***ENTREGA 1 DISEÑO Y ANÁLISIS DE INVESTIGACIÓN***

Benjamín Carreño, Ramón Jara

***Parte 1***

**1.1 Cuando analizamos datos, solemos estar interesados en el valor de un parámetro a nivel de la población, como la proporción de partidarios del candidato A entre todos los votantes de un país. Sin embargo, normalmente sólo tenemos acceso a las estadísticas de una pequeña muestra de observaciones extraídas de la población objetivo, como la proporción de partidarios entre los votantes que respondieron a una encuesta. Discuta si esto es un problema o no es un problema.**

Es un problema si uno se refiere a las limitaciones estadísticas muestrales, ya que las muestras extraídas pueden no ser exactamente iguales a la población, lo que implica que la estimación estadística no sea exactamente el parámetro poblacional. Este problema puede ser mayor toda vez que las muestras son pequeñas y no representativas de la heterogeneidad de la población en cuestión. Pero, si se entiende aquí que se puede utilizar la herramienta fundamental de la estadística, no debería ser un obstáculo insalvable.

Al reconocer y considerar dichas limitaciones estadísticas se puede determinar en gran medida cuán similares o diferentes son las muestras a cierto nivel de significancia estadística (y las medias, proporciones, etc.). Podemos, por tanto, inferir con una certeza definida sobre la población utilizando métodos como intervalos de confianza y pruebas de hipótesis para corregir y corresponder las muestras extraídas. Con todo, siempre debemos tener en cuenta las posibles fuentes de error y sesgos durante el proceso de muestreo y análisis para evitar diferencias extremas entre la muestra extraída de análisis y la población que se intenta corresponder.

**1.2 La variabilidad muestral se refiere al hecho de que el valor de una estadística varía de una muestra a otra porque cada muestra contiene un conjunto diferente de observaciones extraídas de la población objetivo. ¿Cuál de las siguientes afirmaciones es cierta? Justifique su respuesta.**

**a. Un menor tamaño de la muestra suele conllevar una mayor variabilidad del muestreo**  
**b. La variabilidad del muestreo no depende del tamaño de la muestra**  
**c. Un mayor tamaño de la muestra conduce generalmente a una mayor variabilidad del muestreo**  
**d. Ninguna de las anteriores**

Como la variabilidad muestral refiere a la variación de un estadístico dependiendo de la extracción de muestras respecto a una población objetivo, cabe resaltar que mientras más grande sea la muestra (y por tanto, si se asume heterogeneidad en los valores), más se asemejará a las observaciones poblacionales. En este sentido, una muestra pequeña tendrá una mayor variabilidad muestral, ya que la extracción de observaciones será mucho más limitada, tendiendo a una mayor posibilidad de error estándar (según la fórmula de EE) y a una media/proporción incierta respecto al parámetro poblacional. Con todo, se concluye que la afirmación A es la correcta.

**1.3 Gracias a las propiedades de las distribuciones normales, siempre podemos transformar una variable aleatoria distribuida normalmente en una variable aleatoria que se distribuya como la distribución normal estándar. Entonces, si X N ~ (10, 25), entonces Z sería? Desarrolle su respuesta y resuelva el ejercicio.**

Para obtener la variable aleatoria distribuida normalmente y estandarizada por puntaje Z se puede usar la siguiente fórmula: [Z = (X - μ) / Desviación estándar].

Como X N ~ (10, 25) se refiere a “la distribución normal de la variable X que tiene media 10 y desviación estándar 25”, tendríamos que aquella distribución en puntaje Z de la variable X está dada por: Z = (X – 10) / 5.

Un ejemplo del ejercicio sería considerar un X=15 (suponiendo variable aleatoria distribuida normalmente), lo que en este caso significaría que el mismo valor en puntaje Z es de la forma: Z = 15 – 10 / 5 = 1.

Para encontrar un valor se podría tomar cualquier número aleatorio de la muestra de X y considerarlo a partir de Z ~ (0, 1), lo que significa que la distribución normal estándar tiene media 0 y desviación estándar 1. Ahora, a partir del ejemplo, se puede mostrar que para el valor X=15, Z es igual a 1, lo que se interpreta como que el valor 15 de la variable X es 1 desviación estándar por encima de la media de distribución estándar.

