Étude statistique et modélisation de la fréquentation du Centre de Ressources en Langues du Campus Scientifique

Benoît GOZÉ et Baya REZIG

Sous la direction de Gwénaëlle CASTELLAN et Nicole CHAPEL.

Le 10 mai 2019

Sommaire

Remerciements

Nous tenons avant tout à remercier nos deux professeurs accompagnants, Gwénaëlle CASTELLAN (pour son expertise mathématique et son accompagnement pour le développement théorique) et Nicole CHAPEL (pour ses connaissances sur le fonctionnement et l'historique du CRL).
Nous tenons également à remercier Caterina CALGARO, responsable du Master 1 MAS, ainsi que Jean-Michel TIERS, responsable technique du CRL, pour sa collaboration.

Introduction

Le CRL_CS (Centre de Ressources en Langues-Cité Scientifique) de l'université de Lille est un lieu privilégié où l'on peut venir étudier des langues variées, en particulier l'anglais. Pour cela, de nombreuses ressources, notamment informatiques, sont à la disposition des étudiants et personnels de l'université. Parmi ces ressources, nous pouvons citer les manuels d'apprentissage, le logiciel TOEIC Test Simulator, le logiciel Reflex English, ou encore le LAN (Laboratoire Audio-Numérique), avec lequel les utilisateurs peuvent s'exercer à la compréhension orale. Le CRL est ouvert du lundi au vendredi de 9h00 à 18h00 (sauf le vendredi, où l'heure de fermeture est 17h00).

Les étudiants peuvent venir dans le cadre d'une autoformation (tuteurée ou non) ; c'est-à-dire qu'ils ont un nombre d'heures d'apprentissage des langues (plus fréquemment de l'anglais) obligatoire à réaliser par eux-mêmes au CRL. Les étudiants sans autoformation obligatoire viennent au CRL pour diverses raisons, notamment pour s'entraîner en vue de certifications de type TOEIC. Une grande partie de notre étude sera justement consacrée au profil de ces étudiants qui viennent sans obligation d'autoformation.

Malheureusement, nous ne pourrons pas traiter des statistiques liées aux ressources utilisées au CRL dans ce rapport ; en effet, ces dernières ne sont pas fiables puisque, la grande majorité du temps, la ressource sélectionnée par le tuteur d'accueil lors de l'arrivée d'un étudiant est "Autoformation", sans plus de détails.
Il nous a donc fallu trouver d'autres axes d'études, qui soient à la fois intéressants d'un point de vue mathématique, mais aussi qui aient un sens concret et utile pour la responsable du CRL, Madame CHAPEL.

Nous avions plusieurs tableaux de données (bases de données) à notre disposition : soit

ils se trouvaient sur l'ancien poste réservé aux moniteurs du CRL, soit le responsable technique du CRL, Monsieur TIERS, nous les a envoyés suite à nos demandes. Les principaux tableaux que nous avons utilisés sont :

- Base de données générale 2016-2017 : elle regroupe tous les inscrits au CRL sur l'année scolaire sus-mentionnée, ainsi que des informations comme leur niveau CECRL (A1, A2, B1, B2, C1 ou C2), leur année d'étude (L1, L2, etc.), le nom de leur professeur(e) d'anglais...
- Base de données générale 2017-2018 ;
- Tableaux récapitulatifs des conversations pour les années scolaires 2010-2011 à 2017-2018 : le CRL-CS propose des séances de conversation avec des lecteurs natifs de 12h30 à 13h15 en anglais, allemand, espagnol, français langue étrangère et flamand régional, et ce presque tous les jours de la semaine. Ces tableaux permettent de faire le bilan de ces conversations : on peut y voir le nombre d'inscrits en avance (via Moodle), le nombre effectif de présents, le nombre total d'inscrits par mois...
- Nombre de visites au CRL par jour pour l'année 2017-2018.

Cependant, il y avait beaucoup de données manquantes dans ces bases de données (les étudiants ne remplissent pas tous leur fiche en entier). Nous avons donc créé d'autres bases de données : par exemple, pour les tests d'indépendance (voir partie 2), nous avons enlevé tous les utilisateurs n'ayant pas renseigné leur niveau de langue. Nous avons aussi dû faire des suppositions sur le genre (garçon ou fille) de quelques utilisateurs qui ne l'avait pas renseigné sur la fiche, puisque nous effectuons des études sur les différences entre les populations de filles et de garçons au CRL. Nous avons aussi supprimé les doublons présents dans les bases de données générales.

Avant de commencer, nous aimerions mentionner quelques pistes de travail que nous avons essayées mais qui se sont révélées éronnées ou inefficaces :

- La regression linéaire : le R^2 était beaucoup trop petit (de 0.01 à 0.05) alors qu'il devrait être proche de 1. Même avec des dummy variables (transformation de variables qualitatives en variables binaires ou quantitatives), ca ne marchait pas. De plus, nous avons testé la linéarité des données grâce à une superposition de box-plots : on voyait que le caractère des données n'était pas linaire.

— La régression logistique : comme dans nos bases de données générales, nous avions beaucoup de variables qualitatives, nous avons exploré la piste de la régression logistique. Cependant, nous avons vite réalisé qu'utiliser cet outil n'aurait aucun sens : les variables qualitatives étant "Niveau", "Année d'études", "Genre", on ne pouvait pas expliquer ces variables en fonction des variables que nous avions à disposition.

— L'ANOVA : Notre professeure encadrante, Madame CASTELLAN, nous a conseillé d'effectuer une analyse de la variance (ANOVA) à un facteur pour expliquer les différentes variables. Cependant, nous avons été confrontés aux mêmes difficultés que celles évoquées précédemment.
Nous avons donc conclu que, dans le cadre de notre étude, il ne servait à rien de chercher à expliquer une variable en fonction des autres.

— Nous avons aussi exploré des pistes pour essayer d'appliquer un algorithme d'apprentissage supervisé (avec étiquetage), mais les données sont beaucoup trop hétèrogènes pour être séparées en groupes ou pour pouvoir avoir une bonne prédiction de la classe de l'individu (exemple : on a voulu tracer sur un graphe des individus en fonction de leur niveau CECRL (passé en variable quantitative : $A1 = 1$, $A2 = 2$, etc.) en ordonnée et leur temps passé au CRL en abscisse. L' étiquette à prédire était "a augmenté de niveau en 1 an" et "n'a pas augmenté de niveau en 1 an". On s'était logiquement attendu à conclure que plus un étudiant passe de temps au CRL, plus il a de chances d'augmenter son niveau ; cependant, les données étaient beaucoup trop hétérogènes et pas fiables pour une prédiction ou une séparation de classe).

Nous avons donc retenu les axes suivants pour le développement de notre étude :

1. Nous allons d'abord effectuer quelques statistiques descriptives concernant les usagers du CRL, puis comparer les deux années 2016-2017 et 2017-2018 en terme de fréquentation et de profil des usagers.
2. Ensuite, nous allons appliquer des méthodes statistiques plus poussées afin de faire des études approfondies : méthodes d'analyse multivariées, tests d'indépendance...
Notre principal axe d'étude sera l'ensemble des étudiants qui n'ont pas d'autoformation obligatoire, afin de déterminer le profil de ces étudiants et de voir qui serait plus susceptible de venir au CRL_CS.

3. Pour finir, nous allons réaliser une étude sur les séances de conversations.

Chapitre 1

Statistiques descriptives : évolution de la fréquentation au CRL

Nous allons d'abord décrire en détail les bases de données que nous avons appelées "Bases de données générales". Voici les variables qu'elles comportent :

1. Des variables inutiles pour notre étude : Numéro étudiant, date d'inscription, adresse e-mail, numéro de téléphone... Nous les avons supprimées.
2. Des variables qualitatives nominales (c'est-à-dire que l'on ne peut pas classer les individus ; elles représentent des catégories ou des niveaux) : genre, nom du professeur, code formation ;
3. Des variables qualitatives ordinales (c'est-à-dire que l'on peut classer les individus, mais que des calculs classiques type moyenne ou écart-type n'ont aucun sens) : année d'étude, niveau de langue, autoformation ou non ;
4. Des variables quantitatives (la valeur associée est un nombre, on peut effectuer des statistiques descriptives) : temps passé au CRL, nombre d'inscrits en conversation (dans le bilan de conversations)...

Efectuons quelques statistiques qui décriront l'année 2017-2018 :
Nous allons commencer par visualiser le nombre de visites par jour sur l'année 2017-2018 (t représente le numéro du jour d'ouverture dans l'année scolaire) :

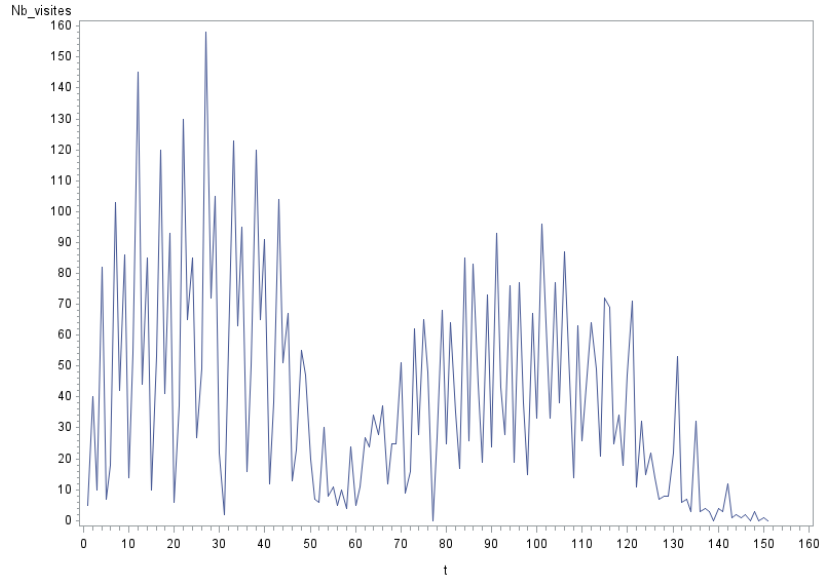


FIGURE 1.1 – Nombre de visites par jour

En première observation, on remarque qu'il y a plus de visites au premier semestre, comparé au deuxième semestre. Ceci peut être expliqué par le fait qu'il y a plus de groupes en autoformation au premier semestre, puis que la majorité des présentations du CRL se déroulent au premier semestre. De plus, la plupart des Master 2 ayant un stage au deuxième semestre, ils ne viennent au CRL que durant le premier semestre.
On remarque plusieurs pics de fréquentation au premier semestre : ils correspondent aux mardis. En effet, ce jour-là, il y avait plusieurs groupes d'autoformation à grand effectif au CRL (notamment les FLE). De plus, certains cours d'anglais avaient lieu directement au CRL par manque de salles disponibles.
Maintenant, explorons plus en détail les statistiques de fréquentation :

— Sur les 151 jours d'ouverture du CRL, il y a eu au total 5958 visites, pour une moyenne de 39 visites par jour.
— L'écart-type des visites vaut 34, ce qui signifie que le nombre de visites est assez hétérogène, ce qui est confirmé par le graphique.
— La médiane (50% des valeurs en dessous de la médiane et 50% au-dessus) est de 28 visites par jour.
— Le maximum de visites est de 158 (le mardi 14 novembre 2017), et le nombre minimal est de 0 (plusieurs dates vers la fin du deuxième semestre, puisque le

CRL était ouvert jusqu'au 15 juin, soit après la fin des cours de la plupart des formations).

Nous allons à présent comparer les années 2016-2017 et 2017-2018 :

	Année 2016-2017	Année 2017-2018
Nombre total d'inscrits	1323	1235
Nombre d'inscrits en autoformation	727	615
Nombre d'inscrits sans autoformation obligatoire	596	620

Il y avait plus d'inscrits an 2016-2017, mais ceci est expliqué par le fait qu'il y avait plus d'étudiants en autoformation.

