Sur quelques tests usuels en statistique bivariée

Sinda Ammous, Olivier Bouaziz, Jerome Dedecker, Jonathan El Methni, Mohamed Mellouk, Florence Muri

Sur quelques tests usuels en statistique bivariée

Sinda Ammous[2], Olivier Bouaziz[1,2], Jérôme Dedecker[1,2], Jonathan El Methni[1,2], Mohamed Mellouk[1,2] et Florence Muri[1]

[1]IUT Paris Descartes, Département STID
[2]Université Paris Descartes, Laboratoire MAP5, CNRS UMR 8145

Résumé

Nous considérons plusieurs tests usuels en statistique bivariée. Nous soulignons les limites de ces tests, dont certaines sont bien connues (problèmes de robustesse ou de calibration), et nous proposons des alternatives simples qui peuvent être facilement présentées aux étudiants.

1 Introduction

Dans ces notes, on considère plusieurs tests usuels en statistique bivariée, qui sont enseignés dans de nombreuses formations scientifiques à divers niveaux (DUT STID, PACES, L2/L3/M1 Maths, L2/L3/M1 Bio, M1 Santé Publique, ...).

Notre but est de montrer, à l'aide de considérations mathématiques et de simulations, que beaucoup de ces tests sont soit très peu robustes (c'est à dire qu'ils ne fonctionnent plus hors du cadre très strict dans lequel ils ont été définis), soit mals calibrés (c'est-à-dire que l'on peut trouver des exemples simples pour lesquels H_0 est vraie, mais le risque de première espèce n'est pas celui annoncé).

Bien sûr, nous ne sommes pas les seuls ni les premiers à avoir remarqué cela, et dans certains cas, des solutions valables ont été proposées (voir par exemple le célèbre article de Welch [14] à propos du test d'égalité de deux espérances dans le cas où les variances sont inégales, le test de Welch étant par ailleurs (asymptotiquement) robuste à la non normalité, cf. Section 5).

Il nous a néanmoins semblé important de revenir sur ces questions pour au moins deux raisons :

- D'abord ces tests ne sont pas seulement des objets mathématiques sur lesquels les étudiants peuvent "se faire les dents", ils sont aussi très souvent utilisés en pratique (par exemple dans des articles de recherche biomédicale). Il est donc important de revenir sur les limites de ces tests, qui ne sont pas toujours bien indiquées (notamment dans le cas des tests dits "non-paramétriques").

- Ensuite, parce qu'il est souvent très facile de modifier ces tests afin de les rendre plus robustes ou (asymptotiquement) bien calibrés. La modification que nous présentons est à chaque fois basée sur le calcul de la variance limite de la statistique de test sous H_0. En renormalisant par un estimateur de l'écart-type, on obtient une version robuste ou asymptotiquement bien calibrée grâce au théorème limite central. Bien sûr, la justification est asymptotique, mais nous verrons que, sur les exemples simulés, on a toujours intérêt à corriger ces tests, même pour des tailles d'échantillons relativement faibles.

Il nous semble que ces tests modifiés, faciles à décrire et à implémenter, devraient être systématiquement signalés aux étudiants. Pour être complets, nous présentons en annexe une fonction R très simple pour chaque modification proposée.

La liste des tests que nous examinons dans ces notes est la suivante : en Section 2, le test de corrélation de Pearson, en Section 3 le test de corrélation de Kendall (avec une remarque sur le test de Spearman), en Section 4 le test de Mann-Whitney, en Section 5 les tests d'égalité des espérances, en Section 6 les tests d'égalité des variances, en Section 7 le test du signe et le test signe et rang.

2 Le test de corrélation de Pearson

Soient $(X_1, Y_1), \ldots, (X_n, Y_n)$ des couples de variables aléatoires quantitatives indépendants et identiquement distribués (i.i.d.). On suppose que X_i et Y_i possèdent un moment d'ordre 2, et on note ρ le coefficient de corrélation entre X_i et Y_i. Pour tester

$$H_0 : \rho = 0 \quad \text{contre} \quad H_1 : \rho \neq 0$$

on utilise souvent la statistique de test de Pearson (introduite par Fisher [6])

$$T_n = \frac{\hat{\rho}_n}{\sqrt{\frac{1-\hat{\rho}_n^2}{n-2}}}$$

où $\hat{\rho}_n$ est le coefficient de corrélation empirique

$$\hat{\rho}_n = \frac{\sum_{k=1}^n (X_k - \bar{X}_n)(Y_k - \bar{Y}_n)}{\sqrt{\sum_{k=1}^n (Y_k - \bar{Y}_n)^2}\,\sqrt{\sum_{k=1}^n (X_k - \bar{X}_n)^2}}\,.$$

L'intérêt de cette statistique étant que, dans le cas où le couple (X_i, Y_i) est Gaussien, sa distribution exacte sous H_0 est connue : c'est la loi de Student à $n-2$ degrés de liberté. Si le couple (X_i, Y_i) n'est pas Gaussien, on peut aussi facilement montrer que, si X_i est indépendante de Y_i, alors T_n converge en loi vers la loi normale centrée réduite.

Par conséquent, si X_i et Y_i sont indépendantes, la zone de rejet $R_{n,\alpha} = \{|T_n| > c_\alpha\}$ où c_α est le quantile d'ordre $1-(\alpha/2)$ de la loi $\mathcal{N}(0,1)$, a bien une probabilité qui tend vers α lorsque n tend vers l'infini. Mais ce n'est pas le cas en général sous H_0, qui bien entendu n'implique pas l'indépendance des variables.

On peut donc poser le constat suivant : **dans un contexte général, le test de Pearson n'est pas bien calibré pour tester** $\rho = 0$. Voir aussi [5] pour un constat similaire.

Hors du cadre strict du modèle linéaire (Gaussien ou non), il est donc préférable d'utiliser la statistique naturelle

$$T'_n = \frac{\sum_{k=1}^n (X_k - \bar{X}_n)(Y_k - \bar{Y}_n)}{\sqrt{\sum_{k=1}^n (Z_k - \bar{Z}_n)^2}}\,.$$

où $Z_i = (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$. Sous la simple hypothèse $\mathbb{E}(X_i^2 Y_i^2) < \infty$, une application directe du théorème limite central et du lemme de Slutsky fournit la convergence en loi sous H_0 de T'_n vers la loi normale centrée réduite. La zone de rejet de H_0 est donc de la forme $R'_{n,\alpha} = \{|T'_n| > c_\alpha\}$ où c_α est le quantile d'ordre $1-(\alpha/2)$ de la loi $\mathcal{N}(0,1)$, ce qui fournit un test asymptotiquement bien calibré. On peut aussi, comme pour la statistique de Pearson, utiliser le quantile d'ordre $1-(\alpha/2)$ de la loi Student$(n-2)$ (ce qu'on fera dans les simulations ci-dessous) car, lorsque n tend vers l'infini, la loi Student$(n-2)$ converge vers la loi $\mathcal{N}(0,1)$. Nous ferons d'autres commentaires sur le choix des quantiles de la loi Student$(n-2)$ à la fin de cette section.

On notera à présent CorTest le test basé sur la statistique T'_n.

Pour illustrer cela, on simule, pour différentes valeurs de n, des couples $(X_i, Y_i)_{1\leq i\leq n}$ i.i.d., selon le modèle

$$Y_i = X_i^2 + 0.3\varepsilon_i \tag{2.1}$$

où les $(X_i)_{1\leq i\leq n}$ et les $(\epsilon_i)_{1\leq i\leq n}$ sont deux suites indépendantes de variables i.i.d. de loi $\mathcal{N}(0,1)$ (voir Figure 1). On peut facilement voir, que, pour ce modèle, $\rho = 0$.

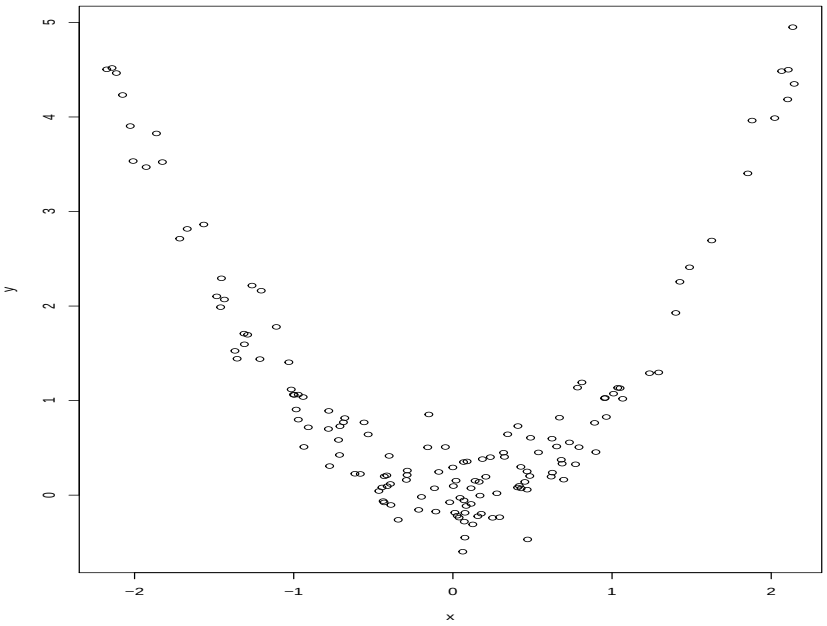

FIGURE 1 – Nuage de point de 150 couples (x_i, y_i) tirés selon le modèle (2.1)

On simule $N = 2000$ échantillons de taille n selon le modèle décrit ci-dessus, et pour chaque test on indique la fréquence de rejet de H_0 au niveau 5%. On considére ici 4 tests : le test CorTest basé sur la statistique T'_n, le test de Pearson, le test de Kendall (qui teste si les variables ont tendance à varier dans le même sens, voir la section suivante pour plus de détails), et enfin le test de Fisher exact (voir [7]), qui teste l'indépendance entre X_i et Y_i (on utilise trois classes pour les variables X_i : $]-\infty, 0.43],]-0.43, 0.43[, [0.43, +\infty[$ et trois classes pour les Y_i : $[0, 0.25],]0.25, 0.95[, [0.95, +\infty[$).

