Document Chapitre 1 - Comparaisons de deux distributions

1 Comparaison de deux distributions sur deux échantillons indépendants *Test de Wilcoxon-Mann-Whitney*

Exemple 1

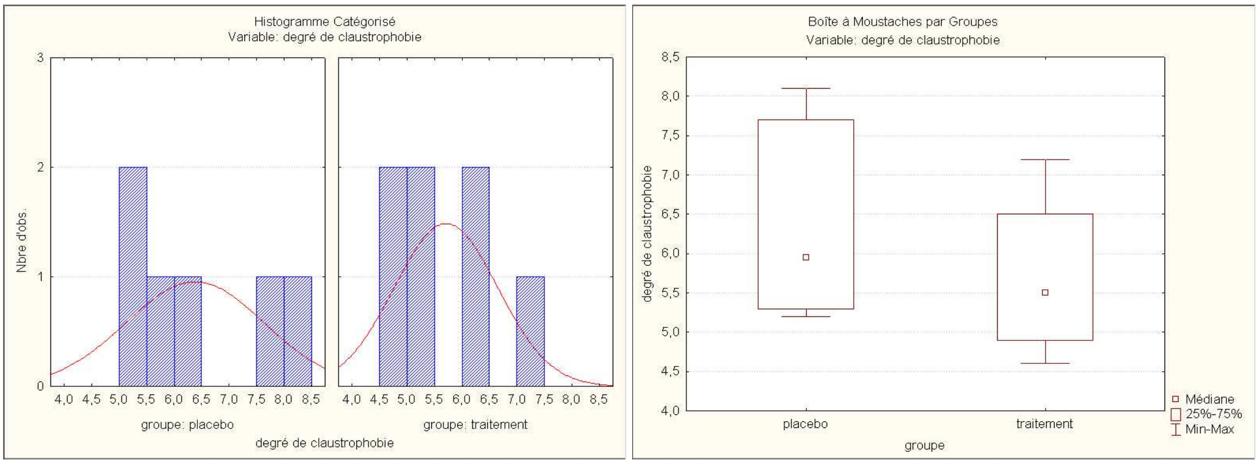
Pour étudier l'efficacité d'un traitement contre la claustrophobie, 13 personnes atteintes de claustrophobie ont été réparties au hasard dans 2 groupes de 6 et 7 personnes.
Les personnes du premier groupe ont reçu un placebo et celles du second groupe le traitement. Après 15 jours de traitement, on a évalué le degré de claustrophobie des 13 personnes, noté dans le tableau suivant

placebo	5,2	5,3	5,6	6,3	7,7	8,1	
traitement	4,6	4,9	5,1	5,5	6,1	6,5	7,2

Peut-on au risque $\alpha = 5\%$, accepter l'hypothèse que le traitement est efficace ?

Analyse descriptive des données

FIGURE 1 – Histogrammes et boîtes à moustaches de X et de Y



Statistiques de Wilcoxon et Mann-Whitney
On range par ordre croissant l'ensemble des $n = 13$ valeurs : les rangs vont de 1 à 13.

placebo	x_i	5,2	5,3	5,6	6,3	7,7	8,1		
	$rang(x,y)$	4	5	7	9	12	13		$w_x = 50$
traitement	y_i	4,6	4,9	5,1	5,5	6,1	6,5	7,2	
	$rang(x,y)$	1	2	3	6	8	10	11	$w_y = 41$

$$\left.\begin{array}{lcl} \text{pour } W_x : w_x & = & 4+5+7+9+12+13 = 50 \\ \text{pour } W_y : w_y & = & 1+2+3+6+8+10+11 = 41 \end{array}\right\} \quad w_x + w_y = 50 + 41 = 91 \text{ et } \frac{n(n+1)}{2} = \frac{13\times 14}{2} = 91$$

W_x varie de 21 à 63 et W_y varie de 28 à 70.

