

# Módulo de Datos - Proyectos

## Estudio de factores sociales que influyen en las pruebas ICFES usando analítica de datos

Oscar Altuve (USB), Victor Guzmán (UCV)

Asesor: Prof. Fabio Martinez (UIS)

Junio de 2022



Latin American alliance for  
Capacity building in Advanced physics  
**LA-CoNGA physics**

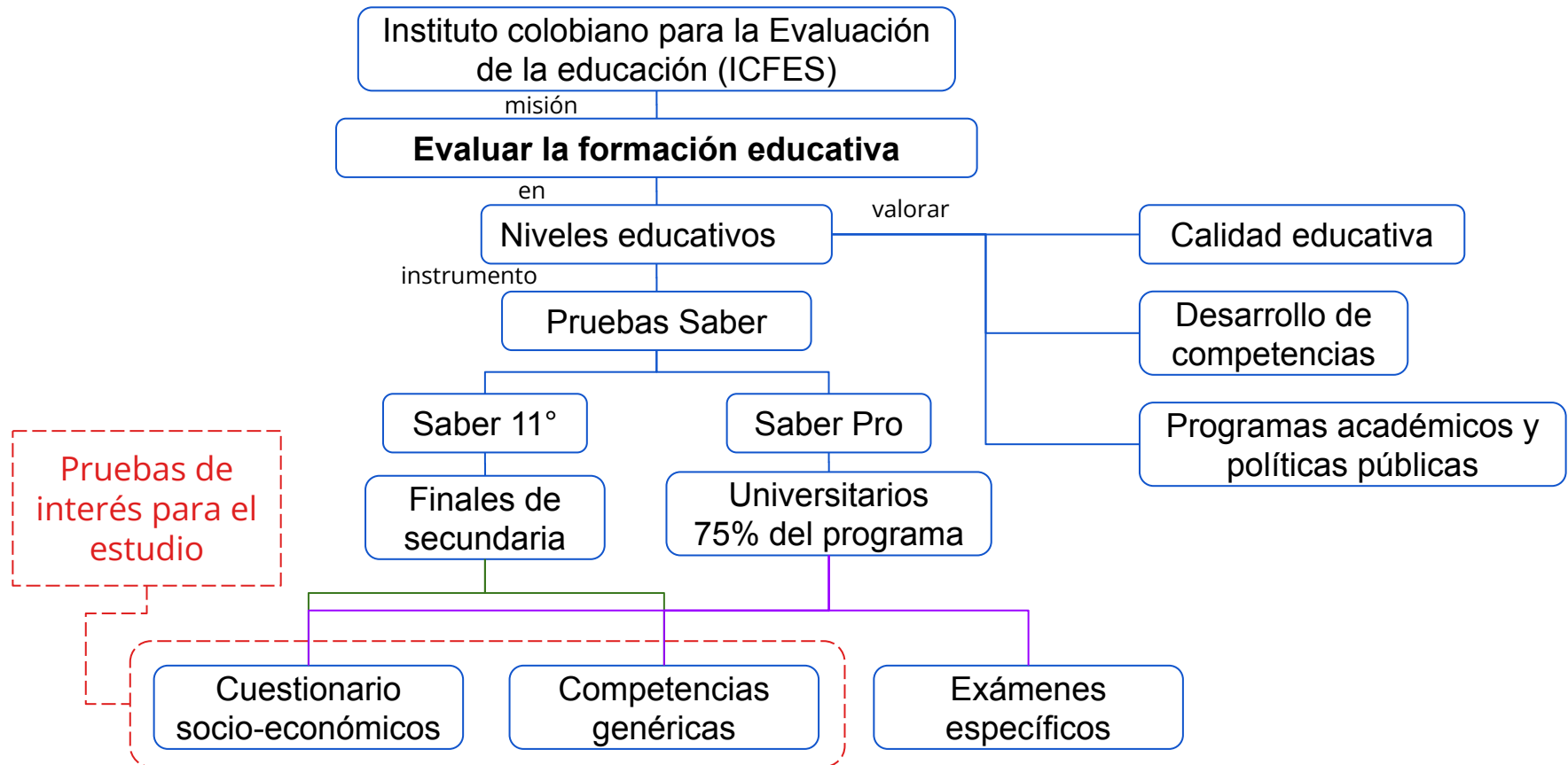


Cofinanciado por el  
programa Erasmus+  
de la Unión Europea





# Introducción





# Resultado y análisis: Variables de interés

- Se emplean resultados públicos de las pruebas Saber 11° (2018-2019) y Saber Pro (2019).
- Evaluación de competencias genéricas y el cuestionario socio-económico.
- Abanico de posibilidades de estudio.
- Análisis en Jupyter Notebook con kernel Python 3.7.

