INFORME I: INFERENCIA

Alejandro Romero Serrano - 2182059
Mauricio Romero Jaimes - 2182049
William Romero Serrano - 2151090

Estadística II
Grupo 6
Profesor: Carlos Mantilla
19 de agosto de 2021

Introducción

En el presente informe se busca resolver preguntas planteadas inicialmente por el grupo de trabajo con base en la información disponible respecto a los exámenes saber pro de años anteriores a disposición en la plataforma por medio de un análisis estadístico básico que comprende muestreo, gráficas e inferencia como se planteó en las clases de estadísticas y en los notebooks disponibles en la plataforma. Para el desarrollo de este trabajo se implementó lenguaje R para analizar la información disponible y obtener los tamaños de muestra ideales para abordar de forma óptima los interrogantes planteados y llegar a conclusiones lógicas sustentadas en los métodos de inferencia sugeridos. A lo largo del informe se explicará de forma detallada cada paso, el código implementado, y el por qué de las conclusiones a las cuales se llegó al final del trabajo.

Planteamiento

La meritocracia es un sistema de gobierno en el cual el poder lo ejercen las personas mejor preparadas según sus méritos, extendiéndose hasta convertirse en una creencia popular según la cual, las personas llegan a donde quieren solo gracias al esfuerzo y dedicación con que abordan sus proyectos y metas. Teniendo esto en cuenta, la información de las pruebas saber-pro de los últimos años representa una oportunidad de estudiar el supuesto de la meritocracia, analizando resultados contra diversas condiciones a las que se enfrentan el conjunto de estudiantes durante el transcurso de sus carreras, por este motivo, se decide abordar un cuestionamiento importante:

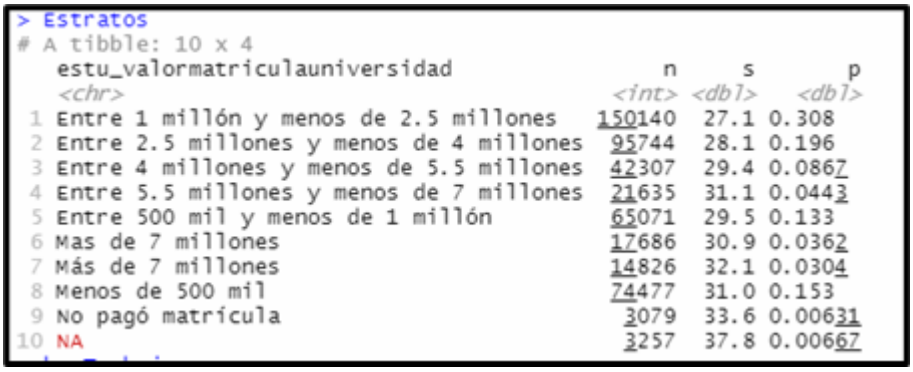
¿Cómo influye el estrato y las horas de trabajo de los estudiantes en su rendimiento en la parte del inglés?

Desarrollo

Al ser atributos explícitamente cualitativos, es necesario discretizar las horas de trabajo y el estrato (según el supuesto de que entre más estrato, mayor valor de matrícula)

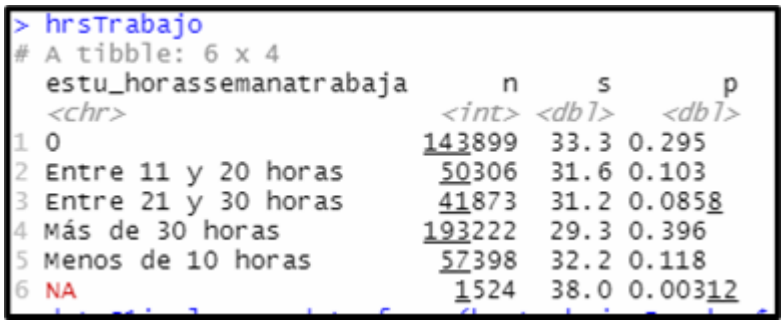
en las opciones que se muestran en la información. Respecto a los estratos, se seleccionan los atributos correspondientes al valor de la matrícula de la universidad, y el puntaje de inglés; Obteniendo la cantidad de estudiantes (n), la desviación estándar(s) y el porcentaje(p) de cada uno de estos estratos para la población total. En el caso de horas de trabajo, se toman las variables, de las horas que trabaja el estudiante a la semana y el puntaje de inglés, agrupando por cantidad de horas, obteniendo los mismos datos que para los estratos. Con base en esta información se busca analizar cómo las horas de trabajo y el estrato socioeconómico de un estudiante pueden afectar su desempeño en diferentes áreas de estudio, para el caso explicado anteriormente se enfoca en la materia de inglés ya que se tenía en cuenta esta variable Nota: Cabe destacar que debido a la tilde en "Mas" para "Mas de 7 millones" en el pago de matrícula, se presentan dos grupos por separado.

Figura 1. Clasificación de estratos de estudiantes por valor de matrícula.



```
> Estratos
# A tibble: 10 x 4
   estu_valormatriculauniversidad                 n     s       p
   <chr>                                      <int> <dbl>   <dbl>
 1 Entre 1 millón y menos de 2.5 millones    150140  27.1 0.308
 2 Entre 2.5 millones y menos de 4 millones   95744  28.1 0.196
 3 Entre 4 millones y menos de 5.5 millones   42307  29.4 0.0867
 4 Entre 5.5 millones y menos de 7 millones   21635  31.1 0.0443
 5 Entre 500 mil y menos de 1 millón          65071  29.5 0.133
 6 Mas de 7 millones                          17686  30.9 0.0362
 7 Más de 7 millones                          14826  32.1 0.0304
 8 Menos de 500 mil                           74477  31.0 0.153
 9 No pagó matrícula                           3079  33.6 0.00631
10 NA                                          3257  37.8 0.00667
```

Figura 2. Clasificación de horas de trabajo de estudiantes.



```
> hrsTrabajo
# A tibble: 6 x 4
  estu_horassemanatrabaja      n     s       p
  <chr>                    <int> <dbl>   <dbl>
1 0                       143899  33.3 0.295
2 Entre 11 y 20 horas      50306  31.6 0.103
3 Entre 21 y 30 horas      41873  31.2 0.0858
4 Más de 30 horas         193222  29.3 0.396
5 Menos de 10 horas        57398  32.2 0.118
6 NA                        1524  38.0 0.00312
```

Para la determinación del tamaño, la media, la desviación estándar, simetría y kurtosis; primero se crean los distintos dataframes enfocados en cada materia, estos contienen las horas de trabajo por semana, el valor de la matrícula y el puntaje respectivo de los estudiantes.

