Revista Colombiana de Estadística
Volumen 28 N^{o} 2. pp. 141 a 154. Diciembre 2005

Metodología para la estimación de dico–razones con el uso de información auxiliar en tablas de contingencia 3×3

Dico–ratio Estimation with Auxiliary Information in 3×3 Contingency Tables

Jairo A. Fúquene P.*

Resumen

Se comparan los diseños M.A.S, P.P.T., ESTMAS y estratificado por el método de Hidiroglou para estimar una razón de totales de variables dicotómicas. En el diseño P.P.T. se muestran las probabilidades de selección que hacen mínima la varianza aproximada. Bajo el diseño ESTMAS, en el caso de asignación proporcional, se compara la eficiencia de utilizar la misma información auxiliar que en el diseño P.P.T. Mediante simulación, se comparan todos los diseños y se obtiene que para un grado de correspondencia medio o alto entre la variable auxiliar y la variable de interés, los estimadores para los diseños ESTMAS y P.P.T. son los más eficientes.

Palabras Claves: Información auxiliar categórica, estimación de una razón de totales de variables dicotómicas, dico–razón, diseño P.P.T, diseño ESTMAS

Abstract

The S.I, P.P.S, STSI and stratified by the Hidiroglou's method to estimate a totals ratio of dichotomic variables are compared. In the P.P.S. design the selection probabilities that make minimum the approximate variance are showed. In the STSI design, in the case of proportional assignment of sample, the efficiency to use the same auxiliary information that in the P.P.S. design one is compared. By simulation, all the designs are compared and is obtained that for a medium or high grade of correspondence between the auxiliary variable and the variable of interest, the estimators for the STSI and P.P.S. designs are more efficient.

Keywords: Auxiliary categorical information, estimation of a totals ratio of dichotomic variables, dicho–ratio, P.P.S. design, STSI design

*Estadístico de la Universidad Nacional de Colombia. Grupo de Investigación en Muestreo. Departamento de Estadística, Sede Bogotá. E-mail: jafuquenep@unal.edu.co

1. Introducción

En muchos estudios de tipo muestral se tiene interés en estimar razones de totales de variables dicotómicas, denominadas también dico–razones. Ejemplo típico es la cifra de desempleo, obtenida como el cociente entre el total de personas que buscan empleo y el tamaño de la población económicamente activa en la región. También es el caso de los resultados arrojados por las encuestas electorales, donde se estima la razón entre el total de quienes apoyan a un determinado candidato sobre la cantidad de quienes votarán en el comicio electoral. Desde el punto de vista teórico este problema no deja de ser una aplicación más de la estimación de razones. Sin embargo en la literatura tanto clásica (Cochran 1963), como la más reciente (Särndal, Swensson & Wretman 1992), no se encuentran indicaciones para tratar el caso de estimar dico–razones con el uso de información auxiliar categórica; como puede ser, en el segundo ejemplo, la situación de empleo o desempleo reportada por la persona el mes anterior.

En particular interesa comparar la eficiencia que se obtiene con el uso de información auxiliar categórica en el estimador de una dico–razón en diseños como P.P.T. o estratificado con muestreo aleatorio simple en cada estrato (ESTMAS). Fúquene & Bautista (2005) propusieron una metodología para estimar por medio de información auxiliar categórica una dico–razón bajo el diseño P.P.T. Es conveniente conocer las propiedades de este estimador y compararlas con las del estimador de Horvitz y Thompson para los muestreos mencionados.

En este trabajo se establece la opción más viable para la estimación de una dico–razón en presencia de información auxiliar categórica para los diseños I.F.–ESTMAS, M.A.S, P.P.T. y ESTMAS y se estudia la precisión de este estimador. En la sección siguiente se muestra la variable auxiliar que se utiliza para hallar las probabilidades del diseño P.P.T. que hacen mínima su varianza. En la sección tres, se estudia la metodología de estratificar utilizando la misma información auxiliar que en el diseño P.P.T. En la cuarta sección se comparan, mediante simulación, las varianzas de los diseños estudiados y en la última sección se presentan algunas conclusiones de tipo práctico.

2. Valores de p_k en diseños P.P.T. que minimizan la varianza del estimador de la dico–razón

Sean $U_y \subseteq U_z \subseteq U$ y $U_{y^c} \subseteq U_z$ y las variables dicotómicas que definen estos subconjuntos, y y z. Gráficamente la situación es la siguiente:

Esta situación se presenta, por ejemplo, en investigación de mercados cuando se desea estudiar la preferencia por una determinada marca. Para este caso:

i. U es el universo de personas.

ii. U_z es el subconjunto de personas que consumen un producto.

