Exercises 2

Introduction to R for Statistics

Jonas-Mika Senghaas

10/02/2021

Exercise 1

The first exercise requires us to import the 'starwars' dataset from the 'dplyr' package. It contains 87 characters (rows) that are described by 13 features (columns). However, there are some missing values. Let's first load the dataset and inspect it.

```
# load the package dpylr, the starwars package is now globally loaded
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# we can get an overview by running head() and summary() on the dataset
# head(starwars)
# summary(starwars)
```

1a. What is the Homeworld of 'Mace Windu'?

To find the homeworld of Mace Windu, we first need to select the row containing Mace Windu. We do this by subsetting. Once we have the row, we can select the column holding the information of his homeworld using the $ operator.

```
mace_windu <- subset(starwars, name=='Mace Windu')
mace_windu$homeworld
```

```
## [1] "Haruun Kal"
```

1b. How many droids are in the dataset?

The number of droids in the dataset is the length of the subset, fitered for all droids.

```
droids <- subset(starwars, species=='Droid')
print(nrow(droids))
```

```
## [1] 6
```

1c. Who are the shortest and tallest humans in the dataset?

We, again fiter for all humans and find the min and max heights. Then we map those to the corresponding names.

```
humans_heights <- subset(starwars, species=='Human')$height
max_height <- max(humans_heights, na.rm = TRUE)
min_height <- min(humans_heights, na.rm = TRUE)

subset(starwars, height==max_height)$name
```

```
## [1] "Darth Vader"
```

```
subset(starwars, height==min_height)$name
```

```
## [1] "Leia Organa" "Mon Mothma"
```

1d. What is the mean and standard deviation of the height of all humans?

```
print(paste('Mean Height of Humans', mean(humans_heights, na.rm=TRUE)))
```

```
## [1] "Mean Height of Humans 176.645161290323"
```

```
print(paste('Standard Deviation of Heights of Humans', sd(humans_heights, na.rm=TRUE)))
```

```
## [1] "Standard Deviation of Heights of Humans 12.5367417008216"
```

Exercise 2

2.1 Create the Dataframe

Now, we should create our own dataframe. We therefore first create vectors corresponding to the columns.

```
name <- c('Astrid', 'Lea', 'Sarina', 'Remon', 'Letizia', 'Babice', 'Jonas', 'Wendy', 'Niveditha', 'Gioi
sex <- c('F', 'F', 'F','M', 'F', 'F', 'M', 'F', 'F', 'F')
age <- c(30, 25, 25, 29, 22, 22, 35, 19, 32, 21)
superhero <- c('Batman', 'Superman', 'Batman', 'Spiderman', 'Batman', 'Antman', 'Batman', 'Superman', '
tattoos <- c(11, 15, 12, 5, 65, 3, 9, 13, 900, 0)

data = data.frame(name, sex, age, superhero, tattoos)
data
```

```
##         name sex age superhero tattoos
## 1     Astrid   F  30    Batman      11
## 2        Lea   F  25  Superman      15
## 3     Sarina   F  25    Batman      12
## 4      Remon   M  29 Spiderman       5
## 5    Letizia   F  22    Batman      65
## 6     Babice   F  22    Antman       3
## 7      Jonas   M  35    Batman       9
## 8      Wendy   F  19  Superman      13
## 9  Niveditha   F  32   Maggott     900
## 10     Gioia   F  21  Superman       0
```

2.2 What was the mean age of female and male pirates separately?

```
males <- subset(data, sex=='M')
females <- subset(data, sex=='F')
```

```
mean(males$age)
```

```
## [1] 32
```

```
mean(females$age)
```

```
## [1] 24.5
```

2.3

Adding a new column

```
tattoos.per.year <- data$tattoos / data$age
data['tattoos.per.year'] <- tattoos.per.year
data
```

```
##         name sex age superhero tattoos tattoos.per.year
## 1     Astrid   F  30    Batman      11        0.3666667
## 2        Lea   F  25  Superman      15        0.6000000
## 3     Sarina   F  25    Batman      12        0.4800000
## 4      Remon   M  29 Spiderman       5        0.1724138
## 5    Letizia   F  22    Batman      65        2.9545455
## 6     Babice   F  22    Antman       3        0.1363636
## 7      Jonas   M  35    Batman       9        0.2571429
## 8      Wendy   F  19  Superman      13        0.6842105
## 9  Niveditha   F  32   Maggott     900       28.1250000
## 10     Gioia   F  21  Superman       0        0.0000000
```

Exercise 2d - What was the median number of tattoos of pirates over the age of 20 whose favorite superhero is Spiderman?