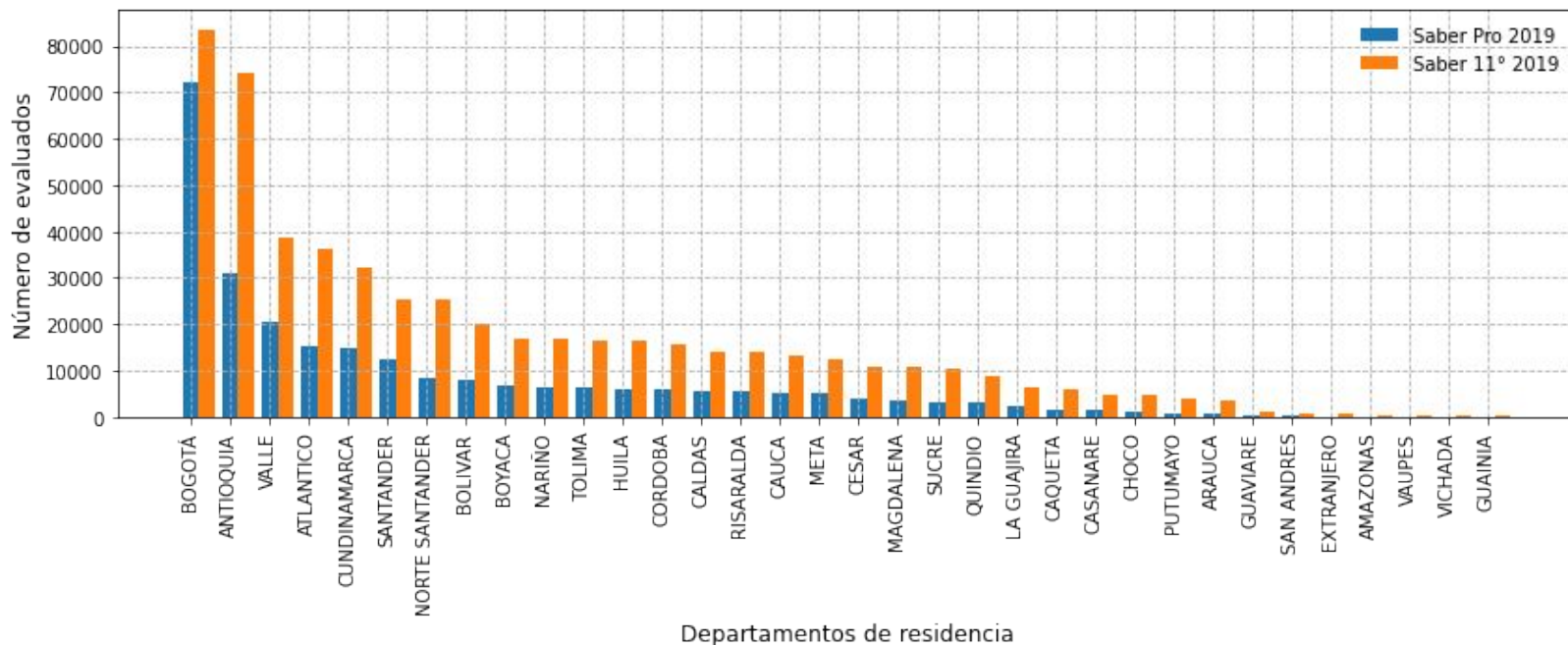
**Tabla 1:** Diccionario de variables empleadas en el estudio, de acuerdo a la prueba Saber 11° 2018-2019 y Saber Pro 2019. [7]

Campo	Descripción del campo	Opciones de respuesta
Información de contacto		
ESTU_DEPTO_RESIDE	Departamento de residencia	Texto
ESTU_COD_RESIDE_DEPTO	Código Dane del departamento de residencia	Númerica [99999 extranjero]
Información socioeconómica		
FAMI_TIENEINTERNET	¿Su hogar cuenta con servicio o conexión a internet?	Si No
Resultados		
PUNT_GLOBAL	Puntaje total obtenido	Númerica - Rango [0, 300]



# Resultado y análisis: evaluados / departamentos

- Cantidad de evaluados de acuerdo al departamento de residencia.
- Influencia educativa.



