Quelques tests non paramétriques[1]

1. Les tests non paramétriques sur un échantillon

Dans cette section nous nous intéressons à deux tests non paramétriques :
– le test du signe et
– le test des rangs signés.
Nous utiliserons de préférence le test des rangs signés dès que les conditions de son utilisation sont remplies, sa puissance étant alors supérieure à celle du test du signe.

1.1. Test du signe

Soit un échantillon indépendant et identiquement distribué $X_1, \ldots, X_n$ d'une loi continue F dont la valeur médiane est notée m_e et la moyenne μ. Le test du signe permet de tester les hypothèses suivantes :

Hypothèses :

$$\boxed{\mathcal{H}_0 : m_e = 0 \quad \text{ou de façon équivalente} \quad \mathbb{P}\left[X_i > 0\right] = \frac{1}{2}}$$

contre

$$\boxed{\mathcal{H}_1 : m_e \neq 0 \quad \text{ou de façon équivalente} \quad \mathbb{P}\left[X_i > 0\right] \neq \frac{1}{2}.}$$

Remarque 1.1. La formulation de ce test est bien sûr la formulation d'un test bilatéral. Nous pouvons envisager les deux tests unilatéraux correspondants. À ce moment là, la formulation de l'hypothèse alternative $\mathcal{H}_1$ est différente et s'écrit soit :

$$\boxed{\mathcal{H}_1' : \mathbb{P}\left[X_i > 0\right] < \frac{1}{2}}$$

soit

[1]Les références [2], [3] et [1] ayant servi à l'élaboration de ce document sont mentionnées dans la bibliographie.

$$\boxed{\mathcal{H}_1^{''} : \mathbb{P}\left[X_i > 0\right] > \frac{1}{2}.}$$

Remarque 1.2. Plus généralement ce test permet de tester l'hypothèse nulle

$$\boxed{\mathcal{H}_0 : m_e = m_0 \quad \text{ou de façon équivalente} \quad \mathbb{P}\left[X_i > 0\right] = p}$$

contre

$$\boxed{\mathcal{H}_1 : m_e \neq m_0 \quad \text{ou de façon équivalente} \quad \mathbb{P}\left[X_i > 0\right] \neq p.}$$

où m_0 est un nombre réel et p est une constante comprise entre 0 et 1,
ou encore, dans la version unilatérale, contre l'hypothèse alternative

$$\boxed{\mathcal{H}_1^{'} : m_e < m_0}$$

ou encore, dans la version unilatérale, contre l'hypothèse alternative

$$\boxed{\mathcal{H}_1^{''} : m_e > m_0}$$

Pour cela il suffit de considérer l'échantillon $Z_1, \ldots, Z_n$ avec $Z_i = X_i - m_0$ et de lui appliquer le test décrit ci-dessous.

Statistique : S_n désigne le nombre de variables X_i, $1 \leqslant i \leqslant n$, qui prennent une valeur positive.

Propriétés 1.1. *Lorsque l'hypothèse nulle $\mathcal{H}_0$ est vraie, la variable aléatoire S_n suit exactement une loi binomiale $\mathcal{B}(n, p)$ de paramètres n et p.*

Concrètement cette hypothèse nulle $\mathcal{H}_0$ signifie que l'effectif de l'échantillon considéré est faible devant celui de la population dont il est issu.

Remarque 1.3. Nous pourrons prendre comme taille limite des échantillons dont les effectifs sont inférieurs à une fraction de 1/10 de la population. Dans ce cas nous pouvons assimiler les tirages réalisés ici à des tirages avec remise.

Cas le plus souvent utilisé : p =1/2. Nous nous proposons de tester :

Hypothèses :

$$\boxed{\mathcal{H}_0 : \mathbb{P}\left[X_i > 0\right] = \frac{1}{2}}$$

contre

$$\boxed{\mathcal{H}_1 : \mathbb{P}\left[X_i > 0\right] \neq \frac{1}{2}.}$$

Statistique : S_n désigne le nombre de variables X_i, $1 \leqslant i \leqslant n$, qui prennent une valeur positive.

Propriétés 1.2. *Lorsque l'hypothèse nulle $\mathcal{H}_0$ est vraie, la variable aléatoire S_n a les trois propriétés suivantes :*

1. *La variable aléatoire S_n suit une loi binomiale $\mathcal{B}(n, 1/2)$ de paramètres n et $1/2$. De ce fait, découle les deux propriétés suivantes :*
2. $\mathbb{E}[S_n] = n/2$.
3. $\mathrm{Var}[S_n] = n/4$.

Cette distribution binomiale est symétrique. Pour n grand ($n \geqslant 40$), nous pouvons utiliser l'approximation normale avec correction de continuité :

$$\mathbb{P}_{\mathcal{H}_0}[S_n \leqslant h] = \mathbb{P}_{\mathcal{H}_0}[S_n \geqslant n-h] = \frac{\Phi(2h+1-n)}{\sqrt{n}}$$

où Φ est la fonction de répartition d'une loi normale centrée réduite.

Décision 1.1. *Pour un seuil α donné (= $\alpha = 5\% = 0,05$ en général), nous cherchons le plus grand entier $s_\alpha^\star$ tel que $\mathbb{P}[Y \leqslant s_\alpha^\star] \leqslant \alpha/2$ où Y suit une loi binomiale $\mathcal{B}(n, 1/2)$ de paramètres n et $1/2$. Alors nous décidons :*

$$\begin{cases} \mathcal{H}_1 \text{ est vraie si} & S_{n,obs} \notin]s_\alpha^\star, n - s_\alpha^\star[\\ \mathcal{H}_0 \text{ est vraie si} & S_{n,obs} \in]s_\alpha^\star, n - s_\alpha^\star[. \end{cases}$$

Remarque 1.4. Le niveau de signification réel du test est alors égal à $2\mathbb{P}[Y \leqslant s_\alpha^\star]$ qui est généralement différent de α.

