Deuxième chapitre: Les tests non paramétriques :

Homogénéité entre échantillons indépendants

Il existe des tests moins "exigeants" en conditions d'applications, notamment en ce qui concerne la taille de l'échantillon, la normalité de la distribution et l'égalité des variances**,** ces tests sont dits non paramétriques. Le principe de base de ces tests est de transformer les données en rangs et à mesurer l'accord entre les rangs observés et ce que devrait être ces rangs sous une hypothèse nulle.

Lors de la comparaison de K = 2 échantillons, nous cherchons à savoir si les observations proviennent de la même population au regard de la variable d'intérêt ; ou, de manière équivalente, la distribution de la variable d'intérêt est la même dans les 2 sous échantillons.

Les tests génériques d'homogénéité visent à détecter toute forme de différenciation entre les 2 distributions empiriques. Cela peut être un décalage entre les distributions, une différence de tendance centrale (paramètre de localisation), une dispersion différente (paramètre d'échelle), des queues de distribution plus ou moins lourdes, une asymétrie ou un aplatissement différent, etc.

Le test d'hypothèse s'écrit :

$$H_0 : F_1(x) = F_2(x)$$

$$H_1 : F_1(x) \neq F_2(x)$$

où $F_k(x)$ est la fonction de répartition de la variable d'intérêt X dans la sous population associée à l'échantillon k.

Remarque 1 (Test unilatéral à gauche ou à droite). Il est possible de définir un test unilatéral. Pour traduire l'idée "les valeurs prises par X ont elles tendance à être plus petites dans la première sous population ?", nous définirons le test de la manière suivante :

$$H_0 : F_1(x) = F_2(x)$$

$$H_1 : F_1(x) > F_2(x)$$

Le sens de l'inégalité dans l'hypothèse alternative peut paraître étrange si l'on se réfère au test que l'on veut mettre en place. Nous préciserons cela sur un exemple concret dans ce qui suit.

Exemple : Écart selon le paramètre de localisation.

Dans la figure 1.1, nous avons généré 2 échantillons de taille n_1 = n_2 = 100. Nous avons tracé les fonctions de densité empiriques respectives. Le premier échantillon -1 (trait en pointillé) correspond à une loi normale $N(0; 1)$, le second -2 (trait continu) à une $N(1; 1)$. Bien évidemment, les valeurs dans le second échantillon sont stochastiquement plus élevées que celles du premier.

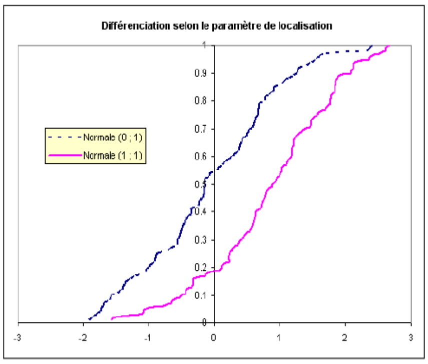


Fig. 1.2. Fonctions de répartition conditionnelles - Différence de paramètre de localisati

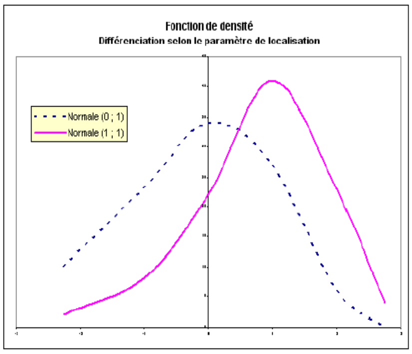


Fig. 1.1. Fonctions de densité conditionnelles - Différence de paramètre de localisation

Concernant les fonctions de répartition, on constate que $F_1(x)$ est systématiquement au dessus de $F_2(x)$. Pour bien comprendre ce mécanisme, prenons comme référence les quantiles de la variable d'intérêt : ils sont systématiquement plus petits dans le premier échantillon. Les 2 fonctions de distributions sont écalées au regard du paramètre de localisation ou de caractéristique de tendance centrale : X a tendance à prendre des valeurs plus élevées dans le deuxième sous échantillon, on a bien $F_1(x) > F_2(x)$.

Parmi ces tests nous allons voir:

- Le test de Kolmogorov et Smirnov, adéquation à une distribution continue (comme introduction), et test de comparaison;

- Le test de Mann-Withney, l'alternative non paramétrique de t de Student pour deux échantillons indépendants;

- Le test de Kruskal-Wallis, l'alternative non paramétrique de l'analyse de variance.

1 – Tests de Kolmogorov et Smirnov

La situation de base est la comparaison de deux échantillons indépendants, notés $Y_1; ... ; Y_{n_1}$ et $Z_1; ... ; Z_{n_2}$. Les Y_i sont supposées indépendantes et de même loi, de fonction de répartition F inconnue, et les Z_j sont supposées indépendantes et de même loi, de fonction de répartition G inconnue. Tester l'hypothèse que les deux échantillons sont issus de la même loi de probabilité, c'est tester :

$$H_0 : F = G \; contre \; H_1 : F \neq G".$$

Le test de comparaison d'échantillons de **Kolmogorov et Smirnov** est une généralisation du test d'adéquation de Kolmogorov.

1.1- Problèmes lié à un seul échantillon: Test d'adéquation de Kolmogorov

On veut tester si les échantillons sont tirés selon une loi de fonction de répartition continue donnée F(x).

Soit X une v.a. de loi de probabilité IP inconnue de fonction de répartition F <u>supposée absolument continue.</u>

Soit $X_1, \dots, X_n$ un n-échantillon de même loi que X. On désire ajuster (test d'adéquation) la loi IP inconnue à une loi donnée IP_0 de fonction de répartition absolument continue F_0 au vue de la fonction de répartition empirique $IF_n(x)$ où $IF_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1I_{]-\infty,x]}(x_i)$

On sait que $IF_n(x) \xrightarrow{p.s} F(x), \forall x$. et que $IF_n(x)$ est un estimateur sans biais de F(x) (chapitre 1).

