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 des 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,

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

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

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

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

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

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

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

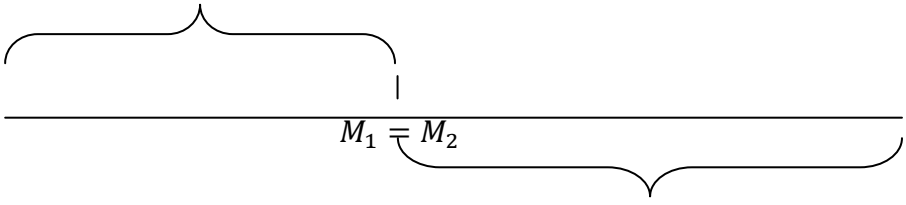
ou de manière équivalente

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

Région de rejet et p-valeur :

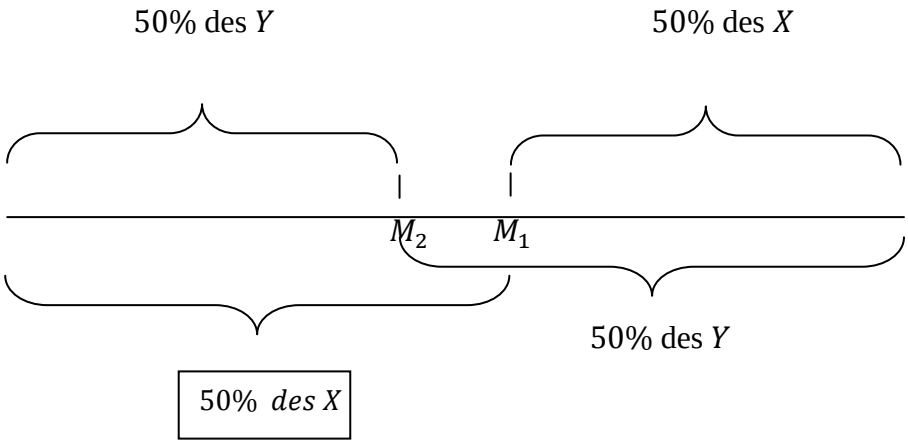
Sous H_0, la somme des rangs est presque la même pour les deux échantillons :

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)



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

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

Approximation normale :

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	d_i
1	22	11	11	0
2	31	13	18	0
3	14	14	3	1
4	19	15	8	1
5	24	16	13	0
6	28	17	17	0
7	27	18	16	0
8	15	19	4	1
9	25	20	14	0
10	13	21	2	0
11	20	22	9	1
12	11	23	1	0
13	23	24	12	1
14	16	25	5	0
15	21	26	10	0
16	18	27	7	1
17	17	28	6	1
18	26	31	15	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 $\alpha = 0.05$:

$$w_{0.05} = \begin{cases} 53 < \min(w_1, w_2) & si\ H_1: M_1 \neq M_2, \\ 56 < w_1 & si\ H_1: M_1 < M_2, \\ 75 = 171 - w_\alpha{}' = 171 - 96 < w_2 & si\ H_1: M_1 > M_2. \end{cases}$$

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.

- Approximation normale sans correction de continuité

  On a $z = \frac{w_1 - 76}{\sqrt{126.66}} = 1.243,\ z_{0.95} = 1.64,\ z_{0.05} = -1.64,\ z_{0.975} = 1.96.$

  $$\begin{cases} z > z_{0.05} & si\ H_1 : M_1 < M_2, \\ z < z_{0.95} & si\ H_1 : M_1 > M_2, \\ |z| < z_{0.975} & si\ H_1 : M_1 \neq M_2. \end{cases}$$

  Conclusion : on accepte H_0.

  Pour les p-valeurs, on obtient

  $$\alpha_0 = \begin{cases} \Phi(1.243) = 0{,}8925 > 0.05 & si\ H_1 : M_1 < M_2, \\ 1 - \Phi(1.243) = 0.1075 > 0.05 & si\ H_1 : M_1 > M_2, \\ 0.215 > 0.05 & si\ H_1 : M_1 \neq M_2. \end{cases}$$

- Approximation normale avec correction de continuité

  $$\alpha_0 = \begin{cases} \Phi(1.28) = 0{,}8997 > 0.05 & si\ H_1 : M_1 < M_2, \\ 1 - \Phi(1.19) = 0.117 > 0.05 & si\ H_1 : M_1 > M_2, \\ 0.234 > 0.05 & si\ H_1 : M_1 \neq M_2. \end{cases}$$

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, au seuil $\alpha = 5\%$, $H_0: M_1 = M_2 \quad vs \quad H_1: M_1 \neq M_2$.

