UNIVERSIDAD DE GRANADA

ESTIMACIÓN DE LA PROPORCIÓN POBLACIONAL EN PRESENCIA DE INFORMACIÓN AUXILIAR

TESIS DOCTORAL

Directores:
Prof. Dr. D. Juan Francisco Muñoz Rosas
Prof. Dr. D. Antonio Arcos Cebrián

Doctoranda:
Encarnación Álvarez Verdejo

DEPARTAMENTO DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA

Granada 2011

Editor: Editorial de la Universidad de Granada
Autor: Encarnación Álvarez Verdejo
D.L.: GR 2392-2011
ISBN: 978-84-694-3598-4

ESTIMACIÓN DE LA PROPORCIÓN POBLACIONAL EN PRESENCIA DE INFORMACIÓN AUXILIAR

Memoria presentada por Encarnación Álvarez Verdejo para aspirar al Título de Doctor con Mención Europea por la Universidad de Granada

DEPARTAMENTO DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA

Fdo. Encarnación Álvarez Verdejo

Vº Bº Directores de tesis:

Fdo. Juan Francisco Muñoz Rosas Fdo. Antonio Arcos Cebrián

UNIVERSIDAD DE GRANADA

Granada 2011

AGRADECIMIENTOS

Estas líneas están dedicadas a todas aquellas personas que me han apoyado y ayudado a realizar la presente tesis doctoral.

Quisiera manifestar mi más profundo agradecimiento a los dos directores de Tesis, el Dr. Arcos y el Dr. Muñoz, por su constante e inestimable aportación a este trabajo y por la confianza depositada en mi. El asesoramiento científico y el constante ánimo que me han transmitido a lo largo de este tiempo han sido fundamentales para llevar a cabo este trabajo, puesto que gracias a su ayuda he adquirido todos mis conocimientos sobre muestreo en poblaciones finitas. Les agradezco también enormemente el tiempo dedicado y la paciencia mostrada durante este periodo, que sin duda han ido más allá de la obligación.

Gracias a mis compañeros del Departamento de Métodos Cuantitativos para la Economía y la Empresa por su apoyo. A Rafael Herrerías y a José Callejón por su inestimable ayuda en mi trabajo como docente.

A mis padres y hermano les dedico esta tesis doctoral, por su apoyo incondicional, por todos y cada uno de los segundos que les he robado para trabajar sobre esta tesis. Por todos los ánimos que me han dado continuamente. Con la esperanza de poder disfrutar algo más de ellos en el futuro.

A ti siempre,

pero hoy más que nunca.

Índice general

Índice de figuras

Índice de Tablas

Capítulo 1

Introducción

Las encuestas por muestreo juegan un papel muy importante en la mayoría de las investigaciones y estudios desarrollados por empresas, instituciones, oficinas de estadística, organismos nacionales, etc. Así, por ejemplo, los principales indicadores demográficos, económicos o sociales de un país están basados en datos muestrales obtenidos a partir de diseños muestrales complejos. En las mencionadas encuestas no sólo se recoge información relacionada con la variable objeto de estudio. Por el contrario, es habitual obtener información auxiliar que permita mejorar la estimación del parámetro en estudio. Esta importancia de la información auxiliar en la fase de estimación ha propiciado que las investigaciones en muestreo en poblaciones finitas de los últimos años se hayan centrado en mejorar la estimación de parámetros en presencia de información auxiliar. Sin embargo, todos los estudios realizados en este sentido están basados en variables cuantitativas, quedando el estudio de variables cualitativas y/o dicotómicas en un segundo plano.

Las variables cualitativas, y en particular las dicotómicas, son también objeto de estudio en la mayoría de las investigaciones por muestreo, especialmente en las ciencias sociales, económicas y de la salud. Al igual que en el caso de las variables cuantitativas, el uso de diseños muestrales complejos y la obtención de información auxiliar son también prácticas habituales. Sin embargo, las técnicas actuales de estimación en el caso de variables cualitativas no tienen en cuenta los aspectos anteriores en la fase de estimación, es decir, tales técnicas de estimación existentes asumen datos extraídos mediante una muestra aleatoria simple en una población infinita y no utilizan información auxiliar en la etapa de estimación. Este hecho no sólo no permite mejorar las estimaciones, sino también puede producir resultados poco satisfactorios tal

como la presencia de importantes sesgos, encarecimiento de los costes de la encuesta para un error fijado de antemano, etc.

En el contexto del muestreo en poblaciones finitas, la estimación eficiente de parámetros ha sido uno de los aspectos más discutidos en los últimos años, tanto por la mejoría en precisión que supone el uso de técnicas más eficientes, como por la reducción de costes en las investigaciones o estudios. En general, son dos las metodologías seguidas para la obtención de estimaciones más precisas y fiables. Por un lado, el uso apropiado de diseños muestrales complejos puede producir importantes beneficios en comparación con esquemas muestrales simples como el muestreo aleatorio simple (en adelante, MAS). En segundo lugar, el uso de información auxiliar también puede producir mejores estimaciones en comparación con los métodos simples de estimación, especialmente cuando la relación entre la variable objeto de estudio y las variables auxiliares sea elevada.

En el campo del muestreo en poblaciones finitas, se han desarrollado numerosos y sofisticados métodos para mejorar la estimación de distintos parámetros asociados a variables cuantitativas. Sin embargo, el problema de la obtención de estimadores de parámetros asociados a variables cualitativas y que posean propiedades deseables es un tema que apenas ha sido discutido en la literatura del muestreo en poblaciones finitas. En concreto, los únicos métodos conocidos para la estimación de parámetros en el caso de variables cualitativas se basan en técnicas tradicionales que en ningún caso utilizan información auxiliar en la etapa de estimación. Destacamos que el principal parámetro objeto de estimación en estas referencias es la proporción de individuos que presentan un determinado atributo.

Teniendo en cuenta las líneas anteriores, en la tesis que se presenta se pretenden lograr los siguientes objetivos:

1. Revisar exhaustivamente la bibliografía más relevante relacionada con la estimación indirecta en poblaciones finitas, la cual está basada en variables cuantitativas.

2. Plantear y definir nuevos estimadores puntuales para una proporción poblacional. Los estimadores propuestos estarán basados en información auxiliar, lo cual puede mejorar el comportamiento de los estimadores propuestos en comparación con los estimadores existentes. En concreto, intentaremos que los estimadores propuestos sean más eficientes y menos sesgados. Debido a las importantes ganancias que obtienen los estimadores basados en información auxiliar en el caso de variables cuantitativas,

resultará interesante conocer el grado de ganancia en el caso de variables cualitativas. Utilizaremos estimadores de tipo razón y regresión, entre otros métodos, para la definición de los nuevos estimadores.

3. Plantear y estudiar otros temas relacionados con la estimación puntual de parámetros, tal como puede ser la estimación de las varianzas de los estimadores propuestos o la construcción de intervalos de confianza.

4. Comparar teóricamente, en términos de varianzas principalmente, los estimadores propuestos con los estimadores ya existentes.

5. Completar el estudio teórico de los estimadores e intervalos de confianza propuestos con estudios empíricos basados en estudios de simulación Monte Carlo que avalen los beneficios obtenidos por los métodos propuestos en comparación con los existentes actualmente en la literatura. Por una parte, los estudios de simulación se basarán en poblaciones simuladas con el fin de abarcar distintos escenarios que pueden presentarse en la práctica. Por ejemplo, consideraremos valores pequeños y elevados para la proporción poblacional que deseamos estimar, así como valores pequeños y elevados para los tamaños muestrales, correlaciones entre las variables, etc. Por otra parte, con el fin de analizar los métodos propuestos en situaciones reales, los estudios de simulación se basarán en datos reales extraídos del ámbito de la Economía y la Empresa.

Esta memoria esta estructurada como sigue:

Este capítulo 1 recoge algunas indicaciones relacionadas con los distintos capítulos de esta memoria. De este modo, en la Sección 1.1 se tratan algunas cuestiones del muestreo en el siglo XXI, en la Sección 1.2 se justifica la importancia del muestreo en la administración de empresas y la economía y se proponen ejemplos de uso de muestreo en ambos escenarios. Finalmente en la Sección 1.3 se presentan numerosos ejemplos de estimadores que hacen uso de la información auxiliar en la etapa de estimación.

En el capítulo 2, en primer lugar, se presenta la nomenclatura y notación básica que se utilizará en el presente trabajo. A continuación, se presentan las aportaciones realizadas en el problema de la estimación de una proporción poblacional mediante estimadores de tipo razón. Estas aportaciones se han realizado tanto en el escenario de un muestreo aleatorio simple como en el caso de muestras extraídas mediante un diseño muestral general. Destacamos que tales estimadores de tipo razón son la base sobre la que se cimienta la investigación central de la presente tesis doctoral, puesto que estos estimadores se utilizarán,

en general, para formular el resto de estimadores propuestos en este trabajo. En este capítulo se ha de destacar el estudio de las propiedades teóricas de los estimadores propuestos y la comparación con el estimador estándar. Además, utilizando esta propiedades se proponen nuevos estimadores de tipo razón con menor sesgo y/o menor varianza. En último lugar, se estudian otros escenarios como la presencia de correlación negativa, la estimación de proporciones pequeñas, etc. La aportación más relevante es la definición del estimador de razón óptimo. Dicho estimador es una combinación lineal de dos de los estimadores de tipo razón descritos en este capítulo. El valor óptimo del peso utilizado en la combinación lineal se determinará mediante el criterio de mínima varianza.

En el capítulo 3 se define el estimador de tipo regresión para una proporción poblacional, tanto para MAS como para un diseño muestral general, y se estudian las propiedades teóricas más importantes. Entre estas propiedades teóricas podemos destacar que posee mínima varianza, es insesgado, coincide, bajo MAS, con el estimador de tipo razón óptimo descrito anteriormente, etc.

Para cada uno de los nuevos estimadores definidos en los capítulos 2 y 3, se estudian desde un punto de vista teórico sus propiedades más importantes. En concreto, se determinan las varianzas reales y estimadas para cada uno de los estimadores definidos. Tales varianzas serán utilizadas para comparar la eficiencia entre los distintos estimadores descritos en este trabajo.

Además de las comparaciones teóricas de los estimadores comentadas anteriormente, se realizan una serie de comparaciones empíricas mediante estudios de simulación Monte Carlo, los cuales están basados en poblaciones simuladas con el fin de abarcar distintos escenarios, así como basados en poblaciones reales del ámbito de la Economía y de la Empresa. Como criterio para la comparación de los distintos estimadores puntuales se utilizarán medidas comúnmente utilizadas como el sesgo relativo y el error cuadrático medio relativo, entre otras. Tanto las comparaciones teóricas como las empíricas muestran que los estimadores propuestos tienen un buen comportamiento en términos de sesgo y error cuadrático medio.

En el capítulo 4 se aborda el problema de la estimación de la proporción mediante intervalos de confianza. Tales intervalos de confianza están basados en los estimadores propuestos en los capítulos 2 y 3. Al igual que en el caso de la estimación puntual, se definen intervalos de confianza en el caso de muestras extraídas bajo MAS y muestras extraídas bajo un diseño muestral general. Se utilizarán varios criterios, como la amplitud media del intervalo, la cobertura empírica, etc, para la comparación empírica de los distintos intervalos de

confianza.

Este trabajo finaliza con un Apéndice donde se describen todas las poblaciones, tanto las simuladas como las basadas en datos reales, utilizadas en los estudios de simulación.

1.1. El muestreo en el siglo XXI

Muestreo "puerta por puerta"

Aunque podemos encontrar referencias de censos y encuestas ya en los relatos bíblicos, las encuestas por muestreo son un fenómeno que se desarrolla rápidamente en el siglo veinte, sobre todo se debió su rápido conocimiento a las encuestas de Gallup y Roper a mediados de los años treinta. En esos momentos la proporción de hogares con teléfono en casa era muy baja y estaba claramente marcada por pertenecer a las familias de altos ingresos.

El mismo sesgo debido a altos ingresos se podía observar en la encuesta por correo llevada a cabo por Literary Digest, que depende de los dueños de teléfono y de los registros de automóviles para su elaboración. La encuesta Literary Digest anunció que Alf Landon ganaría a Franklin Delano Roosevelt en las elecciones presidenciales de EEUU, mientras que Gallup y Roper predijeron de forma correcta que Roosevelt sería el ganador. En ese momento los métodos aplicados por Gallup y Roper adquirieron gran relevancia frente a los métodos usados por Literary Digest que a pesar de ser mayores en número suponían mayor sesgo. Ello evidenció que los tipos de personas seleccionadas eran más importantes que el número de personas seleccionadas.

Muestreo por Cuotas

Los métodos de muestreo que anteriormente se mencionaban aplicados por Gallup y Roper se conocen como "muestreo por cuotas". Dicho método utiliza variantes en la fase de selección de forma geográfica, tales como ciudades, áreas rurales, etc. con probabilidades proporcionales a sus poblaciones. Las grandes ciudades se subdividen un subáreas y se les asigna un entrevistador. Se asignaba un entrevistador por área y además existían otros entrevistadores que podían realizar entrevistas en cualquier zona que deseasen pero debían seguir un control de hombres y mujeres encuestados en cada área, además del número de individuos con altos ingresos y bajos ingresos en cada área.

Además de la encuesta "puerta por puerta"también podían realizar encues-

tas en zonas públicas como por ejemplo parques, calles, puertas de colegios, etc. Ello supuso una mejora en la opinión pública sobre este tipo de encuesta y sus repercusiones en la investigación de mercados.

El muestreo probabilístico.

Al mismo tiempo, justo después del inicio de la segunda guerra mundial, un grupo de estadísticos federales dirigidos por Morrys Hansen se encargaron de llevar a cabo una elaboración de encuestas más controladas para el gobierno americano (Hansen, Dalenius y Tepping, 1985). Un gran reconocimiento de estos trabajos apareció en el artículo de Neyman (1934), que contrastaba el muestreo estratificado con el muestreo por cuotas. Estos procedimientos, los cuales se denominan muestreos probabilísticos por áreas, seleccionaban geográficamente áreas en bloques o en segmentos de pequeño tamaño, con probabilidades proporcionales a la población estimada, y después se realizaban selecciones con probabilidades iguales de hogares de esos bloques o segmentos previamente seleccionados (Cochran, 1977; Hansen, Hurwitz y Madow, 1953; Kish, 1965; Sudman, 1975).

El muestreo probabilístico por áreas elimina las posibles influencias de los entrevistadores sobre la muestra y hace posible especificar la probabilidad de selección para cualquier hogar en la población.

Gallup, Roper, y los investigadores de mercado eran más reticentes a adoptar el muestreo probabilístico por áreas por la sencilla razón del elevado coste que supone si se compara con el muestreo por cuotas. Desde la perspectiva de la sencillez, el muestreo probabilístico por áreas requiere el uso sustancial de mapas así como listín de los bloques de hogares seleccionados o segmentos. Por otra parte, desde la perspectiva de las influencias que ejercen los entrevistadores, el muestreo probabilístico por áreas es más costoso dado que las rellamadas o la repetición de las visitas que se requiere es alta en algunos sectores de hogares que es difícil encontrarlos en sus hogares. Los usuarios del muestreo por bloques generalmente afirman que los resultados que se obtienen son tan satisfactorios como los que ofrecen otros métodos más costosos, y a menudo con las comparaciones que se realizan entre estos métodos es difícil encontrar diferencias significativas que justifiquen el uso de los más costosos.

El principal empuje de los muestreos probabilísticos tuvo lugar después de las elecciones americanas de 1948 cuando la mayoría de las encuestas predijeron de forma errónea una victoria de los republicanos capitaneados por Thomas Dewey sobre el otro candidato Harry Truman. Las encuestas fueron duramente criticadas por sus procedimientos en lo referente a la cuota de muestreo emplea-

da, aunque más tarde los análisis demostraron que los mayores errores fueron debidos al paso de las encuestas demasiado pronto lo cual supuso grandes cambios con respecto a los últimos minutos del día de la votación, de tal forma que la preferencia que lideraba Truman no se observó en las encuestas de a tiempo.

Incluso después de que el muestreo probabilístico por áreas se convirtiese en muestreo estándar, algunos formatos de muestreo por cuotas se seguían llevando a cabo. Gallup y muchos otros conmutaron a métodos que especificaban las características del hogar con el que se había contactado pero ello requería únicamente de un intento. Si en el momento de la encuesta no se encontraba a nadie en el hogar, se contactaba con el siguiente domicilio, es decir estaba sujeto únicamente a mantener las cuotas en lo referente al género. Este método es todavía usado de forma común en Gran Bretaña y en algunas zonas de Estados Unidos, donde los investigadores de mercados consideran que la encuesta "persona a persona"es algo fundamental dentro de sus estudios. La mayoría de problemas que este tipo de encuestación genera es la sobreestimación del número de personas que se pueden encontar en sus hogares, los cuales suelen ser de forma general personas mayores y/o desempleadas. Incluso este inconveniente puede ser solventado mediante el uso de cuotas basadas en la accesibilidad, tales como los métodos "prob-cuota"que se usaron en los sesenta (Sudman, 1966).

Aún así se ha de destacar que hoy en día se le presta muy poca atención al muestreo por cuotas, ya que las encuestas "puerta a puerta"se han convertido en algo poco usado, exceptuando grandes encuestas realizadas por los gobiernos.

Encuestas telefónicas

En las últimas décadas, la mayoría de los métodos usados para obtener estudios de mercado u otro tipo de datos muestrales han sido las encuestas telefónicas. En particular esto se debe a dos motivos fundamentalmente, en primer lugar el muestreo telefónico es una alternativa mucho más atractiva en términos de porcentage de hogares con teléfono desde la segunda guerra mundial, cuando se incremento desde el cincuenta por cierto hasta un noventa por ciento de hogares con teléfono, y ahora se encuentra en torno a dicho valor. El sesgo debido a la falta de respuesta por determinados hogares que no poseen teléfono se ha convertido en una preocupación menor para los investigadores de mercados, pero no ocurre así para las investigaciones llevadas a cabo por los gobiernos ya que en la mayoría de los casos dichos hogares suelen ser hogares con niveles de ingresos bajos y quedan excluidos de la representación del total

del país. La segunda razón por la cual se han incrementado tanto las encuestas teléfonicas es el mayor coste que suponían las encuestas "puerta por puerta", este incremento del coste de las encuestas "puerta por puerta"se debió sobre todo a la falta de respuesta y la disminución de encuestados. El rápido incremento in el porcentage de mujeres trabajadoras había supuesto un consumo de tiempo en sus vidas y una menor accesibilidad para los encuestadores, incluso cuando las entrevistas se hacían en fines de semana o en horario nocturno. Con todo esto, es normal suponer el alto coste fijo de desplazamientos y viajes que no respondían con los resultados esperados en número de encuestas. Las encuestas telefónicas eliminaron todos los costes de desplazamiento y esto supuso un ahorro de más de la mitad.

Marcación aleatoria de números

Inicialmente, los directorios telefónicos se usaban como marcos muestrales para muestreo por teléfono, pero se encontró que el número de hogares que no aparecían en dicho marco muestral era bastante elevado, al menos en áreas urbanas (Glasser y Metzger, 1972).

Esto lideró el uso de los métodos de marcación aleatoria de números a partir de una lista en las cual se elegían los cuatro últimos número de forma aleatoria. El uso de la marcación aleatoria solventó pues el problema de los listines telefónicos que no tenían todos los números de teléfono, pero aún así seguía siendo ineficiente ya que solamente el veinte por ciento de los números aleatorios eran números de hogares.

La siguiente mejora vino de la mano de un nuevo método de selección de números y tras esto generar números aleatorios cambiando las últimas cifras. Esto mejoró la eficiencia de la marcación aleatoria de números sustancialmente. Este método era altamente sesgado, ya que la probabilidad de que se tratase de un número de un banco dependía del número de líneas que el banco tuviese en el listín telefónico que estaba usando.

Desde entonces, dos métodos han sido ideados para mejora de la eficiencia de las muestras en el uso de marcación aleatoria. El procedimiento más ampliamente usado, en especial por los investigadores de mercados, es la compra de listas de números de teléfono de una o varias fuentes. Estas listas son generadas mediante las llamadas a todos los números de los listines telefónicos del país y tras esto realizar una selección sobre los números seleccionados y ordenados. Tras esto se toman las listas del personal que trabaja en varias empresas y se eliminan todas aquellas que coinciden, de tal forma que en muchos casos se observó que la mayoría de los números pertenecían a personal de la empresa.

Otra vía para obtener listas pero evitando en todo momento la presencia de huecos o blancos en las mismas, se basa en el conocido método de Waksberg-Mitofsky (Waksberg, 1978).

Bajo este método la llamada inicial se realiza eligiendo el número de teléfono de un listín actualizado, tras esto si la llamada resulta ser un hogar las siguientes llamadas se harán dentro del mismo grupo de números hasta el número especificado de antemano.

Este método tiene una ventaja muy importante ya que asegura que todos los hogares tienen la misma probabilidad de pertenecer a la muestra y además elimina los posibles problemas generados por llamadas telefónicas a bancos, empresas, etc, pero también requiere algunos costes más.

El muestreo de los individuos que componen un hogar

Otra área de desarrollo en el muestreo telefónico ha sido la referente a el muestreo dentro de cada hogar.

Por razones de composición de hogares y la diversidad de las muestras, la mayoría de las encuestas usan solamente un único individuo de cada hogar, incluso cuando la información concierne a otros individuos. Esto puede crear sesgo en la muestra en favor de los individuos de hogares que son pequeños ya que en un hogar que está compuesto por un único individuo dicho individuo tiene probabilidad total de pertenecer a la muestra, mientras que en un hogar que esté compuesto por más de un individuo, la probabilidad de un individuo n de ser elegido para la muestra es proporcional al número de individuos que componen ese hogar esto es $1/n$, por lo tanto y para dar a cada individuo la misma probabilidad de pertenecer a la muestra, Kish (1949) desarrolló un procedimiento que requería un listín en el cual aparecieran todos y cada uno de los miembros que componían un hogar y de ahí se selecionaba un miembro según las normas de la denominada "Tabla de Kish", este procedimiento llegó a ser el procedimiento estándar para muestreo dentro de los hogares denominado "puerta por puerta".

En las encuestas telefónicas el método de Kish se demostró que era menos eficaz ya que los encuestados no se podían ver mientras se realizaba la encuesta y además estaban menos dispuestos a dar una relación del número de personas que componían el hogar. Una alternativa a dicho método fué usar el método propuesto por Troldahl y Carter en 1964, dicho método implementa su mayor ventaja en que solo necesita dos tipos de información, el número de adultos en el hogar y el numero de hombres (o el número de mujeres). Un individuo se selecciona rotando entre cuatro tablas distintas. Este método elimina la

necesidad de listados sobre el hogar entero, pero genera sesgo ya que no todos los miembros del hogar tienen la misma probabilidad de selección dentro de las cuatro tablas y porque las disparidades en el la composición del género de los adultos solteros que viven solos.

Un método posterior consistía en seleccionar aquel miembro de la familia con la fecha de su cumpleaños más reciente (O´Rourke y Blair, 1983). Este método aunque insesgado en teoría presenta algún sesgo en la práctica ya que las personas que contestaban al teléfono al inicio de la llamada reclamaban ser ellos mismos quienes respondiesen las baterías de preguntas que componían la encuesta. No obstante, este método es ahora uno de los más ampliamente usados en las encuestas por teléfono ya que es muy fácil de gestionar.

Otros acontecimientos históricos: las encuestas por correo

En determinadas poblaciones, en las cuales se dispone de una buena lista de marco poblacional y por algún motivo es fácil obtener cooperación por parte de los encuestados, es factible y útil utilizar encuestas por correo. Sin embargo, las encuestas por muestreo no son comunes en la investigación de mercados o para poblaciones generales de hogares, la principal causa a la que se debe este hecho es la baja cuota de respuesta y por tanto el sesgo que aparece. Muchas investigaciones se han centrado en la mejora de las encuestas correo y ahora es posible describir algunas de la prácticas "mejor valoradas"(Dillman, 1978), pero aún así la baja tasa de respuesta y el requerir la voluntariedad del entrevistado siguen siendo frenos a su uso.

El uso de paneles para correo comercial que se suelen emplear en diversas formas para medir el comportamiento de compra parece ser una excepción a esta generalización. Estos generalmente consiguen altas tasas de respuestas para sus encuestas individuales debido a que los miembros que componen el panel han dado previamente su consentimiento a participar en dicha encuesta.

Las peticiones iniciales de invitación a partipación en el panel, sin embargo suelen tener baja tasa de respuesta en lo que a cooperación de se refiere.

La ventaja más importante de los paneles por correo no reside en su respuesta última sino en la repetición de dichas encuestas que hacen un marco perfecto para conseguir características demográficas mejor detalladas y un punto de referencia para su desarrollo frente a otras vías de recolección de datos.

Encuestas en centros comerciales

Las encuestas en centros comerciales, han incrementado su popularidad so-

bre todo en la investigación de mercados en las últimas décadas, ya que el número de centros comerciales se ha incrementado y la proporción de consumidores que frecuentan este tipo de comercios se ha incrementado hasta en un noventa por ciento. Cuando los centros comerciales se usan para llevar a cabo investigaciones de mercados, o más concretamente para experimentos de marketing, tanto el muestreo entre los centros comerciales como el muestreo dentro de cada centro comercial, es a menudo realizado con numerosos fallos y errores debidos en su mayor parte a descuidos. No obstante, es posible utilizar procedimientos de encuestación estándar para seleccionar cuidadosamente muestras de centros comerciales y de forma similar muestras dentro de esos centros comerciales.

La selección de los centros comerciales se realiza preferentemente desde una lista de todos los centros comerciales disponibles. El mayor problema que se puede presentar es que muchos centros comerciales no permiten las entrevistas en sus locales y por tanto esto impide respetar las probabilidades muestrales. Lo mejor en estos casos sería realizar un muestreo por localización del centro comercial y sustituir aquellos centros comerciales que han negado su cooperación por otros que sí deseen cooperar.

Dentro de los centros comerciales es posible realizar muestreo aleatorio únicamente eligiendo a las personas que entran o salen del centro comercial. Sudman en 1980 propuso procedimientos para realizar dicha elección, en los cuales los visitantes que entraban o salían por determinadas puertas durante unos periodos de tiempo determinados eran tratados como cluster y muestreados de acuerdo con esta consideración.

Se ha de tener en cuenta una consideración importante en lo referente a las unidades de la población que son objeto de estudio. Aunque casi todo el mundo ha visitado un centro comercial en numerosas ocasiones, hay grandes disparidades entre el número de veces que diferentes individuos se pueden encontrar al mismo tiempo en el centro comercial.

Si un individuo es muestreado y no visita el centro comercial muy a menudo, obviamente se deberá determinar la frecuencia de esas visitas para obtener el peso inversamente proporcional de estas frecuencias (Blair, 1983).

Focus Groups y otros métodos de muestreo

Además de todos los desarrollos que se han llevado a cabo para la mejora de la calidad y eficiencia de los métodos de muestreo, también se ha incrementado el uso de los denominados "focus groups". Los focus groups están considerados una excelente herramienta para la mejora del conocimiento que se tiene sobre

el mercado, pero también se ha de tener en cuenta que un grupo de diez o más personas elegidas al azar y en una única localización no se debe esperar que sea representativo del total de la población que componen todos los consumidores existentes. Si nos centramos en los focus groups llevados a cabo en numerosas localizaciones se puede reducir el sesgo potencial, pero sigue siendo insuficiente para eliminar el sesgo totalmente.

Un último tipo de muestreo, sería aquel que se desarrolle en base a encuestas que están presentes en los medios, tales como periódicos, revistas, televisión, y más recientemente la Word Wide Web. Los encuestados que hayan visto o escuchado con anterioridad algo sobre el tipo de encuesta serán invitados a enviar sus cuestionarios, no por correo postal o por llamada telefónica, sino que lo contestarán en el mismo momento a través de un ordenador. Incluso si miles de encuestados completaran los cuestionarios, en los resultados es inevitable la pérdida de información ya que obviamente este conjunto de encuestados representa una proporción minúscula si lo comparamos con el total de población objetivo. Si se compara de forma exahustiva los distintos tipos de muestreo que se pueden realizar al mismo tiempo se observa que los resultados son más sesgados no solo porque existe una gran diferencia en variables demográficas sino porque depende en gran medida del interés del tópico sobre el que se pregunta. Aquellos encuestados que tienen alto conocimiento sobre la temática de la encuesta son más proclives a contestar que otros que tengan un conocimiento medio o bajo sobre la misma, o incluso aquellos que no tienen ningún conocimiento sobre el tema que son más rehacios a contestar.

Los medios de comunicación han reconocido que se encuentran mayores problemas con estas encuestas, a pesar de que su proliferación ha sido rápido debido principalmente a que los usuarios de los medios de comunicación encuentran los resultados de alto interés e ignoran más lo concerniente al método usado.

Muestreo de poblaciones raras

En las últimas décadas se ha llevado a cabo un incremento significativo con respecto al muestreo de poblaciones muy específicas en detrimento del muestreo de la población general. Ejemplos de esta afirmación se pueden presentar por ejemplo en los muestreos de hogares con altos ingresos, por razas, por grupos étnicos tales como los afroamericanos o hispanos, o grupos de personas con una determinada enfermedad como puede ser el cáncer o el asma, o personas que forman parte de un segmento de mercado muy determinado, aquí podríamos poner de ejemplo la pesca recreativa de alta mar.

Para muestrear estas poblaciones raras o específicas, ya que constituyen una proporción ínfima del total de población, lo primero que se han de tener en cuenta si existe un listín adecuado que esté disponible para el estudio. Esto incluye la posibilidad que el pertenecer a un determinado grupo se pueda identificar de forma correcta, aquí podríamos mencionar por ejemplo disponer de datos de la encuesta de población global o bien de tipo panel.

Si esto ocurre, el muestreo y la localización de encuestados resulta sencillo. Sin embargo, esto es la excepción que confirma la regla. En la mayoría de los casos, la selección de partes de la población total es necesario., y el coste de dicha selección puede incluso exceder el coste de las entrevistas.

Los procedimientos que se han desarrollado en los últimos años han reducido sustancialmente costes derivados de situaciones en las cuales la población estaba agrupada geográficamente formando clusters, como ocurre generalmente con los ingresos y grupos étnicos. En estas situaciones puntuales, existe una gran proporción de segmentos geográficos en los cuales los individuos que los componen no forman parte de las denominadas "poblaciones raras", y otros muchos pueden tener algún tipo de relación con dichas poblaciones.

La mayor eficiencia en la selección dentro de la población total puede venir acompañada de la eliminación de determinados segmentos que no van a ser muestreados. Sudman (1985) propuso una adaptación de la marcación digital aleatoria de Mitofsky-Waksberg para lograr este objetivo en las encuestas realizadas. Inicialmente se llama a un número de teléfono entre un grupo de cien números. Si estos números se incrementan de la población objetivo, entonces se ha de realizar otra selección o limpieza en dicho grupo hasta que se obtenga un tamaño del grupo óptimo, normalmente suele estar entre cinco y diez el óptimo de dicho tamaño.

La mayor proporción de segmentos con cero miembros de la población objetivo, representará mayor ahorro en el coste final. Generalmente es mucho más complejo realizar una selección en poblaciones raras donde además no se encuentran grupos geográficamente distribuidos. Dos métodos desarrollados últimamente son los que presentan mejores resultados tanto en eficiencia como en muestreo. En determinadas situaciones, tales como encuestas a poblaciones muy específicas, es natural llevar a cabo la encuesta "in situ", en lugar de realizar encuestas en los hogares. Otros ejemplos que se pueden presentar con características similares podrían ser compradores de determinadas tiendas o centros comerciales, empleados de una marca concreta, pasajeros de un avión o espectadores de un determinado juego o conciertos o eventos deportivos.

Otras formas de mejorar la eficiencia de la selección es a través del denominado "muestreo de redes"el cual incrementa el total de información obtenida en un filtrado de entrevistas. Si se considera una encuesta que trata de obtener información de aquellos individuos que han sido diagnosticado de cáncer. Bajo el típico formato del filtrado de la población, un encuestado podría ser cualquier miembro del hogar al que se le pregunte si existe alguien en su hogar con cáncer y de esta forma cualquiera podría se encuestado dentro de un hogar determinado. En esta red de muestreo, obviamente se incluyen aquellas personas fuera del hogar que son conocidas por el encuestado y que puedan tener cáncer. Esto podría incluir por ejemplo amigos cercanos, familiares, compañeros de trabajo y vecinos del barrio. Para que el muestreo en red sea válido se han de presentar dos condiciones, en primer lugar que el encuestado disponga de la información necesaria para informar a cerca de todos los miembros de la red y en segundo lugar que se pueda obtener de forma precisa una estimación del tamaño de la red de tal forma que los datos obtenidos puedan ser bien determinados en términos de peso de la información sobre el total de información.

Desde el tamaño de la red, que puede variar de un encuestado a otro y ya que la probabilidad de ser identificado es directamente proporcional al tamaño de la red, es necesario conocer los tamaños de los pesos de cada miembro de la red de una población rara. Empíricamente se ha demostrado que las condiciones o requerimientos para que el muestreo en red funcione correctamente, son básicamente dos requisitos, en primer lugar que es mejor cuando la red está formada por parientes cercanos quienes probablemente saben más acerca de los miembros de la población rara y cuando pueden informar con precisión sobre el tamaño de la red.

La ponderación de cualquier muestra aumenta la variabilidad del muestreo, pero la reducción en los costes de selección relacionados con el muestreo de la red es tan grande que el procedimiento sigue siendo altamente rentable.

Gran parte de los primeros trabajos sobre la red de muestreo se realizaron por Sirken y algunos autores más (Levy, 1977; Sirken, 1970; 1972; Sirken y Levy 1974). Un breve resumen sobre este tema se encuentra en Sudman (1985).

Procedimiento CATI: Computer Asisted Telephone Interviewing

Es un sistema que utiliza conjuntamente el teléfono y el ordenador. Una vez seleccionado aleatoriamente el número de teléfono mediante el ordenador y establecido contacto con el mismo, el entrevistador realiza las preguntas de un cuestionario informatizado para grabar al mismo tiempo las respues-

tas obtenidas del entrevistado en una base de datos. A medida que se graba una respuesta, el ordenador guía y muestra la pregunta que sigue, observando así las instrucciones y saltos que se hayan establecido previamente, comprobando además la coherencia de las respuestas. De esta forma se reduce el tiempo necesario para recoger, supervisar y tratar los datos. Es evidente que el ahorro de esfuerzos en estas tareas tan repetitivas y tediosas, permitiendo tener resultados del análisis de forma casi instantánea. Efectivamente el CATI proporciona una gran flexibilidad e individualiza los cuestionarios reduciendo los malentendidos, lo que junto con la velocidad que suministra a la investigación ha hecho que tenga una gran aceptación entre los institutos de investigación. Sin embargo, no es apropiado para introducir preguntas abiertas. En la selección aleatoria de los números de teléfono se observarán determinados requisitos para que los prefijos o códigos barajados se correspondan con los de la población objeto de la encuesta.

CAPI: Computer Asisted Personal Interviewing

Es una estrevista autoadministrada por ordenador; el papel del estrevistador lo asume completamente el ordenador, por lo que el cuestionario ha de ser sencillo y el programa fácil de manejar. El entrevistado responde al cuestionario con la ayuda de un teclado o un ratón mediante el cual se recogen e informatizan las respuestas. El programa proporcionará los mensajes y las ayudas convenientes para facilitar la tarea y así conseguir una comunicación interactiva que intervenga en la calidad de los datos recogidos. Con este sistema el interés por colaborar es grande; a esto contribuye el efecto novedad que siempre supone un fuerte estímulo. Precisamente la innovación que supone es un factor de estímulo para obtener una alta tasa de respuesta.

Paneles

Los paneles son individuos, hogares, organizaciones o cualquier tipo de entidad que informa sobre un mismo tópico durante un periodo de tiempo determinado. El mayor beneficio que ofrecen los paneles es que reduce significativamente los errores de muestreo cuando una de las medidas cambia.

La utilidad de los paneles fue reconocida por la compañía A.C. Nielsen, la cual empezó con los paneles en las tiendas de comestibles durante los años treinta y su objetivo era medir las ventas durante dicha década. Esto fue rápidamente copiado por otras empresas de investigación de mercados, la primera en copiarlo fue Market Research Corporation America, la cual estableció su panel de consumidores (Sudman y Ferber, 1979). Actualmente, los escáner y los paneles por correo se usan para medir una amplia gama de comportamien-

tos del consumidor. También son medidas a diario las audiencias televisivas o radiofónicas, y otros tipos de paneles para obtener patrones de comportamiento y educacionales.

El mayor desarrollo en la investigación por paneles sin duda se ha producido en las últimas décadas y no ha sido únicamente sobre los procedimientos de muestreo aplicados, que en efecto han cambiado un poco, sino más bien sobre la naturaleza y tamaño de dichos paneles. Baratos y de rápido tratamiento informático, han hecho muchísimo más fácil el incremento del tamaño de muestra y el número de variables que se estudian en dichos paneles.

A su vez también se ha de resaltar que los paneles han mejorado las tecnologías para la medición y son capaces de recoger de la forma más idónea posible datos sin mucho esfuerzo por parte de los encuestados.

El muestreo en la empresa

Al igual que con la investigación por paneles, la evolución en las primeras fases de la investigación empresarial desarrollada durante los últimos veinte años se han visto afectados tanto por el volumen de investigación como por los métodos de recopilación de datos. Se ha producido un aumento sustancial en el volumen de investigación, especialmente en las áreas relacionadas con la satisfacción del cliente y la evaluación de nuevos productos, y se han evolucionado de forma satisfactoria procedimientos tales como las encuestas por correo electrónico. Sin embargo, las cuestiones y procedimientos más importantes del muestreo no han cambiado. Hay varias cuestiones que distinguen a las muestras en la empresa de las muestras de consumidores. Con una diferencia muy importante, el más importante es la enorme variabilidad en el tamaño de las empresas. Un investigador tratando de obtener una estimación del potencial de demanda de un producto comercial o un nuevo servicio, pronto reconoce que la demanda está muy afectada por las grandes empresas. Los procedimientos óptimos de muestreo estratificado en las empresas se basan en una afijación proporcional a la desviación estándar de la variable en el estrato. Dado que existe una correlación muy alta entre la desviación estándar y tamaño, en la práctica, estos medios de muestreo suelen tomar por una medida de tamaño de la empresa en la mayoría de las ocasiones las ventas. Otra simplificación se presenta en las grandes empresas que representan la mayor parte de las ventas y la varianza en una categoría depende del número de grandes empresas que formen parte de la competencia en esa categoría (Hansen et al. 1953; Sudman 1975). Una segunda cuestión referente al muestreo es decidir cuál es la unidad de muestreo adecuada dentro de la empresa en estudio. Podría ser una planta, un estudio de oficinas en una región determinada o una línea de negocio. La

elección depende en primer lugar del tema del estudio y si las políticas varían dentro de la empresa. Una tercera cuestión es determinar quién dentro de la organización es el informante adecuado o por el contrario se considera que son necesarios varios informantes para proporcionar resultados precisos. Mientras que obtener la cooperación de las empresas que van a participar en la encuestas requiere habilidad y persistencia, un factor relevante es si la encuesta es importante para el encuestado.

Si se puede acceder al informante idóneo, la probabilidad de la cooperación y la calidad de los datos que se obtendrán son significativamente mejores. Si los datos son recolectados en entrevistas personales, teléfono, correo electrónico, centro comercial, fax, o correo postal, lo esencial es encontrar al informante idóneo. Esto se hace generalmente por teléfono, a partir de la centralita telefónica de la empresa y se pasa desde un teléfono a otro hasta que se consigue contactar con la persona adecuada. Incluso si existe una lista y está disponible, es necesario realizar una comprobación porque las listas de las empresas quedan obsoletas de forma rápida debido a la gran volatilidad de sus plantillas de empleados.

1.2. Aplicaciones del muestreo en la economía y en la administración de empresas

La estadística para los negocios es la *recolección, la organización, el análisis y la elaboración de informes sobre resultados numéricos importantes en una situación o para la toma de decisiones en los negocios.* Es obvio que debido a la gran variedad y diversidad de negocios y métodos de venta disponibles las aplicaciones de la estadística se pueden dar en muy diversos escenarios. (Weiers, 2006).

Lo que sí constituye una realidad hoy en día es que en el entorno actual de la administración y la economía globales disponemos de gran cantidad de información estadística. Los mejores administradores y ejecutivos son quienes pueden comprender la información y usarla eficazmente.

En el mundo empresarial y económico resulta de vital importancia la información, la cual se convierte en una fuente de riqueza y en una ventaja competitiva usada de forma idónea. Para esta recopilación de la información, para su tratamiento y para obtener resultados de la misma es preciso un conocimiento profundo de las técnicas de muestreo, de técnicas de encuestación y de

paquetes informáticos para el tratamiento de datos.

La empresa siempre busca en sus investigaciones reducir al máximo los costes, por ello un cuestionario mal redactado, supone una pérdida de recursos y de tiempo ya que los datos obtenidos no serán de ayuda, o por ejemplo un número de encuestados excesivamente alto supone un gran desembolso también, si la encuesta se realiza a un número insuficiente supondrán resultados poco fiables para posiblemente decisiones de gran importancia estratégica o económica. Con todo esto, queremos reseñar la importancia del muestreo estadístico en todas las fases de la investigación empresarial.

Si bien es cierto que hoy en día la información se puede obtener de bases de datos secundarias que al no ser investigaciones "$ad-hoc$" tienen grandes limitaciones para su tratamiento dentro de la empresa, también es cierto que se pueden contratar con distintas agencias externas en lugar de realizarlas dentro de la propia empresa.

La evolución del muestreo dentro de la empresa ha sido en parte consecuencia de los cambios en el entorno empresarial y económico.

En el pasado, los propietarios de la empresa, eran a su vez los directivos de la misma y encargados de la gestión empresarial. Por ello, estas labores de investigación empresarial, de mercados labores de auditoria las realizaban ellos mismos, lo que se podría denominar el *autocontrol de la función directiva*.

A medida que fué creciendo la complejidad de las empresas en sí mismas y en el entorno de las mismas se fué imponiendo un nivel de competencia y evolución constante, las distintas divisiones del trabajo dentro de la empresa se fueron distribuyendo y especializando por departamentos en cada área en concreto. De tal forma, que a medida que aumentó la complejidad de las empresas se fué imponiendo una división del trabajo que tenía por objeto la especialización de los empleados y de sus funciones. Esto exigía una organización que adecuara los medios necesarios a la consecución de los fines deseados. El crecimiento en el volumen e importe de las operaciones, la mayor dispersión de los activos y, en general, la descentralización de las actividades comerciales y de fabricación han contribuido a un distanciamiento de la Dirección en el control de las operaciones.

Si tal y como comentábamos anteriormente a esto se le añade la modernización y mejora producida en los medios de tratamiento y procesamiento de datos, al sustituir los medios manuales por los sistemas informáticos, en donde diferentes transacciones se procesan para producir estadísticos, informes, facturaciones, gestión de inventario, contabilidad, etc., se llega a la conclusión

de que, efectivamente, el mundo empresarial ha evolucionado y que, en consecuencia, dicha evolución debe ir acompañada de los medios necesarios para garantizar la buena gestión empresarial.

La evolución de las empresas y del entorno de las mismas, que responde al incremento de la complejidad de los negocios con la correspondiente delegación de facultades, exige la implantación de los controles necesarios para conseguir que las responsabilidades delegadas por los directivos y propietarios se conserven íntimamente unidas a los mismos. Por este motivo las empresas establecen planes de organización y un conjunto de métodos y procedimientos que aseguren que los activos están debidamente protegidos, que los registros contables son fidedignos y que la actividad de la entidad se desarrolla eficazmente y se cumplen según las directrices marcadas por la Dirección.

Entre las aplicaciones económicas del muestreo se pueden destacar las siguientes:

Encuestas de presupuestos familiares

En las que las características fundamentales a estudiar son los ingresos y gastos medios familiares y el porcentaje de gastos en diversos artículos o grupos de estos. Las finalidades de estas encuestas pueden ser el análisis de de la demanda, la obtención de ponderaciones actualizadas para el cálculo de los índices de precios, ahorro de familias, etc.

Desde que el INE realizó la primera Encuesta de Presupuestos Familiares (EPF), estas operaciones estadísticas han sido objeto de múltiples cambios metodológicos, que han afectado a todos los aspectos relacionados con este tipo de investigaciones, habiendo adoptado incluso, diversas formas en cuanto a su periodicidad se refiere. No obstante, a pesar de las diferencias entre las sucesivas metodologías, todas las EPF tienen en común que suministran información sobre la naturaleza y destino de los gastos de consumo, así como sobre diversas características relativas a las condiciones de vida de los hogares. Tradicionalmente se han venido realizando dos tipos de EPF, las estructurales o básicas, cada ocho o diez años y las coyunturales o trimestrales.

En 1997 se implantó por primera vez una EPF que trataba de aglutinar los aspectos más positivos de los dos tipos de operaciones que convivían hasta ese momento con el fin de responder a todas las necesidades de los usuarios. De esta forma, se integró en una sola investigación los dos tipos de encuestas realizadas hasta esa fecha.

En los años transcurridos desde 1997 hasta ahora, se han ido generan-

do nuevas exigencias por parte de los diferentes usuarios, así como diversas recomendaciones metodológicas provenientes de los distintos foros internacionales, y de la oficina de estadística de la Unión Europea (EUROSTAT) en particular. Todo ello, unido a la exigencia lógica de toda encuesta permanente de revisar los principales elementos metodológicos que la caracterizan, ha hecho necesario el cambio metodológico cuya principal línea directriz es asegurar la máxima calidad de la información proveniente de la nueva operación estadística. (www.ine.es)

Investigación de mercados

Especialmente se usan en los campos siguientes:

- Estructura del mercado que permite conocer la población por sexo, edades, clases socio-económicas, zona rural, urbana, etc.
- Preferencias en el consumidor para conocer los hábitos, necesidades, gustos, experiencias y conocimientos sobre los productos de venta por parte de los consumidores efectivos o potenciales.
- Técnicas de venta cuyo objetivo es conocer o asegurar el lanzamiento al mercado de nuevos productos, estudios de la motivación y de publicidad.

Encuestas de empresas y establecimientos

En las que interesa conocer características de ellas como son los beneficios, dividendos, sueldos, existencias, equipos, instalaciones, etc. Así como opiniones sobre la situación presente y futura.

Intervención o auditoria y contabilidad

Las empresas de contabilidad pública emplean procedimientos estadísticos de muestreo para llevar a cabo auditorías a sus clientes.

Los auditores comienzan hoy a aplicar en su trabajo, con mayor frecuencia, ciertos conceptos estadísticos que les permiten establecer una base más científica para llevar a cabo las pruebas de auditoria. La incorporación de estos nuevos medios no altera los procedimientos de auditoría que requiere aplicar en cada circunstancia, entendiéndose que mejora la forma de determinar el alcance del trabajo y proporciona evaluaciones cuentificadas matemáticamente de la evidencia obtenida.

De la revisión exhaustiva del auditor de cuentas ha pasado a la prueba selectiva. El auditor tiene a su disposición las técnicas de muestreo estadístico,

en forma limitada pero valiosa, para que puedan emitir juicios científicamente más justificables acerca del resultado de las pruebas de auditoria. (De Agustín Melendro, 1995).

Finanzas

Los asesores financieros recurren a una gama de información estadística para guiarse en sus recomendaciones de inversión. En el caso de las acciones, revisan una variedad de datos financieros, que incluyen relaciones del precio con el redimiento y los dividendos.

Mercadotecnia

Los escáneres en las cajas de los almacenes al detalle se emplean para reunir datos que tienen muchas aplicaciones en la investigación de mercados. Por ejemplo, fuentes de datos como AC Nielsen e Information Resources, Inc., compran datos de punto de venta tomados en almacenes; los procesan y después venden resúmenes estadísticos a los fabricantes. Estos últimos también suelen comprar datos y resúmenes estadísticos acerca de actividades promocionales, como precios especiales y empleo de exhibidores interiores.

Producción

Con el énfasis actual hacia la calidad, el control de calidad es una aplicación importante de la estadística en la producción. Para vigilar el resultado de un proceso de producción se emplean diversas gráficas de control estadístico de la calidad.

Economía

Con frecuencia se pide a los economistas su pronóstico acerca del futuro de la economía o de alguno de sus aspectos, por lo que recurren a información estadística diversa para elaborarlo. Por ejemplo para pronosticar las tasas de inflación usan indicadores como el índice de precios, tasas de desempleo. (Anderson, Sweeney y Williams; 2005)

No se puede olvidar, por su tradición en la actividad histórica de las estadísticas nacionales, los censos de población, viviendas, etc. que siguen siendo objeto de estudios, experiencias y realizaciones y que tienen como finalidad desarrollar políticas económicas y sociales. Se pueden citar, por ejemplo, las estadísticas de servicios, en particular sanitarias, (número de camas, médicos, etc por mil habitantes), las de educación y, más recientemente, las estadísticas relacionadas con el medio ambiente.

En un primer paso las Técnicas Cuantitativas describen un conjunto de mediciones realizadas sobre los elementos de una muestra, construye las tablas y gráficos de distribuciones de frecuencias y obtiene distintas medidas numéricas: de posición, de dispersión, de asimetría, de curtosis, de concentración, etc. Las gráficas de los datos muestrales suponen una estimación empírica de la forma de la población.

En un segundo paso se considera la forma en que se debe realizar la inferencia estadística, en este caso se distingue entre inferencia paramétrica e inferencia no paramétrica.

La investigación en el área de muestreo se ha convertido en un gran reto tanto para el mundo académico como para el mundo empresarial. Ya que es una herramienta básica para las ciencias sociales y disciplinas relacionadas.

Hoy en día las empresas admiten que las encuestas por muestreo son una fuente de obtención de datos estadísticos en un amplio rango de temas tanto a nivel de investigación de la empresa como a niveles administrativos y de investigación y desarrollo de la empresa. Por supuesto que esta evolución y reconocimiento de la utilidad de las encuestas por muestreo tuvo lugar aproximadamente desde 1930 (Kalton et al., 1983). Es un hecho que a partir de 1925 la estadística que se habia considerado como un elemento relativamente pasivo para recopilar información y describirla pasaba a ser descrita como un recurso empresarial activo y últil, que afectaba a decisiones y permitía extraer conclusiones a partir de la información de una muestra.

Según Palacios y Callejón (2004) dentro de las Técnicas estadísticas, la teoría del muestreo ha sido una de las que más aportaciones ha hecho a las Ciencias Económicas y Empresariales. Esencialmente, el problema del muestreo se puede considerar como un problema económico en el que se trata de optimizar la pareja "información-recursos". Se asume un riesgo al afirmar algo de un colectivo a partir de la información obtenida de una parte de él. La decisión óptima consiste en obtener dicha información con una cota mínima exigible de precisión, empleando para ello recursos mínimos.

1.3. Antecedentes del uso de información auxiliar

Si entre dos variables existe una fuerte relación, es posible utilizar la información auxiliar que se tenga de una de las variables, como puede ser la media

o el total poblacional, para estimar la media o el total de la otra variable. Esta circunstancia es importante cuando se pretende estimar el total sin conocer el número de elementos de la población, pero sí se conoce el valor total de la variable que proporciona información auxiliar.

Otro ejemplo que refleja la situación comentada anteriormente es el siguiente. Ya que existe una fuerte relación entre renta y ahorro, se puede estimar el valor total de los ahorros de los empleados de una empresa si se conoce el valor total de las rentas de dichos empleados. Por ejemplo, si se estima que, por término medio, el diez por ciento de la renta se dedica al ahorro y si se conoce la renta total, el ahorro total se estima igual a la décima parte del total de la renta. Observemos que esto se puede llevar a cabo sin necesidad de conocer el número de empleados de la empresa.

Los métodos de muestreo más conocidos y utilizados consideran estimadores que utilizan sólo los valores observados de la característica en estudio. Sin embargo es frecuente que la variable objeto de estudio y, esté altamente relacionada con una característica auxiliar x, cuyos datos están disponibles o son muy fáciles de obtener para todos los elementos de la población. En esta situación es muy útil considerar métodos de estimación que utilizan información auxiliar o suplementaria, relativa a una variable o característica correlacionada con la que es objeto de estudio, para modificar la forma de los estimadores directos o expandidos (los usuales cuando no existe dicha información suplementaria) consiguiendo estimadores más precisos que los calculados a partir de la muestra.

Como información suplementaria puede utilizarse:

- observaciones obtenidas con muestras grandes pero no probabilísticas;
- observaciones obtenidas con muestras grandes probabilísticas pero de tamaño excesivamente pequeño, o bien,
- observaciones obtenidas con muestras relativas a otra población diferente pero relacionada con la que se estudia

Entre estos métodos, llamados métodos de estimación indirecta, hay dos especialmente importantes y conocidos: el método de estimación de razón y el método de estimación de regresión.

La literatura de muestreo de poblaciones finitas es abundante en ejemplos en los cuales estos métodos son utilizados para estimar medias y totales poblaciones. Al respecto, Cochran (1977) hace referencia a que ya en el año 1802, el

procedimientos de Laplace para estimar la población de Francia utiliza un estimador de tipo razón. En la literatura estadística moderna estos procedimientos se han considerado en los últimos 50 años.

Es conocido que para un tamaño de muestra grande, el error cuadrático medio del estimador de razón es más pequeño que la varianza del estimador de expansión simple (si el coeficiente de correlación entre las variables es positivo y alto), pero mayor que el error cuadrático medio del estimador de regresión. No obstante el método de regresión es bastante complejo computacionalmente en el caso de diseños de muestreo multietápicos. Según Yates, "el método de estimación de razón es más sencillo computacionalmente, pero el método de regresión es en ciertas circunstancias más acurado. Cuando la variable auxiliar x representa el tamaño de la unidad, la recta de regresión pasa por el origen". En este último caso el método de razón resulta un estimador óptimo. La superioridad del estimador de razón frente al estimador de expansión simple y su simplicidad respecto al de regresión son dos de las razones por las que se ha extendido el uso de este estimador frente los otros estimadores indirectos.

Representamos por y el carácter o variable que constituye el objeto del estudio propiamente dicho, y por x una variable auxiliar que suponemos correlacionada con la primera.

Si queremos estimar el total poblacional Y y conocemos X (el total poblacional de la variable auxiliar), podemos utilizar, además del estimador directo o expandido $\widehat{Y} = N\overline{y}$, otras estimaciones, las cuales pueden surgir como casos particulares de la expresión general:

$$\widehat{Y}_G = \widehat{Y} + b_0(X - \widehat{X}),$$

donde b_0 puede interpretarse como un coeficiente de corrección para mejorar al estimador $\widehat{Y}$, $\overline{y}$ es la media muestral de la variable y, y N es el tamaño de la población.

Como valores particulares de b_0 tenemos los siguientes casos:

(a)

$$b_0 = 0,$$
$$\widehat{Y}_G = \widehat{Y}$$

(b)

$$b_0 = \frac{\widehat{Y}}{\widehat{X}} = \widehat{R}$$

$$\widehat{Y}_G = \widehat{Y} + \frac{\widehat{Y}}{\widehat{X}}(X - \widehat{X}) = \widehat{Y}_R = \frac{\widehat{Y}}{\widehat{X}}X = N\overline{y}\frac{\overline{X}}{\overline{x}}$$

(c)

$$b_0 = 1,$$
$$\widehat{Y}_G = \widehat{Y} + X - \widehat{X} = \widehat{Y}_D$$

(d)

$$b_0 = b$$

coeficiente de regresión de y_i sobre x_i

$$\widehat{Y}_G = \widehat{Y}_{rg} = \widehat{Y} + b(X - \widehat{X})$$

Si interesa estimar la media poblacional $\overline{Y}$, se obtienen las fórmulas correspondientes sustituyendo $\widehat{Y}$ y $\widehat{X}$ por los estimadores de las medias, es decir:

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

y

$$\widehat{\overline{X}} = \overline{x}.$$

Estimadores de razón

El método de estimación de razón trata de mejorar la precisión del estimador simple, utilizando información sobre una variable auxiliar x, relacionada con la variable en estudio y. En la práctica x suele ser el valor de y en una ocasión anterior en la que se hizo un censo completo.

Consideraremos un esquema de muestreo sin reposición y probabilidades iguales.

Como ya hemos visto anteriormente el estimador de razón para el total poblacional Y es:

$$\widehat{Y}_R = \frac{\overline{y}}{\overline{x}}X.$$

Si el parámetro a estimar es la media poblacional, $\overline{Y}$, el estimador de razón es:

$$\overline{y}_R = \frac{\overline{y}}{\overline{x}}\overline{X}$$

Si interesa estimar una razón entre dos variables, es decir $R = Y/X$, el estimador en este caso es:

$$\widehat{R} = \frac{\overline{y}}{\overline{x}} = \frac{\widehat{Y}}{\widehat{X}}$$

y no es necesario conocer el total de X.

Obviamente, el estimador de razón en general no es insesgado. La expresión exacta del sesgo viene dada por la siguiente expresión:

$$B(\widehat{R}) = -\frac{cov(\widehat{R}, \overline{x})}{\overline{X}}.$$

Este resultado debido a Hartley y Ross (1954) nos permite deducir una cota superior para el estimador de razón. Una cota superior del cociente entre el sesgo y la desviación típica de los estimadores de razón $\widehat{R}$, $\overline{y}_R$ e $\widehat{Y}_R$ viene dada por el cociente de variación de $\overline{x}$, es decir,

$$\frac{\mid B(\widehat{R}) \mid}{\sigma_{\widehat{R}}} \leq C_{\overline{x}}.$$

Entonces si el coeficiente de variación de $\overline{x}$ es pequeño, es decir si no hay demasiada variación de la muestra para la variable auxiliar, el sesgo del estimador es despreciable en comparación con su desviación típica. El mismo límite se aplica al sesgo de los estimadores del total y de la media. Estas expresiones son poco prácticas, de ahí que suela utilizarse normalmente la siguiente aproximación:

$$E(\widehat{R} - R) = \frac{1-f}{n\overline{X}^2}(RS_x^2 - \rho S_y S_x),$$

y una vez obtenidos estimadores de R, S_x, S_y, ρ y $\overline{X}$ de la muestra, se utiliza a veces como una aproximación del tamaño del sesgo en una muestra específica. En la expresión anterior, n representa el tamaño de la muestra y $f = n/N$ es la fracción de muestreo. S_y^2 y S_x^2 son, respectivamente, las cuasi-varianzas poblacionales de las variables y y x.

