1

On veut étudier l'efficacité d'un nouveau traitement (T) contre la leucémie.

On administre T à 50 souris et le traitement de référence (R) à 50 autres souris de la même espèce. On note au bout d'un mois, 33 morts dans le groupe T et 44 morts dans le groupe R.

Peut on conclure à la supériorité de ce nouveau traitement?

1. H0 : Il n'y a pas de différence significative entre les 2 traitements.
2. Traitements = 2 groupes T ou R : variable qualitative 1
3. Etat des souris dans chaque groupe (DCD ou non) : variable qualitative 2
4. Test de comparaison de pourcentages On fixe a priori α = 5%

		Traitements	
		T	**R**
Etat	**DCD**	**33**	**44**
	VIVANTES	**17**	**6**

2

groupe T : % DC = 33/50=66 % groupe R : %DC = 44/50=88%

On peut calculer ε

$$\varepsilon = \frac{0{,}88 - 0{,}66}{\sqrt{\frac{0{,}88 \times 0{,}12}{50} + \frac{0{,}66 \times 0{,}34}{50}}} = 2{,}83$$

ε calculé = 2,83 > ε théorique lu dans la table pour α=5% soit 1,96

On rejette H0. Il existe une diff significative entre les 2 groupes (H1),

α<0,01 à posteriori

On peut conclure : % DC souris traitées par T < % DC souris traitées par R

T est meilleur que R **sur cet échantillon.**

On ne sait rien des groupes, de l'état de santé des souris, et même de l'étude.

On ne peut pas généraliser ce résultat.

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

Etude de la liaison entre 2 caractères qualitatifs.

4

1 - Comparaison de 2 pourcentages observés.

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

$\varepsilon = 1{,}96$ avec $\alpha = 5\ \%$

q = probabilité complémentaire de p = 1 – p

2 - Test du χ^2

$$\chi^2 = \sum \frac{(O_i - C_i)^2}{C_i}$$

χ^2 tabulé

Nb ddl = (nb lignes-1)(nb colonnes-1)

Précisions concernant les ddl

ddl = nb minimal de valeurs d'une série, nécessaire afin de pouvoir calculer les manquants si l'on dispose du total ou des totaux des valeurs de cette série

1 – Test du X^2

Exemple

CAR 1

CAR 2	A	B	Total
C	n_1	n_3	Tot C
D	n_2	n_4	Tot D
Total	Tot A	Tot B	T

ddl = 1

On cherche à savoir si l'exposition professionnelle au benzène peut entrainer une leucémie. On lance une étude dans une grande entreprise, on dénombre les salariés exposés au benzène, et ceux qui ne le sont pas. Au bout de 12 ans, on fait le bilan des leucémies apparues.

	Leucémies	Non Leucémies	Total
Expo	15	485	500
Non Expo	20	980	1000
Total	35	1465	1500

Existe-t-il une relation entre exposition au benzène et leucémies ?

1. H0 : il n'existe pas de lien entre expo benzène et leucémie
2. Variable qualitative 1 : Malades leucémie ou non malades leucémie
3. Variable qualitative 2 : Exposition au benzène ou non exposition
4. Test du $\chi 2$ permet de prendre en compte tous les cas de figure (expo-malades, expo-non malades, non expo-malades, non expo-non malades) et pas seulement deux %.
5. Si répartition au hasard : les nb de leucémies seraient à peu près identiques dans les 2 groupes Expo et Non expo.
6. Nous allons donc construire ce modèle et comparer la situation réelle à ce modèle théorique.

35/1500 = 2,33% malades, expo ou non, et 1465/1500 = 97,66% non malades.

Appliquons ces % aux salariés expo et non expo : **modèle théorique**.

2,33 % de 500 = 11,65 salariés, chiffre théorique de malades chez les expo.

2,33 % de 1000 = 23,35 salariés, chiffre théorique de malades chez les non expo.

	Leucémies	Non Leucémies	Total
Expo	15	485	500
Non Expo	20	980	1000
Total	35	1465	1500
%	2,33	97,66	

Les chiffres calculés (rouge), forment le modèle théorique. Nous allons les comparer aux chiffres observés (noir), à l'aide de la formule ci-dessous :

$$\chi^2 = \sum \frac{(O_i - C_i)^2}{C_i}$$

	Leucémies		Non Leucémies		Total
Expo	15	11,65	485	488,3	500
Non Expo	20	23,35	980	976,7	1000
Total	35	35	1465	1465	1500

$\chi 2 = (15-11,65)^2/11,65 + (20-23,35)^2/23,35 + (485-488,3)^2/488,3 + (980-976,7)^2/976,7$

$\chi^2 = 1,42$

Nb de degrés de liberté = (nb lignes-1)(nb colonnes – 1) = 1

La table du χ^2 indique que χ^2 calculé < χ^2 théorique (α = 5%, soit 3,84)

Nous acceptons H0 :

Il n'existe pas de relation entre expo benzène et apparition des leucémies.

Table du $\chi 2$ (extrait)

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
...									

Etude de la liaison entre 2 caractères qualitatifs.

11

1 - Comparaison de 2 pourcentages observés.

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

ε = 1,96 avec α = 5 %

q = probabilité complémentaire de p = 1 – p

2 - Test du χ^2

$$\chi^2 = \sum \frac{(O_i - C_i)^2}{C_i}$$

χ^2 tabulé

Nb ddl = (nb lignes-1)(nb colonnes-1)

On veut étudier l'efficacité d'un nouveau traitement T contre la leucémie.

On administre T à 50 souris et le traitement de référence R à 50 autres souris de la même espèce. On note au bout d'un mois, 33 DC dans le groupe T et 44 DC dans le groupe R.

Peut on conclure à la supériorité de ce nouveau traitement?

1. On dispose de 2 groupes indépendants traités T ou R : variable qualitative
2. Dénombrements des DC dans chaque groupe (DC ou non) : variable qualitative
3. Comparaison de % (vu précédemment), ou test du χ^2
4. H0 : Il n'y a pas de différence significative entre les 2 traitements.

