UNA CLASE DE ESTIMADORES BASADOS EN UNA RAZÓN: MUESTREO SIMPLE ALEATORIO Y MUESTREO POR CONJUNTOS ORDENADOS

Carlos N. Bouza.[1]
Universidad de La Habana

ABSTRACT
A class of estimators is determined. It contains important families of ratio estimators. The belonging to it is characterized by means of a parametric vector. It allows identifying that more than a dozen of estimators are particular cases. Their behavior when ranked set sampling is used is established by fixing the gains in accuracy of them. Numerical experiments establish which is the performance of them in four applications and by generating normal, exponential and uniform populations

KEY WORDS: Relative precision, Gain in accuracy. Order statistics, moments

MSC: 62D05

RESUMEN
Una clase de estimadores es construida. Esta contiene familias importantes de estimadores de razón. Se caracteriza la pertenencia a ella a partir de un vector de parámetros. Este permite identificar como casos particulares a mas de una docena de estimadores de razón. El comportamiento de estos al utilizar muestro por conjuntos ordenados es establecido al fijar las ganancias en precisión de estos. Experimentos numéricos establecen cual es el comportamiento de ellos en cuatro aplicaciones y al generar poblaciones normales, exponenciales y uniformes.

1. INTRODUCCIÓN

El estimador de razón es muy popularmente utilizado en las aplicaciones de la estadistica. Naturalmente aparecen como razones diversas tasas e índices. Tal es el caso en el desarrollo de encuestas para hacer auditorias, en el estudio de áreas pequeñas, etc. Su uso puede trazarse en el trabajo de J. Graunt en 1662 y en la encuesta desarrollada por el célebre matemático P.S. Laplace. Ambos usaron la razón entre el número de habitantes y el de nacimientos y, a la luz del desarrollo actual, la estimaron. Laplace lo aplicó en la famosa encuesta para estimar la población de Francia en 1820 en que introdujo el uso de la selección aleatoria. Vea Cochran (1978). Una formulación general de la estimación de una razón es considerar que la variable de interés Y se relaciona con una variable auxiliar X, cuyo valor es conocido para todo elemento de la población, mediante el modelo de regresión lineal simple

$$Y=A+BX+e$$

Si $(a\ b)^T$ es el estimador mínimo cuadrático de $(A\ B)^T$ tenemos que

$$b=n(n^{-1}\sum_{i=1}^{n} Y_i-a)/\sum_{i=1}^{n} X_i$$

bajo el supuesto de que

$$\sum_{i=1}^{n} e_i=0.$$

Por lo que si A=0 se tiene que b es la razón entre las medias muestrales de Y y X.
Al considerar que los errores no tienen la misma varianza se utiliza para estimar B la suma ponderada de los errores

$$\sum_{i=1}^{n} W_i(Y_i -bX_i)^2$$

determinando el estimador

$$.b_W = \sum_{i=1}^{n} W_i\ Y_iX_i./\sum_{i=1}^{n} W_iX_i^2$$

Si $V(e^2_i)=X_i^{2t}\ \sigma^2$

[1] bouza@matcom.uh.cu

el peso adecuado es $W_i = X_i^{-2t}$ por lo que para t=0,5 y se triene que b_W es la razón entre los totales de Y y X.

En este trabajo proponemos una clase de estimadores suficientemente amplia como para que diferentes estimadores de razón desarrollados en la literatura pertenezcan a ella como casos particulares. Esto permite un estudio unificado de ellos. Por otra parte proponemos como alternativa, al clásico uso de muestreo simple aleatorio, el del muestreo por conjuntos ordenados (ranked set sampling, rss). Este fue propuesto por McIntire (1952). En este las unidades son rankeadas y se espera haya una ganancia en exactitud asociada a este método. Dell and Clutter (1972) y Takahasi and Wakimoto (1968) demostraron la validez de estas aseveraciones matemáticamente para la estimación de la media. En este trabajo desarrollamos estrategias alternativas rss para los estimadores de razón analizados.

2. UNA CLASE DE ESTIMADORES DE RAZÓN

Proponemos utilizar la clase de estimadores dada por

$$F = \left\{ \bar{y}_\theta = \frac{\bar{Y}_{est} + \alpha}{B\bar{X}_{est} + \lambda}\left(B\bar{X} + \lambda \right); \theta = (\alpha, B, \lambda)^T \in A \times B \times L \right\}$$

donde $\bar{Z}_{est}$, Z=X, Y, estima la media y

$$A = \left\{ 0, \quad b(\bar{X} - \bar{x}), \quad \sigma_x \right\} = \{\alpha_1, \alpha_2, \alpha_3\}$$
$$B = \{1, \quad B_2(x), \quad C_x, \quad \rho\} = \{B_1, B_2, B_3.B_4\}.$$
$$L = \{0, \quad \rho, \quad B_2(x), \quad C_x\} = \{\lambda_1, \lambda_2, \lambda_3.\lambda_4\}$$

Tomando

- $b=s_{xy}/s_x^2$ como el coeficiente de regresión lineal simple muestral,
- $B_2(x)$ como el coeficiente de curtosis de la distribución de X,
- C_x como es el coeficiente de variación de X
- ρ como el coeficiente de correlación lineal entre X y Y.

