CY Tech Sciences et Techniques

# Intermediate Statistics

*Author*
Anh Thu DOAN

January 5, 2022

# Contents



# List of Figures



# List of Tables

# 1 Introduction

The development of technology has been changing training models and influencing the strategy of educational institutions. MOOC (Massive Open Online Course) is an evolving form of distance education. The rapid growth of MOOCs in recent years has made learning easy for anyone, anywhere, and for free. One factor that sets MOOCs apart from traditional distance learning courses is that subscribers can be up to thousands of learners. There are often no limits or conditions for attendance and registration fees.

Learner engagement when participating in MOOCs is a deep concern. There are many factors that influence learner motivation, including future economic benefits, personal and professional identity development, challenges and achievements, and pleasure. Using the Moocs material such as quizzes, assignments, the number of videos watched, etc., to classify the learner's level of engagement. Those who obtained a certificate were called "completers", those who submitted at least one quiz or assignment but did not complete the course were referred to as "disengaging learners". And those who did not submit any quizzes or assignments were referred to as "auditing learners" if they had viewed at least 10 percent of available course videos and "bystanders" if they fell below this threshold.

# 2 Methods

The data set was used from a MOOC organized on Canvas before being on Coursera, the MOOC Affectation. This is a five-week-long entrepreneurship course called Effectuation (Professor Philippe Silberzahn, EMLYON BusinessSchool), referred to as MOOC1. It was hosted by a MOOC agency, which used the open-source LMS Canvas from Instructure.

This study was done in R with the following methods:

- Manipulating the data sets.
- Linear model.
- Logistic regression.
- Survival analysis.

# 3 Results

## 3.1 Describing behavior in the courses

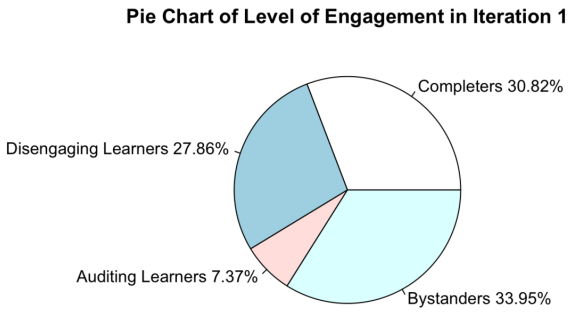


Figure 1: Pie chart of Engagement Level of the learners in Iteration 1

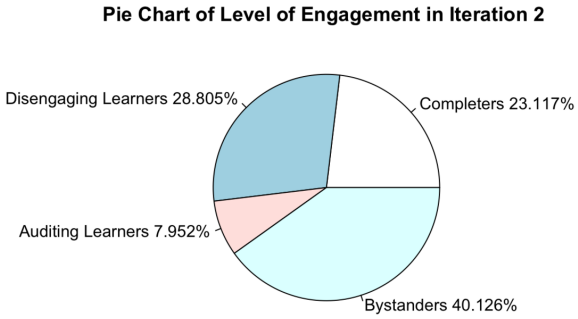


Figure 2: Pie chart of Engagement Level of the learner in Iteration 2

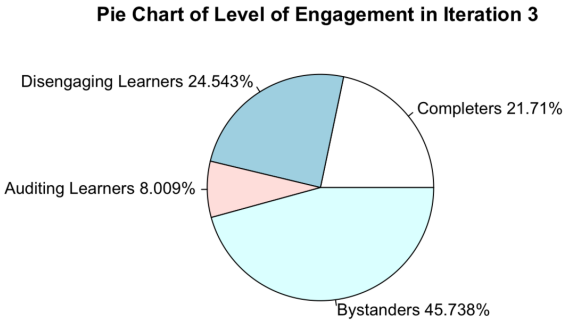


Figure 3: Pie chart of Engagement Level of the learner in Iteration 3

Regarding those iterations Fig.1-2-3 , we have figured two main relevant trends. Indeed, the COMPLETER's trend acknowledge an important decrease while the BYSTANDER'S group increase. The average decrease for the COMPLETER's group is around 4,5% and 5,9% increase for the BYSTANDER's group. When we have a focus on the 2 others categories, they seem adapt stable trend for the 3 iterations. By issuing that analysis, we could assume that the MOOC's level should be more and harder for the candidates who are taking the courses. Then if we go further in our analysis, and look into some key figures among the survey's table. In fact, that may help us to understand with more accuracy, the different trends which have been observed. For instance, if we have a look onto the survey and merged those result with the usage.effect excel table. We would be able to cross those data in order to understand deeper the behavior of those data. For our analysis, as there so much input to take into account. We will propose to steer analysis on one special input. Indeed, as the Data concern the MOOC, first I want to study the influence of the diploma level on the engagement level.