En la práctica se encuentra que el sesgo no tiene importancia aún en muestras de tamaño moderado.

Por último consideramos el caso en que el estimador de razón no tiene sesgo. Esto ocurre cuando la regresión de y sobre x es una línea recta que pasa por el origen:

$$E(y \mid x) = \beta x$$

En una población finita no es probable que esto ocurra exactamente, aunque con frecuencia se tendría una situación aproximada ya que el estimador de razón se utiliza por lo general cuando hay motivos para pensar que la razón y/x es aproximadamente constante.

No se posee una fórmula exacta del error cuadrático medio (ECM) del estimador, sino sólo aproximaciones que son válidas en muestras grandes, como la siguiente:

$$ECM(\widehat{R}) = \frac{1-f}{n\overline{X}^2}\frac{\sum_{1\leq i\leq N}(y_i - Rx_i)^2}{N-1}.$$

En la práctica, como un estimador muestral de:

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

se toma

$$\frac{\sum_{1\leq i\leq n}(y_i - Rx_i)^2}{n-1},$$

y así un estimador de $ECM(\widehat{R})$ es:

$$\widehat{V}(\widehat{R}) = \frac{1-f}{n\overline{X}^2}\frac{\sum_{1\leq i\leq n}(y_i - Rx_i)^2}{n-1}$$

Existen algunos estimadores basados en la razón entre las variables y y x que son insesgados.

A su vez Murthy (1964) proponía otro estimador de la media poblacional de la variable de interés mediante el uso la media poblacional de otra variable auxiliar como estimador del producto. Dicho estimador es más eficiente que la media muestral si la correlación entre x e y es negativa y elevada.

Si realizamos una comparación entre el estimador de producto y el de razón la primera observación que debemos realizar es que en el caso del estimador de razón la correlación entre y y x deber ser positiva y alta (aproximadamente entre 0.5 y 1), mientras que en el caso del estimador de producto dicha correlación deberá ser negativa y alta, encontrándose aproximadamente entre los valores -1 y $-0{,}5$.

Si observamos las rectas de regresión de x/y dadas por cada estimador, en el caso del estimador de razón la recta de regresión pasará por el origen sin

embargo en el caso de la recta de regresión para el estimador de producto se espera que no pase por el origen.

Los valores para los errores cuadráticos medios en el estimador de razón, serán pequeñas si:

(a) El factor de corrección por finitud (f.p.c) $f = \frac{n}{N}$ es alto.

(b) El tamaño muestral n es grande.

(c) La correlación presente entre x e y está muy cercana a la unidad.

(d) Los errores dados por la expresión $\varepsilon_i = (y_i - \overline{Y}) - R(x_i - \overline{X})$ son pequeños.

Los valores para los errores cuadráticos medios en el estimador de producto serán pequeñas si:

(a) El factor de corrección por finitud (f.p.c) $f = \frac{n}{N}$ es alto.

(b) El tamaño muestral n es grande.

(c) La correlación presente entre x e y está muy cercana a -1.

(d) Los errores dados por la expresión $\varepsilon_i = (y_i - \overline{Y}) - R(x_i - \overline{X})$ son pequeños.

El estimador habitual para la aproximación del error cuadrático medio deberá ser bajo si el tamaño muestral es alto, lo cual produce un intervalo de confianza más pequeño.

Obviamente en el caso del estimador de razón se deberá estimar únicamente un parámetro del modelo, así que los grados de libertad para la construcción de la estimación del intervalo serán $n-1$. Mientras que en el caso del estimador de producto si ambas variables son positivas $(x > 0; y > 0)$ pero la correlación es negativa, entonces tendremos que ambos se cortan y presentan la misma pendiente, de tal forma que se deberán usar $n - 2$ grados de libertad al igual que ocurre con el estimador de regresión.

La investigación sobre estimación de razón ha seguido fundamentalmente dos caminos:

- Construir estimadores tipo razón que disminuyan el sesgo como los estimadores insesgados y cuasi-insesgados estudiados por Hartley y Ross (1954), Quenouille (1956) y Ruiz y Santos (1989) entre otros.

- Construir estimadores tipo razón que disminuyan el error cuadrático medio bajo ciertas condiciones, como los estudiados por y Ray y Sahai (1980), Prasad (1986), Menéndez y Ferrales (1989), Srivastava (1980), entre otros.

Estos desarrollos están realizados generalmente bajo un esquema de muestreo aleatorio simple. Otro camino para mejorar la precisión de los estimadores es utilizar estimadores de razón bajo otros diseños muestrales más complejos.

Así hay trabajos sobre estimadores de razón en muestreo estratificado, Hansen, Hurwitz y Gurney (1946) muestreo bifásico Rao (1975a, 1975b, 1981) muestreo polietápico Williams (1961, 1962) y Rao (1964) y muestreo sistemático Swain (1964) y Singh (1966).

Una primera posibilidad para estimar un parámetro poblacional de la variable de interés disponiendo de varias variables auxiliares correladas positivamente dado que el objeto de estudio consiste en determinar la variable con mayor correlación con la variable principal y construir a partir de ella el estimador de razón despreciando el resto de variables. Sin embargo es obvio que este procedimiento implica un desprecio de recursos.

Olkin (1958) inició los trabajos para la utilización de varias variables auxiliares. La idea de su trabajo es construir con cada variable auxiliar el estimador de razón usual y hacer una combinación lineal de ellos con un cierto criterio de óptimo, consiguiendo así un estimador de razón múltiple.

Para intentar solventar algunos de los inconvenientes que presenta el estimador de Olkin, Rueda et al. (1992), proponen un nuevo método de construir estimadores múltiples de razón mediante los estimadores condensados y posteriormente Rueda (1993), desarrolla otro método diferente de construcción de estimadores de razón múltiples que denomina método iterado de razón, que se basa en un procedimiento iterativo en el que el estimador obtenido en un paso se utiliza para construir el estimador en el paso siguiente, presentando grandes ventajas para su puesta en práctica respecto a los estimadores condensados de Olkin.

En otro orden de estimadores que también usan información auxiliar encontramos por ejemplo el propuesto por Srivastava (1967) dicho estimador denominado "estimador de potencia o exponenciación"que está basado en la idea del uso de la media poblacional de la variable auxiliar X para el cálculo de la media poblacional de la variable de interés Y, a través del uso de la transformación dada por:

$$\overline{y}_{PW} = \overline{y}\left(\frac{\overline{x}}{\overline{X}}\right)^{\alpha},$$

donde α es una constante debidamente elegida. Si $\alpha = 1$ entonces $\overline{y}_{PW}$ se reduce a $\overline{y}_P$ (estimador de producto) y si por el contrario $\alpha = -1$ entonces $\overline{y}_{PW}$ se reduce a $\overline{y}_R$ (estimador de razón).

Se puede afirmar que $\alpha = -\rho C_y/C_x$ minimiza el valor de $V(\overline{y}_\alpha)$, donde, ${C_y}^2 = {S_y}^2/\overline{Y}^2$, ${C_x}^2 = {S_x}^2/\overline{X}^2$ y ρ es el coeficiente de correlación entre y y x; nótese que, en todas las derivaciones, α se considerará una constante fijada de antemano en el muestreo.

Sisodia y Dwivedi (1981) sugirieron un estimador de razón modificado para $\overline{Y}$ que presenta la expresión siguiente:

$$\overline{y}_{SD} = \overline{y}\frac{\overline{X} + C_X}{\overline{x} + C_X} = \frac{\overline{y}}{\overline{x}_{SD}}\overline{X}_{SD} = \widehat{R}_{SD}\overline{X}_{SD},$$

donde C_x es el coeficiente de variación de la variable auxiliar y es conocido.

A su vez Upadhyaya y Singh (1999) consideraron los coeficientes de variación y de Kurtosis de forma conjunta en su estimador de tipo razón, cuya expresión presentamos a continuación:

$$\overline{y}_{US1} = \overline{y}\frac{\overline{X}\beta_2(x) + C_X}{\overline{x}\beta_2(x) + C_X} = \frac{\overline{y}}{\overline{x}_{US1}}\overline{X}_{US1} = \widehat{R}_{US1}\overline{X}_{US1}.$$

Estos dos autores también propusieron otro estimador únicamente con la realización del cambio de lugar del coeficiente de Kurtosis y el coeficiente de variación tal y como se presenta a continuación:

$$\overline{y}_{US2} = \overline{y}\frac{\overline{X}C_X + \beta_2(x)}{\overline{x}C_X + \beta_2(x)} = \frac{\overline{y}}{\overline{x}_{US2}}\overline{X}_{US2} = \widehat{R}_{US2}\overline{X}_{US2}$$

Por su parte Ray y Singh (1981), sugieren el estimador de tipo:

$$\overline{y}_{RS} = \frac{\overline{y} + b(\overline{x}^\alpha - \overline{X}^\alpha)}{\overline{x}^\gamma}\overline{X}^\gamma$$

Los autores Kadilar y Cingi (2004) proponían también un estimador de razón basándose en la idea de Sisodia y Dwivedi (1981). Este nuevo estimador presentaba la siguiente expresión:

$$\overline{y}_{pSD} = \frac{\overline{y} + b(\overline{X} - \overline{x})}{\overline{x} + C_X}(\overline{X} + C_X) = \widehat{R}_{pSD}\overline{X}_{SD},$$

donde

$$\widehat{R}_{pSD} = \frac{\overline{y} + b(\overline{X} - \overline{x})}{\overline{x} + C_X}.$$

Este último estimador es más eficiente que los estimadores de Ray y Singh (1981), Sisodia y Dwivedi (1981) y que Upadhayaya y Singh (1999).

Estimadores de regresión

El estimador de razón es apropiado cuando la relación entre y y x es lineal y pasa por el origen. Si existe una relación lineal pero no pasa por el origen, es mejor utilizar la información auxiliar que proporciona x mediante un estimador de regresión.

El estimador de regresión viene dado por la expresión:

$$\widehat{\overline{Y}}_{reg} = \overline{y} + b(\overline{X} - \overline{x})$$

Puede ocurrir que b se desconozca de antemano, o bien que no se conozca, en cuyo caso habrá que estimarlo a partir de los datos de la muestra.

(a) Estimadores de regresión con b fija.

En un muestreo aleatorio simple, si b es una constante prefijada, b_0, el estimador de regresión:

$$\overline{y}_{reg} = \overline{y} + b_0(\overline{X} - \overline{x})$$

es insesgado y su varianza viene dada por:

$$V(\overline{y}_{reg}) = \frac{1-f}{n}(S_y^2 - 2b_0 S_{xy} + b_0^2 S_x^2)$$

cuyo estimador viene dado por:

$$\widehat{V}(\overline{y}_{reg}) = \frac{1-f}{n}(s_y^2 - 2b_0 s_{xy} + b_0^2 s_x^2)$$

(b) Estimadores de regresión cuando b se calcula a partir de la muestra.

En un muestreo aleatorio simple, el estimador de regresión es:

$$\overline{y}_{reg} = \overline{y} + b_0(\overline{X} - \overline{x})$$

donde:

$$b = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Si n es grande:

$$V(\overline{y}_{reg}) = \frac{1-f}{n} S_y^2(1-\rho^2),$$

y esta varianza se estima por:

$$\widehat{V}(\overline{y}_{reg}) = \frac{1-f}{n(n-2)} \sum_{i=1}^{n}[(y_i - \overline{y}) - b(x_i - \overline{x})]^2$$

Otro estimador propuesto por Srivastava (1971) es el estimador de clase general, para estimar la media poblacional $\overline{Y}$ de la variable objeto de estudio en el caso de conocer la media de la variable auxiliar $\overline{X}$. Dicho estimador viene dado por la expresión siguiente:

$$t_g = \overline{y} H(u),$$

donde $u = \dfrac{\overline{x}}{\overline{X}}$ y $H(\cdot)$ es una función paramétrica tal que satisfaga las siguientes condiciones:

1. $H(1) = 1$
2. Las derivadas parciales de primer y segundo orden de H con respecto a **u** existen y son constantes conocidas dadas por el punto $u = 1$.

Srivenkataramana y Tracy (1980) consideraron el estimador dual de razón cuya expresión es la siguiente:

$$\overline{y}_{nsu} = \overline{y}\frac{(N\overline{X} - n\overline{x})}{(N-n)\overline{X}},$$

la cual también se puede expresar como

$$\overline{y}_{nsu} = \overline{y}\frac{\overline{x}^*}{\overline{X}},$$

donde

$$\overline{x}^* = \frac{(N\overline{X} - n\overline{x})}{(N-n)} = \frac{1}{N-n}\sum_{i=1}^{N-n} x_i.$$

Se puede observar aquí que tanto el estimador regresión como el estimador de diferencia no son casos especiales de la clase general de estimadores definidos por Srivastava (1971).

Srivastava define otra clase de estimadores más amplia dada por:

$$t_w = H[\overline{y}, u]$$

donde $H[\overline{y}, u]$ es una función de $\overline{y}$ y u tal que satisfaga las siguientes condiciones de regularidad que se exponen:

(a) El punto $(\overline{y}, u)$ asume valores comprendidos en un subconjunto definido en $\mathbb{R}_2$ del espacio bidimensional que contiene el punto $(\overline{Y}, 1)$;

(b) La función $H(\overline{y}, u)$ es una función continua y definida en el espacio $\mathbb{R}_2$;

(c) $H(\overline{Y}, 1) = \overline{Y}$ y $H_0(\overline{Y}, 1) = \overline{Y}$ denotan las derivadas parciales de primer orden de H con respecto a $\overline{y}$

(d) Las derivadas parciales de primer y segundo orden de $H[\overline{y}, u]$ existen y son continuas y definidas en el espacio de R_2.

Por otra parte también se han de reseñar otros métodos de estimación de la media poblacional que usan la varianza conocida de la variable auxiliar en la fase de estimación, ya que estos métodos pueden ser usados como punto de referencia junto con los métodos que usan la media poblacional conocida o el total poblacional de la variable auxiliar para mejorar el estudio de la media poblacional bajo determinadas circunstancias.

Por ejemplo Srivastava y Jhajj (1981) introdujeron el uso del valor de la varianza conocida de la variable auxiliar para mejorar la eficiencia de los estimadores de la media poblacional. Estos autores consideraban una clase general de estimadores de razón cuya expresión se detalla a continuación:

$$\overline{y}_{SJ} = \overline{y}H(u, v)$$

donde $u = \dfrac{\overline{x}}{\overline{X}}$, $v = \dfrac{s_x^2}{S_x^2}$ y $H(u,v)$ es una función de u y v tal que:

(a) El punto (u,v) asume valores comprendidos un un subconjunto definido en $\mathbb{R}_2$ del espacio bidimensional que contiene el punto $(1,1)$;

(b) La función $H(u,v)$ es una función continua y definida en el espacio de $\mathbb{R}_2$;

(c) $H(1,1) = 1$

(d) Las derivadas parciales de primer y segundo orden de $H(u,v)$ existen y son continuas y definidas en el espacio de $\mathbb{R}_2$.

Otros de los estimadores propuestos por Srivastava y Jhajj (1981) en el cual se introduce el uso de la varianza conocida de la variable auxiliar para mejorar la eficiencia de los estimadores de la media poblacional es el estimador de clase general que se detalla a continuación:

$$\overline{y}_{SJ(w)} = H(\overline{y}, u, v)$$

donde

$u = \dfrac{\overline{x}}{\overline{X}}$, $v = \dfrac{s_x^2}{S_x^2}$ y $H(\overline{y}, u, v)$ es una función de y, u y v tal que:

(a) El punto $(\overline{y}, u, v)$ asume los valores comprendidos en un subconjunto definido en $\mathbb{R}_3$ del espacio tridimensional que contiene el punto $(\overline{Y}, 1, 1)$;

(b) La función $H(\overline{y}, u, v)$ es una función continua y definida en el espacio de R_3;

(c) $H(\overline{Y}, 1, 1) = \overline{Y}$

(d) Las derivadas parciales de primer y segundo orden de $H(\overline{y}, u, v)$ existen y son continuas y definidas en el espacio de $\mathbb{R}_3$.

Llegados a este punto, podemos afirmar que el uso de información auxiliar proporcionada por una variable x altamente correlacionada con la variable principal o de interés y es algo común en la estimación de medias o totales poblacionales tal y como demuestra la revisión de la literatura realizada anteriormente. Los estimadores de razón y regresión utilizan la información que

proporciona la variable x para modificar los estimadores directos, consiguiendo de este modo estimadores más precisos del parámetro en cuestión.

De una forma similar, es razonable suponer que estas técnicas de estimación se pueden utilizar, bajo las condiciones adecuadas, para proporcionar estimadores eficientes de la varianza.

Fuller (1970), propone un estimador de regresión de la varianza del estimador de Horvitz-Thompson del total poblacional usando como variable auxiliar x, las cantidades $\pi_i\pi_j - \pi_{ij}$ y $(\pi_i\pi_j - \pi_{ij})(i-j)^2$, siendo π_i y π_{ij} las probabilidades de inclusión individual y conjunta de cada unidad, respectivamente.

Ogus y Clark (1971) proponen el uso de estimadores de razón y diferencia de la varianza bajo un diseño de muestreo de Poisson, con el propósito de reducir el efecto del tamaño muestral aleatorio, en la estimación de la varianza.

Bajo un diseño muestral en el que se selecciona una unidad en cada estrato con probabilidad proporcional al tamaño (PPS), Hansen, Hurwitz y Madow (1953) proponen el uso de una variable correlacionada junto con la técnica de estratos colapsados, para estimar la varianza, y prueban que el estimador está positivamente sesgado.

Bajo el mismo diseño muestral, Hartley, Rao y Kiefer (1969) proponen un estimador de la varianza basado en suponer una buena regresión entre las verdaderas medias de los estratos y algunas variables auxiliares. Sus ejemplos, que utilizan una sola variable, indican una mejora considerable en términos de sesgo absoluto respecto al estimador propuesto por Hansen, Hurwitz y Madow. No obstante, el primer método es más sencillo de aplicar. Además el estimador de Hartley no se ha comprobado que seo no negativo bajo todas las condiciones.

Posteriormente, Shapire y Bateman (1978) consideran la reducción del sesgo del estimador de la varianza en un diseño con una unidad por estrato, utilizando como estimador de la varianza el estimador de Yates-Grundy, para un diseño con dos unidades por estrato con probabilidades π_{ij} calculadas en base a un esquema de muestreo de Durbin (1967).

Isaki (1983) por su parte propone un método de estimación de regresión para la estimación de la varianza, tanto para muestreo aleatorio simple, muestreo con probabilidades iguales con reemplazamiento y las posteriores mejoras de éste debidas a Prasad y Singh (1990, 1992) para aumentar la precisión y disminuir el sesgo.

Por lo tanto, visto lo anterior ya se ha asumido que solo se dispone de una

variable auxiliar para mejorar la estimación de los distintos parámetros de la variable objeto de estudio. A su vez también se ha asumido que el camino más apropiado para el uso de la variable auxiliar implica tener en cuenta la relación entre las dos variables. Generalizando un poco más, la variable de interés podría depender de un término constante también, o de más de una variable auxiliar, o en muchas ocasiones de ambos. Sin embargo la relación es casi improbable que se pueda representar de forma idónea con un modelo que implica la relevancia de mínimos cuadrados ordinarios.

Una situación en la que el uso de los mínimos cuadrados ordinarios está justificado y puede ser apropiado es donde la variable de interés es el gasto y la variable auxiliar son los ingresos. La relación entre ingresos y gastos, es bien conocida y presenta una relación de dependencia lineal. Pero obviamente los mínimos cuadrados ordinarios asumen homocedasticidad (la varianza de los gastos permanece constante cuando los ingresos aumentan) mientras que es más que probable que la varianza de los gastos se incremente con los ingresos, y de hecho los datos procedentes de la mayoría de las encuestas por muestreo indican claramente la existencia de una medida de heterocedasticidad. Esto por sí mismo es motivo suficiente para hacer cuestionable el uso de los mínimos cuadrados ordinarios en esta situación.

Un estimador comúnmente utilizado para el total en estas circunstancias generales es el *estimador de regresión generalizado*, o GREG (Cassel et al., 1976), el cual se presenta a continuación:

$$\widehat{Y}_{GREG} = \widehat{Y}_{HT} + \sum_{i=1}^{p} \widehat{B}_i (X_i - \widehat{X}_{HTi})$$

siendo X_i el total poblacional de la variable auxiliar x_i para $i = 1, ..., p$.

En estas dos ecuaciones, $\widehat{Y}_{HT}$ es el estimador de Horvitz- Thompson de la variable en estudio, $\widehat{X}_{HT}$ es el estimador de Horvitz-Thompson de la variable auxiliar y $\widehat{B}_i$ es el coeficiente de regresión entre la variable de interés y la i-ésima variable auxiliar, donde la regresión es sobre p variables auxiliares simultáneamente.

Una de esas variables auxiliares debe ser un término constante, en cuyo caso esa será la intersección del estimador en la ecuación.

Capítulo 2

Estimadores de tipo razón para una proporción

2.1. 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 que se sigue a lo largo del trabajo.

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, la función de distribución, etc., a los que se denominan *parámetros poblacionales*.

Existen dos estrategias posibles para la recopilación de datos:

(i) examinar todas las unidades de la población, es decir, realizar un censo,

(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 las características de toda la población.

En la práctica, no siempre resulta posible realizar un censo para obtener determinados parámetros poblacionales. Por está razón, se recurre a una muestra para estimar estos parámetros poblacionales. Así, una muestra es un subconjunto de unidades, s, seleccionados de U 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 tamaño de la muestra s se denotará por n.

Dentro de la población U interesa estudiar ciertas características de una *variable de estudio*, *interés* o *principal*, la cual se denotará como y. Asociado al elemento i de la muestra se conoce exactamente y sin error el valor de la característica de interés. A esta cantidad se le denotará como y_i. 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 muestra, mejorar las estimaciones, etc. Para J variables auxiliares, el vector de variables auxiliares viene dado por $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_J)$, donde $\mathbf{x}_j = (x_{1j}, \ldots, x_{ij}, \ldots, x_{Nj})^t$. 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 $\mathbf{X} = (X_1, \ldots, X_J)$ o $\overline{\mathbf{X}} = (\overline{X}_1, \ldots, \overline{X}_J)$ son conocidas, donde $X_j = \sum_{i=1}^{N} x_{ij}$ y $\overline{X}_j = N^{-1} \sum_{i=1}^{N} x_{ij}$.

La notación y y $\mathbf{x}$ se suele utilizar en variables cuantitativas. Sin embargo, para variables dicotómicas se suele utilizar la variable:

$$A_i = \begin{cases} 1 & \text{si el } \textit{i-ésimo} \text{ elemento presenta la característica de interés } A. \\ 0 & \text{si el } \textit{i-ésimo} \text{ elemento no presenta la característica de interés } A. \end{cases}$$

Por su parte, B denotará un atributo o característica auxiliar relacionada con A, y cuyos valores vienen dados por $B_1, \ldots, B_N$.

La probabilidad de inclusión de primer orden asociada 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.

El objetivo principal de este trabajo es estimar la proporción poblacional de individuos que presentan la característica de interés A, es decir,

$$P_A = \frac{1}{N}\sum_{i=1}^{N} A_i.$$

Esta estimación se realizará mediante estimación puntual y por intervalo de confianza, y con la utilización de nuevos procedimientos que incorporen información auxiliar en la etapa de estimación. Estos estimadores e intervalos de confianza propuestos así como otras propiedades pueden consultarse también en Rueda et al. (2011a), Rueda et al. (2011b) y Muñoz et al. (2011).

2.2. Muestreo aleatorio simple

En esta sección asumiremos que la muestra s, de tamaño n, se selecciona de U mediante muestreo aleatorio simple (MAS). La extensión a un diseño muestral general se tratará en la Sección 2.3. Como se comentó anteriormente, el objetivo es estimar P_A, esto es, la proporción de individuos de la población que poseen el atributo de interés A.

2.2.1. Definición del estimador

El estimador estándar o de expansión simple de P_A viene dado por

$$\widehat{p}_A = \frac{1}{n}\sum_{i\in s} A_i. \tag{2.1}$$

El estimador $\widehat{p}_A$ es insesgado y su varianza viene dada por

$$V(\widehat{p}_A) = \frac{N-n}{N-1}\frac{1}{n}P_A Q_A,$$

donde $Q_A = 1 - P_A$.

Un estimador de $V(\widehat{p}_A)$ es:

$$\widehat{V}(\widehat{p}_A) = \frac{1-f}{n-1}\widehat{p}_A\widehat{q}_A,$$

donde $f = n/N$ es la fracción de muestreo y $\widehat{q}_A = 1 - \widehat{p}_A$. Estas propiedades pueden derivarse fácilmente a partir de las propiedades del estimador simple

$\overline{y} = n^{-1} \sum_{i \in s} y_i$ en el problema de la estimación de la media poblacional $\overline{Y} = N^{-1} \sum_{i \in s} y_i$. En efecto, puede comprobarse (véase, por ejemplo, Cochran 1977, Sección 2.5 y Lohr, 1999, Sección 2.7) que

$$V(\overline{y}) = (1 - f)\frac{S^2}{n} \tag{2.2}$$

y a su vez

$$\widehat{V}(\overline{y}) = (1 - f)\frac{s^2}{n}, \tag{2.3}$$

donde S^2 y s^2 son las cuasivarianzas poblacional y muestral, respectivamente, de la variable y. Dado que

$$V(A) = \frac{1}{N}\sum_{i=1}^{N} A_i^2 - P_A^2 = \frac{1}{N}\sum_{i=1}^{N} A_i - P_A^2 = P_A - P_A^2 = P_A(1 - P_A) = P_A Q_A$$

tenemos que

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

en el caso de variables dicotómicas. Análogamente puede deducirse que

$$s^2 = \frac{n}{n-1}\widehat{p}_A \widehat{q}_A,$$

donde $\widehat{q}_A = 1 - \widehat{p}_A$.

Teniendo en cuenta las expresiones (2.2) y (2.3), podemos concluir que

$$V(\widehat{p}_A) = (1 - f)\frac{S^2}{n} = \frac{N-n}{N}\frac{1}{n}\frac{N}{N-1} P_A Q_A = \frac{N-n}{N-1}\frac{1}{n} P_A Q_A$$

y de forma análoga

$$\widehat{V}(\widehat{p}_A) = (1 - f)\frac{s^2}{n} = (1 - f)\frac{1}{n}\frac{n}{n-1}\widehat{p}_A \widehat{q}_A = \frac{1-f}{n-1}\frac{1}{n}\widehat{p}_A \widehat{q}_A.$$

Podemos observar que el estimador definido en (2.1) no hace uso de la información auxiliar en la fase de estimación. Los estimadores tipo razón para la media poblacional de una variable cuantitativa incorporan información auxiliar en la fase de estimación, y poseen propiedades deseables incluyendo un importante aumento de la eficiencia. Pensamos que estas propiedades deseables también se pueden obtener en el caso de variables dicotómicas. Por tanto, un primer objetivo a seguir en este trabajo es definir nuevos estimadores de tipo razón para la proporción poblacional P_A.

Definimos un primer estimador de tipo razón para P_A mediante:

$$\widehat{p}_r = \widehat{R} P_B, \tag{2.4}$$

donde $\widehat{R} = \widehat{p}_A / \widehat{p}_B$ es un estimador de la razón poblacional $R = P_A / P_B$,

$$\widehat{p}_B = \frac{1}{n} \sum_{i \in s} B_i$$

es la proporción muestral de individuos que presentan el atributo auxiliar B y

$$P_B = \frac{1}{N} \sum_{i=1}^{N} B_i$$

es la homóloga de $\widehat{p}_B$ a nivel poblacional.

Para obtener $\widehat{p}_r$ asumimos que la proporción poblacional de individuos que poseen el atributo B, P_B, es conocido a partir de un censo o se ha estimado sin error. Esta suposición es comúnmente utilizada en el contexto del muestreo en poblaciones finitas cuando el parámetro es la media poblacional (veáse Särndal et al., 1992) o la función de distribución (véase Rao et al., 1990). Además, destacamos que los censos poblacionales llevados a cabo en varios países recopilan información sobre un conjunto de variables auxiliares a nivel poblacional (véase, por ejemplo, Silva y Skinner, 1995). Algunas de las variables son cuantitativas, aunque es bastante común tener también variables categóricas tales como el sexo, estado civil y código postal entre otras. Tales variables pueden ser usadas para el cálculo de P_B, lo cual indica que la suposición anteriormente mencionada puede cumplirse en numerosas situaciones.

Por otro lado, P_B se puede estimar de un censo previo, o bien que se haya calculado en otra ocasión. Por ejemplo, la encuesta canadiense sobre la mano de obra disponible usa un estimador de regresión en el cual algunos de los parámetros poblacionales auxiliares pueden ser estimados. Esta estimación puede tener un impacto sobre la estimación de la varianza. No obstante, Berger et al. (2009) han propuesto un método para la estimación de la varianza que tiene en cuenta la estimación de los parámetros poblacionales auxiliares, y el cual puede aplicarse fácilmente a variables dicotómicas.

Podemos observar que el estimador de tipo razón propuesto en (2.4) se reduce a P_A cuando $A_i = B_i$ para todo $i \in U$, y por tanto la varianza del estimador propuesto será cero en este caso. Esto nos sugiere que $\widehat{p}_r$ podría ser considerablemente más eficiente que el estimador estándar $\widehat{p}_A$ cuando el atributo A esté estrechamente relacionado con el atributo B. Bajo esta situación, tambien se podrían obtener intervalos de confianza de menor amplitud.

2.2.2. Propiedades teóricas

Sean A^c y B^c los atributos complementarios de los atributos A y B respectivamente, y sea la tabla poblacional de doble entrada dada por

	B	B^c	
A	N_{11}	N_{12}	$N_{1\cdot}$
A^c	N_{21}	N_{22}	$N_{2\cdot}$
	$N_{\cdot 1}$	$N_{\cdot 2}$	N

(2.5)

donde $N_{1\cdot} = \sum_{i=1}^{N} A_i$ es el número de individuos en la población que poseen el atributo A, $N_{2\cdot}$ es el número de individuos en la población que poseen el atributo A^c, etc. De forma análoga, N_{11} es el número de individuos en la población que simultáneamente poseen los atributos A y B, N_{12} es el número de individuos en la población que simultáneamente poseen los atributos A y B^c, etc.

La clasificación dada en (2.5) se traduce a nivel muestral en la tabla de doble entrada

	B	B^c	
A	n_{11}	n_{12}	$n_{1\cdot}$
A^c	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	n

(2.6)

Proposición 2.1 *El sesgo del estimador de razón $\widehat{p}_r$ satisface*

$$\frac{\mid B(\widehat{p}_r) \mid}{\sigma_{\widehat{p}_r}} \leq cv_{\widehat{p}_B}, \tag{2.7}$$

donde $\sigma_{\widehat{p}_r}$ es la desviación estándar de $\widehat{p}_r$ y $cv_{\widehat{p}_B}$ es el coeficiente de variación de $\widehat{p}_B$.

□

Demostración

Por un lado se tiene que:

$$B(\widehat{p}_r) = B(\widehat{R}P_B) = P_B B(\widehat{R}) = P_B[E(\widehat{R}) - R] =$$

$$= P_B\left[E(\widehat{R}) - \frac{P_A}{P_B}\right] = P_B\left[E(\widehat{R}) - \frac{E(\widehat{p}_A)}{E(\widehat{p}_B)}\right] =$$

$$= P_B \left[E(\widehat{R}) - \frac{E(\widehat{R}\widehat{p}_B)}{E(\widehat{p}_B)} \right] = P_B \left[\frac{E(\widehat{R})E(\widehat{p}_B) - E(\widehat{R}\widehat{p}_B)}{E(\widehat{p}_B)} \right] =$$

$$= -cov(\widehat{R}, \widehat{p}_B)$$

Tomando cuadrados se tendría que:

$$[B(\widehat{p}_r)]^2 = [cov(\widehat{R}, \widehat{p}_B)]^2 = \frac{[cov(\widehat{R}, \widehat{p}_B)]^2}{V(\widehat{R})V(\widehat{p}_B)} V(\widehat{R})V(\widehat{p}_B) \leq V(\widehat{R})V(\widehat{p}_B) =$$

$$= V(\widehat{p}_r)\frac{V(\widehat{p}_B)}{P_B^2} = V(\widehat{p}_r)[cv_{\widehat{p}_B}]^2.$$

De la última expresión se deduce fácilmente que

$$\mid B(\widehat{p}_r) \mid \leq \sigma_{\widehat{p}_r} cv_{\widehat{p}_B},$$

la cual viene a demostrar la proposición formulada.

□

A partir de la cota superior (2.7) se observa que si el coeficiente de variación de $\widehat{p}_B$ tiende a cero a medida que aumenta el tamaño de la muestra, el sesgo del estimador $\widehat{p}_r$ también tenderá a cero. Destacamos que esta situación se verifica desde un punto de vista teórico al ser el estimador $\widehat{p}_B$ consistente. En otras palabras, el sesgo del estimador de razón $\widehat{p}_r$ generalmente será insignificante a medida que aumenta el tamaño de la muestra.

Proposición 2.2 *Una aproximación del sesgo del estimador de razón propuesto $\widehat{p}_r$ viene dada por*

$$B(\widehat{p}_r) = \frac{N-n}{N-1}\frac{1}{n}\left(\frac{Q_B}{P_B} - \frac{\phi\sqrt{P_A Q_A P_B Q_B}}{P_A} \right), \tag{2.8}$$

donde $Q_B = 1 - P_B$ y

$$\phi = \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1.}N_{2.}N_{.1}N_{.2}}}$$

es el coeficiente V de Cramer basado en la clasificación dada en la tabla de doble entrada (2.5).

□

Demostración

En primer lugar, de la Sección 2.2.1 sabemos que

$$V(\widehat{p}_B) = \frac{N-n}{N-1}\frac{1}{n}P_B Q_B.$$

Por otro lado, las variables $(n_{11}, n_{12}, n_{21}, n_{22}) \simeq HG(N, n, N_{11}, N_{12}, N_{21})$ y

$$cov(n\widehat{p}_A, n\widehat{p}_B) = cov(n_{11}+n_{12}, n_{11}+n_{21}) =$$

$$= V(n_{11}) + cov(n_{11}, n_{12}) + cov(n_{11}, n_{21}) + cov(n_{12}, n_{21}),$$

por lo tanto

$$V(n_{11}) = \frac{N-n}{N-1}n\frac{N_{11}}{N}\left(1-\frac{N_{11}}{N}\right) \quad ; \quad cov(n_{ik}, n_{jl}) = -\frac{N-n}{N-1}n\frac{N_{ik}N_{jl}}{N^2},$$

resultando la covarianza

$$cov(n\widehat{p}_A, n\widehat{p}_B) = \frac{N-n}{N-1}\frac{n}{N^2}(N_{11}N_{22} - N_{12}N_{21}),$$

y de aquí se obtiene fácilmente que

$$cov(\widehat{p}_A, \widehat{p}_B) = \frac{N-n}{N-1}\frac{1}{n}\phi\sqrt{P_A Q_A P_B Q_B}.$$

Finalmente, sea $e_1 = \dfrac{\widehat{p}_A - P_A}{P_A}$ y $e_2 = \dfrac{\widehat{p}_B - P_B}{P_B}$. Podemos expresar el estimador de tipo razón en términos de e_1 y e_2. Asumiendo que $\mid e_2/\widehat{p}_B \mid < 1$ se desarrolla $\widehat{R}$ en términos de e_2:

$$B(\widehat{p}_r) = P_B(E(\widehat{R}) - R) \simeq P_B E(e_1 - e_2 - e_1 e_2 + e_2^2) =$$

$$= P_B\left(-\frac{1}{P_B P_A}cov(\widehat{p}_A, \widehat{p}_B) + \frac{V(\widehat{p}_b)}{P_B^2}\right) = \frac{N-n}{N-1}\frac{1}{n}\left(\frac{Q_B}{P_B} - \frac{\phi\sqrt{P_A Q_A P_B Q_B}}{P_A}\right).$$

□

La expresión (2.8) implica que $\widehat{p}_r$ es asintóticamente insesgado. Sin embargo, el sesgo podría no ser insignificante para tamaños muestrales pequeños. Siguiendo el planteamiento seguido en la Sección 2.2.1 para la obtención de $\widehat{V}(\widehat{p}_A)$, se puede demostrar fácilmente que un estimador de $B(\widehat{p}_r)$ viene dado por

$$\widehat{B}(\widehat{p}_r) = \frac{1-f}{n-1}\left(\frac{\widehat{q}_B}{\widehat{p}_B} - \frac{\widehat{\phi}\sqrt{\widehat{p}_A\widehat{q}_A\widehat{p}_B\widehat{q}_B}}{\widehat{p}_A}\right), \tag{2.9}$$

donde $\widehat{q}_B = 1 - \widehat{p}_B$ y

$$\widehat{\phi} = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}}$$

es el coeficiente V de Cramer basado en la clasificación dada en la tabla de doble entrada (2.6).

Las expresiones de las varianzas de los estimadores son también una cuestión clave en el contexto del muestreo en poblaciones finitas. Por ejemplo, una de sus aplicaciones más importantes es la construcción de intervalos de confianza, tema que se abordará con detalle en el Capítulo 4. En la siguiente proposición se deriva la varianza del estimador de tipo razón propuesto.

Proposición 2.3 *La varianza asintótica del estimador de razón propuesto* $\widehat{p}_r$ *es*

$$AV(\widehat{p}_r) = \frac{N-n}{N-1}\frac{1}{n}\left(P_A Q_A + R^2 P_B Q_B - 2R\phi\sqrt{P_A Q_A P_B Q_B}\right). \qquad (2.10)$$

□

Demostración

La varianza del estimador $\widehat{p}_r$ se deriva utilizando desarrollos en serie de Taylor (véase, por ejemplo, Särndall et al., 1992, pg 178). Utilizando la técnica de linealización de Taylor puede comprobarse que la razón $\widehat{R}$ puede descomponerse como

$$\widehat{R} = R + \frac{1}{nP_B}\sum_{i\in s}(A_i - RB_i) + r_2 = R + \frac{1}{P_B}(\widehat{p}_A - R\widehat{p}_B) + r_2,$$

donde r_2 es el resto de la serie de Taylor de orden 2. Multiplicando la expresión anterior por P_B obtenemos

$$\widehat{p}_r = P_A + (\widehat{p}_A - R\widehat{p}_B) + r_{2r}, \qquad (2.11)$$

donde r_{2r} es el correspondiente resto de orden 2 de la serie de Taylor asociado al estimador $\widehat{p}_r$. Tomando varianzas en (2.11), la varianza asintótica de $\widehat{p}_r$ se puede aproximar por

$$AV(\widehat{p}_r) = V(\widehat{p}_A) + R^2 V(\widehat{p}_B) - 2Rcov(\widehat{p}_A, \widehat{p}_B).$$

Sustituyendo $V(\widehat{p}_A)$, $V(\widehat{p}_B)$ y $cov(\widehat{p}_A, \widehat{p}_B)$ en la expresión anterior por sus correspondientes expresiones deducidas con anterioridad, obtenemos

$$AV(\widehat{p}_r) = \frac{N-n}{N-1}\frac{1}{n}\left(P_A(1-P_A) + R^2 P_B(1-P_B) - 2R\frac{N_{11}N_{22} - N_{12}N_{21}}{N^2}\right) =$$

$$= \frac{N-n}{N-1}\frac{1}{n}\left(P_A Q_A + R^2 P_B Q_B - 2R\phi\sqrt{P_A Q_A P_B Q_B}\right).$$

□

La expresión $AV(\widehat{p}_r)$ depende de parámetros desconocidos y, por tanto, no puede utilizarse, por ejemplo, para la construcción de intervalos de confianza en la práctica. Por esta razón también se ha obtenido un estimador de (2.10), el cual viene dado por

$$\widehat{V}(\widehat{p}_r) = \frac{1-f}{n-1}\left(\widehat{p}_A\widehat{q}_A + \widehat{R}^2\widehat{p}_B\widehat{q}_B - 2\widehat{R}\widehat{\phi}\sqrt{\widehat{p}_A\widehat{q}_A\widehat{p}_B\widehat{q}_B}\right). \tag{2.12}$$

Las expresiones (2.10) y (2.12) nos permitirán obtener, en el Capítulo 4, intervalos de confianza para la proporción poblacional P_A usando diferentes métodos.

2.2.3. Comparación con el estimador de expansión simple

Teorema 2.1 *La varianza del estimador de tipo razón $\widehat{p}_r$ es menor que la varianza del estimador de expansión simple $\widehat{p}_A$, es decir $AV(\widehat{p}_r) < V(\widehat{p}_A)$, cuando*

$$\phi > \frac{1}{2}\frac{cv_B}{cv_A}, \tag{2.13}$$

donde cv_A y cv_B son, respectivamente, los coeficientes de variación de los atributos A y B.

□

Demostración.

Recordamos que las expresiones de las varianzas de los estimadores de razón y expansión simple para la proporción poblacional son

$$AV(\widehat{p}_r) = \frac{N-n}{N-1}\frac{1}{n}(P_A Q_A + R^2 P_B Q_B - 2R\phi\sqrt{P_A Q_A P_B Q_B})$$

y

$$V(\widehat{p}_A) = \frac{N-n}{N-1}\frac{1}{n}P_AQ_A.$$

Tenemos que demostrar cuando $AV(\widehat{p}_r) < V(\widehat{p}_A)$, es decir,

$$\frac{N-n}{N-1}\frac{1}{n}(P_AQ_A + R^2P_BQ_B - 2R\phi\sqrt{P_AQ_AP_BQ_B}) < \frac{N-n}{N-1}\frac{1}{n}P_AQ_A.$$

Simplificando en ambos miembros de la desigualdad anterior nos queda

$$P_AQ_A + R^2P_BQ_B - 2R\phi\sqrt{P_AQ_AP_BQ_B} < P_AQ_A,$$

$$R^2P_BQ_B - 2R\phi\sqrt{P_AQ_AP_BQ_B} < 0$$

$$R\sqrt{P_BQ_B} < 2\phi\sqrt{P_AQ_A},$$

$$\phi > \frac{1}{2}\frac{\sqrt{P_BQ_B}}{\sqrt{P_AQ_A}}\frac{P_A}{P_B} = \frac{1}{2}\frac{\sqrt{P_BQ_B}/P_B}{\sqrt{P_AQ_A}/P_A}.$$

Dado que $V(A) = P_AQ_A$ y $V(B) = P_BQ_B$, según la Sección 2.2.1, se concluye que $AV(\widehat{p}_r) < V(\widehat{p}_A)$ cuando

$$\phi > \frac{1}{2}\frac{cv_B}{cv_A}.$$

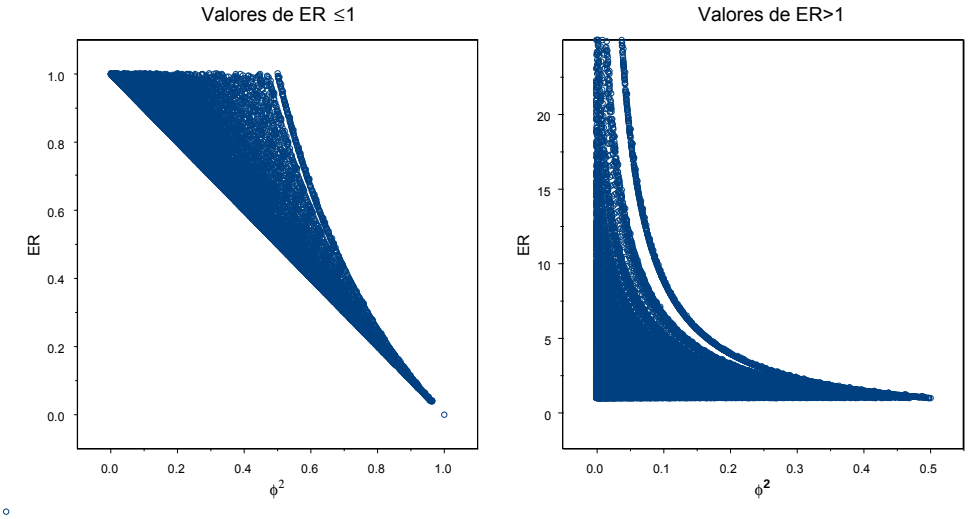
□

A partir de la expresión (2.13) pueden derivarse varias conclusiones. En primer lugar, observamos que

$$\frac{1}{2}\frac{cv_B}{cv_A} \geq 0.$$

Esto implica que la relación entre los atributos A y B deberá ser positiva, es decir $\phi > 0$, para que el estimador de tipo razón $\widehat{p}_r$ sea más eficiente que el estimador estándar $\widehat{p}_A$. En segundo lugar, observamos de (2.13) que si ambos atributos tienen una dispersión, en términos relativos, muy parecida, entonces ϕ tendrá que ser mayor que 0.5 para concluir que el estimador de razón sea más eficiente que el estimador de expansión simple. Por último, reseñar que si la dispersión del atributo B es inferior a la dispersión del atributo A, nos

Figura 2.1: Comparación teórica del estimador de tipo razón $\widehat{p}_r$ con el estimador de expansión simple $\widehat{p}_A$ mediante la eficiencia relativa (ER) entre ambos estimadores.



encontramos ante una situación que favorece el que la cota inferior para ϕ sea menor que la cota anteriormente fijada en 0.5.

El estudio de la comparación teórica entre el estimador de tipo razón propuesto $\widehat{p}_r$ y el estimador estándar $\widehat{p}_A$ se completa a continuación mediante el análisis de la eficiencia relativa entre ambos estimadores en distintas situaciones. Para llevar a cabo este estudio se generaron numerosas tablas de la forma

	B	B^c	
A	P_{11}	P_{12}	P_A
A^c	P_{21}	P_{22}	Q_B
	P_B	Q_B	1

(2.14)

con valores de P_A, P_B y P_{11} entre 0.01 y 0.99. El resto de parámetros requeridos en la tabla de doble entrada dada en (2.14) se completaron a partir de los anteriores ya descritos. Una vez obtenidas estas tablas que reflejan distintas situaciones que se pueden presentar en la práctica, se obtuvo la eficiencia relativa (ER) entre el estimador $\widehat{p}_r$ y el estimador $\widehat{p}_A$, la cual viene definida como

$$ER = \frac{AV(\widehat{p}_r)}{V(\widehat{p}_A)}.$$

Destacamos que la medida anterior no depende ni del tamaño de la población

N ni del tamaño de la muestra n, puesto que la expresión

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

contenida en $AV(\widehat{p}_r)$ y $V(\widehat{p}_A)$ se simplifica al realizar el cociente, es decir, la ER viene dada por

$$ER = \frac{P_A Q_A + R^2 P_B Q_B - 2R\phi\sqrt{P_A Q_A P_B Q_B}}{P_A Q_A}.$$

La Figura 2.1 muestra los resultados obtenidos en el estudio teórico comentado anteriormente, es decir, mostramos los valores de ER en función de ϕ^2. Destacamos que valores de ER menores que 1 indican que el estimador $\widehat{p}_r$ es más eficiente que $\widehat{p}_A$. De la Figura 2.1 observamos que para cualquier valor de ϕ^2 el estimador de razón puede ser más eficiente que el estimador de expansión simple, y siempre que $\phi^2 > 0.5$, $\widehat{p}_r$ es más eficiente que $\widehat{p}_A$. En algunas situaciones, la ganancia en eficiencia entre $\widehat{p}_r$ y $\widehat{p}_A$ puede ser considerable o bien muy similar. Por ejemplo, cuando $\phi^2 = 0.5$, $\widehat{p}_r$ y $\widehat{p}_A$ pueden tener el mismo comportamiento (ER=1), o bien $\widehat{p}_r$ puede ser el doble de eficiente que $\widehat{p}_A$ (ER=0.5). Por último, destacamos que el estimador de tipo razón $\widehat{p}_r$ puede ser considerablemente peor que $\widehat{p}_A$ para valores de ϕ^2 próximos a 0.

2.2.4. Definición de estimadores insesgados y más eficientes

A continuación modificamos el estimador de tipo razón propuesto para construir estimadores bien con menor sesgo o con menor varianza.

En primer lugar, siguiendo la idea de Hartley y Ross (1954), se puede obtener un estimador insesgado de P_A usando la expresión del sesgo $\widehat{B}(\widehat{p}_r)$ derivada en (2.9).

Un segundo estimador propuesto de tipo razón para P_A viene dado por

$$\widehat{p}_{r.u} = \widehat{p}_r - \widehat{B}(\widehat{p}_r),$$

el cual es aproximadamente insesgado.

A continuación se define un nuevo estimador de tipo razón más eficiente que $\widehat{p}_r$. Este estimador de tipo razón se basa en la siguiente idea. El estimador estándar $\widehat{p}_A$ se puede obtener también como $\widehat{p}_A = 1 - \widehat{q}_A$, donde

$\widehat{q}_A = n^{-1}\sum_{i\in s} A_i^c$, lo que implica que $\widehat{p}_A$ tiene el mismo comportamiento en la estimación de P_A que el comportamiento de $\widehat{q}_A$ en la estimación de Q_A. Sin embargo, puede comprobarse fácilmente que esta propiedad no se cumple para el estimador $\widehat{p}_r$, es decir, $\widehat{p}_r \neq 1-\widehat{q}_r$, donde $\widehat{q}_r = \widehat{R}^c Q_B$ es el estimador de tipo razón para Q_A y $\widehat{R}^c = (\widehat{q}_A/\widehat{q}_B)$. Este resultado nos lleva a que un estimador alternativo para P_A es $\widehat{p}_{r.q} = 1-\widehat{q}_r$.

El interés reside en analizar cuándo $\widehat{p}_r$ posee mejores propiedades que $\widehat{p}_{r.q}$ y viceversa. Para dar solución a dicha cuestión se considerará el criterio de mínima varianza.

Teorema 2.2 *La varianza del estimador de tipo razón $\widehat{p}_r$ es menor que la varianza del estimador de tipo razón $\widehat{p}_{r.q}$, es decir, $AV(\widehat{p}_r) < AV(\widehat{p}_{r.q})$, si $P_A < P_B$.*

□

Demostración

Puesto que $AV(\widehat{p}_{r.q}) = AV(1-\widehat{q}_r) = AV(\widehat{q}_r)$, el problema previo es equivalente a determinar cuándo $AV(\widehat{p}_r) < AV(\widehat{q}_r)$.

$AV(\widehat{p}_r) < AV(\widehat{q}_r)$ implica que

$$P_AQ_A+\frac{P_A^2}{P_B^2}P_BQ_B-2\frac{P_A}{P_B}\phi\sqrt{P_AQ_AP_BQ_B} < Q_AP_A+\frac{Q_A^2}{Q_B^2}Q_BP_B-2\frac{Q_A}{Q_B}\phi\sqrt{Q_AP_AQ_BP_B}$$

$$\frac{P_A^2Q_B}{P_B}-\frac{Q_A^2P_B}{Q_B}-2\frac{P_A}{P_B}\phi\sqrt{P_AQ_AP_BQ_B}+2\frac{Q_A}{Q_B}\phi\sqrt{P_AQ_AP_BQ_B} < 0$$

$$\frac{P_A^2Q_B^2-Q_A^2P_B^2-2P_AQ_B\phi\sqrt{P_AQ_AP_BQ_B}+2Q_AP_B\phi\sqrt{P_AQ_AP_BQ_B}}{P_BQ_B} < 0$$

$$(P_AQ_B+Q_AP_B)(P_AQ_B-Q_AP_B)-2\phi\sqrt{P_AQ_AP_BQ_B}(P_AQ_B-Q_AP_B) < 0$$

$$(P_AQ_B - Q_AP_B)(P_AQ_B + Q_AP_B - 2\phi\sqrt{P_AQ_AP_BQ_B}) < 0$$

Puesto que $P_AQ_B - Q_AP_B = P_A - P_AP_B - P_B + P_AP_B = P_A - P_B$, se tiene

$$(P_A - P_B)\left(\left(\sqrt{P_AQ_B} - \sqrt{Q_AP_B}\right)^2 + 2\sqrt{P_AQ_AP_BQ_B} - 2\phi\sqrt{P_AQ_AP_BQ_B}\right) < 0$$

$$(P_A - P_B)\left(\left(\sqrt{P_AQ_B} - \sqrt{Q_AP_B}\right)^2 + 2\sqrt{P_AQ_AP_BQ_B}\,(1-\phi)\right) = K_1K_2 < 0.$$

Dado que $K_2 \geq 0$ se deduce que $AV(\widehat{p}_r) < AV(\widehat{p}_{rq})$ cuando $P_A < P_B$.

□

De forma análoga, se puede deducir que $\widehat{V}(\widehat{p}_r) < \widehat{V}(\widehat{p}_{r.q})$ cuando $\widehat{p}_A < \widehat{p}_B$. Este resultado nos permite definir el siguiente estimador más eficiente que $\widehat{p}_r$:

$$\widehat{p}_{r.e} = \begin{cases} \widehat{p}_r & \text{si } \widehat{p}_A < \widehat{p}_B \\ \\ \widehat{p}_{r.q} & \text{en otro caso} \end{cases} \tag{2.15}$$

Las varianzas de este estimador vienen dadas como sigue. Puesto que

$$AV(\widehat{p}_{r.q}) = AV(1 - \widehat{q}_r) = AV(\widehat{q}_r),$$

se deduce que

$$AV(\widehat{p}_{r.e}) = \begin{cases} AV(\widehat{p}_r) & \text{si } P_A < P_B \\ \\ AV(\widehat{q}_r) & \text{en otro caso} \end{cases}$$

Un estimador de $AV(\widehat{p}_{r.e})$ vendrá dado por

$$\widehat{V}(\widehat{p}_{r.e}) = \begin{cases} \widehat{V}(\widehat{p}_r) & \text{si } \widehat{p}_A < \widehat{p}_B \\ \\ \widehat{V}(\widehat{q}_r) & \text{en otro caso} \end{cases}$$

donde $AV(\widehat{q}_r)$ y $\widehat{V}(\widehat{q}_r)$ pueden determinarse fácilmente a partir de $AV(\widehat{p}_r)$ y $\widehat{V}(\widehat{p}_r)$.

Cuando $\widehat{p}_A = \widehat{p}_B$ se observa que $\widehat{p}_r = \widehat{p}_{r.q}$, lo cual implica que

$$AV(\widehat{p}_r) = AV(\widehat{q}_r)$$

y

$$\widehat{V}(\widehat{p}_r) = \widehat{V}(\widehat{q}_r).$$

2.2.5. Extensión al caso de varias variables auxiliares

Supongamos que el atributo de interés A está asociado con J atributos auxiliares $B_1, \ldots, B_J$. El estimador de razón propuesto en presencia de varios atributos auxiliares vendrá dado por

$$\widehat{p}_{MR} = \sum_{i=1}^{J} w_i \widehat{p}_{ri} = \mathbf{w}\widehat{\mathbf{p}}_r',$$

donde $\widehat{p}_{ri} = \widehat{R}_i P_{B_i}$, $\widehat{R}_i = \widehat{p}_A/\widehat{p}_{B_i}$, $\mathbf{w} = (w_1, \ldots, w_J)$ y $\widehat{\mathbf{p}}_r = (\widehat{p}_{r1}, \ldots, \widehat{p}_{rJ})$.

Los pesos w_i satisfacen la condición $\sum_{i=1}^{J} w_i = 1$ y son calculados de manera que maximicen la precisión del estimador propuesto $\widehat{p}_{MR}$.

La varianza de $\widehat{p}_{MR}$ se puede expresar como $AV(\widehat{p}_{MR}) = \mathbf{w}\mathbf{C}\mathbf{w}'$, donde $\mathbf{C} = (c_{ij})$ es una matriz $J \times J$ definida como $c_{ij} = cov(\widehat{p}_{ri}, \widehat{p}_{rj})$, $i \neq j$, y $c_{ii} = AV(\widehat{p}_{ri})$, con $i, j = 1, \ldots, J$, donde

$$AV(\widehat{p}_{ri}) = \frac{N-n}{N-1}\frac{1}{n}\left(P_A Q_A + R_i^2 P_{B_i} Q_{B_i} - 2R_i\phi_i\sqrt{P_A Q_A P_{B_i} Q_{B_i}}\right),$$

$$cov(\widehat{p}_{ri}, \widehat{p}_{rj}) = \frac{N-n}{N-1}\frac{1}{n}\left(P_A Q_A + R_i R_j \phi_{ij}\sqrt{P_{B_i} Q_{B_i} P_{B_j} Q_{B_j}} - \right.$$

$$\left. R_i\phi_i\sqrt{P_A Q_A P_{B_i} Q_{B_i}} - R_j\phi_j\sqrt{P_A Q_A P_{B_j} Q_{B_j}}\right),$$

ϕ_i es el coeficiente V de Cramer asociado a los atributos de A y B_i y ϕ_{ij} es el coeficiente V de Cramer asociado a B_i y B_j.

