# Laboratorium: Lista 5

## Lokalizacja genów z wykorzystaniem testu Studenta II

Statystyka w zastosowaniach

Sylwia Majchrowska
Matematyka

24 kwietnia 2016

## Spis treści

# 1 Wstęp

W dalszym ciągu zajmujemy się lokalizacją genów z wykorzystaniem testu Studenta z tą różnicą, że tym razem generowana przez nas cecha jest zależna od czynników genetycznych, a dokładnie skorelowana z markerem ulokowanym pośrodku badanego chromosomu (tzn. $\delta = 100\,\text{cM}$). Rozpatrujemy krzyżówkę wsteczną, czyli taką, w której krzyżujemy osobniki homozygotyczne recesywne z heterozygotami, przy czym odpowiedni genotyp kodujemy za pomocą liczb 0 i 1.

# 2 Zadanie pierwsze

Aby wygenerować macierz genotypów dla $n = 500$ osobników z krzyżówki wstecznej na jednym chromosomie o długości 200 cM i odstępie między sąsiednimi markerami $\Delta = 1\,\text{cM}$ posłużono się przykładowym kodem do generacji genotypów na jednym chromosomie krzyżówki wstecznej omawianym na wykładzie.

Oznaczmy przez $M_{j,k}$ genotyp w $j$-tym markerze u $k$-tego osobnika, taki, że

$$M_{j,k} = \begin{cases} 0 & \text{aa} \\ 1 & \text{Aa}. \end{cases}$$

Kolejnym krokiem było wygenerowanie wektora wartości cechy $Y_i$, którego wartości ściśle zależą od genotypu genu zlokalizowanego pośrodku chromosomu (tzn. w $\delta = 100\,\text{cM}$), w taki sposób, że

$$Y_i \sim \begin{cases} N(\beta, 1) & \text{dla } M_{100,i} = 0 \\ N(0, 1) & \text{dla } M_{100,i} = 1, \end{cases}$$

gdzie $\beta \in \{0.2, 0.35, 0.5\}$.

Estymacji położenia genu dokonano zgodnie ze wzorem

$$\hat{\delta} = \arg\max_{i \in \{0,\ldots,200\}} |t_i|,$$

gdzie $t_i$ jest statystyką testu Studenta w położeniu $i\,\text{cM}$ licząc od lewego końca chromosomu. Dodatkowo dla danego estymatora skonstruowany został 95% przedział ufności

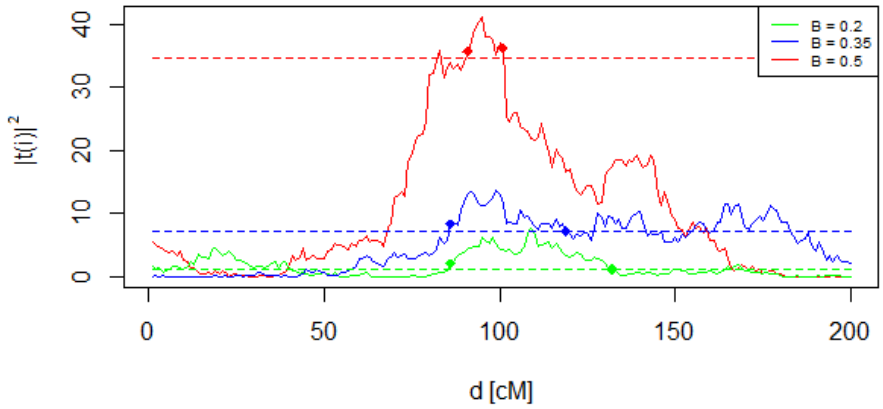
$$\Delta = \{i : t_i^2 > t_{\hat{\delta}}^2 - 6.6\}.$$

Wyznaczono także wielkość efektu genetycznego $\hat{\beta} = \overline{Y}_{aa,\hat{\delta}} - \overline{Y}_{Aa,\hat{\delta}}$, który ilustruje różnicę pomiędzy dwoma grupami genotypów (estymator przyjętego $\beta$).

