Pierwszy projekt zaliczeniowy

Statystyczna analiza danych 2020/2021

Joanna Kęczkowska

26.04.2021

Celem zadania jest statystyczna analiza danych znajdujących się w pliku people.tab. Dane: Są to dane symulowane; opisują wiek (zmienna age), wagę (weight), wzrost (height), płeć (gender), stan cywilny (married), liczbę dzieci (number_of_kids), posiadane zwierzę domowe (pet) oraz miesięczne wydatki (expenses) pewnych osób. We wszystkich zadaniach poniżej zmienna expenses jest zmienną objaśnianą (zależną), a pozostałe zmienne są zmiennymi objaśniającymi (niezależnymi).

1. Wczytaj dane, obejrzyj je i podsumuj w dwóch-trzech zdaniach. Pytania pomocnicze: ile jest obserwacji, ile zmiennych ilościowych, a ile jakościowych? Czy są zależności w zmiennych objaśniających (policz i zaprezentuj na wykresach korelacje pomiędzy zmiennymi ilościowymi, a także zbadaj zależność zmiennych jakościowych). Skomentuj wyniki. Czy występują jakieś braki danych?

```
df <- read.delim("peopletab.txt", header = TRUE, sep='\t')
sprintf("Dane zawierają %d obserwacji i %d cech", dim(df)[1], dim(df)[2])
```

```
## [1] "Dane zawierają 500 obserwacji i 8 cech"
```

```
summary(df)
```

```
##       age            weight           height         gender
##  Min.   :17.00   Min.   : 19.40   Min.   :113.6   Length:500
##  1st Qu.:33.00   1st Qu.: 57.60   1st Qu.:155.6   Class :character
##  Median :39.00   Median : 66.60   Median :169.0   Mode  :character
##  Mean   :39.48   Mean   : 66.39   Mean   :168.2
##  3rd Qu.:45.00   3rd Qu.: 75.30   3rd Qu.:180.1
##  Max.   :72.00   Max.   :107.20   Max.   :235.2
##   married        number_of_kids      pet              expenses
##  Mode :logical   Min.   :0.000   Length:500         Min.   :-685.68
##  FALSE:327       1st Qu.:0.750   Class :character   1st Qu.:  74.51
##  TRUE :173       Median :1.000   Mode  :character   Median : 402.22
##                  Mean   :1.558                      Mean   : 478.60
##                  3rd Qu.:2.000                      3rd Qu.: 802.72
##                  Max.   :6.000                      Max.   :3503.90
```

Dane zawierają **500 obserwacji**.
Zmienne ilościowe: 'age', 'weight', 'height', 'expenses'.
Zmienne jakościowe: 'gender', 'married', 'pet', 'number_of_kids'. Niepokojące są ujemne wartości w cesze 'expanses', jak również factor 'other' w cesze 'gender'. Wartość 'none' w cesze 'pet' interpretuję jako nieposiadanie zwierzęcia. W zmiennej 'gender' potraktuję factor 'other' jako brak informacji.

```
df$gender <- factor(df$gender)
df$pet <- factor(df$pet)
df$married <- factor(df$married)
df$number_of_kids <- factor(df$number_of_kids)
summary(df)
```

