Comparaisons de 2 échantillons

Comparaisons de 2 échantillons indépendants

Le test t pour échantillons indépendants (student t test)

Objectif du test

Tester la différence entre les moyennes de deux échantillons indépendants (les deux échantillons sont composés d'éléments non appariés).

Principe

Soit deux échantillons avec n_1 et n_2 éléments respectivement.
Nous voulons savoir si la différence entre $\bar{x}_1$ et $\bar{x}_2$ reflète une différence significative des moyennes des populations statistiques dont sont extraits les échantillons, ou si l'écart observé n'est dû qu'aux fluctuations naturelles de l'échantillonnage.

Nous calculerons **une statistique *t* de Student** à partir des données et nous déterminerons la probabilité de cette valeur à l'aide de la distribution de Student à **$\upsilon = n_1 + n_2 - 2$ degrés de liberté.**

Cette distribution nous permettra de savoir si la probabilité de rencontrer notre valeur de *t* sous H_0 est plus grande ou plus petite que notre seuil α. Nous prendrons alors la décision de ne pas rejeter ou de rejeter H_0.

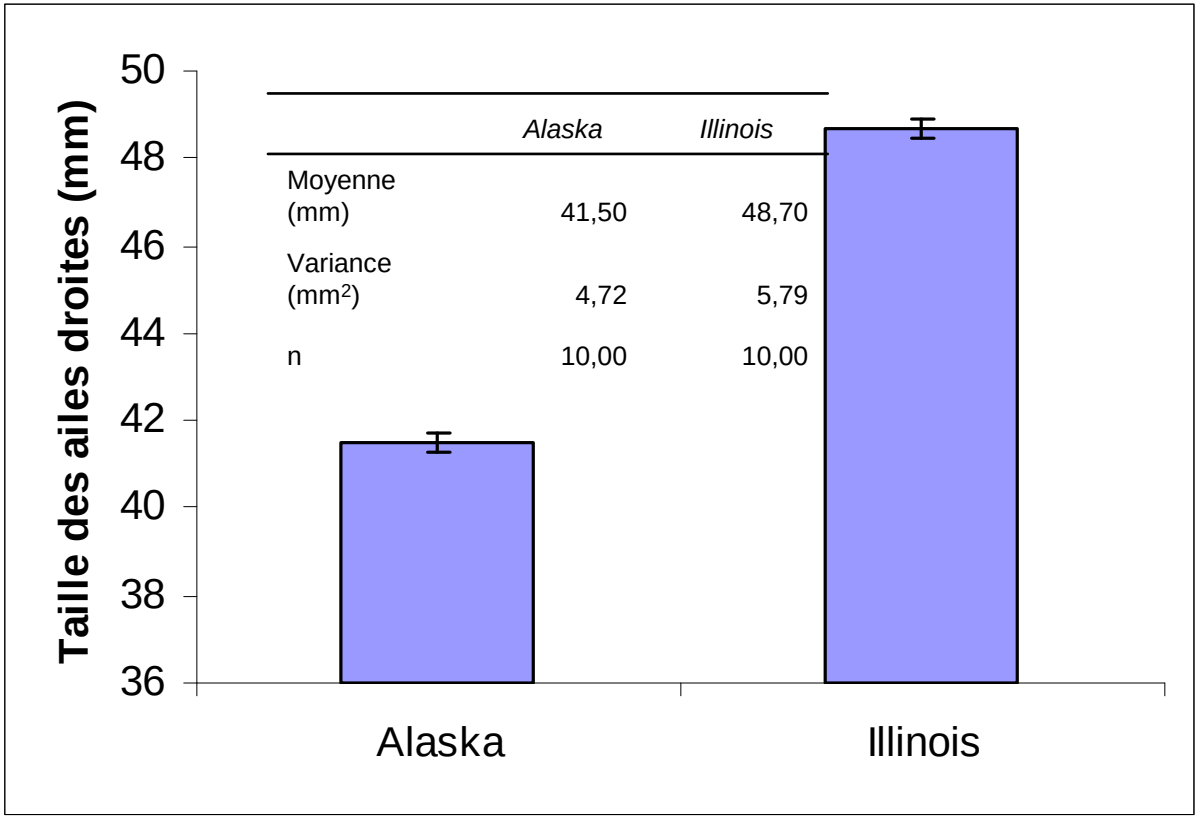
Règles de décision

H_0	H_1	Rejet de H_0 si
	$\mu_1 \neq \mu_2$	$\lvert t_{calc} \rvert \geq \lvert t_{\alpha/2,\nu} \rvert$
$\mu_1 = \mu_2$	$\mu_1 > \mu_2$	$t_{calc} \geq t_{\alpha,\nu}$
	$\mu_1 < \mu_2$	$t_{calc} \leq -t_{\alpha,\nu}$

Robert va à la chasse au *Papilio glaucus*

Longueur des ailes antérieures droites (mm) des mâles de *Papilio glaucus* échantillonnés en Alaska et en Illinois une certaine année.

Alaska	Illinois
42	51
41	48
41	49
37	48
44	47
43	46
43	47
40	47
40	50
44	54



Étape 1: Question biologique

Y a-t-il une différence au niveau de la longueur des ailes des mâles de *Papilio glaucus* en Alaska et en Illinois ? (Le climat de l'Alaska empêche-t-il les papillons de s'y développer autant ?)