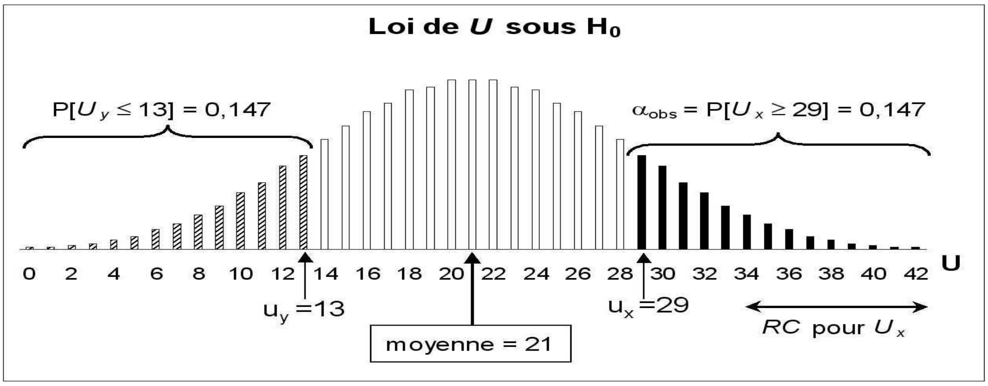
$$\left.\begin{array}{lcl} \text{pour } U_x : u_x & = & 50 - \frac{6\times 7}{2} = 50 - 21 = 29 \\ \text{pour } U_y : u_y & = & 41 - \frac{7\times 8}{2} = 41 - 28 = 13 \end{array}\right\} \textit{vérification} : u_x + u_y = 29 + 13 = 42 \text{ et } n_1 \times n_2 = 6 \times 7 = 42$$

Le domaine de variation de U_x et de U_y : de 0 à $n_1 \times n_2 = 6 \times 7 = 42$.

Lois des statistiques de Mann-Whitney sous H_0 et niveau de signification du test

- **loi exacte**
  Sous $H_0 : X \equiv Y$ U_x et U_y suivent la même loi définie sur $\{0, \ldots, 42\}$ symétrique,
  de moyenne (et milieu) $\mu(U) = \frac{n_1 \times n_2}{2} = \frac{42}{2} = 21$ (Figure 2).
  $\alpha_{obs} = P_{H_0}[U \geq u_x] = P_{H_0}[U \geq 29] \overset{\text{sym}}{=} P_{H_0}[U \leq u_y] = P_{H_0}[U \leq 13] = 0,147$ (table de Mann-Whitney)
- **approximation normale**
  Loi normale de moyenne $\mu(U) = \frac{n_1 n_2}{2} = \frac{6\times 4}{2} = 21$ et de variance $var(U) = \frac{n_1 n_2 (n+1)}{12} = \frac{6\times 7\times 14}{12} = 49$
  La valeur observée de Z : $z_{obs} = \frac{u_x - \mu(U)}{\sqrt{var(U)}} = \frac{29-21}{\sqrt{49}} = \frac{8}{7} \simeq 1,142857$
  $\alpha_{obs} \simeq P[Z \geq z_{obs}] = P[Z \geq 1,142857] = 1 - 0,873451 = 0,126549$ (table de la loi $\mathcal{N}(0,1)$)

FIGURE 2 – Diagramme en bâtons de la loi de U sous H_0 pour $n_1 = 6$ et $n_2 = 7$



Résultats obtenus avec STATISTICA

variable	SommeRgs placebo	SommeRgs traitement	U	Z	niv. p	Z ajusté	niv. p	N Actif placebo	N Actif traitement	2*(1-p) p exact
degré de claustrophobie	50,00000	41,00000	13,00000	1,071429	0,283978	1,071429	0,283978	6	7	0,294872

Test U de Mann-Whitney (exemple1.sta)
Par var. groupe
Tests significatifs marqués à p <,05000

$$z^c_{obs} = \begin{cases} \frac{u_x - \mu(U) - 0,5}{\sqrt{var(U)}} & \text{si} \quad u_x - \mu(U) > 0 \\ \frac{u_x - \mu(U) + 0,5}{\sqrt{var(U)}} & \text{si} \quad u_x - \mu(U) < 0 \end{cases}$$

Exemple 2

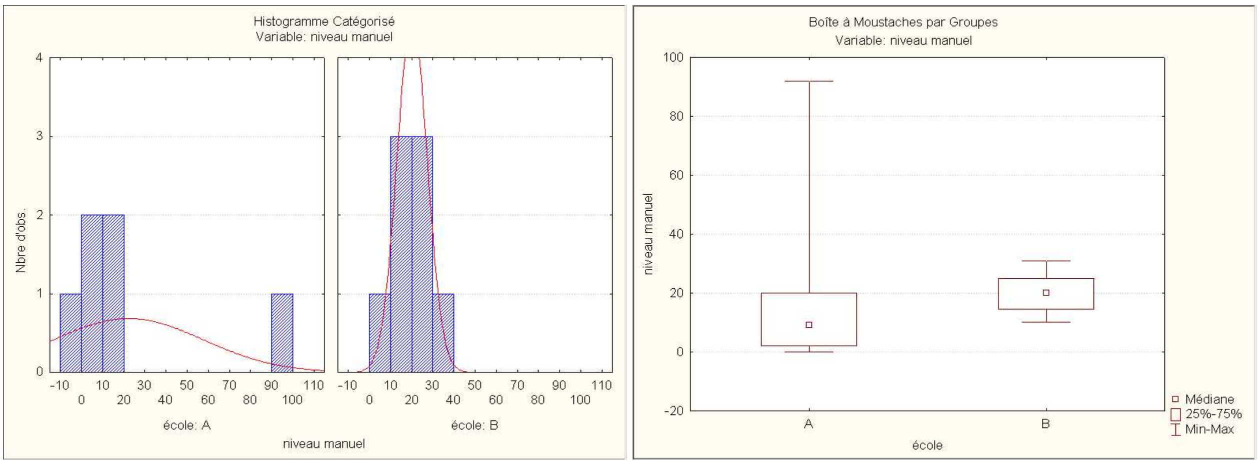
Les niveaux de compétences manuelles des élèves de deux écoles maternelles A et B ont été mesurées pour 6 élèves de l'école A et 8 de l'école B, tirés au sort parmi les élèves de grande section de chaque école, et consignés dans le tableau suivant

