Харинаев Артём 316 группа 11.10.21

1. Оценка максимального правдоподобия

Построим оценку для дисперсии нормального распределения

```
set.seed(151021)
sample <- rnorm(1, mean=10, sd=2)
minus_likelihood <- function(theta){
  dnorm(sample, mean=10, sd=theta)*-1
}
```

```
nlm(minus_likelihood, p=1, stepmax=0.5)$estimate
```

```
## [1] 1.937385
```

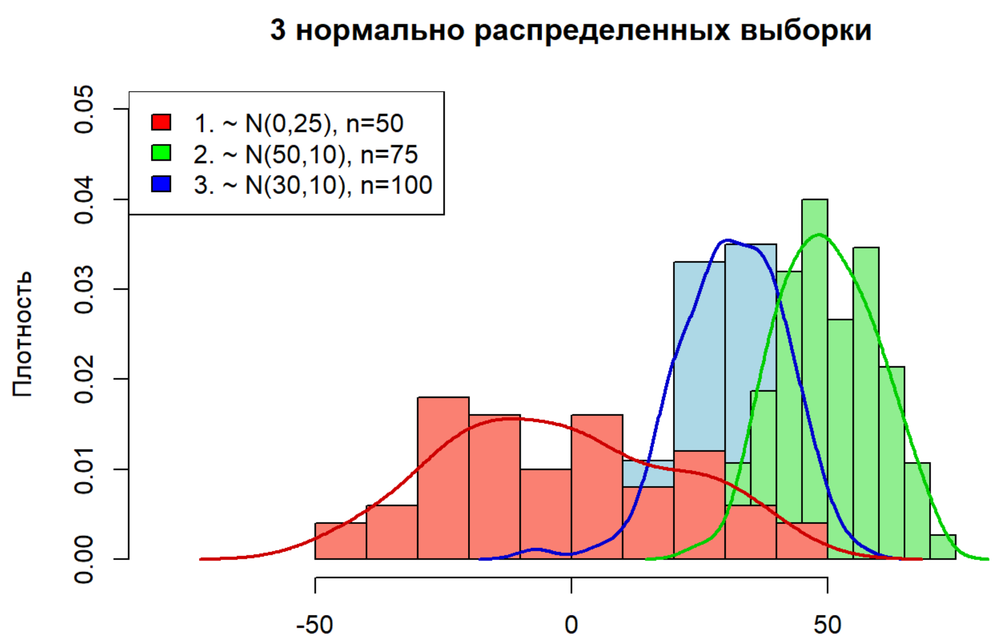
Оценка довольно близка к истинному значению дисперсии (2)

2. Анализ выборок из нормального распределения

2.1 Выборки малого размера

2.1.1 Генерирование и визуализация

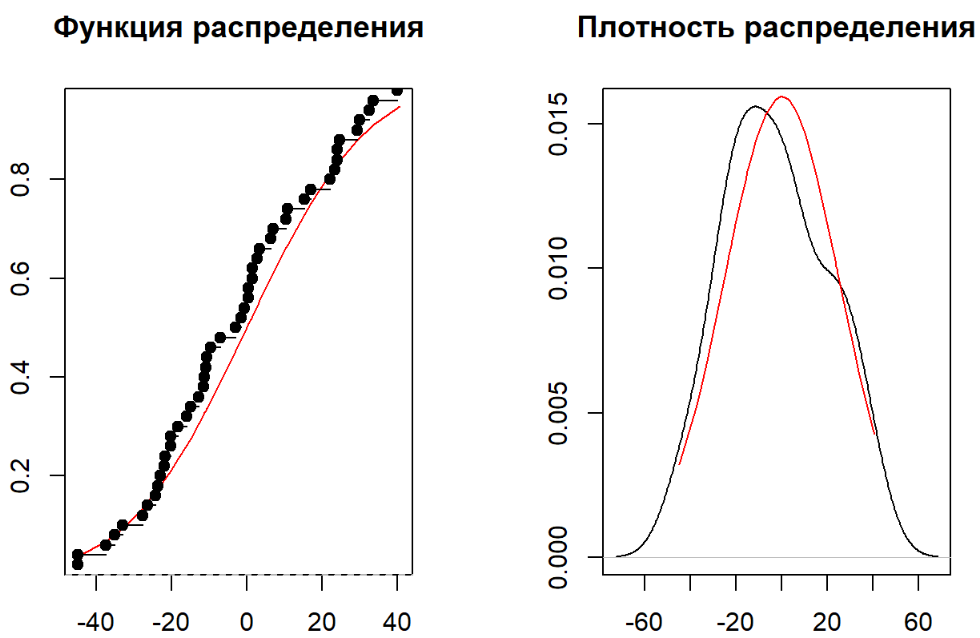
```
set.seed(151021)
s.1 <- rnorm(50, 0, 25)
s.2 <- rnorm(75, 50, 10)
s.3 <- rnorm(100, 30, 10)
hist(s.3, col='lightblue', xlim=c(-80,80), ylim=c(0,0.05), freq=FALSE, main='3 нормально распределенных в
ыборки', xlab='', ylab='Плотность')
hist(s.2, col='lightgreen', add=TRUE, freq=FALSE)
hist(s.1, col='salmon', add=TRUE, freq=FALSE)
lines(density(s.3), col='blue3', lwd=2)
lines(density(s.2), col='green3', lwd=2)
lines(density(s.1), col='red3', lwd=2)
legend('topleft',
       legend=c('1. ~ N(0,25), n=50', '2. ~ N(50,10), n=75', '3. ~ N(30,10), n=100'),
       fill=c('red','green', 'blue'))
```



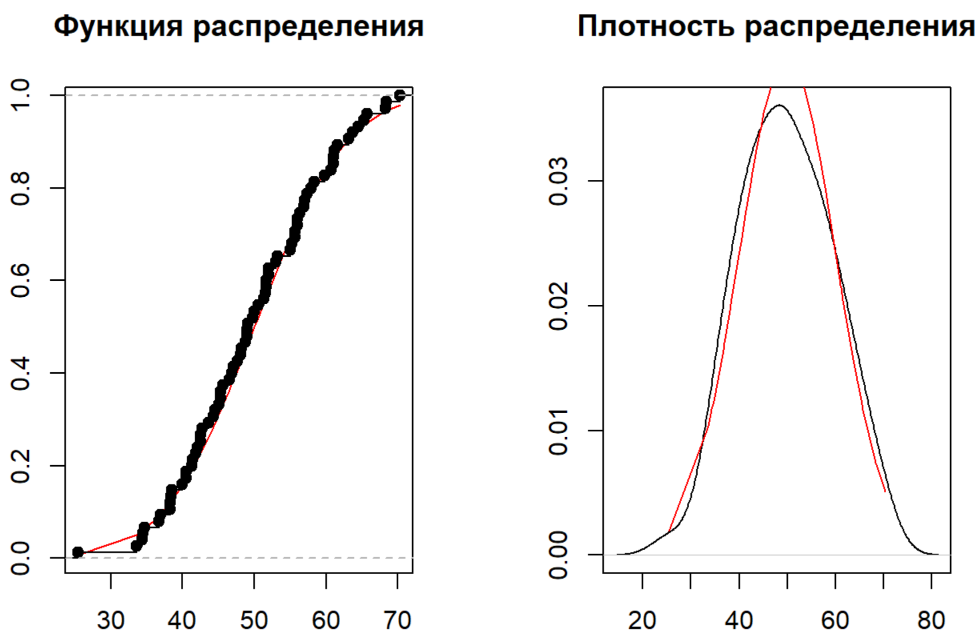
2.1.2 Теоретические и эмпирические функции рапределения и плотности

```
theor_empiric <- function(cur, mean, sd){
  par(mfrow=c(1,2))
  plot(sort(cur), pnorm(sort(cur),mean,sd), type='l', col='red', main='Функция распределения',
        xlab='', ylab='')
  plot(ecdf(cur), add=TRUE)
  plot(density(cur), main='Плотность распределения', xlab='', ylab='')
  lines(sort(cur), dnorm(sort(cur),mean,sd), type='l', col='red')
}
```

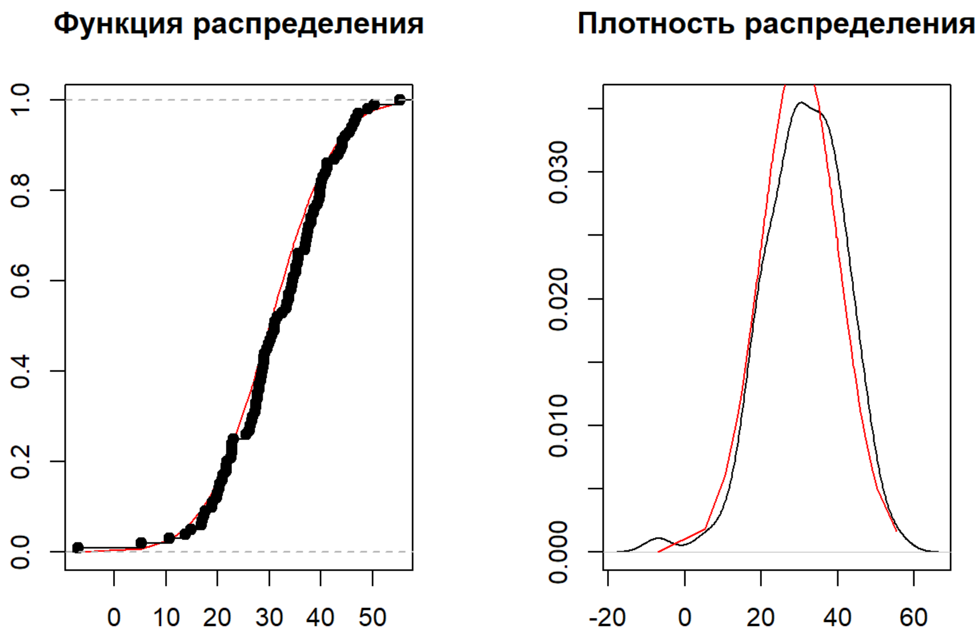
```
theor_empiric(s.1, 0, 25)
```




```
theor_empiric(s.2, 50, 10)
```




```
theor_empiric(s.3, 30, 10)
```

