Chapitre IV

Statistiques

Tous le reste

Plan du cours :

a

Programme (*BO n° 30 du 26-7-2018*) :

— a

Compétences :

— a

INFOX

LE PARADOXE DE SIMPSON — LES CALCULS RÉNAUX

TROISIEME

Un laboratoire médical vient de mettre au point deux traitements pour éliminer les calculs rénaux. Ces traitements ont été testés sur deux groupes de 350 personnes. Il y a deux types de calculs rénaux : les petits calculs rénaux inférieurs à 2 mm et les gros calculs.
Voici les résultats des tests effectués avec ces deux traitements :

Test du Arnaclam 300 mg

	Réussite	Échec	Total
Petits calculs	81	6	
Gros calculs	192		
Total			

Test du Arnodix 750 mg

	Réussite	Échec	Total
Petits calculs		36	
Gros calculs	55		80
Total			

1. Compléter les deux tableaux de résultats en utilisant les données de l'énoncé.

2. Compléter le tableau suivant en calculant en pourcentage à l'unité près les taux de réussite.

	Taux de réussite Arnaclam 300 mg	Taux de réussite Arnodix 750 mg
Petits calculs rénaux		
Gros calculs rénaux		
Total		

3.a Quel est le traitement le plus efficace sur les petits calculs rénaux?
3.b Quel est le traitement le plus efficace sur les gros calculs rénaux?
3.c Quel est le traitement le plus efficace sur l'ensemble de tous les calculs rénaux?

4. Que pensez-vous de cette situation?

La plus lointaine mention d'un cas analogue remonte à 1899, où le mathématicien anglais Karl Pearson décrivit des données équivalentes. Plus tard en 1903, Undy Yule redécouvrit le phénomène et le Britannique Edward Simpson écrivit en 1951 un article où cette singularité statistique était soigneusement étudiée et discutée.
De nombreux cas réels présentent cette inversion de résultat lorsqu'on regroupe plusieurs catégories complémentaires en une seule. De nombreux cas en médecine ont été rapportés. Le paradoxe a aussi été rencontré en démographie, dans l'analyse de match de basket-ball, dans l'étude de risque d'accidents... [4] [1]

INFOX

LE PARADOXE DE SIMPSON — CIGARETTES ET MORTALITÉ

TROISIEME

En 1972, à Whickham, une ville du nord-est de l'Angleterre, un sondage a été effectué afin d'éclairer des travaux sur les maladies cardiaques (Tunbridge et al. 1977). Une suite de cette étude a été menée vingt ans plus tard (Vanderpump et al. 1995). Certains des résultats avaient trait au tabagisme et cherchaient à savoir si les individus étaient toujours en vie lors de la seconde étude. Par simplicité, nous nous restreindrons aux femmes et parmi celles-ci aux 1314 qui ont été catégorisées comme « fumant actuellement »ou « n'ayant jamais fumé ». La survie à 20 ans a été déterminée pour l'ensemble des femmes du premier sondage.

Voici quelques résultats :

18-34 ans	Fumeuses	Non-fumeuses	**Total**
En vie	174	213	387
Décédée	5	6	11
Total	179	219	398

35-50 ans	Fumeuses	Non-fumeuses	Total
En vie	159	145	304
Décédée	36	16	52
Total	195	161	356

51-64 ans	Fumeuses	Non-fumeuses	**Total**
En vie	103		
Décédée		43	
Total		159	318

64 ans et plus	Fumeuses	Non-fumeuses	**Total**
En vie		28	
Décédée	42		
Total	49		

1. Compléter les deux tableaux restants en tenant compte des informations fournies.

2. Pour chacune des quatre tranches d'âge, calculer le taux de mortalité des fumeuses et des non-fumeuses en pourcentages arrondis au dixième près.

Le taux de mortalité est le ratio entre le nombre de personnes décédées et le nombre total de personnes considérées.

3. Que constatez-vous en comparant ces taux de mortalité pour chaque tranche d'âge.

4. En cumulant les données de ces quatre tableaux, déterminer le taux de mortalité des fumeuses et des non-fumeuses sur l'ensemble des 1313 femmes intérrogées.

