2020 ■ Vol. 16 ■ no. 1

La somme des rangs de Wilcoxon, ses moments, sa distribution nulle, avec illustration et compléments

Louis Laurencelle [a]

[a]Université du Québec à Trois-Rivières

Abstract ■ This paper delves into Wilcoxon's sum-of-ranks R statistic (and Mann-Whitney's translated equivalent U) for testing the difference between two groups. The first four raw and central moments and the γ_1 and γ_2 shape indices of the statistic are given, along with a general method for working them out. As for R's null distribution, a fair sample of published tables is offered, plus two continuous (Beta, Normal) approximations. We provide a detailed procedure for establishing individual probabilities of R using a recursive function of the partition of an integer. A worked out example serves to illustrate the mathematical contents. In a conclusive section, we examine the statistical power of the R (and U) test against Student's t reference test, with a somewhat advantageous recommendation for the former. // La statistique R (ou W) attribuée à Wilcoxon (1945), soit la somme des rangs associés aux k observations du groupe 1, $R = r_1 + r_2 + \ldots + r_k$, permet de décider si le niveau des données diffère entre les k observations du groupe 1 et les $n - k$ observations du groupe 2, ce au moyen d'un test non-paramétrique. Algébriquement équivalente au U de Mann et Whitney (1947), elle supplée au test t de la différence entre deux moyennes indépendantes sans reposer sur la condition de normalité des observations. Nous mettons d'abord en place un exemple fictif, avec sa solution classique par le test t de Student. Nous présentons ensuite une procédure d'intérêt général pour établir les moments à l'origine et les moments centraux de R et de U, ce qui nous permet de présenter une expression algébrique simple de ses 3^e et 4^e moments et, grâce à eux, d'offrir une modélisation de la distribution de R et U par la loi Bêta symétrique. Enfin, nous nous intéressons à la distribution nulle de R, ce grâce à une variante originale de la méthode de partition d'un entier : la parenté entre l'élément $f_{n,k}(R)$ de la distribution de fréquences de R et le nombre de partitions d'un entier est étudiée. Nous proposons une fonction $Q(S, k, x)$, à définition récursive, qui dénombre les k partitions de l'entier S formées d'éléments si tels que $1 \leq s_i \leq x$. Posant $S = R - {}^1/_2 k(k-1)$, $k = k$ et $x = n - k + 1$, la fonction Q fournit directement la fréquence $f_{n,k}(R)$. Les calculs (moments, probabilité exacte, modélisation Bêta, approximation normale) sont alors appliqués à l'exemple donné en illustration. L'exposé se conclut par des considérations sur la puissance statistique du test de Wilcoxon, sous différentes conditions.

Keywords ■ Non-parametric tests; Wilcoxon test; rank test. **Tools** ■ Excel.

louis.laurencelle@gmail.com

10.20982/tqmp.16.1.p046

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Ce texte est une version augmentée et remaniée d'un autre texte du même auteur, voir Laurencelle (2005).

Introduction

Posons les groupes G_1 et G_2, de tailles respectives $n_1 = 4$ et $n_2 = 6$ ($n = 10$), avec les données :

G_1 :	30,5	42,6	37,4	32,8		
G_2 :	24,9	37,0	30,9	27,5	24,8	31,6

La question posée : "Est-ce que les éléments du groupe 1 proviennent d'une population à valeurs plus fortes ou moins fortes que celles de la population des éléments du groupe 2?", a pour pendant le modèle de l'hypothèse nulle (H_0), à savoir : "Les éléments et données des deux groupes sont issus de la même population". La procédure de rou-

tine pour vérifier H_0 est d'appliquer le test t de Student pour la différence entre deux moyennes indépendantes, soit :

$$t_{\bar{X}_1-\bar{X}_2} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \times \left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad (1)$$

Les calculs, basés sur les moyennes $\bar{X}_1 = 35,825$ et $\bar{X}_2 = 29,450$ et les écarts-types $s_1 = 5,351$ et $s_2 = 4,687$ produisent $t \approx 1,997$. Selon ses $n_1 + n_2 - 2 = 4 + 6 - 2 = 8$ degrés de liberté, la probabilité extrême (unilatérale) de ce test est de $\approx 0,040$ et, doublée en mode bilatéral, de $\approx 0,081$, la valeur critique au seuil bilatéral $\alpha = 0,05$ étant de 2,306. La preuve ne nous permet pas de rejeter H_0, et nous pouvons considérer que, pour cette mesure, les deux groupes proviennent vraisemblablement de la même population.

Et pourquoi d'autres tests? La validité du test t appliqué ci-dessus à notre question repose essentiellement sur deux conditions : (a) chaque groupe de données doit provenir d'une population homogène à variation (à peu près) normale, et (b) les données des deux populations doivent avoir mêmes variances (condition d'homoscédasticité). Sauf sous une non-normalité sérieuse, des tailles n_1 et n_2 petites ou un ratio n_1/n_2 loin de 1, Kendall et Stuart (1979) soutiennent la robuste validité du test. Il reste que, en termes de puissance, si le t de Student est optimal, c.-à-d. le plus puissant sous des conditions parfaites, il est vite rejoint par le test de Wilcoxon, appelé aussi "test de la somme des rangs", voire il est dépassé par lui lorsque ses conditions d'optimalité s'avèrent sérieusement dégradées.

