Statistiques déductives

I. Généralités sur les tests d'hypothèse

Dans les statistiques **déductives**, contrairement aux statistiques **descriptives**, on essaie, à partir des observations faites, de **tirer des conclusions**. Pour cela, les épidémiologistes utilisent des **tests d'hypothèse**.

1. Les tests de comparaison

Le plus souvent, les tests utilisés en statistiques déductives sont des tests de **comparaison** entre **2 populations** présentant des caractères/paramètres différents. On constitue alors **2 échantillons représentatifs** et on essaie de déterminer s'il existe une différence significative entre ces 2 échantillons pour le caractère étudié. Le but étant d'**extrapoler** le résultat aux 2 populations primitives.

2. La définition des hypothèses

Avant de commencer une étude statistique on **formule des hypothèses** que le test permettra ensuite de **confirmer** ou d'**infirmer**. On définira au début de chaque test 2 hypothèses jouant un rôle **symétrique** :

1. **H0 = hypothèse nulle**
   - Il n'y a pas de différence observée entre les deux groupes
   - Il n'existe pas de lien entre les 2 caractères étudiés, les fluctuations observées sont donc dues au hasard
2. **H1 = hypothèse alternative**
   - Il y a une différence significative entre les deux groupes
   - Il existe bien un lien entre les 2 caractères étudiés, les fluctuations observées ne sont donc pas dues au hasard

Les tests sont donc des techniques permettant de décider si on accepte ou si on rejette H0, en ayant fixé le risque d'erreur α accompagnant cette décision.

3. Les étapes d'un test d'hypothèse

Pour mettre en œuvre un test d'hypothèse, on suivra toujours les étapes suivantes :

1) Définir **H0** et **H1**
2) Déterminer les **caractères des données** (qualitative/quantitative) et choisir le **bon test** en fonction des données
3) Choisir **le risque α a priori** (généralement 5%)
4)
   - Recueillir les données
   - Calculer Z
   - Utiliser la **règle de rejet / décision** (basée sur H0 et α)
   - Fixer le **risque d'erreur réel a posteriori**

5) **Interprétation** des résultats : au niveau de l'**échantillon** (accepter H0 ?) et au niveau de la **population** (extrapolation ?)

4. La notion de risque

Rappel de statistique descriptive : *Lors de l'estimation d'une valeur x par un IC, α représente le risque d'erreur dans l'estimation de x, c'est à dire le risque pour que l'IC ne contienne pas la vraie valeur de x. Il est généralement fixé à 5%*

En statistique déductive on a :

• **α** ou **risque de première espèce** : le risque de **rejeter H0 si H0 est vraie**. Ce risque d'erreur est **maîtrisé**, c'est à dire qu'il est **fixé** (le plus souvent à 5%) **avant l'application** du test statistique.

• **β** ou **risque de seconde espèce** : le risque **d'accepter H0 si H0 est fausse**. Ce risque d'erreur est **négligé** et peut être assez **important**.

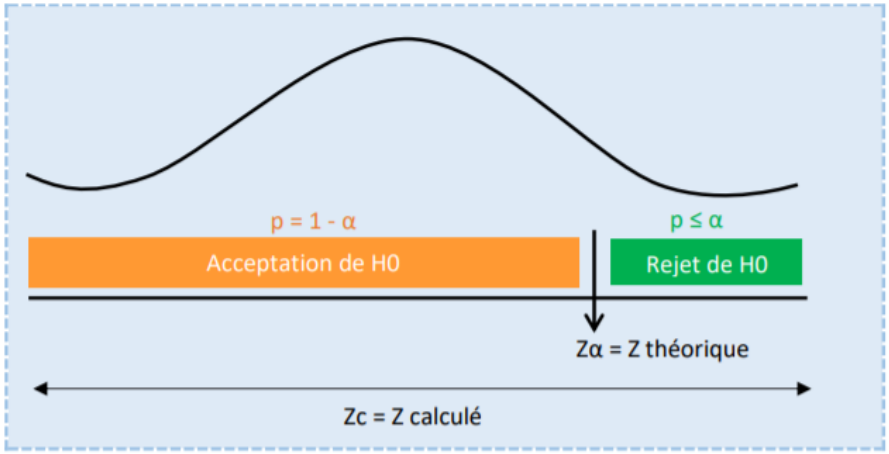
Il y a une **dissymétrie** dans le traitement des deux hypothèses (parce qu'on choisit de maîtriser α quitte à ignorer β)

• **1- β** ou la **puissance du test** : la probabilité de **rejeter H0 si H0 est fausse**.

		Décision du statisticien	
		Rejet H0	**Non rejet H0**
Réalité	**H0 vraie**	α	1 - α
	H1 vraie	1 - β	β

5. Interprétation graphique du risque α

Le paramètre $\mathbf{Z_{calculé}}$ que nous allons apprendre à calculer, suit une **distribution probabiliste** en forme de **courbe de Gauss**.



