

---

# Cálculo del tamaño de muestra para la estimación de una varianza en poblaciones finitas con funciones en $\mathbb{R}^1$

Sample size for the estimation of a finite population variance with R functions

Andrés Gutiérrez<sup>a</sup>  
agutierrez@icfes.gov.co

Hanwen Zhang<sup>b</sup>  
hanwenzhang@usantotomas.edu.co

Cristian Montaña<sup>c</sup>  
cmontano@contratista.icfes.gov.co

---

## Resumen

Estimar la varianza del puntaje de una población finita en un examen estandarizado es un objetivo importante en la evaluación de la educación; sin embargo en la literatura estadística no existe una metodología general que permita determinar el tamaño de muestra mínimo necesario para estimar de forma consistente este parámetro de interés. En este artículo se realizan los desarrollos necesarios para aproximar la varianza del estimador de Hájek de la varianza poblacional por medio de la linealización de Taylor. Además, se proponen diferentes enfoques para calcular el tamaño de muestra mínimo necesario para estimar puntualmente este parámetro o para cotejar un sistema de hipótesis estadísticas. Adicionalmente se proponen funciones computacionales programadas en el *software* R que permiten calcular tamaños de muestra requeridos.

**Palabras clave:** tamaño de muestra, muestreo, pruebas estandarizadas, varianza poblacional.

## Abstract

The estimation of a finite population is a very relevant topic in the context of education assessment. However in the statistical literature there is no available a

---

<sup>1</sup>Gutiérrez, A., Zhang, H., Montaña, C. (2016) Cálculo del tamaño de muestra para la estimación de una varianza en poblaciones finitas con funciones en R. *Comunicaciones en Estadística*, **9**(1), 107-125.

<sup>a</sup>Director de Evaluación. Instituto Colombiano para la Evaluación de la Educación (Icfes)

<sup>b</sup>Profesora. Universidad Santo Tomás, Bogotá, Colombia

<sup>c</sup>Estadístico. Instituto Colombiano para la Evaluación de la Educación (Icfes)

generalized methodology that allows to compute the minimum sample size needed to guarantee accurate variance estimates. This paper provides the approximation for the Hájek estimator of the population variance by means of Taylor linearization. We also find proper expressions for the computation of the minimum sample size required to punctually estimate this parameter of interest along with testing statistical hypothesis. Besides that, we present some computational functions programmed in the R software in order to easily compute proper sample sizes.

**Keywords:** sample size, survey sampling, standardized tests, population variance.

## 1. Introducción

Los exámenes estandarizados son una herramienta de medición de la calidad de la educación que hacen uso de un instrumento o cuestionario compuesto de una cantidad finita de ítems. Desde la teoría psicométrica se ha demostrado que un constructo se mide con mayor precisión a medida que el instrumento contenga más ítems. Por supuesto, en las aplicaciones reales, el estudiante que presenta un examen estandarizado está sujeto a un tiempo limitado para responder los ítems. Lo anterior hace que sea necesario optimizar la cantidad de ítems que se aplican en cada prueba.

Debido a que la cantidad de ítems en una prueba es limitada cualquier prueba estandarizada estará sujeta a un error de medición, que disminuirá a medida que la cantidad de ítems aumente en la prueba. Por ende, cualquier resultado que se desprenda del examen tendrá asociado un error estadístico que permite realizar una inferencia apropiada con respecto a las habilidades de los examinandos.

A partir de la agregación de resultados se generan varias medidas de gran importancia que permiten evaluar la calidad de la educación con el fin de establecer planes de mejoramiento de las diferentes entidades involucradas. De esta manera, para cada aplicación se obtiene una distribución con todos los resultados individuales. Algunos parámetros de interés son: la media, la varianza, el coeficiente de variación, entre otros.

Cuando la aplicación se realiza de forma censal, entonces los parámetros de interés se calculan con base en los resultados individuales de toda la población. Este cálculo responde a expresiones matemáticas específicas. Cuando la prueba se realiza por medio de una muestra probabilística entonces los parámetros de interés deben ser estimados teniendo en cuenta la medida de probabilidad inducida por el diseño de muestreo. Es común que el diseño de muestreo se planee de tal forma que se disminuya el margen de error de estimación de la media poblacional. Sin embargo, además de estimar la media, es necesario también estimar otros parámetros de interés; en particular la varianza.

La varianza de los resultados de un examen es una medida de gran importancia en la evaluación estandarizada puesto que permite identificar la dispersión que tienen

los puntajes y hacer análisis que permitan identificar qué tan lejos del promedio está el puntaje de un estudiante particular, así como de un subgrupo poblacional de interés. Este parámetro está definido como:

$$s_{y_U}^2 = \frac{1}{N-1} \sum_u (y_k - \bar{y}_U)^2 \quad (1)$$

Donde  $y_k$  representa la variable de interés medida sobre el individuo  $k$  (o el puntaje que obtuvo el estudiante  $k$  en un examen estandarizado),  $U$  denota la población finita de tamaño  $N$  y  $\bar{y}_U = \sum_U y_k/N$ . Si una muestra probabilística  $S$  es seleccionada de  $U$  de acuerdo con un diseño de muestra  $p(\cdot)$ , entonces suponiendo que el diseño de muestreo que permitió la selección de la muestra fue aleatorio simple sin reemplazo, la varianza poblacional se puede estimar como:

$$s_{y_s}^2 = \frac{1}{n-1} \sum_s (y_k - \bar{y}_s)^2$$

Donde  $\bar{y}_s = \sum_s y_k/n$ . Särndal et al. (1992, pg.188) afirman que el siguiente estimador es consistente para  $s_{y_U}^2$  bajo cualquier diseño de muestra  $p(\cdot)$  que induzca probabilidad de inclusión  $\pi_k$  para el elemento  $k$ -ésimo:

$$\hat{s}_{yy}^2 = \frac{1}{\hat{N}-1} \sum_s \frac{(y_k - \tilde{y}_s)^2}{\pi_k} \quad (2)$$

Además note que  $\hat{N} = \sum_s 1/\pi_k$  y  $\tilde{y}_s$  es el estimador de Hájek para una media poblacional definido como  $\tilde{y}_s = (\sum_s y_k/\pi_k)/\hat{N}$ . Como es bien sabido este estimador es asintóticamente insesgado; es decir, su sesgo tiende a cero a medida que el tamaño de la población  $N$  junto con el tamaño de muestra  $n$  tienden a ser grandes. En adelante, el estimador definido por (2) será notado como el estimador de Hájek para la varianza poblacional.