Uzyskane wielkości estymatorów:

| $\beta$ | $\hat{\delta}$ | $\Delta = [\min, \max]$ | $\hat{\beta}$ |
|---|---|---|---|
| 0.20 | 109 | [1, 169] | 0.2509961 |
| 0.35 | 99 | [86, 183] | 0.3386601 |
| 0.50 | 95 | [83, 101] | 0.5736259 |

Uzyskane wielkości parametrów są zbliżone do ich wartości rzeczywistej. $\delta = 100$ wpada do każdego z wyznaczonych przedziałów ufności, które wraz ze wzrostem parametru $\beta$ stają się coraz szersze. Wyestymowana wielkość efektu genetycznego jest większa od swojej rzeczywistej wartości.




Rysunek 1: Wykresy zależności kwadratu wartości statystyki Studenta od odległości liczonej od lewego końca chromosomu wraz z naniesionymi prostymi odcięcia (przerywane linie) dla trzech różnych wartości $\beta$.

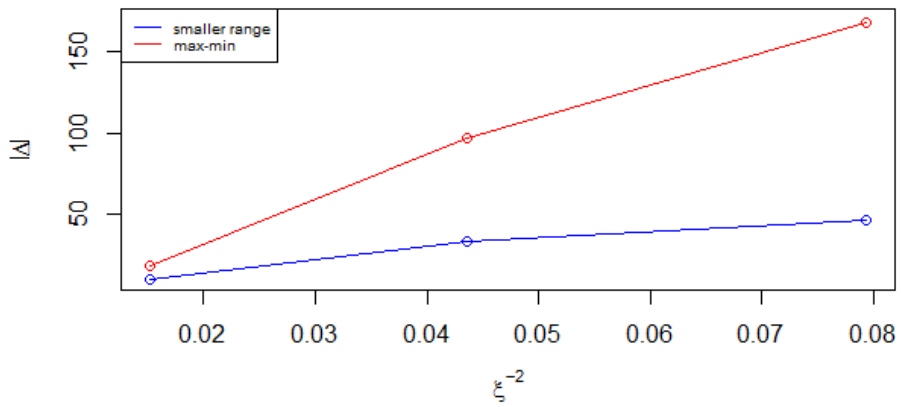
Na wykreślonych wykresach zależności wartości statystyki Studenta od położenia markera widać dość wyraźne zwiększenie jej wartości w okolicach 100 cM. Jest ono tym większe im większy dobierzemy parametr $\beta$. Można także zauważyć, że wyznaczony przedział ufności dla parametru położenia genu nie jest ciągły (odpowiednie statystyki znajdujące się powyżej przerywanych linii). Na wykresie 1 pełnymi punktami zaznaczono miejsca minimów i maksimów podprzedziałów $\Delta$, które zawierają wyznaczony estymator. Dodatkowo sporządzono wykres zależności szerokości przedziału od $\xi^{-2}$, gdzie

$$\xi = \beta\sqrt{n}.$$

Szerokość przedziału ufności $\Delta$ kurczy się wraz ze wzrostem wielkości efektu genetycznego (a tym samym $\xi$) - maleje w tempie $1/n$. Z tego faktu płynie prosty wniosek - im mamy silniejszy sygnał tym lepiej jesteśmy w stanie zlokalizować gen.

Uzyskane szerokości przedziałów:

| typ | $\beta = 0.2$ | $\beta = 0.35$ | $\beta = 0.5$ |
|---|---|---|---|
| min-max | 168 | 97 | 18 |
| smaller range | 46 | 33 | 10 |




Rysunek 2: Szerokość przedziału ufności dla lokalizacji genu zależy proporcjonalnie od $\xi^{-2}$. Kolorem czerwonym zaznaczono szerokość przedziału otrzymaną na podstawie odnalezienia najmniejszej i największej wartości należących do zbioru $\Delta$, niebieskim natomiast - należących do podzbioru opisanego powyżej.

