Revista Colombiana de Estadística
Volumen 28 N° 1. Págs. 99 a 114. Junio 2005

El diseño P.P.T. con variables categóricas para la estimación de dico–razones

Jairo Alberto Fúquene P.*
Jaime Leonardo Bautista S.**

Resumen

Para el diseño con repetición y probabilidad de selección proporcional al tamaño (P.P.T.) se encuentran las probabilidades de selección que minimizan la aproximación de la varianza en la estimación de una razón de totales de variables dicotómicas, dico–razón y se compara su eficiencia frente al diseño M.A.S.

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

Abstract

In probability proportional-to-size designs (P.P.S.) the selection probabilities that minimize the approximate variance of a totals ratio for dichotomic variables - dicho–ratio are found and their efficiency is compared with respect to S.I. design.

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

1. Introducción

Existen estudios muestrales donde las variables de interés son dicotómicas y se intenta estimar la proporción de éxitos, denominada también *dico–razón*; por ejemplo, la proporción de desempleados en Colombia dentro del conjunto de la población económicamente activa, o la proporción de personas que apoyan a un candidato entre los que piensan participar en una votación.

*Estadístico, Universidad Nacional de Colombia. Sede Bogotá. E-mail: jafuquenep@unal.edu.co
**Profesor asociado, Departamento de Estadística. Universidad Nacional de Colombia. E-mail: jlbautistas@unal.edu.co

Tanto en la literatura clásica (Cochran 1963), como en la más reciente (Särndal, Swensson & Wretman 1992), no se encuentra una metodología que explique el tipo de información auxiliar útil y su uso en la estimación de una proporción para diseños muestrales, en particular, para el diseño con probabilidad de selección proporcional al tamaño (P.P.T.). El principal problema que se presenta para este diseño consiste en encontrar las probabilidades de selección que minimizan la aproximación de la varianza del estimador, la expresión del estimador de la proporción y la estimación de su varianza.

Particularmente, es necesario investigar sobre el tipo de información auxiliar que se puede seleccionar y sobre su uso para que el estimador de una proporción sea óptimo cuando se aplica el diseño P.P.T. También es necesario investigar las condiciones bajo las cuales dicho estimador tiene mejores propiedades que uno que no utilice información auxiliar, como lo es el estimador de Horvitz–Thompson bajo muestreo aleatorio simple.

En este trabajo se establece el criterio de selección y utilización de información auxiliar categórica que optimiza la estimación de una proporción bajo el diseño P.P.T. y se estudia la eficiencia de este estimador. Para ello, en la segunda sección, se construye por medio de información auxiliar categórica, la variable auxiliar que se utiliza en el diseño P.P.T. En la sección tres, a través de multiplicadores de Lagrange, se encuentran las probabilidades óptimas para el diseño P.P.T. y en la última sección, se comparan las varianzas de los diseños M.A.S. y P.P.T. para distintos grados de correspondencia entre la variable auxiliar y la variable de interés.

2. Estimación de una dico–razón en diseños P.P.T.

Sean $U_y \subseteq U_z \subseteq U$ con cardinalidades N_y, N_z, N respectivamente y las variables dicotómicas que definen estos subconjuntos, y y z. La dico–razón de interés es $R = N_y/N_z$. La aproximación de la varianza de su estimador, alcanzada mediante linealización de Taylor, implica la determinación de la transformada $u_k = (y_k - Rz_k)/N_z$ que en este caso asume los valores:

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

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 $R = N_x/N_w$ con $u_k^* = (x_k - Rw_k)/N_w$ una variable auxiliar altamente correlacionada con u_k disponible para $k = 1, 2, ..., N$ y que asume los valores:

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

Así, $N_{x^c w}$ 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.

Nomenclatura en el espacio poblacional

Las cardinalidades poblacionales de las intersecciones entre los conjuntos de las variables u_k y u_k^*, son:

Tabla 1: Ilustración del espacio poblacional

		u_k^*			
	Conjunto	$U_x \cap U_w$	$U_{x^c} \cap U_w$	U_{w^c}	Total
	$U_y \cap U_z$	N_{xywz}	$N_{x^c ywz}$	$N_{yw^c z}$	N_{yz}
u_k	$U_{y^c} \cap U_z$	$N_{xy^c wz}$	$N_{x^c y^c wz}$	$N_{y^c w^c z}$	$N_{y^c z}$
	U_{z^c}	N_{xwz^c}	$N_{x^c wz^c}$	$N_{w^c z^c}$	N_{z^c}
	Total	N_{xw}	$N_{x^c w}$	N_{w^c}	N

en donde N_{xywz} es la cantidad de elementos que poseen la característica en x, y, w y z. Para definir las demás cantidades se debe tener en cuenta si los elementos poseen o no la característica en las variables que les corresponden.

Probabilidades de selección

En el caso considerado, las probabilidades de selección de los elementos del marco muestral, no se asignarán de manera individual sino a partir de su pertenencia a un conjunto determinado, de cuyo tamaño depende el valor de la probabilidad.

