# SABEHAFFAKI_MONIZ_code

Antoine et Emile

05-12-2023

STATISTICS PROJET WITH R. INEQUALITIES IN ACCESS TO HIGHER EDUCATION IN FRANCE

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.

Antoine MONIZ, Emile SABEH AFFAKI

#Question 1. Choose a topic, present its interest and collect data for its study.

For our statistics project, we have decided to study a topic that is both current and relevant: the inequalities in access to higher education in France.

To conduct this study, we are using a database that provides us with detailed statistics on programs that were registered on Parcoursup, the French platform for higher education applications. The data was collected in 2022.

To address this topic, we are going to examine several types of inequalities, especially those related to gender, the educational background of candidates, and their financial situation, among others. The data comes from the website data.enseignementsup-recherche.gouv.fr, it includes information on the admission decisions granted by institutions at the end of the Parcoursup allocation process for the 2022 session. Some of the characteristics of our sample describe the institutions (for example: their status : public, private, hybrid, etc), while others provide data on the programs offered (for example, if they're selective or non-selective programs , the percentage of admitted students to the program holding a CROUS scholarship, etc).

As students who joined the Paris Dauphine University earlier this year, we decided to analyse the inequalities in access to higher education partly because of our personal experiences, but mainly because the topic sparked our interest. Right after High school, it would have been harder for us to get admission to Dauphine since it is no secret that most students who attend the school after the Baccalaureate have stellar grades throughout their schooling, and often hold a Bac with honors ("mention Très Bien"). As the university offers very selective programs, there are many occurrences of students getting rejected despite their excellent grades and extra-curricular activities.

We are also very sensible to financial and gender inequalities in access to education, so our personal perceptions comforted us in the choice of our subject, and we decide to study a wider scope of inequalities in access to higher education.

To conduct a thorough analysis, we have taken the initiative to adapt the data set to better match our research criteria. The final database we used includes 13 variables, observed on a sample of 13645 programs offered on Parcoursup in year 2022. The theoretical mean and the theoretical variance are unknown. We chose to translate the variables we used from French to English, but not the data itself, to preserve the french intricacies of the higher education system (for example, mention Très bien, établissement privé sous contrat, etc).

source: https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-parcoursup/information/

Let’s start by clearing our work space.

```
rm(list=ls())
```

Let’s download the necessary packages.

```
library(readxl)
```

Let’s make sure that our numerical result will be rounded off to two decimal places.

```
options(digits=4)
```

Now, we import our database.

```
base <- read_excel('SABEHAFFAKI_MONIZ_base.xlsx')
```

Then, we ask R to show the base, to check for any missing values, and we attach the database so that we don’t need to mention it every time.

```
anyNA(base)
```

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

```
attach(base)
```

#Question 2. Description of data:(a) Define the population, the statistical units and the variables reported in thedatabase (name, type, unit of each variable).

Our population is a sample of the programs offered on Parcoursup in 2022. We will now list our variables and their titles in the original database.

Variables can take different forms:

Discrete Quantitative Variables: These are variables that can take a finite or countably infinite number of values. For example, in our database, the following variable is a discrete quantitative variable : program_capacity = “Capacité de l’établissement par formation.” To represent discrete quantitative variables, we use bar charts.

Continuous Quantitative Variables: These are variables that can take an infinite number of values within a given range. They typically represent measurements. For example, in our database, the following variables are continuous quantitative variables :

percentage_admitted_girls = “% d’admis dont filles”.

percentage_admitted_sameacademy = “% d’admis néo bacheliers issus de la même académie”.

percentage_admitted_withscholarship = “% d’admis néo bacheliers boursiers”.

percentage_admitted_nodistinction = “% d’admis néo bacheliers sans mention au bac”.

percentage_admitted_honors = “% d’admis néo bacheliers avec mention Très Bien au bac”.

percentage_admitted_generalbac = “% d’admis néo bacheliers généraux”.

percentage_admitted_bactechno = “% d’admis néo bacheliers technologiques”.

percentage_admitted_bacpro = “% d’admis néo bacheliers professionnels”.

percentage_acces_rate = “Taux d’accès.

To represent continuous quantitative variables, we use histograms.

Nominal Qualitative Variables: These are variables that describe a category or state without any intrinsic order. These variables are used to label or name attributes.For example, in our database, the following variables are nominal qualitative variables :

institution_status = “Statut de l’établissement de la filière de formation (public, privé. . . )”

institution_region = “Région de l’établissement”

selection_type = “Sélectivité”

To represent them, we use pie charts.

Ordinal Qualitative Variables: These are variables that describe a category or state with a specific order or ranking. They indicate a hierarchy or order in the data. For example, School grades (A, B, C, D, F). To represent them, we use an organ pipe diagram. No ordinal qualitative variables in our dataset.

#Our variables :

the variable “institution_status” which can take the following values : “Public”, “Privé sous contrat d’association”, “Privé enseignement supérieur”, “privé hors contrat”.

the variable “institution_region” which can take as values the names of all the french regions.

the variable “program_capacity” which can take as values any positive integer corresponding to the number of places a program can host.

the variable “percentage_admitted_girls” which can take as values the percentages of girls admitted to a program, from 0 to 100.

the variable “percentage_admitted_sameacademy” which can take as values the percentages of students admitted to a program which takes place in the academy in which they passed their Baccalaureate, from 0 to 100.

the variable “percentage_admitted_withscholarship” which can take as values the percentages of students who hold a scholarship who were admitted to a program, from 0 to 100.

the variable “percentage_admitted_nodistinction” which can take as values the percentages of students who got the Bac diploma without “mention” who were admitted to a program, from 0 to 100.

the variable “percentage_admitted_honors” which can take as values the percentages of students who got the Bac diploma with “mention TB” who were admitted to a program, from 0 to 100.

the variable “percentage_admitted_generalbac” which can take as values the percentages of students who got the General Bac diploma who were admitted to a program, from 0 to 100.

the variable “percentage_admitted_bactechno” which can take as values the percentages of students who got the Technological Bac diploma who were admitted to a program, from 0 to 100.

the variable “percentage_admitted_bacpro” which can take as values the percentages of students who got the Bac Pro diploma who were admitted to a program, from 0 to 100.

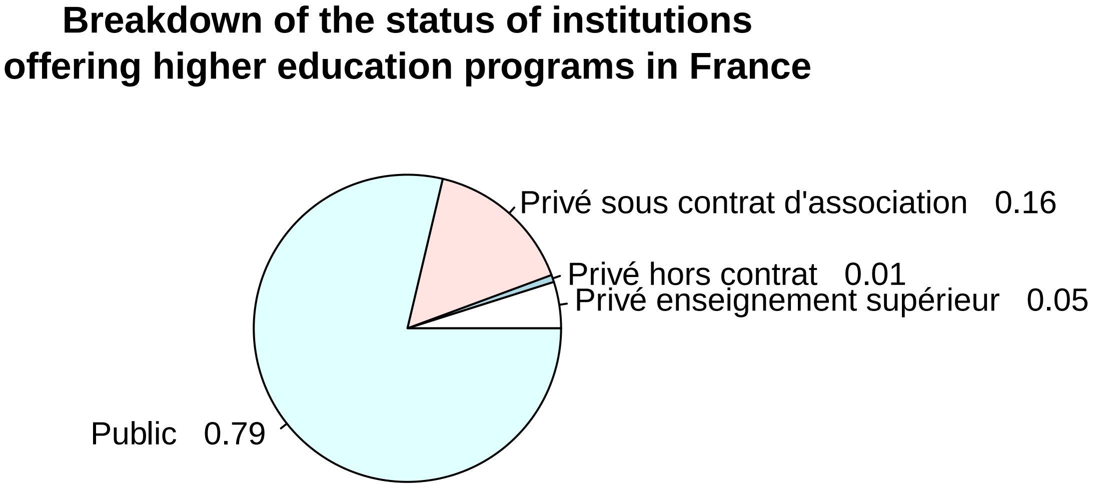
the variable “percentage_acces_rate” which can take as values the percentages of admitted students among those who applied, from 0 to 100.

the variable “selection_type” which can take the values “formation sélective” or “formation non sélective”.

#(b) In order to describe the variables, generate a frequency table, and give a graph of the empirical distribution for each variable (see TP1 and prérentrée R).

