Metodología de Encuestas
Vol 3, Núm 2, 2001, 183-195

SOBRE LA ESTIMACIÓN DE LA RAZÓN DE DOS MEDIAS EN EL MUESTREO EN OCASIONES SUCESIVAS

Eva María Artés Rodríguez
Amelia-V García Luengo
Universidad de Almería

RESUMEN

En este trabajo se desarrolla la teoría de muestreo sucesivo, para construir el estimador óptimo de la razón de dos medias poblacionales en la ocasión actual, considerando para la parte apareada de la muestra un estimador de doble muestreo. Este estimador utiliza la información proporcionada por una variable auxiliar de la primera ocasión. Se obtienen las expresiones para la fracción de apareamiento óptimo, para el estimador combinado, junto con su error, y la curva que nos da la ganancia en precisión del estimador propuesto sobre el estimador simple que no utiliza la información de la primera ocasión. Se compara dicho estimador con otros estimadores y se dan condiciones de eficiencia. Finalmente, se completa el trabajo con un estudio empírico.

Palabras clave:.muestreo sucesivo, estimador, de la razón poblacional, fracción de apareamiento óptimo, ganancia en precisión.

Introducción

En muchas encuestas, la información se recoge periódicamente procedente de la misma población. Existen tres posibles procedimientos de muestreo para llevar a cabo dichas encuestas continuas:

1. Extraer una nueva muestra en todas las ocasiones (muestreo *repetido*).
2. Utilizar la misma muestra en todas las ocasiones (muestreo *panel*).
3. Realizar un reemplazamiento parcial de unidades de una ocasión a otra (muestreo en *ocasiones sucesivas*, o también llamado muestreo *rotativo* cuando los elementos tienen restringido el número de etapas en las que van a formar parte de la muestra, como es caso de la Encuesta de Población Activa, de periodicidad trimestral, y de la mayoría de las encuestas familiares elaboradas por la misma entidad, el Instituto Nacional de Estadística en España).

Esta última posibilidad ha sido estudiada con profundidad, para estimar la media de la ocasión actual (Rao y Mudholkar, 1967; Artés, 1999). En muchas ocasiones prácticas, la estimación de la razón poblacional de dos caracteres, en la ocasión actual, resulta de considerable interés. Por ejemplo, en una encuesta familiar podríamos estar interesados en estimar la razón entre la media de ingresos familiares y la media de gastos de alimentación, o también el total de gastos de alimentación y el número de personas en la familia.

Así, en este trabajo se va a desarrollar la teoría en muestreo sucesivo, para construir el estimador óptimo de la razón de dos medias en la segunda ocasión, considerando la información proporcionada por una variable auxiliar en la primera ocasión, basándonos en el estimador de doble muestreo propuesto por Khare (1991).

Sea
$$R_2 = \frac{\overline{Y}_1}{\overline{Y}_2}$$

la razón de las medias poblacionales y deseamos estimar R_2. La estimación convencional de R_2 es

$$\hat{R}_2 = \frac{\overline{y}_1}{\overline{y}_2}$$

donde $\overline{y}_1$ y $\overline{y}_2$ son las medias muestrales de y_1 e y_2.

Teoría

Sean Y_{1j}, Y_{2j} y X_{1j} los valores de la j-ésima unidad de las características y_1, y_2 y x_1, que asumimos como no negativas. Las medias poblacionales de las características principales y_1, y_2 y la característica auxiliar x_1, se denotan por $\overline{Y}_1$, $\overline{Y}_2$ y $\overline{X}_1$, respectivamente. En el caso en que $\overline{X}_1$ sea desconocida se usa la técnica de doble muestreo. De esta manera, $\bar{x}'_1$ denota la media muestral de la característica auxiliar

x_1, basada en la primera fase muestral de tamaño *n*', e $\bar{y}_1$, $\bar{y}_2$ y $\bar{x}_1$ denotan las medias muestrales de tamaño *n* de las características y_1, y_2 y x_1, respectivamente.

Supongamos que las muestras son de tamaño *n* en ambas ocasiones, que se utiliza muestreo aleatorio simple y que el tamaño de la población *N* es suficientemente grande como para poder ignorar el factor de corrección por finitud.

