Analiza podatko študentov biologije
3. letnik 2011/12

Andrej Blejec

19. januar 2012

Vnos podatkov

```
> data <- read.table(file.path("../data",
+     lfn), sep = "\t", header = TRUE)
> str(data)

'data.frame': 51 obs. of  12 variables:
 $ starost: int  58 23 21 20 21 20 21 21 21 23 ...
 $ mesec  : int  7 3 4 10 1 4 10 4 7 1 ...
 $ spol   : Factor w/ 2 levels "F ženski ","M moški ": 2
 $ masa   : int  88 54 42 60 52 56 58 77 60 95 ...
 $ visina : num  178 174 152 170 171 170 170 169 170 190
 $ roke   : int  175 171 150 155 171 174 169 169 168 201
 $ cevelj : int  44 40 37 39 39 40 39 40 39 46 ...
 $ lasje  : Factor w/ 2 levels "S svetli ","T temni ": 2
 $ oci    : Factor w/ 2 levels "S svetli ","T temni ": 1
 $ majica : Factor w/ 6 levels "L ","M ","S ",..: 1 3 6
 $ mati   : num  150 173 157 160 173 174 155 170 155 174
 $ oce    : int  180 189 159 190 183 186 179 175 189 176
```

Vrste spremenljivk

Opisne spremenljivke

```
> opisne <- which(sapply(data, "class") ==
+     "factor")
> length(opisne)
[1] 4
> names(data)[opisne]
[1] "spol"   "lasje"  "oci"    "majica"
```

Številske spremenljivke

```
> stevilske <- which(!sapply(data, "class") ==
+     "factor")
> length(stevilske)
[1] 8
> names(data)[stevilske]
[1] "starost" "mesec"   "masa"     "visina"
[5] "roke"    "cevelj"  "mati"     "oce"
```

Opisna statistika

```
> summary(data[, 1:6])
```

starost		mesec		spol	
Min.	:20.00	Min.	: 0.000	F ženski	:40
1st Qu.	:21.00	1st Qu.	: 3.500	M moški	:11
Median	:21.00	Median	: 6.000		
Mean	:21.98	Mean	: 5.922		
3rd Qu.	:22.00	3rd Qu.	: 9.000		
Max.	:58.00	Max.	:12.000		

masa		visina		roke	
Min.	:42.00	Min.	:152.0	Min.	: 0.0
1st Qu.	:55.50	1st Qu.	:165.0	1st Qu.	:160.0
Median	:60.00	Median	:170.0	Median	:168.0
Mean	:62.76	Mean	:170.4	Mean	:151.5
3rd Qu.	:65.50	3rd Qu.	:173.0	3rd Qu.	:171.5
Max.	:95.00	Max.	:196.0	Max.	:201.0

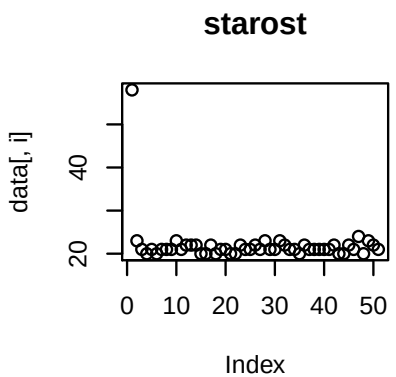
Opisna statistika

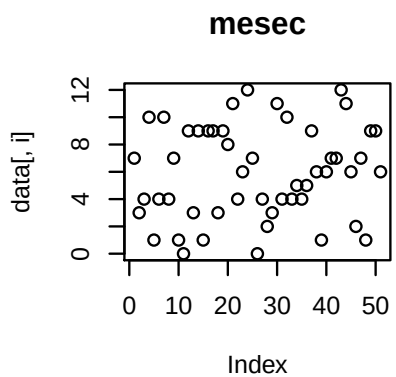
```
> summary(data[, 7:12])
```

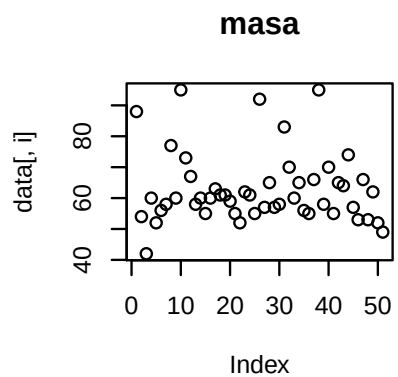
```
     cevelj             lasje              oci
 Min.   :29.00   S svetli :19   S svetli :28
 1st Qu.:38.00   T temni  :32   T temni  :23
 Median :39.00
 Mean   :39.71
 3rd Qu.:41.00
 Max.   :49.00
  majica        mati             oce
 L   : 3   Min.   :  0.0   Min.   :  0.0
 M   :25   1st Qu.:159.0   1st Qu.:175.0
 S   :18   Median :164.0   Median :179.0
 XL  : 1   Mean   :156.2   Mean   :168.9
 XS  : 3   3rd Qu.:170.0   3rd Qu.:182.0
 XXS : 1   Max.   :178.0   Max.   :190.0
```

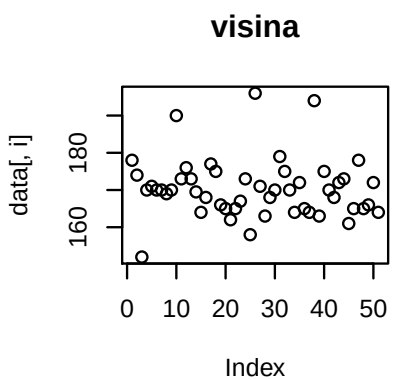
Grafični pregled podatkov

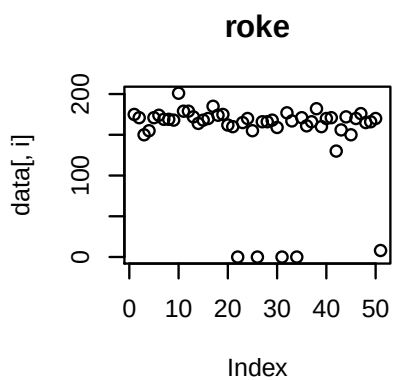
```
> par(mfrow = c(3, 3))
> for (i in stevilske) plot(data[, i], main = names(data
```

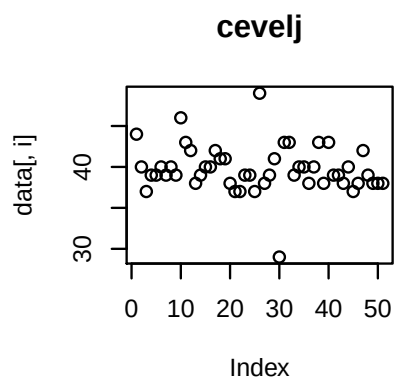












mati oce

Odstranitev enot z manjkajočimi podatki

```
> nas <- which(apply(data, 1, function(x) any(as.numeric
+     0)))
> nas
```

```
[1] 11 22 26 27 31 34
```

```
> if (length(nas) > 0) data <- data[-nas,
+     ]
> dim(data)
```

```
[1] 45 12
```

Odstranitev podatkov

Odstranimo podatke prestarega človeka :)

```
> data <- data[data$starost < 25, ]
```

in še enega kratkorokega

```
> data <- data[data$roke > 50, ]
```

Popravek podatkov

```
> data[, "mati"] <- data[, "mati"] + 100 *
+     (data[, "mati"] < 100)
> data[, "oce"] <- data[, "oce"] + 100 *
+     (data[, "oce"] < 100)
```

Ureditev imen nivojev

Odstranimo presledke v vrednostih

```
> x <- data[, "majica"]
> levels(x)

[1] "L "   "M "   "S "   "XL "  "XS "  "XXS "

> levels(x) <- gsub("(.*) ", "\\1", levels(x))
> levels(x)

[1] "L"   "M"   "S"   "XL"  "XS"  "XXS"
```

Urejenostna merska lestvica

Spremenimo vrstni red nivojev

```
> (data[, "majica"] <- ordered(x, levels = c("XXS",
+     "XS", "S", "M", "L", "XL")))
 [1] S  XXS M  M  S  M  M  S  XL S  M
[12] M  S  M  M  M  M  S  S  S  M  S
[23] M  M  S  L  S  S  S  M  M  M  S
[34] XS S  M  M  L  M  M  XS M  S
Levels: XXS < XS < S < M < L < XL
```

Opisna statistika

```
> summary(data[, 1:6])
```

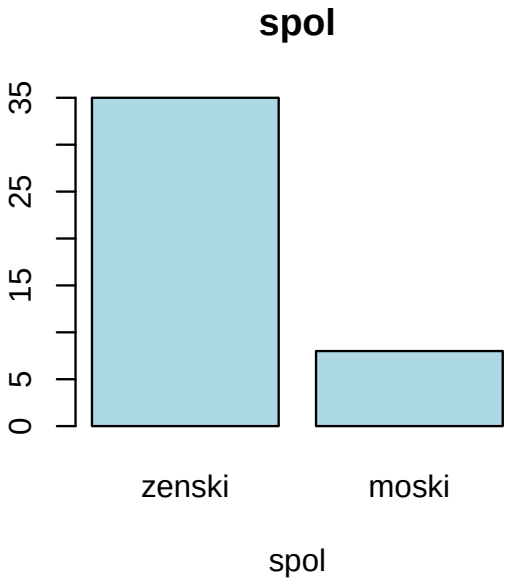
```
    starost          mesec            spol
 Min.   :20.00   Min.   : 1.000   zenski:35
 1st Qu.:21.00   1st Qu.: 3.500   moski : 8
 Median :21.00   Median : 7.000
 Mean   :21.26   Mean   : 6.326
 3rd Qu.:22.00   3rd Qu.: 9.000
 Max.   :24.00   Max.   :12.000
      masa            visina           roke
 Min.   :42.00   Min.   :152.0   Min.   :130.0
 1st Qu.:55.50   1st Qu.:165.0   1st Qu.:163.0
 Median :60.00   Median :170.0   Median :169.0
 Mean   :61.44   Mean   :169.7   Mean   :167.4
 3rd Qu.:64.50   3rd Qu.:173.0   3rd Qu.:171.5
 Max.   :95.00   Max.   :194.0   Max.   :201.0
```