Remarque 1.5. Pour voir un exemple, nous renvoyons à la feuille de travaux dirigés qui sera traitée lors de la première séance de travaux dirigés.

1.2. Test des rangs signés de Wilcoxon

Soit un échantillon indépendant et identiquement distribué $X_1, \ldots, X_n$ d'une loi continue F dont la valeur médiane est notée m_e et la moyenne μ.

Hypothèses : Le test des rangs signés permet de tester l'hypothèse nulle

$$\boxed{\mathcal{H}_0 : \text{La loi continue } F \text{ est symétrique en } 0}$$

contre

$$\boxed{\mathcal{H}_1 : \text{La loi continue } F \text{ n'est pas symétrique en } 0.}$$

De plus, si nous savons que la loi continue F est symétrique, alors le test des rangs signés de Wilcoxon devient

$$\boxed{\mathcal{H}_0 : \mu = \mu_0}$$

contre

$$\boxed{\mathcal{H}_1 : \mu \neq \mu_0.}$$

Ici μ_0 est un nombre réel et ce jeu d'hypothèses permet alors de s'intéresser à la moyenne de la loi continue F.

1.2.1. Cas où il n'y a pas d'ex æquo.

Soit $x_1, \ldots, x_n$ n réalisations de l'échantillon précédent. À chaque x_i nous attribuons le rang r_i^a qui correspond au rang de $|x_i|$ lorsque que les n réalisations sont classées par ordre croissant de leurs valeurs absolues.

Statistique : Nous déterminons alors la somme w des rangs r_i^a des seules observations positives. La statistique W_n^+ des rangs signés de Wilcoxon est la variable aléatoire qui prend pour valeur la somme w. Par conséquent, la statistique W_n^+ des rangs signés de Wilcoxon s'écrit

$$W_n^+ = \sum_{\substack{1 \leqslant i \leqslant n \\ X_i > 0}} R_i^a.$$

Propriétés 1.3. *Lorsque l'hypothèse nulle $\mathcal{H}_0$ est vraie, la variable aléatoire W_n^+ a les trois propriétés suivantes :*

1. *W_n^+ est symétrique autour de sa valeur moyenne $\mathbb{E}\left[W_n^+\right] = n(n+1)/4$.*
2. $\operatorname{Var}\left[W_n^+\right] = n(n+1)(2n+1)/24$.
3. *Elle est tabulée pour de faibles valeurs de n. Pour $n \geqslant 15$, nous avons l'approximation normale avec correction de continuité :*

$$\mathbb{P}\left[W_n^+ \leqslant w\right] = \Phi\left(\frac{w + 0,5 - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}\right)$$

où Φ est la fonction de répartition de la loi normale centrée réduite.

Décision 1.2.
- **Premier cas :** *Pour tester l'hypothèse nulle « $\mathcal{H}_0$: La loi continue F est symétrique en 0 » contre l'hypothèse alternative « $\mathcal{H}_1$: La loi continue F n'est pas symétrique en 0 » pour un seuil donné α, nous cherchons l'entier w_α tel que $\mathbb{P}\left[W_n^+ \leqslant w_\alpha\right] \approx \alpha/2$. Alors nous décidons :*

$$\left\{\begin{array}{ll} \mathcal{H}_1 \textit{ est vraie si} & W_{n,obs}^+ \notin]w_\alpha + 1, n(n+1)/2 - w_\alpha - 1[, \\ \mathcal{H}_0 \textit{ est vraie si} & W_{n,obs}^+ \in]w_\alpha + 1, n(n+1)/2 - w_\alpha - 1[. \end{array}\right.$$

- **Second cas :** *Pour tester l'hypothèse nulle « $\mathcal{H}_0$: $\mu = \mu_0$ », nous introduisons l'échantillon $Z_1, \ldots, Z_n$ avec $Z_i = X_i - \mu$, $1 \leqslant i \leqslant n$.*

1.2.2. Cas où il y a des ex æquo.

Les observations $x_1, \ldots, x_n$ peuvent présenter des ex æquo et *a fortiori* leurs valeurs absolues. Il s'agit en particulier du cas où la loi F est discrète. Deux procédures sont alors employées.

- *Méthode de départition des ex æquo*

  Nous départageons les ex æquo à l'aide d'une table de nombres aléatoires. À chacune des valeurs égales nous associons un entier au hasard puis nous affectons, par ordre croissant de ces entiers, un rang différent à chaque observation. Ainsi chacun des rangs des observations est différent et nous pouvons directement appliquer les résultats du paragraphe précédent.

- *Méthode des rangs moyens*

  En associant à l'observation X_i son rang moyen $R_i^{a\star}$ dans le classement des valeurs absolues et en sommant tous les rangs pour lesquels $X_i > 0$ nous obtenons la statistique :

$$W_n^{+\star} = \sum_{\substack{1 \leqslant i \leqslant n \\ X_i > 0}} R_i^{a\star}.$$

Les valeurs absolues observées $|x_1|, \ldots, |x_n|$ étant ordonnées puis regroupées en classes d'ex æquo, C_0 pour la première classe qui est constituée des nombres $|x_i|$ nuls, s'il en existe, et C_j, $1 \leqslant j \leqslant h$ pour les autres nombres, certaines classes C_j pouvant comporter un seul élément, si cet élément n'a pas d'ex æquo, notons d_j le nombre d'ex æquo de la classe C_j. Nous avons

$$d_0 + \sum_{j=1}^{h} d_j = n.$$

Sous l'hypothèse nulle $\mathcal{H}_0$ et si $n > 15$, il est d'usage d'utiliser l'approximation normale

$$\frac{W_n^{+\star} - m^\star}{\sigma^\star} \approx \mathcal{N}(0,1)$$

où

$$m^\star = \frac{1}{4}\left(n(n+1) - d_0(d_0+1)\right)$$

et

$$(\sigma^\star)^2 = \frac{1}{24}\left(n(n+1)(2n+1) - d_0(d_0+1)(2d_0+1)\right) - \frac{1}{48}\sum_{j=1}^{h}\left(d_j^3 - d_j\right).$$

Dans le cas où nous utilisons cette méthode des rangs moyens, nous ne pouvons pas utiliser les tables statistiques usuelles qui concernent la distribution de la variable aléatoire W_n^+.

