Trabajo Práctico N°1

Fundamentos de Análisis de Datos
Maestría en Ciencia de Datos

Flores, Matías
matflores@itba.edu.ar

Loiseau, Matías
mloiseau@itba.edu.ar

Septiembre 2023

Contents

1 Ejercicio N° 1

En el archivo *Dieta.xlsx* se encuentran los datos correspondientes a 173 personas que están siguiendo una dieta. Para cada una de ellas, se registró el sexo y el consumo de grasas saturadas y de alcohol, así como del total de calorías diarias.

1.1 Primer punto

Consigna: Analizar si existen datos faltantes y, en caso afirmativo, eliminar tales registros.

Antes de arrancar con la consigna cabe aclarar que previo al análisis cargamos las bibliotecas necesarias y el dataset en un data frame llamado *Dieta*.

Para analizar si hay datos nulos utilizamos dos funciones en específico para buscar las posiciones en donde se encuentran dichos datos. Para verificar si los datos fueron borrados correctamente, contamos la cantidad de filas antes y después de ejecutar la función de borrado.

La cantidad de filas del data frame era de 173 y la función devolvió que las posiciones donde los datos eran nulos fueron: 172, 195, 198 y 336 (donde el máximo valor era $173x4$). Luego de aplicar la función de borrado quedaron 169 filas, quedando así un data frame limpio para ser analizado.

1.2 Segundo punto

Consigna: Calcular las siguientes medidas estadísticas descriptivas clásicas del consumo de grasas: rango, media, mediana, desvío estándar y rango intercuartil.

En el siguiente gráfico se puede apreciar las medidas estadísticas descriptivas de todas las variables cuantitativas del dataset Dieta. Creemos pertinente agregar las variables Alcohol y Calorías a la consigna ya que aporta una visión más general del dataset.

Variable	Rango	Media	Mediana	Desvío Estándar	Rango Intercuartil
Grasas	11.82 - 46.36	24.59	24.09	6.40	7.95
Alcohol	0.00 - 40.11	8.66	5.84	8.94	11.19
Calorías	900 - 2376	1585.44	1585	291.43	355

Table 1: Medidas descriptivas de las variables del dataset Dieta.

1.3 Tercer punto

Consigna: Realizar gráficos boxplots de los datos sobre el consumo de calorías en función de la variable categórica. ¿Qué puede observarse?

En la figura XX observamos que ambas cajas son prácticamente parecidas pero el consumo de calorías de los hombres es levemente más dispersa que el de las mujeres. La segunda diferencia notoria es que la mediana en los hombres está centrada a la misma distancia entre los cuartiles Q1 y Q3, mientras que en el caso de las mujeres la mediana se encuentra más cerca al tercer cuartil. Por último se observa que ambos set de datos contienen dos valores atípicos.

AGREGAR BOXPLOT

1.4 Cuarto punto

Consigna: Dividir la cantidad de calorías consumidas en dos categorías: MODERADA (menor o igual a 1700) o ALTA (mayor a 1700). Analizar el consumo de alcohol de acuerdo a la cantidad de calorías consumidas según las categorías definidas.

Creamos dos data frames según la cantidad de calorías consumidas para poder realizar un mejor análisis de cada una por separado. Lo primero que podemos observar que para la categoría moderada quedaron 117 registros mientras que para la categoría alta fueron 52, siendo así casi la mitad de la otra.

Categoría	Filas
Alta	52
Moderada	117

Table 2: Cantidad de filas para ambas categorías.

Los primeros gráficos realizados fueron histogramas para cada una de las categorías. Para visualizar correctamente la distribución de la variable realizamos el cálculo de 3 bins. Los métodos elegidos fueron Sturges, Scott y Freedman-Diaconis. Para ambos casos escogimos el método de Sturges.

GRAFICO 1 Y GRAFICO 2 HISTOGRAMAS

Para ambos casos, a simple vista no se puede observar un supuesto de normalidad conciso. Además se ve que en la categoría moderada el consumo de alcohol no llega a 12, por otro lado para el caso de alta llega a valores cercanos a 40.

Para el caso de moderada, se puede apreciar que hay una mayor distribución entre el 0 y el 12 estando más concentrado en valores cercanos a 0, en cambio en la otra categoría la mayor concentración de datos está a partir del valor 10.

GRAFICO BOXPLOT

En la figura anterior se aprecia que el consumo de alcohol para la categoría alta tiene una mayor cantidad de dispersión que la moderada, además podemos confirmar que las personas con mayor consumo de calorías son también aquellas que consumen más alcohol.

2 Ejercicio N° 2

El archivo *Sociodemograficos.xlsx* contiene datos sobre distintos indicadores socio-demográficos de varios países.

2.1 Primer punto

Consigna: ¿Cuáles son las variables de interés? ¿Cuántos países fueron analizados?

