Проверка на нормальность распределения

Законы распределения вероятностей в R

° **d** (от "*density*", плотность): функции плотности вероятности ("функция распределения масс" для дискретных величин);
° **p** (от "*probability*", *вероятность*): кумулятивные функции Распределения вероятностей;
° **q** (от "*quantile*", квантиль): функции для нахождения квантиле
° **r** (от "*random*", случайный): функции для генерации случайны чисел в соответствии с параметрами того или иного закона распределения вероятностей.

Законы распределения вероятностей (базовая версия) :

° Бета-распределение (dbeta)
° Биномиальное распределение (включая распределение Бернулли) (dbinom)
° Распределение Коши (dcauchy)
° Распределение хи-квадрат (dchisq)
° Экспоненциальное распределение (dexp)
° Распределение Фишера (df)
° Гамма-распределение (dgamma)
° Геометрическое распределение (как частный случай отрицательного биномиального распределения) (dgeom)
° Гипергеометрическое распределение (dhyper)
° Логнормальное распределение (dlnorm)
° Полиномиальное (или мультиномиальное) распределение (dmultinom)
° Отрицательное биномиальное распределение (dnbinom)
° Нормальное распределение (dnorm)
° Распределение Пуассона (dpois)
° Распределение Стьюдента (dt)
° Равномерное распределение (dunif)
° Распределение Вейбулла (dweibull)

Пусть мы имеем дело с непрерывной количественной величиной X, значения которой распределены в соответствии со стандартным нормальным распределением (среднее значение = 0, стандартное отклонение = 1).

Функция плотности вероятности представляет собой такую функцию $f(x)$, что для любых двух значений a и b (при $a \leq b$)

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$

Вероятность того, что некоторая случайная величина X принимает значение, лежаще
интервале $[a, b]$, равна площади под кривой плотности вероятности, ограниченной э
интервалом.

Дифференциальная функция плотности вероятности стандартного нормального распределения в точке x задается уравнением

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

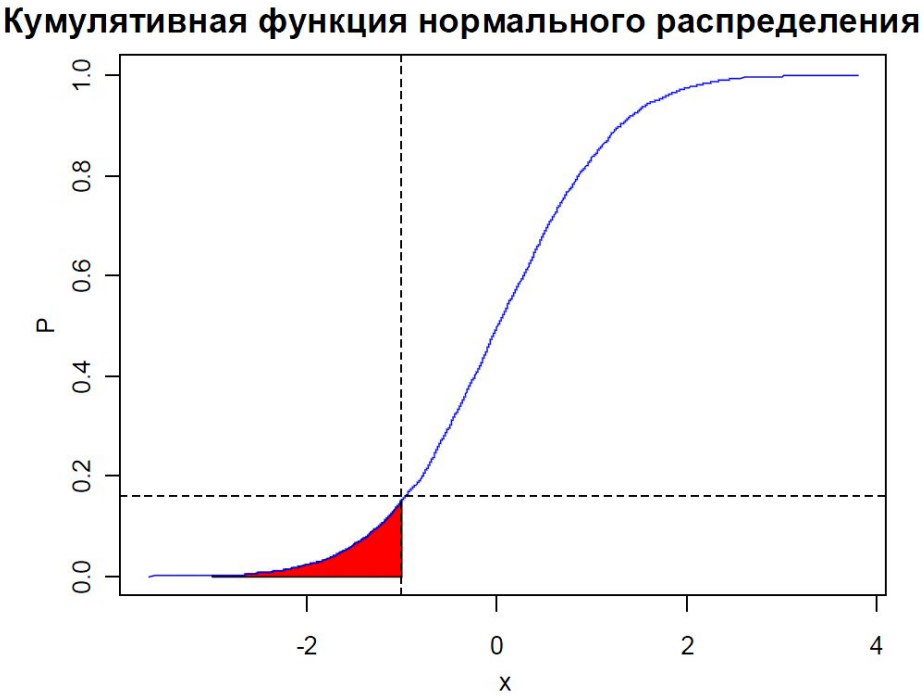
Для *x* = -1 в случае со стандартным нормальным распределением

dnorm(-1)

[1] 0.2419707

pnorm(-1)

[1] 0.1586553



функция qnorm()

Вычислим 1-ый и 3-ий квартили стандартного нормального распределения:

> qnorm(p = c(0.25, 0.75))

[1] -0.6744898 0.6744898

> qnorm(p = c(0.025, 0.975))

[1] -1.959964 1.959964

Функция **rnorm**() служит для случайной генерации совокупностей нормально распределенных чисел.
Сгенерируем совокупность из 10 значений из стандартного нормального распределения:

```
>rnorm(10, mean = 0, sd = 1)
[1] -0.98696489 -0.53126664 -0.23150543 -0.84139429
[5] -1.81401823  0.48510932  0.04734179  0.32588926
[9] -0.36508765 -0.37539185
```

```
> rnorm(8, mean = 13, sd = 3)
[1] 12.65565 18.07006 11.97118 16.21725 15.04990 21.60843 16.14872 16.05072
```

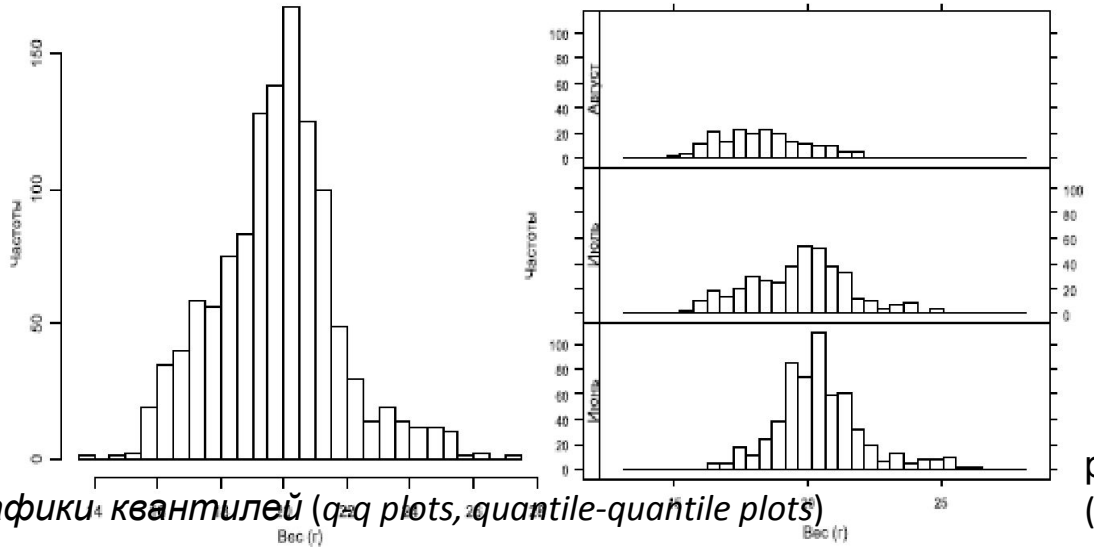
пакеты VGAM, actuar, gamlss и ActuDistns

Проверка на нормальность распределения

Проверка исследуемых переменных на нормальность распределения является важной составной частью разведочного анализа данных

