Student success analyisis

Statistical data analysis project report

Matej Ciglenečki Petar Dragojević Magda Radić Tomislav Prhat

Abstract

The main goal of the project was to create a report. The report had to contain clearly explained concepts of statistical data analysis applied to the existing dataset. There weren't restrictions for the selection of statistical methods, as long as they were applied to the appropriate context and covered in the course's curriculum. The report consists of the test cases selected from the recommended list of test cases provided by faculty personnel or cases made up by team members. One of the requirements was that the R language had to be used to analyze the data and generate the report. The project's grade depended on the quality of the report and oral examination done by faculty personnel which examined the knowledge of the theory of the methods used (why you may or may not use a particular test, assumptions for tests used, details of the statistical methods, etc.). The dataset includes answers to survey questions with grades in mathematics and Portuguese of two students high schools. Collecting data on student achievement in teaching is a prerequisite for analyzing and improving the quality of education system. Details of the dataset are located at pdfs/dataset_documentation.pdf

Grade: 38/40

Contents

Descriptive analysis

Load the data, check dimension, columns, head, and summary

```
students_org <- readxl::read_excel("student_data.xlsx")
dim(students_org)
```

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

```
names(students_org)  # column names
```

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

```
head(students_org[1:7], n = 3)  # Show first 3 rows and first 7 columns
```

```
## # A tibble: 3 x 7
##   school sex     age address famsize Pstatus  Medu
##   <chr>  <chr> <dbl> <chr>   <chr>   <chr>   <dbl>
## 1 GP     F        18 U       GT3     A           4
## 2 GP     F        17 U       GT3     T           1
## 3 GP     F        15 U       LE3     T           1
```

```
summary(students_org[1:4])  # Show details for first 4 columns
```

```
##     school              sex                 age          address
##  Length:370         Length:370         Min.   :15.00   Length:370
##  Class :character   Class :character   1st Qu.:16.00   Class :character
##  Mode  :character   Mode  :character   Median :17.00   Mode  :character
##                                        Mean   :16.58
##                                        3rd Qu.:17.00
##                                        Max.   :22.00
```

Find what's the type of columns: numerical, characters...

```
cat("Numeric columns:", colnames(students_org %>%
    select(where(is.numeric))), fill = TRUE)
```

```
## Numeric columns: age Medu Fedu traveltime studytime failures_mat failures_por
## famrel freetime goout Dalc Walc health absences_mat absences_por G1_mat G2_mat
## G3_mat G1_por G2_por G3_por
```

```
cat("Character columns:", colnames(students_org %>%
    select(where(is_character))), fill = TRUE)
```

```
## Character columns: school sex address famsize Pstatus Mjob Fjob reason guardian
## schoolsup famsup paid_mat paid_por activities nursery higher internet romantic
```

```
# sapply(students_org, class)
```

Checking for invalid data. For example, does the data exceed maximal value specified in the dataset documentation? (it doesn't)

```
colMax <- students_org %>%
    select(where(is.numeric)) %>%
    sapply(., max, na.rm = TRUE)
```

Values of each column do not exceed values specified in the dataset documentation

Removing NaN/NA/null values from the dataset. Luckily, there were no such values.

```
# 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 an another dataset We will continue using
# 'student' variable
students <- students_org %>%
    filter_all(all_vars(!is.na(.) & !is.nan(.) & !is.null(.)))
students_clean <- students
```

Test case: Is parent's education independent from students' success?

author: Petar Dragojević - advised by the rest of the group

Chi-squared test

https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/#:~:text=The%20assumptions%20of%20the%20Chi,the%20variable

The Chi-square statistic is a non-parametric (distribution of the data doesn't matter) test designed to analyze group differences. It's applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance.

A test of independence assesses whether observations consisting of measures on two variables, expressed in a contingency table, are independent of each other (e.g. polling responses from people of different nationalities to see if one's nationality is related to the response).

Chi-squared test assumptions:

- Sample size not less than 50 for a 2x2 contingency table - by using Chi-Squared test on small samples, might end up committing a Type II error
- Expected cell count should be 5 or more for 80% of the cells
- The observations are always assumed to be independent of each other

First, transforming grades to the American grading system is performed:

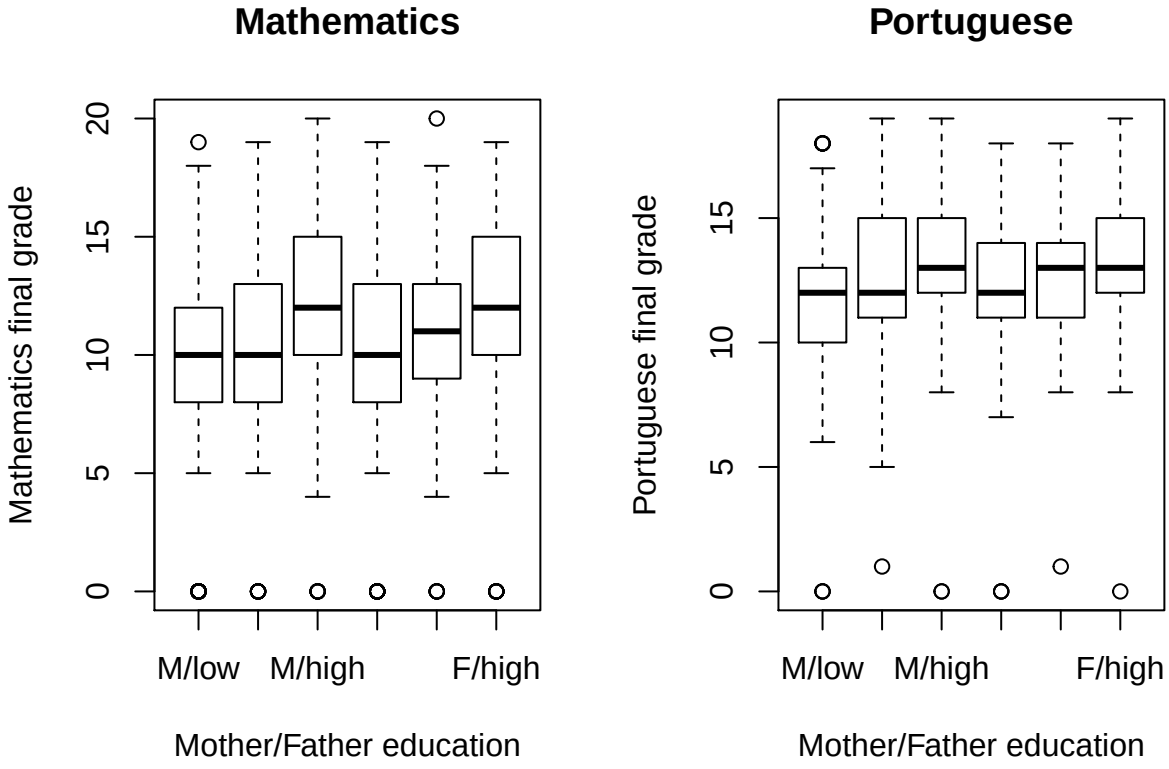
```
students <- students %>%
    mutate(Mat_grade = case_when(G3_mat < 10 ~ "F", G3_mat >= 10 & G3_mat < 14 ~ "C", G3_mat >= 14 &
        G3_mat < 16 ~ "B", G3_mat >= 16 ~ "A"))
students <- students %>%
    mutate(Por_grade = case_when(G3_por < 10 ~ "F", G3_por >= 10 & G3_mat < 14 ~ "C", G3_por >= 14 &
        G3_mat < 16 ~ "B", G3_por >= 16 ~ "A"))
```

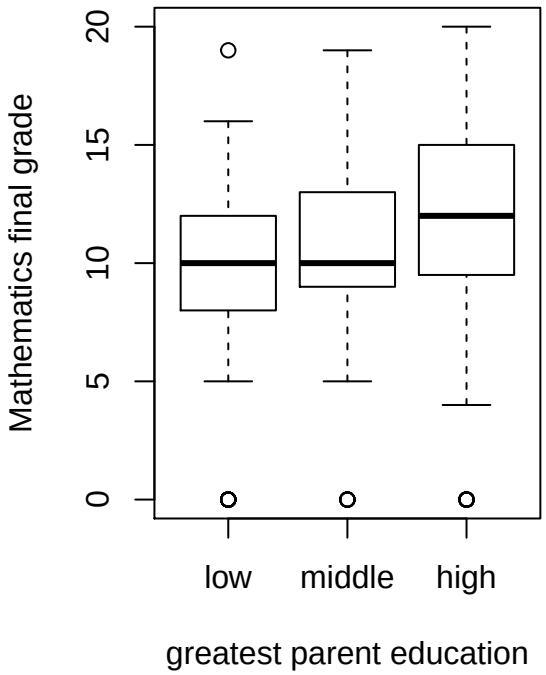
Then father's and mother's education are grouped into larger subgroups

