

Pierwsza praca domowa

Joanna Kęczkowska

30.03.2021

Zbiór laptops.csv zawiera następujące zmienne:

- inches – rozmiar przekątnej w calach
- weight – waga laptopa
- price_euros – cena laptopa w euro
- company – producent laptopa (1 – Acer, 2 – Asus, 3 – Dell, 4 – HP, 5 – Lenovo, 6 – MSI, 7 – Toshiba)
- typename – typ laptopa (1 – 2w1, 2 – gaming, 3 – netbook, 4 – notebook, 5 – ultrabook, 6 – stacja robocza)
- ram – ilość RAM laptopa (1 – 4GB, 2 – 8GB, 3 – 16GB, 4 – 32GB)

```
dataSet <- read.csv(file = "laptops.csv", sep = ";", header = TRUE)
str(dataSet)
```

```
## 'data.frame': 1142 obs. of 6 variables:
## $ inches : num 15.6 15.6 14 14 15.6 ...
## $ weight : num 1.86 2.1 1.3 1.6 1.86 ...
## $ price_euros: num 575 400 1495 770 394 ...
## $ company : int 4 1 2 1 4 4 3 3 5 3 ...
## $ typename : int 4 4 5 5 4 4 4 4 5 ...
## $ ram : int 2 1 3 2 1 1 1 2 2 2 ...
```

```
summary(dataSet)
```

```
## inches weight price_euros company
## Min. :10.10 Min. :0.690 Min. : 209.0 Min. :1.00
## 1st Qu.:14.00 1st Qu.:1.600 1st Qu.: 619.6 1st Qu.:3.00
## Median :15.60 Median :2.060 Median : 986.5 Median :4.00
## Mean :15.08 Mean :2.069 Mean :1128.9 Mean :3.71
## 3rd Qu.:15.60 3rd Qu.:2.330 3rd Qu.:1485.8 3rd Qu.:5.00
## Max. :18.40 Max. :4.700 Max. :4899.0 Max. :7.00
## typename ram
## Min. :1.000 Min. :1.000
## 1st Qu.:2.000 1st Qu.:1.000
## Median :4.000 Median :2.000
## Mean :3.539 Mean :1.874
## 3rd Qu.:4.000 3rd Qu.:2.000
## Max. :6.000 Max. :4.000
```

Należy zweryfikować następujące hipotezy:

a) Stosowana ilość RAM w laptopie jest zależna od jego producenta.

Chi-square test sprawdza zależność między zmiennymi.

dla danej komórki wartość oczekiwana: $e = \frac{\text{row.sum} * \text{col.sum}}{\text{grand.total}}$

Chi-square statistic: $\chi^2 = \sum \frac{(o-e)^2}{e}$, gdzie o - obserwacja, e - wartość oczekiwana

Hipoteza zerowa H_0 : Stosowana ilość RAM w laptopie jest **niezależna** od jego producenta.

Hipoteza alternatywna H_1 : Stosowana ilość RAM w laptopie jest **zależna** od jego producenta.

Założenie - poziomy (kategorie) dla zmiennych są rozłączne/wzajemnie się wykluczają - jest spełnione.

```
alpha <- 0.05 #5% level of significance
```

```
memory <- dataSet$ram
company <- dataSet$company
TAB <- table(company, memory)
TAB
```

```
##          memory
## company    1    2    3    4
##          1  57  33    4    0
##          2  45  61   35    3
##          3  63 161   54    7
##          4  90 142   13    0
##          5  89 141   39    3
##          6   0  22   31    1
##          7  14  25    8    1
```

```
alpha <- 0.1
```

```
total <- sum(TAB)
sumRows <- margin.table(TAB, 1) #rows
sumCols <- margin.table(TAB, 2) #columns
```

```
sumRows <- as.vector(sumRows)
sumCols <- as.vector(sumCols)
```

```
#expected observations
```

```
exp <- matrix(rep(0, 4*7), nrow=7, ncol=4)
exp[] <- 0L
for(i in 1:7) {
  exp[i, ] <- sumRows[i]*sumCols/total
}
```

```
Tab <- data.frame(TAB)
obs <- matrix(Tab[["Freq"]], nrow = 7, ncol = 4)
```

```
chi_sq <- sum((obs-exp)^2/exp) #test statistic
df <- (nrow(obs)-1)*(ncol(obs)-1) #deg of freedom
pval <- pchisq(chi_sq, df, lower.tail=FALSE) #right-tailed
```

```
quantile <- qchisq(alpha, df, lower.tail = FALSE) #quantile of chi-square distribution
```

```
if(alpha > pval) {
```

```

    print("H0 rejected.")
  }else {
    print("There is not enough evidence to suggest an association between RAM and company")
  }