El sesgo del estimador $\widehat{p}_{MR}$ viene dado por

$$B(\widehat{p}_{MR}) = \frac{N-n}{N-1}\frac{1}{n}\sum_{i=1}^{J} w_i\left(\frac{Q_{B_i}}{P_{B_i}} - \frac{\phi_i\sqrt{P_A Q_A P_{B_i} Q_{B_i}}}{P_A}\right).$$

Dado que $\mathbf{C}$ es semidefinida positiva, la desigualdad de Cauchy-Schwarz generalizada puede usarse para demostrar que el valor óptimo de $\mathbf{w}$ que maximiza $AV(\widehat{p}_{MR})$ es

$$\mathbf{w}_{opt} = \frac{\mathbf{e}\mathbf{C}^{-1}}{\mathbf{e}\mathbf{C}^{-1}\mathbf{e}'},$$

donde $\mathbf{e} = (1, \ldots, 1)$. Sustituyendo $\mathbf{w}_{opt}$ en $AV(\widehat{p}_{MR})$ se obtiene la varianza mínima

$$AV_{min}(\widehat{p}_{MR}) = \frac{1}{\mathbf{e}\mathbf{C}^{-1}\mathbf{e}'}.$$

Dado que $\widehat{p}_{MR}$ es desconocido en la práctica, se propone el estimador de razón múltivariante

$$\widehat{p}_{mr} = \widehat{\mathbf{w}}_{opt}\widehat{\mathbf{p}}'_r,$$

donde

$$\widehat{\mathbf{w}}_{opt} = \frac{\mathbf{e}\widehat{\mathbf{C}}^{-1}}{\mathbf{e}\widehat{\mathbf{C}}^{-1}\mathbf{e}'},$$

$\widehat{\mathbf{C}} = (\widehat{c}_{ij})$, $\widehat{c}_{ij} = \widehat{cov}(\widehat{p}_{ri}, \widehat{p}_{rj})$, $i \neq j$, y $\widehat{c}_{ii} = \widehat{V}(\widehat{p}_{ri})$, con $i, j = 1, \ldots, J$, y

$$\widehat{V}(\widehat{p}_{ri}) = \frac{1-f}{n-1}\left(\widehat{p}_A\widehat{q}_A + \widehat{R}_i^2\widehat{p}_{B_i}\widehat{q}_{B_i} - 2\widehat{R}_i\widehat{\phi}_i\sqrt{\widehat{p}_A\widehat{q}_A\widehat{p}_{B_i}\widehat{q}_{B_i}}\right),$$

$$\widehat{cov}(\widehat{p}_{ri}, \widehat{p}_{rj}) = \frac{1-f}{n-1}\left(\widehat{p}_A\widehat{q}_A + \widehat{R}_i\widehat{R}_j\widehat{\phi}_{ij}\sqrt{\widehat{p}_{B_i}\widehat{q}_{B_i}\widehat{p}_{B_j}\widehat{q}_{B_j}} -\right.$$

$$\left.\widehat{R}_i\widehat{\phi}_i\sqrt{\widehat{p}_A\widehat{q}_A\widehat{p}_{B_i}\widehat{q}_{B_i}} - \widehat{R}_j\widehat{\phi}_j\sqrt{\widehat{p}_A\widehat{q}_A\widehat{p}_{B_j}\widehat{q}_{B_j}}\right).$$

De acuerdo con el caso de un único atributo auxiliar, un estimador de razón más eficiente en el caso de varios atributos auxiliares vendría dado por:

$$\widehat{p}_{mr.e} = \widehat{\mathbf{w}}_{opt.e}\widehat{\mathbf{p}}'_{r.e},$$

donde

$$\widehat{\mathbf{w}}_{opt.e} = \frac{\mathbf{e}\widehat{\mathbf{D}}^{-1}}{\mathbf{e}\widehat{\mathbf{D}}^{-1}\mathbf{e}'},$$

$\widehat{\mathbf{p}}_{r.e} = (\widehat{p}_{r1.e}, \ldots, \widehat{p}_{rJ.e})$,

$$\widehat{p}_{ri.e} = \begin{cases} \widehat{p}_{ri} & \text{si } \widehat{p}_A < \widehat{p}_{B_i} \\ 1 - \widehat{q}_{ri} & \text{en otro caso} \end{cases}$$

siendo $\widehat{\mathbf{D}} = (\widehat{d}_{ij})$, $\widehat{d}_{ij} = \widehat{cov}(\widehat{p}_{ri.e}, \widehat{p}_{rj.e})$, $i \neq j$, y $\widehat{d}_{ii} = \widehat{V}(\widehat{p}_{ri.e})$ con $i, j = 1, \ldots, J$.

2.2.6. Otras propiedades

Presencia de correlación negativa

En la práctica, pueden presentarse situaciones en las que el atributo de interés A tenga una relación negativa fuerte con otros atributos auxiliares, y

el uso de esta información auxiliar en la etapa de estimación también puede proporcionar resultados satisfactorios en este escenario. Sin embargo, a partir de las aportaciones realizadas en la Sección 2.2.3 podemos deducir que los estimadores de razón propuestos $\widehat{p}_r$, $\widehat{p}_{r.u}$ y $\widehat{p}_{r.e}$ requieren una relación positiva entre A y el atributo auxiliar B para que tales estimadores puedan ser más eficientes que el estimador de expansión simple.

Cuando el coeficiente V de Cramer ϕ sea negativo, se propone una transformación del atributo auxiliar de forma que el coeficiente V de Cramer entre el atributo de interés y el atributo auxiliar transformado tenga la misma relación pero positiva.

Dado que

$$\phi_{AB} = \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2}}} = -\frac{N_{12}N_{21} - N_{11}N_{22}}{\sqrt{N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2}}} = -\phi_{AB^c},$$

se propone el uso del atributo B^c como atributo auxiliar cuando el coeficiente V de Cramer entre A y B sea negativo.

Para el caso de varios atributos auxiliares relacionados con el atributo de interés A, los estimadores propuestos $\widehat{p}_{mr}$ y $\widehat{p}_{mr.e}$ también asumen un coeficiente V de Cramer positivo entre A y todos los atributos auxiliares. Siguiendo Rao y Mudholkar (1967), se puede usar una combinación de estimadores de tipo razón y de tipo producto cuando algunas relaciones entre A y los atributos auxiliares sean positivas y el resto negativas. Sin embargo, resulta más simple y sencillo utilizar los atributos complementarios como información auxiliar en aquellos casos con una correlación negativa, de acuerdo con el mencionado caso univariante, y entonces usar directamente los estimadores basados en varios atributos auxiliares propuestos en la Sección 2.2.5.

Estimación de proporciones pequeñas

El estimador $\widehat{p}_r$ no se puede calcular cuando $\widehat{p}_B$ toma el valor 0. Sin embargo, esto no es un problema para el estimador propuesto $\widehat{p}_{r.e}$. De hecho, si $\widehat{p}_B = 0$ se tiene que $\widehat{p}_A \geq \widehat{p}_B$, $\widehat{q}_B = 1 - \widehat{p}_B = 1$, y

$$\widehat{p}_{r.e} = 1 - \widehat{p}_{r.q} = 1 - \frac{\widehat{q}_A}{\widehat{q}_B} Q_B = 1 - \widehat{q}_A Q_B \tag{2.16}$$

según (2.15). Según la expresión (2.16) se tiene que $\widehat{p}_{r.e}$ puede ser calculado para pequeñas proporciones sin ningún problema.

Estimación del complementario de P_A

Como se comentó en la Sección 2.2.4, el estimador estándar $\widehat{p}_A$ tiene el mismo comportamiento en la estimación de P_A que $\widehat{q}_A$ en la estimación de Q_A, puesto que $\widehat{p}_A = 1 - \widehat{q}_A$.

Se observa fácilmente que el estimador $\widehat{p}_r$ no posee esa propiedad. Sin embargo, el estimador propuesto $\widehat{p}_{r.e}$ satisface esta propiedad, esto es, $\widehat{p}_{r.e} = 1-\widehat{q}_{r.e}$, lo cual implica que puede calcularse indistintamente como $\widehat{p}_{r.e}$ o $1-\widehat{q}_{r.e}$.

En efecto, $\widehat{q}_{r.e}$ viene dado por

$$\widehat{q}_{r.e} = \begin{cases} \widehat{q}_r & \text{si } \widehat{q}_A < \widehat{q}_B \\ \\ 1 - \widehat{p}_r & \text{en otro caso} \end{cases}$$

y

$$1 - \widehat{q}_{r.e} = \begin{cases} 1 - \widehat{q}_r & \text{si } \widehat{q}_A < \widehat{q}_B \\ \\ \widehat{p}_r & \text{si } \widehat{q}_A \geq \widehat{q}_B \end{cases}$$

Dado $\widehat{q}_A < \widehat{q}_B$ implica que $\widehat{p}_A > \widehat{p}_B$ y $\widehat{p}_r = 1 - \widehat{q}_r$ cuando $\widehat{p}_A = \widehat{p}_B$, se deduce que

$$1 - \widehat{q}_{r.e} = \begin{cases} 1 - \widehat{q}_r & \text{si } \widehat{p}_A \geq \widehat{p}_B \\ \\ \widehat{p}_r & \text{si } \widehat{p}_A < \widehat{p}_B \end{cases}$$

expresión que coincide con el estimador $\widehat{p}_{r.e}$ definido en (2.15).

2.2.7. Definición del estimador de razón óptimo

En secciones anteriores se definieron los estimadores $\widehat{p}_r$ y $\widehat{p}_{r.q} = 1 - \widehat{q}_r$ para la proporción poblacional P_A. En esta sección definiremos un nuevo estimador de tipo razón mediante una combinación lineal de los mencionados estimadores. El valor óptimo del peso utilizado en la combinación lineal se determinará mediante el criterio de mínima varianza.

El nuevo estimador de tipo razón viene dado por

$$\widehat{p}_{r.w} = w\widehat{p}_r + (1 - w)\widehat{p}_{r.q}. \tag{2.17}$$

Teorema 2.3 *El valor óptimo de w en el sentido de mínima varianza dentro la clase de estimadores $\widehat{p}_{r.w}$ viene dado por*

$$w_{opt} = \frac{AV(\widehat{p}_{r.q}) - cov(\widehat{p}_r, \widehat{p}_{r.q})}{AV(\widehat{p}_r) + AV(\widehat{p}_{r.q}) - 2cov(\widehat{p}_r, \widehat{p}_{r.q})}. \tag{2.18}$$

□

Demostración

A continuación determinaremos el valor óptimo de w de forma que se minimize la varianza de $\widehat{p}_{r.w}$. La varianza asintótica de $\widehat{p}_{r.w}$ viene dada por la siguiente expresión

$$AV(\widehat{p}_{r.w}) = AV(w\widehat{p}_r + (1-w)\widehat{p}_{r.q}) =$$

$$= w^2 AV(\widehat{p}_r) + (1-w)^2 AV(\widehat{p}_{r.q}) + 2w(1-w)cov(\widehat{p}_r, \widehat{p}_{rq}).$$

Denotando $V_1 = AV(\widehat{p}_r)$, $V_2 = V(\widehat{p}_{r.q})$ y $C = cov(\widehat{p}_r, \widehat{p}_{r.q})$, la varianza de $\widehat{p}_{r.w}$ puede expresarse como

$$AV(\widehat{p}_{r.w}) = w^2 V_1 + (1-w)^2 V_2 + 2w(1-w)C.$$

La primera derivada de $AV(\widehat{p}_{r.w})$ respecto w viene dada por

$$\frac{\partial AV(\widehat{p}_{r.w})}{\partial w} = 2wV_1 - 2(1-w)V_2 + 2(1-2w)C = 0,$$

en donde simplificando y despejando obtenemos

$$wV_1 - (1-w)V_2 + (1-2w)C = 0;$$

$$wV_1 - V_2 + wV_2 + C - 2wC = 0;$$

$$w(V_1 + V_2 - 2C) = V_2 - C;$$

$$w_{opt} = \frac{V_2 - C}{V_1 + V_2 - 2C}.$$

A continuación se obtiene la segunda derivada

$$\frac{\partial AV(\widehat{p}_{r.w})}{\partial^2 w} = 2V_1 + 2V_2 - 4C = 2(V_1 + V_2 - 2C) = 2AV(\widehat{p}_r + \widehat{p}_{r.q}) > 0,$$

de la cual se concluye que w_{opt} realmente minimiza $AV(\widehat{p}_{r.w})$.

□

Por tanto, el estimador óptimo en el sentido de mínima varianza dentro de la clase (2.17) viene dado por

$$\widehat{p}_{r.OPT} = w_{opt}\widehat{p}_r + (1 - w_{opt})\widehat{p}_{r.q}.$$

En la práctica, el estimador $\widehat{p}_{r.OPT}$ puede ser desconocido, puesto que el peso w_{opt} depende de varianzas poblacionales, la cuales son generalmente desconocidas. En esta situación, emplearemos el estimador

$$\widehat{p}_{r.opt} = \widehat{w}_{opt}\widehat{p}_r + (1 - \widehat{w}_{opt})\widehat{p}_{r.q}, \tag{2.19}$$

donde

$$\widehat{w}_{opt} = \frac{\widehat{V}(\widehat{p}_{r.q}) - \widehat{cov}(\widehat{p}_r, \widehat{p}_{r.q})}{\widehat{V}(\widehat{p}_r) + \widehat{V}(\widehat{p}_{r.q}) - 2\widehat{cov}(\widehat{p}_r, \widehat{p}_{r.q})}. \tag{2.20}$$

Siguiendo Särndal et al. (1992) pg 372, la varianza de $\widehat{p}_{r.w}$ puede expresarse como

$$AV(\widehat{p}_{r.w}) = (V_1 + V_2 - 2C)\left(w - \frac{V_2 - C}{V_1 + V_2 - 2C}\right)^2 + \frac{V_1V_2 - C^2}{V_1 + V_2 - 2C},$$

de la cual podemos deducir que la varianza del estimador óptimo $\widehat{p}_{r.OPT}$ viene dada por

$$AV(\widehat{p}_{r.OPT}) = \frac{V_1V_2 - C^2}{V_1 + V_2 - 2C}.$$

Esta expresión nos sirve también para obtener el siguiente estimador de la varianza del estimador óptimo $\widehat{p}_{r.opt}$

$$\widehat{V}(\widehat{p}_{r.opt}) = \frac{\widehat{V}(\widehat{p}_r)\widehat{V}(\widehat{p}_{r.q}) - \widehat{cov}^2(\widehat{p}_r, \widehat{p}_{r.q})}{\widehat{V}(\widehat{p}_r) + \widehat{V}(\widehat{p}_{r.q}) - 2\widehat{cov}(\widehat{p}_r, \widehat{p}_{r.q})}.$$

Con ayuda de la técnica de linealización de Taylor, el siguiente teorema proporciona una expresión para $C = cov(\widehat{p}_r, \widehat{p}_{r.q})$ bajo MAS.

Teorema 2.4 *La covarianza entre los estimadores* $\widehat{p}_r$ *y* $\widehat{p}_{r.q}$ *viene dada por*

$$cov(\widehat{p}_r, \widehat{p}_{r.q}) = \frac{N-n}{N-1}\frac{1}{n}\left(P_A Q_A + R R_c P_B Q_B - (R + R_c)\phi\sqrt{P_A Q_A P_B Q_B}\right),$$

donde $R_c = Q_A/Q_B$ *es la razón poblacional de las proporciones complementarias de los atributos* A *y* B.

□

Demostración

Siguiendo el desarrollo en serie de Taylor (véase Särndal et al. 1992, pg 178) se obtiene que $\widehat{R}$ se puede descomponer como

$$\widehat{R} \cong R + \frac{1}{nP_B}\sum_{i\in s}(A_i - RB_i) = R + \frac{1}{P_B}(\widehat{p}_A - R\widehat{p}_B),$$

y del mismo modo $\widehat{R}_c = \widehat{q}_A/\widehat{q}_B$ se puede descomponer como

$$\widehat{R}_c \cong R_c + \frac{1}{Q_B}(\widehat{q}_A - R_c\widehat{q}_B).$$

Teniendo en cuenta las expresiones anteriores obtenemos

$$C = cov(\widehat{p}_r, 1-\widehat{q}_r) = -cov(\widehat{p}_r, \widehat{q}_r) = -cov(\widehat{R}P_B, \widehat{R}_c Q_B) =$$

$$= -P_B Q_B cov(\widehat{R}, \widehat{R}_c) = -P_B Q_B cov\left(R + \frac{1}{P_B}(\widehat{p}_A - R\widehat{P}_B), R_c + \frac{1}{Q_B}(\widehat{q}_A - R_c\widehat{q}_B)\right) =$$

$$= -cov(\widehat{p}_A - R\widehat{p}_B, \widehat{q}_A - R_c\widehat{q}_B) =$$

$$= -[cov(\widehat{p}_A, \widehat{q}_A) - R_c cov(\widehat{p}_A, \widehat{q}_B) - R cov(\widehat{p}_B, \widehat{q}_A) + R R_c cov(\widehat{p}_B, \widehat{q}_B)] =$$

$$= -cov(\widehat{p}_A, 1-\widehat{p}_A) + R_c cov(\widehat{p}_A, 1-\widehat{p}_B) + R cov(\widehat{p}_B, 1-\widehat{p}_A) - R R_c cov(\widehat{p}_B, 1-\widehat{p}_B) =$$

$$= V(\widehat{p}_A) - R_c cov(\widehat{p}_A, \widehat{p}_B) - R cov(\widehat{p}_A, \widehat{p}_B) + R R_c V(\widehat{p}_B) =$$

$$= V(\widehat{p}_A) + R R_c V(\widehat{p}_B) - (R + R_c) cov(\widehat{p}_A, \widehat{p}_B) =$$

$$= \frac{N-n}{N-1}\frac{1}{n}\left(P_A Q_A + R R_c P_B Q_B - (R + R_c)\phi\sqrt{P_A Q_A P_B Q_B}\right).$$

□

Un estimador de la covarianza $cov(\widehat{p}_r, \widehat{p}_{r.q})$, el cual necesitaremos para obtener el peso óptimo $\widehat{w}_{opt}$, viene dado por

$$\widehat{cov}(\widehat{p}_r, \widehat{p}_{r.q}) = \frac{1-f}{n-1}\left(\widehat{p}_A\widehat{q}_A + \widehat{R}\widehat{R}_c\widehat{p}_B\widehat{q}_B - (\widehat{R} + \widehat{R}_c)\widehat{\phi}\sqrt{\widehat{p}_A\widehat{q}_A\widehat{p}_B\widehat{q}_B}\right).$$

Teorema 2.5 *El peso óptimo w_{opt} dado en la expresión (2.18) puede obtenerse como*

$$w_{opt} = \frac{R_c - \beta}{R_c - R},$$

donde

$$\beta = \frac{cov(\widehat{p}_A, \widehat{p}_B)}{V(\widehat{p}_B)}.$$

□

Demostración

Teniendo en cuenta que

$$V_1 = V(\widehat{p}_A) + R^2 V(\widehat{p}_B) - 2R cov(\widehat{p}_A, \widehat{p}_B)],$$

$$V_2 = V(\widehat{q}_A) + R_c^2 V(\widehat{q}_B) - 2R_c cov(\widehat{q}_A, \widehat{q}_B) =$$
$$= V(\widehat{p}_A) + R_c^2 V(\widehat{p}_B) - 2R_c cov(\widehat{p}_A, \widehat{p}_B)$$

y

$$C = V(\widehat{p}_A) + R R_c V(\widehat{p}_B) - (R + R_c) cov(\widehat{p}_A, \widehat{p}_B),$$

el numerador y denominador de w_{opt} obtenida en la expresión (2.18) vienen dados por

$$V_2 - C = V(\widehat{p}_B)(R_c^2 - RR_c) - cov(\widehat{p}_A, \widehat{p}_B)[2R_c - (R - R_c)] =$$

$$= V(\widehat{p}_B)R_c(R_c - R) - cov(\widehat{p}_A, \widehat{p}_B)(R_c - R) =$$

$$= (R_c - R)[V(\widehat{p}_B)R_c - cov(\widehat{p}_A, \widehat{p}_B)]$$

y

$$V_1 + V_2 - 2C = V(\widehat{p}_B)(R^2 + R_c^2 - 2RR_c) - cov(\widehat{p}_A, \widehat{p}_B)[2R + 2R_c - 2(R + R_c)] =$$

$$= V(\widehat{p}_B)(R_c - R)^2$$

Sustituyendo dichas expresiones en w_{opt} se obtiene

$$w_{opt} = \frac{V_2 - C}{V_1 + V_2 - 2C} = \frac{(R_c - R)[V(\widehat{p}_B)R_c - cov(\widehat{p}_A, \widehat{p}_B)]}{V(\widehat{p}_B)(R_c - R)^2} =$$

$$= \frac{V(\widehat{p}_B)R_c - cov(\widehat{p}_A, \widehat{p}_B)}{(R_c - R)V(\widehat{p}_B)} = \frac{R_c - \beta}{R_c - R}.$$

□

Teniendo en cuenta el Teorema 2.5, el peso óptimo estimado $\widehat{w}_{opt}$ dado en la expresión (2.20) tendrá la siguiente expresión bajo MAS

$$\widehat{w}_{opt} = \frac{\widehat{R}_c - \widehat{\beta}}{\widehat{R}_c - \widehat{R}}, \tag{2.21}$$

donde

$$\widehat{\beta} = \frac{\widehat{cov}(\widehat{p}_A, \widehat{p}_B)}{\widehat{V}(\widehat{p}_B)}.$$

A partir de la expresión (2.21) concluimos que $\widehat{w}_{opt} = 1$, es decir $\widehat{p}_{r.opt} = \widehat{p}_r$, si $\widehat{\beta} = \widehat{R}$, mientras que $\widehat{w}_{opt} = 0$, es decir $\widehat{p}_{r.opt} = \widehat{p}_{r.q}$, si $\widehat{\beta} = \widehat{R}_c$. En otras palabras, a medida que $\widehat{\beta}$ se aproxime a $\widehat{R}$, el estimador de tipo razón $\widehat{p}_r$

tendrá mas peso en el estimador óptimo $\widehat{p}_{r.opt}$. En el lado opuesto, a medida que $\widehat{\beta}$ se aproxime a $\widehat{R}_c$, el estimador de tipo razón $\widehat{p}_{r.q}$ tendrá mas peso en el estimador de razón óptimo.

Una expresión alternativa para $\widehat{w}_{opt}$ bajo MAS viene dada por

$$\widehat{w}_{opt} = \frac{\widehat{R}_c - \widehat{\phi}\sqrt{\widehat{R}\widehat{R}_c}}{\widehat{R}_c - \widehat{R}},$$

puesto que

$$\widehat{\beta} = \frac{\widehat{cov}(\widehat{p}_A, \widehat{p}_B)}{\widehat{V}(\widehat{p}_B)} = \frac{\widehat{\phi}\sqrt{\widehat{p}_A\widehat{q}_A\widehat{p}_B\widehat{q}_B}}{\widehat{p}_B\widehat{q}_B} = \widehat{\phi}\sqrt{\frac{\widehat{p}_A\widehat{q}_A}{\widehat{p}_B\widehat{q}_B}} = \widehat{\phi}\sqrt{\widehat{R}\widehat{R}_c}.$$

Teorema 2.6 *El estimador de tipo razón óptimo $\widehat{p}_{r.opt}$ dado en la expresión (2.19) puede obtenerse bajo MAS como*

$$\widehat{p}_{r.opt} = \widehat{p}_A + \widehat{\beta}(P_B - \widehat{p}_B).$$

□

Demostración

$$\widehat{p}_{r.opt} = \widehat{w}_{opt}\widehat{p}_r + (1 - \widehat{w}_{opt})\widehat{p}_{r.q} = \frac{\widehat{R}_c - \widehat{\beta}}{\widehat{R}_c - \widehat{R}}\widehat{R}P_B + \left(1 - \frac{\widehat{R}_c - \widehat{\beta}}{\widehat{R}_c - \widehat{R}}\right)(1 - \widehat{R}_c Q_B) =$$

$$= \frac{\widehat{R}_c\widehat{R}P_B - \widehat{\beta}\widehat{R}P_B}{\widehat{R}_c - \widehat{R}} + \frac{\widehat{R}_c - \widehat{R} - \widehat{R}_c + \widehat{\beta}}{\widehat{R}_c - \widehat{R}}\Big(1 - \widehat{R}_c(1 - P_B)\Big) =$$

$$= \frac{\widehat{R}_c\widehat{R}P_B - \widehat{\beta}\widehat{R}P_B - \widehat{R} + \widehat{R}\widehat{R}_c - \widehat{R}_c\widehat{R}P_B + \widehat{\beta} - \widehat{\beta}\widehat{R}_c + \widehat{\beta}\widehat{R}_c P_B}{\widehat{R}_c - \widehat{R}} =$$

$$= \frac{\widehat{\beta}P_B(\widehat{R}_c - \widehat{R}) + (\widehat{R}_c - 1)(\widehat{R} - \widehat{\beta})}{\widehat{R}_c - \widehat{R}} = \widehat{\beta}P_B + \frac{(\dfrac{\widehat{q}_A}{\widehat{q}_B} - 1)(\dfrac{\widehat{p}_A}{\widehat{p}_B} - \widehat{\beta})}{\dfrac{\widehat{q}_A}{\widehat{q}_B} - \dfrac{\widehat{p}_A}{\widehat{p}_B}} =$$

$$= \widehat{\beta}P_B + \frac{\dfrac{1-\widehat{p}_A-1+\widehat{p}_B}{1-\widehat{p}_B}\dfrac{\widehat{p}_A-\widehat{\beta}\widehat{p}_B}{\widehat{p}_B}}{\dfrac{(1-\widehat{p}_A)\widehat{p}_B-\widehat{p}_A(1-\widehat{p}_B)}{(1-\widehat{p}_B)\widehat{p}_B}} =$$

$$= \widehat{\beta}P_B + \frac{(\widehat{p}_B-\widehat{p}_A)(\widehat{p}_A-\widehat{\beta}\widehat{p}_B)}{\widehat{p}_B-\widehat{p}_A\widehat{p}_B-\widehat{p}_A+\widehat{p}_A\widehat{p}_B} =$$

$$= \widehat{p}_A + \widehat{\beta}(P_B-\widehat{p}_B).$$

□

Notamos que la expresión $\widehat{p}_A + \widehat{\beta}(P_B - \widehat{p}_B)$ obtenida para el estimador de tipo razón óptimo también se obtendrá en el Capítulo 3 mediante el método de regresión para la obtención de estimadores de una proporción poblacional. En dicho Capitulo también se analizarán en detalle otras propiedades relevantes para este estimador, como es la insegadez.

Teorema 2.7 *La varianza asintótica del estimador de tipo razón óptimo $\widehat{p}_{r.OPT}$ tiene la siguiente expresión bajo MAS*

$$AV(\widehat{p}_{r.OPT}) = V(\widehat{p}_A)(1-\phi^2).$$

□

Demostración

La varianza asintótica del estimador $\widehat{p}_{r.OPT}$ venía dada por

$$AV(\widehat{p}_{r.OPT}) = \frac{V_1V_2 - C}{V_1 + V_2 - 2C},$$

donde el denominador de dicha expresión, según vimos en la demostración del Teorema 2.5, se puede obtener como

$$V_1 + V_2 - 2C = V(\widehat{p}_B)(R - R_c)^2.$$

A continuación obtendremos el numerador de $AV(\widehat{p}_{r.opt})$. Por razones de claridad en la presentación, denotaremos $V_A = V(\widehat{p}_A)$, $V_B = V(\widehat{p}_B)$ y $C_{AB} = cov(\widehat{p}_A, \widehat{p}_B)$. Recordamos que

$$V_1 = V_A + R^2V_B - 2RC_{AB},$$

$$V_2 = V_A + R_c^2 V_B - 2R_c C_{AB}$$

y

$$C = V_A + RR_c V_B - (R + R_c)C_{AB}.$$

Por un lado,

$$V_1 V_2 = V_A^2 + R_c^2 V_A V_B - 2R_c V_A C_{AB} + R^2 V_A V_B + R^2 R_c^2 V_B^2$$

$$-2R^2 R_c V_B C_{AB} - 2RV_A C_{AB} - 2RR_c^2 V_B C_{AB} + 4RR_c C_{AB}^2.$$

La covarianza al cuadrado se puede expresar como

$$C^2 = V_A^2 + R^2 R_c^2 V_B^2 + (R + R_c)^2 C_{AB}^2 + 2V_A RR_c V_B$$

$$-2(R + R_c)V_A C_{AB} - 2RR_c(R + R_c)V_B C_{AB} =$$

$$= V_A^2 + R^2 R_c^2 V_B^2 + R^2 C_{AB}^2 + R_c^2 C_{AB}^2 + 2RR_c C_{AB}^2 + 2V_A RR_c V_B$$

$$-2RV_A C_{AB} - 2R_c V_A C_{AB} - 2R^2 R_c V_B C_{AB} - 2RR_c^2 V_B C_{AB}.$$

Por lo que el numerador de $AV(\widehat{p}_{r.OPT})$ queda

$$V_1 V_2 - C^2 = V_A V_B(R_c^2 - 2RR_c + R^2) - C_{AB}^2(R^2 + R_c^2 - 2RR_c) =$$

$$= (V_A V_B - C_{AB}^2)(R - R_c)^2.$$

La varianza de $\widehat{p}_{r.OPT}$ se puede obtener también como

$$AV(\widehat{p}_{r.OPT}) = \frac{V_A V_B - C_{AB}^2}{V_B} = \frac{V(\widehat{p}_A)V(\widehat{p}_B) - cov(\widehat{p}_A, \widehat{p}_B)^2}{V(\widehat{p}_B)}. \qquad (2.22)$$

Sustituyendo $V(\widehat{p}_A)$, $V(\widehat{p}_B)$ y $cov(\widehat{p}_A, \widehat{p}_B)$ en la expresión (2.22) por sus correspondientes expresiones bajo MAS, obtenemos

$$AV(\widehat{p}_{r.OPT}) = \frac{N-n}{N-1}\frac{1}{n}\left[\frac{P_A Q_A P_B Q_B - \phi^2 P_A Q_A P_B Q_B}{P_B Q_B}\right] =$$

$$= \frac{N-n}{N-1}\frac{1}{n}P_A Q_A(1 - \phi^2) = V(\widehat{p}_A)(1 - \phi^2).$$

□

La comparación teórica del estimador de tipo razón $\widehat{p}_{r.OPT}$ con respecto el estimador de expansión simple $\widehat{p}_A$ es bastante simple a partir del Teorema 2.7. En efecto, $\widehat{p}_{r.OPT}$ será siempre más eficiente que $\widehat{p}_A$, puesto que $AV(\widehat{p}_{r.OPT}) <= V(\widehat{p}_A)$ al tomar ϕ^2 valores dentro del intervalo [0,1]. Ambos estimadores serán igual de eficientes cuando $\phi^2 = 0$.

Siguiendo el Teorema 2.7, un estimador de la varianza del estimador de tipo razón óptimo $\widehat{p}_{r.OPT}$ puede obtenerse, bajo MAS, como

$$\widehat{V}(\widehat{p}_{r.OPT}) = \widehat{V}(\widehat{p}_A)(1 - \widehat{\phi}^2).$$

2.3. Extensión a un diseño muestral general

2.3.1. Definición del estimador

La estimación puntual y por intervalos de confianza de una proporción poblacional han sido estudiados normalmente bajo la suposición de variables aleatorias independientes e idénticamente distribuidas. Sin embargo, el problema de la estimación de una proporción poblacional cuando las muestras se extraen bajo un diseño de muestreo general ha recibido menos atención.

En la mayoría de las encuestas se asumen diseños muestrales complejos, y el uso de métodos de estimación que tengan en cuenta los pesos muestrales pueden ofrecer una mejor estimación que los enfoques habituales que no tienen el efecto del diseño muestral en consideración.

En esta sección se resuelve el problema de la estimación de una proporción poblacional cuando las muestras son seleccionadas mediante un diseño muestral general. En esta sección asumiremos, por tanto, que la muestra s se ha seleccionado de acuerdo a un diseño muestral específico con probabilidades de inclusión π_i y π_{ij} estrictamente positivas. El estimador estándar $\widehat{p}_A$ así como los estimadores de tipo razón propuestos a lo largo de este capítulo no son apropiados bajo esta suposición de probabilidades de selección desiguales.

Sin usar ninguna información auxiliar, un estimador simple para P_A que tiene en cuenta el diseño muestral viene dado por

$$\widehat{p}_{A.HT} = \frac{1}{N} \sum_{i \in s} d_i A_i, \tag{2.23}$$

donde $d_i = \pi_i^{-1}$ es el peso del diseño. La varianza de $\widehat{p}_{A.HT}$ puede obtenerse como

$$V(\widehat{p}_{A.HT}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\frac{A_i}{\pi_i}\frac{A_j}{\pi_j}, \tag{2.24}$$

mientras que un estimador insesgado de dicha varianza viene dado por

$$\widehat{V}(\widehat{p}_{A.HT}) = \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{A_i}{\pi_i}\frac{A_j}{\pi_j}. \tag{2.25}$$

Hay situaciones en las que el tamaño de la población N es desconocido o incluso siendo conocido, su uso en (2.23) puede llevar a estimaciones de P_A mayores de 1. En esta situación, una solución simple es utilizar un estimador tipo Hájek (véase Hájek, 1964), que viene dado por

$$\widehat{p}_{A.H} = \frac{1}{\widehat{N}}\sum_{i\in s} d_i A_i,$$

donde $\widehat{N} = \sum_{i\in s} d_i$. El uso de $\widehat{N}$ en lugar de N en la expresión (2.23) implica que $\widehat{p}_{A.H} = 1$ cuando $A_i = 1$ para todo $i \in s$, lo que nos garantiza que las estimaciones estarán dentro del intervalo $[0, 1]$.

Usando información auxiliar, el primer estimador de tipo razón que proponemos viene dado por

$$\widehat{p}_{r.HT} = \widehat{R}_{HT}P_B, \tag{2.26}$$

donde $\widehat{R}_{HT} = \widehat{p}_{A.HT}/\widehat{p}_{B.HT}$ y $\widehat{p}_{B.HT}$ se define como (2.23) después de sustituir A_i por el atributo auxiliar B_i.

Destacamos que también se podrían formular estimadores de tipo razón utilizando estimadores de tipo Hájek ($\widehat{p}_{A.H}$ y $\widehat{p}_{B.H}$) en lugar de estimadores de tipo Horvitz-Thompson ($\widehat{p}_{A.HT}$ y $\widehat{p}_{B.HT}$). Sin embargo, la propiedades teóricas de los estimadores de tipo razón obtenidos a partir de estimadores de tipo Hájek serían más difíciles de establecer que si utilizamos estimadores de tipo Horvitz-Thompson, puesto que los estimadores directos de tipo Hájek se definen como el cociente entre dos estimadores.

2.3.2. Propiedades teóricas

Proposición 2.4 *La varianza del estimador de tipo razón $\widehat{p}_{r.HT}$ viene dada por*

$$AV(\widehat{p}_{r.HT}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\frac{A_i - RB_i}{\pi_i}\frac{A_j - RB_j}{\pi_j}. \tag{2.27}$$

□

Demostración.

Partiendo de los desarrollos en la serie de Taylor para un diseño muestral general, la razón $\widehat{R}_{HT}$ puede descomponerse como

$$\widehat{R}_{HT} \cong R + \frac{1}{NP_B}\sum_{i\in s}\frac{A_i - RB_i}{\pi_i}.$$

Multiplicando la expresión anterior por P_B obtenemos

$$\widehat{p}_{r.HT} \cong P_A + \frac{1}{N}\sum_{i\in s}\frac{A_i - RB_i}{\pi_i}. \tag{2.28}$$

Sea I_i la variable indicadora muestral, es decir, $I_i = 1$ si $i \in s$ y $I_i = 0$ en otro caso, de la cual es conocido que $E(I_i) = \pi_i$, $V(I_i) = \pi_i(1-\pi_i)$ y $cov(I_i, I_j) = \pi_{ij} - \pi_i\pi_j$.

A partir de la expresión (2.28), la varianza asintótica de $\widehat{p}_{r.HT}$ se puede obtener como

$$AV(\widehat{p}_{r.HT}) = \frac{1}{N^2}V\left(\sum_{i\in s}\frac{A_i - RB_i}{\pi_i}\right) = \frac{1}{N^2}V\left(\sum_{i=1}^{N}\frac{A_i - RB_i}{\pi_i}I_i\right) =$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N}V\left(\frac{A_i - RB_i}{\pi_i}I_i\right) + \sum_{i\neq j}^{N}cov\left(\frac{A_i - RB_i}{\pi_i}I_i, \frac{A_j - RB_j}{\pi_j}I_j\right)\right] =$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N}\frac{A_i - RB_i}{\pi_i}V(I_i) + \sum_{i\neq j}^{N}\frac{A_i - RB_i}{\pi_i}\frac{A_j - RB_j}{\pi_j}cov(I_i, I_j)\right] =$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\frac{A_i - RB_i}{\pi_i}\frac{A_j - RB_j}{\pi_j}.$$

□

Corolario 2.1 *Un estimador insesgado de* $AV(\widehat{p}_{r.HT})$ *viene dado por*

$$\widehat{V}(\widehat{p}_{r.HT}) = \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{A_i-\widehat{R}_{HT}B_i}{\pi_i}\frac{A_j-\widehat{R}_{HT}B_j}{\pi_j}.$$

□

Proposición 2.5 *Si el diseño muestral es de tamaño fijo, una expresión alternativa de la varianza del estimador de tipo razón* $\widehat{p}_{r.HT}$ *viene dada por*

$$AV(\widehat{p}_{r.HT}) = -\frac{1}{2N^2}\sum_{i\neq j}^{N}(\pi_{ij}-\pi_i\pi_j)\left(\frac{A_i-RB_i}{\pi_i}-\frac{A_j-RB_j}{\pi_j}\right)^2. \qquad (2.29)$$

□

Demostración.

Tenemos que verificar la equivalencia entre las expresiones (2.27) y (2.29) cuando el tamaño del diseño muestra está fijado en n. Desarrollando el cuadrado en (2.29) y sumando obtenemos

$$AV(\widehat{p}_{r.HT}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\frac{A_i-RB_i}{\pi_i}\frac{A_j-RB_j}{\pi_j}$$

$$-\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\left(\frac{A_i-RB_i}{\pi_i}\frac{A_j-RB_j}{\pi_j}\right)^2.$$

Por otra parte,

$$\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\left(\frac{A_i-RB_i}{\pi_i}\frac{A_j-RB_j}{\pi_j}\right)^2 =$$

$$=\sum_{i=1}^{N}\left(\frac{A_i-RB_i}{\pi_i}\frac{A_j-RB_j}{\pi_j}\right)^2\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j).$$

Usando el Resultado 2.6.2 de Särndal et al. (1992), pg. 38 para diseños con tamaño de muestra fijo, obtenemos

$$\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)=\sum_{j=1}^{N}\pi_{ij}-\sum_{j=1}^{N}\pi_i\pi_j = n\pi_i - n\pi_i = 0,$$

lo cual demuestra la equivalencia entre las expresiones (2.27) y (2.29).

□

Corolario 2.2 *Si el diseño muestral es de tamaño fijo, un estimador insesgado de la varianza $AV(\widehat{p}_{r.HT})$ dada en (2.29) viene dado por*

$$\widehat{V}(\widehat{p}_{r.HT}) = -\frac{1}{2N^2}\sum_{i\neq j}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\left(\frac{A_i-\widehat{R}_{HT}B_i}{\pi_i}-\frac{A_j-\widehat{R}_{HT}B_j}{\pi_j}\right)^2$$

□

2.3.3. Definición de otros estimadores

Siguiendo la Sección 2.2.4, podemos definir un estimador de tipo razón más eficiente que $\widehat{p}_{r.HT}$ como

$$\widehat{p}_{r.e.HT} = \begin{cases} \widehat{p}_{r.HT} & \text{si } \widehat{V}(\widehat{p}_{r.HT}) < \widehat{V}(\widehat{p}_{r.q.HT}) \\ \\ \widehat{p}_{r.q.HT} & \text{en otro caso} \end{cases}$$

donde $\widehat{p}_{r.q.HT} = 1-\widehat{q}_{r.HT} = 1-\widehat{R}_{c.HT}Q_B$ es el estimador de tipo razón para P_A obtenido a partir del complementario, $\widehat{R}_{c.HT} = \widehat{q}_{A.HT}/\widehat{q}_{B.HT}$ y

$$\widehat{V}(\widehat{p}_{r.q.HT}) = \widehat{V}(\widehat{q}_{r.HT}) = \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{A_i^c-\widehat{R}_{c.HT}B_i^c}{\pi_i}\frac{A_j^c-\widehat{R}_{c.HT}B_j^c}{\pi_j}.$$

No obstante, como estudiamos en la Sección 2.2.7, resulta más atractivo definir un estimador óptimo en el sentido de mínima varianza a partir de los estimadores $\widehat{p}_{r.HT}$ y $\widehat{p}_{r.q.HT}$. La nueva clase de estimadores para un diseño muestral general viene dado por

$$\widehat{p}_{r.w.HT} = w\widehat{p}_{r.HT} + (1-w)\widehat{p}_{r.q.HT}.$$

A partir del Teorema 2.3, el valor óptimo de w en el sentido de mínima varianza dentro la clase de estimadores $\widehat{p}_{r.w.HT}$ viene dado por

$$w_{opt} = \frac{AV(\widehat{p}_{r.q.HT}) - cov(\widehat{p}_{r.HT},\widehat{p}_{r.q.HT})}{AV(\widehat{p}_{r.HT}) + AV(\widehat{p}_{r.q.HT}) - 2cov(\widehat{p}_{r.HT},\widehat{p}_{r.q.HT})}.$$

Con ayuda del valor óptimo anterior, podemos definir el siguiente estimador óptimo en el sentido de mínima varianza

$$\widehat{p}_{r.OPT.HT} = w_{opt}\widehat{p}_{r.HT} + (1 - w_{opt})\widehat{p}_{r.q.HT}.$$

En la práctica, $\widehat{p}_{r.OPT.HT}$ puede ser desconocido, aunque podemos aproximarlo por el estimador

$$\widehat{p}_{r.opt.HT} = \widehat{w}_{opt}\widehat{p}_{r.HT} + (1 - \widehat{w}_{opt})\widehat{p}_{r.q.HT},$$

donde

$$\widehat{w}_{opt} = \frac{\widehat{V}(\widehat{p}_{r.q.HT}) - \widehat{cov}(\widehat{p}_{r.HT}, \widehat{p}_{r.q.HT})}{\widehat{V}(\widehat{p}_{r.HT}) + \widehat{V}(\widehat{p}_{r.q.HT}) - 2\widehat{cov}(\widehat{p}_{r.HT}, \widehat{p}_{r.q.HT})}.$$

Por último, las expresiones para las varianzas aproximadas de los estimadores óptimos vendrán dadas por

$$AV(\widehat{p}_{r.opt.HT}) = \frac{AV(\widehat{p}_{r.HT})AV(\widehat{p}_{r.q.HT}) - cov^2(\widehat{p}_{r.HT}, \widehat{p}_{r.q.HT})}{AV(\widehat{p}_{r.HT}) + AV(\widehat{p}_{r.q.HT}) - 2cov(\widehat{p}_{r.HT}, \widehat{p}_{r.q.HT})}$$

y su estimador

$$\widehat{V}(\widehat{p}_{r.opt.HT}) = \frac{\widehat{V}(\widehat{p}_{r.HT})\widehat{V}(\widehat{p}_{r.q.HT}) - \widehat{cov}^2(\widehat{p}_{r.HT}, \widehat{p}_{r.q.HT})}{\widehat{V}(\widehat{p}_{r.HT}) + \widehat{V}(\widehat{p}_{r.q.HT}) - 2\widehat{cov}(\widehat{p}_{r.HT}, \widehat{p}_{r.q.HT})}.$$

El siguiente teorema proporciona una expresión para $cov(\widehat{p}_{r.HT}, \widehat{p}_{r.q.HT})$ bajo un diseño muestral general, lo que nos permitirá obtener una estimación de la mencionada covarianza así como la posibilidad de obtener $\widehat{w}_{opt}$ y $\widehat{V}(\widehat{p}_{r.opt.HT})$ en la práctica.

Teorema 2.8 *La covarianza entre los estimadores $\widehat{p}_r$ y $\widehat{p}_{r.q}$ viene dada por*

$$cov(\widehat{p}_{r.HT}, \widehat{p}_{r.q.HT}) = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\frac{A_i - RB_i}{\pi_i}\frac{A_j^c - R_cB_j^c}{\pi_j}.$$

□

Demostración

Teniendo que cuenta que

$$\widehat{p}_{r.HT} \cong P_A + \frac{1}{N}\sum_{i\in s}\frac{A_i - RB_i}{\pi_i}$$

y

$$\widehat{q}_{r.HT} \cong Q_A + \frac{1}{N}\sum_{i\in s}\frac{A_i^c - R_c B_i^c}{\pi_i}$$

obtenemos

$$cov(\widehat{p}_{r.HT}, \widehat{p}_{r.q.HT}) = cov(\widehat{p}_{r.HT}, 1-\widehat{q}_{r.HT}) = -cov(\widehat{p}_{r.HT}, \widehat{q}_{r.HT}) =$$

$$= -cov\left(P_A + \frac{1}{N}\sum_{i\in s}\frac{A_i - RB_i}{\pi_i}, Q_A + \frac{1}{N}\sum_{i\in s}\frac{A_i^c - R_c B_i^c}{\pi_i}\right) =$$

$$= -\frac{1}{N^2}cov\left(\sum_{i\in s}\frac{A_i - RB_i}{\pi_i}, \sum_{i\in s}\frac{A_i^c - R_c B_i^c}{\pi_i}\right) =$$

$$= -\frac{1}{N^2}cov\left(\sum_{i=1}^{N}\frac{A_i - RB_i}{\pi_i}I_i, \sum_{i=1}^{N}\frac{A_i^c - R_c B_i^c}{\pi_i}I_i\right) =$$

$$= -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}cov\left(\frac{A_i - RB_i}{\pi_i}I_i, \frac{A_j^c - R_c B_j^c}{\pi_j}I_j\right) =$$

$$= -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{A_i - RB_i}{\pi_i}\frac{A_j^c - R_c B_j^c}{\pi_j}cov(I_i, I_j) =$$

$$= -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\frac{A_i - RB_i}{\pi_i}\frac{A_j^c - R_c B_j^c}{\pi_j}.$$

□

Un estimador de la covarianza $cov(\widehat{p}_{r.HT}, \widehat{p}_{r.q.HT})$ viene dado por

$$\widehat{cov}(\widehat{p}_{r.HT}, \widehat{p}_{r.q.HT}) = -\frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{A_i - \widehat{R}_{HT}B_i}{\pi_i}\frac{A_j^c - \widehat{R}_{c.HT}B_j^c}{\pi_j}.$$

Siguiendo el Teorema 2.5, el peso óptimo w_{opt} puede obtenerse como

$$w_{opt} = \frac{R_c - \beta_{HT}}{R_c - R},$$

donde

$$\beta_{HT} = \frac{cov(\widehat{p}_{A.HT}, \widehat{p}_{B.HT})}{V(\widehat{p}_{B.HT})},$$

$V(\widehat{p}_{B.HT})$ viene dada por la expresión (2.24) después de sustituir A_i por B_i y

$$cov(\widehat{p}_{A.HT}, \widehat{p}_{B.HT}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \frac{A_i}{\pi_i} \frac{B_j}{\pi_j}.$$

Un estimador de w_{opt} puede obtenerse como

$$\widehat{w}_{opt} = \frac{\widehat{R}_{c.HT} - \widehat{\beta}_{HT}}{\widehat{R}_{c.HT} - \widehat{R}_{HT}},$$

donde

$$\widehat{\beta}_{HT} = \frac{\widehat{cov}(\widehat{p}_{A.HT}, \widehat{p}_{B.HT})}{\widehat{V}(\widehat{p}_{B.HT})},$$

$\widehat{V}(\widehat{p}_{B.HT})$ viene dada por la expresión (2.25) después de sustituir A_i por B_i y

$$\widehat{cov}(\widehat{p}_{A.HT}, \widehat{p}_{B.HT}) = \frac{1}{N^2} \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{A_i}{\pi_i} \frac{B_j}{\pi_j}.$$

Finalmente, siguiendo el Teorema 2.6, podemos concluir que el estimador de tipo razón óptimo $\widehat{p}_{r.opt.HT}$ puede obtenerse bajo un diseño muestral general como

$$\widehat{p}_{r.opt.HT} = \widehat{p}_{A.HT} + \widehat{\beta}_{HT}(P_B - \widehat{p}_{B.HT}). \tag{2.30}$$

Aplicando la técnica de linealización de Taylor en el estimador $\widehat{p}_{r.opt.HT}$ dado en (2.30), la varianza estimada de $\widehat{p}_{r.opt.HT}$ puede obtenerse bajo un diseño muestral general como

$$\widehat{V}(\widehat{p}_{r.opt.HT}) = \frac{1}{N^2} \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{A_i - \widehat{A}_i}{\pi_i} \frac{A_j - \widehat{A}_j}{\pi_j},$$

donde $\widehat{A}_i = \widehat{p}_{A.HT} + \widehat{\beta}_{HT}(B_i - \widehat{p}_{B.HT})$.

Si el diseño muestral tiene tamaño muestral fijo, una expresión alternativa para la varianza anterior es

$$\widehat{V}(\widehat{p}_{r.opt.HT}) = -\frac{1}{2N^2} \sum_{i \neq j}^{n} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{A_i - \widehat{A}_i}{\pi_i} - \frac{A_j - \widehat{A}_j}{\pi_j} \right)^2.$$

Capítulo 3

Estimadores de tipo regresión para una proporción

3.1. Muestreo aleatorio simple

3.1.1. Definición del estimador

En el Capítulo 2 se presentó el estimador de expansión simple para la proporción poblacional P_A, y se definieron los estimadores de razón $\widehat{p}_r$ y $\widehat{p}_{r.q}$. Por último, se definió el estimador de razón óptimo que se obtuvo mediante la combinación lineal de los dos estimadores de tipo razón anteriormente mencionados.

Como se comentó en el Capítulo 1, también se pueden plantear estimadores de tipo regresión para la estimación de una media o total poblacional asociada a una variable cuantitativa. En concreto, el estimador de tipo regresión para la media poblacional $\overline{Y} = N^{-1}\sum_{i=1}^{N} y_i$ viene dado por

$$\overline{y}_{reg} = \overline{y} + \widehat{\beta}(\overline{X} - \overline{x}),$$

donde $\overline{X} = N^{-1}\sum_{i=1}^{N} x_i$ es la media poblacional de la variable auxiliar x, $\overline{y}$ y $\overline{x}$ son los estimadores de expansión simple para $\overline{Y}$ y $\overline{X}$ respectivamente, y

$$\widehat{\beta} = \frac{\widehat{cov}(x,y)}{\widehat{V}(x)}$$

es el coeficiente de regresión estimado basado en el modelo lineal simple.

En esta sección se define, para el caso de variables dicotómicas, el estimador de tipo regresión para la estimación de P_A, y se estudian las propiedades teóricas más importantes.

El estimador de tipo regresión para P_A que se presenta viene dado por

$$\widehat{p}_{reg} = \widehat{p}_A + b(P_B - \widehat{p}_B),$$

donde b es una constante. Como en el caso de los estimadores de tipo razón, se asume que la proporción poblacional de individuos que poseen el atributo B, P_B, es conocida a partir de un censo o ha sido estimada sin error. Esta suposición es la habitual en el muestreo de poblaciones finitas. Además se ha de reseñar que $\widehat{p}_{reg}$ proviene de una clase de estimadores generales que pueden ofrecer otros estimadores. Por ejemplo, el estimador de tipo diferencia se obtiene cuando $b = 1$.

3.1.2. Propiedades teóricas

En esta sección se obtienen las propiedades teóricas más importantes del estimador de regresión propuesto, las cuales nos permitirán la obtención del estimador óptimo dentro de la clase de estimadores dada por el estimador $\widehat{p}_{reg}$ en el sentido de mínima varianza.

Proposición 3.1 *El estimador de regresión propuesto $\widehat{p}_{reg}$ es un estimador insesgado.*

□

Demostración.

Dado que ambos estimadores $\widehat{p}_A$ y $\widehat{p}_B$ son estimadores insesgados de P_A y P_B respectivamente, tenemos que

$$E(\widehat{p}_{reg}) = E(\widehat{p}_A + b(P_B - \widehat{p}_B)) = P_A + bP_B - bP_B = P_A.$$

□

Es conocido que las expresiones para la varianza de un estimador son un problema importante para el cálculo de intervalos de confianza entre otras muchas aplicaciones. En nuestro caso, la varianza del estimador de regresión propuesto se obtendrá en el Teorema 3.1.

Teorema 3.1 *La varianza asintótica del estimador de regresión propuesto* $\widehat{p}_{reg}$ *viene dada por*

$$V(\widehat{p}_{reg}) = \frac{N-n}{N-1}\frac{1}{n}\left(P_AQ_A + b^2P_BQ_B - 2b\phi\sqrt{P_AQ_AP_BQ_B}\right).$$

□

Demostración.

En primer lugar, es fácil comprobar que

$$V(\widehat{p}_{reg}) = V(\widehat{p}_A) + b^2V(\widehat{p}_B) - 2bcov(\widehat{p}_A, \widehat{p}_B), \tag{3.1}$$

donde (véase Capítulo 2)

$$V(\widehat{p}_A) = \frac{N-n}{N-1}\frac{1}{n}P_AQ_A,$$

$$V(\widehat{p}_B) = \frac{N-n}{N-1}\frac{1}{n}P_BQ_B$$

y

$$cov(\widehat{p}_A, \widehat{p}_B) = \frac{N-n}{N-1}\frac{1}{n}\phi\sqrt{P_AQ_AP_BQ_B}.$$

Sustituyendo las expresiones anteriores en (3.1) se obtiene la expresión para $V(\widehat{p}_{reg})$ dada por

$$V(\widehat{p}_{reg}) = \frac{N-n}{(N-1)n}\left(P_AQ_A + b^2P_BQ_B - 2b\phi\sqrt{P_AQ_AP_BQ_B}\right).$$

□

Observamos que $V(\widehat{p}_{reg})$ depende de parámetros poblacionales desconocidos y por lo tanto no se puede aplicar en la práctica. Por ello en el Corolario 3.1 se obtiene un estimador insesgado para $V(\widehat{p}_{reg})$.

Corolario 3.1 *Un estimador insesgado de* $V(\widehat{p}_{reg})$ *viene dado por*

$$\widehat{V}(\widehat{p}_{reg}) = \frac{1-f}{n-1}\left(\widehat{p}_A\widehat{q}_A + b^2\widehat{p}_B\widehat{q}_B - 2b\widehat{\phi}\sqrt{\widehat{p}_A\widehat{q}_A\widehat{p}_B\widehat{q}_B}\right).$$

□

Teorema 3.2 *El valor óptimo de b que minimiza $V(\widehat{p}_{reg})$ viene dado por*

$$b_{opt} = \frac{\phi\sqrt{P_A Q_A}}{\sqrt{P_B Q_B}}.$$

□

Demostración.

Para hallar el valor óptimo es necesario encontrar el valor de b que minimice la función

$$g(b) = P_A Q_A + b^2 P_B Q_B - 2b\phi\sqrt{P_A Q_A P_B Q_B}.$$

La primera derivada de $g(b)$ viene dada por

$$g'(b) = 2bP_B Q_B - 2\phi\sqrt{P_A Q_A P_B Q_B},$$

lo cual implica, después de igualar a 0 dicha expresión y despejar, que

$$b = \frac{\phi\sqrt{P_A Q_A}}{\sqrt{P_B Q_B}}. \tag{3.2}$$

Partiendo de que $g''(b) = 2P_B Q_B > 0$, la solución (3.2) es un mínimo.

□

Por tanto, a partir del Teorema 3.2 podemos definir el estimador óptimo

$$\widehat{p}_{reg}^{OPT} = \widehat{p}_A + b_{opt}(P_B - \widehat{p}_B).$$

Corolario 3.2 *El estimador óptimo $\widehat{p}_{reg}^{OPT}$ tiene mínima varianza dada por*

$$V(\widehat{p}_{reg}^{OPT}) = V(\widehat{p}_A)(1 - \phi^2),$$

donde

$$V(\widehat{p}_A) = \frac{N-n}{(N-1)n} P_A Q_A.$$

□

Demostración.

$$V(\widehat{p}_{reg}^{OPT}) = \frac{N-n}{(N-1)n}\left(P_AQ_A + b_{opt}^2 P_BQ_B - 2b_{opt}\phi\sqrt{P_AQ_AP_BQ_B}\right) =$$

$$= \frac{N-n}{(N-1)n}\left(P_AQ_A + \frac{\phi^2 P_AQ_A}{P_BQ_B}P_BQ_B - 2\frac{\phi\sqrt{P_AQ_A}}{\sqrt{P_BQ_B}}\phi\sqrt{P_AQ_AP_BQ_B}\right) =$$

$$= \frac{N-n}{(N-1)n}P_AQ_A(1+\phi^2-2\phi^2) = V(\widehat{p}_A)(1-\phi^2).$$

□

El Corolario 3.2 indica que $\widehat{p}_{reg}^{OPT}$ es más eficiente que el estimador estándar $\widehat{p}_A$, dado que $0 \leq \phi^2 \leq 1$. El estimador $\widehat{p}_{reg}^{OPT}$ será tan eficiente como $\widehat{p}_A$ cuando $\phi = 0$, la cual es una situación donde el uso de información auxiliar no sería recomendable.

