CÁLCULO DE SIGNIFICANCIA ESTADÍSTICA PARA RESULTADOS DE LAS PRUEBAS SIMCE

Unidad de Análisis Estadístico

División de Evaluación de Logros del Aprendizaje

Agencia de Calidad de la Educación

2018

Índice

1. Antecedentes Generales

Uno de los indicadores más consolidados en los reportes de resultados de las pruebas Simce es la comparación de los puntajes promedio de dos agrupaciones de estudiantes. Por ejemplo, un establecimiento puede comparar su puntaje promedio con el puntaje promedio del grupo socioeconómico en el cual se encuentra clasificado o con el puntaje promedio de todos los estudiantes del país. Realizar estas comparaciones permite a los establecimientos determinar si sus estudiantes demuestran un desempeño superior, similar o inferior al de los estudiantes del grupo de referencia.

Para determinar si la diferencia entre los puntajes promedio de dos agrupaciones de estudiantes es significativa, y no producto de factores aleatorios, se utiliza el método detallado en la primera parte de este documento.

Por otro lado, para el caso de los resultados según Estándares de Aprendizaje surgió la necesidad de contar con un método que permita comparar las distribuciones de estudiantes en dichas clasificaciones. Para esto se buscó una metodología de comparación de la distribución de estudiantes de cada estándar que permitiese determinar si la diferencia entre dos proporciones de estudiantes es significativa o no. Esta metodología es presentada en la segunda parte del presente documento y debe ser utilizada para realizar comparaciones de agregaciones de 1.000 o más estudiantes, como comunas, regiones y grupos socioeconómicos, no siendo adecuada para comparar proporciones de estudiantes en establecimientos.

Dado que las pruebas Simce son de carácter censal, en el documento se hace referencia a poblaciones y no a muestras, siendo estas últimas relacionadas con los Estudios Nacionales.

2. Comparación de Puntajes Promedio

Una medida razonable de la discrepancia entre los datos y la hipótesis nula ($H_0 : \mu_x - \mu_y = 0$) es la diferencia entre el promedio de una agrupación de interés $\overline{x}$, y el promedio con el cual se desea comparar $\overline{y}$ (agregación de referencia). Si $\overline{x}$ e $\overline{y}$ realmente provienen de la misma población, la diferencia debiese tender a ser pequeña. Si provienen de poblaciones diferentes, la diferencia sería más grande. Para esto, se utiliza un método en base al estadístico *t-student*[1].

Una estimación útil es por medio de intervalos, en donde se calculan los valores entre los que se encontrará el parámetro (en este caso la diferencia de promedios: $\overline{x} - \overline{y}$), con un nivel de confianza de 95 %[2].

Un intervalo de confianza de 95 % para la diferencia de medias está dado por:

$$(\overline{x} - \overline{y}) \pm t_{(n;0,975)} \sqrt{\frac{\hat{\varepsilon_1}^2}{n_1} + \frac{\hat{\varepsilon_2}^2}{n_2}}$$

Donde:

- $\overline{x}$ e $\overline{y}$: promedio en cada una de las poblaciones de interés.
- $\hat{\varepsilon_1}^2$ y $\hat{\varepsilon_2}^2$: cuadrados de los errores estándar de medición en cada una de las poblaciones de interés.
- n_1 y n_2: tamaños de las poblaciones a comparar.
- n: grados de libertad del estadístico *t-student*, determinado a partir del tamaño de las poblaciones de interés. $n = n_1 + n_2 - 2$.
- $t_{(n;0,975)}$: valor en la distribución *t-student* con n grados de libertad y con una probabilidad acumulada de 97,5 % (hipótesis bilateral).

2.1. Errores de Estimación de Puntuaciones

En una medición como estas evaluaciones, donde se pretende estimar un rasgo no observable, las estimaciones nunca serán exactas conteniendo cierto grado de error. A partir de ello, tienen limitaciones para determinar si, por ejemplo, existen diferencias entre dos puntajes promedio.

[1] Se utiliza esta distribución porque permite una comparación más robusta en poblaciones de pocos datos. Además, cuando la población es suficientemente grande, se asemeja a una distribución normal

[2] Nivel de confianza es la 'probabilidad' de que el intervalo calculado contenga al verdadero valor del parámetro. Se indica por $1 - \alpha$ y habitualmente se reporta el porcentaje $(1 - \alpha)100\,\%$. Se habla de nivel de confianza y no de probabilidad ya que una vez obtenida la población de interés, el intervalo de confianza podrá contener al verdadero valor del parámetro o no.