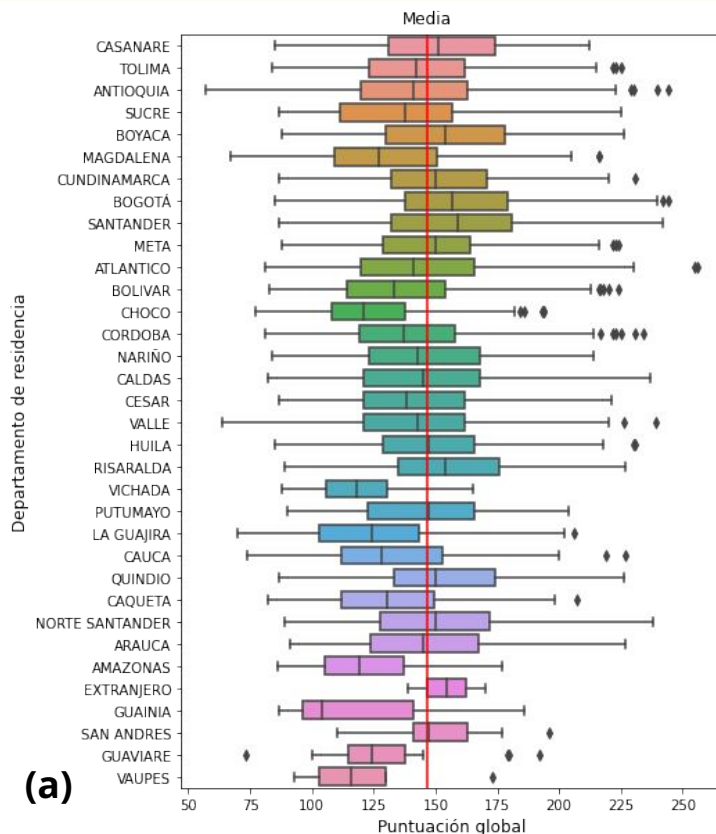
**Figura 1:** Muestra de estudiantes que presentaron las pruebas Saber 11° y Saber Pro en el año 2019 en función del departamento de residencia.



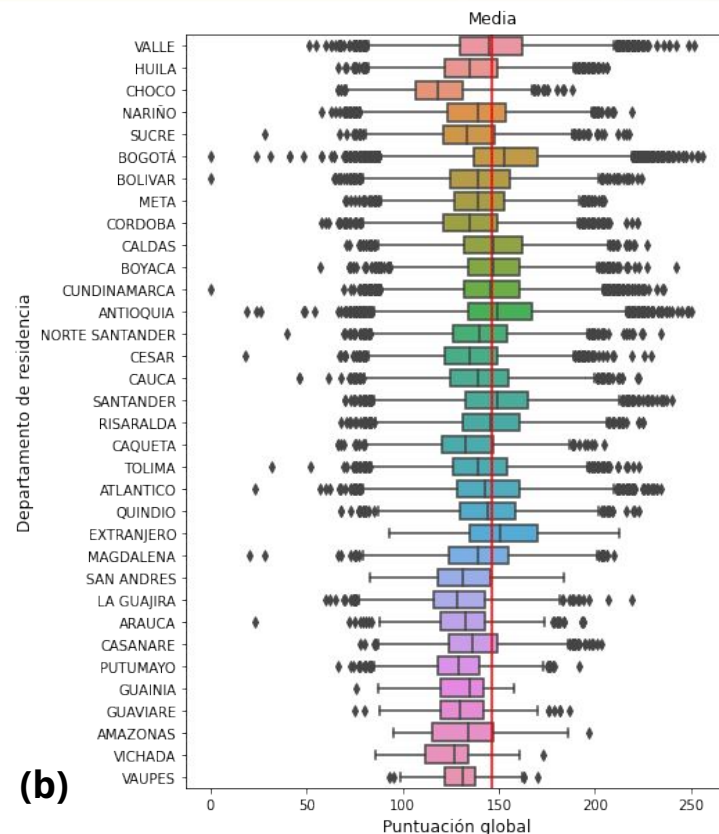
# Resultado y análisis: departamentos / punt. global

- En general las medianas de las muestras se encuentran en el centro de las cajas.
- Prueba Saber Pro, menor dispersión de los datos que la prueba Saber 11°.
- Valores atípicos se representan por los puntos negros.

