Metodología de Encuestas
Vol 4, Núm 1, 2002, 79-86
ISSN: 1575-7803

¿QUÉ ES UN MODELO DE SUPERPOBLACIÓN?

Román A. Pérez-Villalta
Universidad de Sevilla

RESUMEN

El enfoque clásico de las investigaciones por muestreo considera como fuente de aleatoriedad la que se deriva de la selección de la muestra. Cuando se considera como causa de la aleatoriedad la propia naturaleza de los datos y a ésta se le asigna cierta estructura aleatoria, se obtiene un modelo de superpoblación para la población. Este trabajo trata de divulgar estos modelos.

Palabras clave: *Muestreo, enfoque de la población fija, modelo de superpoblación, predicción de variables aleatorias, muestreo aleatorio simple.*

Introducción

El objetivo de este trabajo es explicar los conceptos del modelo de superpoblación y enfoque predictivo a profesionales y estudiosos de la investigación por muestreo, que posean ciertos conocimientos del enfoque clásico del muestreo en poblaciones finitas. En ese sentido, debe entenderse que es un trabajo divulgativo, no de investigación.

Siguiendo a Cassel y col. (1977), la aleatoriedad observada en una muestra puede proceder, básicamente, de tres fuentes:

(a) El método de selección de las unidades.
(b) Los métodos de medición de las variables en las unidades seleccionadas.
(c) El proceso que genera la verdadera medida de la variable para cada unidad.

El enfoque clásico de las investigaciones por muestreo considera como fuente de aleatoriedad la causa (a). Cuando se considera como causa de la aleatoriedad la propia naturaleza de los datos, es decir la fuente (c), y a ésta se le asigna cierta estructura aleatoria, se obtiene un modelo de superpoblación para la población finita objeto de estudio.

Este enfoque fue introducido formalmente por Godambe (1955) para suplir la limitación del enfoque tradicional de no existencia de estimadores óptimos, aunque implícitamente se encuentra en trabajos anteriores sobre estimadores de regresión, muestreo sistemático y comparación de varianzas. Al respecto, ver por ejemplo Cochran (1939, 1946), Deming y Stephan (1941), Madow y Madow (1944).

Para abordar el objetivo de este trabajo, el siguiente apartado en la sección dos se recuerda el enfoque de la población clásico de la población fija y el principio del muestreo repetido, básico en Estadística Matemática y que sirve de apoyo a la definición rigurosa de modelo de Superpoblación que se da a continuación. En el apartado siguiente, exponemos la forma en que, bajo un modelo de superpoblación, se pueden estimar parámetros de la población finita, así como la naturaleza de éstos bajo este enfoque de la inferencia. Por último, en el último apartado se estudia un modelo de superpoblación sencillo que da lugar al muestreo aleatorio simple.

Los enfoques de la población fija y de los modelos de superpoblación

El enfoque clásico del muestreo en poblaciones finitas considera que los valores x_i de la característica de interés asociados a una unidad u_i de una población finita U son fijos aunque desconocidos, salvo para los elementos de la muestra una vez que ha sido obtenida. Por tanto, esos valores no tienen la consideración de aleatorios (no se consideran variables aleatorias).

La aleatoriedad, en este enfoque, es fruto, exclusivamente, de la selección de la muestra y se refleja en el diseño muestral probabilístico (M,P) y en los estimadores utilizados que introducen variables indicadoras de la pertenencia de una unidad a la muestra cuya distribución sólo depende del sistema de selección utilizado. Así, por ejemplo, son profusamente utilizados los estimadores lineales del tipo

$$\hat{\theta} = \sum_{i=1}^{N} w_i x_i e_i$$

donde las variables aleatorias e_i son las indicadoras de la unidad *i*-ésima de la población y w_i son valores adecuados para que el estimador sea insesgado del parámetro a estimar. Al respecto, ver por ejemplo, Azorín y Sánchez Crespo (1986), Mirás (1985), Fernández y Mayor (1995). Esta concepción del muestreo se denomina *enfoque de la población fija.*

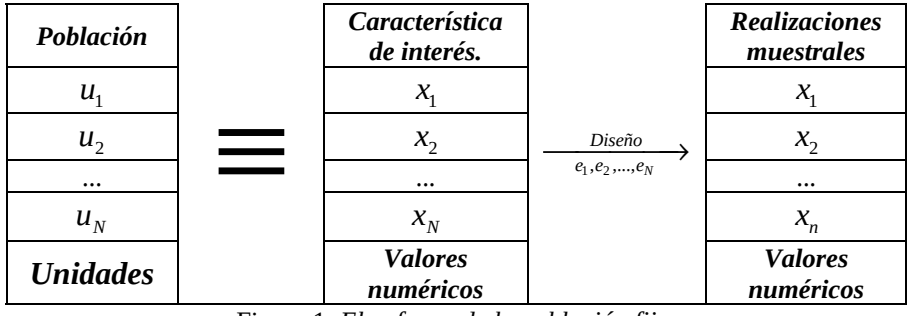

Figura 1: *El enfoque de la población fija*.

