

Projet STT3200

Étude de la relation entre la note finale d'un étudiant et des facteurs environnementaux et sociaux

Auteur : René Picard

Date : 4 mars 2024

Dans cette étude observationnelle on tente de trouver quel facteurs pourrait permettre de créer un modèle futur dans le but de prédire dans quels éléments le ministère de l'éducation du Québec pourrait invertir pour améliorer significativement les notes des étudiants québécois.

Table des matières

Introduction.....	2
Méthodologie.....	2
Résultats.....	3
La variable « Temps de voyage ».....	4
La variable « Temps d'étude ».....	6
La variable « Aide à l'école ».....	9
La variable « Aide familiale ».....	11
La variable « Cours privés ».....	13
La variable « Internet ».....	15
La variable « Consommation d'alcool les jours d'école».....	17
La variable « Consommation d'alcool les fin de semaine».....	20
Conclusion.....	22
Annexe.....	23
Annexe A.....	23
Annexe B.....	34

Introduction

Suite à la demande du ministère de l'éducation du Québec d'évaluer la meilleure méthode pour améliorer de manière significative les notes des étudiants québécois niveau secondaire, nous avons décidé de faire une analyse statistique sur une base de donnée publique qui a été récolté par l'université « University de Camerino ». Cette base de donnée possède une trentaine de prédicteur et trois variables à déduire.

Nous avons tout d'abord décidé de ne prendre que les prédicteurs que l'on peut modifier pour améliorer les notes futures des étudiants. Ensuite, nous avons pris qu'une des trois variables à déduire et fait divers analyses statistiques pour déterminer, quelles variables ont un impact significatif sur la note d'un étudiant. Suite aux conclusions de ce rapport, nous pourrons apporter une analyse plus approfondie, dans lequel nous aurons un modèle permettant de prédire à quel point la modification d'une ou de plusieurs variables aura un impact sur la note futur d'un étudiant.

Méthodologie

Cette étude a employé une approche quantitative pour analyser l'impact de divers facteurs sur les performances académiques des étudiants de niveau secondaire. Les données ont été extraites d'une base

de données publique, recueillie par une institution académique qui comprend des informations sur les élèves, leurs habitudes d'étude, leur environnement familial et éducatif, ainsi que leurs notes finales dans les matières de mathématiques et de portugais.

Parmi la trentaine de prédicteurs disponibles, seul celles qui sont susceptible d'être influencés par des interventions pédagogiques ont été sélectionnée. Ces variables comprennent, entre autres, le temps de voyages entre l'école et la maison, le temps d'étude consacré par l'étudiant pour un cours donné, l'aide à l'école en dehors des cours, l'aide fournit par la famille, les cours privés supplémentaires, l'accès à l'internet à la maison, la consommation d'alcool les jours d'école et la consommation d'alcool les fin de semaine. Pour la variable prédictive, deux ont été choisis, la note finale pour le cours de mathématique et la note de finale pour le cours de portugais. La raison de ce choix est la nature d'un cours de portugais (anologue au cours de français au Québec) est très différente de la nature d'un cours de mathématique.

Ensuite, les données ont été nettoyer. Toutes les données ayant une note finale strictement inférieur à 5 sur 20 a été éliminer de la base de donnée. La raison est qu'il y avait un nombre anormalement élevé de 0/20¹, qui représente très probablement des étudiants qui ont abandonné le cours ou ne ce sont pas présenté, et que la base de donnée ne nous donne pas assez d'information sur le pourquoi ils ont eu 0. Ensuite, pour les notes de 1 à 4, la raison est que sur les presque 1000 étudiants de la base de donné, qu'un seul à une note de 1/20 et personne n'a 2/20, 3/20, et 4/20. Garder ces données causent des problèmes dans les analyses statistiques de ce rapport.

Pour les analyse, on commence par faire des statistiques descriptives pour chaque prédicteur en fonction de la note finale du cours de mathématique et ensuite du cours de portugais. Ensuite créons des graphique permettant de visualiser la relation entre ces 2 variables et comprendre leur distribution. Et ensuite, suite à ces premières analyses, nous utilisons le test adéquat pour déterminer si oui ou non un prédicteur a un impact significatif ou non sur la note finale d'un étudiant d'un cours de mathématique et/ou d'un cours de portugais donné.

Résultats

¹ Voir annexe A

La variable « Temps de voyage »

Tout d'abord, nous analysons le temps de voyage pour le cours de mathématique et ensuite pour le cours de portugais. Le temps de voyage est une variable indépendante à 4 niveaux qui se décompose comme suit :

- Niveau 1 : Cela prend moins de 15 minutes pour se rendre à l'école.
- Niveau 2 : Cela prend entre 15 et 30 minutes pour se rendre à l'école.
- Niveau 3 : Cela prend entre 30 minutes et 1 heure pour se rendre à l'école.
- Niveau 4 : Cela prend plus d'une heure pour se rendre à l'école.

Niveau	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
1	235	11,77	3,28	11	5	20	15
2	95	11,16	3,12	11	5	19	14
3	19	11,21	2,9	11	6	18	12
4	7	10	1,83	10	7	13	6

On peut observer que les moyennes de chaque niveau du temps de voyages sont assez proche, leurs écart-type également, excepté pour le 4^e groupe, qui possède également le moins d'observations. Le 4^e groupe à aussi la moins grande étendu, les notes semblent être moins haute, mais aussi moins basse.

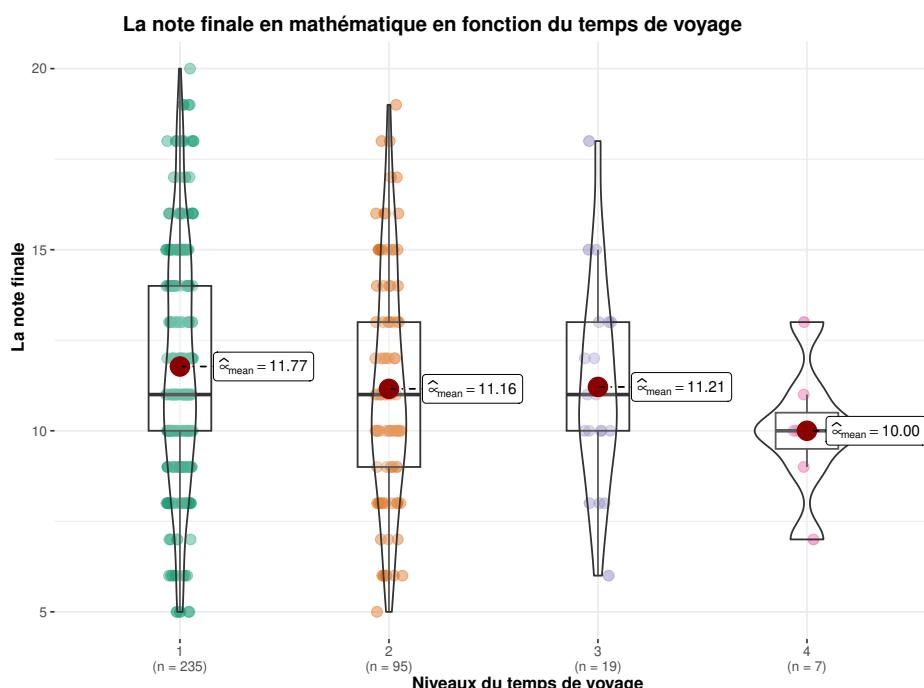


Figure 1: Relation entre la note finale du cours de mathématique et les niveaux du temps de voyages. Voir annexe B pour le code R associé.

La dispersion des données de chaque niveaux sont semblable pour les 3 premiers niveaux, mais pas pour le 4^e. Suite au test de normalité² qui est respecté, on a décidé d'utiliser la méthode de Bonferroni avec un test de Welch. Grâce à ce test, nous pouvons confirmer qu'avec un risque de première espèce α à 5% que le temps de voyage n'affecte pas la moyenne de la note finale pour le cours de mathématique.

Maintenant vérifions pour le cours de portugais :

Niveau	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendue
1	360	12,45	2,72	12	5	19	14
2	205	12,03	2,59	11	7	19	12
3	52	11,6	2,46	11	7	18	11
4	16	10,88	2	11	8	16	8

Comme pour le cours de mathématique. Les moyennes sont assez proche. Par contre la dispersion des données, semblent plus semblable entre elles. Pour finir, pour finir on retrouve également une plus petite étendue des données à mesure que le niveau augmente. Par contre, cela pourrait être dû au fait que plus le niveau augmente, moins il y a d'observations pour le dit niveau. Faisons une comparaison graphique.

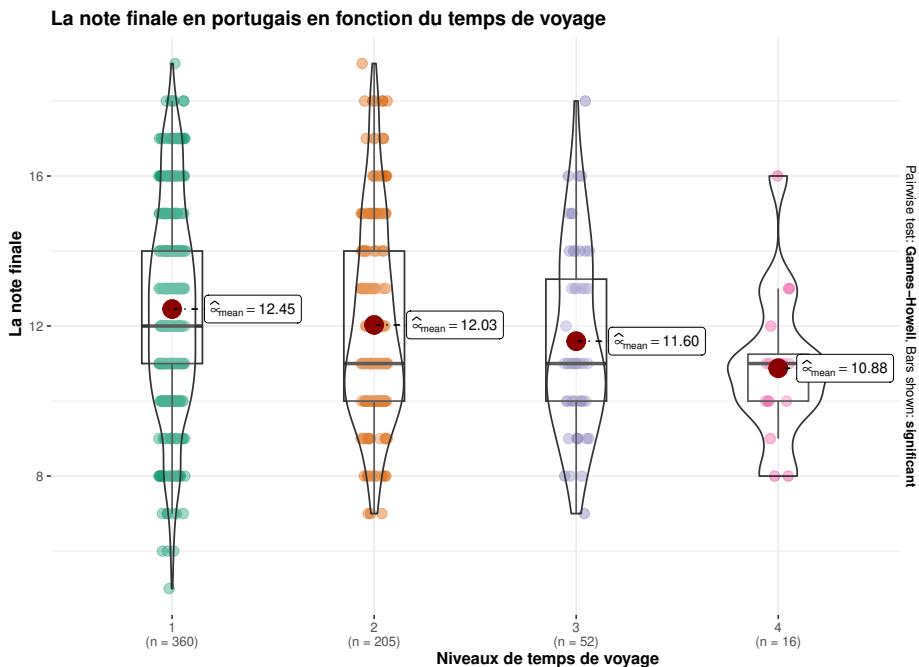


Figure 2: Graphique représentant la relation entre la note finale au cours de portugais et les différents niveaux de temps de voyage pour aller à l'école. Voir annexe B pour le code R

Les variances du niveau 2 et 3 sont semblable, le problème vient du niveau 1 et du niveau 4. Après 2 transformations de variables qui ont été sans succès³, nous faisons un test de normalité⁴ et concluons que les données suivent une loi normale. Donc nous utilisons encore la méthode de Bonferroni avec un test de Welch et nous obtenons que seulement la différence entre le niveau 1 et le niveau 4 possède une p-valeur en-dessous de $\alpha = 0.05$. Nous obtenons 0.0432. Malgré que cette valeur est très proche de α , nous allons considérer que pour le cours de portugais, le temps de voyage pourrait avoir un impacte et garder cette variable pour un futur modèle⁵.

La variable « Temps d'étude »

Maintenant, nous analysons le temps d'étude pour le cours de mathématique. Le temps d'étude est aussi une variable indépendante qui mesure le temps d'étude pour une semaine complète et qui à 4 niveaux qui se décompose comme suit :

- Niveau 1 : 1 à 2 heures exclusivement.
- Niveau 2 : 2 à 5 heures exclusivement.
- Niveau 3 : 5 à 10 heures inclusivement.
- Niveau 4 : Plus de 10 heures d'études.

Niveau	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
1	92	11,47	3,41	11	5	19	14
2	181	11,1	3,03	11	5	19	14
3	59	12,56	2,99	12	7	19	12
4	24	12,67	3,6	12,5	6	20	14

Les moyennes, mis à part pour le 2^e niveau, semblent suivre notre intuition logique comme quoi la moyenne augmente avec le temps d'étude. Pour la dispersion des données, elle semble semblable. Les étendu semble homogène. Regardons graphiquement

³ Voir Annexe A

⁴ Voir Annexe A

⁵ Partie 2, à venir.

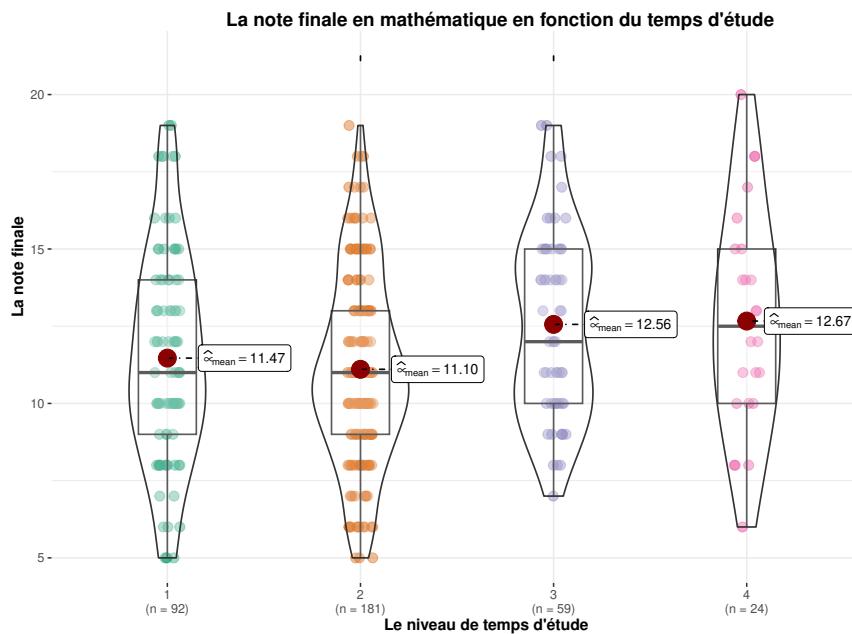


Figure 3: Graphique représentant la relation entre la note finale au cours de mathématique et les différents niveaux de temps d'étude pendant une semaine. Voir annexe B pour le code R pour une décision finale pour cette variable et le cours de mathématique.

Les variances semblent égalent. On peut également observer une relation qui augmente la moyenne de la note finale en mathématique avec une augmentation du temps d'étude. Aussi les données suivent une loi normale⁶. Alors avec un $\alpha=0.05$, on utilise la méthode de Bonferroni couplé avec un test t de student et nous obtenons que seul le niveau 2 et le niveau 3 sont différents significativement avec une p-valeur de 0,0094. Ce qui est possible seulement parce que le niveau 2 est le seul à ne pas suivre la tendance « plus le temps d'étude est grand plus la moyenne de la note finale en mathématique augmente ». Voir la conclusion

Maintenant on passe au cours de portugais :

Niveau	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
1	204	11,27	2,44	11	5	18	13
2	297	12,41	2,61	12	7	19	12
3	97	13,23	2,5	13	8	18	10
4	35	13,06	3,04	13	6	19	13

Nous avons ici des moyennes qui suivent clairement la tendance « plus on augmente le temps d'étude par semaine, plus on augmente la moyenne au cours de portugais ». La dispersion des données est assez semblable, nous vérifierons graphiquement.

6 Voir le test de normalité à l'Annexe A

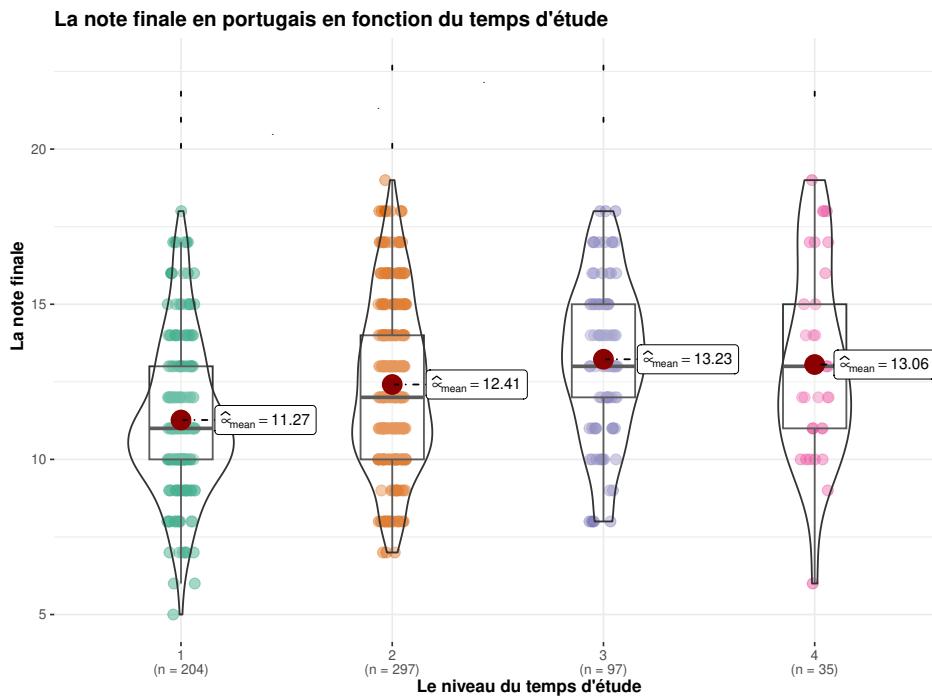


Figure 4: Graphique représentant la relation entre la note finale au cours de portugais et les différents niveaux de temps d'étude pendant une semaine. Voir annexe B pour le code R

temps d'étude va être très importante pour la suite de ce projet⁸.

On peut observer que les variances sont suffisamment égale pour performer un test t de student, et les données suivent la loi normale⁷. Donc nous utilisons la méthode de Bonferroni avec le test t à un $\alpha=0.05$. Et nous obtenons que les niveaux 1 et 2 sont significativement différent avec une p-valeur < 0.0001 , que les niveaux 1 et 3 est aussi différents avec une p-valeur < 0.0001 , ainsi que les groupe 1 et 4 et les groupe 2 et 3, encore avec des p-valeur < 0.0001 . Ce qui veut dire que la variable

La variable « Aide à l'école »

Maintenant, nous analysons l'aide supplémentaire à l'école pour le cours de mathématique. Il n'est pas vraiment expliqué dans les donné ce que représente vraiment l'aide supplémentaire à l'école dans les donnée utilisé pour cette étude. On peut supposer, sans plus, que ce serait l'équivalent de l'aide au devoir qui est souvent proposé au Québec. L'aide à l'école variable indépendante binaire qui se décompose comme suit :

- Niveau 1 : Oui, l'étudiant a reçus de l'aide supplémentaire à l'école.
- Niveau 2 : Non, l'étudiant n'a pas reçus d'aide supplémentaire à l'école.