Se examinan las poblaciones, las cuales son las variables que se definen sobre un universo. De acuerdo a esta información se realizan gráficas de distribución para cada materia (Histogramas), teniendo en cuenta su puntaje y frecuencia respectiva.

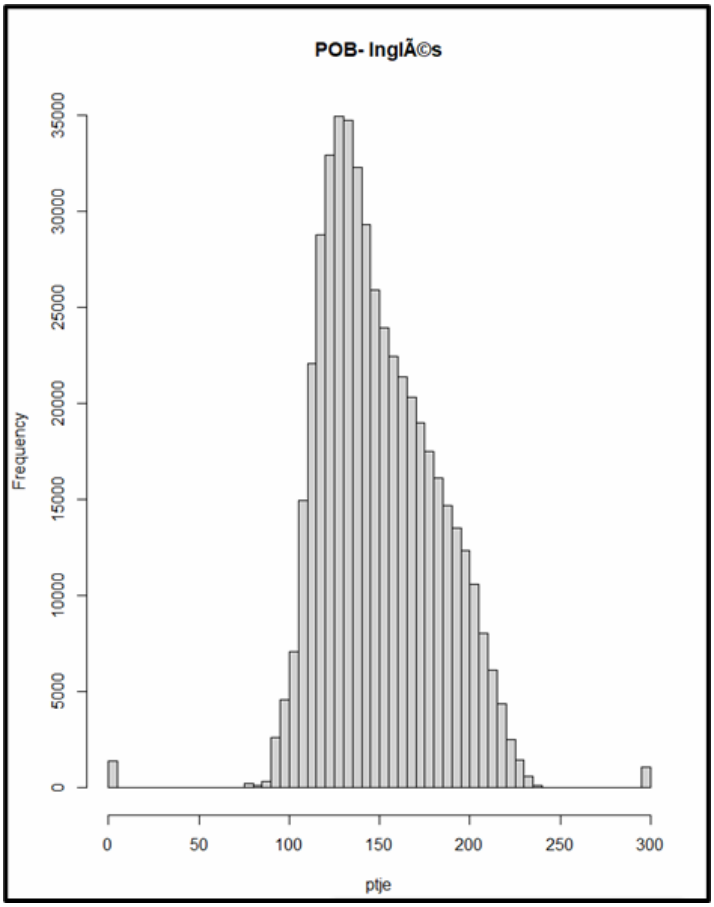
Figura 3. Poblaciones de estudiantes, y respectivas medidas por materia.

	popIngles	popLectura	popciudadana	popCuantitativo	popEscrita
tamaÃ±o	488222.0	488222.0	488222.0	488222.0	473647.0
media	150.2	150.1	146.0	149.3	150.5
desviaciÃ³n estÃ¡ndar	31.8	31.0	32.0	30.5	31.4
simetrÃa	0.3	0.0	-0.1	0.2	0.4
kurtosis	1.4	-0.3	0.0	0.1	1.3

>

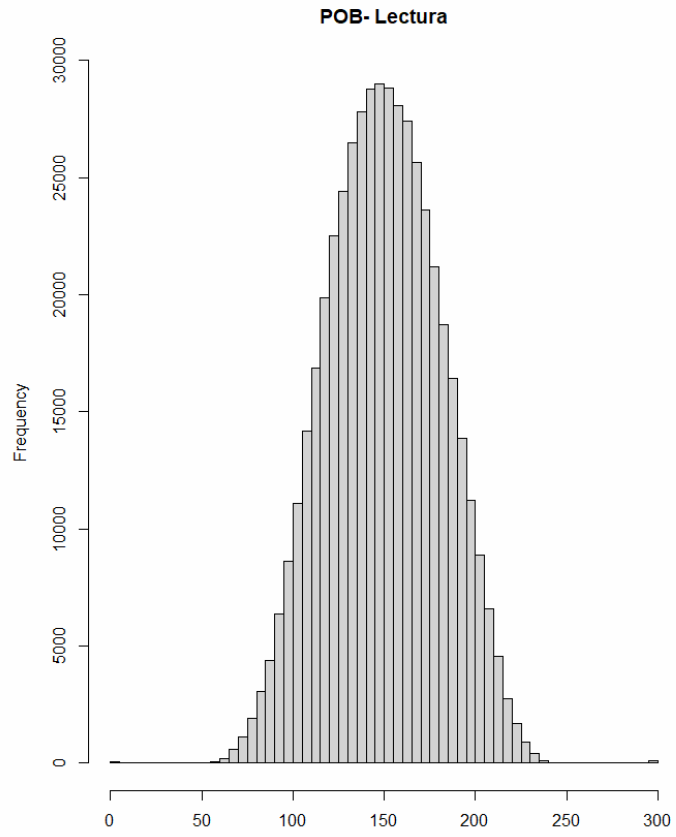
Analizando la gráfica para cada materia se puede decir:
En inglés las frecuencias más altas se encuentran entre 125 y 150, obteniendo una media de 150.2 , la desviación estándar muestra que los datos no son tan dispersos aunque es un poco alta por su primer y último valor dando como resultado 31.8, presenta una simetría más hacia la derecha se puede notar por el valor de 0.3. El dato de curtosis 1.4 indica que es una distribución Leptocúrtica y se aprecia en la gráfica que los datos se encuentran algo concentrados en la media y la curva es apuntada. Ver figura 4.

Figura 4. Población de estudiantes en pruebas de inglés.



Para el caso de lectura se puede observar una gráfica simétrica donde este valor se ve en la tabla como 0, las frecuencias más altas están entre 145 y 155, obteniendo la media en 150.1, la desviación estándar supone que los datos presentan poca dispersión aunque se obtiene 31.0 a causa de que algunos datos se encuentran lejanos a la media. La curtosis -0.3 ligeramente platicúrtica, indica que la gráfica es un poco achatada y sus valores no se concentran tanto en la media. Ver Figura 5.

Figura 5. Población de estudiantes en pruebas de lectura.

