Contenido

Autor: Toni de Andrés Mascaró

Enlace a Github: https://github.com/tonideandres/practicas

Cuestiones

Ejercicio 1

Descripción del dataset. ¿Por qué es importante y que pregunta/problema pretende responder?

El dataset elegido es el usado en la práctica 1. Contiene los datos de la esperanza de vida de diferentes países en el intervalo de años 2000-2015.

Lo campos que contiene son:

- **Country:** nombre del país. Texto.
- **Year:** año de las muestras de información. Numérico(4)
- **Life expectancy Both sexes:** esperanza de vida se ambos sexos. Numérico(2+1)
- **Life expectancy Male:** esperanza de vida de los hombres. Numérico(2+1)
- **Life expectancy Female:** esperanza de vida de las mujeres. Numérico(2+1)
- **Life expectancy both sexes at age 60 (years):** esperanza de vida en años a partir de haber cumplido los 60 para ambos sexos. Numérico(2+1)
- **Life expectancy male at age 60 (years):** esperanza de vida en años a partir de haber cumplido los 60 para hombres. Numérico(2+1)
- **Life expectancy both sexes at age 60 (years):** esperanza de vida en años a partir de haber cumplido los 60 para mujeres. Numérico(2+1)

La pregunta que vamos a tratar de responder es de si la esperanza de vida depende del sexo.

```
> summary(LE)
     hombre          mujer
 Min.   :33.20   Min.   :39.50
 1st Qu.:60.38   1st Qu.:64.40
 Median :68.40   Median :74.50
 Mean   :66.42   Mean   :71.21
 3rd Qu.:72.90   3rd Qu.:78.30
 Max.   :81.30   Max.   :86.80
```

Ejercicio 2

Limpieza de los datos.

2.1. Selección de los datos de interés a analizar. ¿Cuáles son los campos más relevantes para responder al problema?

Dada la inquietud planteada, los datos que nos interesan son aquellos que den por un lado las muestras de la esperanza de vida en los hombres y los datos de la esperanza de vida en mujeres. Los datos más relevantes para responder a la inquietud planteada son:

- **Life expectancy Male:** esperanza de vida de los hombres. Numérico(2+1)
- **Life expectancy Female:** esperanza de vida de las mujeres. Numérico(2+1)

El conjunto de datos inicial tiene también información por país y año, pero como indicado, en este caso esa información no será relevante para la fase de estudio, pero sí la utilizaremos en la fase de preparación para entender bien el conjunto de datos y descubrir si hay valores extremos.

2.2. ¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos?

¿Cómo gestionarías cada uno de estos casos?

Una vez identificados los campos del dataset que queremos usar para conseguir el objetivo marcado, empezamos a trabajar con el entorno R para hacer todo el proceso de preparación de los datos.

```
#Cargamos el dataset
R> library(Rcmdr)
R> LEO <- read.table("C:/Users/Toni/Google Drive/Master
UOC/Asignaturas/Tipologia y ciclo de vida de los
datos/Practica2/LEPC.csv", header=TRUE, sep=";", na.strings="NA",
dec=".", strip.white=TRUE)
R> ls()
R> summary(LEO)
```

```
> summary(LEO)
          Country          Year     Life.expectancy.Both.sexes Life.expectancy..Male Life.expectancy..Female Life.expectancy..Both.sexes.at.age.60..years. Life.expectancy.Male.at.age.60..years.
 Afghanistan       :  16   Min.   :2000   Min.   :36.30          Min.   :33.20        Min.   :39.50          Min.   :10.30                                  Min.   : 8.6
 Albania           :  16   1st Qu.:2004   1st Qu.:62.70          1st Qu.:60.38        1st Qu.:64.40          1st Qu.:16.45                                  1st Qu.:15.4
 Algeria           :  16   Median :2008   Median :71.50          Median :68.40        Median :74.50          Median :18.40                                  Median :17.0
 Angola            :  16   Mean   :2008   Mean   :68.79          Mean   :66.42        Mean   :71.21          Mean   :18.81                                  Mean   :17.4
 Antigua and Barbuda:  16   3rd Qu.:2012   3rd Qu.:75.40          3rd Qu.:72.90        3rd Qu.:78.30          3rd Qu.:21.00                                  3rd Qu.:19.4
 Argentina         :  16   Max.   :2015   Max.   :83.70          Max.   :81.30        Max.   :86.80          Max.   :26.10                                  Max.   :25.0
 (Other)           :2843                  NA's   :11            NA's   :11           NA's   :11
 Life.expectancy.Female.at.age.60..years.
 Min.   :11.20
 1st Qu.:17.40
 Median :19.80
 Mean   :20.09
 3rd Qu.:22.60
 Max.   :28.70
```