**1.4 Si X = {1,1,1,0,0}, ¿Qué es P(X=1), también conocido como p? Desarrolle su respuesta y resuelva el ejercicio.**

P(X=1) significa “la probabilidad de que x sea 1”. En este sentido, p como probabilidad define que x puede tomar cualquier valor de la muestra X determinado por las frecuencias de los valores en el espacio muestral en específico (en este caso un espacio muestral con cinco valores posibles). Con todo, como el valor 1 (posibilidad de que x sea 1) se repite tres veces de un total de cinco posibilidades, se aplica la fórmula de la probabilidad:

[p = n° de veces que ocurre evento 1 / n° total de eventos],

luego, la probabilidad p es igual a 3/5, o, dicho en términos porcentuales, la probabilidad de que x tome valor 1 es de un 60%.

**1.5 Un grupo de investigadores realizo un experimento en el que asignaron aleatoriamente a estudiantes a salas de clases de tamaño pequeño y de tamaño regular. Los investigadores querían estimar el efecto causal de asistir a una clase de tamaño pequeño sobre distintos resultados educativos. Nos centraremos en este análisis en el efecto causal sobre el rendimiento de los estudiantes en la prueba de lectura. Porque el tratamiento fue asignado**  
**aleatoriamente, podemos asumir que los estudiantes que asistieron a la sala pequeña eran comparables (antes de asistir a la escuela) a los estudiantes que asistieron a la sala de tamaño regular. Las variables de interés para este análisis corresponden a: reading, que da cuenta del puntaje obtenido en la prueba de lectura y la variable classtype, la que da cuenta de si los estudiantes fueron asignados a la sala de clase de tamaño pequeño o regular.**  
**En base a estos antecedentes:**

**a) Defina la hipótesis nula y alternativa de los investigadores que realizaron el experimento. Establezca el Alpha con que desarrollara sus análisis.**

La hipótesis nula definida por los investigadores plantea que no existen diferencias significativas entre las medias de los resultados en la prueba de lectura (variable reading), lo que implica que no existe efecto respecto al tamaño de las salas de clase en el rendimiento de los estudiantes. Por otro lado, la hipótesis alternativa dice relación con que sí existen diferencias estadísticamente significativas en los resultados de la prueba de lenguaje entre los estudiantes que estaban en la sala pequeña y los que estaban en la sala regular. Para el análisis que sigue se considera un nivel de confianza del 95% (predeterminado en test realizado en Stata, ver imagen), por tanto, el nivel de significancia establecido (alpha) es igual a 0,05.

**b) Describa los resultados del cuadro**

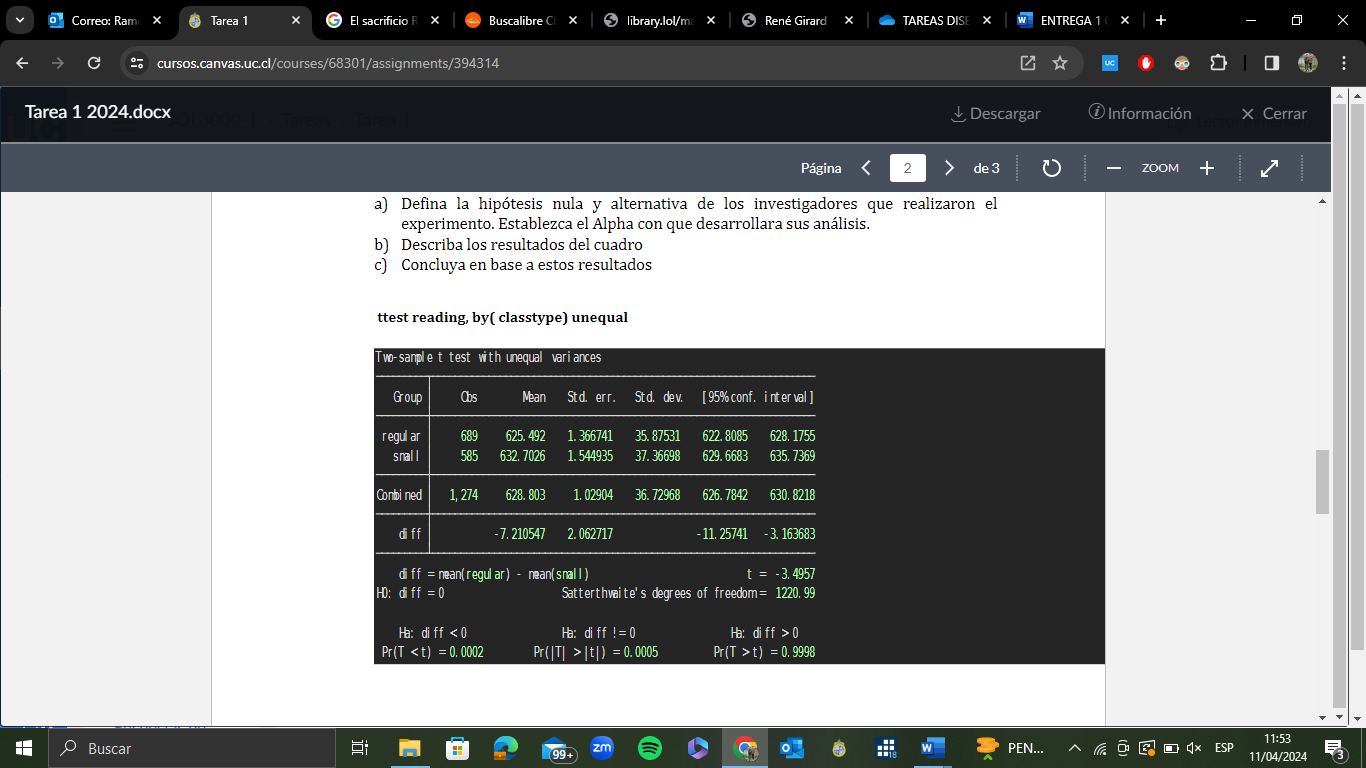
En primer término, se observa que los estudiantes de la sala regular tuvieron de media 625.5, comparados con los de la sala pequeña, que reportan de media 632.7 en la prueba de lenguaje. En principio se observa una diferencia considerable entre ambas medias. Además de esto, se puede reportar que la prueba de hipótesis realizada considera la heteroscedasticidad en los datos obtenidos (diferencia de varianza en los errores estándar).

