Pr. Maignant

Biostatistiques

STATISTIQUES DÉDUCTIVES

Dulclaudiax

Sommaire

Voici la version complète de ce cours :)
J'ai corrigé quelques points et les principaux rajouts concernent les Big Data et la méthode des couples (séries appariées).
Je vous ai tout mis si ce n'est quelques exemples supplémentaires que vous pouvez aller voir dans le diapo du prof (je me suis dit qu'un exemple par test suffirait sinon ce serait trop long, et puis la logique reste la même).
De plus, pour chaque exemple le prof vous met à chaque fois la table dont est issu le paramètre théorique : je ne les ai pas toutes mises étant donné que je vous ai fait des points méthodo pour les trouver, mais si vous voulez vous entrainer à trouver Zt dans la table allez check le diapo :)
Bonne lecture et bon courage !

TESTS D'HYPOTHÈSES

Généralités

Le but des statistiques déductives est de **tirer des conclusions à partir d'observations** *(qu'on aura tiré grâce aux statistiques descriptives)*. Pour cela, on va la majorité du temps comparer 2 groupes pour un caractère donné. *Exemple : pour comparer les notes à l'épreuve de biostats entre 2 années, on se pose la question : y a-t-il une différence entre les 2 groupes ?*

Définition des hypothèses

En statistiques déductives, on travaille à partir de deux hypothèses :

- **Hypothèse H0 ou hypothèse nulle**
  - Il n'y a pas de différence entre les 2 groupes → les fluctuations sont donc dues au hasard
- **Hypothèse H1 ou hypothèse alternative**
  - Il existe une différence significative entre les deux groupes → les fluctuations observées ne sont pas dues au hasard

Définition des tests

Les tests sont des techniques permettant de décider **si on garde ou repousse H0**, en ayant fixé le **risque d'erreur** (α) accompagnant cette décision.

Etapes d'un test d'hypothèses

Théorie

- **Définir H0 et H1** avant le recueil des données. Les deux hypothèses jouent des rôles symétriques
- **Choisir le test** en fonction du nombre et du type de données (qualitatives, quantitatives). Le paramètre qu'on calculera sera **Z**.
- **Choisir le risque α** (en pratique souvent 5%)

Pratique

- **Recueillir les données**
- **Calculer le paramètre Z**

Conclusion

- Utiliser la règle de décision : on **compare** le paramètre Z calculé $\mathbf{Z_c}$ à un paramètre Z théorique $\mathbf{Z_t}$ dont on connait la distribution. Selon le test on **rejette ou accepte H0**
- Fixer le **risque d'erreur réel** attaché à la conclusion (à postériori)
- **Interpréter** les résultats, au niveau statistique puis médical

Notion de risque

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

Le compromis universel est que **α = 5%** *(donc si on ne vous donne pas α, estimez qu'il est égal à 0,05)*.

La **règle du rejet du test** est définie seulement à partie de **α** et de **H0**. Entre 2 alternatives, on choisira pour **H0** l'hypothèse qu'il serait le plus grave de rejeter à tort.

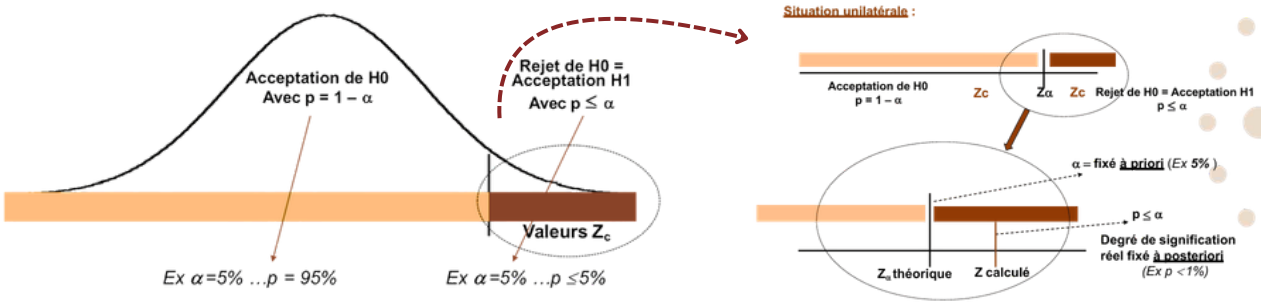
Réalité \ Décision	Rejet H0	Non rejet H0
H0 vraie	α	1 – α
H1 vraie	1 – β	β

Interprétation graphique

Le paramètre calculé Z_c , résultat du test, suit une distribution probabiliste en forme de **courbe de Gauss**. Soit α (risque de 1ère espèce) choisi = 5%. *Les courbes suivantes illustrent donc comment on interprète graphiquement la décision de notre test : en beige → on accepte H0; en marron → on rejette H0.*

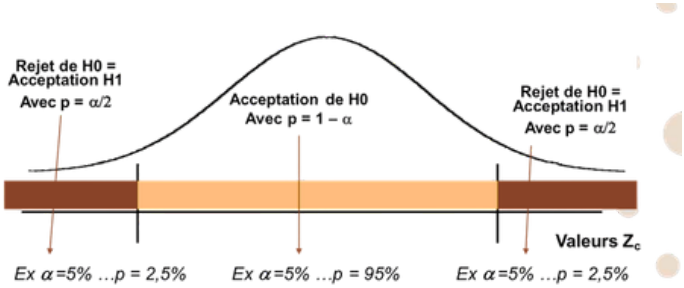
Situation unilatérale

Il y a une seule zone critique, on sait simplement que si Z_c est dedans, on **rejette H0 (= on accepte H1).**



Situation bilatérale

Les zones critiques sont symétriques et nous permettent de dire en plus quelle situation est la **meilleure**.



