Mini-research project

Code

Yarka Lyba (https://www.facebook.com/yarkarybka), Sofiya Hevorhyan (https://www.facebook.com/sofiya.hevorhyan), Danil Shankowskyy (https://www.facebook.com/beeasyit)

December 2018

For our research project we used data from https://www.kaggle.com/ (https://www.kaggle.com/). It contains a lot of interesting social, gender and study information about students of the secondary school.

We decided to analize which factors have the biggest influence on grades of the students. The factors that we will check are: sex, education of parents, romantic relashion, time that students spend on studying and alcohol consumption.

The first step was to read data and look what actually we can analyze.

Hide

```
df <- read.csv("./student-mat.csv",
               header = TRUE)
df
```

school <fctr>	sex <fctr>	age <int>	address <fctr>	famsize <fctr>	Pstatus <fctr>	Medu <int>	Fedu <int>	Mjob <fctr>	Fjob <fctr>
GP	F	18	U	GT3	A	4	4	at_home	teacher
GP	F	17	U	GT3	T	1	1	at_home	other
GP	F	15	U	LE3	T	1	1	at_home	other
GP	F	15	U	GT3	T	4	2	health	services
GP	F	16	U	GT3	T	3	3	other	other
GP	M	16	U	LE3	T	4	3	services	other
GP	M	16	U	LE3	T	2	2	other	other
GP	F	17	U	GT3	A	4	4	other	teacher
GP	M	15	U	LE3	A	3	2	services	other
GP	M	15	U	GT3	T	3	4	other	other

1-10 of 395 rows | 1-10 of 33 columns Previous **1** 2 3 4 5 6 ... 40 Next

Hide

```
final.grade <- df$G3
```

Hide

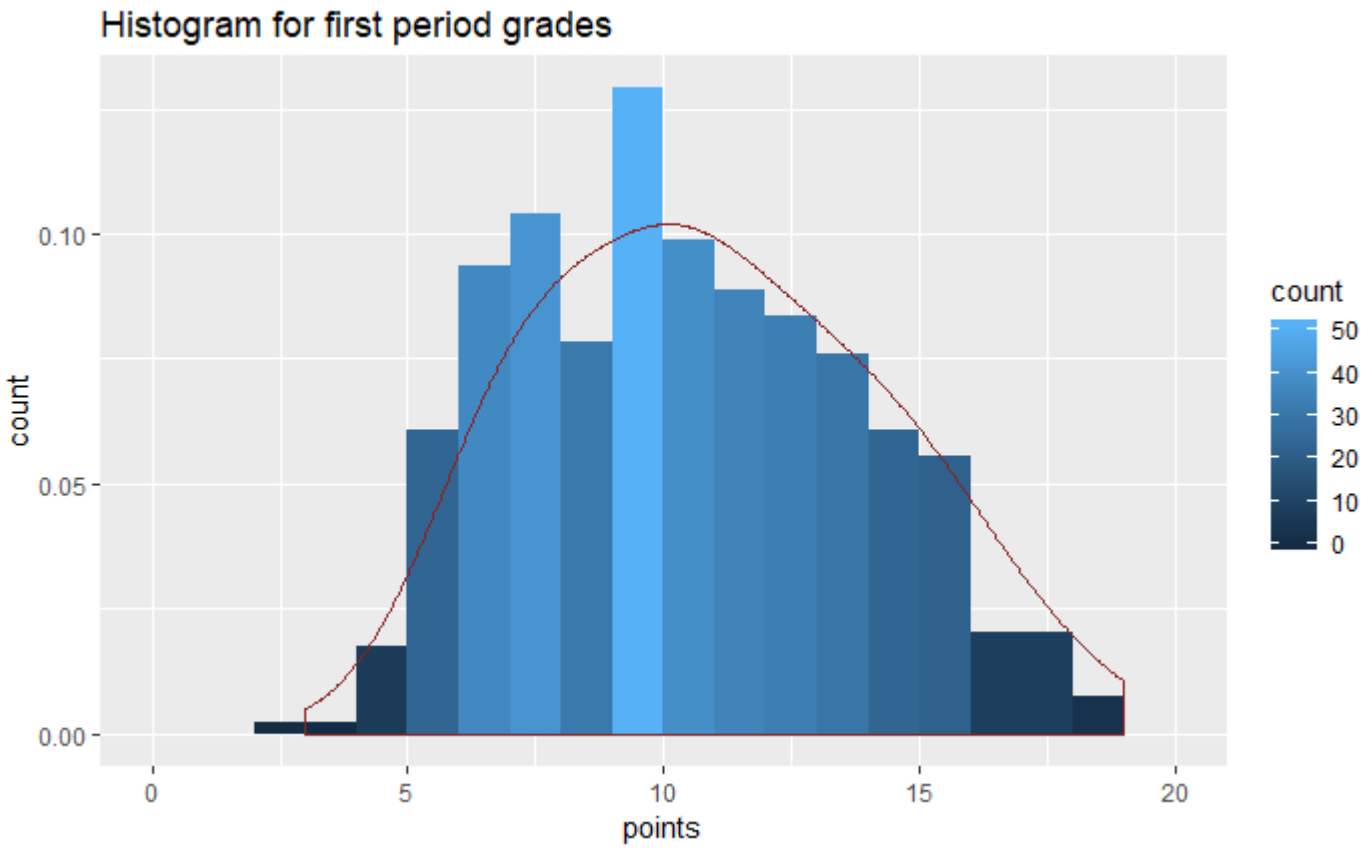
```
library("ggplot2")
library("plyr")
library("dplyr")
library("fitdistrplus")
library("logspline")
```

The main idea is to analyze grades of the students, so as a next step we decided to work with these columns.

We have visualized first period, second period and final grades. Grades are numerical values from 0 to 20.

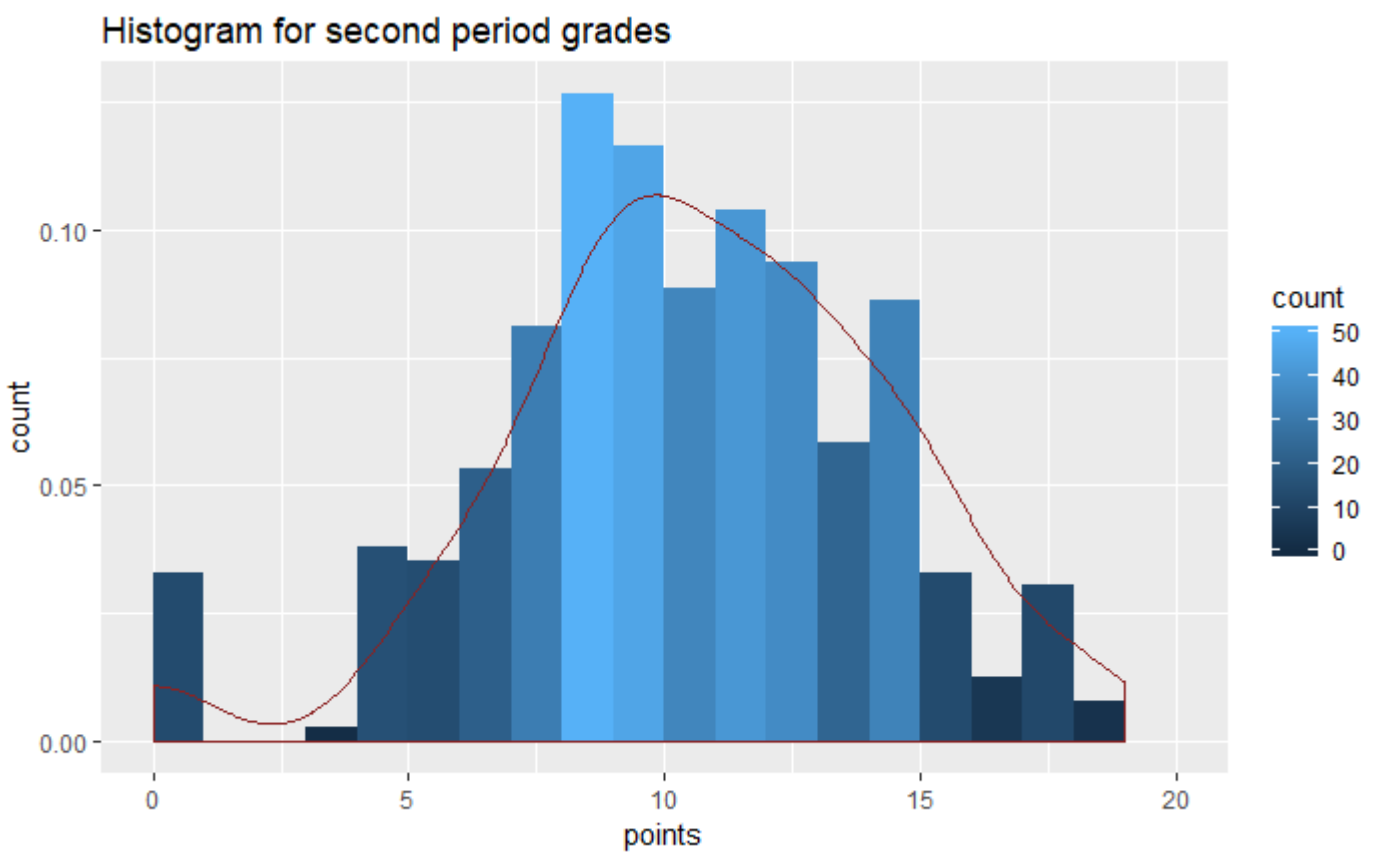
Hide

```
ggplot(data=df, aes(x=df$G1)) +
  geom_histogram(aes(y =..density.., fill=..count..),
                 breaks=seq(0, 20, by =1))+
                 labs(title="Histogram for first period grades") +
                 geom_density(col="brown4", adjust=1.5) +
                 labs(x="points", y="count")
```



