Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Echahid Hamma Lakhdar d'El-Oued
Faculté des Sciences de la Nature et de la Vie
Département de Biologie

Cours de Biostatistique

"Test t & z "

Pour Master 2
Biologie et physiologie végétale

Préparé par
Dr. ALIA Zeid

Année universitaire 2022 - 2023

1. Nature des variables

• Comparer 2 moyennes : tester l'association entre

• 1 variable quantitative continue

• 1 variable qualitative binaire

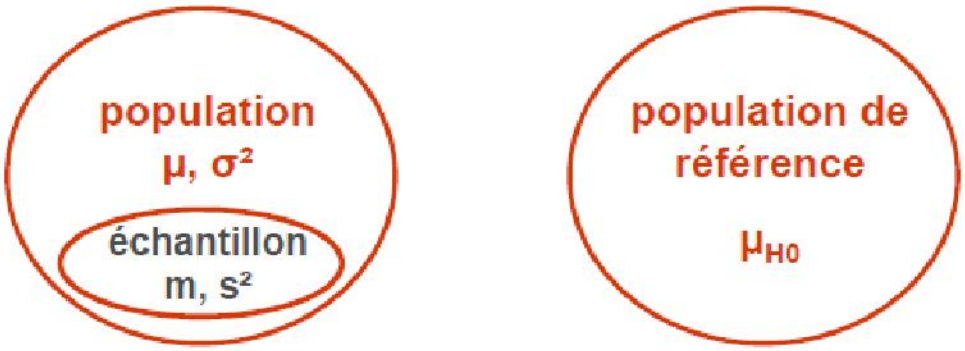
• Exemple : μL1 santé (âge) ≠ μL1 sciences (âge) ?

• âge : variable quantitative continue

• L1 sante versus L1 sciences : variable qualitative binaire (dichotomique)

2. Comparaison d'une moyenne observée à une moyenne théorique échantillon

Comparer une moyenne observée (m) sur un échantillon issu d'une population de moyenne inconnue (μ) à une valeur théorique ou moyenne théorique connue (μH0) d'une population de référence



1. Formulation des hypothèses

H0 : μ = μH0

H1 : μ ≠ μH0

2. Risque α = 0.05 (5%) – a priori

3. Choix du test

Test Z de l'écart réduit (n ≥ 30)

Test t de Student (hypothèse de normalité)

Test z de l'écart réduit

m ≈ μ (fluctuations d'échantillonnage)

Sous H0 : μ = μH0 → μ - μH0 = 0

• Si $n \geq 30$: $m \rightarrow N(\mu, \sigma/\sqrt{n})$

• Rappel : m est une réalisation de la V.A. « moyenne empirique d'un échantillon de taille n » de moyenne μ et d'écart type $\sigma/\sqrt{n}$

Sous H0 : $\mu = \mu H0 \rightarrow \mu - \mu H0 = 0$

• Si $n \geq 30$: $m \rightarrow N(\mu, \sigma/\sqrt{n})$

$m - \mu H0 \rightarrow N(0, \sigma/\sqrt{n})$ (on a « centré » m en lui retranchant sa moyenne μ. Or μ étant inconnue, on lui substitue μH0 qui est connue et dont on sait sous H0 que $\mu = \mu H0$)

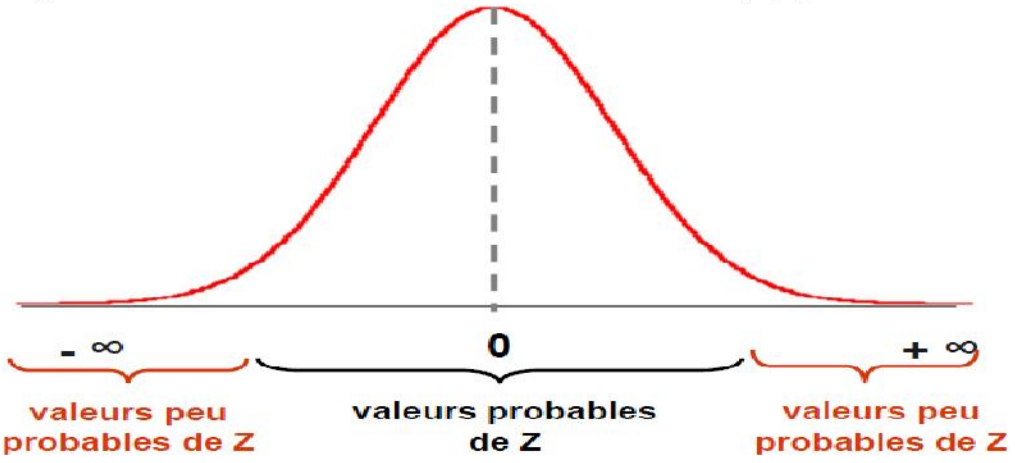
$$Z = \frac{m - \mu_{H0}}{\sqrt{\frac{\sigma^2}{n}}} \rightarrow N(0, 1)$$