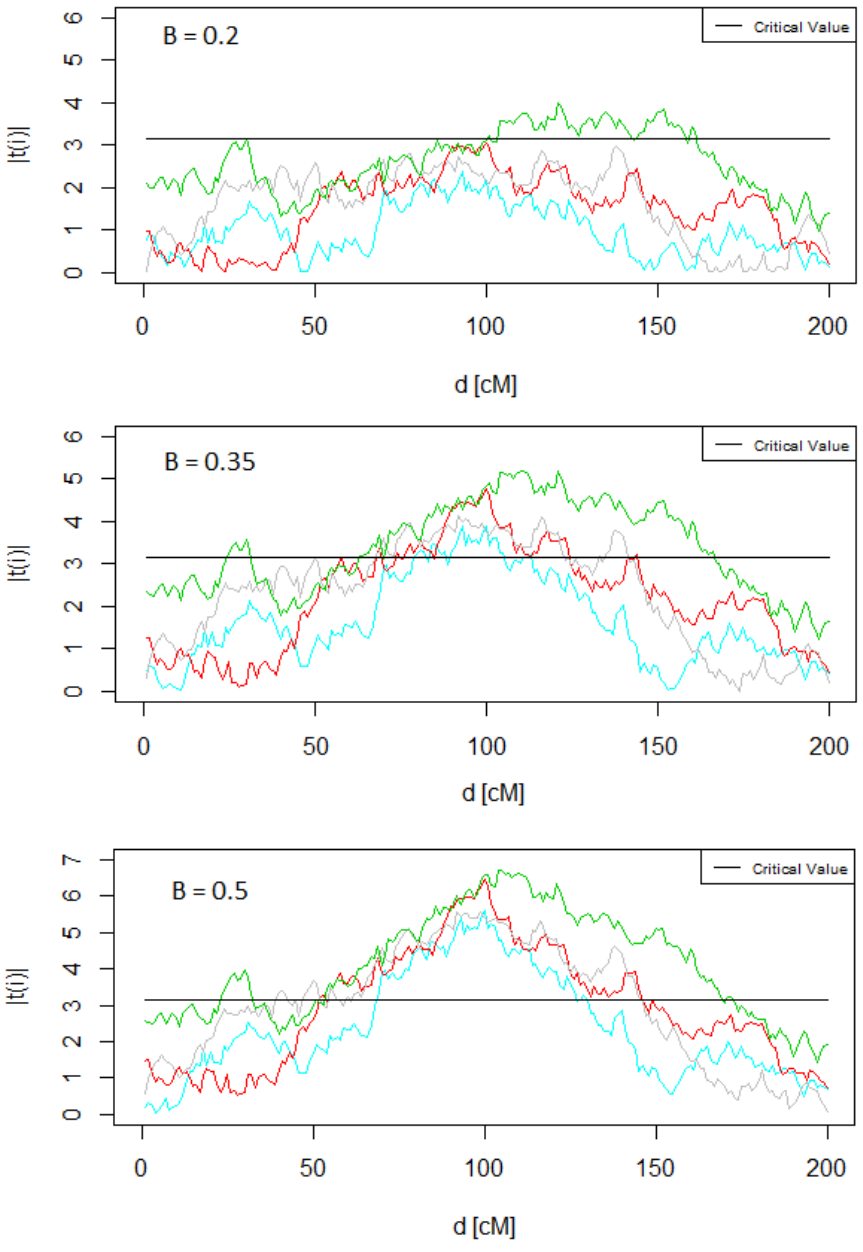
Kod źródłowy zadania 1 podpunktów a) - d) w języku R:

```
# a)
set.seed(1222)
# ------------------------------------------------------------
# Autor: M. Bogdan
# Generacja genotypow na jednym chromosomie z krzyzowki
    wstecznej
n <- 500 # liczba osobnikow - wiersze
L <- 200 # dlugosc chromosomu w cM - liczba markerow,
    kolumny;
d <- 1 # odleglosc miedzy sasiednimi markerami w cM
r <- 0.5*(1-exp(-0.02*d)) # p-stwo rekombinacji miedzy
    sasiednimi markerami
P <- rbinom(n,1,0.5) # genotypy w pierwszym markerze
R <- rbinom(n*(L-1),1,r)
R <- matrix(R,nrow=n,ncol=(L-1)) # macierz rekombinacji
    miedzy sasiednimi markerami
Y <- cbind2(P,R)
Y <- apply(Y,1,'cumsum')
Y <- t(Y)
Y <- Y%%2 # finalna macierz genotypow
# ------------------------------------------------------------

# b)
set.seed(251)
beta = c(0.2,0.35,0.5)
wiersze = cbind(Y[,100],Y[,100],Y[,100])
wiersze[,1] = wiersze[,1] * rnorm(500) + (1 - wiersze
    [,1]) * rnorm(500,beta[1])
```

```
wiersze[,2] = wiersze[,2] * rnorm(500) + (1 - wiersze
    [,2]) * rnorm(500,beta[2])
wiersze[,3] = wiersze[,3] * rnorm(500) + (1 - wiersze
    [,3]) * rnorm(500,beta[3])
# c)
Stat02 = sapply(1:L, function(i){t.test(wiersze[,1]~Y[,i
    ], var.equal = TRUE)$statistic})
Stat035 = sapply(1:L, function(i){t.test(wiersze[,2]~Y[,i
    ], var.equal = TRUE)$statistic})
Stat05 = sapply(1:L, function(i){t.test(wiersze[,3]~Y[,i
    ], var.equal = TRUE)$statistic})