Maintenant, nous allons plus spécifiquement nous attarder sur les élèves qui n'ont pas d'autoformation obligatoire, pour voir quel est le profil type d'un de ces étudiants.
Année d'étude :

Année d'étude	Nombre d'étudiants 2016-2017	Nombre d'étudiants 2017-2018
L1	85	110
L2	100	115
L3	108	85
M1	126	73
M2	86	96
Thèse	3	3
Enseignant	0	4
Personnel	2	4
Élève ingénieur	1	2
Non précisé	83	128

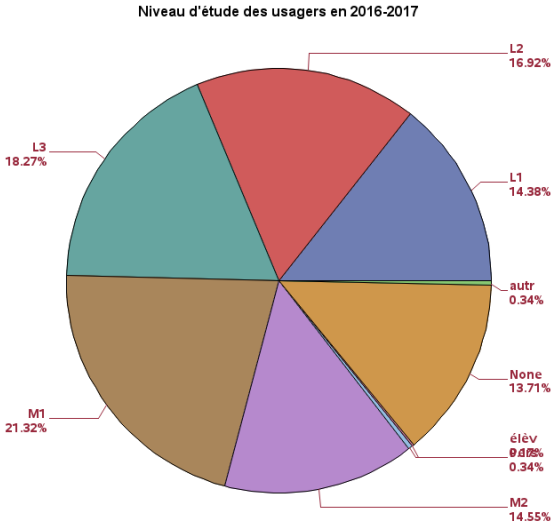


FIGURE 1.2 – Niveau d'étude des usagers en 2016-2017

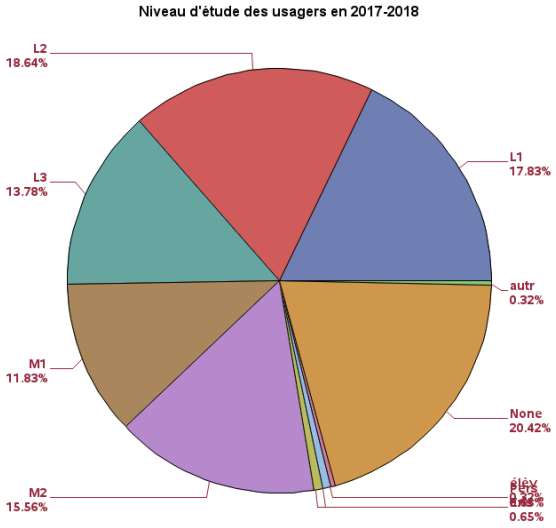
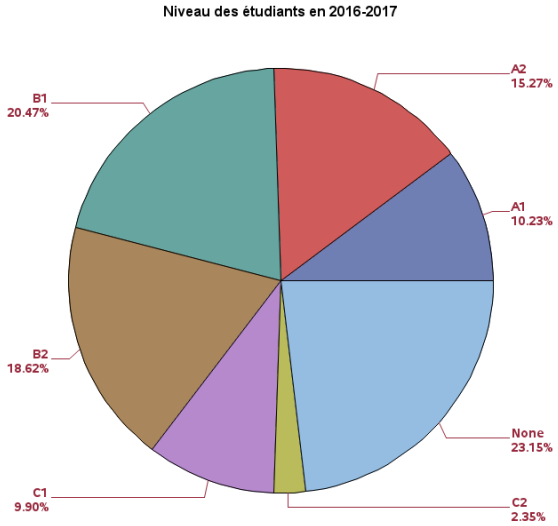


FIGURE 1.3 – Niveau d'étude des usagers en 2017-2018

On remarque qu'il y avait plus de L1 et L2 en 2017-2018, mais plus de L3 et M1 en 2016-2017. Sans doute, les nouveaux étudiants ont été mieux informé de l'existence et de l'utilité du CRL en 2017-2018.
Globalement, des élèves de tous niveaux d'étude sont intéressés par le CRL.

Niveau :

Niveau CECRL	Nombre d'étudiants 2016-2017	Nombre d'étudiants 2017-2018
A1	61	46
A2	91	72
B1	122	90
B2	111	74
C1	59	45
C2	14	12
Non précisé	138	281



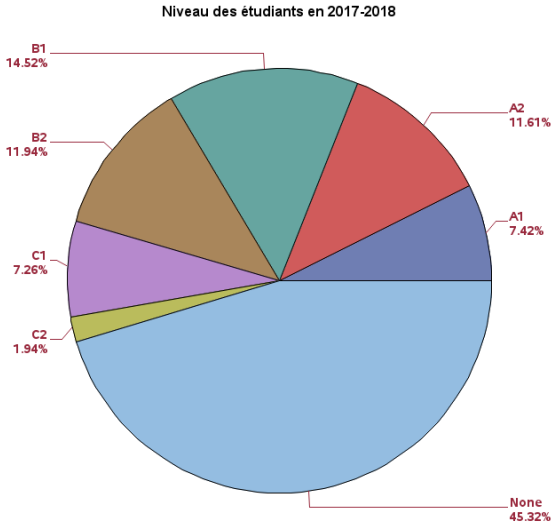


FIGURE 1.4 – Niveau des usagers en 2016-2017 FIGURE 1.5 – Niveau des usagers en 2017-2018

Un très grand nombre d'usagers n'ont pas précisé leur niveau de langue en 2017-2018, ce qui fausse un peu l'étude. Cette anomalie sera sans doute réparée cette année avec l'introduction du test de positionnement SELF, que beaucoup d'étudiants ont passé. Les étudiants les plus présents sans autoformation obligatoire sont ceux ayant le niveau B1, puis ceux ayant les niveaux A2 et B2. Cependant, ces niveaux correspondent aux niveaux les plus représentés parmi l'ensemble des étudiants (on constate une gaussianité des niveaux de langue).

Etudes approfondies sur le temps de présence au CRL :

	Nombre d'étudiants 2016-2017	Nombre d'étudiants 2017-2018
Nombre d'inscrits au total	1323	1235
1 minute de présence ou plus	942	863
2 heures de présence ou plus	649	637
4 heures de présence ou plus	499	476
8 heures de présence ou plus	268	286
16 heures de présence ou plus	91	119

Parmi les étudiants qui n'ont pas d'autoformation obligatoire :

	Nombre d'étudiants 2016-2017	Nombre d'étudiants 2017-2018
Nombre d'inscrits au total	596	620
1 minute de présence ou plus	395	326
2 heures de présence ou plus	183	173
4 heures de présence ou plus	114	114
8 heures de présence ou plus	53	69
16 heures de présence ou plus	10	17

Voici quelques histogrammes qui représentent les temps de présence au CRL, parmi les étudiants qui n'ont pas d'autoformation obligatoire :

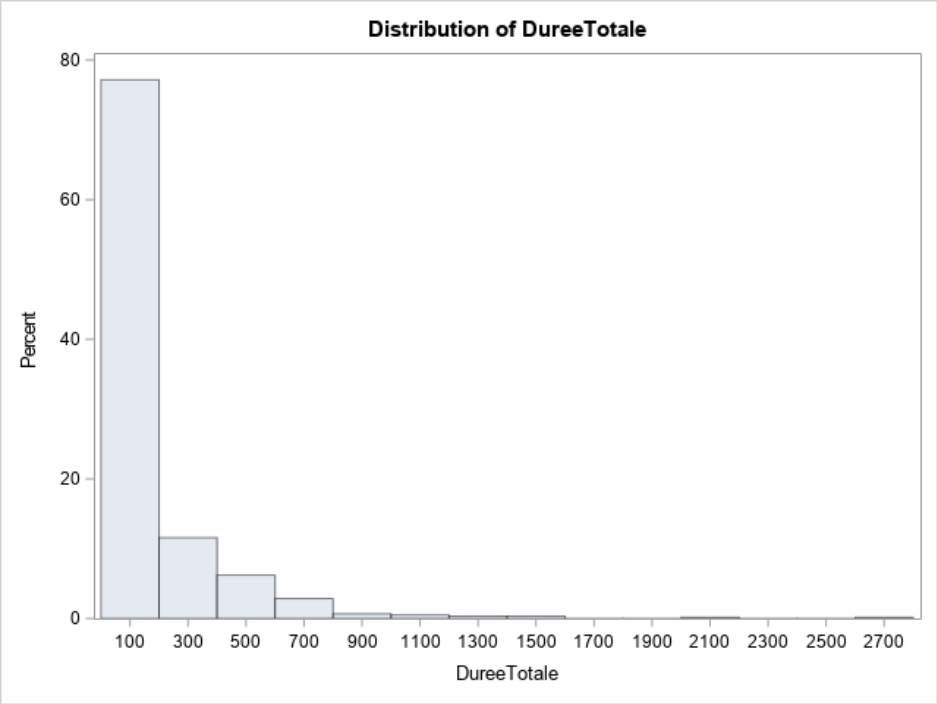


FIGURE 1.6 – Temps de présence en 2016-2017

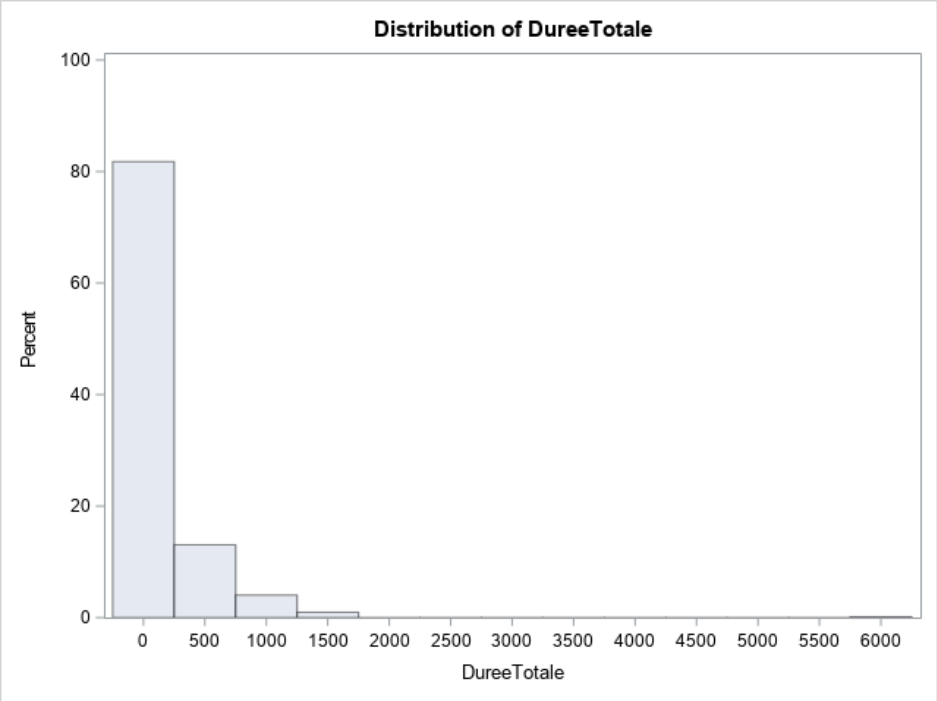
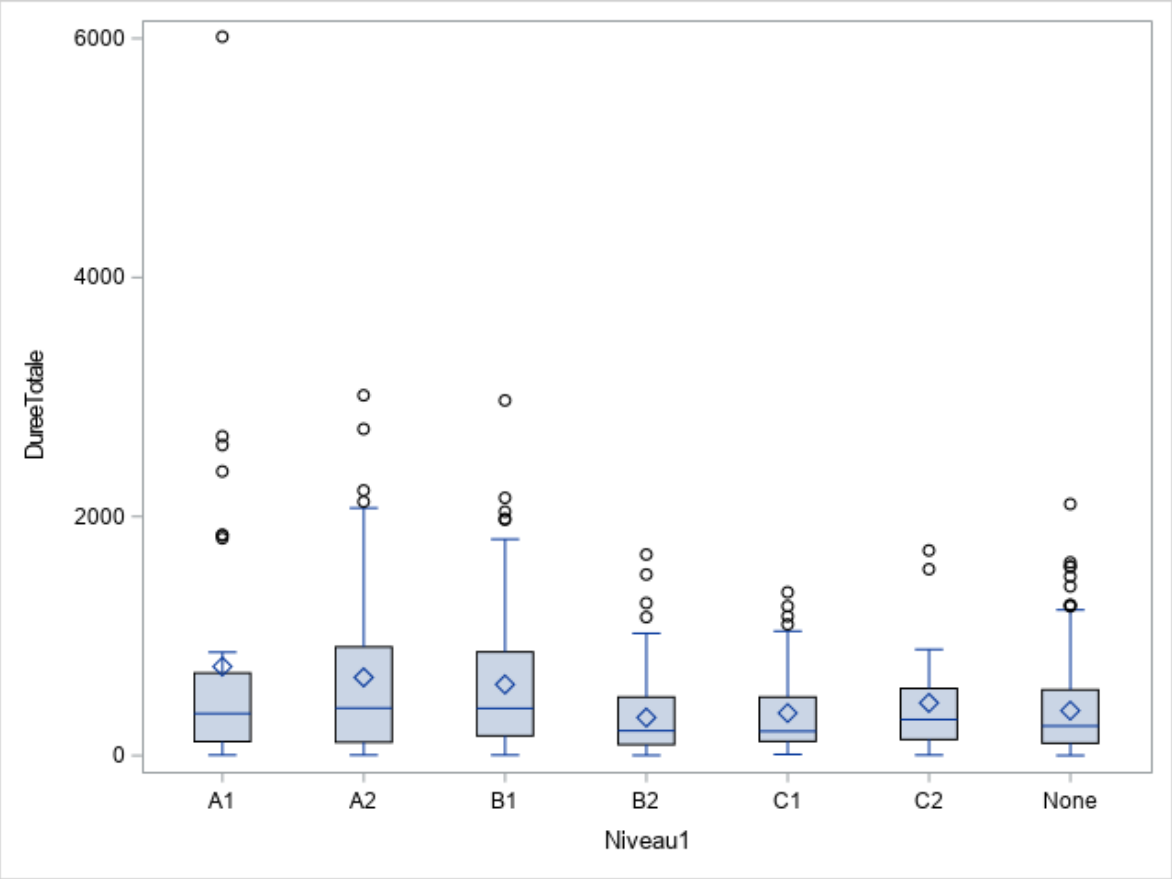