#Frequency table of the institution status

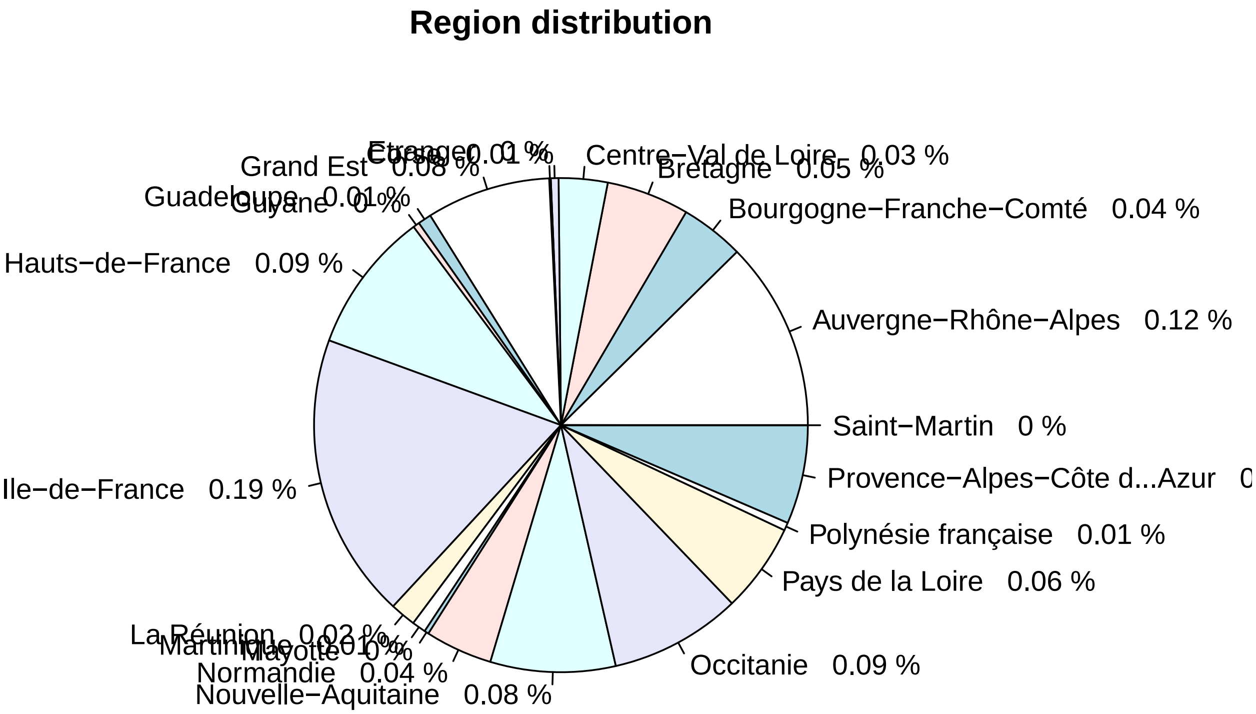
```
Total<-sum
tab<- prop.table(table(institution_status))
pie(prop.table(table(institution_status)),main='Breakdown of the status of institutions
offering higher education programs in France',labels=paste(rownames(tab),' ',round(tab,2),''))
```



We notice that the proportion of public institutions is much higher, indeed there are 79% of public institutions offering higher education programs. In comparison, only 16% of private institutions under contract of association, 4% of private higher education institutions and 1% of private institutions without contract offer programs on Parcoursup. We can conclude that private education in France is uncommon compared to other countries as there is only 20% of higher education training in the private sector, this avoids many inequalities related to the cost of schooling.

#Frequency table of the regions where the institutions are located

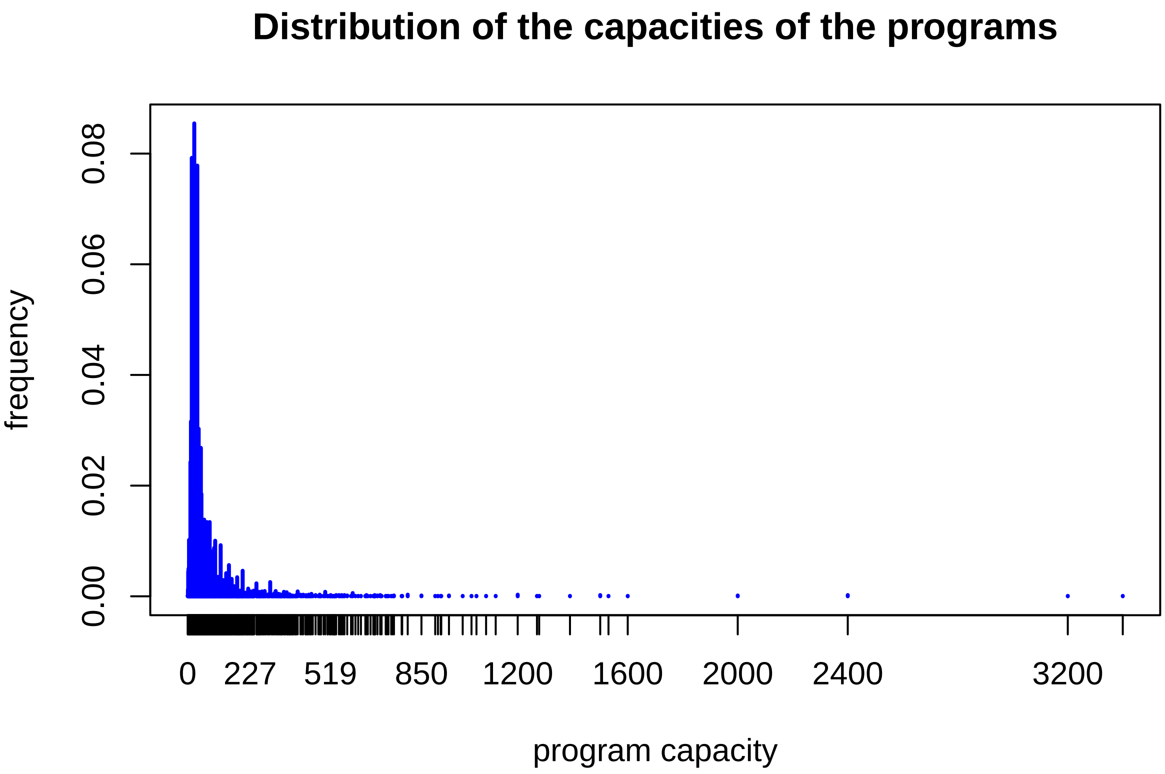
```
Total<-sum
tab1<- prop.table(table(institution_region))
addmargins(tab1,FUN=Total)
pie(tab1,main='Region distribution',labels=paste(rownames(tab1),' ',round(tab1,2),'%'))
```



We notice from this frequency table that the regions offering the most programs are Ile de France with 19% of the offered training, Auvergne-Rhone-Alpes with 12%, followed by Occitanie and Hauts-de-France with 9% each. This represents well the most populated regions in France, so we can conclude that the higher education programs are fairly well spread in the French regions and we do not notice any blatant inequality.

#Frequency table of the capacity of each program offered on the platform

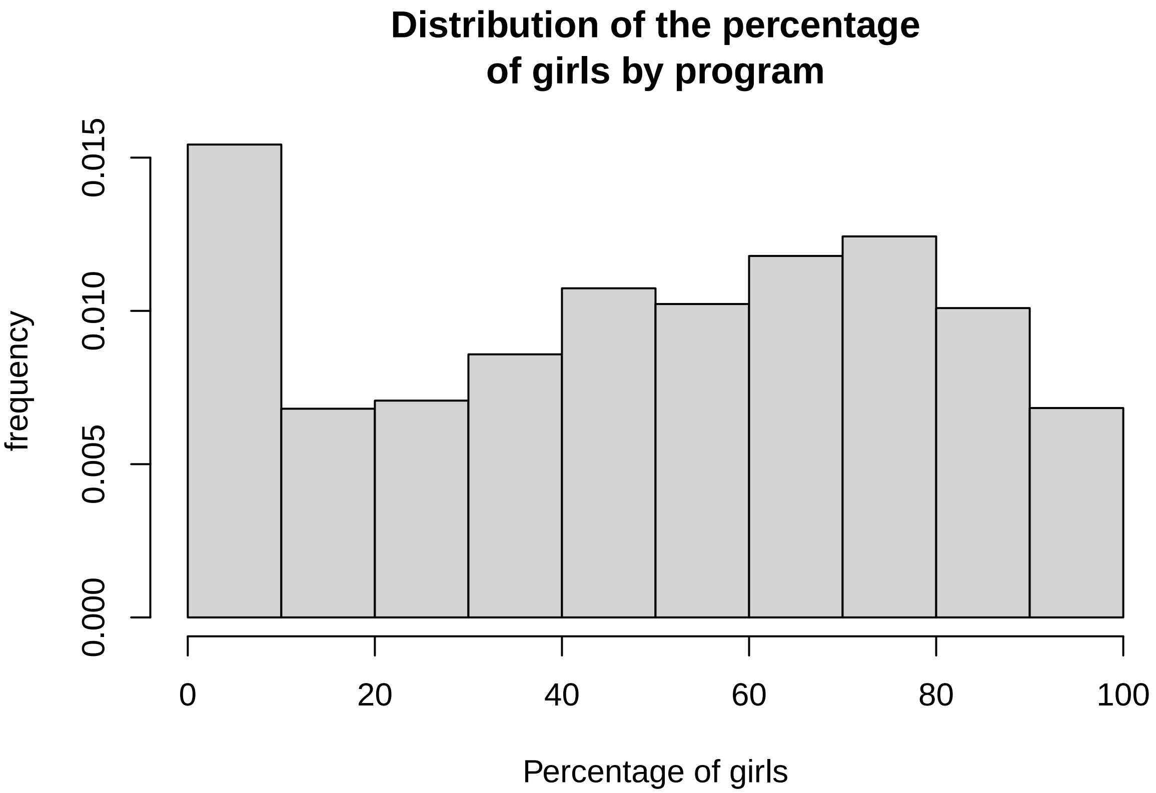
```
tab3<- prop.table(table(program_capacity))
addmargins(tab3,FUN =Total)
plot(tab3,main='Distribution of the capacities of the programs',xlab='program capacity',
ylab='frequency',col='blue')
```



From this frequency table, we observe that the majority of programs have a capacity of less than 100 places, and we also notice a few exceptions that offer thousands of places.

