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 nevažeći 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 minumum 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
```

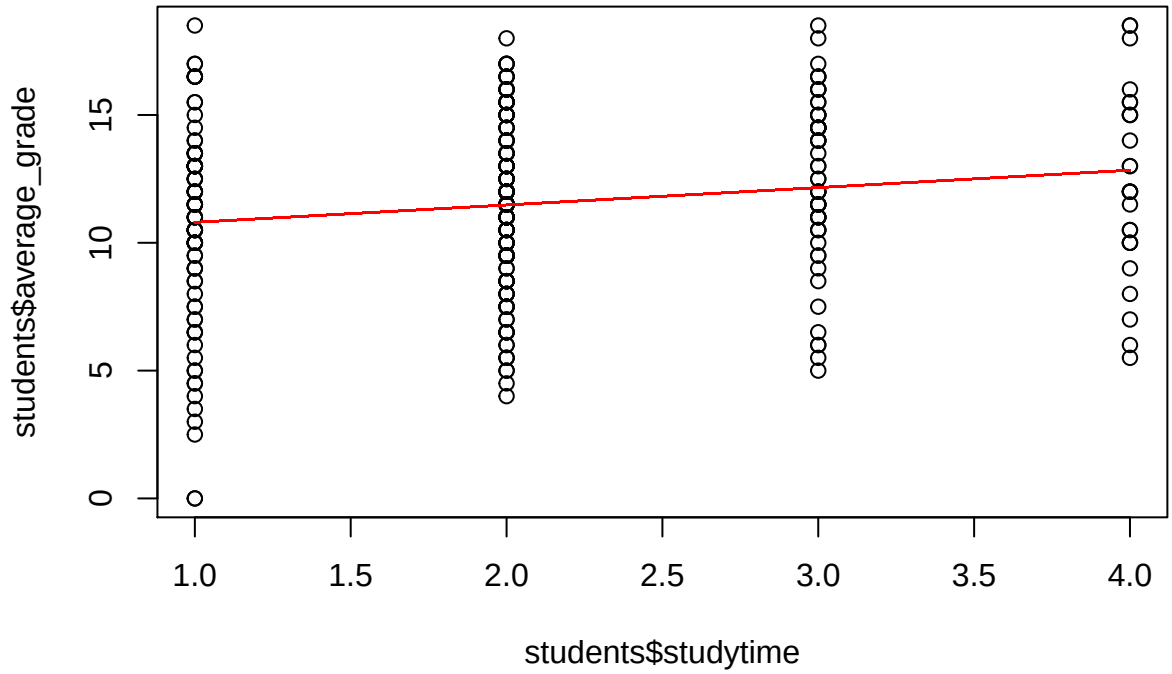
Izbacivanje svih NaN/NA/null vrijednosti iz podatkovnog skupa. Na sreću, takvih vrijednosti nije bilo.

```
# 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

```
students$average_grade <- (students$G3_mat + students$G3_por)/2

fit.studytime = lm(average_grade ~ studytime, data = students)
plot(students$studytime, students$average_grade)
lines(students$studytime, fit.studytime$fitted.values, col = "red")
```




```
# Pearsonov korelacijski koeficijent
cor(students$studytime, students$average_grade)
```

```
## [1] 0.175217
```

```
cor.test(students$studytime, students$average_grade)
```

```
##
##  Pearson's product-moment correlation
##
## data:  students$studytime and students$average_grade
## t = 3.4141, df = 368, p-value = 0.0007113
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07459575 0.27230630
## sample estimates:
##      cor
```

```
## 0.175217
```

```
summary(fit.studytime)
```

```
##
## Call:
## lm(formula = average_grade ~ studytime, data = students)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8003  -1.9801   0.0199   2.1997   7.6997
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.1205     0.4404  22.981  < 2e-16 ***
## studytime     0.6798     0.1991   3.414 0.000711 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.243 on 368 degrees of freedom
## Multiple R-squared:  0.0307, Adjusted R-squared:  0.02807
## F-statistic: 11.66 on 1 and 368 DF,  p-value: 0.0007113
```

```
# procjena modela s dummy varijablama
students.d = dummy_cols(students, select_columns = "studytime")

# procjena modela s dummy varijablama
fit.multi.d = lm(average_grade ~ studytime_1 + studytime_2, students.d)
summary(fit.multi.d)
```

```
##
## Call:
## lm(formula = average_grade ~ studytime_1 + studytime_2, data = students.d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7653  -1.9432   0.0568   2.2347   7.7347
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.4885     0.3475  35.942  < 2e-16 ***
## studytime_1  -1.7232     0.4774  -3.610 0.000349 ***
## studytime_2  -1.0453     0.4213  -2.481 0.013550 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.241 on 367 degrees of freedom
## Multiple R-squared:  0.03468,    Adjusted R-squared:  0.02942
## F-statistic: 6.592 on 2 and 367 DF,  p-value: 0.00154
```