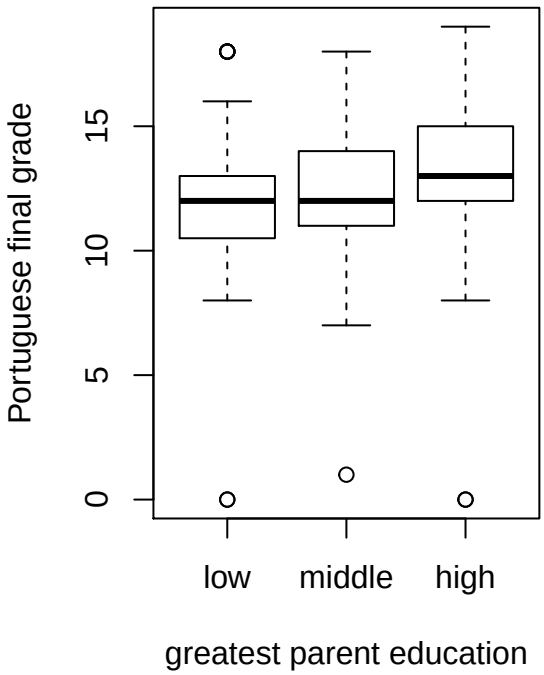
```
students <- students %>%
    mutate(MeduMod = case_when(Medu == "0" | Medu == "1" | Medu == "2" ~ "0", Medu == "3" ~ "1", Medu ==
        "4" ~ "2"))
students <- students %>%
    mutate(FeduMod = case_when(Fedu == "0" | Medu == "1" | Fedu == "2" ~ "0", Fedu == "3" ~ "1", Fedu ==
        "4" ~ "2"))
```

'greatest parent education' is defined as the maximum between father's and mother's education.

```
students$greatestparentedu <- pmax(students$MeduMod, students$FeduMod)
```







Hypothesis	Description
H0	Mathematics grade and greatest parent education are independent
H1	Mathematics grade and greatest parent education are not independent

```
tbl = table(students$greatestparentedu, students$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)$p.value
```

```
## [1] 0.0003224317
```

The p-value of the test is less than 0.05. We reject the H0 hypothesis in favor of the H1 hypothesis and we conclude that **education of the higher educated parent and mathematics grade are dependent attributes**.

Hypothesis	Description
H0	mathematics grade and mother's education are independent
H1	mathematics grade and mother's education are not independent

```
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)$p.value
```

```
## [1] 0.0009897562
```

The p-value of the independence test is less than 0.05. Therefore, we reject the H0 hypothesis in favor of the H1 hypothesis. Conclusion is drawn that **mother's education and mathematics grade are dependent.**

Hypothesis	Description
H0	mathematics grade and father's education are independent
H1	mathematics grade and father's education are not independent

```
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)$p.value
```

```
## [1] 0.1698656
```

The p-value of the independence test is higher than 0.05. Therefore, we do not reject the H0 hypothesis.

Hypothesis	Description
H0	Portuguese grade and greatest parent education are independent
H1	Portuguese grade and greatest parent education are not independent

```
tbl = table(students$greatestparentedu, students$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)$p.value
```

```
## [1] 0.003526118
```

```
for (col_names in colnames(added_margins_tbl)) {
    for (row_names in rownames(added_margins_tbl)) {
        if (!(row_names == "Sum" | col_names == "Sum")) {
            cat("Expected frequency for class", col_names, "-", row_names, ": ", (added_margins_tbl[row
                "Sum"] * added_margins_tbl["Sum", col_names])/added_margins_tbl["Sum", "Sum"], "\n")
        }
    }
}
```

```
## Expected frequency for class A - 0 :  5.752577
## Expected frequency for class A - 1 :  4.391753
## Expected frequency for class A - 2 :  7.85567
## Expected frequency for class B - 0 :  10.2268
## Expected frequency for class B - 1 :  7.80756
## Expected frequency for class B - 2 :  13.96564
## Expected frequency for class C - 0 :  69.98969
## Expected frequency for class C - 1 :  53.43299
## Expected frequency for class C - 2 :  95.57732
## Expected frequency for class F - 0 :  7.030928
## Expected frequency for class F - 1 :  5.367698
## Expected frequency for class F - 2 :  9.601375
```

The p-value of the test is less than 0.05. The H0 hypothesis is rejected in favor of the H1 hypothesis and it's concluded that education of higher educated parent and Portuguese grade are dependent attributes.

Expected frequency for class (grade=A, education=1) 4.391753 could be problematic for Chi-square test of independence. However, the the assumption of the test is that expected frequency should be 5 or more **in at least 80%** of the cells. In which case, Fisher's exact test should be used since it's used for smaller sample sizes.

Hypothesis	Description
H0	Portuguese grade and mother's education are independent
H1	Portuguese grade and mother's education are not independent

```
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)$p.value
```

```
## [1] 0.03370165
```

The p-value of the test is less than 0.05. The H0 hypothesis is rejected in favor of the H1 hypothesis and it's concluded that mother's education and Portuguese grade are dependent.

Hypothesis	Description
H0	Portuguese grade and father's education are independent
H1	Portuguese grade and father's education are not independent

```
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)$p.value
```

```
## [1] 0.04718521
```

```
for (col_names in colnames(added_margins_tbl2)) {
    for (row_names in rownames(added_margins_tbl2)) {
        if (!(row_names == "Sum" | col_names == "Sum")) {
            cat("Expected frequency for class ", col_names, "-", row_names, ": ", (added_margins_tbl2[r
                "Sum"] * added_margins_tbl2["Sum", col_names])/added_margins_tbl2["Sum", "Sum"], "\n")
        }
    }
}
```

```
## Expected frequency for class  A - 0 :  8.103093
## Expected frequency for class  A - 1 :  4.948454
## Expected frequency for class  A - 2 :  4.948454
## Expected frequency for class  B - 0 :  14.4055
## Expected frequency for class  B - 1 :  8.797251
## Expected frequency for class  B - 2 :  8.797251
## Expected frequency for class  C - 0 :  98.58763
## Expected frequency for class  C - 1 :  60.20619
## Expected frequency for class  C - 2 :  60.20619
## Expected frequency for class  F - 0 :  9.90378
## Expected frequency for class  F - 1 :  6.04811
## Expected frequency for class  F - 2 :  6.04811
```

There are two (2) expected frequencies whose value is less than 5. Since 2/12 is close to 20% (16%) Fisher's exact test will be used in this case.

```
fisher.test(tbl2)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tbl2
## p-value = 0.0602
```

```
## alternative hypothesis: two.sided
```

The p-value of the test is higher than 0.05. The H0 hypothesis isn't rejected.

Test case: Which school is better in mathematics and which in Portuguese?

author: Matej Ciglenečki - advised by the rest of the group

two (2) t-tests will be performed on four (4) different datasets. Dataset is split in four (4) different datasets (GP, MS) x (Mathematics, Portuguese): `gp_mat`, `gp_por`, `ms_mat`, `ms_por`

Mean grades for each subject will be used to decide a direction (left or right) of a one-sided t-test. The school's mean with a higher value will be the alternative H1 hypothesis.

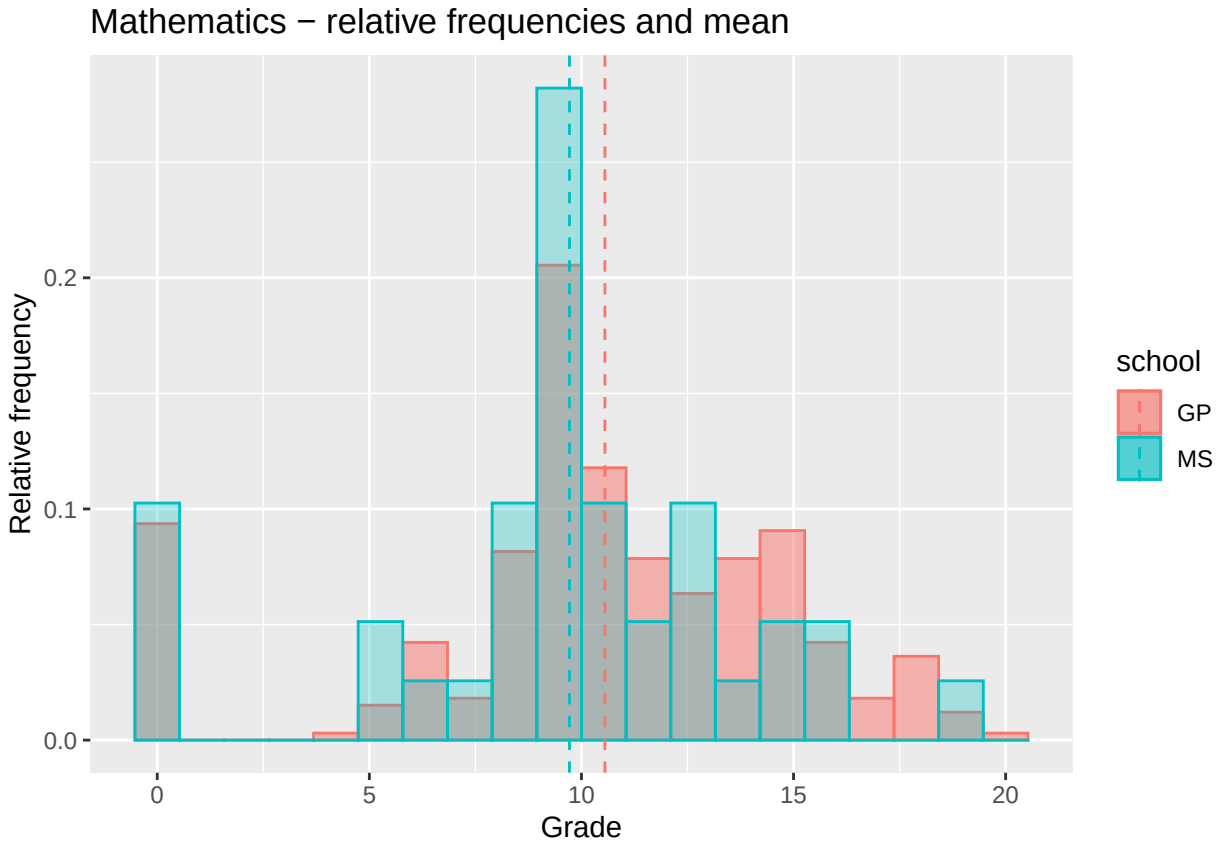
```
# Show average grade for all schools
schools <- students %>%
    select("school") %>%
    distinct(.)
schools  # [GP, MS]
subject_final_grade_names <- names(students)[grepl("G3", names(students))]

# 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)
```