Sea una muestra de tamaño *n* seleccionada en la primera ocasión de una población de tamaño *N*, en donde medimos las características x_1 en la primera ocasión e y_1, y_2 en la segunda ocasión. Sea una muestra aleatoria simple de tamaño $m = pn = n\text{-}u$ $(0 < p < 1)$ submuestreada de las *n* unidades, que se retiene para la segunda ocasión (muestra apareada), y las restantes $u = qn = n\text{-}m$ $(q = 1 - p)$ unidades son reemplazadas por una nueva selección del universo $N - m$ que resulta después de omitir las *m* unidades.

Denotamos por

$$C_i^{\,2} = \frac{S_i^2}{\overline{Y}_{2i}}, \quad S_i^2 = \frac{\Sigma_{j=1}^{N}\left(Y_{ij} - \overline{Y}\right)^2}{N-1}, \quad \text{i} = 1,2$$

$$C_0^2 = \frac{S_0^2}{\overline{X}_1^2} \quad S_0^2 = \frac{\Sigma_{j=1}^{N}\left(X_{ij} - \overline{X}_1\right)^2}{N-1}$$

Los coeficientes de correlación entre (y_1, y_2), (y_1, x_1) y (y_2, x_1) se denotan por ρ_0, ρ_1 y ρ_2, respectivamente.

Así mismo, se denota por

$R_2 = \dfrac{\overline{Y}_1}{\overline{Y}_2}$ la razón de las medias poblacionales en la segunda ocasión.

$\hat{R}_2 = \dfrac{\bar{y}_1}{\bar{y}_2}$ el estimador de la razón de las medias poblacionales en la segunda ocasión.

$\hat{R}_{1m}\left(\hat{R}_{2m}\right)$ el estimador de la razón de las medias poblacionales en la primera (segunda) ocasión, correspondiente a las unidades que son comunes en las dos ocasiones (apareadas).

$\hat{R}_{1u}\left(\hat{R}_{2u}\right)$ el estimador de la razón de las medias poblacionales en la primera (segunda) ocasión, correspondiente a las unidades no apareadas.

Construimos el estimador de la razón poblacional en la ocasión actual, combinando dos estimadores independientes de la razón $\hat{R}'_{2m}$ y $\hat{R}_{2u}$, mediante una media aritmética ponderada, con pesos *Q* y 1-*Q*, inversamente proporcionales a las varianzas de los estimadores independientes y suma igual a la unidad:

$$\hat{R}'_2 = Q\hat{R}_{2u} + (1-Q)\hat{R}'_{2m}$$

Para la parte no apareada, utilizamos un estimador simple de la razón, basado en las *u* unidades de la segunda ocasión.

Para la parte apareada, utilizamos un estimador mejor que el directo, basado sólo en las *m* unidades muestrales de la segunda ocasión. Proponemos un estimador indirecto de doble muestreo (Khare, 1991), dado por:

$$\hat{R}'_{2m} = \int(v,u) \quad \text{siendo} \quad v = \frac{\bar{y}_1}{\bar{y}_2}, \quad u = \frac{\bar{x}'_1}{\bar{x}_1}$$

La función $f(v,u)$ es tal que

$$f(R_2,1) = R_2, \qquad f_1(R_2,1) = 1$$

satisfaciendo también las siguientes condiciones:

1. Cualquiera que sea la muestra escogida, (v,u) son valores que están definidos en el subconjunto convexo cerrado D_1, del espacio real bidimensional que contiene el punto $(R_2,1)$.
2. La función $f(v,u)$, y sus derivadas parciales de primer y segundo orden, existen, son continuas y definidas en D_1.

En estas condiciones y con la ayuda del desarrollo en serie de Taylor (Khare, 1991) alrededor del punto $(R_2,1)$, llegamos a la expresión del sesgo (B):

$$\mathrm{B}\left(\hat{R}'_{2m}\right) = R_2 \frac{f}{m}\left(C_2^2 - \rho_0 C_1 C_2\right) + \left(\frac{f}{m} - \frac{f'}{n}\right) C_0^2 \left(f_2(R_2,1) + \frac{1}{2} f_{22}(R_2,1)\right) +$$
$$+ \frac{1}{2}\left(R_2^2 \frac{f}{m} V_1 f_{11}(R_2,1) - 2R_2\left(\frac{f}{m} - \frac{f'}{n}\right) V_2 C_0 f_{12}(R_2,1)\right)$$

