Biostatistiques

LAS - 2025/26

TUTORAT RENNAIS

SOMMAIRE

SOMMAIRE **3**

INTRODUCTION AUX BIOSTATISTIQUES **5**

L'EPREUVE 5
REPARTITION DES QUESTIONS A L'EXAMEN 5
SUPPORTS 5
EN BIOSTATISTIQUES, VOUS AUREZ 2 PROFESSEURS. 6
LA CALCULATRICE : CALCULATRICE SCIENTIFIQUE TEXAS INSTRUMENTS TI-36X PRO 6
ET LE TUTORAT DANS TOUT ÇA ? 6
LES ED : 6

METHODOLOGIE - BIOSTATISTIQUES **7**

LES EXERCICES 7
LA RÉGULARITÉ 7
PLANIFIER 7
FICHE CALCULATRICE 8
CORRELATION ET REGRESSION : 8
NOTES PERSONNELLES 11

TESTS STATISTIQUES **12**

PRINCIPES DES TESTS STATISTIQUES 12
NOTES PERSONNELLES 13

TESTS DE COMPARAISON DE MOYENNES **14**

EXERCICE DU CONCOURS 2017/2018 15
NOTES PERSONNELLES 19

COMPARAISON DE FREQUENCES **20**

EXERCICE TYPE CORRIGE : COMPARAISON DE FREQUENCE 21
ANNALE 2017-2018 : 21
NOTES PERSONNELLES 25

CORRELATIONS ET REGRESSIONS **27**

CORRELATION : EVALUER LA LIAISON ENTRE 2 VARIABLES QUANTITATIVES 27
REGRESSION LINEAIRE : VOIR SI LA COURBE DES VARIATIONS DE Y EN FONCTION DE X EST UNE DROITE 27
NOTION DE COVARIANCE : 28

POINTS IMPORTANTS A MAITRISER : 29
→ PARTIE CORRELATION 29
→ PARTIE REGRESSION 29
EXERCICE TYPE CORRIGE : CORRELATIONS REGRESSIONS 30
REPRISE D'UN EXERCICE DE L'ANNALE 2014-2015 : 30
METHODE DE RESOLUTION : 31
CHOIX DU TEST 31
METHODE POUR CALCULER LE COEFFICIENT DE SPEARMAN 32
EFFECTUER LE TEST DE SPEARMAN 32
INTERPRETER LE TEST 33
L'ESTIMATION DU COEFFICIENT DE CORRELATION EST DONNEE PAR : 34
NOTES PERSONNELLES 35

EPIDEMIOLOGIE 36
MESURES EN EPIDEMIOLOGIE : 36
INDICATEURS EPIDEMIOLOGIQUES : 38
NOTION D'EPIDEMIOLOGIE : 39
LES ENQUETES DESCRIPTIVES 39
LES ENQUETES ETIOLOGIQUES 39
ENQUETE DE COHORTE (PROSPECTIVE OU RETROSPECTIVE) OU ENQUETE EXPOSE NON-EXPOSE 40
ENQUETE CAS-TEMOINS (RETROSPECTIVE) 41
RESUME 42
NOTES PERSONNELLES 44

INTRODUCTION AUX BIOSTATISTIQUES

Les biostatistiques sont une matière dont le bénéfice est souvent sous-estimé ! Il ne faut pas les négliger, le coefficient reste important (que ce soit en PASS ou en LAS). Et c'est ce qui peut vous permettre de faire la différence ! Même pour ceux qui n'aiment pas les calculs, pas de panique, cette matière reste accessible, les exercices au concours sont souvent similaires, et si vous les maîtrisez bien, vous serez récompensés de votre dur labeur et prendrez même du plaisir à faire vos calculs ;)

L'épreuve

Composée uniquement de **QCMs** (environ 10 QCMs) portant sur le cours. Les années précédentes, il n'y avait pas de calculs. Par ailleurs, il est possible qu'on vous demande d'interpréter des résultats de tests.

Répartition des questions à l'examen

1 - Principes des tests statistiques (Cuggia).....(3 QCM)
2 - Tests de comparaison de moyennes (Hitti).....(2 QCM)
3 - Tests de comparaison de fréquences (Hitti).....(1 QCM)
4 - Corrélations régressions (Cuggia).....(2 QCM)
5 – Corrélation linéaire (Cuggia).....(1 QCM)
6 - Epidémiologie (Cuggia).....(1 QCM)

Supports

- **Vidéos** disponibles sur l'ENT. Voici un tableau pour t'aider à planifier le visionnage !

Chapitre	Temps vidéo 1 (min)	Date	Temps vidéo 2 (min)	Date
Principes des tests statistiques	43		45	
Tests de comparaison de moyennes	31		36	
Tests de comparaison de fréquences	37		20	
Corrélations régressions	52		X	
Corrélation linéaire	52		33	
Epidémiologie	49		46	

- Toutes les **diapositives** qui contiennent ce qu'il faut connaître seront disponibles sur Moodle sur le site de l'ENT.

En biostatistiques, vous aurez 2 professeurs.

Dr. Cuggia : Cet enseignant prend vraiment le temps d'expliquer les choses et donne beaucoup d'exemples.

Ses QCMs sont plus axés sur des questions de cours.

M. Hitti : Ses cours sont majoritairement composés de démonstrations, ne vous inquiétez pas si vous ne comprenez pas tout, les diapositives évoquées en cours sont plus compliquées qu'à l'examen...

Néanmoins, il peut être intéressant d'y assister pour avoir les explications et poser ses questions. Il réalise également la correction d'exercices ce qui peut être utile pour comprendre les étapes de résolution des exercices et certaines astuces.

La calculatrice : Calculatrice scientifique TEXAS INSTRUMENTS TI-36X Pro

Si vous ne vous l'avez pas, il y a une commande groupée organisée par le tutorat tous les ans. Cependant, en attendant de l'avoir, ou s'il vous arrive de l'oublier, pas d'excuses pour ne pas travailler les biostats ! Il existe une application assez fidèle à la calculette : **Calculatrice scientifique 36.**

Et le tutorat dans tout ça ?

Bien sûr, vous n'êtes pas seuls, si vous avez la moindre question sur une conférence ou autre, il y a le **forum**. Et, si vraiment vous voulez une explication plus détaillée, toute **l'équipe de tuteurs** est à votre disposition pendant les conférences !

Les ED :

Ils sont proposés **par des profs** pour les chapitres les plus calculatoires (lois, tests, corrélation - régression). Ils sont réalisés en **petits groupes**. Il s'agit d'une heure de correction d'exercices donnés au préalable, cela vous sera utile pour poser toutes vos questions et reprendre avec le prof la résolution étape par étape de manière détaillée. Ils permettent donc un soutien supplémentaire pour réussir l'examen de biostatistiques !

METHODOLOGIE - BIOSTATISTIQUES

Globalement similaire pour toutes les matières “à calculs” (chimie, biostatistiques et biophysique) :

- Voir les **cours**
- Voir les **compléments de cours** (notamment ce livret) s’il y a des morceaux de cours incompris
- Voir la résolution d’**exercices types**
- Vous **entraîner** avec régularité !
    - Sur les conférences classées, les conférences qui sortent chaque semaine et les exams blancs
    - Essayez si vous le pouvez de bloquer un créneau horaire dédié à la matière

Les biostatistiques sont, comme la chimie ou la physique au deuxième semestre, une **matière à calculs.** Cela veut donc dire que vous ne serez pas évalués sur votre connaissance du cours mais sur votre capacité à résoudre des problèmes. Dans les faits, cela signifie que vous devrez non seulement apprendre vos cours, mais également savoir les appliquer. Pour cela, rien de tels que les **qcms et annales**, ainsi que la sacro-sainte **régularité**.

LES EXERCICES

Le seul moyen de savoir si on a bien compris un cours de biostatistiques, c’est d’appliquer soit-même les formules. C’est pourquoi nous vous recommandons de faire beaucoup de qcms. De plus, les cours de biostatistiques ne changent presque pas d’une année sur l’autre, ce qui veut dire que le contenu des annales correspond toujours au programme. N’hésitez donc pas à **toutes** les faire avant le concours.

La **LAS** étant **récente**, **peu de QCMs** sont **disponibles**. Ainsi, vous pouvez aller vous entraîner sur les **annales ou les QCMs des PASS** sur le site du tutorat, mais faites attention à ne faire que **ceux qui vous concernent**. De plus, pensez bien qu’il n’y aura pas de calculs de tests aussi poussés que pour les PASS dans votre épreuve.

Pro tip : lors de vos sessions d’annales, notez votre note par chapitre, comme ça vous pourrez voir les chapitres qui vous font perdre le plus de points ! Pas besoin de réviser les lois si vous avez eu tous les points sur les questions associées à ce chapitre.

