Revista Colombiana de Estadística
Nº 7 - 1983

EL PROBLEMA DE LA ESTIMACION DE PARAMETROS CUANDO SE TRABAJA CON MARCOS DE MUESTRA SUPERPUESTOS

David Ospina Botero

Profesor Asociado
Universidad Nacional

SEGUNDA PARTE: POBLACION GAMMA Y BINOMIAL, LA IMPORTANCIA DEL VALOR p.

Resumen: El presente artículo presenta algunas recomendaciones acerca de la conveniencia o no de invertir parte del presupuesto de un proyecto en la determinación de los tamaños de los diferentes dominios cuando se trabaja con poblaciones gamma y binomial. También resalta la importancia del valor p utilizado en el cálculo de estimadores de totales y sus respectivas varianzas. La parte ini-

cial del presente tema fue publicada ya en la Revista Colombiana de Estadística Nº 4 en 1981.

Introducción.

Para aplicar las reglas de decisión referentes al problema expuesto ya por D. Ospina (1981) es necesario establecer algunos supuesto acerca de las medias y varianzas de los dominios y del valor de "p" utilizado en los estimadores.

Primer Caso: Población Gamma.

$$\sigma_a^2 = \sigma_b^2 = \sigma_{ab}^2 = \sigma^2$$

$$\bar{y}_a = \bar{y}_b = \bar{y}_{ab} = \bar{y}$$

$$\bar{y} = \sigma/\sqrt{K}$$

$$n_A = n_B, \quad n_A' = n_B'$$

$$C_A = C_B = C_1$$

$$p = \frac{N_A}{N_A+N_B}, \quad q = \frac{N_B}{N_A+N_B}$$

(La notación es exactamente la misma que se utilizó en la Primera Parte).

Este es un caso muy importante. Cuando $K = 1$ tenemos la *distribución gamma* con densidad dada por

$$g(y) = \frac{1}{\lambda} e^{-y/\beta}, \qquad y > 0, \quad \lambda > 0,$$

donde $\lambda = \bar{Y}$.

Podemos tener también la *Distribución de Poisson* con función de probabilidad dada por:

$$f(y) = \frac{e^{-\mu}\,\mu^{y}}{y!}, \qquad y = 0,1,2,\ldots$$

donde $\mu = \bar{Y}$.

Basándonos en las reglas de decisión, si

$$\frac{C_2}{C} > \frac{U}{V_n + U},$$

no debemos pagar C_2 para conocer los tamaños de los dominios.

Reemplazando los valores originales y usando

$$p = \frac{\beta}{\alpha+\beta} \qquad \text{y} \qquad q = \frac{\alpha}{\alpha+\beta}$$

obtenemos

$$\frac{C}{C_2} < 1 + \frac{(1-\alpha)\beta^2(\alpha+\beta)^2+\alpha\beta^4+\alpha^2(1-\beta)(\alpha+\beta)^2+\alpha^4\beta}{K\alpha^2\beta^2\{\alpha(1-\alpha)+\beta(1-\beta)\}}$$

$0 < \alpha < 1, \quad 0 < \beta < 1, \quad K > 0.$

Haciendo $N_A = DN_B$, entonces $\beta = D\alpha$, (D constante), $p = \frac{D}{D+1}$, $q = \frac{1}{D+1}$ y la ecuación anterior se convierte después de algunas simplificaciones en

$$\frac{C}{C_2} < 1 + \frac{(1+D)^2(1+D^2)-3\alpha D^2(1+D)}{K\alpha D^2\{1-\alpha+D-\alpha D^2\}} \tag{1}$$

$0 < \alpha < \frac{1}{D}, \quad K > 0, \quad D > 0.$

El lado derecho de la desigualdad anterior puede considerarse como una función de α si D y K se toman como constantes. Esta función la denominaremos $f(\alpha)$.

De acuerdo con lo anterior, siempre que $\frac{C}{C_2} < f(\alpha)$ no vale la pena tratar de determinar los tamaños de los dominios. El mínimo de $f(\alpha)$ puede usarse para establecer reglas de decisión generales para la escogencia de determinar o no los tamaños de los dominios antes de muestrear de cada marco. Para hacer esto tomamos la derivada de $f(\alpha)$ con respecto a α, igualamos a cero y resolvemos la ecuación resultante.