A esta clase pertenecen estimadores muy conocidos como el estimador clásico en el que $\theta=(0,1,0)=(\alpha_1, B_1, \lambda_1)$ pues

$$\bar{y}_r = \bar{y}_{(0,1,0)} = \frac{\bar{y}}{\bar{x}}\bar{X}$$

El error cuadrático medio de la subclase de *F* a la que pertenece este es caracterizado por

$$M(\bar{y}_{razn}) = \frac{\sigma_y^2 + R^2\sigma_x^2 w(w - 2\gamma)}{n}$$

En el caso del estimador clásico

$$w = 1$$

$$\gamma = \rho\frac{C_y}{C_x} = \rho\frac{\frac{\sigma_y}{\bar{Y}}}{\frac{\sigma_x}{\bar{X}}}$$

por lo que su error es, ver Cochran (1981) por ejemplo,

$$M(\bar{y}_n) = \frac{\sigma_y^2 + R^2\sigma_x^2 - 2\rho R\sigma_y\sigma_x}{n}$$

Singh-Taylor (2003) desarrollaron un estimador alternativo que utilizaba información poblacional sobre el coeficiente de correlación ρ. Este es de la misma famlia pues al fijar que θ=(0, 1, ρ), se obtiene:

$$\bar{y}_{ST} = \frac{\bar{y}}{\bar{x}+\rho}(\bar{X}+\rho)$$

Su error está dado por

$$M(\bar{y}_{ST}) = \left(\bar{y}_{razon} \middle| w = \frac{\bar{X}}{\bar{X}+\rho}, \qquad \gamma = 2\rho C_y / C_x \right) = \frac{\sigma_y^2 + R^2\sigma_x^2 \left(\frac{\bar{X}}{\bar{X}+\rho} \right)^2 - 2\frac{\bar{Y}\rho\sigma_y\sigma_x}{\bar{X}+\rho}}{n}$$

La comparación de estos errores permite ordenarles dentro de la clase.
La preferencia por el clásico la obtenemos al evaluar bajo que situación es válida la relación

$$M(\bar{y}_n) - M(\bar{y}_{ST}) < 0$$

En función del coeficiente de correlación esto es satisfecho cuando

$$\rho < R\sigma_x (w+1)/\sigma_y$$

Kadilar-Cingi en una serie de trabajos {Kadilar-Cingi , 2003, 2004 y 2005} han propuesto estimadores que son clasificables en *F* al denotarles en función de θ mediante $\bar{y}\ (\theta_t)$ donde

- $\theta_1 = (\alpha_2 , B_1 , \lambda_1)$
- $\theta_2 = (\alpha_2 , B_1 , \lambda_3)$
- $\theta_3 = (\alpha_2 , B_1 , \lambda_4)$
- $\theta_4 = (\alpha_2 , B_2 , \lambda_1)$
- $\theta_5 = (\alpha_2 , B_3, \lambda_3)$
- $\theta_6 = (\alpha_2 , B_1 , \lambda_2)$
- $\theta_7 = (\alpha_2 , B_3 \lambda_2)$
- $\theta_8 = (\alpha_2 , B_4 , \lambda_4)$
- $\theta_9 = (\alpha_2 , B_2 , \lambda_2)$
- $\theta_{10} = (\alpha_2 , B_4, \lambda_3)$

En forma explícita estos están dados por

$$\bar{y}(\theta_1)=\frac{\bar{y}+b(\bar{X}-\bar{x})}{\bar{x}}\bar{X},\qquad \bar{y}(\theta_2)=\frac{\bar{y}+b(\bar{X}-\bar{x})}{\bar{x}+B_2(x)}(\bar{X}+B_2(x))$$

$$\bar{y}(\theta_3)=\frac{\bar{y}+b(\bar{X}-\bar{x})}{\bar{x}+C_x}(\bar{X}+C_x),\qquad \bar{y}(\theta_4)=\frac{\bar{y}+b(\bar{X}-\bar{x})}{\bar{x}B_2(x)+C_x}(\bar{X}B_2(x)+C_x)$$

$$\bar{y}(\theta_5)=\frac{\bar{y}+b(\bar{X}-\bar{x})}{\bar{x}C_x+B_2(x)}(\bar{X}C_x+B_2(x)),\qquad \bar{y}(\theta_6)=\frac{\bar{y}+b(\bar{X}-\bar{x})}{\bar{x}+\rho}(\bar{X}+\rho)$$

