Методы статистической обработки информации. Задание 2

Ершов А. С., гр. 22.М04-мм

Вариант 4

Задание:

Распределение хи-квадрат. Промоделировать выборку с заданным законом распределения, построить гистограмму, оценить параметры по методу моментов и максимального правдоподобия, изобразить на гистограмме плотности распределения, соответствующие оценкам из разных методов. Применить статистику хи-квадрат для проверки согласия эмпирического и теоретического распределений.

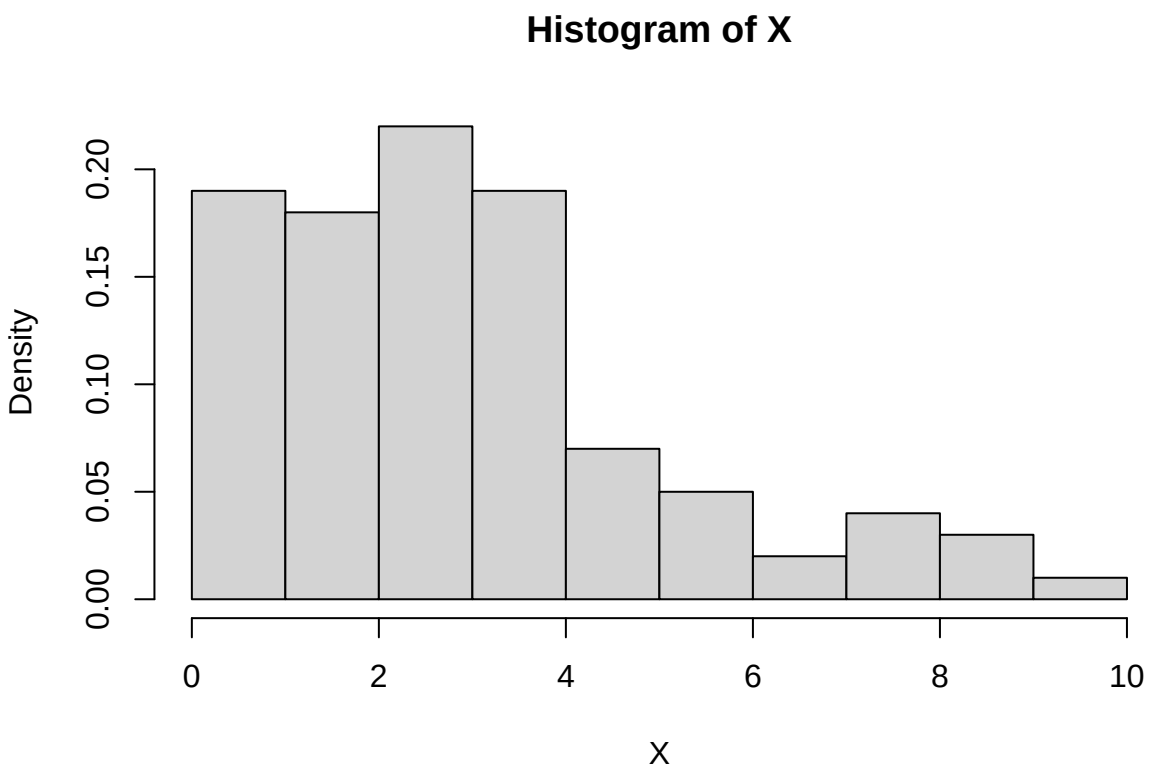
Промоделируем тестовые данные с распределением хи-квадрат с параметром $k = 3$.

```
k = 3
n = 100
set.seed(120)
X <- rchisq(n, k)
X
```

```
##   [1] 1.4889333 3.9921237 1.9240933 4.6461312 2.4774568 0.5994625 0.9957838
##   [8] 3.9093088 2.8326790 3.0277148 3.1221577 7.1180664 3.1959616 2.3373994
##  [15] 8.2013586 0.2236692 3.5235961 5.7175211 1.3876534 0.9001743 5.7124413
##  [22] 2.7615254 0.1610993 0.8059258 2.1070561 2.8361483 4.2560334 3.6345695
##  [29] 0.9047548 0.4373729 8.5076551 0.8678281 1.0559055 1.0086046 0.9953463
##  [36] 2.0794139 7.5286841 1.6143805 2.6383733 1.0795591 2.8980456 2.3429402
##  [43] 0.2780572 3.7153495 4.2616210 9.7261034 2.2245404 0.6350362 3.6755882
##  [50] 3.5177393 2.2880905 2.3816112 1.7416472 2.1284176 3.3481788 0.4363866
##  [57] 7.5048531 1.2850168 3.1789625 1.4196351 6.8707094 0.4709919 3.1974582
##  [64] 2.2419930 1.3733169 5.7917082 0.1601876 4.1934363 2.5649316 1.0191639
##  [71] 4.3124330 3.4325490 3.7072455 1.5561437 2.1721847 2.7213867 4.3636567
##  [78] 2.7159279 5.9090045 1.0409147 1.5114101 7.3263866 5.2562004 2.8879445
##  [85] 0.9931524 0.8623436 0.2832026 3.7681912 2.1783477 2.3063296 1.8666813
##  [92] 1.3340355 1.2938590 4.7784852 3.9809732 0.8814240 3.1577084 8.1286752
##  [99] 3.2845739 6.7128898
```

