

# student-success-analysis

## Deskriptivna analiza

### Osnove

Učitavamo podatke, provjeravamo kojeg je oblika skup podataka i od kojih se stupaca sastoji.

```
students_org <- readxl::read_excel("student_data.xlsx")  
  
# 370 rows, 39 columns  
dim(students_org)  
  
# Show column names  
names(students_org)
```

Provjeravamo prvih par redataka podatkovnog skupa

```
# Show first few rows  
head(students_org)
```

Saznajemo osnovne podatke za svaki stuapc

```
# Show details for each column  
summary(students_org)
```

Provjeravamo koji su stupci kojeg tipa: numerički, kategorički...

```
# Check the class of the column. 'numeric', 'character'...  
sapply(students_org, class)
```

Provjeravamo postoje li nevažeci podaci koji prelaze maksimalne vrijednosti specificirane u uputama o podacima. Sve vrijednosti su dobrom intervalu.

```
# Let's check if any columns exceed the maximum or mininum values specified in  
# the pdf This makes sense only for numerical values  
  
colMax <- students_org %>%  
  select(where(is.numeric)) %>%  
  sapply(., max, na.rm = TRUE)  
colMax  
# Every column has normal maximum value
```

Izbacivanje svih NaN/NA/null vrijednosti iz podatkovnog skupa. Na sreću, takvih vrijednosti nije bilo.

```
# Are there any na values?  
students_org %>%  
  filter(is.na(.))  
sum(apply(students_org, 2, is.nan))  
students_org %>%  
  filter(is.null(.)) %>%  
  summarise(n = n())
```

```

# Drop these values just in case they show up with another dataset We will
# continue using 'student' variable
students <- students_org %>%
  filter_all(all_vars(!is.na(.) & !is.nan(.) & !is.null(.)))

students_clean <- students

```

## Petar Dragojević

### zavisnost između edukacije roditelja i uspješnosti

```

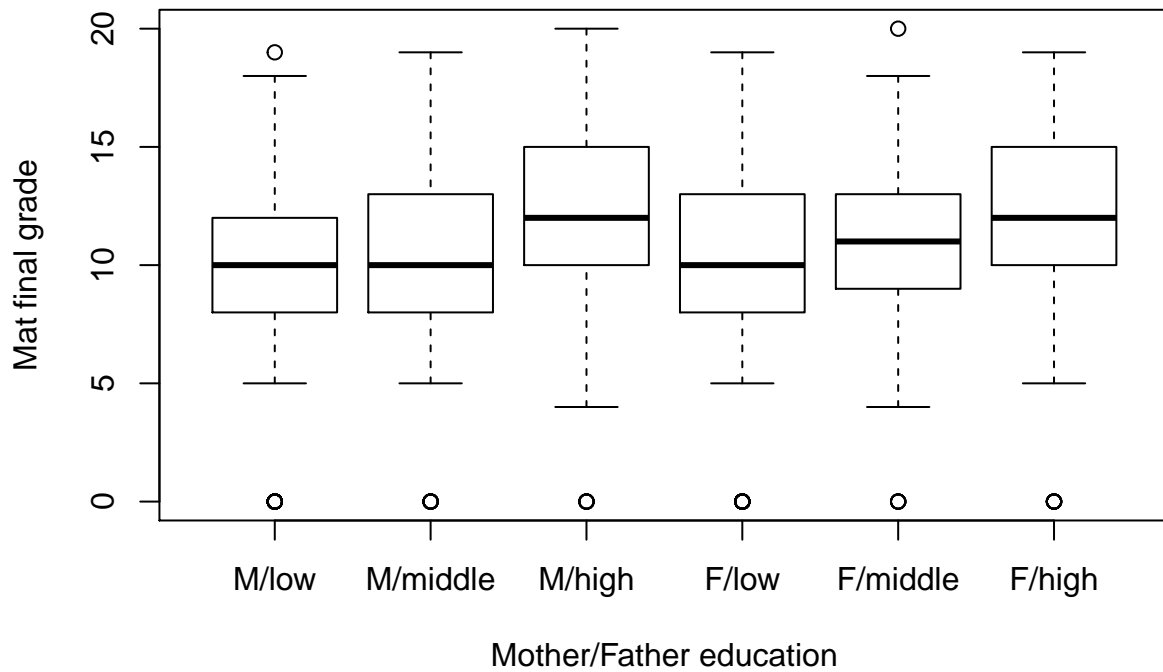
# Mat_grade i Por_grade prebacivanje ocjena u engleski način ocjenjivanja
students <- students %>%
  mutate(Mat_grade = case_when(G3_mat < 10 ~ "F", G3_mat >= 10 & G3_mat < 14 ~
    "C", G3_mat >= 14 & G3_mat < 16 ~ "B", G3_mat >= 16 ~ "A"))
students <- students %>%
  mutate(Por_grade = case_when(G3_por < 10 ~ "F", G3_por >= 10 & G3_mat < 14 ~
    "C", G3_por >= 14 & G3_mat < 16 ~ "B", G3_por >= 16 ~ "A"))

# MeduMod i FeduMod grupiranje edukacije roditelja u veće podgrupe
students <- students %>%
  mutate(MeduMod = case_when(Medu == "0" | Medu == "1" | Medu == "2" ~ "0", Medu ==
    "3" ~ "1", Medu == "4" ~ "2"))
students <- students %>%
  mutate(FeduMod = case_when(Fedu == "0" | Medu == "1" | Fedu == "2" ~ "0", Fedu ==
    "3" ~ "1", Fedu == "4" ~ "2"))

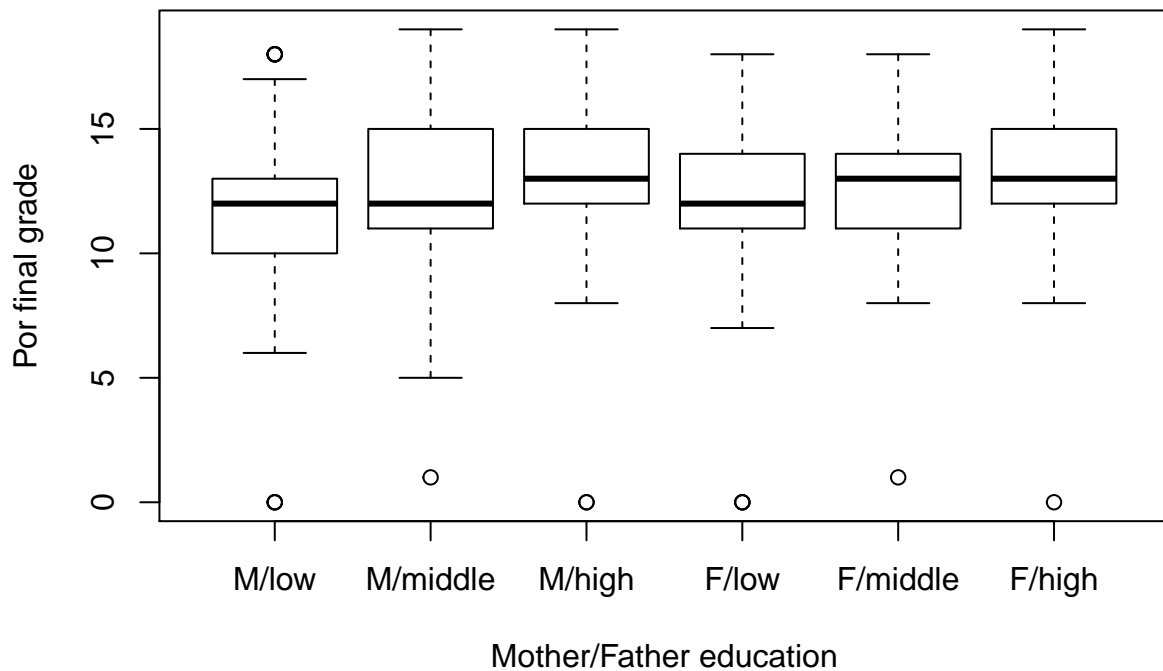
# za edukaciju roditelja uzimamo onu koja je veća
students$highestparentedu <- pmax(students$MeduMod, students$FeduMod)

boxplot(students$G3_mat[students$MeduMod == "0"], students$G3_mat[students$MeduMod ==
  "1"], students$G3_mat[students$MeduMod == "2"], students$G3_mat[students$FeduMod ==
  "0"], students$G3_mat[students$FeduMod == "1"], students$G3_mat[students$FeduMod ==
  "2"], names = c("M/low", "M/middle", "M/high", "F/low", "F/middle", "F/high"),
  xlab = "Mother/Father education", ylab = "Mat final grade")

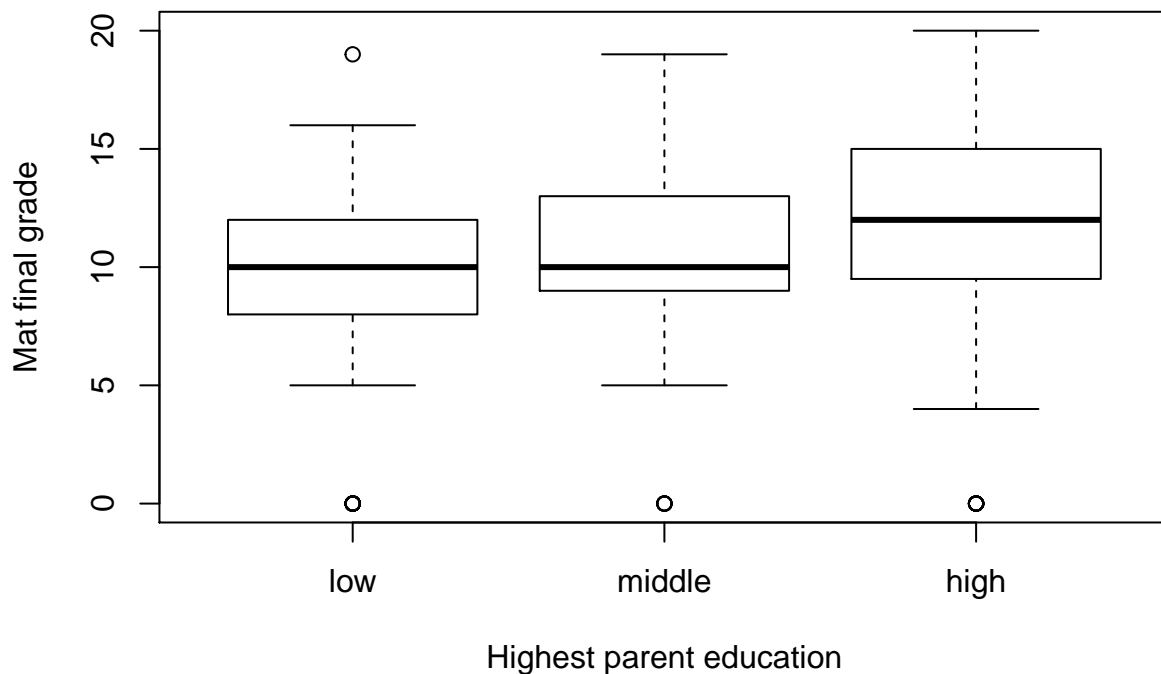
```