Puede mostrarse que las raíces de la ecuación resultante son:

$$\alpha_1 = \frac{(1+D)}{3D^2}\left\{1+D^2+\sqrt{1-D^2+D^4}\right\}$$

y

$$\alpha_2 = \frac{(1+D)}{3D^2}\left\{1+D^2-\sqrt{1-D^2+D^4}\right\}$$

También puede comprobarse que α_2 da el mínimo de la función que en este caso sería:

$$f(\alpha_2) = 1 + \frac{9\,D^2\ \sqrt{1-D^2+D^4}}{[1+D^2-\sqrt{1-D^2+D^4}]\ [3D^2-(1+D^2-\sqrt{1-D^2+D^4})(1+D^2)]} \tag{2}$$

para valores fijos de K y D.

Debemos anotar que el máximo y el mínimo para $f(\alpha)$ siempre existen ya que $D > 0$. Esto puede verse fácilmente.

1. Si $0 \leqslant D \leqslant 1$, $D^2 \leqslant D \leqslant 1$, $\Rightarrow$, $1 - D^2 + D^4 \geqslant D^4 \geqslant 0$
2. Si $1 < D < \infty$, $1 < D^2 < D^4$, $\Rightarrow$, $1 - D^2 + D^4 \geqslant 1 \geqslant 0$

Además, siendo

$$\beta = D\alpha \quad \text{y} \quad 0 \leqslant \beta \leqslant 1, \ \Rightarrow, \ 0 \leqslant \alpha \leqslant \frac{1}{D} \tag{3}$$

Como en la mayoría de los casos no conocemos el valor de α, no podemos utilizar directamente (1). Por otra razón tenemos que utilizar (2) y (3) como condiciones limitantes. Si puede asumirse un rango de valores posibles para α, podemos hallar el número de la función para ese intervalo. En caso contrario (3) nos dará un valor permisible máximo general para C_2/C.

<u>Segundo Caso</u>: Población Binomial.

Definimos:

$\bar{Y}_a$ = Proporción del dominio a con una característica específica.

$\bar{y}_b$ = Proporción del dominio b con una característica específica.

$\bar{y}_{ab}$ = Proporción del dominio ab con una característica específica.

De acuerdo con estas definiciones:

$$\sigma_a^2 = \bar{y}_a(1 - \bar{y}_a)$$
$$\sigma_b^2 = \bar{y}_b(1 - \bar{y}_b)$$
$$\sigma_{ab}^2 = \bar{y}_{ab}(1 - \bar{y}_{ab})$$

ya que

$$y_i = \begin{cases} 1 \text{ si la característica específica esta presente en la } i\text{-ésima unidad.} \\ 0 \text{ en otro caso.} \end{cases}$$

Haciendo las sustituciones correspondientes en la regla de decisión, la condición para no determinar los tamaños de los dominios con parte del presupuesto, C_2 es:

$$\frac{C}{C_2} < 1 + \frac{A + B\alpha}{E\alpha - G\alpha^2}$$

donde $D = \frac{\beta}{\alpha}$

$$A = D^2(1+D^2)\bar{y}_a(1-\bar{y}_a)+(1+D)^2\bar{y}_b(1-\bar{y}_b)$$

$$B = D[(D^3+1)\bar{y}_{ab}(1-\bar{y}_{ab})-D(1+D)^2\bar{y}_a(1-\bar{y}_a)-(1+D)^2\bar{y}_b(1-\bar{y}_b)]$$

$$E = D\{D[(1+D)\bar{y}_a-D\bar{y}_{ab}]^2+[(1+D)\bar{y}_b-\bar{y}_{ab}]^2\}$$

$$G = D^2\{[(1+D)\bar{y}_a-D\bar{y}_{ab}]^2+[(1+D)\bar{y}_b-\bar{y}_{ab}]^2\}.$$

Hagamos ahora:

$$\delta(\alpha) = 1 + \frac{A+B\alpha}{E\alpha - G\alpha^2}, \qquad A,B,E,G \text{ constantes}, \qquad (4)$$

$$0 \leqslant \alpha \leqslant 1.$$

Para analizar esta función debemos hallar los puntos de inflexión por el mismo procedimiento usado en el primer caso. Tomando la derivada con respecto a α, tenemos:

$$\delta'(\alpha) = \frac{-A(E-2G\alpha)}{\alpha^2(E-G\alpha)^2} + \frac{BG}{(E-G\alpha)^2}$$