(On a « réduit » $(m - \mu H0)$ en la divisant par son écart-type $\sigma/\sqrt{n}$)

La variance dans la population σ^2 étant le plus souvent inconnue, on lui substitue son estimateur s^2 (s^2 = estimation de σ^2 à partir de l'échantillon)

$$Z = \frac{m - \mu_{H0}}{\sqrt{\frac{s^2}{n}}} \rightarrow N(0, 1)$$

Densité de probabilité de loi normale centrée réduite N(0,1)

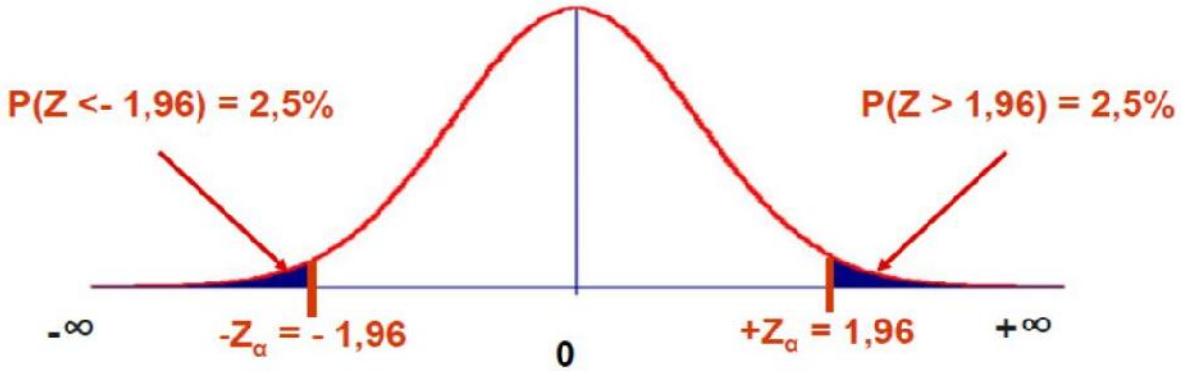


Abscisse : valeurs possibles de Z sous H0

$$Z = \frac{m - \mu_{H0}}{\sqrt{\frac{s^2}{n}}}$$

Densité de probabilité de loi normale centrée réduite (0,1)

α = 5% (0,05)



Abscisse : valeurs possibles de Z sous H_0

$$Z = \frac{m - \mu_{H0}}{\sqrt{\frac{s^2}{n}}}$$

Z_α = valeur de Z pour le risque α

Z_o = valeur observée de Z pour l'échantillon

$Z = \frac{m - \mu_{H0}}{\sqrt{\frac{s^2}{n}}}$

- Z est la variable aléatoire
- Zα est une valeur particulière de la variable aléatoire Z telle que P(Z > Zα) = α
  - (Zα est la valeur de Z pour le risque α)
  - (en santé et biologie, α = 0.05)
- Zo est une réalisation de la variable aléatoire Z
  - (Zo est la valeur observée/calculée de Z sur l'échantillon dont on dispose)

Détermination de la valeur de Zα correspondant à un risque α = 0.05 (5%)

Table de l'écart réduit

La table donne la probabilité α pour que l'écart-réduit dépasse en valeur absolue une valeur donnée ε, c'est-à dire la probabilité extérieure à l'intervalle [-ε,ε]. La probabilité α s'obtient par addition des nombres inscrits en marge

