

Ciencia de datos aplicada al estudio de la Obesidad y otras enfermedades crónicas en Córdoba

Práctico N°1 Análisis y visualización

Grupo 2

- Fernando Apaza
- Edgardo Garrigo
- Nicolas E Ponce

Objetivo y alcance: En esta primera aproximación se pretende comenzar por analizar el conjunto de datos, con el objetivo general de poder conocer las principales características de nuestra población, y su distribución entre subgrupos. De esta forma se pueden presentar los resultados obtenidos, poder plantear posibles relaciones entre variables a través de visualizaciones adecuadas.

Consignas

Parte I

1. Elegir cinco variables del dataset:
2. Analizar los estadísticos clásicos según la naturaleza y distribución de la variable (media, moda, mediana, desviación estándar, etc).
3. Realizar descripción de ellas de manera univariada y en relación a otras a través de la construcción de tablas y gráficos acordes a la naturaleza de cada una.
4. Calcular la FDP de dichas variables. Calcular correlaciones entre variables, entre otras posibilidades a elección.
5. Analizar outliers.

Parte II

1. Describir al menos un insight acompañado por su gráfico.
2. Plantee otros interrogantes si los desea.

Parte I

Las 5 variables elegidas son:

1. sexo : 1.0- masculino; 2.0- femenino
2. IMC: va continua
3. eent: educación del entrevistado 1- primario incompleto, 2- primario completo, 3- secundario incompleto, 4- secundario completo, 5- terciario, 6- universitario, 7- postgrado, 8- terciario incompleto, 9- universitario incompleto.
4. actfis: 1- no; 2- si
5. edad: va continua

Análisis de los estadísticos de las variables

	IMC	eent	edad
Media	25,82	3,900979	43
Mediana	25,10	4	41
Moda	22,77	4	23
Desv Estandar	47,78	1,490349	18
Max	51,11	9	97
Min	14,76	1	18

	sexo	actfis
Mediana	2.0	1.0
Moda	2.0	1.0
Max	2.0	2.0
Min	1.0	1.0

Conclusiones

IMC: el promedio de la muestra poblacional es 25.8 kg/m², que corresponde a IMC de preobesidad (24,9 - 29,9 kg/m²). EL valor más frecuente, moda, es 22.7 kg/m², que corresponde a IMC normal (18,4-24,9 kg/m²).

edad: La edad promedio es de 42,6 años, mientras que la edad más frecuente de la población es 23 años. La dispersión de los datos es bastante grande, pero esto es acorde al amplio rango de edades de la población (18-97 años).

eent: el secundario completo es la educación alcanzada lo más frecuente, valor 4.

sexo: En nuestro dataset, lo más frecuente es el valor 2.0, que indica que las personas son de sexo femenino.

actfis: No realizar actividad física es lo más frecuente, el valor 1.0, en nuestra población de estudio.

Análisis Univariado

Sexo

Tabla 3 - sexo

	index	sexo	número	frecuencia
0	2	femenino	2475	0,5767
1	1	masculino	1817	0,4233



Figura 1: número (A) y frecuencia (B) del sexo, masculino o femenino, de las personas que conforman la base de datos

Tal como se puede apreciar en la **Tabla 1** y en ambos gráficos de la **Figura 1**, femenino es el sexo más frecuente en la base de datos. Con respecto a los outliers, sexo es una variable categórica binaria y no presenta una dispersión significativa (datos no mostrados), por lo tanto no presenta valores anómalos.

actfis: Actividad física

Tabla 4 - actfis

	index	af	número	frecuencia
0	1	No hace actfis	2241	0.5222
1	2	Si hace actfis	2050	0.4777

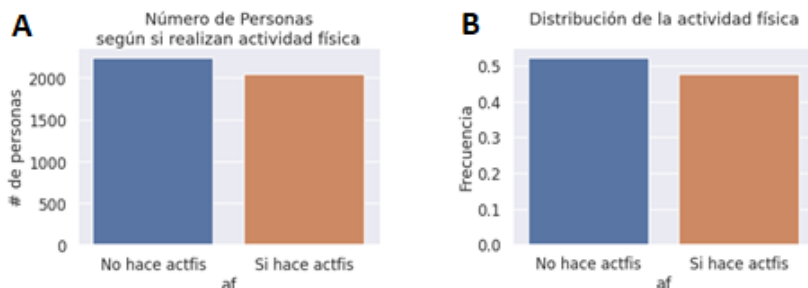


Figura 2: número (A) y frecuencia (B) de la actividad física, hace o no hace, de las personas que conforman la base de datos

