Analiza uspjeha učenika

SAP - projekt

'Drim Tim' - Danijel Kovačević, Luka Panđa, Martin Ante Rogošić, Marko Sršić

1.1.2025.

1. Uvod

Utjecaj različitih čimbenika na školski uspjeh učenika je oduvijek bila jedna od najzanimljivijih tema za istraživače iz raznih područja; kako je pohađanje škole i polaganje ispita sastavni dio života većine ljudi u današnjem modernom dobu, interes za dublje razumijevanje individualnih, sociodemografskih i društvenih karakteristika koje oblikuju akademska postignuća učenika/studenata neprestano raste. Takva istraživanja mogu pridonijeti stvaranju pravednijeg i učinkovitijeg obrazovnog sustava koji može bolje odgovoriti na potrebe različitih skupina učenika.

U sklopu projekta na kolegiju "Statistička analiza podataka", grupa "Drim Tim" pokušat će na temelju podataka o različitim značajkama za više od 350 učenika iz dviju portugalskih srednjih škola odgovoriti na neka zanimljiva istraživačka pitanja koja će čitatelju približiti odnos i utjecaj tih značajki na ispitni uspjeh iz matematike i portugalskog jezika.

Analiza obuhvaća korištenje različitih statističkih metoda i tehnika te je zbog toga podijeljena u nekoliko poglavlja.

U 2. poglavlju se postupcima deskriptivne statistike nastoji dobiti bolji uvid u korišteni podatkovni skup. U tu svrhu će se provjeriti priroda značajki (jesu li kategorijske ili numeričke) i izračunati sažetak 5 brojeva (minimum, maksimum, Q1, medijan, Q3) te srednja vrijednost. U 3. poglavlju su postavljena određena istraživačka pitanja od interesa, poput "razlika u ocjeni iz matematike s obzirom na mjesto stanovanja" ili "predviđanje uspjeha na završnom ispitu iz jezika na temelju sociodemografskih varijabli". Korištenjem deskriptora poput srednje vrijednosti ili varijance i vizualizacijom postavlja se početna hipoteza čija se ispravnost provjerava provedbom odgovarajućeg statističkog testa. Na temelju rezultata tog statističkog testa možemo (ili ne možemo) donijeti zaključak o hipotezi koju smo postavili.
Posljednje, 4. poglavlje donosi sažetak analize i najvažnije/najzanimljivije zaključke.

2. Deskriptivna analiza ulaznog skupa podataka

```
# učitavanje paketa
library(dplyr)
```

Podaci koji će nam poslužiti za statističku analizu uspjeha učenika prikupljeni su školske godine 2005/2006. u dvije portugalske srednje škole - "Gabriel Pereira" (GP) i "Mousinho da Silveira" (MS). Upoznajmo se sa značajkama podatkovnog skupa.

```
df <- read.csv("student_data.csv")
head(df)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
```

```
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures_mat failures_por schoolsup famsup
## 1   mother          2         2            0            0       yes     no
## 2   father          1         2            0            0        no    yes
## 3   mother          1         2            3            0       yes     no
## 4   mother          1         3            0            0        no    yes
## 5   father          1         2            0            0        no    yes
## 6   mother          1         2            0            0        no    yes
##   paid_mat paid_por activities nursery higher internet romantic famrel freetime
## 1       no       no         no     yes    yes       no       no      4        3
## 2       no       no         no      no    yes      yes       no      5        3
## 3      yes       no         no     yes    yes      yes       no      4        3
## 4      yes       no        yes     yes    yes      yes      yes      3        2
## 5      yes       no         no     yes    yes       no       no      4        3
## 6      yes       no        yes     yes    yes      yes       no      5        4
##   goout Dalc Walc health absences_mat absences_por G1_mat G2_mat G3_mat G1_por
## 1     4    1    1      3            6            4      5      6      6      0
## 2     3    1    1      3            4            2      5      5      6      9
## 3     2    2    3      3           10            6      7      8     10     12
## 4     2    1    1      5            2            0     15     14     15     14
## 5     2    1    2      5            4            0      6     10     10     11
## 6     2    1    2      5           10            6     15     15     15     12
##   G2_por G3_por
## 1     11     11
## 2     11     11
## 3     13     12
## 4     14     14
## 5     13     13
## 6     12     13
```

```
dim(df)
```

```
## [1] 370  39
```