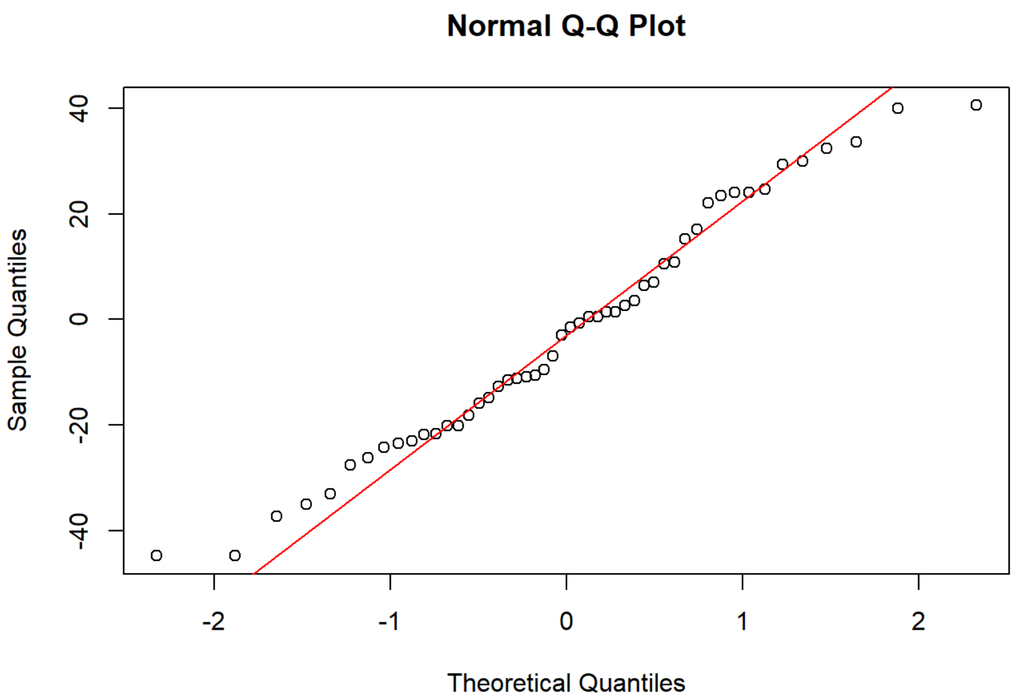


Даже на небольших данных заметно, что с увеличением объема выборки, эмпирические данные приближаются к теоретическим

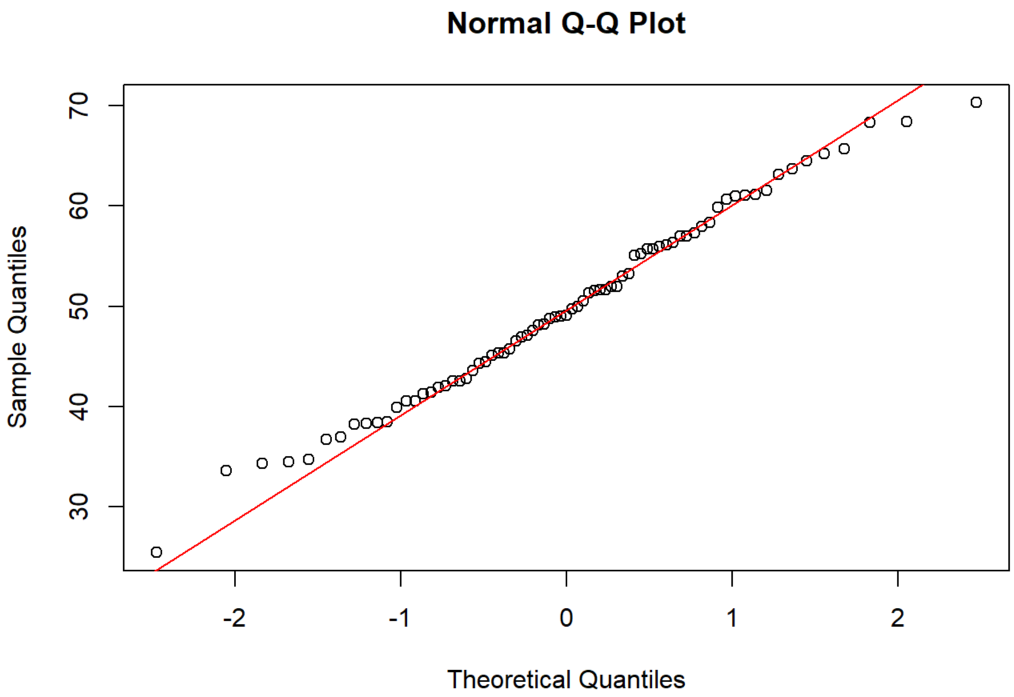
2.1.3 Квантили

```
quantile <- function(x){
  qqnorm(x)
  qqline(x, col='red')
}
```

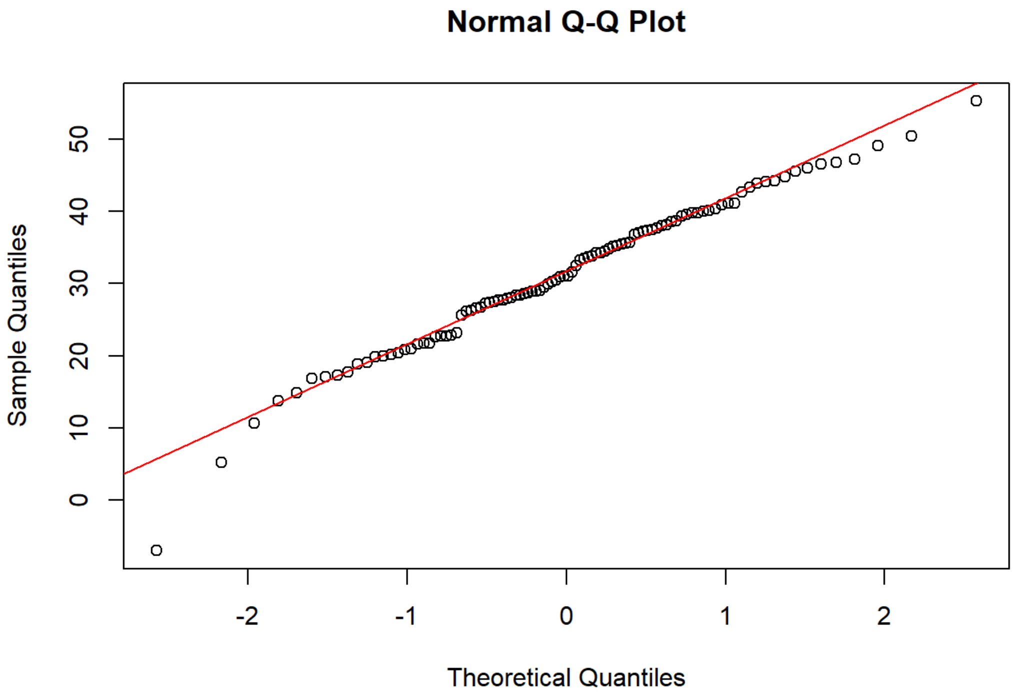
```
quantile(s.1)
```




```
quantile(s.2)
```




```
quantile(s.3)
```

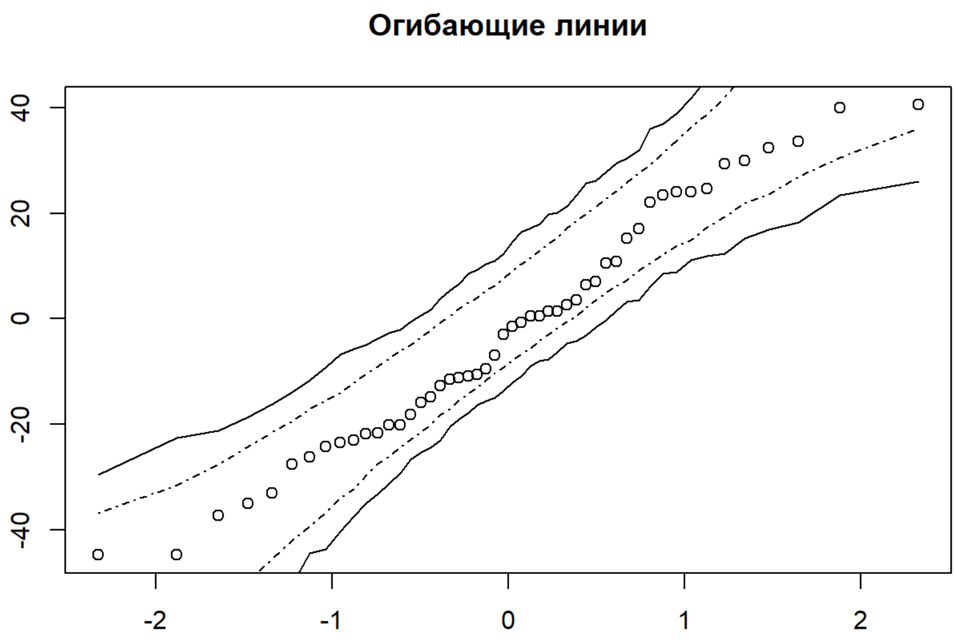


Здесь так же заметна разница от размера выборки (чем больше объем, тем ближе точки к прямой)