El estimador $\widehat{p}_{reg}^{OPT}$ depende de parámetros desconocidos y en la práctica no siempre puede calcularse. Por esta razón proponemos el uso del siguiente estimador que no depende de parámetros desconocidos

$$\widehat{p}_{reg}^{opt} = \widehat{p}_A + \widehat{b}_{opt}(P_B - \widehat{p}_B).$$

donde

$$\widehat{b}_{opt} = \frac{\widehat{\phi}\sqrt{\widehat{p}_A\widehat{q}_A}}{\sqrt{\widehat{p}_B\widehat{q}_B}}.$$

Observamos que el estimador de tipo regresión óptimo $\widehat{p}_{reg}^{opt}$ coincide con el estimador de tipo razón óptimo $\widehat{p}_{r.opt}$ propuesto en el Capitulo 2. Es destacable, en el contexto del muestreo en poblaciones finitas, el hecho de que a partir de una combinación lineal de dos estimadores de tipo razón resulte un estimador de tipo regresión.

Teorema 3.3 *El estimador $\widehat{p}_{reg}^{opt}$, el cual considera $\widehat{b}_{opt}$, tiene la misma distribución asintótica que $\widehat{p}_{reg}^{OPT}$, el cual considera el valor óptimo b_{opt}.*

□

Demostración.

Randles (1982) obtuvo la distribución límite de estadísticos con parámetros estimados. Usando esa notación, se denotará el estimador de tipo regresión $\widehat{p}_{reg}^{opt}$ como $T_n(\widehat{\lambda})$ con $\widehat{\lambda} = \widehat{b}_{opt}$. Sustituyendo el estimador $\widehat{\lambda}$ en $T_n(\cdot)$ con una variable matemática γ nos queda

$$T_n(\gamma) = \widehat{p}_A + \gamma(P_B - \widehat{p}_B)$$

la cual es diferenciable con respecto a γ en $\gamma = \lambda$. Calculado el límite de la esperanza en λ del estadístico $T_n(\gamma)$,

$$\mu(\gamma) = \lim_{n\to\infty} E_\lambda\{T_n(\gamma)\} = P_A.$$

Dado que $\frac{\partial\mu}{\partial\gamma}|_{\gamma=\lambda} = 0$ a partir de Randles (1982), se obtiene que $T_n(\widehat{\lambda})$ $(=\widehat{p}_{reg}^{opt})$ y $T_n(\lambda)$ $(=\widehat{p}_{reg}^{OPT})$ tienen la misma distribución límite y sus varianzas asintóticas coinciden.

□

Corolario 3.3 *Un estimador insesgado para* $V(\widehat{p}_{reg}^{opt})$ *viene dado por*

$$\widehat{V}(\widehat{p}_{reg}^{opt}) = \widehat{V}(\widehat{p}_A)(1 - \widehat{\phi}^2),$$

donde

$$\widehat{V}(\widehat{p}_A) = \frac{1-f}{n-1}\widehat{p}_A\widehat{q}_A.$$

□

Observamos que $\widehat{V}(\widehat{p}_{reg}^{opt})$ no depende de parámetros desconocidos, lo cual implica que se puede usar para la construcción de intervalos de confianza entre otras aplicaciones.

Partiendo de que $P_A = 1 - Q_A$, una cuestión que puede surgir en la práctica sería la obtención del estimador de la proporción a partir de su complementario. Por ejemplo, el estimador habitual de P_A es $\widehat{p}_A$, pero podría también considerarse la idea de la estimación de P_A mediante el estimador $\widehat{p}_{A.q} = 1 - \widehat{q}_A$, donde $\widehat{q}_A = n^{-1}\sum_{i\in s} A_i^c$ es el estimador estándar para Q_A. Se puede apreciar fácilmente que ambos estimadores $\widehat{p}_A$ y $\widehat{p}_{A.q}$ ofrecen la misma estimación. Esta propiedad también implica que $\widehat{p}_A$ tiene el mismo comportamiento en la estimación de P_A que el comportamiento que muestra $\widehat{q}_A$ en la estimación de Q_A. La Proposición 3.2 establece que esta propiedad también se mantiene para el estimador propuesto $\widehat{p}_{reg}^{OPT}$, y en consecuencia para el estimador $\widehat{p}_{reg}^{opt}$.

Proposición 3.2 *Los estimadores* $\widehat{p}_{reg}^{OPT}$ *y* $\widehat{p}_{reg.q}^{OPT} = 1 - \widehat{q}_{reg}^{OPT}$ *coinciden, donde* $\widehat{q}_{reg}^{OPT} = \widehat{q}_A + b_{opt.q}(Q_B - \widehat{q}_B)$ *es el estimador de tipo tipo regresión para* Q_A *y* $\widehat{b}_{opt.q} = \widehat{b}_{opt}$.

□

Demostración.

$$\widehat{p}_{reg}^{OPT} = \widehat{p}_A + b_{opt}(P_B - \widehat{p}_B) = 1 - \widehat{q}_A + b_{opt}\left(1 - Q_B - (1 - \widehat{q}_B)\right) =$$

$$= 1 - \widehat{q}_A - b_{opt}(Q_B - \widehat{q}_B) = 1 - \widehat{q}_{reg}^{OPT}.$$

Queda por demostrar que $b_{opt} = b_{opt.q}$. Considerando que $\widehat{q}_{reg} = \widehat{q}_A + b(Q_B - \widehat{q}_B)$ y siguiendo el Teorema 3.1, está claro que

$$V(\widehat{q}_{reg}) = \frac{N-n}{(N-1)n}\left(P_A Q_A + b^2 P_B Q_B - 2b\phi\sqrt{P_A Q_A P_B Q_B}\right) = V(\widehat{p}_{reg}),$$

lo cual implica que $b_{opt} = b_{opt.q}$ siguiendo el Teorema 3.2.

□

3.1.3. Definición del estimador de diferencia

A partir de la clase de estimadores dada por $\widehat{p}_{reg} = \widehat{p}_A + b(P_B - \widehat{p}_B)$ se pueden definir otros estimadores. En el caso de variables cuantitativas es usual definir estimadores de tipo diferencia. A continuación definiremos, para el problema de la estimación de una proporción, el estimador de tipo diferencia, el cual se obtiene considerando $b = 1$, es decir, el estimador de tipo diferencia para P_A viene dado por

$$\widehat{p}_d = \widehat{p}_A + (P_B - \widehat{p}_B).$$

A partir de las expresiones de las varianzas obtenidas para el estimador de tipo regresión, la varianza del estimador de tipo diferencia viene dada por

$$V(\widehat{p}_d) = \frac{N-n}{(N-1)n}\left(P_A Q_A + P_B Q_B - 2\phi\sqrt{P_A Q_A P_B Q_B}\right).$$

Dado que $V(\widehat{p}_d)$ depende de cantidades usualmente desconocidas, en la práctica usaremos su correspondiente estimador insesgado de la varianza, el

cual viene dado por

$$\widehat{V}(\widehat{p}_d) = \frac{1-f}{n-1}\left(\widehat{p}_A\widehat{q}_A + \widehat{p}_B\widehat{q}_B - 2\widehat{\phi}\sqrt{\widehat{p}_A\widehat{q}_A\widehat{p}_B\widehat{q}_B}\right).$$

3.1.4. Comparación teórica de estimadores

En esta sección se realizan comparaciones teóricas entre los distintos estimadores de tipo razón, regresión, diferencia y expansión simple discutidos en este trabajo. Estas comparaciones se realizarán en términos de varianzas, lo que nos permitirá conocer las condiciones en las que unos estimadores serán más eficientes que otros. Destacamos que los estimadores de tipo regresión, diferencia y expansión simple son insesgados, por lo que la eficiencia se mide a partir de las propias varianzas. Sin embargo, los estimadores de tipo razón, como comentamos en el Capitulo 2, no son insesgados, y de ahí que la eficiencia se mida en términos del error cuadrático medio. No obstante, este sesgo es despreciable para tamaños muestrales elevados, y las comparaciones que se realizarán a continuación en términos de varianzas son por tanto válidas bajo tales circunstancias.

Regresión versus expansión simple

Teorema 3.4 *El estimador de tipo regresión $\widehat{p}_{reg}^{opt}$ es más eficiente que el estimador de expansión simple $\widehat{p}_A$, es decir,*

$$V(\widehat{p}_{reg}^{opt}) < V(\widehat{p}_A).$$

y serán igual de eficientes cuando $\phi^2 = 0$,

□

Demostración.

Tal como se adelantaba en este capítulo,

$$V(\widehat{p}_{reg}^{opt}) = V(\widehat{p}_A)(1-\phi^2),$$

y puesto que $0 \leq \phi^2 \leq 1$, es evidente que $V(\widehat{p}_{reg}^{opt})$ será menor que $V(\widehat{p}_A)$, y tendrán el mismo valor si $\phi^2 = 0$.

□

Regresión versus diferencia

Teorema 3.5 *El estimador de tipo regresión $\widehat{p}_{reg}^{opt}$ es más eficiente que el estimador de tipo diferencia $\widehat{p}_d$, es decir,*

$$V(\widehat{p}_{reg}^{opt}) < V(\widehat{p}_d).$$

□

Demostración.

La varianzas de los estimadores de tipo diferencia y regresión pueden expresarse como

$$V(\widehat{p}_d) = V(\widehat{p}_A + P_B - \widehat{p}_B) = V(\widehat{p}_A) + V(\widehat{p}_B) - 2cov(\widehat{p}_A, \widehat{p}_B)$$

y

$$V(\widehat{p}_{reg}^{opt}) = V(\widehat{p}_A)(1-\phi^2) = V(\widehat{p}_A) - \phi^2 V(\widehat{p}_A).$$

Buscamos comprobar si $V(\widehat{p}_d) - V(\widehat{p}_{reg}^{opt}) > 0$. En efecto,

$$V(\widehat{p}_d) - V(\widehat{p}_{reg}) = V(\widehat{p}_B) - 2cov(\widehat{p}_A, \widehat{p}_B) + \phi^2 V(\widehat{p}_A). \tag{3.3}$$

Puesto que

$$V(\widehat{p}_A) = \frac{N-n}{N-1}\frac{1}{n}P_A Q_A,$$

$$V(\widehat{p}_B) = \frac{N-n}{N-1}\frac{1}{n}P_B Q_B$$

y

$$cov(\widehat{p}_A, \widehat{p}_B) = \frac{N-n}{N-1}\frac{1}{n}\phi\sqrt{P_A Q_A P_B Q_B},$$

la expresión (3.3) puede expresarse como

$$V(\widehat{p}_d) - V(\widehat{p}_{reg}^{opt}) = \frac{N-n}{N-1}\frac{1}{n}\left(P_B Q_B - 2\phi\sqrt{P_A Q_A P_B Q_B} + \phi^2 P_A Q_A\right)$$

$$= \frac{N-n}{N-1}\frac{1}{n}\left(\phi\sqrt{P_A Q_A} - \sqrt{P_B Q_B}\right)^2 \geq 0.$$

A partir de la ecuación anterior es fácil comprobar que la igualdad a 0 se dará cuando

$$\phi = \frac{\sqrt{P_B Q_B}}{\sqrt{P_A Q_A}},$$

es decir,

$$\phi\frac{\sqrt{P_AQ_A}}{\sqrt{P_BQ_B}} = b_{opt} = 1,$$

en cuyo caso el estimador de tipo regresión es el propio estimador de tipo diferencia. La igualdad a 0 también se presentará cuando $n = N$, pero n es menor estricto que N en el contexto que estamos trabajando en este trabajo.

□

Alternativamente, por ser el estimador de diferencia un caso particular con $b = 1$, resulta evidente que el estimador de tipo regresión óptimo es el óptimo en la clase en el sentido de mínima varianza, y por tanto, por definición el estimador de diferencia será menos eficiente que el estimador de tipo regresión óptimo.

Diferencia versus expansión simple

Teorema 3.6 *El estimador de tipo diferencia $\widehat{p}_d$ es más eficiente que el estimador de expansión simple $\widehat{p}_A$, es decir,*

$$V(\widehat{p}_d) < V(\widehat{p}_A),$$

cuando

$$\phi > \frac{1}{2}\frac{\sqrt{V(B)}}{\sqrt{V(A)}}. \tag{3.4}$$

□

Demostración.

En primer lugar recordamos que las expresiones de las varianzas de los estimadores de diferencia y expansión simple para la proporción poblacional P_A son

$$V(\widehat{p}_d) = \frac{N-n}{N-1}\frac{1}{n}(P_AQ_A + P_BQ_B - 2\phi\sqrt{P_AQ_AP_BQ_B})$$

y

$$V(\widehat{p}_A) = \frac{N-n}{N-1}\frac{1}{n}P_AQ_A.$$

El objetivo es demostrar cuando $V(\widehat{p}_d) < V(\widehat{p}_A)$, es decir,

$$\frac{N-n}{N-1}\frac{1}{n}(P_AQ_A + P_BQ_B - 2\phi\sqrt{P_AQ_AP_BQ_B}) < \frac{N-n}{N-1}\frac{1}{n}P_AQ_A.$$

Simplificando en ambos miembros de la desigualdad anterior nos queda

$$P_A Q_A + P_B Q_B - 2\phi\sqrt{P_A Q_A P_B Q_B} < P_A Q_A,$$

$$P_B Q_B - 2\phi\sqrt{P_A Q_A P_B Q_B} < 0$$

$$\sqrt{P_B Q_B} < 2\phi\sqrt{P_A Q_A},$$

$$\phi > \frac{1}{2}\frac{\sqrt{P_B Q_B}}{\sqrt{P_A Q_A}}$$

Dado que $V(A) = P_A Q_A$ y $V(B) = P_B Q_B$, según la Sección 2.2.1, se concluye que $V(\widehat{p}_d) < V(\widehat{p}_A)$ cuando

$$\phi > \frac{1}{2}\frac{\sqrt{V(B)}}{\sqrt{V(A)}}.$$

□

A partir de la expresión (3.4) pueden derivarse varias conclusiones. En primer lugar, observamos que

$$\frac{1}{2}\frac{\sqrt{V(B)}}{\sqrt{V(A)}} \geq 0.$$

Esto implica que la relación entre los atributos A y B deberá ser positiva para que el estimador de tipo diferencia $\widehat{p}_d$ sea más eficiente que el estimador estándar $\widehat{p}_A$. En segundo lugar, observamos de (3.4) que si ambos atributos tienen varianzas parecidas, entonces ϕ tendrá que ser mayor que 0.5 para concluir que el estimador de tipo diferencia sea más eficiente que el estimador de expansión simple. Por último, que la varianza del atributo B sea inferior a la varianza del atributo A favorece que la cota inferior para ϕ sea menor que la cota anterior fijada en 0.5.

Siguiendo la comparación teórica del estimador de tipo razón con el estimador de expansión simple (véase la Sección 2.2.3), a continuación se comparan las varianzas teóricas de los estimadores de tipo diferencia y expansión simple

en función de ϕ. Para ello, se generaron distintas tablas de doble entrada de la forma

	B	B^c	
A	P_{11}	P_{12}	P_A
A^c	P_{21}	P_{22}	Q_B
	P_B	Q_B	1

con valores de P_{11}, P_A y P_B entre 0.01 y 0.99, mientras que el resto de parámetros poblacionales se calcularon a partir de los anteriores. Con este conjunto de tablas, se realizaron estudios de comparación bajo diferentes escenarios: valores pequeños y elevados de ϕ, valores pequeños y elevados de P_A, P_B, P_{11}, etc.

La comparación teórica entre los estimadores $\widehat{p}_d$ y $\widehat{p}_A$ se realizó en términos de eficiencia relativa (ER), donde

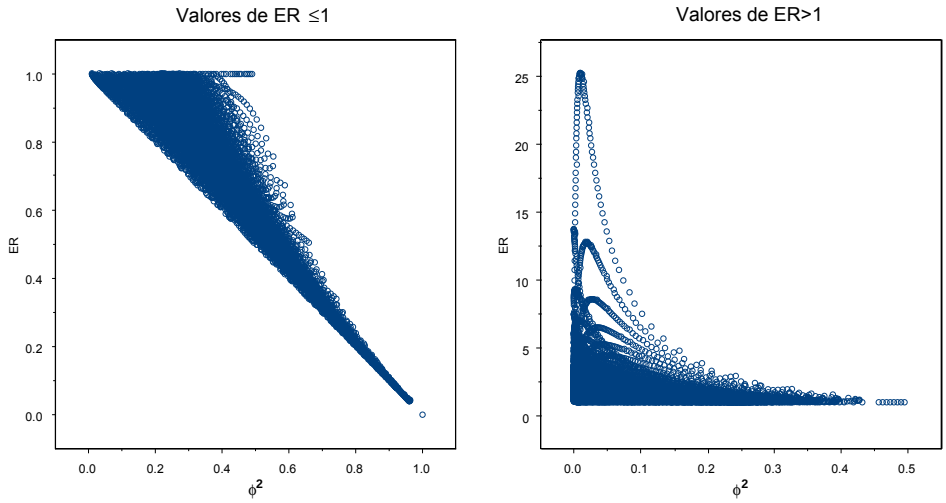
$$ER = \frac{V(\widehat{p}_d)}{V(\widehat{p}_A)}.$$

Podemos observar que esta medida no depende ni del tamaño muestral n ni del tamaño poblacional N, puesto que al simplificar ER nos queda

$$ER = \frac{P_A Q_A + P_B Q_B - 2\phi\sqrt{P_A Q_A P_B Q_B}}{P_A Q_A}.$$

Los resultados obtenidos pueden consultarse en la Figura 3.1. Observamos que para cualquier valor de ϕ^2 el estimador de diferencia puede ser más eficiente que el estimador de expansión simple, y siempre que $\phi^2 > 0.5$, el estimador de diferencia será más eficiente que el estimador de expansión simple. Por tanto, la ganancia en eficiencia de $\widehat{p}_d$ respecto $\widehat{p}_A$ puede ser importante o bien muy similar, por ejemplo cuando ϕ^2 =0.5. En otras palabras, en algunas situaciones el estimador de diferencia y el de expansión simple tienen el mismo comportamiento (ER = 1), mientras que en otras el estimador de diferencia puede ser el doble de eficiente que el estimador de expansión simple (ER = 0.5). Cuando ϕ^2 está próximo a 0, el estimador tipo diferencia puede ser considerablemente peor que $\widehat{p}_A$. Sin embargo, cuando ϕ^2 está próximo a 1, el estimador de tipo diferencia es mucho más eficiente que $\widehat{p}_A$. Finalmente, si comparamos las Figuras 2.1 y 3.1, podemos observar que con mucha frecuencia el estimador de razón es menos eficiente que el estimador de expansión simple (véase Figura 2.1) en comparación con el número de veces que el estimador de tipo diferencia es menos eficiente que $\widehat{p}_A$ (véase Figura 3.1). Además, podemos destacar que la eficiencia relativa en el caso de la comparación entre el estimador de diferencia y expansión simple están acotados superiormente por 25, mientras que en

Figura 3.1: Comparación teórica del estimador de tipo diferencia $\widehat{p}_d$ con el estimador de expansión simple $\widehat{p}_A$ mediante la eficiencia relativa (ER) entre ambos estimadores.



la Figura 2.1 se aprecia que son numerosos los casos en los que la eficiencia relativa está muy por encima de 25.

Regresión versus razón

Teorema 3.7 *El estimador de tipo regresión $\widehat{p}_{reg}^{opt}$ es más eficiente que el estimador de tipo razón $\widehat{p}_r$, es decir,*

$$V(\widehat{p}_{reg}^{opt}) < V(\widehat{p}_r),$$

y serán igual de eficientes cuando $P_{12} = 0$, donde P_{12} es la proporción poblacional de individuos que presentan simultáneamente los atributos A y B^c.

□

Demostración.

Las varianzas de los estimadores de tipo regresión y razón para la proporción poblacional P_A son:

$$V(\widehat{p}_{reg}^{opt}) = \frac{N-n}{N-1}\frac{1}{n}\left(P_A Q_A - \phi^2 P_A Q_A\right)$$

y

$$V(\widehat{p}_r) = \frac{N-n}{N-1}\frac{1}{n}\left(P_AQ_A + R^2P_BQ_B - 2R\phi\sqrt{P_AQ_AP_BQ_B}\right)$$

Tenemos que comprobar si $V(\widehat{p}^{opt}_{reg}) \leq V(\widehat{p}_r)$, es decir,

$$\frac{N-n}{N-1}\frac{1}{n}\left(P_AQ_A - \phi^2P_AQ_A\right) \leq \frac{N-n}{N-1}\frac{1}{n}\left(P_AQ_A + R^2P_BQ_B - 2R\phi\sqrt{P_AQ_AP_BQ_B}\right)$$

$$P_AQ_A - \phi^2P_AQ_A \leq P_AQ_A + R^2P_BQ_B - 2R\phi\sqrt{P_AQ_AP_BQ_B}$$

$$\phi^2P_AQ_A + R^2P_BQ_B - 2R\phi\sqrt{P_AQ_AP_BQ_B} \geq 0. \tag{3.5}$$

Puesto que $R = P_A/P_B$ y

$$\phi = \frac{P_{11}P_{22} - P_{12}P_{21}}{\sqrt{P_AQ_AP_BQ_B}},$$

donde los parámetros poblaciones anteriores vienen explicados en la tabla de doble entrada

	B	B^c	
A	P_{11}	P_{12}	P_A
A^c	P_{21}	P_{22}	Q_A
	P_B	Q_B	1

la desigualdad (3.5) puede expresarse como

$$\frac{(P_{11}P_{22} - P_{12}P_{21})^2}{P_AQ_AP_BQ_B}P_AQ_A + \frac{P_A^2}{P_B^2}P_BQ_B - 2\frac{P_A}{P_B}\frac{P_{11}P_{22} - P_{12}P_{21}}{\sqrt{P_AQ_AP_BQ_B}}\sqrt{P_AQ_AP_BQ_B} \geq 0,$$

donde simplificando y multiplicando por P_BQ_B nos queda

$$(P_{11}P_{22} - P_{12}P_{21})^2 + P_A^2Q_B^2 - 2P_AQ_B(P_{11}P_{22} - P_{12}P_{21}) \geq 0;$$

$$[(P_{11}P_{22} - P_{12}P_{21}) - P_AQ_B]^2 \geq 0,$$

y la condición anterior se cumple al tratarse del cuadrado de una diferencia.

Nos queda comprobar cuando los dos estimadores son igual de eficientes, es decir, cuando

$$P_{11}P_{22} - P_{12}P_{21} - P_AP_B = 0.$$

Puesto que $P_{11} = (P_A - P_{12})$ y $P_{22} = (Q_B - P_{12})$, la ecuación anterior nos queda como

$$(P_A - P_{12})(Q_B - P_{12}) - P_{12}P_{21} - P_A P_B = 0;$$
$$P_A P_B - P_A P_{12} - Q_B P_{12} + P_{12}^2 - P_{12}P_{21} - P_A P_B = 0;$$
$$P_{12}(P_{12} - P_A - Q_B - P_{21}) = 0.$$

Dado que

$$P_{12} - P_A - Q_B - P_{21} = -P_{11} - (1 - P_B) - P_{21} = -P_{11} + P_B - P_{21} - 1$$
$$= -P_{11} + P_{11} - 1 = -1,$$

concluimos que los estimadores $\widehat{p}_{reg}^{opt}$ y $\widehat{p}_r$ serán igual de eficientes cuando $P_{12} = 0$.

□

Observamos que si $P_{12} = 0$, entonces $P_{11} = P_A$ y $P_B = P_A + P_{21}$, lo que implica que $P_A \leq P_B$ y se da la condición necesaria para que $\widehat{p}_r$ sea más eficiente que el estimador $\widehat{p}_{r.q}$, tal como vimos en el Capítulo 2. En otras palabras, si $P_A \leq P_B$ se tiene la siguiente relación entre las varianzas de los distintos estimadores comentados

$$V(\widehat{p}_{reg}^{opt}) \leq V(\widehat{p}_r) \leq V(\widehat{p}_{r.q}).$$

En este sentido, si $P_A > P_B$, el estimador $\widehat{p}_{r.q}$ es más eficiente que $\widehat{p}_r$, por lo que tendremos que estudiar si el estimador de tipo regresión es más eficiente que $\widehat{p}_{r.q}$ en esta situación. El siguiente teorema aclara esta cuestión.

Teorema 3.8 *El estimador de tipo regresión $\widehat{p}_{reg}^{opt}$ es más eficiente que el estimador de tipo razón $\widehat{p}_{r.q}$, es decir,*

$$V(\widehat{p}_{reg}^{opt}) < V(\widehat{p}_{r.q}),$$

y serán igual de eficientes cuando $P_{21} = 0$, donde P_{21} es la proporción poblacional de individuos que presentan simultáneamente los atributos A^c y B.

□

Demostración.

Las varianzas de los estimadores $\widehat{p}_{reg}^{opt}$ y $\widehat{p}_{r.q}$ para la proporción poblacional P_A son:

$$V(\widehat{p}_{reg}^{opt}) = \frac{N-n}{N-1}\frac{1}{n}\left(P_A Q_A - \phi^2 P_A Q_A\right)$$

y

$$V(\widehat{p}_{r.q}) = \frac{N-n}{N-1}\frac{1}{n}\left(P_A Q_A + R_c^2 P_B Q_B - 2R_c\phi\sqrt{P_A Q_A P_B Q_B}\right)$$

Tenemos que comprobar si $V(\widehat{p}_{reg}^{opt}) \leq V(\widehat{p}_{r.q})$, es decir,

$$\frac{N-n}{N-1}\frac{1}{n}\left(P_A Q_A - \phi^2 P_A Q_A\right) \leq \frac{N-n}{N-1}\frac{1}{n}\left(P_A Q_A + R_c^2 P_B Q_B - 2R_c\phi\sqrt{P_A Q_A P_B Q_B}\right)$$

$$P_A Q_A - \phi^2 P_A Q_A \leq P_A Q_A + R_c^2 P_B Q_B - 2R_c\phi\sqrt{P_A Q_A P_B Q_B}$$

$$\phi^2 P_A Q_A + R_c^2 P_B Q_B - 2R_c\phi\sqrt{P_A Q_A P_B Q_B} \geq 0. \tag{3.6}$$

Puesto que $R_c = Q_A/Q_B$ y

$$\phi = \frac{P_{11}P_{22} - P_{12}P_{21}}{\sqrt{P_A Q_A P_B Q_B}},$$

entonces la desigualdad (3.6) puede expresarse como

$$\frac{(P_{11}P_{22} - P_{12}P_{21})^2}{P_A Q_A P_B Q_B} P_A Q_A + \frac{Q_A^2}{Q_B^2} P_B Q_B - 2\frac{Q_A}{Q_B}\frac{P_{11}P_{22} - P_{12}P_{21}}{\sqrt{P_A Q_A P_B Q_B}}\sqrt{P_A Q_A P_B Q_B} \geq 0,$$

donde simplificando y multiplicando por $P_B Q_B$ nos queda

$$(P_{11}P_{22} - P_{12}P_{21})^2 + Q_A^2 P_B^2 - 2Q_A P_B(P_{11}P_{22} - P_{12}P_{21}) \geq 0;$$

$$[(P_{11}P_{22} - P_{12}P_{21}) - Q_A P_B]^2 \geq 0,$$

y la condición anterior se cumple al tratarse del cuadrado de una diferencia.

Nos queda comprobar cuando los dos estimadores son igual de eficientes, es decir, cuando

$$P_{11}P_{22} - P_{12}P_{21} - Q_A P_B \geq 0.$$

Siguiendo los pasos seguidos en la demostración anterior, podemos comprobar fácilmente que los estimadores $\widehat{p}_{reg}^{opt}$ y $\widehat{p}_{r.q}$ serán igual de eficientes cuando $P_{21} = 0$.

□

Al igual que comentamos anteriormente, observamos que si $P_{21} = 0$, entonces $P_{11} = P_B$ y $P_A = P_B + P_{12} \geq P_B$, es decir, se da la condición necesaria para que $\widehat{p}_{r.q}$ sea más eficiente que $\widehat{p}_r$. En otras palabras, si $P_A \geq P_B$ se tiene la siguiente relación

$$V(\widehat{p}_{reg}^{opt}) \leq V(\widehat{p}_{r.q}) \leq V(\widehat{p}_r).$$

Regresión versus razón óptimo

En el caso de MAS, puede observase que el estimador de razón óptimo $\widehat{p}_{r.opt}$ obtenido en el Capítulo 2 mediante la combinación lineal de los estimadores $\widehat{p}_r$ y $\widehat{p}_{r.q}$ y el estimador de regresión óptimo $\widehat{p}_{reg}^{opt}$ obtenido en este capítulo mediante el método de regresión presentan la misma expresión, y por tanto son igual de eficientes, como también lo pone de manifiesto sus correspondientes varianzas obtenidas en cada caso.

Diferencia versus razón

Teorema 3.9 *El estimador de tipo razón $\widehat{p}_r$ es más eficiente que el estimador de tipo diferencia $\widehat{p}_d$, es decir*

$$V(\widehat{p}_r) < V(\widehat{p}_d),$$

cuando $P_A < P_B$, y serán igual de eficientes cuando $P_A = P_B$.

□

Demostración.

Tenemos que comprobar cuando $V(\widehat{p}_r) \leq V(\widehat{p}_d)$, es decir,

$$V(\widehat{p}_A) + R^2 V(\widehat{p}_B) - 2Rcov(\widehat{p}_A, \widehat{p}_B) \leq V(\widehat{p}_A) + V(\widehat{p}_B) - 2cov(\widehat{p}_A, \widehat{p}_B);$$

$$V(\widehat{p}_B)(1 - R^2) - 2cov(\widehat{p}_A, \widehat{p}_B)(1 - R) \geq 0;$$

$$(1 - R)\left(V(\widehat{p}_B)(1 + R) - 2cov(\widehat{p}_A, \widehat{p}_B)\right) \geq 0;$$

$$(1 - R)\frac{N - n}{N - 1}\frac{1}{n}\left(P_B Q_B\left(1 + \frac{P_A}{P_B}\right) - 2\phi\sqrt{P_A Q_A P_B Q_B}\right) \geq 0. \qquad (3.7)$$

Teniendo en cuenta que $P_A = P_{11} + P_{12}$, $P_B = P_{11} + P_{21}$ y $Q_B = P_{12} + P_{22}$ obtenemos que

$$P_B Q_B\left(1 + \frac{P_A}{P_B}\right) - 2\phi\sqrt{P_A Q_A P_B Q_B} = P_B Q_B + P_A Q_B - 2(P_{11}P_{12} - P_{12}P_{21})$$

$$= P_{11}P_{12}+P_{11}P_{22}+P_{12}^2+P_{12}P_{22}+P_{11}P_{12}+P_{11}P_{22}+P_{12}P_{21}+P_{21}P_{22}-2P_{11}P_{22}+2P_{12}P_{21}$$

$$= 2P_{11}P_{12} + P_{12}^2 + P_{12}P_{22} + 3P_{12}P_{21} + P_{21}P_{22} \geq 0, \tag{3.8}$$

puesto que la expresión (3.8) está formada por la suma de proporciones mayores o iguales que 0, lo que implica que $V(\widehat{p}_r) \leq V(\widehat{p}_d)$ cuando $1 - R \geq 0$, o lo que es lo mismo $P_A \leq P_B$. A partir de la expresión (3.7) queda claro que $\widehat{p}_r$ y $\widehat{p}_d$ serán igual de eficientes cuando $P_A = P_B$. Otra posibilidad para que tales estimadores sean igual de eficientes es que la expresión dada en (3.8) sea igual a 0, aunque esta situación se dará únicamente cuando $P_{12} = P_{21} = 0$, en cuyo caso también se verifica que $P_A = P_B$.

□

En conclusión, si $P_A \leq P_B$ se verifica, por un lado, que $V(\widehat{p}_r) \leq V(\widehat{p}_d)$, y también se verifica, por otro lado, que $V(\widehat{p}_r) \leq V(\widehat{p}_{r.q})$. En otras palabras, cuando $P_A < P_B$ es aconsejable utilizar el estimador $\widehat{p}_r$ antes que los estimadores $\widehat{p}_d$ y $\widehat{p}_{r.q}$. Ante este escenario, es oportuno razonar bajo qué circunstancias el estimador de tipo diferencia es más eficiente que el estimador de tipo razón $\widehat{p}_{r.q}$. El siguiente teorema aclara esta situación.

Teorema 3.10 *El estimador de tipo razón $\widehat{p}_{r.q}$ es más eficiente que el estimador de tipo diferencia $\widehat{p}_d$, es decir*

$$V(\widehat{p}_{r.q}) < V(\widehat{p}_d),$$

cuando $P_A > P_B$, y serán igual de eficientes cuando $P_A = P_B$.

□

Demostración.

Tenemos que comprobar cuando

$$V(\widehat{p}_{r.q}) = V(1 - \widehat{q}_r) = V(\widehat{q}_r) \leq V(\widehat{p}_d),$$

esto es,

$$V(\widehat{p}_A) + R_c^2 V(\widehat{p}_B) - 2R_C cov(\widehat{p}_A, \widehat{p}_B) \leq V(\widehat{p}_A) + V(\widehat{p}_B) - 2cov(\widehat{p}_A, \widehat{p}_B)$$

$$V(\widehat{p}_B)(1 - R_c^2) - 2cov(\widehat{p}_A, \widehat{p}_B)(1 - R_c) \geq 0;$$
$$(1 - R_c)\left(V(\widehat{p}_B)(1 + R_c) - 2cov(\widehat{p}_A, \widehat{p}_B)\right) \geq 0;$$

$$(1-R_c)\frac{N-n}{N-1}\frac{1}{n}\left(P_BQ_B\left(1+\frac{Q_A}{Q_B}\right)-2\phi\sqrt{P_AQ_AP_BQ_B}\right)\geq 0. \qquad (3.9)$$

Teniendo en cuenta que $P_A = P_{11} + P_{12}$, $P_B = P_{11} + P_{21}$ y $Q_B = P_{12} + P_{22}$ obtenemos que

$$P_BQ_B\left(1+\frac{Q_A}{Q_B}\right)-2\phi\sqrt{P_AQ_AP_BQ_B} = P_BQ_B + P_BQ_A - 2(P_{11}P_{12} - P_{12}P_{21})$$

$$= P_{11}P_{21}+P_{11}P_{22}+P_{21}^2+P_{21}P_{22}+P_{11}P_{12}+P_{11}P_{22}+P_{21}P_{12}+P_{21}P_{22}-2P_{11}P_{22}+2P_{12}P_{21}$$

$$= P_{11}P_{21} + P_{21}^2 + 2P_{21}P_{22} + 3P_{21}P_{12} + P_{11}P_{12} \geq 0, \qquad (3.10)$$

puesto que la expresión (3.10) es suma de proporciones mayores o iguales que 0, lo que implica que $V(\widehat{p}_{r.q}) \leq V(\widehat{p}_d)$ cuando $1 - R_c \geq 0$, o lo que es lo mismo $Q_A \leq Q_B$ o bien $P_A \geq P_B$. A partir de la expresión (3.9) queda claro que $\widehat{p}_{r.q}$ y $\widehat{p}_d$ serán igual de eficientes cuando $P_A = P_B$. Otra posibilidad para que tales estimadores sean igual de eficientes es que la expresión dada en (3.10) sea igual a 0, aunque esta situación se dará únicamente cuando $P_{12} = P_{21} = 0$, en cuyo caso también se verifica que $P_A = P_B$.

□

En resumen, si $P_A \geq P_B$ se verifica que $V(\widehat{p}_{r.q}) \leq V(\widehat{p}_d)$ y $V(\widehat{p}_{r.q}) \leq V(\widehat{p}_r)$.

A partir de los Teoremas 3.9 y 3.10 podemos concluir que sería aconsejable utilizar el estimador de tipo razón $\widehat{p}_{r.q}$ cuando $P_A > P_B$, y el estimador de tipo razón $\widehat{p}_r$ cuando $P_A < P_B$. Cuando $P_A = P_B$, el estimador de tipo diferencia tendrá el mismo comportamiento en términos de varianzas que los estimadores de tipo razón $\widehat{p}_r$ y $\widehat{p}_{r.q}$.

3.1.5. Comparación empírica de estimadores

En esta sección, todos los estimadores propuestos en este trabajo se compararán numéricamente mediante estudios de simulación Monte Carlo con otros estimadores de la proporción poblacional bajo diferentes situaciones. Los estudios de simulación se basarán, por una parte, en poblaciones simuladas que abarcan distintos escenarios, mientras que la aplicabilidad de los estimadores propuestos podrá también observarse mediante el estudio del comportamiento empírico de tales estimadores en datos reales extraídos del ámbito de la Economía y la Empresa.

Poblaciones simuladas

Se han realizado estudios de simulación basados en 30 poblaciones simuladas, las cuales cubren un amplio rango de posibles escenarios: altas y bajas proporciones, valores pequeños y grandes para el coeficiente V de Cramer, etc. Una breve descripción de cómo se han generado las 30 poblaciones puede consultarse en el Apéndice A.

Para cada población simulada, se seleccionaron $D = 10000$ muestras para comparar los distintos estimadores en términos de sesgo relativo (SR) y eficiencia relativa (ER), donde

$$SR = \frac{E[\widehat{p}] - P_A}{P_A} \quad ; \quad ER = \frac{ECM[\widehat{p}]}{ECM[\widehat{p}_A]},$$

$\widehat{p}$ es un determinado estimador y la esperanza ($E[\cdot]$) y el error cuadrático medio ($ECM[\cdot]$) empíricos vienen dados por:

$$E[\widehat{p}] = \frac{1}{D}\sum_{i=1}^{D}\widehat{p}_i \quad ; \quad ECM[\widehat{p}] = \frac{1}{D}\sum_{i=1}^{D}(\widehat{p}_i - P_A)^2,$$

donde $\widehat{p}_i$ denota el estimador $\widehat{p}$ calculado a partir de simulación i-ésima. Valores de ER menores de 1 indican que el estimador $\widehat{p}$ es más eficiente que el estimador de expansión simple $\widehat{p}_A$, el cual se considera como el estimador de referencia en los estudios de la eficiencia relativa.

A partir del estudio de simulación observamos, en primer lugar, que los valores de SR están dentro de un rango razonable, es decir, todos los estimadores tienen valores de SR menores del 1 %, y están por tanto omitidos.

Como se comentó en secciones anteriores, los estimadores óptimos $\widehat{p}_{r.opt}$ y $\widehat{p}_{reg}^{opt}$ coinciden bajo MAS, y de ahí que bajo este diseño muestral ambos estimadores se identifiquen en las Figuras 3.2 y 3.3 mediante reg, haciendo referencia al estimador de tipo regresión.

La Figura 3.2 muestra la eficiencia relativa (ER) de los distintos estimadores utilizados en el estudio de simulación. Se observa que el estimador de tipo regresión óptimo es el más eficiente en todos los casos, al tener siempre los valores de ER más pequeños. El segundo estimador más eficiente, también en todos los casos, es el estimador $\widehat{p}_{r.e}$. Tal como estudiamos desde un punto de vista teórico, el estimador de tipo razón $\widehat{p}_r$ tiene un buen comportamiento cuando $P_A < P_B$, aunque presenta estimaciones poco eficientes, incluso peores que el estimador $\widehat{p}_A$, en la situación contraria. En concordancia con el estudio teórico desarrollado en secciones anteriores, el estimador de tipo diferencia

Figura 3.2: Valores de eficiencia relativa (ER) para varios estimadores de P_A obtenidos a partir de las poblaciones simuladas. Las muestras son seleccionadas con tamaño $n = 100$. P_A varía de 0.1 a 0.9 y ϕ varía de 0.5 a 0.9.

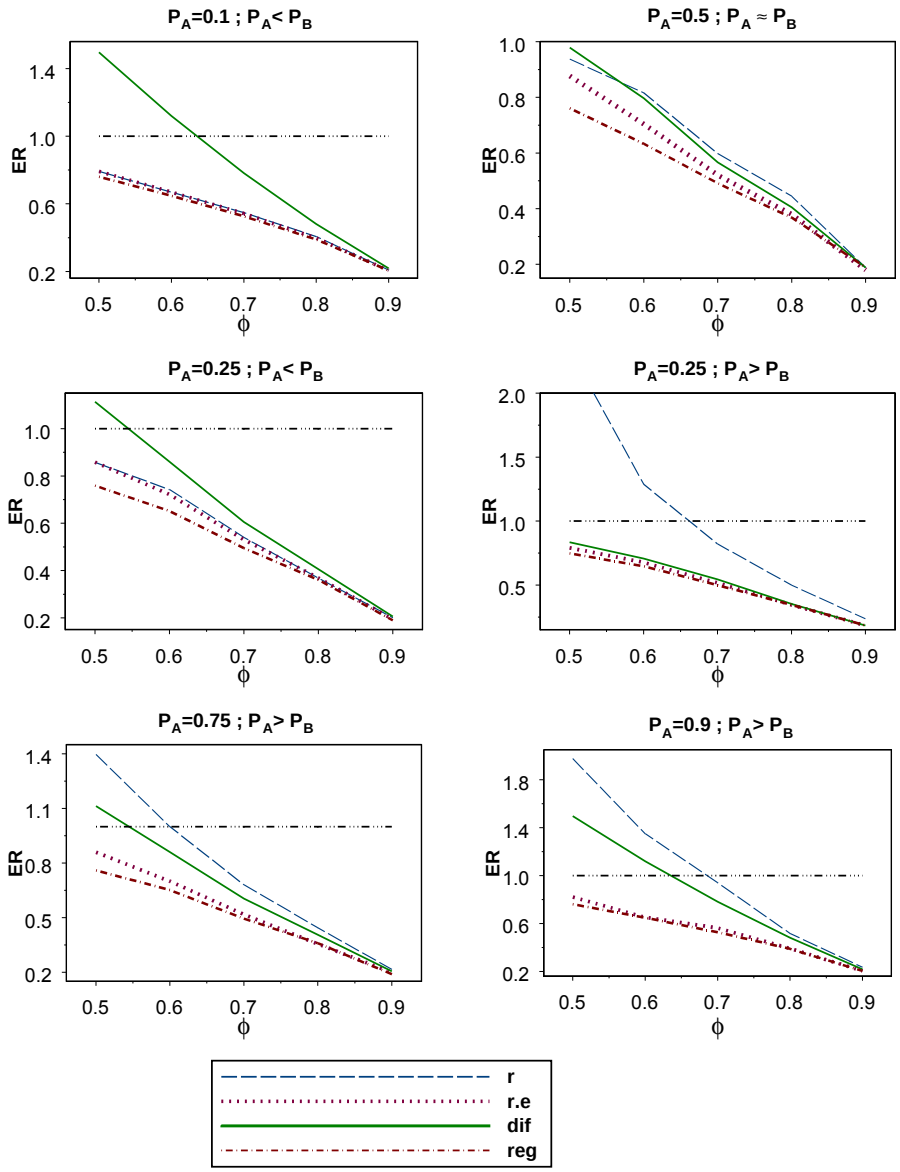
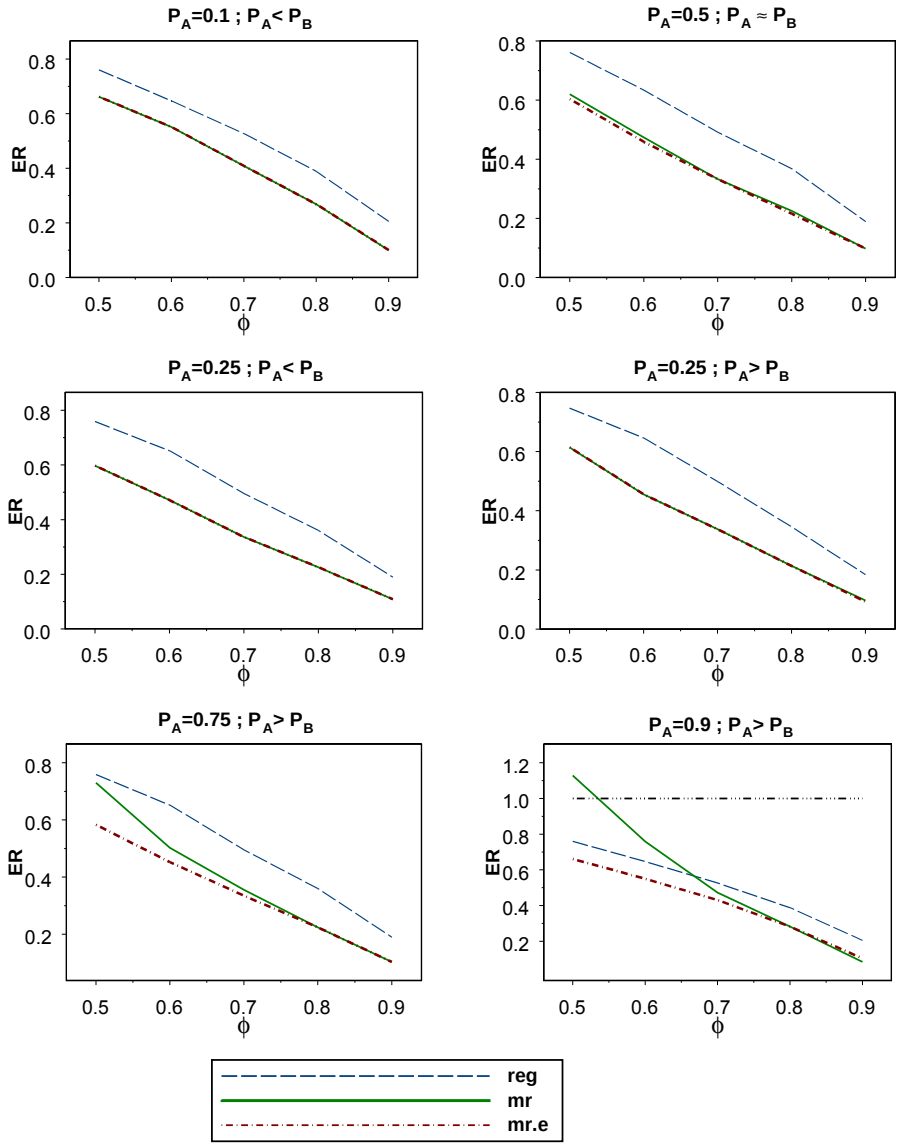


Figura 3.3: Valores de eficiencia relativa (ER) de los estimadores de tipo razón basados en varios atributos auxiliares y del estimador de tipo regresión óptimo obtenidos a partir de las poblaciones simuladas. Las muestras son seleccionadas con tamaño $n = 100$. P_A varía de 0.1 a 0.9 y ϕ varía de 0.5 a 0.9.



$\widehat{p}_d$ es más eficiente que el estimador de tipo razón $\widehat{p}_r$ cuando $P_A > P_B$, es menos eficiente en la situación contraria y ambos estimadores tienen un comportamiento similar cuando $P_A = P_B$. Finalmente y como parece razonable, destacamos que todos los estimadores basados en información auxiliar mejoran en eficiencia con respecto al estimador de expansión simple a medida que aumenta el valor de ϕ.

En la Figura 3.3 se compara la eficiencia del estimador de tipo regresión $\widehat{p}_{reg}^{opt}$ (o bien el $\widehat{p}_{r.opt}$), el cual tiene el mejor comportamiento teórica y empíricamente como hemos comprobado, con los estimadores de tipo razón basados en varios atributos auxiliares. El objetivo de este estudio es analizar empíricamente la ganancia o pérdida de eficiencia que se produce al utilizar un atributo auxiliar en lugar de varios. De la Figura 3.3 observamos que los estimadores de tipo razón $\widehat{p}_{mr}$ y $\widehat{p}_{mr.e}$ tienen una ganancia en eficiencia importante con respecto $\widehat{p}_{reg}^{opt}$, excepto en el caso $P_A = 0{,}9$ y $P_A > P_B$, donde el estimador $\widehat{p}_{mr}$ es menos eficiente que el estimador de tipo regresión cuando ϕ <0.7.

Poblaciones basadas en datos reales

Además del estudio de simulación anteriormente descrito basado en poblaciones simuladas y las cuales reflejan distintos escenarios que pueden presentarse en la práctica, a continuación se lleva a cabo un nuevo estudio de simulación basado en datos reales del ámbito de la Economía y la Empresa. Las poblaciones utilizadas en este estudio de simulación pueden consultarse en el Apéndice A, donde se describen en detalle cada una de ellas. Estas poblaciones están basadas en datos de la Encuesta de Presupuestos Familiares (población EPF), datos de la Encuesta Social Europea (población ESE), datos de lagos estadounidenses (población Lagos) y datos de la Encuesta Nacional de Salud (población ENS).

Siguiendo las medidas de comparación definidas al comienzo de esta Sección 3.1.5, la comparación empírica de los estimadores se realizó en términos de sesgo relativo (SR) y eficiencia relativa (ER). Además, en este estudio de simulación se incorpora como medida de comparación el error cuadrático medido relativo (ECMR), donde

$$ECMR = \frac{\sqrt{ECM[\widehat{p}]}}{P_A}.$$

Los resultados obtenidos en los estudios de simulación asociados a las poblaciones basadas en datos reales se muestran a continuación.

Tabla 3.1: Valores del sesgo relativo (SR), error cuadrático medio relativo (ECMR) y eficiencia relativa (ER) para distintos estimadores de P_A =0.194 en la población EPF. ϕ =0.501 y P_B =0.173.

n	Estimador	SR (%)	ECMR	ER
50	$\widehat{p}_A$	-0.2	28.9	1.00
	$\widehat{p}_r$	6.8	42.4	2.15
	$\widehat{p}_{r.e}$	0.3	26.0	0.81
	$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	0.0	25.8	0.80
	$\widehat{p}_d$	0.0	27.8	0.93
100	$\widehat{p}_A$	-0.4	19.1	1.00
	$\widehat{p}_r$	2.0	21.8	1.30
	$\widehat{p}_{r.e}$	-0.3	18.0	0.89
	$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	-0.5	16.9	0.78
	$\widehat{p}_d$	-0.5	19.1	1.00

En la Tabla 3.1 pueden observarse los valores de SR, ECMR y ER para los distintos estimadores de P_A discutidos en esta memoria. En la mencionada tabla podemos observar que todos los estimadores, a excepción de $\widehat{p}_r$, presentan sesgos razonables, con valores por debajo del 1 %. No obstante, observamos que el sesgo de $\widehat{p}_r$ decrece a medida que aumenta el tamaño de la muestra.

En lo que respecta a la eficiencia, observamos que el estimador de razón es el menos eficiente al tener un mayor valor de eficiencia relativa, mientras que el estimador más eficiente es el estimador de tipo regresión, si bien, es cierto que el estimador $\widehat{p}_{r.e}$ está muy cerca en términos de ER del estimador $\widehat{p}_{reg}^{opt}$. El estimador de tipo diferencia presenta un valor de ER cercano a 1, es decir, el estimador estándar es tan eficiente como el estimador $\widehat{p}_d$.

De la tabla 3.1 también podemos observar que a medida que aumentamos el tamaño de la muestra el estimador de razón se vuelve más eficiente y se acerca en términos de ER al estimador estándar.

En la Tabla 3.2 podemos observar que todos los estimadores tienen sesgos insignificantes con valores por debajo del 1 %. Por otro lado, en lo que respecta a la eficiencia, observamos que el estimador de diferencia es el menos eficiente al tener un mayor valor de eficiencia relativa, seguido del estimador estándar y el estimador de razón. Por su parte, el estimador más eficiente es el estimador de tipo regresión.

Tabla 3.2: Valores del sesgo relativo (SR), error cuadrático medio relativo (ECMR) y eficiencia relativa (ER) para distintos estimadores de P_A =0.496 en la población ESE. ϕ =0.467 y P_B =0.596.

n	Estimador	SR (%)	ECMR	ER
50	$\widehat{p}_A$	0.0	13.9	1.00
	$\widehat{p}_r$	0.6	13.5	0.94
	$\widehat{p}_{r.e}$	0.4	13.1	0.89
	$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	0.0	12.4	0.80
	$\widehat{p}_d$	0.0	14.2	1.04
100	$\widehat{p}_A$	-0.1	9.6	1.00
	$\widehat{p}_r$	0.1	9.2	0.92
	$\widehat{p}_{r.e}$	0.0	9.1	0.90
	$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	-0.2	8.5	0.78
	$\widehat{p}_d$	-0.2	9.8	1.04

En la Tabla 3.3 podemos observar que todos los estimadores presentan sesgos razonables, a excepción de $\widehat{p}_r$ que tiene valores por encima del 8.8 % en el caso más extremo. No obstante, observamos que el sesgo de $\widehat{p}_r$ decrece a medida que aumenta el tamaño de la muestra.

Respecto a la eficiencia de los estimadores, observamos que, tanto para $n = 50$ como para $n = 100$, todos los estimadores mejoran considerablemente con respecto al estimador estándar cuando $\phi = 0{,}9$. Sin embargo, cuando $\phi = 0{,}5$, el estimador de razón es mucho menos eficiente que el estimador estándar, aunque esta pérdida en eficiencia va disminuyendo a medida que aumenta el tamaño muestral. Como se comprobó teórica y numéricamente en secciones anteriores, este hecho se debe a que P_A es claramente mayor que P_B en esta situación. Los estimadores $\widehat{p}_{r.opt}$ y $\widehat{p}_{r.e}$ son de nuevo los más eficientes en todos los casos.

En la Tabla 3.4 todos los estimadores presentan sesgos insignificantes. La ganancia en eficiencia de los estimadores basados en información auxiliar es bastante importante cuando $\phi = 0{,}9$. En esta población, es el estimador de tipo diferencia el que es menos eficiente que el estimador tipo estándar, mientras que el estimador de tipo razón tiene un buen comportamiento cuando $\phi = 0{,}5$. Esta circunstancia se debe a que $P_A < P_B$.

En la simulación realizada para la población Lagos cuando $P_A = 0{,}07$ (Tabla 3.5) podemos observar que cuando $n = 50$ los estimadores $\widehat{p}_r$ y $\widehat{p}_{r.e}$

Tabla 3.3: Valores del sesgo relativo (SR), error cuadrático medio relativo (ECMR) y eficiencia relativa (ER) para distintos estimadores de P_A =0.438 en la población Lagos. P_B =0.44 cuando ϕ =0.9 y P_B =0.163 cuando ϕ =0.5.

n	ϕ	Estimador	SR (%)	ECMR	ER
50	0.9	$\widehat{p}_A$	-0.1	15.3	1.00
		$\widehat{p}_r$	0.3	7.3	0.23
		$\widehat{p}_{r.e}$	0.0	6.9	0.20
		$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	0.0	6.9	0.20
		$\widehat{p}_d$	0.0	7.0	0.21
	0.5	$\widehat{p}_A$	0.2	15.2	1.00
		$\widehat{p}_r$	8.8	39.4	6.72
		$\widehat{p}_{r.e}$	0.1	13.2	0.75
		$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	0.1	13.2	0.75
		$\widehat{p}_d$	0.1	13.7	0.81
100	0.9	$\widehat{p}_A$	-0.1	10.3	1.00
		$\widehat{p}_r$	0.1	4.7	0.21
		$\widehat{p}_{r.e}$	0.0	4.6	0.20
		$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	0.0	4.6	0.20
		$\widehat{p}_d$	0.0	4.7	0.21
	0.5	$\widehat{p}_A$	0.1	10.3	1.00
		$\widehat{p}_r$	3.7	20.7	4.04
		$\widehat{p}_{r.e}$	0.2	8.9	0.75
		$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	0.2	8.9	0.75
		$\widehat{p}_d$	0.2	9.3	0.82

Tabla 3.4: Valores del sesgo relativo (SR), error cuadrático medio relativo (ECMR) y eficiencia relativa (ER) para distintos estimadores de P_A =0.215 en la población Lagos. P_B =0.215 cuando ϕ =0.9 y P_B =0.522 cuando ϕ =0.5.

n	ϕ	Estimador	SR (%)	ECMR	ER
50	0.9	$\widehat{p}_A$	-0.2	25.5	1.00
		$\widehat{p}_r$	0.7	10.9	0.18
		$\widehat{p}_{r.e}$	0.1	9.7	0.14
		$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	0.0	9.8	0.15
		$\widehat{p}_d$	0.0	9.8	0.15
	0.5	$\widehat{p}_A$	-0.8	25.7	1.00
		$\widehat{p}_r$	-0.5	22.5	0.77
		$\widehat{p}_{r.e}$	-0.5	22.5	0.77
		$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	-0.5	22.5	0.77
		$\widehat{p}_d$	-0.1	28.7	1.25
100	0.9	$\widehat{p}_A$	0.1	17.2	1.00
		$\widehat{p}_r$	0.1	7.0	0.17
		$\widehat{p}_{r.e}$	-0.1	6.6	0.15
		$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	-0.1	6.7	0.15
		$\widehat{p}_d$	-0.1	6.7	0.15
	0.5	$\widehat{p}_A$	-0.3	17.3	1.00
		$\widehat{p}_r$	0.0	15.1	0.76
		$\widehat{p}_{r.e}$	0.0	15.1	0.76
		$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	0.0	15.1	0.76
		$\widehat{p}_d$	0.4	19.4	1.26

Tabla 3.5: Valores del sesgo relativo (SR), error cuadrático medio relativo (ECMR) y eficiencia relativa (ER) para distintos estimadores de P_A =0.07 en la población Lagos. ϕ =0.5 y P_B =0.176.

n	Estimador	SR (%)	ECMR	ER
50	$\widehat{p}_A$	0.5	49.0	1.00
	$\widehat{p}_r$	2.3	48.1	0.96
	$\widehat{p}_{r.e}$	2.1	46.7	0.91
	$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	0.6	45.1	0.85
	$\widehat{p}_d$	0.0	65.0	1.76
100	$\widehat{p}_A$	0.3	33.0	1.00
	$\widehat{p}_r$	0.8	29.7	0.81
	$\widehat{p}_{r.e}$	0.8	29.7	0.81
	$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	0.1	29.1	0.78
	$\widehat{p}_d$	-0.2	43.8	1.76

tienen sesgos en torno al 2 %, mientras que el resto de estimadores obtienen sesgo por debajo del 1 %. Al ser $P_A < P_B$, era evidente que el estimador de tipo diferencia iba a ser menos eficiente que el estimador estándar, de acuerdo a los estudios anteriores. La ganancia en eficiencia del resto de estimadores basados en información auxiliar con respecto al estimador estándar no es tan importante como en casos anteriores, especialmente cuando $n = 50$.

