Universidad de La Sabana

Parte Asincrónica

Fecha: 19 de marzo de 2021

Análisis pruebas de estado

Gabriel Gerardo Amaya Becerra

Maestría en Analítica Aplicada

Debido a que el manejo de la segunda lengua es cada vez más indispensable para ingresar al mercado laboral, el acceder a estudios de educación superior internacional, entre otras. El presente análisis busca encontrar si existe una diferencia en el manejo del segundo idioma inglés para las personas que presentaron la prueba saber 11 (año de grado del colegio) y quienes además de ingresar a la educación superior profesional presentaron el examen saber pro, lo que asegura que a la fecha han cursado al menos el 75% del programa académico, con lo cual se espera poder evaluar si la universidad cumplió con su promesa de valor respecto al aseguramiento del aprendizaje, para este caso enfocado en el segundo idioma.

Para lograr este objetivo se ha dividido el análisis en dos partes, donde la primera espera realizar un acercamiento a partir del comparativo de los estudiantes que presentaron el saber 11 en el año 2015 y de este mismo grupo quienes presentaron la prueba saber pro en 2019, teniendo así dos mediciones de la misma variable para una misma persona.

El segundo componente del análisis busca validar si existe una diferencia en las varianzas de los estudiantes que presentaron la prueba saber 11 y saber pro en 2019, esto permitirá observar los indicadores de ingreso (saber 11) vs los indicadores de salida (saber pro).

Antes de iniciar el análisis se realizará una breve descripción de las bases de datos y la metodología de tratamiento de datos usada:

1. Acceso a la base de datos:
    a. La base de datos se encuentra disponible para consulta y descarga en el siguiente enlace: https://www.icfes.gov.co/investigadores-y-estudiantes-posgrado/acceso-a-bases-de-datos
    b. El código utilizado se encuentra disponible en:

2. Composición de la base de datos:
    a. Estructura base saber 11 (aplica para periodos 2015-1, 2015-2, 2019-1, 2019-2):

        La base se encuentra compuesta por alrededor de 82 variables que a nivel agregado contiene 6 variables de información personal como el tipo de documento, el género, la fecha de nacimiento, el código de la prueba saber 11, 57 variables de información de características individuales, 19 variables de información sobre la prueba.

        A partir de lo anterior y enfocados en el objetivo de análisis se realizó la depuración de la base de datos, dejando la siguiente estructura:

        ESTU_CONSECUTIVO: Contiene el código de tipo texto, bajo el cual el estudiante presentó la prueba saber 11 para el periodo respectivo.

        PUNT_INGLES: Corresponde al puntaje en número obtenido por el estudiante en la prueba saber 11, se maneja en una escala de calificación de 0 a 100.

    b. Estructura base saber pro (aplica para el año 2019):

        La base se encuentra compuesta por 105 variables, de forma agregada 6 variables responden a información personal como el tipo de documento, el género, fecha de nacimiento, código de la prueba saber pro, entre otras. 74 variables corresponden a características individuales de la persona, 25 variables corresponden a información de la prueba

        A partir de lo anterior y enfocados en el objetivo de análisis se realizó la depuración de la base de datos, dejando la siguiente estructura:

        ESTU_CONSECUTIVO: Contiene el código de tipo texto, bajo el cual el estudiante presentó la prueba saber pro para el periodo respectivo.
        MOD_INGLES_PUNT: Corresponde al puntaje en número obtenido por el estudiante en la prueba saber pro, se maneja en una escala de calificación de 0 a 300.

    c. Estructura códigos de pruebas: Identificador del código de la prueba saber 11 y código de la prueba saber pro.

        La base de datos está compuesta por dos variables:
        ESTU_CONSECUTIVO_11: Contiene el código de tipo texto, bajo el cual los estudiantes han presentado la prueba saber 11.
        ESTU_CONSECUTIVO_PRO: Contiene el código de tipo texto, bajo el cual los estudiantes presentaron la prueba saber pro.

