Notes de cours Biostatistiques – MIV (L3)
Tests non paramétriques

M. Bailly-Bechet

Université Claude Bernard Lyon 1 – France

1 Intérêts de la variable rang

Dans certains cas, le test de la moyenne ne peut pas être employé. Typiquement, on a les cas où l'échantillon n'est pas distribué normalement, ou est trop petit pour que l'on puisse vérifier statistiquement que la distribution est normale, les tests de normalité acceptant toujours l'hypothèse nulle si on a trop peu de données. Un autre exemple est celui où les variables sont directement ordinales, ou classées qualitativement – et où l'on n'a pas de moyen de revenir à une variable quantitative. Dans ces cas là, on va associer à nos variables $x_i, i = 1..N$ leur rang $R_i = 1..N$, qui est leur position une fois les variables classées par ordre croissant. Voici un exemple où l'on donne les x_i et les R_i associés :

$$x_i = 2, 4, 1, -3, 7, 3 \tag{1}$$
$$R_i = 3, 5, 2, 1, 6, 4. \tag{2}$$

Que ce soit parce que ce sont les seules données dont on dispose, ou bien parce que l'on a choisi de travailler sur les rangs plutôt que sur les données de départ, on va s'intéresser aux rangs des variables plutôt quà leurs valeurs. Un exemple intuitif de l'intérêt de cette procédure consiste à comparer deux procédures de calculs des corrélations, celles de Pearson et de Spearman. On suppose que l'on dispose de 2 échantillons de taille n pour les variables X et Y. On note R_i les rangs des mesures x_i et S_i les rangs des mesures y_i. On définit alors les coefficients de corrélation comme suit :

$$r_{Pearson} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}, \tag{3}$$

et

$$r_{Spearman} = r_{Pearson}(R_i, S_i) = \frac{\sum_{i=1}^{n} (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n} (R_i - \bar{R})^2} \sqrt{\sum_{i=1}^{n} (S_i - \bar{S})^2}}. \tag{4}$$

Ces coefficients mesurent la relation entre les variables x et y ; plus précisément, le coefficient de corrélation de Pearson mesure l'existence d'une relation linéaire entre les deux variables. Le coefficient de Spearman, lui, mesure plus généralement l'existence d'une relation monotone : il suffit que les rangs soient identiques dans les 2 échantillons pour que le coefficient de corrélation vaille 1. Avec les notations ci-dessus, les coefficients de corrélation de Pearson et Spearman des variables $x_i = 1, 2, \ldots, 99, 100$ et $y_i = e^{3x_i}$ sont respectivement :

```
x <- 1:100
y <- exp(3 * x)
cor.test(x, y)
```

```
        Pearson's product-moment correlation

data:  x and y
t = 1.8225, df = 98, p-value = 0.07143
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.01592979  0.36451007
sample estimates:
      cor
0.1810548
```

```
cor.test(x, y, method = "spearman")
```

```
        Spearman's rank correlation rho

data:  x and y
S = 0, p-value < 2.2e-16
```

```
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
  1
```

On peut voir que les coefficients de corrélation sont très différents, et même que le coefficient de corrélation de Pearson n'est pas significativement différent de 0 dans ce cas, au risque de 5%. Ceci est normal : ce dernier coefficient mesure une relation linéaire, ce qui n'est pas du tout le cas ici. Mais cela veut aussi dire que la mesure de Spearman est plus propice à une recherche plus large, et moins sensible à des variations des données, par exemple à cause d'un bruit expérimental. Cette considération se généralise à l'étude des rangs comme variables statistiques : ils doivent être employés dans un premier temps quand la robustesse des résultats est recherchée, ou que le statisticien a peu d'informations sur la nature des données qu'il étudie.