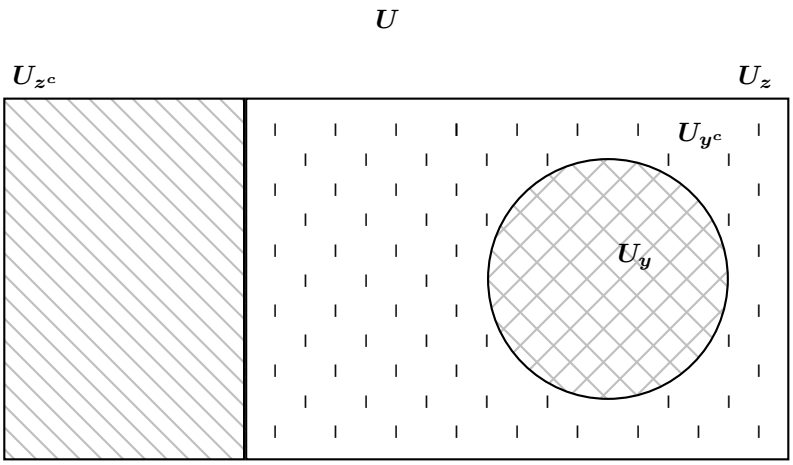

Figura 1: Representación del espacio poblacional en la estimación de una dico–razón.

iii. U_y es el subconjunto de los que consumen una marca dentro de los consumidores del producto.

iv. U_{y^c} es el subconjunto de los que no consumen una marca dentro de los consumidores del producto.

v. U_{z^c} es el subconjunto de personas que no consumen un producto.

El objetivo es estimar la dico–razón $R = \dfrac{N_y}{N_z}$, por ejemplo, la proporción de personas que consumen la marca dentro de los consumidores del producto. La varianza aproximada y el estimador de la varianza de este parámetro se obtienen por medio del método de linealización de Taylor (Särndal et al. 1992) que implica la determinación de la transformada $u_k = \dfrac{1}{N_z}(y_k - Rz_k)$ y en este caso asume los siguientes valores:

$$u_k = \begin{cases} \dfrac{N_{y^c z}}{N_z^2} & \text{si } k \,\epsilon\, U_y \cap U_z \\ -\dfrac{N_{yz}}{N_z^2} & \text{si } k \,\epsilon\, U_{y^c} \cap U_z \\ 0 & \text{si } k \,\epsilon\, U_{z^c} \end{cases} \tag{1}$$

donde $N_{y^c z}$ es la cantidad de elementos que poseen la característica en z y no la poseen en y y, N_z es la cantidad de elementos que poseen la característica en z.

Sean x y w dos variables categóricas auxiliares donde $U_x \subseteq U_w$ y el parámetro $R = \dfrac{N_x}{N_w}$. Sea $u_k^* = \dfrac{1}{N_w}(x_k - Rw_k)$, una variable auxiliar altamente correlacionada con u_k disponible para $k = 1, 2, \ldots, N$. u_k^* asume los valores:

$$u_k^* = \begin{cases} \dfrac{N_{x^c w}}{N_w^2} & \text{si } k \in U_x \cap U_w \\ -\dfrac{N_{xw}}{N_w^2} & \text{si } k \in U_{x^c} \cap U_w \\ 0 & \text{si } k \in U_{w^c} \end{cases} \tag{2}$$

donde, por ejemplo, $N_{x^c w}$ es la cantidad de elementos que poseen la característica en w y no la poseen en x y, N_w es la cantidad de elementos que poseen la característica en w.

Ejemplo 2.1. *Supóngase que en el año 2006 una compañía desea estimar la proporción de colegios que cuentan con servicio de internet dentro de los que tienen sala de cómputo y que para ello cuenta con un censo realizado en el año 2004, en donde se tiene información de N colegios, de los cuales, N_w tenían sala de computo y de ellos, N_{xw} contaban con servicio de internet. Por motivos de costos la compañía decide realizar un muestreo estadístico para conseguir las estimaciones. El parámetro que se quiere estimar es la dico–razón $R = \dfrac{N_y}{N_z}$.*

Las variables auxiliares son el resultado del censo realizado en el año 2004, en el que $N_{x^c w}$ es la cantidad de colegios que no contaban con servicio de internet dentro de los que tenían sala de cómputo y N_{w^c} la cantidad de colegios que no tenían sala de cómputo. Para este caso, el elemento k es el colegio y las variables auxiliares se definen como:

$$w_k = \begin{cases} 1 & \text{si } k \text{ tenía sala de cómputo en el año 2004} \\ 0 & \text{en otro caso} \end{cases}$$

$$x_k = \begin{cases} 1 & \text{si } w_k = 1 \text{ y } k \text{ contaba con servicio de internet} \\ 0 & \text{si } k \text{ no contaba con servicio de internet} \\ 0 & \text{si } w_k = 0 \end{cases}$$

Las variables de estudio son:

$$z_k = \begin{cases} 1 & \text{si } k \text{ tiene sala de cómputo en la actualidad} \\ 0 & \text{en otro caso} \end{cases}$$

$$y_k = \begin{cases} 1 & \text{si } z_k = 1 \text{ y } k \text{ cuenta con servicio de internet} \\ 0 & \text{si } k \text{ no cuenta con servicio de internet} \\ 0 & \text{si } z_k = 0 \end{cases}$$

Por otra parte, N_{xywz} denota la cantidad de colegios que, desde el año 2004 hasta hoy, han contado con servicio de internet y m_{xywz} es la cantidad de colegios que tienen servicio de internet en la muestra y contaban con dicho servicio en el año 2004.

A continuación se muestran las probabilidades denotadas como α_0, β_0 y μ_0 que hacen mínima la varianza aproximada del estimador de la dico–razón bajo el diseño P.P.T. (Fúquene & Bautista 2005).

