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 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: $e = \frac{row.sum*col.sum}{grand.total}$
Chi-square statistic: $\chi^2 = \sum \frac{(o-e)^2}{e}$, gdzie o - obserwacja, e - wartosc 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
}
#observations
#obs <- matrix(rep(0, 4*7), nrow=7, ncol=4)

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
```

```r
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."
```

```r
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]"
```

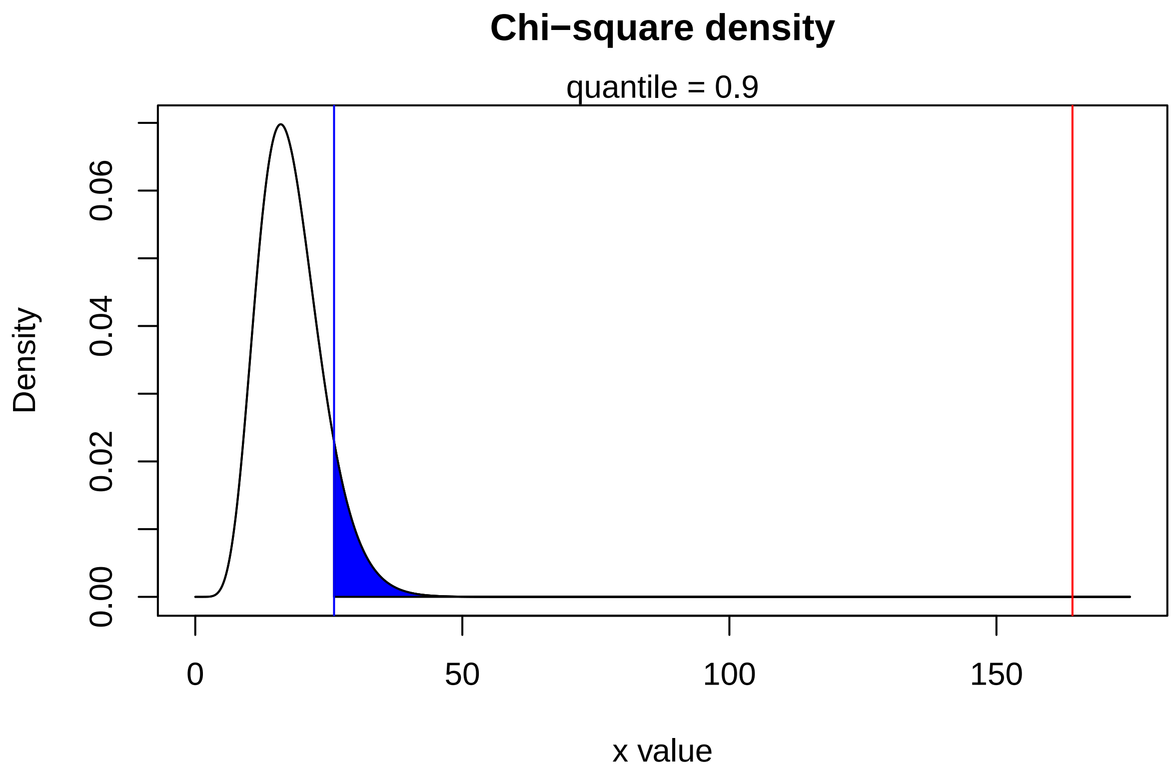
```r
#chisq.test()
test <- chisq.test(TAB)
```

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

```r
test
```

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

```r
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 of Two Variances sprawdza hipotezę, że próby pochodzą z populacji o jednakowych wariancjach.

$F = \frac{S_X^2}{S_Y^2}$, gdzie S_X^2 i S_Y^2 to wariancje próbkowe.