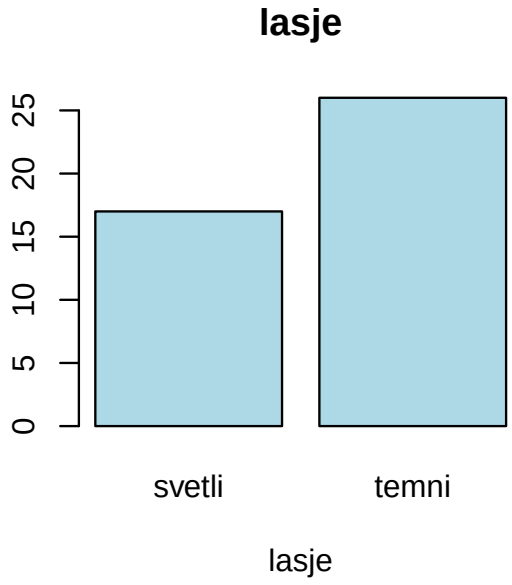
Opisna statistika

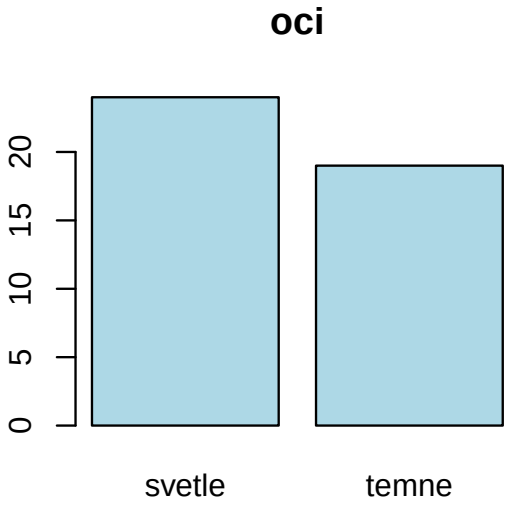
```
> summary(data[, 7:12])
```

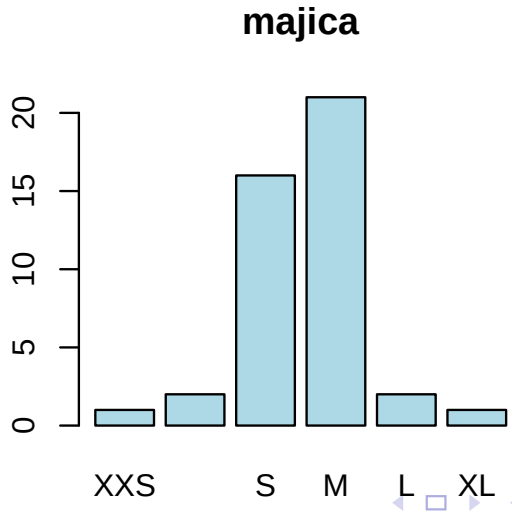
```
     cevelj          lasje          oci        majica
 Min.   :29.00   svetli:17   svetle:24   XXS: 1
 1st Qu.:38.00   temni :26   temne :19   XS : 2
 Median :39.00                           S  :16
 Mean   :39.37                           M  :21
 3rd Qu.:40.00                           L  : 2
 Max.   :46.00                           XL : 1
      mati            oce
 Min.   :150.0   Min.   :153.0
 1st Qu.:160.0   1st Qu.:175.0
 Median :165.0   Median :179.0
 Mean   :165.2   Mean   :178.0
 3rd Qu.:171.0   3rd Qu.:182.5
 Max.   :178.0   Max.   :190.0
```

Porazdelitve opisnih spremenljivk

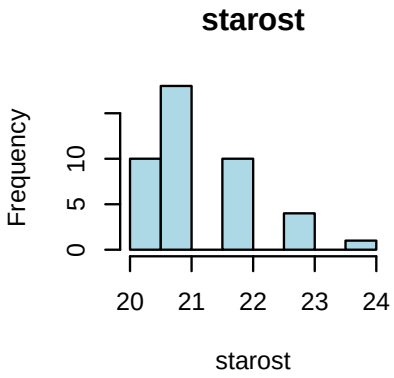


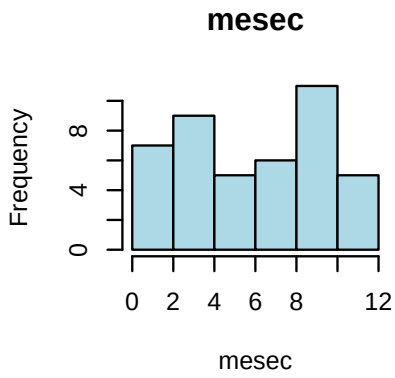


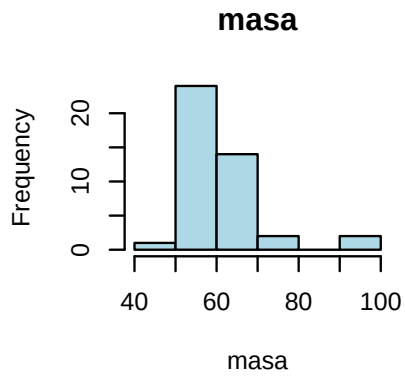


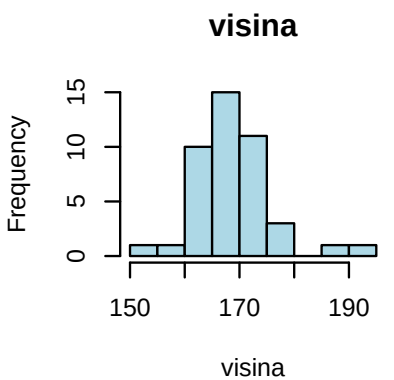


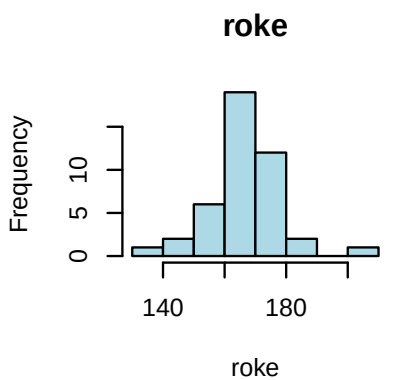
Porazdelitve številskih spremenljivk

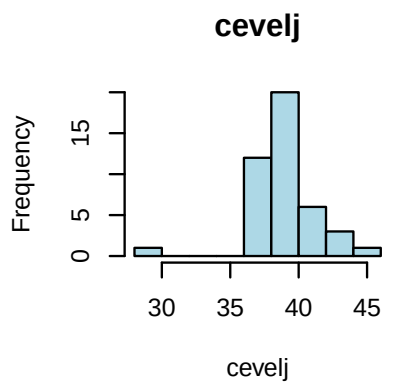


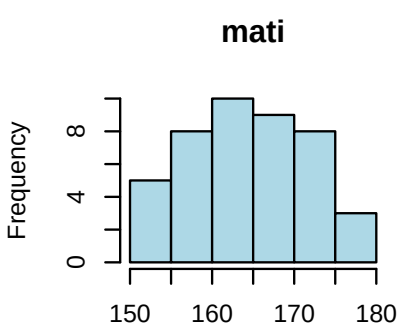


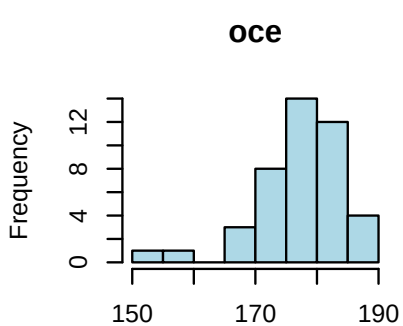




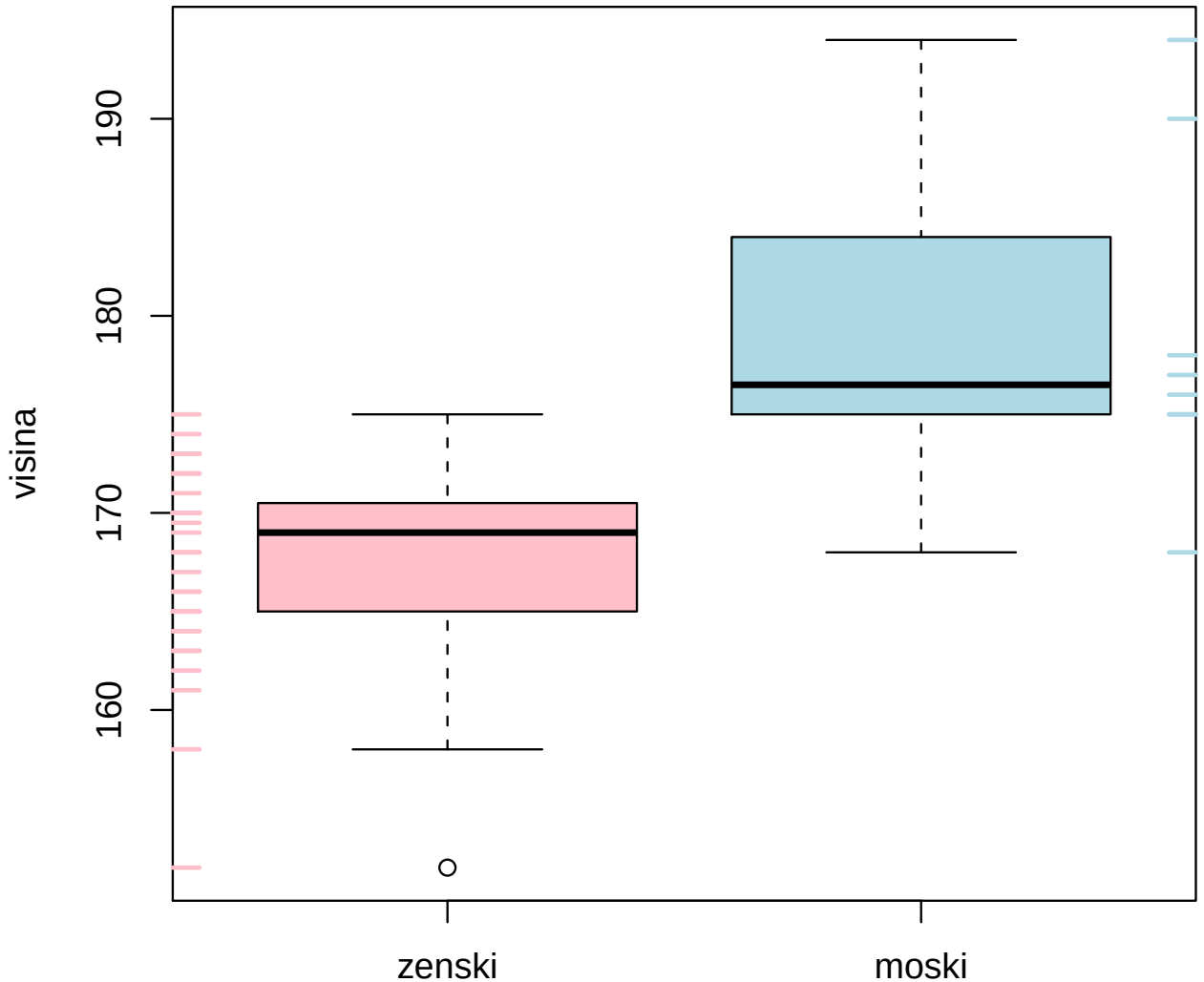








Porazdelitvi



Porazdelitvi - R ukazi

```
> boxplot(visina ~ spol, col = c("pink",
+     "lightblue"), ylab = "visina")
> rug(visina[spol == "zenski"], side = 2,
+     col = "pink", lwd = 2)
> rug(visina[!spol == "zenski"], side = 4,
+     , col = "lightblue", lwd = 2)
```