Imamo podatke o 370 učenika i 39 značajki koje ih opisuju. Te značajke su:

```
names(df)
```

```
##  [1] "school"       "sex"          "age"          "address"      "famsize"
##  [6] "Pstatus"      "Medu"         "Fedu"         "Mjob"         "Fjob"
## [11] "reason"       "guardian"     "traveltime"   "studytime"    "failures_mat"
## [16] "failures_por" "schoolsup"    "famsup"       "paid_mat"     "paid_por"
## [21] "activities"   "nursery"      "higher"       "internet"     "romantic"
## [26] "famrel"       "freetime"     "goout"        "Dalc"         "Walc"
## [31] "health"       "absences_mat" "absences_por" "G1_mat"       "G2_mat"
## [36] "G3_mat"       "G1_por"       "G2_por"       "G3_por"
```

Vidimo da uz osnovne značajke poput škole, spola i godina imamo niz sociodemografskih značajki (veličina obitelji, obrazovanje roditelja, zanimanje roditelja, adresa...), obrazovnih značajki (tjedno vrijeme učenja, dodatne plaćene instrukcije...), društvenih značajki (izlasci, konzumacija alkohola...) te značajke koje opisuju uspjeh učenika na ispitima iz matematike i jezika. Značajke su kategorijskog i numeričkog tipa.

U ovom dijelu radimo samo početni pregled podataka za neke odabrane značajke. Nećemo svaku značajku opisivati zasebno; ako značajka bude korištena u analizi u nekom statističkom testu, tada ćemo navesti opis

te značajke kao i potrebnu deskriptivnu statistiku u sklopu samog testa.

Funkcijom *summary* za 'prave' numeričke podatke možemo dobiti sažetak 5 brojeva i srednju vrijednost:

```
true_numeric <- c('age','absences_mat', 'absences_por',
                  'G1_mat','G2_mat','G3_mat','G1_por','G2_por','G3_por')
summary(df[true_numeric])
```

```
##       age         absences_mat     absences_por        G1_mat
##  Min.   :15.00   Min.   : 0.000   Min.   : 0.000   Min.   : 3.00
##  1st Qu.:16.00   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 8.00
##  Median :17.00   Median : 4.000   Median : 2.000   Median :11.00
##  Mean   :16.58   Mean   : 5.381   Mean   : 3.632   Mean   :10.89
##  3rd Qu.:17.00   3rd Qu.: 8.000   3rd Qu.: 6.000   3rd Qu.:13.00
##  Max.   :22.00   Max.   :75.000   Max.   :32.000   Max.   :19.00
##      G2_mat          G3_mat          G1_por          G2_por
##  Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 5.00
##  1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.:10.00   1st Qu.:11.00
##  Median :11.00   Median :11.00   Median :12.00   Median :12.00
##  Mean   :10.75   Mean   :10.46   Mean   :12.14   Mean   :12.27
##  3rd Qu.:13.00   3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.00
##  Max.   :19.00   Max.   :20.00   Max.   :19.00   Max.   :19.00
##      G3_por
##  Min.   : 0.00
##  1st Qu.:11.00
##  Median :13.00
##  Mean   :12.55
##  3rd Qu.:14.00
##  Max.   :19.00
```

Za značajku *age* vidimo da je raspon godina učenika [15, 22], medijan 17 godina i srednja vrijednost 16.58 godina. Najviše izostanaka imao je jedan učenik na matematici (75), a medijan i srednja vrijednost rezultata učenika veći su od 10 bodova na svim ispitima iz matematike i jezika.

Koristeći IQR (Q3-Q1) možemo provjeriti potencijalne stršeće podatke. To su oni podaci koji su ili manji od 1.5IQR - Q1 ili veći od 1.5IQR + Q3. Stršeći podaci mogu se pojaviti kod značajki koje nisu ograničene, a to su *absences_mat* i *absences_por*.

```
absences <- c('absences_mat', 'absences_por')

find_outliers <- function(column) {
  Q1 <- quantile(column, 0.25, na.rm = TRUE)
  Q3 <- quantile(column, 0.75, na.rm = TRUE)
  IQR_value <- IQR(column, na.rm = TRUE)

  # Donja i gornja granica za outliere
  lower_bound <- Q1 - 1.5 * IQR_value
  upper_bound <- Q3 + 1.5 * IQR_value

  # Vraćanje indeksa outliera
  outliers <- which(column < lower_bound | column > upper_bound)
  return(outliers)
}

outliers_list <- lapply(df[absences], find_outliers)

print('Potencijalni stršeći podaci - izostanci iz matematike:')
```

```
## [1] "Potencijalni stršeći podaci - izostanci iz matematike:"
```

```
print(df$absences_mat[outliers_list$absences_mat])
```

```
##  [1] 25 54 26 56 24 28 22 21 75 22 30 23
```

```
print('Potencijalni stršeći podaci - izostanci iz jezika:')
```

```
## [1] "Potencijalni stršeći podaci - izostanci iz jezika:"
```

```
print(df$absences_por[outliers_list$absences_por])
```

```
##  [1] 16 16 22 16 32 16 16 30 21 16 16 16 22 18 18
```

Jedini podatak koji jako odudara je onaj za učenika koji je imao 75 izostanaka na matematici, ali ni on nije toliko nezamislivo velik. Zasad nećemo ove potencijalne stršeće podatke izbacivati.