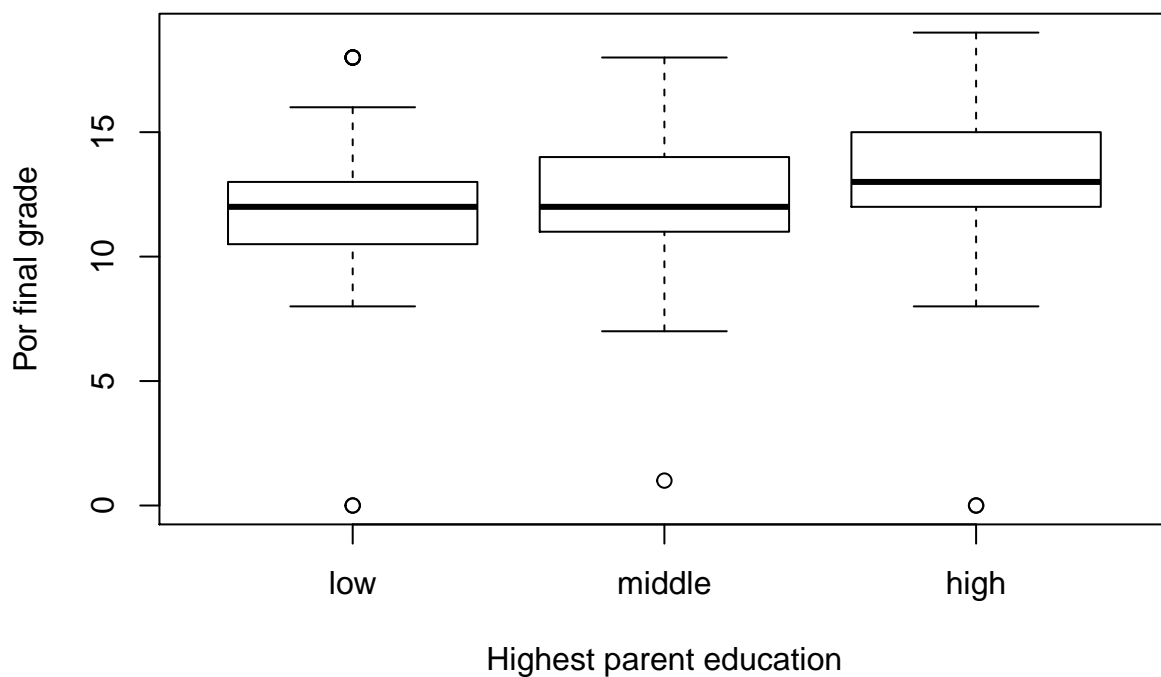
```
boxplot(students$G3_por[students$MeduMod == "0"], students$G3_por[students$MeduMod ==
"1"], students$G3_por[students$MeduMod == "2"], students$G3_por[students$FeduMod ==
"0"], students$G3_por[students$FeduMod == "1"], students$G3_por[students$FeduMod ==
"2"], names = c("M/low", "M/middle", "M/high", "F/low", "F/middle", "F/high"),
xlab = "Mother/Father education", ylab = "Por final grade")
```



```
boxplot(students$G3_mat[students$highestparentedu == "0"], students$G3_mat[students$highestparentedu ==
"1"], students$G3_mat[students$highestparentedu == "2"], names = c("low", "middle",
"high"), xlab = "Highest parent education", ylab = "Mat final grade")
```



```
boxplot(students$G3_por[students$highestparentedu == "0"], students$G3_por[students$highestparentedu ==
"1"], students$G3_por[students$highestparentedu == "2"], names = c("low", "middle",
"high"), xlab = "Highest parent education", ylab = "Por final grade")
```



```
# H0: Ocjena iz matematike i edukacija više educiranog roditelja su nezavisna
# obilježja H1: Ocjena iz matematike i edukacija više educiranog roditelja su
# zavisna obilježja
```

```
tbl = table(students$highestparentedu, students$Mat_grade)
added_margins_tbl = addmargins(tbl)
print(added_margins_tbl)
```

```
##
##      A   B   C   F Sum
##  0    4   9  48  39 100
##  1    7   8  37  25  77
##  2   25  35  54  38 152
## Sum  36  52 139 102 329
```

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

```
##
## Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 25.134, df = 6, p-value = 0.0003224
```

p-value testa iznosi manje od 0.05 stoga odbacujemo hipotezu H0, te zaključujemo da su edukacija više educiranog roditelja i završna ocjena iz matematike zavisne.

```
# H0: Ocjena iz matematike i edukacija majke su nezavisna obilježja H1: Ocjena
# iz matematike i edukacija majke su zavisna obilježja
```

```
tbl = table(students$MeduMod, students$Mat_grade)
added_margins_tbl = addmargins(tbl)
print(added_margins_tbl)
```

```
##
##      A   B   C   F Sum
##  0    7  17  68  56 148
##  1   11  12  37  33  93
##  2   22  30  45  32 129
## Sum  40  59 150 121 370
```

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

```
##
## Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 22.482, df = 6, p-value = 0.0009898
```

```
# H0: Ocjena iz matematike i edukacija oca su nezavisna obilježja H1: Ocjena iz
# matematike i edukacija oca su zavisna obilježja
```

```
tbl2 = table(students$FeduMod, students$Mat_grade)
added_margins_tbl2 = addmargins(tbl2)
print(added_margins_tbl2)
```

```
##
##      A   B   C   F Sum
##  0   11  19  64  49 143
##  1   11  12  40  30  93
##  2   14  21  35  23  93
## Sum  36  52 139 102 329
```

```
chisq.test(tbl2, correct = F)
```

```
##
## Pearson's Chi-squared test
##
## data:  tbl2
```

```
## X-squared = 9.0666, df = 6, p-value = 0.1699
```

p-value Testa nezavisnosti iznosi manje od 0.05 stoga odbacujemo hipotezu  $H_0$ , te zaključujemo da su edukacija majke i završna ocjena iz matematike zavisne, dok do tog zaključka ne možemo doći u slučaju edukacije oca.

```
# H0: Ocjena iz portugala i edukacija više educiranog roditelja su nezavisna  
# obilježja H1: Ocjena iz portugala i edukacija više educiranog roditelja su  
# zavisna obilježja
```

```
tbl = table(students$highestparentedu, students$Por_grade)  
added_margins_tbl = addmargins(tbl)  
print(added_margins_tbl)
```

```
##  
##      A    B    C    F Sum  
## 0      3    3   77   10  93  
## 1      3    6   57    5  71  
## 2     12   23   85    7 127  
## Sum   18   32  219   22 291
```

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

```
## Warning in chisq.test(tbl, correct = F): Chi-squared approximation may be  
## incorrect
```

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

```
## data:  tbl  
## X-squared = 19.409, df = 6, p-value = 0.003526
```

```
# očekivane frekvencije svih razreda moraju biti veće ili jednake 5
```

```
for (col_names in colnames(added_margins_tbl)) {  
  for (row_names in rownames(added_margins_tbl)) {  
    if (!(row_names == "Sum" | col_names == "Sum")) {  
      cat("Očekivane frekvencije za razred ", col_names, "-", row_names, ": ",  
          (added_margins_tbl[row_names, "Sum"] * added_margins_tbl["Sum", col_names])/added_margins_tbl["Sum", "Sum"], "\n")  
    }  
  }  
}
```

```
## Očekivane frekvencije za razred A - 0 : 5.752577  
## Očekivane frekvencije za razred A - 1 : 4.391753  
## Očekivane frekvencije za razred A - 2 : 7.85567  
## Očekivane frekvencije za razred B - 0 : 10.2268  
## Očekivane frekvencije za razred B - 1 : 7.80756  
## Očekivane frekvencije za razred B - 2 : 13.96564  
## Očekivane frekvencije za razred C - 0 : 69.98969  
## Očekivane frekvencije za razred C - 1 : 53.43299  
## Očekivane frekvencije za razred C - 2 : 95.57732  
## Očekivane frekvencije za razred F - 0 : 7.030928  
## Očekivane frekvencije za razred F - 1 : 5.367698  
## Očekivane frekvencije za razred F - 2 : 9.601375
```

```
# Vidimo da postoje očekivane frekvencije manje od 5 pa koristimo fisher.test()  
# umjesto chisq.test()  
fisher.test(tbl)
```

```
##
## Fisher's Exact Test for Count Data
##
## data:  tbl
## p-value = 0.003003
## alternative hypothesis: two.sided
```

chisq.test je nepouzdan pošto su očekivane frekvencije pojedinih razreda manje od 5, radi toga koristimo fisher.test. p-value Fesherovog testa iznosi manje od 0.05 stoga odbacujemo hipotezu  $H_0$ , te zaključujemo da su edukacija više educiranog roditelja i završna ocjena iz portugala zavisne.

```
# H0: Ocjena iz portugala i edukacija majke su nezavisna obilježja H1: Ocjena
# iz portugala i edukacija majke su zavisna obilježja
tbl = table(students$MeduMod, students$Por_grade)
added_margins_tbl = addmargins(tbl)
print(added_margins_tbl)
```

```
##
##      A    B    C    F Sum
## 0      5    9 108   16 138
## 1      6    9  62    8  85
## 2      9   19  72    5 105
## Sum   20   37 242   29 328
```