Les résultats sont présentés dans la Table 1 (niveau $\alpha = 5\%$, simulations réalisées à l'aide de R).

Sur ce jeu de simulations (Table 1), on voit que, pour CorTest, la fréquence de rejet de H_0 est toujours inférieure à 6.5%, comprise entre 4.5% et 6% dès que $n \geq 30$. Comme prévu, il semble bien y avoir convergence vers le niveau 5%.

Le test de Pearson est très mal calibré, puisque sa fréquence de rejet oscille entre 35% et 40%. Cela signifie que, dans plus de 35% des cas, le test de Pearson indique une corrélation significative (au niveau de risque 5%) alors que $\rho = 0$. En d'autres termes, dans le cas du modèle (2.1), ce test produit plus de 35% de faux positifs, au lieu des 5% prévus (risque de première espèce).

Le test de Kendall semble mal fonctionner aussi, puisque sa fréquence de rejet oscille entre 17% et 22%. On verra pourquoi, et comment le corriger, dans la section suivante.

Comme prévu, le test de Fisher exact détecte de mieux en mieux la non-indépendance des variables (dans au moins 90% des cas lorsque $n \geq 30$, systématiquement dès que $n \geq 70$). On constaterait la

n	20	30	40	50	60	70	80	90	100	200
CorTest	0.062	0.056	0.053	0.049	0.055	0.049	0.053	0.051	0.052	0.049
Pearson	0.372	0.391	0.374	0.369	0.367	0.352	0.363	0.382	0.369	0.364
Kendall	0.176	0.179	0.18	0.183	0.202	0.212	0.194	0.186	0.212	0.19
Fisher	0.652	0.903	0.976	0.998	0.999	1	1	1	1	1

TABLE 1 – Fréquence de rejet des différents tests au niveau 5% pour n couples tirés selon le modèle (2.1)

n	5	10	15	20	25	30	35	40	45	50
$\lvert T'_n\rvert$	2.838	2.309	2.186	2.129	2.095	2.072	2.055	2.043	2.032	2.026
Stud(n-2)	3.182	2.306	2.160	2.101	2.069	2.048	2.035	2.024	2.017	2.011
n	60	70	80	90	100	110	120	130	140	150
$\lvert T'_n\rvert$	2.024	2.010	2.003	1.997	1.993	1.991	1.99	1.985	1.985	1.983
Stud(n-2)	2.006	2	1.99	1.987	1.984	1.982	1.98	1.979	1.977	1.976

TABLE 2 – Quantiles d'ordre 95% de la loi de $|T'_n|$ sous H_0 (cas Gaussien) et quantiles d'ordre 97.5% de la loi Student$(n-2)$

même chose pour le test d'indépendance du χ^2 (pourvu que n soit assez grand, de sorte que les conditions d'application du test soient remplies).

Avant de conclure cette section, revenons sur le cas où le coupe (X_i, Y_i) est Gaussien. Dans ce cas, il est très facile de voir que la loi de la statistique T'_n est libre sous H_0, c'est à dire ne dépend pas des espérances et variances de X_i et Y_i. Dans la Table 2, nous donnons une estimation du quantile d'ordre 95% de la loi de $|T'_n|$ sous H_0 par Monte-Carlo ($N = 20 \times 10^6$ simulations) pour différentes valeurs de n, que nous comparons au quantile d'ordre 97.5 % de la loi Student$(n-2)$.

On peut voir grâce à la Table 2 que, dès que $n \geq 10$, les deux quantiles sont très proches. Pour tester $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$ au niveau α, on peut donc utiliser la zone de rejet $R'_{n,\alpha} = \{|T'_n| > c_\alpha\}$ où c_α est le quantile d'ordre $1-(\alpha/2)$ de la loi Student$(n-2)$.

En conclusion, pour tester la corrélation, **nous préconisons de présenter aux étudiants de niveau L2/L3 le test CorTest, plutôt que le test de Pearson.** Il est plus naturel, plus facile à comprendre, et fonctionne mieux en pratique. De plus, il peut aussi être utilisé dans le cas des petits échantillons Gaussiens (à condition d'estimer les quantiles de $|T'_n|$ sous H_0, par exemple par Monte-Carlo (cf. Appendice)), même si, dans ce cas précis, il est moins puissant que le test de Pearson.

3 Le test de corrélation de Kendall

Le contexte est le même qu'en Section 2 : $(X_1, Y_1), \ldots, (X_n, Y_n)$ sont des couples de variables aléatoires quantitatives indépendants et identiquement distribués. On suppose de plus que les variables sont continues.

On cherche à savoir si X_i et Y_i ont tendance à varier dans le même sens ou dans le sens contraire. Notons

$$\pi = \mathbb{P}((X_2 - X_1)(Y_2 - Y_1) > 0) - 0.5$$

On dira que X_i et Y_i sont corrélées au sens de Kendall lorsque $\pi \neq 0$ (corrélation positive si $\pi > 0$ et corrélation négative si $\pi < 0$). Pour tester

$$H_0 : \pi = 0 \quad \text{contre} \quad H_1 : \pi \neq 0$$

Kendall [11] a donc proposé de compter le nombre de paires concordantes (i.e. pour lesquelles le produit $(X_i - X_j)(Y_i - Y_j)$ est strictement positif), ce qui conduit à la statistique

$$T_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left(\mathbf{1}_{(X_i - X_j)(Y_i - Y_j) > 0} - 0.5 \right) .$$

Il est assez facile de voir, que si X_i et Y_i sont indépendantes, alors la loi de T_n est libre (i.e. ne dépend pas de la loi des X_i ni de la loi des Y_i), en se ramenant à deux suites indépendantes $U_1, \ldots, U_n$ et $V_1, \ldots, V_n$ de variables i.i.d. de loi $\mathcal{U}([0,1])$. Par conséquent, si X_i et Y_i sont indépendantes, T_n suit une loi connue et tabulée. Le test de Kendall est construit à partir des quantiles de cette loi.

Mais si X_i et Y_i ne sont pas indépendantes, la loi de la statistique T_n n'a pas de raison d'être libre sous H_0 (elle dépend de la loi jointe de (X_i, Y_i)).

On peut donc poser le constat suivant : **dans un contexte général, le test de Kendall n'est pas bien calibré pour tester** $\pi = 0$.

On peut néanmoins régler (asymptotiquement) ce problème, en considérant la distribution limite de $\sqrt{n}T_n$ sous H_0. En partant de la décomposition de Hoeffding de la U-statistique T_n (voir [8] ou [13], exemple 12.5), on voit que, sous H_0, $\sqrt{n}T_n$ converge en loi vers la loi $\mathcal{N}(0, V)$, avec

$$V = 4\mathrm{Var}(F(X_i, Y_i) + H(X_i, Y_i)) ,$$

où $F(x,y) = \mathbb{P}(X_i < x, Y_i < y)$ et $H(x,y) = \mathbb{P}(X_i > x, Y_i > y)$. Un estimateur naturel de V est donc

$$V_n = \frac{4}{n-1} \sum_{k=1}^{n} \left(F_n(X_k, Y_k) + H_n(X_k, Y_k) - \bar{F}_n - \bar{H}_n \right)^2$$

où

$$F_n(x,y) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{X_k < x, Y_k < y} \qquad H_n(x,y) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{X_k > x, Y_k > y}$$

$$\bar{F}_n = \frac{1}{n} \sum_{k=1}^{n} F_n(X_k, Y_k) \qquad \bar{H}_n = \frac{1}{n} \sum_{k=1}^{n} H_n(X_k, Y_k)$$

Finalement, sous H_0,

$$K_n := \frac{\sqrt{n}T_n}{\sqrt{V_n}} \text{ converge en loi vers la loi } \mathcal{N}(0,1).$$

La zone de rejet de H_0 du test de Kendall corrigé est donc de la forme $R_{n,\alpha} = \{|K_n| > c_\alpha\}$ où c_α est le quantile d'ordre $1 - (\alpha/2)$ de la loi $\mathcal{N}(0,1)$, ce qui fournit un test asymptotiquement bien calibré. On notera à présent CKendall le test basé sur la statistique K_n.

Pour illustrer cela, on se place d'abord dans le cadre du modèle (2.1), pour lequel $\rho = \pi = 0$. On simule $N = 2000$ échantillons de taille n selon le modèle décrit ci-dessus, et pour chaque test on indique la fréquence de rejet de H_0 au niveau 5%. On considére ici 3 tests : le test CKendall basé sur la statistique K_n, le test de Pearson, et le test de Kendall (non corrigé).

Les résultats sont présentés dans la Table 3 (niveau $\alpha = 5\%$).

