Ярылгасимов Султан
Домашнее задание №5

1. Выполнить задания из комментариев к программному коду занятия.

1.1. Оценка максимального правдоподобия

Код:

```
# частный случай
n = 1000
x <- rnorm(n,mean=5,sd=3)
hist(norm, breaks=20)

f <- function(X){c(F1 = (sum(x)-n*X[1])/X[2],
                   F2=-n/(2*X[2]) + sum((x-X[1])^2)/(2*(X[2]^2)))}
multiroot(f, c(1, 3))
```

Вывод:

```
[1] 5.033670 9.157683
```

Комментарий: Решение популярной задачи для нахождения мат. ожидания и отклонения методом максимального правдоподобия. Как видно, все решено верно.

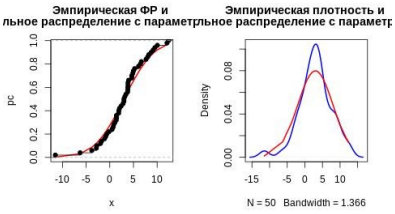
2. Сгенерировать данные из нормального распределения с различными параметрами и провести анализ с помощью графиков эмпирических функций распределений, квантилей, метода огибающих, а также шести стандартных процедур проверки гипотез о нормальности, рассмотренных на семинаре (Колмогорова-Смирнова, Шапиро-Уилка, Андерсона-Дарлинга, Крамера фон Мизеса, Колмогорова-Смирнова в модификации Лиллиефорса и Шапиро-Франсия). Рассмотреть выборки малого (не более 50-100 элементов) и умеренного (1000-5000 наблюдений) объемов. Повторить анализ для собственных данных.

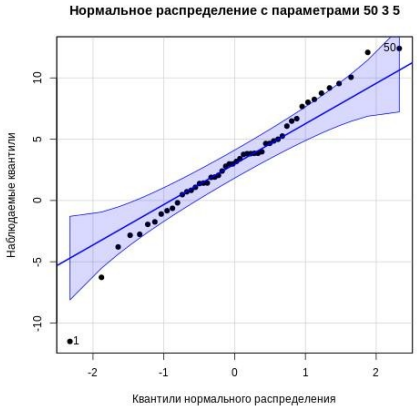
Код:

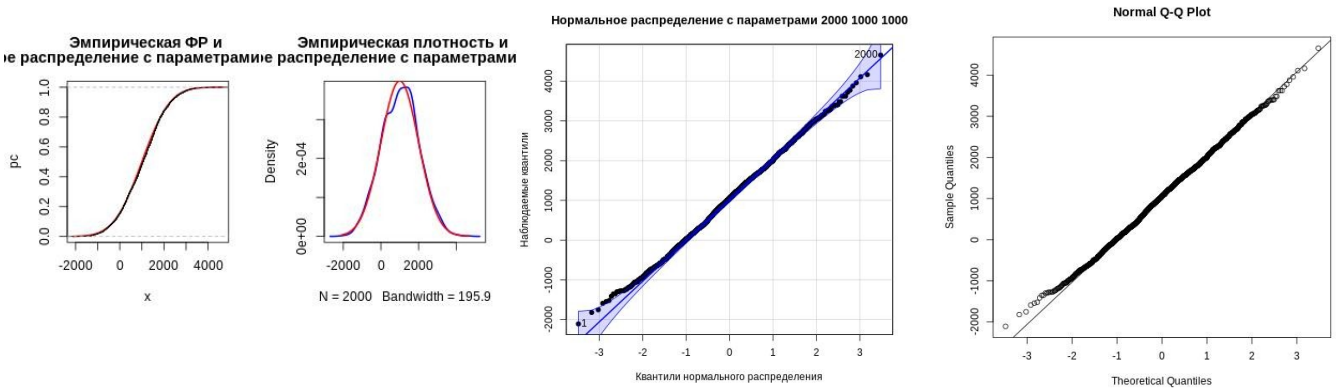
```
graph_distr <- function(x, pc, pd, main_name = "")
{
  op <- par(mfrow = c(1, 1), pty = "s")
  par(mfrow = c(1, 2))
  mn <- paste(c("Эмпирическая ФР и ", main_name))
  plot(x,pc, type = "l", col = "red", lwd = 2, main = mn)
  plot(ecdf(x), add = TRUE)
  mn <- paste(c("Эмпирическая плотность и ", main_name))
  plot(density(x), lwd = 2, col = "blue", main = mn)
  lines(x, pd, col = "red", lwd = 2)
  par(op)
}

library(car)
for (i in c(3, 9, 27)) {
  for (j in c(5, 10, 15)) {
    small = sort(rnorm(50, mean=i, sd=j))
    common_name=paste(50, i, j, sep='_')
    jpeg(paste("pics/small", common_name, 1,sep='_'))
    graph_distr(small, pnorm(small, mean = i, sd = j),
                dnorm(small, mean = i, sd = j),
                paste("нормальное распределение с параметрами", 50, i, j))
    dev.off()

    jpeg(paste("pics/small",common_name,2,sep='_'))
    #text(0.5,0.5,paste("Нормальное распределение с параметрами", 50, i,
j),cex=2,font=2)
```

```
    par(mfrow = c(1,1))
    qqnorm(small, main = paste("Нормальное распределение с параметрами", 50,
i, j))
    qqline(small)
    dev.off()

    jpeg(paste("pics/small",common_name,3,sep='_'))
    qqPlot(small, dist = "norm", col = palette()[1], pch = 19,
           xlab="Квантили нормального распределения",
           ylab="Наблюдаемые квантили",
           main=paste("Нормальное распределение с параметрами", 50, i, j))
    dev.off()
  }
}

for (i in c(10, 100, 1000)) {
  for (j in c(10, 100, 1000)) {
    jpeg(paste("pics/big",common_name,1,sep='_'))
    big = sort(rnorm(2000, mean=i, sd=j))
    graph_distr(big, pnorm(big, mean = i, sd = j),
                dnorm(big, mean = i, sd = j),
                paste("нормальное распределение с параметрами", 2000, i, j))
    #text(0.5,0.5,paste("Нормальное распределение с параметрами", 2000, i,
j),cex=2,font=2)
    jpeg(paste("pics/big",common_name,2,sep='_'))
    par(mfrow = c(1,1))
    qqnorm(big)
    qqline(big)
    dev.off()

    jpeg(paste("pics/big",common_name,3,sep='_'))
    qqPlot(big, dist = "norm", col = palette()[1], pch = 19,
           xlab="Квантили нормального распределения",
           ylab="Наблюдаемые квантили",
           main=paste("Нормальное распределение с параметрами", 2000, i, j))
    dev.off()
  }
}
```

Вывод:







Комментарий: В выводе я указал лишь крайние значения, так как графиков получилось очень много, но очевидно что при больших значениях и меньшей дисперсии, графики больше похожи на графики нормального распределения.

Код:

```
for (k in c (50, 2000)) {
  for (i in c(3, 9, 27)) {
    for (j in c(5, 10, 15)) {
      x = rnorm(k, mean=i, sd=j)
      print(paste("Size:",k,"Mean",i,"Sd",j))
      print(paste("Shapiro", shapiro.test(x)$p.value))
      print(paste("Ad",ad.test(x)$p.value))
      print(paste("Cvm",cvm.test(x)$p.value))
      print(paste("Lillie",lillie.test(x)$p.value))
      print(paste("Sf",sf.test(x)$p.value))
      print("")
    }
  }
}
```

Вывод:

```
[1] "Size: 50 Mean 3 Sd 5"
[1] "Shapiro 0.171900224014044"
[1] "Ad 0.181860912723633"
[1] "Cvm 0.244522184381544"
[1] "Lillie 0.32452464382852"
[1] "Sf 0.19427208363152"
[1] ""
```

```
[1] "Size: 2000 Mean 27 Sd 15"
[1] "Shapiro 0.20282973631302"
[1] "Ad 0.16371797409238"
[1] "Cvm 0.143320754620453"
[1] "Lillie 0.381597305116044"
[1] "Sf 0.259893906419766"
[1] ""
```