Las muestras consideradas son significativamente distintas en cantidad.



(a)



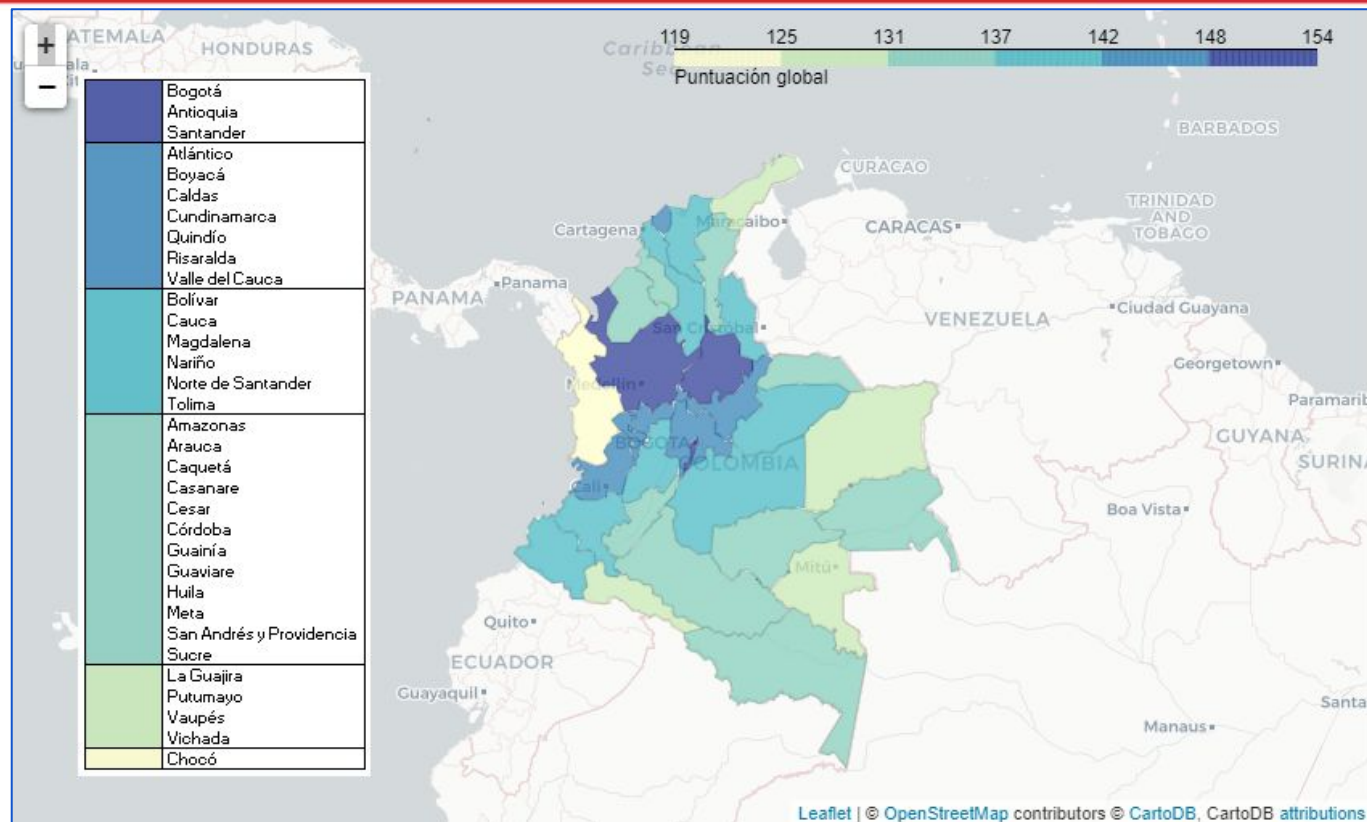
(b)

**Figura 2:** Departamento de residencia en función de la puntuación global de las pruebas Saber en el año 2019. (a) Saber 11° y (b) Saber Pro.