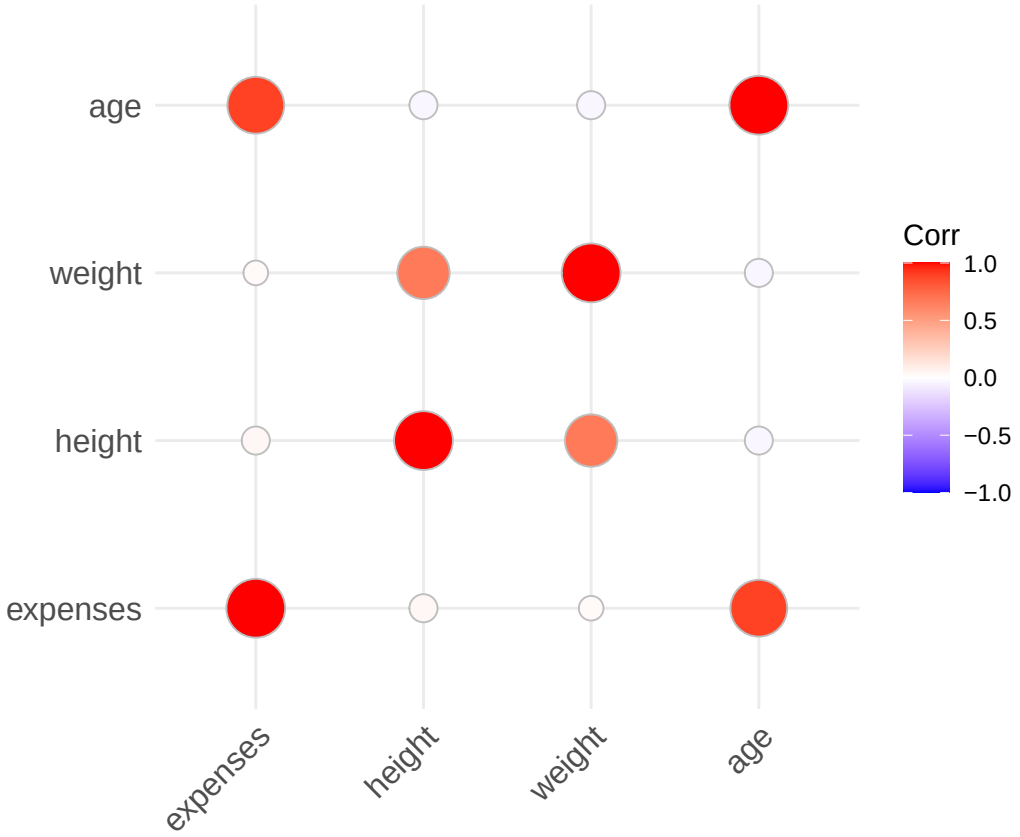
```
##       age            weight           height        gender     married
##  Min.   :17.00   Min.   : 19.40   Min.   :113.6   man  :223   FALSE:327
##  1st Qu.:33.00   1st Qu.: 57.60   1st Qu.:155.6   other: 38   TRUE :173
##  Median :39.00   Median : 66.60   Median :169.0   woman:239
##  Mean   :39.48   Mean   : 66.39   Mean   :168.2
##  3rd Qu.:45.00   3rd Qu.: 75.30   3rd Qu.:180.1
##  Max.   :72.00   Max.   :107.20   Max.   :235.2
##
##  number_of_kids       pet         expenses
##  0:125          cat     :105   Min.   :-685.68
##  1:161          dog     :100   1st Qu.:  74.51
##  2: 99          ferret  : 54   Median : 402.22
##  3: 63          hedgehog: 54   Mean   : 478.60
##  4: 35          none    :187   3rd Qu.: 802.72
##  5: 11                         Max.   :3503.90
##  6:  6
```

Współczynnik korelacji r jest liczbą pomiędzy –1 i 1, która określa, w jakim stopniu dwie zmienne są współzależne. Wartość r = 0 oznacza, że nie ma żadnego powiązania, a wartość 1 lub –1 oznacza idealne powiązanie. Znak współczynnika korelacji wskazuje, czy zmienne są skorelowane dodatnio (większe wartości w jednej zmiennej pokrywają się z większymi wartościami w drugiej), czy też ujemnie (większe wartości w jednej zmiennej pokrywają się z mniejszymi wartościami w drugiej).

```
library(ggcorrplot)
```

```
## Loading required package: ggplot2
```

```
numerical <- df[c("expenses", "height", "weight", "age")]
categorial <- df[c("married", "gender", "pet", "number_of_kids")]

corr <- round(cor(numerical), 2)

ggcorrplot(corr, method = "circle")
```



Zgodnie z intuicją **wiek jest dodatnio skorelowany z zarobkami** i **wzrost jest dodatnio skorelowany z wagą**.