Théorème de Kolmogorov : Si F est absolument continue alors :

$$\lim_{n\to\infty} IP[\sqrt{n} sup_x |IF_n(x) - F(x)| < z] = K(z) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2z^2}, \qquad z > 0$$

Problème de test : $H_0 : F = F_0$

On définit les statistiques sous H_0 qui sont :

$$D_n^+ = sup_x\left(IF_n(x) - F_0(x)\right)$$

$$D_n^- = sup_x\left(F_0(x) - IF_n(x)\right)$$

$$D_n = \max(D_n^+, D_n^-) = sup_x|F_0(x) - IF_n(x)|$$

Algorithme de résolution

Soit $X_1, \dots, X_n$ l'échantillon des observations :

- On les range dans l'ordre croissant : $X_{(1)}, \dots, X_{(n)}$
- On obtient la fonction de répartition empirique $IF_n(x)$
- La fonction de répartition $F_0(x)$ est croissante
- On sait que la fonction $IF_n(x)$ est constante entre deux valeurs consécutives $[x_{(i)}, x_{(i+1)}[$, on calcule :
  $D_n = sup_{1\le i\le n}\left(\left|\frac{i-1}{n} - F_0(x_{(i)})\right| ; \left|\frac{i}{n} - F_0(x_{(i)})\right|\right)$
  Car $IF_n\left(x_{(i)} - 0\right) = \frac{i-1}{n}$ et $IF_n\left(x_{(i)} + 0\right) = \frac{i}{n}$
- On détermine $d_{n,\alpha}$, telle que $K(d_\alpha) = IP\left(\sqrt{n}D_n \le d_{n,\alpha}\right) = 1 - \alpha$ sous H_0.
- Règle de décision :
  On accepte H_0 si $D_n \le d_{n,\alpha}$
  On rejette H_0 si $D_n > d_{n,\alpha}$

Où $d_{n,\alpha}$ est lue dans la table des valeurs critiques de Kolmogorov (Table 1)

Exemple Test de caractère exponentielle $\mathcal{E}(\frac{1}{\theta})$ d'une loi de fiabilité

On dispose d'un échantillon de n matériels identiques et on note les durées de vie en heures pour un échantillon de taille 5.

$$x_1 = 133;\ x_2 = 169;\ x_3 = 8;\ x_4 = 122;\ x_5 = 58$$

Problème : On test H_0 : $F(x) = F_0(x)$ où $F_0(x) = 1 - e^{\frac{-x}{\theta}}, x > 0$

On estime θ par $\hat{\theta} = \bar{x}$ ES B efficace de θ (EMV par exemple).

Une estimation de $\hat{\theta}$ est $\bar{x}$ = 98 heures.

1- On range les observations dans l'ordre croissant

$x_{(1)} = 8;\ x_{(2)} = 58;\ x_{(3)} = 122;\ x_{(4)} = 133;\ x_{(5)} = 169.$

2- On définit $IF_n(\mathrm{x})$ comme suit :

$$IF_n(\mathrm{x}) = \begin{cases} 0 & \text{si} \quad x < 8 \\ 1/5 & \text{si} \quad 8 \le x < 58 \\ 2/5 & \text{si} \quad 58 \le x < 122 \\ 3/5 & \text{si} \quad 122 \le x < 133 \\ 4/5 & \text{si} \quad 133 \le x < 169 \\ 1 & \text{si} \quad x \ge 169 \end{cases}$$

$x_{(i)}$	8	58	122	133	169
$F_0(x_{(i)})$	0.079	0.447	0.711	0.743	0.821
$IF_n(x_{(i)})$	0 1/5	2/5	3/5	4/5	1

3- La statistique de Kolmogorov –Smirnov :

$$D_n = sup_{1\le i\le n}\left(\left|\frac{i-1}{n} - F_0(x_{(i)})\right|;\ \left|\frac{i}{n} - F_0(x_{(i)})\right|\right)$$

$$= \sup(0.079;\ \left|0.079 - \frac{1}{5}\right|;\left|0.447 - \frac{1}{5}\right|;\ \left|0.447 - \frac{2}{5}\right|;\ \left|0.711 - \frac{2}{5}\right|;$$

$$\left|0.711 - \frac{3}{5}\right|;\ \left|0.743 - \frac{3}{5}\right|;\ \left|0.743 - \frac{4}{5}\right|;\ \left|0.821 - \frac{4}{5}\right|\ |0.821 - 1|;$$

=>D_5=0.311

4- Au niveau $\alpha = 0.05$, $IP(\sqrt{n}D_n \le d_{n,\alpha}) = 1 - \alpha$

Pour un test bilatéral, on trouve $d_{5,0.05}$ =0.563 (La table 1)

5- Conclusion : On a $D_5 = 0.311 < d_{5,0.05}$ =0.563 ===➔ On accepte H_0 au niveau de signification 5%

Exercice 1 :

Un calculateur a simulé un échantillon de n= 10 valeurs distribuées selon une loi normale. Les valeurs Xi produites sont rangées par ordre croissant :

$x_{(i)}$	10.8	10.9	11.9	13.5	15.9	16.6	17.4	17.9	18.7	23.0

On va chercher à vérifier si cet échantillon est correct.

1) Donner une estimation de la moyenne et l'écart-type de l'échantillon.

2) Calculer, au moyen d'une table de la loi normale, les valeurs de la fonction de répartition F pour l'échantillon.

3) Exécuter un test de Kolmogorov-Smirnov au seuil de 5% pour décider si la distribution de l'échantillon est en adéquation avec la loi normale.

Exercice 1 :

Pour étudier la solidité d'un type de béton armé, on a pris un n-échantillon d'un certain modèle et on a trouvé les mesures de solidité (en méga pascal) présentées dans le tableau suivant :

Classe de solidité	Fréquence absolue n_i
19- 20	10
20-21	26
21- 22	56
22- 23	64
23- 24	30
24- 25	14

1- Donner une représentation graphique appropriée.
2- Que remarquez-vous ?
3- Tester si cet échantillon provient d'une population normale à un niveau de signification 10%.

1.2- Comparaison de deux échantillons

Le test de Kolmogorov Smirnov vise à détecter toute forme de différenciation entre les distributions. Il repose sur l'écart maximum entre les fonctions de répartition empiriques.

