STATISTIQUES DEDUCTIVES

1. 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**.

a. Tests de comparaison

Le plus souvent, les tests utilisés en statistiques déductives sont des tests de **comparaison** entre **2 populations** présentant des caractères ou des 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.

b. 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** :

H0 = hypothèse nulle	H1 = hypothèse alternative
♥ Il n'y a **pas de différences** entre les 2 groupes. ♥ Il n'existe pas de lien entre les 2 caractères étudiés, et les fluctuations observées sont **dues au hasard.**	♥ Il y a une **différence significative** entre les 2 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.

c. Etapes 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** (qualitatives ou quantitatives) et choisir le bon test en fonction des données. Soit Z le paramètre qui sera calculé.
3. Choisir le **risque α à priori** (généralement **5%**)
4. Recueillir les données

5. Calculer Z
6. Utiliser la **règle de rejet/décision** (basée sur H0 et α) = examiner la position de cette valeur Z par rapport à un modèle théorique dont on connaît la distribution.
7. Fixer le **risque d'erreur réel à postériori**
8. **Interpréter** les résultats : au niveau de l'**échantillon** (est-ce qu'on accepte H0 ?) et au niveau de la **population** (est ce qu'on peut extrapoler ?)

d. Notion de risque

Rappel de statistiques descriptives : 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 statistiques déductives, on a :

- ♥ **α** ou **risque de première espèce** : le risque de rejeter H0 si H0 est vraie. C'est la probabilité de décider que le facteur de risque a un effet alors qu'il n'en a pas. Ce risque d'erreur est maîtrisé, c'est-à-dire qu'il est fixé (le plus souvent α = 5%) **AVANT** l'application du test statistique.

- ♥ **B** ou **risque de deuxième espèce** : le risque de rejeter H0 alors que H0 est faux. C'est la probabilité de ne pas détecter un effet du facteur de risque alors qu'il en existe un. Ce risque d'erreur est négligé, et peut être assez important. (En général β = 20%)

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

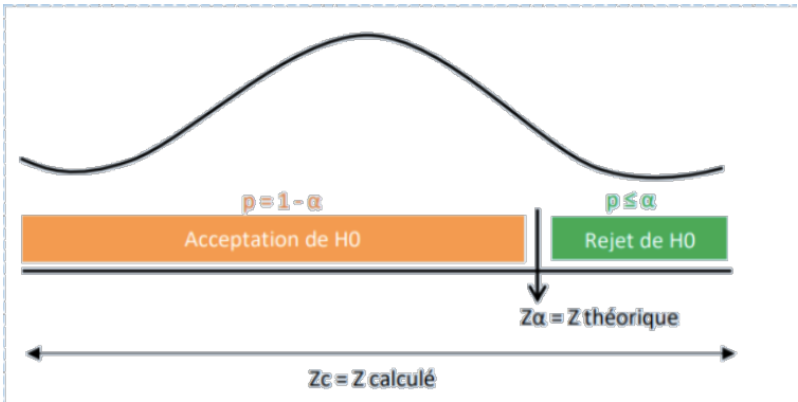
Remarque : On privilégie de maîtriser α quitte à ignorer β.
Remarque Bis : La règle de rejet du test est définie uniquement à partir de **α** et **H0**. Entre 2 alternatives, on choisira pour H0 l'hypothèse qu'il serait le plus grave de rejet à tort.

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

e. Interprétation graphique du risque α

Le paramètre $Z_{calculé}$ 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 doit :



