

# UTILIZACIÓN DE INFORMACIÓN AUXILIAR PARA LA ESTIMACIÓN DE LAGOS CON RIESGO DE ACIDIFICACIÓN

M. Rueda<sup>\*1</sup>, J. F. Muñoz<sup>2\*\*</sup>, A. Arcos<sup>3\*</sup> and E. Álvarez<sup>\*\*4</sup>

<sup>\*</sup>Department of Statistics and O.R., University of Granada, C.P. 18071, Granada, Spain,

<sup>\*\*</sup>Department of Quantitative Methods in Economics and Business, University of Granada, C.P. 18071, Granada, Spain

## ABSTRACT

In this work we have collected estimators of the population proportion through the use of auxiliary information under simple random sampling. We analyzed the main properties of these estimators. Theoretical properties suggest that the estimators with use of auxiliary information can outperform alternative methods, and the results derived from a Monte Carlo simulation study support this view. Has made an application on a population of Risk of Acidification in Lakes in the Northeastern states of the U.S. conducted by the Environmental Monitoring and Assessment Program of the US Environmental Protection Agency. In the application we have compared the different estimators each other with the standard estimators.

**KEYWORDS:** Auxiliary attribute, qualitative variable, ratio estimator, regression estimator

**MSC:** 62P10

## RESUMEN

En este trabajo se han recogido los estimadores de la proporción poblacional mediante el uso de información auxiliar bajo muestreo aleatorio simple. Se han estudiado sus principales propiedades. Los resultados teóricos sugieren que los estimadores que usan información auxiliar superan a métodos alternativos y los resultados que se obtienen con las simulaciones de Monte Carlo avalan también esta conclusión. Se ha realizado una aplicación sobre una población de Lagos utilizando los datos de la encuesta realizada por la Agencia Estadounidense de Protección del Medioambiente.

En la aplicación se han comparado los distintos estimadores entre sí y con el estimador estándar.

## 1 INTRODUCCIÓN

La mayoría de las muestras sobre recursos naturales contienen información auxiliar a nivel poblacional además de los datos muestrales. Esta información auxiliar procede de imágenes de satélite, fotografías aéreas, datos de GPS, etc. A partir de los datos las agencias estadísticas crean conjuntos de datos que muestran tanto el diseño muestral como la información auxiliar. En presencia de información auxiliar métodos como el de razón, diferencia, calibración [2] y [7], o como el de regresión y regresión no paramétrica [1] son métodos basados en distintos diseños muestrales que mejoran la precisión del estimador en la fase de estimación [8]. Dichas técnicas son generalmente más eficientes que las que produce los métodos estándar, ya que dichos métodos no utilizan información auxiliar. Sin embargo las técnicas que usan información auxiliar habían sido abordadas en su mayoría únicamente para variables cuantitativas. Muchos de los estudios de investigación utilizan variables cualitativas extraídas de muestras complejas. Sin embargo, los métodos de estimación tradicionales no tienen en cuenta el diseño muestral, y dichos métodos tampoco hacen uso de la información auxiliar asociada a la variable de interés. El uso de esta información en la fase de estimación puede mejorar considerablemente la precisión de la estimación en el contexto de variables cualitativas. Los primeros trabajos sobre los métodos de regresión, razón y diferencia son recientes [5] y [6]. En el presente trabajo se hace una

---

<sup>1</sup> [mrueda@ugr.es](mailto:mrueda@ugr.es)

<sup>2</sup> [jfmunoz@ugr.es](mailto:jfmunoz@ugr.es)

<sup>3</sup> [arcos@ugr.es](mailto:arcos@ugr.es)

<sup>4</sup> [encarniav@ugr.es](mailto:encarniav@ugr.es)

comparación de estos trabajos y su comportamiento empírico. Uno de los grandes problemas que existen en nuestros días es la escasez de recursos naturales, tales como el agua. Las condiciones ecológicas de los lagos en Estados Unidos constituyen un aspecto muy importante analizado por la Agencia Estadounidense de Protección del Medioambiente. Utilizando los datos extraídos de la encuesta de lagos muestreada por el Programa de Evaluación y Vigilancia del Medioambiente realizada por la Agencia Estadounidense de Protección del Medioambiente se comparan los distintos estimadores y su comportamiento empírico en este contexto. De tal forma que el interés reside en estimar la proporción de lagos con riesgo de acidificación.

