Comparaison de moyennes

I – Comparer une moyenne à une valeur donnée

1 – Test de statistique

Echantillon E de taille n
de moyenne m
de variance s^2

Population $\mathcal{P}$ de moyenne µ

Question : l'échantillon E provient-il de la population $\mathcal{P}$? = Existe-t-il une différence significative entre m et µ ?

Hypothèses :
H_0 : hypothèse nulle = E provient de $\mathcal{P}$
H_1 : hypothèse alternative = E ne provient pas de $\mathcal{P}$

a – Comment faire un test statistique ?

1. Choisir le risque α = risque de première espèce, c'est le risque de rejeter H_0 alors que H_0 est vraie. Généralement α = 5%, parfois α = 1%
2. Poser H_0 et H_1
3. Sous H_0, on calcule un paramètre (U_C) dont on connait la distribution, pour pouvoir comparer m et µ :

$$U_C = \frac{m - \mu}{\frac{s}{\sqrt{n}}}$$

4. Chercher dans la table U_α
5. Comparer $|U_C|$ et U_α :
   - Si $|U_C| > U_\alpha$: au risque α on rejette H_0 et on accepte H_1
   - Si $|U_C| < U_\alpha$: au risque α on accepte H_0

1 – Pour les grands échantillons (n ≥ 30)

Si on connait m, s², µ :
H_0 : il n'existe pas de différence significative entre l'échantillon et la population
H_1 : il existe une différence significative entre l'échantillon et la population

Sous H_0 : $U_C = \frac{m - \mu}{\frac{s}{\sqrt{n}}}$ suit une loi normale

Si α = 5%, U_t = 1,96 / Si α = 1%, U_t = 2,57

Comparaison :
- Si $|U_C| > U_t$: au risque α on rejette H_0 et on accepte H_1
- Si $|U_C| < U_t$: au risque α on accepte H_0 (sans certitude que H_0 est vraie), donc on ne met pas en évidence une différence significative

Si on connait l'écart-type σ de $\mathcal{P}$:

Même procédure sauf que $U_C = \dfrac{m - \mu}{\dfrac{\sigma}{\sqrt{n}}}$

2 – Pour les petits échantillons (n < 30)

Si on connait σ : *cf. grands échantillons*

Si σ inconnu :

Sous H_0, $t_C = \dfrac{m - \mu}{\dfrac{s}{\sqrt{n}}}$ suit une loi de Student à (n-1) ddl (degré de liberté)

On trouve le t_t dans la table de Student

Comparaison :

- Si $|t_C| > t_t$: au risque α on rejette H_0 et on accepte H_1
- Si $|t_C| < t_t$: au risque α on accepte H_0

b – Exemple

On mesure le temps de réaction chez des souris exposées à une stimulation au médicament X.

μ = 23,7 s

n = 100

m = 22,9 s

s^2 = 13,98 s^2

1. Est-ce que le médicament X modifie le temps de réaction des souris ?
2. Pour n=16 ?

1. H_0 = le médicament ne modifie pas le temps de réaction

H_1 = le médicament modifie le temps de réaction

Sous H_0 :

$$U_C = \frac{m - \mu}{\frac{s}{\sqrt{n}}} = \frac{22{,}9 - 23{,}7}{\frac{\sqrt{13{,}98}}{\sqrt{100}}} = -2{,}14$$

α = 5% → U_t = 1,96

$|U_C| > U_t$: au risque 5%, on rejette H_0, donc le médicament modifie le temps de réaction.

2. n=16, H_0 et H_1 sont les mêmes

Sous H_0 :

$$t_C = \frac{m - \mu}{\frac{s}{\sqrt{n}}} = \frac{22{,}9 - 23{,}7}{\frac{\sqrt{13{,}98}}{\sqrt{16}}} = -0{,}86$$

α = 5% et ddl = (n-1) = (16-1) = 15 → t_t = 2,13

$|t_C| < t_t$: au risque 5%, on accepte H_0, donc on ne montre pas que le médicament modifie le temps de réaction.

