PAC 1. Probabilitat i estadística

Amelia Martínez

24/2/2022

Aquesta PAC esta basada en una base de dades obtinguda a partir del llocweb de l'Euroestat Euroestat (Oficina Europea de Estadística). Les dades estan disponibles al fitxer "data_pac" en format csv i xlsx.

Aquesta base de dades conté informació sobre el percentatge de compres que es realitzen per internet de particulars (fins 2019), per a homes i dones d'entre 16 y 74 anys.

Conté les següents variables:

- *geo* = Àrea Geogràfica
- *sex* = Sexe
- *household_goods* = articles domèstics
- *films_music* = películes i música
- *clothes_sports_goods* = roba i articless esportius
- *electronic_equipment* = equips electrònics
- *tickets_for_events* = entrades per a events
- *travel_and_holiday_accommodation* = viatges, vacances i allotjament

Importem les dades, que anomenarem *data_pac1* :

```
## # A tibble: 6 x 8
##   geo      household_goods films_music clothes_sports_goods
electronic_equipment
##   <chr>              <dbl>       <dbl>                <dbl>
<dbl>
## 1 Belgium               30          19                   48
23
## 2 Belgium               28          12                   62
14
## 3 Bulgaria              39           5                   70
22
## 4 Bulgaria              32           3                   82
9
## 5 Czechia               46          15                   64
40
## 6 Czechia               51           9                   79
16
## # ... with 3 more variables: tickets_for_events <dbl>,
## #   travel_and_holiday_accommodation <dbl>, sex <chr>
```

Un cop importades les dades:

Pregunta 1. (25%)

Realitzeu un resum numèric i gràfic per a la variable *electronic_equipment* i comenteu el resultat.

Veiem que les dades corresponents a cada país estan dividides en el dos sexes. Si el que volem es fer un anàlisi basat en les diferències entre països sense tenir en compte el sexe, podem agrupar les dades per països. Farem servir el paquet *dplyr*.

Com les dades numèriques són percentatges, no podem sumar-les. Farem la mitjana entre homes i dones per a cada regió geogràfica.

```
## # A tibble: 6 x 2
##   geo      electronic_equipment
##   <chr>                   <dbl>
## 1 Austria                  34
## 2 Belgium                  18.5
## 3 Bulgaria                 15.5
## 4 Croatia                  26
## 5 Cyprus                   15
## 6 Czechia                  28
```

Ara ja podem fer un resum numèric de la variable.

Estadística descriptiva

Existeixen diferents opcions per obtenir els estadístics descriptius. Amb la funció *summary()*:

```
##      geo            electronic_equipment
##  Length:31          Min.   :15.00
##  Class :character   1st Qu.:21.75
##  Mode  :character   Median :26.50
##                     Mean   :26.42
##                     3rd Qu.:30.50
##                     Max.   :42.50
```

Creant la nostra pròpia funció, afegirem la variància i la desviació típica:

```
##      mitjana  var desv.tip min    Q1   Q2   Q3  max
## [1,]   26.42 54.3     7.37  15 21.75 26.5 30.5 42.5
```

El segon quartil equival a la mediana.

Amb el paquets Hmisc, MVN, pastecs o psych:

```
## EE_paisos$electronic_equipment
##        n  missing distinct     Info     Mean      Gmd      .05
.10
##       31        0       22    0.995    26.42    8.413    15.25
```

```
16.00
##      .25      .50      .75      .90      .95
##    21.75    26.50    30.50    36.50    39.25
##
## lowest : 15.0 15.5 16.0 18.5 19.5, highest: 32.5 34.0 36.5 42.0 42.5

##      nbr.val     nbr.null       nbr.na          min          max
range
##   31.0000000    0.0000000    0.0000000   15.0000000   42.5000000
27.5000000
##          sum       median         mean      SE.mean CI.mean.0.95
var
##  819.0000000   26.5000000   26.4193548    1.3235048    2.7029574
54.3016129
##      std.dev     coef.var
##    7.3689628    0.2789229
```

Test de normalitat:

El **test Kolmogorov-Smirnov** és un test de normalitat numèric la hipòtesi nul·la del qual, H_0, considera que la distribució de la variable seleccionada prové d'una distribució normal. Per exemple, si el nivell de significació o p-valor obtingut al test KS és 0.20, aleshores per a un nivell de significació del 0.05 (el que està fora del 95 % de probabilitats) no rebutgem la hipòtesi nul·la, ja que el p-valor és 0.20>0.05. Per tant, segons aquest test, podem considerar que la distribució de les dades és normal. En resum:

