Project1

Yahui Li

2020/1/13

Step1 Read Data

```
#install.packages("AER")
library(AER)
```

```
## Warning: package 'AER' was built under R version 3.6.2

## Warning: package 'car' was built under R version 3.6.1

## Warning: package 'lmtest' was built under R version 3.6.1

## Warning: package 'zoo' was built under R version 3.6.1

## Warning: package 'sandwich' was built under R version 3.6.1
```

```
data("STAR")
```

Step2 Explore Data

We will only examine the math scores in 1st grade in this project.

```
data <- data.frame(star1 = STAR$star1, math1 = STAR$math1)

sapply(data,class)
```

```
##     star1     math1
##  "factor" "integer"
```

```
sapply(data,summary)
```

```
## $star1
##      regular        small regular+aide         NA's
##         2584         1925         2320         4769
##
## $math1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   404.0   500.0   529.0   530.5   557.0   676.0    4998
```

```
data.star1.na <- data[is.na(data$star1),]
all(is.na(data.star1.na$math1))
```

```
## [1] TRUE
```

Which shows that the math score has not been recorded if class type is not recorded. So we can remove the data where star1 is NA.

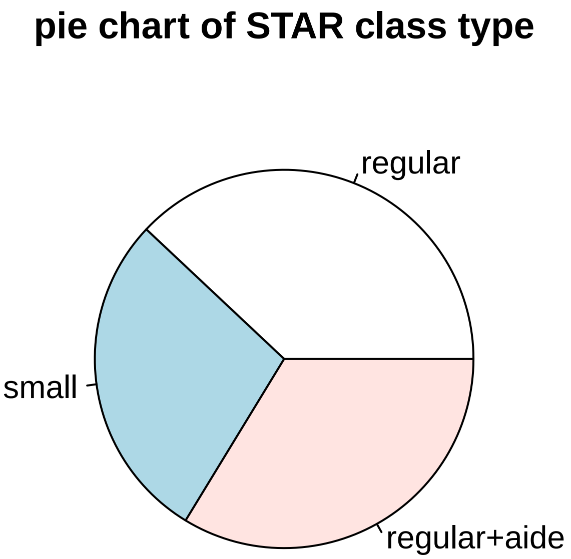
One of the way to deal with NA in math1 is to remove them

```
data_remove_na <- na.omit(data[-is.na(data$star1),])
```

```
table(data_remove_na$star1)
```

```
##
##      regular        small regular+aide
##         2507         1868         2225
```