On considère un couple de variables aléatoires (Y;Z), tel que Y et Z sont supposées indépendantes, de lois respectives P_Y et P_Z. On considère un n_1-échantillon $(Y_1,\dots,Y_{n1})$ de la loi de Y et un n_2-échantillon $(Z_1,\dots,Z_{n2})$ de la loi de Z, supposés indépendants, et on dispose de l'observation $x=(y_1,\dots,y_{n1};z_1,\dots,z_{n2})$ de $X=(Y_1,\dots,Y_{n1}; Z_1,\dots,Z_{n2})$.
On note $\mathbb{F}_{n1}$ la fonction de répartition empirique associée à $(Y_1,\dots,Y_{n1})$ et $\mathbb{G}_{n2}$ celle associée à $(Z_1,\dots,Z_{n2})$ et on introduit la statistique du test de Kolmogorov-Smirnov:

$$D_{n_1,n_2} = sup_x\left|\mathbb{F}_{n_1}(x) - \mathbb{G}_{n_2}(x)\right|$$

Proposition 1: Sous H_0 : « F = G », la statistique D_{n_1,n_2} est libre (c-à-d, sa loi ne dépend que de la taille de l'échantillon et non de la loi commune des observations composant cet échantillon). On notera sa loi $\mathcal{D}(n_1,n_2)$.

Remarque : La loi de Kolmogorov-Smirnov $\mathcal{D}(n_1,n_2)$. est tabulée pour des petites valeurs de n_1 et n_2, elle est aussi accessible facilement par simulation. Pour le cas $n_1 = n_2$, et si $D_{n,n}$ désigne une v.a. de loi $\mathcal{D}(n,n)$, on dispose aussi de la formule (pour tout entier strictement positif z) :

$$\lim_{n\to\infty} IP[\sqrt{n}\, D_{n,n} < z] = K(z) = 2\sum_{j=1}^{[n/z]}(-1)^{j+1}\frac{(n!)^2}{(n-jk)!\,(n+jk)!}$$

Pour les échantillons de grande taille, lorsque n_1 et n_2 tendent vers l'infini, on a le comportement asymptotique suivant :

Proposition 2: Si on pose $\zeta_{n_1,n_2} = \sqrt{\frac{n_1\times n_2}{n_1+n_2}}D_{n_1,n_2}$, lorsque $min(n_1,n_2)\to+\infty$, on a

- Sous H_0, $\forall y > 0, P(\zeta_{n_1,n_2} \le y) \to \sum_{k=-\infty}^{+\infty}(-1)^k e^{-2k^2y^2}$.

- Sous H_1, p.s. ζ_{n_1,n_2} tend vers $+\infty$.

Le test de Kolmogorov-Smirnov consiste à rejeter H_0 : « F = G »si et seulement

$$D_{n_1,n_2} > d_\alpha(n_1, n_2)$$

Où $d_\alpha(n_1, n_2)$ est lue dans la table des valeurs critiques de Kolmogorov-Smirnov de comparaison (Table 2).

Exemple :

Est-ce que la capacité à maintenir son équilibre lorsque l'on est concentré est différente selon l'âge ? Pour répondre à cette question, n= 17 observations ont été recueillies. Des personnes ont été placées sur un plateau mouvant. Elles devaient réagir en appuyant sur un bouton lorsque des signaux arrivaient à intervalles irréguliers. Dans le même temps, elles devaient se maintenir sur le plateau. On a mesuré alors l'amplitude des corrections, d'avant en arrière, effectuées pour rester debout. Les personnes sont subdivisées en 2 groupes: les vieux (n_1=9) et les jeunes (n_2= 8).

numéro	X	groupe	numéro	X	groupe
1	19	vieux	10	25	jeune
2	30	vieux	11	21	jeune
3	20	vieux	12	17	jeune
4	19	vieux	13	16	jeune
5	29	vieux	14	14	jeune
6	25	vieux	15	14	jeune
7	21	vieux	16	22	jeune
8	24	vieux	17	17	jeune
9	50	vieux			

2- Nous déduisons alors la statistique du test en récupérant le maximum de la dernière colonne, soit :
$D_{9,8}$ = max(0.250,0. 375, 0. 625, …,0.000)=0. 625.

3- La valeur critique du test à 5% est lue dans la table, $d_{5\%}(9,8) = 0.639$.

4- Conclusion : puisque : $D_{9,8} = 0.625 < d_{5\%}(9,8) = 0.639$, au risque 5%, nos données ne contredisent pas l'hypothèse nulle d'égalité des fonctions de répartition.

Nous remarquerons néanmoins que nous sommes aux portes de la région critique. Pour un test à 10%, nous rejetterions l'hypothèse nulle.

Exercice 3 :

On souhaite comparer deux médicaments sensés soulager la douleur. On a observé sur 16 patients dont 8 ont pris le médicament A habituel et les 8 autres un médicament B expérimental à base de cannabis, les durées de soulagement suivantes (en heures) :

A	6.8	3.1	5.8	4.5	3.3	4.7	4.2	4.9
B	4.4	2.5	2.8	2.1	6.6	1.1	4.8	2.3

1. Proposer un test paramétrique de comparaison.
2. On veut maintenant effectuer un test non paramétrique de Kolmogorov-Smirnov. Expliquer pourquoi la statistique de test sous l'hypothèse nulle peut être tabulée. Conclure pour un niveau asymptotique 5%.

2. Test de Wilcoxon- Mann-Whitney :

Ce test est destiné à étudier si une variable indépendante nominale **dichotomique** influence une variable dépendante ordinale de scores rangés ou d'intervalle.
Ce test doit être préféré au test t de student lorsque la distribution n'obéit pas à la loi normale (donc remarquablement dissymétrique).

Le test de Mann - Whitney est certainement le plus populaire des tests non paramétriques. Il recouvre en réalité 2 formulations, totalement équivalentes (ils peuvent se déduire l'un de l'autre), le test de Wilcoxon d'une part, le test de Mann-Whitney d'autre part. Nous nous concentrerons sur ce dernier dans un premier temps.

2.1. Principes d'application

Si 2 populations sont très différentes, le cas extrême se produit quand les valeurs de l'une sont inférieures aux valeurs de l'autre. Par contre, si elles sont confondues, il doit y avoir intrication des valeurs de ces 2 populations. Pour comparer deux moyennes, il faut habituellement employer le test t, qui suppose la normalité des distributions et l'égalité des variances, hypothèses invérifiables avec des effectifs faibles.