école A	20	12	0	2	6	92		
école B	31	14	15	10	21	19	28	22

Peut-on, au risque $\alpha = 10\%$, accepter l'hypothèse que les niveaux de compétences manuelles sont différents dans les deux écoles ?

Analyse descriptive des données

FIGURE 3 – Histogrammes et boîtes à moustaches de X et de Y



Statistiques de Wilcoxon et de Mann-Whitney
On range par ordre croissant l'ensemble des $n = 14$ valeurs : les rangs vont de 1 à 14.

école A	x_i	20	12	0	2	6	92			
	$rang(x, y)$	9	5	1	2	3	14			$w_x = 34$
école B	y_i	31	14	15	10	21	19	28	22	
	$rang(x, y)$	13	6	7	4	10	8	12	11	$w_y = 71$

vérification : la somme $w_x + w_y = 34 + 71 = 105$. Or $n = 14$, donc $\frac{n(n+1)}{2} = \frac{14\times15}{2} = 105$.

$$u_x = w_x - \frac{n_1\times(n_1+1)}{2} = 34 - \frac{6\times7}{2} = 34 - 21 = 13$$
$$u_y = w_y - \frac{n_2\times(n_2+1)}{2} = 71 - \frac{8\times9}{2} = 71 - 36 = 35$$

U_x et de U_y variables quantitatives discrètes définies sur : $\{0, 1, \ldots, n_1 \times n_2 = 6 \times 8 = 48\}$
vérification : $u_x + u_y = 13 + 35 = 48$ est égale à $n_1 \times n_2 = 6 \times 8 = 48$, donc $u_x + u_y = n_1 \times n_2$.

Lois sous H_0, région critique et niveau de signification du test

- **loi exacte**
  Sous H_0, U_x et U_y ont la même loi définie sur $\{0, 1, \ldots, 48\}$ et symétrique de moyenne $\mu(U) = \frac{n_1 n_2}{2} = 24$
  $\alpha_{obs} = 2 \times P_{H_0}[U \leq u_{min}] = 2 \times P_{H_0}[U \leq 13] = 2 \times 0,091 = 0,182 = 18,2\%$ (table de Mann-Whitney)
- **approximation normale**
  Loi normale de moyenne $\mu(U) = \frac{n_1 n_2}{2} = 24$ et de variance $var(U) = \frac{6\times8\times15}{12} = 60$
  $z_{obs} = \frac{u_x - \mu(U)}{\sqrt{var(U)}} = \frac{13-24}{\sqrt{60}} = -1,420094$
  $\alpha_{obs} \simeq 2 \times P[Z \geq 1,420094] = 2 \times (1 - 0,922210) = 0,15558$ (table de la loi $\mathcal{N}(0, 1)$)

Résultats obtenus avec STATISTICA

Test U de Mann-Whitney (exemple2.sta)
Par var. école
Tests significatifs marqués à p <,10000

variable	SommeRgs A	SommeRgs B	U	Z	niv. p	Z ajusté	niv. p	N Actif A	N Actif B	2*(1-p) p exact
niveau manuel	34,00000	71,00000	13,00000	-1,35554	0,175245	-1,35554	0,175245	6	8	0,181152

Exemple 3

On a noté les poids de naissance de bébés (en kg) suivant si la mère avait reçu des soins prénataux depuis le 1er trimestre de grossesse (expérimental) ou le 3ème (en routine) pour resp. 10 et 8 bébés, dans le tableau suivant

$3^{\text{ème}}$ trimestre	1,68	3,83	3,11	2,76	1,70	2,79	1,40	2,66		
1^{er} trimestre	3,05	2,77	2,94	3,38	4,90	2,81	2,80	3,21	3,08	2,95

Peut-on, au risque $\alpha = 5\%$, accepter l'hypothèse que les soins prénataux précoces sont bénéfiques sur le poids de naissance du bébé ?