FIGURE 1.7 – Temps de présence en 2017-2018

On voit bien que la plupart des étudiants (près de 80%) ne restent que très peu de temps au CRL (moins de 3 heures). En 2017-2018, il y avait plus d'usagers qui n'ont fait que s'inscrire sans passer de temps au CRL (47.5% contre 34% en 2016-2017) ; mais parmi ceux qui ont passé du temps au CRL, on remarque qu'ils restent plus de temps qu'en 2016-2017.
Il faut aussi prendre en compte les usagers qui ne s'inscrivent au CRL que pour participer aux séances de conversation : en effet, ces derniers sont aussi resté zéro minutes au CRL, mais ont quand même pratiqué les langues (les ateliers de conversation n'ont pas lieu dans le CRL en lui-même, mais dans des salles voisines au SUP).

Traçons les box-plots des niveaux :



On voit que ceux qui restent le plus de temps au CRL sont en général les usagers de niveau A2 ou B1 (même si la moyenne du temps passé, représentée par un carré, est plus élevée chez les A1) : la valeur du troisième quartile (75% des valeurs sont inférieures à ce quartile) est plus élevée dans ces deux populations.
Ce graphique permet aussi de confirmer la non-linéarité des données.

Chapitre 2

Etudes générales

2.1 Détermination du profil des étudiants sans autoformation obligatoire

Nous voulons étudier plus en détail les profils des étudiants qui seraient tentés d'aller au CRL sans autoformation obligatoire. Cela pourrait dépendre de plusieurs facteurs : est-ce que leur professeur d'anglais (ou de langues) leur a longuement parlé du CRL ? Doivent-ils s'entraîner en vue du passage d'une certification en langues ? Ou tout simplement sont-ils passionnés par les langues et veulent-ils approfondir leurs connaissances ?
Tout d'abord, nous allons tenter de voir quelles formations sont les plus représentées parmi les étudiants sans autoformation obligatoire, et voir si ces étudiants sont assidus (s'ils ne viennent pas uniquement pour s'inscrire).

2.1.1 ACP sur les formations

Parmi les étudiants qui se sont autoproclamés "sans autoformation obligatoire", il y en avait certains qui devaient en fait faire de l'autoformation (exemple : MEEF, SIAD...), ce qui a faussé nos résultats dans un premier temps. Nous avons donc supprimé ces formations de notre étude. Nous avons aussi enlevé les ESEAF, car leur nombre d'inscrits (69) est beaucoup plus grand que les autres, et donc l'ACP réalisée avec cette formation ne donnait pas de bons résultats (93% d'inertie sur le premier axe : c'est beaucoup, et tout le monde était regroupé sur le graphique sauf les ESEAF). Cepen-

dant, comme il nous restait peu de formations, nous avons décidé d'élargir nos critères (au début : 8 étudiants inscrits minimum par formation, par la suite 5 étudiants), ce qui nous donna au total 25 formations étudiées (individus), pour 4 variables : nombre d'inscrits, nombre de zéros (ceux qui sont resté zéro minutes), nombre d'assidus (ceux qui sont resté plus de deux heures), temps moyen passé au CRL (par formation) divisé par 10 (pour homogénéiser). Nous avons trouvé les différents étudiants par formation grâce à la colonne "Code Formation" dans la base de données générale.

Maintenant, expliquons l'ACP (Analyse en Composantes Principales) en détail :
Soit un tableau d'observations X de p variables et q individus. L'ACP peut être considérée comme une méthode qui permet de projeter les observations (individus) depuis l'espace de départ des p variables vers un espace réduit de k dimensions, tel qu'un maximum d'information soit conservé.
Le logiciel de statistique SAS effectue automatiquement une ACP normée : dans l'ACP normée, les données sont centrées et réduites (on va appeler le tableau X_{cr}), ce qui les met sur un pied d'égalité. Ensuite, on calcule la matrice de variance-convariance S du tableau de départ ($S = X^T DX - gg^T$, avec D la matrice des poids mis en diagonale et g le centre de gravité des données composé des moyennes des individus selon les variables), et on trouve les valeurs propres (λ_k) de cette derniére (il y a autant de valeurs propres que de variables) ; ensuite, on calcule les vecteurs propres liés au valeurs propres (u tels que $Su = \lambda u$) : ce sont aussi les axes principaux de l'analyse u^j (car la métrique peut être assimilée à l'identité), de même que les facteurs principaux v^j.

La somme des valeurs propres valant l'inertie totale, on veut garder autant d'axes -liés aux valeurs propres- tels qu'ils expliquent le plus d'inertie. Généralement, on utilise la règle de Kaiser : on prend les valeurs propres qui valent plus que la moyenne de celles-ci. On a donc déterminé notre nombre d'axes pour notre analyse en composantes principales.
Après, il faut calculer les composantes principales :

$$c^j = X_{cr} v^j.$$

Ces composantes sont interprétées comme les coordonnées des observations sur le plan factoriel correspondant.

Pour observer la qualité de représentation des individus, on regarde leur contribution à la création d'un axe : plus cette contribution est grande, plus il va aider à créer l'axe. De même, on peut regarder le cos^2, qui nous dit si l'individu est bien représenté sur un axe : plus le cos^2 de l'individu est proche de 1, mieux il est représenté.

Concernant les variables : on peut regarder leur matrice de corrélation : la corrélation peut aller de -1 à 1. Plus elle se rapproche de 1, plus les deux variables concernées sont corrélées positivement, c'est-à-dire que quand la valeur liée à la variable 1 augmente, la valeur liée à la variable 2 augmente aussi. Inversement quand le coefficient de corrélation se rapproche de -1. Quand la corrélation entre ces deux variables se rapproche de 0, il n'y a pas de lien entre ces variables.
On peut regarder le cercle de corrélations : plus une variable, représentée par une flèche se rapproche du bord du cercle, mieux elle sera représentée. On peut aussi regarder quel axe une variable va aider à créer si on regarde la direction de pointage de la variable (si c'est à gauche ou à droite, elle va contribuer à créer l'axe des abscisses ; si c'est en haut ou en bas, elle va contribuer à créer l'axe des ordonnées).

Appliquons la théorie à notre étude :
Statistiques simples :

	Nb d'inscrits	Nb de zero minutes	Nb de plus de 2h	temps divisé par 10
Moyenne	7.84	4.32	1.52	10.08
Écart-type	4.44	4.53	1.05	6.73

La moyenne du nombre d'inscrits par formation étudiée est de 8 élèves, mais il y a une forte hétérogénéité des valeurs : l'écart-type, qui mesure l'écart avec la moyenne, est élevé (4.44). Il en est de même pour ceux qui sont resté zéro minutes au CRL. Le temps moyen de passage au CRL par formation est d'environ 101 minutes.

Matrice de corrélation :

	Nb d'inscrits	Nb de zero minutes	Nb de plus de 2h	temps divisé par 10
Nb d'inscrits	1.00	0.97	0.74	-0.09
Nb de zero minutes	0.97	1.00	0.63	-0.20
Nb de plus de 2h	0.74	0.63	1.00	0.45
temps divisé par 10	-0.09	-0.20	0.45	1.00

Le nombre d'inscrits et le nombre d'étudiants étant resté zéro minutes au CRL est fortement corrélé, tout comme le nombre d'inscrits et le nombre d' "assidus" au CRL. On remarque aussi une corrélation entre le nombre de zéro minutes et le nombre de plus de deux heures, ce qui n'est pas forcément intuitif.

Matrice de covariance :

	Nb d'inscrits	Nb de zero minutes	Nb de plus de 2h	temps divisé par 10
Nb d'inscrits	19.64	18.38	1.13	-7.36
Nb de zero minutes	18.39	20.48	0.035	-11.44
Nb de plus de 2h	1.13	0.04	1.09	4.62
temps divisé par 10	-7.36	-11.44	4.62	45.33

Les valeurs propres de la matrice de covariance sont les suivantes, accompagnées du pourcentage de l'inertie approché qu'elles représentent et du pourcentage cumulé de l'inertie :

	Valeur Propre	Proportion	Cumul
1	55.93	0.65	0.65
2	28.67	0.33	0.98
3	1.64	0.02	1.00
4	0.31	0	1.00

On utilise la règle de Kaiser : la moyenne des valeurs propres vaut 21.64, donc on prend 2 valeurs propres. Le premier plan factoriel représente 98% de l'inertie totale. On est sûr que l'information sera bien représentée.

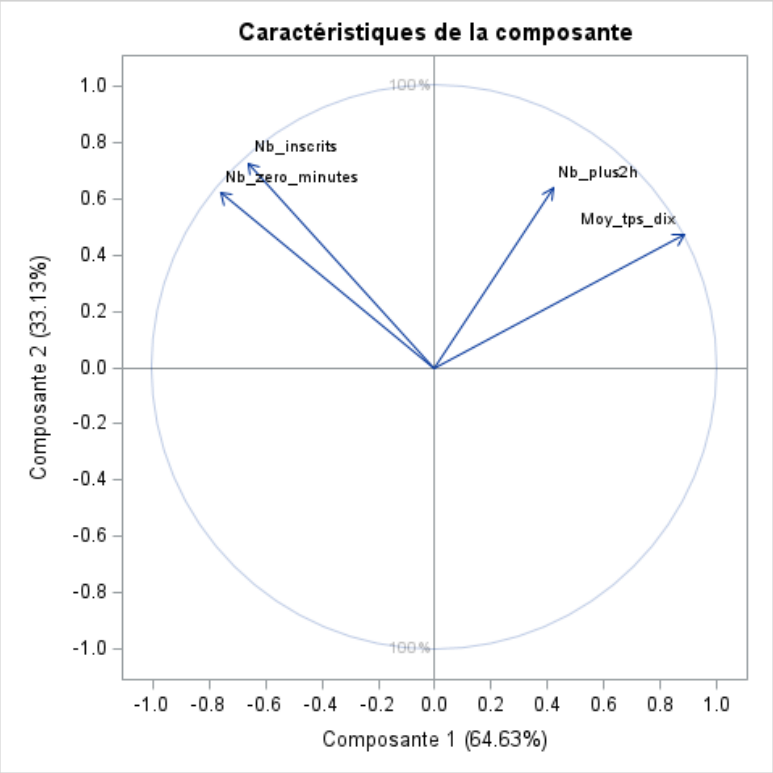


FIGURE 2.1 – Cercle des corrélations

Les variables sont toutes bien représentées sur ce plan, sauf le nombre d'assidus. Le nombre d'inscrits total et le nombre de zéro minutes vont fortement participer à la création de l'axe 2, tandis que la moyenne du temps divisé par 10 va participer un peu à la création de l'axe 1.