Le principe de ce test est: Si les Y_j sont dans l'ensemble supérieurs aux X_i, alors les rangs R_j des Y_j seront dans l'ensemble supérieurs aux rangs R_i des X_i dans l'échantillon complet.

Test bilatéral

La statistique de Mann et Whitney utilise la somme des rangs. Nous retrouvons bien l'idée de décalage entre les distributions basé sur leurs localisations respectives. Pour le test bilatéral

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Définition 1 : La statistique de Wilcoxon $T_{n_1;n_2}$ est la somme des rangs des observations du deuxième échantillon dans l'échantillon complet :

$$T_{n_1;n_2} = \sum_{j=1}^{n_2} R_j$$

Définition 2 . Le test de Wilcoxon est le test de comparaison de deux échantillons basé sur la statistique de Wilcoxon.

Remarques:

- Dans le cas extrême où les Y_j sont tous inférieurs aux X_i, $T_{n_1;n_2} = \sum_{j=1}^{n_2} j = \frac{n_2(n_2+1)}{2}$.
- Inversement, si les Y_j sont tous supérieurs aux X_i, $T_{n1;n2} = \sum_{j=1}^{n_2+n_1} j = \frac{n_2(n_2+1)}{2} + n_1 n_2$.
- Sous H_0, le mélange des deux échantillons est homogène, donc $T_{n1;n2}$ devrait être de l'ordre de

$\frac{n_2(n_2+1)}{2} + \frac{n_1 n_2}{2} = \frac{n_2(n_1+n_2+1)}{2}$.

La statistique de Mann et Whitney utilise la somme des rangs. Nous retrouvons bien l'idée de décalage entre les distributions basé sur leurs localisations respectives.

On peut aborder le problème différemment, en remarquant que, sous H_0, comme les X_i et les Y_j sont indépendantes et de même loi, on a $\forall(i,j);\ P\left(X_i \leq Y_j\right) = \frac{1}{2}$

Définition 3 : La statistique de Mann-Whitney est le nombre de couples $(i;\ j)$ tels que $X_i \leq Y_j$:

$$U_{n_1,n_2} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} 1_{\{X_i \leq Y_j\}}$$

Sous H_0, U_{n_1,n_2} doit être de l'ordre de la moitié des couples $(X_i; Y_j)$ possibles, à savoir $\frac{n_1 n_2}{2}$.

Définition 4 : Le test de Mann-Whitney est le test de comparaison de deux échantillons basé sur la statistique de Mann-Whitney.

Propriété 1 : Sous H_0, $\frac{2U_{n_1,n_2} - n_1 n_2}{\sqrt{(n+1)n_1 n_2}}\sqrt{3} \to \mathcal{N}(0,1)$.

La condition de validité de l'approximation normale est que : $n_1\ \geq 8\ et\ n_2\ \geq 8$.

Propriété 2 . $U_{n_1,n_2}\ =\ T_{n_1;n_2} - \frac{n_2(n_2+1)}{2}$.

2.2. Algorithme de résolution

1er cas: n_A et n_B sont supérieurs à 8

Supposant les donnés suivantes:

A(n_A=12)	11	9	7	12	12	40	5	4	15	10	10	14
B(n_B=11)	13	15	15	14	35	18	13	25	20	6	5	

- Transformer les scores en rangs
- Mélanger les donnés de deux groupes
- Ordonner la série obtenue en ordre croissant
- Accorder des rangs; pour les ex-æquo attribuer à chacun le rang moyen
- Reconstruire les deux groupes avec données et les rangs correspondants

Ce qui donnera la répartition suivante:

A scores	11	9	7	12	12	40	5	4	15	10	10	14
Rangs(R_i)	9	6	5	10.5	10.5	23	2.5	1	17	7.5	7.5	14.5
B scores	13	15	15	14	35	18	13	25	20	6	5	
rangs(R_j)	12.5	17	17	14.5	22	19	12.5	21	20	4	2.5	

2/ calculer la somme des rangs de A et la somme des rangs de B

$T_A = \sum_{i=1}^{n_A} R_i = 114$ et $T_B = \sum_{j=1}^{n_B} R_j = 162$

3/ Calculer $U_1 = n_A n_B + \frac{n_A(n_A+1)}{2} - T_A$ $\quad U_2 = n_A n_B + \frac{n_B(n_B+1)}{2} - T_B$ $\quad U = \min(U_1, U_2)$

4/ Mann et Whitney ont montré que la variable U se distribue selon une loi approximativement normale. Calculer donc: la moyenne de la distribution de U est:

$m_U = \frac{n_A \times n_B}{2}$ et l'écart-type est: $\sigma_U = \sqrt{\frac{(n_A \times n_B)(n_A+n_B+1)}{12}}$

6/ Il suffit de tester l'écart entre U et m_U, soit: $|Z| = \frac{|U-m_U|}{\sigma_U}$

Si nous revenons à notre exemple, nous aurons donc: $U_1 = (12 \times 11) + \frac{12(12+1)}{2} - 114 = 96$ et

$U_2 = (12 \times 11) + \frac{11(11+1)}{2} - 162 = 36$, $U = \min(U_1, U_2) = U_2 = 36$

$$m_U = \frac{12 \times 11}{2} = 66; \qquad \sigma_U = \sqrt{\frac{(12 \times 11)(12+11+1)}{12}} = 16.25$$

On peut vérifier que $m_U = \frac{U_1+U_2}{2} = 66$. Par conséquent $|Z| = \frac{|36-66|}{16.25}$=1.85

7/ Vérifier la signification de la valeur Z :

La région critique du test bilatérale au niveau de signification α est :

$$RC : |Z| \geq u_{(1-\frac{\alpha}{2})}$$

Où $u_{(1-\frac{\alpha}{2})}$ est le quantile d'ordre $(1-\frac{\alpha}{2})$ de la loi normale centrée réduite.