Por otro lado, para conocer la precisión del estimador, es necesario identificar la varianza de estos estimadores. Al conocer esta expresión es posible cuantificar el coeficiente de variación de la estimación puntual (así como el margen de error), construir intervalos de confianza apropiados y calcular la potencia de una prueba de hipótesis. Note que los anteriores elementos proveen al investigador de una estrategia metodológica para hallar expresiones adecuadas con el objetivo de calcular tamaños de muestra necesarios para cumplir los propósitos de una investigación estadística en el contexto educativo. Las primeras referencias de este tema corresponde a Cochran (1977). Por otro lado, Cho (2004) encuentra expresiones apropiadas para la varianza de un estimador de la varianza en términos del cuarto momento teórico (aunque no aborda el problema de la estimación de la varianza en una población finita). Luego, Ardilly & Tillé (2006) desarrollan expresiones apropiadas para poblaciones finitas sobre el supuesto de que la variable de interés tiene distribución normal.

Después de una breve introducción, en la sección 2 se desarrolla el componente teórico de la aproximación de la varianza de  $\hat{s}_{yy}^2$ ; en la sección 3 se encuentran

los cálculos matemáticos necesario que permiten computar tamaños muestrales mínimos necesarios para obtener estimaciones puntuales con valores menores a un coeficiente de variación predefinido y un margen de error predefinido; en la sección 4, para el escenario de pruebas de hipótesis, se encuentra también el tamaño mínimo de muestra para cumplir con un nivel de potencia previamente definida; en la sección 5 se describe brevemente las funciones construidas en el software R que implementa los desarrollos metodológicos de esta investigación y que son incorporados en el paquete `samplesize4surveys` (Gutiérrez 2015); en la sección 6 se aplica esta metodología a los datos de la prueba Saber 11 (ICFES 2016); finalmente, en la sección 6, se brindan algunas conclusiones y recomendaciones.

## 2. Varianza aproximada del estimador de Hájek

Bajo algunas condiciones estadísticas plausibles en cualquier investigación de la evaluación de la educación, el siguiente resultado provee la aproximación para la varianza de  $\tilde{s}_{yy}^2$ .

**Resultado 1.** *Suponiendo que tanto  $N$  como  $n$  son lo suficientemente grandes y que la selección de la muestra está inducida por un diseño de muestreo aleatorio simple, la varianza de  $\tilde{s}_{yy}^2$  se puede aproximar como:*

$$\text{var}(\tilde{s}_{yy}^2) \approx \frac{N^2(NK + 2N + 2)}{n(N-1)^3} \left(1 - \frac{n}{N}\right) s_{yU}^4 \quad (3)$$

Donde  $K$  denota el coeficiente de curtosis<sup>1</sup> de la variable de interés en la población finita definido como:

$$K = \frac{\frac{1}{N} \sum_U (y_k - \bar{y}_U)^4}{\left(\frac{1}{N} \sum_U (y_k - \bar{y}_U)^2\right)^2} - 3$$

*Demostración.* En primer lugar, observamos que el estimador  $\tilde{s}_{yy}^2$  se puede ver como función de dos estimadores de totales: el primero, el estimador del tamaño poblacional  $\hat{N}$  y el segundo, un estimador poblacional de la suma de cuadrados sobre las diferencias de cada puntaje con la estimación del promedio  $\hat{t} = \sum_s \frac{(y_k - \tilde{y}_s)^2}{\pi_k}$ .

Al utilizar la técnica de linealización de Taylor (Gutiérrez 2009, sección 8.1.) de primer orden alrededor de los valores  $\hat{N} = N = \sum_U 1$  y  $\hat{t} = t = \sum_U (y_k - \bar{y}_U)^2$ , tenemos que las derivadas parciales de  $\tilde{s}_{yy}^2 = f(\hat{N}, \hat{t})$  con respecto a cada total

---

<sup>1</sup>Note que cuando  $y$  tiene una distribución perfectamente simétrica, entonces  $K = 0$ .

estimado son:

$$a_1 = \frac{\partial \tilde{s}_{yy}^2}{\partial \hat{N}} \Big|_{\hat{N}=N, \hat{t}=t} = - \frac{\hat{t}}{(\hat{N} - 1)^2} \Big|_{\hat{N}=N, \hat{t}=t} = - \frac{\sum_U (y_k - \bar{y}_U)^2}{(N - 1)^2} = - \frac{s_{yU}^2}{N - 1}$$

$$a_2 = \frac{\partial \tilde{s}_{yy}^2}{\partial \hat{t}} \Big|_{\hat{N}=N, \hat{t}=t} = \frac{1}{(\hat{N} - 1)^2} \Big|_{\hat{N}=N, \hat{t}=t} = - \frac{1}{(N - 1)^2}$$

Luego, al aplicar el teorema de Taylor para linealizar el estimador  $\tilde{s}_{yy}^2$ , se tiene que

$$\tilde{s}_{yy}^2 = f(\hat{N}, \hat{t}) \approx s_{yU}^2 + a_1(\hat{N} - N) + a_2(\hat{t} - t) \tag{4}$$

Por lo anterior, es posible definir una nueva variable linealizada como:

$$E_k = a_1(1) + a_2(y_k - \bar{y}_U)^2 = \frac{1}{N - 1} [(y_k - \bar{y}_U)^2 - s_{yU}^2]$$

Lo anterior conlleva a que una expresión apropiada para aproximar la varianza de  $\tilde{s}_{yy}^2$  está dada por:

$$var(\tilde{s}_{yy}^2) \approx var\left(\sum_s \frac{E_k}{\pi_k}\right) = \sum_{k \in U} \sum_{l \in U} \frac{\Delta_{kl}}{\pi_{kl}} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l} \tag{5}$$

Con  $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$  y  $\pi_{kl}$  definida como la probabilidad de inclusión de segundo orden. En particular, si el diseño de muestreo utilizado es aleatorio simple sin reemplazo, la aproximación de la varianza está dada por:

$$var(\tilde{s}_{yy}^2) \approx \frac{N^2}{n} \left(1 - \frac{n}{N}\right) s_{E_U}^2 \tag{6}$$