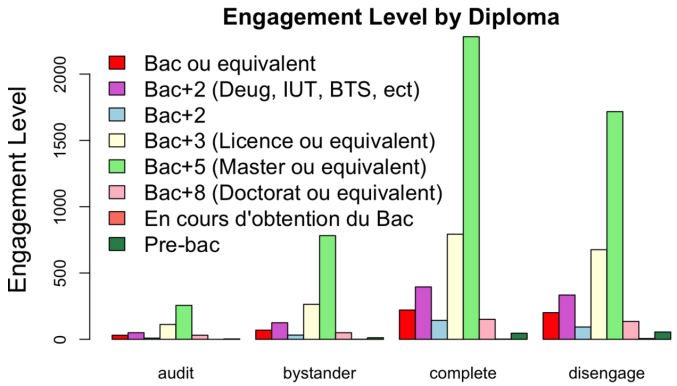


Figure 4: Bar chart describe level of engagement of the learners by Diploma

If we have a deeper look into the Fig 4, we can figure out that a one main highlight. In fact, at first sight, we can see that the major part of MOOC participants hold a Bac+5 or equivalent diploma. Furthermore, if we add input from the four categories, there are almost5000 participant over the total which is an around 30% of total participant. Then regarding the COMPLETER group, we have seen that the Bac+5 diploma acknowledge the most important spike compare to the rest of the other categories. The other key figure will be the rate of people whose are in progress to take the "Baccalaureate" exam. Actually, the most confronting view is the fact that this amount is reaching the zero level. From those findings, we could state two hypothesis. Indeed, the first one regards the master degree diploma. The reason of the important participation of master degree graduated, is the fact we do assume that although they have finish their study, they are constantly in research of new knowledge or need to enhance their skills. Furthermore, as they are freshly graduated, they mind is enough sharp in order to manage the MOOC completion. The second ascertainment regards the low level of people which are categorize as "En cours d'obtention de Bac". Indeed, those people are more focus on their currents main exam which is the "Baccalaureate", then they are not willing to spend a part of their time to participate at such MOOC. These figures would be able to help us to know more the participant profile, especially when we need to readjust the level of the MOOC, in order to be in line with the participant level.

## 3.2 Linear Model

### 3.2.1 From Student's t-test to two-ways ANOVAs

| | estimate | estimate1 | estimate2 | statistic | p.value | parameter | conf.low | conf.high |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.83 | 12.66 | 11.82 | 3.23 | 0.00 | 5819.73 | 0.33 | 1.34 |

Table 1: Welch Two Sample t-test between men and women

To determine whether there is a significant difference between the number of views of videos and genders, we used t-test to find out about it. The result is shown in the Table 1 the p-value of Welch Two Simple t-test is 0.00126 $< 0.05$ implying that the distribution of the data are significantly different. The test also gives us the sample mean in group men was 11.82440 and the sample mean of group woman was 12.65747.

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| HDI | 2 | 655808.90 | 327904.45 | 5101.65 | 0.0000 |
| Residuals | 21274 | 1367370.37 | 64.27 | | |

Table 2: One-Way ANOVA between totals number of views of video vs HDI

In the Table 2, we used One-Way ANOVA test compare the number of views of videos depending on the HDI of the country of origin. From the output, we could have seen that HDI is significant in the 2-degrees of freedom test. However, this will require three tests (TH vs. B, TH vs. I , B vs. I), so we adjust what we consider to be statistically significant to account for this multiplicity of tests.