Построим гистограмму:

```
hist(X, freq = FALSE)
```



Математическое ожидание:

```
mu. <- mean(X)
mu.
```

```
## [1] 2.942099
```

Дисперсия:

```
dispersion <- sd(X) ^ 2
dispersion
```

```
## [1] 4.598514
```

Оценка параметров

Оценим параметры распределения методом моментом и методом максимального правдоподобия.

Метод моментов

$$E[X] = k \Rightarrow k = E[X]$$

Оценим параметр с помощью первого момента:

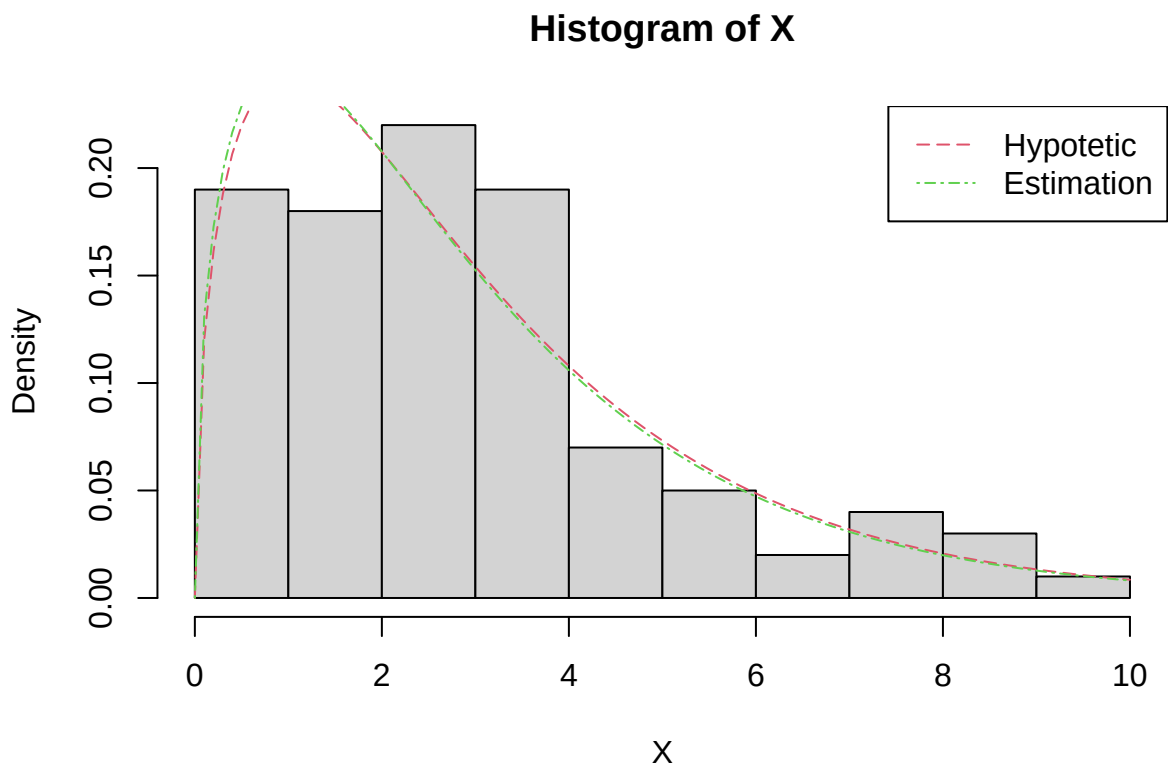
```
k_estimation1 <- mu.
k_estimation1
```

```
## [1] 2.942099
```