Hipoteza zerowa H_0: Równe wariancje.
Hipoteza alternatywna H_1: Różne wariancje.

```
alpha <- 0.1

#independent populations
LenovoRAM <- dataSet[dataSet$company=="5", ]$ram
HPRAM <- dataSet[dataSet$company=="4", ]$ram

n <- length(LenovoRAM)
m <- length(HPRAM)

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

#fisher's test
```

```r
F <- unbiased_estX/(unbiased_estY) # Snedecor's F distribution
pval <- pf(F, n-1, m-1, lower.tail = FALSE)
quantile <- qf(alpha, n-1, m-1, lower.tail = FALSE) #right-tailed

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

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

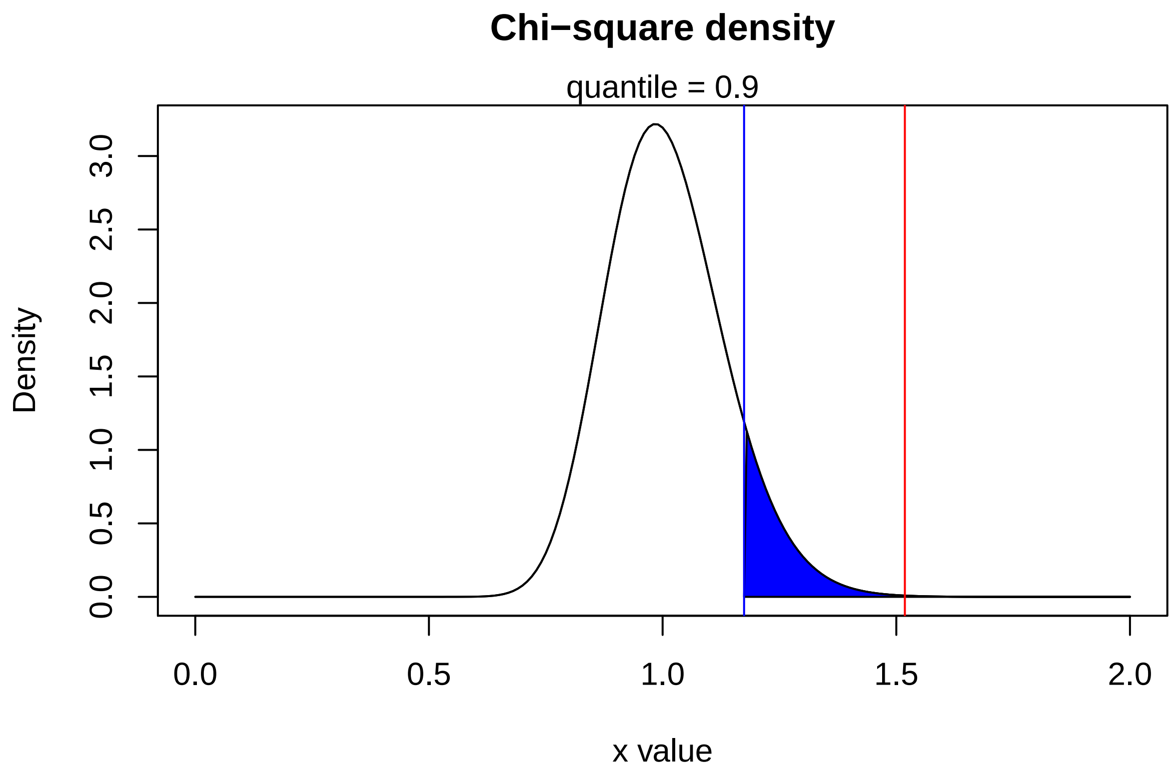
```r
sprintf("test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", F, pval, quantile)
```

```
## [1] "test statistic = 1.518314 , p-value = 0.000456, confidece interval = [-infinity, 1.174316]"
```

```r
#sprawdźmy
var.test(LenovoRAM, HPRAM, 1)
```

```
##
##  F test to compare two variances
##
## data:  LenovoRAM and HPRAM
## F = 1.5183, num df = 271, denom df = 244, p-value = 0.0009113
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.187251 1.938597
## sample estimates:
## ratio of variances
##           1.518314
```