Pour notre exemple :

- Si α=0.05 ; Z =1.85 < 1,96 => on accepte H_0, donc, la différence n'est pas significative au α=0.05
- Si α =0.01 ; Z =1.85 < 2,56 => la différence n'est pas significative au α =0.01.
- Si α=0.10 ; Z =1.85 > 1,64 => on rejette H_0, donc, la différence est significative au α=0.10

Remarque : La correction de continuité

Lorsque les effectifs sont de taille modérée, nous pouvons améliorer l'approximation normale en introduisant la correction de continuité. Notre statistique de test s'écrit pour un test bilatéral :

$$|Z| = \frac{|U - m_U| - 0.5}{\sigma_U}$$

La règle de décision n'est pas modifiée.

Le cas des tests unilatéraux

Les tests unilatéraux ($H_1 : \mu_1 < \mu_2$ ou $H_1 : \mu_1 > \mu_2$) ne posent pas difficultés particulières, il faut simplement introduire correctement la correction de continuité. La statistique générique s'écrit ;

$$Z = \frac{U - m_U \pm 0.5}{\sigma_U}$$

Pour un test unilatéral à gauche, nous rajouterons la valeur 0.5 ; à droite, nous la retrancherons.

La région critique pour un test unilatéral à gauche (resp. à droite) au risque α devient :

RC :Z$\leq u_\alpha$(resp. RC :Z$\geq u_{1-\alpha}$).

2ème cas: n_A et n_B sont inférieurs à 8

Dans ce cas, la distribution n'est pas gaussienne, le modèle précédent ne peut pas être appliqué. Mann et Withney ont construit des tables avec des valeurs critiques qu'il est possible de consulter directement en fonction de:

- de **U**=U_1 si U_1 est inférieur à U_2
- de **U**=U_2 si U_2 est inférieur à U_1

Supposons les mesures de deux groupes et leurs rangs:

A	Scores	6	3	10	5	14	n_1=**5**
	Rangs(R_i)	**7**	**9**	**5**	**8**	**3**	T_A=**32**
B	Scores	12	8	16	18		n_2=**4**
	Rangs(R_j)	**4**	**6**	**2**	**1**		T_B=**13**

$U_1 = 3$ $U_2 = 17$ $U = 3$

La table est consultée en fonction de l'effectif n_2 du plus grand de deux échantillons (ici, n_2 =5).

- Nous devrions comparer cette valeur observée avec les valeurs critiques lues dans la table si nous nous référons au schéma classique. Mais cette dernière est organisée un peu différente dans ce support, elle nous fournit la probabilité $P(MW \leq U)$. Nous pouvons donc obtenir directement la probabilité critique du test en calculant $p = 2\,P(MW \leq U)$. (test bilatéral)
- Dans notre cas, la table nous indique $P(MW \leq 3) = 0.056$, nous pouvons en déduire $p = 2 \times 0.056 = 0.112 > 0.05$. Au seuil 5%,
  Nous ne pouvons pas rejeter l'hypothèse d'égalité des deux moyennes.

Remarque: Pour simplifier les calculs on prendra toujours la somme des rangs dans la situation comportant le moins de sujets. Lorsqu'il y aura le même nombre de sujets dans les deux conditions, il sera possible de prendre l'une ou l'autre des deux conditions pour calculer la somme des rangs.

Exemple

X	Y	R_X	R_Y
3	15	**1**	**12.5**
5	18	**2**	**15**
12	15	**9.5**	**12.5**
10	12	**7**	**9.5**
13	7	**11**	**3**
8	10	**4.5**	**7**
17	10	**14**	**7**
	8		**4.5**
n_1=**7**	n_2=**8**	T_1=**49**	T_2=**71**

$U_1 = 35$ $U_2 = 21$ $U = 21$

Nous avons n_2=**8 ,** n_1=**7,** $U = 21$ nous lisons pour un test bilatéral

$$p = 2\,P(MW \leq 21) = 2(0.232) = 0.464 > 0.05$$

La différence entre les moyennes des rangs est donc non significative au seuil 5%.

Exercice 4

1. Démontrer que $T_A + T_B = (n_A + n_B)(n_A + n_B + 1)/2$.

2. Démontrer que $U_1 + U_2 = n_A n_B$.
3. Démontrer que $T_B + U_2 = n_A n_B + n_B(n_B + 1)/2$.

Exercice 5

On veut comparer les performances de deux groupes d'élèves à des tests d'habilités manuelle. On choisit aléatoirement 8 élèves du 1er groupe et 10 élèves du second groupe. Les performances en min sont

Groupe1	22	31	14	19	24	28	27	28		
Groupe2	25	13	20	11	23	16	21	18	17	26

- Les deux groupes sont-ils de même habilité manuelle ?

Exercice 6

La taille des feuilles de ronces ont été mesurées pour voir s'il y a une différence entre la taille des feuilles qui poussent en plein soleil et celles qui poussent à l'ombre. Les résultats sont les suivants (Largeur des feuilles en cm)

Soleil	6.0	4.8	5.1	5.5	4.1	5.3	4.5	5.1
Ombre	6.5	5.5	6.3	7.2	6.8	5.5	5.9	5.5

- Proposer un test non paramétrique pour donner une réponse à ce problème.

3. Tests de rang dans un modèle de localisation pour $K \geq 2$ populations

Les tests non paramétriques de comparaison de populations peuvent être étendus à K populations ($K \geq 2$). Tout comme le test de Student de comparaison de moyennes peut être généralisé en analyse de variance, une comparaison simultanée de K moyennes, nous devrions pouvoir définir des sortes d'analyse de variance sur les rangs, ou plus précisément sur les scores déduits des rangs.

Dans un premier temps, reconsidérons la formulation du test d'hypothèses : On cherche à savoir si les fonctions de répartitions conditionnelles $F_k(x)$ sont toutes identiques, l'hypothèse nulle s'écrit :

$$H_0 : F_1(x) = \ldots = F_k(x) = \ldots = F_K(x)$$

L'hypothèse alternative est "une des distributions au moins est différente des autres". Si l'on cherche avant tout à s'intéresser au paramètre de localisation θ_k, on peut réécrire l'hypothèse H_0 de la manière suivante :

$$H_0 : \theta_1 = \ldots = \theta_k = \ldots = \theta_K$$

L'hypothèse alternative correspond bien évidemment à "un des θ_k au moins diffère des autres, ou d'un autre". On ne manquera pas de faire le rapprochement avec les hypothèses de l'analyse de variance.

