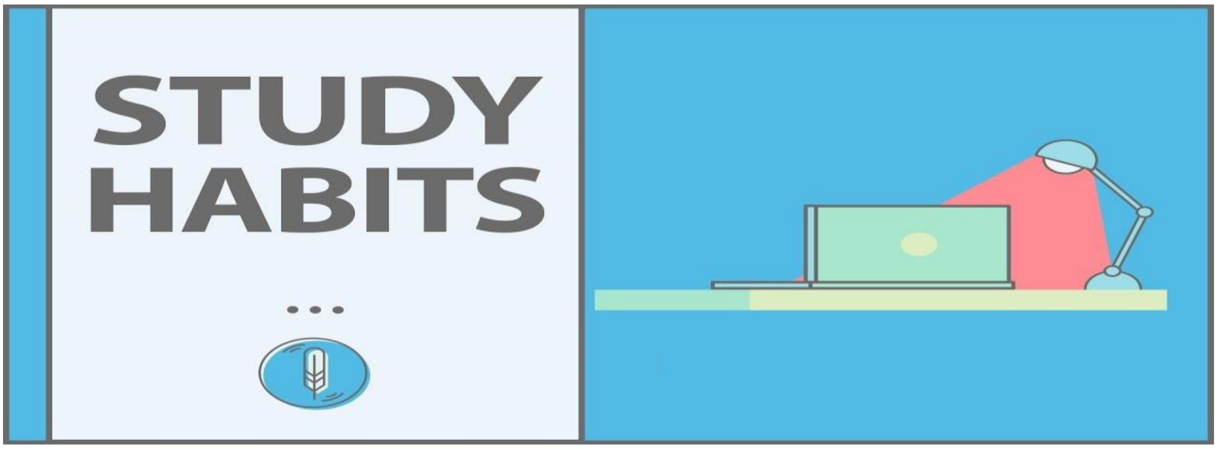
IDA PROJECT
DATA ANALYSIS AND HYPOTHESES



Submitted by:

EL HANAFI Maha
Younes Diba
Achraf Attary
Imane Rammouch
Aymen Bouhair

Supervised by:

GHOGHO Mounir
SBIHI Nada

Contents

Special thanks

This project has been an invaluable opportunity for us to learn more about data analysis.

So we would like to dedicate this work to:

Our teachers

Mr Mounir GHOGHO who introduced us to the field of data analysis, and made us fall in love with it by giving us clear explanations through the semester.

Mme Nada Sbihi who helped us massively with the practical side of this field, by showing us how to use the technologies of data analysis.

Our friends

For their encouragement and support even when times are tough.

Our parents

For being our guiding light in this dark world.

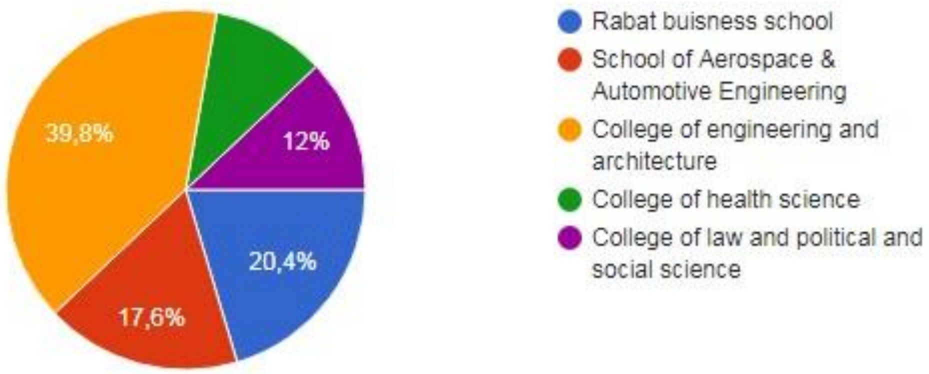
General context for the project

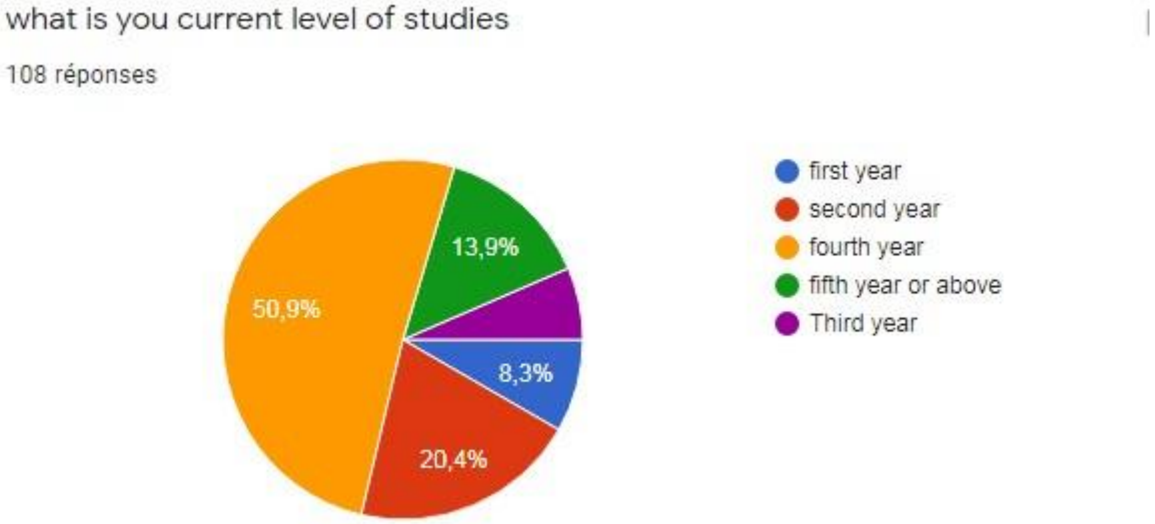
The definition of study habits is the habitual practices one uses to help them study and learn. Good study habits can help students achieve and/or maintain good grades.

The purpose of this project is to help us get valuable data about study habits of the students of the international university of Rabat. The data can be used to find patterns in the behavior of the students, in order for us to identify areas of strength and potential changes. All answers are completely anonymous; no names or email are being collected.

Data Analysis:

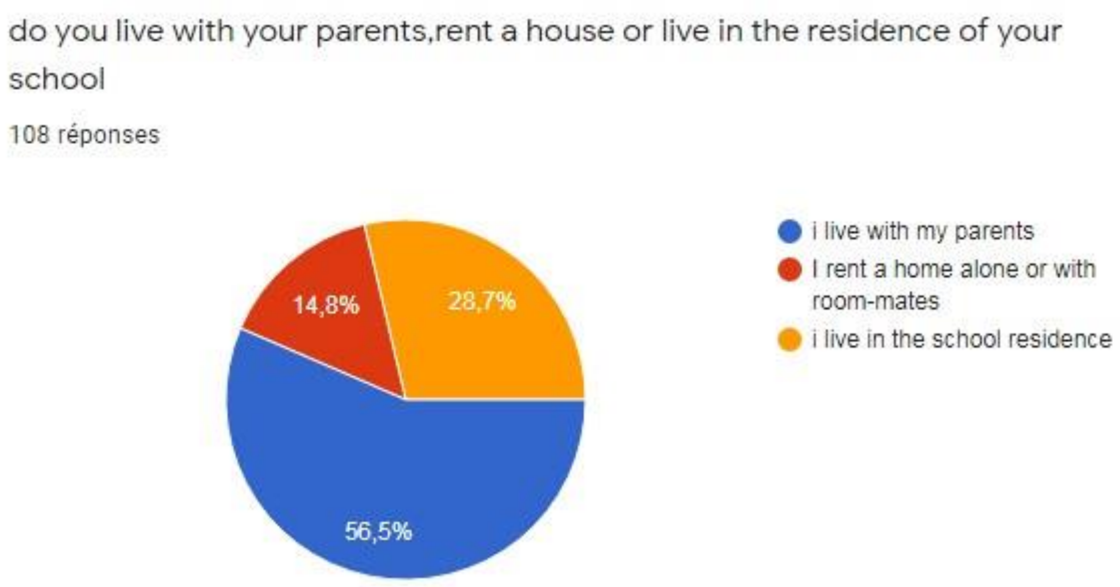
We start by looking at the distribution of the people we surveyed:




what is you current level of studies

108 réponses

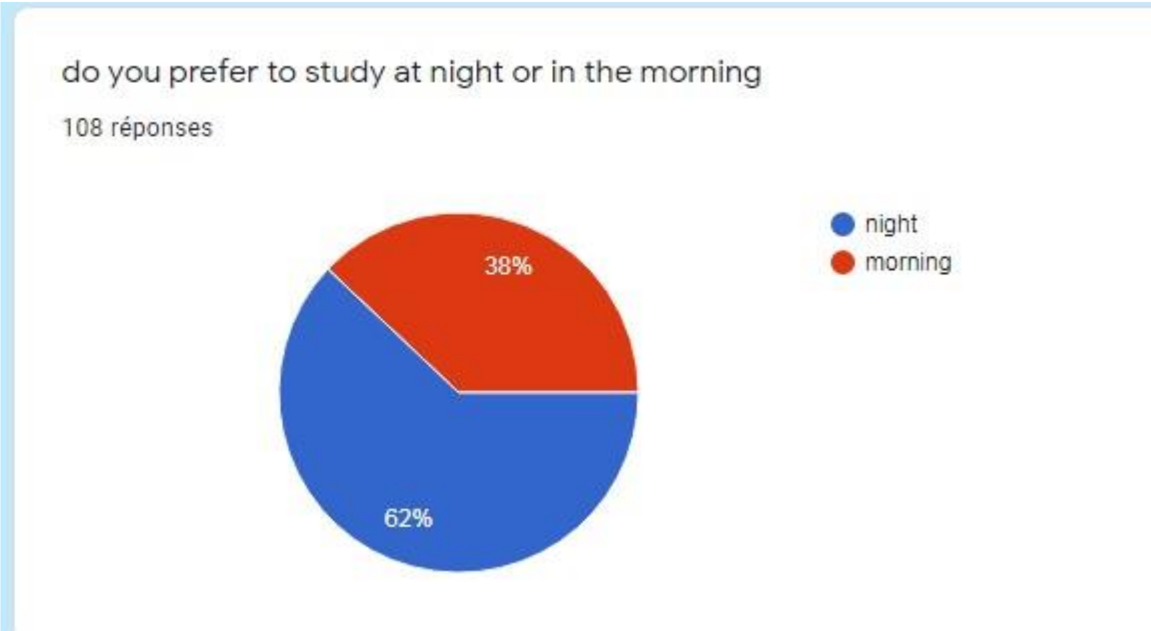
We can see that the majority of people who answered this survey are students in the college of engineering and architecture in their forth year, that’s because those are our classmates.



do you live with your parents,rent a house or live in the residence of your school

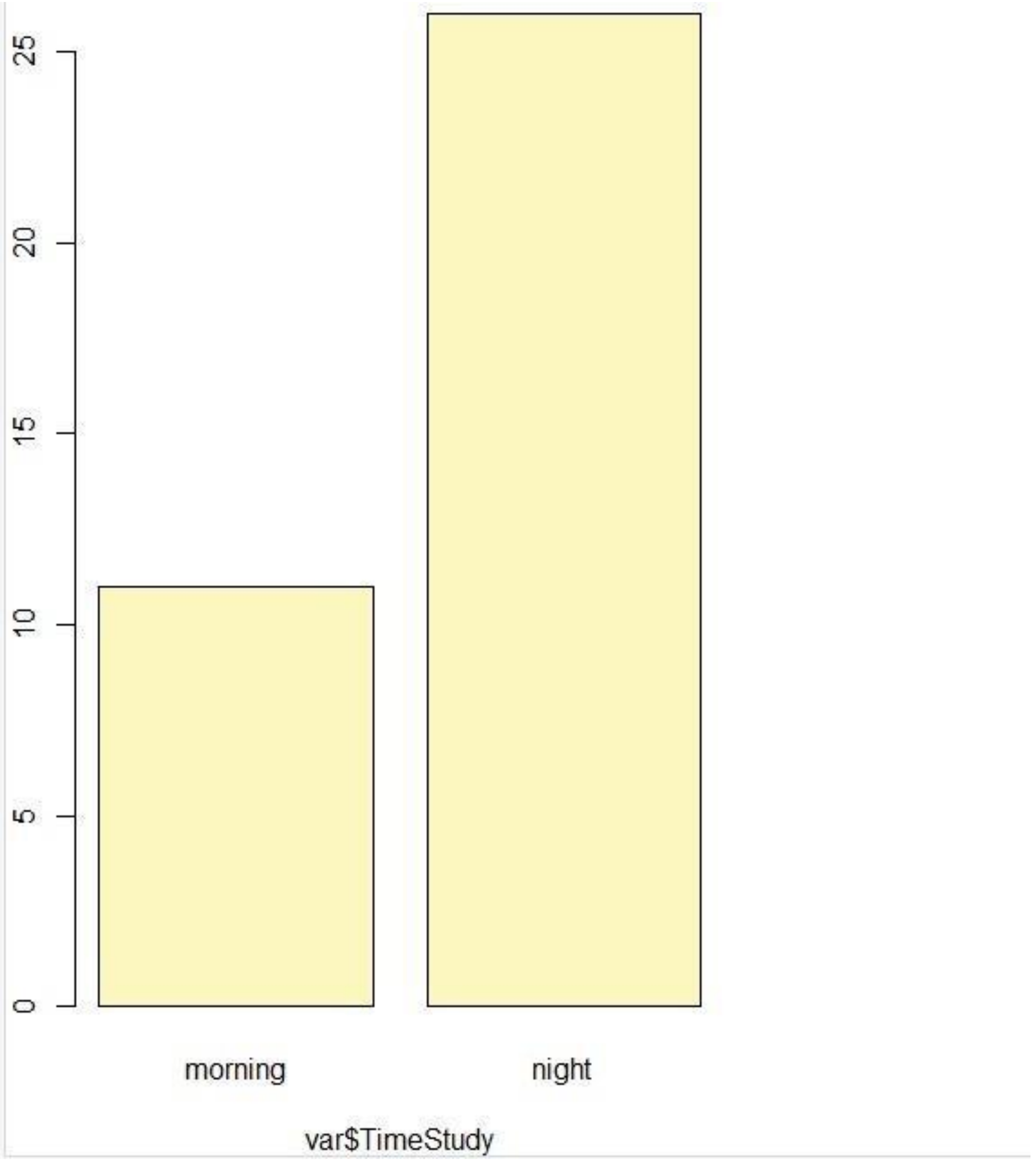
108 réponses

→ More than half of the students live with their parents.



do you prefer to study at night or in the morning

108 réponses

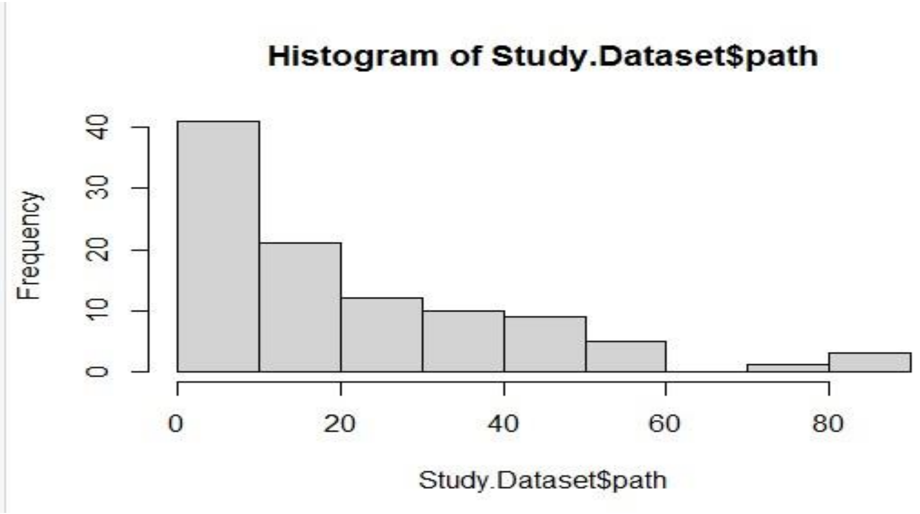


We can see that most students prefer studying at night

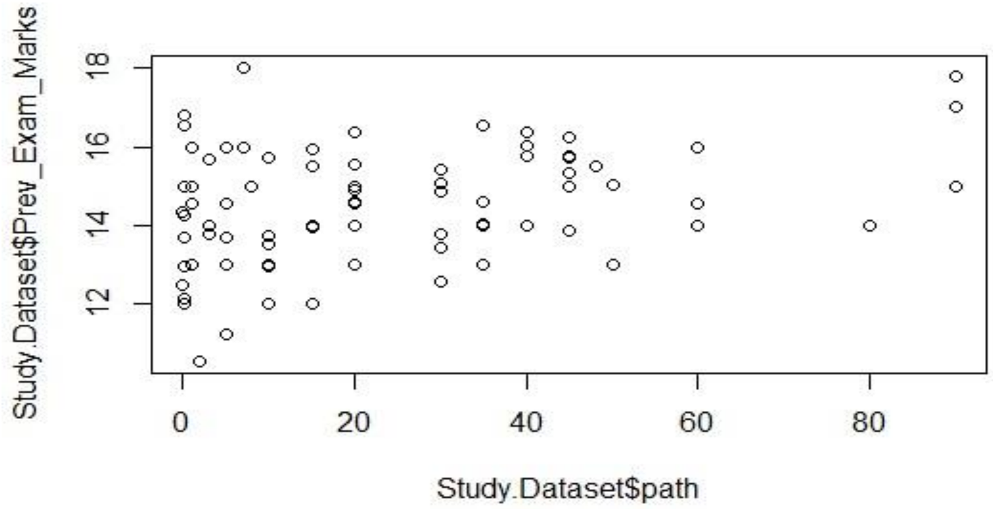
Now that we have a good idea about the population surveyed, we're going to use Rstudio to plot a bunch of graphs to see if there are correlations between our variables:

- First we'll look at the time it takes students to get to school, then we'll see if it has any influence on the final grade.

Here's the distribution of this variable we named "path":

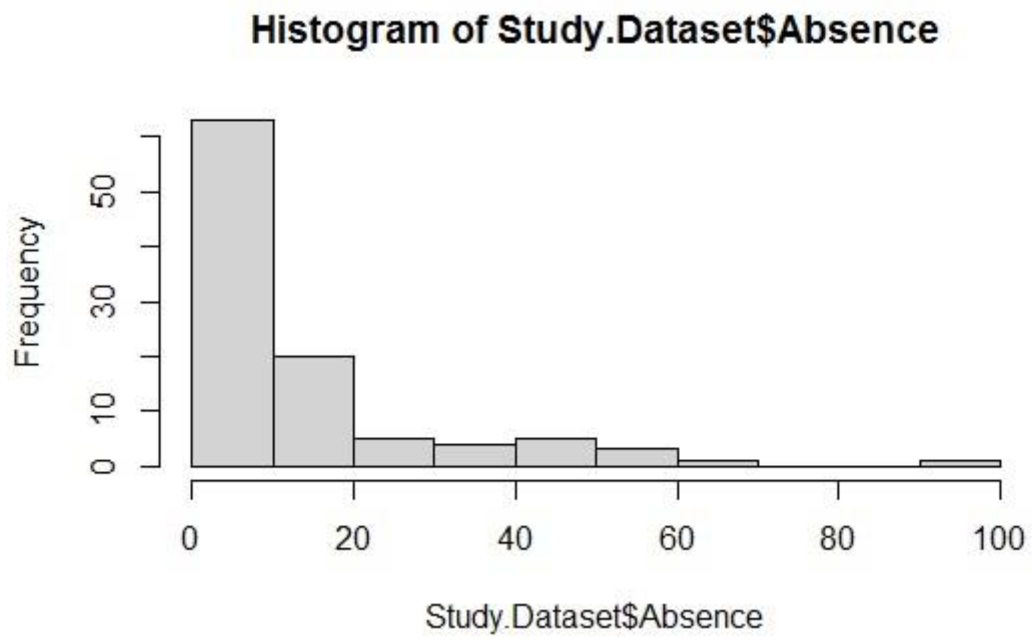


It’s clear that majority of students get to school in 0 to 30 minutes Now we plot the final grade in function of the path:

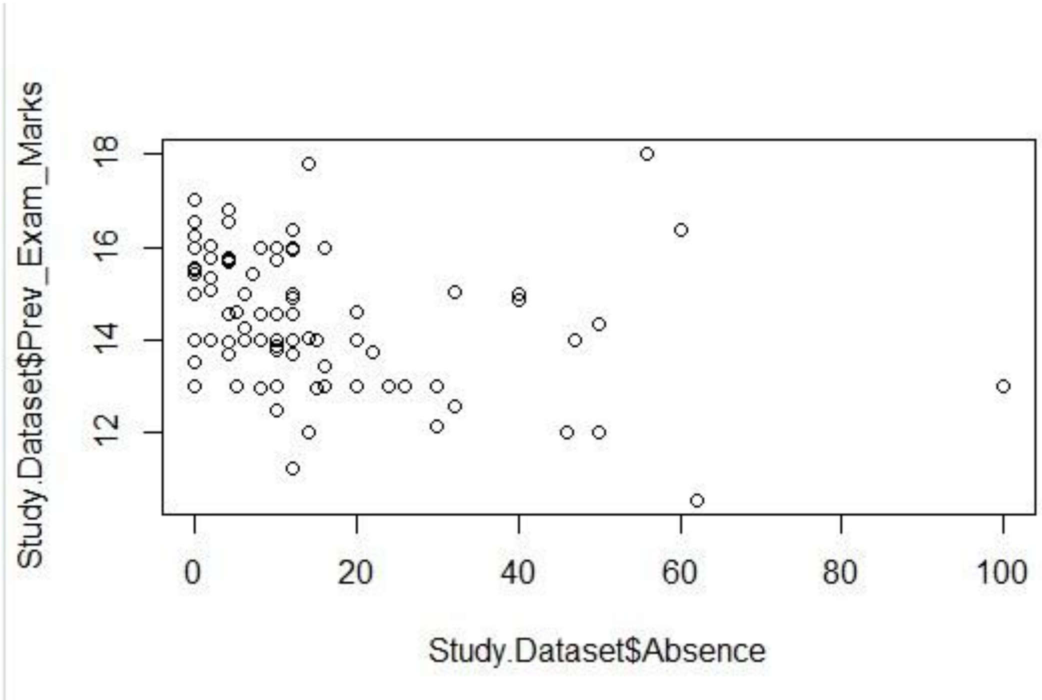


At first glance, there doesn’t seem to be any correlation between grades and the time it takes to get to school.

- Now we do the same thing for number of absences.

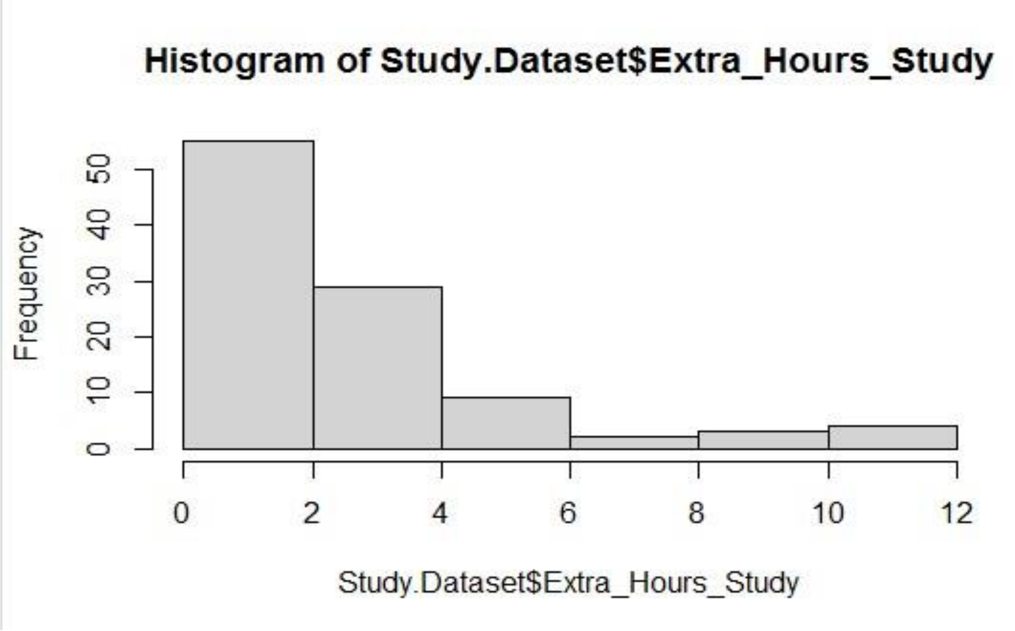


The distribution is right skewed. Most people in this study have between 0 and 20 absences.