Resultado 2.1. Para la dico–razón $R = \dfrac{N_y}{N_z}$, la varianza aproximada bajo el diseño P.P.T. es:

$$AV_{PPT}(\hat{R}) = \frac{1}{m(N-N_{z^c})^4}\left[\frac{N_{xywz}(N_{y^cz})^2 + N_{xy^cwz}(N_{yz})^2}{\alpha_0} + \frac{N_{x^cywz}(N_{y^cz})^2}{\beta_0} + \frac{N_{x^cy^cwz}(N_{yz})^2}{\beta_0} + \frac{N_{yw^cz}(N_{y^cz})^2 + N_{y^cw^cz}(N_{yz})^2}{\mu_0}\right] \quad (3)$$

$$\alpha_0 = (A)\beta_0$$
$$\mu_0 = (B)\beta_0$$
$$\beta_0 = \frac{1}{(N_{xw}(A) + N_{w^c}(B) + N_{x^cw})} \quad (4)$$

Para el resultado anterior A y B se pueden escribir de la siguiente manera:

$$A = \sqrt{\frac{P_{xywz} + P_{xy^cwz}(P_{y^cy})^2}{P_{x^cywz} + P_{x^cy^cwz}(P_{y^cy})^2}} \qquad B = \sqrt{\frac{P_{yw^cz} + P_{y^cw^cz}(P_{y^cy})^2}{P_{x^cywz} + P_{x^cy^cwz}(P_{y^cy})^2}} \quad (5)$$

$$P_{y^cy} = \frac{P_{xy^cwz}(N_{xw}) + P_{x^cy^cwz}(N_{x^cw}) + P_{y^cw^cz}(N_{w^c})}{P_{xywz}(N_{xw}) + P_{x^cywz}(N_{x^cw}) + P_{yw^cz}(N_{w^c})} \quad (6)$$

En lo que sigue se utilizarán los parámetros poblacionales para establecer α_0, β_0 y μ_0. En una aplicación real, el usuario deberá utilizar aproximaciones a partir de alguna fuente de información diferente o en su defecto, estimar a partir de un estudio piloto las proporciones de la tabla siguiente:

Tabla 1: Proporciones para establecer α_0, β_0 y μ_0.

		$\boldsymbol{u_k^*}$		
	Conjunto	$\boldsymbol{U_x \cap U_w}$	$\boldsymbol{U_{x^c} \cap U_w}$	$\boldsymbol{U_{w^c}}$
$\boldsymbol{u_k}$	$\boldsymbol{U_y \cap U_z}$	$P_{xywz} = \dfrac{N_{xywz}}{N_{xw}}$	$P_{x^cywz} = \dfrac{N_{x^cywz}}{N_{x^cw}}$	$P_{yw^cz} = \dfrac{N_{yw^cz}}{N_{w^c}}$
	$\boldsymbol{U_{y^c} \cap U_z}$	$P_{xy^cwz} = \dfrac{N_{xy^cwz}}{N_{xw}}$	$P_{x^cy^cwz} = \dfrac{N_{x^cy^cwz}}{N_{x^cw}}$	$P_{y^cw^cz} = \dfrac{N_{y^cw^cz}}{N_{w^c}}$

Estas proporciones se interpretan en el caso del ejemplo 2.1 de la siguiente manera:

i. $P_{xywz} = \dfrac{N_{xywz}}{N_{xw}}$: proporción de colegios que desde el año 2004 hasta hoy han contado con servicio de internet.

ii. $P_{xy^cwz} = \dfrac{N_{xy^cwz}}{N_{xw}}$: proporción de colegios que en el 2004 contaban con servicio de internet y en la actualidad no.

iii. $P_{x^cywz} = \dfrac{N_{x^cywz}}{N_{x^cw}}$: proporción de colegios que en el 2004 no contaban con servicio de internet y en la actualidad si.

iv. $P_{x^cy^cwz} = \dfrac{N_{x^cy^cwz}}{N_{x^cw}}$: proporción de colegios que ni en el 2004 ni en la actualidad han contado con servicio de internet.

v. $P_{yw^cz} = \dfrac{N_{yw^cz}}{N_{w^c}}$: proporción de colegios que en el 2004 no tenían sala de cómputo y en la actualidad tienen y cuentan con servicio de internet.

vi. $P_{y^cw^cz} = \dfrac{N_{y^cw^cz}}{N_{w^c}}$: proporción de colegios que en el 2004 no tenían sala de cómputo y en la actualidad tienen y no cuentan con servicio de internet.

3. Estimación de una dico–razón en diseños ESTMAS

La resta de las varianzas para los diseños M.A.S. y ESTMAS para el caso de asignación proporcional de muestra, $Nn_h = nN_h$, es:

$$AV_{MAS}(\hat{R}) - AV_{ESTMAS}(\hat{R}) = \frac{N^2}{n}\Big(1 - \frac{n}{N}\Big)S^2_{uU} - N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{H} W_h S^2_{uUh} \quad (7)$$

donde $W_h = \dfrac{N_h}{N}$. Para analizar la expresión anterior se descompone la suma de cuadrados total de la variable u_k que se da en (1) en dos sumas: la suma de cuadrados entre los estratos y la suma de cuadrados dentro de los estratos, así:

$$(N-1)S^2_{uU} = \sum_{U}(u_k - \bar{u}_U)^2 \quad (8)$$

$$= \sum_{h=1}^{H} N_h \bar{u}^2_{Uh} + \sum_{h=1}^{H}(N_h - 1)S^2_{uUh} \quad (9)$$