2.1.1. Simce Censal

Considerando que la estimación de las puntuaciones se realiza utilizando la Teoría de Respuesta al Ítem (TRI o IRT por sus siglas en inglés)[3], se obtiene, para cada estudiante evaluado, un puntaje estimado y su correspondiente error de estimación. Este último permite estimar el intervalo en el cual se encuentra el verdadero valor de la habilidad del estudiante. Así, para obtener una comparación estadística entre dos agrupaciones de interés, el error de medición debe ser tomado en cuenta. Estos errores son incluidos en el estadístico de la siguiente manera:

$$SE^2 = \sum_{i=1}^{n} EP_{Esti}^2$$

Donde EP_{Esti} corresponde al error de estimación de puntaje para el estudiante i, que pertenece a la población de interés (i=1,...,n).

2.1.2. Simce Censal Escritura

A diferencia de las otras pruebas de Simce regular, que son procesadas bajo TRI, la prueba de Escritura aplicada en 6° básico es procesada bajo Teoría Clásica de los Test (TCT), por lo que la construcción del test de comparación no se realiza de la misma forma.

En este caso, el error de medida de la prueba viene definido por el SEM[4], que dado el modelo aplicado, es un valor común para cada una de las formas. Así, la descomposición del error estándar queda definido como:

$$SE^2 = \overline{SEM^2} + S^2$$

Donde:

- $\overline{SEM^2}$: Promedio del error cuadrático de medida correspondiente a las formas que aplican los estudiantes de la población de interés. Considerando k formas, su cálculo viene dado por $\frac{\sum_{j=1}^{k} n_j SEM_j^2}{\sum_{j=1}^{k} n_j}$, donde SEM_j y n_j corresponden al error de medida y el total de estudiantes de la forma j, respectivamente.
- S^2: Varianza de los puntajes obtenidos en la prueba de Escritura, en la población de interés.

2.1.3. Estudios Nacionales

Existen también los Estudios Nacionales, de carácter muestral, que dependiendo del modelo psicométrico con que sean procesados (TRI o TCT), la definición del SE queda definida en parte como se presentó en los puntos

[3]Como textos introductorios se puede consultar por ejemplo, Hambleton & Swaminathan (1985), Item Response Theory. Principles and Applications o Lord & Novick (2008), Statistical Theories of Mental Test Scores.

[4]$SEM = \sigma\sqrt{1-\alpha}$, siendo σ la desviación estándar de los puntajes de la prueba y α la confiabilidad, calculado a partir del indicador Kappa.

anteriores. Sin embargo, dado el carácter muestral de estas pruebas es necesario adicionar otra fuente de variabilidad. Este nuevo componente se denomina "error muestral" y se calcula dependiendo del diseño muestral bajo el cual fue aplicada la prueba.

Para este tipo de pruebas, en general se utiliza un diseño muestral complejo, combinando una etapa de estratificación de establecimientos para posteriormente realizar un muestreo de conglomerados que considera a cada establecimiento como tal. En la etapa de estratificación las escuelas son clasificadas, al menos, de acuerdo a región y dependencia administrativa, para luego muestrear los establecimientos dentro de cada estrato. Para cada establecimiento (conglomerado) se considera a la totalidad de los estudiantes que cursan el nivel respectivo.

Considerando el diseño explicado anteriormente, desde una población de K establecimientos donde K_j es la cantidad de establecimientos del estrato j, con $\sum_j K_j = K$, se obtiene una muestra de k establecimientos donde k_j es la cantidad de establecimientos de la muestra en el estrato j, con $\sum_j k_j = k$. El error muestral, para una agregación de interés i, viene dado por[5]:

$$EM^2 = \sum_{j \in i} \left(\frac{K_j}{K}\right)^2 \left(\frac{K_j - k_j}{K_j k_j}\right) \left(\frac{S_j^2}{k_j - 1}\right)$$

Donde $S_j^2 = \sum_{k \in j} S_k^2$ es la suma de las varianzas de los puntajes de los estudiantes que son parte de los establecimientos del estrato j, que pertenezcan a la agregación de interés i, asumiendo que los establecimientos son independientes unos de otros en sus resultados.