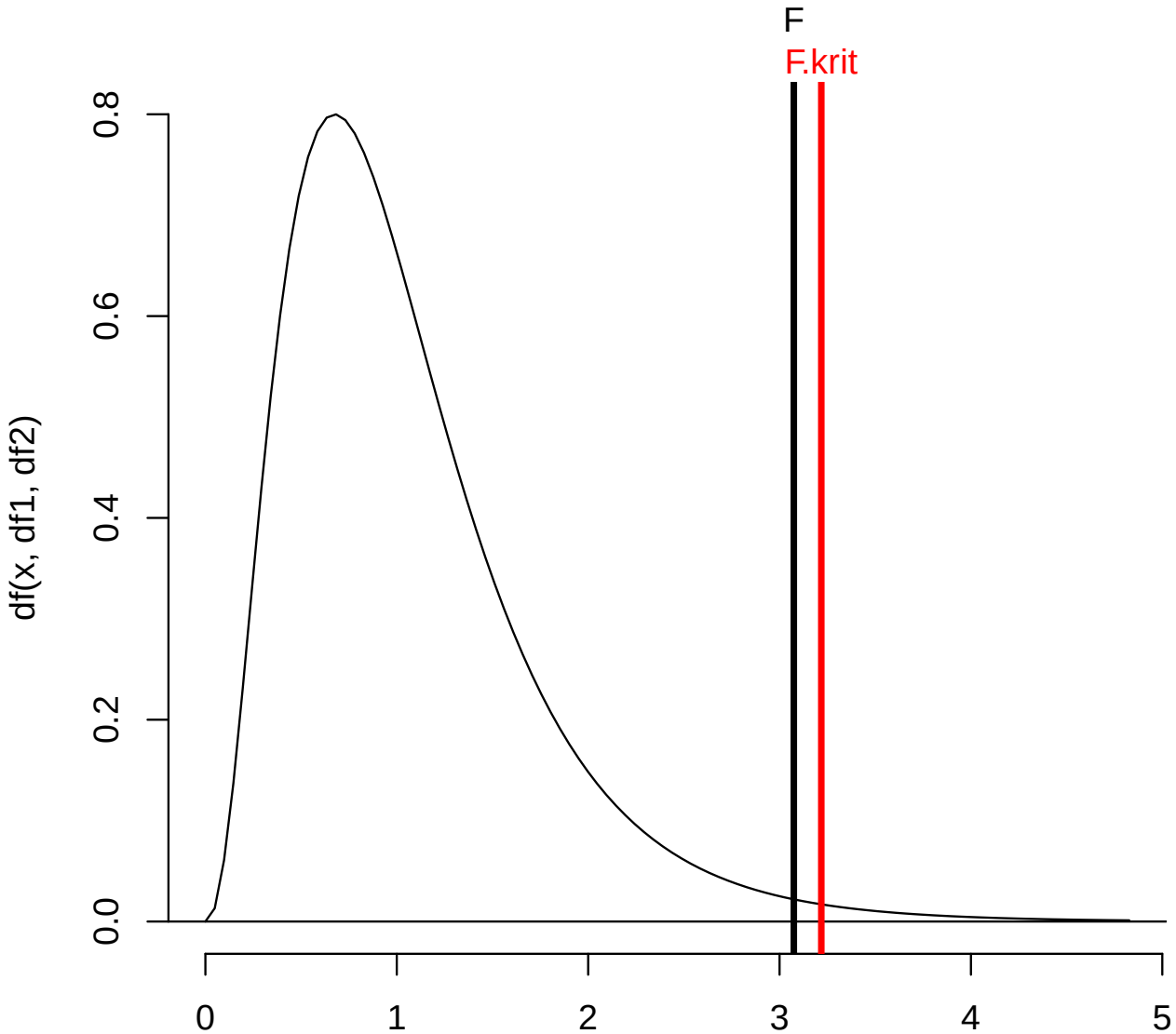
Opisni pregled

```
> y <- visina
> (n <- tapply(y, spol, length))

zenski  moski
    35      8

> (xbar <- tapply(y, spol, mean))

  zenski    moski
167.5857 179.1250

> (s <- tapply(y, spol, sd))

  zenski    moski
4.881667 8.559665
```

Ali sta varianci značilno različni?

```
> alpha <- 0.01
> (v <- sort(s^2))
  zenski    moski
23.83067 73.26786
> ns <- as.vector(n[order(s)])
> (F <- as.vector(v[2]/v[1]))
[1] 3.074519
> (df1 <- ns[2] - 1)
[1] 7
> (df2 <- ns[1] - 1)
[1] 34
> (F.krit <- qf(1 - alpha, df1, df2))
[1] 3.218154
> (p <- 1 - pf(F, df1, df2))
[1] 0.01278561
```

Ali sta varianci značilno različni?



Kako smo narisali sliko?

```
> x <- seq(0, max(F, F.krit) * 1.5, length = 100)
> plot(x, df(x, df1, df2), type = "l", xlab = "F",
+      axes = FALSE)
> axis(1)
> axis(2)
> abline(h = 0)
> abline(v = F, lwd = 3)
> mtext("F", side = 3, line = 1, at = F)
> abline(v = F.krit, col = "red", lwd = 3)
> mtext("F.krit", side = 3, line = 0, at = F.krit,
+      col = "red")
```

Hipoteze in delni rezultati

Uredimo vrstni red delnih rezultatov za test hipotez

$$H_0 : \mu_{moski} = \mu_{zenski} + \Delta$$

$$H_1 : \mu_{moski} > \mu_{zenski} + \Delta$$

```
> ord <- c("moski", "zenski")
> (xbar <- as.vector(xbar[ord]))
[1] 179.1250 167.5857
> (s <- as.vector(s[ord]))
[1] 8.559665 4.881667
> (n <- as.vector(n[ord]))
[1]  8 35
```

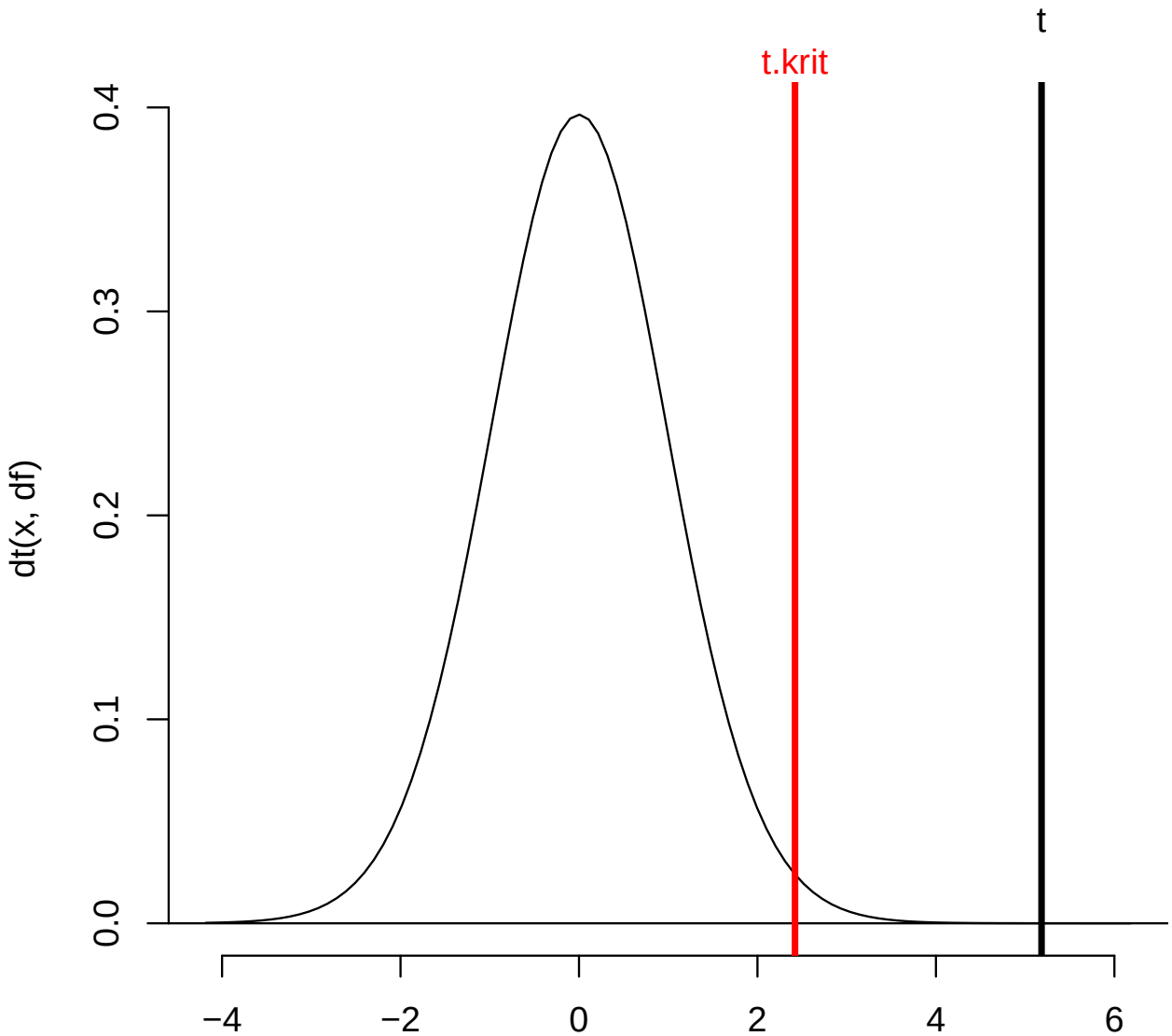
Stopnja tveganja in kritične vrednosti

```
> alpha <- 0.01
> delta <- 0
> (df <- n[1] + n[2] - 2)

[1] 41

> (t.krit <- qt(1 - alpha, df))

[1] 2.420803
```

