Statistiek Exercises Week 2

Jeroen Knoester

Summary statistics

Exercise 1:

a) Create a vector *x* with 98 random values between 1 and 30 (using the "sample" function).

```
x <- sample(1:30, 98, replace = T)
x
```

```
##  [1] 28  8  4 19 24 27 16 26 25  5 23  3 18 16 23 16 13 11  9 11 15 14 18
## [24] 24  4 30  9 20 16 20  2 14  5 20 28 21 24  5 29 22 14 28  8  8  3  3
## [47]  8  1 24 21 29  1 30  5 25  3  1 21 25  8 19 21  4  5 17 15 22 14 13
## [70]  8 13  9 30 10 14 27 23 25  4 18  9 29  5 20 12 21 20 15 12  2  9  3
## [93] 30  1  4 29 26 24
```

b) Create another vector, *x2*, by adding at the end of *x* two values: 50 and -5.

```
x2 <- c(x, -5, 50)
x2
```

```
##   [1] 28  8  4 19 24 27 16 26 25  5 23  3 18 16 23 16 13 11  9 11 15 14 18
##  [24] 24  4 30  9 20 16 20  2 14  5 20 28 21 24  5 29 22 14 28  8  8  3  3
##  [47]  8  1 24 21 29  1 30  5 25  3  1 21 25  8 19 21  4  5 17 15 22 14 13
##  [70]  8 13  9 30 10 14 27 23 25  4 18  9 29  5 20 12 21 20 15 12  2  9  3
##  [93] 30  1  4 29 26 24 -5 50
```

c) Verify that the median of the two vectors is the same. Why does that happen?

```
median(x) == median(x2)
```

```
## [1] TRUE
```

The median of the 2 vectors is the same, because -5 is less than 16 and 50 is more than 16. The median in between stays the same.

d) Compute the interquartile ranges of both vectors and use them to find out if there are any outliers in the two vectors.

```
IQR(x)
```

```
## [1] 15
```

```
IQR(x2)
```

```
## [1] 15.25
```

```
# Check if there is any number in x that is higher then h or lower then l(the highest and lowest boundr
h <- mean(x) + IQR(x) * 1.5
l <- mean(x) - IQR(x) * 1.5
x[x > h]
```

```
## integer(0)
```

```
x[x < l]
```

```
## integer(0)
```

```r
# Same for x2
h2 <- mean(x2) + IQR(x2) * 1.5
l2 <- mean(x2) - IQR(x2) * 1.5
x2[x2 > h2]
```

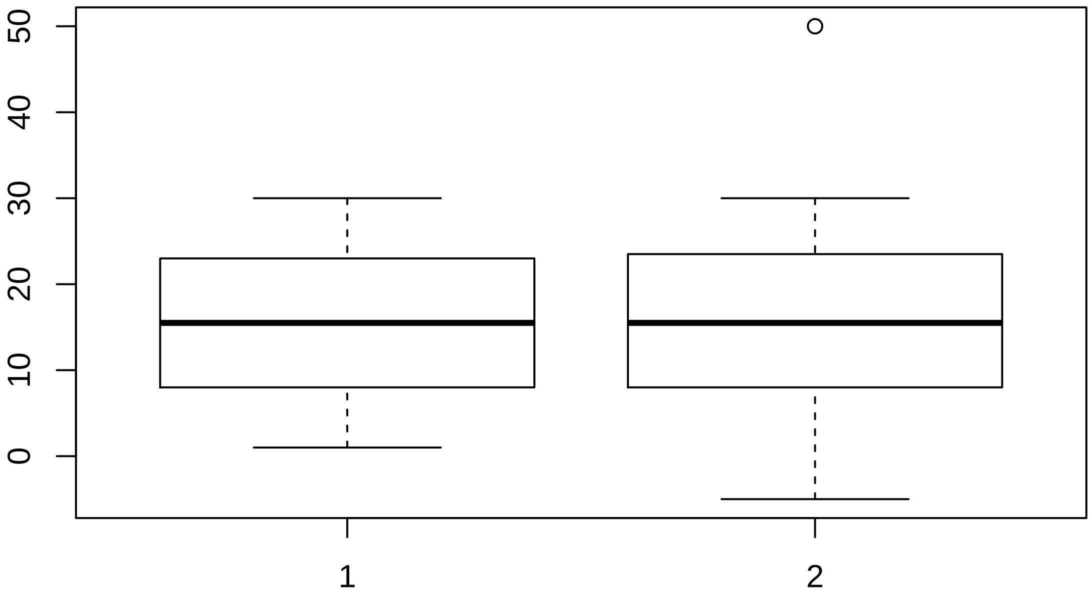
```
## [1] 50
```

```r
x2[x2 < l2]
```

```
## numeric(0)
```