En el primer párrafo se establece que la definición del SE no es exactamente igual a los ya presentados anteriormente, esto debido a que dicha definición queda condicionada al tipo de muestreo y/o reporte, a las comparaciones consideradas para los resultados con sus distintos tipos de pruebas, entre otros aspectos identificados. Considerando cada caso, podría resultar necesario incorporar un factor de ponderación para los errores de medida y las medias de cada agregación. Por ejemplo, en el caso de que se deba proyectar la muestra a la población total al comparar un resultado muestral con uno censal, o en el caso que deba ajustarse una muestra, que no considera distribuciones proporcionales, para obtener resultados representativos de la población, o incluso ambos. Los diseños muestrales respectivos entregarán un lineamiento para el tratamiento de estos ponderadores.

Así, en el caso de utilizar ponderadores muestrales, las definiciones de SE, para cada metodología, viene dado por:

- TRI: $SE_p^2 = \sum_{i=1}^{n} w_i EP_{Esti}^2$

[5]Para detalles sobre la derivación de la ecuación, revisar Thompson (2002), capítulos 11 y 12.

- TCT: $SE_p^2 = \overline{SEM_p^2} + S_p^2$,

con $\overline{SEM_p^2} = \frac{\sum_{i=1}^n w_i SEM_i^2}{\sum_{i=1}^n w_i}$ y $S_p^2 = \frac{\sum_{i=1}^n w_i (x_i - \overline{x}_p)^2}{\sum_{i=1}^n w_i - 1}$, donde $\overline{x}_p = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$,

y w_i siendo el ponderador asignado al estudiante i, definido de acuerdo al diseño muestral de la prueba.

Por último, el nuevo SE^* queda definido como $SE^* = \sqrt{SE^2 + EM^2}$ (o SE_p^2, en caso de usar ponderadores).

2.2. Construcción del Test

Para construir el test y realizar la dócima es necesario computar ciertos indicadores:

1. El promedio simple de puntajes en cada una de las agrupaciones que se desea comparar[6].
2. El número de estudiantes que hay en las respectivas agregaciones.
3. Las desviaciones estándar de cada agregación de interés, dadas de la siguiente forma:

$$DE = \left(\frac{t_{(n-1;0,975)}}{\sqrt{n}} \right) \times SE$$

Donde:

- n: número de estudiantes en las agrupaciones de interés. En el caso de uso de ponderadores, $n = \sum_i w_i$.
- $t_{(n-1;0,975)}$: valor de la distribución $t - student$ con $n - 1$ grados de libertad y una probabilidad acumulada de 97,5 % (hipótesis bilateral).
- SE: indicador que da cuenta de la variabilidad en la agregación. En el caso de prueba muestral, se utiliza SE^*, asumiendo definición de acuerdo al uso o no de ponderadores.

A continuación, para determinar la existencia de una diferencia estadísticamente significativa entre dos agrupaciones, es necesario calcular los límites del intervalo de confianza para la diferencia de sus promedios (LI: límite inferior, LS: límite superior) y evaluar si dicha diferencia es significativa a un nivel de confianza de 95 %. En el caso de las pruebas de Escritura y los Estudios Nacionales, se utiliza la definición anterior, considerando los valores SE determinados en sus secciones correspondientes.

[6]Este promedio simple es calculado a partir de las puntuaciones de todos los estudiantes que pertenecen a dicha agrupación. Un ejemplo sería comparar los puntajes promedio obtenidos por hombres y mujeres en la prueba de Lectura de 4° básico.

Así, para el caso de las pruebas Simce censales, los límites superior e inferior para la diferencia entre los promedios de dos agregaciones, vienen representados por la siguiente ecuación:

$$L = \pm\sqrt{[(DE_1 + E)^2 + (DE_2 + E)^2]} \tag{1}$$

Donde:

- E: error de población[7]. En el caso de comparaciones entre agrupaciones dentro del mismo año, este error toma el valor de 0,5. En el caso de comparar promedios obtenidos en distintos años, este error toma el valor de 3,5.

Para Escritura, los límites del intervalo para una agregación vienen dados por:

$$L = \pm(DE_1 + DE_2) \tag{2}$$

Para el caso de los Estudios Nacionales, la definición del intervalo queda sujeto a los criterios considerados para su análisis, pudiendo utilizarse cualquiera de las dos anteriores.