Considérons à présent un second exemple : on simule, pour différentes valeurs de n, des couples $(X_i, Y_i)_{1 \leq i \leq n}$ i.i.d., selon le modèle

$$Y_i = (X_i \cdot (\varepsilon_i - 0.5))^3 \tag{3.1}$$

n	20	30	40	50	60	70	80	90	100	200
CKendall	0.089	0.072	0.063	0.056	0.056	0.049	0.058	0.053	0.046	0.054
Pearson	0.36	0.361	0.373	0.362	0.378	0.371	0.38	0.371	0.384	0.381
Kendall	0.174	0.173	0.176	0.195	0.173	0.177	0.193	0.2	0.182	0.20

TABLE 3 – Fréquence de rejet des différents tests au niveau 5% pour n couples tirés selon le modèle (2.1)

n	20	30	40	50	60	70	80	90	100	200
CKendall	0.086	0.071	0.066	0.058	0.058	0.056	0.056	0.048	0.056	0.052
Pearson	0.459	0.432	0.416	0.375	0.385	0.395	0.374	0.379	0.377	0.385
Kendall	0.215	0.234	0.231	0.249	0.24	0.251	0.245	0.242	0.245	0.239

TABLE 4 – Fréquence de rejet des différents tests au niveau 5% pour n couples tirés selon le modèle (3.1)

où les $(X_i)_{1\leq i\leq n}$ et les $(\epsilon_i)_{1\leq i\leq n}$ sont deux suites indépendantes, les $(X_i)_{1\leq i\leq n}$ étant des variables i.i.d. de loi $\mathcal{U}([0,1])$, les $(\varepsilon_i)_{1\leq i\leq n}$ étant des variables i.i.d. de loi $\mathcal{B}(0.5)$ (voir Figure 2). On peut facilement voir, que, pour ce modèle, $\rho = \pi = 0$.

Les résultats sont présentés dans la Table 4 (niveau $\alpha = 5\%$).

Les commentaires pour ces deux modèles très différents sont à peu près les mêmes : pour CKendall, la fréquence de rejet de H_0 est en dessous de 7% dès que $n \geq 40$, et entre 4.5% et 6% dès que $n \geq 50$. Comme prévu, il semble bien y avoir convergence vers le niveau 5%.

Le test de Pearson est très mal calibré, avec des fréquences de rejet supérieures à 35%.

Le test de Kendall (non corrigé) est mal calibré, avec des fréquences de rejet supérieures à 17%.

En conclusion, pour tester la corrélation au sens de Kendall, **nous préconisons de présenter aux étudiants de niveau L3/M1 le test CKendall, plutôt que le test de Kendall non corrigé.**

Remarque : Lorsque les variables X_i et Y_i sont continues, mais observées avec un arrondi trop frustre, la statistique T_n et sa correction K_n peuvent mal se comporter s'il y a trop d'ex-aequo (si $x_i = x_j$ ou $y_i = y_j$, l'indicatrice dans t_n vaut 0). Cela peut se corriger par une simple procédure de randomisation (s'il y a un ex-aequo, on tire à pile où face pour s'avoir si l'indicatrice vaut 0 ou 1). Pour nos simulations, cette correction s'est bien sûr avérée inutile. Cette remarque est aussi valable pour le test de Mann-Whitney de la Section 4.

Remarque : On peut aussi corriger le test de corrélation de Spearman (voir [9]), qui teste $H_0 : \rho_S = 0$ contre $H_1 : \rho_S \neq 0$, où ρ_S est le coefficient de corrélation entre les variables $F_X(X_i)$ et $F_Y(Y_i)$ uniformément distribuées sur $[0,1]$ (ici $F_X(x) = \mathbb{P}(X \leq x)$ et $F_Y(y) = \mathbb{P}(Y \leq y)$). Comme les tests de Kendall et Pearson, ce test est mal calibré si les variables X_i et Y_i ne sont pas indépendantes : pour les modèles (2.1) et (3.1), les fréquences de rejet de ce test sont autour de 13%, quelle que soit la taille n de l'échantillon. Comme pour le test de Kendall, Hoeffding [8] a montré que la statistique de test de Spearman peut s'exprimer à l'aide d'une U-statistique, ce qui fournit une expression exacte de la variance limite ; on peut en déduire un estimateur consistant de la variance limite, et obtenir un test de niveau asymptotique α. L'expression de la variance limite étant un peu plus compliquée que celle du test de Kendall, nous nous sommes restreints ici au test de Kendall.

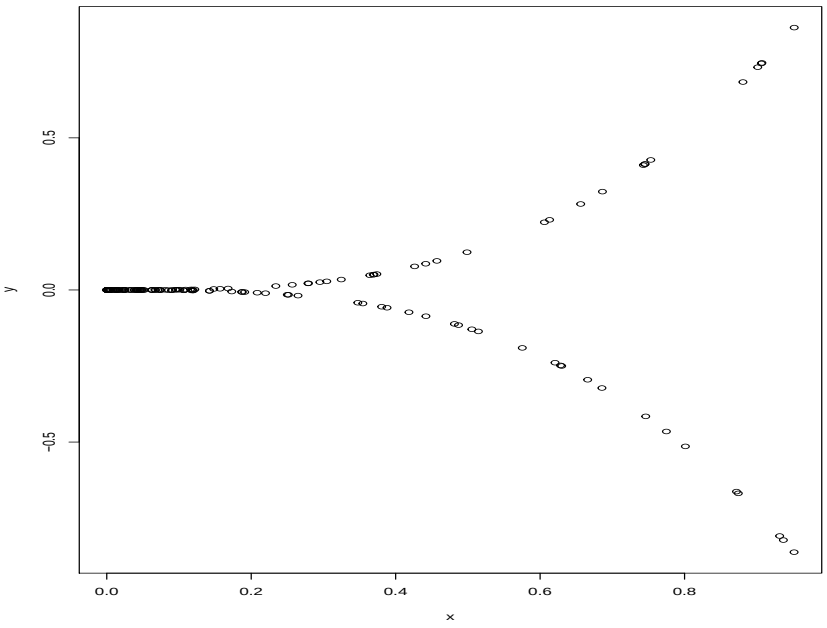

FIGURE 2 – Nuage de point de 150 couples (x_i, y_i) tirés selon le modèle (3.1)

4 Le test de Mann-Whitney

Soient $(X_1, \ldots, X_{n_1})$ et $(Y_1, \ldots, Y_{n_2})$ deux suites indépendantes de variables aléatoires quantitatives continues, indépendantes et identiquement distribuées.

On cherche à savoir si les variables Y_i ont tendance à prendre des valeurs plus grandes ou plus petites que les variables X_i. Notons $\mathrm{Med}(Y - X)$ la médiane de $X_i - Y_i$. Pour répondre à la question, on peut par exemple tester

$$H_0 : \mathrm{Med}(Y - X) = 0 \quad \text{contre} \quad H_1 : \mathrm{Med}(Y - X) \neq 0\,.$$

Mann et Whitney [12] ont proposé la statistique de test

$$T_{n_1,n_2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(\mathbf{1}_{X_i < Y_j} - 0.5 \right) .$$

Il est assez facile de voir, que si X_i et Y_i ont même loi, alors la loi de $T_{n_1,n2}$ est libre (i.e. ne dépend pas de la loi commune des X_i et des Y_i), en se ramenant à deux suites indépendantes $U_1, \ldots, U_{n_1}$ et $V_1, \ldots, V_{n_2}$ de variables i.i.d. de loi $\mathcal{U}([0, 1])$. Par conséquent, si X_i et Y_i ont même loi, T_{n_1,n_2} suit une loi connue et tabulée. Le test de Mann-Whitney est construit à partir des quantiles de cette loi. Il est bien connu que ce test est en fait équivalent au test de la somme des rangs de Wilcoxon.

Mais si X_i et Y_i ne suivent pas la même loi, la loi de la statistique T_{n_1,n_2} n'a pas de raison d'être libre sous H_0.

On peut donc poser le constat suivant : **dans un contexte général, le test de Mann-Whitney n'est pas bien calibré pour tester $\mathrm{Med}(Y - X) = 0$.**

On peut néanmoins régler (asymptotiquement) ce problème, en considérant la distribution limite de $a_{n_1,n_2}T_{n_1,n_2}$ sous H_0, pour une normalisation a_{n_1,n_2} aéquate. En partant de la décomposition de Hoeffding de la U-statistique T_{n_1,n_2} (voir par exemple [13], exemple 12.7), on voit que, sous H_0,

$$\frac{T_{n_1,n_2}}{\sqrt{\frac{V_1}{n_1}+\frac{V_2}{n_2}}} \quad \text{converge en loi lorsque } n_1, n_2 \to \infty \text{ vers la loi } \mathcal{N}(0,1),$$

avec

$$V_1 = \mathrm{Var}(H_Y(X_1)) \quad \text{et} \quad V_2 = \mathrm{Var}(F_X(Y_1))\,,$$

où $F_X(x) = \mathbb{P}(X_1 < x)$ et $H_Y(x) = \mathbb{P}(Y_1 > x)$. Des estimateurs naturels de V_1 et V_2 sont

$$V_{1,n_1,n_2} = \frac{1}{n_1-1}\sum_{k=1}^{n_1}(H_{n_2}(X_k) - \bar{H}_{n_2})^2 \quad \text{et} \quad V_{2,n_1,n_2} = \frac{1}{n_2-1}\sum_{k=1}^{n_2}(F_{n_1}(Y_k) - \bar{F}_{n_1})^2$$

où

$$F_{n_1}(x) = \frac{1}{n_1}\sum_{k=1}^{n_1}\mathbf{1}_{X_k<x} \qquad H_{n_2}(y) = \frac{1}{n_2}\sum_{k=1}^{n_2}\mathbf{1}_{Y_k>y}$$

$$\bar{F}_{n_1} = \frac{1}{n_2}\sum_{k=1}^{n_2}F_{n_1}(Y_k) \qquad \bar{H}_{n_2} = \frac{1}{n_1}\sum_{k=1}^{n_1}H_{n_2}(X_k)$$

Finalement, sous H_0,

$$MW_{n_1,n_2} := \frac{T_{n_1,n_2}}{\sqrt{\frac{V_{1,n_1,n_2}}{n_1}+\frac{V_{2,n_1,n_2}}{n_2}}} \quad \text{converge en loi lorsque } n_1, n_2 \to \infty \text{ vers la loi } \mathcal{N}(0,1).$$

La zone de rejet de H_0 du test de Mann-Whitney corrigé est donc de la forme $R_{n_1,n_2,\alpha} = \{|MW_{n_1,n_2}| > c_\alpha\}$ où c_α est le quantile d'ordre $1-(\alpha/2)$ de la loi $\mathcal{N}(0,1)$, ce qui fournit un test asymptotiquement bien calibré. On notera à présent CMW le test basé sur la statistique MW_{n_1,n_2}.