En la **Tabla 4** y en ambos gráficos de la **Figura 2** se observa que la mayoría de las personas declararon no hacer actividad física. Al igual que la variable sexo, actfis es una variable categórica binaria y no presenta dispersión de datos que generen outliers (datos no mostrados).

eent: Nivel de educación

Tabla 5 - eent

index	eent	educ	número	frecuencia
0	1	primario incompleto	275	0.064073
1	2	primario completo	615	0.143290
2	3	secundario incompleto	439	0.102283
3	4	secundario completo	1787	0.416356
4	5	terciario	371	0.086440
5	6	universitario	773	0.180103
6	7	postgrado	4	0.000932
7	8	Terciario incompleto	0	0
8	9	universitario incompleto	28	0.006524

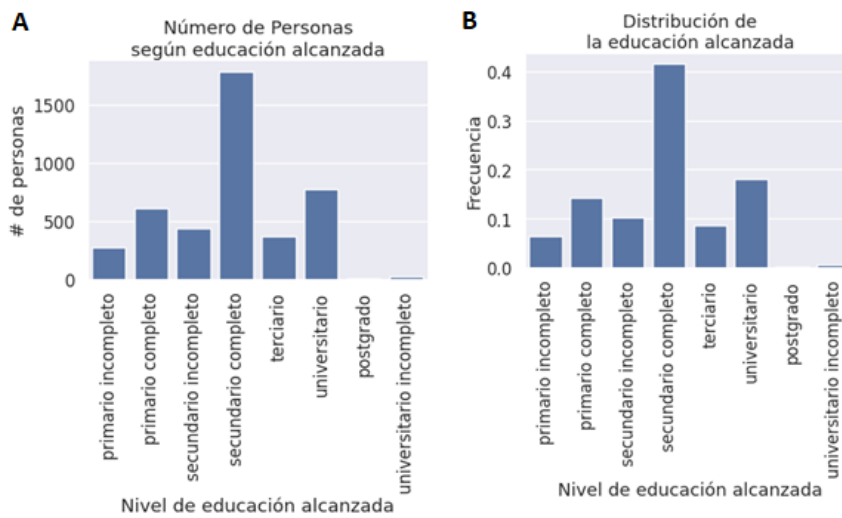


Figura 3: número (A) y frecuencia (B) de las personas según la educación alcanzada: primario incompleto, primario completo, secundario incompleto, secundario completo, terciario, universitario, postgrado, terciario incompleto o universitario incompleto.

En la **Tabla 5** y los gráficos de la **Figura 3** se observa que la mayoría de las personas encuestadas provienen de un secundario completo. En segundo y tercer lugar tenemos personas universitarias y con primario completo, respectivamente.

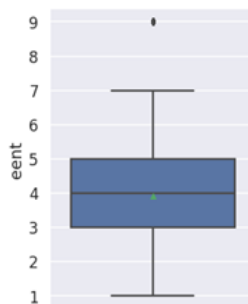


Figura 4: Distribución de la variable eent en un gráfico de cajas. La línea negra en la caja representa la mediana y el triángulo verde la media.

Con respecto a la presencia de outliers, el gráfico de caja de la **figura 4** demuestra que las personas que presentan estudios universitarios incompletos, valor de 9, representan valores anormales respecto a los estudios de las demás personas de la base de datos. De la exploración de datos, surge que ninguna de las personas encuestadas declaró tener un terciario incompleto.

-edad (años)

Tabla 6 - edad

index	i_edad	número	frecuencia
0	18-27	1244	0.289842
1	48-57	700	0.163094
2	28-37	686	0.159832
3	38-47	638	0.148649
4	58-67	521	0.121389
5	68-77	382	0.089003
6	78-87	115	0.026794
7	88-97	6	0.001398