Оценка параметра близка к заданной при моделировании тестовых данных.

Построим графики гипотетического распределения (красный) и распределения с параметром, оцененным методом моментов (зелёный):

```
f <- function(x) dchisq(x, k)
f1 <- function(x) dchisq(x, k_estimation1)
hist(X, freq=FALSE)
curve(f, add=TRUE, col=2, lty=5)
curve(f1, add=TRUE, col=3, lty=6)
legend('topright', c("Hypotetic","Estimation"), col = c(2,3), lty = c(5, 6))
```



По графику видно, что значения распределения с параметром, полученным с помощью метода моментов, близки к значениям гипотетического распрделения.

Метод максимального правдоподобия

Функция правдоподобия:

```
Func.prob.log <- function(x, df) -sum(dchisq(X, df = df, log = TRUE))
```

Оценим параметр:

```
k_estimation2 = optimize(f = Func.prob.log, x = X, interval = c(0, k_estimation1))$minimum
k_estimation2
```

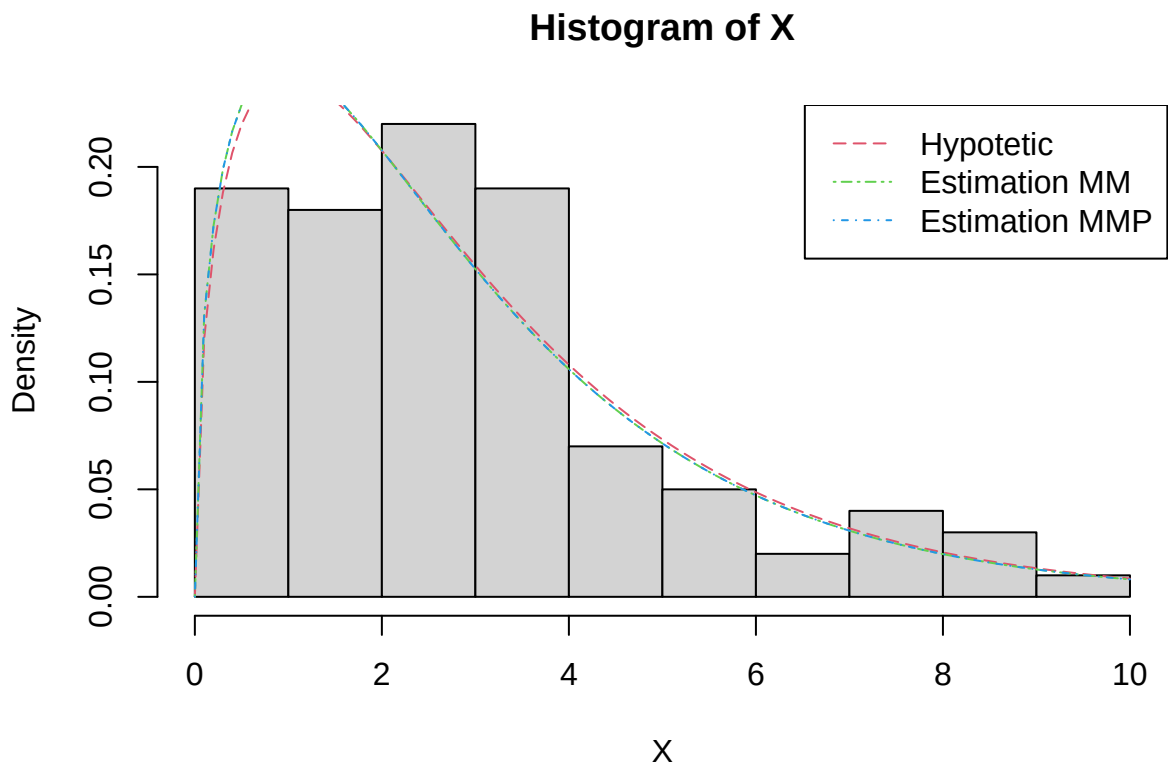
```
## [1] 2.94205
```

Значение параметра близко к гипотетическому (3), но для тестовой выборки метод моментов оценил параметр точнее.

Построим графики гипотетического распределения (красный), распределения с параметром, оцененным методом моментов (зелёный) и распределения с параметром, оцененным методом максимального правдоподобия (синий):

```
f <- function(x) dchisq(x, k)
f1 <- function(x) dchisq(x, k_estimation1)
f2 <- function(x) dchisq(x, k_estimation2)

hist(X, freq=FALSE)
curve(f, add=TRUE, col=2, lty=5)
curve(f1, add=TRUE, col=3, lty=6)
curve(f2, add=TRUE, col=4, lty=4)
legend('topright', c("Hypotetic","Estimation MM","Estimation MMP"), col = c(2,3,4), lty = c(5, 6,4))
```



По графику подтверждаются предыдущие выводы: оценки параметра близки к его значению, метод максимального правдоподобия показал лучшую оценку.

Проверка гипотезы согласия