1.1 Propriétés de la variable rang

On va travailler sur la variable rang. Pour un échantillon unique de n tirages x_i de la v.a. X, dont on ne connaît pas la loi, on a :

$$\sum R_i = \frac{n(n+1)}{2} \tag{5}$$

$$\sum R_i^2 = \frac{n(n+1)(2n+1)}{6} \tag{6}$$

On peut démontrer ces égalités de plusieurs manières. Pour 5, il suffit de remarquer que c'est une suite arithmétique simple, à savoir $1+2+3+\ldots+(n-1)+n$, et de remarquer dans la suite $u_n = 1, 2, 3, 4, \ldots$, la somme des termes $u_i + u_{n-i}$ est constante, et vaut $n+1$. On reformule alors la somme globale en paires constantes, et on doit diviser par deux car on compte chaque terme deux fois. Techniquement, cela donne :

$$\sum_i R_i = 1+ \quad 2+ \quad 3+ \quad \ldots+(n-1) \quad +n \tag{7}$$

$$\sum_i R_i = n+ \quad (n-1)+ \quad (n-2)+ \quad \ldots+2 \quad +1, \tag{8}$$

$$\tag{9}$$

ce qui, par addition des deux lignes, donne

$$2\sum_i R_i = (n+1) + (n+1) + (n+1) + \ldots + (n+1) + (n+1) \tag{10}$$

$$2\sum_i R_i = n(n+1) \tag{11}$$

$$\sum_i R_i = \frac{n(n+1)}{2} \tag{12}$$

On peut démontrer la formule 6 par récurrence à partir de la réponse. De manière plus générale, on peut faire le calcul suivant, se basant sur les propriétés des développements de $(a+b)^3$' :

$$\begin{aligned}
(n+1)^3 &= n^3 + 3n^2 + 3n + 1 && (13)\\
n^3 &= (n-1)^3 + 3(n-1)^2 + 3(n-1) + 1 && (14)\\
&\ldots && (15)\\
1^3 &= 0^3 + 3(0^2) + 3(0) + 1 && (16)
\end{aligned}$$

Par addition des termes en colonnes, si on note $S_{n^k} = \sum_{i=1}^{n} i^k$, on a :

$$S_{(n+1)^3} = S_{n^3} + 3S_{n^2} + 3S_n + (n+1) \tag{17}$$

En faisant passer le terme S_{n^3} à gauche, on a :

$$(n+1)^3 = 3S_{n^2} + 3\frac{n(n+1)}{2} + n + 1 \tag{18}$$

En développant et en transformant on obtient :

$$\begin{aligned}
3S_{n^2} &= n^3 + 3n^2 + 3n + 1 - n - 1 - \frac{3n^2 + 3n}{2}\\
&= n^3 - n + \frac{3n^2 + 3n}{2}\\
S_{n^2} &= \frac{1}{6}\left(2n^3 + 3n^2 + 3n - 2n\right)\\
&= \frac{n}{6}\left\{(2n+1)(n+1)\right\}
\end{aligned} \tag{19}$$

Grâce à ces égalités on peut montrer que, pour une v.a. X dont les rangs des tirages sont notés R_i :

$$\begin{aligned}\mathbb{E}(R) &= \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2} \\ \mathbb{V}(R) &= \frac{1}{n}\sum_{i=1}^{n} R_i^2 - \frac{(n+1)^2}{2} = \frac{n}{6}\left((2n+1)(n+1)\right) - \frac{(n+1)^2}{2} \\ &= \frac{n^2-1}{12}.\end{aligned} \tag{20}$$

Ces propriétés vont nous être utiles par la suite, pour développer un test non paramétrique de comparaison des moyennes.

2 Test de Wilcoxon et White-Manney

On va le plus souvent s'interésser aux rangs dans le cadre d'une comparaison d'échantillons. Dans ce cas on a 2 échantillons venant de v.a. X et Y, de tailles respectives n et m, que l'on va classer *ensemble*. On obtient donc un classement unique avec deux ensembles de rangs $R_i, i = 1..n$ et $S_j, j = 1..m$, tels que l'ensemble des rangs $\{R_i\} \cup \{S_i\} = 1..n+m$. L'idée des tests associés aux rangs est que, si les deux échantillons viennent de la même distribution, les rangs doivent être répartis de manière homogène dans les deux échantillons. Le test de comparaison des rangs a donc pour hypothèse nulle H_0 "Les deux échantillons viennent de distributions ayant la même moyenne", ce qui est logiquement équivalent au fait que les rangs R_i et S_j soient répartis de manière "homogène", aux flucutations près. Voici un exemple simple :