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Gender | 1 | 840.12 | 840.12 | 6.47 | 0.0110 |
| HDI | 2 | 31896.86 | 15948.43 | 122.82 | 0.0000 |
| Gender:HDI | 2 | 372.94 | 186.47 | 1.44 | 0.2379 |
| Residuals | 8179 | 1062079.74 | 129.85 | | |

Table 3: ANOVA table of number of viewed videos with Gender and HDI parameter

To evaluate simultaneously the effect of two grouping variables (A and B) on a response variable we used two-way ANOVA test. From the Table 3 we can conclude that both Gender and HDI are statistically significant, as well as their interaction

### 3.2.2 Model refinement, pairwise comparisons

Add an interaction parameter Gender*HDI to linear model, then summary it to see the interaction of the new parameter as in Table 4.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 7.7391 | 1.1880 | 6.51 | 0.0000 |
| Genderun homme | -2.2974 | 1.3026 | -1.76 | 0.0778 |
| HDII | 1.7185 | 1.4646 | 1.17 | 0.2407 |
| HDITH | 5.7146 | 1.2098 | 4.72 | 0.0000 |
| Genderun homme:HDII | 2.7496 | 1.6757 | 1.64 | 0.1009 |
| Genderun homme:HDITH | 2.0289 | 1.3331 | 1.52 | 0.1280 |

Table 4: Summary Linear model between total views video and Gender*HDI

Then we used stepwise algorithm to assess the performance of various version of the model (forward and backward).Assess the colinearity of all three independant variables of the last model by using a chi-test between HDI and Gender (Table 5). In this table, Pearson's Chi-quare test was use to analyze the correlation between the student's gender and their HDI country level. As we can see from the table, those variables are statistically significantly associated (p-value = 0).

| | statistic | p.value | parameter | method |
|---|---|---|---|---|
| 1 | 74.74 | 0.00 | 2 | Pearson's Chi-squared test |

Table 5: Pearson's Chi-squared test between Gender and HDI

In ANOVA test, a significant p-value indicates that some of the group means are different, but we don't know which pairs of groups are different. To perform that we can use Tukey HSD (Tukey Honest Significant Differences, R function: TukeyHSD()) for performing multiple pairwise-comparison between the means of groups.

| | term | contrast | null.value | estimate | conf.low | conf.high | adj.p.value |
|---|---|---|---|---|---|---|---|
| 1 | HDI | I-B | 0.00 | 8.24 | 7.40 | 9.07 | 0.00 |
| 2 | HDI | TH-B | 0.00 | 11.76 | 11.48 | 12.03 | 0.00 |
| 3 | HDI | TH-I | 0.00 | 3.52 | 2.67 | 4.37 | 0.00 |

Table 6: Tukey multiple comparisons of means

It can be seen from the output (Table 6), that all pairwise comparisons are significant with an adjusted p-value $< 0.05$.

### 3.3 Logistic Regression

By using logistic regression model with "binomial" family, and it does not return directly the class of observations. It allows us to estimate the probability (p) of class membership. We used variable Exam.bin in order to know whether the learner completed the course or not.

| Characteristic | OR[1] | 95% CI[1] | p-value |
|---|---|---|---|
| HDI | | | |
| B | — | — | |
| I | 1.83 | 1.28, 2.62 | <0.001 |
| TH | 1.99 | 1.51, 2.66 | <0.001 |
| Gender | | | |
| un homme | — | — | |
| une femme | 1.14 | 1.01, 1.28 | 0.027 |

[1] OR = Odds Ratio, CI = Confidence Interval

Table 7: Logistic regression table (Gender and HDI)

From the Table 7 and Fig 5, the impact of Gender different on the course's completion was demonstrated. Moreover, it is shown how socioeconomic status impact on the completion of the MOOCs. The higher HDI, the higher would be in the factor. Which means that the more developed of the countries there are higher chance of completing the course of students.