Donde:

$$\begin{aligned} S_{E_U}^2 &= \frac{1}{N - 1} \sum_U E_k^2 \\ &= \frac{1}{N - 1} \sum_U \frac{[(y_k - \bar{y}_U)^2 - s_{yU}^2]^2}{(N - 1)^2} \\ &= \frac{1}{(N - 1)^3} \sum_U [(y_k - \bar{y}_U)^4 - 2s_{yU}^2(y_k - \bar{y}_U)^2 + s_{yU}^4] \\ &= \frac{1}{(N - 1)^3} \left[ N \sum_U (y_k - \bar{y}_U)^4 - (N - 2)s_{yU}^4 \right] \end{aligned}$$

Donde  $\sum_U (y_k - \bar{y}_U)^4$  es el cuarto momento central de la variable  $y$  en la población finita y  $s_{yU}^4 = (s_{yU}^2)^2$ . Ahora, después de un poco de álgebra sobre (6), es posible encontrar que:

$$\text{var}(\tilde{s}_{yy}^2) \approx \frac{N^2}{n(N-1)^3} \left(1 - \frac{n}{N}\right) \left[ N \sum_U (y_k - \bar{y}_U)^4 - (N-2)s_{yU}^4 \right]$$

Por otro lado, el coeficiente de curtosis de la variable de interés en la población finita se define como  $K = \frac{\sum_U (y_k - \bar{y}_U)^4}{s_{yU}^4} - 3$ , de donde  $\sum_U (y_k - \bar{y}_U)^4 = (K+3)s_{yU}^4$ , y por ende se tiene la siguiente aproximación para la varianza de  $\tilde{s}_{yy}^2$ .

$$\text{var}(\tilde{s}_{yy}^2) \approx \frac{N^2(NK + 2N + 2)}{n(N-1)^3} \left(1 - \frac{n}{N}\right) s_{yU}^4$$

□

Nótese que si la distribución de los puntajes individuales es simétrica en la población finita, entonces  $K = 0$ , y por ende la aproximación de la varianza estaría determinada por la siguiente expresión:

$$\text{var}(\tilde{s}_{yy}^2) \approx \frac{2N^2(N+1)}{n(N-1)^3} \left(1 - \frac{n}{N}\right) s_{yU}^4 \quad (7)$$

## 2.1. Verificación empírica de la aproximación

Para ilustrar el comportamiento de la aproximación encontrada anteriormente, se realizaron dos ejercicios de simulación. En cada uno de ellos se simuló una población de tamaño  $N = 100.000$ . Para cada población finita se seleccionaron muestras aleatorias sin reemplazo de tamaño  $n = 100, 110, \dots, 1000$ . Para cada valor de  $n$ , se calculó la aproximación dada en 3 y además se escogieron 1000 muestras de la población finita siguiendo un diseño de muestreo aleatorio simple sin reemplazo. Para cada muestra seleccionada se calculó el estimador de la varianza dado en (2). A este conjunto de 1000 estimaciones se le calculó la varianza y esta cifra se comparó con (3).

El primer ejercicio se realizó con una población finita inducida por  $N$  realizaciones de una distribución normal con media 50 y desviación estándar 10 que induce un coeficiente de curtosis  $K = 0$ . Los resultados de la simulación se muestran en la figura 1, donde podemos ver que la aproximación es acertada. En el segundo ejercicio, la población fue simulada de una distribución gamma con esperanza igual a 200 y desviación estándar igual a 140, en cuyo caso el coeficiente de curtosis es  $K = 3$ . Los resultados de este segundo ejercicio se muestran en la figura 2; podemos ver que en general la aproximación es aceptable, aunque al parecer tiende a sobreestimar levemente la varianza de  $\tilde{s}_{yy}^2$ .

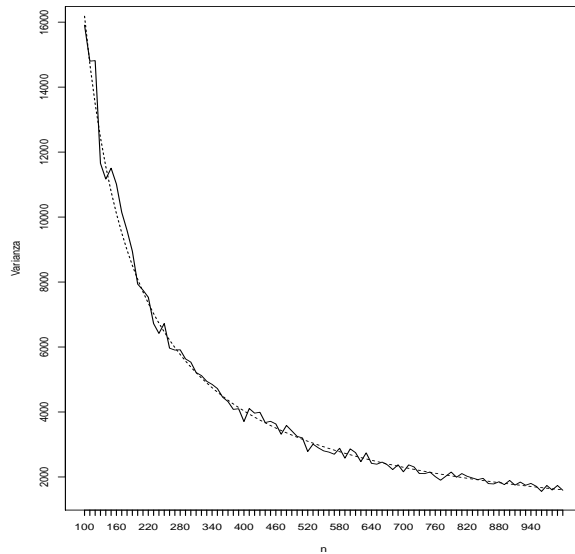


Figura 1: Varianza del estimador  $\tilde{s}_{yy}^2$  (línea continua) y la aproximación dada en el el Resultado 1 (línea punteada) en una población simulada de una distribución normal con  $K = 0$ . Fuente: elaboración propia.

### 3. Tamaño de muestra para estimar puntualmente la varianza

Con los desarrollos encontrados en la sección anterior, es posible definir expresiones apropiadas para calcular el tamaño de muestra mínimo necesario para estimar la varianza de una variable de interés sobre una población finita, sujeta a algunas restricciones como: margen de error absoluto, margen de error relativo o coeficiente de variación. En principio, se desarrollan las expresiones apropiadas sujetas a una estrategia de muestreo que depende de un diseño aleatorio simple sin reemplazo y del estimador  $\tilde{s}_{yy}^2$ . Luego, estas expresiones son generalizadas para poder ser utilizadas bajo cualquier diseño de muestreo arbitrario.

Es importante notar que cuando el diseño de muestra  $p(\cdot)$  es diferente del muestreo aleatorio simple, no es posible dar una expresión general (que permita despejar  $n$ ) para la varianza de  $s_{ys}^2$ . Sin embargo, es posible hacer uso del efecto de diseño DEFF, que se define como:

$$DEFF = \frac{var_p(s_{ys}^2)}{var_{MAS}(\tilde{s}_{yy}^2)}$$

Tenga en cuenta que el DEFF está definido en términos del diseño  $p(\cdot)$  y del estimador  $s_{ys}^2$ . La siguiente expresión es útil para generalizar la varianza del estimador

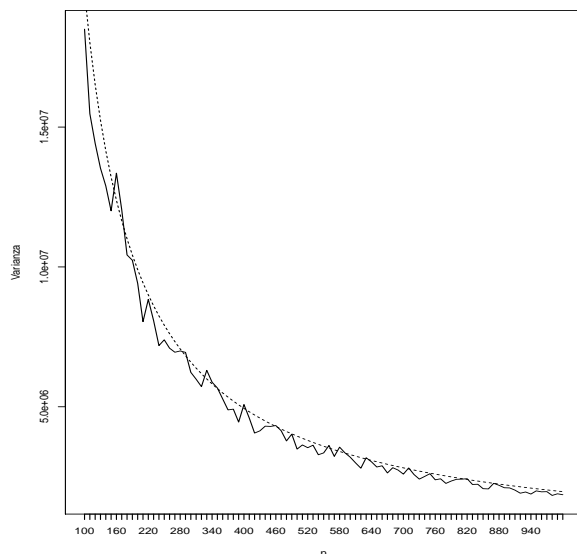


Figura 2: Varianza del estimador  $\tilde{s}_{yy}^2$  (línea continua) y la aproximación dada en el el Resultado 1 (línea punteada) en una población simulada de una distribución gamma con  $K = 3$ . Fuente: elaboración propia.