Étape 2: Déclaration des hypothèses

H_0: Il n'y a pas de différence entre la longueur moyenne des ailes des mâles de *Papilio glaucus* d'Alaska et d'Illinois

$$\mu_{Alaska} = \mu_{Illinois}$$

H_1: La longueur moyenne des ailes des mâles de *Papilio glaucus* d'Alaska est plus **petite** que celle des mâles d'Illinois (hypothèse unilatérale).

$$\mu_{Alaska} < \mu_{Illinois}$$

Étape 3: Choix du test

Comme n_1 **ou** n_2 **est plus petit que 30**, le test statistique utilisé est un test de *t* de Student où:

$$t_{calc} = \frac{\bar{x}_A - \bar{x}_I}{s_{pd}\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_I}\right)}} \qquad \text{où} \qquad s_{pd} = \sqrt{\frac{(n_A - 1)s_A^2 + (n_I - 1)s_I^2}{n_A + n_I - 2}}$$

Si *n*1 **et** *n*2 ≥ 30, on utilise le test Z

Étape 4: Conditions d'application

Indépendance des observations.
Normalité des distributions de données des échantillons.
Équivariance (ou homoscédasticité) des échantillons….????

Pour vérifier l'homoscédasticté, le mini-test de *F*

H_0: les variances des deux échantillons sont égales : $\sigma^2_1 = \sigma^2_2$
H_1: Les variances des deux échantillons sont différentes : $\sigma^2_1 \neq \sigma^2_2$

Si les échantillons sont tirés de populations normales (conditions de normalité), le rapport de leurs variances $\mathbf{s^2_1/s^2_2}$ suivra une distribution de *F* à $\mathbf{\upsilon 1 = n_1 - 1}$ et $\mathbf{\upsilon 2 = n_2 - 1}$ degrés de liberté.

On rejettera H_0 au seuil $\alpha = 0{,}05$ si $F_{\text{calc}} \geq F_{(\alpha/2:\, \upsilon 1\, ,\upsilon 2\,)}$

Attention!!! Il faut **toujours** mettre la **variance la plus grande** au **NUMÉRATEUR**!

MINI-TEST de F!

