Trabajo Practico Nro. 1

Análisis de datos

Mg. Hernan Ezequiel Martínez

v1.0

Índice

EJERCICIO NRO. 1 4

A) IMPORTE LOS DATOS AL R. 4
B) REEMPLACE LOS DATOS FALTANTES CON NA UTILIZANDO EL COMANDO REPLACE. USE EL HELP PARA VER COMO FUNCIONA ESTE COMANDO. 4
C) DESCRIBA LAS PRINCIPALES CARACTERISTICAS QUE PRESENTAN LOS DATOS PARA LAS VARIABLES GRASAS SAT, ALCOHOL Y CALORÍAS. GRAFICAR LAS MEDIDAS DE POSICION, DISPERSION, HISTOGRAMAS Y REALIZAR GRAFICOS BOXPLOTS. EN TODOS LOS CASOS DEBE COMENTAR LOS RESULTADOS. 4
TABLA 1.1 - ESTADÍSTICOS DEL SET DE DATOS 4
FIGURA 1.1 - BOXPLOT DE LAS VARIABLES, PARECERIAN SER SIMETRICAS, A EXCEPCION DEL ALCOHOL. 5
FIGURA 1.2 - EL CONSUMO DE ALCOHOL PARECE ESTAR CONCENTRADO EN VALORES RELATIVAMENTE BAJOS 6
FIGURA 1.3 - FRECUENCIAS DE APARICION DE MILIGRAMOS DE ALCOHOL EN SANGRE EN LA MUESTRA 7
FIGURA 1.4 - SE OBSERVA UNA DISTRIBUCION "NORMALIZADA" HACIA LAS 1600 KCAL. 7
FIGURA 1.5 - GRASAS SATURADAS, PARECERIA TENER LA MISMA DISTRIBUCION QUE LAS CALORIAS. 8
FIGURA 1.6 - EL EJE VERTICAL CORTA EN LA MITAD DE LAS MUESTRAS, EL HORIZONTAL REPRESENTA LA MEDIA. ALLÍ DEBERIA ESTAR EL 'CENTRO DE MASA'. 9
FIGURA 1.7 - SE DENOTA UN CRECIMIENTO LINEAL EN EL CRECIMIENTO DEL CONSUMO DE GRASAS SATURADAS. 10
FIGURA 1.8 - HAY CRECIMIENTO CON UNA PENDIENTE SIMILAR A LAS GRASAS SATURADAS. 11
CONCLUSIONES - PUNTO C 11
D) DESCRIBA EL COMPORTAMIENTO DE LOS DATOS PARA LA VARIABLES GRASAS SAT, ALCOHOL Y CALORÍAS DE ACUERDO A LA VARIABLE CATEGORICA 'SEXO'. COMENTAR LOS RESULTADOS. 12
FIGURA 1.9 - EL CONSUMO DE GRASAS SATURADAS COMPARADO ENTRE SEXOS RESULTA SIMILAR. 12
FIGURA 1.10 - EL CONSUMO DE CALORIAS APAREADO POR SEXOS ES SIMILAR. 13
FIGURA 1.11 - LOS HOMBRES CONSUMEN MAS QUE LAS MUJERES 14
TABLA 1.2 - ESTADÍSTICOS DEL CONSUMO DE ALCOHOL EN MUJERES 14
TABLA 1.3 - ESTADÍSTICOS DEL CONSUMO DE ALCOHOL EN HOMBRES 14
FIGURA 1.12 - EL PATRÓN DE CONSUMO DE ALCOHOL ENTRE SEXOS RESULTA SIMILAR. 15
FIGURA 1.13 - EL CONSUMO DE CALORIAS ENTRE AMBOS SEXOS PARECE SER DIFERENTE 16
TABLA 1.4 - CONSUMO DE CALORIAS POR PARTE DE LAS MUJERES 16
TABLA 1.5 - CONSUMO DE CALORIAS POR PARTE DE LOS HOMBRES 16
FIGURA 1.14 - EL CONSUMO DE GRASAS SATURADAS ENTRE AMBOS SEXOS RESULTA SIMILAR 17
TABLA 1.6 - ESTADÍSTICOS DE CONSUMO DE GRASAS SATURADAS POR PARTE DE LAS MUJERES 17
TABLA 1.7 - ESTADÍSTICOS DE CONSUMO DE GRASAS SATURADAS POR PARTE DE LOS HOMBRES 17
CONCLUSIONES - PUNTO D 17
E) DESCRIBA EL COMPORTAMIENTO DE LOS DATOS PARA 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: MAS DE 1100 HASTA 1700 CALORÍAS CONSUMIDAS, CATE 3: MÍAS DE 1700 CALORÍAS CONSUMIDAS. 18
FIGURA 1.15 - SE OBSERVA CRECIMIENTO CONJUNTO DE AMBAS VARIABLES 19
FIGURA 1.16 - LAS CATEGORIAS CON MAS CONSUMO CALORICO, CONSUMEN MAS ALCOHOL 20
FIGURA 1.17 - LA MAYOR CANTIDAD DE MUESTRAS ESTAN EN LA SEGUNDA CATEGORIA. 21
CONCLUSIONES - PUNTO E 21

EJERCICIO NRO. 2 22

A) DECIDIR SI LAS VARIABLES DEL CONJUNTO DE DATOS SON INDEPENDIENTES. COMENTAR LOS RESULTADOS OBTENIDOS. 22
TABLA 2.1 - SE OBSERVAN DEPENDENCIAS ENTRE ALGUNAS DE LAS VARIABLES 22

TABLA 2.2 - SE OBSERVAN LAS MAGNITUDES DE LAS VARIABLES Y LA CONCENTRACION DE DATOS 23
TABLA 2.3 - LA MATRIZ DE COVARIANZAS NOS PERMITE RESALTAR RELACIONES 24
CONCLUSIONES - PUNTO A **24**
B) BUSCAR LA PRESENCIA DE DATOS ATIPICOS MEDIANTE LA DISTANCIA DE MAHALANOBIS. COMENTAR LOS RESULTADOS OBTENIDOS. **24**
TABLA 2.4 - DISTANCIAS DE MAHALANOBIS ENTRE LOS CANTONES SUIZOS 24
FIGURA 2.1 - CANTONES AGRUPADOS POR SUS DISTANCIAS DE MAHALANOBIS 25
FIGURA 2.2 - SE CONFIRMAN LAS INTERPRETACIONES DEL HISTOGRAMA. 26
FIGURA 2.2 - LOS PUNTOS QUE ESTAN POR SOBRE LA LINEA PUNTEADA SON DATOS ATIPICOS 27
TABLA 2.5 - CANTONES ATIPICOS 27
CONCLUSIONES - PUNTO B **28**

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."
~ Sir. Arthur Conan Doyle,
A Scandal in Bohemia.

Ejercicio Nro. 1

"En el archivo 'Datos trabajo 1.xls' encontrara 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 los datos faltantes es '999,99' para todas las variables."

a) Importe los datos al R.

Ver archivo "EZEQUIEL MARTINEZ.R".

b) Reemplace los datos faltantes con NA utilizando el comando replace. Use el help para ver como funciona este comando.

Ver archivo "EZEQUIEL MARTINEZ.R".

c) Describa las principales características que presentan los datos para las variables Grasas sat, Alcohol y Calorías. Graficar las medidas de posición, dispersión, histogramas y realizar gráficos boxplots. En todos los casos debe comentar los resultados.

A continuación, se realiza el análisis de las variables correspondientes. Se cuenta con 173 muestras.

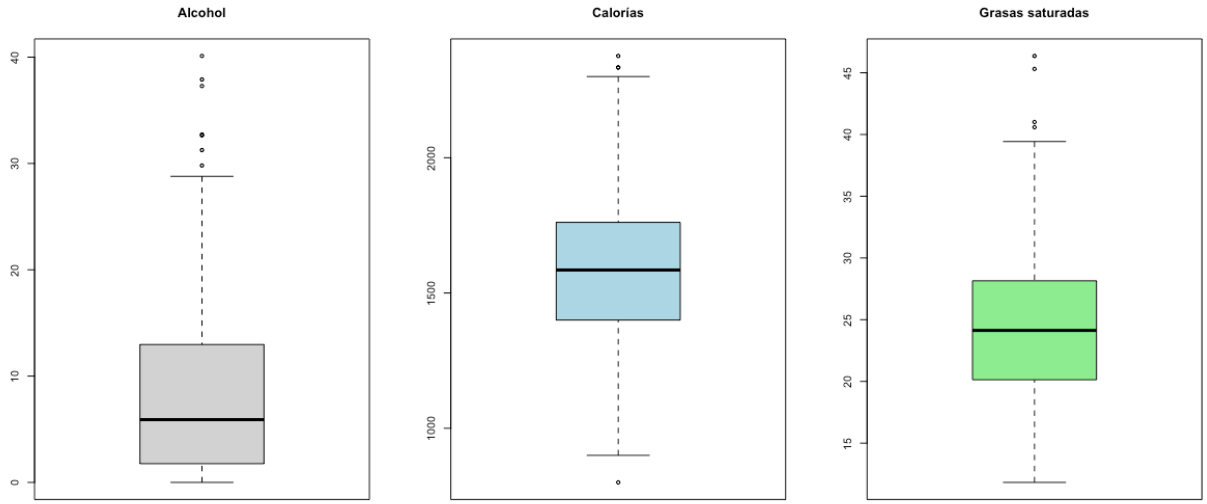
Tabla 1.1 - Estadísticos del set de datos

Grasas_sat	Alcohol	Calorías
Min. :11.82	Min. : 0.000	Min. : 800
1st Qu.:20.17	1st Qu.: 1.780	1st Qu.:1400
Median :24.13	Median : 5.905	Median :1585
Mean :24.77	Mean : 8.832	Mean :1585
3rd Qu.:28.09	3rd Qu.:12.965	3rd Qu.:1761
Max. :46.36	Max. :40.110	Max. :2376
NA's :1	NA's :3	

Observamos en la *tabla 1.1* que la media para "Grasas_sat"[1] y las Calorías se encuentra en valores similares a los de sus respectivas medianas. No siendo así para el alcohol. Esto nos da indicios de que alcohol, podría tener algún tipo de comportamiento asimétrico.

