

Bogotá D.C
Noviembre 2025
ISSN: 2590 - 4663
Publicación Trimestral


Educación


Icfes

SABER AL DETALLE

Edición 18

Réplicas Repetidas Balanceadas (BRR)



Saber Al Detalle: Réplicas Repetidas Balanceadas (BRR)

Presidente de la República
Gustavo Francisco Petro Urrego

Ministro de Educación Nacional
José Daniel Rojas Medellín

Directora General Icfes
Elizabeth Blandón Bermúdez

Director Técnico de Evaluación
Gustavo Andrés Monsalve Londoño

Subdirector de Diseño de Instrumentos
Heider Martínez Mena

Subdirector de Estadísticas
Cristian Fabián Montaña Rincón

Subdirectora de Análisis y Divulgación
Alejandra Neira Aroca

Elaboración del documento
Ana María Mondragón Moreno

Revisado por
John Alexander Calderón Rodríguez

Diseño y diagramación
Linda Nathaly Sarmiento Olaya

Bogotá D.C., noviembre de 2025

Todos los derechos de autor reservados ©.

¿Qué son las réplicas repetidas balanceadas y cómo se aplican en el Icfes?

En el contexto del análisis de encuestas y evaluaciones basadas en diseños muestrales complejos, las réplicas repetidas balanceadas de muestreo, BRR por sus siglas en inglés, constituyen una herramienta fundamental para la estimación precisa de errores estándar y otras medidas de incertidumbre. Estas réplicas consisten en subconjuntos o reponderaciones sistemáticas a los factores de expansión de la muestra original, diseñadas para reflejar la variabilidad inherente al proceso de selección muestral.

El uso de réplicas permite aplicar métodos de remuestreo, como BRR o Jackknife, los cuales generan estimaciones de varianza más realistas al incorporar la estructura del diseño muestral (por ejemplo, estratificación, conglomeración y pesos desiguales). Estos métodos son especialmente útiles en estudios donde la complejidad del diseño impide la aplicación de fórmulas analíticas convencionales para la estimación de la varianza.

Diseños y pesos muestrales en evaluaciones educativas

Cuando el propósito de una evaluación educativa de gran escala es diagnosticar el desempeño de los estudiantes en dominios específicos, por ejemplo, desagregaciones territoriales, grupos étnicos, categorías de discapacidad, etc., en componentes como matemáticas, lectura, habilidades socioemocionales y otros, un censo resulta inviable por su costo, duración y complejidad

logística. Por ello se recurre a muestras probabilísticas de estudiantes, que permiten obtener estimaciones representativas para cada dominio y controlar las probabilidades de inclusión, reduciendo el riesgo de sesgos. En la práctica, las muestras aleatorias simples, donde cada estudiante se selecciona directamente y con igual probabilidad, son poco frecuentes. Esto se debe a que implican costos muy elevados, una operatividad compleja y dificultades para vincular la información individual de los estudiantes con datos contextuales de los colegios en los cuales están matriculados.

Por ello, estas evaluaciones suelen emplear diseños muestrales complejos que combinan varias técnicas: estratificación de la población (subgrupos homogéneos, por ejemplo, zonas urbanas y rurales), selección en dos o más etapas (primero colegios y luego estudiantes dentro de los colegios) y probabilidades de inclusión proporcionales al tamaño de la unidad (PiPT). Estos diseños permiten mejorar la precisión de las estimaciones, optimizar los recursos y facilitar la organización del operativo de campo.

Un componente esencial de estos diseños son los pesos muestrales, que permiten que las estimaciones reflejen correctamente la estructura del muestreo. Estos pesos corrigen por las distintas probabilidades de inclusión y por eventuales pérdidas de información (no respuesta de colegios o estudiantes). De este modo, se evita que los resultados estén sesgados y se garantiza que sean

representativos de la población objetivo. Por esta razón, estudios como PISA o las pruebas Saber 3°, 5°, 7° y 9° aplican y reportan sus resultados usando estos pesos.