plot(abs(Stat05)^2, xlab = 'd [cM]', ylab = '|t(i)|', col
    = 'red', type = 'l')
lines(abs(Stat02)^2, col = 'green', type='l')
lines(abs(Stat035)^2, col = 'blue', type='l')

d02 = which.max(abs(Stat02))
d035 = which.max(abs(Stat035))
d05 = which.max(abs(Stat05))
delta02 = which(Stat02^2 > (Stat02[d02]^2 - 6.6))
delta035 = which(Stat035^2 > (Stat035[d035]^2 - 6.6))
delta05 = which(Stat05^2 > (Stat05[d05]^2 - 6.6))

lines(rep(Stat05[d05]^2 - 6.6,200), type="l", pch=22, lty
    =2, col="red")
lines(rep(Stat035[d035]^2 - 6.6,200), type="l", pch=22,
    lty=2, col="blue")
lines(rep(Stat02[d02]^2 - 6.6,200), type="l", pch=22, lty
    =2, col="green")
points(c(86,132),c(Stat02[86]^2,Stat02[132]^2), cex=1,
    pch=18, col='green')
points(c(86,119),c(Stat035[86]^2,Stat035[119]^2), cex=1,
    pch=18, col='blue')
points(c(91,101),c(Stat05[91]^2,Stat05[101]^2), cex=1,
    pch=18, col='red')
legend("topright", c('B = 0.2', 'B = 0.35', 'B = 0.5'),
    col = c('green','blue','red'), text.col = "black", lty
    = 1, cex=0.6)
# d)
efg02 = mean(wiersze[which(Y[,d02] == 0), 1]) - mean(
    wiersze[which(Y[,d02] == 1), 1])
efg035 = mean(wiersze[which(Y[,d035] == 0), 2]) - mean(
    wiersze[which(Y[,d035] == 1), 2])
efg05 = mean(wiersze[which(Y[,d05] == 0), 3]) - mean(
    wiersze[which(Y[,d05] == 1), 3])

dlugosci2 = c(132-86,119-86,101-91)
dlugosci = c(max(delta02)-min(delta02), max(delta035)-min
    (delta035),max(delta05)-min(delta05))
```