$\tilde{s}_{yy}^2$  bajo cualquier diseño  $p(\cdot)$  en términos del diseño de muestreo aleatorio simple.

$$\text{var}_p(s_{ys}^2) = DEFF \times \text{var}_{MAS}(\tilde{s}_{yy}^2)$$

Donde  $\text{var}_{MAS}(\tilde{s}_{yy}^2)$  está definida en el Resultado 1. Esta técnica es bien conocida en el cálculo de tamaños de muestra cuando el diseño de muestreo es complejo. Por supuesto, si el valor del DEFF es menor que 1, entonces la varianza del estimador bajo ese diseño de muestreo particular es baja, y por ende es de esperarse un menor tamaño de muestra. Por el contrario, si el valor del DEFF es mayor que 1, entonces la varianza del estimador bajo ese diseño de muestreo particular es mayor, y se esperaría un mayor tamaño de muestra.

Por último, los resultados expuestos en este artículo parten del supuesto de que el estimador de Hájek para la varianza tiene distribución normal asintótica; es decir,

$$\frac{\tilde{s}_{yy}^2 - s_{yU}^2}{\sqrt{\text{var}(\tilde{s}_{yy}^2)}} \sim N(0, 1)$$

Este supuesto se encuentra sustentado de forma detallada por Sen (1995) en un ensayo acerca de las propiedades del estimador de Hájek y sus contribuciones al teorema central de límite en poblaciones finitas. Sin embargo, es necesario adver-



tir que para diseños de muestreo que induzcan un tamaño efectivo de muestra aleatorio, este supuesto de normalidad empieza a ser bastante débil.

### 3.1. Minimizando el margen de error absoluto

El margen de error absoluto se define desde la distribución de probabilidad del estimador de interés. En este caso particular se asume que el estimador Hájek para la varianza  $\tilde{s}_{yy}^2$  sigue una distribución normal asintótica con media  $s_{yU}^2$  y varianza  $var(\tilde{s}_{yy}^2)$ . Luego, es posible encontrar un intervalo de confianza de nivel  $(1 - \alpha/2)$  partiendo de la siguiente expresión:

$$1 - \alpha \leq P(|\tilde{s}_{yy}^2 - s_{yU}^2| < MEA) \tag{8}$$

Por lo tanto, es para encontrar el tamaño de muestra  $n$  mínimo necesario para estimar de forma precisa la varianza poblacional  $s_{yU}^2$  se fija un margen de error absoluto  $MEA$  de antemano, restringido a la siguiente expresión:

$$1 - \alpha \leq P\left(\left|\frac{\tilde{s}_{yy}^2 - s_{yU}^2}{\sqrt{var(\tilde{s}_{yy}^2)}}\right| < \frac{MEA}{\sqrt{var(\tilde{s}_{yy}^2)}}\right)$$

Asumiendo normalidad, se tiene que:

$$\frac{MEA}{\sqrt{var(\tilde{s}_{yy}^2)}} \geq z_{1-\alpha/2} \Rightarrow var(\tilde{s}_{yy}^2) \leq \frac{MEA^2}{z_{1-\alpha/2}^2}$$

Retomando la aproximación de  $var(\tilde{s}_{yy}^2)$  en (3) y después de un poco de álgebra, finalmente se obtiene la siguiente expresión que permite obtener un mínimo tamaño de muestra para estimar el parámetro de interés con un margen de error relativo menor a  $MEA$  cuando el diseño de muestreo es aleatorio simple sin reemplazo y cuando el estimador utilizado es  $\tilde{s}_{yy}^2$ .

$$n \geq \frac{\frac{z_{1-\alpha/2}^2 s_{yU}^4}{MEA^2}}{\frac{(N-1)^3}{N^2(NK+2N+2)} + \frac{z_{1-\alpha/2}^2 s_{yU}^4}{MEA^2 N}} \tag{9}$$

Si el diseño de muestreo es diferente al aleatorio simple sin reemplazo, entonces se tiene la siguiente condición para el tamaño de muestra  $n$  en términos del margen de error absoluto:

$$n \geq \frac{\frac{z_{1-\alpha/2}^2 s_{yU}^4 \times DEFF}{MEA^2}}{\frac{(N-1)^3}{N^2(NK+2N+2)} + \frac{z_{1-\alpha/2}^2 s_{yU}^4 \times DEFF}{N \times MEA^2}} \tag{10}$$

Finalmente, también es útil poder cuantificar el margen de error absoluto fijando un tamaño de muestra preestablecido. Nótese que la siguiente expresión es útil para estimar las cotas de un intervalo de confianza sobre  $s_{yU}^2$ .

$$\begin{aligned} MEA &= z_{1-\alpha/2} \sqrt{\text{var}(\tilde{s}_{yy}^2)} \\ &= z_{1-\alpha/2} s_{yU}^2 \frac{N}{N-1} \sqrt{\frac{NK+2N+2}{n(N-1)} \left(1 - \frac{n}{N}\right)} \end{aligned} \quad (11)$$

Es claro que este cálculo es imposible de realizar, por cuanto se necesitaría conocer el valor de  $s_{yU}^2$ . En este caso, es posible estimar el margen de error absoluto reemplazando la anterior expresión por  $\tilde{s}_{yy}^2$ .