Sažetak 5 brojeva i srednju vrijednost nije previše korisno provjeriti za značajke koje su po strukturi numeričke, ali su po prirodi kategorijske (npr. *Medu* poprima vrijednosti od 0 do 4 koje reprezentiraju ordinalnu skalu različitih razina obrazovanja). Podaci su ovako enkodirani kako bi se olakšalo njihovo korištenje s algoritmima strojnog učenja.

Zato ćemo za te numeričke značajke pogledati zastupljenost pojedinih kategorija, zajedno uz kategorijske značajke (za koje nismo mogli dobiti gornji sažetak):

```
# prije provjere zastupljenosti potrebno je faktorizirati kategorijske značajke
df %>% select(-all_of(true_numeric)) -> df_no_numeric

# Kreiranje tablice za prikaz
results <- sapply(names(df_no_numeric), function(n) {
  tab <- table(df[[n]])
  percent <- round(prop.table(tab)*100,2)
  paste0(names(tab), " -- ", tab, " (", percent, "%)",  collapse = " | ")
})

# Prikaz rezultata kao tablice
data.frame(Stupac = names(results), Vrijednosti = results, row.names = NULL)

# Output skriven iz PDF-a radi preglednosti
```

Iz ovih podataka vidimo da prevladavaju učenici iz škole GP (331 naspram 39), raspodjela po spolovima je podjednaka, prevladavaju učenici iz urbanog područja i iz obitelji s više od 3 člana. Što se tiče nekih zanimljivijih značajki, iz podataka vidimo da većina učenika tjedno uči između 2 i 5 sati (*studytime* - 50%), većina mora putovati manje od 15 minuta do škole (*traveltime* - 65.41%), omjer učenika koji uzimaju instrukcije iz matematike i onih koji ne uzimaju instrukcije je podjednak (*paid_mat*), dok samo 6.76% učenika uzima instrukcije iz jezika (*paid_por*). Postotak učenika koji nemaju podršku obitelji u njihovom obrazovanju je 37.57% (*famsup*). Što se tiče konzumacije alkohola, radnim danima je očekivano niska (*Dalc* - gotovo 70% ne konzumira ili konzumira vrlo malo), a vikendom je povećana (*Walc*).

Prije obavljanja bilo kakve analize i provođenja statističkih testova nužno je očistiti i transformirati skup podataka s kojim se radi, ako je to potrebno. Ovdje je prvenstveno fokus na nedostajuće podatke te na stršeće podatke. Skup podataka koji proučavamo u sklopu projekta nema nedostajućih podataka (za svih 370 učenika imamo informacije o svih 39 značajki), a stršeće podatke smo provjerili za značajke *absences_mat* i *absences_por* i pronašli smo jedan podatak koji će se izuzeti iz analize ako to bude potrenbo (učenik sa 75 izostanaka na matematici).

Dio deskriptivne statistike je i vizualizacija podataka - ona je odličan način za dobivanje uvida u prirodu podataka i može pomoći naslutiti na odgovor za neka ključna pitanja: “koji je oblik distribucije podataka?”,

"koje su središnje tendencije?", "koji su rasponi i varijabilnost podataka?", "postoje li stršeće vrijednosti?", itd. U ovom dijelu detaljnu vizualizaciju ispuštamo. Vizualizacija korištenih značajki bit će napravljena u uvodu svakog statističkog testa.

