

# Chapitre 4: Notions sur les statistiques d'ordre et les distributions d'échantillonnage

Léonard Gallardo\*

## 1 Statistiques d'ordre d'un échantillon

### 1.1 Généralités

On a vu dans le chapitre 2 pourquoi la notion d'échantillon est à la base des méthodes statistiques. En particulier pour définir la fonction de répartition empirique, il convient d'ordonner l'échantillon par valeurs croissantes. Cet échantillon ordonné permet de déterminer (plus précisément d'estimer) la plupart des caractéristiques de la loi de probabilité d'où est issu l'échantillon. Nous allons illustrer cette idée avec quelques exemples. Signalons avant de commencer qu'on appelle distribution d'échantillonnage (ce nom figure dans le titre) toute loi de probabilité liée à un échantillon comme par exemple la loi des statistiques d'ordre ou des quantiles empiriques dont nous allons parler dans la suite ainsi que la loi de Wilcoxon, Mann et Whitney qui à la base d'un test fameux servant à comparer deux échantillons.

Soit  $(X_1, \dots, X_n)$  un  $n$ -échantillon issu d'une loi de probabilité de fonction de répartition  $F$ . Pour chaque  $\omega \in \Omega$ , on a une valeur observée

$$(1) \quad (X_1(\omega), \dots, X_n(\omega))$$

de ce  $n$ -échantillon. Ordonnons cette suite de valeurs par ordre croissant

$$(2) \quad X_{k_1(\omega)}(\omega) \leq X_{k_2(\omega)}(\omega) \leq \dots \leq X_{k_n(\omega)}(\omega),$$

où  $k_1(\omega)$  est le numéro de la plus petite valeur, etc...,  $k_n(\omega)$  est le numéro de la plus grande valeur. On note

$$(3) \quad X_{(i)}(\omega) = X_{k_i(\omega)}(\omega),$$

la  $i$ -ième valeur rangée et le vecteur  $(X_{(1)}, \dots, X_{(n)})$  s'appelle l'échantillon rangé.

**Définition 1.1** : Pour tout  $i = 1, \dots, n$ , la variable aléatoire  $X_{(i)}$  s'appelle la  $i$ -ième statistique d'ordre de l'échantillon.

**Exemple** :  $X_{(1)} = \min_{1 \leq i \leq n} X_i$  et  $X_{(n)} = \max_{1 \leq i \leq n} X_i$  sont respectivement la plus petite et la plus grande valeur de l'échantillon. La fonction de répartition de  $X_{(n)}$  est donnée par

$$F_{(n)}(x) = \mathbb{P}(X_{(n)} \leq x) = \mathbb{P}(\cap_{1 \leq i \leq n} [X_i \leq x]) = \prod_{1 \leq i \leq n} \mathbb{P}([X_i \leq x]) = (F(x))^n.$$

---

\*cours de Statistiques, Master 1, Université de Tours, année 2007-2008, Laboratoire de Mathématiques et Physique Théorique-UMR 6083, Parc de Grandmont, 37200 TOURS, FRANCE. email : gallardo@univ-tours.fr

Alors que la fonction de répartition de  $X_{(1)}$  est de la forme

$$\begin{aligned} F_{(1)}(x) &= \mathbb{P}(X_{(1)} \leq x) = 1 - \mathbb{P}(X_{(1)} > x) = 1 - \mathbb{P}(\cap_{1 \leq i \leq n} [X_i > x]) = 1 - \prod_{1 \leq i \leq n} \mathbb{P}([X_i > x]) \\ &= 1 - (1 - F(x))^n. \end{aligned}$$

## 1.2 Quantiles d'une loi et quantiles empiriques

**Définition 1.2** : Soit  $F$  une fonction de répartition continue et strictement croissante. Pour tout  $p \in ]0, 1[$ , on appelle quantile d'ordre  $p$  et on note  $q_p$  la racine (unique) de l'équation  $F(x) = p$ , i.e.  $q_p = F^{-1}(p)$ . En particulier :

si  $p = \frac{1}{2}$ ,  $q_{1/2}$  est appelée la médiane de la loi  $F$ .

si  $p = \frac{1}{4}$ ,  $q_{1/4}$  est appelée le (premier) quartile de la loi  $F$

