UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Statistics on the evaluation of education
June 2016, Vol. 9, No. 1, pp. 99–117

Sample size calculation for the estimation of a variance in finite populations with R functions [1]

Cálculo del tamaño de muestra para la estimación de una varianza en poblaciones finitas con funciones en R

Andrés Gutiérrez[a]
agutierrez@icfes.gov.co

Hanwen Zhang[b]
hanwenzhang@usantotomas.edu.co

Cristian Montaño[c]
cmontano@contratista.icfes.gov.co

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

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

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

[1] Gutiérrez, A., Zhang, H., Montaño, C. (2016) Sample size calculation for the estimation of a variance in finite populations with R functions. *Comunicaciones en Estadística*, **9**(1), 99-117.

[a] Director of Assessment. Colombian Institute for the Assessment of Education, (Icfes)
[b] Professor. Santo Tomás University, Bogotá, Colombia
[c] Statistician. Colombian Institute for the Assessment of Education, (Icfes)

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. Adicionalmete 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.

1 Introduction

The standardized test is a tool to measure the quality of education that use an instrument or questionnaire with a finite amount of items. The psychometric theory has shown that a construct is measured more accurately as the instrument contains more items. Of course, in real applications, the student that presents a standardized test is subject to a limited time to answer the items. This makes it necessary to optimize the amount of items applied in each test.

Because the number of items in a test is limited, any standardized test will be subject to a measurement error, which will decrease as the number of items in the test increases. Therefore, any result that emerges from the examination will be associated with a statistical error that allows an appropriate inference regarding the skills of the examinees.

From the aggregation of results several measures of great importance for assessing the quality of education to establish improvement plans of the different entities involved are generated. Thus, for each application distribution with all individual results are obtained. Some parameters of interest are: median, variance, coefficient variation, among others.

When the application is made as a census, the parameters of interest are calculated based on the individual results of the whole population. This calculation responds to particular mathematical expressions. When the test is done on a random probability sample, the parameters of interest must be estimated taking into account the probability measure induced by the sampling design. It is common that the sampling design is planned so that it decreases the error margin of the estimate of the population median. However, in addition to estime the median, it is also necessary to estimate other parameters of interest; especially variance.

The variance of the results of a test is a measure of great importance in the standardized assessment since it allows to identify the dispersion of scores and make analysis to identify how far of the average is the score of a particular student, as well as how far of a population subgroup of interest. This parameter is defined as:

$$s^2_{y_U} = \frac{1}{N-1}\sum_u (y_k - \bar{y}_U)^2 \tag{1}$$

Where y_k represents the variable of interest measured over individual k (or the score obtained by the student k in a standardized test), U denotes the finite population of size N and $\bar{y}_U = \sum_U y_k/N$. If probability sample S is selected from U according to a sampling design $p(\cdot)$, then assuming that the sampling design that allowed the sample selection was simple random without replacement, the population variance can be estimated as:

$$s^2_{ys} = \frac{1}{n-1}\sum_s (y_k - \bar{y}_s)^2$$

Where $\bar{y}_s = \sum_s y_k/n$. Särndal et al. (1992, pg.188) claim that the next estimator is consistent for $s^2_{y_U}$ under any sampling design $p(\cdot)$ that introduce an inclusion probability π_k for the k-simal element:

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

Besides note that $\hat{N} = \sum_s 1/\pi_k$ and $\tilde{y}_s$ is the estimator of Hájek for a population median defined as $\tilde{y}_s = (\sum_s y_k/\pi_k)/\hat{N}$. As is well known this estimator is asymptotically unbiased; i.e., its bias tends to zero as the size of the population N with the sample size n tend to be large. Thereafter, the estimator defined by (2) will be noticed as the estimator of Hájek for population variance.