Si Sig. (p-valor) > 0.05 acceptem H_0 (hipòtesi nul·la) → distribució normal

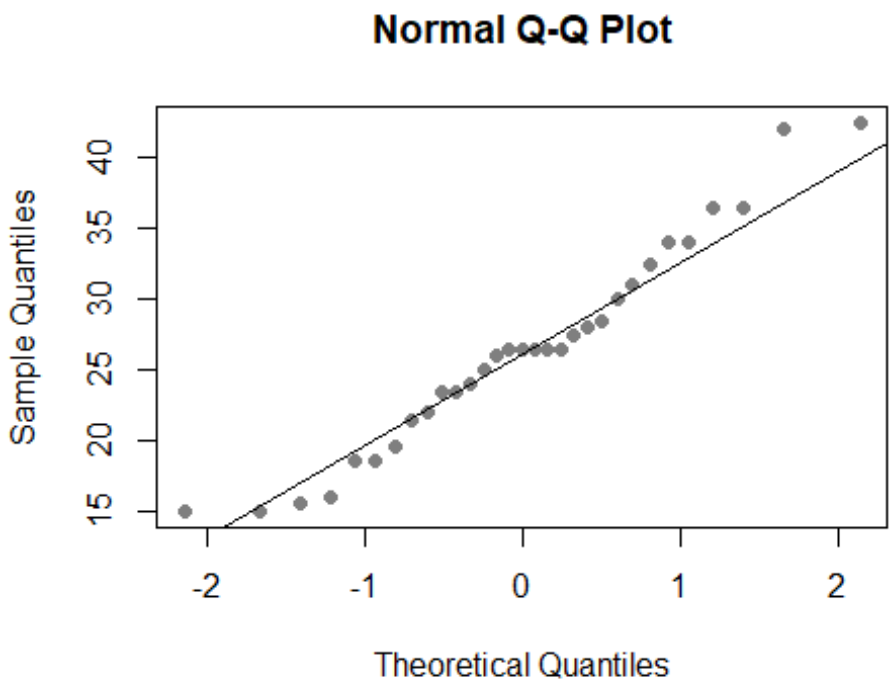
Si Sig. (p-valor) < 0.05 rebutgem H_0 (hipòtesi nul·la) → distribució no normal.

Un altre test que funciona de la mateixa manera és el de **Shapiro-Wilks**, que es fa servir quan les mostres tenen una mida inferior a 50.

```
##
##  Shapiro-Wilk normality test
##
## data:  EE_paisos$electronic_equipment
## W = 0.96193, p-value = 0.3278
```

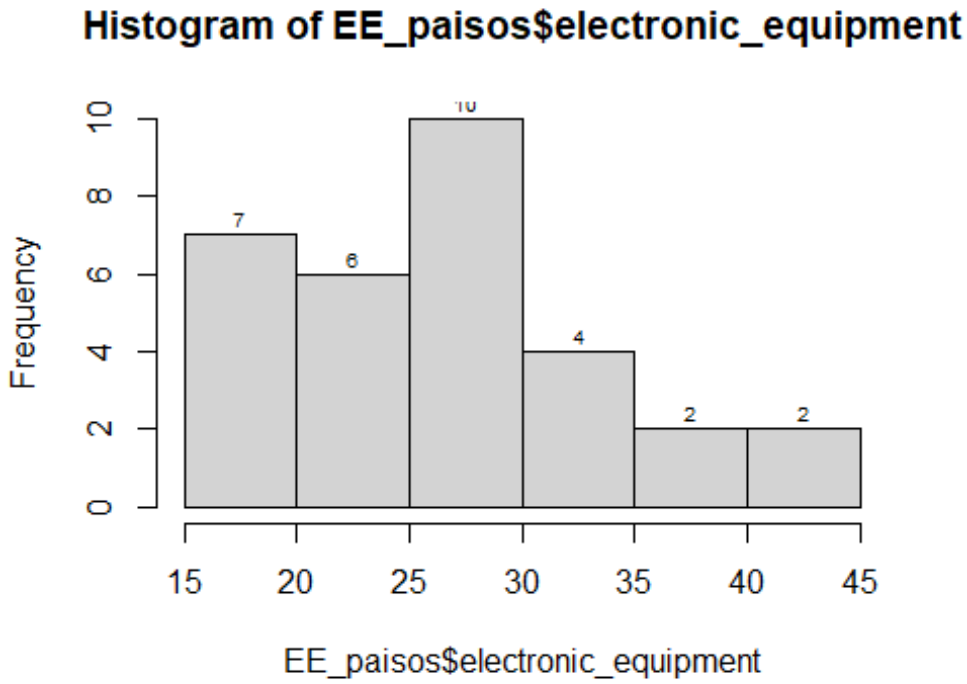
Les dades segueixen una distribució normal.

El gràfic Q-Q normal representa les dades de la variable davant de les dades esperades si la distribució fos normal. Si els punts són a prop de la diagonal podem dir que la distribució és normal.