2.1.4 Огибающие

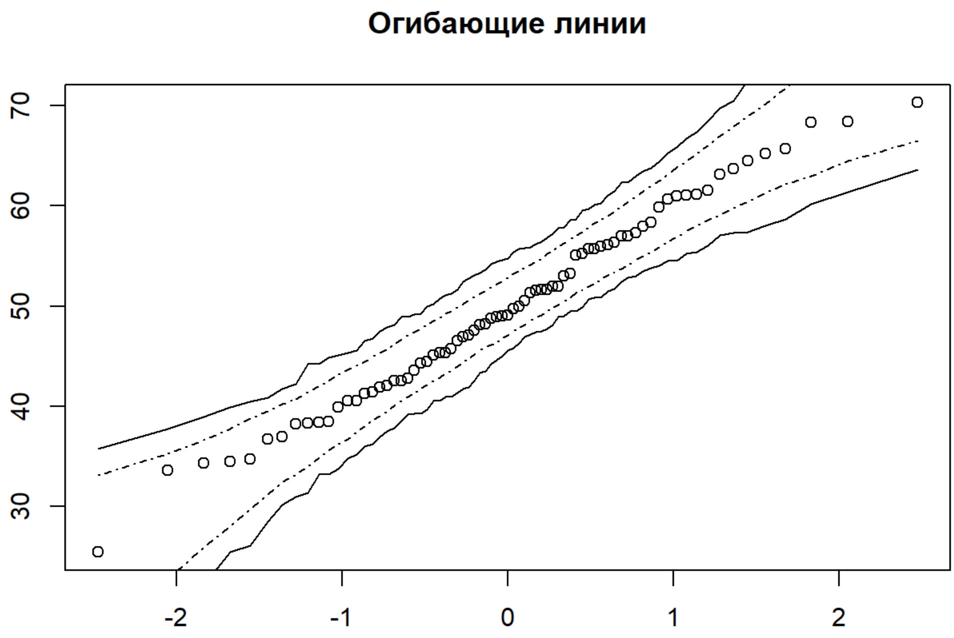
```r
library(boot)
envelopes <- function(s, mean, sd, r){
  x.qq <- qqnorm(s, plot.it = FALSE)
  x.qq <- lapply(x.qq, sort)
  x.gen <- function(dat, mle) rnorm(length(dat), mle[1], mle[2])
  x.qqboot <- boot(s, sort, R = r, sim = "parametric", ran.gen = x.gen, mle=c(mean, sd))
  x.env <- envelope(x.qqboot)
  plot(x.qq, main='Огибающие линии', xlab='', ylab='')
  lines(x.qq$x, x.env$point[1, ], lty = 4)
  lines(x.qq$x, x.env$point[2, ], lty = 4)
  lines(x.qq$x, x.env$overall[1, ], lty = 1)
  lines(x.qq$x, x.env$overall[2, ], lty = 1)
}
```

```r
envelopes(s.1, 0, 25, 2000)
```

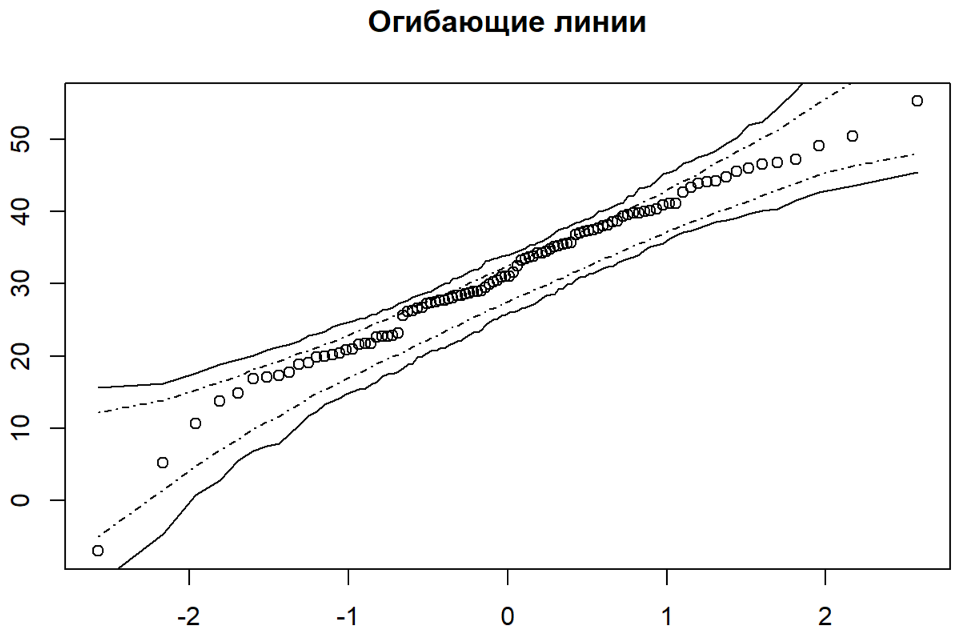



```r
envelopes(s.2, 50, 10, 2000)
```

Огибающие линии




```
envelopes(s.3, 30, 10, 2000)
```



2.1.5 Тесты нормальности

2.1.5.1 Колмогорова-Смирнова

```
ks.test(s.1, pnorm, mean=0, sd=25)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  s.1
## D = 0.108, p-value = 0.5672
## alternative hypothesis: two-sided
```

```
ks.test(s.2, pnorm, mean=50, sd=10)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  s.2
## D = 0.047948, p-value = 0.9921
## alternative hypothesis: two-sided
```

```
ks.test(s.3, pnorm, mean=30, sd=10)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  s.3
## D = 0.097803, p-value = 0.2943
## alternative hypothesis: two-sided
```

Значения статистики Колмогорова-Смирнова достаточно малы, значит выборки действительно из нормального распределения

2.1.5.2 Шапиро-Уилка

```
library(nortest)
shapiro.test(s.1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  s.1
## W = 0.97386, p-value = 0.3299
```

```
shapiro.test(s.2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  s.2
## W = 0.9896, p-value = 0.8024
```

```
shapiro.test(s.3)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  s.3
## W = 0.98052, p-value = 0.1458
```

Тесты Шапиро-Уилка дают результат близкий к 1, что тоже свидетельствует, что выборки из нормального распределения

2.1.5.3 Андерсона-Дарлинга

```
ad.test(s.1)
```

```
##
##  Anderson-Darling normality test
##
## data:  s.1
## A = 0.36838, p-value = 0.4161
```

```
ad.test(s.2)
```

```
##
##  Anderson-Darling normality test
##
## data:  s.2
## A = 0.21112, p-value = 0.8529
```

```
ad.test(s.3)
```

```
##
##  Anderson-Darling normality test
##
## data:  s.3
## A = 0.3207, p-value = 0.5264
```

2.1.5.4 Крамера-фон Мизеса

```
cvm.test(s.1)
```

```
##
##  Cramer-von Mises normality test
##
## data:  s.1
## W = 0.055597, p-value = 0.4252
```

```
cvm.test(s.2)
```

```
##
##  Cramer-von Mises normality test
##
## data:  s.2
## W = 0.030116, p-value = 0.8432
```

```
cvm.test(s.3)
```

```
##
##  Cramer-von Mises normality test
##
## data:  s.3
## W = 0.044388, p-value = 0.5971
```

Статистика достаточно близка к 0, что говорит о нормальности распределения

2.1.5.4 Лиллифорса (вариация Колмогорова-Смирнова именно для нормального распределения)

```
lillie.test(s.1)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  s.1
## D = 0.084471, p-value = 0.4991
```

```
lillie.test(s.2)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  s.2
## D = 0.053154, p-value = 0.8656
```

```
lillie.test(s.3)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  s.3
## D = 0.050117, p-value = 0.7767
```

Статистика близка к 0, значит распределения имеют нормальный вид

2.1.5.5 Шапиро-Франчия

```
sf.test(s.1)
```

```
##
##  Shapiro-Francia normality test
##
## data:  s.1
## W = 0.98186, p-value = 0.5425
```

```
sf.test(s.2)
```

```
##
##  Shapiro-Francia normality test
##
## data:  s.2
## W = 0.99207, p-value = 0.8617
```

```
sf.test(s.3)
```

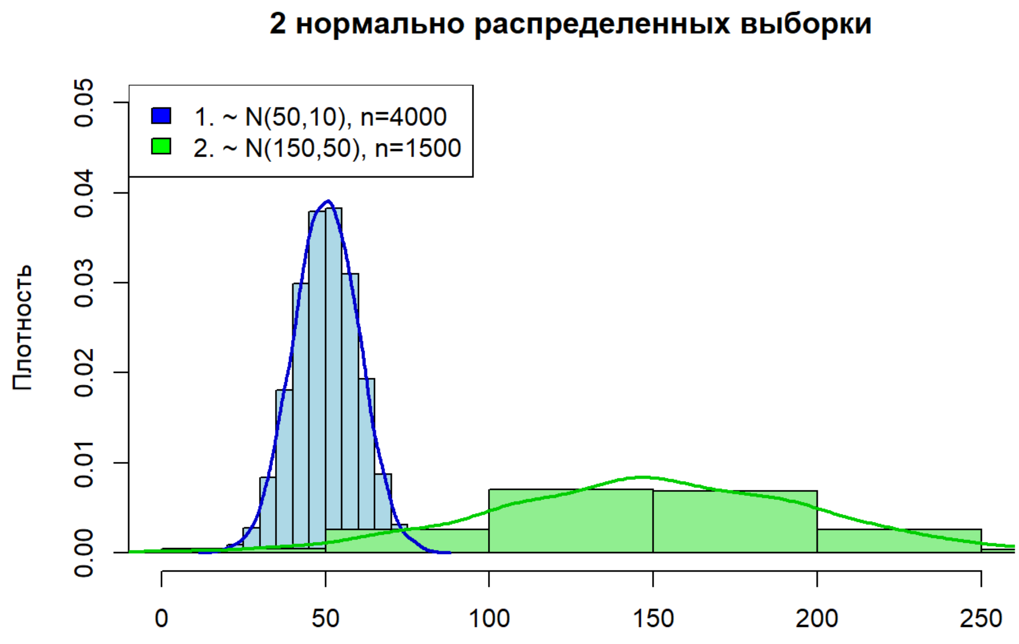
```
##
##  Shapiro-Francia normality test
##
## data:  s.3
## W = 0.97674, p-value = 0.06914
```