#Frequency table of the proportion of girls admitted in a program Slicing every 10%

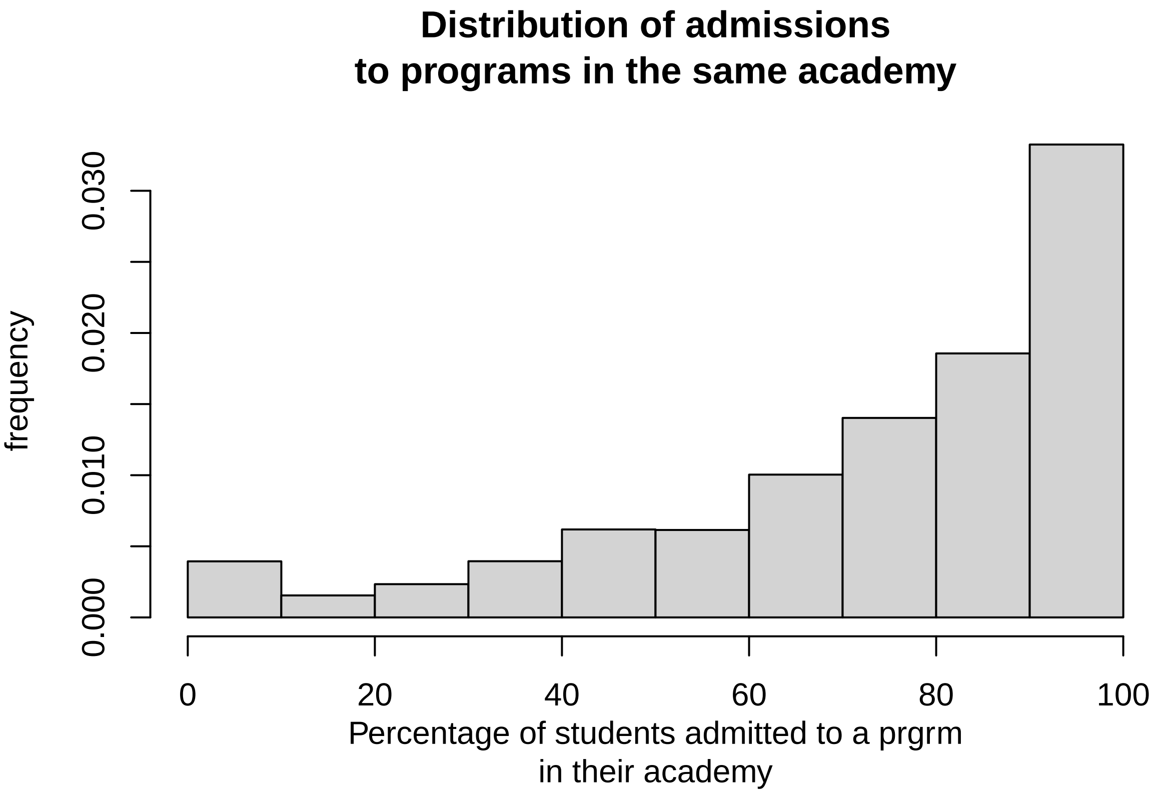
```
borne<- seq(0,100,10)
hist(percentage_admitted_girls,breaks = borne,freq=F,main='Distribution of the percentage
of girls by program',xlab='Percentage of girls',ylab='frequency')
```



We notice that very few institutions have admitted almost no girls (this can be explained by the fact that some trainings are in very male-dominated fields receiving few female applications) Otherwise, we observe that the majority of the trainings admit between 40% and 80% girls We also notice some institutions that have accepted almost only girls (this can be explained by the fact that some trainings are in very female-dominated fields receiving few male applications) but less so than the institutions that have accepted no girls.

#Frequency table of the percentage of students admitted in a program in their academy

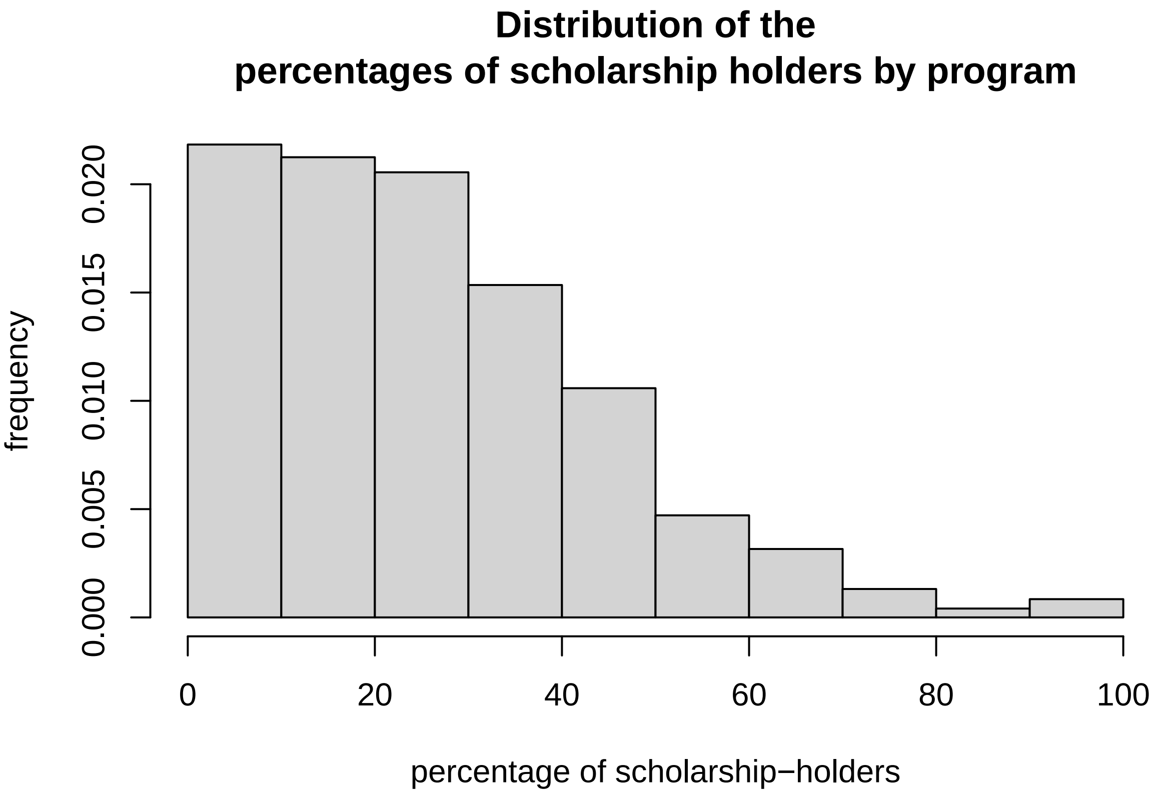
```
hist(percentage_admitted_sameacademy,breaks = borne,freq=F,main='Distribution of admissions
to programs in the same academy',xlab='Percentage of students admitted to a prgrm
in their academy',ylab='frequency')
```



We observe that the majority of training programs admit students from the same academy, which can be explained by the limited mobility of students, for example, for programs that are present across the country. We note a very small minority of establishments that accept very few students from the same academy (example of selective trainings present only in a specific location).

#Frequency table of the percentage of scholarship-holders admitted in a program

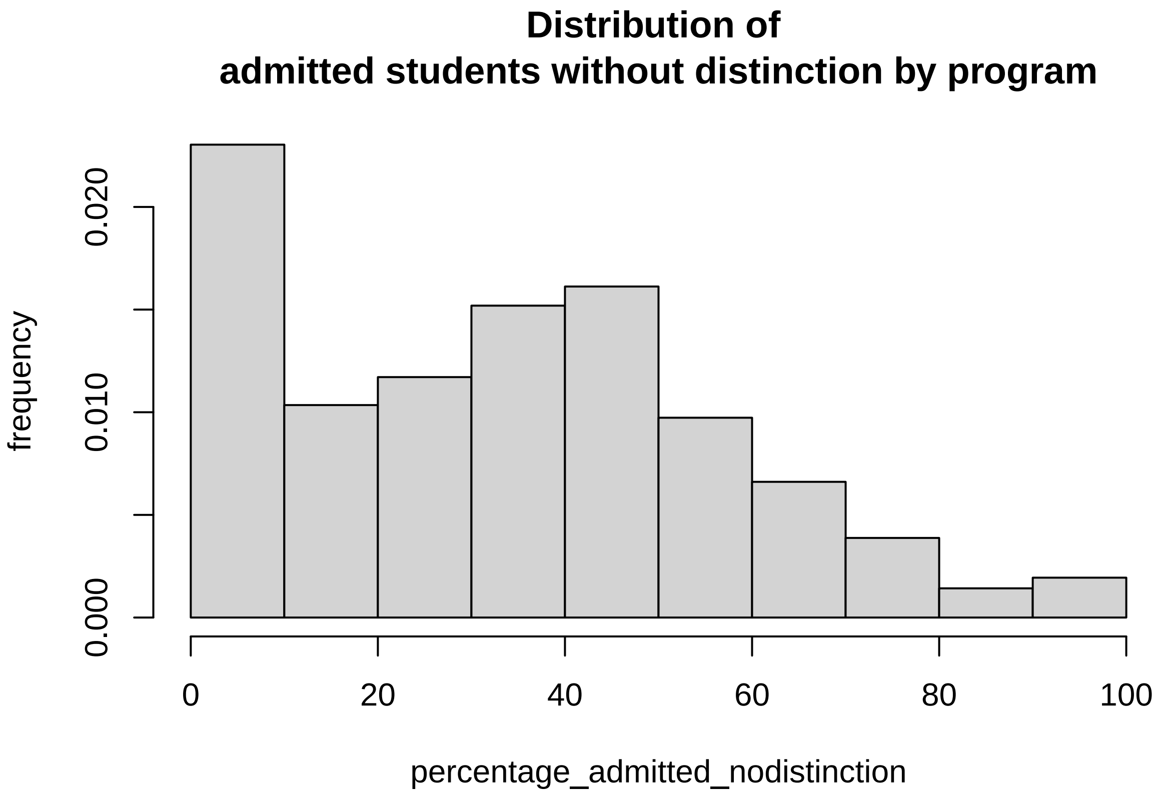
```r
hist(percentage_admitted_withscholarship,breaks = borne,freq=F,main='Distribution of the
percentages of scholarship holders by program',xlab='percentage of scholarship-holders',
ylab='frequency')
```



