République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur

Et de la Recherche Scientifique

جامعة بجاية
Tasdawit n Bgayet
Université de Béjaïa

Faculté des Sciences Exactes

Thème :

Analyse de validité de l'évaluation par les pairs

Pour l'obtention du diplôme de Master II en Mathématiques

Option : Statistiques et Analyse Décisionnelle.

Préparé par :

Mlle AGUECHARIOU Nesrine.

Mlle CHEKKAL Sylia.

Membres de jury:

Mme AMRI Fadhila	*MAA*	*Présidente*
Mr BOURAINE Mohand	*MAA*	*Examinateur*
Mr BOUZIDI L'hadi	*MCB*	*Encadreur*
Mme LAGHA Karima	*MCB*	*Encadreur*

Année universitaire : *2015-2016.*

Remerciements

Nos remerciements s’adressent tout d’abord au bon dieu de nous avoir donné le courage, la volonté, la santé et la patience qui nous ont permis de réaliser ce modeste travail.

Nous adressons ensuite toute notre gratitude à nos deux encadreurs : Mr BOUZIDI et Mme LAGHA, pour leurs disponibilités, leurs patiences, et leurs judicieuses orientations, qui ont contribué à alimenter nos réflexions.

Nous tenons également à remercier l’ensemble des enseignants du département de Mathématiques, qui nous ont transmis des connaissances scientifiques et donnés les outils nécessaires à une meilleure réussite de nos études universitaires.

Nos derniers remerciements vont droit aux amis (es) qui nous ont encouragés en nous témoignant leur sympathie et apportés leur aide morale et intellectuelle tout au long de notre démarche.

Dédicaces

A mon très cher père qui est toujours mon meilleur exemple dans la vie, pour les sacrifices qu'il a consentis pour mon éducation et pour le courage qu'il n'a cessé d'offrir.

A ma mère qui m'a offert la douceur, la tendresse, l'amour et l'affection dont j'ai besoin.

A mes deux frères : Amar et Idris à qui je souhaite des vies pleines de bonheur et de réussite.

A mes sœurs Hakima, Kahina et à mes neveux Axel et Maria.

A mes grands parents.

A mes oncles, tantes cousines et cousins.

A mes chéries : Tata, Rosa, Dida et Sonia.

A Nadir qui m'a beaucoup aidé.

A mes chères amies, qui m'ont créé un milieu de joie et d'ambiance :Samia, Rosa, Nawal, Katia et Amel que j'aime beaucoup.

A ma binôme Sylia.

Dédicaces

À mes trés chérs parents, qui m’ont soutenus tout au long de mes années d’études .

À mes trés chérs fréres et soeurs à qui je souhaite des vies pleines de bonheur et de réussite.

À mes trés chéres belles soeurs.

À tout mes neveux et nièces.

À mes chères amies : Lynda et Fouza.

À tous mes amis sans exception.

À ma binôme Nesrine.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'évaluation des apprentissages est une activité difficile et délicate. Elle est au même temps très importante pour tous les acteurs impliqués dans le processus d'apprentissage. Elle nécessite un temps et un investissement importants pour l'évaluateur. Son caractère répétitif n'est pas rentabilisé et est utilisé uniquement pour produire des notes sans aucun intérêt pédagogique. Ceci a poussé certaines institutions et facultés confrontées à de grands effectifs à avoir recours à des évaluations essentiellement basées sur les questions courtes ou à choix multiples. C'est d'ailleurs, le cas de la faculté de médecine de notre université. Plusieurs chercheurs pensent qu'il y a des solutions à cette situation. La première orientation est l'utilisation de logiciels informatiques dotés d'une intelligence artificielle permettant de corriger de façon automatisée des milliers d'examens composés de questions ouvertes. Ces logiciels commencent déjà à donner des résultats semblables à un correcteur humain dans certains domaines comme celui des langues. La seconde orientation est plutôt basée sur des évaluateurs humains. Il s'agit de l'évaluation par les pairs. L'idée est de faire corriger les copies d'un examen non pas par l'enseignant, mais par les candidats eux-mêmes. Dans cette situation, se pose alors la question de la validité d'une telle évaluation (est-elle équivalente à celle que ferait un enseignant ?). D'un point de vue formatif, cette méthode est excellente puisqu'il est démontré un grand bénéfice pédagogique dans le fait que les étudiants accomplissent une activité les mettant en position de juge (d'évaluateur), leur permettant ainsi plus de responsabilité, mais aussi les confrontant à des situations de conflit cognitifs les amenant à comparer leurs acquis par rapport à ceux des autres [1], [2]. D'un point de vue sommatif, la réponse à la question de validité de l'évaluation par les pairs est plus nuancée [3]. En effet, à la différence de l'évaluation formative, l'évaluation sommative a une conséquence directe sur l'admission ou non des étudiants en années supérieures.

De plus, ce type d'évaluation a un caractère administratif et officiel qui sont régis par des règles et contraintes strictes crées pour tenter d'assurer, le mieux possible, un jugement fiable et valide sur les candidats.

Le travail qui nous a été proposé s'inscrit dans le courant des recherches s'efforçant à répondre à la question de validité de l'évaluation par les pairs. Notre but est d'analyser des données issues de deux cas d'étude concernant deux expériences d'évaluation par les pairs qui se sont déroulées dans deux contextes très différents. La première expérience s'est déroulée à l'Université de Béjaia en 2006 puis en 2007 sur trois examens du domaine des sciences exactes (informatique et génie électrique). L'effectif de 242 étudiants impliqués dans cette première expérience nous paraît suffisant pour entreprendre cette recherche. Dans cette expérience, nous souhaitons vérifier, si, moyennant quelques contraintes, l'évaluation des pairs étaient digne de confiance. La seconde expérience concerne une émission de divertissement diffusée sur la chaîne « France2 » (télévision française). Dans cette émission, 4 artistes connus jugent, au même temps qu'un public composé de 100 personnes, la prestation d'un candidat jouant un sketch (477 candisats). Dans cette expérience, nous voulons vérifier si le jugement d'un nombre important (100) de personnes non expertes est comparable à celui d'experts reconnus.

- Dans le premier chapitre, nous allons présenter les concepts théoriques nécessaires à la compréhension de notre champ d'étude (définition de l'évaluation par les pairs, ses caractéristiques, validité, fiabilité, . . . ,etc.).
- Dans le deuxième chapitre, nous allons présenter les différents outils statistiques en liaison avec notre recherche (docimologie, détection de valeurs aberrantes, comparaison de deux moyennes,. . . etc).
- Dans le troisième chapitre nous allons appliquer quelques outils statistiques sur les différentes données récoltées (analyse préalable, analyse de validité, interprétations et résultats) afin de répondre à la question fondamentale de ce travail qui est dire si l'évaluation par les pairs est valide (digne de confiance).

1

Concepts théoriques

1.1 Définition de l'évaluation par les pairs

L'évaluation par les pairs est définie comme un dispositif dans lequel les individus estiment la quantité, le niveau, la valeur, l'exactitude, la qualité, ou le succès des produits ou des résultats de l'apprentissage des pairs de statut similaire [4].

L'évaluation par les pairs (fréquemment couplée à l'auto-évaluation) est souvent considérée comme une solution à la charge excessive de l'enseignant dans son rôle d'évaluateur, surtout dans le cas des grands effectifs.

L'évaluation par les pairs est utilisée dans divers contextes, avec ou sans technologies. Elle a eu un succès et un regain d'activité cette dernière décennie dans le contexte du e-learning. Cependant, elle n'a eu une véritable exploitation comme évaluation sommative que depuis l'avènement des MOOC. Mais l'intérêt fondamental de cette approche est tout autre : le changement de perspective de l'étudiant, qui se met à la place de l'enseignant, lui fait développer ses

compétences méta-cognitives ; ainsi l'étudiant apprend comment développer un regard critique sur son travail (aussi bien le processus que le résultat). Les travaux réalisés montrent que les résultats de l'évaluation par les pairs et les notes attribuées directement par l'enseignant sont bien corrélés, à condition que les étudiants soient proprement guidés dans la procédure, avec des exemples et une grille d'évaluation détaillée. Il est donc question de définir des compétences à évaluer et des critères clairement énoncés pour chaque niveau de compétence. C'est bien dans cette perspective que le billet " Évaluer ou être évalué, telle est la question " publié par Amaury Daele sur son blog "Pédagogie universitaire - Enseigner et apprendre dans l'enseignement supérieur" nous propose des pistes d'application de l'évaluation par les pairs avec les étudiants. A. Daele, sur la base des résultats d'une expérimentation réalisée sur le sujet, retient trois points essentiels à une utilisation efficace de l'évaluation par les pairs :

- C'est à l'enseignant de définir, de clarifier et de veiller au respect des critères d'évaluation que les étudiants- évaluateurs vont utiliser pour évaluer un travail spécifique de leurs co-apprenants et s'auto-évaluer.
- Il faut notamment insister sur la qualité des feedbacks à transmettre aux pairs, de manière à ce qu'ils fournissent réellement une aide, mais aussi pour que chaque étudiant s'approprie les consignes d'évaluation et améliore sa production.
- L'enseignant doit être convaincu que les étudiants acquièrent grâce à ce processus des habitudes d'auto-évaluation qui influenceront l'ensemble de leurs apprentissages.

Certes, ce processus ne pourra être valide qu'avec l'assistance et le suivi régulier de l'enseignant à qui revient la tâche de la validation finale. Par conséquent, le recours à l'évaluation par les pairs ne devrait pas se faire d'une manière systématique, mais quand l'activité s'y apprête et quand on aurait suffisamment équipé ses étudiants de ce savoir-faire.

1.2 Principe de l'évaluation par les pairs

Considérons le cas de l'evaluation par les pairs dans le domaine pédagogique. Le principe de l'évaluation par les pairs est très simple. A chaque fois qu'un étudiant soumet sa copie, le système lui envoie plusieurs copies d'autres étudiants à évaluer, et c'est à l'enseignant de définir, de clarifier et de veiller au respect des critères d'évaluation que les étudiants- évaluateurs vont utiliser pour évaluer un travail spécifique de leurs co-apprenants et s'auto-évaluer, il faut aussi

insister sur la qualité des feedbacks à transmettre aux pairs, de manière à ce qu'ils fournissent réellement une aide, mais aussi pour que chaque étudiant s'approprie les consignes d'évaluation et améliore son travail.

1.3 Caractéristiques de l'EPP

Les façons de procéder à l'évaluation par les pairs diffèrent selon les cours, et de nombreux paramètres peuvent varier :

- Le nombre de copies à corriger.
- La période de temps dédié à l'évaluation.
- Les mécanismes incitatifs.
- Le procédé de notation et d'évaluation.
- Le guidage dans l'évaluation.
- L'anonymat des évaluateurs.
- La méthode de calcul de la note finale.

Le nombre de copies à corriger :

Il varie en général entre trois et cinq. L'évaluation est plus précise lorsque le nombre d'évaluateur est plus grand. En revanche, plus de copies, c'est aussi plus de travail pour les étudiants. Ceci amène toujours les enseignants à trouver un compromis entre la charge de travail des étudiants et la précision de l'évaluation.

La période de temps dédié à l'évaluation :

C'est la période durant laquelle l'évaluateur peut consacrer du temps pour corriger. Elle s'étend en général de quelques jours à un peu plus d'une semaine. Laisser peu de temps pour corriger les copies, c'est prendre le risque de voir ces évaluations bâclées. Laisser trop de temps, c'est encourager la tendance à remettre le travail à plus tard, et éventuellement courir le risque de voir l'évaluation des productions empiéter sur le suivi du reste du cours.

Les mécanismes incitatifs :

Il faut souligner que cette évaluation n'est pas du goût de tout le monde. Manque de temps et sentiment d'absence de légitimité sont des raisons avancées de manière récurrente. Pour cette raison, les équipes pédagogiques mettent souvent en place des mécanismes incitatifs pour pousser les participants à s'investir dans l'évaluation. Dans certains cours ceux qui ne souhaitent

pas y prendre part se voient retirer des points sur la note finale du devoir ; à l'inverse ceux qui y participent peuvent recevoir des bonus.

Le procédé de notation et d'évaluation :

Le procédé de notation recouvre un certain nombre de concepts (consignes données aux participants, barème, grilles de notation… etc.). La mise au point du barème est une étape délicate, car le nombre et le choix des critères vont être déterminants dans le déroulement de l'évaluation. Les critères proposés peuvent être assez subjectifs et laisser une grande marge de liberté aux apprenants ou au contraire être relativement directifs. Selon l'objectif pédagogique et la démarche de l'enseignant, le barème et les consignes de notation peuvent être rendues visibles en amont de la soumission du devoir. Cette décision dépend avant tout de l'objectif pédagogique et de la philosophie de l'enseignant en charge du cours.

Le guidage dans l'évaluation :

Les barèmes et les grilles de notation sont fondamentales pour réaliser une évaluation par les pairs de qualité. Cependant, ce sont des aspects, des outils assez rudimentaires, car les participants n'étant pas des examinateurs professionnels, peuvent avoir du mal à choisir une note pour un critère donné si on ne les y forme pas. C'est pour cette raison que certaines équipes décident d'aller plus loin et de réaliser de véritables guides d'assistance à l'évaluation pour chaque devoir. Ces guides vont de simple corrigés-type à des séquences où l'enseignant explique sa méthode de notation, en passant par de véritables mini formations où l'évaluateur compare sa notation à celle de l'enseignant.

L'anonymat des copie :

C'est un élément important car il impacte de manière considérable la qualité de la notation et la nature des commentaires laissés par les évaluateurs. Ce phénomène reste marginal, mais c'est l'un des points d'attention importants lors des évaluations sommatives.

La méthode de calcul de la note finale :

Malgré toutes les précautions que l'on pourra prendre pour rendre l'évaluation acceptable et précise, il est inévitable que les styles de notation diffèrent selon les individus, et que certains notent plus sévèrement que les autres ou lorsque le domaine des réponses attendues est plus étendu (subjectivité de l'évaluateur). Ce qui nous amène à la question de la méthode de calcul de la note finale.

1.4 Mise en oeuvre de l'EPP

La répartition des copies aux correcteurs se fait d'une manière aléatoire. Ce sont des plates-formes spécialisées qui s'en chargent de manière automatique. Par exemple la plate-forme Moodle permet de récolter les textes des étudiants, de les répartir aléatoirement et anonymement, de recueillir les évaluations sur base de critères définis par l'enseignant à l'aide la fonctionnalité "Workshop".

1.5 Outils statistiques utilisés

Outils	**Utilisés par**
Coefficient de corrélation de Pearson	Nancy Falchikov et Judy Goldfinch(2000) [6]. L'hadi Bouzidi et al. (2006) [8]. Luc De Grez et al.(2012) [7]. Andrii Vozniuk et al.(2014) [10]. Remi Bachelet et al. (2015) [9].
Coefficient de discrimination	Nancy Falchikov et al. (2000) [6]. L'hadi Bouzidi et al. (2006) [8].
Effect size	Nancy Falchikov al.(2000) [6]. L'hadi Bouzidi et al. (2006) [8].
Coefficient de corrélation intra-classe	Cho-schunn (2006) [2]. Luc De Grez et al. (2012) [7].
Indice de difficulté(p)	L'hadi Bouzidi et al. (2006) [8].
Coefficient alpha de Croncach	L'hadi Bouzidi et al. (2006) [8]. Tyrone Donnon et al.(2013) [11].
Test de Student	L'hadi Bouzidi et al. (2006) [8]. Tyrone Donnon et al. (2013) [11]. Kenneth David Strang (2013) [12].
d de Cohen	L'hadi Bouzidi et al. (2006) [8]. Tyrone Donnon et al. (2013) [11].
Coefficient de corrélation de Spearman	Ian Jones et al. (2012) [13]. Andrii Vozniuk et al. (2013)[10].
Indice de Kappa (k de cohen)	Andrii Vozniuk et al.(2013) [10].
Indice phi (lambda)	Zainab Abolfazli Khonbi et al.(2012) [14].
ANOVA (analyse de la variance)	Zainab Abolfazli Khonbi et al. (2012) [14].

TAB. 1.1 – Outils statistiques utilisés

1.6 Validité de l'EPP

Une épreuve est valide si elle mesure réellement ce qu'elle est censée mesurer [15], [16]. Pour certifier la compétence d'un étudiant à la réalisation d'une activité, technique ou intellectuelle, il faut le mettre dans une situation concrète. Si, par exemple, pour mesurer la capacité à construire un mur, on utilise un examen dans lequel on demande à l'élève de citer la liste des outils et des matériaux à utiliser et le mode d'emploi, on aura enfreint la qualité de validité de l'évaluation puisqu'on ne pourra pas certifier si l'élève pourra ou non construire un vrai mur. Il faut, dans ce cas, le mettre en situation réelle, ce qui n'est pas toujours faisable.

Peu d'études ont été réalisées sur la validité de l'évaluation par les pairs [2], [6], [17], [18], [19], [20] [21]. Dans le cas d'une évaluation par les pairs réalisée par des élèves de sciences dans un lycée des USA au début des années 2000, Sadler et al. [21] ont remarqué un phénomène plutôt normatif de la notation des élèves : les " bons élèves " ont de moins bonnes notes et les " mauvais élèves " de meilleures que celles données par l'enseignant. La proximité entre les évaluations des élèves est celle de l'enseignant est très variable selon les tâches à réaliser, les critères d'évaluation, l'expérience des élèves, etc. En étudiant les évaluations par les pairs réalisées dans un cours d'algorithmique, Chinn dans [22] affirme que les étudiants évaluent de mieux en mieux, à mesure que le cours avance et qu'il existe un lien fort entre la qualité de leur évaluation et leur performance dans les évaluations ordinaires .

D'autre part, certaines études n'ont pas aboutit à trouver de meilleurs résultats en sciences et sciences de l'ingénieur que dans les autres disciplines ni entre les étudiants de premier et de deuxième cycles [23].

D'autres études estiment qu'en général, la corrélation est importante entre les notations des apprenants et celle de l'enseignant, mais d'un point de vue pédagogique, il n'est pas possible de considérer les notes des élèves comme valides sans une révision de ces notes par l'enseignant [21].

Enfin, certaines études ont rapporté que les notes des enseignants et des apprenants évaluant les mêmes travaux sont très fortement corrélées mais pas suffisamment pour que ce soit jugé satisfaisant d'un point de vue pédagogique [21],[24]. Les débats ne sont pas encore tranchés car on a des résultats contradictoires.

2

Outils statistiques en liaison avec notre recherche

Dans ce chapitre nous allons définir et développer quelques outils statistiques qui sont en liaison avec notre recherches, ses outils sont les suivants :

- Docimologie : indice de difficulté, indice de discrimination et le coefficient alpha de cronbach.
- Détermination des valeurs aberrantes : boîte à moustaches et test de dixon.
- Comparaisons des échantillons par des tests : tests paramétriques et tests non paramétriques.

2.1 Docimologie

Le mot de docimologie, construit sur la racine grecque qui signifie épreuve, désigne la science qui s'occupe de la construction des épreuves d'examen et de concours, de l'analyse de notes, des barèmes...etc.

On a trois indices à calculer :

- Indice de difficulté.
- Indice de discrimination.
- Coefficient alpha de cronbach.

2.1.1 Indice de difficulté p

L'indice de difficulté d'une question donnée est le pourcentage d'échec à cette question [26].

L'indice de facilité m est égal à la moyenne des notes des étudiants sur une question donnée divisée par son barème.

$$m = \frac{\frac{\sum_{i=1}^{n} x_i}{n}}{k}. \tag{2.1}$$

Avec :

- n : Nombre des étudiant ayant répondu à la question.
- x_i : Note obtenu par le $i^{ème}$ étudiant sur la question. $(i = 1, ..., n)$
- k : Le barème de la question.

D'où l'indice de difficulté p est égal à :

$$p = 1 - m. \tag{2.2}$$

- Lorsque p est proche de 0, il témoigne d'une question facile.
- Lorsqu'il est proche de 1, il témoigne d'une question difficile.

2.1.2 Indice de discrimination r

Le coefficient de discrimination « r » (variant de -1 à $+1$) représente la corrélation de Pearson entre les réponses à la question et le total des autres questions [26]. On considère qu'il y a début de discrimination à partir de $r = 0,20$ (Normand, 2001). Les items de difficulté moyenne (p allant de 0,40 à 0,60) maximisent la discrimination. Un coefficient de discrimination r égal à

zéro indique qu'il n'y a aucune discrimination. Lorsqu'il est négatif, il signale une incohérence : les meilleurs étudiants échouent à la question, les plus faibles la réussissent.

L'office de l'évaluation scolaire de l'Université de Washington classe le coefficient de discrimination comme suit :

- $(r > 0,30)$ indique que la question est bien formulée.
- $(0.10 < r < 0.30)$ indique que la question est moyennement bien formulée.
- $r < 0.10$ indique que la question est mal formulée.

2.1.3 Consistance interne (Alpha de Cronbach)

α de Cronbach est une méthode souvent utilisée pour mesurer la fiabilité, par exemple pour quantifier la fiabilité d'un résultat (score). La consistance interne des items d'une épreuve peut être analysée en se servant du coefficient α de Cronbach [29].

Le coefficient α de Cronbach est définit par :

$$\alpha = \frac{j}{j-1}[1 - \frac{\sum_i s_i^2}{s_T^2}]. \tag{2.3}$$

Où :

j est le nombre d'items qui composent l'instrument total.

s_T^2 est la variance de l'instrument dans son ensemble.

s_i^2 est la variance de l'item i.

- Cet indice statistique varie entre 0 et 1 et permet d'évaluer l'homogénéité (la consistance ou cohérence interne) d'un instrument d'évaluation composé par un ensemble d'items qui, tous, devraient contribuer à appréhender le niveau de connaissance ou de compétence sur un thème donné.
- Il traduit un degré d'homogénéité d'autant plus élevé que sa valeur est proche de 1.
- Il n'existe pas de distribution statistique connue permettant de conclure si l'alpha de Cronbach est acceptable ou non.

Les seuils empiriques issus de la psychométrie servent de référence : On considère pour une étude exploratoire, que l'alpha de Cronbach est acceptable s'il est compris entre 0,6 et 0,8 [30], [31]. L'office de l'évaluation scolaire de l'Université de Washington classe le coefficient de fiabilité comme suit :

- $\alpha > 0.90$: excellente fiabilité ; elle est au niveau des meilleurs tests standards.
- $0.80 < \alpha < 0.90$: très bonne pour un test en classe.
- $0.70 < \alpha < 0.80$: bonne pour un test en classe ; il y a probablement peu d'items qui nécessiteraient des améliorations.
- $0.60 < \alpha < 0.70$: un peu basse ; le test nécessite d'être accompagné d'autres mesures (autres tests) pour déterminer la note. Il y a probablement peu d'items qui nécessiteraient des améliorations.
- $0.50 < \alpha < 0.60$: suggère une révision du test, sauf si le nombre d'items est réduit (moins de 10). Le test doit nécessairement être renforcé par d'autres mesures (tests) pour déterminer les notes.
- $\alpha < 0.50$: Fiabilité problématique. Il faut le réviser.

2.2 Identification des valeurs aberrantes

En raison de l'évolution rapide des moyens de collecte automatique des données et de leur traitement informatique, le problème des valeurs aberrantes a pris une importance non négligeable durant les dernières décennies. La présence de valeurs anormales peut alors conduire à des estimations biaisées des paramètres des populations et, suite à la réalisation de tests statistiques, à une interprétation des résultats qui peut être erronée. Avant d'exposer des concepts relatifs aux valeurs aberrantes, il est nécessaire de les définir de manière plus précise. De nombreux auteurs ont cherché à décrire le terme de valeur aberrante et les définitions fournies ont évolué au cours du temps. Grubbs [32] définit une valeur aberrante comme étant une observation qui semble dévier de façon marquée par rapport à l'ensemble des autres membres de l'échantillon dans lequel il apparaît. Carletti [33] s'intéresse aux valeurs anormales qu'il définit comme étant une valeur qui paraît suspecte parce qu'elle s'écarte d'une façon importante des autres valeurs de la variable étudiée, ou ne semble pas respecter une norme ou une relation bien définie. Munoz-Garcia et al. dans [34] proposent également une définition du terme valeur aberrante et tentent d'éviter le côté subjectif en ajoutant la condition que l'observation devrait dévier nettement du comportement général par rapport au critère sur lequel l'analyse est réalisée.