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

```
##
## Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 13.658, df = 6, p-value = 0.0337
```

```
# H0: Ocjena iz portugala i edukacija oca su nezavisna obilježja H1: Ocjena iz
# portugala i edukacija oca su zavisna obilježja
tbl2 = table(students$FeduMod, students$Por_grade)
added_margins_tbl2 = addmargins(tbl2)
print(added_margins_tbl2)
```

```
##
##      A    B    C    F Sum
## 0      8   10  99   14 131
## 1      3    7  67    3  80
## 2      7   15  53    5  80
## Sum   18   32 219   22 291
```

```
chisq.test(tbl2, correct = F)
```

```
## Warning in chisq.test(tbl2, correct = F): Chi-squared approximation may be
## incorrect
```

```
##
## Pearson's Chi-squared test
##
## data:  tbl2
## X-squared = 12.75, df = 6, p-value = 0.04719
```

```
# očekivane frekvencije svih razreda moraju biti veće ili jednake 5
for (col_names in colnames(added_margins_tbl2)) {
```

```

for (row_names in rownames(added_margins_tbl2)) {
  if (!(row_names == "Sum" | col_names == "Sum")) {
    cat("Očekivane frekvencije za razred ", col_names, "-", row_names, ": ",
        (added_margins_tbl2[row_names, "Sum"] * added_margins_tbl2["Sum",
        col_names])/added_margins_tbl2["Sum", "Sum"], "\n")
  }
}
}

```

```

## Očekivane frekvencije za razred A - 0 : 8.103093
## Očekivane frekvencije za razred A - 1 : 4.948454
## Očekivane frekvencije za razred A - 2 : 4.948454
## Očekivane frekvencije za razred B - 0 : 14.4055
## Očekivane frekvencije za razred B - 1 : 8.797251
## Očekivane frekvencije za razred B - 2 : 8.797251
## Očekivane frekvencije za razred C - 0 : 98.58763
## Očekivane frekvencije za razred C - 1 : 60.20619
## Očekivane frekvencije za razred C - 2 : 60.20619
## Očekivane frekvencije za razred F - 0 : 9.90378
## Očekivane frekvencije za razred F - 1 : 6.04811
## Očekivane frekvencije za razred F - 2 : 6.04811

```

```

# Vidimo da postoje očekivane frekvencije manje od 5 pa koristimo fisher.test()
# umjesto chiq.test()
fisher.test(tbl)

```

```

##
## Fisher's Exact Test for Count Data
##
## data: tbl
## p-value = 0.03133
## alternative hypothesis: two.sided

```

p-value Testa nezavisnosti iznosi manje od 0.05 stoga odbacujemo hipotezu  $H_0$ , te zaključujemo da su edukacija majke i završna ocjena iz portugala zavisne, kod usporedbe s edukacijom oca koristimo Fisherov test gdje je p-value manji od 0.05 pa odbacujemo  $H_0$  i zaključujemo da su edukacija oca i završna ocjena iz portugala zavisne.

## Tomislav Prhat

1. Jesu li učenici uspješniji u matematici ili glavnom jeziku?

```

students_org %>%
  summarise(Mean.G3_mat = mean(G3_mat), Mean.G3_por = mean(G3_por), ) -> summary.result1
summary.result1

```

```

## # A tibble: 1 x 2
##   Mean.G3_mat Mean.G3_por
##       <dbl>       <dbl>
## 1      10.5       12.6

```

```

students_org %>%
  summarise(Med.G3_mat = median(G3_mat), Med.G3_por = median(G3_por), ) -> summary.result2
summary.result2

```

```

## # A tibble: 1 x 2
##   Med.G3_mat Med.G3_por

```



```
##          <dbl>      <dbl>
## 1          11         13
```

```
students_org %>%
  summarise(Mean.G3_mat = mean(G3_mat, trim = 0.1), Mean.G3_por = mean(G3_por,
    trim = 0.1), ) -> summary.result3
summary.result3
```

```
## # A tibble: 1 x 2
##   Mean.G3_mat Mean.G3_por
##         <dbl>      <dbl>
## 1         10.9        12.6
```

```
(1 - summary.result3/summary.result1) * 100
```

```
##   Mean.G3_mat Mean.G3_por
## 1   -4.016012 -0.7265877
```

Kao što je vidljivo iz podataka, učenici su malo uspješniji u glavnom jeziku (portugalskom), ali ako gleda prema samoj ocjeni obje skupine spadaju u ocjenu "C". Čak i ako uzmemo podrezanu srednju vrijednost (10%), rezultat se promijeni za ~1%.

```
students_org %>%
  summarise(IQR.G3_mat = IQR(G3_mat), IQR.G3_por = IQR(G3_por), ) -> summary.result4
summary.result4
```

```
## # A tibble: 1 x 2
##   IQR.G3_mat IQR.G3_por
##         <dbl>      <dbl>
## 1           6         3
```

```
students_org %>%
  summarise(Var.G3_mat = var(G3_mat), Var.G3_por = var(G3_por), ) -> summary.result5
summary.result5
```

```
## # A tibble: 1 x 2
##   Var.G3_mat Var.G3_por
##         <dbl>      <dbl>
## 1         21.2        8.67
```

```
students_org %>%
  summarise(sd.G3_mat = sd(G3_mat), sd.G3_por = sd(G3_por), ) -> summary.result6
summary.result6
```

```
## # A tibble: 1 x 2
##   sd.G3_mat sd.G3_por
##         <dbl>      <dbl>
## 1         4.61        2.94
```

Ako gledamo raspršenost varijabli vidimo da ocjene iz portugalskog jezika imaju manje sve tri mjere (IQR, varijanca i standardna devijacija) vidimo da se ocjene iz portugalskog manje manje odmiču od srednje vrijednosti nego ocjene iz matematike.

```
boxplot(students_org$G3_mat, students_org$G3_por, names = c("konačna ocjena iz matematike",
  "konačna ocjena iz portugalskog"), main = "Boxplot konačnih ocjena iz matematike i portugala")
```

```
## Warning in axis(side = 1, at = 1:2, labels = c("konačna ocjena iz matematike", :
## conversion failure on 'konačna ocjena iz matematike' in 'mbcsToSbcs': dot
## substituted for <c4>
```

```
## Warning in axis(side = 1, at = 1:2, labels = c("konačna ocjena iz matematike", :
## conversion failure on 'konačna ocjena iz matematike' in 'mbcsToSbcs': dot
## substituted for <8d>

## Warning in axis(side = 1, at = 1:2, labels = c("konačna ocjena iz matematike", :
## conversion failure on 'konačna ocjena iz matematike' in 'mbcsToSbcs': dot
## substituted for <c4>

## Warning in axis(side = 1, at = 1:2, labels = c("konačna ocjena iz matematike", :
## conversion failure on 'konačna ocjena iz matematike' in 'mbcsToSbcs': dot
## substituted for <8d>

## Warning in axis(side = 1, at = 1:2, labels = c("konačna ocjena iz matematike", :
## conversion failure on 'konačna ocjena iz portugalskog' in 'mbcsToSbcs': dot
## substituted for <c4>

## Warning in axis(side = 1, at = 1:2, labels = c("konačna ocjena iz matematike", :
## conversion failure on 'konačna ocjena iz portugalskog' in 'mbcsToSbcs': dot
## substituted for <8d>

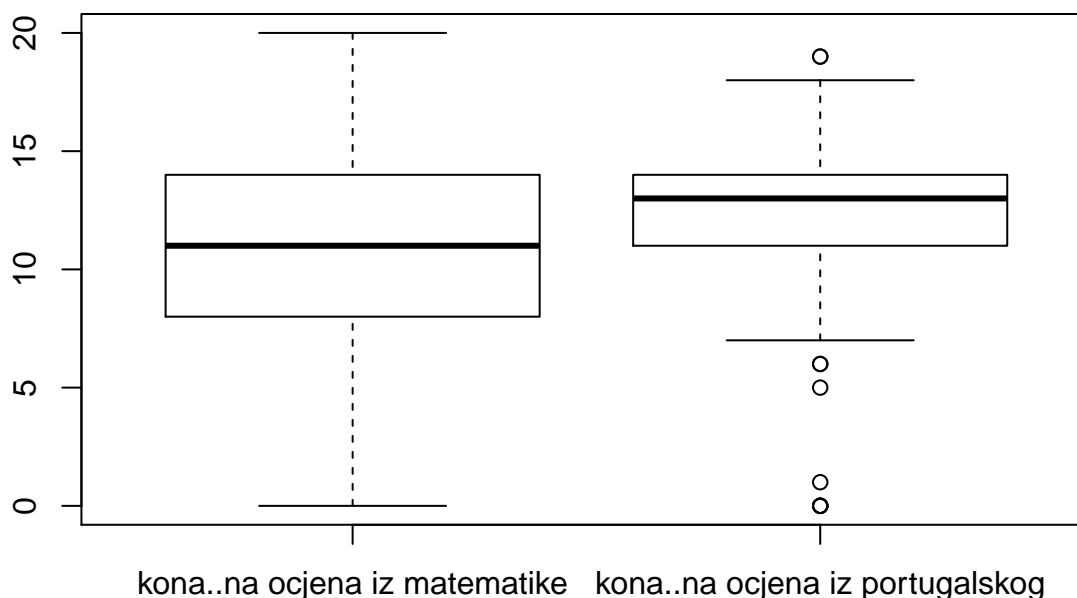
## Warning in axis(side = 1, at = 1:2, labels = c("konačna ocjena iz matematike", :
## conversion failure on 'konačna ocjena iz portugalskog' in 'mbcsToSbcs': dot
## substituted for <c4>

## Warning in axis(side = 1, at = 1:2, labels = c("konačna ocjena iz matematike", :
## conversion failure on 'konačna ocjena iz portugalskog' in 'mbcsToSbcs': dot
## substituted for <8d>

## Warning in (function (main = NULL, sub = NULL, xlab = NULL, ylab = NULL, :
## conversion failure on 'Boxplot konačnih ocjena iz matematike i portugala' in
## 'mbcsToSbcs': dot substituted for <c4>

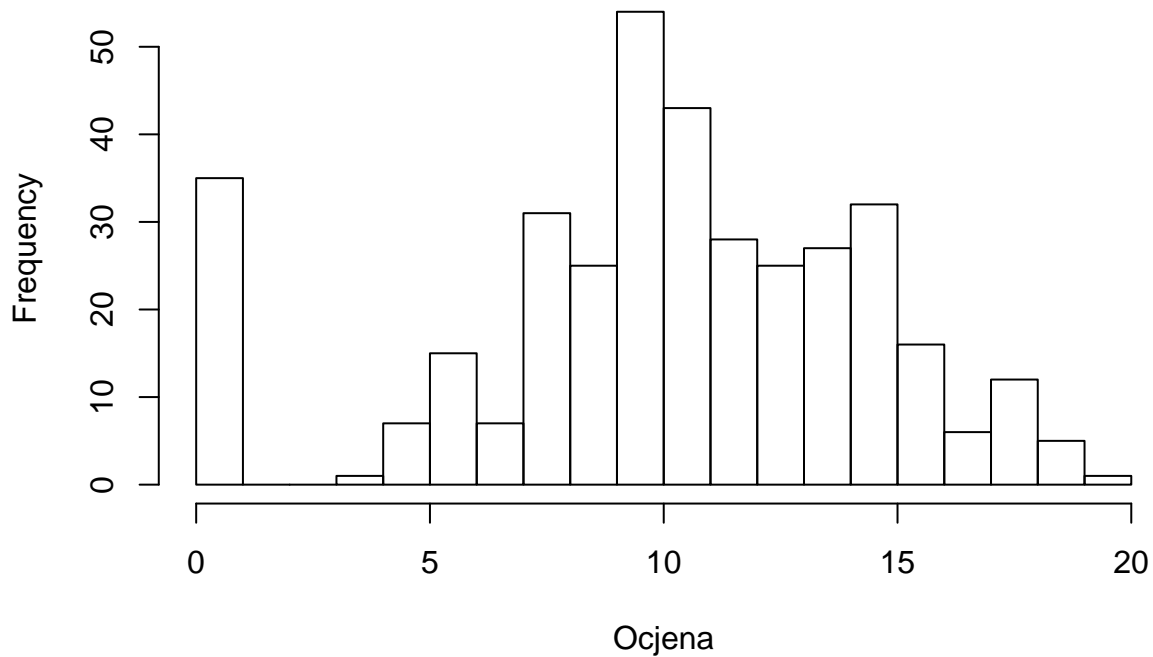
## Warning in (function (main = NULL, sub = NULL, xlab = NULL, ylab = NULL, :
## conversion failure on 'Boxplot konačnih ocjena iz matematike i portugala' in
## 'mbcsToSbcs': dot substituted for <8d>
```

