Analiza bodova studenata u nastavnim aktivnostima

Marko Rajnović, Gordana Borotić, Lucija Bago, Leon Bertol

18.01.2021.

Pretprocesiranje dataseta je jako važno kako bismo na točan i efektivan način mogli analizirati dataset. To uključuje uklanjanje beskorisnih stupaca, Zamjenjivanje nedostajućih vrijednosti u datasetu smislenim vrijednostima i spajanje stupaca gdje to ima smisla.

#PREIMENOVANJE I SPAJANJE DATASETOVA U JEDAN JEDINSTVENI, UČITAVANJE POTREBNIH BIBILIOTEKA

```
knitr::opts_chunk$set(out.height="300px", dpi=120)

library(readxl)
library(tidyverse)
library(dplyr)
library(Rmisc)
library(ggplot2)
library(rapportools)

SAP_2016 <- read_excel("Dataset/SAP 2016L clean.xlsx")
SAP_2017 <- read_excel("Dataset/SAP 2017L clean.xlsx")
SAP_2018 <- read_excel("Dataset/SAP 2018L clean.xlsx")

#dodavanje stupca "godina" u sve datasetove
SAP_2016["godina"] <- "2016"
SAP_2017["godina"] <- "2017"
SAP_2018["godina"] <- "2018"

SAP <- rbind(SAP_2016, SAP_2017, SAP_2018)

colnames(SAP)

#preimenovanje varijabli u nešto što se može lakše oblikovati
SAP <- SAP %>%
  dplyr::rename(
    student = "Student",
    isvu_bodovi = "ISVU Bodovi",
    isvu_ocjena = "ISVU Ocjena",
    isvu_rok = "ISVU Rok",
    kont_bodovi = "Kontinuirana nastava - bodovi",
    kont_prolaz = "Kontinuirana nastava - prolaz",
    projekt_bodovi = "Seminari/projekti - bodovi",
    mi_bodovi = "Međuispit - bodovi",
    zi_bodovi = "Završni ispit: Pismeni - bodovi",
    prvi_rok_bodovi = "1. ispitni rok - bodovi",
    prvi_rok_prolaz = "1. ispitni rok - prolaz",
```

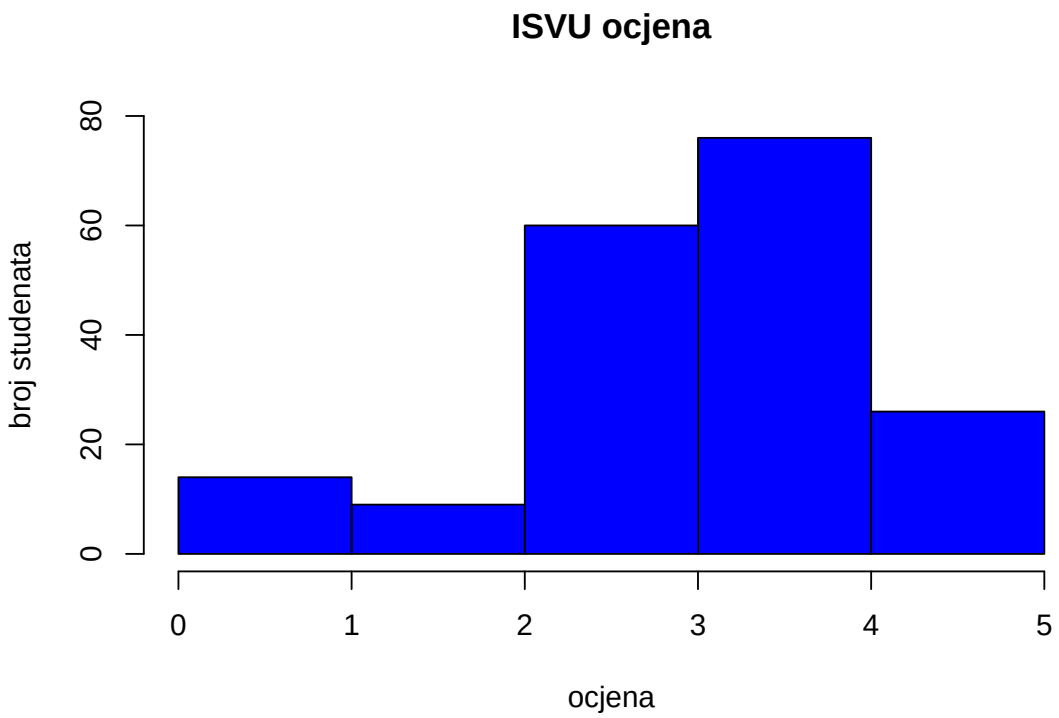
```
    prvi_rok_grupa = "1. ispitni rok - grupa",
    prvi_rok_korisnik = "1. ispitni rok - korisnik",
    prvi_rok_vrijeme = "1. ispitni rok - vrijeme",
    pis_ispit_bodovi = "Pismeni ispit - bodovi",
    pis_ispit_uvjet = "Pismeni ispit - uvjet",
    pis_ispit_grupa = "Pismeni ispit - grupa",
    pis_ispit_vrijeme = "Pismeni ispit - vrijeme",
    drugi_rok_bodovi = "2. ispitni rok - bodovi",
    drugi_rok_prolaz = "2. ispitni rok - prolaz",
    drugi_rok_grupa = "2. ispitni rok - grupa",
    drugi_rok_vrijeme = "2. ispitni rok - vrijeme",
    dek_rok_bodovi = "Dekanski rok - bodovi",
    dek_rok_prolaz = "Dekanski rok - prolaz",
    dek_rok_grupa = "Dekanski rok - grupa",
    dek_rok_vrijeme = "Dekanski rok - vrijeme"
  )
```

#UKLANJANJE NEPOTREBNIH STUPACA, SPAJANJE ONIH KOJE IMA SMISLA SPOJITI I ZAMJENJIVANJE NULL VRIJEDNOSTI SMISLENIMA

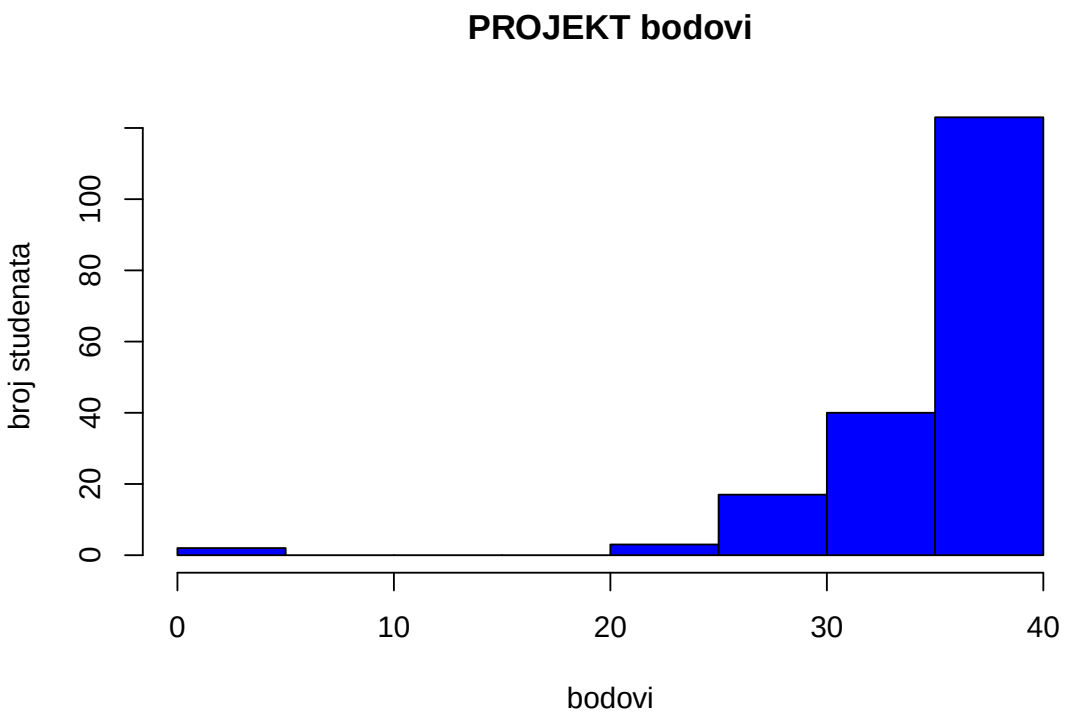
```
#Ako ne piše ništa u kont. prolazu, student nije prošao
SAP$kont_prolaz[is.na(SAP$kont_prolaz)] <- "NE"

#Stupac prvi_rok_grupa je nepotreban, jer sve informacije možemo staviti u prvi_rok_prolaz

#Najprije premještamo vrijednost "Položio ranije"
SAP$prvi_rok_prolaz[SAP$prvi_rok_grupa == "Položio ranije"] <- "Položio ranije"

#Onda premještamo vrijednost "Nije prijavljen""
SAP$prvi_rok_prolaz[SAP$prvi_rok_grupa == "Nije prijavljen"] <- "Nije prijavljen"

#Svi ostali nisu prošli
SAP$prvi_rok_prolaz[is.na(SAP$prvi_rok_prolaz)] <- "NE"

#Uklanjamo stupac prvi_rok_grupa
SAP$prvi_rok_grupa <- NULL

#prvi_rok_korisnik i svi pis_ispit stupci su beskorisni
SAP$prvi_rok_korisnik <- NULL
SAP$pis_ispit_bodovi <- NULL
SAP$pis_ispit_uvjet <- NULL
SAP$pis_ispit_grupa <- NULL
SAP$pis_ispit_vrijeme <- NULL

#Za ostale rokove radimo ono što smo napravili za prvi rok
SAP$drugi_rok_prolaz[SAP$drugi_rok_grupa == "Položio ranije"] <- "Položio ranije"
SAP$drugi_rok_prolaz[SAP$drugi_rok_grupa == "Nije prijavljen"] <- "Nije prijavljen"
SAP$drugi_rok_prolaz[is.na(SAP$drugi_rok_prolaz)] <- "NE"
SAP$drugi_rok_grupa <- NULL

SAP$dek_rok_prolaz[SAP$dek_rok_grupa == "Položio ranije"] <- "Položio ranije"
SAP$dek_rok_prolaz[SAP$dek_rok_grupa == "Nije prijavljen"] <- "Nije prijavljen"
SAP$dek_rok_prolaz[is.na(SAP$dek_rok_prolaz)] <- "NE"
SAP$dek_rok_grupa <- NULL
```

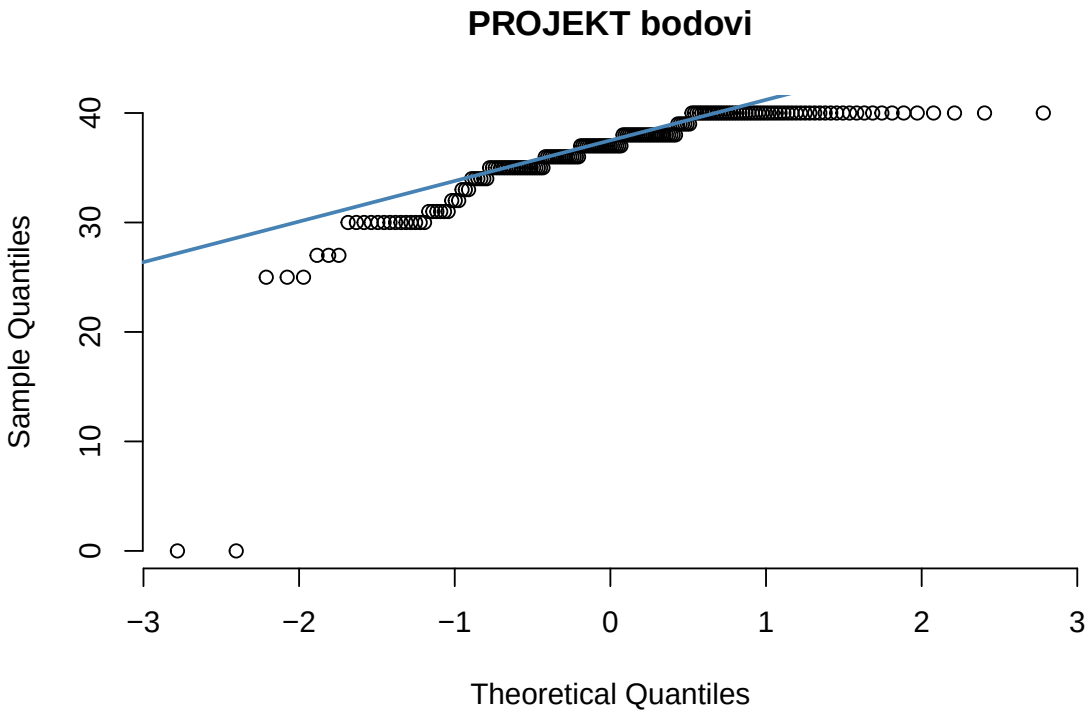
```r
#isvu_ocjena ako je na, onda je 1
SAP$isvu_ocjena[is.na(SAP$isvu_ocjena)] <- 1
```

Najprije, da dobijemo predodžbu naših podataka, pogledajmo histograme naših varijabli. Interpretacija podataka je značajan dio obrade podataka.



Raspored ocjena se čini dobar, većina se nalazi u sredini s manjim ekstremima. Za razdiobu možemo pretpostaviti da je približno normalna, pomalo lijevo zakrenuta, s više boljih ocjena. Ovakav smo raspored naviknuli vidjeti na većini predmeta. Kao i kod ostalih varijabli pretpostavka normalnosti će igrati ključnu ulogu u idućim razmatranjima i testovima.