## 2 ESTIMACIÓN DE LA PROPORCIÓN POBLACIONAL MEDIANTE EL USO DE INFORMACIÓN AUXILIAR

Se considera la población finita  $U = \{1, \dots, N\}$  que contiene  $N$  unidades distintas e identificadas. Donde  $A_1, \dots, A_N$  denota los valores del atributo de interés  $A$ , donde  $A_i = 1$  si la  $i$ -ésima unidad presenta el atributo  $A$  y  $A_i = 0$  en caso contrario. Sea  $B$  un atributo auxiliar asociado con el atributo de interés  $A$  y cuyos valores vienen dados por  $B_1, \dots, B_N$ . Para ello, asumimos que se ha extraído una muestra  $s$ , de tamaño  $n$ , la cual ha sido seleccionada de  $U$  bajo muestreo aleatorio simple o bajo muestreo con probabilidades desiguales.

El objetivo será estimar la proporción poblacional de individuos que poseen el atributo de interés  $A$ , es decir:  $P_A = N^{-1} \sum_{i=1}^N A_i$ . Si asumimos una población finita, el estimador estándar o de expansión simple de  $P_A$  viene dado por:

$$\hat{p}_A = \frac{1}{n} \sum_{i \in s} A_i \quad (1)$$

En [5] se define un primer estimador de tipo razón para  $P_A$  mediante:

$$\hat{p}_r = \hat{R} P_B \quad (2)$$

Donde  $\hat{R} = \frac{\hat{p}_A}{\hat{p}_B}$  es un estimador de la razón poblacional  $R = \frac{P_A}{P_B}$ , con

$\hat{p}_B = \frac{1}{n} \sum_{i \in s} B_i$  es la proporción muestral de individuos que presentan el atributo auxiliar  $B$  y  $P_B = N^{-1} \sum_{i=1}^N B_i$  es la proporción poblacional del atributo  $B$ .

Para obtener  $\hat{p}_r$  asumimos que  $P_B$  es conocido a partir de un censo o se ha estimado sin error.

### 2.1 Propiedades teóricas

Sean  $A^c$  y  $B^c$  los atributos complementarios de los atributos  $A$  y  $B$  respectivamente, y sea la tabla poblacional de doble entrada dada por

	$B$	$B^c$	
$A$	$N_{11}$	$N_{12}$	$N_{1*}$
$A^c$	$N_{21}$	$N_{22}$	$N_{2*}$
	$N_{*1}$	$N_{*2}$	$N$

(3)

Donde  $N_{1*} = \frac{1}{N} \sum_{i=1}^N A_i$  es el número de individuos en la población que poseen el atributo  $A$ ,  $N_{2*}$  es el número de individuos en la población que poseen el atributo  $A^c$ , etc. De forma análoga  $N_{11}$  es el número de individuos que simultáneamente poseen los atributos  $A$  y  $B$ ,  $N_{12}$  es el número de individuos en la población que simultáneamente poseen los atributos  $A$  y  $B^c$ , etc.

La clasificación dada en (3) se traduce a nivel muestral en la tabla de doble entrada

	$B$	$B^c$	
$A$	$n_{11}$	$n_{12}$	$n_{1*}$
$A^c$	$n_{21}$	$n_{22}$	$n_{2*}$
	$n_{*1}$	$n_{*2}$	$n$

(4)

Una aproximación del sesgo del estimador de razón  $\hat{p}_r$  viene dada por

$$B(\hat{p}_r) = \frac{N-n}{N-1} \frac{1}{n} \left( \frac{Q_B}{P_B} - \frac{\Phi \sqrt{P_A Q_A P_B Q_B}}{P_A} \right) \quad (5)$$

Donde  $Q_B = 1 - P_B$  y

$$\Phi = \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N_{1*}N_{2*}N_{*1}N_{*2}}}$$

Es el coeficiente  $V$  de Cramer basado en la clasificación dada en la tabla de doble entrada (3).