2.3. Criterio de Decisión

En este tipo de análisis se combinan los criterios estadísticos de significancia con un criterio de diferencia mínima con el fin de resguardar que diferencias relativamente pequeñas, en los puntajes de dos agregaciones, no sean consideradas significativas (por ejemplo, diferencias de 1 punto). Este criterio es definido según cada una de las pruebas.

2.3.1. Simce Censal

Utilizando la definición del intervalo de la ecuación (1), si la diferencia entre los promedios de ambas poblaciones está entre LI y LS (ambos límites inclusive), no existe una diferencia estadísticamente significativa. Si la diferencia entre los promedios es mayor a LS, esta diferencia es estadísticamente significativa a favor de la población 1. Del mismo modo, si la diferencia de los promedios es menor a LI, esta diferencia es estadísticamente significativa a favor de la población 2. Considerando un intervalo más conservador, se determina comparar las diferencias de los promedios redondeados de ambas poblaciones.

[7]Se considera el error de haber tomado esa cohorte buscando minimizar el error de que los puntajes reflejen las características particulares de los estudiantes evaluados.

Definiendo $D = \overline{X}_1 - \overline{X}_2$, como la diferencia entre las medias redondeadas de dos agregaciones de interés, el algoritmo para calcular la significancia estadística de la diferencia en las pruebas Simce censales, queda definido como[8]:

- $LI \leq D \leq LS \Rightarrow$ *Diferencia no significativa.*
- $LS < D$ y $D > 5 \Rightarrow$ *Diferencia significativa, el promedio de la población 1 es superior al promedio de la población 2.*
- $D < LI$ y $D < -5 \Rightarrow$ *Diferencia significativa, el promedio de la población 2 es superior al promedio de la población 1.*

Así, para que exista diferencia significativa, debe cumplirse el criterio intervalar y que $|D| > 5$. Si uno de estos criterios no se cumple, la diferencia no es considerada significativa.

2.3.2. Simce Censal Escritura

Para el caso de Simce Escritura, si bien es una prueba censal, utilizar el criterio de diferencia mínima anterior resulta ser demasiado estricto, pues la escala de puntajes de esta prueba es un tercio más pequeña que las otras censales. Como solución, se utiliza como criterio de diferencia mínima el indicador *d de Cohen* (Cohen, 1988 y 1992), el cual mide el tamaño del efecto en una población. Este indicador está definido por:

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\sigma}$$

Donde $\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2}{N_1+N_2-2}}$, es la desviación estándar de los puntajes entre ambas agregaciones, considerándolas independientes.

Basándose en la revisión bibliográfica realizada por Morales (2012), se establece que, en el ámbito educacional, cuando $|d| > 0,3$ la diferencia puede ser considerada significativa.

Utilizando los mismos criterios de las pruebas Simce censales para la significancia intervalar, pero considerando la definición de intervalos de la ecuación (2), el algoritmo para el criterio de significancia estadística para dos agregaciones de interés queda definido por:

- $LI \leq D \leq LS \Rightarrow$ *Diferencia no significativa.*

[8]El criterio de los 5 puntos permite ser más conservadores en las conclusiones. El valor escogido viene dado por el error de medida basado en la Teoría Clásica del Test, considerando la confiabilidad de las pruebas Simce igual a 0,99.

- $LS < D$ y $|d| > 0,3 \Rightarrow$ *Diferencia significativa, el promedio de la población 1 es superior al promedio de la población 2.*
- $D < LI$ y $|d| > 0,3 \Rightarrow$ *Diferencia significativa, el promedio de la población 2 es superior al promedio de la población 1.*

2.3.3. Estudios Nacionales

Para el caso de estas pruebas, la construcción de los intervalos depende del modelo de análisis, estableciendo la construcción del intervalo de Simce censal o de Simce Escritura, cuya variación está en el posible uso de ponderadores. Así mismo, el criterio de diferencia mínima también es definido de acuerdo a las características de la prueba. Si se utiliza TRI en la prueba muestral, es posible utilizar el mismo criterio de decisión de las pruebas censales, y si se utiliza TCT también es posible considerar el criterio de decisión utilizado en Escritura. También es posible buscar una nueva definición de acuerdo a criterios que sean establecidos particularmente para cada estudio.

