Le test de Kruskal-Wallis et différents tests post hoc de comparaison par paires

1 - Le test de Wilcoxon Mann Whitney

Une variable *X* a été observée sur deux échantillons indépendants (deux groupes). Dans les deux populations parentes, la médiane est-elle la même ? Autrement dit, l'hypothèse nulle :

$$H_0 : \ \theta_1 = \theta_2$$

est-elle vérifiée ?

Méthode :
On calcule les rangs dans la réunion des deux échantillons. Notations :
Rangs : $r_{11}, \ldots, r_{n1,1}, r_{21}, \ldots, r_{n2,1}$.
Effectifs des deux échantillons : n_1, n_2 ; effectif total : $n = n_1 + n_2$.
Rangs moyens : $\overline{R_1}, \overline{R_2}$

Statistique de test (en l'absence d'ex æquo) :

$$Z = \frac{\overline{R_1} - \overline{R_2}}{E} \quad \text{avec} \quad E^2 = \frac{n(n+1)}{12}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Sous H_0, pour des effectifs suffisants ($n_1 \geq 10$ et $n_2 \geq 10$), Z suit une loi normale centrée réduite.

Prise en compte des ex æquo :
Dans la formule précédente, $\frac{n(n+1)}{12}$ représente la variance de la série (1, 2, …, *n*). Pour tenir compte des ex æquo, on remplace cette valeur par la "vraie" variance :

$$V = \frac{1}{n-1}\left(\sum_{i,j} r_{ij}^2 - n\left(\frac{n+1}{2}\right)^2\right)$$

et on calcule donc :

$$Z = \frac{\overline{R_1} - \overline{R_2}}{E'} \quad \text{avec} \quad E'^2 = V \ \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Autre formule donnant la même statistique de test :

On peut aussi calculer :

$$Z = \frac{\sum r_{i,1} - n_1 \dfrac{n+1}{2}}{E} \quad avec \quad E^2 = \frac{n_1 n_2 (n+1)}{12}$$

La correction pour les ex æquo est la même que précédemment. On obtient alors :

$$Z = \frac{\sum r_{i,1} - n_1 \dfrac{n+1}{2}}{E'} \quad \text{avec} \quad E'^2 = \frac{n_1 n_2}{n(n-1)}\left[\sum_{i,j} r_{ij}^2 - \frac{n(n+1)^2}{4}\right]$$

Mini-exemple :

On dispose des données suivantes :

Groupe	A	A	B	B
Rang	1	3	2	4

On a alors : $\overline{R_1} = 2$; $\overline{R_2} = 3$; $E^2 = \frac{4 \times 5}{12}\left(\frac{1}{2} + \frac{1}{2}\right) = \frac{5}{3}$; $Z = -\sqrt{\frac{3}{5}} = -0{,}7746$

2 - Le test de Kruskal Wallis

La situation est analogue à la précédente, mais la variable X a été observée sur k groupes indépendants G_1, G_2, …, G_k. L'hypothèse nulle est alors l'égalité de l'ensemble des médianes :

$$H_0 : \ \theta_1 = \theta_2 = ... = \theta_k .$$

On calcule les rangs sur la réunion de tous les échantillons. Notations :
Rangs : r_{11}, …, $r_{n1,1}$, r_{21}, …, $r_{n2,1}$, …, $r_{1,k}$, …, $r_{nk,k}$.
Effectifs des échantillons : n_1,n_2, …, n_k ; effectif total : $N = n_1 + n_2 + \ldots + n_k$.
Rangs moyens des groupes : $\overline{R_1}$, $\overline{R_2}$, ...,$\overline{R_k}$.
Rang moyen général : $\overline{R} = \dfrac{N+1}{2}$

Statistique de test (en l'absence d'ex æquo) :

(1) $$K = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(\overline{R_j} - \overline{R} \right)^2$$

Variante : $$K = \left(\frac{12}{N(N+1)} \sum n_j \overline{R_j}^2 \right) - 3(N+1)$$