Se propone asignar una primera probabilidad, denominada α, a todos los elementos que se encuentran dentro del conjunto $U_x \cap U_w$; una probabilidad β a los elementos que pertenecen al conjunto $U_{x^c} \cap U_w$ y una probabilidad μ a los que pertenecen al conjunto U_{w^c}:

$$p_k = \begin{cases} \alpha & \text{si } k \in U_x \cap U_w \\ \\ \beta & \text{si } k \in U_{x^c} \cap U_w \\ \\ \mu & \text{si } k \in U_{w^c} \end{cases} \tag{3}$$

con las restricciones naturales:

$$0 < \alpha < 1 \qquad 0 < \beta < 1 \qquad 0 < \mu < 1 \tag{4}$$

y $\sum_U p_k = 1$, es decir:

$$\sum_U p_k = \sum_{U_x \cap U_w} \alpha + \sum_{U_{x^c} \cap U_w} \beta + \sum_{U_{w^c}} \mu = 1 \tag{5}$$

Sumando sobre cada uno de los correspondientes conjuntos se tiene:

$$N_{xw}\alpha + N_{x^c w}\beta + N_{w^c}\mu = 1 \tag{6}$$

Nomenclatura en el espacio muestral

Cuando se selecciona una muestra de tamaño m, el espacio muestral de la variable u_k en relación con u_k^* se encuentra distribuido de la siguiente forma:

Tabla 2: Ilustración del espacio muestral

		u_k^*			
	Conjunto	$U_x \cap U_w$	$U_{x^c} \cap U_w$	U_{w^c}	Total
	$S_y \cap S_z$	m_{xywz}	$m_{x^c ywz}$	$m_{yw^c z}$	m_{yz}
u_k	$S_{y^c} \cap S_z$	$m_{xy^c wz}$	$m_{x^c y^c wz}$	$m_{y^c w^c z}$	$m_{y^c z}$
	S_{z^c}	m_{xwz^c}	$m_{x^c wz^c}$	$m_{w^c z^c}$	m_{z^c}
	Total	m_{xw}	$m_{x^c w}$	m_{w^c}	m

Todas las cantidades se definen como se hace en el espacio poblacional con la diferencia que la cantidad de elementos están en la muestra ordenada. Por ejemplo, m_{xywz} denota la cantidad de elementos en la muestra ordenada de los que poseen la característica en x, y, w y z.

Resultado 2.1. *El estimador de la dico–razón $R = \dfrac{N_y}{N_z}$, bajo el diseño P.P.T, es:*

$$\widehat{R}_{PPT} = \frac{\widehat{t}_{y_{PPT}}}{\widehat{t}_{z_{PPT}}} = \left[\frac{m_{xywz}}{\alpha} + \frac{m_{x^c ywz}}{\beta} + \frac{m_{yw^c z}}{\mu}\right] \Bigg/ \left[\frac{m_{xywz} + m_{xy^c wz}}{\alpha} + \frac{m_{x^c ywz} + m_{x^c y^c wz}}{\beta} + \frac{m_{yw^c z} + m_{y^c w^c z}}{\mu}\right] \tag{7}$$

La aproximación de la varianza es:

$$AV_{PPT}(\widehat{R}) = \frac{1}{m}\frac{1}{(N-N_{z^c})^4}\left[\frac{N_{xywz}(N_{y^cz})^2 + N_{xy^cwz}(N_{yz})^2}{\alpha} + \frac{N_{x^cy^cwz}(N_{yz})^2}{\beta} + \frac{N_{x^cywz}(N_{y^cz})^2}{\beta} + \frac{N_{yw^cz}(N_{y^cz})^2 + N_{y^cw^cz}(N_{yz})^2}{\mu}\right] \tag{8}$$

Un estimador insesgado para (8) es:

$$\widehat{V}_{PPT}(\widehat{R}) = \frac{1}{m}\frac{1}{(m-1)}\frac{1}{\widehat{t}^2_{z_{PPT}}}\left[\left(\frac{m_{xywz}}{\alpha^2} + \frac{m_{x^cywz}}{\beta^2} + \frac{m_{yw^cz}}{\mu^2}\right)(1-\widehat{R}_{PPT})^2 + \left(\frac{m_{xy^cwz}}{\alpha^2} + \frac{m_{x^cy^cwz}}{\beta^2} + \frac{m_{y^cw^cz}}{\mu^2}\right)(-\widehat{R}_{PPT})^2\right] \tag{9}$$

Demostración. El estimador de la razón bajo el diseño P.P.T. se define como:

$$\widehat{R}_{PPT} = \frac{\widehat{t}_{y_{PPT}}}{\widehat{t}_{z_{PPT}}} = \frac{\frac{1}{m}\sum_{s_y}\frac{y_k}{p_k}}{\frac{1}{m}\sum_{s_z}\frac{z_k}{p_k}} \tag{10}$$

En la muestra, las variables de estudio y y z son dicotómicas y dependen del espacio muestral donde se definen. Entonces (10) queda determinado por:

$$\widehat{R}_{PPT} = \frac{\left[\sum_{s_y\cap s_z}\frac{1}{p_k} + \sum_{s_{y^c}\cap s_z}\frac{0}{p_k}\right]}{\left[\sum_{s_z\cap s_y}\frac{1}{p_k} + \sum_{s_z\cap s_{y^c}}\frac{1}{p_k} + \sum_{s_{z^c}}\frac{0}{p_k}\right]} \tag{11}$$

Sumando sobre las intersecciones de los conjuntos $s_y \cap s_z$ y $s_{y^c} \cap s_z$, es decir, sobre las filas de la tabla 2 donde se define la variable u_k, se obtiene que el numerador de (11) es:

$$\left[\sum_{s_y\cap s_z}\frac{1}{p_k} + \sum_{s_{y^c}\cap s_z}\frac{0}{p_k}\right] = \left[\sum_{(s_y\cap s_z\cap U_x\cap U_w)}\frac{1}{p_k} + \sum_{(s_y\cap s_z\cap U_{x^c}\cap U_w)}\frac{1}{p_k} + \sum_{(s_y\cap s_z\cap U_{w^c})}\frac{1}{p_k}\right] \tag{12}$$