Используем критерий хи-квадрат для проверки гипотезы согласия.

```
h <- hist(X, plot = FALSE)
```

Эмпирические частоты:

```
n.i <- sapply(seq(length(h$breaks) - 1) + 1, function(i)
length(X[X < h$breaks[i] & X >= h$breaks[i - 1]]))
```

Гипотетические частоты:

```
p.i <- sapply(seq(length(h$breaks) - 1) + 1,function(i)
pchisq(h$breaks[i], k_estimation2) - pchisq(h$breaks[i - 1], k_estimation2))
sum(p.i)
```

```
## [1] 0.9823612
```

```
p.i[1] <- pchisq(h$breaks[2], k_estimation2)
p.i[length(p.i)] <- 1 - pchisq(h$breaks[length(h$breaks)-1], k_estimation2)
sum(p.i)
```

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

Проверим условие $n * p_i > 5$:

```
tab <- cbind(h$counts, p.i * length(X))
tab
```

```
##       [,1]      [,2]
##  [1,]   19 20.741765
##  [2,]   18 23.110463
##  [3,]   22 17.982349
##  [4,]   19 12.819693
##  [5,]    7  8.766631
##  [6,]    5  5.849996
##  [7,]    2  3.841167
##  [8,]    4  2.493394
##  [9,]    3  1.604719
## [10,]    1  2.789822
```

```
t1 <- cumsum(tab[,2])
T1 <- min(which(t1 > 5))
t1
```

```
##  [1]  20.74177  43.85223  61.83458  74.65427  83.42090  89.27090  93.11206
##  [8]  95.60546  97.21018 100.00000
```

```
t2 <- cumsum(tab[seq(nrow(tab),1, -1), 2])
t2
```

```
##  [1]   2.789822   4.394541   6.887935  10.729102  16.579098  25.345729
##  [7]  38.165422  56.147772  79.258235 100.000000
```

```
T2 <- min(which(t1 > 5))
T2 <- nrow(tab) - T2
T2
```

```
## [1] 9
```

```
Tab<-apply(tab, 2, function(x) c(sum(x[seq(T1)]), x[c((T1 + 1):T2)], sum(x[c((T2 + 1):nrow(tab))])))
Tab
```

```
##       [,1]      [,2]
##  [1,]   19 20.741765
##  [2,]   18 23.110463
##  [3,]   22 17.982349
##  [4,]   19 12.819693
##  [5,]    7  8.766631
##  [6,]    5  5.849996
##  [7,]    2  3.841167
##  [8,]    4  2.493394
##  [9,]    3  1.604719
## [10,]    1  2.789822
```