2. Les tests non paramétriques sur deux échantillons

2.1. Les échantillons sont indépendants : Test de Mann-Whitney

Nous observons, de manière indépendante, une variable Y, continue, sur deux populations, ou sur une population divisée en deux sous-populations. Nous notons $\mathcal{L}_i$ la loi de Y sur la (sous-)population d'ordre i. Nous allons présenter le test :

Hypothèses :

$$\boxed{\mathcal{H}_0 : \text{Les deux lois } \mathcal{L}_i \text{ sont égales ou encore de façon équivalente : } \mathcal{L}_1 = \mathcal{L}_2}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Les deux lois } \mathcal{L}_i \text{ ne sont pas égales ou encore de façon équivalente : } \mathcal{L}_1 \neq \mathcal{L}_2.}$$

2.1.1. Cas où il n'y a pas d'ex aequo.

Statistique : Pour obtenir la statistique du test notée U en général, nous devons procéder à des étapes successives :

1. En se plaçant sous l'hypothèse nulle $\mathcal{H}_0$, nous classons par ordre croissant l'ensemble des observations des deux échantillons $(x_1, \ldots, x_{n_1})$ et $(y_1, \ldots, y_{n_2})$ de taille respective n_1 et n_2.
2. Nous affectons le rang correspondant.
3. Nous effectuons la somme des rangs pour chacun des deux échantillons, notés R_1 et R_2.
4. Nous en déduisons les quantités U_1 et U_2 qui se calculent ainsi :

$$U_{n_1} = n_1 \times n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{2.1}$$

et

$$U_{n_2} = n_1 \times n_2 + \frac{n_2(n_2+1)}{2} - R_2 = n_1 \times n_2 - U_1. \tag{2.2}$$

La plus petite des deux valeurs U_{n_1} et U_{n_2}, notée U_{n_1,n_2}, est utilisée pour tester l'hypothèse nulle $\mathcal{H}_0$.

Propriétés 2.1. *Lorsque l'hypothèse nulle $\mathcal{H}_0$ est vraie, la variable aléatoire U_{n_1,n_2} a les trois propriétés suivantes :*

1. $\mathbb{E}\left[U_{n_1,n_2}\right] = (n_1 \times n_2)/2$.
2. $\mathrm{Var}\left[U_{n_1,n_2}\right] = (n_1 \times n_2)(n_1 + n_2 + 1)/12$.

3. *La variable aléatoire U_{n_1,n_2} est tabulée pour de faibles valeurs de n. Pour $n \geqslant 20$, nous avons l'approximation normale :*

$$\mathbb{P}\left[U_{n_1,n_2} \leqslant u\right] = \Phi\left(\frac{u - (n_1 \times n_2)/2}{\sqrt{(n_1 \times n_2)(n_1 + n_2 + 1)/12}}\right)$$

où Φ est la fonction de répartition de la loi normale centrée réduite.

Décision 2.1.
- **Premier cas :** *Si les tailles n_1 ou n_2 sont inférieures à 20. Pour un seuil donné α ($= 5\% = 0,05$ en général), la table de Mann-Whitney nous fournit une valeur critique c. Alors nous décidons :*

$$\begin{cases} \mathcal{H}_1 \text{ est vraie si} & U_{n_1,n_2,obs} \leqslant c, \\ \mathcal{H}_0 \text{ est vraie si} & U_{n_1,n_2,obs} > c. \end{cases}$$

- **Second cas :** *Si les tailles n_1 et n_2 sont supérieures à 20, alors la quantité est décrite approximativement par une loi normale et nous utilisons alors le test de l'écart réduit :*

$$Z_{n_1,n_2} = \frac{U_{n_1,n_2} - (n_1 \times n_2)/2}{\sqrt{(n_1 \times n_2)(n_1 + n_2 + 1)/12}}.$$

Pour un seuil donné α ($= 5\% = 0,05$ en général), la table de la loi normale centrée réduite nous fournit une valeur critique c. Alors nous décidons :

$$\begin{cases} \mathcal{H}_1 \text{ est vraie si} & Z_{n_1,n_2,obs} \geqslant c, \\ \mathcal{H}_0 \text{ est vraie si} & Z_{n_1,n_2,obs} < c. \end{cases}$$

2.1.2. Cas où il y a des ex æquo.

Les observations $x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$ peuvent présenter des ex æquo. Il s'agit en particulier du cas où les lois F et G dont sont issus les deux échantillons sont discrètes. Deux procédures sont alors employées.

- *Méthode de départition des ex æquo*

  Nous départageons les ex æquo à l'aide d'une table de nombres aléatoires. À chacune des valeurs égales nous associons un entier au hasard puis nous affectons, par ordre croissant de ces entiers, un rang différent à chaque observation. Ainsi chacun des rangs des observations est différent et nous pouvons directement appliquer les résultats du paragraphe précédent.