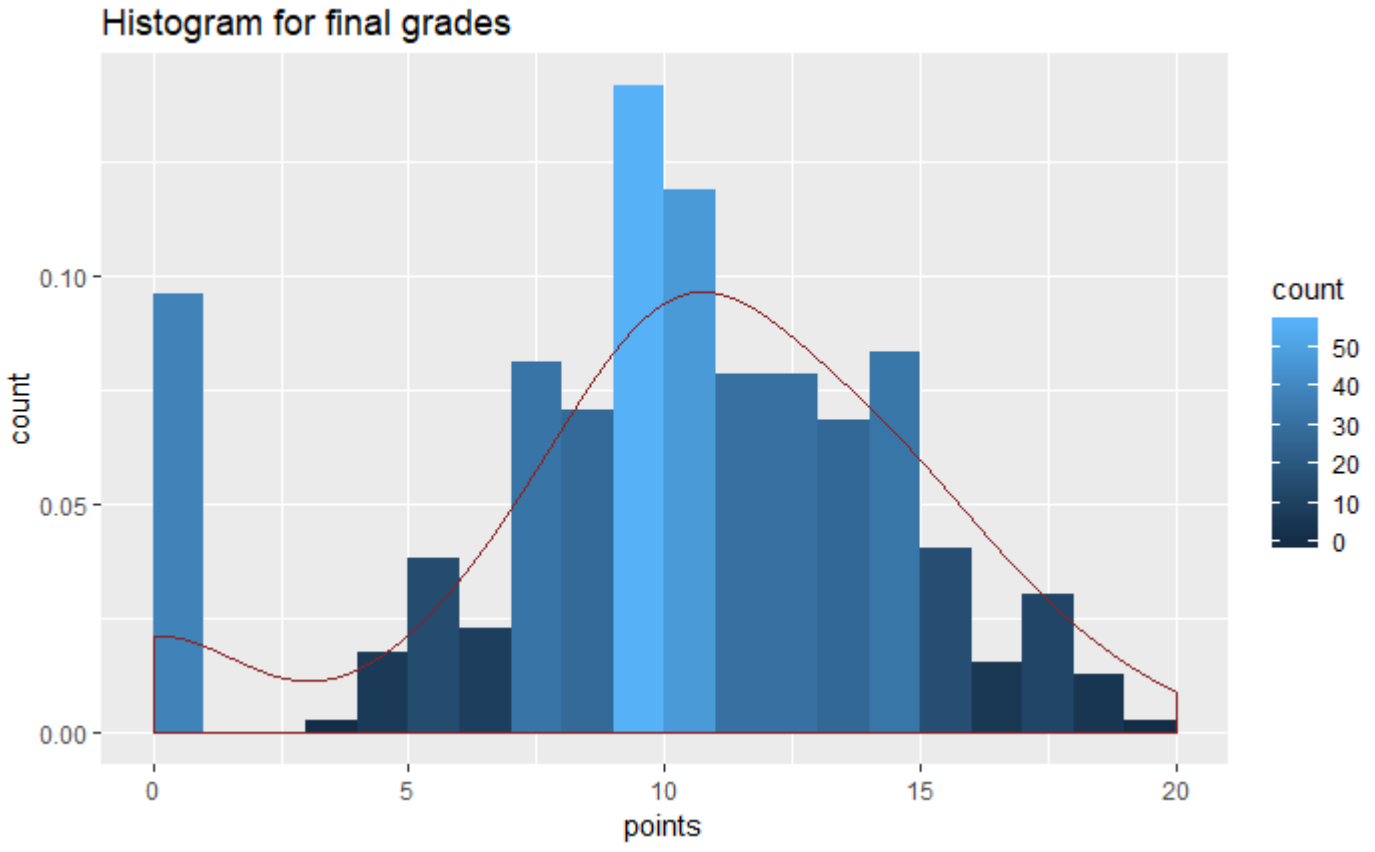
Hide

```
ggplot(data=df, aes(x=df$G2)) +
  geom_histogram(aes(y =..density.., fill=..count..),
                 breaks=seq(0, 20, by =1))+
                 labs(title="Histogram for second period grades") +
                 geom_density(col="brown4", adjust=1.5) +
                 labs(x="points", y="count")
```



Hide

```
ggplot(data=df, aes(x=df$G3)) +
  geom_histogram(aes(y =..density.., fill=..count..),
                 breaks=seq(0, 20, by =1))+
                 labs(title="Histogram for final grades") +
                 geom_density(col="brown4", adjust=1.5) +
                 labs(x="points", y="count")
```