e) Produce a plot containing the two boxplots (together in the same plot) of x and $x2$. Can you see the outliers (if there were any) identified in step d)?

```r
boxplot(x, x2)
```



```r
# Shows the outliner +50 but not the outliner -5
```

f) Compute the standard deviations of both vectors. Which one is smaller and why?

```r
sd(x)
```

```
## [1] 8.957018
```

```r
sd(x2)
```

```
## [1] 9.740242
```

The standard deviation of x is smaller then the sd of $x2$. A low standard deviation is closer to 0. Since $x2$ has outliners, this has a higher standard deviation.

# Visualizations ## Exercise 1 Import the *eba*1977 dataset (from the ISwR package) into *R*. Study the type and content of its variables.

a) Compute the mean of *cases* and produce a boxplot. Are there any outliers? What is the value of the third quartile?

```
library(ISwR)
attach(eba1977)
```

```
## The following objects are masked from eba1977 (pos = 3):
##
##     age, cases, city, pop
```

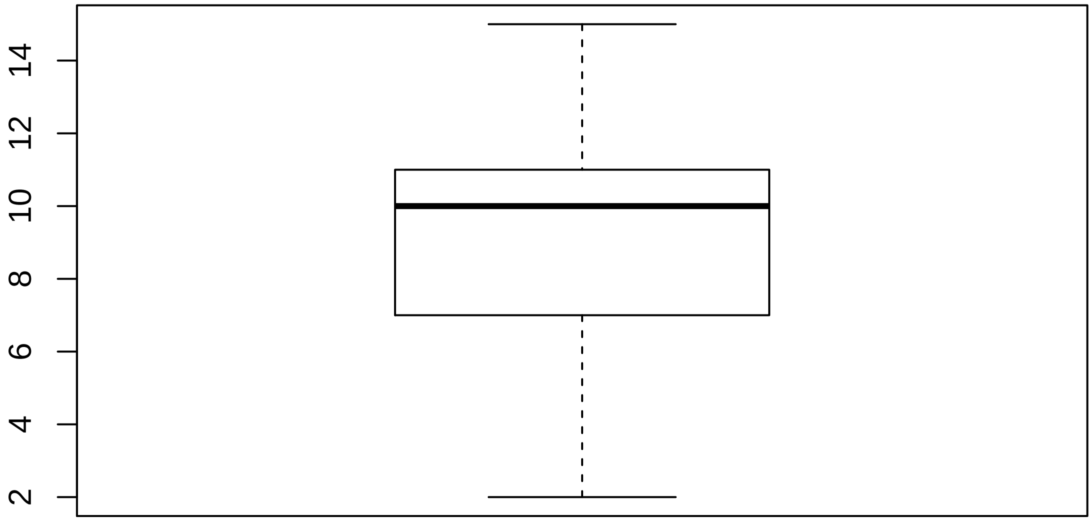
```
## The following objects are masked from eba1977 (pos = 4):
##
##     age, cases, city, pop
```

```
## The following objects are masked from eba1977 (pos = 5):
##
##     age, cases, city, pop
```

```
mean(cases)
```

```
## [1] 9.333333
```

```
boxplot(cases)
```



```
casesSorted <- sort(cases)
(casesSorted[18] + casesSorted[19])/2
```

```
## [1] 11
```

```
eba1977[cases > 17,]
```

```
## [1] city  age   pop   cases
## <0 rows> (or 0-length row.names)
```

```
eba1977[cases< 1, ]
```

```
## [1] city  age   pop   cases
## <0 rows> (or 0-length row.names)
```

There are no outliers, and the value of the thirt quartile = 11 b) Produce the summary of the whole dataset. How many values (rows of the dataset) for each age group were included in the dataset? How many values (rows of the dataset) for each city were included in the dataset?