Cas d'ex-æquo

Si l'échantillon (groupé) considéré $(Z_1, \dots, Z_{n+m})$ présente des ex-æquo et dans le cas où on leur attribue un rang moyen, il faut corriger la variance de W_1 dans l'approximation normale pour les grands échantillons : On remplace le terme $\frac{nm(n+m+1)}{12} = Var(W_1)$ par

$$\frac{nm(n+m+1)}{12} - \frac{nm}{12(n+m)(n+m-1)} \sum t_l(t_l^2 - 1)$$

où t_l est le nombre de valeurs Z_i ex-æquo (Xet/ou Y) ayant le $l^{ième}$ rang.

Exemple

On dispose de deux échantillons de tailles respectives 10 et15 :

Echantillon 1 (X)	80	100	90	110	125	130	70	75	71	83					
Echantillon 2 (Y)	100	120	80	140	130	160	115	120	73	88	135	125	128	95	87.

Tester au seuil 5% si les deux échantillons proviennent de la même population. Quelle hypothèse faites-vous ?

2. Test de Mann-Whitney

On regroupe les $(n+m)$ observations, on obtient l'échantillon $(Z_1, \dots, Z_{n+m}) = (X_1, \dots, X_n, Y_1, \dots, Y_m)$.

On note U_1 le nombre de fois qu'un Y précède un X dans l'échantillon ordonné $\left(Z_{(1)}, \dots, Z_{(n+m)}\right)$.

Si on associe à chaque couple $\left(X_i, Y_j\right)$, la variable indicatrice D_{ij}, définie pour tous $i = \overline{1,n},\ j = \overline{1,m}$, par

$$D_{ij} = \begin{cases} 1 & si\ Y_j < X_i, \\ 0 & sinon, \end{cases}$$

alors

$$U_1 = \sum_{i=1}^{n} \sum_{j=1}^{m} D_{ij}.$$

De même on définit U_2 le nombre de fois qu'un X précède un Y

$$U_2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(1 - D_{ij}\right) = nm - U_1.$$

Notons que $0 \leq U_1 \leq nm$:

- $U_1 = 0$ signifie que chaque X_i précède chaque Y_j,
- $U_1 = nm$ signifie que chaque Y_j précède chaque X_i.

Exemple :

Soient $(x_1, x_2, x_3) = (1,6,10),\ (y_1, y_2, y_3, y_4) = (2,9,3,4), n = 3,\ m = 4, nm = 12.$

On obtient l'échantillon regroupé ordonné

$$\left(z_{(1)}, z_{(2)}, z_{(3)}, z_{(4)}, z_{(5)}, z_{(6)}, z_{(7)}\right) = (1,2,3,4,6,9,10) = (x_1, y_1, y_3, y_4, x_2, y_2, x_3)$$

i	1	1	1	1	2	2	2	2	3	3	3	3
j	1	2	3	4	1	2	3	4	1	2	3	4
D_{ij}	0	0	0	0	1	0	1	1	1	1	1	1

Par suite $U_1 = 7$.

Théorème : Sous H_0

$$E(D_{ij}) = \frac{1}{2}, Var(D_{ij}) = \frac{1}{4}\ , i = \overline{1,n}, j = \overline{1,m},$$

$$Cov(D_{ij}, D_{hk}) = 0, i \neq h\ et\ j \neq k,$$

$$Cov(D_{ij}, D_{ik}) = \frac{1}{12}, j \neq k,$$

$$Cov(D_{ij}, D_{hj}) = \frac{1}{12}, i \neq h,$$

$$E(U_1) = \frac{nm}{2}\ , Var(U_1) = \frac{nm(n+m+1)}{12},$$

$$P(U_1 = k) = \frac{\mu_{n,m}(k)}{C_{n+m}^{n}}$$

où $\mu_{n,m}(k)$ est le nombre d'arrangements distincts des $n\ X$ et $m\ Y$ variables aléatoires pour lesquels on a $U_1 = k$.

