Statistics Project - The correlation between academic success and sport

Anna Bennaim & Chloe Marangon

2023-11-14

Is there a correlation between the practice of a sport and the grades of a student ?

« Nous déclarons sur l'honneur que ce mémoire a été écrit de notre main, sans aide extérieure non autorisée, qu'il n'a pas été présenté auparavant pour évaluation et qu'il n'a jamais été publié dans sa totalité ou en partie. Toutes parties, groupes de mots ou idées, aussi limités soient-ils, y compris des tableaux, graphiques, cartes etc. qui sont empruntés ou qui font référence à d'autres sources bibliographiques sont présentés comme tels, sans exception aucune. »

Contents

A - Introduction

Since the first day in Dauphine, all students are taught that sport is crucial in well-being as well as for academic success. Indeed, the practice of a sport offers a large number of benefits, for example, the improvement of lung and heart function as well as reduce high blood pressure. It is also, medically speaking, beneficial for motor and exercise skills. However, due to lack of time, resources or bad organization, we know that it's often difficult to practice sport regularly.

We have decided to create a study to prove that sport is very important to get good grades, especially for Dauphine's students who need to work hard. Therefore, our project is to verify that the practice of a sport will enhance the capabilities of a student to have better grades.

Our database is extract from a form that we made and publish on the Facebook group of Dauphine's students. Hence, the population is composed of 102 students. Our sample is fairly representative of the population, with 57 women and 45 men. Our sample is based students, all Dauphine students without exception,from first year to master's level without distinction. We don't distinguish between students' different fields of study.

B - Variables description

We are going to analyse three types of data:

- 4 categorical variables: the gender of the student, if they subscribed to PSL Sport, if the student think that sport is accessible in Dauphine and if the practice of a sport influence grades.
- 1 discrete quantitative variable: How many times a week does the student practice a sport?
- 2 continuous quantitatives variables: the grades of the student and the number of hours he use to work per week.

For each variables, we are going to generate a frequency table and give for each a graph of the empirical distribution.

```
rm(list=ls())
library(readxl)
library(ggplot2)
options(digits=2)
```

1) Description of the variable *Gender*

As said earlier, this variable is a categorical value, therefore the graph used for this variable is a pie. This variable allows us to know which type of gender are the students who responded to the form. There are only two options: student can be a man or a woman.

Frequency table associated

```
Gender<-Baseproject$Gender
Total<-sum
gend<-prop.table(table(Gender))
addmargins(prop.table(gend),FUN=Total)
```

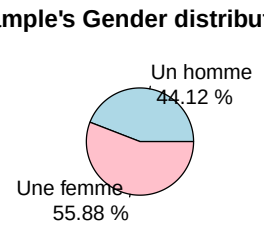
```
## Gender
##  Un homme Une femme     Total
##      0.44      0.56      1.00
```

Graph of the empirical distribution of the "Gender" variable.

```
etiquette<-paste(rownames(gend),'\n',round(gend*100,2),'%')
couleur<-c('lightblue','pink')
addmargins(prop.table(gend),FUN=Total)
```

```
## Gender
##  Un homme Une femme     Total
##      0.44      0.56      1.00
```

```
pie(table(Gender),main="Sample's Gender distribution",col=couleur,labels=etiquette)
```



The graph shows that the majority of survey respondents are women. In fact, 55.88% of our survey respondents are women, compared with 44.12% who are men. However, the number of women is close to the number of men in our sample. This is a good thing to represent fairly the population.

2) Description of the variable *PSLsport*

This variable allows us to know if the students who responded to the form are PSL sport subscribers. There are only two possible answers: Yes or No. This variable is also a categorical variable, therefore the graph used is also a pie.

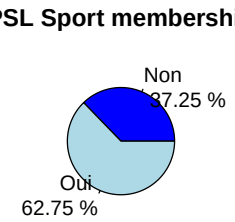
Frequency table associated

```
Total<-sum
adhesion<-prop.table(table(PSLsport))
etiquette<-paste(rownames(adhesion),'\n',round(adhesion*100,2),'%')
couleur<-c('blue','lightblue')
addmargins(prop.table(adhesion),FUN=Total)
```

```
## PSLsport
##   Non   Oui Total
##  0.37  0.63  1.00
```

Graph of the empirical distribution of the "PSLsport" variable.

```r
#graph pSlSport
etiquette<-paste(rownames(adhesion),'\n',round(adhesion*100,2),'%')
couleur<-c('blue','lightblue')
pie(table(PSLsport),main="PSL Sport membership",col=couleur,labels=etiquette)
```



The graph shows that the majority of respondents in this population have subscribed to Psl sport.