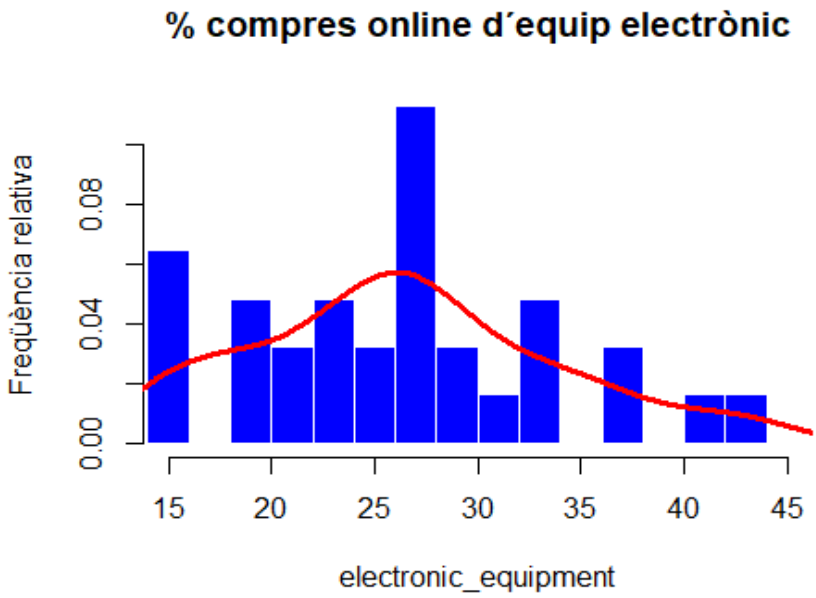
Resum gràfic:

Histograma freqüències absolutes:

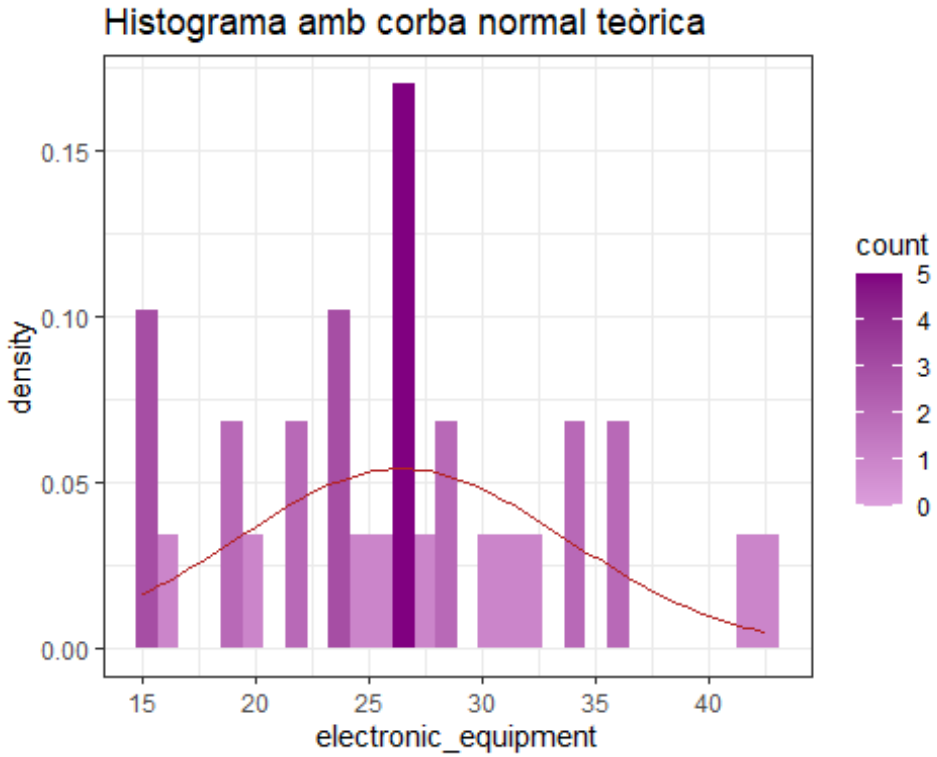


Histograma freqüències relatives

(en tant per u). També hi afegim la corba de densitat:

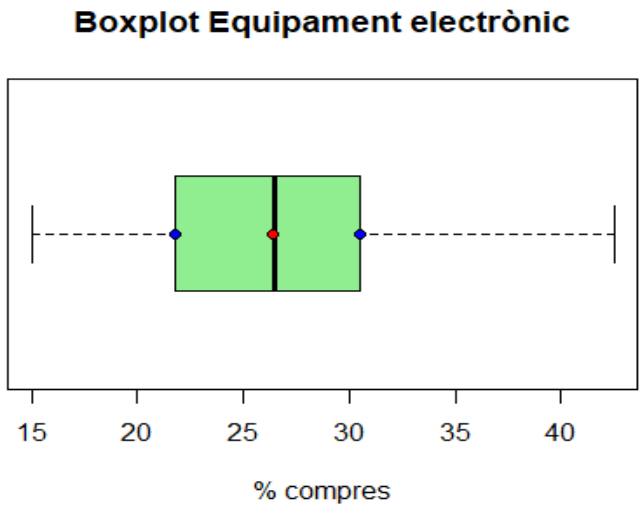


També podem fer servir ggplot2. En aquest cas afegim la corba normal teòrica.