3. Istraživačka pitanja

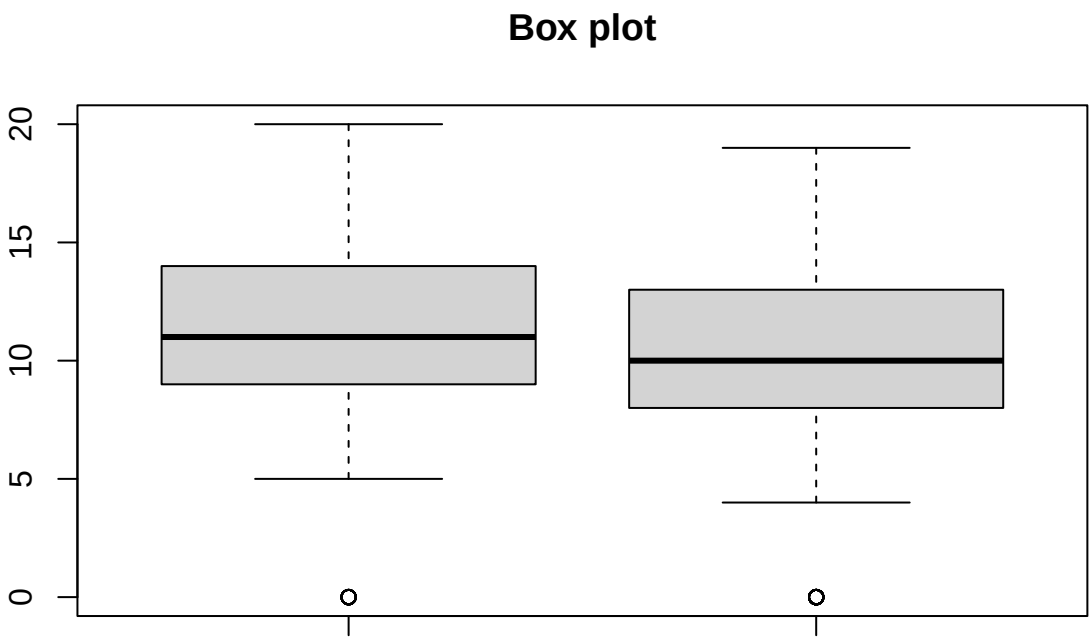
3.1. Jesu li prosječne ocjene iz matematike različite između spolova?

```
data_frame <- read.csv("student_data.csv", header = TRUE)

male_data <- subset(data_frame, sex == "M")

male_data <- male_data$G3_mat


female_data <- subset(data_frame, sex == "F")
female_data <- female_data$G3_mat
```

Kako bismo odgovorili na pitaje postoji li razlika u konačnim ocjenama među spolovima prvo moramo odraditi deskriptivnu statistiku.

Za početak napravimo box plotove za ocjene ovih dvaju populacija.

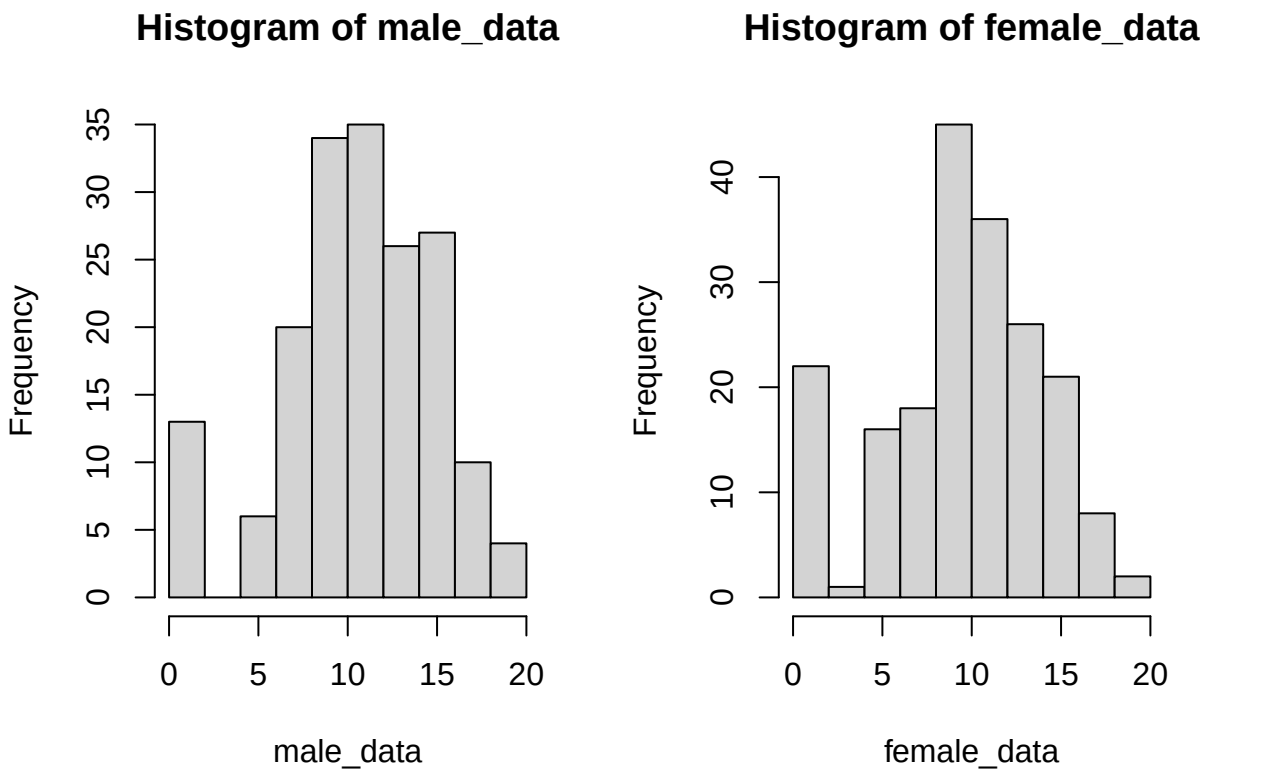
```
boxplot(male_data, female_data, main="Box plot")
```