3. Base de datos para análisis:
    a. Estructura de la base de datos comparativo saber 11 (2015) vs saber pro (2019): Se realizó el siguiente procedimiento para la construcción de la base de datos:
        i. Cruce de la base de datos saber 11 (2015) con la estructura de códigos de pruebas, utilizando la llave correspondiente para extraer la información del puntaje de inglés.
        ii. Cruce de la base de datos saber pro (2019) con la estructura de códigos de pruebas, utilizando la llave correspondiente para extraer la información del puntaje de inglés.
        iii. Los códigos que no quedaron con información en los puntajes fueron retirados de la base de datos, debido a que no presentaron la prueba saber 11 en 2015 o no presentaron la prueba saber pro en 2019, con lo cual no hacen parte de nuestro estudio.

        A partir de lo anterior la estructura de bases de datos que será utilizada es la siguiente:

            i. ESTU_CONSECUTIVO_11: Contiene el código de tipo texto, bajo el cual el estudiante presentó la prueba saber 11 para el periodo 2015.

j. ESTU_CONSECUTIVO_PRO: Contiene el código de tipo texto, bajo el cual el estudiante presentó la prueba saber pro para el periodo 2019.
k. PUNT_INGLES_S11_2015: Corresponde al puntaje en número obtenido por el estudiante en la prueba saber 11, se maneja en una escala de calificación de 0 a 100.
l. PUNT_INGLES_SP_2019: Corresponde al puntaje en número obtenido por el estudiante en la prueba saber pro, se maneja en una escala de calificación de 0 a 300.

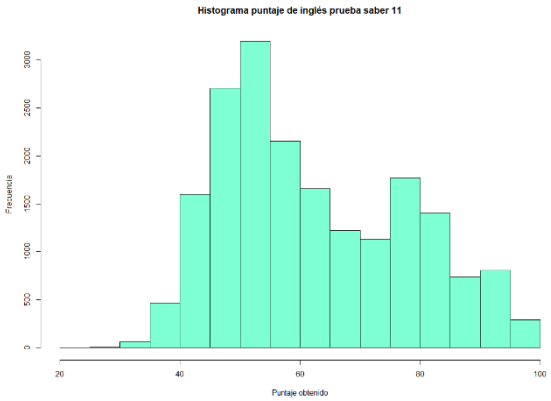
Diferencia de medias Saber 11 (2015) vs Saber Pro (2019)

Para el primer análisis se tomo la base de datos construida presentada previamente, con la cual se busca responder a la siguiente pregunta:

¿Existe evidencia estadística que sugiere que la media obtenida en el resultado en la prueba de inglés del saber pro (2019) es superior a la media obtenida en la prueba saber 11 (2015) para los estudiantes que presentaron estas pruebas en los periodos respectivos?

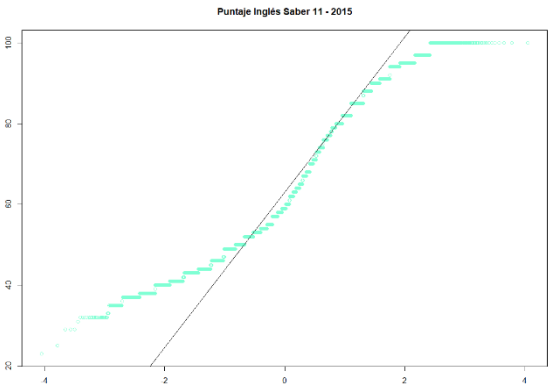
Antes de realizar la prueba de hipótesis empezaremos por validar si ambas variables cumplen con los supuestos de normalidad. Para ello se revisará la distribución de la variable a través de un histograma, se realizará el QQplot que permite observar gráficamente si se aproxima el comportamiento a una normal y se realizarán las pruebas de normalidad de Anderson Darling y Lilliefors, el nivel de significancia (alfa) que será manejado para todas las pruebas será de 5% o 0.05.

a. Variable: Puntaje Inglés Saber 11 – 2015:



El histograma de frecuencias nos permite ver que la variable parece tener un comportamiento normal no simétrico, debido a que la mediana (59) y media (63.1) no son la misma y como se observa en el gráfico se observan mayor cantidad de distribución de registros hacia la derecha que hacia la izquierda.