$$x_i = 1, 3, 5, 6, 8, 9 \tag{21}$$
$$y_i = -6, 2, 4, 7, 18 \tag{22}$$
$$x_i \cup y_i = -6, 1, 2, 3, 4, 5, 6, 7, 8, 9, 18 \tag{23}$$
$$R_i \cup S_i = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 \tag{24}$$

Ici on voit bien que les variables, une fois classées ensemble, alternent globalement, et donc que les moyennes de X et Y doivent être sensiblement les mêmes. Si la moyenne de Y était plus élevée que celle de X, on observerait un décalage du rouge vers la droite, et le contraire si la moyenne de Y était inférieure à celle de X.

On va calculer la statistique $\sum_{i=1}^{n} R_i$, la somme des rangs des n éléments x_i parmi les $m+n$ des deux échantillons. On va pour cela se servir des propriétés que l'on a vues plus haut. On a :

$$\mathbb{E}(\sum_{i=1}^{n} R_i) = \sum_{i=1}^{n} \mathbb{E}(R_i) = \frac{n(n+m+1)}{2} \tag{25}$$

$$\mathbb{V}(\sum_{i=1}^{n} R_i) = \sum_{i=1}^{n} \mathbb{V}(R_i) + \sum_i \sum_{j \neq i} cov(R_i, R_j) \tag{26}$$

Ici, $\mathbb{E}(R_i)$ vaut $\frac{n+m+1}{2}$ car le rang R_i peut prendre toute valeur entre 1 et $n+m$.

Un problème se pose pour calculer la variance, car les covariances des rangs ne sont pas nulles ; intuitivement, si un on sait que les rang de la ième variable est k, on a une informatio partielle sur le rang de la $i+1$ème, qui est forcément supérieur à k. Pour calculer ces covariances, on va d'abord étudier la somme des rangs totaux, S_{RT}. Pour cela on va définir la variable globale de rang $T_i, i = 1..n+m$, avec $T = R \cup S$. On a $S_{RT} = \sum_{i=1}^{n+m} T_i \frac{(n+m)(n+m+1)}{2}$. Cette variable est constante, on a donc $\mathbb{E}(S_{RT}) = \frac{(n+m)(n+m+1)}{2}$ et $\mathbb{V}(S_{RT}) = 0$. On va développer cette dernière égalité :

$$\begin{aligned} \mathbb{V}(S_{RT}) = 0 &= \sum_{i=1}^{n+m} \mathbb{V}(T_i) + \sum_{i=1}^{n+m} \sum_{\substack{j=1 \\ j \neq i}}^{n+m} cov(T_i, T_j) \\ 0 &= (n+m)\frac{(n+m)^2 - 1}{12} + (n+m)(n+m-1)cov(T_i, T_j) \\ cov(T_i, T_j) &= \frac{-((n+m)^2 - 1)}{12(n+m-1)} = \frac{-(n+m+1)}{12} \end{aligned} \tag{27}$$

Le passage à la dernière ligne ayant lieu car les covariances sont toutes égales et indépendantes de i, j. On peut le comprendre intuitivement en se disant que le fait de fixer un rang en particulier i ne fait que restreindre l'espace des possibles pour les autres rangs, créant une relation entre eux, mais n'ayant pas plus d'effet sur les rangs $j > i$ que sur les rangs $j < i$.