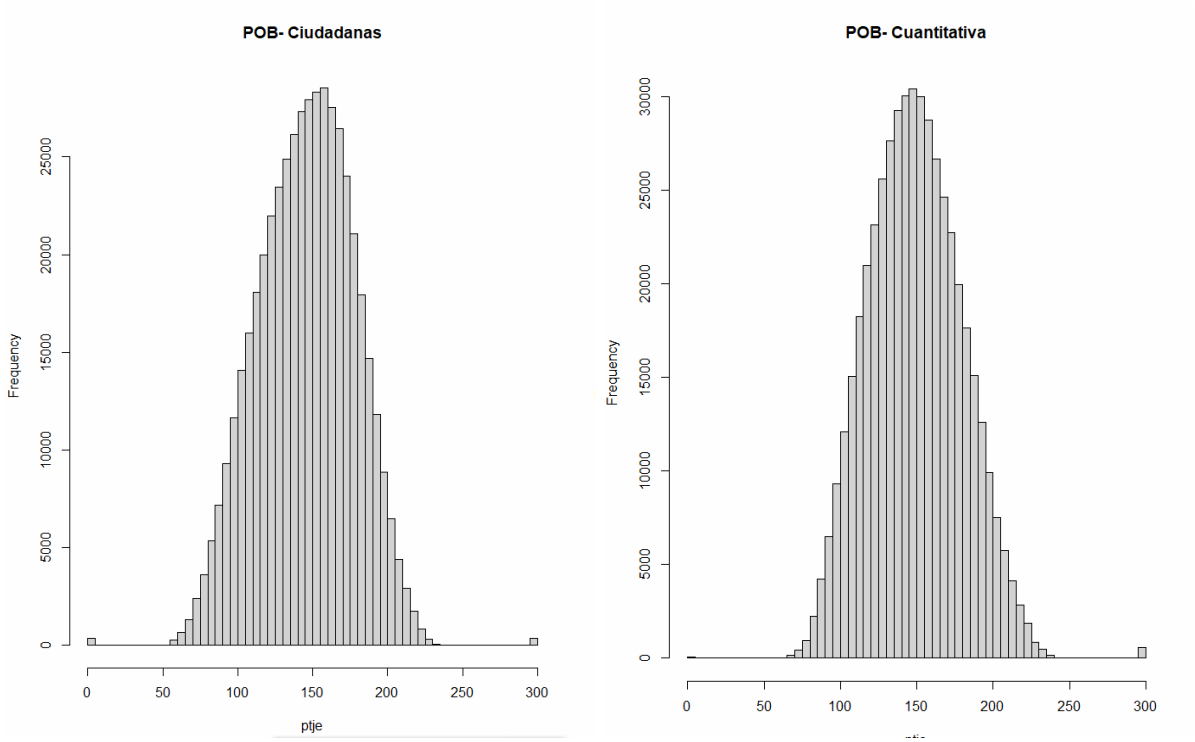


La materia competencias ciudadanas presenta una simetría para la izquierda observada en su resultado -0.1, las mayores frecuencias están entre 140 y 160, teniendo una media de 150.1, el dato de curtosis indica que es mesocúrtica, otorgando una gráfica que se aproxima a una distribución normal. La desviación estándar se comporta como la anterior dando 32.0 por motivo de gran cambio en algunos valores pero se puede decir que los datos tienen una dispersión baja con respecto a la media. Y, respecto a cuantitativa, la gráfica parece algo simétrica aunque no lo es totalmente, esto se puede ver en su valor de simetría de 0.2, las mayores frecuencias están entre 140 y 160, obteniendo un valor de media de 149.3, la curtosis es leptocúrtica con un valor de 0.1 por ende se observa una gráfica apuntada y con sus datos concentrados a la media. La desviación estándar se comporta de la misma manera a las anteriores con un valor de

30.5 a razón de los cambios en algunos valores, aunque surja este resultado se puede suponer que los datos no presentan mucha dispersión con relación a la media. Ver figura 6.

Figura 6. Población de estudiantes en pruebas de ciudadanas y cuantitativa.

(a) **(b)**



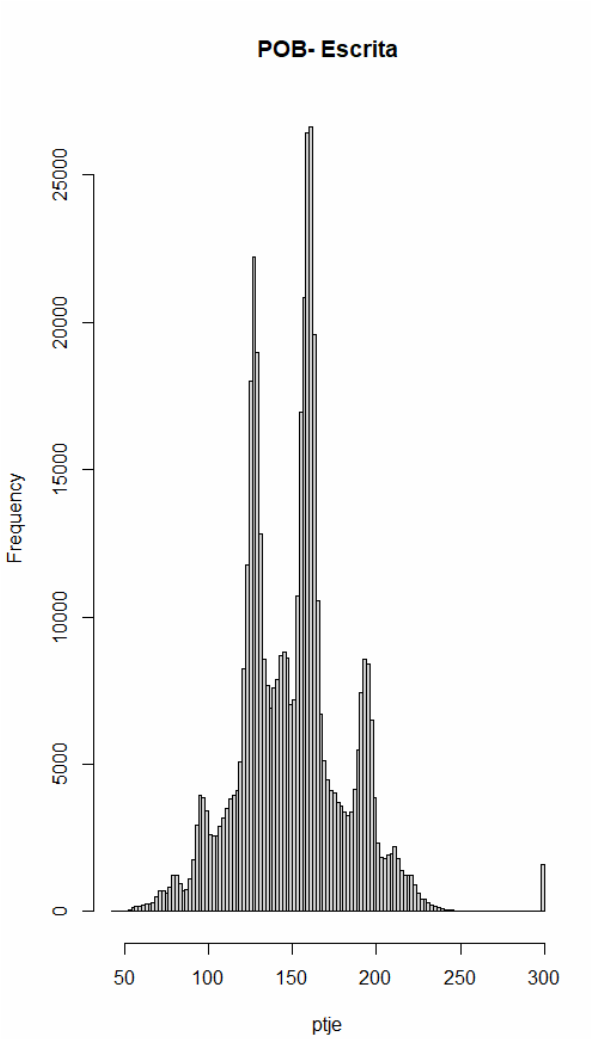
En comunicación escrita, la gráfica no presenta nada de simetría, las frecuencias más altas se encuentran entre 150 y 165, obteniendo una media de 150.5, la desviación estándar de 31.4 indica que los datos no tienen alta dispersión a proporción de la media aunque aparece este resultado por la variabilidad en los datos. La gráfica presenta una simetría hacia a la derecha de un valor 0.4 y su curtosis de 1.3 indica que tiene una distribución leptocúrtica y que sus datos están ligeramente concentrados respecto a la media otorgando una gráfica apuntada. Además para este caso se obtiene una gráfica bimodal, donde hay dos modas, aproximadamente una en 130 y otra en 155.. Ver figura 7.

Para el análisis de cada materia no se utilizó el nivel de confianza por ser la recolección de los datos en bruto, siendo además útil para posteriormente realizar el muestreo.