# Resultado y análisis: departamentos / punt. global

- Dada la generalidad de la mediana de la puntuación global por departamento (figura 2) de la prueba Saber Pro.
- Representación de la media de las puntuaciones globales en cada departamento.



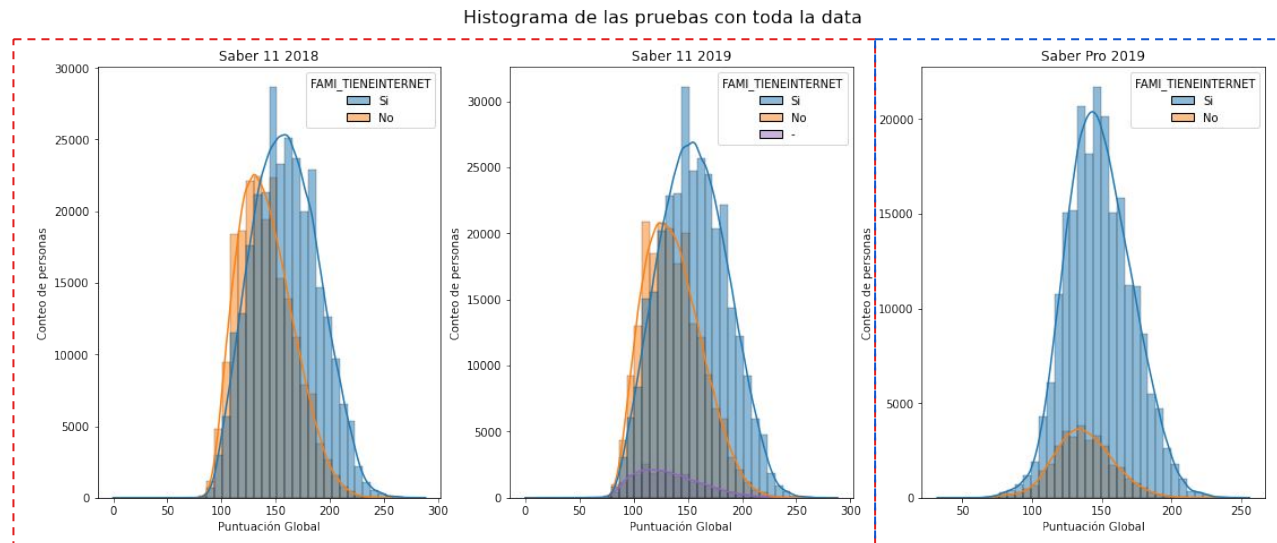
**Figura 3:** Promedio de la puntuación global de los evaluados de acuerdo al departamento de residencia.



# Puntaje total y el internet para toda la muestra

- Relación puntaje total e internet.
- La falta de acceso a internet influye en el puntaje global.
- El puntaje de admisión a las carreras filtra en su mayoría a estudiantes sin acceso a internet.

Bachilleres



Universitarios  
(75% aprobado)

**Figura 4:** Histograma del puntaje global para toda la población, para las pruebas Saber 11 2018, Saber 11 2019 y Saber Pro 2019.