**Remarque** : 1) On considère parfois les autres quartiles : le deuxième  $q_{1/2}$  (qui est la médiane) et le troisième quartile  $q_{3/4}$ . Ces trois quartiles partagent la distribution en parties de même « poids » en ce sens qu'étant donné une variable aléatoire  $X$  de loi  $F$ , on a

$$\mathbb{P}(X < q_{1/4}) = \mathbb{P}(q_{1/4} \leq X < q_{1/2}) = \mathbb{P}(q_{1/2} \leq X < q_{3/4}) = \mathbb{P}(X \geq q_{3/4}) = \frac{1}{4}.$$

Le quantile  $q_{1/10}$  s'appelle parfois (*premier*) *décile*, les autres déciles sont les quantiles  $q_{k/10}$  où  $k = 2, 3, \dots, 9$ ; ils partagent la distribution en 10 régions de même probabilité  $1/10$ .

2) On peut définir les quantiles pour une loi discrète ou une loi continue telle que  $F$  ne soit pas strictement croissante mais il y a forcément un certain arbitraire dans la définition. Par exemple dans le cas où  $F$  est continue, on a recours à l'inverse généralisée de la fonction de répartition pour définir le quantile d'ordre  $p$  par la formule

$$q_p = \inf\{x; F(x) \geq p\}.$$

Pour simplifier, nous n'aborderons cette question des quantiles que dans le cas considéré ici d'une fonction de répartition continue et strictement croissante.

**Définition 1.3** : Soit  $(X_1, \dots, X_n)$  un  $n$ -échantillon issu d'une loi  $F$  et  $(X_{(1)}, \dots, X_{(n)})$  l'échantillon ordonné. Soit  $p \in ]0, 1[$ . La statistique d'ordre  $X_{([np]+1)}$  (où  $[np]$  désigne la partie entière de  $np$ ) s'appelle le quantile empirique d'ordre  $p$  de l'échantillon. En particulier  $X_{([n/2]+1)}$  est la médiane empirique de l'échantillon.

**Exemple** : si  $(X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}, X_{(5)})$  est un échantillon rangé de taille  $n = 5$ , la médiane empirique est  $X_{(3)}$ ; c'est la valeur qui partage l'effectif de l'échantillon en 2 parties de même effectif (il y a deux valeurs plus petites et deux valeurs plus grandes). Noter que si on avait pris  $X_{([n/2])}$  comme définition de la médiane empirique, dans l'exemple précédent la médiane serait  $X_{(2)}$  ce qui n'est pas très logique. Si par contre on a un 4-échantillon rangé  $(X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)})$ , la médiane empirique est  $X_{(3)}$  avec notre définition.

Le théorème suivant montre que le quantile empirique est un estimateur du quantile théorique :

**Théorème 1.4** : Si  $F$  est continue et strictement croissante, on a

$$X_{([np]+1)} \rightarrow q_p \quad \mathbb{P}\text{-presque sûrement si } n \rightarrow +\infty.$$

Autrement dit  $X_{([np]+1)}$  est une estimation de  $q_p$  d'autant meilleure que la taille  $n$  de l'échantillon est grande.

**démonstration** : Soit  $F_n^\omega$  la fonction de répartition empirique de l'échantillon  $(X_1, \dots, X_n)$  ou de l'échantillon ordonné  $(X_{(1)}, \dots, X_{(n)})$ . Par définition de la fonction de répartition empirique, on a

$$(4) \quad F_n^\omega(X_{([np]+1)}) = \frac{1}{n}([np] + 1) \rightarrow p \quad \text{si } n \rightarrow +\infty.$$

Mais par le théorème fondamental de la statistique pour  $\mathbb{P}$ -presque tout  $\omega$ ,  $\sup_{x \in \mathbb{R}} |F(x) - F_n^\omega(x)| \rightarrow 0$  si  $n \rightarrow +\infty$ . Remplaçant  $x$  par  $X_{([np]+1)}(\omega)$  on déduit que  $|F(X_{([np]+1)}(\omega)) - F_n^\omega(X_{([np]+1)}(\omega))| \rightarrow 0$  d'où compte tenu de (4)

$$F(X_{([np]+1)}(\omega)) \rightarrow p \quad (n \rightarrow +\infty),$$

et puisque la fonction  $F^{-1}$  est continue, on a donc

$$X_{([np]+1)}(\omega) \rightarrow F^{-1}(p) = q_p \quad (n \rightarrow +\infty).$$