On the other hand, to know the accuracy of the estimate, it is necessary to identify the variance of these estimators. Knowing this expression makes possible to quantify the variation coefficient of the point estimate (as well as the margin of error), build appropriate confidence intervals and calculate the power of a hypothesis test. Note that the above elements provide the researcher a methodological strategy to find appropriate expressions with the aim of calculating the sample sizes needed to meet the purposes of statistical research in the educational context. The first references to this issue correspond to Cochran (1977). On the other hand, Cho (2004) finds appropriate expressions for the variance of an estimator of the variance in terms of the fourth theoretical moment (though he doesn't address the problem of estimating the variance of a finite population). Then Ardilly & Tillé (2006) develop appropriate expressions for finite populations on the assumption that the variable of interest has a normal distribution.

After a short introduction, in Section 2 we develop the theoretical component of the variance information of $\tilde{s}^2_{yy}$. In Section 3 we present the necessary mathematical calculations to compute minimum sampling sizes needed to obtain point estimates with values that are lower to a predefined variation coefficient and a predefined margin of error. In section 4, for the scenario of hypothesis samples is also the minimum sample size to met with a level of predefined power. In section 5 it's briefly described the built functions in the software `R` implementing the methodological developments of this research and they are incorporated in the `samplesize4surveys` package (Gutiérrez 2015). In section 6 this methodology is applied to the data of the test Saber 11 (ICFES 2016)). Finally, in section 6 some conclusions and recommendations are given.

2 Approximate variance of the Hájek estimator

Under some plausible statistical conditions in any research of the evaluation of education, the next result provides the information for the variance of $\tilde{s}^2_{yy}$.

Result 1. *Assuming that both N and n are large enough and that the selection of the sample is induced by a simple random sampling design, the variance of* $\tilde{s}^2_{yy}$ *can be approximated :*

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

Where K denotes the coefficient of kurtosis[1] *dof the variable of interest in the finite population defined as:*

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

Proof. First, we note that the estimator $\tilde{s}^2_{yy}$ can be viewed as a function of two estimators of totals: the first, the estimate of the population size $\hat{N}$ and the second, a population estimate of the sum of squares over the differences of each score with the estimate of average $\hat{t} = \sum_s \frac{(y_k-\tilde{y}_s)^2}{\pi_k}$.

By using the Taylor linearization technique (Gutiérrez 2009, sección 8.1.) of first order around the values $\hat{N} = N = \sum_U 1$ y $\hat{t} = t = \sum_U (y_k-\bar{y}_U)^2$, we have that the partial derivatives of $\tilde{s}^2_{yy} = f(\hat{N},\hat{t})$ with respect to each estimated total are:

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

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

Then, when applying the theorem of Taylor to linearize the estimator $\tilde{s}^2_{yy}$, we have that

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

Therefore, it is possible to define a new variable linearized as:

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

[1]Note that when y has a perfectly symmetric distribution K = 0.

LThis leads to an appropriate expression to approximate the variance of that $\tilde{s}^2_{yy}$ is given by:

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

With $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ y π_{kl} defined as the probability of inclusion of second order. Particularly, if the sampling design used is simple random without replacement, the variance information is given by:

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

Where:

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

Where $\sum_U (y_k - \bar{y}_U)^4$ is the forth central moment of the variable y and in the finite population y $s^4_{y_U} = (s^2_{y_U})^2$. Now, after a little of algebra about (6), it's possible to find that:

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

On the other hand, the coefficient of kurtosis of the variable of interest in the finite population is defined as $K = \frac{\sum_U (y_k - \bar{y}_U)^4}{s^4_{y_U}} - 3$, from where $\sum_U (y_k - \bar{y}_U)^4 = (K+3)s^4_{y_U}$, and therefore we have the following information for the variance of $\tilde{s}^2_{yy}$.

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

□

Note that if the distribution of individual scores is symmetric in the finite population, then K = 0, and therefore the information of the variance would be deter-

mined by the following expression:

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

2.1 Empirical verification of the information

In order to illustrate the behavior of the information previously found, two exercises of simulation were made. In each of them, a population of size $N = 100.000$ For each finite population random samples without replacement of size $n = 100, 110, \cdots, 1000$ were selected. For each value of n, the information given in 3 was calculated and besides 1000 samples of the finite population were chosen following a simple random sampling design without replacement. For each selected sample the estimator of variance given in (2) was calculated. The variance was calculated for this set of 1000 estimates and this cipher was compared to (3).