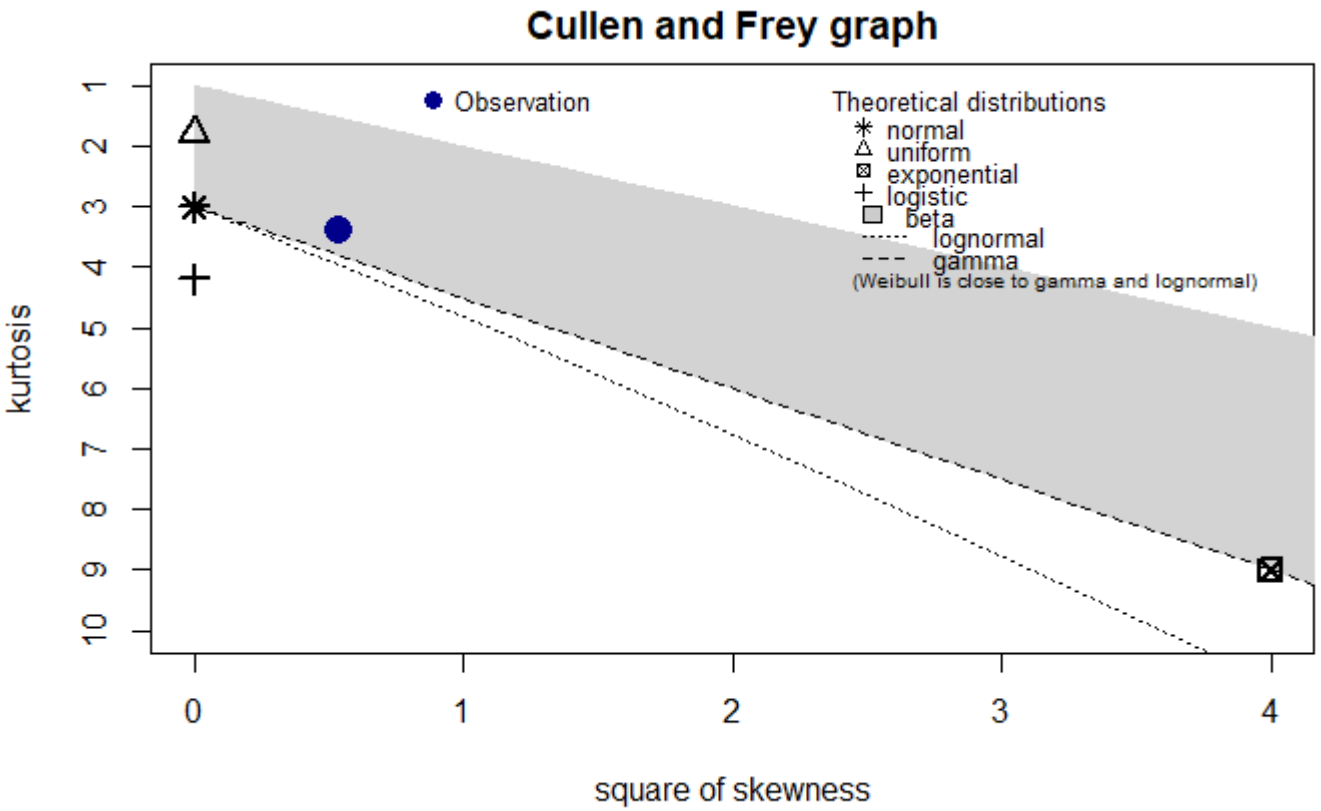


Next step was to determin distribution. Our first and pretty natural assumption was that grades are normally distributed, but we decided to predict other close distribution.

Hide

```
# we use this function to gain some idea about possible candidate distribution
final.grade <- df$G3
descdist(final.grade)
```

```
summary statistics
------
min:  0   max:  20
median:  11
mean:  10.41519
estimated sd:  4.581443
estimated skewness:  -0.7326724
estimated kurtosis:  3.403421
```



It may seem that our distribution is close to lognormal or gamma, however, our data contains zeros, so, in that case, we will fit our data only to normal and logistic, which are also close

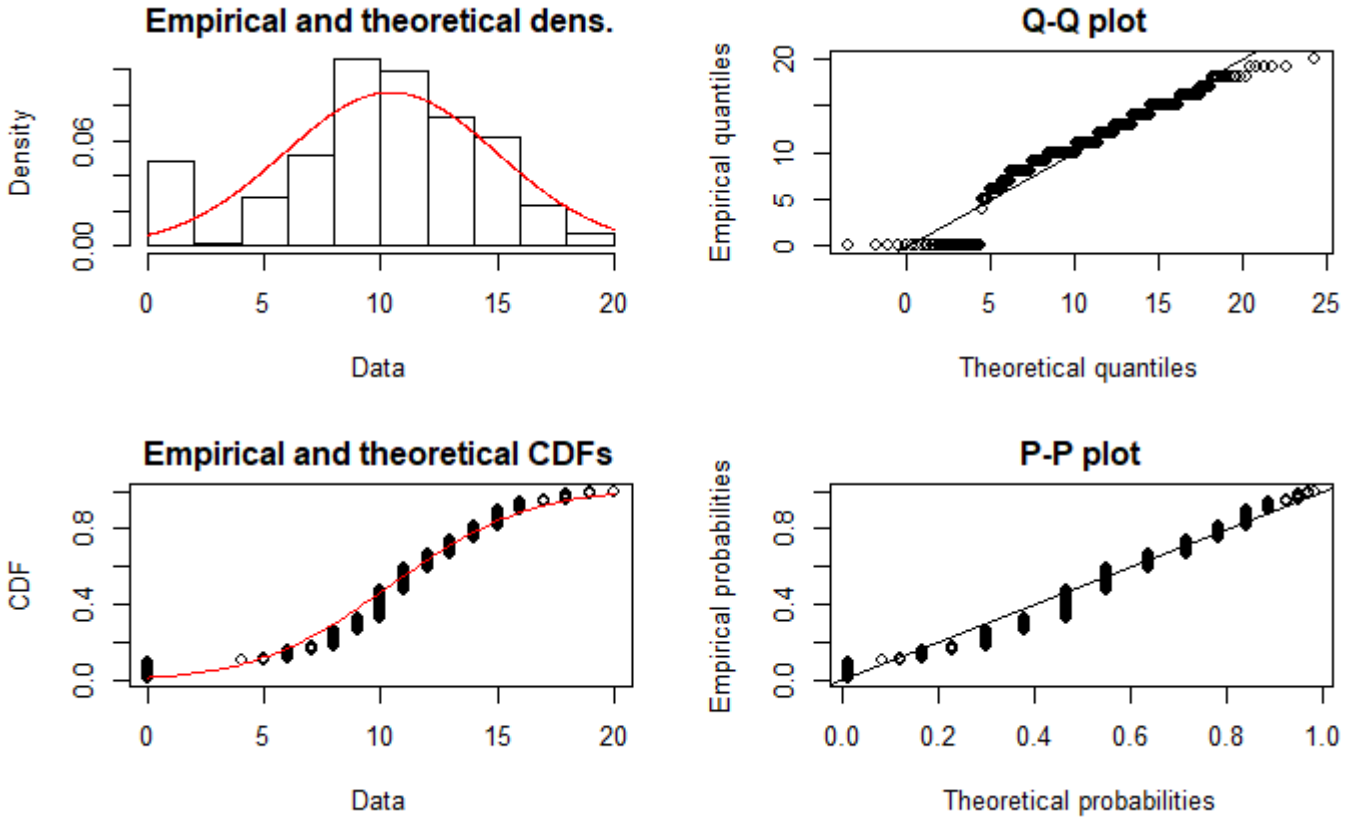
Hide

```
fit.norm <- fitdist(final.grade, "norm")
```

```
NaNs producedNaNs produced
```

Hide

```
plot(fit.norm)
```



Both distributions look good, however, according to plots, logistic may seem more closer

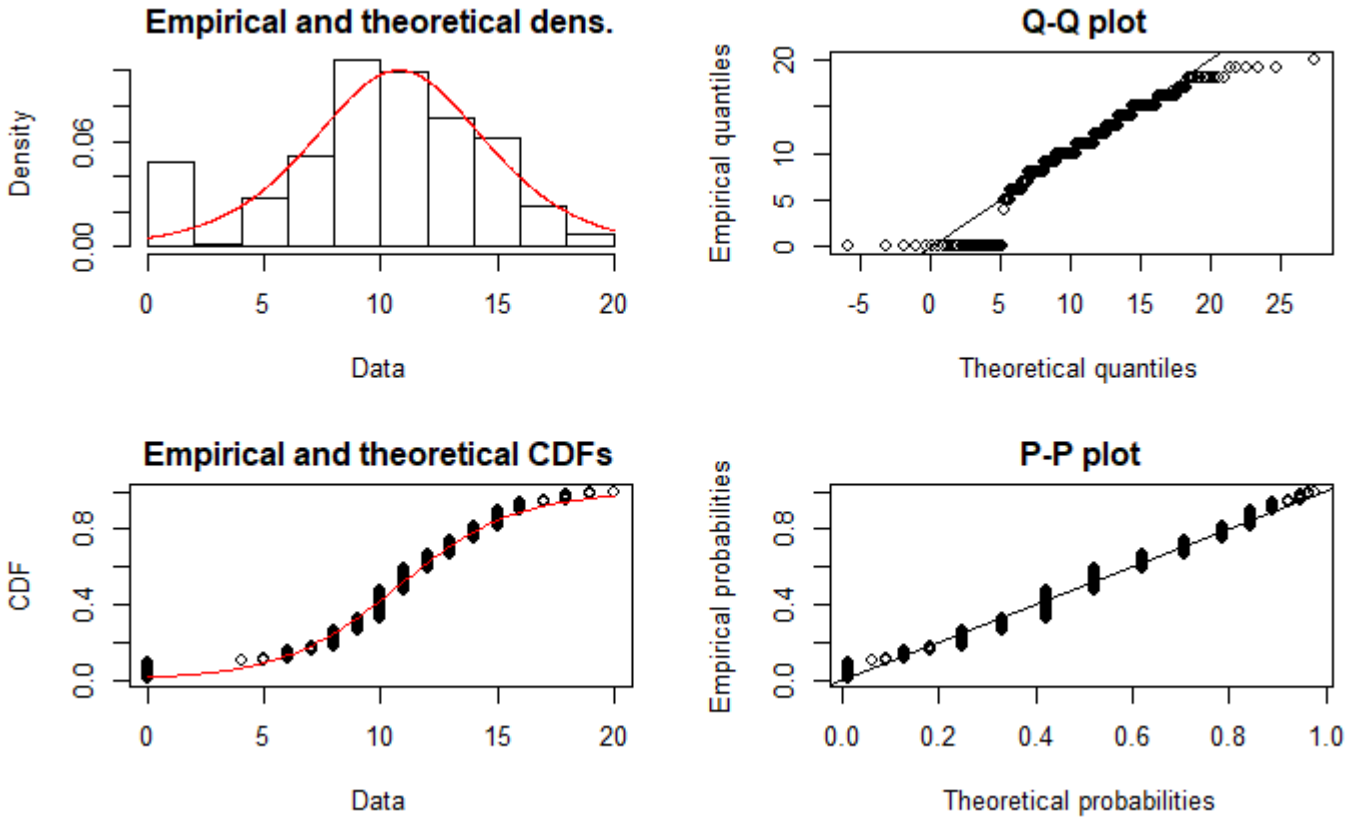
Hide

```
fit.logis <- fitdist(final.grade, "logis")
```

```
NaNs producedNaNs produced
```