LA RÉGULARITÉ

S’il est vrai que les savoirs non utilisés disparaissent plus rapidement de notre mémoire, ça l’est particulièrement pour les matières à calculs. C’est pourquoi nous vous recommandons vivement de vous entraîner régulièrement. Vous pouvez par exemple décider de **bloquer un créneau** réservé aux biostatistiques, et vous dire que vous travaillerez cette matière tous les jours de 13h à 14h, le matin dans l’amphi avant les cours, ou le soir pour bien finir la journée ! D’autres préfèrent une grosse séance le WE, à chacun sa méthode ! Comme ça, pas besoin de réfléchir, les révisions de cette UE deviennent un automatisme, et vous progressez plus rapidement et avec une meilleure rétention des choses apprises.

PLANIFIER

On peut avoir le vertige au début du semestre en voyant la liste des chapitres… Pour vous rassurer, vous pouvez **planifier** au début du semestre **l’ensemble des chapitres**, selon si vous voulez terminer en avance ou pile à temps le programme ! Pour vous organiser, vous pouvez également utiliser le **planning** disponible sur le tutorat pour le visionnage des **vidéos**.

FICHE CALCULATRICE

Corrélation et Régression :

Variables	Signification
n	Nombre de points de données x (1 Var) ou (x,y) (2 Var).
$\bar{x}$ ou $\bar{y}$	Moyenne de toutes les valeurs de x ou y.
Sx ou Sy	Écart type de l'échantillon x ou y.
σx ou σy	Écart type de la population x ou y.
Σx ou Σy	Somme de toutes les valeurs de x ou y.
Σx^2 ou Σy^2	Somme de toutes les valeurs de x^2 ou y^2.
Σxy	Somme de (x…y) pour toutes les valeurs xy.
a (2 Variables)	Coefficient directeur de la droite de régression de Y en X.
b (2 Variables)	Régression linéaire (ordonnée de l'intersection avec l'axe des y).
r (2 Variables)	Coefficient de corrélation linéaire.
x' (2 Variables)	Utilise a et b pour calculer une valeur anticipée de x quand vous saisissez une valeur de y.
y' (2 Variables)	Utilise a et b pour calculer une valeur anticipée de y quand vous saisissez une valeur de x.
MinX	Minimum des valeurs de x.
Q1	Médiane des points entre le minimum et la médiane, ou premier quartile.
Méd	Médiane de tous les points.
Q3	Médiane des points entre le maximum et la médiane, ou troisième quartile.
MaxX	Maximum des valeurs de x.

/!\ Ne pas confondre Sx et σx /!\.

Calcul de r avec PEARSON

Poids en kg	47,8	49,4	50	50,7	55,8	57,7	67,7	70	72,4	73,8	82,2	88,7
Taux de cholestérol	44	62	130	126	62	98	181	237	280	254	318	293

1. Entrer les valeurs dans la calculatrice :

DATA

- L1 pour entrer les valeurs de la première ligne
- L2 pour entrer les valeurs de la deuxième ligne

PS : Pour supprimer les valeurs : appuyer 2 fois sur DATA puis CLEAR L1

2. Pour accéder aux fonctionnalités de calculs : 2nd puis DATA puis STAT-REG puis 2-VAR Stats.

- Choisir L1 pour xDATA;
- L2 pour yDATA;
- ONE pour FRQ,
- Puis CALC.

Descendre et trouver r.

Calcul de r*s* avec SPEARMAN :

1er tableau

[**Créatinine**] *(en mg pour 100 mL)*	**TFG** *(mL/min)*
0.85	90
0.99	45
1.13	70
1.13	103
1.47	60
1.56	35
1.56	53
2.77	27
3.11	25
3.96	15
4.69	15
5.93	5
5.93	10
7.07	10
11.02	12

2ème tableau

Rangs 1	**Rangs 2**
1	14
2	10
3.5	13
3.5	15
5	12
6.5	9
6.5	11
8	8
9	7
10	5.5
11	5.5
12.5	1
12.5	2.5
14	2.5
15	4

A partir du premier tableau, établir les rangs des deux colonnes (deuxième tableau).

- *PS : si on trouve plusieurs valeurs identiques, faire la moyenne de leurs rangs.*

3. Entrer les valeurs dans la calculatrice :

DATA

- L1 pour entrer les valeurs de la première ligne
- L2 pour entrer les valeurs de la deuxième ligne

PS : Pour supprimer les valeurs : appuyer 2 fois sur DATA puis CLEAR L1

4. Pour accéder aux fonctionnalités de calculs :

2nd puis DATA puis STAT-REG puis 2-VAR Stats.

- Choisir L1 pour xDATA;
- L2 pour yDATA;
- ONE pour FRQ,
- Puis CALC.

Descendre et trouver r_s.

Notes personnelles

TESTS STATISTIQUES

PRINCIPES DES TESTS STATISTIQUES

		Réalité	
		Hypothèse H_0 vraie	Hypothèse H_0 fausse
Statistique	Non rejet de H_0 = accepter H_0	Pas d'erreur	Risque β (ou risque de seconde espèce)
	Rejet de H_0 = accepter H_1	Risque α (ou risque de première espèce)	Pas d'erreur = H_1 vraie $1-\beta$ puissance du test

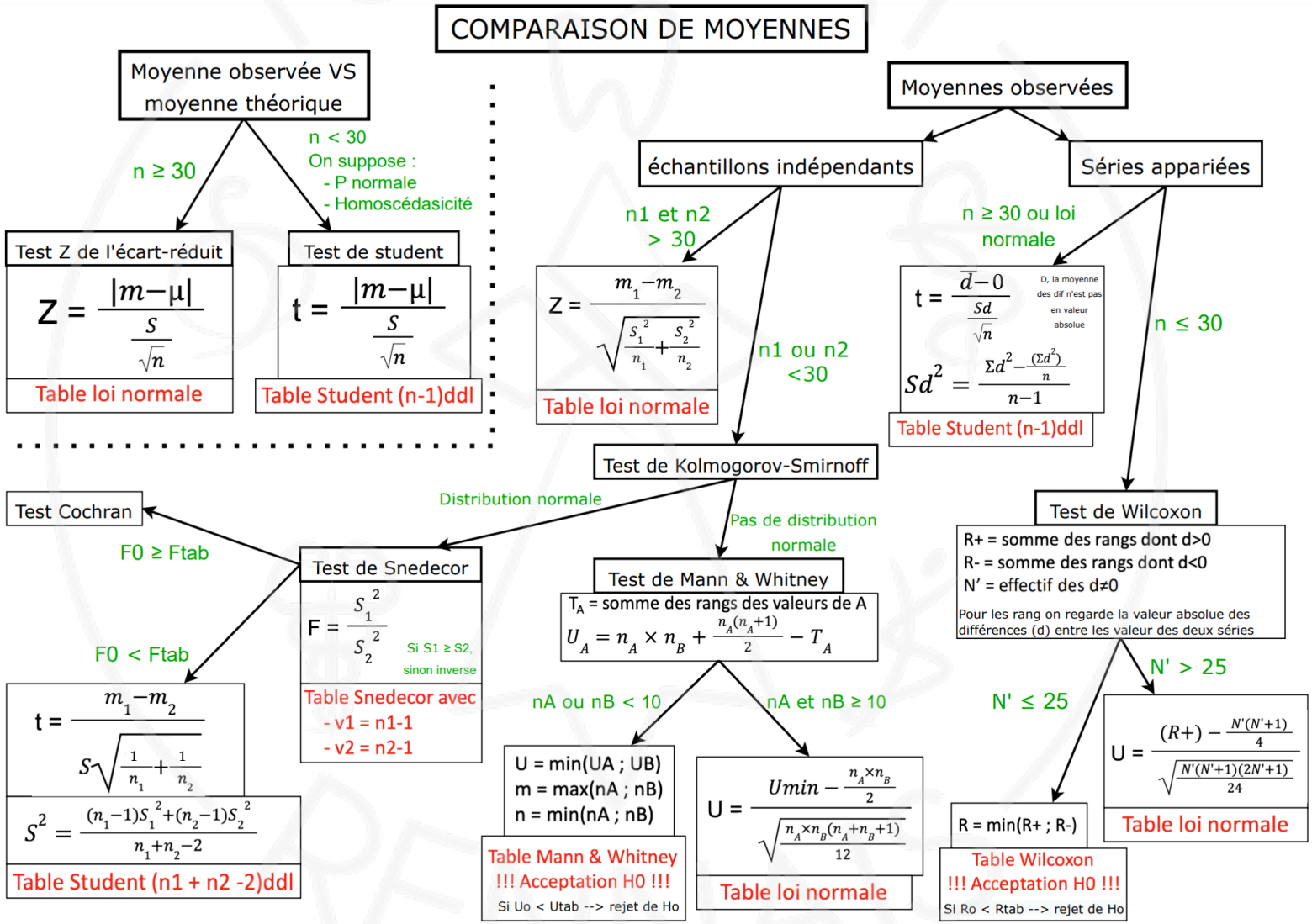
La puissance correspond à la probabilité de rejeter H_0 lorsque H_0 est fausse donc lorsque H_1 est vraie.

- Le **risque** α correspond à la probabilité de rejeter H_0 lorsque H_0 e s t vraie donc H_1 fausse.
- Le **risque** β correspond à la probabilité de ne pas rejeter H_0 alors que H_0 est fausse et H_1 vraie.
    - Plus on diminue α, plus β est grand, plus la puissance du test diminue.