Preuve :

- D_{ij} est une variable de Bernoulli de paramètre

$$p = P(Y_j < X_i) = P(Y < X) = \int_{-\infty}^{+\infty} P(Y < x)\, dF(x) = \int_{-\infty}^{+\infty} G(x)\, dF(x).$$

Sous H_0, $F = G$ donc

$$p = \frac{1}{2}\big(F(x)\big)^2\Big|_{-\infty}^{+\infty} = \frac{1}{2}.$$

Par suite $E(D_{ij}) = \frac{1}{2}$ et $Var(D_{ij}) = \frac{1}{4}$.

- $Si\ i \neq h\ et\ j \neq k$ alors D_{ij} et D_{hk} sont indépendantes et donc $Cov(D_{ij}, D_{hk}) = 0$.

- $Si\ \ j \neq k$ alors

$E(D_{ij}D_{ik}) = P\left((Y_j < X_i) \cap (Y_k < X_i)\right) = \int_{-\infty}^{+\infty} P(Y_j < x, Y_k < x) dF(x) = \int_{-\infty}^{+\infty} (G(x))^2 dF(x).$

Sous H_0

$$E(D_{ij}D_{ik}) = \int_{-\infty}^{+\infty} (F(x))^2 dF(x) = \frac{1}{3}(F(x))^3|_{-\infty}^{+\infty} = \frac{1}{3}.$$

Par suite $Cov(D_{ij}, D_{ik}) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$.

De la même manière, on montre que $Cov(D_{ij}, D_{hj}) = \frac{1}{12}, i \neq h$.

- $$E_{H_0}(U_1) = \sum_{i=1}^{n}\sum_{j=1}^{m} E_{H_0}(D_{ij}) = \frac{nm}{2}.$$

$$Var(U_1) = Cov\left(\sum_{i=1}^{n}\sum_{j=1}^{m} D_{ij}, \sum_{h=1}^{n}\sum_{k=1}^{m} D_{hk}\right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} Var(D_{ij}) + \sum_{i=1}^{n}\sum_{1\leq j\neq k\leq m} Cov(D_{ij}, D_{ik}) + \sum_{j=1}^{m}\sum_{1\leq i\neq h\leq n} Cov(D_{ij}, D_{hj})$$

$$= \frac{nm}{4} + \frac{n}{12}2C_m^2 + \frac{m}{12}2C_n^2 = \frac{nm}{12}(n+m+1)$$

- Sous H_0, les C_{m+n}^n arrangements possibles des n X et m Y sont équiprobables. Ainsi

$$P(U_1 = k) = \frac{\mu_{n,m}(k)}{C_{n+m}^n}$$

où $\mu_{n,m}(k)$ est le nombre d'arrangements pour lesquels $U_1 = k$.

Exemple : $n = 2, m = 3,\ 0 \leq U_1 \leq 6$.

Arrangement	Valeur de U
(X,X,Y,Y,Y)	0
(X,Y,X,Y,Y)	1
(X,Y,Y,X,Y)	2
(X,Y,Y,Y,X)	3
(Y,X,X,Y,Y)	2
(Y,X,Y,X,Y)	3
(Y,X,Y,Y,X)	4
(Y,Y,X,X,Y)	4
(Y,Y,X,Y,X)	5
(Y,Y,Y,X,X)	6

Valeur de U_1 k	Fréquence $\mu_{2,3}(k)$	$P(U_1 = k)$
0	1	$\frac{1}{10}$=0.1
1	1	0.1
2	2	0.2
3	2	0.2
4	2	0.2
5	1	0.1
6	1	0.1

Remarques :

1. La distribution de U_1 est symétrique autour de la moyenne $E(U_1) = \frac{mn}{2}$

$$P_{H_0}(U_1 \leq k) = P_{H_0}(U_1 \geq 2E(U_1) - k), \;\; 0 \leq k \leq E(U_1).$$

2. Il existe une relation récursive pour déterminer la distribution de U_1 :

$$\mu_{n,m}(k) = \mu_{n,m-1}(k) + \mu_{n-1,m}(k - m).$$

Région critique et p-valeur

Pour l'alternative $H_1: M_1 > M_2, H_0$ est rejetée pour une faible valeur de U_2 (forte valeur de U_1).