Frente a este enfoque, se proponen los *modelos de superpoblación* que hacen uso de las ideas durante largo tiempo utilizadas en Estadística Teórica.

Un dato *x* es un número real que procede de una población tras ser investigada o sujeta a experimentación. Si repetimos la investigación o experimento, en general, obtenemos otro dato *x´* y si repetimos esto en muchas ocasiones iremos obteniendo diversos valores *x, x´, x´´...* que pueden ser considerados realizaciones muestrales de cierta variable aleatoria ***X*** (esto recibe el nombre de *principio de muestreo repetido*, ver Azzalini, 1996). En consecuencia, si un dato es desconocido puede ser considerado una variable aleatoria y, cuando lo conozcamos, será una realización de ella. Además, ***X*** tendrá una distribución completa o parcialmente especificada.

Ejemplo 1

Supóngase que lanzamos una moneda, si sale cara lo simbolizamos por un uno y si sale cruz por un cero. Es claro que, antes de lanzar la moneda por primera vez, no sabemos qué vamos a obtener, sólo conocemos que saldrá cara con una probabilidad *p* o cruz con una probabilidad *1-p*; de esta forma, tenemos una variable aleatoria y su distribución, conocida salvo un parámetro, concretamente, $X \sim B(1,p)$. De la misma forma se puede razonar con el resto de tiradas, hasta un total de ***n.*** Después de tirar la moneda se obtienen valores numéricos, ceros o unos, que serán realizaciones de la variable aleatoria correspondiente. En definitiva se obtiene la disposición que se muestra en la cuadro 1.

Tirada	Antes de tirar	Después de tirar
1	$X_1 \sim B(1,p)$	$x_1 \in \{0,1\}$
2	$X_2 \sim B(1,p)$	$x_2 \in \{0,1\}$
…	…	…
N	$X_n \sim B(1,p)$	$x_n \in \{0,1\}$
	Variables aleatorias	Datos numéricos

Cuadro 1: *Experimento 'tirar una moneda'*.

Ejemplo 2

Se tiene una población de $\boldsymbol{N}$ = 2000 individuos y se desea estudiar el tamaño de su unidad familiar. Aquí, antes de preguntar al individuo, tenemos una variable aleatoria con distribución, en principio, desconocida aunque sí sabemos que es discreta y que nos indica el tamaño de la unidad familiar del individuo preguntado. Cuando éste responde (véase el cuadro 2), la variable aleatoria se convertirá en un valor numérico, que es la realización muestral.

Individuo	Antes de preguntar	Después de preguntar
1	$X_1 \sim F$	$x_1 = 0,1,2...$
2	$X_2 \sim F$	$x_2 = 0,1,2...$
…	…	…
N	$X_{2000} \sim F$	$x_n = 0,1,2...$
	Variables aleatorias	Datos numéricos

Cuadro 2: *Realización muestral en una pregunta de la encuesta*.

Nótese que en este caso nuestro conocimiento sobre el modelo es escaso. Los modelos de superpoblación aplican estas ideas al muestreo de poblaciones finitas.

Dada una población $\boldsymbol{U}$ de $\boldsymbol{N}$ elementos $\{u_1, u_2, ... , u_N\}$, los valores de la variable de interés $x_1, x_2, ... , x_N$ nos son desconocidos y, por lo expuesto más arriba, pueden ser considerados variables aleatorias $X_1, X_2, ... , X_N$. De esta forma, cada unidad lleva asociada una variable aleatoria y la población, en cuanto a la característica de interés, puede identificarse con un vector aleatorio

$$\vec{X} = (X_1, X_2, ..., X_N)$$

cuya distribución seguirá un modelo más o menos conocido. A la distribución de $\vec{X}$ se le denomina *modelo de superpoblación*.

Cuando seleccionamos una muestra, estamos seleccionando un conjunto de $\boldsymbol{n}$ variables aleatorias que, para simplificar la notación, denotaremos

$$(X_1, X_2, ..., X_N)$$

Tras estudiarla, obtendremos la realización muestral como conjunto de concreciones de las variables aleatorias que integran la muestra.

Esquemáticamente, el proceso queda como se muestra en la figura 2.