$$\bar{y}(\theta_7)=\frac{\bar{y}+b(\bar{X}-\bar{x})}{\bar{x}C_x+\rho}(\bar{X}C_x+\rho),\qquad \bar{y}(\theta_8)=\frac{\bar{y}+b(\bar{X}-\bar{x})}{\bar{x}\rho+C_x}(\bar{X}\rho+C_x)$$

$$\bar{y}(\theta_9)=\frac{\bar{y}+b(\bar{X}-\bar{x})}{\bar{x}B_2(x)+\rho}(\bar{X}B_2(x)+\rho),\qquad \bar{y}(\theta_{10})=\frac{\bar{y}+b(\bar{X}-\bar{x})}{\bar{x}\rho+B_2(x)}(X\bar{\rho}+B_2(x))$$

Todos ellos son caracterizados por una misma estructura por lo que se determina una subclase en la que el error es:

$$M(\bar{y}(\theta_t))=\frac{R^2(\theta_t)\sigma_x^2+\sigma_y^2(1-\rho^2)}{n},\qquad t=1,...,10$$

Las razones son indexadas por θ_t y ellos son definidas como:

$$R(\theta_1)=R=\frac{\bar{Y}}{\bar{X}},\qquad R(\theta_2)=\frac{\bar{Y}}{\bar{X}+B_2(x)},\qquad R(\theta_3)=\frac{\bar{Y}}{\bar{X}+C_x},\qquad R(\theta_4)=\frac{\bar{Y}B_2(x)}{\bar{X}B_2(x)+C_x}$$

$$R(\theta_5)=\frac{\bar{Y}C_x}{\bar{X}C_x+B_2(x)},\qquad R(\theta_6)=\frac{\bar{Y}}{\bar{X}+\rho},\qquad R(\theta_7)=\frac{\bar{Y}C_x}{\bar{X}C_x+\rho},\qquad R(\theta_8)=\frac{\bar{Y}\rho}{\bar{X}\rho+C_x},$$

$$R(\theta_9)=\frac{\bar{Y}B_2(x)}{\bar{X}B_2(x)+\rho},\qquad R(\theta_{10})=\frac{\bar{Y}B_2(x)}{\bar{X}B_2(x)+\rho},$$

Una comparación entre el estimador clásico y los de la subclase de Kadilar-Cingi nos lleva a preferir los que sugieren estos autores si

$$\rho>\frac{\sigma_x^2\left(R^2w^2-R(\theta_t)\right)}{\sigma_y\left(\sigma_y+2\sigma_x\right)}$$

El valor absoluto del término de la derecha es con mucha frecuencia mayor que 1 lo que haría imposible su preferencia. Por otra parte su comparación con el clásico genera también su aceptación cuando se cumple que

$$\rho>\frac{\sigma_x^2\left(R^2(\theta_t)-R^2\right)}{\sigma_y\left(\sigma_y+2R\sigma_x\right)}$$

con el que podríamos hacer una valoración similar.

3 GANANCIAS EN PRECISIÓN AL USAR MUESTREO DE CONJUNTOS ORDENADOS

3.1 Elementos básicos del muestreo de conjuntos ordenados

Consideremos la selección de m muestras aleatorias independientes de tamaño m:
Rankeando dentro de cada muestra obtenemos

$$Y_{11}, Y_{12}...., Y_{1m}; Y_{21}, Y_{22}...., Y_{2m}; ....; Y_{m\ 1}, Y_{m2}...., Y_{mm}$$

$$Y_{(1:1)}, Y_{(2:1)}...., Y_{(m:1)}; Y_{(1:2)}, Y_{(2:2)}...., Y_{(n:2)}; ....; Y_{(1:m)}, Y_{(2m)}...., Y_{(mm)}$$

Donde $Y_{(j:t)}$ es el estadístico de orden (os) j de la muestra t. Son evaluados solo los elementos en la diagonal : $Y_{(t:t)}$.
Como es bien sabido si usamos muestreo simple aleatorio con reemplazo (msacr) el usual estimador de la media poblacional tiene varianza

$$V[\sum_{i=1}^{m} Y_i /m] = \sigma^2 /m$$

Pero como nuestras inferencias se hacen sobre la base de los estadísticos de orden

$$V[[\ \sum_{i=1}^{m} Y_{(i:i)} /m] = \sum_{i=1}^{m} V(Y_{(i)})/m^2 = \sum_{i=1}^{m} \sigma^2_{(i)} /m^2$$