```

```
## [1] "H0 rejected."
```

```
sprintf("test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", chi_sq, pval, quant
```

```
## [1] "test statistic = 164.234074 , p-value = 0.000000, confidece interval = [-infinity, 25.989423]"
```

```

#chisq.test()
test <- chisq.test(TAB)

```

```
## Warning in chisq.test(TAB): Chi-squared approximation may be incorrect
```

```
test
```

```

##
## Pearson's Chi-squared test
##
## data:  TAB
## X-squared = 164.23, df = 18, p-value < 2.2e-16

```

```

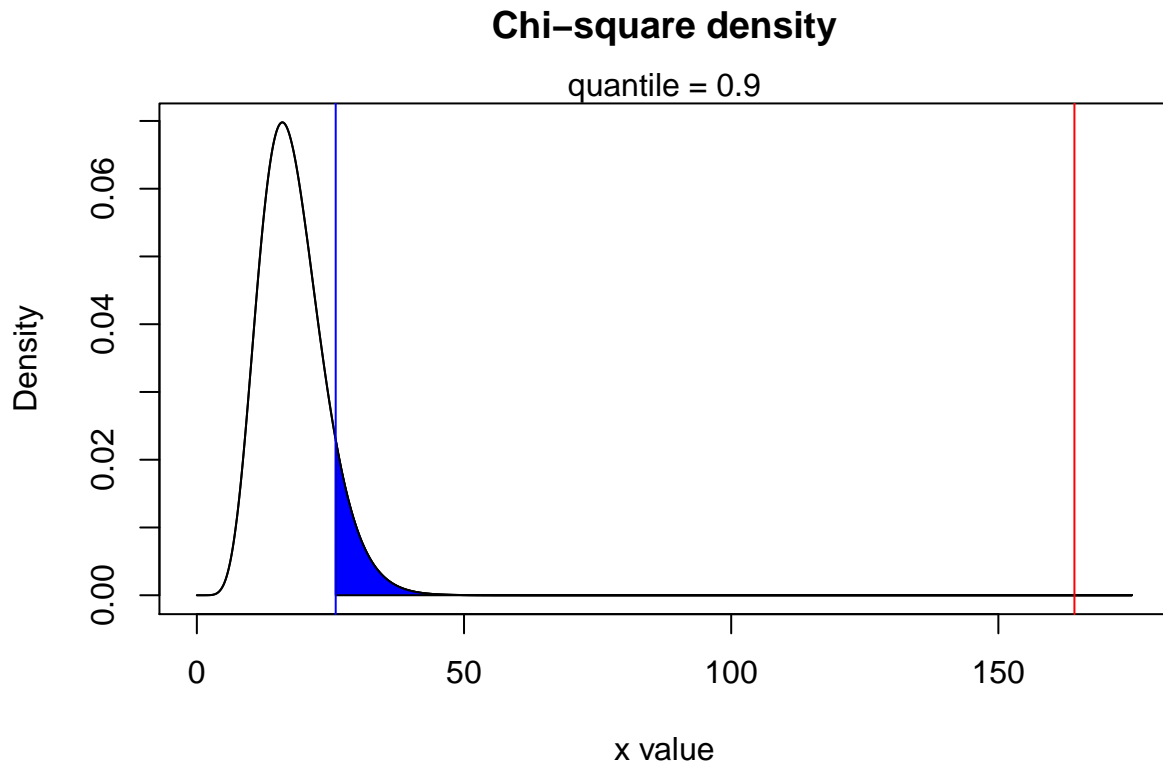
x <- seq(0, 175, by = 0.1)
chi_dense <- dchisq(x, df)

plot(x, chi_dense,type='l', xlab="x value",
     ylab="Density", main="Chi-square density")

i <- x >= quantile
lines(x, chi_dense)
polygon(c(quantile,x[i],175), c(0,chi_dense[i],0), col="blue")

area <- pchisq(quantile, df, lower.tail = TRUE)
result <- paste("quantile =",
               signif(area, digits=3))
mtext(result,3)
abline(v=chi_sq, col="red")
abline(v=quantile, col="blue")