- *Méthode des rangs moyens*

  Les valeurs absolues observées $x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$ étant ordonnées puis regroupées en h classes d'ex æquo C_j, $1 \leqslant j \leqslant h$, certaines classes C_j pouvant comporter un seul élément, si cet élément n'a pas d'ex æquo, notons d_j le nombre d'ex æquo de la classe

C_j. Nous avons $\sum_{j=1}^{h} d_j = n_1 + n_2$.
En associant à l'observation X_i son rang moyen $R_i^\star$ dans ce classement et en sommant tous les rangs de tous les X_i, on obtient la statistique :

$$U^\star_{n_1,n_2} = \sum_{i=1}^{n_2} R_i^\star.$$

Sous l'hypothèse nulle $\mathcal{H}_0$: « X et Y ont la même distribution » et pour $n_1 > 15$ et $n_2 > 15$, il est d'usage d'utiliser l'approximation normale

$$\frac{U^\star_{n_1,n_2} - m^\star}{\sigma^\star} \approx \mathcal{N}(0,1)$$

où

$$m^\star = \frac{1}{2}\left(n_1(n_1 + n_2 + 1)\right)$$

et

$$(\sigma^\star)^2 = \frac{1}{12}\left(n_1 n_2(n_1 + n_2 + 1)\right) - \frac{1}{12}\frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)} \sum_{j=1}^{h} \left(d_j^3 - d_j\right).$$

Dans le cas où nous utilisons cette méthode des rangs moyens nous ne pouvons pas utiliser les tables statistiques usuelles qui concernent la distribution de la variable aléatoire U_{n_1,n_2}.

2.2. Les échantillons sont indépendants : Test de la médiane de Mood

On considère deux échantillons indépendants $(X_1, \ldots, X_{n_1})$ et $(Y_1, \ldots, Y_{n_2})$.
$(X_1, \ldots, X_{n_1})$ est un échantillon indépendant et identiquement distribué d'une loi continue F et $(Y_1, \ldots, Y_{n_2})$ est un échantillon indépendant et identiquement distribué d'une loi continue G.

Après regroupement des $n_1 + n_2$ valeurs des deux échantillons, $n_1 \times M_N$ est le nombre d'observations X_i qui sont supérieures à la médiane des $N = n_1 + n_2$ observations.

Sous l'hypothèse nulle $\mathcal{H}_0$: « Les variables X et Y suivent la même loi continue c'est-à-dire $G = F$ », la variable $n_1 \times M_N$ peut prendre les valeurs $0, 1, \ldots, n_1$ selon la distribution hypergéométrique suivante :

$$\mathbb{P}\left[n_1 \times M_N = k\right] = \frac{C_{n_1}^k C_{n_2}^{N/2-k}}{C_N^{N/2}}.$$

Ainsi on a :

$$\mathbb{E}\left[n_1 \times M_N\right] = \frac{n_1(n_1+n_2-\epsilon_N)}{2N}$$
$$\mathrm{Var}\left[n_1 \times M_N\right] = \frac{n_1 n_2 (n_1+n_2+1)}{4(n_1+n_2-1+\epsilon_N)(n_1+n_2+1-\epsilon_N)},$$

où $\epsilon_N = 0$ si N est pair et $\epsilon_N = 1$ si N est impair.

Lorsque n_1 et n_2 sont grands, c'est-à-dire $n_1 \geqslant 25$ et $n_2 \geqslant 25$, on utilise l'approximation normale :

$$\frac{n_1 \times M_N - \mathbb{E}\left[n_1 \times M_N\right]}{\sqrt{\mathrm{Var}\left[n_1 \times M_N\right]}} \approx \mathcal{N}(0,1)$$

avec correction de continuité.

La distribution est symétrique lorsque N est pair.

Pour tester l'hypothèse nulle $\mathcal{H}_0$: « $G = F$ » contre $\mathcal{H}_1$: « $G \neq F$ » avec un niveau de signification égal à α, on cherche les entiers k_α et $k_\alpha^{'}$ tels que $\mathbb{P}\left[n_1 \times M_N \leqslant k_\alpha\right] \approx \alpha/2$ et $\mathbb{P}\left[n_1 \times M_N \geqslant n_1 - k_\alpha^{'}\right] \approx \alpha/2$, puis on rejette l'hypothèse nulle $\mathcal{H}_0$ si la réalisation de la statistique du test calculée à l'aide de l'échantillon n'est pas dans l'intervalle $[k_\alpha, k_\alpha^{'}]$. Cette statistique permet également de réaliser des tests unilatéraux.

2.3. Les échantillons sont dépendants : Test de Wilcoxon

Nous considérons deux variables aléatoires X et Y, de même nature, observées toutes les deux sur les mêmes unités d'un n-échantillon. Les observations se présentent alors sous la forme d'une suite de couples $(x_1, y_1), \ldots, (x_n, y_n)$. Ce test concerne les lois des deux variables. Pour ce faire nous testons :

Hypothèses :

$$\boxed{\mathcal{H}_0 \text{ : Les deux lois sont égales ou encore de façon équivalente } \mathcal{L}(X) = \mathcal{L}(Y)}$$

contre

$$\boxed{\mathcal{H}_1 \text{ : Les deux lois ne sont pas égales ou encore de façon équivalente } \mathcal{L}(X) \neq \mathcal{L}(Y).}$$

2.3.1. Cas où il n'y a pas d'ex aequo.

Statistique : Pour obtenir la statistique du test notée S^+ en général, nous devons procéder à des étapes successives :

1. Ce test suppose que la loi de la différence entre les deux variables étudiées est symétrique par rapport à 0.

2. Après avoir calculé les différences d_i, nous classons par ordre croissant les $|d_i|$ non nulles, c'est-à-dire les d_i sans tenir compte des signes.
3. Nous attribuons à chaque $|d_i|$ le rang correspondant.
4. Nous restituons ensuite à chaque rang le signe de la différence correspondante.
5. Enfin, nous calculons la somme S^+ des rangs positifs (P) et la somme S^- des rangs négatifs (M).