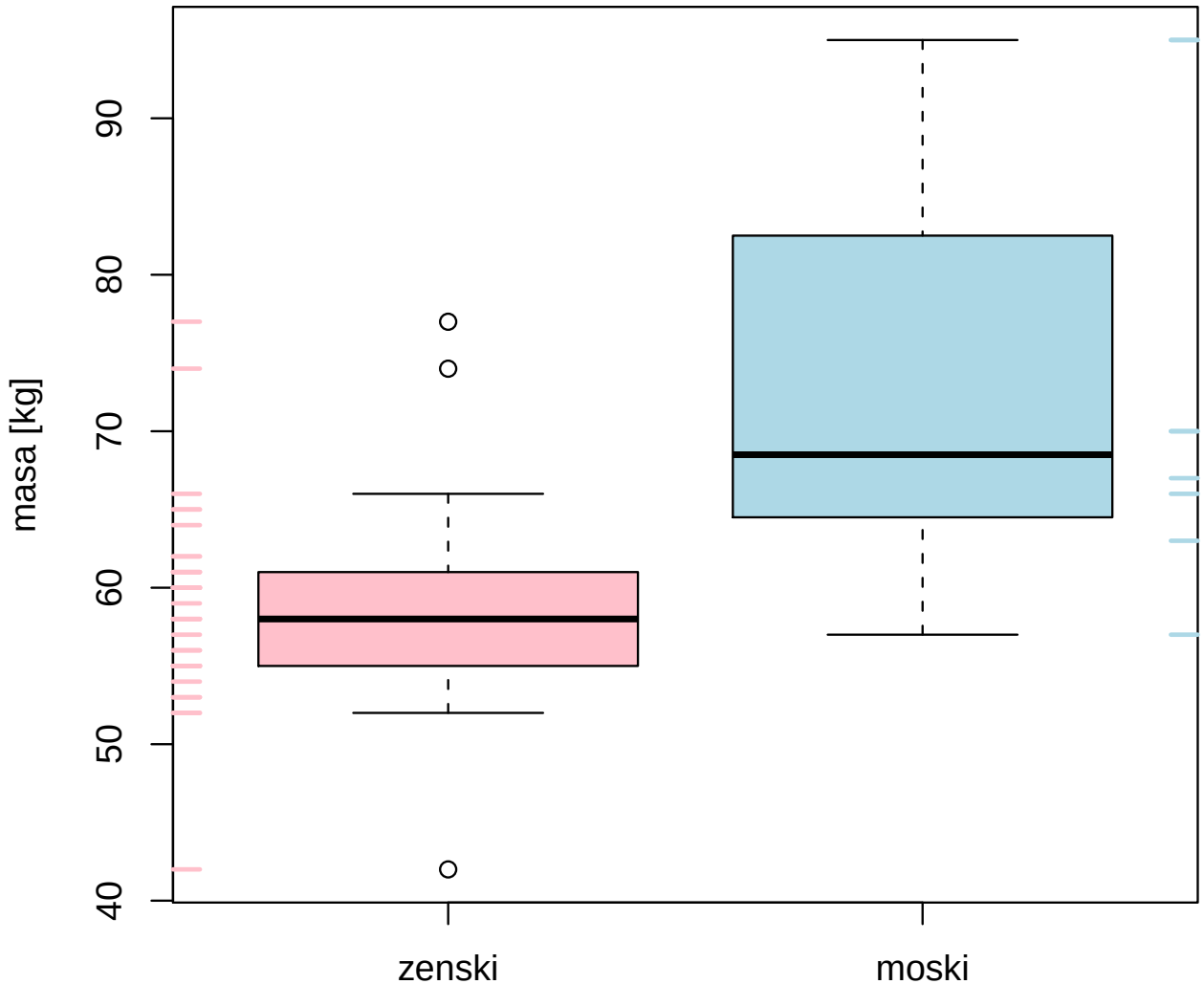
Studentov t-test

```
> xbar[1] - xbar[2]

[1] 11.53929

> s2 <- ((n[1] - 1) * s[1]^2 + (n[2] - 1) *
+     s[2]^2)/(n[1] + n[2] - 2)
> (t <- (xbar[1] - xbar[2] - delta)/sqrt(s2) *
+     sqrt(n[1] * n[2]/(n[1] + n[2])))

[1] 5.18342

> (p <- 1 - pt(t, df))

[1] 3.102809e-06

> if (t < t.krit) cat("Povprečje1 NI statistično značiln
+     round(p, 3), ").\n") else cat("Povprečje1 JE  stat
+     alpha, ") (p =", round(p, 3), ").\n")

Povprečje1 JE  statistično zančilno večje (p < 0.01 ) (p
```

Slika



Porazdelitvi



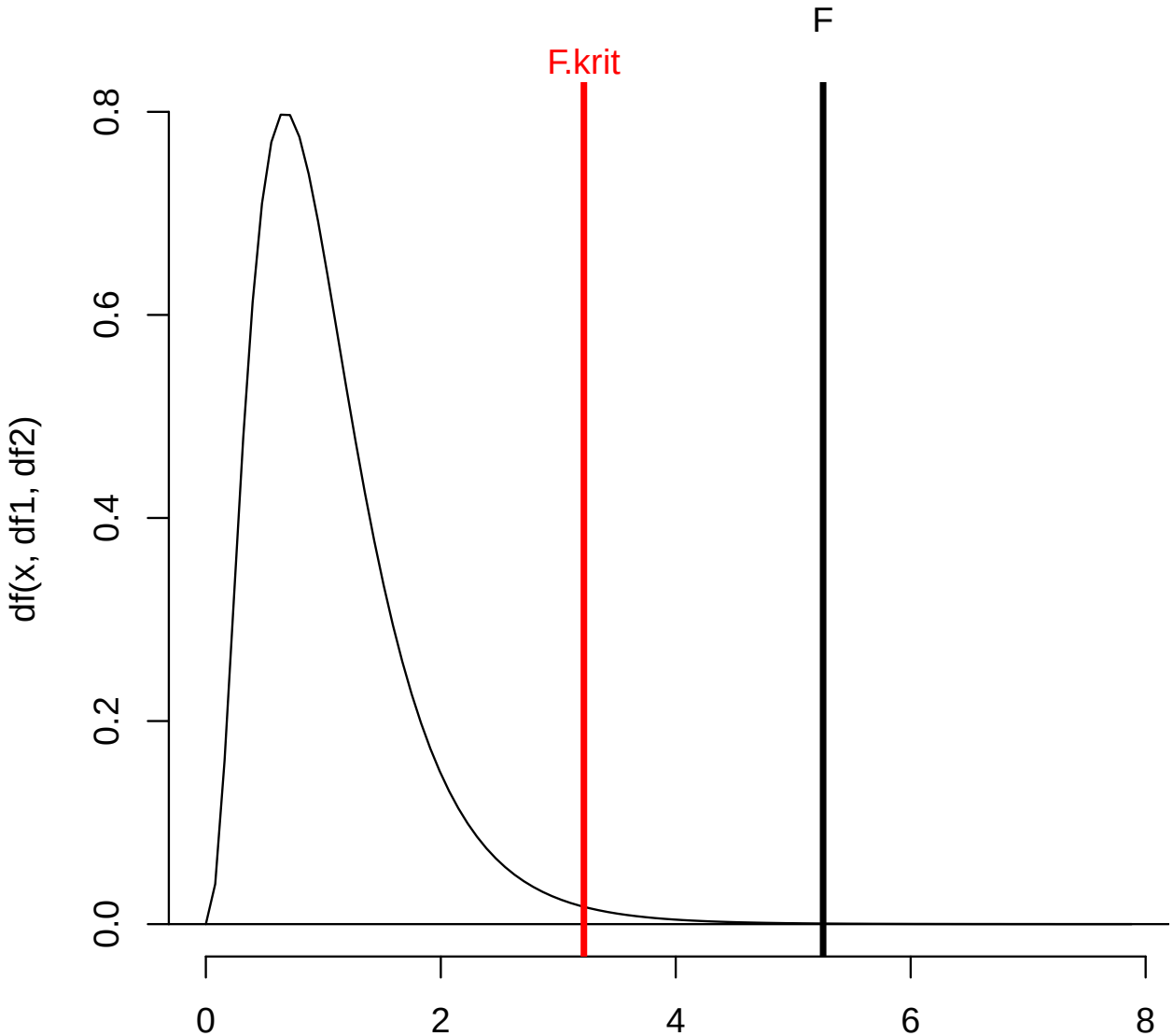
Opisni pregled

```
> (n <- tapply(y, spol, length))

zenski  moski
    35      8

> (xbar <- tapply(y, spol, mean))

  zenski    moski
58.82857 72.87500

> (s <- tapply(y, spol, sd))

   zenski     moski
 6.228425 14.277230
```

Ali sta varianci značilno različni?

```
> alpha <- 0.01
> (v <- sort(s^2))
   zenski     moski
 38.79328 203.83929
> ns <- as.vector(n[order(s)])
> (F <- as.vector(v[2]/v[1]))
[1] 5.2545
> (df1 <- ns[2] - 1)
[1] 7
> (df2 <- ns[1] - 1)
[1] 34
> (F.krit <- qf(1 - alpha, df1, df2))
[1] 3.218154
> (p <- 1 - pf(F, df1, df2))
[1] 0.0003909851
```

Ali sta varianci značilno različni?