## 2.1 Podpunkt e)

Aby wyznaczyć obciążenie oraz odchylenie standardowe wyznaczonych estymatorów położenia i efektu genetycznego, a także prawdopodobieństwo pokrycia dla przedziału ufności doświadczenie powtórzono 1000 dla każdej wartości $\beta$.




Rysunek 3: Wybrane zależności wyliczonych statystyk Studenta od położenia wraz z ustaloną na podstawie wzoru Feingolda granicą odcięcia (kolor czarny).

Jak można zauważyć na wykresach 3 nie wszystkie maksymalne statystyki Studenta dla 1000 powtórzeń eksperymentu przekraczają wartość krytyczną wyznaczoną za pomocą korekty Feingolda i reszty. Dodatkowo dla $\beta = 0.2$ wzrost wartości statystyki w okolicach 100cM jest nieznaczny.

Uzyskane wyniki:

| $\beta$ | $\sigma(\hat{\delta})$ | $E(\hat{\delta}) - \delta$ | $\sigma(\hat{\beta})$ | $E(\hat{\beta}) - \beta$ | p-stwo pokrycia |
|---|---|---|---|---|---|
| 0.20 | 39.06462189 | 0.019 | 0.08368217 | 0.01550971 | $0.905 \pm 0.041$ |
| 0.35 | 16.53864150 | -1.308 | 0.06450578 | 0.01286878 | $0.935 \pm 0.035$ |
| 0.50 | 5.91772046 | -0.402 | 0.062589016 | 0.007980207 | $0.947 \pm 0.032$ |

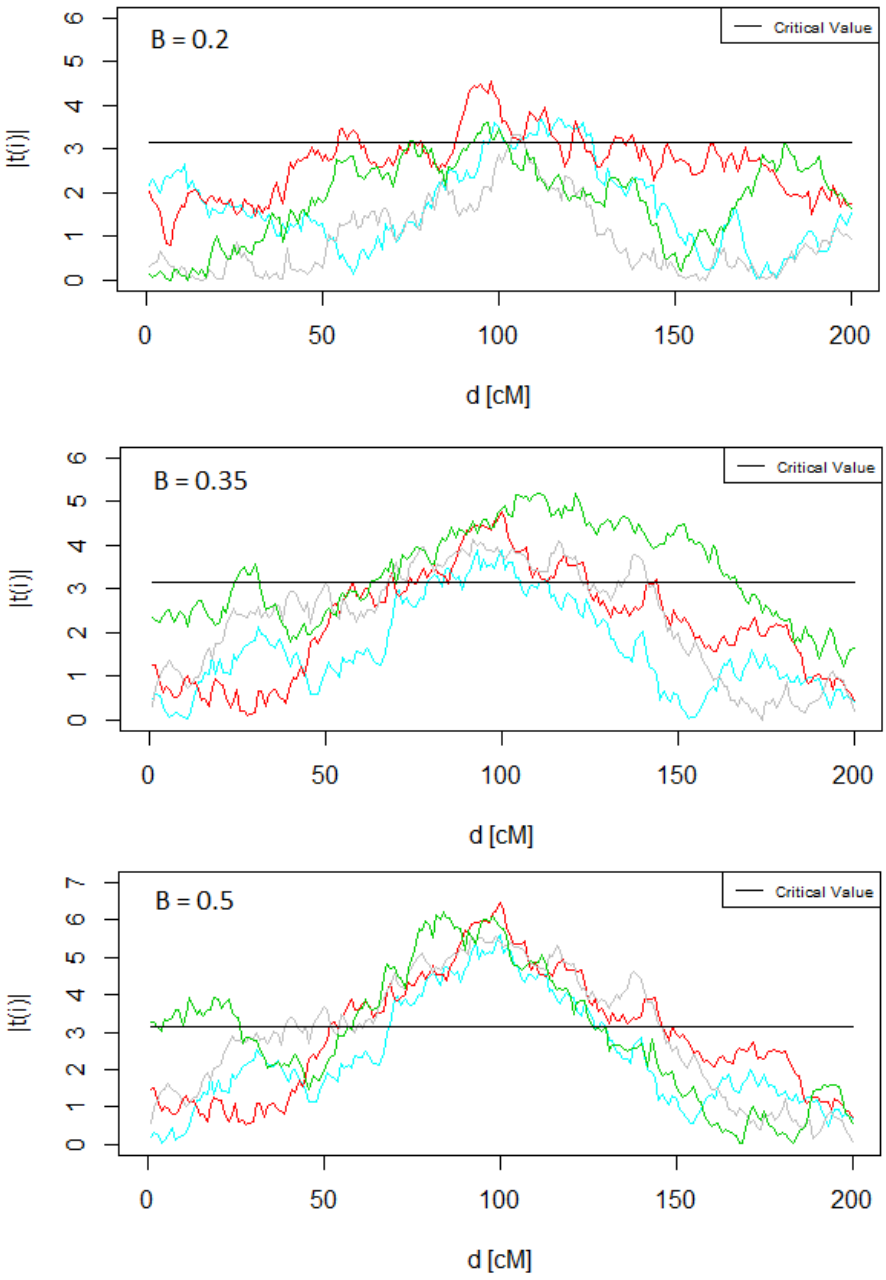
Wraz ze wzrostem parametru $\beta$ maleje odchylenie standardowe estymatora położenia (podobnie jak w przypadku szerokości przedziału ufności). Występuje tu dość szeroki rozrzut wartości $\hat{\delta}$ w przypadku słabych sygnałów. Oszacowane obciążenie dla tego parametru jest bliskie zeru, jednakże ujemna różnica między średnią, a jego wartością rzeczywistą wskazuje na delikatne zaniżenie wielkości w stosunku do szacowanego parametru. W przypadku wielkości efektu genetycznego nie mamy do czynienia z dużą rozbieżnością między wartością oczekiwaną estymatora, a rzeczywistą $\beta$, co świadczy o tym iż dany parametr nie jest obciążony. Jednocześnie różnica ta maleje wraz ze wzrostem $\beta$ i pozostaje dodatnia (zawyżone oceny - przeszacowanie efektu genu). Prawdopodobieństwo pokrycia dla przedziału ufności także rośnie razem z $\beta$ i już dla wartości $\beta = 0.35$ wielkość 0.95 mieści się w wyznaczonym przedziale ufności dla przedziału ufności.