We notice that the majority of the training programs have a low proportion of scholarship recipients.

#Frequency table of the percentage of students without distinction admitted

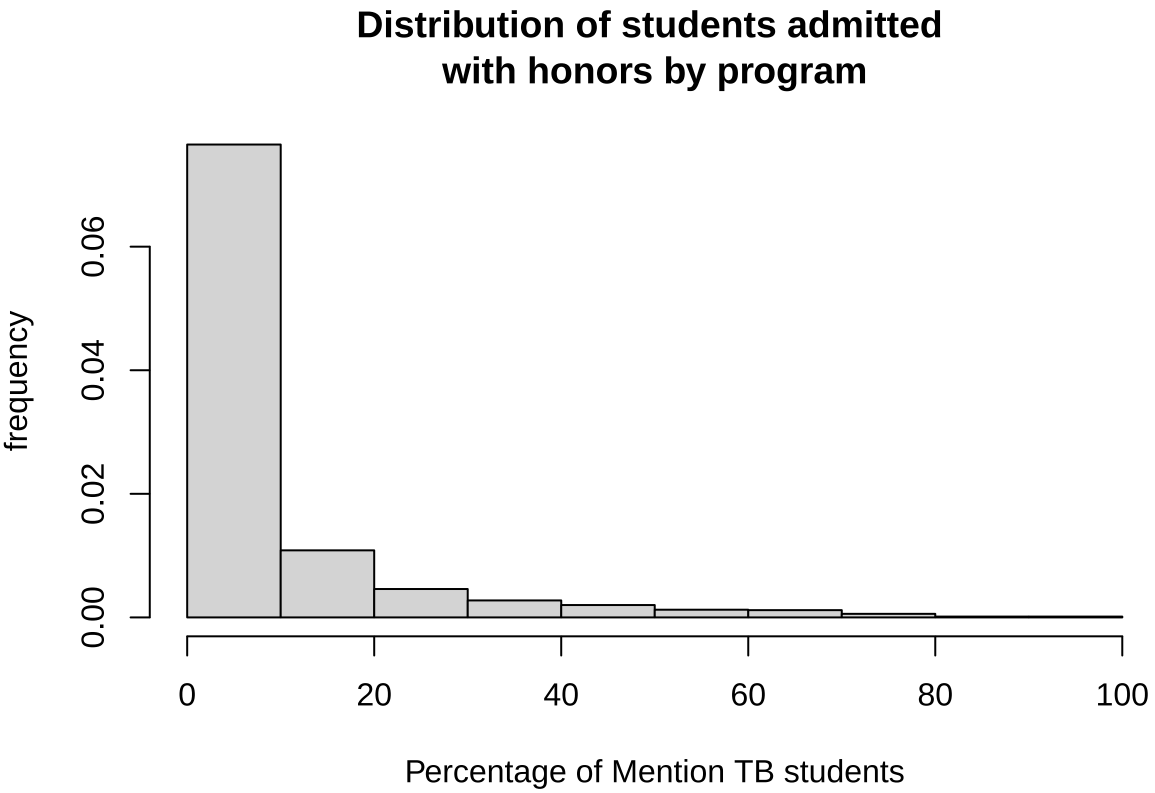
```r
hist(percentage_admitted_nodistinction,breaks = borne,freq=F,main='Distribution of
admitted students without distinction by program',ylab='frequency')
```



We notice that the majority of the training programs have less than 50% of students without honors. A small part have almost only students without honors (less selective training programs).

#Frequency table of the percentage of students who received the 'mention Très Bien' honors admitted in a training program

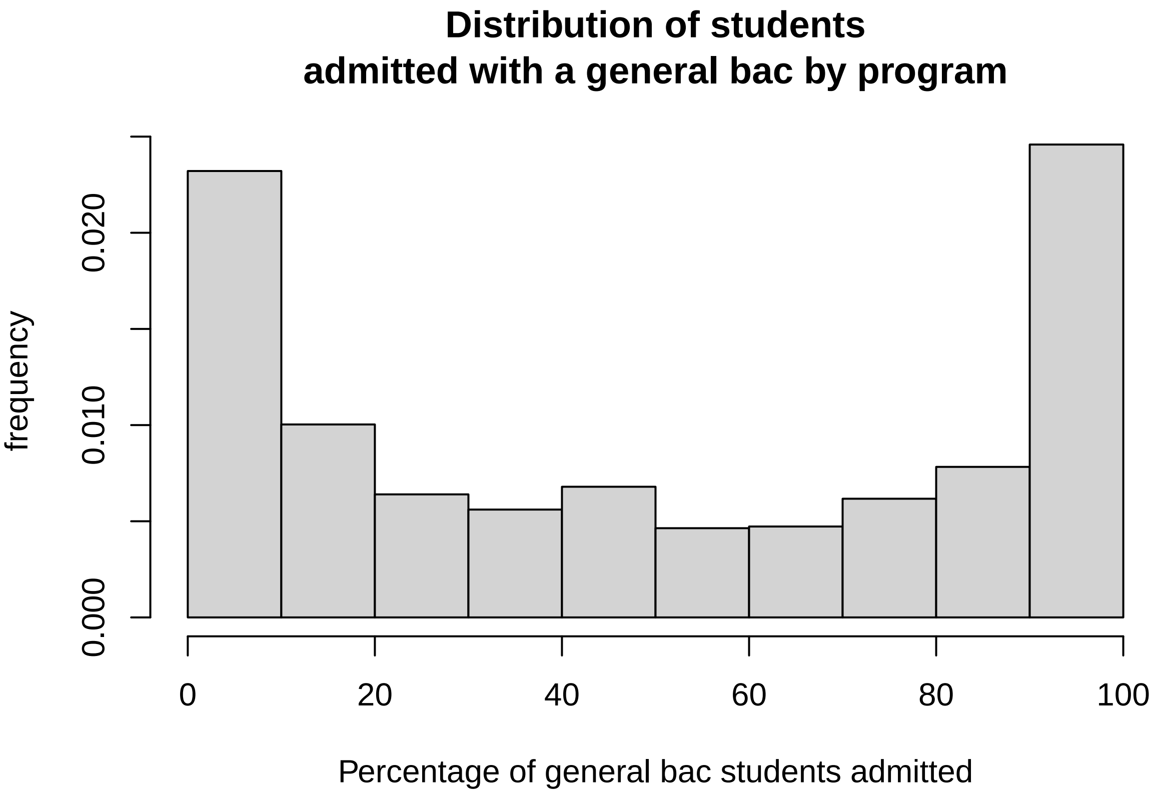
```r
hist(percentage_admitted_honors,breaks = borne,freq=F,main='Distribution of students admitted
with honors by program',xlab='Percentage of Mention TB students',ylab='frequency')
```



We notice that the majority of the training programs have less than 20% of students with 'Very Good' honors.