Pour l'alternative $H_1: M_1 < M_2, H_0$ est rejetée pour une faible valeur de U_1 (forte valeur de U_2).

Pour l'alternative $H_1: M_1 \neq M_2, H_0$ est rejetée pour une faible valeur de U_1 ou U_2.

Si on pose

$$U = \begin{cases} \text{Min}(U_1, U_2) & si\ H_1: M_1 \neq M_2, \\ U_2 & si\ H_1: M_1 > M_2, \\ U_1 & si\ H_1: M_1 < M_2, \end{cases}$$

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

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

Notons que : $(U_2 \leq u_\alpha) \Leftrightarrow (nm - U_1 \leq u_\alpha) \Leftrightarrow (U_1 \geq u_\alpha^{'})$ avec $u_\alpha^{'} = nm - u_\alpha$. Donc dans le cas d'un test bilatéral, la région critique peut se mettre sous la forme

$$(U_1 \leq u_{\alpha/2}) \text{ ou } (U_1 \geq u_{\alpha/2}^{'})$$

où $P_{H_0}(U_1 \leq u_{\alpha/2}) \leq \alpha/2$ et $P_{H_0}(U_1 \geq u_{\alpha/2}^{'}) = P_{H_0}(U_2 \leq u_{\alpha/2}) \leq \alpha/2$.

La p −valeur α_0 pour une valeur observée u de U_1 est donnée par

$$\alpha_0 = \begin{cases} P_{H_0}(U_1 \leq u) & si\ H_1: M_1 < M_2, \\ P_{H_0}(U_1 \geq u) & si\ H_1: M_1 > M_2, \\ 2\min\left(P_{H_0}(U_1 \leq u), P_{H_0}(U_1 \geq u)\right) & si\ H_1: M_1 \neq M_2. \end{cases}$$

Approximation normale :

Pour des tailles d'échantillons n, m assez grandes $(n, m \geq 6)$, on peut utiliser la statistique

$$Z = \frac{U_1 - E(U_1)}{\sqrt{Var(U_1)}} = \frac{U_1 - \frac{nm}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}$$

qui suit la loi normale centrée réduite.

Les régions de rejet et les p −valeurs sont données dans le tableau suivant où u est une valeur observée de U_1 :

Alternative H_1	Région de rejet	p −valeur α_0
$M_1 < M_2$	$Z \leq z_\alpha$	$\Phi\left(\frac{u - \frac{nm)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}\right)$
$M_1 > M_2$	$Z \geq z_{1-\alpha}$	$1 - \Phi\left(\frac{u - \frac{nm}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}\right)$
$M_1 \neq M_2$	$\lvert Z \rvert \geq z_{1-\alpha/2}$	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
$M_1 < M_2$	$U_1 \leq z_\alpha \sqrt{\frac{nm(n+m+1)}{12}} - 0.5 + \frac{nm}{2}$	$\Phi\left(\frac{u + 0.5 - \frac{nm)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}\right)$
$M_1 > M_2$	$W_1 \geq z_{1-\alpha} \sqrt{\frac{nm(n+m+1)}{12}} + 0.5 + \frac{nm}{2}$	$1 - \Phi\left(\frac{u - 0.5 - \frac{nm)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}\right)$
$M_1 \neq M_2$	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 7 individus du premier groupe et 8 du deuxième. Les performances en minutes sont les suivantes :

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

On réordonne les 15 observations par ordre croissant, on obtient

z_i	22	31	14	19	24	28	27	25	13	20	11	23	16	21	18
$z_{(i)}$	11 Y	13 Y	14 X	16 Y	18 Y	19 X	20 Y	21 Y	22 X	23 Y	24 X	25 Y	27 X	28 X	31 X

Le calcul direct du nombre de fois qu'un Y précède un X donne

$$u_1 = 7 + 7 + 6 + 6 + 5 + 5 + 4 + 3 = 43.$$

En effet

11 et 13 précèdent chacune 7 valeurs de X,
16 et 18 précèdent chacune 6 valeurs de X,

20 et 21 précèdent chacune 5 valeurs de X,
23 précède 4 valeurs de X,
25 e précède 3 valeurs de X.