$$SCT = SCE + SCD \quad (10)$$

reemplazando en (7),

$$AV_{MAS}(\hat{R}) - AV_{ESTMAS}(\hat{R}) = N^3\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}\left[\sum_{h=1}^{H} W_h \bar{u}^2_{Uh} - \frac{1}{N}\sum_{h=1}^{H}(1 - W_h)S^2_{uUh}\right] \quad (11)$$

y como las variables y y z son dicotómicas,

$$AV_{MAS}(\hat{R}) - AV_{ESTMAS}(\hat{R}) = \\ N^3\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{N-1}\left[\sum_{h=1}^{H} W_h(P_{zh}(R_h-R)^2) - \frac{1}{N}\sum_{h=1}^{H}(1-W_h)S_{uUh}^2\right] \quad (12)$$

donde $R_h = \dfrac{N_{yh}}{N_{zh}}$ es la dico–razón en el h–ésimo estrato.

Nota 3.1. *Si la dico–razón R_h es igual o aproximadamente igual en todos los estratos, (12) muestra que es posible que la varianza para el diseño ESTMAS sea más grande que la del diseño M.A.S. Por otra parte, si la dispersión de las dico–razones R_h es mayor que la dispersión de la variable u_k en cada uno de los estratos, el diseño ESTMAS resulta más eficiente que el diseño M.A.S.*

Como consecuencia, la mayor ganancia en un diseño ESTMAS se obtiene por medio de una variable auxiliar altamente correlacionada con la variable u_k. En lo que sigue, se estudia el uso de información auxiliar de la forma u_k^* para clasificar en tres estratos, dados por los tres conjuntos donde se puede definir esta variable. En cada estrato se aplica un diseño M.A.S. con asignación proporcional.

Resultado 3.1. Cuando se utiliza la variable u_k^* como información auxiliar, la varianza aproximada bajo el diseño ESTMAS para el estimador de una dico–razón, $R = \dfrac{N_y}{N_z}$, es:

$$AV_{ESTMAS}(\hat{R}) = \frac{1}{N_z^2}\sum_{h=1}^{3} F_h K_h \quad (13)$$

Para el caso de asignación proporcional, F_h y K_h están dados por:

$$F_h = \begin{cases} \dfrac{N}{n}\left(1-\dfrac{1}{N}\right)\dfrac{N_{xw}}{N_{xw}-1} & \text{si } h=1 \\ \dfrac{N}{n}\left(1-\dfrac{1}{N}\right)\dfrac{N_{x^cw}}{N_{x^cw}-1} & \text{si } h=2 \\ \dfrac{N}{n}\left(1-\dfrac{1}{N}\right)\dfrac{N_{w^c}}{N_{w^c}-1} & \text{si } h=3 \end{cases}$$

$$K_h = \begin{cases} N_{xywz}Q_{xywz} - 2(R)N_{xywz}Q_{xwz} + R^2N_{xwz}Q_{xwz} & \text{si } h=1 \\ N_{x^cywz}Q_{x^cywz} - 2(R)N_{x^cywz}Q_{x^cwz} + R^2N_{x^cwz}Q_{x^cwz} & \text{si } h=2 \\ N_{yw^cz}Q_{yw^cz} - 2(R)N_{yw^cz}Q_{w^cz} + R^2N_{w^cz}Q_{w^cz} & \text{si } h=3 \end{cases}$$

con

$$Q_{xywz} = 1 - P_{xywz}; \quad P_{xywz} = \frac{N_{xywz}}{N_{xw}}; \quad Q_{xwz} = 1 - P_{xwz}; \quad P_{xwz} = \frac{N_{xwz}}{N_{xw}}$$
$$Q_{x^cywz} = 1 - P_{x^cywz}; \quad P_{x^cywz} = \frac{N_{x^cywz}}{N_{x^cw}}; \quad Q_{x^cwz} = 1 - P_{x^cwz}; \quad P_{x^cwz} = \frac{N_{x^cwz}}{N_{x^cw}}$$
$$Q_{yw^cz} = 1 - P_{yw^cz}; \quad P_{yw^cz} = \frac{N_{yw^cz}}{N_{w^c}}; \quad Q_{w^cz} = 1 - P_{w^cz}; \quad P_{w^cz} = \frac{N_{w^cz}}{N_{w^c}}$$

Nota 3.2. *Cuando la variable u_k^* discrimina perfectamente los conjuntos de la variable u_k, las proporciones P_{ij} son iguales a uno para $i = j$ e iguales a cero para $i \neq j$:*

Tabla 2: Proporción de elementos de la variable u_k en relación a u_k^*

		$\boldsymbol{u_k^*}$			
	Conjunto	$\boldsymbol{U_x \cap U_w}$	$\boldsymbol{U_{x^c} \cap U_w}$	$\boldsymbol{U_{w^c}}$	$\sum_{j=1}^{3} P_{\cdot j}$
$\boldsymbol{u_k}$	$\boldsymbol{U_y \cap U_z}$	1	0	0	1
	$\boldsymbol{U_{y^c} \cap U_z}$	0	1	0	1
	$\boldsymbol{U_{z^c}}$	0	0	1	1

Para esta clasificación, el efecto de diseño del P.P.T. estimador de la dico–razón está dado por:

$$deff(PPT, \hat{R}_{PPT}) = \frac{4n(N-1)(1-R)RP_z}{m(N-n)} \tag{14}$$

y la varianza aproximada que se da en (13) es igual a cero; por ende, en este caso particular, $\hat{R}_{ESTMAS}$ es un estimador más eficiente que $\hat{R}_{PPT}$.