donde

$$f_1(v,u) = \frac{\partial}{\partial_v} f(v,u); \quad f_2(v,u) = \frac{\partial}{\partial_u} \int(v,u); \quad f_{12}(v,u) = \frac{\partial^2}{\partial u \partial v} f(v,u)$$

$$f_{11}(v,u) = \frac{\partial^2}{\partial_v{}^2} f(v,u); \quad f_{22}(v,u) = \frac{\partial^2}{\partial_u{}^2} f(v,u)$$

$$V_1 = C_1^2 + C_2^2 - 2\rho_o C_1 C_2; \quad V_2 = \rho_1 C_1 - \rho_2 C_2$$

$$f = \frac{N-m}{N}; \quad f' = \frac{N-n}{N}$$

De esta manera, el mínimo error cuadrático medio (ECM), ignorando el factor de corrección por finitud, viene dado por

$$ECM\left(\hat{R}'_{2m}\right)=R_2^2\left[\frac{1}{m}\left(C_1^2+C_2^2-2\rho_0 C_1 C_2\right)-\left(\frac{1}{m}-\frac{1}{n}\right)\left(\rho_1 C_1-\rho_2 C_2\right)^2\right]$$

Siguiendo a Cochran (1977) y asumiendo una población infinita, la aproximación de primer orden, del error catedrático medio es

$$V_{\min}\left(\hat{R}'_2\right)=\frac{V\left(\hat{R}'_{2m}\right)V\left(\hat{R}_{2u}\right)}{V\left(\hat{R}'_{2m}\right)+V\left(\hat{R}_{2u}\right)}=\frac{R_2^2}{n}\left(C_1^2+C_2^2-2\rho_0 C_1 C_2\right)\frac{1-qZ}{1-q^2 Z} \qquad (1)$$

donde

$$Z=\frac{\left(\rho_1 C_1-\rho_2 C_2\right)^2}{C_1^2+C_2^2-2\rho_0 C_1 C_2}$$

Teniendo en cuenta que

$$V\left(\hat{R}_{2u}\right)=\frac{R_2^2}{u}\left(C_1^2+C_2^2-2p_0 C_1 C_2\right)$$

Se obtiene el mejor estimador de R_2 en la segunda ocasión utilizando el valor de Q que minimice $V\left(\hat{R}'_2\right)$,

$$Q_{opt}=\frac{V\left(\hat{R}_{2u}\right)}{V\left(\hat{R}_{2u}\right)+V\left(\hat{R}'_{2m}\right)}=\frac{p}{1-(1-p)^2 Z}$$

Operando adecuadamente

$$V\left(\hat{R}'_{2m}\right)=\frac{R_2^2}{m}\left(C_1^2+C_2^2-2\rho_0 C_1 C_2\right)\left[1-qZ\right]$$

con

$$Z=\frac{\left(\rho_1 C_1-\rho_2 C_2\right)^2}{C_1^2+C_2^2-2\rho_0 C_1 C_2}$$

Puesto que el estimador directo basado en las *m* unidades tiene varianza

$$\frac{R_2^2}{m}\left(C_1^2 + C_2^2 - 2p_0C_1C_2\right)$$

Se obtiene que $\hat{R}_{2m}$ es más preciso que el directo si

$$Z = \frac{(\rho_1 C_1 - \rho_2 C_2)^2}{C_1^2 + C_2^2 - 2\rho_0 C_1 C_2} \geq 0$$

Si suponemos que las muestras son de tamaños distintos, en ambas ocasiones, la varianza tiene la expresión

$$V_{min}\left(\hat{R}'_2\right) = \frac{R_2^2}{n}\left(C_1^2 + C_2^2 - 2\rho_0 C_1 C_2\right)\frac{\theta - (\theta - p)Z}{\theta - Z(1-p)(\theta - p)}$$

donde $\theta = n / n'$.