Statistique de test (en présence d'ex æquo) :

Dans la formule (1) précédente, le facteur $\dfrac{12}{N(N+1)}$ représente l'inverse de la variance de la série des rangs (1, 2, …, N). S'il y a des ex æquo, on calcule cette variance à partir de la série des rangs effectivement observés par :

$$V = \frac{1}{N-1} \left(\sum_{i,j} r_{i,j}^2 - N \frac{(N+1)^2}{4} \right)$$

puis $$K = \frac{1}{V} \sum_{j=1}^{k} n_j \left(\overline{R_j} - \overline{R} \right)^2 .$$

Pour des effectifs suffisamment grands ($n_j \geq 10$ par exemple), K suit approximativement une loi du khi-2 à (k-1) ddl.

Remarque. Lorsque k=2, le test de Kruskal-Wallis est identique au test de Wilcoxon Mann Whitney bilatéral. Les deux statistiques de test vérifient la relation : $K = Z^2$

Exemples.

1. Considérons le mini-exemple suivant :

Groupe	A	A	B	B	C	C
Rang	1	3	2	4	5	6

On a alors :

$$K = \frac{12}{6 \times 7} \left(2(2 - 3{,}5)^2 + 2(3 - 3{,}5)^2 + 2(5{,}5 - 3{,}5)^2 \right) = \frac{26}{7} = 3{,}71$$

2. Pour un échantillon de 510 pêcheurs à pied, on a relevé la consommation moyenne de fruits de mer (variable dépendante) et le nombre de jours de pêche par an, catégorisé en 5 classes, numérotées de 1 à 5 (source : Cindie Picot, thèse en cours).

Les données sont disponibles dans le fichier Conso-Nbjours.csv.

Elles peuvent être importées avec R à l'aide de la commande :

```
Peche <- read.csv2(file.choose())  # Sélectionnez le fichier Conso-Nbjours.csv
Peche[1:5,1]
# [1] 0.08038356 0.16142466 0.17863014 0.19375343 0.24794521
Peche[1:5,2]
# [1] 1 1 1 1 3

> kruskal.test(Peche[,1],Peche[,2])

      Kruskal-Wallis rank sum test

data:  Peche[, 1] and Peche[, 2]
Kruskal-Wallis chi-squared = 279.2922, df = 4, p-value < 2.2e-16
```

3 - Tests post hoc

Plusieurs méthodes sont proposées pour faire des comparaisons par paires, une fois établie l'hypothèse H_1.

3.1 - Test de Steel-Dwass

Ce test est aussi appelé test de Steel-Dwass-Chritchlow-Fligner. La comparaison de deux groupes quelconques *U* et *V* est effectuée de la manière suivante.

- Soient $\sum r_u$ et $\sum r_v$ les sommes des rangs des observations dans les deux groupes, le protocole des rangs étant calculé sur la réunion des deux groupes *U* et *V*.
- Soient n_u et n_v les effectifs des deux groupes et $n=n_u+n_v$.
- La statistique calculée est la suivante (sans ex æquo) :

$$t = \frac{\sum r_u - \left(n_u \ \frac{n+1}{2} \right)}{E} \quad \text{avec} \quad E^2 = \frac{n_u n_v (n+1)}{12}$$

- La statistique $Q_{uv} = t\sqrt{2}$ suit alors une loi des écarts studentisés de Tukey de paramètres *k* (nombre de groupes) et $+\infty$ (ddl).

Remarques.
1) La statistique t ainsi calculée est la statistique *Z* du test de Wilcoxon Mann Whitney appliqué aux deux groupes *U* et *V*.
2) Dans le cas où il y a des ex æquo, on peut calculer directement l'erreur type à partir du protocole des rangs. On calcule alors E^2 par :

$$E^2 = \frac{n_u n_v}{n(n-1)} \left[\sum_{i,\, j=u,v} r_{ij}^2 - \frac{n(n+1)^2}{4} \right]$$

Ce test est notamment proposé par StatsDirect (http://www.statsdirect.com).