L'individu qui contribue le plus à la création de l'axe 1 est SESI : en effet, cette formation compte 26 individus. Elle contribue au côté négatif de l'axe. La formation Ingé Maths va également contribuer dans le même sens. La formation Hygiène et Sécurité contribue dans le sens opposé, tout comme la formation Sciences et Technologies ou encore Personnel.
Concernant l'axe 2, les formations SESI (côté positif), Guide Nature (côté négatif), SVTE Aménagé (côté négatif), Master Génie Civil (côté négatif) sont les formations qui contribuent le plus.

Ces mêmes formations sont aussi celles qui ont les meilleurs cos^2 (pour SESI, c'est le

cos^2 de l'axe 1 qui domine).

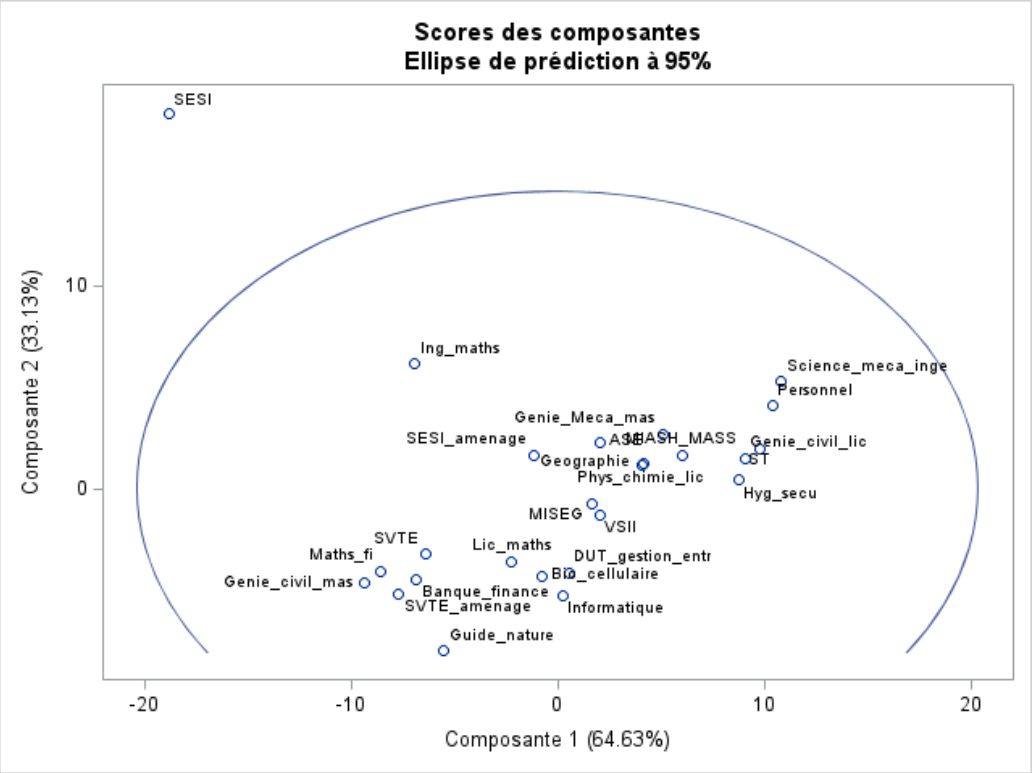


FIGURE 2.2 – Résultats de l'ACP : projection des individus dans le premier plan factoriel

Il y a une certaine homogénéité autour de l'origine : SESI est isolé sur le haut, du fait de son grand nombre d'inscrits par rapport aux autres formations. On remarque tout de même que les formations ne sont pas toutes groupées au même endroit : plus une formation se trouvera en haut, plus elle aura d'inscrits et de "zéro minutes" (comparativement à son nombre d'inscrits). Plus une formation se trouvera vers le bas, moins elle aura d'inscrits.
Plus une formation se trouvera vers la droite, plus on peut dire qu'elle compte des membres sérieux et assidus (en relation avec son nombre total d'inscrits). Inversement, plus une formation se trouvera vers la gauche, plus ses étudiants resteront peu de temps au CRL. C'est l'axe qui compte le plus car il représente 65% de l'inertie totale.
Nous pouvons conclure que les formations les plus assidues sont Science Méca Ingé, Personnel, Licence Génie Civil, Sciences et Technologies ou encore Hygiène et Sécurité

(HSQE). A contrario, les formations les moins assidues sont Master Génie Civil, Maths Finance ou encore SVTE. Les SESI étant nombreux, il est normal qu'ils aient beaucoup de représentants, mais ils ne sont pas forcément assidus.

2.1.2 Classification hiérarchique

Afin d'y voir plus clair qu'avec l'ACP et de peut-être voir des similitudes avec cette dernière, nous allons effectuer une classification hiérarchique :
On considère les distances entre les individus. On veut les séparer en classes différentes. La méthode de Ward consiste à regrouper les classes de façon à ce que l'augmentation de l'inertie interclasse (entre les classes) soit maximale ; donc l'inertie intraclasse sera minimale (puisque variance totale=variance interclasse + variance intraclasse).
Nous avons besoin de savoir en combien de classe nous allons diviser l'ensemble des formations : la statistique du pseudo t^2 peut être utilisée en regardant les maximums (locaux ou globaux) sur le graphique : le nombre de classes sera égal à la valeur pour lequel le pseudo t^2 est maximal, plus un [4]. On peut aussi utiliser les maximums du CCC (critère de classification cubique) ou les maximums de la loi de pseudo F de rang signé (ici, il ne sera pas très fiable car il est toujours croissant).

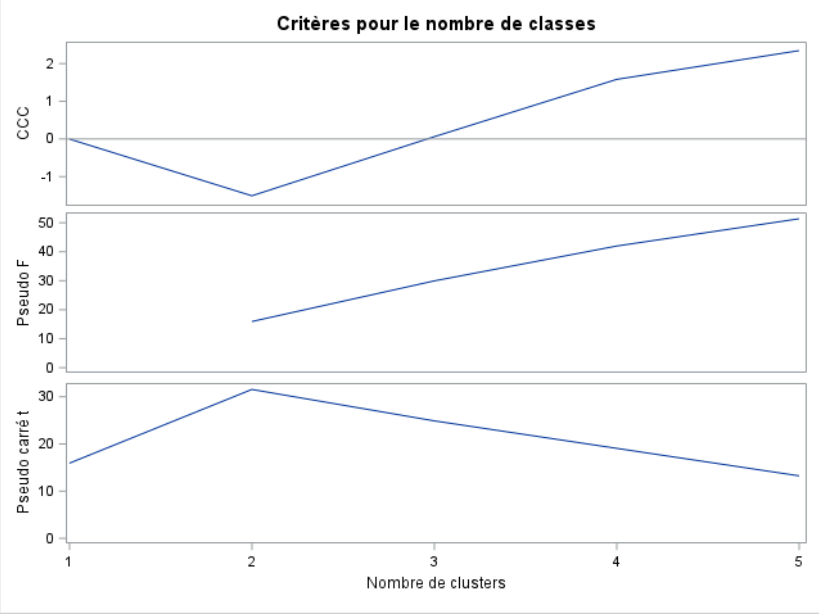


FIGURE 2.3 – Critères pour le nombre de classes

Dans notre cas, on voit que l'indicateur le plus fiable sera le pseudo-t^2. On voit qu'il a un maximum global pour 3 classes (2+1). Le CCC, quant à lui, nous indique 5 classes.

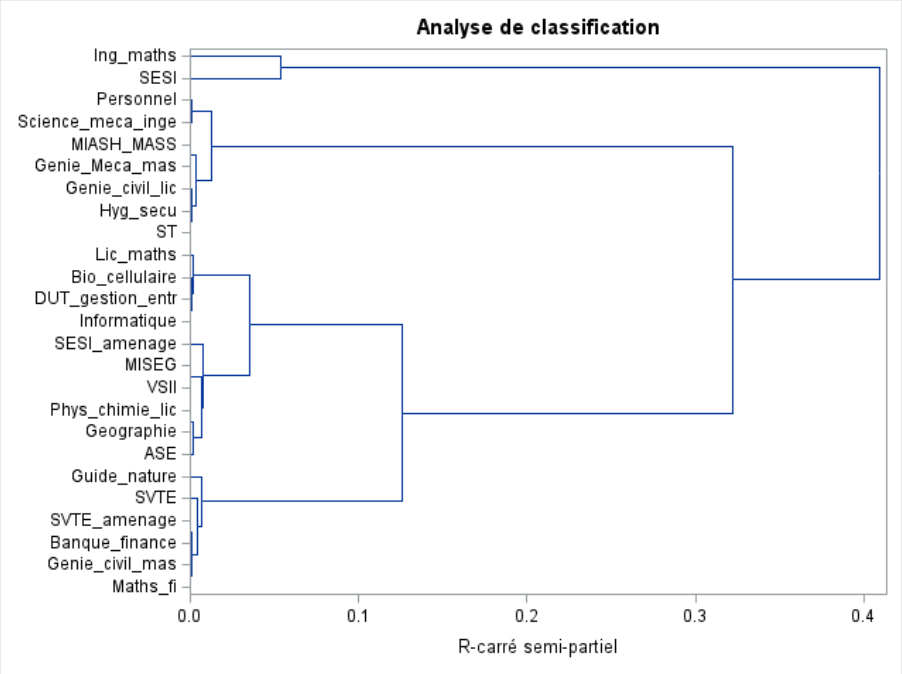


FIGURE 2.4 – Avec la métrique de Ward

Nous allons donc créer nos 3 classes : il faut donc faire une séparation verticale sur le graphique précédent quand le R carré semi-partiel vaut entre 0.13 et 0.32. Nous avons donc nos groupes :

1. Groupe 1 : SESI, Ingé Maths : il est constitué des formations qui ont "beaucoup" d'inscrits (SESI : 26 inscrits ; Ingé Maths : 16 inscrits). Dans la plupart des cas, ce sont des formations dont tous les élèves ont été inscrits lors de leur première visite du CRL.
2. Groupe 2 : Personnel, Science Méca Ingé, MIASH MASS, Master Génie Mécanique, Hygiène et Sécurité, Licence Génie Civil, Sciences et Technologies : c'est la classe des formations "assidues".
3. Groupe 3 : DUT Gestion Entreprise, Biologie Cellulaire, Licence Maths, Informatique, MISEG, VSII, SESI Aménagé, Licence Physique-Chimie, Géographie, ASE, Banque Finance, SVTE, GuideNature, Master Génie Civil, SVTE Aménagé, Maths Finance : ce groupe constitue la "normalité", la "moyenne" au niveau

des formations sans autoformation obligatoire, que ce soit au niveau du nombre d'inscrits et au niveau du temps passé au CRL. Ils ne sont pas très assidus.

Nous avons fait de notre mieux pour classer les formations qui n'ont pas d'autoformation : en général, il y a quand même la majorité des usagers qui s'inscrivent mais ne restent que pas ou peu de temps au CRL. De plus, dès qu'il y a une valeur extrême (dans notre étude initiale, ESEAF), elle a tendance à fausser les résultats. De plus, il a été assez dur d'établir une classification qui ait du sens.
Nous pouvons donc établir notre conclusion sur les étudiants qui n'ont pas d'autoformation : ils sont bien sûr moins assidus en général que ceux qui ont de l'autoformation, mais lorsqu'ils viennent au CRL, ils font du travail sérieux pendant plusieurs heures, parfois plus encore que ceux en autoformation.

2.2 Tests du χ^2 et AFC

En mathématiques, le test du χ^2 d'indépendance sert à vérifier si deux variables sont indépendantes ou non. Mettons en place ce test :
-Il faut d'abord mettre en place les hypothèses de ce test :

$$\begin{cases} H_0 : \text{Les variables sont indépendantes} \\ H_1 : \text{Les variables ne sont pas indépendantes} \end{cases}$$

Maintenant, définissons la statistique de test : soit un tableau de contingence quelconque, avec deux variables X et Y (X a k modalités et Y a l modalités). On obtient des données appariées $(X_i, Y_i)_{\{1 \leq i \leq n\}}$: on notera les k modalités de X par $A_1, \ldots, A_k$, et les l modalités de Y par $B_1, \ldots, B_l$. On définit $N_{cd} = \sum_{i=1,j=1}^{n_1,n_2} \mathbb{1}_{\{X_i \in A_c, Y_i \in B_d\}}$: c'est le nombre d'individus (observations) pour lesquels la variable X prend la modalité A_c et la variable Y prend la modalité B_d.

