Problems

1. Load the `starwars` dataset from the `dplyr` package. The dataset has 87 characters from the StarWars universe with 13 features. **Note** there are missing data in the database represented by `NA` values.

a) What is the homeworld of Mace Windu?

```
require(dplyr)
subset(starwars, name == 'Mace Windu')$homeworld
```

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

b) How many droids are in the dataset?

```
nrow(subset(starwars, species=='Droid'))
```

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

c) Who are the shortest and tallest humans in dataset?

```
humans <- subset(starwars, species=='Human')
subset(humans, height == min(height, na.rm = TRUE))$name
```

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

```
subset(humans, height == max(height, na.rm = TRUE))$name
```

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

d) What is the mean and standard deviation of the height all humans in the starwars database?

```
mean(subset(starwars, species=='Human')$height, na.rm = TRUE)
```

```
## [1] 176.6452
```

```
sd(subset(starwars, species=='Human')$height, na.rm = TRUE)
```

```
## [1] 12.53674
```

2. The following table shows the results of a survey of 10 pirates. In addition to some basic demographic information, the survey asked each pirate "What is your favorite superhero?"" and "How many tattoos do you have?""

a) Combine the data into a single dataframe. Complete all the following exercises from the dataframe!

	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

```
Name <- c('Astrid', 'Lea', 'Sarina', 'Remon', 'Letizia',
          'Babice', 'Jonas', 'Wendy', 'Niveditha', 'Gioia')

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', 'Maggott', 'Superman')

Tattoos <- c(11, 15, 12, 5, 65, 3, 9, 13, 900, 0)

Pirates <- data.frame(Name, Sex, Age, Superhero, Tattoos)
```

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

```
mean(subset(Pirates, Sex=='F')$Age)
```

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

```
mean(subset(Pirates, Sex=='M')$Age)
```

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

c) Add a new column to the dataframe called tattoos.per.year which shows how many tattoos each pirate has for each year in their life. Which pirate had the most number of tattoos per year?

```
Pirates$tattoos.per.year <- Pirates$Tattoos / Pirates$Age
subset(Pirates, tattoos.per.year == max(Pirates$tattoos.per.year))$Name
```

```
## [1] "Niveditha"
```

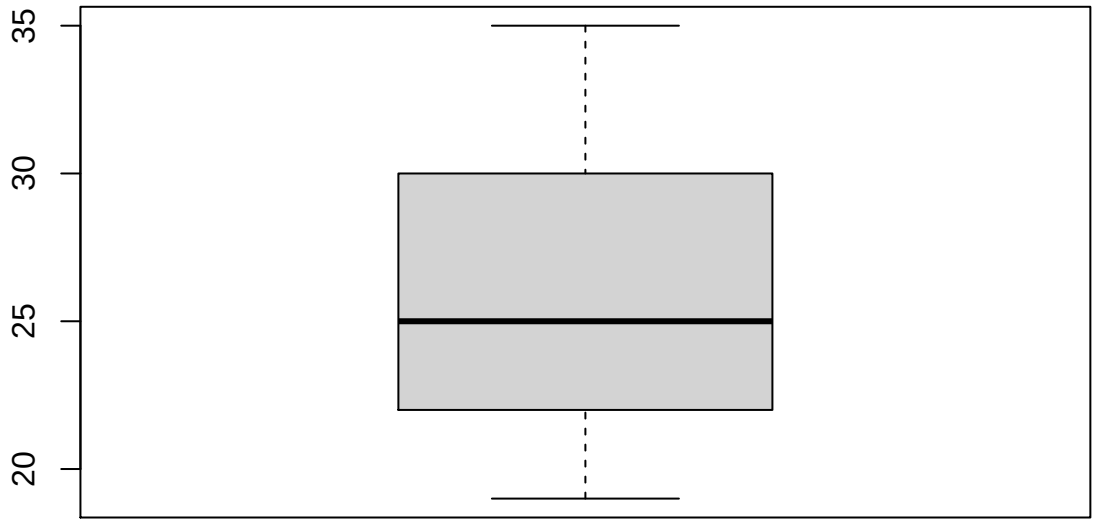
d) What was the median number of tattoos of pirates over the age of 20 whose favorite superhero is Spiderman?

```
median(subset(Pirates, Age>20 & Superhero=='Spiderman')$Tattoos)
```