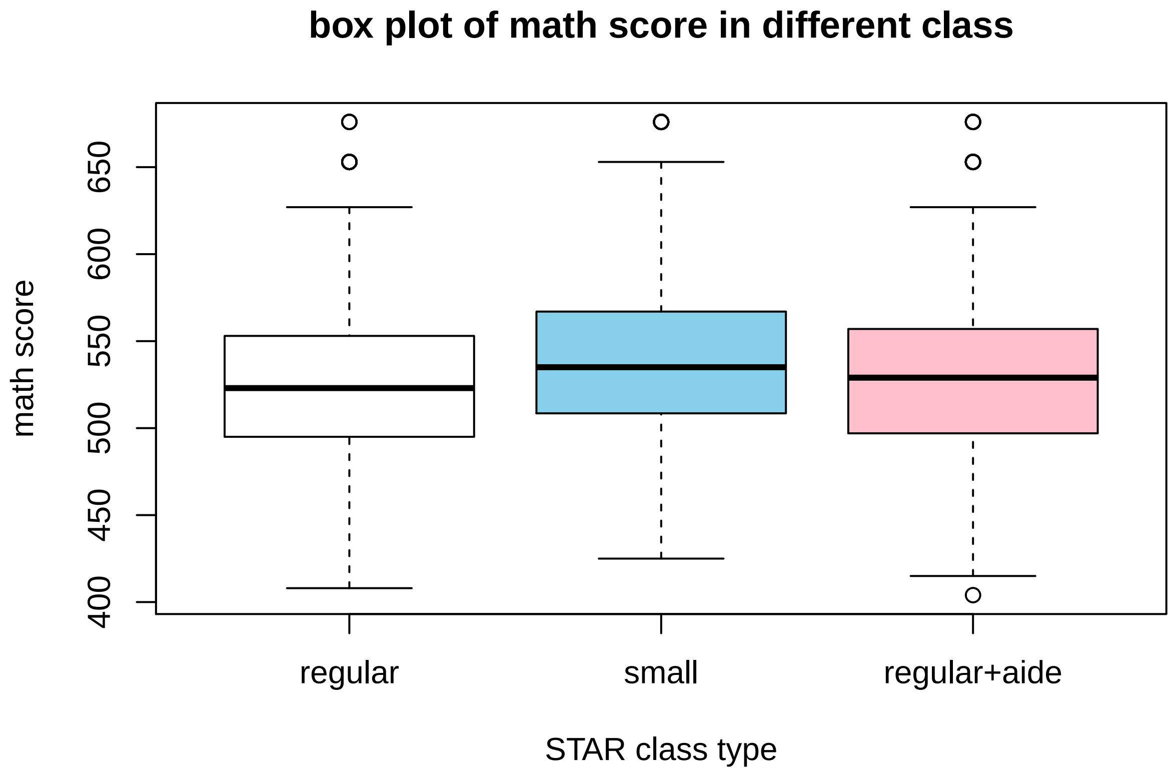
```
pie(table(data_remove_na$star1),main = "pie chart of STAR class type")
```



```
tapply(data_remove_na$math1, data_remove_na$star1,summary)
```

```
## $regular
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   408.0   495.0   523.0   525.3   553.0   676.0
##
## $small
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   425.0   509.2   535.0   538.7   567.0   676.0
##
## $`regular+aide`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   404.0   497.0   529.0   529.6   557.0   676.0
```

```
boxplot(data$math1~data$star1,main = "box plot of math score in different class",
        xlab = "STAR class type", ylab = "math score", col = c("white", "skyblue", "pink"))
```



From the result,

for mean, small > regular+aide > regular;

for all quantile information, small > regular+aide > regular;

for min, small > other two; For max, they are the same.

Something interesting: there are only some certain scores like 601 612 627 653 676.

Step3 One Way ANOVA Model

$$Y_{ij} = \mu_1 + \tau_2 X_{2,ij} + \tau_3 X_{3,ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2), \quad i = 1, 2, 3, j = 1, \cdots, n_i.$$

where $i = 1$ means the class type in 1st grade is regular; $i = 2$ means the class type in 1 st grade is small; $i = 3$ means the class type in 1st grade is regular-with-aide.

From the table in step2, $n_1 = 2507, n_2 = 1868, n_3 = 2225, n = 6600$.

$X_{2,ij} = 1$ if $i = 2$, otherwise $X_{2,ij} = 0$. $X_{3,ij} = 1$ if $i = 3$, otherwise $X_{3,ij} = 0$.

μ_i means the population mean of the i-th type class in 1st grade, $i = 1, 2, 3$.

$\tau_i = \mu_i - \mu_1$ means the difference in population mean between i-th type and first type in 1st grade, $i = 2, 3$.

ϵ_{ij} is the random variable about error which is Normal with 0 mean and σ^2 variance under assumption.

bingdao want to add Y_{ij} denotes the math grade in 1st grade of the j-th experimental unit in the i-th class type.

ϵ_{ij} is the random variable and assumed to be i.i.d. Normal(0,σ^2).

Model Assumption

(a) Response variable residuals are normally distributed.

(b) Variances of populations are equal.

(c) Responses for a given group are independent and identically distributed normal random variables.

All of the assumptions are necessary, because for each population violate the normal distribution, F-tests are not robust. Moreover, if the assumption of homoscedasticity is violated, the Type I error properties degenerate much more severely.[5] (Randolf, E. A.; Barcikowski, R. S. (1989). “Type I error rate when real study values are used as population parameters in a Monte Carlo study”. Paper presented at the 11th annual meeting of the Mid-Western Educational Research Association, Chicago.)

