



---

## Introduction aux statistiques

Diplôme Universitaire Data Analyst enseigné  
par **Dr Matthieu Cisel**

---

**Version 3.0**

8 juillet 2022

---

Geneviève Gleizes

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Description du jeu de données</b>	<b>3</b>
2.1	Types d'apprenants . . . . .	3
2.2	Analyse . . . . .	3
2.3	Interprétation . . . . .	3
<b>3</b>	<b>Variables Genre et IDH : Tests d'indépendance de Chi2 et mosaïc plot</b>	<b>4</b>
3.1	Test d'indépendance de Chi2 . . . . .	4
3.2	Mosaïc plot . . . . .	5
3.3	Interprétation . . . . .	5
<b>4</b>	<b>Modèle linéaire, tests non paramétriques</b>	<b>6</b>
4.1	Nombre de vidéos vues par genre en fonction de l'IDH. Les marqueurs "+" représentent la moyenne. . . . .	6
4.2	Interprétation . . . . .	6
4.3	corrélation entre nombre de vidéos vues et nombre de quizz réalisés . . . . .	6
4.4	Analyse . . . . .	7
4.5	Interprétation . . . . .	8
4.6	Effet de l'IDH et du genre sur le nombre de vidéos vues . . . . .	8
4.7	Degré de liberté des variables IDH et Genre . . . . .	8
4.8	Effet des modalités simples d'IDH et du genre sur le nombre de vidéos vues . . . . .	8
4.9	Nombre de vidéos vues et variables socio-démographiques, effets associés aux modalités croisées . . . . .	9
<b>5</b>	<b>Régression logistique</b>	<b>10</b>
5.1	Odds Ratio . . . . .	10
5.2	Analyse et interprétation . . . . .	10
5.3	Risques Relatifs et Odds Ratio . . . . .	11
5.4	Données de comptage et loi de Poisson . . . . .	11
<b>6</b>	<b>Références</b>	<b>13</b>

## Liste des figures

1	Mosaïc plot du test de Chi2 sur les deux variables socio-démographiques IDH et genre . . . . .	5
2	Distribution de la variable nombre de vidéos vues par genre . . . . .	6
3	Répartition du nombre de quizz réalisés en fonction du nombre de vidéos vues . . . . .	7
4	Répartition du nombre de quizz réalisés en fonction du nombre de vidéos vues avec distinction des points, en rouge : droite de régression linéaire . . . . .	7
5	Obtention de l'examen final et variables socio-démographiques, forest plot des odds ratio, CI = Intervalles de Confiance . . . . .	10
6	Tests graphiques de normalité . . . . .	12
7	Répartition de l'effectif des apprenant.e.s en fonction du nombre de vidéos vues . . . . .	13

## Liste des tableaux

1	Pourcentage de chaque type d'apprenant.e pour les trois itérations du MOOC . . . . .	3
2	Résultat du test d'indépendance de Chi2 sur les deux variables socio-démographiques IDH et genre . . . . .	4
3	Nombre de vidéos vues et variables socio-démographiques, table d'ANOVA . . . . .	8
4	Nombre de vidéos vues et variables socio-démographiques (modalités simples), table d'ANOVA	9
5	Nombre de vidéos vues et variables socio-démographiques (modalités croisées), table d'ANOVA	9
6	Obtention de l'examen final et variables socio-démographiques, régression logistique table des odds ratios . . . . .	10
7	Comparaison Odds Ratio et Risque Relatif de l'effet du genre sur l'obtention de l'examen final	11
8	Corrélation entre le nombre de quizz réalisés et le nombre de vidéos vues résultat de la régression de Poisson . . . . .	13

# 1 Introduction

L'expansion rapide des MOOC (Massive Open Online Course) en France depuis ces dix dernières années pose question. La nature même des MOOC favorise une grande hétérogénéité des inscrits sur de nombreux facteurs comme le pays de résidence et le comportement. Cette étude porte sur les données d'un MOOC et tente d'apporter quelques réponses sur l'influence des facteurs socio-démographiques de l'apprenant.e.

## 2 Description du jeu de données

### 2.1 Types d'apprenants

Le jeu de données du MOOC comporte des données de questionnaire des apprenant.e.s et des données de log. Ces données portent sur trois itérations du MOOC et sont croisées pour l'analyse. Les apprenant.e.s sont classés selon leur degré d'engagement dans le MOOC. Les "Completers" ont obtenu l'examen final. Les "Disengaging learners" ont réalisé au moins un quizz ou ont soumis un devoir. À l'inverse des "Auditing learners" et des "Bystanders" qui ont visualisé respectivement au moins six vidéos et strictement moins de six vidéos.

Type d'apprenant	Itération n° 1	Itération n° 2	Itération n° 3
Bystander	39,4	44,8	50,0
Auditing learner	1,9	2,9	2,5
Disengaging learner	27,6	25,4	21,8
Completer	31,0	26,9	25,6
<b>Nombre d'apprenants</b>	<b>N = 7965</b>	<b>N = 3702</b>	<b>N = 3515</b>

TABLEAU 1 – Pourcentage de chaque type d'apprenant.e pour les trois itérations du MOOC

### 2.2 Analyse

Le Tableau 1 met en évidence une diminution du nombre d'apprenant.e entre la première itération (7965 apprenants) et les suivantes : deuxième (3702 apprenants) et troisième (3515 apprenants). Le pourcentage d'"Auditing learners" est compris entre 1,9% (iteration 1) et 2,9% (itération 2). Pour les "Completers" et les "Disengaging learners", le pourcentage diminue légèrement entre l'itération 1 et l'itération 3 passant respectivement de 31% à 25,9% et de 27,6% à 21,8%. Le nombre de "Bystanders" augmente progressivement et passe de 39,4% en itération 1 à 50% en itération 3.