C'est le résultat annoncé. □

On a même un résultat qui précise la loi asymptotique de la variable aléatoire  $X_{([np]+1)} - q_p$  lorsque la loi  $F$  a une densité de probabilité :

**Théorème 1.5** : *Si la loi  $F$  a une densité de probabilité  $f$  strictement positive sur  $\mathbb{R}$ , alors en posant  $D := \frac{\sqrt{p(1-p)}}{f(q_p)}$ , la variable aléatoire  $\sqrt{n} \left( \frac{X_{([np]+1)} - q_p}{D} \right)$  converge en loi vers la loi normale  $\mathcal{N}(0, 1)$  quand  $n \rightarrow +\infty$ .*

**démonstration** : admis ; pour une preuve détaillée voir le livre de A. Rényi «Calcul des probabilités» p.460.

**Exercice** : Du théorème précédent déduire l'expression d'un intervalle de confiance approché du quantile  $q_p$  au niveau de confiance  $1 - \alpha$  (pour  $\alpha = 0,05$  et  $0,001$ ).

## 2 Le test du signe sur la médiane d'une loi

### 2.1 Tester si 0 est la médiane d'une loi

Si une variable aléatoire  $X$  est de loi continue  $F$ , 0 est la médiane de cette loi si

$$(5) \quad \mathbb{P}(X > 0) = \mathbb{P}(X < 0) = \frac{1}{2}$$

Si maintenant on considère  $(X_1, \dots, X_n)$  un  $n$ -échantillon de la loi  $F$ , alors sous l'hypothèse  $H_0$  : «0 est la médiane de la loi  $F$ », les variables aléatoires

$$(6) \quad S_1 = \sum_{i=1}^n \mathbf{1}_{[X_i > 0]} \quad \text{et} \quad S_2 = \sum_{i=1}^n \mathbf{1}_{[X_i < 0]}$$

«nombres de signes +» (resp. de signes -) dans l'échantillon, sont toutes les deux de loi binomiale  $\mathcal{B}(n, \frac{1}{2})$ .

Si l'hypothèse  $H_0$  était fausse, l'une des quantités  $S_1$  ou  $S_2$  prendrait une valeur supérieure à ce qu'elle devrait prendre normalement si  $H_0$  est vraie. Ceci nous conduit à construire un test de la manière suivante :

On se donne un risque de première espèce  $\alpha$  (par exemple 0,05 ou 0,01) et on détermine à l'aide des tables statistiques, la borne  $s_\alpha$  de la queue d'ordre  $\alpha$  de la loi  $\mathcal{B}(n, \frac{1}{2})$ , c'est à dire  $s_\alpha =$  le plus petit entier tel que

$$(7) \quad \mathbb{P}(S > s_\alpha) \leq \alpha \quad (\text{i.e.} \quad \sum_{k=s_\alpha+1}^n C_n^k \frac{1}{2^k} \leq \alpha),$$

**Le test «du signe»** consiste alors à rejeter  $H_0$  si on observe

$$(8) \quad S_1 > s_\alpha \quad \text{ou} \quad S_2 > s_\alpha.$$

On notera que la probabilité de rejeter à tort  $H_0$  est ici égale à  $2\alpha$ . Pour cette raison ce test s'appelle **le test du signe bilatéral**.

**Remarque :** On peut utiliser le test du signe, pour tester si la loi de  $X$  est symétrique, i.e. si

$$\forall u \in \mathbb{R}, \quad \mathbb{P}(X \geq u) = \mathbb{P}(X \leq u).$$

Si le test du signe donne un résultat de rejet, il est clair qu'il faut rejeter l'hypothèse que la loi est symétrique (car une loi symétrique a une médiane égale à 0).

## 2.2 Intervalle de confiance pour la médiane d'une loi continue

On a vu dans le paragraphe précédent comment tester si 0 est la médiane de la loi (continue) d'une variable aléatoire  $X$ . Si la loi n'est pas centrée, on peut tester si un nombre  $\xi$  est la médiane de  $X$  en faisant la transformation  $X - \xi = X'$  et en appliquant le test du signe bilatéral aux valeurs  $X'$ . On va montrer que cette idée a une application très intéressante qui via l'utilisation des statistiques d'ordre, permet de déterminer un intervalle de confiance pour la médiane d'une loi continue uniquement à partir d'un  $n$ -échantillon issu de cette loi.