α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	∞	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695
0,1	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311
0,2	1,282	1,254	1,227	1,200	1,175	1,150	1,126	1,103	1,080	1,058
0,3	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,860
0,4	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,690
0,5	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,6	0,524	0,510	0,496	0,482	0,468	0,454	0,440	0,426	0,412	0,399
0,7	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,8	0,253	0,240	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,9	0,126	0,113	0,100	0,088	0,075	0,063	0,050	0,038	0,025	0,013

Exemple : Zo = 1.37 → P-value = 0.17

Test t de Student

• Sous H0 : μ = μH0

• Si la distribution de la variable est normale dans la population (et quel que soit l'effectif de l'échantillon n) :

$$T = \frac{m - \mu_{H0}}{\sqrt{\frac{s^2}{n}}} \rightarrow t_{(n-1)\ ddl}$$

Test t de Student : notion de ddl

Sous H0 :

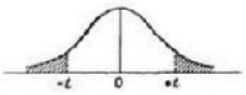
$$t = \frac{m - \mu_{H0}}{\sqrt{\frac{s^2}{n}}} \rightarrow t_{(n-1)\ ddl}$$

La fonction de densité de probabilité de t varie avec l'effectif de l'échantillon (en fait, avec l'effectif de l'échantillon – 1 = n – 1)

→ Il existe autant de lois t de Student qu'il existe d'échantillons d'effectif différent

Table de t (*).

La table donne la probabilité α pour que t égale ou dépasse, en valeur absolue, une valeur donnée, en fonction du nombre de degrés de liberté (d.d.l.).



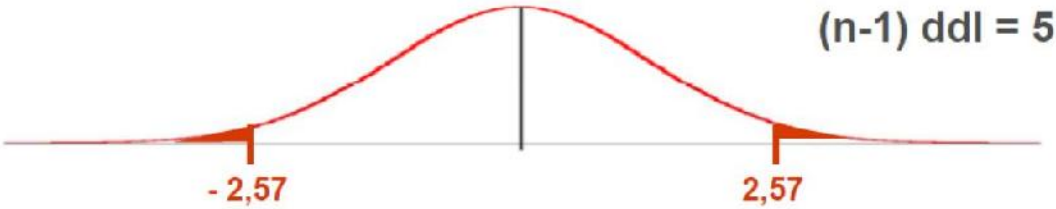
d.d.l. \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646
∞	0,126	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Valeur de t_α pour :

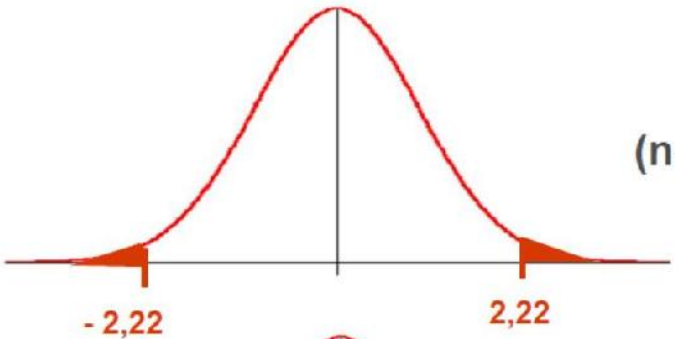
(n-1) ddl = 5 : 2.57

(n-1) ddl = 10 : 2.22

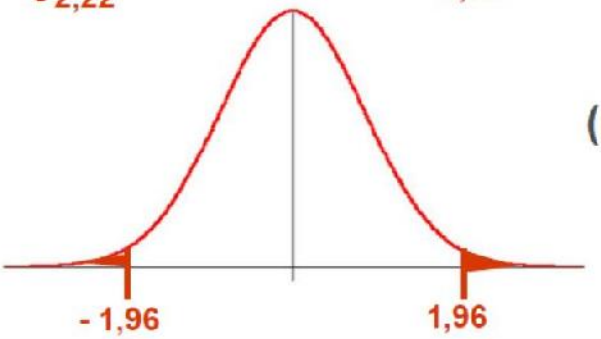
(n-1) ddl > 30 : 1.96



(n-1) ddl = 5



(n-1) ddl = 10



(n-1) ddl > 30

(cf annexe 1)

Condition de validité du test t de Student : la variable continue suit une loi normale

• A l'examen :

• Soit c'est indiqué dans l'énoncé

– (ex : « on suppose les conditions de validité vérifiées »)