```



Wychodzi, że **ilość pamięci RAM zależy od producenta** (wartość statystyki testowej wpada do obszaru krytycznego oraz p-value jest mniejsze niż nasz ustalony poziom istotności).

b) Rozkład stosowanych pamięci RAM w notebookach HP i Lenovo jest taki sam.

Test jak wyżej.

Hipoteza zerowa H_0 : Różne rozkłady.

Hipoteza alternatywna H_1 : Takie same rozkłady.

```
alpha <- 0.05
dataSet <- dataSet[dataSet["typename"]==4, ]

memory <- dataSet$ram
company <- dataSet$company
TAB <- table(company, memory)
TAB <- TAB[4:5, 1:4]

total <- sum(TAB)
sumRows <- margin.table(TAB, 1) #rows
sumCols <- margin.table(TAB, 2) #columns

sumRows <- as.vector(sumRows)
sumCols <- as.vector(sumCols)

#expected observations
```

```

exp <- matrix(rep(0, 4*7), nrow=2, ncol=4)
exp[] <- 0L
for(i in 1:2) {
  exp[i, ] <- sumRows[i]*sumCols/total
}

Tab <- data.frame(TAB)
obs <- matrix(Tab[["Freq"]], nrow = 2, ncol = 4)

chi_sq <- sum((obs-exp)^2/exp) #test statistic
df <- (nrow(obs)-1)*(ncol(obs)-1) #deg of freedom
pval <- pchisq(chi_sq, df, lower.tail=FALSE) #right-tailed

quantile <- qchisq(alpha, df, lower.tail = FALSE) #quantile of chi-square distribution

if(alpha > pval) {
  print("H0 rejected.")
}else {
  print("There is not enough evidence to suggest an association between RAM and company")
}

```

```
## [1] "There is not enough evidence to suggest an association between RAM and company"
```

```
sprintf("test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", chi_sq, pval, quantile)
```

```
## [1] "test statistic = 1.538174 , p-value = 0.673488, confidece interval = [-infinity, 7.814728]"
```

```

#chisq.test()
test <- chisq.test(TAB)

```

```
## Warning in chisq.test(TAB): Chi-squared approximation may be incorrect
```

```
test
```

```

##
## Pearson's Chi-squared test
##
## data: TAB
## X-squared = 1.5382, df = 3, p-value = 0.6735