5. Complétez le tableau de synthèse suivant :

Taux de mortalité en pourcentage	Fumeuses	Non-fumeuses
18-34 ans		
35-50 ans		
50-64 ans		
64 ans et plus		
Ensemble		

6. Que remarquez-vous? Comment pouvez-vous expliquer ce résultat?

La plus lointaine mention d'un cas analogue remonte à 1899, où le mathématicien anglais Karl Pearson décrivit des données équivalentes. Plus tard en 1903, Undy Yule redécouvrit le phénomène et le Britannique Edward Simpson écrivit en 1951 un article où cette singularité statistique était soigneusement étudiée et discutée.

De nombreux cas réels présentent cette inversion de résultat lorsqu'on regroupe plusieurs catégories complémentaires en une seule. De nombreux cas en médecine ont été rapportés. Le paradoxe a aussi été rencontré en démographie, dans l'analyse de match de basket-ball, dans l'étude de risque d'accidents... [4] [1]

Le paradoxe de Simpson — Cigarettes et mortalité —

Correction

Infox

18-34 ans	Fumeuses	Non-fumeuses	**Total**
En vie	174	213	387
Décédée	5	6	11
Total	179	219	398

35-50 ans	Fumeuses	Non-fumeuses	Total
En vie	159	145	304
Décédée	36	16	52
Total	195	161	356

51-64 ans	Fumeuses	Non-fumeuses	**Total**
En vie	103	116	219
Décédée	56	43	99
Total	159	159	318

64 ans et plus	Fumeuses	Non-fumeuses	**Total**
En vie	7	28	35
Décédée	42	165	207
Total	49	193	242

1. Compléter les deux tableaux restants en tenant compte des informations fournies.

2. Pour chacune des quatre tranches d'âge, calculer le taux de mortalité des fumeuses et des non-fumeuses en pourcentages arrondis au dixième près.

Les 18-34 ans

Fumeuses : $\frac{5}{179} \approx 0,03$ soit 3 %

Non-fumeuses : $\frac{6}{219} \approx 0,03$ soit 3 %

Les 35-50 ans

Fumeuses : $\frac{36}{195} \approx 0,18$ soit 18 %

Non-fumeuses : $\frac{16}{161} \approx 0,10$ soit 10 %

Les 51-64 ans

Fumeuses : $\frac{56}{159} \approx 0,35$ soit 35 %

Non-fumeuses : $\frac{43}{159} \approx 0,27$ soit 27 %

Les 64 ans et plus

Fumeuses : $\frac{42}{49} \approx 0,86$ soit 86 %

Non-fumeuses : $\frac{165}{193} \approx 0,85$ soit 85 %

3. Que constatez-vous en comparant ces taux de mortalité pour chaque tranche d'âge.

Globalement le taux de mortalité est supérieur pour les fumeuses que les non-fumeuses.
Pour la population jeune ou très âgée les taux sont similaires pour des raisons simples à comprendre.

4. En cumulant les données de ces quatre tableaux, déterminer le taux de mortalité des fumeuses et des non-fumeuses sur l'ensemble des 1313 femmes interrogées.

Total des fumeuses : $179 + 195 + 159 + 49 = 582$ et nombre de décès dans cette population : $5 + 36 + 56 + 42 = 139$.

Total des non-fumeuses : $219 + 161 + 159 + 193 = 732$ et nombre de décès dans cette population : $6 + 16 + 43 + 165 = 230$.

Le taux de mortalité chez les fumeuses : $\frac{139}{582} \approx 0,24$ soit 24 %.

Le taux de mortalité chez les non-fumeuses : $\frac{230}{732} \approx 0,31$ soit 31 %.