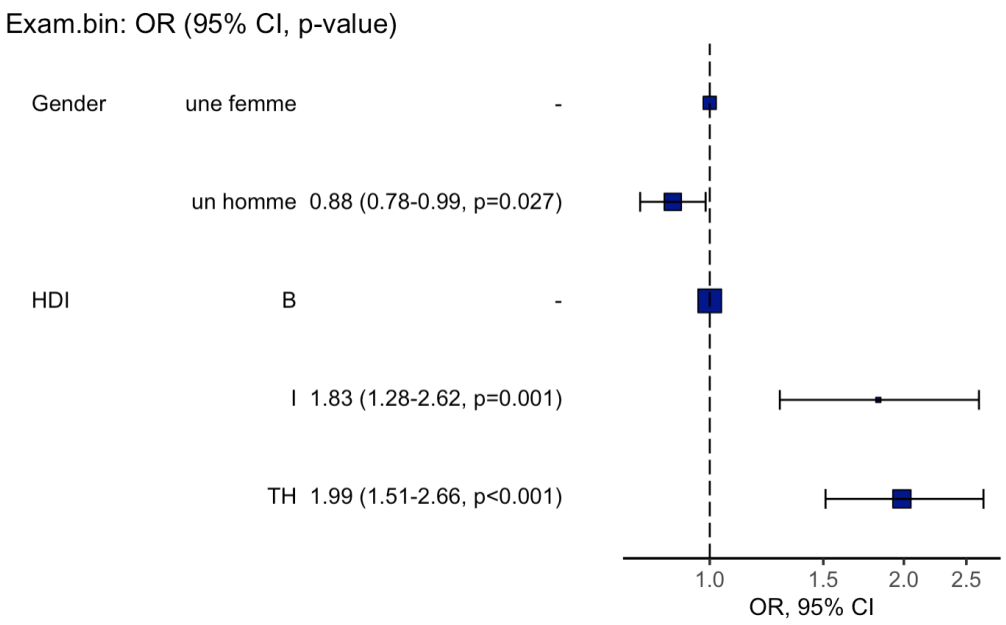


Figure 5: Odd Ratios of MOOC completion by Gender and HDI

| Characteristic | OR[1] | 95% CI[1] | p-value |
|---|---|---|---|
| CSP.fin | | | |
| Artisans, commer&#65533;ants, chefs d'entreprise | — | — | |
| Autre | 0.22 | 0.05, 1.14 | 0.066 |
| Cadres et professions intellectuelles | 0.54 | 0.17, 1.76 | 0.3 |
| CSP.finEmploy<e9>s | 0.26 | 0.06, 1.16 | 0.074 |
| En recherche d'emploi | 1.67 | 0.51, 5.94 | 0.4 |
| Etudiants | 1.09 | 0.35, 3.68 | 0.9 |
| Estimated.hours | | | |
| De 1 &#65533; 2 heures | — | — | |
| Estimated.hoursDe 1 <e0><a0> 2 heures | 5,068 | 1,987, 15,024 | <0.001 |
| Estimated.hoursDe 2 <e0> 4 heures | 0.64 | 0.17, 1.92 | 0.4 |
| Estimated.hoursDe 2 <e0><a0> 4 heures | 9,754 | 2,876, 47,055 | <0.001 |
| Estimated.hoursDe 30 minutes <e0> 1 heure | 0.53 | 0.08, 2.01 | 0.4 |
| Estimated.hoursDe 30 minutes <e0><a0> 1 heure | 1,500 | 572, 4,549 | <0.001 |
| Estimated.hoursDe 4 <e0> 8 heures | 1.86 | 0.28, 7.37 | 0.4 |
| Estimated.hoursDe 4 <e0><a0> 8 heures | 3,285 | 570, 63,372 | <0.001 |
| Moins de 30 minutes | 18.0 | 6.84, 46.9 | <0.001 |
| Plus de 8 heures | 21.1 | 5.25, 72.2 | <0.001 |

[1] OR = Odds Ratio, CI = Confidence Interval

Table 8: Logistic regression table (Jobs, Hours estimated)

To study more about how ocioeconomic status impact on completing the MOOCs, we made another table (Table 8) comparing the job of learner and how many hours they could spend. From the result, people who are finding for a job and students have "positive relationship" with completing the MOOCs. It can be explain by the fact that they want the knowledge to seeking for the job.

## 3.4 Survival Analysis

In this part we used Survival analysis on video consumption with HDI group of country and two level behavior of learners to investigate the time it takes for an event of interest to occur. In Fig 6, we compared video consumption behavior between HDI groups. As easy to see that the median survival time of Very High (TH) HDI and Intermediate (I) was 8 videos, and Low (B) was 4 videos. We can concluded that the the Low HDI had the most significant drop in video consumption and the higher the HDI, the more likely the participant is to survive.