### 2.3 Interprétation

On peut penser que la plupart des apprenants intéressés par le MOOC, le suivent dès la première itération. Ce qui entraîne une diminution du nombre de personnes qui vont s'inscrire aux deuxième et troisième itérations. De plus, l'intérêt de ces personnes va conduire à un meilleur taux de "Completers" en itération 1. Ce taux diminue ensuite à l'inverse du taux de "Bystanders" qui va augmenter au fur et à mesure des itérations du MOOC.

### 3 Variables Genre et IDH : Tests d'indépendance de Chi2 et mosaïc plot

#### 3.1 Test d'indépendance de Chi2

La variable IDH représente l'Index de Développement Humain du pays de résidence de l'apprenant.e. On part de l'hypothèse H0 selon laquelle les deux variables "Genre" et "IDH" sont indépendantes. Un test de chi2 va permettre de vérifier la validité de cette hypothèse. Le risque est fixé à 0,01.

	IDH bas		IDH intermédiaire		IDH très haut	
	Femme	Homme	Femme	Homme	Femme	Homme
Nombre d'apprenant.e.s (valeur observée)	147	883	233	432	2545	4711
Valeur attendue selon H0	336	693	217	447	2371	4884
Écart test de Chi2 <small>(carré de la différence / valeur attendue)</small>	106	52	1,17	0,03	12,76	6,1

TABLEAU 2 – Résultat du test d'indépendance de Chi2 sur les deux variables socio-démographiques IDH et genre

Il y a un lien statistiquement significatif entre le "Genre" et l'"IDH" ( $\text{Chi}^2=179$ ,  $p\text{-value}=1,19 e^{-39}$ , degré de liberté = 2). L'hypothèse H0 est rejetée car la p-value est inférieure à 0,01. Elle est même inférieure à 0,001. Les variables "IDH" et "Genre" sont dépendantes. La Figure 1 permet de visualiser le degré de dépendance des variables.

### 3.2 Mosaïc plot

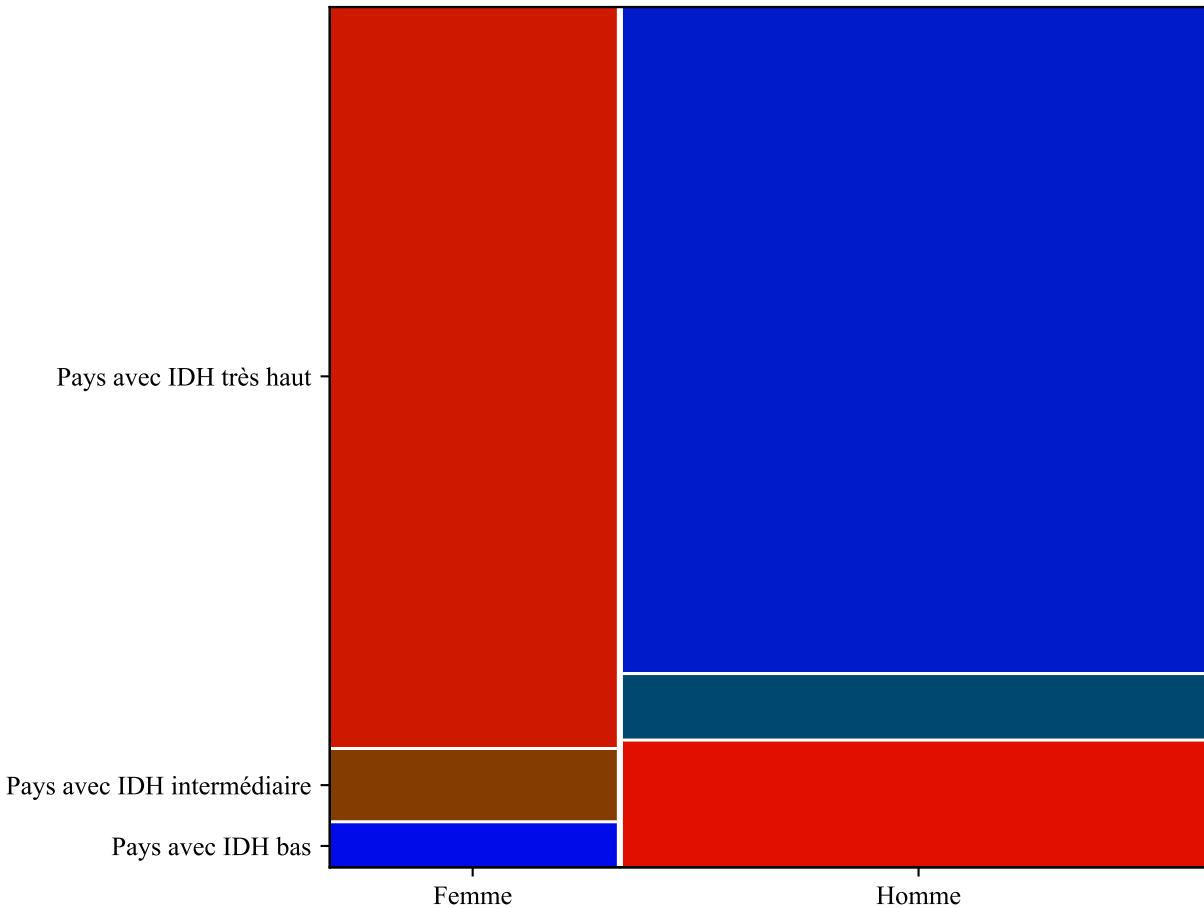


FIGURE 1 – Mosaïc plot du test de Chi2 sur les deux variables socio-démographiques IDH et genre

La Figure 1 nous informe sur le degré de dépendance entre les variables "IDH" et "Genre". Les cases rouges représentent les modalités surreprésentées. Les apprenantes femmes résidant dans un pays avec IDH très haut et les hommes résidant dans un pays avec un IDH bas sont dans ce cas. Les cases bleues représentent les modalités sous représentées. Les hommes résidant dans un pays avec un IDH très élevé et les femmes résidant dans un pays avec un IDH très bas sont dans ce cas. À l'inverse les apprenant.e.s résidant dans un pays avec HDI intermédiaire ne sont que très légèrement en surnombre pour les femmes ( $\text{Chi}^2 = 1,17$ ) ou en sous nombre pour les hommes ( $\text{Chi}^2 = 0,03$ ). Ainsi, les couleurs ne sont pas aussi radicalement bleu ou rouge.

