Отчет по домашнему заданию

Фахртдинов Т. А.

8 апреля 2020 г.

Вторая задача. Оценка параметров. Вариант 4.

1) Промоделировать выборки объемом 100 и 300 имеющие сложно пуассоновское с внутренним геометрическим распределением. Ввиду того, что параметры не были уточнены, я взял произвольные значения параметров распределения, $\lambda = 3$, $p = 0.5$:

```
generate_pois.geom <- function(size){
  samp <- character()
  for (j in 1:size) {
    x <- rpois(1, 3)
    y <- rgeom(x, 0.5)
    samp <- c(samp, sum(y))
  }
  samp <- as.numeric(samp)
  return(samp)
}

first <- generate_pois.geom(100)
second <- generate_pois.geom(300)

n1 <- length(first)
n2 <- length(second)
```

2) Теперь оценим параметры используя метод моментов. Известно, что производящие функции моментов распределения Пуассона и распределения Бернулли соответственно имеют вид:

$$f(s) = e^{-\lambda + s\lambda}$$

$$g(s) = \frac{p}{1 - qs}$$

тогда получаем производящую функцию сложного распределения:

$$h(s) = f(g(s)) = e^{-\lambda + \lambda \frac{p}{1-qs}}$$

$$\alpha_1 = h'(1)$$

$$\eta_2 = h''(1) + h'(1) - (h'(1))^2$$

$h'(1) = \frac{q\lambda}{p} \quad h''(1) = \frac{q^2(\lambda^2 + 2\lambda)}{p^2}$

$\alpha_1 = \frac{q\lambda}{p} \quad \eta_2 = \frac{qp\lambda + 2q^2\lambda}{p^2}$

Из чего мы получаем оценки:

$$\bar{p} = \frac{2}{m_2 + \bar{x}}$$

$$\bar{\lambda} = \frac{2\bar{x}^2}{m_2 - \bar{x}}$$

```
estimate_p <- function(samp){
  return(2 / (var(samp) + mean(samp)))
}

estimate_lmbd <- function(samp){
  return( 2 * mean(samp)^2 / (var(samp) - mean(samp)))
}
```

Для первой:

```
lmbd_f <- estimate_lmbd(first)
p_f <- estimate_p(first)
lmbd_f

## [1] 2.927628

p_f

## [1] 0.1697094
```

Для второй:

```
lmbd_s <- estimate_lmbd(second)
p_s <- estimate_p(second)
lmbd_s

## [1] 2.747303

p_s

## [1] 0.1383145
```

Построим доверительные интервалы.
Начнем с параметра p:

```
DH_p <- function(samp){
  m <- mean(samp)
  a2 <- mean(samp^2)
  a3 <- mean(samp^3)
  a4 <- mean(samp^4)
  m2 <- a2 - m^2
  m3 <- a3 - 3 * a2 * m + 2 * m^3
  m4 <- a4 - 4 * a2 * m + 6 * a2 * m^2 - 3 * m^4
  n <- length(samp)
  dh <- (m2 / n + 2 * (n - 1) * m3 / n^2 + (m4 - m2^2) / n) * 4 / ((m2 + m)^4)
  return(dh)
}

DH_p1 <- DH_p(first)
DH_p2 <- DH_p(second)
```

Для параметра p, 95% доверительный интервал первой и второй выборки соответственно:

```
c(p_f - qnorm(0.975)*sqrt(DH_p1 / n1), p_f + qnorm(0.975)*sqrt(DH_p1 / n1))

## [1] 0.1566909 0.1827280

c(p_s - qnorm(0.975)*sqrt(DH_p2 / n2), p_s + qnorm(0.975)*sqrt(DH_p2 / n2))

## [1] 0.1345357 0.1420933
```