La expresión (5) implica que  $\hat{p}_r$  es asintóticamente insesgado. Sin embargo, el sesgo podría no ser insignificante para tamaños muestrales pequeños. Se puede demostrar fácilmente que un estimador de  $B(\hat{p}_r)$  viene dado por

$$\hat{B}(\hat{p}_r) = \frac{1-f}{n-1} \left( \frac{\hat{q}_B}{\hat{p}_B} - \frac{\hat{\Phi} \sqrt{\hat{p}_A \hat{q}_A \hat{p}_B \hat{q}_B}}{\hat{p}_A} \right) \quad (6)$$

Donde  $\hat{q}_B = 1 - \hat{p}_B$  y

$$\hat{\Phi} = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1*}n_{2*}n_{*1}n_{*2}}}$$

Es el coeficiente  $V$  de Cramer basado en la tabla de doble entrada (4).

La varianza asintótica del estimador de razón  $\hat{p}_r$  es

$$AV(\hat{p}_r) = \frac{N-n}{N-1} \frac{1}{n} \left( P_A Q_A + R^2 P_B Q_B - 2R\Phi \sqrt{P_A Q_A P_B Q_B} \right) \quad (7)$$

y un estimador insesgado de esta varianza viene dado por

$$\hat{V}(\hat{p}_r) = \frac{1-f}{n-1} \left( \hat{p}_A \hat{q}_A + \hat{R}^2 \hat{p}_B \hat{q}_B - 2\hat{R}\hat{\Phi} \sqrt{\hat{p}_A \hat{q}_A \hat{p}_B \hat{q}_B} \right) \quad (8)$$

Los autores definen también un nuevo estimador de tipo razón más eficiente que  $\hat{p}_r$ . Este estimador de tipo razón se basa en la siguiente idea. El estimador estándar  $\hat{p}_A$  se puede obtener también como  $\hat{p}_A = 1 - \hat{q}_A$ , donde  $\hat{q}_A = n^{-1} \sum_{i \in S} A_i^c$ , lo que implica que  $\hat{p}_A$  tiene el mismo comportamiento que  $\hat{q}_A$  en la estimación de  $Q_A$ . Sin embargo, puede comprobarse fácilmente que esta propiedad no se cumple para el estimador  $\hat{p}_r$ , es decir,  $\hat{p}_r \neq 1 - \hat{q}_r$ , donde  $\hat{q}_r = \hat{R}_c Q_B$  es el estimador de tipo razón para  $Q_A$  y  $\hat{R}_c = \frac{\hat{q}_A}{\hat{p}_B}$ . Este resultado nos lleva a que un estimador alternativo para  $P_A$  es  $\hat{p}_{r,q} = 1 - \hat{q}_r$ .

El interés reside en analizar cuando  $\hat{p}_r$  posee mejores propiedades que  $\hat{p}_{r,q}$  y viceversa. Para dar solución a dicha cuestión se considerará el criterio de mínima varianza.

Los autores demuestran que si la varianza del estimador de tipo razón  $\hat{p}_r$  es menor que la varianza del estimador de tipo razón  $\hat{p}_{r,q}$ , es decir,  $AV(\hat{p}_r) < AV(\hat{p}_{r,q})$ , si  $P_A < P_B$ .

Se puede deducir que  $\hat{V}(\hat{p}_r) < \hat{V}(\hat{p}_{r,q})$  cuando  $\hat{p}_A < \hat{p}_B$ . Este resultado nos permite definir el siguiente estimador más eficiente que  $\hat{p}_r$ :

$$\hat{p}_{r,e} = \begin{cases} \hat{p}_r & \text{si } \hat{p}_A < \hat{p}_B \\ \hat{p}_{r,q} & \text{en otro caso} \end{cases} \quad (9)$$