3) Description of the variable *Frequencysport*

The variable "Frequencysport" permits us to know how many times per week the students who responded to the form, practices a sport. There are five answers possible: 0 (no sport at all), 1 (1 time during the week), 2 (2 times during the week), 3 (3 times during the week) and 4 (4 times or more during the week).
We then assessed the frequency of each response in our population. This frequency is measured between 0 and 1.
As this variable is a discrete quantitative variable, therefore the type of graph used is a plot.

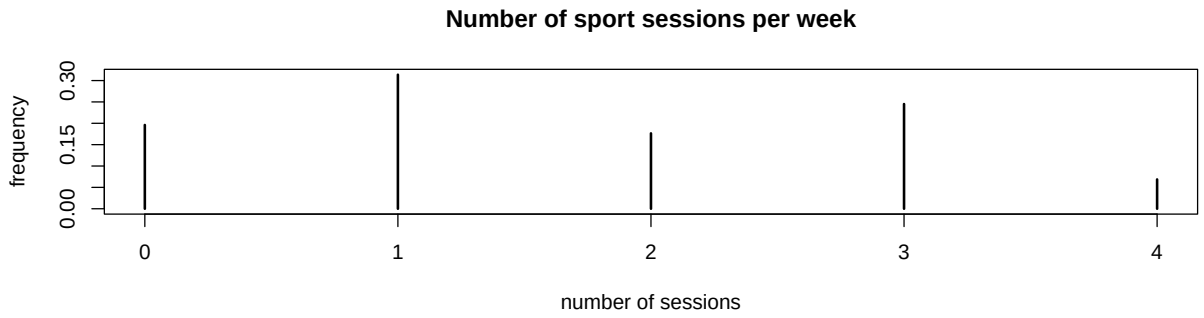
Frequency table associated

```r
Frequencysport<-Baseproject$Frequencysport
Totalsport<-sum
f<-prop.table(table(Frequencysport))
addmargins(prop.table(f),FUN=Totalsport)
```

```
## Frequencysport
##          0          1          2          3          4 Totalsport
##      0.196      0.314      0.176      0.245      0.069      1.000
```

Graph of the empirical distribution of "Frequencysport" variable.

```r
plot(prop.table(table(Frequencysport)),main="Number of sport sessions per week",
     xlab="number of sessions",ylab="frequency")
```



This graph shows that the largest proportion of students take part in sports on average once a week (32%). In second place, we see that almost 25% of students do three sports sessions per week. We can also see that almost 20% of students don't do any sport at all.

4) Description of the variable *Hourswork*

The variable "Hourswork" let us know how many hours of work, the students who responded to the form, does in average per week. There are 4 answers possible: 1 hour or less, 2 hours to 5 hours, 5 hours to 10 hours or 10 hours or more.
As this variable is a continuous quantitative variable, therefore the graph used for this type of variable is an histogram. Here we have use the function barplot to get a more detailed graph.

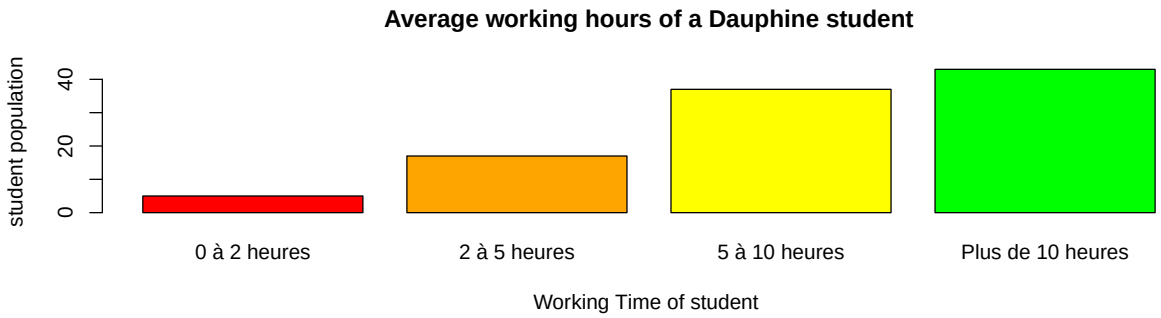
Frequency table associated

```
Hourswork<-Baseproject$Hourswork
tab<-table(Hourswork)
tab
```

```
## Hourswork
##      0 à 2 heures      2 à 5 heures     5 à 10 heures Plus de 10 heures
##                 5                17                37                43
```

Graph of the empirical distribution of "Hourswork" variable.

```
barplot(tab,main="Average working hours of a Dauphine student",xlab="Working Time of student "
        ,ylab="student population",col=c("red","orange","yellow","green"))
```



We can observe that most students work ten hours or more per week.