Adicionalmente, dado que cada muestra implica cierto grado de incertidumbre, es importante calcular el error de muestreo, que mide la precisión de las estimaciones obtenidas. En el caso de diseños muestrales complejos, este cálculo requiere métodos específicos, como los pesos de réplica. Estos métodos permiten estimar de manera más precisa la variabilidad de las estimaciones, reflejando la complejidad real del diseño del estudio.

¿Qué son los pesos de muestreo?

Los pesos de muestreo se utilizan para que los resultados obtenidos a partir de una muestra representen adecuadamente las características de toda la población, garantizando inferencias válidas a nivel poblacional. Los pesos base, o pesos de diseño, se calculan cuando la muestra es probabilística y se extrae de una población finita. Según lo definido en Särndal et al. (1992, cap. 1) y en Valliant et al. (2018), una muestra probabilística se obtiene cumpliendo las siguientes cuatro condiciones:

1. Se puede definir el conjunto de todas las muestras $S = \{s_1, s_2, \dots, s_M\}$ que pueden seleccionarse de una población finita U mediante un procedimiento de muestreo especificado.
2. A cada posible muestra s de S se le asocia una probabilidad de inclusión conocida $p(s)$.
3. Todo elemento de la población objetivo tiene una probabilidad de inclusión

distinta de cero con el procedimiento aleatorio especificado.

4. Una muestra s^* se selecciona mediante un mecanismo aleatorio bajo el cual cada s^* en S recibe la probabilidad $p(s)$.

Una característica fundamental de un diseño probabilístico consiste en que todos los elementos dentro de la muestra tienen asociada una probabilidad π_k mayor a cero de ser seleccionados. Por lo anterior, se define el peso de muestreo o factor de expansión w_k como el inverso de las probabilidades de inclusión de los elementos y se pueden interpretar como el número de elementos fuera de la muestra que son representados por el k -ésimo elemento dentro la muestra.

$\pi_k = \text{Probabilidad de inclusión del elemento } k$.

$$w_k = 1/\pi_k$$

Algunos de los controles de los pesos estimados son (Valliant et al. 2018):

1. Las probabilidades de inclusión deben estar todas dentro del intervalo $(0, 1]$.
2. Los pesos base deben sumar el número total de elementos de la población, o a una estimación razonable de su tamaño. Se deben hacer comprobaciones similares de las sumas de pesos para los principales subgrupos (género, raza/etnia, establecimientos).

Los pesos muestrales se deben calcular en cada una de las etapas del diseño de muestreo en caso de tener un diseño

multietápico. Si llegase a ocurrir que la unidad de muestreo (establecimiento, sede o estudiante) decida no participar y esta no pueda ser remplazada, se debe incorporar un ajuste sobre los pesos de muestreo para compensar la no participación.

A continuación, se describe el cálculo de los pesos para un muestreo aleatorio simple (MAS) y un muestreo proporcional al tamaño (PiPT), para luego profundizar en los diseños de muestreo multietápicos, es decir, aquellos que implican varias etapas de selección.

Pesos de muestreo para un MAS

La probabilidad de inclusión de cada individuo dentro del marco para un MAS es igual para todos los individuos.

$$\pi_k = \frac{n}{N}$$

Donde n es el tamaño de la muestra a seleccionar y N el tamaño poblacional, así el peso de muestreo para cada uno de los individuos en un diseño MAS será:

$$w_k = \frac{N}{n}$$

Pesos de muestreo para un PiPT

Para un diseño muestral en el que la probabilidad de que una unidad primaria de muestreo sea seleccionada en la muestra es proporcional a su tamaño, la probabilidad de inclusión está dada por:

$$\pi_k = \frac{n * x_k}{X}$$

Donde n es el tamaño de la muestra a seleccionar, x_k es la unidad de tamaño de la unidad k (por ejemplo, total de estudiantes matriculados) y $X = \sum x_k$.