On peut alors reprendre le calcul précédent en remplaçant les covariances par leur expression, les covariances calculées sur les T_i étant les mêmes que celles calculées sur les R_i :

$$\mathbb{V}(\sum_{i=1}^{n} R_i) = n\frac{(n+m)^2-1}{12} - n(n-1)(\frac{(n+m+1)}{12}) \tag{28}$$

$$= \frac{1}{12}(n+m+1)(n(n+m-1) - n(n+1)) \tag{29}$$

$$= \frac{nm(n+m+1)}{12}. \tag{30}$$

Cette statistique est directement utilisée dans le test de Wilcoxon. Pour de petits échantillons, on peut calculer numériquement, par permutations, la probabilité pour un classement d'avoir une somme des rangs supérieure à n'importe quelle valeur – et dans ce cas le calcul de la moyenne et de la variance ne sont qu'indicatifs. Si les échantillons sont grands, on applique le théorème central limite à la variable $\sum_{i=1}^{n} R_i$, dont on connaît l'espérance et la variance, et on peut calculer une p-valeur, ou bien appliquer un test avec un risque connu à l'avance, en disant que la somme des n rangs suit une loi normale $\mathcal{N}(\mu = \frac{n(n+m+1)}{2}, \sigma^2 = \frac{(n+m+1)(nm)}{12})$. C'est le *test de Wilcoxon*.

Une variante que l'on observe souvent est de calculer la statistique U de White-Manney, qui est simplement une version centrée de la statistique de Wilcoxon W : $U = W - \mathbb{E}(W) = \sum_{i=1}^{n} R_i - \frac{n(n+m+1)}{2}$.

3 Fonction de répartition

3.1 Définitions et propriétés

La fonction de réparition d'une v.a. X représente la probabilité cumulée pour cette v.a. d'être inférieure à une valeur donnée x. Elle est définie sur l'ensemble du domaine de définition de la v.a. X. La fonction de répartition $P(x)$ d'une loi de probabilité continue de densité $p(x)$ est définie par les égalités suivantes :

$$P(x) = \int_{-\infty}^{x} p(z)dz \quad \text{ou} \quad \frac{dP}{dx} = p(x). \tag{31}$$

Pour une v.a. discrète, on a :

$$P(x) = \sum_{z \leq x} p(z). \tag{32}$$

En anglais, la fonction de répartition est notée *cdf*, pour "cumulated distribution function", tandis que la densité de probabilité est notée *pdf*, pour "probability distribution function". En francçais, on note souvent la fonction de répartition d'une variable par la lettre majuscule associée à sa densité de probabilité.

Les principales propriétés des fonctions de répartition découlent directement de ces définitions. Si X est une v.a. définie sur $[a, b]$, alors :

$$P(x) = 0 \text{ pour } x < a \tag{33}$$

$$P(x) = 1 \text{ pour } x \geq a. \tag{34}$$

Finalement, on peut ajouter que la connaissance de la fonction de répartition d'une v.a. donne autant d'information que la connaissance de sa densité de probabilité. On peut notamment dire que :

$$F(x) = G(x) \Leftrightarrow f(x) = g(x). \tag{35}$$

3.2 Fonction de répartition empirique d'un échantillon

Quand on dispose d'un échantillon $\{x_i\}$ de taille n d'une v.a. X, on peut définir la fonction de répartition empirique de l'échantillon, notée $P_n(x)$. Cette fonction, discontinue par nature, est théoriquement définie en classant les x_i par ordre croissant, et en prenant :

$$P_n(x_i \leq x < x_{i+1}) = \frac{i}{n}. \tag{36}$$

Dans le cadre des tests de comparaisons de fonctions de répartition, ce sont ces fonctions $P_n(x)$ qui vont être comparées entre elles ou à une fonction de répartition théorique. Sur la Fig 1 on peut voir un exemple de fonction de répartition empirique.

3.3 Exemples

L'exemple le plus connu de fonction de répartition est la fonction de répartition de la loi normale centrée réduite. Elle joue un rôle majeur en traitement du signal, à tel point qu'elle est nommée fonction d'erreur, ou en anglais *error function*, notée `erf`. Sur la Fig 2 on retrouve les principales propriétés évoquées plus haut, et il est aisé de faire le lien entre la pente de la fonction de répartition et la densité de probabilité.