$H_0: \sigma^2_1 = \sigma^2_2$ $H_1: \sigma^2_1 \neq \sigma^2_2$

$F_{calc} = s^2_I / s^2_A = 5{,}79 / 4{,}72 = 1{,}227$

$F_{(0{,}025:9,9)} = 4{,}03$

Donc $F_{calc} < F_{(\alpha/2:\, \nu1\, ,\nu2\,)}$ et on ne rejette pas H_0: il y a équivariance

Étape 5: Distribution de la variable auxiliaire

Si H_0 est vraie, la variable auxiliaire t_{calc} suivra une distribution de Student à υ = $n_A + n_I - 2 = 10 + 10 - 2 = 18$ **degrés de liberté**.

Étape 6: Règle de décision

On **rejette H_0** au seuil $\alpha = 0{,}05$ si $t_{calc} \leq -t_{(\alpha:\upsilon)} = -t_{(0.05:18)} = -1{,}734$

Étape 7: Calcul du test

	Alaska	*Illinois*
Moyenne	**41,50**	**48,70**
Variance	**4,72**	**5,79**
n	**10**	**10**
S_{pd}	**5,26**	
υ	**18,00**	
t_{calc}	**-7,022**	

$$t_{calc} = \frac{\bar{x}_A - \bar{x}_I}{s_{pd}\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_I}\right)}} \quad \text{où} \quad s_{pd} = \sqrt{\frac{(n_A - 1)s_A^2 + (n_I - 1)s_I^2}{n_A + n_I - 2}}$$

Étape 8: Décision statistique

Puisque $t_{calc} = -7{,}022 < -t_{(0.05:18)} = -1{,}734$, on rejette H_0 au seuil $\alpha = 0{,}05$.

Étape 9: Interprétation biologique

Les papillons mâles ont en moyenne des longueurs **d'ailes antérieures droites plus petites en Alaska qu'en Illinois** car la saison de croissance y est plus courte ce qui défavorise leur croissance.

Test U de Wilcoxon - Mann - Whitney

Objectif du test *U*

Tester la différence entre les moyennes de 2 échantillons indépendants

Principe

La statistique *U* est basée sur le classement en ordre croissant des éléments des deux échantillons.

Si H_0 est vraie (il n'y a pas de différence entre les moyennes), les éléments des deux groupes devraient être uniformément mélangés dans ce classement.
Ce sont donc les **rangs** des données, et non leurs valeurs précises, qui servent dans ce test.

Exemple 1: deux échantillons A et B

A	B
2	3
4	5
6	7
8	9

Si on classe les données en ordre croissant, ça donne:

A	B	A	B	A	B	A	B
2	3	4	5	6	7	8	9

On voit que les éléments de A et B sont uniformément mélangés: on ne rejetterait probablement pas H_0

Exemple 2: deux échantillons C et D

C	D
2	3
3	6
4	7
5	9

Si on classe les données en ordre croissant, ça donne:

C	C	D	C	C	D	D	D
2	3	3	4	5	6	7	9

On voit que les éléments de D sont en général plus grands: on rejetterait probablement H_0

Déclaration des hypothèses du test de *U*

Hypothèse nulle:

H_0: Les rangs des données des deux groupes sont uniformément distribués.

$$\mathbf{P(x_{i,1} > x_{j,2}) = 0,5}$$

Cela signifie que si $x_{i,1}$ est un élément tiré aléatoirement de la première population et $x_{j,2}$ est un élément tiré aléatoirement de la seconde population, il y a une chance sur deux que $x_{i,1}$ soit plus grand que $x_{j,2}$.

Hypothèses contraires:

- **Bilatérale**

H_1: Les rangs des données des deux groupes ne sont pas uniformément distribués.

$$\mathbf{P(x_{i,1} > x_{j,2}) \neq 0,5}$$

- **Unilatérales**

H_1: Les rangs des données du premier groupe sont décalés vers les grandes valeurs.

$$\mathbf{P(x_{i,1} > x_{j,2}) > 0,5}$$

ou H_1: Les rangs des données du premier groupe sont décalés vers les petites valeurs.

$$\mathbf{P(x_{i,1} > x_{j,2}) < 0,5}$$

tests bilatéral seulement!!

Si 9 ≤ max(*n*1, *n*2)

1) On classe les éléments en **ordre croissant** et on leur assigne un **rang**.