Observamos que hay valores NA y que hay una expectativa de vida muy baja (36,30 y 33,20), por lo que decidimos indagar más sobre estos hechos:

```
R> sapply(LEO, function(x)(sum(is.na(x))))
```

```
> sapply(LEO, function(x)(sum(is.na(x))))
                                Country                                    Year                   Life.expectancy.Both.sexes                         Life.expectancy..Male                       Life.expectancy..Female
                                      0                                       0                                          11                                          11                                          11
Life.expectancy..Both.sexes.at.age.60..years.      Life.expectancy.Male.at.age.60..years.    Life.expectancy.Female.at.age.60..years.
                                            0                                           0                                           0
```

A través del menú del RCommander vemos cuales son los países no informados

LE

	country	year	expectancy
49	Andorra	2013	NA
618	Cook Islands	2013	NA
755	Dominica	2013	NA
1628	Marshall Islands	2013	NA
1693	Monaco	2013	NA
1790	Nauru	2013	NA
1887	Niue	2013	NA
1936	Palau	2013	NA
2145	Saint Kitts and Nevis	2013	NA
2194	San Marino	2013	NA
2691	Tuvalu	2013	NA

Y Comprobamos si tienen valores adyacentes por año para tratar de completar la información.

```
> LE[c(47:51,616:620,753:757,1626:1630),]
```

	country	year	expectancy
47	Algeria	2001	71.4
48	Algeria	2000	71.3
49	Andorra	2013	NA
50	Angola	2015	52.4
51	Angola	2014	51.7
616	Congo	2001	52.7
617	Congo	2000	52.9
618	Cook Islands	2013	NA
619	Costa Rica	2015	79.6
620	Costa Rica	2014	79.5
753	Djibouti	2001	57.7
754	Djibouti	2000	57.4
755	Dominica	2013	NA
756	Dominican Republic	2015	73.9
757	Dominican Republic	2014	73.6
1626	Malta	2001	77.8
1627	Malta	2000	77.5
1628	Marshall Islands	2013	NA
1629	Mauritania	2015	63.1
1630	Mauritania	2014	63.0

Tras la visualización de la información, teniendo presente que es una lista ordenada por país y año, confirmamos que se tratar de valores aislados, por lo que al no haber posibilidad de completarlos se procede a la eliminación de los mismos para evitar distorsiones en las estadísticas. Se dejará anotado en el dataset resultante la decisión tomada.

```
R> LEO <- na.omit(LEO)
R> summary(LEO)
```

```
> LEO <- na.omit(LEO)
> summary(LEO)
                Country          Year      Life.expectancy.Both.sexes Life.expectancy..Male Life.expectancy..Female Life.expectancy..Both.sexes.at.age.60..years. Life.expectancy.Male.at.age.60..years.
 Afghanistan        : 16   Min.   :2000   Min.   :36.30              Min.   :33.20          Min.   :39.50           Min.   :10.3                                Min.   : 8.60
 Albania            : 16   1st Qu.:2004   1st Qu.:62.70              1st Qu.:60.38          1st Qu.:64.40           1st Qu.:16.4                                1st Qu.:15.40
 Algeria            : 16   Median :2008   Median :71.50              Median :68.40          Median :74.50           Median :18.4                                Median :17.00
 Angola             : 16   Mean   :2008   Mean   :68.79              Mean   :66.42          Mean   :71.21           Mean   :18.8                                Mean   :17.39
 Antigua and Barbuda: 16   3rd Qu.:2011   3rd Qu.:75.40              3rd Qu.:72.90          3rd Qu.:78.30           3rd Qu.:21.0                                3rd Qu.:19.40
 Argentina          : 16   Max.   :2015   Max.   :83.70              Max.   :81.30          Max.   :86.80           Max.   :26.1                                Max.   :24.00
 (Other)            :2832
 Life.expectancy.Female.at.age.60..years.
 Min.   :11.20
 1st Qu.:17.40
 Median :19.75
 Mean   :20.08
 3rd Qu.:22.60
 Max.   :28.70
```