Barnett et al. dans [35] définissent une valeur aberrante dans un ensemble de données comme étant une observation (ou un ensemble d'observations) qui semble être inconsistante

avec le reste des données ou d'une autre manière, il y a une valeur aberrante lorsque l'une ou l'autre observation d'un ensemble de données, détonne ou n'est pas en harmonie avec les autres observations.

D'une manière générale, l'objectif d'une méthode statistique destinée à l'examen de valeurs aberrantes de nature aléatoire est de fournir des moyens pour vérifier si une déclaration subjective de la présence d'une valeur aberrante dans un ensemble de données possède des implications objectives importantes pour l'analyse future des données. La bonne attitude à avoir est la suivante : si l'on a pu retrouver la cause de la valeur aberrante (erreur de lecture, faute de calcul...etc), il est tout à fait normal de l'éliminer, en revanche, si aucune cause accidentelle n'a pu être détectée, il est dangereux d'éliminer brutalement la valeur incriminée. Dans ce cas, il faut avoir recours à un test statistique permettant de justifier l'élimination de la valeur aberrante avec une probabilité P, choisie à l'avance, de se tromper.

On applique le test de dixon et la boite à moustache.

2.2.1 Test de Dixon

Le test de Dixon, est parmi les tests qui détectent les valeurs aberrantes. Supposons qu'une expérimentation ait conduit à n observations $(n \geq 3)$. On classe ces observations par ordre croissant :

$$x_1 < x_2 < \ldots < x_{n-1} < x_n. \tag{2.4}$$

où x_i est la $i^{ème}$ valeur obtenue dans la série dotée. Le test permet alors de tester si la première valeur x_1 ou la dernière valeur x_n est aberrante [36]. Pour se faire, suivant le nombre d'observations, on calcule les rapports suivants : Les formules de Q_a et Q_b dépendent de la taille des observations n telles qu'elles sont donné dans la table suivante (référence)

$3 \leq n \leq 7$	$Q_a = \frac{x_2 - x_1}{x_n - x_1}$	$Q_b = \frac{x_n - x_{n-1}}{x_n - x_1}$
$8 \leq n \leq 10$	$Q_a = \frac{x_2 - x_1}{x_{n-1} - x_1}$	$Q_b = \frac{x_n - x_{n-1}}{x_n - x_2}$
$11 \leq n \leq 13$	$Q_a = \frac{x_3 - x_1}{x_{n-1} - x_1}$	$Q_b = \frac{x_n - x_{n-2}}{x_n - x_2}$
$14 \leq n \leq 30$	$Q_a = \frac{x_3 - x_1}{x_{n-2} - x_1}$	$Q_b = \frac{x_n - x_{n-2}}{x_n - x_3}$

TAB. 2.1 – Formules de calcul de Q_a et Q_b

On lit dans la table de Dixon qui donne les valeurs critiques de ces rapports au niveau de risque 10%, 5% et 1%.

La règle à adopter est la suivante : si la valeur du rapport est inférieure à la valeur critique, alors l'élimination de cette observation n'est pas justifiée au risque donné.

2.2.2 Boite à moustache

La boite à moustaches, ou diagramme en boite, (ou encore boxplot en anglais), est un diagramme simple qui permet de représenter la distribution d'une variable [37]. La représentation graphique de la boite à moustaches est mystérieuse lorsqu'on la découvre pour la première fois, sa lecture et son interprétation nécessite de connaître sa construction.

La boite à moustaches utilise 5 valeurs qui résument des données : le minimum (Q_0), les 3 quartiles Q_1, $Q_2(médiane)$, Q_3, et le maximum (Q_4). Selon que l'effectif n des valeurs est pair ou impair, on procédera différemment pour évaluer les quartiles.

Procédure :

1. Classer les n données par ordre croissant.
2. Diviser les données en 2 groupes de tailles égales. On obtient le groupe du bas et le groupe du haut, chacun contenant 50% des observations. Si n est pair : la médiane est la moyenne des 2 points milieu. Si n est impair : la médiane est le point milieu. Dans ce cas il faut, pour permettre les calculs qui vont suivre, reproduire la valeur de ce point dans les 2 groupes.
3. Calculer à nouveau la médiane du groupe du bas. On obtient le quartile Q_1, qui correspond à 25% des observations.
4. Calculer à nouveau la médiane du groupe du haut. On obtient le quartile Q_3, qui correspond à 75% des observations.

L'écart interquartile E (Inter Quartile Range) est utilisé comme indicateur de dispersion. Il correspond à 50% des effectifs situés dans la partie centrale de la distribution.

$$E = Q_1 - Q_3. \tag{2.5}$$

Ce diagramme est composé de :

- Un rectangle qui s'étend du premier au troisième quartile. Le rectangle est divisé par une ligne correspondant à la médiane.
- Ce rectangle est complété par deux segments de droites.
- Pour les dessiner, on calcule d'abord les bornes.

$$Minimum = Q_1 - 1.5E$$

$$Maximum = Q_3 + 1.5E$$

- On identifie ensuite la plus petite et la plus grande observation comprise entre ces bornes. Ces observations sont appelées "valeurs adjacentes".
- On trace les segments de droites reliant ces observations au rectangle.
- Les valeurs qui ne sont pas comprises entre les valeurs adjacentes, sont représentées par des points et sont appelées "valeurs extrêmes".

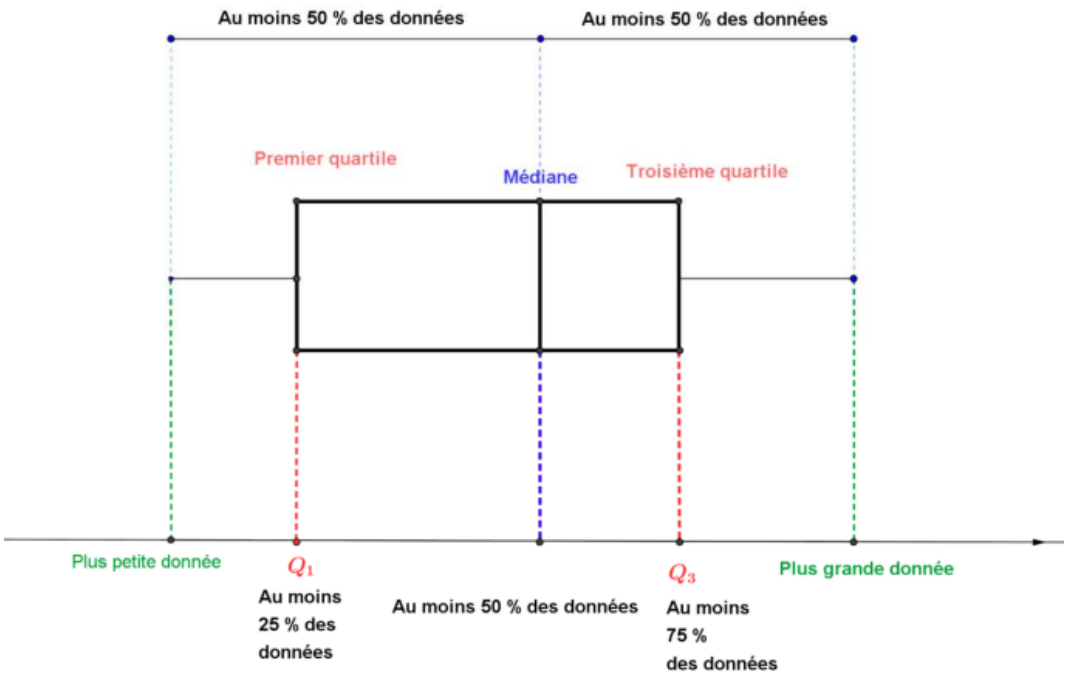

FIG. 2.1 – boite à moustache

2.3 Comparaison des échantillons (cas de deux échantillons)

Ce cas de comparaison est utile lorsque l'on veut établir si deux traitements sont différents ou si un traitement est meilleur qu'un autre.

Tests paramétriques Le terme communément utilisé " tests paramétriques " recouvre les tests statistiques fondés sur des hypothèses sur la loi de distribution (répartition) de la variable étudiée. Il existe de nombreuses lois de distributions que l'on peut résumer par certaines valeurs caractéristiques encore appelées paramètres, d'où ce terme de " tests paramétriques ". Dans la majorité des cas, ces tests paramétriques sont basés sur la loi normale [38].

Tests non paramétriques

Un test non paramétrique est un test qui n'exige pas de distribution particulière de la variable mesurée. Un test non paramétrique est un test dont le modèle ne précise pas les conditions que doivent remplir les paramètres de la population dont a été extrait l'échantillon [25].

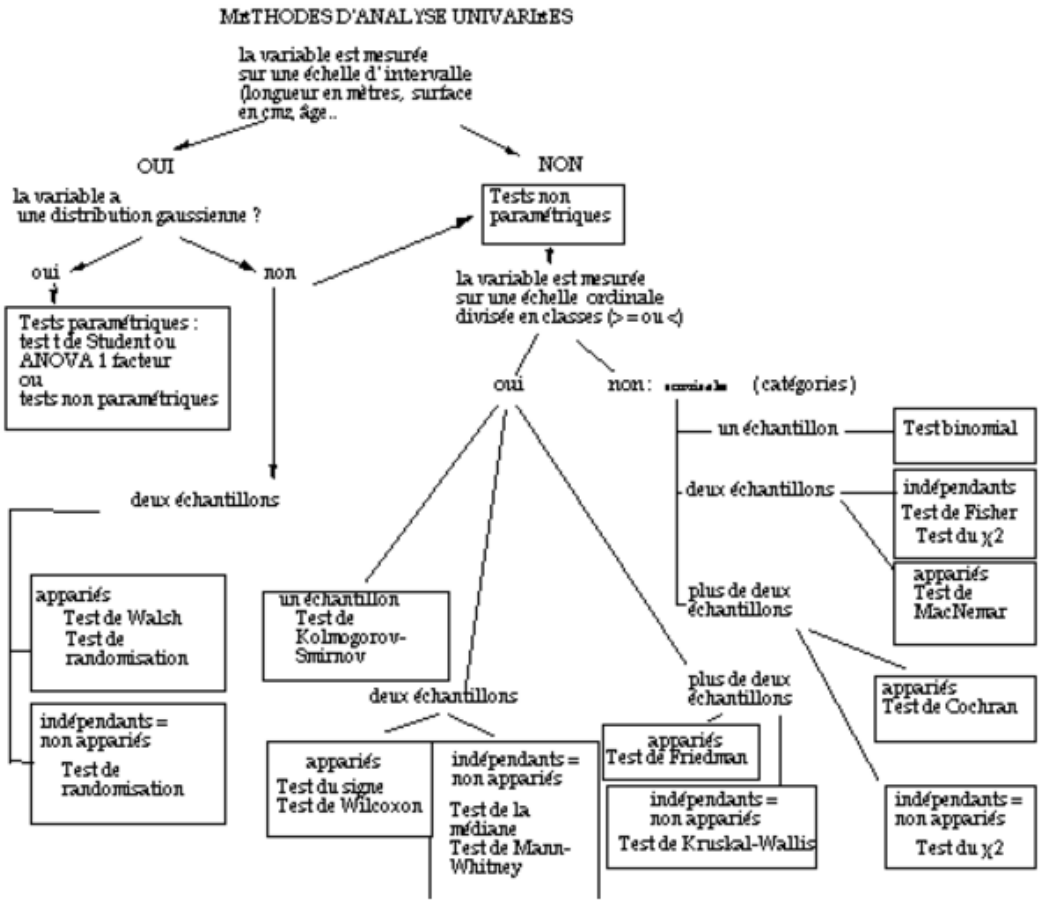

FIG. 2.2 – Schéma montrant comment choisir le test statistique approprié

2.3.1 Echantillons indépendants

Deux groupes sont indépendants si aucun sujet n'appartient aux deux groupes simultanément. On considère donc deux groupes d'individus distincts.

2.3.1.1 Test de student (test paramétrique)

Le test de student est un test d'hypothèse sur la comparaison des moyennes de deux échantillons [41].

1. Les hypothèses que l'on souhaite tester sont :

   Hypothèse nulle : $H_0 : \mu_1 = \mu_2$

   Dans un test d'hypothèse ayant pour but la comparaison de deux échantillons, nous

désirons savoir s'il existe une différence significative entre les moyennes des deux populations dont sont tirés les échantillons.

Hypothèse alternative :

Elle peut prendre les trois formes suivantes :

$H_1 : \mu_1 > \mu_2$ (test unilatéral à droite).

$H_1 : \mu_1 < \mu_2$ (test unilatéral à gauche).

$H_1 : \mu_1 \neq \mu_2$ (test bilatéral).

2. Calcul du ratio :

$$T = \frac{\bar{x_1} - \bar{x_2}}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2.6}$$

où n_1 et n_2 sont les tailles respectives des deux échantillons et S_p est l'écart type pondéré des deux échantillons :

$$S_p = \sqrt{\frac{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x_i})^2}{n_1 + n_2 - 2}}$$

où x_{ij} représente l'observation j de l'échantillon i.

3. Choix du seuil de signification α du test.

4. Comparaison de la valeur calculée de T avec la valeur critique appropriée de t avec $(n_1 + n_2 - 2)$ degrés de liberté. On rejette H_0 si la valeur absolue de $|T|$ est supérieure à cette valeur critique.

Si le test est unilatéral nous prendrons la valeur $t_{n_1+n_2-2,1-\alpha}$ de la table. S'il est bilatéral, nous prendrons la valeur $t_{n_1+n_2-2,1-\frac{\alpha}{2}}$.

Exemple :

Une étude réalisée en vue de comparer la durée de vie moyenne de deux marques de pneus. Un échantillon a été prélevé pour chacune des marques. Les résultats obtenus sont :

– Pour la marque 1 :

$\bar{x_1} = 72000$ km, $S_1 = 3200$ km, $n_1 = 50$.

– Pour la marque 2 :

$\bar{x_2} = 77400$ km, $S_1 = 2400$ km, $n_1 = 40$.

Nous désirons savoir s'il existe une différence significative de durée de vie entre les deux marques de pneus, à un seuil de signification $\alpha = 1\%$. Ceci implique les hypothèses suivantes :

$H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$

L'écart-type pondéré S_p est égal à :

$$\begin{aligned} S_p &= \sqrt{\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij}-\bar{x_i})^2}{n_1+n_2-2}} \\ &= \sqrt{\frac{(n_1-1).S_1^2+(n_2-1).S_2^2}{n_1+n_2-2}} \\ &= \sqrt{\frac{49\times 3200^2+39\times 2400^2}{50+40-2}} \\ &= 2873.07 \end{aligned}$$

Nous pouvons donc calculer le ratio T :

$$\begin{aligned} T &= \frac{\bar{x_1}-\bar{x_2}}{S_p\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \\ &= \frac{72000-74400}{2873.07\times\sqrt{\frac{1}{50}+\frac{1}{40}}} \\ &= -3.938 \end{aligned}$$

Le nombre de degrés de liberté associé au test est égal à :

$$v = n_1+n_2-2 = 88$$

La valeur de $t_{v,1-\alpha}$ trouvée dans la table de student, pour $\alpha = 1\%$ est : $t_{88,0.995} = 2.634$

Comme la valeur absolue de ration calculé est supérieur à cette valeur, $T = |-3.938| > 2.634$, nous rejetons l'hypothèse nulle au profil de l'hypothèse alternative, et concluons qu'à un seuil de signification de 1%, les deux marques de pneus n'ont pas la même durée de vie moyenne.

2.3.1.2 Tests de Mann whitney (test non paramétrique)

Le tes de Mann Whitney est un test non paramétrique visant à tester l'égalité de deux populations. Il est utilisé lorsqu'on est en présence des deux échantillons provenant de deux populations [41].

Soit $(X_1, X_2, ..., X_n)$ un échantillon de taille n provenant d'une population 1 et soit $(Y_1, Y_2, ..., Y_m)$ un échantillon de taille m provenant d'une population 2.

On obtient $N = n+m$ observations que l'on classe dans un ordre croissant en notant chaque fois à quel échantillon appartient l'observation.

La statistique de Mann Whitney, notée U, est défini comme le nombre total de fois qu'un X_i précède un Y_j dans la classification par ordre croissant des N observations.

Quand la taille des échantillons est importante, la détermination de U devient longue et fastidieuse mais on peut utiliser la relation suivante qui donne des résultats identiques :

$$U = mn + \frac{n(n+1)}{2} - T. \tag{2.7}$$

où T est la somme des rangs attribués aux X.

Les hypothèses correspondant au test de Mann-Whitney peuvent être formulées de la façon suivante, selon qu'il s'agit d'un test bilatéral ou d'un test unilatéral :

1. Cas bilatéral :

   $H_0 : P(X < y) = \frac{1}{2}$ contre $H_1 : P(X < y) \neq \frac{1}{2}$

2. Cas unilatéral :

   – a. $H_0 : P(X < y) \leq \frac{1}{2}$ contre $H_1 : P(X < y) > \frac{1}{2}$

   – b. $H_0 : P(X < y) \geq \frac{1}{2}$ contre $H_1 : P(X < y) < \frac{1}{2}$

Dans le cas (1), l'hypothèse nulle correspond à la situation où il n'y a pas de différence entre les deux populations. Dans le cas $(2.a)$, l'hypothèse nulle signifie que la population 1 (d'où est tiré l'échantillon des X) est plus grande que la population 2 (d'où est tiré l'échantillon des Y). Dans le cas $(2.b)$, l'hypothèse nulle indique que la population 1 est plus petite que la population 2.

Si $m, n < 12$, on compare la statistique U avec la valeur trouvée dans la table de Mann-Whitney pour tester H_0.

par contre si $m, n \geq 12$, la distribution d'échantillonnage de U approche très rapidement de la distribution normale avec :

– La moyenne :

$$\mu = \frac{m.n}{2}$$

– L'écart-type :

$$\sigma = \sqrt{\frac{mn(m+n+1)}{12}}$$

Et par conséquent, on peut déterminer la signification d'une valeur observée de U par :

$$Z = \frac{U-\mu}{\sigma} \tag{2.8}$$

où Z est une variable aléatoire centré réduite.

En d'autres termes, on peut comparer la valeur de Z ainsi obtenue avec celle de la table normale.

Par ailleurs, il faut savoir que les règles de décision sont différentes selon les hypothèses posées.

C'est ainsi que l'on a les règles (1), $(2.a)$, $(2.b)$ relatives aux cas précédents (1), $(2.a)$, $(2.b)$.

– Règle de décision 1 :

  On rejette l'hypothèse nulle H_0 au seuil de signification α si U est inférieure à la valeur de la table de Mann-Whitney avec les paramètres $n, m, \frac{\alpha}{2}$ ou si U est plus grand que la valeur de la table pour $n, m, 1-\frac{\alpha}{2}$, c'est-à-dire si :

$$t_{n,m,1-\frac{\alpha}{2}} < U < t_{n,m,\frac{\alpha}{2}}$$

– Règle de décision 2.a :

  On rejette l'hypothèse nulle H_0 au seuil de signification α si U est inférieure à la valeur de la table de Mann-Whitney avec les paramètres n, m, α, c'est-à-dire si :

$$U < t_{n,m,\alpha}$$

– Règle de décision 2.a

  On rejette l'hypothèse nulle H_0 au seuil de signification α si U est supérieure la valeur de la table de Mann-Whitney pour $n, m, 1-\frac{\alpha}{2}$, c'est-à-dire si :

$$U > t_{n,m,1-\alpha}$$

2.3.1.3 Test de Kolmogorov-smirnov (test non paramétrique)

Le test de kolmogorov-smirnov est un test non paramétrique qui vise à determiner si les fonctions de répartition de deux populations sont identiques [25].Il est utilisé lorsqu'on est en présence de deux échantillons provenant de deux populations pouvant être différentes. Contrairement au test de Mann-Whitney ou au test de wilcoxon dont l'objet est de détecter des différences entre deux moyennes ou médianes, le test de Kolmogorov Smirnov a l'avantage de prendre en considération les fonctions de répartition dans leur ensemble.

Soient deux échantillons aléatoires indépendants $X_1, X_2, ..., X_n$, un échantillon de taille n provenant d'une population 1 et $Y_1, Y_2, ..., Y_m$, un échantillon de taille m provenant d'une population 2.

Notant respectivement $F(x)$ et $G(x)$ leur fonction de répartition inconnue.

Si $F(x)$ est la fonction de répartition de la population 1 et $G(x)$ la fonction de répartition de la population 2.

Les hypothèses à tester sont les suivantes :

$H_0 : F(x) = G(x)$ pour tout x. $H_1 : F(x) \neq G(x)$ pour au moins une valeur de x.

Le test statistique T_1 est défini comme la plus grande distance verticale entre les deux fonctions de répartition empiriques :

$$T_1 = sup_x |H_1(x) - H_2(x)| \tag{2.9}$$

On rejette H_0 au seuil de signification α si le test statistique approprié T_1 est supérieure à la valeur de la table de Smirnov ayant pour paramètres n, m et $1-\alpha$, c'est à dire si :

$$T_1 > t_{n,m,1-\alpha}.$$

Les exigences préalables à respecter pour pouvoir effectuer le test de Kolmogorov-Smirnov sont les suivantes :

1. Les deux échantillons sont des échantillons aléatoires tirés de leur population respective.
2. Il y'a indépendance mutuelle entre les deux échantillons.
3. L'échelle de mesure est au moins ordinal.
4. Pour que ce test soit exact, les variables aléatoires doivent être continues, sinon le test est moins précis.

Le test de Kolmogorov-Smirnov peut aussi être utilisé comme test d'adéquation. Dans ce cas on est en présence d'un seul échantillon aléatoire tiré d'une population dont la fonction de répartition est $F(x)$. Son but est de determiner si la fonction de répartition inconnue $F(x)$ est en fait une fonction de répartition spécifique et connue $F_0(x)$.

Les hypothèses sont les mêmes que pour le test à deux échantillons sauf que $F(x)$ et $G(x)$ sont remplacées par $F(x)$ et $F_0(x)$.

$H_0 : F(x) = F_0(x)$ pour tout x. $H_1 : F(x) \neq F_0(x)$ pour au moins une valeur de x.

Si $H(x)$ est la fonction de répartition empirique de l'échantillon aléatoire. Le test statistique T_1 est défini comme suit :

$$T_1 = sup_x|F_0(x) - H(x)| \tag{2.10}$$

La règle de décision est la suivante : Rejeter H_0 au seuil se signification α si T_1 est supérieur à la valeur de la table de Kolmogorov-Smirnov ayant pour paramètres n et $1-\alpha$, c'est à dire si :

$$T_1 > t_{n,1-\alpha}$$

.

2.3.1.4 Test de Khi-deux (test non paramétrique)

Le test d'indépendance de Chi-Carre vise à determiner si deux variables observées sur un échantillons sont indépendantes ou non. Les variables étudiées sont des variables qualitatives catégorielles [41].

Le test d'indépendance du Chi-Carre s'effectue sur la base d'une table de contingence.

Considérons deux variables qualitatives catégorielles X et Y.

On dispose d'un échantillon de n observations sur ces variables.

Les observations obtenues peuvent être représentées par un tableau à deux dimensions appelé table de contingence.

Notons n_{ij}, la fréquence observée pour la catégorie i de la variable X et la catégorie j de la variable Y.

X_1	n_{11}	...	Y_{1c}	$n_{1.}$
...	...	...	...	...
X_r	n_{r1}	...	n_{rc}	$n_{r.}$
$Total$	$n.1$	...	$n_{.c}$	$n..$

TAB. 2.2 – Table de contingence de Khi-deux

La première ligne représente : Catégorie de la variable Y.

La première colonne représente : Catégorie de la variable X.

Les hypothèses à tester sont :

H_0 :"les deux variables sont indépendantes" H_1 :"les deux variables ne sont pas indépendantes"

Étapes du test :

1. Calculer les fréquences estimées notées e_{ij} de chaque case de la table de contingence sous l'hypothèse d'indépendance :

$$e_{ij} = \frac{n_{i.} n_{.j}}{n} \tag{2.11}$$

   Où $n_{i.} = \sum_{k=1}^{c} n_{ik}$

   et $n_{.j} = \sum_{k=1}^{r} n_{kj}$.