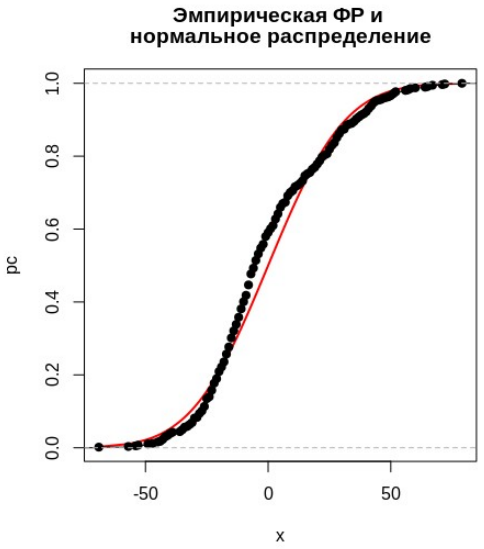
Комментарий: В выводе, снова, указаны лишь граничные значения, но и при всех параметрах не случалось такого, что p-value был ниже 0.05.

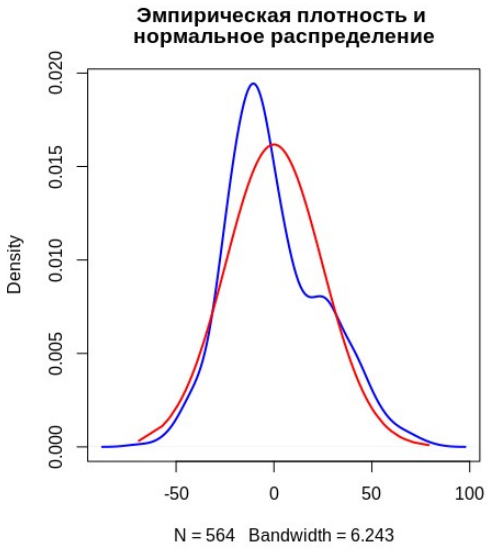
Код:

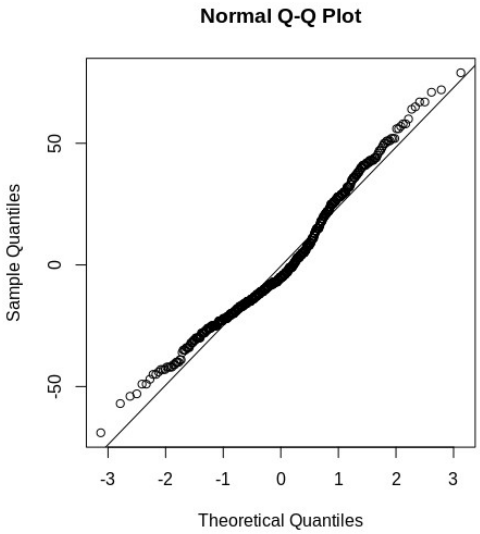
```
gds  # from last hw
gds=sort(gds)
hist(gds, freq=F)
lines(density(gds))
```

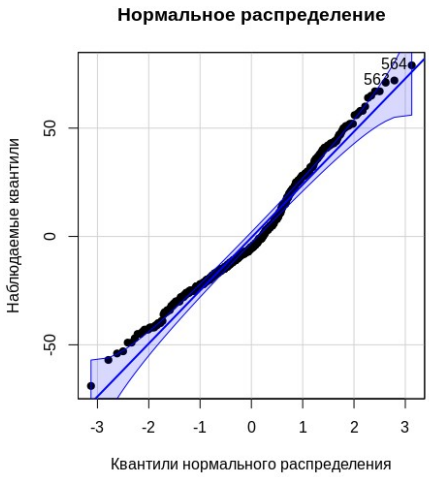
```
shapiro.test(gds)$p.value  # very little p.value\
graph_distr(gds, pnorm(gds, mean = mean(gds), sd = sd(gds)),
            dnorm(gds, mean = mean(gds), sd = sd(gds)),
            "нормальное распределение")

par(mfrow = c(1,1))
qqnorm(gds)
qqline(gds)

qqPlot(gds, dist = "norm", col = palette()[1], pch = 19,
       xlab="Квантили нормального распределения",
       ylab="Наблюдаемые квантили",
       main=paste("Нормальное распределение"))
```

Вывод:









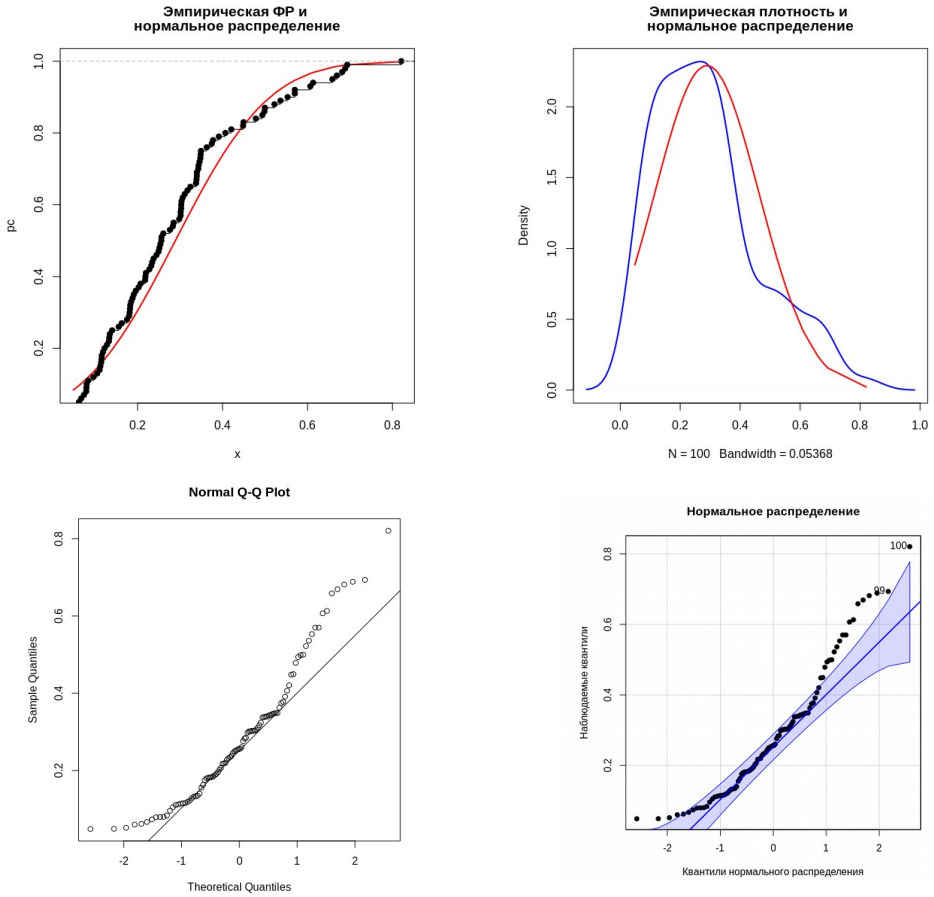
Комментарий: проверил на своих данных и оказалось, что то, что я считал нормальным распределением таковым не является. На графиках это особенно видно при сравнении с нормальным распределением, но явный тест дает более очевидны ответ с очень низким p-value (1.375986e-08).

3. Сгенерировать данные из комбинаций реализованных в R распределений (распределение, параметр которого также является случайным), а затем провести анализ с помощью графиков квантилей, метода огибающих, а также стандартных процедур проверки гипотез о нормальности. Рассмотреть выборки малого и умеренного объемов. Сравнить эффективность методов.

Код:

```
x = sort(rgamma(100, 2, (5 + 2*rpois(10, 1))))
hist(x, freq=F)
lines(density(x))
pars = fitdistr(x, "normal")
ks.test(x, pnorm, mean=pars$estimate[1], sd=pars$estimate[2])
m = pars$estimate[1]
s = pars$estimate[2]
graph_distr(x, pnorm(x, mean = m, sd = s),
            dnorm(x, mean = m, sd = s),
            "нормальное распределение")

par(mfrow = c(1,1))
qqnorm(x)
qqline(x)

qqPlot(x, dist = "norm", col = palette()[1], pch = 19,
       xlab="Квантили нормального распределения",
       ylab="Наблюдаемые квантили",
       main=paste("Нормальное распределение"))

shapiro.test(x)
```

Вывод:

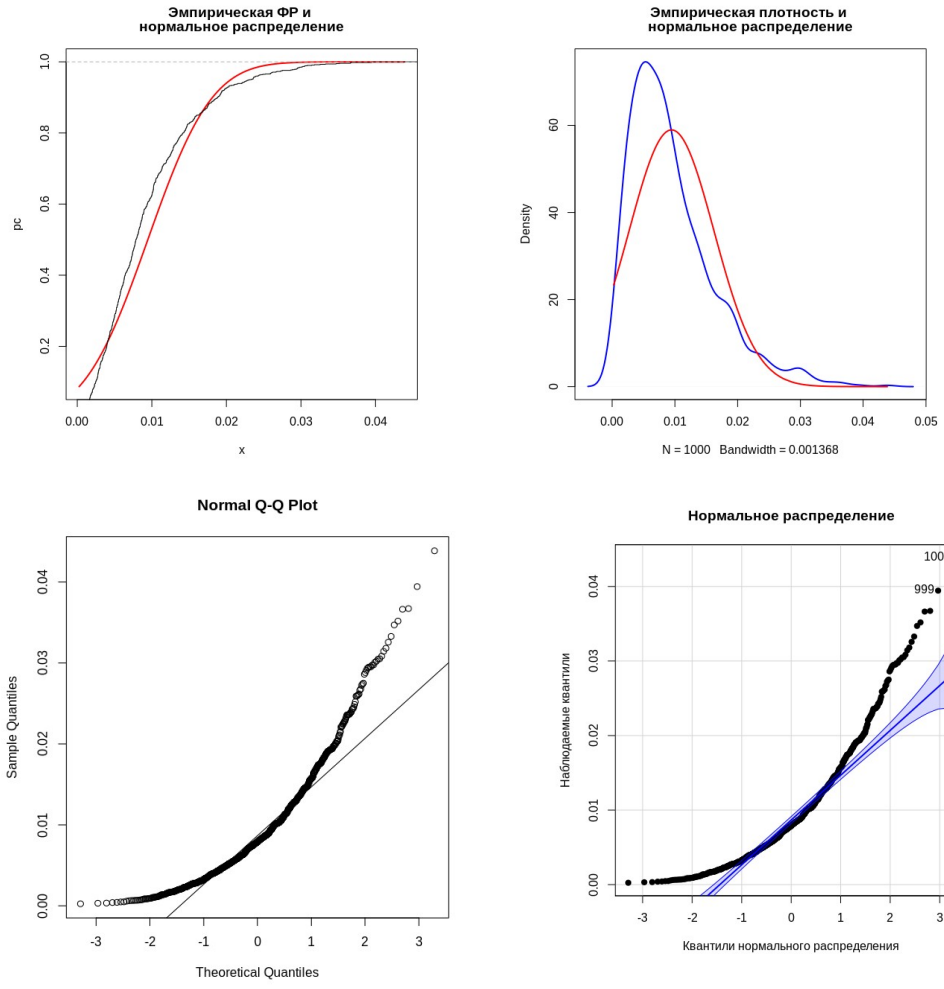


Комментарий: применено Гамма-распределение с Пуассоновскими «атрибутами». ks.test оценил схожесть с нормальным в 14%, а Шапиро дал совсем низкий p-value. Проверим на большем числе наблюдений.

Код:

```
x = sort(rgamma(1000, 2, (5 + 2*rpois(1000, 100))))
hist(x, freq=F)
lines(density(x))
pars = fitdistr(x, "normal")
ks.test(x, pnorm, mean=pars$estimate[1], sd=pars$estimate[2])
m = pars$estimate[1]
s = pars$estimate[2]
graph_distr(x, pnorm(x, mean = m, sd = s),
            dnorm(x, mean = m, sd = s),
            "нормальное распределение")

par(mfrow = c(1,1))
qqnorm(x)
qqline(x)

qqPlot(x, dist = "norm", col = palette()[1], pch = 19,
       xlab="Квантили нормального распределения",
       ylab="Наблюдаемые квантили",
       main=paste("Нормальное распределение"))

shapiro.test(x)
```

Вывод:



Комментарий: на больших данных еще более очевидно, что данное распределение совсем не похоже на нормальное.