TRABAJO DE FIN DE GRADO

ESTIMACIÓN DE PROPORCIONES POBLACIONALES CON INFORMACIÓN AUXILIAR

(Proportional population estimation with auxiliar information)

Autor: Dª. Yolanda Rubio Fernández

Tutor/es: D. Sergio Martínez Puertas

Grado en Marketing e Investigación de mercados

Facultad de Ciencias Económicas y Empresariales

UNIVERSIDAD DE ALMERÍA

Curso Académico: 2010 / 2014

Almería, Julio de 2014

INDICE

La estimación de proporciones en poblaciones finitas tiene un especial interés en distintos ámbitos, tales como el Marketing. Pues en ocasiones, nos puede interesar la proporción de compradores que existe en una determinada población. El objetivo de este trabajo consiste en realizar una revisión de las técnicas existentes en la estimación de una proporción poblacional bajo un determinado diseño muestral empleado en la selección de la muestra, centrándonos principalmente, en aquellas técnicas que hacen uso de la información auxiliar disponible.

Finalmente, las técnicas descritas se emplearán en un estudio de simulación que nos permitirá analizar cuáles son las más adecuadas.

INTRODUCCIÓN

En el presente trabajo el objetivo que se persigue es la incorporación de información auxiliar en la estimación de una proporción poblacional cuando consideramos una muestra tomada de una población finita. La proporción poblacional de un atributo de estudio es un parámetro de gran interés en múltiples disciplinas, tales como la medicina, donde nos puede interesar la proporción de pacientes que mejoran su enfermedad mediante un nuevo tratamiento, la sociología, donde el interés puede recaer en la proporción de votantes de una determinada opción política o el marketing, donde el objetivo puede ser la proporción de consumidores satisfechos con una determinada marca comercial.

A pesar de ello, el uso de información auxiliar en el proceso de estimación de una proporción es un problema menos tratado que el caso de otros parámetros como la media, mediana, etc. En este trabajo se realiza una revisión de las principales técnicas existentes para incorporar información auxiliar disponible en la estimación de una proporción poblacional, centrándonos en el método de calibración y en el uso de modelos de regresión logística.

De este modo, el trabajo está estructurado de la siguiente manera. En el Capítulo 1, centrado en el marco teórico del trabajo, introducimos los conceptos básicos necesarios para abordar el problema de estimación de un parámetro poblacional en muestreo en poblaciones finitas, conceptos que ampliaremos a lo largo de dicho capítulo, donde se abordan los conceptos de espacio muestral y diseño muestral, que son de gran importancia a la hora de realizar estudios muestrales en poblaciones finitas. En este capítulo también se incluye el concepto de probabilidad de inclusión asociada a un determinado diseño muestral y la introducción de conceptos asociados al muestreo en poblaciones finitas. Para finalizar, dedicamos el último punto del capítulo a los conceptos de estadístico, estimador y sus propiedades. En el Capítulo II, metodología, introduciremos los diseños muestrales que vamos a considerar en el presente trabajo. Además, este capítulo está dedicado al objetivo central del trabajo, esto es, la incorporación de información auxiliar en el proceso de estimación de una determinada proporción poblacional mediante estimadores indirectos, centrándonos en un método nuevo y reciente como es el método de calibración y en estimadores basados en modelos de regresión logística. En el presente capítulo, también se estudiaran las propiedades teóricas de los estimadores propuestos. Finalmente, para completar el estudio teórico, en el Capítulo III, resultados, se incluye un estudio de

simulación con datos reales extraídos de la Encuesta de Presupuestos Familiares realizada por el Instituto Nacional de Estadística (INE) correspondiente al año 2012, donde se pone en práctica el uso de los estimadores analizados en el capítulo anterior y donde a través de los resultados obtenidos, podemos observar que la incorporación de información auxiliar mejora la eficiencia de las estimaciones de una proporción poblacional.

CAPÍTULO I. MARCO TEÓRICO

1. Definición de concepto de espacio muestral

Nuestra intención es tomar subconjuntos en *U* para obtener la información que nos permita realizar buenas inferencias. A cualquier subconjunto de *U* le llamaremos muestra. El conjunto de todas las posibles muestras será el conjunto de todos los subconjuntos de *U*, que representaremos por M_U, y cuyo cardinal es 2^N, pues consideramos que dos muestras son iguales si tienen los mismos elementos, y que en principio, no es necesario que los elementos estén repetidos pues sólo nos proporcionan una información. A este conjunto le llamaremos espacio muestral universal.

Como en general M_U es un conjunto muy extenso, consideraremos en nuestros estudios subconjuntos del mismo, $M \subseteq M_U$ que llamaremos espacio muestral, dependiendo del problema abordado. Los elementos de *M* se representarán por *m*.

El conjunto *M* puede tener diferentes propiedades, veamos algunas,

a) ser una partición de *U*. Entonces se denomina espacio muestral partición.
b) tener todas sus muestras con el mismo número de elementos. En tal caso, se denomina espacio muestral de tamaño fijo.
c) contener a cierto subconjunto de *U* en todas sus muestras. Entonces le llamaremos espacio muestral con elementos prefijados.

En cualquier caso, es deseable, y así se supondrá en el futuro, que toda unidad poblacional esté en al menos una muestra.

Por ejemplo, si tenemos la población $U=\{1,2,3\}$, podemos considerar varios espacios muestrales,

$M_1= \{(3,2),(2,1),(2),(1)\}$

$M_2= \{(3,2),(1)\}$ (partición).

$M_3= \{(1,2),(1,3),(2,3)\}$ (tamaño fijo n=2)

$M_4= \{(1,2),(1,3),(1)\}$ (con individuo prefijado u_1).

El número de muestras que tiene el espacio muestral suele llamarse tamaño del soporte muestra, y lo representamos por $n(M)$,

$$M = \{m_1, m_2, \dots, m_{n(M)}\}$$

El número de unidades de cada muestra m se denomina tamaño de la muestra o tamaño muestral y se representa por $n(m)$, tenemos pues,

$$m= \left\{u_{i1}, u_{i2}, \dots, u_{i_{n(m)}}\right\}$$

cuando el tamaño es fijo lo representamos simplemente por n.

1.1. Concepto de diseño muestral

Escogido el espacio muestral, $M = \left\{m_1, m_2, \dots, m_{n(M)}\right\}$, hay que indicar el modo de elegir dichas muestras. Como antes indicábamos, esta elección no se hará de un modo caprichoso sino regida por las leyes del azar. Para ello basta definir una ley de probabilidad discreta sobre *M*,

$$p(\cdot): M \longrightarrow [0,1]$$

tal que,

$$\text{p(m)} > 0 \qquad \forall m \in M$$

$$\sum_{m \in M} p(m) = 1$$

Ejemplo (Diseño muestral de tamaño fijo uniforme)

Conocido también como muestreo aleatorio simple, lo denotaremos con la abreviatura MAS (*N*, n). En este diseño, el espacio muestral lo constituyen todas las muestras de M_U que tienen tamaño fijo n(m) =n. Este espacio muestral, M^n, tiene un tamaño del soporte igual a $\binom{N}{\text{n}}$. La distribución de probabilidad sobre dicho espacio muestral es uniforme, esto es,

$$\text{p(m)} = \frac{1}{\binom{N}{\text{n}}} \qquad \forall m \in Mn$$

2. <u>Probabilidades de inclusión</u>

Como indicábamos anteriormente, toda unidad de la población debe de estar en alguna muestra, pero a veces deseamos que un individuo esté más veces que otro en las diferentes muestras que forman el espacio muestral, o deseamos escoger la muestra a través de la elección de individuos, para ello es necesario conocer las denominadas probabilidades de inclusión.

Sea $m \in M$, dada una unidad u_k, puede que pertenezca o no a dicha muestra. Para representar esta situación definimos la variable indicador de pertenencia a la muestra como una aplicación,

$$I_k : M \mapsto \{0, 1\}$$

tal que,

$$\forall m \in M \quad y \quad k \in U \qquad\qquad I_k(m) = \begin{cases} 1 & si\ u_k \in m \\ 0 & en\ otro\ caso \end{cases}$$

I_k es una variable aleatoria definida sobre el diseño muestral (M, p (·)) y su distribución viene dada por,

$$P\{I_k = 0\} = 1 - P\{I_k = 1\} \tag{1.1}$$

$$P\{I_k = 1\} = \sum_{\substack{m \in M \\ k \in m}} p(m) \triangleq \pi_k \tag{1.2}$$

Así, π_k es la probabilidad de que el elemento *k* esté en la muestra resultante del mencionado experimento aleatorio, y se denomina probabilidad de inclusión de primer orden. Como todo elemento u_k, debe estar en al menos una muestra del diseño, ha de verificarse que $\pi_k > 0$, $\forall k \in U$. Cuando el diseño verifica esta condición el muestreo correspondiente se denomina muestreo probabilístico.