Pour déterminer la **p-value**, on cherche dans la table de **l'écart réduit** la valeur directement **inférieure à** Z_0

NOTES PERSONNELLES

TESTS DE COMPARAISON DE MOYENNES



Pour mieux comprendre comment s'attaquer aux QCMs proposés au concours pour ce chapitre nous allons voir en détail les étapes de résolution de quelques exercices de concours.

Par chance les étapes de résolution de ses exercices sont toujours les mêmes et sont guidées par les différents QCMs qui composent chaque exercice de ce chapitre.

Bon à savoir, le tout est globalement entièrement résumé dans la fiche "comparaison de moyennes".

Bon courage :)

Exercice du concours 2017/2018

On cherche à savoir si l'acuité visuelle des enfants atteints de tumeurs hypophysaires est différente en fonction de la localisation de la tumeur *(intra-sellaire ou supra-sellaire).*

Pour cela, on a mesuré *(sur 10)* l'acuité visuelle de 14 enfants tirés au sort et qui présentent l'un des deux types de tumeurs.

Les résultats sont présentés dans le tableau suivant :

Numéro sujet	Type de tumeur	Acuité visuelle
1	Intra-sellaire	9
2	Intra-sellaire	7
3	Intra-sellaire	8
4	Intra-sellaire	4
5	Intra-sellaire	2
6	Intra-sellaire	9
7	Supra-sellaire	10
8	Supra-sellaire	3
9	Supra-sellaire	10
10	Supra-sellaire	10
11	Supra-sellaire	2
12	Supra-sellaire	8
13	Supra-sellaire	6
14	Supra-sellaire	10

QCM 1 : Sachant que les conditions de normalité ne sont pas respectées, quel test réalisez-vous :

A. Un test de Z de l'écart-réduit
B. Un test de Mc Nemar
C. Un test de Wilcoxon pour séries appariées
D. Un test de Student
E. Toutes les propositions précédentes sont inexactes.

QCM 2 : On réalise le test adapté :

A. La statistique calculée du test vaut 38.
B. La statistique calculée du test vaut 17.
C. Au risque α de 5% et dans une situation bilatérale, l'acuité visuelle est significativement différente en fonction du type de tumeur.
D. Au risque α de 5% et dans une situation bilatérale, on ne peut pas rejeter l'hypothèse nulle.
E. Toutes les propositions précédentes sont inexactes.

Etape 1 *(et résolution du QCM1)* : Déterminer le test adapté à la situation

Pour cela, il suffit de suivre l'arbre de la fiche "comparaison de moyennes" et de s'orienter selon les réponses.

- Dans cette situation, on comprend que le test a pour but de comparer les moyennes d'acuité visuelle observées chez des enfants atteints de tumeurs hypophysaires en fonction de la localisation de la tumeur (intra-sellaire ou supra-sellaire).

On se trouve donc du côté "**moyennes observées**" de l'arbre.

Nb : La partie « Moyenne observée VS moyenne théorique » aurait pu correspondre à un énoncé du type *« On cherche à déterminer si la consommation de gravier à un impact sur la fréquence des vomissements ; dans la population, la moyenne des vomissement est de 3 par semaines et dans notre groupe d'individus consommants du gravier elle est représentée par ce tableau ».*

→ On remarque dans cet énoncé que le but est bien de comparer une moyenne théorique (3/semaines) avec une moyenne observée (chez les consommateurs de gravier)

Après avoir déterminé de quel côté de l'arbre on se trouve on peut continuer avec l'embranchement suivant, ici il s'agit de déterminer si les échantillons sont indépendants ou bien appariés.

- Cela revient à définir s'il s'agit de deux groupes distinct que l'on compare ou bien d'un seul et même groupe qui est étudié avec des paramètres différents. *Par exemple, on compare la moyenne de poids des élèves d'une classe avant et après avoir suivi un régime 100% Mcdonald pendant 30 jours.*

En l'occurrence il s'agit de deux "**échantillons indépendants**" dont les moyennes sont comparées, ceux avec une tumeur intra-sellaire et ceux avec une tumeur supra-sellaire.

Il faut ensuite voir si on peut utiliser la table de la loi normale ou non. Pour cela on regarde "$n1$" et "$n2$" le nombre de composants de nos deux groupes indépendants, comme au moins l'un des deux (les deux en l'occurrence) sont inférieurs à 30, le test de **Kolmogorov-Smirnoff** sera retenu.

Puis, selon si les conditions de normalité sont respectées ou non on s'orientera vers le test de Snedecor ou celui de Mann&Whitney, dans notre cas l'énoncé du QCM1 nous indique que les conditions de normalité ne sont pas respectées, on adopte donc le test de **Mann&Whitney**.

La réponse au QCM1 est donc la E.

Cette étape à été détaillée au mieux pour essayer de tout expliquer mais suivre le tableau en se posant ses questions là rend cette étape très facile et rapide, beaucoup plus qu'avec toutes mes explications :').

Il faut cependant rester vigilant pour ne pas s'y tromper car la résolution du QCM2 en découlera.

Etape 2 (et résolution du QCM2) : Utiliser le test adapté

Maintenant que nous savons quel test utiliser, il ne nous reste plus qu'à le faire, on va donc commencer par calculer **T_A** comme indiqué sur la fiche, pour cela, il faut tout d'abord classer nos valeurs par rangs, le tableau devient ainsi comme suit :

Groupe	Numéro sujet	Type de tumeur	Acuité visuelle	Rangs
A	1	Intra-sellaire	9	9,5
	2	Intra-sellaire	7	6
	3	Intra-sellaire	8	7,5
	4	Intra-sellaire	4	4
	5	Intra-sellaire	2	1,5
	6	Intra-sellaire	9	9,5
B	7	Supra-sellaire	10	12,5
	8	Supra-sellaire	3	3
	9	Supra-sellaire	10	12,5
	10	Supra-sellaire	10	12,5
	11	Supra-sellaire	2	1,5
	12	Supra-sellaire	8	7,5
	13	Supra-sellaire	6	5
	14	Supra-sellaire	10	12,5

Nb : Si deux (ou plus) valeurs ou plus sont identiques, on leur attribue toutes la moyenne des rangs qu'elles auraient obtenus (ex ; trois fois 4 aux rangs 6,7 et 8, la moyenne est 7 donc tous les 4 auront le rang "7").

On fait les rangs sur les 2 groupes en même temps A et B pour les tests de comparaison de moyenne. Ici les deux premiers rangs correspondent aux deux 2,

C'est-à-dire qu'en rang n°1 et en rang n°2 on aura 2 comme valeur.

Pour déterminer le rang quand on a des valeurs identiques, on fait la moyenne des rangs, ici on obtient donc (1+2)/2 = 1,5 pour les deux valeurs 2.

On poursuit ensuite avec la valeur suivante qui est ici 3, elle est unique elle obtient donc le rang 3. Puis on attribut à la valeur 4, le rang 4. Ensuite à la valeur 6, le rang 5 et ainsi de suite...

Comme T_A est la somme des rangs des valeurs de A, on a :

- **T_A**= 9,5 + 6 + 7,5 + 4 + 1,5 + 9,5 = **38**
- **T_B** = la somme des rangs des valeurs du groupe B = **67**

Avec ça, on peut maintenant calculer U_A et U_B avec la formule :

$$U_A = n_A \times n_B + \frac{n_A(n_A+1)}{2} - T_A$$ *Sachant que* ***n_A = 6*** *et* ***n_B = 8***

On a donc $U_A = 6 \times 8 + \frac{6 \times 7}{2} - 38 = 31$

On a aussi $U_B = n_B \times n_A + \frac{n_B(n_B+1)}{2} - T_B = 8 \times 6 + \frac{8 \times 9}{2} - 67 = 17$

Étape 3 (et résolution du QCM2) : Lire le résultat dans la table correspondante

Ensuite, si l'on observe notre fiche de comparaison de moyenne, on remarque qu'après avoir calculer T_A et U_A, deux voies sont possibles, étant donné que n_A (et n_B) sont inférieures à 10, c'est la table de Mann&Whitney qui sera utilisée.

- $U_0 = min\,(U_A, U_B) = U_B = 17$,
- m sera égal à 8, car $n_B > n_A$ et n sera égal à **6** car $n_A < n_B$.

Avec cela nous pouvons enfin lire le résultat dans la table correspondante, pour le seuil demandé, **α de 5%,** la table est la suivante, on a n = 6 et m - n = 2, la statistique tabulée sera donc 8.

Nb : nous sommes dans le cas d'un test **non paramétrique**, donc, si **Uo < Utab → rejet de Ho**.

En l'occurrence U_0 = 17 et U_{tab} = 8, donc $U_0 > U_{tab}$.

On accepte H_0, au risque alpha de 5%, dans une situation bilatérale, l'acuité visuelle n'est pas significativement différente en fonction du type de tumeur.

La réponse au QCM 2 est la réponse BD.