The first exercise was made with a finite population induced by N realizations of a normal distribution with media 50 and standard deviation 10 that induces a coefficient of kurtosis $K = 0$. Results of simulation are shown in figure 1, where we can see that the information is correct. In the second exercise, the population was simulated from a gamma distribution with expectation equal to 200 and standard deviation equal to 140, in which case the coefficient of kurtosis is $K = 3$. Results of this second exercise are shown in figure 2; we can see that in general, the information is acceptable, although apparently it tends to slightly overestimate the variance of $\tilde{s}^2_{yy}$.

3 Sample size for point estimation of the variance

With the developments found in the previous section it's possible to define appropriate expressions to calculate the minimum sample size needed to estimate the variance of a variable of interest over a finite population, subject to some restrictions like margin of absolute error, relative margin of error or coefficient of variation. In principle, appropriate expressions subject to a sampling strategy that relies on a simple random design without replacement and the estimator $\tilde{s}^2_{yy}$ are developed. Then, these expressions are generalized to be used under any arbitrary sampling design.

It is important to note that when the sample design $p(\cdot)$ is different from simple random sampling is not possible to give a general expression (allowing to clear n) for the variance of s^2_{ys}. However, it is possible to use the DEFF design effect, which is defined as:

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

Note that the DEFF is defined in terms of design $p(\cdot)$ and the estimator s^2_{ys}. Following expression is useful to generalize the variance of the estimator $\tilde{s}^2_{yy}$ under

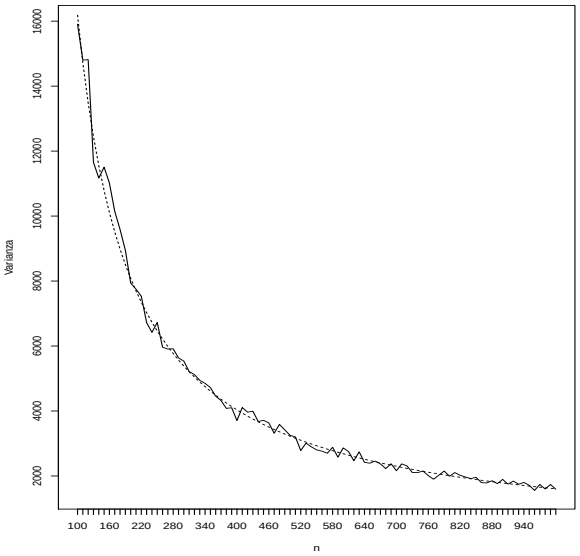

Figure 1: *Variance of estimator $\tilde{s}^2_{yy}$ (continuous line) and the approximation given in the result 1(dotted line)in a simulated population of a normal distribution with K=0. Source: own elaboration.*

any design $p(\cdot)$ in terms of the simple random sampling design.

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

Where $var_{MAS}(\tilde{s}^2_{yy})$ is defined in result 1. This technique is well known in the calculation of sample sizes when sampling design is complex. Of course, if the value of DEFF is less than 1, then the estimator variance under that particular sampling design is low, and therefore it is expected a smaller sample size. Conversely, if the value of DEFF is greater than 1, then the variance of the estimator under that particular sampling design is increased, and a larger sample size would be expected. Finally, the results presented in this article are based on the assumption that the estimator of Hájek for the variance has asymptotic normal distribution; it is,

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

This assumption is supported in detail Sen (1995) in an essay on the properties of the estimator Hájek and its contributions to the central theorem of limit in finite populations. However, it must be noted that for sampling designs that induce an effective random sample size, this assumption of normality begins to be quite weak.

