LA POST-ESTRATIFICACIÓN Y LOS MODELOS DE SUPERPOBLACIÓN

Cristina Aybar Arias
J. Santiago Murgui Izquierdo
Dpto de Economía Aplicada
Universidad de Valencia

RESUMEN

El interés de la estratificación en el diseño de encuestas por muestreo se manifiesta en tres aspectos: permite resultados desagregados, facilita la optimización de los recursos económicos y mejora la precisión de las estimaciones globales.

En muchas ocasiones, a priori no se dispone de la estructura estratificada del colectivo siendo necesario recurrir a la estratificación después de la selección y observación de las unidades muestrales. A esta práctica se la denomina post-estratificación. Las dificultades que presenta el análisis de la post-estratificación están asociadas al hecho de que el tamaño muestral correspondiente a cada estrato particular es variable, pudiéndose encontrar estratos en los que haya una escasa representación muestral.

Este trabajo aporta una aproximación a la metodología propia de la post-estratificación y propone un enfoque alternativo basado en un modelo de superpoblación.

Palabras Clave: Población Finita y Discreta. Post-Estratificación y Superpoblación.

1. Introducción

Una práctica habitual, en el muestreo de poblaciones finitas, es el uso de información auxiliar para mejorar la precisión de las estimaciones, tanto cuando se dispone a priori de dicha información, como cuando, en ocasiones, merece la pena desviar parte de los recursos económicos para conseguirla.

Si la variable auxiliar es de tipo continuo, la literatura estadística propone soluciones ampliamente conocidas, basadas en los estimadores indirectos de tipo regresión y razón. En cambio, cuando la variable auxiliar utilizada es discreta, el planteamiento clásico de los estimadores indirectos ya no es válido, teniendo que recurrir a los modelos de superpoblación (Murgui y Aybar, 1995; Aybar y Murgui, 1999a; 1999b), o bien, planteando diseños basados en la estratificación.

En este trabajo, situándonos en un escenario en el que tanto la variable de interés como la auxiliar sean discretas, vamos a estudiar, en primer lugar, la solución clásica de la estratificación, analizando los casos en los que es necesario recurrir a la post-estratificación o al muestreo doble o en dos fases; y en la segunda parte, se propone una solución alternativa enmarcada en los modelos de superpoblación.

2. La Estratificación

En muchas ocasiones el colectivo sobre el que se pretende observar la variable de interés, presenta una división natural, geográfica, por sexos, por edades, etc., con divisiones que presentan claras características diferenciales. La estratificación con muestreo aleatorio, es un procedimiento que combina la posibilidad de calcular el error cometido en la estimación, al mismo tiempo que aporta el interés de las muestras intencionadas, capaces de incorporar en su diseño toda la información disponible sobre la estructura del colectivo.

Considerando L el número de estratos definidos sobre el colectivo y N_h el número de elementos que integran cada estrato, es bien conocido que un estimador apropiado para la media de la población asociada con una variable Y es:

$$\hat{\mu}_Y = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{L} W_h \bar{y}_h \quad \text{donde} \quad \bar{y}_h = \frac{1}{n_h} \sum_{s_h=1}^{n_h} y_{s_h}$$

Siendo n_h el tamaño de la muestra s_h, seleccionada del estrato h. La estimación de la media poblacional así obtenida se denomina media muestral estratificada y la denotamos $\bar{y}_{st}$. El error de este estimador coincide con su varianza, al ser un estimador insesgado, y su valor está determinado por la expresión:

$$V\left[\bar{Y}_{st}\right] = \sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h} \frac{N_h - n_h}{N_h - 1}$$

donde σ_h^2 es la varianza poblacional de la variable Y dentro de cada estrato h.

Las ventajas del muestreo aleatorio estratificado frente a la selección de una muestra aleatoria simple son, en primer lugar, la posibilidad de obtener, de cada uno de los estratos, estimaciones de las características de interés, además de la estimación global, enriqueciendo así el estudio; y en segundo lugar, una mejora en la precisión de la estimación, respecto al estimador simple de la media muestral, conforme los estratos sean más homogéneos dentro de ellos y heterogéneos entre sí.