	B_1	...	B_q	...	B_l
A_1	N_{11}	...	N_{1q}	...	N_{1l}
$\vdots$	$\vdots$	...	...	...	$\vdots$
A_j	N_{j1}	...	N_{jq}	...	N_{jl}
$\vdots$	$\vdots$	...	...	...	$\vdots$
A_k	N_{k1}	...	N_{kq}	...	N_{kl}

Exemple pourle CRL_CS : dans le test Niveau/Libellé, N_{11} correspondrait au nombre d'individus de niveau A1 en L1.
On peut maintenant calculer la statistique

$$S = \sum_{j=1}^{k} \sum_{q=1}^{l} \frac{(N_{jq} - \frac{N_{j.}N_{.q}}{n})^2}{\frac{N_{j.}N_{.q}}{n}}.$$

Sous H_0, cette statistique converge en loi vers une loi du χ^2 à $(k-1)(l-1)$ degrés de liberté. A contrario, sous H_1, S tend presque sûrement vers $+\infty$ quand n tend vers $+\infty$.

Ceci nous permet de définir une zone de rejet (et plus précisément le seuil) de la forme $\mathcal{R} = \{S \geq t_\alpha\}$: si la valeur de la statistique se trouve dans la zone de rejet, on rejettera H_0 au seuil que l'on aura défini. Pour cela, il faut regarder une table de χ^2 et trouver le quartile d'ordre 0.95.

Rappel sur la p-valeur : c'est la probabilité pour un modèle statistique donné sous l'hypothèse nulle d'obtenir la même valeur ou une valeur encore plus extrême que celle observée. Si cette valeur est inférieure à 0.05, on rejette H_0 au niveau 5 %. SAS la calcule automatiquement.

On se propose de faire deux tests d'indépendance du χ^2 parmi tous les étudiants inscrits au CRL en 2017-2018 : un qui étudie la relation entre le genre et le niveau, et un autre qui étudie la relation entre le niveau et l'année d'étude. On a enlevé tous les étudiants qui n'ont pas renseigné leur niveau.

Premier test :
Table de contingence :

	A1	A2	B1	B2	C1	C2	Total
F	26	80	78	61	43	12	300
M	48	73	115	78	42	13	369
Total	74	153	193	139	85	25	669

La statistique du χ^2 vaut 9.0648 (calcul par SAS).
La p-valeur vaut 0.10 (ou on regarde une table du χ^2 : on trouve que $t_\alpha = 11.07$ au niveau 5% pour une loi du χ^2 à 5 degrés de liberté) : H0 n'est pas rejetée ; on ne rejette pas l'indépendance des variables 'Genre' et 'Niveau' (même si ce n'est pas totalement net car la p-valeur est quand même assez proche de 0.05).
On en déduit que le fait d'être garçon ou fille ne nous prédispose pas à avoir des meilleures aptitudes pour l'étude d'une langue.

L' autre test :
Le test du χ^2 peut ne pas convenir car il y a trop de modalités petit effectif (moins de 5 individus dans la classe). Recréeons une table sans les Th/ed , personnels, enseignants et autres.
Voici la table de contingence obtenue :

	A1	A2	B1	B2	C1	C2	Total
L1	27	40	50	12	13	5	147
L2	16	37	47	29	20	5	154
L3	5	23	24	40	25	7	124
M1	10	18	27	29	18	2	104
M2	8	36	53	40	20	6	163
Total	66	154	201	150	96	25	692

La statistique du χ^2, calculée par SAS, vaut 61.5143. La p-valeur associée au test est proche de 0 (< 0.0001) ; ou, si on regarde une table du χ^2, on trouve que t_α pour 20 degrés de liberté à 5% vaut 31.4.
Résultat : on ne peut pas rejeter l'indépendance entre genre et niveau au seuil 5%, tandis que l'on rejette l'indépendance au seuil 5% entre le niveau et l'année d'étude (calculs réalisés avec SAS).
Ce qui veut dire que, selon notre étude, il y a un lien entre le niveau CECRL et l'année d'études.

Nous allons effectuer une Analyse Factorielle des Composantes (AFC) afin de conforter cette hypothèse.

Explications : on part du tableau de contingence des deux variables étudiées : on crée les tableaux suivants, appelés tableau des profil-lignes et tableau des profil-colonnes : pour créer le tableau des profil-lignes, on divise les observations N_{ij} par la somme des éléments de la ligne correspondante : $\frac{N_{ij}}{N_{i.}}$. Pour créer le tableau des profil-colonnes, on divise les obsevations par la somme des éléments de la colonne correspondante : $\frac{N_{ij}}{N_{.j}}$. Ensuite, pour bien justifier le fait de faire une AFC, on calcule le χ^2 : s'il vaut 0, les variables sont indépendantes et on ne peut pas faire une AFC car il faut que les variables soient dépendantes (dans notre étude, notre χ^2 est bien loin de 0 ; on peut utiliser une AFC).

Faire une AFC correspond à faire deux ACP :

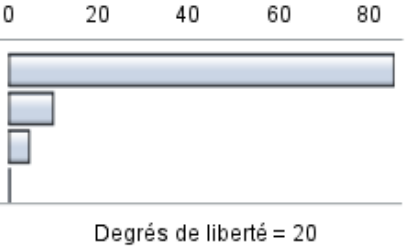
- une avec les profil-lignes : dans ce cas, la métrique est nD_2^{-1}, avec D_2 la matrice des sommes des éléments de chaque colonne en diagonale.
- une avec les profil-colonnes : dans ce cas, la métrique vaut nD_1^{-1}, avec D_1 la matrice avec les sommes des éléments des lignes sur ma diagonale.

Les facteurs principaux dans l'ACP des profil-lignes sont les vecteurs propres de $D_2^{-1}N^TD_1^{-1}N$, avec N la table de contingence.
Les facteurs principaux dans l'ACP des profil-colonnes sont les vecteur propres de $D_1^{-1}ND_2^{-1}N^T$
Les composantes principales dans l'ACP des profil-lignes sont les vecteurs propres de $D_1^{-1}ND_2^{-1}N^T$
Les composantes princiapales dans l'ACP de profil-colonnes sont les vecteurs propres de $D_2^{-1}N^TD_1^{-1}N$.
Aprés, on regardee les contributions et les cos^2.
Il faut que les variables soient liées (non indépendantes).
On a retiré les individus dont l'année d'étude était non précisée. De plus, nous avons également retiré les individus ayant indiqué "Thèse", "Élève ingénieur", "Personnel" ou "Enseignant", puisque, leurs effectifs étant trop petits, ils risquaient de fausser l'étude.
Table de contingence

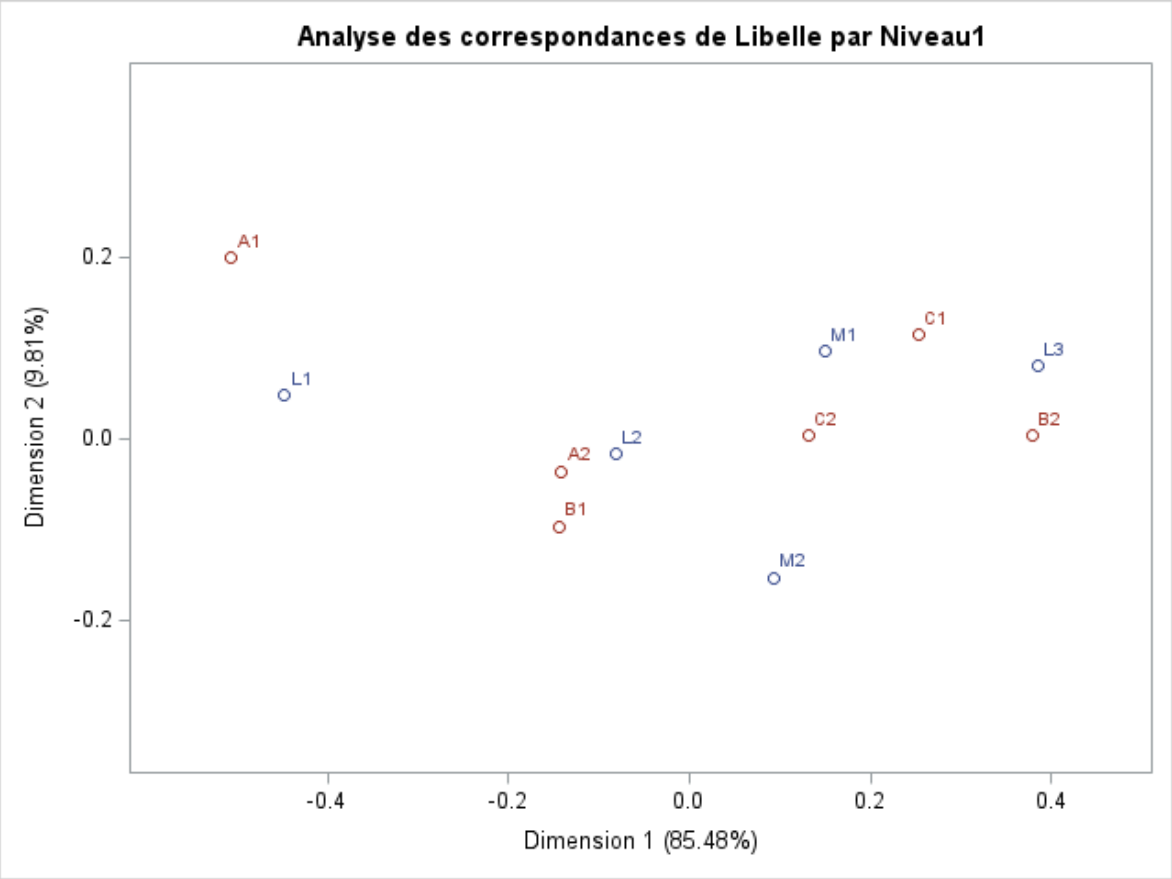
	A1	A2	B1	B2	C1	C2	Effectif
L1	27	40	50	12	13	5	147
L2	16	37	47	29	20	5	154
L3	5	23	24	40	25	7	124
M1	10	18	27	29	18	2	104
M2	8	36	53	40	20	6	163
Effectif	66	154	201	150	96	25	692

Décomposition de l'inertie et du Khi-2

Valeur singulière	Inertie principale	Khi-2	Pourcentage	Pourcentage cumulé
0.27565	0.07599	52.5819	85.48	85.48
0.09338	0.00872	6.0346	9.81	95.29
0.06320	0.00399	2.7636	4.49	99.78
0.01388	0.00019	0.1333	0.22	100.00
	0.08889	61.5134	100.00	



On va garder 2 valeurs propres : on a 95% de l'inertie expliquée.
Contributions partielles à l'inertie des points des lignes : axe 1 : L1 (0.56), L3 (0.35) ; axe 2 : M2 (0.63).
Contributions partielles à l'inertie des points des colonnes : axe 1 : A1 (0.32), B2 (0.41) ; axe 2 : A1 (0.44), B1 (0.32).
Individus bien représentés sur l'axe 1 : L1,L2,L3, A1, A2, B2,C1.
Individus bien représentés sur l'axe 2 : M2, B1.



Sur ce graphique, on voit que l'on peut faire des rapprochements : A1 et L1 sont associés, L2 est associé avec A2 et B1, L3 est associé à B2 et C1, M1 est lié à C1 et C2, puis M2 est lié en partie à C2. Celà montre que les L1 ont une plus grande proportion de A1 que les autres étudiants, que les M1 ont une plus grande proportion de C1 que les autres étudiants, etc.
Ce qui corrobore notre hypothèse.