La somme de rangs (R) de Wilcoxon et le U de Mann-Whitney

C'est en 1945 que Wilcoxon proposa sa fameuse procédure basée sur une somme de rangs afin de tester la différence de niveau entre deux groupes d'observations. Kruskal (1957) en retrace des précurseurs jusqu'en 1914, la statistique testée ayant pris différentes formes. La procédure commune consiste à substituer aux $n = n_1 + n_2$ données originales des deux groupes leurs rangs correspondants, de 1 à n.

Deux statistiques ont survécu, le R de Wilcoxon (1945), soit la somme des rangs du plus petit groupe, R, et le U de Mann et Whitney (1947), lequel est une transformation linéaire du R, soit $U_1 = R_1 - n_1 \cdot (n_1 + 1)/2$ ou $U_2 = R_2 - n_2 \cdot (n_2 + 1)/2$, où n_j et R_j sont respectivement la taille et la somme des rangs du groupe j.[1] Nous traiterons prioritairement la statistique R. Noter que, pour des fins de simplicité, la taille du plus petit groupe sera dénotée k, l'autre groupe ayant la taille $n - k$.

Soit les données du groupe 1: $x_1, x_2, \ldots, x_k$, et celles du groupe 2: $y_{k+1}, y_{k+2}, \ldots, y_n$, $n-k \leq k$. La procédure consiste en trois étapes : d'abord, les n observations doivent être converties en rangs, depuis 1 jusqu'à n, la valeur minimale se voyant attribuer le rang 1, la suivante le rang 2, etc. L'on calcule ensuite $R_{obs} = r(x_1)+r(x_2)+\cdots+r(x_k)$, la somme des rangs associés aux observations du petit groupe ; enfin, on vérifie si la valeur observée (R_{obs}) occupe un rang centile suffisamment extrême dans la distribution nulle de R, auquel cas la différence entre les groupes est déclarée significative au seuil de probabilité choisi.

Note sur les rangs liés. L'attribution des rangs, sous forme de nombres entiers de 1 à n, peut occasionnellement se heurter à une situation de valeurs observées égales, donnant lieu au phénomène dit des "rangs liés" : Kendall et Stuart (1979, p. 537) discutent du problème. Dans le cas d'égalités confinées en un seul groupe, la solution (correcte et exacte) consiste à attribuer les rangs que les valeurs concernées obtiendraient si elles différaient légèrement l'une de l'autre. Dans le cas d'égalités qui enjambent les deux groupes, la solution "Wilcoxon" exacte est impossible. Les solutions palliatives, par l'approximation Bêta ou la normale (voir plus loin), conviennent cependant, au prix de deux interventions : l'attribution à chaque valeur d'un ensemble d'égalité la moyenne des rangs qu'elles auraient obtenus si légèrement différentes, de même qu'une correction appropriée de la variance de R; voir Siegel et Castellan, 1988.

La distribution nulle de R

La distribution nulle de R est constituée de l'ensemble ordonné des valeurs possibles de la statistique résultant de toutes les combinaisons des rangs entre les groupes 1 et 2. Il s'agit donc d'un test permutationnel, aussi nommé " randomisation test " en anglais (Siegel, 1956; Edgington, 1980). La taille de l'ensemble permutationnel correspond au nombre de combinaisons différentes de n objets pris k à la fois, soit $\binom{n}{k}$. Pour obtenir la distribution nulle de R, la méthode directe consiste à énumérer chaque k-groupe de rangs $r_1, r_2, \ldots, r_k$ formable à partir des entiers $1, 2, \ldots, n$, à calculer $R = r_1 + r_2 + \cdots + r_k$, puis à l'incorporer dans la distribution de fréquences, soit $f(R) \leftarrow f(R) + 1/\binom{n}{k}$. Étant donné la correspondance linéaire $U = R - 1/2k(k+1)$ indiquée plus haut, les distributions de probabilités de U et de R sont identiques, à

[1] La documentation présente plutôt la statistique complémentaire $U' = k \cdot (n-k) - U = 1/2k \cdot (2n-k+1) - R$; ainsi, les distributions de U et U' sont symétriques et contrevariantes. Pour les intéressés, signalons que $q = PrX_1 \geq X_2 \approx U_1/(n_1 \cdot n_2) = U/[(k \cdot (n-k)]$, cette expression, de variance proche de $q \cdot (1-q)/n_1$, estimant la probabilité qu'un élément de la population 1 présente une valeur supérieure à un élément de la population 2 (Bradley, 1968, p. 114).

un décalage de la variable près. Mann et Whitney (1947; voir Siegel 1956 et Owen 1962) présentent une fonction récurrente pour déterminer les fréquences cumulatives de U. D'autres fonctions sont aussi possibles (Fix et Hodges 1955; Kendall et Stuart, 1979, p. 521-522). Wilcoxon (1945) fait le lien entre les fréquences de R et les partitions d'un entier. Nous retournons à ces approches plus loin.

Diverses tables utiles pour vérifier la significativité des tests R (ou W) et U sont disponibles, parmi lesquelles nous proposons les suivantes : la mention "Cum" indique qu'il s'agit de probabilités cumulatives, "VC" qu'il s'agit de valeurs critiques :

- Siegel et Castellan (1088), Table I (R), Cum : $m(3-4)$, $n(m-12)$; $m(5-10)$, $n(m-10)$;
- Bradley (1968), Table III (R), VC [0,001, 0,005, 0,01, 0,025, 0,05, 0,10]: $n(1-25)$, $m(n-25)$;
- Siegel (1956), Table J (U) et Mann et Whitney (1947), Table 1, Cum : $n_2(3-8)$, $n_1(1-n_2)$;
- Siegel (1956), Table K (U), VC : $n_2(9-20)$, $n_1(1-20)$;
- Fix et Hodges (1955), Tables 1 et 2 (U), Cum : $m(2-12)$, $n>2$ (calculs à effectuer).
- Rohlf et Sokal (1995), Table 29 (U), VC : $n_1(3-20)$, $n_2(1-n_1)$.

