QÜESTIIÓ, Vol 12, nº 3, pp. 323-332, 1988

UN ENSAYO SOBRE CRITERIOS DE INFERENCIA EN POBLACIONES FINITAS

M. RUIZ ESPEJO

Universidad Complutense de Madrid

Realizamos un análisis de los criterios prácticos para la inferencia en poblaciones finitas, y proponemos un criterio experimental nuevo para decidir el estimador o estrategia preferible basada en una muestra piloto.

Clasificación AMS: 62 D 05, 62 C 99.

An essay on inference criteria in finite populations.

Keywords: Inference criteria, sampling in finite populations.

1. INTRODUCCIÓN

Han sido variados los criterios de inferencia que tratan de dar pautas para la elección de diseño muestral y de estimador que deben utilizarse para estimar ciertas funciones paramétricas de interés como son la media poblacional, la varianza poblacional, la razón de totales poblacionales, el coeficiente de regresión poblacional de la variable de interés sobre cierta variable auxiliar, etc. referidos a una población finita fija de tamaño N, como es usual en la bibliografía (Cassel, Särndal y Wretman, 1977 ó Wolter, 1985).

Actualmente y en la segunda mitad de nuestro siglo estamos presenciando como diversos autores tratan de justificar cierto estimador o clase de estimadores por tal o cual criterio, presentando razones que pueden inducir a su uso. En este camino influyen aspectos de la inferencia clásica y sus técnicas ya conocidas para la elección de estimadores de parámetros de poblaciones infinitas, como elementos de la teoría de la decisión estadística al considerar

–M. Ruiz Espejo - Dept. d'Organització d'Empreses. Facultat de Ciències Econòmiques i Empresarials - Universitat Complutense - 28023 Madrid.

–Article rebut el desembre de 1988.

que existe una incertidumbre total en los posibles valores a observar de la variable de interés en cada unidad de la población finita, o incluso modelos superpoblacionales que encauzan de algún modo dicha incertidumbre mediante la adopción de hipótesis sobre cómo se generan los datos que se observan en cada unidad de la población. Como consecuencia, al lector no introducido en estos estudios puede parecer confuso, incluso a veces paradójico el panorama que se desprende de las distintas propuestas que se justifican de algún modo o bajo los más variados criterios.

Veamos pues una visión general de algunos de estos criterios haciendo oportunos comentarios sobre su alcance y significado en el contexto global de la inferencia en poblaciones finitas.

2. ESTIMABILIDAD

La estimabilidad de una función paramétrica para un diseño dado consiste en la existencia de un estadístico que con tal diseño sea insesgado para la función paramétrica. Este criterio usual en la inferencia clásica fué introducido y estudiado en el contexto de poblaciones finitas por primera vez por Hanurav (1966), quien caracterizó las funciones paramétricas cuadráticas estimables con respecto a un diseño dado. Recientemente Ruiz (1986b) ha extendido este criterio a una variedad mucho más amplia de funciones paramétricas.

No obstante el interés se centra usualmente en ciertas funciones paramétricas como son la media y la varianza poblacionales. Fijada una de estas funciones, un estimador (para un diseño dado) es insesgado si el operador esperanza matemática al actuar sobre nuestro estimador resulta igual a la función paramétrica a estimar. Existen muchos estimadores ya clásicos en teoría de muestras que son sesgadas bajo diseños muy sencillos a utilizar en la práctica, como son los estimadores de razón, producto o de regresión para la media poblacional, sin embargo aunque no son insesgados bajo el diseño de muestreo aleatorio simple poseen en ciertas condiciones unas cualidades muy deseables como son su acuracidad o poca dispersión entorno a la función paramétrica a estimar, cosa que no es exigida por la estimabilidad o insesgación bajo cierto diseño.

Así la estimabilidad de una función paramétrica es debida a la existencia de algún estimador insesgado. La insesgación por sí sola puede ser un mal criterio de selección de estimadores o diseños ya que puede presentarse en la práctica unida a una gran dispersión. Para algunos estadísticos esta dispersión debería ser medida por la desviación media o error absoluto medio, pero sus grandes dificultades en el cálculo real, ha hecho interesarles en la práctica por el error cuadrático medio y, en el caso de ser insesgado el estimador, por su varianza.

