# Домашно 2

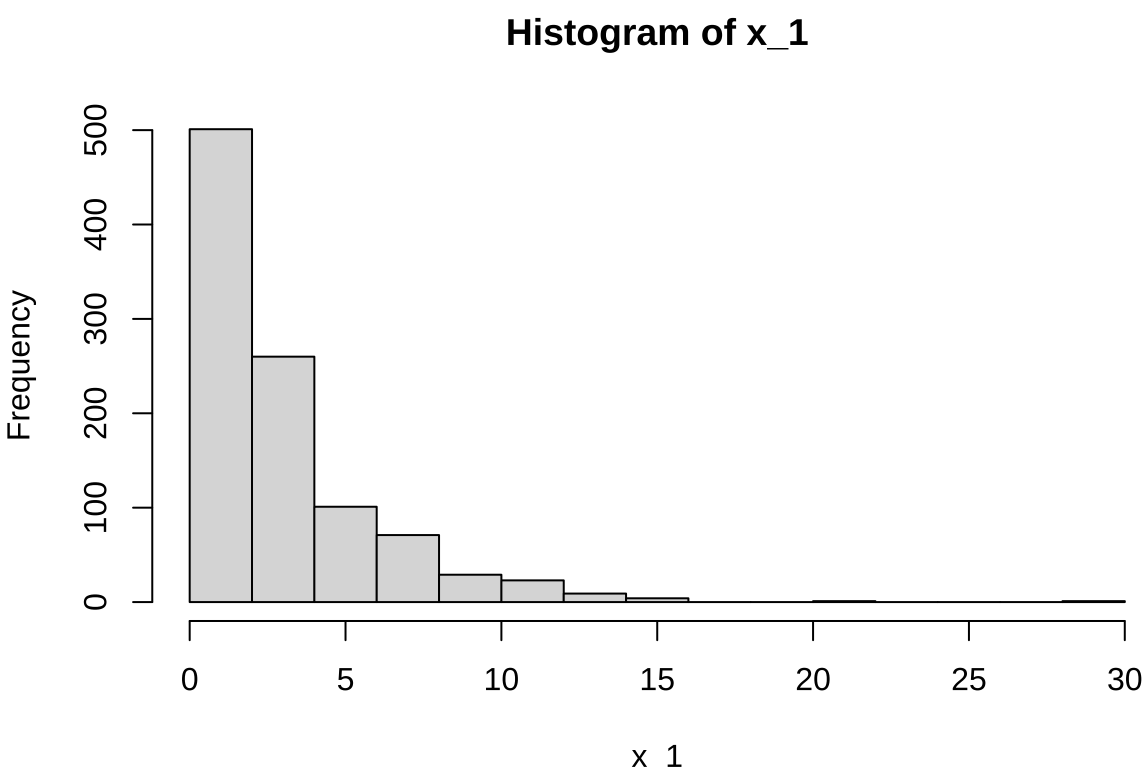
Домашно номер 2

Николай Кормушев 81805

Задача 1

Като начало генерираме експоненциално разпределени случайни величини x1-x500 и ги записваме в simulated_observations, за да се допитваме до тях в по-късен етап. Генерирам и хистограма на x_1, в която се вижда експоненциалното разпределение как изглежда.

```
lambda <- 1 / 3
sim_num <- 500
observation_num <- 1000

simulated_observations <- 1:sim_num %>%
  map(~rexp(observation_num, rate = lambda))

x_1 <- simulated_observations[1] %>%
  unlist()

hist(x_1)
```




####

a) Използвам qexp, да генерирам теоритични квартили при експоненциално разпределение с параметър lambda = 1/3

```
qexp(c(0, 0.25, 0.5, 0.75, 1), rate = lambda)
```

```
## [1] 0.0000000 0.8630462 2.0794415 4.1588831       Inf
```

Емпирични квартили получаваме функцията quantile, като я приложим над генерираните данни.

```
quantile(x_1)
```

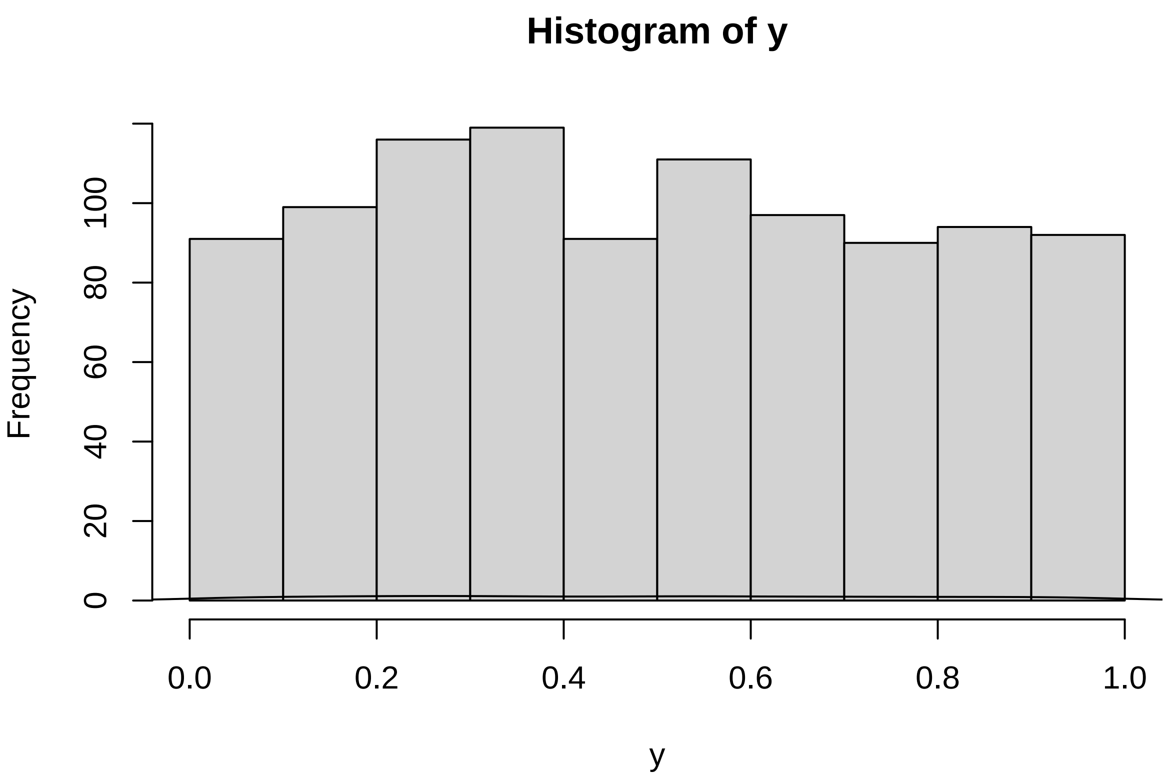
```
##           0%          25%          50%          75%         100%
##  0.006183834  0.844770899  1.993966363  3.885553792 29.645819584
```

б)  В тази точка искаме да видим какво е разпределнието на случайната величина y зададена долу.