A la différence que θ correspond à un indicateur de tendance centrale, qui n'est pas forcément la moyenne. On pourrait penser à la médiane, mais en réalité nous sommes dans un cadre encore plus générique : les paramètres de localisation θ_k servent avant tout à caractériser le décalage entre les fonctions de distribution.

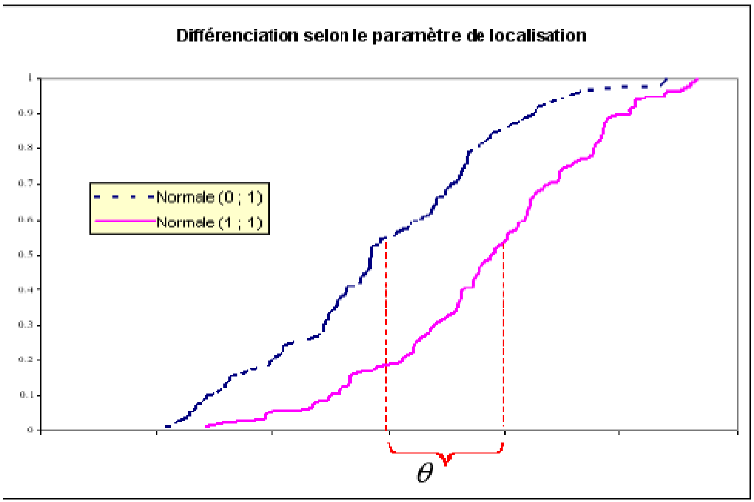


Fig. 1 Paramètre de translation - Décalage entre 2 fonctions de répartition

3.1. Test de Kruskal-Wallis

3.1. 1. Principe, statistique de test et région critique

Le test de Kruskal-Wallis est la généralisation à K populations du test de la somme des rangs de Mann-Whitney bilatéral. On le considère comme l'alternative non paramétrique de l'ANOVA dès que la distribution sous-jacente des données n'est plus gaussienne. Il est extrêmement populaire.

Le rapprochement avec l'analyse de variance est justifié jusque dans la construction de la statistique de test. Soit $\bar{r}$ la moyenne globale des rangs, et $\bar{r_k}$ la moyenne des rangs pour les observations du groupe n^o k (moyenne conditionnelle), la statistique de Kruskal-Wallis est définie de la manière suivante :

$$H = \frac{12}{n(n+1)} \sum_{k=1}^{K} n_k (\bar{r_k} - \bar{r})^2.$$

Remarques : 1/ C'est l'expression d'une variabilité inter-classes, c.-à-d. la dispersion des moyennes conditionnelles autour de la moyenne globale.

2/ $H \geq 0$.

3/ Si l'hypothèse nulle est vérifiée, les moyennes conditionnelles des rangs sont proches de la moyenne globale (à la fluctuation d'échantillonnage près), H prend une valeur proche de 0. La région critique correspond aux grandes valeurs de H. Plus H s'écarte de 0, plus l'hypothèse alternative sera crédible.

Il est possible de simplifier l'expression ci-dessus. On retrouve plus couramment la formule suivante dans la littérature : $H = \frac{12}{n(n+1)} [\sum_{k=1}^{K} \frac{S_k^2}{n_k}] - 3(n+1)$

Où S_k est la somme des rangs des individus appartenant au groupe $n^o k$.

Exemple : (Kruskal-Wallis : Un exemple pour les petits échantillons - La production de capsules).

Une usine veut comparer les performances de 3 réglages différents de machines outils dans la production de capsules : un réglage standard, et 2 autres types de réglages. La variable d'intérêt est le nombre de capsules produites dans une unité de temps. Cet exemple a été utilisé par Kruskal et Wallis pour illustrer leur méthode dans leur article.

Reglage	Production	rang
Standard	340	5
Standard	345	9
Standard	330	1
Standard	342	6
Standard	339	3
Modif_1	339	4
Modif_1	333	2
Modif_1	344	8
Modif_2	347	10
Modif_2	343	7
Modif_2	349	11
Modif_2	349	12

Groupe	Standard	Modif_1	Modif_2
n_k	5	3	4
S_k	24	14	40
S_k^2	576	196	1600

Nous en déduisons la statistique du test avec n=12

$$H = \frac{12}{n(n+1)}[\sum_{k=1}^{K}\frac{S_k^2}{n_k}] - 3(n+1) = \frac{12}{12(12+1)}[\frac{576}{5}+\frac{196}{3}+\frac{1600}{4}] - 3(12+1) = 5.6564$$

Les effectifs étant faibles, nous devons utiliser les tables de Kruskal-Wallis (voir la table de Kruskal-Wallis). Cette dernière est organisée de manière à ce que, pour chaque combinaison de n_k, nous ayons la probabilité critique du test pour une valeur de H donnée. C'est un peu étrange, mais on le comprend dans la mesure où pour les faibles effectifs, le nombre de combinaisons possibles peut être facilement énuméré. Dans notre table, nous avons pour une combinaison de 3 groupes correspondant aux effectifs (5; 4; 3), la probabilité critique associée à la valeur $H = 5.6564$ est $p = 0.049$.

⇨ De fait, pour un test à 5%, nous rejetons (de très peu) l'hypothèse nulle d'égalité des paramètres de localisation.

3.1.2 Distribution asymptotique

Lorsque les effectifs sont assez élevés, en pratique $n_k > 5; \forall k$, la distribution de H peut être approximée par une loi du χ^2 à $(K-1)$ degrés de liberté lorsque H_0 est vrai.