Para un mejor análisis de estos datos, se estudian los boxplots en la Figura 1.1.

Figura 1.1 - Boxplot de las variables, parecerían ser simétricas, a excepción del alcohol.



En la figura 1.1 si bien se observa que los gráficos parecen estar siendo comparados, en realidad las escalas son diferentes.

Rápidamente, podemos comprobar nuestras primeras apreciaciones: si bien cuentan con datos atípicos, Calorías y Grasas_Sat, parecen tener una distribución simétrica, mientras que Alcohol:

[1] Grasas saturadas

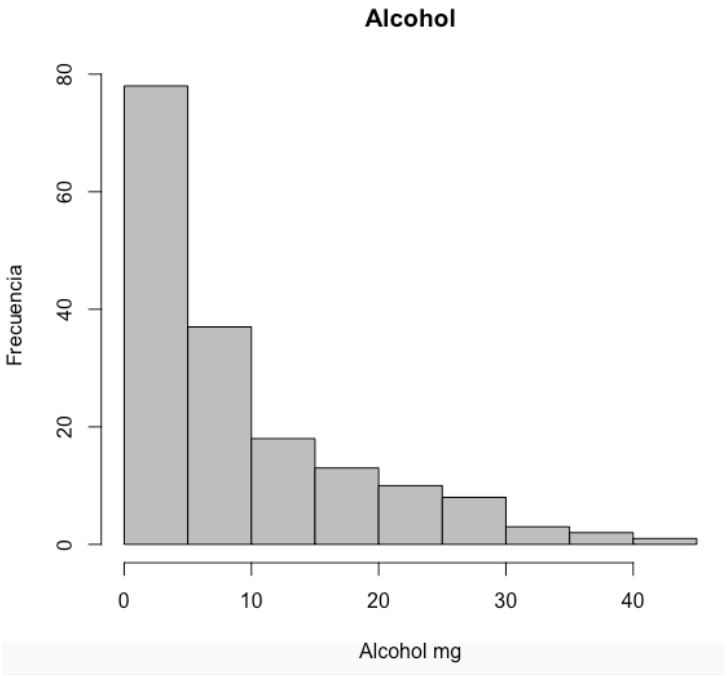
Figura 1.2 - El consumo de alcohol parece estar concentrado en valores relativamente bajos

El consumo de alcohol, parecería contener muchos datos atípicos, y la mediana y los cuartiles, como habíamos anticipado, parecen responder a una distribución binomial o asimétrica (comprobaremos con el histograma).

El resto de las variables, como ya se menciono, parecerían ser simétricas en su distribución con algunos datos atípicos.

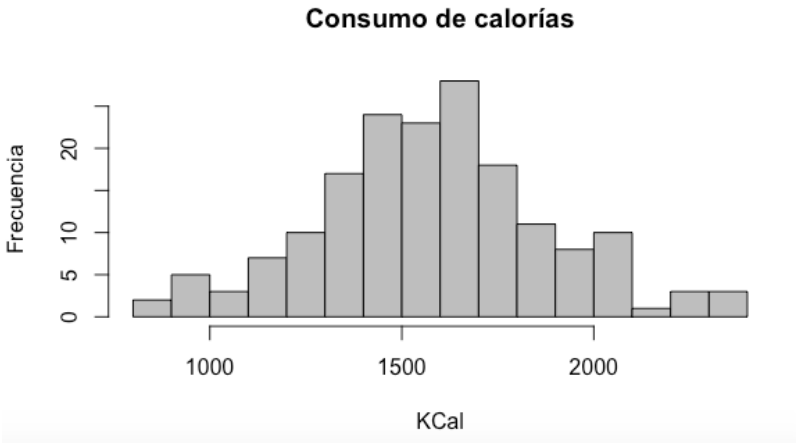
Al realizar los correspondientes histogramas, se confirma que en el caso del consumo de alcohol (Figura 1.3) existe una gran concentración de muestras por debajo de los 20 miligramos; el 50% de todas las observaciones, como se ve en la Tabla 1.1, esta entre los 2 y los 12 miligramos. Esto quiere decir, que muchas personas consumen poco alcohol.

Figura 1.3 - Frecuencias de aparición de miligramos de alcohol en sangre en la muestra



En la Figura 1.4 se observa que el consumo de calorías en la población muestral, podría ser Gaussiana. (un test shapiro revela un p-value de 0,34~). Esto no nos proporciona datos relevantes sobre el comportamiento de la población, sin cruzar con otros datos con los que no se cuentan en este set.[2] Como ser el consumo típico y necesario de calorías para que el cuerpo humano funcione adecuadamente.[3]

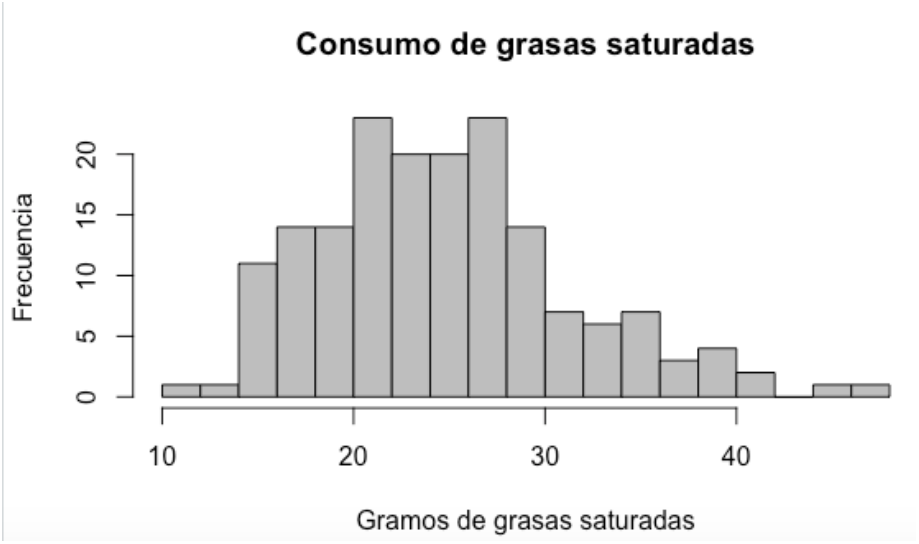
Figura 1.4 - Se observa una distribución "normalizada" hacia las 1600 Kcal.



[2] Se aplico un Test-T, con la idea de identificar si la media de la muestra responde a la de la población, pero el p-value dio muy por debajo del 1%. Conjeturamos que la potencia del test no es la adecuada.

[3] Conjeturamos que estos valores deberían estar entorno a las 2000 kilo-calorías diarias, pero esto excede a los datos y el alcance de este análisis.

Figura 1.5 - Grasas saturadas, parecería tener la misma distribución que las calorías.



La Figura 1.5 aunque parecería pertenecer a una distribución gauss/normal, un test de Shapiro muestra un p-value muy bajo: 0,0014~, lo cual sugiere que no es normal/Gaussiana[4]. De cualquier forma, la forma de la distribución, se asemeja a la del consumo de calorías.

Se estudia la correlación entre ambas: 0.62~, lo cual nos da indicios de que el consumo de calorías podría estar correlacionado con el consumo de grasas saturadas. Conjeturamos que mientras mas calorías se come, probablemente, mas grasas saturadas se ingerirán también.