### 3.3 Interprétation

Les écarts les plus hauts observés (106 pour les femmes et 52 pour les hommes) concernent les apprenant.e.s résidant dans un pays avec un IDH bas. Il se pourrait que l'accès à un ordinateur et à une connexion internet soit complexe et inhabituel pour les femmes des pays avec IDH bas. Pour les hommes de ces pays, cet accès pourrait être plus aisé et offrirait à ces derniers une opportunité de pouvoir suivre une formation gratuite à distance. Lorsque les données relatives aux apprenant.e.s résidant dans un pays avec IDH bas sont retirées du jeu de données, les mêmes tests et représentations graphiques donnent des résultats différents. La p-value est à 1. Les variables sont indépendantes. Toutes les cases du mosaïc plot sont vertes.

## 4 Modèle linéaire, tests non paramétriques

### 4.1 Nombre de vidéos vues par genre en fonction de l'IDH. Les marqueurs "+" représentent la moyenne.

L'étude porte sur la différence dans le nombre de vidéos vues selon le genre de l'apprenant.e. L'hypothèse H0 est celle selon laquelle le nombre de vidéos vues ne diffère pas selon le genre.

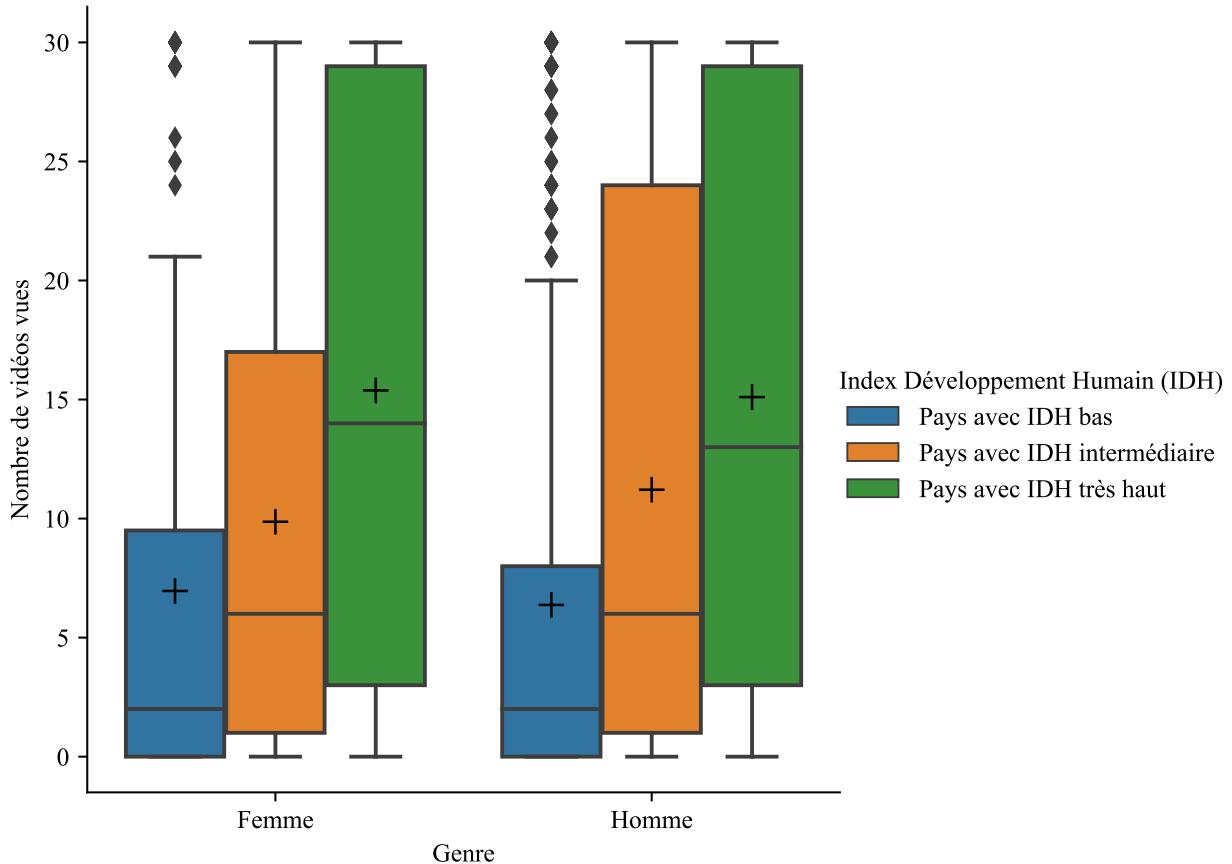


FIGURE 2 – Distribution de la variable nombre de vidéos vues par genre

La Figure 2 permet de voir que la variable nombre de vidéos vues par genre ne suit pas une distribution normale. Le test de Student ne pourra donc pas être utilisé. Le test non paramétrique de Mann-whitney est réalisé ( $U = 8711923.0$ ,  $p\text{-value} = 4,33 \times 10^{-4}$ ). Le risque est fixé à 0,01. La  $p\text{-value}$  obtenue ( $4,33 \times 10^{-4}$ ) est inférieure à 0,01. L'hypothèse H0 est rejetée. Le nombre de vidéos vues diffère selon le genre. Pour déterminer le sens de la différence, le test est réalisé à nouveau avec les options "supérieure" ( $U = 8711923$ ,  $p\text{-value} = 0,99$ ) et "inférieure" ( $U = 8711923$ ,  $p\text{-value} = 2,16 \times 10^{-4}$ ). La  $p\text{-value}$  de l'option "inférieure" ( $2,16 \times 10^{-4}$ ) est inférieure à 0,01. Le nombre de vidéos vues par les hommes est inférieur au nombre de vidéos vues par les femmes.