Графические способы

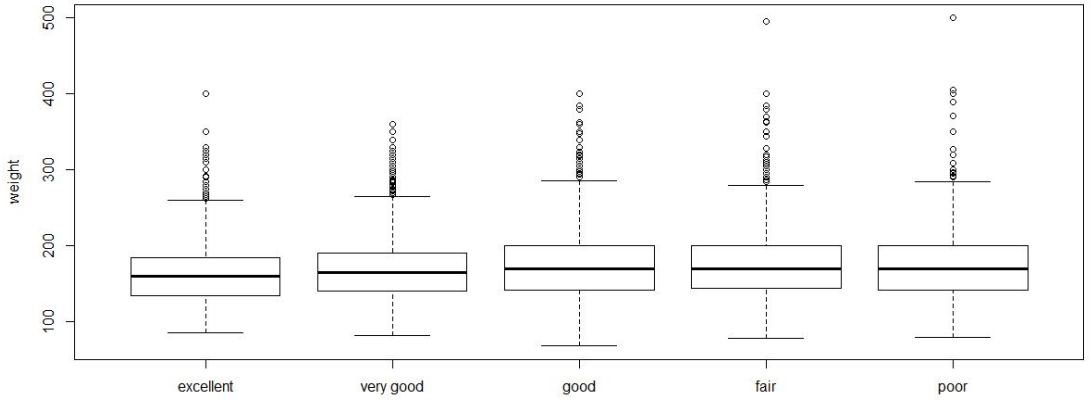
гистограммы



распределение веса 1193 воробьев (Zuur et al., 2010)

Графики квантилей (qq plots, quantile-quantile plots)

Коробчатые графики, боксплоты(boxplots)

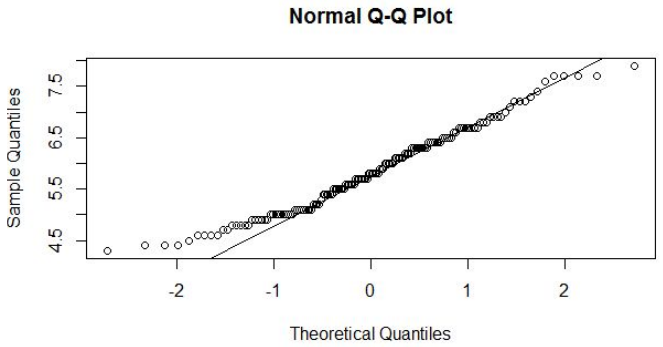


Графики квантилей (*q-q plots, quantile-quantile plots*)

функции **qqnorm()** и **qqplot()**

Квантиль-квантильный график без доверительных огибающих

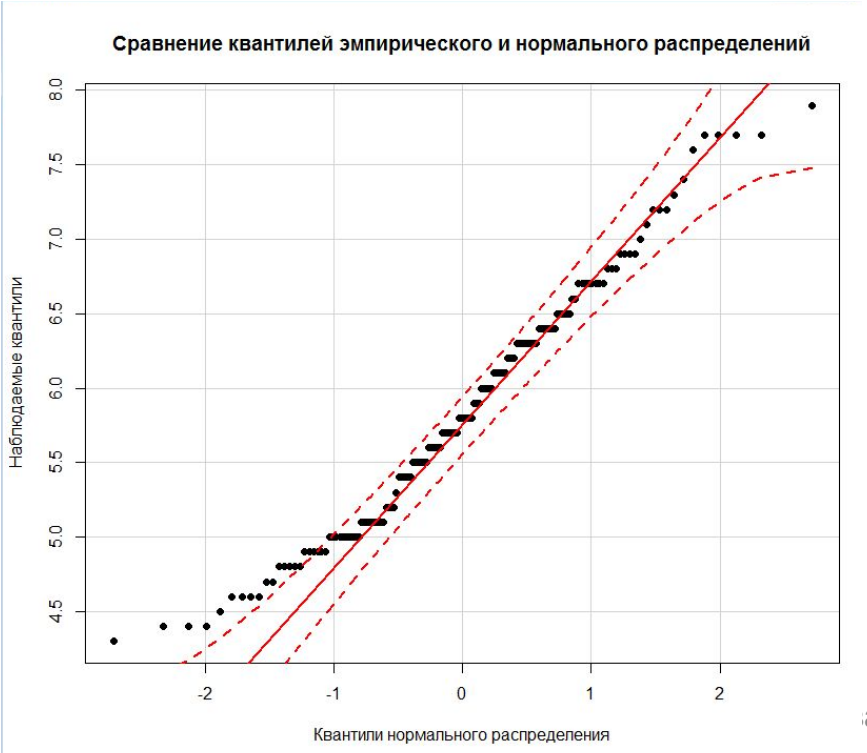
qqnorm(x); **qqline**(x)



Функция **qqPlot()** пакета **car**

для `Sepal.Length` из фрейма `iris`:

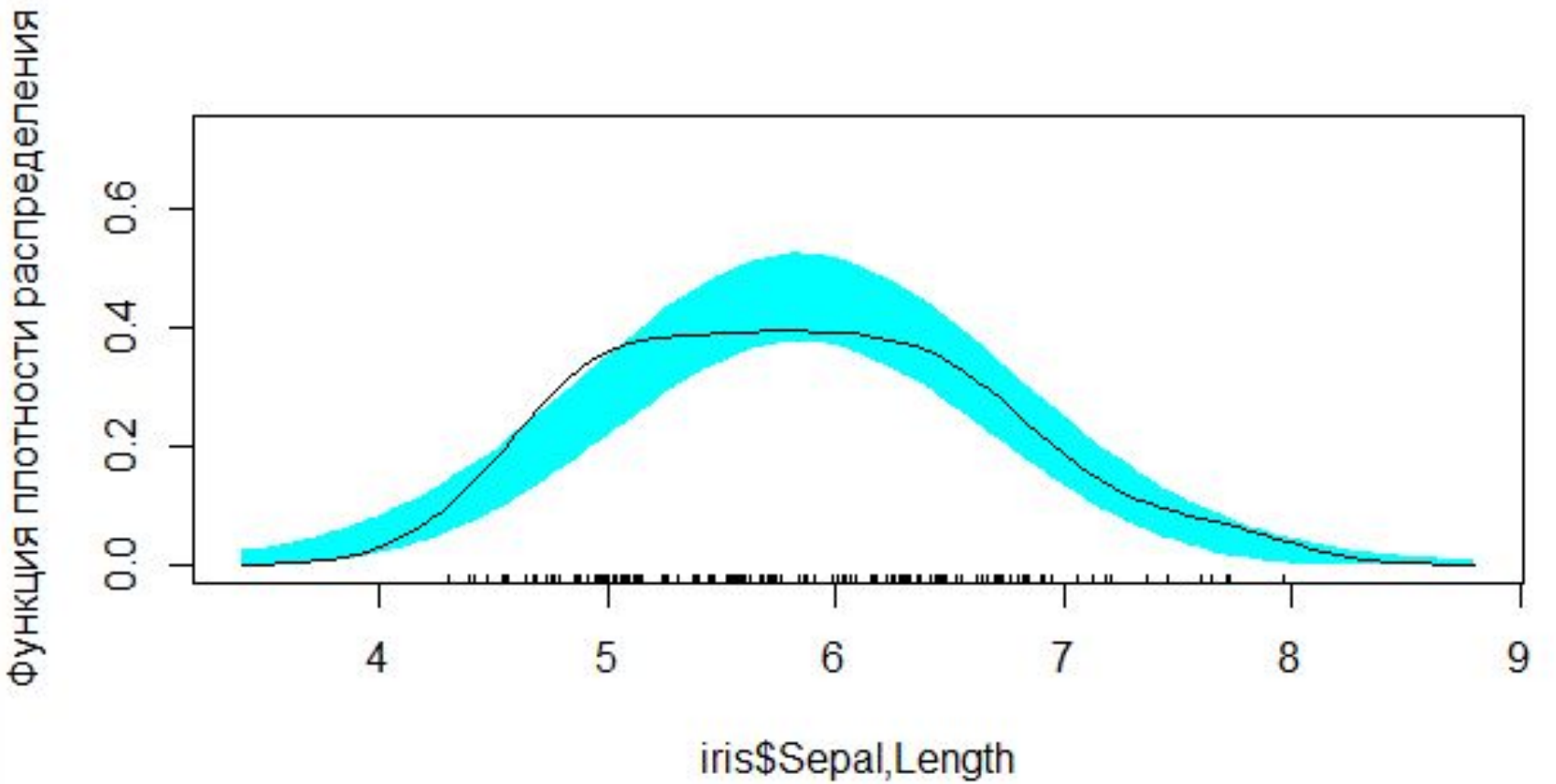
```
>library(car)
>qqPlot(x, dist= "norm", col=palette()[1] , pch=19, xlab="Квантили нормального распределения", ylab="
Наблюдаемые квантили",main="Сравнение квантилей эмпирического и нормального распределений")
```



sm.density() и **sm.density.compare()** из пакета **sm**

>library(sm)

