, y donde

$$V_1 = V(\widehat{Q}_{ym}^r(\beta)),$$
$$V_2 = V(\widehat{Q}_{yu}(\beta)),$$
$$C = Cov(\widehat{Q}_{yu}(\beta), \widehat{Q}_{ym}^r(\beta)).$$

□

Por tanto el valor de w que hace mínima la varianza de $\widehat{Q}_y^R(\beta)$ viene dado por

$$w = \frac{V_2 - C}{V_1 + V_2 - 2C}. \quad (3.35)$$

Partiendo de este resultado, el estimador propuesto será más eficiente que el estimador habitual $\widehat{Q}_{yu}(\beta)$ y el estimador de tipo razón $\widehat{Q}_{ym}^r(\beta)$.

3.3.6. Propiedades teóricas

En esta sección se estudian las propiedades asintóticas del estimador propuesto en (3.34). Los resultados que se establecen se derivan asumiendo las condiciones (C3.4), (C3.5) y (C3.6).

Teorema 3.9 *El estimador compuesto* $\widehat{Q}_y^R(\beta)$ *es asintóticamente insesgado para* $Q_y(\beta)$*.*

Demostración

Para demostrar este resultado usaremos la insesgadez de los dos estimadores en los que se basa el estimador propuesto. En primer lugar, es sabido que el cuantil muestral $\widehat{Q}_{yu}(\beta)$ es asintóticamente insesgado para $Q_y(\beta)$ (véase por ejemplo Särndal *et al.*, 1992), por lo que pasamos a estudiar si dicha propiedad la satisface el estimador de tipo razón $\widehat{Q}_{ym}^r(\beta)$. Para ello, usaremos una aproximación lineal debido a que $\widehat{Q}_{ym}^r(\beta)$ no es una función continua.

El estimador $\widehat{Q}_{ym}^r(\beta)$ puede expresarse asintóticamente como una función lineal de la función de distribución estimada evaluada en el cuantil $Q_y(\beta)$ mediante la representación de Bahadur (véase por ejemplo Chambers y Dunstan, 1986):

$$\widehat{Q}_{ym}^r(\beta) - Q_y(\beta) = \frac{1}{f_y(Q_y(\beta))}(\beta - \widehat{F}_{ym}^r(Q_y(\beta))) + o_p(n^{-1/2}), \quad (3.36)$$

donde $f_y(\cdot)$ denota la derivada del valor límite de $F_y(\cdot)$ cuando $N \longrightarrow \infty$ y $\widehat{F}_{ym}^r(t)$ denota un estimador de tipo razón para $F_y(t)$, es decir

$$\widehat{F}_{ym}^r(t) = \frac{\widehat{F}_{ym}(t)}{\widehat{F}_{xm}(t)}\widehat{F}_x(t).$$

El estimador $\widehat{Q}_{ym}^r(t)$ es asintóticamente insesgado debido a que $\widehat{F}_{ym}^r(t)$ es un estimador insesgado de $F_y(t)$ (véase Rao *et al.*, 1990). De este modo,

$$E(\beta - \widehat{F}_{ym}^r(Q_y(\beta))) = 0,$$

y usando (3.36) puede verse que

$$E(\widehat{Q}_{ym}^r(\beta)) = Q_y(\beta) + O(n^{-1/2}).$$

Puesto que $\widehat{Q}_{ym}^r(\beta)$ y $\widehat{Q}_{yu}(\beta)$ son asintóticamente insesgados para $Q_y(\beta)$, el estimador propuesto $\widehat{Q}_y^R(\beta)$ también lo será. □

Teorema 3.10 *El estimador compuesto* $\widehat{Q}_y^R(\beta)$ *es asintóticamente normal.*

Demostración

La normalidad asintótica de la clase propuesta se deriva fácilmente a partir de la expresión (3.34).

En primer lugar, bajo las condiciones (C3.4), (C3.5) y (C3.6), el cuantil muestral $\widehat{Q}_{yu}(\beta)$ es asintóticamente normal. Este resultado puede consultarse en Gross (1980).

Por otro lado, es sabido que el estimador $\widehat{F}_{ym}^r(t)$ es asintóticamente normal. Asumiendo además la aproximación lineal (3.36), puede derivarse fácilmente la normalidad del estimador $\widehat{Q}_{ym}^r(\beta)$.

Por ultimo, usando los dos resultados anteriores, la linealidad de la expresión (3.34) nos permite establecer la normalidad del estimador compuesto propuesto. □

El siguiente paso en el estudio asintótico del estimador propuesto es la determinación de una expresión para la varianza de dicho estimador. La expression (3.34) del estimador propuesto nos va a permitir computar su varianza asintótica a partir de la varianza del estimador basado en la muestra solapada, la varianza del estimador basado en la muestra no solapada y la covarianza entre ambos. Así

$$V(\widehat{Q}_y^R(\beta)) = w^2V_1 + (1-w)^2V_2 + 2w(1-w)C. \quad (3.37)$$

Estas varianzas y covarianzas toman una forma simple cuando la unidades muestrales se seleccionan mediante muestreo aleatorio simple.

Gross (1980) demostró que una expresión asintótica para la varianza del estimador $\widehat{Q}_{yu}(\beta)$ está dada por

$$V(\widehat{Q}_{yu}(\beta)) = \frac{N-u}{N}\beta(1-\beta)(u)^{-1}\{f_y(Q_y(\beta))\}^{-2}. \quad (3.38)$$

Teorema 3.11 *La varianza del estimador de razón propuesto está dada por*

$$V(\widehat{Q}^r_{ym}(\beta)) = \frac{\beta(1-\beta)}{f_y(Q_y(\beta))^2}\left[\left(\frac{1}{m}-\frac{1}{N}\right)+\left(\frac{1}{m}-\frac{1}{n'}\right)\times\right. \tag{3.39}$$
$$\left.\times R\frac{f_y(Q_y(\beta))}{f_x(Q_x(\beta))}\left\{R\frac{f_y(Q_y(\beta))}{f_x(Q_x(\beta))}+2\left(1-\frac{P_{11}(x,y)}{\beta(1-\beta)}\right)\right\}\right],$$

donde $P_{11}(x,y)$ *denota la proporción de valores en la población para los cuales* $x \leq Q_x(\beta)$ *e* $y \leq Q_y(\beta)$*, y* $R = Q_y(\beta)/Q_x(\beta)$*.*

Demostración

Usando propiedades del muestreo bifásico, la expresión asintótica para $V(\widehat{Q}^r_{ym}(\beta))$ puede obtenerse de

$$\widehat{Q}^r_{ym}(\beta) - Q_y(\beta) \cong$$
$$\widehat{Q}_{ym}(\beta) - Q_y(\beta) + \left(\frac{\widehat{Q}_{xm}(\beta)}{\widehat{Q}_x(\beta)}-1\right)(-Q_y(\beta)) = \tag{3.40}$$
$$Q_y(\beta)e_0 + (e_1-e_2)(-Q_y(\beta)) - e_2(e_1-e_2)(-Q_y(\beta)),$$

con la notación: $e_0 = \dfrac{\widehat{Q}_{ym}(\beta)}{Q_y(\beta)} - 1$, $e_1 = \dfrac{\widehat{Q}_{xm}(\beta)}{Q_x(\beta)} - 1$ y $e_2 = \dfrac{\widehat{Q}_x(\beta)}{Q_x(\beta)} - 1$.

La expresión asintótica de la varianza del estimador $\widehat{Q}^r_{ym}(\beta)$ se obtiene elevando al cuadrado los dos miembros de (3.40) y posteriormente tomando esperanzas (Notamos que solamente se han considerado términos de orden uno):

$$V(\widehat{Q}^r_{ym}(\beta)) = \frac{\beta(1-\beta)}{f_y(Q_y(\beta))^2}\left[\left(\frac{1}{m}-\frac{1}{N}\right)+\left(\frac{1}{m}-\frac{1}{n'}\right)\times\right.$$
$$\times\frac{f_y(Q_y(\beta)}{Q_x(\beta)f_x(Q_x(\beta))}(-Q_y(\beta))\times$$
$$\left.\times\left\{\frac{f_y(Q_y(\beta))}{Q_x(\beta)f_x(Q_x(\beta))}(-Q_y(\beta))+2\left(\frac{P_{11}(x,y)}{\beta(1-\beta)}-1\right)\right\}\right]$$
$$= \frac{\beta(1-\beta)}{f_y(Q_y(\beta))^2}\left[\left(\frac{1}{m}-\frac{1}{N}\right)+\left(\frac{1}{m}-\frac{1}{n'}\right)R\frac{f_y(Q_y(\beta))}{f_x(Q_x(\beta))}\times\right.$$
$$\left.\times\left\{R\frac{f_y(Q_y(\beta))}{f_x(Q_x(\beta))}+2\left(1-\frac{P_{11}(x,y)}{\beta(1-\beta)}\right)\right\}\right].$$

Los valores de $\mathtt{E}[e_0^2]$, $\mathtt{E}[e_1^2]$, $\mathtt{E}[e_2^2]$, $\mathtt{E}[e_0e_1]$ y $\mathtt{E}[e_0e_2]$ pueden verse en Allen *et al.* (2002) y Singh (2003). □

Teorema 3.12 *La covarianza entre los estimadores* $\widehat{Q}_{yu}(\beta)$ *y* $\widehat{Q}^r_{ym}(\beta)$ *está dada por*

$$Cov(\widehat{Q}_{yu}(\beta), \widehat{Q}^r_{ym}(\beta)) =$$
$$= \frac{1}{f_y(Q_y(\beta))^2}\frac{-n}{N-n}\left(1-\frac{n'}{N}\right)\frac{\beta(1-\beta)}{n'}. \tag{3.41}$$

Demostración

Para obtener la covarianza entre los estimadores $\widehat{Q}_{yu}(\beta)$ y $\widehat{Q}^r_{ym}(\beta)$ al primer orden de aproximación, nos basaremos en la propia definición de varianza:

$$Cov(\widehat{Q}_{yu}(\beta), \widehat{Q}^r_{ym}(\beta)) =$$
$$= Cov(E(\widehat{Q}_{yu}(\beta)/s'), E(\widehat{Q}^r_{ym}(\beta)/s')) +$$
$$+E(Cov(\widehat{Q}_{yu}(\beta), \widehat{Q}^r_{ym}(\beta)/s')).$$

Debido a la independencia entre s_u y s_m, el segundo término es cero. En lo que respecta al primer término

$$E(\widehat{Q}^r_{ym}(\beta)/s')) = \widehat{Q}_{ys'}(\beta) + o(m^{-1})$$

y

$$E(\widehat{Q}_{yu}(\beta)/s')) = \widehat{Q}_{ys'^c}(\beta),$$

donde

$$\widehat{Q}_{ys'}(\beta) = inf\{t : \widehat{F}_{ys'}(t) \geq \beta\},$$
$$\widehat{Q}_{ys'^c}(\beta) = inf\{t : \widehat{F}_{ys'^c}(t) \geq \beta\},$$
$$\widehat{F}_{ys'}(t) = \frac{1}{N}\sum_{i\in s'}\frac{\delta(t-y_i)}{\pi'_i}$$

y

$$\widehat{F}_{ys'^c}(t) = \frac{1}{N}\sum_{i\in s'^c}\frac{\delta(t-y_i)}{\pi'^c_i}.$$

Por otro lado, la representación de Bahadur da (véase Kuk y Mak, 1989)

$$\widehat{Q}_{ys'^c}(\beta) - Q_y(\beta) =$$
$$= \frac{1}{f_y(Q_y(\beta))}(\beta - \widehat{F}_{ys'^c}(Q_y(\beta))) + o_p(n^{-1/2}),$$
$$\widehat{Q}_{ys'}(\beta) - Q_y(\beta) =$$
$$= \frac{1}{f_y(Q_y(\beta))}(\beta - \widehat{F}_{ys'}(Q_y(\beta))) + o_p(n^{-1/2}),$$

y de este modo se obtiene

$$Cov(\widehat{Q}_{ys'^c}(\beta), \widehat{Q}_{ys'}(\beta)) \simeq$$
$$\simeq \frac{1}{f_y(Q_y(\beta))^2}Cov(\widehat{F}_{ys'}(Q_y(\beta)), \widehat{F}_{ys'^c}(Q_y(\beta))) =$$
$$\frac{1}{f_y(Q_y(\beta))^2}\frac{-n}{N-n}V(\widehat{F}_{ys'}(Q_y(\beta))) =$$
$$= \frac{1}{f_y(Q_y(\beta))^2}\frac{-n}{N-n}\left(1-\frac{n'}{N}\right)\frac{\beta(1-\beta)}{n'},$$

obteniendo así el resultado (3.41). □

Sustituyendo los valores (3.38), (3.39) y (3.41) en (3.37), se obtiene la siguiente expresión para la varianza del estimador propuesto

$$V(\widehat{Q}^R_y(\beta)) = C_1\frac{\left(\frac{n}{1-\chi}-\frac{1}{N}\right)C_0 - \left[C_1\frac{-n}{N-n}(\frac{1}{n'}-\frac{1}{N})\right]^2}{\left(\frac{n}{1-\chi}-\frac{1}{N}\right)+C_0+2C_1\frac{-n}{N-n}(\frac{1}{n'}-\frac{1}{N})}, \tag{3.42}$$

donde $\chi = m/n$ es la fracción de solapamiento,

$$C_0 = \left(\frac{1}{n\chi}-\frac{1}{N}\right)+C_2\left(\frac{1}{n\chi}-\frac{1}{n'}\right),$$
$$C_1 = \frac{\beta(1-\beta)}{f_y(Q_y(\beta))^2}$$

y

$$C_2 = R\frac{f_y(Q_y(\beta))}{f_x(Q_x(\beta))}\left\{R\frac{f_y(Q_y(\beta))}{f_x(Q_x(\beta))}+2(1-\frac{P_{11}(x,y)}{\beta(1-\beta)})\right\}.$$

El estimador resultante para $f_y(Q_y(\beta))$ junto con $p_{11}(x,y)$ (la proporción de valores en la muestra para los cuales $x \leq \widehat{Q}_x(\beta)$ y $y \leq \widehat{Q}_y(\beta)$) pueden usarse para proporcionar un estimador consistente de las varianzas asintóticas y los valores óptimos w y $1-w$.

Para completar el estudio asintótico en esta sección, analizaremos la ganancia en precisión del estimador propuesto sobre el estimador $\widehat{Q}_{yn}(\beta)$, el cual está basado exclusivamente en las n unidades muestrales para la segunda ocasión. La varianza de este estimador está dada por

$$V(\widehat{Q}_{yn}(\beta)) = \frac{N-n}{N}\beta(1-\beta)(n)^{-1}\{f_y(Q_y(\beta))\}^{-2}. \tag{3.43}$$

De este modo, la ganancia en precisión, G_1, de $\widehat{Q}_y^R(\beta)$ sobre $\widehat{Q}_{yn}(\beta)$ está dada por

$$G_1 = \frac{V(\widehat{Q}_{yn}(\beta)) - V(\widehat{Q}_y^R(\beta))}{V(\widehat{Q}_y^R(\beta))}. \tag{3.44}$$

Esta ganancia en precisión dependerá de los tamaños muestrales, del orden del cuantil y de la población objeto de estudio.

El valor óptimo de u que maximiza (3.44) coincide con el valor que minimiza la varianza asintótica (3.42).

Por tanto, el problema es obtener el mínimo en χ de la función $\phi(\chi) = V(\widehat{Q}_y^R(\beta))$ y verificando la condición natural $0 < \chi < 1$. Esta función es monótona en el intervalo $(0, 1)$. El crecimiento o decrecimiento depende del orden del cuantil y de la población en estudio. Por tanto, los valores óptimos para χ estarán próximos a cero (cuando se renueva completamente la muestra al pasar de una ocasión a otra), o bien, estarán próximos a uno (cuando la misma muestra se conserva de una ocasión a otra). Todos estos resultados asintóticos pueden también consultarse en Rueda y Muñoz (2006b).

3.3.7. Propiedades empíricas

El siguiente paso en el análisis de estimador propuesto en muestreo con dos ocasiones sucesivas y usando diseños probabilísticos desiguales consiste en llevar a cabo un estudio de simulación asumiendo distintos tamaños muestrales en todas las muestras y bajo distintos esquemas de muestreo. Para este análisis se usará la población Counties (véase Apéndice A para una descripción completa de esta población).

Como se ha podido comprobar, para la puesta en práctica de un muestreo con dos ocasiones sucesivas es necesario seleccionar tres muestras diferentes, las cuales pueden obtenerse a partir de distintos diseños muestrales. En concreto, estas tres muestras son la muestra de la primera fase, la muestra solapada y la muestra no solapada. En el estudio de simulación de esta sección se usaran las distintas combinaciones de esquemas de muestreo descritas en la Tabla 3.11. El método de Midzuno se emplea como método de extracción de unidades con probabilidades desiguales, aunque es posible la aplicación del estimador propuesto bajo cualquier otro diseño muestral.

Para cada esquema de muestreo se han generado $B = 1000$ simulaciones con tamaños muestrales $n' = 75$, $n = 25$, $m = 5, \ldots, 15$ y $n' = 75$, $n = 50$, $m = 5, \ldots, 30$. El cumplimiento del estimador propuesto se evalúa para los tres cuartiles, $\beta = 0{,}25, 0{,}50, 0{,}75$, en términos de Sesgo Relativo (SR) y Eficiencia Relativa (ER), donde

$$SR = \frac{1}{B}\sum_{b=1}^{B}\frac{|\widehat{Q}_y^R(\beta)_b - Q_y(\beta)|}{Q_y(\beta)}; ER = \frac{ECM[\widehat{Q}_y^R(\beta)]}{ECM[\widehat{Q}_{yn}(\beta)]},$$

Tabla 3.11: Combinaciones de diseños muestrales usados en muestreo con dos ocasiones sucesivas y probabilidades desiguales.

Acrónimo	Muestra	Tipo de muestreo	
SMS	s'	M. aleatorio simple	(**S**)
	s_m	Método de Midzuno	(**M**)
	s_u	M. aleatorio simple	(**S**)
MSS	s'	Método de Midzuno	(**M**)
	s_m	M. aleatorio simple	(**S**)
	s_u	M. aleatorio simple	(**S**)
MMM	s'	Método de Midzuno	(**M**)
	s_m	Método de Midzuno	(**M**)
	s_u	Método de Midzuno	(**M**)

siendo b la b-ésima simulación,

$$\widehat{Q}_y^R(\beta) = w\widehat{Q}_{ym}^r(\beta) + (1-w)\widehat{Q}_{yu}(\beta),$$

$ECM[\widehat{Q}_y^R(\beta)] = B^{-1}\sum_{b=1}^{B}[\widehat{Q}_y^R(\beta)_b - Q_y(\beta)]^2$, y $ECM[\widehat{Q}_{yn}(\beta)]$ se define análogamente para $\widehat{Q}_{yn}(\beta)$, el estimador estándar para el cuantil poblacional basado en la ocasión más reciente.

Notamos que el valor óptimo para la constante w (3.35) depende de varianzas y covarianzas desconocidas, en concreto depende de $V(\widehat{Q}_{ym}^r(\beta))$, $V(\widehat{Q}_{yu}(\beta))$ y $Cov(\widehat{Q}_{yu}(\beta), \widehat{Q}_{ym}^r(\beta))$. Se usarán técnicas Jackknife (Efron y Tibshirani, 1993) para la estimación de estas expresiones.

Por otro lado, la constante w depende de covarianzas porque la muestra solapada y la no solapada son dependientes, aunque algunos autores ignoran este hecho y consideran tales muestras como independientes, es decir, emplearían la constante

$$w* = \frac{V(\widehat{Q}_{yu}(\beta))}{V(\widehat{Q}_{ym}^r(\beta)) + V(\widehat{Q}_{yu}(\beta))},$$

donde $Cov(\widehat{Q}_{yu}(\beta), \widehat{Q}_{ym}^r(\beta))$ estaría omitida. Con el fin de analizar este hecho en la práctica, el estimador propuesto basado en la constante $w*$ (asumiendo que existe independencia entre las muestras, por lo que se ignoran las covarianzas) ha sido incluido en el estudio de simulación.

En primer lugar analizaremos la eficiencia de los estimadores, la cual puede observarse en las Figuras B.24, B.25 y B.26, en donde se representa la Eficiencia Relativa de los distintos estimadores y combinaciones de diseños y tamaños muestrales. La variación en el cumplimiento de los estimadores desde distintas perspectivas puede por tanto observarse. Notamos que las curvas continuas corresponden al estimador propuesto (usando covarianzas), mientras que las curvas discontinuas corresponden al estimador compuesto que no emplea covarianzas. Las líneas horizontales representan al estimador estándar.

En los tres casos, los resultados obtenidos muestran un buen cumplimiento del estimador propuesto, el cual es siempre más eficiente que el estimador estándar, excepto para el caso de fracciones de solapamiento elevadas. Cuando la fracción de solapamiento aumenta, decrece la

Eficiencia Relativa para el estimador propuesto en comparación con el estimador estándar.

En lo que respecta al comportamiento del uso o no de covarianzas en el estimador propuesto, puede comprobarse que se obtiene una ligera mejoría en eficiencia cuando se tiene en cuenta las covarianzas en la construcción del estimador, teniendo por tanto sentido la hipótesis de dependencia entre el estimador de la muestra no solapada y el estimador propuesto para la parte solapada. Puede además observarse que la ganancia en precisión sobre el estimador que omite las covarianzas es mayor a medida que aumenta el tamaño muestral de la ocasión más reciente. En resumen, estos resultados recomiendan el uso de covarianzas en el estimador propuesto para la estimación de cuantiles bajo un muestreo con dos ocasiones sucesivas y probabilidades desiguales.

El análisis del Sesgo Relativo de los distintos estimadores puede seguirse en las Figuras B.27, B.28 y B.29.

A partir de estas figuras puede observase un similar comportamiento de los estimadores al obtenido en el estudio de la Eficiencia Relativa. Los valores del Sesgo Relativo para los estimadores propuestos están siempre por debajo de 0.2, y en algunas ocasiones son inferiores a 0.1, mientras que el Sesgo Relativo para el estimador estándar es bastante mayor llegando incluso a 0.6.

Por último, analizaremos los valores observados de los estimadores mediante diagramas de cajas con bigotes. Por brevedad, se ha considerado el diseño muestral SMS y los tamaños muestrales $n' = 75$, $n = 50$ y $m = 5, 10, 15, 20$. La Figura B.30 nos da tal información para los tres cuartiles. También en este estudio se comprueba que el estimador propuesto presenta el mejor comportamiento, al obtenerse estimadores menos dispersos en comparación con el estimador estándar y el estimador que omite las covarianzas.

Notamos que se han realizado otras simulaciones con distintos tamaños muestrales a los usados en los estudios anteriores. En todos los casos los resultados confirman el buen comportamiento del estimador propuesto frente a sus competidores. También se ha observado que la ganancia en precisión del estimador propuesto es mejor a medida que el tamaño muestral en la primera ocasión aumenta con respecto al tamaño de la segunda ocasión. Por otro lado, cuando el tamaño muestral en la primera ocasión es menor que el tamaño en la segunda, se obtiene una menor ganancia en precisión, y esta ganancia disminuye a medida que aumenta la diferencia entre tamaños muestrales. Este resultado es lógico porque si n' es mayor en comparación con n, la primera muestra proporcionará mayor información, y el estimador de tipo razón basado en la muestra solapada presentará un menor grado de error, por lo que es de esperar que el estimador propuesto mejore también en precisión. Con el fin de obtener más información sobre la estimación de cuantiles en muestreo con dos ocasiones sucesivas y diseños muestrales arbitrarios, puede también consultarse Rueda y Muñoz (2006b).

3.4. Estimadores bajo el método de verosimilitud empírica

En este apartado se utiliza el método de verosimilitud empírica para la estimación de cuantiles. Para ello, usaremos el estimator de verosimilitud empírica para la función de distribución definido en la Sección 2.4.3. Tomando la inversa de este estimador, podremos obtener estimadores de cuantiles fácilmente. Estos estimadores también se utilizarán para el análisis de algunas medidas de pobreza.