Статистика хи-квадрат:

```
chi2 <- sum(apply(Tab, 1, function(x) (x[1] - x[2]) ^ 2 / x[2]))
chi2
```

```
## [1] 9.787296
```

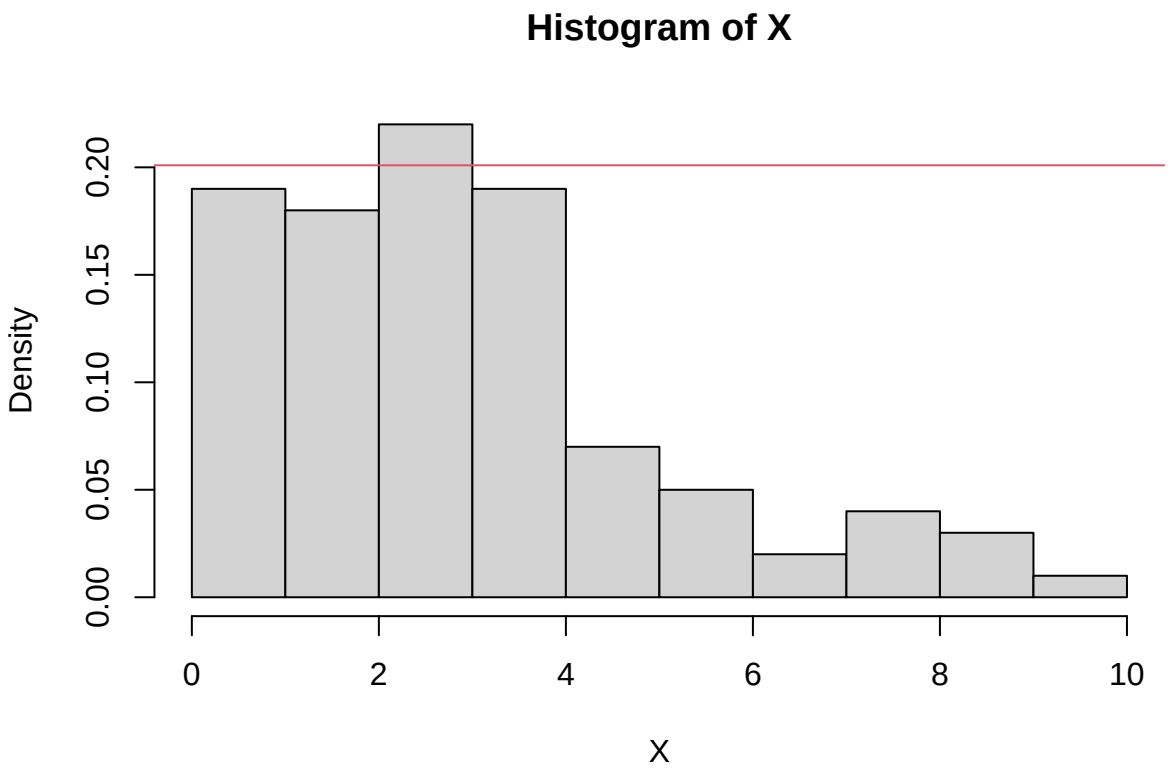
Доверительный уровень вероятности:

```
pvalue <- 1 - pchisq(chi2, nrow(Tab) - 3);
print(paste("p-value", round(pvalue, 4), sep=" = "))
```

```
## [1] "p-value = 0.201"
```

График:

```
hist(X, freq = FALSE)
abline(h = pvalue, col = 2)
```



Значение доверительного интервала больше 0.05, изначальная гипотеза не отвергается.