Continuando la realización del estudio se hizo un muestreo para determinar el tamaño de la muestra para cada materia, el muestreo consiste en determinar las características sobre una población a partir de una muestra, buscando que esta sea representativa a la población, en donde en conjunto con un error a la media del 10 % y luego del 5 % (aunque con un error del 2.5 % la muestra es más confiable), en la materia de

inglés, se usa de 3.755 %. Los datos de la desviación estándar, un nivel de confianza del 95 %, y el tamaño total de la información se obtienen los tamaños respectivos de cada materia para luego usarlos en algún análisis específico que se quiera realizar enfocado a esta información.

Figura 7. Población de estudiantes en pruebas de escrita.



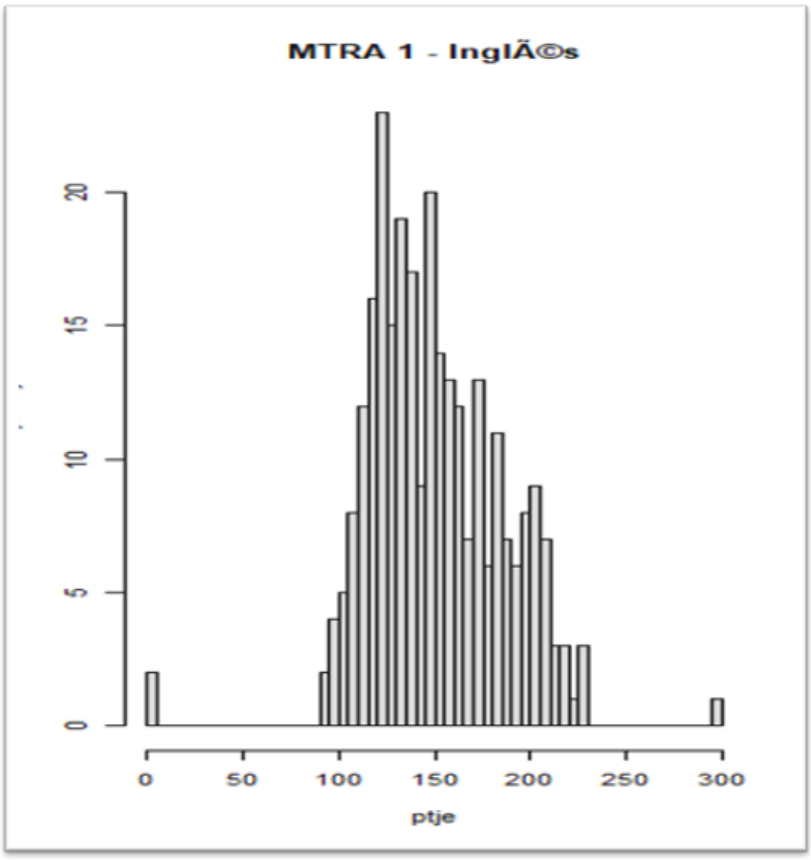
Dichos algunos conceptos preliminares, es adecuado definir la inferencia ya que se va a estar utilizando a lo largo del informe, la inferencia es el conjunto de métodos y técnicas que permiten basados en la muestra, saber cual es el comportamiento de una determinada población con un riesgo de error medible en términos de probabilidad.

Siguiendo la secuencia del código se programaron un diagrama para analizar en inglés, el cual generaba una gráfica del POB que ya fue descrita anteriormente junto a dos gráficas de muestras de las cuales se puede decir lo siguiente:

La primera muestra relaciona la frecuencia con respecto al puntaje que se obtiene

en inglés, analizándolo se puede observar cómo la concentración de puntajes se presenta entre 90 y 230, además hay unas frecuencias bajas existentes para el puntaje más alto y el más bajo. Figura 8.

Figura 8. Primera muestra a la población original de inglés.



La segunda muestra nuevamente relaciona una frecuencia comparándolo con el puntaje de inglés en este caso las mayores frecuencias se concentran entre 80 y 240 aproximadamente, para sus puntajes mínimos y máximos hay una frecuencia pero muy baja; esta gráfica presenta un pequeño grado de similitud respecto a distribución con la población inicial. Ver figura 9.

Como podemos ver las muestras obtenidas tienen una distribución similar a la población, denotando que el muestreo es correcto por el surgimiento de estos parecidos con respecto a la población inicial. Figura 10.

Para continuar el análisis de cómo influye el estrato y las horas de trabajo en el pun-

Figura 9. Segunda muestra de inglés.

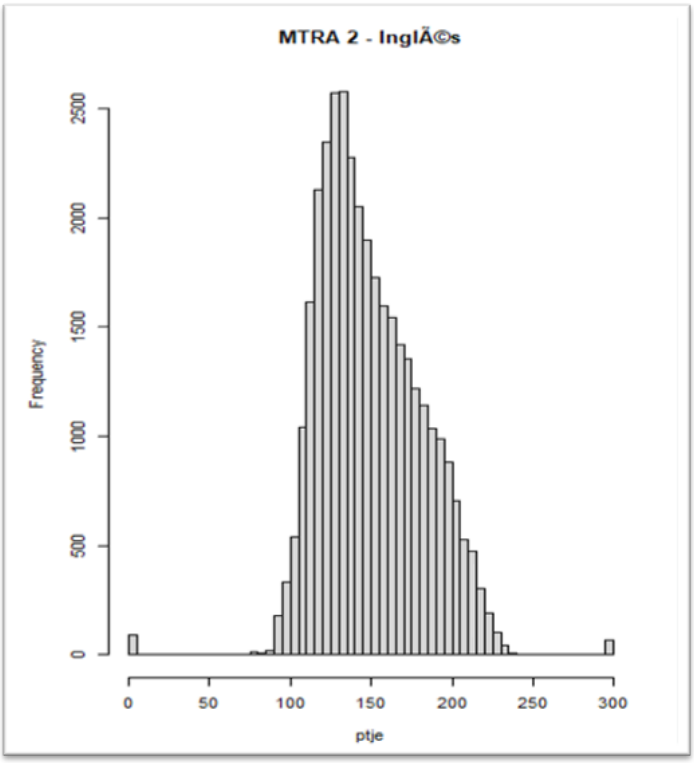
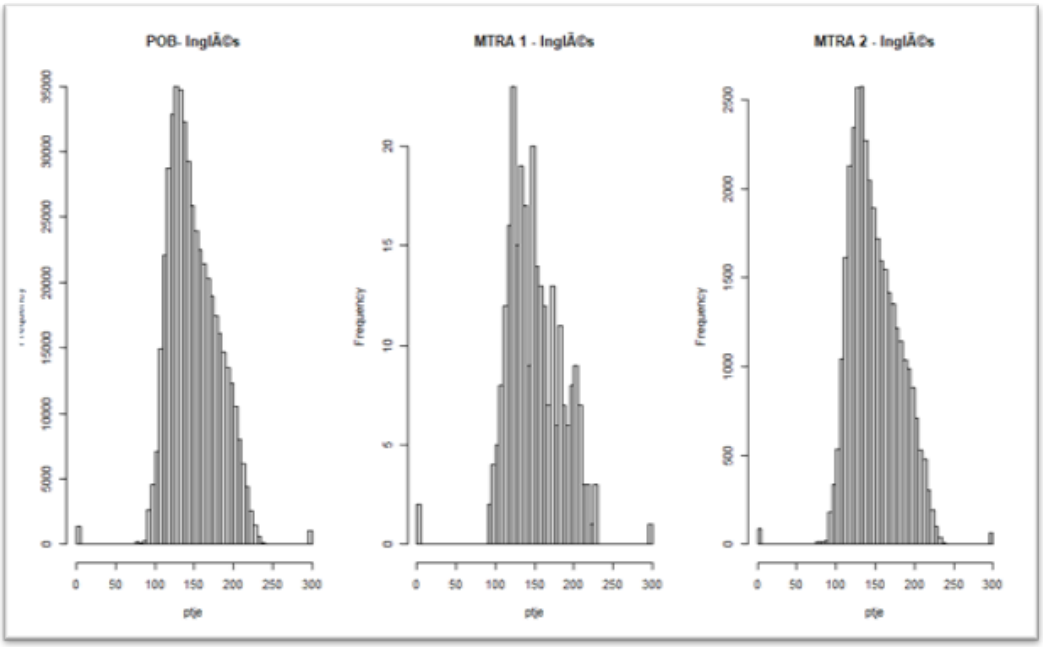