3. SUFICIENCIA

El criterio de suficiencia se basa en reducir los datos observados a una expresión más sencilla que conserve la información esencial para el problema de inferencia concreto. Según justificaron Basu y Ghosh (1967) el dato consistente en las unidades identificadas sin conservar el orden y multiplicidad al ser observadas, es suficiente minimal para el parámetro $\underline{y} = (y_1, ..., y_N)$ y por ello para cualquier función paramétrica basada en $\underline{y}$. Es minimal por ser la reducción de los datos originales más fuerte posible sin dejar de ser suficiente.

Otros tipos de suficiencia son la lineal, estudiada por Godambe (1966), la suficiencia de distribución libre (Godambe, 1966) y la suficiencia bayesiana (Godambe, 1968).

El teorema de Rao-Balckwell adaptado a la teoría de muestras (Ruiz, 1988) permite dar una clase esencialmente completa de estimadores o estrategias (diseño, estimador) en cada problema de inferencia. Todo ello justificaría limitarnos a considerar estimadores que no dependen del orden y multiplicidad de unidades en la muestra. Sin embargo la aceptación de estas conclusiones no ha sido general, debido a que al hacer intervenir costes, el muestreo con reemplazamiento, que considera el orden y multiplicidad de las unidades seleccionadas, puede presentar ventajas económicas como ya indicó Pathak (1962).

4. COMPLETITUD Y ESTIMADORES UMV

El estadístico suficiente minimal visto anteriormente no es completo. Si hubiera sido completo podría verse si existe o no un estimador óptimo en sentido uniformemente de mínima varianza, UMV. El teorema más amplio sobre completitud presentado hasta ahora se debe a Liu (1983), pero se desarrolla en condiciones relativamente restringidas.

Pero la no existencia de estimador uniformemente de mínima varianza fué justificada por Godambe (1955) en la clase de los estimadores lineales homogéneos insesgados y posteriormente Godambe y Joshi (1965) ampliaron dicho resultado a la clase de todos los estimadores insesgados. Este criterio se reduciría a comprobar que existe o no un estimador $t_o \in \mathcal{C}$ verificando

$$(1) \qquad V(t_o) \leq V(t) \qquad \forall t \in \mathcal{C} \quad \text{y} \quad \forall \underline{y} \in \mathbf{R}^N$$

donde $\mathcal{C}$ es una clase de estimadores insesgados y el término $\forall \underline{y} \in \mathbf{R}^N$ significa uniformemente. Han sido varios los autores que han estudiado este criterio de la existencia o no de estimador uniformemente de mínima varianza. En las condiciones más generales, Ruiz (1987b) ha justificado que no existe tan deseable estimador, pues de existir sería la elección natural de la clase de todos los estimadores insesgados, cuando ni siquiera existe estimador uniformemente de mínimo error cuadrático medio en la clase de todos los estimadores.

Otros autores, como Särndal (1976), consiguen demostrar la existencia de estimadores uniformemente de mínima varianza en condiciones restringidas, para ciertas clases de estimadores bajo ciertos diseños, lo que le hace tener poca importancia general en la práctica.

5. ADMISIBILIDAD Y CRITERIOS NO EXPERIMENTALES

Ya que no existe un estimador mejor, o de menor varianza uniformemente, que los demás en la clase de todos los estimadores insesgados de cierta función paramétrica, el interés se centró desde un punto de vista de teoría de la decisión en buscar estrategias admisibles o no superadas uniformemente en su clase de estrategias. Además de los trabajos presentados por Cassel, Särndal y Wretman (1977) podemos citar en esta dirección los de Bellhouse y Joshi (1984), Biyani (1980), Chang (1979, 1982,1983), Chaudhuri (1978), Chaudhuri y Adhikary (1983, 1985), Hege (1966), Joshi (1977, 1979a), Liu (1974), Liu y Thompson (1983), Meeden, Ghosh y Vardeman (1985), Patel y Dharmadhikari (1978), Ruiz (1987a), Sankaranarayanan (1980), Sengupta (1983), Strauss (1982), Tsui (1983), Vardeman y Meeden (1983a, 1983b, 1984). En ellos se considera usualmente fijado el diseño y se trata de encontrar estimadores admisibles; en otros casos es fijado el estimador y se caracterizan los diseños admisibles. En algunos casos se trata de encontrar estrategias (diseño, estimador) admisibles cuando tanto el diseño como el estimador son variables dentro de sus clases respectivas. La conclusión obtenida es que de no existir un estimador UMV pasamos a disponer de clases infinitas de diseños y estimadores (así como de estrategias) admisibles.