El valor óptimo de *u* se obtine minimizando en (1) con respecto a la variaciòn en u, y viene dado por

$$\left(\frac{u}{n}\right)_{opt} = \frac{1-\sqrt{1-Z}}{Z}$$

o lo que es lo mismo, la fracción del apareamiento óptimo vale

$$p_{opt} = \frac{Z-1+\sqrt{1-Z}}{Z}$$

Comparación de estimadores

Estimador simple y estimador combinado

Si se considera el estimador usual de la razón de las medias en la segunda ocasión, $\hat{R}_2$,basado sólo en las *n* unidades muestrales de dicha ocasión, y que no utiliza ninguna información adicional, su varianza toma la siguiente expresión

$$V\left(\hat{R}_2\right) = \frac{R_2^2}{n}\left(C_1^2 + C_2^2 - 2\rho_0 C_1 C_2\right)$$

Así, podemos comparar este método de estimación clásica con aquél que emplea, en la fase de estimación, la información auxiliar disponible. Para ello, podemos

obtener la ganancia en precisión, *G*, del estimador combinado, $\hat{R}'_2$, que utiliza un estimador de la razón de las medias de doble muestreo para la parte apreada de la muestra de la segunda ocasión, sobre el estimador simple, $\hat{R}_2$ mediante la siguiente expresión

$$G = \frac{V\left(\hat{R}_2\right) - V_{\min}\left(\hat{R}_2\right)}{V_{\min}\left(\hat{R}_2\right)} = \frac{Zp(1-p)}{1-(1-p)Z}$$

donde

$$Z = \frac{\left(\rho_1 C_1 - \rho_2 C_2\right)^2}{C_1^2 + C_2^2 - 2\rho_0 C_1 C_2}$$

Por definición $p \leq 1$. Si $p = 1$ (apareamiento total) ó $p = 0$ (reemplazo total), la ganancia vale cero. Para cualquier otro valor de p ($0 < p < 1$), obtendremos una ganancia positiva si $Z \geq 0$.

Caso especial

Para el caso especial

$$\rho_1 = \rho = -\rho_2 ; \quad C_1 = C_2 = C$$

tenemos que

$$V\left(\hat{R}'_{2m}\right) = \frac{R_2^2}{m} 2C^2(1-\rho_0)\left[1 - q\frac{2\rho^2}{1-\rho_0}\right]$$

y por tanto

$$V_{min}\left(\hat{R}'_2\right) = \frac{R_2^2}{n} 2C^2(1-\rho_0)\frac{1-qZ}{1-q^2Z}$$

donde

$$Z = \frac{2\rho^2}{1-\rho_0}$$

El valor óptimo para *u*, en este caso, viene dado por

$$\frac{u}{n} = \frac{1 - \sqrt{1 - \dfrac{2\rho^2}{1-\rho_0}}}{\dfrac{2\rho^2}{1-\rho_0}}$$

La figura 1 muestra el óptimo de la parte común, que disminuye cuando la correlación entre la variable auxiliar y la principal aumenta.

La ganancia en precisión del estimador combinado, $\hat{R}'_2$,sobre el estimador directo, $\hat{R}_2$, se puede obtener mediante la expresión

$$G_1 = \frac{p(1-p)\left(\frac{2\rho^2}{1-\rho_0}\right)}{1-(1-p)\left(\frac{2\rho^2}{1-\rho_0}\right)}$$