Pour pouvoir arriver à une conclusion après une étude statistique on :

1. Fixe **α à priori**

2. Cherche le $\mathbf{Z_{théorique}}$ **sur la table** (cf plus loin pour le trouver)

3. **Calcule** $\mathbf{Z_c}$ grâce aux formules

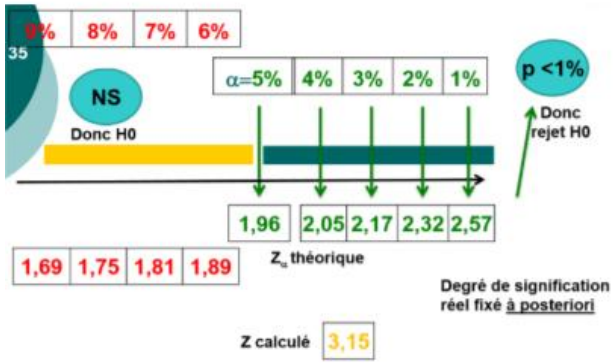
4. **Compare** $\mathbf{Z_c}$ **avec** $\mathbf{Z_t}$ et on peut arriver à **deux conclusions** différentes :

Acceptation de H0	Rejet de H0
Zc < Zt	Zc > Zt
p = 1 - α	p ≤ α

5. **<u>Fixe le degré de signification p a posteriori</u>**

Remarque : α est fixé à priori par le statisticien (=supposition) mais le degré de signification est fixé à postériori (=réel) parce que la précision de l'étude peut s'avérer être supérieure à celle qu'on a supposé.

Exemple :



1. α = **5%**
2. $Z_α$=**1,96**.
3. Zc=3,15.
4. 3,15 > **1,96** donc on rejette H0
5. On voit sur la table (ou le schéma) que pour α=1%, $Z_α$= 2,57.
   Or, 3,15 > 2,57 donc le degré de signification fixé à postériori est **< à 1%** : la précision a donc augmenté.

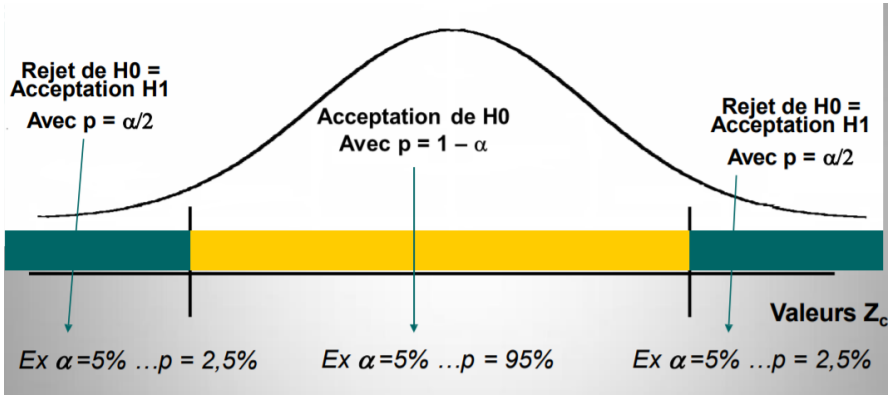
a) <u>Situation unilatérale</u>

Dans une situation **unilatérale**, le **rejet de H0** permet **uniquement** de dire qu'**il existe une différence significative** entre les deux situations.

Par exemple : si on teste l'efficacité de deux traitements A et B, le rejet de H0 en situation unilatérale permet uniquement de dire que l'efficacité de ces deux traitements est différente, mais **on ne saura pas lequel est le meilleur**.

b) Situation bilatérale

Au contraire, dans une situation **bilatérale**, le rejet de H0 permet de dire qu'il existe bien une différence entre les deux situations, mais on peut **aussi déterminer laquelle des deux est la meilleure**. Si l'on reprend l'exemple des traitements, le rejet de H0 permettra, en situation bilatérale, de déterminer lequel des deux traitements sera le plus efficace !



II. L'étude de la liaison entre deux caractères quantitatifs

/!\ Les formules à partir d'ici n'étaient pas à connaitre avant, cette année le Pr. Maignant dit que les formules ne sont toujours pas à connaitre, mais suppose que les formules « simples » (comme le calcul du Chi-2) sont connues.

Soient **2 groupes A et B** et une caractéristique **qualitative** x (couleur des yeux etc.). On se demande si le pourcentage d'individus du groupe A présentant x coïncide avec le pourcentage d'individus du groupe B présentant x.

1. Le test de comparaison des pourcentages (tout effectif)

Le paramètre Z est donné ici par l'écart réduit ε.

On va ainsi comparer :

- ε_t : donné par la **table de l'écart réduit** en fonction de **α**

- $\varepsilon_c = \frac{pA-pB}{\sqrt{\frac{pA*qA}{nA}+\frac{pB*qB}{nB}}}$ avec q=1-p

- Si $\varepsilon_{calculé} > \varepsilon_{théorique}$ → **Rejet de H0**

Comment trouver $\mathcal{E}_t$ sur la table de l'écart réduit ?

On cherche $\mathcal{E}_t$ en fonction du risque α.
On regarde les unités et les dizaine d'α sur les lignes et les centièmes sur les colonnes. ε_t se trouve à l'intersection de la ligne et la colonne

Ainsi, pour α = 5% = 0,05, on est à 0,0 pour la ligne et à 0,05 pour la colonne.

