CIENCIAS PEDAGÓGICAS
Revista electrónica científico-pedagógica
ISSN: 1607 – 5888 RNPS: 1844
No.2 (mayo-agosto), Año 2020, 4ta Época.
Págs. 84-91

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[1]. Investigador Titular.

paul@rimed.cu https://orcid.org/0000-0002-7862-2737

Instituto Central de Ciencias Pedagógicas. La Habana, Cuba.

Recibido: 12 de marzo 2020 *Aprobado: 11 de abril de 2020*

RESUMEN

En este artículo se le dará continuidad al análisis de las posibilidades del entorno informático R de contribuir al perfeccionamiento de la investigación cubana 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

[1] Coordinador Nacional por Cuba del Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE), de la OREALC-UNESCO.

ISSN: 1607 – 5888 RNPS: 1844
No.2 (mayo-agosto), Año 2020, 4ta Época.
Págs. 84-91

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, after the manipulation 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.

… Continuación del número anterior

Dadas las limitaciones de la extracción de “muestras aleatorias simples” en estudios de investigación como estos, se suele emplear -en su lugar- **“muestras estratificadas”**. Para el **muestreo aleatorio estratificado** existen tres formas diferentes de **afijación** de las muestras: de forma uniforme (es decir, asignando igual cantidad de unidades de análisis para cada estrato), proporcional a los tamaños de los estratos, o de forma estratificado óptimo (o sea, atendiendo a la desviación estándar de la variable de interés en cada estrato). Cada una de esas variantes presentan ventajas y desventajas (Ochoa, 2015).

En el presente artículo se trabajará solo la afijación de la muestra **estratificada de forma uniforme** en los estratos. Por ejemplo, se tomarán la cantidad fija de 30 escuelas en cada provincia. Para ello será necesaria la siguiente sub-rutina.

```
size <-rep(25, 16)
ESU1 <-strata(dfDOPI2018, c("IdProv"), size=size, method ="srswor")
str(ESU1)

## 'data.frame':    400 obs. of  4 variables:
##  $ IdProv : int  21 21 21 21 21 21 21 21 21 21 ...
##  $ ID_unit: int  17 28 44 45 56 67 88 117 122 124 ...
##  $ Prob   : num  0.0674 0.0674 0.0674 0.0674 0.0674 ...
##  $ Stratum: int  1 1 1 1 1 1 1 1 1 1 ...
```

ISSN: 1607 – 5888 RNPS: 1844
No.2 (mayo-agosto), Año 2020, 4ta Época.
Págs. 84-91

Se ha obtenido, así, una primera **muestra estratificada** (esta vez **uniforme**) de tamaño 400 (o sea, a razón de 25 de cada "provincia"). El resultado se ha plasmado en una data frame denominado "ESU1" que, como se aprecia arriba, cuenta de cuatro variables: "$IdProv", "ID-unit", "Prob" y "Stratum". La primera de ella es ya conocida (es un identificador numérico único, por "provincias""), creada por el autor como recurso auxiliar.

En cambio las otras tres fueron generadas espontáneamente por **RStudio**, pues (a través de la **biblioteca* "survey"**) comprendió que se deseaba generar una muestra estratificada de tamaños iguales. La **variable "ID-unit"** señala el número de orden que ocupaba la escuela elegida en la base de datos del marco muestral (recuérdese, "dfDOPI2018"); mientras que **variable "Prob"** hace referencia a la probabilidad de haber sido elegido, que es la misma para todas las **unidades muestrales** en este caso, pues todos los estratos son del mismo tamaño; por último, **variable "Stratum"** indica a cuál de los 16 estratos pertenece cada una de las escuelas seleccionadas como parte de esta (primera) muestra.

Ahora bien, el data frame "ESU1" no "arrastra" consigo al resto de las variables del marco muestral "dfDOPI2018", cuando interesa mucho estudiar el comportamiento de importantes variables de este último (como "can_estu3" [cantidad de estudiantes en 3er. grado], "sec_3" [cantidad de aulas de 3er. grado], etc.). Se necesita, por tanto, "mezclar" ambas bases de datos; ello se logra con el siguiente chunk:

```
MuestraESU1=getdata(dfDOPI2018, ESU1)
str(MuestraESU1)

## 'data.frame':    400 obs. of  19 variables:
##  $ id_esc_nac  : int  21012649 21022082 21022131 21022135 21032416 21032463
21042212 21052047 21052894 21052900 ...
##  $ nom_esc     : Factor w/ 3614 levels " S/ I Martha Abreu Arencibia",..: 3604 3388 2552
2136 648 3084 2372 1096 1249 1514 ...
##  $ ubi1_cen_esc: Factor w/ 16 levels "Artemisa","Camagüey",..: 13 13 13 13 13 13 13 13
13 13 ...
##  $ ubi2_cen_esc: Factor w/ 168 levels "Abreus","Aguada de Pasajeros",..: 148 97 97 97
109 109 165 84 84 84 ...
```