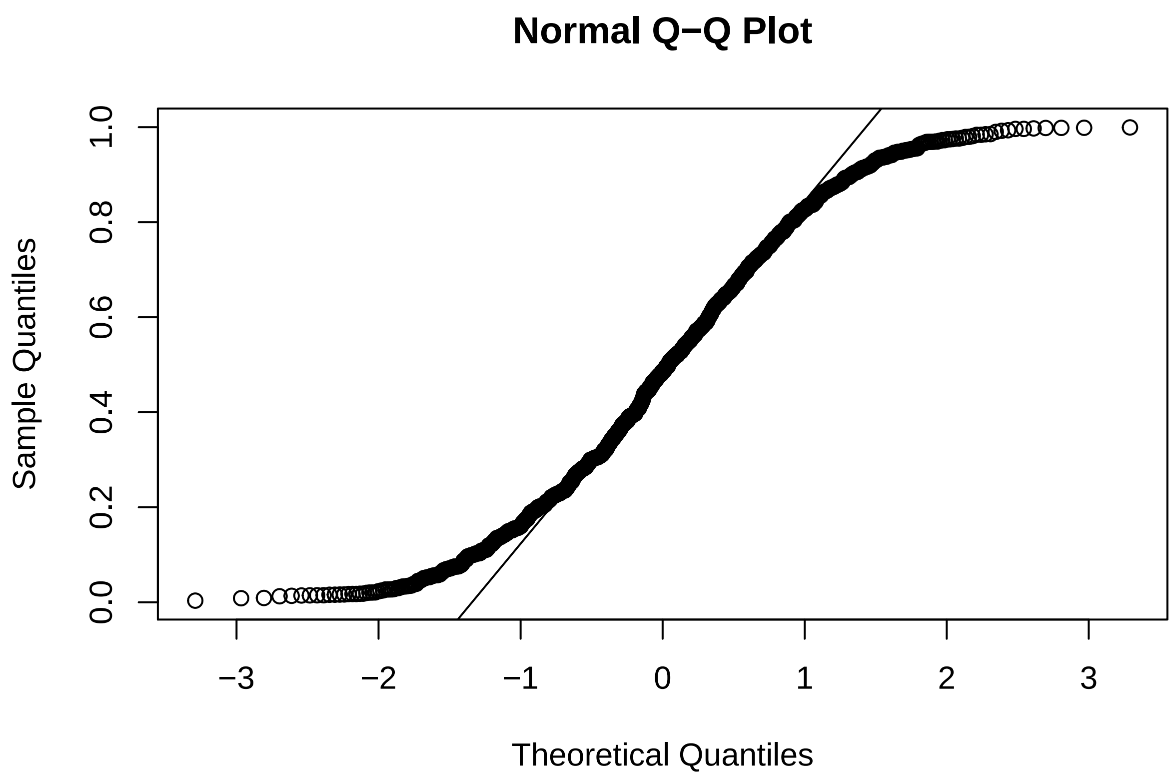
```
x_2 <- simulated_observations[2] %>%
    unlist()

y <- x_1 / (x_1 + x_2)
```

Според диаграмата разпределението изглежда равномерно

```
hist_output <- hist(y)
lines(density(y))
```




```
qqnorm(y)
qqline(y)
```




Ще направя chisq.test, да проверя хипотезата си дали е вярна. Първо разделям данните на секции и гледам колко често попадаме във всяка секция. После в prob смятам теоритичните вероятности и накрая с chisq.test проверявам хипотезата, че теоритичните и емпиричните вероятности съвпадат. Има 26% това да е в сила, което означава, че хипотезата ми не може да се изхвърли и приемам, че е вярна.

```
t <- table(cut(y, breaks = c(0, 1/4, 2/4, 3/4, 1)))

prob <- c(1,1,1,1)
prob[1] <- punif(1/4) - punif(0)
prob[2] <- punif(2/4) - punif(1/4)
prob[3] <- punif(2/4) - punif(3/4)
prob[4] <- punif(3/4) - punif(1)

chisq.test(t, prob)
```

```
## Warning in chisq.test(t, prob): Chi-squared approximation may be incorrect
```

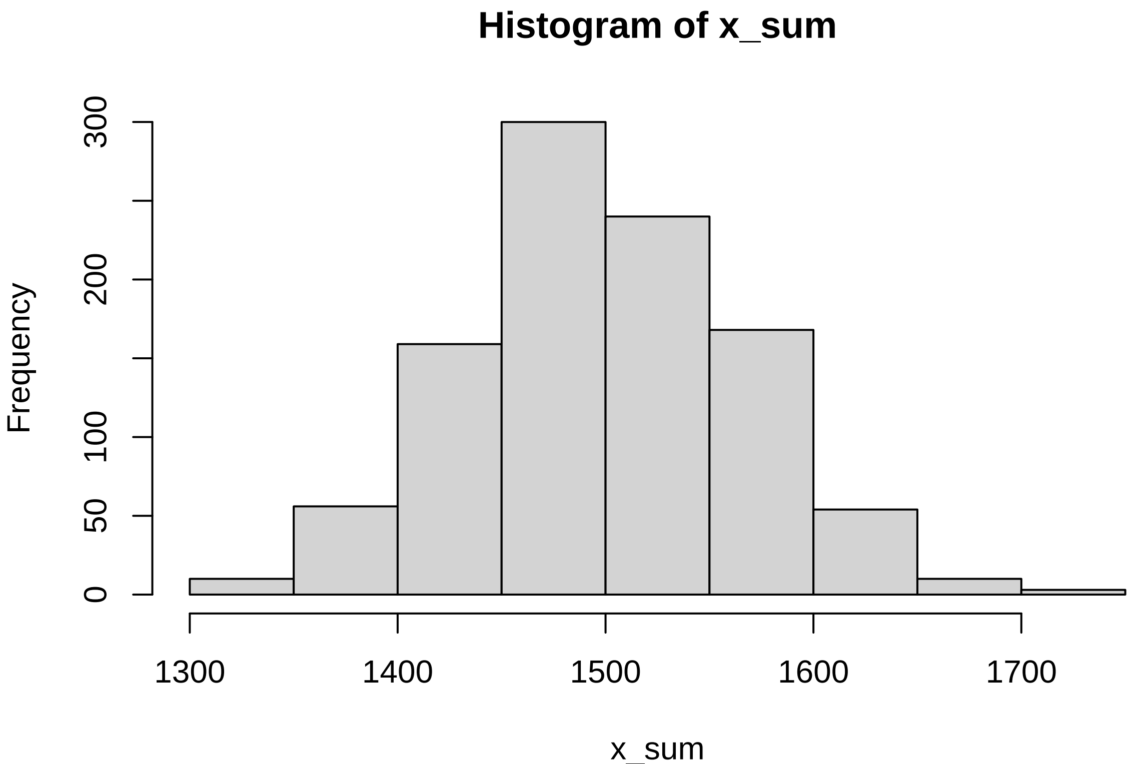
```
##
##  Pearson's Chi-squared test
##
## data:  t and prob
## X-squared = 4, df = 3, p-value = 0.2615
```