2) Si deux éléments ont la **même valeur**, on leur assigne un **rang médian**.

3) On calcule U_1 et U_2 à l'aide des formules suivantes:

$$U_1 = n_1 \times n_2 + \frac{n_1 \times (n_1 + 1)}{2} - R_1$$

$$U_2 = n_1 \times n_2 + \frac{n_2 \times (n_2 + 1)}{2} - R_2$$

où $\boldsymbol{R_1}$ **et** $\boldsymbol{R_2}$ représentent la **somme des rangs** de chacun des échantillons.

$U_1 + U_2 = n_1 \times n_2$

$\boldsymbol{U_{calc}}$ **= min (**$\boldsymbol{U_1}$**,** $\boldsymbol{U_2}$**)** (test bilatéral)

Exemple : 2 échantillons C et D

C	D
2	2
3	4
5	5
5	6
7	6
8	7
9	8
10	11
	12

Calcul de *U* :

Rang	Éléments	C	D
1	2	1,5	
2	2		1,5
3	3	3	
4	4		4
5	5	6	
6	5	6	
7	5		6
8	6		8,5
9	6		8,5
10	7	10,5	
11	7		10,5
12	8	12,5	
13	8		12,5
14	9	14	
15	10	15	
16	11		16
17	12		17
	Sommes	$R_1 = 68,5$	$R_2 = 84,5$

$$U_1 = n_1 \times n_2 + \frac{n_1 \times (n_1 + 1)}{2} - R_1 = 8 \times 9 + \frac{8 \times (8+1)}{2} - 68,5 = 39,5$$

$$U_2 = n_1 \times n_2 + \frac{n_2 \times (n_2 + 1)}{2} - R_2 = 8 \times 9 + \frac{9 \times (9+1)}{2} - 84,5 = 32,5$$

Test bilatéral : U_{calc} = min(U_1, U_2) = 32,5

On compare le U_{cal} au $U_{théorique}$

Si la **valeur de *U* calculée** à partir des données est **plus petite** que la valeur de la table, on **rejette H_0.**

U_{calc} = 32,5
n_C = 8
n_D = 9
α = 0,05 bilatéral

Dans la table, la valeur de *U* critique pour un seuil α = 0,05 bilatéral est $U_{0,05,8,9}$ **= 57**

Puisque le *U* calculé vaut 32,5, et 32,5 < 57 on rejette H_0

Test *U* unilatéral

Si le test est **unilatéral**, il faut **choisir** lequel des deux *U* on doit calculer **en fonction de l'hypothèse contraire H_1**. Le tableau suivant donne la règle de choix:

	H_0: groupe 1 $\leq$ groupe 2	H_0: groupe 1 $\geq$ groupe 2	
——–>	**H_1: groupe 1 > groupe 2**	**H_1: groupe 1 < groupe 2**	<——–
***U* à utiliser**	U_1	U_2	

Si n_1 ou n_2 > 20

U **tend vers une loi normale** ayant pour paramètres:

$$\mu_U = \frac{n_1 \times n_2}{2} \qquad \text{et} \qquad \sigma_U = \sqrt{\frac{n_1 \times n_2 \times (n_1 + n_2 + 1)}{12}}$$

On calcule donc la variable centrée réduite: $$z = \frac{U - \mu_U}{\sigma_U}$$

Test bilatéral : si la valeur de **|*z*|** est **plus grande** que la valeur de ***z* critique** associée au seuil **α*/2***, on **rejette** H_0

Test unilatéral : si la valeur de **|*z*|** est **plus grande** que la valeur de ***z* critique** associée au seuil α, on **rejette** H_0.