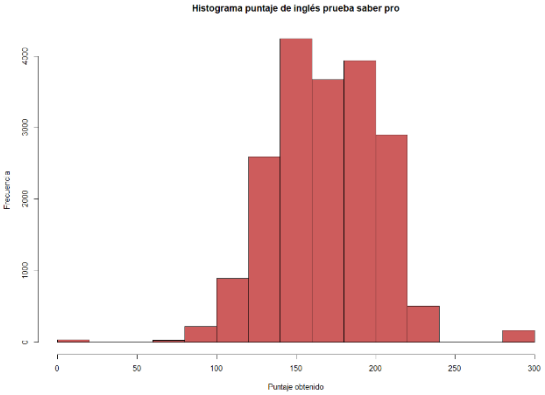
El gráfico QQ por su parte nos muestra que, aunque la distribución no es 100% normal, si tiende el centro de la distribución a estar muy cercano a línea de normalidad. Sin embargo, para no dejar a subjetividad la decisión de si es normal o no, se utilizaron las pruebas de normalidad de Lilliefors y Anderson Darling los cuales tienen como hipótesis nula que la población sigue una distribución normal.



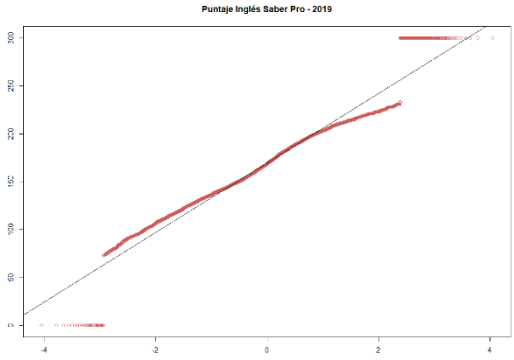

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Pruebas_Estado$PUNT_INGLES_S11_2015
D = 0.11662, p-value < 2.2e-16

> Ander_S11

        Anderson-Darling normality test

data:  Pruebas_Estado$PUNT_INGLES_S11_2015
A = 319.25, p-value < 2.2e-16
```

A partir de los p-values obtenidos y a un nivel de significancia de 0.05 se observa que la hipótesis nula es rechazada, por lo cual la muestra no presenta normalidad.

b. Variable: Puntaje Inglés Saber Pro – 2019:



Revisando el histograma de la segunda variable encontramos que parece seguir más el comportamiento de una distribución normal, y basados en que la media (169.3) y mediana (169) se encuentran con diferencias de decimales podría empezar a intuirse que si presenta un comportamiento normal.



Al revisar el gráfico QQ plot se observa que un grupo grande de datos se encuentra alrededor de la línea de normalidad, sin embargo, se observan algunos datos que están en los limites superiores e inferiores de la curva normal, por lo cual nuevamente se recurrirá a las pruebas de normalidad las cuales ayudarán a reforzar si la variable sigue o no una distribución normal.

Al revisar las dos pruebas de normalidad realizadas observamos nuevamente que la población no sigue una distribución normal a un nivel de significancia del 5%.

```
        Anderson-Darling normality test

data:  Pruebas_Estado$PUNT_INGLES_SP_2019
A = 52.467, p-value < 2.2e-16

> Lillie_SP

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Pruebas_Estado$PUNT_INGLES_SP_2019
D = 0.033558, p-value < 2.2e-16
```

A partir de lo anterior se recurrirá a una técnica de normalización de variables, de tal forma que sea posible realizar la prueba de diferencia de medias. La técnica de normalización está guiada por la siguiente formula:

$$N_x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Una vez realizada la normalización de las variables se procede a realizar la prueba de hipótesis, dónde el sistema de hipótesis es el siguiente y criterio de rechazo es el siguiente:

$$H_0: \mu_d = 0 \qquad H_a: \mu_d > 0 \qquad dónde\ \mu_d = \mu_{saber\_pro} - \mu_{saber\_11}$$

$$Si\ p - valor\ < alfa\ entonces\ existe\ diferencia\ de\ medias\ positiva\ dónde\ alfa\ el\ alfa\ será\ 0.05$$

Es importante resaltar que debido a que sobre un mismo individuo se están realizando dos mediciones entonces estamos ante una prueba pareada, por lo cual no es posible realizar pruebas de validación de varianzas iguales dado que el supuesto de la prueba F en este caso es que ambas muestras son independientes.

Basado en los resultados obtenidos de la prueba se puede concluir que se rechaza Ho debido a que p-value es menor que alfa, lo que significa que a un nivel de significancia del 5% existe evidencia estadística que sugiere que la media de la prueba saber pro en inglés es superior a la prueba saber 11 en inglés, con lo cual la propuesta de valor ofertada por las universidades en términos de segunda lengua parece estarse cumpliendo.