в) Използвам Reduce, да взема сумата на симулираните случайни величини. Искаме да видим какво е нейното разпределение.

```
x_sum <- Reduce(`+`, simulated_observations)
```

По хистограмата изглежда, като да е нормално, а и според централна гранична теорема е логично това да е резултатът.

```r
hist(x_sum)
```




Тестът на Шапиро даде доста високо p-value, което потвърждава хипотезата ми, че е нормално разпределена сумата.

```r
shapiro.test(x_sum)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x_sum
## W = 0.9974, p-value = 0.1103
```

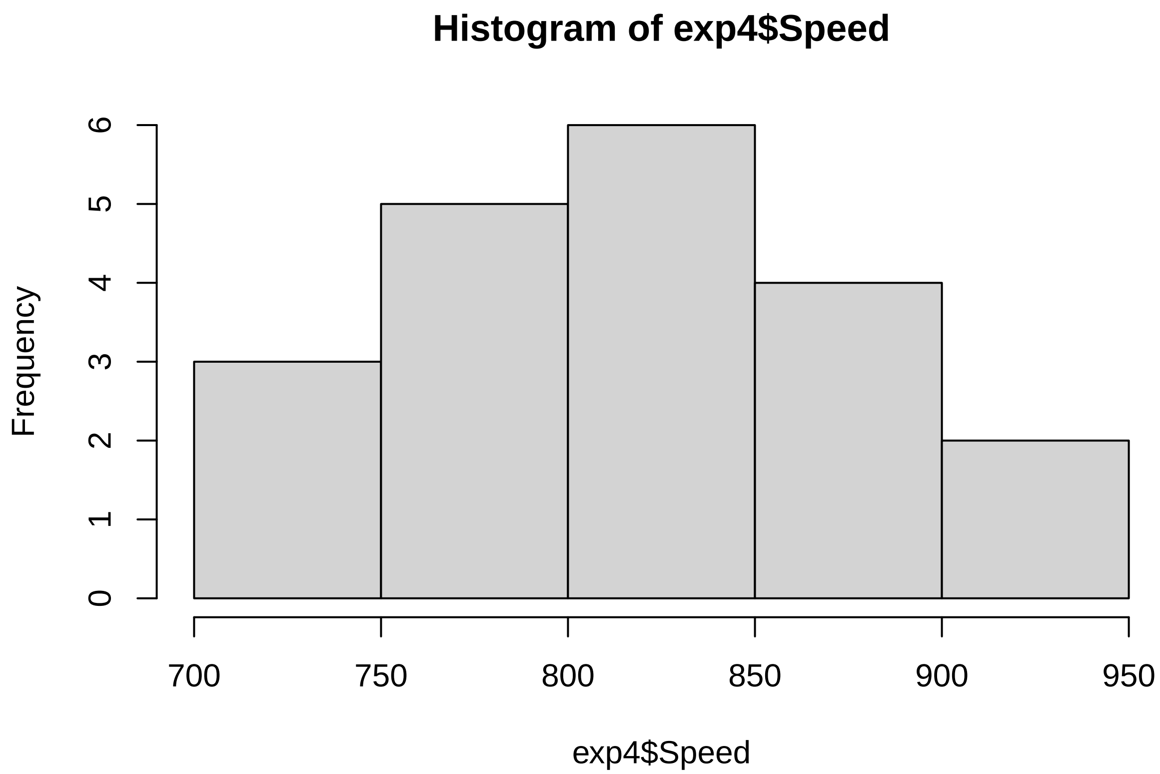
Задача 2

Искаме да построим 97% доверителен интервал за данните morley за скоростта на светлината и по-точно ще разгледаме данните получени при четвъртия експеримент. Като начало отделяме тези данни.

```r
exp4 <- morley[morley$Expt == 4, ]
```

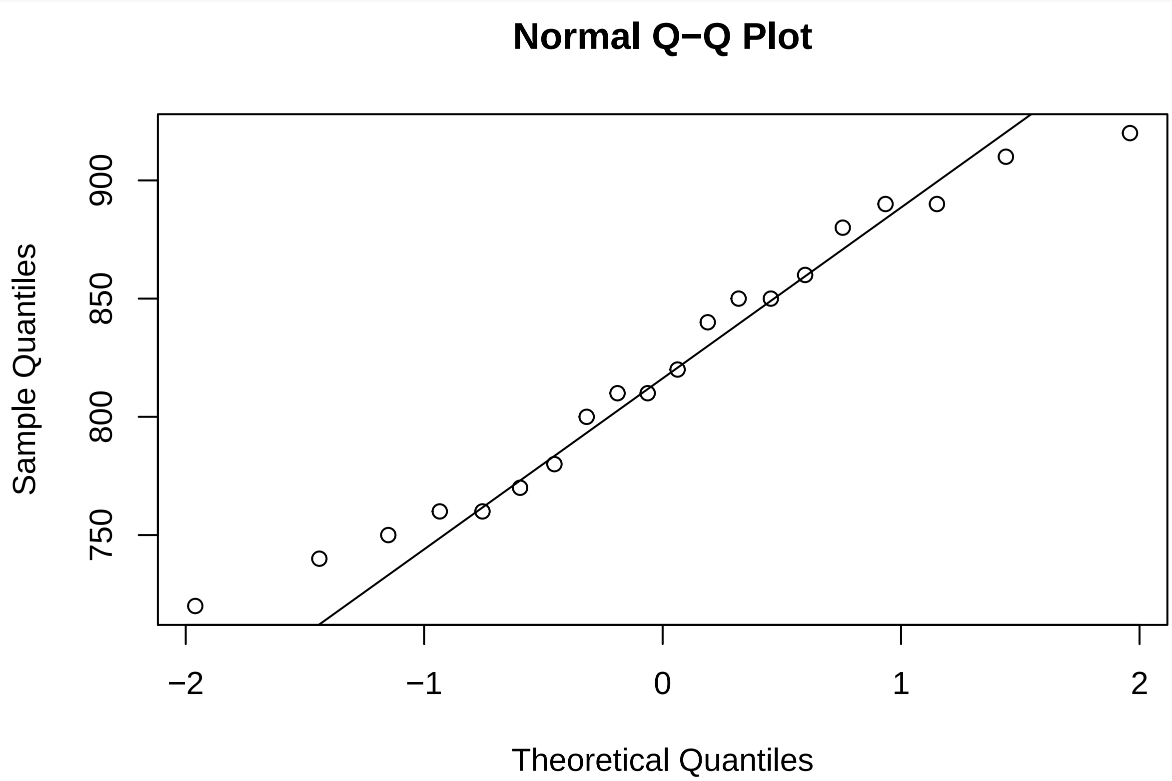
Хистограмата изглежда, като да е нормално разпределена, но нашата извадка е малка(20 теста) и може затова така да се получава.

```r
hist(exp4$Speed)
```