2.3 Etudes de comparaisons entre garçons et filles

2.3.1 Partie théorique

Dans cette partie théorique, nous allons aborder les tests non-paramétriques, et plus particulièrement les tests de Kolmogorov-Smirnov et de la somme des rangs [2]. Ce

sont des tests non paramétriques de comparaison d'échantillons.

Le principe des tests est expliqué dans la section sur les tests d'indépendance (du χ^2). Ce test est justement un test non-paramétrique. Cependant, ce dernier utilise des données appariées (non indépendantes), tandis que les deux tests que nous allons présenter dans cette section utilisent deux échantillons indépendants.

Un test non-paramétrique est un test dans lequel la ou les lois des échantillons sont inconnues, et ces dernières ne sont pas dans des espaces de dimension finie (ou dans un modèle paramétrique). Il est donc moins contraignant qu'un test paramétrique, mais moins puissant.
Quelques conditions : il faut que les échantillons soient aléatoires, et tous les individus sont prélevés indépendamment. Les échantillons doivent aussi être indépendants les uns des autres.

Somme des rangs pour la comparaison de deux populations

Soient $n_1 \in \mathbb{N}^*$ et $n_2 \in \mathbb{N}^*$. Soient l'échantillon de n_1 observations $X = (X_1, \ldots, X_{n_1})$ et l'échantillon de n_2 observations $Y = (Y_1, \ldots, Y_{n_2})$. Regroupons ces échantillons : on aboutit à l'échantillon $Z = (Z_1, \ldots, Z_{n_1}) = (X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ de taille $n_1+n_2 = n$.
Nous allons réordonner Z, c'est-à-dire que nous allons mettre la plus petite observation à la première place, puis la deuxième plus petite, etc. L'échantillon réordonné sera noté $\tilde{Z} = (Z_{(1)}, \ldots, Z_{(n)})$

Exemple. Soient $X = (19, 27, -1)$ et $Y = (6, 7, 13, -4, 3)$. Z sera l'échantillon $(19, 27, -1, 6, 7, 13, -4, 3)$. L'échantillon réordonné est $(-4, -1, 3, 6, 7, 13, 19, 27)$.

À présent, nous pouvons définir le rang :

Définition 2.3.1. Le rang $R(i)$ est la position de l'élément i dans l'échantillon réordonné.
Formulation mathématique : $R(i) = \sum_{j=1}^{n} \mathbb{1}_{\{Z_j \leq Z_i\}}$

Proposition 2.3.2. *R suit une loi uniforme sur l'ensemble des permutations $\mathfrak{S}_n$.*

Proposition 2.3.3. *$R(i)$ suit une loi uniforme sur $\{1, \ldots, n\}$.*

Exemple. Reprenons notre exemple : $R(X_3) = R(-1) = 2$.

Si $X < Y$, on peut logiquement s'attendre à ce que les rangs de X soient plus petits que ceux de Y.
Soient F la fonction de répartition ($P(X \leq t)$) de $(X_i)_{1 \leq i \leq n_1}$ et G la fonction de répartition de $(Y_i)_{1 \leq i \leq n_2}$. On suppose que F et G sont continues et que $(X_i)_{1 \leq i \leq n}$ est indépendant de $(Y_i)_{1 \leq i \leq n}$.
Établissons le test de la somme des rangs, ou test de Wilcoxon [3] :
Pour tester si $X \neq Y$:

$$\begin{cases} H_0 : F = G \\ H_1 : F \neq G \end{cases}$$

Ou encore, pour tester si $X < Y$:

$$\begin{cases} H_0 : F = G \\ H_1 : F > G \end{cases}$$

(en effet : si $X < Y$, alors $F > G$.)
On peut aussi établir le test pour tester si XY.
La statistique que nous allons utiliser pour ces tests est la somme des rangs de X : $W_X = \sum_{i=1}^{n_1} R(i)$.

Proposition 2.3.4. *W_X est à valeurs dans $\left\{ \sum_{k=1}^{n_1} k, \sum_{k=n_2+1}^{n_1+n_2} k \right\}$, avec $\sum_{k=1}^{n_1} k = \frac{n_1(n_1+1)}{2}$ et $\sum_{k=n_2+1}^{n_1+n_2} k = \frac{n_1(n_1+2n_2+1)}{2}$ et ne dépend donc pas de F.*

Proposition 2.3.5. *Sous H_0, $E_0(W_X) = \frac{n_1(n_1+n_2+1)}{2}$ et $\mathrm{Var}_0(W_X) = \frac{n_1 n_2(n_1+n_2+1)}{12}$.*

Démonstration. - Calcul de l'ésperance : $E_0(W_X) = E_0(\sum_{i=1}^{n_1} R(i)) = \sum_{i=1}^{n_1} E_0(R(i))$ par linéarité de l'espérance ;
Sous H_0, R suit une loi uniforme sur l'ensemble des permutations (sur σ_n) : les rangs des éléments de X peuvent en effet aller de 1 à $n_1 + n_2 = n$, et ils ont chacun autant de chance de se produire. L'espérance d'une loi uniforme discrète étant, pour a et b quelconques, $\frac{a+b}{2}$, on en déduit que $E_0(W_X) = \sum_{i=1}^{n} \frac{n_1+n_2+1}{2} = \frac{n_1(n_1+n_2+1)}{2}$. □

Théorème 2.3.6. *Sous H_0, $\frac{W_X - E_0(W_X)}{\sqrt{\mathrm{Var}_0(W_X)}}$ converge en loi vers une loi normale de paramètres 0 (espérance) et 1 (variance) quand n_1 et n_2 tendent vers $+\infty$.*

Démonstration. Admise. [8] □

Pour déterminer la zone de rejet pour le test, il faut connaître la loi de W_X sous H_0, c'est-à-dire calculer $P_0(W_X = k)$ pour $k \in \left\{ \sum_{k=1}^{n_1} k, \sum_{k=n_2+1}^{n_1+n_2} k \right\}$.

Exemple. Nous allons calculer la loi de W_X pour $n_1 = 3$ et $n_2 = 4$. On va donc calculer $P_0(W_X = k)$ pour $k \in \{6, \ldots, 18\}$.
$P_0(W_X = k) = P_0(\sum_{i=1}^{n_1} R(i) = k)$.

Exemple avec $k = 6$: $P_0(\sum_{i=1}^{3} R(i) = 6) = P_0(\text{rang1} = 1, \text{rang2} = 2, \text{rang3} = 3) * 3!$
Le 3! vient des possibilités de placement des éléments : on peut avoir $\{1, 2, 3\}$ (c'est le cas explicité ci-dessus : l'élément 1 de X a le rang 1 dans Z, m'élément 2 de X a le rang 2 dans Z et l'élément 3 de X a le rang 3), $\{1, 3, 2\}$ (l'élément 1 a le rang 1, l'élément 3 a le rang 2,...), $\{2, 1, 3\}$, $\{2, 3, 1\}$, $\{3, 2, 1\}$ ou $\{3, 1, 2\}$. Ces cas ont la même probabilité, qui est égale à $\frac{1}{7} * \frac{1}{6} * \frac{1}{5}$.
Donc $P_0(\sum_{i=1}^{3} R(i) = 6) = \frac{1}{210} * 6 = \frac{3}{105}$.

Exemple avec $k = 9$: $P_0(\sum_{i=1}^{3} R(i) = 9) = P_0(\{\text{rang1} = 1, \text{rang2} = 2, \text{rang3} = 6\} \cup \{\text{rang1} = 2, \text{rang2} = 3, \text{rang3} = 4\} \cup \{\text{rang1} = 1, \text{rang2} = 3, \text{rang3} = 5\}) * 3!$.
Donc $P_0(\sum_{i=1}^{3} R(i) = 6) = \frac{1}{210} * 3 * 6 = \frac{3}{35}$.
A la fin, on a le tableau de la loi de W_X :

k	6	7	8	9	10	11	12	13	14	15	16	17	18
$P_0(W_X = k)$	$\frac{3}{105}$	$\frac{3}{105}$	$\frac{6}{105}$	$\frac{3}{35}$	$\frac{12}{105}$	$\frac{12}{105}$	$\frac{6}{35}$	$\frac{12}{105}$	$\frac{12}{105}$	$\frac{6}{70}$	$\frac{6}{105}$	$\frac{3}{105}$	$\frac{3}{105}$

La somme des probabilités vaut 1 : c'est cohérent.

Maintenant, il faut définir une zone de rejet pour notre test : si la statistique de test se trouve dans cette zone de rejet, on va rejetter H_0 au seuil $\alpha\%$.
Pour le test $F = G$, la zone de rejet sera :

$$\mathcal{R}_\alpha = \left\{ W_X \leq k_\alpha^1 \right\} \cup \left\{ W_X \geq k_\alpha^2 \right\},$$

avec k_α^1 et k_α^2 des seuils à déterminer grâce à une table de valeurs de seuils du test de la somme des rangs (pour de petites valeurs de n_1 et n_2) ; ou on peut utiliser le théorème 2.3.6 pour les grandes valeurs de n_1 et n_2.
Pour le test $F > G$, la zone de rejet sera :

$$\mathcal{R}_\alpha = \{W_X \leq k_\alpha\},$$

avec k_α un seuil à déterminer grâce à une table de valeurs de seuils du test de la somme des rangs.

Exemple. Reprenons notre exemple : on sait que $n_1 = 3$ et $n_2 = 5$. On trouve les seuils dans la table :
- au seuil 5%, $k_\alpha^1 = 6$ et $k_\alpha^2 = 21$. Donc $\mathcal{R}_{5\%} = \{W_X \leq 6\} \cup \{W_X \geq 21\}$.
- au seuil 10%, $k_\alpha^1 = 7$ et $k_\alpha^2 = 20$. Donc $\mathcal{R}_{10\%} = \{W_X \leq 7\} \cup \{W_X \geq 20\}$.
Cherchons les rangs des éléments de X : comme $X = (19, 27, -1)$ et l'échantillon réordonné est $(-4, -1, 3, 6, 7, 13, 19, 27)$, les rangs correspondant aux éléments de X sont 2, 7 et 8. Leur somme vaut 17. Nous ne sommes donc pas dans la zone de rejet du test $F \neq G$.

Tests de Kolmogorov-Smirnov

Avant d'aborder ce sujet, il faut parler de la notion de fonction de répartition empirique : on sait que la fonction de répartition d'une variable aléatoire réelle est définie par : $F(x) = P(X \leq x)$. Cette fonction peut être "approchée" par la fonction de répartition empirique, définie par :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}},$$

avec n le nombre d'individus dans l'échantillon.

Proposition 2.3.7. *À x fixé, $nF_n(x)$ suit une loi binomiale de paramètres n et $F(x)$.*

Proposition 2.3.8. *À x fixé, $F_n(x) \longrightarrow F(x)$ presque sûrement quand $n \to +\infty$.*

Démonstration. Par la loi des grands nombres □

Théorème 2.3.9 (Théorème de Glivenko-Cantelli)**.** *Presque sûrement, la fonction de répartition empirique F_n converge uniformément vers la fonction de répartition F. Autrement dit :*

$$\|F_n - F\|_\infty = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to+\infty]{p.s} 0$$

Adéquation à une loi :

Théorème 2.3.10 (Théorème de Kolmogorov-Smirnov)**.** *1)*$\sqrt{n}\|F_n - F\|_\infty \xrightarrow[n\to+\infty]{loi} D$*, avec D telle que* $P(D \geq \lambda) = 2\sum_{k=1}^{+\infty}(-1)^{k+1}e^{-2k^2\lambda^2}$

2) $\forall n \in \mathbb{N}$, $P(\sqrt{n}\|F_n - F\|_\infty \geq \lambda) \leq 2e^{-2\lambda^2}$ *(on encadre les composantes horizontales de la fonction de répartition empirique). On peut en déduire un intervalle de confiance pour F :*

$$F_n - \frac{\lambda}{\sqrt{n}} \leq F \leq F_n + \frac{\lambda}{\sqrt{n}} \quad \textit{avec probabilité } 1 - 2e^{-2\lambda^2}$$

On va maintenant vraiment parler des tests de Kolmogorov-Smirnov :

Il y a deux types de tests de Kolmogorov-Smirnov : un test pour vérifier qu'un échantillon suit une loi de référence (connue grâce à sa fonction de répartition $P(X \leq t)$) ; puis un test pour vérifier si deux échantillons suivent la même loi (toujours grâce à leur fonction de répartition).
La fonction de répartition de référence est continue.
Concernant le premier test, qui est un test d'adéquation à une loi, il permet de tester :

$$\begin{cases} H_0 : F = F_0 \\ H_1 : F \neq F_0, \end{cases}$$

avec F_0 la fonction de répartition d'une loi de référence : c'est la loi que l'on suppose que notre échantillon suit.
La statistique de test que l'on va utiliser sera :

$$D_n = \sqrt{n}\sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)|.$$

En effet, plus la fonction de répartition empirique s'éloigne de la fonction de répartition de la loi de référence, plus on aura de chances de conclure que l'échantillon ne suit pas cette loi : sous H_1, $\exists x_0 \in \mathbb{R}$ tel que $F(x_0) \neq F_0(x_0)$.