Las propiedades de las probabilidades de inclusión son:

$$E[I_k] = \pi_k$$

$$V[I_k] = \pi_k (1 - \pi_k) \tag{1.3}$$

También tiene especial interés el indicador I_{kl} definido de modo análogo, pero sobre dos elementos muestrales,

$$I_{kl}(m) = \begin{cases} 1 & si\ u_k, u_1 \in m \\ 0 & en\ otro\ caso \end{cases}$$

Se verifica que $Ikl = Ik \cdot Il$ aunque no son independientes. Nuevamente I_{kl} es una variable aleatoria de Bernoulli, *B*e (π_{kl}), siendo,

$$\pi_{kl} = P\{I_{kl} = 1\} = \sum_{\substack{m \in M \\ k,l \in m}} p(m) \tag{1.4}$$

Las cantidades π_{kl} reciben el nombre de probabilidades de inclusión de segundo orden.

Observemos que,

$$Cov\,[I_k\,, I_l]\;=\;E[I_k I_i]\,-\,E[I_k]\,E[I_l]=\;\pi_{kl}\;-\,\pi_k\;\pi_l \qquad (1.5)$$

a esta cantidad se le representa por Δ_{kl}, y aparece frecuentemente en las expresiones de los estimadores y sus errores.

Un diseño deberá verificar, además de $\pi_k\;>0,\;\forall_k \in U$, la condición adicional,

$$\pi_{kl} > 0 \qquad \forall_k \neq l\;\in U$$

pues como ya veremos, para estos diseños podemos dar estimaciones válidas de la varianza, por ello, un muestreo con estas características recibe el nombre de muestreo cuantificable.

3. Estadísticos, estimadores y propiedades básicas

En Teoría se emplea el término estadístico para referirse a una función real cuyos valores varían con las diferentes realizaciones de un experimento. Al estar valorada sobre todos los resultados posibles del experimento, y depender estos del azar, podemos considerar la correspondiente distribución del mismo.

Cuando el estadístico se emplea para producir valores que, para la mayoría de las muestras, están "próximos" a un parámetro desconocido de la población θ(Y) que se desea estimar, recibe el nombre de estimador.

En este caso, el diseño muestral guarda una estrecha relación con el estimador $e(\cdot)$ del parámetro θ(Y) que deseamos estimar, pues el proceso en la práctica será,

1. Escoger una muestra m ∈ M, según la ley de probabilidad $p(\cdot)$.
2. Dada la muestra *m* a través de sus unidades, conocer las valoraciones de las variables en estudio para dicha muestra.
3. Hacer una valoración (estimación) del parámetro θ(Y) a través de la valoración que sobre los valores muestrales efectúe $e(\cdot)$.

   Como para cada muestra m ∈ M, el estimador nos proporciona un valor, que representamos por $e(m)$ de forma resumida, el cual cambiará de muestra a muestra, y dado que cualquiera de ellos puede tomarse como valoración de θ(Y), que es única, parece que debemos de pedirle a la función $e(\cdot)$ ciertas propiedades para que sus valoraciones estén "próximas" al verdadero valor de θ(Y).

Al ser $e(\cdot)$ es una variable aleatoria, tendrá una cierta distribución heredada de $p(\cdot)$ y que llamaremos distribución del estimador en el diseño muestral o, brevemente, distribución del estimador en el muestreo.

Al poder elegir entre muchas posibles funciones, emplearemos aquellas cuyos valores estén más cerca del verdadero valor y posea menores desviaciones y aunque estos requerimientos estén formulados un poco difusamente, la estadística ha estudiado durante décadas estos problemas, cuantificando estas ideas a través del concepto de insesgadez y de error cuadrático medio, disponiendo hoy día de buenos estimadores $e(\cdot)$ para la mayoría de los diseños muestrales y para diferentes parámetros θ(Y).

La distribución en el muestro del estimador $e(\cdot)$ es conocida sólo teóricamente pues en la práctica desconoceremos los valores de la población ($Y_1,\ldots, Y_N$), por ello los parámetros asociados a dicha ley: media, varianza, etc, sólo pueden darse en teoría, y frecuentemente son funciones de los parámetros asociados a la población, por lo que solamente pueden usarse como indicadores para el comportamiento de dichos estimadores. Los parámetros asociados a $e(\cdot)$ como variable aleatoria se determinan del modo usual.

Propiedades de un estimador:

El sesgo de un estimador $\hat{\theta}$ del parámetro θ se define como,

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta \tag{1.6}$$

Diremos que un estimador es insesgado para θ si verifica $B(\hat{\theta}) = 0$ cualquiera que sea el vector poblacional ($Y_1,\ldots, Y_N$).

El error cuadrático medio de $\hat{\theta}$ respecto a θ se define como,

$$\text{ECM}(\hat{\theta}) = \text{E}[(\hat{\theta} - \theta)^2] \tag{1.7}$$

Si se desarrolla la expresión se obtiene,

$$\text{ECM}(\hat{\theta}) = \text{V}[\hat{\theta}] + [\text{B}(\hat{\theta})]^2 \tag{1.8}$$

La anterior igualdad asegura que si el estimador $\hat{\theta}$ es insesgado, entonces verifica,

$$\text{ECM}[\hat{\theta}] = \text{V}[\hat{\theta}] \tag{1.9}$$

Notemos que la insesgadez expresa que la distribución del estadístico se dispersa entorno del verdadero valor del parámetro, y el error cuadrático medio nos indica el grado de dispersión de esta distribución entorno del verdadero valor del parámetro.

Los estimadores que se suelen emplear en la teoría de muestras son estimadores insesgados o aproximadamente insesgados, en el sentido de no ser muy importante el sesgo cuando las muestras son de un tamaño grande.

Entre los diferentes estimadores de un parámetro, $\hat{\theta}_1$, $\hat{\theta}_2$,… se escogerá aquel que sea menos disperso en el sentido de tener un menor error cuadrático medio. Si los estimadores fuesen centrados, la comparación es más precisa y esta se efectúa a través de las varianzas.

A veces es importante emplear el coeficiente de variación del estimador $\hat{\theta}$ que se define como,

$$CV\,[\hat{\theta}] = \frac{\sqrt{V[\hat{\theta}]}}{E[\hat{\theta}]} \qquad (1.10)$$

que da una medida más real de la precisión del estimador, al dividir el error por la medida que estamos empleando. Los efectos que los diseños muestrales pueden tener sobre un estimador, $\hat{\theta}$, de un parámetro, pueden indicarse a través del cociente que el coeficiente de variación tenga en ambos diseños.

Finalmente, destaquemos que aunque la dispersión del estimador $\hat{\theta}$ está dada, al menos teóricamente, a través de su sesgo y su error cuadrático medio, basándonos en ellos se introduce el concepto de intervalo de confianza, que representa un intervalo aleatorio de la recta real, al estar definido sobre cada muestra, que tiene una probabilidad establecida (usualmente próxima a la unidad) de contener al verdadero valor del parámetro,

$$IC(m) \;=\; [e_I(m), e_S(m)] \qquad (1.11)$$

Donde $e_I(\cdot)$ y $e_S(\cdot)$ son dos estadísticos dados, tales que $e_I(\mathrm{m}) \leq e_S(m),\ \forall m \in M$, y dan los extremos inferior y superior del intervalo de confianza, en función de la muestra m elegida.

La probabilidad $P[\theta \in IC\ (m)]$ se conoce de antemano, y recibe el nombre de nivel de confianza. La probabilidad $1 - P[\theta \in IC\ (m)]$ representa el riesgo que corremos de equivocarnos, es decir, la probabilidad de que el parámetro no esté en el intervalo. Como

dicho intervalo se determina a partir de la muestra m, y debe depender de los parámetros $E[\hat{\theta}]$ y V$[\hat{\theta}]$, suelen usarse procedimientos aproximados.