Nous identifierons d'abord les moments à l'origine et les moments centraux de la somme de rangs R, puis nous proposerons deux modélisations de la distribution de R, basées sur ces moments.

Les moments de R

Les premiers moments de R sont connus (Siegel 1956), cependant on ne retrouve nulle part le quatrième moment central ni l'indice d'aplatissement, non plus que l'ensemble des moments à l'origine, ou moments simples, de R: nous en proposons ici un nouveau calcul. Définissons d'abord moments simples (m_s) et moments centraux (μ_s) d'ordre s de la variable R, selon:

$$m_s(R) = \varepsilon(R^s) \tag{2a}$$

$$\mu_s(R) = \varepsilon[(R - m_1(R))^s], \tag{2b}$$

où "ε" dénote ici l'espérance à travers l'ensemble permutationnel de R. Les moments centraux peuvent être obtenus à partir des moments simples (Kendall et Stuart, 1977).

Le multinôme $R_s = (x+y+z+\dots)^s$, comportant k éléments, engendre k^s termes qu'on peut regrouper par homologie, et il y a $p(s)$ catégories d'homologues dans le développement de R^s, $p(s)$ étant déjà le nombre de partitions de l'entier s. Étant donné que, pour la distribution nulle de R, l'échantillonnage se fait sans remise dans l'ensemble fini $E = 1,2,3,\dots,n$, chaque homologue a une espérance calculable. À son tour, chaque catégorie d'homologues apparaît en un nombre déterminé d'exemplaires dans R^s. Posons les espérances "v" pour chaque catégorie d'homologues, telles que $v_1 = \varepsilon x$, $v_2 = \varepsilon x^2$, $v_{11} = \varepsilon xy$, etc. En remplaçant les homologues par leur espérance commune et en multipliant cette espérance par le nombre correspondant d'exemplaires, on obtient le moment m_s, d'où on peut tirer ensuite $\mu = m_1$, $\sigma^2 = \mu_2 = m_2 - \mu^2$, $\mu_3 = m_3 - 3\mu \cdot m_2 + 2\mu^3$, $\mu_4 = m_4 - 4\mu \cdot m_3 + 6\mu^2 m_2 - 3\mu^4$, puis les indices de forme: $\gamma_1 = \mu_3/\sigma^3$ et $\gamma_2 = \mu_4/\sigma^4 - 3$.

Prenons l'exemple simple de $k=3$, $s=2$ et $R_3 = r_1 + r_2 + r_3 = x+y+z$. Nous avons alors $R^2 = (x+y+z)^2 = x^2+xy+xz+yx+y^2+yz+xz+yz+z^2$, soit, en regroupant les termes homologues, $(x^2+y^2+z^2) + 2(xy+xz+yz)$. Substituant aux homologues leurs espérances respectives, nous avons $m_2 = \varepsilon(R^2) = (v^2+v^2+v^2) + 2(v_{11}+v_{11}+v_{11})$, où $v_2 = \varepsilon(r^2)$ et $v_{11} = \varepsilon(r_i r_j)$, et obtenons $m_2 = 3v_2 + 6v_{11}$. En général pour tout k, nous avons:

$$\begin{aligned}
m_1 &= kv_1 \\
m_2 &= kv_2 + k\cdot(k-1)v_{11} \\
m_3 &= kv_3 + 3k\cdot(k-1)v_{21} + k\cdot(k-1)(k-2)v_{111} \\
m_4 &= kv_4 + 4k\cdot(k-1)v_{31} + 3k\cdot(k-1)v_{22} \\
&\quad + 6k\cdot(k-1)(k-2)v_{211} \\
&\quad + k\cdot(k-1)(k-2)(k-3)v_{1111}
\end{aligned} \tag{3}$$

Dans cette écriture, "v_{21}" désigne toute expression réductible à la forme x^2y, incluant donc xy^2 : il en va de même pour v_{31} et v_{211}. Par exemple, pour l'élément v_{21} dans m_3, les termes de forme x^2y et xy^2 apparaissent chacun avec le coefficient binomial "3"; ils sont obtenus par association de deux valeurs r_i et r_j distinctes, chacune portant à son tour l'exposant 2: leur nombre étant $\binom{k}{2}$, les variantes sont au nombre de $2\binom{k}{2}$, et le coefficient final de v_{21} est $3 \times 2\binom{2}{2} = 3k\cdot(k-1)$.