Las varianzas de este estimador vienen dadas por:

$$AV(\hat{p}_{r,q}) = AV(1 - \hat{q}_r) = AV(\hat{q}_r)$$

Se deduce que

$$AV(\hat{p}_{r,e}) = \begin{cases} AV(\hat{p}_r) & \text{si } P_A < P_B \\ AV(\hat{q}_r) & \text{en otro caso} \end{cases} \quad (10)$$

Un estimador de  $AV(\hat{p}_{r,e})$  vendrá dado por

$$\hat{V}(\hat{p}_{r,e}) = \begin{cases} \hat{V}(\hat{p}_r) & \text{si } \hat{p}_A < \hat{p}_B \\ \hat{V}(\hat{q}_r) & \text{en otro caso} \end{cases} \quad (11)$$

Donde  $AV(\hat{p}_{r,e})$  y  $\hat{V}(\hat{q}_r)$  pueden determinarse fácilmente a partir de  $AV(\hat{p}_r)$  y  $\hat{V}(\hat{p}_r)$ . Cuando  $\hat{p}_A = \hat{p}_B$  se observa que  $\hat{p}_r = \hat{p}_{r,q}$  lo cual implica que

$$AV(\hat{p}_r) = AV(\hat{q}_r)$$

$$\text{y } \hat{V}(\hat{p}_r) = \hat{V}(\hat{q}_r).$$

El estimador estándar  $\hat{p}_A$  tiene el mismo comportamiento en la estimación de la  $P_A$  que  $\hat{q}_A$  en la estimación de  $Q_A$ , puesto que  $\hat{p}_A = 1 - \hat{q}_A$ .

Se observa fácilmente que el estimador  $\hat{p}_r$  no posee esa propiedad. Sin embargo el estimador propuesto  $\hat{p}_{r,e}$  satisface esta propiedad, esto es,  $\hat{p}_{r,e} = 1 - \hat{q}_{r,e}$

En [3] se define un nuevo estimador mediante una combinación lineal de los mencionados estimadores. El valor óptimo del peso utilizado en la combinación lineal se determinará mediante el criterio de mínima varianza.

El nuevo estimador de tipo razón viene dado por

$$\hat{p}_{r,w} = w\hat{p}_r + (1 - w)\hat{p}_{r,q} \quad (12)$$

El valor óptimo de  $w$  en el sentido de mínima varianza dentro la clase de estimadores  $\hat{p}_{r,w}$  viene dado por:

$$w_{opt} = \frac{AV(\hat{p}_{r,q}) - cov(\hat{p}_r, \hat{p}_{r,q})}{AV(\hat{p}_r) + AV(\hat{p}_{r,q}) - 2cov(\hat{p}_r, \hat{p}_{r,q})} \quad (13)$$

Por tanto, el estimador óptimo en el sentido de mínima varianza dentro de la clase (12) viene dado por

$$\hat{p}_{r,opt} = w_{opt}\hat{p}_r + (1 - w_{opt})\hat{p}_{r,q}$$

En la práctica, el estimador  $\hat{p}_{r,opt}$  puede ser desconocido, puesto que el peso de  $w_{opt}$  depende de varianzas poblacionales, las cuales son generalmente desconocidas. En esta situación, emplearemos el estimador

$$\hat{p}_{r,opt} = \hat{w}_{opt}\hat{p}_r + (1 - \hat{w}_{opt})\hat{p}_{r,q} \quad (14)$$

Donde

$$\hat{w}_{opt} = \frac{\hat{V}(\hat{p}_{r,q}) - \widehat{cov}(\hat{p}_r, \hat{p}_{r,q})}{\hat{V}(\hat{p}_r) + \hat{V}(\hat{p}_{r,q}) - 2\widehat{cov}(\hat{p}_r, \hat{p}_{r,q})} \quad (15)$$

La varianza del estimador óptimo  $\hat{p}_{r,opt}$  viene dada por

$$AV(\hat{p}_{r,opt}) = \frac{V(\hat{p}_r)V(\hat{p}_{r,q}) - cov^2(\hat{p}_r, \hat{p}_{r,q})}{V(\hat{p}_r) + V(\hat{p}_{r,q}) - 2cov(\hat{p}_r, \hat{p}_{r,q})}$$