Igualando a cero tendremos las soluciones:

$$\alpha_1 = \frac{-A}{B} + \frac{\sqrt{A^2G^2+ABEG}}{BG} \qquad \text{(Máximo)}$$

y

$$\alpha_2 = \frac{-A}{B} - \frac{\sqrt{A^2G^2+ABEG}}{BG} \qquad \text{(Mínimo)}.$$

Para obtener valores reales para α_1 y α_2 necesitamos $A^2G^2 + ABEG \geqslant 0$.

Puesto que A, E y G son siempre no-negativos, necesitamos justamente,

$$AG + BE \geqslant 0.$$

Ahora, para $D \geqslant 1$, $G \geqslant E$, solamente necesitamos

$$A + B \geqslant 0.$$

Usando las definiciones de A y B, debemos tener:

$$D^2(1+D)^2\bar{Y}_a(1-\bar{Y}_a)+(1+D)^2\bar{Y}_b(1-\bar{Y}_b)+D(D^3+1)\bar{Y}_{ab}(1-\bar{Y}_{ab})$$

$$-\ D^2(1+D)^2\bar{Y}_a(1-\bar{Y}_a)-D(1+D)^2\bar{Y}_b(1-\bar{Y}_b) \geqslant 0 \quad ,$$

ó

$$(1-D)(1+D)^2\bar{Y}_b(1-\bar{Y}_b)+D(D^3+1)\bar{Y}_{ab}(1-\bar{Y}_{ab}) \geqslant 0 \ ,$$

y finalmente

$$\frac{\bar{Y}_{ab}(1-\bar{Y}_{ab})}{\bar{Y}_b(1-\bar{Y}_b)} \geqslant \frac{(D-1)(1+D)^2}{D(D^3+1)}$$

Para $D = 1$, tamaños iguales de los marcos ($N_A = N_B$)

$$\frac{\bar{Y}_{ab}(1-\bar{Y}_{ab})}{\bar{Y}_b(1-\bar{Y}_b)} \geqslant 0$$

lo cual siempre ocurre puesto que los $\bar{Y}_s'$ son proporcionales.

Para $D > 1$, ($N_A > N_B$):

Hagamos

$$F = \frac{(D-1)(1+D)^2}{D(D^3+1)}$$

Entonces:

$$\frac{\bar{Y}_{ab}(1-Y_{ab})}{\bar{Y}_b(1-\bar{Y}_b)} \geqslant F \ . \qquad (5)$$

Haciendo $\bar{Y}_b$ constante (entre 0 y 1)

$$\bar{y}_{ab}(1-\bar{y}_{ab}) \geq CF$$

$$-\bar{y}^2_{ab} + \bar{y}_{ab} - CF \geq 0$$

Esta desigualdad se sostiene siempre que:

$$\frac{1}{2} - \frac{\sqrt{1-4CF}}{2} \leq \bar{y}_{ab} \leq \frac{1}{2} + \frac{\sqrt{1-4CF}}{2}$$

Lo anterior siempre producirá una raíz real para $\delta(\alpha)$. Debemos tener en cuenta nuevamente que $\mathcal{D} \geq 1$. Podemos hallar el valor máximo de F el cual es aproximadamente igual a 0.51 para $\mathcal{D} = 1.787$.