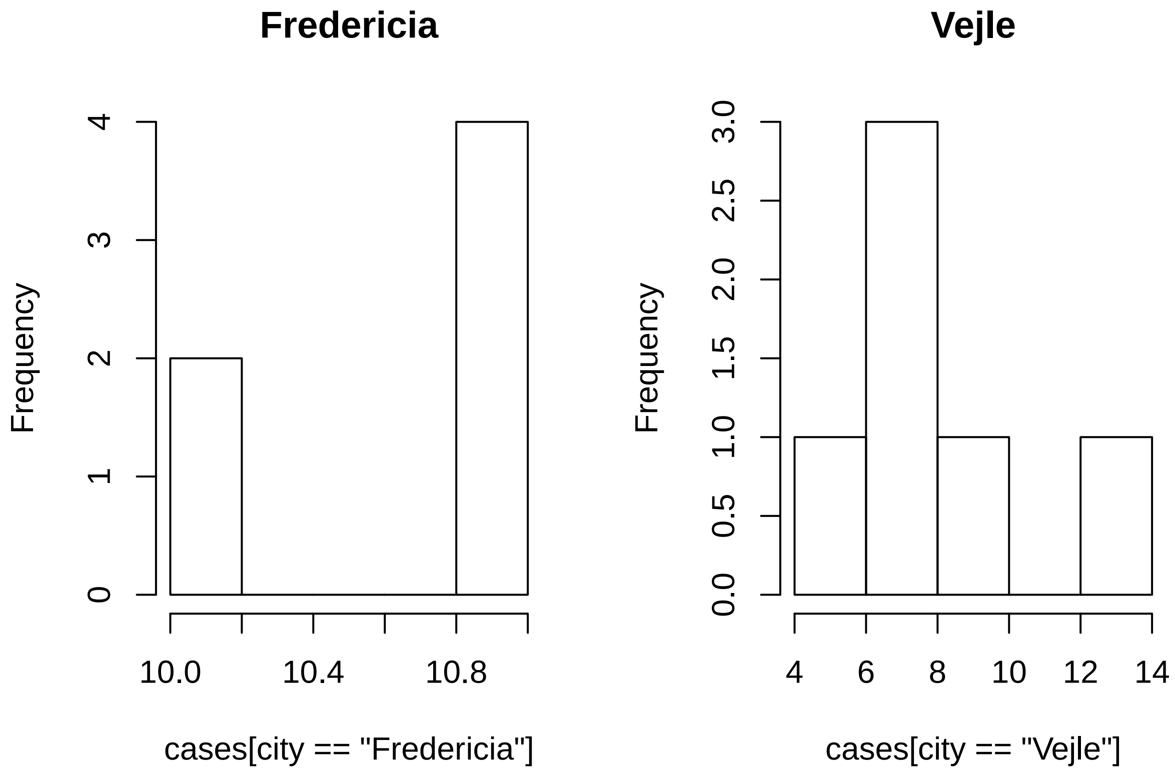
```
summary(eba1977)
```

```
##          city      age          pop             cases       
##  Fredericia:6   40-54:4   Min.   : 509.0   Min.   : 2.000  
##  Horsens   :6   55-59:4   1st Qu.: 628.0   1st Qu.: 7.000  
##  Kolding   :6   60-64:4   Median : 791.0   Median :10.000  
##  Vejle     :6   65-69:4   Mean   :1100.3   Mean   : 9.333  
##                 70-74:4   3rd Qu.: 954.8   3rd Qu.:11.000  
##                 75+  :4   Max.   :3142.0   Max.   :15.000
```

Six rows for each city in the dataset.