Point'tut : la situation bilatérale

La situation bilatérale (selon le Pr. Maignant) désigne une situation où on observe des différences significatives entre le groupe étudié et le groupe témoin, mais où on peut également dire si cette différence est **positive ou négative**.

Comme on peut le voir sur la courbe, dans une situation bilatérale on peut observer **2 zones critiques** : c'est comme si, au lieu d'avoir un seul paramètre théorique Xt, on avait **2 paramètres théoriques** (= seuils critiques) : **-Xt** et **+Xt**.

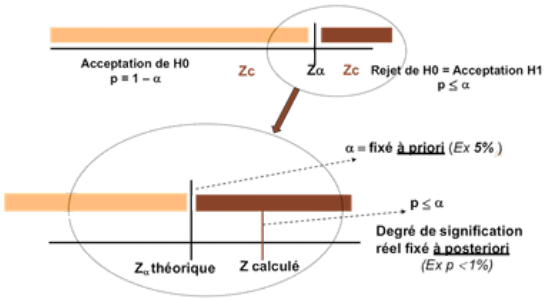
Par exemple, si on a un paramètre calculé Xc = 2,5 et que notre paramètre théorique est de ± 2,2, on aura alors 2 valeurs critiques (=seuils critiques) : 2,2 et -2,2. Ici comme 2,5 > 2,2, on sera plutôt à droite. Si notre paramètre calculé avait été -3,3 par exemple, -3,3 < -2,2 donc on sera plutôt à gauche.

Ainsi, si notre hypothèse nulle H0 est, comme toujours, « il n'y a aucune différence entre les 2 groupes », alors selon notre paramètre calculé on pourra la **rejeter** sous un certain seuil. Si on est à droite ou à gauche, on pourra savoir si la différence est plutôt **négative** *(donc dans le cas où on teste un nouveau médicament, peut-être donc qu'il est nocif plutôt qu'utile)*, ou alors plutôt **positive** *(le médicament fonctionne bien et améliore la maladie)*.

Dans un **test bilatéral**, le **rejet de H0** indique l'existence d'une différence significative et le **côté de la zone critique** dans lequel tombe la statistique de test permet ensuite d'interpréter le **sens** de cette différence (positive ou négative, soit en santé : plutôt délétère ou bénéfique) !

Point'tut : pourquoi fixe-t-on le risque α à priori et p à posteriori ?

Il faut distinguer le **risque α** qu'on **fixe à priori**, et le **degré de signification p** qu'on **observe à postériori** et qui illustre la force de notre conclusion (à quel point on a "raison" de rejeter H0). Il arrive qu'à la fin du test, après avoir rejeté H0 au risque α initial (5% généralement), on se rende compte qu'on puisse rejeter H0 à un degré de signification encore plus petit. Plus p est petit, moins on a de chance de se tromper dans notre déduction. *Donc par exemple, après rejet à α = 5%, une p-value = 1% montre un rejet plus solide que prévu.*



Comment arriver à une conclusion à partir d'un test d'hypothèses ?

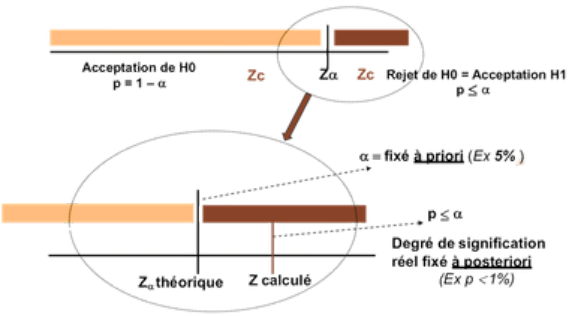
1. Fixer le risque α à priori
2. Chercher Z_t dans la table
3. Calculer Z_c grâce aux formules (dépendantes des tests)
4. Comparer Z_c à Z_t ; on distingue 2 situations :

$Z_c < Z_t$	$Z_c > Z_t$
Acceptation de H0 **p = 1 – α**	**Rejet de H0** **p ⩽ α**

5. Fixer le degré de signification p à postériori

Point'tut : le degré de signification

Le schéma ci-contre illustre les probabilités d'accepter ou de rejeter H0 avec H0 vraie. La zone **beige** représente la probabilité d'accepter H0 avec H0 vraie → elle est donc égale à p = 1 – α (soit 95% la plupart du temps, ce qui signifie qu'on a 95% de chance "d'avoir juste").



La zone **marron** représente la probabilité de rejeter H0 avec H0 vraie, soit une probabilité p ⩽ α (soit qu'on a 5% de chance ou moins de se tromper et rejeter H0 à tort).

Mais alors si notre paramètre calculé tombe dans la zone marron, on rejette H0 à tort ? Notre conclusion n'est alors pas juste ?

Eh bien non : si notre Zc tombe dans la zone critique marron, c'est simplement que les données observées ne colle pas avec la situation dans laquelle H0 **serait** juste : mais on ne peut pas à savoir à l'avance !

Lorsqu'on fait ce graphique, on ne peut **pas** savoir à priori si on rejette ou accepte H0. On crée juste un graphique qui illustre une situation où il serait **probable** de l'accepter, et si justement **malgré tout** notre statistique du test nous dit qu'on le rejette, alors on va chercher à savoir s'il ne serait justement pas vrai de dire qu'il y a une différence significative entre les groupes.

Point'tut : le degré de signification

Le **degré de signification p**, également appelé **p-value**, est **observé à postériori** et est à distinguer du risque de première espèce !