Pour illustrer cela, on simule $N = 2000$ échantillons de taille n_1 selon la loi $\mathcal{U}([-0.5, 0.5])$, et 2000 échantillons de taille n_2 selon la loi $\mathcal{N}(0,(0.04)^2)$ (i.e. écart-type $= 0.04$) selon un rapport de $n_2 = 3n_1$ (voir Figure 3). Pour chaque test on indique la fréquence de rejet de H_0 au niveau 5%. On considére ici 4 tests : le test CMW basé sur la statistique MW_{n_1,n_2}, le test de Mann-Whitney (M-W) non corrigé, le test de Welch (qui teste l'égalité des espérances sans supposer l'égalité des variances, voir Section 5), et le test de Kolmogorov-Smirnov (K-S) à deux échantillons (voir par exemple [4], pages 309-314) qui teste l'égalité des distributions.

Les résultats sont présentés dans la Table 5 (niveau $\alpha = 5\%$).

Sur ce jeu de simulations (Table 5), on voit que, pour CMW et toutes les tailles d'échantillons proposées, la fréquence de rejet de H_0 est en dessous de 7%. Elle se situe entre 5% et 6% dès que $n_1 \geq 50, n_2 \geq 150$. Comme prévu, il semble bien y avoir convergence vers le niveau 5%.

Le test de Welch est également bien calibré, les fréquences de rejet se situant toutes entre 4.7% et 5.8%.

Le test de Mann-Whitney non corrigé est mal calibré, puisque sa fréquence de rejet oscille entre 16% et 18%.

Comme prévu, le test de Kolmogorov-Smirnov à deux échantillons détecte très bien la différence entre les deux distributions, de façon systématique dès que $n_1 \geq 30, n_2 \geq 90$.

En conclusion, pour tester si $\mathrm{Med}(Y-X)=0$, **nous préconisons de présenter aux étudiants de niveau L3/M1 le test CMW, plutôt que le test de Mann-Whitney non corrigé.**

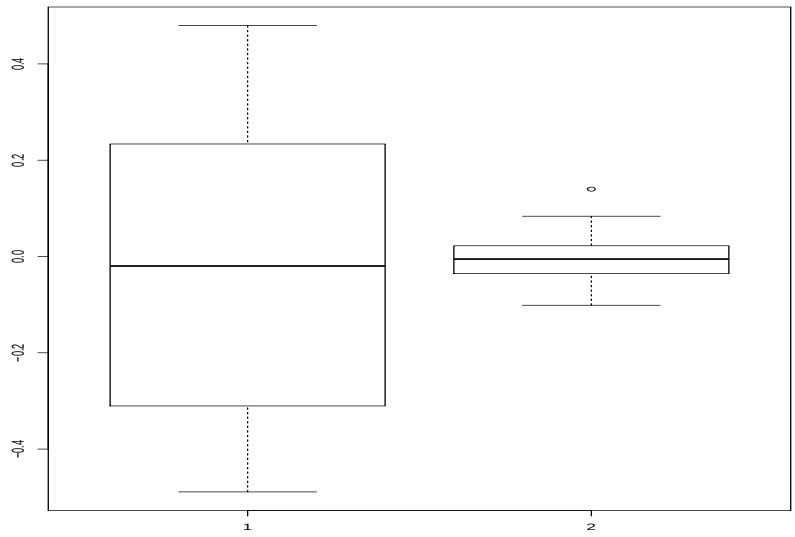

FIGURE 3 – Box plot de 1 : 60 tirages selon la loi $\mathcal{U}([-0.5, 0.5])$, et 2 : 180 tirages selon la loi $\mathcal{N}(0, (0.04)^2)$

$n_1; n_2$	20 ;60	30 ;90	40 ;120	50 ;150	60 ;180	70 ;210	80 ;240	90 ;270	100 ;300
CMW	0.066	0.063	0.057	0.06	0.06	0.055	0.051	0.054	0.052
M-W	0.17	0.166	0.174	0.176	0.16	0.174	0.169	0.169	0.176
Welch	0.047	0.056	0.047	0.048	0.49	0.53	0.051	0.055	0.058
K-S	0.997	1	1	1	1	1	1	1	1

TABLE 5 – Fréquence de rejet des différents tests au niveau 5% pour n_1 couples tirés selon la loi $\mathcal{U}([-0.5, 0.5])$, et n_2 couples selon la loi $\mathcal{N}(0, (0.04)^2)$

5 Tests d'égalité des espérances

Le contexte est le même qu'en Section 4 : $(X_1, \ldots, X_{n_1})$ et $(Y_1, \ldots, Y_{n_2})$ sont deux suites indépendantes de variables aléatoires quantitatives, indépendantes et identiquement distribuées, ayant un moment d'ordre 2. Notons μ_X et σ_X l'espérance et l'écart-type des X_i, μ_Y et σ_Y l'espérance et l'écart-type des Y_i.

Pour tester

$$H_0 : \mu_X = \mu_Y \quad \text{contre} \quad H_1 : \mu_X \neq \mu_Y \, .$$

on utilise souvent la statistique de test de Student (introduite par Fisher [6])

$$T_{n_1,n_2} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \, ,$$

avec

$$S^2 = \frac{1}{n_1 + n_2 - 2} \left(\sum_{k=1}^{n_1} (X_k - \bar{X}_{n_1})^2 + \sum_{k=1}^{n_2} (Y_k - \bar{Y}_{n_2})^2 \right) \, .$$

$n_1; n_2$	5 ;5	5 ;6	5 ;7	5 ;8	5 ;9	5 ;10	5 ;11	5 ;12	5 ;13
Stud.	0.082	0.110	0.138	0.165	0.190	0.215	0.237	0.259	0.278
Welch	0.051	0.050	0.051	0.050	0.050	0.050	0.050	0.050	0.050

TABLE 6 – Fréquence de rejet des tests de Welch et Student au niveau 5% pour n_1 couples tirés selon la loi $\mathcal{N}(0,1)$, et n_2 couples selon la loi $\mathcal{N}(0,(0.04)^2)$

L'intérêt de cette statistique étant que, dans le cas où les variables X_i et Y_i sont Gaussiennes et que $\sigma_X = \sigma_Y$, sa distribution exacte sous H_0 est connue : c'est la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Cependant, comme l'ont noté plusieurs auteurs, dont Welch (voir [14] et références incluses), dans le cas des petits échantillons Gaussiens pour lesquels $\sigma_X \neq \sigma_Y$, la distribution de T_{n_1,n_2} ne suit pas sous H_0 une loi de Student$(n_1 + n_2 - 2)$ (et ceci même si $n_1 = n_2$). Dans ce cas, il faut utiliser la statistique naturelle

$$T'_{n_1,n_2} = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}},$$

avec

$$S_X^2 = \frac{1}{n_1 - 1}\sum_{k=1}^{n_1}(X_k - \bar{X}_{n_1})^2\,, \quad S_Y^2 = \frac{1}{n_2 - 1}\sum_{k=1}^{n_2}(Y_k - \bar{Y}_{n_2})^2\,,$$

et sa loi approchée sous H_0 (une loi de Student de degré ν non entier calculé à partir des observations, voir encore [14]).

Pour illustrer cela, on simule $N = 10^6$ échantillons de taille n_1 selon la loi $\mathcal{N}(0,1)$ et $N = 10^6$ échantillons de taille n_2 selon la loi $\mathcal{N}(0,(0.04)^2)$. On fixe $n_1 = 5$ et on fait varier n_2 de 5 à 13. Pour le test de Student et le test de Welch, on indique la fréquence de rejet de H_0 au niveau 5%.

Les résultats sont présentés dans la Table 6 (niveau $\alpha = 5\%$).

Sur ce jeu de simulations (Table 6), on voit que le test de Welch est toujours parfaitement calibré. Le test de Student est en revanche mal calibré, avec des fréquences de rejet comprises entre 8% et 28%, d'autant plus grandes que le rapport n_2/n_1 est grand.

Considérons à présent le cas où les variables X_i ou Y_i ne sont pas a priori Gaussiennes. D'un point de vue asymptotique, lorsque $\sigma_X = \sigma_Y$ ou lorsque $n_1/n_2 \to 1$, les statistiques T_{n_1,n_2} et T'_{n_1,n_2} convergent en loi sous H_0 vers la loi $\mathcal{N}(0,1)$ lorsque $n_1, n_2 \to \infty$. Mais si $\sigma_X \neq \sigma_Y$ et si n_1/n_2 ne converge pas vers 1, la statistique T_{n_1,n_2} est mal normalisée sous H_0, et sa variance limite n'est pas égale à 1.