Como el estadístico de orden generalmente se basa en el rankeo al juicio si $f_{(i)}(y)$ es la densidad de probabilidad (pdf)

$$P(y) = \sum_{i=1}^{m} f_{(i)}(y)/m$$

y

$$E[Y_{(i)}] = \sum_{t=1}^{m} Y_t f_{(y)}(y)/m = \mu_{(i)}$$

Dada la insesgadez del estimador se tiene que

$$\sum_{i=1}^{m} (\mu_{(i)} - \mu) = \sum_{i=1}^{m} \Delta_{(i)} = 0$$

Un resultado fundamental es la relación general $\sigma^2_{(i)} = \sigma^2 - \Delta^2_{(i)}$. Por lo que

$$\sum_{i=1}^{m} \sigma^2_{(i)} /m^2 = m^{-1} \sigma^2 - m^{-2} \sum_{i=1}^{m} \Delta^2_{(i)}$$

Podemos fijar como procedimiento de selección rss el siguiente

Procedure RSS1

While t<m do
Select a ssu independently from U using srswr.
Each unit in $s_{(t)}$ is ranked and the os's $Y_{(1:t)}$,...., $Y_{(r(t):t)}$ are determined.
END

Este genera la matriz

$Y_{(1:1)}$	$Y_{(2:1)}$	• • •	$Y_{(t:1)}$	• • •	$Y_{(m:1)}$
$Y_{(1:2)}$	**$Y_{(2:2)}$**	• • •	$Y_{(t:2)}$	• • •	$Y_{(m:2)}$
•	•	• • •	•	• • •	•
•	•	• • •	•	• • •	•
•	•	• • •	•	• • •	•
$Y_{(1:t)}$	$Y_{(2:t)}$	• • •	**$Y_{(t:t)}$**	• • •	$Y_{(m:t)}$
•	•	• • •	•	• • •	•
•	•	• • •	•	• • •	•
•	•	• • •	•	• • •	•
$Y_{(1:m)}$	$Y_{(2:m)}$	• • •	$Y_{(t:m)}$	• • •	**$Y_{(m:m)}$**

Se evalúan los elementos en la diagonal $s(j) = \{Y_{(i:i)}, i=1,..,m\}$.
Si este procedimiento es repetido r veces para completar una muestra de tamaño mr=n usaremos el siguiente procedimiento

Procedure RSS for a ranked set sample of size n generation

While j<r do
Procedure RSS1
End.

Entonces tendríamos $Y_{(i:i)j}$, i=1,…,m y j=1,…,r y el estimador sería:

$$\mu_{rss} = \sum_{j=1}^{r} = \sum_{i=1}^{m} Y_{(i:i)} / mr$$

y su error

$$\sum_{i=1}^{m} \sigma^2_{(i)} / rm^2 = (rm)^{-1} \sigma^2 - (rm)^{-2} \sum_{i=1}^{m} \Delta^2_{(i)}$$

3.2. Estrategias rss para los estimadores de razón

Consideremos que rankeamos usando información sobre X. Esto implica que hay un rankeo también en Y. Si la relación entre ellos es descrita por un alto valor de $|\rho|$ habrá ganancia en precisión. Al computar

$$\bar{y}_{rss} = \frac{\sum_{i=1}^{m} y_{(i:i)}}{m}$$

$$\bar{x}_{rss} = \frac{\sum_{i=1}^{m} x_{(i:i)}}{m}$$

tendríamos la alternativa rss del estimador clásico de razón

$$\bar{y}_{r-rss} = \bar{y}_{(0,1,0)-rss} = \frac{\bar{y}_{rss}}{\bar{x}_{rss}} \bar{X}$$

Utilizando la expansión de la varianza de Y bajo rss tenemos que su error cuadrático medio es:

$$M(\bar{y}_{r-rsss}) = \frac{\sigma_y^2 - \sum_{i=1}^{m} \frac{\Delta^2_{Y(i)}}{m} + R^2 \left[\sigma_x^2 - \sum_{i=1}^{m} \frac{\Delta^2_{X(i)}}{m} \right] - 2R\rho \left(\sigma_x^2 - \sum_{i=1}^{m} \frac{\Delta^2_{X(i)}}{m} \right)^{1/2} x \left(\sigma_y^2 - \sum_{i=1}^{m} \frac{\Delta^2_{Y(i)}}{m} \right)^{1/2}}{n}$$

Al compararle con el estimador clásico tenemos que el uso de rss es preferible al del msacr si

$$\delta_{r-rsss} = \frac{\sum_{i=1}^{m} \frac{\Delta^2_{Y(i)}}{m} + R^2 \sum_{i=1}^{m} \frac{\Delta^2_{Y(i)}}{m} - 2R\rho \left[\sigma_X \sigma_Y - \left(\sigma_x^2 - \sum_{i=1}^{m} \frac{\Delta^2_{X(i)}}{m} \right)^{1/2} x \left(\sigma_y^2 - \sum_{i=1}^{m} \frac{\Delta^2_{Y(i)}}{m} \right)^{1/2} \right]}{n} > 0$$

