CIENCIAS PEDAGÓGICAS
Revista electrónica científico-pedagógica
ISSN: 1605– 5888 RNPS: 1844
No.1 (enero-abril), Año 2020, 4ta Época.

MÁS DEL ENTORNO INFORMÁTICO R EN LA INVESTIGACIÓN EDUCATIVA CUBANA: ¿SE PUEDE PREDECIR LA MUESTRA DE CUBA EN EL ERCE 2019?

MORE OF THE COMPUTING ENVIRONMENT R IN THE CUBAN EDUCATIONAL INVESTIGATION: CAN IT PREDICT THECUBA´S SAMPLE IN THE ERCE 2019?

AUTOR:

Dr. C. Paul Antonio Torres Fernández. InvestigadorTitular

paul@rimed.cu

Instituto Central de Ciencias Pedagógicas, La Habana, Cuba. Coordinador Nacional por Cuba del Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE), de la OREALC-UNESCO, Santiago.

Recibido: 24 de noviembre de 2019 *Aprobado: 24 de diciembre de 2019*

RESUMEN

En este artículo se le dará continuidad al análisis de las posibilidades del entorno informático Rde contribuir al perfeccionamiento de la investigacióncubana en el campo educacional. En esta segunda ocasión se utilizará un problema científico real y vigente: la estimación de una muestra aleatoria y representativa del país en el ERCE 2019. Se explica cómo obtener una de carácter complejo, de forma estratificada,previa acomodación de la base de datos de partida. También se explica cómo hacer estimaciones estadísticas del comportamiento de importantes medidas a nivel poblacional, a partir de sus valores análogos en la muestra extraída.

PALABRAS CLAVE**:** investigación educativa, entorno informático R, teoría del muestreo.

ABSTRACT

In this article, it will be given continuity to the analysis of the possibilities of the computing environment R of contributing to the improvement of the Cuban investigation in the educational field. In this second occasion, a real and effective scientific problem will be used: the estimate of an aleatory and representative sample of the country in the ERCE 2019. It is explained how to obtain one of complex character, in stratified way,

ISSN: 1605– 5888 RNPS: 1844
No.1 (enero-abril), Año 2020, 4ta Época.

after themanipulation of the initial database. It is also explained how to make statistical estimates from important measures to populational level, with their similar values in the extracted sample.

KEYWORDS**:** educational investigation, computing environment R, theory of the sampling.

INTRODUCCIÓN

En el artículo primigenio “Lo que todo investigador educativo cubano debiera conocer: el entorno informático R” (Torres, 2018), se realizó una panorámica de las posibilidades de ese potente recurso informático en la investigación educativa, especialmente la que se realiza en Cuba.

Más allá de su potencialidad como herramienta científica auxiliar novedosa y robusta, por demás enclavada en la corriente del software libre y -por tanto- totalmente gratis, está el hecho de que se aviene “como anillo al dedo” a muchos (por no decir todos) de los retos de la investigación educativa cubana actual, expuestos por Torres (2016) (reporte disponible en el link:http://www.cubaeduca.cu/media/www.cubaeduca.cu/medias/evaluador/tesis2dogrado.pdf), todos coherente con los señalamientos críticos de otros investigadores cubanos contemporáneos.

Tal y como se dejó entrever, en el referido artículo sobre el entorno informático R, se dará continuidad a la explicación de las posibilidades de contribuir al perfeccionamiento del proceso investigativo cubano, en el campo educacional, a través de ese software. En esta segunda ocasión se utilizará un problema científico real y vigente: ¿se puede predecir la muestra que le será asignada a Cuba en elEstudio Regional Comparativo y Explicativo de la UNESCO (ERCE-2019) con este recurso informático?

El tema de la determinación de muestras estadísticas no es ajeno al debate (¡y a las urgencias!) de los retos que debe enfrentar la comunidad cubana de investigadores educativos para revertir las insuficiencias acumuladas en esta decisiva actividad del desarrollo económico y social del país.

ISSN: 1605– 5888 RNPS: 1844
No.1 (enero-abril), Año 2020, 4ta Época.