5. Complétez le tableau de synthèse suivant :

Taux de mortalité en pourcentage	Fumeuses	Non-fumeuses
18-34 ans	3 %	3 %
35-50 ans	18 %	10 %
50-64 ans	35 %	27 %
64 ans et plus	85 %	86 %
Ensemble	24 %	31 %

6. Que remarquez-vous ? Comment pouvez-vous expliquer ce résultat ?

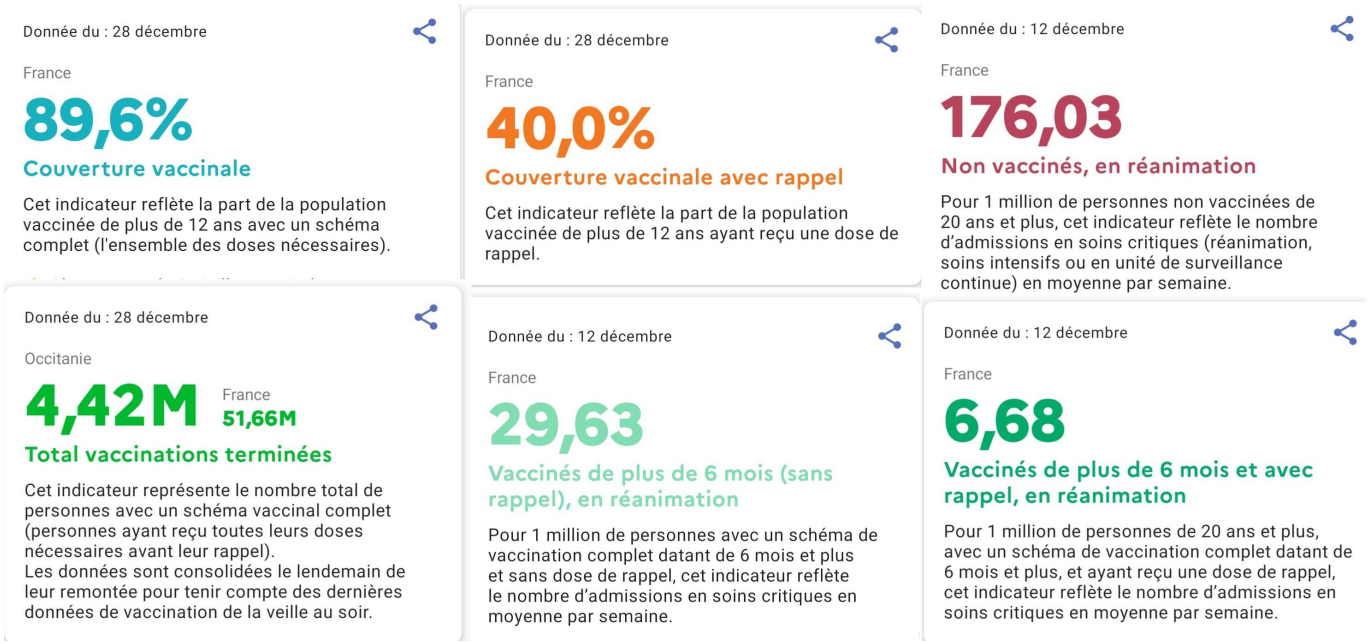
On constate que sur la globalité...

LE PARADOXE DE SIMPSON — COVID ET VACCIN

TROISIEME

INFOX

Voici les chiffres fournis par l'application TousAntiCovid à la date du 29 décembre 2021 :



Sur les réseaux sociaux, certains Anti-vaccins prétendent « qu'il y a plus de vaccinés que de non vaccinés en réanimation aux urgences de l'hopital ». Cela prouverait que le vaccin est inefficace ce qui irait dans le sens de leur opinion.

On souhaite utiliser les informations de l'application TousAntiCovid pour vérifier cette information.

1. Déterminer l'effectif de la part de la population française de plus de 12 ans.

2. Déterminer le nombre de personnes non vaccinés et le nombre de ceux ayant reçu une dose de rappel.

3. Compléter le tableau suivant :

	Effectif total	Effectif en réanimation	Fréquence (%)
Vaccinés			
Dont ayant reçu une dose de rappel			
Non vaccinés			
Total			

4. Que peut-on déduire de ces informations et que peut-on répondre au sujet de cette information?

LE PARADOXE DE SIMPSON — COVID ET VACCIN — Correction

INFOX

Les données fournies par l'application TousAntiCovid mélangent des informations datant du 28 décembre 2021 et d'autres (les taux de malades en réanimation) qui datent du 12 décembre. Il y a donc un biais dans les calculs que nous allons faire. À cette date l'épidémie est dans une phase exponentielle. On peut donc imaginer que nos résultats vont être sous-évalué.