	T	R	TOTAL
Morts	33 (38,5)	44 (38,5)	77
Vivants	17 (11,5)	6 (11,5)	23
TOTAL	50	50	100

77% DC

23% VIVANTS

77% DC
Gr T : 77% x 50 = 38,5
Gr R : 77% x 50 = 38,5
23% VIVANTS
Gr T et R : 23% x 50 = 11,5

$$\chi^2 = \sum \frac{(O_i - C_i)^2}{C_i}$$

Si hasard : même nb de DC et de vivants dans les 2 groupes…

$\chi 2 = 6,83$ ddl = 1

$\chi 2$ théorique = 3,84 ; $\alpha = 5\%$

$\chi 2$ calculé > $\chi 2$ théorique : **on rejette H0**

Il existe une différence significative entre les 2 traitements ($\alpha < 1\%$).

On a le droit d'interpréter le contenu du tableau :

on peut conclure que T est meilleur que R sur cet échantillon.

PLAN GÉNÉRAL DU COURS

1 - Biostatistique

2 - Statistique Descriptive

3 - Statistique Déductive

- ***Liaisons entre caractères qualitatifs***
- ***Liaisons entre caractères qualitatifs et quantitatifs***
- ***Liaisons entre caractères quantitatifs***
- ***Tests non paramétriques***

Liaison entre caractères qualitatifs et quantitatifs

Question : En moyenne la taille des individus d'une population *A* coïncide-t-elle avec la taille des individus d'une population *B* ?

1 - Comparaison de moyennes

n_1 et $n_2 > 30$ "Grands échantillons"

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

Table de l'écart réduit

$\varepsilon = 1{,}96$ avec $\alpha = 5$ %

2 - Test t de Student

n_1 ou $n_2 < 30$ "Petits échantillons"

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

Table t de Student

Nb ddl = $(n_1-1)+(n_2-1)$

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

= écart type sur les 2 échantillons

On cherche à comparer les taux de T3 libre (hormone thyroïdienne) chez des femmes prenant un contraceptif oral (c.o) et chez des femmes n'en prenant pas.

On dispose, après TAS de 2 groupes de femmes :

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 $s_2=0,30$ nmol

Les taux de T3 libre peuvent ils être considérés comme « identiques »

dans les 2 groupes ou bien sont ils significativement différents ?

1. H0 : m_1 et m_2 ne sont guères différentes. Ce sont 2 estimateurs de la valeur moyenne de T3 libre chez la femme, en général.

2. Relation entre caractères qualitatifs (prise ou non de c.o) et quantitatifs (dosages de T3 libre, ici valeur moyenne)

3. n_1 et $n_2 > 30$ ⟹ test de comparaison de moyennes

4. $$\varepsilon = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{2,5 - 2}{\sqrt{\frac{0,35^2}{50} + \frac{0,30^2}{33}}} = 6,94$$

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	3,89059	4,41717	4,89164	5,32672	5,73073	6,10941

$\alpha = 0,0001$ $\varepsilon = 3,89$ Très significatif. On rejette H0 avec $p < 0,0001$

TAS : donc médicalement, résultat généralisable :

La prise de contraceptifs oraux augmente le taux de T3libre

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	3,89059	4,41717	4,89164	5,32672	5,73073	6,10941

19

On teste un antiviral diminuant le nb de jours de symptômes cliniques chez des patients infectés par le virus de la grippe.

Soit 100 sujets non traités, atteints de grippe.

Nb moyen de jours avec symptômes $m_1 = 4,74$ jours et $s_1 = 1$.

Soit 100 autres sujets traités avec l'antiviral et atteints de grippe

Nb moyen de jours avec symptômes $m_2 = 4,2$ jours et $s_2 = 1,7$.

Parmi les propositions suivantes choisir celles qui sont exactes :

A) Le test de comparaison de moyennes permet de rejeter/accepter H0
B) Le test de comparaison de moyennes ne permet pas de rejeter/garder H0
C) On ne pourra pas conclure à l'efficacité du tt à cause du risque de 1ère espèce 5%
D) On ne pourra pas conclure à l'efficacité du tt à cause du risque de 2ème espèce inconnu
E) On ne pourra pas généraliser le résultat car l'étude n'a pas été bien menée

20

A) Le test de comparaison de moyennes permet de rejeter/accepter H0
B) Le test de comparaison de moyennes ne permet pas de rejeter/garder H0
C) On ne pourra pas conclure à l'efficacité du tt à cause du risque de 1ère espèce 5%
D) On ne pourra pas conclure à l'efficacité du tt à cause du risque de 2ème espèce inconnu
E) On ne pourra pas généraliser le résultat car l'étude n'a pas été bien menée

Réponse A. H0 : m1 et m2 ne sont pas significativement différentes
Le test de comparaison de moyennes, comme tous les tests, répond à la question : peut on accepter ou rejeter H0 ?

Réponse E. L'étude aurait due être randomisée (TAS) : Traitement contre Placebo

Les items **B, C, D** sont faux (B), ou sans rapport (C, D)

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	3,89059	4,41717	4,89164	5,32672	5,73073	6,10941

Liaison entre caractères qualitatifs et quantitatifs

Question : En moyenne la taille des individus d'une population A coïncide-t-elle avec la taille des individus d'une population B ?