ISSN: 1607 – 5888 RNPS: 1844
No.2 (mayo-agosto), Año 2020, 4ta Época.
Págs. 84-91

```
##  $ can_estu3   : int  122 8 2 4 5 7 4 6 1 12 ...
##  $ can_estu6   : int  101 10 5 11 6 6 7 2 4 9 ...
##  $ est_hom_3   : int  61 5 0 4 2 3 1 3 1 8 ...
##  $ est_muj_3   : int  61 3 2 0 3 4 3 3 0 4 ...
##  $ est_hom_6   : int  57 5 2 9 4 3 3 0 2 3 ...
##  $ est_muj_6   : int  44 5 3 2 2 3 4 2 2 6 ...
##  $ sec_3       : int  6 1 1 1 1 1 1 1 1 1 ...
##  $ sec_6       : int  6 1 1 1 1 1 1 1 1 1 ...
##  $ can_prof_3  : int  6 1 1 1 1 1 1 1 1 1 ...
##  $ can_prof_6  : int  6 1 1 1 1 1 1 1 1 1 ...
##  $ area_esc    : int  1 2 2 2 2 2 2 2 2 2 ...
##  $ IdProv      : int  21 21 21 21 21 21 21 21 21 21 ...
##  $ ID_unit     : int  17 28 44 45 56 67 88 117 122 124 ...
##  $ Prob        : num  0.0674 0.0674 0.0674 0.0674 0.0674 ...
##  $ Stratum     : int  1 1 1 1 1 1 1 1 1 1 ...
```

Se trata solo de una de las muchas muestras posibles de extraer, con esas características. La biblioteca **"survey"** de **R** permite ir más allá y elaborar **diseños muestrales** (otra de las categorías de la Teoría del Muestro arriba destacadas). Con ellos se pueden realizar **estimaciones de parámetros** de la población acordes con las muestras obtenidas a partir de dicho diseño. Se ilustrarán estos aspectos con la forma de afijación de la muestra anteriormente desarrollada. Primero, el diseño de muestreo con la **función "svydesign"**:

```
ESU1desing <-svydesign(id=~1, strata=~ubi1_cen_esc, data = MuestraESU1, weights =
~Prob)
```

Y ahora, sobre su base, se realizarán las estimaciones de los parámetros: total de la cantidad de estudiantes del 3er. grado y total de aulas de ese grado, por provincias:

```
svytotal(~can_estu3, ESU1desing)
svytotal(~sec_3, ESU1desing)
svyby(~can_estu3, ~sec_3, ESU1desing, svytotal)
```

ISSN: 1607 – 5888 RNPS: 1844
No.2 (mayo-agosto), Año 2020, 4ta Época.
Págs. 84-91

```
##           total     SE
## can_estu3 1367.3 128.41
##       total     SE
## sec_3 80.053 5.0158
##   sec_3  can_estu3        se
## 0     0   0.000000   0.000000
## 1     1 282.215233  32.936281
## 2     2 414.956041  80.808426
## 3     3 429.153705 122.826146
## 4     4 203.438813  98.931757
## 5     5  29.281169  17.479425
## 6     6   8.221024   8.221024
```

También se pueden realizar estimaciones poblacionales de utilidad de los promedios/desviaciones estándar de la cantidad de estudiantes y de aulas del 3er. grado, seleccionadas por escuelas dentro de las provincias:

```
ESU1xProv <-as.data.frame (svyby(~can_estu3+~sec_3,~ubi1_cen_esc, ESU1desing,
svymean))
str(ESU1xProv)
```

```
## 'data.frame':    16 obs. of  5 variables:
##  $ ubi1_cen_esc: Factor w/ 16 levels "Artemisa","Camagüey",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ can_estu3   : num  33 20.8 31.7 21.5 9.8 ...
##  $ sec_3       : num  1.6 1.36 1.72 1.4 0.96 1 1.24 1.8 1.92 1.16 ...
##  $ se.can_estu3: num  5.71 5.2 7.55 4.87 2.78 ...
##  $ se.sec_3    : num  0.163 0.172 0.242 0.141 0.108 ...
##  - attr(*, "svyby")=List of 7
##   ..$ margins  : int 1
##   ..$ nstats   : num 2
##   ..$ vars     : int 1
##   ..$ deffs    : logi FALSE
```

ISSN: 1607 – 5888 RNPS: 1844
No.2 (mayo-agosto), Año 2020, 4ta Época.
Págs. 84-91

```
##   ..$ statistic: chr "svymean"
##   ..$ variables: chr  "can_estu3" "sec_3"
##   ..$ vartype  : chr "se"
##  - attr(*, "call")= language svyby.default(~can_estu3 + ~sec_3, ~ubi1_cen_esc,
ESU1desing, svymean)
```