5) Description of the variable *Accessport*

The variable "Accessport" allows us to know how students evaluate the accessibility of sports at Dauphine. There are 5 answers possible: not accessible at all, not very accessible, moderately accessible, accessible and very accessible.
This variable is a categorical variable so, as said as earlier, we are going to use a pie graph to represent its empirical distribution.

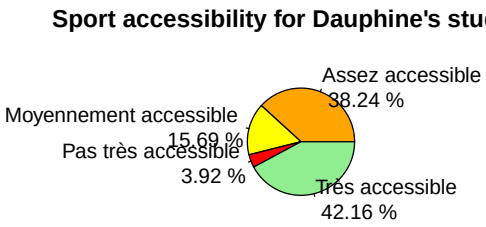
Frequency table associated

```
Total<-sum
accessibilitesport<-prop.table(table(Accessport))
addmargins(prop.table(accessibilitesport),FUN=Total)
```

```
## Accessport
##       Assez accessible Moyennement accessible    Pas très accessible
##                  0.382                  0.157                  0.039
##        Très accessible                  Total
##                  0.422                  1.000
```

Graph of the empirical distribution of "Accessport" variable.

```
etiquette<-paste(rownames(accessibilitesport),'\n',round(accessibilitesport*100,2), '%')
couleur<-c('orange','yellow','red','lightgreen')
pie(table(Accessport),main="Sport accessibility for Dauphine's students",
    col=couleur,labels=etiquette)
```



From our graph, we can see that the majority of people describe sport at Dauphine as very accessible (42.16%), and only 3.92% describe it as not very accessible. On the whole, this shows that sports facilities at Dauphine are quite popular with students.

6) Description of the variable *Grades*

The "Grades" variable allows us to collect the average obtained by each student surveyed. To get a more general idea of the grades obtained, we collected the student's average grade one year ago, all subjects included. There are 5 answers possible: less than 4, between 4 and 8, between 8 and 12, between 12 and 16 and more than 16. Scoring is of course based on 20 points.

This variable is the second continuous quantitative variable of this project. We therefore use a histogram. To facilitate the creation of the histogram, we have also used the barplot function.

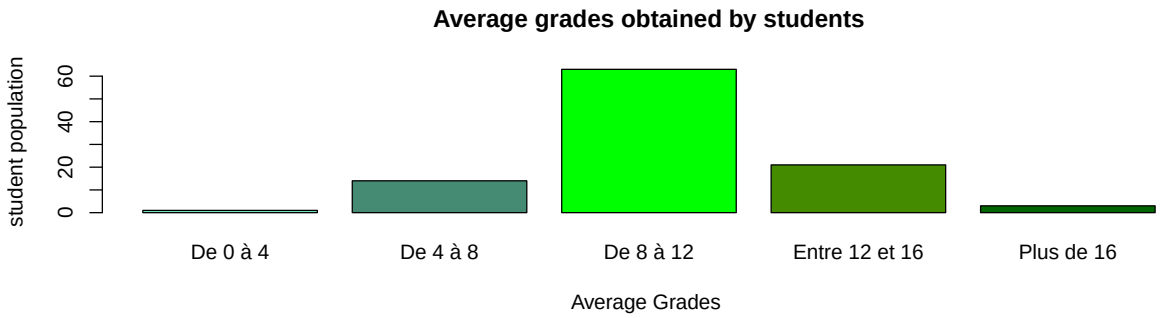
Frequency table associated

```
Grades<-Baseproject$Grades
tab<-table(Grades)
tab
```

```
## Grades
##       De 0 à 4       De 4 à 8      De 8 à 12 Entre 12 et 16     Plus de 16
##              1             14             63             21              3
```

Graph of the empirical distribution of "Grades" variable

```
barplot(tab,main="Average grades obtained by students",xlab="Average Grades ",
        ylab="student population",
        col=c("aquamarine2","aquamarine4","green","chartreuse4","darkgreen"))
```



Our graph shows that by far the majority of students score between 8 and 12.

7) Description of the variable *Influence*

The last variable we'll be using in this project is the "Influence" variable. This variable let us know if the students who responded to the form believes that there is a correlation between practicing a sport and having good grades. There are two possible answers to this variable: Yes or No. Yes means that the student thinks there is a link between doing sport and having good grades, and No means that the student thinks there is no link. This variable is a categorical variable, so the type of graph used is a pie.