Como puede comprobarse en el Apéndice A, la población ENS contiene dos variables auxiliares, y por esta razón se han realizado estudios de simulación bajo esta población donde se incluyen ambas variables en la etapa de estimación, es decir, se obtienen los estimadores $\widehat{p}_{mr}$ y $\widehat{p}_{mr.e}$ basados en más de un atributo auxiliar. Para el resto de estimadores indirectos se utilizará el primer atributo auxiliar (B_1) como información auxiliar en la etapa de estimación. Los resultados obtenidos de esta simulación pueden consultarse en las Tablas 3.6 y 3.7.

En estas dos tablas observamos que los valores de sesgo relativo de los estimadores $\widehat{p}_{mr}$ y $\widehat{p}_r$ son bastantes elevados, llegándose a obtener sesgos que superan el 20 % en términos absolutos. De forma inusual, un incremento en el tamaño muestral no produce mejora para el sesgo relativo del estimador $\widehat{p}_r$, como podemos observar en la Tabla 3.7. En estas tablas también observamos que el estimador estándar, a pesar de ser insesgado, tiene sesgos elevados, especialmente cuando el tamaño de la muestra es menor, donde supera el 10 %. Este hecho se debe a que P_A está muy próximo a 0, y si bien el estimador puede

Tabla 3.6: Valores del sesgo relativo (SR), error cuadrático medio relativo (ECMR) y eficiencia relativa (ER) para distintos estimadores de P_A =0.07 en la población ENS. ϕ_1 =0.583, ϕ_2 =0.57 y $P_{B1} = P_{B2}$ =0.03.

n	Estimador	SR (%)	ECMR	ER
50	$\widehat{p}_A$	10.1	49.7	1.00
	$\widehat{p}_r$	-1.0	52.6	1.12
	$\widehat{p}_{r.e}$	-1.1	41.7	0.70
	$\widehat{p}_{mr}$	-24.1	36.1	0.53
	$\widehat{p}_{mr.e}$	4.4	38.9	0.61
	$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	-0.3	41.5	0.70
	$\widehat{p}_d$	-1.6	41.6	0.70
100	$\widehat{p}_A$	2.4	35.0	1.00
	$\widehat{p}_r$	19.5	70.5	4.06
	$\widehat{p}_{r.e}$	0.5	29.2	0.70
	$\widehat{p}_{mr}$	-7.5	36.7	1.10
	$\widehat{p}_{mr.e}$	1.3	29.5	0.71
	$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	0.6	29.3	0.70
	$\widehat{p}_d$	0.4	29.2	0.70

obtener estimaciones por encima del parámetro sin ningún tipo de acotación (como sucede en cualquier otro caso), las estimaciones de este estimador que quedan por debajo del parámetro están acotadas en 0, y de ahí que este estimador presente un sesgo elevado y positivo, el cual se corrige a medida que aumenta el tamaño de la muestra. Por su parte, el resto de estimadores propuestos tienen sesgos dentro de un rango razonable de valores.

Respecto la eficiencia de los estimadores, los estimadores de tipo razón $\widehat{p}_{mr}$ y $\widehat{p}_r$ pueden ser mucho menos eficientes que el estimador estándar, puesto que $P_A > P_B$ en ambas poblaciones. El resto de estimadores propuestos son siempre más eficientes que el estimador estándar. Por último, destacamos que la introducción de más atributos auxiliares no produce importantes mejorías, y esto puede deberse a que las correlaciones no son muy elevadas en ese caso.

Tabla 3.7: Valores del sesgo relativo (SR), error cuadrático medio relativo (ECMR) y eficiencia relativa (ER) para distintos estimadores de P_A =0.12 en la población ENS. ϕ_1 =0.51, ϕ_2 =0.495 y $P_{B1} = P_{B2}$ =0.04.

n	Estimador	SR (%)	ECMR	ER
50	$\widehat{p}_A$	3.8	38.2	1.00
	$\widehat{p}_r$	13.4	65.3	2.92
	$\widehat{p}_{r.e}$	-0.5	33.8	0.78
	$\widehat{p}_{mr}$	-17.3	37.1	0.94
	$\widehat{p}_{mr.e}$	1.4	34.2	0.80
	$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	0.1	33.8	0.78
	$\widehat{p}_d$	-0.8	33.8	0.78
100	$\widehat{p}_A$	-0.1	27.4	1.00
	$\widehat{p}_r$	21.6	74.6	7.41
	$\widehat{p}_{r.e}$	-0.2	23.8	0.75
	$\widehat{p}_{mr}$	-1.0	41.0	2.24
	$\widehat{p}_{mr.e}$	0.2	23.9	0.76
	$\widehat{p}_{r.opt}$ $(=\widehat{p}_{reg}^{opt})$	-0.1	23.9	0.76
	$\widehat{p}_d$	-0.2	23.9	0.76

3.2. Extensión a un diseño muestral general

3.2.1. Definición del estimador

Definido el estimador de tipo regresión y analizadas sus propiedades bajo MAS, el siguiente paso es extender estos estudios al caso de un diseño muestral general. Por tanto, en esta sección consideraremos que la muestra s se ha seleccionado de acuerdo a un diseño muestral con probabilidades de inclusión de primer y segundo orden dadas, respectivamente, por π_i y π_{ij}.

Prescindiendo del uso de información auxiliar, un estimador simple para P_A que tiene en cuenta el diseño muestral viene dado por

$$\widehat{p}_{A.HT} = \frac{1}{N} \sum_{i \in s} d_i A_i, \tag{3.11}$$

donde $d_i = \pi_i^{-1}$ es el peso básico del diseño. Asumiendo información auxiliar en la etapa de estimación y dada una constante b, el estimador de tipo regresión

bajo un diseño muestral general se puede definir como

$$\widetilde{p}_{reg.HT} = \widehat{p}_{A.HT} + b(P_B - \widehat{p}_{B.HT}),$$

donde $\widehat{p}_{B.HT}$ viene dado por (3.11) después de sustituir A_i por B_i.

El objetivo siguiente será determinar la constante b de forma que el error del estimador $\widetilde{p}_{reg.HT}$ sea lo más pequeño posible. En el caso de un diseño de muestreo general, proponemos un estimador definido a partir del estimador del coeficiente de regresión que incluya los pesos básicos del diseño. Concretamente, el estimador de tipo regresión que se propone es:

$$\widehat{p}_{reg.HT} = \widehat{p}_{A.HT} + \widehat{b}(P_B - \widehat{p}_{B.HT}),$$

donde el coeficiente de regresión viene dado por

$$\widehat{b} = \frac{\sum_{i\in s} d_i(A_i - \widehat{p}_{A.HT})(B_i - \widehat{p}_{B.HT})}{\sum_{i\in s} d_i(B_i - \widehat{p}_{B.HT})^2}.$$

Al igual que en caso de MAS, el estimador de tipo diferencia se puede definir como un caso particular de $\widetilde{p}_{reg.HT}$ cuando $b = 1$. Es decir, el estimador de tipo diferencia en el caso de un diseño muestral general viene dado por

$$\widehat{p}_{d.HT} = \widehat{p}_{A.HT} + P_B - \widehat{p}_{B.HT}.$$

3.2.2. Propiedades teóricas

La varianza asintótica de $\widetilde{p}_{reg.HT}$ viene dada por

$$V(\widetilde{p}_{reg.HT}) = \frac{1}{N^2} \sum_{i<j\in U} (\pi_i\pi_j - \pi_{ij}) \left(\frac{A_i - \widehat{A}_i}{\pi_i} - \frac{A_j - \widehat{A}_j}{\pi_j} \right)^2,$$

donde $\widehat{A}_i = \widehat{p}_{A.HT} + b(B_i - \widehat{p}_{B.HT})$. Un estimador de $V(\widetilde{p}_{reg.HT})$ es

$$\widehat{V}(\widetilde{p}_{reg.HT}) = \frac{1}{N^2} \sum_{i<j\in s} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{A_i - \widehat{A}_i}{\pi_i} - \frac{A_j - \widehat{A}_j}{\pi_j} \right)^2.$$

Para obtener las propiedades del estimador $\widehat{p}_{reg.HT}$, usaremos la técnica de linearización de Taylor (véase Särndal et al., 1992). Dado que $\widehat{p}_{reg.HT}$ puede aproximarse mediante la linearización de Taylor por

$$\widehat{p}_{A.HT} + \frac{\sum_{i\in U}(A_i - P_A)(B_i - P_B)}{\sum_{i\in U}(B_i - P_B)^2}(P_B - \widehat{p}_{B.HT}) =$$

$$\frac{1}{N}\sum_{i\in U}\frac{\sum_{i\in U}(A_i-P_A)(B_i-P_B)}{\sum_{i\in U}(B_i-P_B)^2}B_i+\frac{1}{N}\sum_{i\in s}\frac{E_i}{\pi_i},$$

donde

$$E_i=A_i-\frac{\sum_{i\in U}(A_i-P_A)(B_i-P_B)}{\sum_{i\in U}(B_i-P_B)^2}B_i,$$

el estimador $\widehat{p}_{reg.HT}$ es asintóticamente insesgado. Usando las propiedades de los estimadores tipo Horvitz y Thompson, podemos deducir la expresión de la varianza aproximada de $\widehat{p}_{reg.HT}$:

$$AV(\widehat{p}_{reg.HT})=\frac{1}{N^2}\sum_{i<j\in U}(\pi_i\pi_j-\pi_{ij})\left(\frac{E_i}{\pi_i}-\frac{E_j}{\pi_j}\right)^2. \tag{3.12}$$

Siguiendo los principios de estimación de la varianza y reemplazando los valores no observados E_i por los valores observados $e_i=A_i-\widehat{b}B_i$, obtenemos el estimador de la varianza

$$\widehat{V}(\widehat{p}_{reg.HT})=\frac{1}{N^2}\sum_{i<j\in s}\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\left(\frac{e_i}{\pi_i}-\frac{e_j}{\pi_j}\right)^2.$$

El estudio de normalidad asintótica del estimador de regresión en un diseño de muestreo general es una cuestión compleja. Asumiendo MAS, la condición suficiente y necesaria para la normalidad asintótica de la media muestral viene dada por la condición Lindeberg. La generalización del estimador de Horvitz-Thompson viene dada por Hájek (1964), pero solo para un diseño muestral específico. Asumiendo un diseño de muestreo general, Berger (1998) obtiene una condición suficiente para obtener la normalidad asintótica del estimador de Horvitz-Thompson, la cual se puede usar para demostrar que $\widehat{p}_{reg.HT}$ se distribuye asintóticamente normal. En efecto, $\widehat{p}_{reg.HT}$ se puede expresar como una combinación lineal de dos estimadores de tipo Horvitz-Thompson, los cuales hacen que $\widehat{p}_{reg.HT}$ se distribuya normalmente.

El uso de N en (3.11) puede dar lugar a estimaciones de P_A mayores que 1. Ante esta situación, una solución muy simple sería usar el estimador tipo Hájek $\widehat{p}_{A.H}=\widehat{N}^{-1}\sum_{i\in s}d_iA_i$, donde $\widehat{N}=\sum_{i\in s}d_i$. Se observa que $\widehat{p}_{A.H}=1$ cuando $A_i=1$ para todo $i\in s$, y por tanto se solucionan los problemas de sobreestimación. Se ha de destacar que $\widehat{p}_{A.H}$ es un estimador de tipo razón, y siguiendo la idea de Särndal (2007), podemos concluir que $\widehat{p}_{A.H}$ es asintóticamente insesgado y se puede obtener una expresión aproximada de la varianza. Las propiedades del estimador de regresión basadas en $\widehat{p}_{A.H}$ se pueden también obtener usando la técnica de linearización que se usó al inicio de esta sección.

Respecto a las varianzas del estimador de tipo diferencia, la varianza poblacional de dicho estimador viene dada por

$$V(\widehat{p}_{d.HT}) = \frac{1}{N^2} \sum_{i<j \in U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{A_i - \widetilde{A}_i}{\pi_i} - \frac{A_j - \widetilde{A}_j}{\pi_j} \right)^2,$$

donde $\widetilde{A}_i = \widehat{p}_{A.HT} + (B_i - \widehat{p}_{B.HT})$. Un estimador de esta varianza viene dada por la expresión

$$\widehat{V}(\widehat{p}_{d.HT}) = \frac{1}{N^2} \sum_{i<j \in s} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{A_i - \widetilde{A}_i}{\pi_i} - \frac{A_j - \widetilde{A}_j}{\pi_j} \right)^2.$$

3.2.3. Comparación empírica de estimadores

A continuación se comparan numéricamente los distintos estimadores en diferentes poblaciones y con muestras extraídas bajo un diseño muestral con probabilidades de selección desiguales. En concreto, se ha utilizado muestreo estratificado con estratificación basada en el segundo atributo auxiliar de cada población utilizada en este estudio. El uso de afijación uniforme en este estudio de simulación permitirá la obtención de pesos muestrales con mucha variación y analizar el efecto que pueden tener estos pesos muestrales en los distintos estimadores. Notamos que este esquema de muestreo con probabilidades desiguales se ha utilizado en distintos estudio de simulación con el objetivo de comparar el comportamiento de distintos estimadores, como por ejemplo en Chambers y Dunstan (1986), Rao et al. (1990), etc.

Las poblaciones utilizadas en este estudio de simulación son las poblaciones simuladas, también utilizadas en el estudio anterior, y la población ENS, la cual dispone de dos variables auxiliares. En todos los casos, se utilizará el primer atributo auxiliar como información auxiliar en la etapa de estimación y el segundo atributo auxiliar para estratificar la población. El resto de poblaciones reales no se han utilizado en este estudio porque tan sólo contienen un atributo auxiliar y se necesitan dos para llevar a cabo esta simulación: un atributo que utilizan los estimadores indirectos y otro atributo para estratificar.

Siguiendo también el estudio de simulación anterior, se utilizarán como medidas para comparar el comportamiento de los distintos estimadores el sesgo relativo (SR), el error cuadrático medio relativo (ECMR) y la eficiencia relativa (ER) obtenidos a partir de 10000 muestras.

Figura 3.4: Valores de eficiencia relativa (ER) para varios estimadores de P_A obtenidos a partir de las poblaciones simuladas. Las muestras son seleccionadas con tamaño $n = 100$. P_A varía de 0.1 a 0.9 y ϕ_1 ($=\phi$) varía de 0.5 a 0.9. ϕ_2 (B_2 se utiliza para estratificar) toma los mismos valores que ϕ_1 (B_1 se utiliza en la etapa de estimación).

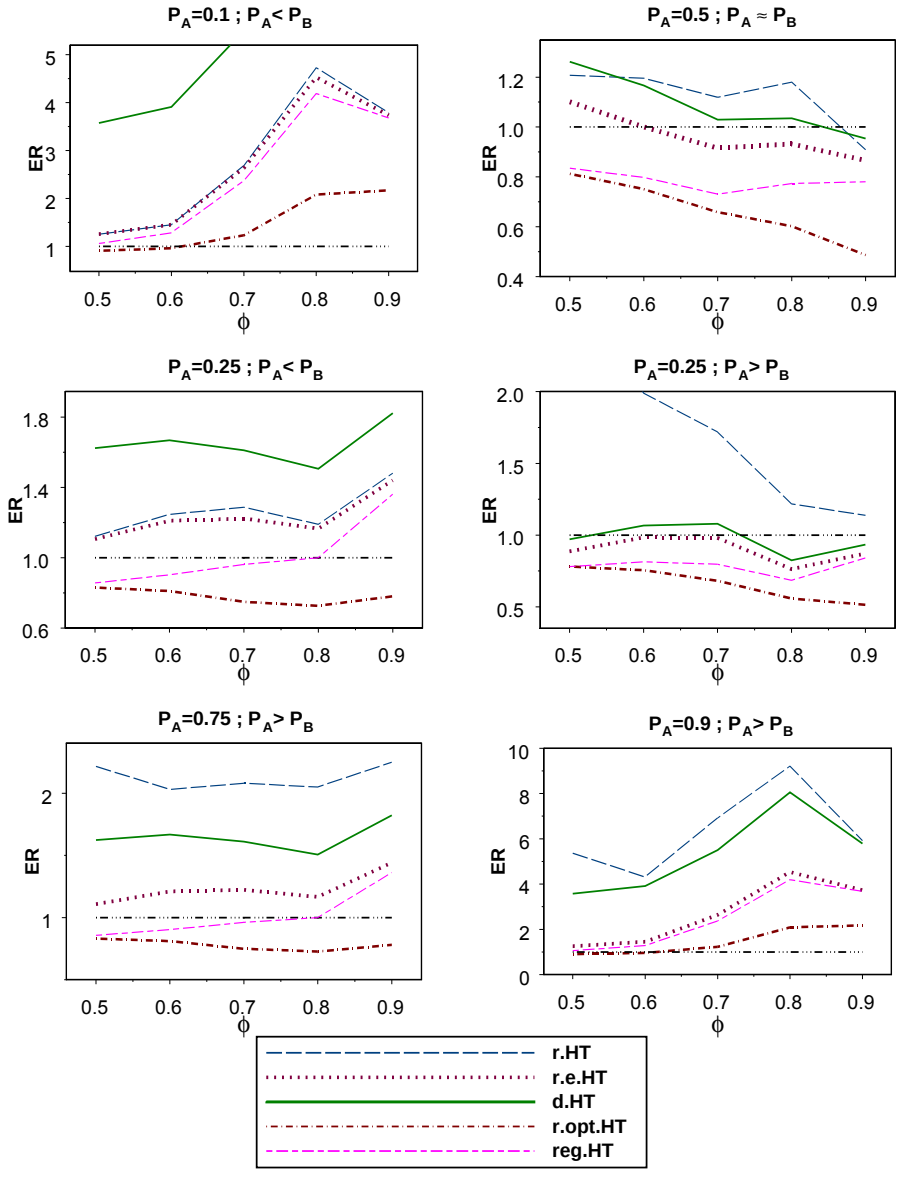
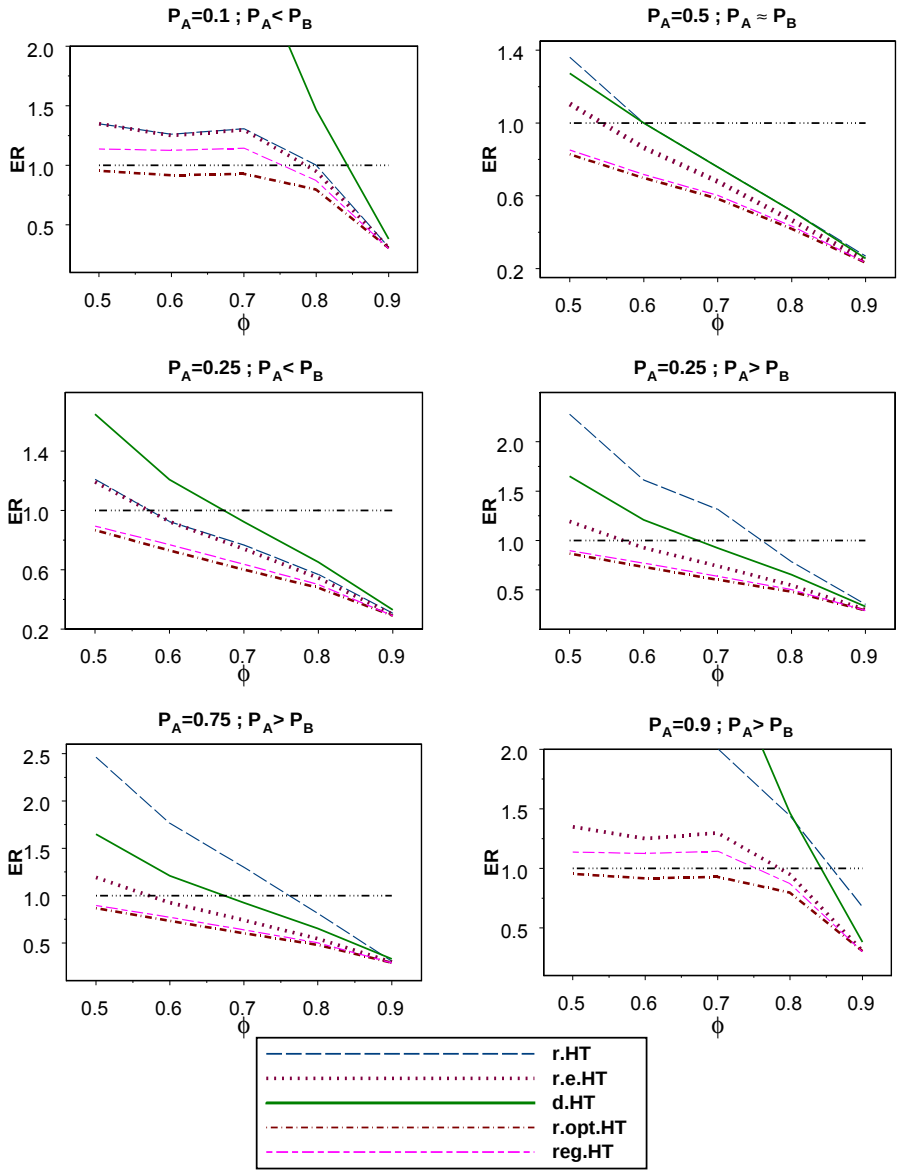


Figura 3.5: Valores de eficiencia relativa (ER) para varios estimadores de P_A obtenidos a partir de las poblaciones simuladas. Las muestras son seleccionadas con tamaño $n = 100$. P_A varía de 0.1 a 0.9 y ϕ_1 (=ϕ) varía de 0.5 a 0.9. ϕ_2 (B_2 se utiliza para estratificar) toma siempre el valor 0.5. B_1 se utiliza en la etapa de estimación.



Las Figuras 3.4 y 3.5 muestran la eficiencia relativa de los diferentes estimadores en el caso de muestreo con probabilidades desiguales. En la Figura 3.4, el atributo utilizado para estratificar tiene siempre el mismo coeficiente V de Cramer con el atributo de interés que el atributo utilizado por los estimadores indirectos en la etapa de estimación. Por su parte, en la Figura 3.5, el atributo utilizado para estratificar tiene fijado el coeficiente V de Cramer en 0.5, mientras que el otro atributo va variando de 0.5 a 0.9.

A partir de estas figuras observamos que la correlación que tenga el atributo utilizado para estratificar tiene un importante impacto en el comportamiento de los estimadores. De este modo, cuando $\phi_2 = \phi_1$ (Figura 3.4) la población queda bien estratificada, el estimador estándar tiene un buen comportamiento y puede ser más eficiente que otros estimadores indirectos. En cualquier caso, el estimador $\widehat{p}_{r.opt.HT}$ es siempre más eficiente que el estimador estándar, excepto cuando P_A está próximo a 0 o bien próximo a 1. En estos casos, la estratificación cuando $\phi_2 > 0{,}7$ es bastante buena y el estimador estándar es el estimador más eficiente. De los estimadores indirectos, el estimador $\widehat{p}_{reg.HT}$ es el que muestra el segundo mejor comportamiento en términos de ER, después del estimador $\widehat{p}_{r.opt.HT}$. Respecto al resto de estimadores indirectos se cumplen las mismas relaciones que las estudiadas bajo MAS, es decir, $\widehat{p}_{r.e.HT}$ es siempre más eficiente que $\widehat{p}_{r.HT}$, $\widehat{p}_{r.HT}$ es más eficiente que $\widehat{p}_{d.HT}$ cuando $P_A < P_B$ y $\widehat{p}_{r.HT}$ es menos eficiente que $\widehat{p}_{d.HT}$ en el caso contrario.

Cuando el coeficiente V de Cramer se fija en 0.5 (Figura 3.5), los estimadores basados en información auxiliar si muestran importantes beneficios en comparación con el estimador estándar. En este caso, el estimador $\widehat{p}_{r.opt.HT}$ es siempre más eficiente que el estimador estándar. El segundo estimador más eficiente es $\widehat{p}_{reg.HT}$, el cual sólo es menos eficiente que el estimador estándar cuando P_A toma los valores 0.1 y 0.9 y $\phi < 0{,}8$.

En conclusión, estos estudios de simulación aconsejan utilizar el estimador $\widehat{p}_{r.opt.HT}$, excepto cuando P_A tome valores cercanos a los extremos del intervalo [0,1], y al mismo tiempo se cumpla que el coeficiente V de Cramer del atributo utilizado para estratificar tome un valor muy alto. Cuando se den estas circunstancias, el estimador estándar podría ser más eficiente que $\widehat{p}_{r.opt.HT}$.

Las Tablas 3.8 y 3.9 muestran los resultados del estudio de simulación para la población ENS. Observamos que tan sólo los estimadores $\widehat{p}_{r.HT}$ y $\widehat{p}_{r.opt.HT}$ obtienen sesgos un poco más elevados que el resto estimadores, en torno al 2 %. A diferencia del caso del MAS, los estimadores indirectos, a excepción del estimador de tipo razón, tienen una eficiencia muy similar al estimador estándar, pero siempre mejoran a dicho estimador. En cualquier caso, observamos con

Tabla 3.8: Para muestreo estratificado, valores del sesgo relativo (SR), error cuadrático medio relativo (ECMR) y eficiencia relativa (ER) para distintos estimadores de P_A =0.07 en la población ENS. ϕ_1 =0.583, ϕ_2 =0.57 y $P_{B1} = P_{B2}$ =0.03.

n	Estimador	SR (%)	ECMR	ER
50	$\widehat{p}_{A.HT}$	-0.6	55.9	1.00
	$\widehat{p}_{r.HT}$	1.1	57.4	1.05
	$\widehat{p}_{r.e.HT}$	0.6	56.3	1.01
	$\widehat{p}_{r.opt.HT}$	-2.1	55.2	0.98
	$\widehat{p}_{reg.HT}$	-0.7	55.1	0.97
	$\widehat{p}_{d.HT}$	-0.6	55.1	0.97
100	$\widehat{p}_{A.HT}$	-0.9	39.8	1.00
	$\widehat{p}_{r.HT}$	0.3	40.8	1.05
	$\widehat{p}_{r.e.HT}$	-0.1	39.7	0.99
	$\widehat{p}_{r.opt.HT}$	-2.3	39.4	0.98
	$\widehat{p}_{reg.HT}$	-0.9	39.2	0.97
	$\widehat{p}_{d.HT}$	-0.9	39.2	0.97

Tabla 3.9: Para muestreo estratificado, valores del sesgo relativo (SR), error cuadrático medio relativo (ECMR) y eficiencia relativa (ER) para distintos estimadores de P_A =0.12 en la población ENS. ϕ_1 =0.51, ϕ_2 =0.495 y $P_{B1} = P_{B2}$ =0.04.

n	Estimador	SR (%)	ECMR	ER
50	$\widehat{p}_{A.HT}$	-0.9	45.2	1.00
	$\widehat{p}_{r.HT}$	1.8	48.0	1.13
	$\widehat{p}_{r.e.HT}$	1.2	46.2	1.04
	$\widehat{p}_{r.opt.HT}$	-2.6	44.7	0.98
	$\widehat{p}_{reg.HT}$	-0.8	44.5	0.97
	$\widehat{p}_{d.HT}$	-0.9	44.7	0.98
100	$\widehat{p}_{A.HT}$	-0.2	32.3	1.00
	$\widehat{p}_{r.HT}$	1.7	34.3	1.13
	$\widehat{p}_{r.e.HT}$	1.1	32.6	1.02
	$\widehat{p}_{r.opt.HT}$	-1.7	32.0	0.98
	$\widehat{p}_{reg.HT}$	-0.1	31.9	0.98
	$\widehat{p}_{d.HT}$	-0.1	31.9	0.98

los coeficiente de V de Cramer no son muy elevados, por lo es razonable pensar que la eficiencia de los estimadores indirectos mejoren considerablemente a medida que aumenta el mencionado coeficiente, tal como sucede en los estudios de simulación de las poblaciones simuladas.

Capítulo 4

Estimación de una proporción mediante intervalos de confianza

4.1. Introducción

Los intervalos de confianza para una proporción poblacional propuestos en la literatura están basados en la suposición de una población infinita. Sin embargo, esta situación puede ser considerada como irreal en la práctica, es decir, es bastante común encontrarse muestras extraídas de una población finita. En esta sección, se derivan los intervalos de confianza descritos en Newcombe (1998), los cuales asumen una población infinita. Sin embargo, en este trabajo se describen dichos métodos bajo el contexto de una población finita y asumiendo MAS.

Asumiendo el estimador estándar para una proporción poblacional, existen varios métodos para construir intervalos de confianza. Por ejemplo, se puede considerar la aproximación Gausiana tradicional (método de Wald en lo sucesivo y de acuerdo con Vollset, 1993). Este método tiene la ventaja de obtener intervalos centrados en la estimación puntual, pero también los inconvenientes de obtener intervalos con extremos fuera del intervalo [0,1], así como pueden aparecer problemas de degeneración, es decir, intervalos formados por un único punto.

Con el fin de solucionar los problemas de degeneración, una opción (utilizada por autores como Newcombe, 1998; Blyth y Still, 1983; etc) es considerar la corrección por continuidad (CC), dada por $1/(2n)$. El uso de esta corrección produce una mejor cobertura en los intervalos de confianza, al obtenerse inter-

valos de confianza más amplios. Sin embargo, esto tiene como inconveniente que los límites o extremos del intervalo se salgan en más ocasiones del intervalo $[0, 1]$.

El método de Wilson o Score (Wilson, 1927) es otro método comúnmente utilizado para la construcción de intervalos de confianza para una proporción poblacional. Este método utiliza la varianza asintótica del estimador puntual para la construcción del intervalo de confianza, y tiene la ventaja de obtener intervalos con extremos dentro del intervalo [0,1].

Supongamos que z denota el cuantil de orden $1 - \alpha/2$ de la distribución normal estándar. En el caso de una población finita y bajo MAS, algunos métodos para construir intervalos de confianza con nivel de confianza $(1 - \alpha)$ y basados en el estimador estándar $\widehat{p}_A$ son los siguientes:

1. Método asintótico (Método de Wald de acuerdo con Vollset, 1993) sin corrección por continuidad

$$\widehat{p}_A \pm z\sqrt{\frac{1-f}{n-1}\widehat{p}_A\widehat{q}_A}.$$

2. Método asintótico con corrección por continuidad (Blyth y Still, 1983)

$$\widehat{p}_A \pm \left(z\sqrt{\frac{1-f}{n-1}\widehat{p}_A\widehat{q}_A} + \frac{1}{2n}\right).$$

3. Método de Wilson o Score (Wilson, 1927). El intervalo de confianza está formado por todo θ que verifica que

$$|\widehat{p}_A - \theta| \leq z\sqrt{\frac{\theta(1-\theta)}{n}\frac{N-n}{N-1}},$$

que proporciona el intervalo:

$$\frac{2nk_1\widehat{p}_A + z^2 \pm z\sqrt{z^2 + 4nk_1\widehat{p}_A\widehat{q}_A}}{2(nk_1 + z^2)},$$

donde $k_1 = (N-1)/(N-n)$.

4. Método Score incorporando la corrección por continuidad (Blyth y Still, 1983; Fleiss et al., 2003). El intervalo de confianza está formado por todo θ que verifica

$$|\widehat{p}_A - \theta| - \frac{1}{2n} \leq z\sqrt{\frac{\theta(1-\theta)}{n}\frac{N-n}{N-1}}.$$

Las expresiones para el límite inferior (L) y el límite superior (U) de este intervalo de confianza son las siguientes:

$$L = \frac{2nk_1\widehat{p}_A + z^2 - 1 - z\sqrt{z^2 - 2 - 1/n + 4\widehat{p}_A(nk_1\widehat{q}_A + 1)}}{2(nk_1 + z^2)}$$

y

$$U = \frac{2nk_1\widehat{p}_A + z^2 + 1 + z\sqrt{z^2 + 2 - 1/n + 4\widehat{p}_A(nk_1\widehat{q}_A - 1)}}{2(nk_1 + z^2)}.$$

5. Método basado en las colas exactas de la binomial (Clopper y Pearson, 1934; Lentner, 1982). El intervalo está dado por $[L, U]$, con $L \leq \widehat{p}_A \leq U$, y tal que para todo θ en el intervalo se verifica

   a) Si $L \leq \theta \leq \widehat{p}_A$, $\quad kp_t + \sum_{j=t+1}^{n} p_j \geq \frac{\alpha}{2}$

   b) Si $\widehat{p}_A \leq \theta \leq U$, $\quad \sum_{j=0}^{t-1} p_j + kp_t \geq \frac{\alpha}{2}$

   respectivamente, donde:

$$p_j = Pr[T = j] = \left(\begin{array}{c} n \\ j \end{array} \right) \theta^j (1-\theta)^{n-j},$$

   $j = 0, 1, \ldots, n$, $t = n_{1\cdot}$ y $k = 1$.

6. Método basado en las colas de la binomial pero con k =0.5 (en adelante denominado "Binomial mid"). Este método puede consultarse, en el caso de poblaciones infinitas, en Miettinen (1985), Cohen y Yang (1994), etc.

7. Método basado en la verosimilitud (Miettinen y Nurminen, 1985). El intervalo está formado por todo θ que satisface

$$t \ln\theta + (n-t)\ln(1-\theta) \geq t \ln\widehat{p}_A + (n-t)\ln(1-\widehat{p}_A) - z^2/2.$$

4.2. Muestreo aleatorio simple

4.2.1. Construcción de intervalos de confianza

En esta sección, se derivan intervalos de confianza para P_A sobre la base del estimador propuesto $\widehat{p}_r$. La extensión al resto de los estimadores propuestos

es bastante directa a partir de las expresiones de las varianzas del resto de estimadores, las cuales se han indicado en esta memoria. Por esta razón se han omitido las expresiones de estos intervalos de confianza.

El método más simple basado en la aproximación asintótica Gaussiana (método de Wald) viene dado por

$$\widehat{p}_r \pm z\sqrt{\widehat{V}(\widehat{p}_r)},$$

donde $\widehat{V}(\widehat{p}_r)$ está definida en (2.12). Este método asintótico con factor de corrección por continuidad (Blyth y Still, 1983) es

$$\widehat{p}_r \pm \left(z\sqrt{\widehat{V}(\widehat{p}_r)} + \frac{1}{2n} \right).$$

Como se comentó en la Sección 2.2.2, el estimador $\widehat{p}_r$ es asintóticamente insesgado, pero el sesgo podría no ser despreciable para tamaños de muestra pequeños. En esta situación, podemos utilizar los siguientes intervalos de confianza que tienen en cuenta el sesgo del estimador propuesto:

$$\widehat{p}_r \pm z\sqrt{\widehat{V}(\widehat{p}_r) + \widehat{B}^2(\widehat{p}_r)}$$

y

$$\widehat{p}_r \pm \left(z\sqrt{\widehat{V}(\widehat{p}_r) + \widehat{B}^2(\widehat{p}_r)} + \frac{1}{2n} \right).$$

Asumiendo el Método Score, el intervalo de confianza basado en el estimador propuesto $\widehat{p}_r$ consiste en considerar los valores θ tales que

$$|\widehat{p}_r - \theta| \leq z\sqrt{\frac{N-n}{(N-1)n}\left(\theta(1-\theta) + \frac{\theta^2}{P_B}Q_B - 2\frac{\theta}{P_B}\widehat{\phi}\sqrt{\theta(1-\theta)P_BQ_B}\right)}.$$

También se puede usar este método Score incorporando el factor de corrección por continuidad. Este intervalo consiste en considerar todo θ tal que

$$|\widehat{p}_r - \theta| - \frac{1}{2n} \leq z\sqrt{\frac{N-n}{(N-1)n}\left(\theta(1-\theta) + \frac{\theta^2}{P_B}Q_B - 2\frac{\theta}{P_B}\widehat{\phi}\sqrt{\theta(1-\theta)P_BQ_B}\right)}.$$

Los intervalos basados en el resto de estimadores de razón propuestos se obtienen sin más que sustituir los estimadores y sus varianzas por las correspondientes expresiones detalladas en el Capítulo 2. A modo de ejemplo, también se presentan los intervalos de confianza bilaterales para P_A basados en el

estimador propuesto $\widehat{p}_{reg}^{opt}$. La extensión al estimador de tipo diferencia también es bastante simple.

En primer lugar, el Método de Wald proporciona el siguiente intervalo

$$\widehat{p}_{reg}^{opt} \pm z\sqrt{\widehat{V}(\widehat{p}_{reg}^{opt})},$$

donde $\widehat{V}(\widehat{p}_{reg}^{opt})$ se definió en el Corolario 3.3 del Capítulo 3.

El intervalo de confianza asintótico que usa la corrección por continuidad viene dado por

$$\widehat{p}_{reg}^{opt} \pm \left(z\sqrt{\widehat{V}(\widehat{p}_{reg}^{opt})} + \frac{1}{2n} \right).$$

También se pueden construir intervalos de confianza basados en el estimador $\widehat{p}_{reg}^{opt}$ y utilizando el método Score (véase Wilson, 1927). El intervalo de confianza propuesto está formado por todo θ que verifica

$$|\widehat{p}_{reg}^{opt} - \theta| \leq z\sqrt{\frac{N-n}{(N-1)n}\theta(1-\theta)(1-\phi^2)}.$$

El Método Score puede incorporar corrección por continuidad. Este intervalo comprende el conjunto de valores θ tal que:

$$|\widehat{p}_{reg}^{opt} - \theta| - \frac{1}{2n} \leq z\sqrt{\frac{N-n}{(N-1)n}\theta(1-\theta)(1-\phi^2)}.$$

4.2.2. Comparación empírica de intervalos de confianza

Poblaciones simuladas

A continuación se llevan a cabo estudios de simulación para evaluar empíricamente los distintos intervalos de confianza descritos en esta memoria. Se han realizado $D = 10000$ replicaciones a partir de las cuales se han obtenidos intervalos de confianza basados en los estimadores $\widehat{p}_A$, $\widehat{p}_r$, $\widehat{p}_{r.e}$, $\widehat{p}_d$ y $\widehat{p}_{reg}^{opt}$, y para cada caso se han utilizado los métodos de Wald y Score y se han construido intervalos de confianza con y sin Corrección por Continuidad (CC).

Al igual que en el caso de la estimación puntual, las simulaciones Monte Carlo para la obtención de los distintos intervalos de confianza así como la evaluación de los mismos se han realizado mediante el Software libre R.

Figura 4.1: Valores de Amplitud media y Cobertura (CI) de los intervalos de confianza (al 95 %) basados en los estimadores $\widehat{p}_A$ (Estándar), $\widehat{p}_{r.e}$ y $\widehat{p}_{reg}^{opt}$ (reg) y obtenidos a partir de las poblaciones simuladas. Las muestras son seleccionadas con tamaño $n = 100$. P_A =0.25 y ϕ varía desde 0.5 a 0.9.

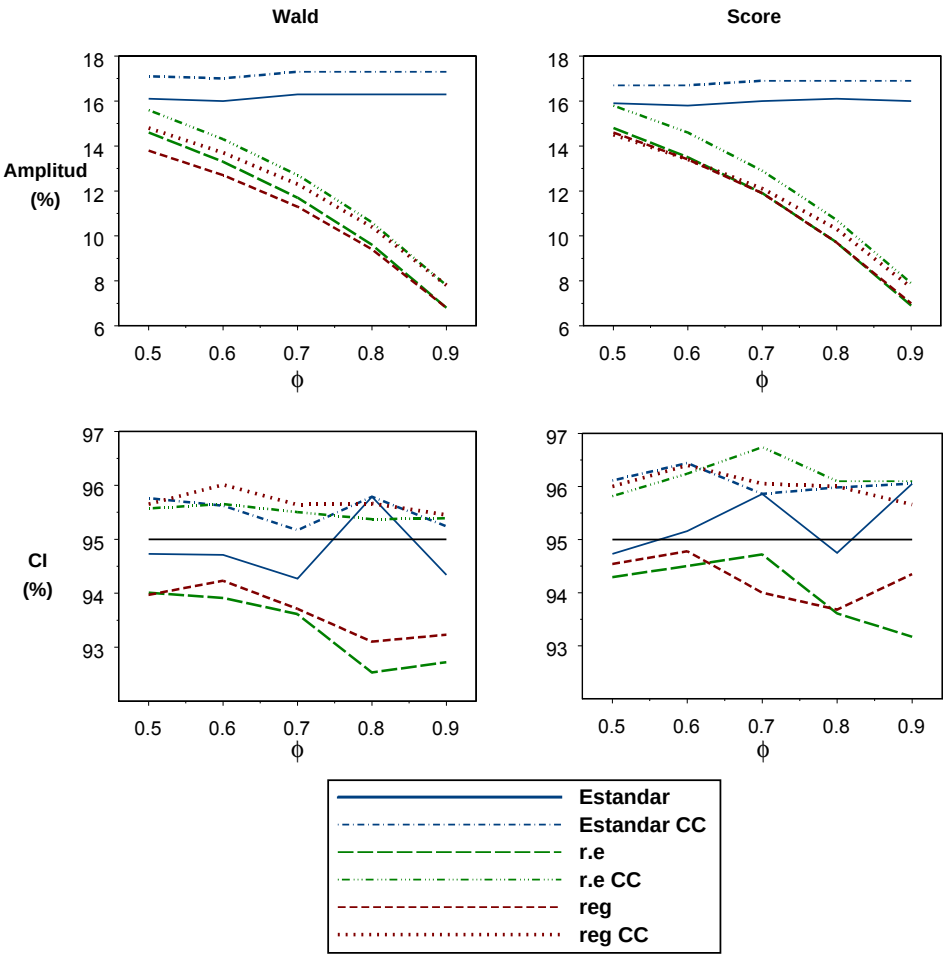
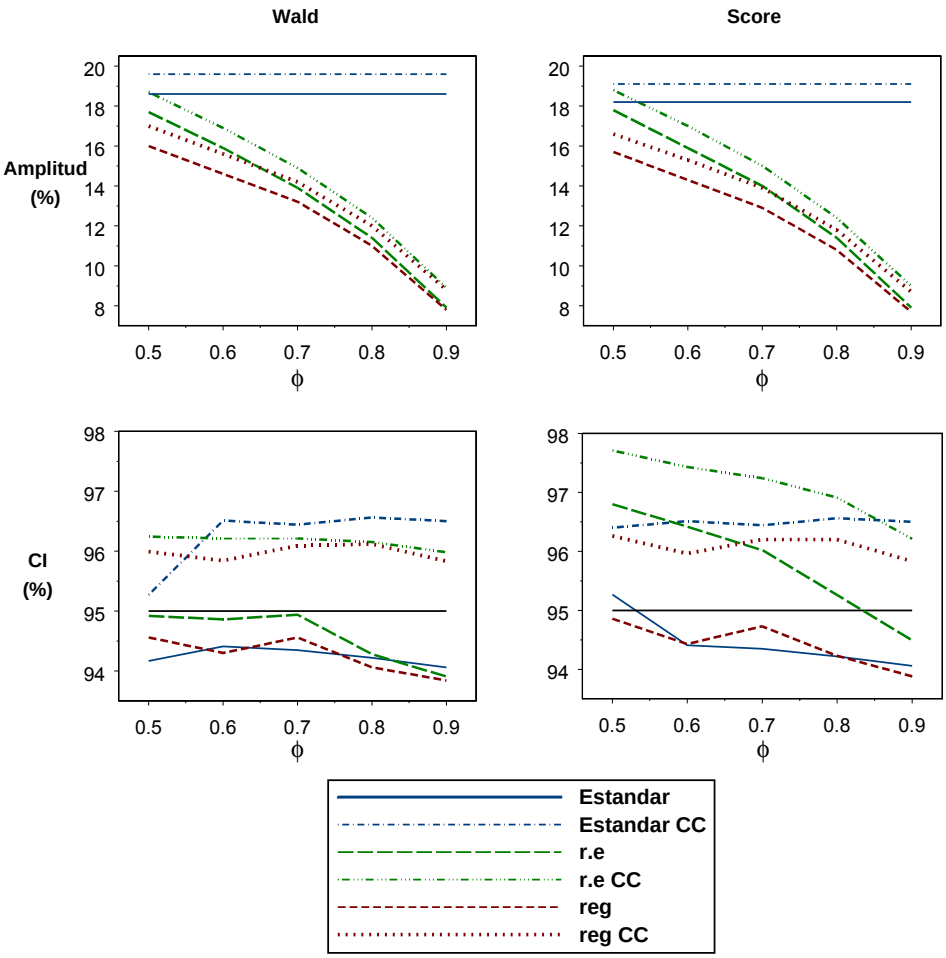


Figura 4.2: Valores de Amplitud media y Cobertura (CI) de los intervalos de confianza (al 95 %) basados en los estimadores $\widehat{p}_A$ (Estándar), $\widehat{p}_{r.e}$ y $\widehat{p}_{reg}$ (reg) y obtenidos a partir de las poblaciones simuladas. Las muestras son seleccionadas con tamaño $n = 100$. P_A =0.5 y ϕ varía desde 0.5 a 0.9.



Para las poblaciones simuladas, los criterios utilizados para evaluar los diferentes intervalos de confianza (construidos con un nivel de confianza del 95 %) son la cobertura empírica del intervalo (porcentaje de veces que los intervalos obtenidos contienen al verdadero parámetro) y la amplitud media (sobre las 10000 réplicas) del intervalo. Es importante reseñar que es deseable la cobertura requerida con la mínima amplitud.

Las Figuras 4.1 y 4.2 muestran los valores de amplitud media y cobertura empírica, en porcentaje, de los intervalos de confianza para algunos de los estimadores considerados y para las proporciones P_A =0.25 y P_A =0.5, respectivamente. El caso P_A =0.75 se omite dado que los intervalos muestran el mismo comportamiento que para el caso P_A =0.25.

Por motivos de espacio en las figuras, sólo se muestran los resultados de los intervalos correspondientes a los estimadores $\widehat{p}_{r.e}$ y $\widehat{p}_{reg}^{opt}$, dado que tales estimadores tienen un mejor comportamiento que los estimadores $\widehat{p}_r$ y $\widehat{p}_d$.

En la Figura 4.1 observamos que los intervalos basados en el estimador estándar (con y sin CC) producen los intervalos más amplios, tanto con el uso del método de Wald como con el método Score. Sin embargo, si prestamos atención a los resultados mostrados para los estimadores $\widehat{p}_{r.e}$ y $\widehat{p}_{reg}^{opt}$, observamos, como era de esperar, que los intervalos son más estrechos a medida que aumenta el Coeficiente V de Cramer. Aunque ligeramente, los intervalos basados en el estimador $\widehat{p}_{reg}^{opt}$ son los proporcionan los intervalos con menor amplitud.

En lo que respecta a la cobertura de los intervalos de confianza y continuando en la Figura 4.1, podemos observar que el estimador que, generalmente, aporta mayor cobertura es el estimador $\widehat{p}_{reg}^{opt}$ para ambos métodos (Wald y Score) y cuando se considera el coeficiente por continuidad (CC). En el lado opuesto, el intervalo que en general proporciona una menor cobertura es el basado en el estimador $\widehat{p}_{r.e}$, lo cual ocurre tanto cuando se utiliza el método Wald como el método Score. En líneas generales, los intervalos que más se aproximan a la cobertura requerida del 95 % son los basados en el estimador estándar, aunque destacamos que el resto de estimadores obtienen intervalos con coberturas también cercanas al 95 % y mejoran considerablemente a los intervalos basados en el estimador estándar en términos de amplitud, especialmente para coeficientes V de Cramer elevados.

Cuando P_A =0.5 (Figura 4.2), los intervalos de confianza muestran un comportamiento similar, en cuanto a la amplitud media de los mismos, que los intervalos obtenidos cuando P_A =0.25. Las principales diferencias con respec-

Figura 4.3: Valores de Amplitud media y Cobertura (CI) de los intervalos de confianza (al 95 %) basados en los estimadores $\widehat{p}_A$ (Estándar), $\widehat{p}_{r.e}$ y $\widehat{p}_{mr.e}$ y obtenidos a partir de las poblaciones simuladas. Las muestras son seleccionadas con tamaño $n = 100$. P_A =0.25 y ϕ varía desde 0.5 a 0.9.

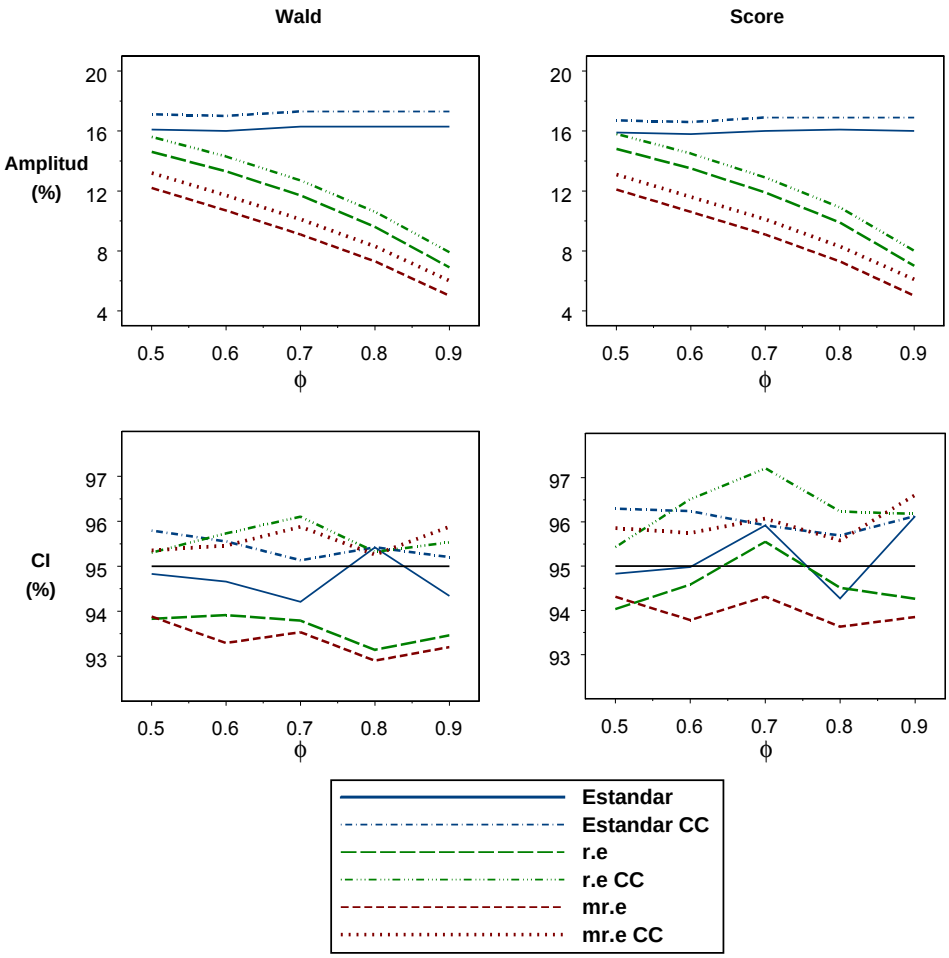
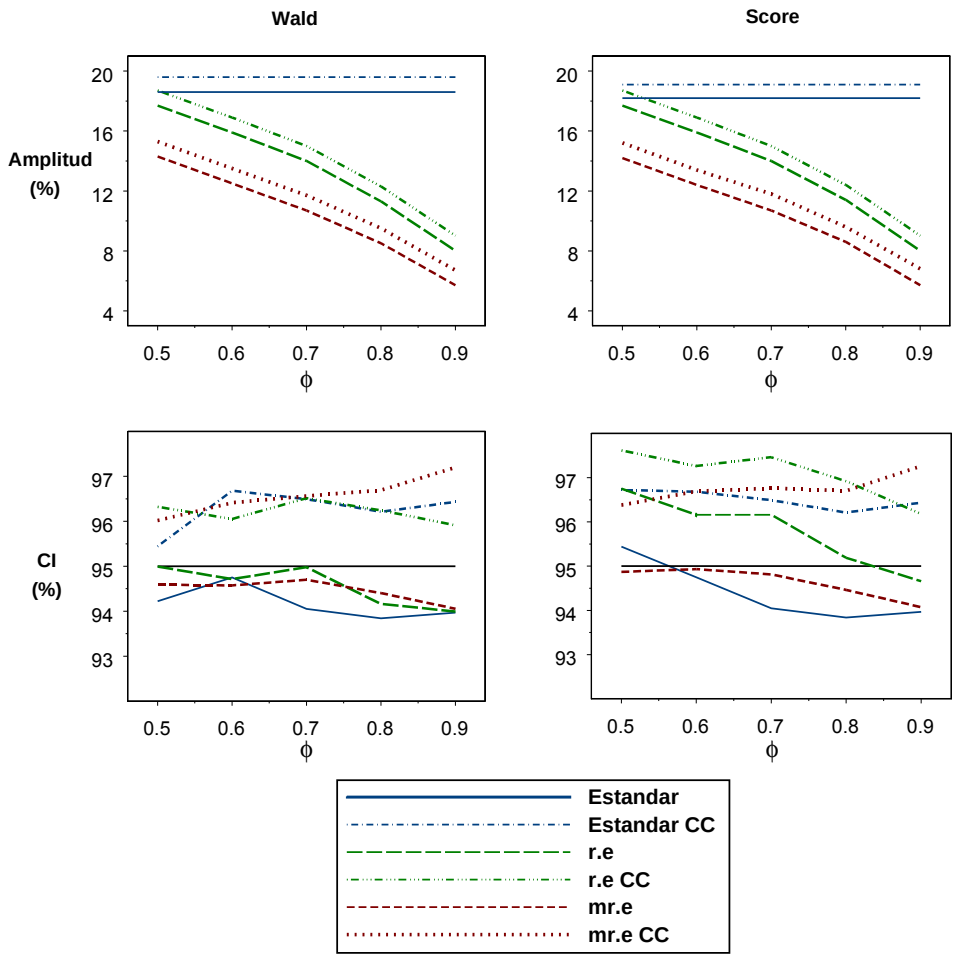


Figura 4.4: Valores de Amplitud media y Cobertura (CI) de los intervalos de confianza (al 95 %) basados en los estimadores $\widehat{p}_A$ (Estándar), $\widehat{p}_{r.e}$ y $\widehat{p}_{mr.e}$ y obtenidos a partir de las poblaciones simuladas. Las muestras son seleccionadas con tamaño $n = 100$. P_A =0.5 y ϕ varía desde 0.5 a 0.9.



to a la figura anterior se reflejan en lo que respecta a la cobertura. En este caso, los intervalos basados en los estimadores $\widehat{p}_{r.e}$ y $\widehat{p}_{reg}^{opt}$ son los que obtienen, en general, coberturas más cercanas al 95 % para los métodos Wald y Score respectivamente.

En las Figuras 4.3 y 4.4 se compara el comportamiento de los intervalos cuando se añade un atributo auxiliar en la etapa de estimación, es decir, se comparan los intervalos basados en el estimador $\widehat{p}_{r.e}$ con los intervalos basados en el estimador $\widehat{p}_{mr.e}$, el cual utiliza dos atributos auxiliares en la etapa de estimación.

Observamos que los intervalos basados en dos atributos auxiliares (los basados en el estimador $\widehat{p}_{mr.e}$) son claramente más estrechos que los intervalos basados en un único atributo auxiliar (los basados en el estimador $\widehat{p}_{r.e}$). Al igual que en las figuras anteriores, los intervalos basados en información auxiliar tienen considerablemente menor amplitud media a medida que aumenta el coeficiente V de Cramer.

Respecto a la cobertura de los intervalos de confianza, observamos para el caso P_A =0.25 (Figura 4.3) que los intervalos basados en $\widehat{p}_{mr.e}$ y que no usan CC son lo que aportan una menor cobertura, aunque esta situación se solventa cuando dichos intervalos incorporan la CC, obteniéndose intervalos en torno al 95.5 % y 96 % para los métodos Wald y Score respectivamente.

Cuando P_A =0.5 (Figura 4.4) y en el caso del método de Wald, los intervalos basados en los estimadores $\widehat{p}_{r.e}$ y $\widehat{p}_{mr.e}$ son los que obtienen coberturas más cercanas al 95 %. Por su parte, cuando se utiliza el método de Score, los intervalos basados en $\widehat{p}_{mr.e}$ son los más próximos al 95 %, aunque los intervalos basados en $\widehat{p}_{r.e}$ se aproximan a esta cantidad a medida que aumenta el coeficiente V de Cramer.

Poblaciones basadas en datos reales

En esta sección se muestran nuevos estudios de simulación para analizar el comportamiento de los diferentes intervalos de confianza. En este caso se consideran poblaciones basadas en datos reales.

Además de las coberturas empíricas y amplitudes medias utilizadas como medidas de comparación en los intervalos de confianza obtenidos para las poblaciones simuladas, en estos estudios de simulación se consideran otras medidas alternativas. En concreto, se ha calculado el porcentaje de casos en los

Tabla 4.1: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo y $P[L < 0]$ de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población EPF. P_A =0.194, P_B =0.173, ϕ =0.501 y muestras seleccionadas con tamaño $n = 50$.