Hide

```
plot(fit.logis)
```



Now we simply use the Akaike information criterion to see which distribution is more precise. It's an estimator of the relative quality of statistical models for our data. The smallest the estimation is, the higher quality has the model. By this criterion we see that logistic distribution is quite closer.

Hide

```
fit.norm$aic
```

```
[1] 2326.351
```

Hide

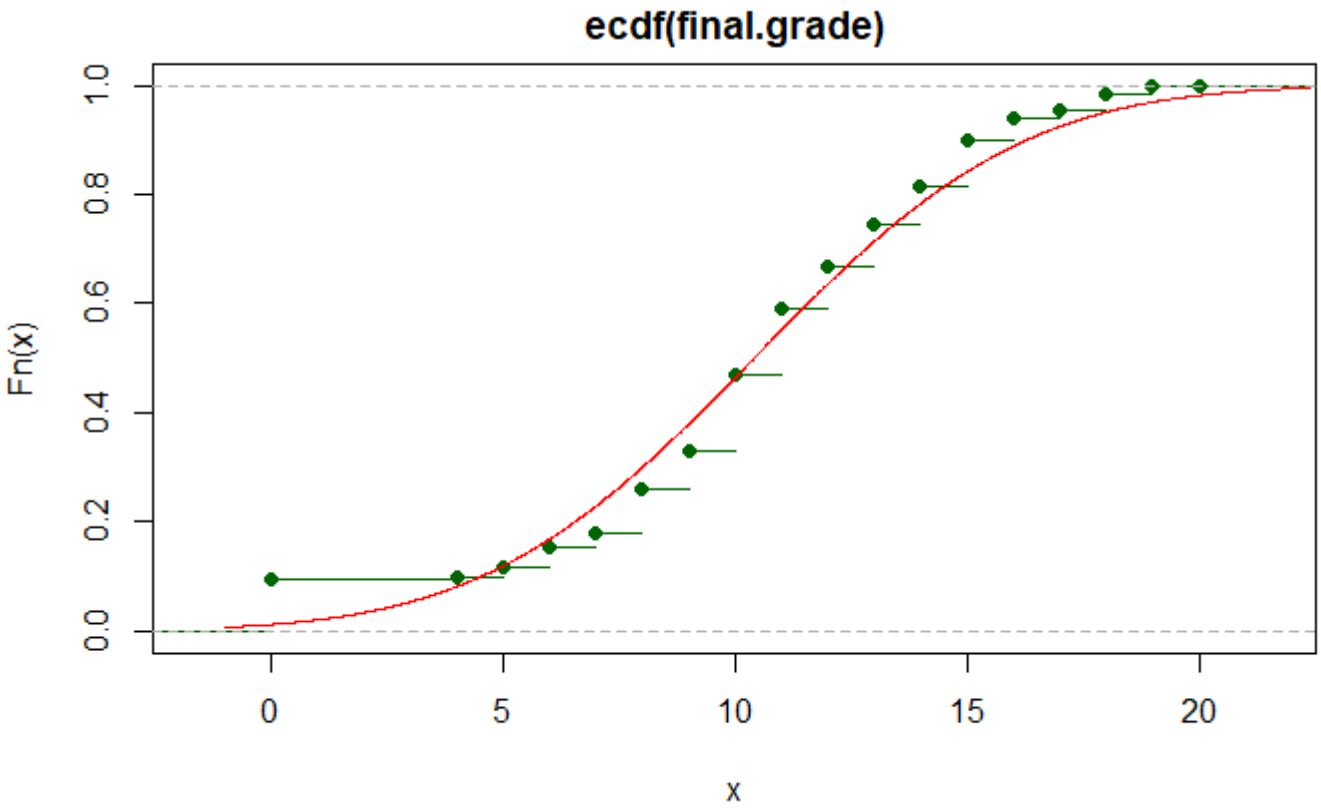
```
fit.logis$aic
```

```
[1] 2315.062
```

As a result we got one more hypothesys to test. We will test logistic distribution in addition to the normal using another method - Kolmogorov-Smirnov test:

Hide

```
N <- 200
mu <- mean(final.grade)
sd <- sd(final.grade)
x <- rnorm(N, mean=mu, sd=sd)
pts <- seq(-1,max(x),by=0.01)
plot(ecdf(final.grade),col="darkgreen")
lines(pts, pnorm(pts, mean=mu, sd=sd), col="red")
```



Hide

```
max(pnorm(pts, mean=mu, sd=sd)-ecdf(final.grade)(pts))
```

```
[1] 0.1339145
```

Hide

```
ks.test(final.grade, "pnorm", mu, sd)
```

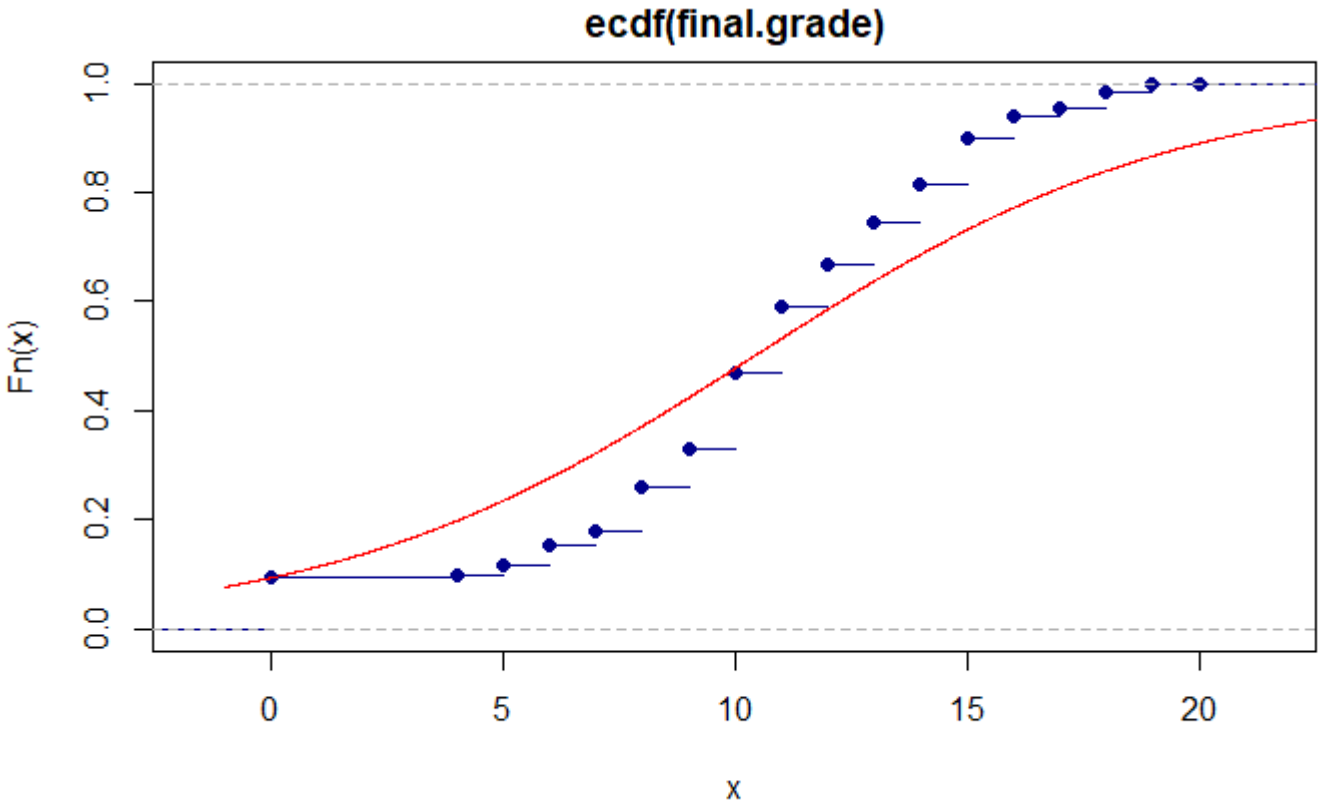
```
ties should not be present for the Kolmogorov-Smirnov test
```

```

    One-sample Kolmogorov-Smirnov test

data:  final.grade
D = 0.13478, p-value = 1.171e-06
alternative hypothesis: two-sided
```