Regardons les premières données :

7 Voir Annexe A

8 Partie 2 à venir

Groupe	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
non	306	11.86	3.20	11.5	5	20	15
oui	50	9.62	2.55	10.0	5	17	12

Les données sont un peu étrange, l'aide supplémentaire à l'école semble être contre-productif. La dispersion des données entre les niveaux semblent être assez différente. Vérifions graphiquement :

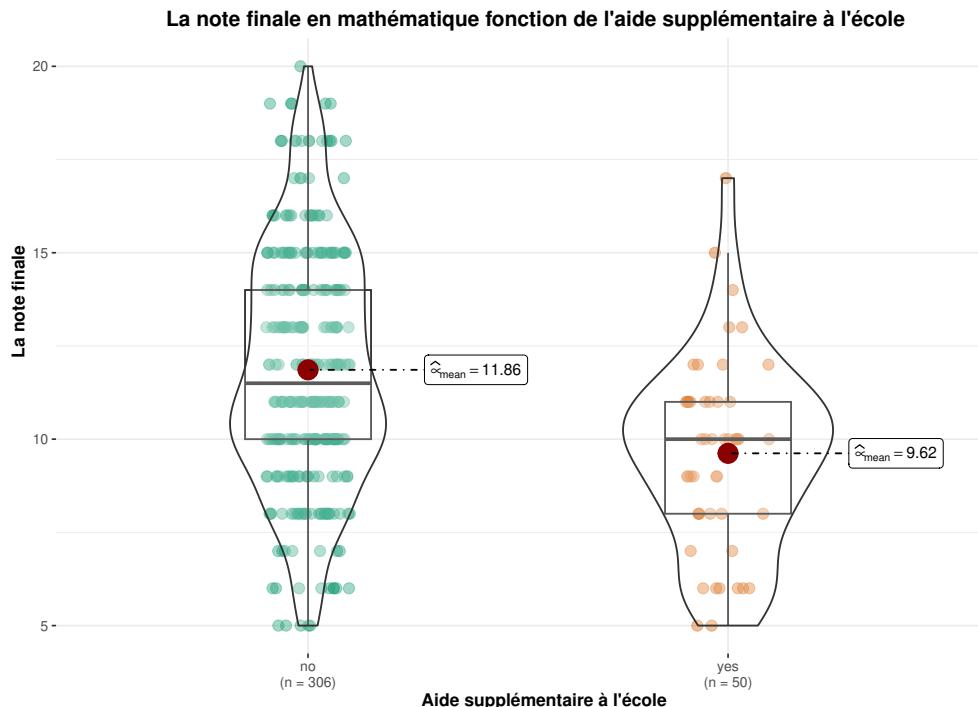


Figure 5: Graphique représentant la relation entre la note finale au cours de mathématique et si l'étudiant a reçus de l'aide à l'école ou non. Voir annexe B pour le code R

Les variances sont clairement différente. Une transformation de variable⁹ a été essayé et a fonctionné. Aussi les données suivent la loi normale¹⁰. Donc un simple test t de student est suffisant pour tester ces 2 niveaux. Et nous trouvons que c'est significatif. Malheureusement, comme nous pouvons le voir, cela semble ne pas être avantageux de donner de l'aide à un étudiant pour le cours de mathématique. Les données fournit nous donne peu d'information sur pourquoi un étudiant a reçus de l'aide supplémentaire en mathématique (ou en portugais). Il se pourrait que ces étudiants soient déjà très mauvais et que l'aide ne leur permet pas de dépasser la moyenne de ceux qui n'avaient pas besoin d'aide au départ. Malgré que cette supposition soit logique et qu'il serait probablement préférable de conserver l'aide à l'école. Nous allons exclure cette variable pour la partie 2 de cette étude.

9 Voir annexe A, et c'était une transformation racine carré.

10 Voir annexe A.

Maintenant pour le cours de portugais :

Groupe	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
non	566	12.30	2.72	12	5	19	14
oui	67	11.45	1.85	11	8	18	10

La moyenne semble montré la même tendance observé pour le cours de mathématique, et ici aussi la dispersion des données entre le groupe du « oui » et le groupe du « non » est assez grande. Regardons avec un graphique :

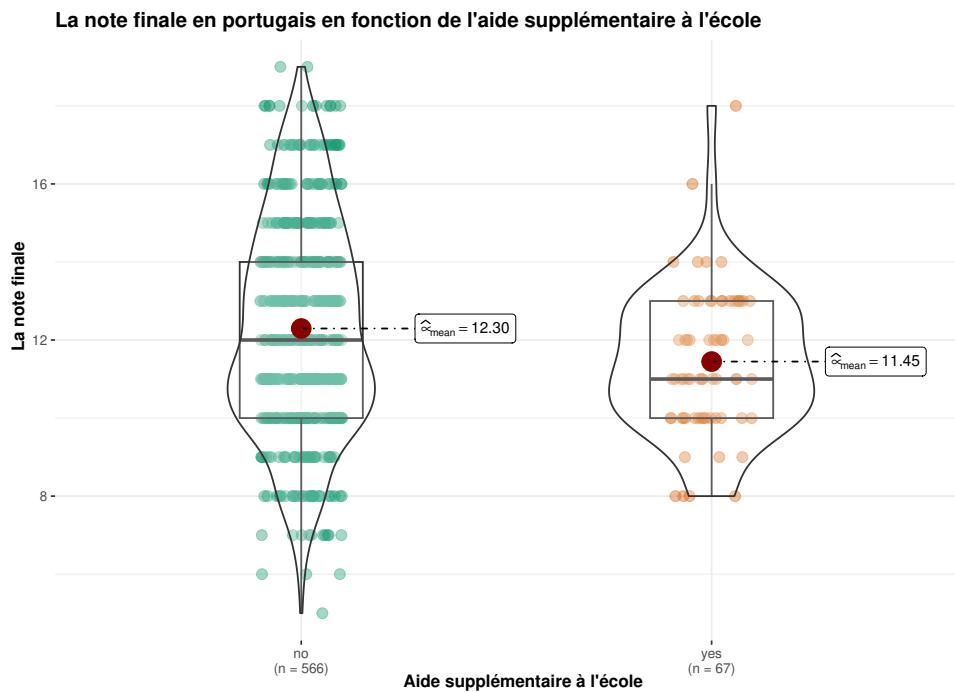


Figure 6: Graphique représentant la relation entre la note finale au cours de portugais et si l'étudiant a reçus de l'aide à l'école ou non. Voir annexe B pour le code R

Les variances sont assez grande, avec la même tendance que la moyenne diminue si l'on a de l'aide à l'école. Par contre, les échecs semblent moins « faible » que les échecs du groupes « non ». Bref, malheureusement après une transformation de donnée¹¹, les variances ne semblent pas s'améliorer. Par contre, les données des 2 groupes suivent une loi normale¹². Donc nous utilisons un test de Welch qui est résistant aux variances différentes et cela nous donne une p-valeur = 0.00116. Par contre nous devons conclure la même chose que pour le cours de mathématique.

11 Voir Annexe A

12 Voir Annexe A

La variable « Aide familiale »

Maintenant, nous analysons l'aide familiale pour le cours de mathématique. Tout comme pour l'aide supplémentaire à l'école, l'aide familiale est très peu expliquée dans ces données. Tout ce qu'on sait, c'est qu'il y a 2 niveaux :

- Niveau 1 : Oui, l'étudiant a reçus de l'aide de la part de sa famille.
- Niveau 2 : Non, l'étudiant n'a pas reçus de l'aide de la part de sa famille.

Commençons pour le cours de mathématique :

Groupe	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
non	137	11.85	3.12	11	5	20	15
oui	219	11.35	3.25	11	5	19	14

Les moyennes sont quasiment identiques et les étendus aussi. La dispersion des données semblent égales pour les 2 groupes. Faisons un graphique pour nous en assurer :

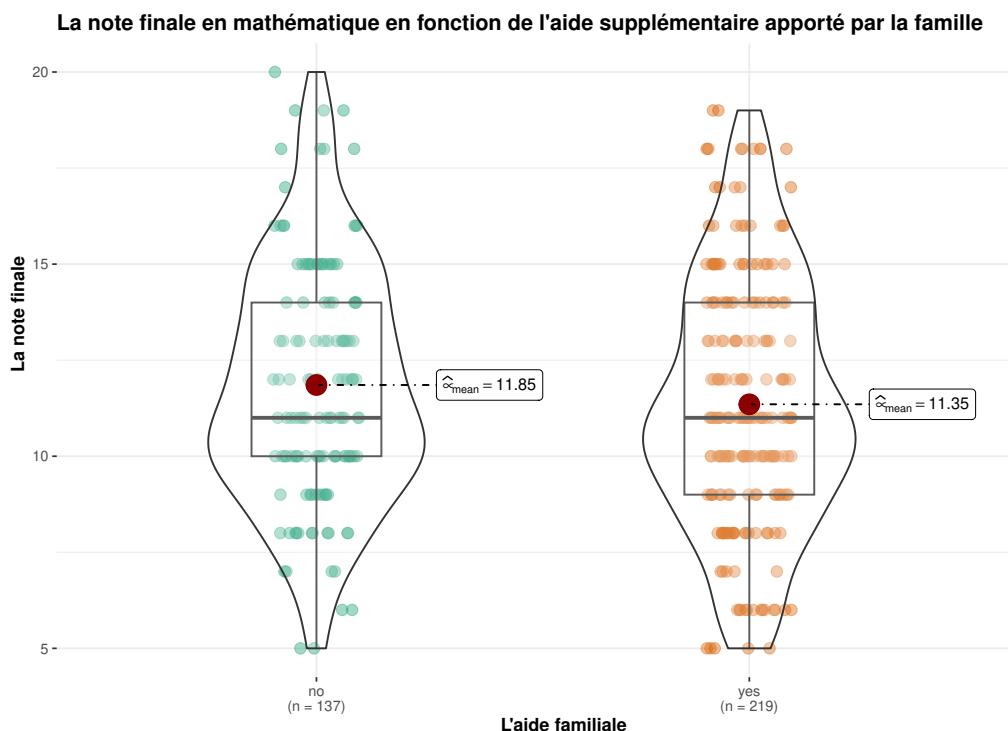


Figure 7: Graphique représentant la relation entre la note finale au cours de mathématique et si l'étudiant a reçus de l'aide familiale ou non. Voir annexe B pour le code R

Les variances apparaissent quasi égale. Par contre, il ne semble pas y avoir de réelle différence entre les 2 graphiques. Les données suivent une loi normale¹³. Donc nous utilisons un test t de student à un niveau $\alpha = 5\%$ et nous obtenons une p-valeur de 0.1469. Donc il n'y a pas de différence significative entre ces 2 niveaux pour la note finale du cours de mathématique.

Regardons maintenant pour le cours de portugais :

Groupe	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
non	240	12.20	2.65	12	5	19	14
oui	393	12.21	2.66	12	6	19	13

Les moyennes sont quasiment identique ainsi que l'écart-type, la médiane et l'étendu (avec les minimum et les maximum). Regardons un graphique :

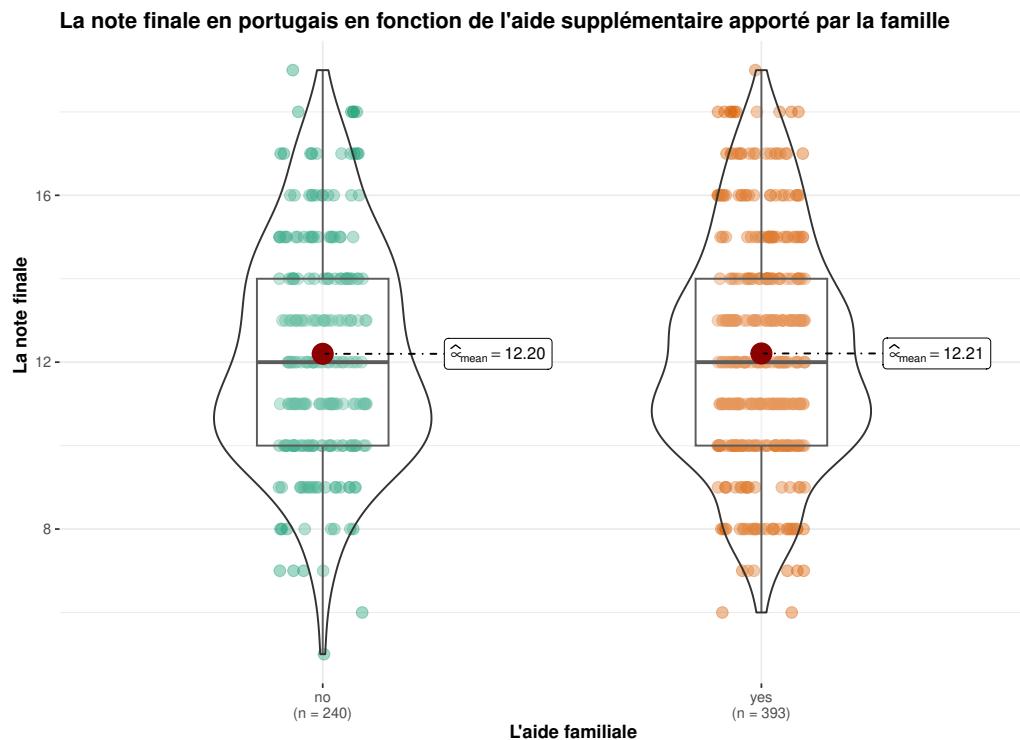


Figure 8: Graphique représentant la relation entre la note finale au cours de portugais et si l'étudiant a reçus de l'aide familiale ou non. Voir annexe B pour le code R

Les variances sont égales, et les données suivent une loi normale¹⁴. Donc on utilise un test t de student et nous obtenons une p-valeur = 0.9683 qui nous montre qu'il n'y a pas de différence significative entre les 2 groupes. L'aide familiale n'aide pas la moyenne des note finale pour le cours de portugais.

13 Voir Annexe A

14 Voir Annexe A

La variable « Cours privés »

Maintenant, nous analysons les cours privés qu'un étudiant a pu prendre pour s'aider lors de son cours de mathématique (ou portugais). Il y a 2 niveaux :

- Niveau 1 : Oui, l'étudiant a eu des cours privés.
- Niveau 2 : Non, l'étudiant n'a pas eu des cours privés.

Regardons pour le cours de mathématique :

Groupe	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
non	184	11.61	3.40	11	5	20	15
oui	172	11.47	2.99	11	5	19	14

Les moyennes sont très proche avec des étendu très proche également. Et la dispersion des données semble proche. Faisons un graphique :

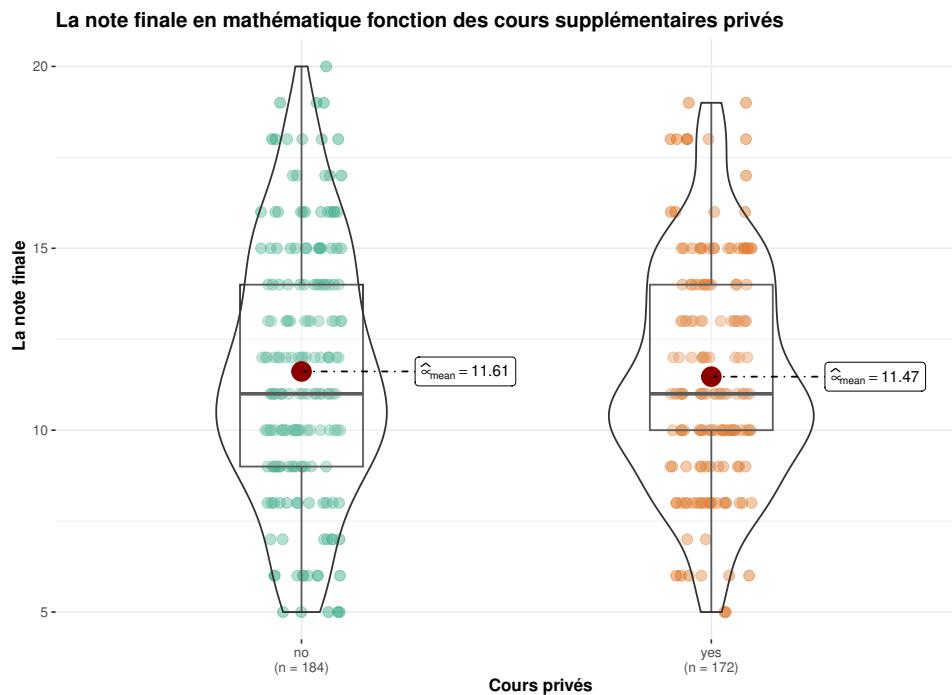


Figure 9: Graphique représentant la relation entre la note finale au cours de mathématique et si l'étudiant a eu des cours privés ou non. Voir annexe B pour le code R

Les variances sont quasi identique et la normalité est respecté¹⁵. Donc nous pouvons utiliser un test t de student. Nous allons utiliser un $\alpha=5\%$ et nous obtenons une p-valeur = 0.6744. Clairement les cours privés en mathématique n'aide pas significativement la moyenne de la note finale du cours de mathématique.

Maintenant regardons le cours de portugais :

Groupe	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
non	595	12.25	2.68	12	5	19	14
oui	38	11.50	2.20	12	7	16	9

Les moyenne semble différer, l'étendu est assez différente. Et il y a eu bien moins grande proportion d'étudiant qui prennent des cours privés pour s'aider dans les cours de portugais que dans les cours de mathématique. La dispersion des données semblent un peu différente. Observons avec le graphique ci-dessous :

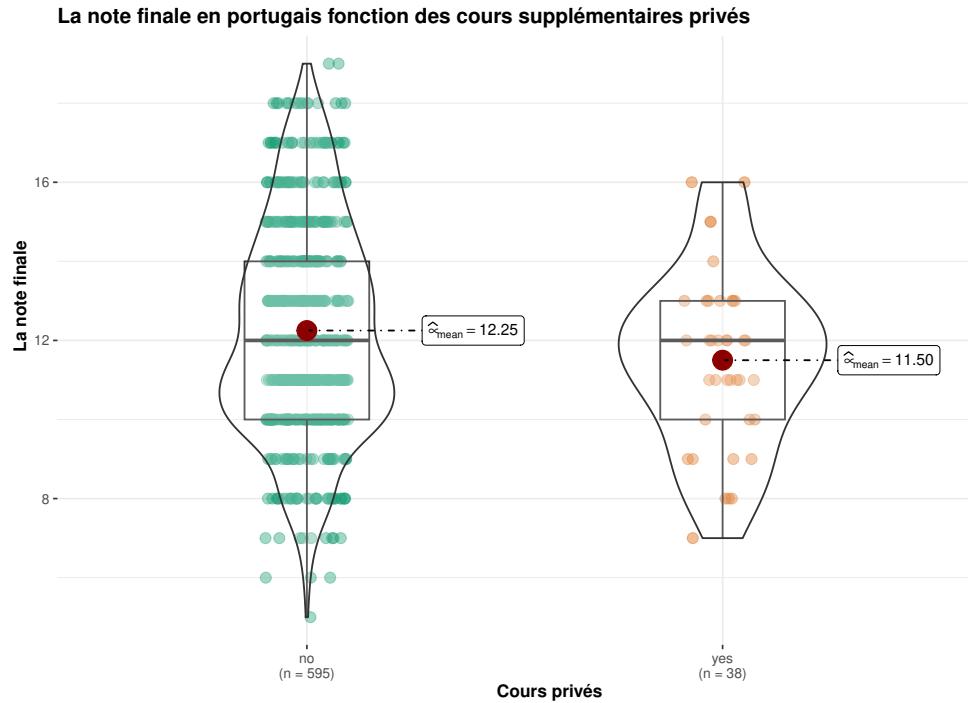


Figure 10: Graphique représentant la relation entre la note finale au cours de portugais et si l'étudiant a eu des cours privés ou non. Voir annexe B pour le code R

Les variances sont vraiment éloigné. Malheureusement les 2 transformations essayées (logarithme et racine carré¹⁶) n'améliore pas la variance des 2 groupes. Par contre les données suivent une loi normale¹⁷. Donc nous pouvons utiliser un test de Welch qui est robuste même lorsque les variances sont différente. Nous obtenons une p-valeur de 0.0508. C'est presque égal à notre risque α . Et vu que la

15 Voir Annexe A

16 Voir Annexe A

17 Voir Annexe A

moyenne des note finale en portugais est inférieur avec des cours privés que sans cours privés. Nous allons considérer que les cours privés ne sont pas un moyen efficace pour améliorer les notes des étudiants québécois. Donc nous n'utiliserons pas cette variable dans le futur¹⁸.

La variable « Internet »

Cette variable est simplement l'accès à l'internet à domicile d'un étudiant. Si il possède internet, on peut supposer qu'il pourrait trouver de l'aide en ligne et donc peut être l'aider dans un cours (mathématique ou portugais). Donc elle a 2 niveaux :

- Niveau 1 : Oui, l'étudiant a l'internet chez lui.
- Niveau 2 : Non, l'étudiant n'a pas l'internet chez lui.