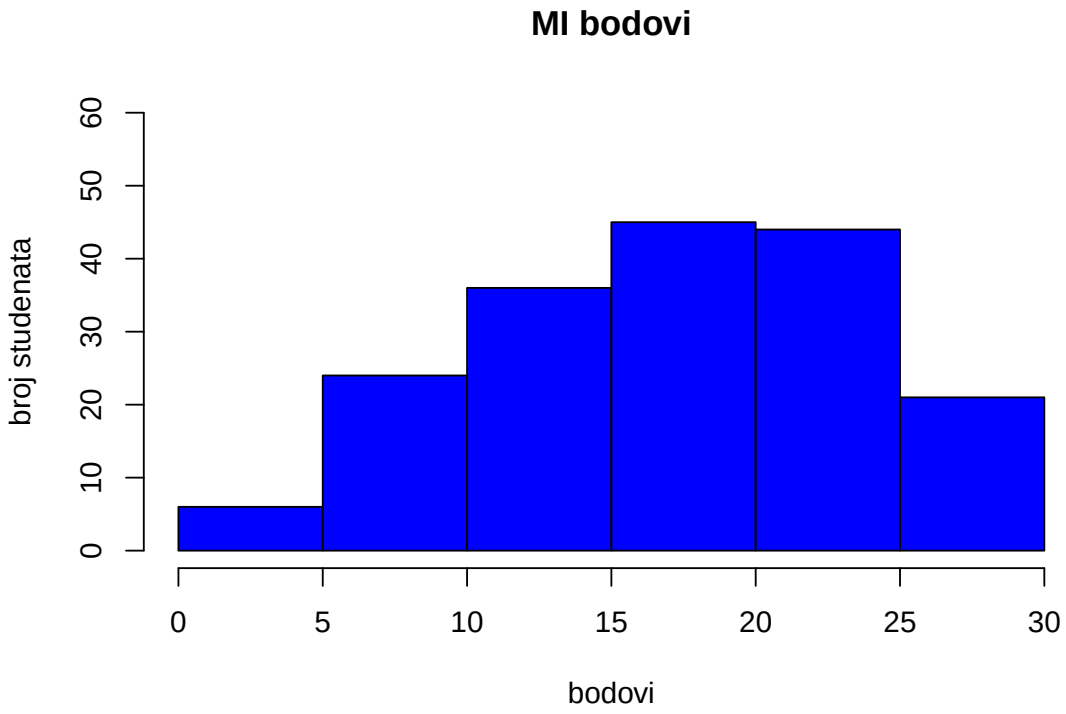


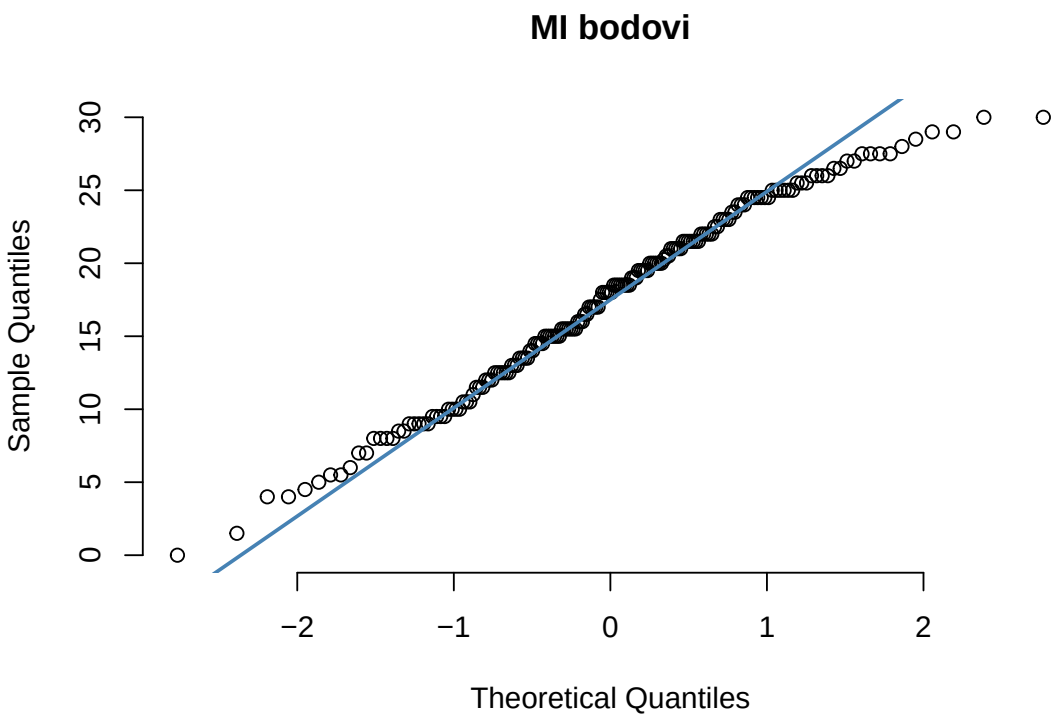



```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  SAP$projekt_bodovi
## W = 0.65725, p-value < 2.2e-16
```

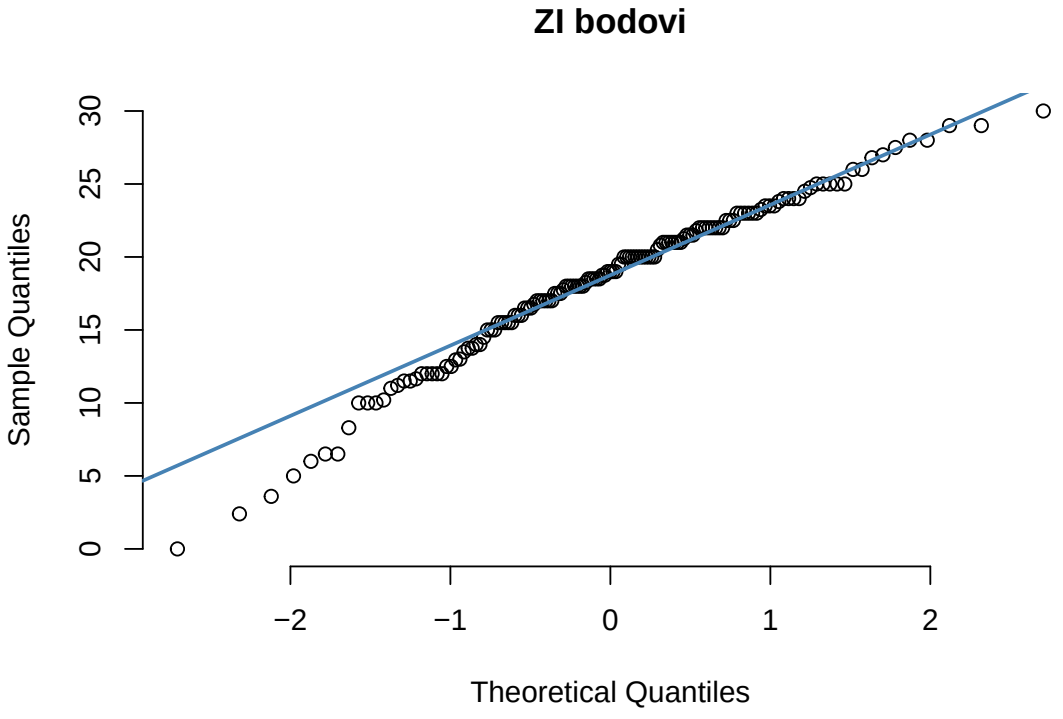
Vidimo da većina učenika jako dobro napiše projekt. Imaju odlične rezultate, a jako mali broj zapravo ne sudjeluje u pisanju projekta. Promatrajući qq plot i histogram možemo jasno vidjeti da se bodovi ne ravnaju po normalnoj razdiobi. Vidimo kako Shapiro-Wilk test daje jako malu p vrijednost, što znači da možemo s velikom pouzdanošću reći da se ova varijabla ne ravna po normalnoj distribuciji. Osim učenika koji imaju 0 bodova i za koje pretpostavljamo da nisu sudjelovali u pisanju projekta, vidimo da svi ostali zadovoljavaju prag od 20 bodova.






```
##
##  Shapiro-Wilk normality test
##
## data:  SAP$mi_bodovi
## W = 0.9824, p-value = 0.02538
```

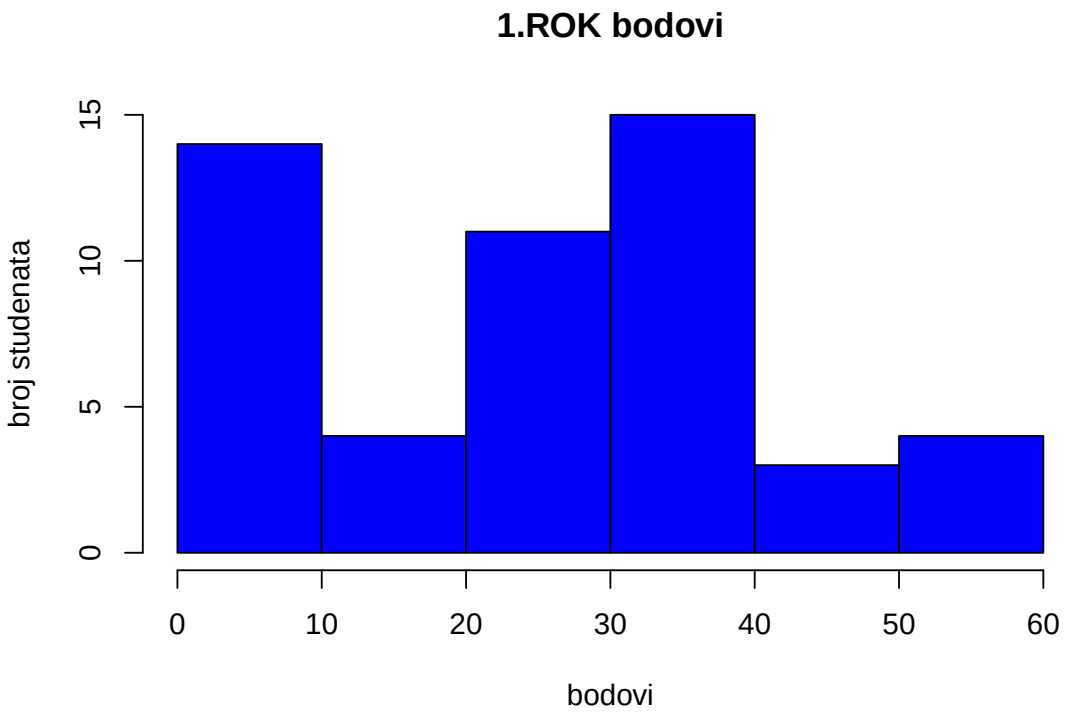
Vidimo da se bodovi s međuispita ravnaju približno normalno.

