DULCLAUDIAX

STATISTIQUES DÉDUCTIVES







Sommaire

TEST D'HYPOTHÈSE

Statistiques déductives → tirer des conclusions à partir des observations

Hypothèse nulle (H0)	Hypothèse alternative (H1)
Pas de différence entre les 2 groupes	Différence signifcative entre les 2 groupes
Les fluctuations observées sont dues au hasard	Les fluctuations observées ne sont pas dues au hasard

ETAPES D'UN TEST

Théorie

- Définir H0 et H1
- Choisir le test (en fonction des données)
- Fixer le risque α (souvent 5%)

Pratique

- Recueillir les données
- Calculer le paramètre Z
- Comparer Z_c à Z_t

Conclusion

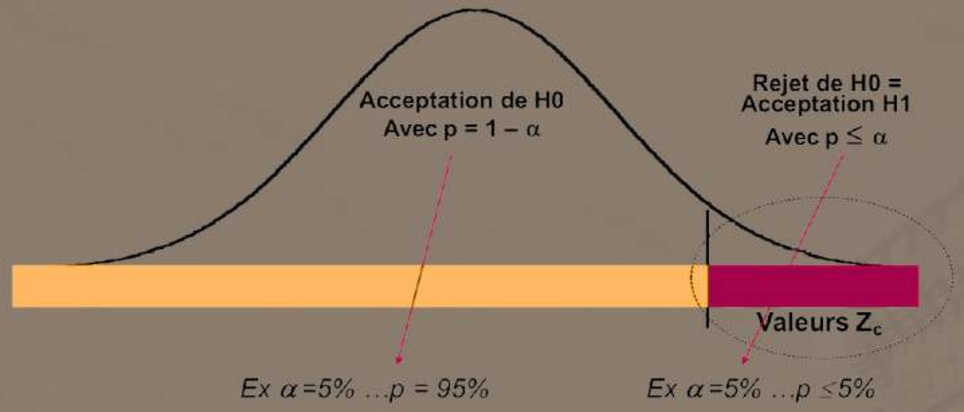
- Accepter ou rejeter H0
- Fixer le risque d'erreur réel β
- Interpréter les résultats : statistiquement et cliniquement

NOTION DE RISQUE

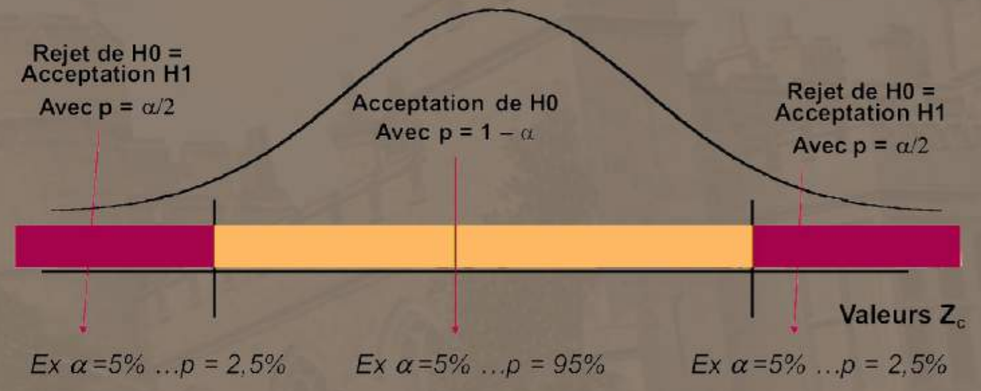
	Rejet H0	Non rejet H0
H0 vraie	α	1-α
H1 vraie	1-β	β

Risque de première espèce Risque α	Risque de seconde espèce Risque β
Probabilité de rejeter H0 si H0 est vraie	Probabilité d'accepter H0 si H0 est fausse
Ce risque est maîtrisé	Ce risque est négligé Il peut être très élevé (en général beta=20%)
Fixé à l'avance	Fixé à postériori
La puissance du test vaut 1-β : probabilité de rejeter H0 avec H1 vraie	

INTERPRÉTATION GRAPHIQUE



Situation unilatérale



Situation bilatérale

Lien entre 2 variables qualitatives

TEST DE COMPARAISON DE POURCENTAGES

Tout effectif

Paramètre Z → **écart-réduit ε**

ε_t vient de la table de l'écart-réduit

$$\varepsilon_c = \frac{p_A - p_B}{\sqrt{\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}}}$$

Si $\varepsilon_c > \varepsilon_t$ → rejet de H0

Méthodologie

Comment trouver le paramètre théorique ?

Table de l'écart réduit

α

		0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	∞	2,576	2,326	2,17	2,054	1,96	1,881	1,812	1,751	1,695
0,1	1,645	1,598	1,555	1,514	1,476	1,44	1,405	1,372	1,341	1,311
0,2	1,282	1,254	1,227	1,2	1,175	1,15	1,126	1,103	1,08	1,058
0,3	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,86
0,4	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,69
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	0,51	0,496	0,482	0,468	0,454	0,44	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	0,24	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,9	0,126	0,113	0,1	0,088	0,075	0,063	0,05	0,038	0,025	0,013

Table pour les petites valeurs de la probabilité

0,001	0,000 1	0,000 01	0,000 001	0,000 000 1	0,000 000 01	0,000 000 001
3,2905	3,8905	4,41717	4,89164	5,32672	5,73073	6,10941

Exemple

Soient 2 groupes de 200 enfants :

→ Crèche : 200 enfants, 130 atteints de rhinopharyngite

→ Maison : 200 enfants, 96 atteints de rhinopharyngite

Le mode de garde influe-t-il sur le risque de rhinopharyngite ?