```

```

x <- seq(0, 20, by = 0.1)
chi_dense <- dchisq(x, df)

plot(x, chi_dense,type='l', xlab="x value",
     ylab="Density", main="Chi-square density")

i <- x >= quantile
lines(x, chi_dense)
polygon(c(quantile,x[i],20), c(0,chi_dense[i],0), col="blue")

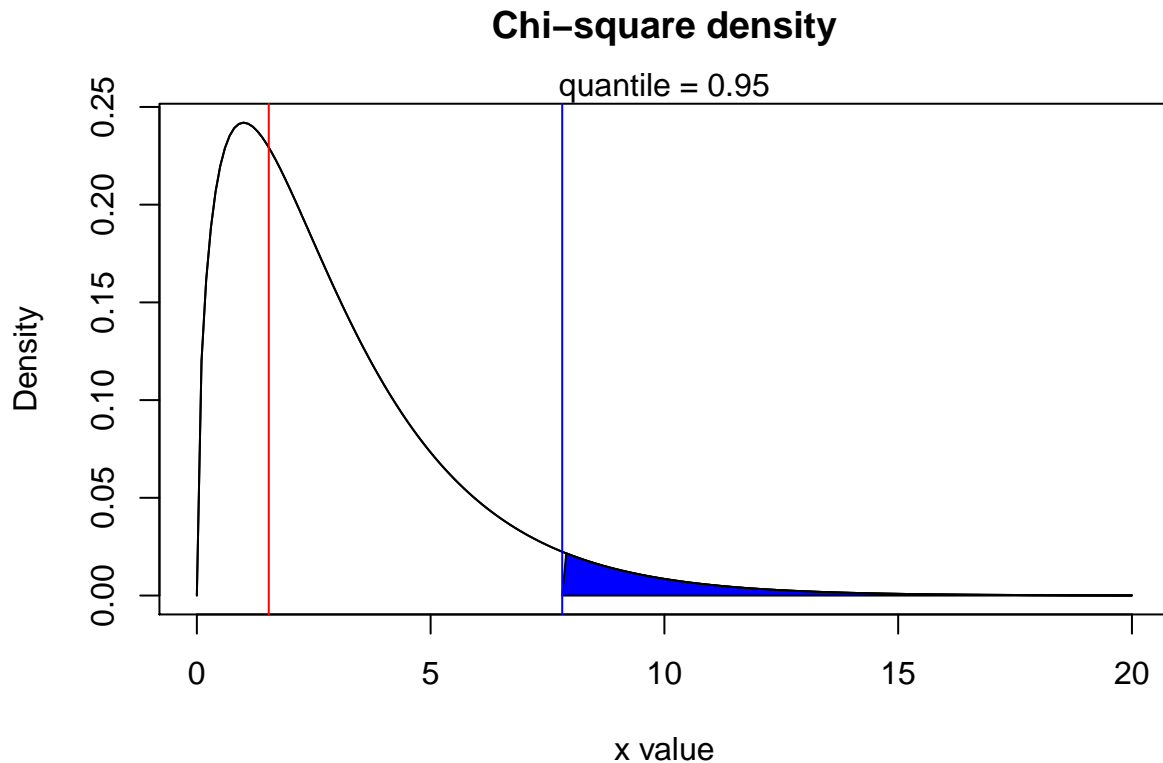
area <- pchisq(quantile, df, lower.tail = TRUE)

```

```

result <- paste("quantile =",
  signif(area, digits=3))
mtext(result,3)
abline(v=chi_sq, col="red")
abline(v=quantile, col="blue")

```



Nie możemy odrzucić hipotezy zerowej ponieważ wartość testu nie wpada do obszaru krytycznego / p-value jest większe niż ustalony poziom istotności.

c) Średnia zlogarytmowana cena notebooka Dell i HP jest równa.

Independent two-sample t-test wykorzystujemy, gdy chcemy porównać dwie grupy pod względem jakiejś zmiennej ilościowej.

$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$, gdzie \bar{X} , \bar{Y} - średnie arytmetyczne, s_X^2 , s_Y^2 - nieobciążone estymatory wariancji, n_X , n_Y - liczby obserwacji

Hipoteza zerowa H_0 : Średnia zlogarytmowana cena notebooka Dell jest taka sama jak średnia zlogarytmowana cena notebooka HP.

Hipoteza alternatywna H_1 : Średnie zlogarytmowane ceny notebooków różnią się.