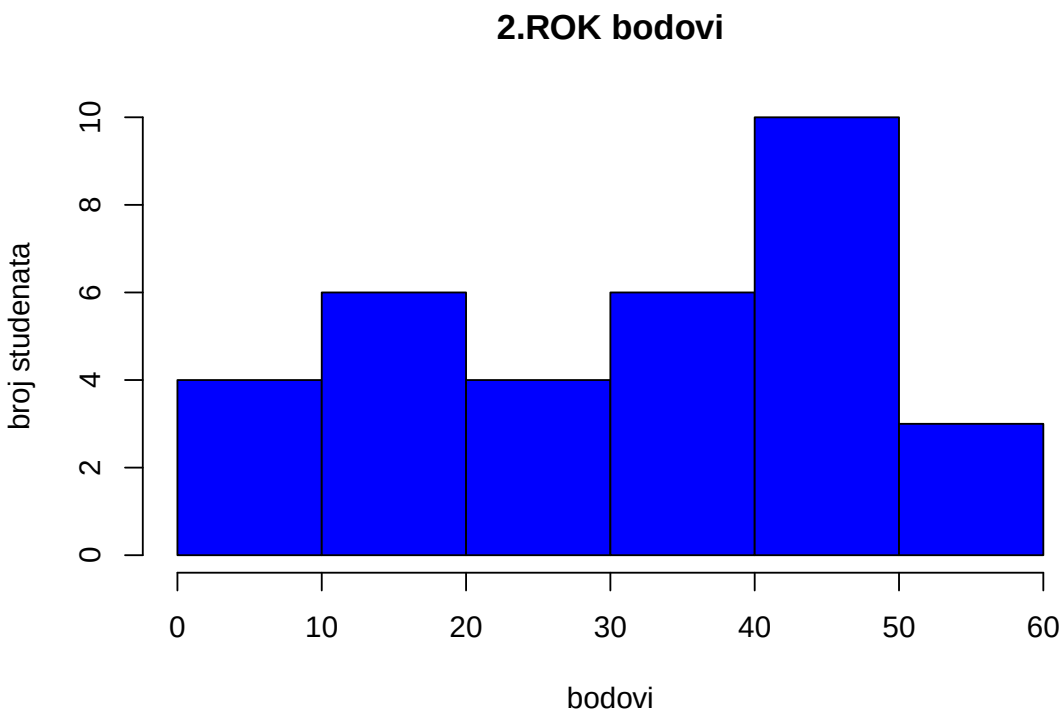


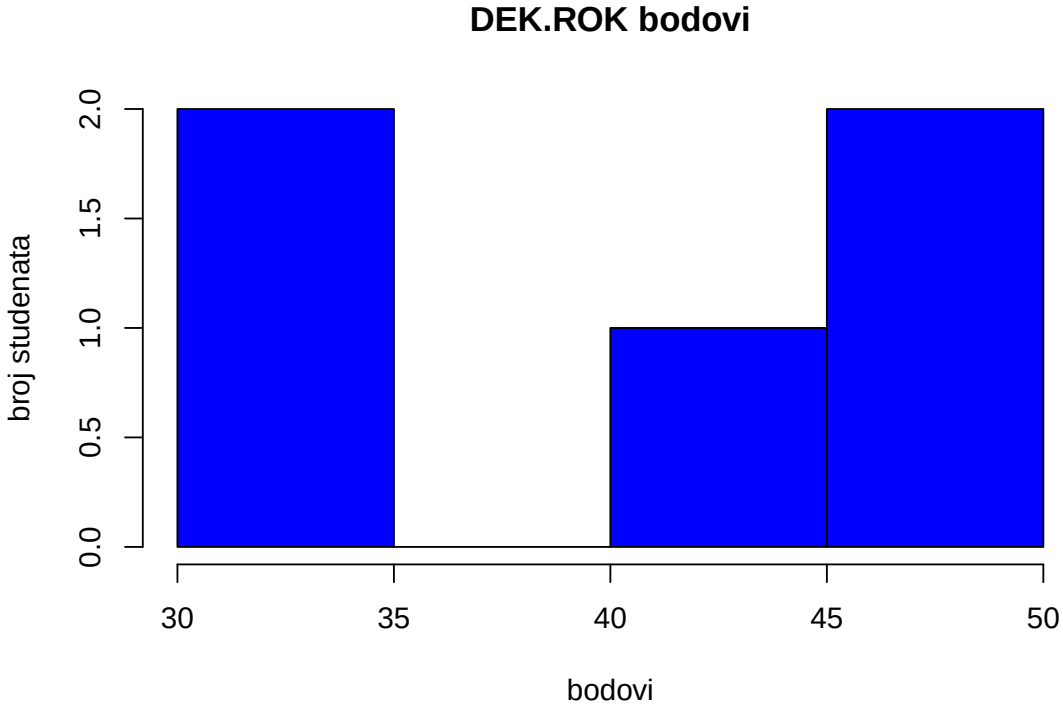

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  SAP$zi_bodovi
## W = 0.97166, p-value = 0.003871
```

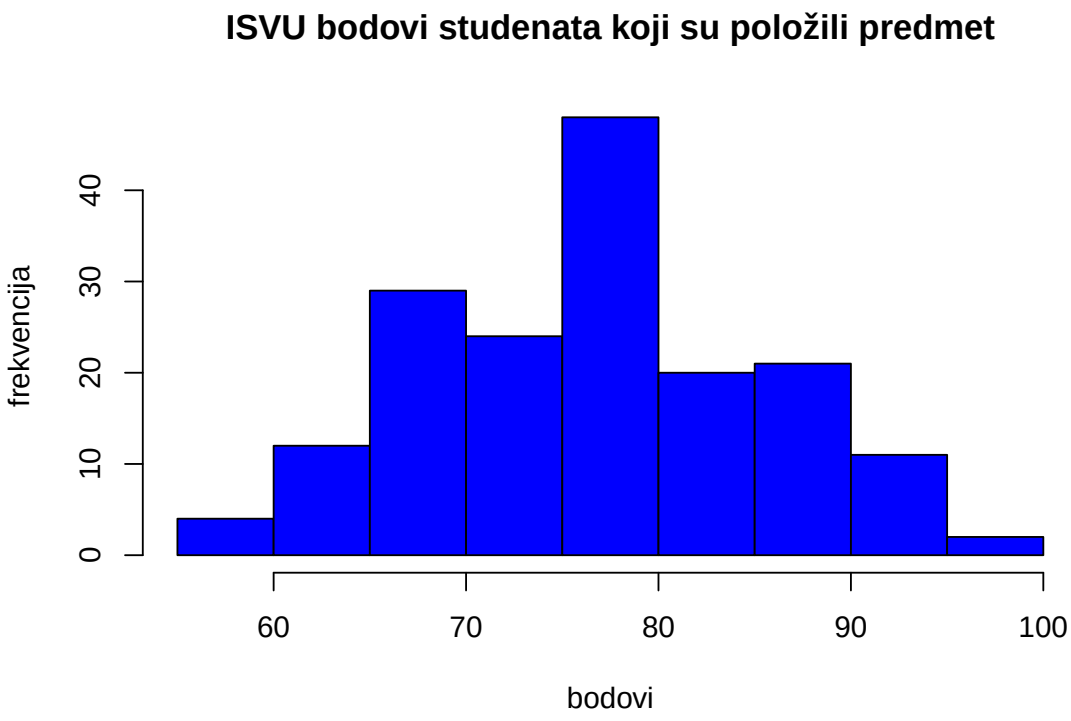
Vidimo da se bodovi s završnog ravnaju približno normalno. Shapiro-Wilk test sa završnog ispita je lošiji nego za međuispit. To možemo možda pripisati tome da su neki učenici jako loše napisali međuispit, pa se nisu trudili na završnome ili pak su predobro napisali međuispit pa su si osigurali prolaz s manje učenja. Ili im znanje iz prvog ciklusa nije sjelo pa su jako loše napisali i drugi ciklus. U nastavku ćemo mi ipak pretpostavljati normalnost ove varijable zbog T testa, jer je on robustan na normalnost.

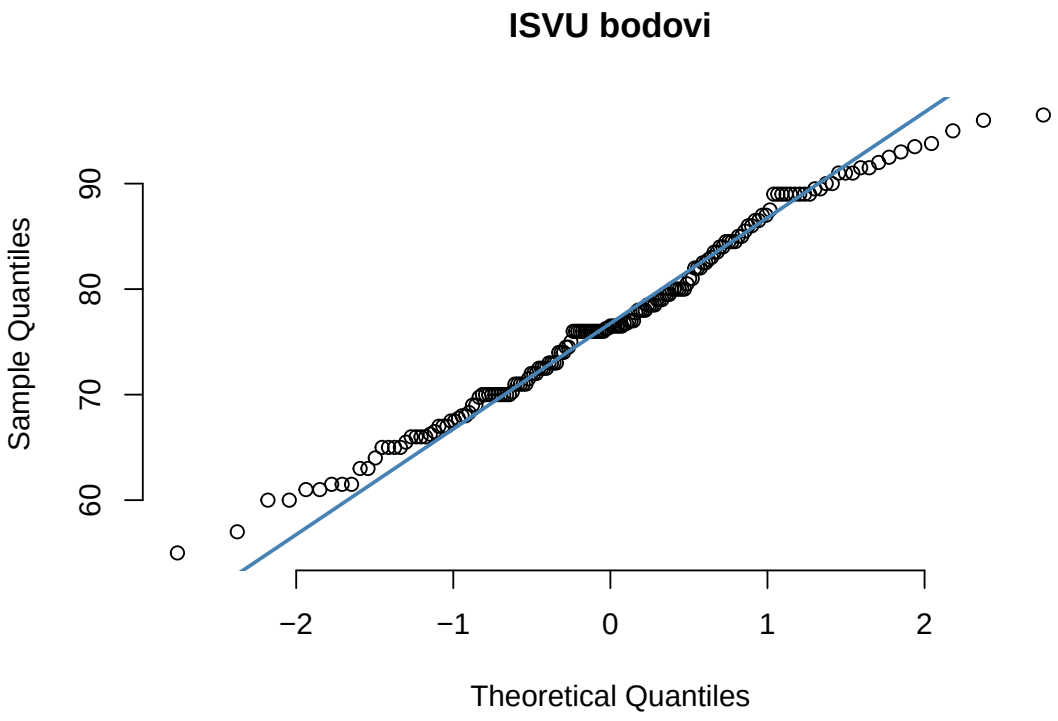






Možemo vidjeti da izgleda kao da je drugi rok bolje napisan od prvog roka, što ćemo kasnije i dokazati. Kod 1.roka distribuciju kvare učenici koji se nisu potrudili. Možemo vidjeti da je na dekanski rok izašlo malo ljudi te stoga na njemu nećemo raditi analize jer imamo jako malo podataka.



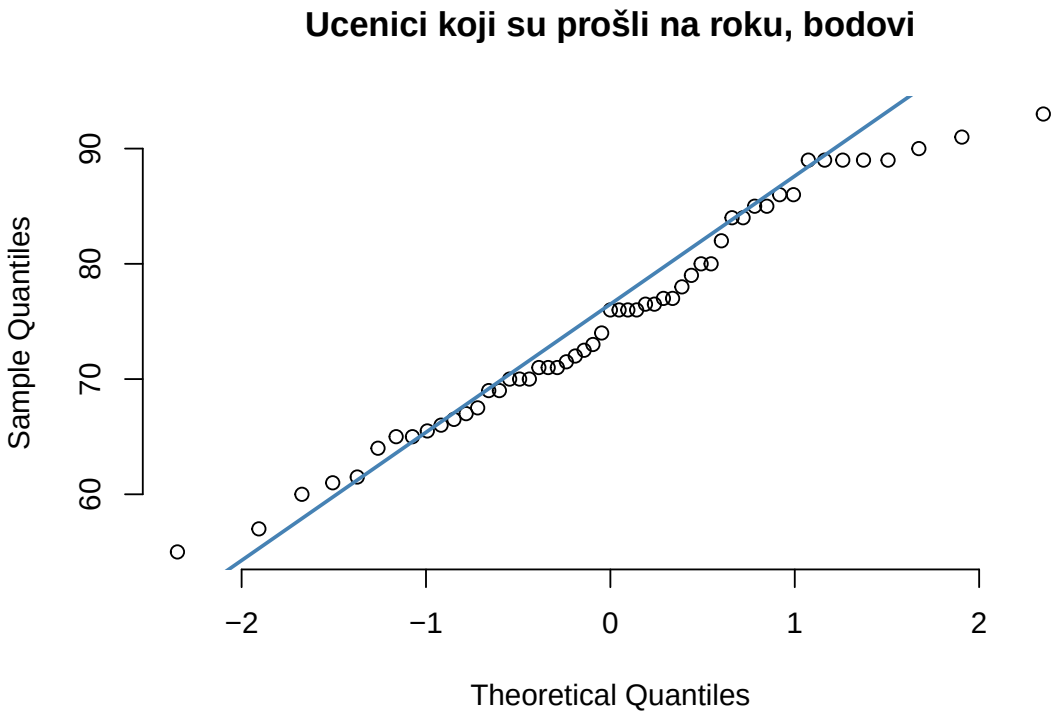



```
##
##  Shapiro-Wilk normality test
##
```

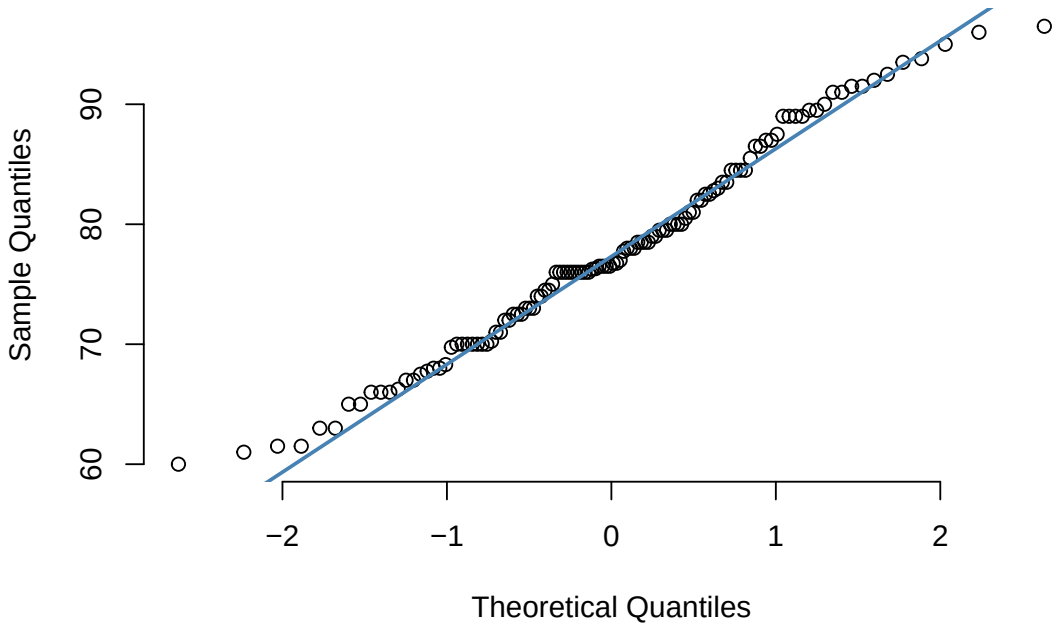
```
## data:  SAP$isvu_bodovi
## W = 0.98673, p-value = 0.1059
```

Pomoću histograma i qq-plota vidimo kako se ISVU bodovi jako dobro ravnaju po normalnoj distribuciji.

Pogledajmo sada malo dublje neke druge značajke. Recimo, one koji su prošli na roku i one koji su prošli kontinuirano. Ravnaju li se i oni po normalnoj distribuciji?




```
##
##  Shapiro-Wilk normality test
##
## data:  prosli_na_roku$isvu_bodovi
## W = 0.97306, p-value = 0.2726
```



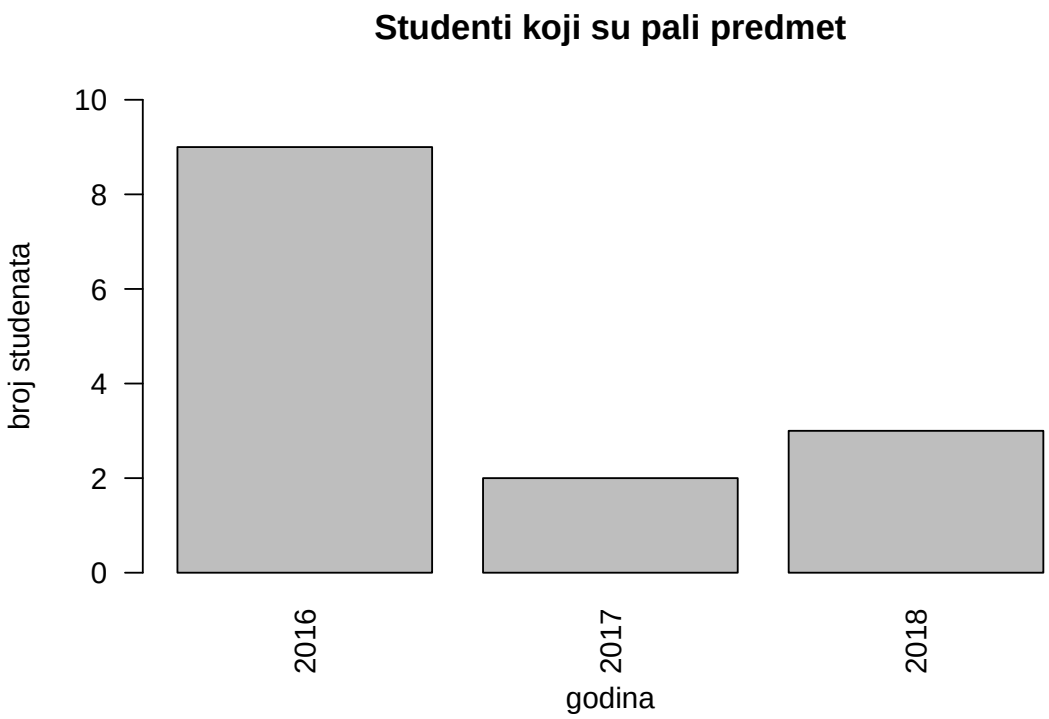

```
##
##  Shapiro-Wilk normality test
##
## data:  prosli_kontinuirano$isvu_bodovi
## W = 0.98308, p-value = 0.1448
```

Shapiro-Wilk test nam kaže da možemo pretpostaviti normalnost u oba slučaja.

Nakon što smo pregledali osnovne varijable i njihove distribucije sada možemo donositi zaključke na određena pitanja koja smo uočili pregledavajući podatke.

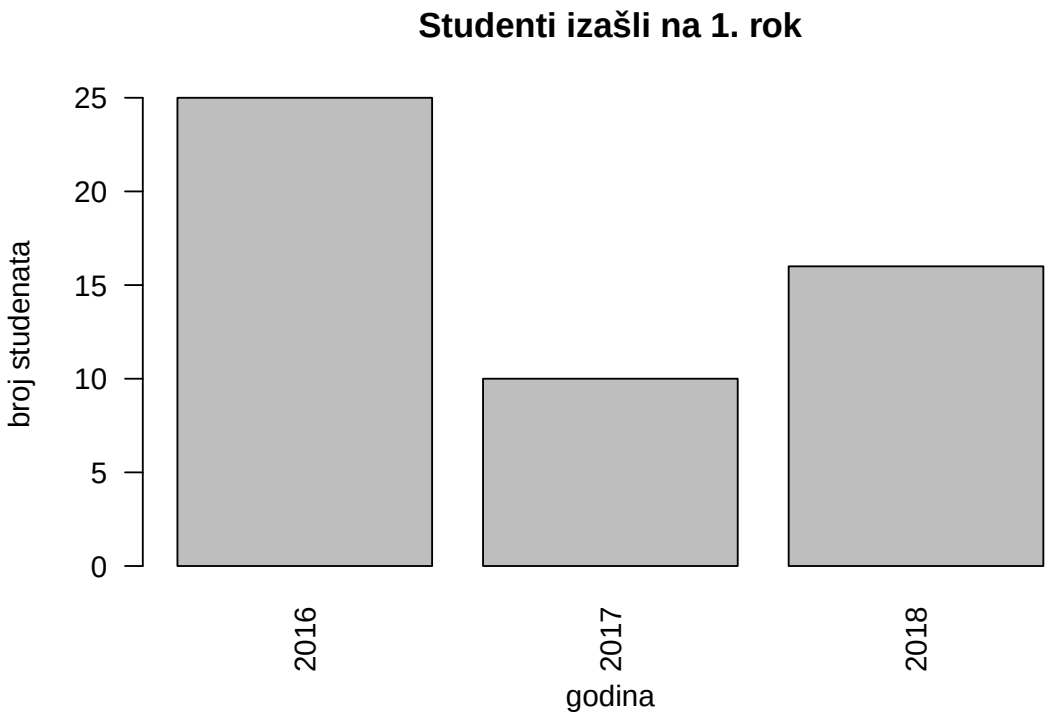
PITANJE: Koliko je studenata palo predmet svake godine?

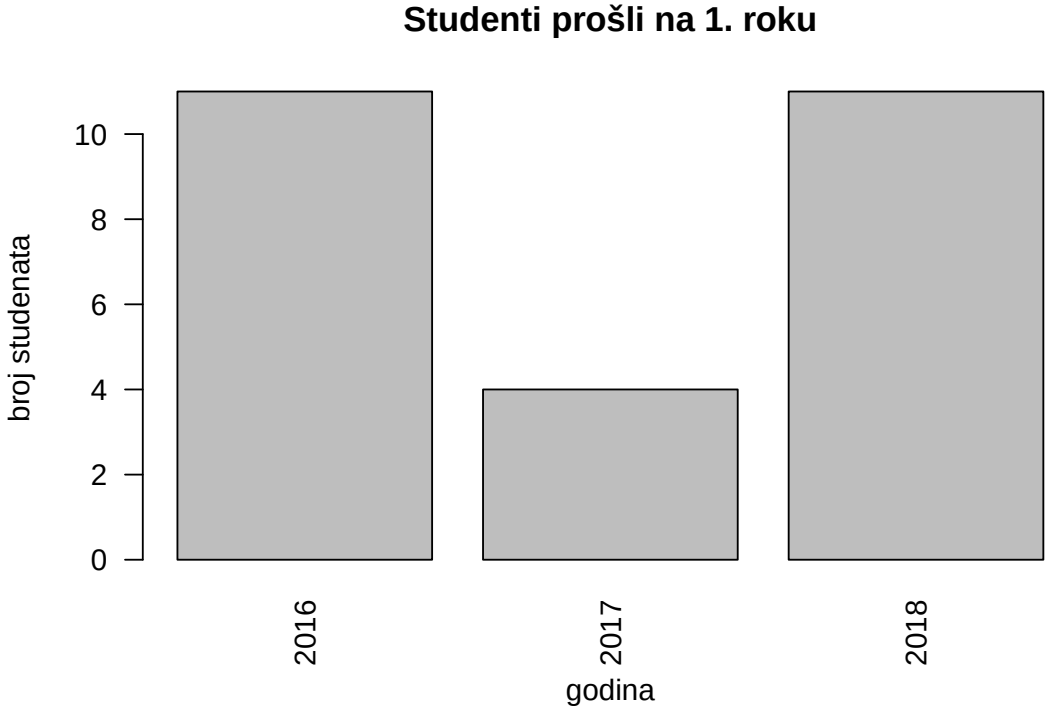
```
studenti_pali_predmet <- SAP[SAP$isvu_ocjena == 1 , ]