Exemple

	Crèche	Domicile
Sain	70	104
Malade	130	96

→ H0 : pas de différence entre les 2 modes de garde vis-à-vis du développement de rhinos

→ H1 : il y a une différence entre les 2 modes de garde

Variables qualitatives → **test de comparaison de pourcentages**

- Zc = 3,4
- Zt = 1,96

Zc > Zt donc on REJETTE H0 au seuil 5%

Ccl : le risque de rhinopharyngite est supérieur chez les enfants gardés en crèche que chez les enfants gardés à domicile.

Lien entre 2 variables qualitatives

Tout effectif

TEST DU KHI^2

Test préféré si tableau de données a + de 2 lignes (ou colonnes)

Paramètre Z → χ^2

χ_t vient de la table du χ^2

DDL = (nb de lignes - 1) * (nb de colonnes - 1)

$$\chi^2 = \sum \frac{(o_i - c_i)^2}{c_i}$$

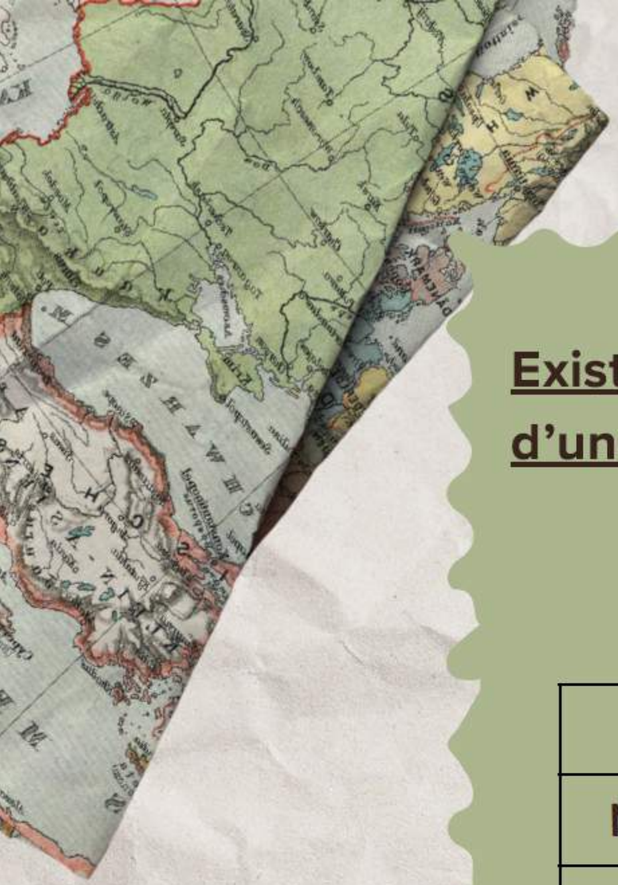
Si $\chi^2_c > \chi^2_t$ → rejet de H0

Méthodologie

Comment trouver le paramètre théorique avec un DDL?

ddl \ α	0,9	0,5	0,3	0,2	0,1	0,05	0,02	0,01	0,001
1	0,016	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5,991	7,824	9,21	13,815
3	0,584	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	1,61	4,351	6,064	7,289	9,236	11,07	13,388	15,086	20,515
6	2,204	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,49	7,344	9,524	11,03	13,362	15,507	18,168	20,09	26,125
9	4,168	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	4,865	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	5,578	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	6,304	11,34	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	7,042	12,34	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	7,79	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	8,547	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	9,312	15,338	18,418	20,465	23,542	26,296	29,633	32	39,252
17	10,085	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,79
...									

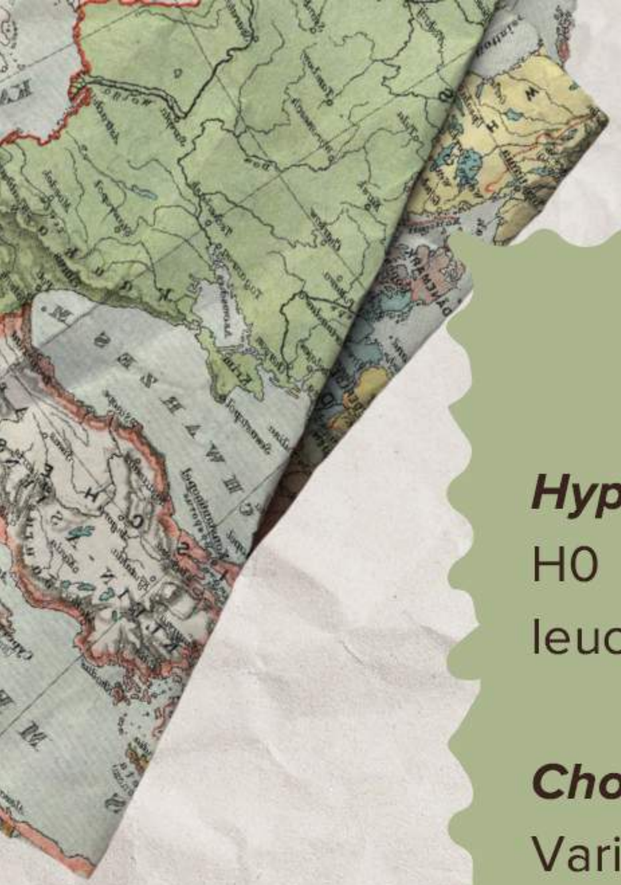
Exemple



Existe-t-il un lien entre l'exposition au benzène et la survenue d'une leucémie ?