### 4.2 Interprétation

Les femmes ont la réputation d'être plus scolaires que les hommes. Ce caractère pourrait transparaître dans l'engagement des femmes à visionner les vidéos du MOOC et expliquer ainsi le résultat du test.

### 4.3 Corrélation entre nombre de vidéos vues et nombre de quizz réalisés

Nous nous intéressons à la corrélation entre le nombre de vidéos vues et le nombre de quizz réalisés. L'hypothèse H0 est qu'il n'existe pas de corrélation entre ces deux variables. Le risque est fixé à 0,01. La Figure 2 permet de voir que la variable nombre de vidéos vues ne suit pas une distribution normale. Ceci écarte l'utilisation du test de Pearson. Un test de Spearman est réalisé (coefficient de corrélation = 0,8,  $p\text{-value} = 0,000$ ).

#### 4.4 Analyse

La p-value est inférieure à 0,01. L'hypothèse H<sub>0</sub> est rejetée. Ce résultat montre que les deux variables sont fortement corrélées (coefficients de 0,8).

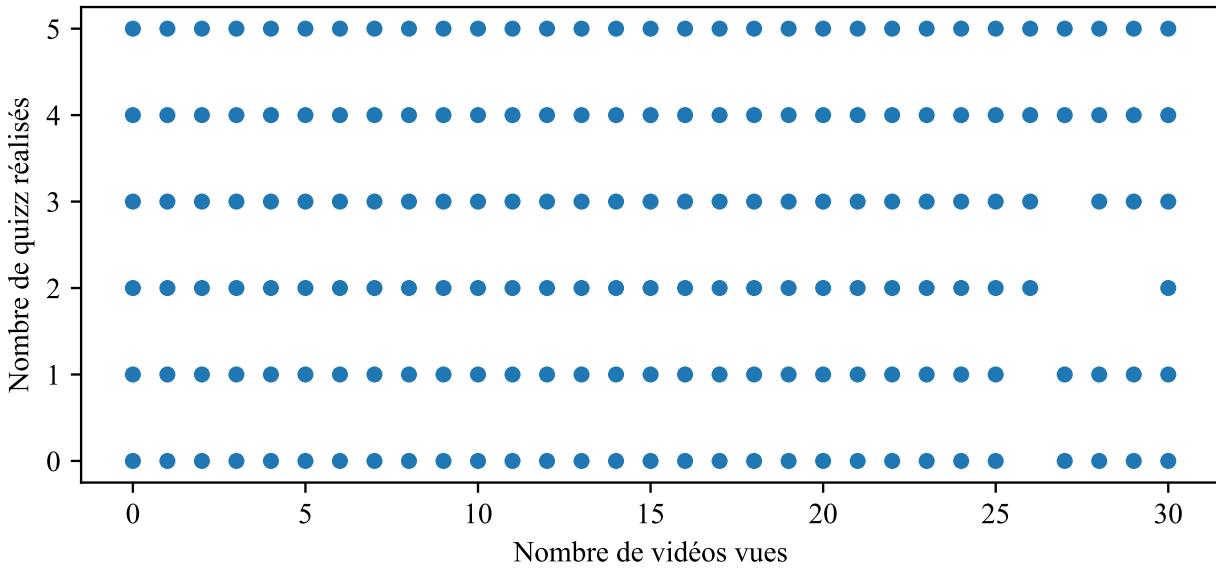


FIGURE 3 – Répartition du nombre de quizz réalisés en fonction du nombre de vidéos vues