barplot(table(studenti_pali_predmet$godina),las=2,main='Studenti koji su pali predmet',
xlab="godina", ylab="broj studenata", ylim = c(0, 10))
```

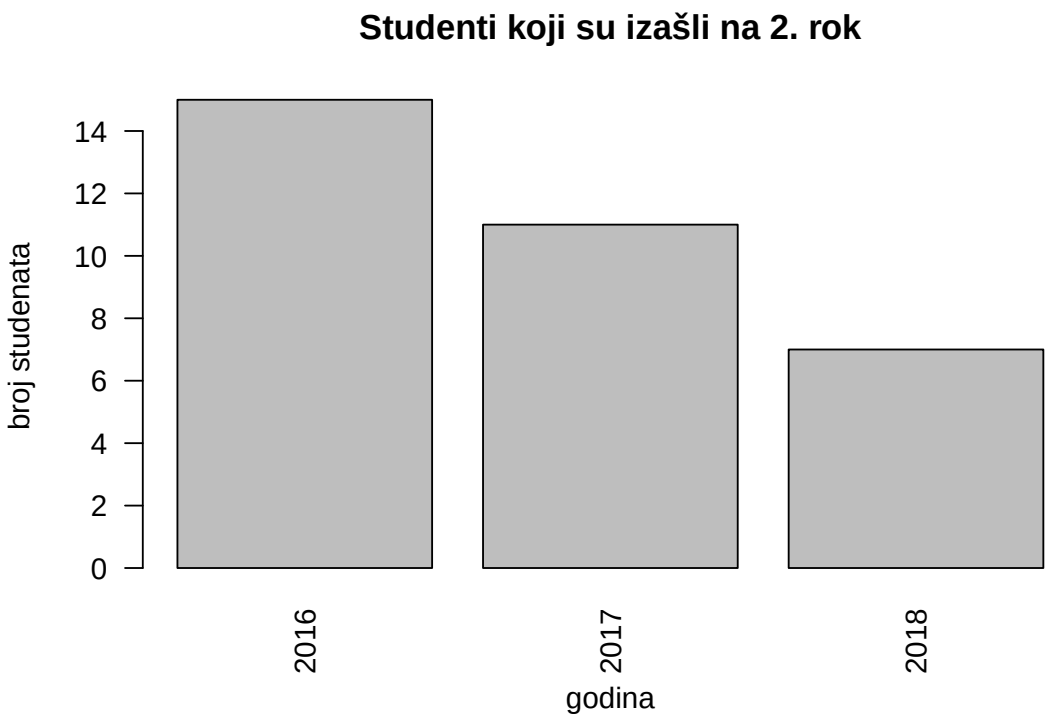


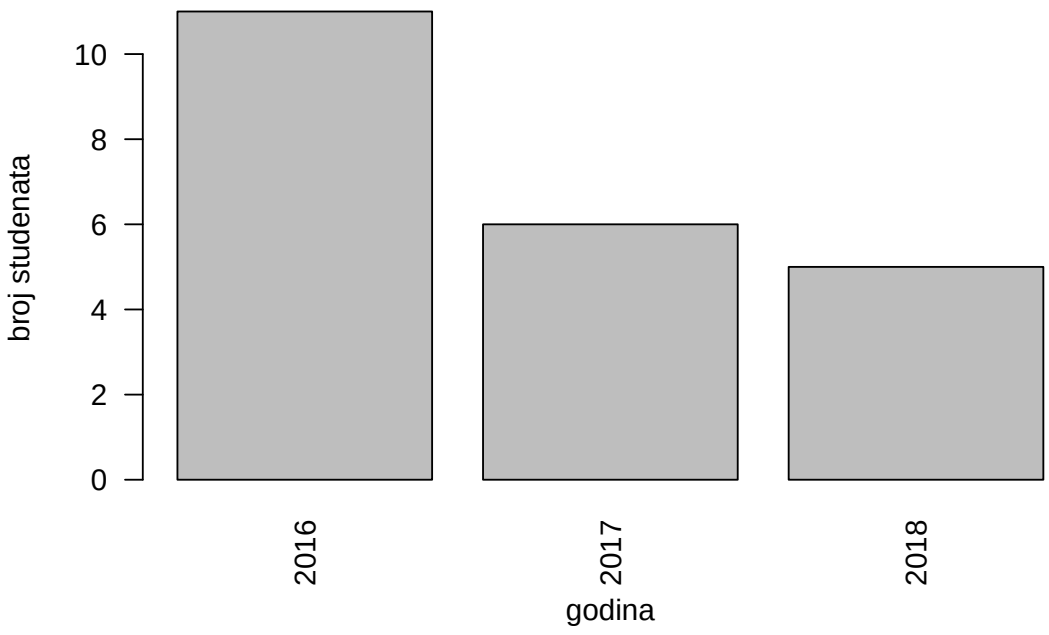
Možemo primijetiti da je broj studenata koji su pali predmet sličan u 2017. i 2018. godini (razlika za 1), a u 2016. godini više odstupa(za otprilike 2/3 više). Razlog tome mogu biti složeniji i teži ispiti u 2016.godini ili jednostavno "lošija" generacija.

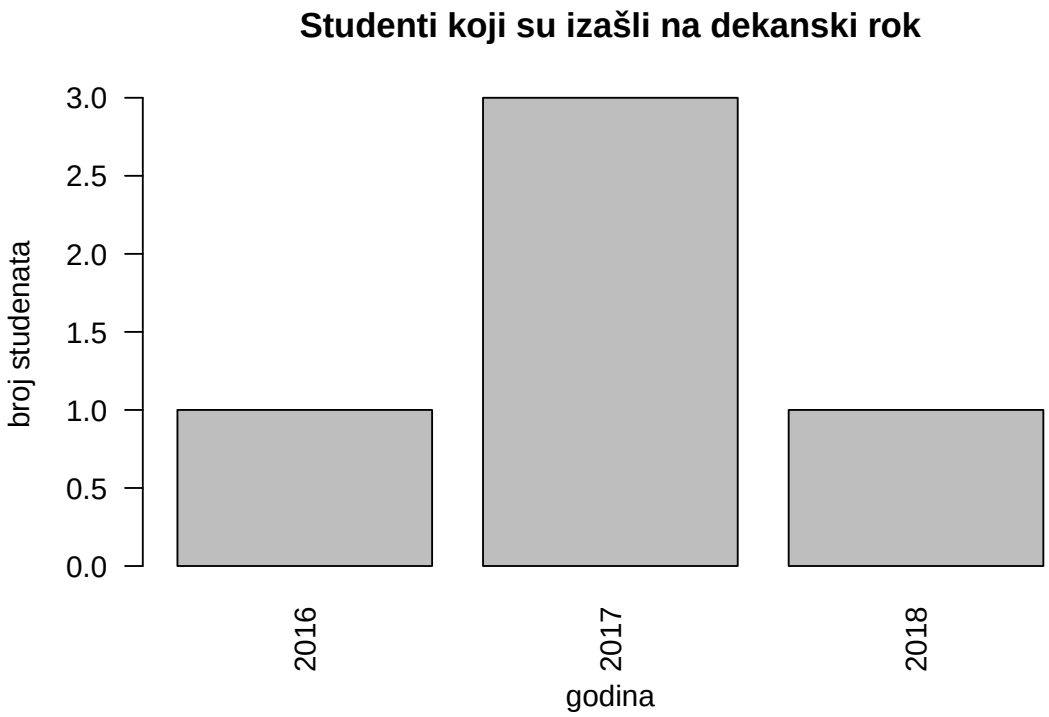
PITANJE: Koliko je učenika izašlo na svaki ispit i koliko ih je prošlo na svakom ispitu?

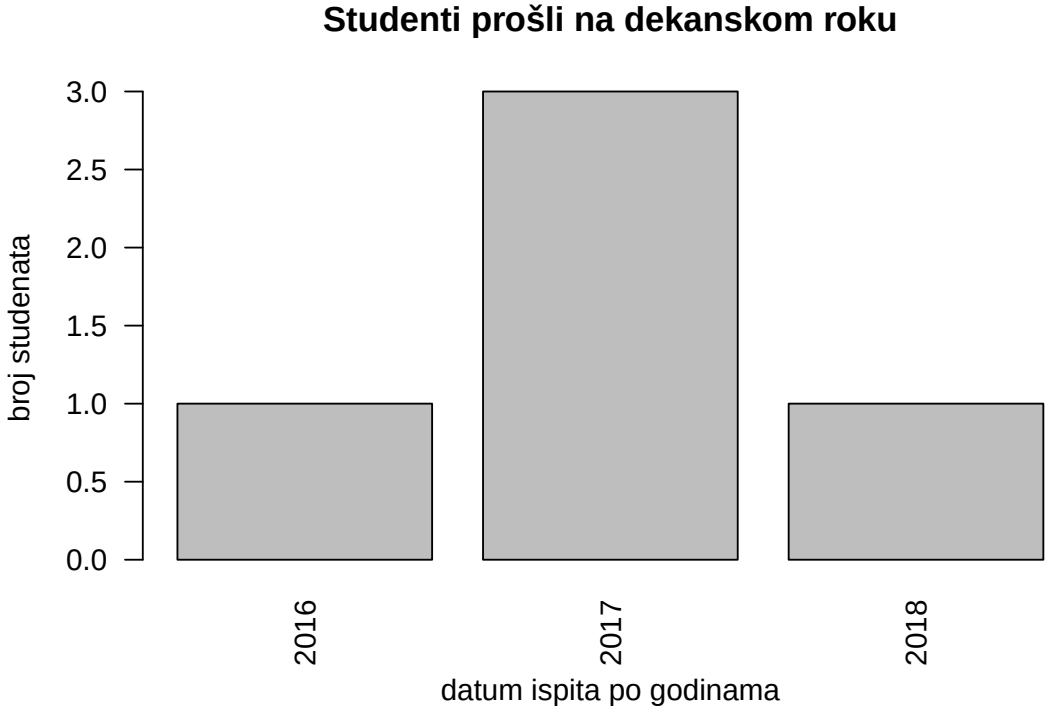

Studenti izašli na 1. rok
broj studenata
0
5
10
15
20
25
2016
2017
2018
godina


Studenti prošli na 1. roku
broj studenata
0
2
4
6
8
10
2016
2017
2018
godina


Studenti koji su izašli na 2. rok
broj studenata
0
2
4
6
8
10
12
14
2016
2017
2018
godina


Studenti prošli na 2. roku
broj studenata
0
2
4
6
8
10
2016
2017
2018
godina





Uspoređujući plotove za dekanski rok, izgleda da su u sve 3 godine svi studenti koji su izašli na rok, ujedno ga i položili. Naravno, to ne znači da su svi studenti svake godine položili predmet jer postoje oni koji nisu prijavili rokove. Na prvom i drugom roku uvijek postoji studenata koji nisu prošli.

PITANJE: Jesu li studenti koji su odbili ocjenu stečenu kontinuirano bolje napisali sljedeći ispit? Koje su ocjene studenti odbijali?

```
kont_odbijeni_bodovi <- SAP[(SAP$kont_bodovi >= "50") &
                              (SAP$mi_bodovi + SAP$zi_bodovi >= "30") &
                              (SAP$projekt_bodovi >= "20") & (SAP$kont_prolaz != "DA") &
                              (!is.na(SAP$mi_bodovi)) & (!is.na(SAP$zi_bodovi)) ,]$kont_bodovi


prvi_pokusaj_bodovi <- na.omit(SAP[(SAP$kont_bodovi >= "50") & (SAP$mi_bodovi + SAP$zi_bodovi >= "30") &

if(!is.empty(kont_odbijeni_bodovi)) {
  t.test(kont_odbijeni_bodovi, prvi_pokusaj_bodovi, paired = TRUE, alternative = "less")

  cat('Bodovi dobiveni kontinuirano koji su odbijeni: ', kont_odbijeni_bodovi,'\n')

  cat('Bodovi dobiveni na prvom sljedećem ispitu    : ', prvi_pokusaj_bodovi,'\n')

} else {
   cat('Nijedan student nije odbio ocjenu.')
}
```

```
## Nijedan student nije odbio ocjenu.
```

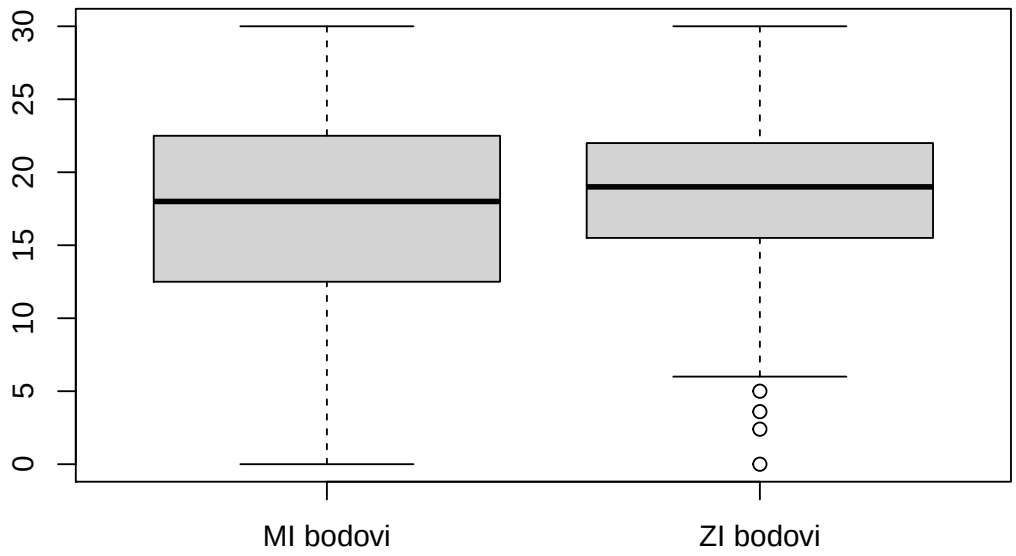
Kada smo pokušali selektirati studente koji su imali sve ispunjene uvjete za prolaz preko kontinuirane nastave, dakle 50% bodova u zbroju međuispita i završnog ispita te 50% bodova iz projekta, nismo pronašli niti jednog koji je odbio ocjenu.

Možemo reći da smo bili iznenađeni ovim ishodom, s obzirom da su u pitanju 3 godine za redom te smo očekivali nekoliko studenata koji su odbili ocjenu ostvarenu na kontinuiranom prolazu. Da su takvi studenti postojali, proveli bismo upareni t-test te na temelju njega zaključili nauče li studenti bolje za sljedeći rok.

PITANJE: Napisu li ucenici zi bolje od mi-a?.

Pogledajmo najprije box plotove.

```
boxplot(SAP$mi_bodovi, SAP$zi_bodovi,
names = c('MI bodovi','ZI bodovi'),
main='Pravokutni dijagram bodova ucenika na ZI i MI')
```



Kao što vidimo, razlika je vrlo mala. Ako pogledamo i aritmetičke sredine, one podupiru činjenicu da nema toliko značajne razlike. Na završnom ispitu možemo uočiti par stršečih vrijednosti.

```
mean(na.omit(SAP$mi_bodovi))
```

```
## [1] 17.47017
```

```
mean(na.omit(SAP$zi_bodovi))
```

```
## [1] 18.43878
```

Upareni t-test nam govori da nemamo dovoljno podataka da odbacimo hipotezu da su vrijednosti na MI manje ili jednake nego na ZI.