En la estratificación, el control sobre la distribución de las unidades muestrales entre los estratos permite diseñar el procedimiento de muestreo de manera que la precisión de las estimaciones mejore, determinando el tamaño de las muestras de forma que se minimice la varianza del estimador o incluso considerando los costes de selección en cada uno de los estratos, en caso de que éstos sean muy diferenciados.

Puesto que la base para una estratificación eficaz es la homogeneidad entre las unidades dentro de cada estrato y la heterogeneidad entre las unidades de diferentes estratos, para llevar a cabo la estratificación satisfactoriamente se requiere el conocimiento inicial sobre la naturaleza de la distribución de la variable de interés, el conocimiento de los tamaños de los estratos y la identificación precisa de las unidades de cada uno.

En general esta información no siempre está disponible en la práctica. En estas circunstancias, la posibilidad de recurrir a una variable auxiliar, relacionada con la de interés, suele aportar soluciones atractivas, tanto desde un punto de vista metodológico como práctico.

Denotamos por X a una variable discreta que toma los valores 1,2,...,L y que especifica la clasificación de las unidades en los distintos estratos. Si Y es la variable de interés, también discreta, que toma K posibles valores, identificamos las observaciones muestrales recurriendo a variables contadoras Q definidas como:

$$Q_{s_h i} = \begin{cases} 1 & \text{si} \quad y_{s_h} = i \\ 0 & \text{si} \quad y_{s_h} \neq i \end{cases} \quad \text{para} \quad i=1,2,...,K \ ; \ h=1,2,..,L$$

A partir de estas variables contadoras, las observaciones de la muestra s_h de tamaño n_h se representarán por:

$$\left[(q_{11},...,q_{1i},...,q_{1K}),.....,(q_{n_h 1},...,q_{n_h i},...,q_{n_h K})\right]$$

Para cada elemento muestral un solo $q_{s_h i}$ tomará el valor 1, siendo 0 el resto de componentes del vector.

Los subcolectivos U_h admiten la misma representación anterior pero para el número de elementos pertenecientes a cada U_h, denotado por N_h. Así, la proporción poblacional θ_i, de unidades pertenecientes a la categoría i de la variable Y se podría obtener como:

$$\theta_i = \frac{1}{N}\sum_{h=1}^{L}\sum_{U_h=1}^{N_h} q_{U_h i} = \sum_{h=1}^{L}\frac{N_h}{N}\theta_{hi} = \sum_{h=1}^{L} W_h \theta_{hi} \quad \text{para} \quad i=1,...,K$$

siendo θ_{hi}, la proporción de elementos pertenecientes a la categoría i de la variable Y, dentro del estrato U_h.

Admitiendo que las W_h son perfectamente conocidas, para estimar θ_i se procede a estimar las proporciones θ_{hi} de los estratos, a partir de las correspondientes proporciones muestrales, obtenidas de la siguiente manera:

$$\hat{\theta}_{hi} = p_{hi} = \frac{\sum_{s_h=1}^{n_h} q_{s_h i}}{n_h} \quad \text{para} \quad i=1,2,...,K$$

siendo, entonces el estimador de θ_i igual a $\hat{\theta}_i = \sum_{h=1}^{L} W_h \hat{\theta}_{hi}$, denotado a partir de ahora por $\hat{\theta}_{i,st}$.

3. La Post-Estratificación

En muchas situaciones prácticas la información sobre la variable auxiliar de clasificación entre los estratos, puede ser bastante incompleta. En un primer escenario podemos conocer sólo los tamaños de los estratos pero no la identificación de los elementos en cada uno. Características personales tales como las opciones políticas no pueden ser conocidas de manera individualizada, aunque sí se puede disponer de los resultados globales obtenidos en las pasadas elecciones, que definen el tamaño de los estratos asociados con las opciones políticas. En un muestreo basado en llamadas telefónicas no podremos conocer a priori el sexo o el estado civil del seleccionado, pero se puede recurrir a la información facilitada por los censos para establecer el tamaño de los estratos. En estos casos, la clasificación de las unidades muestrales en los diferentes estratos sólo es posible después de su observación.