Bajo datos de la Encuesta Continua de Presupuestos Familiares para el primer trimestre del año 1997, mostraremos como tanto el estimador propuesto para los cuantiles como el método bootstrap para la estimación de la varianza, exhiben un buen comportamiento en comparación con otros estimadores alternativos.

3.4.1. Antecedentes

Asumiendo el método de verosimilitud empírica, los únicos estimadores conocidos para cuantiles en la literatura se basan en la aproximación modelo-calibrada, es decir, se usan los estimadores modelo-calibrados para la función de distribución descritos en la Sección 2.4.2. Sea $\widehat{F}_{MCPE}(t)$ uno de estos estimadores cuando se usa el punto $t_0 = \widehat{Q}_{HKy}(\beta)$. Notamos que $\widehat{F}_{MCPE}(t)$ será más eficiente que $\widehat{F}_{HKy}(t)$ para t en las cercanías de $Q_y(\beta)$.

El cuantil $Q_y(\beta)$ puede estimarse mediante inversión directa de $\widehat{F}_{MCPE}(t)$, esto es, $\widehat{Q}_{MCPE}(\beta) = \widehat{F}^{-1}_{MCPE}(\beta)$ para $\beta \in (0, 1)$. Puesto que $\widehat{F}_{MCPE}(t)$ es una verdadera función de distribución, esta inversión es computacionalmente simple.

Notamos que tanto este estimador como su correspondiente varianza asintótica serán usadas en la Sección 3.4.5 para su comparación empírica con el estimador propuesto bajo el método de verosimilitud empírica. Por esta razón, a continuación se resumen las principales propiedades asintóticas de este estimador. Para ello, asumimos que hay una sucesión de poblaciones finitas $\{U_\nu, \nu = 1, 2, \ldots\}$. $F_\nu(t)$ y $Q_\nu(\beta)$ denotan respectivamente $F_y(t)$ y $Q_y(\beta)$, para la población U_ν. Además, sean los diseños muestrales siguientes:

(i) Muestreo aleatorio simple con o sin reemplazamiento.

(ii) Muestreo estratificado aleatorio simple con o sin reemplazamiento.

(iii) Muestreo con probabilidades desiguales de una etapa con reemplazamiento.

(iv) Muestreo de varias etapas con reemplazamiento en la primera etapa.

Notamos que en el caso de diseños con reemplazamiento se usa el estimador de tipo Hansen-Hurwitz (Hansen y Hurwitz, 1943), esto es $\pi_i = nq_i$, donde q_i es la probabilidad de seleccionar la i-ésima unidad.

Una representación de Bahadur para el cuantil $\widehat{Q}_{MCPE}(\beta)$ puede establecerse para estos diseños muestrales. Sean también las condiciones (C2.20), (C2.21) y (C2.22) dadas en la Sección 2.4.2 junto a las siguientes:

(C3.7). Existe una función de distribución $F(t)$ diferenciable de orden 2 con función de densidad $f(t)$, tal que $F_\nu(t) - F(t) = o(1)$, y para cualquier $a_\nu = O(n^{-1/2})$

$$\sup_{|\delta|\leq a_\nu} |[F_\nu(t+\delta) - F_\nu(t)] - [F(t+\delta) - F(t)]| =$$

$$= o(n_\nu^{-1/2}),$$

donde el tamaño muestral $n_\nu \to \infty$ cuando $\nu \to \infty$.

(C3.8). Para un valor fijo $\beta \in (0,1)$, $Q_\nu(\beta) \to Q_0(\beta)$, donde $Q_0(\beta)$ es el cuantil β de $F(t)$ y $f(Q_0(\beta)) > 0$.

El siguiente teorema puede establecerse.

Teorema 3.13 *Bajo los diseños muestrales (i)∼(iv) y las condiciones (C2.20), (C2.21), (C2.22), (C3.7) y (C3.8), se tiene que* $\widehat{Q}_{MCPE}(\beta) - Q_y(\beta) =$

$$= \frac{1}{f(Q_y(\beta))}\left(\beta - \widehat{F}_{MCPE}(Q_y(\beta))\right) + o_p(n^{-1/2}),$$

donde $f(\cdot)$ *es la función densidad de la función de distribución límite de* $F_y(t)$ *cuando* $N \to \infty$.

En consecuencia, la varianza asintótica de $\widehat{Q}_{MCPE}(\beta)$ puede aproximarse por

$$V(\widehat{Q}_{MCPE}(\beta)) \simeq \frac{1}{f(Q_y(\beta))^2} V(\widehat{F}_{MCPE}(Q_y(\beta))) =$$

$$\frac{1}{f(Q_y(\beta))^2}\frac{1}{N^2}\sum_{i<j}\sum_{j=1}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{U_i}{\pi_i} - \frac{U_j}{\pi_j}\right)^2 + o(n^{-1}),$$

donde $U_i = \delta(Q_y(\beta) - y_i) - F_y(Q_y(\beta)) - (w_i^* - \overline{w}^*)B_N$ y $\overline{w}^* = N^{-1}\sum_{i=1}^N w_i^*$. w_i^* viene dada por (2.85), (2.87), (2.90) o (2.93) cuando $t_0 = Q_y(\beta)$.

Esta varianza puede estimarse mediante

$$\widehat{V}(\widehat{Q}_{MCPE}(\beta)) \simeq \frac{1}{f(Q_y(\beta))^2} V(\widehat{F}_{MCPE}(\widehat{Q}_{MCPE}(\beta))) =$$

$$\frac{1}{f(Q_y(\beta))^2}\frac{1}{N^2}\sum_{i<j}\sum_{j=1}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{u_i}{\pi_i} - \frac{u_j}{\pi_j}\right)^2 + o(n^{-1}),$$

donde $u_i = \delta(\widehat{Q}_{MCPE}(\beta) - y_i) - \beta - (w_i - \overline{w})B_N$ y $\overline{w} = N^{-1}\sum_{i=1}^N w_i$. w_i viene dada por (2.86), (2.88), (2.91) o (2.92) cuando $t_0 = \widehat{Q}_{HKy}(\beta)$. $f(Q_y(\beta))$ puede estimarse mediante procedimientos estándares (Silverman, 1986).

La ganancia en eficiencia al usar $\widehat{Q}_{MCPE}(\beta)$ sobre $\widehat{Q}_{HKy}(\beta)$ es comparable a la ganancia de $\widehat{F}_{MCPE}(t)$ sobre $\widehat{F}_{HKy}(t)$. Con la óptima elección $w_i = E_\xi(z_i|\mathbf{x}_i)$, la ganancia máxima de la eficiencia asintótica está garantizada. Así, este método puede aplicarse en diseños muestrales complejos y para un vector multivariante de variables auxiliares.

3.4.2. Aplicación a la estimación de líneas de pobreza

El análisis de las líneas de pobreza es un tema reciente y de gran interés en la sociedad. La proporción oficial de pobreza y el número de personas en pobreza son importantes medidas para el bienestar económico de un país.

El análisis de la estructura de los ingresos y la desigualdad de ingresos son los principales objetivos en los estudios de pobreza. Esto se debe a que la desigualdad de los ingresos puede afectar a la eficiencia del mercado laboral, y a que esto conlleva a una serie de problemas relacionados con la igualdad social, tal como la incidencia de la pobreza o la estratificación social.

La aplicación de una medida de pobreza requiere la especificación de una línea de pobreza, la cual separe a la población en pobres y no pobres. En la literatura, existen distintas formas de especificar una línea de pobreza. Por ejemplo, La Organización para la Cooperación Económica y el Desarrollo (OECD, acrónimo de *Organization for Economic Cooperation and Development*) en el año 1997, definió la línea de bajos ingresos como dos tercios del salario mediano, de modo que un empleado se consideraba que tenía ingresos bajos si recibía un salario inferior al anterior umbral señalado. Sin embargo, Eurostat (2000) define que un empleado en la Unión Europea percibe un salario bajo si su salario mensual es inferior al 60 % del salario mediano de su correspondiente país.

Los empleados con bajos ingresos, en particular, ha sido un centro de investigación con alto interés político (Lucifora y Salverda, 1998). Por un lado, a un nivel macroeconómico, los empleados con bajos ingresos es claramente relevante para la igualdad social, como lo demuestran las razones con alta pobreza en los países donde los empleados con bajos ingresos es relativamente alto (OECD, 1997). Por otro lado, desde una perspectiva microeconómica, existe una relación entre salarios bajos y estado de pobreza de los hogares (OECD,1997, Eurostat, 2000).

En la literatura, existen tres tipos de métodos para determinar las líneas de pobreza: los métodos absolutos, relativos y los subjetivos. Los métodos absolutos obtienen la línea de pobreza como una cantidad mínima de fuentes en un punto del tiempo y ponen al día la línea solamente para cambios de precio sobre el tiempo. La línea de pobreza usada por el estadístico oficial de pobreza de Estados Unidos es un ejemplo de línea de pobreza absoluta. El método relativo especifica la línea de pobreza como un punto en la distribución de ingresos o gastos y, por lo tanto, la línea puede estar sin fecha automáticamente sobre el tiempo para cambios en niveles de vida. En la práctica, los investigadores a menudo especifican la línea de pobreza relativa como un porcentaje del ingreso o gasto medio (Wolfson y Evans, 1989, Johnson y Webb, 1992), como un porcentaje del ingreso o gasto mediano (Smeeding, 1991, Eurostat, 2000) o simplemente como un cuantil (OECD, 1982). El método subjetivo deriva de la línea de pobreza basada en la opinión pública. Comparada con las dos primeras aproximaciones, el método subjetivo es relativamente menos popular y raramente se usa.

Mientras que las líneas de pobreza absolutas han sido usadas en la mayoría de los estadísticos de pobreza de los gobiernos, las líneas de pobreza relativas han ganado recientemente en popularidad y uso tanto en las comparaciones internacionales de pobreza como en análisis nacionales de pobreza a través del tiempo. Preston (1995) estableció las distribuciones muestrales de los estadísti-

cos de pobreza relativos.

La desigualdad entre salarios es requerida a menudo en estudios de pobreza o distribución de la riqueza. Tradicionalmente, La oficina censal de Estados Unidos ha empleado un determinado número de percentiles límite y razones para estudiar cambios en la desigualdad de salarios de los hogares. Entre ellos encontramos la razón de ingresos para un determinado hogar entre el percentil 95 y el percentil 20, el percentil 95 con respecto a la mediana, etc. Derivadas de estos percentiles son también bastantes usados en la literatura de ingresos. Algunos investigadores han propuesto otras medidas alternativas como la razón entre los percentiles 90 y 10 o la razón entre los percentiles de orden 50 y 10. Eurostat (2000) también emplea el salario mediano con respecto al primer decil. Estos valores dan una idea de la extensión de las desigualdades entre salarios. Por ejemplo, la razón entre los percentiles de orden 50 y 10 nos permite ver si la incidencia de empleos con bajos ingresos está fuertemente relacionada con la dispersión de salarios en la cola izquierda de la distribución. En Binder y Kovačevic (1995), Dickens y Manning (2004) pueden consultarse otras medidas desigualdad de ingresos.

La atención dada a este tipo de estadísticos en los medios de comunicación y en los círculos de política es considerable, hasta el punto de que importantes decisiones políticas pueden verse influenciadas por estas medidas.

La característica común de estas medidas es su complejidad. Éstas son funciones no lineales de las observaciones y un alto número de éstas dependen de cuantiles. Como se ha comentado, la literatura relacionada a la estimación de medianas y otros cuantiles, los cuales usan una variable auxiliar, es considerablemente menos extenso que en el caso de medias y totales, y las técnicas habituales, tal como el método de regresión, no tienen una extensión obvia a la estimación de cuantiles. Por tanto, la mayoría de los estudios relacionados con cuantiles han sido desarrollados asumiendo muestreo aleatorio simple o muestreo estratificado (Gross, 1980, Sedransk y Meyer, 1978, Sedransk y Smith, 1988, Kuk y Mak, 1989, Singh *et al.*, 2001), o bien considerando aproximaciones basadas en el modelo (Chambers y Dunstan, 1986, Dorfman y Hall, 1993, Mak y Kuk, 1993), las cuales asumen un modelo de superpoblación, los estimadores son dependientes de dicho modelos y puede llegarse a obtener un pobre cumplimiento de los estimadores bajo una inapropiada especificación del modelo. En la práctica, estas situaciones no son usuales, especialmente para el caso de datos relacionados con ingresos o gastos, los cuales se analizan asumiendo diseños muestrales complejos con probabilidades desiguales y cuyos datos, además, exhiben una alta asimetría, lo que hace muy difícil asociar un modelo de superpoblación a los datos en estudio. El uso de estimadores de cuantiles eficientes basados en información auxiliar y aproximaciones independientes del modelo, puede ayudarnos a obtener una mejoría en la estimación de medidas de pobreza. Notamos que la mayoría de los estudios relacionados con medidas de pobreza han sido llevados a cabo usando estimadores clásicos de la literatura del muestreo en poblaciones finitas.

El propósito de esta sección es desarrollar un estimador de cuantiles que pueda aplicarse a diferentes medidas de pobreza. Para ello, usaremos la aproximación modelo-asistida y el método de verosimilitud empírica para construir nuevos estimadores para un determinado cuantil. En lo que respecta a la estimación de cuantiles usando el método de verosimilitud empírica (véase la Sección 3.4.1), Chen y Wu (2002) propusieron estimadores modelo-calibrados (Wu y Sitter, 2001). Estos estimadores requieren el uso de un modelo de superpoblación apropiado, y son por tanto dependientes de dicho modelo. Además, estos estimadores se construyen por medio de restricciones que requieren el uso de un único valor fijado. Una importante pérdida de eficiencia puede llegar a obtenerse cuando dicho valor fijado se encuentra alejado del cuantil que va a ser estimado.

El estimador propuesto usa de modo efectivo la información auxiliar en la etapa de estimación porque éste está basado en tres valores fijados construidos a partir de la información auxiliar. Estos valores se encuentran bien repartidos dentro de la distribución de datos, resolviendo de este modo la pérdida de eficiencia provocada por la elección de un valor fijado situado a gran distancia de cuantil que se va a estimar. Este estimador propuesto está basado en el estimador para la función de distribución descrito en la Sección 2.4.3.

Debido a la naturaleza específica de los cuantiles y a la complejidad de algunas medidas de pobreza, las varianzas de estos estadísticos complejos no pueden expresarse por simples formulas. Mostraremos como la técnica bootstrap es una posible alternativa en la estimación de la varianza del estimador propuesto.

3.4.3. Estimadores propuestos modelo-asistidos

En este epígrafe se describe el estimador propuesto usando la metodología de verosimilitud empírica. Como se ha comentado, usaremos una perspectiva modelo-asistida debido a que esta proporciona un enfoque en el cual se pueden desarrollar estimadores eficientemente. Para ello, necesitaremos un modelo de superpoblación que describa la relación entre la variable de interés y las variables auxiliares. Este modelo será posteriormente usado para construir estimadores basados en el diseño.

Como resulta habitual, consideraremos el modelo regresión lineal dado por

$$y_i = \beta^t \mathbf{x}_i + v_i \varepsilon_i, \quad i = 1, \ldots, N \tag{3.45}$$

donde v_i es una función conocida de x_i y las cantidades ε_i son variables aleatorias independientes e idénticamente distribuidas con media 0 y varianza σ^2. Notamos que en la práctica los valores del vector β son desconocidos, aunque es sabido que este parámetro puede estimarse eficientemente por mínimos cuadrados (véase por ejemplo Särndal *et al.*, 1992) como

$$\mathbf{B} = \left(\sum_{i \in U} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\sigma^2} \right)^{-1} \cdot \sum_{i \in U} \frac{\mathbf{x}_i y_i}{\sigma^2}. \tag{3.46}$$

Este estimador es óptimo en el sentido de ser el mejor estimador lineal e insesgado para β bajo el modelo (3.45). A

su vez, **B** es una característica poblacional finita, aunque puede estimarse usando los datos muestrales. Esta estimación viene dada por

$$\widehat{\beta} = \left(\sum_{i \in s} \frac{d_i \mathbf{x}_i \mathbf{x}_i^t}{\sigma^2} \right)^{-1} \cdot \sum_{i \in s} \frac{d_i \mathbf{x}_i y_i}{\sigma^2}. \tag{3.47}$$

Como ya sabemos, el método de verosimilitud empírica presenta buenas propiedades asintóticas y empíricas para el problema de la estimación de medias o totales (Chen y Qin, 1993, Chen y Sitter, 1999), funciones de distribución (Chen y Wu, 2002), estimación en presencia de datos faltantes (Rueda, Muñoz, Berger, Arcos y Martínez, 2006, Leung y Qin, 2006), etc. Chen y Wu (2002) propusieron estimadores de verosimilitud empírica modelo-calibrados que requieren el uso de un único valor prefijado. La aplicación de estos estimadores a la estimación de cuantiles resulta posible, aunque este proceso arrastra una importante pérdida de eficiencia cuando dicho valor prefijado está alejado de cuantil que va a ser estimado. Con el propósito de reducir esta pérdida en eficiencia, se proponen estimadores modelo-asistidos para cuantiles usando el método de verosimilitud empírica y tres valores prefijados que ayudarán a reducir tal pérdida de eficiencia.

Asumiendo el método de verosimilitud empírica (Chen y Sitter, 1999), el estimador propuesto para el cuantil β está dado por

$$\widehat{Q}_{MA}(\beta) = \text{ínf}\{t : \widehat{F}_{MA}(t) \geq \beta\}, \tag{3.48}$$

donde

$$\widehat{F}_{MA}(t) = \sum_{i \in s} \widehat{p}_i \delta(t - y_i), \tag{3.49}$$

y las cantidades $\widehat{p}_i$ son las soluciones al problema de maximización de la función de verosimilitud pseudo empírica $\widehat{l}(\mathbf{p}) = \sum_{i \in s} d_i \log(p_i)$ sujeta a

$$\sum_{i \in s} p_i = 1, \quad (p_i > 0), \tag{3.50}$$

$$\sum_{i \in s} p_i \delta(t_{g25} - g_i) = \frac{1}{N} \sum_{k=1}^{N} \delta(t_{g25} - g_k) = F_g(t_{g25}) = 0{,}25, \tag{3.51}$$

$$\sum_{i \in s} p_i \delta(t_{g50} - g_i) = \frac{1}{N} \sum_{k=1}^{N} \delta(t_{g50} - g_k) = F_g(t_{g50}) = 0{,}5, \tag{3.52}$$

$$\sum_{i \in s} p_i \delta(t_{g75} - g_i) = \frac{1}{N} \sum_{k=1}^{N} \delta(t_{g75} - g_k) = F_g(t_{g75}) = 0{,}75, \tag{3.53}$$

donde $t_{g25} = Q_g(0{,}25)$, $t_{g50} = Q_g(0{,}50)$, $t_{g75} = Q_g(0{,}75)$, y $Q_g(\alpha)$ es el cuantil α para la variable $g_i = \widehat{\beta}^t \mathbf{x}_i$.

Notamos que la idea de usar $\delta(t - g_i)$ para cualquier t como una variable de calibración para formar restricciones como las dadas en (3.51), (3.52) y (3.53) fue en primer lugar discutida en Wu y Sitter (2001) y posteriormente elaborada en Chen y Wu (2002). Por otro lado, la elección de los valores t_{g25}, t_{g50} y t_{g75} en (3.51), (3.52) y (3.53) ha sido discutida en la Sección 2.4.

Una vez que se ha definido el estimador de cuantiles, las medidas de pobreza que dependan de tales parámetros podrán ser estimadas. Por ejemplo, la línea de bajos ingresos puede definirse como la fracción α de un cuantil β (Eurostat, 2000, Blackburn, 1990, 1994, Smeeding, 1991, etc.):

$$L_{\alpha,\beta} = \alpha Q_y(\beta), \tag{3.54}$$

y las medidas para cuantificar la desigualdad de ingresos están dadas por la razón entre los cuantiles de órdenes β_1 y β_2 (Eurostat, 2000, U.S. Census Bureau, etc):

$$r_{\beta_1,\beta_2} = Q_y(\beta_1)/Q_y(\beta_2). \tag{3.55}$$

Estas medidas pueden estimarse fácilmente por

$$\widehat{L}_{\alpha,\beta} = \alpha \widehat{Q}_{MA}(\beta), \tag{3.56}$$

para la medida dada en (3.54), y por

$$\widehat{r}_{\beta_1,\beta_2} = \widehat{Q}_{MA}(\beta_1)/\widehat{Q}_{MA}(\beta_2), \tag{3.57}$$

para la medida dada en (3.55).

3.4.4. Propiedades. Estimación de la varianza

El estudio de las propiedades asintóticas del estimador propuesto pasa por analizar tales propiedades para el estimador $\widehat{F}_{MA}(t)$, las cuales se han establecido en la Sección 2.4.4. Queda por tanto describir una expresión para la varianza del estimador propuesto para cuantiles. La determinación de tal expresión es posible, aunque tendría únicamente validez asintótica, es decir, para tamaños muestrales bastantes elevados, situación no siempre presente en la práctica. Por otro lado, por la estructura no lineal del cuantil, se requiere el uso de una aproximación lineal que emplea parámetros poblacionales, por ejemplo densidades, que también tendrían que ser estimados, lo que conlleva a otra pérdida de eficiencia en la etapa de estimación de la varianza.

Si aplicamos el estimador propuesto a la estimación de medidas de pobreza, la determinación de dicha expresión asintótica para la varianza resulta aún más difícil, puesto que la característica común de las medidas de pobreza, como por ejemplo (3.54) y (3.55), es su complejidad. Este hecho puede comprobarse en Shao y Rao (1993), Kovačevik y Binder (1997), Kovačevik y Yung (1997), Zheng, 2001, y Berger y Skinner (2003). Además, los datos de ingresos y gastos provienen usualmente de encuestas complejas (muestreos con probabilidades desiguales de tipo estratificado, con múltiple etapas, por conglomerados, etc), lo que también dificulta la determinación de expresiones asintóticas bajo estas situaciones. La única alternativa en estos casos es el uso de métodos especiales para la estimación de varianzas.

Por estas razones, proponemos el uso de técnicas alternativas para la estimación de la varianza del estimador propuesto. En concreto, se propone la técnica bootstrap que frecuentemente se usa en la estimación de cuantiles, y en particular, para la estimación de las medidas de pobreza. Este hecho queda justificado por los estudios ya llevados a cabo y los cuales resumiremos brevemente a continuación. Puesto que el estudio empírico que llevamos a cabo está basado en algunas medidas de pobreza, centraremos nuestra atención a la estimación de la varianza de medidas de pobreza.

En primer lugar, notamos que en los estudios de pobreza, la variabilidad muestral de las diferentes medidas estimadas presentan un interés particular cuando éstas son comparadas entre países, a través del tiempo o entre subgrupos dentro de un país.

Los métodos tradicionales para aproximar la varianza de un estimador (véase Wolter, 1985), envuelven una de las siguientes estrategias: linealización de Taylor o métodos de replicación tal como bootstrap, jackknife, etc. En los casos donde los estimadores presentan una forma compleja (como en el caso de cuantiles), los métodos de replicación son preferidos por ser más fáciles de implementar, aunque para el caso de cuantiles, el clásico método jackknife da estimadores inconsistentes para la varianza (Kovar *et al.*, 1988, Shao y Wu, 1989). También pueden usarse para la estimación de la varianza otros métodos alternativos tal como linealización y técnicas residuales (Deville, 1999). Una complicación al aplicar el método de linealización en la estimación de cuantiles es que éste requiere la estimación de funciones de densidad de probabilidad para la variable de interés.