Regardons pour le cours de mathématique :

Groupe	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
non	58	10.71	2.97	10	5	18	13
oui	298	11.71	3.23	11	5	20	15

Il semblent y avoir peu de gens qui n'ont pas internet et c'est eux qui ont la pire moyenne des 2 groupes. L'étendu est semblable et la dispersion des données est proche. Regardons ce graphique :

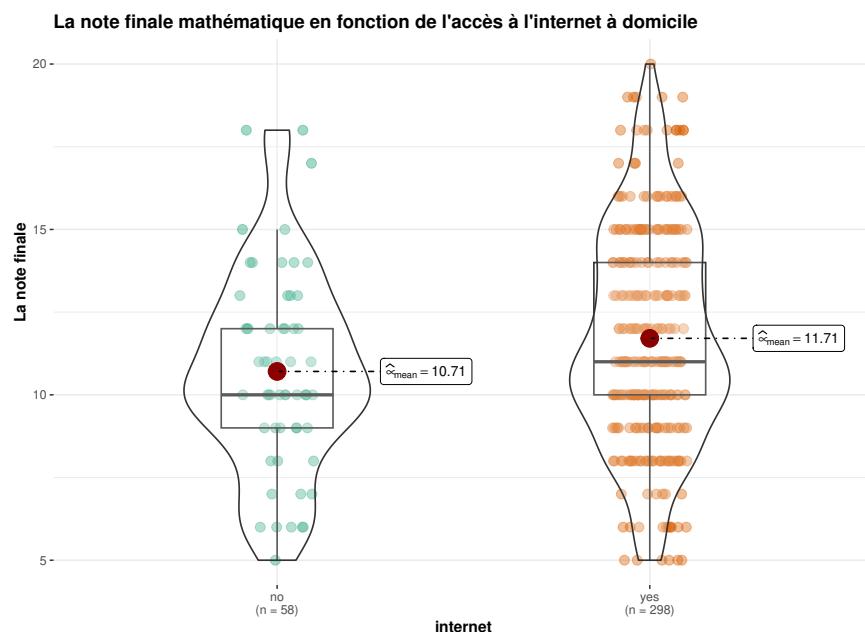


Figure 11: Graphique représentant la relation entre la note finale au cours de mathématique et si l'étudiant a accès à l'internet à domicile. Voir annexe B pour le code R

18 Partie 2 à venir

Les variances sont proche, essayons une transformation de variable logarithme¹⁹ qui a pour propriété de rendre la dispersion des données plus proche et de permettre d'utiliser un test t de student. Cela a effectivement améliorer la similitude des variances. Et la normalité est respecté²⁰. Donc nous pouvons utiliser un test t de student qui nous donne une p-valeur de 0.0353 qui est inférieur à notre α de 5%. Ce qui montre qu'il y a un effet entre la variable internet la note finale en mathématique.

Regardons maintenant pour le portugais :

Groupe	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
non	144	11.56	2.49	11	7	19	12
oui	489	12.39	2.68	12	5	19	14

Il y a un peu plus d'étudiants qui n'ont pas d'internet qui suivent un cours de portugais. Les moyennes sont différente avec une étendu très proche. La dispersion semblent aussi très proche. Observons ce graphique :

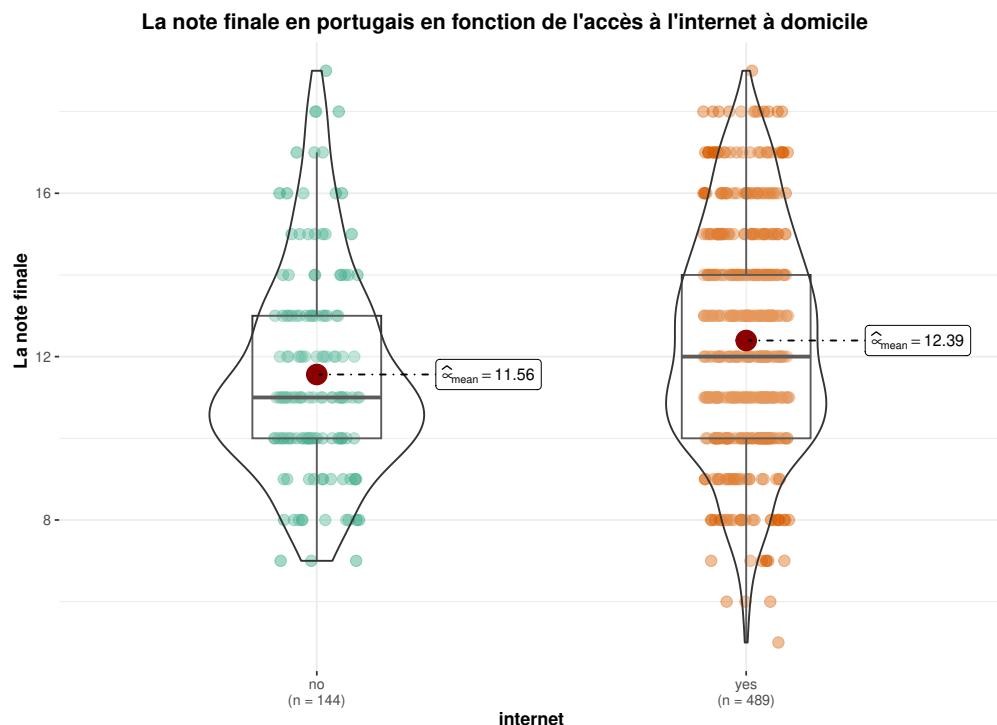


Figure 12: Graphique représentant la relation entre la note finale au cours de mathématique et si l'étudiant a accès à l'internet à domicile. Voir annexe B pour le code R

Les variances sont égale et la normalité est respecté²¹. On peut donc faire un test t de student et nous obtenons une p-valeur de 0.000919 qui est très significatif. Et vu que la moyenne de ceux qui ont accès à internet est supérieur en portugais que ceux qui n'ont pas accès, cela prouve que c'est une variable intéressante pertinente à explorer pour un modèle futur.

19 Voir annexe A

20 Voir Annexe A

21 Voir Annexe A

La variable « Consommation d'alcool les jours d'école»

Cette variable est n'est pas très bien défini quantitativement. C'est une variable qualitative qui est définit qualitativement. Elle possède 4 niveaux qui se décompose comme suit :

- Niveau 1 : Très faible consommation.
- Niveau 2 : Faible consommation.
- Niveau 3 : Moyenne consommation
- Niveau 4 : Grande consommation.
- Niveau 5 : Très grande consommation

Regardons pour le cours de mathématique :

Groupes	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
1	250	11.85	3.30	11	5	20	15
2	63	10.95	3.01	11	5	18	13
3	25	10.92	2.75	10	7	17	10
4	9	9.89	2.62	9	5	13	8
5	9	10.67	2.69	11	5	13	8

Les moyennes sont pour la plupart différentes entre les différents niveaux. Les étendues également. Par contre à partir du niveau 3, il y a clairement moins d'observations. Ce qui pourrait expliquer la variabilité. Pour finir la dispersion des données ne semblent pas égales entre les différents niveaux.

Vérifions avec un graphique :

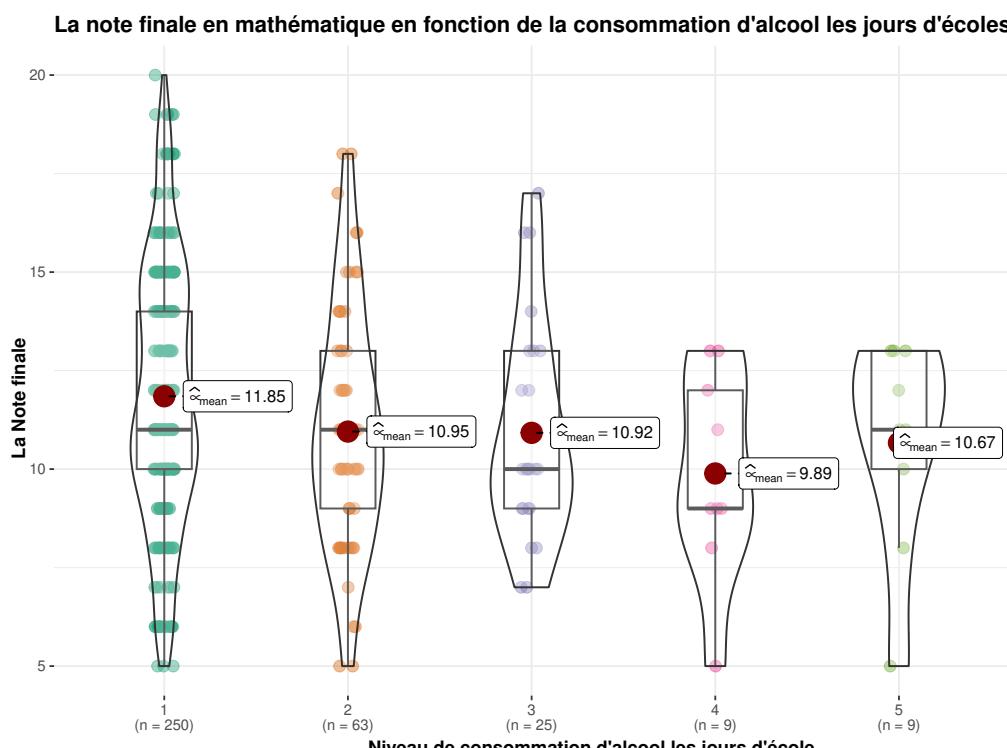


Figure 13: Graphique montrant la relation entre la note finale au cours de mathématique et la consommation d'alcool de l'étudiant les journées où il a cours. Voir annexe B pour le code R

Les variances semblent vraiment différente. Et la consommation d'alcool diminue les notes sauf pour le dernier niveau. Par contre le dernier niveau, ne possède que 9 observations. Une transformation logarithmique a été effectué pour améliorer les variances²². Mais cela n'a pas améliorer les variances. Par contre les données suivent la loi normale²³. Donc vu que nous avons plus de 2 groupes, et des variances inégales, nous utilisons la méthode de Bonferroni et un test de Welch avec un α de 5% et nous obtenons aucune p-valeur significative. Selon nos données, boire de l'alcool la semaine ne semblent pas changer la moyenne des notes finales en mathématique.

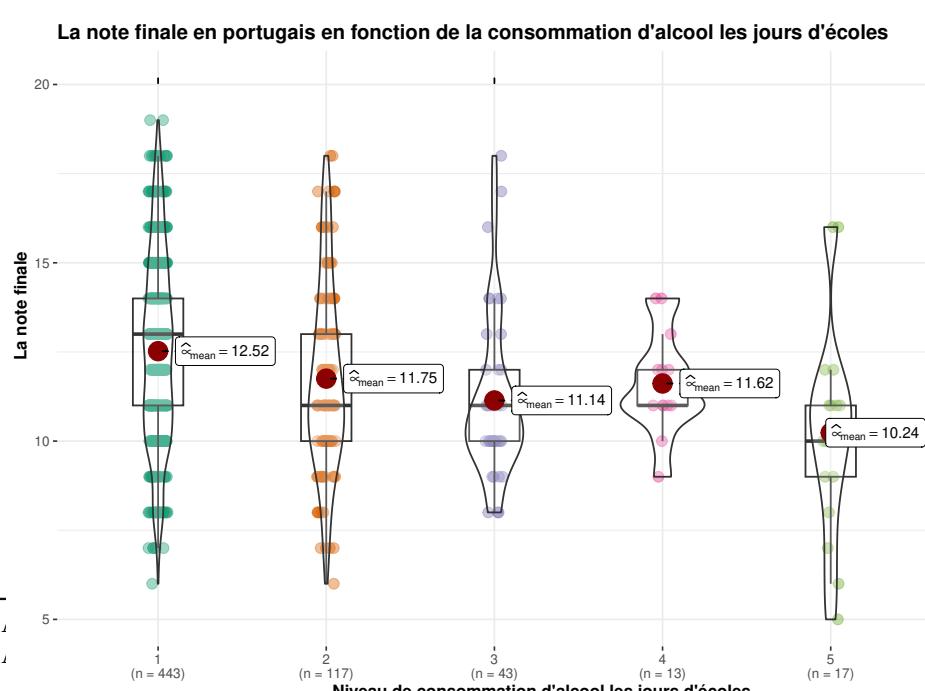
On regarde maintenant pour le cours de portugais :

Groupe	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
1	443	12.52	2.65	13	6	19	13
2	117	11.75	2.62	11	6	18	12
3	43	11.14	2.25	11	8	18	10
4	13	11.62	1.45	11	9	14	5
5	17	10.24	2.95	10	5	16	11

Il y a encore cette dichotomie entre le nombre de buveurs d'alcool de niveau 3 et plus et les niveaux 1 et 2. Plus le niveau augmente, plus la moyenne baisse. Et les étendu sont très semblable, sauf pour le groupe 4, probablement à cause du faible nombre de d'observations. Pareil pour la dispersion des données qui

est plus ou moins proche, sauf pour le niveau 4.

Observons avec un graphique.



22 Voir Annexe .

23 Voir Annexe .

Figure 14: Graphique montrant la relation entre la note finale au cours de portugais et la consommation d'alcool de l'étudiant les journées où il a cours. Voir annexe B pour le code R

Les variances sont très différentes. Et après 2 différentes transformations de variables²⁴, les variances ne s'améliore pas. La normalité est par contre respecté²⁵. Donc on utilise la méthode de Bonferroni et un test de Welch et cela nous donne 2 p-valeur significative. Celle entre le niveau 1 et le niveau 2 avec une p-valeur de 0,0481, qui est très proche de 5%. Et celle entre le niveau 1 et le niveau 3 avec une p-valeur = 0,00358. Bref la variable consommation d'alcool les jours d'école est significative pour les cours de portugais.

La variable « Consommation d'alcool les fin de semaine»

Cette variable est définie de manière identique à la variable précédente (consommation d'alcool les jours d'école). Elle possède 4 niveaux qui se décompose comme suit :

- Niveau 1 : Très faible consommation.
- Niveau 2 : Faible consommation.
- Niveau 3 : Moyenne consommation
- Niveau 4 : Grande consommation.
- Niveau 5 : Très grande consommation

Regardons pour le cours de mathématique :

Groupe	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
1	133	12.19	3.48	12	6	20	14
2	73	11.74	2.98	11	5	19	14
3	77	11.14	3.09	11	5	18	13
4	47	10.43	2.62	10	5	17	12
5	26	10.92	3.08	10	5	18	13

24 Voir Annexe A

25 Voir Annexe A

Le nombre de buveurs est mieux répartis (bien qu'hétérogène) qu'avec la variable « consommation les jours d'école ». Les moyennes descendent au fur et à mesure que le niveau de prise d'alcool augmente (mis à part pour le niveau 4). L'étendu est quasi identique et la dispersion des données semblent proche. Regardons avec ce graphique :

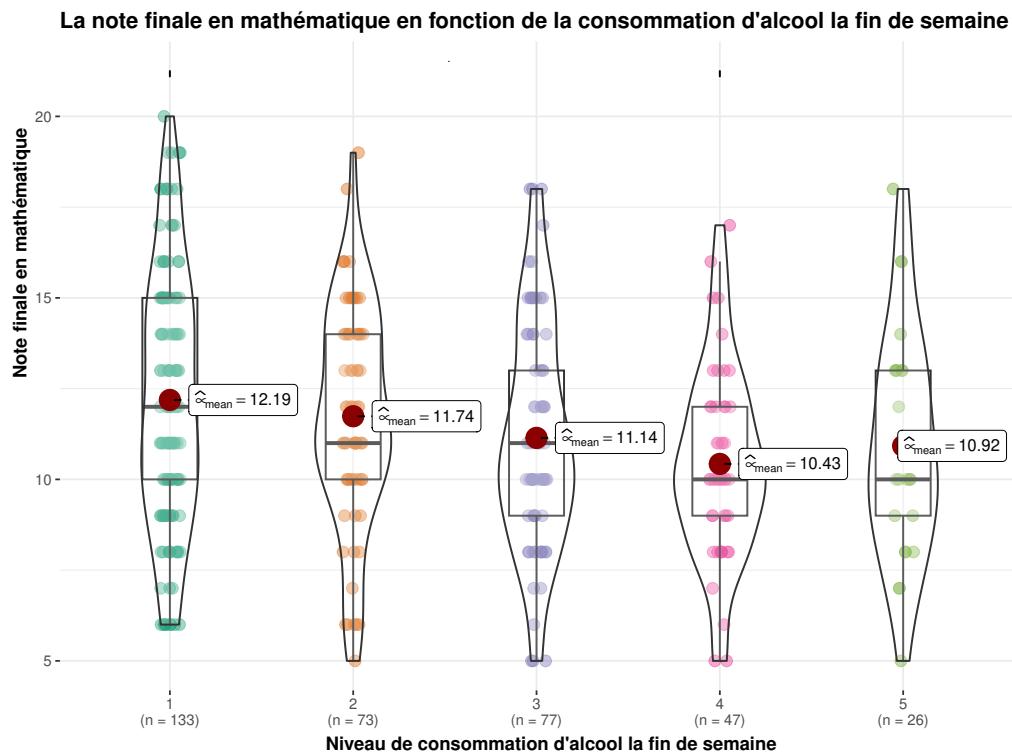


Figure 15: Graphique montrant la relation entre la note finale au cours de mathématique et la consommation d'alcool de l'étudiant les fin de semaines. Voir annexe B pour le code R

Les variances entre les différents niveaux sont clairement égales, et l'hypothèse de normalité est respectée²⁶. On utilise la méthode de Bonferroni avec un test t de student et nous obtenons qu'une seule différence de 2 niveaux significative, le niveau 1 avec le niveau 3 avec une p-valeur de 0,0163. Donc la consommation d'alcool la fin de semaine affecte la moyenne de la note finale en mathématique.

Regardons pour le cours de portugais :

Groupe	N	Moyenne	Écart-type	Médiane	Minimum	Maximum	Étendu
1	243	12.56	2.63	13	7	19	12
2	148	12.43	2.75	12	6	18	12
3	114	12.28	2.52	12	8	18	10
4	85	11.28	2.31	11	7	19	12
5	43	11.05	2.81	11	5	17	12

Encore une fois, les observations sont mieux réparties et les moyennes descendent au fur et à mesure que le niveau de prise d'alcool la fin de semaine augmente. Les étendus sont presque toutes égales et la dispersion des données semblent très proches. Observons avec un graphique.

26 Voir Annexe A

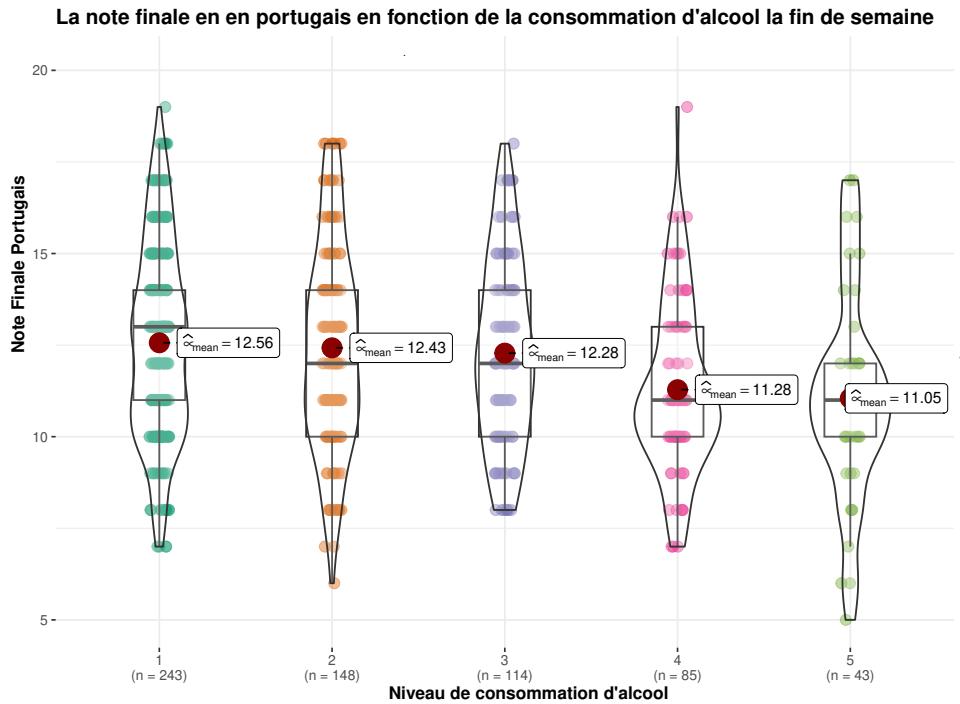


Figure 16: Graphique montrant la relation entre la note finale au cours de portugais et la consommation d'alcool de l'étudiant les fin de semaines. Voir annexe B pour le code R

Les variances sont égales et les données respecte l'hypothèse de normalité²⁷. Donc on peut utiliser la méthode de Bonferroni avec le test t de student et nous obtenons plusieurs différence de niveaux significative. Le niveau 1 et 4 avec une p-valeur < 0.001, le niveau 1 et 5 avec une p-valeur = 0.0059, le niveau 2 et 3 avec une p-valeur de 0.013, le niveau 2 et 4 avec un p-valeur=0.01279, le niveau 2 et 5 avec une p-valeur de 0,04016, le niveau 3 et 4 avec une p-valeur de 0,04213. Bref, boire de l'alcool la fin de semaine n'est clairement pas bon pour la moyenne de la note finale du cours de portugais. C'est une variable très significative.