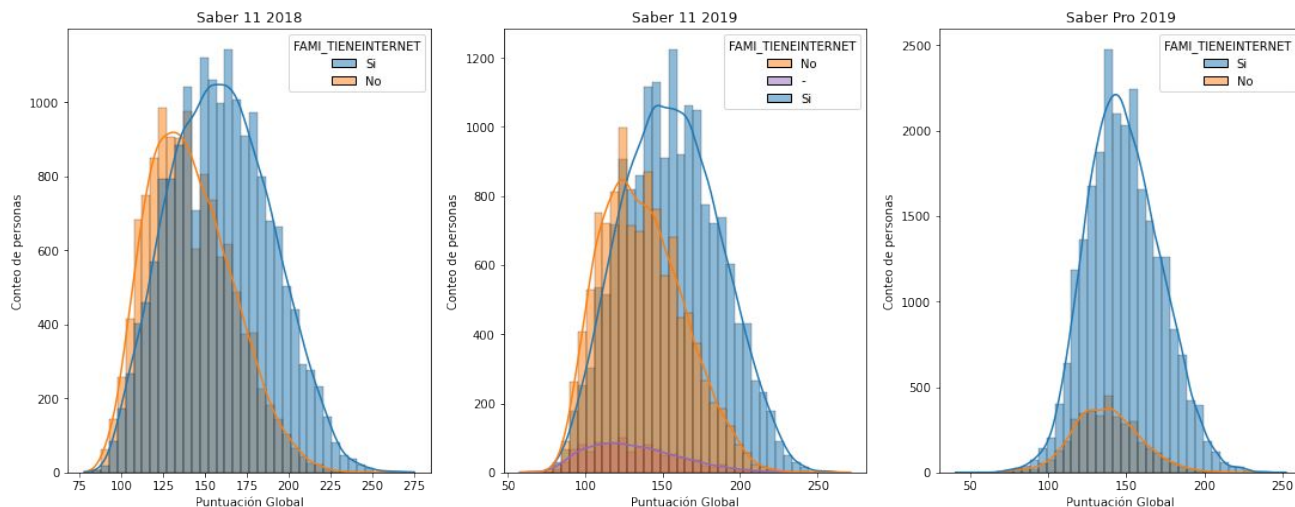


# Comportamiento de una muestra de 30000 datos

- ¿Se observa un comportamiento similar para una muestra de la población?
- ¿Sería válido usar la muestra para inferir sobre la población?

Haciendo un ajuste Gaussiano (asumiendo simetría), podemos caracterizar las distribuciones

Histograma de las pruebas con muestras de 30000 datos



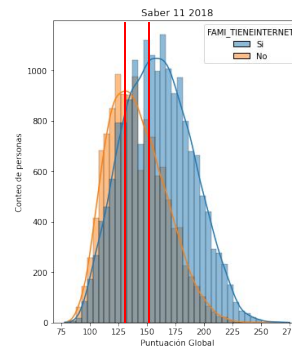
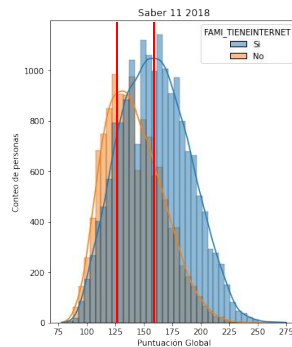
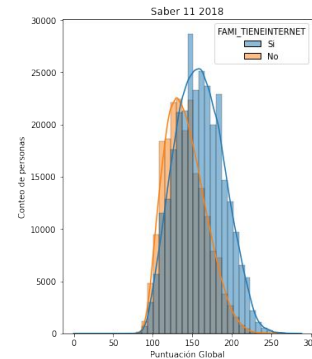
**Figura 5:** Histograma del puntaje global para la muestra de 30000 personas, para las pruebas Saber 11 2018, Saber 11 2019 y Saber Pro 2019.



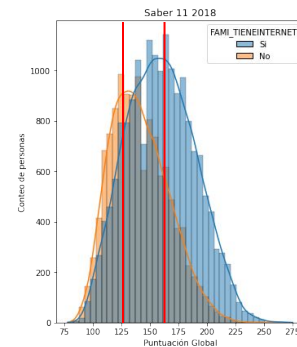
# La media como variable aleatoria

Procedimiento para establecer si la muestra escogida es *suficientemente grande*:

- Cálculo de la media y la desviación estándar para la población total.
- Selección arbitraria de una muestra de  $N$  elementos.
- Generación de  $n$  muestras
- Ajuste Gaussiano a cada muestra