3. Comparación de Proporciones o Porcentajes

Dentro de las comparaciones realizadas en las pruebas, existen las implementadas para comparar proporciones (o porcentajes) de estudiantes en una cierta categoría de clasificación. En las pruebas Simce existen los Estándares de Aprendizaje, que clasifican a los estudiantes en tres niveles progresivos de logro (Insuficiente, Elemental y Adecuado). También, por ejemplo, en el Estudio Nacional de Inglés de III medio (muestral) se realizan comparaciones de acuerdo a la categorización de estudiantes sobre o bajo el nivel A2, referido al Marco Común Europeo de Referencia (CEFR, por sus siglas en inglés). Cada prueba tiene el potencial de poder clasificar a los estudiantes en distintos niveles.

La mayoría de los textos introductorios de estadística sugieren utilizar intervalos de confianza para la diferencia basado en intervalos de Wald. Considerando una población, con n estudiantes, de los cuales r han sido clasificados en una cierta categoría de interés, entonces la proporción de estudiantes en dicha categoría es $p = \frac{r}{n}$. A partir de la ecuación anterior se quiere calcular un intervalo de confianza (IC) para tal proporción en la población. Un IC para p se calcula comúnmente como:

$$p \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

Por lo que la diferencia entre dos proporciones p_1 y p_2, con $D = p_1 - p_2$ tiene como intervalo de confianza:

$$D \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Donde:

- $z_{1-\frac{\alpha}{2}}$: punto de la distribución normal en que se acumula el $100 \times \left(1 - \frac{\alpha}{2}\right)$ % de probabilidad.
- n_1 y n_2: tamaños de las poblaciones 1 y 2 que se comparan.

Por otro lado, hay extensa literatura (Vollset, 1993; Santner, 1998; Agresti & Coull, 1998; Newcombe, 1998; Brown, Cai & DasGupta, 2001) que indica que dicho procedimiento es discutible en particular para tamaños inferiores a 50 y cuando la proporción (o porcentaje) de interés p es cercano a 0 o 1 (0 % o 100 %).

En la misma literatura se sugiere utilizar los intervalos basados en la metodología llamada *Wilson Score Interval* (Brown, Cai & DasGupta, 2001; Newcombe & Merino, 2006) donde el intervalo de confianza para una proporción p viene dado por:

$$\frac{np + \frac{1}{2}z^2}{n + z^2} \pm \frac{z^2\sqrt{n}}{n + z^2} \sqrt{p(1-p) + \frac{z^2}{4n}} \tag{3}$$

Esta metodología solo es válida cuando se cumple la condición[9]: $Min\{np, n(1-p)\} \geq 10$.

[9] Algunos autores relajan la condición a que el mínimo sea mayor a 5.

3.1. Supuestos

Los métodos estadísticos disponibles para la comparación de proporciones se basan en dos supuestos claves:

1. La existencia de una proporción verdadera y desconocida, cuyo valor no es necesariamente constante en el tiempo, de estudiantes que pertenecen a cada categoría de clasificación en cada año, definida por su resultado en una prueba de logro. Este supuesto permite utilizar una *proporción observada*, ya que la prueba Simce entrega un buen estimador de dicha proporción teórica.

2. Que los estudiantes se clasifican en las categorías de manera independiente[10] y que el número de estudiantes es lo suficientemente grande como para aplicar teoría asintótica, permitiendo la determinación probabilística de la confianza de la comparación[11].

Como las categorías de clasificación, por ejemplo los Estándares de Aprendizaje, están construidos usando las pruebas Simce y estas a su vez están construidas y analizadas de modo que sus resultados son comparables año a año, el primer supuesto es satisfecho.

En el caso del segundo supuesto, la situación no es tan sencilla. Por un lado, del análisis de las pruebas Simce, TIMSS y PISA surge evidencia para refutar el supuesto de independencia. Se ha estimado la correlación de los resultados de los estudiantes en aula, tanto para Lectura como para Matemática, y se ha determinado que estas correlaciones son significativamente distintas de cero[12]. Por otro lado, el número de estudiantes por establecimiento que rinde las pruebas Simce es muy pequeño para aplicar teoría asintótica en un gran número de éstos[13]. Para el caso particular de la comparación de proporciones, la aplicación de Teorema Central del Límite exige que se satisfaga una condición que involucra la proporción estimada y el número de individuos usados para estimar la proporción. En este caso dado que el supuesto de independencia no es satisfecho, se utiliza la versión más conservadora de la condición[14]: $Min\{n\hat{p}, n(1-\hat{p})\} \geq 10$.