Conclusion

En conclusion, cette étude observationnelle nous a permis de montré qu'il y avait des variables significative avec la réussite scolaire dans les matières comme les mathématiques ou le portugais et d'autres que non. Parmi les variable significative nous avons le temps de voyage, mais seulement pour le cours de portugais. Nous avons le temps d'étude pour le cours de portugais. Pour le cours de mathématique, il semble avoir une p-valeur significative seulement à cause d'une aberration, le niveau 2 est plus bas que le niveau 1, mais le niveau 3 est plus haut que le niveau 2 et le niveau 1. Mais nous

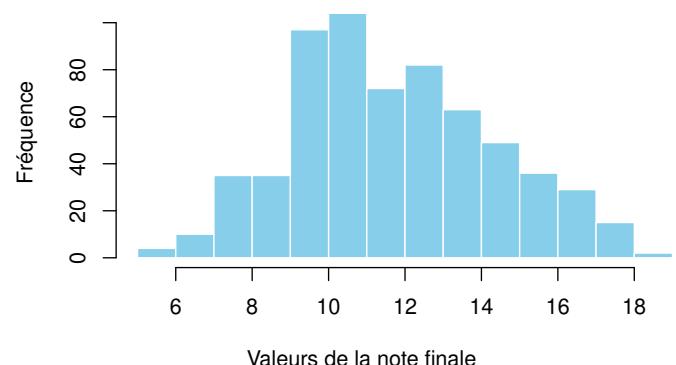
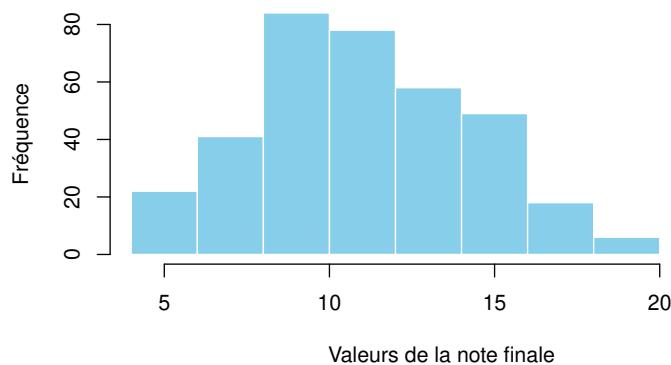
²⁷ Voir Annexe C

allons quand même considérer cette variable comme significative, car dans la partie 2 de cette étude, nous tenterons d'utiliser les variables que nous avons trouvé ici significative, pour en faire un modèle prédisant l'amélioration de la note pour une matière donnée (ici mathématique et portugais) et peut être que la combinaisons du temps d'étude en mathématique avec une autre variable sera incontestablement significative. Ensuite nous avons l'accès à l'internet à domicile qui est significative. Étrangement, seul le cours de portugais était affecté par la consommation d'alcool les jours d'école. Et les 2 cours était affecté par la consommation d'alcool les fin de semaine. Mais, le cours de portugais était beaucoup plus affecté que le cours de mathématique. Bref ceci conclut la partie 1 de cette étude observationnelle.

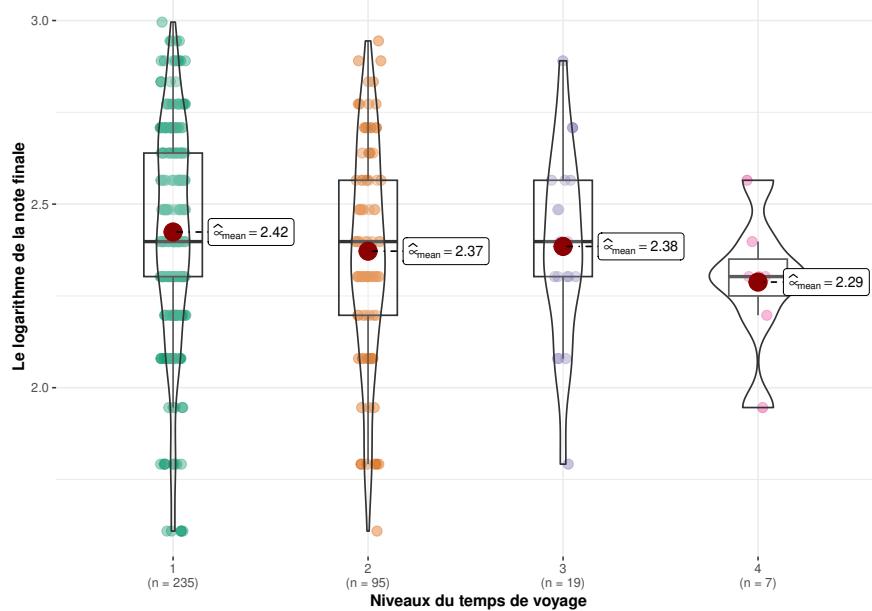
Annexe

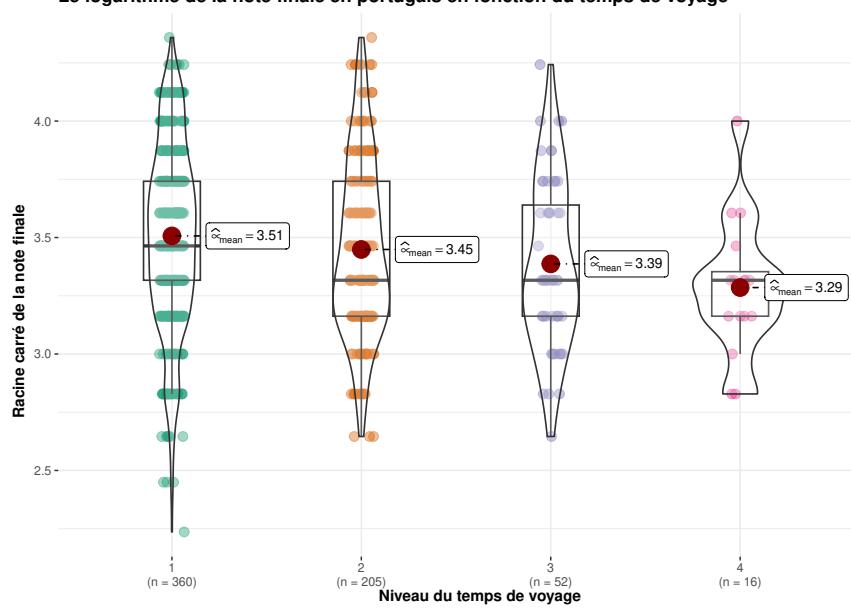
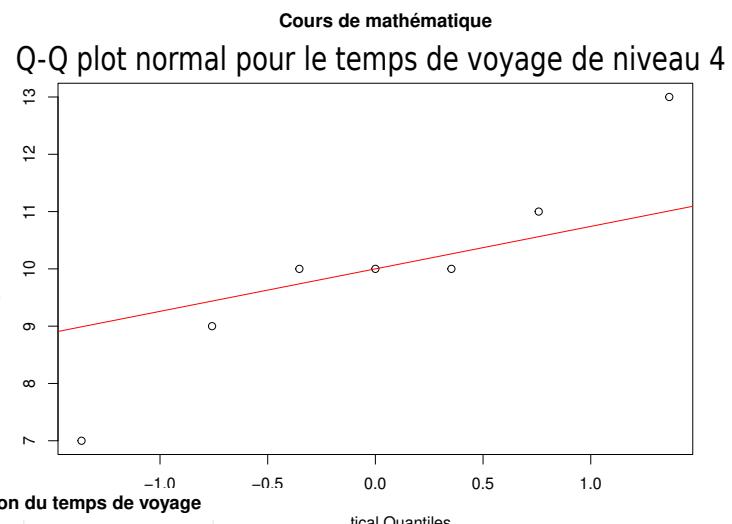
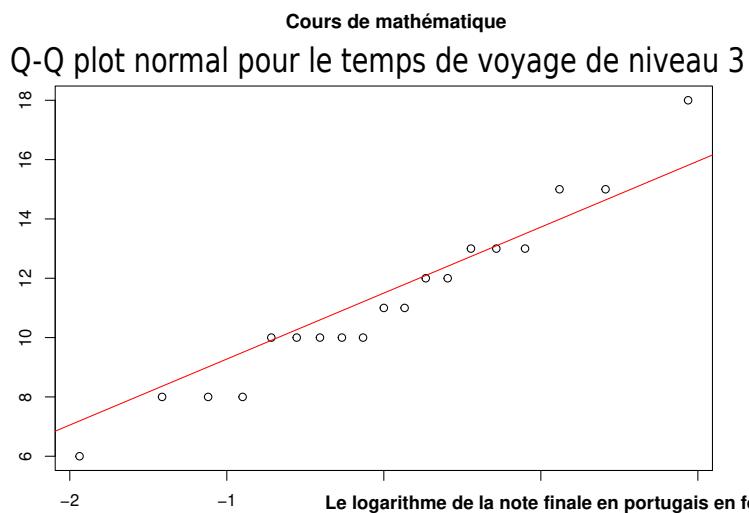
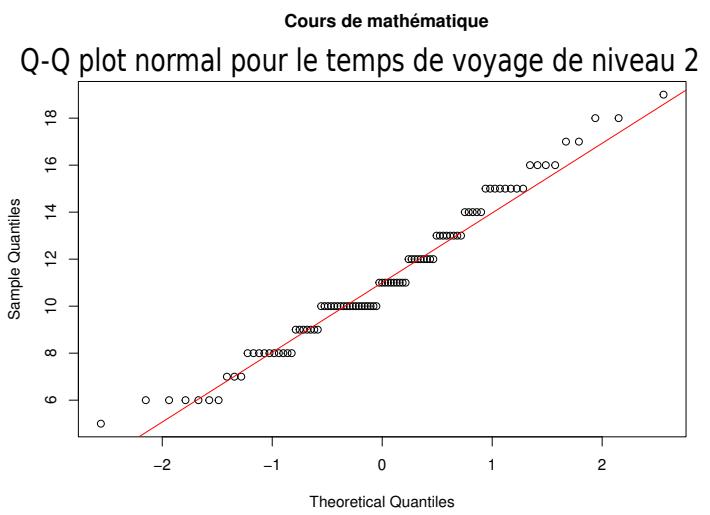
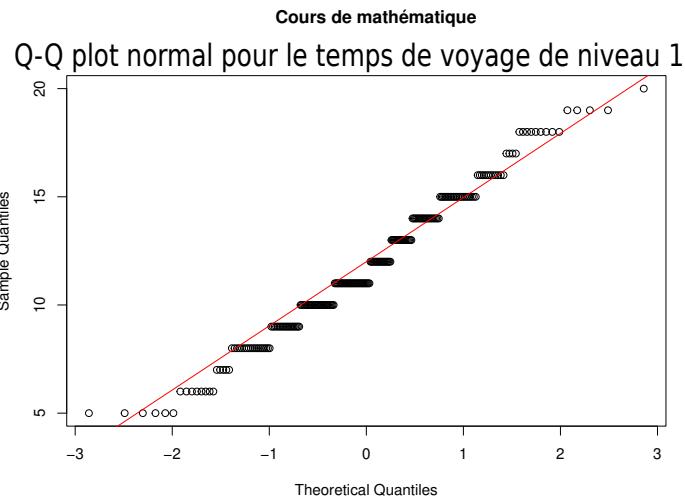
Annexe A

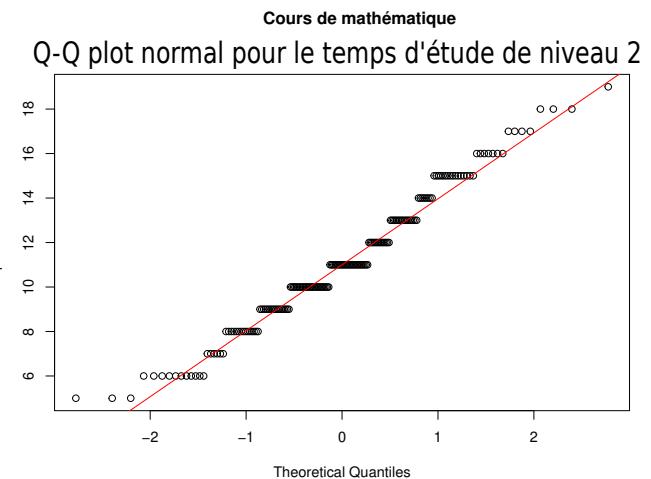
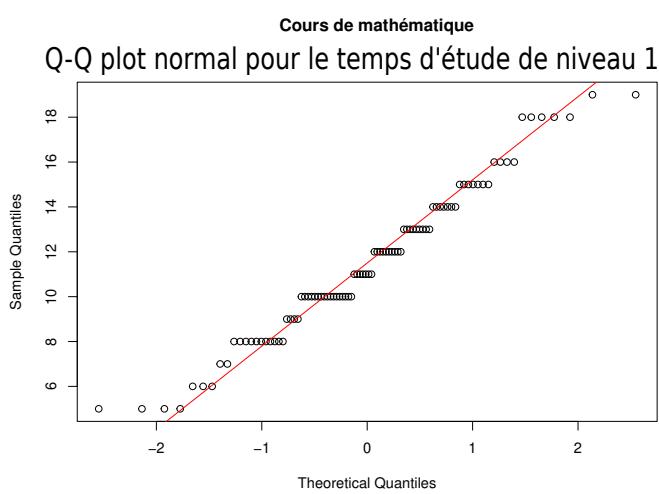
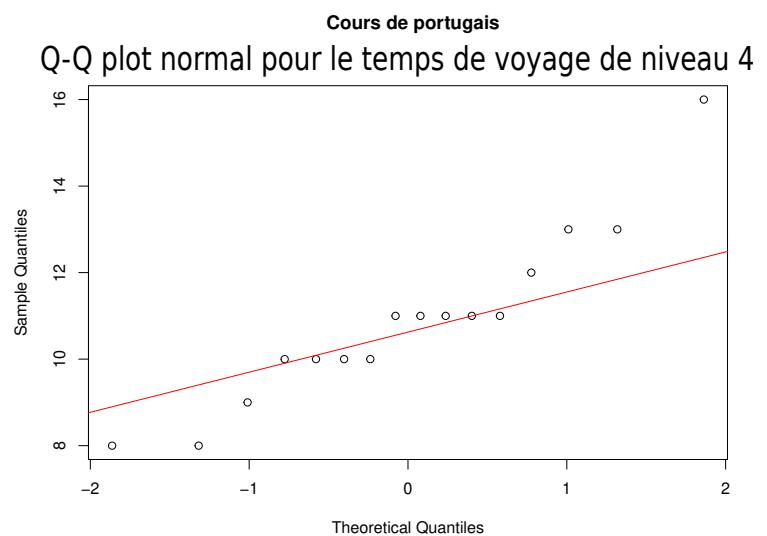
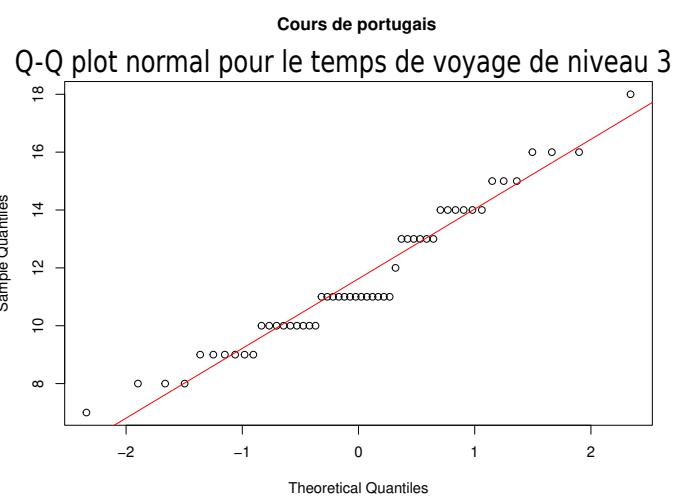
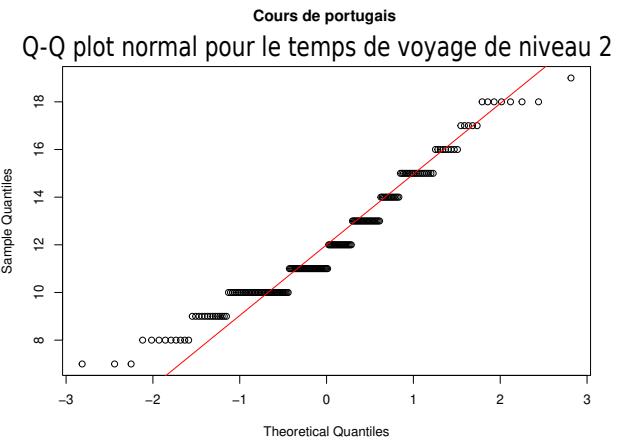
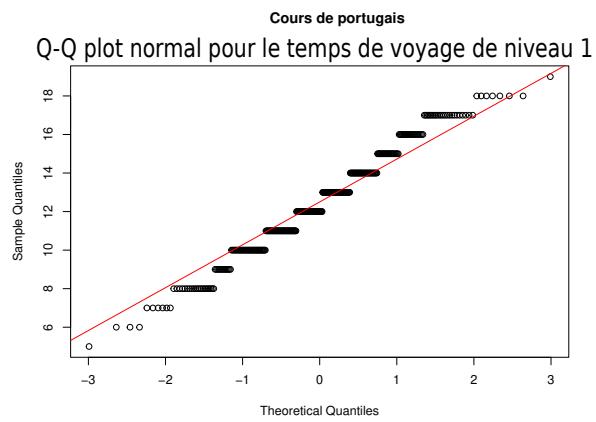
Histogramme de la variable la note finale au cours de mathématique Histogramme de la variable la note finale au cours de portugais