### 3.2. Minimizando el margen de error relativo

Por otro lado, el anterior desarrollo también se puede formular en términos del margen de error relativo (MER), en este caso, la ecuación (8) se convierte en:

$$1 - \alpha \geq P \left( \left| \frac{\tilde{s}_{yy}^2 - s_{yU}^2}{s_{yU}^2} \right| < MER \right) = P \left( |\tilde{s}_{yy}^2 - s_{yU}^2| < MER \times s_{yU}^2 \right) \quad (12)$$

De donde se puede concluir que  $MER \times s_{yU}^2 = MEA$ . Siguiendo la misma lógica y después de un poco de álgebra, fácilmente se concluye que la expresión de  $n$  dada en (9) se convierte en:

$$n \geq \frac{\frac{z_{1-\alpha/2}^2}{MER^2}}{\frac{(N-1)^3}{N^2(NK+2N+2)} + \frac{z_{1-\alpha/2}^2}{N \times MER^2}} \quad (13)$$

Si el diseño de muestreo difiere del aleatorio simple sin reemplazo, entonces la siguiente expresión da cuenta del tamaño de muestra apropiado al minimizar el margen de error relativo:

$$n \geq \frac{\frac{z_{1-\alpha/2}^2 \times DEFF}{MER^2}}{\frac{(N-1)^3}{N^2(NK+2N+2)} + \frac{z_{1-\alpha/2}^2 \times DEFF}{N \times MER^2}} \quad (14)$$

El margen de error relativo (fijando un tamaño de muestra preestablecido) está dado por la siguiente expresión:

$$MER = z_{1-\alpha/2} \frac{N}{N-1} \sqrt{\frac{NK+2N+2}{n(N-1)} \left(1 - \frac{n}{N}\right)} \quad (15)$$

Nótese que en este caso, este cálculo es completamente factible puesto que esta expresión no depende del parámetro que se quiere estimar.

### 3.3. Minimizando el coeficiente de variación estimado

Si la exigencia sobre el tamaño de muestra recae en lograr un coeficiente de variación (CVE) menor a un umbral predefinido, entonces es necesario realizar un desarrollo algebraico simple que empieza por definir apropiadamente este término:

$$\begin{aligned} CVE &= \frac{\sqrt{\text{var}(\hat{s}_{yy}^2)}}{s_{yU}^2} \\ &= \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) \frac{N^2(NK + 2N + 2)}{(N-1)3}} \end{aligned} \quad (16)$$

De donde podemos concluir que el tamaño de muestra requerido debe cumplir la siguiente condición:

$$n \geq \frac{N^2(NK + 2N + 2)}{CVE^2(N-1)^3 + N(NK + 2N + 2)} \quad (17)$$

En situaciones más generales con diseños muestrales complejos, al utilizar el DEFF se concluye que la expresión adecuada para calcular el tamaño de muestra (minimizando el CVE) es la siguiente:

$$n \geq \frac{N^2(NK + 2N + 2) \times DEFF}{CVE^2(N-1)^3 + N(NK + 2N + 2) \times DEFF} \quad (18)$$

## 4. Tamaño de muestra para realizar pruebas de hipótesis sobre la varianza

El cálculo del tamaño de muestra no se reduce únicamente a la estimación puntual de parámetros en una población finita, o incluso a la estimación de intervalos de confianza. También es posible considerar el juzgamiento de prueba de hipótesis sobre la varianza poblacional de los resultados en la prueba estandarizada. Este camino se diferencia bastante del tradicional, puesto que el objetivo de la investigación no es la estimación puntual de  $s_{yU}^2$ . En primer lugar considere el siguiente sistema de hipótesis:

$$H_0 : s_{yU}^2 = s_{y0}^2 \quad \text{vs.} \quad H_a : s_{yU}^2 > s_{y0}^2$$

Equivalentemente, el sistema puede ser reescrito como:

$$H_0 : s_{y_U}^2 - s_{y_0}^2 = 0 \quad \text{vs.} \quad H_a : s_{y_U}^2 - s_{y_0}^2 = D > 0$$

Nótese que  $D$  es el efecto nulo que el investigador considera apropiado definir como umbral que determine que a partir de ese mismo valor  $D$  se considere que la diferencia entre la varianza de la población finita  $s_{y_U}^2$  y el valor nulo  $s_{y_0}^2$  no es despreciable. Ahora, apelando a la normalidad asintótica sobre  $\tilde{s}_{yy}^2$  la regla de decisión con nivel de significación  $\alpha$  es **rechazar**  $H_0$  cuando:

$$\frac{\tilde{s}_{yy}^2 - s_{y_0}^2}{\sqrt{\text{var}(\tilde{s}_{yy}^2)}} > z_{1-\alpha} \quad (19)$$

De esta forma, la función de potencia está dada por:

$$\begin{aligned} \beta(s_{y_U}^2) &= P\left(\frac{\tilde{s}_{yy}^2 - s_{y_0}^2}{\sqrt{\text{var}(\tilde{s}_{yy}^2)}} > z_{1-\alpha}\right) \\ &= P\left(\tilde{s}_{yy}^2 > z_{1-\alpha}\sqrt{\text{var}(\tilde{s}_{yy}^2)} + s_{y_0}^2\right) \\ &= P\left(\frac{\tilde{s}_{yy}^2 - s_{y_U}^2}{\sqrt{\text{var}(\tilde{s}_{yy}^2)}} > z_{1-\alpha} - \frac{s_{y_U}^2 - s_{y_0}^2}{\sqrt{\text{var}(\tilde{s}_{yy}^2)}}\right) \\ &= 1 - \Phi\left(z_{1-\alpha} - \frac{s_{y_U}^2 - s_{y_0}^2}{\sqrt{\text{var}(\tilde{s}_{yy}^2)}}\right) \end{aligned}$$

La potencia, definida como la probabilidad (condicionada a la hipótesis alterna) de detectar una diferencia  $D$  entre  $s_{y_U}^2$  y  $s_{y_0}^2$  puede escribirse de la siguiente manera:

$$\begin{aligned} \beta &< P\left(\frac{\tilde{s}_{yy}^2 - s_{y_0}^2}{\sqrt{\text{var}(\tilde{s}_{yy}^2)}} > z_{1-\alpha} \mid s_{y_U}^2 - s_{y_0}^2 = D\right) \\ &= 1 - \Phi\left(z_{1-\alpha} - \frac{D}{\sqrt{\text{var}(\tilde{s}_{yy}^2)}}\right) \\ &\approx 1 - \Phi\left(z_{1-\alpha} - \frac{D}{s_{y_U}^2 \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) \frac{N^2(NK + 2N + 2)}{(N-1)^3}}}\right) \quad (20) \end{aligned}$$

Por lo anterior, suponiendo que la selección de la muestra está inducida por un diseño de muestreo aleatorio simple con reemplazo, es posible afirmar que:

$$s_{y_U}^4 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{N^2(NK + 2N + 2)}{(N-1)^3} < \frac{D^2}{(z_{1-\alpha} + z_\beta)^2}$$

