

# **ESTIMACION DE LA MEDIANA POR INTERVALOS DE CONFIANZA CON INFORMACION AUXILIAR EN POBLACIONES FINITAS**

A. Arcos Cebrián

M. Rueda Garcia

E. Artés Rodríguez(\*)

Departamento de Estadística e Investigación Operativa.

Universidad de Granada.

(\*) Universidad de Almería.

## **Resumen**

En este trabajo se propone un método alternativo al usual (Woodruff, 1952) para la construcción de intervalos de confianza usando información auxiliar cuando se quiere estimar la mediana de una población finita. Mediante un estudio de simulación se muestra como este método mejora notablemente al clásico.

## **1 Introducción.**

Es conocido que cuando se quiere dar una medida de posición central que proporcione información resumida sobre la tendencia central de determinada variable observada en una población finita, la mediana puede ser una medida que refleje mejor esta tendencia que la media aritmética. Tal es el caso de variables, como los ingresos, que pueden presentar valores extremos que alteran en gran medida el valor obtenido para la media, mientras que la mediana permanece inalterable pues sólo usa la información proporcionada por los valores centrales. En estos casos, el utilizar la media por aprovechar toda la información puede conducir a resultados poco representativos, pues al verse afectada por los valores extremos hace que cuando éstos sean altos, aunque sean pocos, se desplace en el sentido de la asimetría,

perdiendo representatividad a la vez que aumenta la representatividad de la mediana, pues habrá más valores entorno a ella.

En gran número de estudios económicos se dispone de información de una población similar, de información acerca de una variable relacionada con la variable de interés en la misma población, o incluso de la misma variable estudiada para toda las unidades de la población en una ocasión anterior. En estos casos el incorporar esta información auxiliar en la fase de estimación puede producir estimaciones más precisas del parámetro de interés, tal y como ocurre cuando este parámetro es una media, un total o una proporción. En trabajos como los de Chambers y Dunstan (1986), Rao, Kovar y Mantel (1990), Kuk y Mak (1989,94), Mak y Kuk (1993) se pone de manifiesto el interés que recientemente ha recibido la estimación de la mediana de una población finita en presencia de información auxiliar.

Aunque en ocasiones se requiere dar un valor único como estimación de un parámetro de interés (imaginemos qué pasaría si se proporcionará el IPC como un valor comprendido entre 2.3 y 4.1) es adecuado en muchas otras reducir la incertidumbre acerca del parámetro de interés proporcionando un rango de variación de éste a través de un intervalo de confianza. En este sentido la metodología para la construcción de intervalos de confianza para la mediana poblacional puede verse en Krishnaiah y Rao (1988) y concretamente para muestreo aleatorio simple, en Gross (1980), Sedransk y Meyer (1978) y Smith y Sedransk (1983).

Este trabajo comienza, en su segunda sección, con la exposición del método general de Woodruff (1952) para la construcción de intervalos de confianza para la mediana en muestreo aleatorio simple.

En la tercera sección se presenta un método alternativo para construir intervalos de confianza para la mediana cuando se dispone de información auxiliar.

La última sección ilustra mediante un ejemplo de simulación las propiedades de este nuevo método comparándolas con el método tradicional.

## **2 Intervalo de confianza para la mediana de una población finita.**

Como es usual, sean  $y_1, y_2, \dots, y_N$  los valores de la variable de interés en la población. Para cada número real  $y$  se define  $F_y(y)$  como la proporción de individuos en la población para los cuales  $y_k \leq y$ . La mediana poblacional se define como  $M_y = F_y^{-1}(0.5)$ , donde  $F_y^{-1}$  es la función inversa de  $F_y$ , es decir,  $M_y = \inf \{y_k \mid F_y(y_k) \geq 0.5\}$ .