Sa boxplota vidimo da postoji mala razlika u medianima uzoraka. Dodatno zanimljivo je uočiti da za obje populacije boxplot je skoro jednak samo je jedna od populacija translatirana. To nas navodi na hipotezu da su distribucije ovih populacija iste smo ne poravnate.

Pogledajmo jesu li podatci normalni. Za početak pogledajmo histograme.

```
par(mfrow=c(1,2))

hist(male_data)

hist(female_data)
```

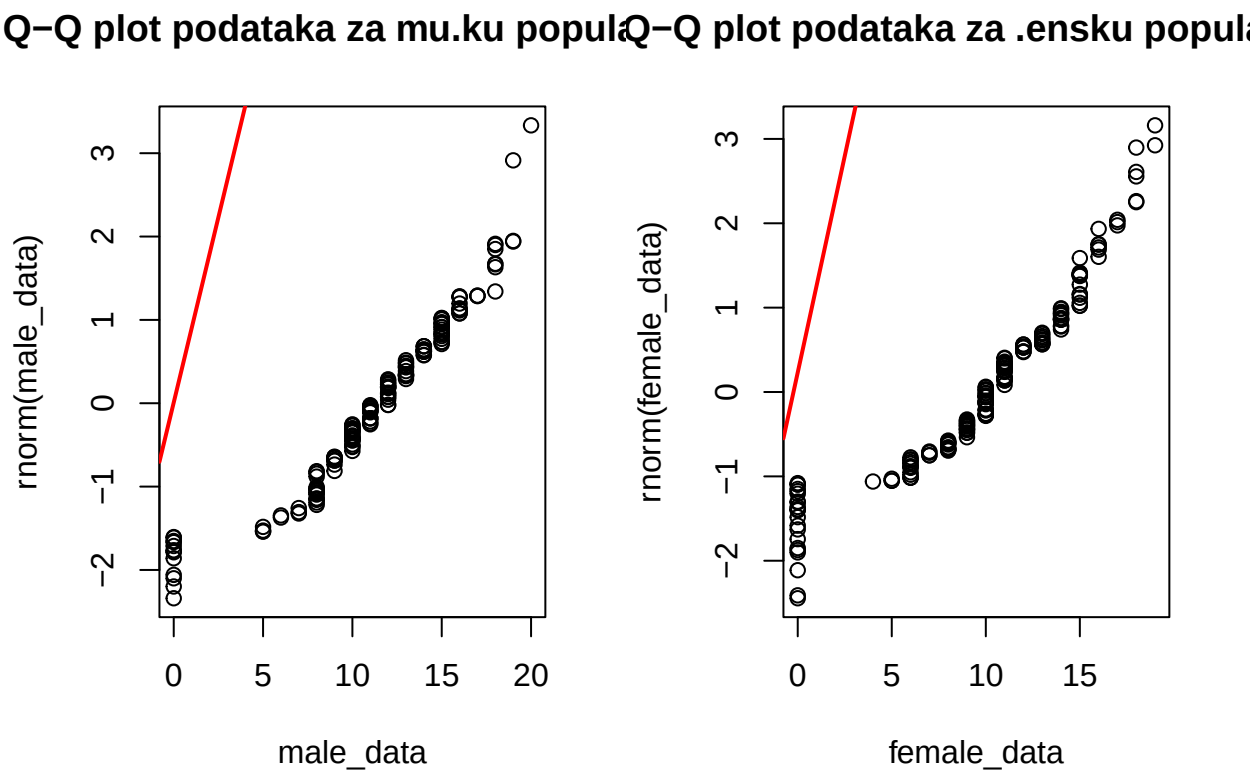



```
par(mfrow=c(1,1))
```

S histograma je jasno da podatci nisu normalni stoga na njih u ovakvom obliku nećemo moći primjeniti parametarske testove koji pretpostavljaju normalnost populacije.