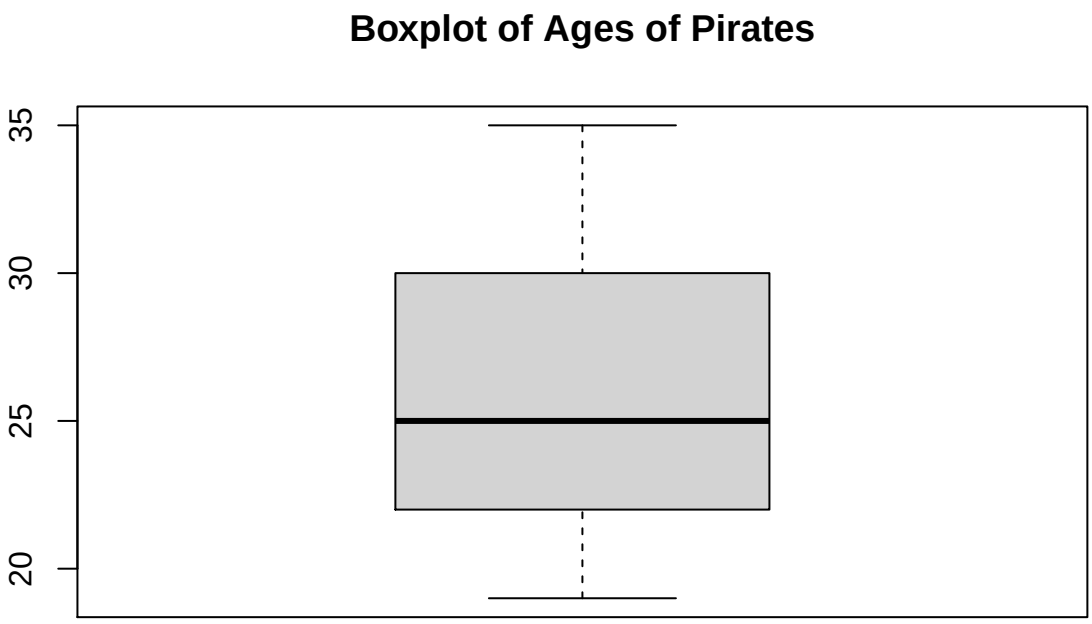
We first filter for all pirate over age 20, whose favorite superhero is Spiderman

```
over.20.and.spiderman <- subset(data, superhero=='Spiderman' & age>20)
mean(over.20.and.spiderman$tattoos)
```

```
## [1] 5
```

Exercise 2e - Make a boxplot of the age distribution of the pirates

```
ages <- data$age
boxplot(ages,
        main='Boxplot of Ages of Pirates')
```

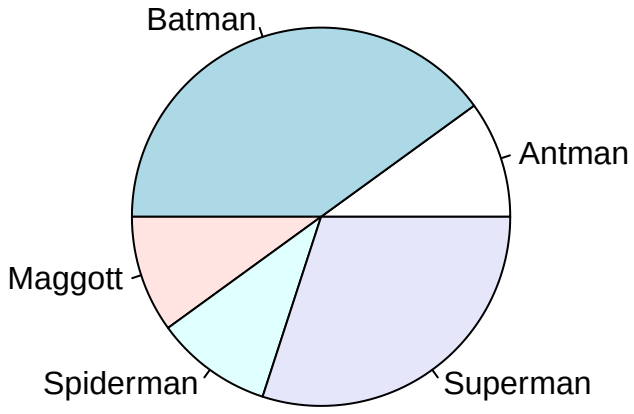


Exercise 2f. Make a piechart showing the number of pirates which has each superhero as their favorite.

```
table(data$superhero)
```

```
##
##    Antman    Batman   Maggott Spiderman  Superman
##         1         4         1         1         3
```

```
pie(table(data$superhero),
    main='Distribution of Favorite Superheros')
```