Ahora bien, pasando al análisis más en detalle. Los intervalos de confianza reportados para el caso de los estudiantes en la sala regular son [622.8 - 628.2]. Por otro lado, los intervalos para aquellos en la sala pequeña son [629.7 - 635.7]. Lo que quiere decir que los intervalos no se solapan entre sí a un 95% de confianza. Esto mismo se corresponde al observar el valor-p (y el estadístico t) para el caso de la hipótesis nula (no existen diferencias entre medias). En este caso, el valor-p reportado es por mucho menor al nivel de significancia establecido (0.0005 < 0.05), lo que también se corresponde con el valor del estadístico t (-3.49) que es grande y no entra en el área sombreada de confianza que aceptan la hipótesis nula.

**c) Concluya en base a estos resultados**

Con todo el análisis desarrollado anteriormente, es posible ahora establecer que i) como los intervalos de confianza no se solapan entre sí y ii) como el valor-p es por mucho menor al valor de significación establecido, en efecto existen diferencias estadísticas significativamente considerables para no aceptar la hipótesis nula, y por tanto para concluir que los resultados de las medias sí son diferentes entre sí. Así, se concluye que sí existen diferencias en los resultados entre quienes aleatoriamente estuvieron en el grupo de la clase regular frente a aquellos en la sala pequeña. Son estos últimos quienes reportaron mejor desempeño en los resultados de la prueba de lenguaje, lo que permite establecer que el factor de las salas pequeñas puede ser un factor de incremento en los resultados estudiantiles.

**ttest reading, by( classtype) unequal**



***Parte 2***

**Para responder a las siguientes preguntas utilice STATA y la base de datos escogida para realizar el trabajo de investigación durante el semestre. Ustedes deben describir esta base de datos brevemente y las variables que han seleccionado para trabajar durante el curso (versión reducida de la base de datos). Es relevante que deben dar cuenta de manera sintética y entendible todas las decisiones metodológicas realizadas.**

**1. Realice un resumen estadístico de las variables escogidas (versión reducida de su base de datos). Indique cuál es su variable dependiente (la que quiere explicar o predecir más adelante en el curso) y las variables independientes (sociodemográficas, de control, etc) con las que trabajaran en sus análisis (Máximo 600 palabras). La variable dependiente debe ser una variable cuantitativa no binaria. Presentar en una tabla editada un resumen estadístico de estas variables. 2. Desarrolle, si es necesario, recodificación de sus variables independientes y limpieza de datos (con casos perdidos y/o sin información). Expliquen estos procedimientos en el do file. Todos los procedimientos realizados deben estar en el do file que adjuntaran a su trabajo.**

Resumen estadístico

La base de datos seleccionada es la Encuesta Bicentenario del 2019. Dado que solo se encuentra publicada hasta la base de datos del 2020 y, como aquella fue aplicada en pandemia, es que preferimos ocupar la base de datos anterior y conocer a profundidad aquella situación pre-pandemia y pre-estallido social. Lo interesante de esta base de datos es que enfatiza en preguntas de contenido cultural, educacional y religioso y en tal sentido quisimos indagar más a fondo estas cuestiones.