	Método	Amplitud	CI	PD	PE	$P[L < 0]$
Wald	$\widehat{p}_A$	21.6	93.2	0.9	5.9	0.5
	$\widehat{p}_A$ CC	23.6	93.2	0.9	5.9	2.2
	$\widehat{p}_r$	22.8	91.2	4.4	4.4	0.6
	$\widehat{p}_r$ CC	24.8	93.4	3.6	3.0	1.6
	$\widehat{p}_{r.e}$	20.0	93.6	2.0	4.4	0.6
	$\widehat{p}_{r.e}$ CC	22.0	95.8	1.2	3.0	1.6
	$\widehat{p}_{r.opt}$	18.2	92.1	2.0	5.9	0.2
	$\widehat{p}_{r.opt}$ CC	20.2	94.8	1.2	4.0	0.6
	$\widehat{p}_d$	20.7	94.7	2.2	3.1	6.1
	$\widehat{p}_d$ CC	22.7	97.3	1.1	1.6	7.4
Score	$\widehat{p}_A$	21.1	95.5	2.3	2.2	0.0
	$\widehat{p}_A$ CC	22.9	97.1	2.3	0.6	0.0
	$\widehat{p}_r$	22.9	91.2	7.7	1.1	0.0
	$\widehat{p}_r$ CC	24.8	92.9	6.6	0.5	0.0
	$\widehat{p}_{r.e}$	23.2	96.4	2.5	1.1	0.0
	$\widehat{p}_{r.e}$ CC	24.3	97.9	1.6	0.5	0.0
	$\widehat{p}_{r.opt}$	17.9	93.7	3.4	2.9	0.0
	$\widehat{p}_{r.opt}$ CC	19.7	96.3	2.2	1.5	0.0
	$\widehat{p}_d$	20.8	94.2	2.8	2.9	0.0
	$\widehat{p}_d$ CC	22.8	96.4	1.8	1.8	0.0
Otros	Binomial	23.1	95.5	2.3	2.2	0.0
	Binomial Mid	21.6	95.5	2.3	2.2	0.0
	Verosimilitud	21.2	95.5	2.3	2.2	0.0

Tabla 4.2: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo y $P[L < 0]$ de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población EPF. P_A =0.194, P_B =0.173, ϕ =0.501 y muestras seleccionadas con tamaño $n = 100$.

	Método	Amplitud	CI	PD	PE	$P[L < 0]$
Wald	$\widehat{p}_A$	15.2	95.7	1.2	3.1	0.0
	$\widehat{p}_A$ CC	16.2	95.7	1.2	3.1	0.0
	$\widehat{p}_r$	15.9	93.0	3.7	3.3	0.0
	$\widehat{p}_r$ CC	16.9	94.6	3.0	2.4	0.0
	$\widehat{p}_{r.e}$	14.4	94.8	1.9	3.3	0.0
	$\widehat{p}_{r.e}$ CC	15.4	96.2	1.3	2.4	0.0
	$\widehat{p}_{r.opt}$	13.0	94.3	1.6	4.2	0.0
	$\widehat{p}_{r.opt}$ CC	14.0	95.7	1.2	3.1	0.0
	$\widehat{p}_d$	14.8	94.8	2.2	3.1	0.1
	$\widehat{p}_d$ CC	15.8	96.7	1.5	1.8	0.2
Score	$\widehat{p}_A$	15.0	96.3	2.2	1.5	0.0
	$\widehat{p}_A$ CC	16.0	96.3	2.2	1.5	0.0
	$\widehat{p}_r$	18.0	92.9	6.1	1.0	0.0
	$\widehat{p}_r$ CC	19.1	94.3	5.1	0.6	0.0
	$\widehat{p}_{r.e}$	14.8	97.0	2.0	1.0	0.0
	$\widehat{p}_{r.e}$ CC	15.9	98.0	1.4	0.6	0.0
	$\widehat{p}_{r.opt}$	12.8	94.8	2.9	2.3	0.0
	$\widehat{p}_{r.opt}$ CC	13.8	96.6	1.9	1.5	0.0
	$\widehat{p}_d$	14.7	94.8	2.3	2.8	0.0
	$\widehat{p}_d$ CC	15.7	96.3	1.7	2.0	0.0
Otros	Binomial	16.2	96.3	2.2	1.5	0.0
	Binomial Mid	15.4	96.3	2.2	1.5	0.0
	Verosimilitud	15.2	94.7	2.2	3.1	0.0

que el parámetro poblacional cae por debajo (PD) y por encima (PE) del intervalo de confianza. Además se ha calculado el porcentaje de veces que el límite inferior del intervalo de confianza (L) es menor que 0, es decir $P(L < 0)$ desde un punto de vista empírico. Destacamos que la situación idónea es que un intervalo tengo la menor amplitud y una cobertura del 95 % y el 5 % restante esté repartido equitativamente a ambos lados del intervalo, es decir, que tenga valores de PD y PE iguales a 2.5 % en ambos casos. Por último, sería deseable que $P(L < 0) = 0$, puesto que por definición una proporción poblacional debería encontrarse dentro del intervalo [0,1].

Las Tablas 4.1 (donde $n = 50$) y 4.2 (donde $n = 100$) muestran el comportamiento empírico de los distintos métodos para la construcción de intervalos de confianza para P_A en la población EPF. Este estudio confirma que los intervalos de confianza propuestos son más estrechos que los intervalos disponibles en la literatura anteriormente, especialmente los intervalos basados en los estimadores óptimos $\widehat{p}_{r.opt}$ y $\widehat{p}_{reg}^{opt}$.

Para $n = 50$ (Tabla 4.1) y el método de Wald, los distintos intervalos de confianza (con y sin CC) ofrecen, en general, coberturas ligeramente más pequeñas del 95 %. Sin embargo, estos intervalos tienen un comportamiento muy pobre en términos de PD y PE, puesto que dichos valores están muy distantes en la mayoría de las ocasiones, llegando en algunos casos a tomar valores de PD en torno al 1 % y valores de PE en torno al 6 %. Esta diferencia entre los valores PD y PE puede estar estrechamente relacionada con el hecho de que los extremos inferiores de los intervalos de confianza estén próximos a 0, puesto que para todos estos intervalos la $P(L < 0)$ no es 0. Atendiendo a esta última medida, los intervalos que tiene un peor comportamiento son los basados en el estimador de tipo diferencia, donde se alcanzan probabilidades superiores al 6 %.

Continuando con la Tabla 4.1 y observando el método Score (con y sin CC), observamos que las coberturas en su mayoría superan el 95 %. En lo que respecta al comportamiento en términos de PD y PE, observamos que en esta ocasión estas medidas están más equilibradas, a excepción del estimador de razón que tiene un comportamiento muy extremo. Este hecho puede estar relacionado con que la $P(L < 0) = 0$ con el método de Score.

El resto de métodos para la construcción de intervalos de confianza obtienen medidas bastantes aceptables, excepto en términos de amplitud, donde se obtienen intervalos más amplios que los propuestos.

Analizando esta población cuando aumentamos el tamaño de la muestra a

$n = 100$ (Tabla 4.2), podemos comprobar, como parece razonable, que todas las medidas mejoran para todos los intervalos de confianza. En primer lugar, los intervalos de confianza basados en el método de Wald tienen menor amplitud media y sus coberturas están más próximas al 95 %. Aunque los valores de PD y PE no se encuentren, en general, muy equilibrados, si es cierto que tiene un mejor comportamiento que cuando $n = 50$. En esta ocasión, tan sólo los intervalos basados en el estimador de tipo diferencia obtienen extremos inferiores por debajo 0, aunque este porcentaje es realmente bajo, inferior al 0.2 %.

En lo que respecta al método Score (con y sin CC), las coberturas en su mayoría superan el 95 %. Respecto al comportamiento en términos de PD y PE, los intervalos son mejores ya que en general están muy equilibrados, a excepción del estimador de razón que tiene un comportamiento muy extremo.

El método Score posee algunas ventajas con respecto al método Wald. Por ejemplo, con el método Score no se obtienen intervalos de confianza fuera del intervalo natural [0,1], y al mismo tiempo estos intervalos de confianza tienen un mejor comportamiento en términos de las medidas PD y PE. En lo que se refiere a la cobertura, el uso de CC proporciona coberturas cercanas al 95 %, y en su mayoría lo superan. Las medidas PD y PE están mejor equilibradas en los intervalos de confianza propuestos, mientras que tales medidas son claramente peores en los intervalos de confianza basados en el estimador estándar.

Los métodos Binomial y de Verosimilitud son ligeramente conservadores y ofrecen intervalos de confianza menos equilibrados en términos de PD y PE que los métodos propuestos basados en el método Score. Estos intervalos son generalmente más amplios que los intervalos de confianza propuestos.

Las Tablas 4.3 y 4.4 muestran el comportamiento de los distintos métodos para la construcción de intervalos de confianza para P_A en la población EES. Este estudio confirma que los intervalos de confianza propuestos son más estrechos que los intervalos disponibles en la literatura anteriormente, como se puede observar para los distintos tamaños de muestra estudiados.

Para $n = 50$ (Tabla 4.3) y usando el método de Wald, los distintos intervalos de confianza con CC ofrecen coberturas ligeramente mayores del 95 %. Por su parte, los intervalos que se han obtenido sin CC están todos, ligeramente, por debajo del 95 %. Estos intervalos tienen un comportamiento aceptable en términos de PD y PE. Mediante el uso del método Score (con y sin CC) las coberturas en su mayoría también están próximas al 95 %, y las medidas PD y PE también están aceptablemente equilibradas. Tanto para el método Wald

Tabla 4.3: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población EES. P_A =0.496, P_B =0.596, ϕ =0.467 y muestras seleccionadas con tamaño $n = 50$.

	Método	Amplitud	CI	PD	PE
Wald	$\widehat{p}_A$	27.0	94.2	2.5	3.3
	$\widehat{p}_A$ CC	29.0	95.9	2.5	1.6
	$\widehat{p}_r$	25.7	94.4	3.0	2.6
	$\widehat{p}_r$ CC	27.7	96.1	2.1	1.8
	$\widehat{p}_{r.e}$	25.6	94.4	3.0	2.6
	$\widehat{p}_{r.e}$ CC	27.6	96.2	2.0	1.8
	$\widehat{p}_{r.opt}$	23.6	94.2	3.1	2.7
	$\widehat{p}_{r.opt}$ CC	25.6	95.9	2.2	1.9
	$\widehat{p}_d$	27.5	94.4	2.9	2.7
	$\widehat{p}_d$ CC	29.5	96.1	2.0	1.9
Score	$\widehat{p}_A$	25.9	94.2	2.5	3.3
	$\widehat{p}_A$ CC	27.6	97.3	1.1	1.6
	$\widehat{p}_r$	25.7	94.0	4.3	1.7
	$\widehat{p}_r$ CC	27.7	95.6	3.4	1.0
	$\widehat{p}_{r.e}$	25.6	96.8	1.5	1.7
	$\widehat{p}_{r.e}$ CC	27.6	98.0	1.0	1.0
	$\widehat{p}_{r.opt}$	22.7	94.6	3.0	2.4
	$\widehat{p}_{r.opt}$ CC	24.6	96.3	2.0	1.7
	$\widehat{p}_d$	26.7	94.9	2.9	2.2
	$\widehat{p}_d$ CC	28.7	96.4	2.1	1.5
Otros	Binomial	28.7	97.3	1.1	1.6
	Binomial Mid	27.1	95.9	2.5	1.6
	Verosimilitud	26.8	95.9	2.5	1.6

Tabla 4.4: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población EES. P_A =0.496, P_B =0.596, ϕ =0.467 y muestras seleccionadas con tamaño $n = 100$.

	Método	Amplitud	CI	PD	PE
Wald	$\widehat{p}_A$	18.6	94.0	3.0	3.0
	$\widehat{p}_A$ CC	19.6	96.5	1.6	1.9
	$\widehat{p}_r$	17.7	94.5	2.8	2.7
	$\widehat{p}_r$ CC	18.7	95.8	2.0	2.2
	$\widehat{p}_{r.e}$	17.7	94.5	2.8	2.7
	$\widehat{p}_{r.e}$ CC	18.7	95.8	2.0	2.2
	$\widehat{p}_{r.opt}$	16.3	94.1	2.9	3.0
	$\widehat{p}_{r.opt}$ CC	17.3	95.7	2.1	2.2
	$\widehat{p}_d$	19.0	94.6	2.8	2.6
	$\widehat{p}_d$ CC	20.0	96.1	2.0	1.9
Score	$\widehat{p}_A$	18.2	94.0	3.0	3.0
	$\widehat{p}_A$ CC	19.1	96.5	1.6	1.9
	$\widehat{p}_r$	17.6	94.0	3.9	2.1
	$\widehat{p}_r$ CC	18.6	95.5	3.0	1.5
	$\widehat{p}_{r.e}$	17.6	94.2	3.7	2.1
	$\widehat{p}_{r.e}$ CC	18.6	95.9	2.6	1.5
	$\widehat{p}_{r.opt}$	16.0	94.2	3.0	2.8
	$\widehat{p}_{r.opt}$ CC	16.9	95.9	2.1	2.0
	$\widehat{p}_d$	18.6	94.7	2.8	2.5
	$\widehat{p}_d$ CC	19.6	95.8	2.4	1.8
Otros	Binomial	20.2	96.5	1.6	1.9
	Binomial Mid	19.4	96.5	1.6	1.9
	Verosimilitud	19.2	96.5	1.6	1.9

como para el método Score, los intervalos que obtienen menor amplitud media son los basados en el estimador de razón óptimo. En este caso era razonable no obtener intervalos con extremos inferiores por debajo de 0, puesto que la proporción poblacional de interés está próxima 0.5.

Cuando $n = 100$ (Tabla 4.4) y para el método de Wald, los distintos intervalos de confianza ofrecen coberturas en torno al 95 %. Se ha de destacar que estos intervalos tienen un mejor comportamiento en términos de PD y PE que los intervalos obtenidos en el caso $n = 50$, puesto que dichos valores están muy equilibrados, siendo la diferencia máxima entre PD y PE de tan solo 0.3 %. Cuando se utiliza el método Score, las coberturas también están en torno al 95 %, mientras que las medidas PD y PE son peores en comparación con las obtenidas con el método de Wald, ya que la mayoría están menos equilibradas y los valores de PD son mayores que los valores de PE.

En general, en esta población observamos una mejoría en las distintas medidas a medida que aumenta el tamaño muestral, y constatamos que de nuevo los intervalos que proponemos son generalmente mejores (en términos de las distintas medidas consideradas para la comparación) que los intervalos disponibles en la literatura hasta el momento.

También se han llevado a cabo estudios de simulación para la comparación de intervalos de confianza en la población Lagos, cuyos resultados pueden consultarse de la Tabla 4.5 a la Tabla **??**. Las conclusiones que se derivan de estos resultados son muy similares a las conclusiones ya comentadas, por lo se destacará únicamente las observaciones más relevantes.

En las Tablas 4.5 y 4.6 observamos que el intervalo con menor amplitud es el basado en el estimador $\widehat{p}_{r.opt}$, y esta amplitud (en torno al 11 %) es claramente menor en comparación con la amplitud de otros métodos (en torno al 27 %) que no utilizan información auxiliar en la etapa de estimación. Este hecho se debe a que el coeficiente V de Cramer es bastante elevado en esta situación. Para los intervalos basados en el estimador $\widehat{p}_{r.opt}$ y cuando $n = 50$, parece preferible el método Score sin CC, puesto que el método Wald sin CC obtiene coberturas muy pequeñas y valores de PD y PE muy descompensados, mientras que el método Score con CC obtiene una cobertura muy alejada del 95 %. Cuando aumenta el tamaño de la muestra ($n = 100$), el método Wald si muestra un comportamiento similar al método Score, por lo que cualquiera de ellos podría aplicarse en la práctica.

Cuando disminuye el coeficiente V de Cramer (Tablas 4.7 y 4.8), observamos que la diferencias de las amplitudes entre los intervalos basados en el

Tabla 4.5: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población Lagos. P_A =0.438, P_B =0.44, ϕ =0.9 y muestras seleccionadas con tamaño $n = 50$.

	Método	Amplitud	CI	PD	PE
Wald	$\widehat{p}_A$	26.3	94.8	2.4	2.8
	$\widehat{p}_A$ CC	28.3	94.8	2.4	2.8
	$\widehat{p}_r$	11.3	89.7	8.5	1.8
	$\widehat{p}_r$ CC	13.3	98.6	0.8	0.6
	$\widehat{p}_{r.e}$	11.0	90.3	7.9	1.8
	$\widehat{p}_{r.e}$ CC	13.0	98.9	0.5	0.6
	$\widehat{p}_{r.opt}$	10.9	90.1	8.0	1.9
	$\widehat{p}_{r.opt}$ CC	12.9	98.8	0.5	0.7
	$\widehat{p}_d$	11.2	91.5	7.3	1.2
	$\widehat{p}_d$ CC	13.2	99.6	0.2	0.2
Score	$\widehat{p}_A$	25.2	94.8	2.4	2.8
	$\widehat{p}_A$ CC	26.9	96.3	2.5	1.2
	$\widehat{p}_r$	12.1	82.4	9.8	7.8
	$\widehat{p}_r$ CC	13.9	98.3	1.4	0.3
	$\widehat{p}_{r.e}$	11.4	84.4	7.8	7.8
	$\widehat{p}_{r.e}$ CC	13.4	99.4	0.3	0.3
	$\widehat{p}_{r.opt}$	11.4	96.6	1.7	1.6
	$\widehat{p}_{r.opt}$ CC	12.6	99.0	0.5	0.5
	$\widehat{p}_d$	11.9	97.9	0.8	1.3
	$\widehat{p}_d$ CC	13.1	99.4	0.4	0.2
Otros	Binomial	28.5	97.6	1.2	1.2
	Binomial Mid	26.9	94.8	2.4	2.8
	Verosimilitud	26.6	94.8	2.4	2.8

Tabla 4.6: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población Lagos. P_A =0.438, P_B =0.44, ϕ =0.9 y muestras seleccionadas con tamaño $n = 100$.

	Método	Amplitud	CI	PD	PE
Wald	$\widehat{p}_A$	17.6	94.3	2.6	3.2
	$\widehat{p}_A$ CC	18.6	96.6	1.5	1.9
	$\widehat{p}_r$	8.0	94.4	3.2	2.4
	$\widehat{p}_r$ CC	9.0	97.4	1.5	1.1
	$\widehat{p}_{r.e}$	7.8	94.7	2.9	2.4
	$\widehat{p}_{r.e}$ CC	8.8	97.7	1.1	1.2
	$\widehat{p}_{r.opt}$	7.7	94.4	3.0	2.6
	$\widehat{p}_{r.opt}$ CC	8.7	97.6	1.2	1.2
	$\widehat{p}_d$	7.9	94.5	3.1	2.4
	$\widehat{p}_d$ CC	8.9	97.6	1.6	0.8
Score	$\widehat{p}_A$	17.3	95.5	2.6	1.9
	$\widehat{p}_A$ CC	18.1	96.6	1.5	1.9
	$\widehat{p}_r$	8.0	93.5	3.9	2.6
	$\widehat{p}_r$ CC	9.0	97.0	2.0	1.0
	$\widehat{p}_{r.e}$	7.9	94.8	2.6	2.6
	$\widehat{p}_{r.e}$ CC	8.9	97.9	1.0	1.0
	$\widehat{p}_{r.opt}$	7.6	94.4	2.8	2.8
	$\widehat{p}_{r.opt}$ CC	8.6	97.5	1.2	1.3
	$\widehat{p}_d$	7.8	94.8	2.8	2.4
	$\widehat{p}_d$ CC	8.8	97.7	1.5	0.8
Otros	Binomial	20.1	98.1	0.8	1.1
	Binomial Mid	19.2	96.6	1.5	1.9
	Verosimilitud	19.1	96.6	1.5	1.9

Tabla 4.7: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población Lagos. P_A =0.438, P_B =0.163, ϕ =0.5 y muestras seleccionadas con tamaño $n = 50$.

	Método	Amplitud	CI	PD	PE
Wald	$\widehat{p}_A$	26.3	95.0	2.5	2.5
	$\widehat{p}_A$ CC	28.3	95.0	2.5	2.5
	$\widehat{p}_r$	45.6	92.4	4.5	3.1
	$\widehat{p}_r$ CC	47.5	93.2	4.4	2.4
	$\widehat{p}_{r.e}$	22.7	94.1	2.2	3.7
	$\widehat{p}_{r.e}$ CC	24.7	95.8	1.4	2.8
	$\widehat{p}_{r.opt}$	22.7	94.1	2.2	3.7
	$\widehat{p}_{r.opt}$ CC	24.7	95.8	1.4	2.8
	$\widehat{p}_d$	23.6	94.9	1.6	3.5
	$\widehat{p}_d$ CC	25.6	94.9	1.6	3.5
Score	$\widehat{p}_A$	25.2	95.0	2.5	2.5
	$\widehat{p}_A$ CC	26.9	96.5	2.4	1.1
	$\widehat{p}_r$	48.3	91.0	8.7	0.3
	$\widehat{p}_r$ CC	50.4	92.1	7.7	0.2
	$\widehat{p}_{r.e}$	22.1	94.6	2.2	3.2
	$\widehat{p}_{r.e}$ CC	24.0	96.2	1.3	2.5
	$\widehat{p}_{r.opt}$	21.9	94.3	2.4	3.3
	$\widehat{p}_{r.opt}$ CC	23.8	95.7	1.5	2.8
	$\widehat{p}_d$	22.9	95.0	1.6	3.4
	$\widehat{p}_d$ CC	24.8	95.4	1.3	3.3
Otros	Binomial	28.5	97.7	1.2	1.1
	Binomial Mid	26.9	95.0	2.5	2.5
	Verosimilitud	26.7	95.0	2.5	2.5

Tabla 4.8: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población Lagos. P_A =0.438, P_B =0.163, ϕ =0.5 y muestras seleccionadas con tamaño $n = 100$.

	Método	Amplitud	CI	PD	PE
Wald	$\widehat{p}_A$	17.6	94.3	2.7	3.0
	$\widehat{p}_A$ CC	18.6	96.7	1.6	1.7
	$\widehat{p}_r$	31.1	93.6	3.9	2.5
	$\widehat{p}_r$ CC	32.1	94.5	3.5	2.0
	$\widehat{p}_{r.e}$	15.2	94.3	2.3	3.4
	$\widehat{p}_{r.e}$ CC	16.2	95.7	1.7	2.6
	$\widehat{p}_{r.opt}$	15.2	94.3	2.3	3.4
	$\widehat{p}_{r.opt}$ CC	16.2	95.7	1.7	2.6
	$\widehat{p}_d$	15.9	94.6	1.5	3.9
	$\widehat{p}_d$ CC	16.9	96.3	1.5	2.2
Score	$\widehat{p}_A$	17.3	95.6	2.7	1.7
	$\widehat{p}_A$ CC	18.1	96.7	1.6	1.7
	$\widehat{p}_r$	36.9	92.7	6.6	0.7
	$\widehat{p}_r$ CC	38.1	93.5	6.0	0.5
	$\widehat{p}_{r.e}$	15.0	94.4	2.3	3.3
	$\widehat{p}_{r.e}$ CC	16.0	96.0	1.7	2.3
	$\widehat{p}_{r.opt}$	14.9	94.4	2.3	3.3
	$\widehat{p}_{r.opt}$ CC	15.9	95.9	1.7	2.4
	$\widehat{p}_d$	15.6	94.1	2.1	3.8
	$\widehat{p}_d$ CC	16.6	96.1	1.5	2.4
Otros	Binomial	20.1	98.1	0.9	1.0
	Binomial Mid	19.2	96.7	1.6	1.7
	Verosimilitud	19.1	96.7	1.6	1.7

Tabla 4.9: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población Lagos. P_A =0.215, P_B =0.215, ϕ =0.9 y muestras seleccionadas con tamaño $n = 50$.

	Método	Amplitud	CI	PD	PE
Wald	$\widehat{p}_A$	21.6	93.7	0.8	5.5
	$\widehat{p}_A$ CC	23.6	96.8	0.8	0.4
	$\widehat{p}_r$	7.0	95.9	2.2	1.9
	$\widehat{p}_r$ CC	9.0	98.6	1.1	0.3
	$\widehat{p}_{r.e}$	6.8	97.9	0.1	2.0
	$\widehat{p}_{r.e}$ CC	8.8	99.7	0.0	0.3
	$\widehat{p}_{r.opt}$	6.7	97.7	0.3	2.0
	$\widehat{p}_{r.opt}$ CC	8.7	99.7	0.0	0.3
	$\widehat{p}_d$	6.9	99.7	0.2	0.1
	$\widehat{p}_d$ CC	8.9	100.0	0.0	0.0
Score	$\widehat{p}_A$	21.1	95.5	2.1	2.4
	$\widehat{p}_A$ CC	22.8	97.1	2.1	0.8
	$\widehat{p}_r$	8.1	43.8	29.1	27.1
	$\widehat{p}_r$ CC	10.1	99.0	1.0	0.0
	$\widehat{p}_{r.e}$	7.5	45.5	27.4	27.1
	$\widehat{p}_{r.e}$ CC	9.5	99.9	0.1	0.0
	$\widehat{p}_{r.opt}$	6.6	98.4	0.1	0.6
	$\widehat{p}_{r.opt}$ CC	8.6	99.9	0.1	0.0
	$\widehat{p}_d$	7.1	99.0	0.4	0.6
	$\widehat{p}_d$ CC	9.1	99.9	0.0	0.1
Otros	Binomial	23.9	98.4	0.9	0.7
	Binomial Mid	22.4	95.5	2.1	2.4
	Verosimilitud	22.1	95.5	2.1	2.4

Tabla 4.10: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población Lagos. P_A =0.215, P_B =0.215, ϕ =0.9 y muestras seleccionadas con tamaño $n = 100$.

	Método	Amplitud	CI	PD	PE
Wald	$\widehat{p}_A$	14.6	93.5	1.7	4.8
	$\widehat{p}_A$ CC	15.6	96.7	0.9	2.4
	$\widehat{p}_r$	5.3	95.1	2.2	2.7
	$\widehat{p}_r$ CC	6.3	98.5	0.6	0.9
	$\widehat{p}_{r.e}$	5.2	96.3	0.9	2.7
	$\widehat{p}_{r.e}$ CC	6.2	99.0	0.1	0.9
	$\widehat{p}_{r.opt}$	5.2	96.3	0.9	2.8
	$\widehat{p}_{r.opt}$ CC	6.2	99.0	0.1	0.9
	$\widehat{p}_d$	5.3	98.8	0.4	0.8
	$\widehat{p}_d$ CC	6.3	98.8	0.4	0.8
Score	$\widehat{p}_A$	14.4	94.5	3.1	2.4
	$\widehat{p}_A$ CC	15.2	97.0	1.7	1.3
	$\widehat{p}_r$	5.7	83.2	9.1	7.7
	$\widehat{p}_r$ CC	6.5	98.6	1.2	0.2
	$\widehat{p}_{r.e}$	5.5	84.6	7.7	7.7
	$\widehat{p}_{r.e}$ CC	6.3	99.4	0.4	0.2
	$\widehat{p}_{r.opt}$	5.4	96.0	2.2	1.8
	$\widehat{p}_{r.opt}$ CC	6.1	99.2	0.5	0.3
	$\widehat{p}_d$	5.6	95.9	1.8	2.3
	$\widehat{p}_d$ CC	6.3	99.1	0.3	0.6
Otros	Binomial	16.8	97.8	0.9	1.3
	Binomial Mid	16.0	97.0	1.7	1.3
	Verosimilitud	15.8	97.0	1.7	1.3

Tabla 4.11: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población Lagos. P_A =0.215, P_B =0.522, ϕ =0.5 y muestras seleccionadas con tamaño $n = 50$.

	Método	Amplitud	CI	PD	PE
Wald	$\widehat{p}_A$	21.6	93.0	1.0	6.0
	$\widehat{p}_A$ CC	23.6	96.3	1.0	2.7
	$\widehat{p}_r$	18.6	92.9	2.7	4.4
	$\widehat{p}_r$ CC	20.6	94.9	1.9	3.2
	$\widehat{p}_{r.e}$	18.6	92.9	2.7	4.4
	$\widehat{p}_{r.e}$ CC	20.6	94.9	1.9	3.2
	$\widehat{p}_{r.opt}$	18.6	92.9	2.7	4.4
	$\widehat{p}_{r.opt}$ CC	20.6	94.9	1.9	3.2
	$\widehat{p}_d$	24.1	92.1	5.4	2.5
	$\widehat{p}_d$ CC	26.0	96.1	2.7	1.2
Score	$\widehat{p}_A$	21.1	95.0	2.3	2.7
	$\widehat{p}_A$ CC	22.7	96.8	2.3	0.9
	$\widehat{p}_r$	18.4	94.3	4.0	1.7
	$\widehat{p}_r$ CC	20.4	96.2	2.7	1.1
	$\widehat{p}_{r.e}$	18.4	94.3	4.0	1.7
	$\widehat{p}_{r.e}$ CC	20.4	96.2	2.7	1.1
	$\widehat{p}_{r.opt}$	18.2	94.0	4.1	1.9
	$\widehat{p}_{r.opt}$ CC	20.0	96.0	2.8	1.2
	$\widehat{p}_d$	24.1	94.1	4.7	1.2
	$\widehat{p}_d$ CC	26.0	96.2	2.7	1.1
Otros	Binomial	23.9	98.1	1.0	0.9
	Binomial Mid	22.3	95.0	2.3	2.7
	Verosimilitud	22.0	95.0	2.3	2.7

Tabla 4.12: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población Lagos. P_A =0.215, P_B =0.522, ϕ =0.5 y muestras seleccionadas con tamaño $n = 100$.

	Método	Amplitud	CI	PD	PE
Wald	$\widehat{p}_A$	14.6	93.3	1.7	5.0
	$\widehat{p}_A$ CC	15.6	96.5	0.8	2.7
	$\widehat{p}_r$	12.6	93.9	2.6	3.5
	$\widehat{p}_r$ CC	13.6	95.4	2.0	2.6
	$\widehat{p}_{r.e}$	12.6	93.9	2.6	3.5
	$\widehat{p}_{r.e}$ CC	13.6	95.4	2.0	2.6
	$\widehat{p}_{r.opt}$	12.6	93.9	2.6	3.5
	$\widehat{p}_{r.opt}$ CC	13.6	95.4	2.0	2.6
	$\widehat{p}_d$	16.4	94.2	4.0	1.8
	$\widehat{p}_d$ CC	17.4	95.7	2.5	1.8
Score	$\widehat{p}_A$	14.4	94.1	3.2	2.7
	$\widehat{p}_A$ CC	15.2	96.8	1.7	1.5
	$\widehat{p}_r$	12.4	94.4	3.5	2.1
	$\widehat{p}_r$ CC	13.4	95.9	2.6	1.4
	$\widehat{p}_{r.e}$	12.4	94.4	3.5	2.1
	$\widehat{p}_{r.e}$ CC	13.4	95.9	2.7	1.4
	$\widehat{p}_{r.opt}$	12.4	94.3	3.5	2.2
	$\widehat{p}_{r.opt}$ CC	13.3	95.7	2.7	1.6
	$\widehat{p}_d$	16.2	94.2	4.0	1.8
	$\widehat{p}_d$ CC	17.2	96.3	2.6	1.1
Otros	Binomial	16.8	97.7	0.8	1.5
	Binomial Mid	16.0	96.8	1.7	1.5
	Verosimilitud	15.8	96.8	1.7	1.5

Tabla 4.13: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo y $P[L < 0]$ de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población Lagos. P_A =0.07, P_B =0.176, ϕ =0.5 y muestras seleccionadas con tamaño $n = 50$.

	Método	Amplitud	CI	PD	PE	$P[L < 0]$
Wald	$\widehat{p}_A$	12.6	88.5	0.4	11.1	50.3
	$\widehat{p}_A$ CC	14.0	88.8	0.1	11.1	73.4
	$\widehat{p}_r$	10.8	85.8	4.7	9.5	31.8
	$\widehat{p}_r$ CC	12.4	91.1	3.3	5.6	49.0
	$\widehat{p}_{r.e}$	10.8	85.8	4.7	9.5	31.8
	$\widehat{p}_{r.e}$ CC	12.4	91.1	3.3	5.6	49.0
	$\widehat{p}_{r.opt}$	10.6	86.0	3.8	10.2	30.9
	$\widehat{p}_{r.opt}$ CC	12.2	90.8	2.6	6.6	47.9
	$\widehat{p}_d$	14.2	92.3	6.4	1.3	64.3
	$\widehat{p}_d$ CC	15.6	93.6	5.8	0.6	70.1
Score	$\widehat{p}_A$	13.9	92.9	5.1	2.0	0.0
	$\widehat{p}_A$ CC	15.6	98.4	1.6	0.0	0.0
	$\widehat{p}_r$	14.1	92.3	7.5	0.2	0.0
	$\widehat{p}_r$ CC	15.7	95.3	4.7	0.0	0.0
	$\widehat{p}_{r.e}$	14.1	92.3	7.5	0.2	0.0
	$\widehat{p}_{r.e}$ CC	15.7	95.3	4.7	0.0	0.0
	$\widehat{p}_{r.opt}$	11.5	92.0	6.0	2.0	0.0
	$\widehat{p}_{r.opt}$ CC	13.3	96.1	3.9	0.0	0.0
	$\widehat{p}_d$	15.8	93.5	4.2	2.3	0.0
	$\widehat{p}_d$ CC	17.1	96.1	2.7	1.2	0.0
Otros	Binomial	15.7	98.4	1.6	0.0	0.0
	Binomial Mid	14.4	98.4	1.6	0.0	0.0
	Verosimilitud	13.9	98.4	1.6	0.0	0.0

Tabla 4.14: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo y $P[L < 0]$ de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población Lagos. P_A =0.07, P_B =0.176, ϕ =0.5 y muestras seleccionadas con tamaño $n = 100$.

	Método	Amplitud	CI	PD	PE	$P[L < 0]$
Wald	$\widehat{p}_A$	9.0	94.3	0.4	5.3	5.2
	$\widehat{p}_A$ CC	9.9	94.3	0.4	5.3	5.2
	$\widehat{p}_r$	7.7	91.7	3.2	5.1	1.6
	$\widehat{p}_r$ CC	8.7	94.6	1.9	3.5	4.0
	$\widehat{p}_{r.e}$	7.7	91.7	3.2	5.1	1.6
	$\widehat{p}_{r.e}$ CC	8.7	94.6	1.9	3.5	4.0
	$\widehat{p}_{r.opt}$	7.6	92.3	2.3	5.4	1.5
	$\widehat{p}_{r.opt}$ CC	8.6	94.6	1.6	3.8	3.8
	$\widehat{p}_d$	10.9	93.9	4.5	1.6	40.0
	$\widehat{p}_d$ CC	11.7	95.7	3.0	1.3	45.0
Score	$\widehat{p}_A$	9.2	95.0	3.3	1.7	0.0
	$\widehat{p}_A$ CC	10.0	95.0	3.3	1.7	0.0
	$\widehat{p}_r$	8.2	93.6	5.1	1.3	0.0
	$\widehat{p}_r$ CC	9.2	96.1	3.4	0.5	0.0
	$\widehat{p}_{r.e}$	8.2	93.6	5.1	1.3	0.0
	$\widehat{p}_{r.e}$ CC	9.2	96.1	3.4	0.5	0.0
	$\widehat{p}_{r.opt}$	7.7	94.0	4.3	1.7	0.0
	$\widehat{p}_{r.opt}$ CC	8.7	96.2	3.0	0.7	0.0
	$\widehat{p}_d$	11.5	94.0	3.2	2.8	0.0
	$\widehat{p}_d$ CC	12.3	96.0	2.3	1.7	0.0
Otros	Binomial	10.9	98.3	1.3	0.4	0.0
	Binomial Mid	10.1	97.0	1.3	1.7	0.0
	Verosimilitud	9.8	97.0	1.3	1.7	0.0

Tabla 4.15: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo y $P[L < 0]$ de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población ENS. P_A =0.07, P_B =0.03, ϕ =0.583 y muestras seleccionadas con tamaño $n = 50$.

	Método	Amplitud	CI	PD	PE	$P[L < 0]$
Wald	$\widehat{p}_A$	14.1	93.3	1.2	5.5	43.3
	$\widehat{p}_A$ CC	15.6	94.2	0.3	5.5	65.3
	$\widehat{p}_r$	15.7	89.3	0.0	10.7	80.3
	$\widehat{p}_r$ CC	16.9	89.3	0.0	10.7	90.6
	$\widehat{p}_{r.e}$	10.5	88.9	0.4	10.7	9.4
	$\widehat{p}_{r.e}$ CC	12.4	89.1	0.2	10.7	10.6
	$\widehat{p}_{r.opt}$	10.4	88.3	0.4	11.3	7.5
	$\widehat{p}_{r.opt}$ CC	12.4	88.5	0.2	11.3	9.6
	$\widehat{p}_d$	10.5	88.7	0.6	10.7	9.7
	$\widehat{p}_d$ CC	12.4	89.1	0.2	10.7	10.6
Score	$\widehat{p}_A$	15.3	96.4	3.3	0.3	0.0
	$\widehat{p}_A$ CC	17.1	96.7	3.3	0.0	0.0
	$\widehat{p}_r$	96.5	99.0	1.0	0.0	0.0
	$\widehat{p}_r$ CC	97.5	99.0	1.0	0.0	0.0
	$\widehat{p}_{r.e}$	23.7	97.2	2.8	0.0	0.0
	$\widehat{p}_{r.e}$ CC	26.4	98.8	1.2	0.0	0.0
	$\widehat{p}_{r.opt}$	11.5	95.4	3.2	1.4	0.0
	$\widehat{p}_{r.opt}$ CC	12.3	96.1	2.7	1.2	0.0
	$\widehat{p}_d$	10.9	86.8	2.7	10.5	0.0
	$\widehat{p}_d$ CC	13.5	88.4	1.2	10.4	0.0
Otros	Binomial	16.5	98.5	1.2	0.3	0.0
	Binomial Mid	15.1	96.7	3.3	0.0	0.0
	Verosimilitud	14.5	96.7	3.3	0.0	0.0

Tabla 4.16: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo y $P[L < 0]$ de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población ENS. P_A =0.07, P_B =0.03, ϕ =0.583 y muestras seleccionadas con tamaño $n = 100$.

	Método	Amplitud	CI	PD	PE	$P[L < 0]$
Wald	$\widehat{p}_A$	10.0	94.1	1.1	4.8	4.8
	$\widehat{p}_A$ CC	11.0	94.7	0.5	4.8	12.4
	$\widehat{p}_r$	13.9	89.7	4.5	5.8	23.2
	$\widehat{p}_r$ CC	14.7	91.4	4.5	4.1	37.5
	$\widehat{p}_{r.e}$	8.0	93.6	0.5	5.9	0.6
	$\widehat{p}_{r.e}$ CC	9.0	93.9	0.4	5.7	1.7
	$\widehat{p}_{r.opt}$	7.9	92.6	0.6	6.8	0.6
	$\widehat{p}_{r.opt}$ CC	8.9	93.2	0.4	6.4	1.7
	$\widehat{p}_d$	8.0	93.6	0.5	5.9	0.7
	$\widehat{p}_d$ CC	9.0	93.9	0.5	5.6	1.8
Score	$\widehat{p}_A$	10.3	95.6	2.8	1.6	0.0
	$\widehat{p}_A$ CC	11.3	96.9	2.8	0.3	0.0
	$\widehat{p}_r$	95.2	87.1	12.9	0.0	0.0
	$\widehat{p}_r$ CC	95.9	91.4	8.6	0.0	0.0
	$\widehat{p}_{r.e}$	9.9	93.4	3.5	3.1	0.0
	$\widehat{p}_{r.e}$ CC	11.3	97.0	1.7	1.3	0.0
	$\widehat{p}_{r.opt}$	8.1	93.0	2.8	4.2	0.0
	$\widehat{p}_{r.opt}$ CC	9.0	94.5	1.7	3.8	0.0
	$\widehat{p}_d$	8.3	92.6	2.5	4.9	0.0
	$\widehat{p}_d$ CC	9.3	95.7	1.6	2.7	0.0
Otros	Binomial	11.1	97.3	1.1	1.6	0.0
	Binomial Mid	10.3	95.6	2.8	1.6	0.0
	Verosimilitud	10.0	95.7	2.7	1.6	0.0

Tabla 4.17: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo y $P[L < 0]$ de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población ENS. P_A =0.12, P_B =0.04, ϕ =0.51 y muestras seleccionadas con tamaño $n = 50$.

	Método	Amplitud	CI	PD	PE	$P[L < 0]$
Wald	$\widehat{p}_A$	17.7	94.9	1.2	3.9	11.9
	$\widehat{p}_A$ CC	19.5	95.6	0.4	4.0	25.7
	$\widehat{p}_r$	25.6	92.8	0.0	7.2	59.8
	$\widehat{p}_r$ CC	26.9	94.3	0.0	5.7	72.5
	$\widehat{p}_{r.e}$	14.8	90.2	0.9	8.9	8.1
	$\widehat{p}_{r.e}$ CC	16.7	90.8	0.4	8.8	13.2
	$\widehat{p}_{r.opt}$	14.6	89.7	1.0	9.3	3.6
	$\widehat{p}_{r.opt}$ CC	16.5	90.6	0.3	9.1	8.9
	$\widehat{p}_d$	14.9	90.0	1.1	8.9	8.2
	$\widehat{p}_d$ CC	16.8	90.8	0.3	8.9	13.4
Score	$\widehat{p}_A$	17.9	96.4	2.8	0.8	0.0
	$\widehat{p}_A$ CC	19.7	97.1	2.8	0.1	0.0
	$\widehat{p}_r$	93.7	89.7	10.3	0.0	0.0
	$\widehat{p}_r$ CC	94.5	94.8	5.2	0.0	0.0
	$\widehat{p}_{r.e}$	18.8	94.4	3.8	1.8	0.0
	$\widehat{p}_{r.e}$ CC	21.2	98.3	1.7	0.0	0.0
	$\widehat{p}_{r.opt}$	14.6	92.5	2.7	4.8	0.0
	$\widehat{p}_{r.opt}$ CC	16.3	95.6	1.9	2.5	0.0
	$\widehat{p}_d$	15.1	92.7	2.3	5.0	0.0
	$\widehat{p}_d$ CC	17.2	96.3	1.3	2.4	0.0
Otros	Binomial	19.4	98.0	1.2	0.8	0.0
	Binomial Mid	18.0	96.4	2.8	0.8	0.0
	Verosimilitud	17.5	96.5	2.8	0.7	0.0

Tabla 4.18: Valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo y $P[L < 0]$ de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población ENS. P_A =0.12, P_B =0.04, ϕ =0.51 y muestras seleccionadas con tamaño $n = 100$.

	Método	Amplitud	CI	PD	PE	$P[L < 0]$
Wald	$\widehat{p}_A$	12.4	94.7	1.0	4.3	0.2
	$\widehat{p}_A$ CC	13.4	94.7	1.0	4.3	0.7
	$\widehat{p}_r$	19.6	89.1	6.8	4.1	5.5
	$\widehat{p}_r$ CC	20.6	91.0	6.0	3.0	9.5
	$\widehat{p}_{r.e}$	10.7	93.9	1.0	5.1	0.1
	$\widehat{p}_{r.e}$ CC	11.7	94.8	0.6	4.6	0.3
	$\widehat{p}_{r.opt}$	10.5	92.7	1.0	6.3	0.0
	$\widehat{p}_{r.opt}$ CC	11.5	93.9	0.6	5.5	0.1
	$\widehat{p}_d$	10.7	94.2	0.8	5.0	0.1
	$\widehat{p}_d$ CC	11.7	94.8	0.6	4.6	0.4
Score	$\widehat{p}_A$	12.5	94.2	4.0	1.8	0.0
	$\widehat{p}_A$ CC	13.4	97.0	2.3	0.7	0.0
	$\widehat{p}_r$	78.5	87.6	12.1	0.3	0.0
	$\widehat{p}_r$ CC	80.1	87.8	12.1	0.1	0.0
	$\widehat{p}_{r.e}$	10.9	93.7	2.7	3.6	0.0
	$\widehat{p}_{r.e}$ CC	11.9	96.1	1.7	2.2	0.0
	$\widehat{p}_{r.opt}$	10.5	93.0	3.1	3.9	0.0
	$\widehat{p}_{r.opt}$ CC	11.5	95.3	2.0	2.7	0.0
	$\widehat{p}_d$	10.8	93.5	2.5	4.0	0.0
	$\widehat{p}_d$ CC	11.8	95.7	1.7	2.6	0.0
Otros	Binomial	13.3	95.9	2.3	1.8	0.0
	Binomial Mid	12.5	95.9	2.3	1.8	0.0
	Verosimilitud	12.3	95.9	2.3	1.8	0.0

estimador de razón óptimo y los métodos que no utilizan información auxiliar disminuye considerablemente, aunque el intervalo basado en $\widehat{p}_{r.opt}$ sigue siendo el de menor amplitud. En este caso, tanto el método Wald como el método Score, tanto para muestras más pequeñas como para muestras más grandes y tanto usando CC con sin hacer uso de esta corrección, los intervalos basados en $\widehat{p}_{r.opt}$ dan resultados bastante aceptables, puesto que en todos los casos las medidas PD y PE están razonablemente equilibradas y las coberturas son ligeramente más pequeñas del 95 % cuando no se usa CC y ligeramente mayores del 95 % cuando se hace uso de la corrección por continuidad.

Cuando P_A = 0,215 y ϕ =0.9 (Tablas 4.9 y 4.10), el intervalo óptimo propuesto también mejora sustancialmente a los intervalos que no utilizan información auxiliar en términos de amplitud media. En este caso no resulta preciso el uso de CC, puesto que la cobertura de tales intervalos sin CC ya superan el 95 %.

Cuando disminuye ϕ (Tablas 4.11 y 4.12), los intervalos óptimos propuestos siguen siendo los más estrechos y se aconseja el uso de CC con el fin de obtener coberturas más cercanas al 95 % requerido.

Cuando la proporción poblacional está próxima a 0 (Tablas 4.13 y 4.14), destacamos que el método Wald, cuando el tamaño muestral es $n = 50$, proporciona intervalos de confianza con extremos inferiores menores que 0 en porcentajes próximos al 50 %, por lo que en esta situación es recomendable utilizar el método Score, donde el intervalo óptimo propuesto que utiliza CC tiene una cobertura del 96.1 %. Los valores de PD y PE no están compensados en esta situación.

Cuando $n = 100$ (Tabla 4.14), el método Score mejora ligeramente en términos de cobertura empírica, y el método Wald obtiene intervalos con extremos inferiores menores que 0 en el 3.8 % de los casos, por lo que ambos métodos aportan intervalos de confianza aceptables.

Por último, se incluye un estudio de simulación basado en los datos de la encuesta nacional de salud (Población ENS). Los resultados de este estudio se presentan en las Tablas 4.15, 4.16, 4.17 y 4.18. Al igual que en casos anteriores, los intervalos óptimos propuestos son los de menor amplitud de entre todos los métodos analizados en el estudio. Tanto para P_A =0.07 como para P_A =0.12, el método Wald proporciona coberturas muy bajas para los intervalos óptimos propuestos, además de intervalos con un mal comportamiento en términos de las medidas PD y PE. Por su parte, el método Score aporta intervalos con coberturas más próximas al 95 % y medidas de PD y PE más equilibradas, por

lo que parece un método más apropiado para poblaciones con características similares a las que posee la población ENS.

4.3. Extensión a un diseño muestral general

4.3.1. Construcción de intervalos de confianza

La estimación puntual y por intervalos de confianza de una proporción han sido estudiados normalmente bajo la suposición de variables aleatorias independientes e idénticamente distribuidas. Sin embargo, el problema de la estimación de la proporción poblacional cuando las muestras se extraen bajo un diseño de muestreo general ha recibido menos atención.

En la mayoría de las encuestas se asumen diseños muestrales complejos, y el uso de métodos de estimación que tengan en cuenta los pesos muestrales pueden ofrecer una mejor estimación que los enfoques habituales que no tienen el efecto del diseño muestral en consideración.

Del mismo modo que en las secciones anteriores, se pueden obtener intervalos de confianza para la proporción poblacional cuando las muestras son seleccionadas mediante un diseño muestral general. Asumiendo el uso del método Wald, la construcción de intervalos de confianza basados en los estimadores propuestos bajo un diseño muestral general es bastante directo, puesto que se han deducido para todos los casos las correspondientes expresiones para las varianzas estimadas. A modo de ejemplo, el intervalo de confianza para P_A utilizando el método Wald y basado en el estimador de tipo razón viene dado por

$$\widehat{p}_{r.HT} \pm z\sqrt{\widehat{V}(\widehat{p}_{r.HT})},$$

donde el estimador de la varianza

$$\widehat{V}(\widehat{p}_{r.HT}) = \frac{1}{N^2}\sum_{i<j\in s}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{A_i - \widehat{R}_{HT}B_i}{\pi_i} - \frac{A_j - \widehat{R}_{HT}B_j}{\pi_j}\right)^2,$$

se estableció en el Capítulo 2. Al igual que en casos anteriores, es posible incorporar en este intervalo la corrección por continuidad para mejorar la cobertura del intervalo de confianza. De forma análoga, se pueden construir intervalos de confianza basados en el resto de estimadores propuestos para un diseño muestral general.

Por otra parte, la extensión del método Score a un diseño de muestreo general no es tan obvia, y requiere una investigación más profunda que se abordará más adelante como una posible futura línea de investigación.

4.3.2. Comparación empírica de intervalos de confianza

A continuación se comparan numéricamente los distintos intervalos de confianza en la población ENS y con muestras extraídas bajo un diseño muestral con probabilidades de selección desiguales. Como se comentó en la Sección 3.2.3, se ha utilizado muestreo estratificado con estratificación basada en el segundo atributo auxiliar, y se utiliza afijación uniforme con el fin de obtener pesos muestrales con mucha variación y analizar el efecto que pueden tener estos pesos muestrales en los distintos intervalos de confianza. Su utiliza únicamente la población ENS, puesto que es la única población real que dispone de dos atributos auxiliares. Se utilizará el primer atributo auxiliar como información auxiliar en la etapa de estimación y el segundo atributo auxiliar para estratificar la población.

Cuando P_A =0.07 (Tabla 4.19), observamos que los estimadores basados en información auxiliar son ligeramente más estrechos que los intervalos basados en el estimador $\widehat{p}_{A.HT}$. La cobertura empírica cuando $n = 50$ está bastante alejada del 95 % para todos los intervalos considerados, los valores PE son claramente superiores a los valores PD y en un gran porcentaje de casos (en torno al 50 %) los distintos intervalos de confianza obtienen extremos inferiores por debajo de 0. Esta situación poco satisfactoria se solventa en gran medida cuando se aumenta el tamaño de la muestra a $n = 100$, donde las coberturas rondan el 90 % y tan solo los intervalos basados en $\widehat{p}_{r.HT}$ obtienen extremos inferiores por debajo de 0. No se observa una mejoría apreciable de los estimadores basados en información auxiliar con respecto al estimador estándar. Como se comprobó en el estudio de comparación de estimadores, este hecho se debe a que ϕ no es muy elevado en esta población, y se podrían obtener resultados más satisfactorios para valores mayores de ϕ.

Cuando P_A =0.12 (Tabla 4.20), los resultados mejoran ligeramente en comparación con el caso P_A =0.07. Por un lado, los intervalos son un poco más amplios cuando P_A =0.12, pero por otro lado, las coberturas cuando P_A =0.12 son mejores, especialmente cuando el tamaño de la muestra es 50. Cuando P_A =0.12 también se reduce considerablemente el número de casos en los que el intervalo de confianza contiene valores negativos.

Tabla 4.19: Para muestreo estratificado, valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo y $P[L < 0]$ de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población ENS. P_A =0.07, P_B =0.03 y ϕ =0.583. Se utiliza el método Wald.

n	Método	Amplitud	CI	PD	PE	$P[L < 0]$
50	$\widehat{p}_{A.HT}$	12.8	69.8	0.1	30.1	55.8
	$\widehat{p}_{A.HT}$ CC	14.2	69.8	0.1	30.1	59.5
	$\widehat{p}_{r.HT}$	12.7	68.5	0.3	31.2	39.4
	$\widehat{p}_{r.HT}$ CC	14.1	68.8	0.0	31.2	58.7
	$\widehat{p}_{r.e.HT}$	12.6	68.6	0.2	31.2	40.4
	$\widehat{p}_{r.e.HT}$ CC	14.1	68.8	0.0	31.2	59.2
	$\widehat{p}_{r.opt.HT}$	12.5	68.7	0.1	31.2	53.8
	$\widehat{p}_{r.opt.HT}$ CC	13.9	68.8	0.0	31.2	59.2
	$\widehat{p}_{reg.HT}$	12.6	68.7	0.1	31.2	46.3
	$\widehat{p}_{reg.HT}$ CC	14.0	68.8	0.0	31.2	59.2
	$\widehat{p}_{d.HT}$	12.6	68.8	0.0	31.2	46.6
	$\widehat{p}_{d.HT}$ CC	14.0	68.8	0.0	31.2	59.2
100	$\widehat{p}_{A.HT}$	10.4	90.1	0.2	9.7	0.0
	$\widehat{p}_{A.HT}$ CC	11.4	90.1	0.2	9.7	0.0
	$\widehat{p}_{r.HT}$	10.3	89.0	0.7	10.3	5.4
	$\widehat{p}_{r.HT}$ CC	11.3	89.3	0.4	10.3	5.6
	$\widehat{p}_{r.e.HT}$	10.2	89.3	0.4	10.3	0.0
	$\widehat{p}_{r.e.HT}$ CC	11.2	89.5	0.2	10.3	0.0
	$\widehat{p}_{r.opt.HT}$	10.2	89.5	0.2	10.3	0.0
	$\widehat{p}_{r.opt.HT}$ CC	11.2	89.5	0.2	10.3	0.0
	$\widehat{p}_{reg.HT}$	10.2	89.5	0.2	10.3	0.0
	$\widehat{p}_{reg.HT}$ CC	11.2	89.5	0.2	10.3	0.0
	$\widehat{p}_{d.HT}$	10.2	89.5	0.2	10.3	0.0
	$\widehat{p}_{d.HT}$ CC	11.2	89.5	0.2	10.3	0.0

Tabla 4.20: Para muestreo estratificado, valores, en porcentaje, de Amplitud media, Cobertura (CI), porcentaje de casos en los el parámetro cae por debajo (PD) y por encima (PE) del intervalo y $P[L < 0]$ de los intervalos de confianza (al 95 %) basados en los distintos estimadores y obtenidos a partir de la población ENS. P_A =0.12, P_B =0.04 y ϕ =0.51. Se utiliza el método Wald.

n	Método	Amplitud	CI	PD	PE	$P[L < 0]$
50	$\widehat{p}_{A.HT}$	19.0	88.5	0.3	11.2	16.7
	$\widehat{p}_{A.HT}$ CC	20.7	88.5	0.3	11.2	44.1
	$\widehat{p}_{r.HT}$	18.8	86.9	0.9	12.2	8.5
	$\widehat{p}_{r.HT}$ CC	20.6	87.3	0.5	12.2	31.2
	$\widehat{p}_{r.e.HT}$	18.7	87.2	0.7	12.2	3.2
	$\widehat{p}_{r.e.HT}$ CC	20.5	87.6	0.2	12.2	31.2
	$\widehat{p}_{r.opt.HT}$	18.6	87.6	0.2	12.2	14.5
	$\widehat{p}_{r.opt.HT}$ CC	20.3	87.6	0.2	12.2	42.1
	$\widehat{p}_{reg.HT}$	18.6	87.5	0.3	12.2	5.2
	$\widehat{p}_{reg.HT}$ CC	20.4	87.6	0.2	12.2	35.7
	$\widehat{p}_{d.HT}$	18.7	87.6	0.2	12.2	5.0
	$\widehat{p}_{d.HT}$ CC	20.4	87.6	0.2	12.2	32.6
100	$\widehat{p}_{A.HT}$	14.2	92.1	0.7	7.2	0.0
	$\widehat{p}_{A.HT}$ CC	15.2	92.4	0.4	7.2	0.0
	$\widehat{p}_{r.HT}$	14.5	90.2	1.7	8.1	3.4
	$\widehat{p}_{r.HT}$ CC	15.5	90.7	1.2	8.1	3.5
	$\widehat{p}_{r.e.HT}$	14.0	91.0	0.9	8.1	0.5
	$\widehat{p}_{r.e.HT}$ CC	15.0	91.3	0.6	8.1	0.9
	$\widehat{p}_{r.opt.HT}$	14.0	91.3	0.6	8.1	0.0
	$\widehat{p}_{r.opt.HT}$ CC	15.0	91.5	0.4	8.1	0.0
	$\widehat{p}_{reg.HT}$	14.0	91.3	0.6	8.1	0.0
	$\widehat{p}_{reg.HT}$ CC	15.0	91.4	0.5	8.1	0.2
	$\widehat{p}_{d.HT}$	14.0	91.3	0.6	8.1	0.6
	$\widehat{p}_{d.HT}$ CC	15.0	91.4	0.5	8.1	1.1

Capítulo 5

Redacción para aspirar a la mención europea en el título de Doctor

5.1. Abstract

Estimation of a proportion is commonly used in areas such as marketing, survey sampling, business, medicine, biopharmaceutical experiments, etc. Estimation of a proportion using auxiliary information has not been investigated in the literature. The present work discusses the estimation of a population proportion in the presence of binary auxiliary information, and some applications to the estimation of proportions in the field of the economy and the business are shown.

Chapter 1 is an introduction to this work, where some applications of the survey sampling on the economy and the business are given. Also, previous references about the use of the auxiliary information are discussed.

In Chapter 2, the ratio estimators for the population proportion are introduced. First, the notation followed in this text and other basic definitions in survey sampling are described in this chapter. Then, various ratio estimators are defined under simple random sampling without replacement. Theoretical properties are established and they are used to define optimum ratio estimators. The extension and the definition of the ratio estimators to a general sampling design is also discussed in this chapter.