Pour achever la solution, il reste à trouver les espérances "v". Notons que, dès que l'espérance comporte plus d'un élément dans $E = 1,2,3,\dots,n$, l'évaluation doit tenir compte qu'on y pratique l'échantillonnage sans remise. En voici la liste:

$$\begin{aligned}
v_1 &= (n+1)/2(4) \\
v_2 &= (n+1)(2n+1)/6 \\
v_{11} &= (n+1)(3n+2)/12 \\
v_3 &= n(n+1)^2/4 \\
v_{21} &= n(n+1)^2/6 \\
v_{111} &= n(n+1)^2/8 \\
v_4 &= (n+1)(2n+1)(3n2+3n-1)/30 \\
v_{31} &= (n+1)(15n^3+21n^2-4)/120 \\
v_{22} &= (n+1)(2n+1)(10n^2+7n-6)/180 \\
v_{211} &= (n+1)(30n^3+35n^2-11n-12)/360 \\
v_{1111} &= (n+1)(15n^3+15n^2-10n-8)/240
\end{aligned} \tag{4}$$

Par exemple, désignant par Σ_1 la somme des n premiers naturels, par Σ_2 la somme de leurs carrés, etc., on a $v_p = \Sigma_p/n$. Quant à v_{11}, considérons que le produit $E \times E$ contient les éléments appropriés "xy" ($x \neq y$), en plus des éléments diagonaux "x^2 " ($x = y$), qui sont à rejeter, d'où $v_{11} = (\Sigma_1^2 - \Sigma_2)/(n_2 - n)$. Ces espérances-ci (ou les sommes correspondantes) se retrouvent aussi dans David et Johnson (1954, en appendice).

Les moments simples de $R = R(n,k)$, la somme de rangs de Wilcoxon, sont donc :

$$\begin{aligned} m_1 =& \varepsilon(R) = \mu_R = k \cdot (n+1)/2 \\ m_2 =& k \cdot (n+1)(3k \cdot n + 2k + n)/12 \\ m_3 =& n \cdot (n+1)^2 \cdot (k+1) \cdot k^2/8 \\ m_4 =& k \cdot (n+1)/240 \times \\ & [n^3(15k^3 + 30k^2 + 5k - 2) \\ & + n^2(15k^3 + 50k^2 + 9k - 2) \\ & - 2nk \cdot (5k^2 - 8k - 1) - 8k^3], \end{aligned} \tag{5}$$

et les moments centraux :

$$\begin{aligned} \sigma_R^2 =& \mu_2 = k(n-k)(n+1)/12 \\ \mu_3 =& 0 \\ \mu_4 =& k(n-k)(n+1) \times \\ & [n^2(5k-2) - n(5k^2 - 7k + 2) - 7k^2]/240. \end{aligned} \tag{6}$$

La statistique R a pour minimum $1/2k \cdot (k+1)$, pour maximum $1/2k \cdot (2n - k + 1)$ et pour étendue $k \cdot (n-k)$. Les indices de forme ($\gamma_1 = \mu_3/\sigma^3$; $\gamma_2 = \mu_4/\sigma^4 - 3$) sont alors :

$$\gamma_1 = 0 \tag{7a}$$

$$\gamma_2 = \frac{-6}{5} \frac{\left(n^2 - n\,(k-1) + k^2\right)}{k\,(n-k)\,(n+1)}, \tag{7b}$$

indiquant une distribution symétrique et légèrement platykurtique. La statistique U présente les mêmes moment que F, sauf le premier, $\varepsilon(U)$, qui est simplement :

$$\mu_U = k \cdot (n-k)/2 \quad \text{ou} \quad n_1 \cdot n_2/2. \tag{8}$$

Modélisation de la distribution nulle de R

La distribution nulle de R est symétrique ($\gamma_1 = 0$) et, bien que discrète, elle peut être modélisée par une loi symétrique continue, telle la loi Bêta symétrique $\beta(p,p)$ (Laurencelle, 2001), qui permet d'en respecter l'indice d'aplatissement, ou par la loi normale, en supposant l'indice d'aplatissement proche de 0.

La variable x_β de loi Bêta $\beta(p,p)$ a pour espérance $\mu_\beta = 1/2$, pour variance $\sigma_\beta^2 = 1/(8p+4)$ et pour indices de forme $\gamma_1 = 0$ et $\gamma_2 = -6/(2p+3)$. En égalisant l'indice d'aplatissement γ_2 à celui de R (éq. 7b) nous déterminons la valeur appropriée du paramètre p, soit :

$$p = \frac{(5n+8) \cdot k(n-k) - 3n(n+1)}{2(n^2 + n - kn + k^2)}; \tag{9}$$

la variable x_β étant ainsi construite, la quantité $R' = (x_\beta - 1/2)/\sigma_\beta \times \sigma_R + \mu_R$ restitue approximativement la distribution de R, à sa discontinuité près. Pour effectuer le test, on calcule d'abord :

$$x_\beta = 1/2 + \frac{R - \mu_R \pm 1/2}{\sigma_R} \times \sigma_\beta, \tag{10}$$

pour consulter ensuite la fonction de répartition de la loi Bêta avec paramètres p et p : le logiciel Excel la fournit.

Le paramètre p ci-dessus est quasi proportionnel à n, et l'indice négatif γ_2, inversement proportionnel à p. Par conséquent, pour des valeurs de n assez élevées ($n \geq 50$) et bien réparties d'un groupe à l'autre ($k \approx 1/2n$), l'indice γ_2 est presque nul et le modèle normal convient à peu près : déjà, pour $n = 50$, $k = 25$, nous avons $\gamma_2 \approx -0,072$. Dans un tel cas, la statistique $(R - \mu_R \pm 1/2)/\sigma_R$ se distribue approximativement selon une normale standard : par exemple, le 95e centile de R serait $R_{95} \approx \mu_R + 1,645 \cdot \sigma_R + 0,5$, en incorporant une correction de continuité.

Fréquences de R et partitions d'un entier

Nous traduirons à présent le problème de la distribution nulle de R, soit la fréquence $f(R)$ associée à chaque valeur possible de la somme de k rangs R, dans le langage des partitions d'un entier. En plus d'établir de belles symétries entre les deux domaines, cette traduction fournit un nouveau moyen pour produire directement la distribution nulle, sans avoir à énumérer les $\binom{n}{k}$ combinaisons de rangs qui y sont incluses.