No siempre se tiene una clasificación uno a uno entre los conjuntos de las variables u_k y u_k^*, se considera ahora desde el punto de vista práctico, la tabla de proporciones 3.

Tabla 3: Proporción de elementos de la variable u_k^* en relación a u_k

		$\boldsymbol{u_k^*}$		
	Conjunto	$\boldsymbol{U_x \cap U_w}$	$\boldsymbol{U_{x^c} \cap U_w}$	$\boldsymbol{U_{w^c}}$
	$\boldsymbol{U_y \cap U_z}$	$P_{xywz}=\frac{N_{xywz}}{N_{xw}}$	$P_{x^cywz}=\frac{N_{x^cywz}}{N_{x^cw}}$	0
$\boldsymbol{u_k}$	$\boldsymbol{U_{y^c} \cap U_z}$	$P_{xy^cwz}=\frac{N_{xy^cwz}}{N_{xw}}$	$P_{x^cy^cwz}=\frac{N_{x^cy^cwz}}{N_{x^cw}}$	0
	$\boldsymbol{U_{z^c}}$	0	0	1
	$\sum_{i=1}^{3} P_{i\cdot}$	1	1	1

Para valores de P_{xywz} y P_{x^cywz} iguales a 0.2, 0.4, 0.6 y 0.8 se comparan, para este escenario, las varianzas de los estimadores $\hat{R}_{PPT}$, $\hat{R}_{ESTMAS}$ y $\hat{R}_{MAS}$ con tamaños poblacionales de 10000, 50000 y 100000 y tamaños muestrales de 100, 500 y 1000. Se obtiene lo siguiente:

i. Cuando P_z es igual a 0.2 o 0.5 la eficiencia relativa de $\hat{R}_{PPT}$ con respecto a $\hat{R}_{ESTMAS}$ se encuentra entre 0.2 y 0.8.

ii. Para P_{xywz} y P_{x^cywz} iguales a 0.2 o 0.8 y $P_z = 0.8$, la eficiencia relativa de $\hat{R}_{PPT}$ con respecto a $\hat{R}_{ESTMAS}$ es igual a 1.24 y en los demás casos de $P_z = 0.8$ dicha eficiencia relativa está entre 0.8 y 0.95.

iii. El efecto de diseño de $\hat{R}_{PPT}$ es igual al valor de P_z.

De lo anterior se puede concluir que si los elementos del conjunto U_{w^c} se clasifican casi en su totalidad en el conjunto U_{z^c}, el estimador $\hat{R}_{PPT}$ tiende a ser más eficiente que $\hat{R}_{MAS}$ y $\hat{R}_{ESTMAS}$.

El método de Hidiroglou para estimar una dico–razón

En el momento de estratificar es necesario decidir sobre la cantidad de estratos y cómo deben ser delimitados. Hidiroglou (1986) propone un método para un tamaño fijo de muestra, que consiste en dividir la población de estudio en dos estratos: uno en el que se aplica un diseño M.A.S y otro en el que todos los elementos hacen parte del estudio. Este procedimiento se basa en que, para el diseño IF–ESTMAS, la varianza del estimador de un total se comporta de manera parabólica con un mínimo que se puede encontrar por un método iterativo. En este trabajo se adapta este método a la estimación de una dico–razón y consiste en:

i. Ordenar los elementos de los subconjuntos $U_y = \{y_1, y_2, \ldots, y_k, \ldots, y_N\}$ y $U_z = \{z_1, z_2, \ldots, z_k, \ldots, z_N\}$ en forma descendente con respecto al valor absoluto de los elementos de la información auxiliar $U_{u^*} = \{u_1^*, u_2^*, \ldots, u_N^*\}$.

ii. Del ordenamiento anterior se tienen t elementos grandes que hacen parte del primer estrato y $(N - t)$ elementos pequeños candidatos a ser estudiados en otro estrato por un muestreo aleatorio simple.

iii. Para una muestra de tamaño n_t, compuesta por t elementos grandes y $(n_t - t)$ elementos pequeños, seleccionados por muestreo aleatorio simple, la varianza aproximada del estimador de la dico–razón se calcula como:

$$AV_{IF--ESTMAS}(\hat{R}) = \frac{(N-t)^2}{(n_t - t)}\left(1 - \frac{n_t - t}{N - t}\right) S^2_{u_{[N-t]}} \tag{15}$$

con

$$u_k = \frac{1}{N_z}(y_k - Rz_k) \tag{16}$$

iv. Se establece n_t fijo y se toma la menor $AV_{IF-ESTMAS}(\hat{R})$ calculada desde $t = 2$ hasta $t = n_t - 2$.