El procedimiento para estimar la mediana  $M_Y$  usando los datos  $y_k, k \in s$ , de una muestra  $s$ , consiste en obtener un estimador de la función de distribución  $\hat{F}_Y(y)$  y estimar  $M_Y = F_Y^{-1}(0.5)$  por  $\hat{M}_Y = \hat{F}_Y^{-1}(0.5)$ , donde la inversa ahora es en la muestra. Para construir un intervalo de confianza para la mediana se razona ahora de la siguiente forma: para cada dos constantes  $d_1$  y  $d_2$ ,

$$P\{d_1 \leq \hat{F}_Y(M_Y) \leq d_2\} = P\{\hat{F}_Y^{-1}(d_1) \leq M_Y \leq \hat{F}_Y^{-1}(d_2)\},$$

de donde se sigue que si se eligen  $d_1$  y  $d_2$  de forma que  $P\{d_1 \leq \hat{F}(M) \leq d_2\} = 1 - \alpha$ , un intervalo de confianza aproximada  $100(1 - \alpha)\%$  para  $M_Y$  es

$$[\hat{F}_Y^{-1}(d_1), \hat{F}_Y^{-1}(d_2)]$$

Si el tamaño de muestra es suficientemente grande,  $\hat{F}_Y(M_Y)$  es aproximadamente normal con esperanza  $F_Y(M_Y) = 0.5$ , y se pueden elegir

$$d_1 = 0.5 - z_{\frac{\alpha}{2}} \left\{ V(\hat{F}_Y(M_Y)) \right\}^{\frac{1}{2}} \quad y \quad d_2 = 0.5 + z_{\frac{\alpha}{2}} \left\{ V(\hat{F}_Y(M_Y)) \right\}^{\frac{1}{2}},$$

estimando  $V(\hat{F}_Y(M_Y))$  una vez obtenida la muestra mediante,  $\hat{V}(\hat{F}_Y(M_Y))$  con lo que resulta

$$\left[ \hat{F}_Y^{-1} \left( 0.5 - z_{\frac{\alpha}{2}} \left\{ \hat{V}(\hat{F}_Y(M_Y)) \right\}^{\frac{1}{2}} \right), \hat{F}_Y^{-1} \left( 0.5 + z_{\frac{\alpha}{2}} \left\{ \hat{V}(\hat{F}_Y(M_Y)) \right\}^{\frac{1}{2}} \right) \right]$$

En muestreo aleatorio simple  $\hat{F}_Y(M_Y)$  sigue una distribución hipergeométrica con  $E(\hat{F}_Y(M_Y)) = F_Y(M_Y) = 0.5$  y

$$V(\hat{F}_Y(M_Y)) = \frac{N-n}{N-1} \frac{1}{n} F_Y(M_Y)(1 - F_Y(M_Y)) \approx \frac{1-f}{n} 0.25,$$

con lo que la aproximación normal  $P(d_1 \leq \hat{F}_Y(M_Y) \leq d_2) = 1 - \alpha$  para

$$d_1 = 0.5 - z_{\frac{\alpha}{2}} \left( \frac{1-f}{n} 0.25 \right)^{\frac{1}{2}} \quad y \quad d_2 = 0.5 + z_{\frac{\alpha}{2}} \left( \frac{1-f}{n} 0.25 \right)^{\frac{1}{2}},$$

y de esta forma un intervalo aproximado de confianza  $100(1 - \alpha)\%$  para  $M_Y$  es  $[\hat{F}_Y^{-1}(d_1), \hat{F}_Y^{-1}(d_2)]$

### 3 Intervalo de confianza para la mediana de una población finita usando información auxiliar.

Supongamos que se dispone de una variable  $x$  que proporciona información auxiliar, que se observan para las  $n$  unidades de la muestra aleatorio simple de la población de tamaño  $N$ ,  $(x_1, y_1), \dots, (x_n, y_n)$  y que la mediana poblacional de la variable  $x$ ,  $M_X$ , es conocida.