Soit une valeur R et un sous-ensemble j de k rangs $\{r_1, r_2, \ldots, r_k\}_j$ qui lui est associé, selon $R = r_1 + r_2 + \cdots + r_k$. Sans perdre de généralité, nous pouvons placer les k valeurs de rangs du sous-ensemble j en ordre décroissant, puis les recoder selon :

$$s_i \leftarrow r_i - k + i, \tag{11}$$

ce qui produit un nouveau sous-ensemble correspondant j' de nouveaux éléments "s", $\{s_1, s_2, \ldots, s_k\}_j$, du domaine des entiers 1 à $n - k + 1$. La somme associée à ce sous-ensemble, $S = s_1 + s_2 + \cdots + s_k$, est visiblement égale à:

$$S = R - 1/2k \cdot (k-1); \tag{12}$$

la correspondance entre les éléments "r" et "s" est bijective. En outre, chaque sous-ensemble d'éléments "s" est une partition, en fait une k-partition de l'entier S.

Table 1 ■ Exemple 1

x	30,5	42,6	37,4	32,8	24,9	37,0	30,9	27,5	24,8	31,6
rang	4	10	9	7	2	8	5	3	1	6

Note. Les valeurs à gauche de la barre centrale sont celles de G_1 alors que celles à droite sont celles de G_2.

Bose et Manvel (1984) donnent un aperçu de la théorie des partitions. Soit $p(S)$, le nombre de partitions de l'entier S. Par exemple, l'entier 5 donne lieu aux partitions suivantes :

$$(5), (41), (32), (311), (221), (2111), (11111),$$

d'où $p(5) = 7$. De plus, les partitions de S peuvent se classer selon le nombre de "morceaux" qu'elles comportent, ce qui donne lieu à l'égalité :

$$p(S) = p_1(S) + p_2(S) + ... + p_S(S) \tag{13}$$

qui relie les partitions d'un entier à l'ensemble de ses k-partitions. Ainsi, (311) et (221) sont deux 3-partitions de l'entier 5, et

$$\begin{aligned} p(5) =& p_1(5) + p_2(5) + p_3(5) + p_4(5) + p_5(5) \\ =& \ 1 \ + \ 2 \ + \ 2 \ + \ 1 \ + \ 1 \ = 7. \end{aligned}$$

On peut montrer que:

$$p_k(S) = p_1(S-k) + p_2(S-k) + \cdots + p_k(S-k), \tag{14}$$

c'est-à-dire que les k-partitions sont calculables récursivement. En effet, soit une k-partition $\lambda_k(S)$ composée de k éléments $s_1 s_2 \ldots s_k$, $s_i > 0$, $\Sigma s_i = S$. Les "s_i" sont arrangés en ordre de valeurs non croissantes, soit $s_1 \geq s2 \geq \cdots \geq s_k$. Recodant chaque "$s_i$" en "$q_i$" selon :

$$q_i = s_i - 1,$$

alors $S' = S - k = \Sigma q_i$ est un entier réduit. Les partitions $p(S') = p(S-k)$ comportent des j-partitions admissibles de $1, 2, \ldots, k$ éléments (en ignorant au besoin des éléments nuls, $q_i = 0$, à droite de la partition), mais elles peuvent comporter aussi des j-partitions trop longues ($j > k$) qu'il faut exclure, d'où l'on obtient:

$$\begin{aligned} p_k(S) =& p(S-k) - p_{k+1}(S-k) - \\ & p_{k+2}(S-k) - \cdots - p_{S-k}(S-k). \end{aligned} \tag{15}$$

L'application de l'égalité (11) complète la preuve du théorème (14). Notons enfin, pour faciliter le calcul récursif de $p_k(S)$, que $p_1(a) = p_a(a) = 1$ pour $a > 0$ et $p_b(a) = 0$ si $a < b$.

Même si le sous-ensemble $\lambda_k(S) = s_1, s_2, \ldots, s_k$, défini par (11) plus haut, est une k-partition de S, toutes les k-partitions de S ne sont pas admissibles car elles ne correspondent pas toutes à des sous-ensembles de rangs de 1 à n : les k-partitions contenant des éléments "s_i" $> n-k+1$ sont impossibles car ces éléments correspondraient à des rangs $r_i > n$. Il faut donc trouver les k partitions plafonnées $\lambda_k(S, x) = s_1 \leq x, s_2 \leq x, \ldots, s_k \leq x$, ou à tout le moins trouver leur nombre $Q(S, k, x)$, où $x = n-k+1$.

La fonction $Q(S, k, x)$ dénombre les partitions de l'entier S en k morceaux, les morceaux étant bornés dans l'intervalle 1 à x. Nous montrons l'égalité récursive:

$$\begin{aligned} Q(S, k, x) = pk(S) &- Q(S-x-1, k-1, x+1) \\ &- Q(S-x-2, k-1, x+2) \\ &- \text{etc.} \end{aligned} \tag{16}$$