En caso de haber existido valores adyacentes, se hubiera procedido a completar el dato con la media del valor inferior y superior.

Ahora quedan valores anormalmente bajos en las variables expectancy, así que vamos a explorar un poco más para tratar de averiguar si necesitan algún tratamiento.

```
#Extraemos los datos con valores anormalmente bajos para hacer una
inspección visual.
anos.bajos<-which(LE$expectancy<50)
anos.bajos
tmp<-LE[anos.bajos,]
```

Tras hacer una inspección visual de los datos, vemos que el país que tiene un valor anormalmente bajo y aislado es Haiti, el cual tiene para el año 2010 un valor de 36,3.

```
541                    Chad 2004       48.5
542                    Chad 2003       48.4
543                    Chad 2002       48.1
544                    Chad 2001       48.0
545                    Chad 2000       47.6
859                 Eritrea 2000       45.3
1121                  Haiti 2010       36.3
1466                Lesotho 2009       49.4
1467                Lesotho 2008       47.8
1468                Lesotho 2007       46.2
1469                Lesotho 2006       45.3
1470                Lesotho 2005       44.5
```

El resto de países tienen una progresión que consideramos normal, si es cierto que alguno de ellos tiene una tendencia creciente muy marcada, pero igualmente es cierto que son países que han pasado por guerras muy duras, lo cual, con la información que se dispone, justifica la tendencia, así que se consideran normales y se procede a trabajar el valor de Haiti.

```
R> LEO[which(as.character(LEO$Country)=="Haiti"),]
```

	Country	Year	Life.expectancy.Both.sexes	Life.expectancy..Male	Life.expectancy..Female	Life.expectancy
1124	Haiti	2015	63.5	61.5	65.5	
1125	Haiti	2014	63.1	61.0	65.2	
1126	Haiti	2013	62.7	60.7	64.8	
1127	Haiti	2012	62.3	60.4	64.3	
1128	Haiti	2011	62.3	60.3	64.3	
1129	Haiti	2010	36.3	33.2	40.3	
1130	Haiti	2009	62.5	60.3	64.7	
1131	Haiti	2008	62.1	59.9	64.4	
1132	Haiti	2007	61.8	59.6	64.0	
1133	Haiti	2006	61.1	59.0	63.3	
1134	Haiti	2005	60.5	58.4	62.7	
1135	Haiti	2004	58.7	56.4	61.1	
1136	Haiti	2003	59.7	57.5	61.9	
1137	Haiti	2002	59.3	57.1	61.5	
1138	Haiti	2001	58.9	56.8	61.1	
1139	Haiti	2000	58.6	56.5	60.8	

```
>
```

Podemos observar que el valor del año 2009 es de 62,5 y el de 2011 es de 62,3, tras un trabajo de documentación se concluye que en el año 2010 Haiti sufrió un terremoto que devastó el país dejando 316.000 muertos directos más todos aquellos que perecieron posteriormente debidos a lo devastado que dejó al país.

Vamos a ver el impacto de los outliers sobre las estadísticas:

```
> summary(haiti_original)
       country         year          leboth           lem             lef
 Haiti      :16   Min.   :2000   Min.   :36.30   Min.   :33.20   Min.   :40.30
 Afghanistan: 0   1st Qu.:2004   1st Qu.:59.20   1st Qu.:57.02   1st Qu.:61.40
 Albania    : 0   Median :2008   Median :61.45   Median :59.30   Median :63.65
 Algeria    : 0   Mean   :2008   Mean   :59.59   Mean   :57.41   Mean   :61.87
 Andorra    : 0   3rd Qu.:2011   3rd Qu.:62.35   3rd Qu.:60.33   3rd Qu.:64.47
 Angola     : 0   Max.   :2015   Max.   :63.50   Max.   :61.50   Max.   :65.50
 (Other)    : 0
```