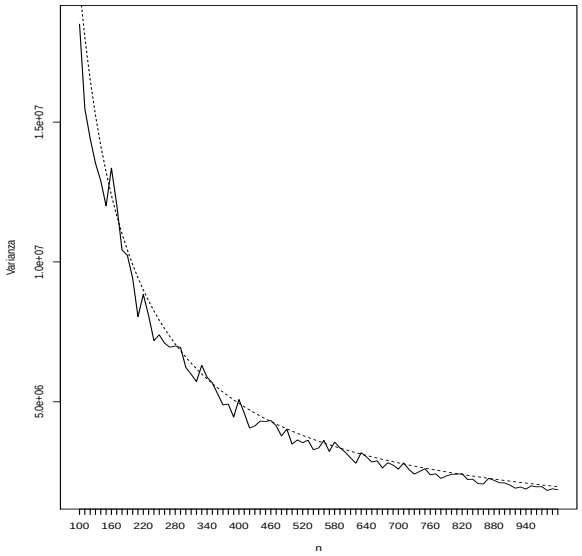

Figure 2: *Variance of estimator $\tilde{s}^2_{yy}$ (continuous line) and the approximation given in the result 1(dotted line) in a simulated population of a gamma distribution with K=3. Source: own elaboration.*

3.1 Minimizing the margin of absolute error

The margin of absolute error is defined from the distribution of probability of the estimator of interest. In this particular case it's assumed that the Hájek estimator for variance $\tilde{s}^2_{yy}$ follows a normal asymptotic distribution with median $s^2_{y_U}$ and variance $var(\tilde{s}^2_{yy})$. Then it's possible to find an interval of confidence of level $(1-\alpha/2)$ based on the following expression:

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

Therefore, it is to find the minimum sample size n necessary to accurately estimate the population variance $s^2_{y_U}$ that a margin of absolute error MEA is fixed beforehand, restricted to the following expression:

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

Assuming normality, we have that:

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

Retaking the information of $var(\tilde{s}^2_{yy})$ in (3) and then a little of algebra, we finally obtain the following expression that let us obtain a minimum sample size to estimate the parameter of interest with a relative margin of error lower than MEA when the sampling design is random without replacement and when the used estimator is $\tilde{s}^2_{yy}$.

$$n \geq \frac{\frac{z^2_{1-\alpha/2} s^4_{y_U}}{MEA^2}}{\frac{(N-1)^3}{N^2(NK+2N+2)} + \frac{z^2_{1-\alpha/2} s^4_{y_U}}{MEA^2 N}} \tag{9}$$

If sampling design is different to the random simple without replacement, then we have the following condition for the sampling size in terms of the margin of absolute error:

$$n \geq \frac{\frac{z^2_{1-\alpha/2} s^4_{y_U} \times DEFF}{MEA^2}}{\frac{(N-1)^3}{N^2(NK+2N+2)} + \frac{z^2_{1-\alpha/2} s^4_{y_U} \times DEFF}{N \times MEA^2}} \tag{10}$$

Finally to quantify the margin of absolute error fixing a preset sample size is also useful. Note that the following expression is useful to estimate the levels of an interval of confidence over $s^2_{y_U}$.

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

It's clear that is impossible to do this calculation because it would be necessary to know the value of $s^2_{y_U}$. In this case it's possible to estimate the margin of absolute error replacing the previous expression with $\tilde{s}^2_{yy}$.

3.2 Minimizing the relative margin of error

On the other hand, the previous development can also be formuled in terms of the relative margin of error (MER), in this case the equation turns into:

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

From where it can be concluded that $MER \times s^2_{y_U} = MEA$. Following the same logic and after a little of algebra, we can easily conclude that the expression of n

given in (9) turns into:

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

If sampling design differs from the simple random without replacement, then the following expression explains the appropriate sampling size when minimizing the relative margin of error:

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

The relative margin of error (fixing a preset simple size) is given by the following expression:

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

Note that in this case this calculation is completely feasible since this expression doesn't depend on the parameter to be estimated.