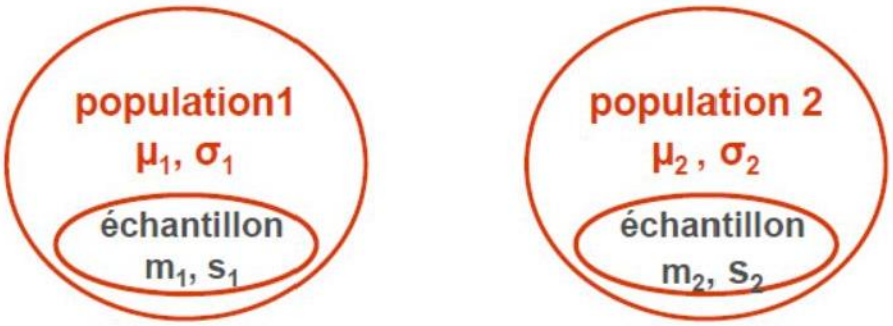
• Soit on vous pose la question

– (ex : « Quelles sont les conditions de validité de ce test ? »)

• Soit on vous demande de vérifier empiriquement qu'on ne s'écarte pas de cette hypothèse (visuellement le plus souvent)

– (ex : on fournit un histogramme dans l'énoncé)

3- Comparaison de 2 moyennes observées sur 2 échantillons indépendants



« Indépendant » signifie que l'échantillon 1 est constitué de manière indépendante de l'échantillon 2 (par opposition aux échantillons appariés) :

• Les sujets de l'échantillon 1 ne sont pas les mêmes que ceux de l'échantillon 2

• Les 2 échantillons peuvent être d'effectifs différents.

1. Formulation des hypothèses

H0 : $\mu 1 = \mu 2$

H1 : $\mu 1 \neq \mu 2$

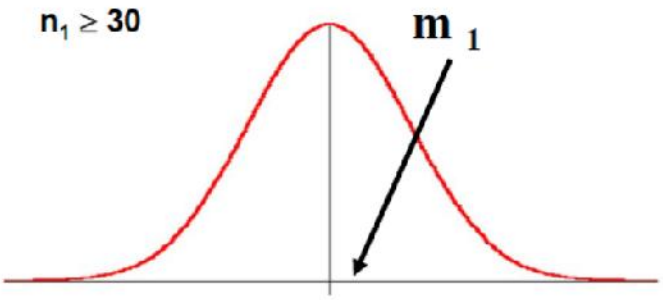
2. Risque $\alpha = 0.05$ (5%) – a priori

3. Choix du test

Test Z de l'écart réduit ($n1 \geq 30$ et $n2 \geq 30$)

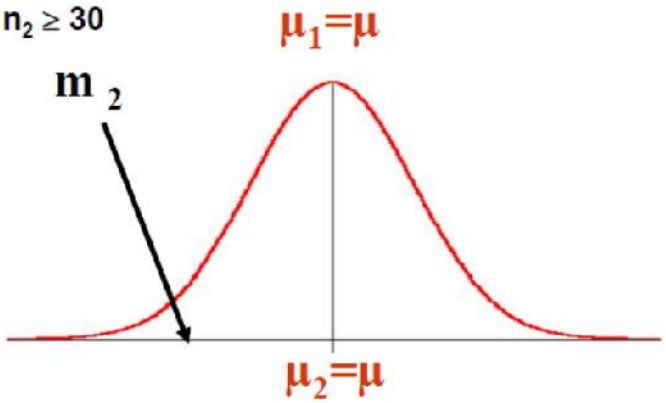
Test t de Student (hypothèse de normalité, variances comparables)

Sous H0 : $\mu_1 = \mu_2 = \mu$



$m_1 \approx \mu_1$

(fluctuations d'échantillonnage)



$m_2 \approx \mu_2$

(fluctuations d'échantillonnage)

$\mu 1 - \mu 2 = 0$ et $m1 - m2 \approx 0$

(fluctuations d'échantillonnage)

Sous H0 : $\mu 1 - \mu 2 = 0$

Rappel : m1 est une réalisation de la V.A. « moyenne empirique d'un échantillon de taille n1 » de moyenne $\mu 1$ et d'écart type $\sigma 1/\sqrt{n1}$: $m1 \rightarrow N(\mu 1, \sigma 1/\sqrt{n1})$

