Test de Kruskal-Wallis

Le test de Kruskal-Wallis est la généralisation à K populations (K échantillons indépendants), $K \geq 3$, du test de Wilcoxon-Mann-Whitney bilatéral. L'objectif étant de comparer la distribution d'une v.a. X continue entre K groupes indépendants. La fonction de répartition de X dans la population i est $F_i, i = 1, \dots, K$. Les distributions F_i sont supposées être identiques à une translation près

$$F_i(x) = F(x - \theta_i), i = ,1, \dots, K.$$

Considérons $\left(X_{i1}, \dots, X_{in_i}\right)$ un échantillon de taille n_i issu de la population i, $i = ,1, \dots, K$. On veut tester :

H_0: *les distributions sont identiques* v.s . H_1: *au moins deux distributions sont différentes*,

$H_0: F_1 = F_2 = \cdots = F_K$ v.s . $H_1: \exists\, i, j, i \neq j, F_i \neq F_j$

$H_0: \theta_1 = \theta_2 = \cdots = \theta_K$ v.s. $H_1: \exists\, i, j, i \neq j, \theta_i \neq \theta_j$,

$H_0: M_1 = M_2 = \cdots = M_K$ v.s. $H_1: \exists\, i, j, i \neq j, M_i \neq M_j$.

On range l'ensemble des $N = \sum_{i=1}^{K} n_i$ observations en une suite unique croissante. Chaque observation est affectée du rang qu'elle occupe dans cette suite. On note

- R_{ij} le rang de $X_{ij}, j = 1, \dots, n_i, i = 1, \dots, K$,
- R_i la somme des rangs des observations de l'échantillon i

$$R_i = \sum_{j=1}^{n_i} R_{ij},$$

- $\bar{R}_i = \frac{R_i}{n_i}$ la moyenne des rangs des observations de l'échantillon i (appelée aussi moyenne conditionnelle),
- $\bar{R}$ la moyenne globale des rangs, $\bar{R} = \frac{1}{N}\sum_{i=1}^{K} R_i = \frac{N+1}{2}$.

Si H_0 est vraie alors

$$E(\bar{R}_i) = \frac{N+1}{2} \Leftrightarrow E(R_i) = \frac{n_i(N+1)}{2},$$

$$V(\bar{R}_i) = \frac{(N+1)(N-n_i)}{12 n_i} \Leftrightarrow V(R_i) = \frac{n_i(N+1)(N-n_i)}{12},$$

$$Cov(\bar{R}_i, \bar{R}_j) = -\frac{N+1}{12} \Leftrightarrow Cov(R_i, R_j) = -n_i n_j \frac{N+1}{12}.$$

Donc si l'hypothèse nulle est vérifiée, les moyennes conditionnelles des rangs sont proches de la moyenne globale.

Kruskal-Wallis (1952) ont proposé une statistique de test basée sur les écarts entre les moyennes conditionnelles et la moyenne globale des rangs. Cette statistique, notée H, est définie par

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{K} n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2 = \frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{1}{n_i} \left(R_i - \frac{n_i(N+1)}{2} \right)^2 .$$

L'hypothèse nulle est rejetée pour les grandes valeurs de H. La région critique sera donc de la forme $\{H \geq h_\alpha\}$ telle que $P_{H_0}(H \geq h_\alpha) \leq \alpha$.

Il est possible de simplifier l'expression de H pour obtenir la formule d'usage pratique

$$\boldsymbol{H = \frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{{R_i}^2}{n_i} - 3(N+1).}$$

Pour déterminer la distribution exacte de H sous H_0, les rangs des N observations sont enregistrés dans une table à K colonnes, où les entrées de la $i$ème colonne sont les n_i rangs affectés aux éléments (observations) du $i$ème échantillon. Ainsi R_i est la somme de la $i$ème colonne. Le nombre de façons d'affecter les rangs $1,2, \dots, N$ est égal au nombre de partitions de N éléments distincts en K groupes (ensembles), le $i$ème groupe contenant n_i éléments, $i = 1, \dots, K$. Le nombre de partitions possibles est égal à $\frac{N!}{\prod_{i=1}^{K} n_i!}$. On calcule la valeur de H pour chaque partition possible. Par suite $P(H = h) = t(h).\frac{\prod_{i=1}^{K} n_i!}{N!}$. où $t(h)$ désigne le nombre de partitions pour lesquelles $H = h$. Les calculs requis sont extrêmement fastidieux et ne seront donc pas illustré ici. Par exemple pour 3 échantillons de tailles $n_i = 2, i = 1,2,3$, il y a 90 partitions possibles des rangs 1,2,3,4,5,6 en 3 groupes , chaque groupe contenant 2 éléments.