>sm.density(x, model = "Normal", xlab=" iris$Sepal.Length", ylab="Функция плотности распределения")



Формальные тесты

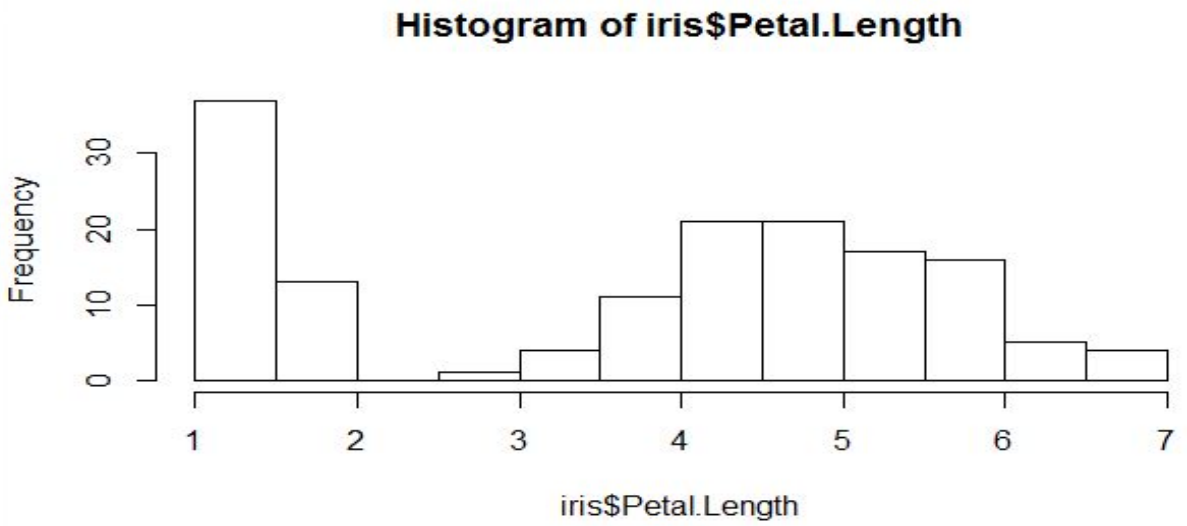
Нулевую гипотезу можно сформулировать так: "анализируемая выборка происходит из генеральной совокупности, имеющей нормальное распределение". Если получаемая при помощи того или иного теста вероятность ошибки p оказывается меньше
некоторого заранее принятого уровня значимости (например, 0.05), нулевая гипотеза отклоняется.
Базовая функция **shapiro.test(),** при помощи которой можно выполнить широко используемый **тест Шапиро-Уилка**.

функции из пакета nortest, реализующие другие распространенные тесты на нормальность:

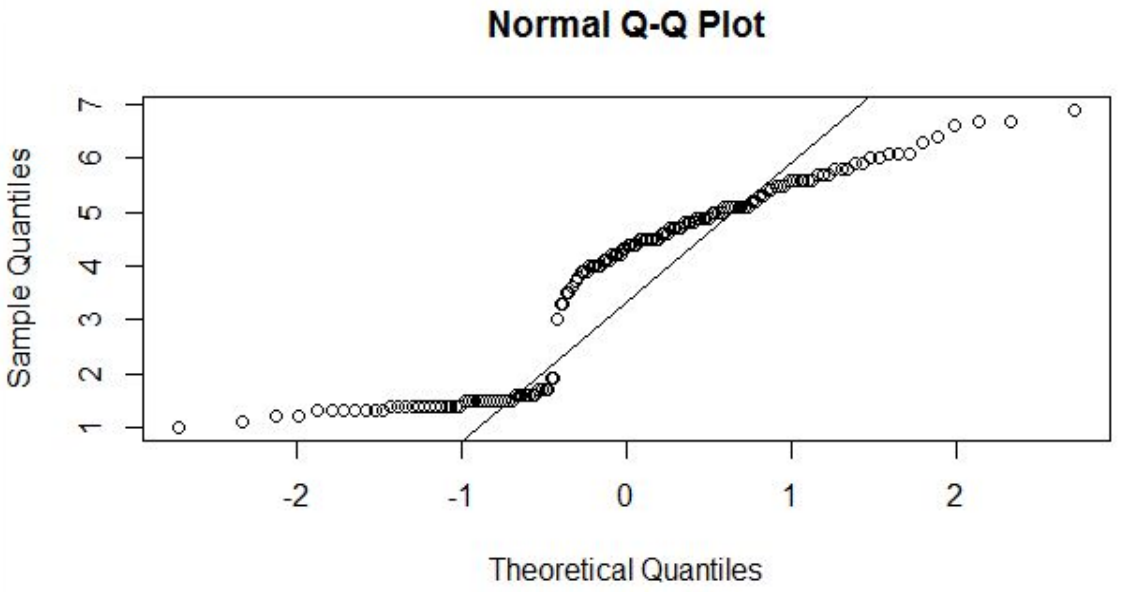
- **ad.test() - тест Андерсона-Дарлинга;**
- **cvm.test() - тест Крамера фон Мизеса;**
- **lillie.test() - тест Колмогорова-Смирнова в модификации Лиллиефорса;**
- **sf.test() - тест Шапиро-Франсия**

```
shapiro.test(x)
Shapiro-Wilk normality test
data: x
W = 0.8986, p-value = 1.219e-06
library(nortest)
ad.test(x)
Anderson-Darling normality test
data: x
A = 2.0895, p-value = 2.382e-05
cvm.test(x)
Cramer-von Mises normality test
data: x
W = 0.3369, p-value = 0.0001219
lillie.test(x)
Lilliefors (Kolmogorov-Smirnov) normality test
data: x
D = 0.1348, p-value = 0.0001225
sf.test(x)
Shapiro-Francia normality test
data: x
W = 0.8936, p-value = 3.617e-06
```