Le nombre $p_k(S)$ dénote les k-partitions de S, incluant potentiellement certaines dont l'élément dominant (i.e. "s_1") serait $x+1$, ou $x+2$, etc. Ces k-partitions excédentaires peuvent être fragmentées en deux parts, soit pour l'une $\{s_1\}_j$, où "s_1" $= x+u$, $u > 0$, et pour l'autre $\{s_2, \ldots, s_k\}_j$, la valeur plafond étant ici $x+u$. La seconde part, $\{s_2, \ldots, s_k\}_j$, désigne manifestement des $(k-1)$-partitions de l'entier $S-x-u$, soit $\lambda_{k-1}(S-x-u, x+u)$, pour $u = 1, 2$, etc. Il s'agit donc de soustraire de $p_k(S)$ ces $k-1$-partitions d'un entier réduit, dont le nombre est encore donné par la fonction Q. L'application du théorème (14) complète la base du calcul récursif de la fonction Q qui, rappelons-le, correspond à l'élément de fréquence de la distribution de Wilcoxon, selon l'équivalence :

$$f_{n,k}(R) = Q[R - 1/2 \cdot k \cdot (k-1), k, n-k+1]. \tag{17}$$

Notons enfin que, la distribution nulle de R étant exactement symétrique, on peut en réduire le calcul aux valeurs supérieures de R ($R > \mu_R$) en appliquant les équivalences:

$$fn, k(R) = fn, k(2\mu_R - R) \tag{18a}$$

et

$$f_{cum(n,k)}(R) = f_{cum(n,k)}(2\mu_R - R - 1). \tag{18b}$$

Reprise de notre illustration, avec $n = 10$ et $k = 4$

Nous revenons maintenant à notre illustration offerte en début d'exposé, qui présente deux groupes de données, de tailles n_1 (ou k) $= 4$ et $n_2 = 6$, pour $n = 4 + 6 = 10$. Les revoici présentées au tableau 1.

Après substitution du rang de 1 à 10 approprié à chaque donnée, on effectue la somme des $k = 4$ rangs

du petit groupe inscrit à la gauche du tableau,[2] ici $R = 4 + 10 + 9 + 7 = 30$.

La statistique $R = R(10, 4)$ a pour minimum $k \cdot (k + 1)/2 = 10$, pour maximum $1/2k \cdot (2n - k + 1) = 34$ et pour étendue $k \cdot (n - k) = 24$. Sa moyenne (éq. 5) et sa variance (éq. 6) sont ici égales, soit $m_1 = \mu_R = 22$ et $\mu_2 = \sigma_R^2 = 22$, d'où $\sigma_R = 4,690$. Le moment central 3 est (toujours) nul ($\mu_3 = \gamma_1 = 0$), le moment central 4 est $\mu_4 = 1262,8$, donnant un indice d'aplatissement $\gamma_2 = \mu_4/\sigma^4 - 3 = -43/110 \approx -0,391$.

L'approximation Bêta. Pour le modèle de la loi Bêta symétrique, $\beta(p, p)$, la valeur du paramètre doublé obtenue par (7b) est $p \approx 6,1744$. La variable x_β standard, avec son paramètre calculé, a pour domaine (0, 1), espérance $\mu_\beta = 1/2$, variance $\sigma_\beta^2 \approx 0,018728$, $\sigma_\beta \approx 0,137$, indices $\gamma_1 = 0$ et $\gamma_2 \approx -0,391$. La valeur transposée de la statistique R en une variable du modèle Bêta (éq. 10) serait donc:

$$\begin{aligned} x_\beta &= 1/2 + \frac{R - \mu_R \pm 1/2}{\sigma_R} \times \sigma_\beta \\ &= 1/2 + \frac{30 - 22 - 1/2}{4,690} \times 0,137 \approx 0,719. \end{aligned}$$

La probabilité extrême de $x_\beta = 0,719$ dans la distribution $\beta(6,1744; 6,1744)$ est de 0,0561, d'où $p \approx 2 \times 0,0561 = 0,1122$, une vraisemblance suffisante pour tolérer l'hypothèse nulle au seuil bilatéral de 0,05. Le lecteur peut vérifier que les valeurs critiques supérieures de R déduites de cette approximation sont $R_{0,05} = 31$, $R_{0,025} = 32$, $R_{0,01} = 33$ et $R_{0,005} = 34$; en miroir, les valeurs correspondantes inférieures s'obtiennent par $k(n + 1) - R$, soit 13, 12, 11 et 10.

L'approximation normale. L'approximation normale usuelle, plus simple, est :

$$z_R = \frac{R - \mu_R \pm 1/2}{\sigma_R} = \frac{30 - 22 - 1/2}{4,690} \approx 1,599, \qquad (19)$$

dont la probabilté extrême, 0,0549, fournit une vraisemblance bilatérale de $2 \times 0,0549 = 0,1098$, suggérant elle aussi de tolérer l'hypothèse nulle.

Cette approximation, assez performante dans les conditions indiquées plus haut, permet de fixer aisément des valeurs critiques supérieures de R au moyen de l'expression $R_\alpha \approx \lceil \mu_R + \sigma_R \times z_{1-\alpha} + 1/2 \rceil$, où $\lceil u \rceil$ dénote l'entier r tel que $r \geq u$, $z_{1-\alpha}$ étant le quantile $1-\alpha$ de la loi normale standard. Selon $\alpha = 0,05, 0,025, 0,01$ et $0,005$ et $z_{1-\alpha} \approx 1,645, 1,960, 2,326$ et $2,576$. nous obtenons pour notre exemple respectivement $R_\alpha = 31, 32, 34$ et 35, cette dernière valeur étant ici impossible. Notre $R_{obs} = 30$ n'atteint pas le $R_{0,05} = 31$ critique.