Si la población finita es homogénea, Joshi (1979b) incluso muestra que la estrategia (mas, $\bar{y}_s$), de tamaño efectivo fijo n y donde $\bar{y}_s$ es la media muestral, es óptima al minimizar (para cualquier función de pérdida convexa) el riesgo máximo y el riesgo medio sobre el conjunto de parámetros que se obtienen permutando los indicadores de las unidades de la población, en la clase de estrategias insesgadas sujetas a la restricción de que el tamaño muestral no exceda n.

Estos criterios anteriores son de tipo no experimental, en el sentido de que no influye en la decisión admisible u óptima los datos observables de la variable de interés. Pero en muchos casos esta previa experimentación es factible mediante una muestra piloto o incluso la misma muestra observada para hacer la inferencia que dará información parcial sobre el parámetro y por tanto es posible considerar a éste no como incierto, sino como incierto parcialmente previamente a la decisión sobre qué estrategia tomar, permitiéndonos delimitar lo "incierto total a priori" (tal y como consideran al parámetro el criterio de admisibilidad, hiperadmisibilidad (Hanurav, 1968), etc.) y conocer datos previos para la toma de decisión. Esto nos induce a considerar una nueva forma de decidir la estrategia a tomar para estimar funciones paramétricas en la práctica.

6. TOMA DE DECISIÓN EXPERIMENTAL

La conclusión de no disponer de un único estimador admisible en las clases más interesantes desde un punto de vista práctico, e incluso de no disponer de una única estrategia admisible, derivó el interés a estudiar el criterio de hiperadmisibilidad para estimar la media poblacional, fué criticado desde muy diversos ángulos.

El criterio más conocido que condiciona la inferencia sobre el parámetro $\underline{y}$ a los datos muestrales es el de máxima verosimilitud, estudiado inicialmente por Godambe (1966, 1975). La conclusión que se sigue de sus estudios es que la única inferencia aportada sobre $\underline{y}$ por el principio de máxima verosimilitud es la trivial que a los componentes y_i para $i \in s$ coinciden con los valores observados, no pudiendo obtener ninguna información sobre los valores no muestreados y_i, con $i \notin s$. Pero estas conclusiones son muy controvertidas ya que no son compartidas por los estadísticos prácticos que disponen de otras informaciones auxiliares y experiencias (Smith, 1976).

Como una solución a estos resultados parcialmente insatisfactorios, vamos a proponer un método para decidir entre un conjunto finito de estrategias para estimar una función paramétrica $f(\underline{y})$ basándonos en un experimento piloto.

La técnica de estimar anticipadamente las varianzas es muy conocida y ha sido utilizada para la determinación de los tamaños muestrales (Cochran, 1977). Haciendo uso de esta técnica podemos sustituir el criterio de comparación no experimental, dado en (1) para una clase de estimadores C insesgados de $f(\underline{y})$, por un nuevo criterio empírico al basarse en el dato d proporcionado por una muestra preliminar s. Es decir, si C es una clase finita de estimadores insesgados, $t_o \in C$ es preferible a $t \in C$ en base al experimento piloto si y sólo si

$$\hat{V}(t_o|d) \leq \hat{V}(t|d) \qquad \forall t \in C$$

siendo d el dato asociado a la muestra piloto s y el acento circunflejo significa que ambas expresiones son estimaciones respectivas de $V(t_o)$ y $V(t)$. Con esta ordenación de estimadores, surge un nuevo criterio de decisión empírico que selecciona al estimador t_o de entre los de la clase finita C si y sólo si

$$\hat{V}(t_o|d) = \min_{t \in C} \hat{V}(t|d) \quad ,$$

decisión que depende de la calidad de los estimadores de las varianzas y de la muestra piloto seleccionada. Modificando la condición de insesgación de los estimadores, podemos sustituir la varianza V por el error cuadrático medio ECM como medida de dispersión.