1 - Comparaison de moyennes

n_1 et $n_2 > 30$ "Grands échantillons"

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

Table de l'écart réduit

$\varepsilon = 1{,}96$ avec $\alpha = 5\ \%$

2 - Test t de Student

n_1 ou $n_2 < 30$ "Petits échantillons"

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

Table t de Student

Nb ddl = $(n_1-1)+(n_2-1)$

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

= écart type sur les 2 échantillons

2 – Série numérique : t de Student

Exemple série 8 valeurs donc n = 8

	2	3	5	12	10	4	7	8	Total = 51
1 valeur manquante	2	3	5	12	10		7	8	Total = 47

Manquant = 51 – 47 = 4 avec n-1 valeurs : on peut calculer la valeur manquante à partir du total

2 valeurs manquantes	2	3		12	10		7	8

Total = 42

Somme des manquants = 51 – 42 = 9 impossible de calculer les 2 valeurs manquantes

Donc ddl = n-1 = 7

t de Student : 2 séries à comparer donc ddl = $(n_1 - 1) + (n_2 - 1)$

Soient un groupe de 15 femmes obèses, et un autre groupe de 12 femmes de poids normal. On a mesuré le taux de corticoïdes sanguins moyens à l'intérieur de ces 2 groupes.

Gr 1 :	$n_1 = 15$	$m_1 = 6,3$	$s_1 = 1,8$
Gr 2 :	$n_2 = 12$	$m_2 = 4,5$	$s_2 = 1,6$

L'obésité a-t-elle une influence sur le taux de corticoïdes ?

1. H0 : m_1 et m_2 ne sont pas différentes dans ces 2 groupes.

2. Relation entre caractères qualitatifs (obèses et non obèses), et quantitatifs (valeurs de dosages sanguins, valeurs moyennes)

3. n_1 et $n_2 < 30$ petits échantillons ⇒ t de student

4. Calcul de l'écart type commun aux 2 groupes. La formule s'écrit :

$$s^2 = \frac{(n_1-1) \times s_1^2 + (n_2-1) \times s_2^2}{(n_1+n_2-2)} = 2,53$$

nb ddl = (15+12) – 2 = 25

$$t = \frac{6,3 - 4,5}{\sqrt{\frac{2,53}{15} + \frac{2,53}{12}}} = 2,92$$

> 2,06 à 5% lu dans la table t Student

On rejette H0 avec α = 1% défini a posteriori.
Il existe une relation entre obésité et augmentation du taux de corticoïdes au niveau de ces échantillons.

TABLE DU t DE STUDENT

α

ddl	0,9	0,5	0,3	0,2	0,1	0,05	0,02	0,01	
1	0,158	1	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,92	4,303	6,965	9,925	31,598
3	0,137	0,765	1,25	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,19	1,533	2,132	2,776	3,747	4,604	8,61
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	1,44	1,943	2,447	3,143	3,707	5,959
7	0,13	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,13	0,706	1,108	1,397	1,86	2,306	2,896	3,355	5,041
9	0,129	0,703	1,1	1,383	1,833	2,262	2,821	3,25	4,781
10	0,129	0,7	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	1,35	1,771	2,16	2,65	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,14
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
17	0,128	0,689	1,069	1,333	1,74	2,11	2,567	2,898	3,965
18	0,127	0,688	1,067	1,33	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	3,85
....									
25	0,127	0,684	1,058	1,316	1,708	2,06	2,485	2,787	3,725

Liaison entre caractères qualitatifs et quantitatifs

Séries appariées
ou
Méthode des couples

1 - Comparaison de moyennes

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

2 - Test t de Student

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

Liaison entre caractères qualitatifs et quantitatifs.

28

Exemple

Comparer deux méthodes de dosage de la glycémie. On dispose de n patients, auxquels on prélève 2 tubes de sang. On dose la glycémie dans chacun de ces tubes par une méthode différente.

On souhaite comparer les valeurs moyennes de ces 2 séries de n résultats. La question posée est :

Ces 2 méthodes de dosage fournissent elles des résultats identiques ?

On calcule **si n>30** $\varepsilon = m_d / \sqrt{\frac{s^2}{n}}$ **si n<30** $t = m_d / \sqrt{\frac{s^2}{n}}$

Avec d=différence des résultats pour un même sujet, m_d=moyenne des d, n=nb de couples, s=variance des différences

Puis la méthodologie est identique aux tests déjà vus : on compare cette valeur calculée aux valeurs dans la table adaptée, et la conclusion se fait de la même manière **en fixant un risque α.**

Cas des séries appariées. Méthode des couples

On souhaite évaluer l'intérêt d'une substance S capable de désintoxiquer les fumeurs. On constitue par T.A.S. 2 groupes de 40 fumeurs, un reçoit S, l'autre reçoit un placebo P. Le traitement dure 2 mois pour les 2 groupes. La consommation de cig/jours (C) est notée avant et après traitement.

	S (n=40)		P (n=40)	
	m_1	s_1^2	m_2	s_2^2
C avant tt	19,5	54,2	16,5	35,6
C après tt	5,4	30,4	3,8	20,1
Variation de C	14,1	9,1	12,7	8,9

Cas des séries appariées. Méthode des couples

	S (n=40)		P (n=40)	
	m_1	s_1^2	m_2	s_2^2
C avant tt	19,5	54,2	16,5	35,6
C après tt	5,4	30,4	3,8	20,1
Variation de C	14,1	9,1	12,7	8,9

1) Quelle est la première précaution à prendre ?

2) Dans le groupe Placebo, la conso moyenne après tt diffère t elle de la valeur avant tt ? Interpréter le résultat.

3) Les 2 groupes diffèrent ils pour leur conso moyenne après traitement?

4) Les 2 groupes diffèrent ils pour la variation de conso avant/après tt ?

Cas des séries appariées. Méthode des couples

	S (n=40)		P (n=40)	
	m_1	s_1^2	m_2	s_2^2
C avant tt	19,5	54,2	16,5	35,6
C après tt	5,4	30,4	3,8	20,1
Variation de C	14,1	9,1	12,7	8,9

1)Quelle est la première précaution à prendre ?

Conso identique dans les 2 groupes?

Les 2 groupes doivent être comparables vis-à-vis des paramètres susceptibles d'influencer la réponse au traitement (âge, sexe, CSP, conso/jour etc..).

Si ce n'est pas le cas, il faut en tenir compte lors des conclusions.