1. On constate que 89,6 % de la population des plus de 12 ans correspond à 55,66 M de personnes soit 55 660 000.

Vaccinés	89,6	55,66 M
Total	100	$\dfrac{55{,}6\text{ M} \times 100}{89{,}6} \approx 62{,}05\text{ M}$

L'échantillon de la population concernée par ces statistiques correspond à 62,05 M soit 62 050 000 de personnes.

2. Comme 89,6 % des personnes sont vaccinés, 100 % − 89,6 % = 10,4 % ne le sont pas.

$\dfrac{10{,}4}{100} \times 62{,}05\text{ M} \approx 6{,}45\text{ M}$ de personnes non vaccinés.

$\dfrac{40}{100} \times 62{,}05\text{ M} = 24{,}82\text{ M}$ de personnes ayant une dose de rappel.

3. Compléter le tableau suivant :

	Effectif total	Effectif en réanimation	Fréquence (%)
Vaccinés	51,66 M	$51{,}66\text{ M} \times \dfrac{29{,}63}{1\text{ M}} \approx 1531$	$\dfrac{1531}{2667} \approx 0{,}57 \approx 57\ \%$
Dont ayant reçu une dose de rappel	24,82 M	$24{,}82\text{ M} \times \dfrac{6{,}68}{1\text{ M}} \approx 166$	$\dfrac{166}{2667} \approx 0{,}06 \approx 6\ \%$
Non vaccinés	6,45 M	$6{,}45\text{ M} \times \dfrac{176{,}03}{1\text{ M}} \approx 1136$	$\dfrac{1136}{2667} \approx 0{,}43 \approx 43\ \%$
Total	62,05 M	1531 + 1136 = 2667	100 %

4. On constate que les vaccinés sont majoritaires en réanimation. 57 % de vaccinés contre 43 % de non-vaccinés.
Cela semble indiquer que le vaccin manque d'efficacité.

Cependant quand on observe les fréquences qui concernent les non-vaccinés, les vaccinés et les doses de rappel, on constate que les non-vaccinés sont présents à $\dfrac{176{,}03}{1\,000\,000}$ dans les réanimations.

En comparant les fréquences on constate que :

— 176,03 non vaccinés sur un million sont en réanimation ;

— 29,63 vaccinés sur un million ;

— 6,68 dose de rappel sur un million.

Comme $\frac{176,03}{29,63} \approx 5,94$: cela signifie qu'il est environ six fois plus probable d'aller en réanimation quand on est vacciné que quand on ne l'est pas !

Comme $\frac{176,03}{6,68} \approx 26,35$: il est environ vingt-six fois plus probable d'aller en réanimation quand on a une dose de rappel que quand on n'est pas vacciné.

Cela semble contredire l'étude des proportions de vaccinés et de non-vaccinés en réanimation !

Cette contradiction apparente est due au fait que la proportion de non vaccinés est faible dans la population. La différence de taille des échantillons entre les vaccinés et les non-vaccinés produit ce paradoxe de type paradoxe de Simpson.

En poussant ce raisonnement à l'extrême, quand 100 % de la population sera vaccinée alors 100 % des personnes en réanimation seront vaccinés (car le vaccin ne garantit pas une totale imunité...).

En effet, la majorité des personnes en réanimation sont des gens vaccinés. Cela ne prouve qu'une seule chose : qu'une majorité de la population est maintenant vaccinée !