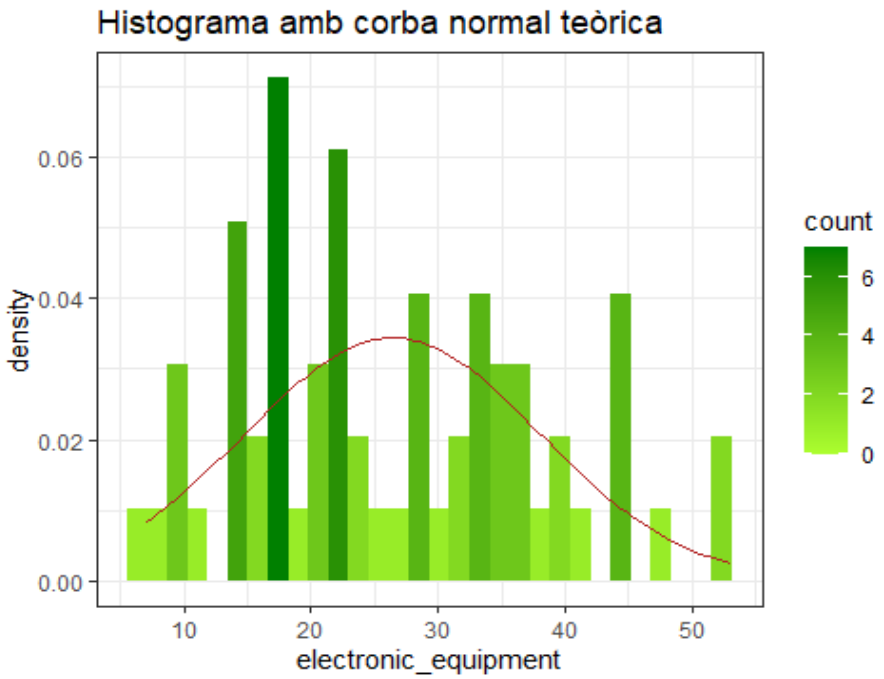
Boxplot:

Els gràfics de Caixa i bigotis s'obtenen a partir de la mitjana. La caixa està definida pel segon i tercer quartil, mentre que els bigotis pel primer i el quart, per la qual cosa dins de la caixa tenim el 50% de les dades de la mostra (mitjana). Aquest tipus de representació també és útil per detectar valors atípics i l´assimetria.



Es pot observar que hi ha assimetria.

Si no haguéssim agrupat les dades, el resultat és molt diferent:



Podem obtenir l´estadística descriptiva i fer una anàlisi de normalitat amb el paquet MVN. Aplicarem el test de normalitat de Shapiro-Wilk:

```
##                        n     Mean  Std.Dev Median Min Max  25th 75th
Skew
## electronic_equipment 62 26.41935 11.58652     24   7  53 17.25   35
0.3549099
##                          Kurtosis
## electronic_equipment -0.7531527

##           Test                          Variable Statistic   p value
Normality
## 4 Shapiro-Wilk       electronic_equipment          0.9688    0.1159
YES
```

Pregunta 2. (25%)

Realitzeu un resum numèric i gràfic per a la variable *sex* i comenteu el resultat.

La variable *sex* és categòrica, no numèrica.

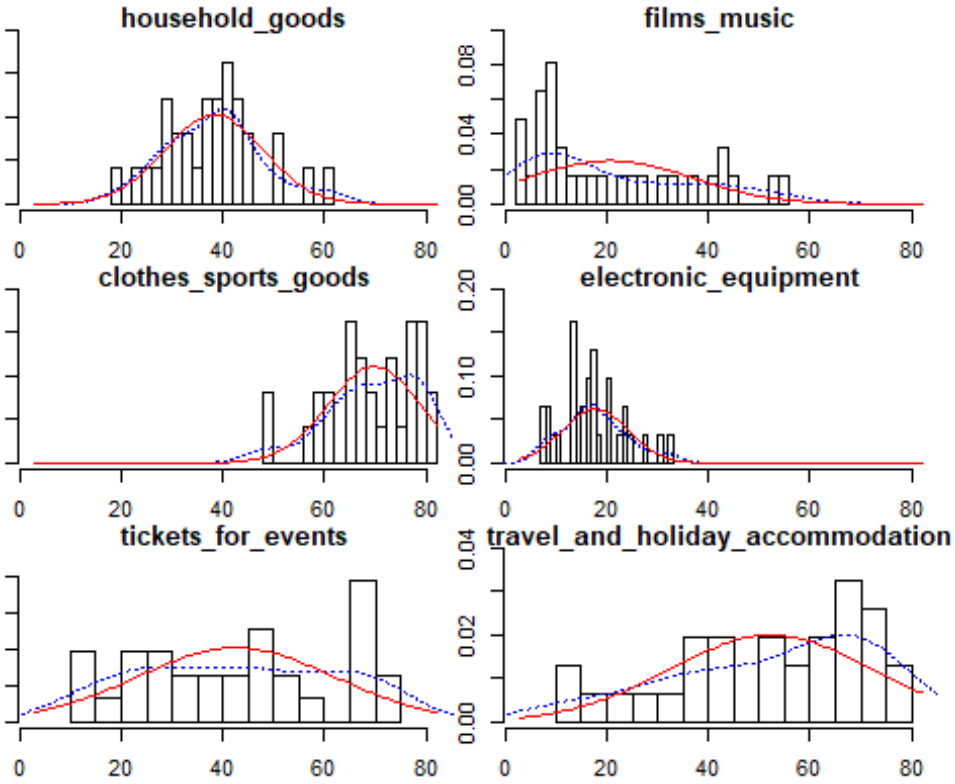
```
## La variable sex és de tipus character . Té una mida de 62

##
## Females   Males
##      31      31
```