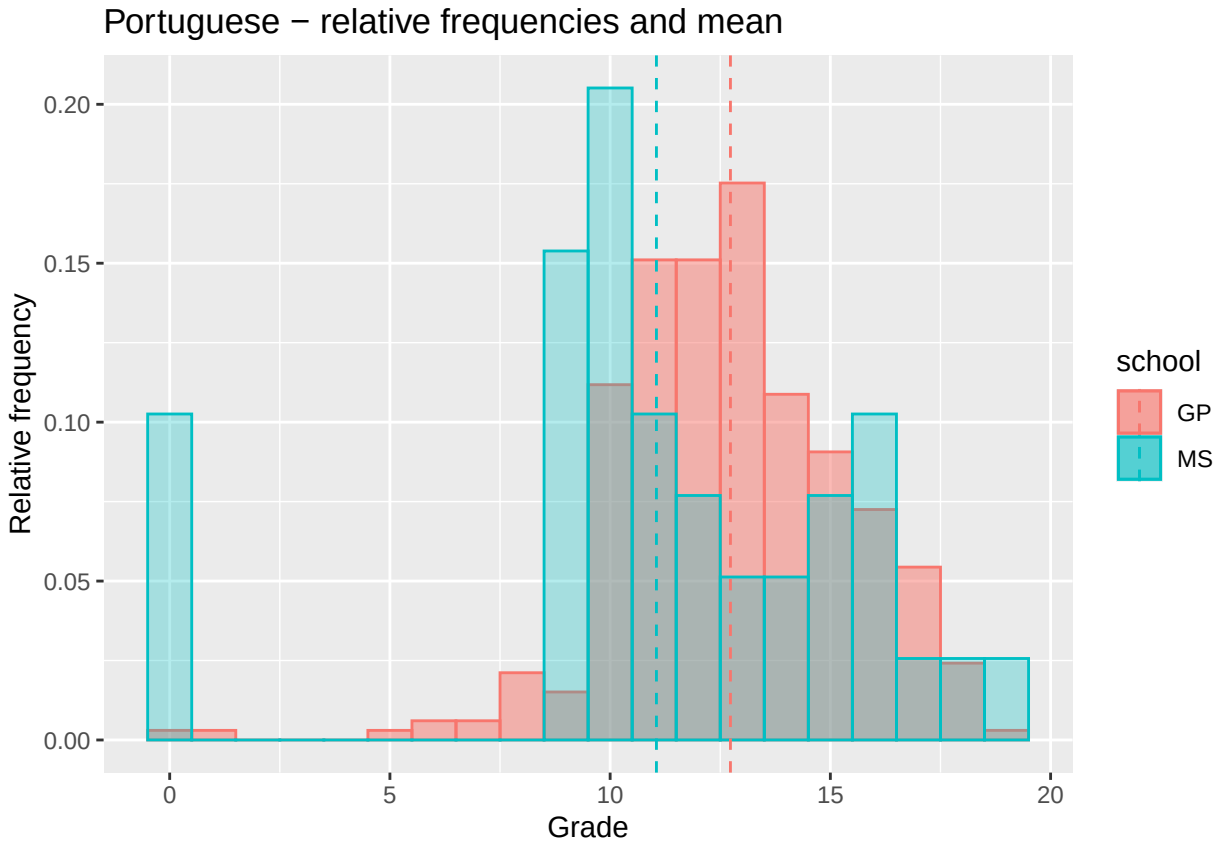
Columns are be renamed for easier usage.

```
# Rename all columns to 'grade'
gp_mat <- gp_mat %>%
    rename(grade = G3_mat)
gp_por <- gp_por %>%
    rename(grade = G3_por)
ms_mat <- ms_mat %>%
    rename(grade = G3_mat)
ms_por <- ms_por %>%
    rename(grade = G3_por)
```

Relative frequencies of subjects

Graphs show relative frequencies and means (vertical dashed lines) in mathematics grades for each school. Means are compared on both graphs. School's mean with a higher value (vertical line to the right) is taken as an alternative to the one-sided t-test. T-test will check a statistical significance between two means.





On both graphs, it's visible that `GP` school has higher a mean of grades in both subjects than `MS` school.

Normality test

Normality can be checked in multiple ways. In the following steps, two (2) methods are used:

- visual (`qqnorm`)
- quantitative decisions / tests (Lilliefors and Kolmogorov-Smirnov tests)

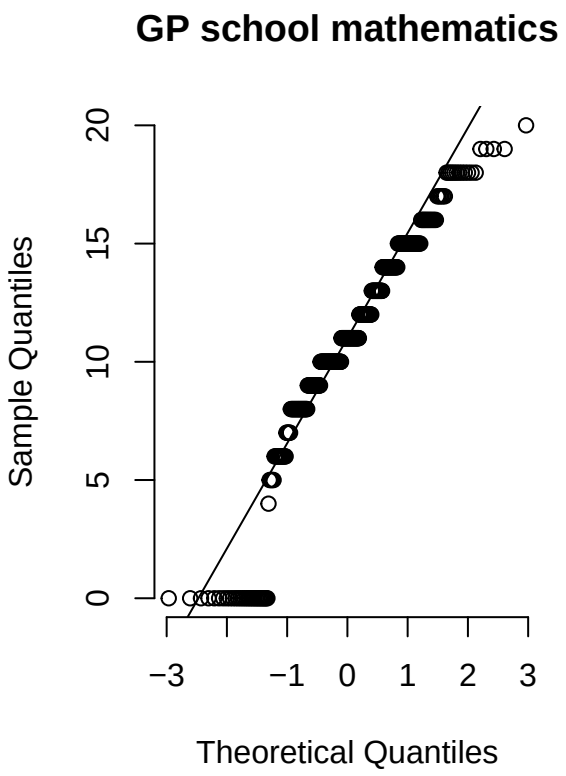
```
nrow(gp_mat)  # == nrow(gp_por)
```

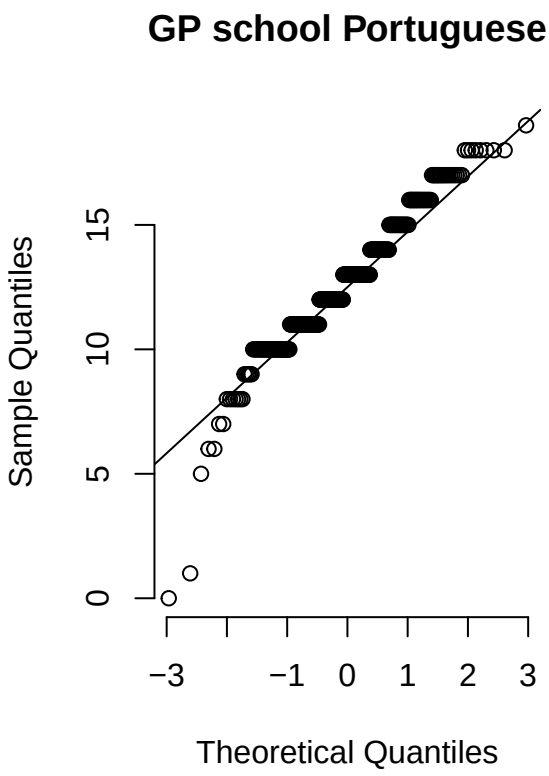
```
## [1] 331
```

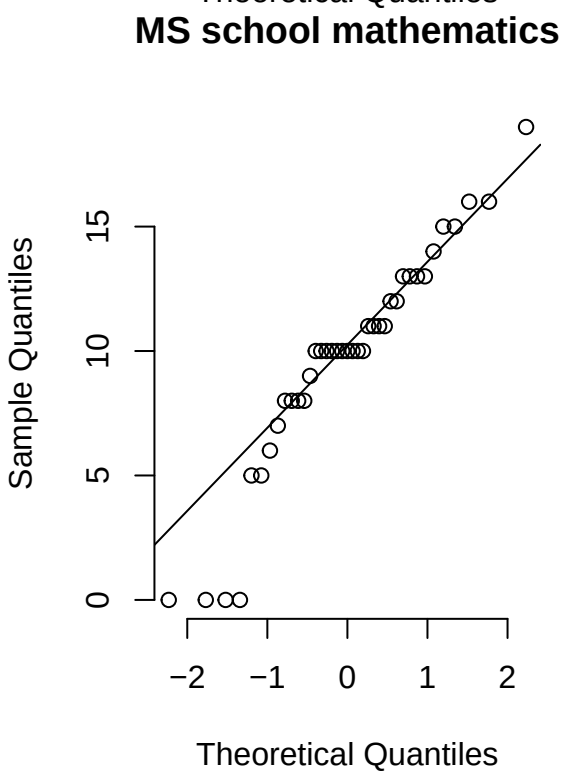
```
nrow(ms_mat)  # == nrow(ms_por)
```

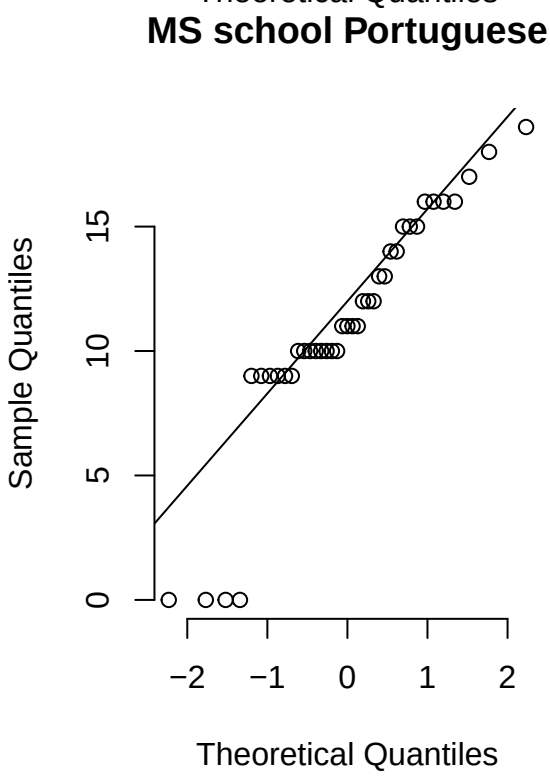
```
## [1] 39
```

`n` - size of the dataset for mathematics is `331` and `39` for Portuguese.










```
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"]
```