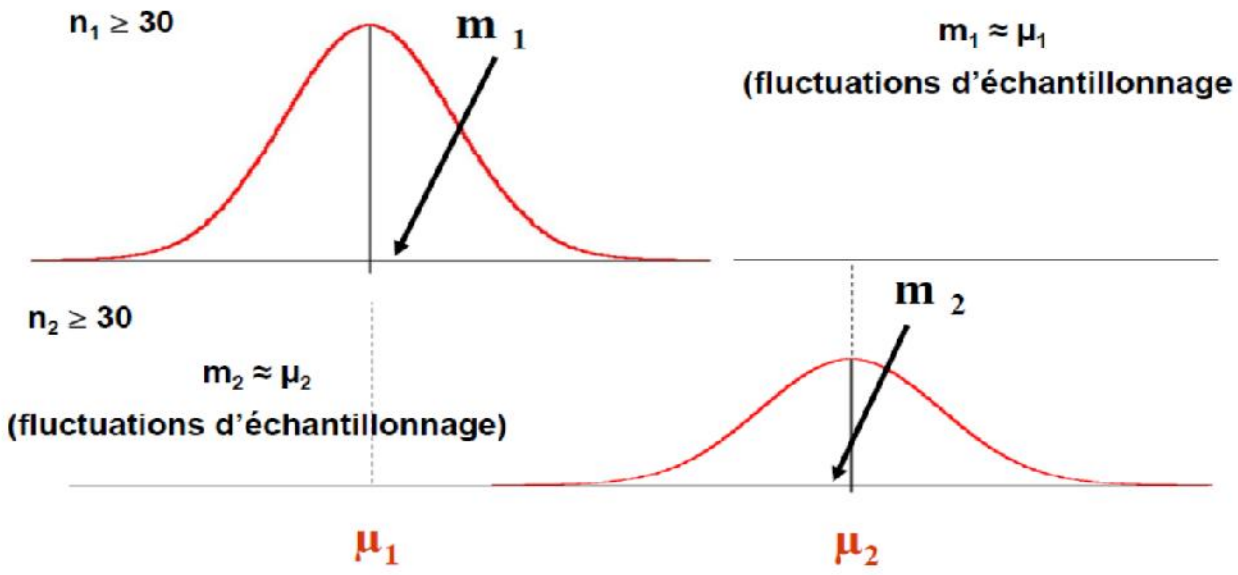
$m2 \rightarrow N(\mu 2, \sigma 2/\sqrt{n2})$

$$(m_1 - m_2) \rightarrow N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \qquad (m_1 - m_2) \rightarrow N\left(0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$\text{var}(m_1 - m_2) = \text{var}(m_1) + \text{var}(m_2) - 2\,\text{cov}(m_1, m_2)$$

$$\text{var}(m_1 - m_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2 \times 0$$

Sous H1 : $\mu_1 \neq \mu_2$



m1 – m2 ≈ μ1 - μ2
(fluctuations d'échantillonnage)

Test Z de l'écart réduit

• Si n1 ≥ 30 et n2 ≥ 30

• Sous H0 : μ1 = μ2 → μ1 - μ2 = 0

$$Z = \frac{m_1 - m_2}{\sqrt{var(m_1 - m_2)}} = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightarrow N(0, 1)$$

s² est un estimateur de σ² → $$Z = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Test t de Student

• Conditions d'application :