Los métodos bootstraps están ganando en popularidad en las investigaciones empíricas. Por ejemplo, en el Instituto Estadístico de Canada se llevó a cabo un estudio de simulación para comparar la eficiencia de varios métodos de remuestreo con respecto al método de estimación de ecuaciones (véase Kovacevic, Yung y Pandher, 1995) en el caso de medidas de desigualdad de ingresos. Para algunos cuantiles, el estimador bootstrap exhibía el menor sesgo relativo, mientras que el método de estimación de ecuaciones junto con el método bootstrap eran los óptimos en el sentido de estabilidad. Estos resultados confirman la ventaja al usar el método bootstrap sobre el resto de aproximaciones. La precisión de las técnicas bootstrap en la estimación de la varianza de cuantiles obtenidos mediante estimadores de tipo razón, diferencia y regresión ha sido discutida en Rueda, Martínez-Miranda y Arcos (2006). Asumiendo también medidas de pobreza, Shao y Chen (1998) también demostraron la consistencia del método bootstrap para la estimación de la varianza. En Bickel y Freedman (1984), Dalgleish (1995), etc, pueden consultarse otros estudios del bootstrap y sus propiedades en muestreo de poblaciones finitas.

3.4.5. Propiedades empíricas

En esta sección se evalúa la precisión del estimador propuesto junto con otros estimadores conocidos. Además, se estudia la eficiencia de estos procedimientos cuando se aplica la estimación de cuantiles a diversas medidas de pobreza. El comportamiento del método bootstrap para la estimación de varianzas será también analizado. Para ello, se calculan las estimaciones bootstrap para los distintos estimadores y comparamos estos resultados con los obtenidos a través de las correspondientes expresiones para la varianza de cada estimador, en aquellos casos que se disponga de tales expresiones. Por simplicidad, se asume muestreo aleatorio simple.

Tabla 3.12: Medidas globales medias de precisión y eficiencia basadas en cuantiles de órdenes $\beta = 0{,}1, 0{,}3, 0{,}5, 0{,}7, 0{,}9$, y muestras de tamaño $n = 500$.

			Varianzas bootstrap				Varianzas asintóticas			
Est.	ERM	SRM	ERM	SRM	CIM	LIM	ERM	SRM	CIM	LIM
MA	0.86	0.25	0.82	14.05	92.9	550.96	–	–	–	–
MA1	0.89	0.23	0.83	12.65	93.2	561.62	–	–	–	–
MCPE	0.92	0.25	0.86	8.72	92.9	563.18	0.78	7.16	93.9	553.87
HK	1.00	0.26	1.00	9.97	92.8	622.32	1.00	9.52	94.0	616.53
r	1.04	0.23	1.08	9.87	93.3	654.58	1.01	3.96	93.2	646.85
d	1.05	0.25	1.06	7.32	92.9	651.83	1.02	3.67	93.3	650.31
dm	0.87	0.21	0.81	12.17	92.7	556.01	0.70	5.27	93.9	548.07
CD	3.58	12.44	0.48	10.24	17.1	436.84	–	–	–	–

Tabla 3.13: Medidas de precisión y eficiencia para la línea de bajos ingresos cuando $\alpha = 0{,}6$, $\beta = 0{,}5$ y se toman muestras de tamaño $n = 500$.

			Varianzas bootstrap				Varianzas asintóticas			
Est.	ER	SR	ER	SR	CI	LI	ER	SR	CI	LI
MA	0.70	-0.10	0.57	16.59	93.8	391.54	–	–	–	–
MA1	0.79	-0.08	0.63	13.03	94.2	410.32	–	–	–	–
MCPE	0.78	-0.11	0.65	14.87	94.0	412.62	0.53	15.81	94.8	423.94
HK	1.00	-0.24	1.00	17.09	93.4	470.88	1.00	18.41	94.2	482.73
r	1.09	-0.00	0.98	7.77	94.6	473.71	0.81	6.97	93.8	481.26
d	1.11	0.01	0.97	6.40	93.8	474.52	0.87	7.45	93.8	486.03
dm	0.74	-0.07	0.49	7.39	93.6	388.18	0.37	8.17	94.8	398.41
CD	1.11	2.23	0.09	0.65	77.2	313.01	–	–	–	–

En este estudio se usa la población ECPF1997 (véase Apéndice A) que está formada por los datos de ingresos y gastos de 3000 familias extraídas de la Encuesta Continua de Presupuestos Familiares del año 1997. Estos datos se han duplicado tres veces para crear una población artificial de $N = 9000$ individuos, a partir de los cuales nos basaremos para llevar a cabo el presente estudio de simulación. Como variable principal se han tomado los ingresos, mientras que como variable auxiliar se consideran los gastos familiares.

El cumplimiento del estimador de cuantiles propuesto y su correspondiente estimación de la varianza obtenida mediante bootstrap se comparará con los estimadores de cuantiles obtenidos a partir de las siguientes funciones de distribución: el clásico estimador de tipo Horvitz-Thompson , $\widehat{F}_{HTy}(t)$, el cual lo usaremos como estimador de comparación para todos los estimadores, los estimadores de tipo razón y diferencia ($\widehat{F}_r(t)$, $\widehat{F}_d(t)$, $\widehat{F}_{dm}(t)$) propuestos en Rao *et al.* (1990), el estimador de Chambers y Dunstan (1986), $\widehat{F}_{CD}(t)$, y $\widehat{F}_{MCPE}(t)$, el estimador propuesto en Chen y Wu (2002). Además, calcularemos el estimador modelo-asistido asumiendo un único valor prefijado. Esto nos permitirá conocer la ganancia en precisión al usar mas de un valor prefijado.

Dado un cuantil de orden β, el comportamiento de todos los estimadores de cuantiles y sus varianzas están medidos por medio del Sesgo Relativo, (SR) y Eficiencia Relativa (ER). Así, para un determinado cuantil, $\widehat{Q}_y(\beta)$, calcularemos

$$\begin{aligned} ER[\widehat{Q}_y(\beta)] &= ECM[\widehat{Q}_y(\beta)]/ECM[\widehat{Q}_{HTy}(\beta)], \\ SR[\widehat{Q}_y(\beta)] &= 100 \times \left(E[\widehat{Q}_y(\beta)] - Q_y(\beta)\right)/Q_y(\beta), \end{aligned} \tag{3.58}$$

y para un estimador de la varianza, $\widehat{V}(\widehat{Q}_y(\beta))$, se obtendrá las medidas dadas por (3.58) después de sustituir $\widehat{Q}_y(\beta)$ y $Q_y(\beta)$ por $\widehat{V}(\widehat{Q}_y(\beta))$ y $V[Q_y(\beta)]$ respectivamente. $E[\cdot]$, $ECM[\cdot]$ y $V[\cdot]$ son las Esperanzas Empíricas, Error Cuadrático Medio y Varianzas basadas en 500 muestras. Notamos que valores de $ER[\widehat{Q}_y(\beta)]$ y $ER[\widehat{V}(\widehat{Q}_y(\beta))]$ menores de 1 indican que $\widehat{Q}_y(\beta)$ y $\widehat{V}(\widehat{Q}_y(\beta))$ son más precisos que $\widehat{Q}_{HTy}(\beta)$ y $\widehat{V}(\widehat{Q}_{HTy}(\beta))$, respectivamente. Asumiendo normalidad, también se ha obtenido la Cobertura de los Intervalos de Confianza (CI) al 95 % y la Longitud Media de cada Intervalo (LI). Todos los estudios se han basado en muestras de tamaño $n = 500$.

Notamos que la precisión de cada estimador depende directamente del cuantil que va a ser estimado. Por ejemplo, el estimador de Chambers y Dunstan es muy eficiente en la estimación de la mediana, aunque generalmente sufre de importantes sesgos en las estimaciones a medida que se estiman cuantiles más alejados de la mediana (véase Rao *et al.*, 1990, Chambers *et al.*, 1993, y Dorfman, 1993). Por este motivo, el primer estudio desarrollado intenta medir la precisión media global de cada estimador a partir de los resultados obtenidos en las estimaciones de los cuantiles de órdenes $\beta = 0{,}1, 0{,}3, 0{,}5, 0{,}7, 0{,}9$. Las medidas usadas para realizar tal medición son el Sesgo Relativo Medio (SRM), dado por

$$SRM = \frac{1}{5}\sum_{i=1}^{5} |SR[\widehat{Q}_y(\beta_i)]|,$$

la raíz cuadrada del valor medio de las medidas ER, es decir,

$$ERM = \sqrt{\frac{1}{5}\sum_{i=1}^{5} ER[\widehat{Q}_y(\beta_i)]},$$

y por último, los valores medios para las medidas CI y LI. Dichas medidas se denotarán como CIM y LIM respectivamente. En la Tabla 3.12 puede observarse las distintas medidas globales para todos los estimadores. A partir de la eficiencia relativa media, podemos comprobar que el estimador propuesto presenta el mejor comportamiento, seguido del estimador de diferencia óptimo (dm). El estimador de Chambers y Dunstan es el menos eficiente, mientras que los estimadores de tipo razón y diferencia también funcionan peor que el estimador estándar. En el estudio de las varianzas observamos que las expresiones asintóticas funcionan ligeramente mejor que la técnica bootstrap, por lo que a tenor de los resultados sería aceptable recurrir a tal procedimiento para la estimación de la varianza. Por último, al estimar todas las varianzas de los estimadores mediante bootstrap, se observa que el estimador propuesto presenta el mejor comportamiento, al estimar los intervalos de confianza con menor longitud y una cobertura similar al resto de estimadores.

El siguiente paso en esta sección es el análisis de la eficiencia del estimador propuesto cuando se aplica a la estimación de medidas de pobreza. En primer lugar analizamos los resultados obtenidos para la estimación de las líneas de bajos ingresos (Tabla 3.13) y a continuación describiremos las conclusiones más importantes en la estimación de razones entre cuantiles para el análisis de la desigualdad entre ingresos (Tablas 3.14 y 3.15).

En primer lugar, notamos que al tratarse de medidas relativas, los resultados obtenidos para las líneas de bajos ingresos en la Tabla 3.13 serán los mismos si se usaran otros valores de α, o bien si se considera la propia mediana. Por tanto, las conclusiones que puedan extraerse de esta tabla se podrían hacer para estos casos comentados.

En la Tabla 3.13 observamos que el estimador propuesto es el más eficiente en términos de eficiencia relativa. Todos los sesgos relativos se encuentran dentro de un rango razonable, excepto el de Chambers y Dunstan con un valor superior al resto, en torno al 2.23 %. Un aspecto importante a tener en cuenta en la estimación de la varianza es que las estimaciones bootstrap son, en términos generales, más precisas que las obtenidas mediante las expresiones asintóticas, puesto que se obtienen para cada estimador sesgos más reducidos, e intervalos de confianza menos amplios con idénticas coberturas. Este resultado nos confirma que la técnica bootstrap es un procedimiento óptimo en la estimación de la varianza de la mediana, y en particular, la estimación de la varianza de las líneas de bajos ingresos. Observando las estimaciones bootstrap podemos comprobar que el estimador diferencia óptimo y el estimador propuesto obtiene las mejores estimaciones para la varianza.

Las Tablas 3.14 y 3.15 nos dan las distintas medidas de precisión y eficiencia para medidas de pobreza dadas por razones de cuantiles. De nuevo, el estimador propuesto se muestra más eficiente en términos de eficiencia relativa. Conclusiones similares pueden derivarse de los re-

Tabla 3.14: Medidas de precisión y eficiencia para la razón de cuantiles cuando $\beta_1 = 0{,}5$, $\beta_2 = 0{,}25$, y se toman muestras de tamaño $n = 500$.

			Varianzas bootstrap			
Est.	ER	SR	ER	SR	CI	LI
MA	0.93	0.05	0.92	18.18	93.6	0.18
MA1	1.04	0.14	1.07	17.75	95.2	0.19
MCPE	1.00	-0.01	1.01	14.68	93.8	0.19
HK	1.00	0.05	1.00	15.91	95.2	0.19
r	1.62	0.34	2.53	14.78	94.4	0.24
d	1.65	0.29	2.16	11.45	94.2	0.23
dm	0.90	0.06	0.80	15.69	93.8	0.18
CD	21.07	14.10	0.05	23.43	0.0	0.08

Tabla 3.15: Medidas de precisión y eficiencia para la razón de cuantiles cuando $\beta_1 = 0{,}95$, $\beta_2 = 0{,}2$, y se toman muestras de tamaño $n = 500$.

			Varianzas bootstrap			
Est.	ER	SR	ER	SR	CI	LI
MA	0.93	0.56	1.01	-0.70	91.4	0.92
MA1	14.66	1.70	–	-82.28	91.4	1.06
MCPE	1.02	0.61	1.07	-3.21	91.6	0.96
HK	1.00	0.27	1.00	-3.04	91.4	0.95
r	1.40	0.95	2.15	0.30	92.6	1.14
d	1.38	0.72	2.01	-3.69	91.4	1.11
dm	1.03	0.61	1.12	-6.12	90.8	0.95
CD	46.52	43.58	–	–	2.4	1.33

sultados obtenidos en la etapa de la estimación de la varianza mediante bootstrap. El estimador de Chambers y Dunstan ofrece el peor comportamiento con importantes sobreestimaciones en la estimación de las razones. Esto se debe a que se están estimando cuantiles alejados de la mediana.

4. Discusión

En este capítulo se hace una discusión conjunta de los resultados obtenidos en todos los capítulos anteriores, resumiendo las principales conclusiones.

4.1. Conclusiones y valoración de resultados

El presente trabajo se divide en dos grandes bloques: estimación bajo el método de verosimilitud empírica (Capítulo 2) y la estimación de cuantiles (Capítulo 3). En estos dos capítulos se han planteado nuevos estimadores en situaciones reales del muestreo en poblaciones finitas.

Así, asumiendo el método de verosimilitud empírica se han propuesto estimadores en presencia de datos faltantes, situación muy usual en la práctica y que no se tiene en cuenta en la mayoría de las investigaciones por muestreo. Las aportaciones hechas en este sentido dan una alternativa para la solución de este problema, puesto que se ha comprobado que puede existir una importante ganancia en eficiencia en las estimaciones de los parámetros desconocidos.

En concreto, se ha usado el método de verosimilitud empírica para estimar una media poblacional cuando en la encuesta nos encontramos con información faltante tanto en la variable de estudio como en la variable auxiliar. Se ha asumido que la muestra puede ser seleccionada mediante un diseño muestral arbitrario, con probabilidades iguales o desiguales.

El estimador propuesto se basa en una clase de estimadores formada por un estimador de verosimilitud empírica y por un estimador de tipo Hájek. Se han derivado las propiedades asintóticas de estos estimadores y el estimador óptimo dentro de la clase propuesta en el sentido de minimizar la varianza asintótica.

El estimador propuesto se ha comparado con otros estimadores en un estudio de simulación, donde se ha comprobado que el estimador óptimo presenta el mejor comportamiento con respecto a sus competidores. La mayor ganancia en eficiencia se presenta cuando el número de valores perdidos es relativamente elevado y la relación lineal entre la variable principal y la auxiliar es débil.

Asumiendo el método de verosimilitud empírica también se han propuesto estimadores modelo-asistidos para la función de distribución. El estimador propuesto posee un importante número de propiedades deseables. Por ejemplo:

- Puede aplicarse fácilmente a diseños muestrales con probabilidades desiguales.
- No es dependiente de un modelo de superpoblación como le ocurre por ejemplo a los estimadores basados en modelos o a los estimadores modelo-calibrados.
- Se establecen las condiciones para la existencia del estimador.
- Bajo ciertas condiciones, el estimador es una verdadera función de distribución. Notamos que esta propiedad no se satisface para un gran número de estimadores en la literatura.
- Se satisfacen también otras propiedades importantes como la insesgadez asintótica, normalidad asintótica, disponibilidad de un estimador de la varianza, etc.

La precisión del estimador propuesto se ha comparado mediante varias medidas con otros estimadores conocidos. Estos estudios han mostrado un comportamiento óptimo por parte del estimador propuesto modelo-asistido. También se ha visto que el estimador de Chambers y Dunstan puede llegar a ser muy eficiente cuando el modelo en el que se basa es apropiado, aunque como se discutió en Rao *et al.* (1990), Chambers *et al.* (1993) y Dorfman (1993), este estimador cumple pobremente cuando se tiene una mala especificación del modelo. Un comentario similar puede hacerse sobre el estimador de verosimilitud empírica modelo-calibrado. Este estimador también sufre una importante pérdida de eficiencia cuando se considera un valor fijado alejado del punto donde va a ser estimada la función de distribución.

Otra propiedad importante que caracteriza al estimador propuesto es el uso eficiente que se hace de la información auxiliar: por un lado porque pueden usarse múltiples variables auxiliares en la etapa de estimación, y por otro porque se usan un conjunto de valores prefijados que poseen una buena distribución y ayudan a mejorar la estimación de la función de distribución, especialmente en las proximidades de algunos de estos puntos. Recordamos también que el hecho de considerar $\mathbf{t}_g$ y $\mathbf{x}$ como valores fijados hacen que los pesos $\widehat{p}_i$ sean independientes de t y puedan establecerse mejores propiedades para el estimador propuesto.

En conclusión, el método de verosimilitud empírica modelo-asistido es una aproximación práctica y simple que incorpora fácilmente información auxiliar en la estimación de la función de distribución. Este estimador presenta un buen cumplimiento y puede ser una alternativa válida a otros estimadores de la función de distribución.

El estudio de la estimación de cuantiles se ha llevado a cabo en el Capítulo 3. Los aportes a la teoría de la estimación de cuantiles se han centrado en tres aspectos:

estimación en muestreo bifásico, estimación en muestreo con dos ocasiones sucesivas y estimación usando el comentado método de verosimilitud empírica.

La mayoría de los procedimientos de muestreo que usan información auxiliar se basan en estimadores que requieren el uso de variables conocidas a nivel poblacional, siendo este hecho poco frecuente en la práctica. Una solución a este problema se presenta con la aplicación de un muestreo bifásico. Por tanto, el problema de la estimación de cuantiles basados en información auxiliar queda resuelto con los estimadores propuestos en este sentido. Con el fin de obtener unas estimaciones más precisas en poblaciones heterogéneas, con una posible distribución en grupos homogéneos, también se han propuesto estimadores para cuantiles en muestreo bifásico y usando un muestreo estratificado en la muestra de la primera fase.

Asumiendo muestreo bifásico bajo cualquier método de extracción de unidades en cada una de las dos fases, se han propuesto estimadores de tipo razón y exponencial. Se ha demostrado la insesgadez de estos estimadores y se han proporcionado expresiones para sus varianzas. Estos resultados nos han servido para poder obtener un estimador óptimo en el estimador de tipo exponencial. Bajo distintos esquemas de muestreo y varios estudios de simulación, se ha comprobado que los estimadores propuestos pueden obtener estimaciones más precisas que el resto de estimadores existentes en la literatura.

Los estimadores propuestos en muestreo bifásico, cuando se usa un muestreo estratificado en la primera fase, están basados en un estimador eficiente para la función de distribución. Se han establecido varias propiedades para este estimador de la función de distribución, por lo que el estimador propuesto para cuantiles posee mejores propiedades. Los resultados teóricos y empíricos que se han llevado a cabo han demostrado que el estimador propuesto puede proporcionar resultados óptimos en este esquema de muestreo.

El muestreo en ocasiones sucesivas es una técnica muy conocida que puede usarse en encuestas continuas para estimar parámetros poblacionales y medidas de diferencia o cambio de una variable de interés. Las encuestas de tipo económico o social llevadas a cabo por la agencias nacionales y otros organismos estadísticos usan este diseño muestral, y la estimación de cuantiles es un problema común en la mayoría de estos estudios. Dentro del muestreo en dos ocasiones sucesivas se han planteado estimadores desde dos perspectivas bastantes usadas dentro del muestreo en poblaciones finitas: asumiendo múltiples variables auxiliares y bajo diseños muestrales probabilísticos con probabilidades desiguales.

Asumiendo múltiples variables auxiliares y muestreo aleatorio simple en cada una de las dos ocasiones, se ha propuesto una clase de estimadores para cuantiles basados en un estimador de tipo razón multivariante y construido a partir de la información obtenida en la parte solapada. Bajo la clase propuesta se ha obtenido la expresión del estimador óptimo en el sentido de mínima varianza asintótica. El estimador propuesto posee un buen número de propiedades deseables, tal como normalidad asintótica, disponibilidad de la varianza del estimador, simplicidad de computación, etc. En los estudios empíricos y teóricos que se han llevado a cabo, el estimador se muestra más preciso que otros estimadores conocidos.

Por otro lado, asumiendo diseños muestrales con probabilidades desiguales en cada ocasión se ha propuesto un estimador compuesto por un estimador de tipo razón (en la porción solapada por ambas muestras) y otro de tipo Hájek (en la parte no solapada de la muestra más reciente). El estimador propuesto es fácil de computar y se ha mostrado bastante preciso en los estudios de simulación. Asumiendo muestreo aleatorio simple en cada una de las dos ocasiones, se ha obtenido la normalidad asintótica del estimador, la cual nos sirve, por ejemplo, para construir intervalos de confianza para los cuantiles.

Por último, se han propuesto estimadores para cuantiles desde una perspectiva modelo-asistida y considerando el método de verosimilitud empírica. La aplicación de estos estimadores a la estimación de algunas medidas de pobreza también ha sido analizada. Se ha propuesto usar la técnica bootstrap para la estimación de la varianza de los estimadores propuestos. La precisión de todos estos procedimientos nuevos ha sido confirmada en estudios de simulación y para el problema de la estimación de cuantiles y medidas de pobreza usadas por numerosos organismos de estadística internacionales y de varios países.

5. Bibliografía

[1] **Adhvaryu, D.** (1978) Successive sampling using multi-auxiliary information. *Sankhya **40**, 167-173*.

[2] **Aitchison, J. y Silvey, S.D.** (1958) Maximum-likelihood estimation of parameter subject to restraints. *Annals of Mathematical Statistics **29**, 813-888*.

[3] **Allen, J., Singh, H.P., Singh, S. y Smarandache, F.** (2002) A general class of estimators of population median using two auxiliary variables in double sampling. INTERSTAT.

[4] **Arcos, A., Rueda, M. y Martínez-Miranda, M.D.** (2005) Using multiparametric auxiliary information at the estimation stage. *Statistical Papers **46**, 339–358*.

[5] [1] **Arcos, A., Rueda, M. y Muñoz, J.F.**(2006) An improved class of estimators of a finite population quantile in sample surveys. *Applied Mathematics Letters*. En prensa.

[6] **Arnab, R. y Okafor, F.C.** (1992) A note on double sampling over two occasions. *Pakistan Journal of Statistics **8**, 9-18*.

[7] **Artés Rodríguez, E.M. y García Luengo A.V.** (2002) *Diseños muestrales en el tiempo.* Monografías, Universidad de Almería.

[8] **Bahadur, R.R** (1966) A note on quantiles in large samples. *Annals of Mathematical Statistics **37**, 577-580*.

[9] **Basu, D.** (1971) *Foundations of statistical inference. A Symposium, eds.* V.P. Godambe and D. A. Sprott, Toronto: Holt Rinehart and Winston.

[10] **Berger, Y.G.** (2004) Variance estimation for measures of change in probability sampling. *The Canadian Journal of Statistics **32**, 451-467*.

[11] [1] **Berger, Y.G., Muñoz, J.F. y Rancourt, E.** (2006) Variance estimation of regression estimators when control total are estimated: an application to the composite estimator. *Survey Methodology*. Aceptado bajo revision.

[12] **Berger, Y.G. y Skinner, C.J.** (2003) Variance estimation for a low income proportion. *Journal of the Royal Statistical Society, Series C **52**, 457-468*.

[13] **Bickel, P.J. y Freedman, D.A.** (1984) Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics **12**, 470-482*.

[14] **Binder, D.A. y Kovačevic** (1995) Estimating some measures of income inequality from survey data: an application of the estimating equation approach. *Survey Methodology **21**, 137-145*.

[15] **Blackburn, M.** (1990) Trends in poverty in the United States, 1967-84. *Review of Income and Wealth **36**, 53-66*.

[16] **Blackburn, M.** (1994) International comparisons of poverty. *American Economic Review **84**, 371-374*.