Comparaison des conso moyennes avant tt dans les 2 groupes:

1. **H0 = les moyennes des conso sont équivalentes dans les 2 groupes.**
2. **Etude liaison entre variables qualitatives (S ou P) et quantitatives (nb cig/j) dans 2 échantillons indépendants**
3. **n> 30 ⇒ Test de comparaison de moyennes**
4. $$\varepsilon = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s^2}{n_2}}} = \frac{19,5 - 16,5}{\sqrt{\frac{54,2}{40} + \frac{35,6}{40}}} = 2,00 > 1,96 \quad (\varepsilon \text{ pour } \alpha = 5\%)$$

On rejette H0, avec un risque α = 5%. Il existe donc une différence significative entre les conso moyennes des 2 groupes : on fume plus dans le groupe S : il faudra en tenir compte lors de l'étude de la variation de cette conso avant/après traitement.

2) Dans le groupe Placebo, la conso moyenne après tt diffère t elle de la valeur avant tt ? Interpréter le résultat.

P (n=40)	
m_2	s_{2_2}
16,5	35,6
3,8	20,1
12,7	8,9

1.Liaison entre variable qualitative (avant / après tt) et quantitative (nb cig/j)

2. Echantillons non indépendants (méthode des couples)

3. n >30 ⇒ **Test de comparaison de moyennes**

$$\varepsilon = m_d / \sqrt{\frac{s^2}{n}} = 12,7 / \sqrt{\frac{8,9}{40}} = 26,9 \quad > 1,96 \text{ au risque } \alpha = 5\%$$

On rejette H0. Il existe une différence très significative (p < 0,001) entre les consommations avant / après tt, dans le groupe P.

Effet psychologique : envie de profiter de l'étude pour arrêter de fumer ?

Cas des séries appariées. Méthode des couples

3) Les 2 groupes diffèrent ils pour leur conso moyenne après traitement?

S (n=40)		P (n=40)	
m_1	s_{1_2}	m_2	s_{2_2}
19,5	54,2	16,5	35,6
5,4	30,4	3,8	20,1
14,1	9,1	12,7	8,9

1. **H0 = les moyennes des conso sont équivalentes dans les 2 groupes.**

2. **Liaison entre variables qualitatives (S ou P) et quantitatives(nb cig/j) dans 2 échantillons indépendants**

3. **n> 30** ⇒ **Test de comparaison de moyennes**

4. $$\varepsilon = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{5,4 - 3,8}{\sqrt{\frac{30,4}{40} + \frac{20,1}{40}}} = 1,42 < 1,96$$

On accepte H0 : il n'existe pas de différence significative entre les 2 groupes pour la consommation après tt.

4) Les 2 groupes diffèrent-ils pour la variation de conso avant/après tt ?

Il faut comparer les variations avant / après tt dans les 2 groupes afin de prouver l'intérêt de la substance S

1.H0 : Il n'existe pas de différence entre les variations de consommation dans les 2 groupes
2.Etude liaison entre variables qualitatives (S ou P) et quantitatives (nb cig/j) dans 2 échantillons indépendants
3. n > 30 >>>> Test de comparaison de moyennes

4. $$\varepsilon = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{14{,}1 - 12{,}7}{\sqrt{\frac{9{,}1}{40} + \frac{8{,}9}{40}}} = 2{,}09 > 1{,}96$$ **au risque 5%**

<u>On rejette H0 :</u> il existe une différence significative entre les variations de conso dans les 2 groupes (p<5%) . Conclusion : efficacité de S. Il y avait eu TAS, donc résultat généralisable.

Conclusion : Pas de différence après tt dans chaque groupe (Quest 3). Mais Gr S fumait + (Quest 1) >>>> Efficacité du traitement S.

PLAN GÉNÉRAL DU COURS

1 - La Biostatistique

2 - Statistique Descriptive

3 - Statistique Déductive

- ***Liaisons entre caractères qualitatifs***
- ***Liaisons entre caractères qualitatifs et quantitatifs***
- ***Liaisons entre caractères quantitatifs***
- ***Tests non paramétriques***

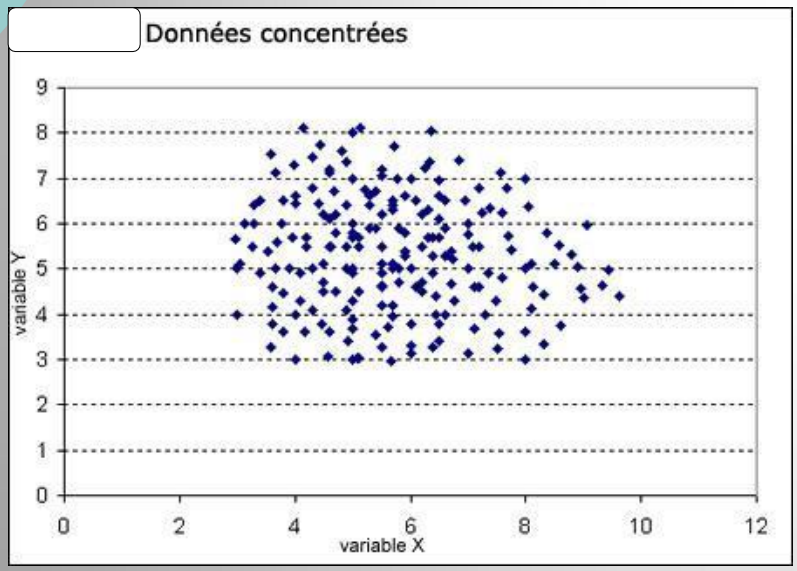
Etude de la liaison entre caractères quantitatifs