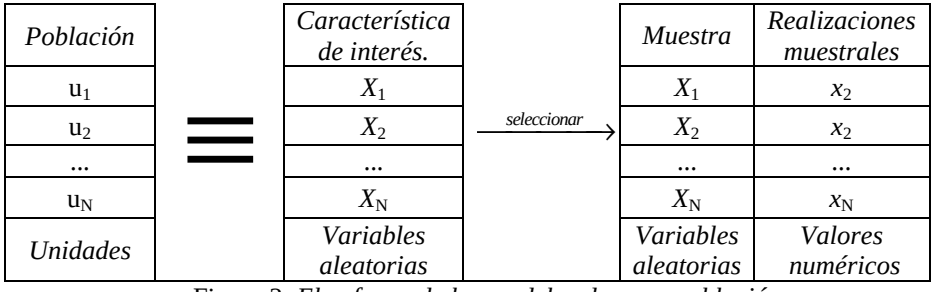

Figura 2: *El enfoque de los modelos de superpoblación.*

Estimación: el enfoque predictivo

Un cambio tan radical en la concepción de la población y de la naturaleza de los valores de la característica objeto de estudio conlleva un cambio, también radical, en el planteamiento del proceso de estimación.

En primer lugar, observemos que cualquier parámetro de la población finita en el que nos interesemos, del tipo

$$\boldsymbol{\theta} = g(x_1, x_2, ..., x_N)$$

es una función de los valores de la característica sobre los individuos de la población. Sin embargo, bajo este enfoque estos se identifican con una variable aleatoria. Por tanto, un parámetro es una función de variables aleatorias y es, en sí mismo, una variable aleatoria que denotaremos

$$Y = g(X_1, X_2, ..., X_N)$$

Dada la muestra

$$\vec{X}_s = (X_1, X_2, ..., X_N)$$

el enfoque predictivo de la estimación trata de aproximarse a $\boldsymbol{\theta}$ mediante la predicción del valor que tome la variable aleatoria *Y* basándonos en alguna función de la muestra $\vec{X}_s$, de forma que se aproxime a *Y* en algún sentido.

Denotemos por $\hat{g}(\vec{X}_s) = \hat{g}(X_1, X_2, ..., X_N)$ a la función de la muestra que predice el valor de ***Y*** (*predictor de Y).* Siguiendo las técnicas habituales de la estimación, le exigiremos que $\hat{g}(\vec{X}_s)$ sea insesgado y que su error cuadrático medio sea lo menor posible, entendiendo aquí que las esperanzas se toman respecto a la distribución del modelo de superpoblación.

La insesgadez, en este contexto, se transforma en la condición:

$$\text{(a)} \quad E\left[\hat{g}(\vec{X}_s)\right] = E\left[g(\vec{X})\right]$$

Pues ambas funciones son variables aleatorias. Esto se denomina *insesgadez respecto al modelo de superpoblación o ξ-insesgadez,* que es distinta de la insesgadez del enfoque de la población fija $E\left[\hat{\theta}\right]=\theta$ que llamaremos *insesgadez respecto al diseño.*

Por su parte, el error cuadrático medio queda:

$$\text{(b)}\quad E\left[\left(g\left(\vec{X}\right)-\hat{g}\left(\vec{X}_s\right)\right)^2\right]$$

que exigiremos sea mínimo.

Obsérvese que estas esperanzas se toman respecto de la distribución del modelo de superpoblación.

Una vez obtenido el predictor y la realización muestral, la estimación del parámetro será $t=\hat{g}\left(x_1,x_2,...,x_n\right)$, que es la predicción del valor de *Y*.

Ejemplo: el muestreo aleatorio simple.

Existen gran variedad de modelos de superpoblación en la literatura. En general, podemos decir que los principales modelos estadísticos, tales como los lineales, polinómicos ..., han sido utilizados en diversas aplicaciones. Véase Trader (1982) para una sucinta relación. A título de ejemplo, estudiaremos uno de los más sencillos, que da lugar al muestreo aleatorio simple.

Supongamos la existencia de una población finita de *N* elementos, de la que estamos interesados en estudiar cierta característica cuantitativa *X,* definida sobre los individuos de la población. Además, supondremos que, por su propia naturaleza, haciendo abstracción de la población objeto de estudio, la característica *X* puede ser modelizada por una variable aleatoria de cuya distribución conocemos la media y la varianza. Concretamente, asumimos el siguiente modelo para la población $\vec{X}$.

$$X_1,X_2,...,X_N \text{ son independientes}$$
$$E\left[X_i\right]=m \qquad \text{var}(X_i)=v \qquad i=1,2,...N$$