De manera análoga a la anterior, se define el denominador de (11) como:

$$\left[\sum_{s_z\cap s_y}\frac{1}{p_k} + \sum_{s_z\cap s_{y^c}}\frac{1}{p_k} + \sum_{s_{z^c}}\frac{0}{p_k}\right] = \left[\sum_{(s_y\cap s_z\cap U_x\cap U_w)}\frac{1}{p_k} + \sum_{(s_y\cap s_z\cap U_{x^c}\cap U_w)}\frac{1}{p_k} + \sum_{(s_y\cap s_z\cap U_{w^c})}\frac{1}{p_k} + \sum_{(s_{y^c}\cap s_z\cap U_x\cap U_w)}\frac{1}{p_k} + \sum_{(s_{y^c}\cap s_z\cap U_{x^c}\cap U_w)}\frac{1}{p_k} + \sum_{(s_{y^c}\cap s_z\cap U_{w^c})}\frac{1}{p_k}\right] \tag{13}$$

Utilizando los p_k que se definen en (3), se obtiene que (12) está dado por:

$$\left[\sum_{s_y\cap s_z}\frac{1}{p_k}+\sum_{s_{y^c}\cap s_z}\frac{0}{p_k}\right]=\Bigg[\sum_{(s_y\cap s_z\cap U_x\cap U_w)}\frac{1}{\alpha}+\sum_{(s_y\cap s_z\cap U_{x^c}\cap U_w)}\frac{1}{\beta} + \sum_{(s_y\cap s_z\cap U_{w^c})}\frac{1}{\mu}\Bigg] \tag{14}$$

De manera análoga (13) es:

$$\left[\sum_{s_z\cap s_y}\frac{1}{p_k}+\sum_{s_z\cap s_{y^c}}\frac{1}{p_k}+\sum_{s_{z^c}}\frac{0}{p_k}\right]=\Bigg[\sum_{(s_y\cap s_z\cap U_x\cap U_w)}\frac{1}{\alpha}+\sum_{(s_y\cap s_z\cap U_{x^c}\cap U_w)}\frac{1}{\beta}+ \sum_{(s_y\cap s_z\cap U_{w^c})}\frac{1}{\mu}+\sum_{(s_{y^c}\cap s_z\cap U_x\cap U_w)}\frac{1}{\alpha}+ \sum_{(s_{y^c}\cap s_z\cap U_{x^c}\cap U_w)}\frac{1}{\beta}+\sum_{(s_{y^c}\cap s_z\cap U_{w^c})}\frac{1}{\mu}\Bigg] \tag{15}$$

De (14) y (15) se obtiene el estimador de la dico–razón:

$$\widehat{R}_{PPT}=\left[\frac{m_{xywz}}{\alpha}+\frac{m_{x^cywz}}{\beta}+\frac{m_{yw^cz}}{\mu}\right]\Bigg/\left[\frac{m_{xywz}}{\alpha}+\frac{m_{x^cywz}}{\beta}+\frac{m_{yw^cz}}{\mu}+ \frac{m_{xy^cwz}}{\alpha}+\frac{m_{x^cy^cwz}}{\beta}+\frac{m_{y^cw^cz}}{\mu}\right] \tag{16}$$

La aproximación de la varianza del estimador del parámetro $R=N_y/N_z$ bajo el diseño P.P.T. es:

$$AV_{PPT}(\widehat{R})=\frac{1}{m}\sum_{U}\frac{u_k^2}{p_k} \tag{17}$$

La variable u_k que se da en (1) tiene tres posibles valores, por ésto (17) queda definido como:

$$\begin{aligned}AV_{PPT}(\widehat{R})=\frac{1}{m}\Bigg[&\sum_{(U_y\cap U_z\cap U_x\cap U_w)}\frac{\left(N_{y^cz}/N_z^2\right)^2}{p_k}+\sum_{(U_y\cap U_z\cap U_{x^c}\cap U_w)}\frac{\left(N_{y^cz}/N_z^2\right)^2}{p_k}\\ &+\sum_{(U_y\cap U_z\cap U_{w^c})}\frac{\left(N_{y^cz}/N_z^2\right)^2}{p_k}+\sum_{(U_{y^c}\cap U_z\cap U_x\cap U_w)}\frac{\left(-N_{yz}/N_z^2\right)^2}{p_k}\\ &+\sum_{(U_{y^c}\cap U_z\cap U_{x^c}\cap U_w)}\frac{\left(-N_{yz}/N_z^2\right)^2}{p_k}+\sum_{(U_{y^c}\cap U_z\cap U_{w^c})}\frac{\left(-N_{yz}/N_z^2\right)^2}{p_k}\\ &+\sum_{(U_{z^c}\cap U_x\cap U_w)}\frac{0}{p_k}+\sum_{(U_{z^c}\cap U_{x^c}\cap U_w)}\frac{0}{p_k}+\sum_{(U_{z^c}\cap U_{w^c})}\frac{0}{p_k}\Bigg]\end{aligned} \tag{18}$$