On peut donc poser le constat suivant : **dans un contexte général, le test de Student basé sur la statistique T_{n_1,n_2} n'est pas bien calibré pour tester $\mu_X = \mu_Y$.**

En revanche, la convergence en loi sous H_0 de T'_{n_1,n_2} vers la loi $\mathcal{N}(0,1)$ a toujours lieu lorsque $n_1, n_2 \to \infty$. La zone de rejet pour tester H_0 est donc de la forme $R_{n_1,n_2,\alpha} = \{|T'_{n_1,n_2}| > c_\alpha\}$ où c_α est le quantile d'ordre $1 - (\alpha/2)$ de la loi $\mathcal{N}(0,1)$, ce qui fournit un test asymptotiquement bien calibré. On peut aussi utiliser le quantile de la loi Student(ν) proposée par Welch (ce qu'on fera dans les simulations ci-dessous) car, lorsque n_1, n_2 tendent vers l'infini, la loi Student(ν) converge vers la loi $\mathcal{N}(0,1)$.

Pour illustrer cela, on reprend l'exemple de la Section 4 : on simule $N = 2000$ échantillons de taille n_1 selon la loi $\mathcal{U}([-0.5, 0.5])$, et 2000 échantillons de taille n_2 selon la loi $\mathcal{N}(0,(0.04)^2)$ (i.e. écart-type

$n_1; n_2$	20 ;60	30 ;90	40 ;120	50 ;150	60 ;180	70 ;210	80 ;240	90 ;270	100 ;300
Stud.	0.246	0.267	0.246	0.243	0.25	0.263	0.253	0.274	0.261
Welch	0.054	0.047	0.051	0.049	0.049	0.051	0.054	0.051	0.05

TABLE 7 – Fréquence de rejet des tests de Welch et Student au niveau 5% pour n_1 couples tirés selon la loi $\mathcal{U}([-0.5, 0.5])$, et n_2 couples selon la loi $\mathcal{N}(0, (0.04)^2)$

$= 0.04$) selon un rapport de $n_2 = 3n_1$. Pour le test de Student et le test de Welch, on indique la fréquence de rejet de H_0 au niveau 5%.

Les résultats sont présentés dans la Table 7 (niveau $\alpha = 5\%$).

Sur ce jeu de simulations (Table 7), on voit que le test de Welch est toujours bien calibré, avec des fréquences de rejet qui oscillent entre 4.7% et 5.4%. Le test de Student est mal calibré, avec des fréquences de rejet qui oscillent entre 24% et 28%.

En conclusion, pour tester l'égalité des espérances, **nous préconisons de présenter aux étudiants de niveau L2/L3 le test de Welch, plutôt que le test de Student**.

Remarque : De même, pour tester l'égalité des espérances de p échantillons de variables i.i.d., nous préconisons d'utiliser la procèdure d'ANOVA corrigée par James [10] et Welch [15] (oneway.test sous R).

6 Tests d'égalité des variances

Le contexte est le même qu'en Section 5 : $(X_1, \ldots, X_{n_1})$ et $(Y_1, \ldots, Y_{n_2})$ sont deux suites indépendantes de variables aléatoires quantitatives, indépendantes et identiquement distribuées, ayant un moment d'ordre 4. Notons σ_X l'écart-type des X_i et σ_Y l'écart-type des Y_i.

Pour tester

$$H_0 : \sigma_X = \sigma_Y \quad \text{contre} \quad H_1 : \sigma_X \neq \sigma_Y \,,$$

on utilise souvent la statistique de Fisher

$$F_{n_1,n_2} = \frac{S_X^2}{S_Y^2}, \quad \text{avec} \quad S_X^2 = \frac{1}{n_1 - 1} \sum_{k=1}^{n_1} (X_k - \bar{X}_{n_1})^2 \,, \quad S_Y^2 = \frac{1}{n_2 - 1} \sum_{k=1}^{n_2} (Y_k - \bar{Y}_{n_2})^2 \,.$$

L'intérêt de cette statistique étant que, dans le cas où les variables X_i et Y_i sont Gaussiennes, sa distribution exacte sous H_0 est connue : c'est la loi de Fisher$(n_1 - 1, n_2 - 1)$. Le test de Fisher d'égalité des variances est construit à partir des quantiles de cette loi.

Mais si X_i ou Y_i ne sont pas des variables Gaussiennes, la loi de F_{n_1,n_2} sous H_0 ne suit pas une loi de Fisher$(n_1 - 1, n_2 - 1)$.

On peut donc poser le constat suivant : **dans un contexte général, le test de Fisher n'est pas bien calibré pour tester l'égalité des variances**. Voir aussi [2] pour un constat similaire.

Notons que la statistique de test F_{n_1,n_2} converge presque sûrement vers σ_X^2/σ_Y^2 (et donc vers 1 sous H_0) lorsque $n_1, n_2 \to \infty$. Mais il ne semble pas aisé de trouver une modification simple de cette statistique de test pour laquelle on ait la convergence en loi sous H_0 vers une loi connue.

Bien entendu, ce problème de l'égalité des variances peut être reformulé comme un problème d'égalité d'espérances : on veut tester si $\mathbb{E}((X_1 - \mu_X)^2) = \mathbb{E}((Y_1 - \mu_Y)^2)$. Notons $A_i = (X_i - \bar{X}_{n_1})^2$

$n_1; n_2$	20 ;40	30 ;60	40 ;80	50 ;100	60 ;120	70 ;140	80 ;160	90 ;180	100 ;200
VFisher	0.003	0.004	0.002	0.005	0.003	0.002	0.003	0.003	0.002
Bartlett	0.004	0.003	0.002	0.005	0.003	0.002	0.003	0.003	0.002
Levene	0.049	0.056	0.051	0.056	0.051	0.051	0.05	0.053	0.052
VWelch	0.055	0.054	0.048	0.058	0.051	0.052	0.051	0.051	0.049

TABLE 8 – Fréquence de rejet des différents tests au niveau 5% pour n_1 couples tirés selon la loi $\mathcal{U}([0,1])$, et n_2 couples selon la loi $\mathcal{U}([0,1])$

$n_1; n_2$	20 ;40	30 ;60	40 ;80	50 ;100	60 ;120	70 ;140	80 ;160	90 ;180	100 ;200
VFisher	0.145	0.141	0.147	0.143	0.159	0.154	0.157	0.158	0.162
Bartlett	0.147	0.14	0.148	0.141	0.158	0.154	0.155	0.159	0.163
Levene	0.08	0.093	0.109	0.135	0.144	0.15	0.165	0.194	0.2
VWelch	0.049	0.054	0.056	0.054	0.053	0.052	0.048	0.051	0.054

TABLE 9 – Fréquence de rejet des différents tests au niveau 5% pour n_1 couples tirés selon la loi $\mathcal{N}(0,1)$, et n_2 couples selon la loi $\chi^2(2)/2$

et $B_i = (Y_i - \bar{Y}_{n_2})^2$. Comme en Section 5, la statistique naturelle s'écrit alors

$$T_{n_1,n_2} = \frac{\bar{A}_{n_1} - \bar{B}_{n_2}}{\sqrt{\frac{S_A^2}{n_1} + \frac{S_B^2}{n_2}}},$$

avec

$$S_A^2 = \frac{1}{n_1 - 1}\sum_{k=1}^{n_1}(A_k - \bar{A}_{n_1})^2\,, \quad S_B^2 = \frac{1}{n_2 - 1}\sum_{k=1}^{n_2}(B_k - \bar{B}_{n_2})^2\,.$$

De la même façon (ou presque) qu'en Section 5, on peut montrer que, sous H_0, T_{n_1,n_2} converge en loi vers la loi $\mathcal{N}(0,1)$ lorsque $n_1, n_2 \to \infty$. Ce test d'égalité des variances est donc un test "de type Welch" : on calcule la statistique de Welch, présentée en Section 5, sur les variables $(A_i)_{1\le i\le n_1}$ et $(B_i)_{1\le i\le n_2}$. On notera à présent VWelch le test basé sur la statistique T_{n_1,n_2}.

Pour illustrer cela, on simule $N = 2000$ échantillons de taille n_1 selon la loi $\mathcal{U}([0,1])$, et 2000 échantillons de taille n_2 selon la loi $\mathcal{U}([0,1])$ selon un rapport de $n_2 = 2n_1$. Pour chaque test on indique la fréquence de rejet de H_0 au niveau 5%. On considére ici 4 tests : le test VFisher d'égalité des variances basé sur la statistique F_{n_1,n_2}, le test de Bartlett d'égalité des variances (voir [1]), le test de Levene (modifié par Brown et Forsythe [3], qui teste en fait l'égalité des écarts absolus à la médiane), et le test VWelch décrit ci-dessus.

Les résultats sont présentés dans la Table 8 (niveau $\alpha = 5\%$).

Sur ce jeu de simulations (Table 8), on voit que les tests de Levene et VWelch ont toujours une fréquence de rejet proche de 5%, et sont donc bien calibrés.

Les tests VFisher et Bartlett se comportent de façon semblable, et ne sont pas bien calibrés, avec une fréquence de rejet autour de 0.3%, très inférieure au niveau attendu.