Si la distribución del estimador $\hat{\theta}$ es aproximadamente normal, con E$[\hat{\theta}] = \theta$, y varianza V$[\hat{\theta}]$, y conocemos un estimador consistente V$[\hat{\theta}]$, para V$[\hat{\theta}]$, se puede dar como intervalo de confianza para θ, al nivel de confianza 1 – α, el definido por,

$$e_I(\mathrm{m}) = \hat{\theta}(\mathrm{m}) - z_{1-\alpha/2} \cdot \sqrt{\hat{V}[\,\hat{\theta}]} \tag{1.12}$$

$$e_S(\mathrm{m}) = \hat{\theta}(\mathrm{m}) + z_{1-\alpha/2} \cdot \sqrt{\hat{V}[\,\hat{\theta}]} \tag{1.13}$$

que usualmente denotaremos por,

$$\hat{\theta}(\mathrm{m}) \pm z_{1-\alpha/2} \cdot \sqrt{\hat{V}[\,\hat{\theta}]} \tag{1.14}$$

Donde $z_{1-\alpha/2}$ es el cuantil correspondiente, para una variable aleatoria $N(0, 1)$.

CAPÍTULO II. METODOLOGÍA

4. Diseños muestrales empleados

En el presente trabajo, destinado a la estimación de proporciones, se van a emplear diversos diseños muestrales en el estudio de simulación incluidos en el trabajo, para analizar de una manera más adecuada las propiedades de los estimadores analizados en el trabajo.

Los diseños muestrales considerados en el presente trabajo son:

- Muestreo Aleatorio Simple (MAS)
- Muestreo Sistemático
- Muestreo Estratificado
- Muestreo con probabilidades de inclusión fijadas (Método de Midzuno)

- **Muestreo Aleatorio Simple (MAS)**

El muestreo aleatorio simple es uno de los diseños más estudiados, y dadas las propiedades que posee frente a las estimaciones de parámetros y de errores de muestreo, es un diseño bastante utilizado.

Como ya hemos comentado en apartados anteriores, en este diseño el espacio muestral, *M*, está formado por todas las muestras de tamaño n fijo, y la distribución de probabilidad $p(\cdot)$ definida en *M* es la ley uniforme. Por ello, puede definirse como,

$$p: M_U \mapsto [0, 1] \text{ tal que } p(m) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{si } n(m) = n \\ 0 & \text{si } n(m) \neq n \end{cases} \quad (2.1)$$

Parámetros

Los parámetros asociados a este diseño muestral son los siguientes,

$$\pi_i = \frac{n}{N} = f \qquad i = 1, \dots, N \quad (2.2)$$

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)} \qquad i \neq j = 1, \dots, N \quad (2.3)$$

Por lo que es un diseño muestral cuantificable. Y dado que es de tamaño fijo, y que,

$$\Delta_{ij} = \pi_{ij} - \pi_i \pi_j = -\frac{f(1-f)}{N-1} \qquad i \neq j \quad (2.4)$$

$$\Delta_{ii} = f\,(1 - f) \quad (2.5)$$

podemos fácilmente tener estimadores no negativos del error.

- **<u>Muestreo Sistemático</u>**

Hasta ahora hemos considerado una población $U = \{u_1, \dots, u_N\}$ en la cual sus unidades están perfectamente identificadas. Imaginemos la población formada por los asistentes a un teatro, si deseamos hacer una encuesta sobre las bondades de la representación, ciertamente, tenemos numerados los individuos que componen la misma por su billete de entrada pero al realizar la elección de la muestra en base a él puede ser difícil escoger los individuos que nos da el mecanismo de azar. Además, los individuos son ese día una

representación de los diferentes individuos que pueden acceder a la función a lo largo del tiempo que la obra esté en cartel.

En estos casos, se suele tomar la muestra usando el denominado muestreo sistemático, que trata de escoger los individuos de la población, para pertenecer a la muestra, de un modo directo, seleccionando las unidades mediante una regla sistemática (o automática), partiendo de una elección inicial o primaria. Esta regla sistemática nos da siempre la misma muestra cuando se parte de la misma unidad primaria por ello las muestras pertenecientes a estos diseños son excluyentes entre sí.

Así, a la salida del teatro podemos fijar la regla de preguntar a una de cada diez personas, partiendo del tercer individuo que salga inicialmente. Con ello podemos escoger muestras de la población.

Observemos, no obstante, que además de los inconvenientes que presenta el que las muestras sean excluyentes, en el plano teórico, es importante la regla de elección de la muestra, ya que si la población presenta algún tipo de tendencia puede ocurrir que la muestra sistemática no sea capaz de recogerla. Por tanto, cuando la población está distribuida al azar esta regla sistemática puede dar resultados similares a los que da el muestreo aleatorio simple, siendo más fácil la elección de la muestra.

Un diseño muestral sistemático es un diseño muestral cuyo espacio muestral está formado por muestras sistemáticas, es decir, las unidades muestreadas se escogen de la población mediante la elección de una unidad de partida y las demás por una regla sistemática de selección.

El diseño sistemático más sencillo, que denotaremos por $MS(N,k)$ es el diseño sistemático uniforme de paso k. Se basa en, partiendo de una lista con las unidades de la población,

$$\{1,2,\dots,k,\dots,N\}$$

en aplicar el siguiente procedimiento,

1. Se escoge un número aleatorio entero $\gamma \sim \mathcal{U}[1..k]$ siendo $k \leq N$ un entero determinado de antemano. El valor de γ nos indica la unidad inicial a seleccionar.
2. Se construye la muestra,

$$m_\gamma = \{\gamma, \gamma + k, \gamma + 2k \dots\} \tag{2.6}$$

hasta agotar la población.

Si consideramos la siguiente representación,

$$N = kn + r, 0 \leq r < k \tag{2.7}$$

tendremos que las muestras son de tamaño n si $r = 0$. Por el contrario, si $r > 0$ entonces el tamaño muestral no es fijo, siendo n si $\gamma \leq r$.

En cualquier caso, existen solo k muestras posibles, que cubren toda la población y que son mutuamente excluyentes.

Las probabilidades de inclusión son fáciles de calcular, y vienen dadas por,

- De primer orden.

$$\pi_i = \frac{1}{k} \quad \forall_i \in U \tag{2.8}$$

luego es un muestreo probabilístico.

- De segundo orden.

$$\pi_{ij} = \begin{cases} 1/k & \text{si } i,j \text{ pertenecen a la misma muestra} \\ 0 & \text{en otro caso} \end{cases}$$

Al no ser $\pi_{ij} > 0, \forall i,j$, no permite una estimación inmediata del error de muestreo, es decir, no es un muestreo cuantificable. Otro inconveniente que presenta este diseño es el poco control que se tiene del tamaño muestral. En efecto, ya hemos visto que dado un valor de k, el tamaño de la muestra puede ser n ó $n + 1$. Pero además, el valor de n puede variar mucho según el k que se tome.

- **<u>Muestreo Estratificado</u>**

La ciencia utiliza la homogeneización por bloques para obtener mayor precisión en sus estudios, pues de este modo puede controlar ciertos aspectos del experimento, para conocer la influencia de los mismos en los resultados. Análogamente, se puede realizar en la Teoría del Muestreo, pues si muestreamos en una población muy heterogénea, se necesita un gran esfuerzo muestral para obtener cierta precisión, muestras que si la población está dividida en grupos (bloques o estratos) internamente homogéneos, el esfuerzo de cada grupo será mínimo, resultando globalmente un esfuerzo menor.

Así por ejemplo, si deseamos obtener una muestra en la Universidad de Almería para conocer la media de horas de estudio que los alumnos han realizado en la pasada semana, problemática (exámenes, horarios de clases, etc) por lo que la muestra será pequeña en cada curso. Pero si deseamos cierta precisión y muestreamos globalmente, la muestra será de gran tamaño para que pueda coger todas las tendencias que existen en las Facultades.

Además de las ventajas de esta homogeneización para reducir el error de muestreo, ya que a un mismo esfuerzo muestral debe corresponder mayor precisión, existen otras ventajas secundarias que aconsejan la división de la población en estratos:

- Facilidad de manejo de los marcos de los estratos.
- Procedimientos muestrales más ágiles en cada estrato.
- Mejor empleo de informaciones especiales.

En principio es fácil definir una población estratificada, y establecer el espacio muestral y las probabilidades asociadas, a partir de los diseños muestrales en cada estrato. Esta división de la población en estratos deberá realizarse con algún patrón de homogeneidad para que dicha división aporte alguna información útil, que produzca una disminución en el error de las estimaciones.