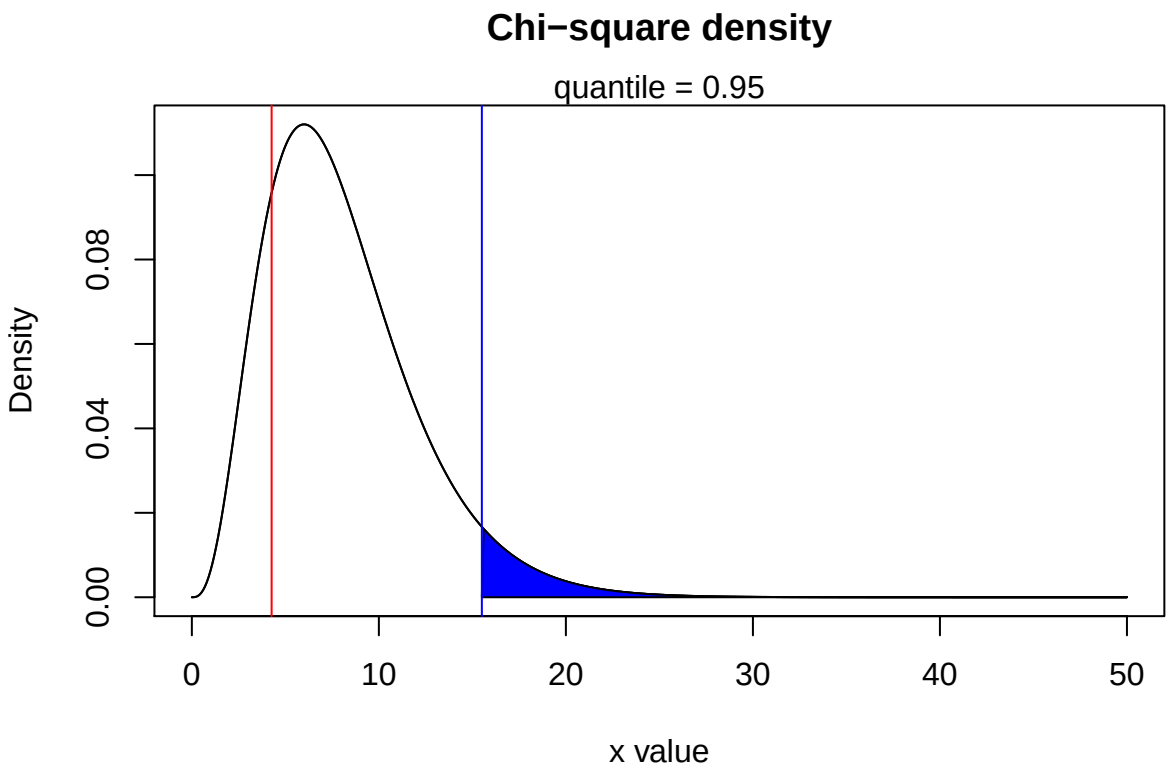
W przypadku zmiennych jakościowych nie możemy zbadać korelacji tak jak dla zmiennych ilościowych - przypisane do nich wartości liczbowe są jedynie symboliczne.
Dla tego typu zmiennych posłużymy się testem zgodności χ^2

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 - wartość oczekiwana

Hipoteza zerowa H_0: Zmienne są **niezależne**.
Hipoteza alternatywna H_1: Zmienne są **zależne**.

```
#funkcja do testowania korelacji zmiennych jakościowych
#przyjmuje dwie kolumny zmiennych kategorycznych, które zamienia na tablicę wielodzielczą

testchi <- function(feature1, feature2, sq = 50) {
  alpha <- 0.05 #5% level of significance
  TAB <- table(feature1, feature2)
  total <- sum(TAB)

  n <- nlevels(feature1)
  m <- nlevels(feature2)

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

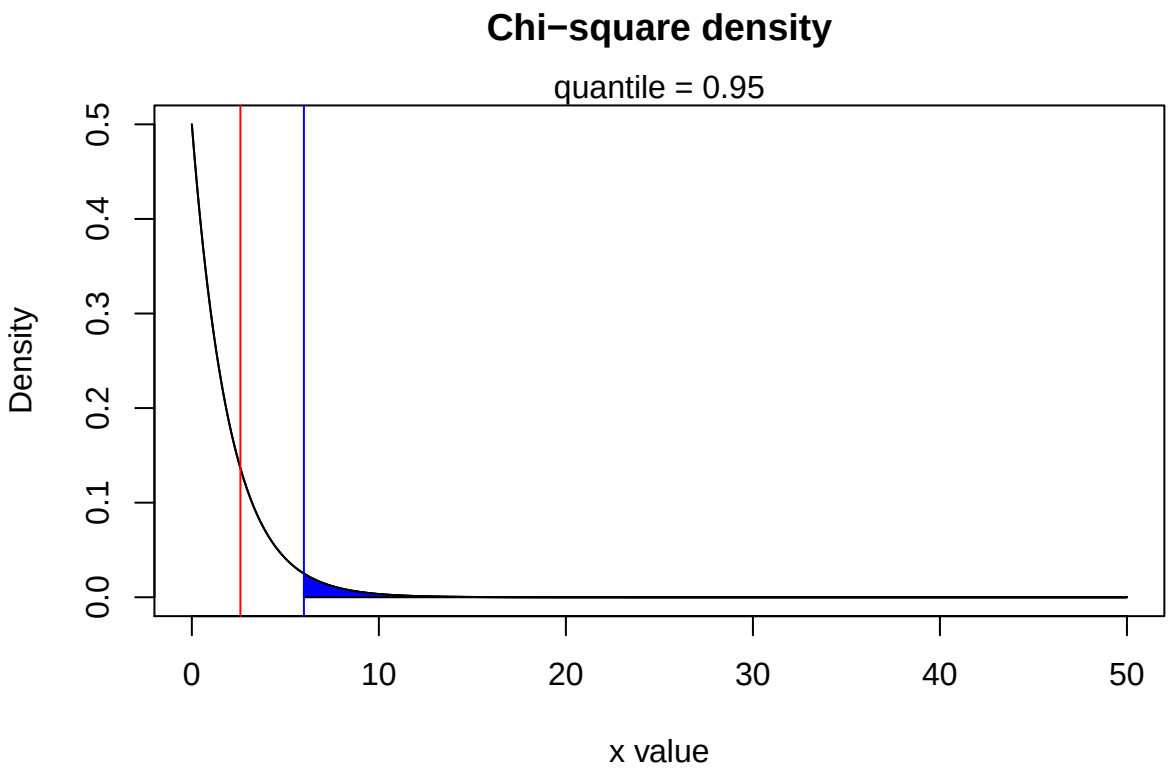
```
  sumRows <- as.vector(sumRows)
  sumCols <- as.vector(sumCols)