Basta recordar que en la referida tesis doctoral (Torres, 2016) se critica el hecho de que: "En la actualidad, la tendencia predominante [en la investigación educativa cubana] es la de selección de muestras intencionales, contraria a la tendencia de la selección de muestras aleatorias, que presenta un comportamiento decreciente y deprimido (…)" (Torres, 2016: 109).

A lo cual se añade que: "(…) solo el 13,8% del total de informes (…) reportan la selección de una muestra aleatoria;[y] llama la atención que en el 43,2% de los casos posibles no se puede precisar en el informe de investigación el tipo de muestra que utilizan, pues no declaran el procedimiento de selección que emplean" (Torres, 2016: 108).

Y para mayor desilusión, se obtuvo el hallazgo de que: "[en lo] referido a la forma de selección de las unidades de análisis (…) [Aunque]son las investigaciones cuantitativas -por mucho- las que declaran emplear muestras estadísticas (84,2% del total de reportes) (…) existen también investigaciones cualitativas que las emplean (…)" (Torres, 2016: 107).

Hay que recordar que se están mencionando aquí solo confusiones y deficiencias metodológicas asociadas a *las muestras estadísticas simples*, que es el único tipo de muestra aleatoria presente en los 1 377 reportes de investigación (de los últimos casi 20 años de trabajo científico en Cuba) que pudieron ser sometidos a análisis crítico en esa meta-investigación (Torres, 2016).

Nada que ver con las *muestras estadísticas complejas*, en lo que a grado de laboriosidad se refiere, y que son las que demandan las investigaciones cercanas a la "frontera de la ciencia", como sucede en los estudios *ERCE*, del Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (*LLECE*).

Y si bien *R* (aquí, a través de *RStudio*) le permite a sus usuarios diseñar y obtener tanto muestras aleatorias de un tipo como del otro, se ha optado en el presente trabajo por hacer foco en las de carácter complejo, y mostrar su utilidad para resolver problemas de naturaleza científica de notable vigencia y pertinencia para las aspiraciones y

ISSN: 1605– 5888 RNPS: 1844
No.1 (enero-abril), Año 2020, 4ta Época.

compromisos del Sistema Nacional de Educación cubano, como es el de la estimación de muestras aleatorias para procesos de investigación educativa que pretenden ir más allá de unidades organizacionales pequeñas, como grupos docentes o instituciones escolares aisladas, aun cuando en ellas también serían aplicables.

DESARROLLO

Se debe comenzar señalando que la *Teoría del Muestreo*, que da sustento al tratamiento de las muestras aleatorias, es una rama de la Estadística sumamente profunda y compleja como para pretender exponerla en toda su extensión.

De modo que aquí se hará referencia solo a rudimentos muy básicos de ella, suficientes para buscar un acercamiento a la respuesta de la pregunta arriba formulada y, con su pretexto, mostrar las potencialidades del *entorno R* también en este componente ineludible, al menos, de las investigaciones científicas que siguen el enfoque cuantitativo; justamente las que predominan en el quehacer investigativo educacional cubano, según se verificó en (Torres, 2016).

Aún con estas acotaciones, resultará inevitable hacer alusión a conceptos básicos de la Teoría del Muestreo, como son los de: marco muestral, método estadístico de muestreo, diseño muestral, asignación de la muestra, factor de expansión y estimación de parámetros. Mas, buscando asequibilidad en su comprensión, no se caracterizarán todos ellos de golpe, sino en la medida en que sean necesarios para avanzar en el proceso de búsqueda de la respuesta al problema de la eventual predicción de la muestra de Cuba en el *ERCE-2019.*

Eso sí, se da por sentado que el lector conoce la diferencia entre los conceptos estadísticos de *población* y de *muestra*, y sobre el carácter básico de ambos en la llamada *Estadística Inferencial*, justo aquella en la que se pretende “hablar” no solo de las unidades de análisis estudiadas, como parte del proceso investigativo, sino también de aquellas otras similares que las contienen y no pudieron ser directamente “observadas”; es decir, realizar inferencias estadísticas desde los estadígrafos

ISSN: 1605– 5888 RNPS: 1844
No.1 (enero-abril), Año 2020, 4ta Época.

muestrales hacia sus respectivos parámetros poblacionales, siempre con una determinada probabilidad, que lo ideal es que sea muy alta.