Consideremos el estimador tipo razón

$$\hat{F}_R(M_Y) = \frac{\hat{F}_Y(M_Y)}{\hat{F}_X(M_X)} F_X(M_X)$$

y dos constantes  $c_1$  y  $c_2$  tales que

$$P\{c_1 \leq \hat{F}_R(M_Y) \leq c_2\} = 1 - \alpha.$$

Entonces, llamando  $r_i = c_i \frac{\hat{F}_X(M_X)^{0.5}}{\hat{F}_Y(M_Y)^{0.5}}$ ,  $i = 1, 2$ ,

$$P(\hat{F}_Y^{-1}(r_1) \leq M_Y \leq \hat{F}_Y^{-1}(r_2)) = 1 - \alpha,$$

y así el intervalo

$$[\hat{F}_Y^{-1}(r_1), \hat{F}_Y^{-1}(r_2)]$$

es aproximadamente un intervalo con el  $100(1 - \alpha)\%$  de confianza para  $M_Y$ .

Si se asume que la distribución de  $\hat{F}_R(M_Y)$  es aproximadamente normal con esperanza  $F_Y(M_Y) = 0.5$  (ver Kuk y Mak, 1989) se pueden elegir

$$c_1 = 0.5 - z_{\frac{\alpha}{2}} \left\{ V(\hat{F}_R(M_Y)) \right\}^{\frac{1}{2}} \quad y \quad c_2 = 0.5 + z_{\frac{\alpha}{2}} \left\{ V(\hat{F}_R(M_Y)) \right\}^{\frac{1}{2}}$$

Para calcular  $V(\hat{F}_R(M_Y))$  consideramos las variables

$$e_0 = \frac{\hat{F}_Y(M_Y) - F_Y(M_Y)}{F_Y(M_Y)}, \quad e_1 = \frac{\hat{F}_X(M_X) - F_X(M_X)}{F_X(M_X)}$$

mediante las cuales se tiene

$$\hat{F}_R(M_Y) = F_Y(M_Y) \frac{1 + e_0}{1 + e_1}.$$

Desarrollando en serie de Taylor, elevando al cuadrado y reteniendo los términos hasta el orden dos, se obtiene

$$\left(\hat{F}_R(M_Y) - F_Y(M_Y)\right)^2 \approx F_Y(M_Y)^2(e_0^2 + e_1^2 - 2e_0e_1),$$

y así

$$\begin{aligned} V(\hat{F}_R(M_Y)) &\approx F_Y(M_Y)^2(E(e_0^2) + E(e_1^2) - 2E(e_0e_1)) = \\ &= \frac{1}{4} \left( 4V(\hat{F}_X(M_X)) + 4V(\hat{F}_Y(M_Y)) - 8Cov(\hat{F}_X(M_X), \hat{F}_Y(M_Y)) \right) = \\ &= \frac{1-f}{n} 0.5 - 2Cov(\hat{F}_X(M_X), \hat{F}_Y(M_Y)) \end{aligned}$$

El problema ahora es evaluar  $Cov(\hat{F}_X(M_X), \hat{F}_Y(M_Y))$ . Para ello consideramos la clasificación a dos vías

	$x_K \leq M_X \quad x_K > M_X$		
$y_k \leq M_Y$	$n_{11} \setminus N_{11}$	$n_{12} \setminus N_{12}$	$N_{1\cdot}$
$y_k > M_Y$	$n_{21} \setminus N_{21}$	$n_{22} \setminus N_{22}$	$N_{2\cdot}$
	$N_{\cdot 1}$	$N_{\cdot 2}$	

donde  $n_i$ , denota el número de unidades en la muestra con  $x \leq M_X$  e  $y \leq M_Y$  y de forma similar  $N_{11}$  es el número de unidades en la población con  $x \leq M_X$  e  $y \leq M_Y$ . Entonces,  $(n_{11}, n_{12}, n_{21}, n_{22})$  sigue una distribución hipergeométrica multivariante  $(n_{11}, n_{12}, n_{21}, n_{22}) \approx HG(N, n, N_{11}, N_{12}, N_{21})$ , y como  $n\hat{F}_Y(M_Y) = n_{11} + n_{12}$ ,  $n\hat{F}_X(M_X) = n_{11} + n_{21}$  se obtiene

$$\begin{aligned} Cov(n\hat{F}_Y(M_Y), n\hat{F}_X(M_X)) &= Cov(n_{11} + n_{12}, n_{11} + n_{21}) = \\ &= V(n_{11}) + Cov(n_{11}, n_{12}) + Cov(n_{11}, n_{21}) + Cov(n_{12}, n_{21}). \end{aligned}$$