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

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

  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

  x <- seq(0, sq, 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],sq), 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")

  c <- list(chi_sq, pval, quantile)
  return (c)

}

gp <- testchi(df$gender, df$pet)
```




```
sprintf("GENDER/PET, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", gp[[1]]
```

```
## [1] "GENDER/PET, test statistic = 4.264540 , p-value = 0.832502, confidece interval = [-infinity, 15
```

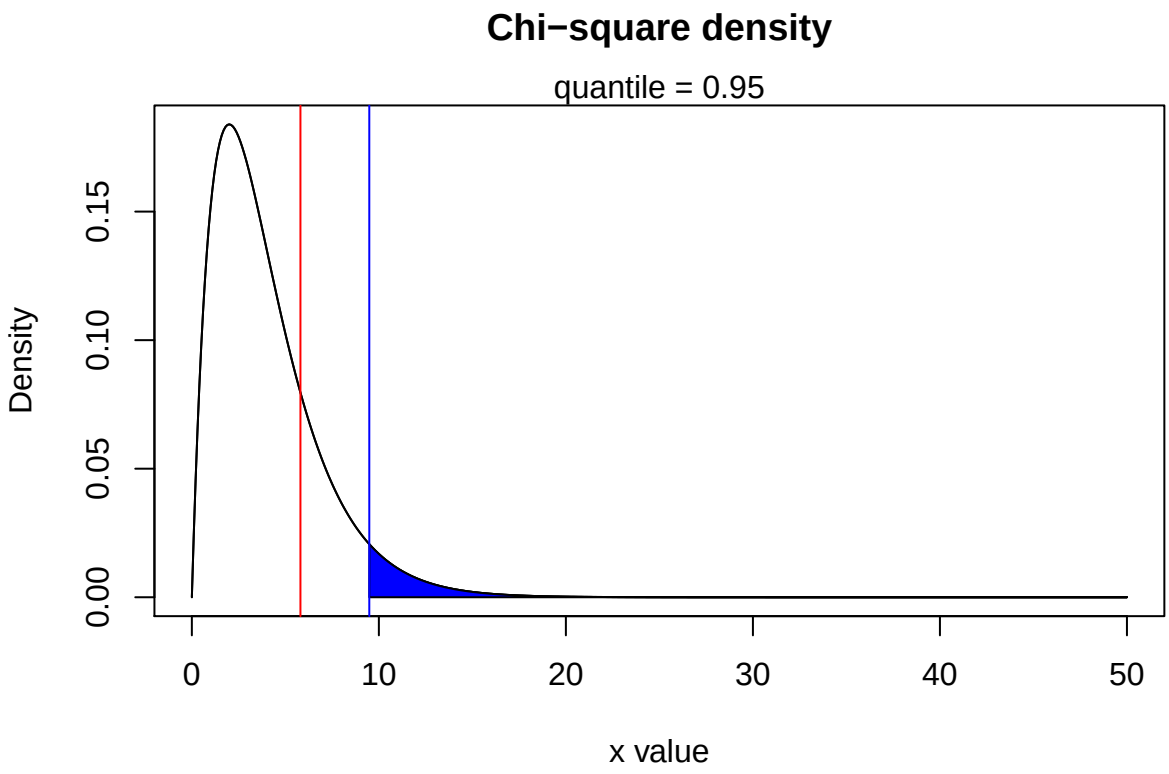
```
gm <- testchi(df$gender, df$married)
```