[17] **Brewer, K.R.W.** (1999) Cosmetic calibration with unequal probability sampling. *Survey Methodology **25**, 205-212*.

[18] **Brewer, K.R.W., Early, L.J. y Joyce, S.F.** (1972) Selecting several samples from a single population. *Australian Journal of Statistics **14**, 231-239*.

[19] **Casell, C.M., Särndal, C.E. y Wretman, J.H.** (1976) Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika **63**, 615-620*.

[20] **Casell, C.M., Särndal, C.E. y Wretman, J.H.** (1977) *Foundations of Inference in Survey Sampling.* New York: Wiley.

[21] **Chambers, R.L., Dorfman, A.H. y Hall, P.** (1992) Properties of estimator of the finite population distribution function. *Biometrika **79**, 577-582*.

[22] **Chambers, R.L., Dorfman, A.H. y Wehrly, T.E.** (1993) Bias robust estimation in finite population using nonparametric calibration. *Journal of the American Statistical Association **88**, 268-277*.

[23] **Chambers, R.L. y Dunstan, R.** (1986) Estimating distribution functions from survey data. *Biometrika **73**, 597-604*.

[24] **Chaudhuri, A. y Vos, J.W.E.** (1988) *Unified theory and strategies of survey sampling.* North-Holland, Amsterdam.

[25] **Chen, H. y Chen, J.** (2000) Bahadur representations of the empirical likelihood quantile processes. *Journal of Nonparametric Statistics **12**, 645-660*.

[26] **Chen, J. y Qin, J.** (1993) Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika **80**, 107-116*.

[27] **Chen, J., Rao, J.N.K. y Sitter, R.R.** (2000) Efficient random imputation for missing data in complex surveys. *Statistica Sinica **10**, 1153-1169*.

[28] **Chen, J. y Sitter, R.R.** (1999) A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica **9**, 385-406*.

[1] Bibliografía correspondiente al doctorando.

[1] Bibliografía correspondiente al doctorando.

[29] **Chen, J., Sitter, R.R. y Wu, C.** (2002) Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika **89**, 230-237*.

[30] **Chen, J. y Wu, C.** (2002) Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica **12**, 1223-1239*.

[31] **Cochran, W.G.** (1977) *Sampling Techniques.* 3rd ed. New York: Wiley

[32] **Cramer, H.** (1946) *Mathematical methods of statistics*. Princenton University Press. Princeton.

[33] **Dalgleish, L. I.** (1995) Software review: Bootstrapping and jackknifing with BOJA. *Statistics and Computing **5**, 165-174*.

[34] **Deng, L.Y. y Wu, C.F.J.** (1987) Estimation of variance of the regression estimator. *Journal of the American Statistical Association **82**, 568-576*.

[35] **Deville, J.C.** (1999) Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology **25**, 193-203*.

[36] **Deville, J.C. y Särndal, C.E.** (1992) Calibration estimators in survey sampling. *Journal of the American Statistical Association **87**, 376-382*.

[37] **Dickens, R. y Manning, A.** (2004) Has the national minimum wage reduced UK wage inequality?. *Journal of the Royal Statistical Society, Series A **167**, 613-626*.

[38] **Dorfman, A.H.** (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *The Australian Journal of Statistics* **35**, 29-41.

[39] **Dorfman, A.H. y Hall, P.** (1993) Estimators of the finite population distribution function using nonparametric regression. *The Annals of Statistics **21** (3), 1452-1475*.

[40] **Eckler, A.R.** (1955) Rotation Sampling. *The Annals of Mathematical Statistics **26** 664-685*.

[41] **Efron, B. y Tibshirani, R.J.** (1993) *An introduction to the Bootstrap*. Chapman & Hall, London.

[42] **Eurostat.** (2000) Low-wage employees in EU countries. Statistics in Focus: Population and Social Conditions. Theme $3-11/2000$. *Office for Official Publications of the EC,* Luxemburgo.

[43] **Fernández García, F.R. y Mayor Gallego, J.A.** (1994) *Muestreo en Poblaciones Finitas: Curso Básico.* P.P.U., Barcelona.

[44] **Fernández Sánchez, M.P., Hernández Bastida, A. y Sánchez González, C.** (2004) Análisis de los ingresos y gastos trimestrales de los hogares españoles usando verosimilitud empírica. *Estudios de Economía Aplicada **22**, 139-150*.

[45] **Francisco, C.A. y Fuller, W.A.** (1991) Quantiles estimation with a complex survey design. *The Annals of Statistics **19**, 454-469*.

[46] **Godambe, V.P.** (1955) A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B **17**, 269-278*.

[47] **Godambe, V.P. y Thompson, M.E.** (1973) Estimation in sampling theory with exchangeable prior distributions. *The Annals of Statistics **1**, 1212-1221*.

[48] **Godambe, V.P. y Thompson, M.E.** (1986) Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review **54**, 127-138*.

[49] **Gordon, L.** (1983) Successive sampling in finite populations. *The Annals of Statistics **11**, 702-706*.

[50] **Gross, S.T.** (1980) Median estimation in sample survey. *Proc. Surv. Res. Meth. Sect. Amer. Statist. Ass. 181-184*.

[51] **Hájek, J.** (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics **35**, 1491-1523*.

[52] **Hall, P.** (1990) Pseudo-likelihood theory for empirical likelihood. *The Annals of Statistics **18**, 121-140*.

[53] **Hall, P. y La Scala, B.** (1990) Methodology and algorithms of empirical likelihood. *International Statistical Review **58**, 109-127*.

[54] **Hansen, M.H. y Hurwitz, W.N.** (1943) On the theory of sampling from finite populations. *Annals of Mathematical Statistics **14**, 333-362*.

[55] **Hanurav, T.V.** (1966) Some aspects of unified sampling theory. *Sankhya, Series A **28**, 175-204*.

[56] **Hartley, H.O. y Rao, J.N.K.** (1968) A new estimation theory for sample surveys. *Biometrika **55**, 547-557*.

[57] **Hedayat, A.S. y Sinha, B.K.** (1991) *Design and Inference in Finite Population Sampling.* John Wiley and Sons.

[58] **Hill, B.M.** (1968) Posterior distribution of percentiles: Bayes theorem for sampling from a population. *Journal of the American Statistical Association **63**, 677-691*.

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

[60] **Huang, E.T. y Fuller,W.A.** (1978) Nonnegative regression estimation for sample survey data. *In Proc. Social Statistics Sec., Am. Statist. Assoc., 300-305* Washington, D.C: American Statistical Association.

[61] **Instituo Nacional de Estadística.** (1992) Encuesta Continua de Presupuestos Familiares. Metodología. *Instituto Nacional de Estadística. Madrid*.

[62] **Isaki, C.T. y Fuller, W.A.** (1982) Survey design under the regression superpopulation model. *Journal of the American Statistical Association **77**, 89-96*.

[63] **Jagers, P.** (1986) Post-stratification against bias in sampling. *International Statistical Review **54**, 159-167*.

[64] **Jessen, R.J.** (1942) Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Statistical Research Bulletin, 304*.

[65] **Jonhson, P. y Webb, S.** (1992) Official statistics on poverty in the United Kingdom. Poverty measurement for economies in transition in eastern european countries. Polish Statistical Association and Polish Central Statistica Office, Warsaw. *Journal of Economics Perspectives **15**, 143-156*.

[66] **Koenker, R. y Hallock, K.F.** (2001) Quantile regression. *Journal of Economics Perspectives **15**, 143-156*.

[67] **Kovačevik, M.S. y Binder, D. A.** (1997) Variance estimation for measures of income inequality and polarization - The estimating equations approach. *Journal of Official Statistics **13**, 41-58*.

[68] **Kovačevik, M.S. y Yung, W.** (1997) Variance estimation for measures of income inequality and polarization - an empirical study. *Survey Methodology **23**, 41-52*.

[69] **Kovačevik, M.S., Yung, W. y Pandher** (1995) Estimating the sampling variances of measures of income inequality and polarization - an empirical study. *Statistic Canada, Methodology Branch Working Paper,* HSMD-95-007E.

[70] **Kovar, J.G., Rao, J.N.K. y Wu, C.F.J.** (1988) Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics **16**, 25-45*.

[71] **Kuk, A.Y.C.** (1993) A kernel method for estimating finite population distribution functions using auxiliary information. *Biometrika **80**, 385-392*.

[72] **Kuk, A.Y.C. y Mak, T.K.** (1989) Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society, Series B **51**, 261-269*.

[73] **Kuk, A.Y.C. y Mak, T.K.** (1994) A functional approach to estimating finite population distribution functions. *Theory Meth.* **23 (3)**, 883-896.

[74] **Kuo, L.** (1988) Classical and Prediction Approaches to Estimating Distribution Functions from Survey Data. Proceeding of the Section on Survey Researh Methods. *American Statistical Association*, 280-285.

[75] **Lahiri, D.B.** (1951) A method of sample selection providing unbiased ratio estimates. *Bulletin of the International Statistical Institute **33**, 133-140*.

[76] **Leung, D.H.Y. y Qin, J.** (2006) Analysing survey data with incomplete responses by using a method based on empirical likelihood. *Journal of the Royal Statistical Society, Series C **55**, 379-396*.

[77] **Little, R.J.A. y Rubin, D.B.** (1987) *Statistical analysis with missing data*. John Wiley, New York.

[78] **Lombardía, M. J., Gónzalez-Manteiga, W. y Prada-Sánchez, J.M.** (2003) Bootstrapping the Chambers-Dunstan estimate of a finite population distribution function. *Journal of Statistical Planning and Inference **116**, 367-388*.

[79] **Lombardía, M. J., Gónzalez-Manteiga W., y Prada-Sánchez, J.M.** (2004) Bootstrapping the Dorfman-Hall-Chambers-Dunstan estimator of a finite population distribution function. *Journal of Nonparametric Statistics **16**, 63-90*.

[80] **Lucifora, C. y Salverda, W.** (1998) *Policies for low wage employment and social exclusion*. Ed. FrancoAngeli.

[81] **Mak, T.K. y Kuk, A.Y.C.** (1993) A new method for estimating finite–population quantiles using auxiliary information. *The Canadian Journal of Statistics **25**, 29-38*.

[82] **Martínez-Miranda, M.D., Rueda, M., Arcos, A., Román, Y. y Gónzalez, S.** (2005) Quantile estimation under successive sampling. *Computational Statistics **20**, 385-399.*

[83] **Midzuno, H.** (1952) On the sampling system with probability proportional to sum of sizes. *Annals of Institute of Statistical Mathematics **3**, 99-107*.

[84] **Molina, C.E.A. y Skinner, C.J.** (1992) Pseudo-likelihood and Quasi-likelihood estimation for complex sampling schemes. *Computational Statistics and Data Analysis **13**, 395-405*.

[85] **Mukhopadhyay, P.** (2000) *Topics in Survey Sampling* Springer.

[86] **Murthy, M.N.** (1967) *Sampling theory and method*. Calcutta: Statistical Publishing Society.

[87] **Narain, R.D.** (1953) On the recurrence formula in sampling on successive occasions. *Journal of the Indian Society of Agricultural Statistics **5**, 96-99*.

[88] **OECD** (1982) The OECD list of social indicators, Paris.

[89] **OECD** (1997) Labour market policies: new challenges policies for low-paid workers and unskilled job seekers. *OECD Working Papers. vol **5**, n^{o} 86* .

[90] **Ogus, J.K. y Clark, D.F.** (1971) The annual survey of manufacturers: A report on methodology. Technical Report No. 2, U.S. Bureau of Census, Washington D.C.

[91] **Olkin, I.** (1958) Multivariate ratio estimation for finite population. *Biometrika **45**, 154-165*.

[92] **Owen, A.B.** (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika **75**, 237-249*.

[93] **Owen, A.B.** (1990) Empirical likelihood confidence regions. *The Annals of Statistics **18**, 90-120*.

[94] **Owen, A.B.** (1991) Empirical likelihood for linear models. *The Annals of Statistics **19**, 1725-1747*.

[95] **Owen, A.B.** (2001) *Empirical likelihood.* Chapman y Hall/CRC.

[96] **Patterson, H.D.** (1950) Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B **12**, 241-255*.

[97] **Pérez, R.A.** (2002) ¿Qué es un modelo de superpoblación?. *Metodología de Encuestas **4 (1)**, 79-86*.

[98] **Polyak, B.T.** (1987) *Introduction to Optimization.* New York: Optimization Software, Inc. Publications Division.

[99] **Prasad, N.G.N. y Thach, T.** (2001) Variance estimation under two-phase sampling. *Working paper, Department of Mathematical Sciences, University of Alberta*.

[100] **Preston, I.** (1995) Sampling distributions of relative poverty statistics. *Journal of the Royal Statistical Society, Series C **44**, 91-99*.

[101] **Qin, J. y Lawless, J.F.** (1994) Empirical likelihood and general estimating equations. *The Annals of Statistics **22**, 300-325*.

[102] **Qin, J. y Lawless, J.F.** (1995) Estimating equations, empirical likelihood and constraints on parameters. *The Canadian Journal of Statistics **23**, 145*.

[103] **Randles, R.H.** (1982) On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics **10**, 462-474*.

[104] **Rao, J.N.K.** (1966) Alternative estimators in PPS sampling for multiple characteristics. *Sankhya Series A **28**, 47-60*.

[105] **Rao, J.N.K.** (1994) Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics **10**, 153-165*.

[106] **Rao, J.N.K., Kovar, J.G. y Mantel, H.J.** (1990) On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika **77**, 365-375*.

[107] **Rao, C.R. y Toutenburg, H.** (1995) *Linear Models: Least Squares and Alternatives*. Springer, New York.

[108] **Royall, R.M. y Cumberland, W.G.** (1981) An empirical study of the ratio estimator and estimator of its variance. *Journal of the American Statistical Association **76**, 66-88*.

[109] **Rubin, D.B.** (1987) *Multiple imputation for nonresponse in sample surveys*. Wiley, New York.

[110] **Rueda, M. y Arcos, A.** (2001) On estimating the median from survey data using multiple auxiliary information. *Metrika **4**, 161-173*.

[111] **Rueda, M. y Arcos, A.** (2002a) The use of quantiles of auxiliary variables to estimate medians. *Biometrical Journal **44** (5), 619-632*.

[112] **Rueda, M. y Arcos, A.** (2002b). Estimación por intervalos de la mediana con estimadores de razón y diferencia. *Estudios de Economía Aplicada **20**, 241-260*.

[113] **Rueda, M., Arcos, A. y Artés, E.** (1997) Improvement on Estimating Quantiles in Finite Population Using Indirect Methods of Estimation. *Lecture Notes in Computer Science **1280**, 491-500*.

[114] **Rueda, M., Arcos, A. y Artés, E.** (1998) Quantile Interval Estimation in Finite Population using a Multivariate Ratio Estimator. *Metrika **47**, 203-213*.

[115] **Rueda, M., Arcos, A. y Martínez-Miranda, M.D.** (2003) Difference estimators of quantiles in finite populations. *Test **12**, 481-496*.

[116] **Rueda, M., Arcos, A., Martínez-Miranda, M.D. y Román, Y.** (2004) Some improved estimators of finite population quantile using auxiliary information in sample surveys. *Computational Statistics and Data Analysis **45**, 825-848*.

[117] [1] **Rueda, M., Arcos, A., Muñoz, J.F. y Singh, S.**(2006) Quantile estimation in two-phase sampling. *Computational Statistics and Data Analysis*. En prensa.

[118] **Rueda, M. y González, S.** (2004) Missing data and auxiliary information in surveys. *Computational Statistic **19**, 551-567*.

[119] **Rueda, M., Martínez-Miranda, M.D., Arcos, A.** (2006) Bootstrap confidence intervals for finite population quantiles in the presence of auxilairy information. *Model Assisted Statistic and Aplications* En prensa.

[120] [1] **Rueda, M. y Muñoz, J.F.** (2005) Una revisión del método de verosimilitud empírica en las encuestas por muestreo. *Investigación Operacional **26**, 225-237*.

[121] [1] **Rueda, M. y Muñoz, J.F.**(2006a) A model-assisted estimator for the distribution function using the pseudo empirical likelihood method. *Statistics and Computing*. En revisión

[122] [1] **Rueda, M. y Muñoz, J.F.** (2006b) Estimating quantiles under sampling in two occasions with unequal probabilities. *Computational Statistics and Data Analysis.* Aceptado bajo revisión.

[123] [1] **Rueda, M. y Muñoz, J.F.** (2006c) Estimating quantiles under two−phase sampling for stratification. *Statistics and Probability Letters.* En revisión.

[124] [1] **Rueda, M. y Muñoz, J.F.** (2006d) Model-assisted estimation of quantiles using empirical likelihood. Applications to different poverty measures. *Journal of the Royal Statistical Society, Series C.* En revisión.

[125] [1] **Rueda, M., Muñoz, J.F., y Arcos, A.** (2006) Estimating quantiles under sampling on two occasions with P auxiliary variables. *Quality and Quantity*. En prensa.

[126] [1] **Rueda, M., Muñoz, J.F., Berger, Y.G., Arcos, A. y Martínez, S.** (2006) Pseudo empirical likelihood method in the presence of missing data. *Metrika*. En prensa.

[127] **Ruspini, E.** (1999) Longitudinal research and the analysis of social change. *Quality and Quantity **33**, 219-227*.

[128] **Sánchez-Crespo, G.** (2002) Introducción a los modelos de superpoblación en las técnicas de muestreo con probabilidades desiguales. *Metodología de Encuestas **4 (1)**, 87-104*.

[1] Bibliografía correspondiente al doctorando.

[1] Bibliografía correspondiente al doctorando.

[129] **Särndal, C.E.** (1980) On π-inverse weighting versus best linear weighting in probability sampling. *Biometrika **67**, 639-650*.

[130] **Särndal, C.E.** (1990) Methods for estimating the precision of survey estimates when imputation has been used. Proceedings of Symposium 1990: Measurement and improvement of data quality, Ottawa, 337-347.

[131] **Särndal, C.E.** (1992) Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology **18**, 241-252*.

[132] **Särndal, C.E., Swensson, B. y Wretman, J.H.** (1989) The weighted technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika **76**, 527-537*.

[133] **Särndal, C.E., Swensson, B. y Wretman, J.H.** (1992) *Model Assisted Survey Sampling.* Springer-Verlag, New York

[134] **Sedransk, J. y Meyer, J.** (1978) Confidence Intervals for the quantiles of a finite populations: simple random and stratified simple random sampling. *Journal of the Royal Statistical Society, Series B **40**, No2, 239-252*.

[135] **Sedransk, J. y Smith, P.J.** (1988) Inference for finite population quantiles. In: Krishnaiah, P.R. and Rao, C. R. (eds.) *Handbook of Statistics **6**, Cap11, 267-289*. North-Holland.

[136] **Sen, A.R.** (1972) Successive sampling with p $(p \geq 1)$ auxiliary variables. *The Annals of Mathematical Statistics **43 (6)**, 2031-2034*.

[137] **Sen, A.R.** (1973) Some theory of sampling on successive occasions. *The Australian Journal of Statistics **15 (2)**, 105-110*.

[138] **Sen, A. R., Sellers, S. y Smith, G.E.J.** (1975) The use of a ratio estimate in successive sampling. *Biometrics **31**, 673-683*.

[139] **Shao, J.** (1994) L-statistics in complex survey problems. *The Annals of Statistics **22**, 946-967*.

[140] **Shao, J. y Chen, Y.** (1998) Bootstrapping sample quantiles based on complex survery data under hot deck imputation. *Statistica Sinica **8**, 1071-1085*.

[141] **Shao, J. y Rao, J.N.K.** (1993) Standard errors for low income proportions estimated from stratified multi-stage samples. *Sankhya Series B **55**, 393-414*.

[142] **Shao, J. y Tu, D.** (1995) *The Jackknife and Bootstrap.* Springer-Verlag, New York.

[143] **Shao, J. y Wu, C.F.J.** (1989) A general theory for jackknife variance estimation. *The Annals of Statistics **17**, 1176-1197*.

[144] **Shao, J. y Wu, C.F.J.** (1992) Asymptotic properties of the balanced repeated replication method for sample quantiles. *The Annals of Statistics **20**, 1571-1593*.

[145] **Silva, P.L.D. y Skinner, C.J.** (1995) Estimating distribution functions with auxiliary information using poststratification. *Journal of Official Statistics **11 (3)**, 277-294*.

[146] **Silverman, B.W.** (1986) *Density estimation for statistics and data analysis*. Chapman and Hall.

[147] **Singh, S.** (2003) *Advanced sampling theory with applications: How Michael Selected Amy.*, Kluwer Academic Publishers, The Netherlands.

[148] **Singh, S., Joarder, A.H. y Tracy, D.S.** (2001) Median estimation using double sampling. *Australian and New Zealand Journal of Statistics **43**, 33-46*.

[149] **Singh, H.P., Singh, H.P. y Singh, V.P.** (1992) A generalized efficient class of estimators of population mean in two phase and successive sampling. *Inter. J. Mgmt. Syst. **8 (2)**, 173-183*.

[150] **Singh, S. y Srivastrava, A.K.** (1973) Use of auxiliary information in two stage successive sampling. *Journal of Indian Society of Agricultural Statistic **25**, 101-104*.

[151] **Sitter, R.R y Wu, C.** (2002) Efficient estimation of quadratic finite population functions in the presence of auxiliary information. *Journal of the American Statistical Association **97**, 535-543*.

[152] **Smeeding, T.M.** (1991) Cross-national comparisons of inequality and poverty position. In: Osberg, L. (Ed.), Economic Inequality and Poverty: International Perspectives, M.E. Sharpe, Inc., Armonk.

[153] **Solga, H.** (2001) Longitudinal surveys and the study of occupational mobility: Panel and retrospective design in comparison. *Quality and Quantity **35**, 291-309*.

[154] **Swamy, P.A.V.B., Tavlas, G.S. y Chang, I.L.** (2005) How stable are monetary police rules: estimating the time-varying coefficient in monetary policy reaction function for the U.S. *Computational Statistics and Data Analysis **49**, 575-590*.

[155] **Théberge, A.** (1999) Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association **94**, 635-644*.

[156] **Toutenburg, H. y Srivastava, V.K.** (1998) Estimation of ratio of population means in survey sampling when some observations are missing. *Metrika **48**, 177-187*.

[157] **Toutenburg, H. y Srivastava, V.K.** (1999) Amputation versus imputation of missing values through ratio method in sample surveys. Unpublished document.

[158] **Toutenburg, H. y Srivastava, V.K.** (2000) Efficient estimation of population mean using incomplete survey data on study and auxiliary characteristic. Unpublished document.

[159] **Tracy, D.S. y Osahan, S.S.** (1994) Random nonresponse on study variable versus on study as well as auxiliary variables. *Statistica **54**, 163-168*.

[160] **Valliant, R., Dorfman, A.H. y Royall, R.M.** (2000) *Finite population sampling and inference: A prediction approach*. Wiley Series in Probability and Statistics, Survey Methodology Section. New York. John Wiley and Sons, Inc.

[161] **Wang, S. y Dorfman, A.H.** (1996) A new estimator for the finite population distribution function. *Biometrika **83**, 639-652*.

[162] **Wolfson, M. y Evans, J.M.** (1989) Statistics Canada's low income cut-offs: metodological concerns and possibilities - a discussion paper. Research Paper Series, Statistical Canada, Ottawa. distribution function. *Biometrika **83**, 639-652*.

[163] **Wolter, K.M.** (1985) *Introduction to Variance Estimation.* Springer-Verlag.

[164] **Woodruff, R.S.** (1952) Confidence intervals for medians and other position measures. *Journal of the American Statistical Association **47**, 635-646*.

[165] **Wu, C.** (2002) Empirical likelihood method for finite populations. *Recent Advances in Statistical Methods,* Y.P. Chaubey, Ed., Imperial College Press, London, *339-351*.

[166] **Wu, C.** (2003) Optimal calibration estimators in survey sampling. *Biometrika **90**, 937-951*.