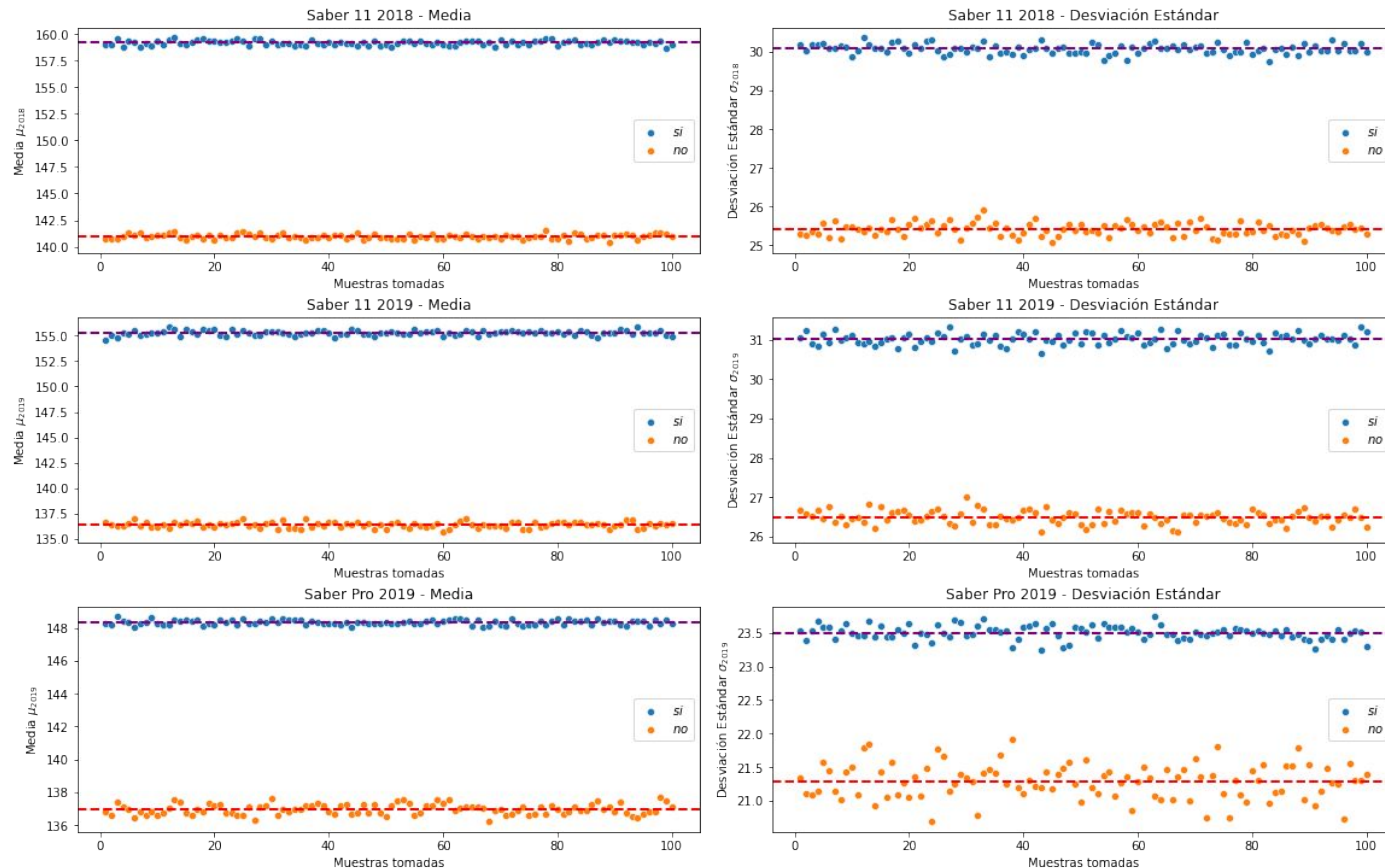


...





# Media y desviación estándar



**Figura 6:** Valor de las medias  $\mu_i$  y las desviaciones estándar  $\sigma_i$  para cada corrida, las líneas punteadas marcan la media aritmética para cada proceso estocástico.

- El error porcentual en la media está por debajo de 0.0065% para 100 muestras de 30000 datos.
- Selección de un treshold de aceptación: 0.05%.
- Se acepta una muestra de 30000 personas como significativa, para caracterizar la población.

**Tabla 2:** Error porcentual  $\epsilon$  de la media tomada de la muestra para cada prueba y dependiendo de si la persona tiene o no internet.

Prueba	Error si (%)	Error no (%)
Saber 11° 2018	0.0050	0.0038
Saber 11° 2019	0.0005	0.0046
Saber Pro 2019	0.0063	0.0441



# Diferencia de puntaje total

$$\Delta\mu_{2018} = \mu_{\text{si},2018} - \mu_{\text{no},2018} = 159,17 - 140,98 = 18,19$$

$$\Delta\mu_{2019} = \mu_{\text{si},2019} - \mu_{\text{no},2019} = 155,31 - 136,41 = 18,90$$

$$\Delta\mu_{\text{pro},2019} = \mu_{\text{si pro},2019} - \mu_{\text{no pro},2019} = 148,36 - 137,08 = 11,28$$

Pruebas a bachilleres

Estudiantes universitarios  
(75% aprobado)

La diferencia en la puntuación global se reduce para los estudiantes universitarios.