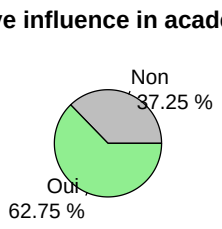
Frequency table associated

```
Influence<-Baseproject$Influence
Total<-sum
influencesport<-prop.table(table(Influence))
addmargins(prop.table(influencesport),FUN=Total)
```

```
## Influence
##   Non   Oui Total
##  0.37  0.63  1.00
```

Graph of the empirical distribution of "Influence" variable

```
etiquette<-paste(rownames(influencesport),'\n',round(influencesport*100,2),'%')
couleur<-c('grey','lightgreen')
pie(table(Influence),main="Do sport have influence in academic results",col=couleur,labels=etiquette)
```



From the graph, we can see that just under 2/3 of the population surveyed believe there is a link between sport and getting good grades. It should be noted that this is subjective data, since it depends solely on personal opinion and experience.

C - Estimation and Confidence Interval

In this section, we have chosen to base our confidence intervals on the proportion. To do this, we have chosen to study the "Pslsport" variable, which represents whether or not students are PSLSport members. This variable follows a Bernoulli distribution (yes or no).

1) Point estimation for the sample Proportion of the variable "Pslsport".

To estimate this variable, we used the Empirical Frequency (F), the estimator of the proportion. The Empirical Frequency is an unbiased and consistent estimator of the proportion. The empirical frequency is calculated by :

$$F(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where Xi=1 if The student i is a PSLSport member and 0 else. With i=0,1,...,102.

With the realization

$$f(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We count the number of "yes" answers in c (on n=102 answers).

```
n<-102
c<-0
v<-Baseproject[,2]
for( i in c(1:102)){
  if (v[i,]=="Oui") {
    c=c+1
    }
}
```

Calculate the point estimation :

```
F<-(1/n)*c
print(F)
```

```
## [1] 0.63
```

We find that F=0.63, which means that 63% of students have subscribed to PSLsport,i.e well over half.

2) Confidence interval estimation at levels 90% for the sample proportion of variable "Pslsport.

We will now make a confidence interval estimate of this "PSLsport" variable with a confidence level of 90%.

We know that our Xi are independent and identically distributed according to a Bernouilli distribution. Moreover, since n=102 is strictly greater than 30, we can use the Central Limit Theorem.

-Pivotal Quantity : From the central limit theorem, we know that F approximately follows a Normal distribution with parameters p and $\sqrt{(p(1-p))/n}$. We get that : $F \sim N(p, \sqrt{(p(1-p))/n})$.

Moreover, we know that: $F \xrightarrow{p} p$. With Slutsky's theorem and by continuity, we get :

$((F-p)/\sqrt{F(1-F)/n}) \sim N(0,1)$.

Our pivotal quantity is also : U=$((F-p)/\sqrt{F(1-F)/n})$.

-Find the quantiles : We are now looking for the quantities of order 5% and 95% in the table of the reduced centered normal distribution such that $P(u_{0,05} < U < u_{0,95}) = 0,90$ With R we find : $u_{0,05} = -1,64$ $u_{0,95} = 1,64$.

```
qnorm(0.95)
```

```
## [1] 1.6
```

```
qnorm(0.05)
```

```
## [1] -1.6
```

-Confidence Interval : The confidence interval for our variable with alpha=10% is of the form :

$$[F - U_{1-\alpha/2} * \sqrt{F(1-F)/n}, F + U_{1-\alpha/2} * \sqrt{F(1-F)/n}]$$

```
alpha<-0.1
IC1<-F-qnorm(0.95)*(sqrt((F*(1-F))/n))
IC2<-F+qnorm(0.95)*(sqrt((F*(1-F))/n))
#l give the length of the interval.
l<-IC2-IC1
print(c(IC1,IC2,l))
```

```
## [1] 0.55 0.71 0.16
```

```
paste(c("A confidence interval of the proportion of pslsport at alpha=90% is ["),
      round(IC1,2),":",round(IC2,2), "]")
```

```
## [1] "A confidence interval of the proportion of pslsport at alpha=90% is [ 0.55 : 0.71 ]"
```

Applying this formula numerically, we obtain IC1, the lower bound of the interval, and IC2, the upper bound of the interval.

3) Confidence interval estimation at levels 95% for the sample proportion of variable "Pslsport"

Repeating the previous reasoning, this time with a confidence level of 95%.
We obtain that : $\alpha = 1 - 0,95 = 0,05$.

The confidence interval is obtained using the same formula as above, but here we need to find the quantities of order 97.5% and 2.5% in the table of the central normal reduced law.