Utilizando los p_k que se definen en (3) se llega a:

$$\begin{aligned} AV_{PPT}(\widehat{R}) = \frac{1}{m}\Bigg[& \sum_{(U_y\cap U_z\cap U_x\cap U_w)} \frac{\left(N_{y^c z}/N_z^2\right)^2}{\alpha} + \sum_{(U_y\cap U_z\cap U_{x^c}\cap U_w)} \frac{\left(N_{y^c z}/N_z^2\right)^2}{\beta} \\ & + \sum_{(U_y\cap U_z\cap U_{w^c})} \frac{\left(N_{y^c z}/N_z^2\right)^2}{\mu} + \sum_{(U_{y^c}\cap U_z\cap U_x\cap U_w)} \frac{\left(-N_{yz}/N_z^2\right)^2}{\alpha} \\ & + \sum_{(U_{y^c}\cap U_z\cap U_{x^c}\cap U_w)} \frac{\left(-N_{yz}/N_z^2\right)^2}{\beta} + \sum_{(U_{y^c}\cap U_z\cap U_{w^c})} \frac{\left(-N_{yz}/N_z^2\right)^2}{\mu} \\ & + \sum_{(U_{z^c}\cap U_x\cap U_w)} \frac{0}{\alpha} + \sum_{(U_{z^c}\cap U_{x^c}\cap U_w)} \frac{0}{\beta} + \sum_{(U_{z^c}\cap U_{w^c})} \frac{0}{\mu}\Bigg] \end{aligned} \tag{19}$$

Sumando sobre los correspondientes conjuntos se obtiene:

$$\begin{aligned} AV_{PPT}(\widehat{R}) = \frac{1}{m}\frac{1}{(N-N_{z^c})^4}\Bigg[& \frac{N_{xywz}(N_{y^c z})^2}{\alpha} + \frac{N_{x^c ywz}(N_{y^c z})^2}{\beta} + \frac{N_{yw^c z}(N_{y^c z})^2}{\mu} \\ & + \frac{N_{xy^c wz}(-N_{yz})^2}{\alpha} + \frac{N_{x^c y^c wz}(-N_{yz})^2}{\beta} + \frac{N_{y^c w^c z}(-N_{yz})^2}{\mu}\Bigg] \end{aligned} \tag{20}$$

La estimación de la varianza del estimador de la razón bajo el diseño P.P.T. es:

$$\widehat{V}_{PPT}(\widehat{R}) = \frac{1}{m}\frac{1}{(m-1)}\sum_{i=1}^{m}\left(\frac{\widehat{u}_{ki}}{p_{ki}}\right)^2 \tag{21}$$

De manera análoga al cálculo de la aproximación de la varianza se obtiene:

$$\begin{aligned} \widehat{V}_{PPT}(\widehat{R}) = \frac{1}{m}\frac{1}{(m-1)}\Bigg[& \sum_{(s_y\cap s_z\cap U_x\cap U_w)} \left(\frac{\widehat{u}_k}{\alpha}\right)^2 + \sum_{(s_y\cap s_z\cap U_{x^c}\cap U_w)} \left(\frac{\widehat{u}_k}{\beta}\right)^2 \\ & + \sum_{(s_y\cap s_z\cap U_{w^c})} \left(\frac{\widehat{u}_k}{\mu}\right)^2 + \sum_{(s_{y^c}\cap s_z\cap U_x\cap U_w)} \left(\frac{\widehat{u}_k}{\alpha}\right)^2 \\ & + \sum_{(s_{y^c}\cap s_z\cap U_{x^c}\cap U_w)} \left(\frac{\widehat{u}_k}{\beta}\right)^2 + \sum_{(s_{y^c}\cap s_z\cap U_{w^c})} \left(\frac{\widehat{u}_k}{\mu}\right)^2 \\ & + \sum_{(s_{z^c}\cap U_x\cap U_w)} \frac{0}{\alpha} + \sum_{(s_{z^c}\cap U_{x^c}\cap U_w)} \frac{0}{\beta} + \sum_{(s_{z^c}\cap U_{w^c})} \frac{0}{\mu}\Bigg] \end{aligned} \tag{22}$$

La transformación para la variable $\widehat{u}_k$ asume los valores siguientes:

$$\widehat{u}_k = \begin{cases} \dfrac{(1-\widehat{R}_{PPT})}{\widehat{t}_{z_{PPT}}} & \text{si } k \,\epsilon\, s_y \cap s_z \\[2ex] -\dfrac{\widehat{R}_{PPT}}{\widehat{t}_{z_{PPT}}} & \text{si } k \,\epsilon\, s_{y^c} \cap s_z \\[2ex] 0 & \text{si } k \,\epsilon\, s_{z^c} \end{cases} \tag{23}$$

Por (23) y sumando sobre los correspondientes conjuntos, (22) queda determinado por:

$$\begin{aligned} \widehat{V}_{PPT}(\widehat{R}) = \frac{1}{m}\frac{1}{(m-1)}\Bigg[&\left(\frac{m_{xywz}}{\alpha^2}+\frac{m_{x^cywz}}{\beta^2}+\frac{m_{yw^cz}}{\mu^2}\right)\left(\frac{1-\widehat{R}_{PPT}}{\widehat{t}_{z_{PPT}}}\right)^2 \\ &+\left(\frac{m_{xy^cwz}}{\alpha^2}+\frac{m_{x^cy^cwz}}{\beta^2}+\frac{m_{y^cw^cz}}{\mu^2}\right)\left(\frac{-\widehat{R}_{PPT}}{\widehat{t}_{z_{PPT}}}\right)^2\Bigg] \quad \square \end{aligned} \tag{24}$$