	Leucémie	Non leucémie	Total
Expo	15	485	500
Non expo	20	980	1000
Total	35	1465	1500

Exemple



Hypothèses :

H0 : il n'existe pas de lien entre l'exposition au benzène et les leucémies

Choix du test :

Variable 1 : leucémie ou non → qualitatif

Variable 2 : exposé au benzène ou non → qualitatif

→ **Test du χ_2**

Calcul du χ_2

Valeurs observées :

	Leucémie	Non leucémie	Total
Expo	15	485	500
Non expo	20	980	1000
Total	35	1465	1500

Valeurs calculées:

	Leucémie	Non leucémie	Total (environ)
Expo	11,65	488,3	500
Non expo	23,35	976,7	1000
Total	35	1465	1500

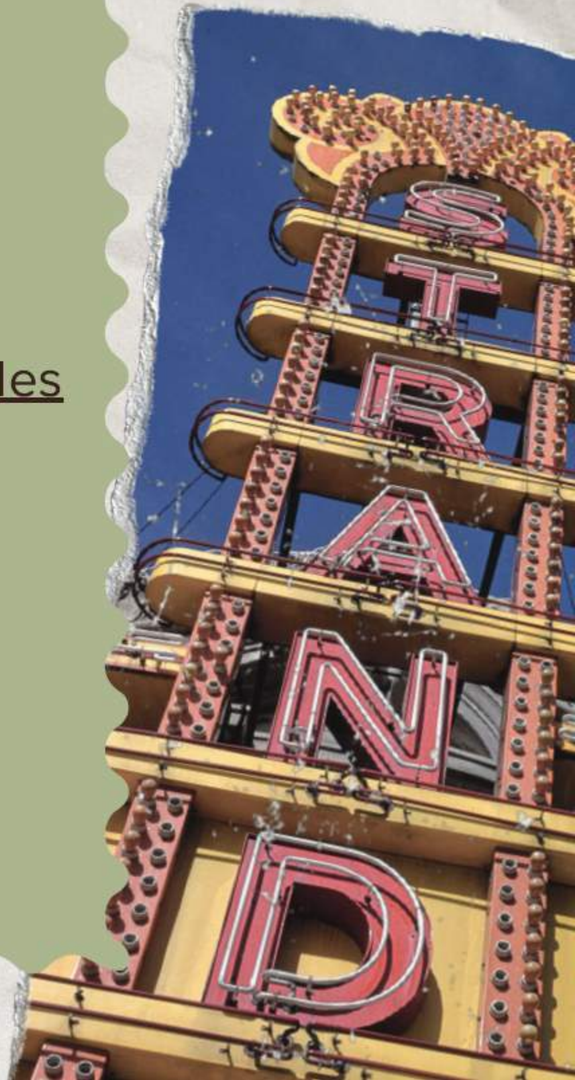
$$\chi^2 = \frac{(15-11,65)^2}{11,65} + \frac{(20-23,35)^2}{23,35} + \frac{(485-488,3)^2}{488,3} + \frac{(980-976,7)^2}{976,7} = 1,42$$

DDL = (nombre de lignes - 1) * (nombre de colonnes - 1) = (2-1) * (2-1) = 1

Conclusion

$\chi^2_c < \chi^2_t$ donc on accepte H0 au seuil 0.05

Ccl : Il n'existe pas de relation entre l'exposition au benzène et les leucémies



Lien entre variables qualitatives et quantitatives

TEST DE COMPARAISON DE MOYENNES

Grands échantillons (n_1 et $n_2 > 30$)

Paramètre Z → **écart-réduit ε**

ε_t vient de la table de l'écart-réduit

$$\varepsilon_c = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Si $\varepsilon_c > \varepsilon_t$ → rejet de H0

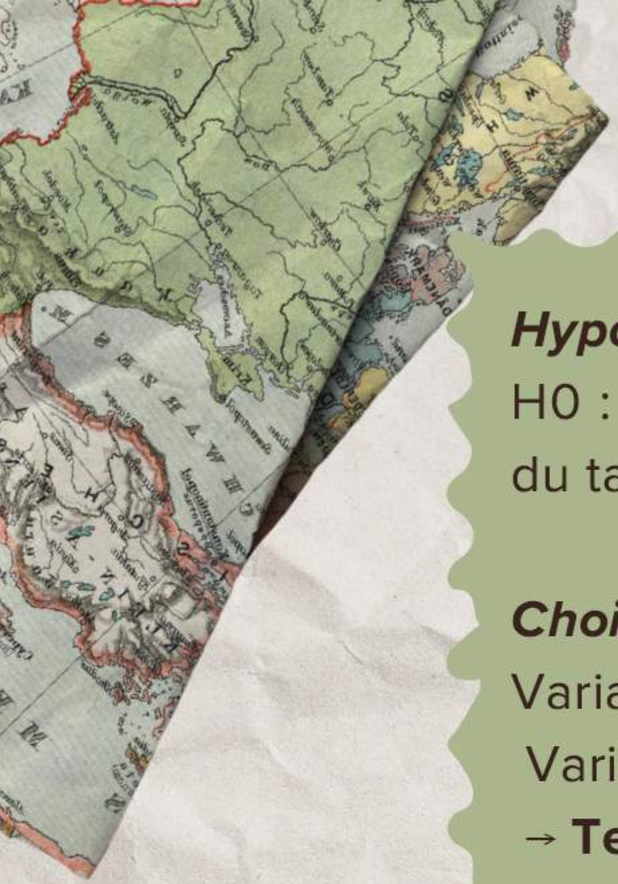
Exemple

On cherche à **comparer les taux de T3 libre chez les femmes prenant un contraceptif oral (c.o.) et chez celles qui n'en prennent pas.** Après un tirage au sort, on obtient :