• - La distribution de la variable continue est normale dans les 2 populations

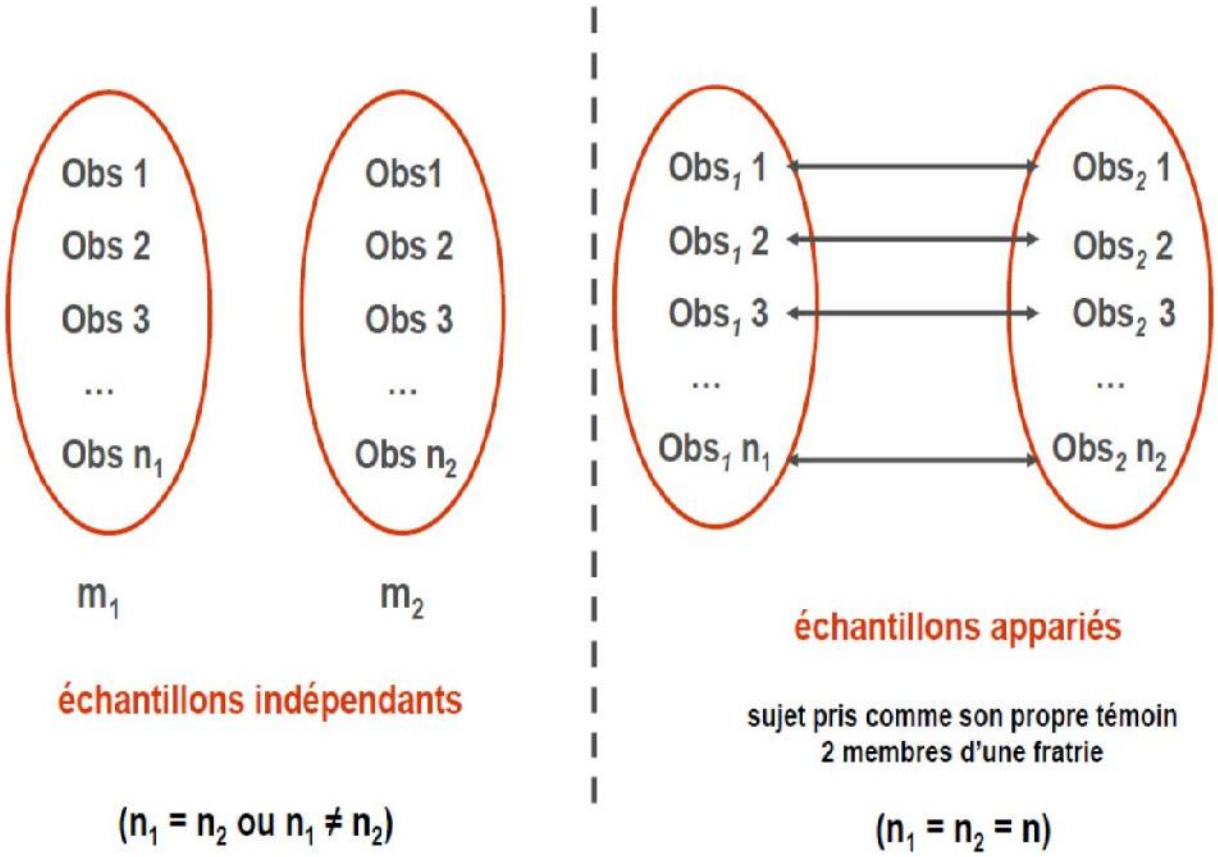
• - Les variances σ1² et σ2² sont comparables (rapport 1 à 3)

• Sous H0 : μ1 = μ2 → μ1 - μ2 = 0

$$T = \frac{m_1 - m_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \rightarrow t_{(n_1 + n_2 - 2)\ ddl}$$

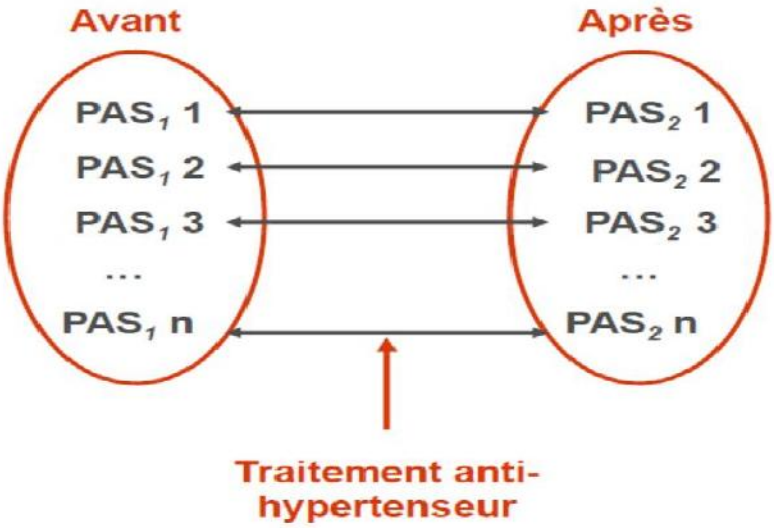
$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

4- Comparaison de deux moyennes observées (échantillons appariés)



