Annexe : Détails d'application du test de Wilcoxon

L'objectif de cette annexe est de rappeler brièvement les détails d'application du test de Wilcoxon (dit de la somme des rangs), utilisé pour comparer les caractéristiques de deux distributions (échantillons). Ce test représente une alternative non paramétrique au test t de Student, reposant uniquement sur l'ordre des observations relatives aux deux échantillons. Le test de Wilcoxon peut apporter des réponses pertinentes à des questions d'intérêt comme "*La forme de deux distributions est-elle identique?*", "*Y a-t-il une différence significative entre les tendances centrales de deux distributions?*" etc.

Soit deux échantillons, A et B, comprenant n_A et n_B observations, respectivement. L'objectif est de tester l'hypothèse *nulle* selon laquelle la distribution de probabilité Ψ d'une variable X est identique dans les deux échantillons, i.e.:

$$\mathbf{H_0}\text{: } \Psi_A = \Psi_B$$

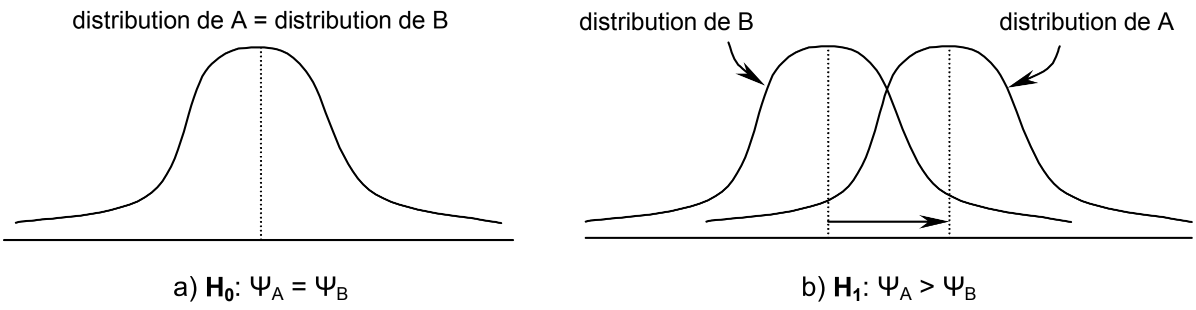
Si cette hypothèse est rejetée, une des hypothèses *alternatives* suivantes sera acceptée:

$\mathbf{H_1}$: $\Psi_A > \Psi_B$, si la distribution de A est déplacée vers la droite par rapport à celle de B (c'est le cas représenté à titre illustratif dans la figure 3);

$\mathbf{H_2}$: $\Psi_A < \Psi_B$, dans le cas contraire;

$\mathbf{H_3}$: $\Psi_A \neq \Psi_B$, l'alternative bilatérale, dans la situation où l'on ne dispose d'aucun a priori concernant la direction du déplacement.

Figure 3: Illustration de $\mathbf{H_0}$: $\Psi_A = \Psi_B$ *vs.* $\mathbf{H_1}$: $\Psi_A > \Psi_B$



Le test de Wilcoxon consiste dans un premier temps à classer dans l'ordre croissant l'ensemble de $n_A + n_B = n$ observations relatives aux deux échantillons réunis. Ensuite, à chaque observation on attribue un 'rang' r_i correspondant à son numéro d'ordre, compris entre 1 et n. Ainsi, le rang 1 est assigné à la plus petite observation, le rang 2 à la suivante etc. Dans le cas particulier où certaines observations sont *ex æquo* (i.e. leur rang est identique), on attribue à chacune d'entre elles un "rang moyen". Celui-ci est défini comme la moyenne des rangs que ces observations auraient si elles étaient différentes. Enfin, la statistique z de Wilcoxon est calculée comme la somme des rangs assignés aux observations provenant de l'un des deux échantillons (soit A par exemple):

$$z_A = \sum_{i=1}^{n_A} r_i$$

Notons z_A la somme des rangs *observés* et Z_A la variable aléatoire correspondante. Afin de déterminer le seuil critique p correspondant à la statistique calculée z_A, on doit d'abord inférer la distribution de probabilité de la somme des rangs, sous l'hypothèse $\mathbf{H_0}$ et $\mathbf{H_1}$, respectivement.