#Frequency table of the percentage of students with a Bac général who were admitted in a training program

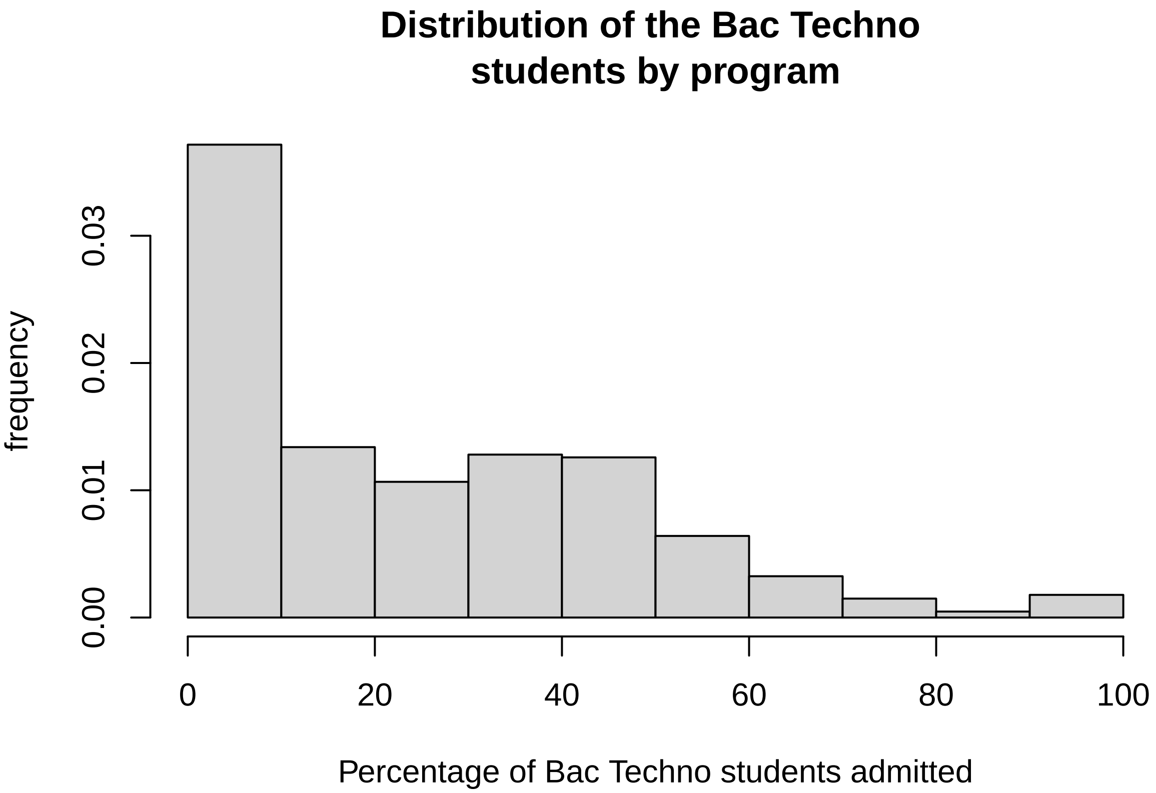
```
hist(percentage_admitted_generalbac,breaks = borne,freq=F,main='Distribution of students
admitted with a general bac by program',xlab='Percentage of general bac students admitted',
ylab='frequency')
```



We notice that there are as many training programs that accept a large number of general stream students as there are those that accept few. This is not surprising because notably, certain programs are tailored and well-aligned with the curriculum of general baccalaureate students, thereby catering to a wider audience with a general baccalaureate focus. Conversely, some are specifically designed for students pursuing technical (bac technologique) or vocational (bac pro) baccalaureate paths.

#Frequency table of the percentage of students with a Bac Technologique admitted to a program

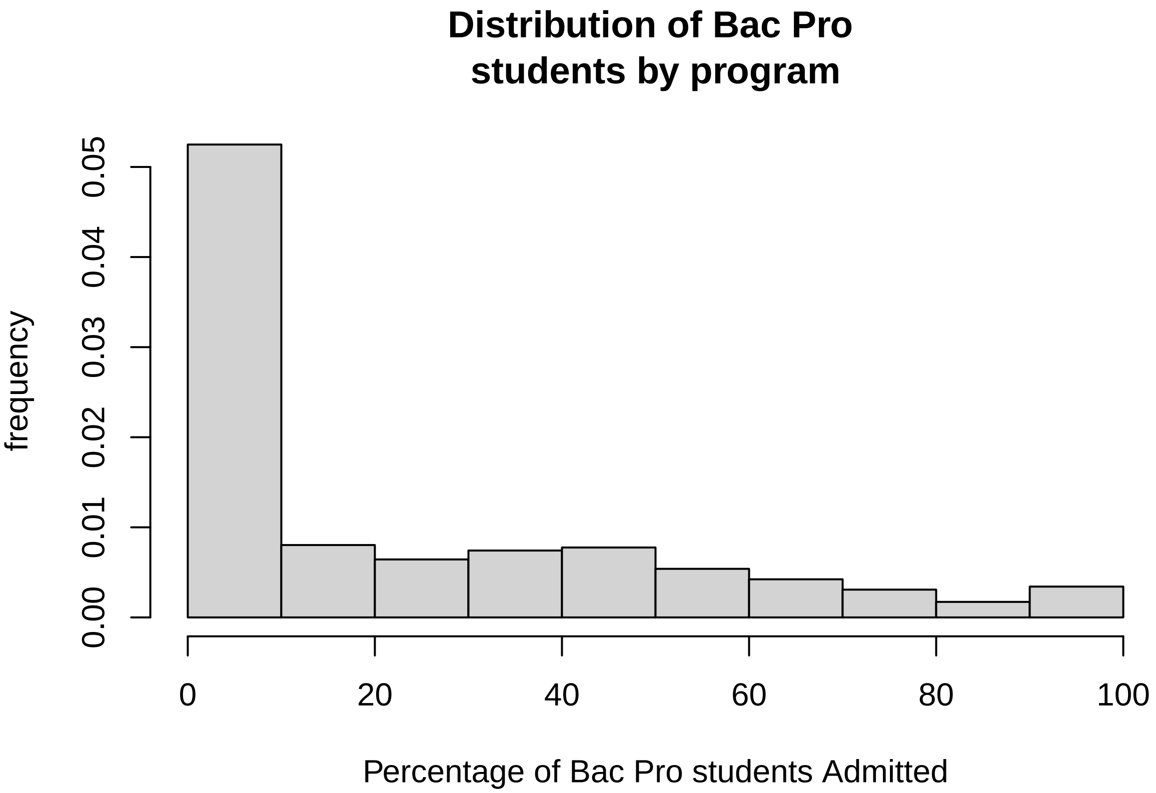
```
hist(percentage_admitted_bactechno, breaks = borne,freq=F,main='Distribution of the Bac Techno
students by program',xlab='Percentage of Bac Techno students admitted',ylab='frequency')
```



We notice that the majority of the training programs admit a low proportion of technology students. This is not surprising because notably, certain programs are tailored and well-aligned with the curriculum of general baccalaureate students, thereby catering to a wider audience with a general baccalaureate focus. Conversely, some are specifically designed for students pursuing technical (bac technologique) or vocational (bac pro) baccalaureate paths.

#Frequency table of the percentage of students who were in vocational high schools and got admitted to a program

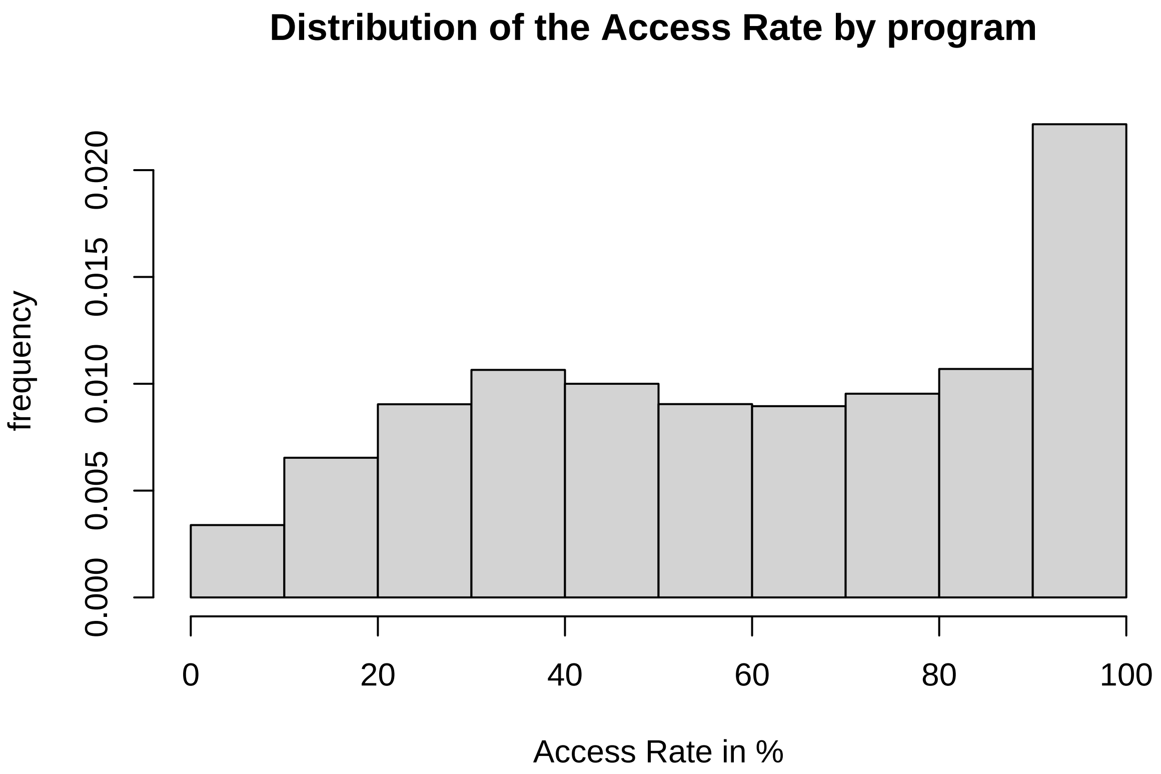
```
hist(percentage_admitted_bacpro,breaks = borne,freq=F,main='Distribution of Bac Pro
students by program',xlab='Percentage of Bac Pro students Admitted',ylab='frequency')
```



We notice that the majority of the training programs admit a low proportion of vocational students. This is not surprising because notably, certain programs are tailored and well-aligned with the curriculum of general baccalaureate students, thereby catering to a wider audience with a general baccalaureate focus. Conversely, some are specifically designed for students pursuing technical (bac technologique) or vocational (bac pro) baccalaureate paths. We may add that the primary objective of a vocational baccalaureate is geared towards preparing individuals for the workforce, emphasizing practical skills and employment rather than serving as a pathway to pursue further studies at the university level.