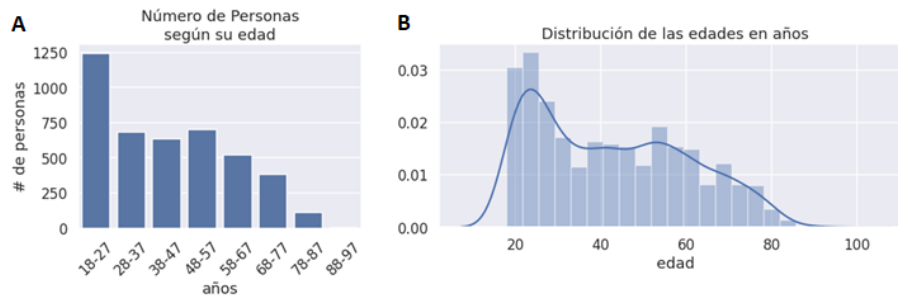


Figura 5: número (A) y distribución (B) de las edades de las personas.

Con el fin de para obtener una mejor visualización, y teniendo en cuenta que los datos mínimo y máximo son 18 y 97 respectivamente, decidimos crear intervalos de 10 años y graficarlos en el eje de la x (i_edad) en la **figura 5A**. Entre 18 y 27 años se encuentra la mayor frecuencia de las personas entrevistadas.

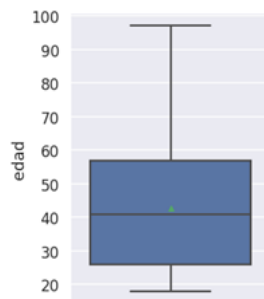


Figura 6: Distribución de la variable edad en un gráfico de cajas. La línea negra en la caja representa la mediana y el triángulo verde la media.

En la **Figura 6** se puede observar que para esta variable continua no hay ningún valor anómalo. Esto se debe a que la desviación estándar de la variable edad tiene un valor de 18 (**tabla 1**) y por lo tanto ningún valor está por fuera de los límites (bigotes).

- IMC: Índice de masa corporal (kg/m²)

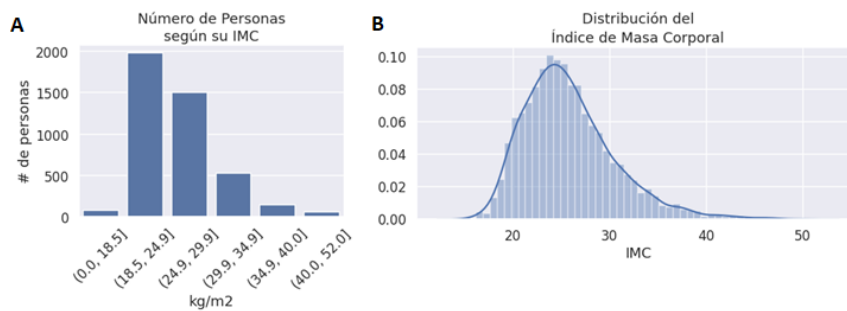


Figura 7: número (A) y distribución (B) del índice de masa corporal (kg/m²) de las personas.

Con el fin de para obtener una mejor visualización, y teniendo en cuenta la clasificación de IMC, decidimos crear intervalos y graficarlos en el eje de la x en la **figura 6A**. El grupo 25 a 29.9 kg /m², preobesidad, es el más frecuente en la población de la base de datos.

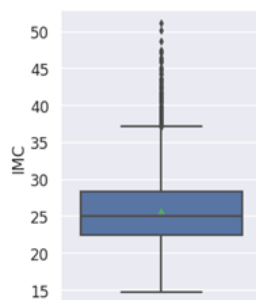


Figura 8: Distribución de la variable IMC en un gráfico de cajas. La línea negra en la caja representa la mediana y el triángulo verde la media.

Con respecto a la presencia de outliers, el gráfico de caja de la **figura 8** demuestra que las personas que un IMC mayor a 2,5 veces la desviación estándar (37.7) representan valores anómalos respecto al IMC del resto de la población.

Correlación entre las variables.

Se elige correlacionar el Índice de masa corporal (IMC) (VA cuantitativa de razón continua) con el nivel educativo del entrevistado (eent) (VA categórica ordinal). De acuerdo a las categorías y características de las variables, el test que nos permite ver si hay correlación es el *test de Spearman*, el cual arrojó los siguientes valores:

Spearman roS = -0.21339217318790546 , p-value = 2.1799063810378494e-45

El roS nos está indicando que existe una baja correlación negativa entre ambas variables. El valor p indica la probabilidad que la no correlación sea interpretada erróneamente, o sea que si exista correlación.