Here we make same test for assumption about logistic distribution. We can see that plot for normal distribution and results of Kolmogorov-Smirnov test show that normal distibution is more likely to fit our data.

Hide

```
x <- rlogis(N, location=mu, scale=sd)
pts <- seq(-1,max(x),by=0.01)
plot(ecdf(final.grade),col="darkblue")
lines(pts, plogis(pts, location=mu, scale=sd), col="red")
```



Hide

```
max(plogis(pts, location=mu, scale=sd)-ecdf(final.grade)(pts))
```

```
[1] 0.1934531
```

Hide

```
ks.test(final.grade, "plogis", mu, sd)
```

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

```

    One-sample Kolmogorov-Smirnov test

data:  final.grade
D = 0.19396, p-value = 2.474e-13
alternative hypothesis: two-sided
```

Having analized our main data we decided to check the influence of other factors on grades. After small analyses that we made drowing histograms, that we will skip here, we understood that education of mother has interesting influence.

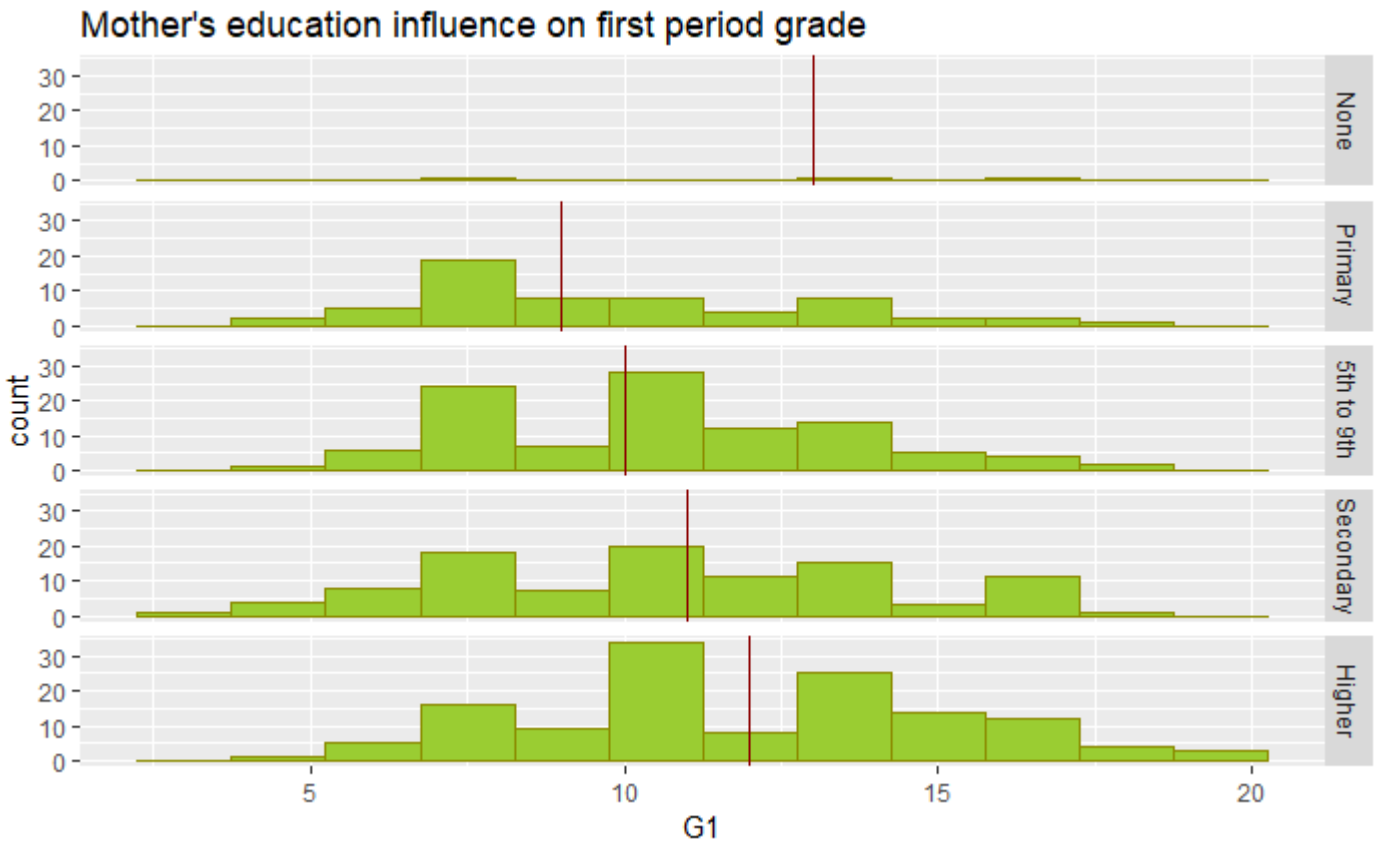
Hide

```
df$Medu <- as.factor(df$Medu)
df$Medu <- mapvalues(df$Medu,
                                from = 0:4,
                                to = c("None", "Primary", "5th to 9th ", "Secondary", "Higher"))
df$M
```

```
NULL
```

Hide

```
ggplot(df, aes(x=G1)) + geom_histogram(fill="yellowgreen", color="yellow4",binwidth = 1.5) +
facet_grid(Medu ~ .)+geom_vline(data=aggregate(df[31], df[7], median), mapping=aes(xintercept=G
1), color="red4")+ggtitle("Mother's education influence on first period grade")
```



Hide

```
df$Medu <- as.factor(df$Medu)
df$Medu <- mapvalues(df$Medu,
                                from = 0:4,
                                to = c("None", "Primary", "5th to 9th ", "Secondary", "Higher"))
```

```
The following `from` values were not present in `x`: 0, 1, 2, 3, 4
```