TEST DE MANN-WHITNEY

m − n \ n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	—	—	—	0	2	5	8	13	17	23	30	37	45	55	64	75	87	99	113	127
1	—	—	—	1	3	6	10	15	20	26	33	41	50	59	70	81	93	106	119	
2	—	—	0	2	5	8	12	17	23	29	37	45	54	64	75	86	99	112		
3	—	—	1	3	6	10	14	19	26	33	40	49	59	69	80	92	105			
4	—	—	1	4	7	11	16	22	28	36	44	53	63	74	85	98				
5	—	—	2	4	8	13	18	24	31	39	47	57	67	78	90					
6	—	0	2	5	9	14	20	26	34	42	51	61	72	83						
7	—	0	3	6	11	16	22	29	37	45	55	65	76							
8	—	0	3	7	12	17	24	31	39	48	58	69								
9	—	0	4	8	13	19	26	34	42	52	62									
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												
15	—	2	7	13	20	29	38													
16	—	2	7	14	22	30	40													
17	—	2	8	15	23	32														
18	—	2	8	16	24	33														
19	—	3	9	17	25															
20	—	3	9	17	27															
21	—	3	10	18																
22	—	3	10	19																
23	—	3	11																	
24	—	4	11																	
25	—	4																		
26	—	4																		

Table V.a

(Seuil α = 0,05)

NOTES PERSONNELLES

COMPARAISON DE FREQUENCES

Comparaison d'effectifs
Tests du X² (Chi²)

Fréquence observée VS Fréquence théorique

Tous les Ci ≥5

X² d'ajustement

$$x^2 = \Sigma \frac{(O_i - C_i)^2}{C_i}$$

Table X² (i-1)ddl

Fréquences observées

Ci <5 : Regrouper les effectifs

Ci ≥5

X² d'homogénéité

$$x^2 = \Sigma \frac{(O_i - C_i)^2}{C_i}$$

Table X² ((c-1)*(l-1))ddl

X² corrigé de Yates

$$x^2 = \Sigma \frac{(|O_i - C_i| - 0{,}5)^2}{C_i}$$

Table X² ((c-1)*(l-1))ddl

Comparaison de pourcentages
Tests de l'écart réduit

Fréquence observée VS Fréquence théorique

n*p et n*q ≥5

$$U = \frac{|p_0 - p|}{\sqrt{\frac{pq}{n}}}$$

Table loi normale

Fréquences observées

nA*p, nB*p, nA*q et nB*q ≥5

$$U = \frac{p_A - p_B}{\sqrt{\frac{pq}{n_A} + \frac{pq}{n_B}}}$$

$$p = \frac{n_A \times p_A + n_B \times p_B}{n_A + n_B}$$

q = 1 - p

Table loi normale

Comparaison de séries appariées

Test de Mac Nemar

$$U = \frac{n_2 - n_3}{\sqrt{n_2 + n_3}}$$

Table loi normale

Après \ Avant	+	-
+	n1	n2
-	n3	n4

n2 + n3 ≥10

$$x^2 = \frac{(n_2 - n_3)^2}{n_2 + n_3}$$

Table X² 1ddl

Légende

i : nombre de colonnes

Oi : effectif observé (donné)

Ci : effectif théorique $C_i = \frac{\text{total ligne} \times \text{total colonne}}{\text{Total général}}$

	G+		G-		
A	45	38,3	55	61,7	100
B	70	76,7	130	123,3	200
	115		185		300

EXERCICE TYPE CORRIGE : COMPARAISON DE FREQUENCE

Annale 2017-2018 :

On souhaiterait savoir si le taux de récidive du cancer est dépendant du traitement proposé aux patients (chimiothérapie, chirurgie et radiothérapie). On constitue au hasard 3 groupes de patients en fonction du traitement qu'ils ont reçu et on regarde au bout de 2 ans combien de patients ont récidivé.

Les résultats sont résumés dans le tableau ci-dessous :

	Chimiothérapie	Chirurgie	Radiothérapie
Pas de récidive	400	350	450
Récidive	120	90	130

QCM 26 : On choisit de réaliser un test du Chi^2 :

A. Il s'agit d'un test sans condition d'application particulière.
B. Son calcul repose sur les effectifs observés des différents groupes.
C. Son calcul repose sur les différences entre les effectifs observés et les effectifs théoriques sous l'hypothèse nulle.
D. Son interprétation peut se faire avec la table du Chi^2.
E. Toutes les propositions précédentes sont inexactes.

QCM 27 : On réalise maintenant le test du Chi^2 :

A. Au risque a de 5 % et dans une situation bilatérale, la statistique calculée du test vaut 7,29 (valeur absolue à 0,01 près)
B. Au risque a de 5 % et dans une situation bilatérale, la statistique tabulée (ou théorique) du test vaut 5,99
C. On peut rejeter l'hypothèse nulle H_0 et conclure que le taux de récidive est significativement différent en fonction du traitement.
D. En rejetant l'hypothèse nulle, on a plus d'une chance sur 10 de se tromper de conclusion
E. Toutes les propositions précédentes sont inexactes.

Les QCMs de ce chapitre vont souvent par pair avec un **premier QCM de cours** et un **second de calculs**.

Dans le cas de **valeurs non appariées**, c'est-à-dire que les données ne sont pas liées puisqu'il s'agit d'un groupe différent pour chaque traitement ou test (on ne réalise pas un test avant/après) :

- Si on a déjà les **valeurs théoriques (Ci)** données dans l'énoncé, on utilise la formule du **Chi^2 d'ajustement** *(cf schéma précédent).* Ces valeurs théoriques sont alors souvent identifiées comme une "moyenne nationale" ou celle que l'on devrait obtenir selon une loi.
- Dans le cas où l'énoncé dispose de deux populations (échantillons) testées différemment, il n'y a aucune donnée préalable, elles sont toutes observées. C'est le cas ici. On utilise la formule du **Chi^2 d'homogénéité.**

Correction QCM 26 : BD

A. FAUX. Le test du Chi2 a bien des conditions d'applications particulières. Il faut que les Ci (valeurs théoriques) soient supérieures ou égales à 5.

B. VRAI. On est dans le cas de fréquences observées uniquement.

C. FAUX. On n'a pas de fréquences théoriques.

D. VRAI. On utilise la table du Chi^2 (*voir fiche précédente*).

CORRECTION 27 : BD

Pour résoudre les exercices : procéder par étape et être bien rigoureux.

	Chimiothérapie	Chirurgie	Radiothérapie	Total
Pas de récidive	**400** **405,19**	**350** **342,86**	**450** **451,95**	1200
Récidive	**120** **114,81**	**90** **97,14**	**130** **128,05**	340
Total	520	440	580	1540

1. Calculer les Ci

Pour cela, il faut d'abord *faire le tableau avec les données* de l'exercice s'il n'est pas donné dans l'énoncé, et *ajouter les totaux*.

On calcule ensuite les effectifs théoriques Ci avec $Ci = \frac{(Total\ ligne) \times (total\ colonne)}{Total\ général}$.

- Exemple : Case "chimiothérapie - pas de récidive" : $\frac{1200 \times 520}{1540}$ = 405,19 (à 0,01 près)
- Exemple : Case "chimiothérapie - récidive" : $\frac{340 \times 520}{1540} = 114{,}81$ (à 0,01 près)

Étape qu'on n'a pas besoin de faire dans le cas du Chi^2 d'ajustement.

2. Tous les $Ci \geq 5$, on applique donc la formule du Chi² d'homogénéité :

(Cf fiche précédente)

$$X^2\ calculé\ = \Sigma \frac{(Oi - Ci)^2}{Ci}$$

- **Oi : valeurs observées**
- **Ci : valeurs théoriques**

Soit $X^2\ calculé$ = $\frac{(400-405{,}19)^2}{405{,}19} + \frac{(120-114{,}81)^2}{114{,}81} + \frac{(350-342{,}86)^2}{342{,}86} + \frac{(90-97{,}14)^2}{97{,}14} + \frac{(450-451{,}95)^2}{451{,}95} + \frac{(130-128{,}05)^2}{128{,}05}$

Soit $X^2\ calculé$ = 0,07 + 0,23 + 0,15 + 0,52 + 0,01 + 0,03 = **1,01**

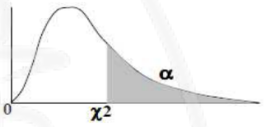
3. On cherche ensuite le $X^2\ tabulé$ dans la table du Chi².

On calcule le nombre de **degré de liberté** (**ddl**).

Nombre de ddl = (nombre de colonnes - 1) × (nombre de lignes - 1) = 2 × 1 = 2

Table de la loi du χ^2

La table donne la probabilité α pour que χ_2 égale ou dépasse une valeur donnée, en fonction du nombre de degrés de liberté (d.d.l.).



ddl \ α	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,001
1	0,0002	0,0010	0,0039	0,0158	2,71	3,84	5,02	6,63	10,83
2	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21	13,82
3	0,11	0,22	0,35	0,58	6,25	7,81	9,35	11,34	16,27
4	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28	18,47
5	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09	20,52
6	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81	22,46
7	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	24,32
8	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	26,12
9	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	27,88
10	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	29,59
11	3,05	3,82	4,57	5,58	17,28	19,68	21,92	24,72	31,26
12	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	32,91
13	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	34,53
14	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	36,12
15	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	37,70
16	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00	39,25
17	6,41	7,56	8,67	10,09	24,77	27,59	30,19	33,41	40,79
18	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,81	42,31
19	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	43,82

Le **risque** α que l'on prend est celui donné dans l'item *(très souvent 5% mais parfois 1%)* et le ddl est calculé selon la formule précédente.