Operando convenientemente, teniendo en cuenta que

$$Cov(n_i, n_j) = -\frac{N-n}{N-1} n \frac{N_i N_j}{N^2}, \quad V(n_{ij}) = \frac{N-n}{N-1} n \frac{N_{ij}}{N} \left( 1 - \frac{N_{ij}}{N} \right)$$

se tiene

$$Cov(n\hat{F}_Y(M_Y), n\hat{F}_X(M_X)) = \frac{N-n}{N-1} \frac{n}{N^2} (N_{11}N_{22} - N_{12}N_{21})$$

y sustituyendo en (3.1) se obtiene

$$V(\hat{F}_R(M_Y)) \approx \frac{1-f}{n} 0.5 - 2 \frac{1-f}{n} \left( \frac{N_{11}N_{22} - N_{12}N_{21}}{N^2} \right) \quad (3.2)$$

Usando el coeficiente V de Cramer

$$f = \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1.}N_{2.}N_{.1}N_{.2}}}$$

(3.2) puede escribirse como

$$V(\hat{F}_R(M_Y)) \approx \frac{1-f}{2n} (1-f).$$

En la práctica  $\phi$  no es observable pues  $M_Y$  es desconocido luego ha de ser estimado a partir de la muestra. Sustituyendo los  $n_{ij}$  por  $\tilde{n}_{ij}$  basados en la clasificación a dos vías

$$\begin{array}{ccc} x_k \leq M_X & x_k > M_X \\ y_k \leq \hat{M}_Y & \tilde{n}_{11} \setminus \tilde{N}_{11} & \tilde{n}_{12} \setminus \tilde{N}_{12} \\ y_k > \hat{M}_Y & \tilde{n}_{21} \setminus \tilde{N}_{21} & \tilde{n}_{22} \setminus \tilde{N}_{22} \end{array}$$

podemos considerar

$$f = \frac{\tilde{n}_{11}\tilde{n}_{22} - \tilde{n}_{12}\tilde{n}_{21}}{\sqrt{\tilde{n}_{1.}\tilde{n}_{2.}\tilde{n}_{.1}\tilde{n}_{.2}}},$$

como un estimador de  $\phi$  y así, llamando

$$\tilde{r}_i = \frac{\hat{F}_X(M_X)}{0.5} \left( 0.5 + (-1)^i z_{\frac{\alpha}{2}} \left\{ \frac{1-f}{2n} (1-f) \right\}^{\frac{1}{2}} \right)$$

el intervalo  $[\hat{F}_Y^{-1}(\tilde{r}_1), \hat{F}_Y^{-1}(\tilde{r}_2)]$  es aproximadamente un intervalo con una confianza del  $100(1-\alpha)\%$  para  $M_Y$ .

Notar que como  $V(\hat{F}(M_Y)) = \frac{1-f}{n} 0.25$ ,  $V(\hat{F}_R(M_Y)) < V(\hat{F}(M_Y)) \Leftrightarrow f > \frac{1}{2}$ , es decir, si el coeficiente de correlación lineal para los factores codificados es mayor que 1, o lo que es lo mismo, existe asociación entre las clases  $y \leq M_Y$  y  $x \leq M_X$ , la precisión en la estimación de  $\hat{F}(M_Y)$  es mayor mediante el estimador tipo razón  $\hat{F}_R(M_Y)$  que proponemos.

## 4 Estudio de simulación.

En esta sección vamos a ilustrar mediante estudios de simulación las propiedades del método que proponemos para construir intervalos de confianza comparándolas con las propiedades del método presentado en la segunda sección siguiendo el método de Woodfruff.