```
izasli_mi_zi <- SAP[!is.na(SAP$mi_bodovi) & (!is.na(SAP$zi_bodovi)),]

t.test(izasli_mi_zi$mi_bodovi, izasli_mi_zi$zi_bodovi, paired = TRUE, alternative = "greater")
```

```
##
##  Paired t-test
##
## data:  izasli_mi_zi$mi_bodovi and izasli_mi_zi$zi_bodovi
## t = 0.94667, df = 146, p-value = 0.1727
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.3399319        Inf
## sample estimates:
## mean of the differences
##               0.4540816
```

PITANJE: Imaju li učenici koji su položili preko kontinuirane nastave bolji prosjek bodova od učenika koji su položili preko rokova?

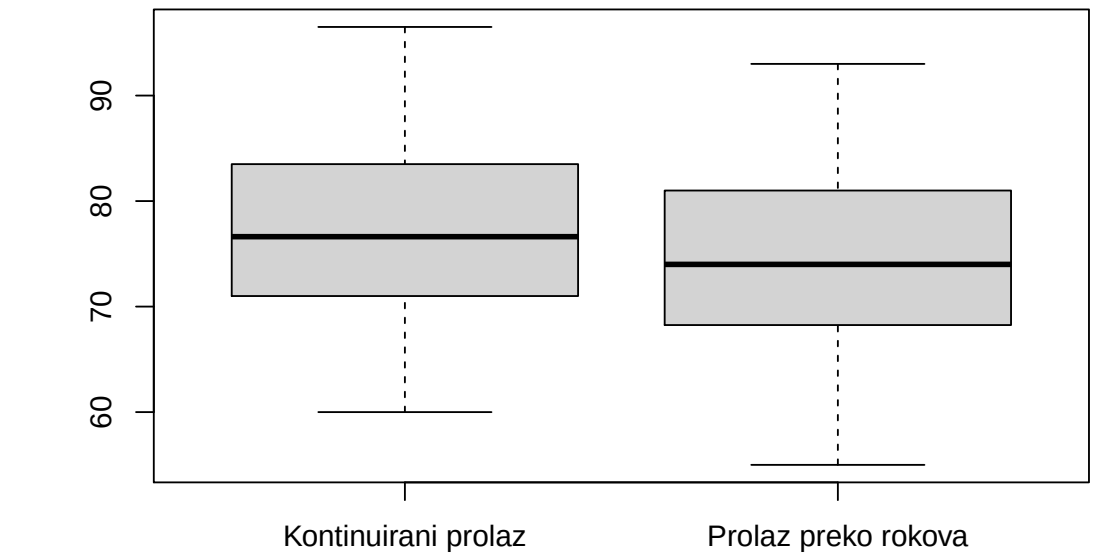
```
ucenici_kont = SAP[(SAP$kont_prolaz == "DA") & (SAP$isvu_bodovi==SAP$kont_bodovi),]
ucenici_rok = SAP[(SAP$kont_prolaz == "NE") & (SAP$isvu_ocjena>1),]


cat('Prosječan broj bodova učenika koji su prošli kontinuirano iznosi ', mean(ucenici_kont$kont_bodovi)
```

```
## Prosjecan broj bodova ucenika koji su prošli kontinuirano iznosi  77.67119
```

```
cat('Prosječan broj bodova učenika koji su prošli na rokovima iznosi ', mean(ucenici_rok$isvu_bodovi),
```

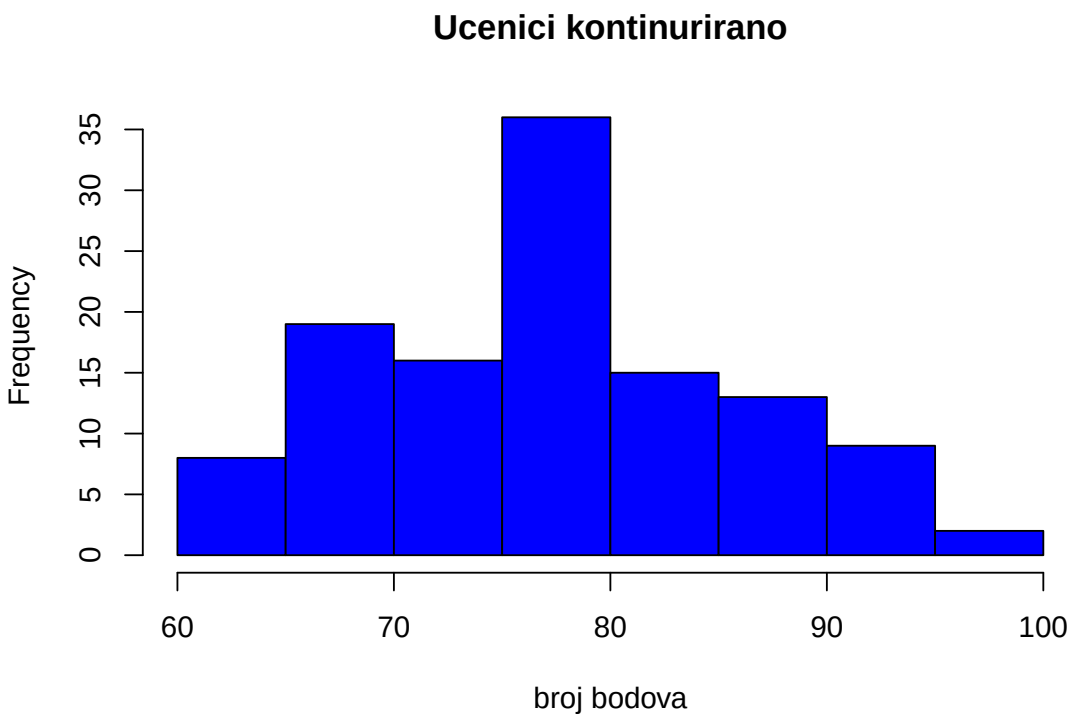
```
## Prosjecan broj bodova ucenika koji su prošli na rokovima iznosi  74.58824
```

```
boxplot(ucenici_kont$kont_bodovi, ucenici_rok$isvu_bodovi,
        names = c('Kontinuirani prolaz','Prolaz preko rokova'),
        main='Pravokutni dijagram bodova učenika koji su prošli kontinuirano i preko roka')
```

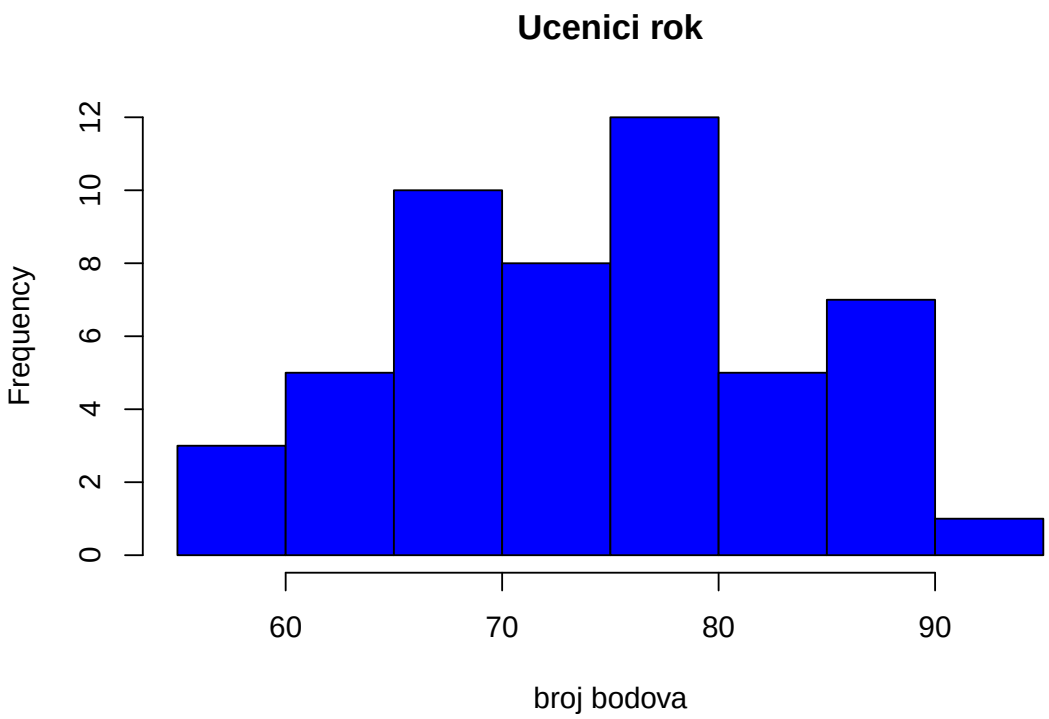


Promatrajući pravokutni dijagram i srednje vrijednosti ostvarenih bodova možemo vidjeti da postoji razlika u bodovima. Kako bi pokazali je li ta razlika statistički značajna koristit ćemo t-test. Kako bi mogli provesti test, moramo najprije provjeriti pretpostavke normalnosti i nezavisnosti uzorka. Pretpostavit ćemo nezavisnost uzoraka jer su promatrani različiti učenici. Za provjeru normalnosti podataka ćemo koristiti histgoram .

```
h = hist(ucenici_kont$kont_bodovi,
         main="Učenici kontinurirano",
         xlab="broj bodova",
         ylab='Frequency',
         col="blue"
         )
```




```
h = hist(ucenici_rok$isvu_bodovi,
         main="Učenici rok",
         xlab="broj bodova",
         ylab='Frequency',
         col="blue"
         )
```



Distribucija nije savršeno normalna, no približno je, a t-test je robustan na normalnost.

Sada ćemo promatrati varijance ova dva uzorka i testom o varijanci usporediti jesu li jednake.

```
var(ucenici_kont$kont_bodovi)
```

```
## [1] 73.4201
```

```
var(ucenici_rok$isvu_bodovi)
```

```
## [1] 86.16706
```

```
var.test(ucenici_kont$kont_bodovi, ucenici_rok$isvu_bodovi)
```

```
##
##  F test to compare two variances
##
## data:  ucenici_kont$kont_bodovi and ucenici_rok$isvu_bodovi
## F = 0.85207, num df = 117, denom df = 50, p-value = 0.4797
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5192832 1.3362964
## sample estimates:
## ratio of variances
##          0.8520669
```

Velika p-vrijednost od 0.4797 nam govori da nećemo odbaciti hipotezu H0 da su varijance naša dva uzorka jednake.

Provedimo sada t-test uz pretpostavku jednakosti varijanci.

```r
t.test(ucenici_kont$kont_bodovi, ucenici_rok$isvu_bodovi, alt = "greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  ucenici_kont$kont_bodovi and ucenici_rok$isvu_bodovi
## t = 2.0933, df = 167, p-value = 0.01892
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.6469815       Inf
## sample estimates:
## mean of x mean of y
##  77.67119  74.58824
```

Mala p-vrijednost nam znači da možemo odbaciti nultu hipotezu u korist hipoteze da učenici koji prođu kontinuirano imaju veći prosjek bodova od učenika koji prođu preko roka što je potvrdilo naše indikacije.

PITANJE: Prosjek broja bodova na prvom roku je manji od prosjeka bodova na drugom roku?

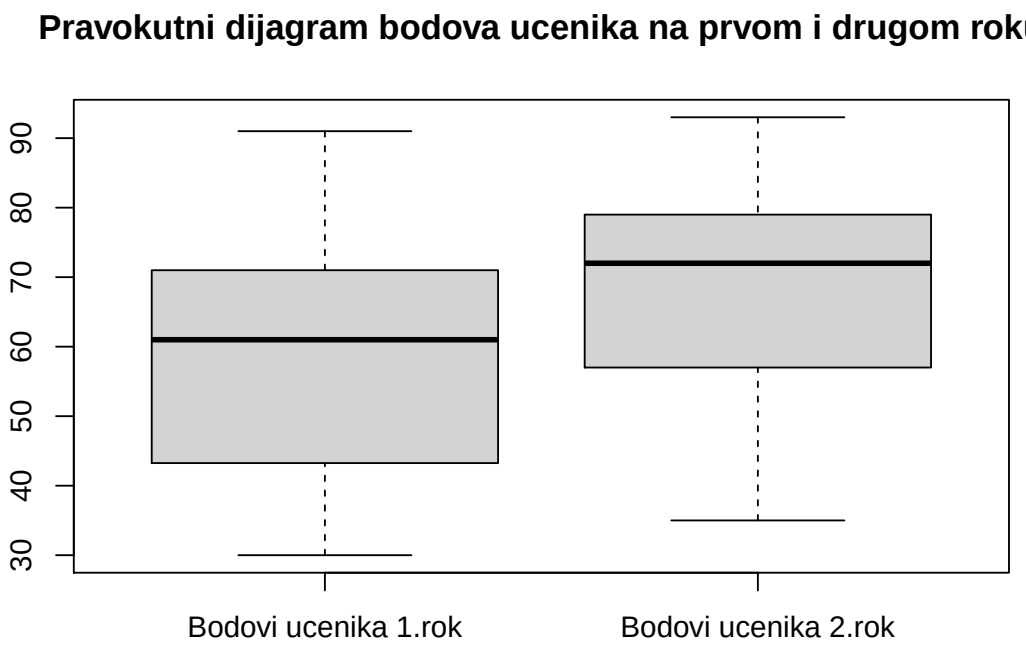
```r
ucenici_prvi_rok = SAP[(SAP$prvi_rok_prolaz != "Položio ranije") & (SAP$prvi_rok_prolaz != "Nije prijavl
ucenici_drugi_rok = SAP[(SAP$drugi_rok_prolaz != "Položio ranije") & (SAP$drugi_rok_prolaz != "Nije pri

cat('Prosječan broj bodova učenika na prvom roku ', mean(ucenici_prvi_rok$prvi_rok_bodovi),'\n')
```

```
## Prosjecan broj bodova ucenika na prvom roku  59.4902
```

```r
cat('Prosječan broj bodova učenika na drugom roku ', mean(ucenici_drugi_rok$drugi_rok_bodovi), '\n')
```

```
## Prosjecan broj bodova ucenika na drugom roku  67.57576
```

```r
boxplot(ucenici_prvi_rok$prvi_rok_bodovi, ucenici_drugi_rok$drugi_rok_bodovi,
        names = c('Bodovi učenika 1.rok','Bodovi učenika 2.rok'),
        main='Pravokutni dijagram bodova učenika na prvom i drugom roku')
```



Iz provedenih pravokutnih dijagrama i izračunatih srednjih vrijednosti možemo naslutiti da je uspjeh na 2.roku bolji. To je intuitivno jasno jer oni učenici koji su pali na prethodnom roku će više učiti za idući rok i imati veću motivaciju jer već su jedan rok pali. Tvrdnju ne možemo statistički potvrditi jer proatrani podaci na 1. i 2.roku nisu nezavisni jer neki učenici su išli na oba roka. Također, ne možemo tvrditi ni da su podaci upareni jer nisu svi učenici išli na oba roka. Promatrajući učenike koji su išli na oba roka jasno je da ćemo dobiti bolje rezultate na 2.roku upravo iz prethodno navedenih razloga što pokazuje i idući dijagram.