Теперь построим доверительный интервал для параметра λ

```
DH_lmbd <- function(samp){
  m <- mean(samp)
  a2 <- mean(samp^2)
  a3 <- mean(samp^3)
  a4 <- mean(samp^4)
  m2 <- a2 - m^2
  m3 <- a3 - 3 * a2 * m + 2 * m^3
  m4 <- a4 - 4 * a2 * m + 6 * a2 * m^2 - 3 * m^4
  n <- length(samp)
  H1 <- 2 * m * (2 * m2 - m) / (m2 - m)
  H2 <- -2 * m^2 / (m2 - m)^2
  dh <- m2 * H1^2 / n + 2 * H1 * H2 * (n - 1) * m3 / n^2 + (m4 - m2^2) * H2^2 / n
  return(dh)
}

DH_lmbd1 <- DH_lmbd(first)
DH_lmbd2 <- DH_lmbd(second)
```

Для параметра λ, 95% доверительный интервал первой и второй выборки соответственно:

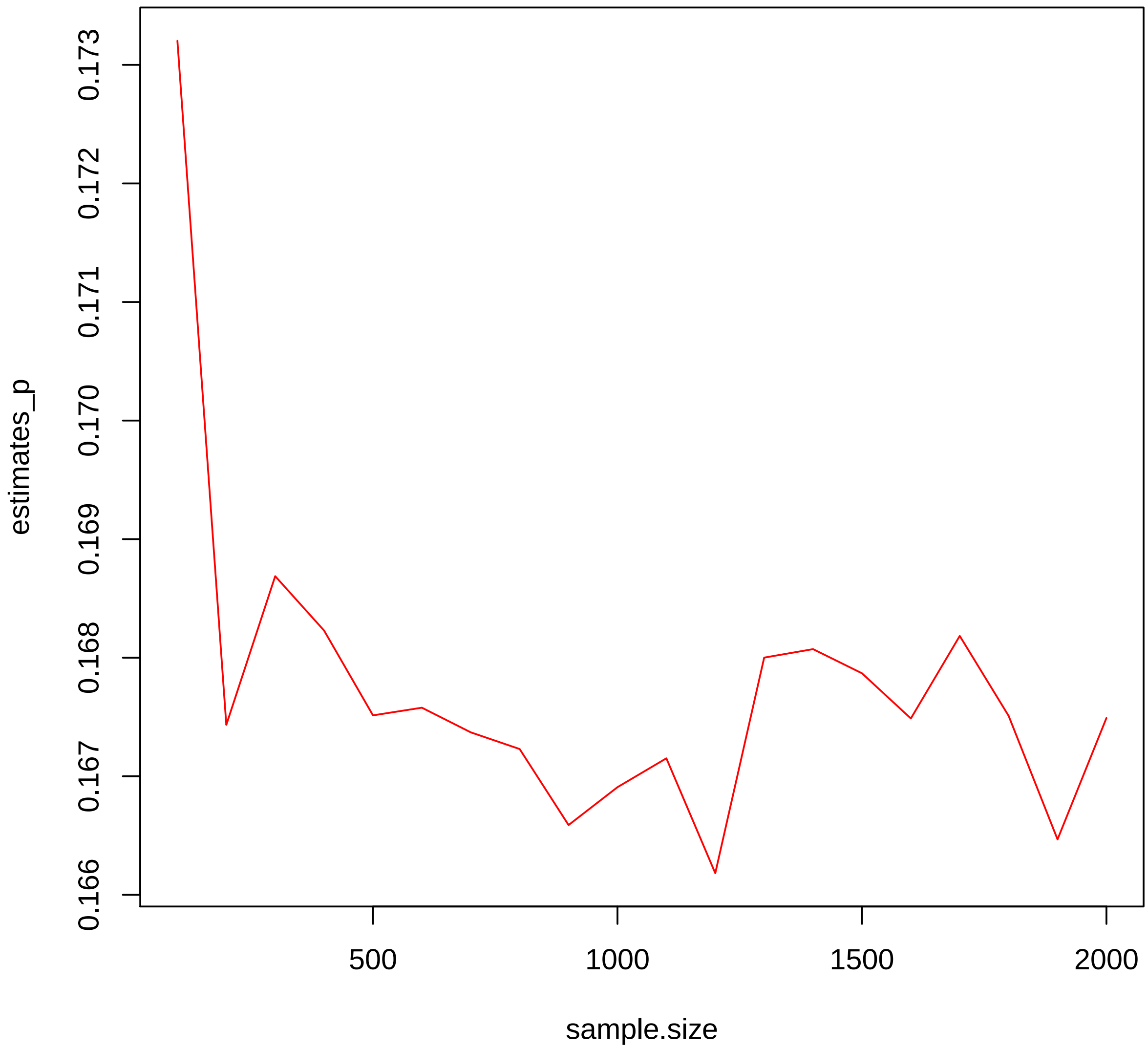
```
c(lmbd_f - qnorm(0.975)*sqrt(DH_lmbd1 / n1), lmbd_f + qnorm(0.975)*sqrt(DH_lmbd1 / n1))

## [1] 2.109157 3.746099

c(lmbd_s - qnorm(0.975)*sqrt(DH_lmbd2 / n2), lmbd_s + qnorm(0.975)*sqrt(DH_lmbd2 / n2))

## [1] 2.420107 3.074500
```