*Pour **α = 0,05 (test unilatéral) – z = 1,645***
*Pour un seuil **α/2 = 0,025 (test bilatéral), z = 1,96.***
*Pour un seuil **α = 0,01, z = 2,33 (test unilatéral) et z = 2,575 (test bilatéral).***

Si plusieurs éléments occupent le même rang, une formule corrigée de σ_U doit être utilisée:

$$\sigma_U = \sqrt{\frac{n_1 \times n_2}{n \times (n-1)} \times \left(\frac{n^3 - n}{12} - \sum_{l=1}^{g} E_l \right)}$$

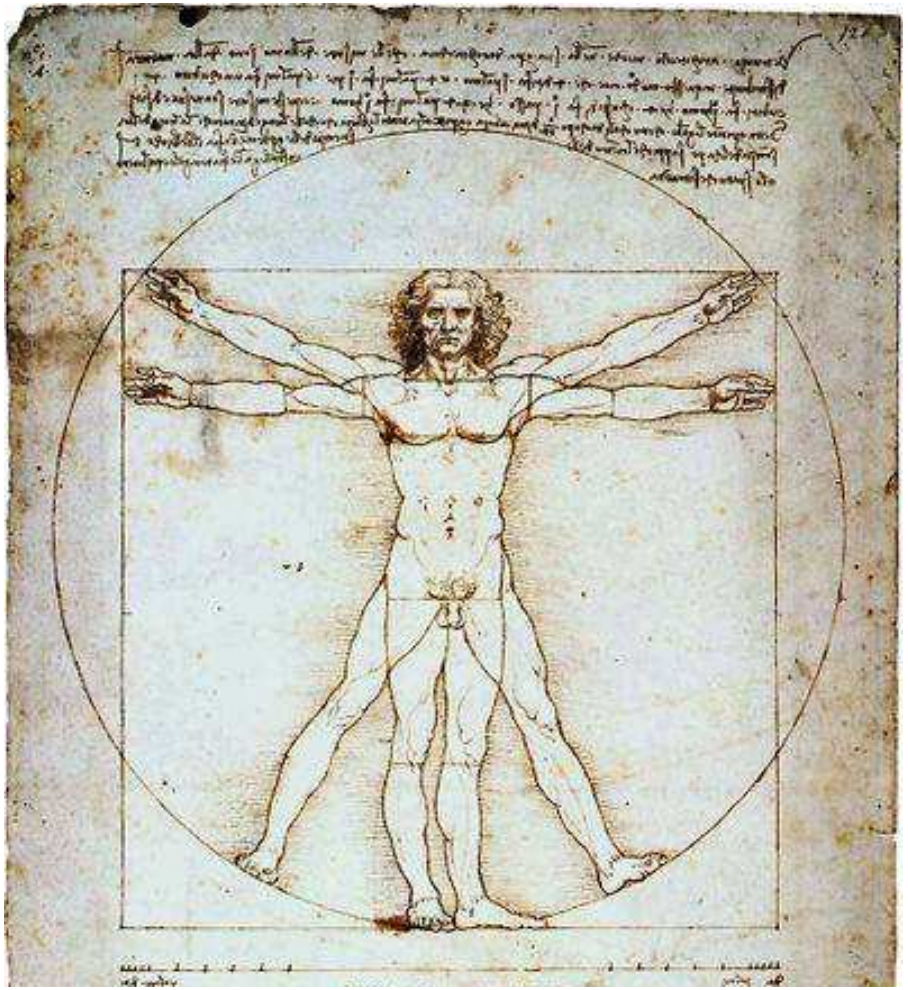
où $n = n_1 + n_2$

g = nombre de rangs avec données ex-aequo

l = le l-ième rang avec ex-aequo

$E_l = \frac{e_l^3 - e_l}{12}$ où e_l = nombre d'observations de rang l

Comparaison de 2 échantillons appariés



Objectif

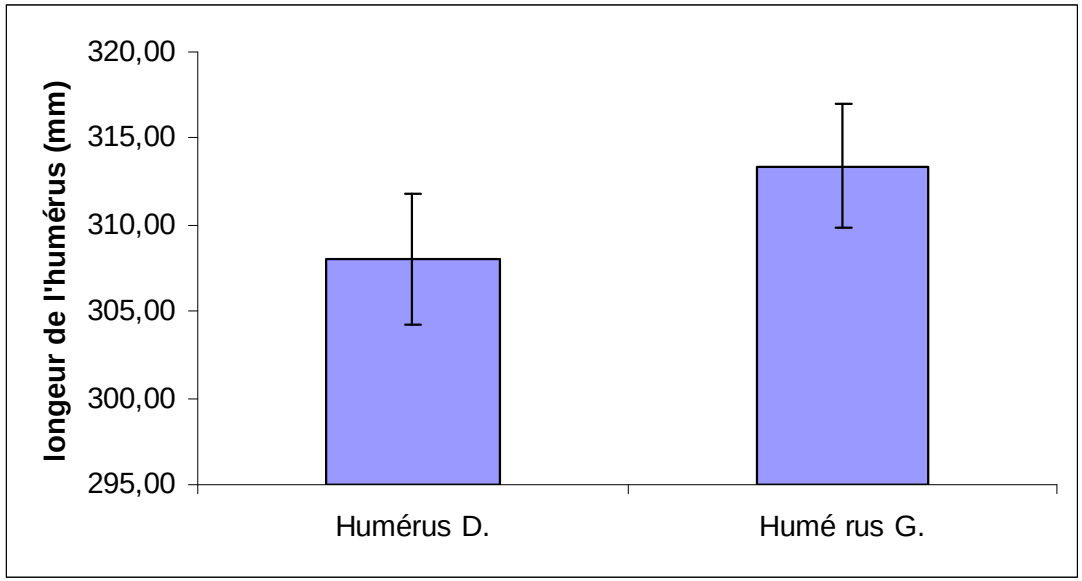
Tester la différence entre les moyennes de 2 échantillons appariés [i.e. les deux échantillons ont les mêmes éléments ou des éléments liés par au moins un critère

Le test t pour échantillons appariés

<u>Principe</u> : analyse des différences ***d*** observées pour chaque paire d'observations

Exemple. Comparaison des longueurs moyennes (mm) des humérus droit et gauche de dix squelettes de femmes. D'après Jolicoeur (1998).

# squelette	Humérus D.	Humérus G.
1	311	315
2	302	306
3	301	311
4	322	333
5	312	316
6	285	292
7	305	308
8	310	318
9	328	326
10	304	309



