Fundamentos de Análisis de Datos

TP1

1) En el archivo "Datos trabajo 1.xls " encontrará los datos correspondientes a 173 personas que están siguiendo una dieta, en las que se ha registrado el sexo (Sexo), el consumo de grasas saturadas (Grasas sat), el consumo de alcohol (Alcohol) y el total de calorías (Calorías). El valor que indica dato faltante es 999, 99 para todas las variables.

Se importaron los datos a Rstudio.

```
library(readxl)
Datos_trabajo_1 <- read_excel("~/Desktop/Algebra lineal/Datos trabajo 1.xls")
```

a) Reemplace los datos faltantes con NA.

Se reemplazaron los valores faltantes 999,99 por NA utilizando el siguiente comando:

```
Datos_trabajo_1$Alcohol[Datos_trabajo_1$Alcohol==999.99]<- NA
Datos_trabajo_1$Grasas_sat[Datos_trabajo_1$Grasas_sat==999.99]<- NA
Datos_trabajo_1$Calorías[Datos_trabajo_1$Calorías==999.99]<- NA
```

b) Describa las principales características que presentan los datos. Realizar gráficos boxplots. En todos los casos debe comentar los resultados.

El conjunto representa a 173 personas que siguieron una dieta. Las variables de estudio corresponden a los registros de consumo de grasas saturadas, alcohol y total de calorías por sexo. Para poder avanzar con el análisis, los registros faltantes o NAs fueron reemplazados por la media de la columna respectiva. La Tabla 1 es un resumen de los principales indicadores estadísticos descriptivos, además a modo de comparación se incluyen los rangos de consumo diario recomendados por la Organización Mundial de la Salud.

Tabla 1

	N	Media	SD[1]	Var[2]	Valor mín	1er cuartil	Mediana	3er cuartil	Valor máx	Valores recomendados[3]
Grasas saturadas	173	24,77	6,51	0,26	11,82	20,20	24,16	28,04	46,36	44 - 76
Alcohol	173	8,83	9,11	1,03	0,00	1,84	5,97	12,95	40,11	20 - 60
Calorías	173	1585	305	0,19	800	1400	1585	1761	2376	1600 - 2000

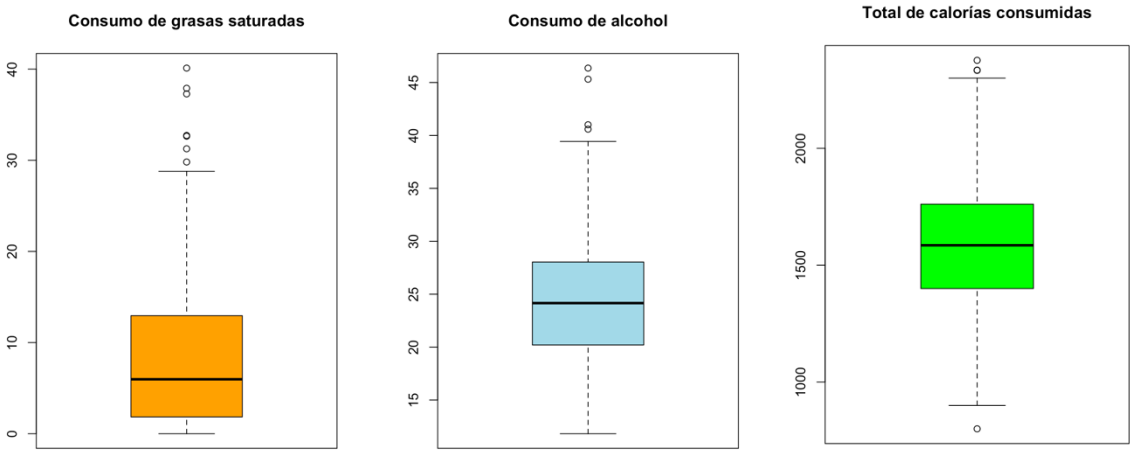
[1] Desviación estándar

[2] Varianza

[3] Consumo diario recomendado por la OMS

La tabla 1 realiza una breve descripción de los datos con los valores mínimos, máximo y cuartiles que luego van a ser representados en el gráfico boxplot. Para analizar la dispersión de los datos se utilizaron la desviación estándar y la varianza, esta última resulta más útil para la comparación de datos con diferentes escalas. Se encuentra una mayor dispersión respecto de la media en el caso del alcohol, mientras que la variabilidad de grasas saturadas y calorías son similares. Por último, permite reconocer que tanto la media como la mediana de las variables (Grasas saturadas, alcohol y calorías) están por debajo de los valores recomendados de consumo.