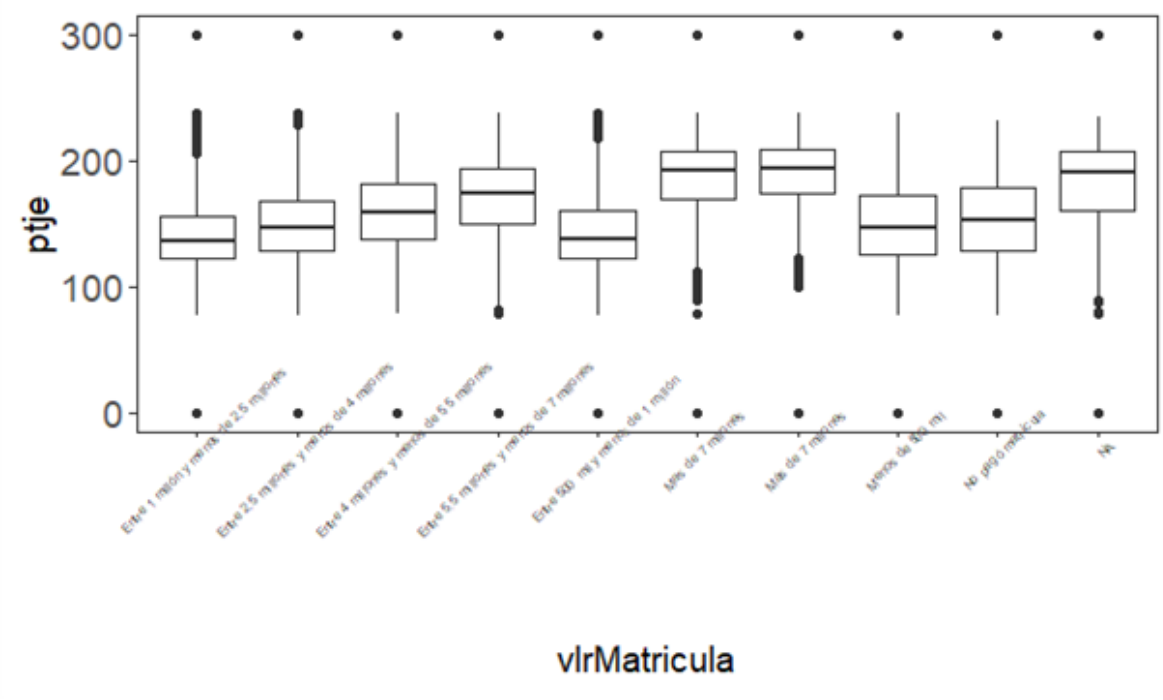
Figura 10. Población original en inglés y dos muestras aleatorias.



taje de inglés, se realizan diagramas de cajas a dependencia de los valores de matrícula (1er y 2do diagrama) y las horas de trabajo (3er y 4to diagrama), se utilizará la población para cada caso y posterior a esta distribución, una muestra representativa correspondiente.

En la figura 11, se reflejan los diagramas de cajas correspondientes a cada caso:

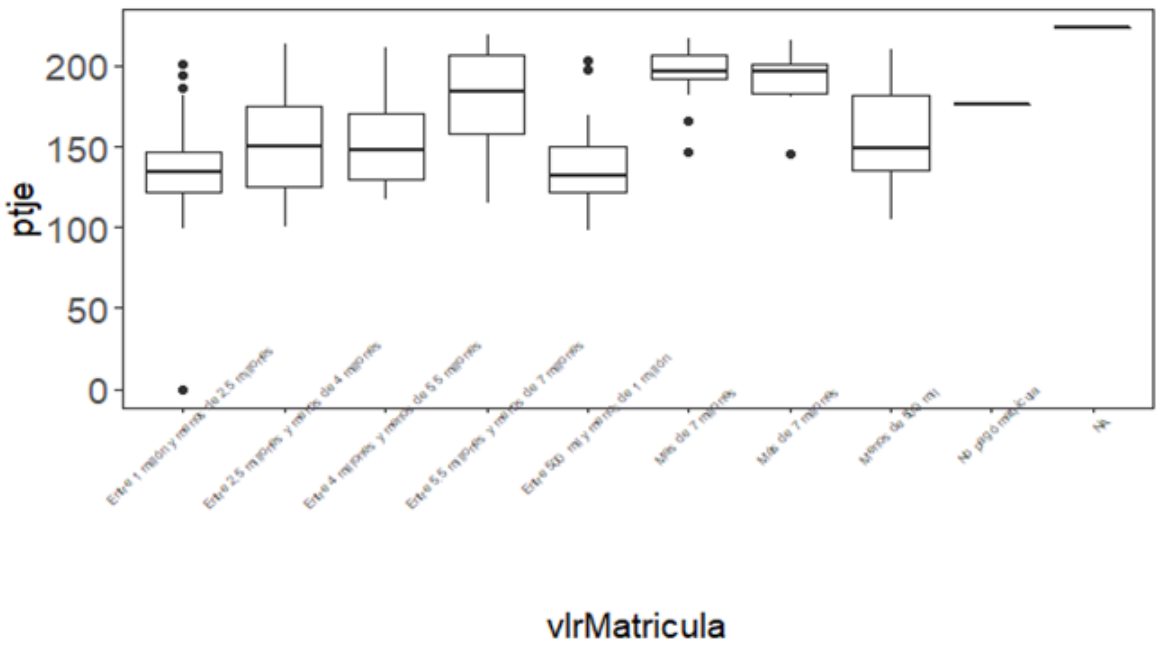
Figura 11. Población de estudiantes en inglés. Puntaje - Valor de matrícula.