We search the quantiles such that : $P(u_{0,025} < U < u_{0,975}) = 0,95$, with U the pivotal quantity We find :

$$u_{0,025} = -1,96$$

$$u_{0,975} = 1,96$$

(as the central normal reduced law is symmetric). With R we find:

```
qnorm(0.975)
```

```
## [1] 2
```

```
qnorm(0.025)
```

```
## [1] -2
```

-Confidence Interval : The confidence interval for our variable with alpha=5% is of the form :

$$[F - U_{1-\alpha/2} * \sqrt{F(1-F)/n}, F + U_{1-\alpha/2} * \sqrt{F(1-F)/n}]$$

With R we obtain :

```
F<-(1/n)*c
IC3<-F-qnorm(0.975)*(sqrt((F*(1-F))/n))
IC4<-F+qnorm(0.975)*(sqrt((F*(1-F))/n))
#l2 give the length of the interval.
l2<-IC4-IC3
print(c(IC3,IC4,l2))
```

```
## [1] 0.53 0.72 0.19
```

```
paste(c("A confidence interval of the proportion of pslsport at alpha=95% is ["),
      round(IC3,2),":",round(IC4,2), "]")
```

```
## [1] "A confidence interval of the proportion of pslsport at alpha=95% is [ 0.53 : 0.72 ]"
```

We can see that the higher the confidence level of the interval is, the wider our interval will be. To verifie this, we calculated l1 and l2, the respective widths of both intervals.

D - Conformity test

We want to know if sport is very important or not for students. Hence, we have chosen to base our conformity test on the mean. To do this, we have chosen to study the "Frequency sport" variable,which represents the number of hours the students practice a sport. We shall assume that elements of this sample are independent and follow the same distribution. This variable follows a Normal distribution.

1) Conformity test at 5% significance level:

We obtain in R:

```r
x_bar=mean(Frequencysport)
print(x_bar)
```

```
## [1] 1.7
```

The mean is equal to 1,7. Since we have n=102>30 and σ is unknown, the test statistic is:

$$\bar{X} \sim \mathsf{N}(m, \frac{S'}{\sqrt{n}})$$

Therefore, we have these hypothesis:

$$H_0 : m = m_0 = 1,7$$
$$H_1 : m = m_1 \neq m_0$$

As you can see, we are doing a bilateral test at α=5%. The critical region(where H_0 is rejected) at α=5% is :

$$W : \bar{X} < c_1$$

or

$$\bar{X} > c_2$$

We want to find c_1 and c_2:

$$P(D_1 \mid H_0) = \alpha$$
$$<=> P(c_1 > \bar{X} > c_2 | m = m_0) = \alpha$$
$$<=> P(D_0 \mid H_0) = 1 - \alpha$$
$$<=> P(c_1 < \bar{X} < c_2 | m = m_0) = 1 - \alpha$$
$$<=> P(\frac{c_1 - m_0}{\frac{S'}{\sqrt{n}}} < \frac{\bar{X} - m_0}{\frac{S'}{\sqrt{n}}} < \frac{c_2 - m_0}{\frac{S'}{\sqrt{n}}}) = 1 - \alpha$$

We have

$$c_2 = m_0 + u_{1-\frac{\alpha}{2}} * \frac{S'}{\sqrt{n}}$$

and

$$c_1 = m_0 - u_{1-\frac{\alpha}{2}} * \frac{S'}{\sqrt{n}}$$

with $u_{1-\frac{\alpha}{2}}$=1,96 for α=5%. Finally, with α=5%, m_0=1,7, S'=1,54 and n=102 we obtain with R:

```r
t.test(Frequencysport,mu=1.7,alterntive='two.sided',conf.level=0.95,correct=FALSE)
```

```
##
##  One Sample t-test
##
## data:  Frequencysport
## t = -0.2, df = 101, p-value = 0.8
## alternative hypothesis: true mean is not equal to 1.7
## 95 percent confidence interval:
##  1.4 1.9
## sample estimates:
## mean of x
##       1.7
```

Therefore, we have c_1=1,4 and c_2=1,9 so $\bar{X}$>c_1 and $\bar{X}$<c_2. We do not reject H_0. Furthermore, we have p-value=0,8 so p-value>α, it prove that we were right and that we do not reject the null hypothesis at α=5%.

2) Conformity test at 10% significance level:

Now, with the same mean and same hypotheses we want to calculate c_1 and c_2 but with α=10%. Using the same formula, we find:

$$c_1 = m_0 + u_{1-\frac{\alpha}{2}} * \frac{S'}{\sqrt{n}}$$

and

$$c_2 = m_0 - u_{1-\frac{\alpha}{2}} * \frac{S'}{\sqrt{n}}$$

with $u_{1-\frac{\alpha}{2}}$=1,645 for α=10%. We obtain in R:

```
t.test(Frequencysport,mu=1.7,alterntive='two.sided',conf.level=0.90,correct=FALSE)
```

```
##
##  One Sample t-test
##
## data:  Frequencysport
## t = -0.2, df = 101, p-value = 0.8
## alternative hypothesis: true mean is not equal to 1.7
## 90 percent confidence interval:
##  1.5 1.9
## sample estimates:
## mean of x
##       1.7
```

Therefore, we have c_1=1,5 and c_2=1,9 so $\bar{X}>c_1$ and $\bar{X}<c_2$. We do not reject H_0. Furthermore, we have p-value=0,8 so p-value$<\alpha$, it prove that we were right and that we do not reject the null hypothesis at α=10%. To conclude, for both α=5% and α=10% we do not reject the null hypothesis H_0. We can deduce that the students are, in average, practicing a sport approximately 2 times a week, which is very good for their health.