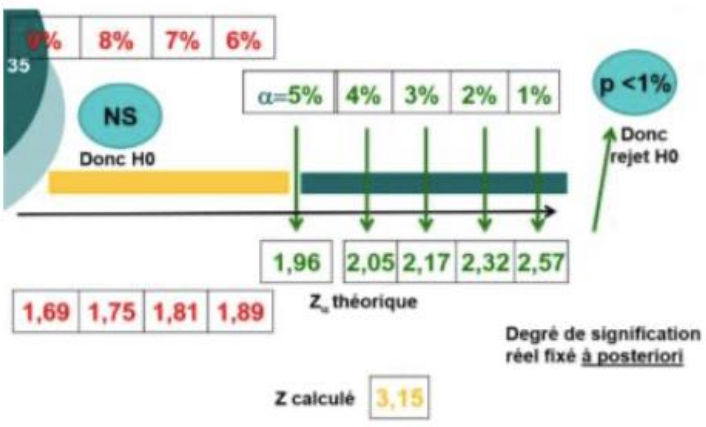
1. Fixer **α à PRIORI**
2. Chercher le **Zthéorique** (= Zt) sur la table (*cf. plus loin pour le trouver*)
3. Calculer **Zcalculé** (= Zc) grâce aux formules
4. **Comparer Zc avec Zt**, et on peut arriver à 2 conclusions différentes :

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

5. **<u>Fixer le degré de signification p à POSTERIORI</u>**

Remarque : **α** est fixé **à priori** par le statisticien (= supposition) mais le **degré de signification** est fixé **à postériori** (= réel) car la précision de l'étude peut s'avérer être supérieure à celle supposée au départ.

Le **degré de signification** est appelé **probabilité p**. Son calcul exact permet de préciser le risque potentiel d'erreur qui accompagne le rejet de l'hypothèse nulle : si p vaut 0,01, on dit que la différence est significative à 1%.



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.

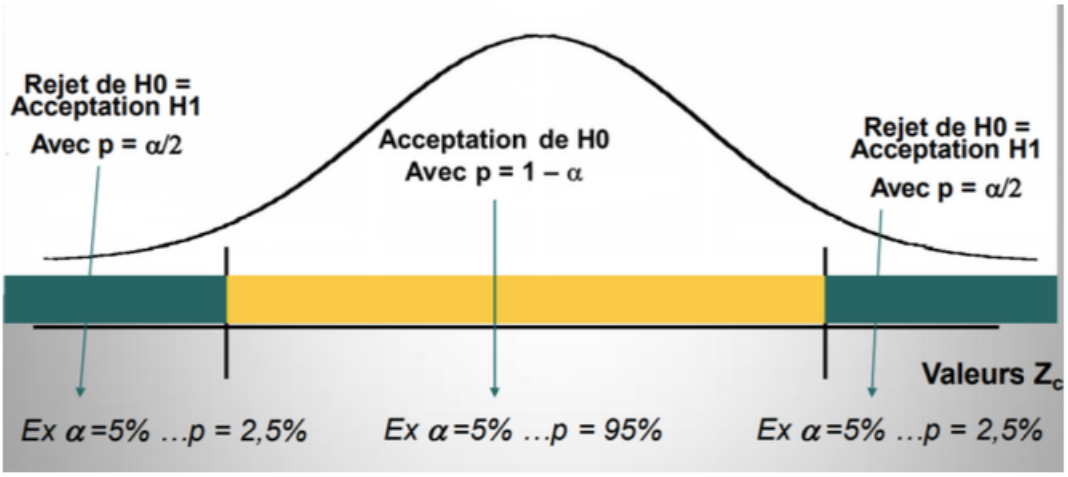
On a 2 façons d'interpréter ce risque :

♠ **Situation unilatérale**

Dans une situation unilatérale, le rejet de H0 permet uniquement de dire qu'il existe une différence significative entre les 2 situations.
Par exemple, on test l'efficacité entre deux traitements A et B, le rejet de H0 permet de dire qu'il existe une différence significative d'efficacité entre les deux traitements. **Cependant, on ne peut pas dire lequel est le meilleur.**

♠ **Situation bilatérale**

Au contraire, dans une situation bilatérale, le rejet de H0 permet de dire qu'il existe bien une différence entre les 2 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 2 traitements sera le plus efficace.



f. Rôle des Big Data

♠ Et si les données étaient le pétrole du 21ème siècle ?

Nous générons et détenons des quantités **d'informations personnelles** : alimentation, achats, contributions sur les réseaux sociaux, goûts, préférences, recherches sur Google, santé connectée, …
Les **données** sont **éparses** mais **captées** par différents intervenants sur Internet.