La zone de rejet sera : $\mathcal{R}_\alpha = \{D_n \geq t_\alpha\}$, t_α à déterminer soit par une table, un logiciel de type SAS ou une loi limite.

Théorème 2.3.11. *La loi de D_n sous H_0 ne dépend pas de F.*

Démonstration. Nous avons besoin du lemme suivant pour démontrer le 1) du théorème [7] :

Lemme 2.3.12. *Soient $(X_i)_{i\geq 1}$ une suite de variables aléatoires identiquement distribuées et indépendantes, de fonction de répartition F, et $(U_i)_{i\geq 1}$ une suite i.i.d de variables aléatoires à valeurs dans $]0,1[$. On va noter F_n la fonction de répartition empirique construite sur $X_1, \ldots, X_n$, et F^{-1} l'inverse généralisé de F défini sur $]0,1[$ par $F^{-1}(u) = \inf\{x \in \mathbb{R}; F(x) \geq u\}$. Pour tout i, on pose $Y_i = F^{-1}(U_i)$ et on note H_n la fonction de répartition empirique construite sur $Y_1, \ldots, Y_n$. Alors pour tout $\omega \in \Omega$,*

$$|H_n(\omega, .) - F|_\infty \leq \sup_{t\in[0,1]} |\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i(\omega)\leq t\}} - t|.$$

De plus, il y a égalité si $]0,1[\subset F(\mathbb{R})$, autrement dit si F est continue sur $\mathbb{R}$.

Comme F est continue dans le cas du test de Kolmogorov-Smirnov, on a égalité dans la formule ci-dessus.
Donc sous l'hypothèse H_0, la loi de D_{n_1,n_2} ne dépend pas de F car les images des X_i par F sont des variables aléatoires de loi uniforme $\mathcal{U}(0,1)$.

□

Ici, nous allons discuter en détail du deuxième test, celui de comparaison de deux échnatillons, car c'est celui-là que nous allons utiliser.
Soient $(X_i)_{1\leq i\leq n_1}$ et $(Y_i)_{1\leq i\leq n_2}$ deux échantillons de variables indépendantes et identiquement distribuées, soyant indépendants. Soient F et G leur fonction de répartition respective, supposées continues.
Nous allons établir le test de Kolmogorov-Smirnov : le test est :

$$\begin{cases} H_0 : F = G \\ H_1 : F \neq G, \end{cases}$$

et la statistique de test est différente :

$$D_{n_1,n_2} = \sup_{x\in\mathbb{R}} |F_{n_1}(x) - G_{n_2}(x)|,$$

avec $F_{n_1}(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}_{\{X_i\leq x\}}$ et $F_{n_2}(x) = \frac{1}{n_2}\sum_{i=1}^{n_1} \mathbb{1}_{\{Y_i\leq x\}}$.
Sous H_0, $|F_{n_1}(x_0) - G_{n_2}(x_0)| \xrightarrow[n_1\to+\infty]{} |F(x_0) - G(x_0)|$.
Sous H_0, $\sup_x |F_{n_1}(x) - G_{n_2}(x)| \leq \sup_x |F_{n_1}(x) - F(x)| + \sup_x |G_{n_2}(x) - F(x)|$, avec F la fonction de répartition supposée commune. Comme $\sup_x |F_{n_1}(x) - F(x)| \longrightarrow 0$ et $|G_{n_2}(x) - F(x)| \longrightarrow 0$, $|F_{n_1}(x) - G_{n_2}(x)| \longrightarrow 0$ (par le théorème de Glivenko-Cantelli).

Théorème 2.3.13. *1. La loi de D_{n_1,n_2} sous H_0 ne dépend pas de F.*

2. Sous H_0, $\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_1,n_2} \xrightarrow[n_1,n_2\to\infty]{loi} D$, avec D définie comme dans le test précédent.

3. Sous H_1, $\exists x_0 \in \mathbb{R}$ tel que $F(x_0) \neq G(x_0)$. D'où $\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_1,n_2} \xrightarrow[n_1,n_2\to\infty]{} +\infty$.

Démonstration. Preuve de 1) : On réutilise le lemme introduit dans la preuve dans le cas de l'adéquation à une loi, un peu modifié. En plus de $(U_i)_{i\geq 1}$ et des $(X_i)_{i\geq 1}$, on introduit un échantillon $(Z_j)_{(j\geq 1)}$ de fonction de répartition G et $(V_j)_{j\geq 1}$, qui est une autre suite i.i.d de variables aléatoires à valeurs dans $]0,1[$. On pose $W_i = G^{-1}(V_i)$ et G_n la fonction de répartition empirique construite sur les W_i. On aura donc :

$$\|F_n(\omega,.) - G_n(\omega,.)\|_\infty \leq \sup_{t\in[0,1]} \left| \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}_{U_i(\omega)\leq t} - \frac{1}{n_2}\sum_{j=1}^{n_2} \mathbb{1}_{V_j(\omega)\leq t} \right|$$

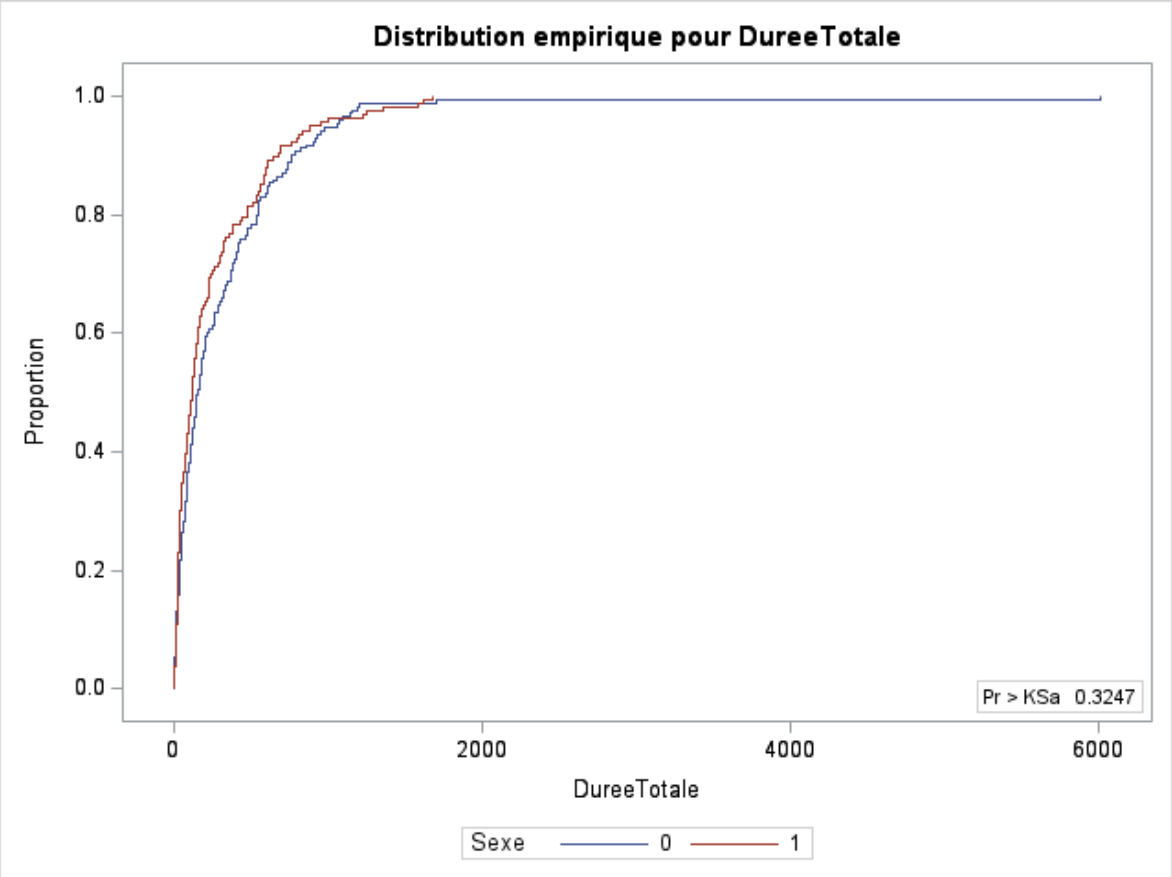
Comme F et G sont continues dans le cas du test de Kolmogorov-Smirnov, on a égalité dans la formule ci-dessus.
Donc sous l'hypothèse H_0, la loi de D_{n_1,n_2} ne dépend pas de F car les images des X_i et des Z_i par F et G sont des variables aléatoires de loi uniforme $\mathcal{U}(0,1)$. □

Note : il existe d'autres tests non-paramétriques, notamment les tests d'Anderson-Darling et Cramer-Von Mises [1] (basés sur les fonctions de répartition empiriques).

2.3.2 Applications

Comme vu dans la partie précédente, le test de Kolmogorov-Smirnov peut servir de comparaison de deux populations. Nous allons donc appliquer directement cet aspect,

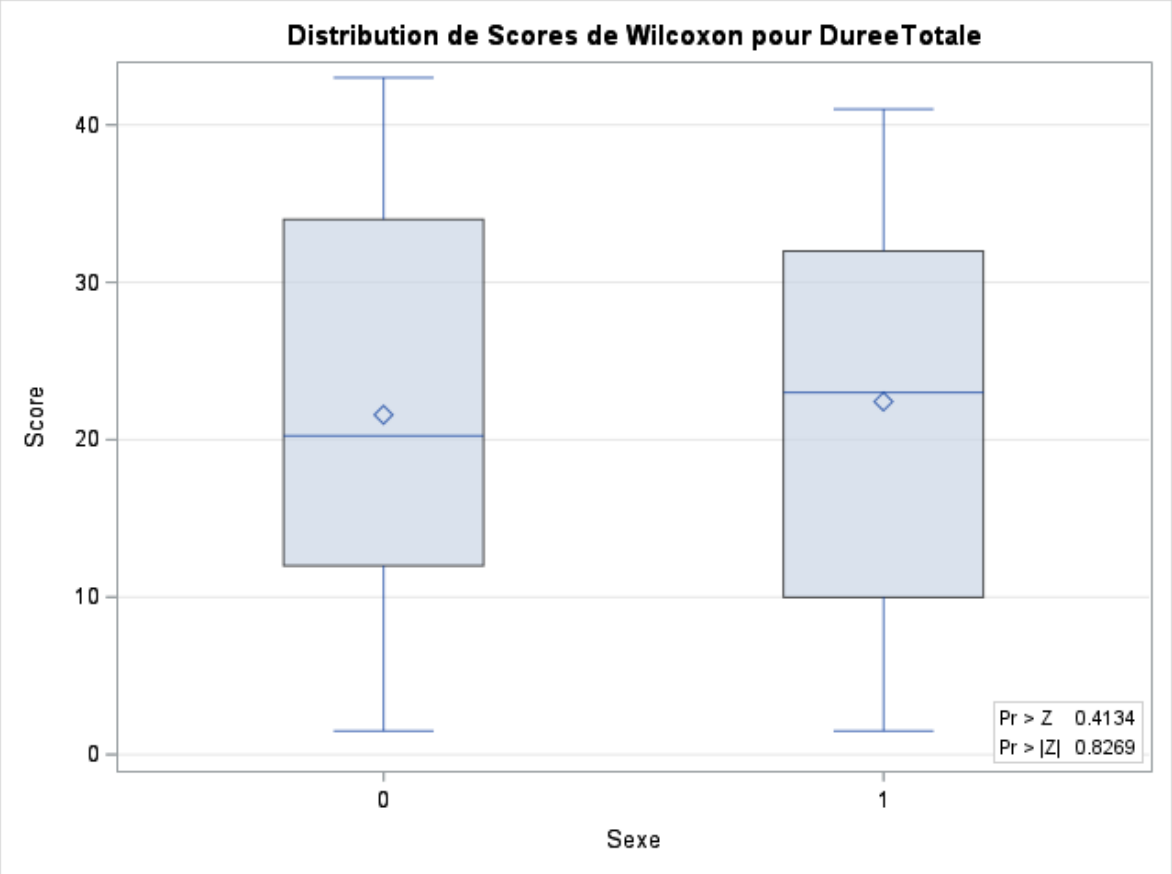
en comparant l'assiduité des garçons et des filles (il faut trouver des variables quantitatives pour avoir de bons résultats) : y a-t-il une différence au niveau de l'assiduité ? Nous allons voir si, parmi les étudiants qui n'ont pas d'autoformation obligatoire, les filles restent plus longtemps au CRL que les garçons, ou inversement (en enlevant ceux qui ne sont resté que zéro minutes).
On utilise la proc npar1way de SAS. [5]
Graphique de la fonction de répartition empirique : en rouge, fonction de répartition empirique de la population des filles ; en bleu, fonction de répartition empirique des garçons :