## Boxplot kona..nih ocjena iz matematike i portugala



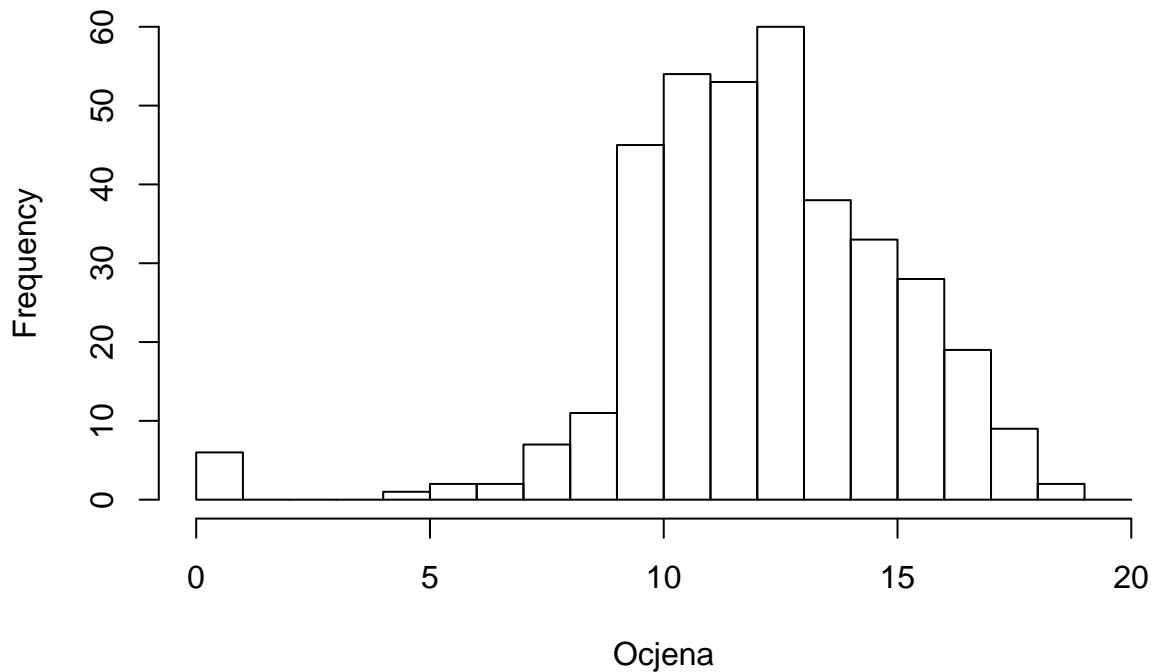
```
hist(students_org$G3_mat, breaks = seq(0, 20), main = "Histogram ocjena iz matematike",
     xlab = "Ocjena")
```

**Histogram ocjena iz matematike**

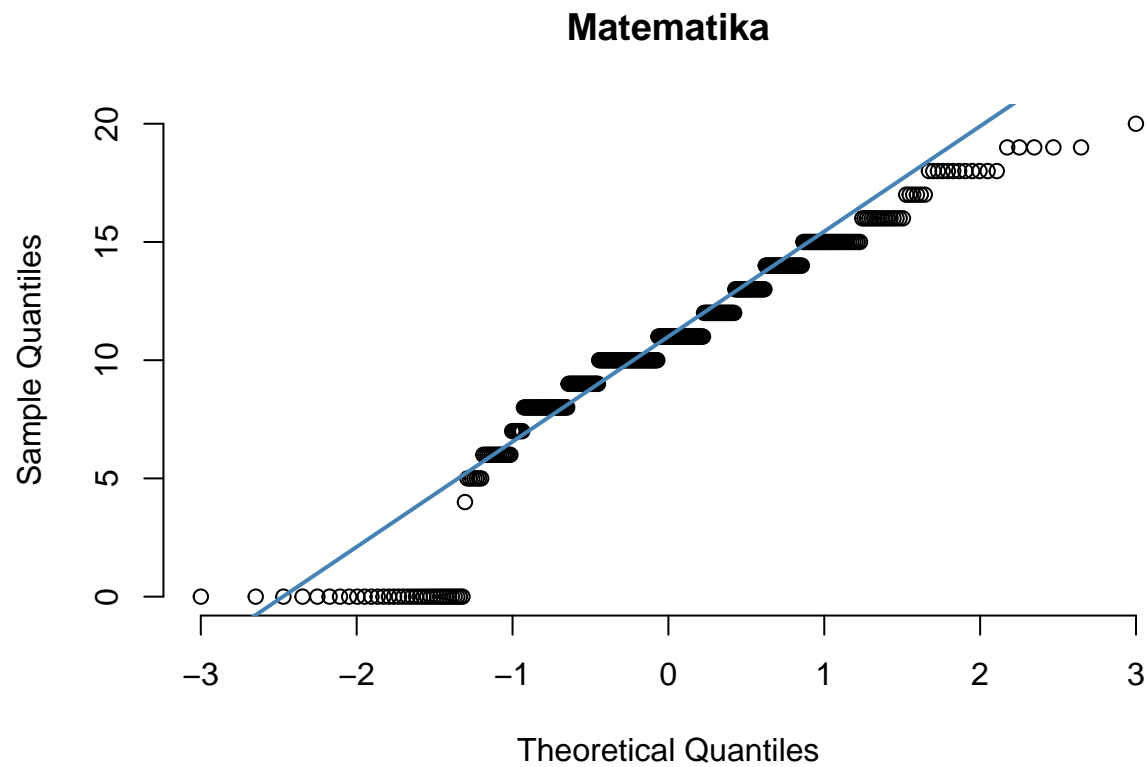


```
hist(students_org$G3_por, breaks = seq(0, 20), main = "Histogram ocjena iz portugalskog",
     xlab = "Ocjena")
```