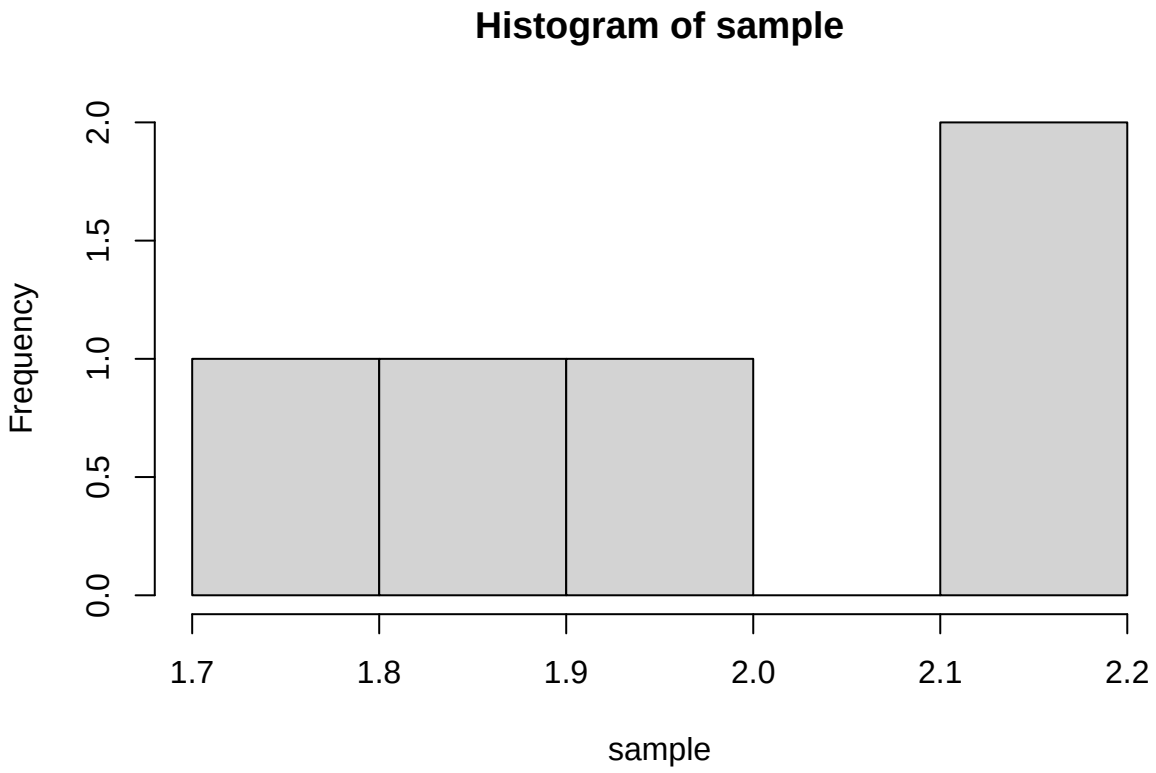
Exercise 3

3a. Normal Distributions

```
sample <- c(rnorm(5, mean=2, sd=1/5))
```

3b. Histogram of Sample

```
hist(sample)
```

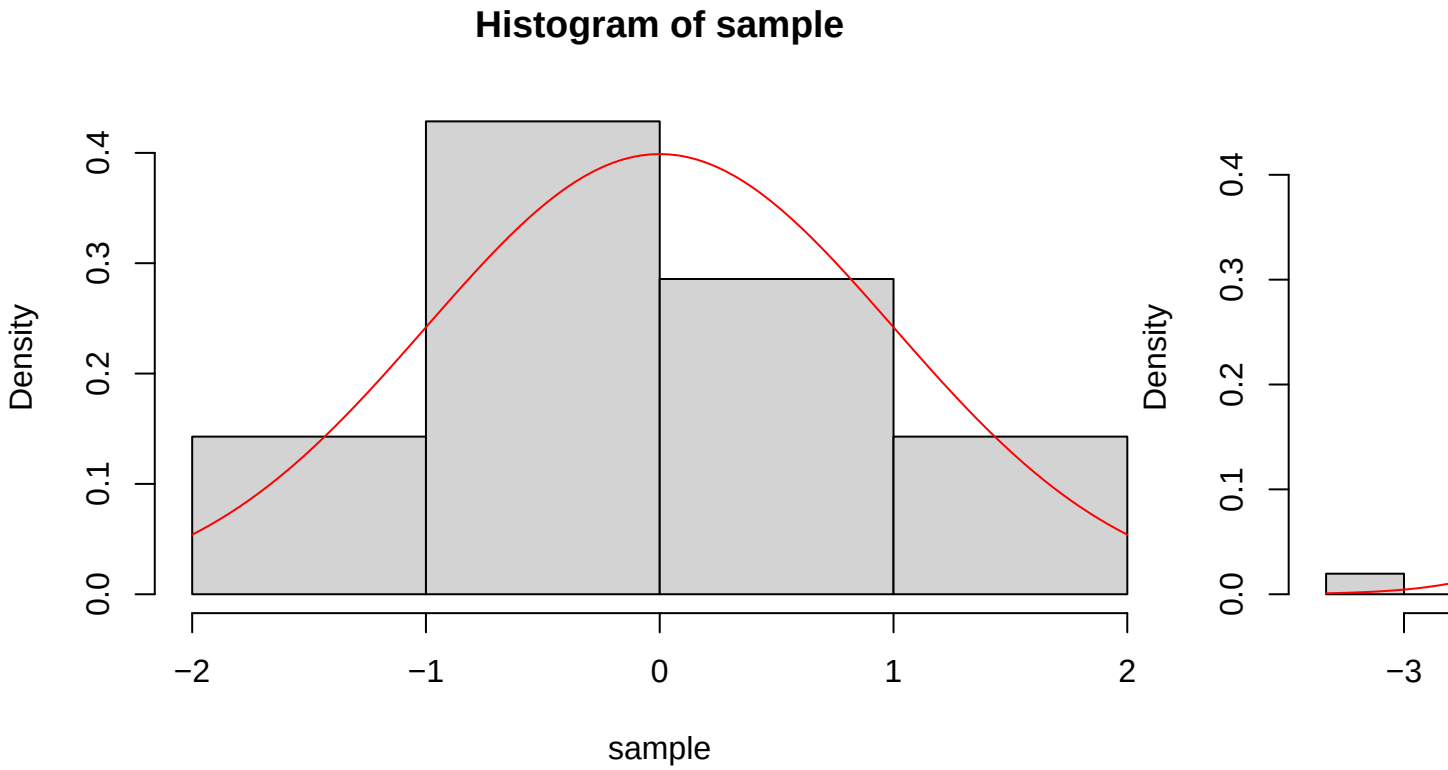


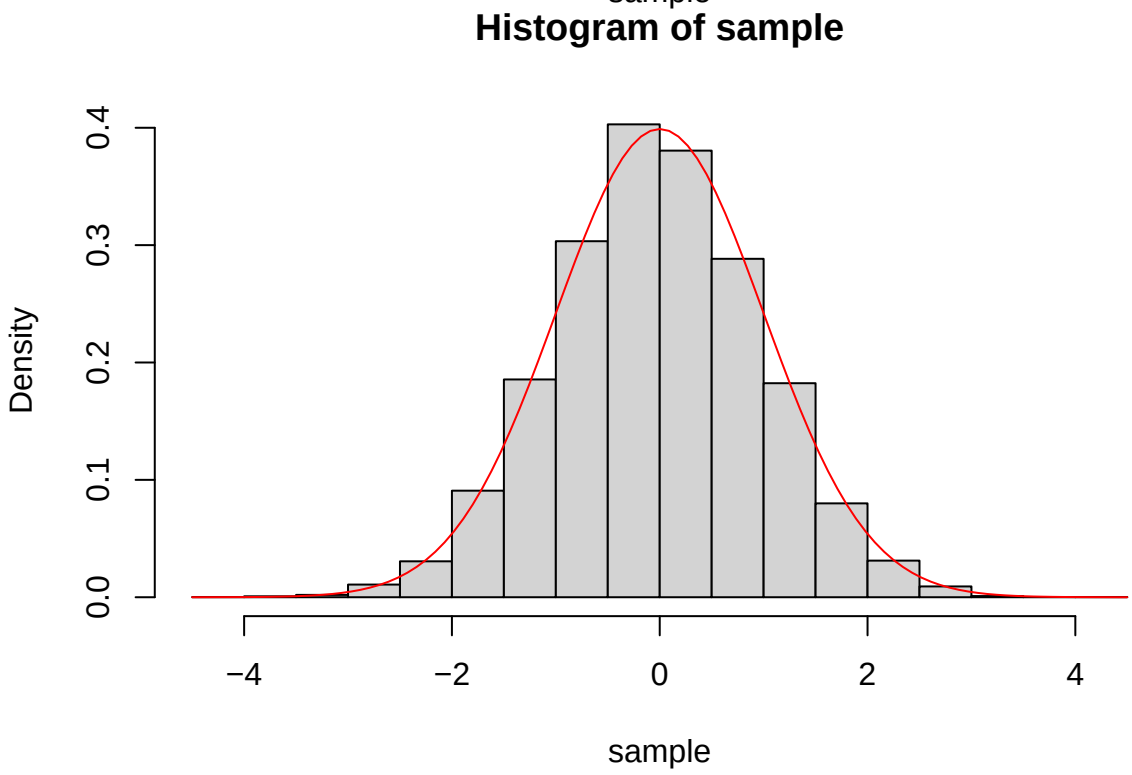
3c. What happens to the mean and standard deviation when you increase the number of samples to 100, how about 10000?

```
plot.normal.distribution <- function(sample.size, with_pdf=TRUE) {
        sample <- c(rnorm(sample.size), mean=2, sd=1/5)
        hist(sample, freq=FALSE)

        if (with_pdf == TRUE) {
                curve(dnorm(x,0,1), add=TRUE, col='red')
        }
}
```

```
for (i in c(5, 100, 10000)) {
 plot.normal.distribution(i)
}
```


Histogram of sample
Density
0.0
0.1
0.2
0.3
0.4
−2
−1
0
1
2
sample


Histogram of sample
Density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
sample