No es acertado, para comparar la eficiencia de los dos métodos, obtener con cada uno de ellos, una vez seleccionada la muestra, un intervalo numérico y seleccionar aquél que para una misma confianza prefijada tenga una longitud menor. Un criterio más acertado en este caso sería observar la longitud media de los intervalos obtenidos con cada uno de los métodos, para un mismo nivel de confianza y para un mismo tamaño de muestra, en distintas muestras, e incluso observar la varianza de estas longitudes como medida de la representatividad de esta longitud media.

Los primeros datos que se han seleccionado para este estudio de simulación corresponden a una población de 270 manzanas de edificios (Kish, 1965). La variable de interés es el número de viviendas alquiladas y la variable auxiliar el número de viviendas de la manzana.

Para cada tamaño de muestra con valores  $n = 30, 35, 40, 45, 50$  y  $100$ , se han calculado intervalos de confianza al 90%, 95% y 99% para 100 muestras aleatorias simples. En la tabla siguiente se muestran las medias,  $\bar{l}$ , y varianzas,  $s_l^2$ , ahí, de las longitudes de los intervalos de confianza para la mediana según el procedimiento usual  $(\hat{F}_Y^{-1}(d_i))$  y el procedimiento propuesto  $(\hat{F}_Y^{-1}(\tilde{r}_i))$ .

**Tabla 1.**

	100(1 - $\alpha$ )%	90%	95%	99%
N		$\bar{l}$ $s_l^2$	$\bar{l}$ $s_l^2$	$\bar{l}$ $s_l^2$
30	$\hat{F}_Y^{-1}(d_i)$	13.76 21.7625	15.49 30.5499	18.69 30.2341
	$\hat{F}_Y^{-1}(\tilde{r}_i)$	6.57 15.8651	8.47 29.3291	10.24 41.3425
35	$\hat{F}_Y^{-1}(d_i)$	12.32 16.4976	12.25 17.9475	17.94 26.2367
	$\hat{F}_Y^{-1}(\tilde{r}_i)$	5.89 13.3379	7.42 19.7436	10.28 21.0615
40	$\hat{F}_Y^{-1}(d_i)$	9.68 9.8576	11.22 16.0516	16.48 27.5295
	$\hat{F}_Y^{-1}(\tilde{r}_i)$	5.66 8.6044	6.52 12.1496	8.84 25.8343

45	$\hat{F}_Y^{-1}(d_i)$	9.88 12.2656	11.47 13.7491	15.85 17.3077
	$\hat{F}_Y^{-1}(\tilde{r}_i)$	5.27 7.5371	6.19 9.2539	7.96 13.6384
50	$\hat{F}_Y^{-1}(d_i)$	9.45 10.9075	11.83 9.8011	15.18 16.0477
	$\hat{F}_Y^{-1}(\tilde{r}_i)$	5.25 5.3675	5.83 8.5811	7.46 12.2484
100	$\hat{F}_Y^{-1}(d_i)$	5.56 2.0264	6.40 2.3600	8.94 4.6964
	$\hat{F}_Y^{-1}(\tilde{r}_i)$	2.70 1.1500	3.23 1.1571	4.44 1.9864

Como se observa, para todos los niveles de confianza y todos los tamaños muestrales el método que proponemos proporciona intervalos de confianza más pequeños que el método usual. La longitud media es muy inferior en nuestro método y la varianza de las longitudes es también menor. Por ejemplo, al 95% de confianza la longitud media de los cien intervalos de confianza construidos con tamaño de muestra  $n = 50$  es el 49% de la longitud del intervalo construido con el método usual para la misma confianza e igual tamaño de muestra. Estos resultados son debidos a que la ordenación de las unidades en  $x$  y en  $y$  está muy ligada. Por tanto, como cabía esperar, este método es preferible al usual para esta población.

La segunda población estudiada corresponde a una población de 1500 familias de una localidad española, tomadas de Fernández y Mayor (1994). En este caso la variable de interés es el gasto anual en alimentación y la variable auxiliar considerada los ingresos familiares anuales. Los resultados obtenidos se muestran en la segunda tabla y admiten comentarios similares a los realizados con la primera población.