Jedyné założenie do sprawdzenia: czy próby pochodzą z rozkładu normalnego.

```

alpha <- 0.1

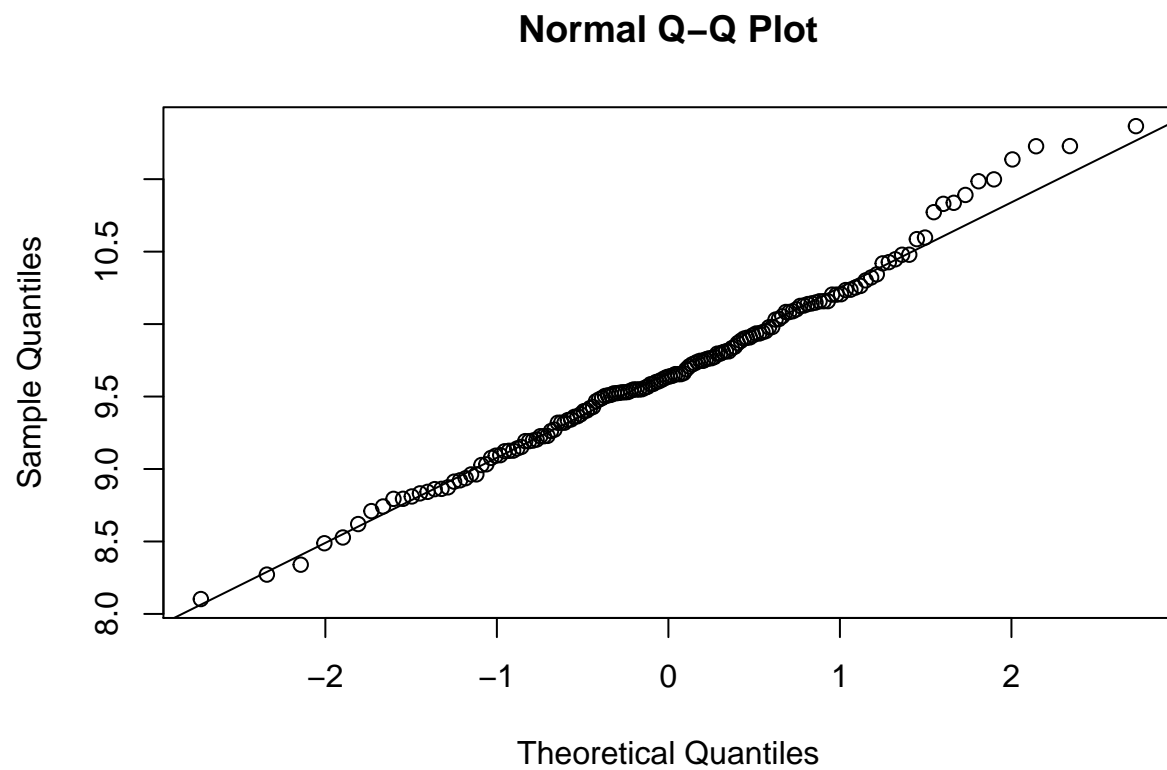
dellPrices <- dataSet[dataSet$company=="3", "price_euros"]
hpPrices <- dataSet[dataSet$company=="4", "price_euros"]