Step4 Appropriate

Before we fit the model, we need to ensure that model is appropriate on this dataset, that is, the response variable satisfies the assumptions of our model. In other words, we will check the normality and equal variance of the response varibales.

We first make a density plot and a Q-Q plot to check the normality of “math1”

```
#hist(data_remove_na$math1, main = "histogram of math socre in 1st grade")
#qqnorm(data_remove_na$math1)
#qqline(data_remove_na$math1)
```

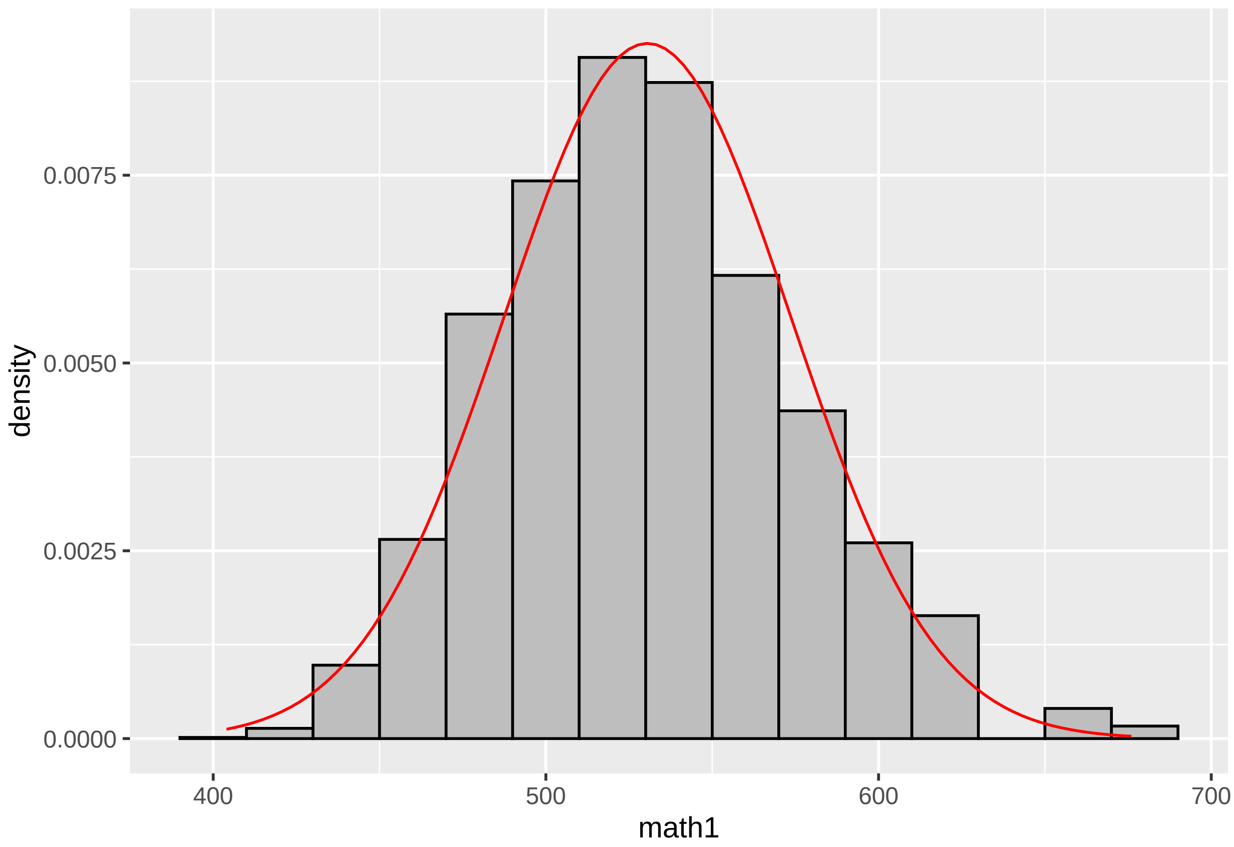
bingdao version

```
library(ggplot2)
```