#Frequency table of the Access Rate by program

```
hist(percentage_access_rate,breaks = borne,freq=F,main='Distribution of the Access Rate by program ',
xlab='Access Rate in %',ylab='frequency')
```



We notice that there are training programs with all access rates, but we can also observe that there are more programs with a high access rate than the opposite.

#Question 3. Conduct a point estimation and a confidence interval estimation at levels 90% and 95% for an unknown parameter of the population (see TP3). Comment on the results.

Point estimate of the theoretical population mean of the percentage of girls admitted

```r
xbar<- mean(percentage_admitted_girls)
print(xbar)
```

```
## [1] 49.39
```

We notice that the percentage of girls admitted in the training programs is on average 49.39%, which appears to be fairly equal, although there is a 1% difference for perfect equality.

A 90% confidence interval estimate of the theoretical mean with sigma unknown and n>30.

$\overline{percentageofadmittedgirls} \sim N\left(m, \frac{\sigma}{\sqrt{n}}\right)$

```r
n<-length(percentage_admitted_girls)
scor<- sd(percentage_admitted_girls)
a1<-xbar -  qnorm (0.95)*scor/sqrt(n)
b1<-xbar+ qnorm (0.95)*scor/sqrt(n)
c(a1,b1)
```

```
## [1] 48.97 49.80
```

We notice that the 90% confidence interval for the percentage of girls admitted in the training programs is between 48.97% and 49.80%, which appears to be fairly equal, although there is a 1% difference for perfect equality.

A 95% confidence interval estimate of the theoretical mean with sigma unknown and n>30.

$\overline{percentageofadmittedgirls} \sim N\left(m, \frac{\sigma}{\sqrt{n}}\right)$

```r
a2<-xbar -  qnorm (0.975)*scor/sqrt(n)
b2<-xbar+ qnorm (0.975)*scor/sqrt(n)
c(a2,b2)
```

```
## [1] 48.89 49.88
```

We notice that the 95% confidence interval for the percentage of girls admitted in the training programs is between 48.89% and 49.88%, which appears to be fairly equal, although there is a 1% difference for perfect equality.

Note that the point and confidence interval estimates are consistent

#Question 4. Conduct a conformity test on a parameter of the population at a 5% significance level and then at a 10% signicance level (see TP4). Comment on the results.

We are studying the average percentage of admitted candidates to a program who hold a scholarship to see if the selection process is satisfactory or not. For the selection process to be fair, we make the conjecture that m = 30 would be the theoretical mean of the percentage of admitted scholarship-holders to a program.

Here we are in case 3 : $\sigma$ unknown, (X1, X2, . . . , Xn) i.i.d. (m, $\sigma$) and n > or = 30

$$\begin{cases} H_0 : m = 30 \quad \text{(selection process is satisfactory)} \\ H_1 : m < 30 \quad \text{(selection process is unsatisfactory and creates inequalities based on financial income)} \end{cases}$$

Test statistic : $\overline{X}$ where $\frac{\overline{X}-m}{s'/\sqrt{n}} \to \mathcal{N}(0,1)$ since n > 30

Decision rule : W: $\overline{X}$ < c

Critical value ? $c = m_0 - U_{1-\alpha} \cdot \frac{S'}{\sqrt{n}}$

```r
x <- base$percentage_admitted_withscholarship
mean(x)
```

```
## [1] 26.49
```

```r
sd(x)
```

```
## [1] 19.5
```

```r
cvalue1 <- 30-1.645*19/sqrt(13644)
cvalue2 <- 30-1.282*19/sqrt(13644)
c(cvalue1,cvalue2)
```

```
## [1] 29.73 29.79
```

For $\alpha = 5\%$ $\overline{X}$ < c because 26.49 < 29.73

For $\alpha = 10\%$ $\overline{X}$ < c because 26.49 < 29.79

We reject $\overline{X}$ in both cases and we conclude that the selection process is unsatisfactory and creates inequalities based on financial income.

Other method using the p-value :

```r
t.test(x,mu=30,alternative='less')
```

```
##
##  One Sample t-test
##
## data:  x
## t = -21, df = 13643, p-value <2e-16
## alternative hypothesis: true mean is less than 30
## 95 percent confidence interval:
##   -Inf 26.76
## sample estimates:
## mean of x
##     26.49
```

```r
t.test(x,mu=30,alternative='less',conf.level=0.9)
```

```
##
##  One Sample t-test
##
## data:  x
## t = -21, df = 13643, p-value <2e-16
## alternative hypothesis: true mean is less than 30
## 90 percent confidence interval:
```

```
##  -Inf 26.7
## sample estimates:
## mean of x
##     26.49
```

This leads us to reject the null hypothesis since the p-value is smaller than $\alpha$ at a 5% significance level and at a 10% significance level. We conclude that the selection process is unsatisfactory with on average less than 30% of candidates holding a CROUS scholarship getting offered admission to the programs in our data base. Looking it up online, we found that in year 2022, the percentage of candidates holding a scholarship getting access to higher education on Parcoursup has hit the lowest level in a ten-year period. Despite the quotas put in place by the ministry of education and the french regions, efforts are to be maintained to decrease inequalities in access to higher education based on financial income.

#Q5 Conduct a test to compare the theoretical characteristics of two subgroups of the population. Again, use a 5% and then a 10% significance level (see TP5). Comment on the results.

We are now researching if selective and non-selective programs admit the same percentage of students holding a scholarship. We are therefore going to conduct difference tests, comparing two means and two variances.

We would like to compare the average percentage of admitted students with a scholarship to selective programs, noted m1, to the average percentage of admitted students with a scholarship to non-selective programs.

We have 2 samples : one sample of 3028 non-selective programs and one sample of 10616 selective programs. We are about to compare the theoretical means of the percentages of students admitted with scholarships in both subsets x1bar and x2bar to decide whether m1=m2 or m1≠m2.

We've seen in class that the decision rule of a comparison of means test differs according to whether theoretical variances of the variable of interest are equal or not. So we need to test for equality of variances first.

```r
base <- read_excel('SABEHAFFAKI_MONIZ_base.xlsx')
base1 <- subset(base, selection_type=='formation non sélective')
base2 <- subset(base, selection_type=='formation sélective')
str(base1) #10616 selective programs
str(base2) #3028 non-selective programs
```

```r
x1 <- base1$percentage_admitted_withscholarship
x2 <- base2$percentage_admitted_withscholarship
xbar1 <- mean(x1)
xbar2 <- mean(x2)
sd1 <- sd(x1)
sd2 <- sd(x2)
c(xbar1,xbar2)
```

```
## [1] 27.92 26.08
```

```r
c(sd1,sd2)
```

```
## [1] 17.49 20.01
```

27.92% of scholarship-holders by program on average in the sample of non-selective programs VS 26.08% of scholarship-holders by program on average in the sample of selective programs. It appears that there are more students admitted with scholarships to non-selective programs than to selective ones. We will test this by conducting comparison of means tests.

We test

$$\begin{cases} H_0 : m_1 = m_2 \\ H_1 : m_1 > m_2 \end{cases}$$

First we test for equality of variances of the variable of interest (the percentage of admitted scholarship-recipients by program).

In our test, the variances are equal under $H_0$ and they are different under $H_1$. We wish to conduct a bilateral test.

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

```
var.test(x1,x2,alternative='two.sided')
```

```
##
##  F test to compare two variances
##
## data:  x1 and x2
## F = 0.76, num df = 3027, denom df = 10615, p-value <2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7218 0.8091
## sample estimates:
## ratio of variances
##             0.7639
```

p-value is below 5%, so we reject the null hypothesis of equality of the variances at $\alpha = 5\%$. We will assume the variances are not equal.

We are in case 4 for the test of equality of means : $\sigma_1^2$ and $\sigma_2^2$ unknown + unequal and n1, n2 > 30

Test statistic (in test comparing two means) :

$$X_1 - X_2 \approx \mathcal{N}(m_1 - m_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

Decision Rule: Reject $H_0$ at threshold $\alpha$ if $\bar{X}_1 - \bar{X}_2 > c$

c? c=

$$U_{1-\alpha}\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}$$

=

$$U_{1-\alpha}\sqrt{\frac{s_1'^2}{n_1} + \frac{s_2'^2}{n_2}}$$

```
c1 <- 1.645*(sqrt(17.49^2/3028+20.01^2/10616))
print(c1) #for alpha=5%
```

```
## [1] 0.6127
```

```r
c2 <- 1.282*(sqrt(17.49^2/3028+20.01^2/10616))
print(c2) #for alpha=10%
```

```
## [1] 0.4775
```

```r
print(xbar1-xbar2)
```

```
## [1] 1.835
```

Conclusion : Here, 1.84 > 0.61 and 1.84 > 0.48 so we reject $H_0$ at threshold $\alpha$=5% and =10%.