Una población U se dice que está estratificada en L estratos, $U_1, U_2, \dots, U_L$, si estos forman partición de la misma, es decir,

- $U = \cup_{i=1}^{L} U_i$
- $U_i \cap U_j = \emptyset, \ \forall i \neq j$

Llamaremos N_h al número de individuos del estrato U_h, y supondremos $N_h \geq 2$ para evitar casos triviales. Se obtiene pues $\sum_h N_h = N$. A los elementos de la población que están en el estrato U_h los representamos por,

$$u_{h1}, u_{h2}, \dots, u_{hNh}$$

y al cociente N_h/N se le llama peso del estrato U_h, representándose por W_h.

Una vez dividida la población en estratos, no se define sobre ella un diseño muestral directamente, sino de un modo indirecto, pues las muestras se escogen de modo independiente en cada uno de los estratos, agrupándose posteriormente para producir las muestras sobre U, siendo esta posibilidad de muestrear en cada estrato la gran ventaja de la división de la población.

Por ello, se define en cada estrato U_h un diseño muestral $d_h = (M_h, p_h(\cdot))$, independientemente de los demás estratos, obteniéndose en él la muestra m_h con probabilidad $p_h(m_h)$, estando m_h compuesta por elementos de U_h únicamente.

Si el tamaño de la muestra m_h es n_h, el tamaño de la muestra m será $n = \sum_h n_h$. Al cociente n_h/N se le denomina fracción de muestreo o tasa de muestreo del estrato U_h, y se le representa por f_h.

Observemos que π_i, la probabilidad de que el individuo u_i esté en la muestra m, es igual a $\pi_i^{d_h}$, probabilidad de inclusión en la muestra m_h si el elemento u_i está en el estrato U_h.

Análogamente, π_{ij} es igual a $\pi_{ij}^{d_h}$, si ambos elementos están en U_h. Pero si están en estratos diferentes, U_h y U_k, entonces,

$$\pi_{ij} = \pi_i^{d_h}\pi_j^{d_k} \tag{2.9}$$

debido a la independencia que existe en la elección de la muestra.
Cuando no exista posibilidad de confusión, nos referiremos a la matriz del diseño II = (π_{ij}) sin utilizar superíndices.
Por consiguiente, definida la estratificación de la población y el diseño muestral en cada estrato, tenemos establecido el diseño muestral de la población completa, y trabajamos con él en el modo usual para realizar la estimación de parámetros, aunque como veremos, nos conduce a una mixtura de las estimaciones que podríamos realizar sobre los parámetros en cada estrato, como si fuera una población independiente. Asimismo, obtendremos estimación de los errores globales a partir de los errores en los estratos.

Diseño muestral aleatorio simple estratificado

En el caso de que en cada estrato se realice un muestreo $MAS(N_h, n_h)$, se dice que el diseño muestral es aleatorio simple estratificado, y se denota $MASE(N, n, L, \{N_h, n_h\}_{h=1,\dots,L})$. Este diseño es uno de los más empleados pues la estratificación suple a veces el empleo de probabilidades variables, por ello vamos a calcular los estimadores y sus errores.
Para el estrato U_h tendremos,

$$\pi_i^h = \frac{n_h}{N_h}, \qquad \pi_{ij}^h = \frac{n_h(n_h-1)}{N_{h(N_{h-1})}} \tag{2.10}$$

A lo largo del presente trabajo, cuando se haga mención al muestreo estratificado, entenderemos que estamos hablando del muestreo aleatorio simple estratificado.

- **Muestreo con probabilidades de inclusión fijadas (Método de Midzuno)**

Hasta ahora hemos impuesto para nuestro problema un diseño muestral en relación al cual estudiamos el estimador que mejor se adecua al parámetro que deseamos estimar, viendo las ventajas e inconvenientes de llevar el mismo a la práctica. Así, hemos estudiado los diseños muestrales aleatorio simple y sistemático.

Sin embargo, partiendo de la base de que vamos a emplear un diseño muestral cuantificable, podemos intentar determinarlo de manera que tengamos menor error de muestreo.

Teniendo presente la expresión de la varianza de los estimadores lineales, podemos optar por dos caminos, actuando en ambos sobre las probabilidades de inclusión:

a) Buscar diseños "equivalentes" a uno dado.
b) Hacer $\pi_k \propto Y_k$.

El problema en ambos casos es determinar un diseño muestral que tenga las probabilidades de inclusión deseadas, y que sea fácil de llevar a la práctica.

El primer camino (a) conduce a los llamados diseños óptimos, pues mediante técnicas de programación matemática pueden determinarse diseños muestrales $(M, p(\cdot))$ tales que tengan alguna propiedad deseable para las muestras pertenecientes al espacio o incluso para las propiedades que se le asignan. La ventaja que ofrece esta metodología es la de poder proporcionar diseños equivalentes, con gran facilidad.

El segundo camino (b) conduce al empleo de información adicional para la determinación de las probabilidades de inclusión, pues de otro modo es utópico poder fijar los valores de la variable que deseamos investigar.

Supongamos que para la población en estudio, $U = \{1,2, \dots, N\}$, conocemos los valores de una variable, $X = (X_1, X_2, \dots, X_N)$, que denominamos auxiliar, la cual creemos que está fuertemente correlacionada con la variable Y, que deseamos estudiar. Si por ejemplo, queremos estudiar el número medio de asignaturas cursadas por los alumnos en el presente curso, podemos tomar como variable auxiliar las asignaturas a las que se presentó de manera efectiva el curso anterior.

De este modo, si nuestro diseño muestral es de tamaño fijo, n, podemos suponer $\pi_k \propto X_k$, y dado que,

$$\sum_{k=1}^{N} \pi_k = n \qquad \text{con} \qquad 0 < \pi_k \leq 1, \ k = 1,2, \dots, N \tag{2.11}$$

ha de ser,

$$\pi_k = \frac{nX_k}{\sum_{k=1}^{N} X_k} \qquad k = 1,2, \dots, N \tag{2.12}$$

Donde X_k debe verificar,

$$0 < nX_k \leq \sum_{k=1}^{N} X_k \qquad k = 1,2,\dots,N \tag{2.13}$$

para que los valores π_k sean aceptables. Esta suposición se mantendrá a lo largo de nuestro estudio.

En caso de que la variable Y_k esté muy dispersa, es posible que X_k también lo esté, por lo que puede ser difícil que se verifique la condición anterior para todos los valores de k. En este caso, se puede incluir en la muestra a todos aquellos individuos que no verifiquen la misma, es decir, hacer π_k=1 para ellos, y realizar el muestreo sobre los restantes elementos. En caso de que hubiera un gran número de individuos en relación a la muestra, se puede dividir la población en grupos homogéneos en relación al tamaño de la variable X, y realizar el muestreo en cada grupo.

A aquellos diseños muestrales cuyas probabilidades de inclusión de primer orden verifiquen la condición,

$$\pi_k \propto X_k \qquad \forall k \in U$$

se les denominará diseños muestrales con probabilidades de inclusión proporcionales al tamaño, denotándolos como IIPS(n, N, X), o simplemente IIPS.

A continuación se imponen las propiedades del diseño muestral,

1. Las probabilidades de inclusión de segundo orden sean de fácil obtención, y verifiquen las probabilidades,

$$\pi_{kl} > 0, \qquad \forall k, l \tag{2.14}$$

$$\pi_{kl} - \pi_k \pi_l < 0, \qquad \forall k, l, k \neq l \tag{2.15}$$

para asegurar la existencia de buenos estimadores del error de muestreo.

2. El diseño muestral sea de fácil manejo, es decir, que podamos obtener las muestras de un modo sencillo, aplicando un proceso probabilístico de poco cálculo.

A pesar de la dificultad que presentan estos métodos, tienen la ventaja de proporcionar generalmente menor error de muestreo y buenas estimaciones del parámetro.

No obstante, existen otras estrategias en el planteamiento del problema, que conducen a otros diseños muestrales para diferentes tipos de estimador. Muchas de estas alternativas se basan en el muestreo con reemplazamiento, por lo que se fija a priori la probabilidad p_k de escoger al elemento k-ésimo en cualquier etapa, tomándose comúnmente $p_k \propto X_k$, por lo

que reciben el nombre de diseños muestrales con probabilidad de elección proporcional al tamaño, $\mathrm{PPS}(N, n, X)$. Estos métodos son de fácil realización pero tienen la dificultad de dar estimaciones del error menos precisas que los anteriores.