```r
x <- seq(0, 2, by = 0.01)
f_dense <- df(x, n-1, m-1)

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

i <- x >= quantile
lines(x, f_dense)
polygon(c(quantile,x[i],2), c(0,f_dense[i],0), col="blue")

area <- pf(quantile, n-1, m-1, lower.tail = TRUE)
result <- paste("quantile =",
   signif(area, digits=3))
mtext(result,3)
abline(v=F, col="red")
abline(v=quantile, col="blue")
```



Odrzucamy hipotezę zerową ponieważ wartość testu wpada do obszaru krytycznego / p-value jest mniejsze niż ustalony poziom istotności. **Rozkład stosowanych pamięci RAM w notebookach HP i Lenovo nie jest taki sam.**

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.

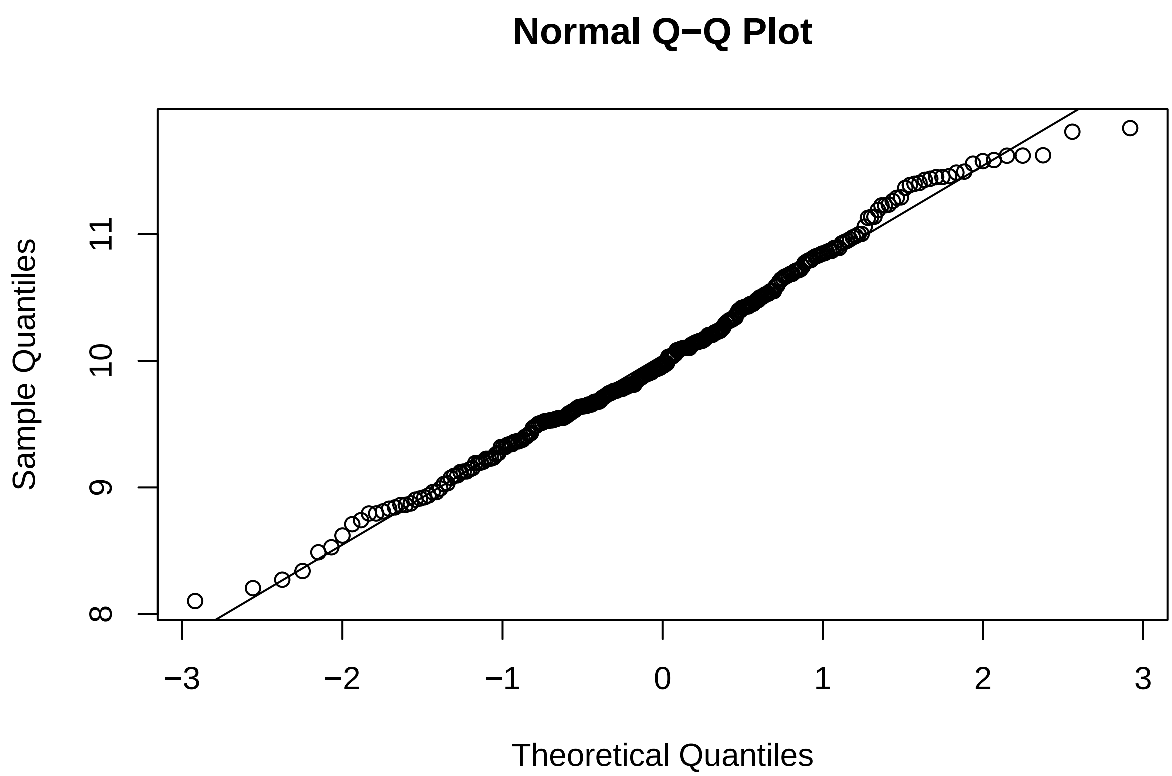
$t = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}}}$, gdzie $\overline{X}$, $\overline{Y}$ - średnie arytmetyczne, s_X^2, s_Y^2 - nieobciążone estymatory wariancji, n_X, n_Y - liczby obserwacji

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.

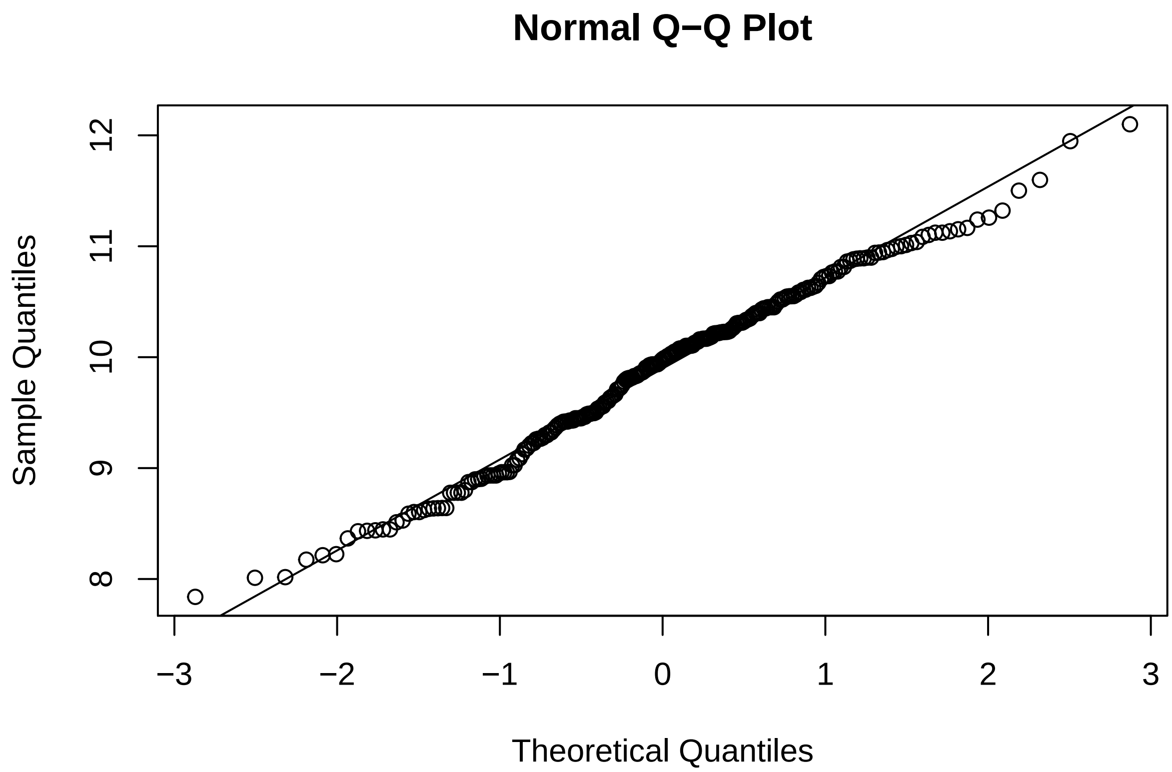
Jedyne założenie do sprawdzenia: czy próby pochodzą z rozkładu normalnego.