A continuación, estudiaremos el comportamiento de las variables en sus Scatter Plots:

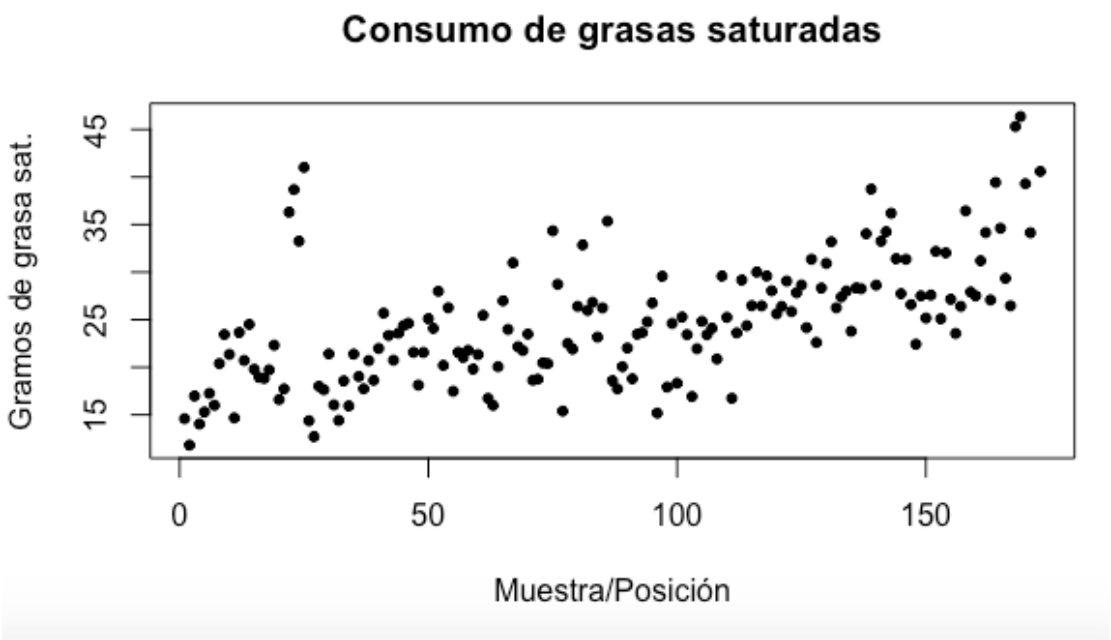
[4] Solo conocemos el uso de este test; entendemos que el test de Kolmogorov podría ser mas efectivo.

Figura 1.6 - El eje vertical corta en la mitad de las muestras, el horizontal representa la media. Allí debería estar el 'centro de masa'.

De los datos en la figura 1.6 se observa que hasta la mitad del n hay un crecimiento lineal del consumo, luego el mismo se vuelve exponencial casi al mismo tiempo que se pasa de la media de consumo del conjunto (8.8~ mg). Esto nos da la idea de que, pasado el umbral de los 9 mg de consumo de alcohol, [5]el mismo se vuelve adictivo muy rápidamente. De allí, el crecimiento exponencial del consumo.

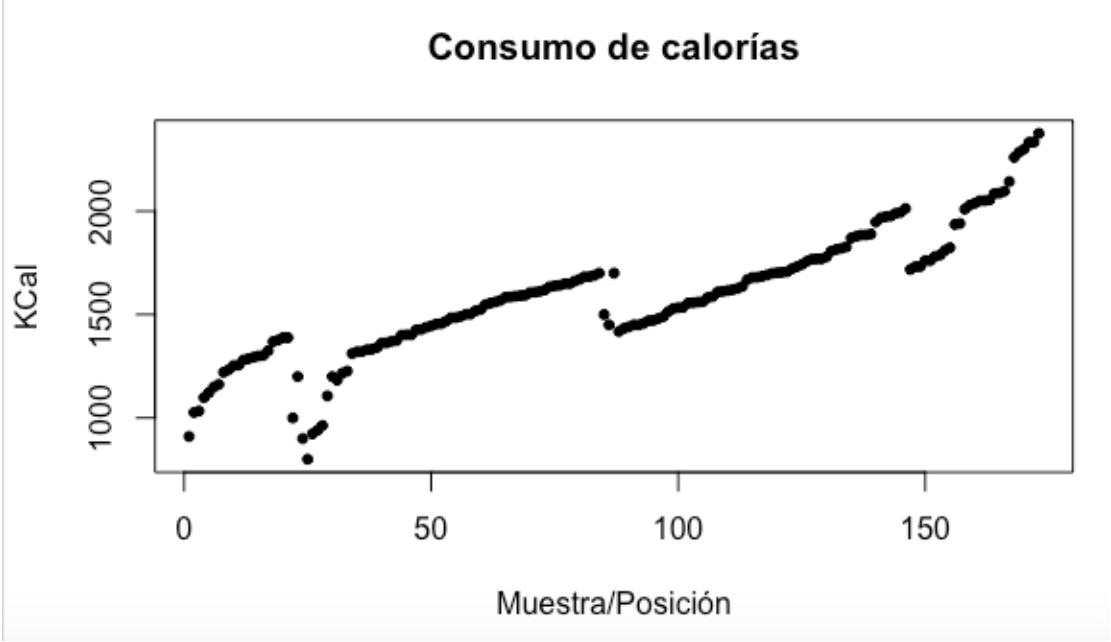
[5] En realidad, aquí hago referencia, nuevamente, a la media del conjunto; se podría argumentar que la mediana (5.9~mg) seria un valor mas robusto para medir esto. Sin embargo, en el grafico, esto no se observa hasta recién pasados los 9 mg! en este caso, las sugerencias teóricas podrían estar detrás de la realidad que presentan los datos.

Figura 1.7 - Se denota un crecimiento lineal en el crecimiento del consumo de grasas saturadas.



La Figura 1.7 muestra que el consumo de grasas saturadas parecería crecer cuando crece el consumo de alcohol, y viceversa. Se computo la correlación entre ambas variables en un 0.66~. Lo cual nos da indicios de que el consumo de grasas saturadas, posiblemente, propicie el consumo de alcohol; o que el consumo de alcohol propicia la ingesta de mas grasas saturadas.

Figura 1.8 - Hay crecimiento con una pendiente similar a las grasas saturadas.



En la Figura 1.8 se observa que las calorías aumentan en consumo de forma similar a como lo hacen las grasas, pero mucho mas suavemente de lo que lo hace el alcohol.

Las grasas saturadas se consumen junto con las calorías, y sin dudas, están correlacionadas (0,62~ es el índice de correlación que se vio previamente). Lo mismo ocurre con el alcohol, que posee una correlación con las calorías de 0,85. Es de suponer, que tanto las grasas saturadas como el alcohol aportan calorías durante su consumo, dado que ambos requieren de ingesta, dado que ambos forman parte de alimentos. Quizás en este sentido, los datos de calorías estén sesgando las muestras; no contamos con datos para afirmar categóricamente esto.

Conclusiones - Punto C

Las grasas saturadas y el alcohol, parecen influenciarse en su consumo mutuamente. Las calorías parecen ser el resultado del consumo de estos. El alcohol parecería dispararse tan pronto pasado el umbral de los 9mg diarios de consumo. [6] Lo cual parecería sugerir que este es el limite diario de consumo para evitar entrar en una vía rápida hacia el alcoholismo.

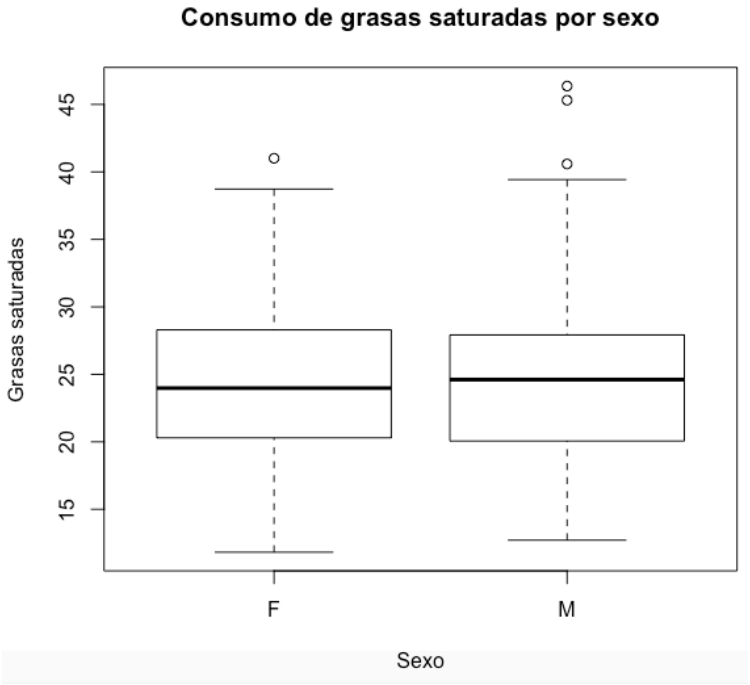
[6] De acuerdo a esta fuente es mas o menos el equivalente a una copa de vino o aperitivo consumida a diario.

d) Describa el comportamiento de los datos para la variables Grasas sat, Alcohol y Calorías de acuerdo a la variable categórica 'Sexo'. Comentar los resultados.

Nos pararemos en las observaciones previas (punto C), de ser necesario, para asistir a nuestras conclusiones.

