Master en Estadística Aplicada y Estadística para el Sector Público

Muestreo en Poblaciones Finitas

Félix Aparicio

Universidad de Alcalá

Instituto Nacional de Estadística

CIFF
Centro Internacional de Formación Financiera

2008
2009

Muestreo con Submuestreo en Poblaciones Finitas

6.1 Métodos Simplificados de Estimación de Varianzas en Encuestas Complejas

6.1.1 Introducción

En encuestas con diseños muestrales de más de una etapa, los métodos teóricos para estimar varianzas son de difícil aplicación. En la práctica se suelen utilizar otros métodos simplificados. A continuación veremos algunos de ellos. Antes vamos a describir una técnica estadística que se empleará al estudiar el método de los conglomerados últimos.

Proposición 6.1 (Método Delta; Rao, 1973) *Sea T_n una sucesión de estadísticos tales que*

$$n^{1/2}(T_n - \theta) \rightarrow^L X \sim N(0, \sigma^2(\theta)),$$

y sea g una función univariante con primera derivada g'. Si $g'(\theta) \neq 0$, entonces

$$n^{1/2}(g(T_n) - g(\theta)) \rightarrow^L X \sim N(0, [g'(\theta)\sigma(\theta)]^2).$$

Si, además g' es continua, entonces

$$\frac{n^{1/2}(g(T_n) - g(\theta))}{g'(T_n)} \rightarrow^L X \sim N(0, \sigma^2(\theta)).$$

Si, además, $\sigma(\theta)$ es continua, entonces

$$\frac{n^{1/2}(g(T_n) - g(\theta))}{g'(T_n)\sigma(T_n)} \rightarrow^L X \sim N(0,1).$$

6.1.2 Método de Grupos Aleatorios

El siguiente resultado es básico:

Proposición 6.2 () *Sea $\hat{\theta}$ un estimador de θ basado en una muestra probabilística. Sean k subconjuntos de la muestra original y un estimador de θ obtenido de cada subconjunto,*

$$\hat{\theta}_1, ..., \hat{\theta}_k$$

Entonces

$$\hat{\theta}^* = \frac{1}{k} \cdot \sum_{i=1}^{k} \hat{\theta}_i$$

es un estimador de θ y si definimos

$$\hat{V}_1 = \frac{1}{k \cdot (k-1)} \cdot \sum_{i=1}^{k} (\hat{\theta}_i - \hat{\theta}^*)^2$$

entonces

$$E(\hat{V}_1) = V(\hat{\theta}^*) - \frac{1}{k \cdot (k-1)} \cdot \sum_{i=1}^{k} \sum_{j=1}^{k} Cov(\hat{\theta}_i, \hat{\theta}_j) + \frac{1}{k \cdot (k-1)} \cdot \sum_{i=1}^{k} (E[\hat{\theta}_i] - E[\hat{\theta}^*])^2$$

donde $Cov(.,)$ indica la covarianza entre dos variables aleatorias y la suma doble está extendida a los $i \neq j$.

6.1.3 Método de Conglomerados Ultimos

En lo que resta de tema supondremos que trabajamos con un diseño muestral estratificado, en cuya primera etapa se toman PSU's (unidades primarias de muestreo) con probabilidades desiguales y con reemplazamiento o bien con probabilidades iguales y sin reemplazamiento. Después de esta primera fase habrá 0,1,2 o más fases de muestreo, con la

condición de que se puedan estimar totales poblacionales en forma insesgada mediante estimadores de Horvitz-Thompson y que el estadístico de interés sea una función suave de estos totales poblacionales. El método de los conglomerados últimos (que engloba al método delta) se puede describir mediante las siguientes ecuaciones.

$$Var(\hat{\theta}) = \sum_{h=1}^{L} \frac{1 - \lambda_h f_h}{n_h} \cdot \Delta g(\hat{Z})^T s_h^2 \Delta g(\hat{Z})$$

donde

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h) \cdot (z_{hi} - \bar{z}_h)^T$$

$$z_{hi} = \sum_{j=1}^{n_{hi}} n_h w_{hij} z_{hij}$$

$$\bar{z}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} z_{hi}$$

$f_h = n_h/N_h$ es la fracción de muestreo de primera etapa y λ_h vale 1 si en la primera etapa las PSU se toman sin reemplazamiento y 0 si se toman con reemplazamiento. w_{hij} es el factor de elevación poblaciónal del indivuduo j del la i-ésima PSU del estrato h, z representa un vector de variables cuyos totales poblacionales estimamos como paso intermedio antes de calcular nuestro estadístico. Este estadstico θ es una función suave g de los totales poblacionales, $\Delta(g)$ indica las derivadas parciales de g respecto de esos totales.