La somme S^+ des rangs positifs (P) permet de tester l'hypothèse nulle $\mathcal{H}_0$.

Décision 2.2.
- **Premier cas :** *Si* $n < 15$, *nous utilisons une table et nous comparons la valeur de* (S^+) *à la valeur critique* c *associée au seuil* α *du test.*
- **Second cas :** *Si* $n \geqslant 15$, *nous utilisons l'approximation normale avec correction de continuité :*

$$\mathbb{P}_{\mathcal{H}_0}\left[S^+ \leqslant h\right] \approx \Phi\left(\frac{h + 0,5 - n(n+1)/4}{\sqrt{(n(n+1)(2n+1))/24}}\right)$$

où Φ *est la fonction de répartition d'un loi normale centrée réduite.*

2.3.2. Cas où il y a des ex æquo.

Il se traite de la même manière que pour la statistique de Wilcoxon pour un échantillon, voir le paragraphe 1.2.

3. Les tests non paramétriques sur k échantillons : 1 facteur

3.1. Les échantillons sont indépendants : Test de Kruskal-Wallis

On suppose que l'on dispose de k échantillons **indépendants** et identiquement distribués $(X_{1,1}, \ldots, X_{1,n_1}), \ldots, (X_{k,1}, \ldots, X_{k,n_k})$. La distribution du $i-$ème échantillon est notée F_i. On admet **a priori** que $F_i(x) = G(x - \alpha_i)$ où G est une fonction de répartition inconnue mais continue de moyenne μ et les α_i sont des nombres réels. Ainsi on suppose que le seul paramètre qui diffère d'une distribution F_i à l'autre est un paramètre de position α_i. C'est pourquoi même lorsque vous effectuez un test de Kruskal-Wallis vous devez vous assurer que vous pouvez au moins supposer que les variances des variables sont égales d'un échantillon à l'autre à l'aide d'un test non paramétrique de Levene d'égalité des variances. Les hypothèses ci-dessus impliquent que l'on peut écrire, pour tout $1 \leqslant i \leqslant k$ la décomposition suivante :

$$X_{i,j} = \mu + \alpha_i + \epsilon_{i,j}, \quad 1 \leqslant j \leqslant n_i,$$

les $N = \sum_{i=1}^{k} n_i$ variables aléatoires $\epsilon_{i,j}$ étant indépendantes et ayant une même distribution inconnue et de moyenne nulle.

3.1.1. Cas où il n'y a pas d'ex æquo.

La variable KW_N de Kruskal-Wallis est utilisée pour tester l'hypothèse

$$\boxed{\mathcal{H}_0 : \alpha_1 = \ldots = \alpha_k = 0}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Il existe au moins un } i_0 \text{ tel que } \alpha_{i_0} \neq 0.}$$

On commence par calculer le rang $R_{i,j}$ de $X_{i,j}$ parmi les N valeurs, puis la somme des rangs associée à chaque échantillon : $R_{i,\bullet} = \sum_{j=1}^{n_i} R_{i,j}$ et enfin la moyenne des rangs de chaque échantillon : $\overline{R_{i,\bullet}} = \dfrac{R_{i,\bullet}}{n_i}$.

La statistique de Krukal-Wallis KW_N prend en compte l'écart entre la moyenne des rangs de chaque échantillon et la moyenne de tous les rangs, qui vaut $(N+1)/2$:

$$KW_N = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\overline{R_{i,\bullet}} - \frac{N+1}{2} \right).$$

Sous l'hypothèse nulle $\mathcal{H}_0$: « $X_1, \ldots, X_k$ ont la même distribution continue », qui dans notre cas est équivalente à $\mathcal{H}_0$: « $\alpha_1 = \ldots = \alpha_k = 0$ », il est possible de déterminer la distribution de KW_N bien que le calcul soit complexe.

– Si l'un des effectifs n_i, $1 \leqslant i \leqslant k$, est inférieur ou égal à 4, on utilise une table spécifique.
– Si $n_i \geqslant 5$, pout tout $1 \leqslant i \leqslant k$ on utilise l'approximation $KW_N \approx \chi^2_{k-1}$.

Pour un seuil de signification de α, on détermine c_α tel que $\mathbb{P}\left[KW_N \geqslant c_\alpha\right] \cong \alpha$ et l'on rejette l'hypothèse nulle $\mathcal{H}_0$ lorsque la valeur prise par KW_N est supérieure à c_α.

3.1.2. Cas où il y a des ex æquo.

• *Méthode de départition des ex æquo*

On départage les ex æquo à l'aide d'une table de nombres aléatoires. À chacune des valeurs égales on associe un entier au hasard puis on affecte, par ordre croissant de ces entiers, un rang différent à chaque observation. Ainsi chacun des rangs des observations est différent et on peut directement appliquer les résultats du paragraphe précédent.

- *Méthode des rangs moyens*

À chaque nombre appartenant à un groupe d'ex æquo on attribue le rang moyen du groupe auquel il appartient puis on détermine la somme $T = \sum_{l=1}^{h}(t_l^3 - t_l)$ où t_l désigne le nombre d'éléments du $l-$ème groupe d'ex æquo. Il est d'usage de substituer à KW_N la variable $KW_N^\star$ définie par :

$$KW_N^\star = \frac{KW_N}{1 - \dfrac{T}{N^3 - N}}.$$

Comparaisons multiples

Si l'on rejette l'hypothèse nulle $\mathcal{H}_0$: « $\alpha_1 = \ldots = \alpha_k = 0$ » d'absence de différence entre les distributions F_i des k échantillons, on peut être amené à se demander quelles sont les distributions qui sont différentes.