Dans le **domaine de la santé**, des **études épidémiologiques** diverses sont lancées (pour le meilleur, et pour le pire aussi ?) : aux USA, les sociétés privées analysent ces données et en tirent des conclusions. Par exemple, on propose à des femmes l'ablation des 2 seins car leur profil génétique est comparé à celui de milliers d'autres femmes, permettant de repérer celles à risque accru de cancer du sein.

Les **objets connectés** (bracelets, balances, t-shirts, fauteuils, iwatch, …) permettent de suivre sa propre forme physique, et de la comparer à ce qu'elle devrait être (mais quelles sont les

normes et qui les définit ?). Mais ils sont aussi la source d'alimentation de manière continue de ces fameuses **Big Data**, comme par exemple les GAFA (référence aux 4 grands géants américains d'Internet : Google, Apple, Facebook, Amazon).

♠ L'utilisation de ces masses de données remet en cause certaines **théories statistiques** et la notion d'**échantillonnage**.

Jusqu'à aujourd'hui, les **données** recueillies dans les **études cliniques** sont des données **démographiques** (sexe, âge), **cliniques** (poids, taille, diagnostique, traitement, dose, durée), **biologiques**, etc… mais jamais de données de type **psychologique** ou **émotionnel**.

Les Big Data vont être une révolution pour les statistiques. On va passer de statistiques inductives à des **statistiques descriptives**. En effet, les Big Data vont permettre de **recouper** et d'**analyser** tous ces types de données et de **remettre en cause** certaines conclusions ou décisions.

De plus, **l'échantillon traditionnel** permettait de sélectionner **uniquement un effectif de quelques dizaines**, au mieux quelques centaines d'individus, supposés représenter des **populations cibles** souvent de plusieurs **centaines de milliers d'individus**.

Grâce aux **Big Data**, l'**effectif de l'échantillon** observé et étudié est **de l'ordre de la population cible**. Et ça, c'est tout de même un vrai bouleversement théorique !

Si la taille de l'échantillon se rapproche désormais de la taille de la population, à terme, cela signifierait que l'on n'aurait plus de problème de nombre de sujets nécessaires à calculer.

2. Etude de la liaison entre 2 caractères qualitatifs

ATTENTION : A partir d'ici les formules ne sont pas à connaître sauf les calculs « simples » comme le Chi-2.

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

a. 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 (le petit « t » veut dire théorique)
-$\varepsilon_c = \frac{pA-Pb}{\sqrt{\frac{pA.qA}{nA}+\frac{pB.qB}{nB}}}$ avec q = 1-p (le petit « c » veut dire calculé)

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

Comment trouver ε_t sur la table de l'écart réduit ?

On cherche ε_t en fonction du risque α.
On regarde les **unités** et les **dizaines** d'α sur les lignes et les **centièmes** sur les colonnes.
εt se trouve à l'intersection de la ligne et de la colonne.
Ainsi, pour α=5%=0,05, on est à **0,0** pour la ligne et à **0,05** pour la colonne → 1.96

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

Rappel : pour α = 5%, ε = 1.96 et en l'occurrence 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 étudie 2 groupes de 200 enfants : Crèche nA = 200 ; Nb rhino = 130 / Domicile nB = 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 ε_t = **1,96** (on le lit sur la table)*
4. *On calcule le paramètre (donné dans l'énoncé) $\varepsilon_{calculé}$ = 3,4*
5. *3,4 > 1,96 donc $\varepsilon_{calculé} > \varepsilon_t$: **on rejette H0** et **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 (p<0,001 défini à posteriori).*
*• On ne peut **pas généraliser** cette conclusion au niveau de tous les enfants car il n'y a **pas eu de TAS** +++*

b. 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** (nombre minimale 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).
- **$X^2_c = \frac{\Sigma(oi-ci)^2}{ci}$ +++++ à connaître**

Cette formule permet de comparer les chiffres calculés C qui forment le modèle théorique aux chiffres observés O. (Donc en gros « oi » = données observées et « ci » = données calculées)

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

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

On cherche X^2_t en fonction d'α et du nombre de DDL
On cherche le nombre de DDL sur les LIGNES et α sur les COLONNES. +++
Exemple : on cherche le 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 entraîner 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