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

```
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"]
```

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

```
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"]
```

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

```
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"]
```

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

Tails are emphasized on the left side of the distribution, which is why the `p` value will almost always be less than 0.05 for the Kolmogovor-Smirnov and Lilliefors' test.

Visually we can see that data comes from the normal distribution but with a strong remark that the left tail is often present. Although normality is assumed, tests sensitive to normality won't be taken into account.

F-test of equality of variances

It's important to emphasize that the F-test of equality of variances is extremely sensitive to normality. The test will be conducted, but its results and conclusions will be discarded. Why? Because the distribution of datasets can't be considered normal in this case (because of left tails).

p – the probability that under the null hypothesis of obtaining the value (of the test statistic) that's as extreme (or more extreme) than the value we got computed from the sample we have

If $p < \alpha$ hypothesis $H0$ is rejected in favor of hypothesis $H1$ * falls under right tail => rejection

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \neg H_0$$

Order of arguments for the `var.test` function doesn't matter. However, in practice, the numerator has the higher value than the denominator:

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

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

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

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

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

Intuitively, it's assumed that the H0 hypothesis for Portuguese will be rejected because the variances are significantly different from each other. Of course, the F-test of equality of variances will be conducted to assure the statistical significance of the difference between two variances.

Construction of the test:

```r
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")
mat_f_test["p.value"]

# 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")
por_f_test["p.value"]
```

```r
var_equal_mat <- if (mat_f_test$p.value < alpha) FALSE else TRUE
cat_reject_h0("Mathematics - F-test of equality of variances:", !var_equal_mat)
```

```
## Mathematics - F-test of equality of variances:
##  We do not reject the H0 hypothesis
```

```r
var_equal_por <- if (por_f_test$p.value < alpha) FALSE else TRUE
cat_reject_h0("Portuguese - F-test of equality of variances:", !var_equal_por)
```

```
## Portuguese - F-test of equality of variances:
##  We reject the H0 hypothesis in favor of the H1 hypothesis
```

Unpaired two sample T-test test of equality of means

Because the `n` is bigger than 30 for both datasets and it is true that the t-test is robust to (non)normality, an unpaired two-sample test of equal means is conducted for both subjects.

With previously calculated means, the one-sided alternative hypothesis is chosen (an alternative that school `GP` has a higher mean)

Again, because of the F-test's sensitivity to normality, it's assumed that variances are unequal for the t-test.

```r
# 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("Mathematics - t-test:", is_gp_mat_better)
## Mathematics - t-test:
##  We do not reject the H0 hypothesis

# 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("Portugueuse - t-test:", is_gp_por_better)
```

```
## Portugueuse - t-test:
##  We reject the H0 hypothesis in favor of the H1 hypothesis
```

Mathematics - The H0 hypothesis is not rejected. It can't be stated that school GP has better math grades than school MS

Portuguese - The H0 hypothesis in favor of the H1 hypothesis from which it's concluded that school `GP` has better grades in Portuguese than school `MS`.

Test case: Are students more successful in mathematics or Portuguese?

author: Tomislav Prhat - advised by the rest of the group

```
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
```

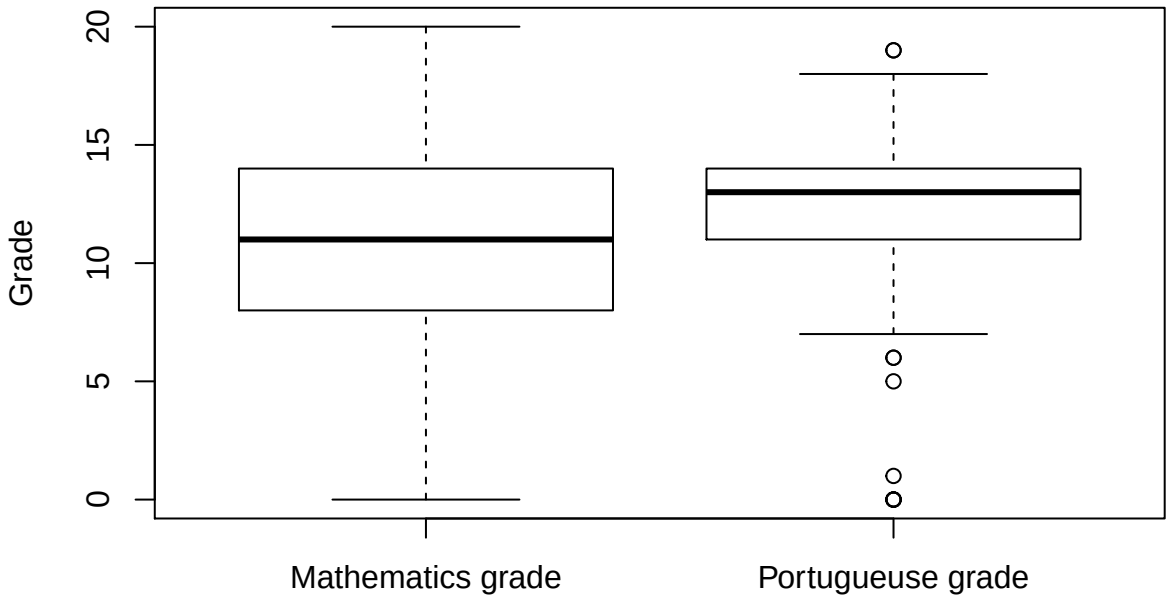
```
students_org %>%
    summarise(Mean.G3_mat = mean(G3_mat, trim = 0.1), Mean.G3_por = mean(G3_por, trim = 0.1), ) ->
    summary.result3
summary.result3
```

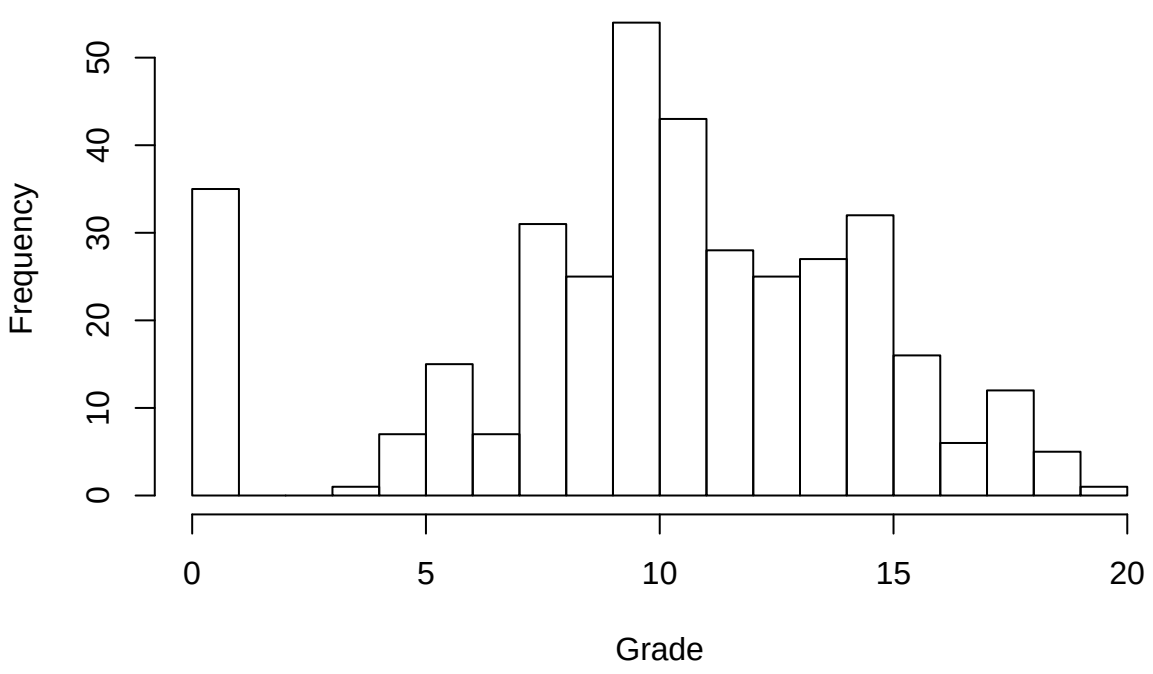
```
## # A tibble: 1 x 2
##   Mean.G3_mat Mean.G3_por
##         <dbl>       <dbl>
## 1        10.9        12.6
```