La loi exacte de H est tabulée (pour les petits échantillons) pour $K = 3,4,5$ pour différentes combinaisons possibles en nombre d'observations par échantillon.

<u>Exemple :</u> Test de Kruskal-Wallis pour les petits échantillons

On veut comparer les performances de 3 réglages différents de machines outils dans la production de capsules : réglage 1, réglage 2 et réglage 3. La variable d'intérêt est le nombre de capsules produites dans une unité de temps. Les données relatives à chaque réglage sont données dans la table suivante :

Réglage 1	Réglage 2	Réglage 3
340	339	347
345	333	343
330	344	349
342		355
338		

Peut-on accepter au seuil 5% l'égalité des performances des 3 réglages ?

On a 3 groupes, $K = 3$, de tailles respectives $n_1 = 5, n_2 = 3$ et $n_3 = 4$ avec un total de $N = 12$ observations.

Les rangs des 12 observations sont les suivants :

	Réglage 1	Réglage 2	Réglage 3	
	5	**4**	**10**	
	9	**2**	**7**	
	1	**8**	**11**	
	6		**12**	
	3			
r_i	**24**	**14**	**40**	
$\frac{r_i^2}{n_i}$	**115.2**	**65.33**	**400**	$\sum_{i=1}^{3} \frac{r_i^2}{n_i} = 580.533$

La valeur de la statistique de Kruskal-Wallis est

$$H_{obs} = \frac{12}{N(N+1)} \sum_{i=1}^{3} \frac{r_i^2}{n_i} - 3(N+1) = \frac{12}{12(13)} \left(\frac{24^2}{5} + \frac{14^2}{3} + \frac{40^2}{4} \right) - 3(13) \cong 5.6564.$$

La région critique du test est $\{H \geq h_\alpha\}$ telle que $P_{H_0}(H \geq h_\alpha) \leq \alpha$.

De la table du test de Kruskal-Wallis, on lit, pour $\alpha = 5\%, K = 3$ et pour une combinaison en nombre d'observations par échantillon (5,3,4), $h_{0.05} = 5.656$.

Comme $H_{obs} > h_{0.05}$, alors on rejette (de très peu) l'hypothèse nulle d'égalité des performances.

<u>Distribution asymptotique de *H* :</u>

Lorsque les effectifs sont assez élevés, en pratique $n_i \geq 6 \; \forall i = 1, \dots, K$, la distribution de H peut être approximée par une loi du Khi-deux à $(K-1)$ degrés de liberté. La région critique du test au seuil α s'écrit donc :

$$H \geq \chi^2_{1-\alpha}(K-1)$$

telle que $P_{H_0}(H \geq \chi^2_{1-\alpha}(K-1)) = \alpha$ ($\chi^2_{1-\alpha}(K-1)$ quantile d'ordre $(1-\alpha)$ d'une χ^2 à $(K-1)$ degrés de liberté).

<u>Exemple :</u> Test de Kruskal-Wallis pour des effectifs assez élevés

On répartit une certaine population en $K = 4$ catégories d'individus, sur lesquels on définit une variable continue X. On prélève un échantillon de taille $n_i = 6$ individus dans chaque catégorie i et l'on obtient les observations suivantes :

Catégorie1	Catégorie2	Catégorie3	Catégorie4
85	95	67	90
73	54	74	68
96	72	84	92
91	81	68	94
88	69	87	66
86	80	77	70

On désire tester au risque $\alpha = 5\%$ H_0 contre H_1

H_0: les 4 catégories sont homogènes via à vis de X,

H_1: l'une au moins des catégories diffère des autres.