Esta expresión nos sirve también para obtener el siguiente estimador de la varianza del estimador óptimo  $\hat{p}_{r,opt}$

$$\hat{V}(\hat{p}_{r,opt}) = \frac{\hat{V}(\hat{p}_r)\hat{V}(\hat{p}_{r,q}) - \widehat{cov}^2(\hat{p}_r, \hat{p}_{r,q})}{\hat{V}(\hat{p}_r) + \hat{V}(\hat{p}_{r,q}) - 2\widehat{cov}(\hat{p}_r, \hat{p}_{r,q})}$$

El peso óptimo de  $\hat{w}_{opt}$  dado en la expresión (13) puede obtenerse como

$$\hat{w}_{opt} = \frac{R_c - \beta}{R_c - R'} \quad (16)$$

Donde

$$\beta = \frac{cov(\hat{p}_A, \hat{p}_B)}{V(\hat{p}_B)}$$

Teniendo en cuenta que el peso óptimo estimado  $\hat{w}_{opt}$  dado en la expresión (15) tendrá la siguiente expresión bajo MAS

$$\hat{w}_{opt} = \frac{\hat{R}_c - \hat{\beta}}{\hat{R}_c - \hat{R}'}$$

donde

$$\hat{\beta} = \frac{\widehat{cov}(\hat{p}_A, \hat{p}_B)}{\hat{V}(\hat{p}_B)}$$

Una expresión alternativa para  $\hat{w}_{opt}$  viene dada por

$$\hat{w}_{opt} = \frac{\hat{R}_c - \hat{\Phi}\sqrt{\hat{R}'\hat{R}_c}}{\hat{R}_c - \hat{R}'}$$

Puesto que

$$\hat{\beta} = \frac{\widehat{cov}(\hat{p}_A, \hat{p}_B)}{\hat{V}(\hat{p}_B)} = \frac{\hat{\Phi}\sqrt{\hat{p}_A\hat{q}_A\hat{p}_B\hat{q}_B}}{\hat{p}_B\hat{q}_B} = \hat{\Phi}\sqrt{\frac{\hat{p}_A\hat{q}_A}{\hat{p}_B\hat{q}_B}} = \hat{\Phi}\sqrt{\hat{R}'\hat{R}_c}$$

El estimador de tipo razón óptimo  $\hat{p}_{r,opt}$  dado en la expresión (14) puede obtenerse bajo MAS como

$$\hat{p}_{r,opt} = \hat{p}_A + \hat{\beta}(P_B - \hat{p}_B)$$

Este estimador se puede obtener de forma alternativa a través del uso del método de regresión. Definiendo el estimador de regresión [6] como:

$$\hat{p}_{reg}^{opt} = \hat{p}_A + \hat{b}_{opt}(P_B - \hat{p}_B)$$

donde

$$\hat{b}_{opt} = \frac{\hat{\Phi} \sqrt{\hat{p}_A \hat{q}_A}}{\sqrt{\hat{p}_B \hat{q}_B}}$$

Observamos que el estimador de tipo regresión óptimo  $\hat{p}_{reg}^{opt}$  coincide con el estimador de tipo razón óptimo  $\hat{p}_{r.opt}$  propuesto anteriormente. Es destacable, en el contexto del muestreo en poblaciones finitas, el hecho de que a partir de una combinación lineal de dos estimadores de tipo razón resulte un estimador de tipo regresión.

Un estimador insesgado para  $V(\hat{p}_{reg}^{opt})$  viene dado por

$$\hat{V}(\hat{p}_{reg}^{opt}) = \hat{V}(\hat{p}_A)(1 - \hat{\Phi}^2)$$

Donde

$$\hat{V}(\hat{p}_A) = \frac{1 - f}{n - 1} \hat{p}_A \hat{q}_A$$

### 3 APLICACIÓN: ACIDIFICACIÓN EN LAGOS

Uno de los grandes problemas que existen en nuestros días es la escasez de recursos naturales, tales como el agua. Las condiciones ecológicas de los lagos en Estados Unidos constituyen un aspecto muy importante analizado por la Agencia Estadounidense de Protección del Medioambiente. En particular existe un gran interés en el conocimiento de la acidez del agua (Pratesi, Ranalli y Salvati, 2008). Utilizando los datos extraídos de la encuesta de lagos muestreada por el Programa de Evaluación y Vigilancia del Medioambiente realizada por la Agencia Estadounidense de Protección del Medioambiente se comparan los distintos estimadores y su comportamiento empírico en este contexto. De tal forma que el interés reside en estimar la proporción de lagos con riesgo de acidificación.