```
ucenici_oba_rok = SAP[(SAP$prvi_rok_prolaz != "Položio ranije") & (SAP$prvi_rok_prolaz != "Nije prijavl
```

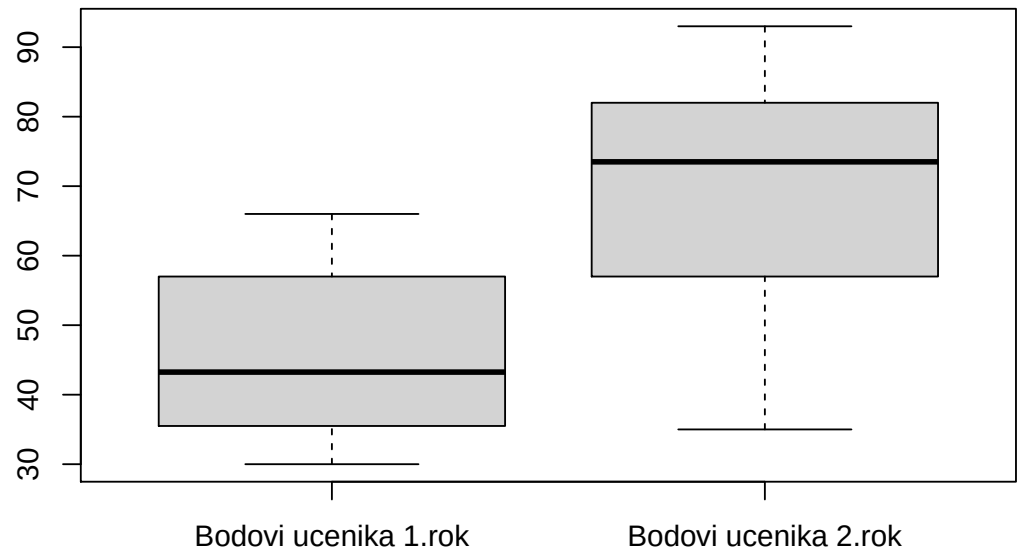
```
cat('Prosječan broj bodova učenika na prvom roku ', mean(ucenici_oba_rok$prvi_rok_bodovi),'\n')
```

```
## Prosjecan broj bodova ucenika na prvom roku  45.83333
```

```
cat('Prosječan broj bodova učenika na drugom roku ', mean(ucenici_oba_rok$drugi_rok_bodovi), '\n')
```

```
## Prosjecan broj bodova ucenika na drugom roku  69.20833
```

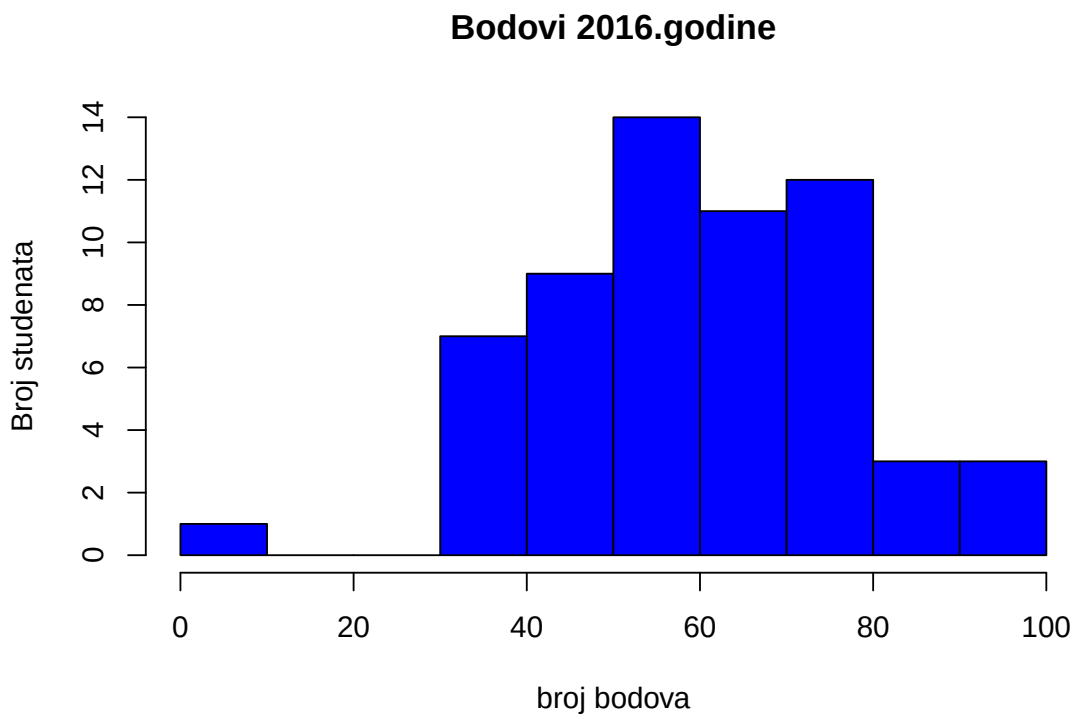
```
boxplot(ucenici_oba_rok$prvi_rok_bodovi, ucenici_oba_rok$drugi_rok_bodovi,
        names = c('Bodovi učenika 1.rok','Bodovi učenika 2.rok'),
        main='Pravokutni dijagram bodova istih učenika na prvom i drugom roku')
```



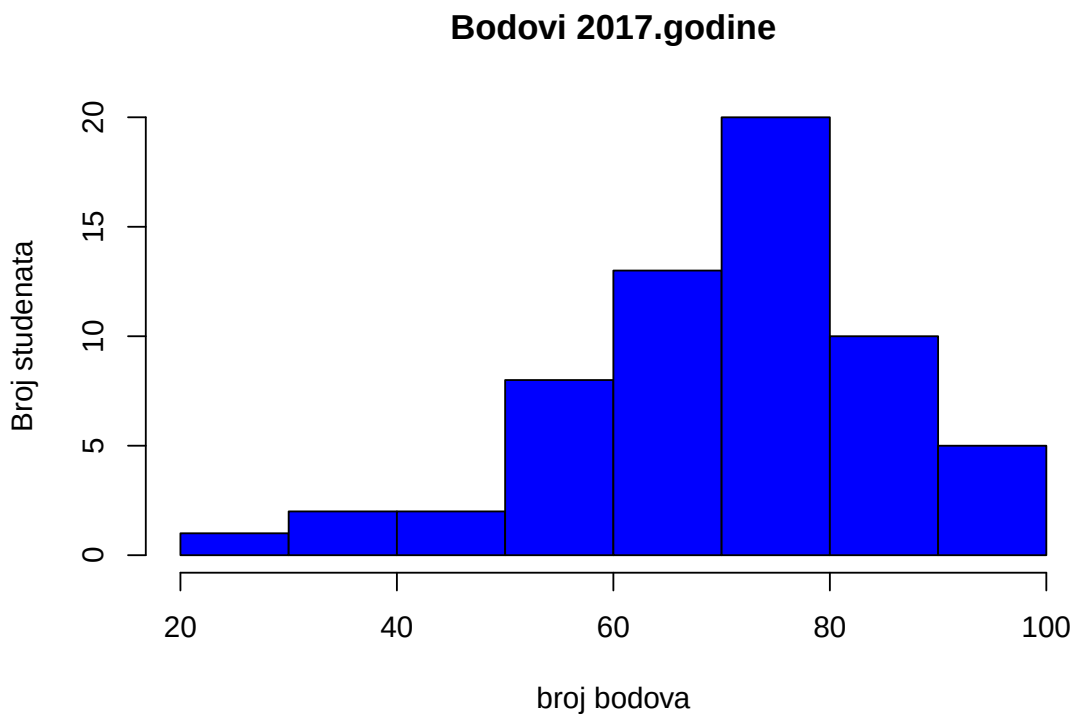
PITANJE: Sada ćemo promatrati razlikuje li se broj ostvarenih bodova kontinuirano između generacija.

Prvo ćemo usporediti histograme i pripadajuće aritmetičke sredine iz godine u godinu.

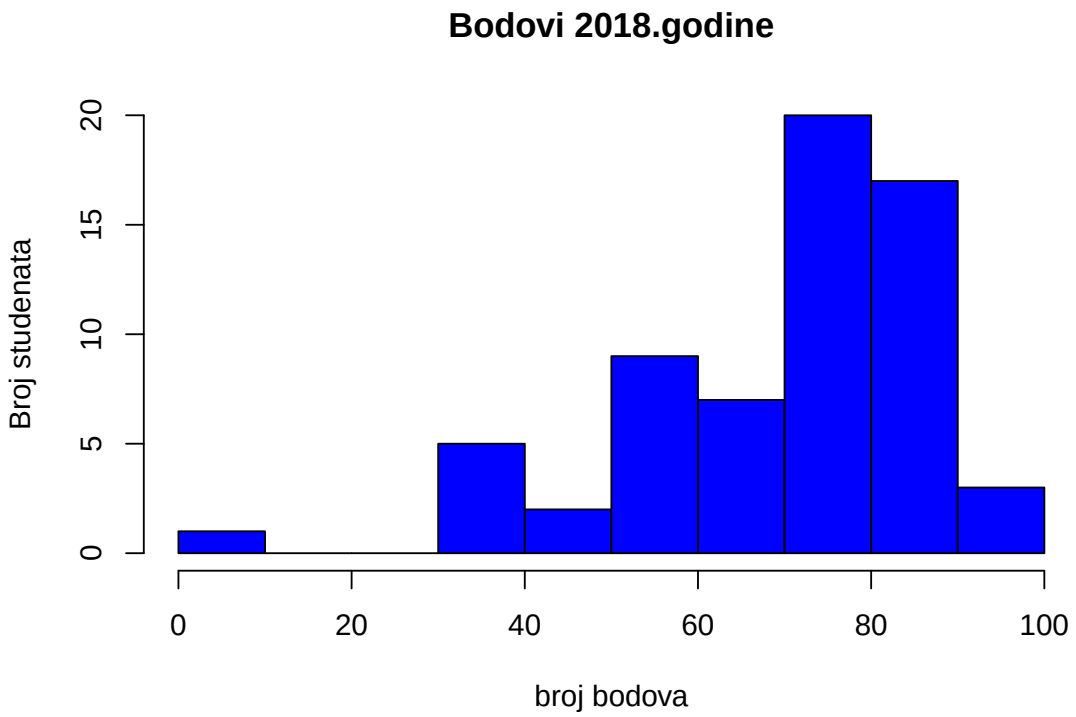
```
h_bodovi_2016 = hist(SAP[SAP$godina == "2016",]$kont_bodovi,
         main="Bodovi 2016.godine",
         xlab="broj bodova",
         ylab='Broj studenata',
         col="blue"
         )
```




```
h_bodovi_2017 = hist(SAP[SAP$godina == "2017",]$kont_bodovi,
         main="Bodovi 2017.godine",
         xlab="broj bodova",
         ylab='Broj studenata',
         col="blue"
         )
```



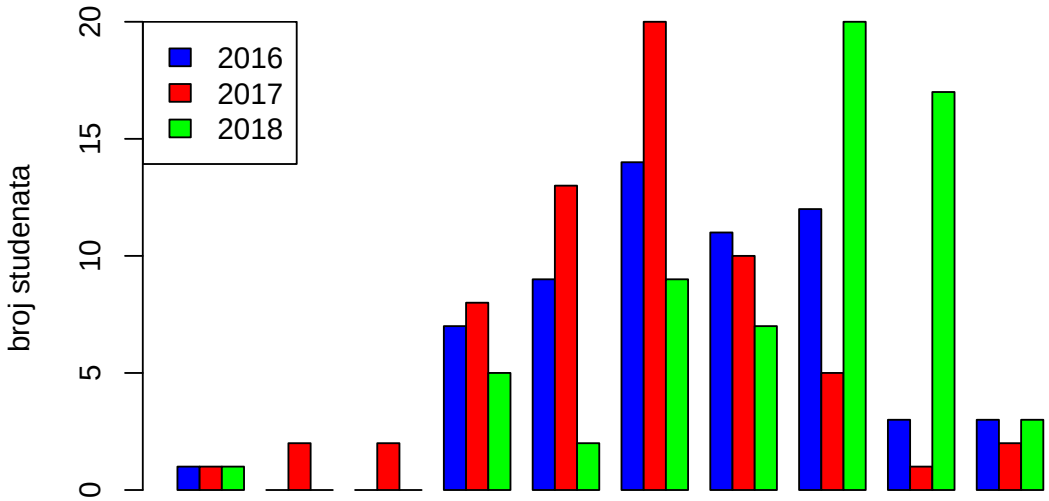

```
h_bodovi_2018 = hist(SAP[SAP$godina == "2018",]$kont_bodovi,
         main="Bodovi 2018.godine",
         xlab="broj bodova",
         ylab='Broj studenata',
         col="blue"
         )
```




```
data <- t(cbind(h_bodovi_2016$counts,h_bodovi_2017$counts,h_bodovi_2018$counts))
```

```
## Warning in cbind(h_bodovi_2016$counts, h_bodovi_2017$counts,
## h_bodovi_2018$counts): number of rows of result is not a multiple of vector
## length (arg 2)
```

```
barplot(data,beside=TRUE, col=c("blue", "red", "green"), ylab="broj studenata")
legend("topleft",c("2016","2017","2018"),fill = c("blue", "red", "green"))
```




```
cat('Prosječan broj bodova učenika 2016 ', mean(SAP[SAP$godina == "2016",]$kont_bodovi),'\n')
```

```
## Prosjecan broj bodova ucenika 2016  60.7625
```

```
cat('Prosječan broj bodova učenika 2017 ', mean(SAP[SAP$godina == "2017",]$kont_bodovi), '\n')
```

```
## Prosjecan broj bodova ucenika 2017  71.07377
```

```
cat('Prosječan broj bodova učenika 2018 ', mean(SAP[SAP$godina == "2018",]$kont_bodovi), '\n')
```

```
## Prosjecan broj bodova ucenika 2018  69.92188
```

Iz podataka aritmetičkih sredina možemom vidjeti kako 2016.godine imamo na prvi pogled dosta manji prosjek bodova, gotovo za jednu ocjenu. Dok su 2017. i 2018. godina tu negdje.

Kod uzoraka možemo vidjeti da se ponašaju približno po normalnoj distribuciji i da su nezavisni iz godine u godinu.

Sada ćemo posebno usporediti 2016. i 2018. godinu. Promatrat ćemo varijance ova dva uzorka i testom o varijanci usporediti jesu li jednake.

```
var(SAP[SAP$godina == "2016",]$isvu_bodovi)
```

```
## [1] 71.94383
```

```
var(SAP[SAP$godina == "2018",]$isvu_bodovi)
```

```
## [1] 78.62268
```

```
var.test(SAP[SAP$godina == "2016",]$isvu_bodovi, SAP[SAP$godina == "2018",]$isvu_bodovi)
```

```
##
##  F test to compare two variances
##
```

```
## data:  SAP[SAP$godina == "2016", ]$isvu_bodovi and SAP[SAP$godina == "2018", ]$isvu_bodovi
## F = 0.91505, num df = 50, denom df = 60, p-value = 0.7508
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5387258 1.5749359
## sample estimates:
## ratio of variances
##          0.9150519
```

Velika p-vrijednost od 0.6471 nam govori da nećemo odbaciti hipotezu H0 da su varijance naša dva uzorka jednake.

Provedimo sada t-test uz pretpostavku jednakosti varijanci.

```
t.test(SAP[SAP$godina == "2018",]$kont_bodovi, SAP[SAP$godina == "2016",]$kont_bodovi, alt = "greater",
```

```
##
##  Two Sample t-test
##
## data:  SAP[SAP$godina == "2018", ]$kont_bodovi and SAP[SAP$godina == "2016", ]$kont_bodovi
## t = 2.8841, df = 122, p-value = 0.002321
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.895714      Inf
## sample estimates:
## mean of x mean of y
##  69.92188  60.76250
```

Mala p-vrijednost nam znači da možemo odbaciti nultu hipotezu u korist hipoteze da je prosječan broj bodova 2018. veći od prosječnog broja bodova 2016. godine.

Isti zaključak slijedi i za 2017. godinu.