Tomislav Prhat

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

```
students_org %>%
    summarise(Mean.G1_mat = mean(G1_mat), Mean.G2_mat = mean(G2_mat), Mean.G3_mat = mean(G3_mat),
```

```
        Mean.G1_por = mean(G1_por), Mean.G2_por = mean(G2_por), Mean.G3_por = mean(G3_por),
        ) -> summary.result1
summary.result1
```

```
## # A tibble: 1 x 6
##   Mean.G1_mat Mean.G2_mat Mean.G3_mat Mean.G1_por Mean.G2_por Mean.G3_por
##         <dbl>       <dbl>       <dbl>       <dbl>       <dbl>       <dbl>
## 1        10.9        10.8        10.5        12.1        12.3        12.6
```

```
students_org %>%
    summarise(Med.G1_mat = median(G1_mat), Med.G2_mat = median(G2_mat), Med.G3_mat = median(G3_mat),
        Med.G1_por = median(G1_por), Med.G2_por = median(G2_por), Med.G3_por = median(G3_por),
        ) -> summary.result2
summary.result2
```

```
## # A tibble: 1 x 6
##   Med.G1_mat Med.G2_mat Med.G3_mat Med.G1_por Med.G2_por Med.G3_por
##        <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1         11         11         11         12         12         13
```

```
students_org %>%
    summarise(Mean.G1_mat = mean(G1_mat, trim = 0.1), Mean.G2_mat = mean(G2_mat,
        trim = 0.1), Mean.G3_mat = mean(G3_mat, trim = 0.1), Mean.G1_por = mean(G1_por,
        trim = 0.1), Mean.G2_por = mean(G2_por, trim = 0.1), Mean.G3_por = mean(G3_por,
        trim = 0.1), ) -> summary.result3
summary.result3
```

```
## # A tibble: 1 x 6
##   Mean.G1_mat Mean.G2_mat Mean.G3_mat Mean.G1_por Mean.G2_por Mean.G3_por
##         <dbl>       <dbl>       <dbl>       <dbl>       <dbl>       <dbl>
## 1        10.8        10.9        10.9        12.1        12.2        12.6
```