**Tabla 2.**

	100(1 - $\alpha$ )%	90%	95%	99%
N		$\bar{l}$ $s_l^2$	$\bar{l}$ $s_l^2$	$\bar{l}$ $s_l^2$
30	$\hat{F}_Y^{-1}(d_i)$	870.13 62765.8242	1109.76 91428.3906	1322.01 93259.2344
	$\hat{F}_Y^{-1}(\tilde{r}_i)$	784.83 113449.8438	931.85 150795.5000	1230.07 194601.6250
35	$\hat{F}_Y^{-1}(d_i)$	853.35 60534.5273	1052.68 72368.8281	1431.23 59549.2422
	$\hat{F}_Y^{-1}(\tilde{r}_i)$	713.51 80955.8750	819.81 86056.6182	1102.49 127262.9297
40	$\hat{F}_Y^{-1}(d_i)$	852.17 46967.4883	978.66 46949.1680	1321.97 66503.6953
	$\hat{F}_Y^{-1}(\tilde{r}_i)$	678.40 72997.3203	751.05 7958.694	1050.354 104220.7031



45	$\hat{F}_Y^{-1}(d_i)$	693.55	32137.9902	789.21	36645.4844	1090.63	51107.8320
	$\hat{F}_Y^{-1}(\tilde{r}_i)$	594.86	40428.5859	738.56	68460.3359	956.86	68661.9606
50	$\hat{F}_Y^{-1}(d_i)$	637.66	26911.5547	783.06	39171.9492	1010.67	49761.1875
	$\hat{F}_Y^{-1}(\tilde{r}_i)$	576.79	43669.7344	665.39	41271.8477	887.14	62667.5313
100	$\hat{F}_Y^{-1}(d_i)$	449.89	8878.517	544.12	13091.5049	742.56	15620.7754
	$\hat{F}_Y^{-1}(\tilde{r}_i)$	399.96	7290.891	454.14	13624.1592	626.39	23972.7578

## 5 Referencias.

Chambers, R. L. y Dunstan, R. (1986) Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Cochran, A. H. (1977) *Sampling Techniques*. Third Edition. Wiley, New York.  
 Fernández García, F. R. y Mayor Gallego, J. A. (1994). *Muestreo en Poblaciones Finitas: Curso Básico*. P.P.U., Barcelona.

Gross, S. T. (1980) Median estimation in sample survey. *Proc. Surv. Res. Meth. Sect. Amer. Statist. Ass.*, 181-184.

Kish, L. (1965) *Survey Sampling*. John Wiley and Sons, New York.

Krishnalal, P. R. y Rao, C. R., eds. (1988), *Handbook of Statistics*, Vol. 6. Elsevier Science Publisher B. V., 267-289.

Kuk, A. Y. C. y Mak, T. K. (1989) Median estimation in presence of auxiliary information. *J. R. Statist. Soc. B*, 51, 261-269.

Kuk, A. Y. C. y Mak, T. K. (1994) A functional approach to estimating finite population distribution functions. *Commun. Statist. Theory Meth.*, 23(3), 883-896.

Mak, T. K. y Kuk, A. Y. C. (1993) A new method for estimating finite population quantiles using auxiliary information. *The Canadian Journal of Statistics*, 21(1), 29-38.

Mak, T. K. y Kuk, A. Y. C. (1994) A functional approach to estimating finite population distribution functions. *Commun. Statist. Theory Meth.*, 23(3), 883-896.

Sedransk, J. y Meyer, J. (1978) Confidence intervals for the quantiles of a finite population: simple random and stratified simple random sampling. *J. Amer. Statist. Assoc.*, 76, 66-77.

Smith, P. y Sedransk, J. (1983) Lower bounds for confidence coefficients for confidence intervals for finite population quantiles. *Commun. Statist. Theor. Meth.*, 12, 1329-1344.

Rao, J. N. K., Kovar, J. G. y Mantel, H. J. (1990) On estimating distribution functions and quantiles from survey data using auxiliary information. -375. *Biometrika*, 77) 365

Woodruff, R. S. (1952) Confidence intervals for medians and other position measures. *J. Amer. Statist. Assoc.*, 47, 635-646.