Analyse descriptive des données
voir Figure 4

Statistiques de Wilcoxon et de Mann-Whitney
On calcule les rangs des individus dans l'ensemble des $n = 18$ individus : les rangs vont de 1 à 18.

$3^{\text{ème}}$ trimestre	x_i	1,68	3,83	3,11	2,76	1,70	2,79	1,40	2,66			
	$rang(x, y)$	2	17	14	5	3	7	1	4			$w_x = 53$
1^{er} trimestre	y_i	3,05	2,77	2,94	3,38	4,90	2,81	2,80	3,21	3,08	2,95	
	$rang(x, y)$	12	6	10	16	18	9	8	15	13	11	$w_y = 118$

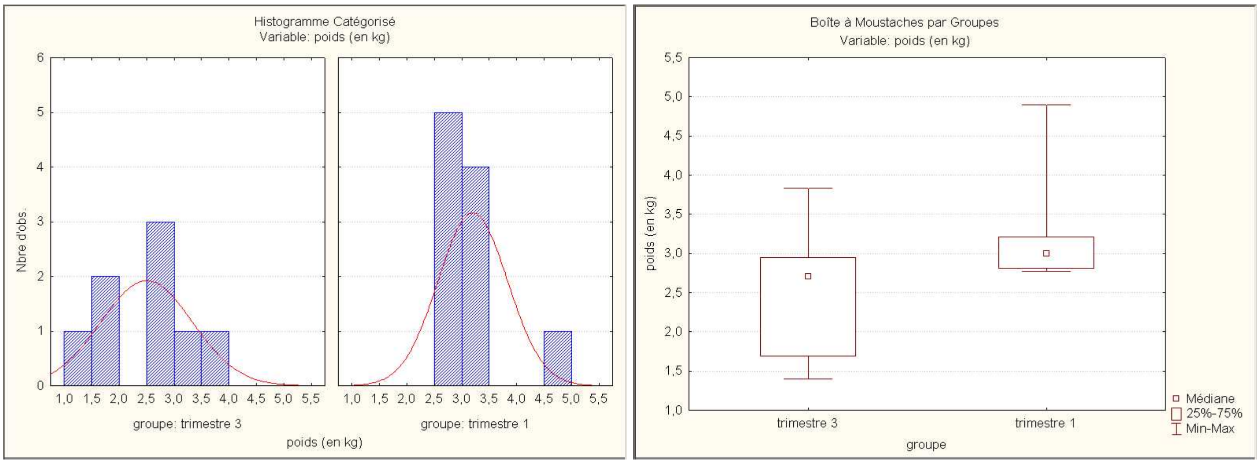
vérification : $w_x + w_y = \frac{n(n+1)}{2}$ puisque $w_x + w_y = 53 + 118 = 171$ et $\frac{n(n+1)}{2} = \frac{18\times19}{2} = 171$.

$$u_x = w_x - \frac{n_1(n_1+1)}{2} = 53 - \frac{8\times9}{2} = 53 - 36 = 17$$
$$u_y = w_y - \frac{n_2(n_2+1)}{2} = 118 - \frac{10\times11}{2} = 118 - 55 = 63$$

vérification : $u_x + u_y = n_1 \times n_2$ puisque $u_x + u_y = 17 + 63 = 80$ et $n_1 n_2 = 8 \times 10 = 80$.

U_x et U_y quantitatives discrètes sur $\{0, \ldots, 80\}$.

Figure 4 – Histogrammes et boîtes à moustaches de X et de Y



Lois sous H_0 région critique et niveau de signification du test

- **loi exacte**
  Sous H_0, U_x et U_y ont la même loi définie sur $\{0, 1, \ldots, 80\}$ et symétrique de moyenne $\frac{n_1 n_2}{2} = 40$.
  $\alpha_{obs} = P_{H_0}[U_x \leq u_x] = P_{H_0}[U \leq 17] = 0,0217 = 2,17\%$ (table de Mann-Whitney)
- **approximation normale**
  Loi normale de moyenne $\mu(U) = \frac{n_1 n_2}{2} = 40$ et de variance $var(U) = \frac{8 \times 10 \times 19}{12} = 95$
  $z_{obs} = \frac{u_x - \mu(U)}{\sqrt{var(U)}} = \frac{17-40}{\sqrt{95}} = -2,04$
  $\alpha_{obs} \simeq P[Z \leq -2,04] = 1 - P[Z \leq 2,04] = 1 - 0,9793 = 0,0207$ (table de la loi $\mathcal{N}(0,1)$)