Para estimar el parámetro poblacional

$$\tau=\sum_{i=1}^{N}x_i$$

desde el enfoque del modelo de superpoblación propuesto, debemos predecir el valor de la variable aleatoria

$$g\left(\vec{X}\right)=g\left(X_1,X_2,...,X_N\right)=\sum_{i=1}^{N}X_i$$

basándonos en la muestra aleatoria, de tamaño *n*, $\vec{X}_m = (X_1, X_2, ..., X_n)$. Para ello, utilizamos una función de la muestra, predictor, $\hat{g}(X_1, X_2, ..., X_n)$. Puesto que el parámetro es de tipo lineal, no parece descabellado buscar un predictor también de tipo lineal.

$$\hat{g} = \hat{g}(\vec{X}_s) = \hat{g}(X_1, X_2, ..., X_n) = \sum_{i=1}^{N} \alpha_i X_i$$

donde $\alpha_1, \alpha_2, ..., \alpha_n \in \Re$

Debemos encontrar ahora los valores de $\alpha_1, \alpha_2, ..., \alpha_n \in \Re$ que hagan que el predictor sea óptimo en el sentido expuesto en la sección anterior: insesgadez respecto al modelo y error cuadrático mínimo. Esto es:

$$\text{(a)} \quad E\left[\hat{g}(\vec{X}_s)\right] = E\left[g(\vec{X})\right]$$

$$\text{(b)} \quad E\left[\left(\hat{g}(\vec{X}_s) - g(\vec{X})\right)^2\right] \text{ mínimo}$$

Es fácil ver que la condición (*a*) es equivalente a $N = \sum_{i=1}^{n} \alpha_i$ y, por tanto, el problema consiste en encontrar $\alpha_1, \alpha_2, ..., \alpha_n \in \Re$ tales que:

$$E\left[\left(\sum_{i=1}^{N} X_i - \sum_{i=1}^{n} \alpha_i X_i\right)^2\right] \text{ mínimo}$$

$$\text{sujeto a } N = \sum_{i=1}^{n} \alpha_i$$

El problema se aborda mediante los multiplicadores de Lagrange que dan lugar a

$$\alpha_1 = \alpha_2 = ... = \alpha_n = \frac{N}{n}$$

y, por tanto, el predictor que buscábamos es

$$\hat{g}(\vec{X}_s) = \sum_{i=1}^{n} \frac{N}{n} X_i = N \frac{\sum_{i=1}^{n} X_i}{n} = N\overline{X}$$

Calculando el error cuadrático se obtiene:

$$EC = E\left[\left(\sum_{i=1}^{N} X_i - \sum_{i=1}^{n} \alpha_i X_i\right)^2\right] = \left(1 - \frac{N}{n}\right) N^2 \frac{\nu}{n}$$

Así, en este modelo de superpoblación, el predictor y su error cuadrático coincide con el estimador y su error cuadrático en el diseño muestral aleatorio simple; por tanto, podemos concluir que ese diseño es adecuado cuando la estructura de los datos sigue el modelo propuesto.

Otros modelos que coinciden, en el mismo sentido que el anterior, con diseños clásicos pueden verse en Lohr (2000).

Referencias.

Azorín. F. y Sánchez-Crespo, J.L. (1986). *Métodos y aplicaciones del muestreo.* Madrid: Alianza.

Azzalini, A.(1996) *Statistical Inference. Based on the Likelihood*. London: Chapman & Hall.

Cassel, C.M., Särdal, E.E. y Wretman, J.H. (1977) *Foundation of Inference in Survey Sampling.* Malabar: Krieger Publishin Company.

Cochran, W.G.(1939) The use of analysis of variance in enumuration of sampling. *Journal of the American Statistical Association,* 34, 492-510.

Cochran, W.G.(1946) Relative accuracy of systematic and stratified random samples for a certain class of population. *Annals of Mathematical Statistics,* 17, 164-177.

Deming W.E. y Stephan F. (1941) On the interpretation of censuses as samples. *Journal of the American Statistical Association,* 36, 45-49.

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

Godambe,V.P. (1955)A unified theorie of sampling from finite population. *Journal of the Royal Statistical Society.* Series B, 17, 269-278.

Lohr, Sh,L.(2000) *Muestreo: diseño y análisis.* México: International Thompson Editores

Madow W.G. y Madow L.H.(1944) On the theory of systematic sampling. *Annales of Mathematics Statistics,* 15, 1-24.

Mirás, J. (1985) *Elementos de muestreo para poblaciones finitas*. Madrid: INE.

Trader, R.L. (1982) Superpopulation´s models. En S. Kotz y L. Norman y N.L. Johnson (eds.) *Encyclopedia of statistical sciences***.** New York: John Wiley and Sons, 1982-1989.