**Histogram ocjena iz portugalskog**

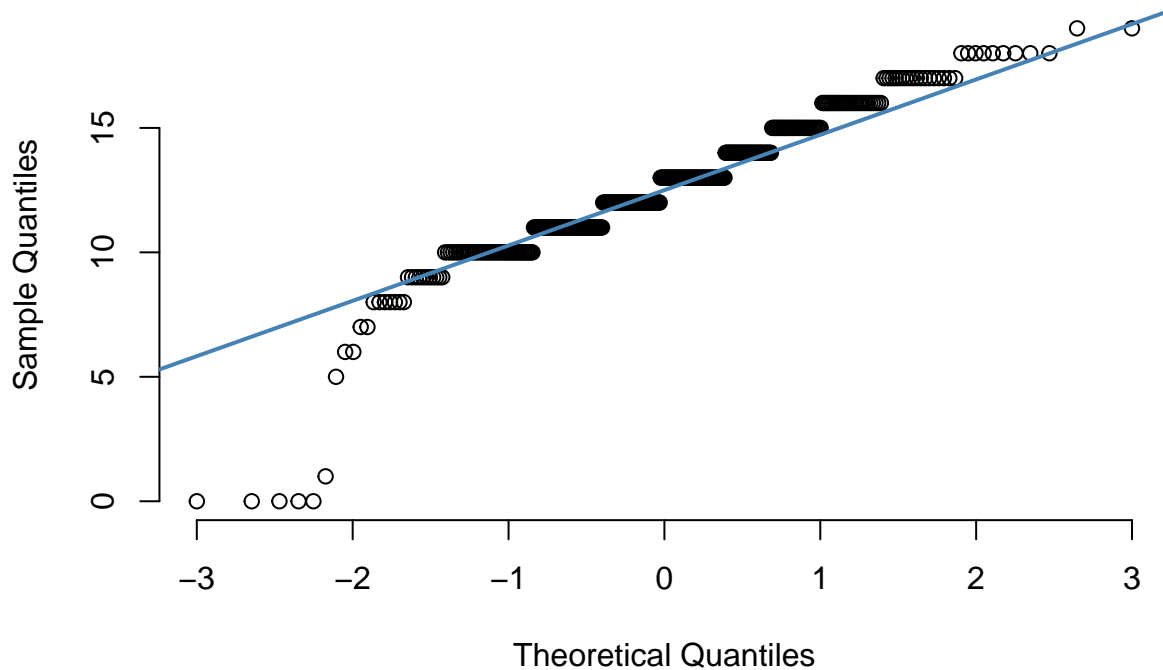


```
qqnorm(students_org$G3_mat, pch = 1, frame = FALSE, main = "Matematika")
qqline(students_org$G3_mat, col = "steelblue", lwd = 2)
```



```
qqnorm(students_org$G3_por, pch = 1, frame = FALSE, main = "Portugalski")
qqline(students_org$G3_por, col = "steelblue", lwd = 2)
```

## Portugalski



```
var(students_org$G3_mat)
```

```
## [1] 21.24131
```

```
var(students_org$G3_por)
```

```
## [1] 8.665092
```

```
var.test(students_org$G3_mat, students_org$G3_por)
```

```
##
```

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

```
##
```

```
## data: students_org$G3_mat and students_org$G3_por
```

```
## F = 2.4514, num df = 369, denom df = 369, p-value < 2.2e-16
```

```
## alternative hypothesis: true ratio of variances is not equal to 1
```

```
## 95 percent confidence interval:
```

```
## 1.998239 3.007245
```

```
## sample estimates:
```

```
## ratio of variances
```

```
## 2.451366
```

Zbog jako male vrijednosti odbacujemo hipotezu  $H_0$  da su varijance dva uzorka jednake.

```
t.test(students_org$G3_por, students_org$G3_mat, paired = TRUE, alternative = "greater",
       var.equal = FALSE)
```

```
##
```

```
## Paired t-test
```

```
##
```

```
## data: students_org$G3_por and students_org$G3_mat
```

```
## t = 9.8854, df = 369, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.740691      Inf
## sample estimates:
## mean of the differences
##                2.089189
```

Zbog jako male p-vrijednosti odbacujemo hipotezu  $H_0$  da su prosjeci ocjena jednaki u korist hipoteze  $H_1$  da je prosjek ocjena iz portugalskog značajno veći od prosjeka ocjena iz matematike.

## Matej Ciglencečki

### Kako vrijeme putovanja do škole utječe na uspjeh učenika?

Na ovo pitanje odgovorit ćemo ANOVA-om. Pretpostavke ANOVA-e su:

- nezavisnost pojedinih podataka u uzorcima
- normalna razdioba podataka
- homogenost varijanci među populacijama

Postavljamo hipotezu  $H_0$  koja glasi, srednja vrijednost grupa su podjednake.

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1 : \neg H_0$$

S obzirom da se radi o različitim školama i različitim predmetima možemo pretpostaviti nezavisnost ocjena.

Ukoliko nakon provedbe ANOVA-e odbacimo  $H_0$  hipotezu možemo zaključiti da su srednje vrijednosti međusobno različite, tj. da vrijeme putovanje utječe na uspjeh učenika.

### Obrada kategoričkih stupaca

Kao grupe koristiti će se vrijednosti iz stupca `traveltime` Prvo je potrebno pretvoriti stupac `traveltime` u kategoričke podatke (s poretkom). `traveltime` se sastoji od 4 mogućih vrijednosti koje definiraju potrebno vrijeme od škole do doma:

- < 15min
- 15 - 30 min
- 30 - 60 min
- > 60 min

Nadalje, zadnju kategoriju (60min+) spojiti ćemo sa predzadnjom kategorijom (30-60min) zbog toga što se u zadnjoj kategoriji nalaze samo 8 podataka dok se u preostalim kategorijama nalazi puno veći broj podataka.

```
count(students, students$traveltime)
```

```
## # A tibble: 4 x 2
##   `students$traveltime`     n
##             <dbl> <int>
## 1                 1   242
## 2                 2    99
## 3                 3    21
## 4                 4     8
```

```
students <- students_clean
students$traveltime <- factor(students$traveltime, ordered = TRUE, labels = c("0 - 15 min",
  "15 - 30 min", "> 30 min", "> 30 min"))
```

Za uspjeh koristiti ćemo zbog varijabli `G[1,2,3]_mat` i `G[1,2,3]_por` koji ćemo spremiti u novu varijablu `G_total`.

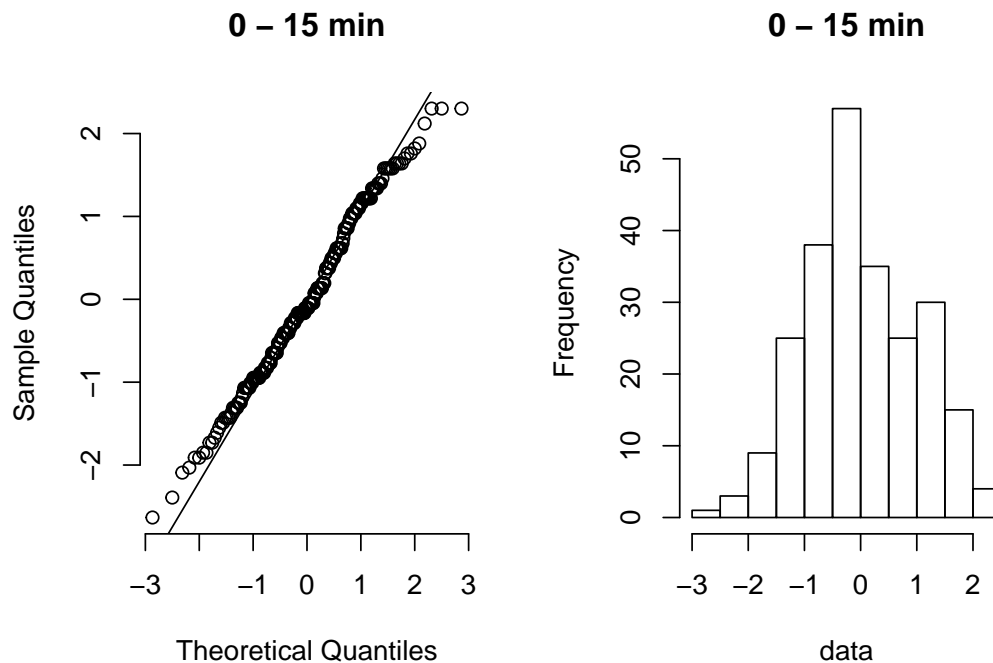
```
students$G3_total <- students$G3_mat + students$G3_por
students$G2_total <- students$G2_mat + students$G2_por
students$G1_total <- students$G1_mat + students$G1_por
students$G_total <- students$G1_total + students$G2_total + students$G3_total
```

ANOVA je robustna na blaga odstupanja što se tiče normalnosti. Svejedno, testirati ćemo normalnost varijable `G_total` nad cijelim podatkovnim skupom, a zatim nad `G_total` za svaku pojedinu grupu `traveltime`-a.

```
model = lm(students$G_total ~ students$traveltime)
```

```
par(mfrow = c(1, 2)) # 2 plots in 1 row
```

```
timeperiod = "0 - 15 min"
data <- rstandard(model)[students$traveltime == timeperiod]
qqnorm(data, pch = 1, frame = FALSE, main = timeperiod)
qqline(data)
hist(data, main = timeperiod)
```



```
lillie.test(data)["p.value"]
```

```
## $p.value
## [1] 0.008983716
```