En effet, n'oublions pas que les sommes de rangs S_k sont asymptotiquement normaux (plus généralement la somme des scores). De fait, toute statistique de la forme

$$\sum_{k=1}^{K}\frac{[S_k - E(S_k)]^2}{V(S_k)} \sim {\chi^2}_{(K-1)}$$

Compte tenu du fait que les quantités S_k sont reliées par une relation linéaire. La région critique du test au risque α s'écrit :

$$\text{R.C.} : H \geq {\chi^2}_{(K-1,1-\alpha)}$$

3.1.3 Traitement des ex-aequo

Lorsque les données comportent des ex-æquo, nous utilisons le principe des rangs moyens et la statistique du test devra être corrigée. Soit G le nombre de valeurs distinctes dans le fichier ($G \leq n$). Pour la valeur $n^o g$, nous observons t_g valeurs. La statistique ajustée s'écrit

$$\widetilde{H} = \frac{H}{1-\frac{\sum_{g=1}^{G}(t_g^3 - t_g)}{n^3 - n}}$$

☼ Attention, la statistique H est calculée sur les rangs modifiés (c.-à-d. avec les rangs moyens lorsqu'il y a des ex-æquo).

Remarques:

1- S’il n’y a aucun ex-æquo, alors t=0 et le facteur de correction vaut 1, si bien qu’aucune correction n’est alors nécessaire.

2- Dans la pratique, la correction est le plus souvent négligeable, c.-à-d., qu'elle n'est pas suffisante pour modifier la conclusion de l'analyse.

Exemple : Supposant 4 groupes de sujets reçoivent un enseignement selon quatre méthodes différentes.

On souhaite comparer leurs résultats sur la base des données suivantes:

Groupes	1	2	3	4
Scores	8	15	18	4
	20	14	16	7
	13	7	15	12
	14	9	19	10
	17	12		8
		10		6
				11
Effectifs	n_1=5	n_2=6	n_3=4	n_4=7

1/ On mélange les 4 groupes (K=4) et on ordonne les scores:

4 - 6 - 7 - 7 - 8 - 8 - 9 - 10 - 10 - 11 - 12 - 12 - 13 - 14 - 14 - 15 - 15 - 16 - 17 - 18 - 19 - 20

2/ Calculons les sommes des rangs

Groupe1		Groupe2		Groupe3		Groupe4	
Scores	Rangs	Scores	Rangs	Scores	Rangs	Scores	Rangs
8	5.5	15	16.5	18	20	4	1
20	22	14	14.5	16	18	7	3.5
13	13	7	3.5	15	16.5	12	11.5
14	14.5	9	7	19	21	10	8.5
17	19	12	11.5			8	5.5
		10	8.5			6	2
						11	10
S_1=**74**		S_2=**61.5**		S_3=**75.5**		S_4=**42**	

Notons que $\sum S_k$ (somme des totaux des rangs) $=\frac{n(n+1)}{2}=$ 253

3/ On applique la formule de Kruskal et Wallis:

$$H = \frac{12}{n(n+1)}\left[\sum_{k=1}^{K}\frac{S_k^2}{n_k}\right] - 3(n+1)=11.64$$

⇨ H=11.64

4/ La valeur critique $\chi^2_{(3-1,0.95)} = 7.81$. Alors la R.C. : H ≥ 7.81

5/ Conclusion La valeur calculée est **supérieur** à la valeur théorique, donc, on **rejette** H_0 au niveau de signification 5% . Autrement dit, il **existe** des différences entre les moyennes des 4 groupes.

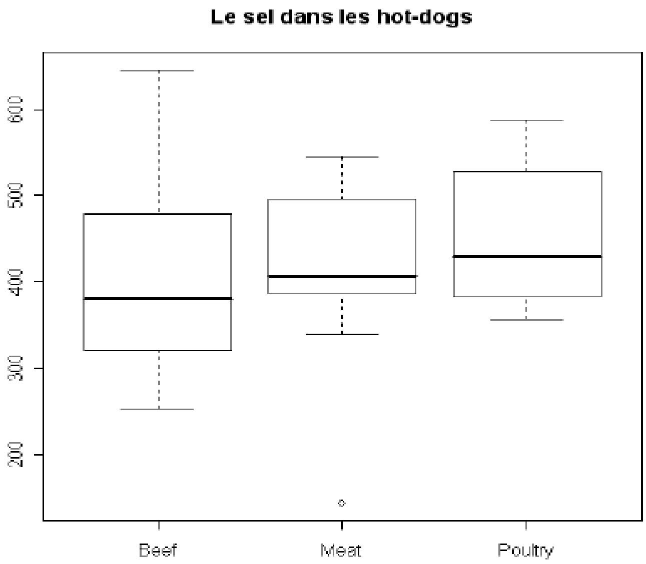
Remarque : Puisqu'il y des ex-æquo, faisons la correction Avec n=22, G=6 et $t_g = 2, \forall g = \overline{1,6}$. Le facteur de correction est :

$$1- \frac{\sum_{g=1}^{G}(t_g^3 - t_g)}{n^3 - n} = 1 - \frac{6(2^3-2)}{22^3-22} = 1 - \frac{36}{10626} = 0.9966 => \widetilde{H} = \frac{11.64}{0.9966} = 11.68$$

On conclue qu'il n'y a pas de grande différence entre H et $\widetilde{H}$ et la conclusion est la même.

Exercice 7: (Kruskal-Wallis : Distribution asymptotique - Les hot-dogs plus ou moins salés).

On s'intéresse à la teneur en sel (variable d'intérêt "Sodium") de "Hot-Dogs". Il y a 3 catégories, selon la viande qu'ils contiennent : bœuf, volaille et un mélange de viande. Visuellement, il semble qu'il y ait *un petit quelque chose* si l'on se réfère aux boîtes à moustaches conditionnelles (Figure).



Il faut confirmer ou infirmer cela avec les calculs statistiques, en proposons un test statistique d'égalité des moyennes à un risque $\alpha = 5\%$, puis à $\alpha = 10\%$,.

Les données ont été regroupées en bloc d'appartenance comme suit :

TYPE	Beef	Meat	Poultry
	235	144	357
	298	339	358
	300	360	359
	317	372	376
	319	386	383
	322	387	388
	324	393	396
	330	405	426
	370	406	430
Scores	375	428	513

	385	458	515
	401	473	522
	425	496	528
	440	506	542
	477	507	546
	479	511	581
	482	545	588
	495		
	587		
	645		

Groupe	Beef	Meat	Poultry
n_k	20	17	17
S_k	440	478	367
S_k^2	193600	228484	321489

$$H = \frac{12}{n(n+1)}\left[\sum_{k=1}^{K}\frac{S_k^2}{n_k}\right] - 3(n+1) = 4.8236$$

La valeur critique $\chi^2{}_{(3-1,0.95)} = 5.9915$. Alors la R.C. : H ≥ 5.9915 .