O sea si

$$\rho < \frac{\sum_{i=1}^{m} \frac{\Delta^2_{Y(i)}}{m} + R^2 \sum_{i=1}^{m} \frac{\Delta^2_{Y(i)}}{m}}{2R \left[\sigma_X \sigma_Y - \left(\sigma_x^2 - \sum_{i=1}^{m} \frac{\Delta^2_{X(i)}}{m} \right)^{1/2} x \left(\sigma_y^2 - \sum_{i=1}^{m} \frac{\Delta^2_{Y(i)}}{m} \right)^{1/2} \right]}$$

Los términos dentro de las raíces cuadradas son positivos por ser la expresión de varianzas.
Podemos expresar también esta relación mediante

$$\rho < \frac{\sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m} + R^2\sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m}}{2R\left[\sigma_X\sigma_Y\left(1-\left(1-\sum_{i=1}^{m}\frac{\Delta_{X(i)}^2}{m\sigma_x^2}\right)^{1/2}\left(1-\sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m\sigma_y^2}\right)^{1/2}\right)\right]}$$

El término a la derecha de la ecuación es positivo por lo que la existencia de una correlación negativa entre X y Y permite aseverar que el uso de rss es preferible a la del msacr.
Por otra parte aunque usemos X para rankear , como esta variable es conocida, podemos computar la media de esta en las rm^2 =mn unidades seleccionadas para confeccionar las muestra rss. Sea esta media

$$\bar{\bar{x}} = \frac{\sum_{t=1}^{r}\sum_{i=1}^{m}\sum_{j-1}^{m}x_{(i,j)t}}{rm^2}$$

Entonces podríamos usar

$$\bar{y}_{r-rss2} = \bar{y}_{(0,1,0)-rss2} = \frac{\bar{y}_{rss}}{\bar{\bar{x}}}\bar{X}$$

lo que nos permite determinar que

$$M(\bar{y}_{r-rsss2}) = \frac{\sigma_y^2 - \sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m} + R^2\frac{\sigma_x^2}{r} - 2R\rho\frac{\sigma_x}{\sqrt{r}_x}\left(\sigma_y^2 - \sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m}\right)^{1/2}}{n}$$

Al compararle con el estimador clásico de razón tenemos que el uso de rss es preferible al uso de msacr si

$$\delta_{r-rsss2} = \frac{\sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m} - 2R\rho\left[\frac{\sigma_X\sigma_Y}{\sqrt{r}} - \left(1-\sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m\sigma_Y}\right)^{1/2}\right]}{n} > 0$$

lo que simplifica el análisis de la preferencia del rss pues nos queda que

$$\rho < \frac{\sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m}}{2R\left[\frac{\sigma_X\sigma_Y}{\sqrt{r}} - \left(1-\sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m\sigma_Y}\right)^{1/2}\right]}$$

Si comparamos ambas estrategias rss tenemos que si

$M(\bar{y}_{r-rsss}) - M(\bar{y}_{r-rsss2}) > 0$

preferimos rss2, o sea cuando se cumple que

$$\delta_{rss,rss2} = \frac{R^2\sigma_X^2\left(1-\frac{1}{r}\right) - R^2\left[\sum_{i=1}^{m}\frac{\Delta_{X(i)}^2}{m}\right] - 2R\rho\sigma_X\left[\left(1-\sum_{i=1}^{m}\frac{\Delta_{X(i)}^2}{m\sigma_x}\right)^{1/2} - \frac{1}{\sqrt{r}}\right]\left(\sigma_y^2 - \sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m}\right)^{1/2}}{n}$$

es mayor que cero. Entonces por lo que en general utilizar toda la información de X no genera un incremento en la precisión,
La versión rss del estimador propuesto por Singh-Taylor (2003) es .

$$\bar{y}_{ST-rss} = \frac{\bar{y}_{rss}}{\bar{x}+\rho}(\bar{X}+\rho)$$

cuyo error cuadrático medio es expresado mediante

$$M(\bar{y}_{ST-rss}) = \frac{\sigma_y^2 - \sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m} + R^2\left(\sigma_x^2 - \sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m}\right)\left(\frac{\bar{X}}{\bar{X}+\rho}\right)^2 - 2\xi}{n}$$

donde

$$\xi = \frac{\bar{Y}\rho\left(\sigma_y^2 - \sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m}\right)^{1/2}\left(\sigma_x^2 - \sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m}\right)^{1/2}}{\bar{X}+\rho}$$

cuya comparación con el error del basado en mascr nos lleva a que la opción rss sea preferible cuando