Résultats obtenus avec STATISTICA

	Test U de Mann-Whitney (exemple3.sta) Par var. groupe Tests significatifs marqués à p <,10000									
variable	**SommeRgs trimestre 3**	SommeRgs trimestre 1	U	Z	niv. p	Z ajusté	niv. p	N Actif trimestre 3	N Actif trimestre 1	2*(1-p) p exact
poids (en kg)	53,00000	118,0000	17,00000	-1,99918	0,045590	-1,99918	0,045590	8	10	0,043421

2 Comparaison de deux distributions sur deux échantillons appariés

2.1 *Test du signe pour deux échantillons appariés*

Exemple 4

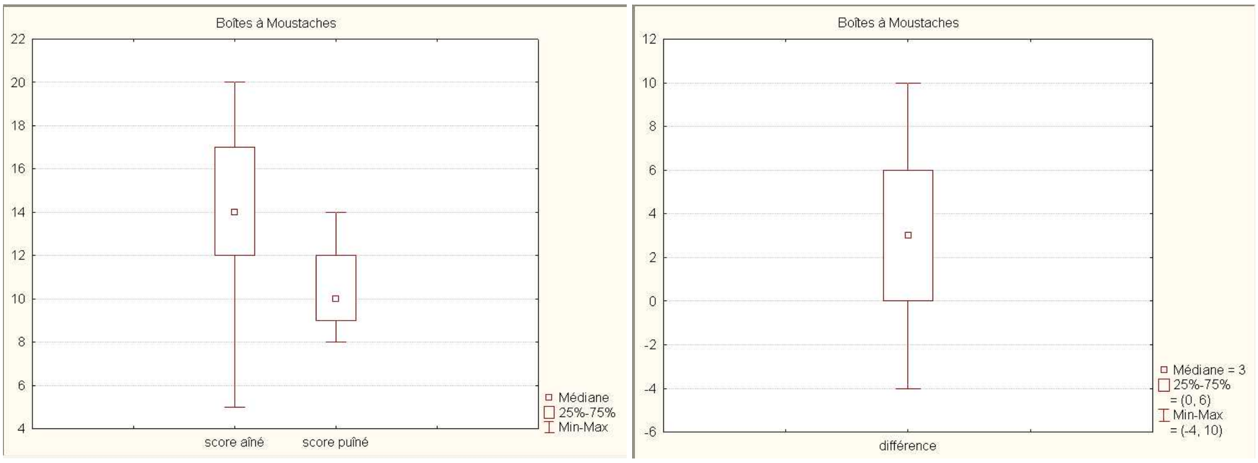
On cherche à savoir si dans les familles, les aînés ont tendance à être plus indépendants que leurs cadets. Lors d'une étude on a procédé à l'évaluation sur une échelle d'indépendance en 25 points, de 9 aînés et du frère ou de la sœur qui suit directement chacun des aînés. On a obtenu les résultats suivants

score de l'aîné	8	12	14	15	5	13	18	20	17
score du puîné	9	10	14	12	9	8	12	10	10

Peut-on répondre à la question, au risque $\alpha = 0,05$?

Analyse descriptive des données

FIGURE 5 – Boîtes à moustaches de X et Y, et de $D = X - Y$



Statistique de test

score de l'aîné	x_i	8	12	14	15	5	13	18	20	17
score du puîné	y_i	9	10	14	12	9	8	12	10	10
différence	$d_i = x_i - y_i$	-1	2	0	3	-4	5	6	10	7
	$signe(d_i)$	$-$	$+$	$\times$	$+$	$-$	$+$	$+$	$+$	$+$

$s_{obs} = 6$ signes '+' pour les $n = 8$ différences non nulles.

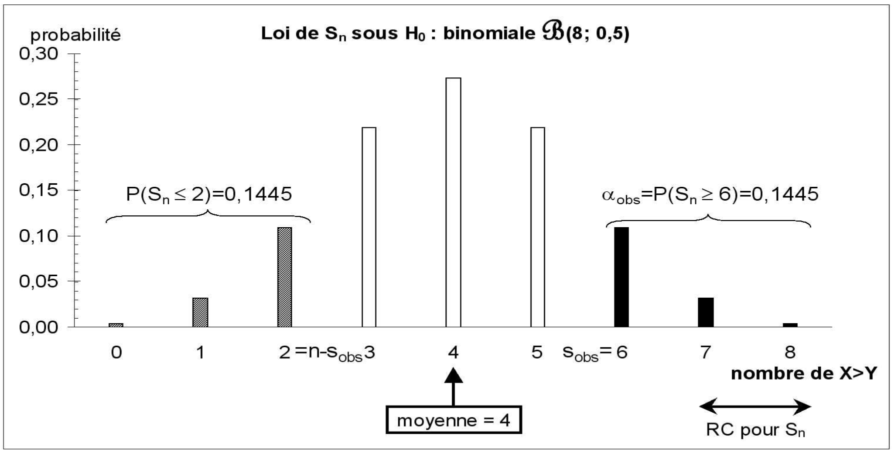
Lois de la statistique de test sous H_0 et niveau de signification du test