   Avec c représentant le nombre de colonnes (ou nombre de catégories de la variable X dans la table de contingence) et r, le nombre de lignes (ou nombre de catégories de la variable Y).

2. Calculer la valeur de la statistique χ^2 (Chi-Carre) qui est réalité une mesure de l'écart entre les fréquences observées n_{ij} et les fréquences estimées e_{ij}.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}. \tag{2.12}$$

3. Choisir le seuil de signification α du test et comparer la valeur de χ^2 calculée avec la valeur $\chi^2_{v,\alpha}$ qui peut être obtenue dans la table du Chi-Carre.

Le nombre de degrés de liberté correspond au nombre de cases du tableau pour lesquelles il est possible de donner des valeurs arbitraires ; les autres valeurs étant imposées par les totaux sur les lignes et sur les colonnes. Ainsi, le nombre de degrés de liberté et égal à :

$$v = (r-1)(c-1).$$

4. Si le χ^2 calculé est inférieur au $\chi^2_{v,\alpha}$, de la table, on ne rejette pas l'hypothèse H_0 :on considère que les deux variables sont indépendantes.

   Si, par contre, le χ^2 calculé est supérieur au $\chi^2_{v,\alpha}$ de la table, on rejettera l'hypothèse nulle H_0 au profit de l'hypothèse alternative H_1. On conclura alors que les deux variables ne sont pas indépendantes.

Le test de Khi-deux peut aussi être utilisé comme test d'adéquation :

Soient $X_1, ..., X_n$, un échantillon de n observations.

Les étapes d'un test d'adéquation du chi-deux sont les suivantes :

1. Poser les hypothèses. L'hypothèse nulle sera de la forme

   $$H_0 : F = F_0$$

   où F_0 est la fonction de répartition présumée de la distribution.

2. Répartir les observations en k-classes disjointes $[a_{i-1}, a_i]$.

   On note n_i le nombre d'observation contenues dans la i-ème classe.

3. Calculer les probabilités théoriques pour chaque classe sur la base de la fonction de répartition présumées F_0 :

   $$p_i = F_0(a_i) - F_0(a_{i-1})$$

   .

4. En déduire les fréquences estimées pour chaque classe

   $$e_i = n.p_i$$

   Où n est la taille de l'échantillon.

5. Calculer la statistique χ^2

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$$

Si H_0 est vrai, la statistique χ^2 suit une loi du khi-deux avec v degrés de liberté où :

$$v = (k - 1 - \text{ nombre de paramétres estimés}).$$

6. Rejeter H_0 si l'écart entre les fréquences observées et les fréquences estimées est grand, c'est à dire :

$$\chi^2 > \chi^2_{v,\alpha}$$

La valeur du $\chi^2_{v,\alpha}$ est donnée dans la table du khi-deux pour un seuil de signification α particulier.

2.3.2 Echantillons appariés

Ce cas se présente chaque fois que l'on compare deux méthodes de mesures, en soumettant les mêmes individus à ces 2 méthodes. A chacune des méthodes correspond alors une population de mesures, mais ces populations et les échantillons que l'on peut en extraire ne sont pas indépendants. Il est aussi possible de soumettre les mêmes sujets à deux traitements différents [25].

2.3.2.1 Test de student (test paramétrique)

Le test de student pour observations pairées sert à comparer les moyennes de deux populations, dont chaque élément de l'une des populations est mis en relation avec un élément de l'autre [25].

Par exemple, il peut s'agir de comparer deux traitements, les données étant considérées comme des pairs d'observations (première observation de la paire recevant le traitement 1 et deuxième observation de la paire recevant le traitement 2).

Soit x_{ij} l'observation j pour la paire i ($j = 1,2$) et ($i = 1,2,...,n$).

Pour chaque paire d'observations on calcule la différence :

$$d_i = x_{i_2} - x_{i_1}. \tag{2.13}$$

Puis nous cherchons l'erreur-type estimé de la moyenne des d_i :

$$s_{\bar{d}} = \frac{1}{\sqrt{n}}.S_d \tag{2.14}$$

où S_d est l'écart-type des d_i :

$$S_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}$$

Le test statistique résultant est défini par :

$$T = \frac{\bar{d}}{s_{\bar{d}}}. \tag{2.15}$$

Le test de student pour observations pairées est un test bilatéral.

Les hypothèses sont :

$H_0 : \delta = 0$ contre $H_1 : \delta \neq 0$

où δ représente la différence entre les moyennes des deux populations ($\delta = \mu_1 - \mu_2$).

On accepte l'hypothèse nulle au seuil de signification α si :

$$|T| < t_{n-1,\frac{\alpha}{2}}$$

où $t_{n-1,\frac{\alpha}{2}}$ est la valeur de la table de student avec $n-1$ degrés de liberté.

Exemple :

Supposons que deux traitements soient appliqués sur 10 pairs d'observations. Les données obtenues et les différences correspondantes notées d_i se trouvent dans le tableau suivant :

La moyenne est égale à :

$$\bar{d} = \frac{1}{10}\sum_{i=1}^{10} d_i = \frac{45}{10}$$

L'écart type se calcule ainsi :

$S_d = \sqrt{\frac{\sum_{i=1}^{10}(d_i - \bar{d})^2}{10-1}} = \sqrt{\frac{340.5}{9}} = 6.15$

et l'erreur type : $s_{\bar{d}} = \frac{1}{\sqrt{n}}.S_d = \frac{1}{\sqrt{10}} \times 6.15 = 1.94$

Le test statistique est donc égal à :

$$T = \frac{\bar{d}}{s_{\bar{d}}} = \frac{4.5}{1.94} = 2.32$$

Si l'on choisit un seuil de signification $\alpha = 0.05$, la valeur de $t_{9,0.025}$ est 2.26. Par conséquent, l'hypothèse nulle $H_0 : \delta = 0$ doit être rejetée puisque $|T| > t_{9,0.025}$

Paire i	Traitement 1	Traitement 2	$d_i = x_{i2}$ - x_{i1}
1	110	118	8
2	99	104	5
3	91	85	-6
4	107	108	1
5	82	81	-1
6	96	93	-3
7	100	102	2
8	87	101	14
9	75	84	9
10	108	111	3

TAB. 2.3 – Tableau représentant les deux traitements

2.3.2.2 Test de signes (test non paramétrique)

Il tire son nom du fait qu'il utilise les signes $(+)$ et $(-)$, au lieu de données quantitatives. Il est basé uniquement sur l'étude des signes des différences observées entre les paires d'individus, quelles que soient les valeurs de ces différences [25].

Les seules contraintes de ce test sont que la variable considérée ait une distribution continue et que les échantillons soient appariés.

L'hypothèse nulle H_0 peut s'écrire :

$$H_0 : "P(+) = P(-) = \frac{1}{2}"$$

.

Où $P(+)$: est la probabilité d'observer une différence positive.

et $P(-)$: est la probabilité d'observer une différence négative.

Sous l'hypothèse H_0, le nombre de différences positives (ou négatives) est une variable binomiale, de loi $B(N, \frac{1}{2})$.

Le test permet de comparer, grâce à cette distribution, le nombre observé de signes « plus » (ou « moins ») et le nombre attendu $\frac{N}{2}$.

Quand certaines différences sont nulles, les paires d'observations correspondantes sont

écartées de l'analyse et la valeur de N est réduite en conséquence.

Le test des signes peut être unilatéral lorsque l'on prédit quel signe, $(+)$ ou $(-)$, sera le plus fréquent ; ou bilatéral, lorsque les fréquences des deux signes seront simplement différentes.

Petits échantillons : Lorsque $N < 25$, la table de l'Annexe A donne les probabilités associées aux valeurs de X obtenues, sous H_0.Où x est le nombre des signes les moins fréquents.

Si $P(H_0) = P(X \leq x) <$ à la valeur lu sur la table, alors on accepte H_0.

Exemple : Un bûcheron doit abattre 12 arbres. Accompagné de son ami le géomètre, il lui demande d'estimer la hauteur des 12 arbres avant de les faire tomber. Celle-ci est calculée par une mesure trigonométrique.

Une fois abattus, ces mêmes arbres sont mesurés au sol. Cette seconde mesure est évidemment précise et considérée comme juste. La mesure trigonométrique effectuée par le géomètre était-elle bonne ?

Il s'agit de tester :

H_0 :" Il n'y a pas de différences entre les deux mesures" Contre H_1 :" il y a une différence significative ".

Le seuil de signification est fixé à $\alpha = 5\%$.

Mesures trigonométriques : $20.4, 25.4, 25.6, 25.6, 26.6, 28.6, 28.7, 29, 29.8, 30.5, 30.9, 31.1$

Mesures effectuées au sol : $20.7, 26.3, 26.8, 28.1, 26.2, 27.3, 29.5, 32, 30.9, 32.3, 32.3, 31.7$

Différences entre les mesures : -0.3, -0.9, -1.2, -2.5, 0.4, 1.3, -0.8, -3.0, -1.1, -1.8, -1.4, -0.6

$N = 12$ (nombre de différences non nulles).

$x = 2$ (nombre de différences pour le signe le moins fréquent).

$P(H_0) = P(X \leq 2) = 2 \times 0.019 = 3.8\% \leq \alpha = 5\%$

La probabilité $P(X \leq 2)$ est donnée par la table des valeurs critiques du test binomial (nous ne pouvons pas utiliser les quartiles sur une fonction de test de loi discrète) ; on multiplie par 2 car le test est bilatéral.

Conclusion :

Au niveau $\alpha = 5\%$, nous concluons à H_1. Les deux méthodes de mesures ne donnent pas les mêmes résultats.

Grands échantillons :

Lorsque $N > 25$ (N grand), on peut utiliser l'approximation normale, en faisant intervenir une correction de continuité. Il suffit de calculer la valeur :

$$Z = \frac{X \pm 0.5 - \frac{N}{2}}{\frac{\sqrt{N}}{2}}$$

où : $X + 0,5$ est utilisé lorsque $X < \frac{N}{2}$

et $X - 0,5$ est utilisé lorsque $X > \frac{N}{2}$.

La signification de Z peut être déterminée par référence à la table de l'Annexe B. Cette table donne la probabilité unilatérale d'obtenir des valeurs aussi extrêmes que le Z observé. Pour un test bilatéral, la probabilité donnée par la table 1 doit être doublée.

2.3.2.3 Test des rangs de Wilcoxon (test non paramétrique)

Le test de wilcoxon est un test non-paramétrique. Il est utilisé lorsqu'on est en présence de deux échantillons provenant de deux populations.

Son but est de vérifier s'il existe des différences entre les deux populations sur la base d'échantillons aléatoires tirés de ces populations.

Soit $(X_1, X_2, ..., X_n)$ un échantillon de taille n provenant d'une population 1, et soit $(Y_1, Y_2, ..., Y_n)$ un échantillon de taille m provenant d'une population 2.

On obtient ainsi $N = n + m$ observations que l'on va classer dans un ordre croissant sans tenir compte de l'appartenance aux échantillons. on attribue ensuite un rang de 1 à la plus petite valeur, un rang de 2 à la valeur juste supérieure et ainsi de suite jusqu'au rang N attribué à la plus grande valeur.

On note $R(X_i)$ le rang attribué à X_i, $i = 1, ..., n$.

Si plusieurs observations ont exactement la même valeur, on leur attribuera un rang moyen.

La statistique T du test est définie par :

$$T = \sum_{i=1}^{n} R(X_i) \tag{2.16}$$

Si la taille des échantillons est grande $(m+n > 12)$, on utilise une approximation en utilisant la statistique T_1 supposée suivre une loi normale standard $N(0,1)$:

$$T_1 = \frac{T - \mu}{\sigma} \sim N(0,1) \tag{2.17}$$

Où σ et μ sont respectivement, la moyenne et l'écart type de la variable aléatoire T.
$\mu = \frac{n(N+1)}{2}$ et $\sigma = \sqrt{\frac{mn(N+1)}{12}}$.

Si l'on utilise l'approximation pour les grands échantillons et qu'il ya des rangs ex aequo parmi les N observations, on remplace l'écart type par :

$$\sigma = \sqrt{\frac{mn}{12}(N+1-\frac{\sum_{i=1}^{g} t_i(t_i^2-1)}{N(N-1)})}$$

Où g est le nombre de groupes de rangs aequo et t_i la taille du groupe i.

Hypothèses :

$H_0 : p(X < Y) = \frac{1}{2}$.

$H_1 : p(X < Y) \neq \frac{1}{2}$.

Règle de décision :

On rejettera l'hypothèse nulle H_0 au seuil de signification α si T est inférieure à la valeur de la table de Wilcoxon avec les paramétres n, m et $\frac{\alpha}{2}$ ou si T est plus grand que la valeur de la table pour n, m et $1-\frac{\alpha}{2}$, c'est à dire si :

$$T < t_{n,m,\frac{\alpha}{2}}$$

ou

$$T > t_{n,m,1-\frac{\alpha}{2}}$$

.

Si le test utilise la statistique T_i, la comparaison se fait avec la table normale :

$$T_1 < z_{\frac{\alpha}{2}}$$

ou

$$T_1 > z_{1-\frac{\alpha}{2}}$$

.

Les exigences préalables à l'utilisation du test de Wilcoxon sont les suivantes :

1. Les deux échantillons sont des échantillons aléatoires tirés de leur population respective.
2. En plus de l'indépendence à l'intérieur de chaque échantillon, il ya indépendance mutuelle entre les deux échantillons.

3. L'echele de mesure est au moins ordinale.

Remarques 1 :

– Si le score de l'un des membres d'une paire peut être déclaré "plus grand" que le score de l'autre membre de la même paire (échelle ordinale), le test des signes est applicable.

– Quand les mesures sont réalisées dans une échelle ordinale à la fois dans les paires et entre elles, le test de Wilcoxon doit être utilisé.

Remarques 2 : D'autres tests non paramétriques peuvent être utilisés : test de McNemar, test de Walsh et test de randomization.

– Le test de McNemar peut être utilisé lorsque les données sont mesurées dans l'échelle nominale. Il n'a pas d'équivalent dans le cas de deux échantillons appariés.

– Le test de Walsh est applicable à de petits échantillons $(N < 15)$ quand il est possible d'affirmer que les échantillons observées proviennent de populations symétriques et continues et que les données sont mesurées dans une échelle d'intervalle.

– Le test de randomization n'est applicable que lorsque N est suffisamment petit et que les mesures sont, au moins, dans une échelle d'intervalle. Ce test prend en compte toute l'information des échantillons et il est donc aussi efficace qu'un test de student t.

2.4 Coefficient de corrélation de Pearson

Étudier la corrélation entre deux ou plusieurs variables revient à étudier l'intensité du lien entre ces variables. La mesure de la corrélation linéaire entre deux variables aléatoires se fait par le calcul du coefficient de corrélation linéaire (coefficient de corrélation de Pearson). Le coefficient de corrélation de Pearson permet de détecter la présence ou l'absence d'une relation linéaire entre deux variables aléatoires indépendantes et quantitatives X et Y, ayant une variance finie [39]. Il est définit par :

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_x.\sigma_y}. \tag{2.18}$$

où :

$Cov(X,Y)$ est la covariance des variables X et Y. σ_x est l'écart type de la variable X. σ_y est l'écart type de la variable Y.

Le coefficient de corrélation de Pearson varie entre -1 et $+1$. Son interprétation est la suivante :

- Si r est proche de 0, il n'y a pas de relation linéaire entre X et Y.
- Si r est proche de -1, il existe une forte relation linéaire négative entre X et Y.
- Si r est proche de 1, il existe une forte relation linéaire positive entre X et Y.

Telle que :

- **Une relation linéaire est positive :** si les deux caractères varient dans le même sens.
- **Une relation linéaire est négative :** si les deux caractères varient en sens inverse.

Corrélation	**Négative**	**Positive**
Faible	de -0.5 à 0.0	de 0.0 à 0.5
Forte	de -1.0 à 0.5	de 0.5 à 1.0

TAB. 2.4 – Interprétation du coefficient de corrélation de Pearson.

2.5 Coefficient de corrélation de Spearman

Le coefficient de corrélation de rang (appelé coefficient de Spearman) examine s'il existe une relation entre le rang des observations pour deux caractères X et Y, ce qui permet de détecter l'existence de relations monotones (croissante ou décroissante). Ce coefficient est donc très utile lorsque l'analyse du nuage de points révèle une forme curviligne dans une relation qui semble mal ajustée à une droite. On notera également qu'il est préférable au coefficient de Pearson lorsque les distributions X et Y sont dissymétriques et/ou comportent des valeurs exceptionnelles. Il est généralement désigné soit par la lettre grecque rho (ρ) , ou r_s. Soit $X = (x_1, x_2, ..., x_n)$ et $Y = (y_1, y_2, ..., y_n)$ deux classements observés sur n individus. Par définition, x_i (resp.y_i) désigne le rang de l'observation i pour la variable X (resp. variable Y). On suppose qu'il n'y a pas d'ex-aequo.

En 1904 le psychologue spearman a proposé de définir la corrélation entre deux classements [39]. si on pose $d_i = x_i - y_i$, ce coefficient peut s'écrire :

$$r_s = 1 - \frac{6.\sum_{i=1}^{N} d_i^2}{n(n^2-1).} \tag{2.19}$$

Propriétés et interprétation de $\rho(XY)$ On peut démontrer que ce coefficient varie entre -1 et +1. Son interprétation est la suivante :

- Si r est proche de 0, il n'y a pas de relation monotone entre X et Y.
- Si r est proche de -1, il existe une forte relation monotone négative entre X et Y.
- Si r est proche de 1, il existe une forte relation monotone positive entre X et Y.

3

Application aux données statistiques

Dans ce chapitre, on procède à l'analyse de la validité de l'évaluation par les pairs dans deux cas d'études différents. Nous procédons dans le premier cas d'étude par une comparaison de l'évaluation effectuée par l'enseignant à celle réalisée par les pairs (les étudiants), dans un contexte d'évaluation assistée par ordinateur. Les données utilisées sont issues d'une expérience authentique d'évaluation par les pairs qui a eu lieu à l'Université de Béjaia en 2006 et en 2007. Dans le second cas d'étude, les données sont issues d'une émission télévisée de divertissement diffusée sur "France 2" années $(2011 - 2012)$. Nous avons utilisé pour le traitement des données, le logiciel R.

3.1 Application dans le domaine de l'enseignement

Deux cent quarante deux (242) étudiants ont participé à une expérimentation organisée par notre encadreur (Mr BOUZIDI L'hadi) en 2006 et en 2007, à l'Université de Béjaia. Trois

(3) épreuves différentes du cycle ingénieur du domaine des sciences exactes ont été choisies : Deux épreuves d'architecture des ordinateurs (AO-2006 et AO-2007), destinées aux étudiants de deuxième année ingénieurs en informatique, et une épreuve d'électrotechnique générale (ET-2007), destinée aux étudiants de troisième année ingénieurs en électrotechnique et en électromécanique. Le processus d'évaluation mis en œuvre s'est déroulé selon les phases suivantes :

1. Organiser un examen en classe.
2. Récupérer puis numériser les copies d'examens.
3. Traiter les copies numérisées pour assurer l'anonymat des étudiants.
4. Déposer les copies numérisées dans le campus virtuel pour chaque étudiant (mettre les copies en ligne).
5. Préparer l'activité sur le campus virtuel en indiquant le barème, les critères d'évaluation et divers paramètres pour le calcul des notes.
6. Permettre à l'enseignant de corriger les copies en ligne.
7. Affecter pour chaque copie 4 étudiants pour la corriger.
8. Permettre à chaque étudiant de corriger sa propre copie et 4 choisis au hazard parmi celles de ses camarades.
9. Récupérer les notes attribuées à chaque copie (la note de l'enseignant, les quatre notes des pairs et la note d'auto-évaluation).
10. Récupérer les notes données à chaque question pour chaque copie de chacune des épreuves (épreuve AO-2006, épreuve AO-2007, épreuve ET-2007).
11. Remettre les notes et les feedback aux étudiants et ouvrir les discussions.

Ce processus a été réalisé à l'aide de la plateforme d'elearning Moodle.

Voici un exemple d'aperçu d'une évaluation par les pairs qui a été réalisée. Pour chaque question on a associé une solution, un barème, une consigne de notation et une note maximale, sachant que la note totale de l'épreuve est de 20 points.

Questions	Barème	Consignes de notations
Q1 - 15_{12} vaut combien en base 10?	1	résultat trouvé, donner note complète sinon 0.
Q2 - $200,5_6$ vaut combien en base 12?	1.5	partie entière + partie décimale correct : donner note complète partie entière seulement correcte donnez 0.5 partie entière incorrecte peut importe le reste donner 0 partie entière correcte + partie décimal incomplète: donner 1 point
Q3 - $200,5_8$ vaut combien en base 2?	1	résultat trouvé, donner note complète sinon 0.
Q4 - En virgule flottante normalisée simple précision (IEEE754) codez le nombre suivant : 15,575	1	résultat trouvé, donner note complète sinon 0.
Q5 - Faire, en complément à 2, les 03 opérations...	1.5	compter 0,5 pour chaque opération réalisée à 100% juste (les nombres trouvés en décimal doivent être juste, faire la somme correctement, préciser s'il y a une retenu et un débordement).
...	...	...

FIG. 3.1 – Indication d'une solution claire avec des consignes de notation.

La Figure (3.2) permet de mieux comprendre comment le processus d'évaluation a été réalisé. Cette figure présente une fenêtre comportant deux cadres : un cadre destiné à l'évaluation et un autre permettant l'observation de la copie d'examen à évaluer. Le premier cadre comporte une zone d'identification de la copie d'examen et de l'évaluateur et une zone énumérant les différents éléments d'évaluation (ou question). Pour chaque élément, on permet à l'évaluateur d'indiquer une note et un feedback. La note est indiquée par un clic sur une zone de boutons à option allant de " correcte " jusqu'à " incorrecte ". La valeur de la note associée à un élément d'évaluation est calculée à base d'un barème et d'un coefficient fixé à l'avance lors de la création de l'atelier d'évaluation par les pairs. La note finale de l'examen est calculée automatiquement, par l'addition des notes affectées à tous les éléments d'évaluation multipliées par leurs coefficients respectifs. Le second cadre permet à l'évaluateur d'observer la copie d'examen.

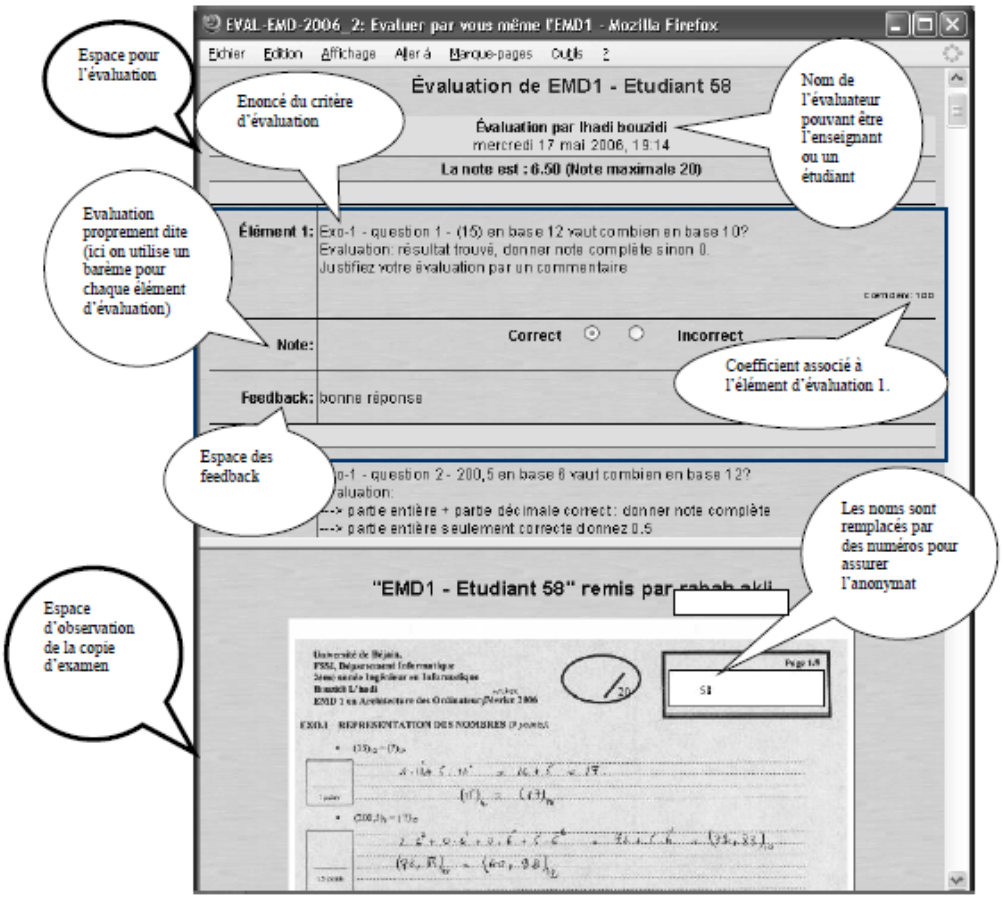

FIG. 3.2 – Exemple d'évaluation par les pairs.