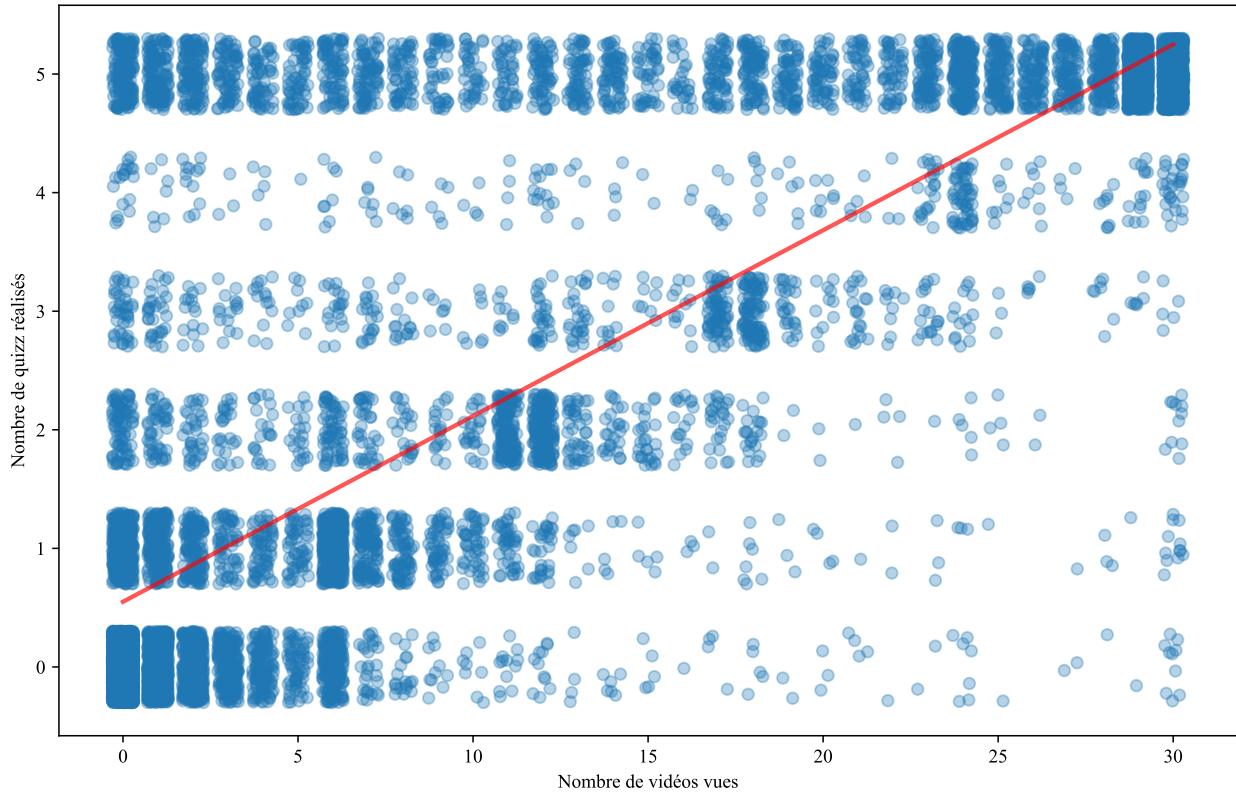


FIGURE 4 – Répartition du nombre de quizz réalisés en fonction du nombre de vidéos vues avec distinction des points, en rouge : droite de régression linéaire

Alors que la Figure 3 ne donne pas d'indication précise sur la répartition des données de comptage, la Figure 4 permet d'évaluer le volume des différentes modalités d'apprenant.e.s. Les zones de forte concentra-

tion sont observées pour les apprenant.e.s n'ayant vu aucune vidéo et réalisé aucun quizz ainsi qu'à l'extrême inverse pour les apprenant.e.s ayant vu toutes les vidéos et réalisé tous les quizz. On observe également une forte concentration d'apprenant.e.s ayant réalisé tous les quizz et n'ayant vu aucune vidéo.

## 4.5 Interprétation

La concentration importante d'apprenant.e.s ayant réalisé les cinq quizz peut s'expliquer par le côté ludique offert par les quizz qui motive sans doute de nombreux apprenant.e.s du MOOC. Il est également possible que des apprenant.e.s aient déjà les connaissances voulues et souhaitent seulement les vérifier au travers des quizz.

## 4.6 Effet de l'IDH et du genre sur le nombre de vidéos vues

Dans ce chapitre, nous étudions l'effet de plusieurs variables qualitatives sur une variable quantitative. Pour se faire, nous utilisons l'analyse de variance (ANOVA). L'hypothèse H0 est que les variables socio-démographiques IDH et Genre n'ont pas d'influence sur le nombre de vidéos vues. Le risque est fixé à 0,01.

Variable à expliquer : nombre de vidéos vues

ddl : degré de liberté

	<b>ddl</b>	<b>Somme des carrés</b>	<b>Carré moyen</b>	<b>F</b>	<b>PR(&gt;F)</b>	
<b>Genre</b>	1	1,87 e <sup>+03</sup>	1,87 e <sup>+03</sup>	14	1,50 e <sup>-04</sup>	***
<b>IDH</b>	2	7,41 e <sup>+04</sup>	3,70 e <sup>+04</sup>	284	1,40 e <sup>-120</sup>	***
<b>Résidus</b>	8947	1.16 e <sup>+06</sup>	130			

p-value <0,001 : \*\*\*, p-value <0,01 : \*\*, p-value <0,05 : \*

TABLEAU 3 – Nombre de vidéos vues et variables socio-démographiques, table d'ANOVA

Le Tableau 3 permet d'observer l'analyse de la variance. Les p-value (respectivement  $1,50 \text{ e}^{-04}$  et  $1,40 \text{ e}^{-120}$  pour le Genre et l'IDH) sont inférieures à 0,01 (et même inférieures à 0,001). L'hypothèse H0 est rejetée. Le genre et l'IDH du pays de résidence de l'apprenant.e ont chacun un effet significatif sur le nombre de vidéos vues.

## 4.7 Degré de liberté des variables IDH et Genre

L'analyse de la variance est réalisée pour chaque variable agissant séparément. Le degré de liberté représente le potentiel de variation. Il est calculé pour chaque source de variation. L'IDH peut prendre trois valeurs. Son potentiel de variation est donc de deux (trois auquel on retire la valeur de référence). De même, le genre peut prendre deux valeurs. Son potentiel de variation est donc de un.

## 4.8 Effet des modalités simples d'IDH et du genre sur le nombre de vidéos vues

L'hypothèse H0 est que les variables socio-démographiques IDH et Genre n'ont pas d'influence sur le nombre de vidéos vues. Le risque est fixé à 0,01.

Variable à expliquer : nombre de vidéos vues

Modalité de référence (Intercept) : Femme résidant dans un pays avec IDH bas

	<b>Coef</b>	<b>Err. stand.</b>	<b>t</b>	<b>P&gt; t </b>
Référence Femme - Pays avec HDI bas	6,60	0,42	15,74	0,000 ***
Genre Homme	-0,17	0,26	-0,66	0,507
Pays avec IDH intermédiaire	4,24	0,57	7,44	0,000 ***
Pays avec IDH très haut	8,70	0,38	22,68	0,000 ***

p-value <0,001 : \*\*\*, p-value <0,01 : \*\*, p-value <0,05 : \*

TABLEAU 4 – Nombre de vidéos vues et variables socio-démographiques (modalités simples), table d'ANOVA

Le Tableau 4 montre que la variation de genre n'a pas d'influence significative par rapport à la modalité de référence (intercept : Femme résidant dans un pays avec IDH bas) car la p-value (0,507) est supérieure à 0,01. En revanche, la variation de niveau de l'IDH (pays avec IDH intermédiaire ou très haut) a un impact significatif par rapport à la modalité de référence. Les p-value de ces modalités sont égales à 0.