```
        Paired t-test

data:  Pruebas_Estado$Norm_Saber_Pro_Ing and Pruebas_Estado$Norm
_Saber_11_Ing
t = 45.906, df = 19202, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than
 0
95 percent confidence interval:
 0.04195432        Inf
sample estimates:
mean of the differences
             0.04351353
```

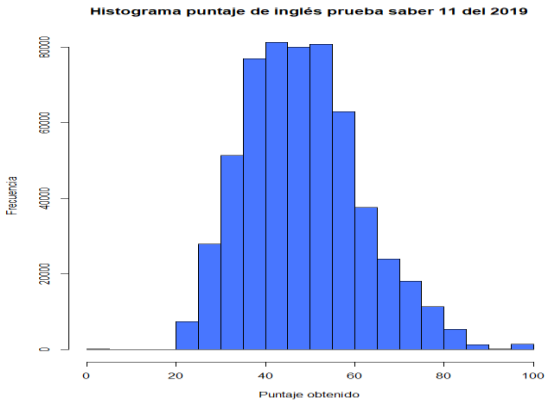
Diferencia de varianzas Saber 11 (2019) vs Saber Pro (2019)

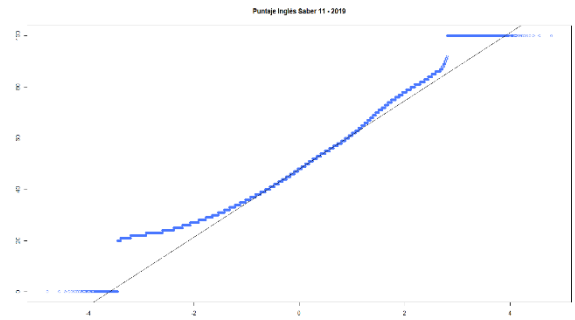
Para la segunda parte del análisis se realizará a partir de comparar los resultados de todos los estudiantes que presentaron el saber 11 para el año 2019 y los estudiantes que presentaron el saber pro en el 2019, la pregunta a responder en este caso es:

¿Existe evidencia estadística que sugiere que la varianza de los resultados de las pruebas de ingles del saber 11 y saber pro del año 2019 son diferentes a 1?

Para este caso se hará uso de la prueba F, sin embargo, debido a que esta prueba es altamente sensible si no existe normalidad en las distribuciones, antes de realizar la prueba se validará si las variables siguen una distribución normal utilizando el mismo método de la sección anterior. Dos aspectos importantes a resaltar es que en el tratamiento de datos es que se retiraron aquellas personas que no presentaban ningún resultado en la prueba, de igual forma que se generó una columna denominada selección a través de números aleatorios para la base saber 11 con la cual se realizará el test F.

a. Variable: Saber 11 2019





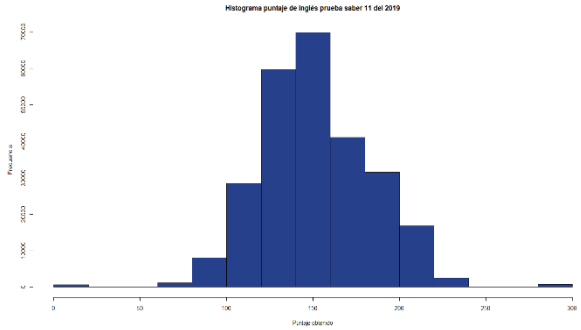
El histograma para este caso nos muestra una distribución normal asimétrica debido a que nuevamente tenemos una mayor cola hacia la parte derecha de la distribución, presentando de nuevo datos extremos de 0 y 100, pero siendo en frecuencia de mayor número los ubicados hacia el valor máximo.

Al revisar el QQ-plot, observamos que los datos atípicos que se discutían previamente se encuentran en el extremo de la distribución y que a nivel general la parte inferior de los puntos se encuentra alejada de la línea de normalidad, y cada uno de estos por encima de la misma. Motivo por el cual se usará el resultado de las pruebas para definir la existencia de normalidad.

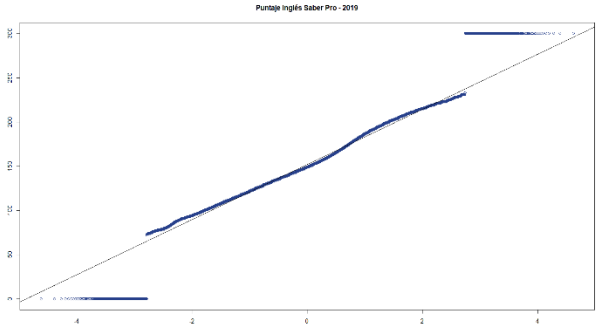
Basado en el test de normalidad se identificó que la distribución no sigue una normal, debido a que el p-value es inferior al nivel de significancia determinado, motivo por el cual se recurrirá a realizar un proceso de normalización de la variable antes de realizar la prueba de varianzas.