Considérons à présent un second exemple : on simule $N = 2000$ échantillons de taille n_1 selon la loi $\mathcal{N}(0,1)$, et 2000 échantillons de taille n_2 selon la loi $\chi^2(2)/2$ selon un rapport de $n_2 = 2n_1$ (voir Figure 4).

Les résultats sont présentés dans la Table 9 (niveau $\alpha = 5\%$).

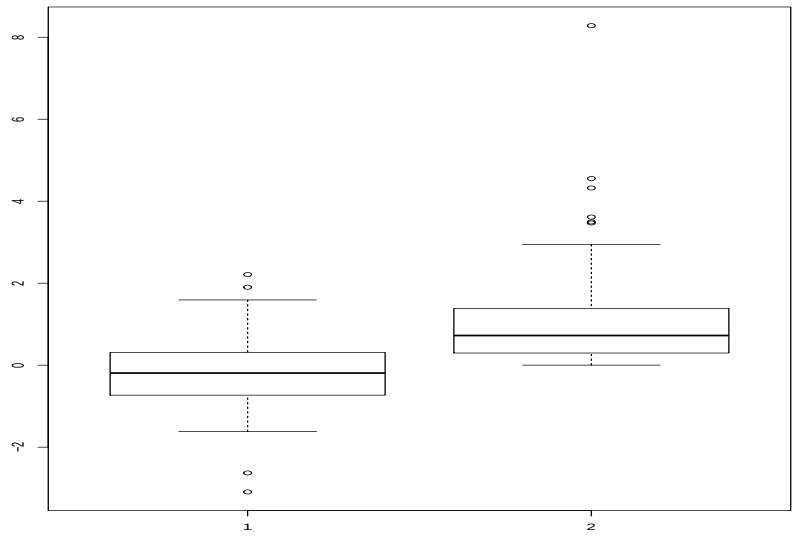

FIGURE 4 – Box plot de 1 : 60 tirages selon la loi $\mathcal{N}(0,1)$, et 2 : 120 tirages selon la loi $\chi^2(2)/2$

Sur ce jeu de simulations (Table 9), on voit que le test VWelch a toujours une fréquence de rejet proche de 5%, et est donc bien calibré.

Le test de Levene a une fréquence de rejet toujours supérieure à 8% qui a tendance à augmenter (de l'ordre de 20% pour $n_1 = 100, n_2 = 200$). Ce n'est pas étonnant car (contrairement à ce qu'annonce le titre de l'article de Brown et Forsythe [3]), ce n'est pas un test d'égalité des variances, mais un test d'égalité des écarts absolus à la médiane ; or, pour les variables de cet exemple, les variances sont identiques, mais les écarts absolus à la médiane sont légèrement différents.

Les tests VFisher et Bartlett se comportent de façon semblable, et ne sont pas bien calibrés, avec des fréquences de rejet qui oscillent entre 14% et 16.5%.

En conclusion, pour tester l'égalité des variances, **nous préconisons de présenter aux étudiants de niveau L2/L3 le test VWelch, plutôt que le test VFisher**. Le test VFisher ne devrait être présenté que dans le cadre strict des échantillons Gaussiens.

Remarque : Ce que nous avons décrit ici peut facilement s'étendre à plus de deux échantillons. Pour tester l'égalité des variances de p échantillons de variables i.i.d., on peut utiliser la procèdure d'ANOVA corrigée par James [10] et Welch [15] (oneway.test sous R) sur le carré des variables recentrées empiriquement. Nous proposons en appendice une commande R, Voneway, qui permet de réaliser ce test.

7 Test du signe et test signe et rang de Wilcoxon

Le contexte est le même qu'en Sections 2 et 3 : $(X_1, Y_1), \ldots, (X_n, Y_n)$ sont des couples de variables aléatoires quantitatives indépendants et identiquement distribués. On suppose de plus que les variables sont continues.

Comme en Section 4, on cherche à savoir si les variables Y_i ont tendance à prendre des valeurs plus grandes ou plus petites que les variables X_i. Mais ici, les variables X_i et Y_i ne sont a priori pas

indépendantes. De façon classique, on considère la série des différences $D_i = Y_i - X_i$, et on se ramène donc à un problème de statistique univariée.

Pour répondre à la question, on peut par exemple tester

$$H_0 : \mathrm{Med}(D) = 0 \quad \text{contre} \quad H_1 : \mathrm{Med}(D) \neq 0\,.$$

La statistique du test du signe est

$$T_n = \sum_{i=1}^{n} \mathbf{1}_{D_i>0}$$

qui suit sous H_0 une loi binômiale de paramètres $(n, 0.5)$. Modulo le fait que la loi de T_n est discrète, de sorte que le niveau α ne peut pas être atteint exactement, ce test sera donc bien calibré (il faut utiliser une procédure de randomisation pour obtenir un test de niveau exact α).

On propose souvent comme alternative au test du signe, le test signe et rang de Wilcoxon, dont la statistique s'écrit

$$W_n = \sum_{i=1}^{n} R_i \mathbf{1}_{D_i>0}\,,$$

où R_i est le rang de $|D_i|$ dans l'échantillon $(|D_k|)_{1\leq k\leq n}$. Si les variables D_i sont symétriques (ce qui signifie que D_i a même loi que $-D_i$), Wilcoxon [16] a montré que $|D_i|$ et $\mathbf{1}_{D_i>0}$ sont indépendantes ; on peut alors facilement en déduire que la loi de W_n est libre (i.e. ne dépend pas de la loi de D_i). Par conséquent, si les D_i sont des variables symétriques, la loi de W_n est connue et tabulée. Le test signe et rang de Wilcoxon est construit à partir des quantiles de cette loi.

Mais si les variables D_i ne sont pas symétriques, la loi de W_n n'a pas de raison d'être libre sous H_0.

On peut donc poser le constat suivant : **dans un contexte général, le test signe et rang de Wilcoxon n'est pas bien calibré pour tester Med(D)=0.** On reviendra en fin de section sur l'hypothèse nulle "naturelle" du test signe et rang, et comment le corriger.

Pour tester si Med(D)=0, on peut aussi utiliser l'estimateur naturel de la médiane, à savoir la médiane empirique $\mathrm{Med}(D)_n$. Sous l'hypothèse (peu restrictive) que la densité f_D de D_i existe et est strictement positive en 0, on sait que, sous H_0,

$$\sqrt{n}\,\mathrm{Med}(D)_n \text{ converge en loi vers la loi } \mathcal{N}\left(0, 1/(2f_D(0))^2\right)$$

(voir par exemple [13], Corollaire 21.5). Si $\hat{f}_n(0)$ est un estimateur consistant de $f_D(0)$, on peut donc proposer la statistique

$$T'_n = 2\sqrt{n\hat{f}_n(0)}\,\mathrm{Med}(D)_n$$

qui converge en loi sous H_0 vers la loi $\mathcal{N}(0,1)$. Bien sûr, l'estimation de $f_D(0)$ est un problème délicat, et ce n'est pas le lieu d'en discuter ici. Pour les simulations qui vont suivre, on a choisi un estimateur à noyau, avec le noyau rectangulaire, et on a laissé le logiciel R choisir la fenêtre (choix par défaut).

Enfin, une troisième façon de procéder consiste à utiliser l'intervalle de confiance pour la médiane basé sur la statistique d'ordre $(D_{(i)})_{1\leq i\leq n}$. Pour simplifier, nous ne donnons ici que l'intervalle de niveau asymptotique $1-\alpha$. Soit

$$k_{n,\alpha} = \left[-c_\alpha \frac{\sqrt{n}}{2} + \frac{n}{2}\right], \quad \ell_{n,\alpha} = \left[c_\alpha \frac{\sqrt{n}}{2} + \frac{n+1}{2}\right],$$

les crochets désignant la partie entière, et c_α le quantile d'ordre $1-(\alpha/2)$ de la loi $\mathcal{N}(0,1)$. Alors l'intervalle de confiance

$$IC_{1-\alpha} = [D_{(k_{n,\alpha})}, D_{(\ell_{n,\alpha})}]\,, \tag{7.1}$$

n	30	40	50	60	70	80	90	100	300	500
Signe	0.039	0.044	0.036	0.038	0.051	0.037	0.048	0.042	0.043	0.044
Med0	0.025	0.032	0.037	0.042	0.04	0.042	0.048	0.044	0.045	0.05
IC0	0.059	0.052	0.051	0.052	0.045	0.046	0.052	0.048	0.051	0.05
W-Test	0.094	0.117	0.131	0.153	0.172	0.185	0.195	0.219	0.504	0.741

TABLE 10 – Fréquence de rejet des différents tests au niveau 5% pour n variables D_i tirées selon la loi $\chi^2(4)-\text{Med}(\chi^2(4))$

est un intervalle de confiance de niveau asymptotique $1-\alpha$ pour $\text{Med}(D)$ (voir par exemple [13], exemple 21.8). Pour tester si $\text{Med}(D)=0$, le test ayant pour zone de rejet $R_{n,\alpha} = \{0 \notin [D_{(k_{n,\alpha})}, D_{(\ell_{n,\alpha})}]\}$ est donc un test de niveau asymptotique α. On notera ce test IC0.

Pour illustrer cela, on simule $N = 2000$ échantillons de taille n selon la loi $\chi^2(4)-\text{Med}(\chi^2(4))$ (loi des variables D_i, voir Figure 5). Pour chaque test on indique la fréquence de rejet de H_0 au niveau 5%. On considére ici 4 tests : le test du signe basé sur la statistique T_n, le test basé sur la médiane empirique et la statistique T'_n (noté Med0), le test IC0 basé sur l'intervalle de confiance (7.1), et le test signe et rang de Wilcoxon (W-Test) basé sur la statistique W_n.