$$\delta_{ST} - M(\bar{y}_{ST}) - M(\bar{y}_{ST-rss}) > 0$$

$$M(\bar{y}_{r-rsss}) - M(\bar{y}_{r-rsss2}) > 0$$

preferimos rss2, o sea cuando se cumple que

$$\delta_{rss,rss2} = \frac{R^2\sigma_X^2\left(1-\frac{1}{r}\right) - R^2\left[\sum_{i=1}^{m}\frac{\Delta_{X(i)}^2}{m}\right] - 2R\rho\sigma_X\left[\left(1-\sum_{i=1}^{m}\frac{\Delta_{X(i)}^2}{m\sigma_x}\right)^{1/2} - \frac{1}{\sqrt{r}}\right]\left(\sigma_y^2 - \sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m}\right)^{1/2}}{n}$$

Lo que nos lleva a que

$$\delta_{ST} = \frac{\sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m} + 2\frac{\bar{Y}\rho\sigma_y\left(1-\sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m\sigma_y^2}\right)^{1/.2}\sigma_x}{\bar{X}+\rho}}{n}$$

Asumiendo que

$$1 > \sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m\sigma_y^2}$$

esto se puede expresar como

$$\rho > -\frac{\bar{X}\sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m}}{\sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m} + 2\bar{Y}\sigma_y\left(1-\sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m\sigma_y^2}\right)^{1/.2}\sigma_x}$$

que es válido siempre también cuando X y Y están correlacionados.

Un estudio similar para los estimadores del tipo Kadilar-Cingi nos lleva a los estimadores alternativos rss al sustituir los estimadores del msa por los rss
Cuando usamos rss el rankeo nos lleva a que el error sea :

$$M(\bar{y}(\theta_{t-rss})) = \frac{R^2(\theta_{t-rss})\left[\sigma_x^2 - \sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m}\right] + \left[\sigma_y^2 - \sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m}\right](1-\rho^2)}{n}, \qquad t = 1,...,10$$

Por tanto el estimador propuesto es mejor que los de Kadilar-Cingi si

$$M(\bar{y}(\theta_t)) - M(\bar{y}(\theta_{t-rss})) > 0$$

O sea cuando

$$\delta_t = \frac{R^2(\theta_t)\sum_{i=1}^{m}\frac{\Delta_{X(i)}^2}{m}}{n} + \frac{\sum_{i=1}^{m}\frac{\Delta_{Y(i)}^2}{m}(1-\rho^2)}{n} > 0, \qquad t = 1,..,10$$

Esto es válido siempre que $\rho \neq 1$. De ahí que, en términos de la correlación, prefiramos la estrategia rss al cumplirse que

$$\rho_t < \sqrt{\frac{R^2(\theta_t)\sum_{i=1}^{m}\Delta_{X(i)}^2 + \sum_{i=1}^{m}\Delta_{Y(i)}^2}{\sum_{i=1}^{m}\Delta_{Y(i)}^2}}$$

Indexando por t el caso en que el estimador es θ_t

4. ESTUDIO NUMÉRICO

No es posible establecer condiciones generales para valorar que estrategia es mejor aunque hay evidencias de que las alternativas rss son mejores que las basadas en msacr. Para establecer cuales funcionan más adecuadamente se desarrollaron varios experimentos usando datos reales. Analizaremos las siguientes bases de datos

B1 Ventas en supermercados: Usando datos brindados por Castro (1997) analizamos los datos de las ventas de dos supermercados.
B2. Infestación: Se utilizaron los resultados reportados por Bouza-Schubert (2003) en los niveles de infestación de campos de caña de azúcar. Y es el número de adultos de la plaga y X el de huevos.
B3. . Mejoría en el tratamiento de la soriasis: Una investigación desarrollada por Viada et.al. (2004) sobre el área afectada por la soriasis en pacientes bajo tratamiento fue analizada. Y fue el área afectada en la evaluación 2 y X el de la evaluación inicial.
B4. Análisis de sangre: la evaluación de hemoglobina fue Y y X el de WBC desarrollada en Xalapa en 1998.

Estos datos fueron usados para determinar el error de muestreo de cada estimador en la estrategia rss y la msa. La precisión relativa RP=MSE (rss)/MSE(msa) fue computada para cada alternativa. Los resultados se presentan en la Tabla 1. En ella se considero r=5 y m=4.
Los resultados de la misma sugieren que es más recomendable el uso de rss excepto para θ_6 y θ_7 en el estudio de la composición de la sangre. En este la distribución de ambas variables es normal. La mayor ganancia fue la del estimador de Singh-Taylor seguido por la del estimador clásico.
A partir de los resultados con los datos reales se consideró necesario valorar el papel de distribuciones paramétricas. Consideramos las distribuciones normal (N(0,1), para Y y N(1,1) para X, así como la exponencial con parámetro 1, Exp(1) y la uniforme en (0,1), U(0,1) para Y y X. El caso de la normal estándar en ambas variables generaría razones con valor indefinido. Los momentos de los estadísticos de orden fueron calculados utilizando la usual aproximación a partir de Series de Taylor, ver Droesbeke-Fine (1996).