Le degré de signification représente la **probabilité d'obtenir un résultat au moins aussi extrême que celui observé si l'hypothèse nulle est vraie.**

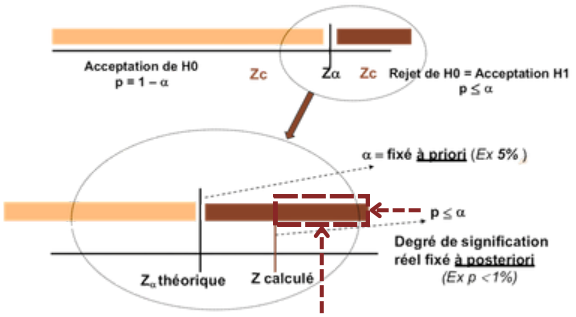
Attention : ce n'est PAS la probabilité que l'hypothèse nulle soit vraie.

En d'autres termes, c'est la probabilité que les résultats observées soient simplement dûs au **hasard**, et qu'on se trouve dans une situation extrême (comme si on était dans les queues d'une courbe de Gauss : on est dans un cas rare), mais tout de même bien dans une situation où H0 serait vraie !

On comprend donc que plus la p-value est petite, mieux c'est car **moins on a de chance de se tromper dans notre conclusion de rejet de H0**.

En pratique, on la compare à un seuil α souvent fixé à 0,05 :

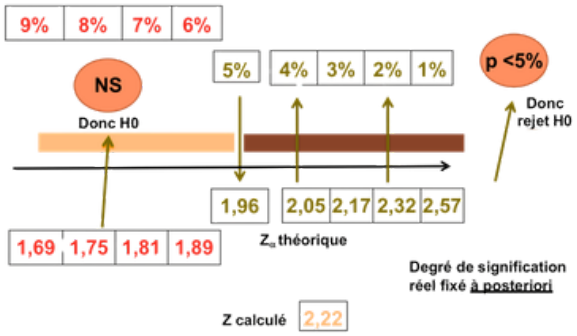
- si **p ⩽ 0,05**, le résultat est dit statistiquement significatif et on rejette l'hypothèse nulle ;
- si **p > 0,05**, le résultat n'est pas significatif.



Sur le schéma, la p-value serait donc la zone marron se trouvant à droite de Zc, le paramètre calculé !

Exemple :

1. α = 5%
2. Z_t = 1,96
3. Z_c = 2,22
4. 2,22 > 1,96 (Zc > Zt) donc **on rejette H0**



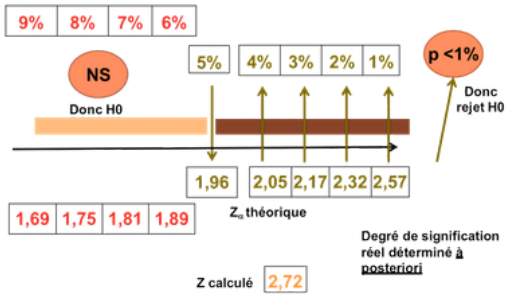
→ On cherche à savoir si on peut aussi rejeter pour α = 1%. Avec 1 % on a donc Z_t = 2,57, or 2,22 < 2,57 (Zc < Zt) donc **on ne rejette pas H0 à 1%**. La précision n'a pas augmenté.

6. On a donc p < 5%

On peut même dire dans ce cas-là qu'on rejette H0 à 3% car 2,17 < 2,22 < 2,32 comme visible sur le schéma ci-dessus qui représentent les Zα.

Cependant **en pratique on utilise seulement 1% et 5%.** *Si on rejette ou accepte H0 à tous les seuils, le test n'est pas très discriminant ou non significatif.*

Pour Z_c = 2,72 le raisonnement est le même, mais on peut ici rejeter H0 à 1% car 2,72 > 2,57.



Après un test d'hypothèses, on peut se retrouver face à 2 situations :

Situation unilatérale	Situation bilatérale
Le rejet d'H0 permet seulement de dire qu'il y a une différence significative entre les 2 situations. C'est la situation la plus fréquente.	**L'acceptation de H1 permet de déterminer laquelle des situations est la meilleure.**

Exemple : on compare 2 traitements A et B. En rejetant H0 :

- En situation **unilatérale** : on pourra uniquement dire qu'il y a une différence significative entre les 2 traitements
- En situation **bilatérale** : on pourra dire qu'il y a une différence significative **et** que le traitement A est meilleur que le B (ou inversement)

Big Data (Données massives)

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

Nous générons et détenons quantités d'informations personnelles *(alimentation, achats, contributions, 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, **diverses études épidémiologiques** ont été lancées ("pour le meilleur et pour le pire ?"). Par exemple, des sociétés privées aux USA analysent ces data et en tirent des conclusions : ils proposent donc à des femmes l'ablation des 2 seins car leur profil génétique est comparé à celui de milliers d'autres femmes permettant ainsi de **détecter un risque accru de cancer du sein**.

On peut également évoquer les **objets connectés** (bracelets, balances, tee-shirts, fauteuils, iwatch,…) qui permettent de suivre sa propre forme physique, la comparer à ce qu'elle devrait être (!) mais en alimentant aussi de manière continue ces fameuses Big Data.

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, diag, trait, dose, durée)*, **biologiques**,… Il n'y a jamais de données de type psy, émotionnel, ..

Les **Big Data** permettent de **recouper et analyser TOUS ces types de données** et de **remettre en cause certaines conclusions ou décisions**..

De plus, un **échantillon traditionnel** comprend des effectifs de quelques dizaines, au mieux quelques centaines d'individus, représentant des populations cibles souvent de plusieurs **centaines de milliers** d'individus. Est-ce vraiment le schéma le plus performant ?