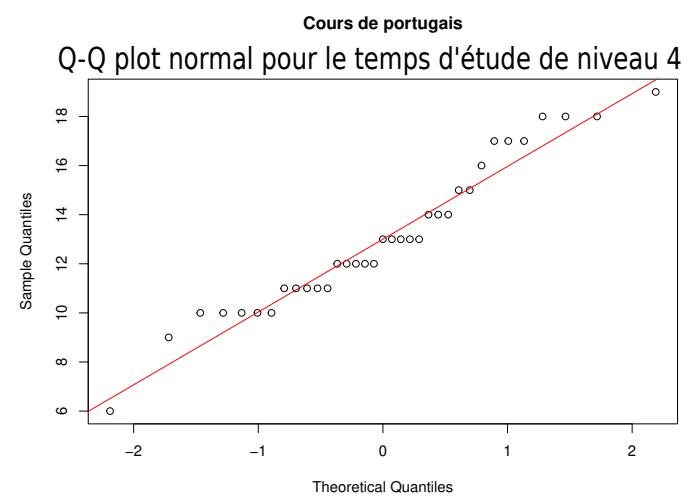
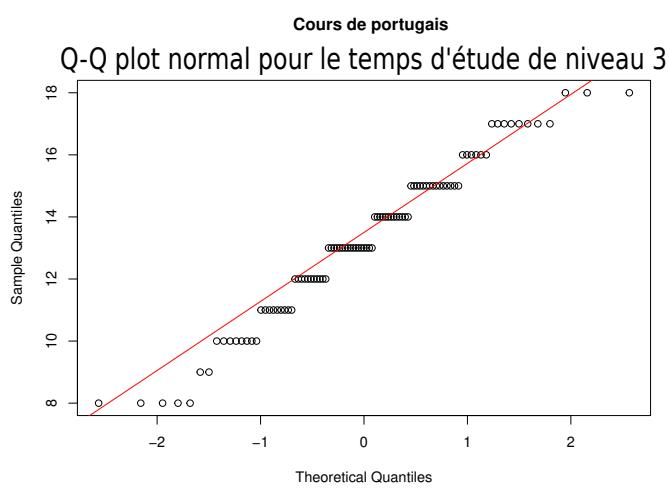
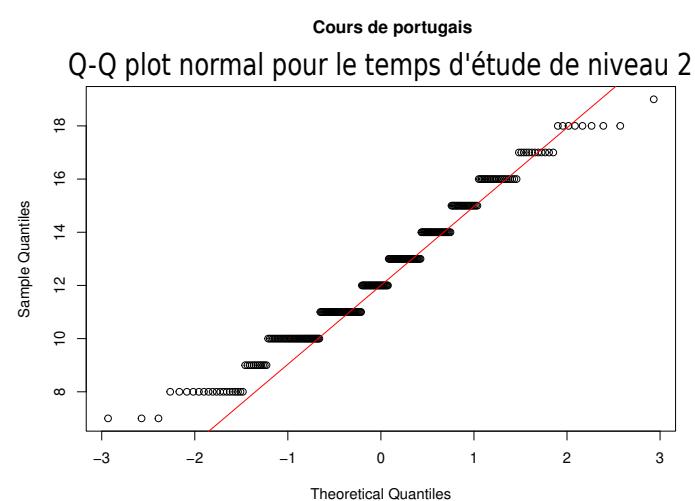
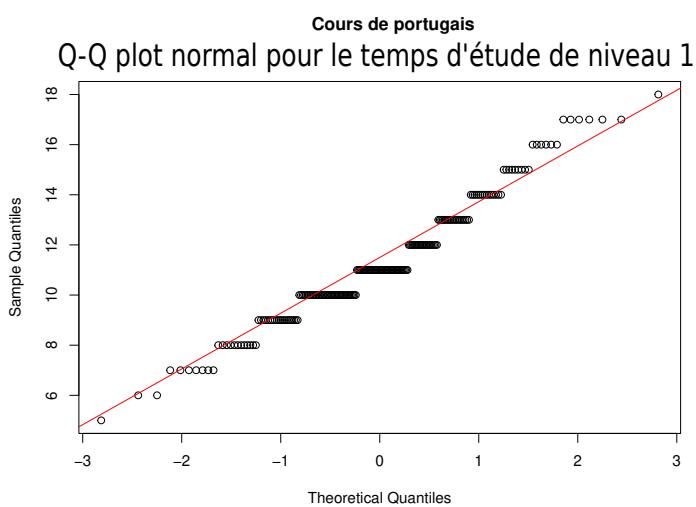
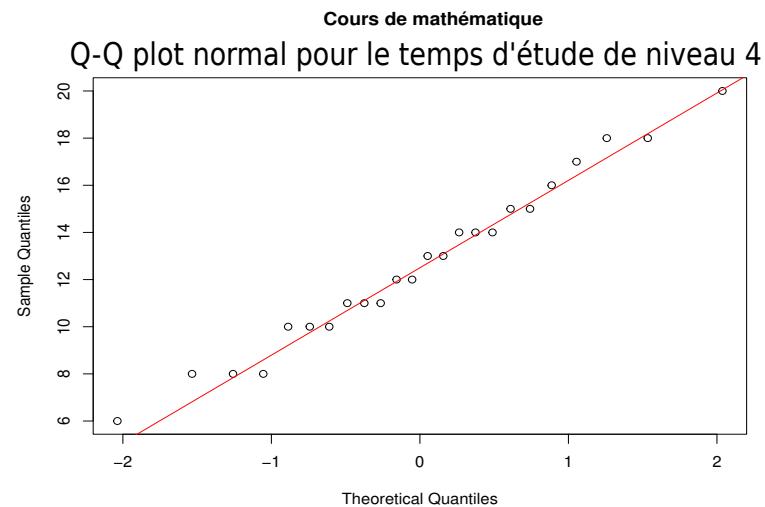
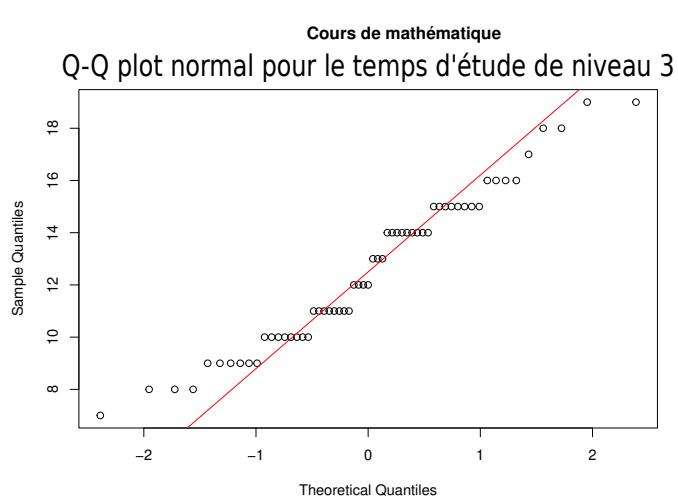


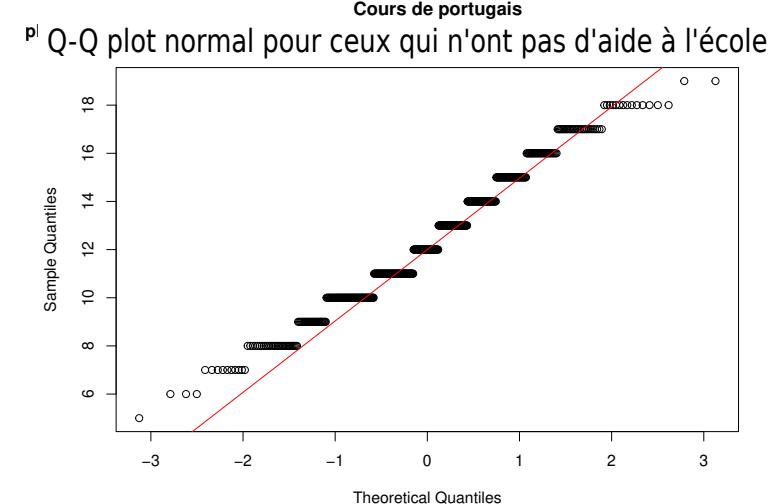
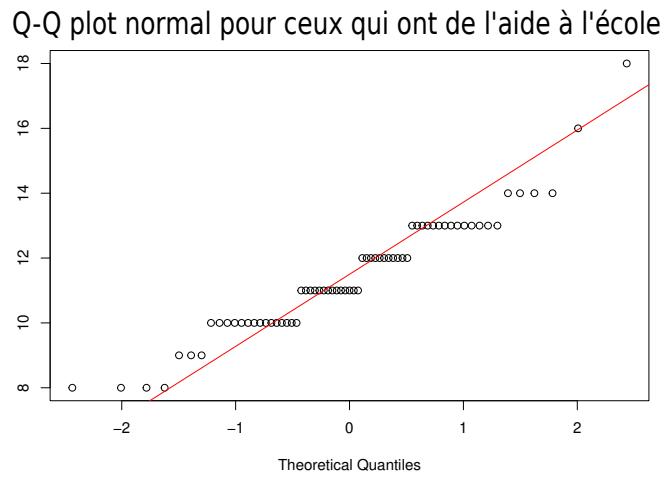
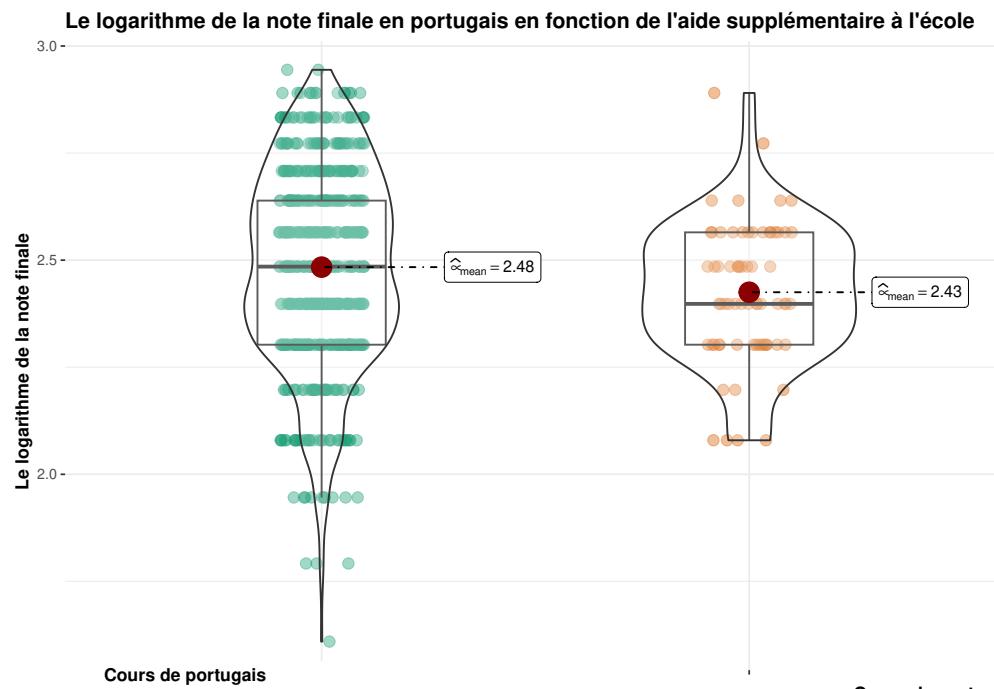
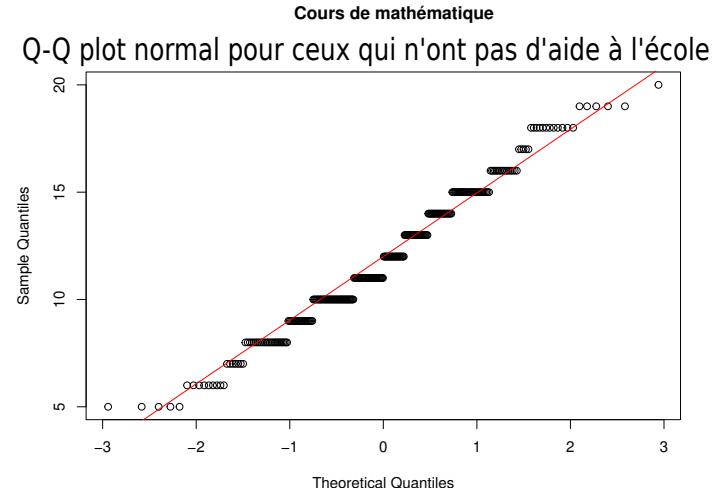
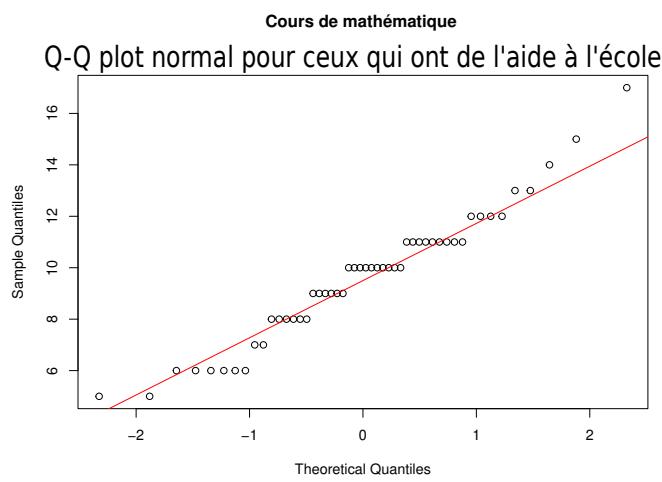
Le logarightme de la note finale en mathématique en fonction du temps de voyage

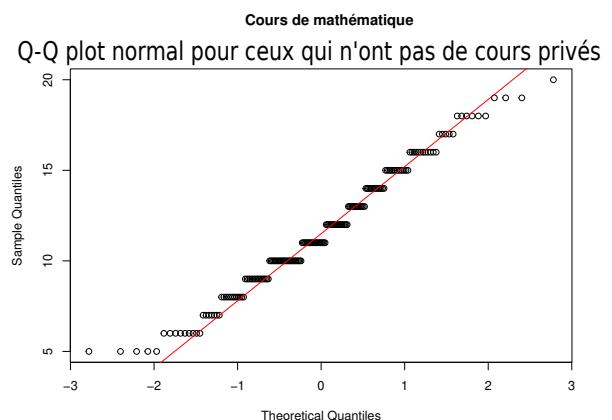
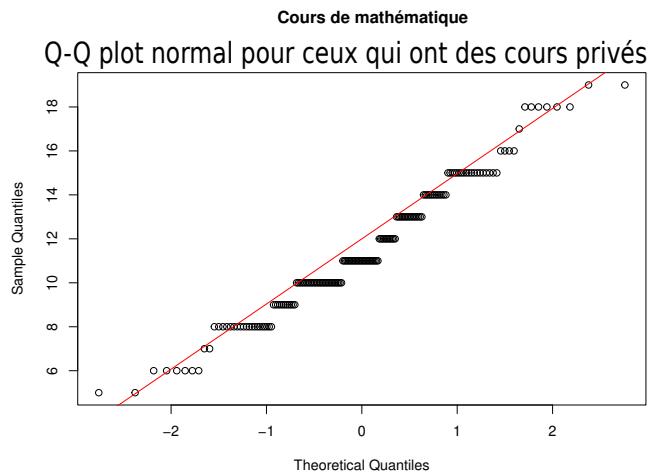
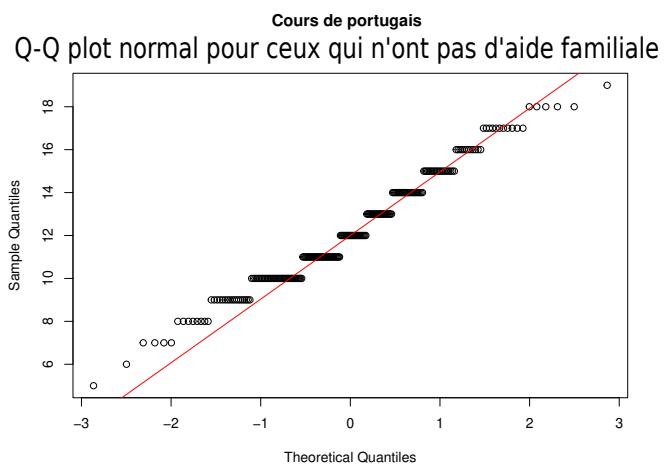
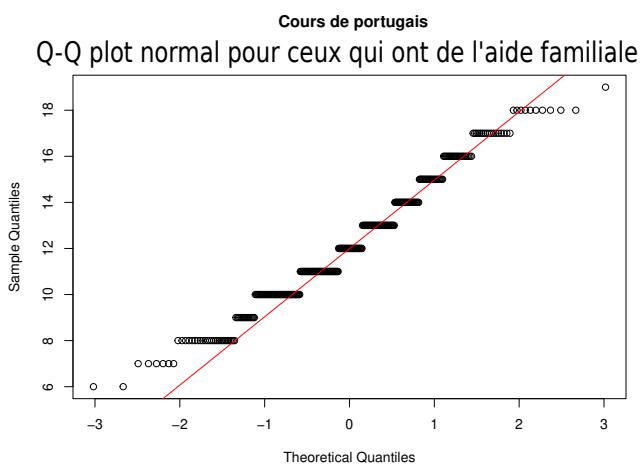
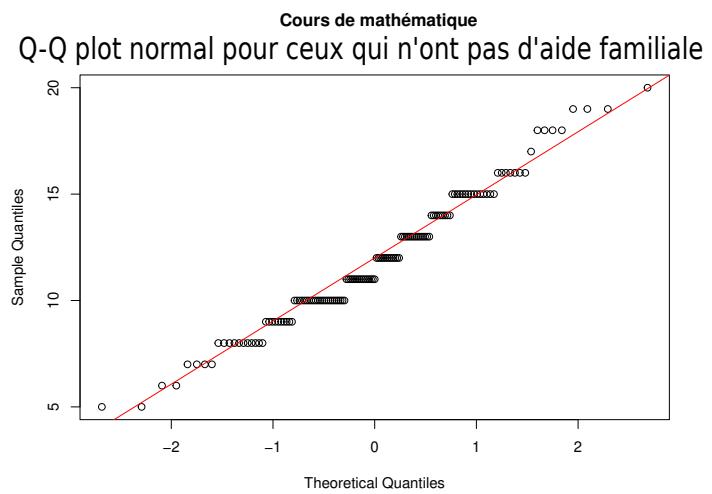
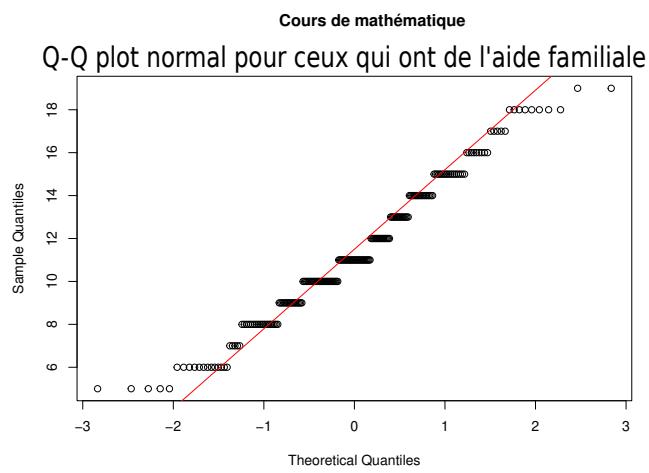