**Théorème 2.1 :** Soit  $X = (X_1, \dots, X_n)$  un  $n$ -échantillon issu d'une loi  $F$  continue et  $\tilde{X} = (X_{(1)}, \dots, X_{(n)})$  l'échantillon rangé. Soit  $0 < \alpha < \frac{1}{2}$  (fixé) et soit  $s_\alpha$  le seuil d'ordre  $\alpha$  pour le test du signe défini en (7). Alors l'intervalle

$$(9) \quad [X_{(n-s_\alpha)}, X_{(s_\alpha+1)}]$$

est un intervalle de confiance de la médiane de  $F$  au niveau de confiance  $1 - 2\alpha$ .

**démonstration :** Notons tout d'abord que l'hypothèse faite sur  $\alpha$  nous assure déjà que  $s_\alpha$  est un entier plus grand que  $n/2$ , en particulier ceci implique  $n - s_\alpha < s_\alpha + 1$  donc l'intervalle (9) a un sens. Notons  $\xi$  la médiane de la loi  $F$  et posons

$$(10) \quad X'_i = X_i - \xi \quad (i = 1, \dots, n).$$

On suppose qu'on fait le test du signe sur les variables  $X'_i$  et qu'on travaille sous l'hypothèse  $H_0 :=$  "la médiane de  $F$  est  $\xi$ ".

Examinons les cas suivants qui peuvent se produire :

1) Si on observe

$$(11) \quad \xi < X_{(n-s_\alpha)},$$

i.e.  $0 < X_{(n-s_\alpha)} - \xi$ . Alors<sup>1</sup> on a aussi  $0 < X_{(i)} - \xi$  pour tout  $i \geq n - s_\alpha$ . Donc parmi les  $X'_i$  définies en (10), il y en a plus de  $s_\alpha$  qui sont  $> 0$ . On devrait donc rejeter l'hypothèse  $H_0$ .

<sup>1</sup>comme les  $X_{(i)}$  forment une suite croissante.

Autrement dit en termes d'événements, on a l'inclusion  $[\xi < X_{(n-s_\alpha)}] \subset [\text{on rejette } H_0]$ , ce qui implique qu'on a

$$(12) \quad \mathbb{P}_{H_0}(\xi < X_{(n-s_\alpha)}) \leq \alpha.$$

2) De même si

$$(13) \quad X_{(s_\alpha+1)} < \xi,$$

on voit (en utilisant le test bilatéral) qu'on devrait aussi rejeter  $H_0$  donc

$$(14) \quad \mathbb{P}_{H_0}(X_{(s_\alpha+1)} < \xi) \leq \alpha.$$

On déduit de (12) et (14) qu'on a

$$(15) \quad \mathbb{P}_{H_0}(X_{(s_\alpha+1)} \leq \xi \leq X_{(n-s_\alpha)}) \geq 1 - 2\alpha,$$

ce qui prouve le théorème. □

### 2.3 Le test du signe dans le cas des grands échantillons

Le nombre  $S$  de signes  $+$ , suit la loi  $\mathcal{B}(n, \frac{1}{2})$  et on a  $\mathbb{E}(S) = \frac{n}{2}$  et  $\text{Var}(S) = \frac{n}{4}$ . D'après le théorème limite central, si  $n$  est grand, la variable aléatoire

$$(16) \quad \mathcal{N} = \frac{1}{\sqrt{n/4}}(S - \frac{n}{2})$$

suit approximativement la loi  $\mathcal{N}(0, 1)$ . On utilisera alors la variable  $\mathcal{N}$  pour faire le test du signe. Précisément, la décision de rejet sera prise si

$$(17) \quad \mathcal{N} > \tilde{s}_\alpha,$$

où  $\tilde{s}_\alpha$  est la borne de la queue d'ordre  $\alpha$  de la loi  $\mathcal{N}(0, 1)$  i.e.

$$\int_{\tilde{s}_\alpha}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \alpha.$$

On considère généralement que dès que  $n \geq 50$ , l'approximation normale est bonne.

## 3 La loi et le test de Wilcoxon, Mann et Whitney

Nous allons à titre d'exemple montrer comment on peut comparer deux échantillons à partir de leurs valeurs rangées par ordre croissant.