Método de Midzuno (1952)

El siguiente esquema descrito, expone la obtención de una muestra usando probabilidades proporcionales al tamaño. Para extraer una muestra de tamaño n se sigue en dos pasos,

1. El primer elemento se extrae con probabilidad variable,

$$\alpha_i = \frac{X_i}{T(X)} \qquad i = 1,2,\dots,N \tag{2.16}$$

2. Los restantes elementos se extraen mediante un diseño $MAS(N-1, n-1)$, a partir de la población obtenida eliminando el elemento extraído en el paso anterior.

   Con ello,

$$\pi_i = \alpha_i + (1-\alpha_i)\tfrac{n-1}{N-1} = \tfrac{N-n}{N-1}\alpha_i + \tfrac{n-1}{N-1} \tag{2.17}$$

$$\pi_{ij} = \alpha_i \frac{n-1}{N-1} + \alpha_j \frac{n-1}{N-1} + \left(1-\alpha_i-\alpha_j\right)\frac{n-1}{N-1}\frac{n-2}{N-2} = \tag{2.18}$$

$$\frac{n-1}{N-1}\left[\frac{N-n}{N-2}\left(\alpha_i+\alpha_j\right)+\frac{n-2}{N-2}\right], \qquad i \neq j$$

Y con estos valores se tiene,

$$\Delta_{ij} = -\frac{N-n}{(N-1)^2}\left[(N-n)\alpha_i\alpha_j + (1-\alpha_i-\alpha_j)\frac{n-1}{N-2}\right] < 0 \tag{2.19}$$

5. Estimadores de calibración

5.1. Definición del estimador de calibración

Dada una muestra s de tamaño n extraída de una población finita $U = \{1,2,\dots,N\}$ de tamaño N, mediante un determinado diseño muestral d, con probabilidades de inclusión de primer y segundo orden π_k y π_{kl} que consideraremos estrictamente positivas. Dado un atributo de estudio A definido en la población U, de forma que A_k=1 cuando una unidad k de la población U presenta el atributo A y A_k=0 en otro caso. La proporción de la población que presenta el atributo A en la población U viene dado por:

$$P_A = \frac{1}{N}\sum_{k\in U} A_k \tag{2.20}$$

Para su estimación, podemos considerar el estimador Horvitz- Thompson dado por

$$\hat{P}_{AH} = \frac{1}{N}\sum_{k\in U} d_k A_k \tag{2.21}$$

donde $d_k = 1/\pi_k$.

Si consideramos un atributo auxiliar B en el que el valor B_k es conocido para toda unidad k en la muestra s y de forma que P_B también es conocido, resulta que el estimador de Horvitz-Thompson no puede incorporar la información proporcionada por el atributo B en el proceso de estimación de la proporción del atributo A. Una manera de incorporar información auxiliar en el proceso de estimación es mediante el reemplazamiento de los pesos básicos d_k por nuevos pesos ω_k, usando técnicas de calibración.

La calibración es un método reciente que sirve para la incorporación de información auxiliar (Särndal, 2007) por las siguientes razones:

- Proporciona una forma sistemática de incorporación de información auxiliar disponible.
- Es un medio de obtener estimaciones consistentes para las variables auxiliares con totales conocidos.
- Es usado por agencias de estadística para estimar diferentes parámetros de poblaciones finitas. Varias agencias nacionales de estadística tienen softwares desarrollados y diseñados para calcular medidas de calibración basadas en información auxiliar disponible en registros de población y otras fuentes. Dichas agencias incluyen CLAN (Estadística de Suecia) y BASCULA (Oficina Central de Estadística, Países Bajos).

Siguiendo a Deville y Särndal (1992), para obtener un estimador de calibración para el atributo A basado en el atributo B, calculamos los nuevos pesos ω_k minimizando la distancia chi-cuadrado

$$\chi = \sum_{k\in s} \frac{(\omega_k - \omega_k)^2}{d_k q_k} \tag{2.22}$$

sujeto a la condición

$$P_B = \frac{1}{N}\sum_{k\in U} B_k = \frac{1}{N}\sum_{k\in s} \omega_k B_k \tag{2.23}$$

donde q_k son constantes positivas conocidas no relacionadas para d_k y $0 < P_B < 1$.

Al minimizar

$$\chi = \sum_{k\in s} \frac{(\omega_k - \omega_k)^2}{d_k q_k} \tag{2.24}$$

sujeto a la condición

$$P_B = \frac{1}{N}\sum_{k\in U} B_k = \frac{1}{N}\sum_{k\in s} \omega_k B_k \tag{2.25}$$

los nuevos pesos ω_k vienen dados por:

$$\omega_k = d_k + \frac{\lambda d_k q_k B_k}{N} \tag{2.26}$$

donde λ denota al multiplicador de Lagrange, dado por

$$\lambda = \frac{N^2\left(P_B - \hat{P}_{BH}\right)}{\sum_{k\in s} d_k q_k B_k} \tag{2.27}$$

y $\hat{P}_{BH}$ es el estimador Horvitz- Thompson usual para el atributo B.
Con los nuevos pesos de calibración (2.26) y asumiendo que $\sum_{k\in s} d_k q_k B_k \neq 0$, el estimador obtenido es:

$$\hat{P}_{AW} = \frac{1}{N}\sum_{k\in s} \omega_k A_k = \hat{P}_{AH} + \frac{(P_B - \hat{P}_{BH})}{\sum_{k\in s} d_k q_k B_k} \cdot \sum_{k\in s} d_k q_k B_k A_k \tag{2.28}$$

Por (2.25), cuando el nuevo estimador es aplicado para estimar la proporción poblacional del atributo B, coincide con P_B.

5.2. Propiedades del estimador propuesto

Siguiendo a Deville and Särndal (1992), podemos demostrar que el estimador $\hat{P}_{AW}$ es asintóticamente insesgado para P_A y su varianza asintótica dada por

$$AV\left(\hat{P}_{AW}\right) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U} \Delta_{kl}\,(d_k E_k)(d_l E_l) \tag{2.29}$$

donde

$$\Delta_{kl} = \pi_{kl} - \pi_k \pi_l;\;\; E_k = A_k - D \cdot B_k \tag{2.30}$$

y

$$D = \frac{\sum_{k\in U} q_k B_k A_k}{\sum_{k\in U} q_k B_k} \tag{2.31}$$

Un estimador para esta varianza es

$$\hat{V}\left(\hat{P}_{AW}\right) = \sum\nolimits_{k\in s}\sum\nolimits_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}}(d_k e_k)(d_l e_l) \tag{2.32}$$

con $e_k = A_k - B_k \cdot \widehat{D}$ and $\widehat{D} = \frac{\sum_{k\in s} d_k q_k B_k A_k}{\sum_{k\in s} d_k q_k B_k}$

Ejemplo. Bajo SRSWOR y $q_k = 1$ para todo $k \in U$ el estimador $\hat{P}_{AW}$ es:

$$\hat{P}_{AW} = \hat{p}_A + (P_B - \hat{p}_B) \cdot \frac{\hat{p}_{AB}}{\hat{p}_B} \tag{2.33}$$

donde

$$\hat{p}_A = \frac{1}{n}\sum_{k\in s} A_k; \quad \hat{p}_B = \frac{1}{n}\sum_{k\in s} B_k \quad \text{y} \quad \hat{p}_{AB} = \frac{1}{n}\sum_{k\in s} A_k B_k \tag{2.34}$$

Según Rueda et al (2007), el estimador $\hat{P}_{AW}$ tiene el mismo comportamiento asintótico que:

$$\hat{P}_{AVW} = \hat{p}_A + (P_B - \hat{p}_B)\cdot D = \hat{p}_A + (P_B - \hat{p}_B)\cdot \frac{P_{AB}}{P_B} \text{ con } P_{AB} = \frac{1}{n}\sum_{k\in s} A_k B_k \tag{2.35}$$

Así

$$AV\left(\hat{P}_{AW}\right) = V\left(\hat{P}_{AVW}\right) = V(\hat{p}_A) + D^2 V(\hat{p}_B) - 2DCov(\hat{p}_A, \hat{p}_B) =$$
$$= \frac{(1-f)}{n}\frac{N}{N-1}\left[P_A Q_A + \left(\frac{P_{AB}}{P_B}\right)^2 \cdot P_B Q_B - 2\left(\frac{P_{AB}}{P_B}\right)(P_{AB} - P_A P_B)\right] \tag{2.36}$$

donde $Q_A = 1 - P_A$; $Q_B = 1 - P_B$ and $f = \frac{n}{N}$. Esta varianza puede ser estimada por

$$\hat{V}\left(\hat{P}_{AW}\right) = \frac{1-f}{n-1}\left[\hat{p}_A\hat{q}_A + \left(\frac{\hat{p}_{AB}}{\hat{p}_B}\right)^2 \cdot \hat{p}_B\hat{q}_B - 2\left(\frac{\hat{p}_{AB}}{\hat{p}_B}\right)(\hat{p}_{AB} - \hat{p}_A\hat{p}_B)\right] \tag{2.37}$$

con

$$\hat{q}_A = \frac{1}{n}\sum_{k\in s}(1 - A_k) \text{ y } \hat{q}_B = \frac{1}{n}\sum_{k\in s}(1 - B_k) \tag{2.38}$$

Ahora bien, bajo muestreo aleatorio simple, el estimador de Horvitz-Thompson cumple la propiedad $\hat{p}_A = 1 - \hat{q}_A$. Esta propiedad tiene importancia en el ámbito de estimaciones de proporciones pues equivale a que el estimador $\hat{p}_A$ tiene el mismo comportamiento con respecto a P_A que el estimador $\hat{q}_A$ con respecto a Q_A. Esta propiedad recibe el nombre de propiedad complementaria y en general, el estimador $\hat{P}_{AW}$ no cumple esta deseable propiedad.