4. Comparación de las estimaciones

Con el fin de medir la precisión de las estimaciones de la dico–razón, se utiliza la metodología para distribuciones discretas expuesta en Martín, Ríos & Ríos (2000) para generar 125 poblaciones con $N = 10000$ mediante simulación. Las 125 poblaciones corresponden al cruce de 5 casos de $P_z : 0.2, 0.4, 0.5, 0.7, 0.9$ con 5 casos de $R = \dfrac{N_y}{N_z} : 0.1, 0.3, 0.5, 0.7, 0.9$ con 5 valores de coeficientes de contingencia $\rho_p : 0.1, 0.2, 0.4, 0.6, 0.8$.

Precisión de los estimadores estudiados

Para cada una de las 125 poblaciones simuladas se compara la precisión de los estimadores $\hat{R}_{MAS}$, $\hat{R}_{PPT}$, $\hat{R}_{ESTMAS}$ y $\hat{R}_{Hidiroglou}$ mediante el coeficiente de variación poblacional. De los resultados obtenidos se concluye lo siguiente:

i. Cuando se tiene una débil correspondencia entre u_k y u_k^* (ρ_p igual a 0.2), los estimadores que se estudian alcanzan la misma precisión. Para los demás grados de correspondencia se obtiene una mayor ganancia con los estimadores $\hat{R}_{PPT}$ y $\hat{R}_{ESTMAS}$.

ii. Por lo general, se alcanzan los mismos coeficientes de variación con el estimador $\hat{R}_{Hidiroglou}$ que con el estimador $\hat{R}_{MAS}$.

iii. La mayor ganancia lograda con el estimador $\hat{R}_{ESTMAS}$ en comparación con $\hat{R}_{MAS}$ y $\hat{R}_{PPT}$ se obtiene cuando el grado de correspondencia entre las variables u_k y u_k^* es medio o alto (ρ_p entre 0.4 y 0.8) y R_p está entre 0.3 y 0.7. Esta ganancia aumenta con el tamaño de muestra m, P_z y la razón poblacional R_p.

iv. La mayor ganancia que se alcanza utilizando el estimador $\hat{R}_{PPT}$ con respecto a $\hat{R}_{MAS}$ y $\hat{R}_{ESTMAS}$ se da cuando el valor de R_p es cercano a 0.1 o 0.9 y se tiene el mismo grado de correspondencia que en iii. La precisión de $\hat{R}_{PPT}$ crece junto a la razón poblacional R_p, el tamaño de muestra m y P_z.

v. Para obtener alguna ganancia con el estimador $\hat{R}_{PPT}$ cuando R_p es cercano a 0.1 y el grado de correspondencia entre las variables u_k y u_k^* es alto ($\rho \approx 0.8$) es necesario un tamaño de muestra de $n = 1000$ individuos.

vi. Cuando la razón poblacional R_p es cercana a 0.1 y el coeficiente de contingencia entre las variables u_k y u_k^* está entre 0.2 y 0.6, incluso para tamaños de muestra altos, el valor del coeficiente de variación para los estimadores $\hat{R}_{PPT}$ y $\hat{R}_{MAS}$ es mayor al 10 %. Lo anterior conduce a que utilizar los estimadores $\hat{R}_{PPT}$ y $\hat{R}_{MAS}$ en este caso particular no parece recomendable.

Para ilustrar se muestran los resultados en las siguientes gráficas:

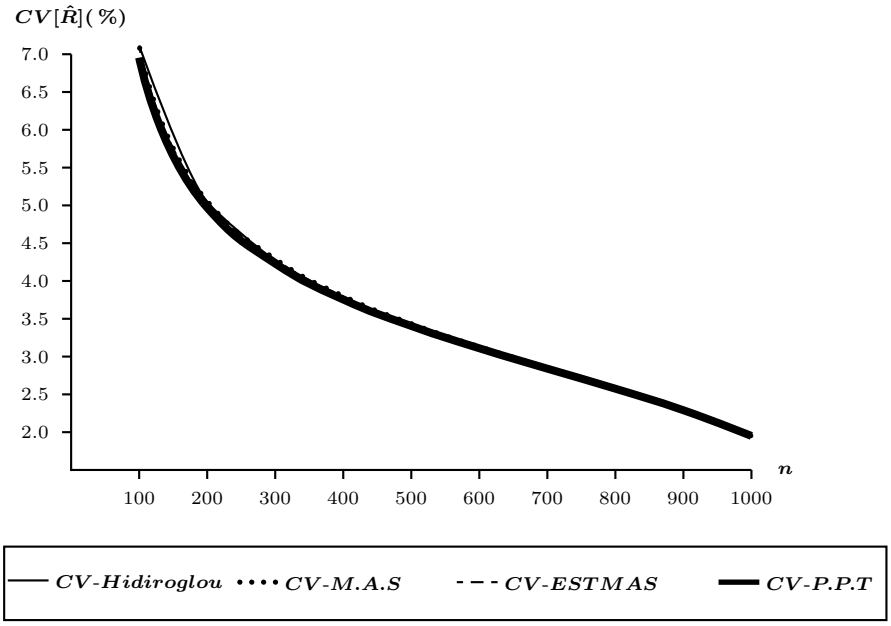

Figura 2: Coeficiente de contingencia (ρ_p) : 0.2 - Dico-razón: 0.9 - $P_z = 0.2$

