CÁLCULO DE SIGNIFICANCIA ESTADÍSTICA PARA RESULTADOS DE LAS PRUEBAS SIMCE

Unidad de Análisis Estadístico

División de Evaluación de Logros de Aprendizaje

Agencia de Calidad de la Educación

2013

Índice

1. Antecedentes Generales **1**

2. Comparación de puntajes promedios **2**

2.1. Errores de estimación de puntuaciones 3

2.2. Construcción del test 3

2.3. Criterio de decisión 4

3. Comparación de proporciones o porcentajes **6**

3.1. Supuestos 6

3.2. Construcción del test 8

3.3. Criterios de decisión 9

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, con la incorporación de los resultados según estándares de aprendizaje surgió la necesidad de contar con un método que permita comparar las distribuciones de estudiantes en dichos estándares. 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 en un determinado estándar 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), por lo tanto no es 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.

2. Comparación de puntajes promedios

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 (agregación de referencia), $\overline{y}$. Si $\overline{x}$ e $\overline{y}$ realmente provienen de la misma población, la diferencia tendería a ser pequeña. Si provienen de poblaciones diferentes, la diferencia sería más grande.

Cuando no se puede asumir que las dos poblaciones en estudio tienen varianzas homogéneas entonces se utiliza un método en base al estadístico *t-student*[1].

Una estimación útil es por intervalos, en donde se calculan los dos 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 correspondiente al 95 % para la diferencia de medias está dado por:

$$(\overline{x} - \overline{y}) \pm t_{(n,0,95)} \sqrt{\frac{\hat{\varepsilon_1^2}}{n_1} + \frac{\hat{\varepsilon_2^2}}{n_2}}$$

Donde:

- $\overline{x}$ y $\overline{y}$: promedio en cada una de las poblaciones de interés.
- ε_1^2 y ε_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.
- $t_{(n,0,95)}$: valor en la distribución *t-student* con n grados de libertad y con una probabilidad acumulada de 0,95.

[1] Se utiliza esta distribución porque además, permite una comparación más robusta en poblaciones de pocos datos.

[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 contendrá al verdadero valor del parámetro o no.

2.1. Errores de estimación de puntuaciones

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

Considerando que la estimación de las puntuaciones se realiza utilizando la teoría de respuesta al ítem[3] (IRT), 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:

$$SE = \sqrt{EP_{Est1}^2 + EP_{Est2}^2 + EP_{Est3}^2 + \ldots + EP_{Esti}^2}$$

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

2.2. 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[4].
2. El número de estudiantes que hay en las respectivas agregaciones.
3. La desviación estándar, dada de la siguiente forma:

$$DE = SE \cdot \frac{t_{(n-1,0,95)}}{n}$$

Donde:

- n: número de estudiantes en las agrupaciones de interés.
- $t_{(0,95,n-1)}$: valor de la distribución con $n-1$ grados de libertad y con una probabilidad acumulada de 0,95.
- SE: indicador que da cuenta de la variabilidad en la agregación.

[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]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.

A continuación, para determinar la existencia de una diferencia estadísticamente significativa entre esas dos agrupaciones, es necesario calcular los límites del intervalo de confianza, para posteriormente determinar si la diferencia, entre los promedios de puntajes de las poblaciones de interés, es significativa con un nivel de confianza de 95 %.

Así, los límites superior e inferior, para la diferencia entre los promedios de las dos agregaciones vienen representados por las siguientes ecuaciones:

$$LS = \sqrt{[(DE_{pobl1} + E)^2 + (DE_{pobl2} + E)^2]}$$
$$LI = -1 \cdot \sqrt{[(DE_{pobl1} + E)^2 + (DE_{pobl2} + E)^2]}$$

Donde:

- DE_{pobli}: desviación estándar de la población i (i=1,...,n).
- E: error de población[5].
- LS: límite superior del intervalo de confianza.
- LI: límite inferior del intervalo de confianza.

2.3. Criterio de decisión

Si la diferencia entre los promedios de ambas poblaciones es menor o igual al LS o es mayor o igual al LI, no existe una diferencia estadísticamente significativa. Si la diferencia entre los promedios es mayor al LS, esta diferencia es estadísticamente significativa a favor de la población 1, del mismo modo si la diferencia de los promedios es menor al LI esta diferencia también es estadísticamente significativa a favor de la población 2.

[5]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. En el caso de comparaciones entre agrupaciones para una medición dentro del mismo año este error toma el valor de 0,5, en el caso de comparar promedios obtenidos en mediciones ocurridas en distintos años este error toma el valor de 3,5.

Considerando un intervalo más conservador, se determinó comparar las diferencias de los promedios de ambas poblaciones redondeados. Finalmente el criterio queda de la siguiente manera[6]:

- $LI \leq \overline{Población_1} - \overline{Población_2} \leq LS \Rightarrow$ *Diferencia no significativa.*
- $LS < \overline{Población_1} - \overline{Población_2}$ y $5 < \overline{Población_1} - \overline{Población_2} \Rightarrow$ *Diferencia significativa, el promedio de la población 1 es superior al promedio de la población 2.*
- $\overline{Población_1} - \overline{Población_2} < LI$ y $\overline{Población_1} - \overline{Población_2} < -5 \Rightarrow$ *Diferencia significativa, el promedio de la población 2 es superior al promedio de la población 1.*

[6]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. $EM = SD\sqrt{1 - confiabilidad}$.