Al despejar el valor  $n$  de la anterior desigualdad, tenemos que el tamaño de muestra mínimo para conservar una potencia de por lo menos  $1 - \beta$ , al detectar una diferencia  $D$ , está dado por:

$$n > \frac{s_{yU}^4}{\frac{D^2}{(z_{1-\alpha} + z_\beta)^2} \frac{(N-1)^3}{N^2(NK + 2N + 2)} + \frac{s_{yU}^4}{N}} \quad (21)$$

Cuando el diseño de muestra es complejo, podemos hacer uso del DEFF, de la misma forma que en las anteriores secciones. En este caso, la condición para  $n$  se convierte en:

$$n > \frac{s_{yU}^4 \times DEFF}{\frac{D^2}{(z_{1-\alpha} + z_\beta)^2} \frac{(N-1)^3}{N^2(NK + 2N + 2)} + \frac{s_{yU}^4 \times DEFF}{N}} \quad (22)$$

## 5. Funciones computacionales

El paquete `samplesize4surveys` de R contiene funciones que permiten calcular el tamaño de muestra para la estimación de una proporción, una media, diferencia de dos proporciones y diferencia de dos medias. También permite el cálculo del error de muestreo y del nivel de potencia para un tamaño de muestra fijo.

A continuación se presentan cuatro funciones para la estimación de una varianza poblacional y para la realización de pruebas de hipótesis estadísticas sobre este parámetro de interés. Enseguida se presenta la descripción de estas funciones:

- La función `ss4S2` permite calcular el tamaño de muestra para la estimación de  $s_{yU}^2$  sujeto a un valor predefinido del coeficiente de variación o del margen de error relativo. Adicionalmente ofrece al usuario la opción de graficar el coeficiente de variación y el margen de error como una función del tamaño de muestra, con el fin de facilitar la determinación de  $n$ .
- La función `ss4S2H` permite calcular el tamaño de muestra para la estimación de  $s_{yU}^2$  sujeto al nivel de potencia para el detectar una varianza poblacional mayor al valor de la hipótesis nula. También ofrece al usuario la opción de graficar el nivel de potencia en función del tamaño de muestra.
- La función `e4S2` permite calcular el coeficiente de variación y el margen de error para un tamaño de muestra fijo. También permite obtener una gráfica similar a la de `ss4S2`.
- La función `b4S2` permite calcular el nivel de potencia para un tamaño de muestra fijo. También permite obtener una gráfica similar a la de `ss4S2H`.

Para poder utilizar las anteriores funciones es necesario instalar y cargar el paquete que las contiene en el **Comprehensive R Archive Network**. Para lo cual, desde la consola en R es necesario digitar las siguientes líneas de código:

```
install.packages("samplesize4surveys")
library(samplesize4surveys)
```

Por otro lado, como el paquete está en permanente actualización, sus autores han dispuesto de un repositorio itinerante en el cual los usuarios pueden utilizar las funciones más recientes e interactuar con la comunidad académica con el fin de corregir posibles errores en los códigos computacionales, mejorar la eficiencia de las funciones, entre otros aspectos. Para acceder a esta versión de control desde R, es necesario digitar las siguientes líneas.

```
library(devtools)
install_github("psirusteam/samplesize4surveys")
```

Por ejemplo, la siguiente línea de código arroja el tamaño de muestra necesario para estimar la varianza de una característica de interés en una población finita con un coeficiente de curtosis de uno para alcanzar un coeficiente de variación estimado de a lo más 5% y un margen de error relativo del 3%.

En la figura 3 se observa que para una población de diez mil personas, con un coeficiente de curtosis de uno, es necesario seleccionar una muestra de al menos 1937 estudiantes para que el coeficiente de variación estimado sea de menos de 5%; también es necesario seleccionar una muestra de al menos 7193 para que el margen de error relativo sea de a lo más 3%.

Por otro lado, si la necesidad del estudio radica, no en la estimación puntual de la varianza poblacional, sino en el juzgamiento de una hipótesis estadística, entonces el tamaño de muestra mínimo estará dado por la función `ss4S2H`. En particular, suponga una población simétrica (curtosis nula) de diez mil estudiantes, en donde la selección de la muestra se realiza con un diseño aleatorio simple sin reemplazo, y sobre la cual se quiere probar el siguiente sistema de hipótesis:

$$H_0 : s_{yU}^2 = 110 \quad vs. \quad H_a : s_{yU}^2 > 110$$

Las expresiones encontradas en este artículo indican que el tamaño de muestra depende del efecto nulo  $D$  y de  $s_{yU}^4$ , lo cual implica que es necesario conocer<sup>2</sup> una estimación cercana al valor de  $s_{yU}^2$ . En este caso particular asuma que el efecto nulo es de 10 puntos y que una estimación plausible de la varianza poblacional es 120. Por lo tanto, para una confianza del 95% y una potencia del 80%, el tamaño de muestra mínimo necesario para probar el anterior sistema de hipótesis es de 1512 estudiantes. La curva del tamaño de muestra se observa en la figura 4.

<sup>2</sup>En el contexto de la evaluación de la educación esto no supone un reto mayor, puesto que se acostumbra a realizar exámenes censales cada cierto tiempo. Por lo anterior, una estimación cercana de  $s_{yU}^2$  será el valor calculado de la varianza poblacional para el último examen censal.

```
ss4S2(N = 10000, K = 1, cve = 0.05, me = 0.03, DEFF = 2, plot = TRUE)
```

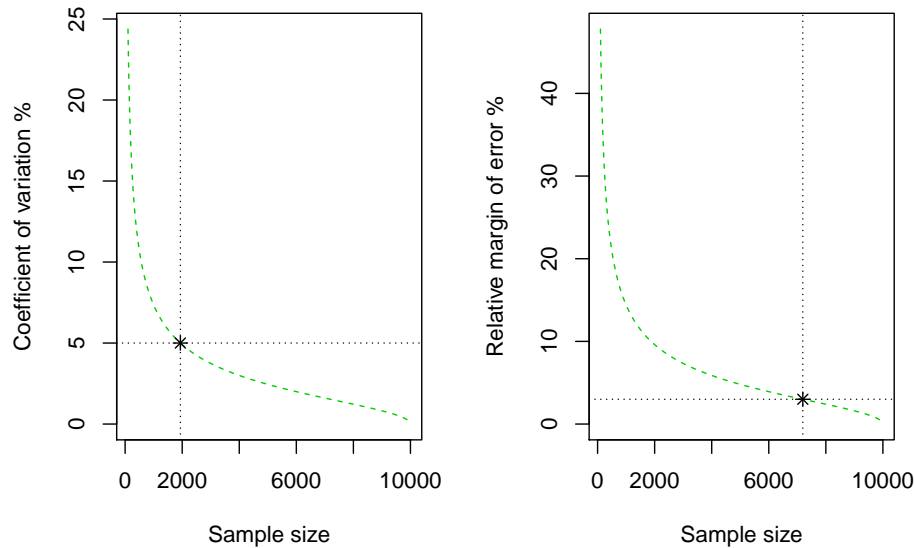


Figura 3: *Tamaños de muestra necesarios para minimizar el coeficiente de variación (izquierda) y el margen de error (derecha). Fuente: elaboración propia.*