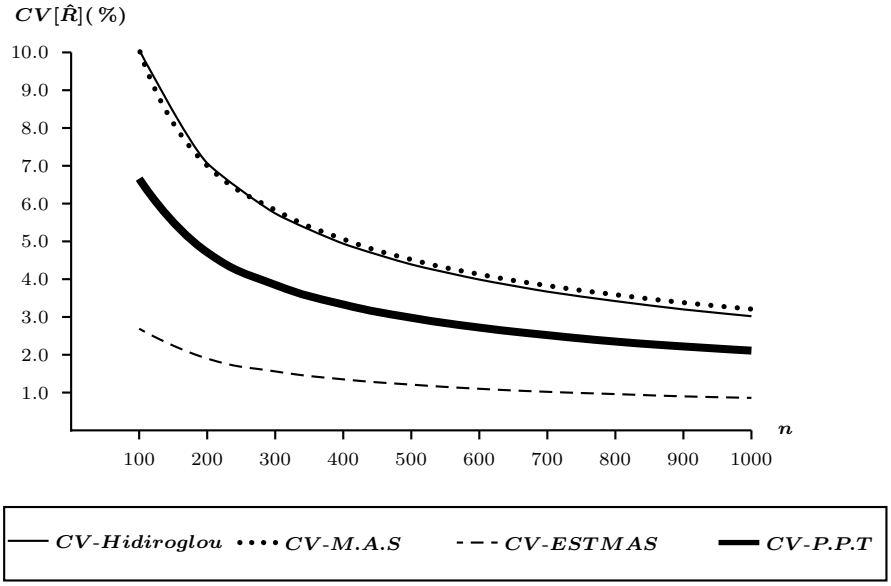

Figura 3: Coeficiente de contingencia (ρ_p) : 0.8 - Dico-razón: 0.7 - $P_z = 0.4$

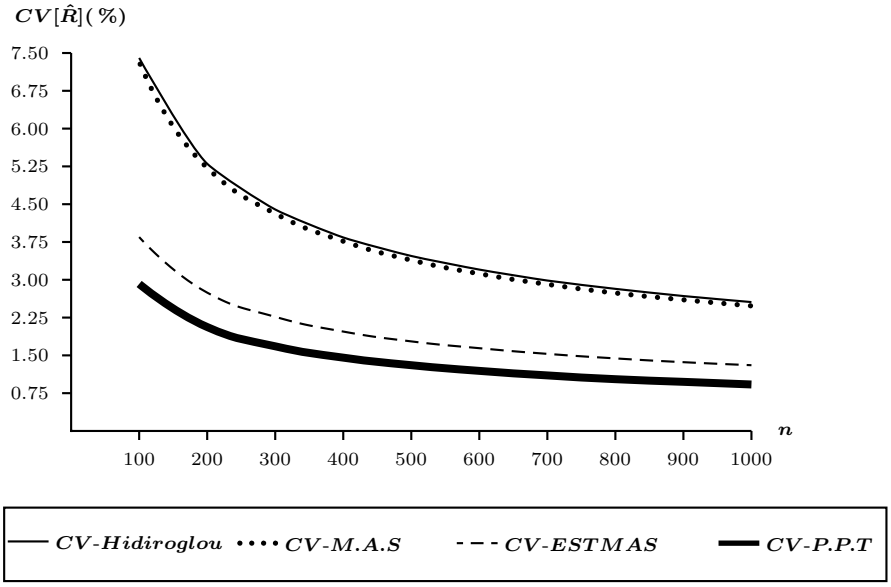

Figura 4: Coeficiente de contingencia (ρ_p) : 0.8 - Dico-razón: 0.9 - $P_z = 0.2$

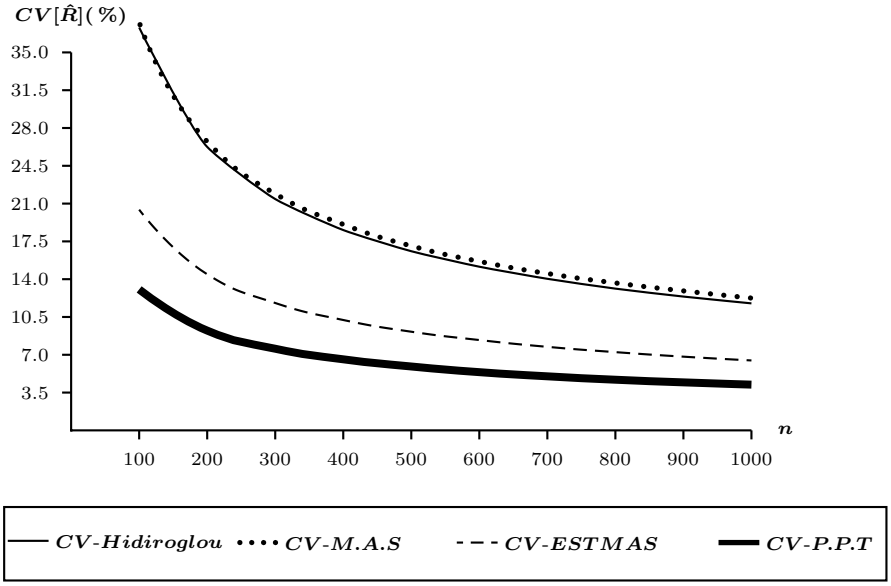

Figura 5: Coeficiente de contingencia (ρ_p) : 0.8 - Dico-razón: 0.1 - $P_z = 0.7$