Considerando lo anterior se entregan comparaciones de proporciones o porcentajes de estudiantes en categorías de clasificación solo para poblaciones estadísticamente grandes (1.000 o más estudiantes). Es decir, las com-

[10] El número de estudiantes que se clasifica en una cierta categoría sigue una distribución binomial porque cada estudiante se clasifica en un nivel siguiendo una distribución Bernoulli.

[11] Aproximación de la distribución binomial a la distribución normal usando el Teorema Central del Límite.

[12] Una condición necesaria para independencia es que las correlaciones sean cero.

[13] No hay reglas estrictas para la aplicación del Teorema Central del Límite, sin embargo hay consenso en la literatura estadística en que para tamaños superiores a 50 funciona muy bien, entre tamaños de 20 a 50 funciona bastante bien, tamaños menores que 10 no siempre funciona y no debe aplicarse para tamaños menores a 5.

[14] El efecto neto de la no-independencia entre observaciones es que se pierden grados de libertad.

paraciones de proporciones o porcentajes son robustas si la base para la cual se calculan es lo suficientemente grande.

3.2. Construcción del Test

Considerando las restricciones de la metodología indicada, se exponen a continuación las ecuaciones utilizadas en la comparación de proporciones o porcentajes de estudiantes por categoría de clasificación, metodología denominada *Wilson Score Interval* (1927).

Primeramente, se han de calcular tres cantidades, determinadas a partir de la ecuación (3), realizando cálculos algebraicos para su simplificación:

$$A = 2r + z^2$$
$$B = z\sqrt{z^2 + 4r\left(1 - \frac{r}{n}\right)}$$
$$C = 2\left(n + z^2\right)$$

En donde:

- z: valor en la distribución normal donde se acumula el 97,5 % de la distribución ($\alpha = 5\,\%$, hipótesis bilateral). Su valor corresponde a 1,96.
- r: número de estudiantes en una categoría de clasificación.
- n: número de estudiantes en la población.

Los componentes de las ecuaciones anteriores dan como resultado la estimación de un intervalo de 95 % de confianza para una proporción. Esta representación es equivalente a la presentada en la ecuación (3) pero tiene un manejo operacional más simple. Luego, el intervalo de confianza viene dado por $\frac{(A \pm B)}{C}$.

Utilizando el intervalo de confianza presentado en el punto anterior, se deben calcular l_1 y u_1, que son los límites inferior y superior que definen el intervalo de confianza al 95 % para la población de estudiantes 1 (población de interés), y l_2 y u_2, que son los límites inferior y superior de la población de estudiantes 2 (población de referencia)[15]. Ambos calculados a partir de la metodología de Wilson antes presentada.

[15]Una comparación sería, por ejemplo, determinar si la proporción de estudiantes en el Estándar de Aprendizaje Insuficiente de algún grupo socioeconómico es significativamente menor con respecto al mismo Estándar de Aprendizaje en otro grupo socioeconómico.

Los límites del intervalo de confianza de la diferencia de proporciones de estudiantes en alguna categoría de clasificación están dados por (Newcombe, 1998b):

$$LI : D - \sqrt{(p_1 - l_1)^2 + (u_2 - p_2)^2}$$
$$LS : D + \sqrt{(p_1 - l_1)^2 + (u_2 - p_2)^2}$$

Donde $D = (p_1 - p_2)$.

Cabe destacar que las comparaciones presentadas en este documento son bloque a bloque. Por ejemplo, proporción de estudiantes en un Estándar de Aprendizaje versus otra población de estudiantes en el mismo Estándar de Aprendizaje.

Esta metodología es aplicable tanto a pruebas censales como a muestrales. Sin embargo, para el caso de las pruebas muestrales es necesario realizar una variación en la ecuación (3) de tal manera que, realizando ejercicio algebraico se llega a:

$$\frac{p + \frac{z^2}{2n} \pm z\sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

Así, el término $\frac{p(1-p)}{n}$, que corresponde a la varianza poblacional, debe reemplazarse por el error muestral de una proporción (EM_p), de acuerdo al diseño utilizado. Además, en caso de utilizarse ponderadores, las proporciones p y el total n debe reemplazarse por sus respectivos valores ponderados:

$$n_w = \sum_i w_i$$
$$p_w = \frac{\sum_i w_i I_i}{\sum_i w_i}$$

Donde I_i es el indicador de pertenencia del estudiante i-ésimo a la categoría de interés.