Una primera alternativa, sólo válida para muestreo aleatorio simple, para obtener un estimador calibrado que satisfaga la propiedad anterior consiste en hacer uso de la siguiente condición:

$$Q_B = 1 - P_B = \frac{1}{N}\sum_{k\in U} \omega_k(1 - B_k) \tag{2.39}$$

El resultado del estimador, aceptando que $\hat{q}_B \neq 0$, puede estar expresado por

$$\hat{Q}_{AW} = \hat{q}_A + (Q_B - \hat{q}_B)\cdot \frac{\hat{q}_{AB}}{\hat{q}_B} \tag{2.40}$$

con

$$\hat{q}_{AB} = \frac{1}{n}\sum_{k\in s}(1 - B_k)(1 - A_k) \tag{2.41}$$

De la misma manera como con el estimador $\hat{P}_{AW}$ en el ejemplo, la varianza asintótica de $\hat{Q}_{AW}$ viene dada por:

$$AV\left(\hat{Q}_{AW}\right) = \frac{(1-f)}{n}\frac{N}{N-1}\left[P_A Q_A + \left(\frac{Q_{AB}}{Q_B}\right)^2 \cdot P_B Q_B - 2\left(\frac{Q_{AB}}{Q_B}\right)(Q_{AB} - Q_A Q_B)\right] \quad (2.42)$$

y tenemos que $AV\left(\hat{P}_{AW}\right) < AV\left(\hat{P}_{AQ}\right)$ cuando ocurre:

$$\frac{P_{AB}}{P_B} < \frac{Q_{AB}}{Q_B} \quad (2.43)$$

Por lo tanto, asintóticamente, un estimador más eficiente para la proporción de población P_A es

$$\hat{P}_{AT} = \begin{cases} \hat{P}_{AW} & si \ \ \frac{\hat{p}_{AB}}{\hat{p}_B} < \frac{\hat{q}_{AB}}{\hat{q}_B} \text{ ó } \hat{q}_B = 0 \\ \hat{P}_{AQ} & si \ \ \frac{\hat{p}_{AB}}{\hat{p}_B} \geq \frac{\hat{q}_{AB}}{\hat{q}_B} \text{ ó } \hat{p}_B = 0 \end{cases}$$

Teniendo en cuenta que la varianza asintótica de $\hat{P}_{AT}$ es

$$AV(\hat{P}_{AT}) = \begin{cases} AV\left(\hat{P}_{AW}\right) & si \ \frac{P_{AB}}{P_B} < \frac{Q_{AB}}{Q_B} \\ AV\left(\hat{P}_{AQ}\right) & en\ otro\ caso \end{cases}$$

La segunda alternativa, para obtener un estimador calibrado que cumpla la propiedad complementaria, es considerar un estimador calibrado que minimice la distancia (2.24) sujeto a las condiciones:

$$\begin{cases} P_B = \frac{1}{N}\sum_{k\in U} \omega_k B_k \\ Q_B = \frac{1}{N}\sum_{k\in U} \omega_k (1 - B_k) \end{cases}$$

y de igual forma que con el estimador $\hat{P}_{AW}$, se obtiene que el nuevo estimador calibrado $\hat{P}_{AR}$ viene dado por:

$$\hat{P}_{AR} = \hat{P}_{AH} + \left(P_B - \hat{P}_{BH}\right) \cdot \hat{B}_1 + \left(Q_B - \hat{Q}_{BH}\right) \cdot \hat{B}_2 \quad (2.44)$$

donde

$$\hat{B}_1 = \frac{\sum_{k\in s} d_k q_k B_k A_k}{\sum_{k\in s} d_k q_k B_k} \ ; \ \ \hat{B}_2 = \frac{\sum_{k\in s} d_k q_k (1 - B_k) A_k}{\sum_{k\in s} d_k q_k (1 - B_k)} \quad (2.45)$$

El nuevo estimador así obtenido, sí cumple la propiedad complementaria. En cuanto al comportamiento asintótico del estimador, podemos decir que es asintóticamente insesgado y su varianza asintótica viene dada por:

$$AV\left(\hat{P}_{AR}\right) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U} \Delta_{kl}\,(d_k U_k)(d_l U_l) \quad (2.46)$$

donde $U_k = A_k - B_1 \cdot B_k - B_2 \cdot (1 - B_k)$ y

$$B_1 = \frac{\sum_{k \in U} q_k B_k A_k}{\sum_{k \in s} q_k B_k} \;\; ; \;\; B_2 = \frac{\sum_{k \in U} q_k (1 - B_k) A_k}{\sum_{k \in U} q_k (1 - B_k)} \tag{2.47}$$

la cual puede ser estimada mediante:

$$\hat{V}(\hat{P}_{AR}) = \sum_{k \in S} \sum_{l \in S} \frac{\Delta_{kl}}{\pi_{kl}} (d_k u_k)(d_l u_l) \tag{2.48}$$

donde $u_k = A_k - \hat{B}_1 \cdot B_k - \hat{B}_2 \cdot (1 - B_k)$

Una de las alternativas posibles a la hora de incorporar la información auxiliar en la estimación de una proporción, es considerar el empleo de modelos de regresión logística. Puesto que el atributo de estudio A, es una variable dicotómica, parece más natural considerar un modelo de regresión logística para incorporar la información auxiliar, que en este caso consideraremos que viene dada en forma de vector de variables auxiliares de dimensión P,

$$\vec{x}_k = (x_{1k}, x_{2k}, \dots, x_{Pk}) \tag{2.49}$$

que es conocido para todas las unidades de la población U y donde las variables que lo componen, pueden ser de tipo dicotómico o bien de tipo continuo.

De este modo, al ser el valor del vector conocido para todas las unidades podemos considerar el modelo dado por:

$$pl_k = P(A_k = 1) = \frac{\exp(\vec{x}_k \hat{\beta})}{1 + \exp(\vec{x}_k \hat{\beta})} \tag{2.50}$$

y $\hat{\beta}$ es un estimador del parámetro de regresión logística β.

Así, con este modelo tenemos el estimador LGREG, que viene dado por:

$$\hat{P}_{ALGREG} = \frac{1}{N} \sum_{U} pl_k + \sum_{s} \frac{A_k - pl_k}{\pi_k} \tag{2.51}$$

cuya varianza asintótica viene dada por:

$$AV(\hat{P}_{ALGREG}) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (d_k H_k)(d_l H_l) \tag{2.52}$$

donde $H_k = A_k - pl_k$.