- **loi exacte**
  Sous $H_0 : X \equiv Y$ ou $H_0 : p = \frac{1}{2}$ la statistique de test S_n suit une loi binomiale $\mathcal{B}(8, \frac{1}{2})$ définie sur $\{0, 1, \ldots, n = 8\}$ symétrique autour de sa moyenne (et milieu) $\frac{n}{2} = \frac{8}{2} = 4$ (Figure 6).
  $\alpha_{obs} = P_{H_0}[S_n \geq 6] \overset{\text{sym}}{=} P_{H_0}[S_n \leq 2] = 0,1445$ (table de la loi binomiale $\mathcal{B}(8, \frac{1}{2})$)
- **approximation normale**
  Loi normale de moyenne $\mu(S_n) = \frac{n}{2} = 4$ et de variance $var(S_n) = \frac{n}{4} = \frac{8}{4} = 2$
  $z_{obs} = \frac{s_{obs} - \mu(S_n)}{\sqrt{var(S_n)}} = \frac{6-4}{\sqrt{2}} \simeq 1,4142$
  $\alpha_{obs} \simeq P[Z \geq 1,4142] = 1 - P[Z \leq 1,4142] = 1 - 0,92135 = 0,07865$ (table de la loi $\mathcal{N}(0,1)$)

Résultats obtenus avec STATISTICA

	Test des Signes (exemple4.sta) Tests significatifs marqués à p <,10000			
Couples de variables	Nb. Non ex-aequo	%age v < V	Z	niv. p
score aîné & score puîné	8	25,00000	1,060660	0,288844

FIGURE 6 – Diagramme en bâtons de la loi de S_n sous H_0 pour $n = 8$: $\mathcal{B}(8, \frac{1}{2})$



Exemple 5

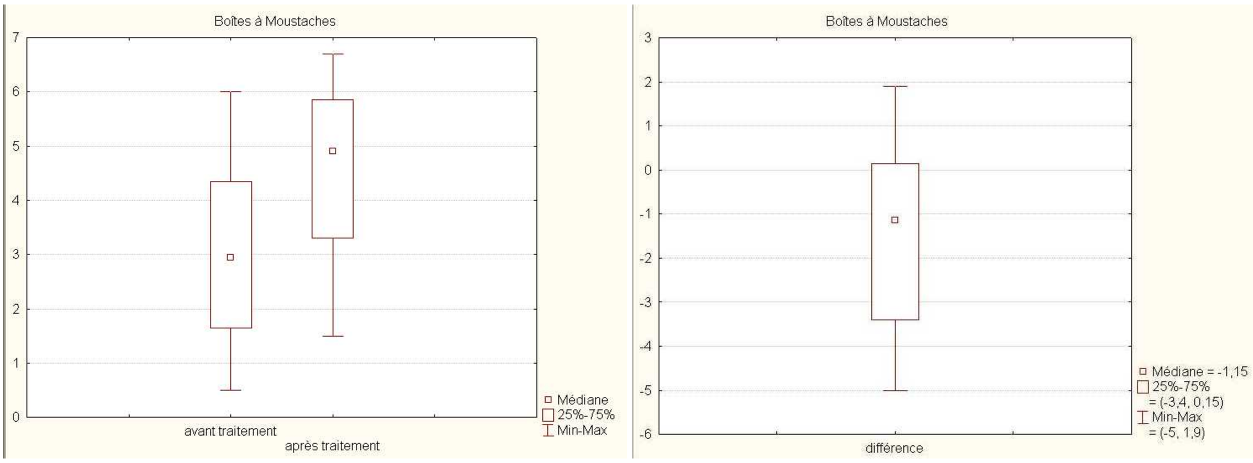
Huit adolescentes de 13 à 18 ans présentant une carence en fer sans anémie ont reçu un traitement pour améliorer leur fonction cognitive. On mesure cette fonction par le score au test d'apprentissage verbal de Hopkin avant et après traitement. On obtient les valeurs du tableau suivant

score avant traitement	5	3,1	2,8	1,5	1,8	0,5	3,7	6
score après traitement	3,1	6	6,7	1,5	3,5	5,5	4,3	5,7

Peut-on conclure, au risque $\alpha = 0,10$ que le traitement est efficace ?