Kako smo narisali sliko?

```
> x <- seq(0, max(F, F.krit) * 1.5, length = 100)
> plot(x, df(x, df1, df2), type = "l", xlab = "F",
+      axes = FALSE)
> axis(1)
> axis(2)
> abline(h = 0)
> abline(v = F, lwd = 3)
> mtext("F", side = 3, line = 1, at = F)
> abline(v = F.krit, col = "red", lwd = 3)
> mtext("F.krit", side = 3, line = 0, at = F.krit,
+      col = "red")
```

Hipoteze in delni rezultati

Uredimo vrstni red delnih rezultatov za test hipotez

$$H_0 : \mu_{moski} = \mu_{zenski} + \delta$$

$$H_1 : \mu_{moski} > \mu_{zenski} + \delta$$

```
> ord <- c("moski", "zenski")
> (xbar <- as.vector(xbar[ord]))
[1] 72.87500 58.82857
> (s <- as.vector(s[ord]))
[1] 14.277230  6.228425
> (n <- as.vector(n[ord]))
[1]  8 35
```

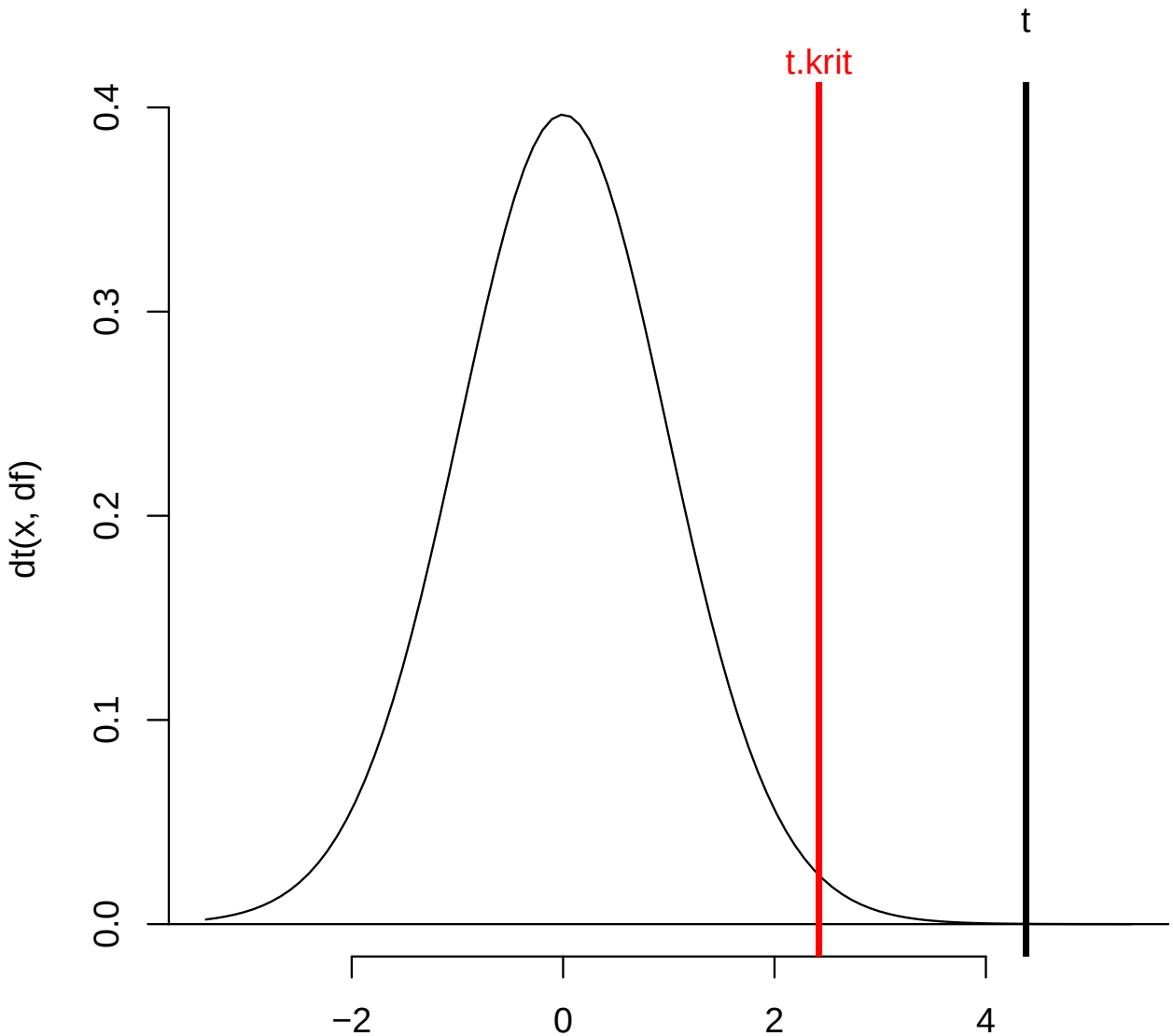
Stopnja tveganja in kritične vrednosti

```
> alpha <- 0.01
> delta <- 0
> (df <- n[1] + n[2] - 2)

[1] 41

> (t.krit <- qt(1 - alpha, df))

[1] 2.420803
```

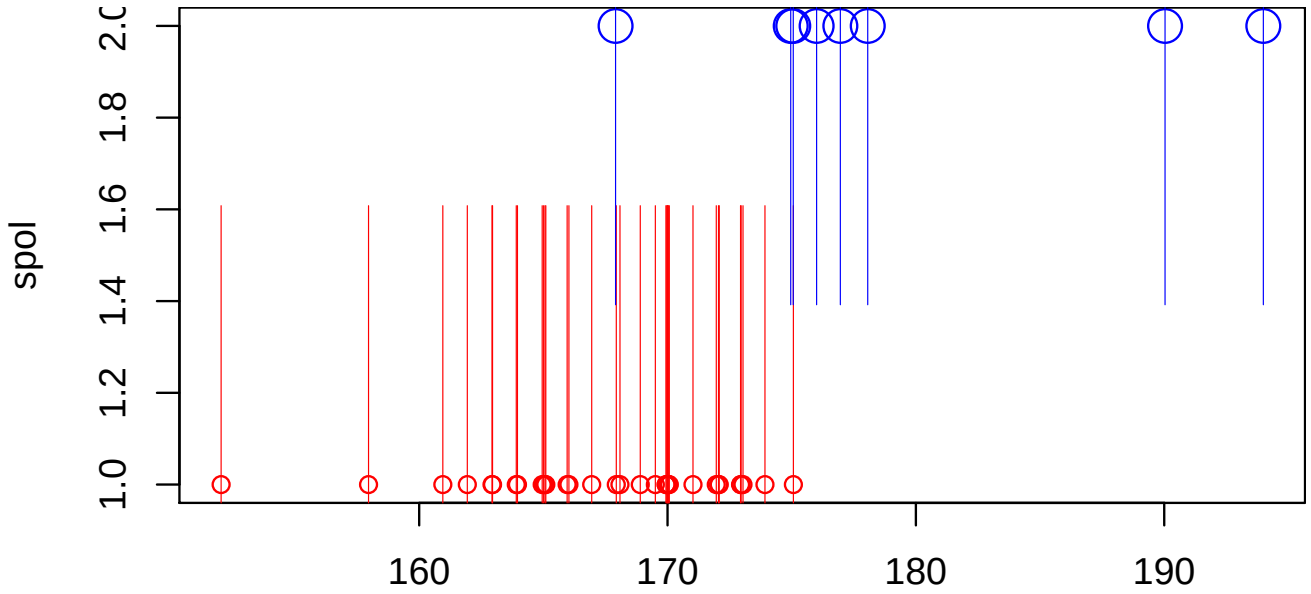
Studentov t-test

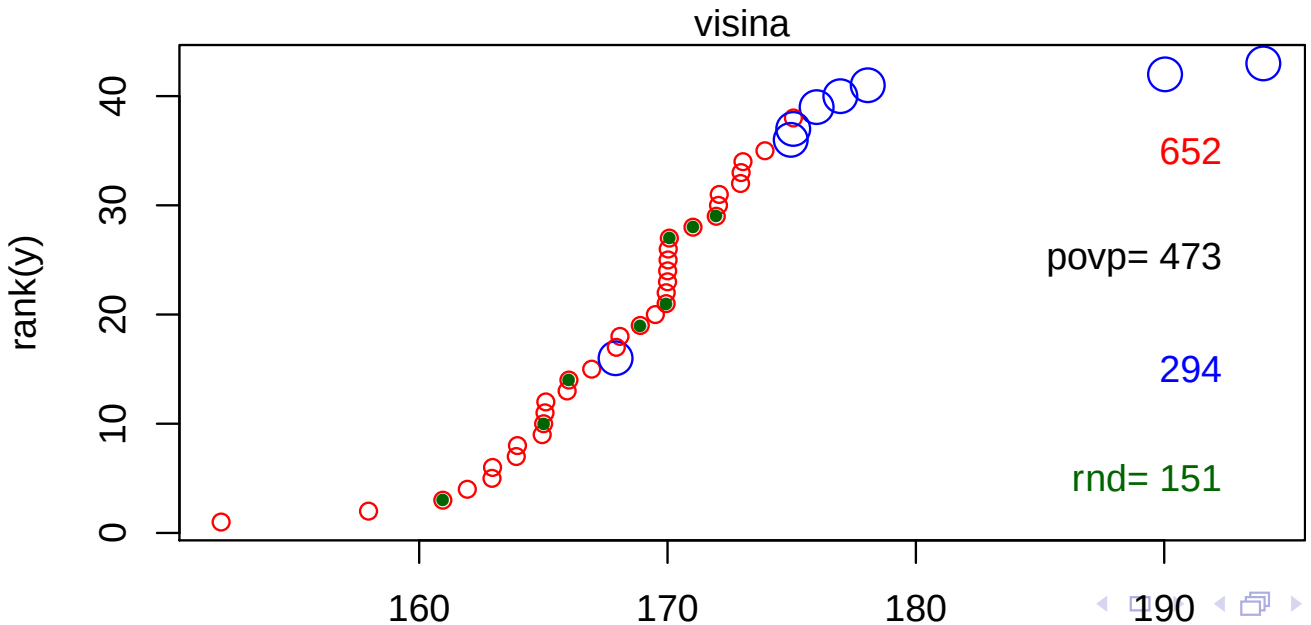
```
> xbar[1] - xbar[2]

[1] 14.04643

> s2 <- ((n[1] - 1) * s[1]^2 + (n[2] - 1) *
+     s[2]^2)/(n[1] + n[2] - 2)
> (t <- (xbar[1] - xbar[2] - delta)/sqrt(s2) *
+     sqrt(n[1] * n[2]/(n[1] + n[2])))

[1] 4.379903

> (p <- 1 - pt(t, df))

[1] 4.012688e-05

> if (t < t.krit) cat("Povprečje1 NI statistično značiln
+     round(p, 3), ").\n") else cat("Povprečje1 JE  stat
+     alpha, ") (p =", round(p, 3), ").\n")

Povprečje1 JE  statistično zančilno večje (p < 0.01 ) (p
```