Les mesures PAS1 et PAS2 du sujet 1 ne sont pas indépendantes Les 2 mesures ont été effectuées sur le même sujet : si PAS1 était très élevée, il est probable que PAS2 restera élevée (mais moins que PAS1 si le traitement est efficace)

Le test doit prendre en compte cette dépendance des observations PAS1 et PAS2 (En revanche, les mesures PAS2 du sujet 1 et PAS2 du sujet 2 sont indépendantes)



$$H0 : m_{PASavant} = m_{PASaprès}$$

Z pour échantillons indépendants

m1 – m2

var (m1– m2) = var (m1) + var (m2) – 2 cov(m1, m2) = var (m1) + var (m2)

$$Z = \frac{m_1 - m_2}{\sqrt{var(m_1) + var(m_2)}}$$

• Z pour échantillons appariés

m1 – m2

var (m1– m2) = var (m1) + var (m2) – 2 cov (m1, m2)

$$Z = \frac{m_1 - m_2}{\sqrt{var(m_1) + var(m_2) - 2\,cov(m_1, m_2)}}$$

z apparié > z indépendant → gain de puissance

Echantillons appariés

$$Z = \frac{m_1 - m_2}{\sqrt{var(m_1) + var(m_2) - 2\,cov(m_1, m_2)}}$$

$\frac{S_1^2}{n_1}$ $\frac{S_2^2}{n_2}$

ne peut pas être estimée directement car on ne dispose que d'une mesure de m_1 et une mesure de m_2

Echantillons appariés

$$Z = \frac{m_1 - m_2}{\sqrt{\text{var}(m_1) + \text{var}(m_2) - 2\,\text{cov}(m_1, m_2)}} = \frac{m_d}{\sqrt{\text{var}(m_d)}}$$

- $(m_1 - m_2) = m_d$, avec $m_d = \dfrac{\sum_{i=1}^{n} d_i}{n}$

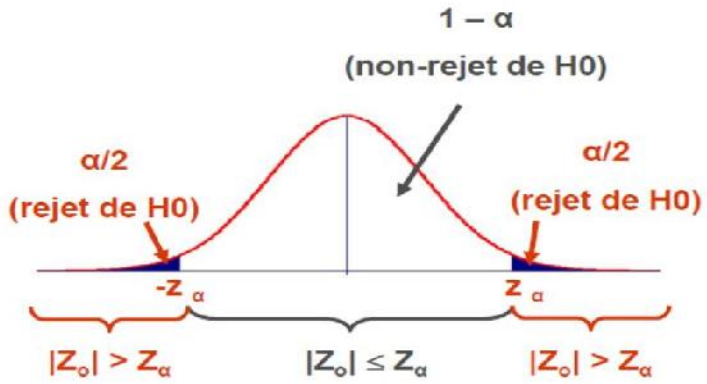
- $\text{var}(m_d) = s_d^2 / n$, avec $s_d^2 = \dfrac{\sum_{i=1}^{n} (d_i - m_d)^2}{(n-1)}$

md est une réalisation de la V.A. « moyenne empirique des différences d'un échantillon de taille n » de moyenne μd et d'écart type σd/√n

Test Z de l'écart réduit pour échantillons appariés

- H0 : μd = 0 (μ1 = μ2)
- H1 : μd ≠ 0 (μ1 ≠ μ2)
- Si n ≥ 30 paires

$$Z_o = \frac{m_d}{\sqrt{\frac{s_d^2}{n}}}$$



Test t de Student pour échantillons appariés

- H0 : μd = 0 (μ1 = μ2)
- H1 : μd ≠ 0 (μ1 ≠ μ2)
- Si la distribution des différences individuelles est normale

$$t_o = \frac{m_d}{\sqrt{\frac{s_d^2}{n}}}$$

Resumé

m1	m2	effectif	test	conditions
observée	théorique	$n \geq 30$	Z	-
		n	$t_{(n-1)\ ddl}$	normalité
observée	observée	$n_1, n_2 \geq 30$	Z	-
(indépendantes)		n_1, n_2	$t_{(n1+n2\ -2)\ ddl}$	normalité σ^2 comparables
observée	observée	$n \geq 30$ paires	Z	-
(appariées)		n paires	$t_{(n-1)\ ddl}$	normalité d_i