С qqnorm и qqline също се вижда, че има линейна зависимост и че данните са нормално разпределени.

```
qqnorm(exp4$Speed)
qqline(exp4$Speed)
```




За всеки случай реших и един shapiro test да пусна И той е потвърждава, че данните са нормално разпределени.

```
shapiro.test(exp4$Speed)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  exp4$Speed
## W = 0.96113, p-value = 0.5667
```

Горните заключения означават, че можем да използваме t.test, да получим доверителния интервал. Излиза, че има 97% шанс скоростта на светлината да е в доверителния интервал (789.008, 851.992)

```
t.test(exp4$Speed, conf.level = 0.97)
```

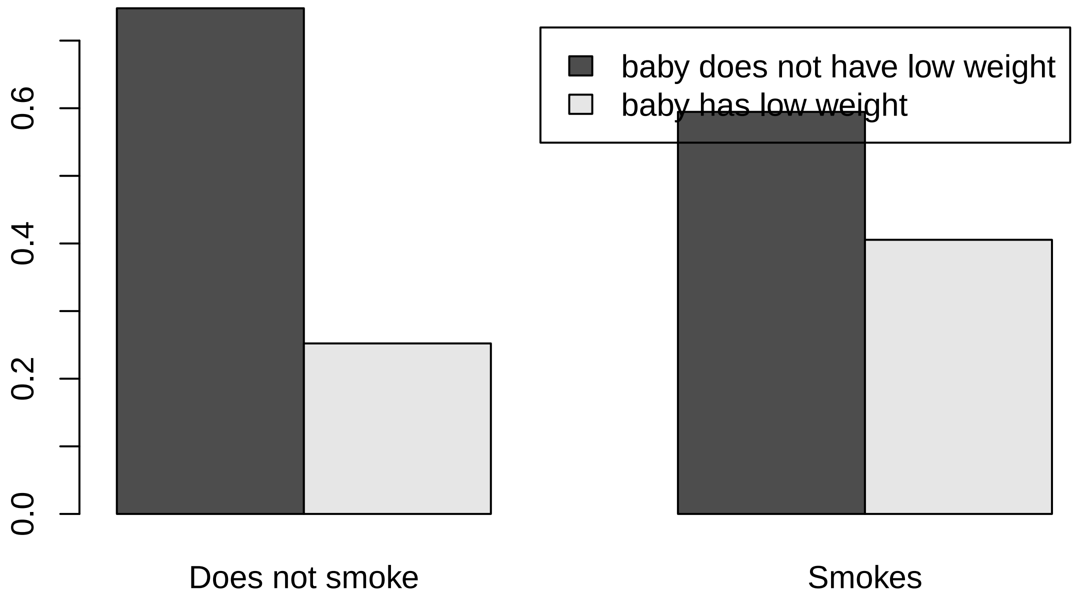
```
##
##  One Sample t-test
##
## data:  exp4$Speed
## t = 61.114, df = 19, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 97 percent confidence interval:
##  789.008 851.992
## sample estimates:
## mean of x
##     820.5
```

Задача 3

Искаме да видим, като разгледаме данните birthwt от пакета MASS дали има връзка между теглото на детето при раждане и това дали майката пуши. За целта ще разгледаме колонката low, която показва дали детето е с тегло под 2.5 кг при раждането.

Като начало реших да направя barplot, който показва, в случай, че майката пуши, каква е вероятността да има бебе с ниско тегло и ако не пуши, каква е вероятността да стане същото. На диаграмата се вижда, че се увеличава значително вероятността да имаш дете с ниско тегло, ако пушиш, както би се очаквало.

```
barplot(prop.table(table(birthwt$low, birthwt$smoke), 2),
        legend = T, names.arg = c("Does not smoke", "Smokes"),
        legend.text = c("baby does not have low weight", "baby has low weight"), beside = T)
```




За да потвърдя резултатите от диаграмата ще използвам prop.test. Като начало ще вземем множеството от пушачите.

```
smokers <- birthwt[birthwt$smoke == T, ]
```

И множеството от непушачите.

```
non_smokers <- birthwt[birthwt$smoke == F, ]
```

Гледам броя случаи, в които пушач е имал дете с ниско тегло.

```
smokers_low_babies <- sum(smokers$low == T)
```

Взимаме и общия брой на пушачите.

```
smoker_count <- nrow(smokers)
```

Аналогично за непушачите.

```
non_smokers_low_babies <- sum(non_smokers$low == T)
non_smoker_count <- nrow(non_smokers)
```

Правя prop.test, в който проверям дали вероятността на това пушач да има дете с тегло под 2.5 кг е по-голяма от тази при непушач. Резултатът е категорично да с 98% вероятност, от което можем да заключим, че е добра идея майките да не пушат, докато са бременни, ако вече не беше ясно.

```
prop.test(c(smokers_low_babies, non_smokers_low_babies),
          c(smoker_count, non_smoker_count), alternative = "less")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(smokers_low_babies, non_smokers_low_babies) out of c(smoker_count, non_smoker_count)
## X-squared = 4.2359, df = 1, p-value = 0.9802
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000  0.2794441
## sample estimates:
##    prop 1    prop 2
```

```
## 0.4054054 0.2521739
```

В последствие видях, че мога да използвам и chisq.test да видя дали са независими две случайни величини. В случая величините са дали майката пуши или не и броя на децата с тегло под 2.5. За целта създаваме матрица с данните за пушачите на първия ред и данните на непушачите на втория ред. Първата колона е броя на майки без проблеми, а втората броя майки с деца с тегло под нормата. Подаваме матрицата на chisq.test и виждаме, че вероятността да са независими е под 5%, т.е. и така потвърдихме, че има значение дали майката пуши или не. Мисля, че и по двата начина става и затова ще оставя и двата, макар и да може да греша и да трябва да се използва chisq.test

```
m <- matrix(c(smoker_count - smokers_low_babies, smokers_low_babies,
       non_smoker_count - non_smokers_low_babies, non_smokers_low_babies),
       nrow = 2, byrow = T)
chisq.test(m)
```