		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

Pour l'écart réduit, **si α = 5% alors ε_t = 1,96 ++++**

Exemple :

Soient 2 populations : la première où les enfants vont à la **crèche** et la deuxième où ils restent à la **maison**. On cherche à savoir si le mode de garde (crèche ou domicile) modifie le risque de rhinopharyngite des enfants.

On fait une étude sur 2 groupes de 200 enfants :

Crèche n_A=200 Nb rhino = 130

Domicile n_B=200 Nb rhino = 96

Le mode de garde influe-t-il sur le risque d'avoir une rhinopharyngite ?

1. H0 : pas de différence entre les 2 modes de garde vis-à-vis des rhinopharyngites
H1 : différence entre les 2 modes de garde

2. Caractère 1 : garde en crèche ou à domicile = qualitatif

Caractère 2 : avoir une rhinopharyngite ou non = qualitatif

3. α = 5% défini **à priori** donc $\varepsilon_{théorique}$ = 1,96 (on lit sur la table)

4. On calcule le paramètre (donné dans l'énoncé) $\varepsilon_{calculé}$ = 3,4

5. 3,4 > 1,96 : **on rejette H0 donc on accepte H1**

• Au niveau de **l'échantillon**, on peut en conclure que le risque de rhinopharyngites est supérieur chez les enfants gardés en crèche

• **On ne peut pas généraliser** cette conclusion au niveau de tous les enfants car il n'y a **pas eu TAS**

2. Le test du X^2 (tout effectif)

Le **paramètre Z** est donné ici par le **X^2**. On va donc comparer :

- **X^2_t** : donné par la **table du X^2** en **fonction d'α et du nombre de ddl**

- **$X^2_c = \frac{\sum (oi-ci)^2}{ci}$** **NEW : La formule est à connaitre !**

Cette formule permet de comparer les **chiffres calculés C** qui forment le modèle théorique, aux **chiffres observés O**

ddl = (nb lignes – 1) (nb colonnes – 1)

Le **test du X^2** permet de prendre en compte **tous les cas de figure** et **pas seulement deux %.**

Comment trouver X^2_t sur la table du X^2 ?

On cherche X^2_t en fonction du risque α et du nb de ddl.
Sur les lignes on trouve le nombre de ddl et sur la colonne le risque α. X^2_t se trouve à l'intersection des deux.

nb ddl = (nb lignes – 1) (nb colonnes – 1)
on ne compte pas les lignes et colonnes « total »

Ici par exemple, on cherche X^2_t pour α=5% et ddl=1

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 :
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 : pas de lien entre l'exposition au benzène et les leucémies
2. Variable 1 : leucémie ou non = qualitatif
   Variable 2 : exposition au benzène ou non = qualitatif
3. α = 5% défini à priori
   nb ddl = (2 – 1) (2 – 1) = 1 → $X^2_t = 3,841$
4. On calcule le paramètre (donné dans l'énoncé) $X^2_c = 1,42$
5. 1,42 < 3,84 donc on accepte H0

III. Liaison entre caractères qualitatifs et quantitatifs

Problématique : En moyenne la taille des individus d'une population A coïncide-telle avec la taille des individus d'une population B ?

1. Comparaison de moyenne : n1 et n2 > 30 "Grands échantillons"

On utilise : La table de l'écart réduit (ε<1,96 et α<5%). $\varepsilon=\frac{m1-m2}{\sqrt{\frac{s1^2}{n1}+\frac{s2^2}{n2}}}$

➔Si $\varepsilon calculé > \varepsilon théorique$ Rejet de H0

Exemple : On teste un antiviral diminuant le nombre de jours de symptômes cliniques chez des patients infectés par le virus de la grippe. Soit 100 sujets non traités, atteints de grippe. Le nombre moyen de jours avec symptômes est m1= 4,74 jours et l'écart-type : s1= 1. Soit 100 autres sujets traités avec l'antiviral et atteints de grippe le nombre moyen de jours avec symptômes est m2 = 4,2 jours et l'écart-type est s2=1,7.

→On fera ici un test de comparaison de moyenne qui nous permettra de répondre à la question : Peut-on accepter ou rejeter H0 ?

2. Série numérique : t de Student n1 ou n2 < 30 "Petits échantillons"

On utilise : La **table t de Student**, avec $(n_1-1)+(n_2-1)$ddl. $t=\frac{m1-m2}{\sqrt{\frac{s^2}{n1}+\frac{s^2}{n2}}}$

- Si $t_{calculé} > t_{théorique}$ → Rejet H0

Remarque : Ici on calcule un écart-type « s » sur les deux échantillons au lieu d'utiliser s1 et s2 (comme pour les comparaisons de moyennes). En effet ici s1 et s2 sont moins significatifs que pour les tests de comparaison de moyenne avec $s=\sqrt{\frac{\sum(xi-m1)^2+\sum(xj-m2)^2}{(n1-1)+(n2-1)}}$

Précision sur les ddl : On prend une série de 8 valeurs donc n=8 :

Toutes les valeurs :	2	3	5	12	10	4	7	8	Total=51
Avec 1 valeur manquante :	2	3	5	12	10		7	8	Total=47
Avec 2 valeurs manquante :	2	3		12	10		7	8	Total=42

S'il nous manque une donnée : mais que **nous connaissons le total**=51 **et le total sans la valeur** on peut retrouver la valeur manquante : 51-47=4.