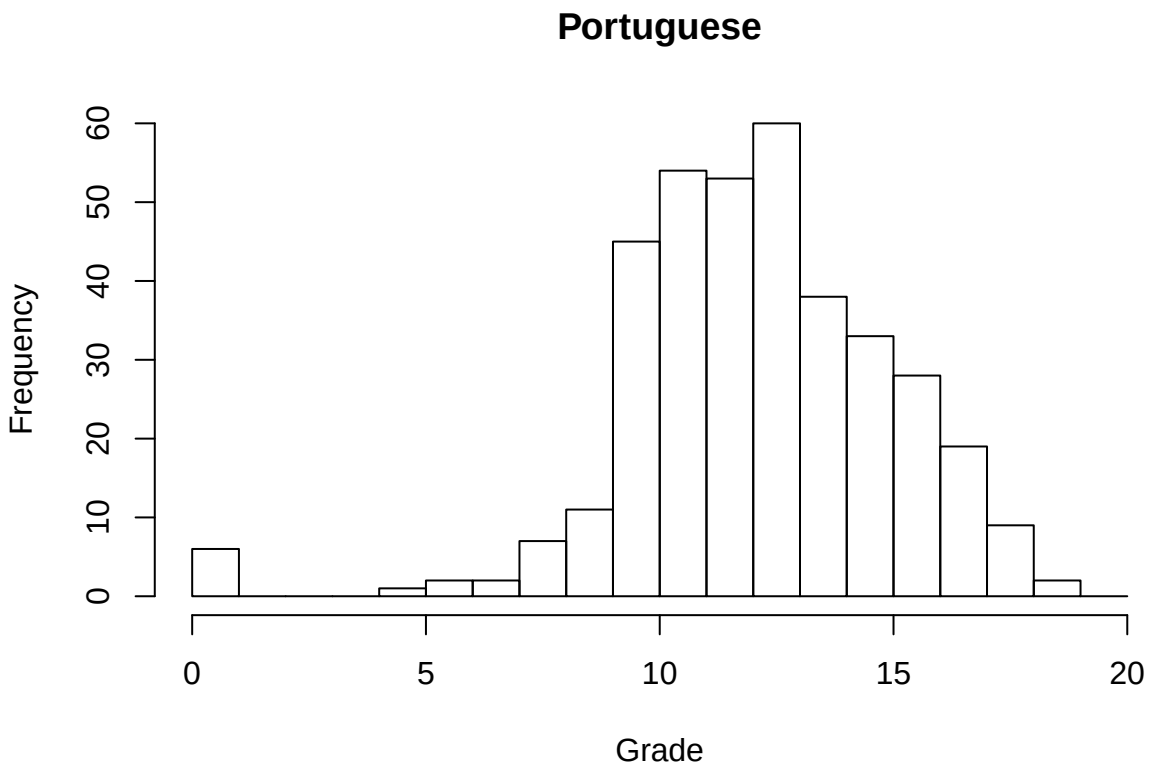
```
(1 - summary.result3/summary.result1) * 100
```

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

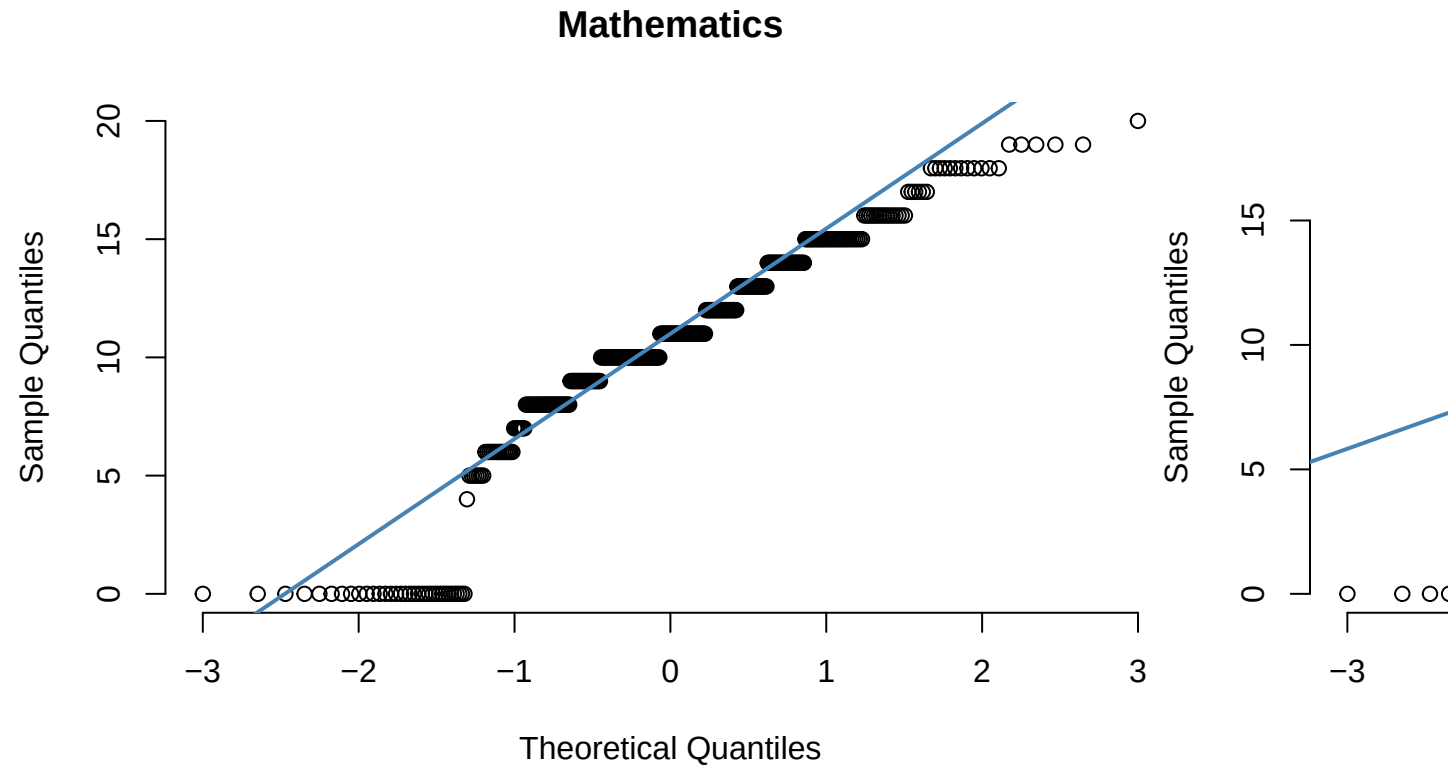
Portuguese grade's mean, median and trimmed mean (10%) is higher than mathemathic's


Grade
0
5
10
15
20
Mathematics grade
Portugueuse grade


Mathematics
Frequency
0
10
20
30
40
50
0
5
10
15
20
Grade



Before the t-test, normality is checked visually and with Kolmogorov-Smirnov's and Lilliefors' tests:




```
cat("Mathematics p-value (Lillie):", unlist(lillie.test(students_org$G3_mat)["p.value"]))
```

```
## Mathematics p-value (Lillie): 2.653956e-17
```

```
cat("Mathematics p-value (KS):", unlist(ks.test(students_org$G3_mat, "pnorm", mean(students_org$G3_mat)
    sd(students_org$G3_mat))["p.value"]))
```

```
## Mathematics p-value (KS): 4.296089e-06
```

```
cat("Portuguese p-value (Lillie):", unlist(lillie.test(students_org$G3_por)["p.value"]))
```

```
## Portuguese p-value (Lillie): 1.236014e-12
```

```
cat("Portuguese p-value (KS):", unlist(ks.test(students_org$G3_por, "pnorm", mean(students_org$G3_por),
    sd(students_org$G3_por))["p.value"]))
```

```
## Portuguese p-value (KS): 0.0001241436
```

Small p-values are the result of left tails. Visually we can see that data comes from the normal distribution but with a strong remark that the left tail is often present. Although normality is assumed, tests sensitive to normality won't be taken into account.

F-test of equality of variances

Because of the already mentioned extreme sensitivity to normality, the F-test of equality of variances will be conducted, but its results and conclusions won't be taken into account.

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \neg H_0$$

```
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
```

Because of the small p-value, the H0 hypothesis is rejected in favor of the H1 hypothesis. Variances are different for each subject.

T-test for equality of grade means

For both tests, the alternative case is the higher mean grade is in Portuguese compared to mathematics.

```
# H0 - Mean grades are the same (Mat=Por) H1 - Por > Mat
por_mat_t_test <- t.test(students_org$G3_por, students_org$G3_mat, alternative = "greater", var.equal =

is_por_higher <- if (por_mat_t_test$p.value < alpha) TRUE else FALSE
cat_reject_h0("T-test for equality of grade means:", is_por_higher)
```

```
## T-test for equality of grade means:
##  We reject the H0 hypothesis in favor of the H1 hypothesis
```

Test case: How does travel time affect students' success?

ANOVA will be performed to answer this question.

ANOVA's assumptions are: * independence of sample cases * the population from which samples are drawn should be normally distributed * homogeneity of variance (variance among the groups should be approximately equal)

H0 hypothesis - mean value of groups are equal

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

$$H_1 : \neg H_0$$

We can assume independence because the schools are different.