Femmes sans c.o. : $n_1 = 50$; $m_1 = 2$ nmol ; $s_1 = 0,35$ nmol

Femmes avec c.o. : $n_2 = 33$; $m_2 = 2,5$ nmol ; $s_2 = 0,3$ nmol

Exemple



Hypothèses :

H0 : les moyennes ne sont pas différentes, ce sont 2 estimateurs du taux de T3 libre chez la femme en général

Choix du test :

Variable 1 : prise ou non de la pilule → qualitatif

Variable 2 : dosage de T3 → quantitatif

→ **Test de comparaison de moyennes**

$\varepsilon_t = 1{,}96$

$\varepsilon_c = 6.94$

$\varepsilon_c > \varepsilon_t$ donc on rejette H0 avec $p < 0.0001$

Lien entre variables qualitatives et quantitatives

Petits échantillons (n_1 et $n_2 < 30$)

TEST T DE STUDENT

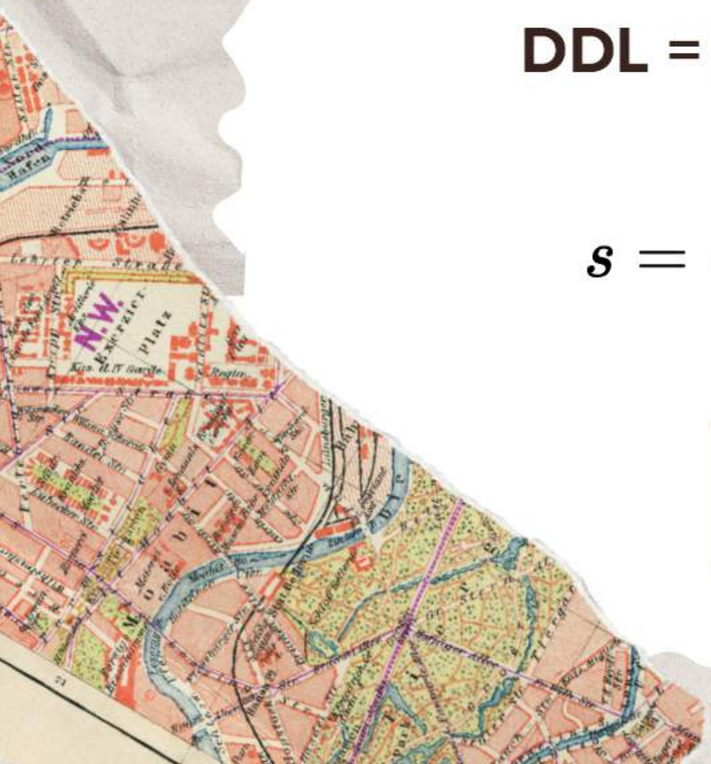
Paramètre Z → **t**

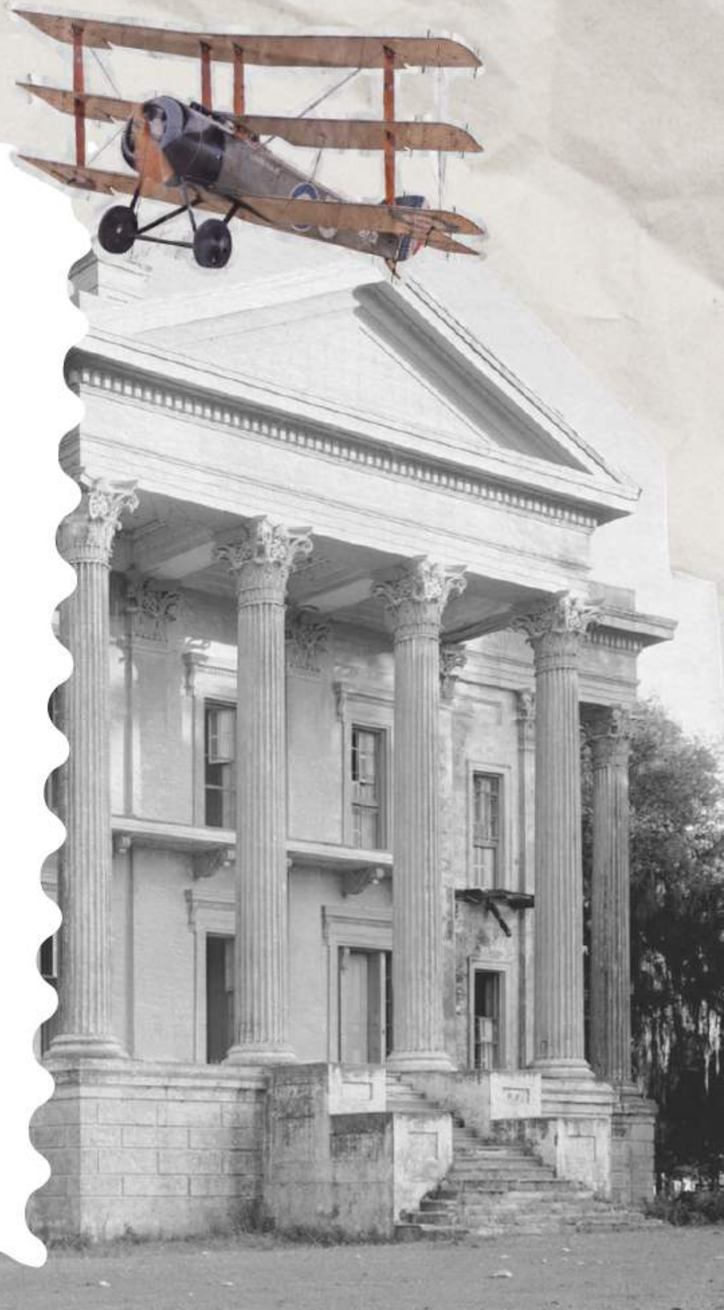