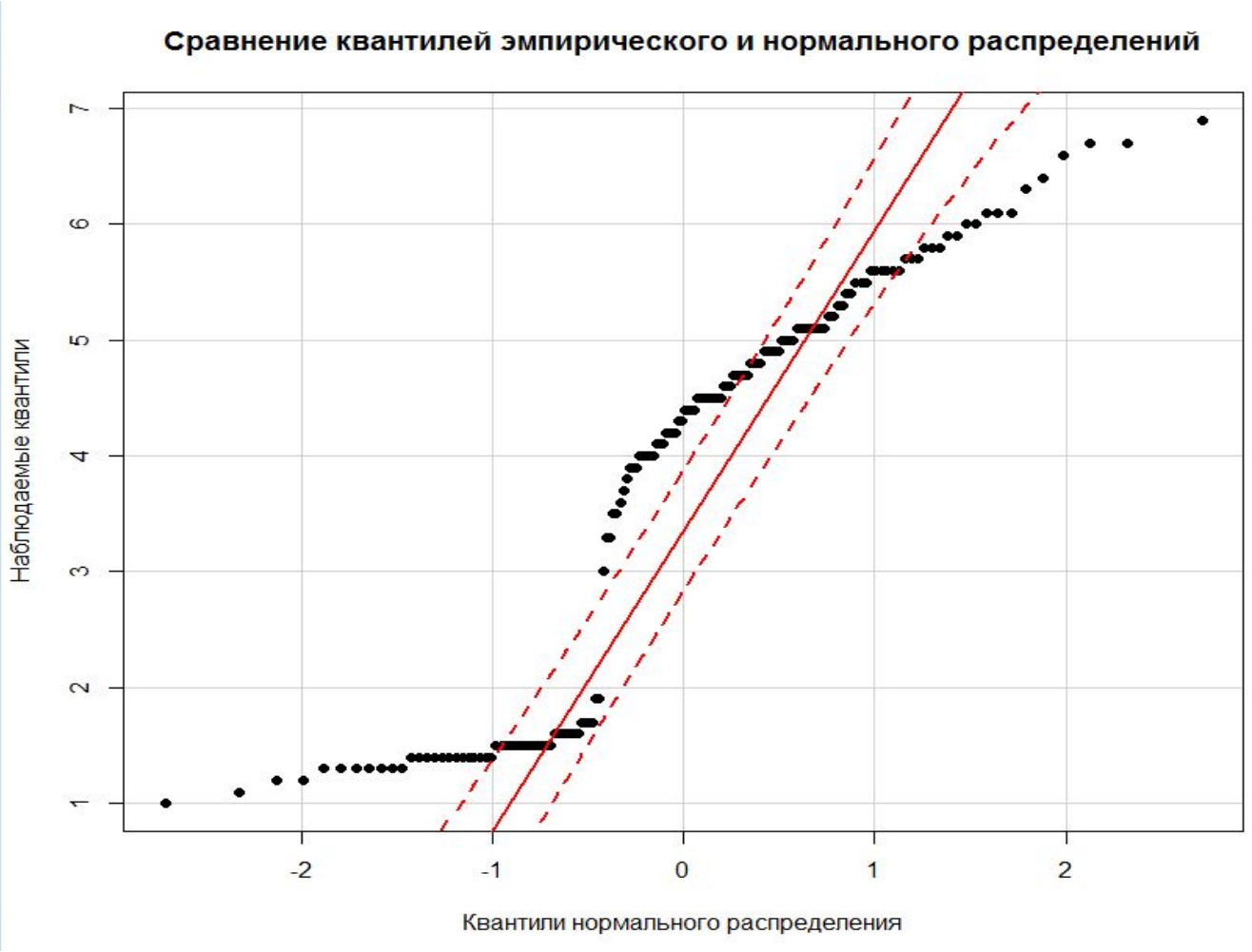
```
> hist(iris$Petal.Length)
```



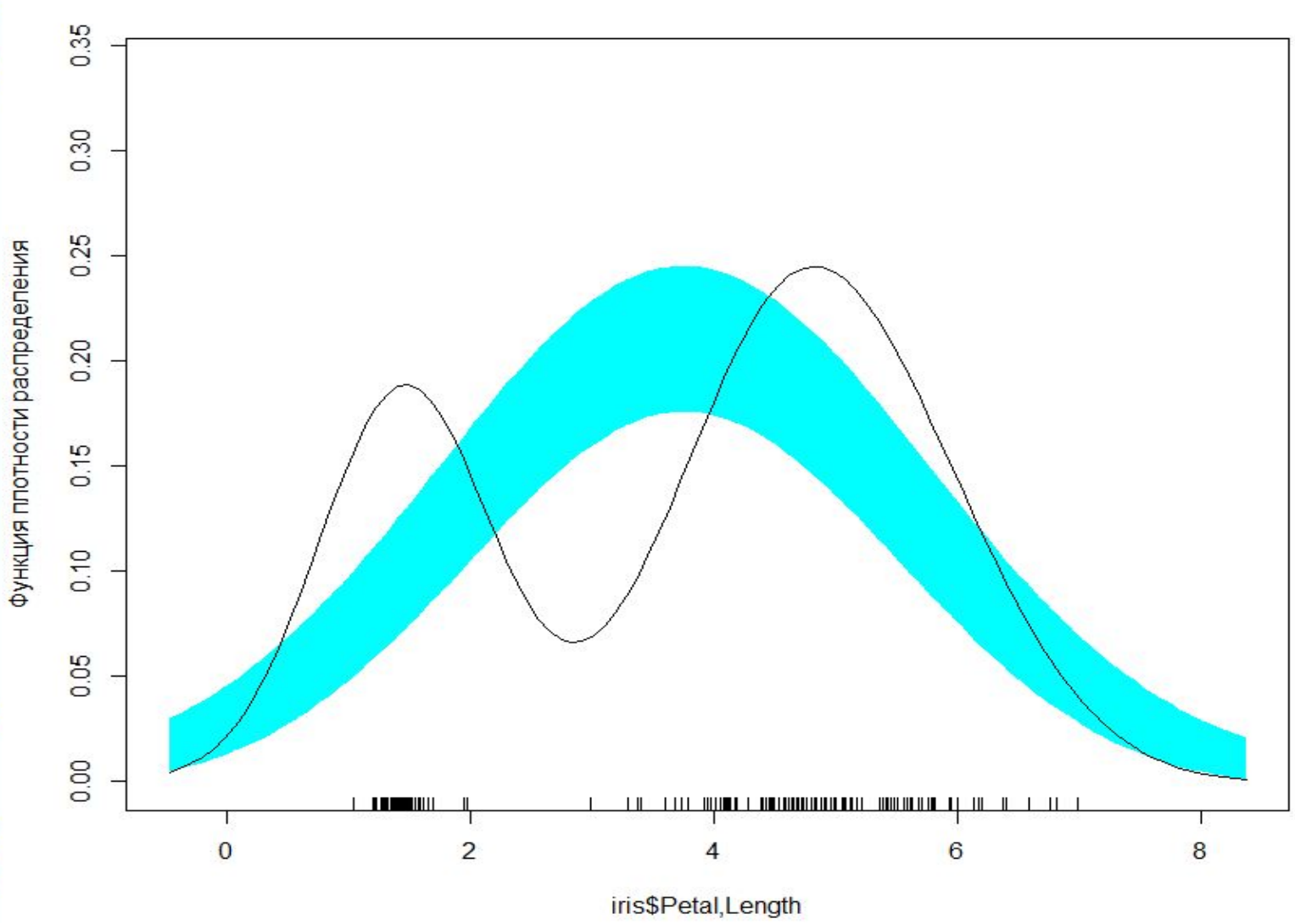

```
> qqnorm(iris$Petal.Length)
> qqline(iris$Petal.Length)
```



```
> library(car)
> qqPlot(iris$Petal.Length, dist= "norm", col=palette()[1] , pch=19,
+ xlab="Квантили нормального распределения",
+ ylab="Наблюдаемые квантили",
+ main="Сравнение квантилей эмпирического и нормального распределений")
```