[167] **Wu, C.** (2004a) Weighted empirical likelihood inference. *Statistics and Probability Letters **66/1**, 67-79*.

[168] **Wu, C.** (2004b) Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica **14**, 1057-1067*.

[169] **Wu, C.** (2004c) Combining information from multiple surveys through empirical likelihood method. *The Canadian Journal of Statistics **32**, 15-26*.

[170] **Wu, C.** (2005) Algorithms and R Codes for the Pseudo Empirical Likelihood Method in Survey Sampling. *Survey Methodology, **31**, 239-243*.

[171] **Wu, C. y Luan, Y .** (2003) Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics **19**, 119-131*.

[172] **Wu, C. y Sitter, R.R.** (2001) A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association **96**, 185-193*.

[173] **Zheng, B.** (2001) Statistical inference for poverty measures with relative poverty lines. *Journal of Econometrics **101**, 337-356*.

[174] **Zhong, C.X.B., Chen, J. y Rao, J.N.K.** (2000) Empirical likelihood inference in the presence of measurement error. *The Canadian Journal of Statistics **28**, 841*.

[175] **Zhong, C.X.B. y Rao, J.N.K.** (1996) Empirical likelihood inference for finite populations with auxiliary information using stratified random sampling. *Proceeding of the Section on Survey Research Methods, Am. Statist. Assoc., 793-803*. Washington, DC: American Statistical Association.

[176] **Zhong, C.X.B. y Rao, J.N.K.** (2000) Empirical likelihood inference under stratified random sampling using auxiliary information. *Biometrika **87**, 929-938*.

A. Descripción de poblaciones finitas

En este apéndice se detallan las distintas poblaciones que han sido usadas en este trabajo con objeto de estudiar el comportamiento de los estimadores propuestos y su precisión con respecto a otros estimadores existentes en las literatura. Notamos que las poblaciones basadas en datos reales han sido utilizadas por otros autores en diferentes estudios de simulación, siendo estas poblaciones apropiadas para el estudio del comportamiento de estimadores en muestreo de poblaciones finitas. Las poblaciones que han sido simuladas siguen los modelos propuestos por otros autores, o bien, se han simulado de manera que pueda ser posible la extracción de muestras en los diseños muestrales más complejos que han sido tratados en este trabajo. De esta forma, se dispone de una estructura de datos apropiada para la obtención de tanto los estimadores propuestos como del resto de estimadores existentes en la literatura.

A.1. Poblaciones naturales

A.1.1. Fam1500

Esta población consta de $N = 1500$ familias de Andalucía y fue usada por primera vez por Fernández y Mayor (1994). Numerosos estudios posteriores (por ejemplo, Rueda *et al.*, 2006a, 2006b, Rueda y González, 2004, etc.) han usado esta población en sus estudios de simulación. La característica de interés, y, son los gastos de alimentación, mientras que las variables auxiliares x_1 y x_2 son, respectivamente, los ingresos familiares y otros gastos. En la Tabla A.1 puede consultarse información adicional sobre las variables de la población Fam1500, mientras que la Figura B.31 muestra los diagramas de dispersión correspondientes a dichas variables.

A.1.2. Counties

Las poblaciones Counties60 y Counties70 son poblaciones habitualmente usadas en muestreo de poblaciones finitas. Fueron usadas por primera vez en Royall y Cumberland (1981). Posteriormente, se ha usado en numerosos trabajos, como por ejemplo en Valliant *et al.* (2000). La población Counties60 consta de $N = 304$ ciudades de Carolina del Norte, Carolina del Sur y Georgia con menos de 100000 hogares en el año 1960. La variable y es la población de cada ciudad, excluyendo los barrios de grupos de residentes. Como variable auxiliar, x, se tiene el número de hogares en 1960.

Por otro lado, la población Counties70 está formada por la variable de interés y que denota la población de 304 ciudades de Carolina del Norte, Carolina del Sur y Georgia con menos de 100000 hogares en el año 1970, excluyendo los barrios de grupos de residentes y por las variables auxiliares x_1 y x_2, que coinciden con las variables x e y, respectivamente, de la población anterior.

Los datos de esta población pueden descargarse de:

ftp://ftp.wiley.com/public/sci_tech_med/finite_populations

Además, un breve resumen descriptivo de estas poblaciones puede consultarse en las Tablas A.2 y A.3. La Figura B.32 nos da los diagramas de dispersión entre las distintas variables de estas poblaciones. Puede observarse que estas poblaciones exhiben una mejor relación lineal entre las variables que la población Fam1500, lo que nos ha permitido comprobar en los distintos estudios el grado de ganancia en precisión en función de una mayor o menor relación lineal entre la variable principal y las auxiliares.

A.1.3. Hospitals

Esta población es una muestra nacional de hospitales en Estados Unidos. Esta muestra también fue considerada como una población en los estudios llevados a cabo por Royall y Cumberland (1981) y Valliant *et al.* (2000). El tamaño poblacional es de $N = 393$ hospitales de corta estancia con menos de 1000 camas, la variable de interés, y, es el número de pacientes dados de alta, mientras que la variable auxiliar es el número de camas que dispone el hospital.

El resumen descriptivo de las variables de esta población puede consultarse en la Tabla A.4. El diagrama de dispersión dado por la Figura B.33 nos permite profundizar en la estructura que presentan los datos de las variables de la población Hospitals.

Tabla A.1: Análisis descriptivo para las variables de la población Fam1500

V.	Min	Q_1	Me	Media	Q_3	Max	Cv	ρ_{yx}
y	5045	7358	8136	8181.94	8941	11795	0.14	
x_1	30052	36660	40200	40283.96	43700	55379	0.12	0.848
x_2	2116	3515	4001	4044.40	4538	6990	0.19	0.546

Tabla A.2: Análisis descriptivo para las variables de la población Counties60

V.	Min	Q_1	Me	Media	Q_3	Max	Cv	ρ_{yx}
y	1876	9787	18330	32916	38690	266623	1.24	
x	482	2502	4886	8931	10410	76887	1.30	0.998

Tabla A.3: Análisis descriptivo para las variables de la población Counties70

V.	Min	Q_1	Me	Media	Q_3	Max	Cv	ρ_{yx}
y	1924	9613	19080	36984	42560	409644	1.38	
x_1	482	2502	4886	8931	10410	76887	1.30	0.982
x_2	1876	9787	18330	32916	38690	266623	1.24	0.982

Tabla A.4: Análisis descriptivo para las variables de la población Hospitals

V.	Min	Q_1	Me	Media	Q_3	Max	Cv	ρ_{yx}
y	14	311	713	814.65	1186	2844	0.72	
x	1	102	233	274.70	393	986	0.78	0.911

A.1.4. Murthy

La población Murthy es apropiada para observar el efecto de una mala especificación de un modelo de superpoblación en los estimadores, y poder proporcionar, por tanto, una indicación de la robustez de tales estimadores. Esta población consta de 80 fábricas donde la variable de interés, y, es la producción, y como variable auxiliar, x, se ha considerado el número de trabajadores. Esta población se usó previamente en Murthy (1967), Kuk y Mak (1989) y Kuk y Mak (1994).

En la Figura B.34 puede comprobarse que una hipótesis de linealidad no sería válida para las variables de esta población. Un estudio más exhaustivo sobre las características de las variables de la población Murthy puede obtenerse a partir de la Tabla A.5.

A.1.5. Turismos

Esta población se ha obtenido a partir del número de turismos recogidos en los años 2002 y 2003 por el Instituto de Estadística de Andalucía en los distintos municipios de Andalucía. Estos datos pueden descargarse en la página web del Instituto de Estadística de Andalucía:

http:\\www.juntadeandalucia.es\institutodeestadistica

Por tanto, La población Turismos está formada por el número de turismos en $N = 770$ municipios de Andalucía. La variable principal, y, es el número de turismos por municipio en el año 2003. Se dispone de cuatro variables auxiliares: x_1, x_2, x_3 y x_4 que corresponden al número de turismos en el año 2002 con capacidad cilíndrica de clase 1, 2, 3 y 4, respectivamente.

El objetivo que tiene el uso de esta población es comprobar la ganancia en eficiencia de las estimaciones cuando se aumenta de manera paulatina el número de variables auxiliares.

En el análisis descriptivo de la Tabla A.6 se muestran las características más importantes de las variables de la población Turismos. En estas variables destaca la presencia de una alta asimetría y una importante variabilidad en los datos, como reflejan los correspondientes coeficientes de variación. Los diagramas de dispersión asociados a estas variables están disponibles en la Figura B.35.

A.1.6. ECPF1997

La última población natural que se ha considerado en este trabajo se corresponde con los datos muestrales procedentes del primer trimestre del año 1997 de la Encuesta Continua de Presupuestos Familiares ($ECPF$). Véase Instituto Nacional de Estadística (1992) para una consulta detallada de la metodología. Esta población ha sido también analizada en Fernández *et al.* (2004).

Notamos que el objetivo de esta encuesta es proporcionar estimaciones acerca de los gastos de consumo y de los ingresos para el conjunto nacional, según varias variables de clasificación. La población consta de $N = 3000$ hogares españoles, donde se ha considerado que la variable de interés, y, son los ingresos totales trimestrales por hogar (en euros), mientras que los gastos trimestrales por hogar (en euros) será la variable auxiliar.

El correspondiente análisis descriptivo de las variables de esta población está dado por la Tabla A.7. Observamos que en este caso no existe una fuerte relación lineal entre la variable principal y la auxiliar. Este hecho es frecuente entre datos correspondientes a variables tales como ingresos o gastos, donde la alta presencia de valores extremos habitualmente dificulta la interpretación de algunas medidas como la media.

En cualquier caso, el objetivo al usar esta población es comprobar el comportamiento real de distintos estimadores en situaciones donde no pueda aceptarse una fuerte relación lineal entre las variables. En la Figura B.36 se muestra el correspondiente diagrama de dispersión.

A.2. Poblaciones simuladas

A.2.1. Pop06, Pop07, Pop08 y Pop09

Paralelamente a Wu y Sitter (2001), se han generado cuatro poblaciones de $N = 2000$ unidades mediante muestras independientes e idénticamente distribuidas mediante el modelo

$$y = \theta_0 + \theta_1 x + \epsilon, \tag{A.1}$$

donde $x \sim Gamma(1, 1)$, $\epsilon \sim N(0, \sigma^2)$ y $\theta_0 = \theta_1 = 1$. Estas poblaciones se han generado escogiendo diferentes valores de σ^2, de modo que los coeficientes de correlación entre y y x están dados por 0.6, 0.7, 0.8 y 0.9. Las poblaciones se han llamado Pop06, Pop07, Pop08 y Pop09, respectivamente. La Figura B.37 muestra los diagramas de dispersión de estas poblaciones, mientras que los distintos estudios descriptivos están dados por las Tablas A.8, A.9 ,A.10 y A.11.

A.2.2. Pob098 y Pob080

Por último, se han generado dos poblaciones (Pob098 y Pob080) de tamaño $N = 1000$ mediante el modelo

$$y_i = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \epsilon_i, \tag{A.2}$$

donde $\theta_0 = \theta_1 = \theta_2 = 1$ y las variables x_{1i} y x_{2i} se han generado de distribuciones Gamma con parámetros de forma y escala dados por 4 y 1, respectívamente. Las cantidades ϵ_i son variables aleatorias independientes e idénticamente distribuidas con distribución Normal de parámetros 0 y σ^2. El valor de σ^2 se ha seleccionado de modo que el coeficiente de correlación entre y_i e $\widehat{y}_i = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i}$ es 0.98 para la primera población (Pob098) y 0.80 para la segunda población (Pob080). Los análisis descriptivos de estas poblaciones están dados por las Tablas A.12 y A.13, mientras que los diagramas de dispersión los encontramos en las Figuras B.38 y B.39.

Tabla A.5: Análisis descriptivo para las variables de la población Murthy

V.	Min	Q_1	Me	Media	Q_3	Max	Cv	ρ_{yx}
y	1176	3727.0	5105	5183.0	6754.0	9250	0.35	
$x1$	51	86.5	148	285.1	445.3	1095	0.94	0.915

Tabla A.6: Análisis descriptivo para las variables de la población Turismos

V.	Min	Q_1	Me	Media	Q_3	Max	Cv	ρ_{yx}
y	11	343.3	894.0	3967.8	2483.5	308738	4.23	
x_1	5	73.0	176.5	810.2	464.0	61176	4.41	0.994
x_2	4	101.0	263.0	1313.7	749.3	111977	4.55	0.998
x_3	1	123.0	338.0	1373.1	957.5	102710	4.04	0.998
x_4	0	22.0	61.0	295.9	174.8	24023	4.26	0.961

Tabla A.7: Análisis descriptivo para las variables de la población ECPF1997

V.	Min	Q_1	Me	Media	Q_3	Max	Cv	ρ_{yx}
y	240.4	2745	4037	4660	5842	61320	0.67	
x	107.6	2609	3845	4527	5654	27730	0.66	0.594

Tabla A.8: Análisis descriptivo para las variables de la población Pop06

V.	Min	Q_1	Me	Media	Q_3	Max	Cv	ρ_{yx}
y	-2.4588	0.87	1.93	1.98	2.96	9.33	0.81	
x	0.0008	0.27	0.66	0.96	1.32	8.10	1.03	0.6

Tabla A.9: Análisis descriptivo para las variables de la población Pop07

V.	Min	Q_1	Me	Media	Q_3	Max	Cv	ρ_{yx}
y	-2.349	1.02	1.88	2.00	2.86	10.03	0.71	
x	0.001	0.30	0.70	0.99	1.36	8.22	0.98	0.7

Tabla A.10: Análisis descriptivo para las variables de la población Pop08

V.	Min	Q_1	Me	Media	Q_3	Max	Cv	ρ_{yx}
y	-2.243	1.15	1.81	1.99	2.63	8.54	0.64	
x	0.001	0.25	0.67	0.98	1.34	7.36	1.04	0.8

Tabla A.11: Análisis descriptivo para las variables de la población Pop09

V.	Min	Q_1	Me	Media	Q_3	Max	Cv	ρ_{yx}
y	-0.374	1.23	1.73	1.96	2.43	11.80	0.57	
x	0.002	0.29	0.67	0.98	1.33	10.51	1.02	0.9

Tabla A.12: Análisis descriptivo para las variables de la población Pob098

V.	Min	Q_1	Me	Media	Q_3	Max	Cv	ρ_{yx}
y	-0.207	5.07	7.33	7.99	9.97	25.65	0.52	
x_1	0.003	0.90	2.26	3.08	4.37	22.32	0.96	0.71
x_2	0.081	1.80	3.17	3.85	5.34	17.55	0.72	0.67
$\widehat{y}$	1.615	4.97	7.23	7.93	10.03	25.08	0.51	0.98

Tabla A.13: Análisis descriptivo para las variables de la población Pob080

V.	Min	Q_1	Me	Media	Q_3	Max	Cv	ρ_{yx}
y	-0.097	6.61	8.69	8.89	11.00	19.98	0.37	
x_1	0.480	2.46	3.67	3.98	5.15	11.86	0.50	0.60
x_2	0.417	2.54	3.59	3.89	5.00	12.20	0.48	0.53
$\widehat{y}$	3.316	6.88	8.65	8.87	10.47	20.84	0.30	0.80

B. Representaciones gráficas

Figura B.1: Eficiencia Relativa para los estimadores $\overline{y}_{PE}^{A}$ (Pemle 1), $\overline{y}_{PE}^{AB}$ (Pemle 12), $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ (Alpha óptimo), $\overline{y}_{Reg}$ (Regresión) y $\overline{y}_{T3}$ (Toutenburg 3). Se toman muestras de tamaño $n = 200$.

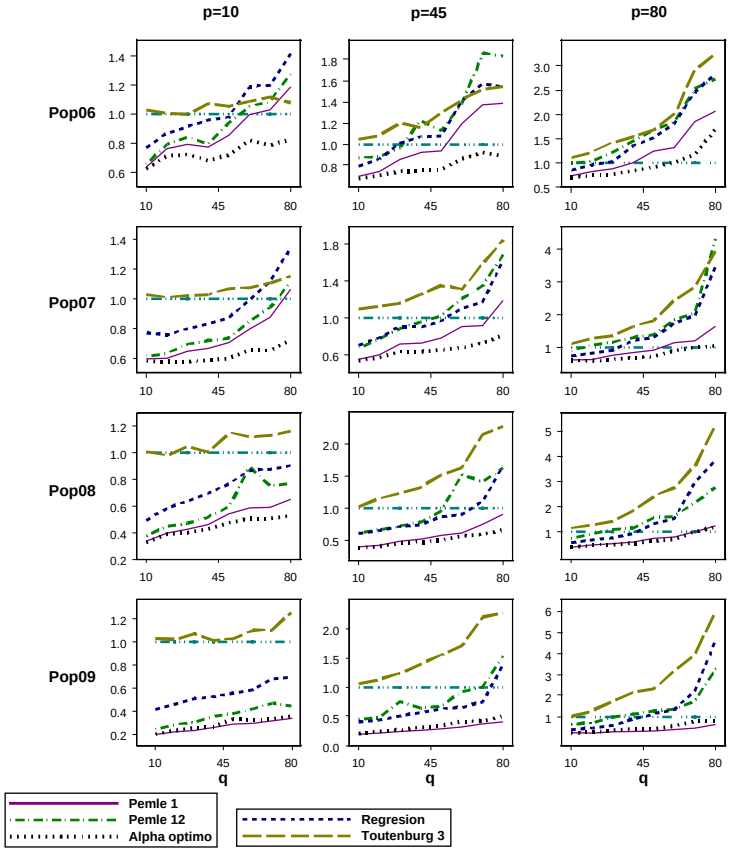

Figura B.2: Eficiencia Relativa para los estimadores $\overline{y}_{PE}^{A}$ (Pemle 1), $\overline{y}_{PE}^{AB}$ (Pemle 12), $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ (Alpha óptimo), $\overline{y}_{Reg}$ (Regresión) y $\overline{y}_{T3}$ (Toutenburg 3). Se considera la población Fam1500 y muestras de tamaño $n = 150$.

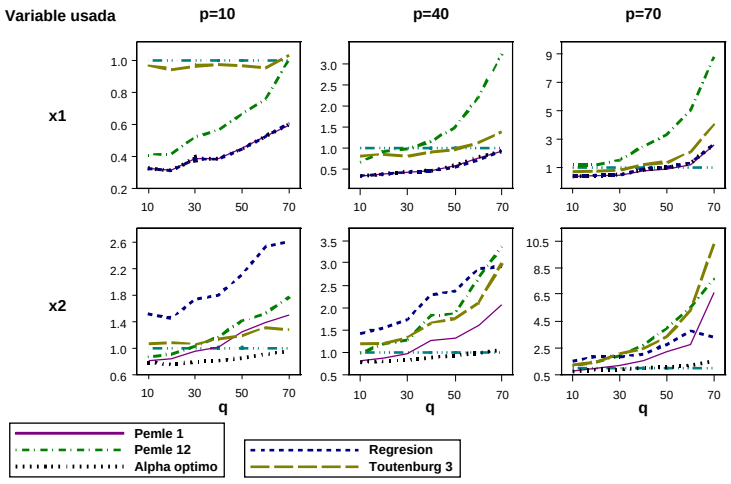

Figura B.3: Eficiencia Relativa para los estimadores $\overline{y}_{PE}^{A}$ (Pemle 1), $\overline{y}_{PE}^{AB}$ (Pemle 12), $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ (Alpha óptimo), $\overline{y}_{Reg}$ (Regresión) y $\overline{y}_{T3}$ (Toutenburg 3). Se considera la población Hospitals y muestras de tamaño $n = 100$.

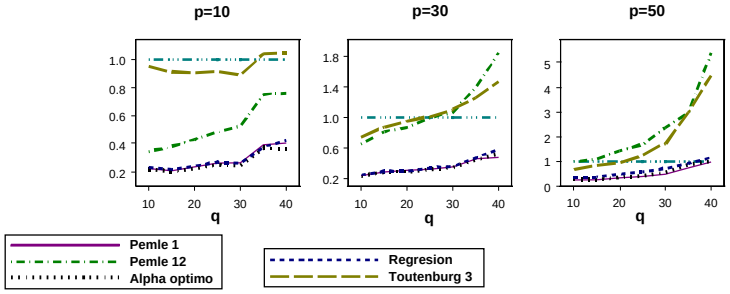

Figura B.4: Sesgo Relativo para los estimadores $\overline{y}_{PE}^{A}$ (Pemle 1), $\overline{y}_{PE}^{AB}$ (Pemle 12), $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ (Alpha óptimo), $\bar{y}_{w}^{AC}$ (estándar), $\overline{y}_{Reg}$ (Regresión) y $\overline{y}_{T3}$ (Toutenburg 3). Se toman muestras de tamaño $n = 200$.

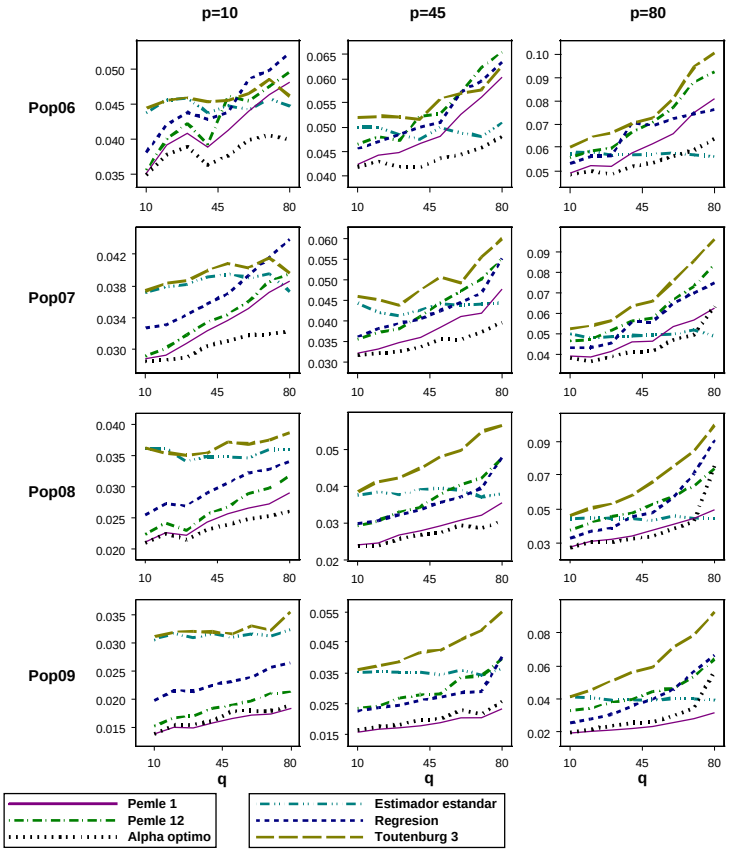

Figura B.5: Sesgo Relativo para los estimadores $\overline{y}_{PE}^{A}$ (Pemle 1), $\overline{y}_{PE}^{AB}$ (Pemle 12), $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ (Alpha óptimo), $\bar{y}_{w}^{AC}$ (estándar), $\overline{y}_{Reg}$ (Regresión) y $\overline{y}_{T3}$ (Toutenburg 3). Se considera la población Fam1500 y muestras de tamaño $n = 150$.

