Test de Wilcoxon-Mann-Whitney : (pour échantillons non-appariés)

Ce test regroupe deux tests équivalents : le test U de Mann-Whitney et le test W de Wilcoxon appelé test de la somme de rangs de Wilcoxon. Les deux tests se déduisent l'un de l'autre. Cependant, le test de Wilcoxon est plus facile.

Soient $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_m)$ deux échantillons indépendants issus respectivement de X et Y supposées absolument continues de fonctions de répartition respectives F et G. On suppose que $n \leq m$ et que $F\ et\ G$ sont identiques à une translation près : $F(x) = G(x-\theta), \forall x \in \mathbb{R}$, θ inconnu (X et Y ont la même forme de distribution et différent seulement dans leurs paramètres de position ou de tendance centrale). Soient $M_1\ et\ M_2$ les médianes de X et Y respectivement (μ_1 et μ_2 les moyennes respectives).

On veut tester

$$H_0: M_1 = M_2 \quad vs \quad H_1: M_1 \neq M_2 \quad (H_1: M_1 > M_2 \text{ ou } H_1: M_1 < M_2).$$

$$H_0: \mu_1 = \mu_2 \quad vs \quad H_1: \mu_1 \neq \mu_2 \quad (H_1: \mu_1 > \mu_2 \text{ ou } H_1: \mu_1 < \mu_2)$$

$$H_0: F = G \quad vs \quad H_1: F \neq G \quad (H_1: F > G \text{ ou } H_1: F < G)$$

$$H_0: \theta = 0 \quad vs \quad H_1: \theta \neq 0 \quad (H_1: \theta < 0 \text{ ou } H_1: \theta > 0)$$

1.Test de la somme de rangs de Wilcoxon

On regroupe les deux échantillons, on obtient ainsi un échantillon $(Z_1, \cdots, Z_{n+m}) = (X_1, \cdots, X_n, Y_1, \cdots, Y_m)$ de taille $n+m$. On ordonne les Z_i par ordre croissant et on associe à Z_i son rang dans $\left(Z_{(1)}, \cdots, Z_{(n+m)}\right)$. On note W_1 et W_2 les sommes des rangs des observations provenant des deux échantillons : W_1 la somme des rangs des observations $X_i, i = \overline{1,n}$, W_2 la somme des rangs des observations $Y_i, i = \overline{1,m}$. Les sommes W_1 et W_2 sont liées par la relation

$$W_1 + W_2 = \sum_{i=1}^{n+m} i = \frac{(n+m)(n+m+1)}{2}.$$

La plus petite valeur de W_1est $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ et sa plus grande valeur est $\sum_{i=m+1}^{n+m} i = \frac{n(2m+n+1)}{2}$

$\left(\frac{n(n+1)}{2} \leq W_1 \leq \frac{n(2m+n+1)}{2}\right)$.

Si on associe à $Z_{(i)}$ la variable aléatoire indicatrice D_i définie par

$$D_i = \begin{cases} 1 & si\ Z_{(i)} \text{ est un } X, \\ 0 & si\ Z_{(i)} \text{ est un Y,} \end{cases} \quad \overline{i = 1, n+m},$$

alors W_1 peut etre exprimée comme une combinaison linéaire des D_i

$$W_1 = \sum_{i=1}^{n+m} iD_i.$$

Exemple:

Soient les observations $(x_1, x_2, x_3) = (1,6,10)$ et $(y_1, y_2, y_3, y_4) = (2,9,3,4)$. Les z_i, $z_{(i)}, et\ d_i$ sont données dans le tableau suivant

i	1	2	3	4	5	6	7
z_i	1	6	10	2	9	3	4
$z_{(i)}$	1	2	3	4	6	9	10
d_i	1	0	0	0	1	0	1.

Par suite $w_1 = 1 + 5 + 7 = 13$ et $w_2 = \frac{7(8)}{2} - w_1 = 15$.