A una estrategia de investigación en la que el plan de muestreo sea el aleatorio simple pero el procedimiento de estimación sea el asociado con el muestreo aleatorio estratificado, se le conoce como *estratificación después del muestreo o post-estratificación.*

Formalmente el procedimiento de post-estratificar consiste en seleccionar una muestra aleatoria simple del colectivo completo de tamaño n, efectuar la clasificación de las unidades observadas en los estratos, y utilizar para la estimación, la expresión del estimador estratificado $\overline{Y}_{st}$ anteriormente definido.

La expresión del estimador de θ_i con el procedimiento post-estratificado, que denotamos por $\hat{\theta}_{i,pst}$, coincide con la del estimador estratificado $\hat{\theta}_{i,st}$, antes

definido, con la diferencia de que ahora los tamaños muestrales n_h no han podido ser fijados de antemano.

Para deducir la expresión del error del estimador $\hat{\theta}_{i,pst}$ se considera la doble aleatorización asociada con la variable de interés Y, y con el tamaño aleatorio de la muestra perteneciente a cada estrato. Para esta última aleatorización es fácil comprobar el ajuste a un modelo de distribución hipergeométrica.

Teniendo en cuenta la distribuciones asociadas con esta doble aleatorización, se comprueba que la varianza del estimador se determina desarrollando la siguiente relación:

$$V\left[\hat{\theta}_{i,pst}\right] = V_{(s,v)}\left[\hat{\theta}_{i,pst}\right] = E_s\left[V\left[\hat{\theta}_{i,pst} \mid s\right]\right] + V_s\left[E\left[\hat{\theta}_{i,pst} \mid s\right]\right]$$

Donde los subíndices s y v indican la aleatoriedad respecto a los tamaños muestrales n_h y respecto a la variable de interés Y, respectivamente. Una aproximación a la varianza del estimador $\hat{\theta}_{i,pst}$, debida a Stephan (1945) y desarrollada en Fernández, F.R. y Mayor, J.A. (1994), adopta la siguiente expresión:

$$V\left[\hat{\theta}_{i,pst}\right] \approx \frac{1}{n}\sum_{h=1}^{L} W_h \sigma_h^2 + \frac{1}{n^2}\sum_{h=1}^{L}(1 - W_h)\sigma_h^2$$

El primer sumando de la varianza aproximada coincide con la varianza de la estimación de la media poblacional mediante muestreo aleatorio estratificado, asumiendo que el tamaño de las muestras en cada estrato es proporcional al tamaño de los mismos. El segundo término se puede considerar como la variación adicional introducida por la post-estratificación, y refleja el hecho de que los n_h son variables aleatorias.

La selección de una única muestra aleatoria simple de tamaño considerablemente elevado conduce a pensar con optimismo, que al clasificar los elementos con posterioridad a su observación, el reparto entre los estratos será casi igual al obtenido procediendo con un muestreo aleatorio estratificado con asignación proporcional. Así pues, se puede interpretar también el segundo sumando de la varianza aproximada del estimador post-estratificado como una corrección a este ajuste natural.

Uno de los problemas más comunes de la estratificación es la existencia de estratos de tamaño reducido. La asignación proporcional, en estos casos, tiene el inconveniente de fijar un tamaño muestral que podría ser tan pequeño que el error de su estimación sea excesivamente alto. La solución práctica que se utiliza en estos casos es destinar una parte del total de elementos de la muestra a repartir

entre todos los estratos por partes iguales, de esta manera, se estaría efectuando una corrección a la proporcionalidad en favor de los estratos pequeños.

Ante un procedimiento de post-estratificación, este problema requiere soluciones diferentes. En ocasiones la solución consiste en agregar dos o más estratos antes de realizar la estimación, teniendo que sacrificar con este procedimiento riqueza en los resultados del trabajo. En Chang y otros (1999) se propone un procedimiento de muestreo secuencial, en el que la selección se realiza sucesivamente hasta que los tamaños muestrales de los post-estratos alcancen una cantidad mínima y siempre controlando que el número de unidades totales seleccionadas no supere una cantidad máxima. El problema de este interesante procedimiento, tal y como sus autores comentan, es la obtención de expresiones tediosas, tanto para el estimador como para la determinación del tamaño máximo a seleccionar.

En un segundo escenario, cuando existe una clasificación del universo que puede ser de interés para la investigación de una variable Y, pero se desconocen los tamaños de los estratos así como la clasificación de las unidades, puede diseñarse un plan de actuación alternativo al descrito anteriormente en el primer escenario.