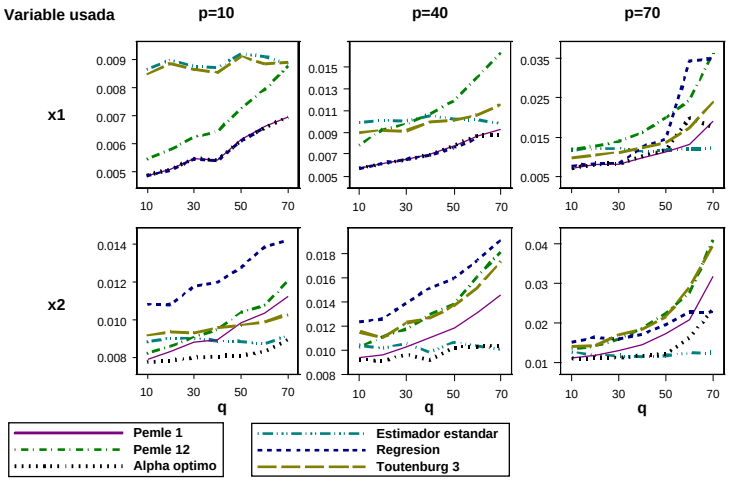

Figura B.6: Sesgo Relativo para los estimadores $\overline{y}_{PE}^{A}$ (Pemle 1), $\overline{y}_{PE}^{AB}$ (Pemle 12), $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ (Alpha óptimo), $\bar{y}_{w}^{AC}$ (estándar), $\overline{y}_{Reg}$ (Regresión) y $\overline{y}_{T3}$ (Toutenburg 3). Se considera la población Hospitals y muestras de tamaño $n = 100$.

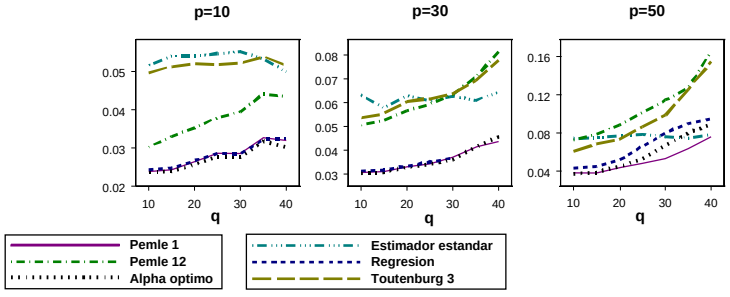

Figura B.7: Eficiencia Relativa de distintos estimadores en las poblaciones Pob098 y Pob080.

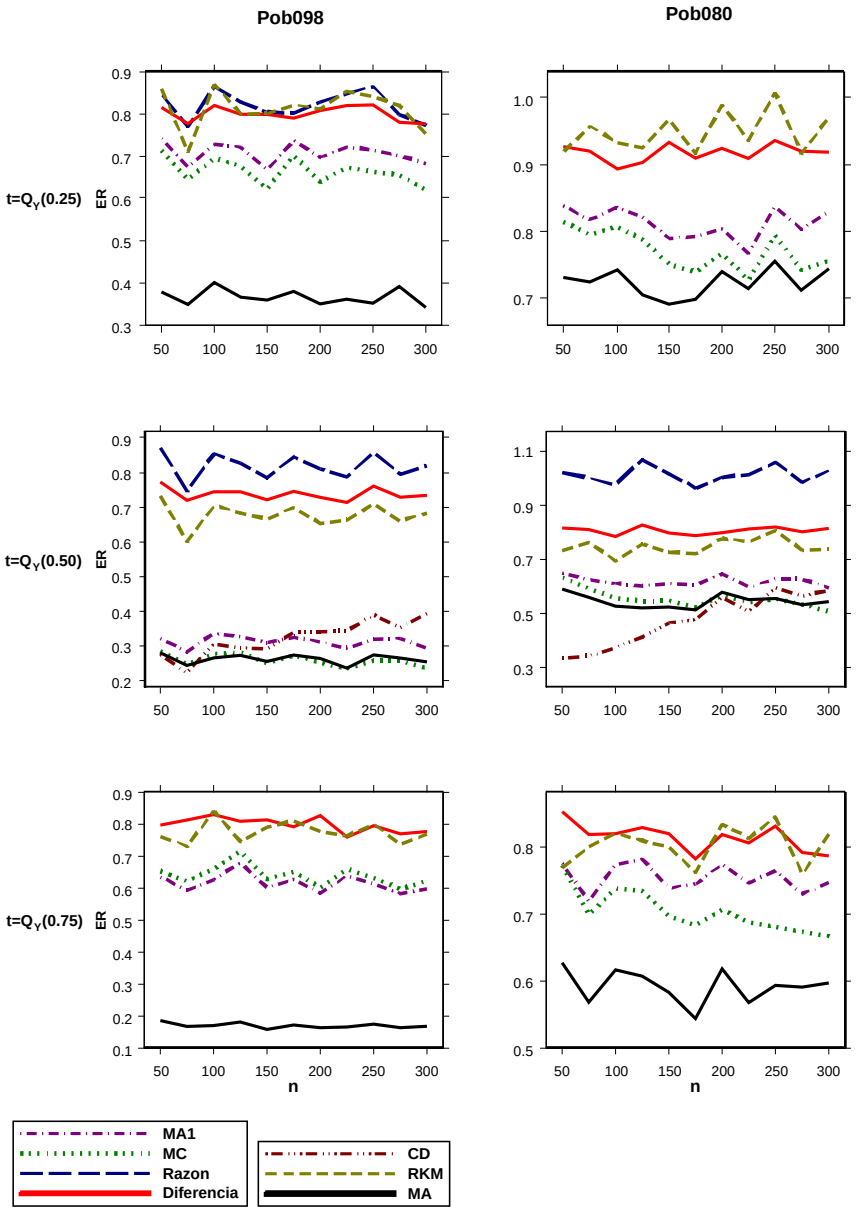

Figura B.8: Eficiencia Relativa de distintos estimadores en la población Murthy.

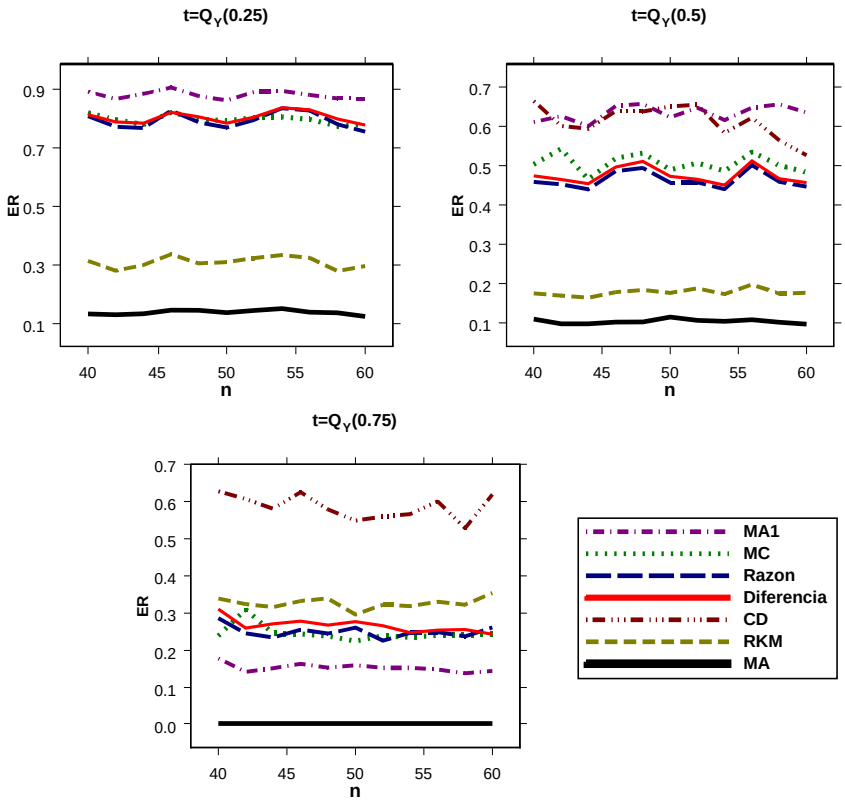

Figura B.10: Eficiencia Relativa Media de distintos estimadores en las poblaciones Pob098, Pob080 y Murthy.

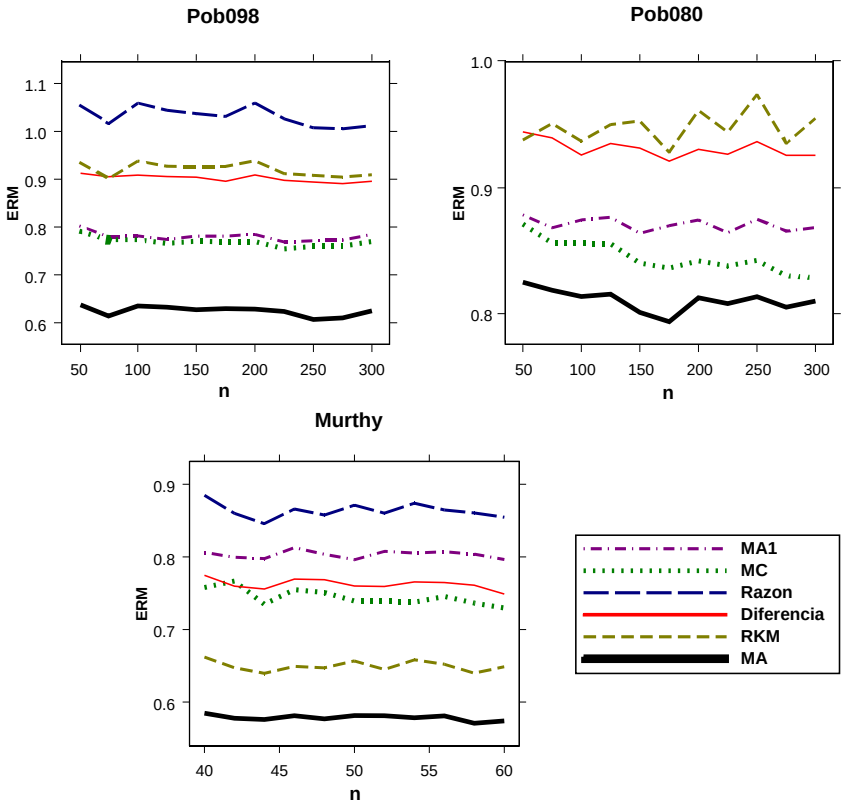

Figura B.11: Diagramas de cajas con bigotes de las Desviaciones Absolutas Medias de distintos estimadores en las poblaciones Pob098 (con $n = 100$), Pob080 (con $n = 100$) y Murthy (con $n = 50$).

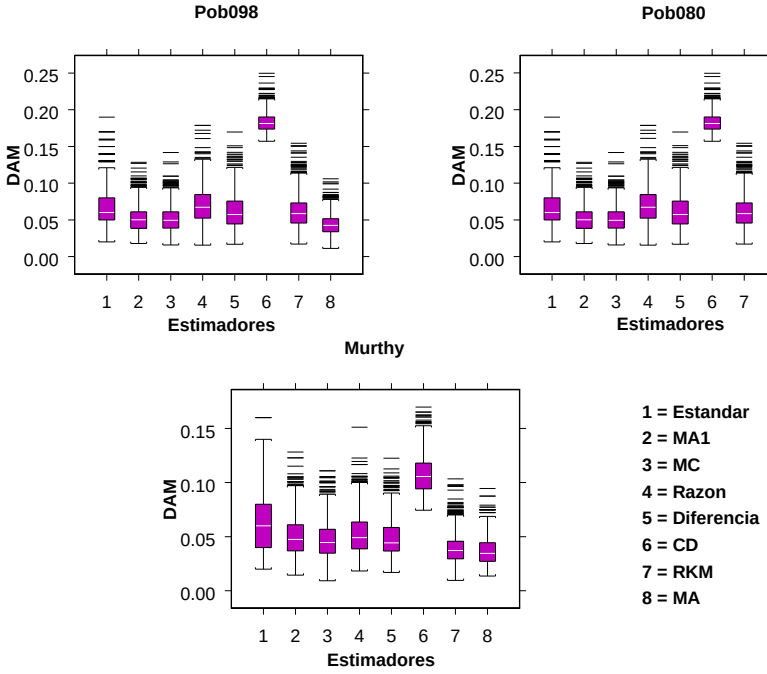

Figura B.12: Eficiencia Relativa para la población Fam1500 y bajo el diseño muestral $Mas.Midzuno$. $n' = 150$.

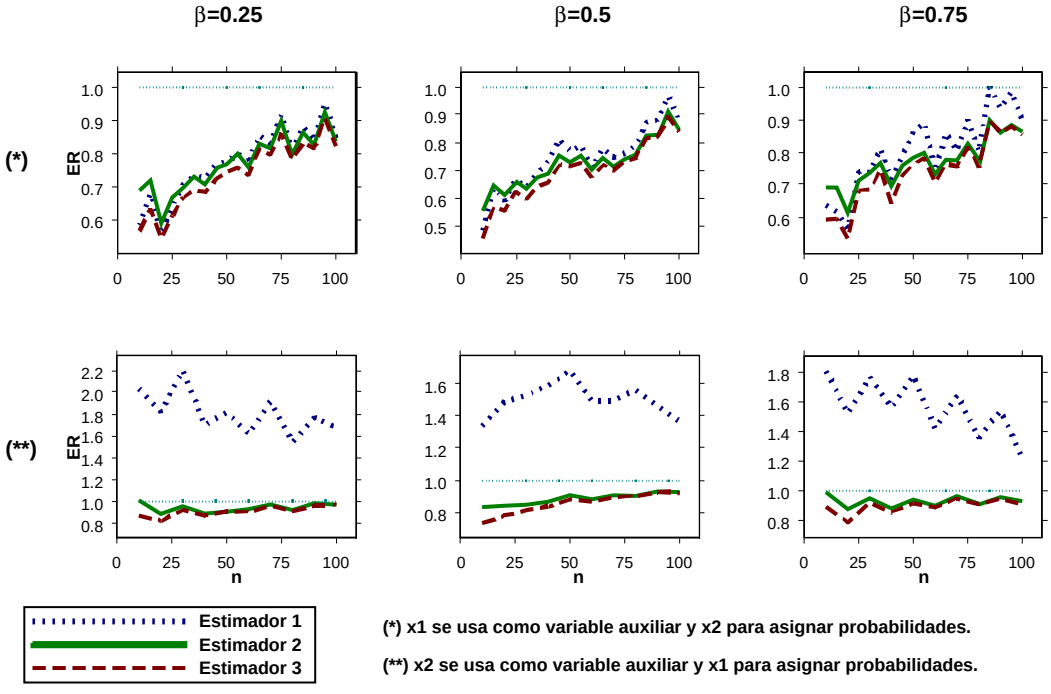

Figura B.13: Eficiencia Relativa para la población Fam1500 y bajo el diseño muestral $Mas.Poisson$. $n' = 150$.

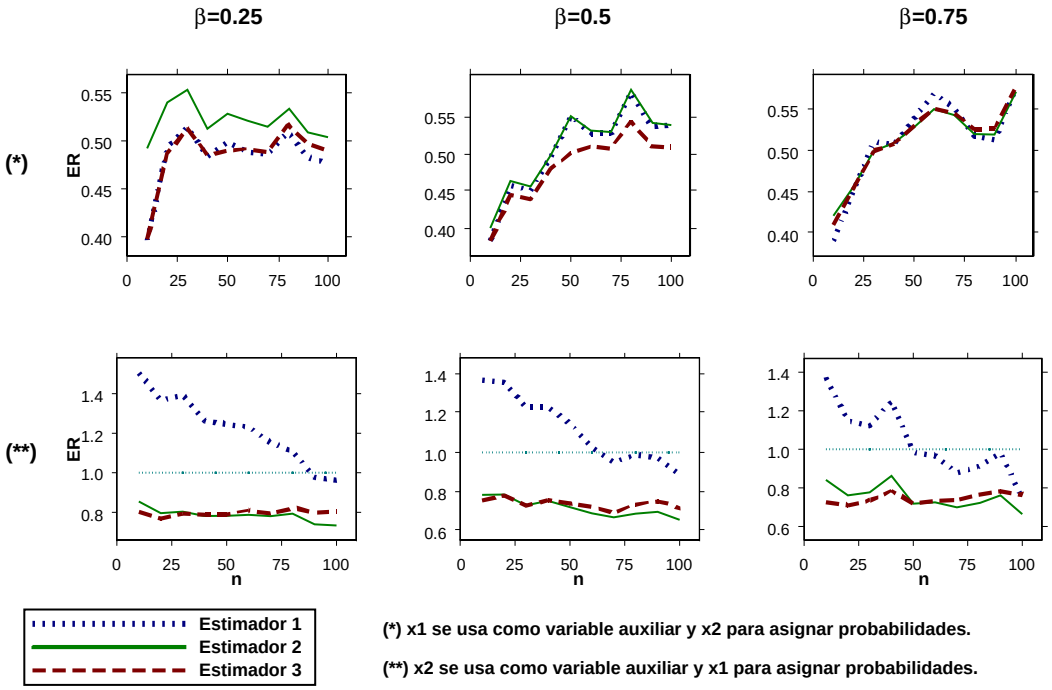

Figura B.14: Eficiencia Relativa para la población Counties y bajo el diseño muestral $Mas.Midzuno$. $n' = 150$.

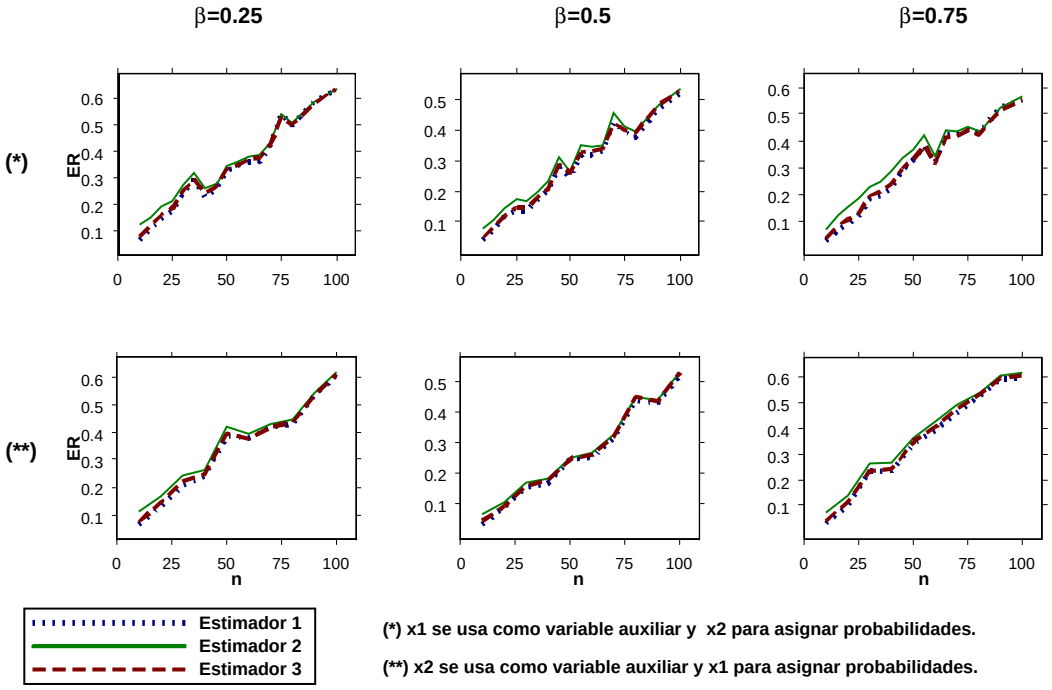

Figura B.15: Eficiencia Relativa para la población Counties y bajo el diseño muestral $Mas.Poisson$. $n' = 150$.

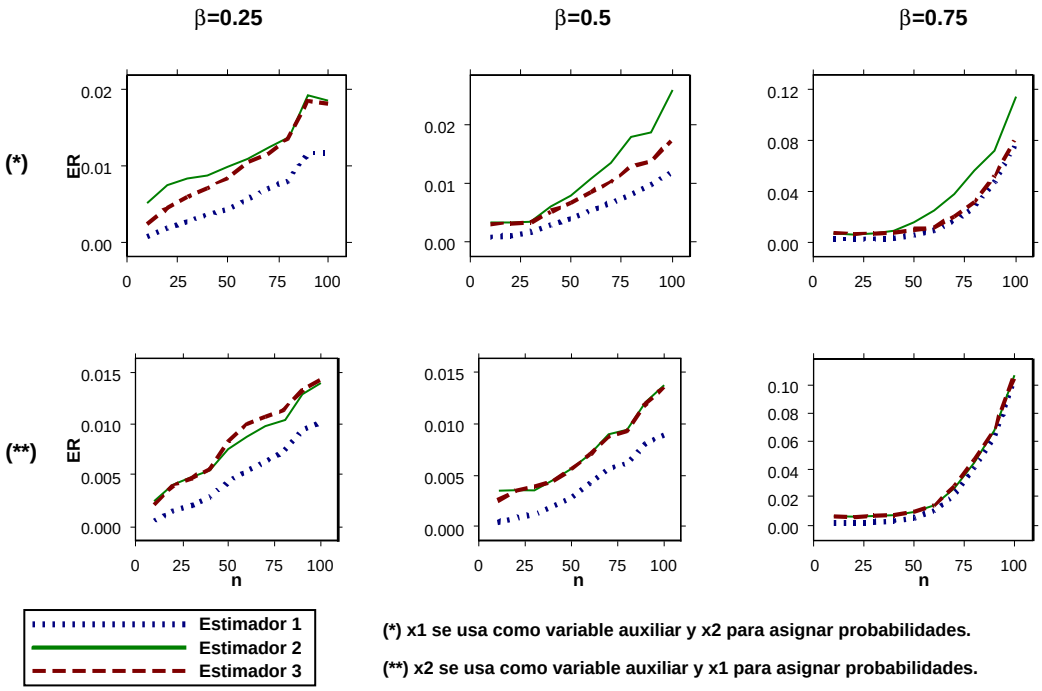

Figura B.16: Sesgo Relativo en porcentaje para la población Fam1500 cuando $x1$ se usa como variable auxiliar y $x2$ para asignar probabilidades. $n' = 150$.

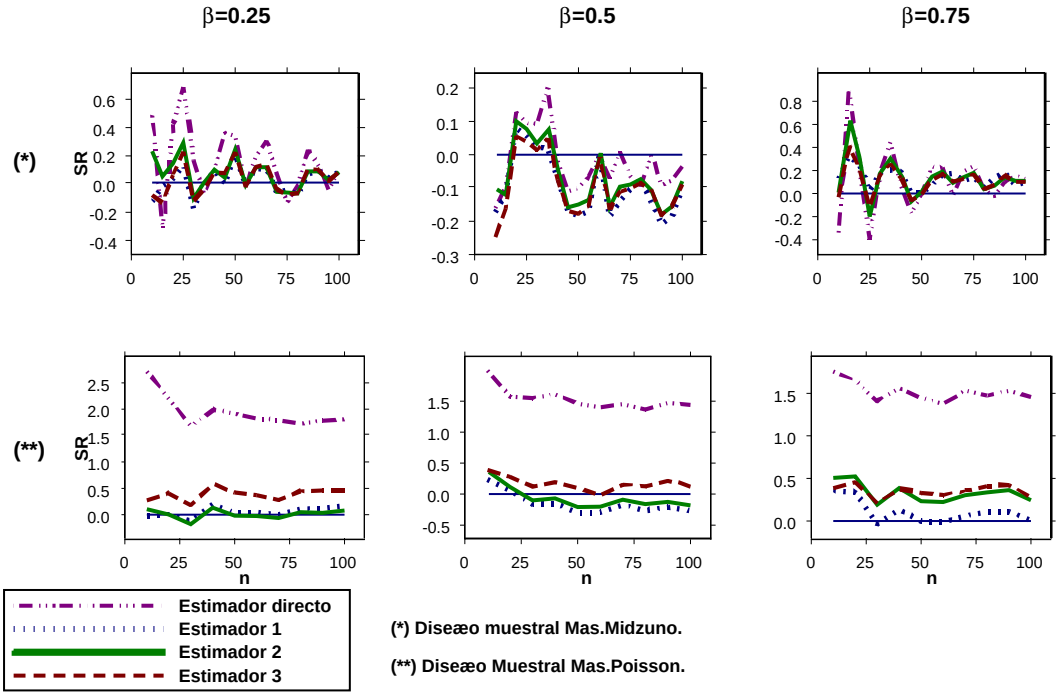

Figura B.17: Sesgo Relativo en porcentaje para la población Counties cuando $x1$ se usa como variable auxiliar y $x2$ para asignar probabilidades. Los valores SR para el estimador directo en (**) son mayores de 97.6 %, 74.6 % y 21.5 % para $\beta = 0{,}25, 0{,}5$ y 0.75, respectivamente, y están omitidos. $n' = 150$.