Finalmente, el factor de expansión está dado por:

$$w_k = \frac{X}{n * x_k}$$

Pesos de muestreo para un diseño multietápico

Usamos diseños de muestreo multietápicos cuando la población que queremos estudiar es demasiado grande y está muy dispersa. En lugar de intentar llegar a todos desde el principio, seleccionamos grupos grandes primero —como los colegios— y luego, dentro de ellos, escogemos a los estudiantes. Esto facilita el trabajo logístico y hace posible obtener resultados que representen a toda la población sin tener que medir a cada persona. En resumen, resulta más sencillo seleccionar un número limitado de colegios y, dentro de ellos, evaluar a los estudiantes, que intentar visitar una gran cantidad de colegios para medir estudiantes dispersos.

A las unidades seleccionadas en cada etapa se les da un nombre específico: las de la primera etapa se denominan unidades primarias de muestreo (UPM); las de la segunda etapa, unidades secundarias de muestreo (USM); y, si el diseño continúa en más niveles, se pueden identificar unidades terciarias y así sucesivamente, hasta llegar a la unidad final de observación, que corresponde a quienes efectivamente son medidos o encuestados.

En un diseño multietápico, cada unidad de análisis tiene una probabilidad de inclusión que es el producto de las probabilidades condicionadas en cada etapa $\pi^{(1)}$ para la unidad de primera etapa, $\pi^{(21)}$ para la sub-unidad dada la selección de la primera, etc. El peso base se define como el inverso de esa probabilidad acumulada:

$$w_k = \frac{1}{\pi_i^{(1)} \pi_{kl1}^{(2)} \dots}$$

Según Valliant et al. (2018) para multietapas la construcción de pesos sigue una receta secuencial: se arranca con el peso base del primer nivel (por ejemplo, el colegio), se multiplica por el inverso de la probabilidad condicional del segundo nivel (estudiante | colegio), y después se aplican, si procede, ajustes de elegibilidad desconocida, de no respuesta y de calibración, cada uno calculado sobre el peso resultante de la fase previa.

Ejemplo de diseño muestral y pesos en evaluación educativa

Supongamos que se desea estimar la habilidad de los estudiantes en Lenguaje para una región del territorio nacional. Se requiere que las estimaciones sean representativas por sector educativo (oficial y no oficial) y por zona de ubicación del colegio (rural o urbana). Para reflejar diferencias clave en el sistema educativo por estas variables, la población se organiza en tres estratos:

- » **Estrato 1:** colegios no oficiales.
- » **Estrato 2:** colegios oficiales urbanos.
- » **Estrato 3:** colegios oficiales rurales.

Esta estratificación permite representar adecuadamente tanto al sector privado como al oficial, diferenciando además a los colegios

oficiales según su zona de ubicación. De esta forma, se garantiza que cada uno de estos subgrupos esté cubierto en la muestra y que las comparaciones posteriores sean válidas.

En la primera etapa se seleccionan colegios (UPM) dentro de cada estrato definido. Dado que la probabilidad de ser incluido puede estar relacionada con la cantidad de matrícula que el colegio atiende, se utiliza un esquema proporcional al tamaño (PiPT). Esto significa que un colegio con más estudiantes matriculados tiene una mayor probabilidad de ser incluido en la muestra que uno más pequeño. En la segunda etapa, dentro de cada colegio seleccionado, se elige una muestra de estudiantes mediante muestreo aleatorio simple (MAS), de manera que todos los estudiantes del colegio tengan la misma probabilidad de ser seleccionados.

Diseño y asignación de la muestra

Muestra total de colegios (UPM):

Supongamos una asignación proporcional de tamaño de muestra proporcional al número de colegios en cada estrato M_h , es decir, los estratos con un mayor número de colegios tendrán una muestra proporcionalmente más grande.

- » **Estrato 1:** n_1
- » **Estrato 2:** n_2
- » **Estrato 3:** n_3

Cálculo de pesos

1) Primera etapa (selección de colegios dentro de cada estrato, PiPT)

Dentro de cada estrato (h), se seleccionan (n_h), es decir, los estratos con un mayor

número) colegios con **probabilidad proporcional a una medida de tamaño** (PiPT). Denotemos:

- » M_h : número de colegios en el estrato (h)
- » X_{hi} : medida de tamaño del colegio (i) en el estrato (h), es decir número de estudiantes matriculados en el colegio.
- » $X_h = \sum_{i=1}^{M_h} X_{hi}$