t_t vient de la table du t de Student

DDL = (n_1 - 1) + (n_2 - 1)

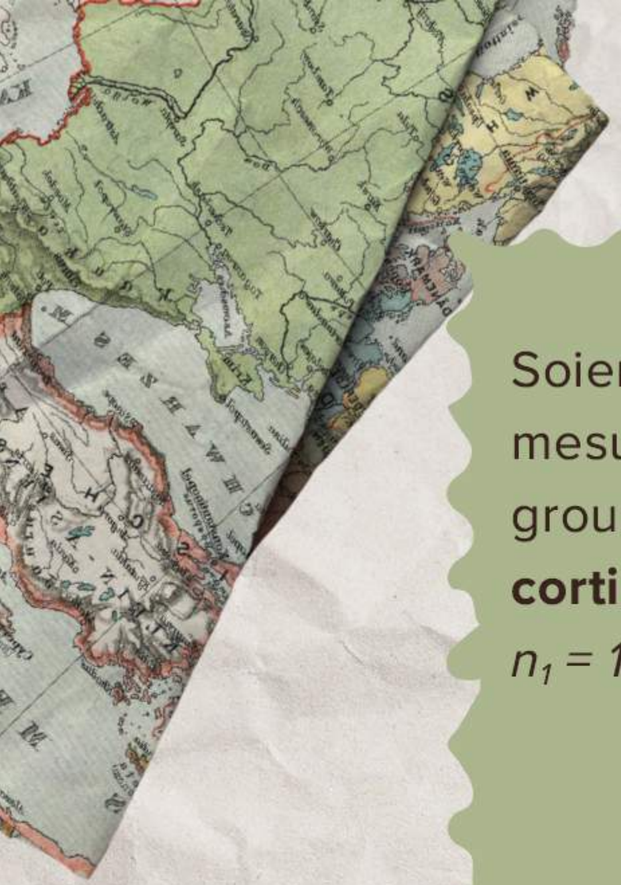
$$s = \sqrt{\frac{\sum (x_i - m_1)^2 + \sum (x_j - m_2)^2}{(n_1 - 1) + (n_2 - 1)}}$$

Si $t_c > t_t$ → rejet de H0





Exemple



Soient 15 femmes obèses et 12 femmes de poids normal. On mesure le taux de corticoïdes sanguins moyens dans chaque groupe. **L'obésité a-t-elle une influence sur le taux de corticoïdes ?**

$n_1 = 15$; $m_{1=} 6.3$; $s_1 = 1.8$ / $n_2 = 12$; $m_2 = 4.5$; $s_2 = 1.6$

Exemple

Hypothèses :

H0 : m_1 et m_2 ne sont pas différents dans les 2 groupes

Choix du test :

Variable 1 : obèse ou non → qualitatif

Variable 2 : taux de corticoïdes → quantitatif

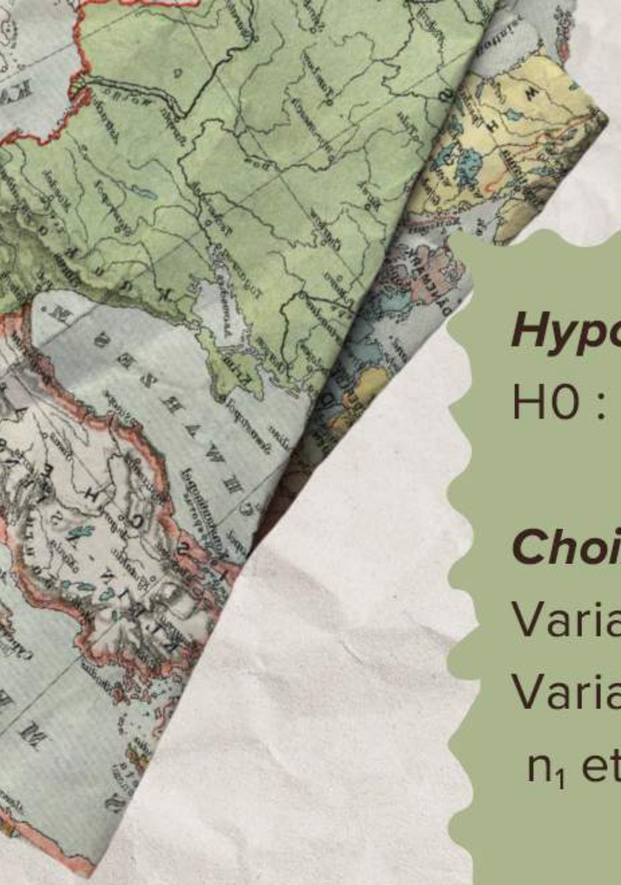
n_1 et $n_2 < 30$ → **test t de Student**