En el presente trabajo los métodos propuestos se evalúan numéricamente usando datos de la encuesta de lagos muestreada. Estos conjuntos de datos proceden de 334 lagos con un total de medidas de 557. El objetivo es estimar la proporción de lagos en riesgo de acidificación. Nosotros consideraremos las variables de interés relacionadas con la capacidad de neutralización del ácido (ANC). Un valor de ANC menor que cero indica que el agua es ácida. Si los valores de ANC se aproximan a cero el lago pierde su capacidad de almacenamiento en el búfer. Los valores de ANC entre 50 y 200 identifican al lago en cuestión en situación de pre-alarma y en otras ocasiones críticas. Valores de ANC mayores de 500 reflejan un lago con un bajo riesgo de acidificación.

El estudio de simulación realizado consiste en considerar un total de 557 medidas de diferentes lagos procedentes de una población de la cual las muestras han sido seleccionadas bajo MAS. El interés reside en estimar la proporción de lagos cuyos valores de ANC sean menores de 0, y la proporción de lagos cuyos valores de ANC sean mayores de {200; 500}. Tales proporciones se consideran en este trabajo puesto que existe un relevante interés en la práctica.

Las proporciones poblacionales  $P_A = 0.07$  para  $ANC < 0$ ;  $P_A = 0.22$  para  $ANC > 500$  y  $P_A = 0.44$  para  $ANC > 200$  son las que se han analizado. Como se comentaba anteriormente, las proporciones mayores de 0.5 no son consideradas, puesto que este problema es equivalente a estudiar las proporciones menores que 0.5, estudiadas en este estudio. Hemos usado la concentración de calcio en cada lago (CA) para determinar los atributos auxiliares. Para  $P_A = 0.07$ , consideramos  $B_i = 1$  si el  $i$ -ésimo lago tiene un valor de CA menor que 78 y por otra parte  $B_i = 1$ . Para  $P_A = 0.22$ , consideramos valores mayores que {189, 480} como atributos

auxiliares. Para  $P_A = 0.44$  consideramos valores de CA mayores de  $\{654, 230\}$  como atributos auxiliares. Nótese que se podrían haber usado como atributos auxiliares otros valores de CA o el uso de otras variables. Siguiendo las medidas de comparación generalmente utilizadas la comparación empírica de los estimadores se realizó en términos de sesgo relativo (SR) y ganancia en eficiencia (GE), el error cuadrático medido relativo (ECMR), donde

$$ECMR = \frac{\sqrt{ECM(\hat{p})}}{P_A}$$

y  $GE = \frac{ECM(\hat{p})}{ECM(\hat{p}_*)}$  donde  $ECM(\hat{p}_*)$  ser el error cuadrático medio de alguno de los estimadores que se han presentado anteriormente.

Los resultados obtenidos en los estudios de simulación asociados a las poblaciones basadas en datos reales se muestran a continuación.

**Tabla 1.** Valores del sesgo relativo (SR), error cuadrático medio relativo (ECMR) y la ganancia en eficiencia (GE) para distintos estimadores de  $P_A = 0.438$  en la población Lagos.  $P_B = 0.44$  cuando  $\Phi = 0.9$  y  $P_B = 0.163$  cuando  $\Phi = 0.5$ .