Свойства оценок метода моментов:

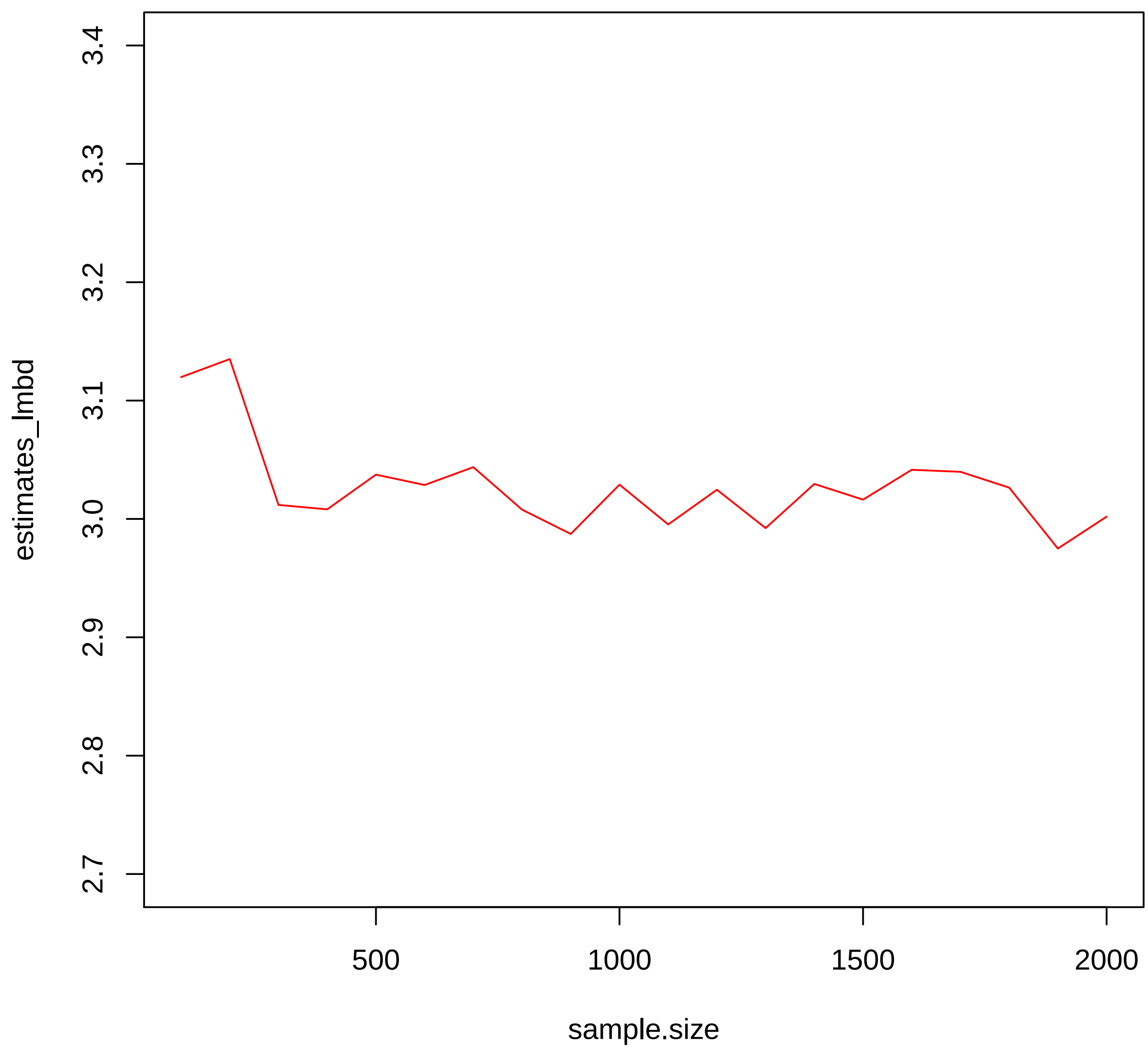
Проверим несмещенность оценки. Посторим график математического ожидания оценки, для разных размеров выборки.

```
sample.size <- seq(100, 2000, by = 100)

estimates_p <- character()
estimates_lmbd <- character()
estimates_lmbd_dh <- character()

for (size in sample.size) {
  temp_p <- character()
  temp_lmbd <- character()
```

```r
  temp_dh_lmbd <- character()
  for (i in 1:100) {
    temp <- generate_pois.geom(size)
    temp_p <- c(temp_p, estimate_p(temp))
    temp_lmbd <- c(temp_lmbd, estimate_lmbd(temp))
    temp_dh_lmbd <- c(temp_dh_lmbd, DH_lmbd(temp))
  }

  temp_p <- as.numeric(temp_p)
  temp_lmbd <- as.numeric(temp_lmbd)
  temp_dh_lmbd <- as.numeric(temp_dh_lmbd)

  estimates_p <- c(estimates_p, mean(temp_p))
  estimates_lmbd <- c(estimates_lmbd, mean(temp_lmbd))
  estimates_lmbd_dh <- c(estimates_lmbd_dh, mean(temp_dh_lmbd))
}

estimates_p <- as.numeric(estimates_p)
estimates_lmbd <- as.numeric(estimates_lmbd)


plot(sample.size, estimates_p, type = "l", col = "red")
```