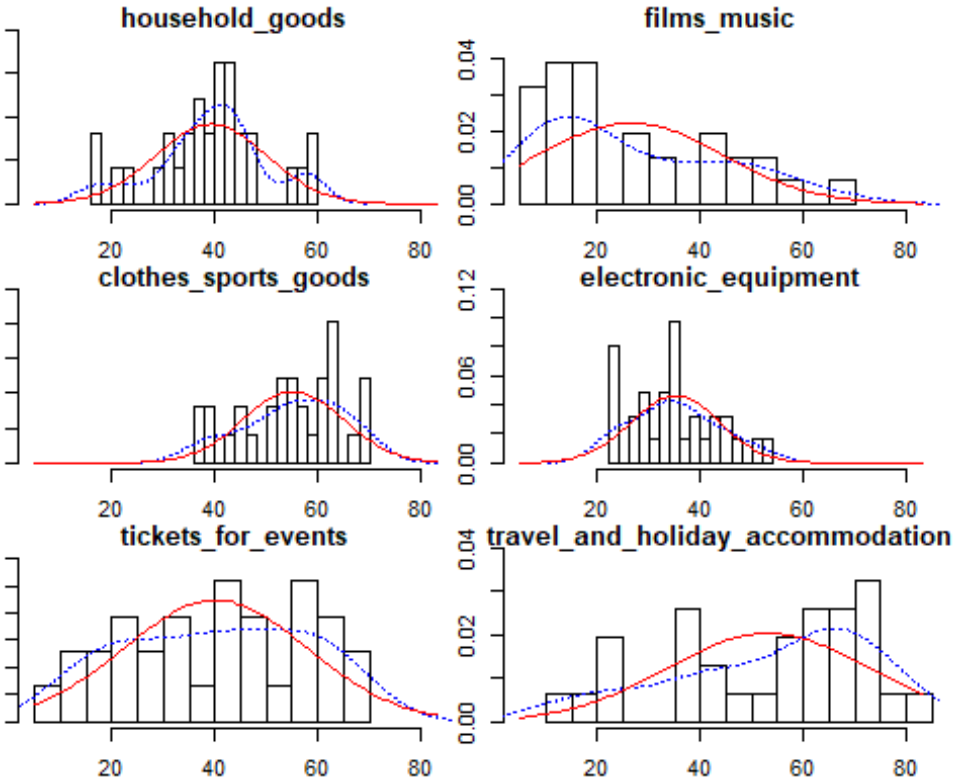
Podem comprovar que, per cada país, hi ha dues entrades de la variable *sex*, una per Females i altra per Males:

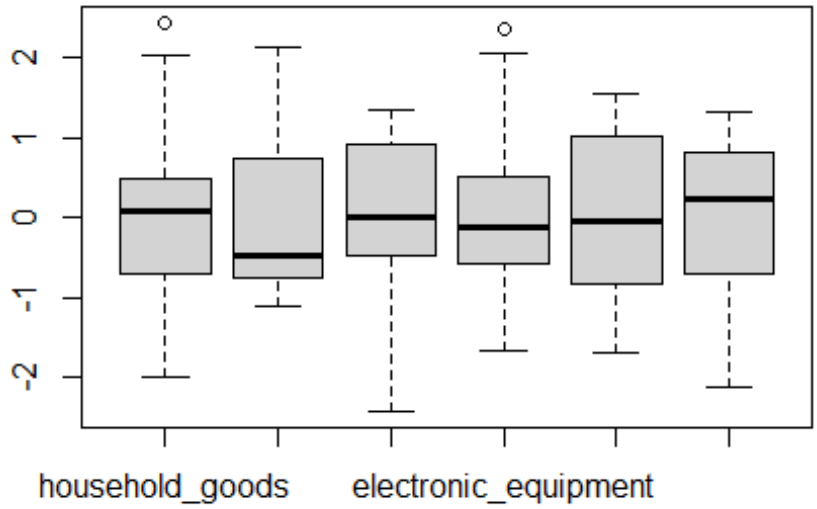
```
##
## TRUE
##   31
```

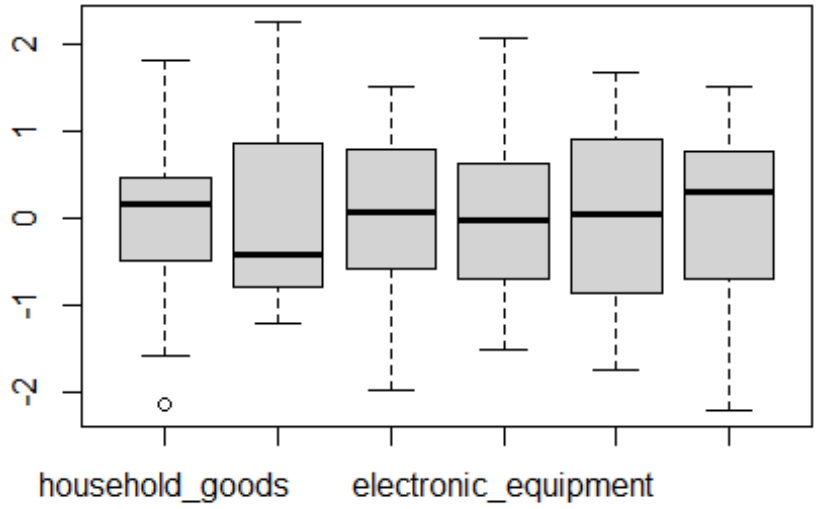
Podem veure com es distribueixen les diferents variables en funció del sexe. Per les dones primer:


household_goods
films_music
clothes_sports_goods
electronic_equipment
tickets_for_events
travel_and_holiday_accommodation

I pels homes:


household_goods
films_music
clothes_sports_goods
electronic_equipment
tickets_for_events
travel_and_holiday_accommodation






```
## $multivariateNormality
## $multivariateNormality$Females
##      Test       H     p value MVN
## 1 Royston 20.7739 0.001313527  NO
```

```
##
## $multivariateNormality$Males
##      Test        H    p value MVN
## 1 Royston 14.70283 0.01750686  NO
##
##
## $univariateNormality
## $univariateNormality$Females
##               Test                          Variable Statistic   p
value
## 1 Anderson-Darling         household_goods             0.2760
0.6336
## 2 Anderson-Darling           films_music               1.5805
0.0004
## 3 Anderson-Darling       clothes_sports_goods          0.6186
0.0979
## 4 Anderson-Darling       electronic_equipment          0.3447
0.4633
## 5 Anderson-Darling        tickets_for_events           0.5579
0.1373
## 6 Anderson-Darling travel_and_holiday_accommodation    0.7414
0.0478
##   Normality
## 1    YES
## 2    NO
## 3    YES
## 4    YES
## 5    YES
## 6    NO
##
## $univariateNormality$Males
##               Test                          Variable Statistic   p
value
## 1 Anderson-Darling         household_goods             0.4593
0.2450
## 2 Anderson-Darling           films_music               1.0725
0.0069
## 3 Anderson-Darling       clothes_sports_goods          0.3803
0.3818
## 4 Anderson-Darling       electronic_equipment          0.2885
0.5938
## 5 Anderson-Darling        tickets_for_events           0.4118
0.3206
## 6 Anderson-Darling travel_and_holiday_accommodation    0.8811
0.0212
##   Normality
## 1    YES
## 2    NO
## 3    YES
## 4    YES
```

```
## 5    YES
## 6    NO
##
##
## $Descriptives
## $Descriptives$Females
##                                    n     Mean   Std.Dev Median Min Max
25th 75th
## household_goods                   31 38.29032  9.723489     39  19  62
31.5 43.0
## films_music                       31 20.70968 16.098020     13   3  55
8.5 32.5
## clothes_sports_goods              31 69.80645  9.027437     70  48  82
65.5 78.0
## electronic_equipment              31 17.70968  6.461133     17   7  33
14.0 21.0
## tickets_for_events                31 42.70968 19.528088     42  10  73
26.5 62.5
## travel_and_holiday_accommodation 31 52.48387 20.051386     57  10  79
38.5 69.0
##                                           Skew   Kurtosis
## household_goods                    0.332358108 -0.1073851
## films_music                        0.728410932 -0.9073540
## clothes_sports_goods              -0.669647464 -0.2998745
## electronic_equipment               0.438021632 -0.3300847
## tickets_for_events                -0.001914759 -1.3665731
## travel_and_holiday_accommodation -0.599342245 -0.8140714
##
## $Descriptives$Males
##                                    n     Mean   Std.Dev Median Min Max
25th 75th
## household_goods                   31 39.35484 10.895101     41  16  59
34.0 44.5
## films_music                       31 26.58065 17.903024     19   5  67
12.5 42.0
## clothes_sports_goods              31 55.32258  9.775436     56  36  70
49.5 63.0
## electronic_equipment              31 35.12903  8.628410     35  22  53
29.0 40.5
## tickets_for_events                31 40.22581 17.813309     41   9  70
25.0 56.5
## travel_and_holiday_accommodation 31 53.32258 19.581262     59  10  83
39.5 68.5
##                                           Skew   Kurtosis
## household_goods                   -0.23139973 -0.1568906
## films_music                        0.61797155 -0.9374219
## clothes_sports_goods              -0.36360599 -0.8994271
## electronic_equipment               0.24753488 -0.7951519
## tickets_for_events                -0.07586825 -1.3078947
## travel_and_holiday_accommodation -0.61675934 -0.8004333
```

Pregunta 3. (25%)

Distingint entre homes i dones, realitzeu un resum numèric i gràfic per a la variable *electronic_equipment*. Compareu això amb el que s'ha observat a l'apartat 1.