El primer diagrama refleja el valor de matrícula a comparación del puntaje en inglés utilizando la población de los estudiantes, analizando se puede concluir, las personas que pagan mayor matrícula obtienen mejores puntajes en esta materia, pues la mediana de las cajas para estos estudiantes está en un puntaje alto, los estudiantes entre 500 mil y menos de 2.5 millones, presentan los puntajes más bajos con las medias más baja, y para los demás estratos tienen un rendimiento regular; notando que entre el precio de la matrícula va aumentando el desempeño de los estudiantes es mejor, de manera directamente proporcional, aunque exceptuando a las personas sin pago de matrícula porque estas tienen buen desempeño. Figura 11.

El segundo diagrama refleja nuevamente el valor de matrícula a comparación del puntaje en inglés, pero ahora se utiliza una muestra de la población, al comparar este diagrama con el anterior tiene un análisis similar, a diferencia de los tamaños de las cajas pues cambian un poco, pero nuevamente se puede decir a mayor pago de matrícula es mejor el puntaje obtenido en esta materia, exceptuando los estudiantes sin pago de matrícula las cuales se desempeñan adecuadamente, notando que las personas de

Figura 12. Muestra de estudiantes en inglés. Puntaje - Valor de matrícula.

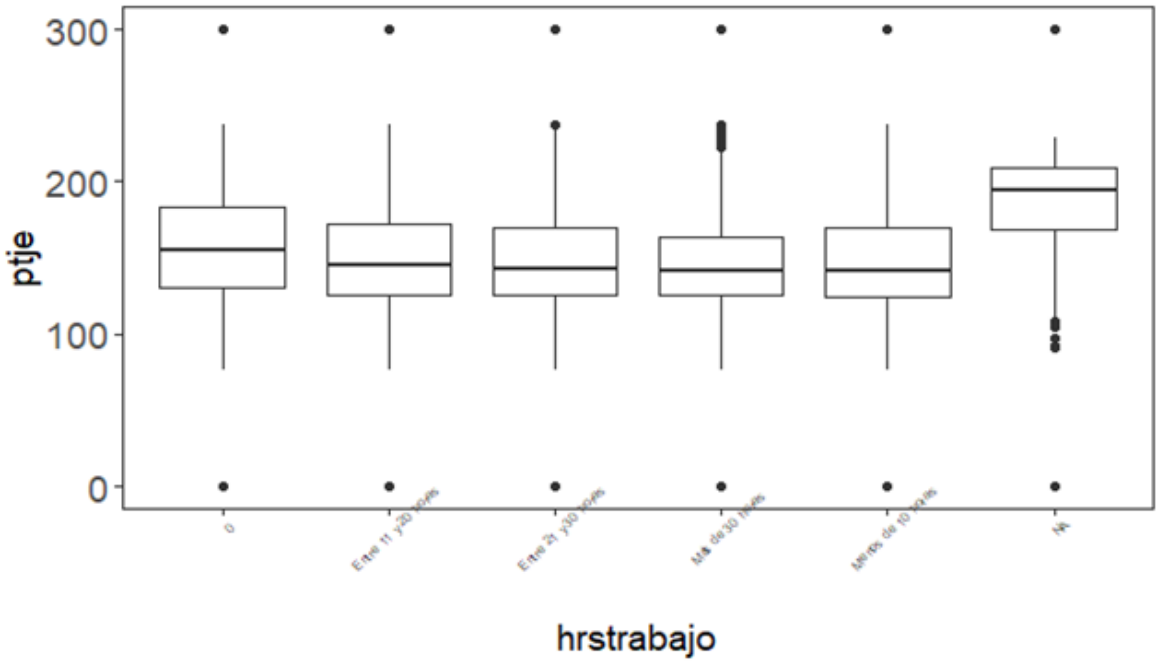


estratos altos tienen una mayor concentración en sus cajas y sus valores como medias se encuentran en buenos puntajes, las personas entre 500 mil y menos de 2.5 millones presentan rendimiento de una manera regular y cada vez que se observan las demás cajas este desempeño aumenta un poco. Figura 12.

La figura 13 es un análisis con respecto a la población, en donde se compara el puntaje en la materia de inglés con respecto a las horas de trabajo del estudiante, denotando que las cajas son un poco similares, los estudiantes que no tienen horas de trabajo presentan un desempeño un poco mayor con respecto de los que trabajan, y los estudiantes NA tienen mayor desempeño que todos desconociendo si presentan o no horas de trabajo, y las personas que tienen horas de trabajo tienen unos puntajes aceptables con una dispersión y mediana mínimamente parecida, por ende con puntajes afines. Y, el último diagrama, nuevamente es una paridad entre horas de trabajo y el puntaje de inglés, para este caso se toma la muestra de la población de los estudiantes, los que no trabajan presentan una alta dispersión pero algunos con buenos puntajes, entre 21 y 30 horas la caja tiene bastante concentración con algunos puntos atípicos y con un desempeño aceptable, la caja de NA es la que tiene mayor concentración y allí se encuentran las personas con mejores puntajes, las cajas sobrantes tienen una leve similitud en su alta dispersión y en su mediana Figura 14.

La varianza de la muestra es 3.275294, y la desviación estándar es de 1.809777.

Figura 13. Población de estudiantes en inglés. Puntaje - horas de trabajo.