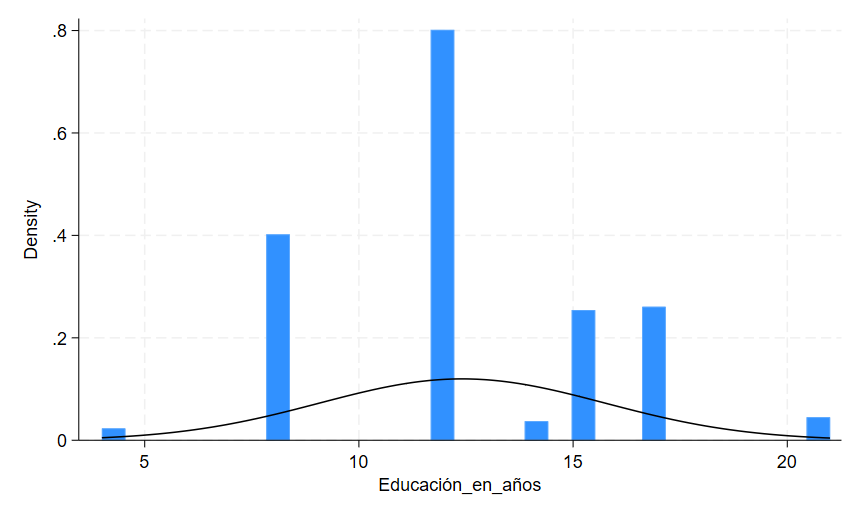
Nuestra variable dependiente de interés es la educación medida en años (y por tanto posibilitada de analizarse como una variable continua no dicotómica). En este sentido, luego de borrar los missings de dicha variable, procedimos a recodificarla dado que la pregunta del cuestionario se construía en base a las categorías de la educación (educación media completa, técnica incompleta, etc.) y por tanto debíamos transformar estas categorías a datos numéricos para construir una variable continua. La decisión metodológica utilizada fue traducir las categorías etiquetadas como “completas” al año educacional que se aparejaba (Básica completa: 8 años, universitaria completa: 17 años, et.) y traducir aquellas etiquetadas como “incompletas” a la suma de la última categoría completa más la mitad de la duración estándar de la categoría incompleta (para más especificaciones, ver do file).

Las variables de control/sociodemográficas que fueron escogidas son: variable sexo, nivel socioeconómico (dicotomizada como “clase alta/media y clase baja”) y la posición política. También, como posible variable dependiente dejamos la variable Asistencia religiosa, para detallar la descripción y establecer posibles relaciones entre el perfil religioso, político y sociodemográfico de quienes no tienen duraciones extensas de años educacionales y quienes sí.

El N total de la muestra, luego de la elección de variables y la limpieza de los datos, es de 1.420 casos. Como variable continua se tiene la Educación por años (variable dependiente), además de contar con dos variables binomiales que son Sexo/Género y NSE (variables sociodemográficas) y dos variables categóricas, Asistencia religiosa y Posición política (variables de control).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Media e intervalos de confianza de variables de interés | | | |  |  |
| Variable | N | Mean | Std. err. | [95% conf. | interval] |
| Educación por años | 1.420 | 12,41 | 0,09 | 12,24 | 12,59 |
| Asistencia religiosa | 1.420 | 4,05 | 0,03 | 3,98 | 4,11 |
| Sexo | 1.420 | 1,61 | 0,01 | 1,58 | 1,63 |
| NSE | 1.420 | 4,01 | 0,02 | 3,97 | 4,05 |
| Posición política | 1.420 | 2,91 | 0,03 | 2,85 | 2,97 |

**3. Genere una buena visualización de su variable dependiente, indique si esta variable tiene una distribución normal. Exporte el gráfico en formato png.**



El gráfico expuesto presenta la distribución de la variable de la educación en años, representado en el eje x, mientras que el eje y la densidad de la frecuencia. Ahora para poder distinguir si es una distribución normal hay que considerar una serie de conceptos estadísticos. Estos conceptos son la forma, la curtosis y la asimetría. El primero referido a la distribución de los datos en conjunto, en una distribución simétrica, los datos están equilibrados alrededor de un punto central.