### 3.1 La statistique de Wilcoxon, Mann et Whitney

Soient  $X = (X_1, \dots, X_m)$  un  $m$ -échantillon d'une loi  $F_0$  et  $Y = (Y_1, \dots, Y_n)$  un  $n$ -échantillon d'une loi  $F_1$ . On suppose que ces deux échantillons sont indépendants et que les lois  $F_0$  et  $F_1$  sont continues. Le problème est de déterminer si, dans de l'ensemble, les valeurs de  $Y$  ont tendance à être supérieures à celles de  $X$ . La méthode de Wilcoxon consiste à réunir les deux échantillons  $X$  et  $Y$  en un seul et à ordonner toutes les valeurs obtenues par ordre croissant mais en gardant en mémoire la nature de chaque valeur (i. e. si elle appartient à  $X$  ou  $Y$ ).

On a donc classé par ordre croissant les deux échantillons  $X$  et  $Y$  en un seul échantillon rangé de  $N = m + n$  valeurs. Soit  $Y_{(i)}$  la  $i$ -ème statistique d'ordre de  $Y$ .

**Définition 3.1** : On appelle nombre d'inversions de  $Y_{(i)}$  le nombre

$$(18) \quad I_{(i)} = \sum_{j=1}^m \mathbf{1}_{[Y_{(i)} > X_j]},$$

de valeurs  $X_j$  de l'échantillon  $X$  qui sont avant  $Y_{(i)}$  dans l'échantillon rangé (global). **Le nombre total d'inversions relativement à l'échantillon  $Y$**  est appelée statistique  $U$  de Wilcoxon-Mann-Whitney. Elle est égale à

$$U = \sum_{i=1}^n I_{(i)} = \sum_{i=1}^n \sum_{j=1}^m \mathbf{1}_{[Y_{(i)} > X_j]},$$

Il est clair<sup>2</sup> qu'on a aussi

$$(19) \quad U = \sum_{i=1}^n \sum_{j=1}^m \mathbf{1}_{[Y_i > X_j]} = \sum_{(i,j) \in [1,n] \times [1,m]} \mathbf{1}_{[Y_i > X_j]},$$

**Sous l'hypothèse  $H_0 : F_0 = F_1$  d'égalité des lois des échantillons  $X$  et  $Y$** , on a pu déterminer la loi de  $U$ . Les lois de  $U = U_{m,n}$  sont tabulées suivant les valeurs de  $m$  et  $n$ .

**Exemple** : Supposons qu'on ait un  $m = 5$ -échantillon de valeurs  $X$  et un  $n = 4$  échantillon de valeurs  $Y$ . On range les 9 valeurs par ordre croissant et on obtient :

$$xxyxyxyy.$$

On notera qu'il est inutile de mentionner les valeurs exactes observées mais seulement l'ordre dans lesquelles apparaissent les valeurs de l'échantillon  $X$  ou  $Y$ . Il y a

$$\begin{aligned} I_{(1)} &= 2 && \text{valeurs de } X \text{ qui précèdent le premier } y, \\ I_{(2)} &= 3 && \text{valeurs de } X \text{ qui précèdent le deuxième } y, \\ I_{(3)} &= 5 && \text{valeurs de } X \text{ qui précèdent le troisième } y, \\ I_{(4)} &= 5 && \text{valeurs de } X \text{ qui précèdent le quatrième } y, \end{aligned}$$

Donc  $U = 2 + 3 + 5 + 5 = 15$  inversions.

**Remarque** : Si on considère  $V$  = **le nombre total d'inversions relativement à l'échantillon  $X$** , c'est à dire

$$(20) \quad V = \sum_{(i,j) \in [1,n] \times [1,m]} \mathbf{1}_{[X_j > Y_i]},$$

la somme  $U + V$  est constante, plus exactement on a

$$(21) \quad U + V = mn,$$

car pour tout couple  $(i, j) \in [1, n] \times [1, m]$ , on a  $\mathbf{1}_{[X_j > Y_i]} + \mathbf{1}_{[Y_i > X_j]} = 1$  puisque, la loi  $F$  étant continue, on a  $\mathbb{P}([Y_i = X_j]) = 0$  donc  $\mathbf{1}_{[X_j = Y_i]} = 0$ . Ainsi dans l'exemple précédent on avait trouvé  $U = 15$  ; on voit aussi que  $V = 0 + 0 + 1 + 2 + 2 = 5$  d'où  $U + V = 15 + 5 = 20 = mn$ .

