Applied Statistics – Exercise 2

Problems

1.

a) Compute $\frac{74+5}{(2 \cdot 3)^5}$ and assign the name `calculation` to the result. Print `calculation` to the console.

```
calculation <- (74+5) / (2*3)^5
print(calculation)
```

```
## [1] 0.01015947
```

b) Define a vector months containing the numbers 29, 63, 7, 23, 84, 10 and 9. Compute a vector years from it by dividing months by 12.

```
months <- c(29, 63, 7, 23, 84, 10, 9)
years <- months/12
```

c) Check whether the string "R rules!" is equal to "r rules!" for R.

```
"R rules!" == "r rules!"
```

```
## [1] FALSE
```

d) In a fictitious medical study patients should be excluded from the study if they weigh more than 90 kg or if they are either younger than 18 years or older than 60 years. Define the variable age as `age <- c(50, 17, 39, 27, 90)` and the variable weight as `weight <- c(80, 75, 92, 105, 60)`. Then write a logical statement involving these two variables that tests for the exclusion criteria.

```
age <- c(50, 17, 39, 27, 90)
weight <- c(80, 75, 92, 105, 60)

age < 18 | age > 60 | weight > 90
```

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

2. The data set `rivers` contains the lengths of 141 major rivers in North America.

a) What proportion are less than 500 miles long?

```
sum(rivers<500)/length(rivers)
```

```
## [1] 0.5815603
```

b) What proportion are less than the mean length?

```
sum(rivers<mean(rivers)) / length(rivers)
```

```
## [1] 0.6666667
```

c) What is the 0.75 quantile?

```
quantile(rivers)[4]
```

```
## 75%
## 680
```

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`)

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

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

```
mean(x)
```

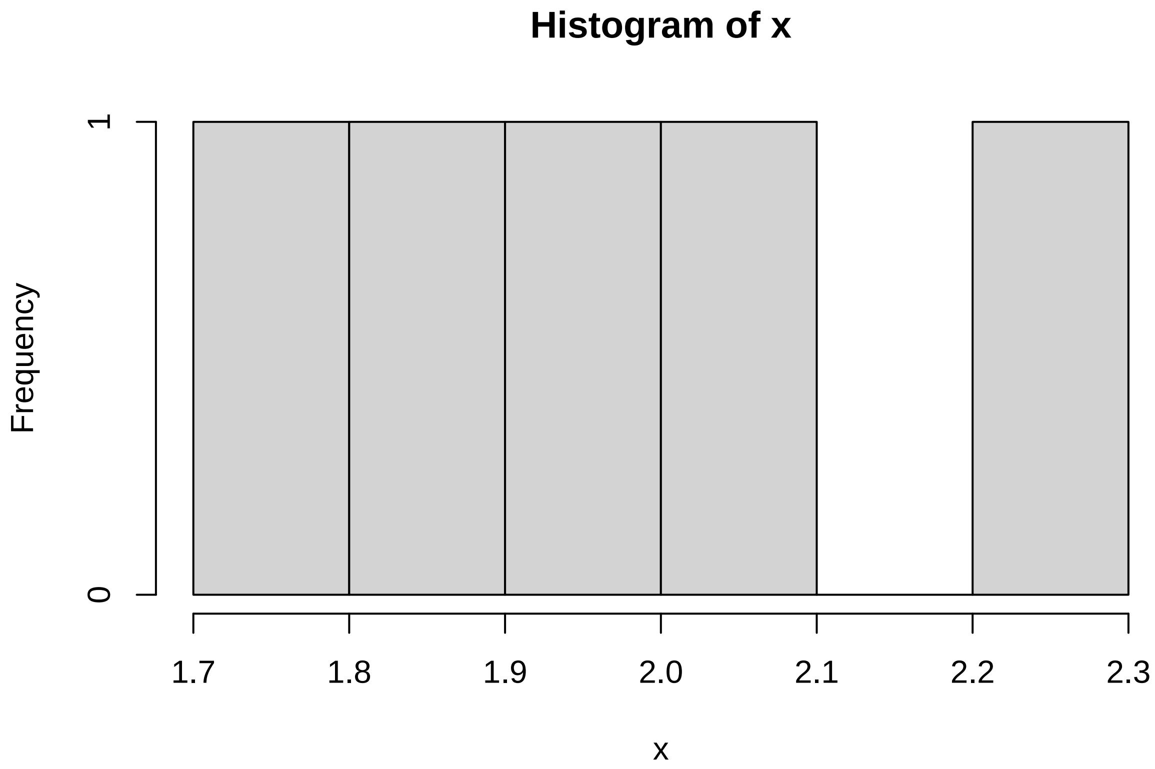
```
## [1] 1.964101
```

```
sd(x)
```

```
## [1] 0.1865696
```