Les copies sont notées sur 20 points. L'épreuve AO-2006 est composée de 17 items, AO-2007 de 13 items et ET-2007 de 9 items.

épreuve		Nombre d'étudiant	Nombre d'item
identifiant	année		
AO-2006	2006	68	17
AO-2007	2007	94	13
ET-2007	2007	80	09

Voici un exemple d'une épreuve ayant eu lieu : Il s'agit de l'examen d'architecture d'ordinateur AO-2007. L'énoncé est clair, le barème est clairement indiqué et les consignes de notation sont explicitement indiquées. (Voir la figure 3.3).

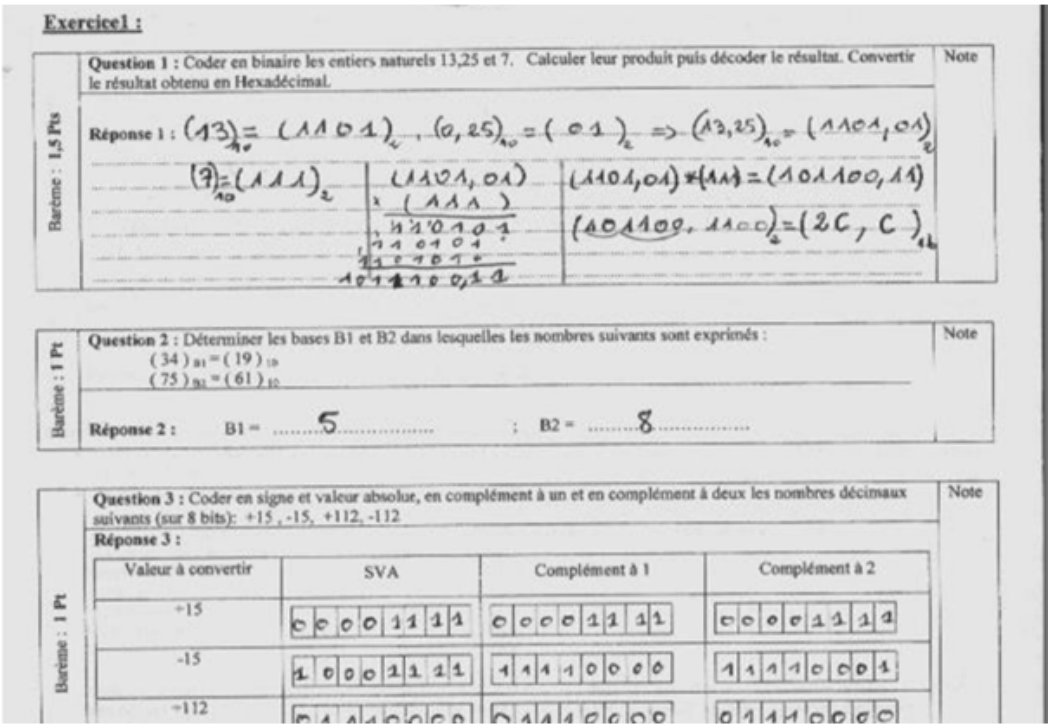

FIG. 3.3 – Aperçu de l'épreuve AO-2007.

Nous avons récupéré, pour chaque copie et pour chaque item, une note d'auto-évaluation, trois à quatre notes attribuées par les pairs et une note attribuée par l'enseignant. Une fois ces données récoltées, nous avons procédé à la sommation des notes associées aux items (questions) pour calculer les notes de tous les évaluateurs.

On note : " NA " les notes produites par une auto-évaluation, et " NE " les notes produites par l'enseignant. Nous avons ensuite calculé la moyenne " NP " et " NPA ", où : " NP " est la moyenne des notes produites par les pairs en ignorant l'auto-évaluation et " NPA " est la moyenne des notes produites par les pairs en incorporant l'auto-évaluation. Voir la figure (3.4).

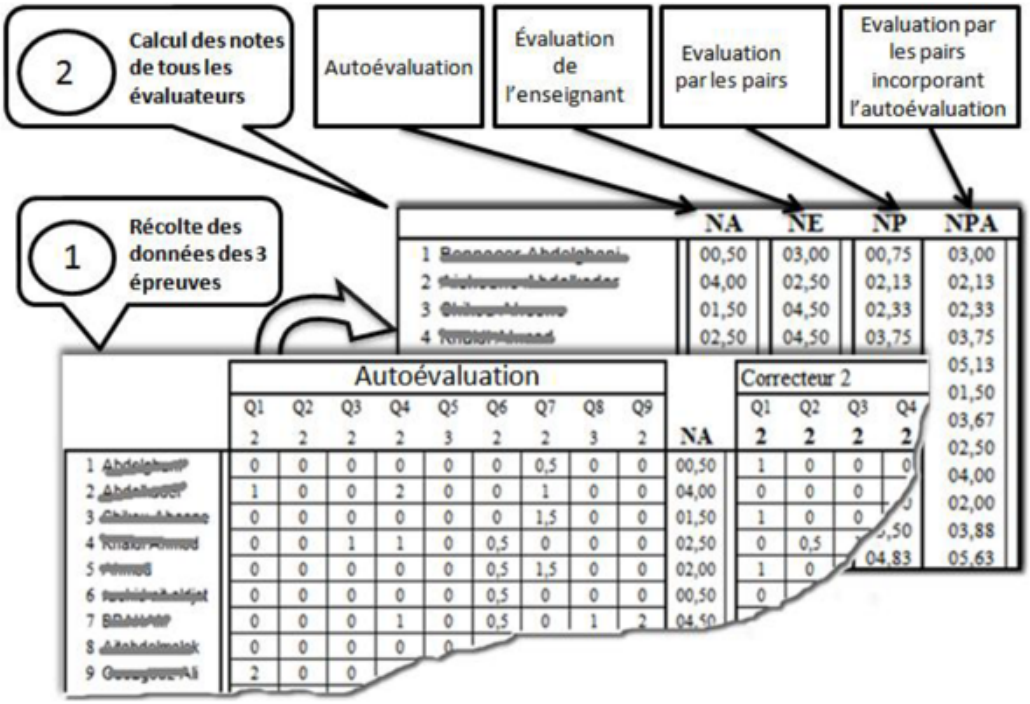

FIG. 3.4 – Récolte et traitement préalable des données.

3.1.1 Traitement statistique des données

Avant d'effectuer l'analyse comparative entre l'évaluation de l'enseignant et celle des pairs nous devons nous assurer de la cohérence à priori des examens, en prenant en compte les anomalies pouvant survenir au niveau de l'énoncé de la question, des consigne de notation ou bien au niveau de la correction de l'enseignant et des pairs. Par conséquent nous proposons de suivre les étapes suivantes :

1. **Analyse d'item :** Cette étape vise à vérifier la validité d'un examen en analysant ses items (questions). Elle permet aussi d'identifier d'éventuelles anomalies au niveaux des items. S'il s'avère qu'un examen, pour une raison ou une autre, n'est pas valide ou comporte des anomalies au niveau de ses questions, il est tout à fait logique que l'évaluation par les pairs qui en découlerait serait biaisée. Pour cela, on calcule les indices suivants, pour chacune des épreuves :
   - Indice de difficulté.
   - Indice de discrimination.
   - Coefficient alpha de cronbach.

Les items comportant des anomalies seront alors éliminées de l'étude. Le nombre d'items à traiter peut alors être reduit.

2. **Détection des évaluations aberrantes :** on utilise le test de dixon pour détecter les évaluations (notes) aberrantes, puis les éliminer.
3. **Analyse de validité de l'EPP** : dans cette étape on utilise les tests de comparaison des moyennes entre deux populations (paramétriques et/ou non paramétriques).
   – Test de Kolmogorve-Smirnov : ce test nous permet de vérifier la normalité (distribution normale) de nos données, qui est une condition nécessaire pour appliquer un test paramétrique (t-test)
   – Test de Student.

   Si les résultats des tests ne sont pas satisfaisant, nous appliquons le test non-paramétrique de Wilcoxon.
4. **Analyser les résultats obtenus et tirer des conclusions**.

Les méthodes de calcul des indices de difficulté, discrimination, alpha de cronbach et la boite à moustache sont dans les figures(3.5), (3.6), (3.7) respectivement.

– Indice de difficulté :

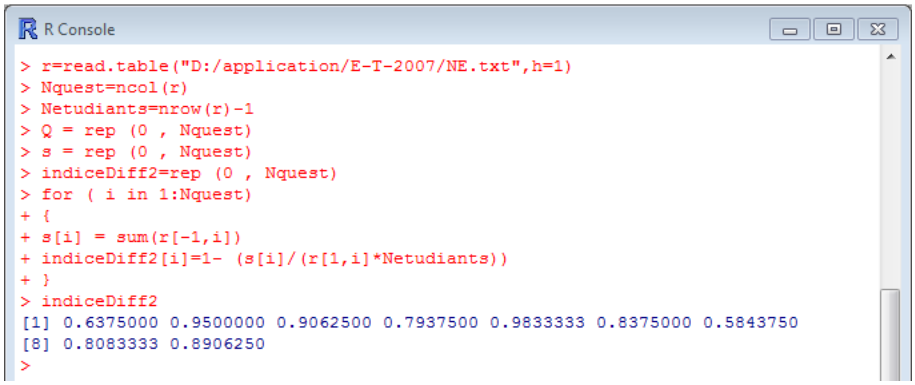

FIG. 3.5 – Indice de difficulté obtenu avec le logiciel R

– Indice de discrimination :

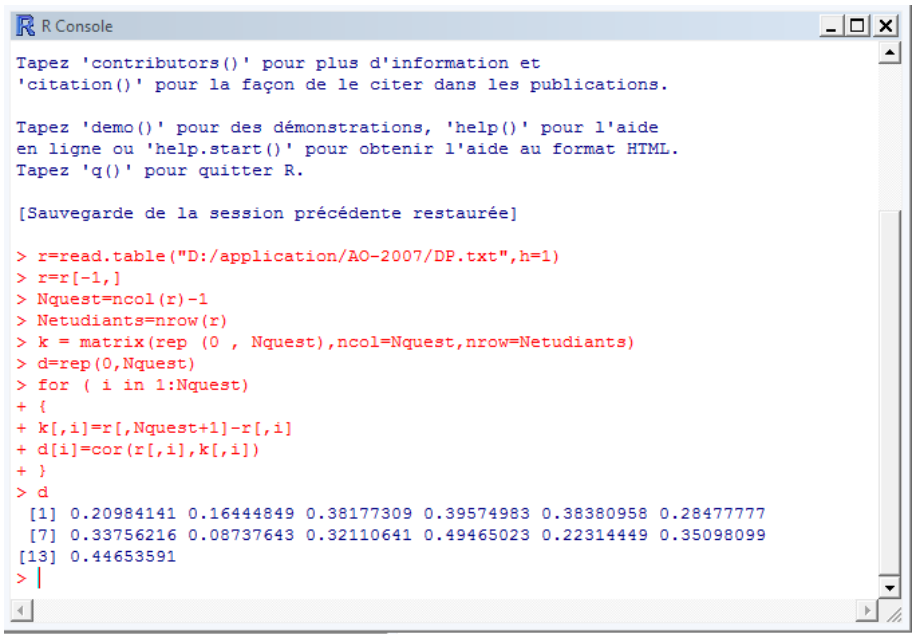

FIG. 3.6 – Indice de discrimination obtenu avec le logiciel R

– Alpha de cronbach :

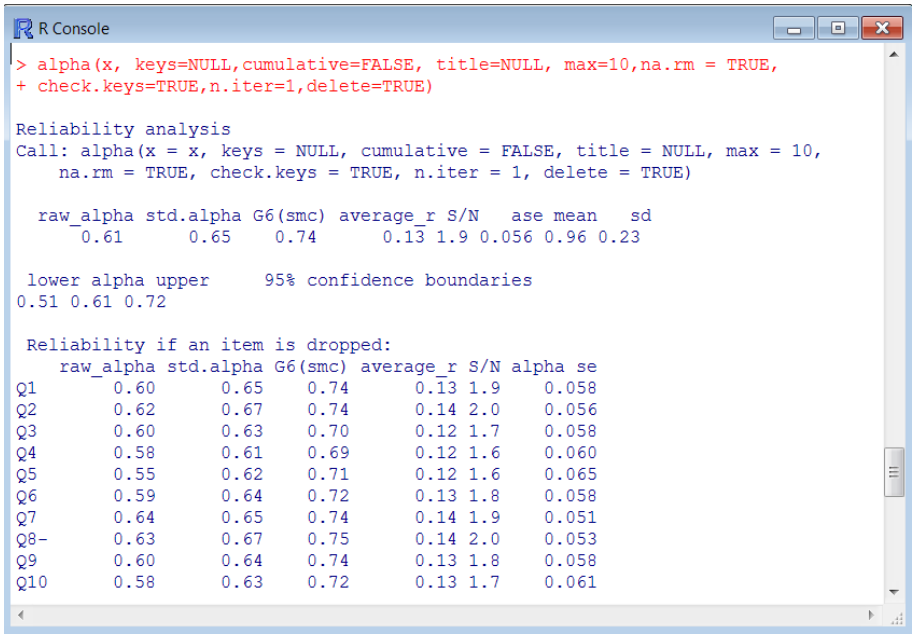

FIG. 3.7 – Alpha de cronbach obtenu avec le logiciel R

– Boite à moustache :

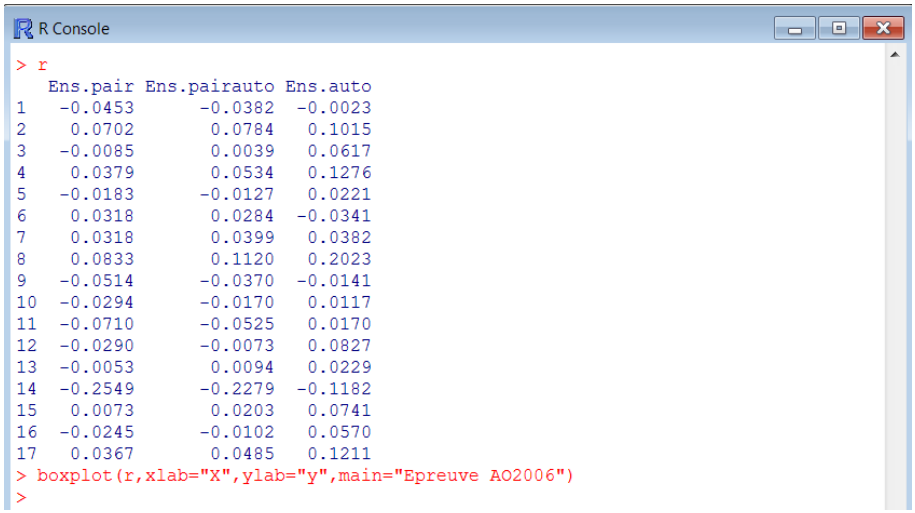

FIG. 3.8 – Boîte à moustache obtenu avec R

Et les résultats correspondants à chaque épreuves sont résumés dans les tableaux (3.1), (3.5), (3.9).

3.1.2 Epreuve AO2006

3.1.2.1 Analyse d'item

Le tableau (3.1) présente à la fois l'indice de difficulté, l'indice de discrimination et le coefficient alpha de cronbach pour chaque item et pour chacune des évaluations suivantes :

– Evaluation par l'enseignant (NE).

– Moyennes des évaluations des pairs sans auto-évaluation (NP).

– Moyennes des évaluations des pairs avec auto-évaluation (NPA).

Notes enseignant				Notes des pairs			Notes des pairs avec autoevaluation		
Item	Diff	Discrim	Alpha	Diff	Discrim	Alpha	Diff	Discrim	Alpha
Q1	0,12	-0,05	0,62	0,16	-0,05	0,56	0,16	-0,05	0,58
Q2	0,59	0,17	0,59	0,52	0,1	0,56	0,51	0,1	0,59
Q3	0,16	0,12	0,60	0,17	0,09	0,55	0,16	0,1	0,56
Q4	0,87	0,25	0,58	0,83	0,23	0,55	0,81	0,22	0,55
Q5	0,64	0,45	0,53	0,66	0,47	0,51	0,65	0,51	0,49
Q6	0,71	0,23	0,58	0,67	0,1	0,56	0,68	0,11	0,57
Q7	0,83	0,22	0,59	0,81	0,31	0,53	0,8	0,28	0,53
Q8	0,88	0,03	0,61	0,8	0,04	0,55	0,77	0,07	0,59
Q9	0,71	0,4	0,55	0,76	0,32	0,49	0,74	0,32	0,52
Q10	0,66	-0,06	0,62	0,69	0,04	0,54	0,68	0,03	0,57
Q11	0,27	0,23	0,58	0,34	0,2	0,55	0,32	0,23	0,52
Q12	0,36	0,15	0,59	0,39	0,08	0,54	0,37	0,15	0,53
Q13	0,6	0,5	0,52	0,61	0,52	0,47	0,59	0,56	0,43
Q14	0,16	0,25	0,58	0,42	0,15	0,54	0,39	0,21	0,53
Q15	0,79	0,19	0,58	0,79	0,28	0,52	0,77	0,26	0,51
Q16	0,9	0,56	0,55	0,92	0,35	0,53	0,91	0,4	0,53
Q17	0,94	0,42	0,58	0,9	0,29	0,52	0,89	0,38	0,52

TAB. 3.1 – Tableau AO2006

	Alpha de cronbach total
NE	0,61
NP	0,55
NPA	0,55

TAB. 3.2 – Tableau des alpha de cronbach total pour l'épreuve AO2006

D'après les résultats du tableau (3.1), on remarque que :

1. **Indice de difficulté " p " :** il est basé uniquement sur l'évaluation de l'enseignant. L'examen comporte : 4 questions faciles $(p < 0.27)$, 4 questions moyennes $(0.36 < p < 0.66)$ et 13 questions difficiles $(p > 0.71)$. Ce qui permet de conclure que l'examen est globalement difficile.
2. **Indice de discrimination " r " :** les valeurs de l'indice de discrimination obtenues pour chaque question dans les trois évaluations sont proches. Nous avons $(r < 0.1)$ pour les questions (Q1, Q8, Q10), cela veut dire que ces questions sont mal conçues, $(0.1 < r < 0.3)$ pour les questions (Q2, Q3, Q4, Q6, Q7, Q11, Q12, Q14, Q15), cela veut dire que ces questions sont acceptables, et $(r > 0.3)$ pour les questions (Q5, Q9, Q13, Q16, Q17), cela veut dire qu'elles sont bien conçues et permettent de distinguer les bons des mauvais élèves.

3. **Coefficient alpha de Cronbach :** en se référant au coefficient de Cronbach total obtenu pour chaque évaluateur, on remarque que ce coefficient varie entre 0.55 et 0.61, ce qui le situe dans un examen plus ou moins acceptable. Cependant, en se référant à l'évaluation de l'enseignant, ce coefficient est de 0.61. Donc, on peut considérer que l'examen AO2006 a une fiabilité acceptable, mais il y'a probablement peu d'items qui nécessitent des améliorations.
4. Pour mesurer la sévérité de chaque correcteur de l'épreuve AO2006, nous proposons de calculer les différences des indices de difficulté de chaque item entre les correcteurs (enseignant-pairs sans auto-évaluation) et (enseignant-pairs avec auto-évaluation). Ce choix de faire la différence des indices de difficulté entre deux évaluateurs est justifié par le fait que cette différence va annuler l'effet de la question et représentera uniquement la différence entre les sévérités des deux évaluateurs. Les résultats sont consignés dans le tableau suivant :

Items	Enseignant/Pairs	Enseignant/Pairs+auto-évaluation
Q1	-0,04	-0,04
Q2	0,07	0,08
Q3	-0,01	0
Q4	0,04	0,06
Q5	-0,02	-0,01
Q6	0,04	0,03
Q7	0,02	0,03
Q8	0,08	0,11
Q9	-0,05	-0,03
Q10	-0,03	-0,02
Q11	-0,07	-0,05
Q12	-0,03	-0,01
Q13	-0,01	0,01
Q14	-0,26	-0,23
Q15	0	0,02
Q16	-0,02	-0,01
Q17	0,04	0,05

TAB. 3.3 – Sévérité AO-2006

Remarque : Les différences des indices de difficulté obtenus pour chaque question et pour chaque deux évaluations sont proches dans l'ensemble, mais on constate un écart important pour la question Q14 entre la sévérité de l'enseignant et celle des pairs (sans et avec auto-évaluation). Ceci veut dire que l'enseignant et les pairs avec et sans auto-évaluation, ont noté cette question d'une manière très différente. Ce qui indique qu'il y'a une anomalie (ou bien une ambiguïté)

au niveau de cette question (cela peut être au niveau de consignes de notations, barème...). Une analyse plus profonde sur cette question est alors nécessaire afin de décider s'il faut ou non la retirer de l'étude et la considérer comme valeur aberrante.

La boite à moustache présentée dans la figure (3.9) nous permet de voir clairement les questions aberrantes :

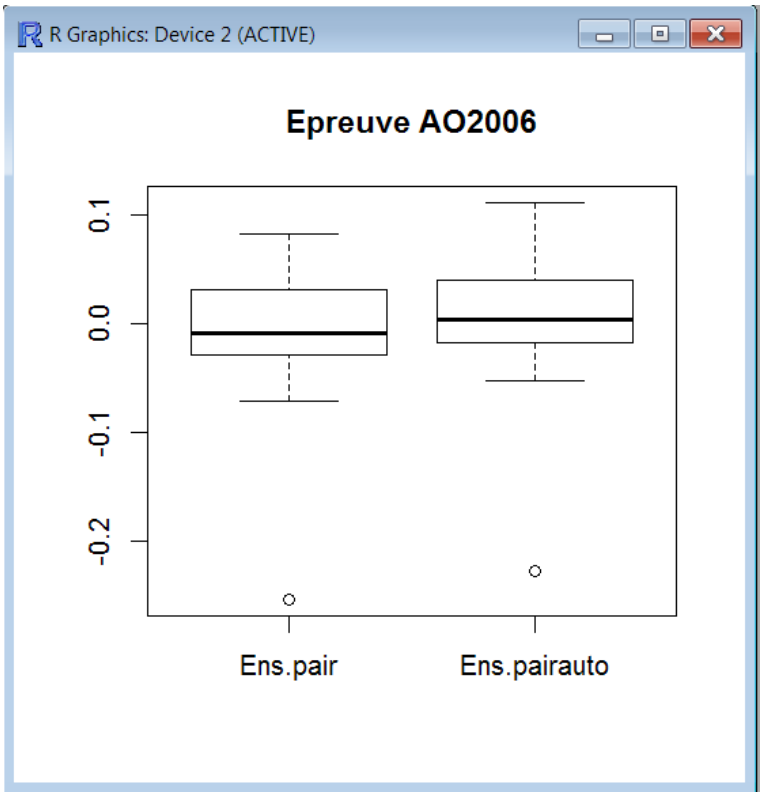

FIG. 3.9 – Boite à moustache AO2006

Ce graphe regroupe les 2 boites à moustache construites à base de différence des indices de difficulté (enseignant-pairs, enseignant-pairs avec auto-évaluation) présenté dans le tableau (3.3). Les valeurs extrêmes (les valeurs qui sont hors de la boite à moustache) sont considérées comme aberrantes.