Le calcul des rangs donne

	Catégorie1	Catégorie2	Catégorie2	Catégorie2	
	15	**23**	**4**	**19**	
	9	**1**	**10**	**2**	
	24	**8**	**14**	**21**	
	20	**13**	**5**	**22**	
	18	**6**	**17**	**3**	
	16	**12**	**11**	**7**	
r_i	**102**	**63**	**61**	**74**	
$\frac{r_i^2}{n_i}$	**1734**	**661.5**	**620.2**	**912.7**	$\sum_{i=1}^{4} \frac{r_i^2}{n_i} = 3928.4$

La statistique H de Kruskal-Wallis a pour valeur (avec $N = 24$)

$$H_{obs} = \frac{12}{24(25)} \mathbf{3928.4 - 3(25) = 3.568}.$$

Sous H_0, la statistique du test suit une χ^2 à $K - 1 = 3$ degrés de liberté ($\boldsymbol{n_i} > 5, i = \overline{\mathbf{1}, \dots, \mathbf{4}}$). La région critique est de la forme

$$\boldsymbol{H \geq \chi^2_{0.95}(3)}.$$

De la table du χ^2, on lit $\boldsymbol{\chi^2_{0.95}(3) = 7.82}$. Comme $H_{obs} < 7.82$ alors on accepte $H_0, c.\text{à}.d.$ il n'y a pas d'influence de la catégorie sur X.

<u>Cas d'ex-æquo :</u>

Lorsque les données comportent des ex-æquo, on utilise généralement la méthode des rangs moyens et la statistique du test devra être corrigée. La statistique ajustée $\boldsymbol{\tilde{H}}$ s'écrit

$$\boldsymbol{\tilde{H} = \frac{H}{1 - \frac{\sum t_l (t_l^2 - 1)}{N(N^2 - 1)}}}$$

où H est calculée avec les rangs moyens lorsqu'il ya des ex-æquo et $C = \mathbf{1} - \frac{\sum t_l (t_l^2 - 1)}{N(N^2 - 1)}$ est le facteur de correction, t_l étant le nombre d'ex-æquo ayant le l −ème rang.

Dans le cas d'approximation, $\tilde{H}$ suit asymptotiquement une χ^2 à $(K - 1)$ degrés de liberté.

<u>Exemple :</u> **Test de Kruskal-Wallis dans le cas d'ex-aequo**

Dans cet exemple la variable d'intérêt est le poids à la naissance d'un animal particulier. On dispose de 56 observations réparties en 8 catégories c1, c2,…,c8 (groupes) :

Obs	Catégorie
1.1	c1
1.2	c7
1.2	c7
1.4	c8
1.6	c2
1.9	c1
2.0	c1
2.0	c2
2.0	c5
2.0	c5
2.1	c5
2.2	c7
2.2	c7
2.3	c2
2.4	c2
2.4	c8
2.5	c4
2.5	c6
2.5	c7
2.5	c8

Obs	Catégorie
2.6	c3
2.6	c4
2.6	c5
2.6	c5
2.6	c7
2.8	c1
2.8	c1
2.8	c2
2.8	c4
2.9	c3
2.9	c4
2.9	c5
2.9	c6
3.0	c8
3.1	c3
3.1	c6
3.1	c6
3.2	c1
3.2	c2
3.2	c3

Obs	Catégorie
3.2	c3
3.2	c3
3.2	c4
3.2	c4
3.3	c1
3.3	c1
3.3	c3
3.3	c3
3.3	c4
3.3	c4
3.4	c3
3.5	c2
3.5	c2
3.6	c1
3.6	c3
4.4	c1

Les rangs des 56 observations sont

1.1	c1	1
1.2	c7	2.5
1.2	c7	2.5
1.4	c8	4
1.6	c2	5
1.9	c1	6
2.0	c1	8.5
2.0	c2	8.5
2.0	c5	8.5
2.0	c5	8.5
2.1	c5	11
2.2	c7	12.5
2.2	c7	12.5
2.3	c2	14
2.4	c2	15.5
2.4	c8	15.5
2.5	c4	18.5
2.5	c6	18.5
2.5	c7	18.5
2.5	c8	18.5