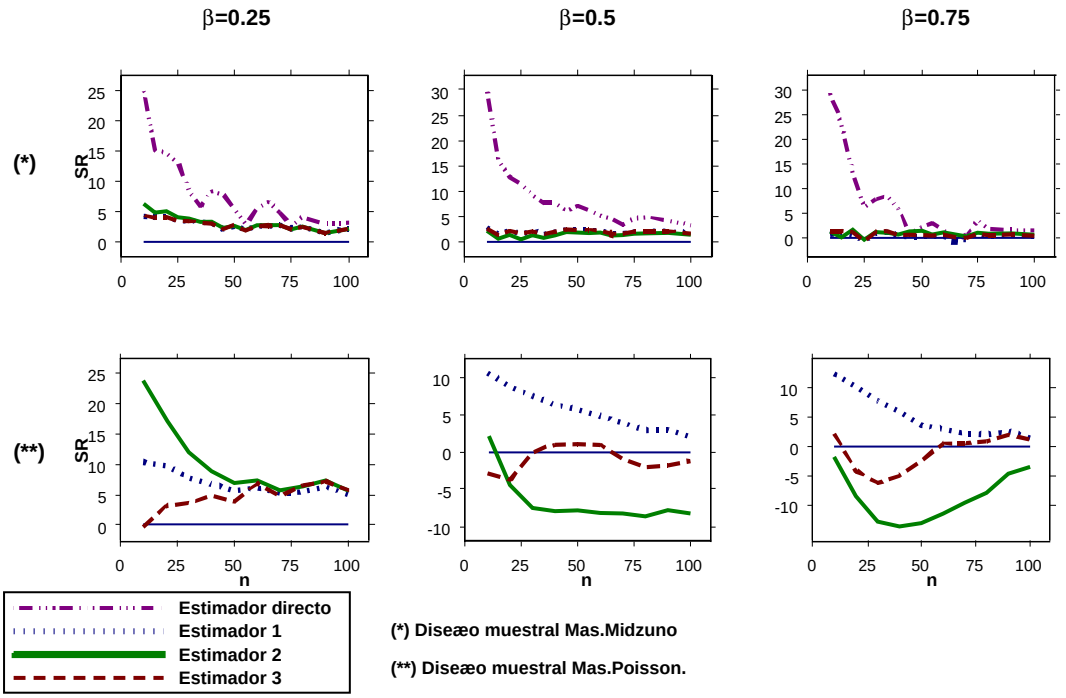

Figura B.18: Ratios Teóricos entre la varianza del estimador óptimo propuesto y la varianza del estimador estándar bajo la población Counties y el cuantil de orden $\beta = 0{,}5$.

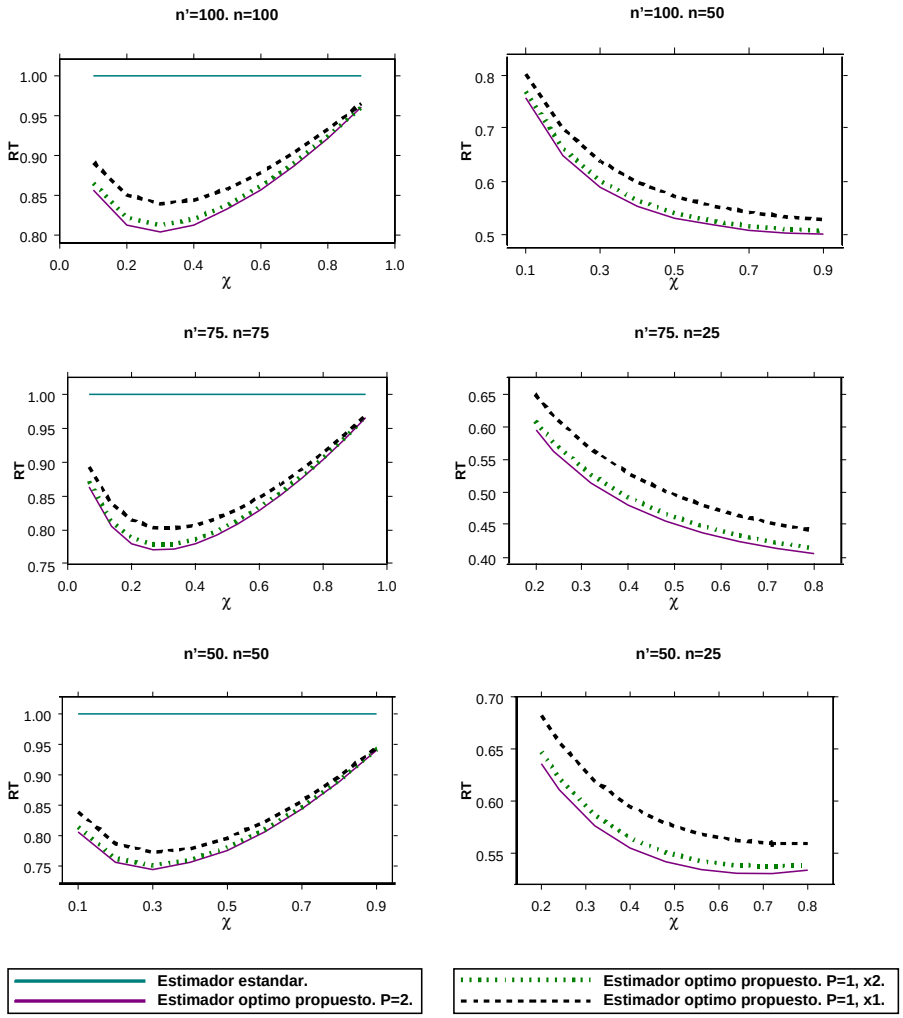

Figura B.19: Ratios Teóricos entre la varianza del estimador óptimo propuesto y la varianza del estimador estándar bajo la población Turismos y el cuantil de orden $\beta = 0{,}5$.

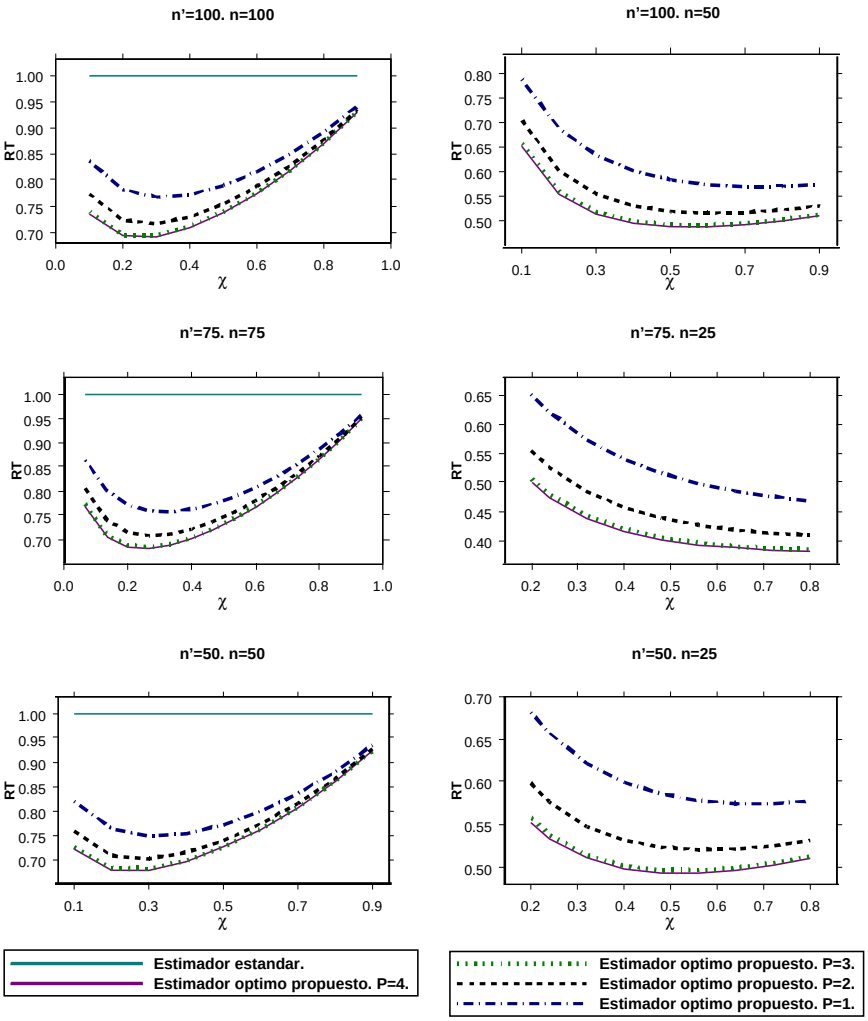

Figura B.20: Eficiencia Relativa para los estimadores óptimo propuesto y estándar en la población Counties y para el cuantil de orden $\beta = 0{,}5$.

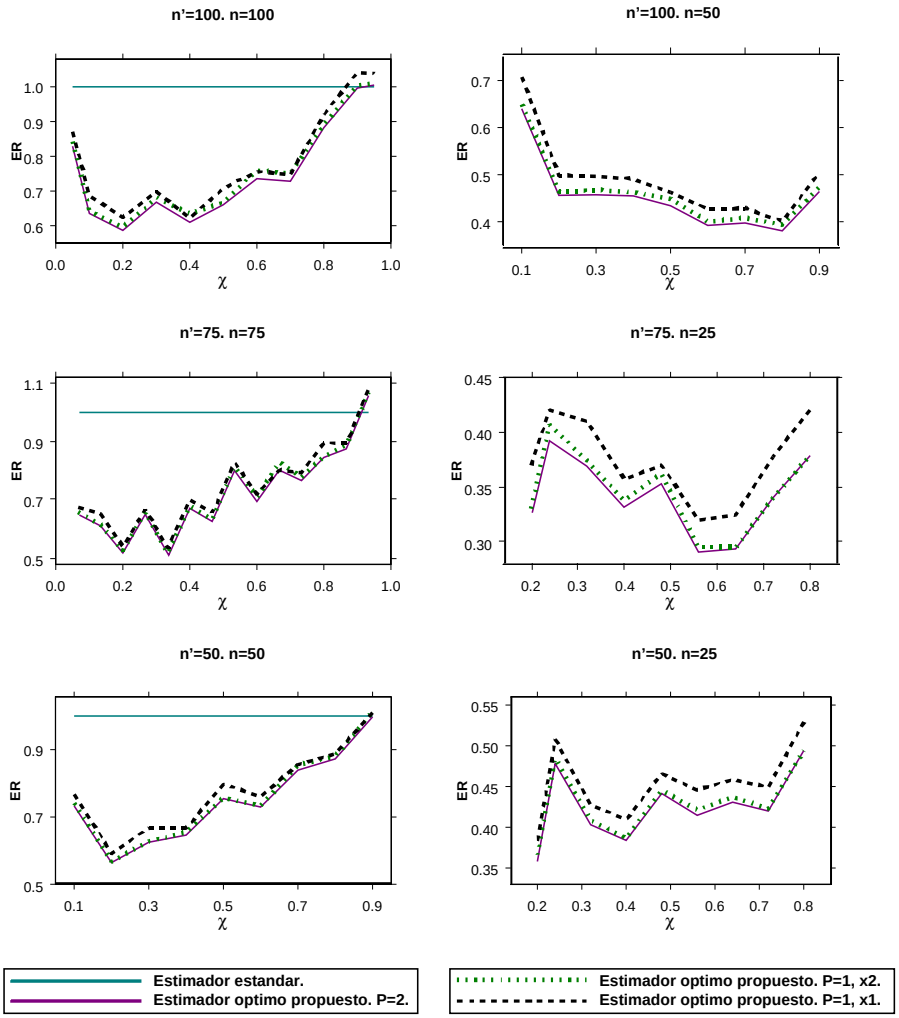

Figura B.21: Eficiencia Relativa para los estimadores óptimo propuesto y estándar en la población Turismos y para el cuantil de orden $\beta = 0{,}5$.

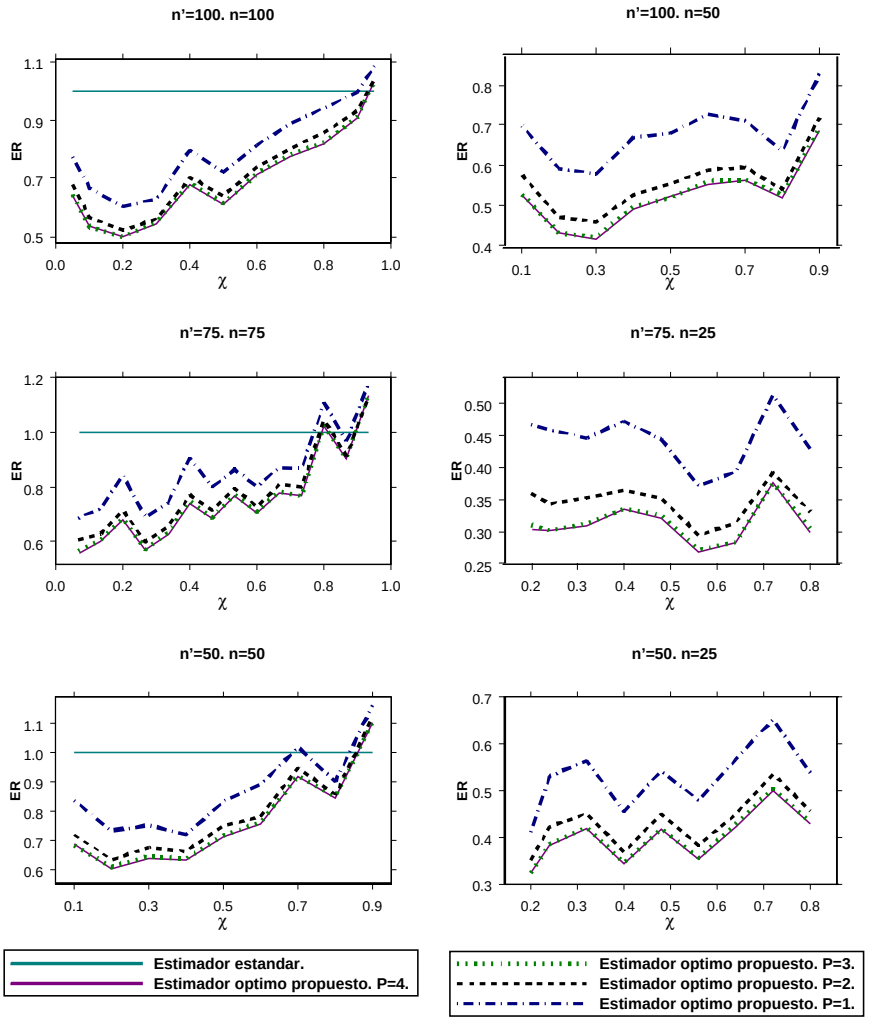

Figura B.22: Evolución de los valores W_{opt} usados por el estimador óptimo propuesto en la población Counties y para el cuantil de orden $\beta = 0{,}5$.

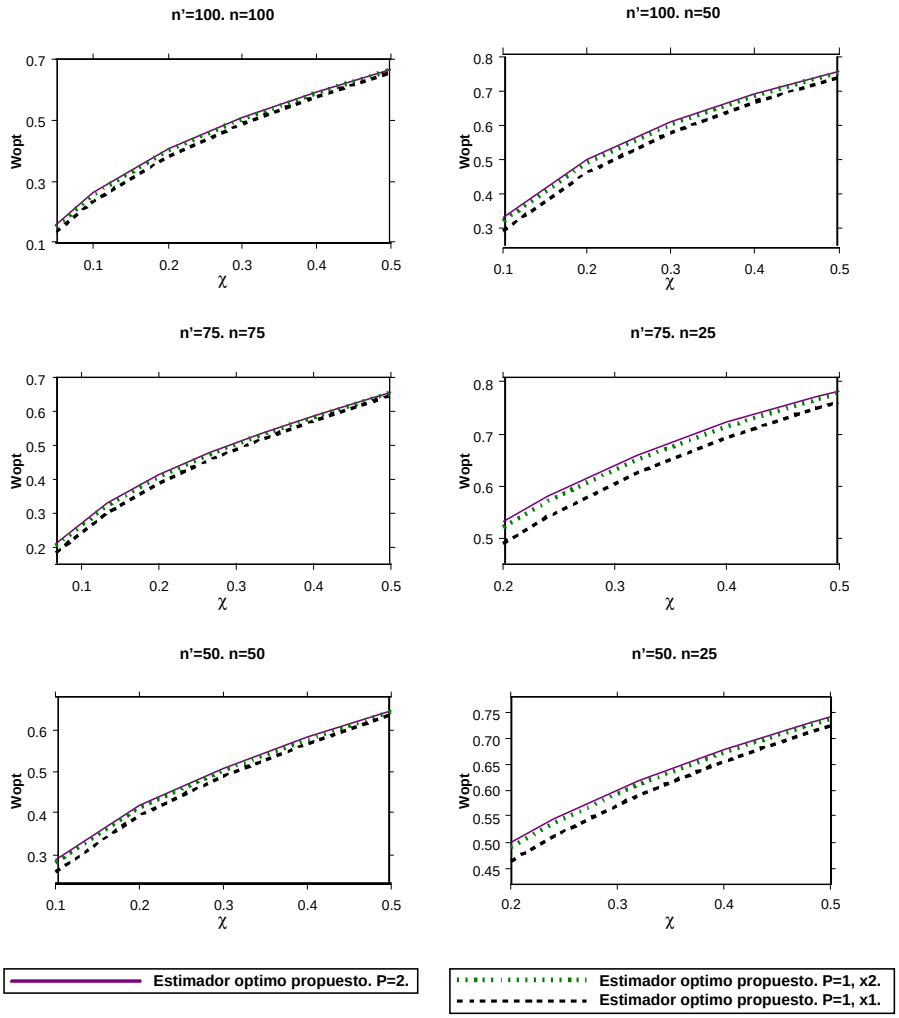

Figura B.23: Evolución de los valores W_{opt} usados por el estimador óptimo propuesto en la población Turismos y para el cuantil de orden $\beta = 0{,}5$.

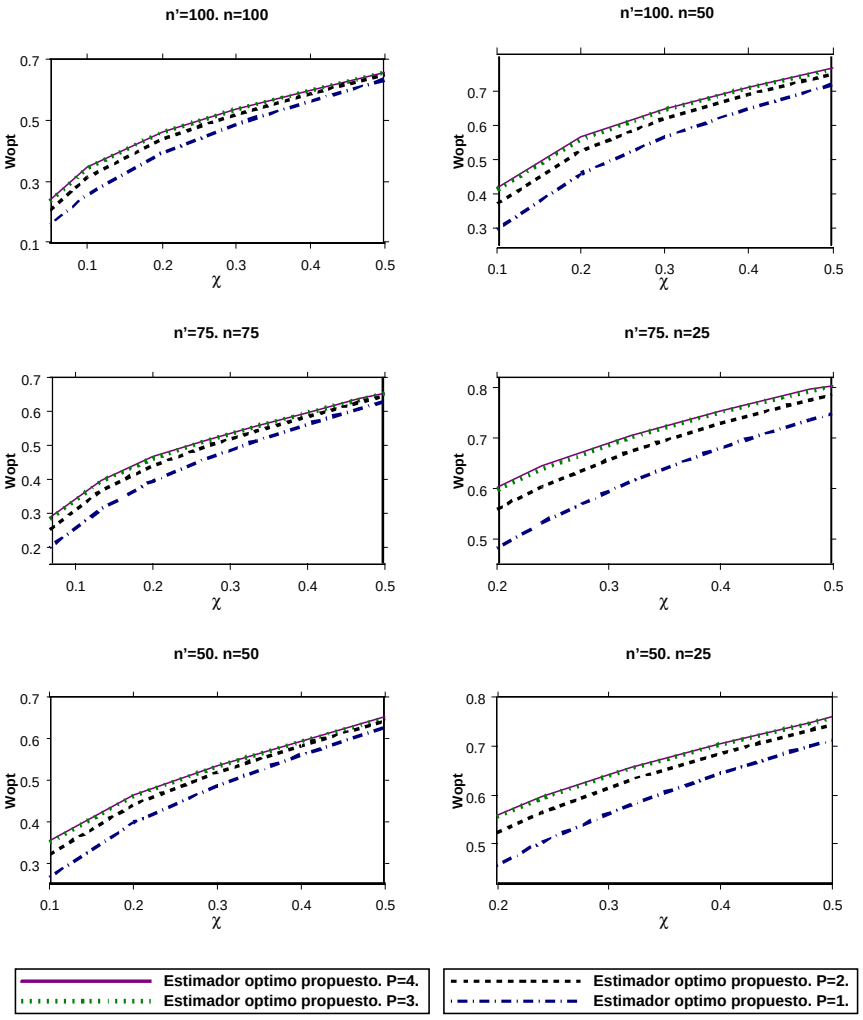

Figura B.24: Eficiencia Relativa para el diseño muestral SMS.

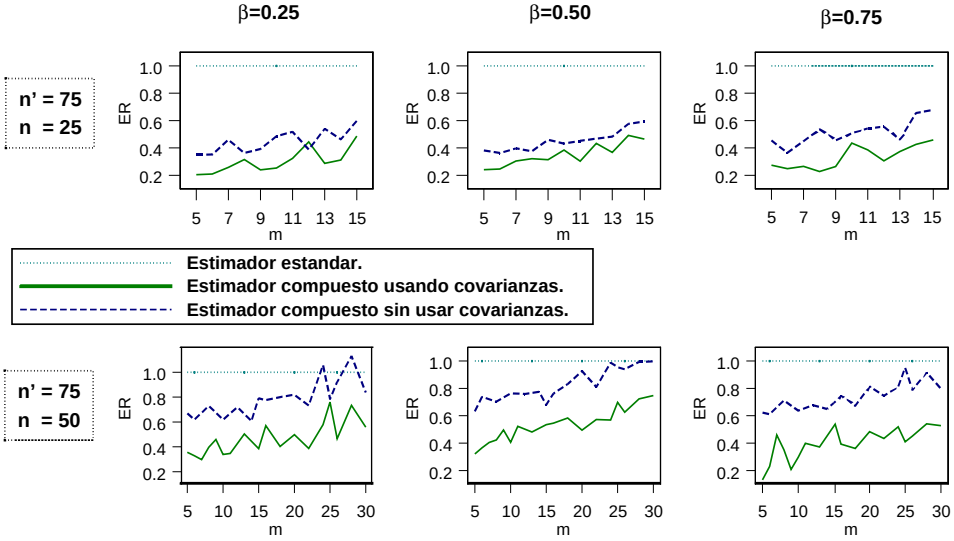

Figura B.25: Eficiencia Relativa para el diseño muestral MSS.

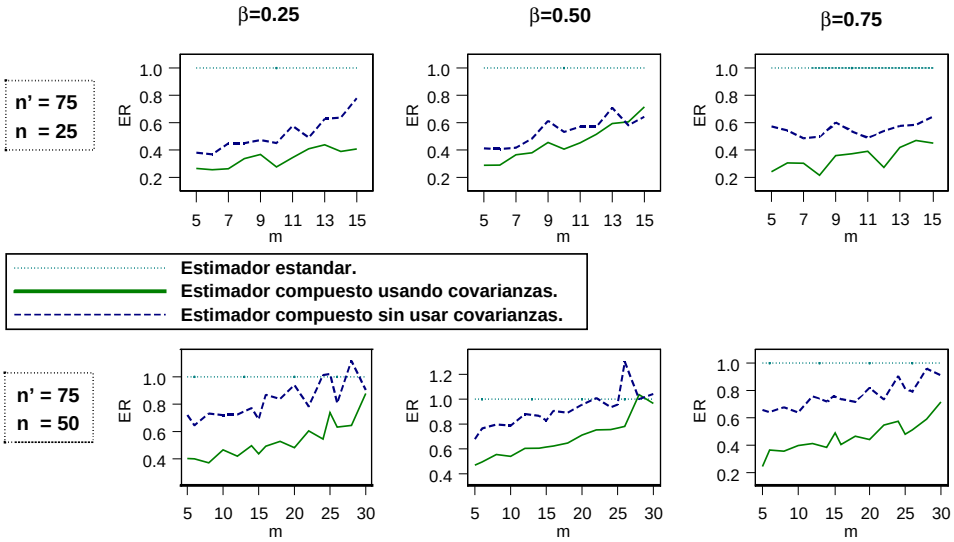

Figura B.26: Eficiencia Relativa para el diseño muestral MMM.