Slika

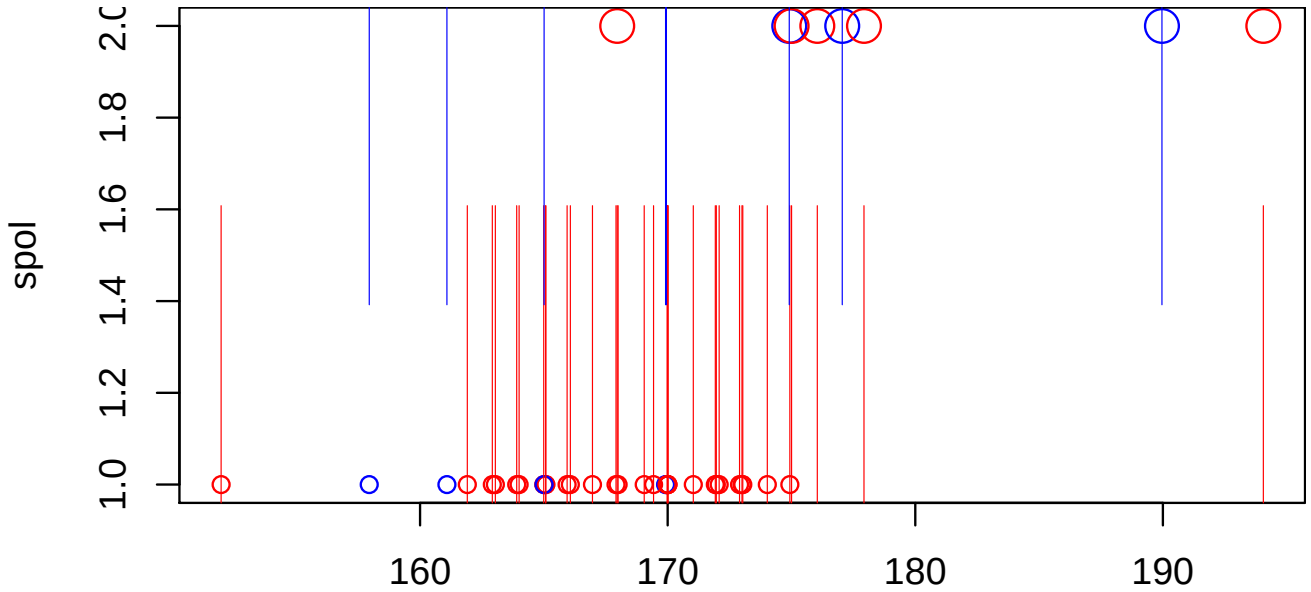


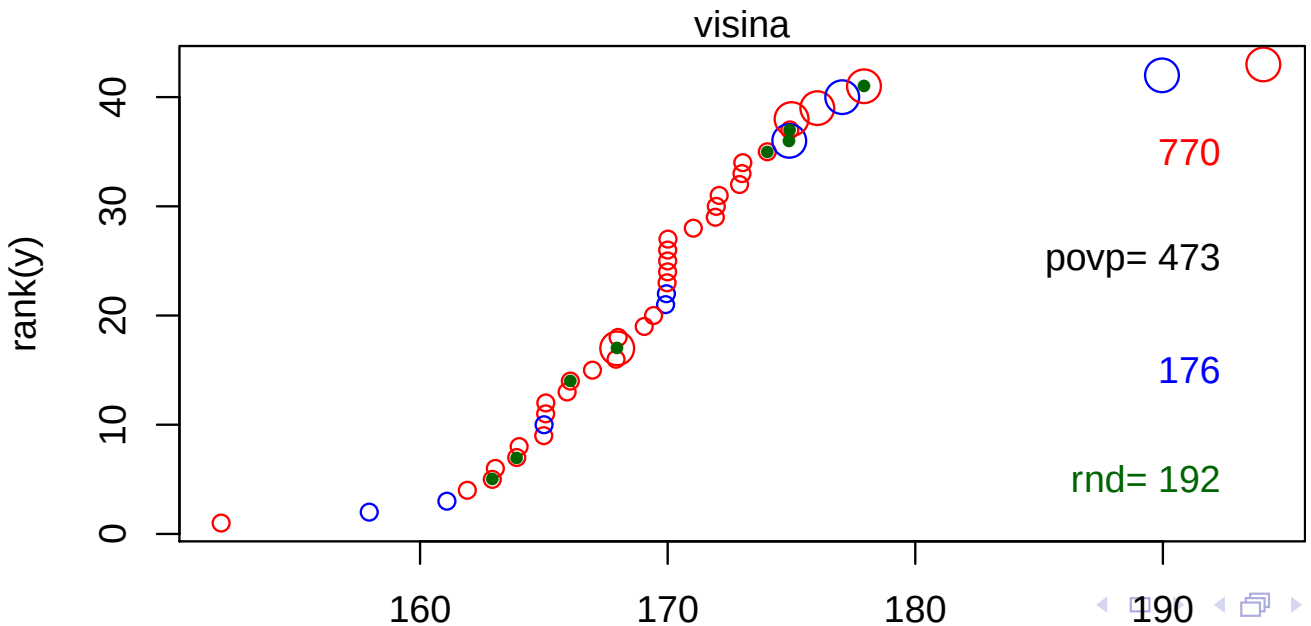
Neparametrični test - Wilcoxon





Randomizacijski test - vsota rangov





Wilcoxon test v R

```
> wilcox.test(jitter(visina) ~ spol)

Wilcoxon rank sum test

data:  jitter(visina) by spol
W = 19, p-value = 2.296e-05
alternative hypothesis: true location shift is not equal
```

Barva las in oči

```
> f <- table(lasje, oci)
> f
        oci
lasje    svetle temnе
  svetli      14     3
  temni       10    16

> addmargins(f)
        oci
lasje    svetle temnе Sum
  svetli      14     3  17
  temni       10    16  26
  Sum         24    19  43
```

Pričakovane frekvence

```
> e <- outer(rowSums(f), colSums(f))/sum(f)
> addmargins(e)
```

```
          svetle     temne Sum
svetli  9.488372  7.511628  17
temni  14.511628 11.488372  26
Sum    24.000000 19.000000  43
```

Razlika opaženega in pričakovanega

```
> f - e

         oci
lasje         svetle      temne
  svetli   4.511628 -4.511628
  temni   -4.511628  4.511628

> (f - e)^2/e

         oci
lasje        svetle    temne
  svetli 2.145235 2.709770
  temni  1.402654 1.771773

> 1.96^2

[1] 3.8416
```

Test χ^2

```
> alpha <- 0.05
> (df <- (ncol(f) - 1) * (nrow(f) - 1))
[1] 1
> (X2.krit <- qchisq(1 - alpha, df))
[1] 3.841459
> (X2 <- sum((f - e)^2/e))
[1] 8.029432
> (p <- 1 - pchisq(X2, df))
[1] 0.004602328
> if (X2 < X2.krit) cat("Spremenljivki NISTA odvisni (p
+     round(p, 4), ").\n") else cat("Spremenljivki STA o
+     alpha, ") (p =", round(p, 4), ").\n")
Spremenljivki STA odvisni (p < 0.05 ) (p = 0.0046 ).
```

Funkcija za test v R

```
> chisq.test(lasje, oci, correct = FALSE)

Pearson's Chi-squared test

data:  lasje and oci
X-squared = 8.0294, df = 1, p-value =
0.004602
```

Barva las in oči

	oci		
lasje	svetle	temne	Sum
svetli	14	3	17
temni	10	16	26
Sum	24	19	43

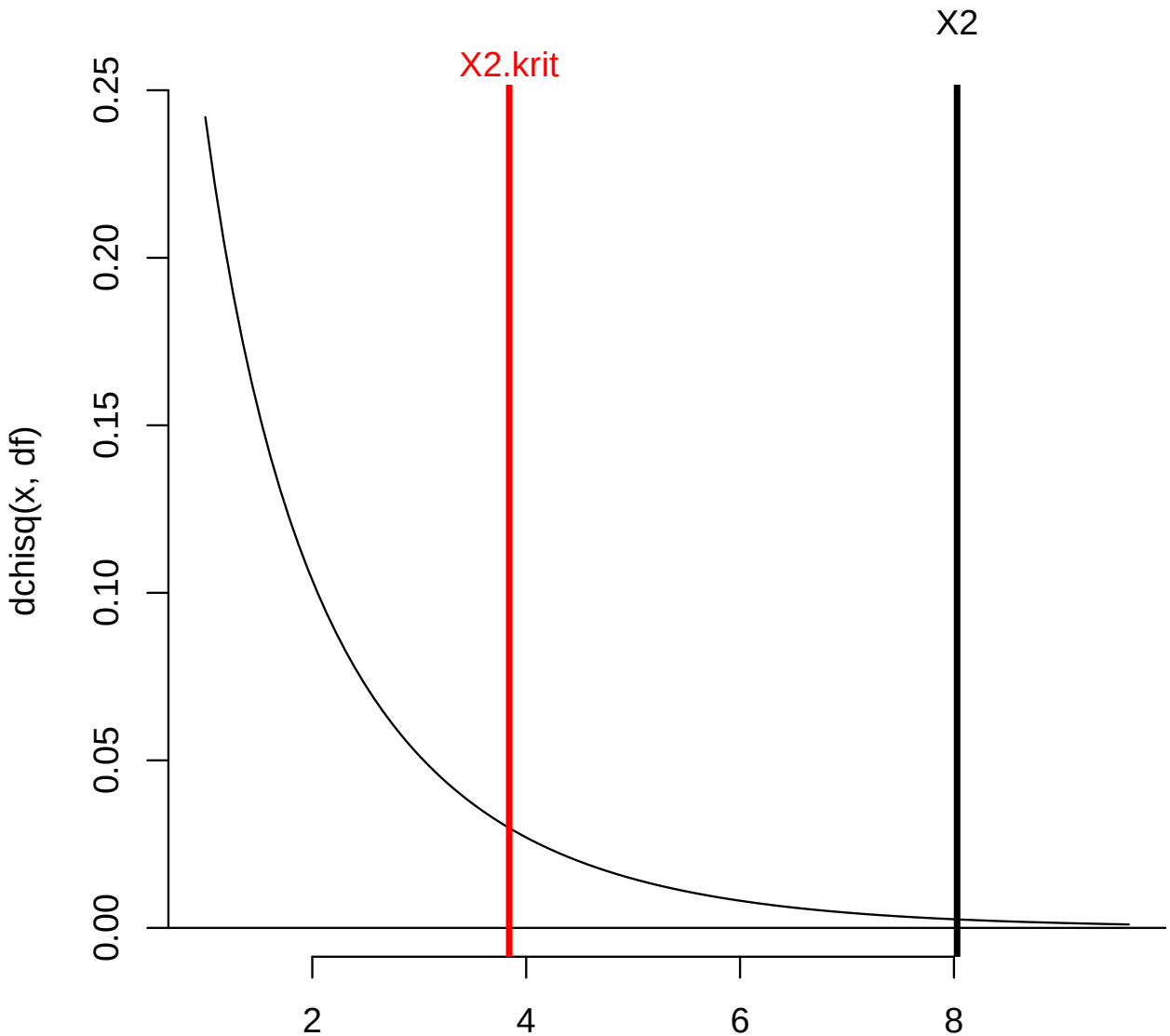
Asociacija (2×2)