Las variables de interés son: Tasa de natalidad, Tasa de mortalidad, Tasa de mortalidad infantil, Tasa de personas mayores de 65 años, Expectativa de vida al nacer, Expectativa de vida al nacer en varones, Expectativa de vida al nacer en mujeres y Población urbana. Fueron analizados 26 países.

2.2 Segundo punto

Consigna: ¿Cuáles son los países con menor y mayor tasa de natalidad?

El país con menor tasa de natalidad es Austria con 9. En cambio, el país con mayor tasa de natalidad es Afganistán con 47.

2.3 Tercer punto

Consigna: Realizar un diagrama de dispersión con las tasas de natalidad y de mortalidad infantil. ¿Qué puede observarse? Justificar lo observado a partir del gráfico con una medida cuantitativa.

GRAFICO XX

En el diagrama de dispersión podemos observar que la mayor concentración de países está en la parte inferior izquierda de la figura. Además, se visualiza un patrón lineal que se puede analizar con la correlación de Pearson. En este caso nos dio un valor de 0.9198, esto quiere decir que existe una correlación positiva fuerte entre la tasa de natalidad y la tasa de mortalidad infantil.

2.4 Cuarto punto

Consigna: Calcular el vector de medias y medianas.

En la siguiente tabla se muestra las medias y medianas de todas las variables.

Variable	Media	Mediana
Tasa de Natalidad	18.73	14.50
Tasa de Mortalidad	8.84	8.00
Tasa de Mortalidad infantil	26.42	16.50
Tasa de personas mayores de 65 años	8.84	8.00
Expectativa de vida al nacer	70.81	73.00
Expectativa de vida al nacer en varones	68.31	71.00
Expectativa de vida al nacer en mujeres	73.35	76.00
Población urbana	45333423	8673000

Table 3: Media y Mediana de cada una de las variables del set de datos.

2.5 Quinto punto

Consigna: Calcular las matrices de covarianzas y de correlaciones. A partir de estas matrices dar un ejemplo de dos variables fuertemente correlacionadas positivamente, de dos variables fuertemente correlacionadas negativamente y de dos variables no correlacionadas.

FIGURA

A partir de la figura XX se observa que la expectativa de vida al nacer y la expectativa de vida al nacer en mujeres tiene una fuerte correlación positiva. En cambio, la expectativa de vida al nacer en mujeres y tasa de mortalidad infantil tiene una fuerte correlación negativa. Por último podemos ver que no existe una correlación alguna entre la tasa de personas mayores a 65 años y tasa de mortalidad.

3 Ejercicio N° 3

Vamos a considerar el conjunto de datos *swiss* disponible en R.

3.1 Primer punto

Consigna: Cargar la base de datos y explorarla. ¿Cuántos registros y cuántas variables tiene? Describir las variables de estudio.

A partir del análisis del dataste podemos determinar que tiene 47 registros correspondientes a las provincias de suiza y 6 variables de estudio. Las mismas son: fertilidad, agricultura, examinación, educación, catolicismo y mortalidad infantil. Todas las variables se encuentran en porcentajes.

3.2 Segundo punto

Consigna: Se desea comparar las provincias entre sí. ¿Es adecuado utilizar la distancia Euclídea para realizar la comparación? Justificar la respuesta.

Para poder determinar si utilizar la distancia euclidiana para hacer la comparación, realizamos una matriz de correlación entre las variables del conjunto de datos.

GRAFICO MATRIZ DE CORRELACION

En la figura anterior podemos observar que las variables tienen algún tipo de correlación lineal entre ellas. Si bien esta relación no es muy fuerte, consideramos que dichos valores no son despreciables. En resumen, determinamos que la distancia Euclídea no es la mejor manera de comparar las provincias entre sí.

3.3 Tercer punto

Consigna: Buscar la presencia de datos atípicos mediante la distancia de Mahalanobis. Comentar los resultados obtenidos.

Para buscar la presencia de datos atípicos lo primero que hicimos fue calcular la distancia de Mahalanobis. Luego establecemos los puntos de corte usando la distribución de chi cuadrado con 6 grados de libertad para distintos niveles de significación (0.90, 0.95 y 0.99).

Nivel de significación	Valor atípico
0.90	Geneve, La Vallee, Porrentruy, Sierre, Neuchatel, Rive Droite
0.95	Geneve, La Vallee, Porrentruy
0.99	Geneve

Table 4: Valores atípicos según la distancia de Mahalanobis con varios niveles de significancia.

4 Ejercicio N° 4

El Departamento de Psicología de una universidad ubicada en una ciudad céntrica realizó un estudio sobre la asistencia a clases teóricas no obligatorias dependiendo de la localidad de residencia del estudiantado. Para tal fin, se seleccionaron 40 estudiantes en la Ciudad A, 40 estudiantes en la Ciudad B y 40 estudiantes en la Ciudad C, y se contabilizó la cantidad de clases a las que cada uno/a asistió. Los resultados obtenidos se muestran en la siguiente tabla.