La idea del método es estimar los totales poblacionales mediante una técnica del tipo de la de los grupos aleatorios, tomando como grupos a cada PSU. Después, se utiliza el método delta para relacionar la varianza del estádístico de interés con la de los totales.

6.1.4 Método de Semimuestras Reiteradas

Expondremos la técnica en el caso de que haya exactamente dos conglomerados en cada estrato. En Wolter (1985) se puede encontrar la extensión a otros casos.

Se trata de tomar R semimuestras. cada semimuestra se define como un subconjunto de la muestra formado por exactamente una de las dos

PSU que hay en la muestra de cada estrato. El número R se toma como el mínimo número entero múltiplo de 4 que sea mayor que el número de estratos. La forma de elegir los conglomerados en la semimuestra viene dada por los elementos de R columnas cualesquiera, excepto la primera de una matriz de Hadamard de orden R (las matrices de Hadamard están formadas por ceros y unos, se emplean en diseño de experimentos y se pueden encontrar por ejemplo en el libro de Wolter (1985)). La idea es tomar las semimuestras en forma equilibrada, de manera que con pocas semimuestras se tenga una varianza reducida. Despues, se calcula el estadístico de interés para cada semimuestra y finalmente se estima la varianza del estadístico como la varianza de los estadísticos calculados para cada semimuestra.

6.1.5 Método Jackknife

La idea es aplicar el jackknife clásico quitando cada PSU de la muestra y estimando con el resto de PSU's. Las ecuaciones son:

$$\hat{Z}_{h'i'} = \sum_{h \neq h'} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{hij} z_{hij} + \frac{n'_h}{n'_h - 1} \cdot \sum_{i \neq i'} \sum_{j=1}^{n_{h'i}} w_{h'ij} z_{h'ij}$$

$\hat{Z}_{h'i'}$ es la estimación de los totales quitando el conglomerado i' del estrato h'. Despues calculamos nuestro estadístico de interes como $\hat{\theta}_{h'i'} = g(\hat{Z}_{h'i'})$. Las medias del estadístico por estrato se definen como $\hat{\theta}_{h'} = (1/n_{h'}) \cdot \sum_{k=1}^{n'_h} \hat{\theta}_{h'k}$ y finalmente el jackknife nos estima la varianza como

$$Cov_J(\hat{\theta}) = \sum_{h=1}^{L} \frac{(1 - \lambda_h f_h)(n_h - 1)}{n_h} \cdot \sum_{i=1}^{n_h} (\hat{\theta}_{hi} - \hat{\theta}_h)(\hat{\theta}_{hi} - \hat{\theta}_h)^T$$

6.1.6 Métodos Bootstrap

Existen varias versiones del bootstrap adaptadas a la estimación de varianzas en encuestas con diseños complejos. Sin embargo, su aplicación al caso de diseños polietápicos es mas complicada que la de las técnicas anteriormente citadas.

Algunas de las técnicas bootstrap que se pueden emplear son el bootstrap con reemplazamiento, el booststrap poblacional o el Mirror Match booststrap.

Chapter 7

Otras Técnicas Especiales de Muestreo

7.1 Algunas Técnicas Especiales de Muestreo

En esta sección estudiaremos algunas técnicas de muestreo que se emplean en circunstancias no tan habituales como las de las técnicas que hemos visto hasta ahora.

7.1.1 Muestreo Doble o Bifásico

Cuando el marco que empleamos para diseñar la muestra contiene poca información sobre la población es difícil realizar un diseño eficiente. En este caso se puede intentar hacer un diseño muy sencillo y aumentar en tamaño muestral, o reunir mas información sobre la población y despues realizar un diseño mas eficiente. Sin embargo, debido al coste elevado o a la falta de tiempo puede que ninguna de estas dos opciones sea viable.