Результат близок к 1, что говорит о нормальности распределения

2.2 Выборки большого объема

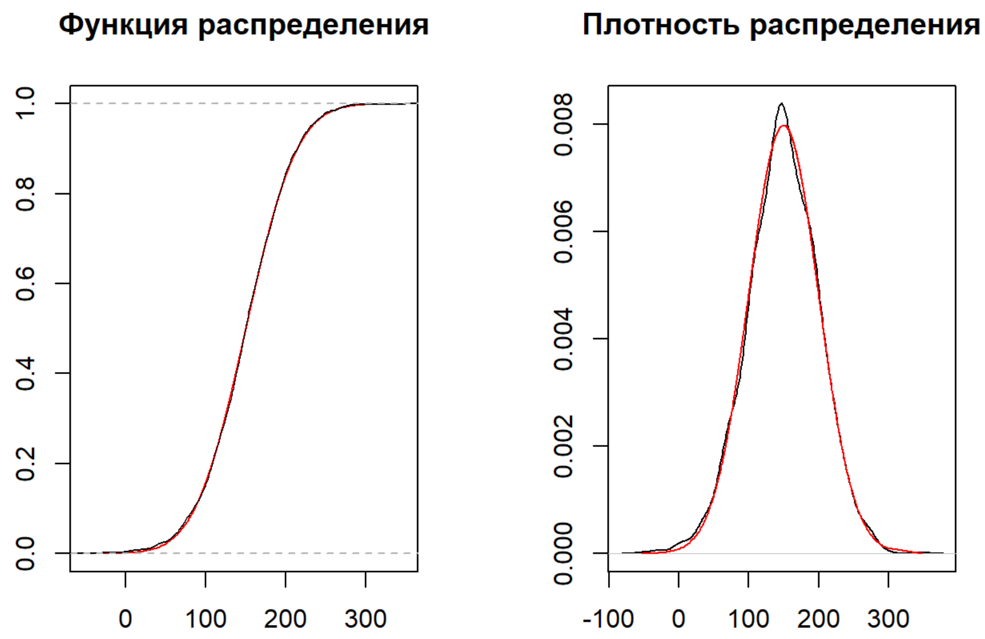
2.2.1 Генерирование и визуализация

```
set.seed(151021)
b.1 <- rnorm(2000, 150, 50)
b.2 <- rnorm(4000, 50, 10)
hist(b.2, col='lightblue', xlim=c(0,250), ylim=c(0,0.05), freq=FALSE, main='2 нормально распределенных вы
борки', xlab='', ylab='Плотность')
hist(b.1, col='lightgreen', add=TRUE, freq=FALSE)
lines(density(b.2), col='blue3', lwd=2)
lines(density(b.1), col='green3', lwd=2)
legend('topleft',
       legend=c('1. ~ N(50,10), n=4000', '2. ~ N(150,50), n=1500'),
       fill=c('blue', 'green'))
```

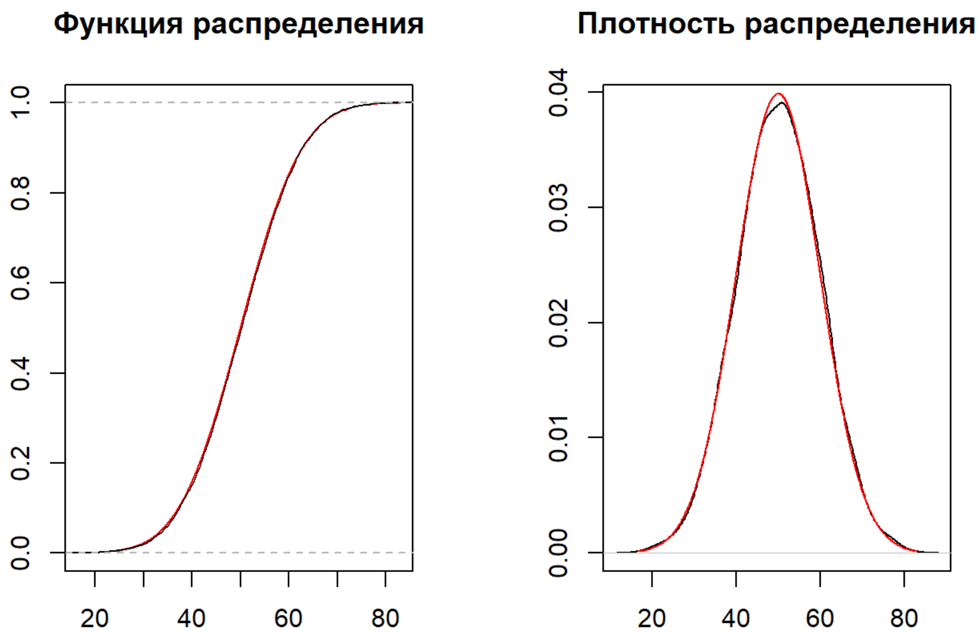


2.2.2 Теоретические и эмпирические функции рапределения и плотности

```
theor_empiric(b.1, 150, 50)
```




```
theor_empiric(b.2, 50, 10)
```

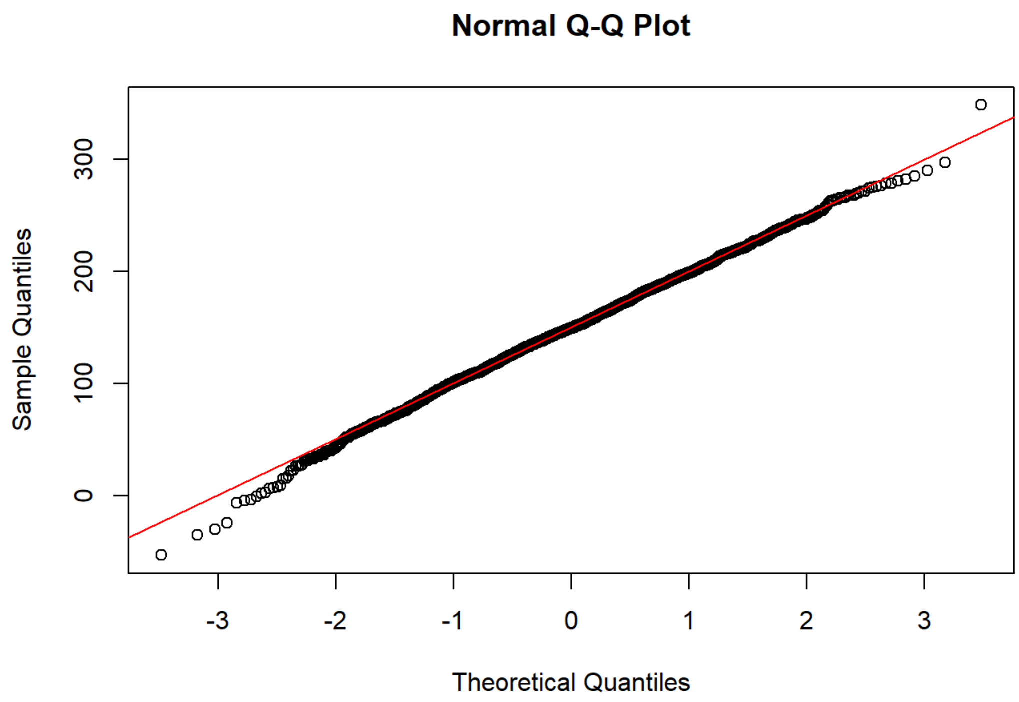


С увеличением объема эмпирические значение значительно приблизились к теоретическим