Figuras 1, 2 y 3

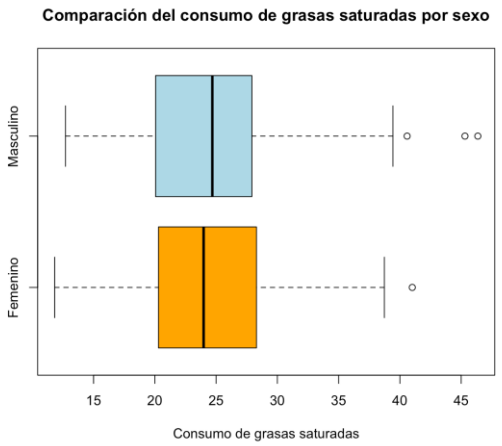


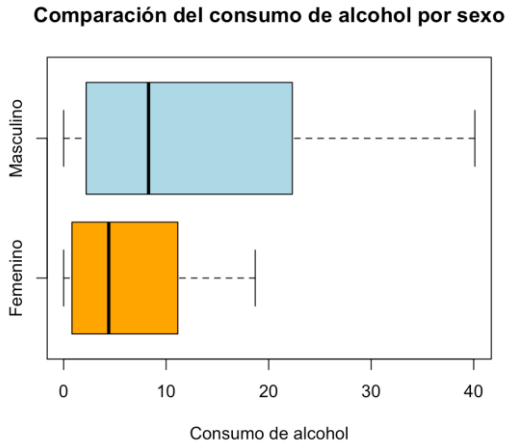
En la figura 1, el consumo de grasas saturadas cuenta con datos concentrados hacia los valores inferiores y con bigotes más largos hacia arriba, dando una asimetría positiva o sesgada a la derecha. En cambio, las variables de alcohol y calorías consumidas cuentan con los datos centrados y las medianas situadas en el centro de la caja reflejando una distribución simétrica. Si se realiza una comparación con la Tabla 1, para el caso de las calorías y el alcohol, las medias coinciden con las medianas.

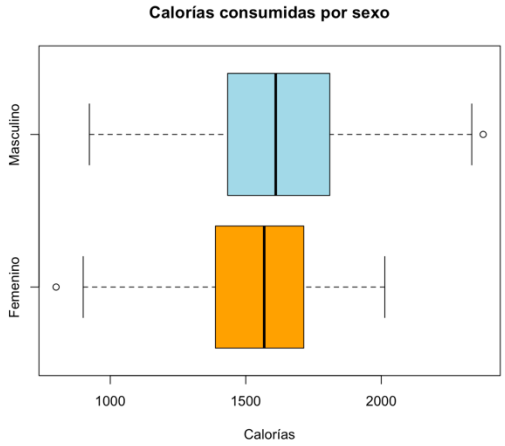
Si bien, a simple vista, en los gráficos boxplots se observan outliers, calculando la distancia intercuartil y multiplicándola por 1.5 o 3 es posible detectar aquellas variables con outliers moderados y severos. En este caso, todas las variables del conjunto presentan ambos tipos de outliers, lo que indica una gran dispersión de los datos.

c) Analice los datos para la variables Grasas sat, Alcohol y Calorías de acuerdo a la variable categorica Sexo. Comentar los resultados.