Other way to solve this : same thing, right-tailed test with $H_0$ selective and non-selective programs admit on average the same percentage of students holding CROUS scholarships and $H_1$ non-selective programs admit a greater percentage of students holding CROUS scholarships than selective programs. If we are able to reject $H_0$, we will confirm the observation we made earlier and we will be able to conclude that non-selective programs admit a greater percentage of students holding CROUS scholarships than selective programs.

```r
t.test(x1,x2,alternative = 'greater', var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = 4.9, df = 5491, p-value = 4e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.222   Inf
## sample estimates:
## mean of x mean of y
##     27.92     26.08
```

p-value smaller than 5% so we can reject $H_0$. The 2 methods we used lead us to the same result. Non-selective programs admit a greater percentage of students holding CROUS scholarships than selective programs. This highlights the social inequalities in access to higher education, students coming from low-income households who are eligible for CROUS scholarships have a harder time getting admission to selective programs since those programs tend to admit less scholarship-holders on average than the non-selective programs. This limits scholarship holders in their choices. We notice that the quotas of scholarship holders are reinforced by non-selective programs better than by selective programs, which raises concerns about equality in access to selective programs in higher education.

#Q6 Conduct a chi-squared test of independence for two variables at a 5% and then at a 10% significance level (if necessary, group consecutive classes). Comment on the results.

We are trying to understand the factors that explain the percentage of admitted graduates with honors to a program. Hence, we would like to know if the access rate to a higher education program influences the percentage of admitted graduates with honors (mention Très bien).

```r
base_independence <- read_excel('SABEHAFFAKI_MONIZ_base.xlsx', sheet="independence")
Taux_accès <- base_independence$percentage_access_rate
proportion_Très_Bien <- base_independence$percentage_admitted_honors
Total <- sum
Tout <- sum
```

Let's construct the observed contingency table of the variables access rate (percentage_access_rate) and percentage of admitted students with honors (percentage_admitted_honors)

Those are 2 continuous quantitative variables : they can take any real non-negative number with zero probability (so many values that we cannot count them). So we need to regroup them in classes.

```
summary(percentage_access_rate) #to help us construct our intervals / classes
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##     1.0    36.0    62.0    60.7    88.0   100.0      95
```

```
summary(percentage_admitted_honors)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    0.00    2.00    8.39   10.00  100.00
```

```
percentage_access_rate_borne <- seq(0,100,20) #values ranging from 0 to 100 with an interval of 20
percentage_access_rate_borne
```

```
## [1]   0  20  40  60  80 100
```

```
percentage_admitted_honors_borne <- seq(0,100,20)
percentage_admitted_honors_borne
```

```
## [1]   0  20  40  60  80 100
```

```
#We are covering all the possible values.
percentage_access_rate_cut <- cut(percentage_access_rate,percentage_access_rate_borne)
percentage_admitted_honors_cut <- cut(percentage_admitted_honors,percentage_admitted_honors_borne)
#we regroup the variables of interest according to the classes they belong to
#which are defined by these vectors here.
percentage_access_rate[1]
```

```
## [1] 23
```

```
percentage_access_rate_cut[1]
```

```
## [1] (20,40]
## Levels: (0,20] (20,40] (40,60] (60,80] (80,100]
```

```
#To verify that the first program with an access rate of 23% is indeed in the class (20,40]
TC_emp <- table(percentage_access_rate_cut,percentage_admitted_honors_cut)
addmargins(TC_emp,FUN=Total)
```

```
## Margins computed over dimensions
## in the following order:
## 1: percentage_access_rate_cut
## 2: percentage_admitted_honors_cut
##                            percentage_admitted_honors_cut
## percentage_access_rate_cut (0,20] (20,40] (40,60] (60,80] (80,100] Total
##                   (0,20]      452     221     162     109       14   958
##                   (20,40]    1209     234     129     104       12  1688
##                   (40,60]    1140     162      80      24        1  1407
##                   (60,80]     962     181      46       2        4  1195
##                   (80,100]   1696     204      26       1        5  1932
##                   Total      5459    1002     443     240       36  7180
```

```
#or another way to construct our observed contingency table is to name the columns and the rows,
#this way we do not need to use FUN=Total
TC_emp2 <- addmargins(TC_emp)
colnames (TC_emp2)<-c('very low','low','medium','high','very high','Total')
rownames (TC_emp2)<-c('very low','low','medium','high','very high','Total')
TC_emp2
```

```
##                            percentage_admitted_honors_cut
## percentage_access_rate_cut very low  low medium high very high Total
##                  very low       452  221    162  109        14   958
##                  low           1209  234    129  104        12  1688
##                  medium        1140  162     80   24         1  1407
##                  high           962  181     46    2         4  1195
##                  very high     1696  204     26    1         5  1932
##                  Total         5459 1002    443  240        36  7180
```

We have a better presentation this way. We get the observed contingency table of the 2 variables of interest and we can read the joint frequencies of the two variables : for example we have 452 programs with a very low access rate for which the percentage of admitted students with honors is very low.

The last line represents the marginal frequencies of the variable percentage_admitted_honors_cut (percentage of admitted students with honors) so we can say for example that 240 programs have a high percentage of admitted students who graduate their Baccalaureate with honors.

The last column represents the marginal frequencies of the variable percentage_access_rate_cut (access rate) so we can say for example that 958 programs have a very low acceptance rate.

Let's draw up a table of line-profiles (percentage of admitted students with honors / access rate) and comment it.

```
addmargins(prop.table(addmargins(TC_emp,1,FUN=Tout),1),2,FUN=Total)
```

```
##                            percentage_admitted_honors_cut
## percentage_access_rate_cut    (0,20]   (20,40]   (40,60]   (60,80]  (80,100]
##                   (0,20]   0.4718163 0.2306889 0.1691023 0.1137787 0.0146138
##                   (20,40]  0.7162322 0.1386256 0.0764218 0.0616114 0.0071090
##                   (40,60]  0.8102345 0.1151386 0.0568586 0.0170576 0.0007107
##                   (60,80]  0.8050209 0.1514644 0.0384937 0.0016736 0.0033473
##                   (80,100] 0.8778468 0.1055901 0.0134576 0.0005176 0.0025880
##                   Tout     0.7603064 0.1395543 0.0616992 0.0334262 0.0050139
##                            percentage_admitted_honors_cut
## percentage_access_rate_cut     Total
##                   (0,20]   1.0000000
##                   (20,40]  1.0000000
##                   (40,60]  1.0000000
##                   (60,80]  1.0000000
##                   (80,100] 1.0000000
##                   Tout     1.0000000
```

```
#or another way to do it is:
TPL <- addmargins(prop.table(addmargins(TC_emp,1),1),2)
colnames(TPL)<-c('very low','low','medium','high','very high','Total')
rownames(TPL)<-c('very low','low','medium','high','very high','Tout')
TPL
```

```
##                            percentage_admitted_honors_cut
## percentage_access_rate_cut  very low       low    medium      high very high
##                  very low  0.4718163 0.2306889 0.1691023 0.1137787 0.0146138
##                  low       0.7162322 0.1386256 0.0764218 0.0616114 0.0071090
##                  medium    0.8102345 0.1151386 0.0568586 0.0170576 0.0007107
##                  high      0.8050209 0.1514644 0.0384937 0.0016736 0.0033473
##                  very high 0.8778468 0.1055901 0.0134576 0.0005176 0.0025880
##                  Tout      0.7603064 0.1395543 0.0616992 0.0334262 0.0050139
##                            percentage_admitted_honors_cut
## percentage_access_rate_cut     Total
##                  very low  1.0000000
##                  low       1.0000000
##                  medium    1.0000000
##                  high      1.0000000
##                  very high 1.0000000
##                  Tout      1.0000000
```

The last line represents the marginal distribution of the percentage of admitted students with honors independently from the program's acceptance rate.

76% of programs have a very low percentage of admitted students with honors. 14% of programs have a low percentage of admitted students with honors. 6% of programs have a medium percentage of admitted students with honors. 3% of programs have a high percentage of admitted students with honors. 1% of programs have a very high percentage of admitted students with honors.

The conditional distributions of programs admitting students with honors conditional to the acceptance rate seem to differ from one another and also differ from the marginal distribution of programs admitting students with honors. For example, we have 47% of programs admitting a very low percentage of students who graduates with honors in the sample of programs with a very low acceptance rate. However, we have 76% of programs admitting a very low percentage of students in our whole sample.

It seems that the two variables are not independent but we will test this hypothesis rigorously later on.