Reemplazando en (5)

$$\frac{\bar{y}_{ab}(1-\bar{y}_{ab})}{\bar{y}_b(1-\bar{y}_b)} \geq 0.51$$

ó

$$\bar{y}_{ab}(1-\bar{y}_{ab}) \geq 0.51\ \bar{y}_b(1-\bar{y}_b).$$

Pero sabemos que el valor máximo para $\bar{y}_b(1-\bar{y}_b)$ es 0.25. Entonces:

$$\bar{y}_{ab}(1-\bar{y}_{ab}) \geq 0.1275$$

que tiene la solución

$$0.15 \leq \bar{y}_{ab} \leq 0.85.$$

Como conclusión podemos decir que la función (4) tendría un máximo y un mínimo siempre que $\bar{y}_{ab}$ se encuentre entre 0.15 y 0.85. Intervalos como el anterior para $\bar{Y}_{ab}$ son funciones de D.

El Valor " p "

El valor de "p" utilizado en $\dot{Y}$ y $\hat{Y}$ (ver Rev. Col. de Est. N° 4, pp.19-22) influenciará no sólo las estimaciones de los totales sino sus varianzas. De todas maneras el valor de p que minimiza la varianza será una función de los parámetros de la situación. Sin embargo, en la práctica, el p óptimo no está disponible debido a la naturaleza de la información necesaria para calcularlo. En la discusión anterior se escogió p como $\frac{N_A}{N_A+N_B}$ para todas las consideraciones. En esta sección se estudiarán varios valores "convenientes" para p. La influencia de estos valores en la regla de decisión serán de especial interés. La siguiente tabla muestra los resultados correspondientes a los diferentes valores asignados a p (incluído $\frac{N_A}{N_A+N_B}$).

Tabla 1

$p = 0$	$\alpha_2 = \dfrac{(D^2+1) - \sqrt{D^2+1}}{D^2}$
	$f(\alpha) = 1 + \dfrac{(D^2+1) - \alpha D^2}{KD^2(1-\alpha)}$

$$p = \frac{N_B}{N_A+N_B} = \frac{1}{D+1} \qquad \alpha_2 = \frac{(1+D^3)}{D(1+D+D^2)}$$

$$f(\alpha) = 1 + \frac{(1+D^2)(1+D)^2-\alpha D(1+D)(1+D+D^2)}{KD\alpha\left[1+D^3-\alpha D(1+D^2)\right]}$$

$$p = \frac{1}{2} \qquad \alpha_2 = \frac{4(D^2+1)-\sqrt{10D^4-12D^3+20D^2-12D+10}}{3D(D+1)}$$

$$f(\alpha) = 1 + \frac{4(D^2+1)-3\alpha D(D+1)}{KD\alpha(D+1-2D\alpha)}$$

$$p = \frac{N_A}{N_A+N_B} = \frac{D}{D+1} \qquad \alpha_2 = \frac{(1+D)}{3D^2}\left(1+D^2-\sqrt{1-D^2+D^4}\right)$$

$$f(\alpha) = 1 + \frac{(1+D)^2(1+D^2)-3\alpha D^2(1+D)}{K\alpha D^2(1-\alpha+D-\alpha D^2)}$$

$$p = 1 \qquad \alpha_2 = \frac{(D^2+1)-D\sqrt{D^2+1}}{D}$$

$$f(\alpha) = 1 + \frac{(D^2+1)-D\alpha}{KD\alpha(1-D\alpha)}$$

El objetivo de esta sección es encontrar, cuando sea posible, el p óptimo para algunos de los casos ya considerados. Esto nos permitirá hacer algunas afirmaciones generales relacionadas con la función de decisión. El método utilizado para hacer esto, dado por R. Cochran (1965), es:

1. Expresar la función de decisión como una función de p, digamos $g(p)$.
2. Seleccionar 3 valores "convenientes" de p (0.00, 0.50 y 1.00 generalmente sirven).
3. Ajustar una función cuadrática en p de la forma:

$$g(p) = ap^2 + bp + c$$

4. Hallar el valor de p que minimice la función anterior. Este valor puede considerarse como el óptimo aproximado para p. Si en el paso 2 tomamos p como .00, .50 y 1.00, el p óptimo estará dado por:

$$p = \frac{1}{2} + \frac{g(1.00) - g(0.00)}{4\left[g(1.00)-2g(0.50)+g(0.00)\right]}$$