To potvrđuju i Q-Q plotovi ovih uzoraka.

```
par(mfrow=c(1,2))

qqplot(male_data,rnorm(male_data), main="Q-Q plot podataka za mušku populaciju")
qqline(rnorm(male_data), col="red", lwd=2)

qqplot(female_data,rnorm(female_data), main="Q-Q plot podataka za žensku populaciju")
qqline(rnorm(female_data), col="red", lwd=2)
```

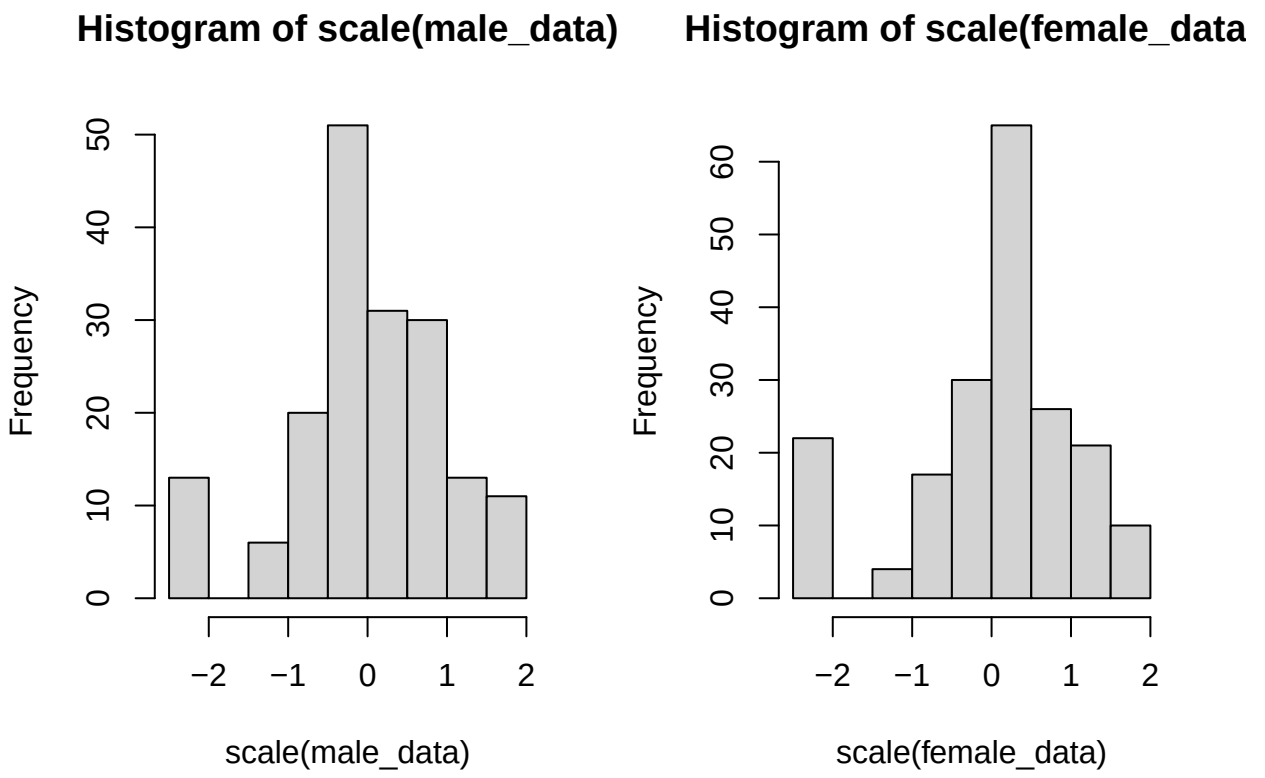



```
par(mfrow=c(1,1))
```

Q-Q plotovi definitivno potvrđuju da podatci nijednog od uzoraka nisu normalni.

S obzirom da podatci nisu normalni, prvo ćemo se poslužiti nekim neparametarskim postupkom. Nakon toga pokušati ćemo izmijeniti uzorke (izbacivanjem ekstrema) i pokušati primjeniti neki parametarski postupak. Idealno bi bilo koristiti Kolmogorov-Smirnovljev test, međutim zbog diskretnosti podataka on nije prikladan, iako bi on izravno dao odgovor na postavljeno pitanje. Umjesto toga upotrijebit Mann-Whitney U test koji će nam reći postoji li statistički zančajna razlika u medijanima dviju distribucija sličnog oblika. Budući da iz histograma vidimo da su distribucije sličnog oblika on U test je ovdje prikladan. Dodatno normalizacijom podataka možemo vidjeti da su histogrami sličniji ako standardiziramo podatke, što će smanjiti vjerojatnost pogreške

```
par(mfrow=c(1,2))
hist(scale(male_data))
hist(scale(female_data))
```




```
par(mfrow=c(1,1))
```