If the H0 hypothesis we conclude that mean values are unequal. In other words, we conclude that travel time affects the mean of students' grade (success).

Handling categorical values

Groups are defined by the attribute `traveltime`. It's necessary to transform the values from attribute `traveltime` to categorical continuous data (factors with an order). `traveltime` attribute has 4 possible values which define the travel time from school to students' home: * < 15min * 15 - 30 min * 30 - 60 min * > 60 min

The last category (60min+) will be merged with the second to last category (30-60min) because only 8 data points are contained within the last group (60min+), which is significantly smaller compared to the size of other groups.

```
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
```

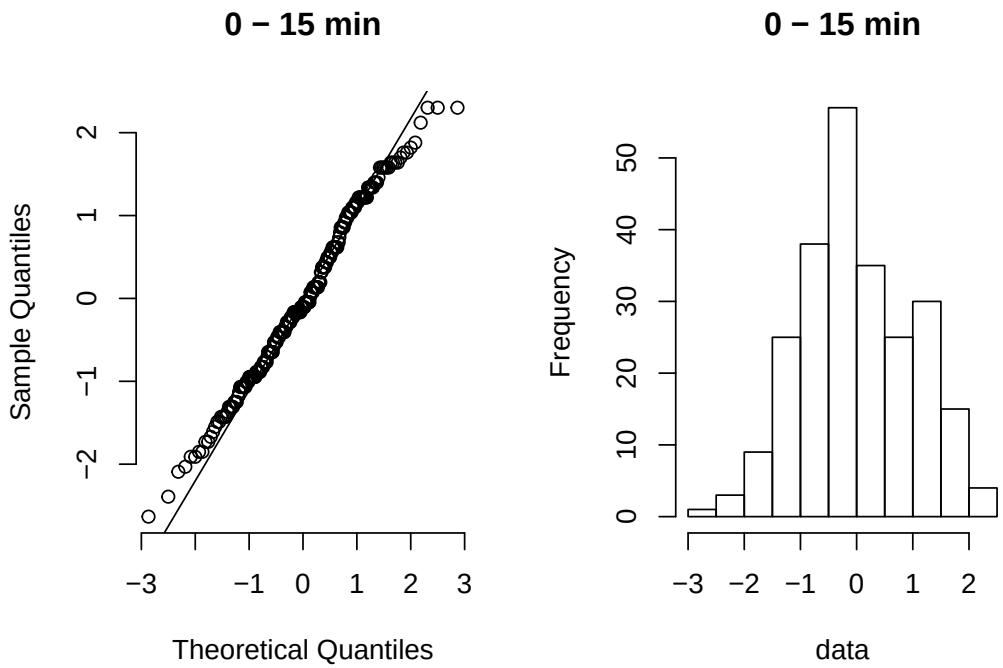
```
students <- students_clean
students$traveltime <- factor(students$traveltime, ordered = TRUE, labels = c("0 - 15 min", "15 - 30 mi
    "> 30 min", "> 30 min"))
```

Term 'success' (G_total) is defined as sum of `G[1,2,3]_mat` i `G[1,2,3]_por`

```
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$G_por_total <- students$G1_por + students$G2_por + students$G3_por
students$G_mat_total <- students$G1_mat + students$G2_mat + students$G3_mat

students$G_total <- students$G1_total + students$G2_total + students$G3_total
```

ANOVA is robust to slight irregularities in normality. Nonetheless, normality for `G_total` will be tested for the whole dataset and then for each group independently.

```
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)
```



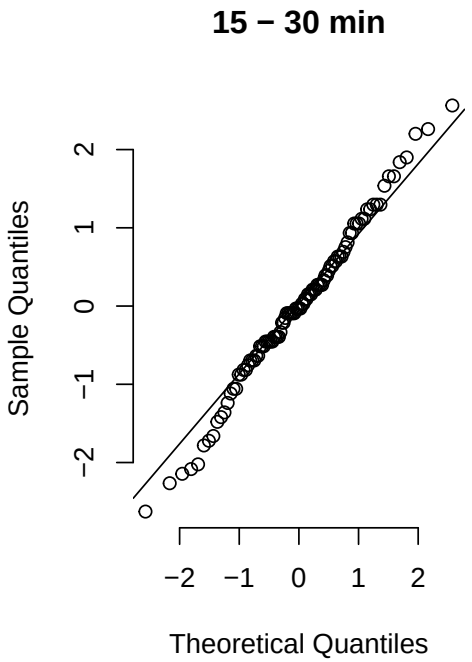

```
lillie.test(data)["p.value"]
```

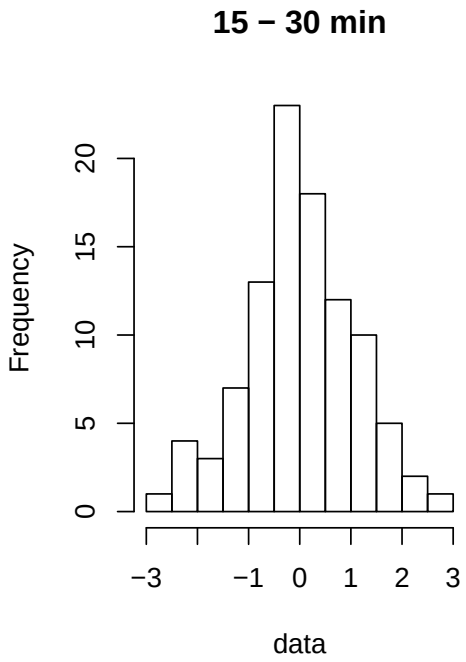
```
## $p.value
## [1] 0.008983716
```

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

```
## $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)
```





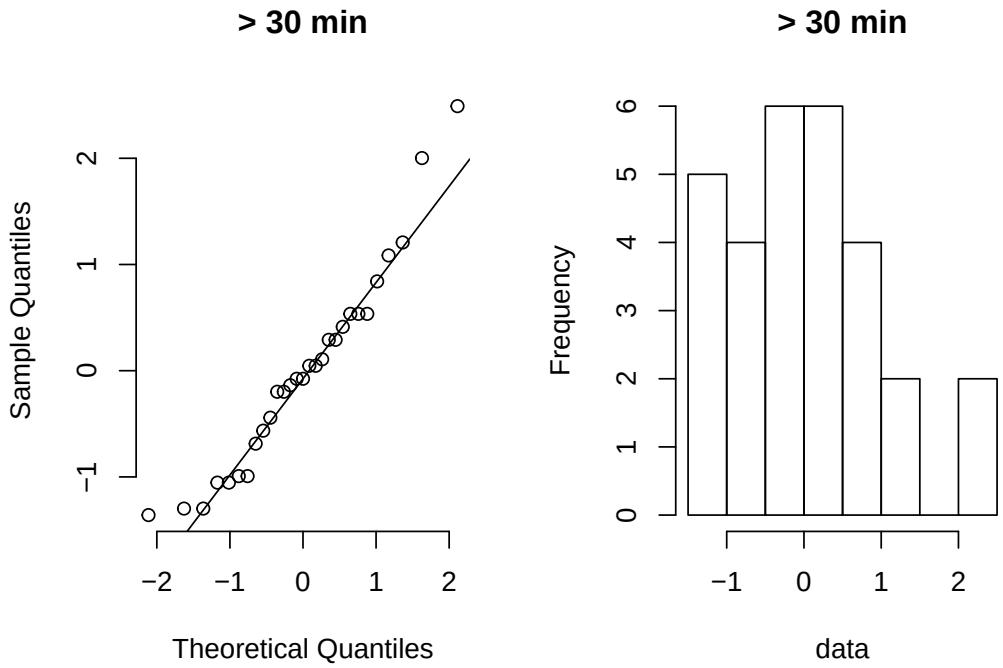

```
lillie.test(data)["p.value"]
```

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

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

```
## $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"]
```

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

On the graph, it's visible that data is normally distributed with a few outliers (left tail). `p` value of the Lilliefors' test sometimes goes below 0.05, however, it's always above 0.05 for the Kolmogorov-Smirnov test.

Lilliefors' test is used if the variance and mean of the population are unknown, which is true for this dataset. It's known that Lilliefors is more conservative compared to the Kolmogorov-Smirnov test, meaning that it's more likely to reject the H0 hypothesis.

Taking everything into account, normality is assumed. Deviations from normality are small and `p` values that are below 0.05 are relatively close to 0.05.