Théorème:

Sous l'hypothèse nulle H_0,

$$E(D_i) = \frac{n}{n+m}, Var(D_i) = \frac{nm}{(n+m)^2},\ \overline{i = 1, n+m},$$

$$Cov(D_i, D_j) = \frac{-nm}{(n+m)^2(n+m-1)}, i \neq j,$$

$$E(W_1) = \frac{n(n+m+1)}{2}, \quad Var(W_1) = \frac{nm(n+m+1)}{12},$$

$$P(W_1 = k) = \frac{w_{n,m}(k)}{C_{n+m}^n}, \qquad k = \overline{\frac{n(n+1)}{2}, \frac{n(2m+n+1)}{2}},$$

où $w_{n,m}(k)$ est le nombre d'arrangements de n uns et m zéros pour lesquels $W_1 = k$.

Preuve :

- La variable indicatrice D_i suit une loi de Bernoulli de paramètre $p = P(D_i = 1) = \frac{n}{n+m}$. Ainsi

$$E(D_i) = \frac{n}{n+m}, \text{ et } Var(D_i) = \frac{n}{n+m}\left(1 - \frac{n}{n+m}\right) = \frac{nm}{(n+m)^2}.$$

- Pour $i \neq j$, $D_i D_j$ est aussi une variable de Bernoulli, donc

$$E(D_i D_j) = P(D_i D_j = 1) = P(D_i = 1, D_j = 1) = C_n^2 = \frac{n(n-1)}{(n+m)(n+m+1)}$$

et $Cov(D_i, D_j) = \frac{n(n-1)}{(n+m)(n+m+1)} - \frac{n^2}{(n+m)^2} = \frac{-nm}{(n+m)^2(n+m-1)}$.

- $E(W_1) = \sum_{i=1}^{n+m} i\, E(D_i) = \frac{n}{n+m}\sum_{i=1}^{n+m} i = \frac{n}{n+m}\frac{(n+m)(n+m+1)}{2} = \frac{n(n+m+1)}{2}$.

$$Var(W_1) = \sum_{i=1}^{n+m} i^2\, Var(D_i) + \sum_{i \neq j} \sum ij\, Cov(D_i, D_j)$$

$$= \frac{nm}{(n+m)^2}\sum_{i=1}^{n+m} i^2 - \frac{-nm}{(n+m)^2(n+m-1)} \sum_{i \neq j} \sum ij$$

$$= \frac{nm}{(n+m)^2(n+m-1)}\left((n+m-1)\sum_{i=1}^{n+m} i^2 - \sum_{i \neq j} \sum ij\right)$$

$$= \frac{nm}{(n+m)^2(n+m-1)}\left((n+m)\sum_{i=1}^{n+m} i^2 - \sum_{i=1}^{n+m}\sum_{i=1}^{n+m} ij\right)$$

$$= \frac{nm}{(n+m)^2(n+m-1)}\left((n+m)\sum_{i=1}^{n+m} i^2 - \left(\sum_{i=1}^{n+m} i\right)^2\right)$$

$$= \frac{nm}{(n+m)^2(n+m-1)}\left((n+m)\frac{(n+m)(n+m+1)(2(n+m)+1)}{6} - \frac{(n+m)^2(n+m+1)^2}{4}\right).$$

Après simplification, on obtient $Var(W_1) = \frac{nm(n+m+1)}{12}$.

- La distribution exacte de W_1 sous H_0 dépend de celle de $(D_1, \dots, D_{n+m})$ qui prend ses valeurs dans l'ensemble des arrangements possibles de n uns (X) et m zéros (Y) qui sont au nombre de $C_{n+m}^{n} = C_{n+m}^{m}$. Ces arrangements sont, sous H_0, équiprobables, c.à.d. $P(D_1 = d_1, \dots, D_{n+m} = d_{n+m}) = \frac{1}{C_{n+m}^{n}}, \forall\ (d_1, \dots, d_{n+m})$ disposition possible de n uns et m zéros . Par suite la loi de W_1 sous H_0 est déterminée par énumération directe. Les valeurs de W_1 sont calculées pour chaque disposition $(d_1, \dots, d_{n+m})$ et $P(W_1 = k) = \frac{w_{n,m}(k)}{C_{n+m}^{n}}$ où $w_{n,m}(k)$ est le nombre d'arrangements $(d_1, \dots, d_{n+m})$ pour lesquels $W_1 = k$.