3.3 Minimizing the estimated coefficient of variation

If the requirement about the sample size falls on achieving a coefficient of variation (CVE) less than a predefined threshold, then it is necessary to perform a simple algebraic development that begins by properly define this term:

$$\begin{aligned} CVE &= \frac{\sqrt{var(\tilde{s}^2_{yy})}}{s^2_{y_U}} \\ &= \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) \frac{N^2(NK+2N+2)}{(N-1)3}} \end{aligned} \tag{16}$$

From where we can conclude that the sample size required must met the following condition:

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

In more general situations with complex sampling designs, when using the DEFF it's concluded that the appropriate expression to calculate the sample size (minimizing the CVE) is the following:

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

4 Sample size for hypothesis test for variance

The calculation of sample size it's not reduced only to the point estimate of the parameter in a finite population, or even to the estimate of intervals of confidence. It's also possible to consider the judgment of hypothesis test over the population variance of the results of the standardized test. This way is very different from the traditional, since the research's goal is not the point estimate of $s^2_{y_U}$. In first place consider the following hypothesis system:

$$H_0 : \ s^2_{y_U} = s^2_{y_0} \quad vs. \ \ H_a : \ s^2_{y_U} > s^2_{y_0}$$

The system can be equivalently rewritten as:

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

Note that D is the null effect that researcher consider appropriate to define as threshold to determine that from that same value D is considered that the difference between the variance of the finite population $s^2_{y_U}$ and the null value $s^2_{y_0}$ is not negligible. Now, appealing to asymptotic normality over $\tilde{s}^2_{yy}$ the decision rule with significance level α is to **reject** H_0 when:

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

So, the power function is given by:

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

Power, defined as the probability (subject to alternative hypothesis) of detecting

a difference D between $s^2_{y_U}$ and $s^2_{y_0}$ can be written as follows:

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

Therefore, assuming that the selection of the sample is induced by a design of simple random sampling with replacement, one can say that:

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

Clearing the value n of the above inequality, we have that the minimum sample size to keep a power of at least $1-\beta$, when detect inga difference D, it is given by:

$$n > \frac{s^4_{y_U}}{\frac{D^2}{(z_{1-\alpha}+z_\beta)^2}\frac{(N-1)^3}{N^2(NK+2N+2)} + \frac{s^4_{y_U}}{N}} \tag{21}$$

When sample design is complex we can use the DEFF as in the above sections. In this case the condition for n turns into:

$$n > \frac{s^4_{y_U} \times DEFF}{\frac{D^2}{(z_{1-\alpha}+z_\beta)^2}\frac{(N-1)^3}{N^2(NK+2N+2)} + \frac{s^4_{y_U} \times DEFF}{N}} \tag{22}$$

5 Computational functions

The package `samplesize4surveys` of `R` contain functions that allow us to calculate the sample size for the estimates of a proportion, a median, difference of two proportions and difference of two medians. It also allows the calculation of sample error and of the power level for a fixed sample size.

Now four functions are presented for the estimation of a population variance and for conducting statistical hypothesis testing on this parameter of interest. Right away is the description of these functions:

Function `ss4S2` allows calculating the sample size for estimates of s^2_{yU} subject to a preset value of the coefficient of variation or the relative margin of error. Additionally, it offers the user the option of mapping the coefficient of variation and the margin of error as a function of the sample size, to make easier the determination of n.

Function `ss4S2H` allows calculating the sample size for the estimates of s^2_{yU} subject to the power level to detect a population variance greater than the value of the null hypothesis. It also offers to the user the option of mapping the power level in function of the sample size.

Function `e4S2` allows calculating the coefficient of variation and the margin of error for a fixed sample size. It also allows obtaining a mapping similar to the one of `ss4S2`.

Function `b4S2` allow calculating the power level for a fixed sample size. It also allows obtaining a mapping similar to the one of `ss4S2H`.