Konačno provedimo test

```
wilcox.test(scale(male_data), scale(female_data))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  scale(male_data) and scale(female_data)
## W = 16207, p-value = 0.4046
## alternative hypothesis: true location shift is not equal to 0
```

Iz p vrijednosti 0.406 možemo zaključiti da ne postoji statistički značajna razlika u završnim ocjenama iz matematike između splova. Zanimljivo je međutim uočiti da U test nad ne standardiziranim podatcima daje drugačije rezultate.

```
wilcox.test(male_data, female_data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  male_data and female_data
## W = 19719, p-value = 0.009412
## alternative hypothesis: true location shift is not equal to 0
```

S ovom novom p vrijednosti zaključili bismo da postoji značaja razlika u ocjenama. No budući da je oblik distribucije sličniji za normalizirane podatke to je rezultat kojeg odabiremo.

3.2. Postoji li razlika u prvoj ocjeni iz matematike s obzirom na mjesto stanovanja?

Pregled podataka

```
student_data <- read.csv('student_data.csv')
head(student_data)
```

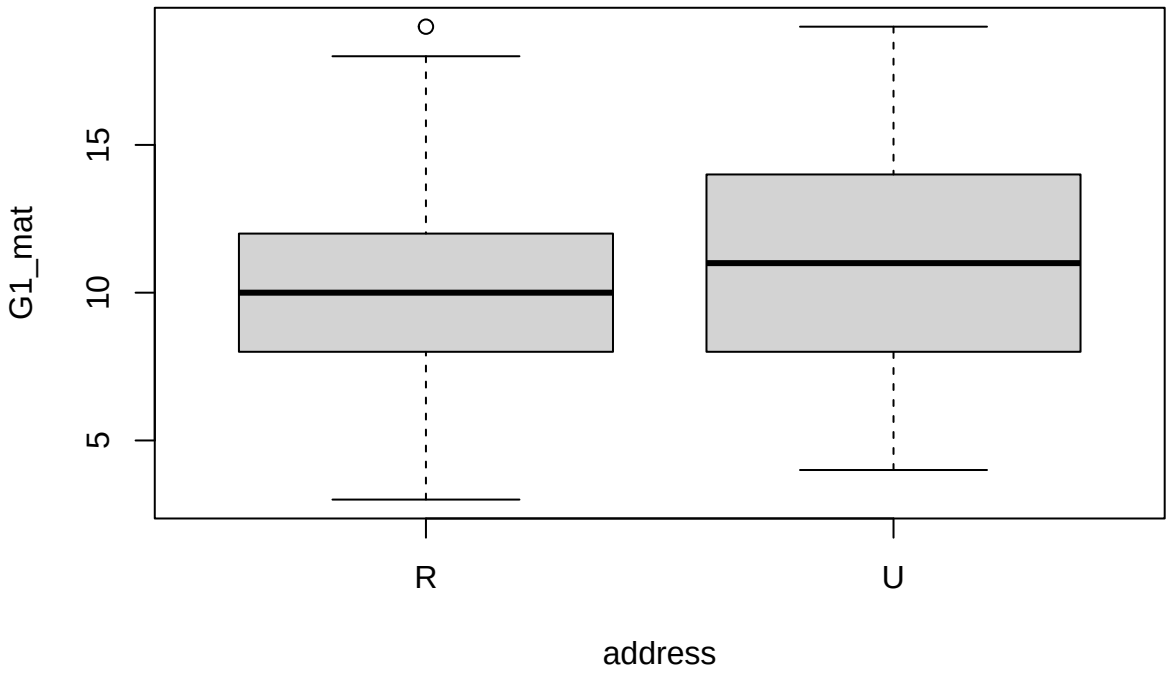
```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures_mat failures_por schoolsup famsup
## 1   mother          2         2            0            0       yes     no
## 2   father          1         2            0            0        no    yes
## 3   mother          1         2            3            0       yes     no
## 4   mother          1         3            0            0        no    yes
## 5   father          1         2            0            0        no    yes
## 6   mother          1         2            0            0        no    yes
##   paid_mat paid_por activities nursery higher internet romantic famrel freetime
## 1       no       no         no     yes    yes       no       no      4        3
## 2       no       no         no      no    yes      yes       no      5        3
## 3      yes       no         no     yes    yes      yes       no      4        3
## 4      yes       no        yes     yes    yes      yes      yes      3        2
## 5      yes       no         no     yes    yes       no       no      4        3
## 6      yes       no        yes     yes    yes      yes       no      5        4
##   goout Dalc Walc health absences_mat absences_por G1_mat G2_mat G3_mat G1_por
## 1     4    1    1      3            6            4      5      6      6      0
## 2     3    1    1      3            4            2      5      5      6      9
## 3     2    2    3      3           10            6      7      8     10     12
## 4     2    1    1      5            2            0     15     14     15     14
## 5     2    1    2      5            4            0      6     10     10     11
## 6     2    1    2      5           10            6     15     15     15     12
##   G2_por G3_por
## 1     11     11
## 2     11     11
## 3     13     12
## 4     14     14
## 5     13     13
## 6     12     13
```

```
columns <- c('G1_mat', 'address')
grades <- student_data[columns]
```