Figuras 4, 5 y 6





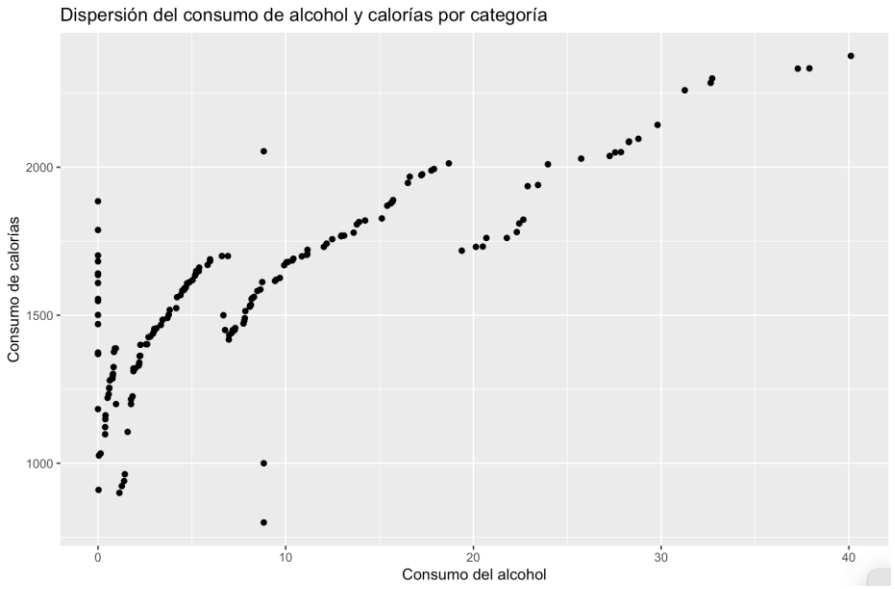


Al comparar las variables según el sexo se observan: medianas similares para el consumo de grasas saturadas; un consumo de alcohol notablemente más elevado en hombres que en mujeres y una diferencia leve en el consumo de calorías. Esta última variable presenta una mediana del conjunto de datos para hombres de 1610 kilocalorías y mujeres de 1568Kcal. Mientras que los valores recomendados oscilan en un rango de 1800 a 2100 para mujeres y de 2.000 a 2.400 para hombres. Indicando nuevamente que los valores del conjunto de datos están por debajo de lo recomendado. En cuando al consumo de alcohol la OMS recomienda como un consumo regular diario entre 20 y 40g en mujeres y 40 a 60 en hombres. Si bien este consumo tiene una media de 8,2, un valor mucho mayor a la media de consumo en mujeres (4,4), es un indicador de que esta población tiene un consumo sano y por debajo de lo recomendado.

d) Analice la variable Alcohol de acuerdo a la cantidad de calorías consumidas, tomando 3 categorías para la variable Calorías: CATE 1:1100 o menos calorías consumidas, CATE 2: más de 1100 hasta 1700 calorías consumidas, CATE 3: más de 1700 calorías consumidas.

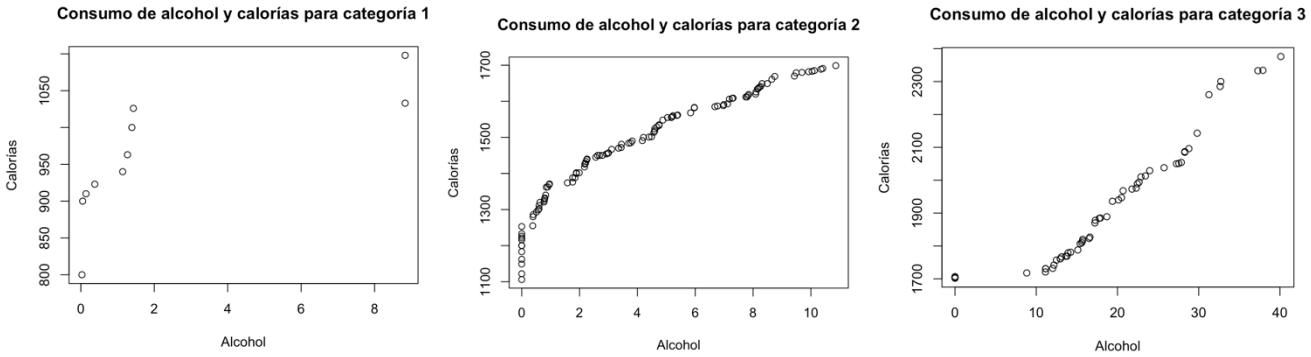
La variable alcohol con el consumo de calorías tiene una correlación de 0.8 indicando una alta correlación y una dependencia entre las variables.

Figura 7



Si bien un bajo o moderado consumo de alcohol grafica de forma evidente la relación entre las variables, un consumo mayor a 10 permite visualizar una distribución más alineada de los registros.

Figuras 8, 9 y 10



Las figuras 8, 9 y 10 representan la división de las calorías consumidas en tres categorías. Para la categoría 3, a partir de las 1700 calorías se grafica una relación con tendencia lineal acompañados de un consumo de alcohol por encima de 10.

2.En Suiza, los cantones constituyen el ente político y administrativo sobre el que se construye el Estado-nación. La llamada Confederación Helvética, de carácter fuertemente federal adoptó su condición actual en 1848, fecha hasta la cual cada uno de los cantones entonces existentes poseía sus propias fronteras, ejército y moneda. Suiza se encuentra en el cruce de algunas de las grandes culturas europeas, las cuales han influenciado fuertemente el idioma y la cultura del país. Suiza tiene tres idiomas oficiales (alemán, francés, italiano) y uno parcialmente oficial, el romanche.

El país ha estado históricamente dividido entre católicos y protestantes, con una compleja mezcla de territorios con mayorías católicas y protestantes por todo el país. Las ciudades más grandes (Berna, Zúrich y Basilea) son predominantemente protestantes. El centro del país, así como el Tesino, son tradicionalmente católicos. En 1980 se votó una iniciativa para separar completamente la iglesia y el Estado, pero fue rechazada, con solo el 21,1 % de la población a favor.
En la revisión de la constitución de 1874 la escuela primaria se hace obligatoria.
R contiene un dataset de nombre swiss que se puede cargar mediante el comando data(swiss). Los datos corresponden a seis variables medidas en los 47 cantones suizos en el año 1888.