Il ne reste plus qu'à lire dans le tableau.

On trouve ainsi $X^2\ tabulé$ = **5,99**.

- *Avec* $\alpha = 5\%$ *et 2ddl*

4. Interprétation

Dans une situation **bilatérale** :

- H_0 = **homogénéité** entre les différents traitements
- H_1 = **différence** entre les différents traitements (pas de notion de + ou - entre les traitements).

On trouve : $X^2\ calculé < X^2\ tabulé$ (1,01 < 5,99).

Donc au risque $\alpha = 5\%$, on accepte Ho et on rejette H1. On ne montre pas de différence significative entre les différents traitements proposés concernant le taux de récidive.

5. P-value

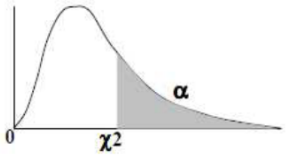
On cherche la **p-value** dans la table du Chi^2.

On cherche **dans la ligne du ddl** calculé précédemment *(soit ici **2**ddl)* la valeur qui se rapproche le plus du X^2 *calculé* = 1,01.

Elle se trouve entre deux valeurs : 0,21 < **1,01** < 4,61

Table de la loi du χ^2

La table donne la probabilité α pour que χ_2 égale ou dépasse une valeur donnée, en fonction du nombre de degrés de liberté (d.d.l.).



ddl \ α	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,001
1	0,0002	0,0010	0,0039	0,0158	2,71	3,84	5,02	6,63	10,83
2	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21	13,82
3	0,11	0,22	0,35	0,58	6,25	7,81	9,35	11,34	16,27
4	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28	18,47
5	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09	20,52
6	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81	22,46
7	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	24,32
8	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	26,12
9	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	27,88
10	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	29,59
11	3,05	3,82	4,57	5,58	17,28	19,68	21,92	24,72	31,26
12	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	32,91
13	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	34,53
14	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	36,12

On lit dans la ligne de α les valeurs de la p-value correspondante :

- La p-value se trouve entre 0,900 et 0,100 soit 90% et 10%.
- 0,900 > p-value > 0,100
- 90% > p-value > 10% (1/10)

On a donc plus d'une chance sur dix de se tromper.

Plus exactement, la p-value vaut **50%**, on a donc une chance sur deux de se tromper.

1,1 - 0,21 = 0,8

Or, on constate dans le début de la ligne qu'une différence de 5% d'alpha correspond à une différence de 0,10 dans la valeur de X^2

$$\frac{0,8 \times 0,05}{0,1} = 0,4$$

0,9 - 0,4 = 0,5 = **50%**.

RAPPEL :

- *Lorsque X^2 calculé > X^2 tabulé , on rejette Ho.*
- *Lorsque X^2 calculé < X^2 tabulé , on accepte Ho.*

Exception : test de Mann & Whitney et Wilcoxon : c'est l'inverse !

NOTES PERSONNELLES

Fiche : Comparaison de moyenne / de fréquence

COMPARAISON DE MOYENNES : CHOIX DES TESTS

MOY OBS X MOY THÉO	*n >/= 30*	Test Z de l'écart réduit			
	n < 30	Test T de student			
	On suppose : normalité / homoscédasticité				
MOY OBS X MOY OBS	Séries indépendantes	*n1 et n2 >/= 30*	Test Z de l'écart réduit		
		n1 ou n2 < 30	Pas de normalité	Test de Mann et Whitney	
			Normalité	*Egalité des variances* (Test de snedecor -> non rejet de H0)	Test T de student
			Test Kolmogorov Smirnoff	*Non égalité des variances* (rejet H0 Snedecor)	*Test de Cochran*
	Séries dépendantes	Normalité	Test T de student		
	--> échantillons liés --> séries appariées	Pas de normalité	Test de Wilcoxon		

COMPARAISON DE FREQUENCES : CHOIX DES TESTS

Comparaison d'effectifs	OBS X OBS	*ci >/= 5*	Test du chi2 d'homogénéité
		ci < 5	Regroupement de classe
			Test du chi2 corrigé de Yates
	OBS X THÉO	chi2 d'ajustement	
Comparaison de %	Test Z de l'écart réduit		
Comparaison sur séries appariées	*n2 et n3 >/= 10*	Test du chi2	
	n2 ou n3 < 10	Test de Mac Nemar	

Violet : tests non paramétriques
Vert : tests paramétriques : distribution normale

⚠ *On ne prend en compte que les valeurs discordantes (n2 n3)*

CORRELATIONS et REGRESSIONS

CORRELATION : EVALUER LA LIAISON ENTRE 2 VARIABLES QUANTITATIVES

<table>
<tr><th>Test paramétrique de PEARSON</th><th>Test non paramétrique de SPEARMAN</th></tr>
<tr><td>Estimation de ρ (coef de corrélation linéaire) par r :
• Si ρ < 0 : Les variables sont liées et évoluent en sens opposé
• Si ρ > 0 : Les variables sont liées et évoluent dans le même sens
• Si ρ = 0 : Les variables sont indépendantes (-1 ≤ p ≤ 1)</td><td>Coef de corrélation r_s sur les rangs

Intensité du lien monotone croissant ou décroissant (suivant le signe du coef).</td></tr>
<tr><td>Bi-normalité Linéaire</td><td>Test de rangs
Monotone croissant ou décroissant</td></tr>
<tr><td>Calcul de r (qui est une estimation de p)</td><td>Calcul de r_s</td></tr>
<tr><td>Calculatrice*[1]
Data L1 pour la première ligne et L2 pour la deuxième
*Pour supprimer vos données appuyer 2 fois sur Data Clear L1.
• Faire stat-reg/distr puis 2-Var Stats
• Choisir L1 pour xDATA, L2 pour yDATA et ONE pour FRQ puis CALC
• Descendre : on trouve r
$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{(\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2)(\sum_{i=1}^{n} {y_i}^2 - n\bar{y}^2)}}$</td><td>Etablir des rangs pour les 2 colonnes. Ce sont les rangs de chaque valeur qu'il faudra désormais rentrer.
* Si on trouve plusieurs valeurs identiques : Faire la moyenne de leurs rangs
Calculatrice
Rentrer ces rangs comme données dans la calculatrice :
• Suivre la même méthode*[1]
Descendre : on trouve r_s</td></tr>
<tr><td colspan="2">Test Statistique : Est-ce que r est significativement ≠ de 0 ?
H_0 : ρ = 0 et H_1 : ρ ≠ 0
Statistique de test $t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
t_α sur la table de Student à (n-2) ddl au risque α
• Si $|t_0| \geq t_\alpha$: On rejette H_0. Il existe un lien linéaire
• Si $|t_0| < t_\alpha$: On ne rejette pas H_0. On ne peut pas affirmer de lien linéaire.</td><td>$n \leq 10$
On compare directement r_s à la valeur du coef de Sperman à n ddl

$n > 10$
$t_0 = \frac{r_s\sqrt{n-2}}{\sqrt{1-{r_s}^2}}$
Table de Student à n-2 ddl.

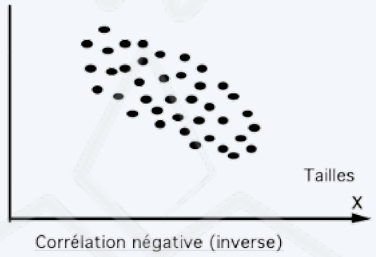
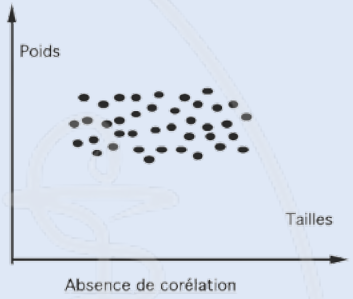
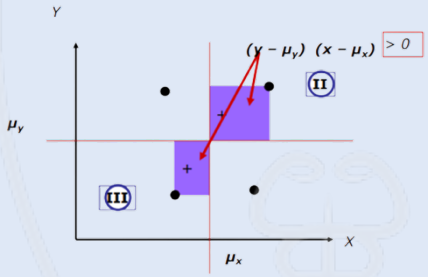
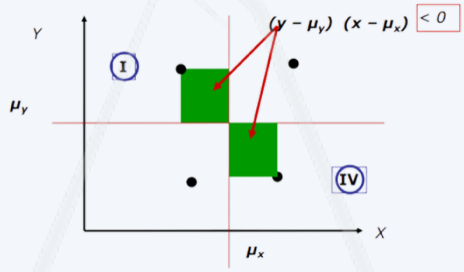
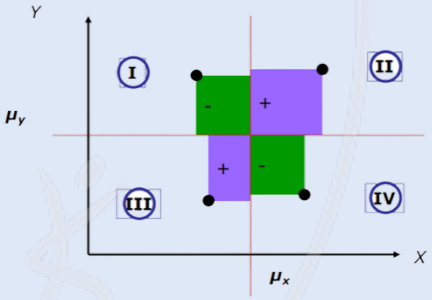
Test Bilatéral : 1α
Test unilatéral : Faire ×2 sur α</td></tr>
</table>