4. On calcule le paramètre (donné dans l'énoncé) X2c = 1,42
5. 1,42 < 3,84 donc **on accepte H0** : il n'existe pas de relation entre l'exposition au benzène et l'apparition des leucémies. (Le « 3.84 » vient de la table du X^2)

3. Etude de la liaison entre caractères quantitatifs et qualitatifs

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

a. Comparaison de moyennes : 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 $\varepsilon_c > \varepsilon_t$ → 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 moyennes*** *qui nous permettra de répondre à la question : Peut-on accepter ou rejeter H0 ?*

b. Série numérique : T de student : N1 ou N2 < 30 (petits échantillons)

On utilise la **table t de Student** avec **(n1-1) + (n2-1) DDL :** $t = \frac{m1-m2}{\sqrt{\frac{s1^2}{n1}+\frac{s2^2}{n2}}}$

Si $t_{calculé} > t_{théorique}$ → rejet de 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 comparaisons de moyennes avec $s = \sqrt{\frac{\Sigma(xi-m1)^2+\Sigma(xj-m2)^2}{(n1-1)+(n2-1)}}$

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

♠ 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 autres valeurs** manquantes.

Dans le test t de Student, on compare 2 valeurs, donc : DDL = (n1 – 1) + (n2-1)

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 inférieurs à 30 c'est donc un petit échantillon. Choix du test : test t de Student.*
4. ***Écart-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 vous sera toujours 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.**
Attention : on ne peut pas généraliser ce test à toute la population car il n'y a pas eu de tirage au sort. On conclura seulement sur les échantillons.

c. 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. G1consommera 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 groupes avant traitement et après traitement sont identiques. Ils ne sont pas indépendants car ils forment un seul et même groupe.

- ♠ Si n > 30, on utilise le test de comparaison de moyennes : $\varepsilon = \frac{m_d}{\sqrt{\frac{s^2}{n}}}$
- ♠ Si n < 30, on utilise un test t de Student : $t = \frac{m_d}{\sqrt{\frac{s^2}{n}}}$

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 : $\varepsilon = \frac{m_d}{\sqrt{\frac{s^2}{n}}}$ et si n < 30 : $t = \frac{m_d}{\sqrt{\frac{s^2}{n}}}$

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 TAS 2 groupes de 40 fumeurs. L'un reçoit la substance S, l'autre reçoit un placebo P. Le traitement dure 2 mois pour les 2 groupes. La consommation de cigarettes par jour (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 va falloir en tenir compte lors des conclusions.

Comparaison des consommations moyennes avant traitement dans les 2 groupes :

- H0 = Les moyennes des consommations sont équivalentes dans les 2 groupes.
- Étude de la liaison entre variables **qualitatives** (S ou P) et **quantitatives** (nb de cigarettes par jour) dans des échantillons **indépendants**.
- n>30 donc on utilise le **test de comparaison de moyennes**.
- ε=2,00 > 1,96 (1,96 est le ε pour α = 5%)

On **rejette H0** avec un risque α = 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 consommation avant et après traitement.

2. **Dans le groupe Placebo, la consommation moyenne après traitement diffère-t-elle de la valeur avant traitement ? Interpréter le résultat.**

- Liaison entre une variable **qualitative** (avant et après traitement) et **quantitative** (nb de cigarettes par jour).
- Les échantillons sont **non indépendants** donc on va utiliser la **méthode des couples**.
- n>30 donc on va utiliser le **test de comparaison de moyennes**.
- 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 et après traitement dans le groupe P. Il y a sûrement 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 ?

- ♠ H0 = Les moyennes des consommations sont équivalentes dans les 2 groupes.
- ♠ Liaison entre variables **qualitatives** (S ou P) et **quantitatives** (nb de cigarettes par jour) dans 2 échantillons **indépendants**.
- ♠ n>30 donc on utilise le **test de comparaison de moyennes**.
- ♠ ε=1,42 < 1,96.

On **accepte H0**. Il n'existe **pas de différence significative** entre les 2 groupes pour la consommation après traitement.