In order to use the above functions it's necessary to install and charge the package that contains them in the Comprehensive `R` Archive Network, that for, it's necessary to type the following code lines from the console:

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

On the other hand, as the package is in constant update, the authors have arranged a traveling repository in which users can use the latest features and interact with the academic community in order to correct possible errors in computer codes and improve the efficiency of functions, among others. To access to this control version from 'R', it's necessary to type the following lines.

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

For example, the following code line throws the sample size necessary to estimate the variance of a characteristic of interest in a finite population with a coefficient of kurtosis of one to reach an estimate coefficient of variation of maximum 5% and a relative margin of error of 3%

Figure 3 shows that for a population of ten thousand people with a kurtosis coefficient of one, it's necessary to select a sample of at least 1937 students in order to get a coefficient of variation less than 5%. It's also necessary to select a sample of at least 7193 in order to get a relative margin of error of maximum 3%.

On the other hand, if the necessity of the study doesn't lie in the point estimate of the population variance, but in the judgment of a statistical hypothesis, then the minimum sample size would be given by the function `ss4S2H`. In particular,

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

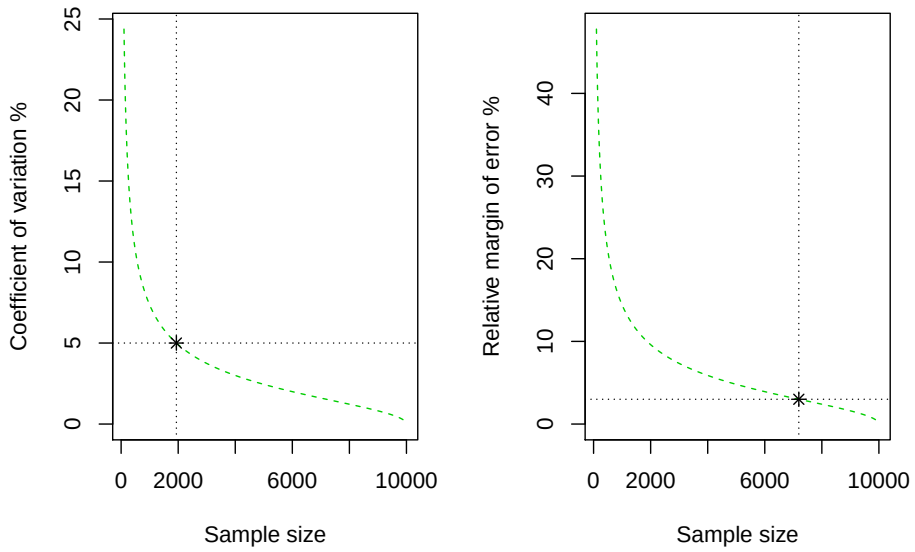

Figure 3: *Sample sizes needed to minimize the coefficient of variation (left) and the margin of error (right). Source: own*

assume an asymmetric population (null kurtosis) of ten thousand students, where the sample selection is made with a simple random design without replacement, and over which you want to prove the following hypothesis system:

$$H_0 : \ s^2_{y_U} = 110 \quad vs. \quad H_a : \ s^2_{y_U} > 110$$

The expressions found in this paper show that the sample size depend on the null effect of D and $s^4_{y_U}$, which implies that it's necessary to know [2] an estimates close to the value of $s^2_{y_U}$. In this particular case assume that the null effect is of 10 points and that a plausible estimates of the population variance is 120. Therefore, for a confidence of 95% and a power of 80%, the minimum sample size needed to prove the above hypothesis system is of 1512 students. The curve of sample size is in figure 4.

[2]In the field of the evaluation of education this is not a major challenge, since it's usual to make census tests every so often. That's why, an estimates close to $s^2_{y_U}$ will be the calculated value of the population variance for the last census test.