On décide que **deux distributions** F_i et $F_{i'}$ sont significativement différentes au seuil $\boldsymbol{\alpha}$ si :

$$\left|\overline{R_{i,\bullet}} - \overline{R_{i',\bullet}}\right| \geqslant \sqrt{\chi^2(k-1, 1-\alpha)}\sqrt{\frac{N(N+1)}{12}}\sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}},$$

où $\chi^2(k-1, 1-\alpha)$ est le $100(1-\alpha)$ quantile de la loi du χ^2 à $k-1$ degrés de liberté.

On décide qu'**au seuil global** $\boldsymbol{\alpha}$ deux distributions F_i et $F_{i'}$, parmi les $\boldsymbol{k(k-1)}$ **comparaisons** que l'on va faire, sont significativement différentes si :

$$\left|\overline{R_{i,\bullet}} - \overline{R_{i',\bullet}}\right| \geqslant u\left(1 - \frac{\alpha}{k(k-1)}\right)\sqrt{\frac{N(N+1)}{12}}\sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}},$$

où $u\left(1 - \dfrac{\alpha}{k(k-1)}\right)$ est le $100\left(1 - \dfrac{\alpha}{k(k-1)}\right)$ quantile de la loi normale centrée réduite. Il s'agit d'une application des inégalités de Bonferroni[2]. Cette procédure est plus puissante que la précédente.

On décide qu'**au seuil global** $\boldsymbol{\alpha}$ deux distributions F_i et $F_{i'}$, parmi les $\boldsymbol{k(k-1)}$ **comparaisons** que l'on va faire, sont significativement différentes si :

[2] On pourra consulter le cours sur les modèles d'analyse de la variance pour plus de détails sur les procédures de comparaisons multiples.

$$\left|\overline{R_{i,\bullet}} - \overline{R_{i',\bullet}}\right| \geqslant q(k,+\infty,1-\alpha)\sqrt{\frac{N(N+1)}{12}}\sqrt{\frac{1}{2}\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)},$$

où $q(k,+\infty,1-\alpha)$ est le $100(1-\alpha)$ quantile de la loi de l'étendue studentisée pour k moyennes et $+\infty$ degrés de liberté. Il s'agit d'une procédure analogue à celle de Tukey-Kramer[2] dans le cas paramétrique et valide asymptotiquement. Elle est généralement plus puissante que les deux approches précédentes.

3.2. Les échantillons sont indépendants : Test de Jonckheere-Terpstra

La statistique J_N de Jonckheere-Terpstra permet de raffiner l'approche de la statistique KW_N de Kruskal-Wallis : supposons que les k modalités du facteur pour lequel on a réalisé les expériences soient naturellement ordonnées. C'est par exemple le cas dans la situation suivante : vous souhaitez trouver la dose optimale d'engrais à utiliser pour améliorer un rendement. Vous allez donc réaliser des expériences avec des doses de plus en plus importantes d'engrais et les modalités de votre facteur explicatif seront donc naturellement ordonnées par la quantité croissante d'engrais utilisé.

La statistique J_N de Jonckheere-Terpstra permet de tester l'hypothèse :

$$\boxed{\mathcal{H}_0 : \alpha_1 = \ldots = \alpha_k = 0}$$

contre

$$\boxed{\mathcal{H}_1 : \alpha_1 \leqslant \ldots \leqslant \alpha_k = 0 \text{ et il existe au moins un } i_0 \text{ tel que } \alpha_{i_0} < \alpha_{i_0+1}.}$$

3.2.1. Cas où il n'y a pas d'ex æquo.

La statistique J_N est construite à l'aide de toutes les variables de Mann-Whitney $U_{i,j}$, associées à l'échantillon i et l'échantillon j, lorsque $1 \leqslant i < j \leqslant k$:

$$J_N = \sum_{1 \leqslant i < j \leqslant k} U_{i,j}.$$

Sous l'hypothèse nulle $\mathcal{H}_0$: « $\alpha_1 = \ldots = \alpha_k = 0$ » :

– L'espérance et la variance de la statistique J_N sont :

$$\begin{aligned}
\mathbb{E}[J_N] &= \frac{N^2 - \displaystyle\sum_{i=1}^{k} n_i^2}{4},\\
\mathrm{Var}[J_N] &= \frac{1}{72}\left(N^2(3+2N) - \sum_{i=1}^{k} n_i^2(3+2n_i)\right).
\end{aligned}$$

- Les valeurs critiques de la statistique J_N sont tabulées pour de faibles valeurs de k et des n_i.
- Lorsque $n_i \geqslant 5$, pour tout $1 \leqslant i \leqslant k$, on a l'approximation normale avec correction de continuité :

$$\frac{J_N - \mathbb{E}\left[J_n\right]}{\sqrt{\operatorname{Var}\left[J_N\right]}} \approx \mathcal{N}\left(0,1\right).$$

On cherche l'entier ϕ_α tel que $\mathbb{P}\left[J_N \geqslant \phi_\alpha\right] \approx \alpha$ puis on rejette l'hypothèse nulle $\mathcal{H}_0$ au seuil α si la valeur prise par la statistique J_N est supérieure ou égale à ϕ_α.

3.2.2. Cas où il y a des ex æquo.

On peut utiliser une méthode de départition des ex æquo ou des tests de Mann-Whitney basés sur des rangs moyens, l'inconvient de la seconde méthode étant qu'on ne peut utiliser les mêmes tables qu'en absence d'ex æquo.

3.3. Les échantillons ne sont pas indépendants : Test de Friedman

On se place ici dans le cas où les échantillons utilisés pour tester l'influence d'un facteur **ne sont pas indépendants**.