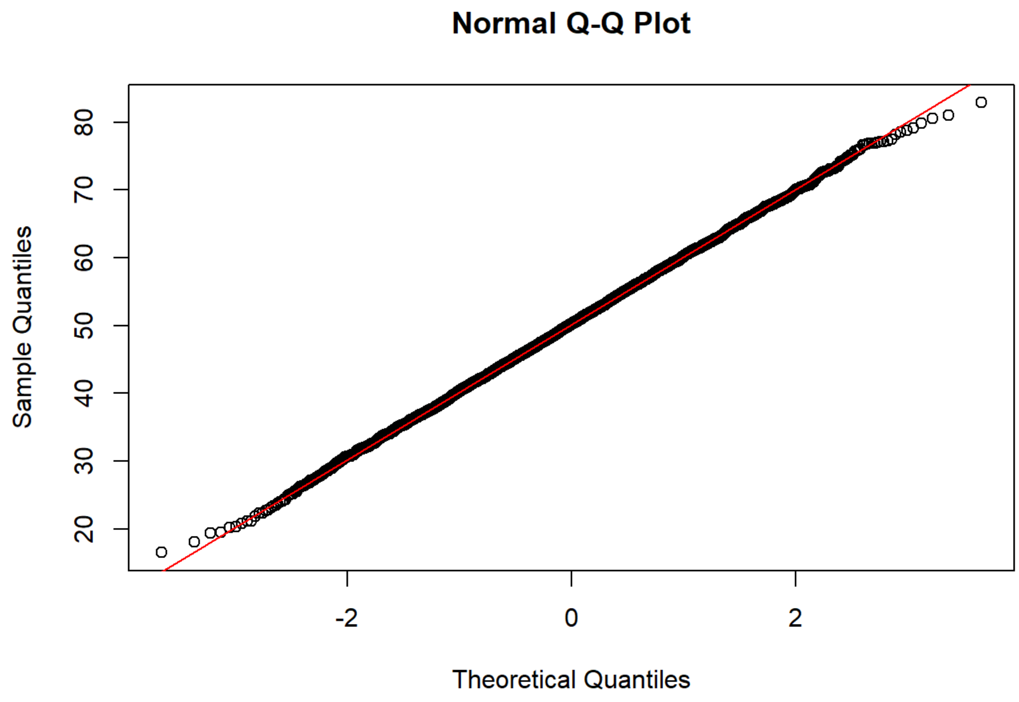
2.2.3 Квантили

```
quantile <- function(x){
  qqnorm(x)
  qqline(x, col='red')
}
```

```
quantile(b.1)
```



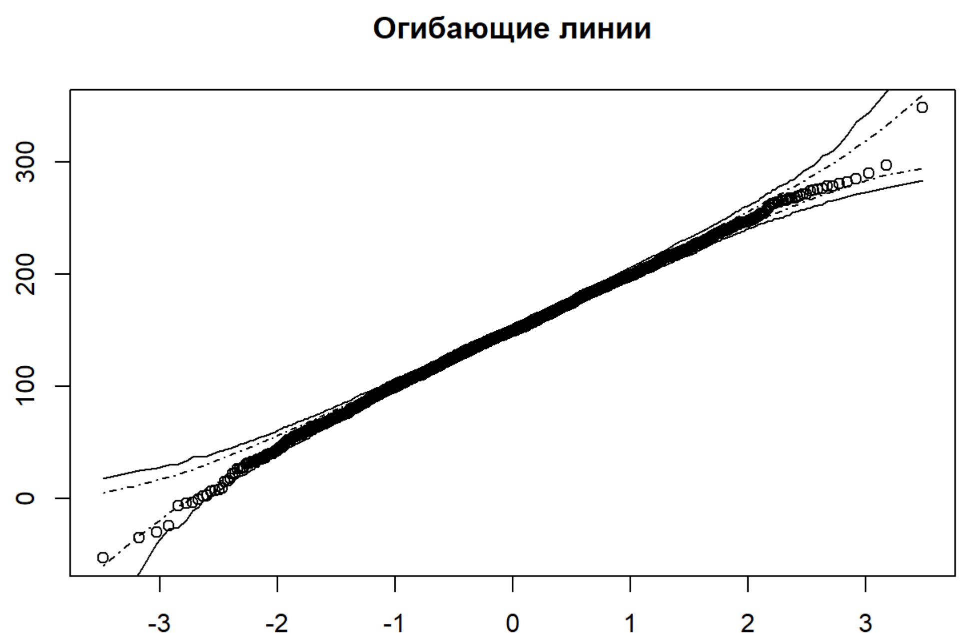

```
quantile(b.2)
```



На Q-Q графиках так же заметно, что с увеличением выборки точки ближе прижимаются к прямой

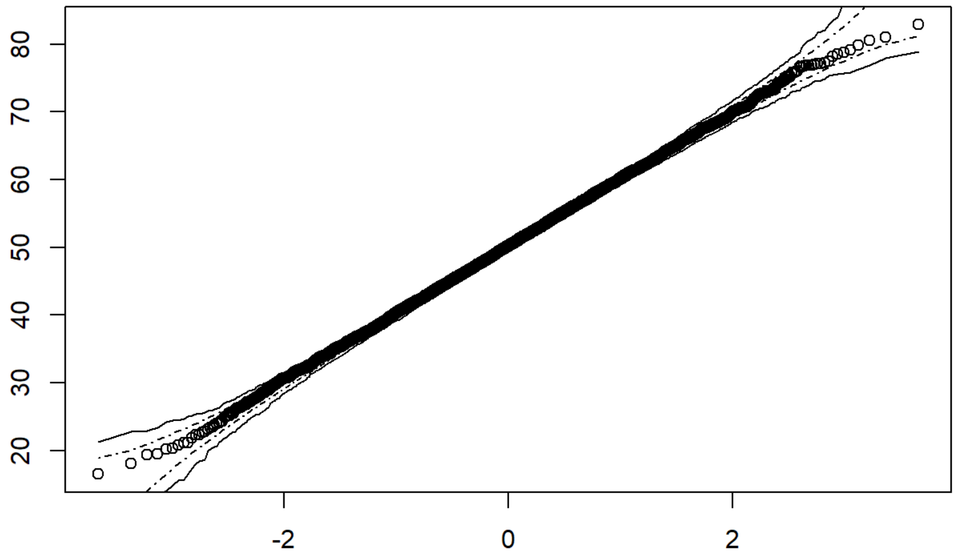
2.2.4 Огибающие

```
envelopes(b.1, 150, 50, 5000)
```




```
envelopes(b.2, 50, 10, 5000)
```

Огибающие линии



2.2.5 Тесты нормальности

2.2.5.1 Колмогорова-Смирнова

```
ks.test(b.1, pnorm, mean=150, sd=50)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  b.1
## D = 0.0145, p-value = 0.7944
## alternative hypothesis: two-sided
```

```
ks.test(b.2, pnorm, mean=50, sd=10)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  b.2
## D = 0.011115, p-value = 0.7062
## alternative hypothesis: two-sided
```

Значения статистики Колмогорова-Смирнова достаточно малы, значит выборки действительно из нормального распределения

2.2.5.2 Шапиро-Уилка

```
library(nortest)
shapiro.test(b.1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  b.1
## W = 0.99806, p-value = 0.01704
```

```
shapiro.test(b.2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  b.2
## W = 0.99973, p-value = 0.9116
```

Тесты Шапиро-Уилка дают результат близкий к 1, что тоже свидетельствует, что выборки из нормального распределения

2.2.5.3 Андерсона-Дарлинга

```
ad.test(b.1)
```

```
##
##  Anderson-Darling normality test
##
## data:  b.1
## A = 0.65818, p-value = 0.08567
```

```
ad.test(b.2)
```

```
##
##  Anderson-Darling normality test
##
## data:  b.2
## A = 0.14036, p-value = 0.9739
```

2.2.5.4 Крамера-фон Мизеса

```
cvm.test(b.1)
```

```
##
##  Cramer-von Mises normality test
##
## data:  b.1
## W = 0.10527, p-value = 0.09514
```

```
cvm.test(b.2)
```

```
##
##  Cramer-von Mises normality test
##
## data:  b.2
## W = 0.020743, p-value = 0.9617
```

Статистика достаточно близка к 0, что говорит о нормальности распределения

2.2.5.5 Лиллифорса (вариация Колмогорова-Смирнова именно для нормального распределения)

```
lillie.test(b.1)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  b.1
## D = 0.018715, p-value = 0.09157
```

```
lillie.test(b.2)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  b.2
## D = 0.0060825, p-value = 0.975
```

Статистика близка к 0, значит распределения имеют нормальный вид

2.2.5.6 Шапиро-Франчия

```
sf.test(b.1)
```

```
##
##  Shapiro-Francia normality test
##
## data:  b.1
## W = 0.99786, p-value = 0.008931
```

```
sf.test(b.2)
```

```
##
##  Shapiro-Francia normality test
##
## data:  b.2
## W = 0.99982, p-value = 0.9825
```

Результат близок к 1, что говорит о нормальности распределения

2.3 Анализ своих данных

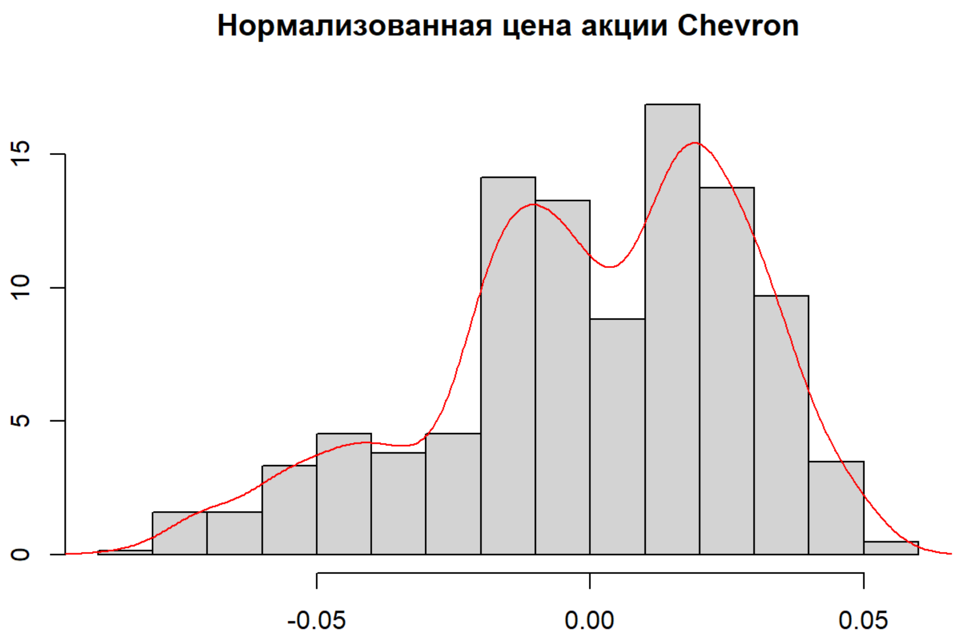
Нормализованные данные по цене акций компании Chevron (CVX)

```
data <- read.csv(file='..\\dataset.csv')
data$date <- as.Date(data$date)
```

```
chevron <- subset(data, data['Name'] == 'CVX')
chevron$norm <- sapply(chevron$open,FUN=function(x){(x-mean(chevron$open))/(sd(chevron$open)*sqrt(nrow(ch
evron)))})
```