4. Les 2 groupes diffèrent-ils pour la variation de consommation avant et après traitement ?

Il faut comparer les variations avant et après traitement dans les 2 groupes afin de prouver l'intérêt de la substance S.

- ♠ H0 =Il n'existe pas de différence entre les variations de consommation dans les 2 groupes.
- ♠ On étudie la liaison entre une variable **qualitative** (S ou P) et une variable **quantitative** (nombre de cigarettes par jour) dans 2 échantillons **indépendants**.
- ♠ n>30 donc on va utiliser le **test de comparaison de moyennes**.
- ♠ ε=2,09 > 1,96 au risque α = 5%.

On **rejette H0**. Il existe une **différence significative** entre les variations de consommation dans les 2 groupes (p<5%).

Conclusion : efficacité de S. Il y avait eu un **TAS**, donc le résultat est **généralisable**.

Conclusion générale : pas de différence après traitement dans chaque groupe (cf question 3), mais le groupe S fumait plus (cf question 1) donc efficacité du traitement S.

4. Etude de la liaison entre caractères quantitatifs

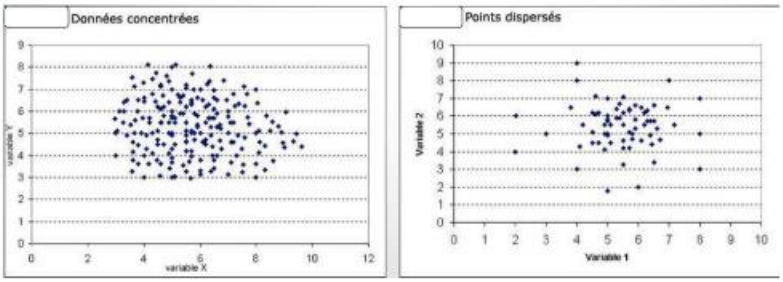
a. Corrélation et régression

Corrélation : évaluation de la liaison entre 2 variables quantitatives

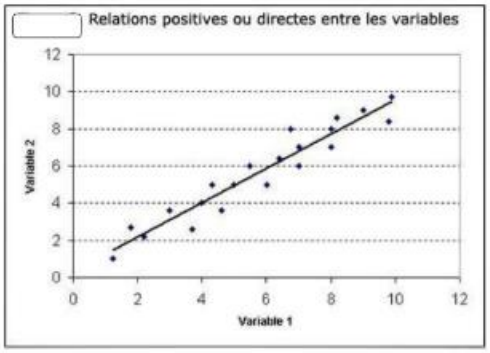
Régression : méthode mathématique expliquant les relations entre variables observées

b. Représentation des données

Nuage de points :



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

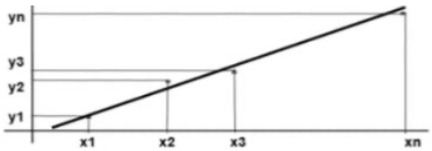


La droite de régression est aussi appelée **droite des moindres carrés**, car **elle passe au plus près de chaque point du graphe**. (Dans ce cours, on ne parle que de régression linéaire).
Une droite de régression peut permettre de **prédire certaines valeurs de y à partir d'une valeur x**.

c. Etude de la liaison entre caractères quantitatifs

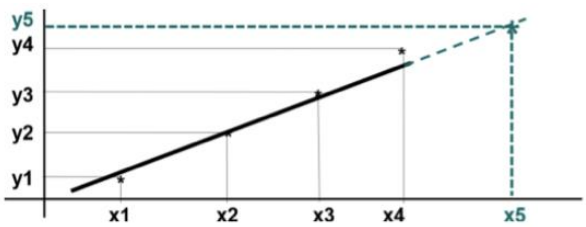
Exemple

- 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 sont liés, alors y=f(x) et on a une **droite de régression de y en x**.