#### 4.9 Nombre de vidéos vues et variables socio-démographiques, effets associés aux modalités croisées

L'hypothèse H0 est que les variables socio-démographiques IDH et Genre n'ont pas d'influence croisée sur le nombre de vidéos vues. Le risque est fixé à 0,01.

Variable à expliquer : nombre de vidéos vues

Modalité de référence (Intercept) : Femme résidant dans un pays avec IDH bas

En italique les lignes non exploitables du résultat (modalités simples)

	<b>Coef</b>	<b>Err. stand.</b>	<b>t</b>	<b>P&gt; t </b>
Référence Femme - Pays avec HDI bas	6,95	0,94	7,39	0,000 ***
<i>Genre Homme</i>	<i>-0,58</i>	<i>1,01</i>	<i>-0,57</i>	<i>0,56</i>
<i>Pays avec IDH intermédiaire</i>	<i>2,90</i>	<i>1,20</i>	<i>2,41</i>	<i>0,016 *</i>
<i>Pays avec IDH très haut</i>	<i>8,42</i>	<i>0,96</i>	<i>8,70</i>	<i>0,000 ***</i>
Genre Homme - Pays avec IDH intermédiaire	1,93	1,37	1,40	0,16
Gender Homme - Pays avec IDH très haut	0,30	1,05	0,29	0,77

p-value <0,001 : \*\*\*, p-value <0,01 : \*\*, p-value <0,05 : \*

TABLEAU 5 – Nombre de vidéos vues et variables socio-démographiques (modalités croisées), table d'ANOVA

Le Tableau 5 nous permet de voir que les p-value des modalités croisées Homme de Pays avec IDH intermédiaire (0,16) et Homme de Pays avec IDH très haut (0,77) sont supérieures à 0,01. Le test ne permet pas de rejeter H0. Les effets croisés des variables Genre et IDH n'ont pas d'influence significative sur le nombre de vidéos vues. L'effet de la variation de genre a diminué entre le Tableau 4 et le Tableau 5, coefficient respectivement de -0,17 et -0,58. Dans le Tableau 5, l'effet du genre est réparti sur les différentes modalités de l'analyse : modalités simples qui apparaissent en *italique* et modalités croisées (deux dernières lignes du Tableau 5). Lors de l'analyse des effets associés, les effets des modalités simples sont impactés. Le résultat des lignes relatives aux effets des modalités simples n'est plus exploitable. Les lignes en *italique* sont exclues de l'analyse.

## 5 Régression logistique

### 5.1 Odds Ratio

Nous étudions l'effet des deux variables socio-démographiques IDH et genre sur l'obtention de l'examen final. L'hypothèse H0 est que les variables IDH et genre n'ont pas d'effet sur l'obtention de l'examen final. Le risque est fixé à 0,01.

Variable à expliquer : obtention de l'examen final

Modalité de référence (Intercept) : Femme résidant dans un pays avec IDH bas

	Coef	IC 0,025	IC 0,975	p_value
Référence Femme - Pays avec IDH bas	0,19	0,15	0,23	$8,47 e^{-62}$ ***
Pays avec IDH intermédiaire	1,12	0,85	1,47	$4,15 e^{-01}$
Pays avec IDH très haut	1,37	1,14	1,65	$7,86 e^{-04}$ ***
Genre Homme	0,89	0,80	1,00	$4,72 e^{-02}$ *

p-value <0,001 : \*\*\*, p-value <0,01 : \*\*, p-value <0,05 : \*

TABLEAU 6 – Obtention de l'examen final et variables socio-démographiques, régression logistique table des odds ratios

### 5.2 Analyse et interprétation

Le Tableau 6 met en évidence des modalités ayant une influence significative sur l'obtention de l'examen final. Ainsi pour la modalité résider dans un pays avec IDH très haut ( $p\text{-value} = 7,86 e^{-04}$ ) et pour la modalité être un homme ( $p\text{-value} = 4,72 e^{-02}$ ), la p-value est inférieure à 0,01. Pour ces modalités l'hypothèse H0 est rejetée. À l'inverse, la modalité résider dans un pays avec IDH intermédiaire n'a pas d'influence significative car la p-value (0,415) est supérieure à 0,01. Résider dans un pays avec IDH très haut est un facteur favorisant l'obtention de l'examen final avec un coefficient de 1,37. Être un homme est un facteur défavorable à l'obtention de l'examen final avec un coefficient de 0,89. L'IDH haut avait déjà un effet positif et significatif sur le nombre de vidéos vues. Il en va de même pour l'obtention de l'examen final.

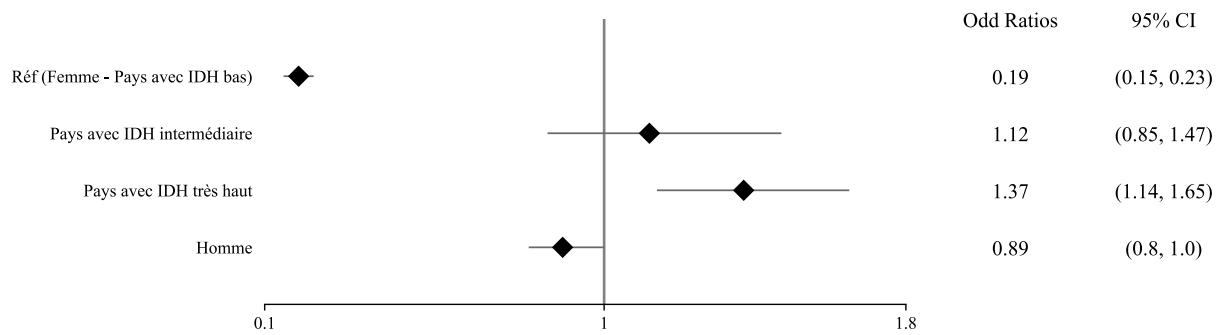


FIGURE 5 – Obtention de l'examen final et variables socio-démographiques, forest plot des odds ratio, CI = Intervalles de Confiance

La Figure 5 permet de visualiser un intervalle de confiance très faible (0,15 - 0,23) pour la modalité apprenante Femme résidant dans un pays à l'IDH bas et faible pour la modalité Homme (0,8 - 1). Les intervalles de confiance sont plus larges pour la modalité IDH très haut (1,14 - 1,65) et pour la modalité IDH intermédiaire (0,85 - 1,47).