On peut voir graphiquement qu'il n y a pas beaucoup de différences entre les deux fonctions de répartition empirique. On peut s'attendre à ce que les deux populations suivent la même loi. Vérifions cette hypothèse par des calculs :

Parmi les 326 étudiants de l'échantillon, il y a 170 garçons (code=0) et 156 filles (code=1). D_{n_1,n_2} vaut 0.1056.
La p-valeur, calculée par SAS, est 0.3247.
Avec le test de Kolmogorov-Smirnov, $P - valeur > 0.05$: on ne rejette pas H0
Donc on ne rejette pas le fait qu'il n'y a pas de différence.

Test de Wilcoxon (somme des rangs) :
Test de la somme des rangs sur les individus sans autoformation obligatoire et étant resté 4 heures ou plus au CRL : qui est le plus assidu ?
Dans cet échantillon, il y a 21 filles et 22 garçons. La somme totale des rangs vaut donc 946. On va calculer la somme des rangs des garçons : elle peut aller de 253 (somme des entiers de 1 à 22) à 715 (somme des entiers de 22 à 43).
La somme des rangs calculée par SAS vaut 471. On est à peu près au milieu de ces valeurs extrêmes [6].



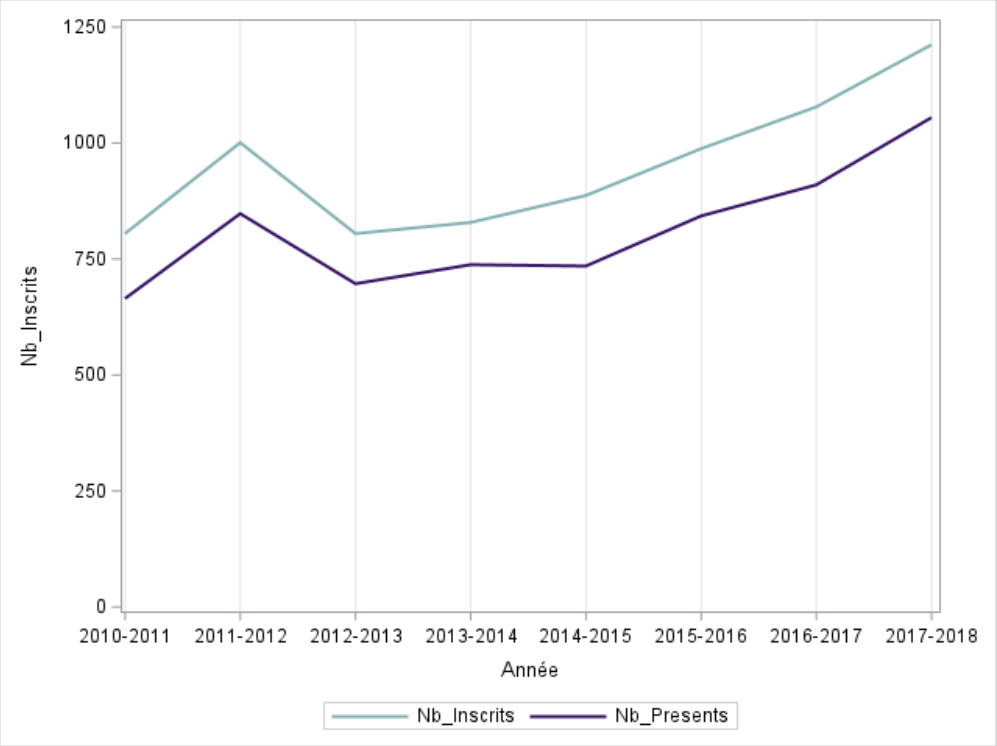
La p-valeur pour le test unilatéral vaut 0.4167. Pour le test bilatéral, elle vaut 0.8334. Dans les deux cas, on ne rejette pas H_0. On voit aussi que les box-plots sont homogènes. On ne peut pas rejeter l'hypothèse disant qu'il n y a pas de différence entre les populations des garçons et celle des filles.

Chapitre 3

Modèles de prediction

3.1 Prédiction des présents en conversations

Tout d'abord, analysons l'évolution de la fréquentation des séances de conversation d'année en année (courbe du haut : nombre d'inscrits ; courbe du bas : nombre de présents) :



Il y a un décalage logique entre le nombre d'inscrits et le nombre de présents : en effet, l'inscription pour les séances de conversation se faisant en ligne, quelquefois plusieurs jours en avance, il se peut que certains étudiants ne puissent pas y participer le jour-même.
Analysons l'évolution proprement dite : on observe un pic de fréquentation en 2011-2012, puis une diminution soudaine et une augmentation régulière du nombre d'inscrits. Le minimum en 2012-2013 s'explique par l'absence de conversations en allemand, alors qu'il y en a les autres années ; l'augmentation régulière, quant à elle, peut s'expliquer par l'augmentation du nombre de lecteurs en anglais (2010 : 3 2011 : 2 2012 : 2 2013 : 3 2014 : 3 2015 : 3 2016 : 4 2017 : 3), puis par une petite augmentation annuelle du nombre total de conversations (plus de demande donc plus de conversations).
Maintenant, passons à la prédiction proprement dite : nous avons en effet effectué une régression linéaire avec les données : peut-on prévoir le nombre de présents en conversation en fonction du nombre d'inscrits ?
Dans une régression linéaire, on va expliquer une variable Y (appelée variable à expliquer en fonction d'autres variables X_j (appelées variables explicatives). Si la relation était parfaitement linéaire, cela se traduirait par des points alignés sur le plan et on pourrait écrire la relation suivante : $Y = b_0 + b_1 X$. Mais, en général, $Y_i = b_0 + b_1 X_i + \epsilon_i$ (les ϵ_i sont appelés résidus : c'est la différence entre valeur prédite et vraie valeur). On suppose que les résidus suivent une loi gaussienne centrée.
Pour déterminer b_0 et b_1, on peut utiliser la méthode des MCO (Moindres Carrés Ordinaires) :

$$(\hat{\beta}_0, \hat{\beta}_1) = \text{argmin}_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{,} (Y_i - \beta_0 - \beta_1 x_i)^2.$$

D'où $\hat{\beta}_0 = \bar{y} - \beta_1 \text{`} \bar{x}$ et $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$.
Notion de R^2 : plus ce coefficient est proche de 1, meilleure est la régression.

$$R^2 = \frac{SCE}{SCT},$$

avec SCE = somme des carrés expliquée $= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$ et SCT = somme des carrés totale $= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$.

Le modèle, produit par SAS, est le suivant :

$$\text{Nombre de présents} = -20.476 + 0.875 * \text{Nombre d'inscrits} + \textit{résidus}$$

Le R^2 vaut 0.9787, c'est donc une régression fiable.
Nous allons vérifier si ce modèle est efficace :

Année	Nombre de présents en vrai	Nombre de présents prédits
2010-2011	665	684
2011-2012	848	855
2012-2013	697	684
2013-2014	738	705
2014-2015	735	756
2015-2016	843	844
2016-2017	910	923
2017-2018	1055	1040

Notre modèle prédit assez bien le nombre de présents en conversations.
Cependant, ce modèle n'a pas beaucoup d'utilité en pratique car il nécessite le nombre total d'inscrits sur une année entière.

Nous allons donc faire une deuxième régression qui va tenter de prédire le nombre de présents en conversations en fonction du nombre d'inscrits sur Moodle, mais séance par séance. Par exemple : combien d'élèves seront présents s'il y a 9 inscrits pour telle ou telle séance ?
Nous allons faire notre régression sur l'ensemble des conversations d'anglais depuis l'année 2010-2011 jusque l'année 2017-2018. Nous aurons donc beaucoup de conversations dans notre base de données (836).
Le modèle, produit par SAS, est le suivant :

$$\text{Nombre de présents} = -0.02721 + 0.85285 * \text{Nombre d'inscrits} + \textit{résidus}$$

Le R^2 de cette régression vaut 0.8958 ; il est assez bon : on peut faire confiance en ce modèle.
Voici ce que ce modèle prédit, en fonction du nombre d'inscrits en séance de conversation anglaise sur Moodle (nous avons arrondi le nombre de présents prédit car cela n'a

pas de sens d'avoir par exemple 3.38 inscrits) :

Inscrits Moodle	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Présents prédits	1	2	3	3	4	5	6	7	8	9	9	10	11	12

Plus il y a d'inscrits en avance sur Moodle, plus il y a de chances que certaines personnes ne viennent pas.

Conclusion

Grâce à ce TER, nous avons pu expérimenter et appliquer de nombreuses méthodes statistiques que nous avons vues en cours. Ainsi, cela nous a donné un aperçu de ce que sera notre activité professionnelle dans le domaine des statistiques.

Cependant, les données à traiter ayant été assez hétérogènes, il nous a été assez difficile de chercher des axes d'études et des applications concrètes à effectuer. Les résultats que nous avons obtenus n'ont pas toujours été satisfaisants d'un point de vue pratique et fonctionnel. Cela nous a confronté à une des tâches principales d'un statisticien : le tri et le traitement de données afin de les rendre exploitables, et surtout la démarche d'essais et d'erreurs (trial and error) concernant l'utilité de l'application d'une méthode donnée.

Un autre aspect important du travail de statisticien est la réponse à la demande d'un client ; ici, le client était le CRL. Nous avons donc fourni des statistiques en rapport avec le CRL. Cependant, nous savons que certains résultats ne seront pas très utiles et ne pourront pas être exploités facilement. Si nous avions eu, par exemple, des données exploitables sur les ressources utilisées par les utilisateurs qui n'ont pas d'autoformation obligatoire, nous aurions pu dresser un portrait plus révélateur de cette population.

Bibliographie

[1] Gilbert Colletaz. Statistique non paramétrique, 2017. https ://www.univ-orleans.fr/deg/masters/ESA/GC/sources/CoursNP.pdf.

[2] Didier DaCunha-Castelle and Marie Duflo. *Probabilités et Statistiques : 1.Problèmes à temps fixe*. Masson, 1989.

[3] Valérie Monbet. Tests statistiques : Notes de cours, 2009. https ://perso.univ-rennes1.fr/valerie.monbet/doc/cours/Cours_Tests_2009.pdf.

[4] SAS. Proc cluster : Crude birth and death rates. https ://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htmstatug_cluster_sect026.htm.

[5] SAS. Proc npar1way : Edf statistics and edf plot. https ://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htmstatug_npar1way_sect021.htm.

[6] SAS. Proc npar1way : Exact wilcoxon two-sample test. https ://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htmstatug_npar1way_sect022.htm.

[7] Charles Suquet. Initiation à la statistique, 2009. http ://math.univ-lille1.fr/ suquet/Polys/IS.pdf.

[8] Jian-Feng Yao. Statistiques empiriques, 2008. https ://perso.univ-rennes1.fr/jian-feng.yao/pedago/modps/empirique.pdf.