Now, Let's draw up a table of column-profiles (access rate/ percentage of admitted students with honors) and comment it.

```
TPC <- addmargins(prop.table(addmargins(TC_emp,2),2),1)
```

First we add a column in the end that will display the marginal distribution of the programs' acceptance rates. Then with prop.table we display all of the table's elements as relative frequencies. With addmargins we add a line in the end that will sum up the table's elements in each column. The last column is the marginal distribution of the access independently from the percentage of students who graduated with honors and were admitted to the program. In the last line, we calculate the sum of the elements in each column, so it is indeed a Total.

```
colnames(TPC)<-c('very low','low','medium','high','very high','Tout')
rownames(TPC)<-c('very low','low','medium','high','very high','Total')
TPC
```

```
##                            percentage_admitted_honors_cut
## percentage_access_rate_cut very low      low   medium     high very high
##                  very low  0.082799 0.220559 0.365688 0.454167  0.388889
##                  low       0.221469 0.233533 0.291196 0.433333  0.333333
##                  medium    0.208829 0.161677 0.180587 0.100000  0.027778
##                  high      0.176223 0.180639 0.103837 0.008333  0.111111
##                  very high 0.310680 0.203593 0.058691 0.004167  0.138889
##                  Total     1.000000 1.000000 1.000000 1.000000  1.000000
##                            percentage_admitted_honors_cut
## percentage_access_rate_cut     Tout
##                  very low  0.133426
##                  low       0.235097
##                  medium    0.195961
##                  high      0.166435
##                  very high 0.269081
##                  Total     1.000000
```

We read in every column the conditional distributions of the variable access rate by group of programs admitting a certain percentage of students who graduated with honors, for example, in column one, by group of programs admitting a very low percentage of students who graduated with honors. In the group of programs admitting a very low percentage of students who graduated with honors, we have 8.28% of programs with very low access rates, 22,15% of them with low access rates, 20.88% of them with mid access rates, 31.07% of them with high access rates and 31.07% of them with very high access rates.

We observe that the column-profiles are relatively different. The distributions of the access rates are quite dissimilar in every subset and in the whole sample.

It seems that the 2 variables (access rate and percentage of admitted students with honors) are not independent, let’s verify this statement by conduction a chi-square independence test.

```
resu <- chisq.test(percentage_admitted_honors_cut, percentage_access_rate_cut, correct=FALSE)
```

```
## Warning in chisq.test(percentage_admitted_honors_cut,
## percentage_access_rate_cut, : Chi-squared approximation may be incorrect
```

```
#(correct=FALSE parce qu'on ne fait pas de correction de continuité)
resu
```

```
##
##  Pearson’s Chi-squared test
##
## data:  percentage_admitted_honors_cut and percentage_access_rate_cut
## X-squared = 863, df = 16, p-value <2e-16
```

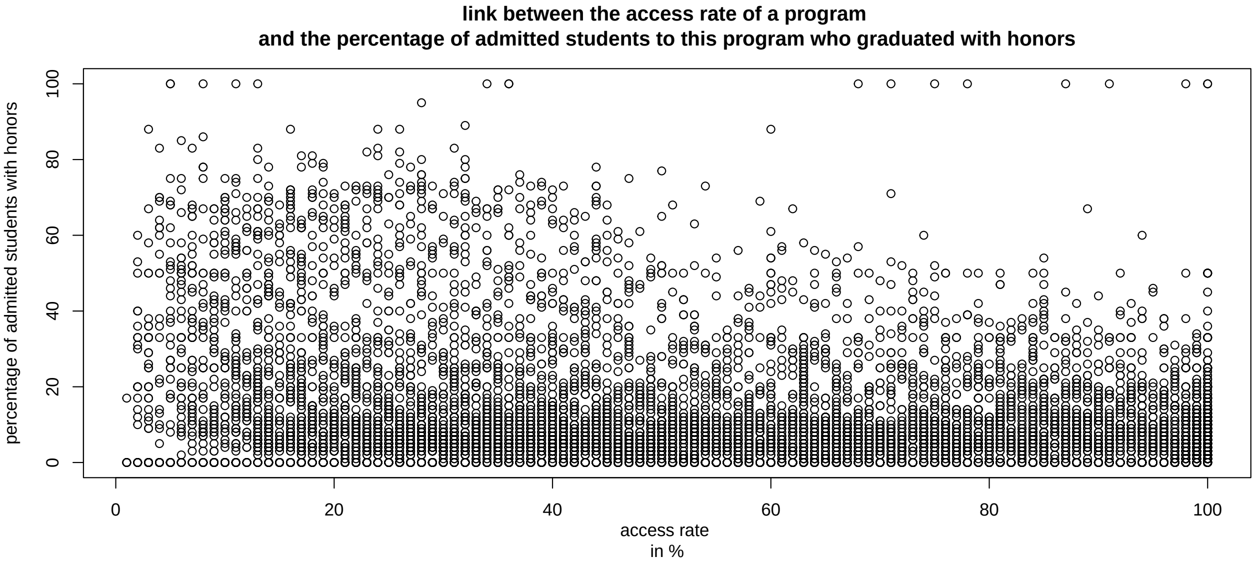
```
resu$expected #joint frequencies > 4 so we can use the chi-square law in the test.
```

```
##                                percentage_access_rate_cut
## percentage_admitted_honors_cut  (0,20]  (20,40]  (40,60] (60,80] (80,100]
##                         (0,20]   728.374 1283.397 1069.751 908.566 1468.912
##                         (20,40]  133.693  235.568  196.353 166.767  269.619
##                         (40,60]   59.108  104.148   86.811  73.731  119.203
##                         (60,80]   32.022   56.423   47.031  39.944   64.579
##                         (80,100]   4.803    8.464    7.055   5.992    9.687
```

Critical value < 5% so it is necessarily <10% which leads us to reject the null hypothesis that claims that the two variables are independent. These results confirm our intuition based on the observation of the line-profiles and column-profiles tables : the access rate to a higher education program influences the percentage of admitted graduates with highest honors. So we showed that these variables are not independent at a 5% nor 10% significance level.

Scatter plot of the 2 variables :

```
plot(percentage_access_rate,percentage_admitted_honors,main='link between the access rate of a program
and the percentage of admitted students to this program who graduated with honors',xlab='access rate
in %',ylab='percentage of admitted students with honors')
```



Each point represents a program with in abscissa the access rate and in ordinate the percentage of admitted students who graduated with honors.

We see a slight decreasing trend. Let's see if the function cor gives us the same results.

```
# Create a new data frame without missing values
data <- data.frame(percentage_access_rate, percentage_admitted_honors)
data <- na.omit(data)
# Calculate correlation on the new data frame
correlation <- cor(data$percentage_access_rate, data$percentage_admitted_honors)
correlation
```

```
## [1] -0.2986
```

Which confirms that there is a weak negative correlation between the access rate and the percentage of admitted students who graduated with honors.

Programs with low access rates receive a significant number of applications compared to the places they offer, and tend to admit more students who graduated with honors. This favors merit over equality, and it raises questions : on one hand, elite private high schools and some elite public schools tend to have very high rates of honor graduates, offering great learning environments which are conducive to students' high achievements. On the other hands, ZEP (zone d'éducation prioritaire) schools historically achieve lower rates of honor graduates, and the learning environments are much less ideal, with very limited resources. By admitting more honor graduates to low access rate programs, we create social reproduction which leads to more inequalities in access to higher education. Some greats students with very high potential who struggled during high school due to their life circumstances could thrive in higher education if they are in the right environment. However, even though there is a negative correlation between the access rate and the percentage of admitted students who graduated with honors, it is relatively weak (-0.3 with correlation coefficient ranging from -1 to 1).

#Conclusion

In this R project for our statistics class, we analysed the inequalities in access to higher education in France using our database extracted from the Ministry of Education's website. The database we used includes 13 variables, observed on a sample of 13644 programs offered on Parcoursup in year 2022.

We found that the percentage of admitted girls to higher education programs hasn't reached 50% yet, but it is between 48.89% and 49.88% for a 95% confidence level, so the government has almost reached perfect equality and therefore it must pursue efforts towards gender equality in higher education.

Regarding financial inequalities, we have found that respectively at a 5% then a 10% significance level, on average less than 30% of candidates receiving a CROUS scholarship were offered admission to the programs in our data base, so the government hasn't achieved equality in access to education based on income yet. However, it is crucial to highlight that France is a developed country offering excellent higher education programs for a significantly lower cost than its other European and international counterparts. All in all, the government needs to reverse the backwards trend in the recent years and reinforce the existing quotas of scholarship-recipients in higher education programs.

Finally, we also found that there is a weak negative correlation between access rate and the percentage of students who graduated with honors. It isn't a major issue since the correlation between the two variables is weak and we do realize that selective programs can only select students fairly based on their academic achievements so it is no surprise that they would pick more students with "mention Très Bien" at their Baccalaureate exam. However, it would be interesting to implement more programs like the one called "Egalité des chances" at Dauphine to avoid social reproduction, recruit a more diverse set of students from different geographical locations and social classes, and to integrate bright students who weren't offered the same opportunities to thrive in high school.

As a closing statement, we would like to say that France has gone a long way to minimize inequalities in access to higher education. Meritocracy is a key element in the higher education system in France and while some inequalities in access to higher education have been considerably reduced, like gender inequalities, others still need to be worked on, such as the percentage of scholarship-recipients by program

Thank you for your time.

Emile SABEH AFFAKI and Antoine MONIZ