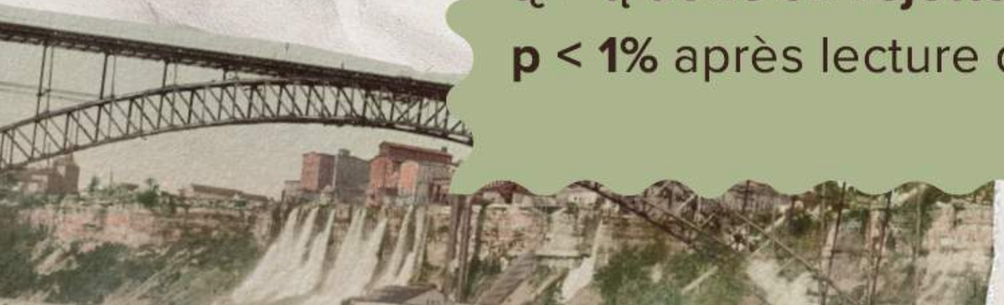
tc = 2,92

DDL = 15 + 12 - 2 = **25** donc d'après la table **tt = 2,06**

$t_c > t_t$ donc on rejette H0 au seuil 5%

p < 1% après lecture dans la table. **On rejette H0 à 1% à postériori**



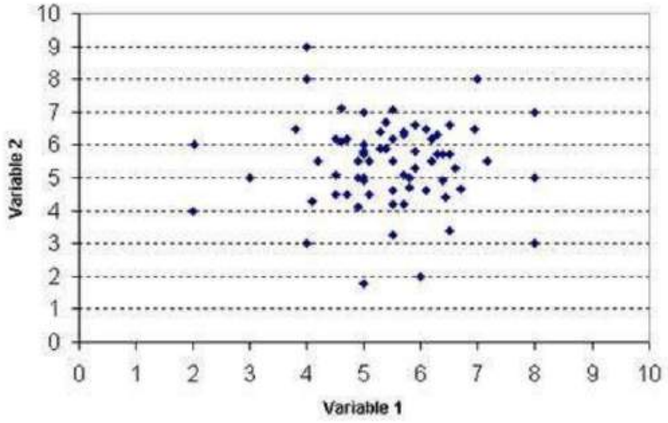


Lien entre 2 variables quantitatives

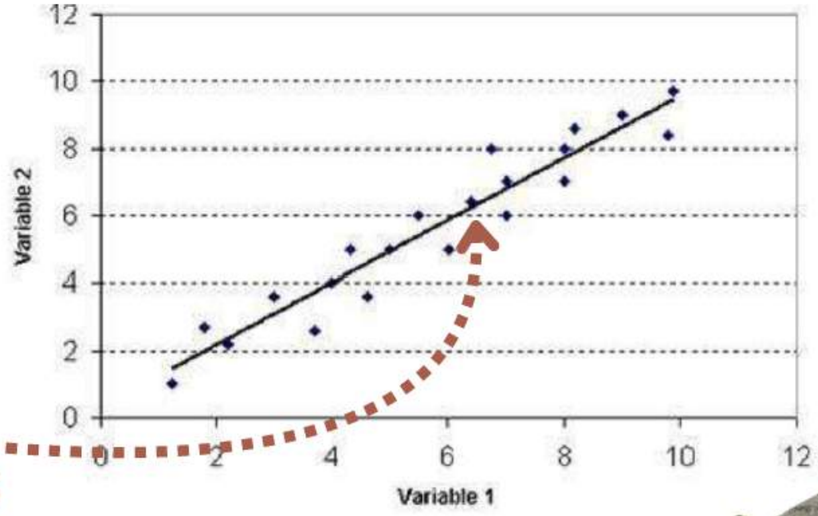
CORRÉLATION ET RÉGRESSION

Corrélation = évaluation de la liaison entre 2 variables quantitatives

Régression = méthode mathématique expliquant les relations entre variables observées



Corrélation ≠ causalité !