```
sprintf("GENDER/MARRIED, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", gm[
```

```
## [1] "GENDER/MARRIED, test statistic = 2.597089 , p-value = 0.272929, confidece interval = [-infinity
```

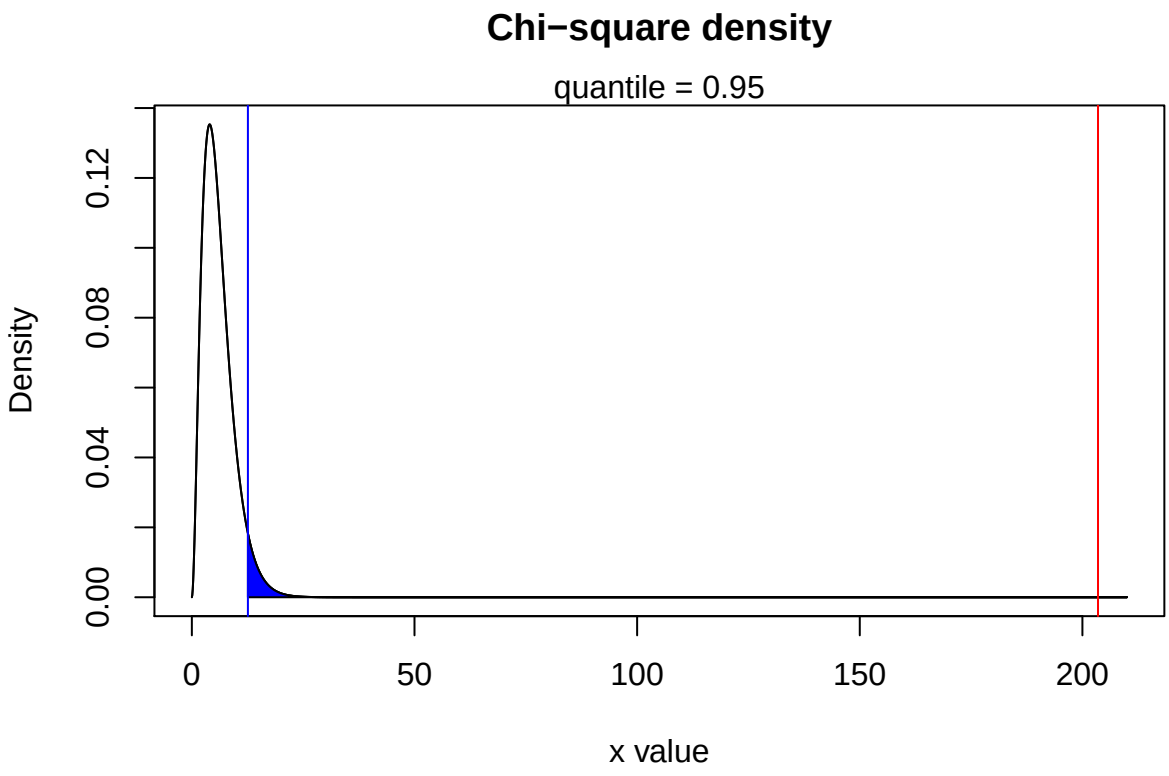
```
pm <- testchi(df$pet, df$married)
```




```
sprintf("PET/MARRIED, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", pm[[1]]
```