Kod źródłowy zadania 1 podpunktu e) w języku R:

```
# plot(1000, 100, xlim=c(0,200), ylim=c(0,7), xlab = 'd [
    cM]', ylab = '|t(i)|', col = 'red', type = 'l')
doswiadczenie <- function(b, sumNum, genotypy, nr_markera
    ){
  set.seed(6532)
  osobniki = length(genotypy[,1])  # 500
  L = length(genotypy[1,])  # 200 = dlugosc chromosomu w
      cM
  cecha = matrix(rnorm(osobniki * sumNum), osobniki,
      sumNum)
  wiersze = cbind(genotypy[,nr_markera],genotypy[,nr_
      markera])
  for (ind in 3:sumNum){
    wiersze = cbind(wiersze,genotypy[,nr_markera])
  }
  for (i in 1:sumNum){
    wiersze[,i] = wiersze[,i] * rnorm(500) + (1 - wiersze
        [,i]) * rnorm(500,b)
  }
```

```
  Stat = matrix(0,L,sumNum)
  for(j in 1:sumNum){
    cecha = wiersze[,j]
    for (i in 1:L){
      gr1 = (genotypy[,i] == 0) * cecha
      gr1 = gr1[gr1 != 0]
      gr2 = (genotypy[,i] == 1) * cecha
      gr2 = gr2[gr2 != 0]
      Stat[i,j] = t.test(gr1,gr2, var.equal = TRUE)$
          statistic
    }
  }
  # lines(abs(Stat[,5]), col = 5, type='l')
  # lines(abs(Stat[,100]), col = 2, type='l')
  # lines(abs(Stat[,200]), col = j, type='l')
  # lines(abs(Stat[,1000]), col = 3, type='l')

  d = apply(abs(Stat), 2, which.max)
  efg = rep(0,sumNum)
  for (indeks in 1:sumNum) {
    efg[indeks] = mean(wiersze[which(genotypy[,d[indeks]]
        == 0), indeks])
    - mean(wiersze[which(genotypy[,d[indeks]] == 1),
        indeks])
  }

  prawd = 0.0
  for (inde in 1:sumNum){
    s = Stat[,inde]
    delta = which(s^2 > (s[d[inde]]^2 - 6.6))
    if (nr_markera %in% delta){
      prawd = prawd + 1
    }
  }
  prawd = prawd/sumNum

  return (c(sd(d), mean(d) - nr_markera, sd(efg), mean(
      efg) - b, prawd))
}

w1 = doswiadczenie(0.2, 1000, Y, 100)
w2 = doswiadczenie(0.35, 1000, Y, 100)
w3 = doswiadczenie(0.5, 1000, Y, 100)

bladI = qnorm(1-0.05/2)*sqrt(w1[5]*(1-w1[5])/200)
bladII = qnorm(1-0.05/2)*sqrt(w2[5]*(1-w2[5])/200)
bladIII = qnorm(1-0.05/2)*sqrt(w3[5]*(1-w3[5])/200)
```