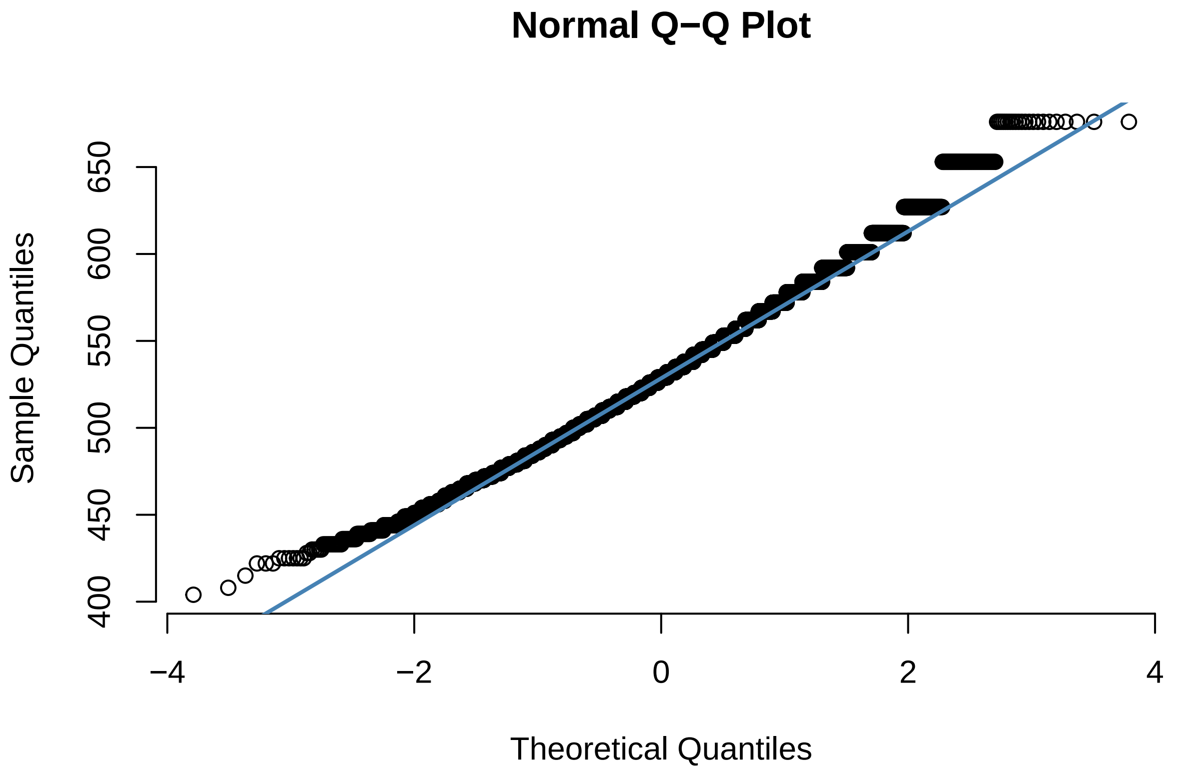
```
## Warning: package 'ggplot2' was built under R version 3.6.1
```

```
x <- seq(404, 676, length.out=100)
df <- with(data_remove_na, data.frame(x = x, y = dnorm(x, mean(math1), sd(math1))))

ggplot(data_remove_na, aes(x=math1, y = ..density..)) +
  geom_histogram(binwidth = 20, fill = "grey", color = "black") +
  geom_line(data = df, aes(x = x, y = y), color = "red")
```