La distribution exacte. D'entrée de jeu, rappelons que la distribue du U est identique à celle du R, sous l'équivalence $f(U) = f[R - 1/2k \cdot (k + 1)]$. La théorie des partitions développée plus haut nous permet de trouver les éléments $f(R)$ de la distribution de fréquences de notre statistique. Reprenant l'égalité (17), nous avons ici: $f_{10,4}(R) = Q(R - 1/2k(k - 1); k; x) = Q(R - 6; 4; 7)$, où $x = n - k + 1 = 10 - 4 + 1 = 7$, grâce à laquelle toute la distribution peut être trouvée (en exploitant ou non la symétrie présente). Nous en donnons deux exemples.

Soit $R = 10$, la valeur minimum; il s'agit de trouver $f_{10,4}(10)$. Cette fréquence équivaut à $Q(4, 4, 7)$, soit les partitions de l'entier 4 en 4 morceaux n'excédant pas 7. Il n'y a évidemment qu'une seule partition de 4 en 4 morceaux, soit $\lambda_4(4, 7) = 1111$, d'où $f(10) = 1$. Rappelons d'ailleurs que $Q(k, k, x) = 1$, puisque évidemment $pk(k) = 1$.

Soit $R = 20$, une situation plus corsée. La fréquence $f_{10,4}(20)$ égale Q(14,4,7), qu'on obtient par :

$$\begin{aligned} Q(14, 4, 7) = p_4(14) &- Q(6, 3, 8) \\ &- Q(5, 3, 9) \\ &- Q(4, 3, 10) \\ &- Q(3, 3, 11), \end{aligned}$$

série qui se termine au dernier terme présenté, le terme suivant, $Q(2, 3, 12)$, ne contenant évidemment aucune partition (car $S = 2 < k = 3$). Le nombre dominant, $p_4(14)$, s'évalue récursivement au moyen du théorème (14). Ainsi, $p_4(14) = p_1(10) + p_2(10) + p_3(10) + p_4(10)$. Par définition, $p_1(10) = 1$; $p_2(10)$ enclenche une chaîne récursive totalisant 5 partitions, $p_3(10) = 8$ et $p_4(10) = 9$, d'où $p_4(14) = 1 + 5 + 8 + 9 = 23$.

Quant à $Q(6, 3, 8)$, il égale $p_3(6) - Q(-3, 2, 9) -$ etc., les partitions à entier nul ou négatif étant évidemment vacantes. Alors $Q(6, 3, 8) = p_3(6) = 3$ [car $p_3(6) = p_1(3) + p_2(3) + p_3(3) = 1 + 1 + 1$, le premier et le dernier par définition, et le médian étant $p_2(3) = p_1(1) + p_2(1) = 1 + 0$]. En résumé, $Q(6, 3, 9) = 3$, $Q(5, 3, 9) = 2$, $Q(4, 3, 10) = 1$ et $Q(3, 3, 11) = 1$. On obtient ainsi que $f_{10,4}(20) = Q(14, 4, 7) = 23 - 3 - 2 - 1 - 1 = 16$. Le reste de la distribution peut s'obtenir de même.[3] La distribution

[2] L'utilisation du "petit groupe" (de taille k, ou n_1), commodité traditionnellement suggérée parce qu'elle économise du calcul, n'est pas requise et est retenue ici pour simplifier l'exposé. Noter que, hormis l'espérance μ_R (éq. 5), qui deviendrait $(n - k)(n + 1)/2$ si on utilisait le "grand groupe" de taille $n - k$, les autres calculs, *mutatis mutandis*, resteraient les mêmes.

[3] Le calcul *récursif* décrit ici est avantageusement remplacé par un calcul cumulatif dans un programme d'ordinateur. Il consiste à préparer un tableau triangulaire d'indices k ($1 \leq k \leq n$) et n. Au moment de quérir la valeur $p_k(n)$, l'algorithme de calcul vérifie d'abord si la valeur est déjà disponible au tableau, sinon la chaîne récursive est (ré-)engagée, la valeur calculée, puis inscrite et marquée disponible au tableau. En plus d'être économique et naturelle, cette méthode *constructive* assure une rapidité optimale du calcul des fonctions $p_k(n)$ et $Q(n, k, x)$.

nulle complète de R (ou U) pour $n = 10$, $k = 4$, débute à $R = 10$ (ou $U = 0$) : 1, 1, 2, 3, 5, 6, 9, 10, 13, 14, 16, 16, 18, 16, 16, 14, 13, 10, 9, 6, 5, 3, 2, 1, 1, la série se terminant pour $R = 34$ (ou $U = 24$).

La distribution nulle ci-dessus permet de déterminer les valeurs critiques exactes, supérieures, soit $R_{0,05} = 31$, $R_{0,025} = 32$, $R_{0,01} = 33$ et $R_{0,005} = 34$, coïncidant ici avec celles trouvées plus haut par l'approximation Bêta.

Quant à notre valeur $R = 30$ observée, elle n'atteint pas la valeur critique $R_{0,05} = 31$; sa probabilité extrême, $Pr(R \geq 30) = \sum_{R=30}^{34} f(R) / \binom{n}{k} = (5 + 3 + 2 + 1 + 1) / \binom{n}{k} = 12/210 \approx 0,0571$, doublée (pour un test bilatéral) à 0,114, déborde 0,05, en accord ici avec les autres tests.

Le lecteur intéressé trouvera sur le site du journal une feuille de calcul Excel jointe à l'article et téléchargeable, laquelle réalise tous les calculs documentés ci-dessus.