Les données sont compatibles avec H_0 d'égalité de teneur en sel des « Hot-dogs ».

Probabilités associées aux valeurs aussi extrêmes que les valeurs de U observées du test de Mann-Whitney

n2 = 3				n2 = 4				
n1 / U	1	2	3	n1 / U	1	2	3	4
0	0,250	0,100	0,050	0	0,200	0,067	0,028	0,014
1	0,500	0,200	0,100	1	0,400	0,133	0,057	0,029
2	0,750	0,400	0,200	2	0,600	0,267	0,114	0,057
3		0,600	0,350	3		0,400	0,200	0,100
4			0,500	4		0,600	0,314	0,171
5			0,650	5			0,429	0,243
				6			0,571	0,343
				7				0,443
				8				0,557

n2 = 5							n2 = 6						
n1 / U	1	2	3	4	5		n1 / U	1	2	3	4	5	6
0	0,167	0,047	0,018	0,008	0,004		0	0,143	0,036	0,012	0,005	0,002	0,001
1	0,333	0,095	0,036	0,016	0,008		1	0,286	0,071	0,024	0,010	0,004	0,002
2	0,500	0,190	0,071	0,032	0,016		2	0,428	0,143	0,048	0,019	0,009	0,004
3	0,667	0,286	0,125	0,056	0,028		3	0,571	0,214	0,083	0,033	0,015	0,008
4		0,429	0,196	0,095	0,048		4		0,321	0,131	0,057	0,026	0,013
5		0,571	0,286	0,143	0,075		5		0,429	0,190	0,086	0,041	0,021
6			0,393	0,206	0,111		6		0,571	0,274	0,129	0,063	0,032
7			0,500	0,278	0,155		7			0,357	0,176	0,089	0,047
8			0,607	0,365	0,210		8			0,452	0,238	0,123	0,066
9				0,452	0,274		9			0,548	0,305	0,165	0,090
10				0,548	0,345		10				0,381	0,214	0,120
11					0,421		11				0,457	0,268	0,155
12					0,500		12				0,545	0,331	0,197
13					0,579		13					0,396	0,242
							14					0,465	0,294
							15					0,535	0,350
							16						0,409
							17						0,469
							18						0,531

n2 = 7							
n1 U	1	2	3	4	5	6	7
0	0,125	0,028	0,008	0,003	0,001	0,001	0,000
1	0,250	0,056	0,017	0,006	0,003	0,001	0,001
2	0,375	0,111	0,033	0,012	0,005	0,002	0,001
3	0,500	0,167	0,058	0,021	0,009	0,004	0,002
4	0,625	0,250	0,092	0,036	0,015	0,007	0,003
5		0,333	0,133	0,055	0,024	0,011	0,006
6		0,444	0,192	0,082	0,037	0,017	0,009
7		0,556	0,258	0,115	0,053	0,026	0,013
8			0,333	0,158	0,074	0,037	0,019
9			0,417	0,206	0,101	0,051	0,027
10			0,500	0,264	0,134	0,069	0,036
11			0,583	0,324	0,172	0,090	0,049
12				0,394	0,216	0,117	0,064
13				0,464	0,265	0,147	0,082
14				0,538	0,319	0,183	0,104
15					0,378	0,223	0,130
16					0,438	0,267	0,159
17					0,500	0,314	0,191
18					0,562	0,365	0,228
19						0,418	0,267
20						0,473	0,310
21						0,527	0,355
22							0,402
23							0,451
24							0,500
25							0,549

n2 = 8										
n1 U	1	2	3	4	5	6	7	8	t	Normal
0	0,111	0,022	0,006	0,002	0,001	0,000	0,000	0,000	3,308	0,001
1	0,222	0,044	0,012	0,004	0,002	0,001	0,000	0,000	3,203	0,001
2	0,333	0,089	0,024	0,008	0,003	0,001	0,001	0,000	3,098	0,001
3	0,444	0,133	0,042	0,014	0,005	0,002	0,001	0,001	2,993	0,001
4	0,556	0,200	0,067	0,024	0,009	0,004	0,002	0,001	2,888	0,002
5		0,267	0,097	0,036	0,015	0,006	0,003	0,001	2,783	0,003
6		0,356	0,139	0,055	0,023	0,010	0,005	0,002	2,678	0,004
7		0,444	0,188	0,077	0,033	0,015	0,007	0,003	2,573	0,005

8		0,556	0,248	0,107	0,047	0,021	0,010	0,005	2,468	0,007
9			0,315	0,141	0,064	0,030	0,014	0,007	2,363	0,009
10			0,387	0,184	0,085	0,041	0,020	0,010	2,258	0,012
11			0,461	0,230	0,111	0,054	0,027	0,014	2,153	0,016
12			0,539	0,285	0,142	0,071	0,036	0,019	2,048	0,020
13				0,341	0,177	0,091	0,047	0,025	1,943	0,026
14				0,404	0,217	0,114	0,060	0,032	1,838	0,033
15				0,467	0,262	0,141	0,076	0,041	1,733	0,041
16				0,533	0,311	0,172	0,095	0,052	1,628	0,052
17					0,362	0,207	0,116	0,065	1,523	0,064
18					0,416	0,245	0,140	0,080	1,418	0,078
19					0,472	0,286	0,168	0,097	1,313	0,094
20					0,528	0,331	0,198	0,117	1,208	0,113
21						0,377	0,232	0,139	1,102	0,135
22						0,426	0,268	0,164	0,998	0,159
23						0,475	0,306	0,191	0893	0,185
24						0,525	0,347	0,221	0,788	0,215
25							0,389	0,253	0,683	0,247
26							0,433	0,87	0,578	0,282
27							0,478	0,323	0,473	0,318
28							0,522	0,360	0,368	0,356
29								0,399	0,263	0,396
30								0,439	0,158	0,437
31								0,480	0,052	0,481
32								0,520		

Modifiée d'après Siegel 1956.