```
var.test(SAP[SAP$godina == "2016",]$kont_bodovi, SAP[SAP$godina == "2017",]$kont_bodovi)
```

```
##
##  F test to compare two variances
##
## data:  SAP[SAP$godina == "2016", ]$kont_bodovi and SAP[SAP$godina == "2017", ]$kont_bodovi
## F = 1.2683, num df = 59, denom df = 60, p-value = 0.3613
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7596987 2.1199039
## sample estimates:
## ratio of variances
##           1.268341
```

```
t.test(SAP[SAP$godina == "2018",]$kont_bodovi, SAP[SAP$godina == "2016",]$kont_bodovi, alt = "greater",
```

```
##
##  Two Sample t-test
##
## data:  SAP[SAP$godina == "2018", ]$kont_bodovi and SAP[SAP$godina == "2016", ]$kont_bodovi
## t = 2.8841, df = 122, p-value = 0.002321
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.895714      Inf
## sample estimates:
```

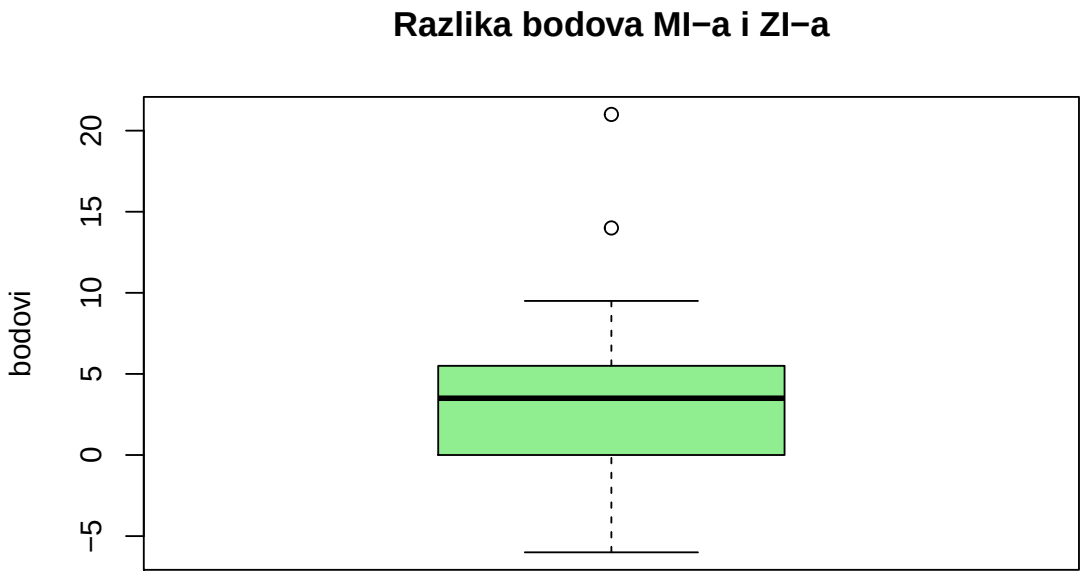
```
## mean of x mean of y
##  69.92188  60.76250
```

PITANJE: Zanima nas jesu li rezultati uspješnih studenta na međuispitu jednaki njihovim rezultatima na završnom ispitu. Smatramo kako je sve iznad 20 bodova jako dobar rezultat na međuispitu te ćemo zbog toga koristiti tu skupinu studenata kako bismo provjerili našu tvrdnju.

```
mi_zi <- do.call(rbind, Map(data.frame, MI=SAP$mi_bodovi, ZI=SAP$zi_bodovi)) #uparivanje
mi_zi <- mi_zi[mi_zi$MI>=20,] #filtriranje: uvjet mi>=20
mi_zi <- mi_zi[mi_zi$ZI>0,]
```

Prvo provjerimo kako nam izgledaju podatci za uparenu razliku broja bodova MI-ja i ZI-ja studenta koji su ostvarili više od 20 bodova na MI-ju.

```
boxplot(mi_zi$MI - mi_zi$ZI,

        col="lightgreen",
        main='Razlika bodova MI-a i ZI-a',
        ylab='bodovi')
```



Na box-plot dijagram možemo vidjeti da je medijan nešto veći od 0 što nam sugerira da provjerimo jesu li bodovi osvareni na MI-ju zaista statistički značajno veći od bodova ostvarenih na ZI-ju. Također možemo vidjeti dva outliera za koje ćemo kasnije provjeriti smetaju li nam prlilikom provedbe statističkog testa.

Koristiti ćemo upareni t-test na razini značajnosti od 5% zato što gledamo razliku broja bodova međuispita i završnog ispita istog studenta, odnosno uzorak nam je zavisan. Postaviti ćemo sljedeće statističke hipoteze:

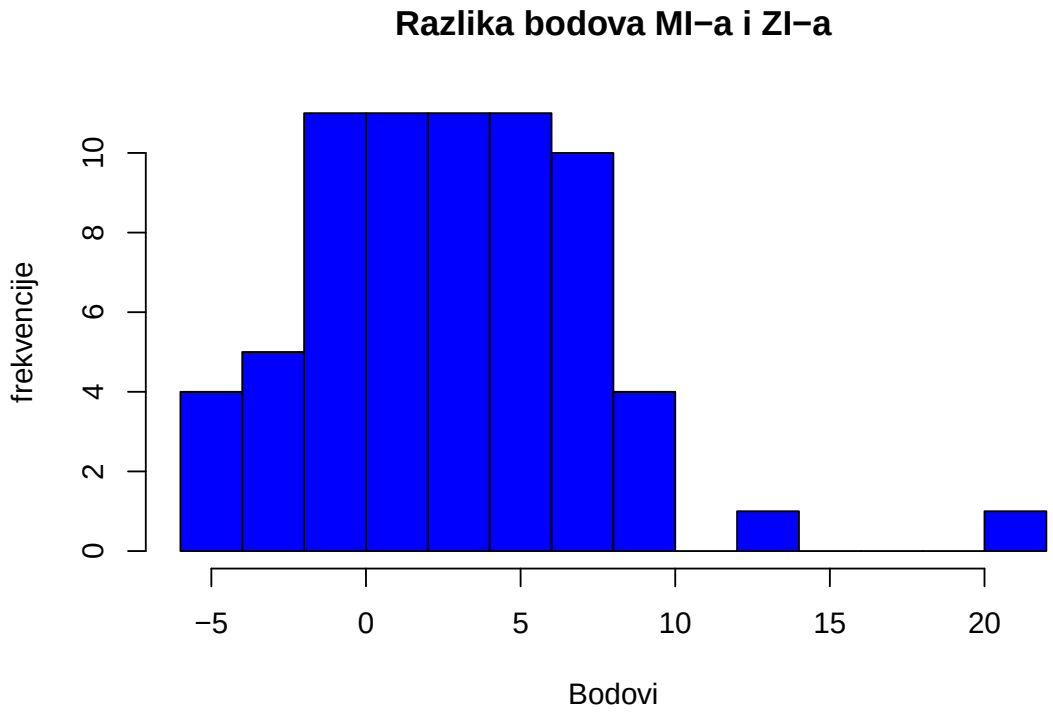
H0... u = 0 H1... u > 0

Za alternativnu hipotezu smo odabrali da studenti koji napišu MI za 20 bodova ili bolje napišu lošije ZI zato što pretpostavljamo da su manje motivirani učiti za ZI pošto su već ostvarili dosta bodova na MI-ju. Također na prethodnom box-plot diajgramu možemo vidjeti da je medijan nešto veći od 0 što nam sugerira da provjerimo jesu li bodovi ostvareni na MI-ju zaista statistički značajno veći od bodova ostvarenih na ZI-u

Kako bismo proveli test, prvo moramo provjeriti ravnaju li se podatci po normalnoj razdiobi ili razdiobi sličnoj normalnoj zato što t-test nije toliko osjetljiv na normalnost.

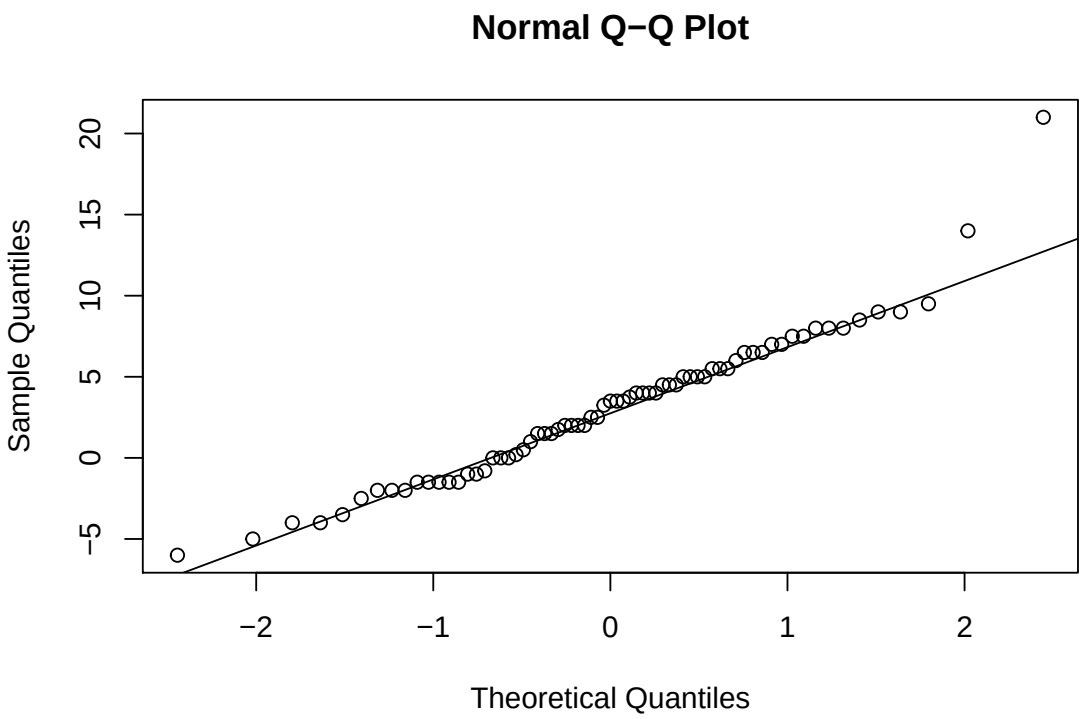
Pogledajmo prvo histograme razdiobe razlike aritmetičkih sredina broja bodova na međuispitu i završnom ispitu studenta:

```
h = hist(mi_zi$MI-mi_zi$ZI,
         main="Razlika bodova MI-a i ZI-a",
         breaks=15,
         xlab="Bodovi",
         ylab='frekvencije',
         col="blue"
)
```



Iz histograma vidimo da distribucija nije baš skroz normalna te ima par "stršećih" vrijednosti pa ćemo provesti još neke testove normalnosti

```
qqnorm(mi_zi$MI-mi_zi$ZI)
qqline(mi_zi$MI-mi_zi$ZI)
```



Na ovom grafu također vidimo da imamo par stršećih vrijednosti, no kao što smo spomenuli t-test nije toliko osjetljiv na normalnost pa ćemo ga provesti.

```
t.test(mi_zi$MI, mi_zi$ZI, alt = "greater", paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  mi_zi$MI and mi_zi$ZI
## t = 5.5918, df = 68, p-value = 2.163e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.162807      Inf
## sample estimates:
## mean of the differences
##                3.081884
```

S obzirom na malu p-vrijednost (približno 0) možemo odbaciti hipotezu H0 na razini značajnosti od 5% u korist alternativne hipoteze H1 koja nam govori da studenti koji ostvare 20 ili više bodova na MI-ju, ostvare lošiji rezultat na ZI-ju nego na MI-ju.

PITANJE: Želimo ispitati jesu li rokovi nezavisni s ocjenama koje studenti postižu na kraju akademske godine? Odnosno zanima nas je li odlazak na specifičan rok imao utjecaj u konačnoj ocjeni studenta.

U tu svrhu provesti ćemo kategorijski test nezavisnosti. Kategorije koje ćemo promatrati su ocjene i rokovi.

Napravimo tablicu koja sadrži informaciju o konačnim ocjenama i mogućim rokovima.

```
rok1<- do.call(rbind, Map(data.frame, ocjena=SAP$isvu_ocjena, rok=SAP$prvi_rok_prolaz))
rok2<- do.call(rbind, Map(data.frame, ocjena=SAP$isvu_ocjena, rok=SAP$drugi_rok_prolaz))
rok3<- do.call(rbind, Map(data.frame, ocjena=SAP$isvu_ocjena, rok=SAP$dek_rok_prolaz))

rok1<-rok1[rok1$rok=="DA",]
rok2<-rok2[rok2$rok=="DA",]
rok3<-rok3[rok3$rok=="DA",]

rok1$rok[rok1$rok=="DA"] <- "prvi_rok"
rok2$rok[rok2$rok=="DA"] <- "drugi_rok"
rok3$rok[rok3$rok=="DA"] <- "dekanski_rok"

rokovi<-rbind(rok1, rok2, rok3)
tbl2 <- table(rokovi)
tbl2
```

```
##       rok
## ocjena dekanski_rok drugi_rok prvi_rok
##      2            1         0        4
##      3            1         8       12
##      4            2        12        5
##      5            1         2        5
```

Retci tablice odgovaraju pripadnoj ocjeni dok stupci tablice odgovaraju rokovima na kojima su studenti položili predmet.

Vidimo da su frekvencije u tablicama poprilično malene zbog čega ćemo umjesto aproksimativnog hi-kvadrat testa iskoristiti egzaktni Fisherov test.

Hipoteze koje ćemo koristiti za test su:

H0... konačna ocjena i vrsta roka na kojem je student položio predmet su nezavisni H1... konačna ocjena i vrsta roka na kojem je student položio predmet nisu nezavisni

Kako bismo proveli test sumirati ćemo stupce i retke tablice.

```
added_margins_tbl2 = addmargins(tbl2)
print(added_margins_tbl2)
```

```
##       rok
## ocjena dekanski_rok drugi_rok prvi_rok Sum
##    2              1         0        4   5
##    3              1         8       12  21
##    4              2        12        5  19
##    5              1         2        5   8
##    Sum            5        22       26  53
```

Sada možemo provesti Fisherov egzaktni test:

```
test <- fisher.test(tbl2,alternative="greater")
test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tbl2
## p-value = 0.07202
## alternative hypothesis: greater
```

Naš test nam ukazuje da na razini značajnosti od 5% ne možemo odbaciti hipotezu H0 da su konačna ocjena i vrsta roka na kojem je student položio predmet nezavisni.

PITANJE: Vidjeli smo da na razini značajnosti od 5% ne možemo odbaciti hipotezu da su konačna ocjena i vrsta roka na kojem je student položio predmet nezavisni. No, sada nas zanima postoji li zavisnost između konačne ocjene i načina na koji je student položio predmet (kontinuiran ili nekontinuirano) zato što se rokovi obično smatraju težim od kontinuiranih provjera.

U tu svrhu opet ćemo provjesti kategorijski test nezavisnosti, no ovaj put sa kategorijama ocjena i način polaganja predmeta, odnosno kontinuirano ili nekontinuirano.

Radimo sljedeću tablicu

```
kategorije <- SAP %>%
  select(isvu_ocjena, kont_prolaz)
kategorije<-kategorije[kategorije$isvu_ocjena>1,]


tablica <- table(kategorije$isvu_ocjena,kategorije$kont_prolaz)
tablica
```

```
##
##     DA NE
##   2  4  5
##   3 39 21
##   4 57 19
##   5 20  6
```

Stupci tablice označavaju je li student prošao kontinuirano ili nije, a redci tablice označavaju konačnu ocjenu koju je student ostvario.

Ovaj put vidimo da frekvencije u tablicama nisu male pa umjesto Fisherovog egzaktnog testa možemo provjesti aproksimativan hi-kvadrat test.

Hipoteze koje ćemo koristiti za test su:

H0... konačna ocjena i načina polaganja predmeta su nezavisni H1... konačna ocjena i način polaganja predmeta nisu nezavisni

Sumirajmo redke i stupce tablice kako bismo mogli provesti naš test:

```
added_margins_tbl = addmargins(tablica)
print(added_margins_tbl)
```

```
##
##         DA  NE Sum
##   2     4   5   9
##   3    39  21  60
##   4    57  19  76
##   5    20   6  26
##   Sum 120  51 171
```

Sada moramo provjeriti jesu li sve relativne frekvencije iz tablice veće od 5, zato što je to predpostavka da hi-kvadrat test da ispravne rezultate.

```
for (col_names in colnames(added_margins_tbl)){
  for (row_names in rownames(added_margins_tbl)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      cat('Ocekivane frekvencije za razred ',col_names,'-',row_names,': ',(added_margins_tbl[row_names,
    }
  }
}
```

```
## Ocekivane frekvencije za razred  DA - 2 :  6.315789
## Ocekivane frekvencije za razred  DA - 3 :  42.10526
## Ocekivane frekvencije za razred  DA - 4 :  53.33333
## Ocekivane frekvencije za razred  DA - 5 :  18.24561
## Ocekivane frekvencije za razred  NE - 2 :  2.684211
## Ocekivane frekvencije za razred  NE - 3 :  17.89474
## Ocekivane frekvencije za razred  NE - 4 :  22.66667
## Ocekivane frekvencije za razred  NE - 5 :  7.754386
```

Vidimo da za razred nekontinuirano i ocjena 2 imamo relativnu frekvenciju manju od 2.

Kako bismo to ispravili spojiti ćemo razrede ocjena 2 i 3 u jedan razred 3. Odnosno ocjene 2 i 3 smatrati ćemo jednakima.