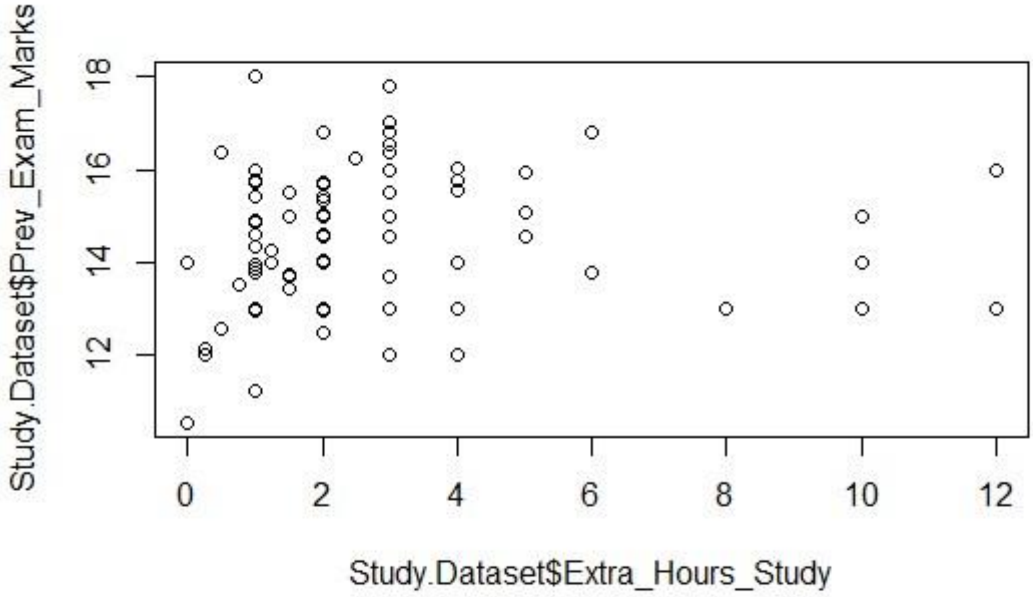


We can see that between 0 and 20 hours of absence there doesn’t seem to be any correlation between absences and grades. But once we cross that threshold we can observe a clear drop-off in grades. This means that too much absence can affect one’s grades.

- Now we look at the number of hours spend studying in a day:

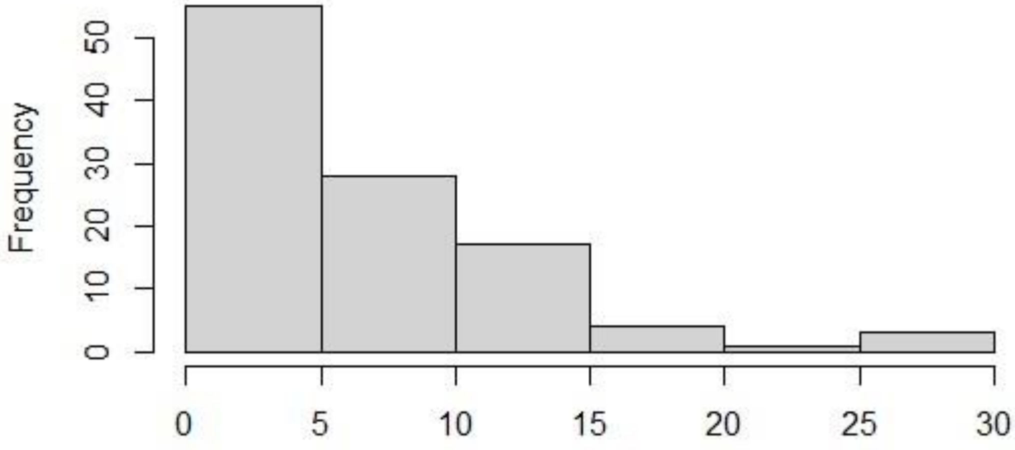


The distribution is right skewed. Most people study 0 to 2 hours a day (17.6% spend 1 hour, and 20% spend 2 hours studying)

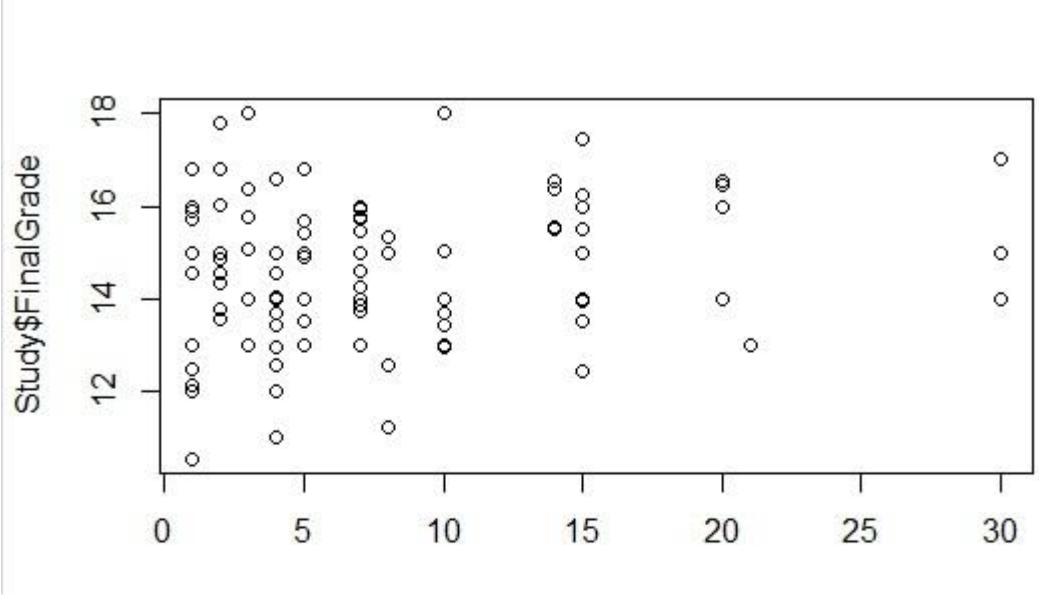


We can see that the general trend here is that the grades tend to get better with the increase of the time spent studying. We can also see that studying past 5 hours doesn't seem to wield tangible results.

- Now we look at the number of days before students start preparing for an exam:



The distribution is right skewed. The majority of people study less than 10 days before an exam.



➡ We can see that the number of days before studying for an exam doesn’t really have an effect on the grades.

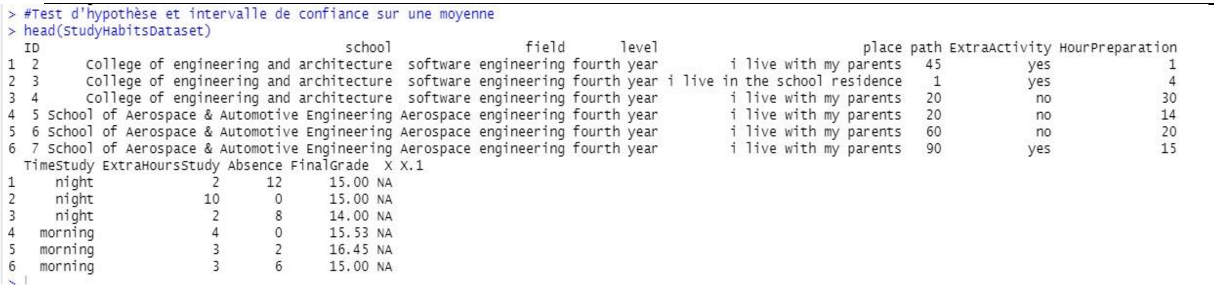
Hypothesis:

H0: There is no relation between grades and study habits

HA: Study habits have real effects on grades

Hypothesis testing

1. Test d’hypothèse et intervalle de confiance sur une moyenne

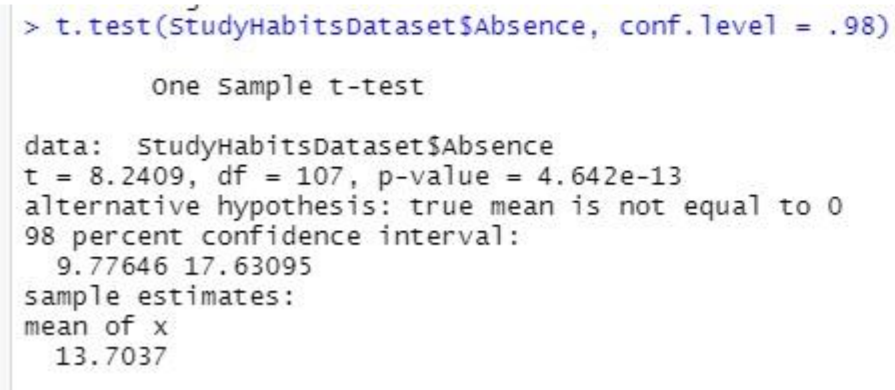


```
> #Test d'hypothèse et intervalle de confiance sur une moyenne
> head(StudyHabitsDataset)
  ID                                       school                field       level                          place path ExtraActivity HourPreparation
1  2      College of engineering and architecture  software engineering fourth year         i live with my parents   45           yes               1
2  3      College of engineering and architecture  software engineering fourth year i live in the school residence    1           yes               4
3  4      College of engineering and architecture  software engineering fourth year         i live with my parents   20            no              30
4  5 School of Aerospace & Automotive Engineering Aerospace engineering fourth year         i live with my parents   20            no              14
5  6 School of Aerospace & Automotive Engineering Aerospace engineering fourth year         i live with my parents   60            no              20
6  7 School of Aerospace & Automotive Engineering Aerospace engineering fourth year         i live with my parents   90           yes              15
  TimeStudy ExtraHoursStudy Absence FinalGrade  X X.1
1     night               2      12      15.00 NA
2     night              10       0      15.00 NA
3     night               2       8      14.00 NA
4   morning               4       0      15.53 NA
5   morning               3       2      16.45 NA
6   morning               3       6      15.00 NA
```

We see that the data set is 108 observations of final grade and average of absence. There is also school, field, level information attached, that we will ignore for the purposes of this example.

Let us find a 98-confidence interval for the mean Absence of students. For this to be valid, we are assuming that we have a *random sample* from all students and that the average of absence of students is normally distributed.

The R command that finds a confidence interval for the mean in this way is:

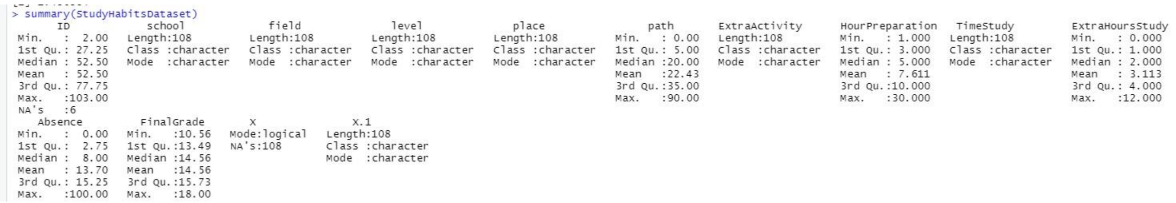


```
> t.test(StudyHabitsDataset$Absence, conf.level = .98)

        One Sample t-test

data:  StudyHabitsDataset$Absence
t = 8.2409, df = 107, p-value = 4.642e-13
alternative hypothesis: true mean is not equal to 0
98 percent confidence interval:
  9.77646 17.63095
sample estimates:
mean of x
  13.7037
```

- ✓ We get a lot of information, but we can pull out what we are looking for as the confidence interval [9.78, 17.64]. So, we are 98 confident that the mean Absence of students is greater than 9.78 and less than 17.64.

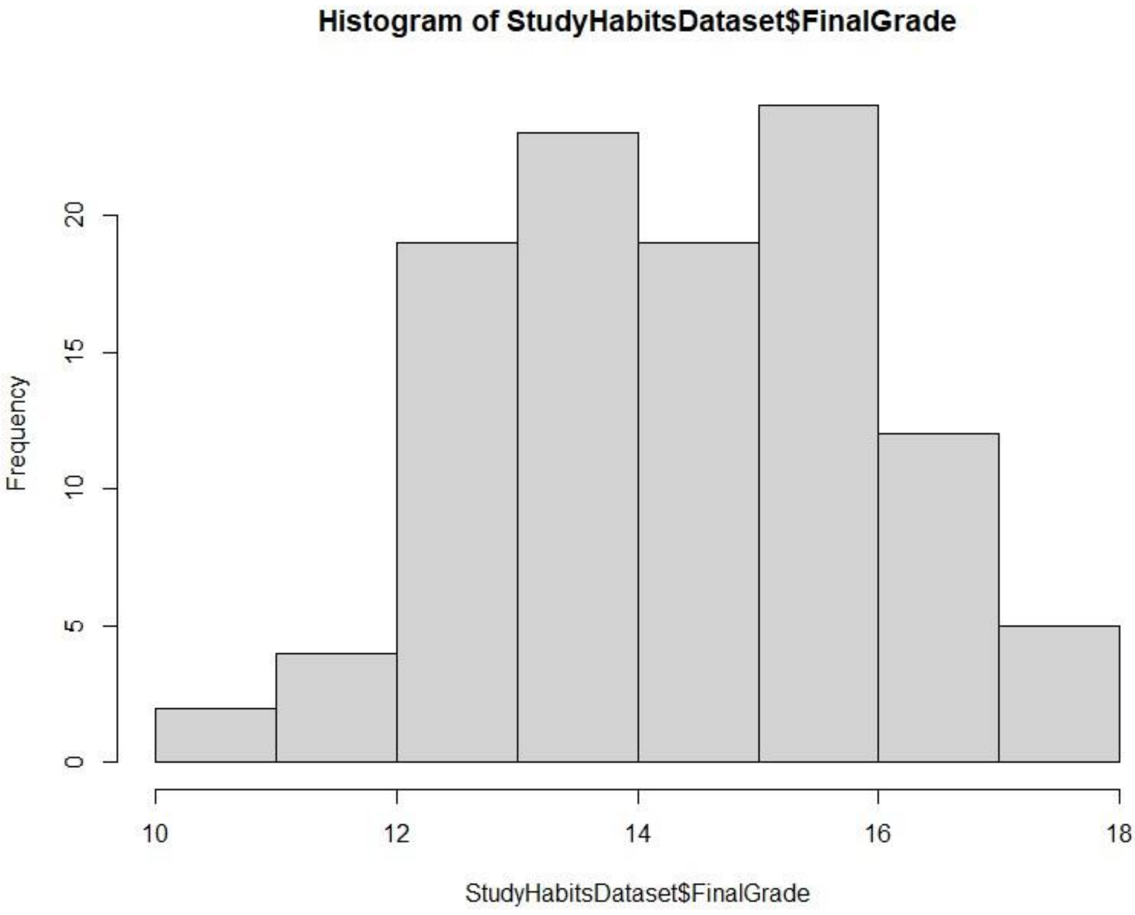
2. Test d’hypothèse et intervalle de confiance sur la différence entre deux moyennes

Summary of our dataSet :

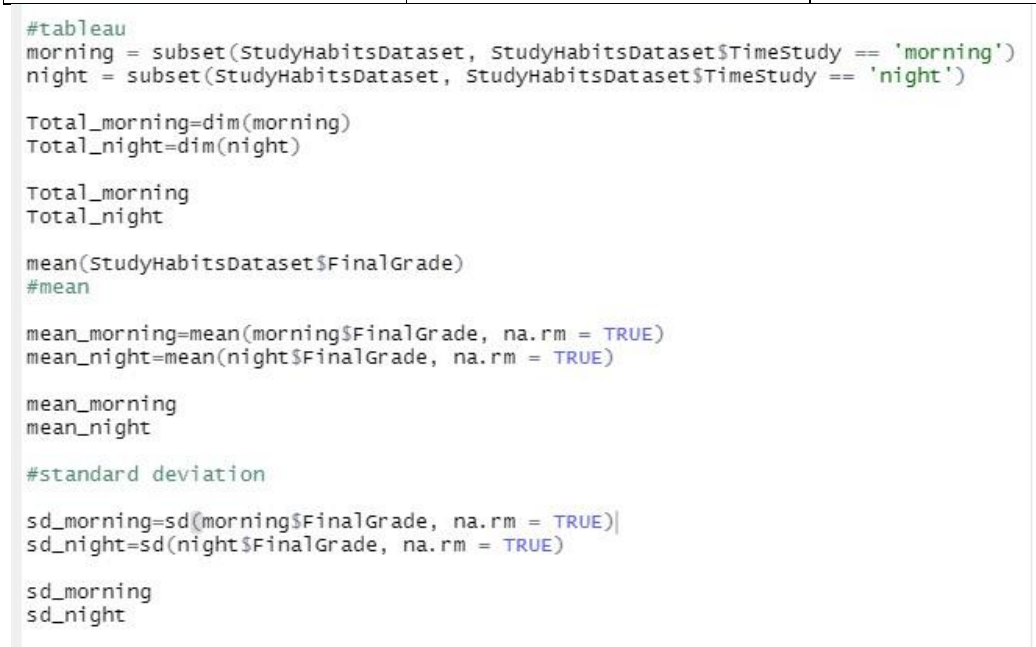


```
> summary(StudyHabitsDataset)
       ID             school              field               level               place                path         ExtraActivity       HourPreparation   TimeStudy          ExtraHoursStudy
 Min.   :  2.00   Length:108         Length:108         Length:108         Length:108         Min.   : 0.00   Length:108         Min.   : 1.000   Length:108         Min.   : 0.000
 1st Qu.: 27.25   Class :character   Class :character   Class :character   Class :character   1st Qu.: 5.00   Class :character   1st Qu.: 3.000   Class :character   1st Qu.: 1.000
 Median : 52.50   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :20.00   Mode  :character   Median : 5.000   Mode  :character   Median : 2.000
 Mean   : 52.50                                                                               Mean   :22.43                      Mean   : 7.611                      Mean   : 3.113
 3rd Qu.: 77.75                                                                               3rd Qu.:35.00                      3rd Qu.:10.000                      3rd Qu.: 4.000
 Max.   :103.00                                                                               Max.   :90.00                      Max.   :30.000                      Max.   :12.000
 NA's   :6
    Absence          FinalGrade        X                 X.1
 Min.   :  0.00   Min.   :10.56   Mode:logical   Length:108
 1st Qu.:  2.75   1st Qu.:13.49   NA's:108       Class :character
 Median :  8.00   Median :14.56                  Mode  :character
 Mean   : 13.70   Mean   :14.56
 3rd Qu.: 15.25   3rd Qu.:15.73
 Max.   :100.00   Max.   :18.00
```