b) Make a histogram of the generated samples.

```
hist(x)
```



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

```
y <- rnorm(100, mean = 2, sd = 1/5)
mean(y)
```

```
## [1] 1.998059
```

```
sd(y)
```

```
## [1] 0.2145026
```

```
z <- rnorm(1000, mean = 2, sd = 1/5)
mean(z)
```

```
## [1] 2.006469
```

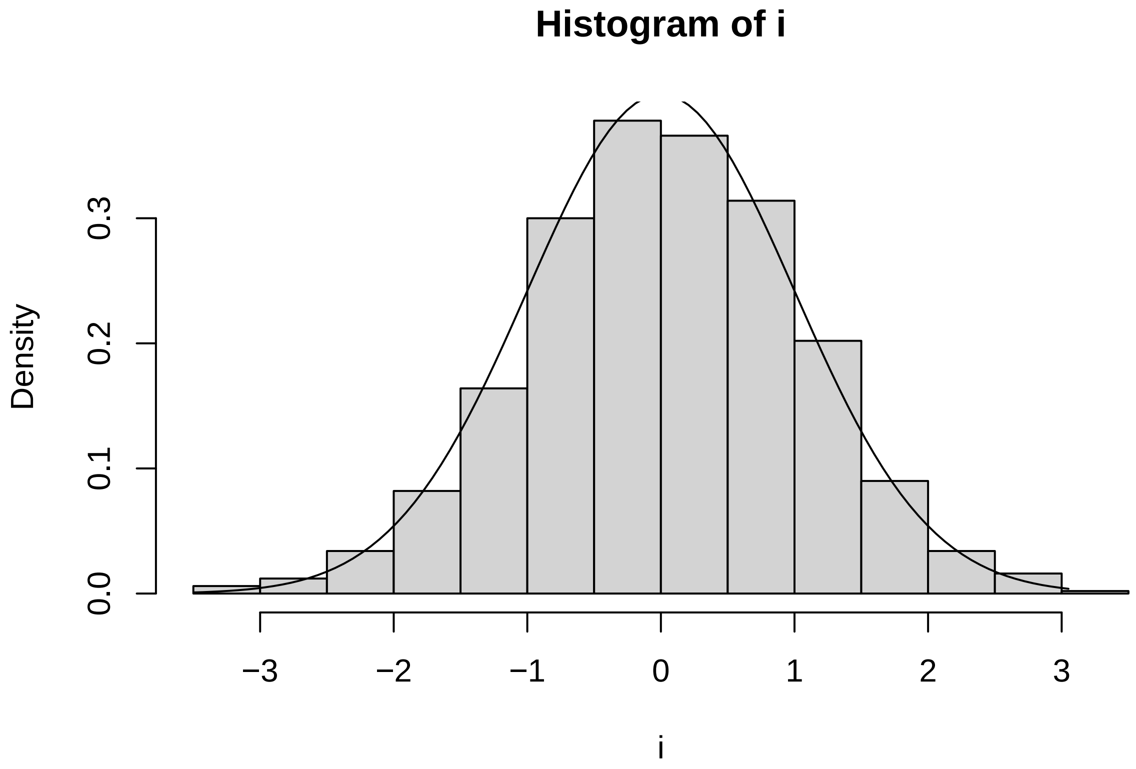
```
sd(z)
```

```
## [1] 0.1970801
```

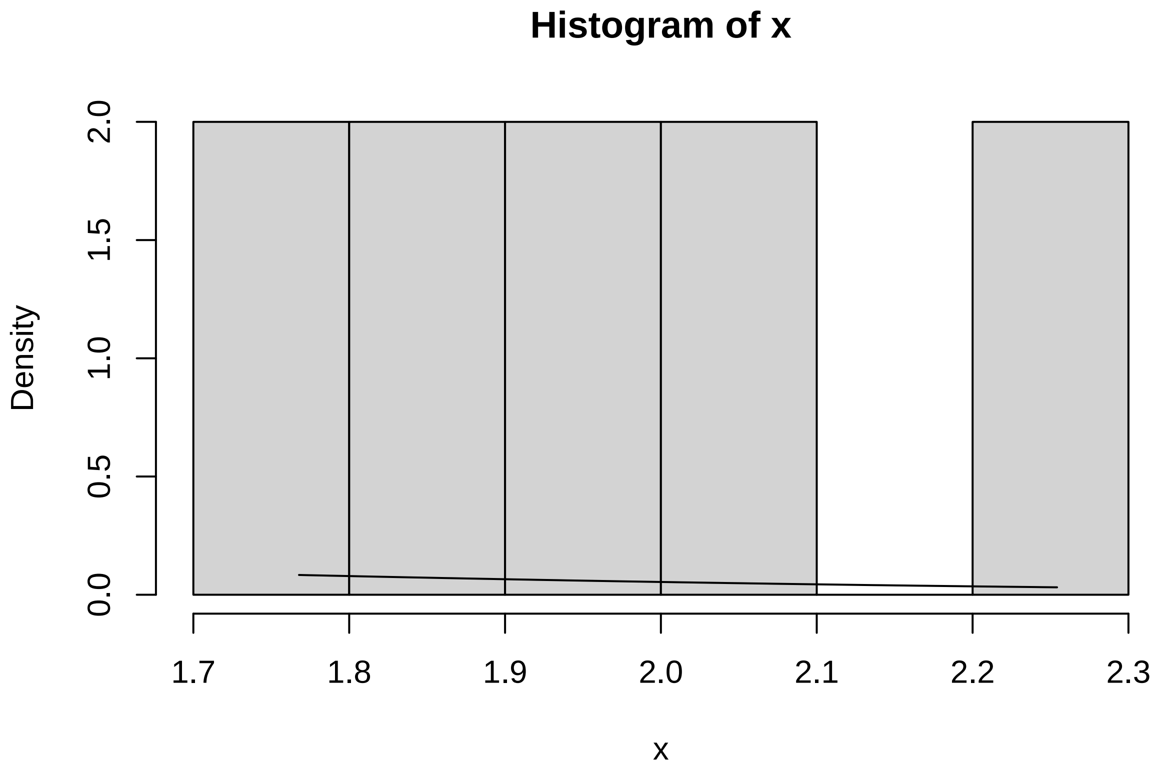
ANSWER: Standard deviation gets closer to 1/5 and mean gets closer to 2.

d) Add the theoretical distribution to the plot using the `lines` function.
QUESTION: Is it about the sub

```
i <- rnorm(1000)

hist(i, freq = FALSE)
fit1 <- seq(min(i), max(i), length = 100)
fit2 <- dnorm(fit1)

lines(fit1, fit2)
```



```
hist(x, freq = FALSE)
fitb1 <- seq(min(x),max(x), length = 100)
fitb2 <- dnorm(fitb1)

lines(fitb1, fitb2)
```



(**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.)