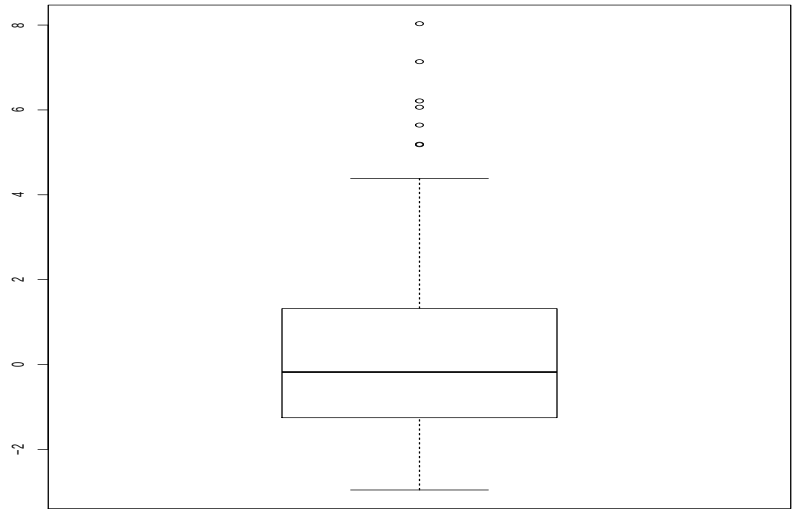

FIGURE 5 – Box plot de 100 tirages selon la loi $\chi^2(4)-\text{Med}(\chi^2(4))$

Les résultats sont présentés dans la Table 10 (niveau $\alpha = 5\%$).

Sur ce jeu de simulations (Table 10), on voit que le test du signe et Med0 se comportent de façon comparable, les fréquences de rejet se situant toutes entre 2.5% et 5%, supérieures à 4% dès que $n \geq 90$. Il faut donc attendre assez longtemps pour se rapprocher du seuil 5%. Pour le test du signe, cela est dû au fait que sa loi est discrète ; on peut régler ce problème (pour les échantillons de petite taille) en construisant un test randomisé de niveau exact 5%. Pour le test Med0, cela est sans doute dû au fait que la vitesse de convergence de $\hat{f}_n(0)$ vers $f_D(0)$ est assez lente (plus lente que $\sqrt{n}$).

Le test IC0 est lui bien calibré, avec des fréquences de rejet toujours comprises entre 4.5% et 6%.

n	30	40	50	60	70	80	90	100	300	500
Signe	0.131	0.158	0.171	0.187	0.254	0.256	0.333	0.338	0.80	0.95
Med0	0.118	0.153	0.2	0.235	0.278	0.288	0.343	0.372	0.829	0.965
IC0	0.075	0.096	0.172	0.187	0.195	0.256	0.258	0.338	0.80	0.95

TABLE 11 – Fréquence de rejet des différents tests au niveau 5% pour n variables D_i tirées selon la loi $\chi^2(4)-\text{Med}(\chi^2(4))+0.5$

Le test signe et rang (W-Test) n'est pas bien calibré pour tester $\text{Med}(D)=0$, avec des fréquences de rejet supérieures à 15% dès que $n \geq 60$, et une fréquence de rejet d'environ 75% pour $n = 500$.

Pour compléter, nous proposons une étude de puissance pour comparer le test du signe, le test Med0 et le test IC0. Cette fois, on simule $N = 2000$ échantillons de taille n selon la loi $\chi^2(4)-\text{Med}(\chi^2(4))+0.5$ (loi des variables D_i).

Les résultats sont présentés dans la Table 11 (niveau $\alpha = 5\%$).

Sur ce jeu de simulations (Table 11), le test Med0 semble légèrement plus puissant que le test du signe et que le test IC0 dès que $n \geq 50$.

Notre première conclusion est donc : **pour tester Med(D)=0, nous préconisons de présenter le test du signe, le test IC0 ou le test basé sur la médiane empirique (pour des tailles d'échantillons pas trop petites). Le test signe et rang n'est pas bien calibré pour tester cette hypothèse.**

En fait, le test signe et rang permet de tester

$$H_0' : \text{Med}(D_1 + D_2) = 0 \quad \text{contre} \quad H_1 : \text{Med}(D_1 + D_2) \neq 0\,.$$

Mais là encore, il n'est pas bien calibré pour tester cette hypothèse. On peut néanmoins régler (asymptotiquement) ce problème, en considérant la distribution limite de $\sqrt{n}(2U_n/(n(n-1)) - 0.5)$ sous H_0', avec

$$U_n = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathbf{1}_{D_i+D_j>0}\,.$$

Notons que

$$W_n = U_n + \sum_{i=1}^{n} \mathbf{1}_{D_i>0}\,,$$

le second terme à droite dans l'égalité étant asymptotiquement négligeable par rapport à U_n. En partant de la décomposition de Hoeffding de la U-statistique U_n (voir par exemple [13], exemple 12.4), on voit que la variable $\sqrt{n}(2U_n/(n(n-1)) - 0.5)$ converge en loi sous H_0' vers la loi $\mathcal{N}(0, V)$, où

$$V = 4\text{Var}(F(-D_1))$$

et F est la fonction de répartition des variables D_i. Un estimateur naturel de V est donc

$$V_n = \frac{1}{n-1} \sum_{i=1}^{n} (F_n(-D_i) - \bar{F}_n)^2\,,$$

où

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{D_i \leq x} \quad \text{et} \quad \bar{F}_n = \frac{1}{n}\sum_{i=1}^{n} F_n(-D_i)\,.$$

Finalement, sous H_0',

$$W_n' = \frac{\sqrt{n}}{\sqrt{V_n}}\left(\frac{2U_n}{n(n-1)} - 0.5\right) \quad \text{converge en loi lorsque } n \to \infty \text{ vers la loi } \mathcal{N}(0,1).$$

La zone de rejet de H_0' du test signe et rang corrigé est donc de la forme $R_{n,\alpha} = \{|W_n'| > c_\alpha\}$ où c_α est le quantile d'ordre $1-(\alpha/2)$ de la loi $\mathcal{N}(0,1)$, ce qui fournit un test asymptotiquement bien calibré. On notera à présent CWTest le test basé sur la statistique W_n'.

Pour illustrer cela, on simule $N = 2000$ échantillons de taille n selon la loi

$$\Gamma(1/10,1)\text{-Med}(\Gamma(1/5,1))/2$$

(loi des variables D_i, voir Figure 6), de sorte que $\text{Med}(D_1 + D_2)=0$, mais $\text{Med}(D) \neq 0$ (rappelons que la somme de deux variables indépendantes de loi $\Gamma(1/10,1)$ suit une loi $\Gamma(1/5,1)$). Pour chaque test on indique la fréquence de rejet de H_0' au niveau 5%. On considère ici 5 tests : le test du signe basé sur la statistique T_n, le test basé sur la médiane empirique et la statistique T_n' (noté Med0), le test IC0 basé sur l'intervalle de confiance (7.1), le test signe et rang de Wilcoxon (W-Test) basé sur la statistique W_n, et le test signe et rang corrigé (CWTest) basé sur la statistique W_n'.

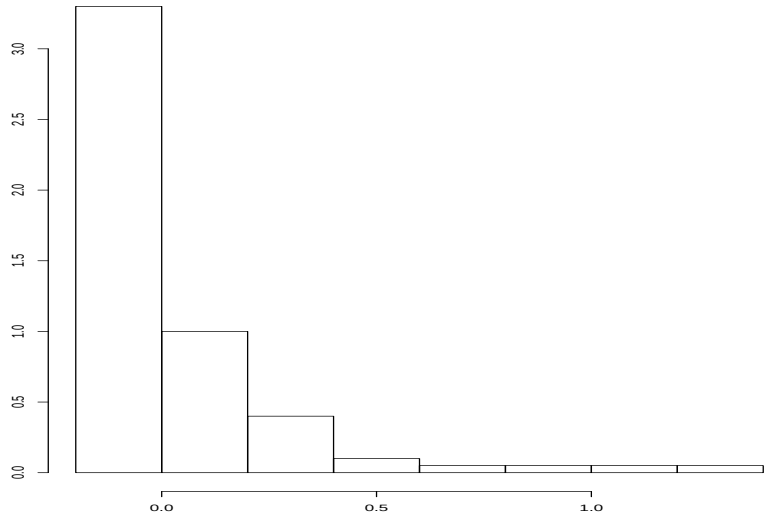

FIGURE 6 – Histogramme (densité de fréquence) de 100 tirages selon la loi $\Gamma(1/10,1)$-Med$(\Gamma(1/5,1))/2$

Les résultats sont présentés dans la Table 12 (niveau $\alpha = 5\%$).

Sur ce jeu de simulations (Table 12), on voit que, pour CWTest, les fréquences de rejet sont toutes inférieures à 7.1%, comprises entre 5% et 6% dès que $n \geq 50$.