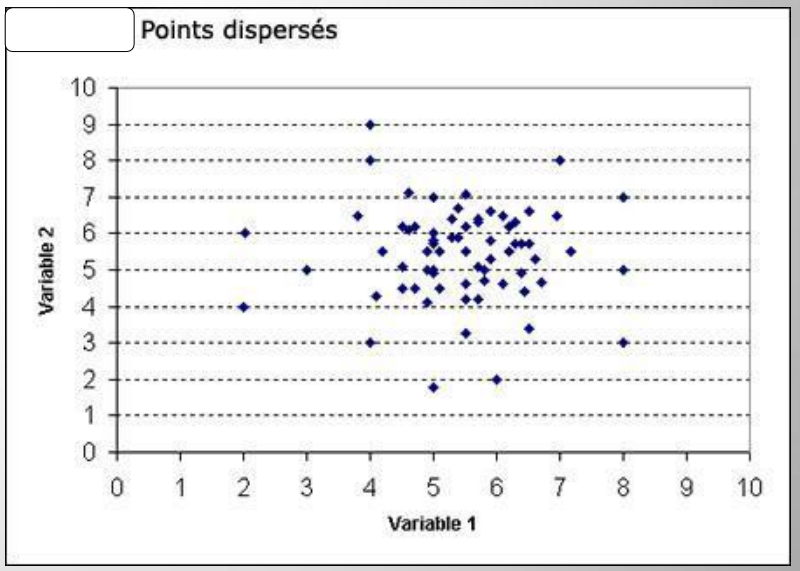
Corrélation et régression

Corrélation = Evaluation de la liaison entre 2 variables quantitatives

Régression = Méthode mathématique expliquant les relations entre variables observées.

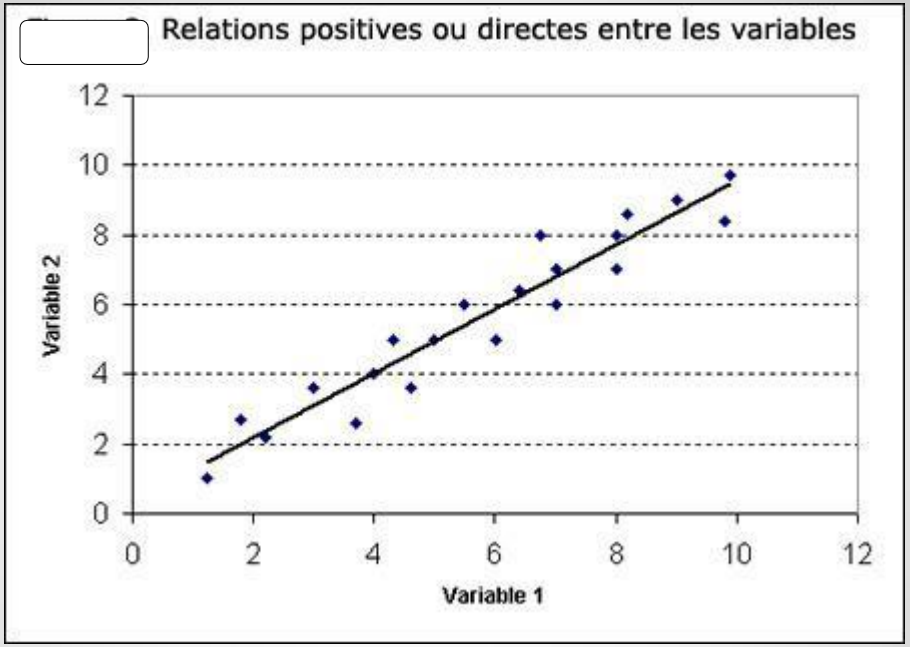
Représentation des données : **nuages de points**





Etude de la liaison entre caractères quantitatifs

Nuages de points



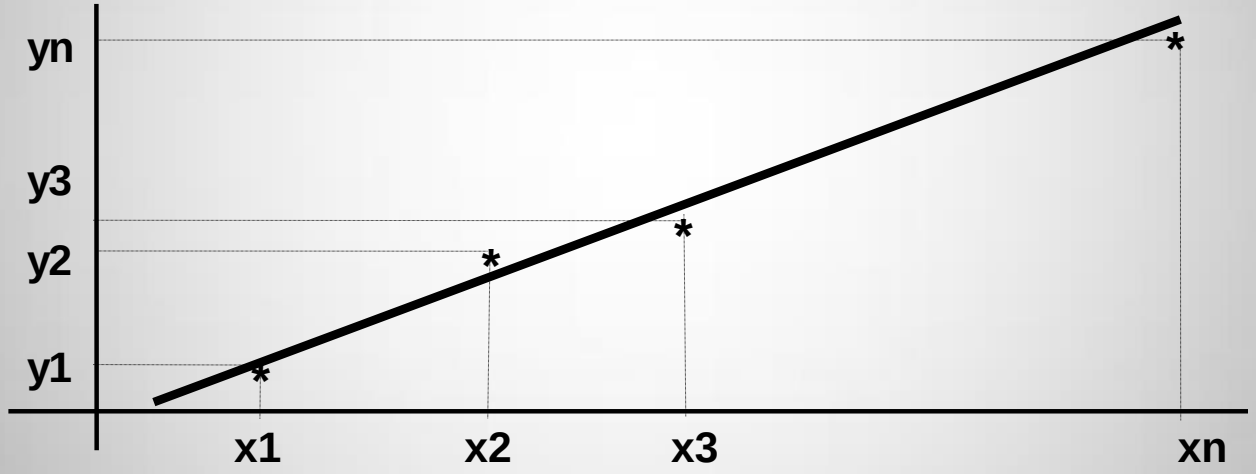
La droite de régression permet de visualiser si une des 2 variables est dépendante de l'autre..