Un caso particular de esta técnica puede verse en Ruiz (1986a).

7. ALGUNOS EJEMPLOS DEL CRITERIO PROPUESTO

Veamos algunos ejemplos prácticos sencillos en que comparamos varios métodos de estimación con la estrategia (mas, $\bar{y}_s$) justificada por Joshi (1979).

a) El estimador de traslación t_T para diseños postcensales es

$$t_T = \bar{y}_s + \bar{x} - \bar{x}_s = \frac{1}{n} \sum_{k \in s} (y_k + \bar{x} - x_k) = \bar{z}_s$$

siendo la variable auxiliar x la variable de interés en un censo reciente de la misma población, e y la variable de interés en la investigación actual, y $z_k = y_k + \bar{x} - x_k (k = 1, 2, ..., N)$. Entonces

$$\hat{V}(t_T|d) = \frac{N-n}{N} \cdot \frac{\hat{S}_z^2}{n} \quad \text{y} \quad \hat{V}(\bar{y}_s|d) = \frac{N-n}{N} \cdot \frac{\hat{S}_y^2}{n}$$

y por tanto t_T es preferible a $\bar{y}_s$ cuando $\hat{S}_z^2 \leq \hat{S}_y^2$, siendo S^2 la cuasivarianza poblacional, estimable de modo insesgado por la cuasivarianza muestral.

b) El estimador de razón bifásico $t'_R = \bar{y}_{s_2} \bar{x}_{s_1} / \bar{x}_{s_2}$ es aproximadamente insesgado (Sukhatme y Sukhatme, 1970) y su varianza aproximada es

$$V(t'_R) = \frac{N-n_1}{Nn_1} S_y^2 + \frac{n_1 - n_2}{n_1 \cdot n_2} \left(S_y^2 + R^2 S_x^2 - 2RS_{yx}\right)$$

pudiéndose calcular los tamaños óptimos de n_1 y n_2

$$n_1 = \frac{C - c_o}{c_1 + \sqrt{\dfrac{c_1 \cdot c_2}{2RS_{yx} - R^2 S_x^2}}} \quad \text{y} \quad n_2 = \frac{C - c_o}{c_2 + \sqrt{\dfrac{c_1 c_2 (2RS_{yx} - R^2 S_x^2)}{S_y^2 + R^2 S_x^2 - 2RS_{yx}}}} \quad ,$$

que minimizan $V(t'_R)$ sujetos a la función de coste

$$C = c_o + c_1 n_1 + c_2 n_2 \quad ,$$

que son estimables por los datos obtenidos de la muestra piloto de presupuesto c_o, siendo C el presupuesto total, c_1 el costo de observar la variable auxiliar en una unidad y c_2 el costo de una observación de la variable de interés y. Así utilizando los estimadores $\hat{R}$, $\hat{S}_{xy}$, $\hat{S}_x^2$ y $\hat{S}_y^2$ es inmediato estimar n_1 y n_2, concluyendo que sólo si

$$\left(\frac{c_2}{C - n_o c_1} - \frac{1}{\hat{n}_2}\right) \hat{S}_y^2 \geq \left(\frac{1}{\hat{n}_2} - \frac{1}{\hat{n}_1}\right) \left(\hat{R}^2 \hat{S}_x^2 - 2\hat{R}\hat{S}_{yx}\right)$$

es preferible t'_R a $\bar{y}_s$ de tamaño muestral $(C - n_o c_1)/c_2$.