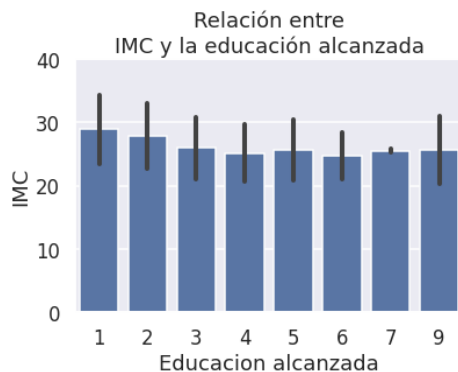
Por lo tanto se rechaza la hipótesis nula. Como el valor de p es tan bajo podríamos decir que el resultado es estadísticamente significativo

Por lo tanto concluimos que existe una baja correlación negativa entre las variables IMC y eent.

Parte II

Las primeras preguntas que nos planteamos son

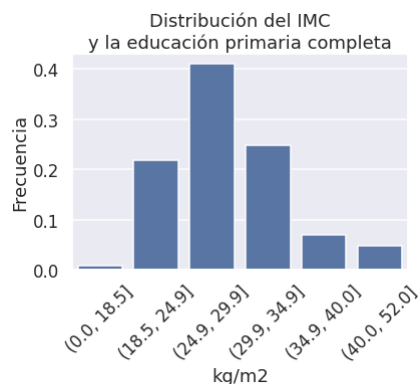
¿Qué relación existe entre un persona con un IMC Normal y sus estudios alcanzados? ¿y si una persona solo a adquirido el nivel más bajo de educación? ¿y si esta pregunta la filtramos por sexo femenino o masculino?



Este gráfico muestra que aquellas personas que tienen educación primaria completa (eent=1.0), presentan una media de IMC mayor que los otros grupos. Esto se correlaciona con los estudios de correlación efectuados anteriormente.

Teniendo en cuenta que son la población más afectada, decidimos estudiar puntualmente a la población de personas con eent=1.0, es decir con educación primaria completa.

index	i_imc	frecuencia
0	[24.9, 29.9]	0.410909
1	[29.9, 34.9]	0.247273
2	[18.5, 24.9]	0.218182
3	[34.9, 40.0]	0.069091
4	[40.0, 52.0]	0.047273
5	(0.0, 18.5]	0.007273



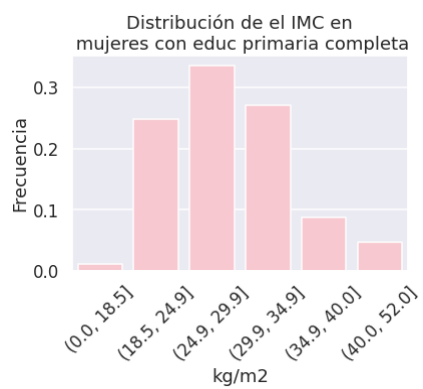
Este gráfico muestra que, dentro de la población con eent=1.0, con mayor frecuencia las personas presentan un IMC de pre-obesidad.

Si calculamos la probabilidad condicional de tener un IMC Normal, dado que tenés estudios primarios, no encontramos que el valor es 21.8%. Por lo que, la probabilidad de tener un IMC alterado (sea considerado un IMC bajo, pre-obesidad, obeso tipo 1, tipo2 o tipo 3) es de 78.2%.

Luego, nos propusimos evaluar si esta probabilidad de tener un IMC normal dado que tenes educación primaria completa, cambia segun el sexo.

En el caso del sexo femenino, obtuvimos los siguientes datos:

	Femenino	
index	i imc	frecuencia
0	(24.9, 29.9]	0.335294
1	(29.9, 34.9]	0.270588
2	(18.5, 24.9]	0.247059
3	(34.9, 40.0]	0.088235
4	(40.0, 52.0]	0.047059
5	(0.0, 18.5]	0.011765

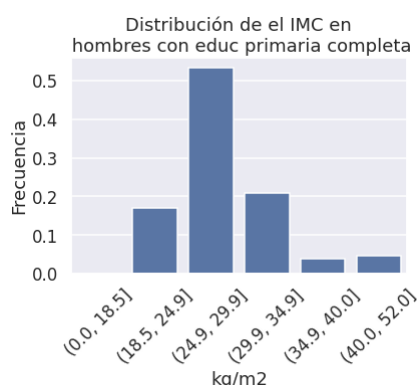


Si siguiendo las tendencias del gráfico anterior, las mujeres con $eent=1.0$ presentan con mayor frecuencia un IMC de pre-obesidad, obesidad tipo I y normal, respectivamente.