Ilustración 1. Conjunto Haiti original

```
> summary(haiti)
       country         year          leboth           lem             lef
 Haiti      :16   Min.   :2000   Min.   :58.60   Min.   :56.40   Min.   :60.80
 Afghanistan: 0   1st Qu.:2004   1st Qu.:59.60   1st Qu.:57.40   1st Qu.:61.80
 Albania    : 0   Median :2008   Median :61.95   Median :59.75   Median :64.15
 Algeria    : 0   Mean   :2008   Mean   :61.27   Mean   :59.11   Mean   :63.39
 Andorra    : 0   3rd Qu.:2011   3rd Qu.:62.55   3rd Qu.:60.33   3rd Qu.:64.70
 Angola     : 0   Max.   :2015   Max.   :63.50   Max.   :61.50   Max.   :65.50
 (Other)    : 0
```

Ilustración 2. Conjunto Haiti modificado

Vemos como los valores de las medias y cuartiles no se ven afectados significativamente pasando las medias de:

- Leboth: 59,59 a 61,27 años
- Lem: 57,41 a 59,11 años
- Lef: 61,87 a 63,39 años

Observamos que el impacto de los outliers es significativo, pero que al unir estos datos con el resto de países será despreciable por la gran cantidad de muestras existente. Debido a los motivos expuestos, se decide dejar los outliers con su valor original y dejar una notación en los datos dejando constancia de este hecho por si posteriormente se tiene que hacer algún trabajo con el conjunto de datos resultante

Ejercicio 3

Análisis de los datos.

3.1. Selección de los grupos de datos que se quieren analizar/comparar.

Para realizar los estudios se trabaja el conjunto de datos para que contenga las variables deseadas y los valores que se han decidido tras el estudio de valores no informados y extremos.

Seleccionamos las variables de hombre y mujeres y sus observaciones:

```
R> LE<-LEO
R> head(LE)
R> LE<-LE[,4:5]
R> colnames(LE)<-c("hombre","mujer")
R> sapply(LE, function(x)(sum(is.na(x))))
R> LE <- na.omit(LE)
R> sapply(LE, function(x)(sum(is.na(x))))
R> summary(LE)
```

```
> summary(LE)
     hombre           mujer
 Min.   :33.20   Min.   :39.50
 1st Qu.:60.38   1st Qu.:64.40
 Median :68.40   Median :74.50
 Mean   :66.42   Mean   :71.21
 3rd Qu.:72.90   3rd Qu.:78.30
 Max.   :81.30   Max.   :86.80
```

En este caso, de la observación de las medias se infiere que los hombres viven 4,79 años menos que las mujeres.

Los cuartiles están próximos a las medias, lo cual nos denota que hay poca dispersión de los datos.

Para continuar se modifica el conjunto para que esté según el formato Tidy data. Para ello se construyen dos columnas, la primera que contendrá el nombre de las variables y la segunda el valor de las variables.

```
R> library(tidyr)
R> LE<-gather(LE,sexo,edad)
R> head(LE)
R> summary(LE)
```

```
> summary(LE)
     sexo                edad
 Length:5856        Min.   :33.20
 Class :character   1st Qu.:62.10
 Mode  :character   Median :70.70
                    Mean   :68.82
                    3rd Qu.:76.40
                    Max.   :86.80
```

Observamos que los valores de la media y cuartiles se han ajustado a un valor medio de los dos grupos de variables.

3.2. Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.

En este vamos a ver en que medida las distribuciones que tenemos se pueden ajustar a la curva normal para decidir posteriormente los métodos a utilizar a la hora de comprobar la relación entre la esperanza de vida de los hombres y las mujeres. Vamos a realizar contrastes de normalidad para comprobar si se verifica la hipótesis de normalidad necesaria para que el resultado de algunos análisis sea de fiable.