	Humérus D.	*Humérus G.*
Moyenne	**308,00**	**313,40**
Variance	**140,44**	**126,71**

Étape 1. Question biologique

Y a-t-il une différence de longueur entre l'humérus droit et l'humérus gauche chez les femmes ?

Étape 2. Déclaration des hypothèses

H_0: Il n'y a pas de différence entre la longueur moyenne des humérus gauche et droit chez les femmes

$$\mu_D = \mu_G$$

H_1: Il y a une différence entre la longueur moyenne des humérus gauche et droit chez les femmes

$$\mu_D \neq \mu_G$$

Étape 3. Choix du test

Le test statistique utilisé est un test de *t* de Student pour échantillons appariés:

$$t_{\bar{d}} = \frac{\bar{d}}{s_{\bar{d}}} \quad \text{où} \quad s_{\bar{d}} = \frac{s_d}{\sqrt{n}} \quad \text{et} \quad \bar{d} = \bar{x}_1 - \bar{x}_2$$

Étape 4. Conditions d'application

Les échantillons sont appariés.
Les différences de longueur suivent une distribution normale.

Étape 5: Distribution de la variable auxiliaire

Si H_0 est vraie, la variable auxiliaire t_d suivra une distribution de Student à $\upsilon = n - 1 = 10 - 1 = 9$ ddl, où ***n*** est le nombre de ***différences*** de longueur.

Étape 6. Règle de décision

On rejette H_0 au seuil $\alpha = 0{,}05$ si $|t_d| \geq |t_{(\alpha/2:\upsilon)}|$ où $t_{(\alpha/2:\upsilon)} = t_{(0{,}025:9)} = 2{,}262$.