Se realiza un remuestreo con la finalidad de saber si el estimador es insesgado o no, o sea realizar la medición del sesgo, planteando 10.000 combinaciones y buscando la media como mediana de estas muestras, para luego comparar cual de estas dos presenta un menor sesgo, denotando así cual varianza es más eficiente. A pesar de que la mediana es más robusta que la otra medida de tendencia central, en este cálculo se puede observar como la media es más insesgada con un valor de -0.001687681 mientras que la mediana es más amplia de 0.11385.
En los cálculos de eficiencia para cada estimador podemos comprobar lo dicho anteriormente, la media es más útil en este análisis con un resultado de eficiencia de 3.305218 mientras que la mediana arroja un valor de 7.727086.

Por último se hizo un cálculo de consistencia, realizando distintas muestras en las cuales se va aumentando su tamaño lo que permite observar que a medida que se esto se aumenta el estimador se va pareciendo más al parámetro, algo similar sucede con la distribución, a medida que la muestra crece la distribución de probabilidad tiende a ser semejante a la distribución normal, si esto sucede se puede decir que el estimador presenta consistencia.

En la figura 15 se puede observar la población donde se encuentran la cantidad de los estudiantes en comparación al puntaje de inglés, y a continuación la siguientes

Figura 14. Muestra de estudiantes en inglés. Puntaje - horas de trabajo.

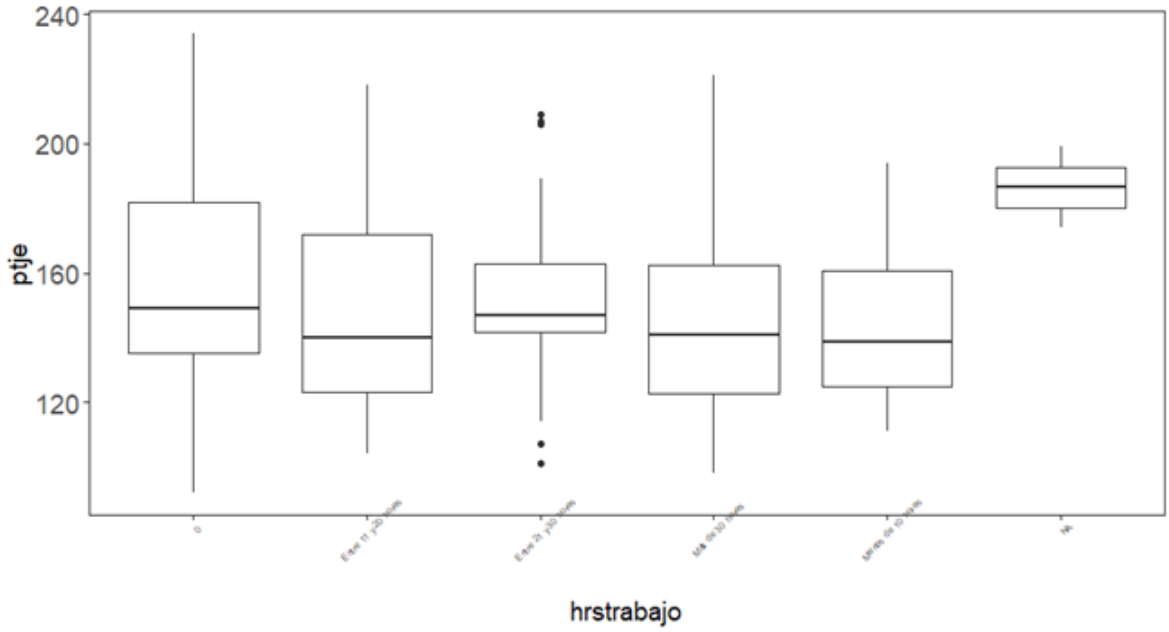
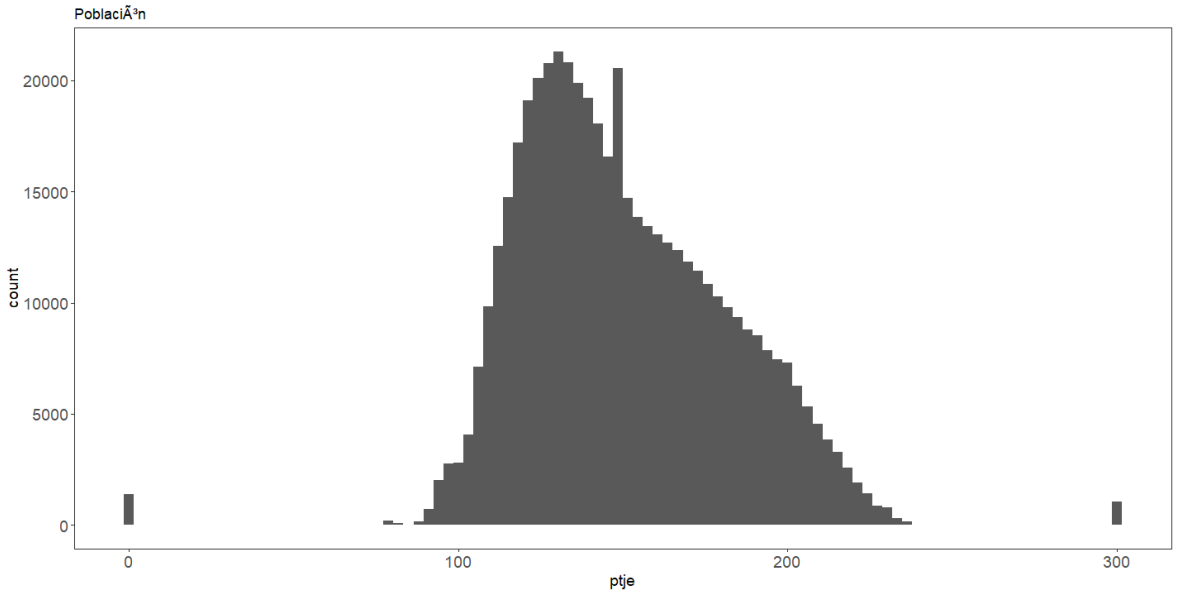


Figura 15. Población en consistencia.



gráficas son tres muestras diferentes donde se va variando la muestra para demostrar lo descrito anteriormente.

En la primera muestra con un tamaño de 100 no presentan mucha consistencia a comparación de la población. Figura 16.

Figura 16. Muestra tamaño 100 en consistencia.