## 2.2 Podpunkt f)




Rysunek 4: Wybrane zależności wyliczonych statystyk Studenta przekraczających wartość krytyczną od położenia wraz z ustaloną na podstawie wzoru Feingolda granicą odcięcia (kolor czarny).

W tym przypadku dołożono jeden krok testowania - położenie genu oraz jego efekt estymowano tylko wtedy, gdy wartość statystyki Studenta przekroczyła wartość krytyczną. Na tej podstawie oszacowano moc detekcji genu dla różnych $\beta$, otrzymując:

- dla $\beta = 0.20$ -> P(odrzucamy $H_0$) = 0.370,
- dla $\beta = 0.35$ -> P(odrzucamy $H_0$) = 0.877,
- dla $\beta = 0.50$ -> P(odrzucamy $H_0$) = 0.998.

Jak można się było spodziewać moc detekcji rośnie wraz ze wzrostem wielkości efektu genetycznego.

Uzyskane wyniki:

| $\beta$ | $\sigma(\hat{\delta})$ | $E(\hat{\delta}) - \delta$ | $\sigma(\hat{\beta})$ | $E(\hat{\beta}) - \beta$ | p-stwo pokrycia |
|---|---|---|---|---|---|
| 0.20 | 29.95413386 | 1.100 | 0.04922820 | 0.06551125 | 0.800 ± 0.056 |
| 0.35 | 14.48013762 | -1.097 | 0.05651464 | 0.02378205 | 0.927 ± 0.037 |
| 0.50 | 5.88736395 | -0.380 | 0.062427231 | 0.008217024 | 0.947 ± 0.032 |

W tym przypadku oszacowane parametry zachowują się podobnie jak dla podpunktu e) z tą różnicą, że są trochę mniejsze, co jest związane ze zmniejszeniem liczby obserwacji. Jednocześnie można zauważyć, że wyznaczony efekt genetyczny przy przekroczeniu ustalonej bariery jest generalnie większy niż prawdziwa wartość. Potwierdzono więc informację podaną na wykładzie, że estymacja i lokalizacja nie powinna odbywać się na tych samych danych, gdyż może to doprowadzić do przeszacowania parametrów.