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

```
##   Mean.G1_mat Mean.G2_mat Mean.G3_mat Mean.G1_por Mean.G2_por Mean.G3_por
## 1    1.085608    -1.08723   -4.016012   0.1670379    0.715859  -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.G1_mat = IQR(G1_mat), IQR.G2_mat = IQR(G2_mat), IQR.G3_mat = IQR(G3_mat),
        IQR.G1_por = IQR(G1_por), IQR.G2_por = IQR(G2_por), IQR.G3_por = IQR(G3_por),
        ) -> summary.result4
summary.result4
```

```
## # A tibble: 1 x 6
##   IQR.G1_mat IQR.G2_mat IQR.G3_mat IQR.G1_por IQR.G2_por IQR.G3_por
##        <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1          5          4          6          4          3          3
```

```
students_org %>%
    summarise(Var.G1_mat = var(G1_mat), Var.G2_mat = var(G2_mat), Var.G3_mat = var(G3_mat),
        Var.G1_por = var(G1_por), Var.G2_por = var(G2_por), Var.G3_por = var(G3_por),
        ) -> summary.result5
summary.result5
```

```
## # A tibble: 1 x 6
##   Var.G1_mat Var.G2_mat Var.G3_mat Var.G1_por Var.G2_por Var.G3_por
##        <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1       11.2       14.4       21.2       6.51       6.08       8.67
```

```
students_org %>%
    summarise(sd.G1_mat = sd(G1_mat), sd.G2_mat = sd(G2_mat), sd.G3_mat = sd(G3_mat),
        sd.G1_por = sd(G1_por), sd.G2_por = sd(G2_por), sd.G3_por = sd(G3_por), ) ->
    summary.result6
summary.result6
```

```
## # A tibble: 1 x 6
##   sd.G1_mat sd.G2_mat sd.G3_mat sd.G1_por sd.G2_por sd.G3_por
##       <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1      3.35      3.80      4.61      2.55      2.47      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.

Matej Ciglenečki

Kako vrijeme putovanja do škole utječe na uspjeh učenika?

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

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

Postavljamo hipotezu H0 koja glasi, srednja vrijednost grupa su podjednake.

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

$$H_1 : \neg H_0$$

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

Ukoliko nakon provedbe ANOVA-e odbacimo H0 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
```

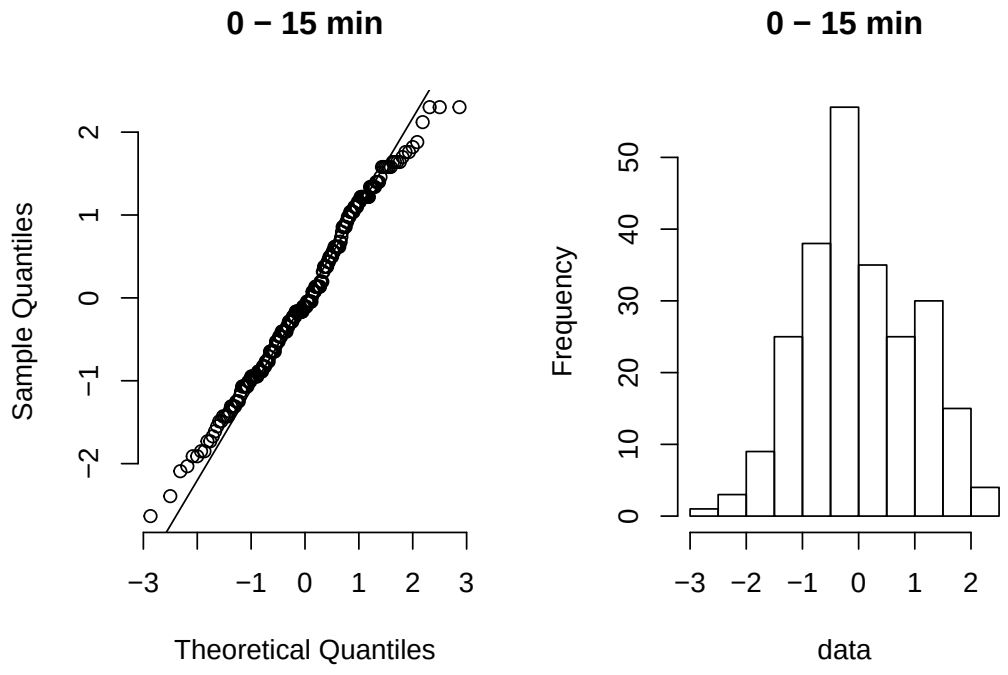
```
students <- students_clean
students$traveltime <- factor(students$traveltime, ordered = TRUE, labels = c("0 - 15 min",
    "15 - 30 min", "> 30 min", "> 30 min"))
```