### 5.3 Risques Relatifs et Odds Ratio

	Obtention de l'examen		Comparaison OR et RR	
	Oui	Non	Odds Ratio	Risque Relatif
Homme	N = 1073	N = 5030	ORH = 0,87 $\frac{1073}{5030}$ 588 / 2402	RRH = 0,89 $\frac{1073}{(1073 + 5030)}$ 588 / (588 + 2402)
Femme	N = 588	N = 2402	ORF = 1,14 $(= 1 / \text{ORH})$	RRF = 1,12 $(= 1 / \text{RRH})$

TABLEAU 7 – Comparaison Odds Ratio et Risque Relatif de l'effet du genre sur l'obtention de l'examen final

Comme on peut le voir dans le détail de la formule mathématique figurant dans le Tableau 7, le Risque Relatif se rapporte à l'effectif global de la variable explicative. Alors que l'Odds Ratio (rapport des chances) est simplement le rapport entre l'effectif de chaque modalité. Le sens du Risque Relatif peut être plus intuitif pour certaines personnes (Bruno Falissard 2017). L'Odds Ratio comporte plusieurs avantages : il est utilisé dans la méthode de la régression logistique et permet ainsi d'identifier les effets de chaque variable binaire sur la variable à expliquer. Nous avons pu le constater dans l'étude de l'effet du genre et de l'IDH sur l'obtention de l'examen final. De plus, de par sa nature (sans rapport avec l'effectif global de la variable explicative), il peut servir dans le cas d'une enquête témoin. Dans le cas où la modalité étudiée est rare (pour une maladie, prévalence de moins de 5%, Bruno Falissard 2017), Odds Ratio et Risque Relatif convergent. Cet effet s'explique par la formule mathématique car dans ce cas, les différences de dénominateur sont très faibles. Dans le cas de notre étude de l'effet du genre et de l'IDH sur l'obtention de l'examen, la probabilité globale d'obtenir l'examen final est autour de 22%. La différence entre Odds Ratio (0,87 pour les hommes et 1,14 pour les femmes) et Risque Relatif (0,89 pour les hommes et 1,12 pour les femmes) est faible.

### 5.4 Données de comptage et loi de Poisson

La Figure 6 regroupe quatre Figures qui permettent de tester graphiquement la normalité de la régression linéaire du nombre de quizz réalisés (la variable à expliquer) par le nombre de vidéos vues (variable explicative). Si la régression linéaire suivait une distribution normale dans la Figure 6a les points seraient dispersés de façon homogène de part et d'autre de la ligne de résidus zéro. On peut voir dans cette figure que la répartition est biaisée car les points plongent du haut à gauche vers le bas à droite. La moitié gauche de la figure comporte une majorité de résidus positifs tandis que la moitié droite de la figure comporte une majorité de résidus négatifs. On reconnaît, dans la Figure 6a et dans la Figure 6c, la répartition représentée dans la Figure 3, pente ajoutée. L'homoscédasticité est observée lorsque la variance des erreurs de la régression est la même pour chaque observation. L'homoscédasticité est nécessaire pour pouvoir appliquer une régression linéaire. La Figure 6a met en évidence une hétéroscédisticité de la régression. Concernant la Figure 6b, les points bleus devraient suivre la diagonale tracée en rouge dans le cas d'une distribution normale. On voit dans cette figure que les points s'éloignent à plusieurs endroits de la diagonale rouge. Ces éléments permettent de confirmer que les conditions de validité pour appliquer une régression linéaire ne sont pas respectées.