$n$	$\Phi$	Estimador	SR(%)	ECMR	GE
50	0.9	$\hat{p}_A$	-0.1	15.3	0
		$\hat{p}_r$	0.3	7.3	334.78
		$\hat{p}_{r,e}$	0.0	6.9	400
		$\hat{p}_{r,opt}$ (= $\hat{p}_{reg}^{opt}$ )	0.0	6.9	400
	0.5	$\hat{p}_A$	0.2	15.2	0
		$\hat{p}_r$	8.8	39.4	-85.11
		$\hat{p}_{r,e}$	0.1	13.2	33.33
		$\hat{p}_{r,opt}$ (= $\hat{p}_{reg}^{opt}$ )	0.1	13.2	33.33
100	0.9	$\hat{p}_A$	-0.1	10.3	0
		$\hat{p}_r$	0.1	4.7	376.19
		$\hat{p}_{r,e}$	0.0	4.6	400
		$\hat{p}_{r,opt}$ (= $\hat{p}_{reg}^{opt}$ )	0.0	4.6	400
	0.5	$\hat{p}_A$	0.1	10.3	0
		$\hat{p}_r$	3.7	20.7	-75.24
		$\hat{p}_{r,e}$	0.2	8.9	33.33
		$\hat{p}_{r,opt}$ (= $\hat{p}_{reg}^{opt}$ )	0.2	8.9	33.33

En la Tabla 1 podemos observar que todos los estimadores presentan sesgos razonables, a excepción de  $\hat{p}_r$  que tiene valores por encima del 8.8% en el caso más extremo. No obstante, observamos que el sesgo de  $\hat{p}_r$  decrece a medida que aumenta el tamaño de la muestra.

Respecto a la eficiencia de los estimadores, observamos que, tanto para  $n = 50$  como para  $n = 100$ , todos los estimadores mejoran considerablemente con respecto al estimador estándar cuando  $\Phi = 0.9$ . Sin embargo, cuando  $\Phi = 0.5$ , el estimador de razón es mucho menos eficiente que el estimador estándar, aunque esta pérdida en eficiencia va disminuyendo a medida que aumenta el tamaño muestral. Los estimadores  $\hat{p}_{r,opt}$  y  $\hat{p}_{r,e}$  son los más eficientes en todos los casos.

En la Tabla 2 todos los estimadores presentan sesgos insignificantes. La ganancia en eficiencia de los estimadores basados en información auxiliar es bastante importante cuando  $\Phi = 0.9$ . En esta población, es el

estimador de tipo estándar el que es menos eficiente, mientras que el estimador de tipo razón tiene un buen comportamiento cuando  $\phi = 0.5$ . Esta circunstancia se debe a que  $P_A < P_B$ .

**Tabla 2** Valores del sesgo relativo (SR), error cuadrático medio relativo (ECMR) y la ganancia en eficiencia (GE) para distintos estimadores de  $P_B = 0.215$  en la población Lagos. cuando  $\phi = 0.9$  y  $P_B = 0.522$  cuando  $\phi = 0.5$ .

$n$	$\phi$	Estimador	SR(%)	ECMR	GE
50	0.9	$\hat{p}_A$	-0.2	25.5	0
		$\hat{p}_r$	0.7	10.9	455.55
		$\hat{p}_{r,e}$	0.1	9.7	614.28
		$\hat{p}_{r,opt}$ (= $\hat{p}_{reg}^{opt}$ )	0.0	9.8	566.66
	0.5	$\hat{p}_A$	-0.8	25.7	0
		$\hat{p}_r$	-0.5	22.5	29.87
		$\hat{p}_{r,e}$	-0.5	22.5	29.87
		$\hat{p}_{r,opt}$ (= $\hat{p}_{reg}^{opt}$ )	-0.5	22.5	29.87
100	0.9	$\hat{p}_A$	0.1	17.2	0
		$\hat{p}_r$	0.1	7.0	488.23
		$\hat{p}_{r,e}$	-0.1	6.6	566.66
		$\hat{p}_{r,opt}$ (= $\hat{p}_{reg}^{opt}$ )	-0.1	6.7	566.66
	0.5	$\hat{p}_A$	-0.3	17.3	0
		$\hat{p}_r$	0.0	15.1	31.57
		$\hat{p}_{r,e}$	0.0	15.1	31.57
		$\hat{p}_{r,opt}$ (= $\hat{p}_{reg}^{opt}$ )	0.0	15.1	31.57

**Tabla 3** Valores del sesgo relativo (SR), error cuadrático medio relativo (ECMR) y la ganancia en eficiencia (GE) para distintos estimadores de  $P_A = 0.07$  en la población Lagos.  $\phi = 0.5$  y  $P_B = 0.176$ .