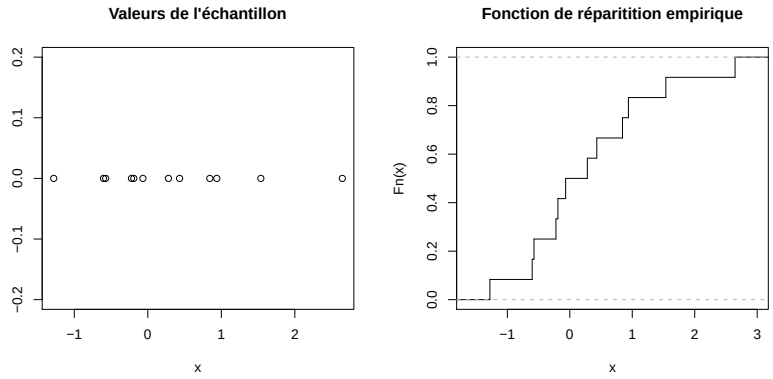

FIGURE 1 – 12 tirages d'une v.a. normale et sa fonction de répartition empirique.

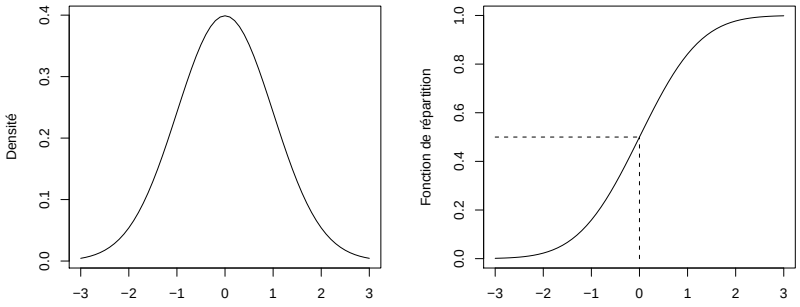

FIGURE 2 – Densité de probabilité et fonction de répartition de la loi normale.

Finalement, sur la Fig 3, on peut voir les densités de probabilité et les fonctions de répartition pour d'autres fonctions : uniforme, piquée, et finalement deux exemples de fonctions asymétriques.