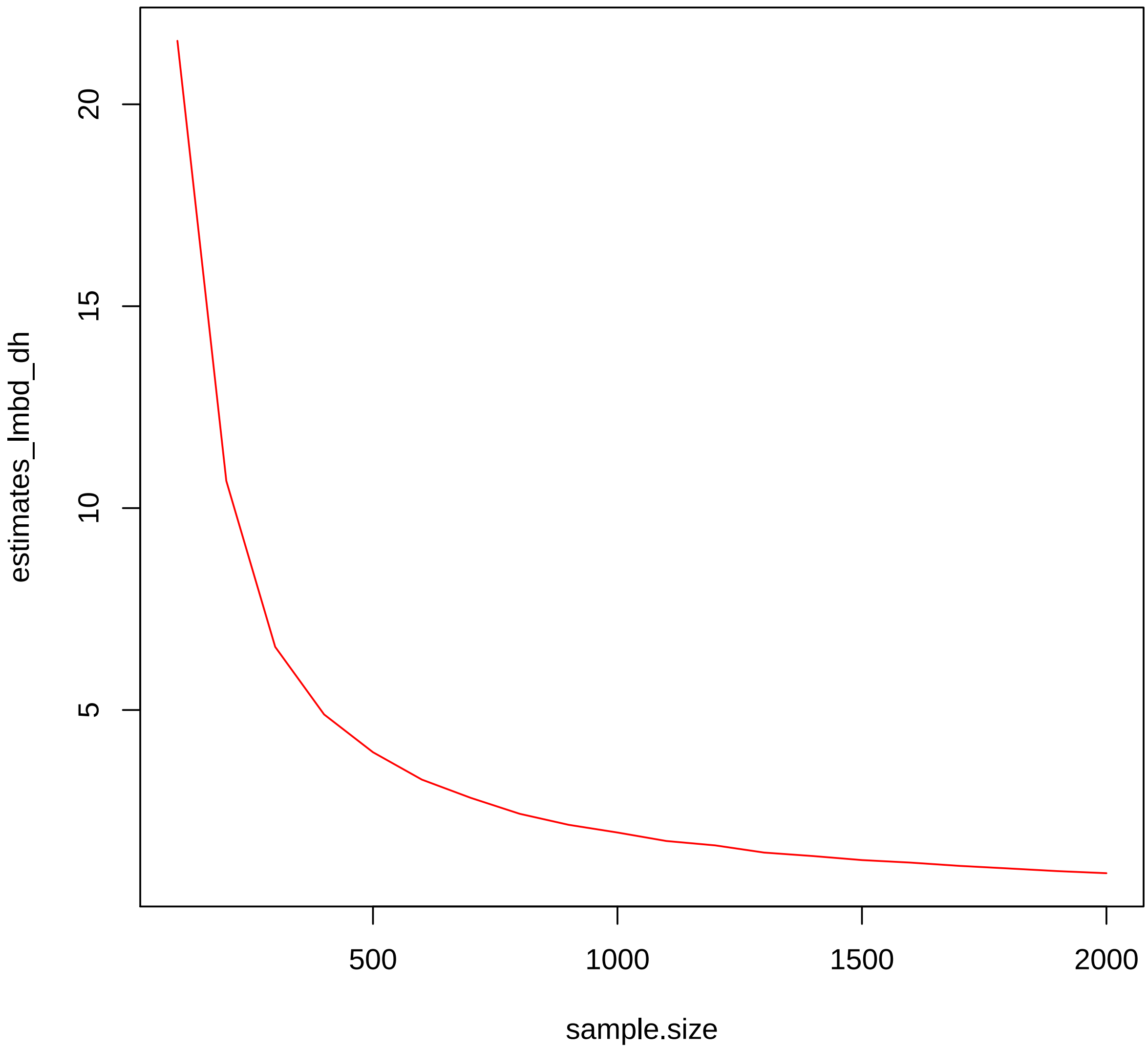

```
plot(sample.size, estimates_lmbd, type = "l", col = "red", ylim=c(2.7, 3.4))
```