Avec **n-1** valeurs **on peut calculer la valeur manquante** à partir du total.
Avec **n-2** valeurs il est **impossible de trouver les deux valeurs** manquantes.

Dans le test t de Student on compare 2 valeurs, donc **ddl=(n_1-1)+(n_2-1)**

Exemple : 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. Pour le groupe 1 : n1 = 15 ; m1 = 6,3 ; s1 = 1,8 et pour le groupe 2 : n2 = 12 ; m2 = 4,5 ; s2 = 1,6. L'obésité a-t-elle une influence sur le taux de corticoïdes ?

Méthode :

1. **On pose H0** : m1 et m2 ne sont pas différentes dans ces 2 groupes.
2. **Type de caractères étudiés** : Relation entre caractères qualitatifs (obèses et non obèses), et quantitatifs (valeurs de dosages sanguins, valeurs moyennes).
3. **Taille de l'échantillon** : n1=15 et n2=12 les deux sont < 30 c'est donc un petit échantillon. **Choix du test : test t de Student**.
4. **Ecart-type** : Ici on doit **calculer l'écart-type commun aux deux groupes** car on est en t de Student

Aparté : on ne vous demandera pas de calculer l'écart type, la formule est bien trop compliquée donc sa valeur sera donnée dans l'énoncé

S^2=2,53 ; ddl=(15-1)+(12-1)=25 ; t=2,92
On cherche donc t dans la table t de Student :

Ddl/α	0,9	0,5	0,3	0,2	0,1	0,05	0,02	0,01	0,001
25	0,127	0,684	1,058	1,316	1,708	2,06	2,485	2,787	3,725

t=2,92 ∈[2,787 ; 3,725]

→On a **α<1%** (après lecture dans la table) donc on peut **rejeter H0** et conclure à une relation entre obésité et augmentation du taux de corticoïdes **au niveau de ces échantillons** (on ne peut conclure que sur les échantillons et non pas sur la population générale car il n'y a pas eu de TAS).

3. Séries appariées ou méthode des couples

On utilise la méthode des couples lorsqu'on étudie la liaison entre deux variables **qualitatives et quantitatives** dans **2 échantillons non indépendants**.

Série indépendante : Lorsque les deux groupes comparés sont **distincts** et indépendants (**=sans lien**).

Exemple : On tire au sort un groupe 1 puis un groupe 2, G1 consommera un placebo et G2 le nouveau médicament à tester.

Série appariée : Lorsque les deux groupes comparés ne sont pas distincts et indépendants (**=liés**).

Exemple : On compare les résultats avant traitement puis après traitement : c'est donc une observation sur le même groupe : les groupe avant traitements et après traitements sont identiques. Ils ne sont pas indépendants car ils forment un seul et même groupe.

Test utilisés :

Si **n>30** $\varepsilon = m_d / \sqrt{\frac{s^2}{n}}$ avec un **test de comparaison de moyenne**.

Si **n<30** $t = m_d / \sqrt{\frac{s^2}{n}}$ avec un **test t de Student**

Exemple : On veut 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, md=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.

Autre exemple : 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

1) Quelle est la première précaution à prendre ? La consommation est-elle 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 consommations sont équivalentes dans les 2 groupes.

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

3. n> 30 →Test de comparaison de moyennes

4. ε=2,00 > 1,96 (1,96 est le ε pour α=5%).

On rejette H0, avec un risque a = 5%. Il existe donc une différence significative entre les consommations 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 traitement diffère-t-elle de la valeur avant traitement ? **Interpréter le résultat**.

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

4. ε=26,9 > 1,96 au risque α=5%.

On rejette H0. Il existe une différence très significative (p < 0,001) entre les consommations avant / après tt, dans le groupe P. Il y a surement eu un effet psychologique : envie de profiter de l'étude pour arrêter de fumer ?

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

1. H0 = les moyennes des consommations 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. ε=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 traitement? 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. ε=2,09 > 1,96 au risque de 5%

On rejette H0 : 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 (cf Question 3). Mais le Groupe S fumait plus (cf Question 1) → Efficacité du traitement S.

IV. Liaison entre caractères quantitatifs

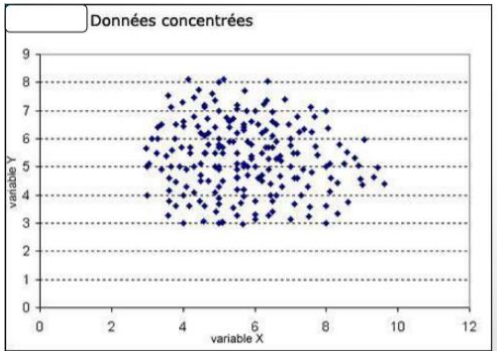
1. Corrélation et régression

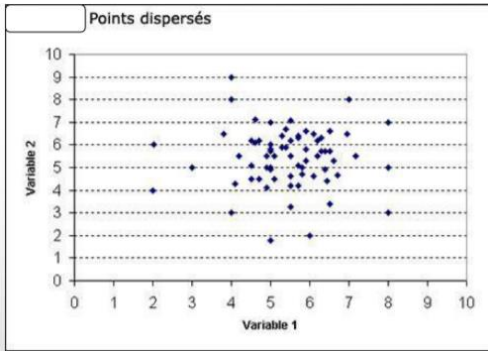
Corrélation = Evaluation de la liaison entre 2 variables quantitatives