Ciudad A	Ciudad B	Ciudad C
11	13	6
14	10	7
7	12	3
15	7	5
11	5	9
13	10	6
11	10	1
16	16	6
10	9	0
15	7	2
18	7	5
12	2	6
9	6	11
9	9	6
10	9	7
10	8	0
15	8	5
10	10	7
14	3	5
10	6	4
10	5	7
12	2	4
14	9	2
12	3	8
15	4	9
7	5	6
13	10	1
6	8	4
10	5	7
15	9	7
20	10	8
10	8	9
13	13	7
10	10	5
6	0	1
14	2	6
8	1	9
10	1	4
8	0	7
11	4	16

Table 5: Tabla de asistencia a clases teóricas no obligatorias.

4.1 Primer punto

Consigna: Armar un *data frame* en R con los datos de la tabla anterior, creando dos variables: una que represente la cantidad de asistencias a las clases teóricas no obligatorias y otra que represente la localidad de residencia. ¿Qué tipo de variable es cada una?

Parte del data frame creado se puede visualizar en la siguiente tabla

Ciudad	Asistencia
A	11
A	14
B	7
C	5
..	..

Table 6: Variables sobre la asistencia a clases teóricas no obligatorias según la localidad.

La variable ciudad es de tipo categórica y la variable asistencia es de tipo numérica.

4.2 Segundo punto

Consigna: Analizar los datos de la muestra mediante gráficos y medidas estadísticas descriptivas. ¿Se observan diferencias en los valores promedios por localidad?

Ciudad	Media	Mediana	Rango	Desvío Estándar
A-B-C	8.0	8.0	$0 - 20$	4.22
A	11.6	11.0	$6 - 20$	3.16
B	6.9	7.5	$0 - 16$	3.84
C	5.7	6.0	$0 - 16$	3.13

Table 7: Medidas estadísticas descriptivas.

GRAFICO BOXPLOT

Observando la tabla 7 y la figura XX, podemos determinar que la ciudad A tiene un mayor promedio de asistencia a clases teóricas no obligatorias. Además, las ciudades B y C son relativamente parecidas pero la B tiene un mayor desvío estándar.

4.3 Tercer punto

Consigna: Realizar un test ANOVA para comparar las medias de las 3 poblaciones. Plantear las hipótesis nula y alternativa del test, informar los resultados obtenidos y la decisión tomada.

Para realizar el test de ANOVA primero vamos a realizar los test de normalidad y el de homocedasticidad. La hipótesis nula para el test de ANOVA es: $h_0 = \mu_A = \mu_B = \mu_C$ mientras que la alternativa es que existe al menos un par $i \neq j$ tal que $\mu_i \neq \mu_j$ significativamente.

Primero analizamos la normalidad de los datos de asistencia a partir de un gráfico de qqplot, ademas realizamos un test de Lilliefors para confirmar si los datos corresponden a una distribución normal. Utilizamos este test y no el de Shapíro-Willk porque estamos trabajando con más de 50 datos.

El qqplot nos dio bastante normal siendo que los puntos se encuentran en el rango con algunos fuera de este en el margen superior derecho. Pero realizando el test de Lilliefors nos dio un p-value menor al nivel de significancia por lo cual determinamos que el conjunto no sigue una distribución normal. A pesar de esto, decidimos continuar con el análisis ya que suponemos que los resultados obtenidos no se encuentran tan significativamente fuera del rango de normalidad.

Continuamos con el análisis de homocedasticidad de los datos, para esto utilizamos el test de Levene. En nuestro caso nos dio 0.15 por ende no rechazamos la hipótesis nula que supone la igualdad de las varianzas en los grupos.

Procedemos a realizar el test de ANOVA sobre la asistencia según la ciudad. Luego de obtener los resultados calculamos el tamaño del efecto para compararlos con los niveles de clasificación más comunes.

En la siguiente tabla podemor ver los indicadores del test de ANOVA.

Indicadores	Resultado
Grado de Libertad	2
Estadístico F	33.72
P-Value	$2.74e^{-12}$
Tamaño del Efecto	0.36

Table 8: Resultados de ANOVA.

El P-Value obtenido es más bajo que el corte usual que el 0.05. El tamaño del efecto nos dio mayor que el nivel de clasificación de 0.14 por lo cual deducimos que la varianza de la asistencia se encuentra explicada en gran medida por el tipo de ciudad en la que se encuentra. A partir de los resultados obtenidos podemos concluir que existen diferencias significativas entre las medias de las asistencias en las distintas ciudades.

4.4 Cuarto punto

Consigna: Si se han obtenido diferencias significativas entre las localidades, determinar cuáles son esas diferencias utilizando el test de Tukey.

Intersecciones	Diff	Lower	Upper	P-Value
B-A	-4.7	-6.5	-2.8	0
C-A	-5.9	-7.7	-4.1	0
C-B	-1.2	-3.0	-0.6	0.258

Table 9: Resultados de Tukey.

Podemos concluir que existen diferencias estadísticamente significativas en las medias de las asistencias entre las ciudades A-B y A-C, donde los valores del P-Value son menores a 0.05.