Analyse descriptive des données

FIGURE 7 – Boîtes à moustaches de X et Y, et de $D = X - Y$



Statistique de test

score avant traitement	x_i	5	3,1	2,8	1,5	1,8	0,5	3,7	6
score après traitement	y_i	3,1	6	6,7	1,5	3,5	5,5	4,3	5,7
différence	$d_i = x_i - y_i$	1,9	$-2,9$	$-3,9$	0	$-1,7$	-5	$-0,6$	0,3
	$signe(d_i)$	+	−	−	╳	−	−	−	+

$s_{obs} = 2$ signes '+' pour les $n = 7$ différences non nulles.

Lois sous H_0, région critique et niveau de signification du test

- **<u>loi exacte</u>**
  Sous H_0, S_n quantitative discrète sur $\{0, 1, \ldots, 7\}$ suit une loi $\mathcal{B}(7, \frac{1}{2})$ de moyenne 3,5.
  $\alpha_{obs} = P_{H_0}[S_n \leq s_{obs}] = P_{H_0}[S_n \leq 2] = 0,2266$ (table de la loi binomiale $\mathcal{B}(7, \frac{1}{2})$)

- **approximation normale**

  Loi normale de moyenne $\frac{7}{2} = 3,5$ et de variance $\frac{7}{4} = 1,75$

  $z_{obs} = \frac{s_{obs}-3,5}{\sqrt{1,75}} = \frac{2-3,5}{\sqrt{1,75}} = -1,1339$

  $\alpha_{obs} \simeq P_{H_0}[Z \leq -1,1339] = 1 - 0,87158 = 0,12842$ (table de la loi $\mathcal{N}(0,1)$)

Résultats obtenus avec STATISTICA

	Test des Signes (exemple5.sta) Tests significatifs marqués à p <,20000			
Couples de variables	Nb. Non ex-aequo	%age v < V	Z	niv. p
avant traitement & après traitement	7	71,42857	0,755929	0,449692

2.2 *Test de Wilcoxon (signes et rangs) pour deux échantillons appariés*

Exemple 4

Statistiques de Wilcoxon

score de l'aîné	x_i	8	12	14	15	5	13	18	20	17
score du puîné	y_i	9	10	14	12	9	8	12	10	10
différence	$d_i = x_i - y_i$	-1	2	0	3	-4	5	6	10	7
	$signe(d_i)$	$-$	$+$	χ	$+$	$-$	$+$	$+$	$+$	$+$
	$\lvert d_i \rvert$	1	2	0	3	4	5	6	10	7
	$rang(\lvert d_i \rvert)$	1	2	χ	3	4	5	6	8	7

$$\begin{aligned} v^+ &= 2+3+5+6+7+8 \\ &= 31 \\ v^- &= 1+4=5 \end{aligned}$$

Ici : $v^+ = 31$, $v^- = 5$ donc $v^+ + v^- = 36$. Or $n = 8$ donc $\frac{n(n+1)}{2} = \frac{8\times 9}{2} = 36$.
V^+ et V^- sont quantitatives discrètes et varient sur le domaine $\{0, 1, \ldots, 36\}$.

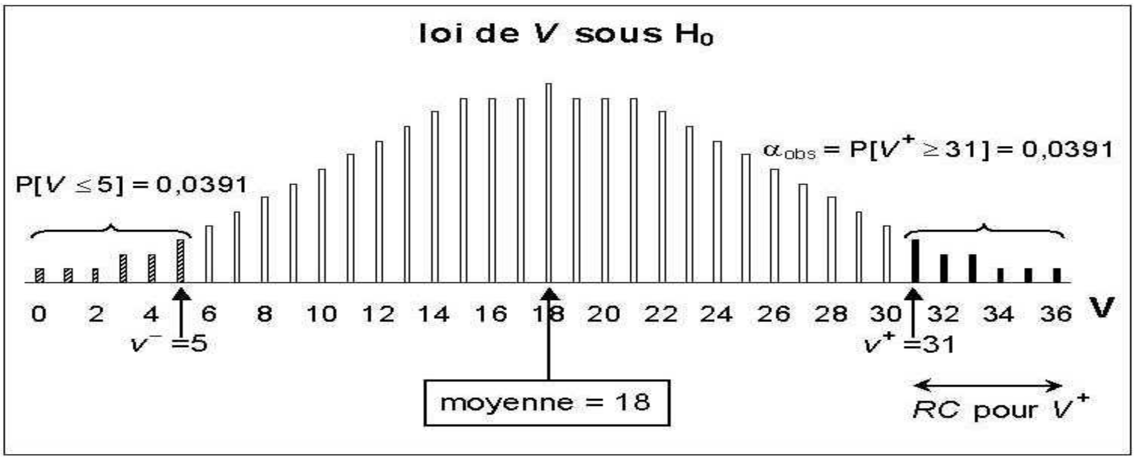
Lois sous H_0, région critique et niveau de signification du test