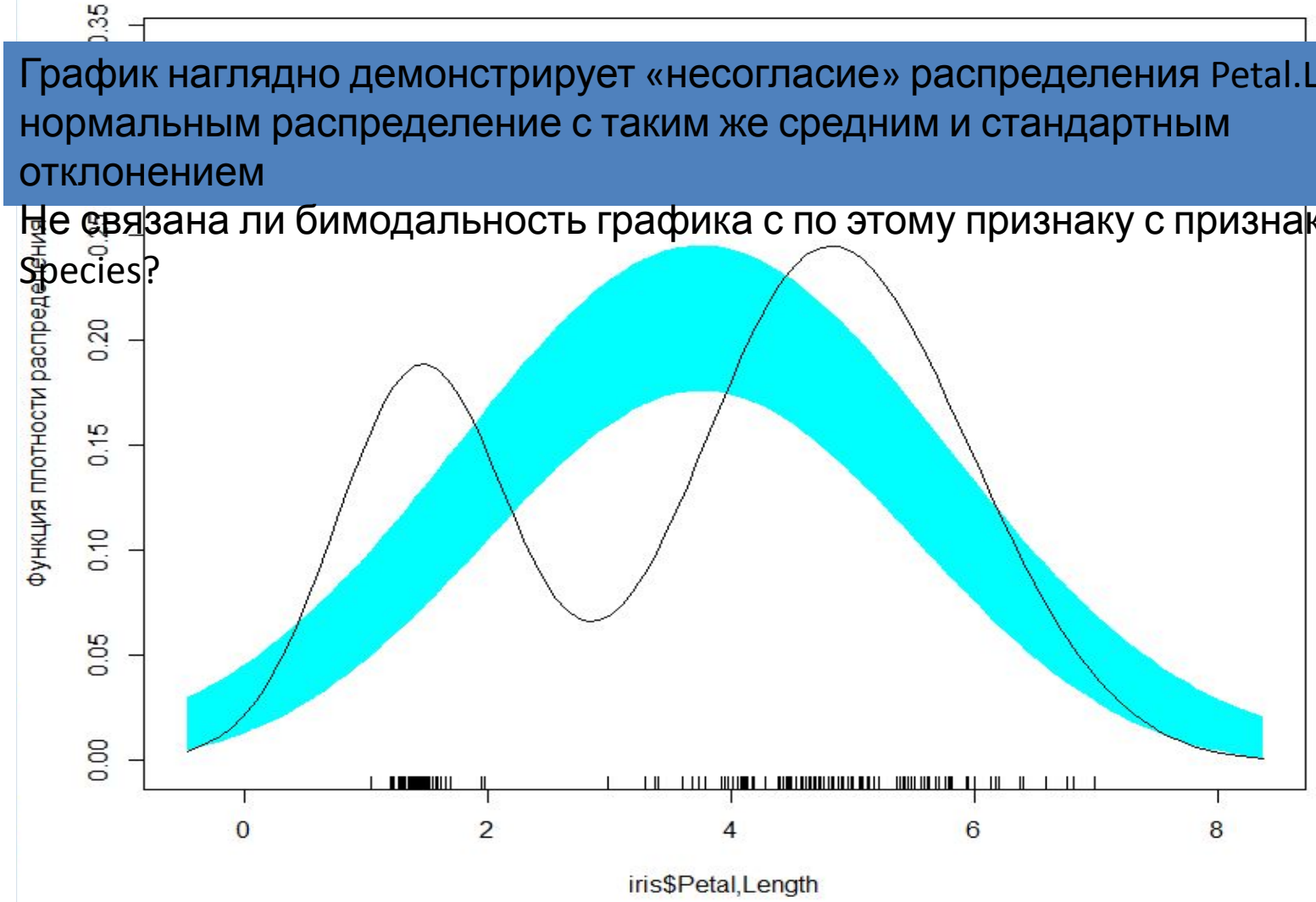
```
> library(sm)
> sm.density(iris$Petal.Length, model = "Normal", xlab="iris$Petal,Length", ylab="Функция плотности распределения")
```



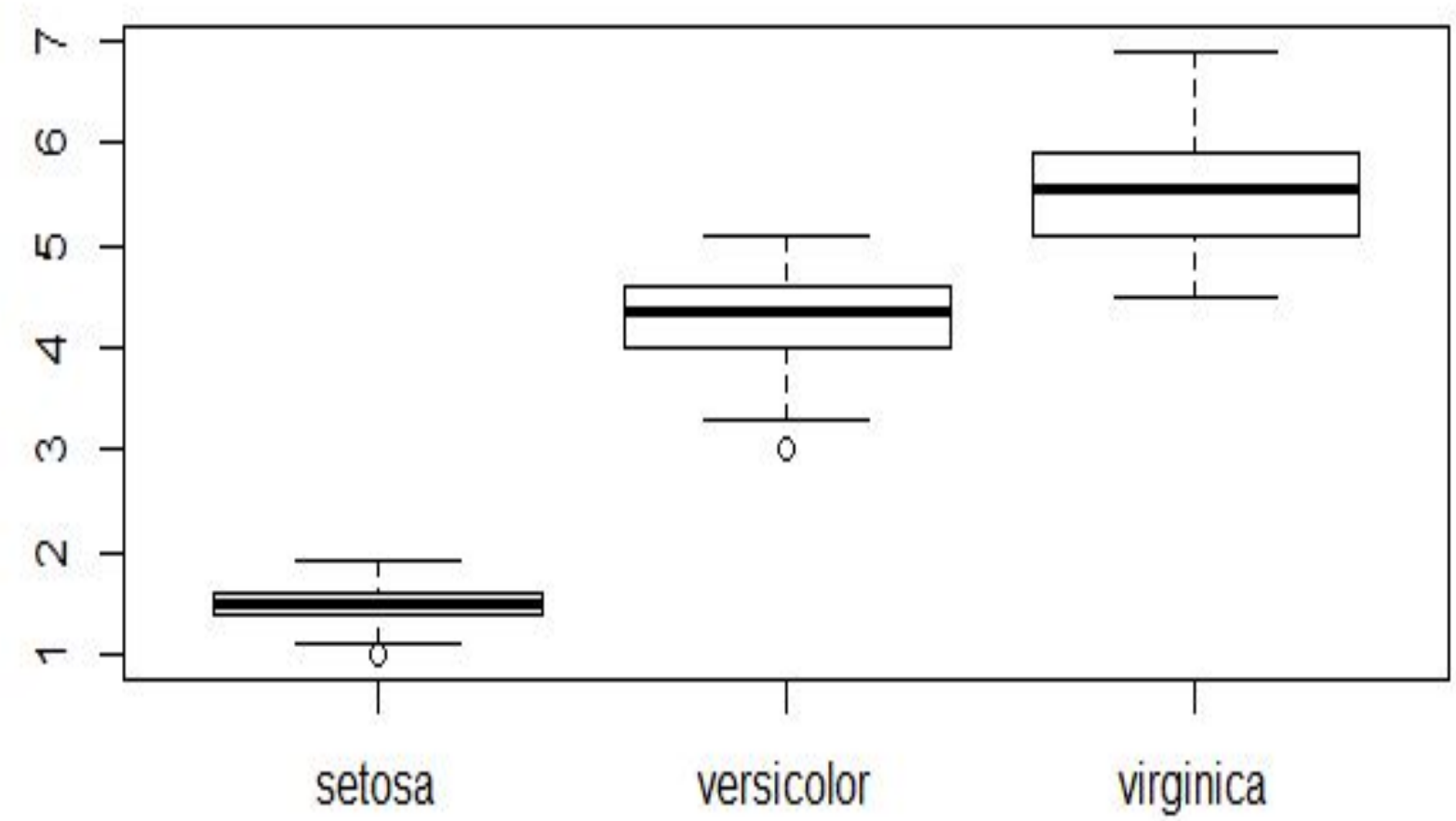
```
> library(sm)
> sm.density(iris$Petal.Length, model = "Normal", xlab="iris$Petal,Length", ylab="Функция плотности распределения")
```

График наглядно демонстрирует «несогласие» распределения Petal.Length с нормальным распределение с таким же средним и стандартным отклонением