Vizualizacija

Box plot prve ocjene iz matematike s obzirom na mjesto stanovanja.

```
grades$address <- factor(grades$address)

plot(G1_mat ~ address, grades)
```



Uz pomoć našeg box plot-a možemo očekivati da neće biti razlike u prvoj ocjeni iz matematike s obzirom na mjesto stanovanja. No kako bismo to statistički zaključili, provodimo dva različita testa, hi-kvadrat test nezavisnosti/homogenosti podataka, te ANOVA-u.

hi-kvadrat test nezavisnosti/homogenosti podataka

Kako bismo mogli primjeniti hi-kvadrat test nezavisnosti/homogenosti podataka, sve očekivane frekvencije moraju imati vrijednost veću ili jednaku 5. Iz tog razloga prvo provjeravamo broj očekivanih vrijednosti manjih od 5. Ako jedna ili više takvih vrijednosti postoji, ne možemo primjeniti ovaj test, te moramo razmatrati alternativu.

```
group <- grades$address
g1_mat <- grades$G1_mat

contingency_table <- table(group, g1_mat)

N <- length(group)
num <- table(group)
num_R <- num['R']
num_U <- num['U']

counter <- 0

for (i in 1:ncol(contingency_table)) {
  value <- 0
  for (j in 1:nrow(contingency_table)) {
    value <- value + contingency_table[j, i]
  }

  expected_R <- N * (value / N) * (num_R / N)
  expected_U <- N * (value / N) * (num_U / N)
```

```
  if (expected_R < 5) {
    counter <- counter + 1
  }

  if (expected_U < 5) {
    counter <- counter + 1
  }
}

print(paste("Postoji ", counter, " očekivanih vrijednosti s vrijednošću manjom od 5, te ne možemo proves
```

```
## [1] "Postoji  12  očekivanih vrijednosti s vrijednošću manjom od 5, te ne možemo provesti hi-kvadrat
```

ANOVA

Prvi korak prije samog provođenja ANOVA-e je provođenje Bartlett-ovog testa nad podacima kako bismo testirali homogenost varijanci uzoraka.

Želimo testirati:
H0: sve su varijance jednake
H1: barem dvije varijance se razlikuju
alpha = 0.05

```
# Testiranje homogenosti varijanci uzoraka
# Bartlettov test

bartlett_test_result <- bartlett.test(G1_mat ~ address, grades)

print(bartlett_test_result)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  G1_mat by address
## Bartlett's K-squared = 0.14685, df = 1, p-value = 0.7016
```

Dobivši p-vrijednost 0.7016, tj. značajno veću vrijednost od pretpostavljenje razine značajnosti (0.05), možemo zaključiti da su varijance jednake te da nad ovim podacima možemo provesti ANOVA-u.

Prije same ANOVA-e također promatramo koliko različitih uzoraka imamo za svaku kategoriju (U - urban te R - rural).

```
table(grades$address, useNA = 'always')
```

```
##
##    R    U <NA>
##   81  289    0
```

Želimo testirati:
H0: sve su srednje vrijednosti jednake
H1: barem dvije srednje vrijednosti se razlikuju
alpha = 0.05

```
anova_result <- aov(G1_mat ~ address, grades)

summary(anova_result)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## address       1     20   19.63   1.758  0.186
## Residuals   368   4110   11.17
```

p-vrijednost našeg ANOVA testa je zadana kao Pr(>F), te iznosi $0.186 > 0.05$, te iz tog razloga ne možemo odbaciti nultu hipotezu. Koristeći ANOVA test, statistički zaključujemo da prva ocjena iz matematike ne ovisi o mjestu stanovanja.

3.3. Možemo li predvidjeti prolaz iz završnog ispita iz jezika na temelju sociodemografskih varijabli poput spola, obrazovanja roditelja i veličine obitelji?

3.4. Postoji li razlika u broju izostanaka iz matematike između učenika koji dolaze iz manjih i većih obitelji?

3.5. ...

3.6. ...

4. Zaključak