Finalmente, podemos emplear conjuntamente el método de calibración y los modelos de regresión logística para obtener un nuevo estimador. Para ello, consideraremos la minimización de la distancia (2.24) sujeto a la condición de calibración:

$$\overline{PL} = \frac{1}{N} \cdot \sum_{k \in U} pl_k = \frac{1}{N} \cdot \sum_{k \in S} \omega_k \cdot pl_k \tag{2.53}$$

El estimador así obtenido viene dado por

$$\hat{P}_{AP} = \frac{1}{N} \sum_{k \in s} \omega_k A_k = \hat{P}_{AH} + \frac{(\overline{PL} - \hat{P}_{LH})}{\sum_{k \in s} d_k q_k pl_k} \cdot \sum_{k \in s} d_k q_k pl_k A_k \tag{2.54}$$

donde $\hat{P}_{LH}$ denota el estimador de Horvitz-Thompson para la variable pl_k

El estimador $\hat{P}_{AP}$ tiene por varianza asintótica

$$AV(\hat{P}_{AP}) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \, (d_k M_k)(d_l M_l) \tag{2.55}$$

donde

$$M_k = A_k - G \cdot pl_k \tag{2.56}$$

y

$$G = \frac{\sum_{k \in U} q_k pl_k A_k}{\sum_{k \in U} q_k pl_k} \tag{2.57}$$

la cual puede ser estimada por

$$\hat{V}(\hat{P}_{AP}) = \sum\nolimits_{k \in s} \sum\nolimits_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} (d_k m_k)(d_l m_l) \tag{2.58}$$

con $m_k = A_k - \hat{G} \cdot pl_k$ y $\hat{G} = \frac{\sum_{k \in s} d_k q_k pl_k A_k}{\sum_{k \in s} d_k q_k pl_k}$ (2.59)

CAPÍTULO III. RESULTADOS

6. Estudio de simulación

En este apartado, se ha llevado a cabo un estudio de simulación para analizar el comportamiento de los estimadores indirectos de proporciones analizados en el apartado anterior. El estudio de simulación ha sido programado en Rgui, y ha sido necesario desarrollar código nuevo para poder computar los estimadores a comparar.

Concretamente, la población empleada en el estudio de simulación se corresponde con una población real de N=21800 familias españolas, cuyos datos han sido extraídos de la Encuesta de Presupuestos Familiares realizada por el Instituto Nacional de Estadística (INE)

correspondiente al año 2012 y donde hemos considerado como variable de estudio *A*="Régimen de tenencia de vivienda" de manera que el atributo A toma el valor 1 para una familia con vivienda en propiedad sin hipoteca, y 0 para el resto de casos, esto es, nuestro objetivo es estimar la proporción P_A de familias con vivienda en propiedad sin hipoteca y donde hemos considerado las siguientes variables auxiliares:

- B_1="Estudios del sustentador principal". Atributo que toma el valor 1 para aquellas familias cuyo sustentador principal tiene estudios superiores y 0 en el resto de casos.
- B_2="Tipo casa". Atributo que toma el valor 1 para aquellas familias con un piso o casa media y cero para el resto de casos.
- X_3="Gastos totales anuales familiares". Variable numérica que indica el gasto anual de la familia.
- X_4="Ingresos mensuales familiares". Variable numérica que indica los ingresos mensuales de la familia.

De este modo, los estimadores calibrados $\hat{P}_{AW}$, $\hat{P}_{AT}$ y $\hat{P}_{AR}$ que sólo consideran un atributo auxiliar, fueron obtenidos empleando el atributo B_1 correspondiente al nivel de estudios del sustentador principal.

En el caso de los estimadores $\hat{P}_{AP}$ y $\hat{P}_{AGREG}$, basados en modelos de regresión logística, hemos empleado todas las variables auxiliares en su construcción.

Para llevar a cabo el estudio hemos considerado los diseños muestrales descritos en el Capítulo II, esto es:

- Muestreo Aleatorio Simple (MAS)
- Muestreo Sistemático
- Muestreo Estratificado
- Muestreo mediante método de Midzuno

donde en el muestreo estratificado hemos empleado como estratos las comunidades autónomas a la que pertenece cada familia y en el muestreo de Midzuno, hemos empleado la variable "Edad del sustentador principal" como variable para la definición del diseño muestral.

Con cada uno de los diseños muestrales, hemos seleccionado 10000 muestras para cuatro tamaños muestrales distintos. Concretamente, los tamaños considerados fueron n=500, n=550, n=600 y n=650, excepto para el muestreo sistemático, donde hemos considerado, en lugar del

tamaño muestral, cuatro pasos distintos, paso k=70, k=60, k=50 y k=40. De este modo, para cada estimador disponemos de 10000 estimaciones distintas de la proporción P_A con los cuatro tamaños muestrales considerados.

Para evaluar el comportamiento de los estimadores propuestos después de haber realizado las simulaciones, es necesario tener en cuenta criterios para evaluar el rendimiento de los resultados obtenidos. La comparación de los resultados simulados con los verdaderos valores utilizados para simular los datos proporciona una medida del comportamiento y la precisión de los estimadores en el proceso de simulación. Las medidas consideradas deben incluir una evaluación del sesgo y una evaluación de la eficiencia. Los métodos que resultan en una estimación insesgada con gran variabilidad o en una estimación sesgada con poca variabilidad pueden ser considerados de poca utilidad práctica. Consideremos ahora las medidas de rendimiento más utilizadas.

Evaluación del sesgo. El sesgo es la desviación de una estimación respecto de la verdadera cantidad a estimar, y puede indicar el rendimiento de los métodos que se está evaluando. Una evaluación del sesgo es la diferencia entre la estimación promedio y el verdadero valor. Otro enfoque consiste en calcular el sesgo como un porcentaje del valor real, supuesto que el valor verdadero no es igual a cero. El sesgo como porcentaje puede ser más informativo que el primer enfoque. Un sesgo normalizado superior a 40 por ciento en cualquier dirección tiene un impacto adverso apreciable sobre las tasas de eficiencia y error.

Evaluación de la precisión. El error cuadrático medio proporciona una útil medida de la precisión global, ya que incorpora medidas de sesgo y de la variabilidad.

El comportamiento de cada estimador se ha medido y comparado a través del sesgo relativo (RB)

$$RB\left(\hat{P}_A\right) = \frac{1}{P_A}\left[\frac{1}{10000}\sum_{r=1}^{10000}\left(\hat{P}_A{}^{(r)} - P_A\right)\right] \tag{3.1}$$

y de la eficiencia relativa (RE)

$$RE\left(\hat{P}_A\right) = \frac{ECM\left(\hat{P}_A\right)}{ECM\left(\hat{P}_{AHT}\right)} \tag{3.2}$$

siendo ECM el error cuadrático medio de un estimador definido como:

$$ECM(\hat{P}_A) = \frac{1}{10000}\sum_{r=1}^{10000}\left(\hat{P}_A{}^{(r)} - P_A\right)^2 \qquad (3.3)$$

donde $\hat{P}_A$ es un estimador cualquiera de la proporción poblacional considerada y donde $\hat{P}_A{}^{(r)}$ es la estimación r-ésima realizada con el estimador $\hat{P}_A$, esto es, la estimación llevada a cabo con el estimador mediante la muestra r-ésima tomada. La eficiencia relativa RE es la eficiencia relativa de cada estimador con respecto al estimador de Horvitz-Thompson. Un valor de RE>100 significa que el estimador es ineficiente con respecto al estimador de Horvitz-Thompson.

En la tabla 1 podemos examinar los resultados obtenidos con muestreo aleatorio simple para todos los tamaños de muestra considerados

Tabla 3.1

Tamaño de muestra n=500						
Estimador	YHT	AW	AT	AR	AP	AGREG
RB	0.00058	-0.00043	-0.00040	0.00057	-0.00051	-0.00051
RE	1	1.16261	1.14544	1.00272	0.97914	0.97895
Tamaño de muestra n=550						
RB	0.00025	0.00036	0.00038	0.00026	0.00027	0.00027
RE	1	1.15535	1.13894	0.99977	0.97021	0.97013
Tamaño de muestra n=600						
RB	-0.00039	-0.00043	-0.00039	0.00037	-0.00042	-0.00042
RE	1	1.14921	1.13462	1.00212	0.96456	0.96451
Tamaño de muestra n=650						
RB	0.00045	0.00056	0.00057	0.00045	0.00051	0.00051
RE	1	1.16483	1.14988	1.00255	0.97076	0.97069

Los resultados obtenidos con muestreo aleatorio simple son satisfactorios desde el punto de vista del sesgo relativo, pues para todos los estimadores y para todos los tamaños muestrales considerados el sesgo es prácticamente despreciable. No ocurre lo mismo desde el punto de vista de la eficiencia, pues en este caso los estimadores calibrados $\hat{P}_{AW}$, $\hat{P}_{AT}$ y $\hat{P}_{AR}$ son menos

eficientes que el estimador de Horvitz-Thompson y solamente los estimadores basados en modelos de regresión logística $\hat{P}_{AP}$ y $\hat{P}_{AGREG}$ son los que presentan una eficiencia mejor que el estimador de Horvitz-Thompson, si bien la ganancia en eficiencia conseguida por estos estimadores es mínima, produciéndose en el mejor de los casos una mejora en la eficiencia en torno al 3%. Así, el uso de información auxiliar mediante estos estimadores con muestreo aleatorio simple no es del todo eficiente

En la tabla 2, podemos observar los resultados obtenidos con muestreo sistemático. Al igual que ocurre con el muestreo aleatorio simple, los sesgos obtenidos con todos los estimadores es prácticamente nulo, si bien en este caso la eficiencia relativa presenta unos mejores resultados con respecto al muestreo aleatorio simple, pues en la mayoría de los casos los estimadores calibrados sí hacen un uso eficiente de la información auxiliar y presenta una eficiencia relativa ligeramente mejor que el estimador de Horvitz-Thompson, con una ganancia en torno al 10% en el mejor de los casos. Por otro lado, los estimadores basados en modelos de regresión logística mejoran también y presentan una eficiencia superior al estimador de Horvitz-Thompson, con una ganancia de aproximadamente un 13%.