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

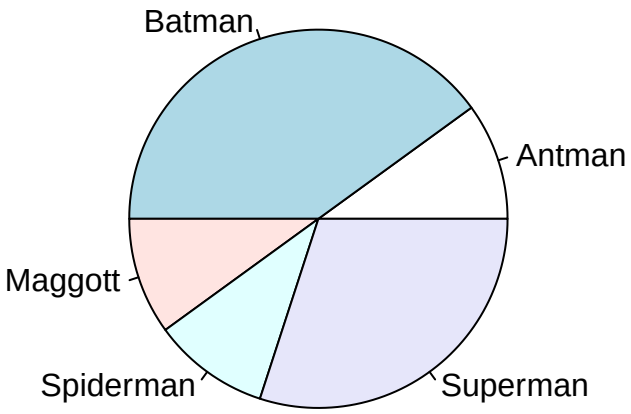
e) Make a boxplot of the age distribution of the pirates

```
boxplot(Pirates$Age)
```



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

```
pie(table(Pirates$Superhero))
```



3. Sample 5 random numbers from the normal (Gaussian) distribution with a mean of 2 and a standard deviation of 1/5. (**Hint** look up the help file using `?rnorm`)

a) Calculate the mean and standard deviation of the generated samples.

```
samples <- rnorm(5, 2, 1/5)
mean(samples)
```

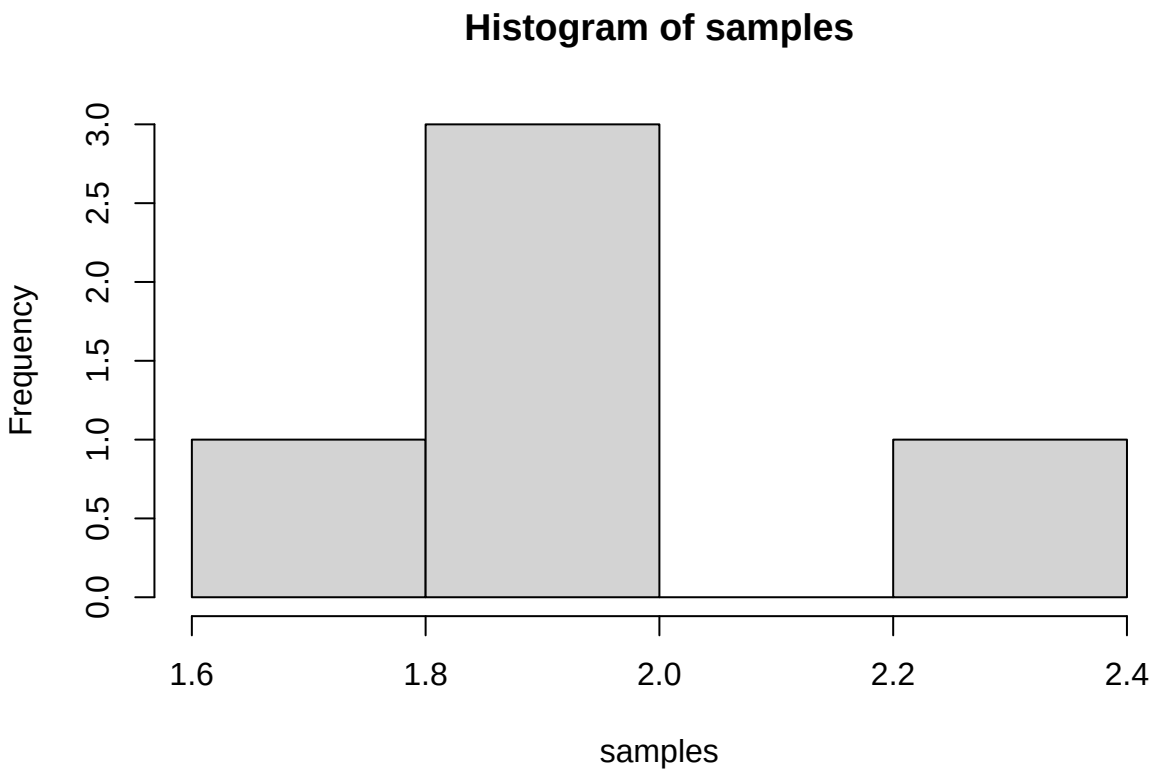
```
## [1] 1.972555
```

```
sd(samples)
```

```
## [1] 0.2257559
```

b) Make a histogram of the generated samples.

```
hist(samples)
```



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

```
sample <- rnorm(100, 2, 1/5)
mean(sample)
```

```
## [1] 1.975072
```

```
sd(sample)
```

```
## [1] 0.2112159
```

```
sample <- rnorm(1000, 2, 1/5)
mean(sample)
```

```
## [1] 2.010981
```

```
sd(sample)
```

```
## [1] 0.193689
```

d) **Optional** Add the theoretical distribution to the plot using the `lines` function.

(**Hint** First define a suitable interval in a `vector` then get the corresponding probabilities with the `dnorm` function. You need to use the `freq = FALSE` argument in the hist function to produce a normalized histogram.)

```
hist(sample, freq = FALSE)

xs <- seq(1, 3, by=0.01)
probs <- dnorm(xs, 2, 1/5)
lines(xs, probs, col='red')
```