Individu	Facteur A		
	1	$\cdots$	n
1	$x_{1,1}$	$\cdots$	$x_{n,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$x_{1,k}$	$\cdots$	$x_{n,k}$

On construit alors le tableau des rangs :

Individu	Facteur A			Total
	1	$\cdots$	n	
1	$r_{1,1}$	$\cdots$	$r_{n,1}$	$n(n+1)/2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$n(n+1)/2$
k	$r_{1,k}$	$\cdots$	$r_{n,k}$	$n(n+1)/2$
Total	$r_{1,\bullet}$	$\cdots$	$r_{n,\bullet}$	$kn(n+1)/2$

Si on est en présence de répétitions $x_{i,j,k}$ on remplace $x_{i,j}$ par la moyenne $\overline{x_{i,j}}$ des valeurs pour chaque cas où il y a des répétitions.

On cherche alors à tester l'hypothèse :

$$\boxed{\mathcal{H}_0 : \text{Les niveaux du facteur ont tous la même influence}}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Les niveaux du facteur n'ont pas tous la même influence.}}$$

3.3.1. Cas où il n'y a pas d'ex æquo.

La statistique de Friedman $F_{k,n}$ est définie par :

$$F_{k,n} = \frac{12k}{n(n+1)} \sum_{i=1}^{n} \left(\frac{R_{i,\bullet}}{k} - \frac{n+1}{2} \right)^2 = \frac{12}{kn(n+1)} \sum_{i=1}^{n} R_{i,\bullet}^2 - 3k(n+1).$$

On admet que sous l'hypothèse nulle $\mathcal{H}_0$: « Les niveaux du facteur ont tous la même influence » les distributions pour chaque individu ne diffèrent que par un paramètre de position, ce que l'on peut vérifier par un test non paramétrique de Levene par exemple.

- Pour de petites valeurs de k on utilise une table spécifique. Il se peut que l'on vous fournisse une table du coefficient de concordance $W_{k,n}$ de Kendall car la statistique de Friedman $F_{k,n} = k(n-1)W_{k,n}$.
- Pour des valeurs de k assez grandes on utilise l'approximation asymptotique suivante :

$$F_{k,n} \approx \chi^2_{n-1}.$$

On rejettera l'hypothèse nulle $\mathcal{H}_0$ si la valeur prise par $F_{k,n}$ est trop grande.

3.3.2. Cas où il y a des ex æquo.

• *Méthode de départition des ex æquo*

On départage les ex æquo à l'aide d'une table de nombres aléatoires. À chacune des valeurs égales on associe un entier au hasard puis on affecte, par ordre croissant de ces entiers, un rang différent à chaque observation. Ainsi chacun des rangs des observations est différent et on peut directement appliquer les résultats du paragraphe précédent.

• *Méthode des rangs moyens*

Dans chaque classement présentant des ex æquo on attribue à chacun de ceux-ci le rang moyen du groupe d'ex æquo auquel il appartient et qui n'est pas nécessairement un entier. Lorsque le classement numéro m a h_m groupes d'ex æquo, on lui attribue la somme $T_m = \sum_{l=1}^{h_m} \left(t_{l,m}^3 - t_{l,m} \right)$ où $t_{l,m}$ désigne le nombre d'éléments du $l-$ème de ces h_m groupes. S'il n'y a pas d'ex æquo on a évidemment $T_m = 0$ puisque la répartition des n entiers du classement en classes de nombres égaux donne $h_m = n$ et $t_{l,m} = 1$ pour tout l. Alors la statistique de Friedman corrigée est définie par :

$$\begin{aligned} F^{\star}_{k,n} &= \frac{12k(n-1)}{\left(n^3-n\right) - \frac{1}{k}\sum_{m=1}^{k} T_m} \sum_{l=1}^{n} \left(\frac{R_{l,\bullet}}{k} - \frac{n+1}{2}\right)^2 \\ &= \frac{1}{1 - \frac{1}{(n^3-n)}\frac{1}{k}\sum_{m=1}^{k} T_m} \frac{12k}{n(n+1)} \sum_{l=1}^{n} \left(\frac{R_{l,\bullet}}{k} - \frac{n+1}{2}\right)^2. \end{aligned}$$

On en déduit que :

$$F^{\star}_{k,n} = \frac{F_{k,n}}{1 - \frac{1}{(n^3-n)}\frac{1}{k}\sum_{m=1}^{k} T_m}.$$

4. Les tests non paramétriques sur nk échantillons : 2 facteurs

4.1. Les échantillons sont indépendants : Test de Friedman

On se place ici dans le cas où les échantillons utilisés pour tester l'influence d'un facteur **sont indépendants**.

Facteur B	Facteur A		
	1	$\cdots$	n
1	$x_{1,1}$	$\cdots$	$x_{n,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$x_{1,k}$	$\cdots$	$x_{n,k}$

On construit alors le tableau des rangs :

Facteur B	Facteur A			Total
	1	$\cdots$	n	
1	$r_{1,1}$	$\cdots$	$r_{n,1}$	$n(n+1)/2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$n(n+1)/2$
k	$r_{1,k}$	$\cdots$	$r_{n,k}$	$n(n+1)/2$
Total	$r_{1,\bullet}$	$\cdots$	$r_{n,\bullet}$	$kn(n+1)/2$

Si on est en présence de répétitions $x_{i,j,k}$ on remplace $x_{i,j}$ par la moyenne $\overline{x_{i,j}}$ des valeurs pour chaque cas où il y a des répétitions.

On admet **a priori** que l'influence des couples de niveaux (A_i, B_j) des facteurs A et B, pour $1 \leqslant i \leqslant n, 1 \leqslant j \leqslant k$, se traduit par une décomposition de la forme :

$$X_{i,j} = \mu + \alpha_i + \beta_j + \epsilon_{i,j}, \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant k$$

avec $\sum_{i=1}^{n} \alpha_i = 0$ et $\sum_{j=1}^{k} \beta_j = 0$. Les $N = \sum_{i=1}^{n} k = n \times k$ variables aléatoires $\epsilon_{i,j}$ étant indépendantes et ayant une même distribution inconnue et de moyenne nulle.