REGRESSION LINEAIRE : voir si la courbe des variations de y en fonction de x est une droite

$b = \bar{y} - a\bar{x}$	$a = r\frac{Sy}{Sx}$
Test : est-ce que a est différent de 0 ? $H_0 : a = 0$; $H_1 : a \neq 0$	
$t_0 = \frac{a\sqrt{n-2}}{\sqrt{\frac{{S_y}^2}{{S_x}^2} - a^2}}$	
Table de Student à $n-2$ ddl	

NOTION DE COVARIANCE :

- Indicateur de mesure de **liaison entre 2 variables X et Y**. (= mesure de la variation simultanée)
- Moyenne des produits des écarts entre X et Y divisée par leurs moyennes respectives µx et µy
- La covariance peut être illustrée par la somme des aires des rectangles

Diagramme de dispersion (nuage de points de coordonnées (x ; y)		
Corrélation positive (directe)	Corrélation négative (inverse)	Absence de corrélation
		 *Nuage de points diffus <=> Absence de l° entre les caractères étudiés*
 *Les rectangles Mauves dans les zones II et III représentent les produits positifs.*	 *Les rectangles verts dans les zones I et IV représentent les produits négatifs*	
$Cov > 0 \Leftrightarrow \sum aires > 0$	$Cov < 0 \Leftrightarrow \sum aires < 0$	$Cov = 0 \Leftrightarrow \sum aires = 0$
X et Y évoluent dans le même sens **=> plus** $X \nearrow$ *plus* $Y \nearrow$	X et Y évoluent dans le sens opposé **=> plus** $X \nearrow$ *plus* $Y \searrow$	X et Y indépendantes *=> les aires se compensent : les points sont répartis uniformément dans les quadrants*
Cas du test unilatéral à **droite** : Test $H_0 : \rho = 0$ **contre** H_1 **: ρ > 0** *Rejeter* H_0 *revient alors à mettre en évidence une corrélation linéaire positive.*	Cas du test unilatéral à **gauche** : Test $H_0 : \rho = 0$ **contre** H_1 **: ρ < 0** *Rejeter* H_0 *revient alors à mettre en évidence une corrélation linéaire négative*	

POINTS IMPORTANTS A MAITRISER :

→ Partie corrélation

1. **Analyser un graphique** pour identifier le lien entre les variables et savoir quel calcul de coefficient (Pearson / Spearman) est suggéré.
2. **Calculer les coefficients :**
    - **À partir de la formule**, c'est ce que l'on retrouve désormais le plus souvent à l'examen.
        - *(Cf annales 2020-2021 et 2021-2022)*
    - /!\ Sachez aussi le faire **à partir du tableau** grâce à la calculatrice (comme dans l'exemple ici)
        - *(Cf méthodo calculatrice)*
3. **Effectuer le test**
    - /!\ Être très vigilant avec les ddl et les risques /!\
4. **Interpréter le test**

→ Partie régression

1. **Calculer le coefficient de régression**
2. **Calculer l'ordonnée à l'origine**
3. **Faire le test**

NB : le test de la **pente** α de la droite de régression est strictement le même que le test du coefficient de Pearson.

4. **Interpréter le test**
5. **Identifier si un point appartient ou non à la droite de régression.**

EXERCICE TYPE CORRIGE : CORRELATIONS REGRESSIONS

Reprise d'un exercice de l'annale 2014-2015 :

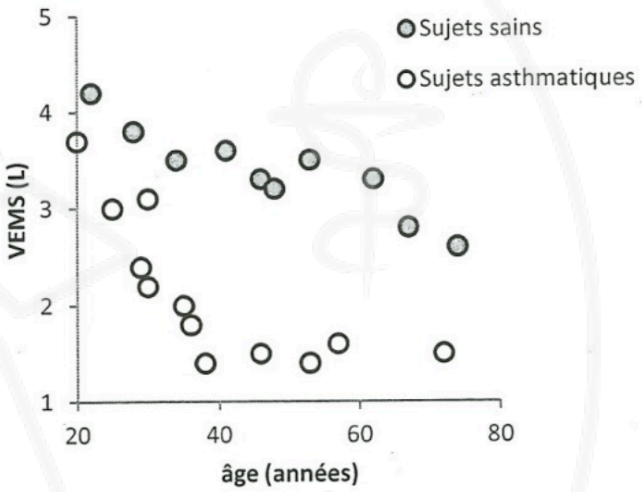
On a mesuré le **VEMS** (*Volume Expiratoire Maximal Seconde*) chez 10 sujets sains et 12 sujets souffrant d'asthme. On s'intéresse au lien entre le VEMS et **l'âge** dans chacune des populations (sujets sains et sujets asthmatiques). Les mesures sont présentées dans le tableau et le graphique ci-dessous.

Le VEMS est distribué selon une loi normale dans chaque population (sujets sains et sujets asthmatiques). Chez les sujets sains les conditions d'application de l'analyse par régression linéaire sont satisfaites. Chez les sujets asthmatiques le diagramme de dispersion fait douter de la linéarité du lien entre VEMS et âge.

Dans la suite toutes les valeurs numériques proposées sont telles que les âges sont exprimés en années et les VEMS en litres.

Sujets sains	
Âge (années)	VEMS (L)
22	4,2
28	3,8
34	3,5
41	3,6
46	3,3
48	3,2
53	3,5
62	3,3
67	2,8
74	2,6

Sujets asthmatiques	
Âge (années)	VEMS (L)
20	3,7
25	3
29	2,4
30	2,2
30	3,1
35	2
36	1,8
38	1,4
46	1,5
53	1,4
57	1,6
70	1,5



QCM 1 : Corrélation Chez les sujets asthmatiques.

A. Le doute émis sur la linéarité du lien entre le VEMS et l'âge impose le choix d'un test paramétrique pour évaluer l'existence d'un lien entre ces variables.
B. Le coefficient r_s de Spearman vaut 0,865 (à 0,001 près).
C. La statistique du test sur le coefficient de Spearman vaut 4,87 en valeur absolue (à 0,01 près).
D. Au seuil de risque 5%, on conclut qu'il existe un lien monotone décroissant entre les variables. La valeur de la table de Student utilisée pour faire cette conclusion est 1,812.
E. Au seuil de risque 5%, on conclut qu'il n'existe pas de lien monotone entre les variables. La valeur de la table de Student utilisée pour faire cette conclusion est 2,228.
F. Toutes les réponses sont inexactes.

Correction QCM1 : D

A. FAUX, un test NON paramétrique.
B. FAUX, le coefficient r_s de Spearman vaut -0,865 (à 0,001 près). **/!\ SIGNES /!**
C. FAUX : $|t_0| = 5,44$.
D. VRAI.
E. FAUX.

QCM 2 : Régression : Chez les sujets sains, concernant la droite de régression du VEMS en fonction de l'âge :

A. La pente de la droite vaut -0,025.
B. L'ordonnée à l'origine de la droite vaut 161,46 (à 0,01 près).
C. La droite de régression linéaire permet de prédire le VEMS pour un âge donné.
D. La droite de régression linéaire passe par le pont de coordonnées (47,5 ; 3,38).
E. La droite de régression linéaire passe par le point de coordonnées (22 ; 4,2).
F. Toutes les réponses sont inexactes.

METHODE DE RESOLUTION :

Choix du test

Conditions d'application des tests des coefficients :

- **Pearson** (*paramétrique*) : au moins une de ces 2 conditions doit être remplie :
  - La distribution du couple de variables aléatoires (X ; Y) doit être **bi-normale**, c'est-à-dire que pour chaque valeur de X, la distribution de Y est normale et vice-versa.
  - La distribution de **Y est normale** et de **variance constante** pour chaque valeur de X ou vice versa ET la relation entre **X et Y est linéaire**.

*NB : Cela est difficile à vérifier en pratique**, le plus souvent on vérifie la normalité de X ou de Y et le caractère monotone de leur relation** (examen visuel des données).*

Le test est robuste c'est à dire qu'il accepte quelques écarts à ces conditions d'applications.

- **Spearman** (*non-paramétrique*) : Aucune
  - On réalise donc le test de Spearman quand il est impossible de remplir les conditions du test de Pearson ou qu'on ne peut pas les vérifier.

Méthode pour calculer le coefficient de Spearman

Pour réaliser le test de **Spearman**, on n'utilise pas les paramètres des 2 variables contrairement à celui de Pearson, mais on **attribue des rangs** à chaque valeur en les classant par ordre croissant ce qui donne le tableau suivant.

Nb : *si on a 2 valeurs identiques, on fait la moyenne.*

- Ici, le 4ème et le 5ème ont 30, donc les 30 ont un rang de $\frac{(4+5)}{2} = \mathbf{4,5}$

On rentre ensuite dans la **calculatrice**, non pas les valeurs, mais les rangs attribués (cf méthode du livret).