Za uspjeh koristiti ćemo zboj varijabli `G[1,2,3]_mat` i `G[1,2,3]_por` koji ćemo spremitit 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"]
```

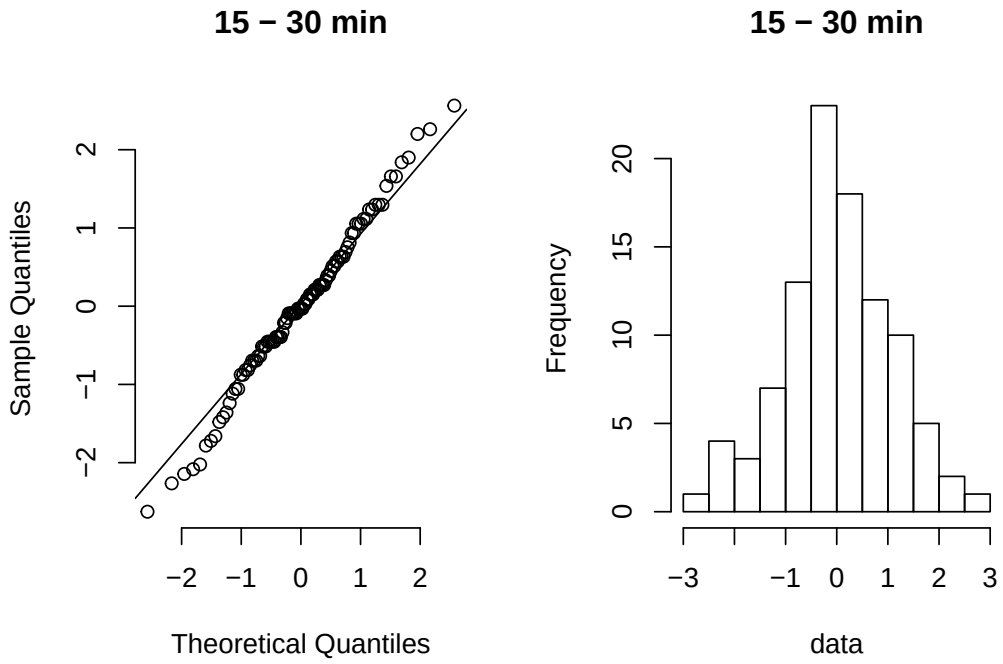
```
## $p.value
## [1] 0.008983716
```

```
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"]
```

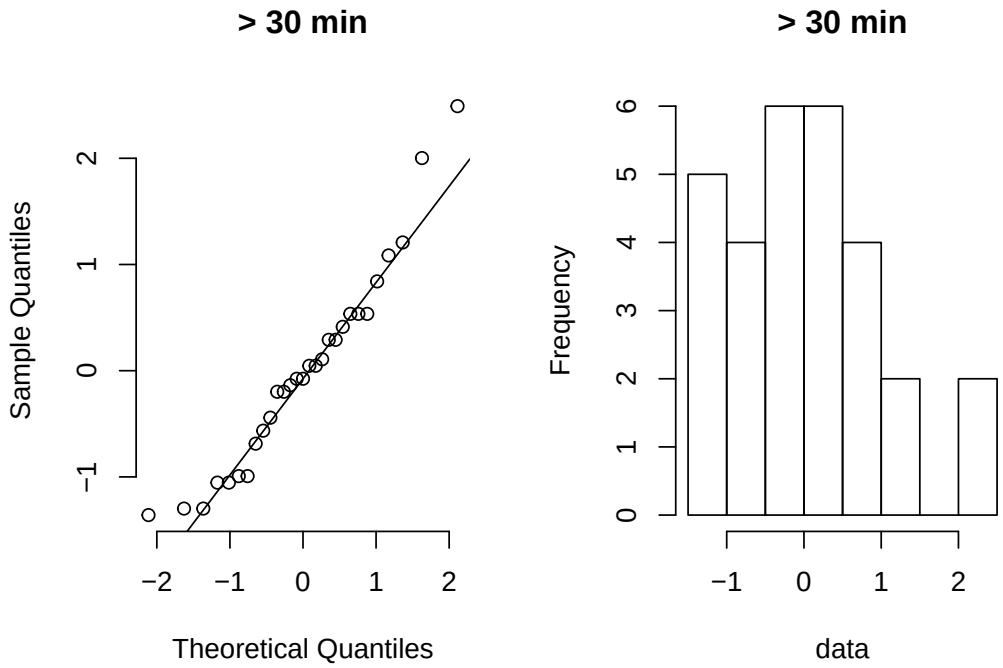
```
## $p.value
## [1] 0.5782076
```