Régression = Méthode mathématique expliquant les relations entre variables observées

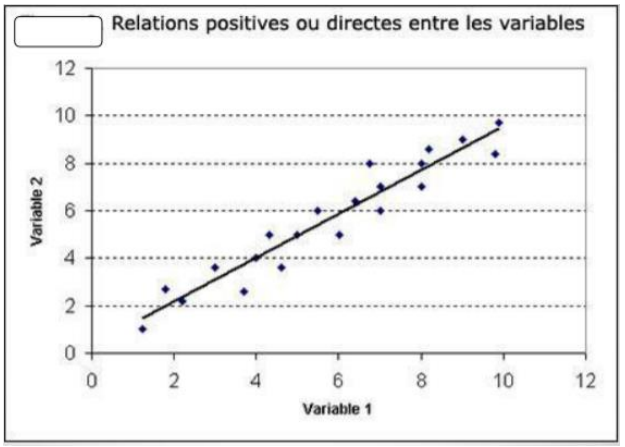
2. Représentation des données

Nuage de points :





Droite de régression : permet de visualiser si une des 2 variables est **dépendante** de l'autre



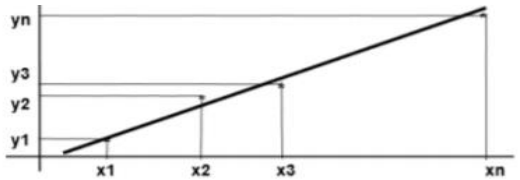
La **droite de régression** est aussi appelée **droite des moindres carrés** car elle passe au « plus près » de chaque point du graphe. *(Dans ce cours, on ne parle que de régression linéaire).* **Une droite de régression peut permettre de prédire certaines valeurs de y à partir d'une valeur de x**.

3. Etude de la liaison entre caractères quantitatifs

Exemple :

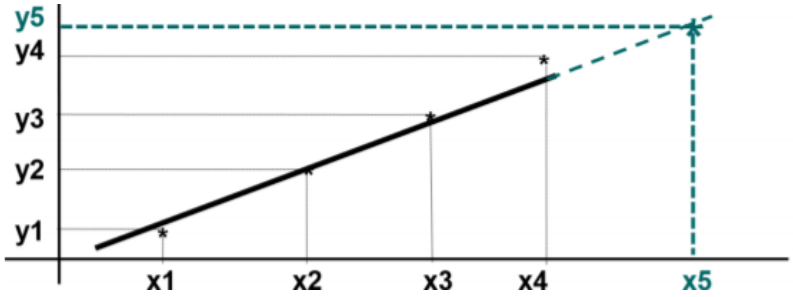
➢ La capacité respiratoire est-elle dépendante de la consommation de cigarettes ?

➢ 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) on a une droite de régression de y en x !



Remarque : On dit que y peut être peut-être « expliqué » en fonction de x.

Prédiction : La droite de régression permet de prédire des valeurs de y pour un certains x. *Il suffit pour cela de prolonger la droite ! ici on arrive à prédire le point en turquoise grâce à la droite.*



Coefficient de corrélation : pente de la droite

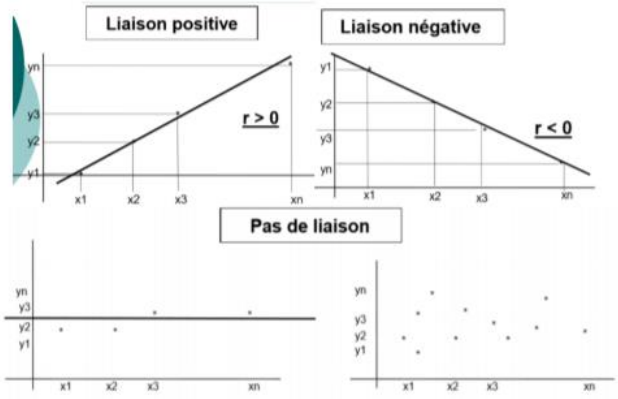
On utilise : La **Table du coefficient de corrélation**, avec **n-2 ddl**.

$$r = \frac{\sum(xi-mx)(yi-my)}{\sqrt{\sum(xi-mx)^2 \sum(yi-my)}} \quad r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{(\sum x^2 - \frac{(\sum x)^2}{n})(\sum y^2 - \frac{(Ey)^2}{n}}} \quad (osef)$$

→Si **r > 0** : liaison **positive** donc x et y varient dans le **même sens**.

→Si **r < 0** : liaison **négative** donc x et y varient en **sens inverse**.