La calculatrice nous donne alors les différents paramètres qui nous intéressent, dont le **coefficient** r_s de Spearman.

Age (années)	Rang	VEMS(L)	Rang
20	1	3,7	12
25	2	3	10
29	3	2,4	9
30	**4,5**	2,2	8
30	**4,5**	3,1	11
35	6	2	7
36	7	1,8	6
38	8	1,4	1,5
46	9	1,5	3,5
53	10	1,4	1,5
57	11	1,6	5
70	12	1,5	3,5

/!\ Bien prendre le tableau concernant les asthmatiques /!

Effectuer le test de Spearman

$n > 10$, donc on applique la formule vue ci-dessus avec $t_0 = \frac{r_s\sqrt{n-2}}{\sqrt{1-{r_s}^2}}$ Table **de student** $(n - 2ddl)$

- $r_s = -0{,}865$ et $n = 12$
- Au risque $2 * \alpha$ Unilatéral

Alors, $|t_0|$ **= 5,44**

Interpréter le test

$n > 10$, on réalise donc un **test unilatéral**.

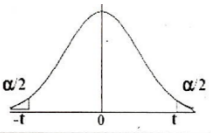
Pour trouver la valeur dans la table, comme $n > 10$, on prend :

- **n-2ddl**, donc 12-2 = **10ddl**
- Au risque **2α**, soit **2 x 0,05 = 0,1**

Table du *t* de Student

La table donne la probabilité α pour que *t* égale ou dépasse, en valeur absolue, une valeur donnée, en fonction du nombre de degrés de liberté (d.d.l.).

2α



ddl \ α	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,599
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
n-2ddl → 10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437

- On a donc $t_{2\alpha}$ **= 1,812**
- $|t_0| > t_{2\alpha}$ 5,44 > **1,812**

On rejette H_0, il existe bien un lien **monotone** (croissant / décroissant).

De plus, le coefficient de Spearman étant négatif, c'est donc un lien **monotone <u>décroissant</u>**

Correction QCM2 : ACE

A. **VRAI**. Utiliser la calculatrice, il faut rentrer dans un tableau les 2 variables concernant les sujets sains, de la même manière que pour les coefficients de Pearson et Spearman. Puis, il faut identifier *a* dans les valeurs proposées.
B. **FAUX**, $b = \bar{y} - a\bar{x}$
On prend les valeurs moyennes de x et y qu'on obtient en rentrant le tableau dans la calculatrice de la même manière que vu précédemment.
Donc b = 3,38 - (-0,025) x 47,5 = 4,569
C. **VRAI**, grâce à l'équation y = ax + b avec a pente de la droite et b l'ordonnée à l'origine, soit y = -0,025x + 4,569 avec x l'âge et y le VEMS.
Par exemple, pour une personne âgée de 32 ans, on peut évaluer que le VEMS sera de 3,77 (à 0,01 près). y = ax + b = -0,025 x 32 + 4,569 = 3769
D. **FAUX**, pour x = 22, on a y = 4,02 car y = -0,025 x 22 + 4,569
E. **VRAI**, pour x = 32, on a y = 3,77 car y = -0,025 x 32 + 4,569

QCM BONUS : Questions de cours les plus fréquentes

A. Le coefficient directeur de la droite de régression est inversement proportionnel au coefficient de corrélation linéaire (Pearson)
B. La droite de régression a pour particularité d'avoir une ordonnée à l'origine nulle.
C. La corrélation permet d'évaluer 2 variables qualitatives.
D. Le coefficient de corrélation de Spearman se situe toujours dans l'intervalle [-1 ; 1].
E. Le coefficient de corrélation de Spearman est toujours égal au coefficient de corrélation de Pearson.
F. Le calcul du coefficient de corrélation de Pearson repose sur la notion de covariance.

Correction QCM Bonus : ADF

A. VRAI. $a = \frac{s_y}{s_x}$ ou $r = a\frac{s_x}{s_y}$

B. FAUX. $b = \bar{y} - a\bar{x}$
C. FAUX, quantitatives.
D. VRAI, comme celui de Pearson.
E. FAUX, il prend en compte les rangs, donc le résultat change parfois.
F. VRAI, cette formule est à connaître par ❤️.

L'estimation du coefficient de corrélation **est donnée par :**

$$r = \frac{cov(x,y)}{\sqrt{{s_x}^2 {s_y}^2}}$$

$$r = \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{(\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2)(\sum_{i=1}^{n} {y_i}^2 - n\bar{y}^2)}}$$

NOTES PERSONNELLES

EPIDEMIOLOGIE

Attention : les QCM sur ce chapitre peuvent porter sur le cours, ne vous contentez pas d'apprendre les formules.

MESURES EN EPIDEMIOLOGIE :

	Définition	Numérateur & Dénominateur	Résultat	Exemple
Proportion	Similaire à la fréquence. *Soit A = population et B une partie de A,* La proportion **de** B **par rapport à** A est définie par : $= \frac{b}{a}$ $= \frac{effectif\ de\ B\ (nb\ d'éléments\ de\ B)}{effectif\ de\ A\ (nb\ d'éléments\ de\ A)}$	Le numérateur est <u>une partie</u> du dénominateur. Ils sont de MÊME nature.	Pourcentage Compris entre 0 et 1 $(a > b)$.	*Dans une population de 490 élèves, 30 sont en antenne. Quelle est la* **proportion** *d'élèves en antenne ?* *30/490 = 0,061 = 6,1%*
Ratio	Rapport entre les effectifs de **2 classes** d'<u>une même variable</u>. *Soit A, une population et B, une seconde population, toutes deux faisant partie d'un même groupe.* Le ratio de B par rapport à A est défini par : $\frac{b}{a} = \frac{effectif\ de\ B}{effectif\ de\ A}$	Le numérateur et le dénominateur sont <u>exclusifs</u> l'un de l'autre. Ils ne font pas partie de la même classe. Ils sont de MÊME nature.	Sans unité	*Dans une population de 210 élèves, 150 sont des PASS et 60 sont des LAS.* *Quel est le* **ratio** *PASS/LAS ?* *150/60 = 2,5*

Cote	**Ratio de la probabilité** de **survenue** d'un évènement sur la probabilité de **non-survenue** de cet événement. $\simeq Pari$ La cote d'un événement est définie par $\frac{P(survenue\ de\ l'événement)}{P(non\ survenue\ de\ l'événement)}$	Le numérateur et le dénominateur sont **exclusifs** l'un de l'autre. Ils ne font pas partie de la même classe. Ils sont de MÊME nature.	Probabilité Compris entre 0 et 1 ($a > b$).	*Dans une population de 100 PASS, il y a 20 places pour passer en 2ème année.* *Quelle est la* **côte** *pour passer en 2ème année ?* *20/80 = 0,25*
Indice	**Rapport** de 2 effectifs qui sont **de natures différentes.** Soit A une population et B une seconde population d'objets distincts. L'indice de B par rapport à A est défini par : $\frac{b}{a} = \frac{effectif\ de\ B}{effectif\ de\ A}$	Le numérateur et le dénominateur sont exclusifs. Ils ne sont pas de la même nature.	Sans unité	*Dans un hôpital, on a 5 médecins et 20 lits. Quel est* **l'indice** *du nombre de lits par médecin ?* *20/5 = 4*
Taux	**Rapport** qui prend en compte la notion du **temps**, c'est-à-dire la probabilité qu'un évènement se réalise au cours du temps. Le taux se définit par : $\frac{Effectif^{\ qui\ a\ connu\ un\ événement}}{Effectif^{\ susceptible\ de\ connaître\ l'événement}}$	Le <u>dénominateur</u> est l'ensemble des individus **susceptible de connaître** cet événement au cours du temps. Le <u>numérateur</u> sont les individus **qui ont connu** cet événement **au cours du temps**.	Sans unité	**Taux** *brut de natalité =* $\frac{Nb^{\text{ naissances viables de l'année}}}{Pop^{\text{au milieu de l'année}}}$ **Taux** *global de fécondité =* $\frac{Nb^{\text{annuel de naissances}}}{Nb^{\text{femmes en âge de procréer}}}$

INDICATEURS EPIDEMIOLOGIQUES :

	Morbidité = fréquence de la maladie	Mortalité = fréquence des décès
À l'instant t	**Prévalence** : $\frac{nombre\ de\ cas_{observé\ à\ l'instant\ t}}{population\ à\ risque\ _{à\ l'instant\ t}}$ Elle ne prend pas en compte la durée de la maladie (malade depuis 3 jours ou 3 mois, pas de différence). Elle est utilisée pour les maladies chroniques ++.	**Létalité :** $\frac{nombre\ de\ décès_{dû\ à\ une\ maladie\ X}}{population_{atteinte\ de\ la\ maladie\ X}}$ Elle décrit la gravité de la maladie et la qualité de sa prise en charge.
Sur une période	**Incidence cumulée :** $\frac{nb\ de\ nouveaux\ cas_{sur\ une\ période\ \Delta t}}{population\ à\ risque_{sur\ une\ période\ \Delta t}}$ Plus la maladie est contagieuse, plus l'IC est grand.	**Mortalité globale :** $\frac{nb\ de\ décès_{sur\ une\ période\ \Delta t}}{population}$ Correspond à l'incidence cumulée mais pour les décès. **Mortalité spécifique** : $\frac{nb\ de\ décès_{dû\ à\ la\ cause\ X\ sur\ une\ période\ \Delta t}}{population\ _{étudiée\ sur\ une\ période\ \Delta t}}$ Elle peut être également sur la mortalité d'une certaine classe d'âge (mortalité infantile). **Mortalité proportionnelle** : $\frac{nb\ de\ décès\ dû\ à\ la\ cause\ X_{sur\ une\ période\ \Delta t}}{nb\ de\ décès_{sur\ une\ période\ \Delta t}}$ Elle permet de rendre compte du poids d'une maladie dans le nombre de décès.