3.2 Test indiqué par Siegel et Castellan

Ce test est décrit dans Siegel et Castellan éd. 1988, équations 8.7 et 8.8. C'est aussi le test fait par Statistica (à partir de la version 7.1).

Comparaison de toutes les paires de groupes

Pour comparer les groupes *U* et *V*, avec des hypothèses nulle et alternative ayant la forme suivante :

$$H_0 : \ \theta_u = \theta_v \qquad H_1 : \ \theta_u \neq \theta_v$$

- on conserve le protocole des rangs calculé sur l'ensemble des groupes ;
- on calcule la statistique :

$$Z_{u,v} = \frac{\left|\overline{R_u} - \overline{R_v}\right|}{E} \quad \text{avec} \quad E^2 = \frac{N(N+1)}{12}\left(\frac{1}{n_u} + \frac{1}{n_v}\right)$$

N.B. *N* désigne ici l'effectif total des *k* groupes, et les rangs sont calculés sur l'ensemble des *k* groupes.
- pour obtenir un résultat au seuil α, la statistique $Z_{u,v}$ est comparée à la valeur critique de la loi normale centrée réduite au seuil unilatéral $\frac{\alpha}{k(k-1)}$ (car il y a $\frac{k(k-1)}{2}$ comparaisons possibles).
- ou, de manière alternative, on détermine $P(Z > Z_{uv})$ puis on calcule la p-value par : $p = k\,(k-1)\,P(Z > Z_{uv})$; cette dernière méthode est celle utilisée par Statistica.

Ce test peut donner des résultats notablement différents de ceux donnés par le test de Steel-Dwass. Voir ci-dessous l'exemple de "pêche à pied".

Comparaison des groupes avec un groupe contrôle

Pour comparer le groupe contrôle *C* et le groupe *U*, avec des hypothèses nulle et alternative ayant la forme suivante :

$$H_0 : \ \theta_c = \theta_u \qquad H_1 : \ \theta_c \neq \theta_u$$

- on conserve le protocole des rangs calculé sur l'ensemble des groupes ;
- on calcule la statistique :

$$Z_u = \frac{\left|\overline{R_c} - \overline{R_u}\right|}{E} \quad \text{avec} \quad E^2 = \frac{N(N+1)}{12}\left(\frac{1}{n_c} + \frac{1}{n_u}\right)$$

- pour obtenir un résultat au seuil α, la statistique Z_u est comparée à la valeur critique de la loi normale centrée réduite au seuil unilatéral $\frac{\alpha}{2(k-1)}$ (car il y a $(k-1)$ comparaisons possibles).

Pour comparer le groupe contrôle *C* et le groupe *U*, avec des hypothèses nulle et alternative ayant la forme suivante :

$$H_0 : \ \theta_c = \theta_u \qquad H_1 : \ \theta_c > \theta_u$$

- on conserve le protocole des rangs calculé sur l'ensemble des groupes ;
- on calcule la statistique :

$$Z_u = \frac{\overline{R_c} - \overline{R_u}}{E} \quad \text{avec} \quad E^2 = \frac{N(N+1)}{12}\left(\frac{1}{n_c} + \frac{1}{n_u}\right)$$

- pour obtenir un résultat au seuil (unilatéral) α, la statistique Z_u est comparée à la valeur critique de la loi normale centrée réduite au seuil unilatéral $\frac{\alpha}{k-1}$.

3-3 Test de Conover-Inman

La méthode utilisée par la procédure de Conover-Inman ressemble à celle de Siegel et Castellan. En outre, on pose :
- S^2 : variance de la série des rangs dans l'ensemble des groupes. On a : $S^2 = \frac{N(N+1)}{12}$ s'il n'y a pas d'ex æquo.
- *T* : valeur de la statistique de Kruskal-Wallis.