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

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

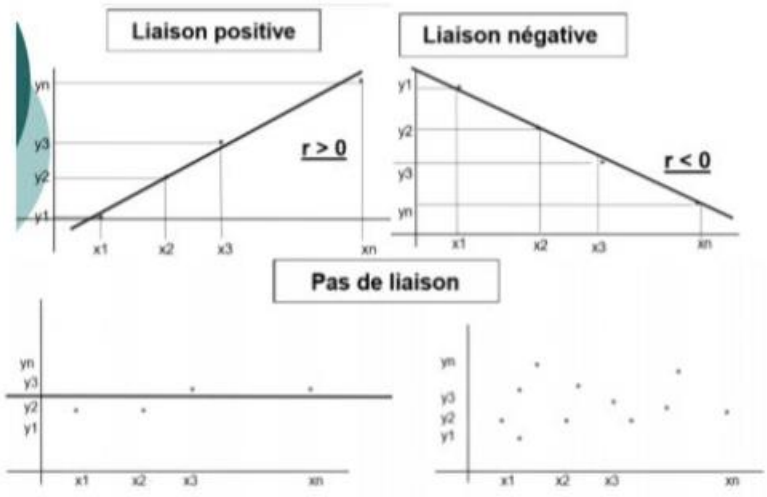


Coefficient de corrélation = pente de la courbe

On utilise la **table du coefficient de corrélation** avec : DDL = n-2

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

♠ 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**

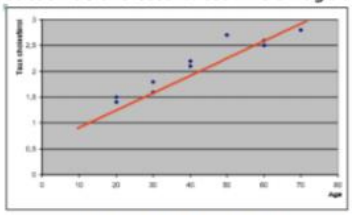


⚠ **ATTENTION** : r est toujours inférieur à 1 ⚠

Exemple : Sur un échantillon de 10 sujets d'âge 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. *On a 2 variables **quantitatives**, donc on va utiliser le **test du coefficient de corrélation**.*
3. *rcalculé = 0,955 > rthéorique = 0,76 avec DDL = 10-2 = 8 pour 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 n'est pas généralisable car il n'y a pas eu de TAS.*

⚠ **Attention** : corrélation ≠ causalité ⚠

♠ Si on établit une **corrélation** entre 2 variables cela veut dire qu'il existe un **lien** entre les 2 (*ex : l'âge et le cholestérol sont* liés → on ne dit pas que l'un cause l'autre +++).
♠ Si on établit une **causalité** entre 2 variables, cela veut dire que l'une est la **cause** de l'autre *(ex : l'âge cause le cholestérol*).

Autre exemple : 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 tension artérielle avant et après traitement : la question posée peut se traduire par y=f(x) ?

1. H0 = Les 2 séries x e y sont indépendantes, et il n'existe pas de relation entre elles.
2. On a 2 variables **quantitatives** donc on utilise le **test du coefficient de corrélation**.
3. $r_{calculé}$ = |-0,83| > $r_{théorique}$ = |0,76| au risque de 1%

Si $r_{calculé} > r_{théorique}$ → rejet de H0
⚠ On compare les r théoriques et calculés en valeur absolue ++

Conclusion : **Rejet de H0**. Il existe une **relation entre x et y** (elles sont « corrélées »), avec p<1%, rcalculé = -0,83, le signe moins indiquant que plus les valeurs d'une série augmentent, plus les valeurs de l'autre série diminuent. Elles varient donc en **sens inverse**.

5. Tests non paramétriques

Test paramétrique = test avec modèle à **forte contrainte**, car il n'est fiable que si les données associées à chaque échantillon suivent une distribution selon une loi normale. Il est **difficile à réaliser sur de petits effectifs.**
(ex : test t de Student, test du Chi-2, test de comparaison de moyennes, …)

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 : test de U Mann et Whitney, test r' de Spearman, test de Wilcoxon, …)
Il est beaucoup plus **robuste** car il ne se base pas sur des distributions statistiques.

Important : On utilise obligatoirement un test non paramétrique lorsque les effectifs sont faibles (4<n<12) ++++

Les tests non paramétriques 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**.

a. Les différents tests non paramétriques

	Test	
	Paramétriques	Non paramétrique
Comparaison de 2 échantillons indépendants	♠ **Test t de Student** ♠ **Test de comparaison de moyennes**	♠ **Test de Mann-Whitney**
Comparaison de 2 échantillons appariés	♠ **Test t de Student pour séries appariées** ♠ **Test de comparaison de moyennes pour séries appariées**	♠ **Test de Wilcoxon**
Test de corrélation	♠ **Test du coefficient r**	♠ **Test du 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 :

Âges	14	17	28	30
Rang	1	2	3	4

b. U-MANN et WHITNEY 4<N<12

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 2 groupes de données sont proches**.