Resum numèric per a dones:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   14.00   17.00   17.71   21.00   33.00

## La desviació estàndard i la variància són: 6.461133 i 41.74624
respectivament.
```

i per a homes:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.00   29.00   35.00   35.13   40.50   53.00

## La desviació estàndard i la variància són: 8.62841 i 74.44946
respectivament.
```

Si comparem amb les dades anteriors, observem que són diferents, ja que provenen de fer la mitjana entre homes i dones:

```
##      geo            electronic_equipment
##  Length:31          Min.   :15.00
##  Class :character   1st Qu.:21.75
##  Mode  :character   Median :26.50
##                     Mean   :26.42
##                     3rd Qu.:30.50
##                     Max.   :42.50

## La desviació estàndard i la variància són: 7.368963 i 54.30161
respectivament.
```

Test de normalitat

Dones:

```
##
##  Shapiro-Wilk normality test
##
## data:  dones$electronic_equipment
## W = 0.96607, p-value = 0.4178
```

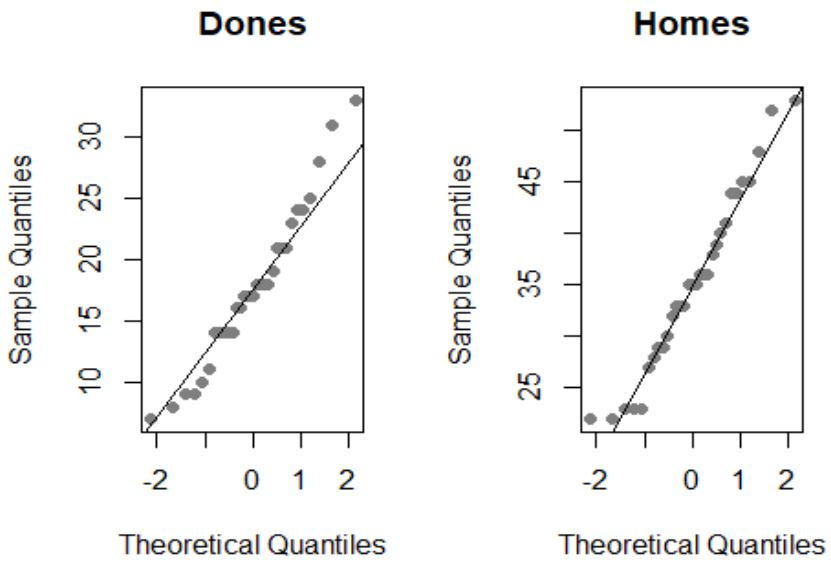
Homes:

```
##
##  Shapiro-Wilk normality test
##
## data:  homes$electronic_equipment
## W = 0.96317, p-value = 0.3529
```

Les dades segueixen una distribució normal.

El gràfic Q-Q normal.