En estas circunstancias, se selecciona también una muestra aleatoria simple del colectivo completo, se identifican los L estratos U_h correspondientes a las L categorías determinadas por la covariable X y además, se estiman los tamaños N_h de los estratos.

El siguiente paso a realizar va a depender exclusivamente de las restricciones presupuestarias. Así, cuando el coste de observar la variable de clasificación no sea significativamente distinto al de la variable de interés, se procederá a obtener información de la variable de interés sobre todas las unidades muestrales seleccionadas. Es decir, el procedimiento apropiado sería el post-estratificado propuesto en el primer escenario, pero con la estimación adicional de los pesos W_h.

Cuando el coste de observar la variable de clasificación sea significativamente menor que el de la variable de interés puede diseñarse un procedimiento de muestreo en dos fases. En una primera fase se selecciona una muestra aleatoria M_1 de tamaño n, se identifican los L estratos y se estiman sus tamaños N_h, tal y como se procedió anteriormente. En una segunda fase, se selecciona, de cada uno de los L estratos, una muestra aleatoria M_{2h} de tamaño n'_h, con $\sum_h n'_h < n$, y a partir de ellas, se obtiene la información de la variable de interés Y.

Formalmente, el diseño muestral en dos fases descrito, implica desarrollar el proceso de estimación en un plan secuencial que se ajustará al siguiente esquema: en la primera fase se estiman los pesos W_h, a través de $p_{1h} = \frac{n_h}{n}$. En la segunda fase, se estiman las proporciones poblacionales θ_{hi} a partir de las proporciones muestrales p_{2hi}, calculadas recurriendo de nuevo a las variables contadoras Q, pero ahora definidas sobre las unidades muestrales de M_{2h}:

$$p_{2hi} = \frac{\sum_{M_{2h}=1}^{n'_h} q_{M_{2hi}}}{n'_h}$$

Finalmente, el estimador de las proporciones θ_i denotado, en este caso, por $\hat{\theta}_{i,md,st}$ se define como $\hat{\theta}_{i,md,st} = \sum_{h=1}^{L} p_{1h} p_{2hi}$.

En la práctica, el hecho de conocer a priori los tamaños de los estratos o no conocerlos, además de conducir a procedimientos de post-estratificación alternativos, implica otra diferencia. Si son conocidos, podríamos anticiparnos al problema de tener estratos pequeños, que implican la selección de una muestra aleatoria suficientemente grande para garantizarnos su representatividad, procediendo a una reclasificación inicial controlada. Cuando no se conocen los tamaños de los estratos a priori, el problema anterior puede plantearse después de seleccionada la muestra, y por tanto nos veríamos obligados a reagrupar los estratos inicialmente propuestos, de acuerdo a los resultados obtenidos.

El cálculo de la varianza del estimador en el muestreo doble implica efectuar un desarrollo en el que debe intervenir la aleatorización sobre la variable auxiliar de clasificación entre los estratos y sobre la variable de interés. Operando por condicionalidad en una estructura encadenada de las dos fases, se obtiene la expresión:

$$V\left[\hat{\theta}_{i,md,st}\right] = \left(\sum_{h=1}^{L} \sigma_h^2 W_h\right)\mathbf{A} + \left(\sum_{h=1}^{L} \sigma_h^2 (W_h - 1)\right)\mathbf{B} + \left(\sum_{h=1}^{L} (\theta_{hi} - \theta)^2 W_h\right)\mathbf{C}$$

donde:

$$\mathbf{A} = \frac{1}{nn'_h}\frac{N-n}{N-1} + \frac{N_h}{N-1}\frac{n-1}{nn'_h} - \frac{n}{N} - \frac{n-1}{(N-1)^2} + \frac{n-1}{N^2(N-1)^2}\sum_h N_h^2$$

$$\mathbf{B} = \frac{1}{nN}\frac{N-n}{N-1}$$

$$\mathbf{C}=\frac{N-n}{n(N-1)}-\frac{n-1}{N^2(N-1)^2}(N^2-\sum_h N_h^2)$$

Esta expresión presenta bastante analogía con una de las que Cochran (1981) propone, diferenciándose en los coeficientes **A** y **C,** a pesar de que este autor se apoya en argumentaciones diferentes.