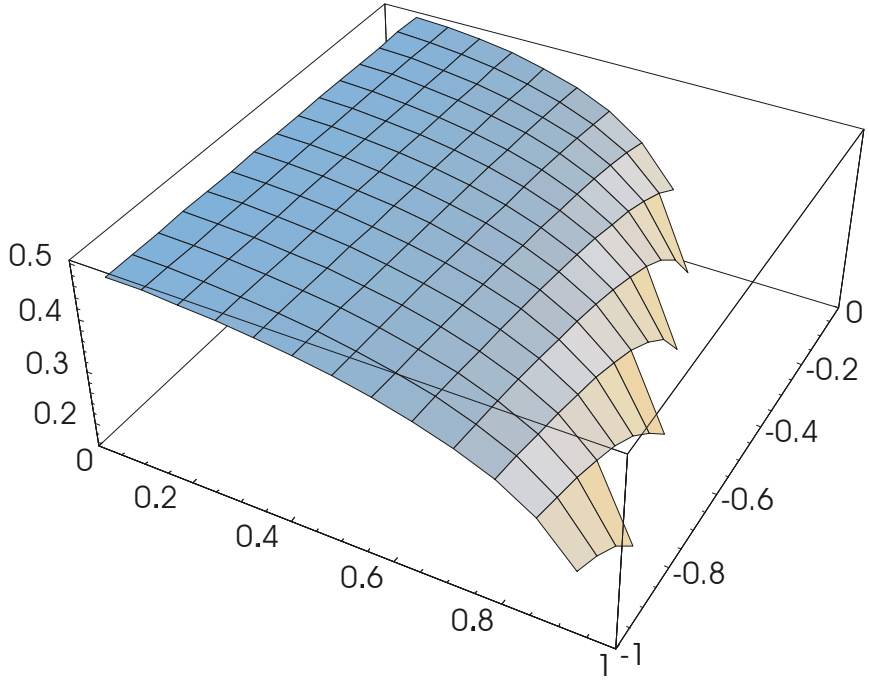

Figura 1: *fracción óptima de apareamiento.*

A partir de esta expresión deducimos que, para cualquier valor de p ($0 < p < 1$), el estimador combinado que utiliza un estimador de la razón de las medias para la parte apareada de la muestra, $\hat{R}'_2$, es más preciso que el estimador usual, $\hat{R}_2$, siempre que

$$\frac{2\rho^2}{1-\rho_0} \geq 0$$

Por tanto, se puede concluir que la ganancia en precisión de $\hat{R}'_2$ sobre $\hat{R}_2$ será óptima cuanto mayor sea la dependencia entre la variable x_1 con las variables objeto del estudio y_1 e y_2 (ρ crece) y, al mismo tiempo, cuanto más incorreladas estén y_1 e y_2 entre sí (ρ_0 decrece), como se muestra en la figura 2.

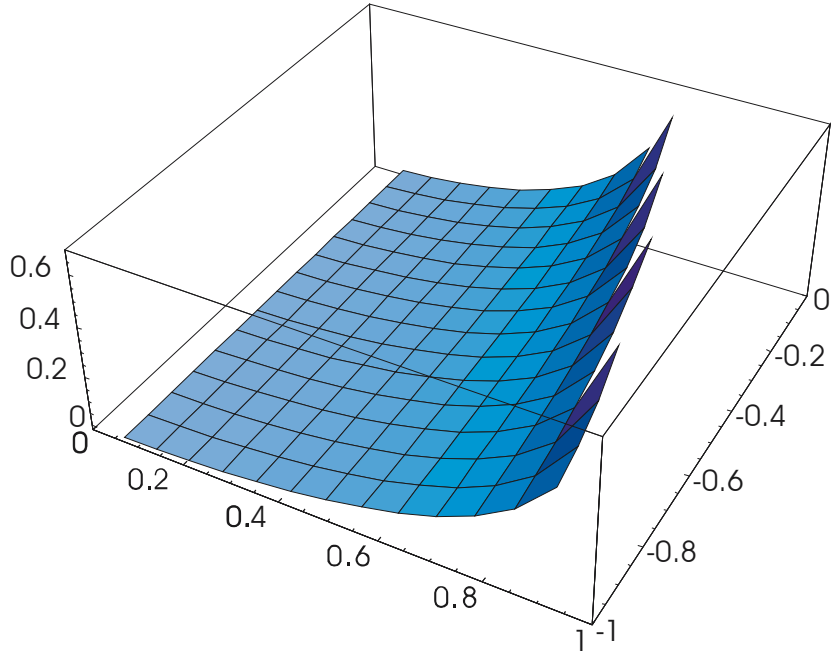

Figura 2: *ganancia en precisión.*

La tabla 1 indica el porcentaje de ganancia en precisión, G_1, del estimador $\hat{R}'_2$ sobre el estimador de la razón de las medias, que sólo utiliza la información sobre la ocasión actual.

Tabla 1: *porcentaje de ganancia en precisión, G_1.*

		ρ = 0,3			ρ = 0,5			ρ = 0,9		
	q =	0,3	0,5	0,7	0,3	0,5	0,7	0,3	0,5	0,7
ρ_0 =	-0,3	3,1	3,7	3,3	12,3	16,6	16,1	25,8	40,0	51,0
	-0,6	2,3	2,9	2,5	9,5	12,5	11,6	22,1	33,0	38,0
	-0,9	2,1	2,6	2,2	6,9	8,8	7,9	18,6	26,9	28,8

Comparación con otras estrategias

Se ha estudiado la precisión de los estimadores indirectos de Okafor (1992), $\hat{R}_2(i)$, i=1,2,3,4,5, y $\hat{R}'_2$ a partir de sus varianzas.