```
kategorije$isvu_ocjena[kategorije$isvu_ocjena==2]<-3

tablica <- table(kategorije$isvu_ocjena,kategorije$kont_prolaz)

added_margins_tbl = addmargins(tablica)
print(added_margins_tbl)
```

```
##
##        DA  NE Sum
##   3    43  26  69
##   4    57  19  76
##   5    20   6  26
##   Sum 120  51 171
```

Sada opet moramo provjeriti relativne frekvencije da vidimo jesmo li uklonili problem.

```
for (col_names in colnames(added_margins_tbl)){
  for (row_names in rownames(added_margins_tbl)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      cat('Ocekivane frekvencije za razred ',col_names,'-',row_names,': ',(added_margins_tbl[row_names,
    }
  }
}
```

```
## Ocekivane frekvencije za razred  DA - 3 :  48.42105
## Ocekivane frekvencije za razred  DA - 4 :  53.33333
## Ocekivane frekvencije za razred  DA - 5 :  18.24561
## Ocekivane frekvencije za razred  NE - 3 :  20.57895
## Ocekivane frekvencije za razred  NE - 4 :  22.66667
## Ocekivane frekvencije za razred  NE - 5 :  7.754386
```

Vidimo da su nam sada sve relativne frekvencije veće od 5, stoga možemo provesti testiranje.

```
chisq.test(tablica,correct=F)
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  tablica
## X-squared = 3.4458, df = 2, p-value = 0.1785
```

P-vrijednost testa iznosi 17.85% što nam ukazuje da na razini značajnosti od 5% ne možemo odbaciti nultu hipotezu da su konačna ocjena i način na koji je student položio predmet nezavisni.

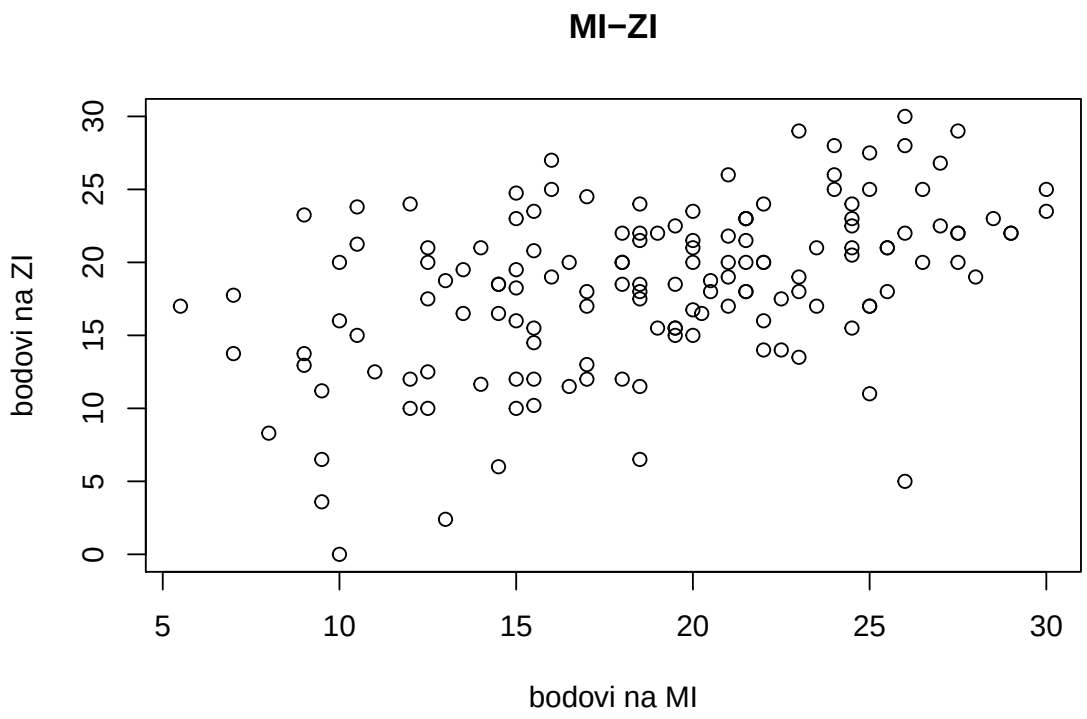
PITANJE: Želimo ispitati možemo li na temelju rezultata međuispita nekog studenta predvidjeti njegov rezultat na završnom ispitu? U tu svrhu pokušati ćemo napraviti regresiju.

Za početak spajamo podatke međuispita i završnog ispita:

```
mi_zi_r<-do.call(rbind, Map(data.frame, MI=SAP$mi_bodovi, ZI=SAP$zi_bodovi)) #uparivanje
mi_zi_r<- mi_zi_r[!is.na(mi_zi_r$ZI),]
mi_zi_r<- mi_zi_r[!is.na(mi_zi_r$MI),]
```

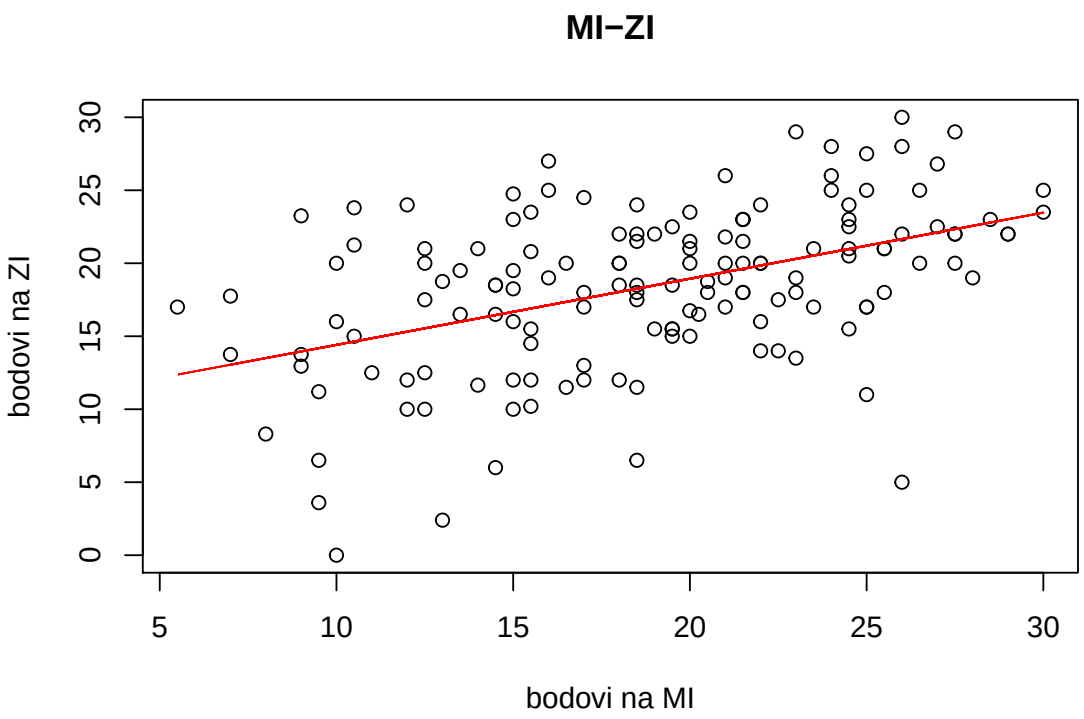
Kako bismo provjerili ima li uopće smisla raditi regresiju napraviti ćemo raspršeni dijagram.

```
plot(x=mi_zi_r$MI, y=mi_zi_r$ZI, main="MI-ZI", xlab = "bodovi na MI", ylab = "bodovi na ZI")  # scatter
```



Vidimo da imamo neki rastući trend pa ćemo provesi regresiju da vidimo možeo li opisati rezultate zi-a na temelju rezultata mi-a.

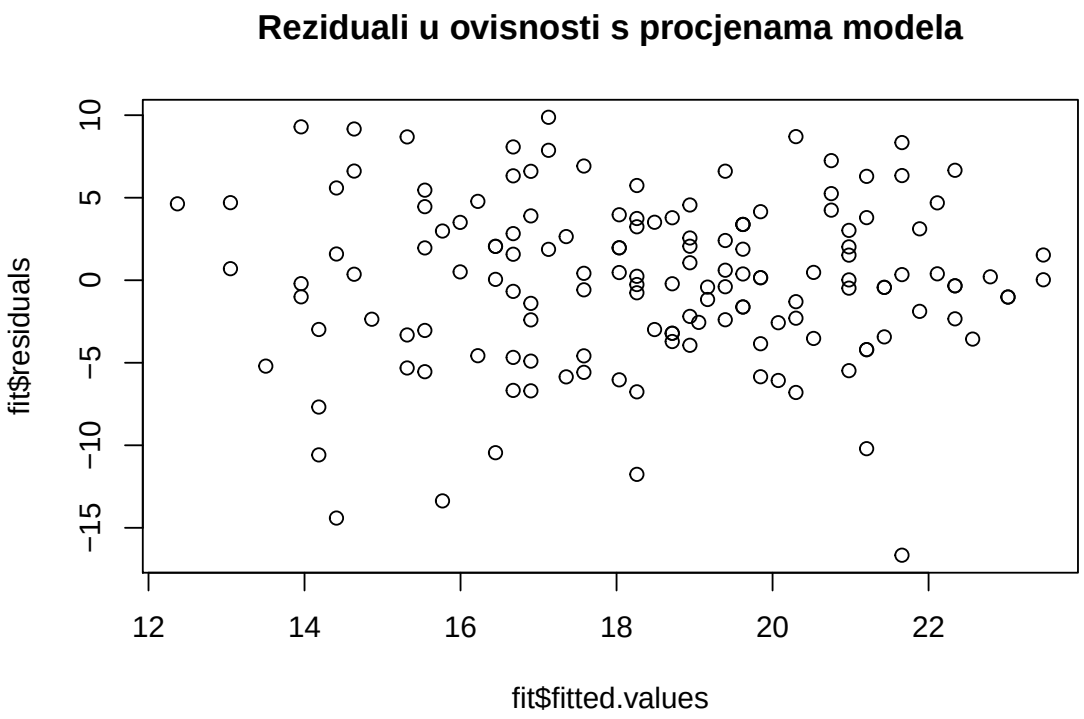
```
fit = lm(ZI~MI,data=mi_zi_r)
#fit$coefficients #koeficijenti linearnog modela
plot(mi_zi_r$MI,mi_zi_r$ZI,xlab = "bodovi na MI", ylab = "bodovi na ZI", main="MI-ZI") #plot podataka
lines(mi_zi_r$MI,fit$fitted.values,col='red') #plot fitanih vrijednosti
```



Kako bismo dalje na ovom modelu mogli provoditi bilo kakve testove osnova je da nam se reziduali ravnaju prema normalnoj distribuciji ili distribuciji približno normalnoj zato što t-test nije toliko osjetljiv na normalnost.

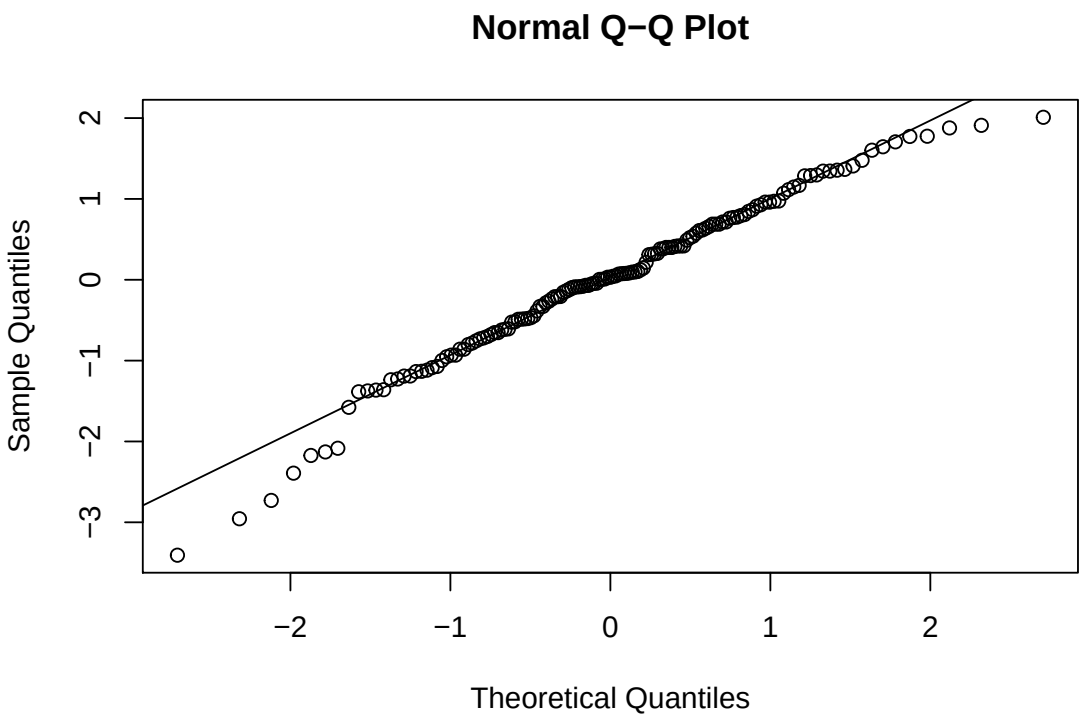
Pogledajmo kako nam izgledaju reziduali u ovisnosti s procjenama modela:

```
plot(fit$fitted.values,fit$residuals,main="Reziduali u ovisnosti s procjenama modela")
```



Iz ovog grafa možemo naslutiti normalnost podataka, odnosno da reziduali nisu ovisni o procjenama modela, no provesti ćemo detaljnije testove.

```
qqnorm(rstandard(fit))
qqline(rstandard(fit))
```



Na grafu Q-Q plota vidimo da imamo stršeće vrijednosti pa ćemo provesti i Kolmogorov-Smirnovljev test normalnosti kako bismo se stvarno uvjerili da nam se podatci ravnaju po normalnoj distribuciji ili distribuciji sličnoj normalnoj.

```
ks.test(rstandard(fit),'pnorm')
```

```
## Warning in ks.test(rstandard(fit), "pnorm"): ties should not be present for the
## Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit)
## D = 0.059678, p-value = 0.6718
## alternative hypothesis: two-sided
```

P-vrijednost od 67.18% nam ukazuje da na razini značajnosti od 5% ne možemo odbaciti hipotezu da nam se podatci ravnaju po normalnoj razdiobi, stoga možemo dalje nastaviti sa testiranjem našeg modela.

```
summary(fit)
```

```
##
## Call:
## lm(formula = ZI ~ MI, data = mi_zi_r)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.6589  -3.0148   0.1534   3.3800   9.8719
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  9.87867    1.41758   6.969 1.04e-10 ***
## MI           0.45309    0.07188   6.304 3.30e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.933 on 145 degrees of freedom
## Multiple R-squared:  0.2151, Adjusted R-squared:  0.2097
## F-statistic: 39.74 on 1 and 145 DF,  p-value: 3.295e-09
```

Provevši razne testove na modelu možemo zaključiti sljedeće:

Na razini značajnosti od 5% možemo odbaciti hipotezu da su koeficijenti b0 i b1 jednaki nuli u korist alternative da su koeficijenti b0 i b1 različiti od 0.

R kvadrat vrijednost nam govori da smo sa našim modelom uspjeli objasniti 20.16% varijabilnosti u podatcima. S obzirom na to da su nam podatci zasnovani na ljudskom faktoru to i nije toliko loš rezultat te smo zadovoljni dobivenim modelom.