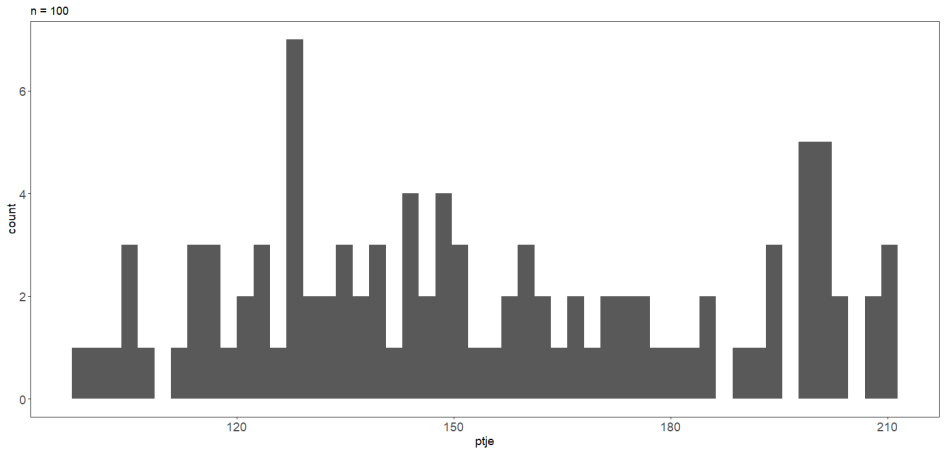
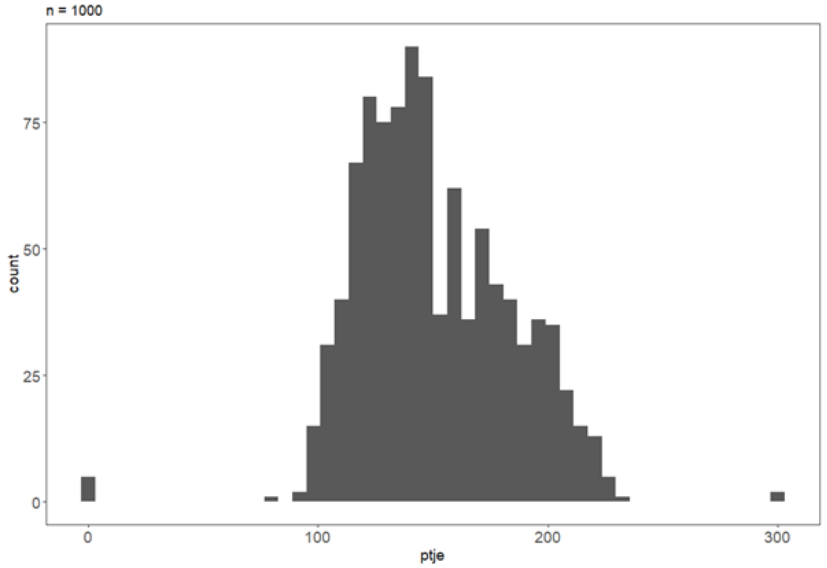


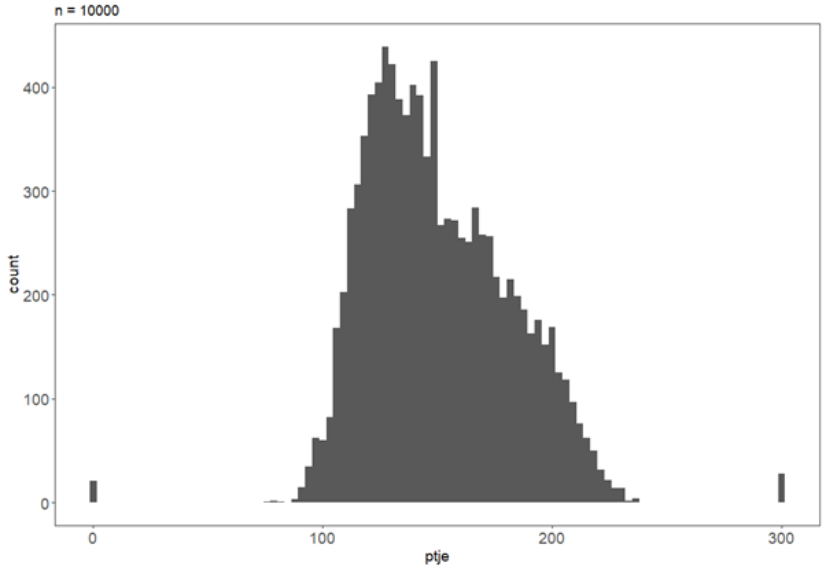
Figura 17. Muestra tamaño 1000 en consistencia.



La segunda muestra la cual se aumentó a un tamaño de 1000 ya presenta mejor consistencia y algo de similitud en la distribución de probabilidad con la distribución normal, se va pareciendo al parámetro. Figura 17.
La última muestra presenta una buena consistencia, ya que su distribución de probabilidad es demasiado similar a la distribución normal, y su estimador también es muy

parecido al parámetro, concluyendo lo dicho, a medida que aumentamos el tamaño la muestra final se parecerá a la población. Figura 18.

Figura 18. Muestra tamaño 10000 en consistencia.



Conclusiones

Al finalizar la práctica experimental relacionada con los datos estadísticos ofrecidos de las pruebas SABER, es posible concluir:

- El estrato y los puntajes en la materia de inglés está relacionado de manera inversamente proporcional cómo es posible identificar en nuestro primer y segundo diagrama de cajas, pues la distribución de los puntajes va aumentando a dependencia de la razón de crecimiento adjudicada a el valor de la matrícula de los estudiantes, exceptuando los casos de los estudiantes que no realizan pago de matrícula

- La distribución utilizada ofrece una buena consistencia ya que mientras se va aumentando el tamaño de la muestra, se aprecia como esta se asemeja a la distribución normal original.

- Para los datos empleados se demostró que la medida más eficiente y con menor varianza fue la media, la cual se empleó a pesar de la robustez que tiene la mediana respecto a los puntos atípicos.

- En los diagramas de cajas de puntaje respecto a horas de trabajo, se denota que las personas que no presenten horas de trabajo tienen un desempeño bueno pero no hace gran diferencia con respecto a las personas que trabajan por diversas horas, entonces todos presentan un desempeño aceptable y algunos valores tiene dispersiones altas aunque en general mantienen una mediana relativamente similar

- Se denota cómo comprobar si un muestreo es correcto, ya que, realizando varias muestras aleatorias, podemos destacar que se encuentran características gráficas muy similares a la de la población inicial.

Por tanto, el informe en general demuestra qué hay cierta variación en los resultados de los estudiantes en las pruebas de inglés en SABER, en dependencia de los estratos de los estudiantes y de las horas de trabajo. Esta variación, en un caso demuestra correspondencia proporcional y en el otro no, en su mención respectiva.