$n$	Estimador	SR(%)	ECMR	GE
50	$\hat{p}_A$	0.5	49.0	0
	$\hat{p}_r$	2.3	48.1	4.16
	$\hat{p}_{r,e}$	2.1	46.7	9.89
	$\hat{p}_{r,opt}$ (= $\hat{p}_{reg}^{opt}$ )	0.6	45.1	17.64
100	$\hat{p}_A$	0.3	33.0	0
	$\hat{p}_r$	0.8	29.7	23.45
	$\hat{p}_{r,e}$	0.8	29.7	23.45
	$\hat{p}_{r,opt}$ (= $\hat{p}_{reg}^{opt}$ )	0.1	29.1	28.20

En la simulación realizada para la población Lagos cuando  $P_A = 0.07$  (Tabla 3) podemos observar que cuando  $n = 50$  los estimadores  $\hat{p}_r$  y  $\hat{p}_{r,e}$  tienen sesgos en torno al 2%, mientras que el resto de estimadores obtienen sesgo por debajo del 1%. La ganancia en eficiencia del resto de estimadores basados en información auxiliar con respecto al estimador estándar no es tan importante como en casos anteriores, especialmente cuando  $n = 50$  ya que todos los valores correspondientes a la ganancia en eficiencia están por debajo del 30%.

#### 4 CONCLUSIONES



El uso de información auxiliar para la estimación de la proporción poblacional de variables cualitativas no es muy usual. Pero puede producir un aumento sustancial de precisión de los estimadores, especialmente cuando la correlación existente entre el atributo de interés y el atributo auxiliar es alta.

Observar que la obtención de la información auxiliar procedente del atributo  $B$  es sencilla y poco costosa puesto que solo se refiere a  $P_B$ , dado que se puede obtener de censos o encuestas similares anteriormente realizadas.

RECEIVED OCTOBER , 2011

REVISED MAY, 2012

#### REFERENCIAS

- [1] BREIDT, F.J., J.D. OPSOMER (2000): Local polynomial regression estimators in survey sampling, **Ann. Stat.** 28 1026-1053.
- [2] DEVILLE, J.C. and C.E. SÄRNDAL. (1992) Calibration Estimators in Survey Sampling. **Journal of the American Statistical Association**, 87, 376-382.
- [3] MUÑOZ, J. F., A.ARCOS; E. ÁLVAREZ; M.M. RUEDA; S. GONZALEZ y A. SANTIAGO (2011): Optimum ratio estimators for the population proportion, **International Journal of Computers Mathematics**. En prensa.
- [4] PRATESI, M., M.G. RANALLI, N. SALVATI (2008): , Semiparametric quantile regression for estimating the proportion of acidic lakes in digit HUCs of the Northeastern US, **Environmetrics** 19, 687–701.
- [5] RUEDA M.M., MUÑOZ, J.F., ARCOS, A., ÁLVAREZ-VERDEJO, E. y MARTNEZ, S. (2011a) Estimators and confidence intervals for the proportion using binary auxiliary information with applications to pharmaceutical studies. **Journal of Biopharmaceutical Statistics**; 21, 526–554.
- [6] RUEDA M.M., MUÑOZ J.F., ARCOS A. y ÁLVAREZ-VERDEJO, E. (2011b) Indirect estimation of proportions in natural resource surveys. **Mathematics and Computers in Simulation**; En prensa.
- [7] SÄRNDAL, C.E (2007): The calibration approach in survey theory and practice. **Survey Methodology**. 33, 99-119.
- [8] SINGH,S (2003): **Advanced sampling theory with applications: How Michael "selected" Amy**, Kluwer Academic Publisher, The Netherlands.