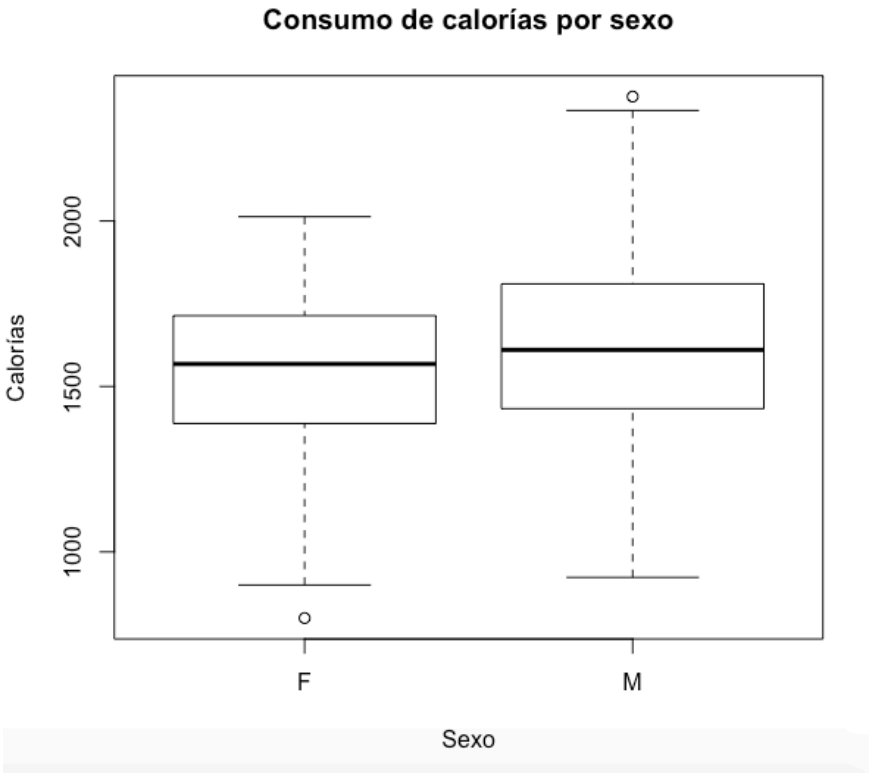
Considérese que en todos los gráficos la letra "F" significa "Femenino" y la "M", "Masculino".

Figura 1.9 - El consumo de grasas saturadas comparado entre sexos resulta similar.



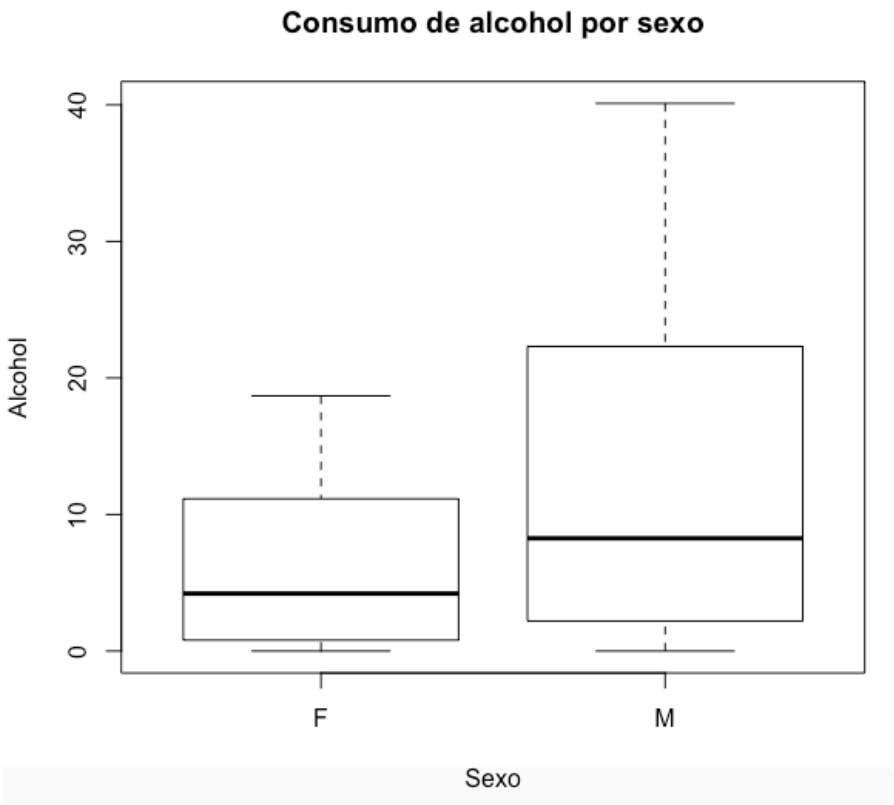
En la Figura 1.9 se evidencia que el consumo es similar entre ambos sexos, aunque en los hombres hay algunos casos extremos adicionales.

Figura 1.10 - El consumo de calorías apareado por sexos es similar.



En la Figura 1.10, los hombres y las mujeres parecen consumir las mismas cantidades, aunque entre los hombres parecería haber un rango mas amplio de consumo.

Figura 1.11 - Los hombres consumen mas que las mujeres



Se observa en la Figura 1.11, que el consumo en los hombres es de rango mas amplio. El 50% de los valores contenidos en el boxplot de los hombres ("la caja") casi contiene a todos los valores de consumo del boxplot de las mujeres. Esto es casi el doble de miligramos de alcohol de lo que consumen las mujeres, conteniendo este a casi todo el rango de consumo de las mujeres.[7]

Esto supone una conjetura. Analicemos los datos duros:

Tabla 1.2 - Estadísticos del consumo de alcohol en mujeres

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.000	0.800	4.210	5.844	11.150	18.690	2

Tabla 1.3 - Estadísticos del consumo de alcohol en hombres

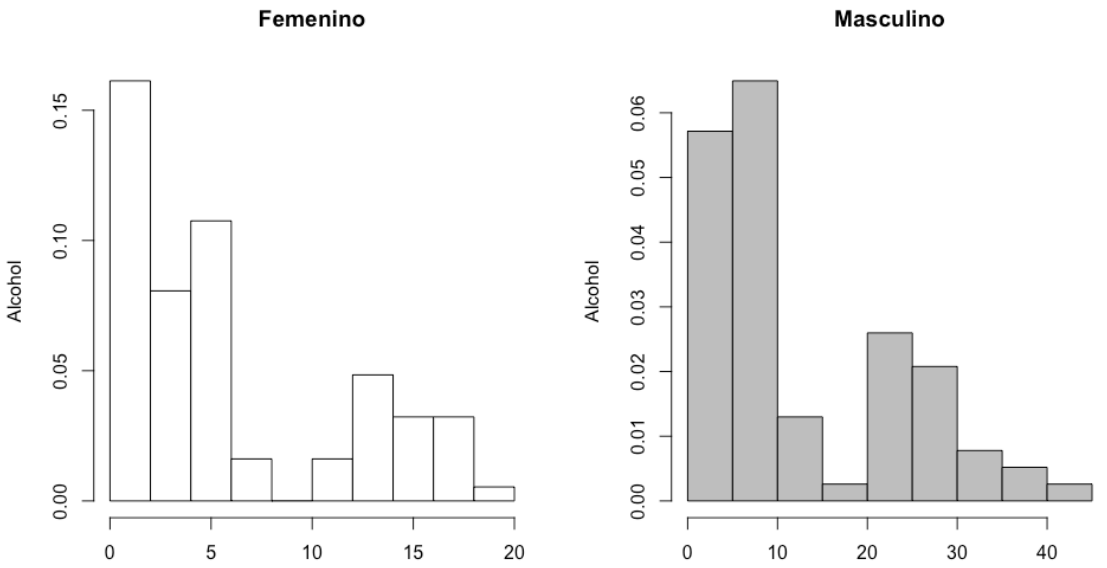
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.00	2.20	8.26	12.44	22.31	40.11	1

Comparando la Tabla 1.2 con la Tabla 1.3, vemos que efectivamente las medias y las medianas difieren mucho. Por ejemplo, la media de los hombres es casi el doble que la de las mujeres, así como también los valores máximos. [8]

[7] Surge del análisis visual de los gráficos.
[8] Asimismo, se observan pocos valores faltantes 2/95 (2%~)en el caso de las mujeres y 1/78 (1,3%~) en el de los hombres, evidenciando poco peso en los resultados.

Los datos sugieren que los casos de consumo extremos, se dan mas entre los hombres.