```
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 stršećih 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 H0 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 H0 i H1:

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

$$H_1 : \neg H_0$$

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

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

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

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

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

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

```r
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
```

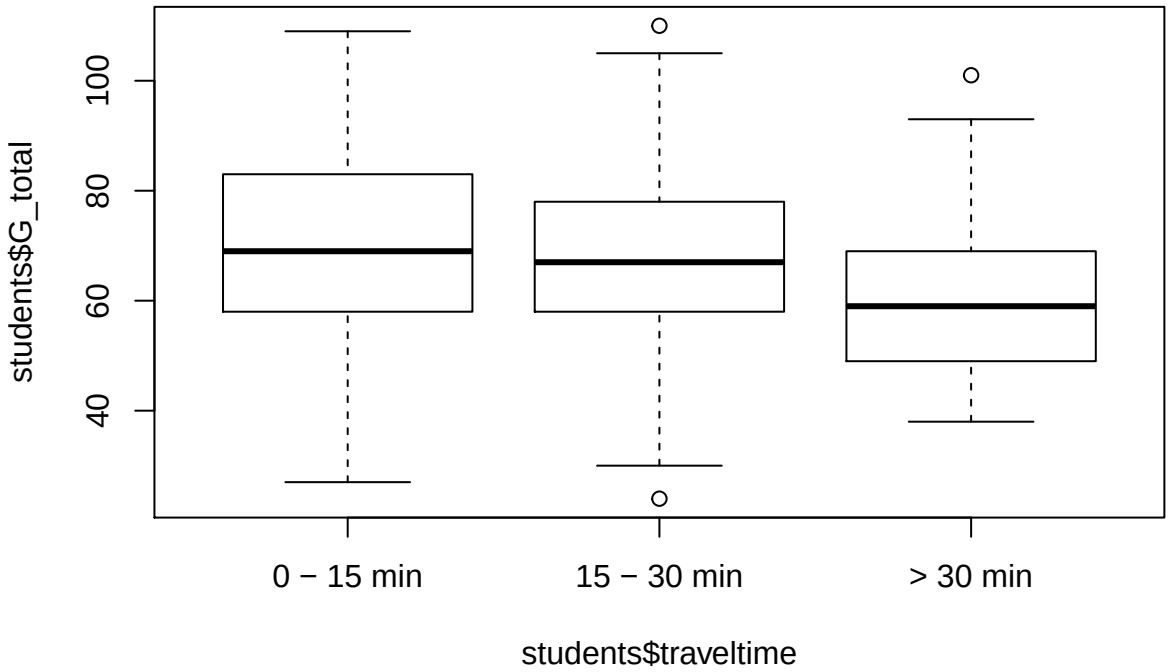
Vidimo da su vrijednosti varijance slične. S obzirom da je `p` vrijednost testa veća od 0.05 ne odbacujemo H0 čime zadovoljavamo ANOVA pretpostavku o homogenosti varijanca.

ANOVA - Jesu li srednje vrijednosti za različite grupe drugačije?