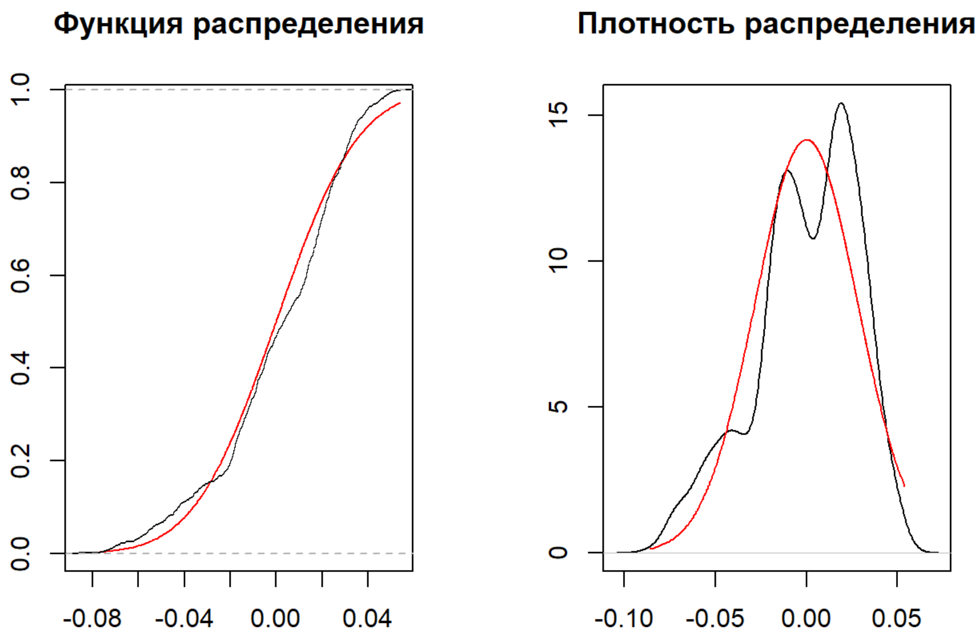
2.3.1 Визуализация данных

```
hist(chevron$norm, freq=FALSE, xlab='', ylab='', main='Нормализованная цена акции Chevron')
lines(density(chevron$norm), col='red')
```



2.3.2 Теоретическая и эмпирическая функция рапределения и плотности

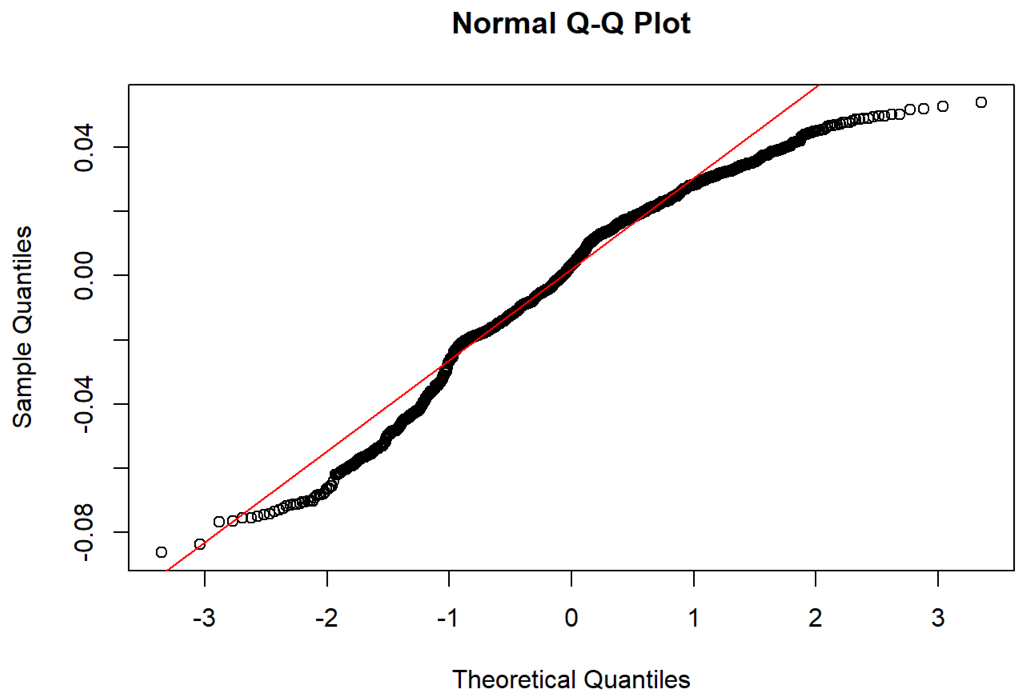
```
theor_empiric(chevron$norm, mean(chevron$norm), sd(chevron$norm))
```



Визуально нормализация дает результат довольно близкий к теоретическому распределению

2.3.3 Квантили

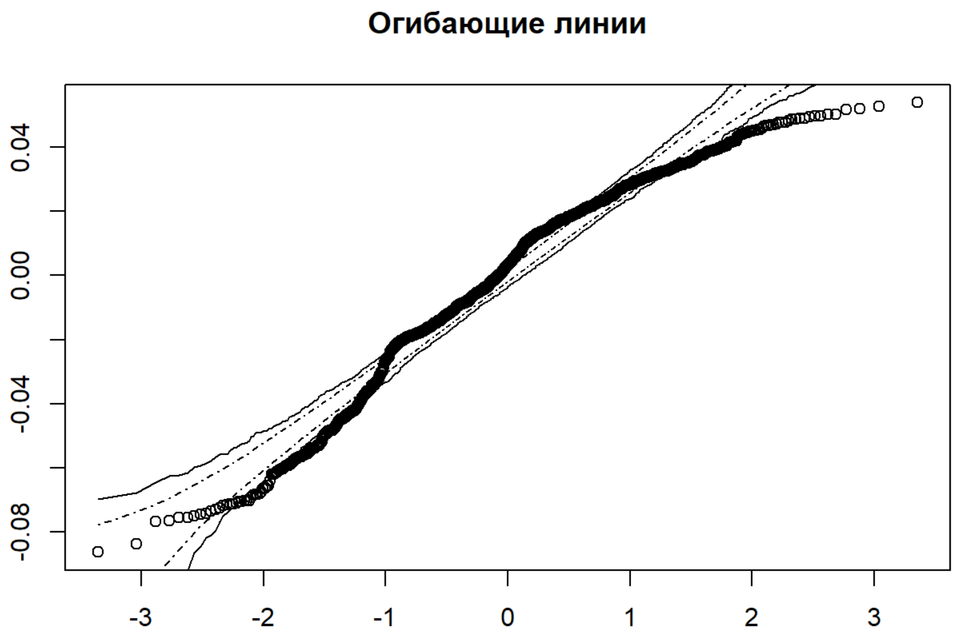
```
quantile(chevron$norm)
```



На Q-Q графиках так же заметно, что значения, близкие к среднему выборки, нормализовались лучше, чем “хвосты”

2.3.4 Огибающие

```
envelopes(chevron$norm, mean(chevron$norm), sd(chevron$norm), 3000)
```



Хвосты достаточно сильно выбиваются из “коридора” огибающих линий, так как данные не сгенерированы

2.3.5 Тесты нормальности

2.3.5.1 Колмогорова-Смирнова

```
library(MASS)
x <- unique(chevron$norm)
fit <- fitdistr(x,"normal")
cvx.mean <- fit$estimate[1]
cvx.sd <- fit$estimate[2]
ks.test(x, pnorm, mean=cvx.mean, sd=cvx.sd)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.07802, p-value = 4.093e-06
## alternative hypothesis: two-sided
```

Значение статистики Колмогорова-Смирнова достаточно мало, значит данные примерно удовлетворяют нормальному распределению

2.2.5.2 Шапиро-Уилка

```
library(nortest)
shapiro.test(chevron$norm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  chevron$norm
## W = 0.96423, p-value < 2.2e-16
```

Тест Шапиро-Уилка даёт результат близкий к 1, что тоже свидетельствует, что данные удовлетворяют нормальному распределению

2.3.5.3 Лиллифорса (вариация Колмогорова-Смирнова именно для нормального распределения)

```
lillie.test(chevron$norm)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  chevron$norm
## D = 0.085204, p-value < 2.2e-16
```

Статистика близка к 0, значит данные имеют нормальный вид

2.3.5.4 Шапиро-Франчия

```
sf.test(chevron$norm)
```

```
##
##  Shapiro-Francia normality test
##
## data:  chevron$norm
## W = 0.96489, p-value = 2.473e-15
```