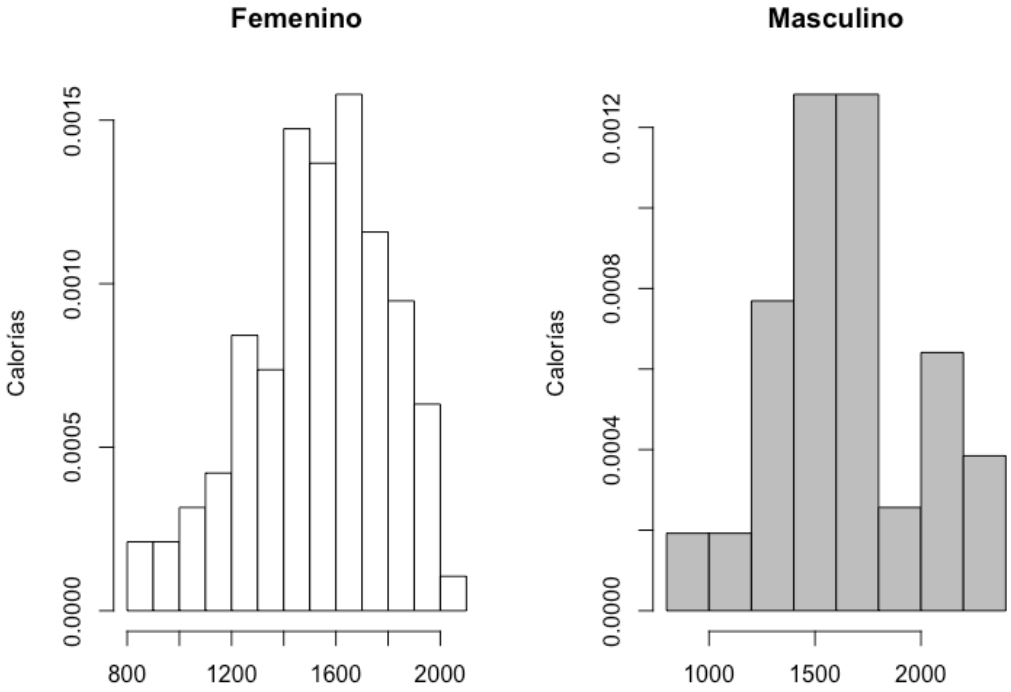
Figura 1.12 - El patrón de consumo de alcohol entre sexos resulta similar.



La Figura 1.12, parece sugerir que los "comportamientos" en el consumo de alcohol entre ambos sexos son similares dado que, comparativamente, las distribuciones parecerían comportarse de forma similar. Debido a esto, es de suponer, que ambos conjuntos de muestras pertenezcan a la misma distribución.[9]

[9] Deberíamos proporcionar un test relevante para confirmar esto, pero preferimos no ahondar en estos detalles por cuestiones de tiempo (y porque aun no lo vimos en la materia).

Figura 1.13 - El consumo de calorías entre ambos sexos parece ser diferente



Como se ve en la Figura 1.13, los hombres parecerían consumir mas calorías. Esto confirma nuestras apreciaciones de la Figura 1.10. Veamos los datos duros, descriptivos:

Tabla 1.4 - Consumo de calorías por parte de las mujeres

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
800	1388	1568	1537	1714	2013

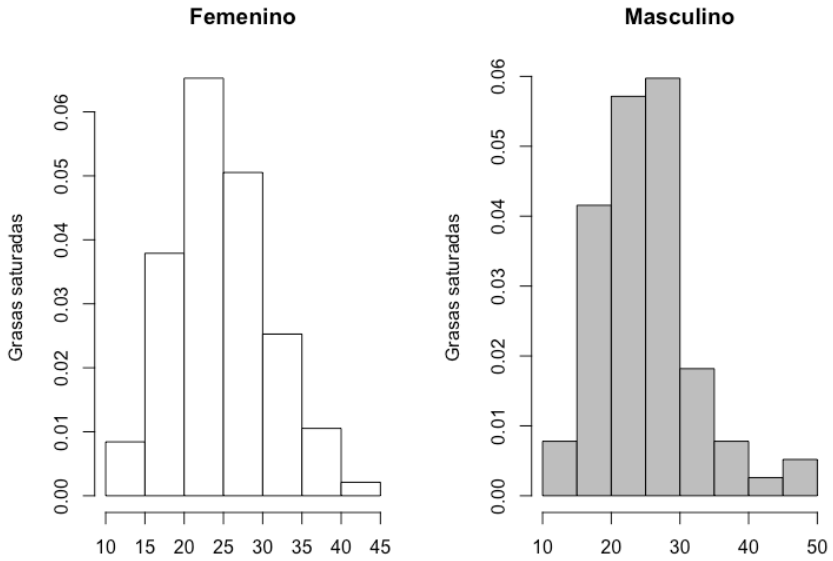
Tabla 1.5 - Consumo de calorías por parte de los hombres

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
923	1435	1610	1642	1804	2376

Si bien las medias provistas en las Tabla 1.4 y Tabla 1.5 son similares, en todos los valores los hombres parecerían consumir un poco mas. Asimismo, los valores máximos difieren, siendo los hombres quienes en algunos casos consumen hasta un 18%~ mas que las mujeres. [10]

[10] 2376/2013 - 1 = 0,18032787~

Figura 1.14 - El consumo de grasas saturadas entre ambos sexos resulta similar



En esta Figura 1.14, se evidencia que el comportamiento es similar en el consumo de grasas saturadas. Ambos sexos tienen el mismo patrón de consumo. Esto ya se había visto en el consumo de alcohol.

Tabla 1.6 - Estadísticos de consumo de grasas saturadas por parte de las mujeres

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
11.82	20.30	23.98	24.59	28.30	41.01

Tabla 1.7 - Estadísticos de consumo de grasas saturadas por parte de los hombres

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
12.71	20.06	24.61	25.01	27.92	46.36	1

Se observan medias y medianas muy similares. Asimismo, los cuartiles, máximos y mínimos se mantienen en valores muy similares. Esto sugiere que las conclusiones obtenidas desde los gráficos podrían ser validas.[11]

Conclusiones - Punto D

Los hombres parecerían consumir mas alcohol que las mujeres, pese a que sus patrones de alimentación son similares. Asimismo, se observaron algunos valores tope un poco mas extremos en los hombres que en las mujeres en el consumo de calorías. Esto no debería sugerir nada nuevo, considerando que en el punto C, ya habíamos identificado que el consumo de calorías y alcohol están fuertemente correlacionados.

[11] Deberíamos aplicar un test que confirme o rechace estas hipótesis. Aun no se cual deberíamos aplicar.

e) Describa el comportamiento de los datos para 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: mas de 1100 hasta 1700 calorías consumidas, CATE 3: mías de 1700 calorías consumidas.

En esta sección, lo que vamos a hacer es estudiar el consumo de alcohol en función de las calorías. Esto ya lo veníamos haciendo desde los puntos precedentes. Aun así, lo que vemos aquí, es que se han definido tres categorías para el consumo:

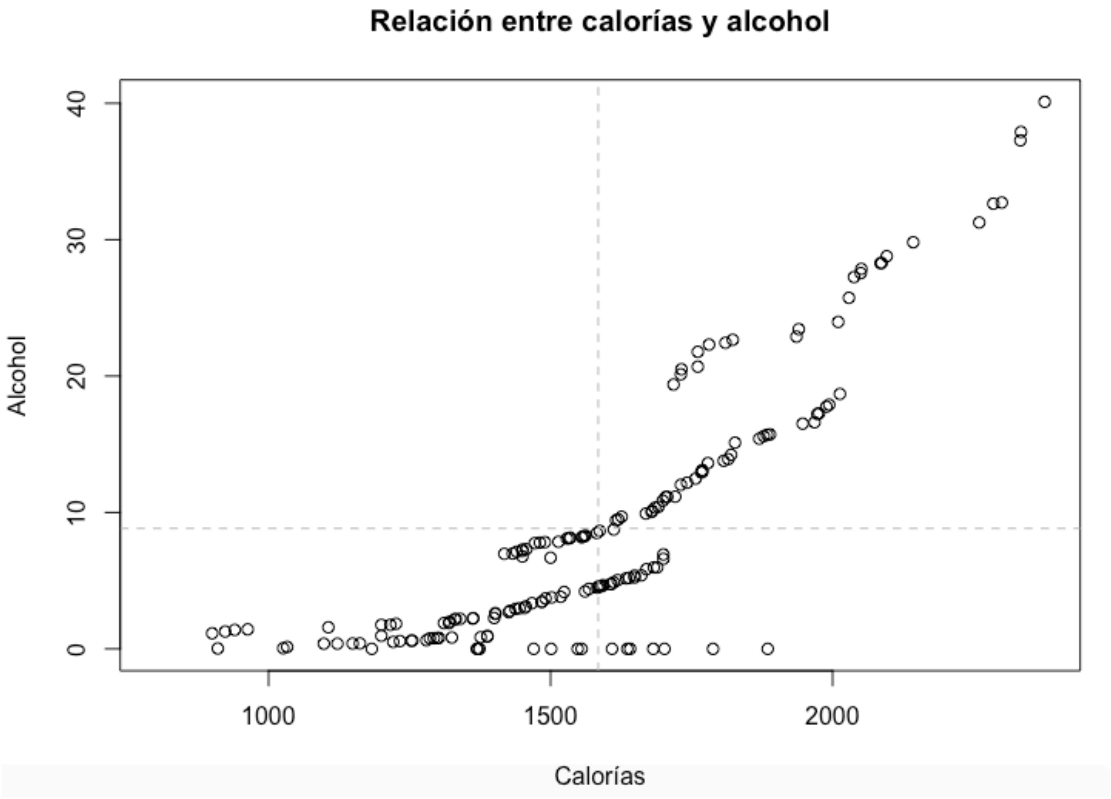
- **CATE 1: 0 a 1100 calorías:** Aquí se encuentran las persona que consumen pocas calorías diarias, posiblemente estén a una dieta mas rigurosa o incluso bajo ayuno.
- **CATE 2: mayor a 1100 y hasta 1700 inclusive:** Estas personas están en el rango medio. Según vimos en la Figura 1.1 y Figura 1.8, es el grupo donde se encuentra el 50% de las muestras observadas.[12]
- **CATE 3: mayor a 1700:** Son las personas que mas consumen.