On remarque qu'il y a 2 valeurs qui sont hors des deux boites à moustache. Ces valeurs correspondent à la question Q14. Afin de nous assurer que Q14 est réellement aberrante, une analyse plus profonde a été opérée avec notre promoteur sur cette question, et effectivement une incohérence dans les consignes de notations a été constatée. De ce fait, cette question doit être retirée de l'étude afin de ne pas biaiser les résultats de notre étude.

3.1.2.2 Détection des évaluations aberrantes

Après avoir écarté les questions aberrantes de notre étude, nous devons aussi éliminer toute évaluation (attribuée par un correcteur, excepté la note de l'enseignant) éloignée des notes attribuées par les trois autres pairs pour une même copie. Ceci dans le cas où la copie est évaluée par 4 évaluateurs. Dans le cas où la copie est évaluée par trois évaluateurs, celle-ci sera écartée et corrigée seulement par l'enseignant. Pour cela nous appliquons le test de dixon.

Voici une méthode d'application sur R :

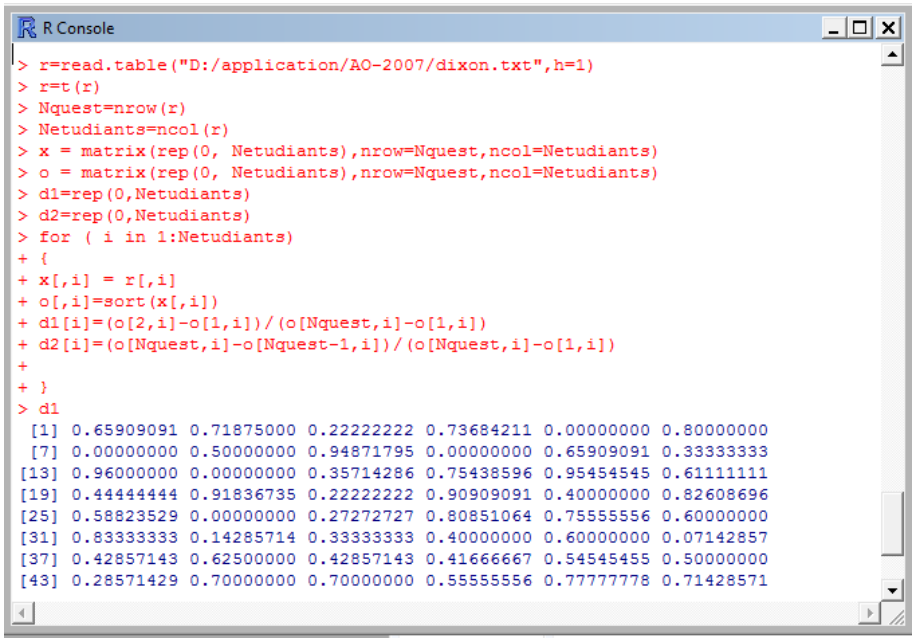

FIG. 3.10 – Tableau AO2006

	Nombre de copies corrigées	Nombre d'évaluations aberrantes
4 Pairs	30	1
3 Pairs	26	9

TAB. 3.4 – Tableau des résultats obtenus avec le test de dixon pour l'épreuve AO2006

Après avoir trié nos données en éliminant toute question et copie pouvant invalider les résultats de notre étude, nous entamons l'analyse de validité des pairs.

3.1.2.3 Validité de l'EPP

Nous considérons qu'on est dans le cas de comparaison de deux échantillons appariés. Nous effectuons deux comparaisons par rapport aux notes données par l'enseignant.

La première comparaison, sera effectuée entre les notes attribuées par l'enseignant, et les moyennes des notes attribuées par les pairs sans auto-évaluation.

La deuxième comparaison sera effectuée entre les notes attribuées par l'enseignant, et les moyennes des notes attribuées par les pairs avec auto-évaluation.

A cet effet, nous allons procéder à l'application du test de student de comparaison des moyennes pour deux échantillons appariées. Mais avant cela, nous devons d'abord vérifier la normalité des échantillons cités ci-dessus.

1. **Test de normalité (Kolmogorov-Smirnov)**

   Nous appliquons le test de Kolmogorov-Smirnov sur chaque variable (évaluation de l'enseignant, évaluation des pairs et évaluation des pairs avec auto-évaluation), pour voir la distribution de chaque variable.

   Nous avons utilisé la droite de Henry pour visualiser la normalité des variables.

   Les résultats trouvés avec le logiciel R pour chaque échantillon sont présentés dans les figures (3.11), (3.12), (3.13), (3.14), (3.15), (3.16), (3.17).

   (a) **Variable NE (Notes de l'Enseignant)** :

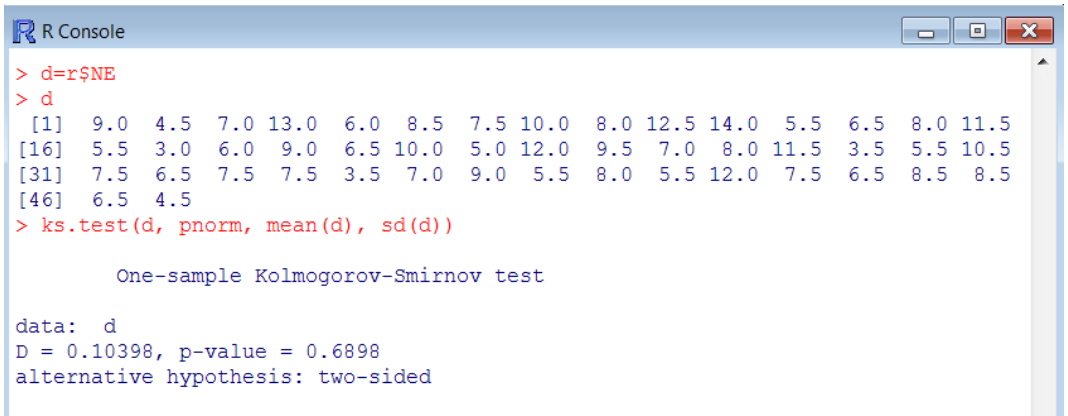

FIG. 3.11 – Test de normalité des notes de l'enseignant

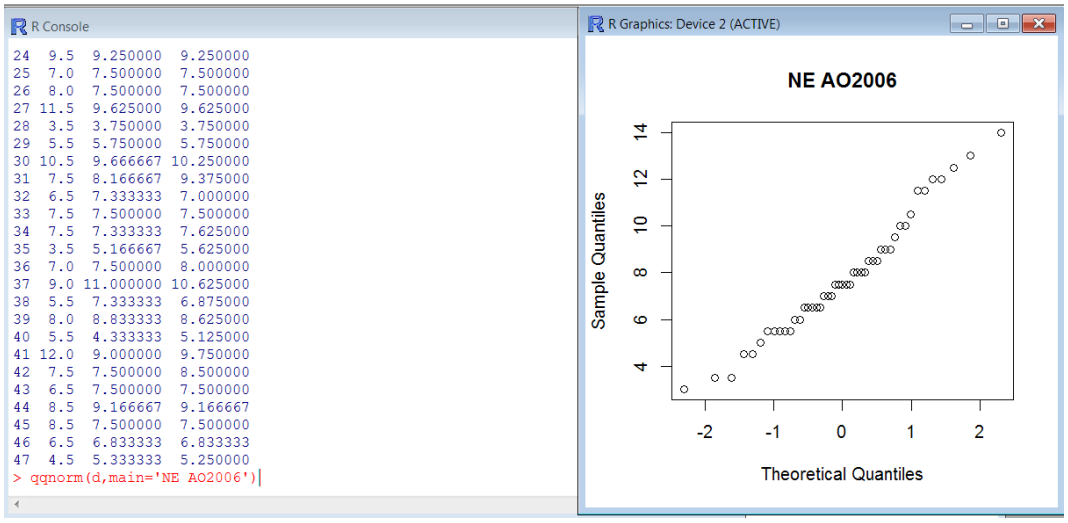

FIG. 3.12 – Droite d'henry des notes de l'enseignant

D'après les résultats obtenus sur R, nous avons :

– $p - value = 0,6899 > \alpha = 0,05 \Rightarrow$ Nous ne rejetons pas l'hypothèse H_0 :"L'échantillon NE a une distribution normale."

– La statistique $D = 0.10398 < k = 0.1983 \Rightarrow$ Nous acceptons l'hypothèse H_0.

Avec k est la valeur lu sur la table de K-S.

On déduit donc que les notes de l'enseignant sont normalement distribuées.

(b) **Variable NP (Notes des Pairs)** :

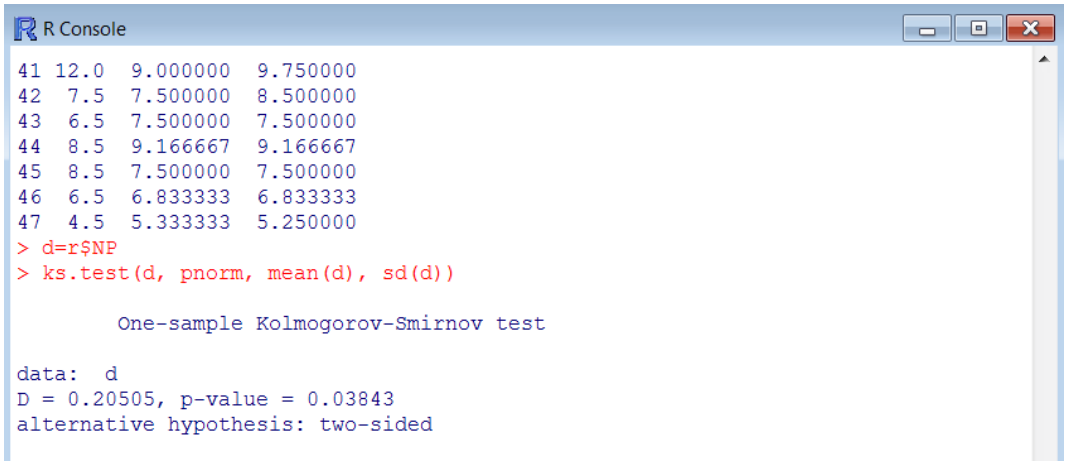

FIG. 3.13 – Test de normalité des notes des pairs

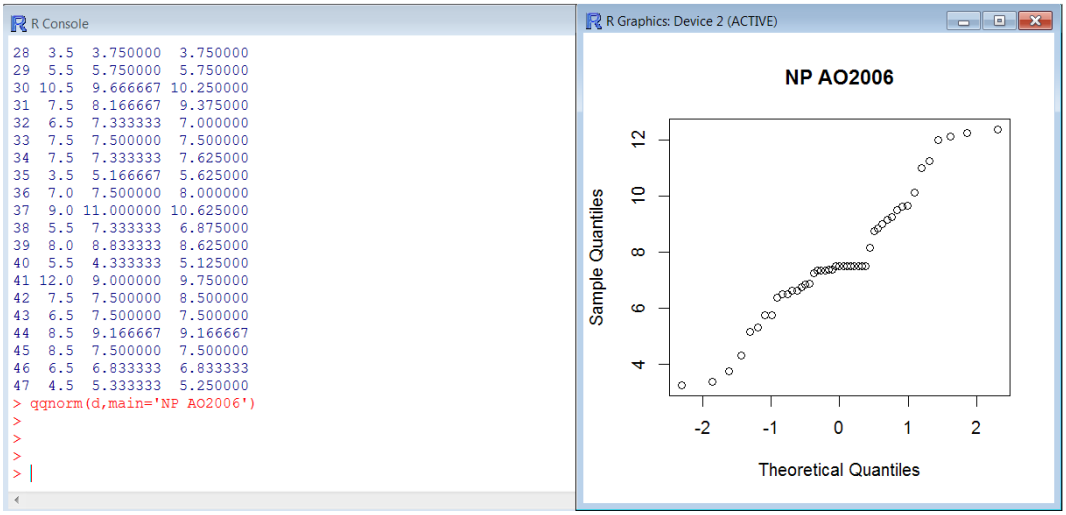

FIG. 3.14 – droite d'henry des notes des pairs

D'après les résultats obtenus sur R, nous avons :

– $p-value = 0,03843 < \alpha = 0,05 \Rightarrow$ Nous rejetons l'hypothèse H_0 :"L'échantillon NE a une distribution normale."

– La statistique $D = 0.2050 < k = 0.1983 \Rightarrow$ Nous rejetons l'hypothèse H_0.

Avec k est la valeur lu sur la table de K-S.

On déduit donc que les moyennes des notes des pairs n'ont pas une distribution normale. Mais le graphe d'henry nous suggère la normalité, alors nous proposons d'appliquer le test de khi-deux pour vérifier la normalité.

les résultats obtenus sur R sont les suivants :

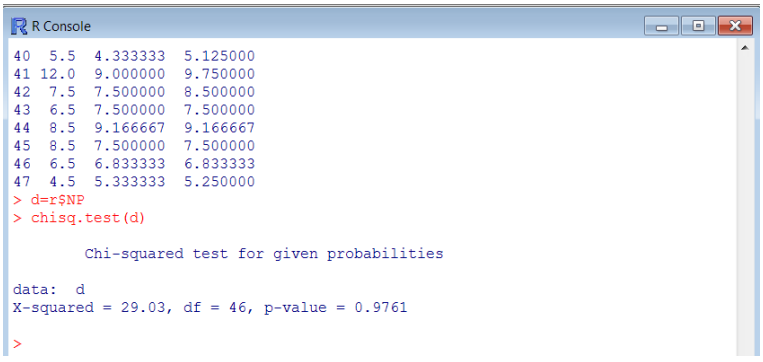

FIG. 3.15 – Test de normalité de khi-deux sur les notes des pairs (Khi-deux)

D'après les résultats obtenus sur R, nous avons :

– $p - value = 0,9761 > \alpha = 0,05 \Rightarrow$ Nous ne rejetons pas l'hypothèse H_0 :"L'échantillon NE a une distribution normale."

– La statistique $X - squared = 29.03 < s = 43.77 \Rightarrow$ Nous acceptons l'hypothèse H_0.

Avec s est la valeur lu sur la table de Khi deux a 46 degrés de liberté..

On déduit donc que les moyennes des notes de l'enseignant sont normalement distribuées.

(c) **Variable NPA (Notes des Pairs avec Auto-évaluation)** :

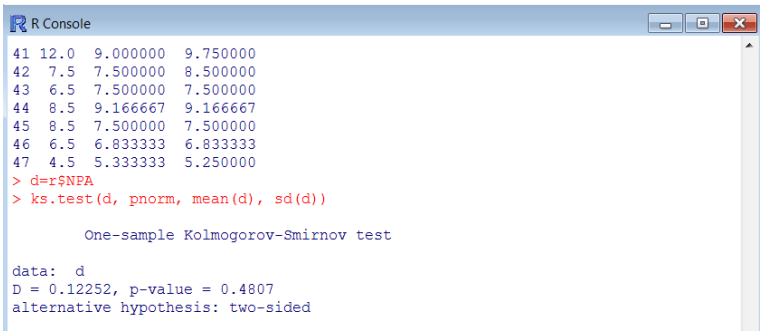

FIG. 3.16 – Test de normalité des notes des pairs avec auto-évaluation

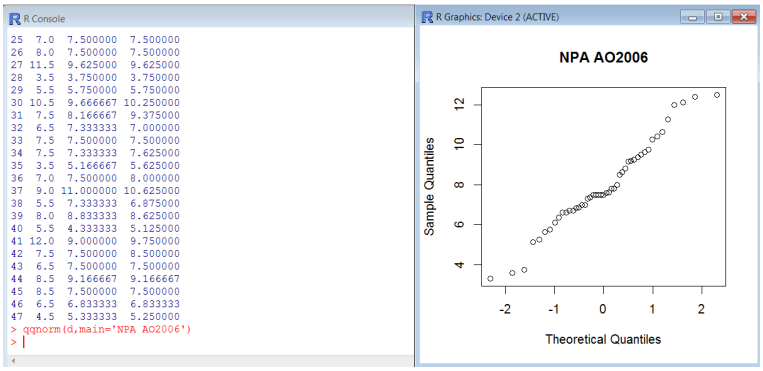

FIG. 3.17 – Droite d'henry des notes des pairs avec auto-évaluation

D'après les résultats obtenus sur R, nous avons :

– $p - value = 0,4807 > \alpha = 0,05 \Rightarrow$ Nous ne rejetons pas l'hypothèse H_0 :"L'échantillon NE a une distribution normale."

– La statistique $D = 0.12252 < k = 0.1983 \Rightarrow$ Nous acceptons l'hypothèse H_0.

Avec k est la valeur lu sur la table de K-S.

On déduit donc que l'échantillon NPA est normalement distribué.

Vu que les trois échantillons (NE, NP, NPA) ont une distribution normale, nous appliquons le test de Student..

2. **Test de Student :**

   (a) **Première comparaison :** (Comparaison entre l'évaluation de l'enseignant et celles des pairs).

   Voici l'application sur R :

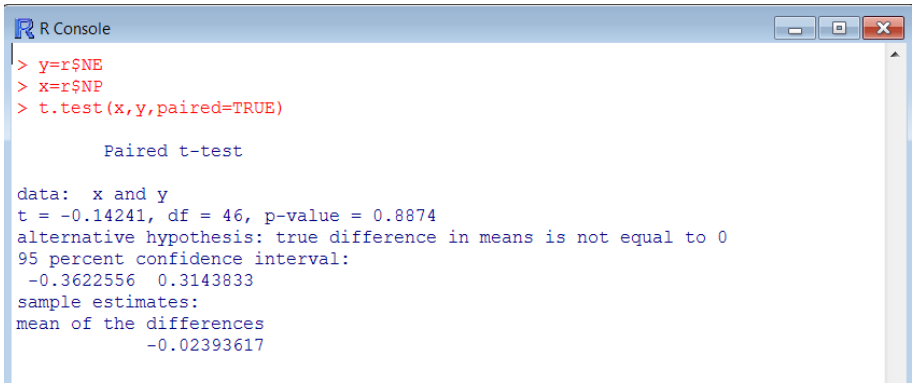

FIG. 3.18 – Test de student de comparaison entre NE et NP

– **Interprétation :**

Pour $\alpha = 0,05$

$p - value = 0,8874 > \alpha = 0,05 \Rightarrow$ Nous ne rejetons pas $H_0 : \mu_{NE} = \mu_{NP}$

$|t| = 0,14241 < t_{46;0,025} = 2,009 \Rightarrow$ Nous acceptons H_0.

Donc les moyennes des notes attribuées par les pairs sont équivalentes aux notes attribuées par l'enseignant.

(b) **Deuxième comparaison :** (Comparaison entre l'évaluation de l'enseignant et celles des pairs avec auto-évaluation).

Voici l'application sur R :

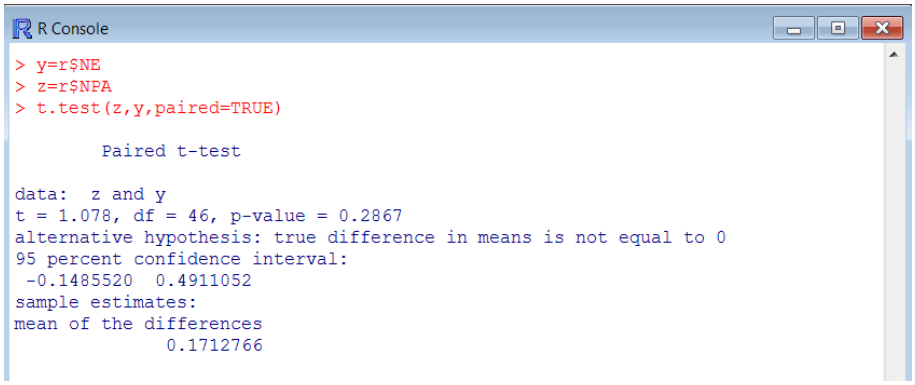

FIG. 3.19 – Test de student de comparaison entre NE et NPA

– **Interprétation :**

Pour $\alpha = 0,05$

$p - value = 0,2867 > \alpha = 0,05 \Rightarrow$ Nous ne rejetons pas $H_0 : \mu_{NE} = \mu_{NPA}$

$|t| = 1,078 < t_{46;0,025} = 2,009 \Rightarrow$ Nous acceptons H_0.

Donc les moyennes des notes attribuées par les pairs avec auto-évaluation sont équivalentes aux notes attribuées par l'enseignant.

3.1.3 Epreuve AO-2007

3.1.3.1 Analyse d'item

Le tableau donné ci-dessous présente à la fois l'indice de difficulté, l'indice de discrimination et le coefficient alpha de cronbach pour chaque item et pour chacune des évaluations suivantes :

– Evaluation par l'enseignant (NE).

– Moyenne des évaluations des pairs sans auto-évaluation (NP).

– Moyenne des évaluations des pairs avec auto-évaluation (NPA).

	Notes enseignant			Notes des pairs			Notes des pairs avec autoévaluation		
Item	Diff	Discrim	Alpha	Diff	Discrim	Alpha	Diff	Discrim	Alpha
Q1	0,24	0,29	0,61	0,26	0,21	0.67	0,25	0,23	0.67
Q2	0,04	0,13	0,64	0,06	0,16	0.67	0,05	0,14	0.68
Q3	0,2	0,3	0.61	0,24	0,38	0.65	0,24	0,38	0.66
Q4	0,39	0,36	0.60	0,38	0,4	0.64	0,38	0,4	0.65
Q5	0,64	0,44	0.59	0,66	0,38	0.64	0,64	0,4	0.65
Q6	0,48	0,16	0.62	0,48	0,28	0.66	0,47	0,28	0.67
Q7	0,59	0,24	0.67	0,74	0,34	0.65	0,72	0,34	0.66
Q8	0,14	0,03	0.67	0,13	0,09	0.69	0,14	0,07	0.70
Q9	0,59	0,34	0.62	0,59	0,32	0.65	0,58	0,31	0.66
Q10	0,4	0,5	0.58	0,39	0,49	0.62	0,35	0,53	0.63
Q11	0,78	0,27	0.63	0,84	0,22	0.67	0,82	0,25	0.68
Q12	0,57	0,23	0.63	0,53	0,35	0.66	0,52	0,37	0.67
Q13	0,53	0,5	0.57	0,53	0,45	0.63	0,5	0,49	0.63

TAB. 3.5 – Les indices de l'expérience AO2007

	Alpha de cronbach total
NE	0.64
NP	0,67
NPA	0.68

TAB. 3.6 – Tableau des alpha de cronbach total pour l'épreuve AO2007

D'après les résultats du tableau (3.5), on remarque que :

1. **Indice de difficulté " p " :** En se référant à l'évaluation de l'enseignant, on constate que cet examen regroupe 3 questions très facile $(p < 0.2)$, 8 questions moyennes $(0.2 < p < 0.6)$ et 2 questions difficiles $(p > 0.6)$, on conclut que cet examen est moyennement difficile.
2. **Indice de discrimination " r " :** Les valeurs de l'indice de discrimination obtenu pour chaque question dans les trois évaluations sont proches. Nous avons 2 questions problématiques : Q2 et Q8 avec $(r < 0.2)$. Les autres questions avec $(r > 0.23)$ s'avèrent discriminantes et bonnes.
3. **Coefficient alpha de Cronbach :** Les coefficients obtenus pour les trois évaluations sont proches, cet indice varie entre 0.64 et 0.68 avec un un alpha de 0.64 pour l'évaluation de l'enseignant. Cette dernière (évaluation de l'enseignant) témoigne d'une bonne fiabilité de l'examen, mais il y'a probablement peu d'items qui nécessitent des améliorations.