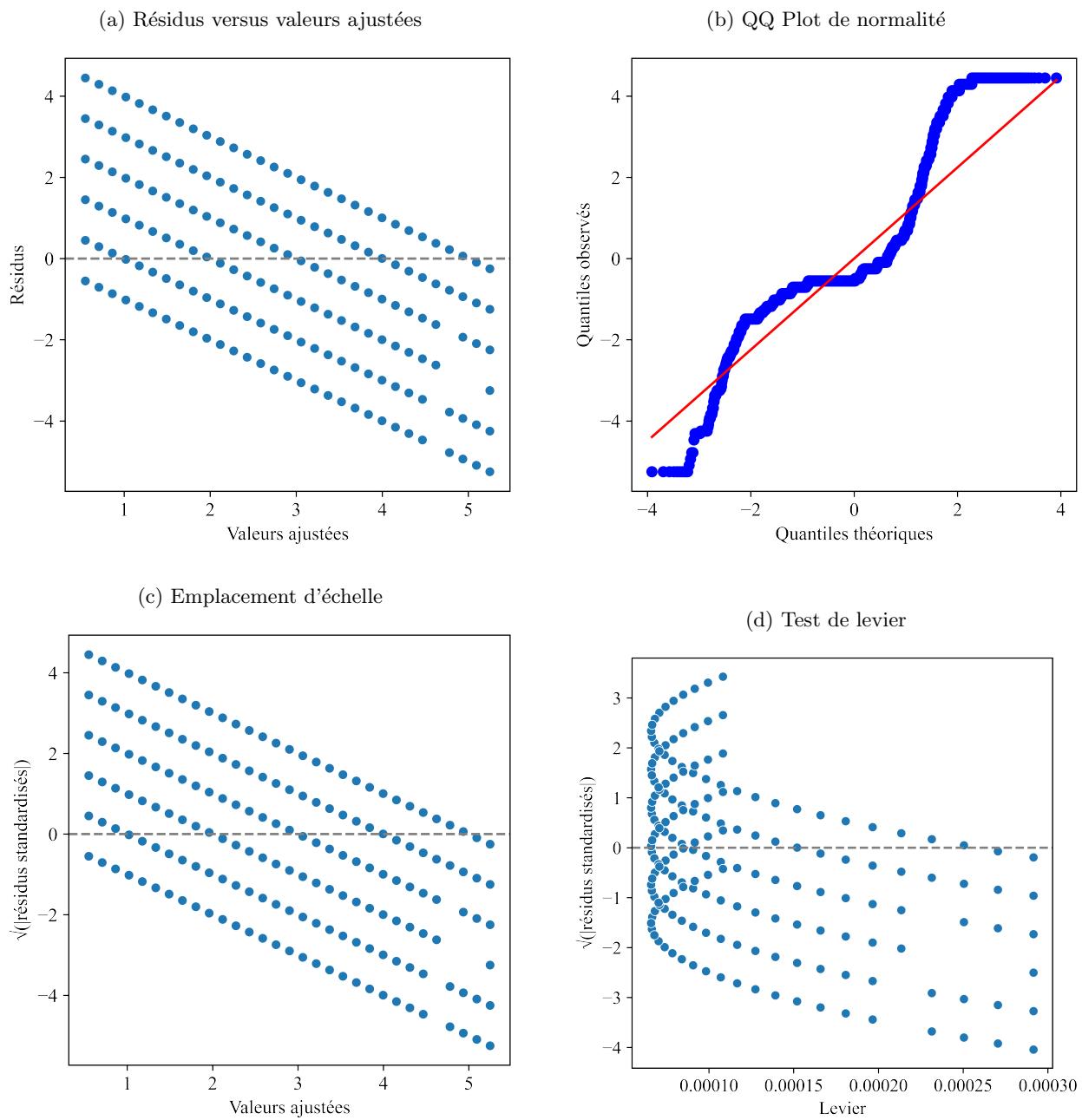


FIGURE 6 – Tests graphiques de normalité

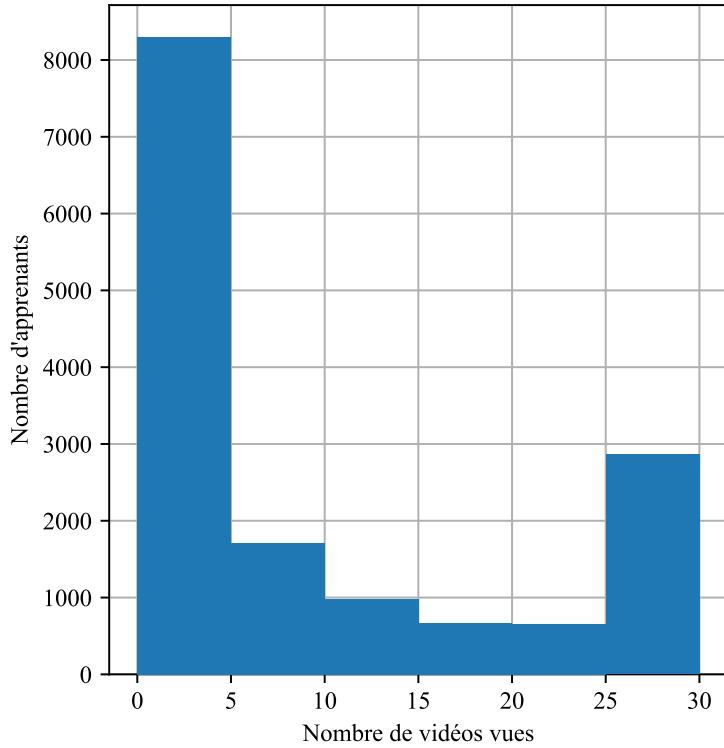


FIGURE 7 – Répartition de l'effectif des apprenant.e.s en fonction du nombre de vidéos vues

La Figure 7 permet de voir que la répartition du nombre de vidéos vues suit une courbe descendante depuis zéro jusqu'à 25 vidéos vues. La courbe descend fortement dans la première partie car 8200 apprenant.e.s ont vu entre zéro et cinq vidéos alors que 1700 apprenant.e.s ont vu entre cinq et dix vidéos. La courbe descend ensuite doucement jusqu'à 700 apprenant.e.s qui ont vu entre 20 et 25 vidéos. On observe ensuite une ascendance dans la répartition : 2900 apprenant.e.s ont vu entre 25 et 30 vidéos. Cette répartition fait penser à une courbe de loi de poisson, souvent observée pour la distribution des données de comptage. On décide donc d'appliquer une régression de poisson.

	Coef	Err. Stand.	z	IC (0,025 - 0,975)	P> t	
Référence	-0,2725	0,011	-24,92	-0,29 -0,25	0,000	***
Nombre de vidéos	0,0676	0,000	140	0,06 0,07	0,000	***

p-value <0,001 : \*\*\*, p-value <0,01 : \*\*, p-value <0,05 : \*

TABLEAU 8 – Corrélation entre le nombre de quizz réalisés et le nombre de vidéos vues résultat de la régression de Poisson

L'hypothèse H0 est que le nombre de quizz réalisés est indépendant du nombre de vidéos vues. Le risque est fixé à 0,01. Le résultat de la régression de Poisson est présenté dans le Tableau 8. Coefficient = 0,81 (calcul :  $\exp(-0,2725) * \exp(0,0676)$ ), p-value = 0,000, Intervalle de Confiance = 0,06 à 0,07. La p-value est de 0,000. Elle est inférieure à 0,01 (et même inférieure à 0,001). L'hypothèse H0 est donc rejetée. Le nombre de quizz réalisés dépend du nombre de vidéos vues.

## 6 Références

Falissard B. (2017), MOOC Introduction à la statistique avec R