Así, se verifica que

$$V\left(\hat{R}_2(i)\right) - V_{min}\left(\hat{R}'_2\right) \geq 0$$

siempre que

$$DF + A_i E \geq 0 \quad \text{para} \quad i = 1,2,3,4,5,$$

siendo

$$E = 1 \qquad F = \frac{2\rho^2}{1-\rho_0} \qquad D = 1-\rho_0$$

y A_i, i = 0 1,2,3,4,5

Para cada uno de los estimadores, ocurre:

- Estimación de una razón tipo *razón* $\hat{R}_2(1)$

$$A_1 = 1-\rho_0+2\rho$$

- Estimación de una razón tipo *producto* $\hat{R}_2(2)$

$$A_2 = 1-3\rho_0+2\rho$$

- Estimación de una razón tipo *diferencia* $\hat{R}_2(3)$

$$A_3 = \rho\left(2\rho_0 - \rho - \rho\rho_0\right)$$

- Estimación de una razón tipo *diferencia con razón* $\hat{R}_2(4)$

$$A_4 = \frac{1}{2}\left(1-\rho^2\right) - \rho + \rho_0$$

- Estimación de una razón tipo *diferencia con producto* $\hat{R}_2$ (5)

$$A_5 = \frac{1}{2}\left(1-\rho^2+2\rho\right)+\rho_0\left(2\rho-1\right)$$

Gráficos de la ganancia en precisión y del apareamiento óptimo

Para el caso especial

$$\rho_1 = -\rho_2 = -\rho_0 = \rho\ ; \quad C_1 = C_2 = C$$

tenemos que

$$V\left(\hat{R}'_{2m}\right) = \frac{R_2^2}{m}\,2C^2\left(1+\rho\right)\left[1-q\,\frac{2\rho^2}{1-\rho}\right]$$

y por tanto

$$V_{\min}\left(\hat{R}'_2\right) = \frac{R_2^2}{n}\,2C^2\left(1+\rho\right)\frac{1-qZ}{1-q^2Z}$$

donde

$$Z = \frac{2\rho^2}{1+\rho}$$

Las figuras 3 y 4 muestran el porcentaje óptimo de la muestra en la parte común y la ganancia en precisión del estimador propuesto respecto a un estimador sin parte común, para distintos valores del coeficiente de correlación.

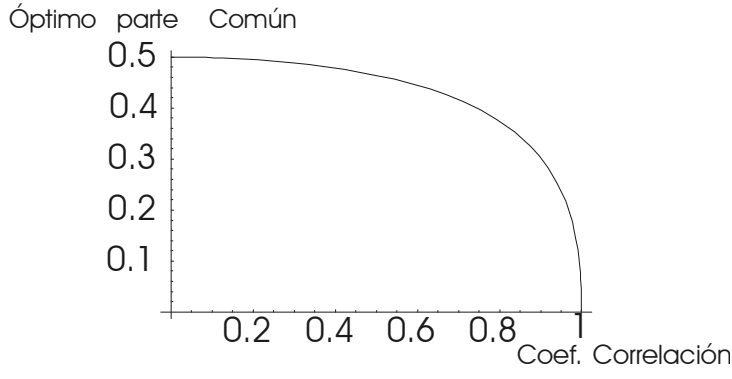

Figura 3: *fracción óptima de apareamiento cuando* $\rho_1 = -\rho_2 = -\rho_0 = \rho$.

En la figura 3 puede observarse que el mejor porcentaje en parte común no excede del 50 por 100 y decrece rápidamente cuando ρ toma valores próximos a la unidad.