4. Pour mesurer la sévérité de chaque correcteur de l'épreuve AO2007, nous avons procédé exactement comme pour l'examen AO2006. Ainsi, nous avons obtenus les résultats du tableau (3.7)

Items	Enseignant/Pairs	Enseignant/Pairs+auto-évaluation
Q1	-0,02	-0,01
Q2	-0,02	-0,01
Q3	-0,04	-0,04
Q4	0,01	0,01
Q5	-0,02	0
Q6	0	0,01
Q7	-0,15	-0,13
Q8	0,01	0
Q9	0	0,01
Q10	0,01	0,05
Q11	-0,06	-0,04
Q12	0,04	0,05
Q13	0	0,03

TAB. 3.7 – Sévérité AO2007

Remarque : Les différences des indices de difficulté obtenus pour chaque question et pour chacune des deux évaluations sont proches dans l'ensemble, mais un écart important dans la question Q_7 entre la sévérité de l'enseignant et celle des pairs (sans et avec auto-évaluation). Ceci veut dire que l'enseignant, les pairs avec et sans auto-évaluation, ont noté cette question d'une manière très différente. Ce qui indique qu'il y a une anomalie (ou bien une ambiguïté) au niveau de cette question (cela peut être au niveau de consignes de notations, barème...).

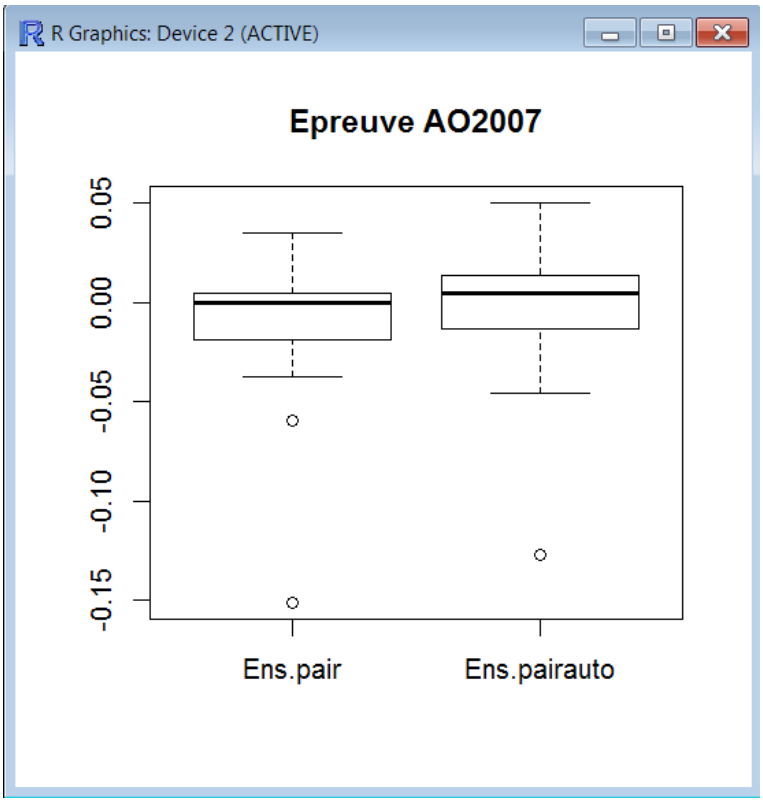

FIG. 3.20 – Boite à moustache AO2007

Les deux boites à moustaches données dans la figure 3.20 construites à base des différences des indices de difficulté (enseignant-pairs, enseignant-pairs avec auto-évaluation) présentées dans le tableau (3.7). Les valeurs extrêmes sont considérées aberrantes.

On remarque qu'il y a 3 valeurs qui sont hors des deux boites à moustache. Deux valeurs correspondent à la question Q_7 et une appartient à la question Q_{11}. Afin de nous assurer que Q_7 et Q_{11} sont réellement aberrantes, une analyse plus profonde a été opérée avec notre promoteur sur ces questions, et effectivement une incohérence dans les consignes de notations a été constatée pour la question Q_7 mais pas pour la question Q_{11}. De ce fait, la question Q_7 doit être retirée de l'étude afin de ne pas biaiser les résultats de notre étude, en revanche la question Q_{11} ne peut être écartée.

3.1.3.2 Détection des notes aberrantes

Comme dans le cas de l'épreuve précédente, on écarte les notes aberrantes, en utilisant le test de Dixon.

Voici le résultat de l'application sur R :

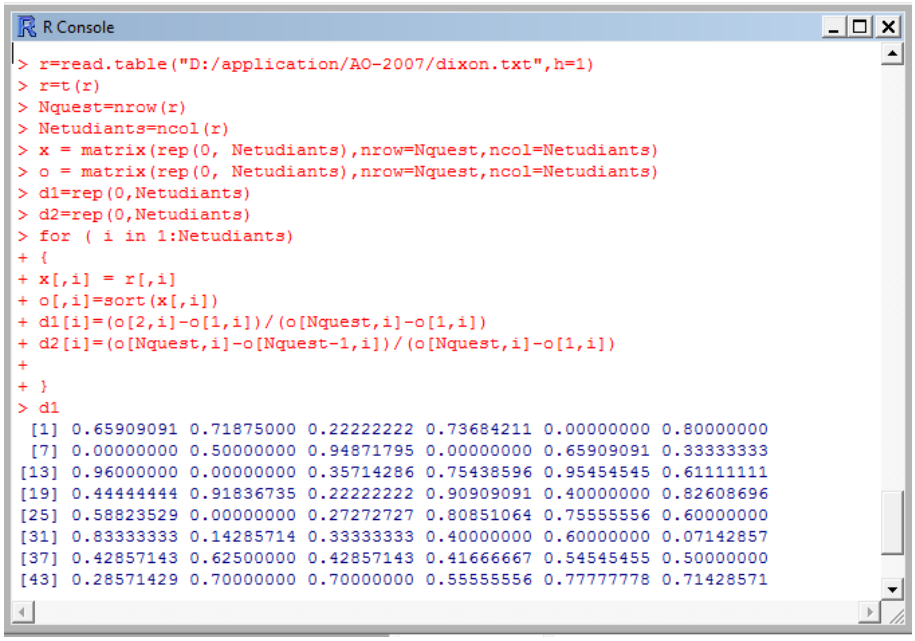

FIG. 3.21 – Test de dixon

	Nombre de copies corrigées	Nombre d'évaluations aberrantes
4 Pairs	60	2
3 Pairs	25	6

TAB. 3.8 – Tableau des résultats obtenus avec le test de dixon pour l'épreuve AO2007

Après l'élimination de toute question et copie pouvant invalider les résultats de notre étude, nous entamons l'analyse de validité de l'EPP.

3.1.3.3 Analyse de validité de l'EPP

Comme dans l'épreuve précédente, on applique le test de student de comparaison des moyennes pour deux échantillons appariés. vérifions d'abord la normalité des trois échantillons suivants : NE, NP et NPA.

1. **Test de normalité (Kolmogorov-Smirnov)**

   Nous appliquons le test de Kolmogorov-Smirnov sur chaque variable (NE,NP et NPA). Là encore, on utilise la droite d'Henry.

   (a) **Variable NE (Notes de l'Enseignant)** :

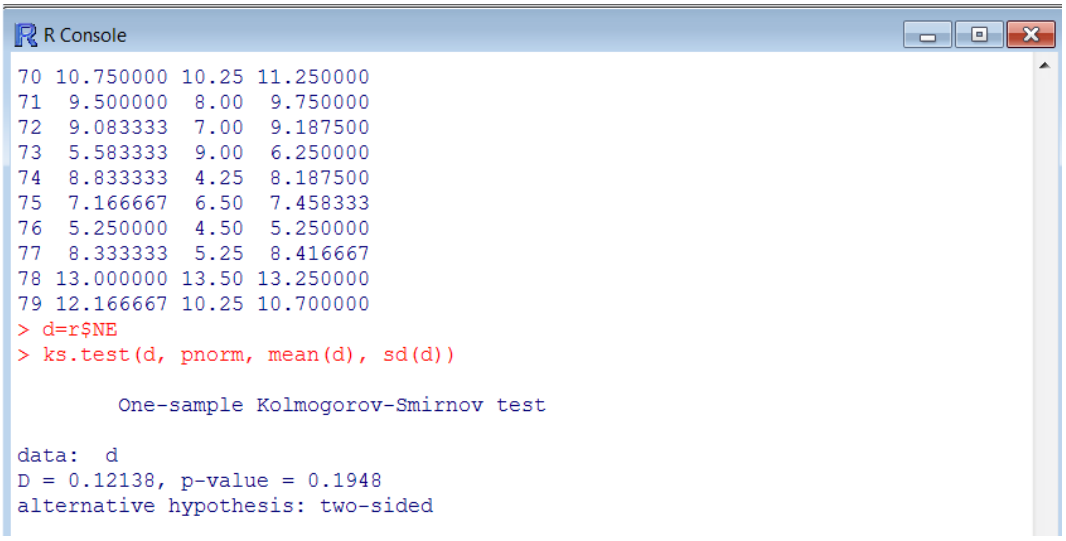

FIG. 3.22 – Test de normalité sur les notes de l'enseignant

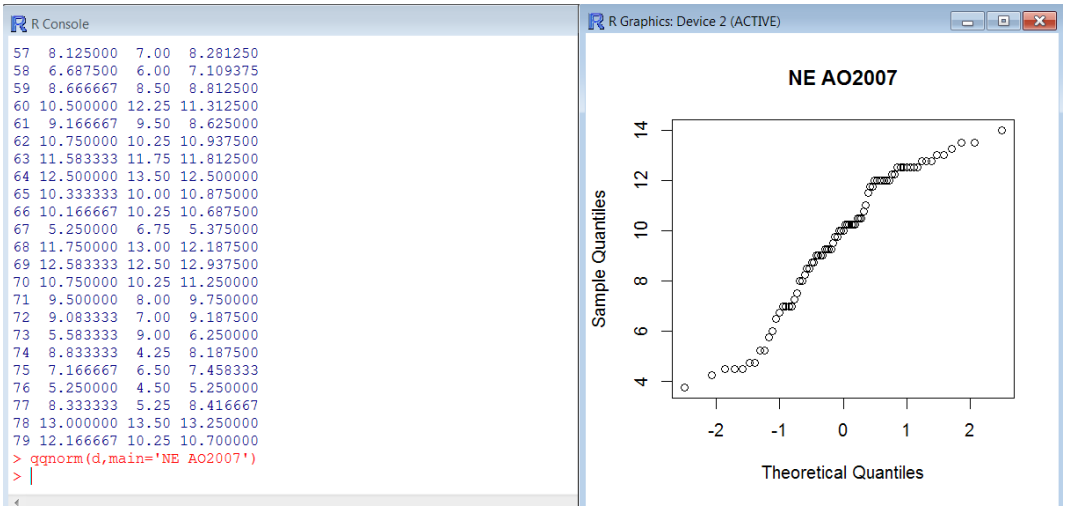

FIG. 3.23 – Droite d'Henry sur les notes de l'enseignant

(b) **Variable NP (Notes des Pairs)** :

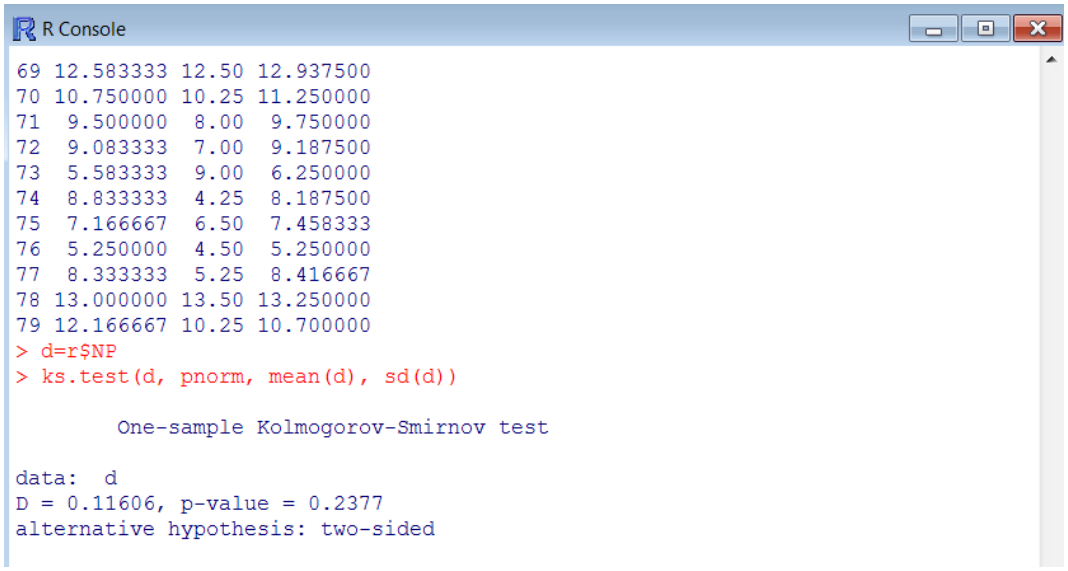

FIG. 3.24 – Test de normalité sur les notes des pairs

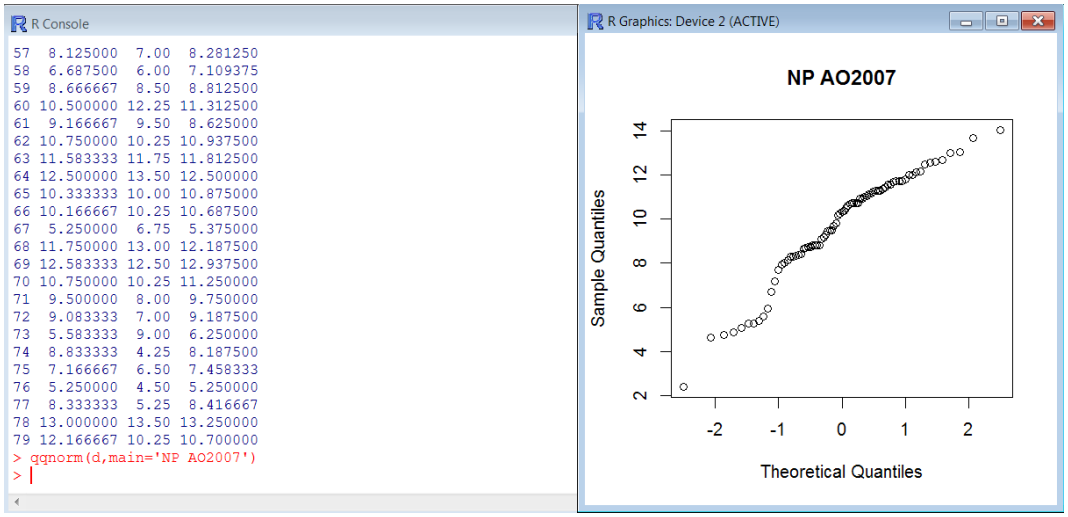

FIG. 3.25 – Droite d'Henry sur les notes des pairs

(c) **Variable NPA (Note des Pairs avec Auto-évaluation)** :

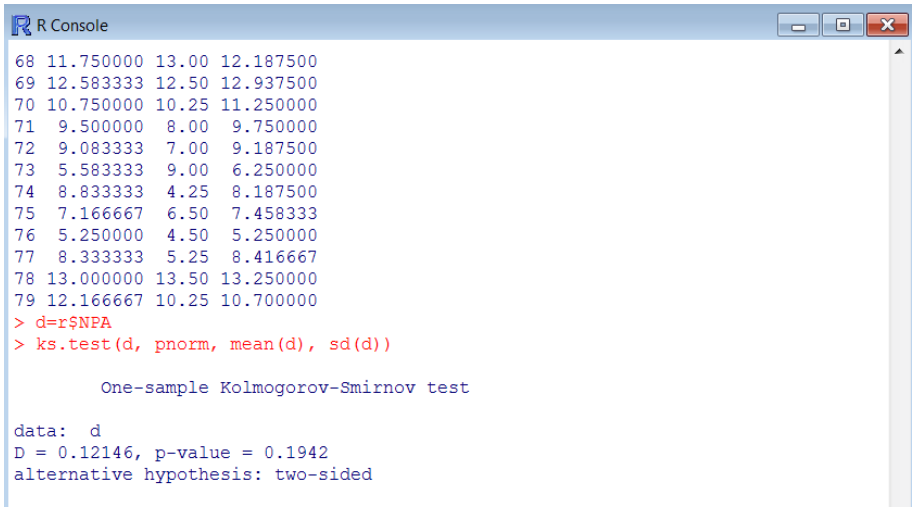

FIG. 3.26 – Test de normalité sur les notes des pairs avec au-évaluation

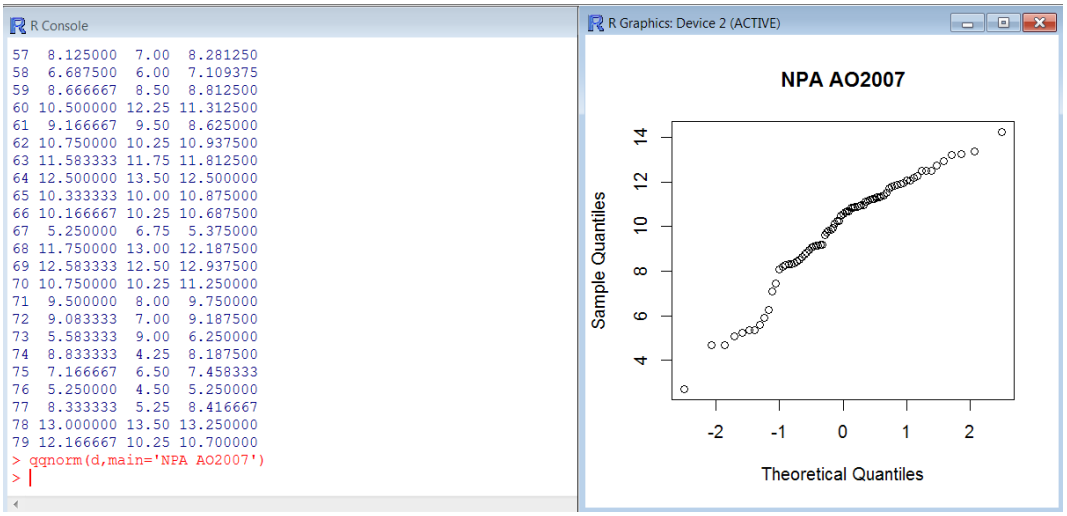

FIG. 3.27 – Droite d'Henry sur les notes des pairs avec au-évaluation

Les résultats de l'application sur R, donnés dans les figures (3.22-3.27) permettent de conclure la normalité des distributions des variables NE, NP et NPA. Nous pouvons donc appliquer le test de student.

2. **Test de Student :**

   (a) **Première comparaison :** (Comparaison entre l'évaluation de l'enseignant et celles des pairs).

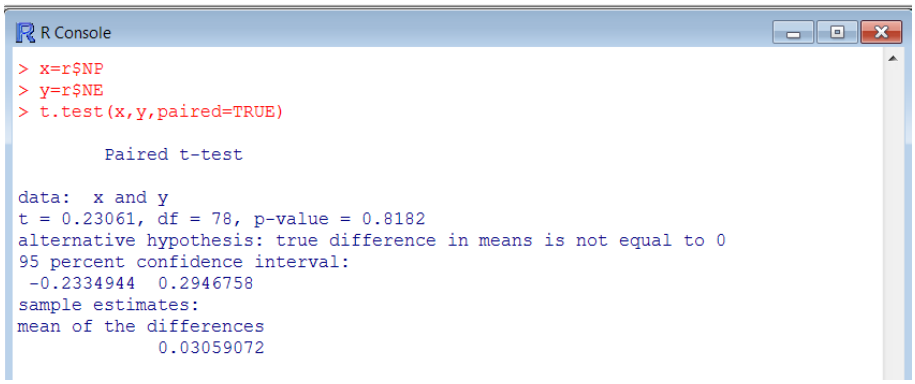

FIG. 3.28 – Test de student de comparaison entre NE et NP

   – **Interprétation :**

   Pour $\alpha = 0,05$

   $p-value = 0,8182 > \alpha = 0,05 \Rightarrow$ Nous ne rejetons pas $H_0 : \mu_{NE} = \mu_{NP}$

   $|t| = 0.23061 < t_{78;0,025} = 1,98 \Rightarrow$ Nous acceptons H_0.

   Donc les moyennes des notes attribuées par les pairs sont équivalentes aux notes attribuées par l'enseignant.

   (b) **Deuxième comparaison :** (Comparaison entre l'évaluation de l'enseignant et celles des pairs avec auto-évaluation)

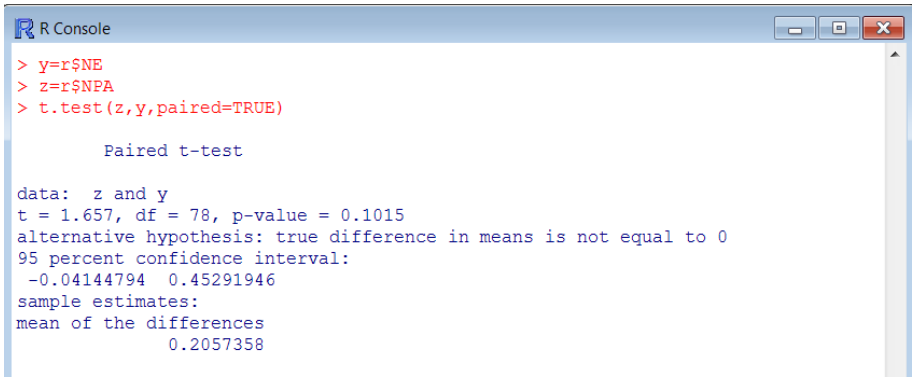

FIG. 3.29 – Test de student de comparaison entre NE et NP

– **Interprétation :**

Pour $\alpha = 0,05$

$p - value = 0,1015 > \alpha = 0,05 \Rightarrow$ Nous ne rejetons pas $H_0 : \mu_{NE} = \mu_{NPA}$

$|t| = 1,657 < t_{78;0,025} = 1,98 \Rightarrow$ Nous acceptons H_0.

Donc les moyennes des notes attribuées par les pairs avec auto-évaluation sont équivalentes aux notes attribuées par l'enseignant.

3.1.4 Epreuve ET2007

3.1.4.1 Analyse d'item

Le tableau donné ci-dessous présente à la fois l'indice de difficulté, l'indice de discrimination et le coefficient alpha de cronbach pour chaque item et pour chacune des évaluations NP, NE et NPA.

	Notes enseignant			Notes des pairs			Notes des pairs avec autoevaluation		
Item	Diff	Discrim	Alpha	Diff	Discrim	Alpha	Diff	Discrim	Alpha
Q1	0,64	0,11	0.50	0,78	0,27	0.58	0,75	0,29	0.58
Q2	0,95	0,16	0.47	0,92	0,34	0.58	0,9	0,22	0.60
Q3	0,91	-0,09	0.53	0,89	0,16	0.61	0,88	0,03	0.63
Q4	0,79	-0,05	0.52	0,78	0,21	0.65	0,77	0,2	0.62
Q5	0,98	0,01	0.50	0,97	0,13	0.61	0,96	0,09	0.61
Q6	0,84	0,37	0.40	0,75	0,45	0.54	0,72	0,5	0.51
Q7	0,58	0,32	0.46	0,67	0,59	0.50	0,63	0,53	0.48
Q8	0,81	0,48	0.33	0,83	0,38	0.55	0,81	0,43	0.50
Q9	0,89	0,2	0.50	0,94	0,32	0.58	0,92	0,33	0.56

TAB. 3.9 – Tableau ET2007

	Alpha de cronbach total
NE	0,50
NP	0,61
NPA	0,60

TAB. 3.10 – Tableau des alpha de cronbach total pour l'épreuve ET2007

D'après les résultats du tableau (3.9) obtenus, on remarque que :

1. **Indice de difficulté " p " :** D'après l'évaluation de l'enseignant, on constate que cet examen comporte deux questions moyennement difficiles ($p = 0.50$ et $p = 0.64$) et tout le

reste des questions sont difficiles $(p > 79)$. On conclut donc que l'examen est globalement très difficile. .