Lo primero que hacemos es construir un nuevo set de datos que contemple estas categorías, y desde allí, iniciamos el análisis.

Lo primero que debemos recordar, es que la correlación entre el alcohol y las calorías esta en 0,82~, lo cual resulta alto y nos habla de que las variables se influencian mutuamente: mas consumo de alcohol -> mas calorías o viceversa.

[12] La "caja" en el boxplot de calorías.

Figura 1.15 - Se observa crecimiento conjunto de ambas variables

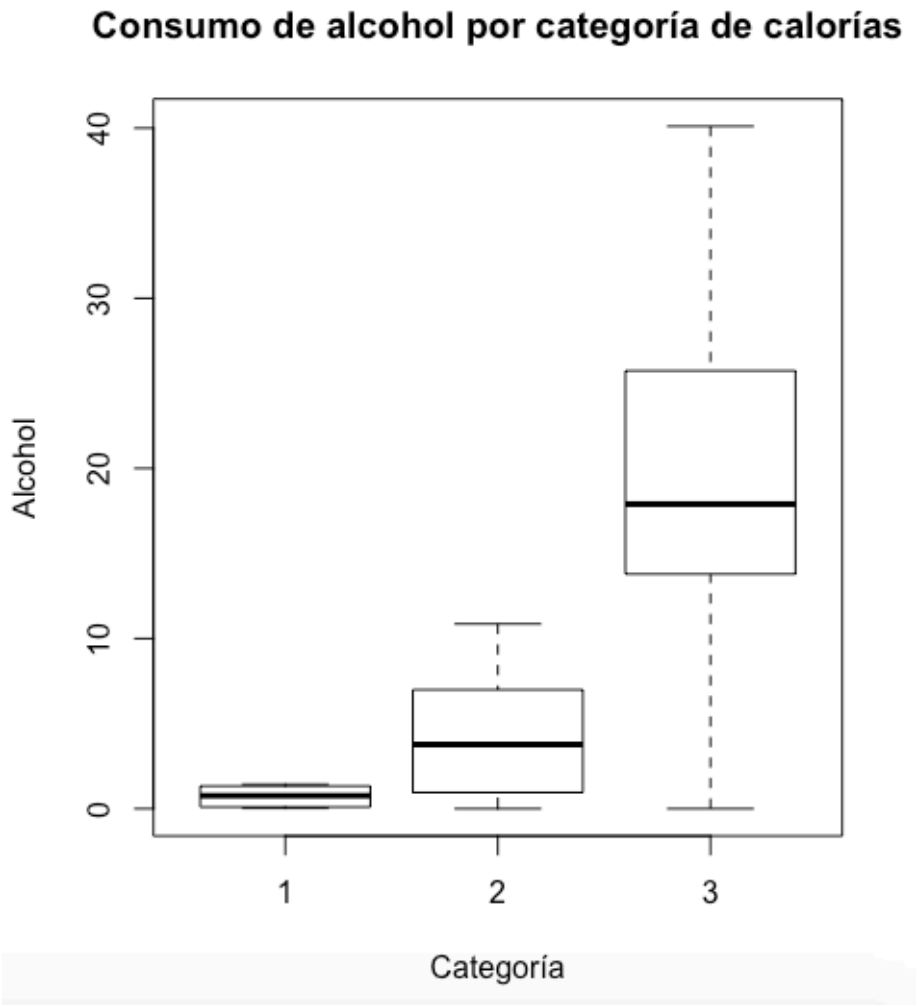


En la Figura 1.15, observamos una aglomeración en el cuadrante inferior izquierdo y en el superior derecho. Lo cual nos da la idea de que el crecimiento de ambas variables es lineal. En calculo, esto significa, que una crece en función de la otra; es decir, en conjunto y de forma directamente proporcional. Aun así, y como ya vimos previamente, el centro de masa parece estar en el par (1535; 8,84~)[13].

Esto nos sugiere que el comportamiento "mas común" del conjunto muestral, debería estar en estos valores. En otras palabras, si las muestras respondieran al comportamiento de la población, este seria el consumo diaria de calorías y alcohol mas típico.

[13] Según puede verificarse en la Tabla 1.1.

Figura 1.16 - Las categorías con mas consumo calórico, consumen mas alcohol



En la Figura 1.16, encontramos algo destacable: se ve muy claramente como, la tercera categoría, consume mucho mas alcohol que las otras dos. Esta conclusión se obtiene de comparar visualmente la mediana [14]de la tercera categoría con la de las otras dos: esta muy por encima de los bigotes de la segunda categoría.

Vale la pena destacar que la desviación estándar de la tercera categoría esta en 8.9~. Esto resulta en una dispersión alta[15]; como puede evidenciarse en el boxplot de la tercera categoría, hay personas que consumen las mismas cantidades de alcohol que los que pertenecen a la segunda.

Algo mas a destacar, es que la primera categoría posee un consumo muy bajo de alcohol y de calorías. Habrá que ver si solo se midieron los alimentos de la dieta o todo lo que las personas consumen en el día.

[14] El cuartil inferior de la tercera categoría, también esta muy por encima, lo cual profundiza esta apreciación.
[15] La media esta en 19,62~, es casi un 50% de variación en el consumo.

Figura 1.17 - La mayor cantidad de muestras están en la segunda categoría.

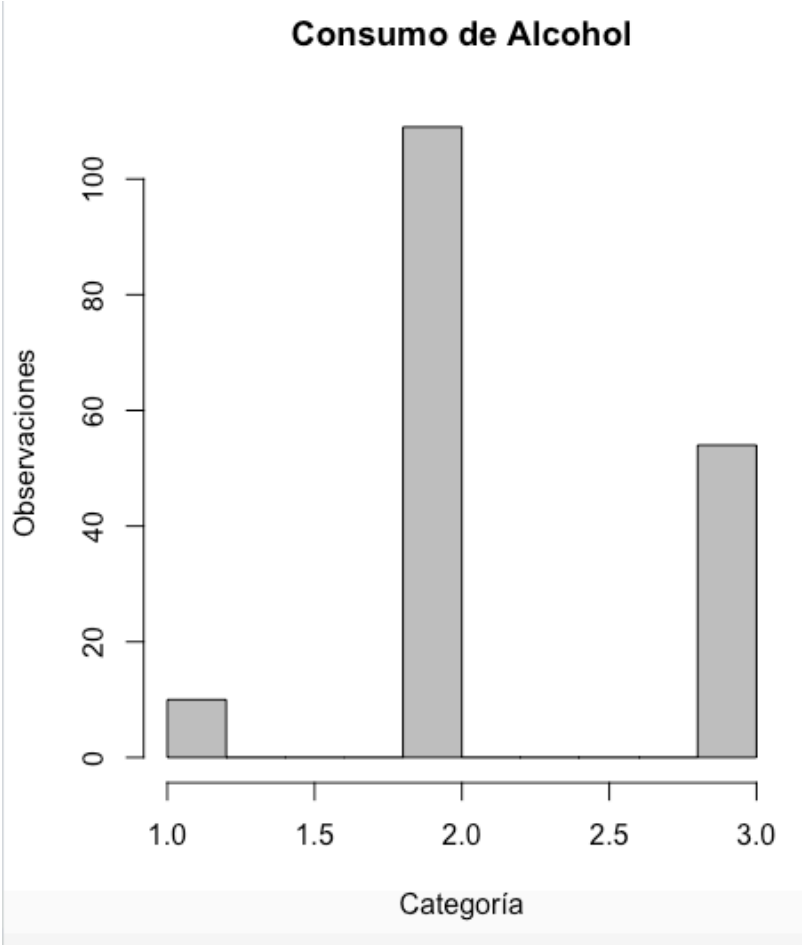


Figura 1.17, donde se identifica la segunda categoría como la poseedora de la mayor cantidad de muestras[16]. Esto nos permite identificar que, habiendo mas observaciones en la segunda categoría que en la tercera, los valores de consumo pesan mucho mas en la tercera categoría, dadas las conclusiones obtenidas con la Figura 1.16.