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

$$H_1 : \neg H_0$$

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



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

```r
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?

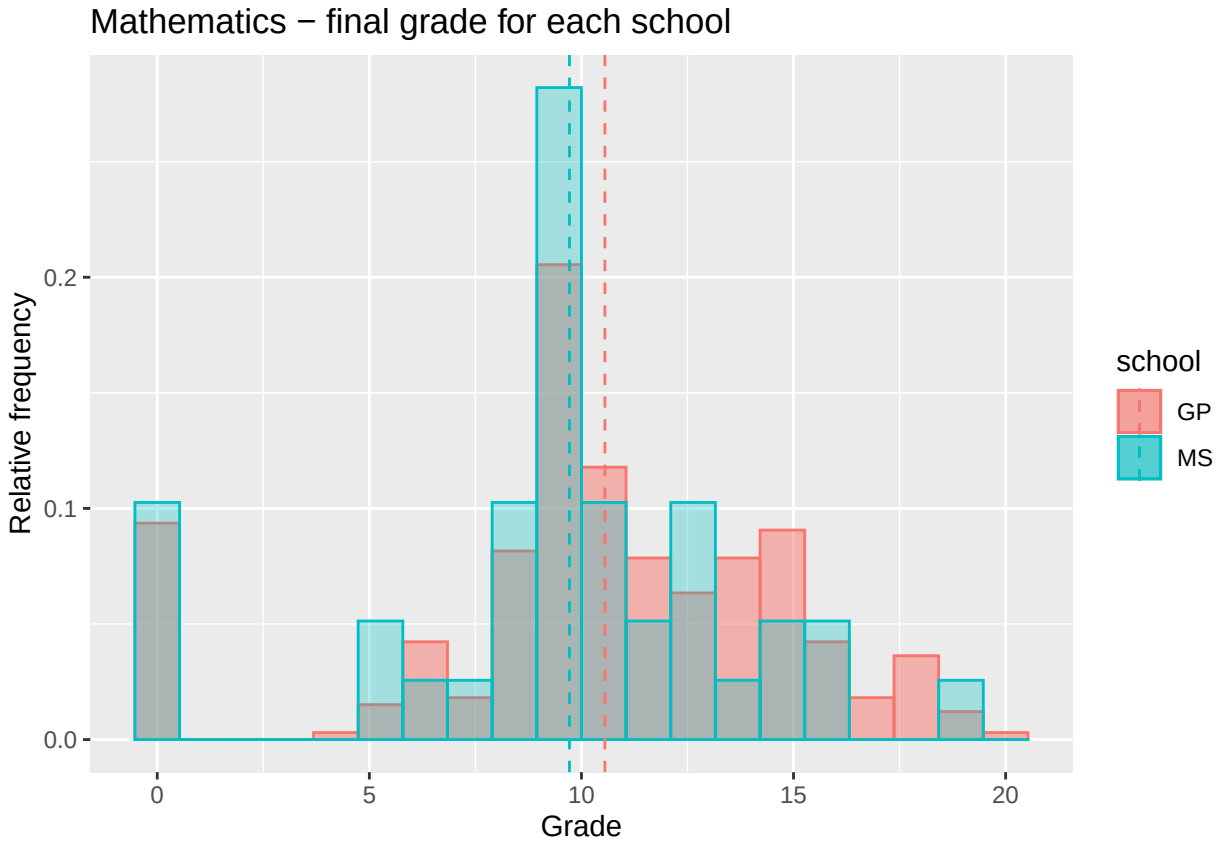
Na ovo pitanje odgovoriti ćemo provedbom t-testa koristeći 4 različita podatkovna skupa. Razdvajanje podatkovnog skupa na dvije škole (GP, MS) te na dva predmeta (matematika i portugalski) dobivamo sljedeće podatkovne skupove: `gp_mat`, `gp_por`, `ms_mat`, `ms_por`