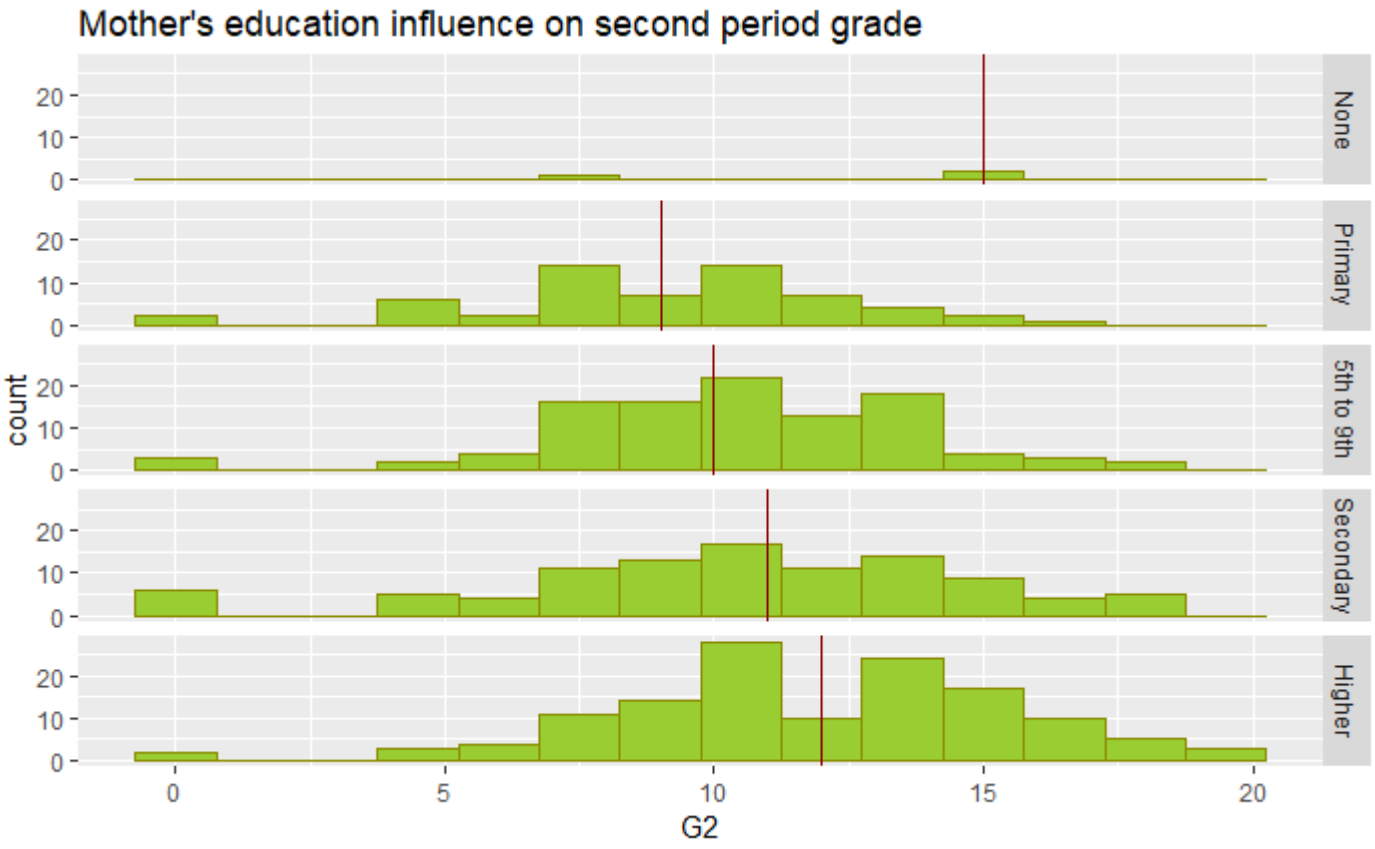
Hide

```
df$M
```

```
NULL
```

Hide

```
ggplot(df, aes(x=G2)) + geom_histogram(fill="yellowgreen", color="yellow4",binwidth = 1.5) +
facet_grid(Medu ~ .)+geom_vline(data=aggregate(df[32], df[7], median), mapping=aes(xintercept=G
2), color="red4")+ggtitle("Mother's education influence on second period grade")
```



Hide

```
df$Medu <- as.factor(df$Medu)
df$Medu <- mapvalues(df$Medu,
                                from = 0:4,
                                to = c("None", "Primary", "5th to 9th ", "Secondary", "Higher"))
```

```
The following `from` values were not present in `x`: 0, 1, 2, 3, 4
```

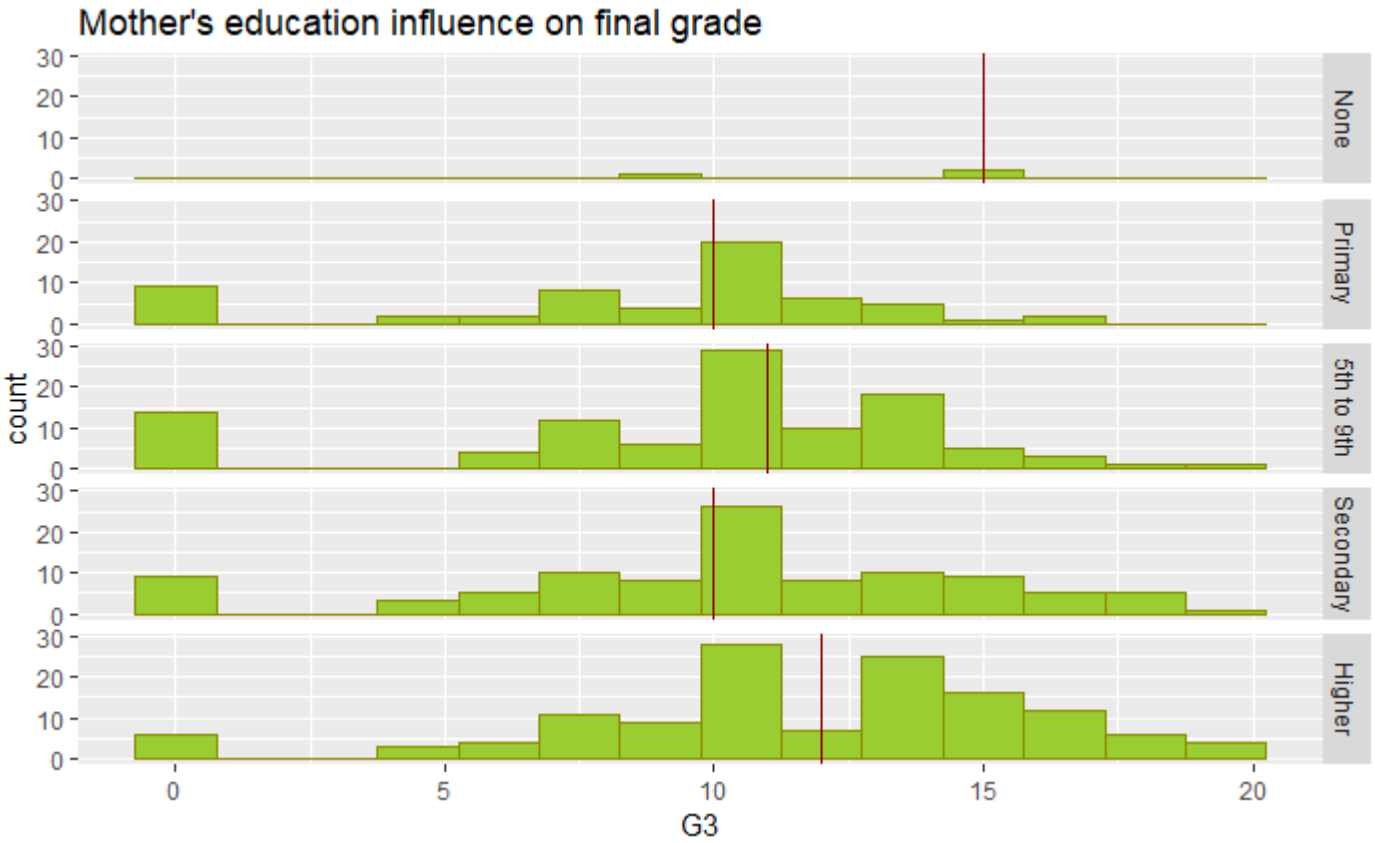
Hide

```
df$M
```

```
NULL
```

Hide

```
ggplot(df, aes(x=G3)) + geom_histogram(fill="yellowgreen", color="yellow4",binwidth = 1.5) +
facet_grid(Medu ~ .)+geom_vline(data=aggregate(df[33], df[7], median), mapping=aes(xintercept=G
3), color="red4")+ggtitle("Mother's education influence on final grade")
```

