Chapter 3

Descriptive measures for univariate distributions

Exercise 6 *Answer* :

- ```
  > incomes<-c(17305,478320,45678,18980,17408,
  + 25676,28906,12500,24540,33450,
  + 12500,33855,37450,20432,28956,
  + 34983,36540,250921,36853,16430,
  + 32654,98213,48980,94024,35671)
  > cat("mean =",mean(incomes),"\ n")
  [1] mean = 60849
  > cat("median =",median(incomes),"\ n")
  [1] median = 33450
  > cat("mode =",names(sort(-table(incomes)))[1],"\ n")
  [1] mode = "12500"
  ```

- Technically, the person was telling the truth because the average is (generally) defined to be the mean. However, of the three measures of central tendency, it seems clear that the median is most representative. The mean is inflated by the two highest salaries.

Exercise 7 *Answer* :

1. For this data, the mean is 4.23, the median is 3, and the mode is 2. Of these, the mode is probably the most representative.

2. For this data, the mean and median are each 5 and the modes are 1 and 9 (the distribution is bimodal). Of these, the mean or median is the most represen- tative.

3. For this data, the mean is 4.59, the median is 5, and the mode is 1. Of these, the mean or median is the most representative.

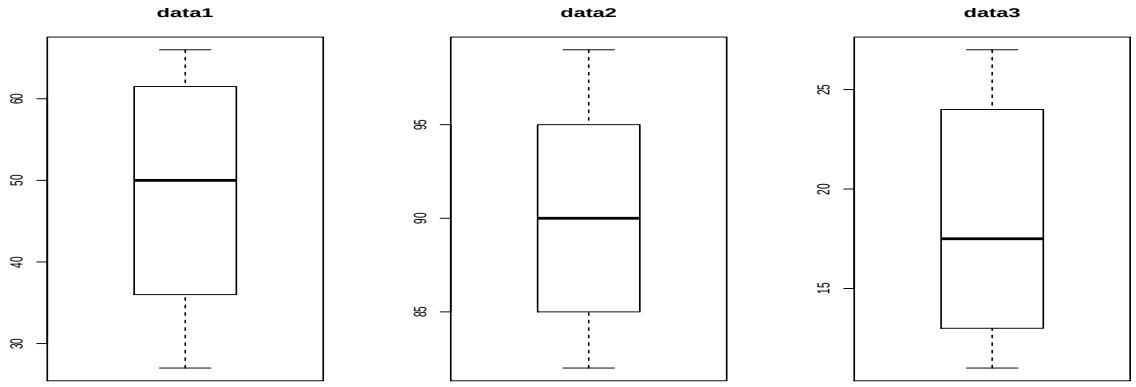
Exercise 8 *Answer* :

```
> quantile(data,c(0.25,0.75),type=2)
  25%   75%
 15.5  24.5
```

The median of the entire set is 19. The median of the six numbers that are less than 19 is 15. So, the lower quartile is 15. The median of the six numbers that are greater than 19 is 25. So, the upper quartile is 25.
(see http://tolstoy.newcastle.edu.au/R/e17/help/att-1067/Quartiles_in_R.pdf)
```

Exercise 9 *Answer* :

```
> data1<-c(27, 28, 30, 42, 45, 50, 50, 61, 62, 64, 66)
> data2<-c(82, 82, 83, 85, 87, 89, 90, 94, 95, 95, 96, 98, 99)
> data3<-c(11, 13, 13, 15, 17, 18, 20, 24, 24, 27)
> par(mfrow=c(1,3))
> boxplot(data1,main="data1")
> boxplot(data2,main="data2")
> boxplot(data3,main="data3")
```



In the plot, notice that five numbers are listed: the smallest number, the lower quartile, the median, the upper quartile, and the largest number. Also notice that the numbers are spaced proportionally, as though they were on a real number line.

Exercise 10 *Answer* :

1. Of the three sets, the numbers in set A are grouped most closely to the center and the numbers in set C are the most dispersed. So, set A has the smallest standard deviation and set C has the largest standard deviation.

2. Because of the symmetry of each bar graph, we can conclude that each has a mean of $\overline{x} = 4$. Next we use R to determine the standard deviation of each set.

```
> tab<-data.frame(data=c(1,2,3,4,5,6,7),nA=c(1,2,3,5,3,2,1),
+ nB=c(2,2,2,2,2,2,2),nC=c(5,4,3,2,3,4,5))
> attach(tab)
> sd(rep(data,nA))
[1] 1.581139
> sd(rep(data,nB))
[1] 2.075498
> sd(rep(data,nC))
[1] 2.262742
```

Exercise 11 *Answer* :

1.
```
> students<-data.frame(Names=c("Leonie","Luka","Lea","Leon","Laura","Luis"),
+ Degree=c("Master","Master","Bachelor","Bachelor","Bachelor","Bachelor"),
+ mat.nr=c(1111,1112,1113,1114,1115,1116))
```

```
> students
     Names    Degree  mat.nr
 1  Leonie    Master    1111
 2    Luka    Master    1112
 3     Lea  Bachelor    1113
 4    Leon  Bachelor    1114
 5   Laura  Bachelor    1115
 6    Luis  Bachelor    1116
> names(students)
[1] "Names" "Degree" "mat.nr"
> str(students)
'data.frame':  6 obs.  of 3 variables:  $ Names :  Factor w/ 6 levels "Laura","Lea",..:
4 6 2 3 1 5
$ Degree:  Factor w/ 2 levels "Bachelor","Master":  2 2 1 1 1 1
$ mat.nr:  num 1111 1112 1113 1114 1115 ...
> summary(students)
    Names      Degree        mat.nr
 Laura :1  Bachelor:4   Min.  :1111
   Lea :1   Master :2  1st Qu.:1112
  Leon :1              Median :1114
 Leonie:1                Mean :1114
  Luis :1              3rd Qu.:1115
  Luka :1               Max.  :1116
```

2. ```
   > students$mat.nr[5]
   [1] 1115
   ```

3. ```
   > Degree
   Error :  object 'Degree' not found
   > attach(students)
   > Degree
   [1] Master Master Bachelor Bachelor Bachelor Bachelor
   Levels:  Bachelor Master
   ```

4. ```
   > ba.students<-subset(students,Degree=="Bachelor",select=-Degree)
   > ba.students
      Names  mat.nr
   3    Lea    1113
   4   Leon    1114
   5  Laura    1115
   6   Luis    1116
   ```

5. ```
   > write.table(students,"studentsfile.txt",sep="\t")
   > read.table("studentsfile.txt",sep="\t",header=T)->students2
   > all(students==students2)
   [1] TRUE
   ```

Exercise 12 *Answer* :

1. We apply the function kurtosis from the "e1071 package" to compute the kurtosis of eruptions. As the package is not in the core `R` library, it has to be installed and loaded into the R workspace.

   ```
   > library(e1071)
   > duration = faithful$eruptions
   ```

```
> kurtosis(duration)
[1] -1.5116
```

2.
```
> attach(faithful)
> kurtosis(waiting)
[1] -1.156263
```

Exercise 13 *Answer* :

```
> install.packages("ineq")
> library(ineq)
> data(AirPassengers)
> ineq(AirPassengers,type="Gini")
> plot(Lc(AirPassengers)) # or
> plot(Lc(AirPassengers),col="darkred",lwd=2)
```