3. Valores de p_k que minimizan la varianza aproximada del estimador de la dico–razón

En esta sección se busca construir las probabilidades α, β y μ que hagan mínima la varianza del estimador de la dico–razón bajo el diseño P.P.T. Para ello, se estudian los puntos extremos de la aproximación de la varianza como función de las probabilidades:

$$f(\alpha,\beta,\mu) = AV_{PPT}(\widehat{R}) \tag{25}$$

con (25) limitada por el siguiente plano:

$$(g(\alpha,\beta,\mu)-1) = N_{xw}\alpha + N_{x^cw}\beta + N_{w^c}\mu = 1 \tag{26}$$

donde N_{xw}, N_{x^cw} y N_{w^c} son las cardinalidades que se dan en la tabla 1. Se utiliza entonces el método de los multiplicadores de Lagrange. Se igualan a cero las derivadas parciales de la función auxiliar siguiente:

$$h(\alpha,\beta,\mu) = f(\alpha,\beta,\mu) - \lambda(g(\alpha,\beta,\mu)-1) \tag{27}$$

Para (27) se obtiene las siguientes ecuaciones:

$$-\frac{1}{m}\frac{1}{N_z^2}\left[\frac{N_{xywz}(N_{y^cz})^2 + N_{xy^cwz}(N_{yz})^2}{\alpha^2}\right] = (\lambda)N_{xw} \tag{28}$$

$$-\frac{1}{m}\frac{1}{N_z^2}\left[\frac{N_{x^cywz}(N_{y^cz})^2 + N_{x^cy^cwz}(N_{yz})^2}{\beta^2}\right] = (\lambda)N_{x^cw} \tag{29}$$

$$-\frac{1}{m}\frac{1}{N_z^2}\left[\frac{N_{yw^cz}(N_{y^cz})^2 + N_{y^cw^cz}(N_{yz})^2}{\mu^2}\right] = (\lambda)N_{w^c} \tag{30}$$

$$N_{xw}\alpha + N_{x^cw}\beta + N_{w^c}\mu - 1 = 0 \tag{31}$$

Las soluciones de este sistema de ecuaciones son:

$$\alpha = (A)\beta; \qquad \mu = (B)\beta; \qquad \beta = \frac{1}{(N_{xw}A + N_{w^c}B + N_{x^cw})} \tag{32}$$

donde,

$$\begin{aligned} A &= \pm\sqrt{\frac{N_{x^cw}(N_{xywz}(N_{y^cz})^2 + N_{xy^cwz}(N_{yz})^2)}{N_{xw}(N_{x^cywz}(N_{y^cz})^2 + N_{x^cy^cwz}(N_{yz})^2)}} \\ B &= \pm\sqrt{\frac{N_{x^cw}(N_{yw^cz}(N_{y^cz})^2 + N_{y^cw^cz}(N_{yz})^2)}{N_{w^c}(N_{x^cywz}(N_{y^cz})^2 + N_{x^cy^cwz}(N_{yz})^2)}} \end{aligned} \tag{33}$$

Nota 3.1. *Para (33) se tienen los siguientes comentarios:*

i. *Se deben descartar las raíces con signo negativo, debido a que las probabilidades que se dan en (32) deben ser mayores a cero.*

ii. *Todos los elementos que están dentro de las raíces son mayores o iguales a cero, pues ellos denotan la cantidad de elementos en los diferentes conjuntos.*

iii. *Si A o B tienen denominador igual a cero se tiene una indeterminación. Esto sucede cuando, desde el punto de vista práctico, se tienen los siguientes casos extremos:*

   a) $N_{xw} = N_{w^c} = N_z = N_{x^cw} = 0$

   b) $N_{x^cw} = N_{x^cwz^c}$

Se aplica el criterio de la segunda derivada para saber si (α, β, μ) determina un mínimo, un máximo o un punto silla; para ello se calcula el determinante del hessiano limitado para la función auxiliar $h(\alpha,\beta,\mu) = f(\alpha,\beta,\mu) - \lambda(g(\alpha,\beta,\mu) - 1)$ y luego, se examinan los determinantes de las submatrices diagonales de orden ≥ 3 en los puntos críticos de h, como sigue:

$$
\begin{aligned}
|\bar{H}_2| &= \begin{bmatrix} 0 & -\dfrac{\partial g}{\partial \alpha} & -\dfrac{\partial g}{\partial \beta} \\ -\dfrac{\partial g}{\partial \alpha} & \dfrac{\partial^2 h}{\partial \alpha^2} & \dfrac{\partial^2 h}{\partial \alpha \partial \beta} \\ -\dfrac{\partial g}{\partial \beta} & \dfrac{\partial^2 h}{\partial \alpha \partial \beta} & \dfrac{\partial^2 h}{\partial \beta^2} \end{bmatrix} \\
&= -\frac{2}{m}\frac{1}{(N-N_{z^c})^4}[(A)N_{xw}^2 + (B)N_{x^cw}^2] \qquad (34) \\
|\bar{H}_3| &= \begin{bmatrix} 0 & -\dfrac{\partial g}{\partial \alpha} & -\dfrac{\partial g}{\partial \beta} & -\dfrac{\partial g}{\partial \mu} \\ -\dfrac{\partial g}{\partial \alpha} & \dfrac{\partial^2 h}{\partial \alpha^2} & \dfrac{\partial^2 h}{\partial \alpha \partial \beta} & \dfrac{\partial^2 h}{\partial \alpha \partial \mu} \\ -\dfrac{\partial g}{\partial \beta} & \dfrac{\partial^2 h}{\partial \alpha \partial \beta} & \dfrac{\partial^2 h}{\partial \beta^2} & \dfrac{\partial^2 h}{\partial \beta \partial \mu} \\ -\dfrac{\partial g}{\partial \mu} & \dfrac{\partial^2 h}{\partial \alpha \partial \mu} & \dfrac{\partial^2 h}{\partial \beta \partial \mu} & \dfrac{\partial^2 h}{\partial \mu^2} \end{bmatrix} \\
&= -\frac{2}{m}\frac{1}{(N-N_{z^c})^4}[(AC)N_{xw}^2 + (BC)N_{x^cw}^2 + BDA] \qquad (35)
\end{aligned}
$$