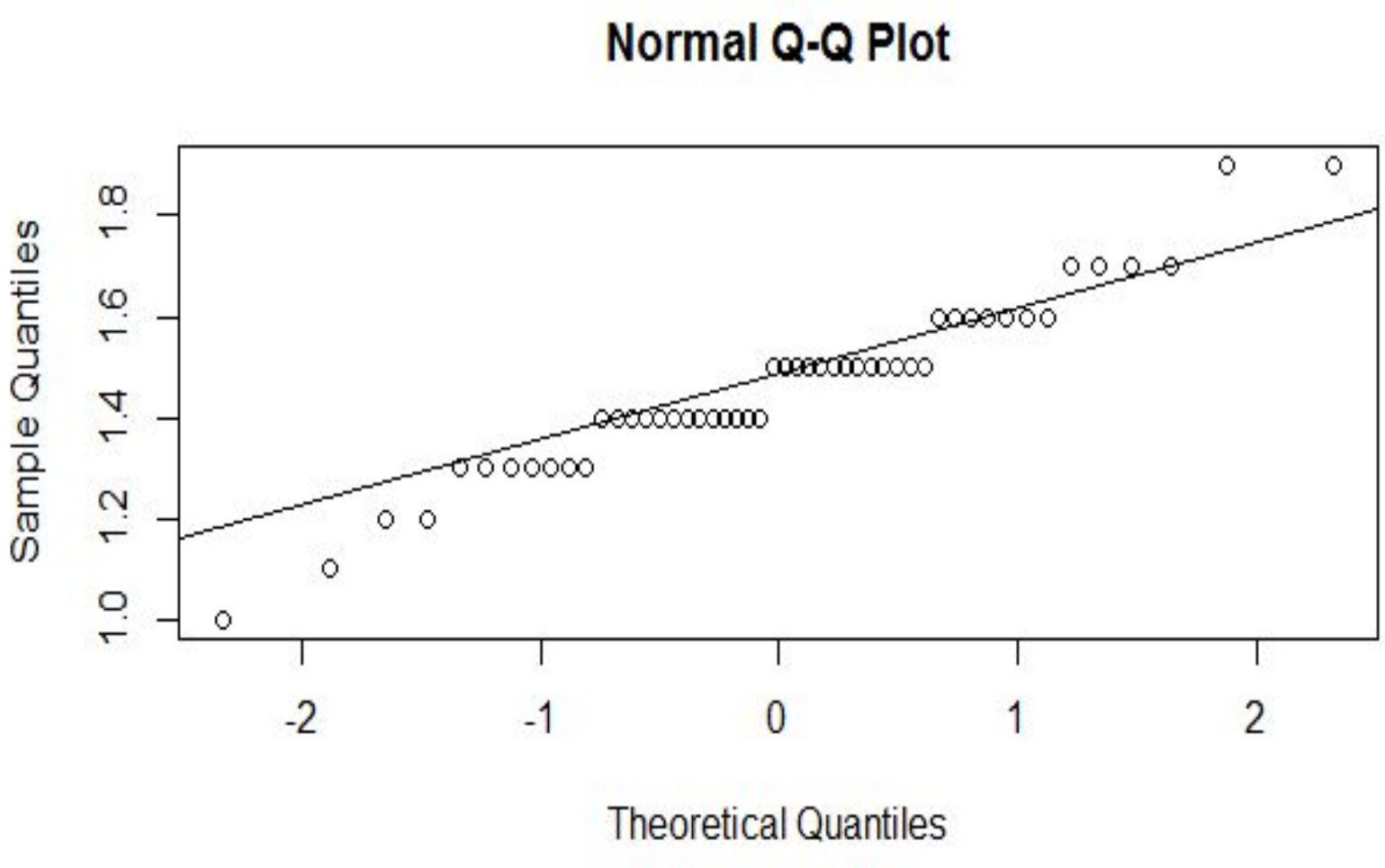
Не связана ли бимодальность графика с по этому признаку с признаком Species?



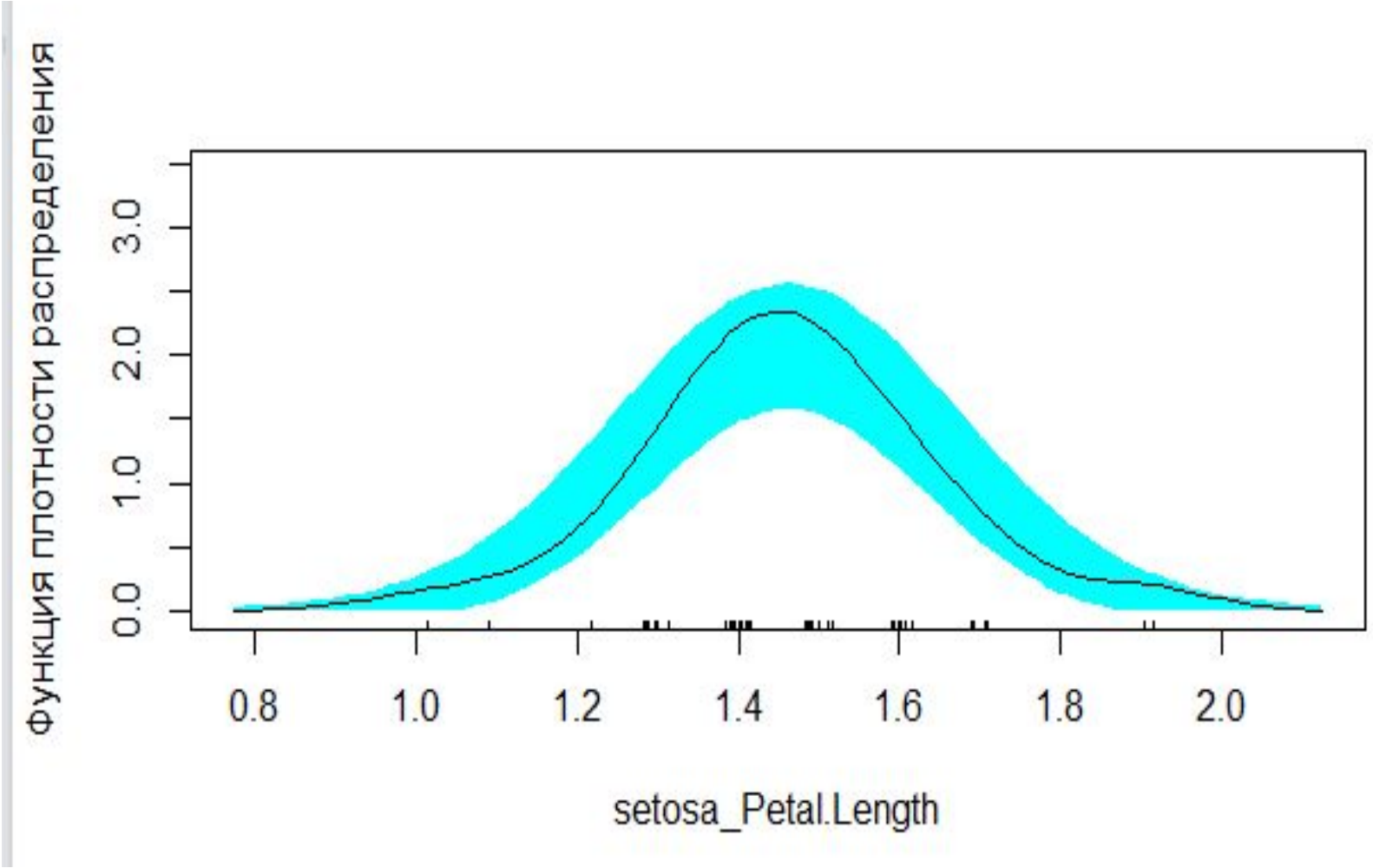
```
> boxplot(iris$Petal.Length~iris$Species)
```