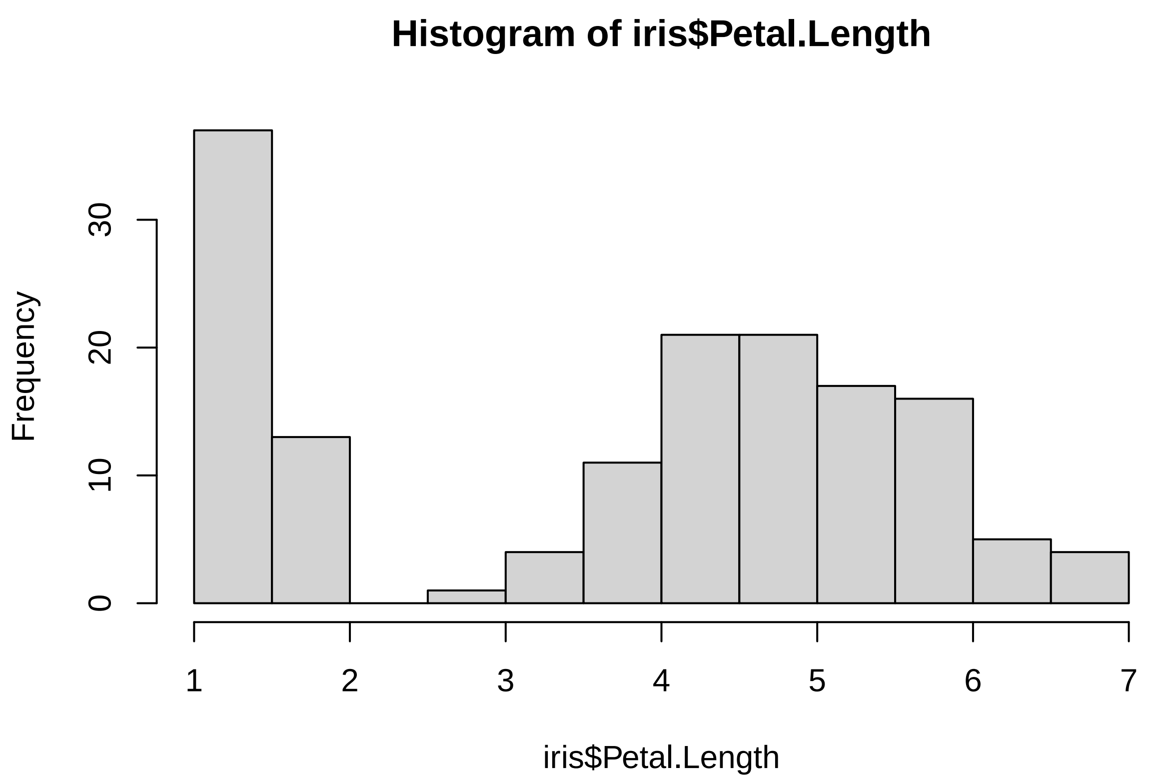
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  m
## X-squared = 4.2359, df = 1, p-value = 0.03958
```

Задача 4

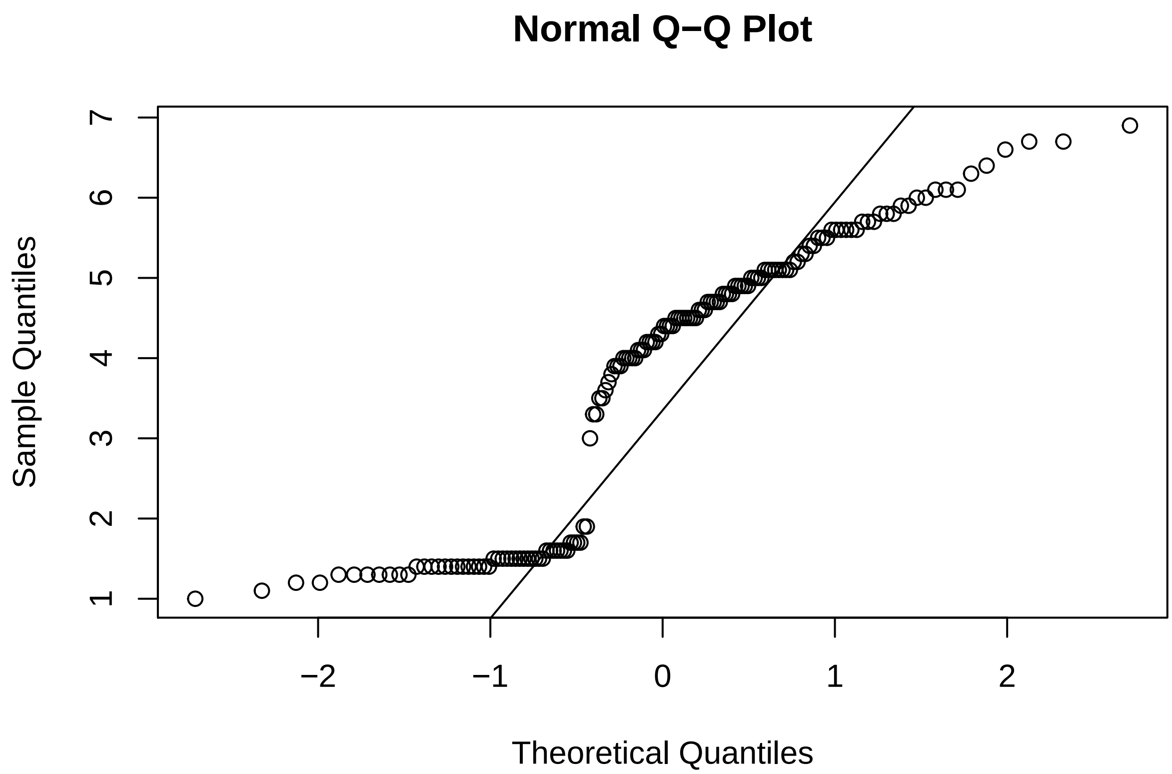
а) Искаме да видим дали спрямо данните от таблиците iris можем да заключим, че дължината на венчелистчетата е равна на три пъти ширината им.

Като начало гледаме хистограмата, qqnorm, qqplot и за здраве правим и един shapiro.test, от което явно се вижда, че данните не са нормално разпределени.