Grâce aux **Big Data**, les effectifs des échantillons observés et étudiés sont **de l'ordre de la population cible**, et ça, c'est tout de même un vrai bouleversement théorique !

Je ne pense pas que ça tombera texto cours mais comprenez bien l'intérêt des Big Data surtout en santé :)

LIEN ENTRE DEUX VARIABLES QUALITATIVES

On se demande si le pourcentage d'individus possédant un caractère x dans un **groupe A** **est le même** que le pourcentage d'individus possédant le caractère x dans un **groupe B**.

Le caractère x est ici **qualitatif** (couleur des yeux, porteur de lunettes...)

Comparaison de pourcentages

(tout effectif)

- Paramètre Z → **écart-réduit ε**
- ε_t vient de la table de l'écart réduit
- **ε = 1,96** avec α = 5%

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

avec $q_A = 1 - p_A$

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

Comment trouver Zt dans la table ?

*Pour les tests **sans** DDL*

- On cherche ε (Z_t) en fonction d'α
- On regarde le **dixième de la valeur d'α** sur les **lignes** et le **centième de la valeur d'α** sur les **colonnes**.
- **ε** sera à **l'intersection**.

Exemple : Pour α = 5% = 0,05

- On regarde 0,**0**0 pour les **lignes** et 0,0**5** pour les **colonnes**
- **ε = 1,96**

Pour α = 0,1% = 0,001

- On regarde la table des petites valeurs
- **ε = 3,29**

Table de l'écart réduit

	α 0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	∞	2,576	2,326	2,17	2,054	1,96	1,881	1,812	1,751	1,695
0,1	1,645	1,598	1,555	1,514	1,476	1,44	1,405	1,372	1,341	1,311
0,2	1,282	1,254	1,227	1,2	1,175	1,15	1,126	1,103	1,08	1,058
0,3	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,86
0,4	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,69
0,5	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,6	0,524	0,51	0,496	0,482	0,468	0,454	0,44	0,426	0,412	0,399
0,7	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,8	0,253	0,24	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,9	0,126	0,113	0,1	0,088	0,075	0,063	0,05	0,038	0,025	0,013

Table pour les petites valeurs de la probabilité

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

Exemple : Soient 2 groupes de 200 enfants :

→ Crèche : 200 enfants, 130 atteints de rhinopharyngite

→ Maison : 200 enfants, 96 atteints de rhinopharyngite

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

	Crèche	Domicile
Sain	70	104
Malade	130	96

1. Hypothèses :

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

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

2. Variable 1 : gardé en crèche / gardé à domicile —> qualitatif

Variable 2 : développer ou non une rhinopharyngite —> qualitatif

→ **Test de comparaison des pourcentages**

3. Risque de première espèce α = 5%

4. Recueil des données

5. $p_A = 130/200 = 0,65$ $p_B = 96/200 = 0,48$ $\varepsilon_c = \frac{0,65 - 0,48}{\sqrt{\frac{0,65\times0,35}{200} + \frac{0,48\times0,52}{200}}} = 3,4$

6. **3,4 > 1,96** (εt pour α=5%) donc on rejette H0 au seuil 5%

On peut même aller plus loin car **3,4 > 3,3** (εt pour α=0,1%)

7. **On rejette donc H0 au seuil 0,1%**

8. Pour conclure, sur cet échantillon, le risque de rhinopharyngite est supérieur chez les enfants gardés en crèche que chez les enfants gardés à domicile.

On ne peut cependant **pas généraliser** car il n'y a pas eu de tirage au sort et il manque des informations sur les enfants (précision sur le mode de garde à domicile, sur les revenus des parents...)

Test du χ^2
(tout effectif)

On utilise ce test de préférence si notre tableau de données a plus de 2 lignes ou 2 colonnes.

- Paramètre Z → **χ^2**
- χ^2_t vient de la table du χ^2
- **DDL = (nombre de lignes - 1) * (nombre de colonnes - 1)**

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

avec o_i les **données observées** et c_i les **données calculées**

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

Point'tut : le degré de liberté

Le **DDL** ou **degré de liberté** est le **nombre minimal de valeurs nécessaire dans une série pour pouvoir retrouver/calculer toutes les autres**.

Prenons un exemple avec un DDL venant du test T de Student *(que vous verrez juste après)* de formule DDL = n - 1
→ On a une série de 8 valeurs donc n = 8

2	3	5	12	10	4	7	8	Total : 51

En calculant le DDL, on a : **DDL = n - 1 = 7**
Imaginons qu'on n'ait pas accès à toutes ces valeurs et qu'il en manque une (4). Le total de la série est alors de 47. On peut calculer tout de même la valeur manquante à partir du total (51-47=4).
Cependant, s'il en manque 2 maintenant (4 et 5).
Le total est alors de 42. 51-42=9 mais il est **impossible** de retrouver les 2 valeurs manquantes *(ça peut être 7 et 2, 1 et 8...)*
C'est donc pour ça qu'on doit avoir **au moins 7 valeurs** (**DDL = 7**) pour avoir accès à la série complète.

Comment trouver Zt dans la table ?

*Pour les tests **avec** DDL*

- On cherche χ^2_t (Z_t) en fonction d'α
- On cherche le **DDL sur les lignes** et **α sur les colonnes**
- **χ^2_t** sera à **l'intersection**.

Exemple :

Si α = 5% et DDL = 1 alors χ^2_t = 3,8

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 : Exposition au benzène et leucémie

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

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

2. Variable 1 : leucémie ou non → qualitatif
   Variable 2 : exposé au benzène ou non → qualitatif

→ **Test du χ_2**

3. α = 5%
4. Valeurs observées (cf. tableau ci-dessus) : 15; 20; 485 et 980
5. Valeurs calculées (obtenues par un modèle théorique)

Cette partie est un peu compliquée à comprendre. Je vous la détaille si ça vous intéresse mais vous n'aurez sûrement pas à calculer tout ça à l'examen.