Kod źródłowy zadania 1 podpunktu f) w języku R:

```
ni <- function(t){
  wn = (2/t)*(pnorm(t/2)-0.5)/((t/2)*pnorm(t/2)+dnorm(t/
      2))
  return (wn)
}

FBS <- function(C,L,delta,tc){
  ff = 1 - exp(-2*C*(1-pnorm(tc))-0.04*L*tc*dnorm(tc)*ni(
      tc*sqrt(0.04*delta)))
  return (ff)
}

tt = seq(1.001, 4.001, by = 0.0001)
fun = rep(0, length(tt))
for (index in 1:length(tt)){
  fun[index] = FBS(1, 200, 1, tt[index])
}
```

```
plot(tt,fun, type = 'l', xlab = expression(t[c]), ylab =
    expression(alpha))

szukane = 0.05
pozycja = 1
for (kk in 1:length(fun))
{
  if(abs(fun[kk] - szukane) < abs(fun[pozycja] - szukane)
      ){
    pozycja = kk
  }
}
critical_value = tt[pozycja] # tt[21378] = 3.1387
lines(critical_value, fun[pozycja], type ='p', col='red')

# plot(1000, 100, xlim=c(0,200), ylim=c(0,7), xlab = 'd [
    cM]', ylab = '|t(i)|', col = 'red', type = 'l')
doswiadczenie <- function(b, sumNum, genotypy, nr_markera
    ){
  set.seed(6532)
  osobniki = length(genotypy[,1])  # 500
  L = length(genotypy[1,])  # 200 = dlugosc chromosomu w
      cM
  cecha = matrix(rnorm(osobniki * sumNum), osobniki,
      sumNum)
  wiersze = cbind(genotypy[,nr_markera],genotypy[,nr_
      markera])
  for (ind in 3:sumNum){
    wiersze = cbind(wiersze,genotypy[,nr_markera])
  }

  for (i in 1:sumNum){
    wiersze[,i] = wiersze[,i] * rnorm(500) + (1 - wiersze
        [,i]) * rnorm(500,b)
  }

  Stat = matrix(0,L,sumNum)
  for(j in 1:sumNum){
    cecha = wiersze[,j]
    for (i in 1:L){
      gr1 = (genotypy[,i] == 0) * cecha
      gr1 = gr1[gr1 != 0]
      gr2 = (genotypy[,i] == 1) * cecha
      gr2 = gr2[gr2 != 0]
      Stat[i,j] = t.test(gr1,gr2, var.equal = TRUE)$
          statistic
    }
  }
```

```
  maxwar = apply(Stat, 2, max)
  pwar = rep(0, length(maxwar))
  for (ind in 1:sumNum){
    pwar[ind] = FBS(1,L,1,abs(maxwar[ind]))
  }
  numery = which(pwar < 0.05)  # kolumny Stat, dla
      ktorych odrzucamy H0
  moc = length(numery)/sumNum
  sumNum = length(numery)
  nStat = Stat[,numery[1]]
  nwiersze = wiersze[,numery[1]]
  for (indy in numery[2:(length(numery))]){
    nStat = cbind(nStat,Stat[,indy])
    nwiersze = cbind(nwiersze,wiersze[,indy])
  }
  Stat = nStat
  rm(nStat)
  wiersze = nwiersze
  rm(nwiersze)
  # lines(abs(Stat[,5]), col = 5, type='l')
  # lines(abs(Stat[,100]), col = 2, type='l')
  # lines(abs(Stat[,200]), col = j, type='l')
  # lines(abs(Stat[,300]), col = 3, type='l')
  d = apply(abs(Stat), 2, which.max)
  efg = rep(0,sumNum)
  for (indeks in 1:sumNum) {
    efg[indeks] = mean(wiersze[which(genotypy[,d[indeks]]
        == 0), indeks])
    - mean(wiersze[which(genotypy[,d[indeks]] == 1),
        indeks])
  }
  prawd = 0.0
  for (inde in 1:sumNum){
    s = Stat[,inde]
    delta = which(s^2 > (s[d[inde]]^2 - 6.6))
    if (nr_markera %in% delta){
      prawd = prawd + 1
    }
  }
  prawd = prawd/sumNum
  return (c(moc, sd(d), mean(d) - nr_markera, sd(efg),
     mean(efg) - b, prawd))
}
w1 = doswiadczenie(0.2, 1000,Y,100)
w2 = doswiadczenie(0.35, 1000,Y,100)
w3 = doswiadczenie(0.5, 1000,Y,100)
bladI = qnorm(1-0.05/2)*sqrt(w1[6]*(1-w1[6])/200)
bladII = qnorm(1-0.05/2)*sqrt(w2[6]*(1-w2[6])/200)
bladIII = qnorm(1-0.05/2)*sqrt(w3[6]*(1-w3[6])/200)
```