Мы видим, что оценка параметра p — смещенная, а следовательно не состоятельная. Оценка параметра λ — несмещенная.

Теперь построим график дисперсии параметра λ, чтобы оценить состоятельность.

```
plot(sample.size, estimates_lmbd_dh, type = "l", col = "red")
```



Дисперсия с ростом объема выборки стремиться к нулю из чего делаем вывод о состоятельности оценки λ.

Чтобы оценить достаточность оценок, рассмотрим функцию правдоподобия:

$$L(k) = \prod_{s=1}^{n} \sum_{j=1}^{\infty} \frac{\lambda^j e^{-\lambda}}{j!} \frac{(j+k_s-1)!}{(j-1)!k_s!} p^j (1-p)_s^k$$

Невозможно свести функцию правдоподобия к виду, для параметров λ и p:

$$h(k)g(\bar{\theta}(k), \theta)$$

Делаем вывод о том, что оценки не являются достаточными.

Метод правдоподобия.

Сделаем оценку методом правдоподобия для первого варианта.
Возьмем гамма распределение gamma(k, r) с параметрами: $k = 3$, $r = 2$.
Функция правдоподобия имеет вид:

$$L(x) = \frac{e^{n\bar{x}r^{-1}}}{r^{nk}\Gamma^n(k)}(\prod_{i=1}^{n} x_i)^{k-1}$$

Получаем уравнения:

$$\frac{\partial \log(L)}{\partial k} = \log(r) + \psi(k) - \frac{1}{n}\sum_i \log(x_i) = 0$$

$$\frac{\partial \log(L)}{\partial r} = \frac{n\bar{x}}{r^2} - \frac{nk}{r} = 0$$

Тогда $r = \frac{\bar{x}}{k}$, а k выразим через уравнение:

$$\log(\bar{x}) - \log(k) + \psi(k) - \frac{1}{n}\sum_i \log(x_i) = 0$$

```
estimate_mmp <- function(samp){
  C <- log(mean(samp)) - sum(log(samp)) / length(samp)

  Credibility_func <- function(al, const){
    log(al) - digamma(al) + const
  }

  k <- uniroot(Credibility_func, const = -C,
  lower = 10e-10, upper = 1000, extendInt = "yes")$root
  r <- mean(samp)/k
  return(c(k, r))
}

first <- rgamma(100, shape = 3, scale = 2)
second <- rgamma(300, shape = 3, scale = 2)
```

Оценки параметров для первой выборки:

```
k_f <- estimate_mmp(first)[1]
r_f <- estimate_mmp(first)[2]
c(k_f, r_f)

## [1] 3.484572 1.811735
```

Оценки параметров для второй выборки:

```
k_s <- estimate_mmp(second)[1]
r_s <- estimate_mmp(second)[2]
c(k_s, r_s)

## [1] 3.396228 1.737584
```

Построим доверительный интервал для оценок:

```
D_k <- 1 / (length(first) * (psigamma(k_f) - 1 / k_f))
D_r <- r_f^2 / (length(first) * (k_f - 1 / psigamma(10)))

D_k2 <- 1 / (length(second) * (psigamma(k_s) - 1 / k_s))
D_r2 <- r_s^2 / (length(second) * (k_s - 1 / psigamma(k_s)))
```

Для параметра k, 95% доверительный интервал первой и второй выборки соответственно:

```
c(k_f - qnorm(0.975)*sqrt(D_k / n1), k_f + qnorm(0.975)*sqrt(D_k / n1))

## [1] 3.462809 3.506335

c(k_s - qnorm(0.975)*sqrt(D_k2 / n2), k_s + qnorm(0.975)*sqrt(D_k2 / n2))

## [1] 3.388801 3.403655
```