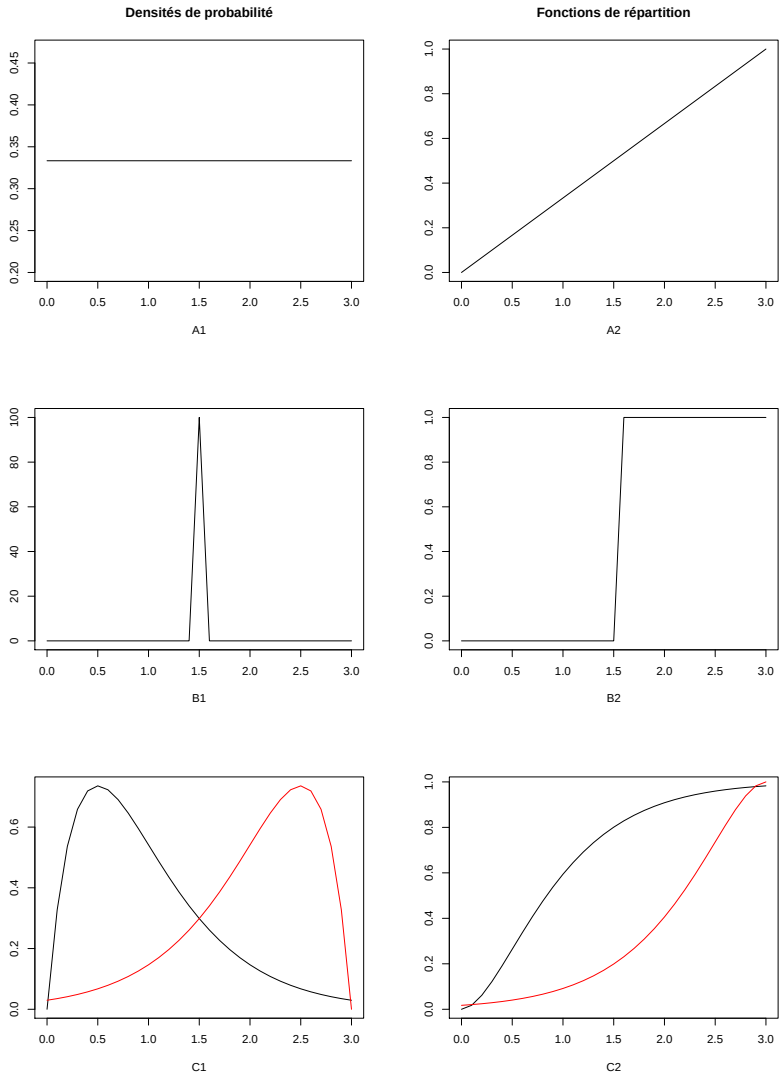

FIGURE 3 – Loi uniforme (A), piquée (B) et deux lois Gamma avec des paramètres différents (C). Dans le dernier cas on voit bien comment l'asymétrie de la densité se retrouve dans la fonction de répartition.

4 Tests basés sur la comparaison des fonctions de répartition

L'intérêt des tests basés sur la comparaison de fonctions de répartition (que ce soit la comparaison de deux échantillons ou bien d'un échantillon à une fonction de répartition théorique) est d'utiliser au mieux toute l'information présente dans l'échantillon, et de ne pas en perdre en la résumant à quelques valeurs comme la moyenne ou lécart-type. Ceci n'a de sens que si l'on ne peut pas faire d'hypothèses sur la nature de la fonction de répartition, et donc de la densité de probabilité sous-jacente à l'échantillon : on est donc bien dans le domaine des tests non paramétriques. Un des intérêts en particulier est le domaine de l'inférence statistique sur des données très bruitées : dans ce cadre, l'usage des fonctions de répartition permet de "lisser" le bruit et de minimiser son impact sur les analyses. Finalement, la comparaison de la fonction de répartition empirique d'un échantillon à une fonction de répartition théorique est une façon de faire très adaptable, et qui permet de poser des questions extrêmement variées en utilisant le même formalisme.

L'idée générale des tests de comparaison de fonctions de répartition consiste à définir une "distance" entre les deux fonctions de répartition. Ceci peut être fait de très nombreuses manières, et il existe en conséquence de très nombreux tests ce comparaison, plus ou moins complexes.

4.1 Test de Cramér-von Mises (CVM)

Ce test est basé sur la distance euclidienne entre fonctions. Soit un échantillon $\{x_i\}$ de taille n. On note $F_n(x)$ sa fonction de répartition empirique, que l'on veut comparer à une fonction de répartition connue à l'avance $F(x)$. On calcule la statistique suivante :

$$\omega^2 = \int_{-\infty}^{+\infty} \left(F(x) - F_n(x)\right)^2 dF \tag{37}$$

Une approximation utile de cette statistique, qui permet de ne pas avoir besoin de calculer explicitement l'intégrale et donne une formule analytique, est :

$$T = n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F(x_i) \right) \tag{38}$$

On peut relier cette formule à la définition précédente 37 en voyant que, pour des x_i ordonnés, on a remplacé $F_n(x) = \frac{i}{n}$ par $F_n(x) = \frac{i-\frac{1}{2}}{n}$, pour des questions de symétrie (les rangs effectifs vont alors non pas de 1 à n mais de 0.5 à $n - \frac{1}{n}$).

Dans le cas de la comparaison de deux fonctions de répartition empiriques (pour répondre à la question "Ces deux échantillons proviennent-ils de la même population" ?), on peut employer l'approximation suivante :

$$T = n\omega^2 = \frac{U}{nm(n+m)} - \frac{4mn-1}{6(m+n)}, \tag{39}$$

avec

$$U = n\sum_{i=1}^{n}(r_i - i)^2 + m\sum_{j=1}^{m}(s_j - j)^2, \tag{40}$$

où les r_i et les s_j sont respectivement les rangs des variables $\{x_i\}$ et $\{y_j\}$ dans l'échantillon combiné de taille $n+m$. Dans la pratique, ces calculs sont implémentés dans R par la commande `cvm.test`.