Bartlett's test of homogeneity of variancies

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

$$H_1 : \neg H_0$$

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

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

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

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

```
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
```

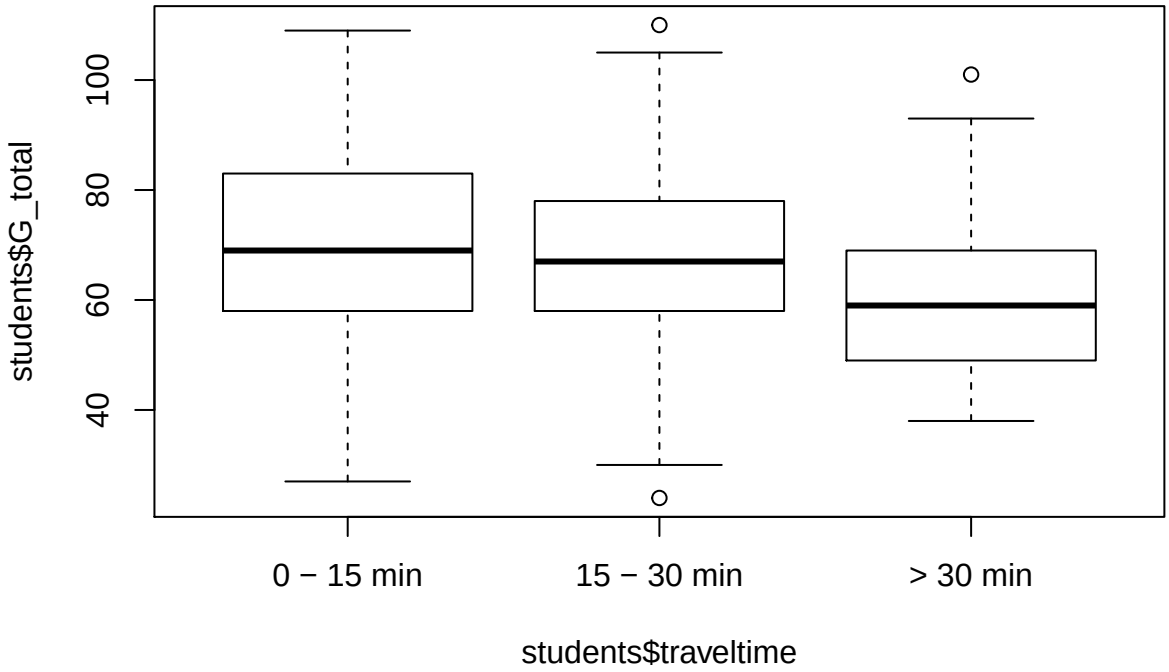
Values of variances are similar. p value of the test is above 0.05 because of which the H0 hypothesis is not rejected. With this, it's confirmed the dataset does not violate ANOVA's assumption for homogeneity of variances.

Analysis of variance (ANOVA) test of equality of means

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

$$H_1 : \neg H_0$$

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



Visually, we can assume that travel time does affect students' success. However, it's necessary to perform the ANOVA test to confirm if the difference is statistically significant.

```
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 suggests that there is a difference between groups `traveltime`. Although the difference isn't enormous, the `p` (between `0.001` and `0.01`) value still suggests statistical significance. The conclusion follows: different `traveltime` groups influence students' success.

Test case: Which variables best predict students' success?

author: Magda Radić - advised by the rest of the group

First, categorical data is one-hot-encoded.

```
require(fastDummies)
students_org
students_dummies = dummy_cols(students_org, remove_first_dummy = TRUE, remove_selected_columns = TRUE)
students_dummies

students_dummies$G_por_total <- students_dummies$G1_por + students_dummies$G2_por + students_dummies$G3

students_dummies$G_mat_total <- students_dummies$G1_mat + students_dummies$G2_mat + students_dummies$G3

students_dummies$G_total <- students_dummies$G_por_total + students_dummies$G_mat_total
```

Coefficient of determination

Coefficient of determination $R^2 \in [0, 1]$, pronounced "R squared", is a statistical measure that represents the proportion of the variance for a dependent variable (`G_total`) that's explained by an independent variable or variables in a regression model.

Individual linear regressions are performed where `G3_mat` and `G3_por` are dependent variables and other variables are regressors. R^2 and p-values of the F-tests are saved to an array and will be used later to check which regressors give the minimum R^2 value.

```
filtered_col_names = c()
r_squares = c()
ps = c()

for (i in 1:ncol(students_dummies)) {

    col_names = colnames(students_dummies)
    col_name = col_names[i]

    if (!startsWith(col_name, "G")) {
        # skip grades

        model = lm(students_dummies$G_total ~ students_dummies[[col_name]])

        summary_model = summary(model)

        # appending values
        filtered_col_names <- append(filtered_col_names, col_name)

        r_squares <- append(r_squares, summary_model$r.squared)
        ps <- append(ps, pf(summary_model$fstatistic[1], summary_model$fstatistic[2], summary_model$fst
            lower.tail = FALSE))
    }
}
```

```
df_g_squares = data.frame(filtered_col_names, r_squares, ps)

head(df_g_squares, n = 3)
##   filtered_col_names  r_squares           ps
## 1                age 0.01943650 7.236672e-03
## 2               Medu 0.06082286 1.569225e-06
## 3               Fedu 0.03959779 1.165177e-04
```

Top predictors for dependent variable `G_total`

A predictor is also referred to as:

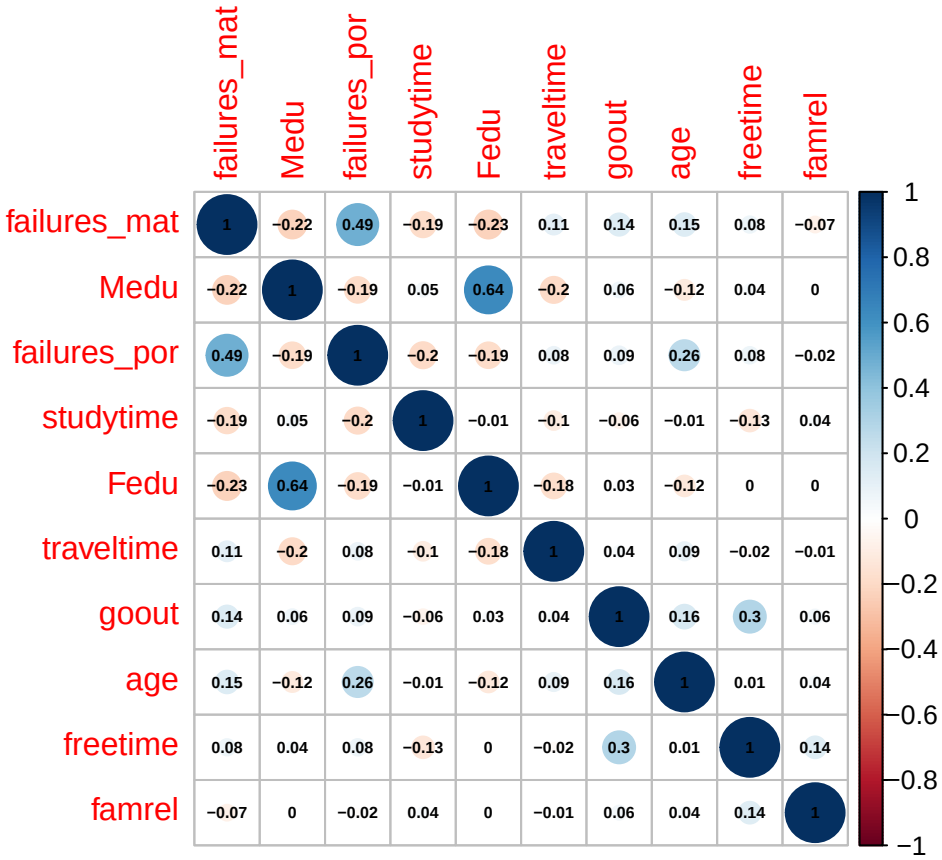
- A **regressor**
- An explanatory variable
- An independent variable
- A manipulated variable
- A feature

Which variables are the best predictors (regressors) for `G_total`? Predictors are sorted by the coefficient of determination R^2. Predictor with greatest R^2 value are the best predictors. In this case, 10 best predictors are taken into a consideration.

```
df_top_predictors = df_g_squares[order(-df_g_squares$r_squares[1:10]), ]
top_10_predictors = as.vector(df_top_predictors$filtered_col_names)
df_top_predictors
##    filtered_col_names    r_squares           ps
## 6        failures_mat 0.1776510687 2.285554e-17
## 2                Medu 0.0608228554 1.569225e-06
## 7        failures_por 0.0501116774 1.382143e-05
## 5           studytime 0.0400388296 1.065453e-04
## 3                Fedu 0.0395977880 1.165177e-04
## 4          traveltime 0.0301058215 8.031659e-04
## 10              goout 0.0247917030 2.386251e-03
## 1                 age 0.0194364986 7.236672e-03
## 9            freetime 0.0023387926 3.535949e-01
## 8              famrel 0.0002020802 7.852153e-01
```

From the top 10 predictors, it might be desirable to ditch predictors which highly correlate with another predictor, as both of them describe similar variability. The decision is performed with a visual and quantitative review of the correlation matrix. If there is any pair of predictors whose absolute correlation value is higher than 0.7, one of the predictors from the pair is ditched. Preferably, it would be a predictor whose sum of absolute correlations coefficients with other predictors is greater.

```
df_student_success <- students_dummies[, top_10_predictors]
corrplot(cor(df_student_success), addCoef.col = "black", number.cex = 0.5)
```



All predictors are taken into account since there isn't a pair of predictors whose absolute correlation value exceeds 0.7.