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

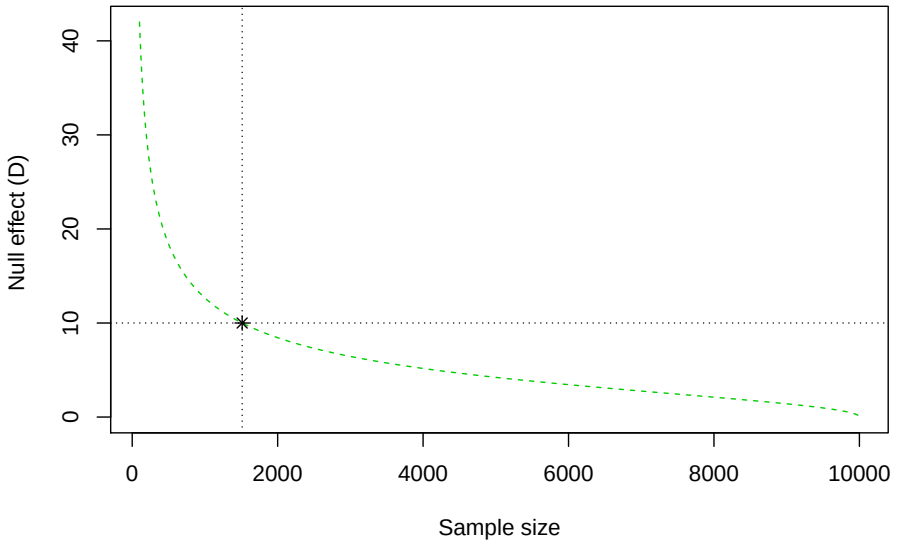

Figure 4: *Sample sizes needed to prove a statistical hypothesis system about population variance. Source: own elaboration.*

6 Application: The controlled sample for Saber 11 test

The Colombian Institute of evaluation of education (ICFES) is the entity in charge of measuring the quality of education in Colombia through the use of standardized tests; particularly, and it applies every six months the state test Saber 11 (ICFES 2016), in which students of the last grade of secondary education are tested. This test is also used by the institutions of high education as an admission filter.

Saber 11 test evaluate all the students inscribed in the test in the area of mathematics, critical writing, social and citizenship sciences, natural sciences and English. Therefore, if it were intended to obtain some population characteristics as the median or the variance of the results in each area, it would be enough to calculate these values and it wouldn't be necessary to make any estimate. However, the ICFES must guarantee that it exists a reference sample, over which some measurements of logistic control are applied, and with which population values are replied. So the score's process can be supported in case that some logistical drawbacks occur when developing the application in the field. For this, a probability sample

that seeks to estimate the mean and variance of the results in different areas is selected. In particular, it is important to find a sample size that ensures compliance with some admissible values over the discussed parameters in this paper for the estimation of variance. To make the selection of the sample, the sampling frame is consolidated based on information from student enrollment and with the assignment of people to places and rooms where they're going to present the test. The sampling design that proposed for the selection is performed in the following three stages:

First stage: It begins with the selection of the places of application, a systematic sampling design with variables of implicit stratification (municipality of presentation of the test and number of students who will submit the test at each place) is used. It's intended to ensure that in the sample are large and small places scattered across the country.

Second stage: Within each application site selected in the first stage, all classrooms with students are chosen, i.e. census is conducted. It optimizes the provision of logistic resources within the test.

Third stage: It ends by choosing a sample (simple random without replacement) of students within each classroom, in turn, included in each selected site in the previous stages.

Note that the interest of this study focuses on the point estimate of the population variance. Therefore, it's pretended to define the overall size of the sample; it is, to find the total number of students to be selected for which auxiliary information coming from the previous more recent application will be used. The auxiliary variable is defined as the results of the test of mathematics (with the range between zero and one hundred) in the immediately preceding census application, which has an average of 55.1, a population variance of 241.5, and a kurtosis coefficient of -0026. Based on this information and using the `samplesize4surveys` package, we proceed to estimate an adequate sample size for estimation of variance using `ss4S2` and `ss4S2H` functions.