4.1.1. Cas où il n'y a pas d'ex æquo.

La variable $F_{k,n}$ de Friedman est utilisée pour tester l'hypothèse

$$\boxed{\mathcal{H}_0 : \alpha_1 = \ldots = \alpha_n = 0}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Il existe au moins un } i_0 \text{ tel que } \alpha_{i_0} \neq 0.}$$

On commence par calculer le rang $R_{i,j}$ de $X_{i,j}$ parmi les n valeurs de la colonne i, puis la somme des rangs associée à chaque colonne : $R_{i,\bullet} = \sum_{j=1}^{k} R_{i,j}$ et enfin la moyenne des rangs de chaque colonne : $\overline{R_{i,\bullet}} = \dfrac{R_{i,\bullet}}{k}$.

La statistique de Friedman $F_{k,n}$ est définie par :

$$F_{k,n} = \frac{12k}{n(n+1)} \sum_{i=1}^{n} \left(\frac{R_{i,\bullet}}{k} - \frac{n+1}{2} \right)^2 = \frac{12}{kn(n+1)} \sum_{i=1}^{n} R_{i,\bullet}^2 - 3k(n+1).$$

– Pour de petites valeurs de k on utilise une table spécifique. Il se peut que l'on vous fournisse une table du coefficient de concordance $W_{k,n}$ de Kendall car $F_{k,n} = k(n-1)W_{k,n}$.
– Pour des valeurs de k assez grandes on utilise l'approximation asymptotique suivante :

$$F_{k,n} \approx \chi^2_{n-1}.$$

On rejettera l'hypothèse nulle $\mathcal{H}_0$ si la valeur prise par la statistique de Friedman $F_{k,n}$ est trop grande.
Si l'on voulait également tester l'influence du facteur B on aurait analysé les tableaux ci-dessous avec la même méthode.

Facteur A	Facteur B		
	1	$\cdots$	n
1	$x_{1,1}$	$\cdots$	$x_{n,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$x_{1,k}$	$\cdots$	$x_{n,k}$

Facteur A	Facteur B			Total
	1	$\cdots$	n	
1	$r_{1,1}$	$\cdots$	$r_{n,1}$	$n(n+1)/2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$n(n+1)/2$
k	$r_{1,k}$	$\cdots$	$r_{n,k}$	$n(n+1)/2$
Total	$r_{1,\bullet}$	$\cdots$	$r_{n,\bullet}$	$kn(n+1)/2$

Comme on a échangé le rôle du facteur A et du facteur B on teste maintenant :

$$\boxed{\mathcal{H}_0 : \text{Les niveaux du facteur } B \text{ ont tous la même influence}}$$

contre

$$\boxed{\mathcal{H}_1 : \text{Les niveaux du facteur } B \text{ n'ont pas tous la même influence.}}$$

On ne peut pas tester l'existence d'une interaction par cette méthode puisque le modèle utilisé ne comporte pas de terme d'interaction. Il existe d'autres tests pour étudier l'existence d'une interaction.

4.1.2. Cas où il y a des ex æquo.

• *Méthode de départition des ex æquo*

On départage les ex æquo à l'aide d'une table de nombres aléatoires. À chacune des valeurs égales on associe un entier au hasard puis on affecte, par ordre croissant de ces entiers, un rang différent à chaque observation. Ainsi chacun des rangs des observations est différent et on peut directement appliquer les résultats du paragraphe précédent.

• *Méthode des rangs moyens*

Dans chaque classement présentant des ex æquo on attribue à chacun de ceux-ci le rang moyen du groupe d'ex æquo auquel il appartient et qui n'est pas nécessairement un entier. Lorsque le classement numéro m a h_m groupes d'ex æquo, on lui attribue la somme $T_m = \sum_{l=1}^{h_m} \left(t_{l,m}^3 - t_{l,m}\right)$ où $t_{l,m}$ désigne le nombre d'éléments du $l-$ème de ces h_m groupes. S'il n'y a pas d'ex æquo on a évidemment $T_m = 0$ puisque la répartition des n entiers du classement en classes de nombres égaux donne $h_m = n$ et $t_{l,m} = 1$ pour tout l. Alors la statistique de Friedman corrigée est définie par :

$$\begin{aligned} F^{\star}_{k,n} &= \frac{12k(n-1)}{\left(n^3-n\right) - \dfrac{1}{k}\displaystyle\sum_{m=1}^{k} T_m} \sum_{l=1}^{n} \left(\frac{R_{l,\bullet}}{k} - \frac{n+1}{2} \right)^2 \\ &= \frac{1}{1 - \dfrac{1}{(n^3-n)} \dfrac{1}{k} \displaystyle\sum_{m=1}^{k} T_m} \frac{12k}{n(n+1)} \sum_{l=1}^{n} \left(\frac{R_{l,\bullet}}{k} - \frac{n+1}{2} \right)^2 . \end{aligned}$$

On en déduit que :

$$F^{\star}_{k,n} = \frac{F_{k,n}}{1 - \dfrac{1}{(n^3-n)} \dfrac{1}{k} \displaystyle\sum_{m=1}^{k} T_m}.$$

Table des matières

Références

[1] B. Falissard. *Comprendre et utiliser les statistiques dans les sciences de la vie*. Abrégés. Masson, Paris, $3^{\text{ème}}$ edition, 2005.

[2] S. Kotz, C. B. Read, and N. Balakrishnan, editors. *Encyclopædia Of Statistical Sciences*. Wiley-Interscience, 1st edition, 1996.

[3] G. Pupion and P.-C. Pupion. *Tests non paramétriques*. Statistique mathématique et probabilité. Economica, Paris, 1998.