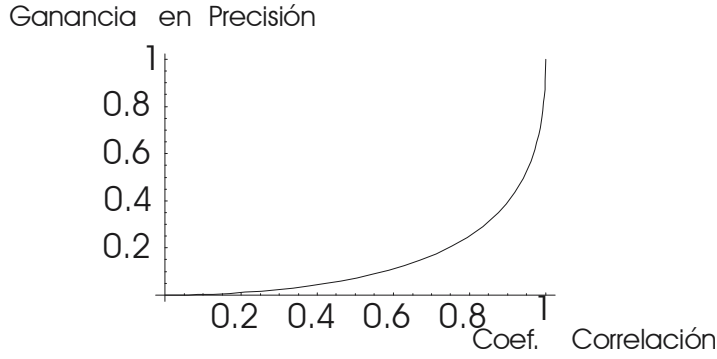

Figura 4: *ganancia en precisión cuando* $\rho_1 = -\rho_2 = -\rho_0 = \rho$.

En la figura 4 se observa que la mayor ganancia en precisión es total cuando $\rho = 1$. A menos que el coeficiente de correlación sea grande, la ganancia será modesta aunque apreciable.

Estudio empírico

Se han utilizado los datos recogidos en una investigación sobre hábitos saludables y nivel de condición física (Casimiro, 1999).

Con el objetivo de proporcionar estimadores más precisos de las variables estudiadas, se ha desarrollado un plan de muestreo basado en el principio del muestreo *sucesivo* de la misma población. Ha consistido en dos conjuntos de muestras aleatorias independientes: 1) una muestra de 135 escolares seleccionados en la primera ocasión (Abril de 1998), entre los 2.681 escolares que formaban la población, y 2) una segunda muestra de 202 escolares, seleccionada en la segunda ocasión (Junio de 1998), entre los 2.546 escolares que no formaron parte de la muestra apareada.

Hemos considerado la variable auxiliar *componente endomorfo* (x_1) en la primera ocasión, tomando como variables el *índice de masa corporal* (y_1) y *el volumen máximo de oxígeno* (y_2) de la segunda ocasión.

Los datos muestrales sobre el número de escolares y parámetros obtenidos en las dos ocasiones han sido los siguientes:

Primera ocasión: (Abril 1998), gran muestra $n = 337$.
Segunda ocasión: (Junio 1998), muestra apareada $m = 135$, muestra no apareada $u = 202$.

Los datos muestrales sobre el número de escolares y parámetros obtenidos en las dos ocasiones han sido los siguientes:

$$s_0 = 1{,}54 \qquad \bar{x}_1 = 3{,}67 \qquad \hat{\rho}_1 = 0{,}71$$

$$s_1 = 3{,}71 \qquad \bar{y}_1 = 21{,}37 \qquad \hat{\rho}_0 = -0{,}20$$
$$s_2 = 6{,}87 \qquad \bar{y}_2 = 39{,}4 \qquad \hat{\rho}_2 = -0{,}56$$

A partir de los datos, obtenemos que:

$$\hat{V}_{min}\left(\hat{R}'_2\right) = 0{,}78\frac{R_2^2}{n}\left(C_1^2 + C_2^2 - 2\rho_0 C_1 C_2\right) < \frac{R_2^2}{n}\left(C_1^2 + C_2^2 - 2\rho_0 C_1 C_2\right) = \hat{V}\left(\hat{R}_2\right)$$

lo que supone un 28,20% de ganancia en precisión del estimador propuesto sobre el estimador usual.

Se ha calculado también la fracción del apareamiento óptimo:

$$\hat{p}_{opt} = 36{,}41\%$$

Referencias

Artés, E.; Rueda, M. y Arcos, A. (1999) Aportaciones al muestreo sucesivo, *Metodología de Encuestas*, 1 (1) 19-28.

Casimiro, A.J. (1999) *Comparación, evolución y relación de hábitos saludables y nivel de condición física-salud en escolares, entre final de Educación Primaria (12 años) y final de Educación Secundaria Obligatoria (16 años)*. Tesis doctoral. Universidad de Granada.

Cochran, W.G. (1977) *Sampling techniques*, third edition. John Wiley & Sons, New York.

Khare, B.B. (1991) Determination of sample sizes for a class of two phase sampling estimators for ratio and product of two population means using auxiliary character. *Metron: revista internazionale di statistica*, 49, 185-197.

Okafor, F.C. (1992) The theory and application of sampling over two occasions for the estimation of current population ratio. *Statistica*, 1, 137-147.

Rao, P.S.R.S. y Mudholkar, G.S. (1967) Generalized multivariate estimators for the mean of finite populations. *Journal of the American Statistical Association*, 62, 1008-1012.