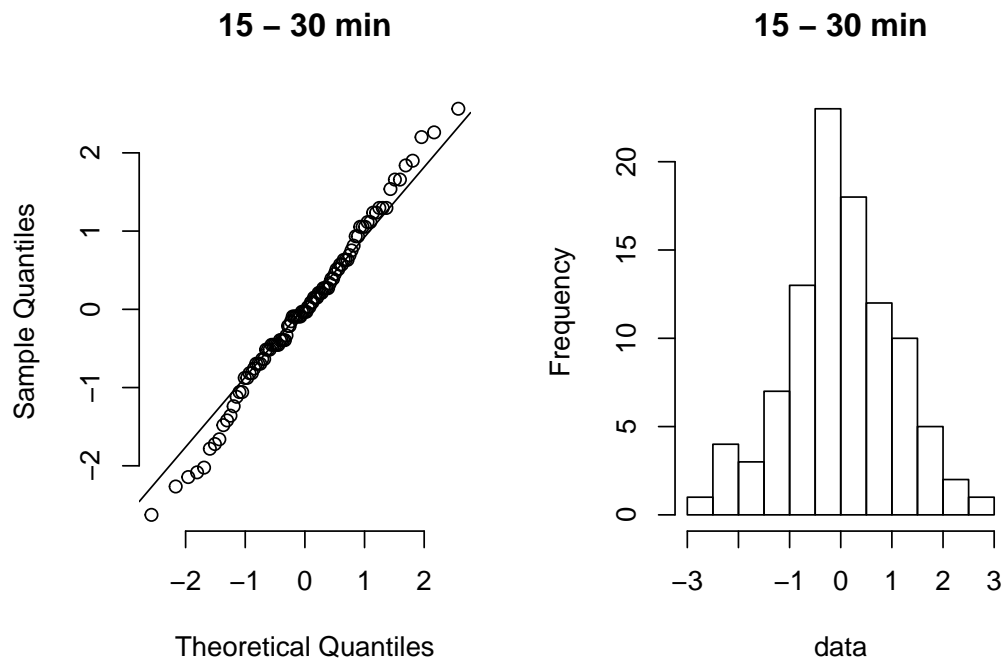
```
ks.test(data, "pnorm", mean = mean(data), sd = sd(data))["p.value"]
```

```
## Warning in ks.test(data, "pnorm", mean = mean(data), sd = sd(data)): ties should
## not be present for the Kolmogorov-Smirnov test
```

```
## $p.value
## [1] 0.2157153
```

```
timeperiod = "15 - 30 min"
data <- rstandard(model)[students$traveltime == timeperiod]
```

```
qqnorm(data, pch = 1, frame = FALSE, main = timeperiod)
qqline(data)
hist(data, main = timeperiod)
```



```
lillie.test(data)["p.value"]
```

```
## $p.value
## [1] 0.5782076
```

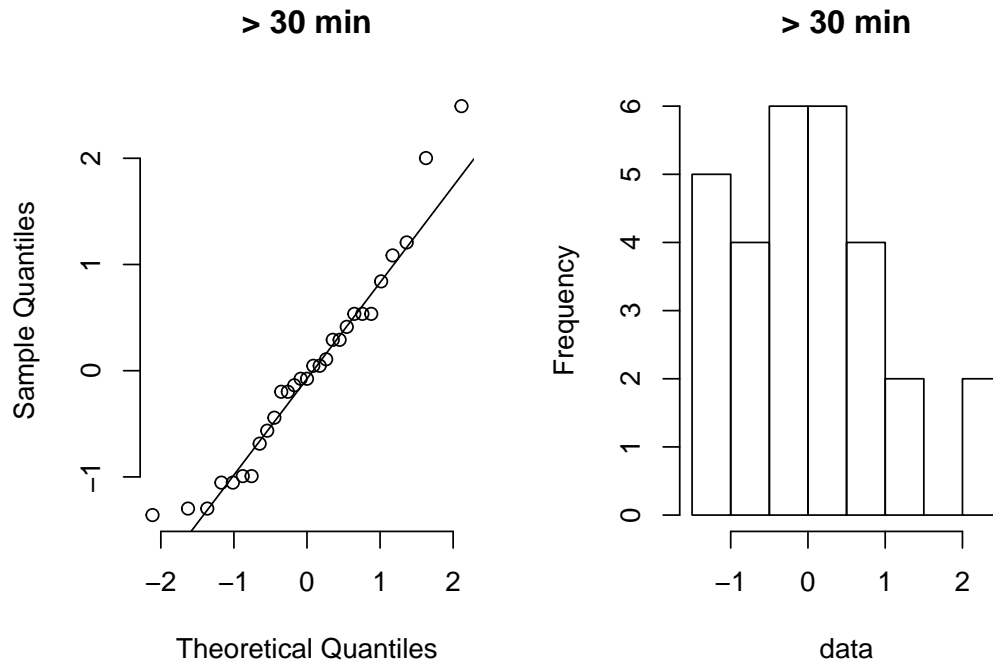
```
ks.test(data, "pnorm", mean = mean(data), sd = sd(data))["p.value"]
```

```
## Warning in ks.test(data, "pnorm", mean = mean(data), sd = sd(data)): ties should
## not be present for the Kolmogorov-Smirnov test
```

```
## $p.value
## [1] 0.897279
```

```
timeperiod = "> 30 min"
data <- rstandard(model)[students$traveltime == timeperiod]
qqnorm(data, pch = 1, frame = FALSE, main = timeperiod)
qqline(data)
hist(data, main = timeperiod)
```





```
lillie.test(data)["p.value"]
```

```
## $p.value
## [1] 0.4329395
```

```
ks.test(data, "pnorm", mean = mean(data), sd = sd(data))["p.value"]
```

```
## Warning in ks.test(data, "pnorm", mean = mean(data), sd = sd(data)): ties should
## not be present for the Kolmogorov-Smirnov test
```

```
## $p.value
## [1] 0.8440515
```

Na svakom grafu možemo vidjeti da podaci uglavnom prate normalnu distribuciju uz manji broj stršućih vrijednosti (lijevi rep). Nadalje, p vrijednosti Lillieforsovog testa nisu uvijek iznad 0.05 međutim za sve Kolmogorov-Smirnov testove p vrijednosti su iznad 0.05.

Lilliefors koristimo ako nam nije poznata varijanca i srednja vrijednost populacije, što je s ovim podacima i slučaj. Poznato je da Lilliefors konzervativniji i da odbacuje hipotezu  $H_0$  češće nego Kolmogorov-Smirnov.

S obzirom na manja odstupanja, ne toliko male p vrijednosti i grafički izgled `qqnorm`-a i histograma pretpostaviti ćemo da su podaci uzrokovani iz normalne distribucije.

## Homogenost varijanci - Bartlettov test

Prvo je potrebno postaviti hipoteze  $H_0$  i  $H_1$ :

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

$$H_1 : \neg H_0$$

```
var(students$G_total[students$traveltime == "> 30 min"])
```

```
## [1] 241.6897
```

```
var(students$G_total[students$traveltime == "15 - 30 min"])
```

```
## [1] 296.1703
```

```
var(students$G_total[students$traveltime == "> 30 min"])
```

```
## [1] 241.6897
```

```
bartlett.test(students$G_total ~ students$traveltime)
```

```
##
## Bartlett test of homogeneity of variances
##
## data:  students$G_total by students$traveltime
## Bartlett's K-squared = 0.48546, df = 2, p-value = 0.7845
```

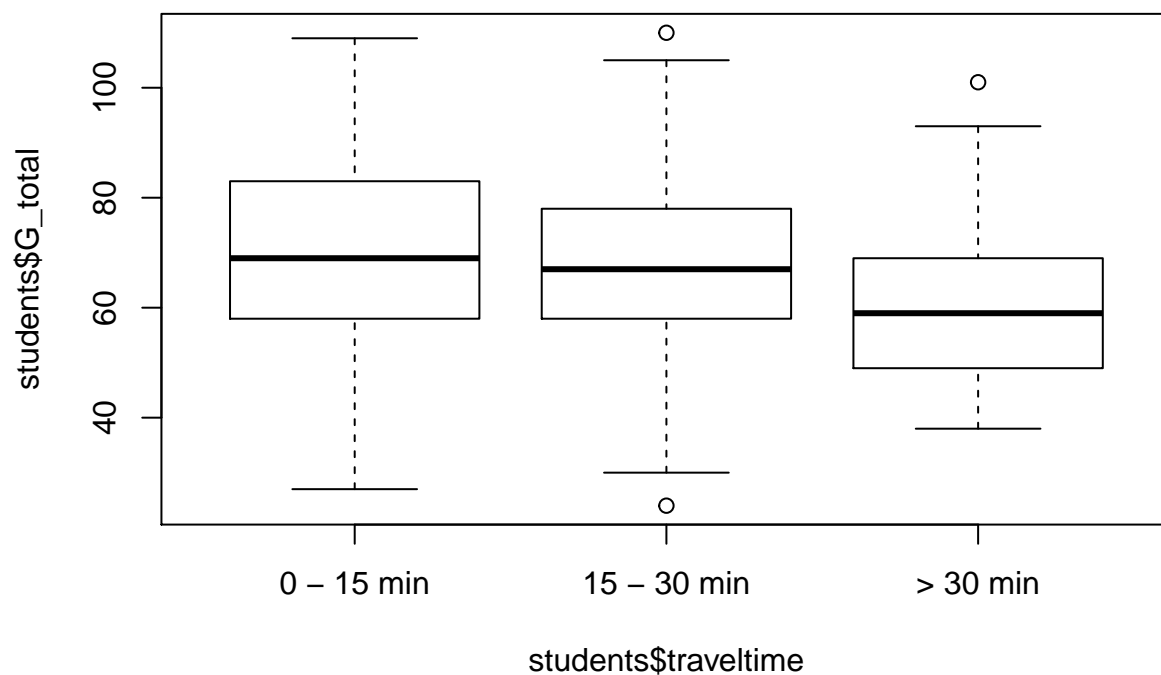
Vidimo da su vrijednosti varijance slične. S obzirom da je p vrijednost testa veća od 0.05 ne odbacujemo  $H_0$  čime zadovoljavamo ANOVA pretpostavku o homogenosti varijanca.