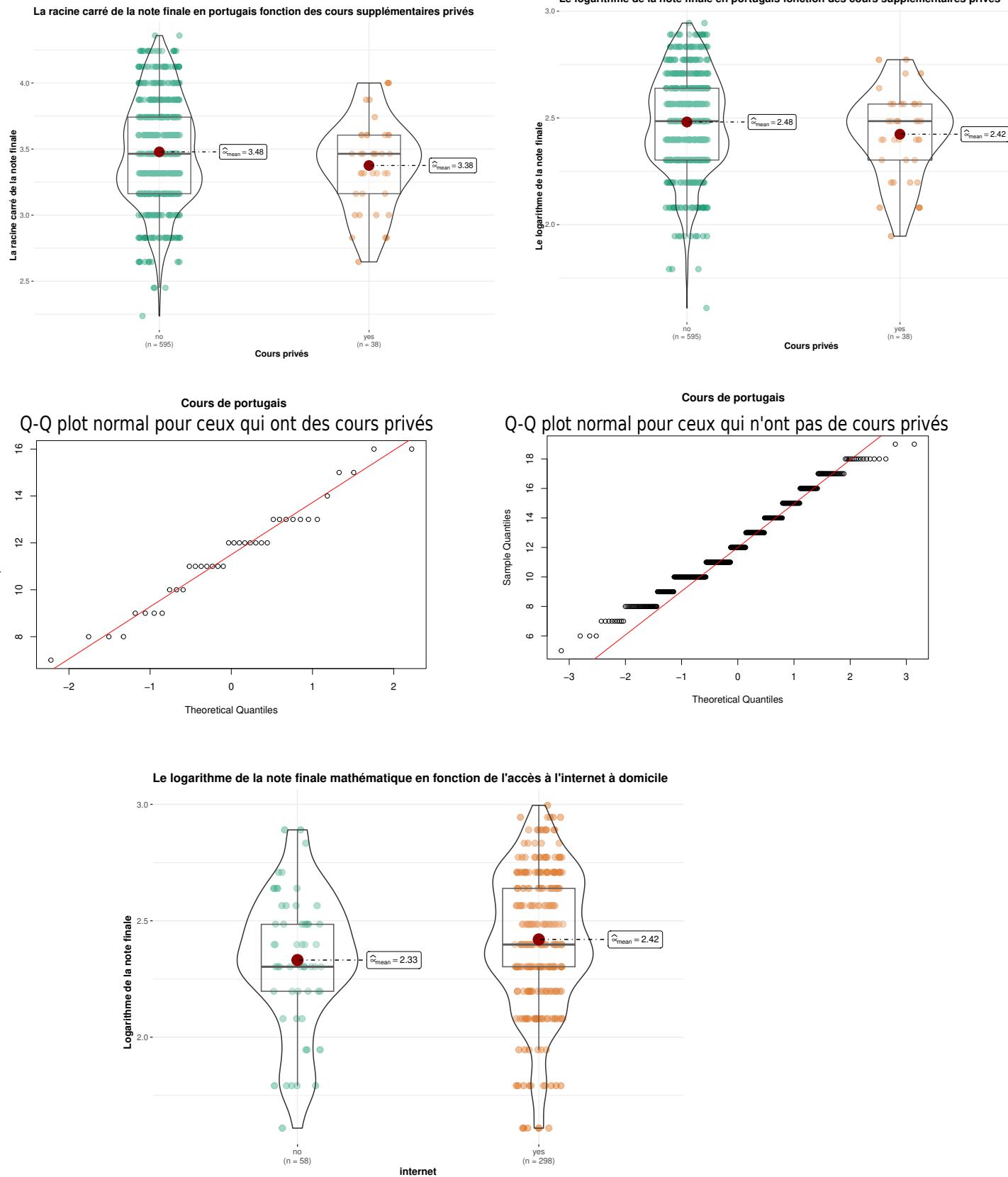


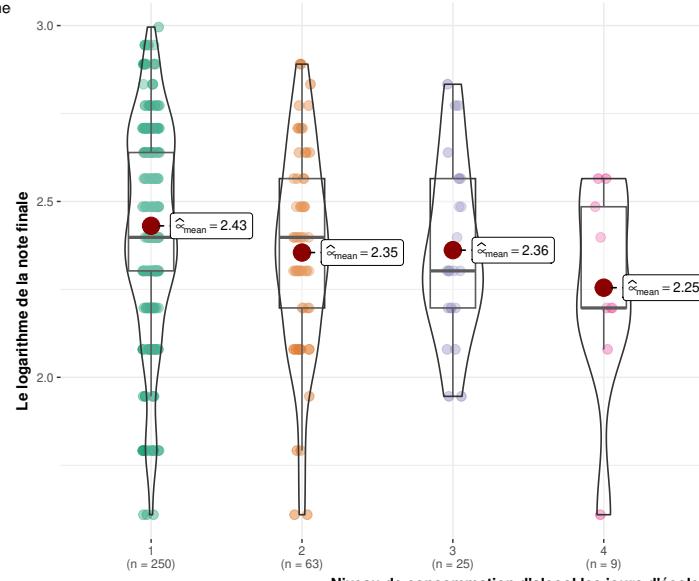
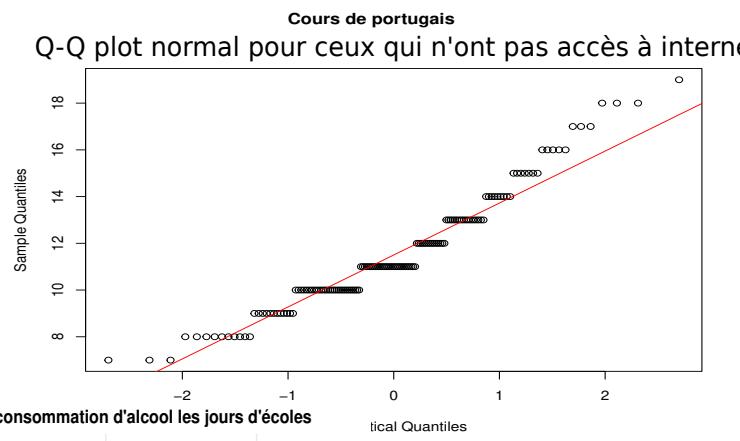
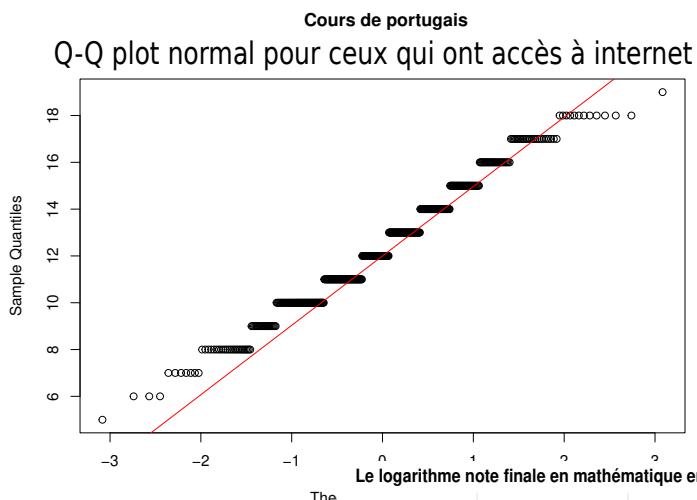
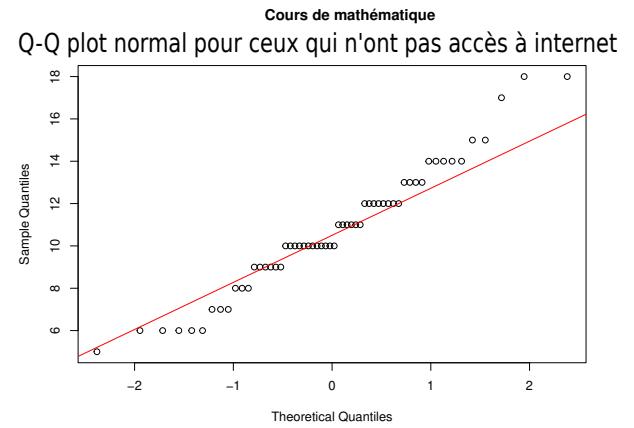
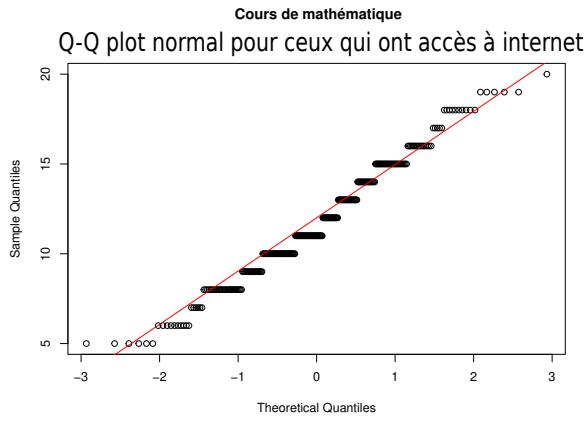


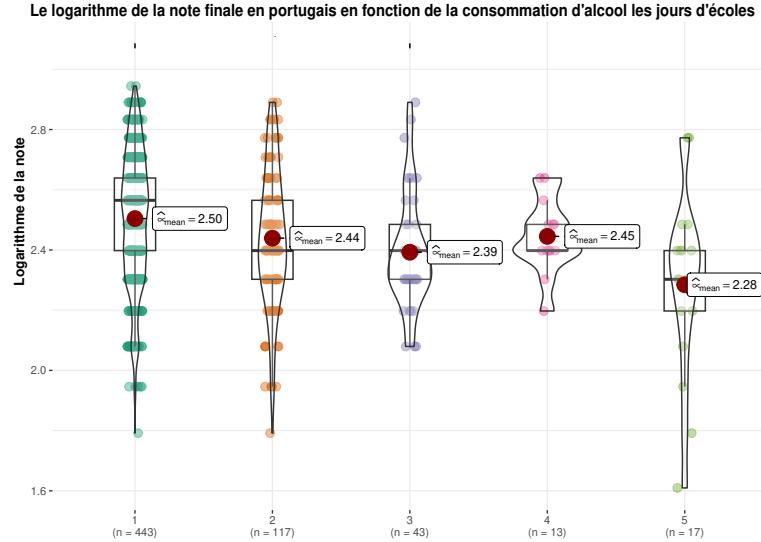
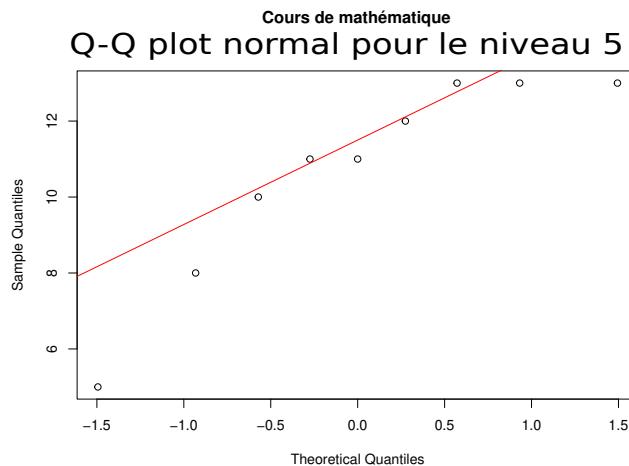
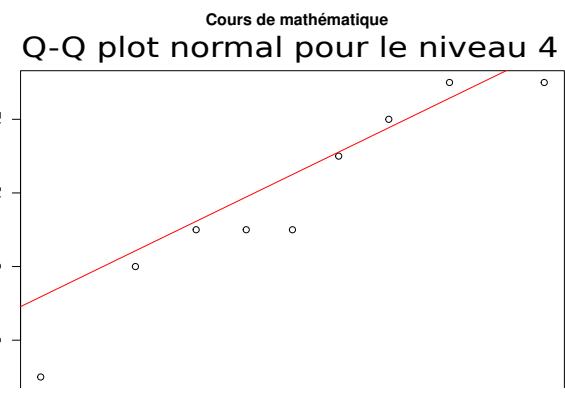
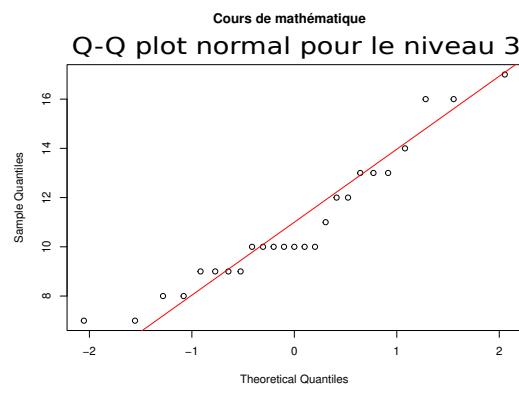
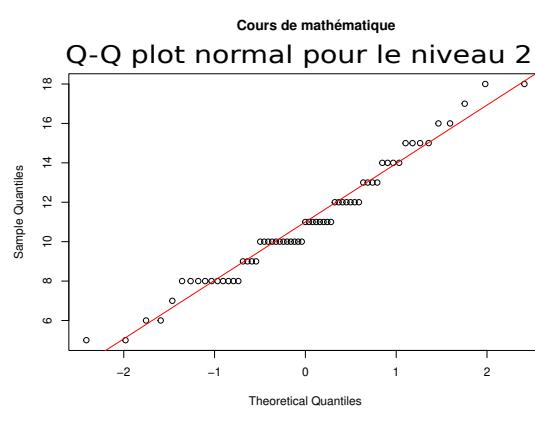
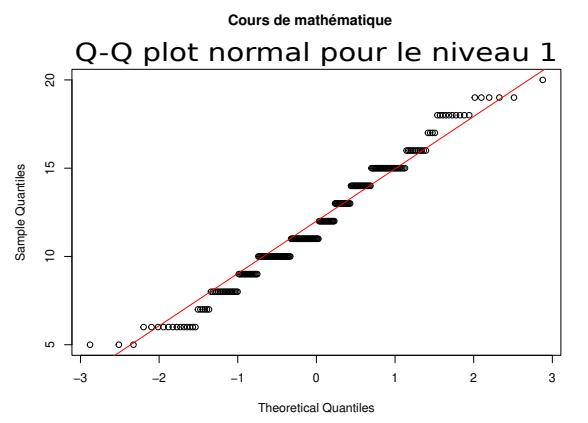


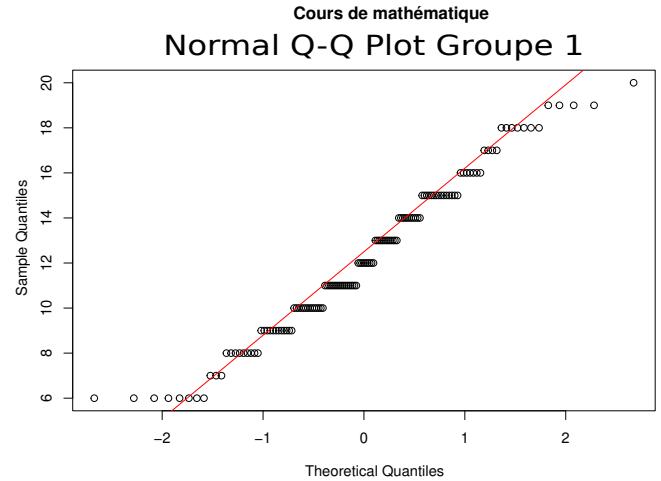
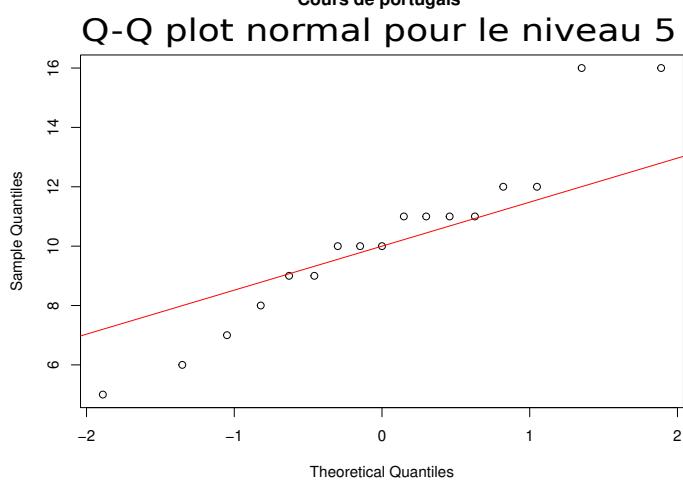
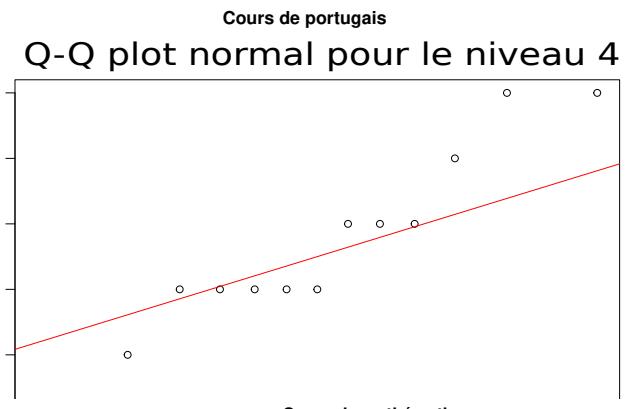
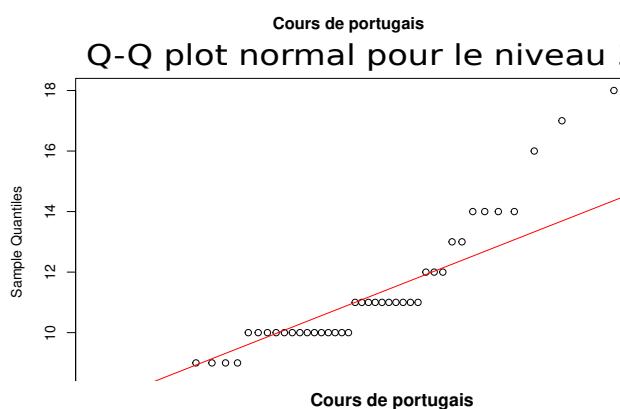
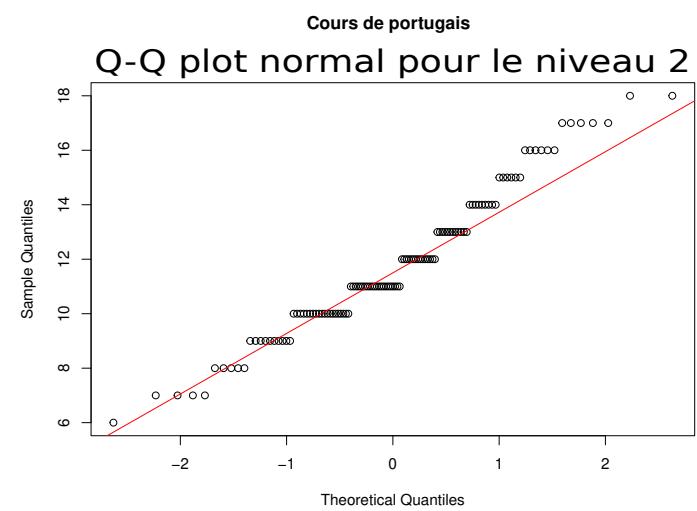
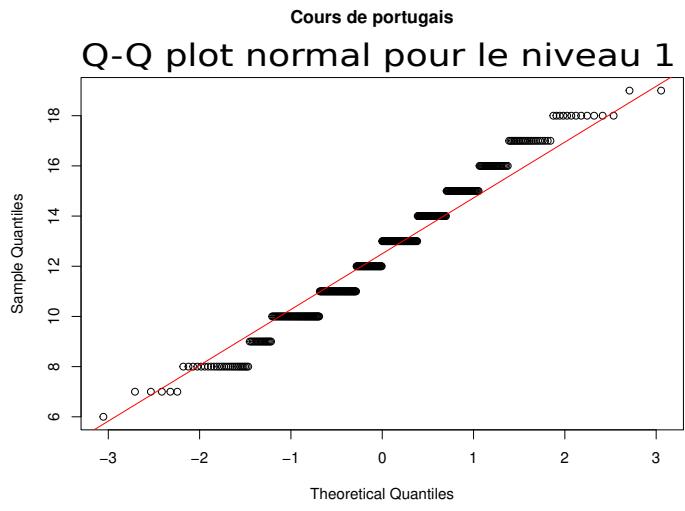