In Chapter 3, regression type estimators for the population proportion are defined. Assuming simple random sampling without replacement, it is proved that the optimum regression estimator coincides with the optimum ratio estimator defined in Chapter 2. Theoretical properties are analyzed and a theoretical comparison of the different estimators described in this text is carried out. Estimators are also compared via simulation studies based on different populations. The relative bias and the relative root mean square error are the measures used to compare the various estimators. This chapter finishes with the definition of the regression estimator under a general sampling design. Theoretical and empirical studies are also analyzed.

Assuming the proposed ratio and regression estimators, two-sided confidence intervals are derived in Chapter 4. Both cases of simple random sampling without replacement and a general sampling design are also considered. Empirical studies are used to compare the different confidence intervals in terms of interval width, coverage, etc.

The present Chapter 5 has been prepared in order to obtain the European Mention in the PhD. According to recent law, this chapter includes the present abstract and the main conclusions derived from this work. In addition, some ratio estimators and the corresponding confidence intervals are presented. Note that this work was recent published in the Journal of Biopharmaceutical Statistics (See Rueda et al., 2011).

This work finishes with the Appendix A, which describes the various populations used in the simulation studies.

5.2. Ratio estimators and confidence intervals for the proportion

Ratio estimators of the population proportion and two-sided confidence intervals based upon auxiliary information are derived in this section. Real data extracted from a National Health Survey are used to demonstrate the application of the proposed methods in the estimation of prevalences. Results derived from simulation studies show that proposed estimators are more efficient than the traditional estimator. Proposed confidence intervals outperform the alternative methods, specially in terms of interval width.

5.2.1. Introduction

Estimation of a single proportion is a commonly used statistic in many practical situations and studies derived from many areas (e.g. medical statistics, biopharmaceutical experiments, clinical research, marketing research, survey sampling, etc.). For the problem of the estimation of a proportion, two different tools for presenting results from medical or biopharmaceutical studies are the confidence intervals (see for example, Newcombe, 1998; Gardner and Altman, 1989) and the hypothesis tests (see for example, Han, 2008; Dann and Koch, 2008). Confidence intervals and hypothesis tests are also used by topics related to the estimation of a single proportion, such as the difference of proportions (Chen et al., 2004; Schaarschmidt et al., 2009), the ratio of two proportions (Dann and Koch, 2005) and correlated proportions (May and Johnson, 1998). However, various medical studies indicate that confidence intervals are generally preferred to p-values for several arguments which are set out by Gardner and Altman (1989). For example, Newcombe (1998) argues that a major advantage of confidence intervals in the presentation of results is that interval estimates, in common with point estimates, are relatively close to the data, being on the same scale of measurement, whereas the p-value is a probabilistic abstraction.

The customary sample proportion is calculated as the percentage of individuals with a specific attribute divided by the total number of individuals in the sample. Assuming this simple estimation, there exist many confidence intervals for the population proportion proposed in the literature. For example, the simple asymptotic Gaussian approximation (Wald method in Vollset, 1993) has the advantage of producing intervals centred on the point estimate, but also this method leads to two obvious defects or aberrations, namely overshoot and degeneracy. In practice, the use of a continuity correction (CC) avoids degeneracy but leads to more instances of overshoots (see Newcombe, 1998). The Wilson (Wilson, 1927) score method is another commonly used method for the construction of confidence intervals of a proportion that uses the asymptotic variance associated to the sample estimation. It has the theoretical advantage amongst asymptotic methods of being derived from the efficient score approach (Cox and Hinkley, 1974).

Most methods for estimating a population proportion and forming confidence intervals, including the aforementioned, are based on the assumption of a simple random sample drawn from a large population. However, this scenario is not always presented in practice, i.e., many surveys assume a finite population with samples extracted from complex sampling designs. For example, the

National Health and Nutrition Examination Survey (NHANES) carried out by the National Center for Health Statistics in US uses a complex sampling design (stratified multistage probability sample). In this situation, the use of estimation methods involving sampling weights can provide better estimates than the customary approaches. Also, when the population size is known or the sampling fraction is not small enough, the assumption of a finite population can provide better results than the assumption of a large population.

In sample survey, auxiliary information is often used at the estimation stage to increase the precision of estimators of totals or means (Särndal et al., 1992; Singh, 2003), distribution functions (Rao et al., 1990), quantiles (Rueda et al., 2007) and other parameters. When the interest is the estimation of a proportion, it is also common to have auxiliary information related to the attribute of interest. However, existing procedures for the estimation of a proportion and construction of confidence intervals do not take the auxiliary information into consideration. The use of auxiliary information into the estimation stage can provide better interval and point estimates. We address the incorporation of the auxiliary information into the estimation stage for the problem of the interval and point estimation of a population proportion. Proposed estimators and confidence intervals take the sampling weights into account, as they are based on a general sampling design.

In Section 5.2.2, we define a ratio estimator of the population proportion under simple random sampling without replacement (SRSWOR). Note that SRSWOR is one of the most commonly used sampling designs in practice, as it is well known to the samplers for its ready and simple applicability, and forms the basis of many other sampling designs encountered in theory and practice. Some properties of the proposed ratio estimator are obtained, and they allow us to define a novel ratio estimator more efficient than the previous one. The extension to the presence of multivariate auxiliary information is also addressed in this section. In Section 5.2.3, proposed methods are implemented under the case of a general sampling design, and other problems that might be present in practice are also addressed. The confidence intervals commonly used under the assumption of a large population are described in Section 5.2.4. However, we define them under the context of a finite population. In Section 5.2.5, we propose confidence intervals based on the ratio estimators discussed in Section 5.2.2. Assuming different scenarios (sample sizes, sampling designs, correlations, etc), proposed point estimators and confidence intervals are evaluated empirically in Section 5.2.6, and we observe that conclusions are consistent with the theoretical properties derived in previous sections. Assuming real data extracted from the Spanish National Health Survey, a practical situation related to the estimation of prevalences (asthma and allergy) is described in

Section 5.2.7. We observe that proposed estimators and confidence intervals can be easily applied in this situation, and desirable results in comparison to the existing ones are achieved. Assuming stratified random sampling, the performance of the proposed interval and point estimates is studied in Section 5.2.8. Empirical studies show that proposed ratio estimators are more efficient than the traditional estimator, and confidence intervals with the least width are achieved.

5.2.2. Proposed estimators for the population proportion

We consider the scenario of a finite population $U = \{1, \ldots, N\}$ containing N units. Let $A_1, \ldots, A_N$ denote the values of an attribute of interest A, where $A_i = 1$ if ith unit possesses the attribute A and $A_i = 0$ otherwise. Let B denote an auxiliary attribute associated with A and values given by $B_1, \ldots, B_N$, where $B_i = 1$ if ith unit possesses the attribute B and $B_i = 0$ otherwise. We also assume that a sample s, of size n, is selected from U according to SRSWOR. The extension to a general sampling design is addressed in Section 5.2.3. The aim is to estimate the population proportion of individuals that posses the attribute A, i.e. $P_A = N^{-1} \sum_{i=1}^{N} A_i$. This problem will be addressed via both interval and point estimation, and using new procedures involving auxiliary information.

The customary estimator of P_A is given by

$$\widehat{p}_A = \frac{1}{n} \sum_{i \in s} A_i. \tag{5.1}$$

Note that estimator (5.1) makes no use of the auxiliary information at the estimation stage. Ratio type estimators of the population mean are known methods involving auxiliary information that posses desirable properties including an important gain in efficiency. We define novel ratio estimators to estimate the population proportion P_A.

A first ratio estimator of P_A is

$$\widehat{p}_r = \widehat{R} P_B, \tag{5.2}$$

where $\widehat{R} = \widehat{p}_A / \widehat{p}_B$ is the estimation of the population ratio $R = P_A / P_B$, $\widehat{p}_B = n^{-1} \sum_{i \in s} B_i$ and $P_B = N^{-1} \sum_{i=1}^{N} B_i$.

We assume that the population proportion of individuals that posses the attribute B, P_B, is known from a census or estimated without sampling errors. This assumption is commonly used in the survey sampling context when the parameter is the mean (see Särndal et al., 1992) or the distribution function (see Rao et al., 1990). Note that population census carried out by many countries record information of many auxiliary variables at the population level (see, for example, Nascimento-Silva and Skinner, 1995). Some of the auxiliary variables are quantitative, although it is quite common to have categorical variables such as sex, marital status, residential district, etc. Such variables can be used to calculate P_B, which indicates that the aforementioned assumption can be satisfied in many practical situations. On the other hand, P_B can be estimated from a previous census, wave or occasion. For example, the Canadian Labour Force Survey uses a regression estimator whereby some of the population auxiliary parameters are estimated. This estimation can have an impact on the variance estimation. However, Berger et al. (2009) proposed a variance estimator that takes the estimation of the population auxiliary parameters into account, and which can be easily applied to binary variables.

The proposed estimator (5.2) reduces to P_A when $A_i = B_i$ for all $i \in U$, and hence the variance becomes zero in this case. This suggests that $\widehat{p}_r$ might lead to a considerable gain in efficiency over the customary estimator $\widehat{p}_A$ when A_i is closely related to B_i. Confidence intervals with the least width are also likely to be achieved.

Properties of the proposed ratio estimator

Let A^c and B^c denote the complementary attributes of A and B, and consider the population two-way table given by

	B	B^c	
A	N_{11}	N_{12}	$N_{1\cdot}$
A^c	N_{21}	N_{22}	$N_{2\cdot}$
	$N_{\cdot 1}$	$N_{\cdot 2}$	N

(5.3)

where $N_{1\cdot} = \sum_{i=1}^{N} A_i$ is the number of units in the population that posses the attribute A, $N_{2\cdot}$ is the number of units in the population that not to posses the attribute A, etc. Analogously, N_{11} is the number of units in the population that simultaneously posses the attributes A and B, N_{12} is the number of units in the population that simultaneously posses the attributes A and B^c, etc.

Classification (5.3) can be also defined at the sample level as

	B	B^c	
A	n_{11}	n_{12}	$n_{1\cdot}$
A^c	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	n

(5.4)

We first show that $\widehat{p}_r$ is a biased estimator of P_A. Since both $\widehat{p}_A$ and $\widehat{p}_B$ are unbiased estimators of P_A and P_B respectively, we may write

$$B(\widehat{p}_r) = P_B \left(E(\widehat{R}) - \frac{E(\widehat{R}\widehat{p}_B)}{E(\widehat{p}_B)} \right) = -cov(\widehat{R}, \widehat{p}_B).$$

The above result enables us to obtain an upper bound to the bias in the ratio estimator. We see that

$$\frac{\mid B(\widehat{p}_r) \mid}{\sigma_{\widehat{p}_r}} \leq \frac{\mid cov(\widehat{R}, \widehat{p}_B) \mid}{\sigma_{\widehat{p}_r}} = C_{\widehat{p}_B},$$

where $\sigma_{\widehat{p}_r}$ is the standard deviation of $\widehat{p}_r$ and $C_{\widehat{p}_B}$ is the coefficient of variation of $\widehat{p}_B$. From this bound we see that if n is sufficiently large, the bias in the ratio estimator is negligible in comparison to its standard deviation.

An approximation to the bias of the ratio estimator is

$$B(\widehat{p}_r) = \frac{N-n}{(N-1)n} \left(\frac{Q_B}{P_B} - \frac{\phi \sqrt{P_A Q_A P_B Q_B}}{P_A} \right), \tag{5.5}$$

where $Q_A = 1 - P_A$, $Q_B = 1 - P_B$ and

$$\phi = \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2}}}$$

is the Cramer's V coefficient based on the two-way classification (5.3).

Proof

The approximation (2.8) to the bias of the ratio estimator $\widehat{p}_r$ is now achieved. We consider $e_1 = \frac{\widehat{p}_A - P_A}{P_A}$ and $e_2 = \frac{\widehat{p}_B - P_B}{P_B}$.

We express the ratio estimator in terms of e_1 and e_2. Assuming that $\mid e_2/\widehat{p}_B \mid < 1$ we may expand $\widehat{R}$ as a series in powers of e_2:

$$B(\widehat{p}_r) = P_B(E(\widehat{R}) - R) \simeq P_B E(e_1 - e_2 - e_1 e_2 + e_2^2)$$

$$= P_B \left(-\frac{1}{P_B P_A} cov(\widehat{p}_A, \widehat{p}_B) + \frac{V(\widehat{p}_b)}{P_B^2} \right) = \frac{N-n}{(N-1)n} \left(\frac{Q_B}{P_B} - \frac{\phi\sqrt{P_A Q_A P_B Q_B}}{P_A} \right).$$

□

Expression (5.5) implies that $\widehat{p}_r$ is asymptotically unbiased. However, bias might not be negligible for small sample sizes. An estimator of $B(\widehat{p}_r)$, derived analogously to the results in Lohr (1999, pg. 34) and Särndal et al. (1992, pg. 70), is

$$\widehat{B}(\widehat{p}_r) = \frac{1-f}{n-1} \left(\frac{\widehat{q}_B}{\widehat{p}_B} - \frac{\widehat{\phi}\sqrt{\widehat{p}_A \widehat{q}_A \widehat{p}_B \widehat{q}_B}}{\widehat{p}_A} \right), \tag{5.6}$$

where $f = n/N$ is the sampling fraction, $\widehat{q}_A = 1 - \widehat{p}_A$, $\widehat{q}_B = 1 - \widehat{p}_B$ and

$$\widehat{\phi} = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}}$$

is the Cramer's V coefficient based on the two-way classification (5.4).

Expressions of variances are a key issue for the construction of confidence intervals. We now derive the variance of the proposed ratio estimator. The asymptotic variance of $\widehat{p}_r$ is

$$AV(\widehat{p}_r) = \frac{N-n}{(N-1)n} \left(P_A Q_A + R^2 P_B Q_B - 2R\phi\sqrt{P_A Q_A P_B Q_B} \right). \tag{5.7}$$

Proof

The variance of estimator $\widehat{p}_r$ is derived by using the Taylor series. We can write to a first approximation

$$AV(\widehat{p}_r) = V(\widehat{p}_A) + R^2 V(\widehat{p}_B) - 2R cov(\widehat{p}_A, \widehat{p}_B).$$

The variables $(n_{11}, n_{12}, n_{21}, n_{22}) \simeq HG(N, n, N_{11}, N_{12}, N_{21})$ and

$$cov(n\widehat{p}_A, n\widehat{p}_B) = cov(n_{11} + n_{12}, n_{11} + n_{21}) =$$

$$V(n_{11}) + cov(n_{11}, n_{12}) + cov(n_{11}, n_{21}) + cov(n_{12}, n_{21}),$$

hence

$$V(n_{11}) = \frac{N-n}{N-1} n \frac{N_{11}}{N} \left(1 - \frac{N_{11}}{N} \right) \quad ; \quad cov(n_{ik}, n_{jl}) = -\frac{N-n}{N-1} n \frac{N_{ik}N_{jl}}{N^2}.$$

Therefore, we have

$$cov(n\widehat{p}_A, n\widehat{p}_B) = \frac{N-n}{N-1}\frac{n}{N^2}(N_{11}N_{22} - N_{12}N_{21})$$

On substituting, we obtain the approximate expression for the variance of the ratio estimate:

$$AV(\widehat{p}_r) = \frac{N-n}{(N-1)n}\left(P_A(1-P_A) + R^2P_B(1-P_B) - 2R\frac{N_{11}N_{22} - N_{12}N_{21}}{N^2}\right) =$$

$$= \frac{N-n}{(N-1)n}\left(P_A(1-P_A) + R^2P_B(1-P_B) - 2R\phi\sqrt{P_A(1-P_A)P_B(1-P_B)}\right).$$

□

Expression (5.7) depends on unknown parameters and hence it can not be used to construct confidence intervals in practice. For this purpose, we also derive an estimator of (5.7), which is given by

$$\widehat{V}(\widehat{p}_r) = \frac{1-f}{n-1}\left(\widehat{p}_A\widehat{q}_A + \widehat{R}^2\widehat{p}_B\widehat{q}_B - 2\widehat{R}\widehat{\phi}\sqrt{\widehat{p}_A\widehat{q}_A\widehat{p}_B\widehat{q}_B}\right). \tag{5.8}$$

Expressions (5.7) and (5.8) allow us to derive, in Section 5.2.5, confidence intervals for the single proportion P_A using different methods. Some existing methods to construct confidence intervals are described in Section 5.2.4.

An efficient ratio estimator

Properties derived in Section 5.2.2 allow us to achieve a novel ratio estimator, which is more efficient than the estimator (5.2). This suggested ratio estimator is based on the following idea. The customary estimator $\widehat{p}_A$ can be also obtained as $\widehat{p}_A = 1 - \widehat{q}_A$, where $\widehat{q}_A = n^{-1}\sum_{i\in s} A_i^c$, hence $\widehat{p}_A$ has the same performance in the estimation of P_A than the performance of $\widehat{q}_A$ in the estimation of Q_A. However, this property is not satisfied by $\widehat{p}_r$, i.e. it can be easily seen that $\widehat{p}_r \neq 1 - \widehat{q}_r$, where $\widehat{q}_r = \widehat{R}^cQ_B$ and $\widehat{R}^c = (\widehat{q}_A/\widehat{q}_B)$. This implies that an alternative estimator for P_A is $\widehat{p}_{r.q} = 1 - \widehat{q}_r$. The question is to analyze when $\widehat{p}_r$ possesses better properties than $\widehat{p}_{r.q}$ and viceversa. To solve this question we consider the criterion of minimal variance, which is commonly used in the comparison between estimators. We now deduce that $AV(\widehat{p}_r) < AV(\widehat{p}_{r.q})$ when $P_A < P_B$.

Since $AV(\widehat{p}_{rq}) = AV(1-\widehat{q}_r) = AV(\widehat{q}_r)$, the previous problem is equivalent to study when $AV(\widehat{p}_r) < AV(\widehat{q}_r)$. If $AV(\widehat{p}_r) < AV(\widehat{q}_r)$ then we have

$$P_AQ_A + \frac{P_A^2}{P_B^2}P_BQ_B - 2\frac{P_A}{P_B}\phi\sqrt{P_AQ_AP_BQ_B} < Q_AP_A +$$

$$\frac{Q_A^2}{Q_B^2}Q_BP_B - 2\frac{Q_A}{Q_B}\phi\sqrt{Q_AP_AQ_BP_B},$$

$$\frac{P_A^2Q_B}{P_B} - \frac{Q_A^2P_B}{Q_B} - 2\frac{P_A}{P_B}\phi\sqrt{P_AQ_AP_BQ_B} + 2\frac{Q_A}{Q_B}\phi\sqrt{P_AQ_AP_BQ_B} < 0,$$

$$\frac{P_A^2Q_B^2 - Q_A^2P_B^2 - 2P_AQ_B\phi\sqrt{P_AQ_AP_BQ_B} + 2Q_AP_B\phi\sqrt{P_AQ_AP_BQ_B}}{P_BQ_B} < 0,$$

$$(P_AQ_B + Q_AP_B)(P_AQ_B - Q_AP_B) - 2\phi\sqrt{P_AQ_AP_BQ_B}(P_AQ_B - Q_AP_B) < 0,$$

$$(P_AQ_B - Q_AP_B)(P_AQ_B + Q_AP_B - 2\phi\sqrt{P_AQ_AP_BQ_B}) < 0,$$

Since $P_AQ_B - Q_AP_B = P_A - P_AP_B - P_B + P_AP_B = P_A - P_B$, we have

$$(P_A - P_B)\left(\left(\sqrt{P_AQ_B} - \sqrt{Q_AP_B}\right)^2 + 2\sqrt{P_AQ_AP_BQ_B} - 2\phi\sqrt{P_AQ_AP_BQ_B}\right) < 0,$$

$$(P_A - P_B)\left(\left(\sqrt{P_AQ_B} - \sqrt{Q_AP_B}\right)^2 + 2\sqrt{P_AQ_AP_BQ_B}\,(1-\phi)\right) = K_1K_2 < 0.$$

Since $K_2 \geq 0$ we deduce that $AV(\widehat{p}_r) < AV(\widehat{p}_{rq})$ when $P_A < P_B$.

Analogously, it can be easily seen at the sample level that $\widehat{V}(\widehat{p}_r) < \widehat{V}(\widehat{p}_{r.q})$ when $\widehat{p}_A < \widehat{p}_B$, hence a more efficient estimator for the population proportion P_A is given by

$$\widehat{p}_{r.e} = \begin{cases} \widehat{p}_r & \text{if } \widehat{p}_A < \widehat{p}_B \\ \\ \widehat{p}_{r.q} & \text{otherwise} \end{cases} \tag{5.9}$$

Note that $\widehat{p}_{r.q}$ is defined for theoretical reasons, therefore the suggested estimator is only $\widehat{p}_{r.e}$. Since $AV(\widehat{p}_{r.q}) = AV(1-\widehat{q}_r) = AV(\widehat{q}_r)$, we deduce that $AV(\widehat{p}_{r.e}) = AV(\widehat{p}_r)$ if $P_A < P_B$ and $AV(\widehat{p}_{re}) = AV(\widehat{q}_r)$ otherwise. An estimator of $AV(\widehat{p}_{r.e})$ is given by $\widehat{V}(\widehat{p}_{r.e}) = \widehat{V}(\widehat{p}_r)$ if $\widehat{p}_A < \widehat{p}_B$ and $\widehat{V}(\widehat{p}_{r.e}) = \widehat{V}(\widehat{q}_r)$ otherwise, where $AV(\widehat{q}_r)$ and $\widehat{V}(\widehat{q}_r)$ can be easily defined following (5.7) and (5.8). When $\widehat{p}_A = \widehat{p}_B$ we see that $\widehat{p}_r = \widehat{p}_{r.q}$, which implies that $AV(\widehat{p}_r) = AV(\widehat{q}_r)$ and $\widehat{V}(\widehat{p}_r) = \widehat{V}(\widehat{q}_r)$.

Extension to multivariate auxiliary information

We now assume that the attribute of interest A is associated to J auxiliary attributes $B_1, \ldots, B_J$. The proposed ratio estimator in the presence of multivariate auxiliary attributes is given by

$$\widehat{p}_{MR} = \sum_{i=1}^{J} w_i \widehat{p}_{ri} = \mathbf{w}\widehat{\mathbf{p}}_r',$$

where $\widehat{p}_{ri} = \widehat{R}_i P_{B_i}$, $\widehat{R}_i = \widehat{p}_A/\widehat{p}_{B_i}$, $\mathbf{w} = (w_1, \ldots, w_J)$ and $\widehat{\mathbf{p}}_r = (\widehat{p}_{r1}, \ldots, \widehat{p}_{rJ})$. The weights w_i satisfy the condition $\sum_{i=1}^{J} w_i = 1$ and are to be calculated to maximize the precision of the proposed estimator $\widehat{p}_{MR}$.

The variance of $\widehat{p}_{MR}$ can be written as $AV(\widehat{p}_{MR}) = \mathbf{w}\mathbf{C}\mathbf{w}'$, where $\mathbf{C} = (c_{ij})$ is a $J \times J$ matrix defined as $c_{ij} = cov(\widehat{p}_{ri}, \widehat{p}_{rj})$, $i \neq j$, and $c_{ii} = AV(\widehat{p}_{ri})$, with $i, j = 1, \ldots, J$, where

$$AV(\widehat{p}_{ri}) = \frac{N-n}{(N-1)n}\left(P_A Q_A + R_i^2 P_{B_i} Q_{B_i} - 2R_i\phi_i\sqrt{P_A Q_A P_{B_i} Q_{B_i}}\right),$$

$$cov(\widehat{p}_{ri}, \widehat{p}_{rj}) = \frac{N-n}{(N-1)n}\left(P_A Q_A + R_i R_j \phi_{ij}\sqrt{P_{B_i} Q_{B_i} P_{B_j} Q_{B_j}} - \right.$$

$$\left. R_i\phi_i\sqrt{P_A Q_A P_{B_i} Q_{B_i}} - R_j\phi_j\sqrt{P_A Q_A P_{B_j} Q_{B_j}}\right),$$

ϕ_i is the Cramer's V coefficient associated to attributes A and B_i and ϕ_{ij} is the Cramer's V coefficient associated to B_i and B_j.

Estimator $\widehat{p}_{MR}$ has a bias given by

$$B(\widehat{p}_{MR}) = \frac{N-n}{(N-1)n}\sum_{i=1}^{J} w_i\left(\frac{Q_{B_i}}{P_{B_i}} - \frac{\phi_i\sqrt{P_A Q_A P_{B_i} Q_{B_i}}}{P_A}\right),$$

which can be taken into account for the construction of confidence intervals.

Since $\mathbf{C}$ is positive semidefinite, the generalized Cauchy-Schwarz inequality can be used to show that the optimum $\mathbf{w}$ that minimizes $AV(\widehat{p}_{MR})$ is $\mathbf{w}_{opt} = \mathbf{e}\mathbf{C}^{-1}/\mathbf{e}\mathbf{C}^{-1}\mathbf{e}'$, where $\mathbf{e} = (1, \ldots, 1)$. Insertion of $\mathbf{w}_{opt}$ in $AV(\widehat{p}_{MR})$ yields $AV_{min}(\widehat{p}_{MR}) = 1/\mathbf{e}\mathbf{C}^{-1}\mathbf{e}'$.

Since $\widehat{p}_{MR}$ is unknown in practice, we propose to use the multivariate ratio estimator

$$\widehat{p}_{mr} = \widehat{\mathbf{w}}_{opt}\widehat{\mathbf{p}}_r',$$

where $\widehat{\mathbf{w}}_{opt} = \mathbf{e}\widehat{\mathbf{C}}^{-1}/\mathbf{e}\widehat{\mathbf{C}}^{-1}\mathbf{e}'$, $\widehat{\mathbf{C}} = (\widehat{c}_{ij})$, $\widehat{c}_{ij} = \widehat{cov}(\widehat{p}_{ri}, \widehat{p}_{rj})$, $i \neq j$, and $\widehat{c}_{ii} = \widehat{V}(\widehat{p}_{ri})$, with $i, j = 1, \ldots, J$, where

$$\widehat{V}(\widehat{p}_{ri}) = \frac{1-f}{n-1}\left(\widehat{p}_A\widehat{q}_A + \widehat{R}_i^2\widehat{p}_{B_i}\widehat{q}_{B_i} - 2\widehat{R}_i\widehat{\phi}_i\sqrt{\widehat{p}_A\widehat{q}_A\widehat{p}_{B_i}\widehat{q}_{B_i}}\right),$$

$$\widehat{cov}(\widehat{p}_{ri}, \widehat{p}_{rj}) = \frac{1-f}{n-1}\left(\widehat{p}_A\widehat{q}_A + \widehat{R}_i\widehat{R}_j\widehat{\phi}_{ij}\sqrt{\widehat{p}_{B_i}\widehat{q}_{B_i}\widehat{p}_{B_j}\widehat{q}_{B_j}} - \right.$$
$$\left.\widehat{R}_i\widehat{\phi}_i\sqrt{\widehat{p}_A\widehat{q}_A\widehat{p}_{B_i}\widehat{q}_{B_i}} - \widehat{R}_j\widehat{\phi}_j\sqrt{\widehat{p}_A\widehat{q}_A\widehat{p}_{B_j}\widehat{q}_{B_j}}\right).$$

According to the univariate case, a more efficient multivariate ratio estimator is given by

$$\widehat{p}_{mr.e} = \widehat{\mathbf{w}}_{opt.e}\widehat{\mathbf{p}}_{r.e}',$$

where $\widehat{\mathbf{w}}_{opt.e} = \mathbf{e}\widehat{\mathbf{D}}^{-1}/\mathbf{e}\widehat{\mathbf{D}}^{-1}\mathbf{e}'$, $\widehat{\mathbf{p}}_{r.e} = (\widehat{p}_{r1.e}, \ldots, \widehat{p}_{rJ.e})$,

$$\widehat{p}_{ri.e} = \begin{cases} \widehat{p}_{ri} & \text{if } \widehat{p}_A < \widehat{p}_{B_i} \\ 1 - \widehat{q}_{ri} & \text{otherwise} \end{cases}$$

$\widehat{\mathbf{D}} = (\widehat{d}_{ij})$, $\widehat{d}_{ij} = \widehat{cov}(\widehat{p}_{ri.e}, \widehat{p}_{rj.e})$, $i \neq j$, and $\widehat{d}_{ii} = \widehat{V}(\widehat{p}_{ri.e})$, with $i, j = 1, \ldots, J$.

5.2.3. Additional extensions and properties

This section discusses some relevant aspects related to the estimation of proportions that may be present in many practical situations. For example, the estimation in the presence of a negative relationship between the attribute of interest and the auxiliary attributes, or the extension to a general sampling design are some topics addressed in this section.

Proposed methods in the presence of a negative Cramer'S V Coefficient

There exist many situations where the attribute of interest A may have a strong negative relationship with other auxiliary attributes, and the use of this auxiliary information into the estimation stage also can provide satisfactory results in this scenario. However, proposed ratio estimators $\widehat{p}_r$ and $\widehat{p}_{r.e}$ assume a positive relationship between A and the auxiliary attribute B, and they can give undesirable results in this case.

When the Cramer's coefficient is negative, we propose to transform the auxiliary attribute such that the Cramer's coefficient between the attribute of interest and the transformed attribute has the same relationship, but such relationship will be positive.

Let

	B^c	B	
A	N_{12}	N_{11}	$N_{1\cdot}$
A^c	N_{22}	N_{21}	$N_{2\cdot}$
	$N_{\cdot 2}$	$N_{\cdot 1}$	N

(5.10)

be the population two-way table associated to the attributes A and B^c, where the values N_{12}, N_{11}, etc. in (5.10) are defined in (5.3). Since

$$\phi_{AB} = \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2}}} = -\frac{N_{12}N_{21} - N_{11}N_{22}}{\sqrt{N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2}}} = -\phi_{AB^c},$$

we propose to use the attribute B^c as the auxiliary attribute when the Cramer's V coefficient between A and B is negative.

For the multivariate case, the proposed estimators $\widehat{p}_{mr}$ and $\widehat{p}_{mr.e}$ also assume a positive Cramer's V coefficient between A and the auxiliary attributes. Following Rao and Mudholkar (1967), we can use a combination of ratio and product estimators when some relationships among A and the auxiliary attributes are positive and the other are negative. However, it is simpler to use the complementary attributes as the auxiliary information when the relationship between the attributes is negative, according to the aforementioned univariate case, and then we directly can use the proposed multivariate estimators.

Proposed methods in the estimation of low proportions

For low proportions such as prevalences, estimator $\widehat{p}_r$ cannot be calculated when $\widehat{p}_B$ equal 0. However, it is not a problem for the suggested estimator $\widehat{p}_{r.e}$. Indeed, if $\widehat{p}_B = 0$ we have that $\widehat{p}_A \geq \widehat{p}_B$, $\widehat{q}_B = 1 - \widehat{p}_B = 1$, and

$$\widehat{p}_{r.e} = 1 - \widehat{p}_{r.q} = 1 - \frac{\widehat{q}_A}{\widehat{q}_B} Q_B = 1 - \widehat{q}_A Q_B \tag{5.11}$$

according to (5.9). From expression (5.11) we see that $\widehat{p}_{r.e}$ can be calculated for low proportions without any complication.

Proposed methods in the estimation of P_A and Q_A

As commented in Section 5.2.2, the customary estimator $\widehat{p}_A$ has the same performance in the estimation of P_A than the performance of $\widehat{q}_A$ in the estimation of Q_A, since $\widehat{p}_A = 1 - \widehat{q}_A$. It can be easily seen that estimator $\widehat{p}_r$ does not posses this property. However, the suggested estimator $\widehat{p}_{r.e}$ satisfies this property, i.e., $\widehat{p}_{r.e} = 1 - \widehat{q}_{r.e}$, which implies that it can be calculated as $\widehat{p}_{r.e}$ or $1 - \widehat{q}_{r.e}$. Indeed, $\widehat{q}_{r.e}$ is given by

$$\widehat{q}_{r.e} = \begin{cases} \widehat{q}_r & \text{if } \widehat{q}_A < \widehat{q}_B \\ \\ 1 - \widehat{p}_r & \text{otherwise} \end{cases}$$

and

$$1 - \widehat{q}_{r.e} = \begin{cases} 1 - \widehat{q}_r & \text{if } \widehat{q}_A < \widehat{q}_B \\ \\ \widehat{p}_r & \text{if } \widehat{q}_A \geq \widehat{q}_B \end{cases}$$

Since $\widehat{q}_A < \widehat{q}_B$ implies that $\widehat{p}_A > \widehat{p}_B$ and $\widehat{p}_r = 1 - \widehat{q}_r$ when $\widehat{p}_A = \widehat{p}_B$, we deduce that

$$1 - \widehat{q}_{r.e} = \begin{cases} 1 - \widehat{q}_r & \text{if } \widehat{p}_A \geq \widehat{p}_B \\ \\ \widehat{p}_r & \text{if } \widehat{p}_A < \widehat{p}_B \end{cases}$$

which coincides with the estimator $\widehat{p}_{r.e}$ defined on (5.9).

Proposed methods in the presence of a general sampling design

Interval and point estimation related to a single proportion is a topic that has been studied extensively in the literature from many areas. Studies related

to the estimation of other parameters such as means or totals in the context of a general sampling design are also quite prominent. However, the problem of the estimation of a population proportion when samples are extracted under a general sampling design has received less attention than the previous aspects. As commented in Section 5.2.1, many surveys assume complex sampling designs, and the use of estimation methods involving sampling weights can provide better estimates than the customary approaches, which do not take the effect of the sampling design into consideration. We now discuss the estimation of a population proportion when samples are selected under a general sampling design. Point estimators and confidence intervals are derived under both cases of absence and presence of auxiliary information.

Assume that the sample s is selected according to a specified sampling design with inclusion probabilities π_i and π_{ij} assumed to be strictly positive. The customary estimator $\widehat{p}_A$ and estimators proposed in Section 5.2.2 are not appropriate under this assumption, as they do not account for varying survey weights. Confidence intervals associated to the previous estimators also can provide undesirable results under a general sampling design.

Without using any auxiliary information, a simple estimator for P_A that takes the sampling design into account is given by

$$\widehat{p}_{A.HT} = \frac{1}{N} \sum_{i \in s} d_i A_i, \tag{5.12}$$

where $d_i = \pi_i^{-1}$ is the basic design weight. Using auxiliary information, a ratio estimator can be defined as

$$\widehat{p}_{r.HT} = \widehat{R}_{HT} P_B, \tag{5.13}$$

where $\widehat{R}_{HT} = \widehat{p}_{A.HT}/\widehat{p}_{B.HT}$ and $\widehat{p}_{B.HT}$ is defined as (5.12) after substituting A_i by B_i. The remaining proposed estimators can be defined easily following (5.12) and (5.13).

We now derive expressions for variances of estimator (5.13), whereas the extension to the remaining estimators is also quite straightforward. The asymptotic variance of the ratio estimator $\widehat{p}_{r.HT}$ is

$$AV(\widehat{p}_{r.HT}) = \frac{1}{N^2} \sum_{i<j \in U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{A_i - R \times B_i}{\pi_i} - \frac{A_j - R \times B_j}{\pi_j} \right)^2,$$

An unbiased estimator of $AV(\widehat{p}_{r.HT})$ is given by

$$\widehat{V}(\widehat{p}_{r.HT}) = \frac{1}{N^2} \sum_{i<j \in s} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{A_i - \widehat{R}_{HT} B_i}{\pi_i} - \frac{A_j - \widehat{R}_{HT} B_j}{\pi_j} \right)^2.$$

The use of N in (5.12) may leads to instances of overshoot for P_A approaching one. In this situation, a simple solution is to use a Hájek type estimator (Hájek, 1964), which is given by

$$\widehat{p}_{A.H} = \frac{1}{\widehat{N}} \sum_{i \in s} d_i A_i,$$

where $\widehat{N} = \sum_{i \in s} d_i$. The use of $\widehat{N}$ instead of N in expression (5.12) implies that $\widehat{p}_{A.H} = 1$ when $A_i = 1$ for all $i \in s$, and problems of overshoot are thus addressed.

5.2.4. Traditional confidence intervals under SRSWOR

Traditional confidence intervals are based on the assumption of a large population. However, this situation can be unrealistic in practice, i.e., it is quite common to assume that the sample is extracted from a finite population. In this section, we derive the confidence intervals described in Newcombe (1998), which assume a large population, under the context of a finite population.

Following Newcombe (1998), Blyth and Still (1983), etc., we also make use of a continuity correction (CC) $1/(2n)$, since it improves coverage and avoids degeneracy. However, CC leads to wider mean interval width, and hence to more instances of overshoot. Let z be the $1 - \alpha/2$ point of the standard Normal distribution. Assuming a finite population, some confidence intervals with confidence level $1 - \alpha$ based on the customary estimator $\widehat{p}_A$ are given by

1. Simple asymptotic method (Wald method in Vollset, 1993) without continuity correction
$$\widehat{p}_A \pm z \sqrt{\frac{1-f}{n-1} \widehat{p}_A \widehat{q}_A}$$

2. Asymptotic method with continuity correction (Blyth and Still, 1983)
$$\widehat{p}_A \pm \left(z \sqrt{\frac{1-f}{n-1} \widehat{p}_A \widehat{q}_A} + \frac{1}{2n} \right)$$

3. Wilson Score method (Wilson, 1927) using asymptotic variance
$$\frac{\theta(1-\theta)}{n} \frac{N-n}{N-1}$$

and solving for θ

$$\frac{2nk_1\widehat{p}_A + z^2 \pm z\sqrt{z^2 + 4nk_1\widehat{p}_A\widehat{q}_A}}{2(nk_1 + z^2)},$$

where $k_1 = (N-1)/(N-n)$.

4. Score method incorporating continuity correction (Blyth and Still, 1983; Fleiss et al., 2003). The interval consists of all θ such that

$$|\widehat{p}_A - \theta| - \frac{1}{2n} \leq z\sqrt{\frac{\theta(1-\theta)}{n}\frac{N-n}{N-1}}$$

Expressions for the lower (L) and upper (U) limits in closed form are given by

$$L = \frac{2nk_1\widehat{p}_A + z^2 - 1 - z\sqrt{z^2 - 2 - 1/n + 4\widehat{p}_A(nk_1\widehat{q}_A + 1)}}{2(nk_1 + z^2)}$$

and

$$U = \frac{2nk_1\widehat{p}_A + z^2 + 1 + z\sqrt{z^2 + 2 - 1/n + 4\widehat{p}_A(nk_1\widehat{q}_A - 1)}}{2(nk_1 + z^2)}.$$

5. Method using exact binomial tail areas (Clopper and Pearson, 1934; Lentner, 1982). The interval is $[L, U]$, with $L \leq \widehat{p}_A \leq U$, such that for all θ in the interval:

$a)$ if $L \leq \theta \leq \widehat{p}_A$, $\quad kp_t + \sum_{j=t+1}^{n} p_j \geq \frac{\alpha}{2}$

$b)$ if $\widehat{p}_A \leq \theta \leq U$, $\quad \sum_{j=0}^{t-1} p_j + kp_t \geq \frac{\alpha}{2}$

respectively, where

$$p_j = Pr[T = j] = \binom{n}{j}\theta^j(1-\theta)^{n-j},$$

$j = 0, 1, \ldots, n$, T denoting the random variable of which $t = n_{1\cdot}$ is the realization, and $k = 1$.

6. Method using mid-p binomial tail areas (Miettinen, 1985; Cohen and Yang, 1994). As method 5, but with $k = 1/2$.

7. Likelihood-based method (Miettinen and Nurminen, 1985). The interval comprises all θ satisfying

$$t\ln\theta + (n-t)\ln(1-\theta) \geq t\ln\widehat{p}_A + (n-t)\ln(1-\widehat{p}_A) - z^2/2.$$

5.2.5. Proposed confidence intervals

In this section, we derive two-sided confidence intervals for P_A based on the proposed estimator $\widehat{p}_r$. The extension to the remaining proposed estimators is quite straightforward since expressions of variances are available in this text, hence it is omitted.

The simpler method based on the asymptotic Gaussian approximation (Wald method) is given by

$$\widehat{p}_r \pm z\sqrt{\widehat{V}(\widehat{p}_r)},$$

where $\widehat{V}(\widehat{p}_r)$ is defined by (5.8). This asymptotic method with continuity correction (Blyth and Still, 1983) is

$$\widehat{p}_r \pm \left(z\sqrt{\widehat{V}(\widehat{p}_r)} + \frac{1}{2n}\right)$$

The confidence interval based on the Score method consists of all θ such that

$$|\widehat{p}_r - \theta| \leq z\sqrt{\frac{N-n}{(N-1)n}\left(\theta(1-\theta) + \frac{\theta^2}{P_B}Q_B - 2\frac{\theta}{P_B}\widehat{\phi}\sqrt{\theta(1-\theta)P_BQ_B}\right)}.$$

We can also use the Score method incorporating continuity correction. This interval consists of all θ such that

$$|\widehat{p}_r - \theta| - \frac{1}{2n} \leq z\sqrt{\frac{N-n}{(N-1)n}\left(\theta(1-\theta) + \frac{\theta^2}{P_B}Q_B - 2\frac{\theta}{P_B}\widehat{\phi}\sqrt{\theta(1-\theta)P_BQ_B}\right)}.$$

For methods based on the Wald method, the extension of proposed confidence intervals to a general sampling design is quite straightforward, as variances of estimators are available. However, the extension of proposed confidence intervals to a general sampling design and using the Score method is a topic that needs further research.

5.2.6. Monte carlo studies

In this section, the proposed ratio estimators are compared numerically with the customary estimator used in many practical situations. Proposed

confidence intervals are also compared with existing ones in the literature. Simulation studies are based on several simulated populations which cover a wide number of possible scenarios, including small and large proportions, small and large Cramer's V coefficients between the attribute of interest and the auxiliary attributes, etc. Assuming data extracted from the Spanish National Health Survey, Section 5.2.7 describes a practical situation in which the proposed estimators can be applied in the estimation of prevalences, and desirable results are also achieved. Simulated populations are briefly described as follows.

A total of 30 populations of $N = 1000$ units were generated to study the effect of different aspects on the estimators of a population proportion. Populations were generated as a random sample of 1000 units from a Bernoulli distribution with parameter $p = \{0{,}1, 0{,}25, 0{,}5, 0{,}75, 0{,}9\}$, and the attributes of interest were thus achieved with the aforementioned population proportions. Auxiliary attributes were also generated by using the same distribution, but we randomly change a given proportion of values in order to the Cramer's V coefficient between the attribute of interest and the auxiliary attribute goes from 0.5 to 0.9. Since $P_A < P_B$ when $P_A = 0{,}25$, we also generated populations with $P_A = 0{,}25$ and $P_A > P_B$, which allow us to study the effect of the relation between P_A and P_B on the proposed estimators, specially on the estimators $\widehat{p}_r$ and $\widehat{p}_{mr}$.

For each of the 30 populations, $D = 10000$ samples were selected to compare the various estimators in terms of relative bias (RB) and relative efficiency (RE), where

$$RB = \frac{E[\widehat{p}] - P_A}{P_A} \quad ; \quad RE = \frac{MSE[\widehat{p}]}{MSE[\widehat{p}_A]},$$

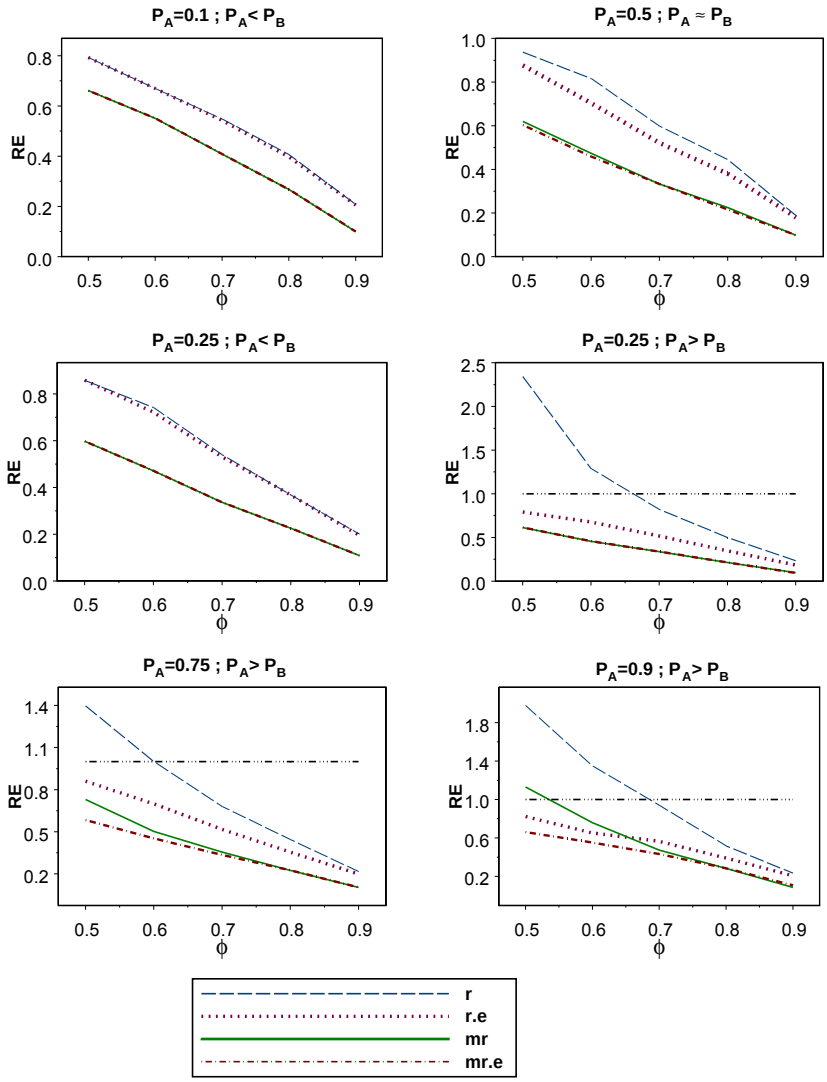
$\widehat{p}$ is a given estimator and the empirical expectation ($E[\cdot]$) and the empirical mean square error ($MSE[\cdot]$) are given by

$$E[\widehat{p}] = \frac{1}{D}\sum_{d=1}^{D}\widehat{p}_d \quad ; \quad MSE[\widehat{p}] = \frac{1}{D}\sum_{d=1}^{D}(\widehat{p}_d - P_A)^2,$$

$\widehat{p}_d$ denotes the estimator $\widehat{p}$ calculated at the dth simulation run. Values of RE less than 1 indicate that the estimator $\widehat{p}$ is more efficient than the customary estimator $\widehat{p}_A$, which is considered as the reference estimator in the efficiency studies.

We considered the proposed estimators $\widehat{p}_r$, $\widehat{p}_{r.e}$, $\widehat{p}_{mr}$ and $\widehat{p}_{mr.e}$. Following Hartley and Ross (1954), an unbiased ratio estimator can be defined by incorporating the bias estimator (5.6) into the estimation stage. However, we

Figura 5.1: Estimated RE of the proposed estimators for simulated populations. Samples with size $n = 100$ are selected under SRSWOR. P_A ranges from 0.1 to 0.9 and ϕ ranges from 0.5 to 0.9.



observed in a simulation study not presented here that the estimator $\widehat{p}_r$ is more efficient than this unbiased estimator, and both estimators had a similar RB. For this reason, we decided not to use the unbiased estimator in this simulation study.

As far as the confidence interval is concerned, the criteria by which we evaluate the different methods are given as follows. We computed the empirical coverage probability (CP) of the 95 % confidence intervals. Another commonly used criterion for evaluation of confidence intervals is the empirical average width achieved from the $D = 10000$ confidence intervals calculated by each method. Note that it is desirable to achieve the required coverage with the least width. We constructed confidence intervals (with and without CC) based on the estimators $\widehat{p}_A$ (customary), $\widehat{p}_{r.e}$ and $\widehat{p}_{mr.e}$, and using both Wald and Score methods. In Section 5.2.7, we examine more criteria for confidence intervals and compare proposed confidence intervals to other alternative methods.

Monte Carlo studies were carried out by using the freeware statistical program R (R Development Core Team, 2009). For application of the proposed estimators and confidence intervals in this software, R codes are available from the authors on request.

Values of RB in this simulation study are within a reasonable range, i.e, they are all less than 1 % and are thus omitted. Figure 5.1 reports the RE of the proposed estimators for simulated populations and samples selected under SRSWOR. First, we observe that proposed estimators have a larger gain in efficiency in comparison to the customary estimator $\widehat{p}_A$ as the Cramer's V coefficient increases. This is due to the fact that proposed estimators make an appropriate use of the auxiliary information at the estimation stage.

Figure 5.1 also indicates when the population proportion P_A is larger than P_B, since it may have an impact on the proposed estimators $\widehat{p}_r$ and $\widehat{p}_{mr}$. When $P_A < P_B$ we observe that the ratio estimator $\widehat{p}_r$ is as efficient as the suggested ratio estimator $\widehat{p}_{r.e}$. This also occurs for the multivariate ratio estimators $\widehat{p}_{mr}$ and $\widehat{p}_{mr.e}$, which are more efficient than $\widehat{p}_r$ and $\widehat{p}_{r.e}$, which in turn are more efficient than the customary estimator. As expected, the suggested ratio estimators $\widehat{p}_{r.e}$ and $\widehat{p}_{mr.e}$ are more efficient than $\widehat{p}_r$ and $\widehat{p}_{mr}$ when $P_A \geq P_B$. We observe that this gain in efficiency is larger as ϕ decreases. This is due to the fact that P_A and P_B are quite apart from each other as ϕ decreases, whereas P_A and P_B are closer as ϕ increases. For cases where $P_A > P_B$, the estimator $\widehat{p}_{mr}$ seems closer to $\widehat{p}_{mr.e}$ than $\widehat{p}_r$ to $\widehat{p}_{r.e}$. In other words, the efficiency of the efficient estimator does not improve as much in the multivariate case as in the univariate case. In particular, when $P_A = 0{,}25$, the estimator $\widehat{p}_{mr.e}$ is almost

equal to $\widehat{p}_{mr}$, whereas $\widehat{p}_{r.e}$ and $\widehat{p}_r$ are quite apart from each other. We observe that $\widehat{p}_{mr.e}$ is the most efficient estimator in Figure 5.1, and proposed estimators $\widehat{p}_{r.e}$ and $\widehat{p}_{mr.e}$ are always more efficient than the customary estimator $\widehat{p}_A$.

Conclusions derived from Figure 5.1 support the theoretical properties obtained in Section 5.2.2. Indeed, proposed estimators make an effective use of the auxiliary information and better estimates in terms of RE are achieved in comparison to the customary estimator. On the other hand, Section 5.2.2 shows that $\widehat{p}_{r.e}$ is more efficient than $\widehat{p}_r$ when $\widehat{p}_A \geq \widehat{p}_B$, and they have the same performance otherwise. Empirical studies obtained in this section also support this theoretical result. Finally, the multivariate ratio estimators are more efficient than the ratio estimators based on a single auxiliary variable. This is due to the fact that multivariate ratio estimators use more auxiliary information than $\widehat{p}_r$ and $\widehat{p}_{r.e}$ at the estimation stage. According to the conclusions derived in this section, we decided not to use estimators $\widehat{p}_r$, $\widehat{p}_{mr}$ hereafter.

We now turn our attention to the confidence intervals calculated by different methods. Figures 5.2 and 5.3 show the mean coverage probabilities (CP), and the mean width for simulation populations. Samples with size $n = 100$ are selected under SRSWOR. Population proportions $P_A = 0{,}25$ and $P_A = 0{,}5$ are respectively considered in Figures 5.2 and 5.3. According to Section 5.2.3, the case $P_A = 0{,}75$ is not presented here because the methods analyzed have the same performance in terms of CP and mean width than such methods under the case $Q_A = 1 - P_A = 0{,}25$, which is showed in Figure 5.2.

From Figures 5.2 and 5.3 we observe that the main advantage of the proposed confidence intervals is that they clearly are narrower than confidence intervals based on the customary estimator. The average width of the proposed confidence intervals is smaller as ϕ increases, and they are always narrower than confidence intervals based on the customary estimator. For example, with $\phi = 0{,}9$ we observe that some proposed confidence intervals are 3 times narrower than the customary confidence interval. This desirable property is directly related to the gain in efficiency showed in Figure 5.1. Indeed, suggested ratio estimators are clearly more efficient than the customary estimator, and it implies that suggested estimators have a smaller variance than the customary estimator. This result is also due to the fact that the biases are negligible. With smaller variances it is obvious that confidence intervals will be narrower than intervals based on the customary estimator.

Figure 5.2 ($P_A = 0{,}25$) shows that the overall average CP ranges from 93 % to 96 % for methods based on the Wald method, and CP approximately ranges from 94 % to 97 % for methods based on the Score method. On average

Figura 5.2: Estimated CP (%) and Width (%) for 95 % confidence intervals calculated by various methods in simulated populations. Samples with size $n = 100$ are selected under SRSWOR. $P_A = 0{,}25$ and ϕ ranges from 0.5 to 0.9.

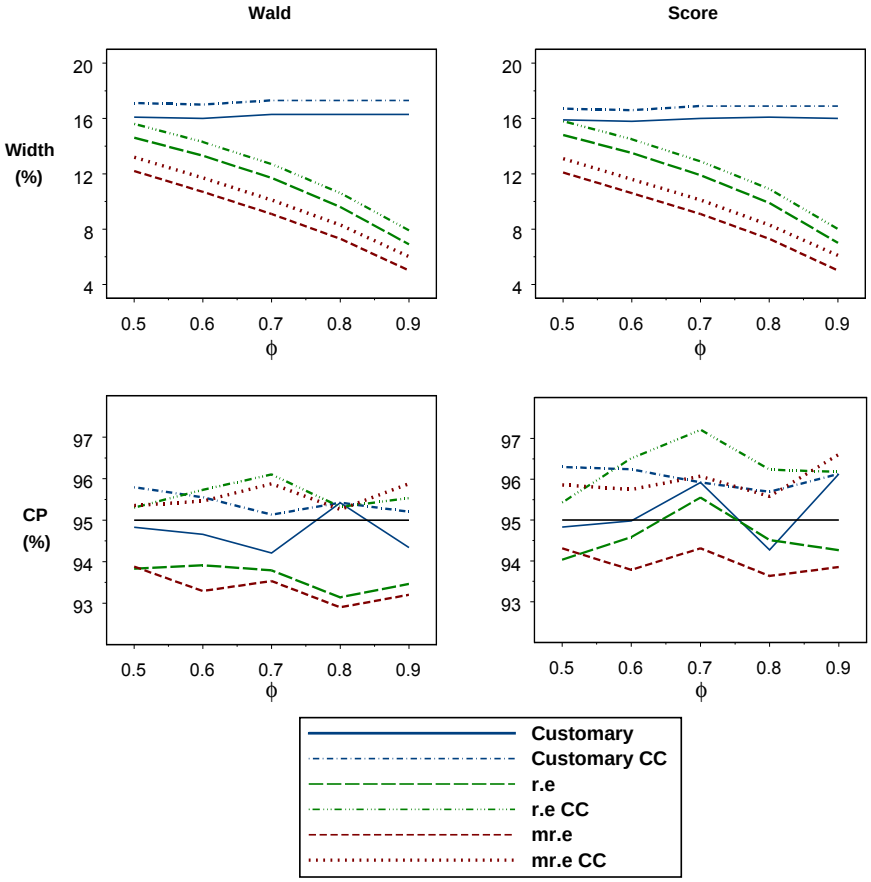
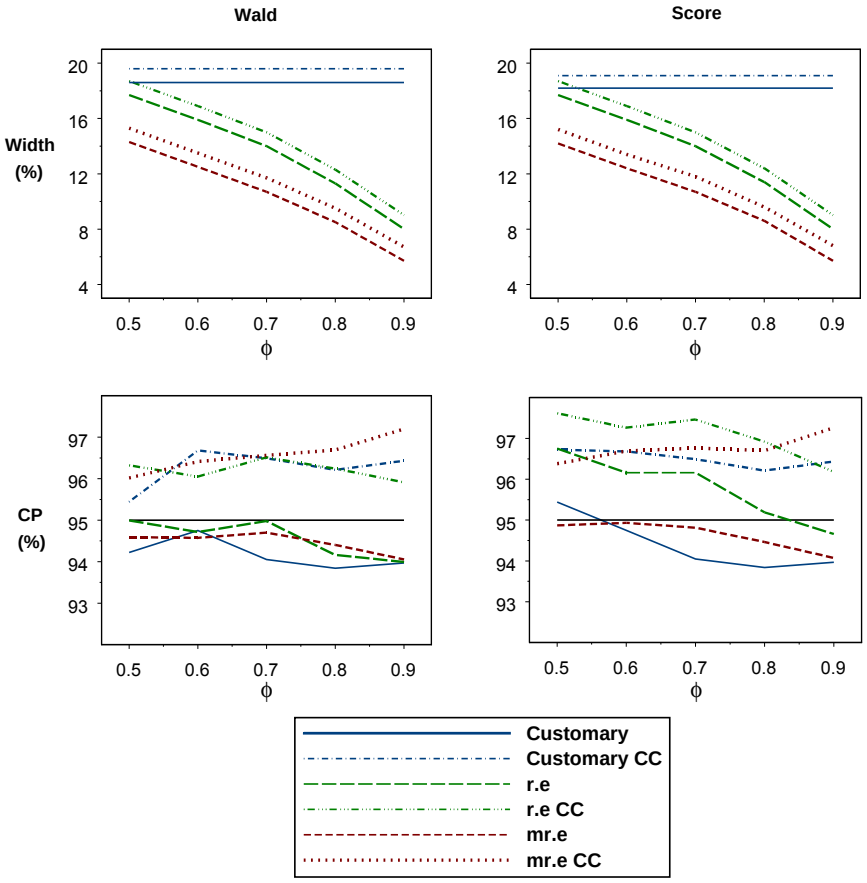


Figura 5.3: Estimated CP (%) and Width (%) for 95 % confidence intervals calculated by various methods in simulated populations. Samples with size $n = 100$ are selected under SRSWOR. $P_A = 0{,}5$ and ϕ ranges from 0.5 to 0.9.



confidence intervals based on the estimator $\widehat{p}_{mr.e}$ are slightly anti-conservative. However, this method achieves desirable CP when continuity correction is incorporated. A similar situation is showed by the ratio estimator $\widehat{p}_{r.e}$, although conservative confidence interval on average can be achieved when both Score method and continuity correction are considered.