De más interéses determinar los **intervalos de confianza** de algunos de esos parámetros; es decir, la determinación de los valores mínimos y máximos entre los cuales se ubicará el parámetro buscado, con una alta probabilidad. En el chunk que sigue se determina el intervalo de confianza del promedio de estudiantes por escuelas elegidas dentro de las provincias, que se extiende de 20 a 30 estudiantes, como puede apreciarse en la última línea del **chunk** siguiente:

```
prom_est <-svymean(~can_estu3, ESU1desing, deff=TRUE, na.rm=TRUE)
confint(prom_est)

##              2.5 %   97.5 %
## can_estu3 21.91207 31.79937
```

Lo presentado hasta aquí es apenas es un anticipo de las muchas cosas que se pudieran hacer con **R**, en materia de selección de muestras aleatorias, además de suficientemente representativas de la población de la que se extraen (algo imprescindible para la generalización de los **resultados estadísticos** finales de la investigación); así como también en términos de inferencia de sus estadígrafos muestrales a los parámetros poblacionales(valga la redundancia conceptual).Queda pendiente la ilustración de otras formas de afijación de muestras estratificadas, e incluso de otros métodos de muestreo más potentes, como el **sistemático**(mejor aún, con arranque en un número aleatorio **sembrado con una semilla**). Pero ello no será posible en este primer trabajo. Quedará pendiente para otros que habrán de continuarle.

ISSN: 1607 – 5888 RNPS: 1844
No.2 (mayo-agosto), Año 2020, 4ta Época.
Págs. 84-91

CONCLUSIONES

Con el presente artículo se ha pretendido ahondar en la importancia y fortalezas del **entorno informático R** para el perfeccionamiento de la investigación educativa cubana; en esta ocasión, desde la perspectiva de las **muestras estadísticas aleatorias**.

El lector, seguramente, habrá podido percibir que, si bien el lenguaje de programación empleado por **R** resulta, a primera vista, algo incómodo, los recursos de ese poderoso entorno informático son de gran valía y utilidad para la investigación científica, pues son capaces de generar productos y análisis estadísticos profundos y aportadores, con relativa facilidad. El complemento a esta aproximación estaría en el desarrollo de **cursos introductorios del manejo de R** y, claro, mucho estudio independiente y perseverancia por parte de los iniciados.

Al mismo tiempo, se espera que el lector haya podido consolidar nociones epistemológicas esenciales de la actividad investigativa, como que **en el quehacer científico no existen grandes y cómodas calzadas; todo lo contrario, está plagado de caminos angostos y empedrados**. Debiera tomarse distancia de quienes aseguren haber obtenido resultados científicos trascendentes con recursos metodológicos simples y en plazos de tiempo relativamente cortos.

No obstante, y en lo relativo a lo aquí tratado, hay que tener en cuenta que **el entorno informático R no es una bola de cristal**; se pueden obtener sobre su base nociones (escenarios) de una probable muestra representativa de una población; la muestra definitiva seguramente no la tienen ni Zeus, ni su hija Atenea, diosa de la sabiduría. Como se conoce, las muestras se apoyan en números aleatorios, propio de la modelación de procesos nada predecibles. Esta situación se agudiza en el caso de las llamadas **muestras complejas**, que son las que utilizan los estudios internacionales de evaluación educativa, como ocurre con los **ERCE** del **LLECE**. No hay nada seguro en torno a ellas, pero tampoco se ubican totalmente en la zona de lo ignoto; especialmente, si existe la voluntad de aproximarse a ellas,… con la ayuda incalculable de **R**.

ISSN: 1607 – 5888 RNPS: 1844
No.2 (mayo-agosto), Año 2020, 4ta Época.
Págs. 84-91

BIBLIOGRAFÍA

Campirán, E. (2016). *Muestreo estratificado en R (parte 2)* (Recuperado de: https://www.youtube.com/watch?v=lgzLVIyDeDw)

Maguiña, M. E. (2016). *Muestreo estratificado en R* (Recuperado de: https://www.youtube.com/watch?v=302NJmbz9Pk&t=232s)

Ochoa, C. (2015). *Muestreo probabilístico: muestreo estratificado.*

Ortiz, E. (2015). *Problemas que afectan la calidad de las tesis doctorales en Ciencias Pedagógicas.* Pedagogía Universitaria 20(2), 23-38.

Torres, P. (2016). *Retos de la investigación educativa cubana actual. Aportes a su tratamiento.* La Habana, Cuba: Universidad en Ciencias Pedagógicas "Enrique J. Varona". (http://www.cubaeduca.cu/media/www.cubaeduca.cu/medias/evaluador/tesis2dograd o.pdf)

_____ (2018). *Lo que todo investigador educativo cubano debiera conocer: el entorno informático R.* Atenas 4(44), 1-27. (Recuperado de: http://atenas.mes.edu.cu)

Cómo citar el artículo:
Torres, P (2020, mayo). Más del entorno informático R en la investigación educativa cubana: ¿Se puede predecir la muestra de cuba en el erce 2019? Ciencias Pedagógicas, No.2 (2020).p. 84. www.cienciaspedagogicas.rimed.cu