```
alpha <- 0.1

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

logDellPrices <- log2(dellPrices)
```

```r
logHpPrices <- log2(hpPrices)

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

qqnorm(logDellPrices)
qqline(logDellPrices)
```




```r
qqnorm(logHpPrices)
qqline(logHpPrices)
```



Linia prosta na wykresach QQ mówi nam, że nasze 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)

#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
```

```r
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] "H0 rejected."
```

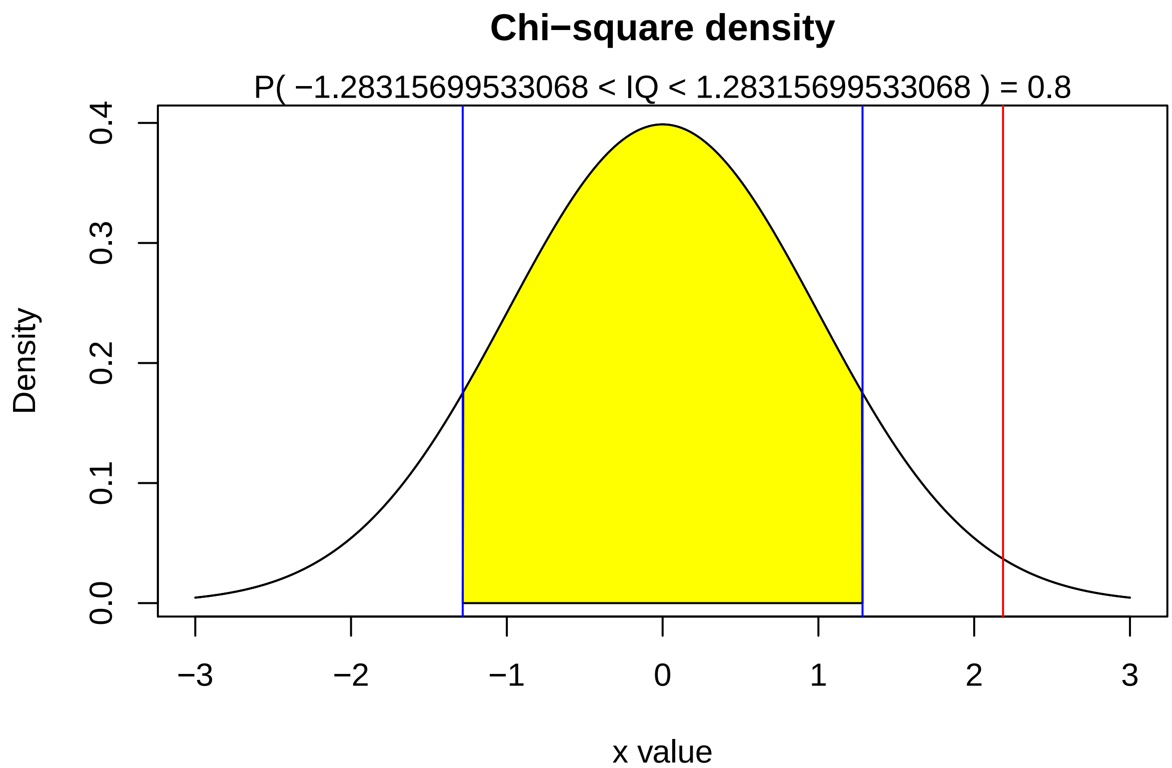
```r
sprintf("test statistic = %f , p-value=%f, confidence interval = (%f, %f)", t, pval, lowerBound, upperBo
```

```
## [1] "test statistic = 2.185085 , p-value=0.029321, confidence interval = (-1.283157, 1.283157)"
```

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

```
##
##  Welch Two Sample t-test
##
## data:  logDellPrices and logHpPrices
## t = 2.1851, df = 503.74, p-value = 0.02934
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01507785 0.28388985
## sample estimates:
## mean of x mean of y
##  10.03851   9.88903
```

```r
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="Chi-square density")

i <- x >= lowerBound & x <= upperBound
lines(x, t_dense)
polygon(c(lowerBound,x[i],upperBound), c(0,t_dense[i],0), col="yellow")

area <- pt(upperBound, n+m-2) - pt(lowerBound, n+m-2)
result <- paste("P(",lowerBound,"< IQ <",upperBound,") =",
   signif(area, digits=3))
mtext(result,3)
abline(v=t, col="red")
abline(v=lowerBound, col="blue")
abline(v=upperBound, col="blue")
```



Odrzucamy hipotezę zerową - wartość testu wpada do obszaru krytycznego/ p-value mniejsze niż ustalony poziom istotności - na rzecz hipotezy alternatywnej. **Średnie zlogarytmowane ceny notebooków różnią się.**