Pasamos a comprobar la normalidad de las variables mediante la hipótesis nula

H0: la muestra proviene de una distribución normal

Y hacemos los siguientes test a los datos normalizados

- Test de Shapiro-Wilk
- Test de Anderson-Darling
- Test de Cramer-von Mises
- Test de Lillie
- Test de Pearson
- Test de Shapiro-Francia

Test de Shapiro-Wilk:

El test de Shapiro falla por exceder de 5.000 registros.

Test de Anderson-Darling

```
> ad.test(edad)$p.value
[1] 3.7e-24
```

Test de Cramer-von Mises

```
> cvm.test(edad)$p.value
Warning in cvm.test(edad) :
  p-value is smaller than 7.37e-10, cannot be computed more accurately
[1] 7.37e-10
```

Test de Lillie

```
> lillie.test(edad)$p.value
[1] 3.388353e-120
```

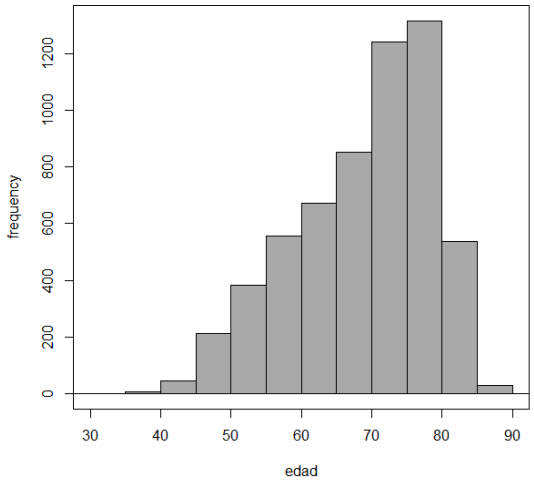
Test de Pearson

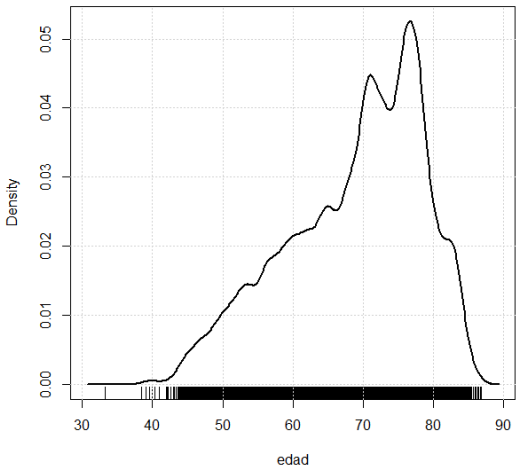
```
> pearson.test(edad)$p.value
[1] 4.47199e-191
```

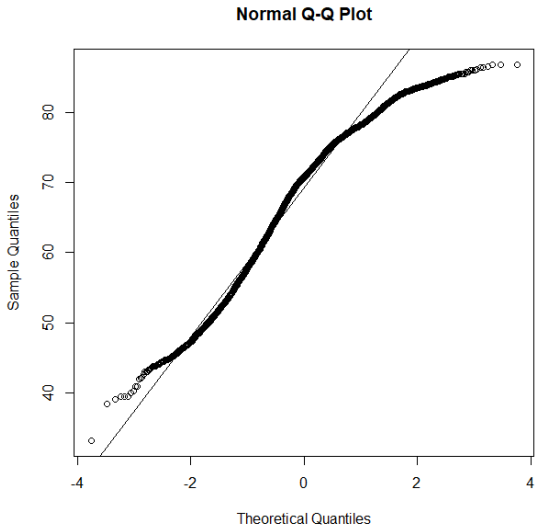
Test de Shapiro-Francia

```
> sf.test(edad)$p.value
Error in sf.test(edad) : sample size must be between 5 and 5000
```

Observamos que todos los test devuelven un valor P muy bajo, por debajo del 5%, por lo que concluimos que no se cumple la hipótesis nula y sabemos que no estamos ante una distribución normal. Sacamos un par de gráficos para comprobarlo visualmente:







Las gráficas confirman visualmente lo que los números ya indicaban. Tanto el histograma como la gráfica de densidad muestran una distribución que no se parece a la normal. La forma en S de la última gráfica también denota que la muestra no se adecua a la normal.