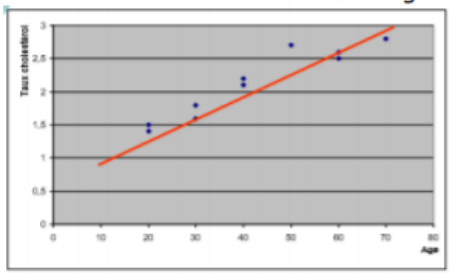
/!\ r est toujours inférieur à 1 /!



Exemple : 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-il de 2 séries totalement indépendantes ?

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

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

2. 2 variables quantitatives → Test du coefficient de corrélation

3. r calculé = 0,955 > r théorique= 0,76 avec 10-2 = 8 ddl on voit dans la table : (α = 1%)

Conclusion : **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. Cependant le résultat est non généralisable (**pas de TAS**).

/!\ Corrélation ≠ Causalité /!

Si on établit **corrélation** entre deux variables cela veut dire qu'**il existe un lien** entre les deux. Ex : l'âge et le cholestérol sont *liés* (on ne dit pas que l'un cause l'autre). Si on établit une **causalité** entre deux variables cela veut dire que **l'une est la cause de l'autre**. Ex : l'âge *cause* le cholestérol.

Exemple 2 : On teste un traitement favorisant la baisse de tension artérielle. On cherche à savoir si l'effet de ce traitement est lié à l'âge des patients. Dans ce but, on effectue 2 prises de tension : une avant traitement et une autre 1h après le traitement. On note la différence entre les 2 valeurs. Soient X la série des âges des patients et Y la série des différences de T.A avant – après traitement : la question posée peut se traduire par Y = f(X) ?

x = âge	y = diff	xy
64	1	64
52	9	468
56	9	504
60	6	360
62	2	124
58	4	232
57	3	171
61	1	61
58	6	348
55	9	495

$\sum x = 583 \quad \sum y = 50 \quad \sum xy = 2827$

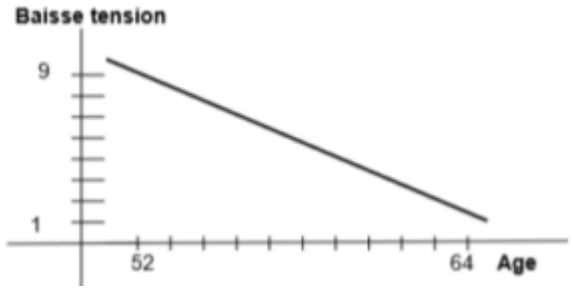
1. H0 : les 2 séries x et y sont indépendantes, et il n'existe pas de relation entre elles.

2. 2 variables quantitatives → Test du coefficient de corrélation

3. $r_{calculé} = |-0,83| > r_{théorique} = |0,76|$ au risque de 1%

NB : on compare les r théoriques et calculés en valeur absolue !

Conclusion : Nous pouvons **rejeter H0**. Il existe une **relation entre x et y** (elles sont « **corrélées** »), avec p < 1% r calculé -0,83 : le signe – indique que plus les valeurs d'une série augmentent, plus les valeurs de l'autre série diminuent. Elles varient en **sens inverse**.



V. Tests non paramétriques

Test paramétrique : test avec modèle à **forte contrainte**, difficile à réaliser sur de petits effectifs. (ex : t de Student ; Chi-2 ; comparaison de moyenne…)

Test non paramétrique : test dont le modèle ne précise pas les conditions que doivent remplir les paramètres de la population dont a été extrait l'échantillon. (ex : U Mann et Whitney ; r' de Spearman ; Wilcoxon)

On utilise obligatoirement un test non paramétrique si les effectifs sont trop faibles (4<X<12)

- Ils sont utilisés pour des variables quantitatives lorsque les effectifs sont trop faibles (les populations ne sont pas distribuées normalement).
- Ces tests présentent une excellente robustesse.

NEW : (*le prof a rajouté 2 diapos explicatifs sur U Mann et Whitney)*

Le test de **Wilcoxon-Mann-Whitney** (ou **test U de Mann-Whitney** ou encore **test de la somme des rangs de Wilcoxon**) est un test statistique **non paramétrique** qui permet de tester l'hypothèse selon laquelle les médianes de chacun de deux groupes de données sont proches.

-2 échantillons indépendants E1 et E2 de taille n1 et n2

On souhaite **tester l'hypothèse H0** disant que les moyennes expérimentales dans les 2 échantillons sont égales **µ1=µ2**

On **trie les valeurs** obtenues dans la réunion des 2 échantillons **par ordre croissant**. Pour chaque valeur xi issue de E1, on compte le nombre de valeurs issues de E2 situées après lui dans la liste ordonnée (celles qui sont égales à xi ne comptent que pour 1/2)

On note **u1 la somme des nombres** ainsi associés aux différentes valeurs issues de E1.

On fait de même en **échangeant les rôles des deux échantillons**, ce qui donne la somme **u2**. Soit **u la plus petite des deux sommes** obtenues : u = min{u1 ; u2}.

On note U la variable aléatoire associée.

Pour n1 et n2 quelconques, on lit dans les tables du test de Mann et Whitney le nombre mα tel que, **sous (H0), P(U ≤ mα) = α.**

On **rejette (H0)** au risque d'erreur α **si u ≤ mα**. Autrement on accepte (H0).