On calcule alors la statistique :

$$t_{u,v} = \frac{\left|\overline{R_u} - \overline{R_v}\right|}{E_{u,v}} \quad \text{avec} \quad E_{u,v}^2 = S^2\,\frac{N-1-T}{N-k}\left(\frac{1}{n_u} + \frac{1}{n_v}\right)$$

et on compare cette statistique à la valeur $t_{1-\frac{\alpha}{2}}$ d'une distribution *t* de Student à *N* - *k* degrés de liberté.

Remarques.
1. Cette méthode est notamment fournie par StatsDirect.
2. Selon l'aide en ligne de StatsDirect, ce test est simplement l'application du test LSD de Fisher sur les rangs.
3. Les différences les plus importantes par rapport à la procédure de Siegel et Castellan sont l'intervention de la valeur de la statistique de Kruskal-Wallis dans le calcul (rapport $\frac{N-1-T}{N-k}$) et l'évaluation du niveau de significativité.

4 - Programmes R implémentant ces tests et exemples

4.1 - Programme R pour le test de Steel-Dwass

```
# N.B. Les groupes doivent être désignés par 1, 2, 3, etc.
# source("http://aoki2.si.gunma-u.ac.jp/R/src/Steel-Dwass.R", encoding="euc-jp")

Steel.Dwass <- function(data, group)
{       OK <- complete.cases(data, group)
        data <- data[OK]
        group <- group[OK]
        n.i <- table(group)
        ng <- length(n.i)
        t <- combn(ng, 2, function(ij) {
                i <- ij[1]
                j <- ij[2]
                r <- rank(c(data[group == i], data[group == j]))
                R <- sum(r[1:n.i[i]])
                N <- n.i[i]+n.i[j]
                E <- n.i[i]*(N+1)/2
                V <- n.i[i]*n.i[j]/(N*(N-1))*(sum(r^2)-N*(N+1)^2/4)
                return(abs(R-E)/sqrt(V))
        })
        p <- ptukey(t*sqrt(2), ng, Inf, lower.tail=FALSE)
        result <- cbind(t, p)
        rownames(result) <- combn(ng, 2, paste, collapse=":")
        return(result)
}
```

4.2 - Programme R pour la procédure indiquée par Siegel et Castellan

```
Siegel.Castellan <- function(data, group)
{       OK <- complete.cases(data, group)
        data <- data[OK]
        group <- group[OK]
        n.i <- table(group)
        ng <- length(n.i)
        r <- rank(data)
        N <- length(data)
        t <- combn(ng, 2, function(ij) {
                i <- ij[1]
                j <- ij[2]
                rim <- mean(r[group == i])
                rjm <- mean(r[group == j])
                V <- (sum(r^2)-N*(N+1)^2/4)/(N-1)*(1/n.i[i]+1/n.i[j])
                return(abs(rim-rjm)/sqrt(V))
        })
         p1 <- 1 - pnorm(t)
         p <- p1 * ng * (ng - 1)
        result <- cbind(t, p)
        rownames(result) <- combn(ng, 2, paste, collapse=":")
```

```
        return(result)
}
```

4.3 - Programme R pour la procédure indiquée par Conover-Inman

```
Conover.Inman <- function(data, group)
{       OK <- complete.cases(data, group)
        data <- data[OK]
        group <- group[OK]
        n.i <- table(group)
        ng <- length(n.i)
        r <- rank(data)
        N <- length(data)
        T <- as.real(kruskal.test(data,group)$statistic)
        t <- combn(ng, 2, function(ij) {
                i <- ij[1]
                j <- ij[2]
                rim <- mean(r[group == i])
                rjm <- mean(r[group == j])
                V <- (sum(r^2)-N*(N+1)^2/4)/(N-1)*(1/n.i[i]+1/n.i[j])*(N-1-T)/(N-ng)
                return(abs(rim-rjm)/sqrt(V))
        })
         p <- (1 - pt(t, N-ng))*2
        result <- cbind(t, p)
        rownames(result) <- combn(ng, 2, paste, collapse=":")
        return(result)
}
```