For $P_A = 0{,}5$ (Figure 5.3) and the Wald method, proposed confidence intervals without CC have coverage probabilities close below 0.95, whereas the customary confidence interval is more anti-conservative than the previous ones. The use of CC yields to conservative confidence intervals for the three methods. Confidence intervals have a similar performance under the Score method. We point out that the confidence intervals based on $\widehat{p}_{r.e}$ are conservative, although this problem is solved as ϕ increases.

5.2.7. Simulations results based on real data

In this section, proposed estimators and proposed confidence intervals were applied to the estimation of prevalences for the chronic diseases asthma and allergy. The performance of the proposed methods can be thus observed for low proportions. This study is based on $N = 9063$ real data extracted from the Spanish National Health Survey 2006 (NHS2006), i.e., we consider the sample data as a population which samples will be extracted. Simulations were thus conducted to compare the performance of the various estimators and confidence intervals in this practical situation. Detailed information of the NHS2006 is described as follows.

The NHS2006 assumes a complex sampling design where the first stage units are the census sections (2236 census sections are selected in the sample). The first stage units are grouped into strata in agreement with the size of the municipality. The second stage units are main family dwellings. Within the second stage units no sub-sampling is carried out, and all dwellings with usual residence are investigated. Within each household, an adult (16 years old or over) is selected to complete the adults questionnaire and if there are children (0 to 15 years old) in the household, one of them is selected as well to complete the children's questionnaire. Data used in this simulation study are referred to the children's questionnaires. A total of 31300 households were selected in the sample and $N = 9063$ children fill the questionnaires.

Databases of national health statistics contain data on health indicators, including basic demographic and socioeconomic indicators; some lifestyle- and environment-related indicators and health care resources, utilization and ex-

penditure. Chronic diseases as asthma and allergy are investigated by the NHS2006. World Health Organization recognizes that asthma is of major public health importance. The Organization plays a role in coordinating international efforts against this disease. International Study of Asthma and Allergies in Childhood found that the prevalence of both deseases is rising in European children. The aim is to use alternative and more precise point estimates for the commented prevalences and the construction of confidence intervals with desirable coverage and the least width.

Data used in simulations involve three variables. The variable of interest is a dichotomous variable (attribute A) indicating whether the child has ever suffered from the disease (asthma or allergy). Auxiliary information is provided by the variables "has the doctor prescribed drugs to the child for asthma/allergy?"(auxiliary attribute B_1), and "has the child consumed drugs in the last two weeks for asthma/allergy?"(auxiliary attribute B_2). Note that $P_A = 0{,}07$, $\phi_1 = 0{,}583$ and $\phi_2 = 0{,}570$ for the asthma case, and $P_A = 0{,}12$, $\phi_1 = 0{,}510$ and $\phi_2 = 0{,}495$ for the allergy. Proposed estimator $\widehat{p}_{r.e}$, which is based on a single auxiliary variable, used the attribute B_1 as the auxiliary information.

Note that the Health Information System of the Spanish National Health System has information on the prescribed drugs, which is classified by therapeutical groups, age, gender, disease, etc. The electronic prescription will be implanted in the very near future, and this information will be greater. On the other hand, studies derived from private and public agencies such as the Official College of Pharmacists or the Business Federation of Spanish Pharmacists also have relevant information on prescribed drugs, specially related to child health. These arguments indicate that the proposed estimators and confidence intervals can be applied in this situation, since the population proportion of prescribed drugs is known. Furthermore, the proportion of consumed drugs can be obtained by the Health Information System. However, our interest is to study the empirical performance of the proposed estimators and confidence intervals by assuming real data.

Following Section 5.2.6, the evaluation of point estimates is realized in terms of RB and RE and using 10000 simulations. However, the lower and upper tail error rates of the 95 % confidence intervals are now computed for the evaluation of confidence intervals. These rates are commonly named, respectively, as the distal ($DNCP$) and mesial ($MNCP$) non-coverage probabilities. It is desirable that $DNCP$ and $MNCP$ should be equal. Since population proportions are close to 0, we also computed the empirical probability that the lower limit (L) calculated by each method is less than 0, i.e., we calculated $P[L < 0]$.

Tabla 5.1: Estimated RB (%) and RE of various estimators for the NHS2006 population. Samples are selected under SRSWOR.

			RB (%)			RE		
	n	f (%)	$\widehat{p}_A$	$\widehat{p}_{r.e}$	$\widehat{p}_{mr.e}$	$\widehat{p}_A$	$\widehat{p}_{r.e}$	$\widehat{p}_{mr.e}$
Asthma	50	0.6	10.6	-0.6	4.8	1.00	0.70	0.59
	100	1.1	1.8	-0.2	-0.1	1.00	0.68	0.69
	250	2.8	-0.2	0.0	-0.4	1.00	0.67	0.67
	500	5.5	0.1	0.0	-0.3	1.00	0.66	0.66
	750	8.3	0.0	0.0	-0.2	1.00	0.65	0.65
	1000	11.0	0.1	0.0	-0.3	1.00	0.68	0.68
Allergy	50	0.6	3.2	-1.0	0.3	1.00	0.78	0.77
	100	1.1	-0.1	-0.3	-0.7	1.00	0.77	0.77
	250	2.8	0.1	0.0	-0.2	1.00	0.76	0.76
	500	5.5	-0.2	-0.1	-0.3	1.00	0.75	0.74
	750	8.3	0.0	0.0	-0.2	1.00	0.73	0.73
	1000	11.0	0.0	0.1	0.0	1.00	0.75	0.73

Table 5.1 reports the estimated RB and RE of estimators for the NHS2006 population. We observe that the customary estimator $\widehat{p}_A$ has a large bias when the sampling fraction is small, but this estimator gives small biases as f increases. Biases of the suggested estimator $\widehat{p}_{r.e}$ are negligible for small and large sampling fractions, i.e., values of RB are all less than 1 % in absolute values. The proposed multivariate ratio estimator $\widehat{p}_{mr.e}$ has a large bias when $f = 0{,}6\,\%$ and the variable is the Asthma. It is clear from Table 5.1 that in terms of RE the proposed estimators $\widehat{p}_{r.e}$ and $\widehat{p}_{mr.e}$ are more efficient than $\widehat{p}_A$, even in this population where ϕ is small. As commented in Section 5.2.6, this gain in efficiency would increase if ϕ is larger. Estimators $\widehat{p}_{r.e}$ and $\widehat{p}_{mr.e}$ have a similar performance in the estimation of the proportions of both diseases. This is likely due to the fact that both auxiliary attributes provide an analogous auxiliary information. However, it is obvious that $\widehat{p}_{r.e}$ is computationally simpler than the estimator $\widehat{p}_{mr.e}$, hence $\widehat{p}_{r.e}$ is a preferable estimator in comparison to $\widehat{p}_{mr.e}$ in this situation.

Table 5.2 shows the performance of various methods of constructing a confidence interval for P_A in the population NHS2006 and the Asthma variable. This study confirms that proposed confidence intervals are the narrowest among the different methods analyzed.

For $n = 100$ and the Wald method, the various confidence intervals (with

Tabla 5.2: Estimated Width, CP, $DNCP$, $MNCP$ and $P[L < 0]$ in percentage for 95 % confidence intervals calculated by various methods in NHS2006 population and the Asthma variable. Samples are selected under SRSWOR.

n		Method	Width	CP	$DNCP$	$MNCP$	$P[L < 0]$
100	Wald	Customary	10.0	93.5	1.5	5.0	5.0
		Customary CC	10.9	94.3	0.7	5.0	12.7
		r.e	8.0	93.6	0.4	6.0	0.7
		r.e CC	9.0	93.9	0.3	5.8	1.8
		mr.e	8.0	93.6	0.8	5.6	1.5
		mr.e CC	9.0	94.2	0.6	5.2	2.5
	Score	Customary	10.3	95.6	2.9	1.5	0.0
		Customary CC	11.2	96.8	2.9	0.3	0.0
		r.e	9.7	92.8	3.7	3.5	0.0
		r.e CC	11.2	96.9	1.7	1.4	0.0
		mr.e	9.5	92.6	3.8	3.6	0.0
		mr.e CC	10.6	95.9	1.9	2.2	0.0
	Others	Binomial	11.1	97.0	1.5	1.5	0.0
		Binomial Mid	10.3	95.6	2.9	1.5	0.0
		Likelihood	9.9	95.6	2.9	1.5	0.0
500	Wald	Customary	4.4	94.2	1.7	4.1	0.0
		Customary CC	4.6	94.8	1.0	4.2	0.0
		r.e	3.6	93.9	1.4	4.7	0.0
		r.e CC	3.8	95.3	1.1	3.6	0.0
		mr.e	3.6	93.7	1.3	5.0	0.0
		mr.e CC	3.8	95.1	0.9	4.0	0.0
	Score	Customary	4.4	94.6	3.8	1.6	0.0
		Customary CC	4.6	95.7	2.6	1.7	0.0
		r.e	3.5	93.6	3.3	3.1	0.0
		r.e CC	3.7	95.1	2.5	2.4	0.0
		mr.e	3.5	93.4	3.5	3.1	0.0
		mr.e CC	3.7	94.9	2.8	2.4	0.0
	Others	Binomial	4.7	96.6	1.7	1.7	0.0
		Binomial Mid	4.5	94.7	2.6	2.7	0.0
		Likelihood	4.4	94.7	2.6	2.7	0.0

Tabla 5.3: Estimated Width, CP, $DNCP$, $MNCP$ and $P[L < 0]$ in percentage for 95 % confidence intervals calculated by various methods in NHS2006 population and the Allergy variable. Samples are selected under SRSWOR.

n		Method	Width	CP	$DNCP$	$MNCP$	$P[L < 0]$
100	Wald	Customary	12.4	94.4	1.0	4.6	0.2
		Customary CC	13.4	94.4	1.0	4.6	0.6
		r.e	10.7	93.5	1.0	5.5	0.2
		r.e CC	11.7	94.5	0.6	4.9	0.4
		mr.e	10.5	92.8	0.9	6.3	0.1
		mr.e CC	11.5	94.1	0.4	5.5	0.5
	Score	Customary	12.5	94.1	3.9	2.0	0.0
		Customary CC	13.4	97.3	2.1	0.6	0.0
		r.e	10.8	93.6	2.5	3.9	0.0
		r.e CC	11.9	96.3	1.6	2.1	0.0
		mr.e	11.1	93.4	4.3	2.3	0.0
		mr.e CC	12.1	96.3	2.2	1.5	0.0
	Others	Binomial	13.3	95.9	2.1	2.0	0.0
		Binomial Mid	12.5	95.9	2.1	2.0	0.0
		Likelihood	12.3	95.9	2.1	2.0	0.0
500	Wald	Customary	5.5	95.0	1.6	3.4	0.0
		Customary CC	5.7	95.0	1.6	3.4	0.0
		r.e	4.7	94.4	1.7	3.9	0.0
		r.e CC	4.9	95.4	1.4	3.2	0.0
		mr.e	4.7	94.3	1.7	4.1	0.0
		mr.e CC	4.9	95.1	1.4	3.5	0.0
	Score	Customary	5.5	94.6	2.9	2.5	0.0
		Customary CC	5.6	95.4	2.1	2.5	0.0
		r.e	4.6	94.1	2.7	3.2	0.0
		r.e CC	4.8	95.2	2.2	2.6	0.0
		mr.e	4.7	94.2	3.2	2.6	0.0
		mr.e CC	4.9	95.4	2.6	2.0	0.0
	Others	Binomial	5.8	95.9	1.6	2.5	0.0
		Binomial Mid	5.6	95.4	2.1	2.5	0.0
		Likelihood	5.5	95.4	2.1	2.5	0.0

and without CC) are slightly anti-conservative. However, intervals have a poor performance in terms of distal and mesial non-coverage probabilities, since the $DNCP$ values are close to 0 and the $MNCP$ values are all larger than 5 %. On the other hand, the 5 % (without CC) and the 12,7 % (with CC) of the confidence intervals based on $\widehat{p}_A$ obtain a lower limit L below 0. Then confidence intervals based on $\widehat{p}_{mr.e}$ violates this bound in the 1,5 % and the 2,5 % of the cases. According to the violation of bounds criterion, the estimator $\widehat{p}_{mr.e}$ provides the best confidence intervals, since only the 0,7 % and the 1,8 % of the intervals give a limit L less than 0.

The Score method has some advantages in comparison to the Wald method. For example, with Score methods violations of the boundaries are not achieved, whereas confidence intervals have a better performance in terms of distal and mesial non-coverage probabilities. However, wider confidence intervals are now reported in comparison to the Wald method. As far as the coverage is concerned, the use of CC achieves coverage probabilities close to 95 %. Distal and mesial aspects of non-coverage are reasonably closely balanced for the proposed methods, whereas $DNCP$ and $MNCP$ are not balanced for the confidence intervals based on the customary estimator.

Binomial and Likelihood methods are slightly conservative and provide confidence intervals less balanced than proposed methods based on the Score method. These confidence intervals are generally wider than proposed confidence intervals.

As the sample size increases ($n = 500$), confidence intervals generally perform very well in terms of coverage, since CP values, in general, are close to the required 95 % coverage. For the Wald method we observe that there is no violations of bounds, though the $MNCP$ values are again much larger than the $DNCP$ values. Proposed methods based on the Score method are very well balanced and furthermore they achieve confidence intervals narrower than those based on Wald method. Proposed methods perform better than existing ones in terms of interval width.

From Table 5.3 (the Allergy variable) we observe that better properties are generally observed in comparison to Table 5.2. For example, with $n = 100$ there is less cases of violations of bounds and CP values are generally closer to 95 % than those in Table 5.2.

5.2.8. Evaluation of methods under a general sampling design

We now turn to the application of the proposed methods in situations where samples are selected under a sampling design more complex than the SRSWOR. For this purpose, we first compare the proposed estimators involving sampling weights ($\widehat{p}_{A.HT}$, $\widehat{p}_{r.e.HT}$ and $\widehat{p}_{r.HT}$) to the customary estimator ($\widehat{p}_A$) that ignores them into the estimation stage. In this case, variations of sampling weights are very different to show the impact on the estimation of a proportion of estimator $\widehat{p}_A$. Then we study the performance of proposed methods under stratified random sampling and use the size of the municipality as the variable for the formation of strata, i.e., we carried out simulations with the same conditions used by the Spanish National Health Survey.

The relative bias (RB) defined in Section 5.2.6 and the relative efficiency $RE_{HT} = MSE[\widehat{p}]/MSE[\widehat{p}_{A.HT}]$ are the measures used to study the performance of point estimators, whereas the coverage probability (CP), the distal ($DNCP$) and mesial ($MNCP$) non-coverage probability, the interval width and the empirical probability that L is less than zero are the criteria for evaluation of confidence intervals.

For the first simulation, we used stratified random sampling with stratification based on the auxiliary attribute B_2 described in Section 5.2.7. The use of equal allocation in this simulation allows us to achieve sampling weights very different, and the effect that the sampling design may have on the various estimators can be observed. The multivariate ratio estimator is not calculated here because the NHS2006 population only have two auxiliary attributes, and the second one is already used at the design stage.

Table 5.4 reports the evaluation of the various estimators in terms of RB and RE_{HT} under stratified random sampling with equal allocation. As expected, the customary estimator $\widehat{p}_A$ has a poor performance in terms of RB and RE_{HT}, since it does not take the sampling weights into consideration. Values of RB of estimators $\widehat{p}_{A.HT}$ and $\widehat{p}_{r.e.HT}$ are negligible, i.e. they are less than 1 % in absolute terms. The estimator $\widehat{p}_{r.e.HT}$ is the most efficient. These results are consistent with those obtained in Section 5.2.6. A simulation study not presented here confirms that the Binomial and Likelihood methods for the constructions of confidence intervals give coverage probabilities less than 60 % in this simulation. This is due to fact that the estimator $\widehat{p}_A$ has a poor performance.

Assuming the same variable for stratification (size of the municipality)

Tabla 5.4: Estimated RB (%) and RE_{HT} of various estimators for the NHS2006 population. Samples are selected under stratified random sampling with stratification based on the attribute B_2 and equal allocation.

		RB (%)			RE_{HT}		
	n	$\widehat{p}_A$	$\widehat{p}_{A.HT}$	$\widehat{p}_{r.e.HT}$	$\widehat{p}_A$	$\widehat{p}_{A.HT}$	$\widehat{p}_{r.e.HT}$
Asthma	50	579.4	0.5	0.5	103.24	1.00	0.97
	100	579.2	0.2	0.3	212.37	1.00	0.97
	250	579.0	0.2	0.2	524.10	1.00	0.98
	500	579.4	0.4	0.4	1059.54	1.00	0.97
Allergy	50	312.6	0.1	0.2	48.05	1.00	0.97
	100	313.0	-0.5	-0.5	95.69	1.00	0.94
	250	313.2	0.2	0.2	245.71	1.00	0.98
	500	312.8	-0.1	-0.1	492.78	1.00	0.98

Tabla 5.5: Estimated RB (%) and RE_{HT} of various estimators for the NHS2006 population. Samples are selected under stratified random sampling with stratification based on the size of the municipality and proportional allocation.

		RB (%)		RE_{HT}	
	n	$\widehat{p}_{A.HT}$	$\widehat{p}_{r.e.HT}$	$\widehat{p}_{A.HT}$	$\widehat{p}_{r.e.HT}$
Asthma	50	10.3	-0.8	1.00	0.70
	100	1.7	-0.4	1.00	0.68
	250	0.1	0.2	1.00	0.68
	500	0.1	0.1	1.00	0.67
Allergy	50	2.7	-1.1	1.00	0.78
	100	0.7	0.2	1.00	0.75
	250	0.0	-0.1	1.00	0.75
	500	0.0	-0.1	1.00	0.75

Tabla 5.6: Estimated Width, CP, $DNCP$, $MNCP$ and $P[L < 0]$ in percentage for 95 % confidence intervals calculated by various methods in NHS2006 population and the Asthma variable. Samples are selected under stratified random sampling with stratification based on the size of the municipality and proportional allocation.

n		Method	Width	CP	$DNCP$	$MNCP$	$P[L < 0]$
100	Wald	A.HT	10.0	94.2	0.8	5.0	5.0
		A.HT CC	10.9	94.5	0.5	5.0	12.2
		r.e.HT	8.0	93.2	0.5	6.3	0.6
		r.e.HT CC	8.9	93.6	0.3	6.1	1.9
	Others	Binomial	11.1	97.4	1.2	1.4	0.0
		Binomial Mid	10.3	95.8	2.7	1.5	0.0
		Likelihood	9.9	95.9	2.7	1.4	0.0
500	Wald	A.HT	4.4	94.0	1.6	4.4	0.0
		A.HT CC	4.6	95.4	1.2	3.4	0.0
		r.e.HT	3.6	93.8	1.5	4.7	0.0
		r.e.HT CC	3.8	95.1	1.1	3.8	0.0
	Others	Binomial	4.7	96.7	1.6	1.7	0.0
		Binomial Mid	4.5	94.9	2.4	2.7	0.0
		Likelihood	4.4	94.9	2.4	2.7	0.0

Tabla 5.7: Estimated Width, CP, $DNCP$, $MNCP$ and $P[L < 0]$ in percentage for 95 % confidence intervals calculated by various methods in NHS2006 population and the Allergy variable. Samples are selected under stratified random sampling with stratification based on the size of the municipality and proportional allocation.

n		Method	Width	CP	$DNCP$	$MNCP$
100	Wald	A.HT	12.4	94.6	1.2	4.2
		A.HT CC	13.4	94.9	0.9	4.2
		r.e.HT	10.7	93.8	1.1	5.1
		r.e.HT CC	11.7	95.1	0.6	4.3
	Others	Binomial	13.4	96.2	2.2	1.6
		Binomial Mid	12.6	96.2	2.2	1.6
		Likelihood	12.3	96.2	2.2	1.6
500	Wald	A.HT	5.5	95.4	1.5	3.0
		A.HT CC	5.7	95.4	1.5	3.0
		r.e.HT	4.7	94.7	1.6	3.7
		r.e.HT CC	4.9	95.7	1.2	3.1
	Others	Binomial	5.8	96.5	1.5	2.0
		Binomial Mid	5.6	95.9	2.1	2.0
		Likelihood	5.5	95.9	2.1	2.0

than the Spanish National Health Survey, Tables 5.5, 5.6 and 5.7 show the performance of estimators and confidence intervals under stratified random sampling. We consider proportional allocation because it provides the same results for both estimators $\widehat{p}_A$ and $\widehat{p}_{A.HT}$, and methods of constructing confidence intervals based on $\widehat{p}_A$ (Binomial, Binomial Mid and Likelihood) are thus comparable to those based on $\widehat{p}_{r.e.HT}$.

From Table 5.5 we see that the estimator $\widehat{p}_{r.e.HT}$ is clearly more efficient than $\widehat{p}_{A.HT}$. Biases are negligible for the estimator $\widehat{p}_{r.e.HT}$, whereas $\widehat{p}_{A.HT}$ has large values of RB for $n = 50$.

Tables 5.6 and 5.7 show that methods based on the estimator $\widehat{p}_{r.e.HT}$ give confidence intervals with the least width. Coverage probabilities are close to the required 95 %, specially as the sample size increases.

5.2.9. Discussion

Estimation of a proportion is a commonly used statistic for summarizing data such as prevalences, clinical adverse experiences in clinical trials, and other practical situations in medical and biopharmaceutical studies. The customary proportion estimator does not involve auxiliary information at the estimation stage, and the sampling design is not taken into consideration. When presenting results in medical and biopharmaceutical studies, confidence intervals are generally given in common with point estimates.

Ratio estimators of a population proportion have been proposed in this work. Proposed estimators address the incorporation of the auxiliary information at the estimation stage and take the sampling design into consideration. Some useful theoretical properties and other issues related to the estimation of a proportion have been also addressed. Several simulation studies show that the proposed ratio estimators are clearly more efficient than the customary estimator. An application to the estimation of prevalences confirm that proposed ratio estimators can obtain an important gain in efficiency in comparison to the customary estimator. This gain in efficiency over the customary estimator, which is the basis of many other confidence intervals, allows us to construct narrower confidence intervals in comparison to those achieved by alternative methods. Proposed confidence intervals are based on both Wald and Score methods and provide desirable coverage probabilities with the least width. However, proposed confidence intervals based on the Score method have the advantage of providing more balanced confidence intervals in terms of distal and mesial non-coverage probabilities.

5.3. Conclusions

Various estimators for the population proportion have been proposed in this work. Proposed estimators address the incorporation of the auxiliary information at the estimation stage and take the sampling design into consideration. Proposed estimators are based on the ratio, regression and difference methods.

Assuming simple random sampling without replacement, a ratio estimator is defined and some theoretical properties are established. These properties have allowed to define unbiased and more efficient ratio estimators. Then, the extension to several auxiliary variables is discussed. Finally, an optimum ratio estimator is defined under simple random sampling without replacement. Assuming a general sampling design, other ratio estimators are also defined. Theoretical properties are also established.

The regression method is also considered to define estimators for the population proportion. This estimator is defined under simple random sampling without replacement and the most important theoretical properties are analyzed, which are used to define the optimum regression estimator. The difference type estimator is also defined. A relevant theoretical comparison of the different estimators is made. This comparison includes the standard estimator for the population proportion. This study gives some relevant conclusions. For example, the optimum ratio estimator coincides with the optimum regression estimator under simple random sampling without replacement, and both estimators are more efficient than alternative estimators. Empirical studies based upon different populations confirm the theoretical results. Simulated populations were considered in the Monte Carlo studies in order to study different situations, which can occur in practice. Also, populations based upon real data sets are considered. The data sets were taken from the area of the economy and the business. Results derived from the different simulation studies indicate that the proposed estimators have biases within a reasonable range, and the proposed optimum estimator was the most efficient estimator in each situation. The standard ratio and difference estimators can be less efficient than the customary estimator of the population proportion. However, this situation is theoretically studied and the condition to know when both estimators perform worse than the standard estimator is established.

Assuming a general sampling design, a regression estimator is also defined. Note that this estimator does not coincide with the optimum ratio estimator defined under a general sampling design. The various estimators proposed under a general sampling design are also numerically compared. This empirical

study indicates that the optimum ratio estimator is the most efficient estimator, which is followed by the regression estimator. The standard ratio and difference estimators can be less efficient than the customary estimator of the population proportion.

Once the estimators are defined and theoretical and empirically analyzed, the next step was to define confidence interval based upon the proposed estimators. First, simple random sampling was considered and interval was defined by using two methods: Wald and Score. Also, confidence interval was defined by using a continuity correction. Empirical studies were used to compare the proposed estimators with the alternative estimators existing in the literature. Result derived from these studies indicate that the proposed optimum confidence intervals outperform the alternative methods, specially in terms of interval width. This gain is better as the Cramer's V coefficient increases. Relevant results are obtained under a general sampling design.

Bibliografía

[1] Anderson, Sweeney y Williams. Estadística para Administración y Economía. *Octava Edición* 2005; **31**:4–16.

[2] Bello, A.L. (1993). Choosing among imputation techniques for incomplete multivariate data: a simulation study. *Comunication in Statistics*, **22**:823–877.

[3] Berger, Y.G. (1998) Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *J. Stat. Plan. Infer.* 67 209-226.

[4] Berger, Y. G., Muñoz, J. F., Rancourt, E. (2009). Variance estimation of survey estimates calibrated on estimated control totals - An application to the extended regression estimator and the regression composite estimator. *Comput. Stat. Data An.* 53:2596–2604.

[5] Blair, E. Sampling issues in trade area maps drawn from shopping surveys *Journal of Marketing* 1983; **47**:98–106.

[6] Blyth, C.R., Still, H.A. Binomial confidence intervals. *Journal of the American Statistical Association* 1983; **78**:108–116.

[7] Brewer, K.R.W Design-based or prediction-based inference? Stratified random vs stratified balanced sampling, *Int. Statist. Rev.* 1999b; **67**:35–47.

[8] Cassel, C.M., Särndal and J.H. Wretman, Some results on generalized difference estimation and generalized regression estimation for finite populations, *Biometrika* 1976; **63**:615–620.

[9] Chen, J. y Shao, J. Nearest neighbour imputation for surveydata. *Journal of Official Statistics* 2000; **16**:113–131.

[10] Chen H., Stasny E.A., Wolfe D.A. Ranked set sampling for efficient estimation of a population proportion. *Statistics in Medicine* 2005; **24**:3319–3329.

[11] Clopper C.J., Pearson E.S. The use of confidence or fiducial limits illustrated in the case of binomial. *Bometrika* 1934; **26**:404–413.

[12] Cochran W.G.

Sampling Techniques. New York: John Wiley 1977;

[13] Cohen M.P. A new approach to imputation. *American Statistical Association Proceding of the Section on Survey Research Methods* 1996; **13**:293–298.

[14] Cohen G.R., Yang S.Y. Mid-p confidence intervals for the Poisson expectation. *Statistics in Medicine* 1994; **13**:2189–2203.

[15] Cox D.R., Hinkley D.V. *Theoretical Statistics.* Chapman and Hall, London, 1974.

[16] De Agustín Melendro,J.A. *Guía Práctica de Aplicación del Muestreo Estadístico a la Auditoría* 1995; Madríd, REA

[17] Dillman, D. *Mail and Telephone Surveys,* 1978; New York . John Wiley

[18] Durbin, J. Desing of Multi-stage Surveys for the Estimation of Sampling Error, *Applied Statistics* 1954; **16**:152–164.

[19] Durbin, J. Test of serial independence based on the cummulated periodogram. , *Paper presented at the 36th Session of the International Statistical Institute, Sydney* 1967;

[20] Fay, R.E. (1991). A design-based perspective on missing data variance. In Proc. Seventh Annual Res. Conf., Washington, D.C.: U.S. Bureau of the Census. 429-440.

[21] **Fernández García, F.R. y Mayor Gallego, J.A.** (1994) *Muestreo en Poblaciones Finitas: Curso Básico.* P.P.U., Barcelona.

[22] Fleiss J.L., Levin B., Paik M.C. *Statistical methods for rates and proportions* (3rd edn). Wiley, New Yersey, 2003.

[23] Fuller,W.A. Sampling with Random Stratum Boundaries. *Journal of the Royal Statistical Society. Ser B* 1970; **32**:203–226.

[24] Gardner M.J., Altman D.G. (Eds). *Statistics with Confidence.* British Medical Journal, London, 1989.

[25] Glasser, G.L. y Metzger, G.D. Ramdon Digit Dialing as a Method of Telephone Sampling *Journal of Marketing Research* 1972; **9**:59–64.

[26] Hájek J. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics* 1964; **35**:1491–1523.

[27] Hansen, M. H.; Dalenius T.; Tepping, B.J. "The development of Sample Surveys of Finite Populations"in a Celebration of Statistics. Eds. Anthony C. Atkinson an Stephen E. Fienberg. *NEW YORK. Springer-Verlag.* 1985; **1**:327–354.

[28] Hansen M.H., Hurwitz W.N. On the theory of sampling from finite populations. *Annals of Mathematical Statistics* 1943; **14**:333–362.

[29] Hansen, H.O., Hurwitz W.N. y Gurney, M. Problems and methods of the sample surveys of business *J. Amer. Statis. Assoc.* 1946; **41**

[30] Hansen, H.O., Hurwitz W.N. y Madow W.G. *Sample Survey Methods and Theory. New York. John Wiley* 1953; **2 vol**

[31] Healy, M.J.R. y Westmacott, M. Missing values in experiments analysed on automatic computers *Appled Statis.* 1956; **5**:203–206.

[32] Hartley, H. O.; Rao, J.N.K, y Kiefer, G. Variance Estimation with one Unit pero Stratum, *Journal of the American Statistical Association* 1969; **64**: 841–851.

[33] Hartley, H. M. y Ross, A. Unbiased ratio estimators, *Nature* 1954; **174**:270–271.

[34] Hartley H.O., Ross A. Unbiased ratio estimators. *Nature* 1954; **174**:270–271.

[35] Horvitz D.G., Thompson D.J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; **47**:663–685.

[36] **Instituo Nacional de Estadística.** (1992) Encuesta Continua de Presupuestos Familiares. Metodología. *Instituto Nacional de Estadística. Madrid.*

[37] Isaki, C. T. Variance Estimation Using Auxiliary Information. *Journal of the American Statistical Association* 1983; **78**:117–123.

[38] Kadilar, C. and Cingi H. Ratio estimator in simple random sampling *Applied Mathematics and Computation* 2004; **151**:893–902.

[39] Kalton, G. Introduction to Survey Sampling *SAGE Publications* 1983; **31**:1–6.

[40] Kalton, G. y Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology* **12** 1–16.

[41] Kish, L. A Procedure for Objetive Respondent Selection Within the Household, *Journal of the American Statistical Association* 1949; **44**:380–387.

[42] Kish, L. *Survey Sampling.* New York. John Wiley 1965;

[43] Lentner C. (ed). *Geigy Scientific Tables.* (8th edition, volume 2). Ciba-Geigy, Basle, 1982.

[44] Levy, P.S. Optimum Allocation in Stratified Random Network Sampling for Estimating the Prevalence of Attributees in Rare Populations, *Journal of the American Statistical Association* 1977; **72**:758–763.

[45] Little, R.J.A. y Rubin, D.B. (2002). *Statistical analysis with missing data. 2^{nd} edition.* New York: John Wiley & Sons, Inc.

[46] Lorh, S.L. (ed). *Sampling: design and analysis* Duxbury, 1999.

[47] Menendez, E. y Ferrales J.

El estimador de razón generalizado. *Trabajos de Estadística* 1989; **4 (1)**

[48] Miettinen O.S. *Theoretical Epidemiology.* Wiley, New York, 1985.

[49] Miettinen O.S., Nurminen M. Comparative analysis of two rates. *Statistics in Medicine* 1985; **4**:213–226.

[50] Muñoz J.F., Álvarez-Verdejo, E., Arcos A., Rueda M.M., González, S. Optimum ratio estimators for the population proportion. *International Journal of Computer Mathematics* 2011; En prensa.

[51] Murthy, M.N Product Method of Estimation *The Indian Journal of Statistics, Series A* 1964; **26**:64–74.

[52] Newcombe R.G. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 1998; **17**:857–872.

[53] Neyman, J. On the two different aspects of the Representative Method: The method of Stratified sampling and the method of puposive selection. *Journal of the royal Statistics Society* 1934; **97**:558-606.

[54] Ogus, J. L. y Clark, D . F. The Annual Survey of Manufactures: A report on Methodology, *U.S. Bureau of the Census Tecnical Paper. U.S Goberment Printing Office, Washington,DC* 1971; **24**

[55] Olkin, I. Multivariate ratio estimation for finite population. *Biometrika* 1958; **45**:154–165.

[56] O´Rourke, D. y Blair, J. Improving Random Respondent Selection in Telephone Surveys, *Journal of Marketing Reseach* 1983; **20**:428–432.

[57] Palacios, F. y Callejón, J. (ed). *Técnicas Cuantitativas para el Análisis Regional.* (Editorial Universidad de Granada). Universidad de Granada, Facultad de Ciencias Económicas y Empresariales, 2004.

[58] Prasad, B. Some unbiased estimators versus mean per unit and ratio estimators in finite population sample surveys. *Commun. Statist. Theory and Meth.* 1986; **15 (12)**:3647–3657.

[59] Prasad, B. y Singh H.P. Some improved ratio-type estimators of finite population variance in sample surveys. *Comunications in Statistics. Theory and Methods* 1990; **19(3)**:1127–1139.

[60] Prasad, B. y Singh H.P. Unbiased estimators of finite population variance using auxiliary information in samples surveys. *Comunications in Statistics. Theory and Methods* 1992; **21(5)**:1367–1376.

[61] Pratesi, M., Ranalli, M.G y Salvati, N. Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US.. *Environmetrics* 2008; **19**:687–701.

[62] Quenouille, M.H. Notes on bias in estimation *Biometrika* 1956; **43**:353–360.

[63] Randles, R.H.

On the asymptotic normality of statistics with estimated parameters *Ann. Stat.* 1982; **14**:462–474.

[64] Rao J.N.K. Unbiased ratio and regression estimators in multi-stage sampling. *Journal of the Indian Society of Agricultural Statistics* 1964; **14**:175–188.

[65] Rao J.N.K. On variance estimation with imputed survey data (with discussion). *Journal of the American Statistical Association* 1996; **91**:499–520.

[66] Rao J.N.K. y Shao, J. Jackknife Variance Estimation With Survey Data Under Hot-Deck Imputation. *Biometrika* 1992; **91**:811–822.

[67] Rao J.N.K., Kovar J.G., Mantel H.J. On estimating distribution function and quantiles from survey data using auxiliary information. *Biometrika* 1990; **77**:365–375.

[68] Rao P.S.R.S. On the two-phase ratio estimator in finite population. *Journal of the American Statistical Association* 1975; **70**:839–845.

[69] Rao P.S.R.S. Hartley- Ross type estimator with two phase sampling. *Sankhya Serie C* 1975; **37**:140–146.

[70] Rao P.S.R.S. Efficiences of the nine two-phase ratio estimators fo the mean. *Journal of the American Statistical Association* 1981; **76**:434–442.

[71] Rao P.S.R.S., Mudholkar G.S. Generalized multivariate estimator for the mean of finite population. *Journal of the American Statistical Association* 1967; **62**:1009–1012.

[72] Ray, S.K. and Singh, R.K Differece-cum-ratio type estimators. *Journal of Indian Satatistical Association* 1981; **19**:147–151.

[73] Rubin, D.B. (1996). Mutiple imputation after 18+ years. *Journal of the American Statistical Association*, **91** 473–489.

[74] Rueda M., Ruiz, M. y Arcos, A. Estimadores condensados de razón. *Official Journal of the Chilean Statistical Society* 1992; **9**:15–27.

[75] Rueda M. Aportaciones a la teoría de estimadores de razón. *Tesis Doctoral - Universidad de Granada* 1993;

[76] Rueda M., Muñoz JF, González S., Arcos A. Estimating quantiles under sampling on two occassions with arbitrary sampling designs. *Computational Statistics and Data Analysis* 2007; **51**:6596–6613.

[77] Rueda M.M., Muñoz J.F., Arcos A., Álvarez-Verdejo, E., Martínez, S. Estimators and confidence intervals for the proportion using binary auxiliary information with applications to pharmaceutical studies. *Journal of Biopharmaceutical Statistics* 2011a; **21**:1–29.

[78] Rueda M.M., Muñoz J.F., Arcos A., Álvarez-Verdejo, E. Indirect estimation of proportions in natural resource surveys. *Mathematicss and Computers in Simulation* 2011b; En prensa.

[79] Ray , S.K. y Sahai, A. Efficient families of ratio and product-type estimators. *Biometrika* 1980; **67**:211–215.

[80] Ruiz, M. y Santos J. Unbiased mean-of-the-ratio estimators. *Statistica* 1989; **50**:285–288.

[81] Särndal, C.E. (2007). The calibration approach in survey theory and practice. *Survey Methodology* 33 99-119.

[82] Särndal C.E., Swensson B, Wretman J.H. *Model Assisted Survey Sampling.* Springer-Verlag, New York, 1992.

[83] Sedransk, J. (1985). The objetive and practice of imputation. In *Proc. First Annual Res. Conf.*, Washington, D.C.: Bureau of the Cencus. 445–452.

[84] Shapire, G. M y Bateman, D. V. A better alternative to the Collapse Stratum Variance Estimate, *Proceding of the Social Statistics Section. American Statistical Association* 1978; :451–456.

[85] Sedransk, J.. The objetives and practice of imputation, *Proceding of the First Annual Research Conference* 1985; Washington, DC: United States **Bureau of the Census**:445–452.

[86] Nascimento-Silva, P.L.D, Skinner, C.J. (1995). Estimating distribution functions with auxliary information using Poststratification. *J. Offic. Stat.* 11:277–294.

[87] Silverman B.W. *Density estimation for statistics and data analysis.* Chapman and Hall. 1986.

[88] Singh, M.P. Efficient use of systematic sampling in ratio and product estimation. *Metrika* 1966; **10**:199–205.

[89] Singh S. *Advanced sampling theory with applications: How Michael Selected Amy.* Kluwer Academic Publishers, The Netherlands, 2003.

[90] Sirken, M.G. Households Surveys With Multiplicity, *Journal of American Statistical Association* 1970; **65**:257–266.

[91] Sirken, M.G. Stratified Sample Surveys With Multiplicity, *Journal of American Statistical Association* 1972; **67**:224–227.

[92] Sirken, M.G., and Levy, P.S. Multiplicity Estimation of Proportions Based on Ratios of Random Variables, *Journal of American Statistical Association* 1974; **69**:68–73.

[93] Sisodia, B.V.S. and Dwivedi, V.K. A modified ratio estimator using coefficient of variation of auxiliary variable. *Journal of Indian Society Agricultural Statistics* 1981; **33**:13–18.

[94] Srivastava, S. K. An estimator using auxiliary information in sample surveys. *Calcutta Statist. Assoc. Bull* 1967; **16**:121–132.

[95] Srivenkataramana,T. and Tracy, D.S. An Alternative to Ratio Method in Sample Surveys. *Ann. Inst. Statist. Math.* 1980; **32-Part A**:111–120.

[96] Srivastava, S. K. A class of estimators using auxiliary information in sample surveys. *Canad. J. Statist.* 1980; **8**:253–254.

[97] Srivastava, S. K. and Jhajj, H. S. A Class of Estimators of the Population Mean in Survey Sampling Using Auxiliary Information. *Biometrika.* 1981; **68**:341–343.

[98] Sudman, S. Probability Sampling With Quotas. *Journal of the American Statistical Association* 1966; **61**:749–771.

[99] Sudman, S. *Applied Sampling. Orlando, F.L: Academic Press* 1975;

[100] Sudman, S. Improving the quality of Shopping Center Sampling, *Journal of Marketing Research* 1980; **17**:423–431.

[101] Sudman, S. Efficient Screening Methods fo the Sampling of Geographically Clustered Special Populations, *Journal of Marketing Research* 1985; **22**:20–29.

[102] Sudman, S. and Blair, E. Sampling in the Twenty-First Century, *Journal of the Academy of Marketing Science* 1999; **27**:269–277.

[103] Sudman, S. and Ferber, R. *Consumer Panels ; Chicago: American Marketing Association* 1979;

[104] Swain, A.K.P.C. The use of systematic sampling in ratio estimate. *Journal of the Indian Statistical Association* 1964; **2**:160–164.

[105] Troldahl, V.C, y Carter, R. E Random Selection of Respondents Within Households in Phone Surveys, *Journal of Marketing Reseach*1964; **1**:71–76.

[106] Upadhyaya, L.N. and Singh, H.P Use a transformed auxiliary variable in estimating the finite population mean, *Biometrical Journal* 1999; **41** :627–636.

[107] Vollset S.E. Confidence interval for a binomial proportion. *Statistics in Medicine* 1993; **12**:809–824.

[108] Waksberg, J. Sampling Methods for Random Digit Dialing, *Journal of the American Statistical Association* 1978; **73**:40–46.

[109] Weiers, Ronald M. Introducción a la estadística para los negocios *Thomson* 2006; **31**:2–3.

[110] Williams, W.H. Generating ubiased ratio and regression estimators. *Biometrics* 1961; **17**:267–274.

[111] Williams, W.H. On two methods of unbiased estimation with auxiliary variates *Journal of the American Statistical Association* 1962; **57**:184–186.

[112] Wilson, E.B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 1927; **22**:209–212.

[113] Wu, C. Algorithms and R codes for the Pseudo Empirical Likelihood method in survey sampling. *Survey Methodology* 2005; **31**:239–243.

Apéndice 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.

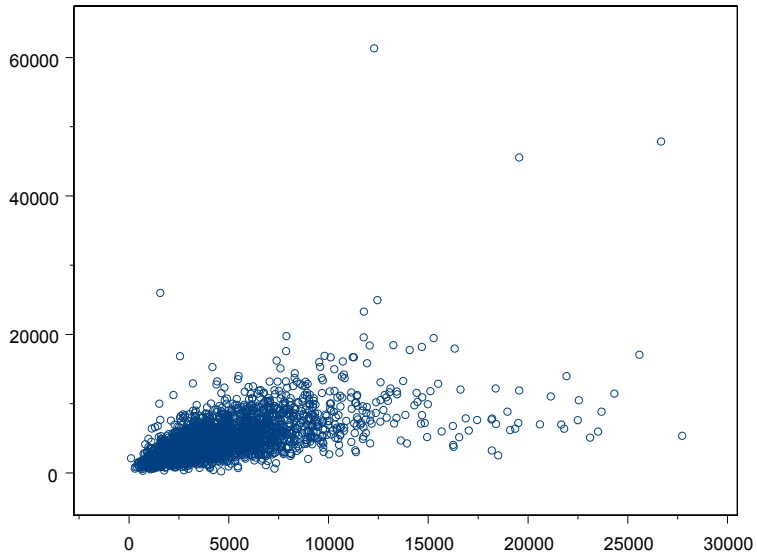
A.1. Poblaciones basadas en datos reales

A.1.1. Población EPF

Esta población está basada en los datos muestrales procedentes del primer trimestre del año 1997 de la Encuesta de Presupuestos Familiares (EPF). Véase Instituto Nacional de Estadística (1992) para una consulta detallada de la metodología. Esta población ha sido también analizada en Fernández *et al.* (2004).

La población EPF está formada por $N = 3114$ familias a las cuales se les preguntaron sobre sus ingresos y gastos familiares. La nube de puntos asociada a estas dos variables de la población EPF puede consultarse en la Figura A.1. La proporción poblacional de interés es la proporción de familias con gastos familiares inferiores a $0.6Q_y(0{,}5)$, donde $Q_y(0{,}5)$ representa la mediana de la variable y. Notamos que este tipo de proporciones representa en varios países la proporción de familias que se encuentran por debajo del umbral de pobreza. Como atributo auxiliar consideramos el valor 1 si una familia tiene ingresos por debajo de $0.6Q_x(0{,}5)$ y 0 en caso contrario. Notamos que esta información se podría conocer a nivel poblacional a partir de la declaración de la renta. Las proporciones poblacionales de los atributos A y B son, respectivamente $P_A = 0{,}194$ $P_B = 0{,}173$ y el valor correspondiente para el coeficiente V de Cramer es de $\phi = 0{,}501$

Figura A.1: Nube de puntos de la población EPF.



El correspondiente análisis descriptivo de las variables de esta población está dado por la Tabla A.1. 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

Tabla A.1: Análisis descriptivo para las variables de la población EPF

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

de algunas medidas como la media.

En cualquier caso, el objetivo es estimar la proporción de individuos con gastos por debajo del umbral de pobreza, y la única condición necesaria para que los estimadores de la proporción propuestos tengan un buen cumplimiento es que la relación entre los atributos principal y auxiliar no sea demasiado baja, tal como se ha comprobado durante este trabajo.

A.1.2. Población ESE

El estudio está realizado por METROSCOPIA, una empresa dedicada a estudios de mercado y opinión. En esta encuesta se tratan sobre todo temas políticos y sociales que afectan a todo el país y también a toda Europa. Este estudio se está realizando en 30 países europeos y está coordinado y dirigido por un grupo de investigadores europeos. Cuenta con el apoyo económico y el respaldo de la Unión Europea y del Ministerio de Educación y Ciencia.

La Encuesta Social Europea (ESE) es una encuesta social, académico-mecánica diseñada para trazar y explicar la interacción entre las instituciones cambiantes de Europa y las actitudes, creencias y patrones de comportamiento de sus diversas poblaciones. Ahora preparada para su quinta ronda, la encuesta abarca a más de 30 países y emplea las metodologías más rigurosas. La encuesta se ha financiado a través de Marco de la Comisión Europea de programas, la Fundación Europea de la Ciencia y los organismos nacionales de financiación en cada país. El folleto informativo ESE proporciona información de base de la encuesta. Además las principales conclusiones de las tres primeras rondas de la encuesta están también disponibles.

El objetivo de la ESE es diseñar, desarrollar y ejecutar un estudio conceptual bien anclado y metodológicamente robusto para observar los cambios en las actitudes sociales y valores. El logro de estos objetivos en un contexto trasnacional requiere çomparabilidad óptima"en la puesta en marcha del estudio en todos los países participantes.

Este "principio de igualdad o equivalencia "se aplica a la traducción de selección muestral de la encuesta, y todos los métodos y procesos.

Los atributos utilizados proceden de estas dos cuestiones:

Utilizando esta tarjeta, si suma los ingresos provenientes de todo tipo de fuentes, ¿qué letra describe mejor los ingresos totales de su hogar después de descontar los impuestos y otras deducciones obligatorias? Si no conoce la cantidad exacta, por favor díganos una cantidad aproximada. Utilice la sección de la tarjeta que mejor conozca: ingresos semanales, mensuales o anuales.

- J 01
- R 02
- C 03
- M 04
- F 05
- S 06
- K 07
- P 08
- D 09
- H 10
- No contesta (No sugerir) 77
- No sabe (No sugerir) 88

¿Cuál de las afirmaciones en esta tarjeta describe mejor cómo se siente con respecto a los ingresos de su hogar en la actualidad?

- Con los ingresos actuales vivimos cómodamente 1
- Con los ingresos actuales nos llega para vivir 2
- Con los ingresos actuales tenemos dificultades 3
- Con los ingresos actuales tenemos muchas dificultades 4

- No sabe (No sugerir) 8

Ambas cuestiones se trasladan en nuestro estudio de forma siguiente:

Se tiene un total de observaciones $N = 990$, la proporción poblacional del atributo A es $P_A = 0{,}496$ y la proporción poblacional del atributo B tiene un valor de $P_B = 0{,}596$, con un coefeciente V de Cramer $\phi = 0{,}467$. El atributo B consiste en la variable $B_i = 1$ si la i-ésima familia gana 04 o menos y $B_i = 0$ en caso contrario.

Por su parte el atributo A, se traduce en la variable $A_i = 1$ si la i-ésima familia encuentra dificultades o muchas dificultades con los ingresos familiares, mientras que $A_i = 0$ en caso contrario.

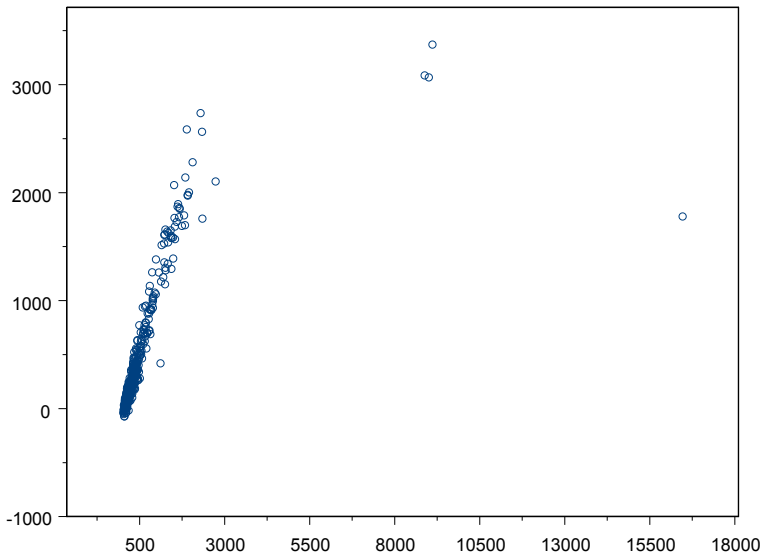
A.1.3. Población Lagos

En el estudio referido a los lagos estadounidenses, los métodos propuestos se evalúan numéricamente usando datos de la encuesta de lagos muestreada por el Programa de Evaluación y Vigilancia del Medioambiente realizado por la Agencia Estadounidense de Protección del Medioambiente. Estos conjuntos de datos proceden de 334 lagos con un total de medidas de 557. El objetivo es estimar la proporción de lagos en riesgo de acidificación. Las condiciones ecológicas de los lagos en Estados Unidos constituyen un aspecto muy importante analizado por la Agencia Estadounidense de Protección del Medioambiente. En particular, en la acidez del agua reside un gran interés (ver por ejemplo Pratesi, Ranalli y Salvati, 2008). Nosotros consideraremos las variables de interés relacionadas con la capacidad de neutralización del ácido, (ANC). Un valor de ANC menor que cero indica que el agua is ácida. Si los valores de ANC se aproximan a cero el lago pierde su capacidad de almacenamiento en el búfer. Los valores de ANC entre 50 y 200 identifican al lago en cuestión en situación de pre-alarma y en otras ocasiones críticas. Valores de ANC mayores de 500 reflejan un lago con un bajo riesgo de acidificación.

El estudio de simulación realizado consiste en considerar un total de 557 medidas de diferentes lagos procedentes de una población de la cual las muestras han sido seleccionadas bajo MAS. El interés reside en estimar la proporción de lagos cuyos valores de ANC sean menores de 0, y la proporción de lagos cuyos valores de ANC sean mayores de 200 ; 500. Tales proporciones son observadas aquí ya que existe un relevante interés en la práctica y a su vez nosotros deseamos estudiar los métodos propuestos para diferentes valores de P_A. Las proporciones poblacionales $P_A = 0{,}07$ para ANC< 0 ; $P_A = 0{,}22$ para ANC $>$

500 y $P_A = 0{,}44$ para ANC > 200 son las que se han analizado. Como se comentaba anteriormente, las proporciones mayores de 0.5 no son consideradas, se han estudiado para $P_A = 0{,}07, 0{,}22, 0{,}44$. Hemos usado la concentración de calcio en cada lago (CA) para determinar los atributos auxiliares. Para $P_A = 0{,}07$, consideramos $B_i = 1$ si el i-ésima lago tiene un valor de CA menor que 78 y por otra parte $B_i = 0$. Para $P_A = 0{,}22$, consideramos valores mayores que 189, 480 como atributos auxiliares. Para $P_A = 0{,}44$ consideramos valores de CA mayores de 654, 230 como atributos auxiliares. Nótese que se podrían haber usado como atributos auxiliares otros valores de CA o el uso de otras variables.

Figura A.2: Nube de puntos de la población Lagos.



A.1.4. Población ENS

Para la población referida a la Encuesta Nacional de Salud, los estimadores propuestos y los intervalos de confianza que proponemos se aplican a la estimación de prevalencias para las enfermedades crónicas de asma y alergia. El desarrollo de ambos métodos puede ser observado para proporciones pequeñas.

Este estudio está basado en una población de tamaño $N = 9063$ compuesta por datos reales extraídos de la Encuesta Nacional de Salud Española del año 2006, (ENS), por ejemplo nosotros consideramos los datos muestrales como una población en el cual las muestras se pueden extraer sin problema.

Las simulaciones realizadas nos conducen a una comparación del funcionamiento de varios estimadores e intervalos de confianza para los cuales es aplicable esta situación práctica. Los detalles más reseñables sobre la información contenida en la población ENS se describen a continuación. La población ENS asume un diseño muestral complejo donde las unidades de muestreo de la primera etapa son secciones censales (2236 secciones censales se han seleccionado en la muestra). Las unidades de la primera etapa se agrupan en estratos de acuerdo con el tamaño del municipio. Las unidades de la segunda etapa son las viviendas familiares principales. Dentro de las unidades de la segunda etapa no se lleva a cabo un submuestreo, y las viviendas con residentes se encuestan. Dentro de cada hogar, un adulto (de una edad igual o superior a 16 años) se selecciona y se completa el cuestionario de adultos, mientras que si hay menores, (de 0 a 15 años) en el hogar, uno de ellos se selecciona y se completa el cuestionario de los menores. Los datos usados en el estudio de simulación se refieren a los cuestionarios de menores. Se seleccionaron un total de 31300 hogares donde se rellenaron en total $N = 9063$ cuestionarios sobre menores.

Las bases de datos de estadística nacional de salud contienen datos de indicadores de salud, incluyendo indicadores básicos demográficos y socioeconómicos, algunos indicadores están relacionados con los estilos de vida, el entorno y los cuidados de la salud, utilización y gastos en medicación. La dolencia del asma crónica y la alergia se investigaron en la ENS. La Organización Mundial de la Salud reconoce que el asma es una de las mayores preocupaciones de la salud pública. La Organización Mundial de la Salud juega el rol de coordinadora internacional de todos los esfuerzos contra esta dolencia. El estudio internacional del asma y las alergias en niños y adolescentes reveló que las mismas siguen en crecimiento en los niños europeos. En estos momentos el objetivo básico es el uso alternativo de estimadores puntuales más precisos para las comentadas prevalencias y la construcción de intervalos de confianza con cobertura deseable y la mínima anchura. Los datos usados en las simulaciones se refieren a tres variables. La variable de interés (atributo A) indica si el niño ha sufrido de la dolencia (asma o alergia). Por otro lado la información auxiliar procede de la variable "ha recetado el médico medicamentos al niño para el asma o para la alergia?"(atributo auxiliar B_1) y "ha consumido el niño los medicamentos para la alergia o para el asma en las últimas dos semanas?"(atributo B_2) Nótese que $P_A = 0{,}07$, $P_{B_1} = P_{B_2} = 0{,}04$ $\phi_1 = 0{,}583$

y $\phi_2 = 0{,}570$ para el caso del asma y $P_A = 0{,}12$,$P_{B_1} = P_{B_2} = 0{,}03$ $\phi_1 = 0{,}510$ y $\phi_2 = 0{,}495$ para la alergia. El estimador propuesto $\widehat{p}_{r.e.}$ el cual está basado en una variable auxiliar, usa el atributo B_1 como información auxiliar.

Nótese que el Sistema de Información de la Salud del Sistema Nacional de Salud Española posee toda la información referida a los medicamentos prescritos, los cuales son clasificados por grupos terapéuticos, edad, sexo, dolencias, etc. La prescripción electrónica se implantará en un futuro muy cercano y su información será mayor. Por otro lado, los estudios llevados a cabo por agencias públicas o privadas tales como la del Colegio Oficial de Farmacéuticos o la Federación de Farmacéuticos Españoles también tienen información muy relevante en lo que respecta a las recetas médicas, especialmente en el área de salud infantil. Estos argumentos indican que los estimadores propuestos y los intervalos de confianza dados se pueden aplicar en estas situaciones, siempre y cuando la proporción poblacional de recetas médicas sea conocida. Por otra parte, la proporción de medicamentos consumidos se puede obtener del Sistema de Información de la Salud. Sin embargo, nuestro interés reside en el funcionamiento de los estimadores propuestos y los intervalos de confianza usados con datos reales.

A.2. Poblaciones simuladas

Un total de 30 poblaciones, de $N = 1000$ individuos cada una, fueron generadas para el estudio del efecto de diferentes situaciones sobre distintos estimadores de la proporción poblacional.

Las poblaciones se generaron de forma aleatoria a través de muestras de 1000 unidades obtenidas de una distribución de Bernoulli con parámetro $p = \{0.1, 0.25, 0.5, 0.75, 0.9\}$. Estos valores representan las correspondientes proporciones poblacionales P_A en cada población. Por su parte, los atributos auxiliares se generaron a partir de las poblaciones anteriores. Para ello se cambiaron aleatoriamente un determinado porcentaje de valores de tales poblaciones, de forma que los coeficientes V de Cramer entre los atributos de interés (las poblaciones anteriores) y los atributos auxiliares (las variables transformadas aleatoriamente) oscilan entre 0.5 y 0.9.