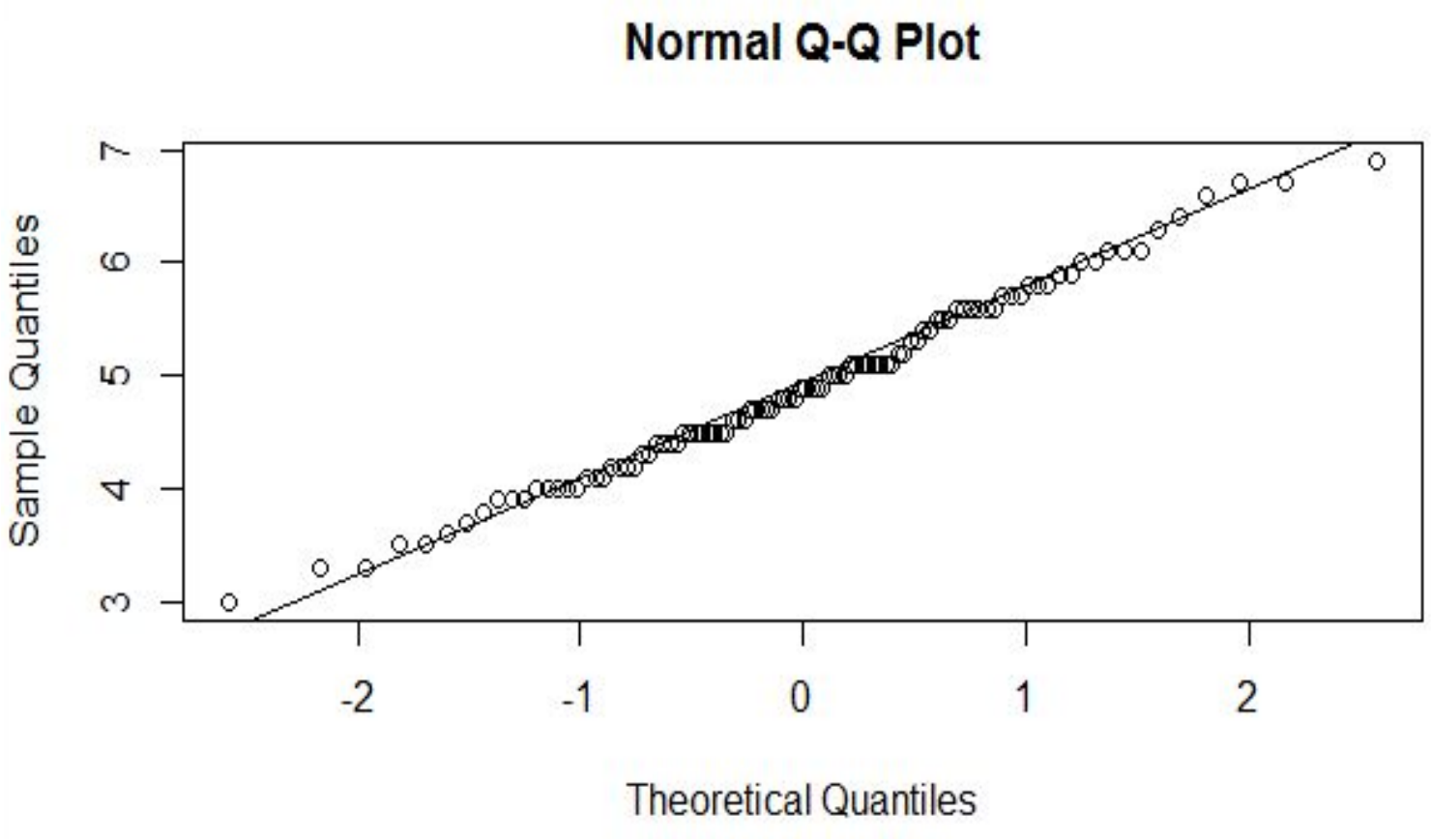
```
> setosa<-subset(iris, iris$Species=="setosa")$Petal.Length
> qqnorm(setosa)
> qqline(setosa)
```



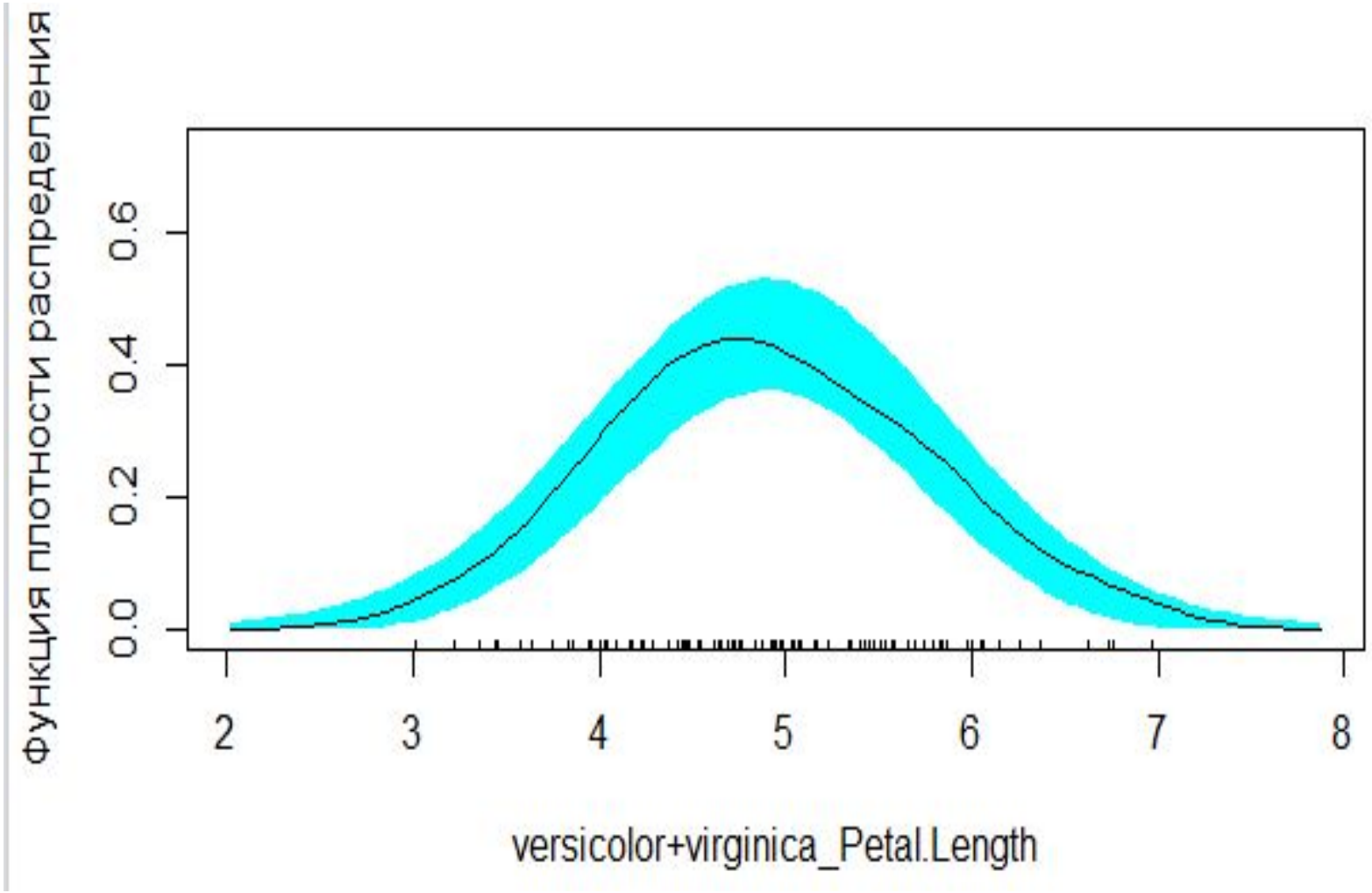
```
> sm.density(setosa, model = "Normal", xlab="setosa_Petal.Length",
+ ylab="Функция плотности распределения")
```