Ahora vamos a comprobar la normalidad y homogeneidad de la varianza:

HO: rechazamos la hipótesis nula y buscamos un p-valor lo más alto posible.

Prueba de Bartlett:

```
> bartlett.test(edad~sexo, data=LE)

        Bartlett test of homogeneity of variances

data:  edad by sexo
Bartlett's K-squared = 34.759, df = 1, p-value = 0.000000003731
```

Prueba de Levene

```
> levene.test(edad,sexo)

        modified robust Brown-Forsythe Levene-type test based on the absolute deviations from the median

data:  edad
Test Statistic = 14.862, p-value = 0.0001169
```

3.3. Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.

H0: la esperanza de vida no depende del sexo

Anteriormente hemos comprobado si la muestra se asemeja a una normal y hemo concluido que no. Este hecho condiciona las pruebas que tenemos que realizar a continuación para aceptar la hipótesis nula o la alternativa.

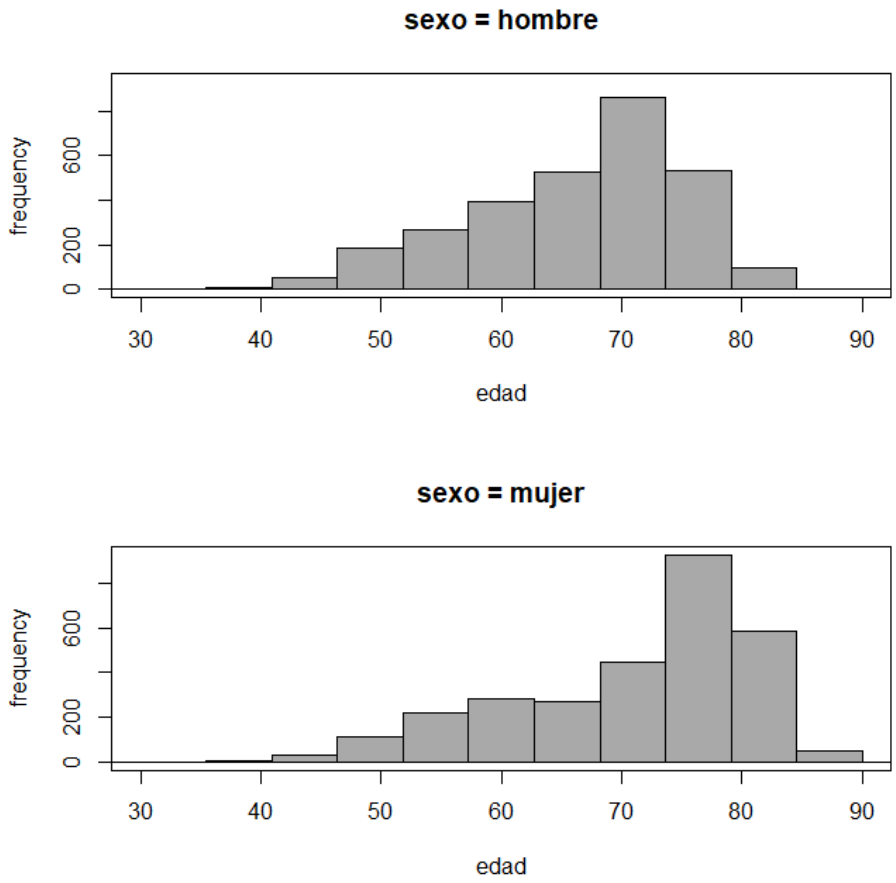
Para comprobar la hipótesis nula, en este caso al tener un tipo de datos no normales y ser variables dependientes, hacemos el test de Wilcoxon. Este test es un test no paramétrico que se basa en la comparación de medianas y trabaja sobre un rango de orden. En este caso la H0 será que la mediana de las diferencias es igual a 0. Para aceptar la hipótesis nula la P debería estar por encina del 5%, o lo que es lo mismo P>0,05

```
> wilcox.test(edad ~ sexo, alternative="two.sided", data=LSF)

        Wilcoxon rank sum test with continuity correction

data:  edad by sexo