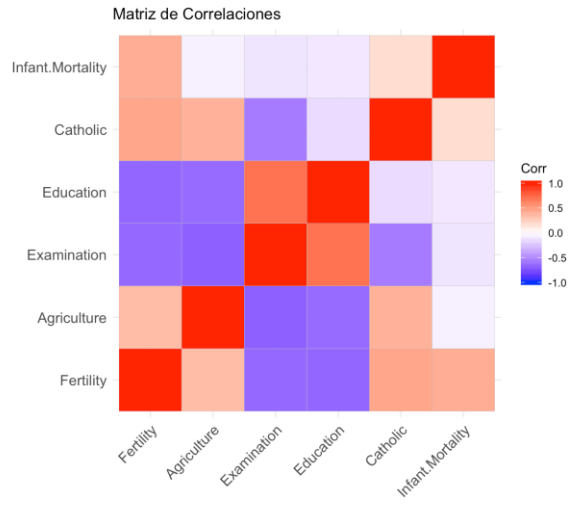
Las variables son las siguientes, todas ellas están en el intervalo [0, 100]:

a. Fertility: es una medida de la fertilidad del suelo del cantón (cuanto más cerca de 100, mayor fertilidad y cuanto más cerca de 0, menor fertilidad).
b. Agriculture: porcentaje de hombres trabajando en agricultura.
c. Examination: porcentaje de reclutas que reciben la calificación más alta en un examen del ejército.
d. Education: porcentaje de reclutas con estudios superiores a primaria.
e. Catholic: porcentaje de católicos.
f. Infant Mortality: porcentaje de nacidos que viven menos de 1 año de vida.

Se pide:

a)Decidir si las variables del conjunto de datos son independientes. Comentar los resultados obtenidos.

Figura 11

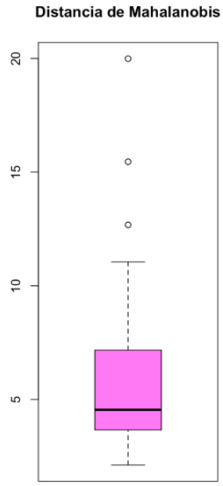


La figura 11 grafica todas las variables del conjunto de datos en una matriz de correlaciones. La educación y la variable examinación están altamente correlacionadas entre sí, con una correlación

lineal positiva cercana a 1. La correlación evidencia que los valores de ambas variables tienden a aumentar al mismo tiempo. En este caso, la examinación como el porcentaje de reclutas que reciben la calificación más alta en un examen del ejército y la educación porcentaje de reclutas con estudios superiores a primaria. Contrariamente la agricultura y la fertilidad presentan una correlación negativa con ambas variables, examinación y educación, es decir que habría una relación inversa. Ambas relaciones reflejan la realidad de la contraposición entre la educación y la agricultura donde la obligatoriedad de la educación primaria previene el trabajo infantil.

b) Buscar la presencia de datos atípicos mediante la distancia de Mahalanobis. Comentar los resultados obtenidos.

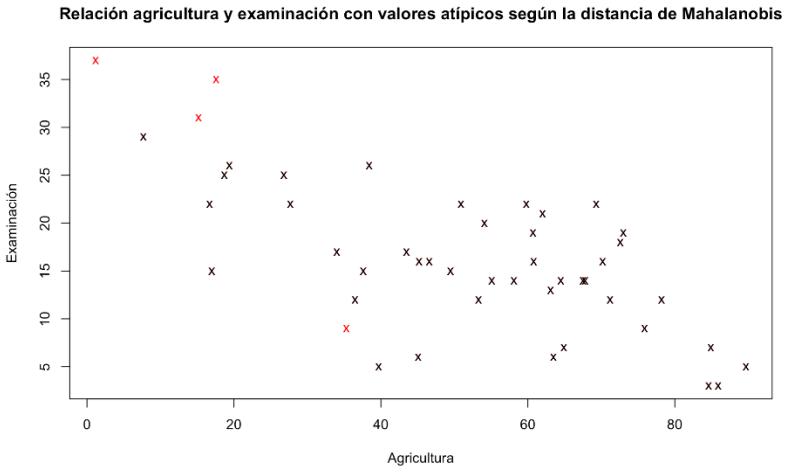
Figura 12



La figura 12 es un boxplot de la distancia de Mahalanobis, tiene el objetivo de determinar la similitud entre las variables aleatorias multidimensionales. Se observan valores atípicos, es decir puntos con una distancia grande y alejados del centro de masa alcanzando como valor máximo 19,9. Sin embargo, la mayoría de los registros se posicionan debajo del 3er cuartil (7,1) y por encima del primero (3,6).

A partir del boxplot se tomó como definición de outlier aquellos valores superiores a 11 con el objetivo de quitarles peso en la próxima figura.

Figura 13



Por último, se graficaron las variables Agricultura y Examinación con correlación inversa y se marcaron en rojo aquellos valores atípicos que superan el valor 11 según la distancia de Mahalanobis dando como resultado una manor variabilidad en el gráfico.