c) Con el estimador de regresión bifásico $t'_{RG} = \bar{y}_{s_2} + b\left(\bar{x}_{s_1} - \bar{x}_{s_2}\right)$, con

$$b = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(x_i - \bar{x}_{s_2}\right)\left(y_i - \bar{y}_{s_2}\right) \Big/ \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(x_i - \bar{x}_{s_2}\right)^2 \quad ,$$

se tiene que es aproximadamente insesgado (Sukhatme y Sukhatme, 1970) y su varianza aproximada es

$$V(t'_{RG}) = \frac{N - n_1}{N n_1} S_y^2 + \frac{n_1 - n_2}{n_1 n_2} S_y^2 \left(1 - \rho_{yx}^2\right)$$

que puede optimizarse al variar n_1 y n_2 sujetos a la función de coste

$$C = c_o + c_1 n_1 + c_2 n_2 \quad ,$$

obteniendo los valores óptimos

$$n_1 = \frac{C - c_o}{c_1 + \sqrt{\dfrac{c_1 c_2 \left(1 - \rho_{yx}^2\right)}{\rho_{yx}^2}}} \quad \text{y} \quad n_2 = \frac{C - c_o}{c_2 + \sqrt{\dfrac{c_1 c_2\, \rho_{yx}^2}{1 - \rho_{yx}^2}}}$$

que son estimables en base al estadístico $\hat{\rho}_{yx}^2$. Por tanto si

$$\frac{c_2}{C - n_o c_1} \geq \frac{1}{\hat{n}_2}\left(1 - \hat{\rho}_{yx}^2\right) + \frac{1}{\hat{n}_1}\hat{\rho}_{yx}^2$$

es preferible t'_{RG} a $\bar{y}_s$ de tamaño muestral $(C - n_o c_1)/c_2$.

8. BIBLIOGRAFÍA

[1] **Basu, D.** y **Ghosh, J.K.** (1967). "Sufficient statistics in sampling from a finite universe". Proc. 36th Session Internat. Statist. Inst. 850-859.

[2] **Bellhouse, D.R.** y **Joshi, V.M.** (1984). "On the admissibility of the regression estimator". J. Roy Statist. Soc. B 46, 268-269.

[3] **Biyani, S.H.** (1980). "On inadmissibility of the Yates-Grundy variance estimator in unequal probability sampling". J. Amer. Statist. Assoc. 75, 709-712.

[4] **Cassel, C.M., Sãrndal, C.E.** y **Wretman, J.H.** (1977). "Foundations of Inference in Survey Sampling". Wiley, New York.

[5] **Cochran, W.G.** (1977). "Sampling Techniques" (3rd. edition). Wiley, New York.

[6] **Chang, H.J.** (1979). "On a class of uniformly admissible strategies in sampling survey". Soochow J. Math. 5,185-191.

[7] **Chang, H.J.** (1982). "A class of admissible regression estimator's in sampling from finite populations". Tamkang J. Management Sci. 3, 7-16.

[8] **Chang, H.J.** (1983). "A class of admissible estimators of a finite population total and their uniform admissibility". Tamkang J. Management Sci. 4, 13-22.

[9] **Chauduri, A.** (1978). "On estimating the variance of a finite population". Metrika 25, 65-76.

[10] **Chauduri, A.** y **Adhikary, A.K.** (1983). "On optimality of double sampling strategies with varying probabilities". J. Statist. Plann. Inference 8, 257-265.

[11] **Chauduri, A.** y **Adhikary, A.K.** (1985). "Some results on admissibility and uniform admissibility in double sampling". J. Statist. Plann. Inference 12, 199-202.

[12] **Godambe, V.P.** (1985). "A unified theory of sampling from finite populations". J. Roy. Statist. Soc. B 17, 269-278.

[13] **Godambe, V.P.** (1966). "A new approach to sampling from finite populations I, II.". J. Roy. Statist. Soc. B 28, 310-328.

[14] **Godambe, V.P.** (1968). "Bayesian sufficiency in survey sampling". Ann. Inst. Statist. Math. 20, 363-373.

[15] **Godambe, V.P.** (1975). "Likelihood principle and randomization". Unpublished report.

[16] **Godambe, V.P.** y **Joshi, V.M.** (1965). "Admissibility and Bayes estimation in sampling finite populations I.". Ann. Math. Statist. 36, 1707-1722.

[17] **Hanurav, T.V.** (1966). "Some aspects of unified sampling theory". Sankhyā Ser. A 28, 175-204.

[18] **Hanurav, T.V.** (1968). "Hyperadmissibility and optimum estimators for sampling finite populations". Ann. Math. Statist. 39, 621-642.