Le nombre de fois qu'un X précède un Y est $u_2 = 6 + 4 + 2 + 1 = 13$.

On a aussi $u_2 = nm - u_1 = 56 - 43 = 13$.

Au seuil $\alpha = 5\%$, on a

$$u_\alpha = \begin{cases} 13 < u_1 & si\ H_1 : M_1 < M_2, \\ 13 = u_2 & si\ H_1 : M_1 > M_2, \\ 10 < \min(u_1, u_2) & si\ H_1 : M_1 \neq M_2, \end{cases}$$

Donc on accepte H_0 égalité des performances. Ceci est confirmé par la p −valeur :

$$\alpha_0 = \begin{cases} P_{H_0}(U_1 \leq 43) = 1 - P_{H_0}(U_1 \geq 42) = 1 - P_{H_0}(U_1 \leq 14) = 1 - 0.06 = 0.94 > 0.05 & si\ H_1 : M_1 < M_2, \\ P_{H_0}(U_1 \geq 43) = P_{H_0}(U_1 \leq 13) = 0.047 & si\ H_1 : M_1 > M_2, \\ 2\min\left(P_{H_0}(U_1 \leq 43), P_{H_0}(U_1 \geq 43)\right) = 0.094 > 0.05 & si\ H_1 : M_1 \neq M_2. \end{cases}$$

Remarque

Le test de Mann –Whitney est équivalent à celui de la somme des rangs de Wilcoxon puisqu'il existe une relation linéaire entre les statistiques des deux tests. En effet dans $U_1 = \sum_{i=1}^{n}\sum_{j=1}^{m} D_{ij}$, la somme $\sum_{j=1}^{m} D_{ij}$ représente le nombre de valeurs de j pour lesquelles $Y_j < X_i$ (le nombre de $Y_j < X_i$). Autrement dit, $\sum_{j=1}^{m} D_{ij}$ représente le rang de X_i réduit par n_i le nombre de X inférieures ou égales à X_i. Ainsi

$$U_1 = \sum_{i=1}^{n}(rg(X_i) - n_i) = \sum_{i=1}^{n} rg(X_i) - \sum_{i=1}^{n} n_i = \sum_{i=1}^{n} rg(X_i) - \sum_{i=1}^{n} i = W_1 - \frac{n(n+1)}{2}.$$

De la même manière, on obtient

$$U_2 = W_2 - \frac{m(m+1)}{2}.$$

On a également les relations suivantes :

$$U_1 = W_1 - \frac{n(n+1)}{2} = \frac{(n+m)(n+m+1)}{2} - W_2 - \frac{n(n+1)}{2} = nm + \frac{m(m+1)}{2} - W_2,$$

$$U_2 = W_2 - \frac{m(m+1)}{2} = nm + \frac{n(n+1)}{2} - W_1.$$

Cas d'ex-æquo

Dans le cas d'ex-æquo, on utilise souvent la statistique de Mann-Whitney standardisée U_T définie par

$$U_T = \sum_{i=1}^{n}\sum_{j=1}^{m} D_{ij}^*$$

où

$$D_{ij}^* = \begin{cases} 1 & si\ Y_j < X_i, \\ 0 & si\ Y_j = X_i, \\ -1 & si\ X_i < Y_j. \end{cases}$$

La statistique U_T est asymptotiquement normale. De plus on a

$$E_{H_0}(U_T) = 0,$$

$$Var_{H_0}(U_T) = \frac{nm(n+m+1)}{12} - \frac{nm}{12(n+m)(n+m-1)}\sum t_l(t_l^2 - 1).$$

Exemple

On dispose de deux échantillons de tailles respectives 10 et15 :

Echantillon 1 (X)	80	100	90	110	125	130	70	75	71	83					
Echantillon 2 (Y)	100	120	80	140	130	160	115	120	73	88	135	125	128	95	87.

Utiliser le test de Mann-Whitney pour tester au seuil 5% si les deux échantillons proviennent de la même population.

Quelle hypothèse faites-vous ?