2.6	c3	23
2.6	c4	23
2.6	c5	23
2.6	c5	23
2.6	c7	23
2.8	c1	27.5
2.8	c1	27.5
2.8	c2	27.5
2.8	c4	27.5
2.9	c3	31.5
2.9	c4	31.5
2.9	c5	31.5
2.9	c6	31.5
3.0	c8	34
3.1	c3	36
3.1	c6	36
3.1	c6	36
3.2	c1	41
3.2	c2	41
3.2	c3	41

3.2	c3	41
3.2	c3	41
3.2	c4	41
3.2	c4	41
3.3	c1	47.5
3.3	c1	47.5
3.3	c3	47.5
3.3	c3	47.5
3.3	c4	47.5
3.3	c4	47.5
3.4	c3	51
3.5	c2	52.5
3.5	c2	52.5
3.6	c1	54.5
3.6	c3	54.5
4.4	c1	56

La réorganisation des rangs selon les catégories donne

	C1	C2	C3	C4	C5	C6	C7	C8	
	1	5	23	18.5	8.5	18.5	2.5	4	
	6	8.5	31.5	23	8.5	31.5	2.5	15.5	
	8.5	14	36	27.5	11	36	12.5	18.5	
	27.5	15.5	41	31.5	23	36	12.5	34	
	27.5	27.5	41	41	23		18.5		
	41	41	41	41	31.5		23		
	47.5	52.5	47.5	47.5					
	47.5	52.5	47.5	47.5					
	54.5		51						
	56		54.5						
r_i	317	216.5	414	277.5	105.5	122	71.5	72	
$\frac{r_i^2}{n_i}$	10048.9	5859.03125	17139.6	9625.78125	1855.04166	3721	852.04166	1296	$\sum_{i=1}^{4} \frac{r_i^2}{n_i} = 50397.39577$

La valeur de la statistique de Kruskal-Wallis non corrigée est

$$H_{obs} = \frac{12}{56(57)} 50397.39577 - 3(57) = 18.4638.$$

Le facteur de correction (facteur d'ajustement):

$$c = 1 - \frac{\sum_{l=1}^{22} t_l ({t_l}^2 - 1)}{N(N^2 - 1)}$$

où t_l est le nombre d'observations ayant le $l^{ème}$ rang.

Les observations sont réparties en 22 ensembles d'ex-æquo (22 valeurs distinctes)

Rang l	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
t_l	1	2	1	1	1	4	1	2	1	2	4	5	4	4	1	3	7	6	1	2	2	1

$$\sum_{l=1}^{22} t_l ({t_l}^2 - 1) = 960$$

$$c = 1 - \frac{960}{56(56^2 - 1)} = 0.994$$

La statistique du test ajustée

$$\widetilde{H_{obs}} = \frac{H_{obs}}{c} = \frac{18.4638}{0.994} = 18.575.$$

Sous H_0, la statistique ajustée (corrigée) $\widetilde{H}$ suit une loi du χ^2 à 7 degrés de liberté. La région critique est

$$\widetilde{H} \geq \chi^2_{1-\alpha}(7).$$

De la table du χ^2, on lit $\chi^2_{0.95}(7) = 14.067$ ($\cong 14.07$).

Comme $\mathbf{18.575} >$ 14.07 alors on rejette l'hypothèse nulle.

<u>Détermination de la source des écarts :</u>

Lorsque le test de Kruskal-Wallis aboutit au rejet de l'hypothèse nulle, une procédure de comparaison multiple est menée pour rechercher les couples de groupes différents. Cette comparaison multiple consiste à comparer les paramètres de localisation θ_i de tous les groupes.

Le test s'écrit : $H_0: \theta_i = \theta_j$ v.s . $H_1: \theta_i \neq \theta_j$

(ou de manière équivalente

$H_0: F_i = F_j$ v.s . $H_1: F_i \neq F_j$

$H_0: M_i = M_j$ v.s. $H_1: M_i \neq M_j$.)