4. Estimadores basados en un Modelo de Superpoblación

El problema de estimar proporciones incorporando información sobre una variable auxiliar, se puede resolver mediante un procedimiento distinto al clásico presentado en los apartados anteriores, recurriendo a un modelo estocástico que describa la relación existente entre las observaciones de las variables de interés y la auxiliar. Para describir dicha relación se supone que el modelo adecuado es el que especifican las siguientes hipótesis:

$$H_1: P(Y_u = i \mid x_u = h) = \alpha_{hi} \quad , \quad h=1,2,...,L \quad y \quad i=1,2,...,K$$

$$H_2: C[Y_u, Y_{u'} \mid x_1,...,x_N] = 0$$

Donde $P(Y \mid x)$ expresa una probabilidad condicionada, C la covarianza y verificándose las relaciones $\sum_{i=1}^{K} \alpha_{hi} = 1$ para h=1,2,...,L.

Como un caso particular, este modelo se adaptaría a aquellas situaciones en las que X e Y expresan una misma variable medida en dos ocasiones distintas de tiempo. En la primera ocasión se suponen parcialmente conocidos los resultados de una investigación exhaustiva, planteándose la revisión de los resultados censales en una segunda ocasión, mediante una investigación por muestreo.

La primera hipótesis indica que la probabilidad de que en la segunda ocasión una unidad adopte una opción, viene explicada por la opción que tal unidad adoptó en la ocasión precedente.

La determinación del estimador para una proporción de unidades con una opción determinada sobre la variable de interés, depende de las propuestas para los parámetros del modelo, ya que las proporciones de interés están relacionadas con los parámetros a través de la siguiente expresión:

$$\theta_i = P(Y_u = i) = \sum_{h=1}^{L} P(Y_u = i / x_u = h) P(X_u = h) = \sum_{h=1}^{L} \alpha_{hi} \Pi_h$$

En Aybar y Murgui (1999b) se presentan los estimadores basados en este modelo, así como sus propiedades. Las expresiones de los estimadores para las proporciones de interés, obtenidos maximizando la función de verosimilitud, son:

$$\hat{\theta}_i = \sum_{h=1}^{L} \Pi_h \hat{\alpha}_{hi} = \sum_{h=1}^{L} \Pi_h \frac{\sum_s z_{sh} q_{si}}{\sum_s z_{sh}} \qquad \text{para} \qquad i=1,2,...,K$$

Donde s es la muestra seleccionada de tamaño n y (Z,Q) son las variables dicotómicas introducidas para contar el número de elementos en la muestra que toman un determinado valor para cada una de las variables. Así:

$$Q_{si} = \{1 \text{ si } y_s = i \text{ y } 0 \text{ si } y_s \neq i \;;\; i = 1,.,K\}$$

$$Z_{sh} = \{1 \text{ si } x_s = h \text{ y } 0 \text{ si } x_s \neq h \;;\; h = 1,.,L\}$$

Es importante destacar que a pesar de que el modelo propuesto se define sobre el conocimiento censal de los valores adoptados por la covariable X, los estimadores propuestos únicamente dependen de agregados poblacionales, de ahí la referencia anterior al conocimiento parcial de la investigación exhaustiva.

La insesgadez de $\hat{\alpha}_{hi}$ con respecto a α_{hi} en base al modelo propuesto, garantiza que $\hat{\theta}_i$ también será un estimador insesgado con respecto a θ_i. Asimismo, se comprueba que su varianza está determinada por:

$$V\left[\hat{\theta}_i\right] = \sum_{h=1}^{L} (\alpha_{hi} - \alpha_{hi}^2) E_1 \left[\frac{\Pi_h^2}{\sum_{s=1}^{n} z_{sh}} \right] \quad , \quad \text{para} \quad i=1,2,...,K$$

Siendo E_1 la esperanza con respecto a las variables consideradas en la primera ocasión. Un estimador insesgado de la citada varianza es el que define la expresión:

$$e(\hat{\theta}_i) = \sum_{h=1}^{L} (\hat{\alpha}_{hi} - \hat{\alpha}_{hi}{}^2) \frac{\Pi_h^2}{\sum_s z_{sh} - 1}$$