On est en présence de 2 échantillons indépendants E1 et E2 de taille n1 et n2.

On souhaite **tester l'hypothèse H0** selon laquelle 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.). *Voir exemple plus loin pour comprendre toute cette partie (et celle à venir). Mais en soi, on peut compter les*

valeurs après ou avant : ça revient au même car on choisira la plus petite valeur entre les deux u (vraiment, regarde l'exemple plus bas pour comprendre).

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

On fait de même en **échangeant les rôles des 2 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).

☞ Avec ce test, 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.

On compare le **plus petit des deux U** (soit UAB soit UBA) à une valeur théorique lue dans la table (à l'intersection de la ligne nA-nB et de la colonne n).
Cet U théorique est la **limite maximale 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 une **imbrication** et donc **accepter H0** (→ si les données sont imbriquées c'est qu'elles sont issues d'un même ensemble).
Rappel : H0 = Les moyennes des 2 échantillons est égale.

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

Exemple : On dispose de 2 groupes (groupe 1 de 6 malades et groupe B de 5 sujets non malades). On dose une certaine hormone dans le sang de ces 11 sujets. La question que l'on se pose est : Il y a-t-il une différence significative entre ces 2 groupes du point de vue de cette hormone ?

Groupe A : 11 ; 21 ; 25 ; 52 ; 71 ; 79
Groupe B : 22 ; 43 ; 72 ; 92 ; 116

*On constate qu'il y a peu de valeurs à comparer, et que le test « ressemble » au test de comparaison de moyennes, car on compare une variable **quantitative** (valeur du dosage de l'hormone) à une variable **qualitative** (malade ou sain).*
Soit les dosages diffèrent en fonction du groupe malades/non malades, soit ils sont tous équivalents.

1. *H0 = Il n'y a pas de différence significative entre les 2 groupes pour cette hormone.*
2. *On étudie une liaison éventuelle entre des données **qualitatives/quantitatives**.*
3. *On est en présence d'un **faible effectif**, donc on va s'orienter vers le choix du **test de U Mann & Whitney**.*
4. *On **range** toutes les valeurs issues du groupe A ou du groupe 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 paramètre UBA : pour chaque membre du groupe A, on cumule le nombre de membres du groupe B qui sont classés avant lui. De même, on calcule U_{AB}.

UBA = 0 + 0 + 1 + 2 + 2 + 3 = **8** et UAB = 2 + 3 + 5 + 6 + 6 + 6 = **22**

Remarque : UAB + UBA = nA x nB = 6x5 = 30

On compare le **plus petit des deux U** (soit UBA) à une valeur théorique lue dans la table (à l'intersection entre la ligne nA-nB = 6-5 = 1 et la colonne n = 5).

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

→ La valeur **U théorique = 3**

*UBA > 3 donc **l'imbrication de ces 2 groupes est considérée comme importante**. Les données sont issues d'un **même ensemble**, c'est-à-dire que les moyennes des valeurs du dosage de l'hormone sont identiques chez les malades et les non malades, donc il n'y a pas de différence significative entre les 2 groupes = on **accepte H0**.*

Méthode :

1. 1ère possibilité = 2 groupes indépendants sont extraits d'une population unique.
   2ème possibilité = 2 groupes indépendants viennent de deux populations distinctes.
2. Type de caractères étudiés : **qualitatifs** (*groupe A ou B*) et **quantitatifs** (*dosage de l'hormone*)
3. Taille de l'échantillon : *nA=6* et *nB=5*, les deux sont **<12**. Il faut donc un **test non paramétrique**. Choix du test : **test de U Mann et Whitney**.
4. Calcul du paramètre et consultation de la table.

c. R' de Spearman

On utilise la **table du r de Spearman** : $r' = 1 - \frac{6\Sigma di^2}{n(n^2-1)}$

Si $r'_{calculé} > r'_{théorique}$ → rejet de H0

Exemple : On recense pour 6 étudiants les notes obtenues au concours PACES en Biostatistiques, et le classement final à ce même concours. On cherche à établir s'il existe une relation entre cette note et le classement final.