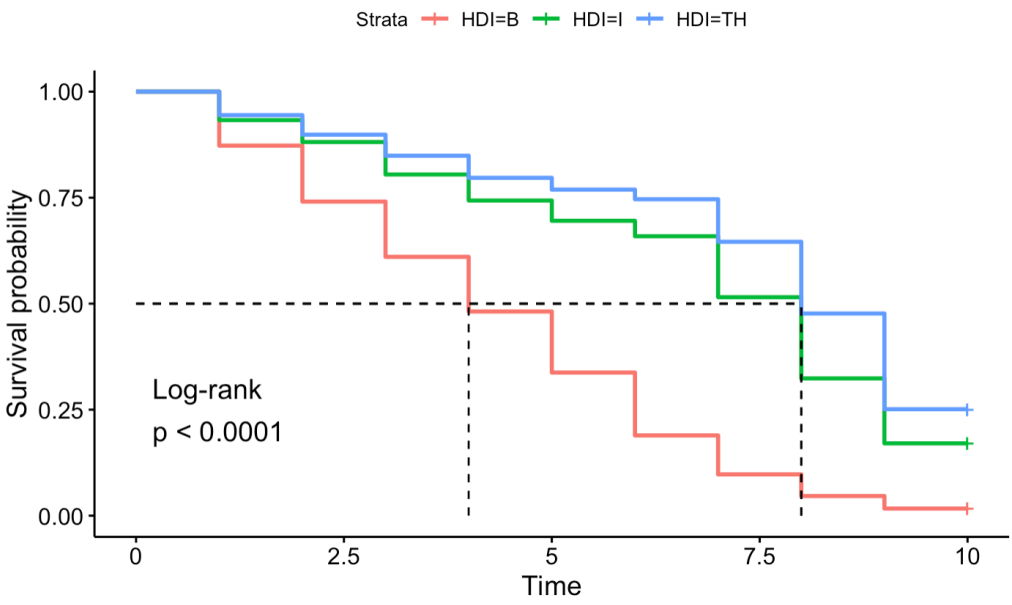


Figure 6: Survival plot between HDI group of country

In the Fig 7, we did the same with two level of Engagement learner. Consequently, it had surprisingly result when Disengaging and Auditing have exactly the same median survival which was around 7 videos.

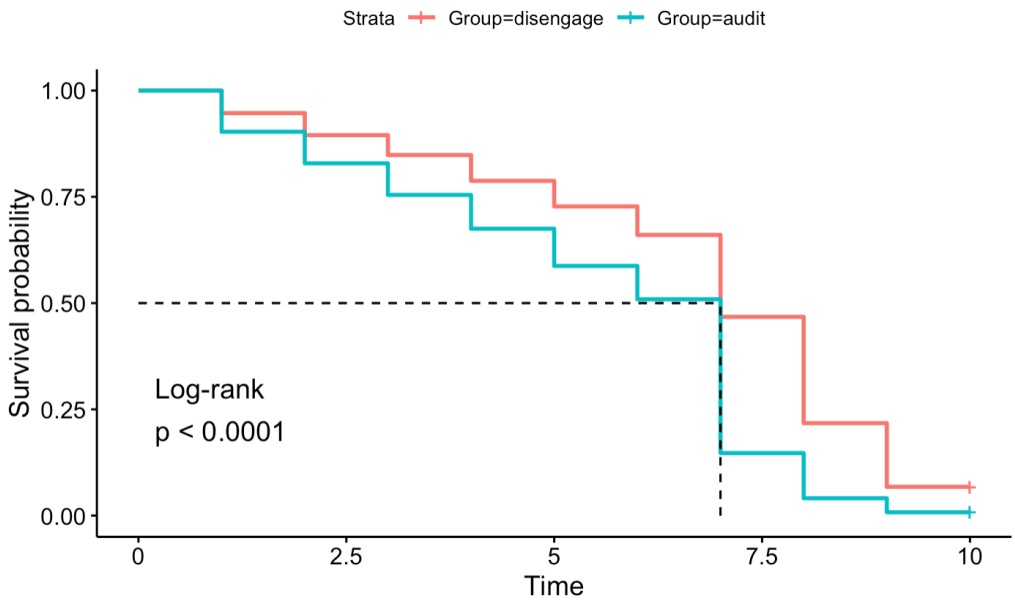


Figure 7: Survival plot between engagement level of learner

# 4 Discussion

Formed and developed over the past decade, MOOC is considered a phenomenon and a new trend in modern education. It has left many profound imprints in the development of the Education industry, with the criterion of bringing quality open knowledge to many learners. With the great benefits brought by MOOCs, a series of large systems have been born not only in MOOC's homeland, the United States but also spread to other countries around the world. Although predicted as a leap in teaching technology, MOOC itself still has many inherent limitations that have not been overcome. As we have studied above, there are many factors impact on the engagement level of learners, from gender to HDI level of country to their background, etc. Even we found out there was a lot of connection between those variables, the impact of characteristics such as socioeconomic background or country of residence on completion rates was discovered primarily in the management course's most challenging track.