```
# 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

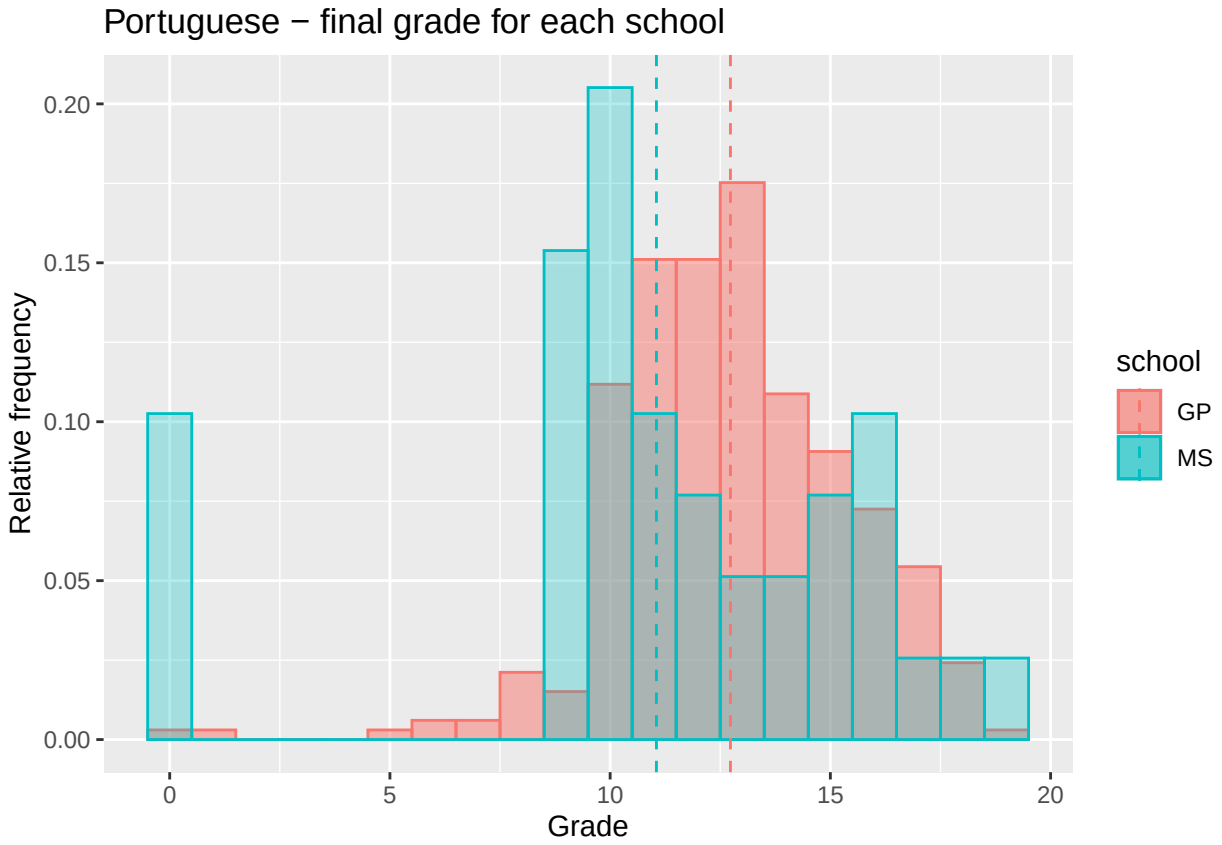
Iz grafa relativne frekvencije možemo usporediti vertikalne crte koje određuju srednju vrijednost ocjene za pojedinu školu i također dobiti osjećaj za normalnost podataka. Konstruirati ćemo jednosmjerni T-test a alternativa će ići u korist škole koja ima veću srednju vrijednost čime ćemo provjeriti je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