logDellPrices <- log2(dellPrices)
logHpPrices <- log2(hpPrices)

n <- length(dellPrices)
m <- length(hpPrices)

qqnorm(logDellPrices)
qqline(logDellPrices)

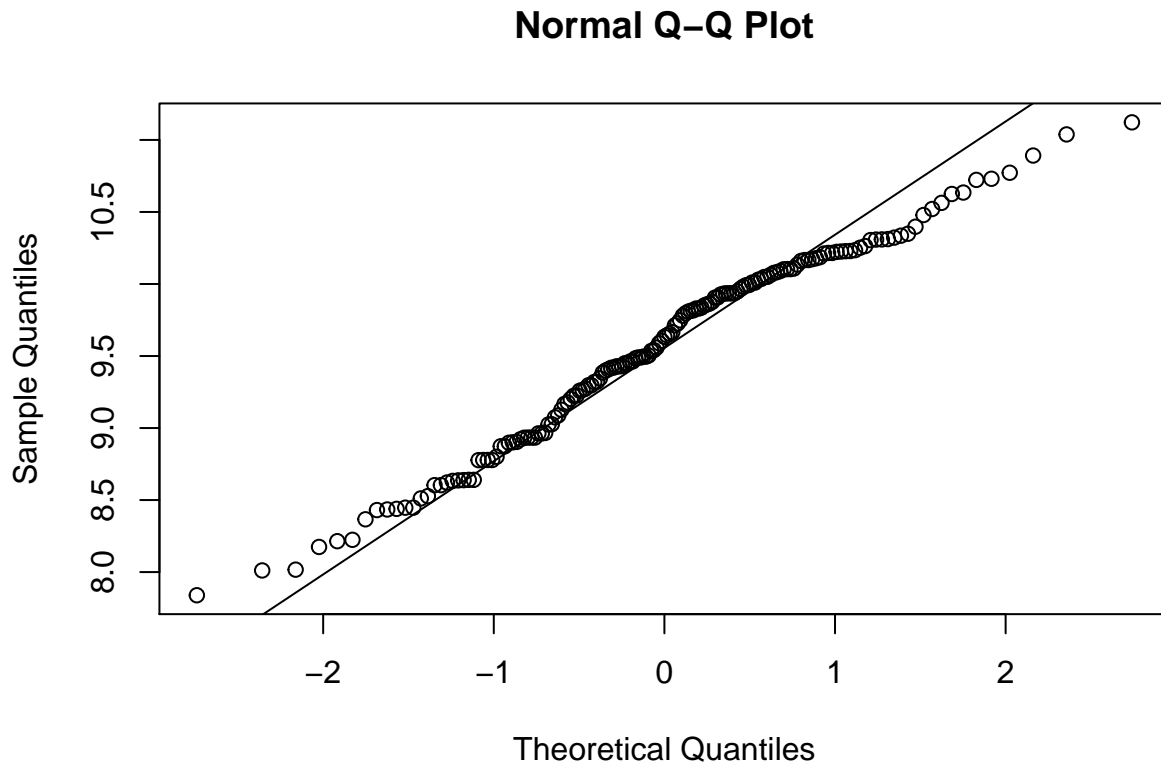
```



```

qqnorm(logHpPrices)
qqline(logHpPrices)