```
## [1] "PET/MARRIED, test statistic = 5.806971 , p-value = 0.214035, confidece interval = [-infinity, 9
```

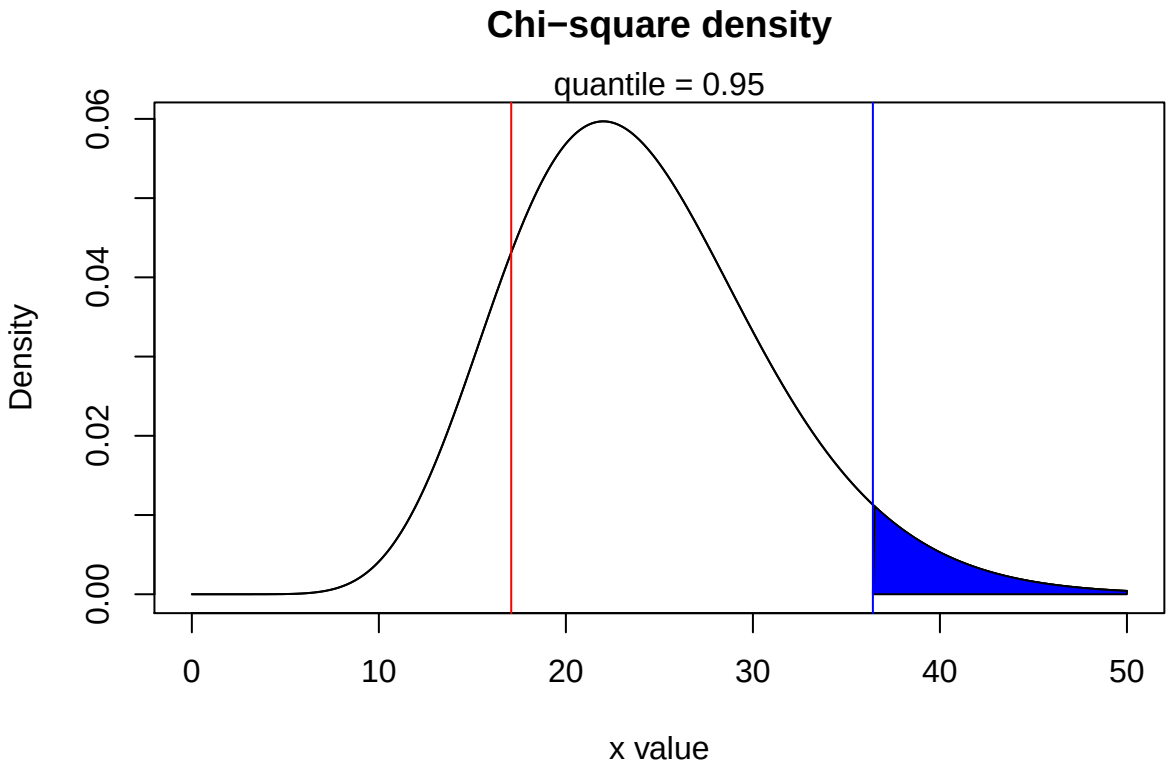
```
nm <- testchi(df$number_of_kids, df$married, sq=210)
```




```
sprintf("KIDS/MARRIED, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", nm[[1
```