```
model_top_pred <- lm(students_dummies$G_total ~ ., df_student_success)
summary_top_pred <- summary(model_top_pred)
summary_top_pred
##
## Call:
## lm(formula = students_dummies$G_total ~ ., data = df_student_success)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.958  -9.738  -0.315   9.818  37.775
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   78.8829    12.4851   6.318 7.84e-10 ***
## failures_mat  -8.2133     1.2912  -6.361 6.11e-10 ***
## Medu           2.1615     0.9449   2.287   0.0228 *
## failures_por   0.8885     1.8836   0.472   0.6374
## studytime      2.4008     0.9545   2.515   0.0123 *
## Fedu           0.3121     0.9392   0.332   0.7399
## traveltime    -2.0171     1.1424  -1.766   0.0783 .
## goout         -1.5871     0.7389  -2.148   0.0324 *
## age           -0.7149     0.6950  -1.029   0.3043
## freetime       0.3195     0.8416   0.380   0.7045
## famrel        -0.1828     0.8651  -0.211   0.8328
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.91 on 359 degrees of freedom
## Multiple R-squared:  0.2397, Adjusted R-squared:  0.2185
## F-statistic: 11.32 on 10 and 359 DF,  p-value: < 2.2e-16

ps_top_pred <- summary_top_pred$coefficients[, 4]
ps_top_pred[order(ps_top_pred)]  # $coefficients[,4] -> p-values
## failures_mat  (Intercept)    studytime         Medu        goout   traveltime
## 6.109956e-10 7.839594e-10 1.233313e-02 2.275028e-02 3.238780e-02 7.830779e-02
##          age failures_por     freetime         Fedu       famrel
## 3.043017e-01 6.374389e-01 7.044750e-01 7.398662e-01 8.327764e-01
```

The model is simplified so that it doesn't have two (2) regressors whose p-value is the greatest. Regressors with the greatest p-values. A higher p-value indicates a weaker explanation of variance.

```
top_pred_trim <- top_10_predictors[1:(length(top_10_predictors) - 2)]
df_student_success_trim <- df_student_success[, top_pred_trim]

model_top_pred_trim <- lm(students_dummies$G_total ~ ., df_student_success_trim)
summary_top_pred_trim <- summary(model_top_pred_trim)
summary_top_pred_trim
##
## Call:
## lm(formula = students_dummies$G_total ~ ., data = df_student_success_trim)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.839  -9.548  -0.452   9.747  37.391
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   79.3586    11.8657   6.688 8.63e-11 ***
## failures_mat  -8.1889     1.2837  -6.379 5.46e-10 ***
## Medu           2.1787     0.9416   2.314   0.0212 *
## failures_por   0.9115     1.8774   0.486   0.6276
## studytime      2.3530     0.9448   2.491   0.0132 *
## Fedu           0.2992     0.9362   0.320   0.7495
## traveltime    -2.0362     1.1383  -1.789   0.0745 .
## goout         -1.5157     0.7058  -2.148   0.0324 *
## age           -0.7323     0.6919  -1.058   0.2906
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.87 on 361 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2225
## F-statistic:  14.2 on 8 and 361 DF,  p-value: < 2.2e-16

ps_top_pred_trim <- summary_top_pred_trim$coefficients[, 4]  # $coefficients[,4] -> p-values
ps_top_pred_trim[order(ps_top_pred_trim)]
##  (Intercept) failures_mat    studytime         Medu        goout   traveltime
## 8.625439e-11 5.458482e-10 1.320284e-02 2.124413e-02 3.241815e-02 7.450261e-02
##          age failures_por         Fedu
```

```
## 2.906085e-01 6.275990e-01 7.494700e-01
```

The R^2 is smaller however, the adjusted R^2 is larger than the previous model, indicating that unnecessary regressors were discarded. This linear model represents the proportion of the variance (22.25%) for a dependent variable (`G_total`) that's explained by the top 8 variables.

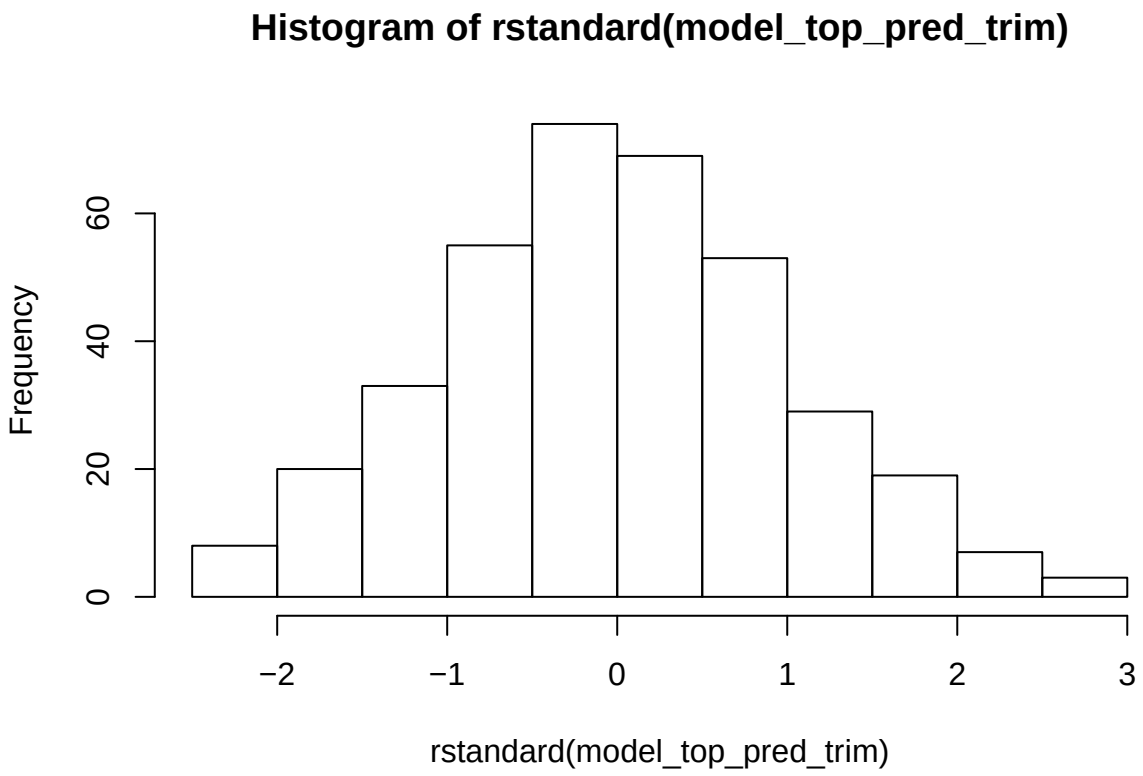
Normality of residuals

https://analyse-it.com/docs/user-guide/fit-model/linear/residual-normality#:~:text=Normality%20is%20the%20assumption%

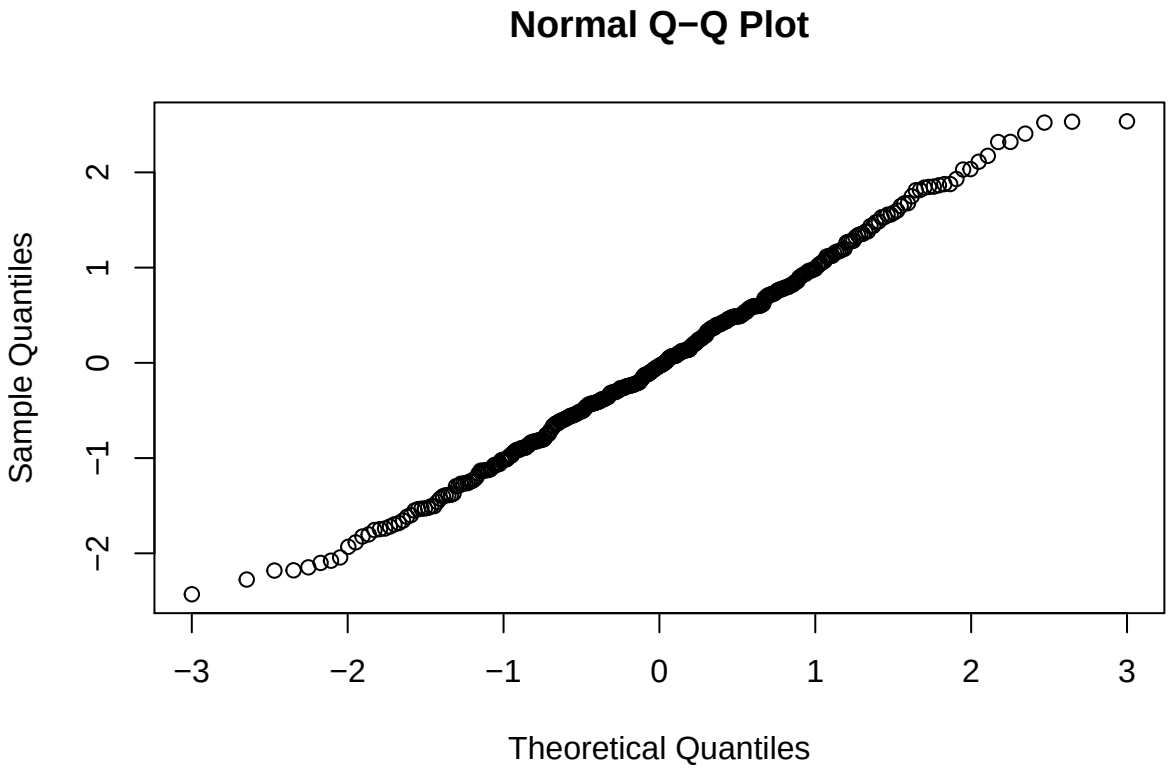
> Violation of the normality of residuals assumption only becomes an issue with small sample sizes. For large sample sizes, the assumption is less important due to the central limit theorem, and the fact that the F and t-tests used for hypothesis tests and forming confidence intervals are robust to modest departures from normality.

On graphs, it's visible that residuals are normally distributed.

```
hist(rstandard(model_top_pred_trim))
```




```
qqnorm(rstandard(model_top_pred_trim))
```




```
ks.test(rstandard(model_top_pred_trim), "pnorm")
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(model_top_pred_trim)
## D = 0.031334, p-value = 0.8607
## alternative hypothesis: two-sided
```