Pour cette distribution non symétrique, le test W-Test est légèrement mal calibré avec des fréquences de rejet comprises entre 6.8% et 8.2%.

n	30	40	50	60	70	80	90	100	150	200
Signe	0.407	0.522	0.577	0.675	0.769	0.815	0.888	0.90	0.983	0.996
Med0	0.208	0.297	0.394	0.5	0.559	0.657	0.69	0.739	0.905	0.967
IC0	0.564	0.649	0.688	0.767	0.838	0.871	0.926	0.932	0.989	0.999
W-Test	0.068	0.069	0.08	0.07	0.077	0.074	0.074	0.074	0.074	0.082
CWTest	0.071	0.066	0.059	0.057	0.058	0.055	0.055	0.055	0.054	0.058

TABLE 12 – Fréquence de rejet des différents tests au niveau 5% pour n variables D_i tirées selon la loi $\Gamma(1/10,1)$-Med$(\Gamma(1/5,1))/2$

Comme prévu, le test du signe et les tests Med0 et IC0 détectent de mieux en mieux le fait que Med$(D)\neq 0$, avec une fréquence de rejet qui tend vers 1 lorsque $n \to \infty$. On notera que, sur ce jeu de simulations, le test du signe et le test IC0 sont bien plus puissants que le test Med0.

Notre seconde recommandation est donc de **bien préciser aux étudiants de L3/M1 l'hypothèse nulle du test signe et rang, à savoir Med$(D_1 + D_2)$=0. Pour tester cette hypothèse, nous préconisons de présenter le test CWTest, plutôt que le test signe et rang non corrigé.**

Appendice : Commandes R

Dans cet appendice, nous donnons les fonctions R correspondant aux différents tests que nous avons présentés. Ces fonctions ont été utilisées pour réaliser les simulations.

Cortest :
Correspond au test décrit en Section 2. La p-valeur est calculée à partir des quantiles de la loi Student$(n-2)$.

```
Cortest=function(X,Y)
{
  n <- length(X)
  R <- (X-mean(X))*(Y-mean(Y))
  T <- sqrt(n)*((n-1)/n)*cov(X,Y)/sqrt(((n-1)/n)*var(R))
  Pval <- 1-pt(abs(T),n-2)+pt(-abs(T),n-2)
  return(Pval)
}
```

Comme indiqué en fin de Section 2, on peut aussi utiliser les quantiles de la loi de $|T'_n|$ sous H_0 dans le cas Gaussien, mais il est alors nécessaire d'estimer ces quantiles. Dans la fonction qui suit, nous les estimons par Monte-Carlo ($N = 400000$). L'avantage est que ce test peut aussi être utilisé dans le cas des petits échantillons Gaussiens (même s'il est, dans ce cas précis, moins puissant que le test de Pearson). Le temps d'execution de cette commande est d'environ 30 secondes.

```
Cortest=function(X,Y)
{
  n<- length(X)
  T1 <- vector(mode = "numeric", length = 400000)
  for(i in 1:400000)
  {
    X1<-rnorm(n)
    Y1<-rnorm(n)
    R1 <- (X1-mean(X1))*(Y1-mean(Y1))
    T1[i]<-abs(sqrt(n)*((n-1)/n)*cov(X1,Y1)/sqrt(((n-1)/n)*var(R1)))
  }
  F<-ecdf(T1)
  R <- (X-mean(X))*(Y-mean(Y))
  T <-sqrt(n)*((n-1)/n)*cov(X,Y)/sqrt(((n-1)/n)*var(R))
  Pval<- 1-F(abs(T))
  return(Pval)
}
```

CKendall :
Correspond au test décrit en Section 3.
Attention : cette fonction ne tient pas compte des possibles ex-aequo.

```
CKendall=function(X,Y)
{
  n <- length(X)
  R <- array(0,dim=c(n,n))
  S <- array(0,dim=c(n,n))
  for(i in 1:n)
  {
    for(j in 1:n)
    {
      R[i,j]<-((X[j]-X[i])*(Y[j]-Y[i]))>0
      S[i,j]<-(X[j]>X[i]) & (Y[j]>Y[i])
    }
  }
  H <- vector(mode = "numeric", length = n)
  H <- 2*(apply(S,1,mean)+apply(S,2,mean))
  V <- var(H)
  T <- sqrt(n)*(sum(R)/(n*(n-1))-0.5)/sqrt(V)
  Pval <- 1-pnorm(abs(T))+pnorm(-abs(T))
  return(Pval)
}
```

CMW :
Correspond au test décrit en Section 4.
Attention : cette fonction ne tient pas compte des possibles ex-aequo.

```
CMW=function(X,Y)
{
  n <- length(X)
  m <- length(Y)
  R <- array(0,dim=c(n,m))
  for(i in 1:n)
  {
    for(j in 1:m)
    {
      R[i,j]<-(Y[j]>X[i])
    }
  }
  H <- vector(mode = "numeric", length = n)
  H <- apply(R,1,mean)
  G <- vector(mode = "numeric", length = m)
  G <- apply(R,2,mean)
  V <- var(H)/n + var(G)/m
  T <- (mean(R)-0.5)/sqrt(V)
  Pval <- 1-pnorm(abs(T))+pnorm(-abs(T))
  return(Pval)
}
```

VWelch :
Correspond au test décrit en Section 6.

```
VWelch=function(X,Y)
{
    X2 <- (X-mean(X))^2
    Y2 <- (Y-mean(Y))^2
    return(t.test(X2,Y2)$p.value)
}
```

Voneway :
Correspond au test décrit dans la remarque de la Section 6. Ici X est une variable quantitative, et F un facteur, c'est à dire une variable qualitative dont la $i$ème valeur indique à quel sous-échantillon appartient la variable X_i. Si F possède seulement deux modalités A et B, ce test est équivalent au test VWelch ci-dessus appliqué aux variables X_i du groupe A et aux variables X_j du groupe B.

```
Voneway=function(X,F)
{
  F<-as.factor(F)
  reg<-lm(X~F)
  return(oneway.test((reg$residuals)^2~F)$p.value)
}
```

ICMed :
Cette fonction donne l'intervalle de confiance asymptotique pour la médiane présenté en Section 7 (voir (7.1)). Le niveau de confiance asymptotique du test est noté `level`. Le test IC0 de niveau asymptotique 1- `level` est obtenu ainsi : rejet de H_0 si 0 n'appartient pas à ICMed (voir Section 7).

```
ICMed=function(D,level)
{
  n <- length(D)
  Y <- sort(D)
  return(c(Y[floor(-qnorm(0.5+level/2)*sqrt(n)/2 + n/2)],
          Y[floor(qnorm(0.5+level/2)*sqrt(n)/2 + n/2 + 0.5)]))
}
```

Med0 :
Correspond au test décrit en Section 7.

```
Med0=function(D)
{
    n <- length(D)
    z <- density(D, kernel="rectangular",n=2000)
    xab <- which.min(abs(z$x))
    T <- 2*(z$y[xab])*sqrt(n)*median(D)
    Pval <- 1-pnorm(abs(T))+pnorm(-abs(T))
    return(Pval)
}
```

CWTest :
Correspond au test décrit en Section 7.

```
CWTest=function(D)
{
  n <- length(D)
  R <- array(0,dim=c(n,n))
  Diag <- vector(mode = "numeric", length = n)
  for(i in 1:n)
  {
    for(j in 1:n)
    {
      R[i,j]<-(D[j]+D[i]>0)
      Diag[i]=R[i,i]
    }
  }
  H <- vector(mode = "numeric", length = n)
  H <- apply(R,1,mean)
  V <- var(H)
  T <- sqrt(n)*((sum(R)-sum(Diag))/(n*(n-1))-0.5)/(2*sqrt(V))
  Pval <- 1-pnorm(abs(T)) +pnorm(-abs(T))
  return(Pval)
}
```

Références

[1] M. S. Bartlett, (1937). *Properties of sufficiency ans statistical tests.* Proceeding of the Royal Statistical Society, **160** 268-282.

[2] G. E. P. Box, (1953). *Non-Normality and Tests on Variances.* Biometrika, **40** 318-335.

[3] M. B. Brown and A. B. Forsythe, (1974). *Robust Tests for the equality of Variances.* Journal of the American Statistical Association, **69** 364-367.

[4] W. J. Conover, (1971). *Practical Nonparametric Statistics.* Wiley & Sons, New York.

[5] S. E. Edgell and S. M. Noon, (1984). *Effect of violation of normality on the t test of the correlation coefficient.* Psychological Bulletin, **95** 576-583.

[6] R. A. Fisher, (1925). *Applications of "Student" Distribution.* Metron, **5** 90-104.

[7] R. A. Fisher, (1935). *The design of experiments.* Oliver & Boyd, Oxford, England.

[8] W. Hoeffding, (1948). *A class of statistics with asymptotically normal distribution* Ann. Math. Statistics, **19** 293-325.

[9] H. Hotelling and M. R. Pabst, (1936). *Rank correlation and test of significance involving no assumptions of normality.* Ann. Math. Statistics, **7** 29-43.

[10] G. S. James, (1951). *The Comparison of Several Groups of Observations When the Ratios of the Population Variances are Unknown.* Biometrika, **38** 324-329.

[11] M. Kendall, (1938). *A New Measure of Rank Correlation.* Biometrika, **30** 81-89.

[12] H. B. Mann and D. R. Whitney, (1947). *On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other.* Annals of Mathematical Statistics, **18** 50-60.

[13] A. W. van der Vaart, (1998) *Asymptotic statistics.* Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press, Cambridge, xvi+443 pp.

[14] B. L. Welch, (1947). *The generalization of "Student" problem when several different population variances are involved*, Biometrika, **34** 28-35.

[15] B. L. Welch (1951). *On the comparison of several mean values : an alternative approach*, Biometrika, **38** 330-336.

[16] F. Wilcoxon, (1945). *Individual comparisons by ranking methods.* Biometrics Bulletin, **1** 80-83.