Voici le tableau des valeurs calculées qu'on obtient :

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

- Il y a 35 malades pour 1500 personnes au total, soit 2,33%. On applique ce pourcentage aux exposés et aux non exposés :
  - 2,33% de 500 (les exposés) = 11,65 malades chez les exposés (chiffre théorique)
  - 2,33% de 1000 (les non exposés) = 23.35
- Il y a 1465 non malades pour 1500 personnes au total, soit 97,67%. On applique ce pourcentage aux exposés et aux non exposés :
  - 97,67% de 500 = 488,3
  - 97,67% de 1000 = 976,7

Ainsi, avec les données observées et données calculées, on peut calculer notre χ^2_c :

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

Donc **χ^2_c = 1,42**

6. Pour trouver χ^2_t , on doit calculer le DDL. Pour cela on se base sur le 1er tableau (données observées)

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

D'après la table du χ^2 , on a donc **χ^2_t = 3,84**

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

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

LIEN ENTRE VARIABLES QUALITATIVES ET QUANTITATIVES

On peut se demander, par exemple, si en moyenne la taille des individus d'une population A coïncide avec la taille des individus d'une population B.

Comparaison de moyennes

(grands échantillons : n_1 et $n_2 > 30$)

- Paramètre Z → **écart-réduit ε**
- ε_t vient de la table de l'écart réduit
- **ε = 1,96** avec α = 5%

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

avec n_1 et n_2 les effectifs des différents groupes, m_1 et m_2 les moyennes et s_1 et s_2 les écarts-types

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

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

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

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

1. H0 : les moyennes ne sont pas différentes, ce sont 2 estimateurs du taux de T3 libre chez la femme en général
2. Variable 1 : prise ou non de la pilule → qualitatif
   Variable 2 : dosage de T3 → quantitatif
   → **Test de comparaison de moyennes**
3. α = 5%
4. $\varepsilon_t = 1{,}96$
5. $\varepsilon_c = 6.94$ *(je vous laisse calculer à l'aide de la formule et des données de l'énoncé si vous voulez vérifier)*
6. **$\varepsilon_c > \varepsilon_t$** donc on **rejette H0**
7. $p < 0.0001$

Il y a eu TAS donc le résultat est **généralisable** : la prise de c.o. augmente le taux de T3 libre

Test T de Student

(petits échantillons : n_1 et $n_2 < 30$)

- Paramètre Z → **t**
- t_t est lu dans la table du t de Student
- **DDL = = $(n_1 - 1) + (n_2 - 1) = (n_1 + n_2) - 2$**

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

Si $t_c > t_t$ —> rejet de H0

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

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

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

1. H0 : m_1 et m_2 ne sont pas différents dans les 2 groupes
2. Variable 1 : obèse ou non → qualitatif
   Variable 2 : taux de corticoïdes → quantitatif

   n_1 et $n_2 < 30$ → **test t de Student**
3. α = 5 %
4. **DDL** = 15 + 12 – 2 = **25** donc d'après la table t Student **t_t = 2.06**
5. En calculant avec la formule, on trouve **t_c = 2.92**
6. **$t_c > t_t$** donc on **rejette H0 au seuil 5%**
7. p < 1% après lecture dans la table. On **rejette H0 à 1% défini à postériori**
8. Conclusion : il existe une relation entre obésité et augmentation du taux de corticoïdes au niveau de ces échantillons.

Méthode des couples

Cas des séries appariées

On utilise la **méthode des couples** lorsque **les deux échantillons étudiés ne sont pas indépendants** *(par exemple, si on teste l'efficacité d'un médicament, alors on va observer ses effets avant sa prise puis après sa prise sur un même groupe de patients)*.

Séries indépendantes

Les 2 groupes comparés sont **distincts et indépendants** (sans lien) *Ex : Par TAS on prend un groupe 1 à qui on fait une prise de sang puis un groupe 2 à qui on fait aussi une prise de sang. Il n'y a pas de lien entre le groupe 1 et le groupe 2*

Séries appariées

Les 2 groupes comparés ne sont **pas distincts et indépendants** (liés) *Ex : On fait une prise de sang à un groupe puis une prise de sang à ce même groupe 6 mois plus tard. Il y a un lien entre les premiers et les derniers résultats car l'analyse sanguine est propre à chacun*

Comparaison de moyennes
Si n > 30

$$\varepsilon = \frac{m_d}{\sqrt{\frac{s^2}{n}}}$$

Test T de Student
Si n < 30

$$t = \frac{m_d}{\sqrt{\frac{s^2}{n}}}$$

avec d : différence de résultat pour un même sujet, m_d : moyenne des d, s : variance des d et n : nombre de couples

Le reste de la méthodologie reste **identique** : 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.

Exemple : On souhaite évaluer l'effet d'une substance S capable de désintoxiquer les fumeurs. On considère par TAS 2 groupes de 40 fumeurs. L'un reçoit la substance S, l'autre reçoit le placebo P. Le traitement dure 2 mois. La consommation de cigarette par jour (C) est notée avant et après traitement (TTT).

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

La consommation est-elle identique dans les 2 groupes ?

1ère précaution à prendre → Les 2 groupes doivent être comparables vis-à-vis des paramètres susceptibles d'influencer la réponse au traitement des paramètres

(âge, sexe, catégorie socio-professionnelle, conso/jour etc..). Si ce n'est pas le cas, il faut en tenir compte lors des conclusions.