Droite de régression = droite des moindres carrés

Lien entre 2 variables quantitatives

TEST DE CORRÉLATION DE PEARSON

Paramètre Z → **r**

r_t vient de la table du coefficient de corrélation

DDL = n - 2

Si $r_c > r_t$ → rejet de H0

Tests non paramétriques

$4 < n < 12$

TEST U DE MANN ET WHITNEY

Hypothèse testée : **les moyennes de deux groupes de données sont proches**

Lien entre variables **quantitatives et qualitatives**

Paramètre Z → **u**

u_t vient des tables du test de Mann et Whitney

Si $u_c > u_t$ → acceptation de H0

Méthodologie

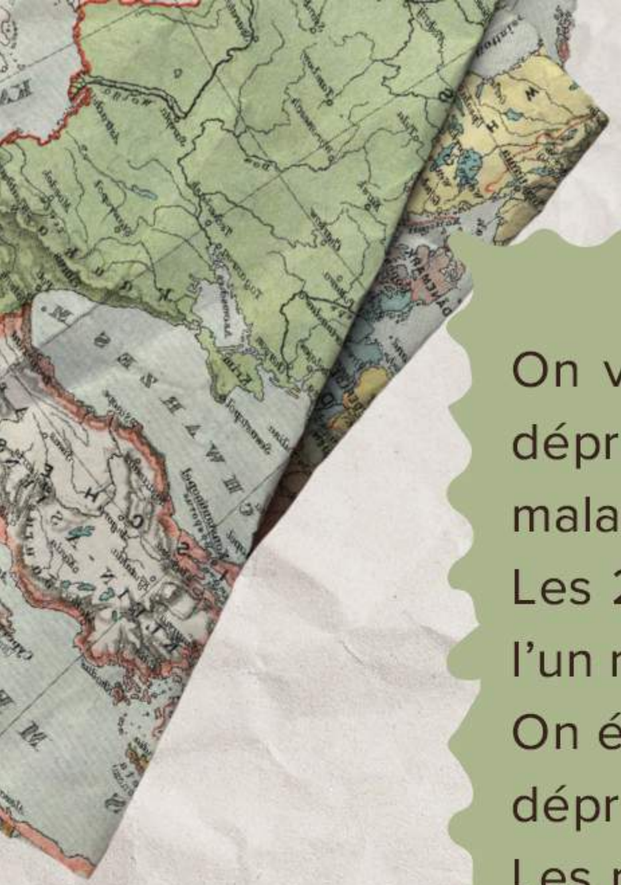
Comment trouver le paramètre théorique pour un test non paramétrique ?

n_1 est le plus petit des 2 effectifs, U le plus petit des 2 U calculés

n_1

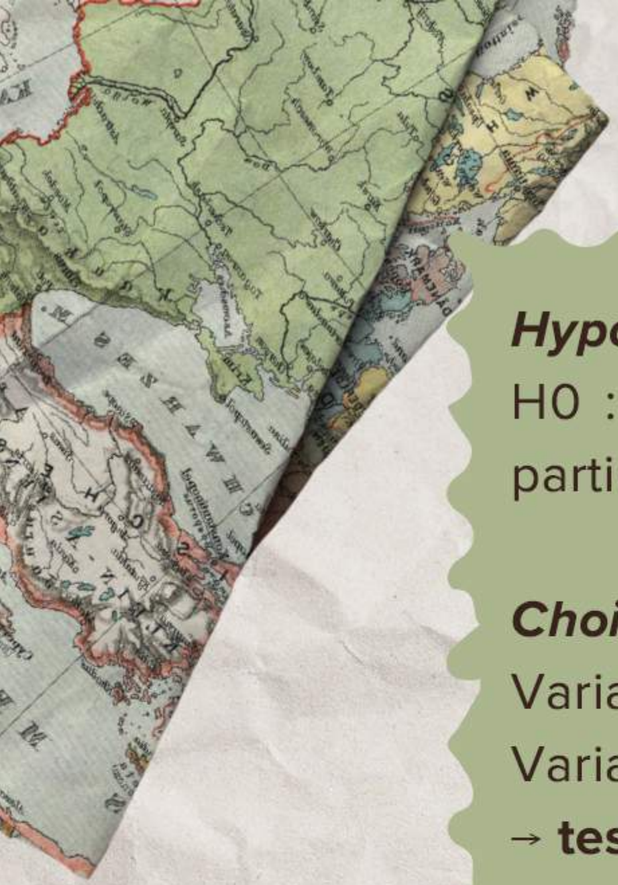
n_2-n_1	1	2	3	4	5	6	7	8	9	10
0	-	-	-	0	2	5	8	13	17	23
1		-	-	1	3	6	10	15	20	26
2		-	0	2	5	8	12	17	23	29
3		-	0	3	6	10	14	19	26	33
4		-	1	4	7	11	16	22	28	36
5		-	2	4	8	13	18	24	31	39
6		0	2	5	9	14	20	26	34	42
7		0	3	6	11	16	22	29	37	45
8		0	3	7	12	17	24	31	39	48
9		0	4	8	13	19	26	34	42	52
10		1	4	9	14	21	28	36	45	55
11		1	5	10	15	22	30	38	48	
12		1	5	11	17	24	32	41	50	
13		1	6	11	18	25	34	43		
14		1	6	12	19	27	36	45		
...										
18		2	8	16	24	33				
19		3	9	17	25					
20		3	9	17	27					

Exemple



On veut savoir si une nouvelle molécule présente un effet anti-dépresseur. Pour cela, on organise un essai portant sur 20 malades dépressifs, répartis en 2 groupes.
Les 20 malades sont répartis par TAS en 2 groupes de 10 sujet : l'un recevant la **nouvelle molécule**, l'autre recevant le **placébo**.
On évalue les patients à l'aide d'une échelle numérique de **0** (non déprimé) à **50** (très déprimé). Le groupe témoin reçoit le placébo.
Les patients des 2 groupes sont évalués **avant le traitement** puis **après le traitement** au bout de 28 jours.
Le traitement est-il efficace ?