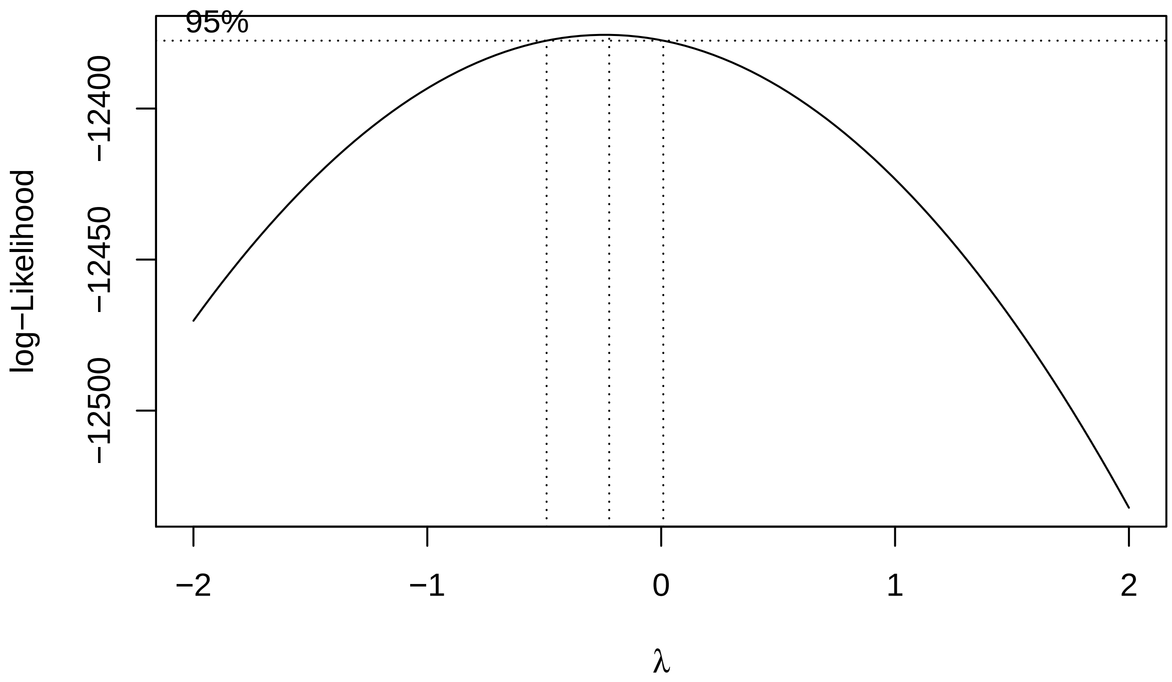
```r
qqnorm(data_remove_na$math1, pch = 1, frame = FALSE)
qqline(data_remove_na$math1, col = "steelblue", lwd = 2)
```



The histogram shows that it seems normal distributed. The Q-Q plot shows the the distribution of math score is right-skewed.

So we use box cox method to make a transformation on math1.

```
library(MASS)
boxcox(math1 ~ star1 , data = data_remove_na)
```



It indicates that we need make a log-transformation for "math1"

```
#hist(log(data_remove_na$math1), main = "histogram of math socre in 1st grade")
#qqnorm(log(data_remove_na$math1))
#qqline(log(data_remove_na$math1))
```