Sobre la curtosis, describe cómo los datos se distribuyen en relación con la normalidad. La curtosis de una distribución leptocúrtica es alta y sus colas son más pesadas, mientras que la de una distribución platicúrtica es baja y sus colas son más ligeras. Con respecto al tercer concepto se indica la falta de simetría en una distribución de datos, pudiendo ser positiva o negativa.

Considerando lo mencionado, la variable se distribuye de manera normal. Pero es sesgada a la derecha, ya que hay más casos hacia ese lado, por lo tanto, la cola de ese lado se extiende más allá. Por otro lado, se acumulan muchos casos al centro.

**4. Estime la media e intervalos de confianza para dos variables cuantitativas y dos proporciones. Desarrolle estos análisis a distintos niveles de confianza. Presente sus resultados en una tabla y explique por qué el intervalo de confianza cambia en la medida que aumenta el nivel de confianza (Máximo 700 palabras).**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Medias e intervalos de confianza a 95% y 99% de confianza | | | |  |  |  |  |
| Variable | N | Mean | Std. err. | [95% conf. | interval] | [99% conf. | interval] |
| Educación por años | 1.420 | 12,4127 | 0,0883 | 12,2394 | 12,5860 | 12,1848 | 12,6405 |
| Asistencia religiosa | 1.420 | 4,0500 | 0,0300 | 3,9800 | 4,1100 | 3,9600 | 4,1300 |
|  |  |  |  |  |  |  |  |
| Proporciones e intervalos de confianza a 95% y 99% de confianza | | | | |  |  |  |
| Variable | N | Proportion | Std. err. | [95% conf. | interval] | [99% conf. | interval] |
| Género | 1.420 | 0,6092 | 0,0129 | 0,5832 | 0,6346 | 0,5751 | 0,6424 |
| NSE (clase alta/baja) | 1.420 | 0,7923 | 0,0108 | 0,7702 | 0,8131 | 0,7632 | 0,8193 |

La variable de interés "Educación por años" registró 1420 casos, con una media de 12.4127 y un error estándar aproximado de 0,0883. Por otro lado, se estimó que la verdadera media poblacional, para un intervalo del 95%, estaba entre 12.2394 y 12.5860. La media se encuentra entre 12.1848 y 12.6405 para un intervalo del 99%.

Basado en 1420 observaciones, presenta una media de 4.0500 y un error estándar de 0.0300 para la asistencia religiosa. Con una confianza del 95%, la media es de 3.9800 a 4.1100. La media para el nivel de confianza del 99 % puede oscilar entre 3.9600 y 4.1300.

Con un total de 1420 observaciones, se estimó una proporción de aproximadamente 0.6092 para la variable "Género", con un error estándar de 0.0129. Con un nivel de confianza del 99 %, el intervalo de confianza calculado oscila entre aproximadamente 0,5751 y 0,6424. Sin embargo, para un nivel de confianza del 95%, los intervalos oscilan entre 0.5832 y 0.6346.

Se estimó una proporción de aproximadamente 0,7923 para la variable "NSE (clase alta/baja)", con un error estándar de 0,0108, en cuanto al nivel socioeconómico (NSE). Utilizando un nivel de confianza del 99 %, el intervalo de confianza calculado va desde aproximadamente 0.7632 hasta 0.8193. El intervalo, con un nivel de confianza del 95%, va desde 0.7702 hasta 0.8131.

Finalmente, el intervalo de confianza se calcula utilizando una variedad de elementos. El tamaño de la muestra, que afecta la precisión de la estimación, es uno de ellos. El intervalo de confianza generalmente será más estrecho a medida que aumenta el tamaño de la muestra porque la variabilidad estimada será menor.

Otra parte importante es el error estándar. Este es un cálculo de la variabilidad de la estimación utilizando la desviación estándar y el tamaño de la muestra. La estimación será más precisa y el intervalo de confianza será más pequeño cuanto menor sea el error estándar.

El tercer componente es el valor crítico, que está relacionado con la distribución de la población y el nivel de confianza seleccionado. A medida que aumentamos el nivel de confianza, aumenta el valor crítico correspondiente. La amplitud del intervalo de confianza se determina con este valor crítico. Por lo tanto, estamos eligiendo un valor crítico más extremo al aumentar el nivel de confianza, lo que resulta en un intervalo de confianza más amplio para permitir una mayor seguridad en nuestra estimación.