Si n1 et n2 sont assez grands (**≥20** en général), sous (H0), **U suit approximativement la loi normale N(μ,σ)**.

1. Les différents tests non paramétriques

	Test	
	Paramétriques	Non paramétriques
Comparaison de 2 échantillons indépendants	T de Student Comparaison de moyenne	Mann-Whitney
Comparaison de 2 échantillons appariés	T de Student Comparaison de moyenne	Wilcoxon
Test de corrélation	Coefficient r	Coefficient r' de Spearman

Principe :

Pour réaliser ces tests il est nécessaire de **transformer les données quantitatives en données de mesures ordinales** (les rangs). On prend les données quantitatives et on les ordonne.

Exemple : Si on étudie la variable quantitative « âge » on va ranger les âges des patients du plus jeune au plus âgé :

Ages	14	17	28	30
Rangs des âges	1	2	3	4

2. U Mann et Whitney 4<n<12

On utilise : La table de U Mann et Whitney (α<5%)

/!\ Si $U_{calculé} > U_{théorique}$ → Acceptation de H0 /!

Ces paramètres **résument les données et traduisent leur imbrication**. Si les données sont **très imbriquées** il n'y a **pas de différence significative** entre les 2 groupes. Si les données ne sont pas ou peu imbriquées il y a une différence significative entre les 2 groupes. On compare **le plus petit des deux U** (soit U_{AB} soit U_{BA}) à une valeur théorique lue dans la table (intersection de la ligne nA-nB et colonne n). Cet U théorique est la limite max au-delà de laquelle l'imbrication est considérée comme importante. C'est-à-dire que lorsque le U calculé est plus important que le U théorique (qui est cette limite) on peut **conclure qu'il y a imbrication et donc accepter H0** (si les données sont imbriquées c'est qu'elles sont issues d'un même ensemble).

Le test de Mann-Whitney permet de **tester si 2 groupes indépendants sont extraits d'une population unique** (2de possibilité) **ou non** (1ère possibilité).

Exemple : 1) On dispose de 2 groupes (groupe A de 6 malades, et groupe B de 5 sujets non malades). On dose une certaine hormone dans le sang de ces 11 sujets. Y a-t-il une différence significative entre ces 2 groupes du point de vue de cette hormone ?

Gr A 11 ; 21 ; 25 ; 52 ; 71 ; 79

Gr B 22 ; 43 ; 72 ; 92 ; 116

Nous constatons :

- Peu de valeurs à comparer
- Le test « ressemble » à la comparaison de moyennes car on compare une variable quantitative à une variable qualitative.
- Soit les dosages diffèrent en fonction du groupe malades/non malades, soit ils sont tous équivalents.

1. H0 : pas de différence significative entre les 2 groupes pour cette hormone.

2. Etude d'une liaison éventuelle entre données qualitatives (malades ou non), et données quantitatives (valeurs des dosages).

3. Comparaison de moyennes, effectif faible → U de Mann & Whitney

4. On range toutes les valeurs A ou B par ordre croissant :

	11	21	22	25	43	52	71	72	79	92	116
	A	A	B	A	B	A	A	B	A	B	B
U_{AB}	0	0	2	1	3	2	2	5	3	6	6

On va calculer le <u>paramètre U_{BA}</u> : pour chaque membre du groupe A on cumule le nombre de membres du groupe B qui sont classés avant lui. On peut calculer de même U_{AB}.

U_{BA} = 0 + 0 + 1 + 2 + 2 + 3 = 8 U_{AB} = 2 + 3 + 5 + 6 + 6 +6 = 22

Remarque : $U_{AB} + U_{BA} = nA \times nB = 6x5 = 30$

On compare le plus petit des deux U : soit U_{BA} à une valeur théorique lue dans la table (intersection de la ligne nA-nB =1 et colonne n = 5).

	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

→La valeur **U théorique = 3**

U_{BA} > 3 donc <u>l'imbrication des 2 groupes est considérée comme importante</u> = les données sont issues d'un **même ensemble** = il n'y a pas donc de différence significative entre les 2 groupes = **on accepte H0**

<u>Méthode :</u>

1. 1 ère possibilité : 2 groupes indépendants sont extraits d'une population unique. 2 ème possibilité : 2 groupes indépendants viennent d'une même population.

2. Type de caractères étudiés : Relation entre caractères qualitatifs (Groupe A ou B), et quantitatifs (dosage de l'hormone).

3. Taille de l'échantillon : n1=6 et n2=9 les deux sont < 12 Il faut donc un test non paramétrique. Choix du test : test U Mann et Whitney.