[19] **Hege, V.S.** (1966). "Admissible estimates for any sampling design". Bull. Calcutta Statist. Assoc. 15, 120-126.

[20] **Joshi, V.M.** (1977). "A note on estimators in finite polulations". Ann. Statist. 5, 1051-1053.

[21] **Joshi, V.M.** (1979a). "Joint admissibility of the sample means as estimators of the means of finite populations". Ann. Statist. 7, 995-1002.

[22] **Joshi, V.M.** (1979b). "The best strategy for estimating the mean of a finite population". Ann. Statist. 7, 531-536.

[23] **Liu, T.P.** (1974). "A general unbiased estimation for the variance of a finite population. Sankhyā 36, 23-32.

[24] **Liu, T.P.** (1983). "A general completeness theorem in sampling theory". J. Roy. Statist. Soc. B 45, 369-372.

[25] **Liu, T.P.** y **Thompson, M.E.** (1983). "Properties of estimators of quadratic finite populations functions: the batch approach". Ann. Statist. 11, 275-285.

[26] **Meeden, G.**; **Ghosh, M.** y **Vardeman, S.** (1985). "Some admissible nonparametric and related finite population sampling estimators". Ann. Statist. 13, 811-817.

[27] **Patel, H.C.** y **Dharmadhikari, S.W.** (1978). "Admissibility of Murthy's and Midzuno's estimators within the class of linear unbiased estimators of finite populations totals". Sankhyā Ser. C 40, 21-28.

[28] **Pathak, P.K.** (1962). "On simple random sampling with replacement". Sankhyā Ser. A. 24, 287-302.

[29] **Ruiz, M.** (1986a). "El estimador de traslación para muestreos post-censales". Estadíst. Española 111, 81-86.

[30] **Ruiz, M.** (1986b). "Funciones paramétricas estimables en teoría de muestras". Estadíst. Española 112-113, 69-73.

[31] **Ruiz, M.** (1987a). "Diseños muestrales admisibles para el estimador Horvitz-Thompson". Trabajos Estadíst. 2, 45-50.

[32] **Ruiz, M.** (1987b). "Sobre estimadores UMV y UMECM en poblaciones finitas". Estadíst. Española 115, 105-111.

[33] **Ruiz, M.** (1988). "El teorema de Rao-Blackwell en poblaciones finitas". Actas XVII Reunión Nacional de Estadística, I.O. e Informática. Benidorm.

[34] **Sankaranarayanan, K.** (1980). "A note on the admissibility of some non-negative quadratic estimators". J. Roy. Statist. Soc. B 42, 387-389.

[35] **Särndal, C.E.** (1976). "On uniformly minimum variance estimation in finite populations". Ann. Statist. 4, 993-997.

[36] **Sengupta, S.** (1983). "Admissibility of unbiased estimators in finite populations sampling for samples of size at most two". Bull. Calcutta Statist. Assoc. 32, 91-102.

[37] **Smith, T.M.F.** (1976). "The foundations of survey sampling: a review". J. Roy. Statist. Soc. A 139, 183-204.

[38] **Strauss, I.** (1982). "On the admissibility of estimators for the finite population variance". Metrika 29, 195-202.

[39] **Sukhatme, P.V.** y **Sukhatme, B.V.** (1970). "Sampling Theory of Surveys with Applications". (2nd. edition). Iowa State University Press, Ames.

[40] **Tsui, K.W.** (1983). "A class of admissible estimators of a finite population total". Ann. Inst. Statist. Math. 35, 25-30.

[41] **Vardeman, S.** y **Meeden, G.** (1983a). "Admissible estimators in finite population sampling employing various types of prior information". J. Statist. Plann. Inference 7, 329-341.

[42] **Vardeman, S.** y **Meeden, G.** (1983b). "Admissible estimators of the population total using trimming and Winsorization". Statist. Probab. Lett. 1, 317-321.

[43 **Vardeman, S.** y **Meeden, G.** (1984). "Admissible estimators for the total of a stratified population that employ prior information". Ann. Statist. 12, 675-684.

[44] **Wolter, K.M.** (1985). " Introduction to Variance Estimation". Springer-Verlag, New York.