II – Comparer les moyennes observées sur des échantillons indépendants

1 – Deux grands échantillons

E_1 : m_1, s_1^2, n_1
E_2 : m_2, s_2^2, n_2
$\mathcal{P}$: µ

Si n_1 et $n_2 \geq 30$:

- H_0 : les 2 échantillons sont de la même population, il n'y a pas de différence significative entre les 2 moyennes
- H_1 : les 2 échantillons ne sont pas de la même population, il existe une différence significative entre les 2 moyennes

Sous H_0,

m_1 et m_2 suivent une loi normale $\mathcal{N}(\mu ; \frac{\sigma^2}{n})$

m_1 suit une loi normale $\mathcal{N}(\mu ; \frac{s_1^2}{n_1})$

m_2 suit une loi normale $\mathcal{N}(\mu ; \frac{s_2^2}{n_2})$

$d = m_1 - m_2$
$E(d) = \mu - \mu = 0$
$s^2_d = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$

$U = \frac{d-0}{s_d}$ suit une loi $\mathcal{N}(0 ; 1)$ d'où $U_C = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Soit α = 5% → $U_t = 1{,}96$
Comparaison :

- Si $|U_C| > U_t$: au risque α on rejette H_0 et on accepte H_1 : les 2 moyennes proviennent de 2 populations différentes, il existe une différence significative entre les 2 moyennes
- Si $|U_C| < U_t$: au risque α on accepte H_0

Exemple

Poids des nouveau-nés dans une maternité :

♂	$n_1 = 41$	$m_1 = 3{,}4$ kg	$s_1 = 0{,}385$ kg
♀	$n_2 = 65$	$m_2 = 3{,}36$ kg	$s_2 = 0{,}363$ kg

H_0, H_1
Sous H_0 :

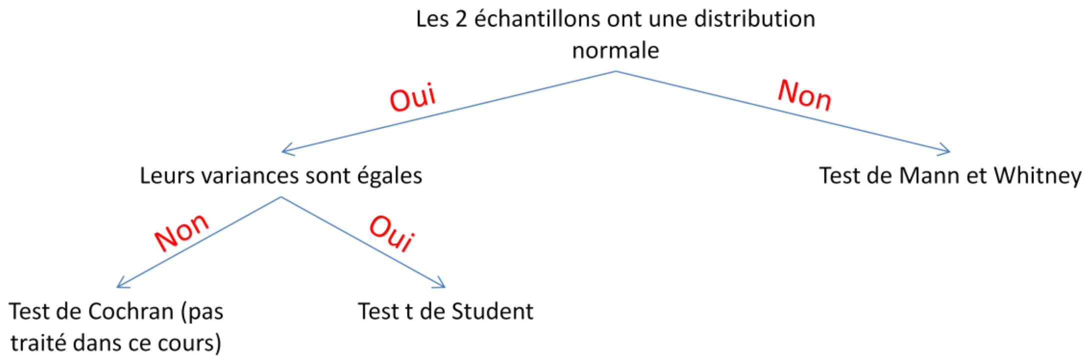
$$U_C = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{3{,}4 - 3{,}36}{\sqrt{\frac{0{,}385^2}{41} + \frac{0{,}363^2}{65}}} = 0{,}54$$

α = 5% → $U_t = 1{,}96$
$|U_C| < U_t$: au risque 5%, on ne montre pas de différence significative entre les poids des nnés ♂ et ♀

2 – Deux petits échantillons

2 échantillons :

E_1 : m_1, s_1^2, n_1
E_2 : m_2, s_2^2, n_2



a – Les deux échantillons ont une distribution normale

On compare alors les variances :
H_0 : égalité des variances
H_1 : inégalité des variances

Test de Snédécor ou test F :

Paramètre $F_C = \frac{S_1^{\ 2}}{S_2^{\ 2}}$ si $s_1 > s_2$ ou $F_C = \frac{S_2^{\ 2}}{S_1^{\ 2}}$ si $s_1 < s_2$

Sous H_0 et $\alpha = 5\% \rightarrow p = \alpha/2 = F_T$
Table de Snédécor :

v_1 = nombre ddl au numérateur : $(n_1 - 1)$ si $s_1 > s_2$ ou $(n_2 - 1)$ si $s_1 < s_2$
v_2 = nombre ddl au dénominateur : $(n_2 - 1)$ si $s_1 > s_2$ ou $(n_1 - 1)$ si $s_1 < s_2$

Comparaison :

- Si $F_C > F_T$: au risque α on rejette H_0 donc les variances ne sont pas égales $\rightarrow$ test de Cochran
- Si $F_C < F_T$: au risque α on accepte H_0 donc les variances sont égales $\rightarrow$ test T de Student

1 – Les variances ne sont pas égales

On s'arrête là, on ne sait pas faire après.