On compare donc les consommations moyennes avant ttt dans les 2 groupes

- H0 : les moyennes de consommation sont équivalentes dans les 2 groupes
- Variable 1 : S ou P = qualitative, Variable 2 : C = quantitative
- Échantillons indépendants → test de comparaison des moyennes
- $\varepsilon_c = 2 > 1{,}96$
- On rejette H0 avec un risque $\alpha = 5\%$.

Il y a donc une **différence significative** de la consommation moyenne de cigarette par jour dans les 2 groupes. On fume plus dans le groupe S (situation bilatérale). Il faut en tenir compte lors de l'étude de la variation de consommation avant et après ttt.

Dans le groupe placebo, la consommation moyenne diffère-t-elle avant et après ttt ?

- Variable 1 : avant après ttt = qualitatif, Variable 2 : C = quantitative.
- Échantillon non indépendants → méthode des couples ; n > 30 → test de comparaison des moyennes
- $\varepsilon_c = 26{,}9 > 1{,}96$: rejet de H0

Il y a une **différence très significative** ($p < 0{,}001$) entre C avant et après ttt dans le groupe **placebo**. Il y a un effet psychologique : l'envie de profiter de l'étude pour arrêter de fumer.

Les 2 groupes diffèrent-ils dans leurs conso moyenne après ttt ?

- H0 : les moyennes de consommation sont les mêmes dans les 2 groupes
- Variable 1 : S ou P = qualitative, Variable 2 : C = quantitative
- Échantillons indépendants et n > 30 → test de comparaison des moyennes
- $\varepsilon_c = 1{,}42 < 1{,}96$: on accepte H0 au seuil 5%

Il n'existe **pas de différence significative** entre les 2 groupes pour la consommation après ttt.

Les 2 groupes diffèrent-ils pour la variation de consommation avant et après ttt ?

Il faut comparer les variations dans les 2 groupes pour prouver l'efficacité de la substance S

- H0 : il n'existe pas de différence entre les variations de consommation dans les 2 groupes
- Variable 1 : S ou P = qualitative, Variable 2 : C = quantitative
- n > 30 → test de comparaison des moyennes
- $\varepsilon_c = 2{,}09 > 1{,}96$: rejet de H0 au risque 5%

Il existe une **différence significative** entre les variations de consommation dans les 2 groupes (p<5%). Conclusion : Il y a eu TAS donc le résultat est généralisable.

Conclusion générale

Il n'y a **pas de différence de consommation après traitement** mais il y avait une **différence avant traitement** (le groupe S fumait plus). On peut donc dire qu'il y a une **efficacité du traitement S pour désintoxiquer les fumeurs**.

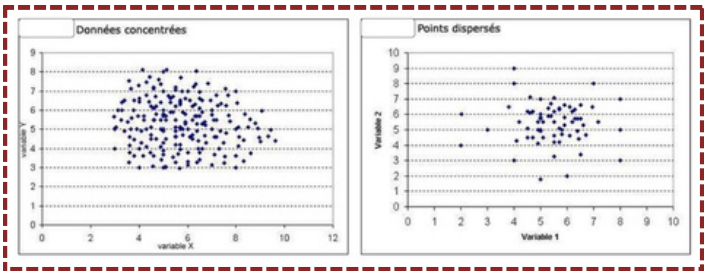
LIEN ENTRE DEUX VARIABLES QUANTITATIVES

Corrélation et régression

La **corrélation** est l'évaluation de la liaison entre 2 variables quantitatives

La **régression**, elle, est une méthode mathématique permettant d'expliquer les relations entre les variables observées. *Elle va nous permettre d'objectiver la corrélation qu'on observe entre 2 variables*

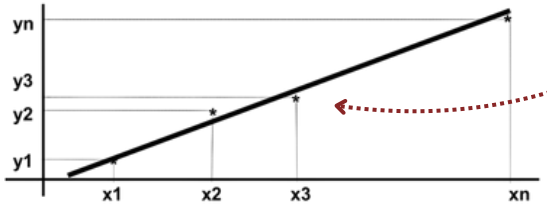
Représentation des données : nuages de points



En variable **x** *(abscisse)*, on met la **variable explicative**.

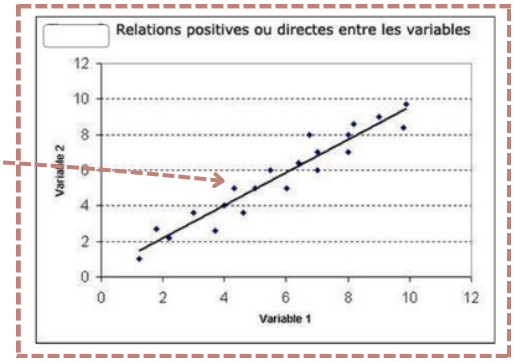
En variable **y** *(ordonnée)*, on met la **variable à expliquer**.

La **droite de régression** permet de visualiser si l'une des 2 variables est dépendante de l'autre. Elle est aussi appelée **droite des moindres carrés** car elle passe au plus près de chaque point du graphe. *Elle cherche à **minimiser** la somme des **carrés des distances** entre les **points observés** et la **droite** ajustée.*



Corrélation ≠ causalité

*Ici la corrélation est **positive** : les variables évoluent dans le même sens (quand l'une augmente, l'autre aussi).*
*La **droite de régression** passe au plus près des points du nuage*



Test de corrélation (de Pearson)

Ce test et ses modalités ne sont pas organisés comme ça dans le cours du prof mais je vous l'ai arrangé à ma sauce pour que vous puissiez faire le parallèle avec les autres tests.