4.4 - Comparaison des résultats obtenus sur les données "Pêche à pied"

```
> Peche <- read.csv2(file.choose())  # Sélectionnez le fichier Conso-Nbjours.csv

> Steel.Dwass(Peche[,1],Peche[,2])
            t            p
1:2  6.991718 2.718747e-11
1:3 10.199030 4.463097e-14
1:4 10.860478 4.984901e-14
1:5 10.301552 5.129230e-14
2:3  6.812099 9.620282e-11
2:4 10.242801 4.096723e-14
2:5  9.858654 4.762857e-14
3:4  6.304335 2.892627e-09
3:5  8.120584 6.583623e-14
4:5  4.350374 1.324461e-04
>

> Siegel.Castellan(Peche[,1],Peche[,2])
            t            p
1:2  4.060759 4.891342e-04
1:3  9.035646 0.000000e+00
1:4 12.429676 0.000000e+00
1:5 13.945588 0.000000e+00
2:3  5.179702 2.222402e-06
2:4  9.138862 0.000000e+00
2:5 10.875183 0.000000e+00
3:4  4.767805 1.862441e-05
3:5  6.813212 9.544365e-11
4:5  2.102772 3.548572e-01

> Conover.Inman(Peche[,1],Peche[,2])
            t            p
1:2  6.020950 3.339128e-09
1:3 13.397291 0.000000e+00
1:4 18.429672 0.000000e+00
1:5 20.677338 0.000000e+00
2:3  7.680024 8.304468e-14
2:4 13.550332 0.000000e+00
```

```
2:5 16.124801 0.000000e+00
3:4  7.069298 5.207390e-12
3:5 10.102054 0.000000e+00
4:5  3.117812 1.925887e-03
```

Le résultat de Siegel.Castellan() peut être retrouvé à l'aide de Statistica (menu Statistiques - Tests non paramétriques - Comparaison de plusieurs échantillons indépendants (groupes) - Comparaison multiple des rangs moyens de tous les groupes) :

Dépend. : Valeurs z des comp. multiples ; Conso (Conso-Nbjours dans Classeur1)
Conso Test de Kruskal-Wallis : H (4, N= 510) =279,2922 p =0,00

	1 R:76,500	2 R:166,29	3 R:263,82	4 R:355,45	5 R:401,67
1		4,06075	9,035634	12,42966	13,94557
2	4,06075		5,179695	9,13885	10,87517
3	9,03563	5,17970		4,76780	6,81320
4	12,42966	9,13885	4,767798		2,10277
5	13,94557	10,87517	6,813203	2,10277	

Dépend. : Valeurs p des Comp. Multiples (bilatéral) ; Conso (Conso-Nbjours dans Classeur1)
Conso Test de Kruskal-Wallis : H (4, N= 510) =279,2922 p =0,00

	1 R:76,500	2 R:166,29	3 R:263,82	4 R:355,45	5 R:401,67
1		0,000489	0,000000	0,000000	0,000000
2	0,000489		0,000002	0,000000	0,000000
3	0,000000	0,000002		0,000019	0,000000
4	0,000000	0,000000	0,000019		0,354860
5	0,000000	0,000000	0,000000	0,354860	