4. Calcul du paramètre et consultation de la table

3. r' de Spearman (corrélation)

On utilise : La table du r de Spearman (r'=0,89 avec α<5%, r'=1 avec α<1%)

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

Si $r_{calculé} > r_{théorique}$→Rejet de H0

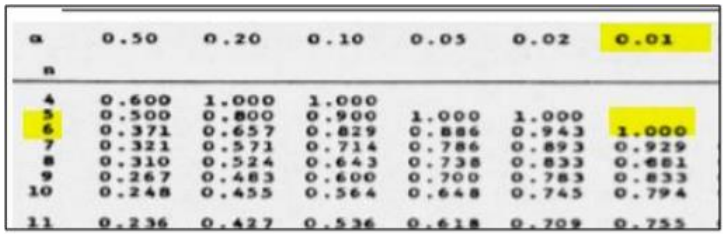
<u>Exemple :</u> On a recensé pour 6 étudiants les notes obtenues au concours PACES en Biostatistique, et le classement final à ce même concours. On cherche à établir s'il existe une relation entre cette note et le classement final.
<u>Rappel :</u> Variable Classement = variable « pseudo quantitative ».

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

1. H0 : Il n'y a pas de lien entre ces 2 séries de valeurs numériques. Il s'agit de 2 séries indépendantes.

2. Etude d'une liaison éventuelle entre données quantitatives (classement et note en Biostat') : coeff de corrélation mais avec de faibles effectifs → r' Spearman

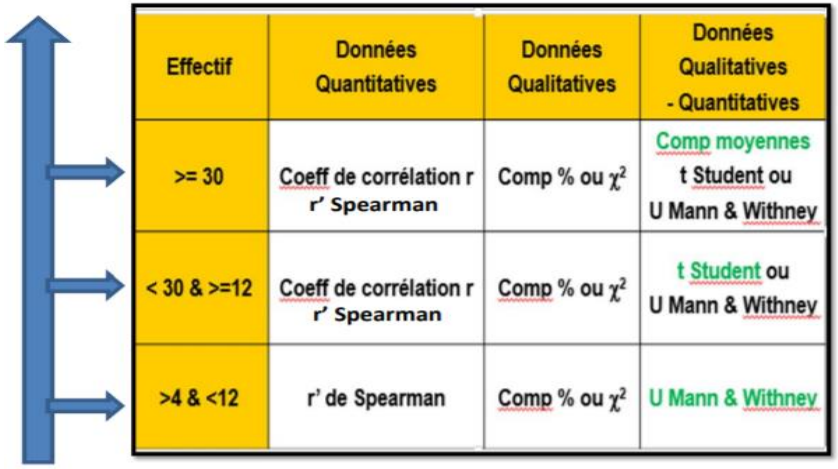
3. On calcule le r4 (vous n'aurez pas à le faire au cc) : r'calculé = - 1 On regarde dans la table du r' de Spearman l'intersection entre l'effectif et α :



n \ α	0.50	0.20	0.10	0.05	0.02	0.01
4	0.600	1.000	1.000			
5	0.500	0.800	0.900	1.000	1.000	
6	0.371	0.657	0.829	0.886	0.943	1.000
7	0.321	0.571	0.714	0.786	0.893	0.929
8	0.310	0.524	0.643	0.738	0.833	0.881
9	0.267	0.483	0.600	0.700	0.783	0.833
10	0.248	0.455	0.564	0.648	0.745	0.794
11	0.236	0.427	0.536	0.618	0.709	0.755

r'calculé =| - 1| = |r'théorique|, on repousse donc H0 (p<1%), on met en évidence un lien très significatif entre ces 2 séries. Il s'agit donc de 2 séries corrélées : plus la note de Biostat est élevée, plus petit est le rang de classement (d'où le signe – pour r').

VI. Récap tests+++



Effectif	Données Quantitatives	Données Qualitatives	Données Qualitatives - Quantitatives
>= 30	Coeff de corrélation r r' Spearman	Comp % ou χ^2	Comp moyennes t Student ou U Mann & Withney
< 30 & >=12	Coeff de corrélation r r' Spearman	Comp % ou χ^2	t Student ou U Mann & Withney
>4 & <12	r' de Spearman	Comp % ou χ^2	U Mann & Withney

NB : la flèche qui est orientée vers le haut indique que l'on peut utiliser le test pour des effectifs supérieurs. Attention l'inverse n'est pas vrai on ne peut pas utiliser de test pour des effectifs inférieurs !
NB bis : Le test écrit en vert est celui que l'on préfèrera utiliser.

Dédicaces :

- A mes cotuts d'amour qui sont un peu trop fans de cheddar <3 <3
- A mes vieilles, rien à dire, vous avez déjà compris que c'est les best
- A mes fillotes du turfu (surtout Audrey pour le soutien moral <3)
- Au KL
- A la team 12
- Au disco bus (meilleur bus), disco bungal (meilleur bungal) et disco balais (meilleur balais)
- A gre et leur table (RIP <3)
- Aux P1 de monteb (surtout le bâtiment Azur je vous kiff)
- A Tomy le meilleur danseur que la fac de médecine ait connu
- A Marine et Clara les meilleures <3
- A Léa qui me rappelle que je devrai commencer mes cours de P2
- Aux bg de la P1 : Diegz, Arthuros, Ines, Cha, Vic, Quentin, Alexis, Kevin
- La meilleure pour la fin : à Lisa qui ne lira surement jamais cette dédicace mais qui est tellement plus forte qu'elle croit, qui m'impressionne tout le temps et qui va sans doute jamais arrêter d'être la meilleure dans tout ce qu'elle fait