Exemple : $n = 2, m = 3$

On a $3 \leq W_1 \leq 9$ et $E(W_1) = 6$. Il ya $C_5^2 = 10$ arrangements possibles de 2 uns (X) et 3 zéros (Y)

Disposition	Rangs des X	Valeur de W_1
$(X,X,Y,Y,Y) = (1,1,0,0,0)$	1,2	3
$(1,0,1,0,0)$	1,3	4
$(1,0,0,1,0)$	1,4	5
$(1,0,0,0,1)$	1,5	6
$(0,1,1,0,0)$	2,3	5
$(0,1,0,1,0)$	2,4	6
$(0,1,0,0,1)$	2,5	7
$(0,0,1,1,0)$	3,4	7
$(0,0,1,0,1)$	3,5	8
$(0,0,0,1,1)$	4,5	9

Valeur de W_1 k	Fréquence $w_{2,3}(k)$	$P(W_1 = k)$
3	1	1/10=0.1
4	1	0.1
5	2	0.2
6	2	0.2
7	2	0.2
8	1	0.1
9	1	0.1

Remarques :

1. La distribution de W_1, sous H_0, est symétrique par rapport à $E(W_1)$:

$$P_{H_0}(W_1 - E(W_1) = w) = P_{H_0}(W_1 - E(W_1) = -w),$$

et $$P_{H_0}(W_1 \leq k) = P_{H_0}(W_1 \geq 2E(W_1) - k), \ \frac{n(n+1)}{2} \leq k \leq E(W_1).$$

2. Il existe des relations récursives pour déterminer la distribution de W_1 sous H_0. Si $w_{n,m}(k)$ désigne le nombre d'arrangements de n uns (X) et m zéros (Y) tel que $W_1 = k$ alors

$$w_{n,m}(k) = w_{n-1,m}(k - (n+m)) + w_{n,m-1}(k)$$

$$et\ P_{H_0}(W_1 = k) = \ P_{n,m}(k) = \frac{w_{n-1,m}(k - (n+m)) + w_{n,m-1}(k)}{C_{n+m}^{n}}$$

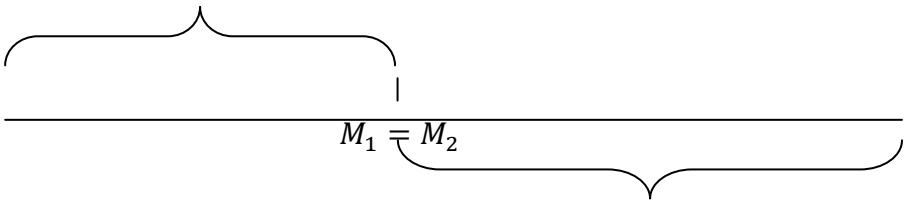
ou de manière équivalente

$$(n+m)P_{n,m}(k) = nP_{n-1,m}\big(k - (n+m)\big) + mP_{n,m-1}(k).$$

Région de rejet et p-valeur :

Sous H_0, la somme des rangs est presque la même pour les deux échantillons

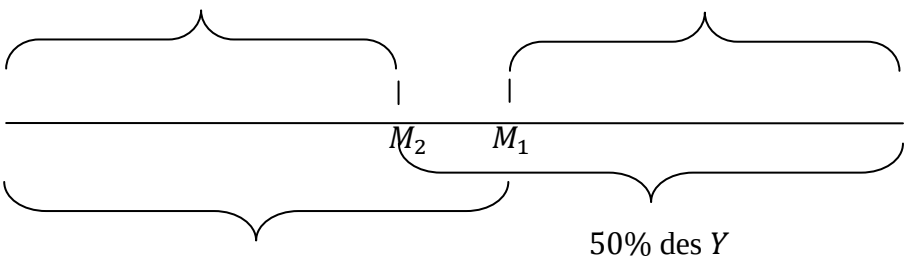
50% des X, 50%des Y



50% des X, 50%des Y

Pour l'alternative $H_1: M_1 > M_2$, H_0 est rejetée pour une trop forte valeur de W_1(trop faible valeur de W_2)

50% des Y 50% des X



50% des Y

50% $des\ X$

Pour l'alternative $H_1: M_1 < M_2$, H_0 est rejetée pour une trop faible valeur de W_1(trop forte valeur de W_2).