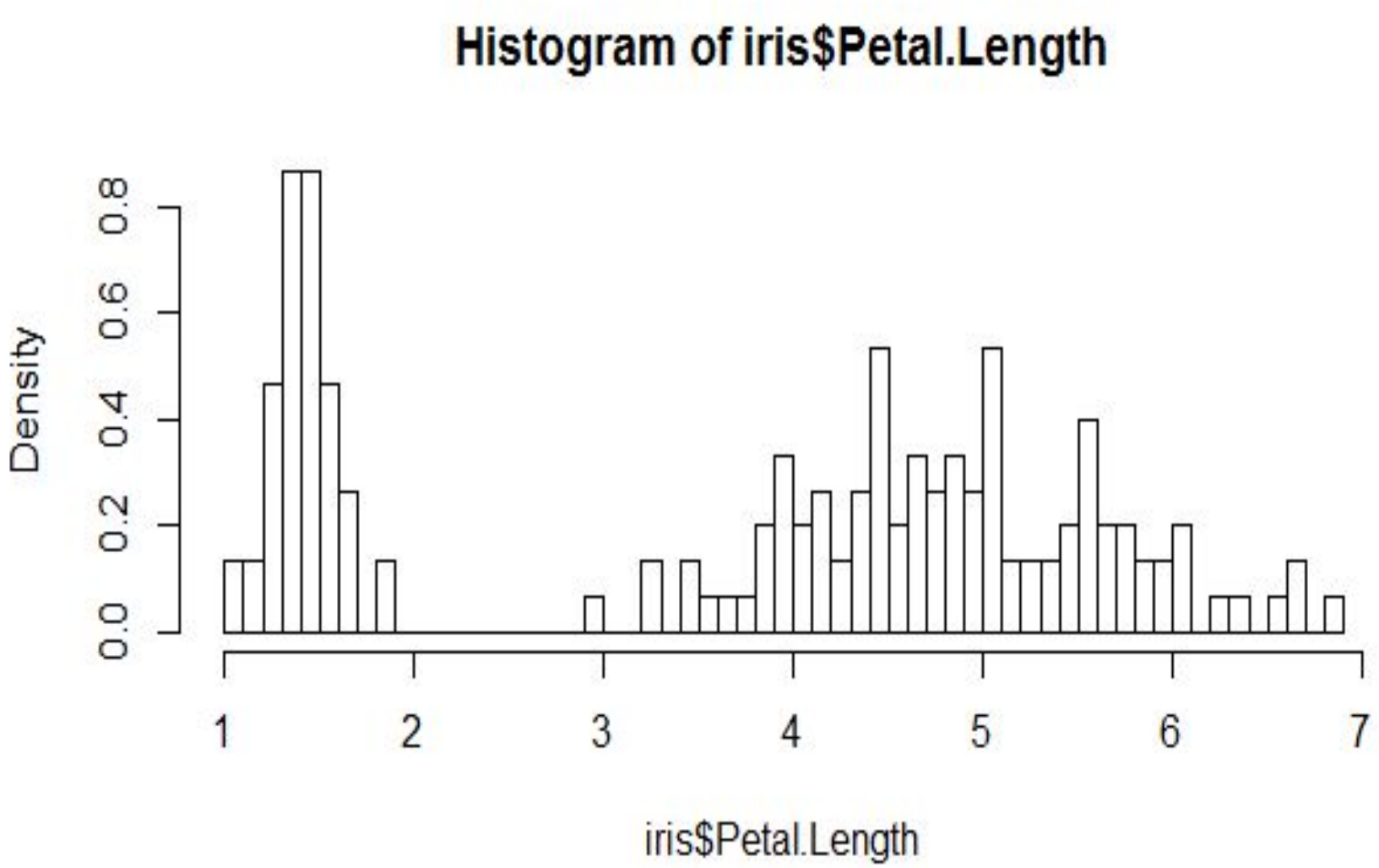
```
> versicolor<-subset(iris, iris$Species=="versicolor")$Petal.Length
> virginica<-subset(iris, iris$Species=="virginica")$Petal.Length
> ver_vir<-c(versicolor,virginica)
> qqnorm(ver_vir)
> qqline(ver_vir)
```



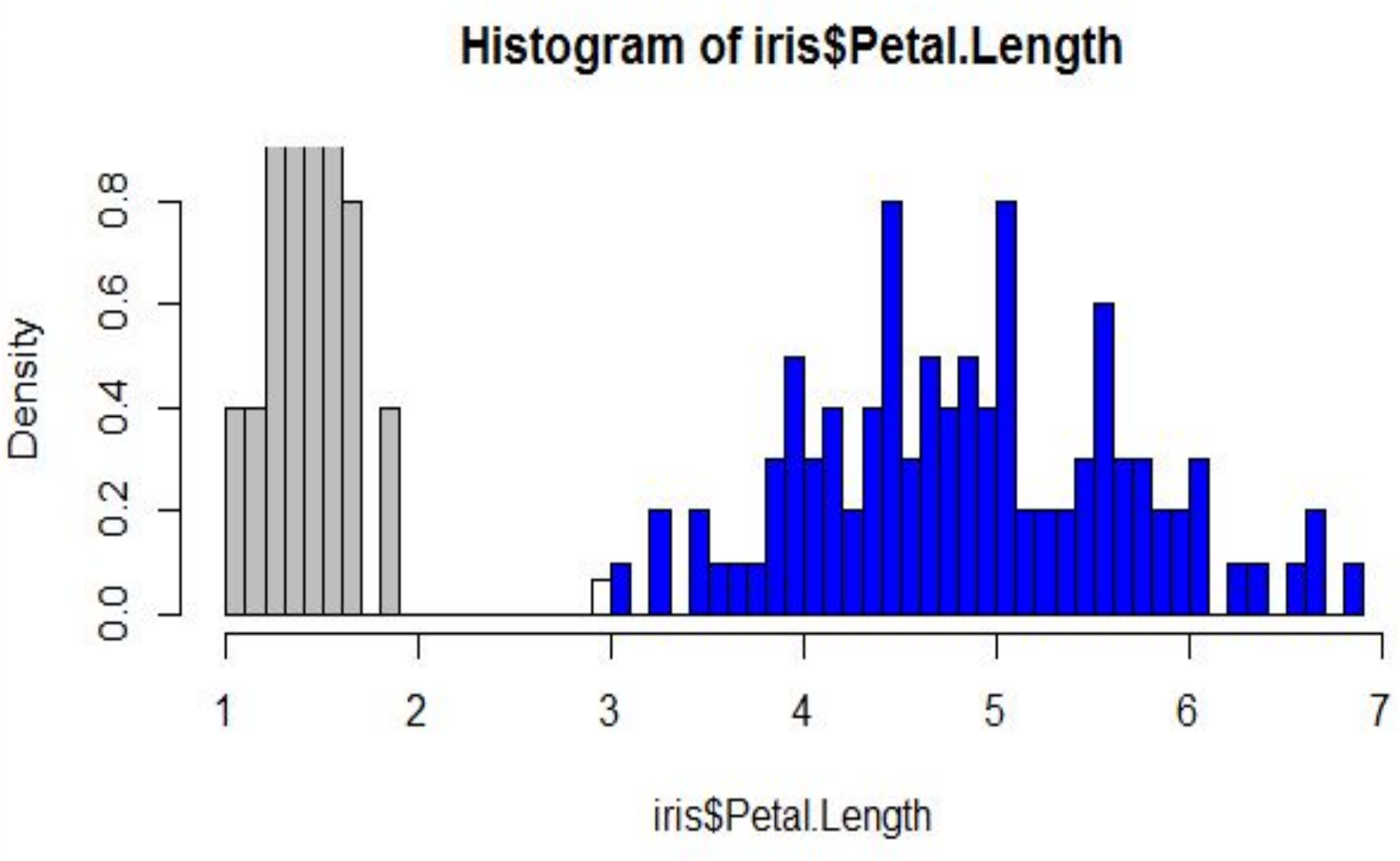
```
> sm.density(ver_vir, model = "Normal", xlab="versicolor+virginica_Petal.Length",
+ ylab="Функция плотности распределения")
```



```
> hist(iris$Petal.Length, breaks=50, freq=F)
```



```
> hist(setosa, breaks=8, freq=F, col="grey", add=T )
> hist(ver_vir, breaks=50, freq=F, col="blue", add=T )
```



```
> boxplot(iris$Petal.Length,setosa,ver_vir)
> legend("top",c("1-iris,2-setosa,3-ver+vir"))
```