Tabla 3.2

Paso de la muestra k=70						
Estimador	YHT	AW	AT	AR	AP	AGREG
RB	0.00051	0.00072	0.00040	0.00047	-0.00107	-0.00109
RE	1	1.25732	1.22499	1.00483	0.91303	0.91287
Paso de la muestra k=60						
RB	-0.00014	$-9.67\cdot10^{-5}$	-0.00030	0.00016	-0.00038	-0.00038
RE	1	1.22429	1.19178	0.97300	0.96862	0.96840
Paso de la muestra k=50						
RB	0.00044	-0.00034	-0.00029	$-1.33\cdot10^{-5}$	$3.18\cdot10^{-5}$	$4.23\cdot10^{-5}$
RE	1	0.98110	0.97708	0.99457	0.91301	0.91350
Paso de la muestra k=40						
RB	0.00041	0.00047	0.00052	0.00037	-0.00041	-0.00043
RE	1	0.92566	0.89522	0.90167	0.87934	0.86773

En los resultados correspondientes a muestreo estratificado (Tabla 3), podemos observar como el empleo de información auxiliar mejora sustancialmente las estimaciones, no sólo ya en el caso de los sesgos que siguen siendo prácticamente nulos, sino también con respecto a la eficiencia, donde la ganancia de los estimadores calibrados $\hat{P}_{AW}$, $\hat{P}_{AT}$ y $\hat{P}_{AR}$ es ya bastante considerable con un máximo de mejora en torno al 25%. De igual forma, los resultados obtenidos con los estimadores $\hat{P}_{AP}$ y $\hat{P}_{AGREG}$ mejoran con respecto a los muestreos anteriores presentando una eficiencia relativa mejor que la del estimador de Horvitz-Thompson en un 30% aproximadamente.

Tabla 3.3

Tamaño de muestra n=500						
Estimador	YHT	AW	AT	AR	AP	AGREG
RB	0.00051	0.00046	0.00042	0.00052	-0.00051	-0.00052
RE	1	0.95772	0.92871	0.90573	0.88344	0.87266
Tamaño de muestra n=550						
RB	-0.00044	-0.00047	-0.00043	0.00036	-0.00037	-0.00037
RE	1	0.90269	0.87861	0.84448	0.80278	0.80250
Tamaño de muestra n=600						
RB	-0.00034	-0.00040	-0.00035	0.00032	-0.00032	-0.00032
RE	1	0.87558	0.83934	0.78385	0.75772	0.74588
Tamaño de muestra n=650						
RB	0.00031	0.00027	0.00032	0.00028	-0.00031	-0.00031
RE	1	0.82481	0.79960	0.72581	0.70377	0.70221

Los resultados obtenidos mediante el muestreo por método de Midzuno vienen recogidos en la Tabla 4

Tabla 3.4

Tamaño de muestra n=500						
Estimador	YHT	AW	AT	AR	AP	AGREG
RB	0.00048	0.00049	0.00046	0.00043	0.00037	0.00037
RE	1	0.90552	0.90172	0.89288	0.87266	0.86899
Tamaño de muestra n=550						
RB	-0.00042	-0.00041	-0.00040	-0.00038	0.00035	0.00035
RE	1	0.89114	0.86227	0.80284	0.78728	0.78175
Tamaño de muestra n=600						
RB	-0.00035	-0.00033	-0.00032	0.00030	-0.00031	-0.00031
RE	1	0.83475	0.81285	0.74276	0.72017	0.71990
Tamaño de muestra n=650						
RB	0.00032	0.00027	0.00025	0.00026	-0.00029	-0.00029
RE	1	0.80904	0.79122	0.68335	0.65488	0.64883

En este caso, los resultados obtenidos mejoran incluso con respecto a los obtenidos con muestreo estratificado, es decir, con el muestreo mediante método de Midzuno la incorporación de información auxiliar mediante estimadores indirectos mejora notablemente la eficiencia de las estimaciones. Así, podemos ver que los estimadores $\hat{P}_{AP}$ y $\hat{P}_{AGREG}$ siguen ofreciendo los mejores resultados con una mejora en la eficiencia del 35% aproximadamente en el mejor de los casos con respecto al estimador de Horvitz-Thompson y los estimadores calibrados $\hat{P}_{AW}$, $\hat{P}_{AT}$ y $\hat{P}_{AR}$ presentan una mejora del 30% aproximadamente. Los sesgos, al igual que el resto de casos, siguen siendo prácticamente nulos para todos los estimadores.

CONCLUSIONES

En este trabajo hemos abordado el problema de la estimación de una proporción poblacional en el ámbito del muestreo en poblaciones finitas. La proporción poblacional de un atributo es un parámetro de gran interés en múltiples disciplinas, si bien la incorporación de información

auxiliar en la estimación de dicho parámetro es un problema menos estudiado que el caso de otros parámetros de interés como la media. El presente trabajo realiza una revisión de las principales técnicas para incorporar la información auxiliar en el proceso de estimación de una proporción poblacional centrándose principalmente en el método de calibración y en modelos de regresión logística, que permiten obtener estimadores indirectos para el parámetro estudiado de una manera eficiente. También, el trabajo incluye un estudio de las propiedades teóricas de los estimadores indirectos considerados e incluye un estudio de simulación para analizar el comportamiento de estos estimadores desde un punto de vista práctico, en donde podemos comprobar que estos estimadores no sólo poseen buenas propiedades teóricas sino que también producen una mejora en la eficiencia de las estimaciones, es decir, la incorporación de información auxiliar mediante los estimadores considerados se realiza de una manera eficiente.

BIBLIOGRAFÍA

Alvarez, E., Arcos, A., González, S., Muñoz, J.F., & Rueda, M. (2013). Estimating population proportions in the presence of missing data. *Journal of Computational and Applied Mathematics. 237*, 470-476.

Cassel, C.M., Särndal, C.E., & Wretman, J.H. (1977). *Foundations of inference in survey sampling*. New York , Wiley.

Deville, J.C., & Särndal, C.E. (1992). Calibration estimators in survey sampling. *J. Am. Stat. Assoc. 87*, 376–382.

Duchesne, P. (2003). Estimation of a proportion with survey data. *Journal of Statistics Education. [On line]*, *11*.

Fernández, F. R., & Mayor, J. A. (1995). *Muestreo en poblaciones finitas: curso básico.* Barcelona, Ediciones Universitarias de Barcelona (EUB).

Lehtonen, R., & Veijanen, A. (1998) Logistic generalized regression estimators. *Survey Methodology. 24*, 51-55.

Muñoz, J. F., Arcos, A., Álvarez-Verdejo, E., Rueda, M., & Martínez-Puertas, S. (2011) Estimators and confidence intervals for the proportion using auxiliary information with applications to the estimation of prevalences. *Journal Of Biopharmaceutical Statistics. 1*, 1-32.

Rueda, M., Martínez, S., Martínez, H., & Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference. 137*, 435-448.

Rueda,M., Muñoz, J.F., Arcos, A., & Álvarez, E. (2011). Estimators and confidence intervals for the proportion using binary auxiliary information with applications to pharmaceutical studies. *Journal of Biopharmaceutical Statistics. 21*, 526-554.

Särndal, C.E. (2007). The calibration approach in survey theory and practice. *Survey Methodology. 33*, 99-119.

Särndal, C.E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York, Springer.

Singh, S. (2003). *Advanced sampling theory with applications: How Michael "selected"*. The Netherlands, Amy. Kluwer Academic Publisher.

Wu, C., & Sitter, R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association. 96,* 185-193.