Pour l'alternative $H_1: M_1 \neq M_2$, H_0 est rejetée pour une trop forte valeur ou une trop faible valeur de W_1 (trop faible valeur de W_2 ou trop faible valeur de W_1).

Si on pose

$$W = \begin{cases} \text{Min}(W_1, W_2) & si\ H_1: M_1 \neq M_2, \\ W_2 & si\ H_1: M_1 > M_2, \\ W_1 & si\ H_1: M_1 < M_2, \end{cases}$$

alors H_0 est rejetée au seuil α si $W \leq w_\alpha$ telle que $P_{H_0}(W \leq w_\alpha) \leq \alpha$.

Les valeurs critiques w_α sont tabulées dans le cas de tests bilatéral et unilatéral (pour α, n, m fixés).

Notons que : $(W_2 \leq w_\alpha) \Leftrightarrow \left(\frac{(n+m)(n+m+1)}{2} - W_1 \leq w_\alpha\right) \Leftrightarrow (W_1 \geq w'_\alpha)$, $w'_\alpha = \frac{(n+m)(n+m+1)}{2} - w_\alpha$. Donc dans le cas d'un test bilatéral, la région critique est de la forme

$$(W_1 \leq w_{\alpha/2}) \text{ ou } (W_1 \geq w'_{\alpha/2})$$

où $P(W_1 \leq w_{\alpha/2}) \leq \alpha/2$ et $P(W_1 \geq w'_{\alpha/2}) \leq \alpha/2$

Autrement dit, le niveau de signification d'un test bilatéral est égal à 2 fois le niveau de signification d'un test unilatéral.

La $p-$valeur α_0 pour une valeur observée w de W_1 est donnée par

$$\alpha_0 = \begin{cases} P_{H_0}(W_1 \leq w) & si\ H_1: M_1 < M_2, \\ P_{H_0}(W_1 \geq w) & si\ H_1: M_1 > M_2, \\ 2\min\left(P_{H_0}(W_1 \leq w), P_{H_0}(W_1 \geq w)\right) & si\ H_1: M_1 \neq M_2. \end{cases}$$

<u>Approximation normale :</u>

Pour des tailles d'échantillons n, m assez grandes $(n, m \geq 8)$, on peut utiliser la statistique

$$Z = \frac{W_1 - E(W_1)}{\sqrt{Var(W_1)}} = \frac{W_1 - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}$$

qui suit la loi normale centrée réduite.

Les régions de rejet et les p −valeurs son données dans le tableau suivant où w est une valeur observée de W_1 :

Alternative H_1	Région de rejet	p −valeur α_0
$M_1 < M_2$	$Z \leq z_\alpha$	$\Phi\left(\frac{w - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}\right)$
$M_1 > M_2$	$Z \geq z_{1-\alpha}$	$1 - \Phi\left(\frac{w - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}\right)$
$M_1 \neq M_2$	$\lvert Z \rvert \geq z_{1-\alpha/2}$	2(la plus petite des deux ci-dessus)

Avec correction de continuité, on obtient les régions critiques et p −valeurs suivantes :

Alternative H_1	Région de rejet	p −valeur α_0
$M_1 < M_2$	$W_1 \leq z_\alpha \sqrt{\frac{nm(n+m+1)}{12}} - 0.5 + \frac{n(n+m+1)}{2}$	$\Phi\left(\frac{w + 0.5 - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}\right)$
$M_1 > M_2$	$W_1 \geq z_{1-\alpha} \sqrt{\frac{nm(n+m+1)}{12}} + 0.5 + \frac{n(n+m+1)}{2}$	$1 - \Phi\left(\frac{w - 0.5 - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}\right)$
$M_1 \neq M_2$	Les deux ci-dessus avec α remplacé par $\alpha/2$	2(la plus petite des deux ci-dessus)

Exemple :

On veut comparer les performances de deux groupes d'élèves à des tests d'habileté manuelle. On choisit aléatoirement 8 individus du premier groupe et 10 du deuxième. Les performances en minutes sont les suivantes :