donde

$$
A = \frac{N_{x^cywz}(N_{y^cz})^2 + N_{x^cy^cwz}(N_{yz})^2}{\beta^3} \qquad B = \frac{N_{xywz}(N_{y^cz})^2 + N_{xy^cwz}(N_{yz})^2}{\alpha^3}
$$
$$
C = \frac{N_{yw^cz}(N_{y^cz})^2 + N_{y^cw^cz}(N_{yz})^2}{\mu^3} \qquad D = (B)\frac{3}{\alpha}
$$

Es claro que $|\bar{H}_2|$ y $|\bar{H}_3|$ son menores que cero; por lo tanto, el punto (α, β, μ) es un mínimo local para la función $f(\alpha, \beta, \mu) = AV_{PPT}(\widehat{R})$, limitada por el plano $N_{xw}\alpha + N_{x^cw}\beta + N_{w^c}\mu = 1$.

Resultado 3.1. *Para la dico–razón* $R = \dfrac{N_y}{N_z}$, *el estimador bajo el diseño P.P.T. con la mínima aproximación de la varianza es:*

$$
\begin{aligned}
\widehat{R}_{PPT} = \left[\frac{m_{xywz}}{\alpha_0} + \frac{m_{x^cywz}}{\beta_0} + \frac{m_{yw^cz}}{\mu_0}\right] \Bigg/ \Bigg[&\frac{m_{xywz} + m_{xy^cwz}}{\alpha_0} \\
&+ \frac{m_{x^cywz} + m_{x^cy^cwz}}{\beta_0} + \frac{m_{yw^cz} + m_{y^cw^cz}}{\mu_0}\Bigg] \qquad (36)
\end{aligned}
$$

con la siguiente aproximación de la varianza:

$$
\begin{aligned}
AV_{PPT}(\widehat{R}) = \frac{1}{m}\frac{1}{(N-N_{z^c})^4}\Bigg[&\frac{N_{xywz}(N_{y^cz})^2 + N_{xy^cwz}(N_{yz})^2}{\alpha_0} + \frac{N_{x^cywz}(N_{y^cz})^2}{\beta_0} \\
&+ \frac{N_{x^cy^cwz}(N_{yz})^2}{\beta_0} + \frac{N_{yw^cz}(N_{y^cz})^2 + N_{y^cw^cz}(N_{yz})^2}{\mu_0}\Bigg] \qquad (37)
\end{aligned}
$$

El siguiente es un estimador insesgado para (37):

$$\widehat{V}_{PPT}(\widehat{R}) = \frac{1}{m}\frac{1}{(m-1)}\frac{1}{\widehat{t}^2_{z_{PPT}}}\Bigg[\left(\frac{m_{xywz}}{\alpha_0^2}+\frac{m_{x^cywz}}{\beta_0^2}+\frac{m_{yw^cz}}{\mu_0^2}\right)(1-\widehat{R}_{PPT})^2 + \left(\frac{m_{xy^cwz}}{\alpha_0^2}+\frac{m_{x^cy^cwz}}{\beta_0^2}+\frac{m_{y^cw^cz}}{\mu_0^2}\right)(-\widehat{R}_{PPT})^2\Bigg] \tag{38}$$

donde

$$\alpha_0 = (A)\beta_0; \qquad \mu_0 = (B)\beta_0 \qquad \beta_0 = \frac{1}{(N_{xw}(A)+N_{w^c}(B)+N_{x^cw})} \tag{39}$$

con

$$A = \sqrt{\frac{N_{x^cw}(N_{xywz}(N_{y^cz})^2 + N_{xy^cwz}(N_{yz})^2)}{N_{xw}(N_{x^cywz}(N_{y^cz})^2 + N_{x^cy^cwz}(N_{yz})^2)}}$$

$$B = \sqrt{\frac{N_{x^cw}(N_{yw^cz}(N_{y^cz})^2 + N_{y^cw^cz}(N_{yz})^2)}{N_{w^c}(N_{x^cywz}(N_{y^cz})^2 + N_{x^cy^cwz}(N_{yz})^2)}} \tag{40}$$

4. Efecto de diseño del P.P.T. estimador de la dico–razón

Para el resultado 3.1, A y B se pueden escribir de la manera siguiente:

$$A = \sqrt{\frac{P_{xywz}+P_{xy^cwz}(P_{y^cy})^2}{P_{x^cywz}+P_{x^cy^cwz}(P_{y^cy})^2}} \qquad B = \sqrt{\frac{P_{yw^cz}+P_{y^cw^cz}(P_{y^cy})^2}{P_{x^cywz}+P_{x^cy^cwz}(P_{y^cy})^2}} \tag{41}$$

con

$$P_{y^cy} = \frac{P_{xy^cwz}(N_{xw})+P_{x^cy^cwz}(N_{x^cw})+P_{y^cw^cz}(N_{w^c})}{P_{xywz}(N_{xw})+P_{x^cywz}(N_{x^cw})+P_{yw^cz}(N_{w^c})} \tag{42}$$

Para establecer α_0, β_0 y μ_0, el usuario debe contar con información que le permita aproximarse a la tabla 3, en donde P_{xywz} es la proporción de elementos que poseen la característica en x, y, w y z. Para definir las demás proporciones debe tener en cuenta si los elementos poseen o no la característica en las variables que les corresponden.