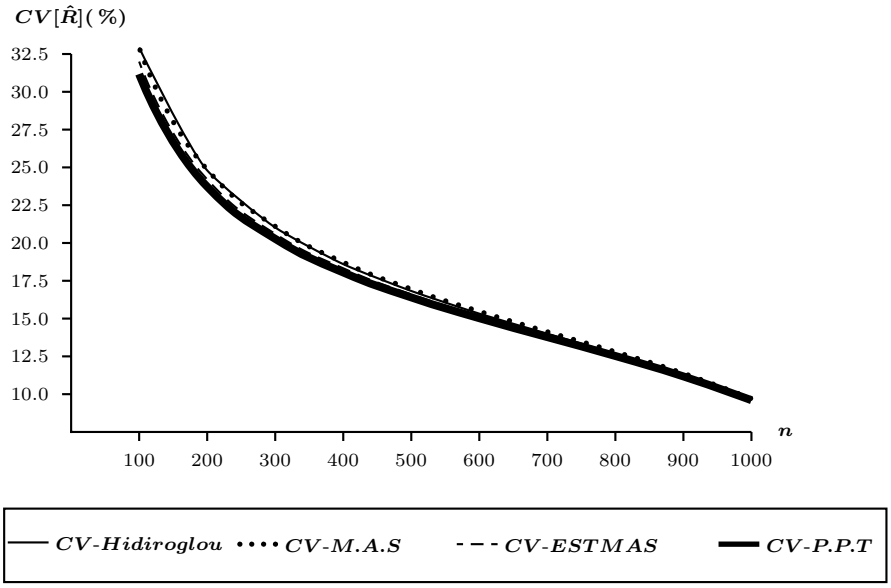

Figura 6: Coeficiente de contingencia (ρ_p) : 0.4 - Dico-razón: 0.1 - $P_z = 0.7$

5. Conclusiones

Como producto de los ejercicios desarrollados en este trabajo para la estimación de una dico–razón, se propone, en primer lugar, construir una variable auxiliar a partir de dos variables categóricas que tenga las mismas características de la variable u_k. En segundo lugar, con base en información proveniente de una fuente auxiliar o de un estudio piloto, establecer el coeficiente de contingencia entre las variables u_k y u_k^*. En caso de conseguir lo anterior algunos criterios para seleccionar el diseño más apropiado en la estimación de una dico–razón son:

1. Si la variable auxiliar discrimina casi perfectamente los conjuntos de la variable u_k se debe utilizar un diseño en tres estratos. Los estratos se construyen a partir de los tres conjuntos de la variable auxiliar y en cada uno se aplica un diseño M.A.S. con asignación proporcional de muestra.

2. Si la población es de N≈ 10000, el coeficiente de contingencia entre la variable auxiliar y u_k es medio o alto ($0.4 \leq \rho \leq 0.8$) y si se cuenta con una dico–razón poblacional entre 0.3 y 0.7 se puede utilizar un diseño ESTMAS con el fin de alcanzar una buena eficiencia. Para un grado de correspondencia entre 0.4 y 0.8 y una dico–razón cercana a 0.9 es aconsejable aplicar un diseño P.P.T. Si la variable auxiliar no se relaciona con la variable u_k ($\rho \leq 0.2$) y la dico–razón poblacional es distinta de 0.1 se debe utilizar un diseño M.A.S.

3. Para N≈ 10000 individuos, una dico–razón poblacional cercana a 0.1 y un grado de correspondencia entre la variable auxiliar y u_k alto ($\rho \approx 0.8$) se debe tomar un tamaño de muestra de 1000 individuos y utilizar un diseño P.P.T.

4. Cuando se puede establecer que los elementos del conjunto U_{w^c} se clasifican casi en su totalidad en el conjunto U_{z^c} independiente de los valores del coeficiente de contingencia y de P_z se recomienda utilizar un diseño P.P.T. En el ejemplo 2.1 este caso es equivalente a que los colegios que no contaban con sala de cómputo en el año 2004 en la actualidad tampoco cuentan con ello.

Recibido: 21 de Mayo de 2005
Aceptado: 6 de Octubre de 2005

Referencias

Cochran, W. G. (1963), *Sampling Techniques*, second edn, Wiley, New York.

Conover, W. J. (1980), *Practical Nonparametric Statistics*, second edn, John Wiley and Sons.

Fúquene, J. & Bautista, L. (2005), 'El diseño p.p.t. con variables categóricas para la estimación de dico-razones', *Revista Colombiana de Estadística* **28**, 99–114.

Fúquene, J. (2004), Criterios de selección y utilización de información auxiliar para optimizar la estimación de una razón de variables dicotómicas, Trabajo de grado, Universidad Nacional de Colombia.

Fúquene, J. (2005), Estratificación sesgo y eficiencia en la estimación de una proporción aplicando un diseño estratificado de muestreo, *in* 'Tercer Coloquio Regional de Estadística', Universidad Nacional de Colombia, Medellín.

Hidiroglou, M. (1986), 'The construction of a self-representing stratum of large units in survey design', *The American Statistician* **40**, 27–31.

Martín, J., Ríos, D. & Ríos, S. (2000), *Simulación, Métodos y Aplicaciones*, Ra-Ma, Madrid.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer Verlag, New York.