Il y a $C_K^2 = \frac{K(K-1)}{2}$ comparaisons à réaliser. L'idée importante est de conserver un niveau de test global compatible avec celui qui a été utilisé pour le test de Kruskal-Wallis . Si la région critique du test de Kruskal-Wallis a été définie au risque α, alors on peut décider que les paramètres de localisation de 2 groupes (par exemple i, j) sont différents si

$$Z_{ij} = \frac{\left|\bar{R}_i - \bar{R}_j\right|}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \geq z_{1-a} \Leftrightarrow \left(\left|\bar{R}_i - \bar{R}_j\right| \geq z_{1-a}\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}\right)$$

où z_{1-a}est le quantile d'ordre $(1-a)$de la loi $N(0,1)$ et

$$\boldsymbol{a = \frac{\alpha}{K(K-1)}}.$$

Cette formulation des comparaisons est valable uniquement si les tailles des échantillons sont assez grandes.

<u>Exemple :</u> Test de Kruskal-Wallis pour la détermination de la source des écarts

On dispose de $N = 54$ observations représentant la teneur en sel (sodium) des sandwiches. Il ya 3 catégories selon la viande qu'ils contiennent : Bœuf, volaille, viande mélange des deux. Les données sont les suivantes :

Bœuf	Volaille	Mélange
253	357	144
298	358	339
300	359	360
317	376	372
319	383	386
322	388	387
324	396	393
330	426	405
370	430	406
375	513	428
385	515	458
401	522	473
425	528	486
440	542	506
477	546	507
479	581	511
482	588	545
495		
587		
645		

La transformation des observations en rangs donne

	Boeuf	Volaille	Mélange	
	2	11	1	
	3	12	10	
	4	13	14	
	5	18	16	
	6	19	21	
	7	23	22	
	8	25	24	
	9	30	27	
	15	32	28	
	17	44	31	
	20	45	34	
	26	46	35	
	29	47	40	
	33	48	41	
	36	50	42	
	37	51	43	
	38	588	49	
	39			
	52			
	54			
$\boldsymbol{r_i}$	**440**	**567**	**478**	
$\boldsymbol{\frac{{r_i}^2}{n_i}}$	**9680**	**18911.117**	**13440.235**	$\sum_{i=1}^{3} \frac{{r_i}^2}{n_i} = \mathbf{42031.3529}$

- $H_{obs} = \frac{12}{54(55)} \mathbf{42031.3529 - 3(55) = 4.8236.}$
- $Sous\ H_0$, la statistique de Kruskal-Wallis suit une χ^2 à 2 degrés de liberté ($n_1 = 20 > 5, n_2 = 17 > 5, n_3 = 17 > 5$).

- La région critique du test est : $H \geq \chi^2_{1-\alpha}(2)$.
- Pour $\alpha = 10\%$, $\chi^2_{0.90}(2) = 4.61$.
- Au seuil $\alpha = 10\%$, $H_{obs} > \chi^2_{0.90}(2)$ donc on rejette H_0. Ce résultat est confirmé par la p-valeur $p - valeur = P(H_0 \geq \mathbf{4.8236}) = 0.0897 < 0.1$.
- Le test de Kruskal-Wallis est significatif au seuil $\alpha = 0.1$. Pour les comparaisons deux à deux (qui sont au nombre de 3), nous calculons le risque

$$a = \frac{\alpha}{K(K-1)} = \frac{0.1}{54(53)} = 0.01666 \cong 0.0167.$$

- Le quantile d'ordre $1 - a = 0.9833$ de la loi $N(0,1)$ est $z_{0.9833} = 2.128$.
- Les moyennes conditionnelles des rangs sont données dans le tableau suivant :

$\boldsymbol{r_i}$	**440**	**567**	**478**
$\boldsymbol{\bar{r}_i}$	**22**	**33.3529**	**28.117**

- Les calculs requis pour les comparaisons sont résumés dans le tableau suivant

$\boldsymbol{\bar{r}_i}$	$\boldsymbol{\bar{r}_j}$	$\lvert\boldsymbol{\bar{r}_i - \bar{r}_j}\rvert$	$z_{0.9833}\sqrt{\frac{54(55)}{12}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$	Significatif
22 (bœuf)	33.3529 (volaille)	**11.3529**	**11.043**	oui
22	28.117	**6.117**	**11.043**	non
33.3529	28.117	**5.2359**	**11.482**	non

Conclusion : Le seul écart significatif survient lors de l'opposition entre « bœuf » et « volaille ».