**5. Realice test de hipótesis para las variables utilizadas en la pregunta anterior. Defina las hipótesis nula y alternativa, el valor t/z observado (según corresponda) y sus conclusiones estadísticas y sustantivas a un 95% y 99% de nivel de confianza para cada una de las variables en cuestión. Analice o genere intervalos a un mismo nivel de confianza para validar sus resultados. ¿Qué puede concluir a partir de estos resultados? (Máximo 600 palabras)**

Primeramente, se testea la hipótesis que relaciona la Educación por años con el Sexo/Género de los encuestados. En este sentido, la hipótesis nula refiere a que no existen diferencias en las medias de los años educacionales alcanzados, por ende, tanto hombres como mujeres reportan años educativos iguales (medias de 12,58 y 12,3 años educativos, respectivamente). La hipótesis alternativa es que sí existen diferencias significativas entre las medias de los años educativos mediados por género. Los resultados de la prueba de hipótesis invitan a aceptar la hipótesis nula tanto al 95% como al 99% de confianza. Esto porque en ambas situaciones los intervalos de los hombres y las mujeres se solapan, además de que el valor-p es mayor al nivel de significancia (0,12>0,05) y el valor t observado es pequeño (1,51; lo que significa que está dentro de los intervalos de significancia del área bajo la curva de +/-2,58 –al 99%- y de +/-1,96 –al 95%).

La segunda hipótesis por testear refiere a la tendencia de los años educativos respecto al nivel sociodemográfico. En este caso, como la variable NSE está dicotomizada en “clase alta/media” y “clase baja”, la hipótesis nula refiere a que no existen diferencias significativas entre las medias de los años educativos de los individuos de la clase alta y aquellos de clase baja. La hipótesis alternativa corresponde a la afirmación de que sí existen diferencias estadísticas entre las medias de ambos grupos. Las medias en cuestión son de 14,7 años educativos para la clase alta/media y de 11,8 para la clase baja. La hipótesis nula no se puede aceptar porque los intervalos son diferentes entre sí, ya que no se solapan. El intervalo para la clase alta es [14,2-15,2] y de la clase baja [11,6-12,1], al 99% de confianza, y menos será, por tanto, a un 95% de confianza, ya que los intervalos serán más restringidos en ese caso. Esto se corresponde en que el valor-p señalado es 0,000 (mucho menor al nivel de significancia del 0,05 –al 95%- y del 0,01 –al 99%), lo que también se confirma al analizar que el estadístico t es 13,5, muy por sobre cualquier restrictivo del área bajo la curva a cualquiera de los dos niveles de confianza.

Finalmente, se realiza un test de hipótesis de la Educación por años referido a la variable Asistencia religiosa. Esta última se dicotomizó para una mejor utilización, en donde 0 refería a personas que asistían al menos una vez a la semana a un evento religioso regular, mientras que 1 a aquellos que nunca asistían. La hipótesis nula dice relación con que no existen diferencias en las medias de los años educativos de quienes asisten versus quienes no asisten. La hipótesis alternativa, por otro lado, apunta a que sí existen diferencias significativas en las medias de los años educacionales entre quienes asisten y no asisten a eventos religiosos periódicos. En este caso particular, la hipótesis nula no se puede aceptar al 95%, es decir, existen diferencias significativas, pero sí se acepta a un 99% de confianza, es decir, no existen diferencias significativas entre las medias. Al 95%, los intervalos de confianza no se solapan ([11,8-12,3] para quienes asisten y [12,4-12,9] para quienes no asisten), y se concluye i) que la hipótesis nula no puede aceptarse, ii) que las medias de los años educacionales son significativamente diferentes y iii) en específico, que aquellos que no asisten a eventos religiosos periódicamente tienen de media más años educacionales que aquellos que sí asisten. Sin embargo, al 99% de confianza, y como en este caso los intervalos extienden su rango, ahora los intervalos en cuestión sí se solapan y se acepta la hipótesis nula de que no hay diferencias significativas en las medias consideradas. Tiene sentido este cambio debido al nivel de confianza al analizar que tanto el valor-p (0,001) y el estadístico t (-3,28) se reportan como valores muy cercanos a las fronteras de los niveles de confianza estándar, lo que produce que a un cambio del nivel de confianza puedan reportarse conclusiones de hipótesis distintas.