Una alternativa es el muestreo doble, que consiste en diseñar una muestra sencilla de tamaño grande y recoger para esa muestra solo información que no sea costosa, y despues, en una segunda fase, ayudándose de esa información, seleccionar una submuestra, recogiendo las variables de interés solo para la submuestra. Para que este tipo de muestreo

resulte eficiente tendremos que ser capaces de obtener en la primera fase la suficiente información para que, en el conjunto de las dos fases y con un coste dado, podamos estimar las carácterísticas de interés con mas precisión que si hubiéramos utilizado un diseño de una sola fase.

Aplicación a la Estratificación

Si intentamos utilizar muestreo estratificado y no conocemos los valores de los W_h de los estratos, podemos diseñar en la primera fase una muestra aleatoria simple de n' unidades y estimar $\hat{W}_h = n'_h/n'$, donde n'_h es el número de unidades que pertenecen al estrato h.

En la segunda fase podemos utilizar muestreo estratificado. Si, por poner un ejemplo sencillo, en la segunda fase utilizamos m.a.s. de tamaño n_h dentro de cada estrato tomando una submuestra de la muestra de la primera fase , podemos estimar $\bar{y}'_{es} = \sum_{h=1}^{L} \hat{W}_h \bar{y}_h$, donde es indica estratificado. La diferencia con el caso de una fase es que $\hat{W}_h$ son variables aleatorias. Este estimador es insesgado y su varianza es

$$Var(\bar{y}'_{es}) = \frac{N-n'}{N} \cdot \frac{S_y^2}{n'} + \sum_{h=1}^{L} (\frac{1}{\phi_h} - 1) \cdot W_h \frac{S_{yh}^2}{n'}$$

donde los $\phi_h = n_h/n'_h$ se supone que están fijos y que pertenecen al intervalo $(0, 1]$.

Aplicación a los estimadores de la razón

Se trata ahora de utilizar la m.a.s. de tamaño n' de la primera fase para estimar la media poblacional de la variable auxiliar en un estimador de la razón. Llamamos a este estimador $\hat{\bar{x}}'$ Suponiendo que en la segunda fase tomamos una submuestra aleatoria simple de tamaño n de la muestra de la primera fase tendremos $\bar{y}'_R = \frac{\bar{y}}{\bar{x}} \cdot \hat{\bar{x}}'$ Se puede demostrar que su varianza es

$$Var(\bar{x}'_R) = \frac{N-n'}{N} \cdot \frac{S_y^2}{n'} + \frac{n'-n}{n'} \cdot \frac{S_y^2 + R^2 S_x^2 - 2RS_{xy}}{n}$$

7.1.2 Modelos de captura-recaptura

Supongamos que se accede a N_a unidades de una población y que estas son marcadas. Si despues se accede a n unidades de la misma población y n_a resultan estar marcadas, un estimador del total de individuos en la población es $\hat{N} = N_a \cdot n/n_a$. Este estimador se llama Indice de Lincoln

y es sesgado, pero consistente. La hipótesis de este método es que el proceso de captura y recaptura supone una muestra aleatoria simple sin reemplazamiento. En este caso, podemos pensar en este estimador como un estimador de la razón, con variable de interés $Y = 1$ para todos los individuos de la población y variable auxiliar X que vale 1 para los individuos que han sido marcados y cero en otro caso. El estimador del total de Y será $\hat{Y}_r = X \cdot \hat{y}/\hat{x}$ que, bajo la hipótesis citada, coindice con $\hat{N}$.

7.1.3 Estimadores en dominios de estudios y pequeñas áreas

Llamaremos dominios a cada uno de los D grupos disjuntos U_d, $d = 1, ..., D$ en que dividimos a una población. Nos interesa estimar totales o medias de determinadas variables en estos dominios. La diferencia esencial con otras particiones de la población como los estratos es que los dominios no intervienen en el diseño. Por este motivo, no podemos controlar los errores de nuestras estimaciones.