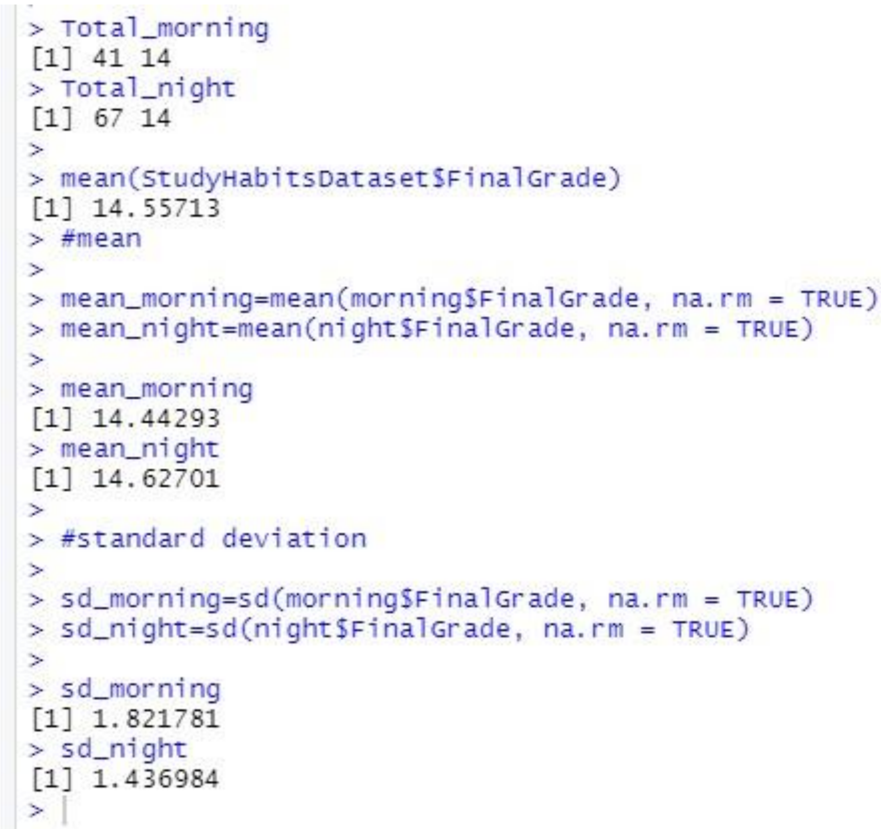
distribution of variable FinalGrade:



	morning	**night**
Mean_Final_grade	14.44293	14.62701
SD_Finale_grade	1.821781	1.436984
n	41	67



```
#tableau
morning = subset(StudyHabitsDataset, StudyHabitsDataset$TimeStudy == 'morning')
night = subset(StudyHabitsDataset, StudyHabitsDataset$TimeStudy == 'night')

Total_morning=dim(morning)
Total_night=dim(night)

Total_morning
Total_night

mean(StudyHabitsDataset$FinalGrade)
#mean

mean_morning=mean(morning$FinalGrade, na.rm = TRUE)
mean_night=mean(night$FinalGrade, na.rm = TRUE)

mean_morning
mean_night

#standard deviation

sd_morning=sd(morning$FinalGrade, na.rm = TRUE)
sd_night=sd(night$FinalGrade, na.rm = TRUE)

sd_morning
sd_night
```



```
> Total_morning
[1] 41 14
> Total_night
[1] 67 14
>
> mean(StudyHabitsDataset$FinalGrade)
[1] 14.55713
> #mean
>
> mean_morning=mean(morning$FinalGrade, na.rm = TRUE)
> mean_night=mean(night$FinalGrade, na.rm = TRUE)
>
> mean_morning
[1] 14.44293
> mean_night
[1] 14.62701
>
> #standard deviation
>
> sd_morning=sd(morning$FinalGrade, na.rm = TRUE)
> sd_night=sd(night$FinalGrade, na.rm = TRUE)
>
> sd_morning
[1] 1.821781
> sd_night
[1] 1.436984
>
```

Parameter of interest: mean of the difference between final grades of students who study at night and students who prefer study at the morning

µm-µn

Point estimate: mean of the difference between final grades of students who study at night and students who prefer study at the morning of our sample

Xm-Xn

Hypotheses:

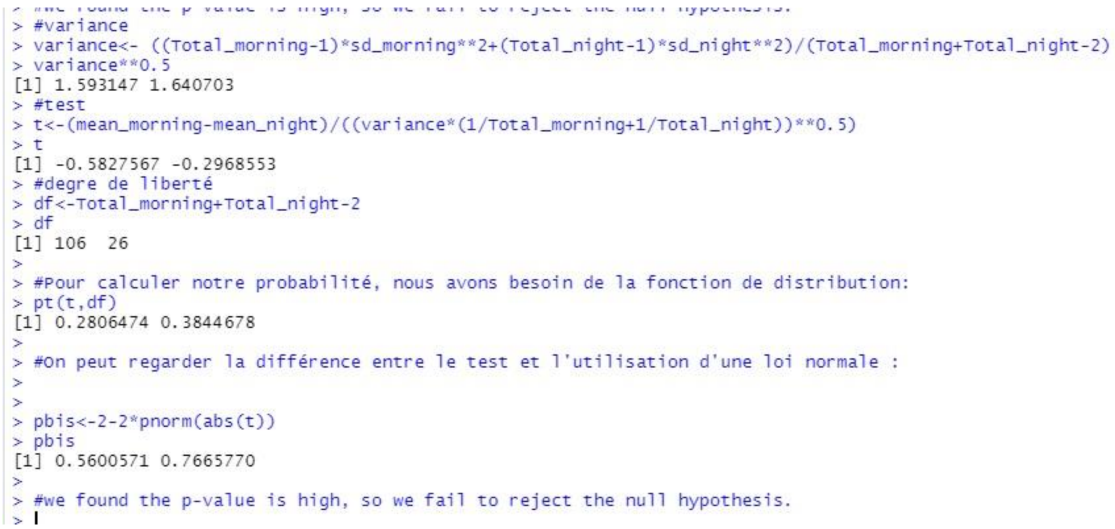
H0: µm = µn

HA: µm < µn

Conditions:

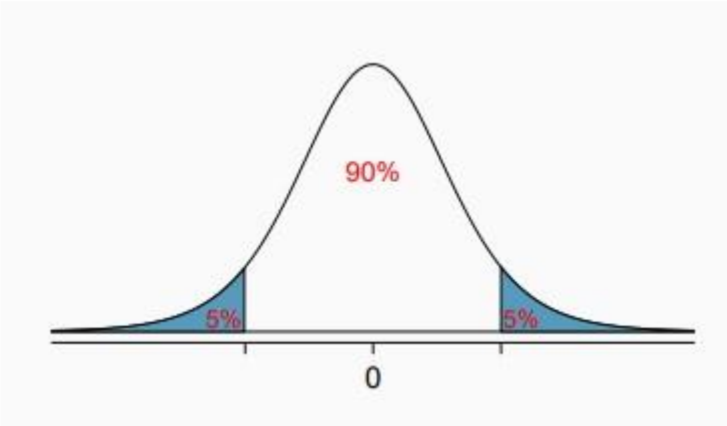
- Independence between those groups
- Both sample sizes should be at least 30