Tabla 1. Precisión relativa de estrategias rss versus msa

Estimador	Ventas en supermercados	Infestación en campos	Área afectada	Evaluación de composición de sangre
Clásico	0.56	0.66	0.60	0.89
Singh-Taylor	0.35	0.55	0.59	0.88
θ_1	0.84	0.81	0.68	0.95
θ_2	0.66	0.73	0.61	0.92
θ_3	0.92	0.83	0.72	0.77
θ_4	0.97	0.72	0.61	0.85
θ_5	0.98	0.84	0.60	0.98
θ_6	0.89	0.87	0.55	1.07
θ_7	0.80	0.68	0.64	1.09
θ_8	0.84	0.75	0.66	0.97
θ_9	0.81	0.68	0.62	0.91
θ_{10}	0.79	0.77	0.7-	0.91

Se utilizó la correspondiente distribución conjunta con valores del coeficiente de correlación $\rho \in \{-0.9, -0.5, -0.1, 0.1, 0.5, 0.9\}$. Cada uno de los 12 estimadores fue comparados para las correspondientes distribuciones en las Tablas subsiguientes.

Tabla 2. Precisión relativa de estrategias rss versus msa al usar el estimador de : razón clásico

Distribución	ρ=-0.9	ρ=-0.5	ρ=-0.1	ρ=0.1	ρ=0.5	ρ=0.9
Normales	0.33	0.46	0.52	0.67	0.71	0.82
E(1)	0.21	0.30	0.34	0.39	0.43	0.48
U(0,1)	0.46	0.49	0.51	0.54	0.55	0.59

Los resultados para el criterio clásico establecen (Tabla 2) que la mayor ganancia está presente cuando la distribución es la exponencial y para valores negativos de ρ. Esto último fue un resultado deducido anteriormente para establecer la preferencia del rss.

Tabla 3. Precisión relativa de estrategias rss versus msa al usar el estimador de : Singh-Taylor

Distribucion	ρ=-0.9	ρ=-0.5	ρ=-0.1	ρ=0.1	ρ=0.5	ρ=0.9
Normales	0.43	0.43	0.43	0.43	0.43	0.43
E(1)	0.28	0.23	0.24	0.18	0.13	0.11
U(0,1)	0.74	0.72	0.69	0.62	0.68	0.72

Los resultados para el estimador de Singh-Taylor (2003) en la Tabla 3 reflejan el hecho de que, en el caso en que la esperanza de la variable es cero, el coeficiente de correlación no interviene en la expresión de la ganancia en precisión. Tal es el caso de la distribución N(0,1). Este si lo hace en lo otros dos casos. El resultado teórico obtenido establece que si ρ es no negativo siempre rss es mejor. Esto se puede valorar analizando las ganancias para correlaciones mayores que cero. Nuevamente es la exponencial la distribución más favorable para el uso de rss.

La tabla 4 brinda los resultados de la comparación de las estrategias alternativas msa y la rss para los estimadores del tipo Kandilar-Cingi. En todos los casos lo más recomendable es usar rss.

Es de apuntar que se pueden hacer tablas para diversas combinaciones de m y r haciendo los cálculos requeridos de los momentos de los estadísticos de orden necesarios. Esto es un problema numérico relativamente simple y que permite valorar con que estimador se obtendría un mejor resultado valorando los valores posibles de la correlación y la distribución. Como en la práctica se recomienda usar m pequeño y r grande para satisfacer que rm=n el número de posibles tablas no tiene que ser necesariamente grande. En el caso de as distribuciones estudiadas es relativamente fácil realizar tales cálculos.

Tabla 4. Precisión relativa de estrategias rss versus msa al usar estimadores de Kadilar-Cingi θ_1