$$\frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

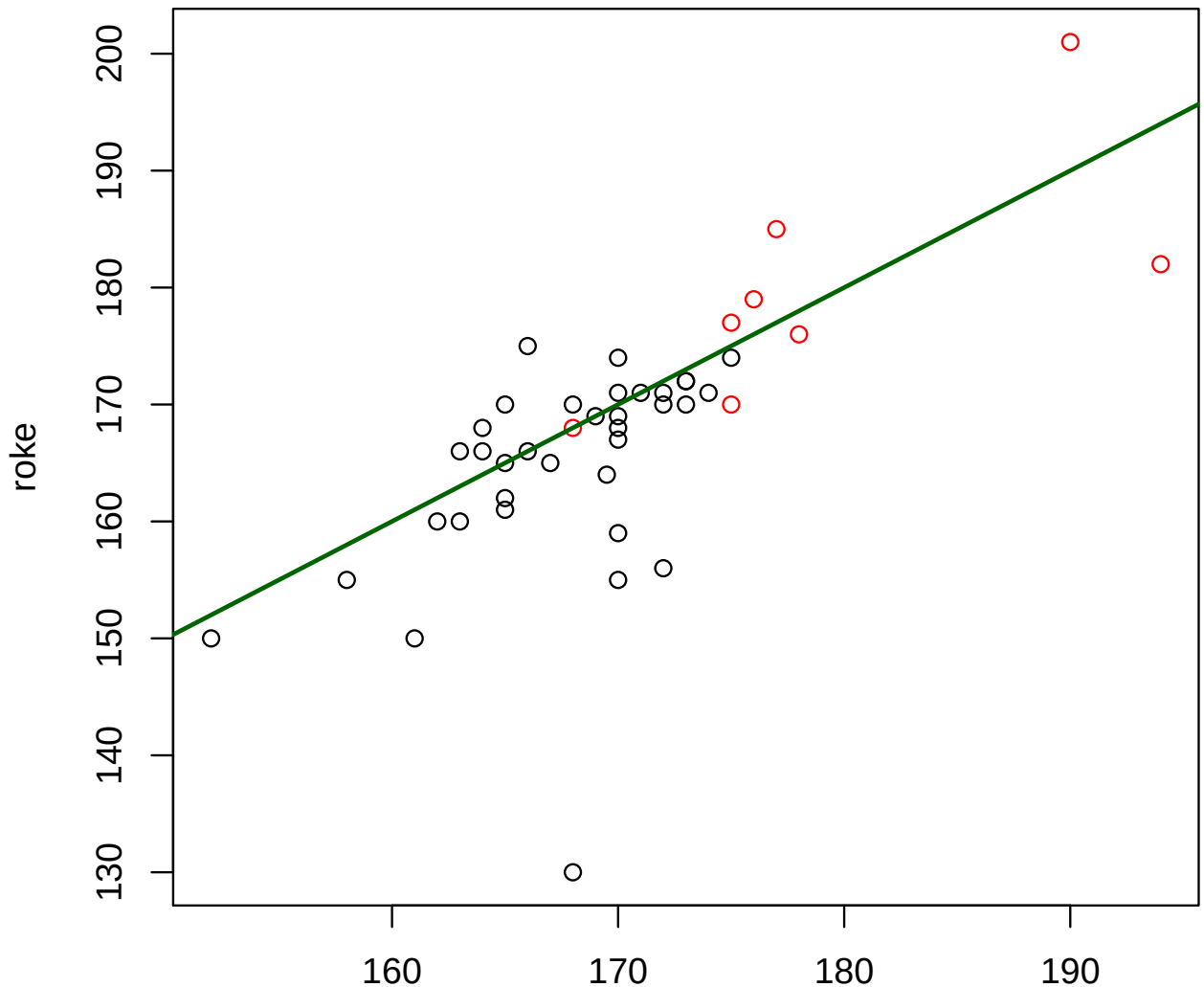
```
> sum(f) * det(matrix(f, 2, 2))^2/prod(colSums(f),
+      rowSums(f))
```

```
[1] 8.029432
```

Slika



Virtruvian man: Roke = Višina



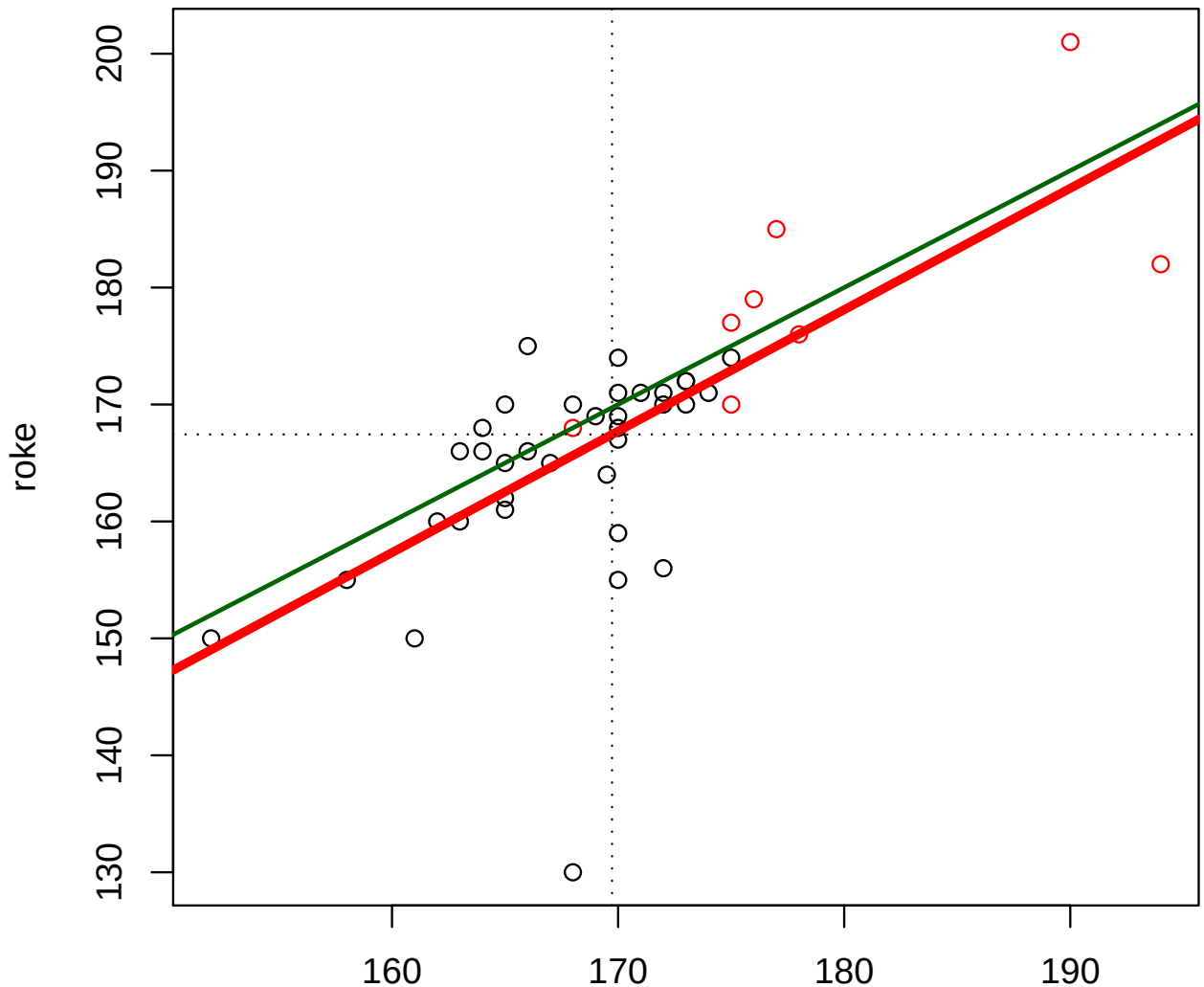
Spremenljivke

```
> select <- roke > 0
> x <- visina[select]
> y <- roke[select]
> spol <- spol[select]
> n <- length(x)
```

Osnovni račun

```
> (xbar <- mean(x))
[1] 169.7326
> (ybar <- mean(y))
[1] 167.4419
> sum((x - xbar) * (y - ybar))/(n - 1)
[1] 54.1567
> cov(x, y)
[1] 54.1567
> (r <- cov(x, y)/(sd(x) * sd(y)))
[1] 0.6906298
> r^2
[1] 0.4769695
> (b <- cov(x, y)/var(x))
[1] 1.038539
> (a <- mean(y) - b * mean(x))
[1] -8.832009
```