Pero no se dilatará más el inicio del trabajo con *RStudio* en el empeño de predecir la muestra cubana en el *ERCE-2019*. Para ello se utilizarán los *paquetes* específicos *"survey", "sampling" y "dplyr",* los que deberán ser previamente cargados en la consola de *RStudio,* como se muestra en el siguiente *chunk*, o segmento de secuencia de programación.

```
setwd("C:/datos")
library("survey")
library("sampling")
library("dplyr")
```

Cargadas las *bibliotecas* apropiadas, sigue ahora la activación de la base de datos que hace referencia a la población sujeta a estudio. Ahora bien, suele suceder que no todas las unidades de análisis que conforman la población pueden (o interesan) ser estudiadas. Por ejemplo, en el caso del *ERCE-2019* no se considerarán las escuelas especiales (pues se requeriría de apoyos adicionales para concretar la participación de los estudiantes con NEE), como tampoco el 2% de las escuelas más pequeñas de cada país.

La base de datos que contiene las unidades de análisis de la población, salvo las exclusiones prefijadas, se le suele denominar *marco muestral*. En el ejemplo que se sigue en el presente trabajo, que es un hecho real, se utilizó el marco muestral proporcionado por el Mined para el estudio piloto del *ERCE-2019*, de mayo del 2018; ello, en espera de su actualización para el presente curso escolar. El chunk siguiente permite realizar ese paso primario, además de ordenar las escuelas que integran la base de datos por el código nacional único, además de fijar sus variables de estudio y mostrar un resumen de ellas, como se presentará más abajo, precedidos de la simbología "##".

ISSN: 1605– 5888 RNPS: 1844
No.1 (enero-abril), Año 2020, 4ta Época.

```
dfDOPI2018=read.table("C:/datos/dfDOPI2018.csv", header = T, sep =",")
dfDOPI2018 <-dfDOPI2018[order(dfDOPI2018$id_esc_nac), ]
attach(dfDOPI2018)
str(dfDOPI2018)

## 'data.frame':    6654 obs. of  15 variables:
##  $ id_esc_nac  : int  21012050 21012160 21012161 21012168 21012169 21012170
21012183 21012184 21012187 21012194 ...
##  $ nom_esc     : Factor w/ 3614 levels " S/ I Martha Abreu Arencibia",..: 2906 528
3250 537 1837 1228 1289 656 2939 753 ...
##  $ ubi1_cen_esc: Factor w/ 16 levels "Artemisa","Camagüey",..: 13 13 13 13 13 13
13 13 13 13 ...
…..
```

¡Un alto para reducir tensiones entre los noveles!… La información devuelta por *RStudio,* al ejecutar la última instrucción del chunk anterior, lo que está destacando es que el *marco muestral*está conformado por 6 654 unidades de análisis (escuelas primarias) y 15 variables o atributos caracterizadores de ellas. Después comienza a relacionar las variables y su tipo: “$id_esc_nac" como números enteros ("int") que determinan el código nacional único, "$nom_esc” como los nombres de las escuelas en forma de *factor* con 3 614 niveles diferentes, “$ubi1_cen_esc” como nombres de las provincias, también en forma de factor (ahora de 16 niveles), etc.

Antes de seguir adelante con la selección de un método estadístico de muestreo, se deberán hacer todavía algunos procesos asociados a la base de datos del marco muestral. Nótese,por ejemplo, que no se dispone de una variable numérica que designe a las provincias, pues esa función la realiza un *factor de caracteres* (“$ubi1_cen_esc”), el cual dificultaría operaciones aritméticas, eventualmente necesarias, más adelante.