La probabilidad condicional de tener un IMC Normal, dado que tienes estudios primarios y sexo femenino, es de 24.7%. Por lo tanto, la probabilidad de tener un IMC alterado (sea IMC de bajo, pre-obesidad, obeso tipo 1, tipo 2 y tipo 3) es de 75.3%.

Continuamos con los estudios en el sexo masculino:

masculino		
index	i_imc	frecuencia
0	(24.9, 29.9]	0.533333
1	(29.9, 34.9]	0.209524
2	(18.5, 24.9]	0.171429
3	(34.9, 40.0]	0.047619
4	(40.0, 52.0]	0.038095
5	(0.0, 18.5]	0.000000



Al igual que el sexo femenino, los hombres presentan mayor frecuencia de IMC pre-obesidad, obesidad tipo I y normal, respectivamente. Pero el IMC de pre-obesidad es mucho más frecuente comparado al que presenta el sexo femenino.

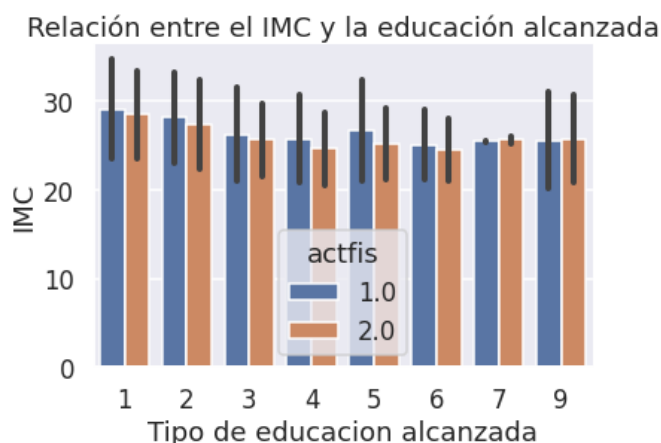
La probabilidad condicional de tener un IMC Normal, dado que tienes estudios primarios y sexo masculino, es de 17.1%. Por lo que la probabilidad de tener un IMC alterado (bajo, pre-obesidad, obeso tipo 1, tipo 2 y tipo 3) es de 82.9%.

Conclusión del insight planteado

Este análisis aporta la nueva información sobre la relación entre el nivel de educación alcanzado por una persona y su IMC. Las personas con estudio primario completo presentan una media del IMC más alta comparada al resto de los niveles educativos. Adicionalmente, la probabilidad de tener un IMC alterado aumenta en las personas con sexo masculino comparado al femenino, ambos con nivel de estudio primario completo.

Luego nos planteamos otras preguntas,

¿Qué relación existe entre un persona con un IMC y la actividad física? ¿y si tenemos en cuenta el nivel educativo en esta pregunta?



Uno puede presuponer que la actividad física podría tener impacto en el IMC.

De este gráfico, de los gráficos anteriores y de la correlación que hicimos previamente podemos decir que si bien se observa que a mayor nivel educativo existe un menor IMC. Esta correlación negativa entre ellas es muy pequeña.

Sin embargo, lo que nos llamó la atención en los datos es que no existe diferencia marcada en el IMC entre la población que hace y no hace actividad física, para cada nivel educacional.

Aunque en la mayoría de los grupos, se observa una disminución en la media del IMC en aquellas personas que realizan actividad física comparados a los que no lo hacen, su impacto pareciera no ser significativo si tenemos en cuenta la dispersión de los datos ($ci = "sd"$). Además, en el caso $eent = 7$, tienen la misma media de IMC.

Conclusión del insight planteado

Con los datos obtenidos no podemos comprobar que la variable hacer actividad física por sí sola no genera un cambio en el IMC de las personas, dependiente o independiente de su nivel de educación. Sin embargo, planteamos la hipótesis a futuro de que la actividad física en conjunto con otras variables, como alimentación, conducirán a un IMC normal.

Pero comprobar esa hipótesis necesitamos acceso a más variables y adentrarse más al dataset.