Probabilidad de inclusión (1.ª etapa):

$$\pi_{hi} = \frac{n_h X_{hi}}{X_h}$$

Por ende, su peso de muestreo en la primera etapa será:

$$w_{hi}^{(1)} = \frac{1}{\pi_{hi}}$$

2) Segunda etapa (selección de estudiantes dentro de cada colegio seleccionado)

En cada colegio (i) seleccionado del estrato (h), se elige una muestra de estudiantes por **MAS** (muestreo aleatorio simple). Sea:

- » N_{hi} : total de estudiantes en el colegio i del estrato h .
- » m_{hi} : número de estudiantes seleccionados en ese colegio

Probabilidad condicional (2.ª etapa):

$$\pi_{j|hi} = \frac{m_{hi}}{N_{hi}}$$

Por ende, el peso de muestro del estudiante en el colegio i del estrato h será:

$$w_{j|hi}^{(2)} = \frac{1}{\pi_{j|hi}}$$

3) Peso total (dos etapas)

Para un estudiante j en el colegio del estrato h :

$$w_{hij} = w_{hi}^{(1)} \times w_{j|hi}^{(2)}$$

El problema de la estimación de la varianza en encuestas complejas

En el análisis de encuestas con diseños muestrales complejos, es decir, diseños estratificados, multietápicos, entre otros; la estimación precisa de la varianza de los estimadores de un parámetro¹ constituye un desafío metodológico. A diferencia de los diseños muestrales simples, en los que la varianza puede calcularse mediante fórmulas convencionales, en los diseños complejos estas fórmulas no resultan aplicables debido a la dependencia entre observaciones dentro de conglomerados y al uso de pesos ajustados a la no respuesta.

La subestimación de la varianza es un riesgo frecuente si se ignora la estructura del diseño, lo que puede llevar a intervalos de confianza demasiado estrechos y a conclusiones estadísticamente inválidas. Para enfrentar esta dificultad, se han desarrollado técnicas de remuestreo que simulan la variabilidad muestral sin requerir la recolección de nuevas muestras.

¹ Un **parámetro** es una característica numérica de la población, generalmente desconocida. Un **estimador** es una estadística basada en la muestra que se utiliza para inferir el valor de ese parámetro. (Särndal et al., 1992)

Según Valliant et al. (2018) se presentan tres grandes familias de métodos para el cálculo del error:

1. **Fórmulas exactas:** aplicables a un número limitado de diseños y estimadores. Destaca el estimador de conglomerados últimos, exacto cuando las unidades primarias de muestreo (PSU) se seleccionan con reemplazo y buena aproximación cuando la fracción de muestreo es pequeña.
2. **Linealización (Taylor):** convierte estimadores no lineales en aproximaciones lineales para derivar varianzas a partir de las diferencias entre totales ponderados de las PSU. Resulta versátil cuando no existen fórmulas exactas.
3. **Métodos de remuestreo (Jackknife, Réplicas Repetidas Balanceadas y Bootstrap):** Generan réplicas de la muestra (submuestras de la unidad primaria de muestreo) y calculan la varianza a partir de la dispersión entre las réplicas.

Teniendo en cuenta las limitaciones de las fórmulas exactas y de la linealización en diseños muestrales complejos en evaluaciones educativas, nos centramos en el uso de técnicas de remuestreo a la que pertenece el BRR.