# Conclusiones

- **Distribución geográfica:** Los departamentos de residencia con mayor participación de las pruebas Saber 11° y Saber Pro, son coincidentes con los departamentos más competitivos de Colombia. La mayor cantidad de evaluados se observó en la capital del país.
- **Tamaño de la muestra:** Una muestra arbitraria de 30000 estudiantes puede ser usada para describir la población.
- **Acceso a internet y un mejor desempeño académico:** Los estudiantes con acceso a internet tienen un mejor rendimiento en las pruebas Saber 11 y Saber Pro.
- **Reducción de la brecha en el puntaje global:** La brecha que separa las puntuaciones de los estudiantes con y sin internet se ve reducida para los estudiantes universitarios con un 75% de la carrera aprobada.



# Estudios a futuro

- **Distribución geográfica:** Ajustes a la densidad de probabilidad PDF's sobre las puntuaciones globales para corroborar los resultados en los diagramas de caja.
- **Relación entre las pruebas Saber 11 y Saber Pro:** Vincular las pruebas Saber Pro con pruebas Saber 11 de 3-4 años pasados.
- **Asimetría de los histogramas:** Realizar un ajuste que permita caracterizar la asimetría de las distribuciones y estudiar el fenómeno (¿cambia la asimetría de una prueba a otra?, ¿a qué se debe?)
- **Puntaje global y estrato socioeconómico:** Se observó aunque no fue incorporado en el presente trabajo, que los estudiantes en un estrato socioeconómico superior tienden a sacar notas más altas. Puede verificarse para cada estrato, la sensibilidad del puntaje global ante otras variables.



# Referencias

- [1] ICFES, “Guía de orientación saber 11°,” 2022. [Online]. Available: <https://eservicioseducativos.com/wp-content/uploads/2022/02/Guia-de-orientacion-Saber-11-2022-1.pdf>
- [2] —, “Guía de orientación saber pro: Módulos de competencias genéricas,” jul 2019. [Online]. Available: <https://www.studocu.com/co/document/universidad-de-san-buenaventura/introduccion-a-las-relaciones-internacionales/guia-de-orientacion-modulos-de-competencias-genericas-saber-pro-2019/14138800>
- [3] J. Mantilla, H. Gonzalez, and L. Jaimes, “Estudio de factores sociales determinantes para los resultados de las pruebas icfes saber 11 2018,” 2021. [Online]. Available: <https://github.com/juandmantilla/Estudio-de-Factores-Sociales-Icfes>
- [4] ICFES, “Saber 11° 2019-2,” dic 2019. [Online]. Available: <https://www.datos.gov.co/Educaci%C3%B3n/Saber-11-2019-2/ynam-yc42>
- [5] —, “Resultados saber pro competencias genericas 2019-2,” feb 2020. [Online]. Available: <https://www.datos.gov.co/Educaci-n/Resultados-Saber-Pro-Competencias-Genericas-2019-2/2x55-9wxm>
- [6] O. A. Altuve and V. Guzmán, “Estudio de factores sociales que influyen en las pruebas icfes usando analítica de datos,” 2022. [Online]. Available: <https://gitmilab.redclara.net/altuveo/estudio-de-factores-sociales-que-influyen-en-las-pruebas-icfes-usando-analitica-de-datos>



# Referencias

- [7] ICFES, “Diccionario de variables saber-pro periodo 2012-2019.” [Online]. Available: <https://fdocuments.es/document/diccionario-de-variables-saber-pro-periodo-2012-2019-2020-05-27-1-diccionario.html>
- [8] Portafolio, “Los departamentos más competitivos del país,” Nov 2019. [Online]. Available: <https://www.portafolio.co/economia/gobierno/los-departamentos-mas-competitivos-del-pais-535920>
- [9] B. Øksendal, Stochastic differential equations: an introduction with applications, 6th ed., ser. Universitext. Springer, 2010.