E - Comparison test

1) Samples generating:

Now we are conducting a test to compare the proportions of two subgroups of the population. We are doing this test on the variable grades. Hence, we want to compare the number of people who have more than 12 in average grades in the population of people working more than 10 hours per week (f1) to the number of people who have more than 12 in average grades in the population of people who works less than 10 hours per week (f2). It is interesting because it shows if the number of working hours is the only parameter for students to have good grades. These variables follows a Bernoulli distribution. We have in R:

```
#f1= frequency of people who have 12 or more in average grades in the population
#.   of people working more than 10 hours per week.
#f2= frequency of people who have 12 or more in average grades in the population of
#.   people who works less than 10 hours per week.

f1=12/43
f2=12/59
print (c(f1,f2))
```

```
## [1] 0.28 0.20
```

We get f_1=0,28 and f_2=0,20.

2) Comparison test at 5% significance level :

We want to know if the proportion of the group 1 is bigger than the proportion of the group 2: it's a Left-tailed comparative test. Therefore, we obtain these hypothesis:

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 > p_2 = 0,20$$

To realize this test, we will use the empirical frequencies:

The test statistic is: $F_1 - F_2$, with

$$F_1 = \frac{\sum_{i=1}^{n_1} x_i}{n_1}$$

and

$$F_2 = \frac{\sum_{i=1}^{n_2} y_i}{n_2}$$

Since $n_1 > 30, n_2 > 30$, according to the TCL: $F_1 - F_2 \rightarrow N(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}})$

Hence, we reject H_0 at α=5% if $F_1 - F_2$<c. We want to find c, the critical value. Therefore:

$$P(D_1|H_0) = \alpha <=> P(D_0|H_0) = 1 - \alpha <=> P(F_1 - F_2 > c|p_1 = p_2) = 1 - \alpha$$

$$<=> P(\frac{F_1 - F_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} > \frac{c - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}|p_1 = p_2) = 1 - \alpha$$

Thank to the Slutsky's and continuity theorems, we get:

$$P(\frac{F_1 - F_2}{\sqrt{F(1-F) * (\frac{1}{n_1} + \frac{1}{n_2})}} > \frac{c}{\sqrt{F(1-F) * (\frac{1}{n_1} + \frac{1}{n_2})}}) = 1 - \alpha$$

So we find,

$$c = u_\alpha * (\sqrt{F(1-F)(\frac{1}{n_1} + \frac{1}{n_2})})$$

with

$$F = \frac{n_1 * F_1 + n_2 * F_2}{n_1 + n_2}$$

Since $-u_{1-\alpha} = u_\alpha$, u_α is equal to -1,645 at α=5% level. In R, we have:

```
prop.test(c(12,12),c(43,59),alternative='greater',conf.level=0.95, correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(12, 12) out of c(43, 59)
## X-squared = 0.8, df = 1, p-value = 0.2
```

```
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.066  1.000
## sample estimates:
## prop 1 prop 2
##   0.28   0.20
```

So we can establish that c=0,14. We obtain that $F_1 - F_2$=0,28-0,20=0,08<c so we do not reject H_0. Moreover, we find that p-value=0,2 so p-value>0,05=α, it indicates that we do not reject the null hypotheses.

3) Comparison test at 10% significance level:

Now, we want to do the same but with α=10%. We use p_1=0,26 and p_2=0,14. So we find,

$$c = -u_{1-\alpha} * (\sqrt{F(1-F)(\frac{1}{n_1} + \frac{1}{n_2})})$$

with

$$F = \frac{n_1 * F_1 + n_2 * F_2}{n_1 + n_2}$$

Since $u_{1-\alpha}$ it is equal to -1,282 at α=10%. In R we have:

```
prop.test(c(12,12),c(43,59),alternative='greater',conf.level=0.9,
correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(12, 12) out of c(43, 59)
## X-squared = 0.8, df = 1, p-value = 0.2
## alternative hypothesis: greater
## 90 percent confidence interval:
##  -0.035  1.000
## sample estimates:
## prop 1 prop 2
##   0.28   0.20
```

So we can establish that c=0,11. We obtain that $F_1 - F_2$=0,28-0,20=0,08<c so we do not reject H_0. Furthermore, we find that p-value=0,2 as p-value>0,1=α, it confirms that we do not reject the null hypotheses. To conclude, for both α=5% and α=10% we do not reject the null hypothesis, H_0. We can deduce that the proportion of students who have 12 or more in average grades in the population of people who works less than 10 hours per week is bigger than the proportion of students who have 12 or more in average grades in the population of people working more than 10 hours per week. It means that there is at least another parameter to have good grades, which means that sport is probably a good parameter to consider.