```
hist(iris$Petal.Length)
```



```
qqnorm(iris$Petal.Length)
qqline(iris$Petal.Length)
```




```
shapiro.test(iris$Petal.Length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  iris$Petal.Length
## W = 0.87627, p-value = 7.412e-10
```

Затова използваме wilcox.test и смятаме дали разликата между дължината и ширината * 3 е нула. Вижда се, че вероятността това да е така е 16%, което означава, че не можем да отхвърлим тази хипотеза.

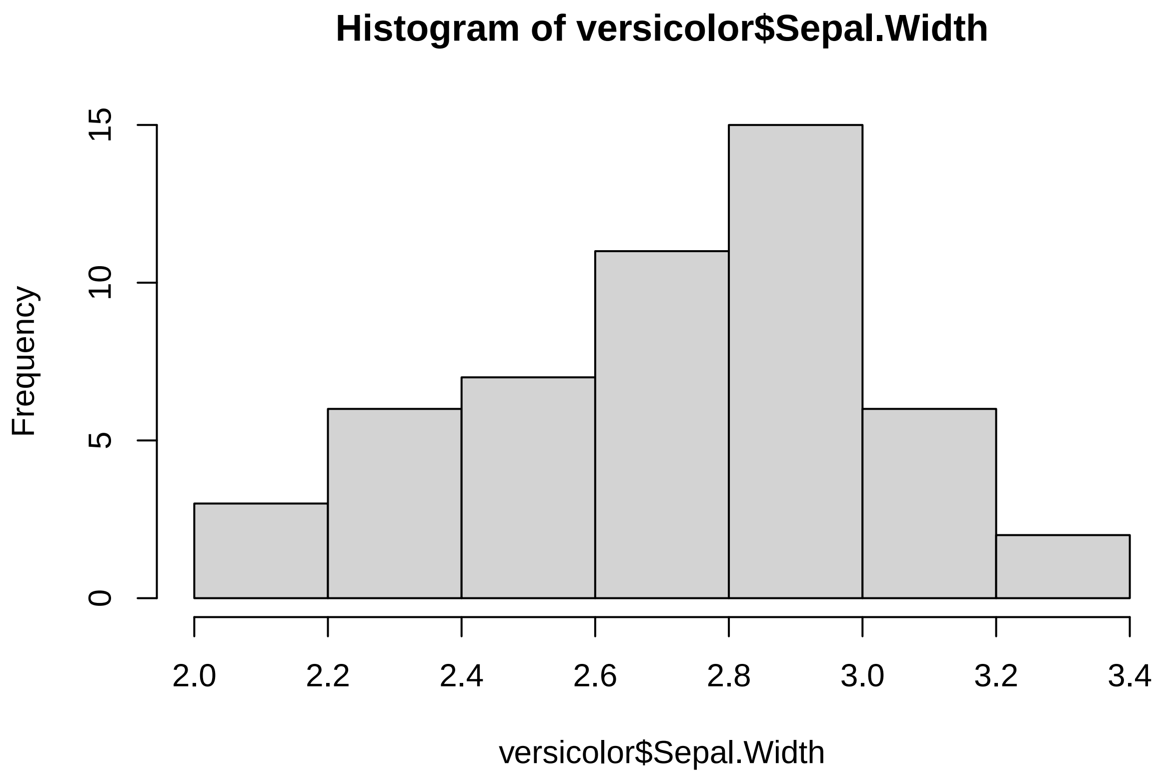
```
wilcox.test(iris$Petal.Length, 3 * iris$Petal.Width,
            alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  iris$Petal.Length and 3 * iris$Petal.Width
## W = 12295, p-value = 0.164
## alternative hypothesis: true location shift is not equal to 0
```

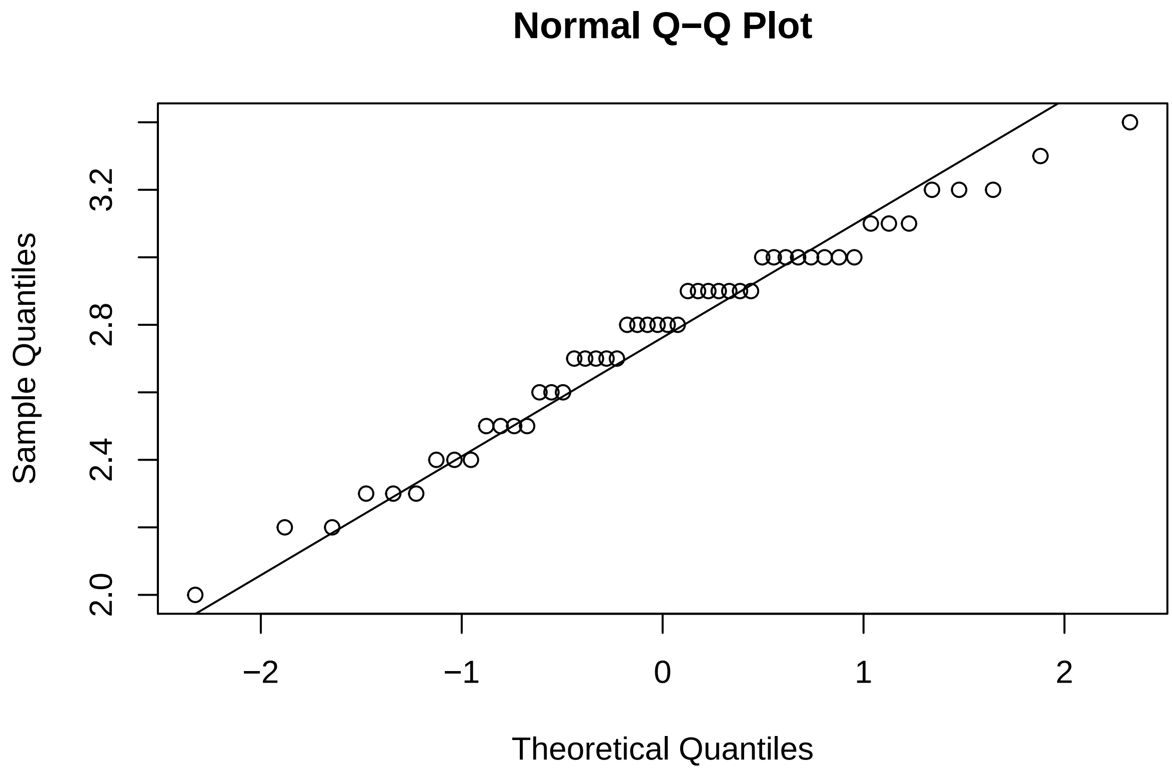
б) Тествам дали ширината на чашелистчетата на ирисите от сорт versicolor е по-голяма от ширината на чашелистчетата на сорт virginica.

Започнах с проверка дали данните са нормално разпределени Работих аналогично на а). Тук според диаграмите и тестът на шапиро се вижда, че имаме нормално разпределени данни и можем да използваме t.test.