W = 2875300, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

El valor de P está muy por debajo del 0,05, por lo que rechazamos la hipótesis nula y concluimos que no hay evidencias que demuestren que la esperanza de vida no depende del sexo

Ejercicio 4

Representación de los resultados a partir de tablas y gráficas.



Histograma donde se observa que las distribución de la esperanza de vida entre los hombres y las mujeres es casi idéntica con un desplazamiento hacia la derecha en el caso de las mujeres, lo que significa que tienen mayor esperanza de vida.

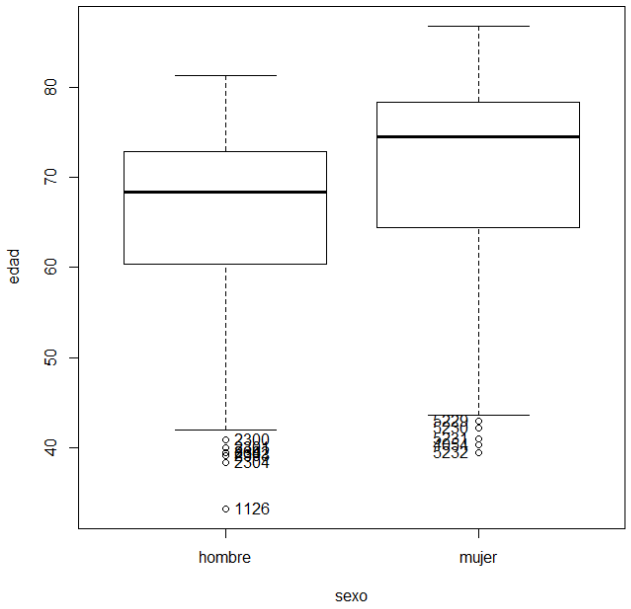
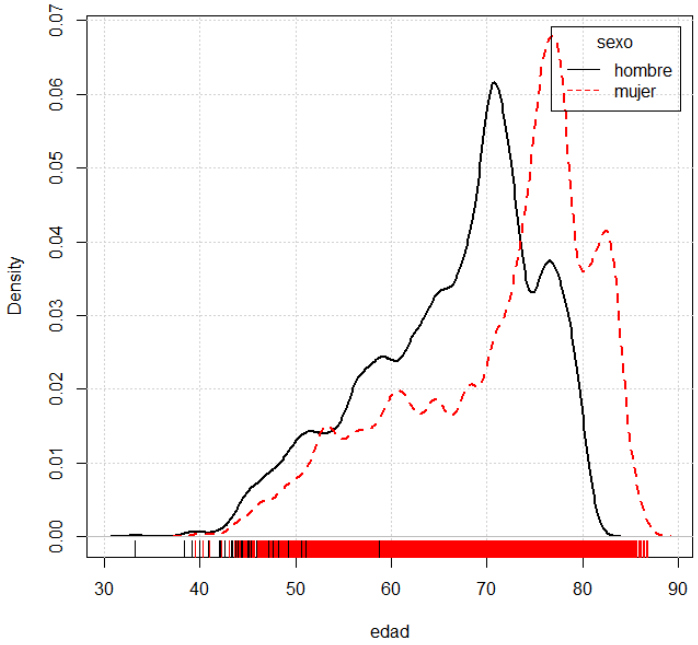


Gráfico de cajas donde se observa que las mujeres tienen una esperanza de vida superior a la de los hombres, sin embargo, el quartil por encima de la media es menos en el caso de las mujeres que en el de los hombres, lo que significa que de media viven más, pero que una vez superada la media tienen menos esperanza de vida.



Al igual que en caso del histograma, estamos antes dos curvas muy similares con un desplazamiento hacia la izquierda de la esperanza de vida de las mujeres.

Ejercicio 5

Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Tal como se ha comentado en el apartado 3.3, la prueba realizada arroja un valor de P muy por debajo del 0,05, por lo que rechazamos la hipótesis nula y concluimos que no hay evidencias que demuestren que la esperanza de vida no depende del sexo. El análisis gráfico también nos hace llegar a la misma conclusión.