2 – Les variances sont égales, on compare les moyennes

Il faut trouver une variance commune (puisqu'elles sont égales) : s² commune.

$$s^2 = \frac{(n_1 - 1)\, s_1^{\ 2} + (n_2 - 1)\, s_2^{\ 2}}{n_1 + n_2 - 2}$$

H_0 : pas de ≠ significative entre les moyennes = les échantillons proviennent de la même population
H_1 : il existe une ≠ significative entre les moyennes = les 2 échantillons proviennent de populations ≠

Sous H_0 :

$$t_C = \frac{m_1 - m_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = \frac{m_1 - m_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$ suit une loi de Student à $(n_1 + n_2 - 2)$ ddl

On pose α et t_t à $(n_1 + n_2 - 2)$ ddl

Comparaison :

- Si $|t_C| > t_t$: au risque α on rejette H_0 et on accepte H_1 : il existe une différence significative entre les moyennes
- Si $|t_C| < t_t$: au risque α on accepte H_0 : on ne montre pas de différence significative entre les moyennes

b – Les 2 échantillons n'ont pas une distribution normale : test non paramétrique de Mann et Whitney (rangs : *distribution free*)

Exemple :

Comparaison des notes obtenues aux tests psychomoteurs pour des patients atteints par la maladie A ou par la maladie B.

Maladie A : 48, 60, 42, 58, 50, 31, 42 $n_A = 7$

Maladie B : 31, 41, 23, 28, 42 $n_B = 5$

H_0 : pas de différence significative

H_1 : il existe une différence significative

On classe les données par ordre croissant :

A =			31		42, 42	48	50	58	60
B =	23	28	31	41	42				
Rangs :	1	2	3,5	5	7	9	10	11	12

$T_A = 3,5 + 7x2 + 9 + 10 + 11 + 12 = 59,5$

$T_B = 1 + 2 + 3,5 + 5 + 7 = 18,5$

Pour vérifier si on n'a pas fait d'erreur, on calcule $T_A + T_B$ et ça doit être égal à $\frac{(n_A + n_B)\,(n_A + n_B + 1)}{2}$.

Paramètres :

$U_A = n_A \times n_B + \frac{n_A\,(n_A + 1)}{2} - T_A = 3,5$ et $U_B = n_A \times n_B + \frac{n_B\,(n_B + 1)}{2} - T_B = 31,5$

1 – Si n_A <u>ou</u> n_B < 10

$U_C = \min(U_A\,;\,U_B)$ (ici, U_A)

Table : $n = \min(n_A\,;\,n_B)$

$m = \max(n_A\,;\,n_B)$

Ici : n = 5, m = 7, m-n = 2, α = 5%

Comparaison :

- Si $|U_C| < U_t$: au risque α on rejette H_0 et on accepte H_1 : il existe une différence significative
- Si $|U_C| > U_t$: au risque α on accepte H_0 : on ne montre pas de différence significative

Ici, $U_C = 3,5$ $U_t = 5$ → $U_C < U_t$: au risque 5%, on rejette H_0 : il existe une différence significative

2 – Si n_A et $n_B \geq 10$

On montre que U_A et U_B suivent une loi normale $\mathcal{N}(\frac{(n_A.n_B)}{2} ; \frac{(n_A.n_B)(n_A+n_B+1)}{12})$.

$U_C = \frac{U_A - \frac{n_A.n_B}{2}}{\sqrt{\frac{n_A.n_B(n_A+n_B+1)}{12}}}$ ou $U_C = \frac{U_B - \frac{n_A.n_B}{2}}{\sqrt{\frac{n_A.n_B(n_A+n_B+1)}{12}}}$ selon que $U_A < U_B$ ou non.

On pose α, U_t

Comparaison :

- Si $|U_C| > U_t$: au risque α on rejette H_0 et on accepte H_1 : il existe une différence significative
- Si $|U_C| < U_t$: au risque α on accepte H_0 : on ne montre pas de différence significative

c – Exercice

N° ampoule	1	2	3	4
Ø induration	10	13	13	7
	11	13	13	11
	11	14	14	12
	13	18	19	13
	14	18	22	13
	15	21		14
	17			

					Totaux
n_i	8	7	6	7	**28**
Σx_{ij}	103	111	98	82	**394**
m_i	12,87	15,86	18,33	11,71	
$(\Sigma x_{ij})^2$	10 609	12 321	9 604	6 724	
$\frac{(\Sigma x_{ij})^2}{n_i}$	1 326,12	1 760,14	1 600,66	960,57	**5 647,49**
Σx_{ij}^2	1 365	1 819	1 668	992	**5 844**

$M = \frac{8\,.\,12{,}87 + 7\,.\,15{,}86 + \ldots}{28} = \frac{394}{28} = 14{,}07$

$\sum_{i=1}^{4} n_i\,(m_i - M)^2 \quad = \sum_{i=1}^{4} \frac{(\Sigma\, xij)^2}{n_i} - MN^2$