Voici quelques précisions sur la loi  $U$  :

<sup>2</sup>puisque l'on fait la somme sur toutes les valeurs de  $i$

**Théorème 3.2** : La loi de la variable aléatoire  $U$  est concentrée sur les entiers  $0, 1, \dots, mn$  et sous l'hypothèse  $H_0$  elle est symétrique par rapport à  $\frac{mn}{2}$ , i.e.

$$(22) \quad \forall k = 0, 1, \dots, mn, \mathbb{P}(U = k) = \mathbb{P}(U = mn - k).$$

De plus on a

$$(23) \quad \mathbb{E}(U) = \frac{mn}{2} \quad \text{et} \quad \text{Var}(U) = \frac{1}{12}mn(m + n + 1).$$

**démonstration** : Le fait que  $U$  est concentrée sur  $0, 1, \dots, mn$  est clair car la configuration qui donne le nombre maximum d'inversions est celle où les valeurs de  $X$  occupent les  $m$  premières places du rangement et les  $Y$  les  $n$  dernières places (par exemple  $xxxxxyyyy$  lorsque  $m = 5$  et  $n = 4$ ) et dans ce cas  $U = mn$ .

D'autre part sous l'hypothèse  $H_0$ , toutes les variables composant les échantillons  $X$  et  $Y$  sont indépendantes et de même loi  $F$ . Par symétrie des événements  $[X_j > Y_i]$  et  $[Y_i > X_j]$ , il est clair d'après les expressions (19) et (20) que les variables aléatoires  $U$  et  $V$  ont la même loi. Ainsi d'après la relation (21), pour tout  $k = 1, \dots, mn$ , on a

$$\mathbb{P}(U = k) = \mathbb{P}(V = mn - k) = \mathbb{P}(U = mn - k),$$

ce qui prouve que la loi de  $U$  est symétrique par rapport à la valeur  $\frac{mn}{2}$ .

De plus par l'argument de symétrie déjà utilisé, on a  $\mathbb{P}(X_j > Y_i) = \mathbb{P}(Y_i > X_j) = \frac{1}{2}$ . D'où par l'additivité de l'espérance mathématique, il vient immédiatement

$$(24) \quad \mathbb{E}(U) = \sum_{(i,j) \in [1,n] \times [1,m]} \mathbb{E}(\mathbf{1}_{[Y_i > X_j]}) = \frac{1}{2}mn,$$

puisque tous les termes de la somme valent  $\frac{1}{2}$ .

Le calcul de la variance est plus délicat. Posons  $T_{ij} = \mathbf{1}_{[Y_i > X_j]}$  pour alléger les notations. On a donc  $U - \mathbb{E}(U) = \sum_{(i,j)} (T_{ij} - \frac{1}{2})$ . D'où

$$(25) \quad \text{Var}(U) = \mathbb{E}((U - \mathbb{E}(U))^2) = \sum_{(i,j)} \sum_{(k,l)} \mathbb{E} \left( (T_{ij} - \frac{1}{2})(T_{kl} - \frac{1}{2}) \right).$$

On doit alors distinguer les cas  $i = k$  et  $j = l$  ( $mn$  cas) pour lesquels on peut montrer que  $\mathbb{E}((T_{ij} - 1/2)(T_{kl} - 1/2)) = 1/4$

et les cas  $i = k$  et  $j \neq l$  ou  $i \neq k$  et  $j = l$  ( $mn(n-1) + nm(m-1)$  cas) pour lesquels on peut voir que

$$\mathbb{E}((T_{ij} - 1/2)(T_{kl} - 1/2)) = 1/12.$$

Dans tous les autres cas, on a  $\mathbb{E}((T_{ij} - 1/2)(T_{kl} - 1/2)) = \mathbb{E}(T_{ij} - 1/2)\mathbb{E}(T_{kl} - 1/2) = 0$

D'où  $\text{Var}(U) = \frac{1}{4}mn + \frac{1}{12}(mn(n-1) + nm(m-1)) = \frac{1}{12}mn(m+n+1)$ .

### 3.2 Le test de Wilcoxon, Mann et Whitney

Comme nous l'avons déjà dit, les tables de la loi  $U$  sont fabriquées sous l'hypothèse  $H_0$  que les lois  $F_0$  et  $F_1$  coïncident. Considérons l'hypothèse alternative  $H_1$  : «les valeurs de

$Y$  sont globalement supérieures aux valeurs prises par  $X$ ». Ceci peut être traduit sur les fonctions de répartition par la condition :

$$F_1(x) \leq F_0(x)$$

(On pourra dessiner les fonctions de répartition pour se convaincre qu'une variable aléatoire  $Y$  de fonction de répartition  $F_1$  prendra en général des valeurs supérieures à une variable aléatoire  $X$  de fonction de répartition  $F_0$ ).