Exemple



Hypothèses :

H0 : il n'y a pas de différence de niveau de déprime entre les participants prenant le placébo et ceux prenant la molécule

Choix du test :

Variable 1 : placébo ou molécule → qualitatif

Variable 2 : score de déprime→ pseudo-quantitatif

→ **test U de Mann et Whitney**

uc = 9

ut = 23

$u_c < u_t$ donc on rejette H0 au risque 5%

CCl : **le traitement est efficace contre la dépression.**

Tests non paramétriques

TEST R' DE SPEARMAN

$4 < n < 12$

Paramètre Z → **r'**

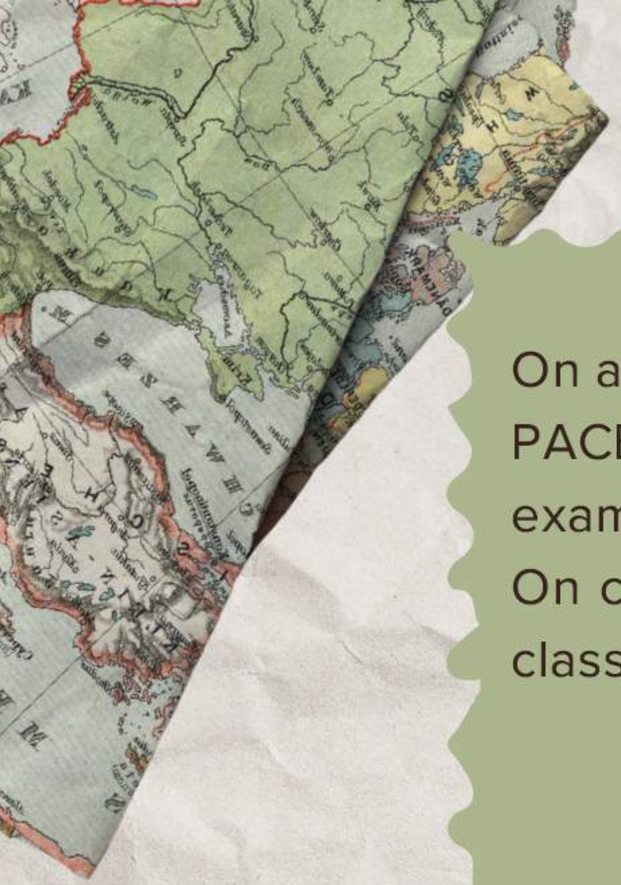
Lien entre variables **quantitatives**

r'_t vient de la table du r' de Spearman

$$r' = 1 - \frac{6 \sum d_i^2}{n\left(n^2 - 1\right)}$$

Si $r'_c > r'_t$ → acceptation de H0

Exemple

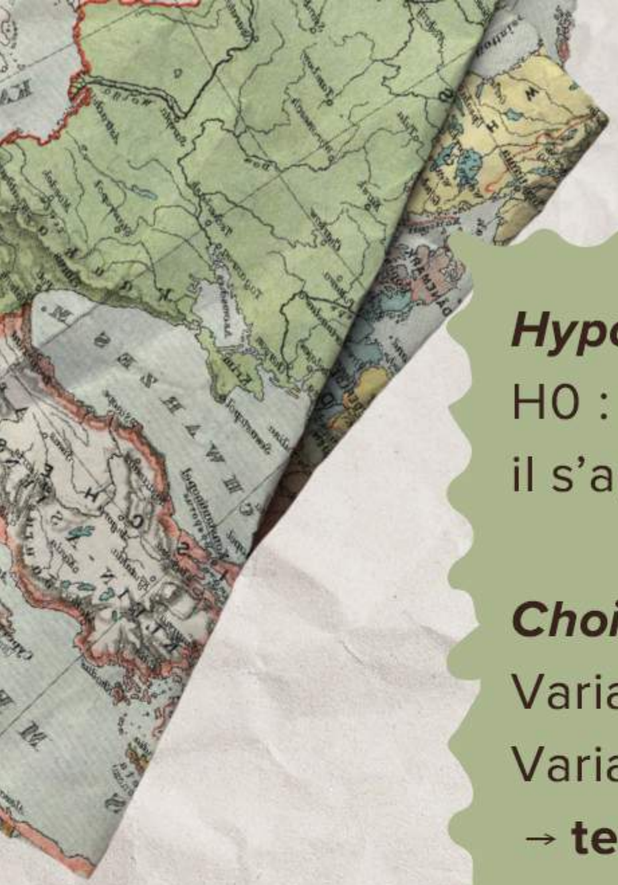


On a recensé pour 6 étudiants les notes obtenues au concours de PACES en biostatistiques, et le classement final à ce même examen.

On cherche à établir s'il existe une relation entre cette note et le classement final.

X Biostats	12,4	4,9	18,1	5,4	19,4	16
Y Classement	210	555	6	445	5	14

Exemple



Hypothèses :

H0 : il n'y a pas de lien entre ces 2 séries de valeurs numériques, il s'agit de 2 séries indépendantes

Choix du test :

Variable 1 : note → quantitatif

Variable 2 : classement→ pseudo-quantitatif

→ **test r' de Spearman**

r'c = -1

r't = 0,89 (α = 5%) et r't = 1 (α = 1%)

$r'_c < r'_t$ donc on rejette H0 au risque 1%

Il s'agit de 2 séries corrélées. Plus la note de biostat est élevée, plus petit est le rang de classement (d'où le signe - pour r').

Tableau récap

Effectif	Données quantitatives	Données qualitatives	Données quantitatives et qualitatives
$4 < n < 12$	r' de Spearman	Comp % ou χ^2	U de Mann & Whitney
$12 \leq n < 30$	Coeff de corrélation r	Comp % ou χ^2	t de Student
$n \geq 30$	Coeff de corrélation r	Comp % ou χ^2	Comp moyennes