Matematika - prikaz relativnih frekvencija i srednjih vrijednosti



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`

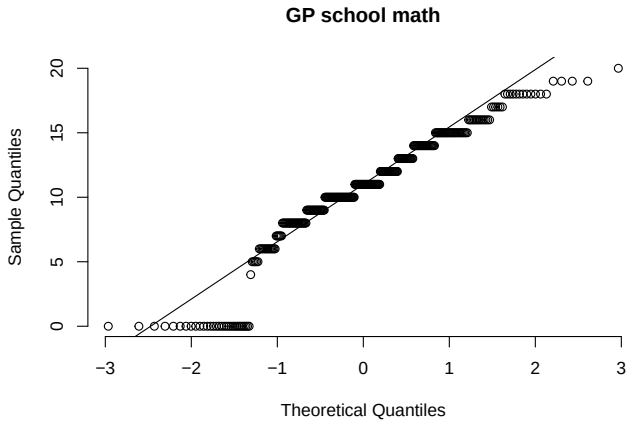
Provjera normalnosti

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"]
```

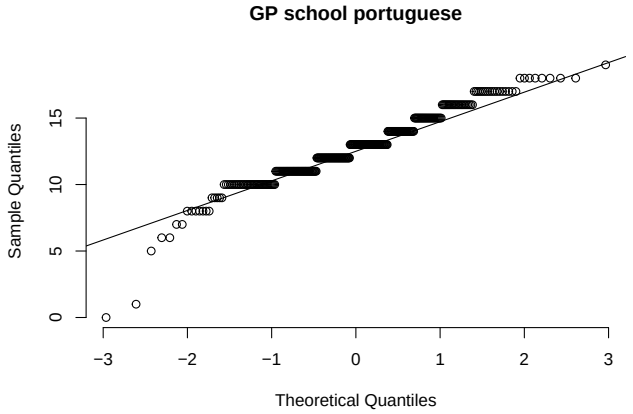
```
## $p.value
## [1] 7.814771e-14
```

```
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"]
```

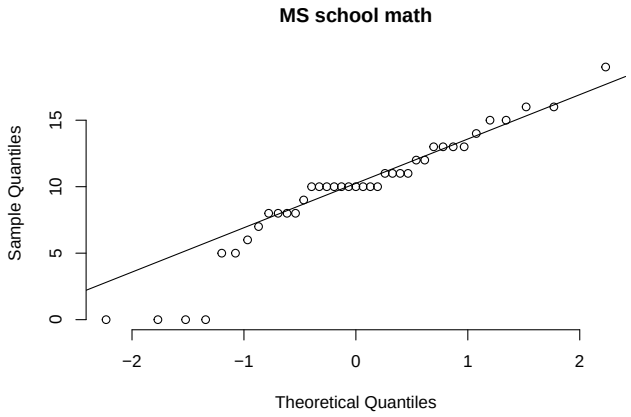
```
## $p.value
## [1] 1.673428e-09
```

```
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"]
```

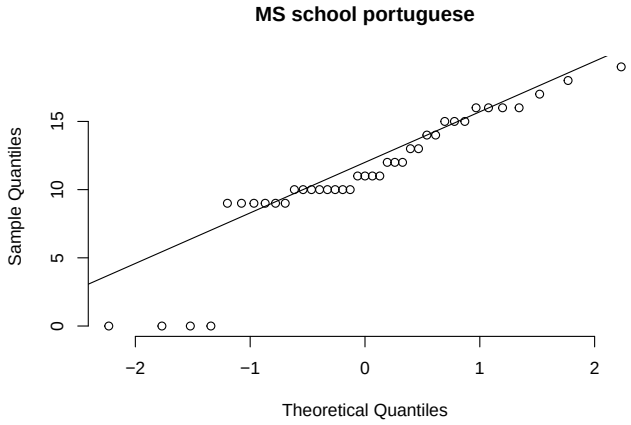
```
## $p.value
## [1] 0.0009170632
```

```
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 ordeđujemo da za sve skupove vrijedi

da proizlaze iz normalne distribucije ali s opaskom da postoje stršeće vrijednosti na lijeovj 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 H0 dobijemo vrijednost koja je jednako ili više ekstremna nego vrijednost koji bi dobili izračunom iz uzorka kojeg imamo

Ako je $p < \alpha$, odbacujemo hipotezu $H0$ u korist hipoteze $H1$:

- 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 H0 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 alternativa hipoteza.

Ponovno, zbog toga što test o varijanci nije robustan na nenormalnost pretpostaviti ćemo da vraijance 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 H0 i zbog čega ne možemo zaključiti da škola GP ima bolje ocjene iz matematike od škole MS.

Za portugalski, odbacujemo hipotezu H0 u korist hipoteze H1 i zaključujemo da je škola `GP` ima bolje ocjene iz portgualskog od škole `MS`.