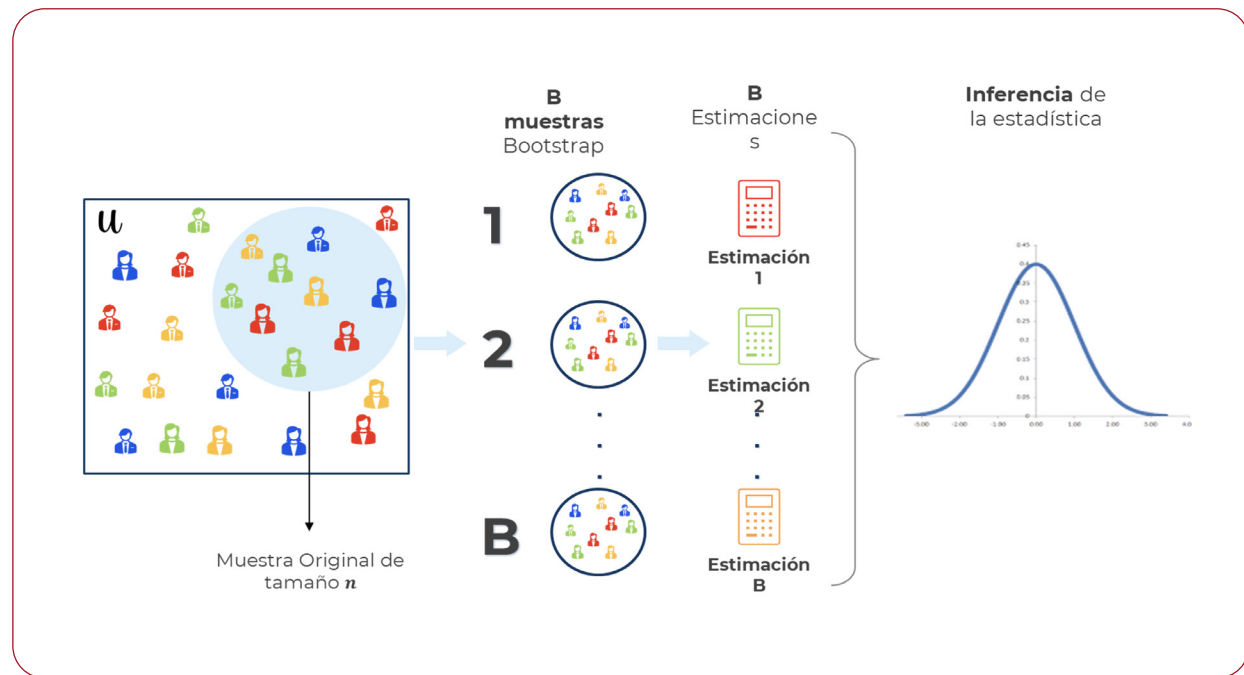
Técnicas de remuestreo

Los métodos de remuestreo son útiles para estimar intervalos de confianza, realizar pruebas de hipótesis, evaluar la capacidad predictiva de los modelos estadísticos, o mejorar la precisión de las estimaciones cuando la muestra es pequeña o cuando los datos tienen una distribución no normal o desconocida.

En el marco de la estimación del error de estimación o de la varianza del estimador, las técnicas de remuestreo consisten en generar submuestras de la unidad primaria de muestreo para calcular el error de estimación a partir de la dispersión entre las réplicas. El remuestreo implica tomar muestras aleatorias repetidas de los datos disponibles y utilizarlas para estimar estadísticas y evaluar la precisión de los resultados.

Bootstrap

El método de bootstrap es una técnica estadística utilizada para estimar la distribución de probabilidad de un estadístico a partir de una muestra de datos. La idea principal detrás del método es simular una gran cantidad de muestras de la población original a partir de la muestra disponible y utilizar estas muestras simuladas para estimar la distribución de probabilidad del estadístico de interés.