Sin embargo, es sencillo dar fórmulas para hacer estas estimaciones. Para estimar totales, si el tamaño del dominio es desconocido utilizaremos el estimador de Horvitz-Thompson:

$$\hat{Y}_d = \sum_{i \in U_d} w_i \cdot Y_i$$

donde w_i es el factor de elevación poblacional de la unidad i, y la suma está extendida a los individuos de la muestra que pertenecen al dominio. Si conocemos el tamaño del dominio N_d podemos emplear el estimador de la razón

$$\tilde{Y}_d = N_d \cdot \frac{1}{\hat{N}_d} \cdot \sum_{i \in U_d} w_i \cdot Y_i$$

donde

$$\hat{N}_d = \sum_{i \in U_d} w_i$$

Para estimar la media en un dominio, tanto si conocemos N_d como si no emplearemos

$$\bar{Y}_d = \frac{1}{\hat{N}_d} \cdot \sum_{i \in U_d} w_i \cdot Y_i$$

Si los dominios son muy pequeños, entonces estos estimadores tendrán una varianza grande. Existen técnicas especiales para estimar en este tipo de dominios. Muchas de estas técnicas se basan en modelos que relacionan las variables de interés con otras variables auxiliares que están disponibles para los dominios.

7.2 Muestreo en Ocasiones Sucesivas

7.2.1 Introducción

Hay muchas encuestas que se repiten en el tiempo, por ejemplo, mensual o trimestralmente, midiendo las mismas variables cada vez. Para este tipo de encuestas una pregunta fundamental es si la muestra debe ser distinta cada vez, o solaparse con las de veces anteriores.

Nos ocuparemos aqui de una encuesta que se repita en dos ocasiones. La respuesta a la pregunta anterior depende de si estamos interesados en estimar el nivel de la variable de interés en cada ocasión o el cambio de esa variable de la primera a la segunda ocasión o la media global entre las dos ocasiones.

7.2.2 Estimadores del Cambio y del Nivel

Supongamos que tenemos una m.a.s. sin reemplazamiento de tamaño n. Si llamamos Y_t a la variable de interés en el instante $t \in \{1, 2\}$, y estamos interesados en la media de esta variable, considerando el estimador simple del cambio $\hat{\delta} = \bar{y}_2 - \bar{y}_1$, tenemos los siguientes resultados:

$$Var(\hat{\delta}) = \frac{2S^2}{n} \cdot (1 - \rho_{12}\phi_c)$$

donde ϕ_c es la proporción de unidades de la muestra comunes en las dos ocasiones y ρ_{12} es el coeficiente de correlación entre las estimaciones de las medias basadas en esas unidades comunes.

De esta expresión se deduce que si, como cabe esperar, ρ_{12} vale aproximadamente 1, es mejor mantener la misma muestra en las dos ocasiones.

Por el contrario, si estimamos la media global entre las dos ocasiones mediante $\bar{y} = (1/2) \cdot (\bar{y}_1 + \bar{y}_2)$, es

$$Var(\bar{x}) = \frac{S^2}{2n} \cdot (1 + \rho_{12}\phi_c)$$

En este caso resulta mejor emplear muestras indepedientes en las dos ocasiones.

7.2.3 Estimadores de Mínima Varianza

Estimación del cambio

Utilizamos un estimador combinado de la forma $\hat{\theta} = w_1(\bar{y}_{2c} - \bar{y}_{1c}) + w_2(\bar{y}_{2n} - \bar{y}_{1n})$ donde c indica la parte común de la muestra y n la no común; ha de ser $w_1 + w_2 = 1$. Se demuestra que el valor óptimo (de varianza mínima) se alcanza en

$$w_1 = \frac{V(\bar{y}_{2n} - \bar{y}_{1n})}{V(\bar{y}_{2n} - \bar{y}_{1n}) + V(\bar{y}_{2c} - \bar{y}_{1c})}$$

con una varianza de

$$Var(\hat{\theta}) = \frac{2S^2(1 - \rho_{12})}{n(1 - \rho_{12}(1 - \phi_c))}$$

Estimación de la media en la segunda ocasión

En este caso se obtiene la misma precisión manteniendo la misma muestra que cabiándola por completo en la segunda ocasión.

7.2.4 Rotación de la Muestra con Solapamiento Parcial

Se trata de intentar conciliar los intereses contrapuestos que hemos visto en los apartados anteriores. Con este fin, se suelen emplear en encuestas continuas esquemas de muestreo en los que hay solapamiento parcial entre las muestras de sucesivos instantes temporales. Esto quiere decir que de un instante del tiempo al siguiente se mantiene una parte de la muestra y se sustituye otra parte.

Por ejemplo, en algunas de las encuestas que ha realizado el INE se han empleado diseños en los que una sexta parte de la muestra se sustituye cada trimestre. De esta forma cinco sextos de la muestra son comunes de un trimestre al siguiente.