- Paramètre Z → **r**
- r_t est retrouvé dans la table du coefficient de corrélation
- **DDL = n - 2**

Si $r_c > r_t$ —> rejet de H0

Exemple : Sur un échantillon de 10 sujets, on recueille leur âge (années) et leur 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 ?

1. H0 : le taux de cholestérol n'est pas lié à l'âge
2. Variable 1 : âge → quantitatif
   Variable 2 : taux de cholestérol → quantitatif
   → **Test du coefficient de corrélation**
3. α = 1%
4. DDL = n - 2 = 10 - 2 = 8 donc d'après la table, **r_t = 0,76**
5. D'après la formule *(que le prof ne donne pas)* **r_c = 0,955**
6. **$r_c > r_t$** donc on **rejette H0 au seuil 1%**
7. Conclusion : Plus l'âge augmente, plus le cholestérol augmente.

⚠ Le résultat n'est **pas généralisable** (10 individus sans tirage au sort)

- **Corrélation ≠ causalité** : si d'un point de vue mathématique, on a obtenu une corrélation entre des paramètres statistiques, cela n'implique **pas** une relation de cause à effet entre ces paramètres *(on ne peut pas dire que l'âge cause l'augmentation du taux de cholestérol). C'est le rôle des statistiques et des essais cliniques de déterminer si ce lien de corrélation est un lien de causalité ou non.*

TESTS NON PARAMÉTRIQUES

Tests paramétriques

Tests à forte contraire car ils ne sont fiables que si les données suivent une distribution selon une loi normale. *Pour remplir cette condition ils nécessitent l'utilisation d'effectifs suffisamment **grands**. La majorité des tests que nous avons vu jusqu'à présent sont des **tests paramétriques**.*

Tests non paramétriques

Tests qui ne précisent pas les conditions que doivent remplir les paramètres de la population dont a été extrait l'échantillon. *Ici, ce sont les tests qu'on utilise pour de **très petits effectifs**.* Ils présentent une excellente robustesse.

On utilise **obligatoirement** un **test non paramétrique** quand les effectifs sont très faibles : **4 < n < 12**.

Pour les variables quantitatives, on en utilise un obligatoirement si les effectifs sont **inférieurs à 5** (car les populations ne sont plus distribuées normalement).

Test U de Mann et Whitney

4 < n < 12

Le test U de Mann et Whitney (ou test de Wilcoxon-Mann-Whitney ou test de la somme de rangs de Wilcoxon) permet de **tester l'hypothèse** selon laquelle les **moyennes de deux groupes de données sont proches**.

- Lien entre données **quantitatives et qualitatives**
- Paramètre Z → **u**
- mα (ou u_t) est retrouvé dans les tables du test de Mann et Whitney

Si $u_c > u_t$ (ou $u_c >$ mα) —> acceptation de H0 ⚠ *≠ des tests paramétriques !*

Si les effectifs sont **grands** (n_1 et n_2 > 20 en général), U suit approximativement la **loi normale**. *Et ce, comme la plupart des tests non paramétriques : on ne peut pas utiliser un test pour un effectif inférieur à ses conditions de base, mais on peut toujours utiliser un test pour des effectifs supérieurs.*

Point'tut

Pourquoi parle-t-on de mα et des tables de test ?

→ Dans un **test paramétrique**, il n'y a qu'une seule table universelle (paramétrée par le degré de liberté et le risque α).

→ Dans un **test non paramétrique** (comme Mann et Whitney), il faut **plusieurs tables spécifiques à chaque combinaison d'effectifs** (n_1 et n_2).

mα, quant à lui, correspond à la valeur critique qu'on aura trouvé dans la table qui correspond à notre test. C'est tout simplement $u_{théorique}$, la valeur critique au risque alpha qu'on a choisi.

Comment trouver u (Zt) dans la table ?

Pour un test non paramétrique

- On regarde la **différence $n_2 - n_1$** sur les **lignes** et le **plus petit effectif** sur les **colonnes**.
- U sera à **l'intersection**.

Exemple : $n_1 = 10$ et $n_2 = 10$ donc $u_t = 23$

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

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

Méthodologie du test

On a 2 échantillons indépendants E_1 et E_2 de taille n_1 et n_2.

H0 : les moyennes expérimentales dans les 2 groupes sont égales ($\mu_1 = \mu_2$)

Pour mieux comprendre, on va procéder avec un exemple : on veut savoir si une nouvelle molécule présente un effet anti-dépresseur. Pour cela, on organise un essai portant sur **20 malades dépressifs, répartis en 2 groupes**

Les 20 malades sont répartis par TAS en 2 groupes de 10 sujet : l'un recevant la nouvelle molécule, l'autre recevant le placébo.

On évalue les patients à l'aide d'une échelle numérique de 0 (non déprimé) à 50 (très déprimé). Le groupe témoin reçoit le placébo.

Les patients des 2 groupes sont évalués avant le traitement puis après le traitement au bout de 28 jours.

Témoins	J0	34	30	25	27	31	24	28	30	35	26
	J28	31	28	26	25	24	25	26	27	32	25
Traités	J0	27	32	30	28	25	33	29	31	32	29
	J28	22	25	23	26	20	27	21	26	25	23

Le traitement est-il efficace ?

Pour cela on calcule les **différences entre les scores** de déprime avant le traitement et après le traitement, pour chacun des groupes (témoins et traités)

Témoins d = J0 - J28	3	2	-1	2	7	-1	2	3	3	1
Traités d = J0 - J28	5	7	7	2	5	6	8	5	7	6

C'est sur ces données qu'on va travailler.