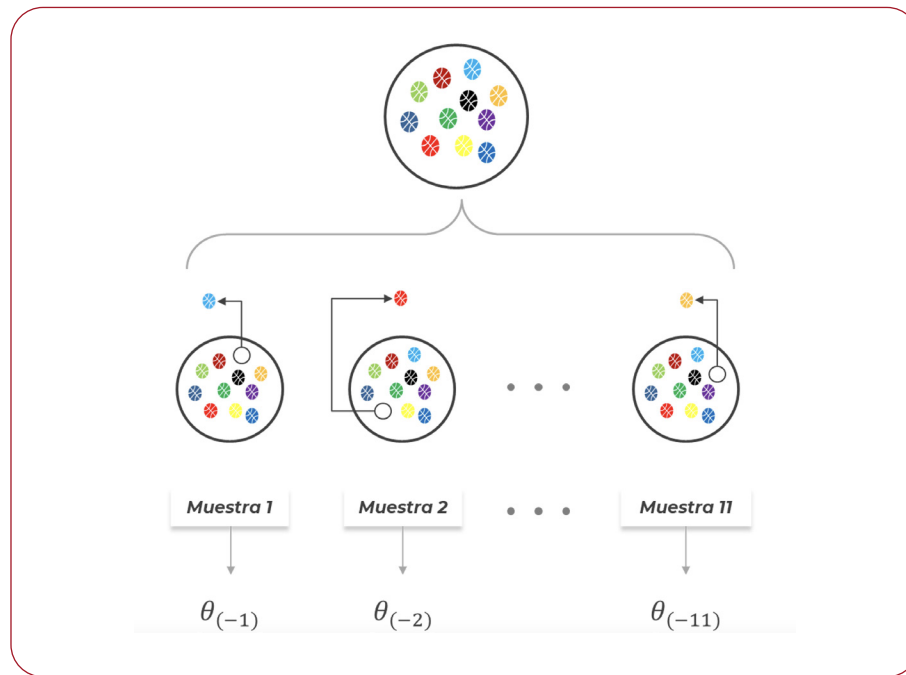
En Valliant (2018), se expone este método que parte de la muestra completa y, dentro de cada estrato, vuelve a seleccionar las unidades primarias de muestreo (UPM) con reemplazo y el mismo tamaño de las que aparecen originalmente. Cada UPM puede ser re-muestreada cero, una o varias veces; el conteo de cuántas veces fue elegida determina un factor de expansión específico. Con ese factor se construyen pesos “réplica” para todas las unidades de la primera etapa de muestreo, no sólo para las que quedaron en el subconjunto: la fórmula ajusta el peso original mediante una constante que depende del tamaño del estrato y del número total de UPM extraídas, de modo que el conjunto de pesos permanezca no negativo y conserve, en promedio, la estructura del diseño.

Una vez calculados los pesos de la réplica, se estima con ellos la estadística de interés.

El procedimiento se repite muchas veces (B replicaciones), cada vez con una nueva re-selección de UPM y sus correspondientes pesos. Al finalizar, se dispone de B estimaciones paralelas; la dispersión de todas esas estimaciones alrededor de la que se obtuvo con la muestra original y los pesos finales, se toma como estimador bootstrap de la varianza.

Jackknife

El método de remuestreo de Jackknife, también conocido como el método de “dejar uno fuera” (leave-one-out), es una técnica estadística cuya idea principal es repetir el análisis de los datos n veces, cada vez dejando fuera una observación diferente, para luego evaluar el impacto de cada observación en la estimación del estadístico.



En la versión clásica del Jackknife para muestreo de una sola etapa se genera una colección de réplicas eliminando, una a una, cada unidad de la muestra. Cuando se descarta la unidad i , a todas las restantes se les incrementa su peso original por el factor $n/(n-1)$ (donde n es el tamaño muestral) de modo que la suma de los pesos siga representando al universo completo. Con esos pesos reajustados se recalcula el total (o la estadística) de interés y se guarda como la estimación de la réplica correspondiente.

El procedimiento se repite para las n unidades, obteniéndose así n estimaciones replicadas. Para cuantificar la variabilidad debida al muestreo, se comparan estas n estimaciones con la que se obtuvo usando todos los datos: se mide cuánto se desvían las réplicas, se promedia el cuadrado de esas desviaciones y se aplica un factor de corrección de $(n-1)/n$. El resultado es el estimador Jackknife de la varianza, que refleja la sensibilidad de la estadística a la exclusión de cada unidad de la muestra.

Fundamentos del método BRR

El método de Réplicas Repetidas Balanceadas (BRR), desarrollado por McCarthy (1969), permite estimar la varianza de manera práctica a partir de lo que se llaman “medias muestras”. Estas consisten en submuestras construidas a partir de la muestra original, donde en cada estrato se conserva solo una unidad primaria de muestreo, se duplica su peso y se descarta la otra, de modo que la submuestra siga representando al total. En la práctica, a veces se selecciona solo una UPM por estrato para cubrir más territorio; en esos casos, los estratos se pueden reagrupar en pares, formando pseudo-estratos, que se tratan como si tuvieran dos UPM, y sobre ellos se aplica el BRR. En teoría, este procedimiento generaría 2^H medias muestras distintas (donde H es el número de pseudo-estratos), pero como procesarlas todas sería inviable, se utiliza solo un subconjunto balanceado que mantiene el mismo efecto.