Etude de la liaison entre caractères quantitatifs

- La capacité respiratoire est elle dépendante de la consommation de cigarettes?
- Le poids des bébés à la naissance est il lié à l'âge de la mère?

si x et y liées alors ⇒ y = f(x) ***droite de régression de y en x***

y peut être « expliqué » en fc de x

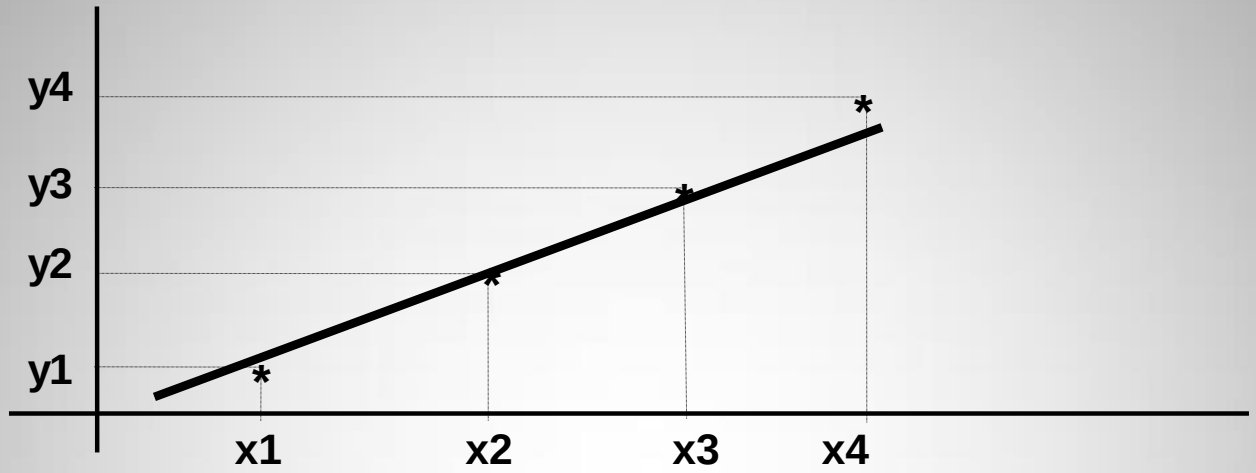


Droite de régression:

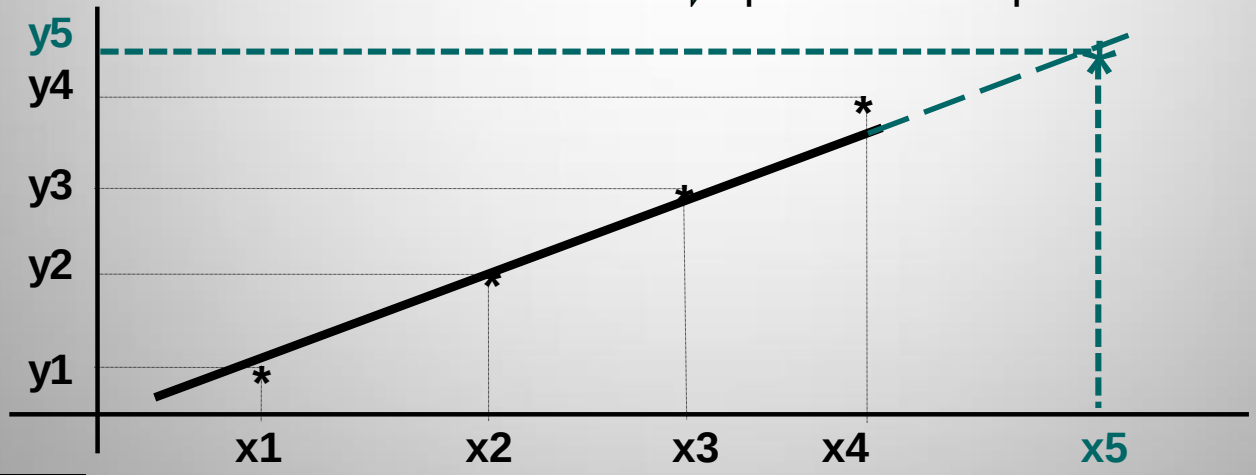
Droite des moindres carrés, passe au «plus près» de chaque point du graphe.

Dans ce cours, on ne parle que de **régression linéaire**.

Etude de la liaison entre caractères quantitatifs



La prédiction avec une droite de régression :
Pour nouvelle valeur de x ➡ quelle valeur possible de y ?



41

coefficient de corrélation = **pente** de cette droite. Soient (x_i,y_i) les couples dont on cherche à étudier la corrélation, m_x et m_y sont les moyennes des x_i et y_i respectivement

$$r = \frac{\sum (x_i - m_x)(y_i - m_y)}{\sqrt{\sum (x_i - m_x)^2 \sum (y_i - m_y)}}$$

$$r = \frac{\Sigma xy - \frac{\Sigma x . \Sigma y}{n}}{\sqrt{(\Sigma x^2 - \frac{(\Sigma x)^2}{n})(\Sigma y^2 - \frac{(\Sigma y)^2}{n})}}$$

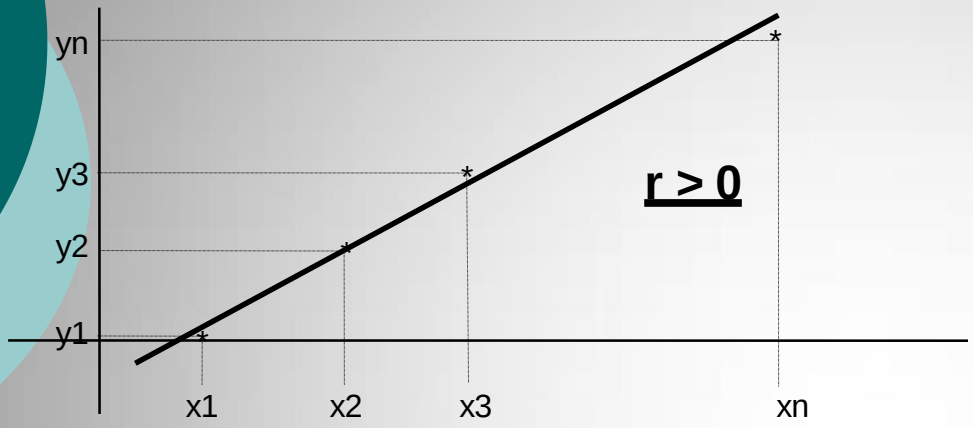
Si r > 0 liaison positive : x et y varient dans le même sens
Si r < 0 liaison négative : x et y varient en sens inverse