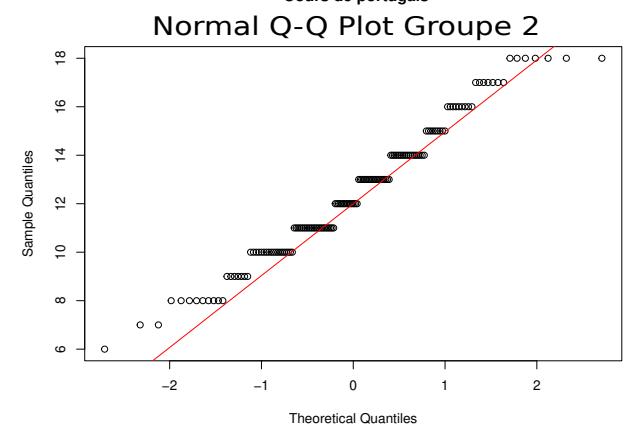
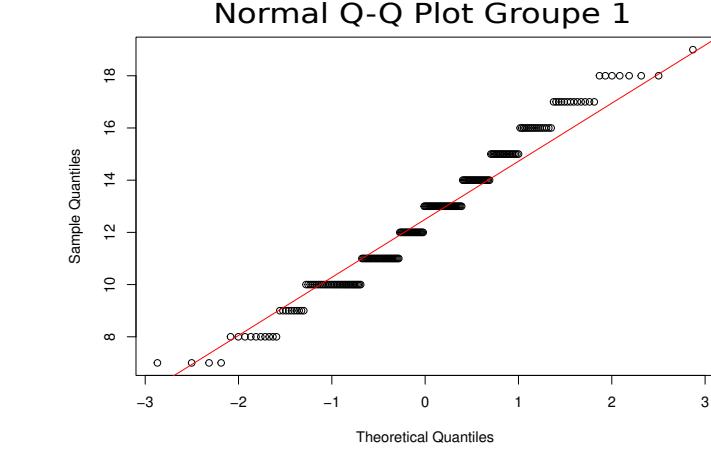
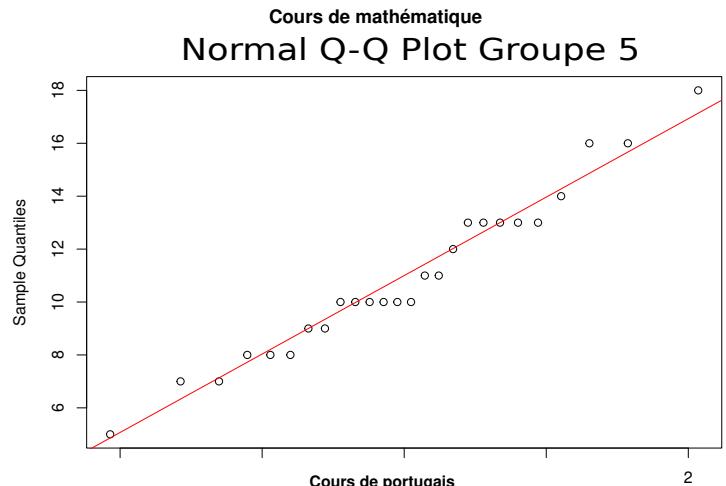
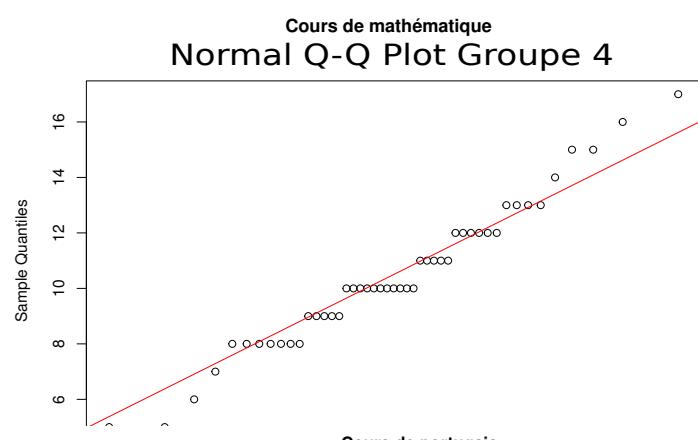
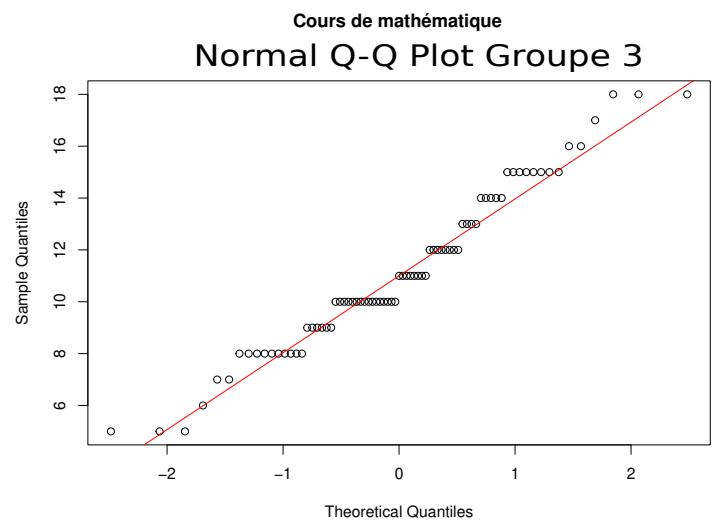
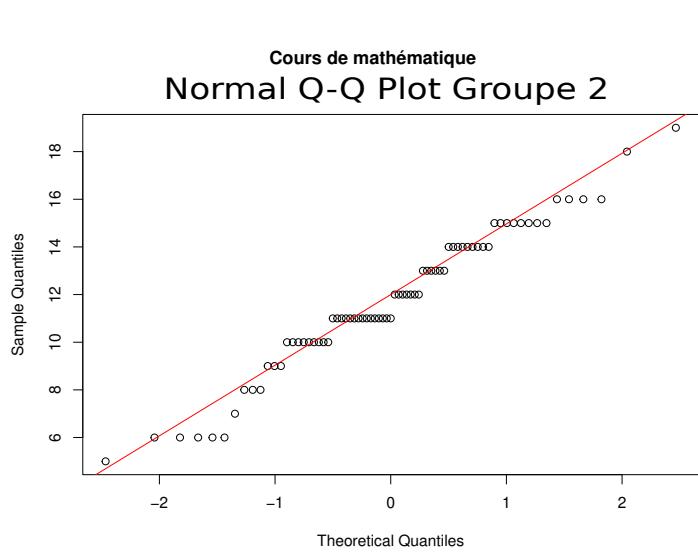


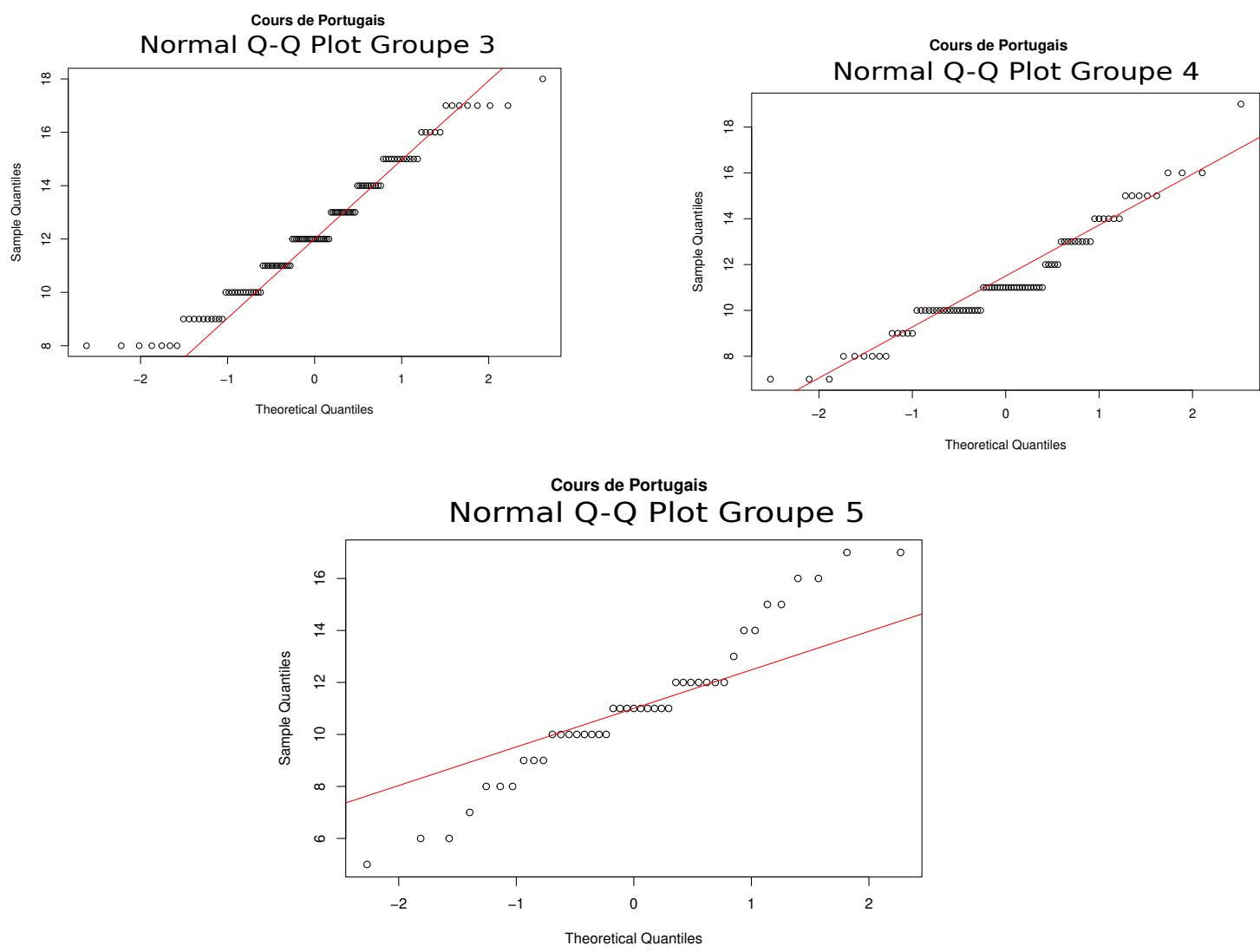












Annexe B

Le code R utilisé pour ce projet :

```

library(psych)
library(ggstatsplot)

#données: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption
mathematique <- read.table("student-mat.csv", header = T, sep = ",")
#View(mathematique)

portugais <- read.table("student-por.csv", header = T, sep = ",")
#View(portugais)
dim(portugais)

#####
#####je trouve qu'il y a vraiment trop de 0 dans les données, c'est ma démarche ici
# Créer un histogramme pour la variable G1
hist(mathematique$G1,
     main = "Histogramme de la variable G1",
     xlab = "Valeurs de G1",
     ylab = "Fréquence",
     col = "skyblue",
     border = "white")
# Créer un histogramme pour la variable G1
hist(mathematique$G2,
     main = "Histogramme de la variable G2",
     xlab = "Valeurs de G2",
     ylab = "Fréquence",
     col = "skyblue",
     border = "white")
# Créer un histogramme pour la variable G3
hist(mathematique$G3,
     main = "Histogramme de la variable G3",
     xlab = "Valeurs de G3",
     ylab = "Fréquence",
     col = "skyblue",
     border = "white")

# Créer un histogramme pour la variable G1
hist(portugais$G1,
     main = "Histogramme de la variable G1",
     xlab = "Valeurs de G1",

```

```

ylab = "Fréquence",
col = "skyblue",
border = "white")

hist(portugais$G2,
  main = "Histogramme de la variable G2",
  xlab = "Valeurs de G2",
  ylab = "Fréquence",
  col = "skyblue",
  border = "white")

hist(portugais$G3,
  main = "Histogramme de la variable G3",
  xlab = "Valeurs de G3",
  ylab = "Fréquence",
  col = "skyblue",
  border = "white")

print(mathematique$G3[mathematique$G3 >= 0 & mathematique$G3 <= 5])
print(mathematique$G2[mathematique$G2 >= 0 & mathematique$G2 <= 5])
print(mathematique$G1[mathematique$G1 >= 0 & mathematique$G1 <= 5])
print(portugais$G1[portugais$G1 >= 0 & portugais$G1 <= 5])
print(portugais$G2[portugais$G2 >= 0 & portugais$G2 <= 5])
print(portugais$G3[portugais$G3 >= 0 & portugais$G3 <= 5])

#Fin de la démarche pour les 0, on va éliminer les 0 des données
#####
#####Élimination des 0 dans les données
#####

which(mathematique$G3 < 6)
mathematique <- mathematique[mathematique$G3 >= 5, ]
which(mathematique$G3 < 5)
which(mathematique$G2 == 0)
which(mathematique$G1 == 0)

which(portugais$G3 < 6)
portugais <- portugais[portugais$G3 >= 5, ]
which(portugais$G3 < 6)
which(portugais$G2 == 0)
which(portugais$G1 == 0)

#Fin de l'élimination des 0 dans les données
#####

```

```
#####
mTravelTime <- mathematique$traveltime
pTravelTime <- portugais$traveltime

mStudyTime <- mathematique$studytime
pStudyTime <- portugais$studytime

mFailures <- mathematique$failures
pFailures <- portugais$failures

mSchoolSup <- mathematique$schoolsup
pSchoolSup <- portugais$schoolsup

mFamSup <- mathematique$famsup
pFamSup <- portugais$famsup

mPaid <- mathematique$paid
pPaid <- portugais$paid

mInternet <- mathematique$internet
pInternet <- portugais$internet

mFreeTime <- mathematique$freetime
pFreeTime <- portugais$freetime

mDalc <- mathematique$Dalc
pDalc <- portugais$Dalc

mWalc <- mathematique$Walc
pWalc <- portugais$Walc

mG3 <- mathematique$G3
pG3 <- portugais$G3

#####
#Début de l'analyse

# analyse du temps pour aller à l'école en premier
```

```

psych::describeBy(mG3, group = mTravelTime)

#la variance semble etre ok sans plus, peut être un peu différente

ggstatsplot::ggbetweenstats(
  data = mathematique, x = traveltimes, y = G3,
  title = "La note finale en mathématique en fonction du temps de voyage",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)

```

#Les variances sont clairement différentes et les n_i sont problématique, on devrait faire un test de randomisation si possible

```

mathematique$logG3 <- log(mG3)

ggstatsplot::ggbetweenstats(
  data = mathematique, x = traveltimes, y = logG3,
  title = "Le logarightme de la note finale en mathématique en fonction du temps de voyage",
  mean.ci = TRUE,
  type="p",
  var.equal = T,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)

```

#Le log n'aide pas, mais vérifions tout de même la normalité

```

qqnorm(mG3[mTravelTime == 1])
qqline(mG3[mTravelTime == 1], col = "red")
qqnorm(mG3[mTravelTime == 2])
qqline(mG3[mTravelTime == 2], col = "red")
qqnorm(mG3[mTravelTime == 3])
qqline(mG3[mTravelTime == 3], col = "red")
qqnorm(mG3[mTravelTime == 4])
qqline(mG3[mTravelTime == 4], col = "red")

```

#Les données semblent suivre une loi normales

```
# Méthode de Bonferroni
```

```
# Création de sous-ensembles pour chaque niveau de studytime
```

```
groupe1 <- mG3[mTravelTime == 1]
```

```
groupe2 <- mG3[mTravelTime == 2]
```

```
groupe3 <- mG3[mTravelTime == 3]
```

```
groupe4 <- mG3[mTravelTime == 4]
```

```
p_values <- c(
```

```
  t.test(groupe1, groupe2, var.equal = F)$p.value,
```

```
  t.test(groupe1, groupe3, var.equal = F)$p.value,
```

```
  t.test(groupe1, groupe4, var.equal = F)$p.value,
```

```
  t.test(groupe2, groupe3, var.equal = F)$p.value,
```

```
  t.test(groupe2, groupe4, var.equal = F)$p.value,
```

```
  t.test(groupe3, groupe4, var.equal = F)$p.value
```

```
)
```

```
# Ajuster les p-valeurs pour les comparaisons multiples (Bonferroni)
```

```
p_values_adjusted <- p.adjust(p_values, method = "bonferroni")
```

```
# Afficher les résultats ajustés
```

```
p_values_adjusted
```

pour le cours de mathématique il n'y a aucune corrélation entre le temps de voyage et la note finale G3.

```
#Maintenant on vérifie pour portugais
```

```
# analyse du temps pour aller à l'école en premier
```

```
psych::describeBy(pG3, group = pTravelTime)
```

```
#la variance semble très semblable, peut être un peu différente
```

```
ggstatsplot::ggbetweenstats(
```

```
  data = portugais, x = traveltimes, y = G3,
```

```
  title = "La note finale en portugais en fonction du temps de voyage",
```

```
  mean.ci = TRUE,
```

```
  type="p",
```

```
  var.equal = F,
```

```
  bf.message = FALSE,
```

```
  results.subtitle = FALSE,
```

```
  pairwise.comparisons = FALSE
```

```
)
```

```
#La variance est assez différente, essayons une transformation racine carré
```

```

portugais$racineG3 <- sqrt(portugais$G3)

ggstatsplot::ggbetweenstats(
  data = portugais, x = traveltimes, y = racineG3,
  title = "La racine carré de la note finale en portugais en fonction du temps de voyage",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)
#ça améliore un peu, essayons avec une transformation log, voir ce qui est le mieux

portugais$logG3 <- log(portugais$G3)

ggstatsplot::ggbetweenstats(
  data = portugais, x = traveltimes, y = racineG3,
  title = "Le logarithme de la note finale en portugais en fonction du temps de voyage",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)
#même effet que la racine carré.

#Vérifions la normalité

qqnorm(pG3[pTravelTime == 1])
qqline(pG3[pTravelTime == 1], col = "red")
qqnorm(pG3[pTravelTime == 2])
qqline(pG3[pTravelTime == 2], col = "red")
qqnorm(pG3[pTravelTime == 3])
qqline(pG3[pTravelTime == 3], col = "red")
qqnorm(pG3[pTravelTime == 4])
qqline(pG3[pTravelTime == 4], col = "red")

#Les données semblent suivre une loi normale

#réutilisons la méthode de bonferroni avec un test de welch

```

```

groupe1 <- pG3[pTravelTime == 1]
groupe2 <- pG3[pTravelTime == 2]
groupe3 <- pG3[pTravelTime == 3]
groupe4 <- pG3[pTravelTime == 4]

p_values <- c(
  t.test(groupe1, groupe2, var.equal = F)$p.value,
  t.test(groupe1, groupe3, var.equal = F)$p.value,
  t.test(groupe1, groupe4, var.equal = F)$p.value,
  t.test(groupe2, groupe3, var.equal = F)$p.value,
  t.test(groupe2, groupe4, var.equal = F)$p.value,
  t.test(groupe3, groupe4, var.equal = F)$p.value
)
# Ajuster les p-valeurs pour les comparaisons multiples (Bonferroni)
p_values_adjusted <- p.adjust(p_values, method = "bonferroni")

# Afficher les résultats ajustés
p_values_adjusted

##### Pour le cours de français, il semble y avoir une corrélation entre le temps de voyage et la note finale G3
#Mais cette corrélation est très faible, à peine inférieur à 0.05
#[1] 0.40263836 0.14093051 0.04318712 1.00000000 0.25417042 1.00000000, c'est entre le groupe 1 et 4.
#Selon le graphique, ça indiquerait que le pire temps de voyages est significatif avec le plus petit temps.

#####
##### Nous sommes rendus à la variable StudyTime

psych::describeBy(mG3, group = mStudyTime)

#Les variances ont l'air pas mal égales, vérifions visuellement

ggstatsplot::ggbetweenstats(
  data = mathematique, x = studytime, y = G3,
  title = "La note finale en mathématique en fonction du temps d'étude",
  mean.ci = TRUE,
  type="p",
  var.equal = T,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)

```

)

#Les variances sont pas mal égale,

```
qqnorm(mG3[mStudyTime == 1])
qqline(mG3[mStudyTime == 1], col = "red")
qqnorm(mG3[mStudyTime == 2])
qqline(mG3[mStudyTime == 2], col = "red")
qqnorm(mG3[mStudyTime == 3])
qqline(mG3[mStudyTime == 3], col = "red")
qqnorm(mG3[mStudyTime == 4])
qqline(mG3[mStudyTime == 4], col = "red")
```

#les données suivent une loi normales

Méthode de Bonferroni

Création de sous-ensembles pour chaque niveau de studytime

```
groupe1 <- mG3[mStudyTime == 1]
groupe2 <- mG3[mStudyTime == 2]
groupe3 <- mG3[mStudyTime == 3]
groupe4 <- mG3[mStudyTime == 4]
```

Effectuer des tests t pour chaque paire de niveaux de studytime

```
p_values <- c(
  t.test(groupe1, groupe2)$p.value,
  t.test(groupe1, groupe3)$p.value,
  t.test(groupe1, groupe4)$p.value,
  t.test(groupe2, groupe3)$p.value,
  t.test(groupe2, groupe4)$p.value,
  t.test(groupe3, groupe4)$p.value)
```

)

Ajuster les p-valeurs pour les comparaisons multiples (Bonferroni)

```
p_values_adjusted <- p.adjust(p_values, method = "bonferroni")
```

Afficher les résultats ajustés

```
p_values_adjusted
```

```
# p_values_adjusted: [1] 1.000000000 0.241111816 0.902398797 0.009943247 0.310055949
1.000000000
```

#il semble que le niveau 3 d'étude est différent du niveau 2

#mais le niveau 1 ne l'est pas du niveau 3.