Distribución	ρ=-0.9	ρ=-0.5	ρ=-0.1	ρ=0.1	ρ=0.5	ρ=0.9
Normales	0.53	0.45	0.40	0.40	0.45	0.53
E(1)	0.41	0.38	0.27	0.27	0.38	0.41
U(0,1)	0.69	0.55	0.49	0.49	0.55	0.69
			θ_2			
Normales	0.51	0.46	0.38	0.38	0.46	0.51
E(1)	0.43	0.39	0.24	0.24	0.39	0.43
U(0,1)	0.68	0.57	0.51	0.51	0.57	0.68
			θ_3			
Normales	0.55	0.49	0.40	0.40	0.49	0.55
E(1)	0.40	0.38	0.27	0.27	0.38	0.40
U(0,1)	0.66	0.58	0.52	0.52	0.58	0.66
			θ_4			
Normales	0.47	0.40	0.38	0.38	0.40	0.47
E(1)	0.41	0.35	0.30	0.30	0.35	0.41
U(0,1)	0.78	0.74	0.71	0.71	0.74	0.78
			θ_5			
Normales	0.56	0.52	0.48	0.48	0.52	0.56
E(1)	0.39	0.36	0.33	0.33	0.36	0.39
U(0,1)	0.67	0.64	0.61	0.61	0.64	0.67
			θ_6			
Normales	0.55	0.49	0.40	0.40	0.49	0.55
E(1)	0.40	0.38	0.27	0.27	0.38	0.40
U(0,1)	0.66	0.58	0.52	0.52	0.58	0.66
			θ_7			
Normales	0.55	0.45	0.40	0.40	0.45	0.55
E(1)	0.40	0.38	0.35	0.35	0.38	0.40
U(0,1)	0.66	0.58	0.53	0.53	0.58	0.66
			θ_8			
Normales	0.66	0.49	0.40	0.40	0.49	0.66
E(1)	0.40	0.37	0.27	0.27	0.37	0.40
U(0,1)	0.69	0.67	0.62	0.62	0.67	0.69
			θ_9			
Normales	0.48	0.40	0.36	0.36	0.40	0.48
E(1)	0.41	0.35	0.30	0.30	0.35	0.41
U(0,1)	0.58	0.54	0.51	0.51	0.54	0.58
			θ_{10}			
Normales	0.57	0.50	0.38	0.38	0.50	0.57
E(1)	0.51	0.35	0.30	0.30	0.35	0.51
U(0,1)	0.68	0.65	0.60	0.60	0.65	0.68

Agradecimientos : Expresamos nuestro reconocimiento a dos referees anónimos los que recomendaron mejoras que e implementaron en los experimentos numéricos desarrollados y en la redacción del trabajo. Este trabajo fue soportado parcialmente por el proyecto CAPES/MES-CUBA-PROJETOS CGPR No. Edital 046/2013 y al PNCB del CITMA

RECEIVED AUGUST, 2014
REVISED NOVEMBER, 2014

REFERENCIAS

[1] BOUZA, C.N. and L. SCHUBERT (2003): The estimation of biodiversity and the characterization of the dynamics : an application to the study of a pest. **Rev. Mat. E Stat.** 21, 85-98.

[2] BOUZA, C. (2000): Model assisted ranked set sampling. **Biometrical J.** 42, 9-19.
[3] CASTRO, C. (1997): Introducción al análisis exploratorio en la investigación de Mercado. Universidad Veracruzana, **Reporte Técnico PFOMES9631113.**
[4] COCHRAN, W. G. (1977): **Sampling Techniques**. Wiley and Sons, N. York.
[5] COCHRAN , W. G. (1978): **Laplace's ratio estimator**. En "Contributions to Survey Sampling and Applied Statistics", H.A. David (editor). Academic Press, N. York.
[6] DELL, T.R. and CLUTTER, J.L. (1972): Ranked set sampling theory with order statistics. Background. **Biometrics** 28, 545-55.
[7] DROESBEKE, J.J. et J. FINE (1996**): Inferénce non Paramétrique**. **Les statitiques de rangs**. Edition Ellipses, Bruxelles.
[8] KADILAR, C. and CINGI, H. (2004): Ratio estimators in simple random sampling. **Applied Math. And Computation**.151, 893-902.
[9] KADILAR, C. and CINGI, H . (2005): A new ratio estimator in stratified random sampling. **Comm. in Stat:. Theory and Methods**. 34, 597-602.
[10] MCINTYRE, G.A. (1952): A method of unbiased selective sampling using ranked sets. **Australian J. Agricultural Research.** 3, 385-390.
[11] PATIL, G.P (2002): **Ranked set sampling. In Encyclopedia of Enviromentrics**. (A.H. El-Shaarawi and W.W. Pieegoshed, (Editors). Vol. 3,1684-1690. Wiley, Chichester.
[12] SINGH, H.P. and TAYLOR, R. (2003): Use of known correlation coefficient in estimating finite population means **Statistics in Transition**. 6, 555-560.
[13] TAKAHASI K. and WAKIMOTO, K. (1968): On unbiased estimates of the population mean based on sample stratified by means of ordering. **Annals of the Inst. of Statistical Mathematics.** 20, 1-3.
[14] VIADA GONZÁLEZ, CARMEN E., CARLOS N. BOUZA HERRERA , FRANZ TORRES BARBOSA , y OLGA TORRES GEMEIL (2004): Estudio estadístico de ensayos clínicos de un medicamento para la Psoriasis Vulgar usando técnicas de imputación, **Inv. Operacional**, 25 243-255