Redefiniendo las tres cantidades anteriores (pudiendo utilizarse p ó p_w y n ó n_w):

$$A = p + \frac{z^2}{2n}$$
$$B = z\sqrt{EM_p + \frac{z^2}{4n^2}}$$
$$C = 1 + \frac{z^2}{n}$$

Se vuelve a obtener el intervalo para la proporción muestral de manera $\frac{(A \pm B)}{C}$. Es importante considerar que, en el caso de utilizar ponderadores muestrales, el criterio de $Min\{np, n(1-p)\} \geq 10$ debe utilizarse sobre los totales muestrales de origen y no los ponderados. A su vez, si no se utilizan ponderadores, es importante evaluar la condición de que se realice este procedimiento a agregaciones de 1.000 o más estudiantes.

3.3. Criterios de Decisión

Si la diferencia entre las proporciones de ambas poblaciones está contenida dentro del intervalo, no existe una diferencia estadísticamente significativa. Si la diferencia entre las proporciones es superior a LS, esta diferencia es estadísticamente significativa a favor de la población 1. Del mismo modo, si la diferencia de las proporciones es inferior a LI, esta diferencia es estadísticamente significativa a favor de la población 2. Otra forma de comprobar la significancia es verificando si el cero (0) está contenido entre los límites definidos del intervalo de confianza. De modo análogo que para promedios, se determina un valor mínimo, en este caso de 3 puntos porcentuales de diferencia, para construir intervalos más conservadores[16]. Así, la regla de decisión, la cual aplica para todas las pruebas, es la siguiente:

1. Si $LI \leq 0 \leq LS$, la diferencia no es estadísticamente significativa.
2. Si $LS < 0$ y $|D| \geq 3\,\%$, entonces la proporción en la población 2 es significativamente mayor a la proporción de estudiantes en la población 1.
3. Si $LI > 0$ y $|D| \geq 3\,\%$, entonces la proporción en la población 1 es significativamente mayor a la proporción de estudiantes en la población 2.

[16] El valor de 3 % recoge el error de clasificación de estudiantes basado en la metodología utilizada para ello.

Referencias

[1] Agresti A. y Coull B. A. (1998), *Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions*, The American Statistician, vol 52:2, pp 119-126.

[2] Brown L. D., Cai T. T. y DasGupta A. (2001), *Interval Estimation for a Binomial Proportion*, Statistical Science, vol 16:2, pp 101-133.

[3] Cohen J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates.

[4] Cohen J. (1992), *A Power Primer*, Psychological Bulletin, vol 112:1, pp 155-159.

[5] Hambleton K. y Swaminathan H. (1985), *Item Response Theory: Principles and Applications*, Springer Netherlands.

[6] Lord F. y Novick M. (1968), *Statistical Theories of Mental Test Scores*, MA: Addison-Wesley.

[7] Morales P. (2012), *El Tamaño del Efecto (Effect Size): Análisis Complementarios al Contraste de Medias*, Universidad Pontificia Comillas.

[8] Newcombe R. (1998a), *Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods*, Statistics in Medicine, vol 17:8, pp 857-872.

[9] Newcombe R. (1998b), *Interval Estimation for the Difference between Independent Proportions: Comparison of Eleven Methods*, Statistics in Medicine, vol 17:8, pp 873-890.

[10] Newcombe R. y Merino C. (2006), *Intervalos de Confianza para las Estimaciones de Proporciones y las Diferencias entre Ellas*, Interdisciplinaria, vol 23:2, pp 141-154.

[11] Santner T. (1998), *Teaching Large-Sample Binomial Confidence Intervals*, Teaching Statistics, vol 20:1, pp 20-23.

[12] Thompson S. K. (2002), *Sampling*, John Wiley & Sons Inc.

[13] Vollset S. E. (1993), *Confidence Intervals for a Binomial Proportion*, Statistics in Medicine, vol 12, pp 809-824.