2. **Indice de discrimination " r " :** Les valeurs de l'indice de discrimination obtenu pour chaque question dans les trois évaluations sont proches. Nous avons 6 questions problématiques : Q_1, Q_2, Q_3, Q_4, Q_5 et Q_9 avec $(r \leq 0.2)$ (c'est à dire elles ne sont pas discriminantes) et 3 questions moyennement bien formulées $(0.32 < r < 0.48)$ (ou encore, elles sont moyennement discriminantes).
3. **Coefficient alpha de Cronbach :** Les coefficients obtenus pour les trois évaluations sont proches, cet indice varie entre 0.5 et 0.61 avec un un alpha de 0.5 pour l'évaluation de l'enseignant. Ceci suggère une révision de l'examen.
4. Pour mesurer la sévérité de chaque correcteur de l'épreuve ET2007, nous avons procédé exactement comme pour les examens AO2006 et AO2007. Ainsi, nous avons obtenus les résultats suivants :

Items	Enseignant/Pairs	Enseignant/Pairs+auto-évaluation
Q1	-0,14	-0,11
Q2	0,03	0,05
Q3	0,02	0,03
Q4	0,01	0,02
Q5	0,01	0,02
Q6	0,09	0,12
Q7	-0,09	-0,05
Q8	-0,02	0
Q9	-0,05	-0,03

TAB. 3.11 – Sévérité ET2007

Les différences des indices de difficultés obtenus pour chaque question et pour chaque deux évaluations, sont proches dans l'ensemble. Mais un écart important dans la question Q_1 entre la sévérité de l'enseignant et celle des pairs (sans et avec auto-évaluation). Ceci veut dire que l'enseignant, les pairs avec et sans auto-évaluation, ont noté cette question d'une manière très différente. Ce qui indique qu'il y'a une anomalie (ou bien une ambiguïté) au niveau de cette question (cela peut être au niveau de consignes de notations, barème...). Ceci nécessite une analyse plus profonde sur cette question afin de décider s'il faut ou non la retirer de l'étude.

La boite à moustache présentée ci-dessous nous permet de voir clairement les questions aberrantes :

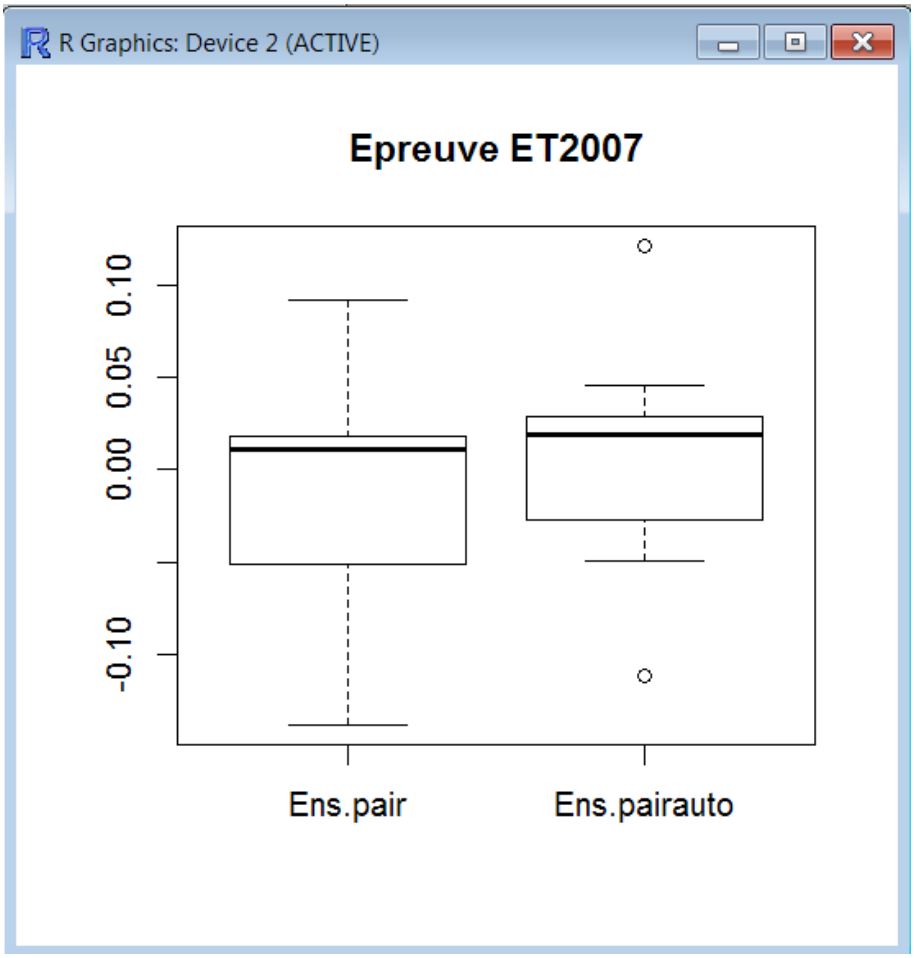

FIG. 3.30 – Boite à moustache ET2007

D'après la figure (3.30), on constate qu'il y a deux valeurs en dehors des deux boites à moustache de (Ens.pairauto). Ces valeur correspondent aux questions Q_1 et Q_6. Afin de décider s'il faut ou non écarter ces questions de notre étude, une analyse plus profonde a été opérée avec notre promoteur et une décision a été prise de les conserver pour la suite de notre étude, vue qu'elles ne signalent pas d'incohérences ou de confusions.

3.1.4.2 Détection des notes aberrantes

Là encore applique le test de Dixon pour éliminer toute note atypique.

Voici un exemple d'application sur R :

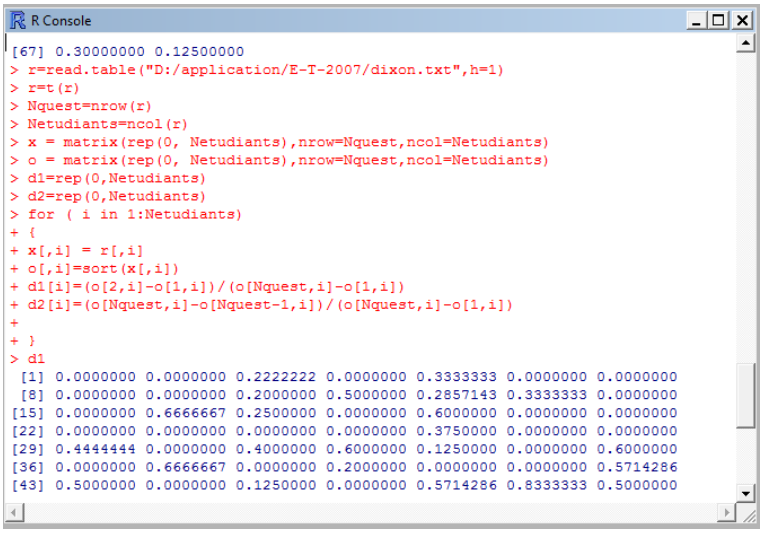

FIG. 3.31 – Test de Dixon

	Nombre de copies corrigées	Nombre d'évaluations aberrantes
4 Pairs	27	1
3 Pairs	25	8

TAB. 3.12 – Tableau des résultats obtenus avec le test de dixon pour l'épreuve ET2007

Après avoir trié les données, en éliminant toute question et copie pouvant invalider les résultats de notre étude, nous entamons l'analyse de validité de l'EPP.

3.1.4.3 Analyse validité de l'EPP

Là encore, on procède comme dans les deux épreuves. On applique le test de student de comparaison des moyennes pour deux échantillons appariées. On vérifie la normalité des trois échantillons suivants : notes attribuées par l'enseignant, notes attribuées par les pairs et les notes attribuées par les pairs avec auto-évaluation.

1. **Test de normalité (Kolmogorov-Smirnov)**

   Nous appliquons le test de Kolmogorov-Smirnov sur chaque variable (évaluation de l'enseignant, évaluation des pairs et évaluation des pairs avec auto-évaluation).

   Voici un exemple d'application sur le logiciel R :

   (a) **Variable NE (Note de Enseignant)** :

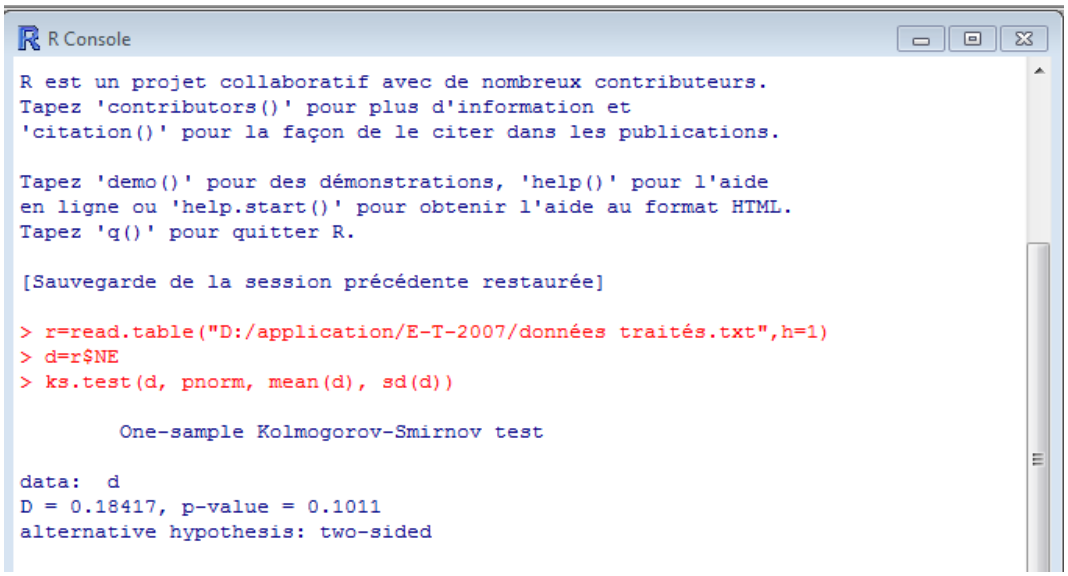

FIG. 3.32 – Test de normalité des notes de l'enseignant

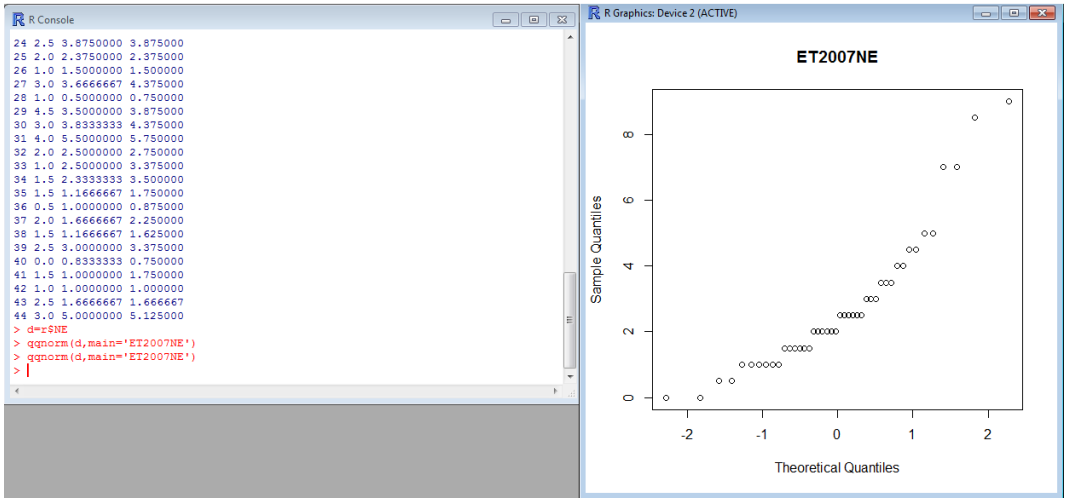

FIG. 3.33 – Droite d'Henry des notes de l'enseignant

(b) **Variable NP (Note des Pairs)** : Test de Kolmogorov avec R :

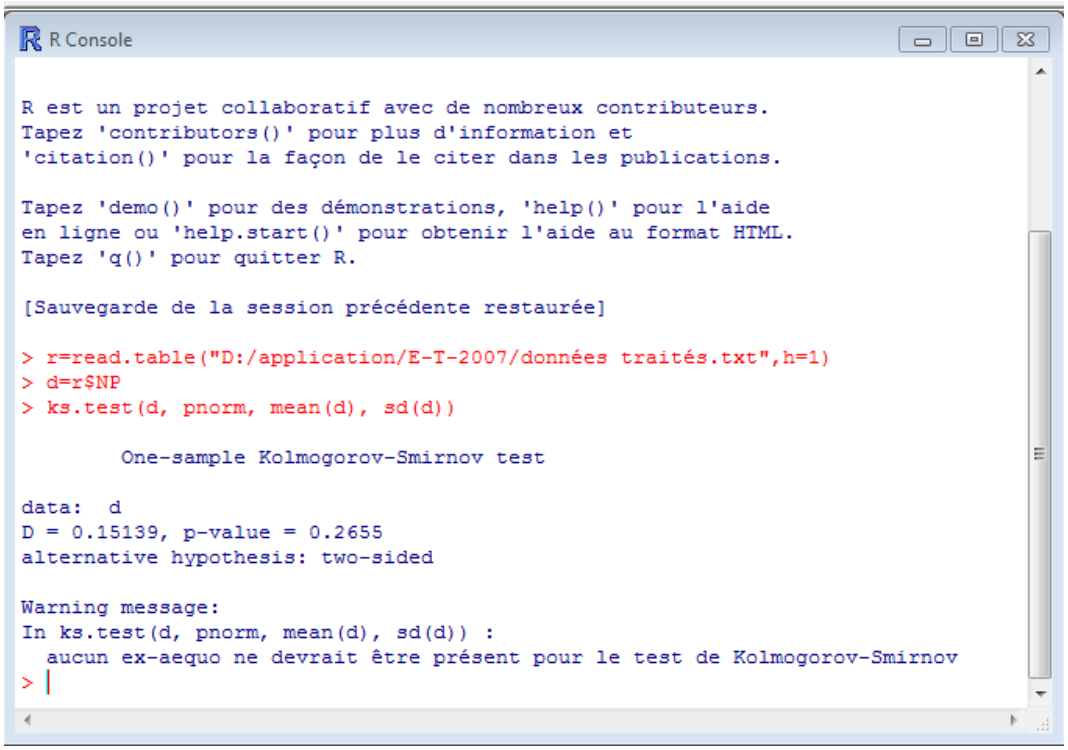

FIG. 3.34 – Test de normalité des notes des pairs

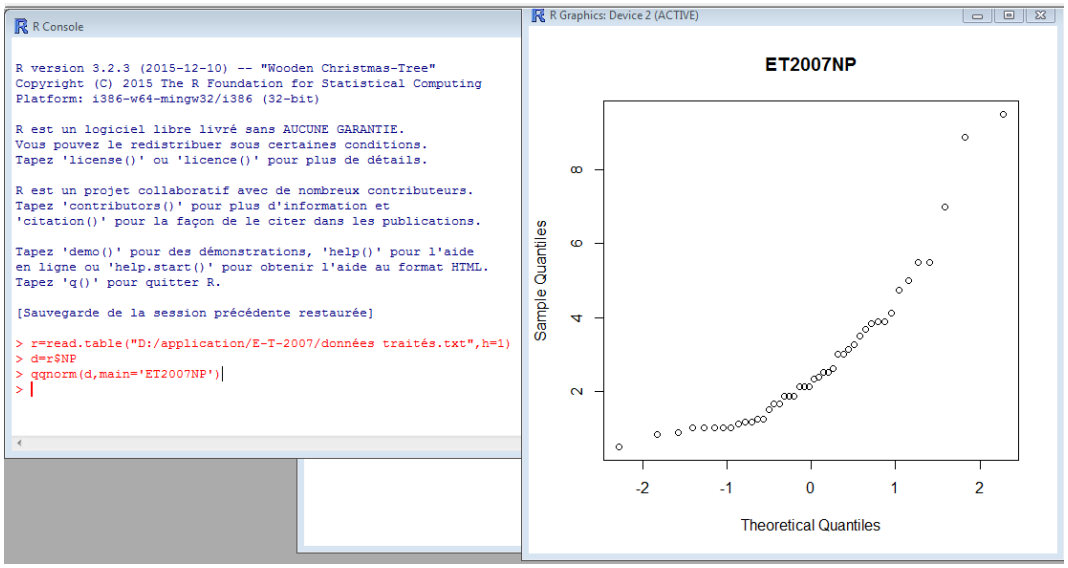

FIG. 3.35 – Droite d'Henry des notes des pairs

(c) **Variable NPA (Note des Pairs avec Auto-évaluation) :**

Test de Kolmogorov avec R :

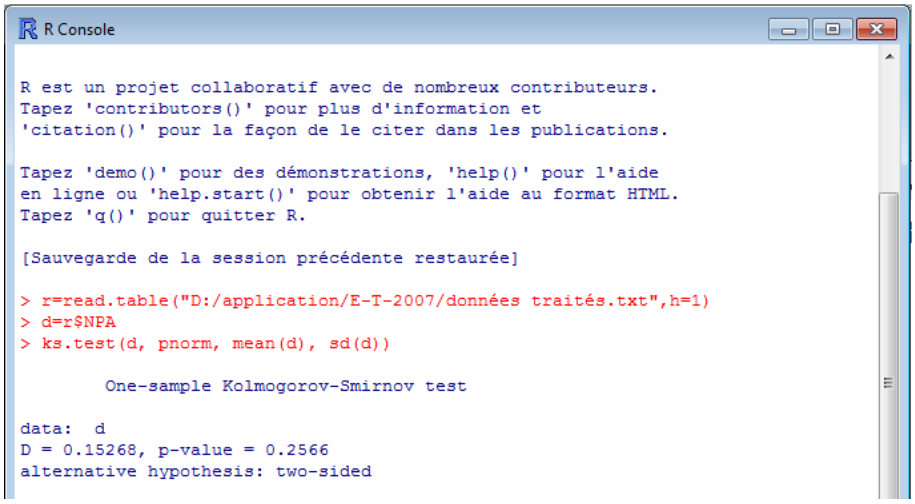

FIG. 3.36 – Test de normalité des notes des pairs avec auto-évaluation

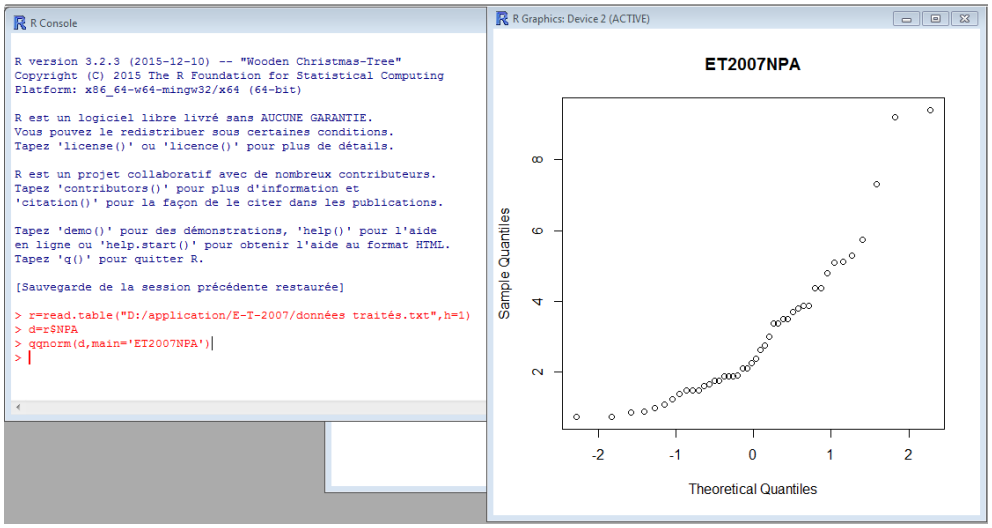

FIG. 3.37 – Droite d'Henry des notes des pairs avec auto-évaluation

Les résultats de l'application sur R, donnés dans les figures (3.32-3.37) permettent de conclure la normalité des distributions des variables NE, NP et NPA. Nous pouvons donc appliquer le test de student.

2. **Test de Student**

(a) **Première comparaison :** (Comparaison entre l'évaluation de l'enseignant et celles des pairs).

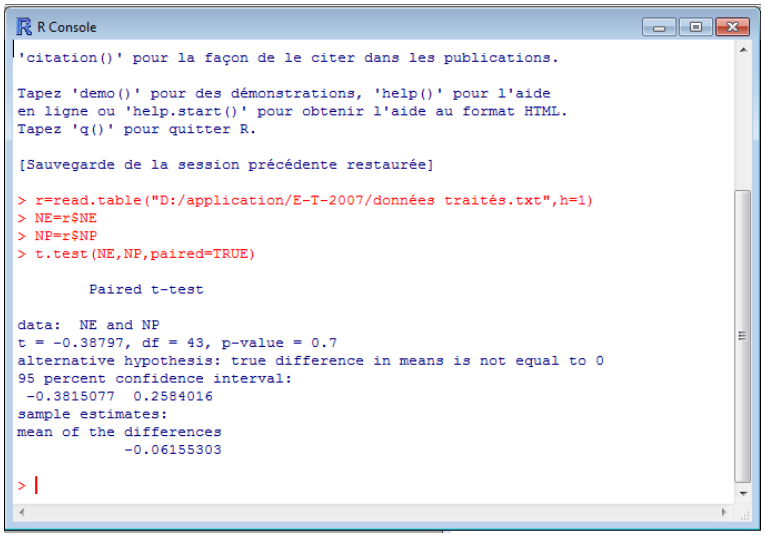

FIG. 3.38 – Test de student de comparaison entre NE et NP

– **Interprétation :**

Pour $\alpha = 0,05$

$p - value = 0,7 > \alpha = 0,05 \Rightarrow$ Nous ne rejetons pas $H_0 : \mu_{NE} = \mu_{NP}$

$|t| = 0.38797 < t_{43;0,025} = 2,009 \Rightarrow$ Nous acceptons H_0.

Donc les moyennes des notes attribuées par les pairs sont équivalentes aux notes attribuées par l'enseignant.

(b) **Deuxième comparaison :** (comparaison entre l'évaluation de l'enseignant et celles des pairs avec auto-évaluation)

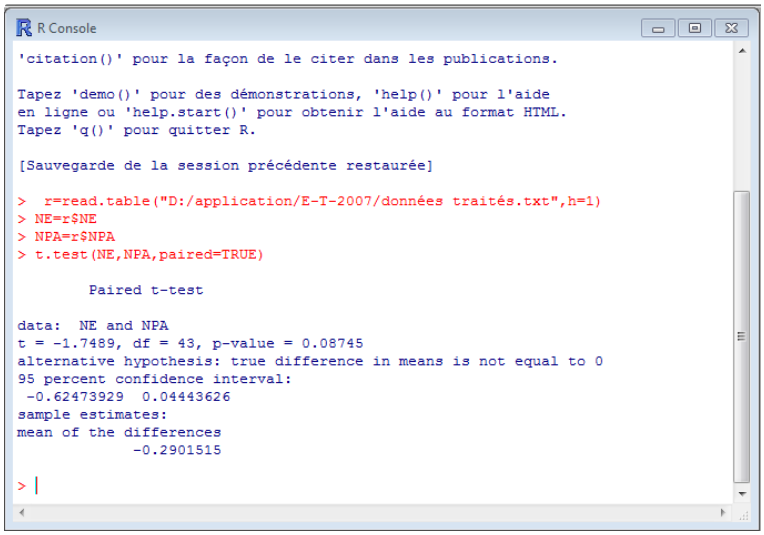

FIG. 3.39 – Test de student de comparaison entre NE et NPA

– **Interprétation :**

Pour $\alpha = 0,05$

$p - value = 0,08745 > \alpha = 0,05 \Rightarrow$ Nous ne rejetons pas $H_0 : \mu_{NE} = \mu_{NPA}$

$|t| = 1,7489 < t_{43;0,025} = 2,009 \Rightarrow$ Nous acceptons H_0.

Donc les moyennes des notes attribuées par les pairs avec auto-évaluation sont équivalentes aux notes attribuées par l'enseignant.