4.2 Test de Kolmogorov-Smirnov (KS)

Ce test est basé sur la distance maximale entre deux fonctions de répartition. Pour le même échantillon que précédemment, on définit D_n comme la distance maximale entre $F_n(x)$ et $F(x)$:

$$D_n = \max(\|F_n(x) - F(x)\|). \tag{41}$$

Cette distance, de par les propriétés des fonctions de répartition, est comprise entre 0 si les deux fonctions sont identiques en tous points (ce qui n'est possible que pour des n théoriquement infinis, car F_n est discontinue), et 1 si l'ensemble des valeurs observées dans l'échantillon est en dehors (et du même coté) du domaine de définition de F. La probabilité d'observer une valeur de D_n donnée, c-à-d. la p-value associée à une valeur de D_n, est donnée par :

$$p\,(z) = 1 - 2\sum_{\nu=1}^{\infty}(-1)^{\nu-1}e^{-\nu^2 z^2} \quad z = \frac{D_n}{\sqrt{n}} \tag{42}$$

Cette p-valeur peut se calculer asymptotiquement numériquement.

Si l'on veut comparer deux échantillons de taille n et m, auquel cas l'hypothèse nulle est "Les deux échantillons proviennent de la même population", alors on calcule $D_{m,n}$ commme :

$$D_{m,n} = \max(\|F_n(x) - F_m(x)\|), \tag{43}$$

et la p-value se calcule comme précedemment, mais avec :

$$z = \frac{D_{m,n}}{\sqrt{\frac{nm}{n+m}}}. \tag{44}$$

La commande pour ce test est `ks.test`.

4.3 Tests de normalité

Les tests de normalité sont une sous-catégorie des tests basés sur les fonctions de répartition. Leur hypothèse nulle est : "L'échantillon testé provient d'une loi normale". Si en théorie on pourrait employer les tests CVM ou KS pour poser cette question, il a été démontré que des tests *ad hoc* sont plus puissants. En particulier, le test de Shapiro-Wilks est très employé. Il consiste à calculer la statistique suivante :

$$W = \frac{T^2}{ns^2} \quad T = \left(\sum_{i=1}^{n} a_i \left(x_{n+i-1} - x_i\right)\right)^2 \tag{45}$$

Les coefficients a_i sont calculés à partir de la matrice de covariance des données attendues pour un échantillon de taille n ; cela signifie que les coefficients a_i ne dépendent que de n, et pas des données elles-mêmes. s^2 est la variance observée $\sum_{i=1}^{n}(x_i - \bar{x})^2$. On peut voir que T^2 et σ^2 vont être globalement proportionnels, puisque σ^2 est la variance et T^2 se construit à l'aide des carrés des étendues $x_{n+i-1} - x_i$. Si l'échantillon suit une loi normale, les étendues théoriques sont connues, et prsentent l'avantage de moins fluctuer que les valeurs maximales et minimales prises indépendamment. Alors $W \to 1$. Mais si l'échantillon n'est pas normalement distribué, les étendues vont être disproportionnées avec les a_i correspondants, conduisant à des valeurs de W trop faibles qui pourront être interprétées comme une non-normalité de l'échantillon testé. La loi suivie par W n'est pas analytiquement connue, et il faut calculer numériquement la p-value, comme pour les autres tests non paramétriques.

Le test de Shapiro-Wilks (commande `shapiro.test`) n'est pas le seul test de normalité utilisable, mais c'est le plus recommandé. D'autres tests existent : Anderson-Darling, ... En particulier, pour avoir une idée de test de normalité non baśe sur les fonctions de répartition, on peut se reporter au test de Jarque-Bera, présenté dans la section "Moments et fonction génératrice" de ce polycopié [1].

1. Cette présentation sera écrite ultérieurement.