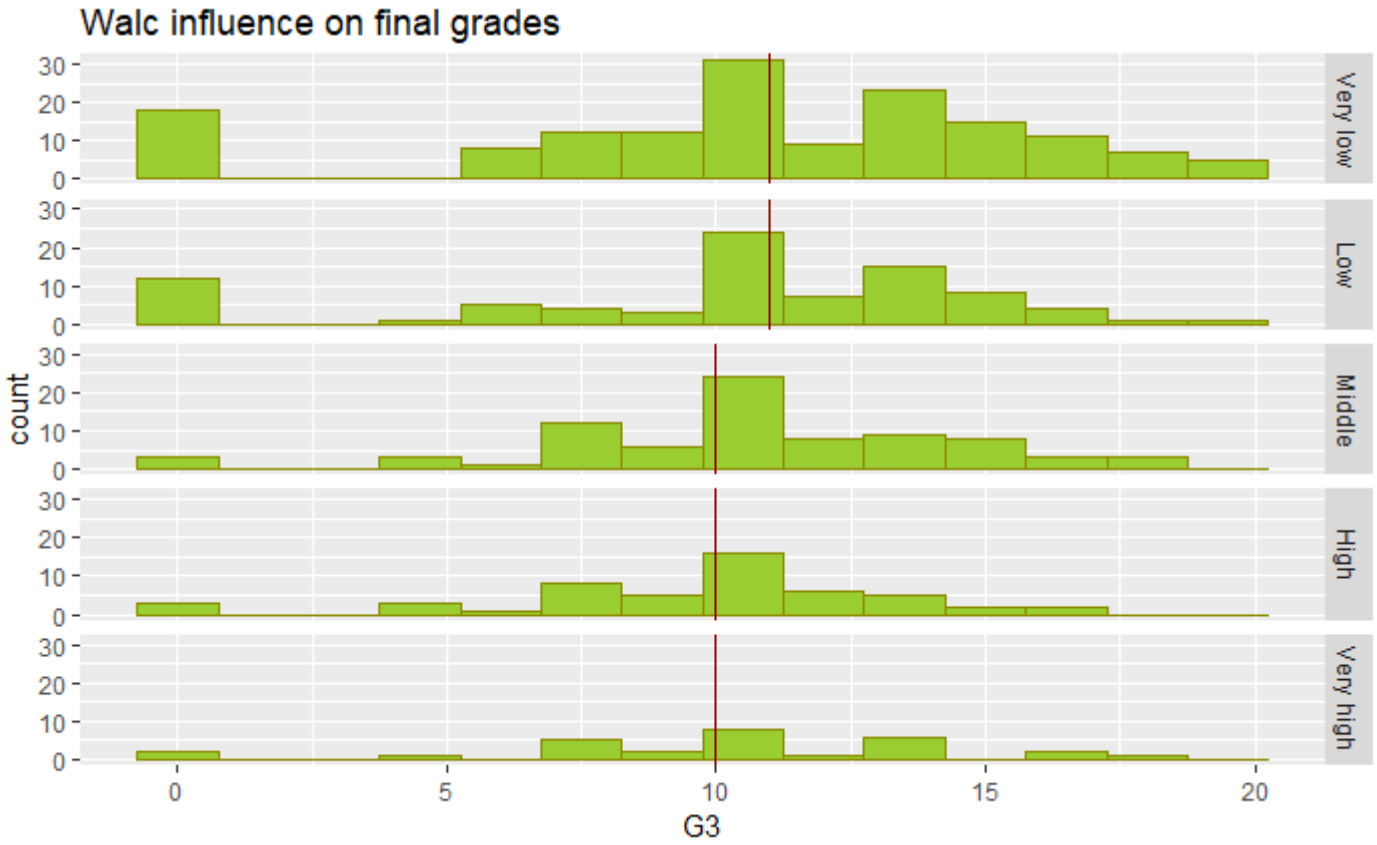


Visualization shows us that education of mother has big influence on grades of students.

We made same analyses to check influence of alcohol comsumption on children.

Hide

```
df$Walc <- as.factor(df$Walc)
df$Walc <- mapvalues(df$Walc,
                             from = 1:5,
                             to = c("Very low", "Low", "Middle", "High", "Very high"))
ggplot(df, aes(x=G3)) + geom_histogram(fill="yellowgreen", color="yellow4",binwidth = 1.5) +
facet_grid(Walc ~ .)+geom_vline(data=aggregate(df[33], df[28], median), mapping=aes(xintercept=G
3), color="red4")+ggtitle("Walc influence on final grades")
```

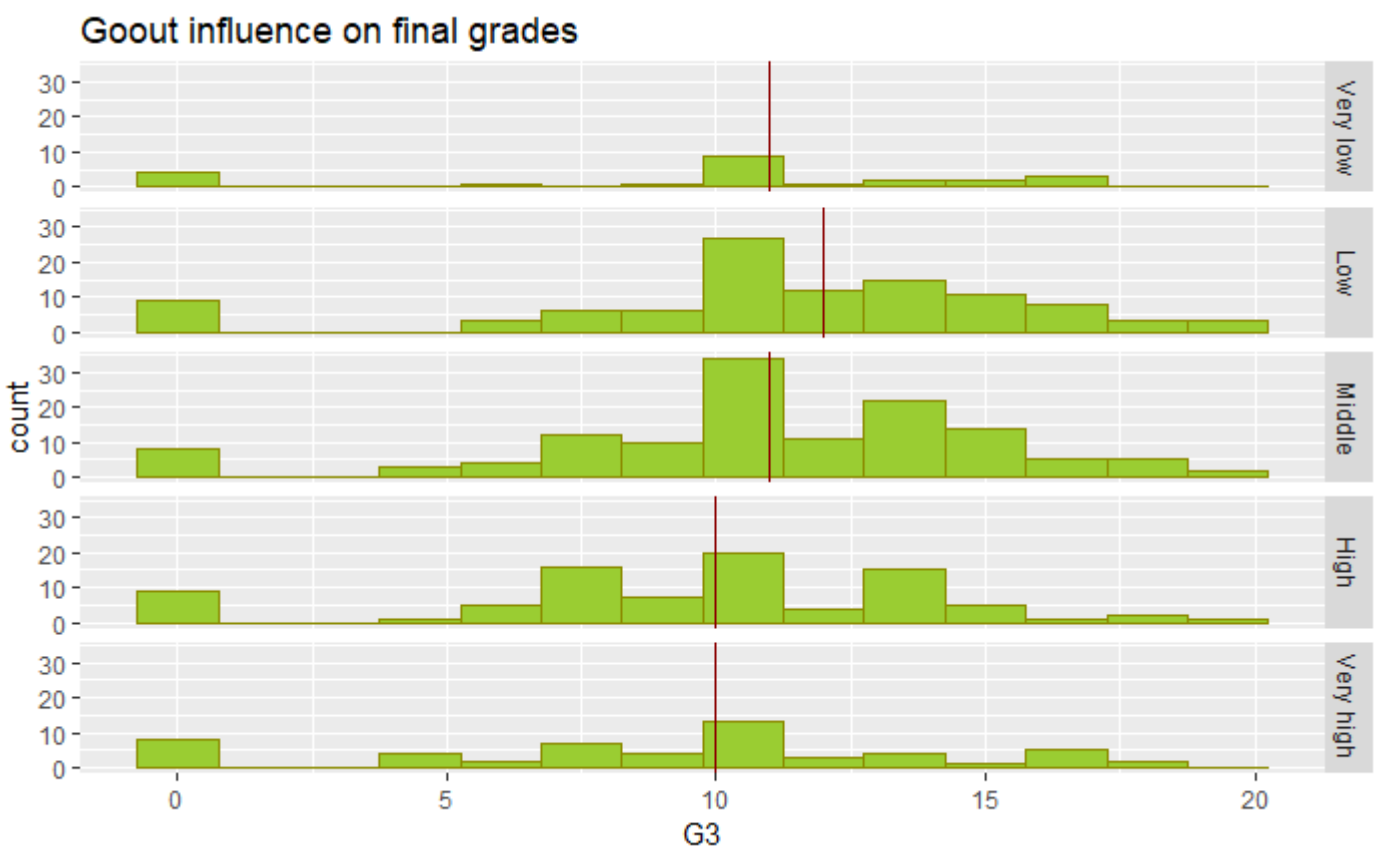


Here we see only a small progress in studing by children whose level of weekly alcohol consumption is very low.

The next step was to analyze influence of going out with friends.