Una solución racional a ese obstáculo es obtener ese código numérico, representativo de cada provincia, a partir del código único de cada escuela, toda vez que este último inicia con dos dígitos que corresponden, justamente, a esa división político-

ISSN: 1605– 5888 RNPS: 1844
No.1 (enero-abril), Año 2020, 4ta Época.

administrativa. En el siguiente chunk se presentará la sub-rutina programada por este autor para lograr ese efecto, conservando su producto en el *objeto de R* denominado "IdProv".

```
IdProv<-(signif(dfDOPI2018$id_esc_nac,2)*10)/10^8
IdProv<-round((IdProv*10),2)
dfDOPI2018$IdProv <-IdProv
IdProv
```

De haberse programado, aparecerían tras la corrida varias decenas con los primeros valores generados, pero no se previó así por razones de espacio. Todos ellos fueron agregados en una nueva columna de la base de datos *"dfDOPI2018",* bajo la denominación *"IdProv".* Así, a las escuelas de "Pinar del Río" le corresponderán, en lo adelante, el "IdEsc" número 21, a las de "Artemisa" el número 22, a las de "La Habana" el número 23, y así sucesivamente hasta el número 35 que corresponderá a "Guantánamo". A las del municipio especial "Isla de la Juventud" les corresponderá el "IdEsc" número 40, respetando la voluntad de los autores de la base de datos.

Para más seguridad, se convertirá ese nuevo objeto de R en un vector de números enteros, con una sub-rutina más. No hay que alarmarse por la dilación del trabajo con el muestreo, propiamente dicho; la "acomodación" de las bases de datos previamente al cumplimiento del objetivo principal trazado no solo es usual, sino además aconsejable.

```
dfDOPI2018 <-dfDOPI2018%>%
mutate(IdProv=as.integer(IdProv))
```

Hecho esto (y registrado el cambio en la base de datos "dfDOPI2018"), llegó el momento de analizar el segundo concepto de la Teoría del muestreo requerido: el *método estadístico de muestreo*. En el *LLECE* está definido que para el *ERCE-2019* se empleará un método de *muestreo aleatorio bietápico* (o sea, en dos etapas, una para seleccionar escuelas y la otra para elegir grupos docentes); en la primera fase se

ISSN: 1605– 5888 RNPS: 1844
No.1 (enero-abril), Año 2020, 4ta Época.

utilizará el tipo de muestreo *estratificado óptimo* y en la segunda el de selección *porconglomerados.*

Puesto que a los efectos de este trabajo lo que interesa es la primera etapa, no se redundará en el muestreo por conglomerado (de las aulas). Basta conocer que, una vez definida la muestra de la primera etapa (las escuelas seleccionadas dentro de cada provincia), la selección de las siguientes unidades de análisis (las aulas dentro de estas últimas) puede realizarse, o bien a través de un *muestreo aleatorio simple* (a razón de un grupo por cada grado de interés, por ejemplo), o sencillamente se toman todas la unidades de análisis de cada escuela elegida en la primera fase; en la utilización de una de esas dos posibilidades consiste precisamente el *muestreo por conglomerados*, toda vez que todas las unidades (aulas) están dispuestas (dentro de cada escuela) de manera natural.

Por tanto, se estará hablando -en lo adelante- esencialmente de *muestreo aleatorioestratificado*. En esta otra modalidad del método estadístico de muestreo, la base del procedimiento consiste en particionar el marco muestral en grupos homogéneos de unidades de análisis, de manera que -al mismo tiempo- sean agrupaciones diferentes entre sí (Ochoa, 2015); es lo que se conoce en Teoría del muestreo como *estratos.*

Pues bien, las escuelas primarias de cada provincia pueden ser consideradas estratos, atendiendo a determinados atributos propios (disposición geográfica de las escuelas, singularidades del gobierno provincial, cobertura docente, tradiciones y cultura local, etc.).

Esta variante de muestreo retrotrae el trabajo al punto anterior. Es decir, hay que laborar nuevamente sobre la base de datos del marco muestral, esta vez para disponer (en una nueva base de datos, que aquí se denominará “Provincias”) los 16 estratos a tener en cuenta en la resolución del problema asumido.