3. Comparación de proporciones o porcentajes

La comparación de proporciones es un problema clásico, la mayoría de los textos introductorios de estadística sugieren utilizar intervalos de confianza para la diferencias basado en intervalos de Wald. Considerando una población, con n estudiantes, de los cuales r han sido clasificados en un estándar de aprendizaje de interés[7], entonces la proporción de estudiantes en dicho estándar de aprendizaje 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:

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

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

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

Donde:

- $z_{\frac{\alpha}{2}}$: punto de la distribución normal en que se acumula el $1 - \frac{\alpha}{2}$ 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 indican 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:

$$IC : \frac{np + \frac{1}{2} \cdot z^2_{\frac{\alpha}{2}}}{n + z^2_{\frac{\alpha}{2}}} \pm \frac{\sqrt{n} \cdot z_{\frac{\alpha}{2}}}{n + z^2_{\frac{\alpha}{2}}} \sqrt{[p(1-p) + \frac{z^2_{\frac{\alpha}{2}}}{4n}]}$$

Esta metodología solo es válida cuando se cumple la condición[8]: $Min\{np, n(1-p)\} \geq 10$.

3.1. Supuestos

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

[7] De los tres posibles: Adecuado, Elemental e Insuficiente.

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

1. La existencia de una proporción verdadera y desconocida[9] de estudiantes que pertenecen a cada estándar de aprendizaje de cada establecimiento en cada año, definido por su resultado en una prueba de logro. Este supuesto permite utilizar una *proporción observada*, suponiendo que la prueba SIMCE es el mejor estimador de dicha proporción teórica.

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

Como 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 establecimientos[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. Esta es[14]:

$$Min\{n\hat{p}, n(1-\hat{p})\} \geq 10$$

Considerando lo anterior se entregan comparaciones de proporciones o porcentajes de estudiantes en estándares de aprendizaje sólo para poblaciones estadísticamente grandes (1.000 o más estudiantes). Es decir, las comparaciones de proporciones o porcentajes son robustas si la base para la cual se calculan es lo suficientemente grande.

[9] Cuyo valor no es, necesariamente, constante en el tiempo.

[10] El número de estudiantes que se clasifica en un estándar de aprendizaje 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 este 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.

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 estándar de aprendizaje, metodología denominada *Wilson Score Interval* (1927).

Primeramente, se han de calcular tres cantidades:

$$A = 2 \cdot r + z^2_{\frac{\alpha}{2}}$$
$$B = z_{\frac{\alpha}{2}} \cdot \sqrt{z^2_{\frac{\alpha}{2}} + 4 \cdot r(1 - \frac{r}{n})}$$
$$C = 2 \cdot (n + z^2_{\frac{\alpha}{2}})$$

En donde:

- $z_{\frac{\alpha}{2}}$: valor en la distribución normal donde se acumula el 97,5 % de la distribución ($\alpha = 5\,\%$), su valor es 1,96.
- r: número de estudiantes en el estándar de aprendizaje.
- n: número de estudiantes en la población.

Los componentes de las ecuaciones anteriores dan como resultado la estimación de un intervalo de confianza al 95 % para una proporción. Esta representación es equivalente a la presentada en el inicio del punto 3 pero tiene un manejo operacional más simple. Luego, el intervalo de confianza está dado por:

$$\text{IC: } \frac{(A \pm B)}{C}$$

Utilizando el intervalo de confianza presentado en el punto anterior, se deben calcular l_1 y u_1: 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 son los límites inferior y superior, de la población de estudiantes 2, con la cual es comparada (población de referencia)[15]. Ambos calculados partir de la metodología de Wilson antes presentada.

Los límites del intervalo de confianza de la diferencia de proporciones de estudiantes en algún estándar de aprendizaje está dado por (Newcombe, 1998b):

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

[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 respecto al mismo estándar de aprendizaje en otro grupo socioeconómico.

Donde D es la diferencia de proporciones: $D = (p_1 - p_2)$.

Cabe destacar que las comparaciones presentadas en este documento son bloque a bloque: proporción de estudiantes en un estándar de aprendizaje versus otra población de estudiantes en el mismo estándar de aprendizaje.

3.3. Criterios de decisión

Si la diferencia entre las proporciones de ambas poblaciones es menor o igual al LS y mayor o igual al LI, no existe una diferencia estadísticamente significativa. Si la diferencia entre las proporciones es superior al LS, esta diferencia es estadísticamente significativa a favor de la población 1. Del mismo modo, si la diferencia de las proporciones es inferior al 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 y, 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 es la siguiente:

1. Si el intervalo de confianza contiene el valor cero, es decir $LI \leq 0 \leq LS$, la diferencia no es estadísticamente significativa.

2. Si el límite superior es menor al valor cero, es decir $LS < 0$ y además $|D| \geq 3\,\%$, entonces la proporción en la población 2 es significativamente mayor a la proporción de estudiantes en la población 1.

3. Si el límite inferior es mayor al valor cero, es decir $LI > 0$ y además $|D| \geq 3\,\%$, 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.