Cuando la información respecto a las proporciones poblacionales Π_h, asociadas a la variable auxiliar X son desconocidas, y se considera de interés su estimación para incorporarla al proceso inferencial, se puede también proceder con un muestreo en dos fases. En la primera fase se estiman dichas proporciones y en la segunda fase, considerando las hipótesis del modelo antes planteado y la

estimación de Π_h se construye la estimación de las proporciones de interés. El estimador que se propone es:

$$\hat{\theta}_{i,md} = \sum_{h=1}^{L} \frac{p_{2hi}}{p_{2h}} p_{1h} = \sum_{h=1}^{L} \frac{\sum_{M_2=1}^{n'} z_{2h} q_{2i}}{\sum_{M_2=1}^{n'} z_{2h}} \sum_{M_1=1}^{n} z_{1h} \qquad \text{para} \qquad i=1,2,...,K$$

Donde las variables (Z,Q) son variables contadoras análogas a las definidas anteriormente pero diferenciando cuando se observan en la segunda muestra, introduciendo el subíndice 2 y en la primera, con el subíndice 1.

El error de este estimador, se comprueba que adopta la expresión:

$$V\left[\hat{\theta}_{i,md}\right] = \sum_{h=1}^{L} \alpha_{hi}^2 V_1[p_{1h}] + \sum_{h=1}^{L} (\alpha_{hi} - \alpha_{hi}^2) E_1\left[\frac{p_{1h}^2}{p_{2h}}\right]$$

La varianza es similar a la del estimador $\hat{\theta}_i$ con la salvedad de la estimación en la primera fase de las proporciones poblacionales de la covariable X y el primer sumando, debido precisamente al error cometido en esta estimación.

5. Comparación de ambos Enfoques

Si se comparan las expresiones de los estimadores planteados bajo los dos enfoques, el clásico basado en el diseño aleatorio estratificado, $\hat{\theta}_{i,st}$ y el de los modelos, $\hat{\theta}_i$:

$$\hat{\theta}_{i,st} = \sum_{h=1}^{L} W_h \frac{\sum_{s_h=1}^{n_h} q_{s_h i}}{n_h} \quad , \quad \hat{\theta}_i = \sum_{h=1}^{L} \Pi_h \frac{\sum_s z_{sh} q_{si}}{\sum_s z_{sh}}$$

se puede observar que coinciden, ya que, los pesos W_h se pueden identificar con las proporciones Π_h asociadas a la variable auxiliar. El sumatorio $\sum_s z_{sh}$ indica el número de elementos en la muestra s para los que X=h, coincidiendo con n_h. En cuanto a los numeradores, la suma $\sum_s z_{sh} q_{si}$ determina el número de unidades de la muestra s que toman el valor X=h e Y=i, lo mismo que el sumatorio

$\sum_{s_h=1}^{n_h} q_{s_h i}$, ya que cuenta el número de unidades en las que Y=i para el conjunto de elementos en los que X=h.

Por otro lado, teniendo en cuenta que $V[Y/x=h]=\alpha_{hi}-\alpha_{hi}^2$, también se puede observar que la varianza estimada de $\hat{\theta}_i$ es análoga a la varianza estimada para el estimador estratificado $\hat{\theta}_{i,st}$.

A pesar de que coincidan las expresiones de los estimadores, e incluso que las estimaciones de sus errores sean similares, los procesos inferenciales requieren interpretaciones diferentes. En un caso, se debe someter a contraste el modelo de superpoblación, mientras que en el enfoque clásico nos apoyamos en el diseño de la muestra.

Cabe llamar la atención que cuando se pueda aceptar la existencia de una estructura diferenciada en el colectivo, el modelo conduce a un estimador con un nivel de precisión comparable al estratificado, pero sin necesidad de conocer la identificación de las unidades a los estratos, sino simplemente los tamaños de los estratos, situación que, como antes se ha comentado, se da con bastante frecuencia en la práctica.