1. Supposons d'abord que l'hypothèse $\mathbf{H_0}$: $\Psi_A = \Psi_B$ soit vraie. Alors, l'ensemble de $n = n_A + n_B$ observations proviennent d'une même loi de probabilité. Par conséquent, les rangs attribués à ces n observations sont les nombres naturels $1, 2, \ldots, n$ et tout se passe comme si on avait tiré de manière aléatoire n_A observations de l'échantillon A et le reste de l'échantillon B. Dans ce premier cas la distribution de probabilité de la somme des rangs Z_A est connue et des tables spéciales en existent.

2. Si l'hypothèse $\mathbf{H_1}$: $\Psi_A > \Psi_B$ est vraie, l'échantillon A contient certainement les observations dont les rangs assignés sont les plus élevés. Par conséquent, la somme des rangs observée z_A est significativement plus importante que celle calculée sous l'hypothèse $\mathbf{H_0}$ Ainsi, le seuil p sera donnée par:

$$p = \text{Prob}\,[Z_A \geq z_A]$$

   où la probabilité est calculée en utilisant la distribution de Z_A sous l'hypothèse $\mathbf{H_0}$.

3. Si $\mathbf{H_1}$: $\Psi_A < \Psi_B$ est vraie, en appliquant un raisonnement *a pari*, on peut en déduire que

$$p = \text{Prob}\,[Z_A \leq z_A]$$

Bien évidemment, si on teste les alternatives unilatérales, la direction de l'inégalité utilisée dans le calcul du seuil critique p est la même que celle définissant l'hypothèse alternative, i.e. $Z_A > z_A \Leftrightarrow A > B$ et *vice versa*. En revanche, pour le test bilatéral, i.e. $\mathbf{H_0}$: $\Psi_A = \Psi_B$ *vs.* $\mathbf{H_1}$: $\Psi_A \neq \Psi_B$, une somme des rangs supérieure <u>ou</u> inférieure à la valeur tabulée signifie le rejet de $\mathbf{H_0}$. Dans ce cas, on calcule d'abord la probabilité que Z_A se trouve à proximité de la valeur observée z_A (soit à gauche, soit à droite) et ensuite on double cette probabilité. Par exemple, si z_A se situe à l'extrémité gauche de la queue de distribution, alors

$$p = 2 \cdot \text{Prob}\,[Z_A \geq z_A]$$

Dans le cas des échantillons de petite taille ($n_A, n_B < 12$ observations), certaines probabilités relatives à la distribution de Z_A, i.e. de la somme des rangs assignés dans l'échantillon A sous l'hypothèse $\mathbf{H_0}$: $\Psi_A = \Psi_B$, sont tabulées pour divers choix de n_A et n_B. Lorsque les observations relatives aux deux échantillons ne sont pas *appariées* (i.e. $n_A \neq n_B$), les tables de la loi de Wilcoxon sont conçues pour la somme des rangs calculée par rapport à l'échantillon le plus petit (noté A dans cette annexe).

Lorsque la taille des échantillons est suffisamment importante, la distribution de probabilité de Z_A (sous l'hypothèse $\mathbf{H_0}$) peut être approchée, avec une précision raisonnable, par une distribution normale $\mathcal{N}(\mu_A, \sigma_A^2)$, de moyenne

$$\mu_A = n_A(n_A + n_B + 1)/2$$

et variance

$$\sigma_A^2 = n_A n_B (n_A + n_B + 1)/12$$

En d'autres termes,

$$\text{Prob}[Z_A \leq z_A] \simeq \text{Prob}[Z_A^* \leq z] = \mathbf{\Phi}(z),$$

où $Z_A^* = (Z_A - \mu_A)/\sigma_A$, $z = (z_A - \mu_A)/\sigma_A$ et $\mathbf{\Phi}(\centerdot)$ représente la fonction de répartition de la loi normale centrée réduite.

Notes:

1. À la différence du test t de Student, le test de Wilcoxon peut être appliqué quelle que soit la distribution de probabilité caractérisant les données (i.e. normale ou non). De plus, il s'avère moins sensible aux valeurs aberrantes que le test classique de Student.

2. La distribution de Z_A est symétrique, i.e.

$$\text{Prob}[Z_A \leq z_A] = \text{Prob}[Z_A \geq n_A(n_A + n_B + 1) - z_A]$$