Con el siguiente chunk se creará esa nueva base de datos, de manera que cuente el número de escuelas disponibles en el marco muestral por provincias (o sea, por

ISSN: 1605– 5888 RNPS: 1844
No.1 (enero-abril), Año 2020, 4ta Época.

estratos). Ciertamente, los estratos se conforman habitualmente de acuerdo con una de las variables características de la población; en esta ocasión se ha ponderado la variable "can_estu3" (cantidad de estudiantes en 3er. grado), que es una *variable numérica continua.*

```
Provincias<-dfDOPI2018%>%
select(IdProv, can_estu3)%>%
group_by(IdProv)%>%
summarise(n=n(),m=mean(can_estu3))%>%
mutate(p=n/sum(n))
str(Provincias)

## Classes 'tbl_df', 'tbl' and 'data.frame':    16 obs. of  4 variables:
##  $ IdProv: int  21 22 23 24 25 26 27 28 29 30 ...
##  $ n     : int  371 216 463 135 257 403 208 279 232 425 ...
##  $ m     : num  15.1 24.3 40.6 26.8 27 ...
##  $ p     : num  0.0558 0.0325 0.0696 0.0203 0.0386 ...
```

El nuevo objeto de R, denominado "Provincias" es un *dataframe* (es decir, una base de datos) que está compuesta por 16 casos (las 15 provincias y el municipio especial "Isla de la Juventud") y 4 variables (el identificador de cada provincia, creado anteriormente ["IdProv"], el número de escuelas por provincias ["n"], el promedio de estudiantes de 3er. grado por escuelas de cada provincia ["m"], y la proporción que representa el número de escuelas de cada provincia con relación al total del marco muestral ["p"]). En él se podría apreciar que hay cuatro provincias con más de 500 escuelas en su marco muestral: Holguín, Granma, Santiago de Cuba y Guantánamo; los tamaños de las tres primeras superan por mucho a los tamaños de los restantes estratos del país.

Aclarado lo anterior, se regresará entonces a la tarea central de pretender predecir la muestra de Cuba en el *ERCE-2019*, a partir del marco muestral disponible. Es obvio que un *muestreo aleatorio simple* no es una solución adecuada a ese problema, pues no hay garantía de que las características del marco muestral queden adecuadamente representadas en la muestra extraída con ese método estadístico de muestreo, al

ISSN: 1605– 5888 RNPS: 1844
No.1 (enero-abril), Año 2020, 4ta Época.

punto que puede suceder que -por ejemplo- existan provincias que no queden representadas en la muestra, o que la proporción de escuelas urbanas y rurales no sean las adecuadas para ciertas provincias, etc.

Véase la siguiente muestra aleatoria simple, contenida en el objeto de R "ProvinciasMAS1":

```
N <-length(dfDOPI2018$IdProv)
MAS <-sample(1:N, 400)
MAS1 <-dfDOPI2018[MAS,]
ProvinciasMAS1 <-MAS1%>%
select(IdProv, ubi1_cen_esc, can_estu3)%>%
group_by(ubi1_cen_esc)%>%
summarise(n=n(), prom3=round(mean(can_estu3), 0))%>%
mutate(p=round(n/sum(n)*100, 1))
ProvinciasMAS1
```

```
## # A tibble: 16 x 4
##    ubi1_cen_esc          n prom3     p
##    <fct><int><dbl><dbl>
##  1 Artemisa             11    18   2.8
##  2 Camagüey             24    13   6
##  3 Ciego de Ávila       14    20   3.5
##  4 Cienfuegos           16    24   4
##  5 Granma               47    14  11.8
##  6 Guantánamo           31    10   7.8
##  7 Holguín              57    11  14.2
##  8 Isla de la Juventud   1    14   0.2
##  9 La Habana            37    35   9.2
## 10 Las Tunas            32    15   8
## 11 Matanzas             12    50   3
## 12 Mayabeque             9    42   2.2
```

ISSN: 1605– 5888 RNPS: 1844
No.1 (enero-abril), Año 2020, 4ta Época.

```
## 13 Pinar del Río          30    18   7.5
## 14 Sancti Spíritus        14    17   3.5
## 15 Santiago de Cuba       39    15   9.8
## 16 Villa Clara            26    15   6.5
```