El modelo de superpoblación permitiría obtener una estimación desagregada de la población, igual que en la estratificación, a través de la estimación de los parámetros α_{hi}. Por ello se podría decir, que el planteamiento de este modelo tiene las ventajas del procedimiento estratificado y del post-estratificado, pero no sus inconvenientes. Es decir, la ventaja del estratificado en cuanto al enriquecimiento de sus resultados, por desagregados, y por la incorporación de información auxiliar, pero no el inconveniente de necesitar conocer la identificación precisa de las unidades a cada uno de los estratos. Por ello, tiene la ventaja de la post-estratificación que sólo requiere conocer los resultados globales de una variable auxiliar, pero sin implicar su desventaja, en cuanto a la pérdida de precisión que ello significa.

Destacamos también las analogías entre los estimadores obtenidos en el muestreo doble o en dos fases, para la estratificación o bajo los modelos de superpoblación, pero se debe resaltar la sencillez analítica con la que se deducen los resultados obtenidos bajo los modelos de superpoblación, frente a la complejidad que plantea el cálculo de las expresiones de la varianza en el caso de la post-estratificación, junto con la falta de acuerdo que existe en cuanto a su expresión.

Además, como se puede observar a lo largo de este trabajo, las hipótesis de los modelos permanecen invariables ante los distintos escenarios producidos por los distintos niveles de conocimiento sobre la variable auxiliar, mientras que,

en una inferencia basada en el diseño, este cambia según el escenario que se contemple, llegando a una situación, en el muestreo doble, en la que se combinan dos tipos de selección aleatoria correspondientes a las dos fases definidas.

6. Conclusiones

La determinación de diversas soluciones utilizando diferentes métodos inferenciales, debe verse como algo que enriquece los planteamientos. Con los dos enfoques planteados, el clásico basado en el diseño estratificado, y el de los modelos de superpoblación, se pueden obtener soluciones a la estimación de proporciones con información auxiliar también discreta.

Los desarrollos analíticos para determinar las expresiones, de los errores de los estimadores basados en los modelos, son mucho menos complejas que en el caso del diseño. Además, también cabe destacar que en el supuesto de disponer de la proporción poblacional de la variable auxiliar, el modelo propuesto resulta más atractivo de utilizar que el diseño estratificado, por las ventajas, tanto prácticas como estadísticas que su estimador asociado conlleva.

Con este trabajo hemos pretendido aportar una reflexión sobre una técnica práctica, simple y habitual, como es la post-estratificación, a la que, tal y como señalan Holt, D. y Smith, T.M.F. (1979), se le ha destinado poco espacio en la literatura, proponiendo además soluciones alternativas. Nuestro siguiente paso, en esta línea de investigación, va a consistir en el planteamiento de una simulación en la que se puedan recoger empíricamente los resultados comparados de la aplicación de ambos enfoques en un conjunto amplio de escenarios.

Bibliografía

Aybar, C. y Murgui, J.S. (1999a). Estimación de Proporciones con Encuestas Repetidas en Poblaciones Dicotómicas. *Estadística Española* (pendiente de publicación).

Aybar, C. y Murgui, J.S. (1999b). Estimación de Proporciones con Información Auxiliar de Variables Discretas. *Qüestiió* (pendiente de publicación).

Chang, K., Liu, J. y Han, C. (1998). Multiple inverse sampling in post-stratification. *Journal of Statistical Planning and Inference*, **69**, 209-227.

Chang, K., Han, C. y Hawkins, D. (1999).Truncated multiple inverse sampling in post-stratification. *Journal of Statistical Planning and Inference,* **76,** 215-234.

Cochran, W.G. (1981). *Técnicas de Muestreo Estadístico,* Wiley, New York.

Fernandez F.R. y Mayor J.A. (1994). *Muestreo en Poblaciones Finitas: Curso Básico.* Promociones y Publicaciones Universitarias, S.A. Barcelona.

Holt, D. y Smith, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society, Ser. A*, **142**, 33-46.

Jagers, P. (1986). Post-stratification against bias in sampling. *International Statistical Review,* **54**, 159-167.

Kish, Lesly (1965). *Survey Sampling.* John Wiley & Sons, Inc. New York.
Levy, P.S. y Lemeshow, S. (1991). *Sampling of Populations.* John Wiley & Sons, Inc. New York.
Murgui, J.S. y Aybar, C. (1995). Estimadores de Regresión y Razón para proporciones. *Estadística Española*, **138**, 5-13.
Smith, T.M.F. (1990). Post-stratification. *The Statistician*, **40**, 315-323.