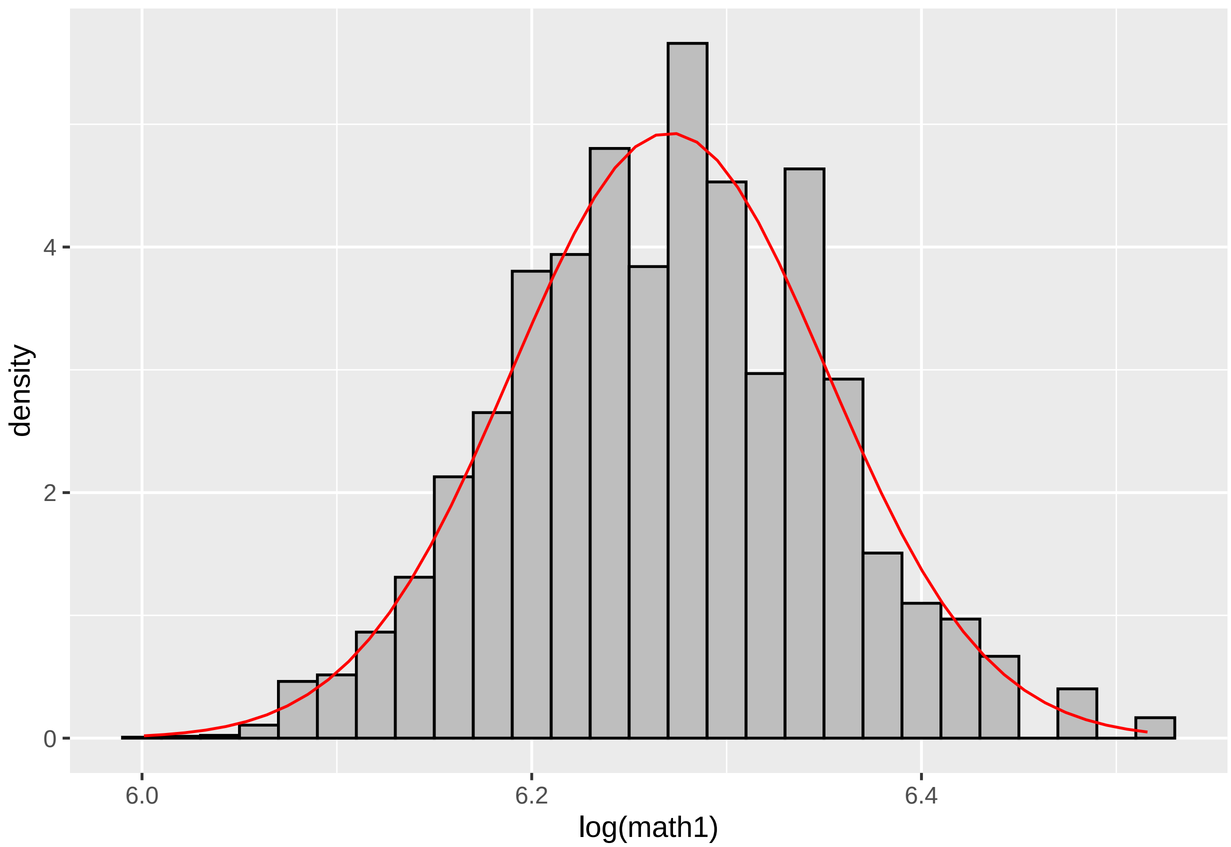
bingdao version

```
summary(log(data_remove_na$math1))
```

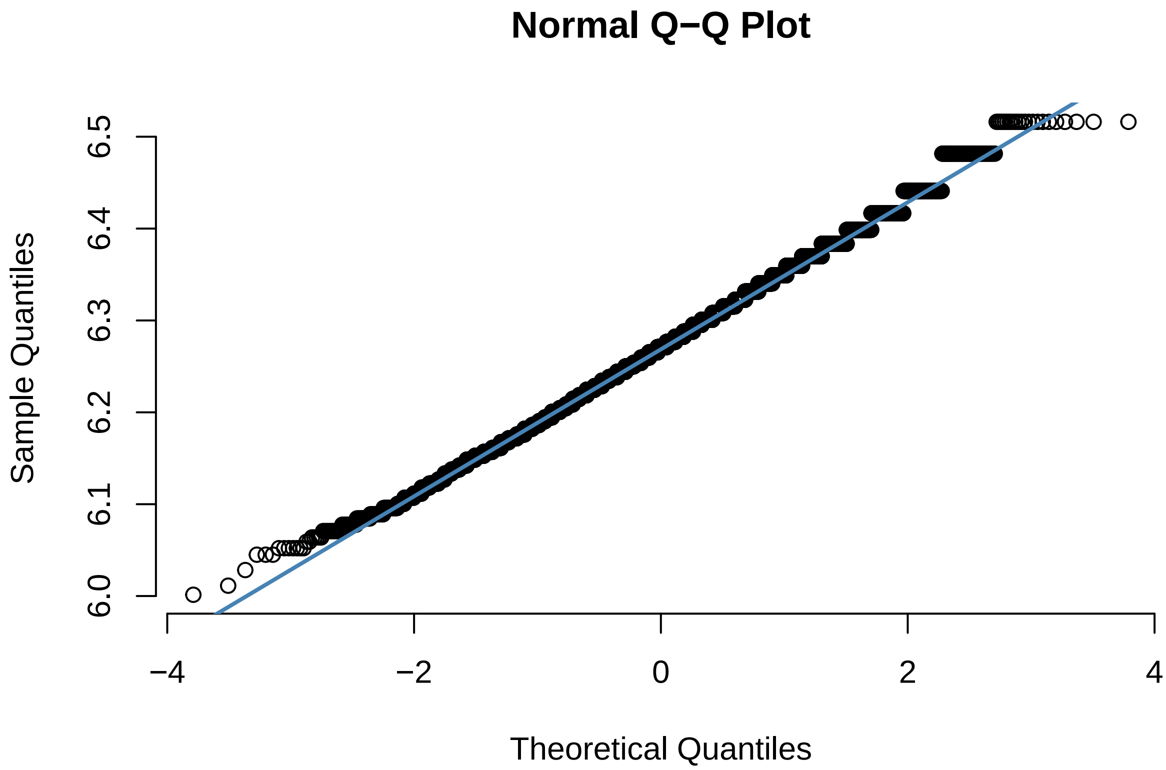
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.001   6.215   6.271   6.271   6.323   6.516
```