Resultado 4.1. *El efecto de diseño del estimador de la dico–razón bajo el diseño P.P.T. está dado por:*

$$deff(PPT,\widehat{R}_{PPT}) = \frac{n(N-1)t_2}{mN(N-n)\beta_0^2(N_{yz}+N_{y^cz})^3N_{x^cw}R(1-R)} \tag{43}$$

donde

$$n = N - [N_{xw}(1-\alpha_0)^m + N_{x^cw}(1-\beta_0)^m + N_{w^c}(1-\mu_0)^m] \tag{44}$$

Tabla 3: Proporciones para establecer α_0, β_0 y μ_0

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

con

$$t_2 = N_{x^cywz}(N_{y^cz})^2 + N_{x^cy^cwz}(N_{yz})^2$$

y β_0 *como en (39).*

Demostración. La varianza del estimador de la dico–razón, $R = \dfrac{N_y}{N_z}$, bajo el diseño M.A.S. se obtiene mediante el desarrollo de la aproximación de Taylor, para lo cual se aplica la transformación $u_k = \dfrac{1}{N_z}(y_k - Rz_k)$:

$$u_k = \begin{cases} \dfrac{1}{N_z}(1-R) & \text{si } y_k = 1 \quad z_k = 1, \\ \\ \dfrac{1}{N_z}(0-R) & \text{si } y_k = 0 \quad z_k = 1, \\ \\ 0 & \text{si } y_k = 0 \quad z_k = 0. \end{cases}$$

El caso $y_k = 1$, $z_k = 0$ no se puede dar puesto que $U_y \subseteq U_z$. Debido a que y y z son variables $0, 1$, se cumple que:

$$\sum\nolimits_U y_k = N_y; \qquad \sum\nolimits_U z_k = N_z; \qquad \bar{y}_U = P_y;$$
$$\bar{z}_U = P_z; \qquad Q_y = 1 - P_y; \qquad Q_z = 1 - P_z$$

Se tiene entonces

$$S^2_{yU} = \frac{1}{N-1}\sum\nolimits_U (y_k - \bar{y}_U)^2 = \frac{N}{N-1}(P_yQ_y)$$

y análogamente

$$S^2_{zU} = \frac{N}{N-1}(P_zQ_z)$$

Como $U_y \subseteq U_z$, entonces $\sum_U y_k z_k = N_y$, para la covarianza poblacional:

$$S_{yzU} = \frac{1}{N-1}\left(\sum\nolimits_U y_k z_k - \frac{\sum_U y_k \sum_U z_k}{N}\right) = \frac{N}{N-1}(P_y Q_z)$$

Por lo anterior la aproximación de la varianza queda definida como:

$$AV_{MAS}(\widehat{R}) = \frac{N}{N-1}\frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{P_z}R(1-R) \tag{45}$$

Dividiendo (37) entre (45), se obtiene el efecto de diseño de la dico–razón para el diseño P.P.T.:

$$deff(PPT, \widehat{R}_{PPT}) = \frac{n(N-1)t_2}{mN(N-n)\beta_0^2(N_{yz}+N_{y^c z})^3 N_{x^c w} R(1-R)} \tag{46}$$

Puesto que el efecto de diseño tiene sentido sólo bajo el supuesto de tamaños de muestra equiparables, es decir:

$$\sum_U \pi_k = E(n_S) = n \tag{47}$$

Como para el diseño P.P.T. se cumple que $\pi_k = 1-(1-p_k)^m$, se tiene:

$$\begin{aligned} E(n_S) &= n \\ &= \sum_U 1-(1-p_k)^m \\ &= N - \left[\sum_{U_x \cap U_w}(1-\alpha)^m + \sum_{U_{x^c} \cap U_w}(1-\beta)^m + \sum_{U_{w^c}}(1-\mu)^m\right] \\ &= N - [N_{xw}(1-\alpha)^m + N_{x^c w}(1-\beta)^m + N_{w^c}(1-\mu)^m] \end{aligned} \tag{48}$$

El tamaño de muestra anterior corresponde a un diseño sin reposición que se equipara a uno con reposición de tamaño m. Al utilizar el punto mínimo local $(\alpha_0, \beta_0, \mu_o)$ que se encuentra en (39) se tiene que:

$$n = N - [N_{xw}(1-\alpha_0)^m + N_{x^c w}(1-\beta_0)^m + N_{w^c}(1-\mu_0)^m]$$

Eficiencia y sesgo del estimador de la dico–razón

Con el fin de medir la eficiencia y el sesgo del estimador de la dico–razón, se utiliza la metodología para distribuciones discretas expuesta en (Martín, Ríos & Ríos 2000) para generar 75 poblaciones con $N = 10000$ mediante simulación. El 75 corresponde al cruce de 3 casos de P_z con 5 casos de $R = N_y/N_z$ con 5 tipos de coeficientes de contingencia[1]. Entonces, para cada uno de los 75 casos, se toman 5000 muestras de tamaño $m = 100$. Los estimadores de interés son:

[1] Este coeficiente se calcula entre las variables u_k y u_k^* y se denota como ρ.