```
## [1] "KIDS/MARRIED, test statistic = 203.501380 , p-value = 0.000000, confidece interval = [-infinity
```

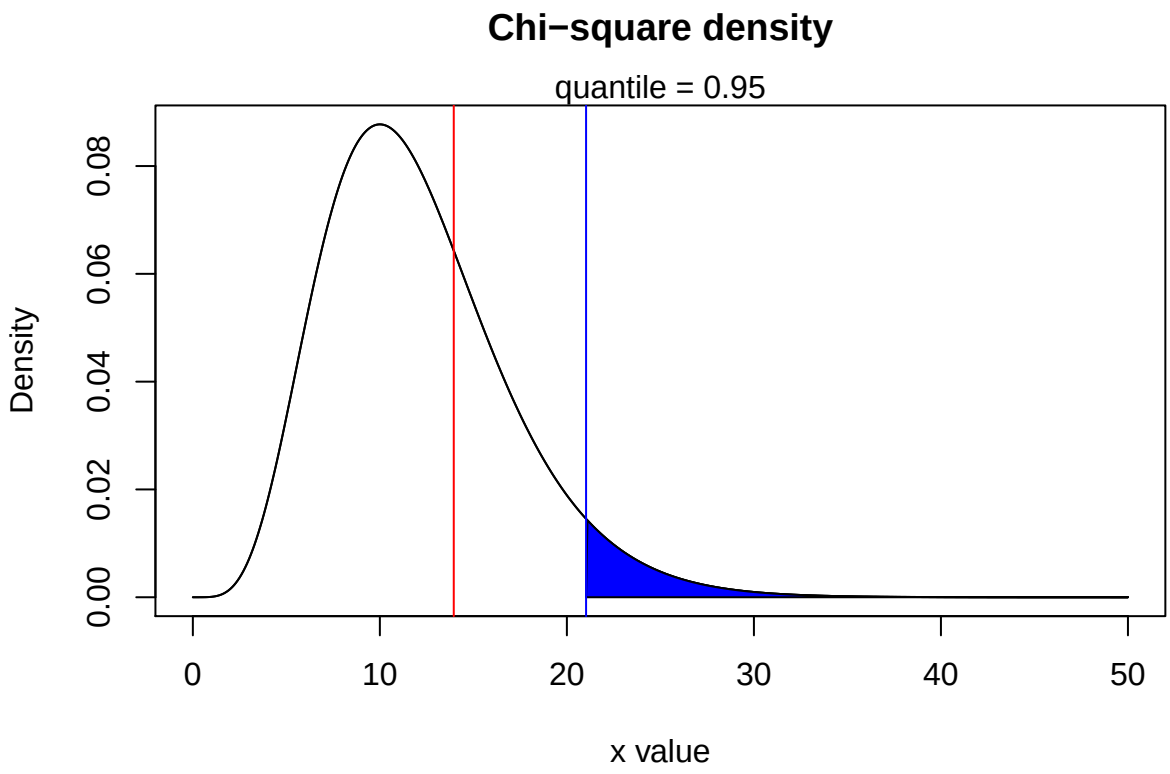
```
np <- testchi(df$number_of_kids, df$pet)
```




```
sprintf("KIDS/PET, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", np[[1]], n
```

```
## [1] "KIDS/PET, test statistic = 17.076893 , p-value = 0.845364, confidece interval = [-infinity, 36.4
```

```
gn <- testchi(df$gender, df$number_of_kids)
```




```
sprintf("PET/MARRIED, test statistic = %f , p-value = %f, confidece interval = [-infinity, %f]", gn[[1]]
```

```
## [1] "PET/MARRIED, test statistic = 13.950817 , p-value = 0.303860, confidece interval = [-infinity, 1
```

Jedyne dwie skorelowane zmienne jakościowe to 'number_of_kids' i 'married' - statystyka testowa wpada do obszaru krytycznego. W przypadku pozostałych par zmiennych nie mamy podstawy do odrzucenia hipotezy zerowej. Żadne dwie inne zmienne nie wydają się być skorelowane.

Jeszcze tylko szybkie sprawdzenie:

```
#sprawdźmy
chisq.test(table(df$gender, df$pet))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$gender, df$pet)
## X-squared = 4.2645, df = 8, p-value = 0.8325
```

```
chisq.test(table(df$gender, df$married))
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  table(df$gender, df$married)
## X-squared = 2.5971, df = 2, p-value = 0.2729
```

```
chisq.test(table(df$pet, df$married))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$pet, df$married)
## X-squared = 5.807, df = 4, p-value = 0.214
```

```
chisq.test(table(df$number_of_kids, df$married))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$number_of_kids, df$married)
## X-squared = 203.5, df = 6, p-value < 2.2e-16
```

```
chisq.test(table(df$number_of_kids, df$pet))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$number_of_kids, df$pet)
## X-squared = 17.077, df = 24, p-value = 0.8454
```

```
chisq.test(table(df$gender, df$number_of_kids))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$gender, df$number_of_kids)
## X-squared = 13.951, df = 12, p-value = 0.3039
```

2. Podsumuj dane przynajmniej trzema różnymi wykresami. Należy przygotować: **a)** wykres typu scatter-plot (taki jak na wykładzie 6, slajd 3) dla wszystkich zmiennych objaśniających ilościowych i zmiennej objaśnianej. **b)** Wykresy typu pudełkowy (boxplot) dla jednej wybranej zmiennej ilościowej. **c)** Wykres typu słupkowy (barplot) dla jednej wybranej zmiennej jakościowej. Dodatkowe wykresy wg własnej inwencji (np. histogram, punktowy, liniowy, mapa ciepła...).

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(numerical, title="Correlogram of numerical features")
```