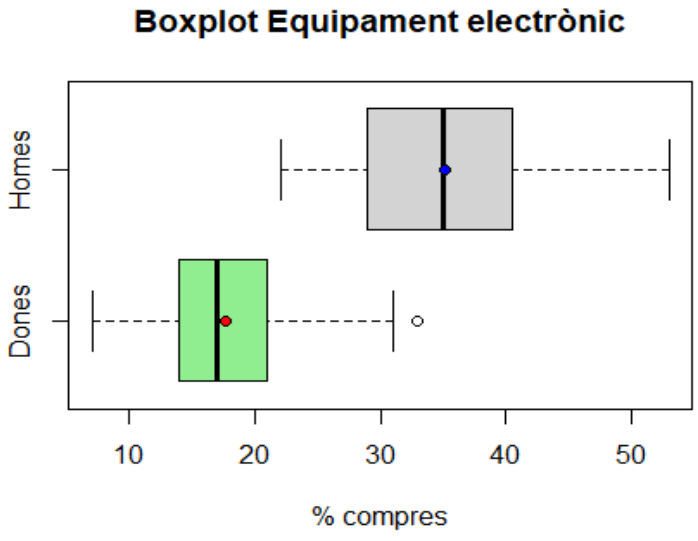
Dones



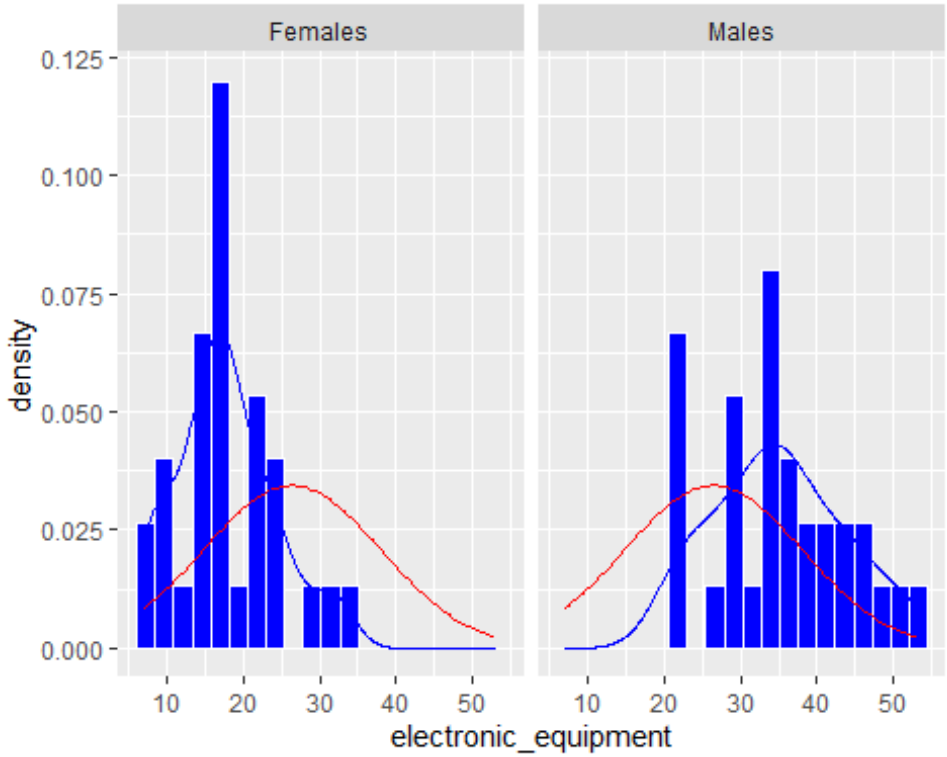
Resum gràfic:

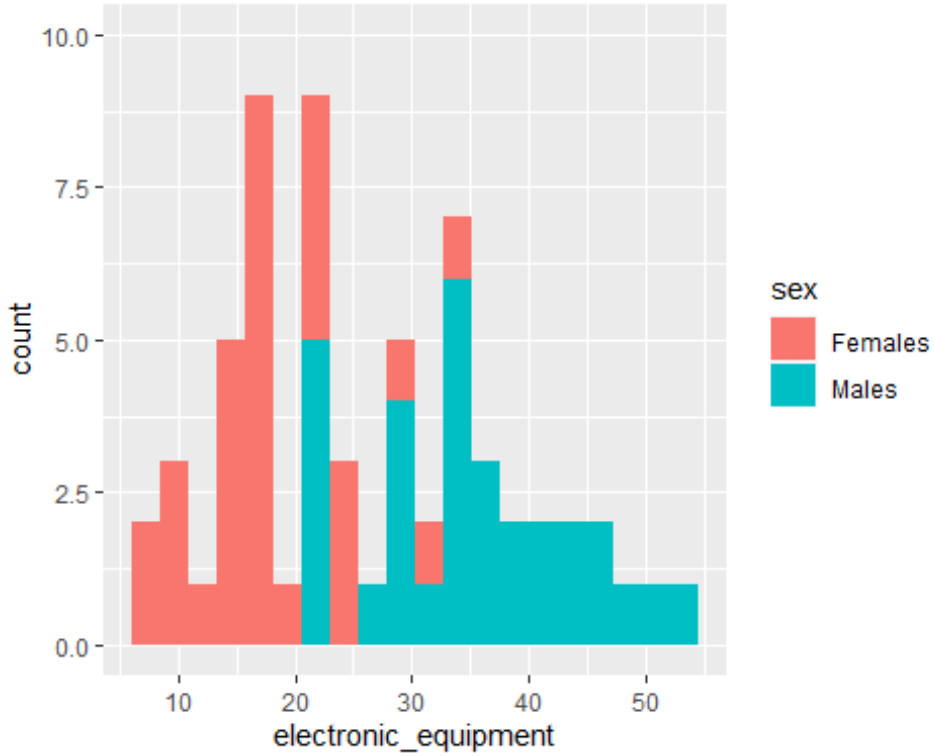
Boxplot

Al boxplot es marquen amb punts les mitjanes, observem que són molt diferents. La mitjana obtinguda al primer apartat és, la mitjana d´aquestes dues. Però sí que són més simètriques.



Histograma


Females
Males
density
0.125
0.100
0.075
0.050
0.025
0.000
10
20
30
40
50
electronic_equipment


count
10.0
7.5
5.0
2.5
0.0
10
20
30
40
50
electronic_equipment
sex
Females
Males

Pregunta 4. (25%)

Trobeu el valor mínim i màxim de la variable *travel_and_holiday_accommodation* i els corresponents països on es dona aquest valor.

```
## El valor mínim de la variable travel_and_holiday_accommodation és 10 , que correspòn als paisos: Croatia Romania

## El valor màxim de la variable travel_and_holiday_accommodation és 83 , que correspòn al país: Switzerland
```