F - Chi-square test

To answer this question, we decided to study the relationship between two variables: Grades and Frequencysport. We think that it is relevant to show if there is a correlation between the grades of the students and how often they practice a sport. We first decided to create the empirical contingency table of our two variables:

```
total <- sum
tab_obs <- addmargins(table(Grades,Frequencysport), FUN=total)
```

```
## Margins computed over dimensions
## in the following order:
## 1: Grades
## 2: Frequencysport
```

```
colnames(tab_obs)=c('no sport at all','1 time per week','2 times','3 times','4 times or more','total')
tab_obs
```

```
##                 Frequencysport
## Grades           no sport at all 1 time per week 2 times 3 times
##   De 0 à 4                     1               0       0       0
##   De 4 à 8                     3               9       2       0
##   De 8 à 12                   13              19      12      15
##   Entre 12 et 16               3               4       2      10
##   Plus de 16                   0               0       2       0
##   total                       20              32      18      25
##                 Frequencysport
## Grades           4 times or more total
##   De 0 à 4                     0     1
##   De 4 à 8                     0    14
##   De 8 à 12                    4    63
##   Entre 12 et 16               2    21
##   Plus de 16                   1     3
##   total                        7   102
```

Then we decided to do the theoretical contingency table for both variables:

```
resul <- chisq.test(Grades,Frequencysport,correct=FALSE)
```

```
## Warning in chisq.test(Grades, Frequencysport, correct = FALSE): Chi-squared
## approximation may be incorrect
```

```
tab_exp <- resul$expected
tab_exp2 <- addmargins(tab_exp,FUN=total)
```

```
## Margins computed over dimensions
## in the following order:
## 1: Grades
## 2: Frequencysport
```

```
colnames(tab_exp2)=c('no sport at all','1 time per week','2 times','3 times','4 times or more','total')
tab_exp2
```

```
##                 Frequencysport
## Grades           no sport at all 1 time per week 2 times 3 times
##   De 0 à 4                 0.20            0.31    0.18    0.25
##   De 4 à 8                 2.75            4.39    2.47    3.43
```

```
##   De 8 à 12               12.35          19.76   11.12   15.44
##   Entre 12 et 16           4.12           6.59    3.71    5.15
##   Plus de 16               0.59           0.94    0.53    0.74
##   total                   20.00          32.00   18.00   25.00
##                Frequencysport
## Grades          4 times or more total
##   De 0 à 4                0.069     1
##   De 4 à 8                0.961    14
##   De 8 à 12               4.324    63
##   Entre 12 et 16          1.441    21
##   Plus de 16              0.206     3
##   total                   7.000   102
```

A rather interesting comment to make here concerns students with grades between 8 and 12 who mostly play sports once a week. By comparing the empirical table with the theoretical table we notice that overall the values are not always close, a sign that the variables are probably not independent. We will confirm or refute this hypothesis using the chi-square test. However, before carrying out our test, we notice that our table contains theoretical values less than 4, we therefore seek to group these variables with others before carrying out our test.

```
#We make a grouping because 2 theoretical values are strictly under 4.
#We group rows 1 and 2
regrpmt<-tab_exp2
regrpmt[1,]<-tab_exp2[1,]+tab_exp2[2,]
#We delete row 1
regrpmt<-regrpmt[-2,]
#We group rows 4 and 3, then we delete row 4
regrpmt[3,]<-tab_exp2[4,]+tab_exp2[3,]
regrpmt<-regrpmt[-4,]
rownames(regrpmt)=c('De 0 à 8','De 8 à 12','De 12 à 20','total')
regrpmt
```

```
##             Frequencysport
## Grades       no sport at all 1 time per week 2 times 3 times 4 times or more
##   De 0 à 8               2.9             4.7     2.6     3.7             1.0
##   De 8 à 12             12.4            19.8    11.1    15.4             4.3
##   De 12 à 20            16.5            26.4    14.8    20.6             5.8
##   total                 20.0            32.0    18.0    25.0             7.0
##             Frequencysport
## Grades       total
##   De 0 à 8      15
##   De 8 à 12     63
##   De 12 à 20    84
##   total        102
```

1) Chi-square test of independence at 5% significance level:

We can now move on to testing the independence of our two variables, by performing a Chi-Square test at the confidence level 5% then 10%. We define the following hypotheses :

$$\begin{cases} H_0 : independent \\ H_1 : variablesnotindependent \end{cases}$$

Test statistic : $T = \sum_{i=1}^{3} \sum_{j=1}^{5} \frac{(n_{ij}-t_{ij})^2}{t_{ij}} \xrightarrow{d} \chi^2_8$ if $t_{ij} \geq 4$ for all i, j.