Tableau récapitulatif :

	NE-NP					NE-NPA				
	n	t obs	tc	p-value	moy diff	n	t obs	tc	p-value	moy diff
ET2007	43	-0,388	2,009	0,7	-0,0616	43	-1,7489	2,009	0,0875	-0,2902
AO2006	46	-0,14241	2,009	0,8874	0,0239	46	1,078	2,009	0,2867	0,1713
AO2007	78	0,2306	1,98	0,8182	0,0306	78	1,657	1,98	0,1015	0,2057

FIG. 3.40 – Tableau récapitulatif regroupant les résultats obtenus pour les trois épreuves

Conclusion

Dans les trois épreuves (AO2006, AO2007 et ET2007) nous avons constaté que l'évaluation par les pairs est valide que se soit avec ou sans auto-évaluation, mais il reste à savoir quelle

évaluation est la meilleure et la plus proche de celle de l'enseignant. Pour se faire, on s'appuie sur les moyennes des différence de chaque évaluation. Cependant, nous avons remarqué que la différence des moyennes dans le cas de l'évaluation par les pairs est inférieure à celle de l'évaluation par les pairs avec auto-évaluation pour les trois épreuves (Voir le tableau 3.40). Ce qui veut dire que l'évaluation par les pairs (sans auto-évaluation) est la plus proche à celle de l'enseignant.

3.2 Application aux données d'une émission télévisée

Dans cette deuxième expérience, cent (100) personnes du public et quatres (4) juges ont participé à une émission de divertissement (ONDAR) diffusée sur la chaîne « France2 » (télévision française). Il s'agit de juger (donner une note compris entre 0 et 20) 477 candidats jouant un sketch. Dans cette expérience, on veut vérifier si le jugement d'un nombre important (100) de personnes non expertes est comparable à celui d'experts reconnus. Nous allons suivre les étapes suivantes :

- Identification des notes aberrantes : dans cette étape nous allons appliquer la boite à moustache.
- Analyse de validité de l'EPP : Comme dans la première expérience, nous allons appliquer un test statistique de comparaison de deux échantillon. Pour savoir quel test appliquer (paramétrique ou non paramétrique), nous devons vérifier si les deux variables (jugement du public "JP" et jugement des jury "JJ") suivent une loi normale. De ce fait, nous allons appliquer le test de kolmogorov-Smirnov et appliquer le test adéquat en conséquence.

3.2.1 Identification des valeurs aberrantes

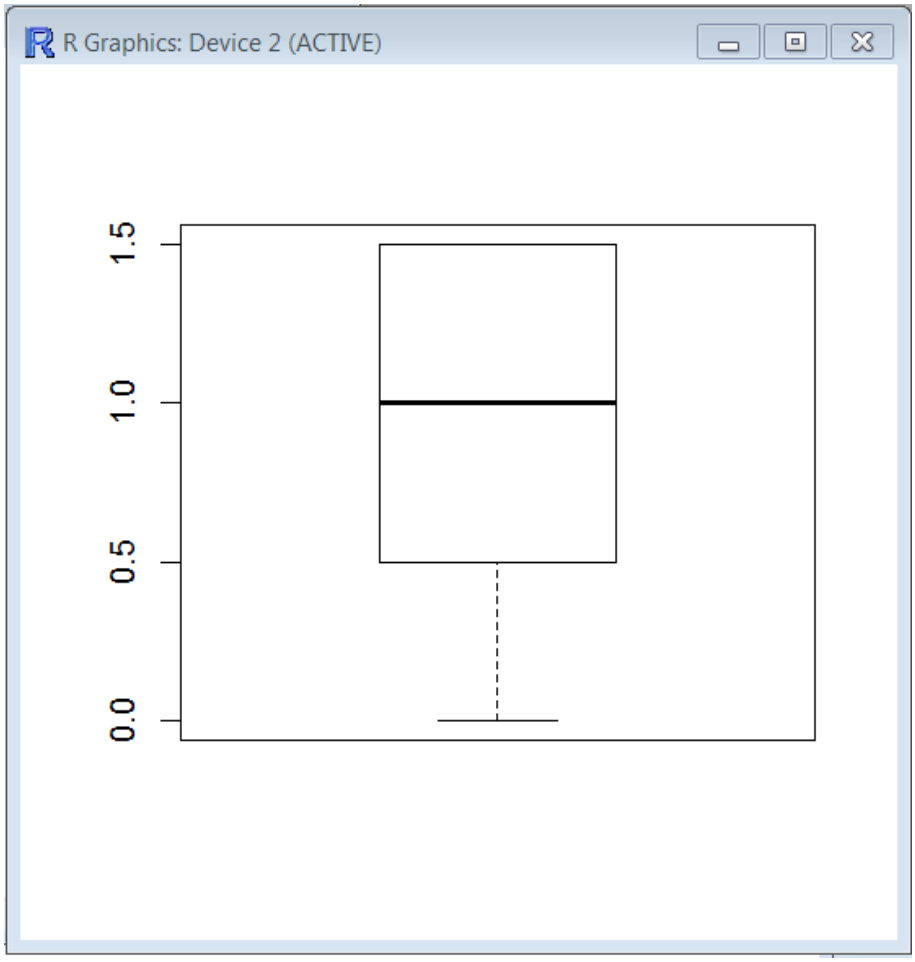

FIG. 3.41 – Boîte à moustaches des moyennes des notes du public

La boite à moustache (la figure 3.41) nous montre clairement qu'il n'y a pas de valeurs aberrantes (hors de la boite à moustache).

3.2.2 Analyse de validité de l'EPP

Nous vérifions si les deux variables (jugement du public et jugement des jury) sont normalement distribuées. Nous appliquons le test de kolmogorov-Smirnov.

1. **Test de Kolmogorov-Smirnov :**

   Pour la variable moyenne des notes attribuées par le public.

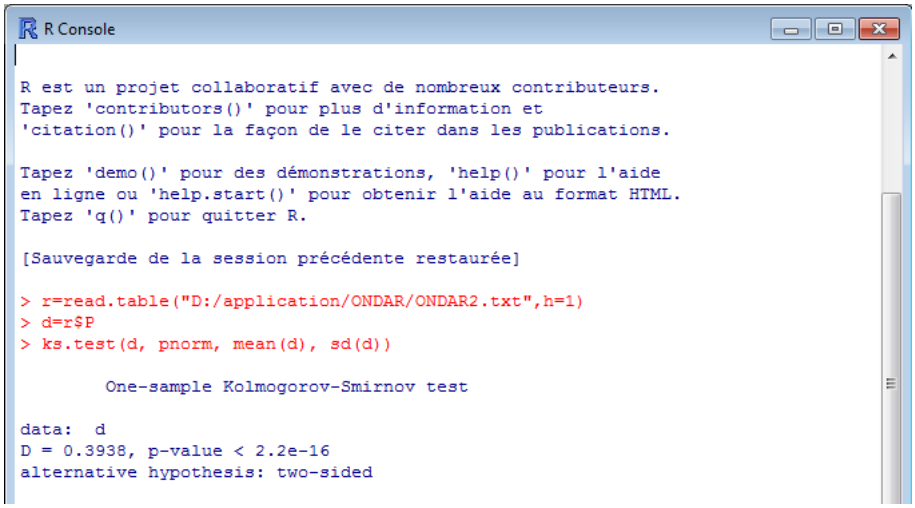

FIG. 3.42 – Test de normalité des notes du public

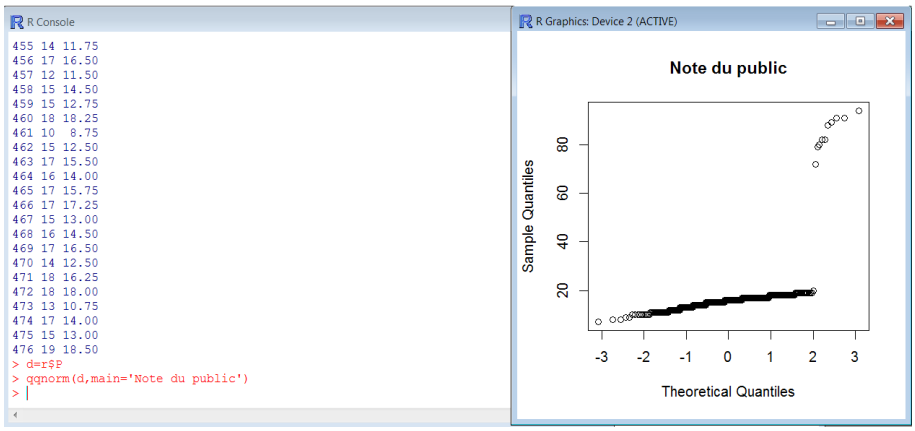

FIG. 3.43 – Droite de Henry des notes du public

D'après les résultats obtenus sur R, nous avons :

– $p-value = 2.2e^{-16} < \alpha = 0,05 \Rightarrow$ Nous rejetons l'hypothèse H_0 :"L'échantillon JP a une distribution normale."

– La statistique $D = 0.3938 < k = 0.0623 \Rightarrow$ Nous rejetons l'hypothèse H_0.

Avec k est la valeur lu sur la table de K-S.

On déduit donc que les notes du public ne sont pas normalement distribuées. Pour vérifier les résultats obtenus avec le test de K-S, nous appliquons le test de khi-deux.

Les résultat obtenus avec R sont dans la figure (3.44).

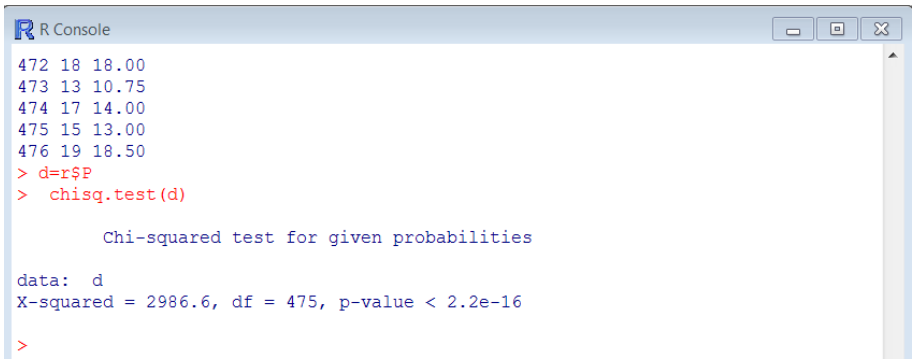

FIG. 3.44 – Test de normalité de khi-deux

– $p-value = 2.2e^{-16} < \alpha = 0,05 \Rightarrow$ Nous rejetons l'hypothèse H_0 :"L'échantillon JP a une distribution normale."

– La statistique $X-squared = 2986,6 > d = 273.88 \Rightarrow$ Nous rejetons l'hypothèse H_0.

On conclut donc que l'échantillon JP n'est pas normalement distribué.

Nous appliquons donc le test de wilcoxon sur les deux échantillons.

2. Test de wilcoxon :

   Voici l'application dans R :

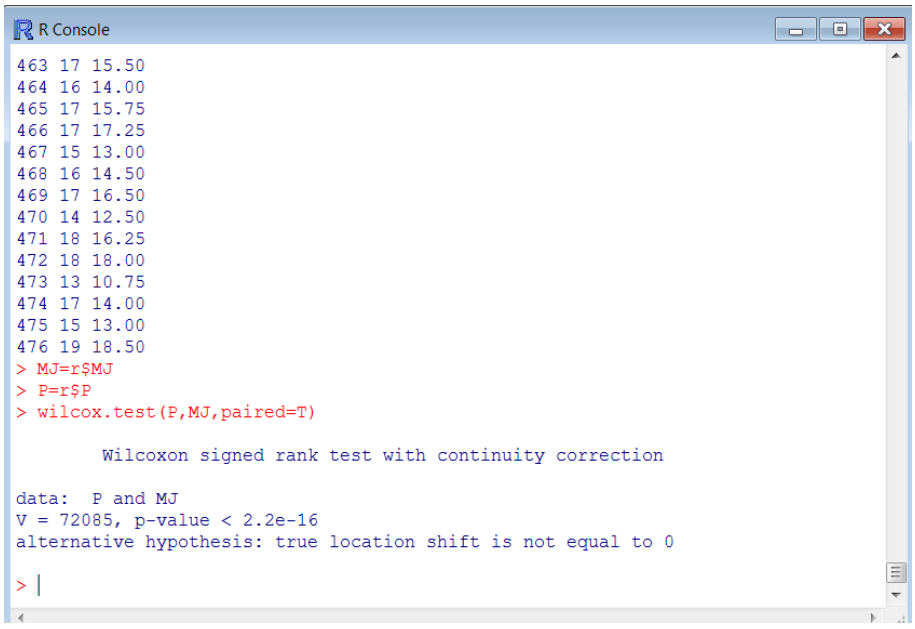

FIG. 3.45 – Test de wilcoxon de comparaison entre les moyennes des notes du public et celles des experts

D'après les résultats obtenus sur R, nous avons :

- $p - value = 2.2e^{-16} < \alpha = 0,05 \Rightarrow$ Nous ne rejetons pas l'hypothèse H_0 :"L'échantillon NE a une distribution normale."
- La statistique $V = 72085 > l = 0.480 \Rightarrow$ Nous acceptons l'hypothèse H_0.

On déduit donc que les notes du public ne sont pas équivalentes à celles des experts.

Conclusion

Le but de cette deuxième application était de vérifier si l'augmentation du nombre d'évaluateurs (4 experts et 100 non experts) peut compenser les écarts pouvant être occasionnel par la subjectivité sachant que l'expérience que nous avons choisis impliquait une évaluation complètement subjective.

A notre étonnement, les résultats ont montré clairement que l'impact la subjectivité est beaucoup plus fort que celui de l'augmentation du nombre d'évaluateurs.

Conclusion

Dans ce travail, nous avons traité la problématique de la validité de l'évaluation par les pairs. A cet effet, nous avons entrepris deux applications d'analyse statistiques correspondant à deux expériences d'évaluation par les pairs très différentes.

Dans la première application, nous avons exploité les données issues d'une expérience d'évaluation par les pairs ayant eu lieu à l'université de Abederrahmane Mira en 2006 et 2007 sur trois les épreuves. Pour chaque épreuve, nous avons opéré trois étapes : Analyse d'items, identification des évaluations atypiques et enfin l'analyse de validité de l'évaluation par les pairs proprement dite.

La première étape nous a permis une analyse profonde de la qualité des épreuves et d'identifier les items atypiques pour chaque épreuve. Dans la deuxième étape, nous avons considéré les notes globales (sur 20) pour chaque copie et nous avons procédé à la détection des évaluations aberrantes effectuées par les pairs. Cette étape, a permet d'exclure un ensemble de copies de l'étude. Dans la dernière étape, nous avons utilisé des tests statistiques (kolmogorov, student), pour effectuer une comparaison de l'évaluation de l'enseignant et celle des pairs (avec et sans auto-évaluation), dans le but de savoir si l'évaluation par l'enseignant peut être remplacée par celle des pairs. Les résultats obtenus montrent que l'évaluation par les pairs est valide, pour les trois épreuves. Ceci nous amène à conclure que l'on pourrait faire confiance aux notes des pairs à condition de respecter les recommandations fixées lors de l'expérience : Validité des épreuves, validité des barèmes, validité et clarté des consignes de notation, questions fortement objectives, 4 à 5 évaluations des pairs pour chaque copie.

Un paramètre clés de cette première expérience est le nombre limité d'évaluateurs étudiant (les pairs) impliqués pour corriger une copie. Un second paramètre est l'objectivité des

épreuves considérées. Nous avons voulu analyser une situation, où ces deux paramètres sont complètement différents de cette première expérience : Un nombre important d'évaluateurs et une très grande subjectivité. C'est ce qui a donné lieu à notre seconde application qui concerne une émission de divertissement (ONDAR) qui a été diffusée à la télévision (France 2). A notre surprise, les résultats obtenus, contredisent l'hypothèse de la validité de l'évaluation par les pairs. Malgré que nous avons considéré non pas un seul expert, mais la moyenne de 4 experts et 100 évaluateurs non expert (public) pour chaque copie (sketch), l'évaluation du public n'est pas digne de confiance. Ceci témoigne du poids de la subjectivité dans l'évaluation !

Pour répondre à notre problématique, l'évaluation par les pairs n'est valide qu'avec certaines conditions : Bonne qualité de l'épreuve, des barèmes précis, et des consignes de notations claires et une faible subjectivité.

Annexes

Annexe A

x / N	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5	031	188	500	812	969	*										
6	016	109	344	656	891	984	*									
7	008	062	227	500	773	938	992	*								
8	004	035	145	363	637	855	965	996	*							
9	002	020	090	254	500	746	910	980	998	*						
10	001	011	055	172	377	623	828	945	989	999	*					
11		006	033	113	274	500	726	887	967	994	*	*				
12		003	019	073	194	387	613	806	927	981	997	*	*			
13		002	011	046	133	291	500	709	867	954	989	998	*	*		
14		001	006	029	090	212	395	605	788	910	971	994	999	*	*	
15			004	018	059	151	304	500	696	849	941	982	996	*	*	*
16			002	011	038	105	227	402	598	773	895	962	989	998	*	*
17			001	006	025	072	166	315	500	685	834	928	975	994	999	*
18			001	004	015	048	119	240	407	593	760	881	952	985	996	999
19				002	010	032	084	180	324	500	676	820	916	968	990	998
20				001	006	021	058	132	252	412	588	748	868	942	979	994
21				001	004	013	039	095	192	332	500	668	808	905	961	987
22					002	008	026	067	143	262	416	584	738	857	933	974
23					001	005	017	047	105	202	339	500	661	798	895	953
24					001	003	011	032	076	154	271	419	581	729	846	924
25						002	007	022	054	115	212	345	500	655	788	885

FIG. 3.46 – Table des valeurs critiques du test binomial

Annexe B

Valeur critique de z	Probabilité sous H0 de z	
	unilatéral	bilatéral
1,65	0,0495	0,099
1,96	0,0250	0,05
2,33	0,0099	0,0198
2,58	0,0049	0,0098
3,08	0,0010	0,0020
3,30	0,0005	0,001

FIG. 3.47 – Table des valeurs critiques de Z

Bibliographie

[1] Ferguson, G., Sheader, E., & Grady, R. (2008). Computer-assisted and peer assessment : A combined approach to assessing first year laboratory practical classes for large numbers of students. Bioscience Education, 11(1), 1-16.

[2] Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. Journal of Educational Psychology, 98(4), 891.

[3] Bostock, S. (2000). Student peer assessment. Learning Technology.

[4] Topping, K. (1998). Peer assessment between students in colleges and universities. Review of educational Research, 68(3), 249-276.

[5] Zenaidi, G. L'évaluation par les pairs dans les universités : utile aux chercheurs comme aux étudiants [en ligne]. (Créé le mercredi 13 janvier 2010, Mise à jour le jeudi 15 mars 2012)Disponible sur : "http ://goo.gl/oDXh1b" (consulté le 03 février 2016).

[6] Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education : A meta-analysis comparing peer and teacher marks. Review of educational research, 70(3), 287-322.

[7] De Grez, L., Valcke, M., & Roozen, I. (2012). How effective are self-and peer assessment of oral presentation skills compared with teachers' assessments ?. Active Learning in Higher Education, 13(2), 129-142.

[8] Bouzidi, L., Jaillet, A. (2006). Prolongement virtuel de l'examen sur table par la mise en place d'une évaluation par les pairs dans un but formatif. 1-11.

[9] Bachelet, R., Zongo, D., & Bourelle, A. (2015). Does peer grading work ? How to implement and improve it ? Comparing instructor and peer assessment in MOOC GdP. In European MOOCs Stakeholders Summit 2015.

[10] Vozniuk, A., Holzer, A., & Gillet, D. (2014). Peer assessment based on ratings in a social media course. In Proceedings of the Fourth International Conference on Learning Analytics And Knowledge (pp. 133-137). ACM.

[11] Donnon, T., McIlwrick, J., & Woloschuk, W. (2013). Investigating the reliability and validity of self and peer assessment to measure medical students' professional competencies. Creative Education, 4(06), 23.

[12] Strang, K. D. (2013). Does cooperative e-learning improve graduate student project outcomes ?. International Journal of Technology Enhanced Learning, 5(1), 42-55.

[13] Jones, I., & Alcock, L. (2012). Summative peer assessment of undergraduate calculus using Adaptive Comparative Judgement. Mapping University Mathematics Assessment Practices. Norwich : University of East Anglia.

[14] Khonbi, Z. A., & Sadeghi, K. (2012). The Effect of Assessment Type (self vs. peer vs. teacher) on Iranian University EFL Students' Course Achievement. Language Testing in Asia, 2(4), 1-28.

[15] Laveault, D., & Grégoire, J. (2002). Introduction aux théories des tests : en psychologie et en sciences de l'éducation. De Boeck Supérieur.

[16] De Ketele, J. M., & Gerard, F. M. (2005). La validation des épreuves d'évaluation selon l'approche par les compétences. Mesure et évaluation en éducation, 28(3), 1-26.

[17] Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self assessments. Assessment and Evaluation in Higher Education, 11(2), 146-166.

[18] Mowl, G., & Pain, R. (1995). Using self and peer assessment to improve students' essay writing : A case study from geography. Programmed Learning, 32(4), 324-335.

[19] Stefani, L. A. (1994). Peer, self and tutor assessment : relative reliabilities. Studies in Higher Education, 19(1), 69-75.

[20] Cheng, W., & Warren, M. (1999). Peer and teacher assessment of the oral and written tasks of a group project. Assessment & Evaluation in Higher Education, 24(3), 301-314.

[21] Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. Educational assessment, 11(1), 1-31.

[22] Chinn, D. (2005, June). Peer assessment in the algorithms course. In ACM SIGCSE Bulletin (Vol. 37, No. 3, pp. 69-73). ACM.

[23] Saito, H., & Fujita, T. (2009). Peer-assessing peers' contribution to EFL group presentations. RELC Journal, 40(2), 149-171.

[24] Kulkarni, C., Pang-Wei, K., Le, H., Chia, D., Papadopoulos, K., Koller, D., & Klemmer, S. R. (2013). Scaling self and peer assessment to the global design classroom. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

[25] Ramousse, R., Berre, M. L., & Guelte, L. L. (1996). Introduction aux statistiques. disponible sur le lien : http ://www. consdev. org/elearning/stat/index. html.

[26] Morissette, D. (1997). Guide pratique de l'évaluation sommative : gestion des épreuves et des examens. De Boeck Supérieur.

[27] Lavire, C., Louis, D., Perrière, G., Briolay, J., Normand, P., & Cournoyer, B. (2001). Analysis of pFQ31, a 8551-bp cryptic plasmid from the symbiotic nitrogen-fixing actinomycete Frankia. FEMS microbiology letters, 197(1), 111-116.

[28] Kim, J., Doop, M. L., Blake, R., & Park, S. (2005). Impaired visual recognition of biological motion in schizophrenia. Schizophrenia research, 77(2), 299-307.

[29] Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. psychometrika, 16(3), 297-334.

[30] Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. Educational and psychological measurement, 53(1), 33-49.

[31] Nunnally, J. C., Bernstein, I. H., & Berge, J. M. T. (1967). Psychometric theory (Vol. 226). New York : McGraw-Hill.

[32] Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. Technometrics, 11(1), 1-21.

[33] Carletti, G. (1988). Comparaison empirique de methodes statistiques de detection de valeurs anormales a une et a plusieurs dimensions.

[34] Munoz-Garcia, J., Moreno-Rebollo, J. L., & Pascual-Acosta, A. (1990). Outliers : A formal approach. International Statistical Review/Revue Internationale de Statistique, 215-226.

[35] Barnett, V. (1994). T. Lewis Outliers in statistical data. John Wiley & Sons, 920, l.

[36] Konieczka, P., & Namiesnik, J. (2009). Quality assurance and quality control in the analytical chemical laboratory : a practical approach. CRC Press.

[37] Bressoud, É., & Kahané, J. C. (2008). Statistique descriptive : applications avec Excel et la calculatrice. Pearson education.

[38] Preux, P. M., Druet-Cabanac, M., Dalmay, F., & Vergnenègre, A. (2003). Qu'est-ce qu'un test paramétrique ?. Revue des maladies respiratoires, 20(6-C1), 952-954.

[39] Bennani-Dosse, M. (2011). Statistique bivariée avec R (pp. 296-p). Presses universitaires de Rennes.

[40] Legras, B., Kohler, F., & Kohler, F. (1998). Eléments de statistique : à l'usage des étudiants en médecine et en biologie : cours et exercices corrigés. Ellipses.

[41] Morgenthaler, S. (2007). Introduction à la statistique. PPUR presses polytechniques.