Les deux autres résultats peuvent être retrouvés à l'aide de StatsDirect 2.7.2 (http://www.statsdirect.com).

```
Kruskal-Wallis test

Variables: Conso_NbJoursAn_1, Conso_NbJoursAn_3, Conso_NbJoursAn_2,
Conso_NbJoursAn_5, Conso_NbJoursAn_4

Groups = 5
df = 4
Total observations = 510

T = 279,29143
P < 0,0001

Adjusted for ties:

T = 279,292176
P < 0,0001

At least one of your sample populations tends to yield larger observations than at
least one other sample population.


Kruskal-Wallis: all pairwise comparisons (Dwass-Steel-Chritchlow-Fligner)
Critical q (range) = 3,857656

Conso_NbJoursAn_1 vs. Conso_NbJoursAn_3       significant
(|-14,423606| > 3,857656)     P < 0,0001
Conso_NbJoursAn_1 vs. Conso_NbJoursAn_2       significant
(|-9,887783| > 3,857656)      P < 0,0001
Conso_NbJoursAn_1 vs. Conso_NbJoursAn_5       significant
(|-14,568594| > 3,857656)     P < 0,0001
Conso_NbJoursAn_1 vs. Conso_NbJoursAn_4       significant
(|-15,359036| > 3,857656)     P < 0,0001
Conso_NbJoursAn_3 vs. Conso_NbJoursAn_2       significant
```

```
(|-9,633763| > 3,857656)      P < 0,0001
Conso_NbJoursAn_3 vs. Conso_NbJoursAn_5       significant
(|11,48424| > 3,857656)       P < 0,0001
Conso_NbJoursAn_3 vs. Conso_NbJoursAn_4       significant
(|8,915675| > 3,857656)       P < 0,0001
Conso_NbJoursAn_2 vs. Conso_NbJoursAn_5       significant
(|13,942242| > 3,857656)      P < 0,0001
Conso_NbJoursAn_2 vs. Conso_NbJoursAn_4       significant
(|14,485507| > 3,857656)      P < 0,0001
Conso_NbJoursAn_5 vs. Conso_NbJoursAn_4       significant
(|6,152359| > 3,857656)       P = 0,0001
```

N.B. Les valeurs indiquées en première colonne (|-14,42|, etc) sont les valeurs t du programme R, multipliées par $\sqrt{2}$.

```
Kruskal-Wallis: all pairwise comparisons (Conover-Inman)
Critical t (505 df) = 1,964673

Conso_NbJoursAn_1 and Conso_NbJoursAn_3       significant
(187,319728 > 27,469958)      P < 0,0001
Conso_NbJoursAn_1 and Conso_NbJoursAn_2       significant
(89,785714 > 29,297712)       P < 0,0001
Conso_NbJoursAn_1 and Conso_NbJoursAn_5       significant
(325,174699 > 30,896807)      P < 0,0001
Conso_NbJoursAn_1 and Conso_NbJoursAn_4       significant
(278,94898 > 29,737102)       P < 0,0001
Conso_NbJoursAn_3 and Conso_NbJoursAn_2       significant
(97,534014 > 24,950829)       P < 0,0001
Conso_NbJoursAn_3 and Conso_NbJoursAn_5       significant
(137,854971 > 26,810458)      P < 0,0001
Conso_NbJoursAn_3 and Conso_NbJoursAn_4       significant
(91,629252 > 25,465333)       P < 0,0001
Conso_NbJoursAn_2 and Conso_NbJoursAn_5       significant
(235,388985 > 28,680271)      P < 0,0001
Conso_NbJoursAn_2 and Conso_NbJoursAn_4       significant
(189,163265 > 27,427003)      P < 0,0001
Conso_NbJoursAn_5 and Conso_NbJoursAn_4       significant
(46,225719 > 29,128977)       P = 0,0019
```

N.B. On lit en en-tête : "Critical t = 1,964673". L'inégalité en première colonne représente la comparaison entre le numérateur et le dénominateur multiplié par la valeur "Critical t". Autrement dit, on retrouve les valeurs t du programme R à l'aide de calculs du genre :

$$t = \frac{187,319728}{\frac{27,469958}{1,964673}} = 13,3973$$

5 - Conclusion

Les trois méthodes vues au paragraphe 4 fournissent des résultats notablement différents. Laquelle vaut-il mieux utiliser ? D'après l'aide en ligne de StatsDirect, en l'absence d'indications contraires, c'est le test de Steel-Dwass.