- **loi exacte**

  Sous $H_0 : X \equiv Y$ V^+ et V^- ont la même distribution, définie sur $\{0, 1, \ldots, \frac{8\times 9}{2} = 36\}$ symétrique autour de la moyenne (et milieu) $\frac{n(n+1)}{4} = \frac{36}{2} = 18$ (Figure 8).

  $\alpha_{obs} = P_{H_0}[V^+ \geq 31] \overset{\text{sym}}{=} P_{H_0}[V \leq 5] = 0,0391$ (table de Wilcoxon)

FIGURE 8 – Diagramme en bâtons de la loi de V sous H_0 pour $n = 8$



- **approximation normale**

  Loi normale de moyenne $\mu(V) = \frac{n(n+1)}{4} = \frac{8\times 9}{4} = 18$ et de variance $var(V) = \frac{n(n+1)(2n+1)}{24} = \frac{8\times 9\times 17}{24} = 51$

  $z_{obs} = \frac{V^+ - \mu(V)}{\sqrt{var(V)}} = \frac{31-18}{\sqrt{51}} \simeq 1,82036$

  $\alpha_{obs} \simeq P[Z \geq 1,82036] = 1 - P[Z \leq 1,82036] = 1 - 0,96565 = 0,03435$ (table de la loi $\mathcal{N}(0,1)$)

Résultats obtenus avec STATISTICA

	Test de Wilcoxon pour Ech. Appariés (exemple4.sta) Tests significatifs marqués à p <,10000			
Couples de variables	N Actifs	T	Z	valeur p
score aîné & score puîné	8	5,000000	1,820364	0,068704

Exemple 5

Statistiques de Wilcoxon

score avant trait.	x_i	5	3,1	2,8	1,5	1,8	0,5	3,7	6
score après trait.	y_i	3,1	6	6,7	1,5	3,5	5,5	4,3	5,7
différence	$d_i = x_i - y_i$	1,9	$-2,9$	$-3,9$	0	$-1,7$	-5	$-0,6$	0,3
	$signe(d_i)$	+	$-$	$-$	╳	$-$	$-$	$-$	+
	$\|d_i\|$	1,9	$2,9$	$3,9$	╳	$1,7$	5	$0,6$	0,3
	$rang(\|d_i\|)$	4	5	6	╳	3	7	2	1

$$\begin{aligned} v^+ &= 4+1=5 \\ v^- &= 2+3+5+6+7 \\ &= 23 \end{aligned}$$

Vérification : $v^+ = 5$, $v^- = 23$ donc $v^+ + v^- = 28$. Or $n = 7$ donc $\frac{n(n+1)}{2} = \frac{7\times 8}{2} = 28$.
V^+ et V^- sont quantitatives discrètes et varient sur le domaine $\{0, 1, \ldots, 28\}$.

Lois sous H_0, région critique et niveau de signification du test

- **loi exacte**

  Sous $H_0 : X \equiv Y$ $\quad V^+$ et V^- ont la même distribution, définie sur $\{0, 1, \ldots, 28\}$ symétrique autour de la moyenne (et milieu) $\frac{n(n+1)}{4} = \frac{28}{2} = 14$.

  $\alpha_{obs} = P_{H_0}[V^+ \leq v^+] = P_{H_0}[V^+ \leq 5] = 0,0781$ (table de Wilcoxon)

- **approximation normale**

  La loi normale de moyenne $\mu(V) = \frac{n(n+1)}{4} = \frac{7\times 8}{4} = 14$ et de variance $var(V) = \frac{7\times 8\times 15}{24} = 35$

  $z_{obs} = \frac{V^+ - \mu(V)}{\sqrt{var(V)}} = \frac{5-14}{\sqrt{35}} \simeq -1,521278$

  $\alpha_{obs} \simeq P[Z \leq -1,521278] = 1 - P[Z \leq 1,521278] = 1 - 0,9359 = 0,0641$ (table de la loi $\mathcal{N}(0,1)$)

Résultats obtenus avec STATISTICA

	Test de Wilcoxon pour Ech. Appariés (exemple5.sta) Tests significatifs marqués à p <,20000			
Couples de variables	N Actifs	T	Z	valeur p
avant traitement & après traitement	7	5,000000	1,521278	0,128191