## ANOVA - Jesu li srednje vrijednosti za različite grupe drugačije?

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1 : \neg H_0$$

```
boxplot(students$G_total ~ students$traveltime)
```



Grafički možemo pretpostaviti da se vrijeme putovanja utječe na uspjeh učenika. Naravno, ANOVA-om je potrebno provjeriti koliko je ta razlika statistički značajna.

```
model = lm(students$G_total ~ students$traveltime)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: students$G_total
##              Df Sum Sq Mean Sq F value    Pr(>F)
## students$traveltime  2    3185   1592.35    5.7419 0.003504 **
## Residuals          367   101777    277.32
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA nam govori da postoji razlika između grupa `traveltime`. Iako nije strogo značajna i dalje se radi o značajnoj p vrijednosti koja se nalazi između 0.001 i 0.01. Možemo zaključiti da za različite grupe vremena putovanja imaju utjecaj na učenikov uspjeh.

## Koja škola je bolja u matematici a koja u portugalskom?

Na ovo pitanje odgovoriti ćemo provedbom t-testa koristeći 4 različita podatkovna skupa. Razdvajanje podatkovnog skupa na dvije škole (GP, MS) te na dva predmeta (matematika i portugalski) dobivamo sljedeće podatkovne skupove: `gp_mat`, `gp_por`, `ms_mat`, `ms_por`

```
# Show average grade for all schools
schools <- students %>%
  select("school") %>%
  distinct(.)
schools # [GP, MS]
subject_final_grade_names <- names(students)[grepl("G3", names(students))]

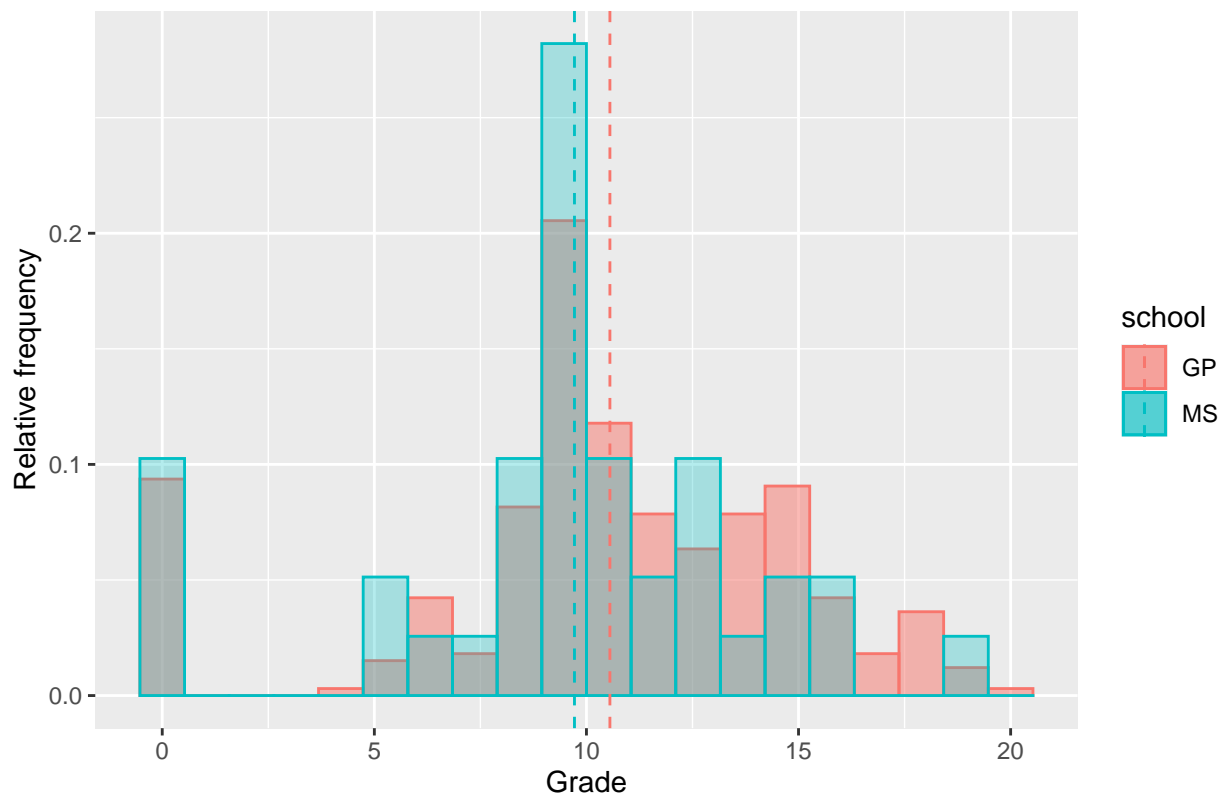
# all_of Note: Using an external vector in selections is ambiguous. Use
# `all_of(vars)` instead of `vars` to silence this message.
students_final_grade <- students %>%
  select("school", all_of(subject_final_grade_names))

# Select only the subject grade and school
gp_mat <- students_final_grade %>%
  filter(school == "GP") %>%
  select(G3_mat, school)
gp_por <- students_final_grade %>%
  filter(school == "GP") %>%
  select(G3_por, school)
ms_mat <- students_final_grade %>%
  filter(school == "MS") %>%
  select(G3_mat, school)
ms_por <- students_final_grade %>%
  filter(school == "MS") %>%
  select(G3_por, school)
```

## Prikaz relativnih frekvencija predmeta i škola

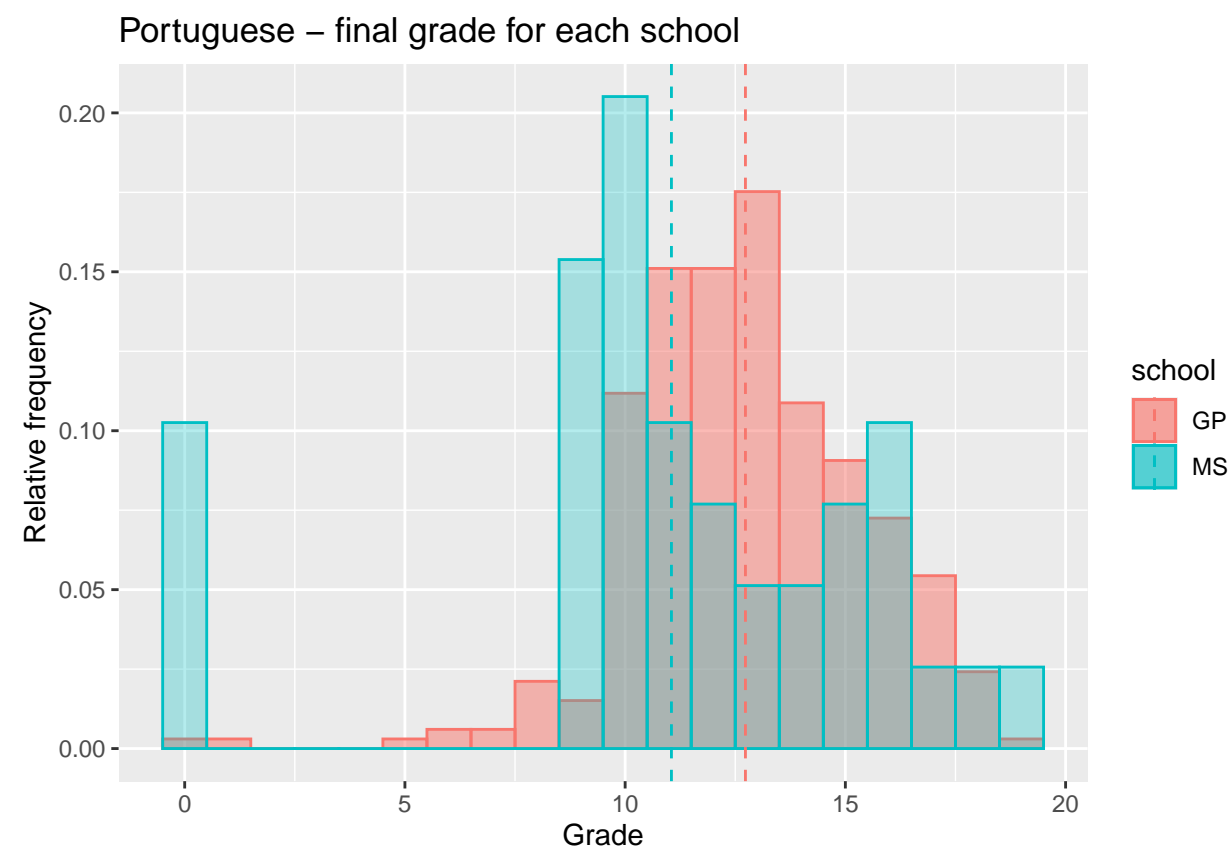
Iz grafa relativne frekvencije možemo usporediti vertikalne crte koje određuju srednju vrijednost ocjene za pojedinu školu i također dobiti osjećaj za normalnost podataka. Konstruirati ćemo jednosmjerni T-test a alternativa će ići u korist škole koja ima veću srednju vrijednost čime ćemo provjeriti je li ta škola statistički značajno bolja u matematici/portugalskom.

Matematika - prikaz relativnih frekvencija i srednjih vrijednosti  
 Mathematics – final grade for each school



Na grafu za matematiku vidi se da škola GP ima veću srednju vrijednost od škole MS

## Portugalski - prikaz relativnih frekvencija i srednjih vrijednosti



Na grafu za portugalski vidi se da škola GP ima veću srednju vrijednost od škole MS

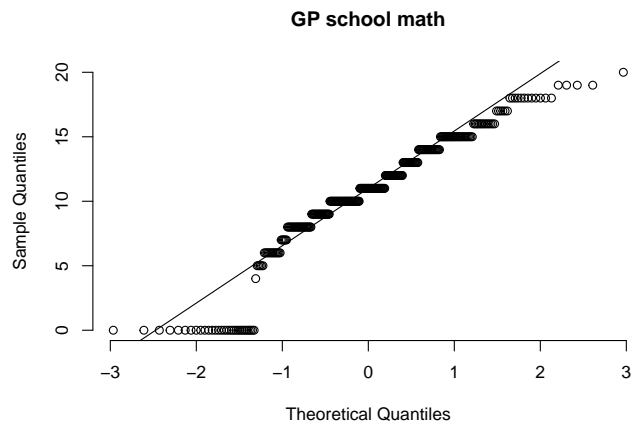
## Provjera normalnosti

Normalnost se provjerva na više načina. U sljedećim koracima biti će prikazani `qqnorm` grafovi i provedeni Lilliefors i Kolmogorov-Smirnov testovi na temelju kojih će se pretpostaviti (ne)normalnost.