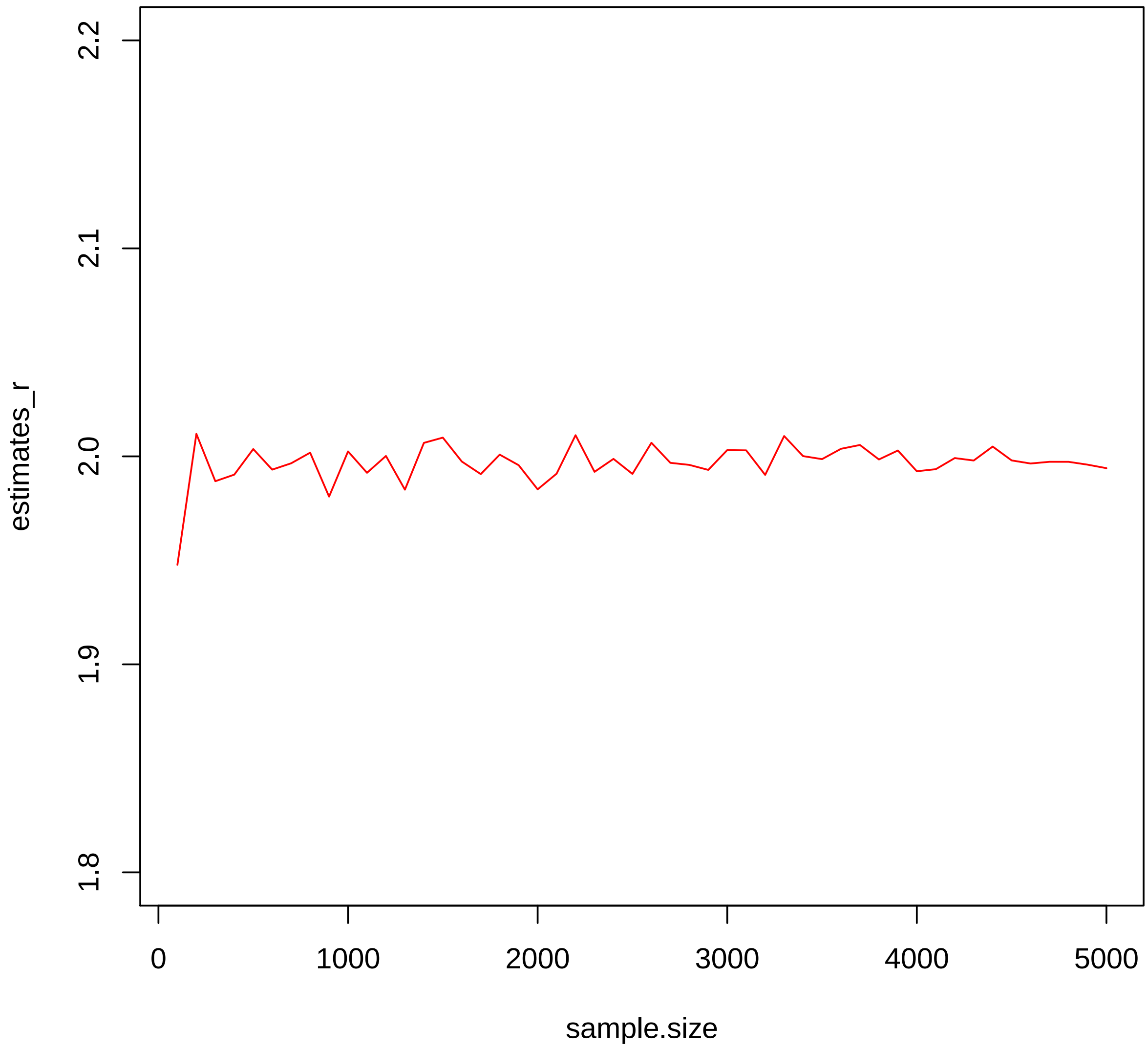
Для параметра r, 95% доверительный интервал первой и второй выборки соответственно:

```
c(r_f - qnorm(0.975)*sqrt(D_r / n1), r_f + qnorm(0.975)*sqrt(D_r / n1))

## [1] 1.791370 1.832099

c(r_s - qnorm(0.975)*sqrt(D_r2 / n2), r_s + qnorm(0.975)*sqrt(D_r2 / n2))

## [1] 1.730346 1.744822
```