Decision rule : We reject H_0 with α if T>c. Moreover, $c = \chi^2_{(I-1)(d-1);1-\alpha}$ where I = 3 and d = 5. We find, with $\alpha = 5$, $c = \chi^2_{8;0.95} = 15.51$:

```
qchisq(0.95,8)
```

```
## [1] 16
```

2) Chi-square test of independence at 10% significance level:

Then, with $\alpha = 10$, we find $c = \chi^2_{8;0.9} = 13.36$

```
qchisq(0.9,8)
```

```
## [1] 13
```

We can then calculate T : $T = \sum_{i=1}^{3} \sum_{j=1}^{5} \frac{(n_{ij}-t_{ij})^2}{t_{ij}} = 30$.

We therefore observe that T>c for the two thresholds. We therefore reject the null hypothesis and we can conclude that the variables "Grades" and "Frequencysport" are not independent, so are dependent. In addition, we can calculate the p-value which gives us the following result:

```
chisq.test(Grades,Frequencysport,correct=FALSE)
```

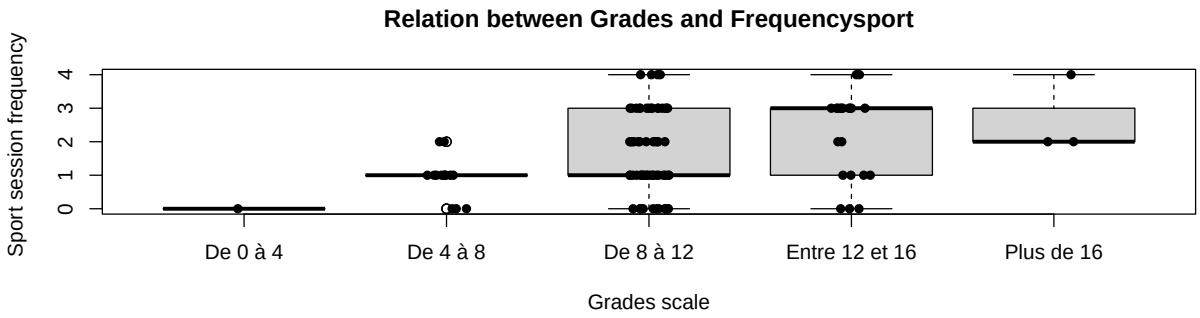
```
## Warning in chisq.test(Grades, Frequencysport, correct = FALSE): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Grades and Frequencysport
## X-squared = 30, df = 16, p-value = 0.02
```

The p-value is equal to 0.02. The p-value is therefore less than alpha, for alpha=5% and alpha=10%. It confirms once again that our two variables are not independent (are dependent).

As the variable Grades is a continuous quantitative variable, here is the equivalent of the scatter plot that we can obtain.

```
boxplot(Frequencysport ~ as.factor(Grades),main="Relation between Grades and Frequencysport",
        xlab="Grades scale",ylab="Sport session frequency ")
stripchart(Frequencysport ~ as.factor(Grades), vertical = TRUE, pch = 16,
           metho = "jitter", add = TRUE)
```




```
cor(as.numeric(as.factor(Grades)),Frequencysport)
```

```
## [1] 0.35
```

Here cor(Grades,Frequencysport) = 0.35>0. We observe that the correlation between Frequencysport and Grades is positive, it means that practicing a sport is a factor to take into account to get good grades. This factor is of course not the only and main one since working time or other factors can be taken into account.

G - Conclusion and comments on the observations made

To conclude, the goal of this project was to link the practice of a sport and good grades for a Dauphine's student. We put in place various tests to see step by step if it was true. And we concluded, thanks to the final test, that for a Dauphine's student practice a sport will enhance its grades unlike a student who doesn't practice at all a sport. We now can say that, for a student, to practice a sport is more than a distraction or a passion, it is a way to evacuate all the stress and enhance their motors and neurological skills. It is safe to say that sport is very important in every one life.

To see if students' opinions were similar to ours, we asked the following question in our questionnaire: "Do you think sport has an impact on academic results?". To this question, 62.75% of students answered yes. This shows that students can feel the impact of a lack of sport on their results in one way or another.

With this project, we were able to show that this intuition was quite correct even if sport isn't the only factor to consider.

So, to achieve good academic results, it's essential to stay in good mental and physical health. Therefore, there is a correlation between the practice of a sport and the grades of a student.