Test statistic :

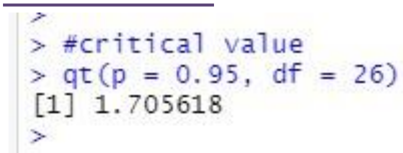


```
> #we found the p-value is high, so we fail to reject the null hypothesis.
> #variance
> variance<- ((Total_morning-1)*sd_morning**2+(Total_night-1)*sd_night**2)/(Total_morning+Total_night-2)
> variance**0.5
[1] 1.593147 1.640703
> #test
> t<-(mean_morning-mean_night)/((variance*(1/Total_morning+1/Total_night))**0.5)
> t
[1] -0.5827567 -0.2968553
> #degre de liberté
> df<-Total_morning+Total_night-2
> df
[1] 106  26
>
> #Pour calculer notre probabilité, nous avons besoin de la fonction de distribution:
> pt(t,df)
[1] 0.2806474 0.3844678
>
> #On peut regarder la différence entre le test et l'utilisation d'une loi normale :
>
>
> pbis<-2-2*pnorm(abs(t))
> pbis
[1] 0.5600571 0.7665770
>
> #we found the p-value is high, so we fail to reject the null hypothesis.
> |
```

What is the equivalent confidence level for a one-sided hypothesis test at α = 0.05? **90%**



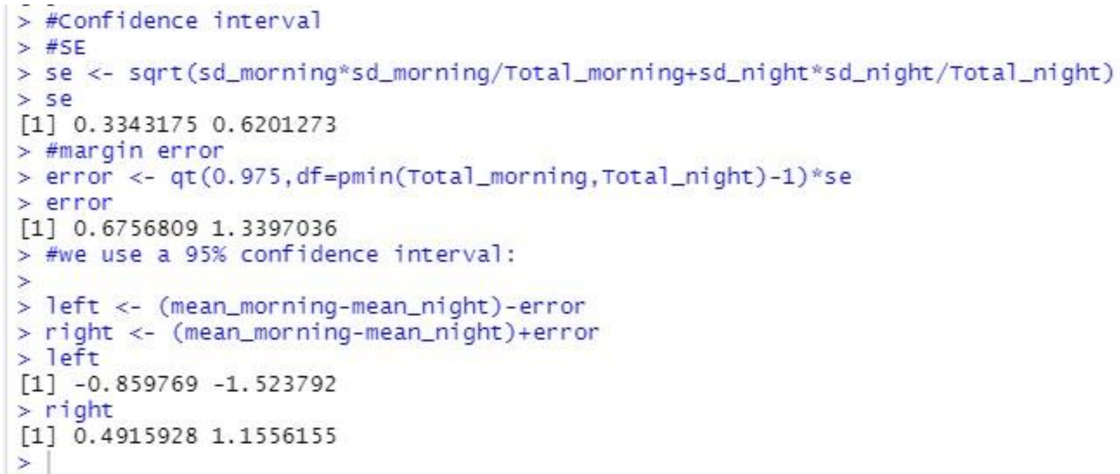
Critical value:



```
>
> #critical value
> qt(p = 0.95, df = 26)
[1] 1.705618
>
```

Confidence interval:

point estimate ± ME



```
> #Confidence interval
> #SE
> se <- sqrt(sd_morning*sd_morning/Total_morning+sd_night*sd_night/Total_night)
> se
[1] 0.3343175 0.6201273
> #margin error
> error <- qt(0.975,df=pmin(Total_morning,Total_night)-1)*se
> error
[1] 0.6756809 1.3397036
> #we use a 95% confidence interval:
>
> left <- (mean_morning-mean_night)-error
> right <- (mean_morning-mean_night)+error
> left
[1] -0.859769 -1.523792
> right
[1] 0.4915928 1.1556155
> |
```

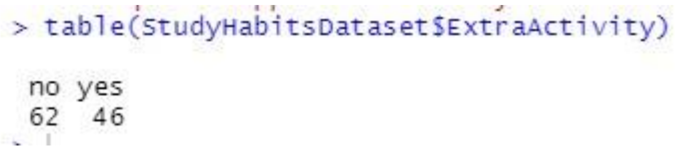
- ✓ This gives the confidence intervals for each of the three tests. For example, in the first experiment the 95% confidence interval is between -0.86 and 0.5 assuming that the random variables are normally distributed, and the samples are independent.

3. Test d’hypothèse et intervalle de confiance sur une proportion

We conduct this test on the proportion of people who take part in extracurricular activities. We assume that the null probability is 0.5.

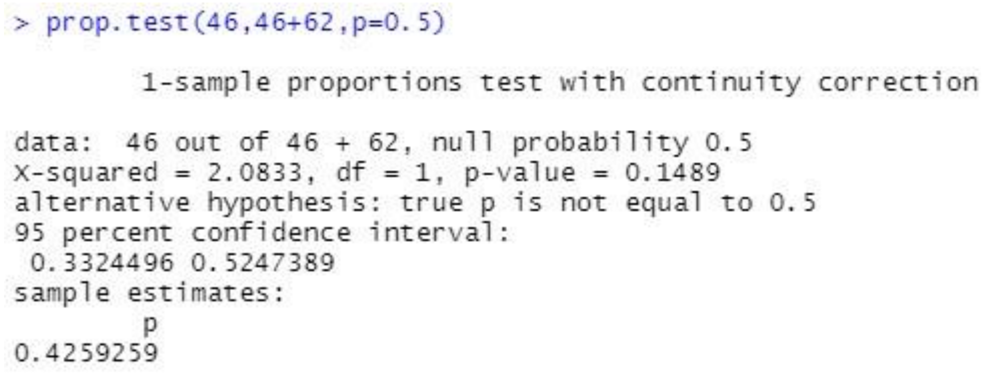
H0: p equal to 0.5

HA: p not equal to 0.5



```
> table(StudyHabitsDataset$ExtraActivity)

 no yes
 62  46
```

We consider the success to be “yes”



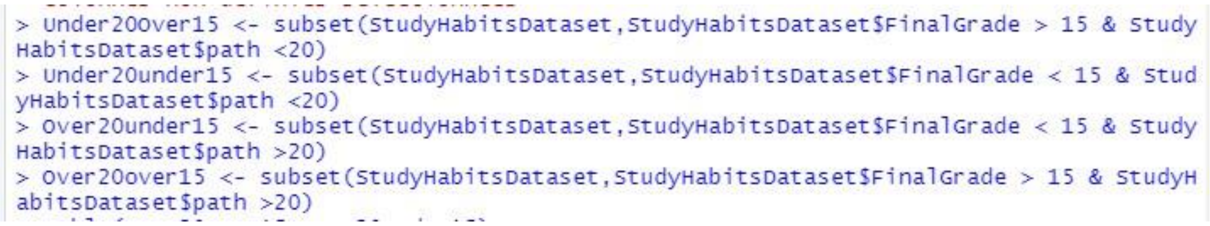
```
> prop.test(46,46+62,p=0.5)

	1-sample proportions test with continuity correction

data:  46 out of 46 + 62, null probability 0.5
X-squared = 2.0833, df = 1, p-value = 0.1489
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3324496 0.5247389
sample estimates:
        p
0.4259259
```

The p-value is greater than 0.05, so we fail to reject the null hypothesis. The confidence interval is : (0.3324496 , 0.5247389)

4. Test d’hypothèse et intervalle de confiance sur deux proportions

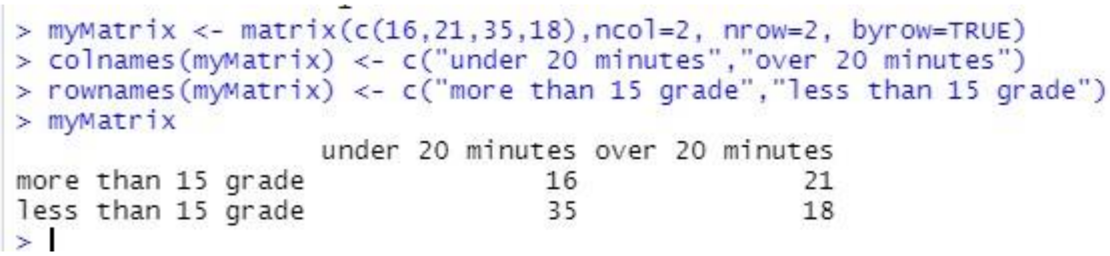
We conduct this test to compare two proportions: the proportion of student who got more than 15 on their final grade, who have more than 20 minutes to get to school vs less than 20 minutes to get to school.