## 6. Aplicación: la muestra controlada para el examen Saber 11

El Instituto Colombiano para la Evaluación de la Educación (Icfes), es la entidad encargada de medir la calidad de la educación en Colombia mediante el uso de exámenes estandarizados; en particular, y de manera semestral realiza la aplicación del examen de estado Saber 11 (ICFES 2016), en el cual se evalúan a estudiantes del último grado de la educación media. Este examen también es utilizado por las instituciones de educación superior como filtro de admisión.

En Saber 11 se evalúan a todos los estudiantes inscritos a la prueba en las áreas de matemáticas, lectura crítica, sociales y ciudadanas, ciencias naturales e inglés. Por ende, si se quisieran obtener algunas características poblacionales como la media o la varianza de los resultados en cada área, bastaría con calcular estos valores y no sería necesario realizar ninguna estimación. Sin embargo, el Icfes debe garantizar que exista una muestra de referencia, sobre la cual se extremen algunas medidas de control logístico, y con la cual se logra replicar los valores poblacionales. De esta forma se podrá respaldar el proceso de calificación en el caso de que ocurran

```
ss4S2H(N = 10000, S2 = 120, S20 = 110, plot = TRUE)
```

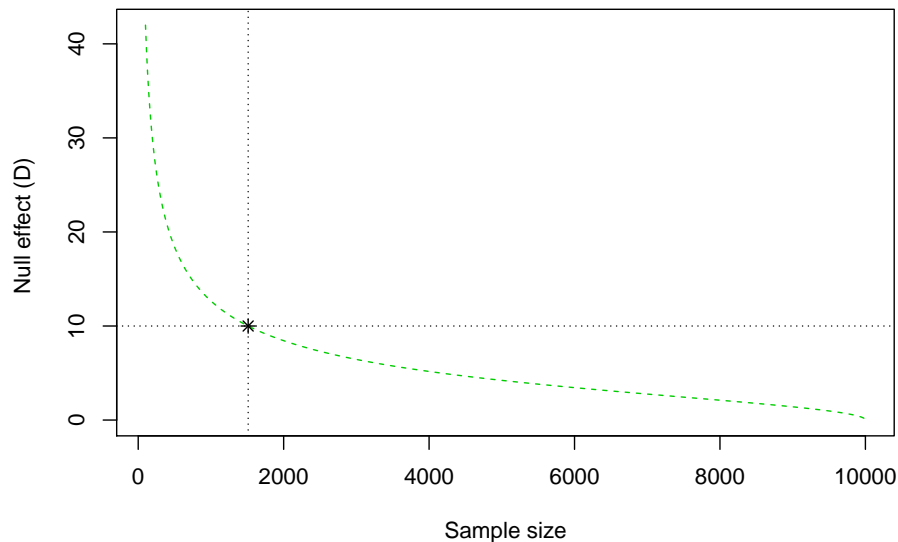


Figura 4: *Tamaños de muestra necesarios para probar un sistema de hipótesis estadístico sobre la varianza poblacional. Fuente: elaboración propia.*

inconvenientes de tipo logístico al momento de desarrollar la aplicación en campo.

Para lo anterior, se selecciona una muestra probabilística que busca estimar la media y la varianza de los resultados en las distintas áreas. En particular, es importante encontrar un tamaño de muestra que garantice el cumplimiento de algunos valores admisibles sobre los parámetros tratados en el presente artículo para la estimación de la varianza. Para realizar la selección de la muestra, se consolida el marco de muestreo con base en la información de inscripción de los estudiantes y con la asignación de las personas a los sitios y salones a donde van a presentar la prueba. El diseño de muestreo que se plantea para la selección se realiza en las siguientes tres etapas:

- **Primera etapa:** comienza con la selección de los sitios de aplicación, se utiliza un diseño de muestreo sistemático con variables de estratificación implícita (municipio de presentación de la prueba y cantidad de estudiantes que presentarán la prueba en cada sitio). Se busca garantizar que en la muestra se encuentren sitios grandes y pequeños distribuidos en todo el país.
- **Segunda etapa:** dentro de cada sitio de aplicación seleccionado en la pri-



mera etapa, se escogen todos los salones con estudiantes, es decir, se realiza censo. Esto optimiza la disposición de recursos logísticos dentro de la prueba.

- **Tercera etapa:** se finaliza seleccionando una muestra (aleatoria simple sin reemplazo) de estudiantes dentro de cada salón, a su vez incluido en cada sitio seleccionado en las etapas anteriores.

Nótese que el interés de este estudio se centra en la estimación puntual de la varianza poblacional. Por lo tanto, interesa definir el tamaño global de la muestra; es decir, encontrar el número total de estudiantes que se deben seleccionar, para lo cual se utilizará información auxiliar que proviene de la aplicación anterior más reciente. La variable auxiliar se define como los resultados en la prueba de matemáticas (con rango entre cero y cien) en la aplicación censal inmediatamente anterior, la cual presenta una media de 55.1, una varianza poblacional de 241.5, y un coeficiente de curtosis de -0.026. Con base en esta información y usando el paquete `samplesize4surveys`, se procede a estimar un tamaño de muestra adecuado para la estimación de la varianza utilizando las funciones `ss4S2` y `ss4S2H`.

Para fijar los valores admisibles, se analizó qué tan grande podía ser el DEFF dada la importancia del estudio y las implicaciones logísticas de utilizar un tamaño de muestra grande. De acuerdo con lo anterior y usando simulaciones tipo Montecarlo con la información auxiliar, se estima un valor de DEFF igual a 4.5; además, se obtiene un tamaño de muestra de referencia. Sin embargo, el tamaño final depende de los valores que se definan para el CVE, el error relativo y la potencia de la prueba.