1. El valor absoluto de la estimación del sesgo relativo del estimador de la dico–razón se obtiene mediante:

$$\widehat{B}_r[\widehat{R}_{PPT}] = \left| \frac{\bar{\widehat{R}}_{PPT} - R}{\sqrt{S^2_{\widehat{R}_{PPT}}}} \right| \tag{49}$$

donde $\bar{\widehat{R}}_{PPT} = \dfrac{1}{5000} \sum_{j=1}^{5000} \widehat{R}_j$, con $\widehat{R}_j$ la estimación de la dico–razón para la j-ésima muestra de tamaño $m = 100$. $S^2_{\widehat{R}_{PPT}} = \dfrac{\sum_{j=1}^{5000}(\widehat{R}_j - \bar{\widehat{R}}_{PPT})^2}{5000-1}$, que es una estimación de la aproximación de la varianza de la dico–razón, $AV_{PPT}(\widehat{R})$.

2. La estimación del efecto de diseño $deff$, se calcula como:

$$\widehat{Deff}(\widehat{R}, PPT, n) = \frac{S^2_{\widehat{R}_{PPT}}}{S^2_{\widehat{R}_{MAS}}} \tag{50}$$

donde $S^2_{\widehat{R}_{MAS}}$ se calcula bajo el diseño M.A.S. y es análoga a $S^2_{\widehat{R}_{PPT}}$.

Los resultados se muestran en la tabla 4.

5. Conclusiones

Del estudio de simulación se concluye lo siguiente:

1. En el proceso de simulación se puede observar que, para $P_z = 0{,}2$, $\widehat{R}_{PPT}$ presenta un sesgo que, en general, se puede considerar insignificante, pues en todos los casos el valor absoluto del sesgo relativo es menor a 0,1. En cuanto a la eficiencia, la simulación muestra que en los escenarios donde $P_z = 0{,}2$ y se presenta una débil correspondencia con la variable auxiliar, $\widehat{R}_{PPT}$ es aproximadamente igual de eficiente a $\widehat{R}_{MAS}$. Como era de esperarse, la eficiencia de $\widehat{R}_{PPT}$ aumenta con el grado de correspondencia entre las variables u_k y u_k^*.

2. Cuando $P_z = 0{,}5$ de nuevo el valor absoluto del sesgo relativo para $\widehat{R}_{PPT}$ es menor que 0,1. En este escenario $\widehat{R}_{PPT}$ es al menos tan eficiente como $\widehat{R}_{MAS}$, pues la estimación del efecto de diseño de $\widehat{R}_{PPT}$ es siempre menor o igual a uno.

3. La simulación muestra que cuando $P_z = 0{,}9$, el sesgo relativo de $\widehat{R}_{PPT}$ puede considerarse despreciable. De otro lado, $\widehat{R}_{PPT}$ es en el 95 % de los casos al menos tan eficiente como $\widehat{R}_{MAS}$ y esta eficiencia aumenta con el grado de correspondencia.

Tabla 4: Resultados de una simulación de 5.000 muestras de tamaño $m = 100$ para 75 poblaciones de $N = 10000$.

ρ	R	P_z	$\widehat{B}_r[\widehat{R}_{PPT}]$	$\widehat{Deff}$	P_z	$\widehat{B}_r[\widehat{R}_{PPT}]$	$\widehat{Deff}$	P_z	$\widehat{B}_r[\widehat{R}_{PPT}]$	$\widehat{Deff}$
	0,1		0,008	0,9		0,007	0,98		0,014	1,0
	0,3		0,021	1,1		0,019	0,95		0,018	0,94
0,1	0,5	0,2	0,007	1,1	0,5	0,021	1,0	0,9	0,033	1,0
	0,7		0,016	1,1		0,082	0,96		0,000	0,95
	0,9		0,031	1,1		0,010	1,0		0,026	1,0
	0,1		0,036	1,0		0,009	0,94		0,001	0,92
	0,3		0,008	1,1		0,008	0,91		0,017	0,98
0,2	0,5	0,2	0,024	1,0	0,5	0,031	0,98	0,9	0,022	0,96
	0,7		0,002	1,1		0,024	0,91		0,018	1,1
	0,9		0,012	1,1		0,001	0,93		0,029	0,90
	0,1		0,032	0,76		0,032	0,75		0,041	0,83
	0,3		0,005	0,86		0,019	0,79		0,027	0,92
0,4	0,2	0,2	0,035	0,81	0,5	0,037	0,98	0,9	0,046	0,95
	0,7		0,025	0,88		0,019	0,89		0,006	0,90
	0,9		0,015	0,84		0,025	0,78		0,010	0,81
	0,1		0,066	0,59		0,042	0,28		0,020	0,66
	0,3		0,018	0,62		0,017	0,83		0,032	0,83
0,6	0,2	0,2	0,006	0,57	0,5	0,052	0,86	0,9	0,056	1,0
	0,7		0,000	0,56		0,001	0,79		0,028	0,86
	0,9		0,021	0,53		0,020	0,25		0,016	0,70
	0,1		0,085	0,12		0,097	0,35		0,063	0,39
	0,3		0,095	0,22		0,046	0,58		0,031	0,82
0,8	0,2	0,2	0,012	0,22	0,5	0,098	0,61	0,9	0,001	0,85
	0,7		0,094	0,18		0,024	0,54		0,057	0,73
	0,9		0,069	0,12		0,059	0,22		0,080	0,39

Bibliografía

Cochran, W. G. (1963), *Sampling Techniques*, 2 edn, Wiley, New York.

Conover, W. J. (1980), *Practical Nonparametric Statistics*, 2 edn, John Wiley and Sons.

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

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.