```
        Anderson-Darling normality test

data:  Saber_11_2019$PUNT_INGLES
A = 2007.2, p-value < 2.2e-16

> Lillie_S11_2019

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Saber_11_2019$PUNT_INGLES
D = 0.048109, p-value < 2.2e-16
```

b. Variable: Saber Pro-2019



El histograma nos permite observar una distribución más similar a la normal, con algunos datos extremos tanto en el nivel inferior como en el nivel superior, esta distribución presenta una media de 152 y una mediana de 149 por lo cual no puede afirmarse que es completamente simétrica.



Por su parte el gráfico QQ-Plot nos muestra que los registros para este caso si están más cercanos a la línea de normalidad, sin embargo, que existe un número importante de datos en los externos superior e inferior de la distribución normal, ante lo cual será mejor terminar de validar la hipótesis de normalidad usando las pruebas respectivas.

Para este caso nuevamente las pruebas presentan un p-value inferior al nivel de significancia establecido ante lo cual se realizará un procedimiento de normalización de la variable previo a realizar la prueba de varianzas.

```
        Anderson-Darling normality test

data:  Saber_Pro_2019$MOD_INGLES_PUNT
A = 819.17, p-value < 2.2e-16

> Lillie_SP_2019

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Saber_Pro_2019$MOD_INGLES_PUNT
D = 0.049918, p-value < 2.2e-16
```

Una vez realizado el proceso de normalización de las variables se procede a usar la prueba F para determinar si existe diferencia en las varianzas de los resultados obtenidos entre las pruebas.

El sistema de hipótesis y criterio de rechazo que se utilizará en este caso es:

$$H_o: \frac{\sigma^2_{Saber\_Pro}}{\sigma^2_{Saber\_11}} = 1 \qquad H_a: \frac{\sigma^2_{Saber\_Pro}}{\sigma^2_{Saber\_11}} \neq 1$$

$$Si\ p - valor\ < alfa\ entonces\ la\ relación\ de\ las\ varianzas\ \ es\ diferente\ de\ 1\ dónde\ alfa\ \ es\ 0.05$$

```
        F test to compare two variances

data:  Saber_Pro_2019$Norm_Saber_Pro_2019 and Saber_11_2019_Modified$Norm_Saber_11_2019
F = 0.69786, num df = 260674, denom df = 260674, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6925203 0.7032363
sample estimates:
ratio of variances
         0.6978577
```

Teniendo en cuenta el p-value obtenido y al contrastarlo con el alfa se rechaza la hipótesis nula por lo cual se puede concluir que con un nivel significancia del 5% que las varianzas entre las dos pruebas es diferente de 1, siendo por tanto diferentes los resultados obtenidos entre las dos pruebas presentadas por estudiantes de bachilleres vs los estudiantes pregrado.