Groupe 1	22	31	14	19	24	28	27	15		
Groupe 2	25	13	20	11	23	16	21	18	17	26

On réordonne les 18 observations par ordre croissant, on obtient

i	z_i	$z_{(i)}$	Rang de z_i	Groupe	d_i
1	22	11	11	2	0
2	31	13	18	2	0
3	14	14	3	1	1
4	19	15	8	1	1
5	24	16	13	2	0
6	28	17	17	2	0
7	27	18	16	2	0
8	15	19	4	1	1
9	25	20	14	2	0
10	13	21	2	2	0
11	20	22	9	1	1
12	11	23	1	2	0
13	23	24	12	1	1
14	16	25	5	2	0
15	21	26	10	2	0
16	18	27	7	1	1
17	17	28	6	1	1
18	26	31	15	1	1

La somme des rangs des individus du premier groupe est :

$$w_1 = 11 + 18 + 3 + 8 + 13 + 17 + 16 + 4 = 90$$

La somme des rangs des individus du deuxième groupe est

$$w_2 = \frac{18(19)}{2} - w_1 = 171 - 90 = 81$$

$$w_2 = 14 + 2 + 9 + 1 + 12 + 5 + 10 + 7 + 6 + 15 = 81$$

Soit H_0: $M_1 = M_2$ $(\mu_1 = \mu_2)$. Si H_0 est vraie alors $E(W_1) = \frac{8(19)}{2} = 76$ et $Var(W_1) = \frac{8(10)19}{12} = 126.66$.

Région critique $\begin{cases} W_1 \leq w_\alpha & si\ H_1: M_1 < M_2 \\ W_2 \leq w_\alpha \Leftrightarrow W_1 \geq w_\alpha{}' & si\ H_1: M_1 > M_2 \\ Min(W_1, W_2) \leq w_\alpha \Leftrightarrow W_1 \leq w_{\alpha/2}\ ou\ W_1 \geq w_{\alpha/2}{}' & si\ H_1: M_1 \neq M_2 \end{cases}$

- Valeur critique au seuil $\alpha = 0.05$:

$$w_{0.05} = \begin{cases} 53 & si\ H_1: M_1 \neq M_2, \\ 56 & si\ H_1: M_1 < M_2, \\ 75 = 171 - w_\alpha{}' = 171 - 96 & si\ H_1: M_1 > M_2. \end{cases}$$

Conclusion : on accepte H_0 au seuil α =0.05 c. à. d. les deux groupes ont les mêmes performances.

- p-valeur

$$\alpha_0 = \begin{cases} P_{H_0}(W_1 \leq 90) & si\ H_1: M_1 < M_2, \\ P_{H_0}(W_1 \geq 90) & si\ H_1: M_1 > M_2, \\ 2\min\left(P_{H_0}(W_1 \leq 90), P_{H_0}(W_1 \geq 90)\right) & si\ H_1: M_1 \neq M_2. \end{cases}$$

$$\alpha_0 = \begin{cases} 1 - P_{H_0}(W_1 \geq 91) = 1 - 0.102 = 0.898 & si\ H_1: M_1 < M_2, \\ 0.118 & si\ H_1: M_1 > M_2, \\ 0.236 & si\ H_1: M_1 \neq M_2. \end{cases}$$

$\alpha_0 > \alpha$ =0.05 → on accepte H_0.

Exemple :

On a mesuré dans deux forêts les hauteurs de 27 arbres choisis au hasard et indépendamment (13 arbres choisis dans la forêt 1 et 14 arbres choisis dans la forêt 2). On veut vérifier si les hauteurs médianes sont égales ou pas (vérifier si les distributions des hauteurs des arbres des deux forêts sont ou ne sont pas égales). Les hauteurs observées sont les suivantes :

Forêt 1 X	Forêt 2 Y
23.4	22.5
24.6	23.7
25.0	24.3
26.3	25.3
26.6	26.1
27.0	26.7
27.7	27.4
24.4	22.9
24.9	24.6
26.2	24.5
26.5	26.0
26.8	26.4
27.6	26.9
	28.5

On veut tester $H_0: M_1 = M_2 \quad vs \quad H_1: M_1 \neq M_2$.