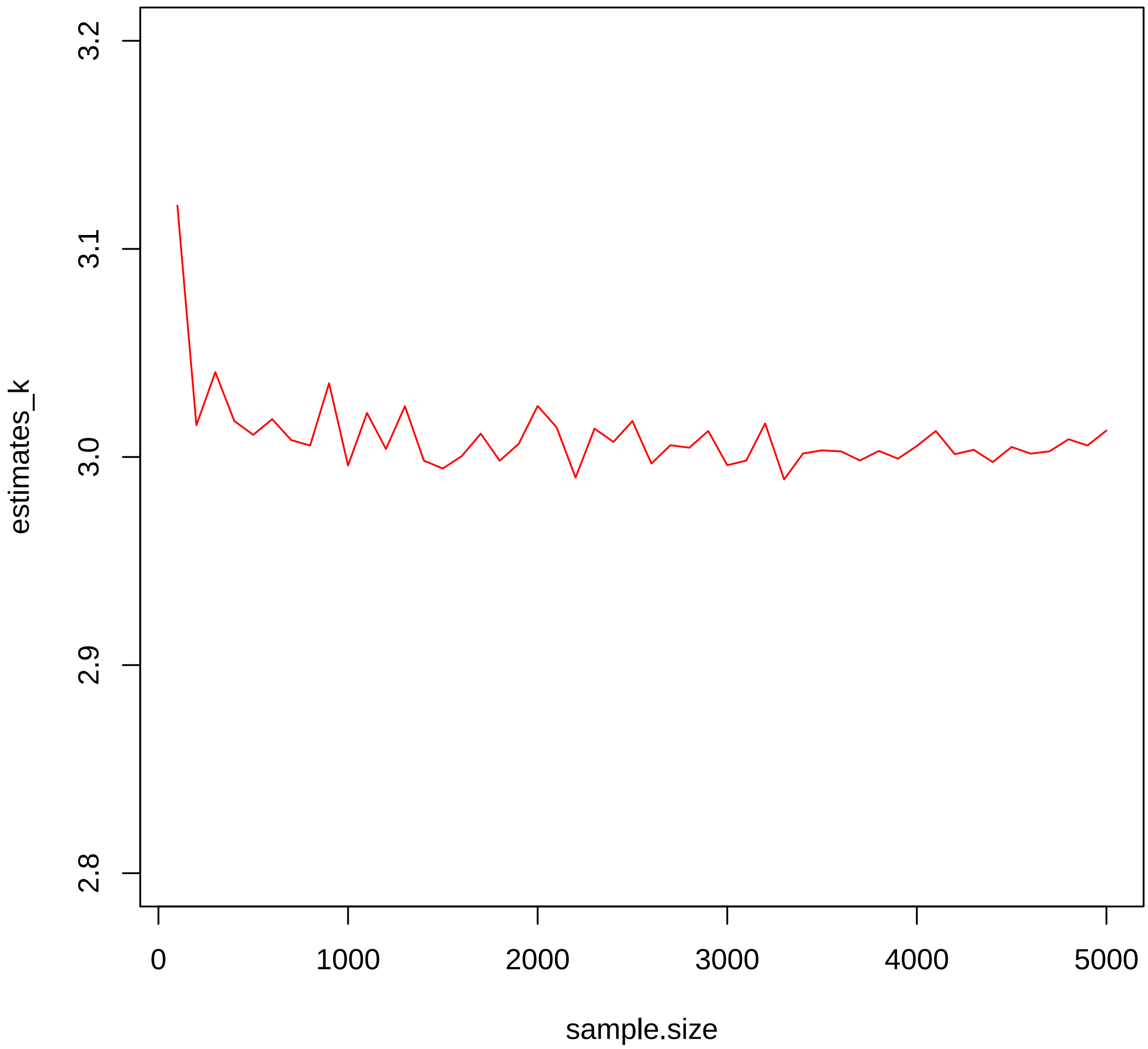
оценим несмещенность оценок:

```
sample.size <- seq(100, 5000, by = 100)

estimates_r <- character()
estimates_k <- character()

for (size in sample.size) {
  temp_r <- character()
  temp_k <- character()
  for (i in 1:100) {
    temp <- rgamma(size, shape = 3, scale = 2)
    y <- estimate_mmp(temp)
    temp_k <- c(temp_k, y[1])
    temp_r <- c(temp_r, y[2])
  }

  temp_k <- as.numeric(temp_k)
  temp_r <- as.numeric(temp_r)

  estimates_r <- c(estimates_r, mean(temp_r))
  estimates_k <- c(estimates_k, mean(temp_k))
}

estimates_k <- as.numeric(estimates_k)
```

```r
estimates_r <- as.numeric(estimates_r)

plot(sample.size, estimates_r, type = "l", col = "red", ylim=c(1.8, 2.2))
```




```r
plot(sample.size, estimates_k, type = "l", col = "red", ylim=c(2.8, 3.2))
```



Для обеих оценок из графика устанавливаем, что при увеличении объема выборки, среднее сходится к истинному значению, делаем вывод, что оценки несмещенные.

Для оценок ММП известно, что они являются состоятельными и ассимптотически эффективными.