```
x <- seq(6.001, 6.516, length.out=50)
df <- with(data_remove_na, data.frame(x = x, y = dnorm(x, mean(log(math1)), sd(log(math1)))))

ggplot(data_remove_na, aes(x=log(math1), y = ..density..)) +
  geom_histogram(binwidth = 0.02, fill = "grey", color = "black") +
  geom_line(data = df, aes(x = x, y = y), color = "red")
```



```r
qqnorm(log(data_remove_na$math1), pch = 1, frame = FALSE)
qqline(log(data_remove_na$math1), col = "steelblue", lwd = 2)
```



The graph shows the distribution of log math score in 1st grade is Normal-like.

bingdao version Then we calculate the variance of math grade in 1st grade of each class type.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.1
```

```
## -- Attaching packages --------------------------------------------- tidyverse 1.2.1 --
```

```
## v tibble  2.1.3     v purrr   0.3.2
## v tidyr   1.0.0     v dplyr   0.8.3
## v readr   1.3.1     v stringr 1.4.0
## v tibble  2.1.3     v forcats 0.4.0
```

```
## Warning: package 'tibble' was built under R version 3.6.1
```

```
## Warning: package 'tidyr' was built under R version 3.6.1
```

```
## Warning: package 'readr' was built under R version 3.6.1
```

```
## Warning: package 'purrr' was built under R version 3.6.1
```

```
## Warning: package 'dplyr' was built under R version 3.6.1
```

```
## Warning: package 'stringr' was built under R version 3.6.1
```

```
## Warning: package 'forcats' was built under R version 3.6.1
```

```
## -- Conflicts ------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::recode() masks car::recode()
## x dplyr::select() masks MASS::select()
## x purrr::some()   masks car::some()
```

```
data_remove_na %>%
  group_by(star1) %>%
  summarize(var_math1 = var(log(math1), na.rm = T))
```

```
## # A tibble: 3 x 2
##   star1         var_math1
##   <fct>             <dbl>
## 1 regular         0.00625
## 2 small           0.00666
## 3 regular+aide    0.00650
```

The result shows that they are very small and nearly equal to each other. Therefore, it is appropiate to run our model on this dataset.

Step5 Fit Model

```
anova.fit<- aov(log(math1)~star1,data=data_remove_na)
summary(anova.fit)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## star1          2   0.68  0.3391   52.56 <2e-16 ***
## Residuals   6597  42.55  0.0065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova.fit$coefficients
```

```
##       (Intercept)        star1small star1regular+aide
##        6.26079457        0.02499081        0.00812069
```

From the result, the fitted model we get is:

$$\log \hat{Y}_{ij} = 6.2608 + 0.0250X_{2,ij} + 0.0081X_{3,ij}$$

with means when the type is regular, the estimate math score is $e^{6.2608} = 523.6377$; when the type is small, the estimate math score is $e^{6.2608+0.0250} = 536.8936$; when the type is regular-with-aide, the estimate math score is $e^{6.2608+0.0081} = 527.8964$.