r est toujours < 1 avec n – 2 ddl

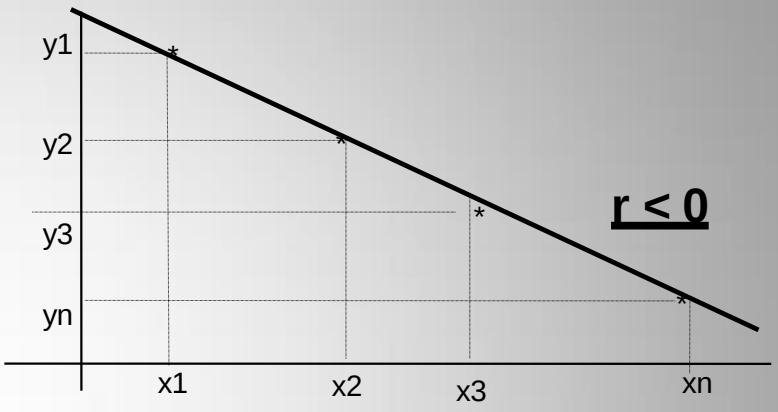
	X	Y
n1	10	4
n2	12	5
n3	5	23
n4	9	**b**
n5	22	4
n6	**a**	6
Total	**63**	**47**

Corrélation

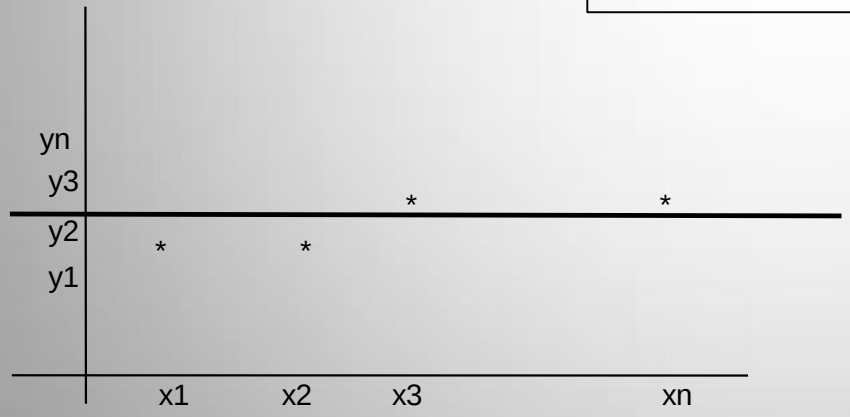
Liaison positive

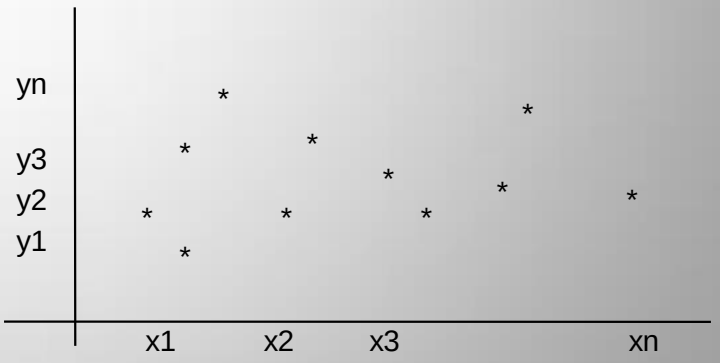


Liaison négative



Pas de liaison





43

Sur un échantillon de 10 sujets d'âges différents, on recueille les données suivantes : âge (années) et concentration de cholestérol dans le sang (g/L)

X âges	30	60	40	20	50	30	40	20	70	60
Y chol	1,6	2,5	2,2	1,4	2,7	1,8	2,1	1,5	2,8	2,6

Le taux de cholestérol est il lié à l'âge ?

Existe-t-il un lien entre ces 2 séries de données? Ou bien s'agit t il de 2 séries totalement indépendantes?

H0 = Le taux de cholestérol est indépendant de l'âge

H1 = Le taux de cholestérol est lié à l'âge

2 variables quantitatives >>> **Test du coefficient de corrélation.**

$$r = \frac{\sum xy - \frac{\sum x . \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

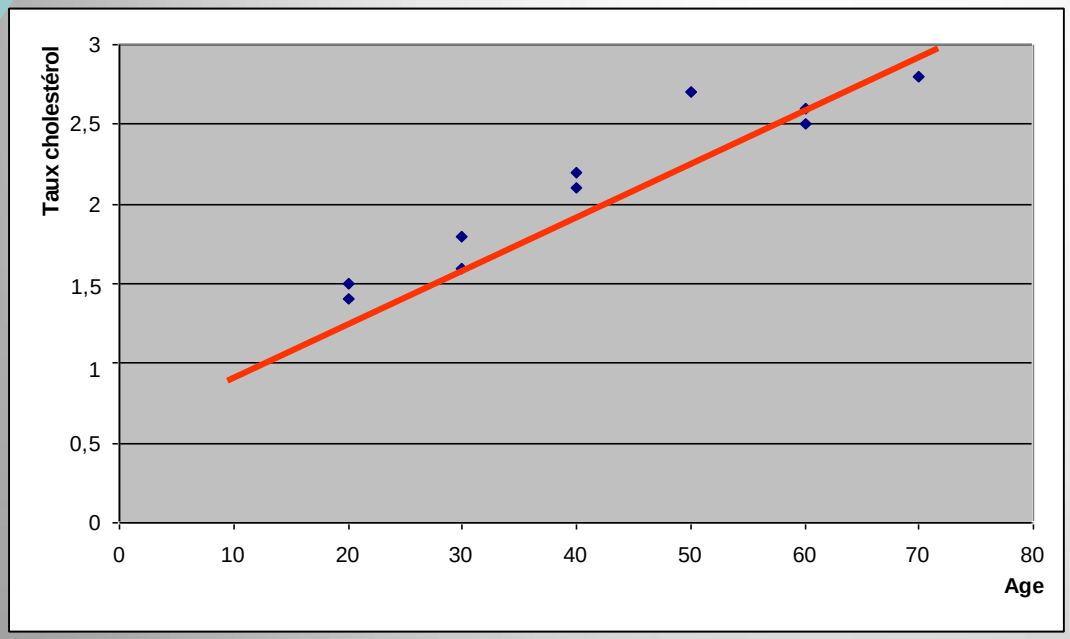
X âges	30	60	40	20	50	30	40	20	70	60
Y chol	1,6	2,5	2,2	1,4	2,7	1,8	2,1	1,5	2,8	2,6