Etape 1 : tri des valeurs

1. **Fusionner** les deux échantillons et **trier** les données par ordre croissant

Précision *: dans son cours, le prof rajoute une étape qui consiste à attribuer à chaque valeur un* ***rang****. A l'aide de ses rangs, on calcule* u_1 *et* u_2 *avec la formule suivante :* $u = R - \frac{n(n+1)}{2}$ *(R étant la somme des rangs du groupe correspondant). Cependant, on utilise, dans le reste du cours, une autre méthode où les rangs sont inutiles. Je vous le mets quand même mais pour la suite on se basera sur les valeurs sans rang, triées par ordre croissant.*

Calcul du rang : on a **4** valeurs identiques, qui ont comme rangs de base : **4, 5, 6 et 7**. Le rang est donc : (4 + 5 + 6 + 7) / 4 = 22 / 4 = **5.5**

Différence	-1	-1	1	2	2	2	2	3	3	3
Rang	1,5	1,5	3	5,5	5,5	5,5	5,5	9	9	9

5	5	5	6	6	7	7	7	7	8
12	12	12	14,5	14,5	17,5	17,5	17,5	17,5	20

Etape 2 : calcul de u_1 et u_2

2. Pour chaque valeur x_i issue de E_1, **compter le nombre de valeurs issues de E_2 situées après lui** dans la liste ordonnée. Le nombre obtenu est noté **u_1.**

3. Faire la même chose pour trouver la somme **u_2**.

-1	-1	1	2	2	2	2	3	3	3

5	5	5	6	6	7	7	7	7	8

Après avoir trié les valeurs par ordre croissant, on va **compter** , pour **chaque valeur**, le nombre de valeurs de l'autre série qui se trouvent après elle.

Par exemple, après le premier -1 qui fait partie de la liste bleue (groupe témoin), il y a 10 valeurs (du groupe traité).

On **additionne** ensuite ces nombres pour chaque groupe. On obtient u_1 et u_2.

Exemple du groupe traité : $u_2 = 4 + 1 + 1 + 1 + 1 + 1 + 0 + 0 + 0 + 0 = 9$

On trouve donc **$u_1 = 91$** (100 - 9) et **$u_2 = 9$**

Etape 3 : choix de u et conclusion

4. Choisir la valeur **la plus faible** entre u_1 et u_2 : c'est **u**. On note **U la variable aléatoire associée**

5. **Comparer u_c à u_t** trouvé dans la table

6. **Conclure** sur H0.

On choisit donc le plus petit des u : c'est notre u_c. On a donc **$u_c = 9$**.

Après avoir regardé la table théorique du test de Whitney correspondante ($\alpha = 5\%$, $n_1 = 10$, $n_2 = 10$), on trouve **$u_t = 23$**

$u_c < u_t$ → on **rejette H0 au risque 5%**

Les différences sont donc significativement plus importantes pour le traitement que pour le placébo : **le traitement est efficace** contre la dépression.

Test R' de Spearman

n ⩽ 5

- Lien entre données **quantitatives**
- Paramètre Z → **r'**
- r'_t vient de la table du r' de Spearman

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

Si $r'_c > r'_t$ —> acceptation de H0

Remarque : on a donc toujours r ou r' dans un test ne traitant que de données quantitatives !

Exemple :

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

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

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

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

2. Variable 1 (X) : note → quantitative

Variable 2 (Y) : classement → pseudo-quantitative

→ **test r' de Spearman**

Remarque : même si dans les faits pseudo-quantitatif = qualitatif, dans le r' de Spearman on traite ces variables comme si elles étaient quantitatives !

3. On associe à chaque X et à chaque Y un **rang**. On calcule $\mathbf{d_i}$ la différence entre le rang X et le rang Y, et $\mathbf{d_i^2}$. (cf formule)

X Biostat	12,4	4,9	18,1	5,4	19,4	16
Rang X	3	1	5	2	6	4
Y Classement	210	555	6	445	5	14
Rang Y	4	6	2	5	1	3
d_i	-1	-5	3	-3	5	1
d_i^2	1	25	9	9	25	1

D'après la formule, on obtient **$r'_c = -1$**

4. Dans la table théorique, avec n = 6, **$r'_t = 0{,}89$** avec α = 5% et **$r'_t = 1$** avec α = 1%.

5. **$r'_c < r'_t$** : on **rejette donc H0** (p < 1%)

6. Conclusion : on met en évidence un lien très significatif entre ces 2 séries. Il s'agit de 2 séries **corrélées**. Plus la note de biostat est élevée, plus petit est le rang de classement (d'où le signe - pour r').

Point'tut : le coefficient de corrélation des rangs

Le **coefficient de corrélation des rangs r'** de Spearman mesure à quel point nos variables sont corrélées :

- Sa **valeur absolue** représente la **force** de la corrélation → plus c'est proche de 1, plus le lien est fort
- Son **signe** représente le **sens** de la corrélation → positif : les variables évoluent dans le même sens; négatif : les variables évoluent dans des sens opposés

Méthodologie d'utilisation des tests

Effectif	Données quantitatives	Données qualitatives	Données qualitatives et quantitatives
4 < n < 12 ou n < 5	r' de Spearman	Comp % ou χ^2	U de Mann & Whitney
12 ⩽ n < 30	Coeff de corrélation r	Comp % ou χ^2	t de Student
n ⩾ 30	Coeff de corrélation r	Comp % ou χ^2	Comp moyennes

On peut utiliser un test pour des **effectifs supérieurs** mais **pas** pour des ~~effectifs inférieurs~~.

Remarque : le choix du test le plus approprié ne dépend pas que de l'effectif, il y a d'autres facteurs à prendre en compte (que l'on ne vous demande pas de connaître). Cela explique pourquoi le prof peut utiliser un test t de Student avec un effectif de 10.