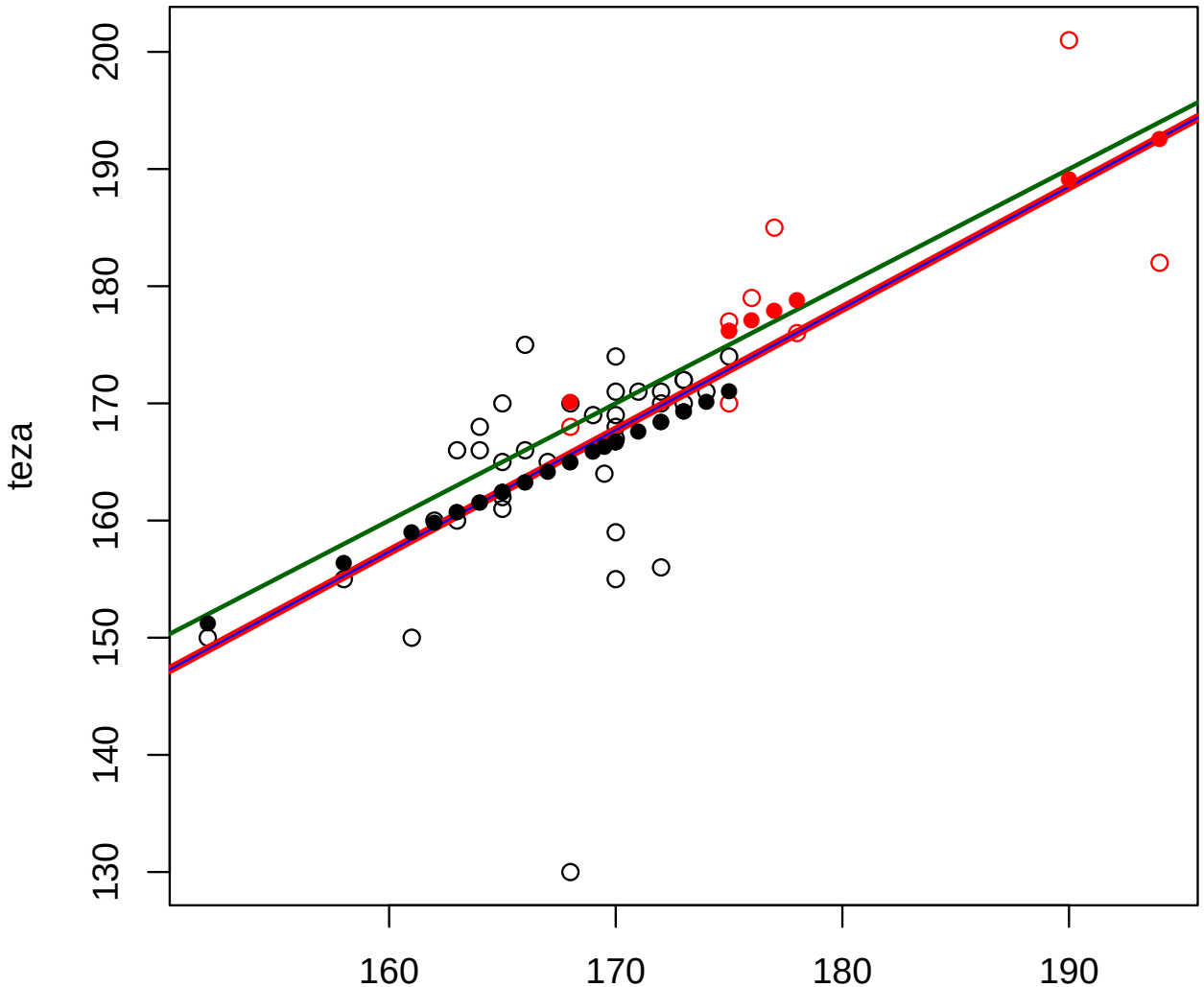
Virtruvian man: Roke = Višina



Regresija v R

```
> cor(x, y)
[1] 0.6906298
> lsfit(x, y)$coefficients
Intercept         X
-8.832009  1.038539
> lm(y ~ x)$coefficients
(Intercept)           x
  -8.832009    1.038539
> lm(y ~ x * spol)

Call:
lm(formula = y ~ x * spol)

Coefficients:
(Intercept)           x    spolmoski
  20.648518    0.859143     4.642732
x:spolmoski
   0.003153
```

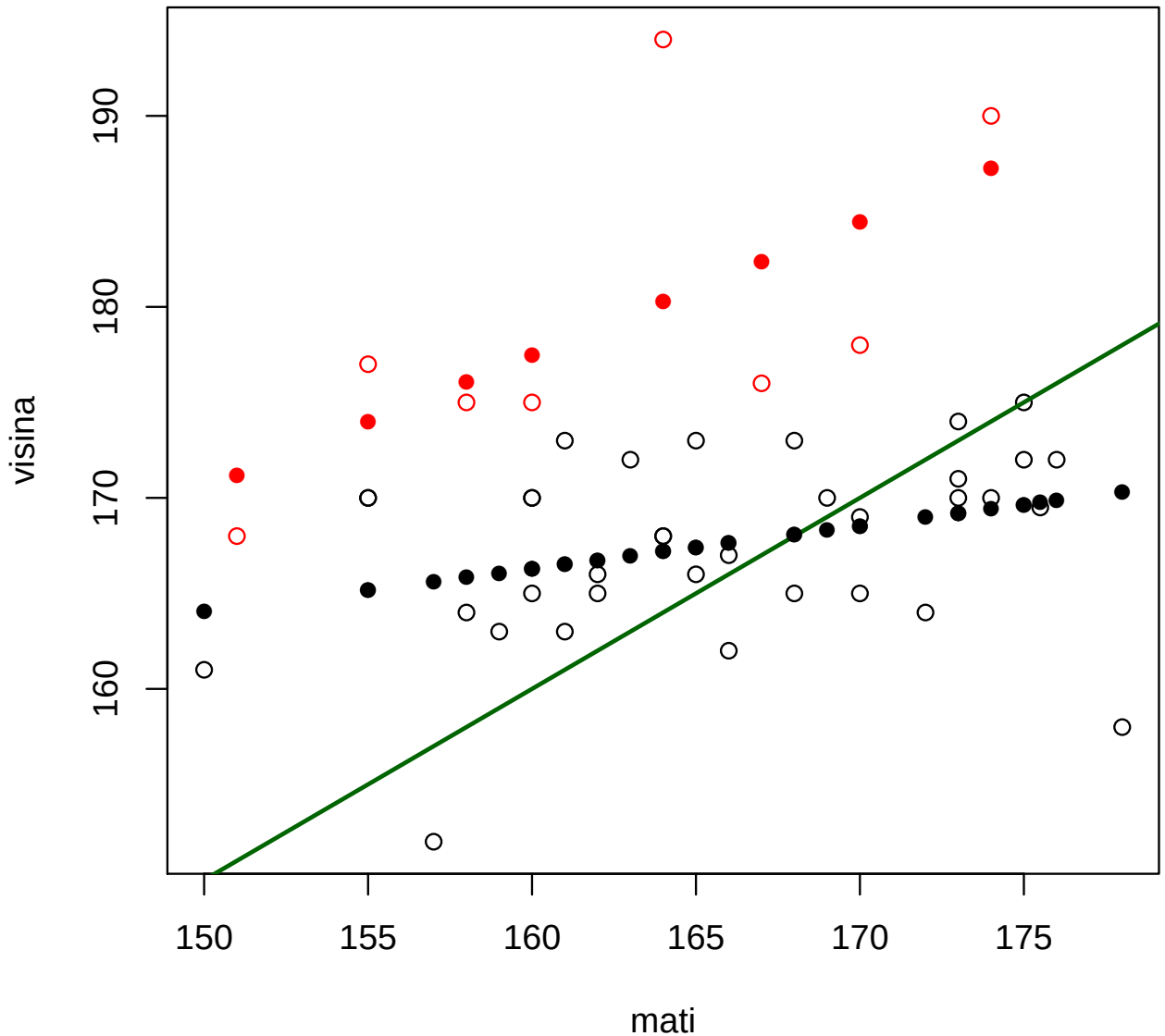
Virtruvian man: Roke = Višina



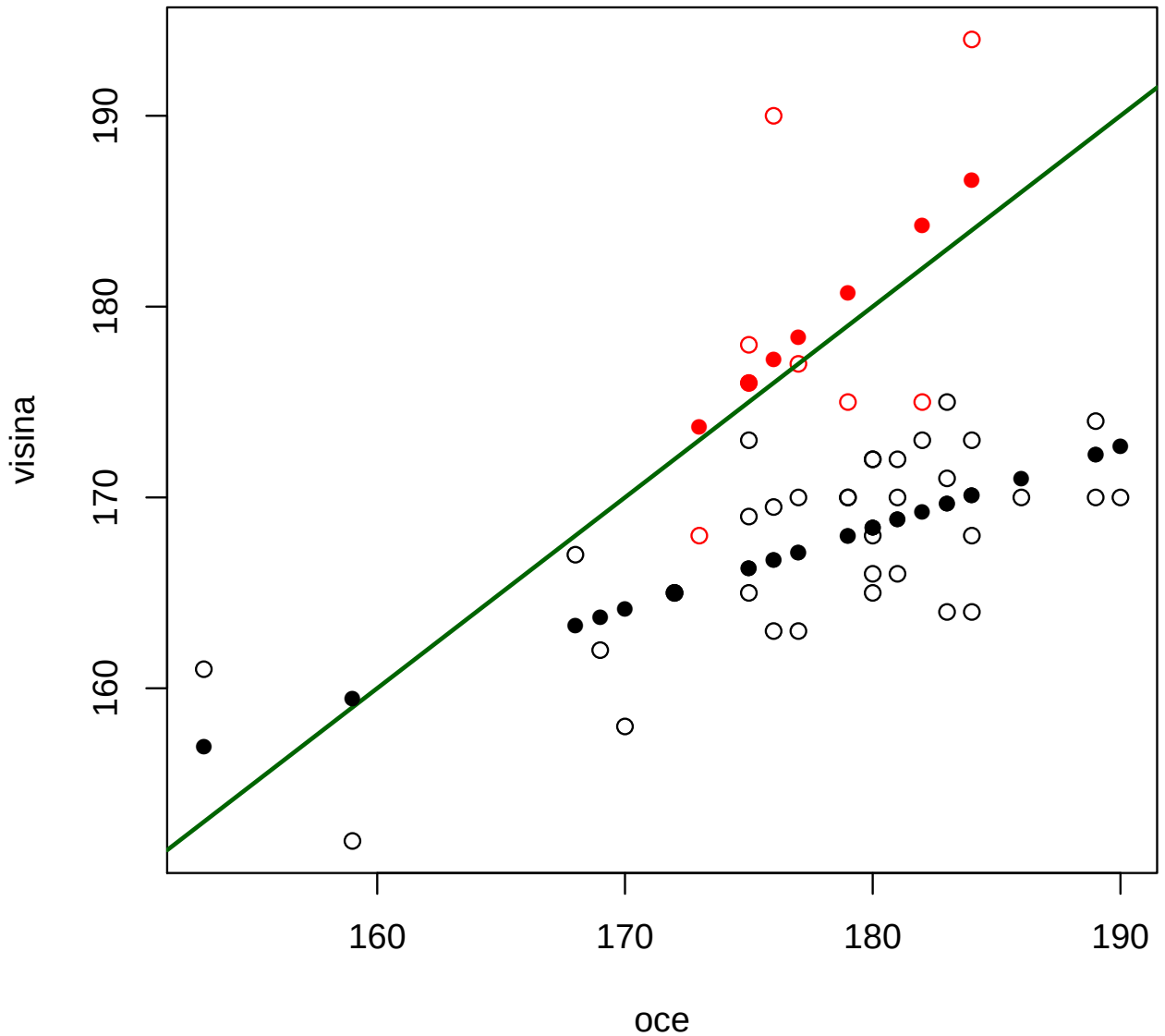
Analiza variance za linearni model

```
> anova(lm(roke ~ visina * spol))
Analysis of Variance Table

Response: roke
             Df Sum Sq Mean Sq   F value    Pr(>F)
visina        1 2362.2  2362.2   37.0845 3.887e-07
spol          1  106.1   106.1    1.6657    0.2044
visina:spol   1    0.0     0.0 4.902e-05    0.9944
Residuals    39 2484.3    63.7

visina       ***
spol
visina:spol
Residuals
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Primerjava z višino matere



Primerjava z višino očetov



Primerjava višin očetov in mam