Étape 7. Calcul du test

# squelette	Humérus D.	Humérus G.	**d**
1	311	315	**-4**
2	302	306	**-4**
3	301	311	**-10**
4	322	333	**-11**
5	312	316	**-4**
6	285	292	**-7**
7	305	308	**-3**
8	310	318	**-8**
9	328	326	**2**
10	304	309	**-5**

$$\bar{d} = -5{,}4$$

$$s_d = 3{,}78$$

$$s_{\bar{d}} = 1{,}19$$

$$t_{\bar{d}} = -4{,}52$$

Étape 8. Décision statistique

Puisque $|t_d| > |t_{(\alpha:\upsilon)}|$, on rejette H_0 au seuil $\alpha = 0,05$.

Étape 9. Interprétation biologique

Les femmes ont des humérus de longueurs différentes car le bras …à vous d'imaginer la conclusion☺

Le test T de Wilcoxon pour échantillons appariés (Wilcoxon T test)

Si n ≤ 60

Exemple : 12 sujets sont entraînés à la réalisation d'une tâche. L'épreuve permettant de mesurer leur efficacité apporte les résultats suivant :

sujets	A	B	C	D	E	F	G	H	I	J	K	L
Av Ent	15	20	2	42	3	10	29	20	30	45	60	19
Ap Ent	45	60	20	45	24	17	28	57	58	47	56	38

Étape 1. Question biologique

Peut-on conclure à l'efficacité de l'entraînement ?

Étape 2. Déclaration des hypothèses

Hypothèse nulle:

H_0: Les rangs des données des deux groupes sont uniformément distribués.

$$\mathbf{P(x_{i,1} > x_{j,2}) = 0,5}$$

- **Bilatérale**

H_1: Les rangs des données des deux groupes ne sont pas uniformément distribués.

$$\mathbf{P(x_{i,1} > x_{j,2}) \neq 0,5}$$

- **Unilatérales**

H_1: Les rangs des données du premier groupe sont décalés vers les grandes valeurs.

$$\mathbf{P(x_{i,1} > x_{j,2}) > 0,5}$$

ou H_1: Les rangs des données du premier groupe sont décalés vers les petites valeurs.

$$\mathbf{P(x_{i,1} > x_{j,2}) < 0,5}$$

Étape 3. Choix du test

Test T de Wilcoxon. On détermine la valeur des différences (Di = après - avant) et on classe les différences dans l'ordre croissant de leur valeur absolue. On cherche s'il y a une différence entre la somme des différences des rangs > 0 **(T+)** et celles des rangs < 0 **(T-).**

Étape 4. Conditions d'application

- 2 échantillons appariés

Étape 5: Distribution de la variable auxiliaire

Si H_0 est vraie, la variable auxiliaire T_{obs} suivra une distribution de T à **$\upsilon = n$** = 12 ddl, où ***n*** est le nombre de ***différences***

Étape 6: Règle de décision

On rejette H_0 au seuil $\alpha = 0{,}05$ (bilatéral) si $\boldsymbol{T_{obs} \leq T_{(0,05,12)} = 13}$

Étape 7: Calcul du test

sujets	A	B	C	D	E	F	G	H	I	J	K	L
Di	30	40	18	3	21	7	- 1	37	28	2	- 4	19
rang	10	12	6	3	8	5	1	11	9	2	4	7

T+ = 10 + 12 + 6 + 3 + 8 + 5 + 11 + 9 + 2 + 7 = **73**
T- = 1 + 4 **= 5**

T_{obs} = la plus petite valeur entre *T+* et *T-* = *T-* = *5*

Étape 8. Décision statistique

Puisque $T_{obs} < T_{(0,05,12)}$ on rejette H_0 au seuil α = 0,05.

Étape 9. Interprétation biologique

L'entraînement permet d'améliorer la réalisation de la tâche