Conclusiones - Punto E

Todo parece indicar que cuanto mas calorías se consumen, mas alcohol se toma. Pero, además de esto, se identifica un punto de quiebre: pasadas las 1700 calorías de consumo, el crecimiento se vuelve mucho mas alto. No sabemos si en este estudio se restringió el consumo de calorías totales que la persona consume en el día, o si solo se observaron los valores relacionados a la dieta[17].

En cualquier caso, debemos concluir que, al aproximarse el consumo de calorías a los valores recomendados de energía diaria[18], esta dieta, aumenta el consumo de alcohol significativamente, quizás, propiciando el alcoholismo.

---

[16] Casi el doble de las de la tercera categoría y mas de cinco veces la primera.
[17] De ser así, el estudio estaría contaminado.
[18] 2.000 calorías, de acuerdo a la OMS. En este estudio la tercera categoría empieza en los 1700 calorías.

Ejercicio Nro. 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 adopto su condición actual en 1848, fecha hasta la cual cada uno de los cantones entonces existentes poseía sus propias fronteras, ejercito 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 los católicos y los protestantes, con una compleja mezcla de territorios con mayorías católicas y protestantes por todo el país. Las ciudades mas grandes (Berna, Zúrich y Basilea) son predominantemente protestantes. El centro del país, así como el Tesino, son tradicionalmente católicos. En 1980 se voto 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 mas cerca de 100, mayor fertilidad y cuanto mas cerca de 0, menor fertilidad).
b) **Agriculture**: porcentaje de hombres trabajando en agricultura.
c) **Examination**: porcentaje de reclutas que reciben la calificación mas alta en un examen del ejercito.
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.

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

Se estudia la correlación entre las diferentes variables.

Tabla 2.1 - Se observan dependencias entre algunas de las variables

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Fertility	1.0000000	0.35307918	-0.6458827	-0.66378886	0.4636847	0.41655603
Agriculture	0.3530792	1.00000000	-0.6865422	-0.63952252	0.4010951	-0.06085861
Examination	-0.6458827	-0.68654221	1.0000000	0.69841530	-0.5727418	-0.11402160
Education	-0.6637889	-0.63952252	0.6984153	1.00000000	-0.1538589	-0.09932185
Catholic	0.4636847	0.40109505	-0.5727418	-0.15385892	1.0000000	0.17549591
Infant.Mortality	0.4165560	-0.06085861	-0.1140216	-0.09932185	0.1754959	1.00000000

Los índices de correlación, se observan, en algunos casos, muy alejados del 0. Lo que se interpreta como correlación positiva (se comportan de forma proporcional ambas variables) o correlación negativa (inversamente proporcional), dependiendo de si se trata de valores negativos o positivo. Se observa, como es lógico, que la diagonal de la matriz que compone la tabla tiene valores iguales a 1, como es lógico[19].

Hay un solo par de variables con una correlación positiva por encima de 0,5:

- La "examination" contra "education": a mayor aptitud física[20], mayor educación.

Las correlaciones negativas, o que están por debajo de -0,5:

- La "education" con la "fertility": a mayor fertilidad, menor educación.
- La "examination" con la "fertility": a mayor fertilidad, menor aptitud física.
- La "agriculture" con "examination" y "education": a mayor trabajo en el campo, menor educación y menor aptitud física.

Analizamos los datos estadísticos duros de cada variable:

Tabla 2.2 - Se observan las magnitudes de las variables y la concentración de datos

Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Min. :35.00	Min. : 1.20	Min. : 3.00	Min. : 1.00	Min. : 2.150	Min. :10.80
1st Qu.:64.70	1st Qu.:35.90	1st Qu.:12.00	1st Qu.: 6.00	1st Qu.: 5.195	1st Qu.:18.15
Median :70.40	Median :54.10	Median :16.00	Median : 8.00	Median : 15.140	Median :20.00
Mean :70.14	Mean :50.66	Mean :16.49	Mean :10.98	Mean : 41.144	Mean :19.94
3rd Qu.:78.45	3rd Qu.:67.65	3rd Qu.:22.00	3rd Qu.:12.00	3rd Qu.: 93.125	3rd Qu.:21.70
Max. :92.50	Max. :89.70	Max. :37.00	Max. :53.00	Max. :100.000	Max. :26.60

La Tabla 2.2 no aporta datos significativos sobre la correlación, pero queríamos identificar si encontrábamos magnitudes que pudieran resultar disruptivas. Observando las diferentes medias, valores máximos y quintiles, creemos que el uso de la distancia de Mahalanobis o algún otro método de estandarización, podría beneficiarnos en el análisis.

[19] Una variable, comparada consigo misma, deber tener correlación positiva perfecta.
[20] En realidad, es "mayor aptitud para la milicia", ya que el examen medico esta sesgado en este sentido. Pero no creo que nos sirva este nivel de rigurosidad en este informe.

Tabla 2.3 - La matriz de covarianzas nos permite resaltar relaciones

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Fertility	156.04250	100.169149	-64.366929	-79.729510	241.56320	15.156193
Agriculture	100.16915	515.799417	-124.392831	-139.657401	379.90438	-4.025851
Examination	-64.36693	-124.392831	63.646623	53.575856	-190.56061	-2.649537
Education	-79.72951	-139.657401	53.575856	92.456059	-61.69883	-2.781684
Catholic	241.56320	379.904376	-190.560611	-61.698830	1739.29454	21.318116
Infant.Mortality	15.15619	-4.025851	-2.649537	-2.781684	21.31812	8.483802

Lo que no pudimos ver en la Tabla 2.1, lo encontramos aquí, en la Tabla 2.3. Vemos magnitudes de co-variacion grandes, que en las correlaciones no habíamos visto, por ejemplo:

- La "Fertility" tiene covarianza mas alta respecto de "Agriculture" y "Catholic": nos da la idea de que en Cantones católicos y rurales de Suiza, se tienen mas hijos y la población crecería mas. Asimismo, tiene covarianza negativa frente a "Education", lo cual nos da la idea de que en Suiza, mayor educación, supone tener menos hijos.

- La "Education" es fuerte covariancia inversa contra "Catholic": lo que nos da la idea de que a mayor educación, menor cantidad de católicos.
- La "Infant. Mortality" donde mas resalta es en "Catholic", lo cual parece indicar que el ser católico, aumenta las probabilidades de tener mortalidad infantil.

Se observan otras relaciones entre las variables, pero las mas prominentes, son estas.

Conclusiones - Punto A

Las variables no son independientes entre si, cada una influencia en el comportamiento de la otra, a veces mas, otras veces menos. Algunas tienen una correlación y covarianzas muy altas entre ellas, pero siempre van en algún sentido u otro: la influencia esta siempre.

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

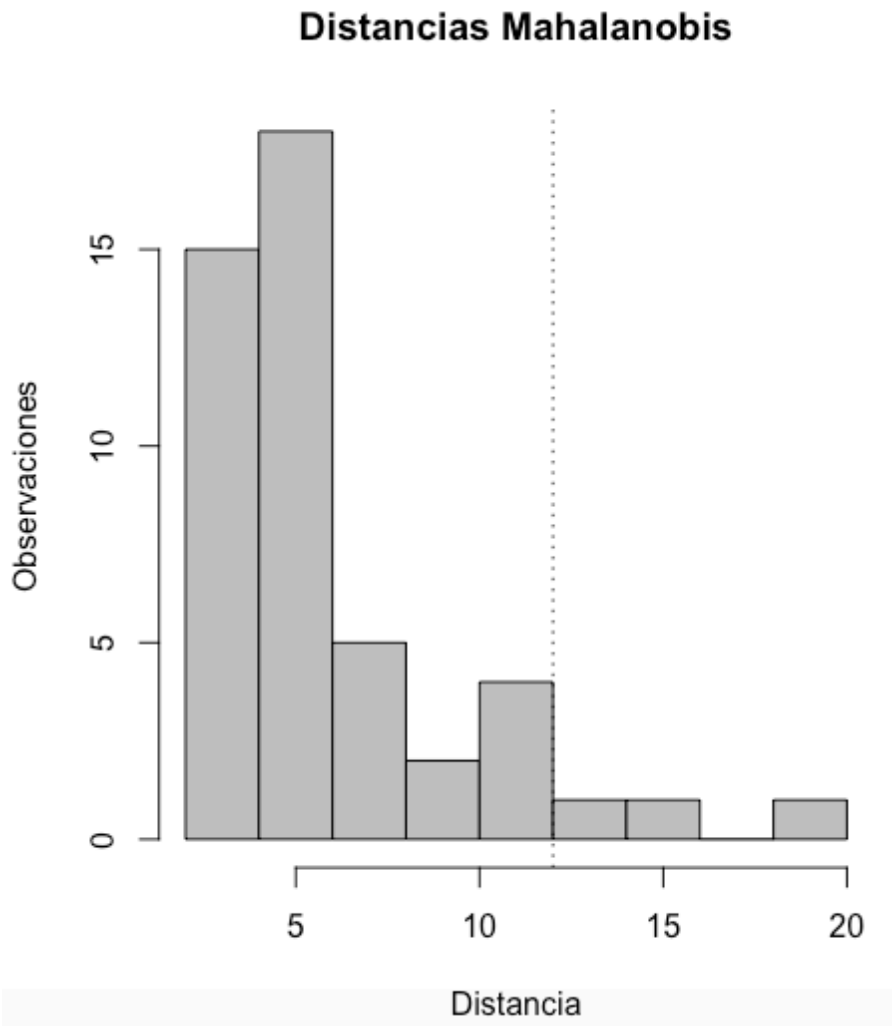
Se estudian los datos aplicando las distancias de Mahalanobis.