Rappel : La variable « classement » est une variable pseudo-quantitative.

X Biostat	12.4	4.9	18.1	5.4	19.4	16
Y classement	210	555	6	445	5	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. Étude d'une liaison éventuelle entre des données **quantitatives** (classement et note en Biostat) → on est dans le cas du test de corrélation, mais avec de faibles effectifs on va utiliser le **test r' de Spearman**.
3. On calcule le r' (il vous sera donné dans l'énoncé, vous n'aurez pas à le calculer) : $r'_{calculé} = -1$.
4. On regarde dans la table du r' de Spearman l'intersection entre l'**effectif** et le **risque** α.

n \ α	0.2	0.1	0.05	0.02	0.01	0.002
4	1.000	1.000	–	–	–	–
5	0.800	0.900	1.000	1.000	–	–
6	0.657	0.829	0.886	0.943	1.000	–
7	0.571	0.714	0.786	0.893	0.929	1.000
8	0.524	0.643	0.738	0.833	0.881	0.952
9	0.483	0.600	0.700	0.783	0.833	0.917
10	0.455	0.564	0.648	0.745	0.794	0.879
11	0.427	0.536	0.618	0.709	0.755	0.845
12	0.406	0.503	0.587	0.678	0.727	0.818
13	0.385	0.484	0.560	0.648	0.703	0.791
14	0.367	0.464	0.538	0.626	0.679	0.771
15	0.354	0.446	0.521	0.604	0.654	0.750
16	0.341	0.429	0.503	0.582	0.635	0.729
17	0.328	0.414	0.488	0.566	0.618	0.711

$r'_{calculé} = -1 = r'_{théorique}$ au risque α = 1%. On **rejette donc H0** (p<1%), et 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 moins pour r').

6. Récap des tests

Effectif	Données quantitatives	Données qualitatives	Données qualitatives et quantitatives
n ≥ 30	• **Coefficient de corrélation r** • r' de Spearman	• Comparaison de pourcentages • Chi-2	• **Comparaison de moyennes** • T de Student • U Mann et Whitney
30 > n ≥ 12	• **Coefficient de corrélation r** • r' de Spearman	• Comparaison de pourcentages • Chi-2	• **T de Student** • U Mann et Whitney
12 > n > 4	• **r' de Spearman**	• Comparaison de pourcentages • Chi-2	• U Mann et Whitney

On peut utiliser un test **pour des effectifs supérieurs** (*ex : r' de Spearman pour un effectif supérieur à 30*), mais l'inverse n'est pas vrai : on ne peut pas utiliser un test pour des effectifs inférieurs. +++

Le test écrit en gras est le test qu'on préfèrera utiliser. ++

Remarque : En fait, pour le choix du test le plus approprié, on prend en compte plusieurs critères (*dont le prof ne parle pas cette année, par souci de simplification*) et l'effectif n'en est qu'un parmi tant d'autres. Donc des fois, il se peut que vous voyiez le prof utiliser, par exemple, un test T de Student malgré un effectif à n=10 et ce n'est pas faux ! Sa décision repose sur un tas d'autres paramètres que les statisticiens prennent en compte, mais qui ne sont pas à votre programme. *On ne vous demandera rien que vous ne puissiez pas deviner, rassurez-vous.*

FIIIIIIIIIN

Et voilà fin de ce cours très long mais super cool ! Cours super important avec en moyenne 4 QRU le jour de l'épreuve ! Presque toutes les valeurs théoriques et calculées vous seront données dans l'énoncé (hormis le khi²) donc no stress ! Il faut comprendre le cheminement et surtout comment interpréter les résultats !
On vous souhaite tout pleins de courage et la biostat veille sur vous <3

Ma seule dédi est pour vous les champions ! Déchirez tout !!!!!!!!!!!!!

Bosse comme un boss pour être un boss, j'ai toutes les ~~options~~ connaissances dans le ~~gamos~~ cerveau !