$= 5\,647{,}49 - \frac{(\Sigma\, xij)^2}{N}$

$= 5\,647{,}49 - \frac{394^2}{28}$

$= 103{,}39$

$\sum_i \sum_j (x_{ij} - m_i)^2 \quad = \sum_i [\sum_j x_{ij}^2 - n_i m_i^2]$

$= 5\,844 - 5\,647{,}49$

$= 196{,}51$

1^e estimation $= \frac{196{,}51}{24} = 8{,}19$

2^e estimation $= \frac{103{,}35}{4-1} = 34{,}45$

⇨ $F_C = \frac{34,45}{8,19} = 4,2$

$\alpha = 2\% \rightarrow p = 1\%$
$F_T = 4,72$
$v_1 = 3$ddl
$v_2 = 24$ ddl
$F_C < F_T$: au risque 2%, on ne montre pas de différence significative entre les vaccins.

III – Séries appariées

1 – Test paramétrique t

di = différence entre valeur mesurée avant et après un traitement
TA = tension artérielle

Patients :	1	2	3	4	5	...	N
TA avant :	18	16	15	...			
TA après :	16	17	14	...			
Différence :	2	-1	1	...			

Pas indépendantes car réalisées sur mêmes patients

Différence moyenne : $\bar{d} = \frac{\Sigma\, di}{N}$

$s_d^2 = \frac{\Sigma\, di^2}{N-1} - \frac{(\Sigma\, di)^2}{N(N-1)}$

H_0 : pas de différence significative entre TA avant et TA après
H_1 : il existe une différence significative

Sous H_0 :

$t_c = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$ suit une loi de Student à (N-1) ddl :

- si $N \geq 30$
- si $N < 30$ mais d_i suivent une loi $\mathcal{N}$ormale
- sinon : Wilcoxon

α

Comparaison :

- Si $|t_C| > t_t$: au risque α on rejette H_0 et on accepte H_1 : il existe une différence significative
- Si $|t_C| < t_t$: au risque α on accepte H_0 : on ne montre pas de différence significative

2 – Test de Wilcoxon

H_0, H_1
Calculer les d_i
$|d_i|$: les classer par ordre croissant en éliminant les $d_i = 0$, et leur donner des rangs

R^+ : somme des rangs des $d_i > 0$
R^- : somme des rangs des $d_i < 0$
N' : nombre des d_i non nuls

Si N' ≤ 25 ⟶ table de Wilcoxon
R table
α

Comparaison :
- Si R_{table} > min (R^+ ; R^-) : au risque α on rejette H_0 et on accepte H_1 : il existe une différence significative
- Si R_{table} < min (R^+ ; R^-) : au risque α on accepte H_0 : on ne montre pas de différence significative

Si N' > 25 :
On ne sait pas faire.

3 – Exercice

Patient	Antalgique 1	Antalgique 2	di	di²
1	260	250	+10	100
2	248	240	+8	64
3	270	240	+30	900
4	320	310	+10	100
5	270	240	+30	900
6	310	210	+100	10 000
7	332	260	+72	5184
8	240	290	-50	2500
9	250	260	-10	100
10	380	220	+160	25600
11	290	330	-40	1600
12	220	210	+10	100
13	280	300	-20	400
14	400	380	+20	400
15	370	300	+70	4900
16	284	230	+54	296
17	340	220	+120	14400
Totaux :			574	70 164

Sous H_0,

$$t_c = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$$

$$\bar{d} = \frac{\Sigma\, di}{N} = \frac{574}{17} = 33{,}76$$

$$s_d^2 = \frac{70\,164}{16} - \frac{574^2}{16\,.\,17} = 3\,173{,}94$$

$s_d = 56{,}34$

$t_c = 2{,}47$

α = 5% ⟶ $t_t = 2{,}12$

$t_c > t_t$ donc au risque 5% il existe une ≠ significative d'efficacité entre les 2 antalgiques

α = 1% ⟶ $t_t = 2{,}92$

$t_c < t_t$ donc au risque 1% on ne montre pas de ≠ significative d'efficacité entre les 2 antalgiques

Avec Wilcoxon :

Rangs - = 31
Rangs + = 122

α = 5% ⟶ R table = 35
$R_C = \min (R^- ; R^+) = 31$
R table > R_C : au risque 5%, on conclue à une différence significative

α = 1% ⟶ R table = 23
$R_C = \min (R^- ; R^+) = 31$
R table < R_C : au risque 5%, on ne montre pas de différence significative