Tabla 2.4 - Distancias de Mahalanobis entre los cantones suizos

Courtelary	Delemont	Franches-Mnt	Moutier	Neuveville	Porrentruy	Broye	Glane
6.916428	4.667792	7.957319	4.445657	5.593227	12.675925	5.971568	8.076917
Gruyere	Sarine	Veveyse	Aigle	Aubonne	Avenches	Cossonay	Echallens
2.723267	4.190818	5.575705	3.409203	2.887389	4.257768	5.092576	4.477044
Grandson	Lausanne	La Vallee	Lavaux	Morges	Moudon	Nyone	Orbe
2.174126	3.702675	15.454100	4.204265	2.679660	5.379107	2.514257	5.036103
Oron	Payerne	Paysd'enhaut	Rolle	Vevey	Yverdon	Conthey	Entremont
3.662571	4.865432	7.334404	2.705033	2.118165	3.645059	7.008403	4.542607
Herens	Martigwy	Monthey	St Maurice	Sierre	Sion	Boudry	La Chauxdfnd
4.040018	4.272835	2.886064	4.703667	10.704194	4.281045	3.059886	10.310001
Le Locle	Neuchatel	Val de Ruz	ValdeTravers	V. De Geneve	Rive Droite	Rive Gauche	
3.986104	11.053171	2.970473	7.591800	19.990642	10.752144	9.453386	

Estos datos en crudo, no nos sirven para identificar datos atípicos. Vamos a utilizar otras herramientas graficas, como el Histograma.

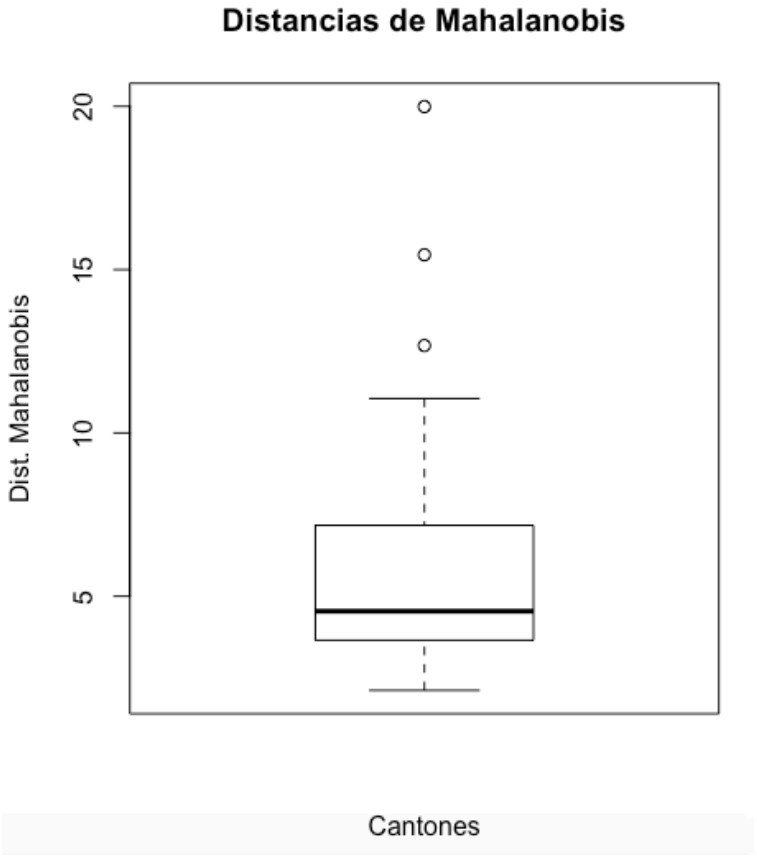
Figura 2.1 - Cantones agrupados por sus distancias de Mahalanobis



En la Figura 2.1, se observa que a partir de la distancia 12, se empiezan a reducir significativamente el numero de observaciones. Vamos a identificar, entonces, esos datos como datos atípicos.

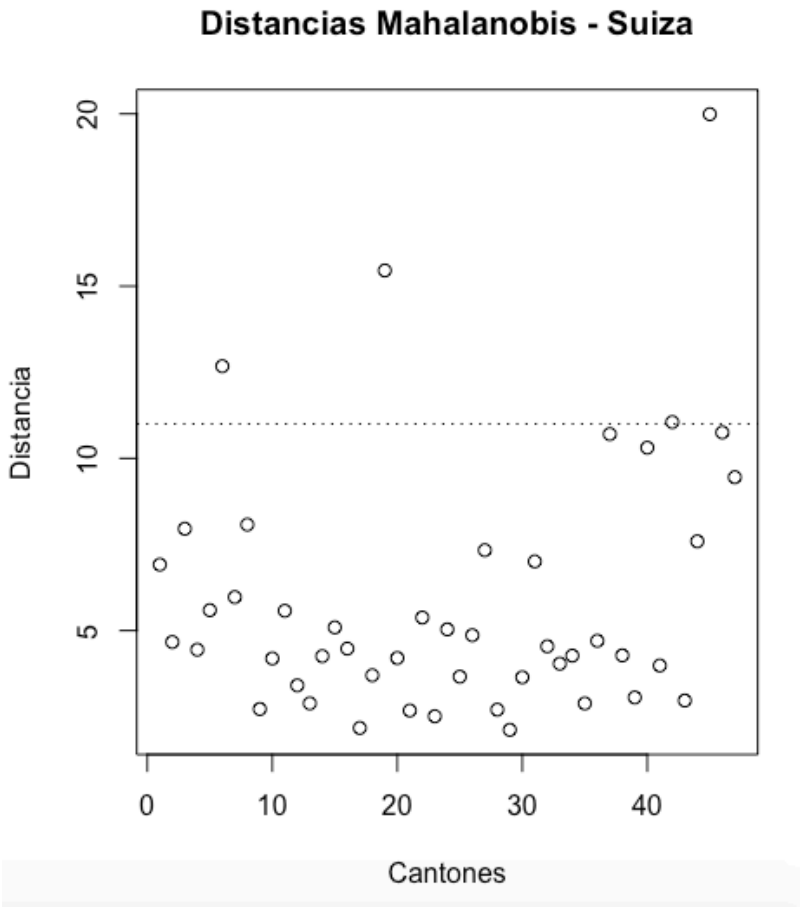
Para mayor seguridad, o como segunda herramienta, graficamos el boxplot del vector:

Figura 2.2 - Se confirman las interpretaciones del histograma.



Efectivamente, en la Figura 2.2, se confirman nuestras sospechas: a partir del valor 11 o 12, comienzan a aparecer los datos atípicos. Debido a este grafico, vamos a probar como punto de corte el 11, en vez del 12, aunque estimamos que no debería haber un impacto significativo en los resultados.

Figura 2.2 - Los puntos que están por sobre la línea punteada son datos atípicos



La Figura 2.2 nos muestra que hay tres cantones con una distancia atípica, y cuatro mas que están en una situación de frontera[21]. La distancia de Mahalanobis tiene en cuenta las covarianzas y las medias de todas las variables, así como las magnitudes de los datos. Con lo cual, la distancia funcionaria como medida de "resumen" respecto de lo similares que los datos pueden ser, contemplando todas las variables disponibles. Esto quiere decir que los siguientes tres cantones, están muy alejados del comportamiento del resto:

Tabla 2.5 - Cantones atípicos

```
Porrentruy
  12.67593
La Vallee
  15.4541
V. De Geneve
    19.99064
```

(La Tabla 2.5, se construye seleccionando los cantones con distancia Mahalanobis mayor a 12 desde la Tabla 2.4.)

[21] "Borderline"

Creemos que estos tres cantones son los que deberían ser considerados verdaderos datos atípicos, los otros cuatro (Entre los que se cuentan Neuchatel, con un valor ligeramente por encima de 11) están cerca pero no lo son.

Conclusiones - Punto B

Los datos atípico, u "outliers", son datos con valores tan alejados de la media, que puede sospecharse que no pertenecen al conjunto muestral. En este caso identificamos tres cantones. La conclusión mas fuerte seria que, dada la naturaleza de las variables: la cultura y costumbres de esos tres lugares podría ser muy diferente de la del resto, quizás hasta el extremo de no comportarse como "típicos" Suizos.