Para resolver la dificultad de manejar todas las combinaciones posibles, McCarthy propuso construir un subconjunto reducido de réplicas usando una matriz matemática llamada de Hadamard. El número de réplicas necesarias, denominado A , corresponde al múltiplo de cuatro más cercano al número de estratos; por ejemplo, con 23 estratos se necesitan 24 réplicas, con 42 estratos se requieren 44, y así sucesivamente. Gracias al equilibrio que asegura esta matriz, cada UPM aparece el mismo número de veces con peso duplicado y con peso cero, lo que hace que la varianza estimada con este subconjunto sea equivalente a la que se obtendría usando todas las 2^H medias muestras posibles, pero con un esfuerzo de cálculo mucho menor. De esta manera, el BRR ofrece un balance eficiente entre fidelidad al diseño y carga de procesamiento, razón por la cual se ha convertido en un método estándar en las evaluaciones académicas internacionales y en pruebas nacionales como Saber 3°, 5°, 7° y 9°.

Uso de la Matriz de Hadamard

La matriz de Hadamard es un arreglo cuadrado formado únicamente por $+1$ y -1 cuyas filas son ortogonales entre sí; esta propiedad garantiza que cada columna (pseudo-estrato) y cada fila (réplica) queden perfectamente balanceadas. Para generar las réplicas:

Se construye una matriz cuya orden sea el múltiplo de 4 más pequeño que no sea inferior al número de pseudo-estratos. Cada fila se convierte en una réplica: $+1$ se traduce en “inflar” el peso del establecimiento (factor 2) y -1 en “deflactarlo” a 0 (OECD, 2009).

En la siguiente tabla se observa este patrón: dentro de cada pseudo-estrato una escuela alterna 2/0 y la otra 0/2, siguiendo la plantilla que proporciona la matriz de Hadamard. De esta forma se reducen drásticamente las combinaciones necesarias sin perder representatividad.

Pseudo-estrato	Establecimiento	R1	R2	R3	R4	...	R84
1	1	2	0	2	0	...	0
1	2	0	2	0	2	...	2
2	3	2	0	0	2	...	0
2	4	0	2	2	0	...	2
3	5	2	0	0	2	...	2
3	6	0	2	2	0	...	0
4	7	2	0	0	2	...	0
4	8	0	2	2	0	...	0

Corrección de Fay

Con este método de replicación, cada réplica utiliza solo la mitad de las observaciones

disponibles. Esta fuerte reducción del tamaño muestral puede resultar problemática al estimar un estadístico para una subpoblación poco común. De hecho, el número de casos

restantes puede ser tan pequeño o llegar a cero, que la estimación del parámetro poblacional para una réplica determinada sea imposible. Para superar esta desventaja, Fay desarrolló una variante del método BRR. En lugar de multiplicar los pesos de las escuelas por 0 o por 2, Fay sugirió multiplicarlos por un factor de deflación k entre 0 y 1, haciendo que el factor de inflación sea igual a $2 - k$.

Por ejemplo, si el factor de deflación $k = 0.5$, el factor de inflación será $2 - k$, es decir, 1.5 (OECD, 2009).

El patrón de signos de la matriz se mantiene, pero ahora ningún establecimiento recibe peso 0: el registro con -1 pasa a 0.5 y el de $+1$ a 1.5. El equilibrio global sigue siendo válido gracias a la ortogonalidad de la matriz.