```
nrow(gp_mat)
nrow(gp_por)
nrow(ms_mat)
nrow(ms_por)
```

n - broj podataka za matematiku je 331 a za portugalski 39

```
qqnorm(gp_mat$grade, pch = 1, frame = FALSE, main = "GP school math")
qqline(gp_mat$grade)
```



```
lillie.test(gp_mat$grade)["p.value"]
```

```
## $p.value
## [1] 7.814771e-14
```

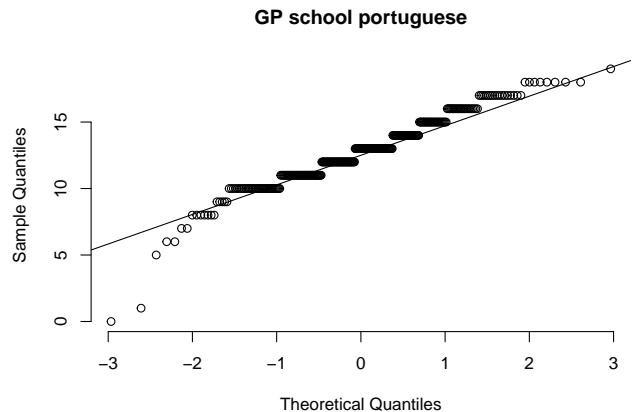
```
ks.test(gp_mat$grade, "pnorm", mean(gp_mat$grade), sd(gp_mat$grade))["p.value"]
```

```
## Warning in ks.test(gp_mat$grade, "pnorm", mean(gp_mat$grade), sd(gp_mat$grade)):
```

```
## ties should not be present for the Kolmogorov-Smirnov test
```

```
## $p.value
## [1] 5.330255e-05
```

```
qqnorm(gp_por$grade, pch = 1, frame = FALSE, main = "GP school portuguese")
qqline(gp_por$grade)
```



```
lillie.test(gp_por$grade)["p.value"]
```

```
## $p.value
## [1] 1.673428e-09
```

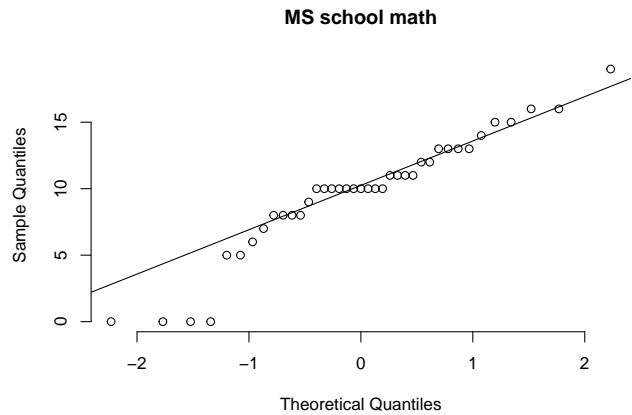
```
ks.test(gp_por$grade, "pnorm", mean(gp_por$grade), sd(gp_por$grade))["p.value"]
```

```
## Warning in ks.test(gp_por$grade, "pnorm", mean(gp_por$grade), sd(gp_por$grade)):
```

```
## ties should not be present for the Kolmogorov-Smirnov test
```

```
## $p.value
## [1] 0.001247681
```

```
qqnorm(ms_mat$grade, pch = 1, frame = FALSE, main = "MS school math")
qqline(ms_mat$grade)
```



```
lillie.test(ms_mat$grade)["p.value"]
```

```
## $p.value
## [1] 0.0009170632
```

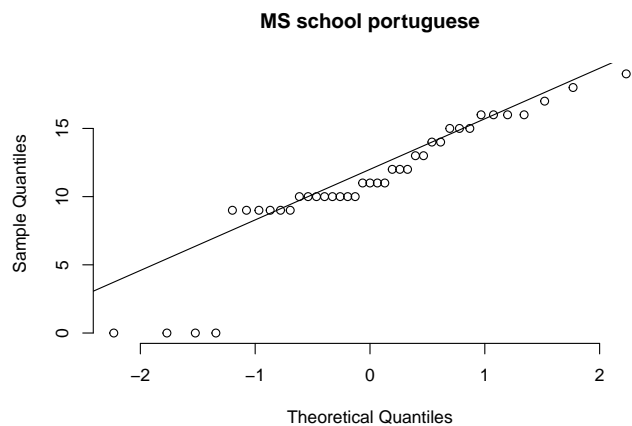
```
ks.test(ms_mat$grade, "pnorm", mean(ms_mat$grade), sd(ms_mat$grade))["p.value"]
```

```
## Warning in ks.test(ms_mat$grade, "pnorm", mean(ms_mat$grade), sd(ms_mat$grade)):
```

```
## ties should not be present for the Kolmogorov-Smirnov test
```

```
## $p.value
## [1] 0.1131777
```

```
qqnorm(ms_por$grade, pch = 1, frame = FALSE, main = "MS school portuguese")
qqline(ms_por$grade)
```



```
lillie.test(ms_por$grade)["p.value"]
```

```
## $p.value
## [1] 1.951046e-05
```

```
ks.test(ms_por$grade, "pnorm", mean(ms_por$grade), sd(ms_por$grade))["p.value"]
```

```
## Warning in ks.test(ms_por$grade, "pnorm", mean(ms_por$grade), sd(ms_por$grade)):
```

```
## ties should not be present for the Kolmogorov-Smirnov test
```

```
## $p.value
## [1] 0.03355273
```

Repovi su prisutni na lijevoj strani podataka zbog čega je p vrijednost skoro uvijek manja od 0.05 za Kolmogorov-Smirnov i Lillieforsov test. Grafički, na temelju rezultata određujemo da za sve skupove vrijedi

da proizlaze iz normalne distribucije ali s opaskom da postoje stršeće vrijednosti na lijevoj strani distribucije.

## F-test - test o jednakosti varijanca

Važno je napomenuti da je test o varijanci iznimno osjetljiv na normalnost. Test će biti proveden zbog vježbe ali njegov **rezultat se neće uzeti u obzir** jer podaci nisu normalno distribuirani.

$p$  – vjerojatnost da pod  $H_0$  dobijemo vrijednost koja je jednako ili više ekstremna nego vrijednost koji bi dobili izračunom iz uzorka kojeg imamo

Ako je  $p < \alpha$ , odbacujemo hipotezu  $H_0$  u korist hipoteze  $H_1$ :

- pada u desni ili lijevi rep => odbacivanje

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \neg H_0$$

Poredak argumenata za `var.test` nije bitna ali generalno vrijedi:

$$\frac{\sigma_1^2}{\sigma_2^2}, \quad \sigma_1^2 > \sigma_2^2$$

```
cat("Mathematics variances", var(gp_mat$grade), var(ms_mat$grade))
```

```
## Mathematics variances 21.38735 19.89204
```

```
cat("Portugeuse variances", var(gp_por$grade), var(ms_por$grade))
```

```
## Portugeuse variances 6.839605 22.1552
```

Na prvi pogled čini se da će  $H_0$  hipoteza za portugalski biti odbačena zbog toga što su varijance značajno drugačije. Potrebno je provesti f-test da se uvjerimo da se radi o statistički značajnoj razlici varijanci.

Konstruirajmo i provedimo testove o varijanci:

```
alpha <- 0.05
```

```
# H0 - Variance of GP_MAT and MS_MAT are equal H1 - not H0
```

```
mat_f_test <- var.test(gp_mat$grade, ms_mat$grade, alternative = "two.sided") # F = 1.0752, p = 0.817
```

```
# H0 - Variance of GP_POR and MS_MAT are equal H1 - not H0
```

```
por_f_test <- var.test(gp_por$grade, ms_por$grade, alternative = "two.sided") # F = 0.30871, p = 1.217
```

```
var_equal_mat <- if (mat_f_test$p.value < alpha) FALSE else TRUE
```

```
cat_reject_h0("Matematika - test o jednakosti varijanca:", !var_equal_mat)
```

```
## Matematika - test o jednakosti varijanca:
```

```
## Ne odbacujemo hipotezu H0
```

```
var_equal_por <- if (por_f_test$p.value < alpha) FALSE else TRUE
```

```
cat_reject_h0("Portugalski - test o jednakosti varijanca:", !var_equal_por)
```

```
## Portugalski - test o jednakosti varijanca:
```

```
## Odbacujemo hipotezu H0 u korist hipoteze H1
```



## T-test - testiranje jednakosti srednje vrijednosti ocjena za dvije škole uz nepoznate varijance

Uz to što je  $n$  veći od 30 za oba podatkovna skupa i uz činjenicu da je t-test robustan na (ne)normalnost provodimo t-test srednje vrijednosti za oba predmeta.

Zbog prethodno dobivenih srednje vrijednosti o ocjenama (koje idu u korist škole GP) postavljena je jednosmjerna alternativna hipoteza.

Ponovno, zbog toga što test o varijanci nije robustan na nenormalnost pretpostaviti ćemo da varijance uzoraka nisu jednake.

```
# H0 - GP school has equal grades to in mathematics to MS (GP=MS) H1 - GP>MS
mat_t_test <- t.test(gp_mat$grade, ms_mat$grade, alt = "greater", var.equal = FALSE)
is_gp_mat_better <- if (mat_t_test$p.value < alpha) TRUE else FALSE
cat_reject_h0("Matematika - t-test:", is_gp_mat_better)
## Matematika - t-test:
## Ne odbacujemo hipotezu H0

# H0 - GP school has equal grades to in Portuguese to MS (GP=MS) H1 - GP>MS
por_t_test <- t.test(gp_por$grade, ms_por$grade, alt = "greater", var.equal = FALSE)
is_gp_por_better <- if (por_t_test$p.value < alpha) TRUE else FALSE
cat_reject_h0("Portugalski t-test:", is_gp_por_better)
## Portugalski t-test:
## Odbacujemo hipotezu H0 u korist hipoteze H1
```

Za matematiku, nismo odbacili hipotezu  $H_0$  i zbog čega ne možemo zaključiti da škola GP ima bolje ocjene iz matematike od škole MS.

Za portugalski, odbacujemo hipotezu  $H_0$  u korist hipoteze  $H_1$  i zaključujemo da je škola GP ima bolje ocjene iz portugalskog od škole MS.