First, we create subsets of based of whether or not the student has more than 15 on his grade, and the time it takes to get to school:



```
> Under20Over15 <- subset(StudyHabitsDataset,StudyHabitsDataset$FinalGrade > 15 & Study
HabitsDataset$path <20)
> Under20under15 <- subset(StudyHabitsDataset,StudyHabitsDataset$FinalGrade < 15 & Stud
yHabitsDataset$path <20)
> Over20under15 <- subset(StudyHabitsDataset,StudyHabitsDataset$FinalGrade < 15 & Study
HabitsDataset$path >20)
> Over20over15 <- subset(StudyHabitsDataset,StudyHabitsDataset$FinalGrade > 15 & StudyH
abitsDataset$path >20)
```

Now, we can create a matrix to clearly see the proportions:



```
> myMatrix <- matrix(c(16,21,35,18),ncol=2, nrow=2, byrow=TRUE)
> colnames(myMatrix) <- c("under 20 minutes","over 20 minutes")
> rownames(myMatrix) <- c("more than 15 grade","less than 15 grade")
> myMatrix
                   under 20 minutes over 20 minutes
more than 15 grade               16              21
less than 15 grade               35              18
> |
```

We can conduct the hypothesis test, and we assume that success is when a student gets more than 15 on his/her final grade.

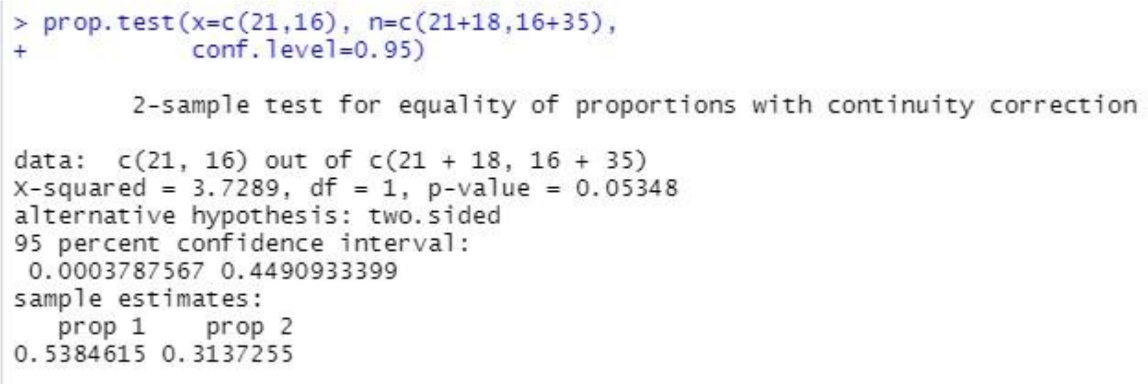
P1: more than 15, and takes more than 20 minutes to get to school

P2: more than 15, and takes less than 20 minutes to get to school

We start by a two-sided test:

H0 : p1=p2

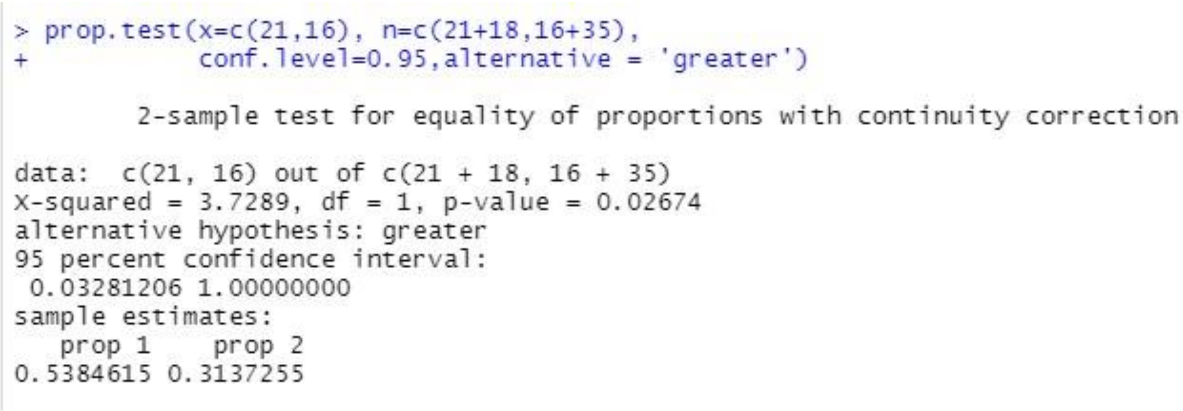
HA : p1 != p2



```
> prop.test(x=c(21,16), n=c(21+18,16+35),
+           conf.level=0.95)

        2-sample test for equality of proportions with continuity correction

data:  c(21, 16) out of c(21 + 18, 16 + 35)
X-squared = 3.7289, df = 1, p-value = 0.05348
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0003787567 0.4490933399
sample estimates:
   prop 1    prop 2 
0.5384615 0.3137255 
```

The p-value is fairly small, which means we can reject the null hypothesis.

Now we conduct a one sided test to see which proportion is likely to be greater:



```
> prop.test(x=c(21,16), n=c(21+18,16+35),
+           conf.level=0.95,alternative = 'greater')

        2-sample test for equality of proportions with continuity correction

data:  c(21, 16) out of c(21 + 18, 16 + 35)
X-squared = 3.7289, df = 1, p-value = 0.02674
alternative hypothesis: greater
95 percent confidence interval:
 0.03281206 1.00000000
sample estimates:
   prop 1    prop 2 
0.5384615 0.3137255 
```

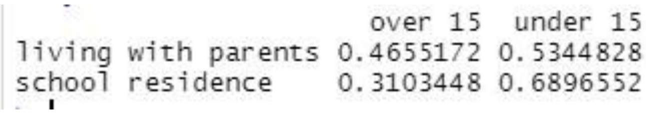
We have a p-value of 0.02674 < 0.05 so we can reject the null hypothesis.

This indicates that we have a high level of confidence that p1 will be greater than p2, which means that student who take more time to get to school tend to get better grades.

5. Test Chi-2 :

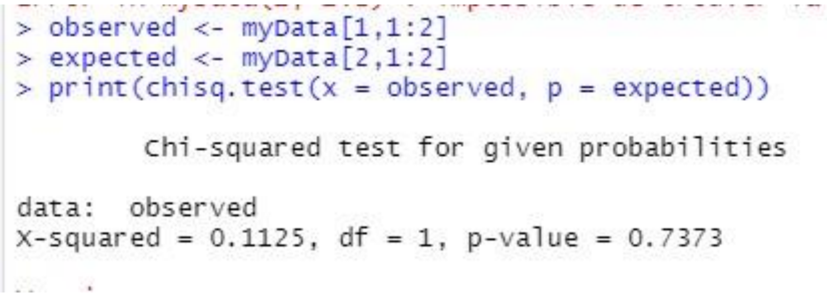
We want to know if living in the school residence has any effect on one's grades. For this, we use the chi-2 test, to see if the proportion of students that got more than 15 on their final grade changes depending where the student lives.

First we create a matrix with the needed values:



```
                      over 15  under 15
living with parents 0.4655172 0.5344828
school residence    0.3103448 0.6896552
```

Then we conduct the chi-2 test:



```
> observed <- myData[1,1:2]
> expected <- myData[2,1:2]
> print(chisq.test(x = observed, p = expected))

        Chi-squared test for given probabilities

data:  observed
X-squared = 0.1125, df = 1, p-value = 0.7373
```

The p-value of 0.7373 indicates that there is a 73% chance the the null hypothesis is correct. This fairly high value means that we fail to reject the null hypothesis. In this case, that means that cannot have a high level confidence that living in the school residence has an effect on the final grade.

Conclusion

The initial analysis of our data gave us some pretty encouraging results, like the fact that grades tend to get better with the increase of the time spent studying, or that absences appear to have negative effects on grade once the student gets past a certain threshold. However, once we started conducting hypothesis tests to verify our assumptions about the data, we were pretty disappointed. Almost all tests indicated that study habits don’t have tangible effects on grades, because we got faily high p-values. The only conclusive test was the test on two proportions, which showed us that students who take more time to get to school tend to get better grades, and that was confirmed by the chi-square test.