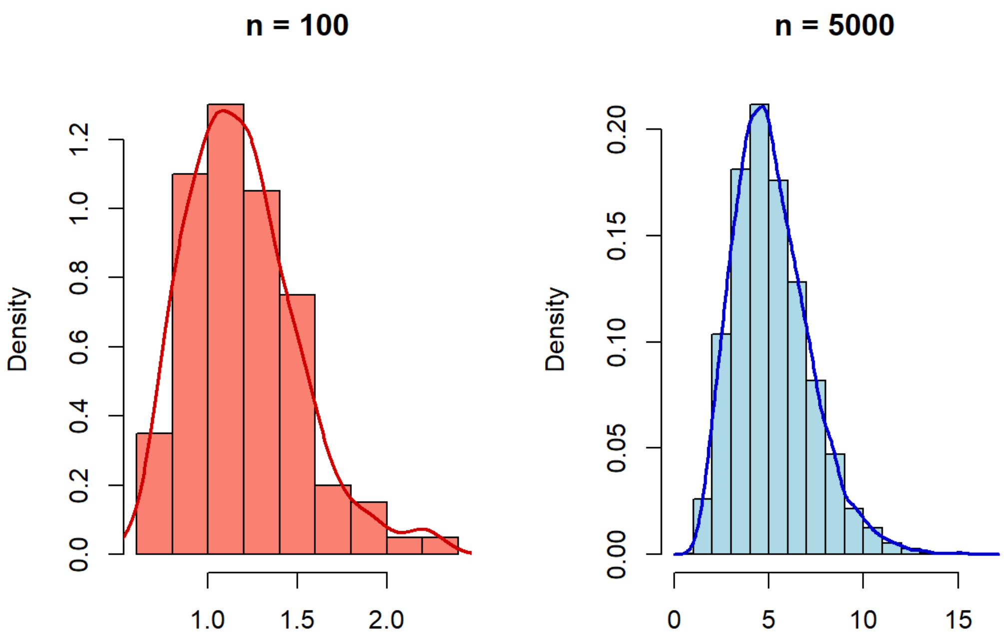
Значение статистики близко к 1, значит данные имеют нормальный вид

3. Выборки со случайными параметрами

3.1. Генеирование и визуализация

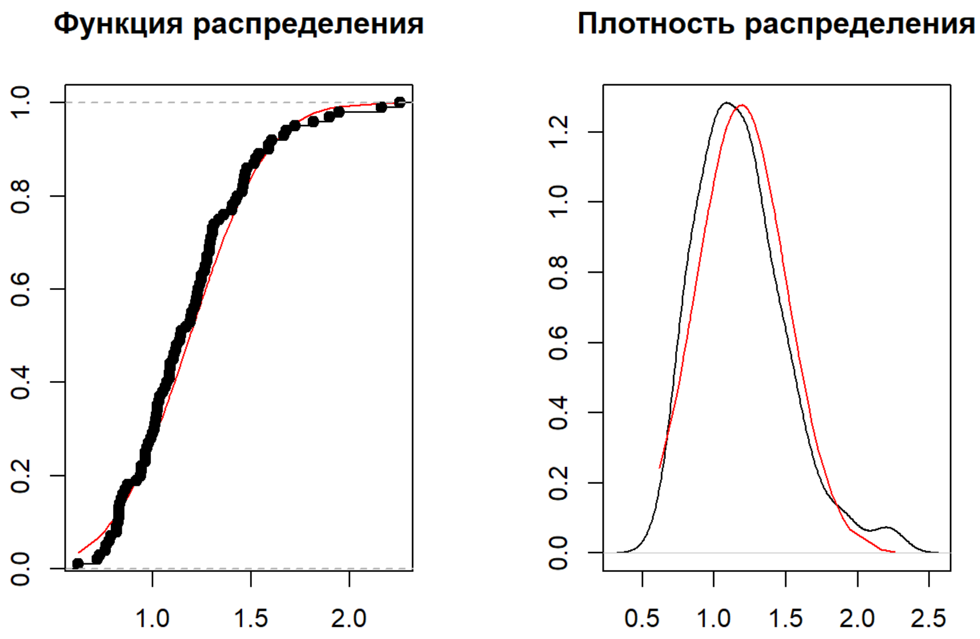
2 выборки из гамма распределения с параметрами из нормального распределения

```
set.seed(16102021)

par <- rnorm(10, 10, 10)

sm <- rgamma(100, abs(par[1]), abs(par[2]))
bg <- rgamma(5000, abs(par[3]), abs(par[4]))

par(mfrow=c(1,2))

hist(sm, freq=FALSE, col='salmon', main='n = 100', xlab='')
lines(density(sm), col='red3', lwd=2)

hist(bg, freq=FALSE, add=FALSE, col='lightblue', main='n = 5000', xlab='')
lines(density(bg), col='blue3', lwd=2)
```

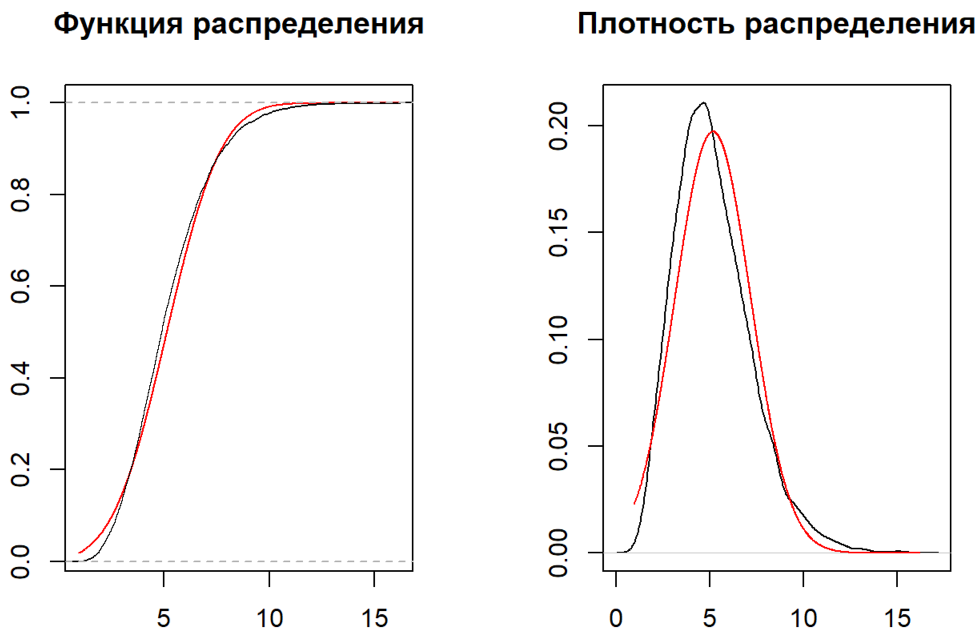


3.2. Теоретические и эмпирические функции рапределения и плотности

```
fit.sm <- fitdistr(sm,"normal")
sm.mean <- fit.sm$estimate[1]
sm.sd <- fit.sm$estimate[2]

fit.bg <- fitdistr(bg,"normal")
bg.mean <- fit.bg$estimate[1]
bg.sd <- fit.bg$estimate[2]

theor_empiric(sm, sm.mean, sm.sd)
```



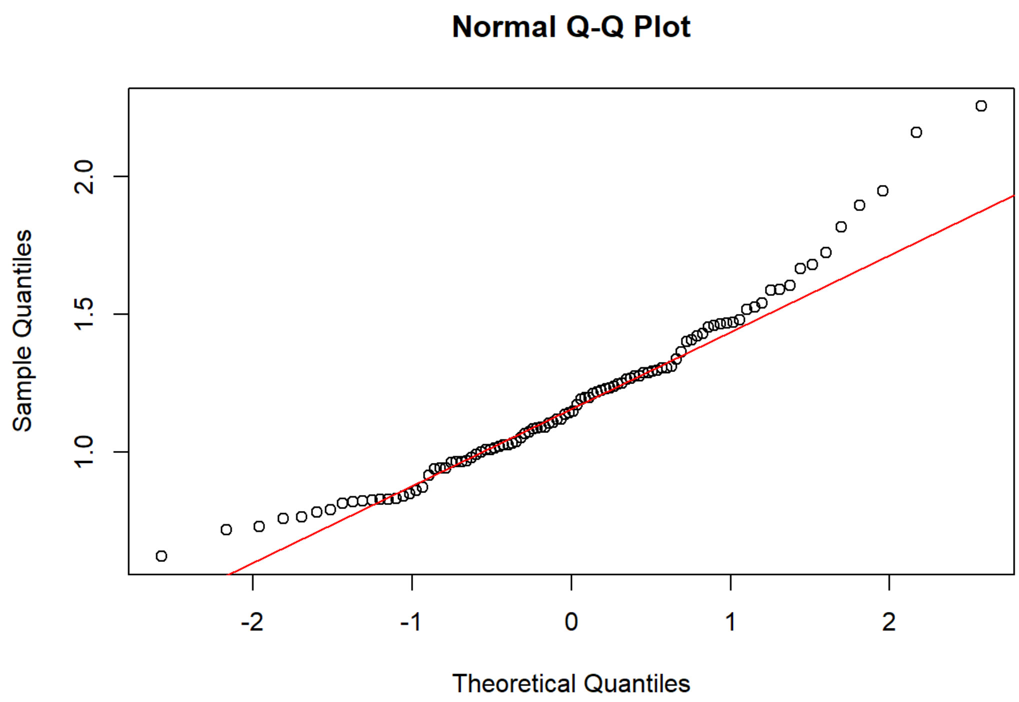

```
theor_empiric(bg, bg.mean, bg.sd)
```