Si c'est l'hypothèse  $H_1$  qui est vraie, la variable aléatoire  $U$  qu'on a considéré au paragraphe précédent, aura tendance à prendre des valeurs plus grandes que celles qu'elle prendrait sous l'hypothèse  $H_0$ . Ceci nous conduit au

**Test de Wilcoxon, Mann et Whitney :**

Etant donné le risque de première espèce  $\alpha$  (en général égal à 0,05 ou 0,01), on détermine le seuil de rejet

$$u_\alpha = \inf \{t; \mathbb{P}_{H_0}(U > t) \leq \alpha\}$$

(on doit définir ainsi le seuil car la loi de  $U$  est discrète).

Alors on décide de rejeter  $H_0$  au profit de  $H_1$  si on observe  $U > u_\alpha$ .

On pourra faire facilement la vérification du test qui montre que  $\mathbb{P}_{H_0}(\text{rejeter } H_0) \leq \alpha$ .

**Remarque :** Une des principales applications du test concerne les comparaisons de performances entre deux produits : un ancien et un nouveau dont on veut savoir s'il est équivalent ou meilleur que l'ancien.

**Exemple :** On veut comparer les performances de deux types de carburants automobiles et tester l'hypothèse  $H_0$  : «les carburants A et B sont équivalents» contre l'hypothèse  $H_1$  : «le carburant B est supérieur à A». Voici les performances des deux types d'essence en kilomètres parcourus avec 10 litres de carburant

X : essence A	170	178	152	168	184	162	183	181	173	
Y : essence B	186	188	171	195	176	190	157	198	175	180

L'échantillon rangé se présente comme suit :

152, 157, 162, 168, 170, 171, 173, 175, 176, 178, 180, 181, 183, 184, 186, 188, 190, 195, 198,

où on a souligné les valeurs provenant de l'échantillon  $Y$ . La première valeur de  $Y$  a un nombre d'inversions égal à 1, la deuxième un nombre d'inversions égal à 4, etc. ... Dans ce cas on a

$$(26) \quad U = 1 + 4 + 5 + 5 + 6 + 9 + 9 + 9 + 9 + 9 = 66.$$

Au risque  $\alpha = 0,05$ , la table indique  $\mathbb{P}(U \leq 24) \leq 0,05$  ce qui équivaut à  $\mathbb{P}(U \geq 66) \leq 0,05$ , i.e.  $\mathbb{P}(U > 65) \leq 0,05$  donc  $u_\alpha = 65$ . On rejette donc l'hypothèse que les carburants sont équivalents au profit de l'hypothèse que le carburant B est supérieur.

### 3.3 Comportement asymptotique de la loi de $U$

Pour les grandes valeurs de  $m$  et  $n$  la loi  $U = U_{m,n}$  n'est pas tabulée. La raison est la suivante : Posons

$$(27) \quad \tilde{U}_{m,n} = \frac{U - \mathbb{E}(U)}{\text{Var}(U)},$$

où  $\mathbb{E}(U)$  et  $\text{Var}(U)$  sont données dans le théorème précédent. On a alors

**Théorème 3.3** :  $\tilde{U}_{m,n} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$  si  $m \rightarrow +\infty$  et  $n \rightarrow +\infty$ .

**démonstration** : résultat admis.

**Remarque** : Les statisticiens admettent que si  $m \geq 10$  et  $n \geq 10$ , l'approximation de  $\tilde{U}_{m,n}$  par la loi normale  $\mathcal{N}(0, 1)$  est suffisamment exacte. Le seuil de rejet pour le test de Wilcoxon, Mann et Whitney est calculé à l'aide de la table de la loi  $\mathcal{N}(0, 1)$ . C'est le nombre  $\tilde{u}_\alpha$  tel que

$$(28) \quad \mathbb{P}(\mathcal{N}(0, 1) > \tilde{u}_\alpha) = \alpha,$$

et on rejette alors l'hypothèse  $H_0$  si  $\tilde{U}_{m,n} > \tilde{u}_\alpha$ .