Teniendo en cuenta lo expuesto anteriormente, se usó el paquete `samplesize4surveys` para calcular los tamaños de muestra mínimos necesarios. En particular, se definió que CVE de 4.5% es un umbral apropiado, en cuyo caso es necesario seleccionar una muestra de al menos 3305 estudiantes; mientras que para obtener un margen de error relativo de menos del 9% se debe escoger una muestra de tamaño 3434. Nótese que partiendo de la expresión (12), un margen de error relativo del 9% implica un margen de error absoluto de  $0.09 * 241.5 = 21.7$  puntos sobre los puntajes de matemáticas de la aplicación anterior; es decir, que las cotas inferior y superior del intervalo de 95% confianza son 219.8 y 263.2, respectivamente. En términos de la desviación estándar poblacional, las cotas inferior y superior del intervalo de 95% confianza son 14.8 y 16.2. Estos valores se han considerado como pertinentes, teniendo en cuenta que alcanzar intervalos de confianza de menor longitud induce un aumento significativo en el tamaño de la muestra.

Finalmente, y con base en lo anteriormente mencionado, para dar alcance a los objetivos planteados en esta aplicación, se define un tamaño de muestra de 3.434 estudiantes, al ser este un valor que permitiría cumplir con todos los valores deseables para las estadísticas que se han trabajado en este artículo para la estimación de la varianza de los resultados nacionales.

## 7. Conclusiones

En este artículo se aborda el problema de la estimación de una varianza poblacional en un estudio por muestreo. Se encuentra una expresión aproximada para la varianza del estimador, y también las expresiones para calcular el tamaño de muestra sujeto al error de muestra en términos del coeficiente de variación y el margen de error o sujeto al nivel de potencia cuando se trata del juzgamiento de hipótesis. En cuanto a los detalles computacionales, se crearon cuatro funciones nuevas para el paquete `samplesize4surveys` del *software* estadístico R.

En la mayoría de la literatura que se encuentra en la actualidad se aborda principalmente el problema de la estimación para poblaciones finitas, de parámetros como totales o medias poblacionales y se han propuesto desarrollos matemáticos para realizar la estimación de sus respectivas varianzas. Sin embargo, existen otras medidas de bastante utilidad e interés; en particular en el actual artículo se encontraron expresiones para aproximar la estimación de la varianza del estimador de Hájek de un muestreo aleatorio simple sin remplazo. Con base en esto, se deducen expresiones teóricas para calcular el tamaño mínimo de muestra en diseño más complejos haciendo uso del efecto de diseño DEFF, puesto que no es posible encontrar una expresión general para el cálculo del tamaño de muestra en este tipo de diseños complejos.

Con estos desarrollos, es posible definir tamaños de muestra para realizar estimaciones puntuales sobre la varianza en poblaciones finitas cuando se fijan un margen de error absoluto, margen de error relativo de estimación o coeficiente de variación. De manera análoga, con estos resultados es posible realizar pruebas de hipótesis sobre la varianza, para así garantizar que se conserva una potencia específica sobre un efecto nulo  $D$ . Adicionalmente, con los resultados encontrados se pueden determinar tamaños de muestra para funciones de la varianza como, por ejemplo, la desviación estándar.

En el paquete `samplesize4surveys` se implementaron las funciones que permiten desarrollar el cálculo de los tamaños de muestra para la varianza poblacional utilizando los criterios descritos en este artículo. De esta manera se facilita la aplicación de esta metodología en los problemas cotidianos de planteamiento de los diseños de muestreo y definición de tamaños muestrales. Además, en términos de eficiencia computacional, claramente esta metodología supera la definición de los tamaños de muestra utilizando por ejemplo simulaciones de tipo montecarlo.

En el caso particular del exámen Saber 11 que administra el Icfes, para respaldar el proceso de calificación, se deben obtener estimaciones precisas tanto de la media como de la varianza poblacional, y con las expresiones desarrolladas fue posible determinar eficientemente un tamaño de muestra que cumpliera con los umbrales particulares sobre los errores de estimación. Similarmente para otros tipos de prueba desarrolladas por el Icfes los resultados que se presentan en este artículo serán de gran utilidad, en particular en la prueba Saber 359 para la cual<sup>3</sup> se selecciona

---

<sup>3</sup>En esta prueba se evalúa a estudiantes de grado 3, 5 y 9 en todo el país

una muestra controlada de sedes y con la cual se realizan estimaciones nacionales de vital importancia en la toma de decisiones en el sector educativo.

Los resultados que se presentaron se pueden extender a cualquier tipo contexto o estudio en el cual se defina una población finita y se requiera determinar un tamaño de muestra bajo un margen de error de estimación permitido; por ejemplo en estudios de estimación de ingresos, de estimación de desempleo, en la evaluación de políticas públicas, etc.

A pesar de que en el presente artículo se desarrolló una metodología matemática para definir un tamaño de muestra en una población finita, se debe aclarar que existen otros factores que influyen sobre la decisión final, como, por ejemplo, los costos del estudio, en las expresiones presentadas no se está teniendo en cuenta esta componente y en futuras investigaciones pueden llegar a ser un aporte significativo dada su importancia. De igual manera, se pueden realizar avances análogos en la determinación del tamaño de muestra cuando se desean estimar otro tipo de parámetros de interés.

**Recibido: 1 de marzo de 2016**

**Aceptado: 15 de abril de 2016**

## Referencias

- Ardilly, P. & Tillé, Y. (2006), *Sampling Methods. Exercises and Solutions*, Springer.
- Cho, E. (2004), 'The Variance of Sample Variance for a Finite Population', *ASA Section on Survey Research Methods* .
- Cochran, W. G. (1977), *Sampling Techniques*, Wiley.
- Gutiérrez, H. A. (2009), *Estrategias de Muestreo. Diseño de encuestas y estimación de parámetros*, Universidad Santo Tomás.
- Gutiérrez, H. A. (2015), *samplesize4surveys: Sample Size Calculations for Complex Surveys*. R package version 2.4.0.900.  
\*<https://CRAN.R-project.org/package=samplesize4surveys>
- ICFES (2016), *Información de la prueba Saber 11*.  
\*<http://www.icfes.gov.co/index.php>
- Särndal, C.-E., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer.
- Sen, P. K. (1995), 'The Hájek Asymptotics for Finite Population Sampling and Their Ramifications.', *Kybernetika* **31**, 151 – 268.