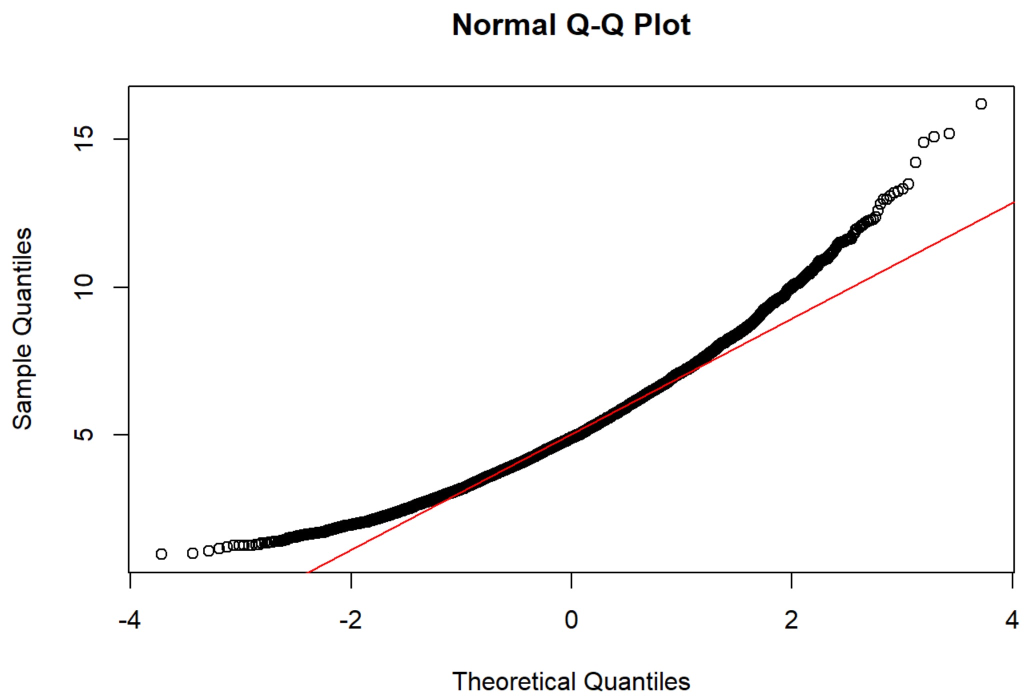
Данные достаточно хорошо приближаются нормальным распределением

3.3 Квантили

```
quantile(sm)
```



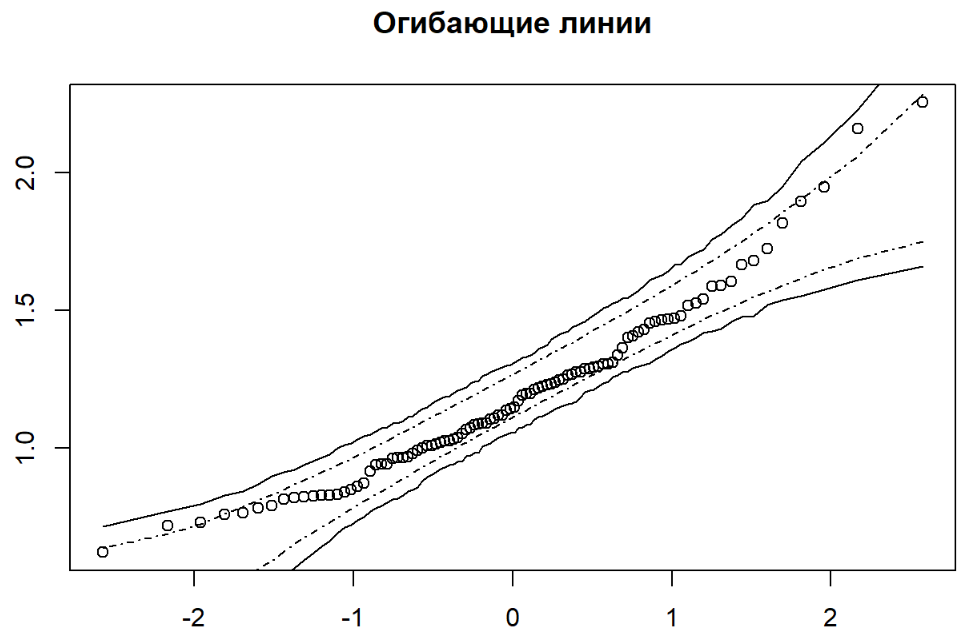

```
quantile(bg)
```



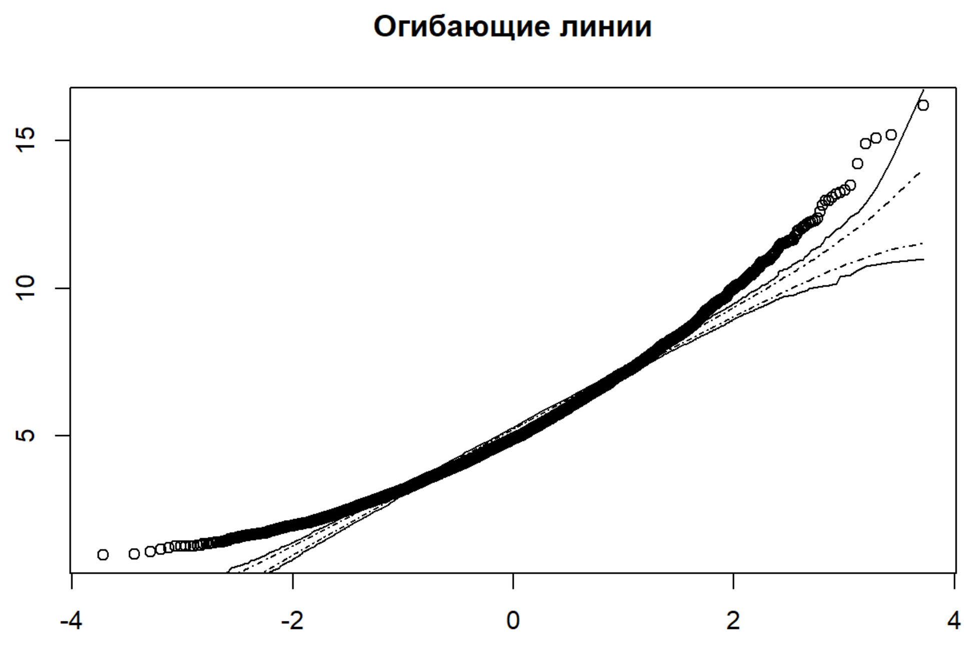
На Q-Q графиках заметна вогнутость линии, значит, правый хвост распределения больше, чем левый (что справедливо для гамма-распределения)

3.4 Огибающие

```
envelopes(sm, sm.mean, sm.sd, 5000)
```




```
envelopes(bg, bg.mean, bg.sd, 5000)
```



Небольшая выборка хорошо аппроксимируется нормальным распределением, большая же - плохо, квантили выходят за “коридор” огибающих

3.5 Тесты нормальности

3.5.1 Колмогорова-Смирнова

```
ks.test(sm, pnorm, mean=sm.mean, sd=sm.sd)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  sm
## D = 0.087238, p-value = 0.432
## alternative hypothesis: two-sided
```

```
ks.test(bg, pnorm, mean=bg.mean, sd=bg.sd)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  bg
## D = 0.057169, p-value = 1.277e-14
## alternative hypothesis: two-sided
```

Значения статистики Колмогорова-Смирнова достаточно малы, значит выборки близки к нормальному распределению

3.5.2 Шапиро-Уилка

```
library(nortest)
shapiro.test(sm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sm
## W = 0.95205, p-value = 0.001134
```

```
shapiro.test(bg)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bg
## W = 0.96558, p-value < 2.2e-16
```

Тесты Шапиро-Уилка дают результат близкий к 1, но меньший, чем они давали на заведомо нормальных данных (здесь уже можно усомниться в нормальности данных)

3.5.3 Андерсона-Дарлинга

```
ad.test(sm)
```

```
##
##  Anderson-Darling normality test
##
## data:  sm
## A = 0.98453, p-value = 0.01288
```

```
ad.test(bg)
```

```
##
##  Anderson-Darling normality test
##
## data:  bg
## A = 32.959, p-value < 2.2e-16
```

На большой выборке тест дает слишком большой результат для нормального распределения

3.5.4 Крамера-фон Мизеса

```
cvm.test(sm)
```

```
##
##  Cramer-von Mises normality test
##
## data:  sm
## W = 0.13837, p-value = 0.03329
```

```
cvm.test(bg)
```

```
## Warning in cvm.test(bg): p-value is smaller than 7.37e-10, cannot be computed
## more accurately
```

```
##
##  Cramer-von Mises normality test
##
## data:  bg
## W = 5.1663, p-value = 7.37e-10
```

Статистика на малой выборке достаточно близка к 0, что говорит о прибилженности к нормальному распределению

На большой же выборке результат сразу указывает на ненормальность данных

3.5.5 Лиллифорса (вариация Колмогорова-Смирнова именно для нормального распределения)

```
lillie.test(sm)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  sm
## D = 0.087965, p-value = 0.0543
```

```
lillie.test(bg)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  bg
## D = 0.057169, p-value < 2.2e-16
```

Статистика близка к 0, значит распределения имеют близкий к нормальному вид

3.5.6 Шапиро-Франчия

```
sf.test(sm)
```

```
##
##  Shapiro-Francia normality test
##
## data:  sm
## W = 0.9514, p-value = 0.001649
```

```
sf.test(bg)
```

```
##
##  Shapiro-Francia normality test
##
## data:  bg
## W = 0.96557, p-value < 2.2e-16
```

Результат близок к 1, что говорит о близости к нормальному распределению

3. Вывод

Выборки небольшого объема хорошо аппроксимируются нормальным распределением

Сделать вывод о нормальности выборки большого объема лучше всего помогут:

1. Q-Q график
2. Метод огибающих
3. Метод Шапиро-Уилка
4. Тест Андерсона-Дарлинга
5. Тест Крамера-фон Мизеса