r calculé = 0,955

r théorique (α = 1%) avec 10-2 = 8 ddl = 0,76 Table

r calculé > r théorique **Rejet de H0**

Il existe une relation significative (α= 1%) entre l'âge et le taux de cholestérol



Plus l'âge augmente, plus le taux de cholestérol augmente
Résultat non généralisable

Corrélation n'est pas **causalité**

df \ α	0.2	0.1	0.05	0.02	0.01	0.001
1	0.951057	0.987688	0.996917	0.999507	0.999877	0.999999
2	0.800000	0.900000	0.950000	0.980000	0.990000	0.999000
3	0.687049	0.805384	0.878339	0.934333	0.958735	0.991139
4	0.608400	0.729299	0.811401	0.882194	0.917200	0.974068
5	0.550863	0.669439	0.754492	0.832874	0.874526	0.950883
6	0.506727	0.621489	0.706734	0.788720	0.834342	0.924904
7	0.471589	0.582206	0.666384	0.749776	0.797681	0.898260
8	0.442796	0.549357	0.631897	0.715459	0.764592	0.872115
9	0.418662	0.521404	0.602069	0.685095	0.734786	0.847047
10	0.398062	0.497265	0.575983	0.658070	0.707888	0.823305
11	0.380216	0.476156	0.552943	0.633863	0.683528	0.800962
12	0.364562	0.457500	0.532413	0.612047	0.661376	0.779998
13	0.350688	0.440861	0.513977	0.592270	0.641145	0.760351
14	0.338282	0.425902	0.497309	0.574245	0.622591	0.741934
15	0.327101	0.412360	0.482146	0.557737	0.605506	0.724657
16	0.316958	0.400027	0.468277	0.542548	0.589714	0.708429
17	0.307702	0.388733	0.455531	0.528517	0.575067	0.693163
18	0.299210	0.378341	0.443763	0.515505	0.561435	0.678781
19	0.291384	0.368737	0.432858	0.503397	0.548711	0.665208
20	0.284140	0.359827	0.422714	0.492094	0.536800	0.652378
21	0.277411	0.351531	0.413247	0.481512	0.525620	0.640230
22	0.271137	0.343783	0.404386	0.471579	0.515101	0.628710
23	0.265270	0.336524	0.396070	0.462231	0.505182	0.617768
24	0.259768	0.329705	0.388244	0.453413	0.495808	0.607360
25	0.254594	0.323283	0.380863	0.445078	0.486932	0.597446
26	0.249717	0.317223	0.373886	0.437184	0.478511	0.587988
27	0.245110	0.311490	0.367278	0.429693	0.470509	0.578956
28	0.240749	0.306057	0.361007	0.422572	0.462892	0.570317
29	0.236612	0.300898	0.355046	0.415792	0.455631	0.562047
30	0.232681	0.295991	0.349370	0.409327	0.448699	0.554119

df \ α	0.2	0.1	0.05	0.02	0.01	0.001
35	0.215598	0.274611	0.324573	0.380976	0.418211	0.518898
40	0.201796	0.257278	0.304396	0.357787	0.393174	0.489570
45	0.190345	0.242859	0.287563	0.338367	0.372142	0.464673
50	0.180644	0.230620	0.273243	0.321796	0.354153	0.443201
60	0.164997	0.210832	0.250035	0.294846	0.324818	0.407865
70	0.152818	0.195394	0.231883	0.273695	0.301734	0.379799
80	0.142990	0.182916	0.217185	0.256525	0.282958	0.356816
90	0.134844	0.172558	0.204968	0.242227	0.267298	0.337549
100	0.127947	0.163782	0.194604	0.230079	0.253979	0.321095
125	0.114477	0.146617	0.174308	0.206245	0.227807	0.288602
150	0.104525	0.133919	0.159273	0.188552	0.208349	0.264316
175	0.096787	0.124036	0.147558	0.174749	0.193153	0.245280
200	0.090546	0.116060	0.138098	0.163592	0.180860	0.229840
250	0.081000	0.103852	0.123607	0.146483	0.161994	0.206079
300	0.073951	0.094831	0.112891	0.133819	0.148019	0.188431
350	0.068470	0.087814	0.104552	0.123957	0.137131	0.174657
400	0.064052	0.082155	0.097824	0.115997	0.128339	0.163520
450	0.060391	0.077466	0.092248	0.109397	0.121046	0.154273
500	0.057294	0.073497	0.087528	0.103808	0.114870	0.146436
600	0.052305	0.067103	0.079920	0.094798	0.104911	0.133787
700	0.048427	0.062132	0.074004	0.087789	0.097161	0.123935
800	0.045301	0.058123	0.069234	0.082135	0.090909	0.115981
900	0.042711	0.054802	0.065281	0.077450	0.085727	0.109385
1000	0.040520	0.051993	0.061935	0.073484	0.081340	0.103800
1500	0.033086	0.042458	0.050582	0.060022	0.066445	0.084822
2000	0.028654	0.036772	0.043811	0.051990	0.057557	0.073488
3000	0.023397	0.030027	0.035775	0.042457	0.047006	0.060027
4000	0.020262	0.026005	0.030984	0.036773	0.040713	0.051996
5000	0.018123	0.023260	0.027714	0.032892	0.036417	0.046512