Esta práctica es de uso general también en otros paises. Algunos de los esquemas de rotación de la muestra son bastante complejos, por ejemplo en la CPS de Estados Unidos la muestra se divide en 8 grupos, cada uno de ellos permanece cuatro períodos consecutivos, para salir luego durante 8 períodos consecutivos y finalmente volver a estar en la muestra otros 4 períodos mas.

Estos esquemas de rotación producen estimaciones cuyos errores están autocorrelados en el tiempo. Por ejemplo, en el caso de un solapamiento parcial de cinco sextos de la muestra, cabe esperar autocorrelaciones no nulas con retardos de hasta orden 5 inducidas por la rotación parcial de la muestra.

7.3 Encuestas Panel

7.3.1 Introducción

El diseño y análisis de encuestas a lo largo del tiempo es un tema de gran interés. Los motivos son, principalmente dos. el primero se ellos es que las carácterísticas de los individuos pueden cambiar en el tiempo, por ejemplo, la renta de una familia no ha de ser la misma todos los años. El segundo motivo es que la misma población puede cambiar en el tiempo. Si no fuera por estos problemas, una única encuesta realizada en un solo instante nos daria toda la información necesaria.

Los distintos tipos de encuestas que se pueden diseñar teniendo en cuenta la dimensión temporal son basicamente cuatro, encuesta repetida, panel, panel rotatorio y panel dividido.

Encuestas Repetidas. Consiste en realizar diseños de encuestas individuales en cada instante del tiempo, sin forzar a que las mismas unidades pertenezcan a la muestra en instantes sucesivos.

Panel. Se recoge básicamente la misma información a lo largo del tiempo sobre una muestra única y fija.

Panel Rotatorio. Es semejante al panel, pero una parte de la

muestra se renueva de un instante al siguiente.

Panel Dividido. Se utiliza un panel y a la vez una encuesta repetida en cada instante temporal.

7.3.2 Problemas de los Paneles

La utilización de un panel tiene ventajas e inconvenentes. Las ventajas son una mejor estimación del cambio de las características de la población en el tiempo y la posibilidad de agregar la muestra en el tiempo, por ejemplo, recogiendo algunas variables estáticas distintas cada vez. Los inconvenientes son la creación de sesgos, debido a que la muestra teórica se va reduciendo en el tiempo y el condicionamiento de las unidades encuestadas, debido a que se les pide información varias veces sobre unas mismas variables.

7.3.3 Estimación de datos con Paneles

La estimación con paneles se realiza básicamente igual que la de una única encuesta o la de encuestas repetidas en dos ocasiones. Sin embargo, se deben calcular los factores de elevación adecuados según se explica en la siguiente sección.

Otro tipo de estimaciones basadas en paneles tienen en cuenta la dimensión temporal de la información y emplean técnicas propias de la econometria y del análisis de series temporales.

7.3.4 Análisis Longitudinal y Transversal

Un panel es transversalmente representativo si permite estimar las características de la población en cada uno de los instantes de tiempo. Un panel es representativo longitudinalmente si en todo instante de tiempo permite estimar las características de la población inicial, a pesar de la mortalidad y otros problemas de no respuesta.

Para poder efectuar análisis transversales con un panel, se debe seguir la evolución de la población en cada instante de tiempo, por ejemplo nacimientos, muertes y cambios de estrato, añadir una muestra suplementaria ,en cada instante de tiempo, representativa de los nacimientos y reajustar de manera acorde los factores de elevación.

La muestra longitudinal se debe considerar como una medición múltiple de las mismas variables (para las mismas unidades iniciales) a lo largo del tiempo. Por tanto, en principio se deben emplear los factores de elevación iniciales. Sin embargo, en la realidad habrá que ir modificándolos para corregir la falta de respuesta que se vaya produciendo a lo largo del tiempo.

7.3.5 Análisis de Supervivencia

El análisis de supervivencia intenta estimar las probabilidades de que las unidades de la muestra del panel permanezcan en éste a lo largo del tiempo. Con este fin se emplean distintas técnicas, entre las que cabe destacar dos.

1. La regresión logística y otros modelos de análisis de datos categóricos. En estos modelos, la probabilidad de supervivencia es la variable explicada, y las variables explicativas pueden ser tanto variables de diseño como otras recogidas en los distintos instantes temporales.
2. El estimador de Kaplan-Meier de la función de supervivencia.