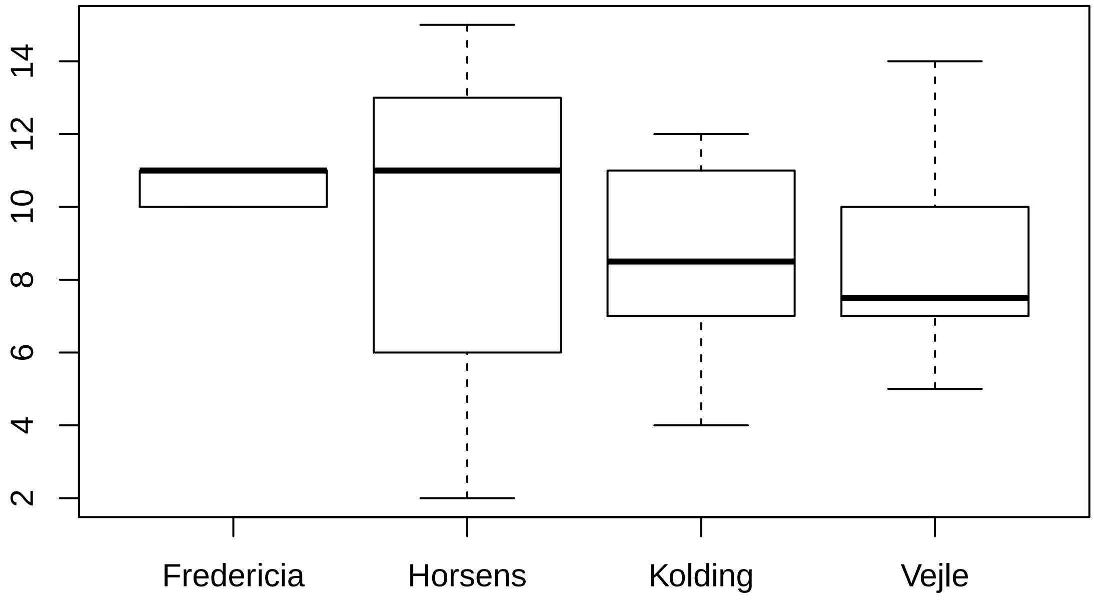
c) Produce two histograms: one with the number of cases in the city of Fredericia and one with the city of Vejle. Put the two histograms in the same plot. What can you say by comparing these two histograms?

```
par(mfrow = c(1:2))
hist(cases[city == "Fredericia"], main = "Fredericia")
hist(cases[city == "Vejle"], main = "Vejle")
```

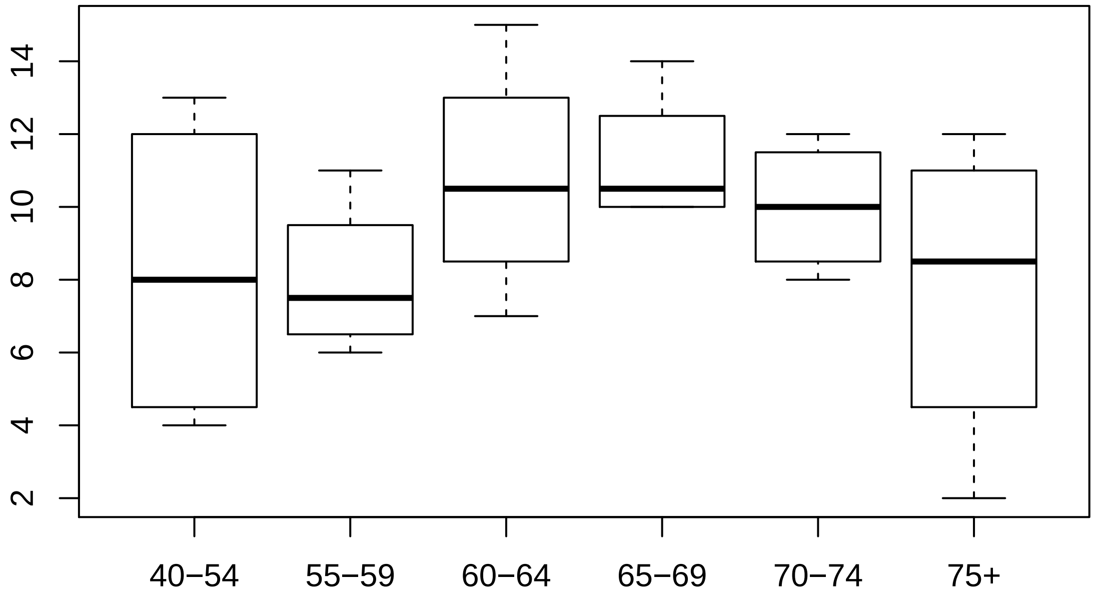


d) Using just one R command, put in a single plot the four boxplots representing the number of cases for each of the four cities. Do the same with the age groups.

```
layout(1)
boxplot(cases ~ city)
```



```r
boxplot(cases~age)
```



e) With just one R command, compute the average of cases for each of the four cities. Do the same with the ages (i.e., compute the average for each of the age groups). Which is the city with the highest average? Which is the age group with the highest average?

```
tapply(cases, city, FUN = mean)
```

```
## Fredericia    Horsens    Kolding      Vejle
##  10.666667   9.666667   8.500000   8.500000
```

```
tapply(cases, age, FUN = mean)
```

```
## 40-54 55-59 60-64 65-69 70-74   75+
##  8.25  8.00 10.75 11.25 10.00  7.75
```

Fredericia has the highest average cases of the 4 cities, the ages between 65-69 has most average cases