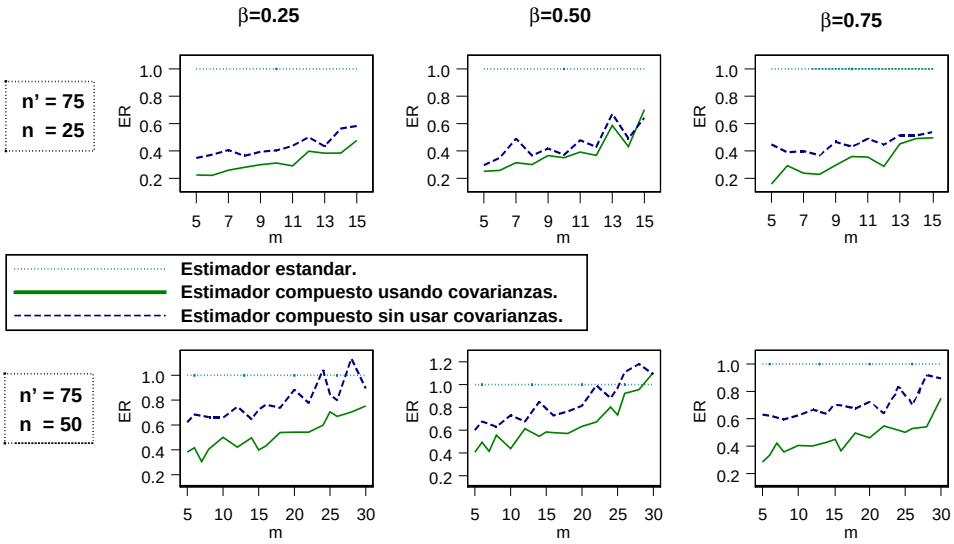

Figura B.27: Sesgo Relativo para el diseño muestral SMS.

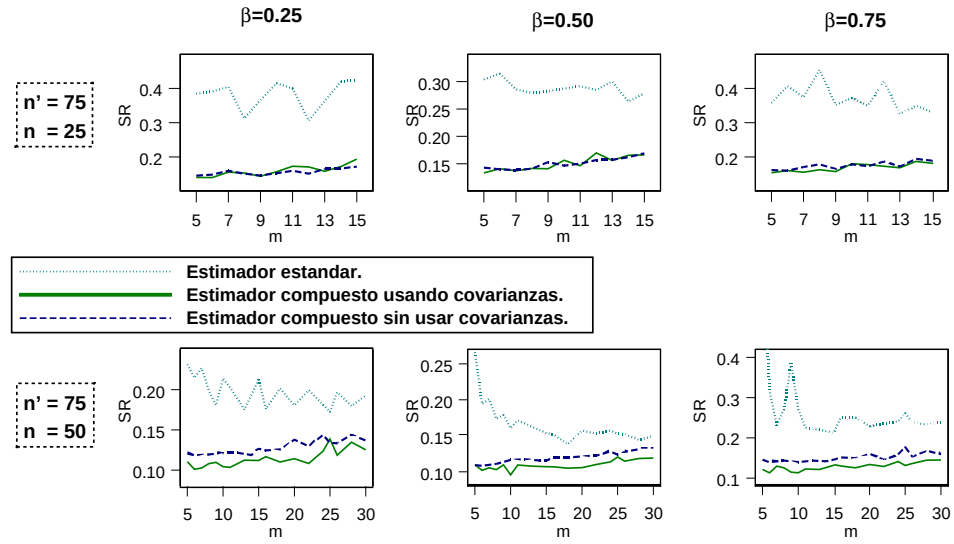

Figura B.28: Sesgo Relativo para el diseño muestral MSS.

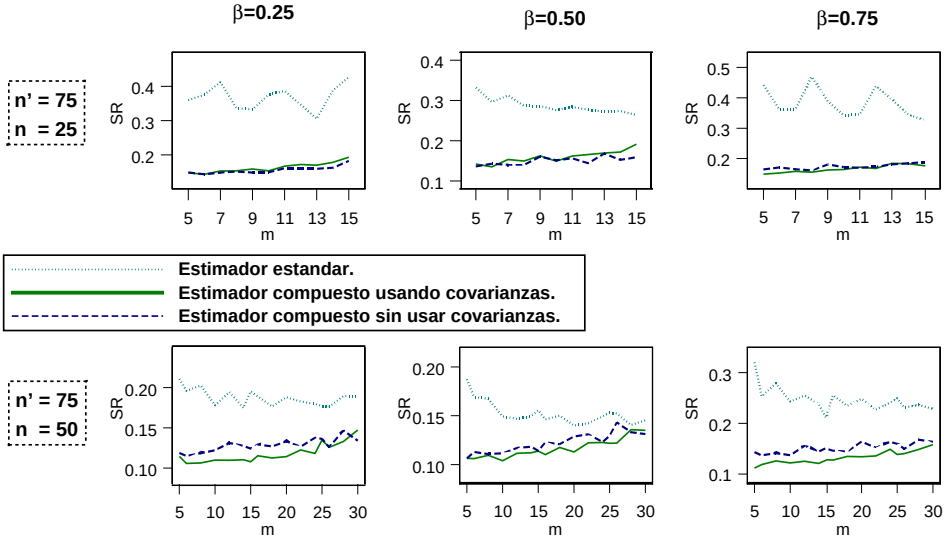

Figura B.29: Sesgo Relativo para el diseño muestral MMM.

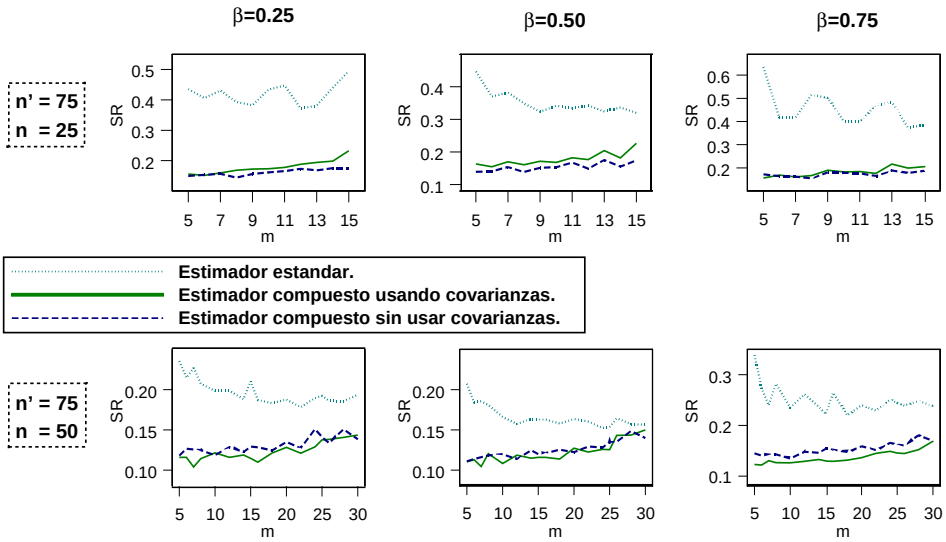

Figura B.30: Diagrama de caja con bigotes para los valores de los distintos estimadores. Se asume el diseño muestral SMS y tamaños muestrales $n' = 75$ y $n = 50$.

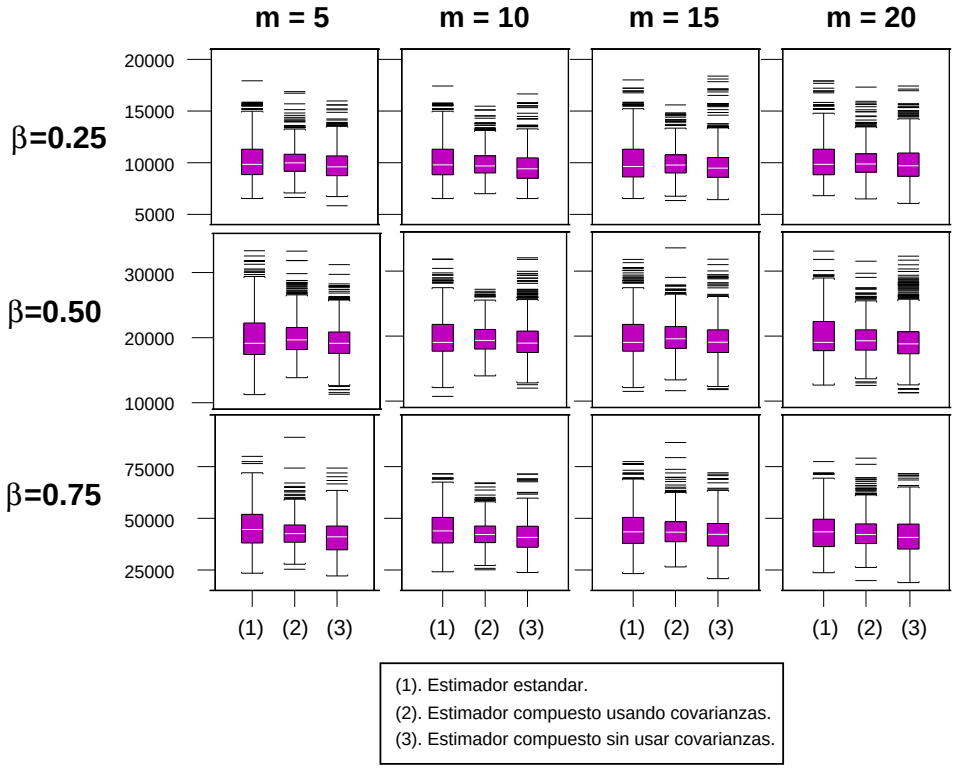

Figura B.31: Diagramas de dispersión de la población Fam1500

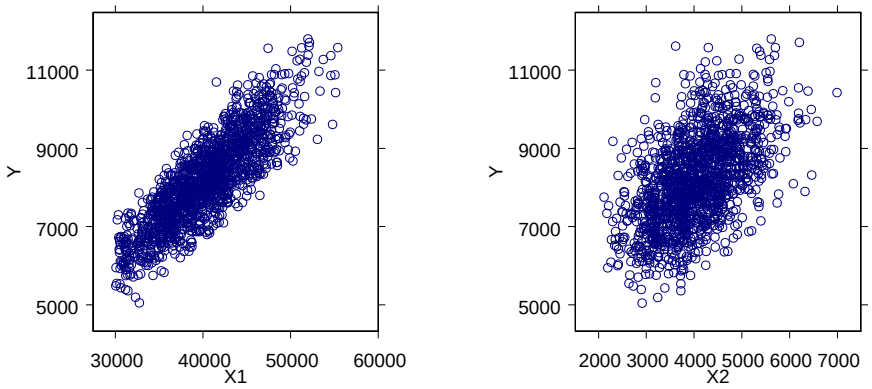

Figura B.32: Diagramas de dispersión de las poblaciones Counties70 y Counties60.

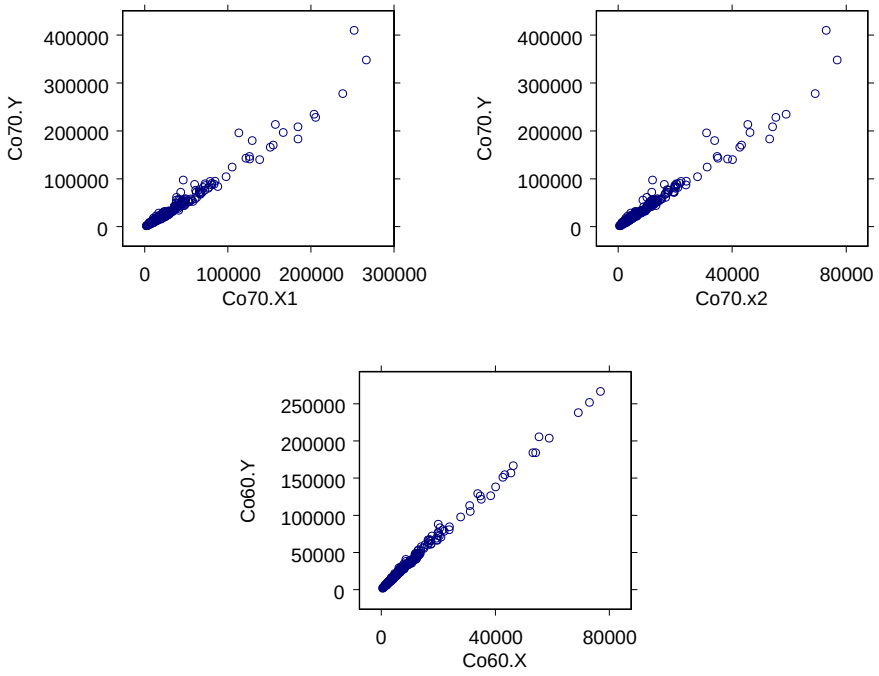

Figura B.33: Diagrama de dispersión de la población Hospitals.

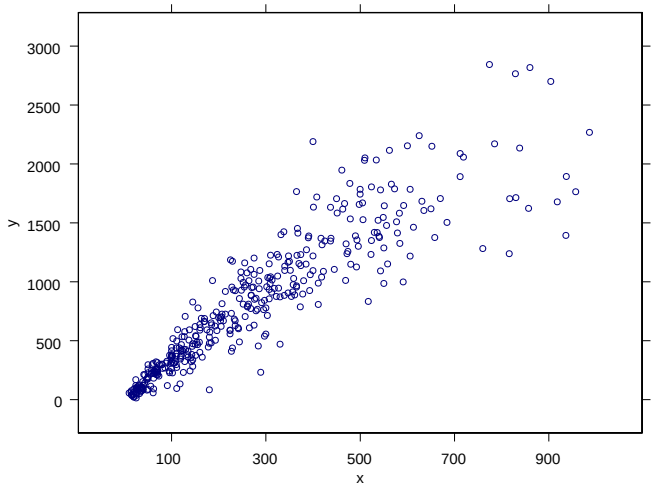

Figura B.34: Diagrama de dispersión de la población Murthy.

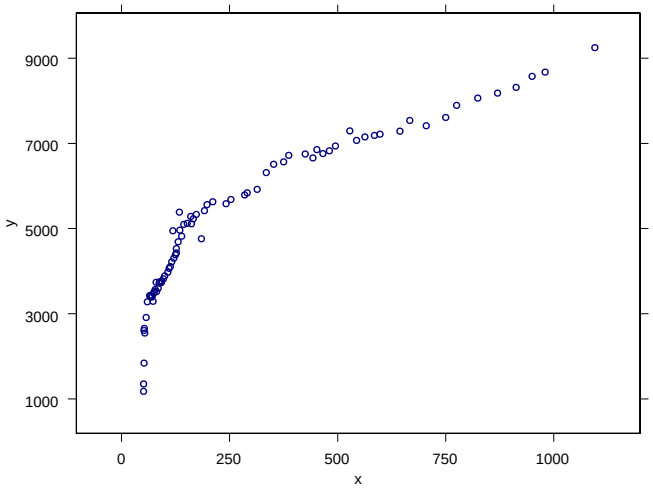

Figura B.35: Diagramas de dispersión de la población Turismos.

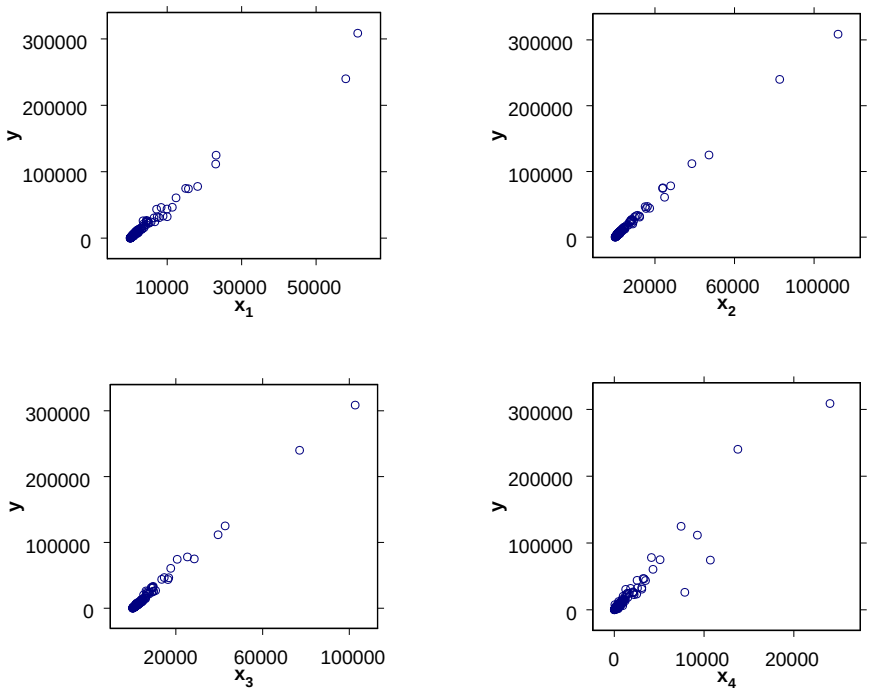

Figura B.36: Diagrama de dispersión de la población ECPF1997.

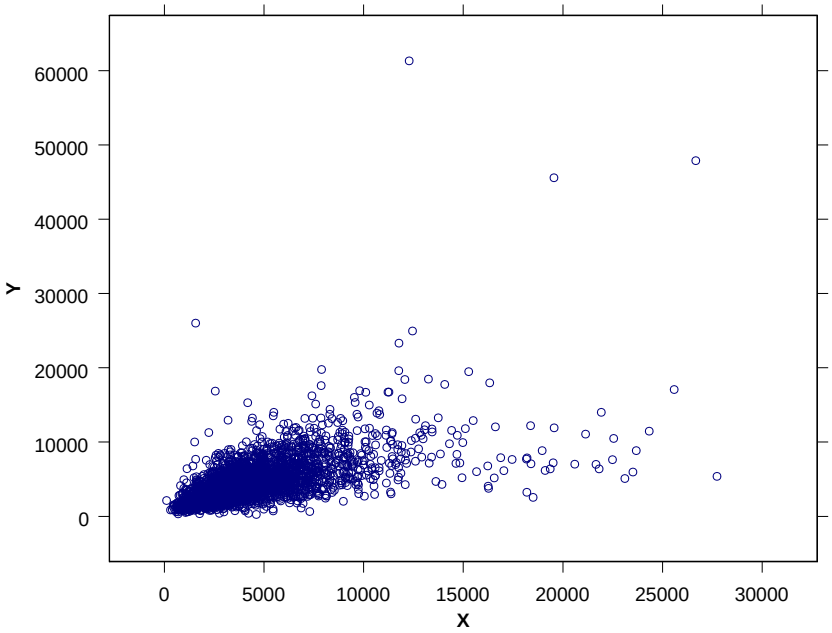

Figura B.37: Diagramas de dispersión de las poblaciones Pop06, Pop07, Pop08 y Pop09

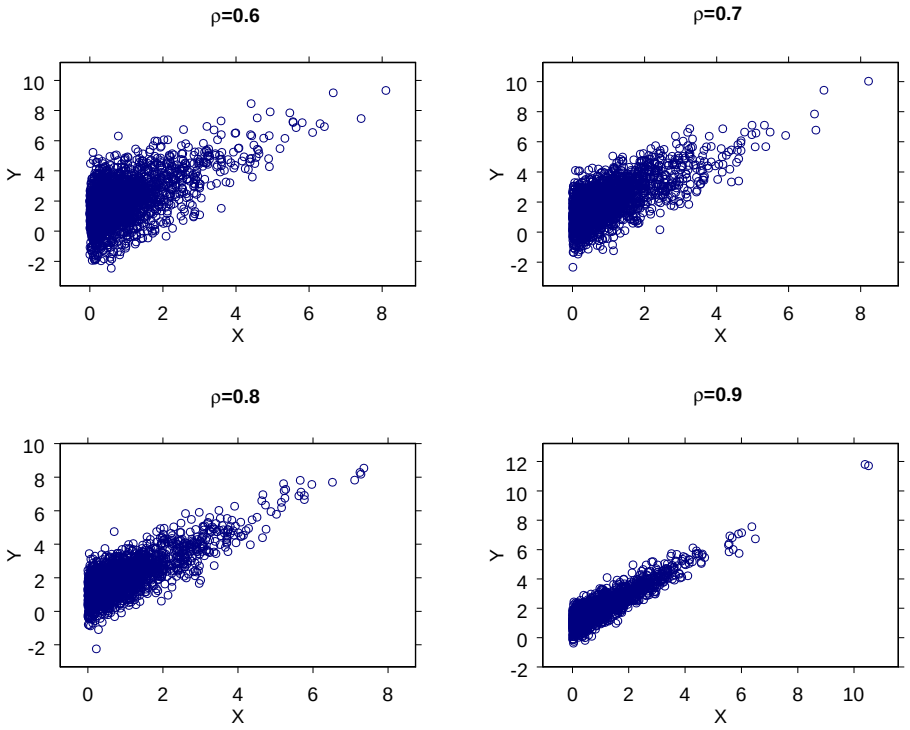

Figura B.38: Diagramas de dispersión de la población Pob098

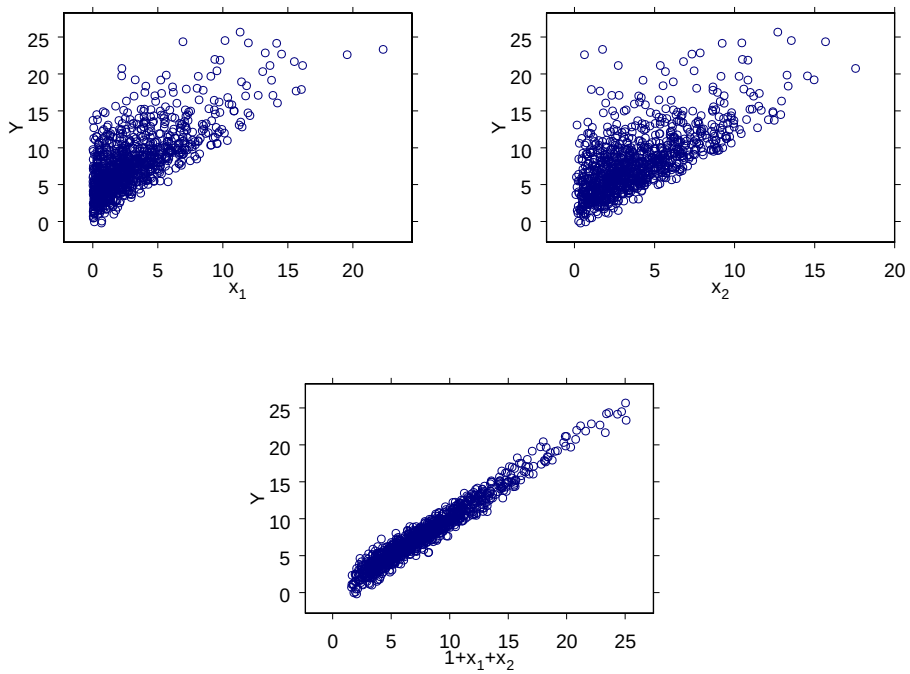

Figura B.39: Diagramas de dispersión de la población Pob080