Este valor minimizará la diferencia entre las varianzas de los dos estimadores ($\hat{Y}$ y $\dot{Y}$) pero no la varianza de cada uno de ellos separadamente. Sin embargo, el p óptimo para la varianza de $\hat{Y}$ fué aproximado para analizar la consistencia de los dos valores. El valor óptimo de p en el último caso está dado por Hartley (1962) como una de las soluciones de la expresión bicuadrática:

$$\frac{C_A\ p^2}{C_B\ q^2} = \frac{\sigma_a^2(1-\alpha)+\alpha p^2\sigma_{ab}^2}{\sigma_b^2(1-\beta)+\beta q^2\sigma_{ab}^2}$$

Debido a las condiciones laterales de que p y q deban ser positivos y tales que $p+q = 1$, debe escogerse la raíz de la expresión anterior cuyo valor esté entre cero y uno.

Los casos considerados fueron:

1. Para $D = 1$ ($N_A = N_B$) el p óptimo es 0.50 tanto para $g(p)$ como para $V(\hat{Y})$.

2. Para $D = 2$ ($N_A = 2N_B$):

   $K = \frac{1}{2}$, p óptimo para $g(p) = 0.8496$

   $K = 1$, p óptimo para $g(p) = 0.8636$.

   $K = 4$, p óptimo para $g(p) = 0.9444$

   El p óptimo para $V[\hat{Y}]$ depende del valor de $= \frac{N_{ab}}{N_A}$.

   Algunos p_s' óptimos son:

Para: $\alpha = 0.05$, p óptimo $= 0.505$

$\alpha = 0.10$, p óptimo $= 0.510$

$\alpha = 0.45$, p óptimo $= 0.654$

$\alpha = 0.48$. p óptimo $= 0.720$

(*) $\alpha = 0.50$, p óptimo $= 1$

(*) Para $D = 2$ el valor mayor de α es 0.50. Esto ocurre cuando el marco B está incluído totalmente dentro del marco A.

Como podemos ver, los valores óptimos de p difieren de alguna manera para $g(p)$ y $V[\hat{Y}]$. Sin embargo el p óptimo para $g(p)$ siempre está dentro del intervalo de los p_s' óptimos para $V[\hat{Y}]$. Estos resultados pueden considerarse consistentes. Tomando p como criterio para la función de decisión pueden establecerse algunas reglas generales para decidir la proporción máxima del presupuesto que debe utilizarse para determinar los tamaños de los dominios (ver Tabla 2). El p óptimo aumenta con el valor de D y con el valor de K. Sin embargo, la proporción permisible máxima del presupuesto para hallar los tamaños de los dominios disminuye con un incremento de D.

Tabla 2

$D = \frac{N_A}{N_B}$	$K = \frac{1}{\text{Coef.Var.}(\bar{Y})^2}$	Porcentaje máximo de C a pagar
1 (p = 0.5000)	0.5	5.26
	1	10.00
	4	30.77
2 (p = 0.8496)	0.5	2.20(*)
(p = 0.8636)	1	4.50(*)
(p = 0.9444)	4	18.00(*)

(*) Valores calculados en base a las gráficas de las respectivas funciones.

Otras situaciones pueden considerarse para valores diferentes de K y D. El método de aplicación es el mismo. También sería de gran interés analizar el caso Binomial, discutido anteriormente, y algunos otros casos donde los supuestos iniciales puedan modificarse.

* *

BIBLIOGRAFIA

Cochran, R.S., "Theory and Application of Multiple Frame Surveys" Unpublished Ph. D. dissertation, Dep. of Statistics, Iowa State University, Ames, Iowa, 1965.

Hartley, H.O., "Multiple Frame Surveys" Proceeding of Social Science Section of American Statistical Association. Mectings, Minneapolis, Minnesota, 1962.

Ospina, D., "El problema de la estimación de parámetros cuando se trabaja con marcos de muestra superpuestos", Revista Col. de Estadística Nº 4, 1981.

*

N.del C.E. Este artículo es una continuación de "El Problema de la Estimación de Parámetros cuando se trabaja con Marcos de Muestra Superpuestos" (Ospina 1981) cuya primera parte es una adaptación del artículo de Hartley (1962) hecha con el fin de facilitar a los lectores la comprensión del tema, dada la dificultad de adquisición de dicha bibliografía.

* * *