```r
versicolor <- iris[iris$Species == "versicolor", ]
virginica <- iris[iris$Species == "virginica", ]

hist(versicolor$Sepal.Width)
```



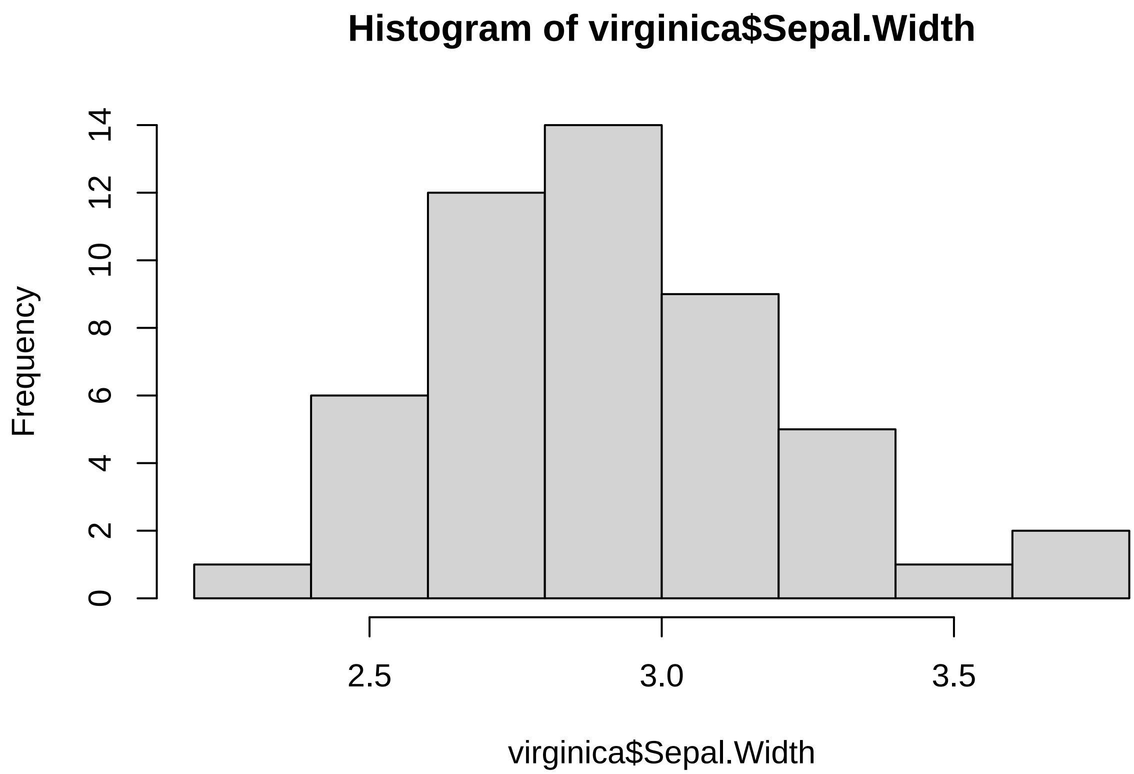

```r
qqnorm(versicolor$Sepal.Width)
qqline(versicolor$Sepal.Width)
```



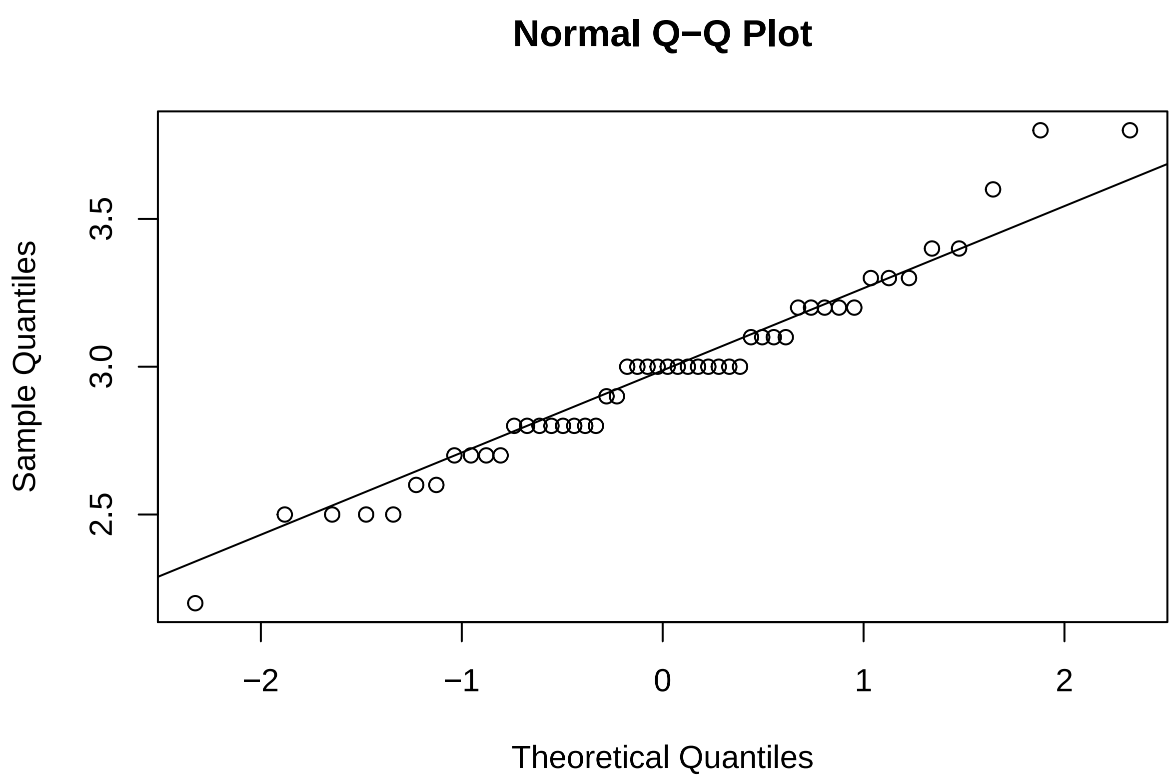

```
shapiro.test(versicolor$Sepal.Width)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  versicolor$Sepal.Width
## W = 0.97413, p-value = 0.338
```

```
hist(virginica$Sepal.Width)
```




```
qqnorm(virginica$Sepal.Width)
qqline(virginica$Sepal.Width)
```



```
shapiro.test(virginica$Sepal.Width)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  virginica$Sepal.Width
## W = 0.96739, p-value = 0.1809
```

Т тестът показва, че вероятността versicolor да са с по-голяма дължина от virginica е под 1%, което означава, че хипотезата ни е грешна и можем да я отхвърлим.