#en fait le niveau 2 à une moyenne inférieure au niveau 1...

```
#####-----#####

```

#Maintenant portugais

```
psych::describeBy(pG3, group = pStudyTime)
```

#Les variances ont l'air pas mal égale, vérifions visuellement

```
ggstatsplot::ggbetweenstats(
```

```
  data = portugais, x = studytime, y = G3,
  title = "La note finale en portugais en fonction du temps d'étude",
  mean.ci = TRUE,
  type="p",
  var.equal = T,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)
```

#Les variances sont pas mal égale,

```
qqnorm(pG3[pStudyTime == 1])
qqline(pG3[pStudyTime == 1], col = "red")
qqnorm(pG3[pStudyTime == 2])
qqline(pG3[pStudyTime == 2], col = "red")
qqnorm(pG3[pStudyTime == 3])
qqline(pG3[pStudyTime == 3], col = "red")
qqnorm(pG3[pStudyTime == 4])
qqline(pG3[pStudyTime == 4], col = "red")
```

#Les données semblent suivent la loi normale

```
groupe1 <- pG3[pStudyTime == 1]
groupe2 <- pG3[pStudyTime == 2]
groupe3 <- pG3[pStudyTime == 3]
groupe4 <- pG3[pStudyTime == 4]
```

Effectuer des tests t pour chaque paire de niveaux de studytime
`p_values <- c(`

```
  t.test(groupe1, groupe2)$p.value,
  t.test(groupe1, groupe3)$p.value,
  t.test(groupe1, groupe4)$p.value,
  t.test(groupe2, groupe3)$p.value,
```

```

t.test(groupe2, groupe4)$p.value,
t.test(groupe3, groupe4)$p.value
)

# Ajuster les p-valeurs pour les comparaisons multiples (Bonferroni)
p_values_adjusted <- p.adjust(p_values, method = "bonferroni")

# Afficher les résultats ajustés
p_values_adjusted

# p_values_adjusted: [1] 4.641942e-06 7.755940e-09 1.180120e-02 3.998751e-02 1.000000e+00
# 1.000000e+00

#####
# Maintenant la variable SchoolSup

psych::describeBy(mG3, group = mSchoolSup)

#grosse différence dans les variances

ggstatsplot::ggbetweenstats(
  data = mathematique, x = schoolsup, y = G3,
  title = "La note finale en mathématique fonction de l'aide supplémentaire à l'école",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)

#encore une grosse différence dans les variances, essayons une racine carré
mathematique$racineG3 <- sqrt(mathematique$G3)

ggstatsplot::ggbetweenstats(
  data = mathematique, x = schoolsup, y = racineG3,
  title = "La racine carré de la note finale en mathématique en fonction de l'aide supplémentaire à l'école",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)#je n'ai pas ce graphique

```

```
#ça améliore les choses
#Testons la normalité
```

```
qqnorm(mG3[mSchoolSup == "yes"])
qqline(mG3[mSchoolSup == "yes"], col = "red")
qqnorm(mG3[mSchoolSup == "no"])
qqline(mG3[mSchoolSup == "no"], col = "red")
```

#Les données suivent une loi normale

#ça améliore un peu les données, faisons un test t

```
t.test(racineG3 ~ schoolsup, data = mathematique, var.equal = T)
```

#p-value = 2.831e-06, donc oui l'école en plus est très significative, mais ça n'améliore pas la note, au contraire.

#-----

#Maintenant portugais

```
psych::describeBy(pG3, group = pSchoolSup)
```

#grosse différence dans les variances

```
ggstatsplot::ggbetweenstats(
  data = portugais, x = schoolsup, y = G3,
  title = "La note finale en portugais en fonction de l'aide supplémentaire à l'école",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)
```

#assez grosse différence dans les variances, essayons un log

```
portugais$logG3 <- log(portugais$G3)
```

```
ggstatsplot::ggbetweenstats(
  data = portugais, x = schoolsup, y = logG3,
  title = "Le logarithme de la note finale en portugais en fonction de l'aide supplémentaire à l'école",
  mean.ci = TRUE,
```

```

type="p",
var.equal = F,
bf.message = FALSE,
results.subtitle = FALSE,
pairwise.comparisons = FALSE
)

```

#N'améliore pas les variances

```

qqnorm(pG3[pSchoolSup == "yes"])
qqline(pG3[pSchoolSup == "yes"], col = "red")
qqnorm(pG3[pSchoolSup == "no"])
qqline(pG3[pSchoolSup == "no"], col = "red")

```

#ça suit une loi normale, donc faisons un test de welch

```
t.test(G3 ~ schoolsup, data = portugais, var.equal = F)
```

#C'est significatif (p-value = 0.001166), mais l'aide à l'école n'aide pas vraiment les notes.

```
#####
#On est rendu à la variable mFamSup
#####
```

```
psych::describeBy(mG3, group = mFamSup)
```

#variance très proche, regardons visuellement

```

ggstatsplot::ggbetweenstats(
  data = mathematique, x = famsup, y = G3,
  title = "La note finale en mathématique en fonction de l'aide supplémentaire apporté par la famille",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)

```

#variance très semblable aussi, vérifions la normalité

```
qqnorm(mG3[mFamSup == "yes"])
```

```

qqline(mG3[mFamSup == "yes"], col = "red")
qqnorm(mG3[mFamSup == "no"])
qqline(mG3[mFamSup == "no"], col = "red")

#Les données suivent une loi normale, donc faisons un test t

t.test(G3 ~ famsup, data = mathematique, equal.var = T)

#p-value = 0.1469, donc il n'y a pas de différence avec la famille qui aide à la maison pour les math

#-----#
# Portugais

psych::describeBy(pG3, group = pFamSup)

#Variance quasi identique

ggstatsplot::ggbetweenstats(
  data = portugais, x = famsup, y = G3,
  title = "La note finale en portugais en fonction de l'aide supplémentaire apporté par la famille",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)

#Variance quasi identique

qqnorm(pG3[pFamSup == "yes"])
qqline(pG3[pFamSup == "yes"], col = "red")
qqnorm(pG3[pFamSup == "no"])
qqline(pG3[pFamSup == "no"], col = "red")

#suivent loi normale

t.test(G3 ~ famsup, data = portugais, equal.var = T)

#p-value = 0.9683, pour le portugais également, ça n'aide pas l'aide à la maison, ça ne change rien.

#####
#Variable mPaid

psych::describeBy(mG3, group = mPaid)

```

#Variance très proche

```
ggstatsplot::ggbetweenstats(
  data = mathematique, x = paid, y = G3,
  title = "La note finale en mathématique fonction des cours supplémentaires privés",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)
```

#Variance toujours très proche

```
qqnorm(mG3[mPaid == "yes"])
qqline(mG3[mPaid == "yes"], col = "red")
qqnorm(mG3[mPaid == "no"])
qqline(mG3[mPaid == "no"], col = "red")
```

Les données suivent une loi normale

```
t.test(G3 ~ paid, data = mathematique, var.equal = T)
```

#p-value = 0.6744, aucune différence pour mpaid

```
psych::describeBy(pG3, group = pPaid)
```

#Variance un peu éloigné

```
ggstatsplot::ggbetweenstats(
  data = portugais, x = paid, y = G3,
  title = "La note finale en portugais fonction des cours supplémentaires privés",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)
```

#Les variances sont assé éloigné

#Essayons une racine carré

```
ggstatsplot::ggbetweenstats(
  data = portugais, x = paid, y = racineG3,
  title = "La racine carré de la note finale en portugais fonction des cours supplémentaires privés",
```

```

mean.ci = TRUE,
type="p",
var.equal = F,
bf.message = FALSE,
results.subtitle = FALSE,
pairwise.comparisons = FALSE
)

#essayons un log
ggstatsplot::ggbetweenstats(
  data = portugais, x = paid, y = logG3,
  title = "Le logarithme de la note finale en portugais fonction des cours supplémentaires privés",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)

```

#aucune transformation ne fait vraiment mieux que les données sans transformations, vérifions la normalité

```

qqnorm(pG3[pPaid == "yes"])
qqline(pG3[pPaid == "yes"], col = "red")
qqnorm(pG3[pPaid == "no"])
qqline(pG3[pPaid == "no"], col = "red")

```

#Les données suivent une loi normale

```

t.test(G3 ~ paid, data = portugais, var.equal = F)
# p-value = 0.05082, on pourrait rejeter le test, mais on est presque pile sur le alpha = 5%, j'irais pour un rejet
#car les médiane sont égales et que la médiane résiste aux valeurs extrêmes, il y a aussi le fait que cela n'améliore pas
#la moyenne de la note, au pire, si c'était significatif, paid diminue la note moyenne des étudiants.

```

Variable Internet

```
psych::describeBy(mG3, group = mInternet)
```

#Variance proche

```
ggstatsplot::ggbetweenstats(
  data = mathematique, x = internet, y = G3,
  title = "La note finale mathématique en fonction de l'accès à l'internet à domicile",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)
```

#Variance proche

```
ggstatsplot::ggbetweenstats(
  data = mathematique, x = internet, y = logG3,
  title = "Le logarithme de la note finale mathématique en fonction de l'accès à l'internet à domicile",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)
```

#variance encore plus proche

```
qqnorm(mG3[mInternet == "yes"])
qqline(mG3[mInternet == "yes"], col = "red")
qqnorm(mG3[mInternet == "no"])
qqline(mG3[mInternet == "no"], col = "red")
```

Les données suivent une loi normale

```
t.test(logG3 ~ internet, data = mathematique, var.equal = T)
```

#p-value = 0.0353, l'internet semble montré une différence, elle semble aider!

```
psych::describeBy(pG3, group = pInternet)
```

#Variance peu éloigné

```
ggstatsplot::ggbetweenstats(
  data = portugais, x = internet, y = G3,
  title = "La note finale en portugais en fonction de l'accès à l'internet à domicile",
  mean.ci = TRUE,
  type="p",
```

```

var.equal = F,
bf.message = FALSE,
results.subtitle = FALSE,
pairwise.comparisons = FALSE
)

```

#Les variances sont proche

#vérifions la normalité

```

qqnorm(pG3[pInternet == "yes"])
qqline(pG3[pInternet == "yes"], col = "red")
qqnorm(pG3[pInternet == "no"])
qqline(pG3[pInternet == "no"], col = "red")

```

#Les données suivent une loi normale

```
t.test(G3 ~ internet, data = portugais, var.equal = T)
```

#p-value = 0.0009197, très significatif, l'internet semblent améliorer les notes de portugais

```
#####
# Variable mDalc
```

```
psych::describeBy(mG3, group = mDalc)
```

#Les variances semblent différent

```

ggstatsplot::ggbetweenstats(
  data = mathematique, x = Dalc, y = G3,
  title = "La note finale en mathématique en fonction de la consommation d'alcool les jours d'écoles",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)

```

```

ggstatsplot::ggbetweenstats(
  data = mathematique, x = Dalc, y = logG3,
  title = "Le logarithme note finale en mathématique en fonction de la consommation d'alcool les jours d'écoles",
  mean.ci = TRUE,

```

```

type="p",
var.equal = F,
bf.message = FALSE,
results.subtitle = FALSE,
pairwise.comparisons = FALSE
)

```

```

#la transformation log améliore, mais pas beaucoup
#on vérifie la normalité

```

```

qqnorm(mG3[mDalc == 1])
qqline(mG3[mDalc == 1], col = "red")
qqnorm(mG3[mDalc == 2])
qqline(mG3[mDalc == 2], col = "red")
qqnorm(mG3[mDalc == 3])
qqline(mG3[mDalc == 3], col = "red")
qqnorm(mG3[mDalc == 4])
qqline(mG3[mDalc == 4], col = "red")
qqnorm(mG3[mDalc == 5])
qqline(mG3[mDalc == 5], col = "red")

```

```
#les données suivent une loi normale, difficile d'exclure pour les niveaux 4 et 5 (peu d'observations)
```

```

groupe1 <- mG3[mDalc == 1]
groupe2 <- mG3[mDalc == 2]
groupe3 <- mG3[mDalc == 3]
groupe4 <- mG3[mDalc == 4]
groupe5 <- mG3[mDalc == 5]

```

```

p_values <- c(
  t.test(groupe1, groupe2, var.equal = F)$p.value,
  t.test(groupe1, groupe3, var.equal = F)$p.value,
  t.test(groupe1, groupe4, var.equal = F)$p.value,
  t.test(groupe1, groupe5, var.equal = F)$p.value,
  t.test(groupe2, groupe3, var.equal = F)$p.value,
  t.test(groupe2, groupe4, var.equal = F)$p.value,
  t.test(groupe2, groupe5, var.equal = F)$p.value,
  t.test(groupe3, groupe4, var.equal = F)$p.value,
  t.test(groupe3, groupe5, var.equal = F)$p.value
)

```

```
# Ajuster les p-valeurs pour les comparaisons multiples (Bonferroni)
p_values_adjusted <- p.adjust(p_values, method = "bonferroni")
```

```
# Afficher les résultats ajustés
```

```
p_values_adjusted
```

```
#0.3705751 1.0000000 0.5144004 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  
#Il y a aucune différence significative, boire la semaine ne semblent pas changer les notes de  
mathématique
```

```
#-----  
#portugais
```

```
psych::describeBy(pG3, group = PDalc)
```

```
#Les variances sont assez différentes
```

```
ggstatsplot::ggbetweenstats(  
  data = portugais, x = Dalc, y = G3,  
  title = "La note finale en portugais en fonction de la consommation d'alcool les jours d'écoles",  
  mean.ci = TRUE,  
  type="p",  
  var.equal = F,  
  bf.message = FALSE,  
  results.subtitle = FALSE,  
  pairwise.comparisons = FALSE  
)
```

```
ggstatsplot::ggbetweenstats(  
  data = portugais, x = Dalc, y = logG3,  
  title = "Le logarithme de la note finale en portugais en fonction de la consommation d'alcool les jours  
d'écoles",  
  mean.ci = TRUE,  
  type="p",  
  var.equal = F,  
  bf.message = FALSE,  
  results.subtitle = FALSE,  
  pairwise.comparisons = FALSE  
)
```

```
#la transformation log améliore, mais pas beaucoup  
#on vérifie la normalité
```

```
qqnorm(pG3[PDalc == 1])  
qqline(pG3[PDalc == 1], col = "red")  
qqnorm(pG3[PDalc == 2])  
qqline(pG3[PDalc == 2], col = "red")  
qqnorm(pG3[PDalc == 3])  
qqline(pG3[PDalc == 3], col = "red")
```

```

qqnorm(pG3[PDalc == 4])
qqline(pG3[PDalc == 4], col = "red")
qqnorm(pG3[PDalc == 5])
qqline(pG3[PDalc == 5], col = "red")

#les données suivent une loi normale, difficile d'exclure pour les niveaux 4 et 5 (peu d'observations)

groupe1 <- pG3[PDalc == 1]
groupe2 <- pG3[PDalc == 2]
groupe3 <- pG3[PDalc == 3]
groupe4 <- pG3[PDalc == 4]
groupe5 <- pG3[PDalc == 5]

p_values <- c(
  t.test(groupe1, groupe2, var.equal = F)$p.value,
  t.test(groupe1, groupe3, var.equal = F)$p.value,
  t.test(groupe1, groupe4, var.equal = F)$p.value,
  t.test(groupe1, groupe5, var.equal = F)$p.value,
  t.test(groupe2, groupe3, var.equal = F)$p.value,
  t.test(groupe2, groupe4, var.equal = F)$p.value,
  t.test(groupe2, groupe5, var.equal = F)$p.value,
  t.test(groupe3, groupe4, var.equal = F)$p.value,
  t.test(groupe3, groupe5, var.equal = F)$p.value
)
)

# Ajuster les p-valeurs pour les comparaisons multiples (Bonferroni)
p_values_adjusted <- p.adjust(p_values, method = "bonferroni")

# Afficher les résultats ajustés
p_values_adjusted

#[1] 0.048103725 0.003579952 0.435058880 0.052687767 1.000000000 1.000000000 0.524709947
1.000000000 1.000000000

#Difficile à interpréter, il semble avoir une différence significative entre 1 et 2 et entre 1 et 3, mais
presque pas
#significative entre 1 et 5, alors que la moyenne et la médiane sont toutes 2 bien inférieur au groupe 1

#####
# variable walc

psych::describeBy(mG3, group = mWalc)

#Les variances semblent différent

ggstatsplot::ggbetweenstats(

```

```

data = mathematique, x = Walc, y = G3,
title = "La note finale en mathématique en fonction de la consommation d'alcool la fin de semaine",
mean.ci = TRUE,
type="p",
var.equal = F,
bf.message = FALSE,
results.subtitle = FALSE,
pairwise.comparisons = FALSE
)

```

#Les variance semblent finalement très proche
#on vérifie la normalité

```

qqnorm(mG3[mWalc == 1])
qqline(mG3[mWalc == 1], col = "red")
qqnorm(mG3[mWalc == 2])
qqline(mG3[mWalc == 2], col = "red")
qqnorm(mG3[mWalc == 3])
qqline(mG3[mWalc == 3], col = "red")
qqnorm(mG3[mWalc == 4])
qqline(mG3[mWalc == 4], col = "red")
qqnorm(mG3[mWalc == 5])
qqline(mG3[mWalc == 5], col = "red")

```

#les données suivent une loi normale,

```

groupe1 <- mG3[mWalc == 1]
groupe2 <- mG3[mWalc == 2]
groupe3 <- mG3[mWalc == 3]
groupe4 <- mG3[mWalc == 4]
groupe5 <- mG3[mWalc == 5]

```

```

p_values <- c(
  t.test(groupe1, groupe2, var.equal = T)$p.value,
  t.test(groupe1, groupe3, var.equal = T)$p.value,
  t.test(groupe1, groupe4, var.equal = T)$p.value,
  t.test(groupe1, groupe5, var.equal = T)$p.value,
  t.test(groupe2, groupe3, var.equal = T)$p.value,
  t.test(groupe2, groupe4, var.equal = T)$p.value,
  t.test(groupe2, groupe5, var.equal = T)$p.value,
  t.test(groupe3, groupe4, var.equal = T)$p.value,
  t.test(groupe3, groupe5, var.equal = T)$p.value
)

```

Ajuster les p-valeurs pour les comparaisons multiples (Bonferroni)
p_values_adjusted <- p.adjust(p_values, method = "bonferroni")

```

# Afficher les résultats ajustés
p_values_adjusted

#[1] 1.00000000 0.27039427 0.01628743 0.77862114 1.00000000 0.13375697 1.00000000
1.00000000 1.00000000
#Il y a une différence entre les groupe, certain entre 1 et 3

#-----
#portugais

psych::describeBy(pG3, group = pWalc)

#Les variances semblent proche

ggstatsplot::ggbetweenstats(
  data = portugais, x = Walc, y = G3,
  title = "La note finale en portugais en fonction de la consommation d'alcool la fin de semaine",
  mean.ci = TRUE,
  type="p",
  var.equal = F,
  bf.message = FALSE,
  results.subtitle = FALSE,
  pairwise.comparisons = FALSE
)

qqnorm(pG3[pWalc == 1])
qqline(pG3[pWalc == 1], col = "red")
qqnorm(pG3[pWalc == 2])
qqline(pG3[pWalc == 2], col = "red")
qqnorm(pG3[pWalc == 3])
qqline(pG3[pWalc == 3], col = "red")
qqnorm(pG3[pWalc == 4])
qqline(pG3[pWalc == 4], col = "red")
qqnorm(pG3[pWalc == 5])
qqline(pG3[pWalc == 5], col = "red")

#les données suivent une loi normale

groupe1 <- pG3[pWalc == 1]
groupe2 <- pG3[pWalc == 2]
groupe3 <- pG3[pWalc == 3]
groupe4 <- pG3[pWalc == 4]
groupe5 <- pG3[pWalc == 5]

```

```
p_values <- c(  
  t.test(groupe1, groupe2, var.equal = T)$p.value,  
  t.test(groupe1, groupe3, var.equal = T)$p.value,  
  t.test(groupe1, groupe4, var.equal = T)$p.value,  
  t.test(groupe1, groupe5, var.equal = T)$p.value,  
  t.test(groupe2, groupe3, var.equal = T)$p.value,  
  t.test(groupe2, groupe4, var.equal = T)$p.value,  
  t.test(groupe2, groupe5, var.equal = T)$p.value,  
  t.test(groupe3, groupe4, var.equal = T)$p.value,  
  t.test(groupe3, groupe5, var.equal = T)$p.value  
)  
  
# Ajuster les p-valeurs pour les comparaisons multiples (Bonferroni)  
p_values_adjusted <- p.adjust(p_values, method = "bonferroni")  
  
# Afficher les résultats ajustés  
p_values_adjusted  
  
#1.0000000000 1.0000000000 0.0007616568 0.0058612511 1.0000000000 0.0127945398  
0.0401602842 0.0421373683 0.0802960388  
#au moins un des groupe est différent, le groupe 1 et (1,4) et (1,5). Boire de l'alcool les fds c'est pas bon  
pour le portugais
```