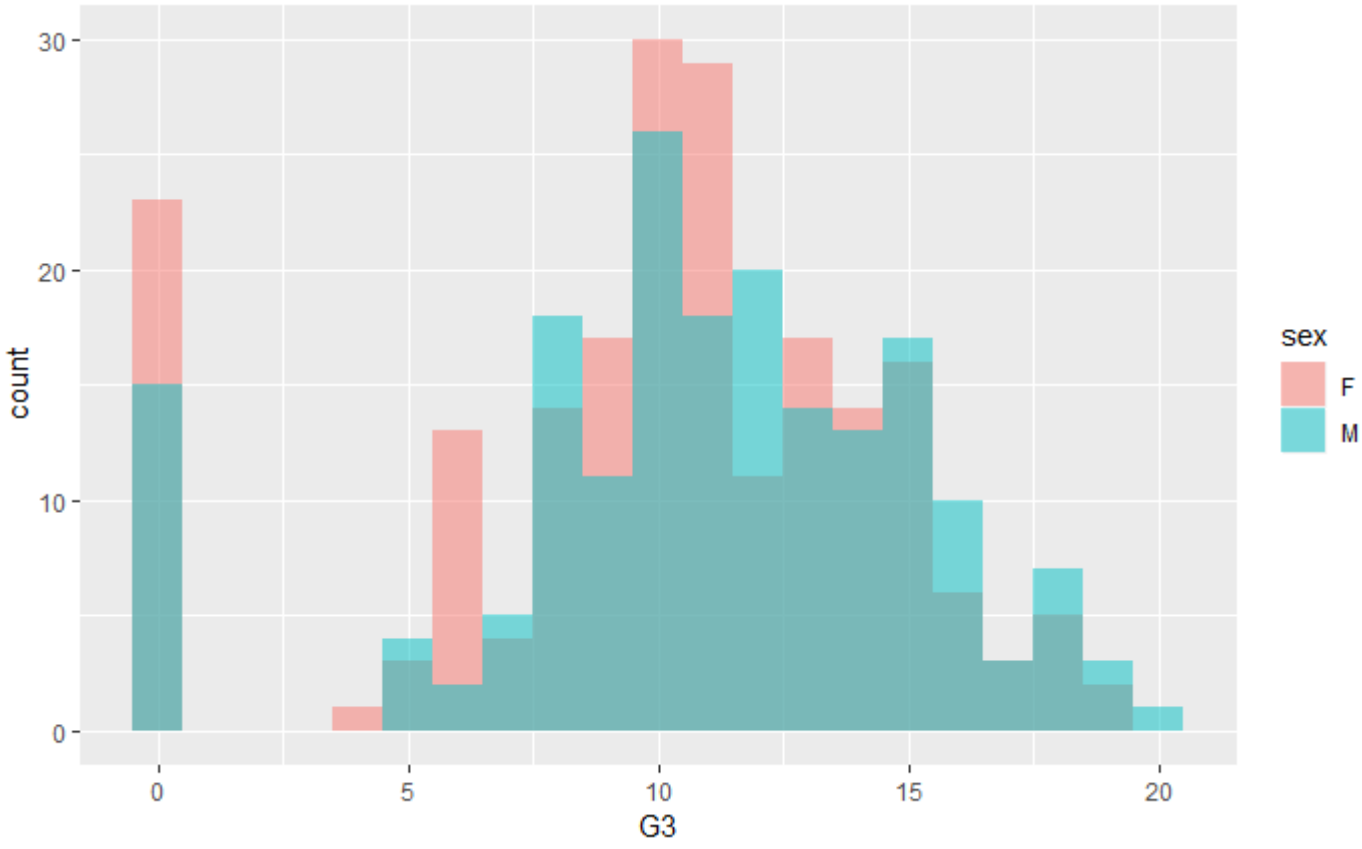
Hide

```
df$goout <- as.factor(df$goout)
df$goout <- mapvalues(df$goout,
                              from = 1:5,
                              to = c("Very low", "Low", "Middle", "High", "Very high"))
ggplot(df, aes(x=G3)) + geom_histogram(fill="yellowgreen", color="yellow4",binwidth = 1.5) +
facet_grid(goout ~ .)+geom_vline(data=aggregate(df[33], df[26], median), mapping=aes(xintercept=
G3), color="red4")+ggtitle("Goout influence on final grades")
```



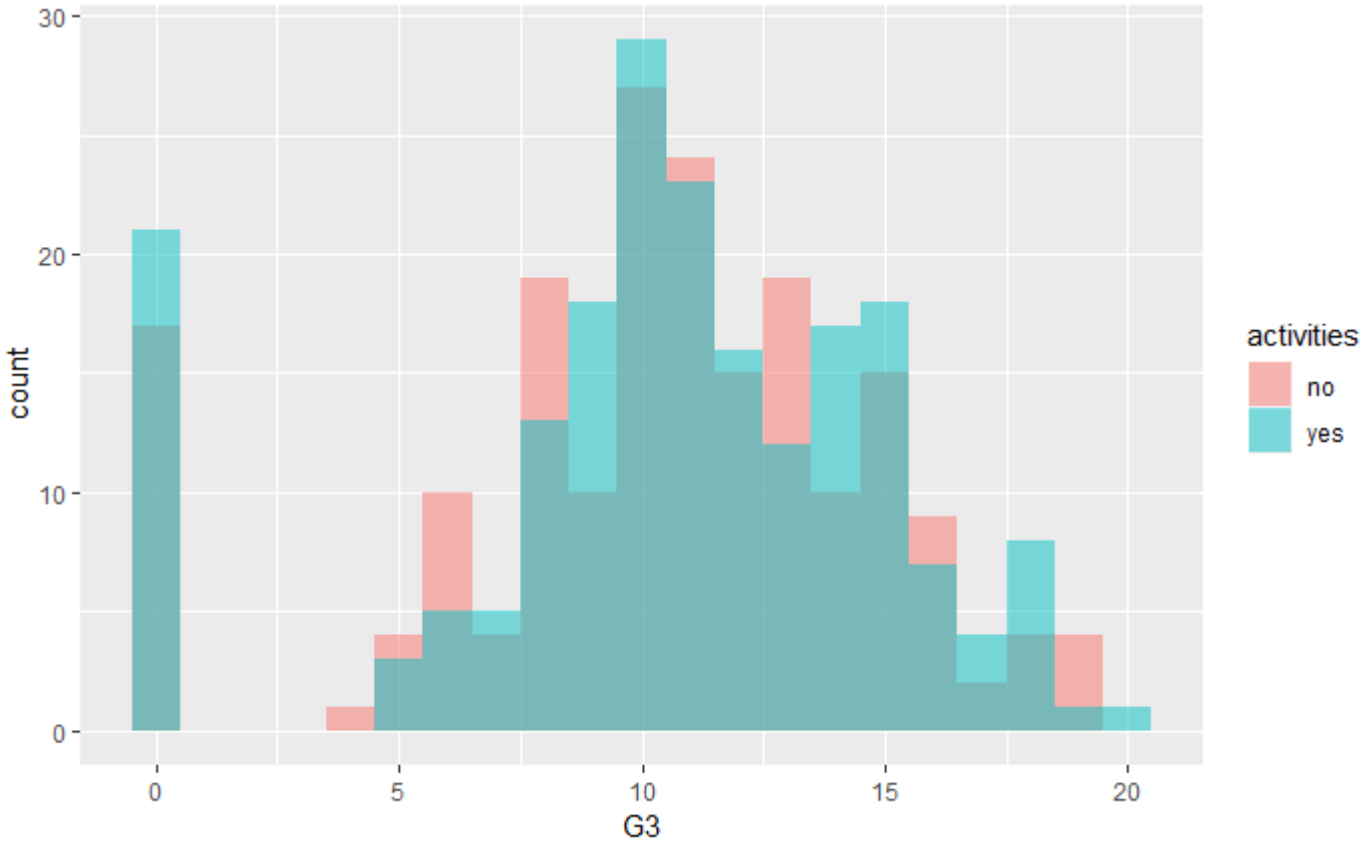
Hide

```
ggplot(df, aes(x=G3, fill=sex)) +
geom_histogram(position="identity", alpha=0.5, binwidth=1)
```



Hide

```
ggplot(df, aes(x=G3, fill=activities)) +
geom_histogram(position="identity", alpha=0.5,binwidth=1)
```


30
20
10
0
count
0
5
10
15
20
G3
activities
no
yes