De la puissance et d'autres commentaires

La puissance statistique du test de la somme des rangs de Wilcoxon (R) a d'emblée fait l'objet d'études, qu'il est loisible de résumer comme suit. Sous des conditions de variation normale et homoscédastique, conditions pour lesquelles le test t (1) se voit attribuer la puissance de référence, le test de Wilcoxon a une puissance relative asymptotique de $3/\pi \approx 0,955$. Les données d'Allaire (1993), basées sur de petits échantillons de tailles égales à 4, 5 et 6, suggèrent que cette valeur est minimale, les puissances relatives enregistrées (au seuil avoisinant 0,05) étant de 0,970, 0,965 et 0,967. Sous des conditions de variation non normales, le t perd aussitôt sa validité stricte puisque le respect du seuil α n'y est plus garanti, ce qui n'est pas le cas du R, quel que soit le modèle de variation de la variable testée. Dans ces cas, la puissance relative varie d'un modèle à l'autre, parfois au-delà de 1.[4] Finalement, Hodges et Lehmann (1956; voir aussi Kendall et Stuart, 1979, p. 524-525) établissent que, sous quelque condition que ce soit, la puissance relative n'est jamais inférieure à $0,864$. Pour obtenir une solution exacte et à défaut de valeurs critiques ou d'une solution programmée par notre fonction $Q(S, k, x)$ ou par d'autres fonctions récursives, l'usager peut recourir à "l'algorithme (très) efficace" de combinaisons complètes élaboré à cette fin et décrit dans Ferland et Laurencelle (2012), algorithme indifféremment applicable aux valeurs originales, leurs rangs ou leurs scores normaux.[5]

De très bonne efficience, donc, et protégeant toujours son seuil α, au contraire du test t, le R de Wilcoxon, basé sur la transposition des données en rangs de 1 à n, cède tout de même la vedette à un autre test permutationnel de même forme, celui-là basé sur une transposition en n scores normaux équivalents, c.-à-d. les espérances des statistiques d'ordre d'un échantillon de n données normales standard (Hoeffding, 1952; Bradley, 1968) Ce test, toujours valide comme le R, a une puissance minimale de 1 et se montre aussi plus robuste que le t (Kendall et Stuart, 1979, p. 526-531). Toutefois, même s'il s'avère le meilleur test pour comparer le niveau des données de deux groupes, le test de la somme des scores normaux n'est pas d'utilisation commode et exige des calculs laborieux : aucun logiciel n'en est communément disponible, à notre connaissance.

References

Allaire, D. (1993). La puissance du test t, du test de wilcoxon et du test permutaitonnel pour la comparaison de deux moyennes indépendantes, sous quatre populations. *Lettres Statistiques*, *9*, 25–67.

Bose, R. C., & Manvel, B. (1984). *Introduction to combinatorial theory*. Wiley: New York.

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs (NJ): Prentice-Hall.

David, F. N., & Johnson, N. L. (1954). Statistical treatment of censored data, part i. *Fundamental formulae. Biometrika*, *41*, 228–240.

Edgington, E. S. (1980). *Randomization tests*. Marcel Dekker: New York.

Ferland, P., & Laurencelle, L. (2012). Un algorithme efficace pour la comparaison de deux moyennes indépendantes par combinatoire exhaustive. *The Quantitative Methods for Psychology*, *8*, 137–150. doi:10.20982/tqmp.08.3.p137

Fix, E., & Hodges, J. L., Jr. (1955). Significance probabilities of the wilcoxon test. *Annals of mathematical statistics*, *26*, 301–312.

Hodges, J. L., Jr., & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics*, *27*, 324–325.

Hoeffding, W. (1952). The large-sample power of tests based on permtations of observations. *Annals of Mathematical Statistics*, *23*, 169–192.

Kendall, M. G., & Stuart, A. (1977). New York: Macmillan.

Kendall, M. G., & Stuart, A. (1979). New York: Macmillan.

Kruskal, W. H. (1957). Historical notes on the wilcoxon unpaired two-sample test. *Journal of the American Statistical Association*, *52*, 356–360.

[4] Dans le cas de variation non normale, pour laquelle le taux de rejet α n'est pas garanti (et est souvent violé), la valeur de puissance calculable (relativement au t) est explicitement biaisée.

[5] La version en QBasic est disponible auprès de l'auteur.

Laurencelle, L. (2001). *Hasard, nombres aléatoires et méthode monte carlo*. Presses de l'Université du Québec: Sainte-Foy.

Laurencelle, L. (2005). La distribution nulle de la somme des rangs de wilcoxon et ses moments. *Lettres Statistiques*, *12*, 99–108.

Mann, H. H., & Whitney, D. R. (1947). On a test of whether one of two variables is stochas¬tically larger than the other. *Annals of mathematical statistics*, *18*, 50–60.

Owen, D. B. (1962). *Handbook of statistical tables*. Addison-Wesley: Reading (Mass.)

Rohlf, F. J., & Sokal, R. R. (1995). *Statistical tables (3e édition)*. New York: Freeman.

Siegel, S. (1956). *Nonparametric statistics*. McGraw-Hill: New York.

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences (2e édition)*. McGraw-Hill: New York.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, *1*, 80–83.

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on the journal's web site.

Citation

Laurencelle, L. (2020). La somme des rangs de wilcoxon, ses moments, sa distribution nulle,avec illustration et compléments. *The Quantitative Methods for Psychology*, *16*(1), 46–53. doi:10.20982/tqmp.16.1.p046

Received: 19/11/2019 ~ Accepted: 01/01/2020