NOTION D'EPIDEMIOLOGIE :

On a deux types d'enquêtes :

- Les **enquêtes descriptives**
- Les **enquêtes étiologiques**

Les enquêtes descriptives

Objectifs : avoir des données sur la maladie à un instant T. Donc on va soit utiliser **l'incidence** soit la **prévalence**.

Les enquêtes étiologiques

Objectifs : établir des liens entre les maladies et les facteurs de risques (*est-ce que fumer provoque plus de cancer du poumon).*

Donc dans ces enquêtes ont va surtout comparer des groupes de personnes (non-fumeur et fumeur).

ATTENTION : on n'établit pas de relation causale mais une présomption c'est-à-dire que l'on n'est jamais certain à 100% mais que l'on le pense fortement.

Il existe 3 types d'enquêtes étiologiques établies selon la chronologie : **prospective, rétrospective, transversale.**

Prospective (cohorte) :

- Mesure de l'exposition avant l'apparition de la maladie :
- Exposition → maladie ?
- Non exposition → maladie ?

Rétrospective (cas-témoins ou cohorte) :

- Recherche de l'exposition après constat (ou non) de la maladie
- Malades → exposition ?
- Non malades → exposition ?

Transversale :

- Mesure de l'exposition et de la maladie simultanément

Enquête de cohorte (prospective ou rétrospective) ou enquête exposé non-exposé

Dans les enquêtes de cohorte, on va suivre les sujets sur le temps avec des **sujets non-exposés** et des **sujets exposés**. On aura donc à la fin de l'enquête, les **incidences** de la maladies parmi les exposés et les non-exposés.

Présentation des données :

	Malade	Non Malade
Exposés	a	b
Non exposés	c	d

Mesures :

- **Incidence** : risque de survenue de maladie dans chaque groupe

$$I_e = \frac{a}{(a+b)}$$

$$I_{ne} = \frac{c}{(c+d)}$$

- **Différence de risque** entre exposés et non-exposés : $DR = I_e - I_{ne}$
- **Risque relatif (RR)** : rapport entre l'incidence chez les exposés sur l'incidence chez les non-exposés : **RR** = $\frac{I_e}{I_{ne}}$

Le **RR** est un nombre sans unité compris entre 0 et l'infini. Un RR "nul" est égal à 1. Plus le RR est éloigné de 1, plus l'association entre la survenue de la maladie et la présence du facteur étudié est forte.

Interprétation de l'intervalle de confiance du RR à 95%

(L'intervalle sera donné pour ce type d'exercice) :

- S'il contient 1 : pas de relation démontrée entre la maladie et l'exposition
- S'il est > 1 : excès de risque dans le groupe exposé
- S'il est < 1 : moins de risque de contracter la maladie s'il y a exposition (facteur protecteur)

Enquête cas-témoins (rétrospective)

Les études cas-témoins sont des études **rétrospectives** : c'est-à-dire que l'on fait l'enquête une fois que les sujets sont malades. On réalise l'enquête avec des sujets sains (sujet témoins) et des sujets malades et on regarde leur **fréquence d'exposition** à la possible cause de leur maladie. Le but final étant de savoir si l'exposition est à l'origine potentielle de la maladie.

Présentation des données :

	Cas *(malade)*	**Témoins *(non malade)***
Exposés	a	b
Non exposés	c	d

Mesures :

- On a choisi le nombre de témoins : on ne peut pas calculer l'incidence
- On calcule les fréquences d'exposition au facteur :

$$Exp_c = \frac{a}{(a+c)}$$

$$Exp_t = \frac{b}{(b+d)}$$

- On peut comparer les fréquences avec le χ^2 (*voir fiche comparaison de fréquence*).

Les cotes d'exposition :

- La cote d'exposition chez les cas : $\frac{a}{c}$
- La cote d'exposition chez les témoins : $\frac{b}{d}$

Rapport des cotes : **Odd ratio** (**OR**) : permet d'évaluer la liaison entre l'exposition et la maladie.

$$OR = \frac{\left(\frac{a}{c}\right)}{\left(\frac{b}{d}\right)}$$

OR : nombre sans unité entre 0 et l'infini :

- *OR "nul" : 1*
- *Plus OR est éloigné de 1 : plus l'association entre la maladie et la présence du facteur étudié est forte*

Interprétation de l'intervalle de confiance du RR à 95% :

- 1 est compris dans l'IC : pas de relation démontrée entre la maladie et l'exposition
- OR significativement > 1 (borne inf de l'IC >1) : fréquence d'exposition supérieure chez les cas
- OR significativement < 1 (borne sup de l'IC <1) : Exposition : facteur protecteur

RESUME

	Enquête Cohorte	Enquête Cas-témoins
Avantages	Bien adaptée pour étudier : • Les risques (incidences) • Des expositions rares • Plusieurs maladies • La séquence exposition-maladie	Bien adaptée pour étudier : • Des maladies rares • Plusieurs facteurs de risque
	Peu de biais de sélection	Etude courte, coût faible
	Peu de biais de mémorisation	Rapidité d'exécution Échantillon de taille modérée
Inconvénients	Non adaptée pour étudier : • Des maladies rares • Plusieurs expositions	Non adaptée pour étudier : • Des expositions rares • Plusieurs maladies • La séquence temporelle exposition-maladie
	Coût élevé	Pas de calcul de taux d'incidence
	Longue période de latence	OR biaisé si prévalence élevée

	ÉPIDÉMIOLOGIE			
	Enquête par observation ! On ne choisi pas qui est exposé ou non à un FdR ***≠ des études expérimentales*** *(ex : essai thérapeuthique avec groupe exposé (reçoivent mdt) / groupe témoin (reçoivent placébo)*			
	ENQUÊTE ÉTIOLOGIQUE			**ENQUÊTE DESCRIPTIVE**
	Visent à établir un lien entre un facteur d'exposition et la maladie			*Décrire la mortalité/morbidité aux travers d'indicateurs (incidence, prévalence)*
	Cohorte *prospective*	**Cas témoin** *rétrospective*	**Transversale** *silmultanée*	
AVANTAGES	• Etudier les risques (incidences) • Expositions rares • Plusieurs maladies • Séquence expo-maladie • Peu de biais de sélection	• Maladies rares • Plusieurs FdR • Courte durée • Coû faible / faciles • Taille modéré échantillon	• Liaison entre FdR et présence (et non pas survenue) maladie • On ne sait pas si maladie = conséquence ou cause • Moins puissante => indication pour réaliser les autres	
INCONVÉNIENTS	• Maladies rares • Plusieurs expositions • Coût élevé • Longues latences avant résultats	• Expositions rares • Plusieurs maladies • Séquence temporelle peu sur (biais mémorisation) • Pas de calcul d'incidence		

NOTES PERSONNELLES

Page de FIN.

Bon courage à vous en espérant que ce livret vous aidera à performer et à comprendre les biostats !
Je tiens à remercier Lola et Pierre pour la magnifique mise en forme, et mes amis et la tut'team pour les relectures du livret !

Léa JÉGOREL

Tutrice 2023/24

Mot de la fin de Swann :
♥ Impasse UE4 ♥

Merci à toutes les personnes ayant pu permettre la réalisation de ce livret, et à ceux qui, dans le futur, l'entretiendront et le feront vivre. Et pour toi, cher PASS arrivé jusqu'ici, bon courage pour cette année si riche !

Estelle RENARD

Tutrice 2023/24

Avec toute la bienveillance qu'on a mis dans ce livret, un entraînement régulier et des supers tuteurs, vous serez tous de vrais spécialistes des biostats à la fin du semestre même si les maths ne vous passionnent pas !
Rendez-vous en conférence !

Azéline LELIEVRE

Tutrice LAS 2023/24

Merci à tout le monde pour la réalisation de ce livret et bon courage pour ce second semestre.

Timothée LIN

Tuteur LAS 2023/24

Aurélie LIN

Tutrice LAS 2025/26

Bon courage pour vos partiels, et que les biostats soient avec vous !

Héléna VARON

Tutrice LAS 2025/26

Jules ONGHENA

Tuteur 2022/24

Brieuc BOUFFORT

Tuteur 2022/2024

Emelyne ROBERT

VP Design 2023

Lola LE GUELL

VP Site 2023/24

Pierre GELGON

Pilote 2023/24