```



Linia prosta na wykresach QQ mówi nam, że nasze próby pochodzą z rozkładu normalnego.

```
alpha <- 0.05

dellPrices <- dataSet[dataSet$company=="3", "price_euros"]
hpPrices <- dataSet[dataSet$company=="4", "price_euros"]

logDellPrices <- log2(dellPrices)
logHpPrices <- log2(hpPrices)

n <- length(dellPrices)
m <- length(hpPrices)

#unbiased variance estimators
unbiased_estX <- 1/(n-1)*sum((logDellPrices-mean(logDellPrices))^2)
unbiased_estY <- 1/(m-1)*sum((logHpPrices-mean(logHpPrices))^2)

a <- unbiased_estX/n + unbiased_estY/m
t <- (mean(logDellPrices) - mean(logHpPrices))/sqrt(a) #test statistic

df <- a^2/(1/(n-1)*(unbiased_estX/n)^2+1/(m-1)*(unbiased_estY/m)^2) #deg of freedom

#two-tailed hypothesis
pval <- 2*pt(t, n+m-2, lower.tail = FALSE)

#confidence interval
```



```
lowerBound <- qt(alpha, n+m-2)
upperBound <- qt(1-alpha, n+m-2)
```

```
if(alpha > pval) {
  print("H0 rejected.")
}else {
  print("There is not enough evidence to reject H_0")
}
```

```
## [1] "There is not enough evidence to reject H_0"
```

```
sprintf("test statistic = %f , p-value=%f, confidence interval = (%f, %f)", t, pval, lowerBound, upperBound)
```

```
## [1] "test statistic = 1.481162 , p-value=0.139557, confidence interval = (-1.649675, 1.649675)"
```

```
#t.test
t.test(logDellPrices, logHpPrices)
```

```
##
## Welch Two Sample t-test
##
## data: logDellPrices and logHpPrices
## t = 1.4812, df = 315.63, p-value = 0.1396
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.03515752 0.24930117
## sample estimates:
## mean of x mean of y
## 9.670211 9.563139
```

```
x <- seq(-3, 3, by = 0.01)
t_dense <- dt(x, n+m-2)

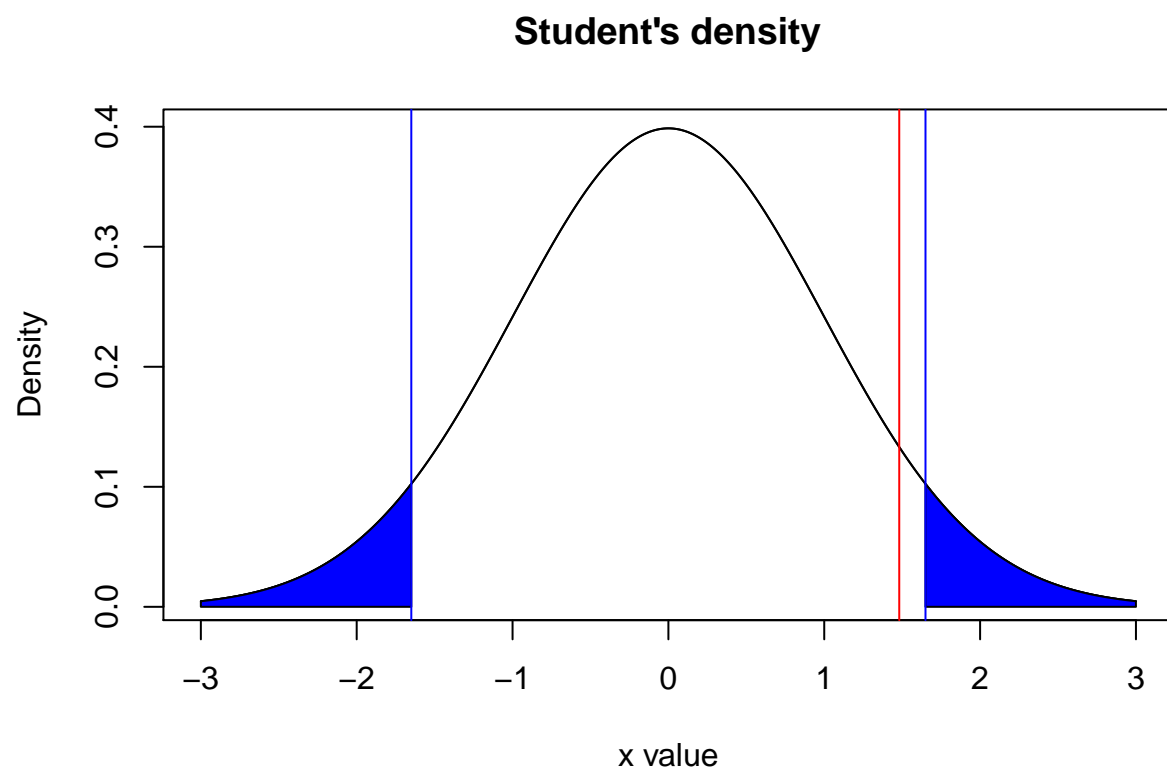
plot(x, t_dense, type='l', xlab="x value",
     ylab="Density", main="Student's density")

i <- x >= lowerBound & x <= upperBound
lines(x, t_dense)

polygon(x = c(-3, seq(-3, lowerBound, 0.01), lowerBound),
       y = c(0, dt(seq(-3, lowerBound, 0.01), n+m-2), 0),
       col = 'blue')

polygon(x = c(upperBound, seq(upperBound, 3, 0.01), 3),
       y = c(0, dt(seq(upperBound, 3, 0.01), n+m-2), 0),
       col = 'blue')

area <- pt(upperBound, n+m-2) - pt(lowerBound, n+m-2)
abline(v=t, col="red")
abline(v=lowerBound, col="blue")
abline(v=upperBound, col="blue")
```



Nie możemy odrzucić hipotezy zerowej - wartość testu nie wpada do obszaru krytycznego/ p-value większe niż ustalony poziom istotności.