```
t.test(versicolor$Sepal.Width, virginica$Sepal.Width, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  versicolor$Sepal.Width and virginica$Sepal.Width
## t = -3.2058, df = 97.927, p-value = 0.0009097
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.09832934
## sample estimates:
## mean of x mean of y
##     2.770     2.974
```

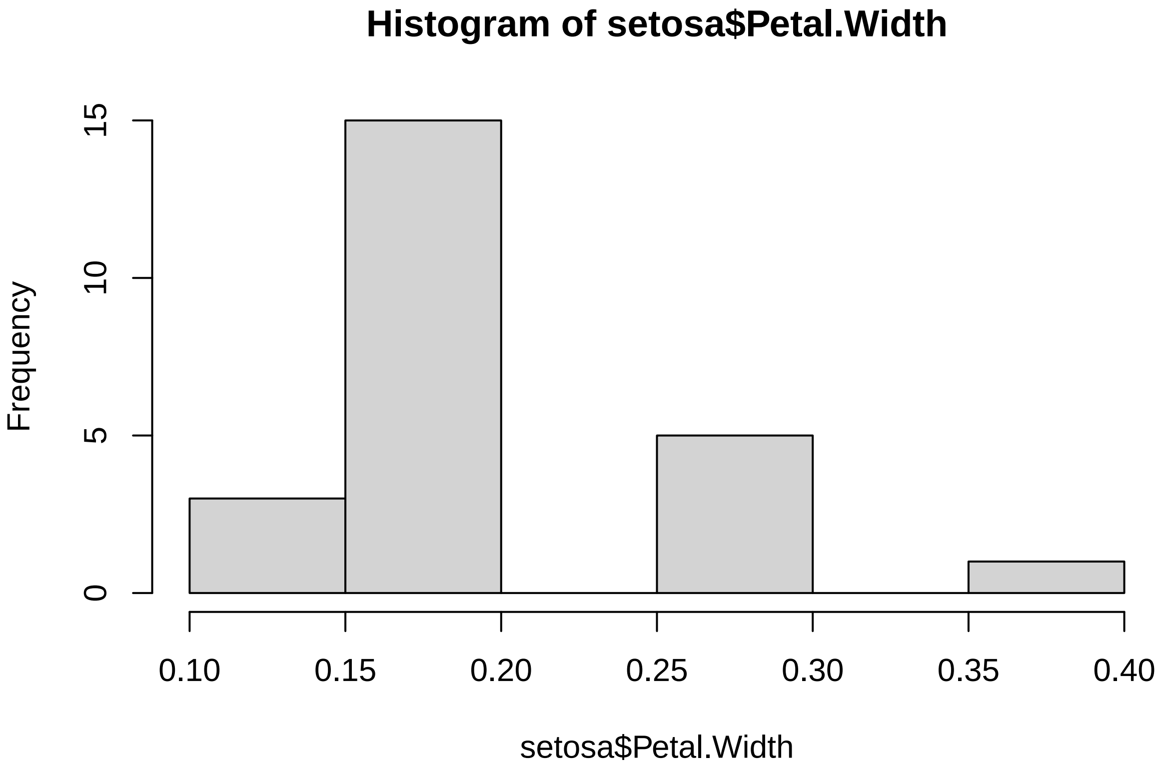
в) Проверявам дали ирисите от сорт setosa с дължина на венчелистчетата по-малко от 1.4 имат ширина на венчелистчетата по-голяма от 0.26.

Първо взимам iris-ите от тип setosa с желната дължина.

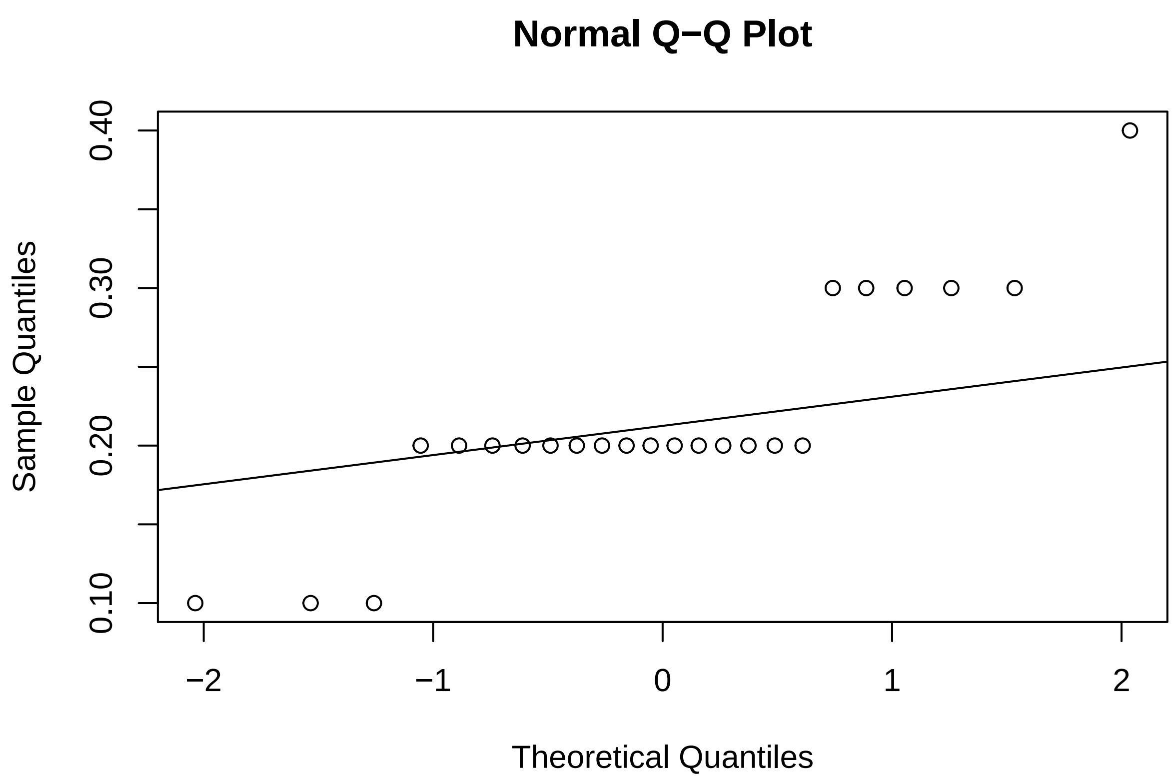
```
setosa <- iris[iris$Petal.Length <= 1.4 & iris$Species == "setosa", ]
```

Според графиките и shapiro.test не са нормално разпределени данните. Затова ще използвам wilcox.test.

```
hist(setosa$Petal.Width)
```




```
qqnorm(setosa$Petal.Width)
qqline(setosa$Petal.Width)
```



```r
shapiro.test(setosa$Petal.Width)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  setosa$Petal.Width
## W = 0.80724, p-value = 0.0003877
```

Тестът показва, че вероятността венчелистчетата да са с ширина по-голямо от 0.26 е много малка (под 1%), т.е хипотезата, че ирисите от този тип са с дължина на венчелистчетата по-голяма от 0.26 може да се отхвърли.

```r
wilcox.test(setosa$Petal.Width, mu = 0.26, alternative = "less")
```

```
## Warning in wilcox.test.default(setosa$Petal.Width, mu = 0.26, alternative =
## "less"): cannot compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  setosa$Petal.Width
## V = 36, p-value = 0.0004129
## alternative hypothesis: true location is less than 0.26
```