To set the allowable values, we analyzed how big could be the DEFF given the importance of the study and the logistical implications of using a large sample size. According to the above and using Monte Carlo type simulations with auxiliary information, a DEFF value equal to 4.5 is estimated; besides, a reference sample size is obtained. However, the final size depends on the values defined for the CVE, the relative error and the power of the test.

Considering the above mentioned, the `samplesize4surveys` package was used to calculate the minimum sample sizes needed. In particular, it was determined that CVE of 4.5% is an appropriate threshold, in which case it is necessary to select a sample of at least 3305 students; while for obtaining a relative margin of error less than 9% a sample size of 3434 should be chosen. Note that based on the expression (12), a relative margin of error of 9% implies an absolute error margin of 0.09 * 241.5 = 21.7 points on math scores

of the previous application; i.e., that the lower and upper bounds of the 95% confidence interval are 219.8 and 263.2 respectively. Regarding the population standard deviation, the lower and upper bounds of the 95% confidence interval are 14.8 y 16.2. These values have been considered relevant taking into account that reaching lower confidence intervals induces a significant increase in the sample size.

Finally, and based on the above mentioned to reach the planned objectives in this application, a sample size of 3.434 students is defined, being this a value that allows to met all the desirable values for the statistics treated in this paper for the estimation of the variance in the national results.

7 Conclusions

This article addresses the problem of the estimation of a population variance in a sample study. An approximate expression for the variance of the estimator is found, and also the expressions to calculate the sample size subject to sampling error concerning coefficient of variation and the margin of error or subject to the power level when it comes to judging hypothesis. As for the computational details, four functions were created for the `samplesize4surveys` package of the statistical *software* `R`.

Most of the literature available now address mainly the problem of estimation for a finite population, of parameters as total or population means and mathematical developments to estimate their respective variances have been proposed. However, there are other quite useful and interesting measures; particularly in the current paper expressions to approximate the variance estimate of the Hájek estimator of a simple random sampling without replacement is found.Based on this, theoretical expressions are derived to calculate the minimum sample size in most complex designs using the design effect $DEFF$, since it is not possible to find a general expression for calculating the sample size in this type of complex designs.

With these developments, it is possible to define sample sizes for making point estimates on the variance in finite populations when a margin of absolute error and a relative margin of error of estimation or coefficient of variation are fixed. Analogously, with these results, it is possible to test hypotheses about the variance in order to ensure that a specific power on a null D effect is keeped. Additionally, with the results found sample sizes for variance functions as, for example, the standard deviation can be determined.

In the samplesize4surveys package functions that allow to develop the calculation of sample sizes for the population variance using the criteria described in this paper were implemented. In this way the application of this methodology in the everyday problems of approach of sampling designs and definition of sample sizes becomes easier. Furthermore, in terms of computational ef-

ficiency, this methodology clearly exceeds the definition of the sample sizes using for example montecarlo type simulations.

In the particular case of the Saber 11 test managed by the ICFES, accurate estimates of both average as the population variance should be obtained in order to support the score process, and with the developed expressions it was possible to determine efficiently a sample size that meets the particular thresholds about the estimation errors. Similarly for other types of tests developed by the ICFES the results presented in this paper will be very useful, particularly in the test Saber 359 for which[3] a controlled sample of sites is selected and with which national estimates of vital importance in decision-making in the education sector is performed.

The results presented can be extended to any context or study in which a finite population is defined and be necessary to determine a sample size under an allowed margin of estimation error; for example in studies of income estimation, unemployment estimation, in the evaluation of public policies, etc.

Although in this paper a mathematical methodology was developed to define a sample size in a finite population, it should be clarified that there are other factors that influence the final decision. For example, the costs of the study, in the submitted expressions this component was not considered and in future researches they can become a significant contribution given its importance. In the same way, analogous contributions in determining the sample size when you want to estimate other parameters of interest can be made.

Received: March 1, 2016
Accepted: April 15, 2016

References

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

[3]This test evaluates students in grades 3, 5 and 9 all over the country.

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.