Pseudo-estrato	Establecimiento	R1	R2	R3	R4	...	R84
1	1	1.5	0.5	1.5	0.5	...	0.5
1	2	0.5	1.5	0.5	1.5	...	1.5
2	3	1.5	0.5	0.5	1.5	...	0.5
2	4	0.5	1.5	1.5	0.5	...	1.5
3	5	1.5	0.5	0.5	1.5	...	1.5
3	6	0.5	1.5	1.5	0.5	...	0.5
4	7	1.5	0.5	0.5	1.5	...	0.5
4	8	0.5	1.5	1.5	0.5	...	0.5

Cálculo del error de muestreo

Sea $\hat{\mu}$ la estimación con el peso final, $\hat{\mu}_i$ la estimación con la réplica i . Con G réplicas y corrección de Fay k , constante entre 0 y 1, la varianza de muestreo se obtiene:

$$\sigma_{\hat{\mu}}^2 = \frac{1}{G(1-k)^2} \sum_{i=1}^G (\hat{\mu}_i - \hat{\mu})^2$$

En Colombia, el examen Saber 3°, 5° 7° y 9°, además de utilizar el método de réplicas repetidas balanceadas, utiliza la estimación de valores plausibles. En esta prueba se hace uso de 84 réplicas y una corrección de Fay de $k = 0.5$. Así, el error de muestreo del estadístico $\hat{\mu}$ está dado por:

$$\sigma_{\hat{\mu}}^2 = \frac{1}{84(1-0.5)^2} \sum_{i=1}^{84} (\hat{\mu}_i - \hat{\mu})^2$$

$$\sigma_{\hat{\mu}}^2 = \frac{1}{21} \sum_{i=1}^{84} (\hat{\mu}_i - \hat{\mu})^2$$

Estándares internacionales y uso en pruebas nacionales

En pruebas internacionales como PISA, PIRLS o TIMSS el error de estándar se calcula mediante la metodología de Réplicas Repetidas Balanceadas con la corrección de Fay. El consorcio que compila estos estudios publica 80 pesos de réplica construidos a partir de la metodología anteriormente descrita.

En Colombia, en el examen Saber3°, 5°, 7° y 9° realizado por el Icfes, emplea esta misma metodología, pero adaptada a su propio diseño probabilístico (proporcional al tamaño

en tres etapas). Para estas pruebas, el Icfes generó 84 réplicas balanceadas (Peso_R1... Peso_R84) aplicando exactamente la misma corrección de Fay con $k = 0.5$. Los microdatos de las aplicaciones muestrales 2022 y 2023, junto con el Manual de procesamiento y los macros en R que automatizan el uso de estos 84 pesos, están disponibles en el portal Data Icfes; allí se detallan tanto la estructura de las réplicas como las pautas para combinar los valores plausibles y reportar los errores estándar coherentes con el diseño.

Conclusiones

El cálculo del error de muestreo es un componente esencial en toda evaluación basada en muestras probabilísticas, ya que permite cuantificar la incertidumbre de las estimaciones y otorgarles solidez estadística. Ignorar este componente puede conducir a intervalos de confianza demasiado estrechos y a conclusiones engañosas, subestimando la variabilidad real del diseño muestral. En este contexto, el método de Réplicas Repetidas Balanceadas (BRR) constituye una alternativa práctica y eficiente frente a la complejidad de los diseños multietápicos y estratificados. Al construir subconjuntos balanceados de la muestra, este procedimiento reproduce la variabilidad esperada del muestreo sin necesidad de aplicar fórmulas analíticas sofisticadas. Además, su implementación es directa, ya que basta con utilizar los pesos de réplica

generados y hacer uso de herramientas disponibles para su procesamiento.

Gracias a estas ventajas, el BRR se ha consolidado como un estándar tanto en evaluaciones internacionales (como PISA o TIMSS) como en pruebas nacionales en Colombia (Saber 3°, 5°, 7° y 9°). Su adopción garantiza estimaciones más precisas y comparables, fortaleciendo la credibilidad de los resultados y ofreciendo a investigadores y responsables de política educativa una herramienta robusta para la toma de decisiones informadas.

Bibliografía

- McCarthy, P. J. (1969). Pseudo-replication: Half-samples. *Review of the International Statistical Institute*, 37, 239–264.
- Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.
- Valliant, R., Dever, J. A., & Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples* (2.ª ed.). Springer.
- Organisation for Economic Co-operation and Development (OECD). (2009). *PISA Data Analysis Manual: SAS* (2.ª ed.). OECD Publishing. oecd.org