LES ÉQUIPES DE BASKET-BALL

TROISIEME

SITUATION INITIALE

Voici la taille en centimètres des joueurs de deux équipes de basket-ball (même s'il y a cinq joueurs sur le terrain en même temps, il faut tenir compte des remplaçants) :

Lakers de la Ramée : 178 cm – 196 cm – 165 cm – 209 cm – 162 cm – 198 cm – 196 cm – 197 cm – 163 cm – 176 cm – 196 cm
Celtics de Tibaous : 183 cm – 185 cm – 181 cm – 187 cm – 196 cm – 183 cm – 176 cm – 188 cm – 207 cm – 185 cm – 165 cm

On souhaite comparer la taille des joueurs de ces deux équipes.

1. Calculer la moyenne des tailles en centimètres de chacune des deux équipes. Que pouvez-vous dire de ces résultats ?

2. Déterminer la plus petite taille, la plus grande taille et l'écart entre la plus petite et la plus grande taille pour chacune des deux équipes. Que pouvez-vous dire de ces résultats ?

3. Pour chacune de ces deux équipes, classer les tailles dans l'ordre croissant. Que pouvez-vous dire de ce classement ?

4. Compléter le tableau suivant :

Analyse des tailles des Lakers de La Ramée

Taille (cm)	[160;170[	[170;180[	[180,190[	[190;200[	[200;210[	Total
Effectif						
Fréquence						

Analyse des tailles des Celtics de Tibaous

Taille (cm)	[160;170[	[170;180[	[180,190[	[190;200[	[200;210[	Total
Effectif						
Fréquence						

Que pouvez-vous en dire ?

5. Voici les tailles de l'équipe des Hornets des Pradettes :

Taille (cm)	[160;170[	[170;180[	[180,190[	[190;200[	[200;210[	Total
Effectif	3	2	1	2	3	
Fréquence						

Complétez ce tableau.
Pouvez-vous comparer cette équipe avec les deux précédentes ?

SITUATION INITIALE

I — Moyenne arithmétique et pondérée

II — Annexes

1 Exercices

Exercice n° 4.1 : Un exercice

STATISTIQUES

VOCABULAIRE

Une **série statistique** est une liste de valeurs obtenues en étudiant une **population** (des élèves, des plantes, des factures...). Pour chaque **individu** de la population étudiée on peut observer un ou plusieurs **caractères** (tailles, masse, âge, prix, couleur...), c'est à dire une information. Un caractère peut être **qualitatif** (couleur, difficulté, goût...) ou **quantitatif** (quantité, nombre, prix...).
On connaît parfois toutes les valeurs d'une série statistiques. Quelquefois on ne connaît que la **répartition** des valeurs étudiées.
L'effectif total d'une série désigne le nombre total d'individu étudié. Dans un tableau de répartition on utilise le mot **effectif** pourle nombre d'individu concerné par une valeur du caractère.
La fréquence d'une valeur du caractère étudié correspond au quotient de l'effectif de ce caractère sur l'effectif total. Une fréquence peut s'exprimer sous forme d'une fraction, d'un pourcentage ou d'un nombre décimal approché ou non.

EXEMPLES :
Voici une première série qualitative : la couleur des yeux de 10 personnes :

Bleu – Bleu – Vert – Vert – Vert – Marron – Marron – Marron – Marron – Noir

Voici une seconde série quantitative : les notes d'un groupe de 9 élèves au diplome de fin d'année :

10 – 05 – 15 – 20 – 11 – 15 – 15 – 03 – 17

Voici une troisième série quantitative : la répartition des notes sur les 156 élèves de dernière année :

Notes	[0;5[	[5;10[	[10;15[	[15;20]
Effectif	26	54	60	16

MOYENNE ARITHMÉTIQUE ET PONDÉRÉE

La **moyenne** ou **moyenne arithmétique** de la série de n valeurs : x_1, x_2, x_3, ... x_n est :

$$\frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

La **moyenne pondérée** de la série de n valeurs x_1, x_2, x_2, ... x_n pondérées par les nombres a_1, a_2, a_3, ... a_n est :

$$\frac{a_1 \times x_1 + a_2 \times x_2 + a_3 \times x_3 + ... + a_n \times x_n}{a_1 + a_2 + a_3 + ... + a_n}$$

La moyenne d'une série statistique est un nombre qui correspond à un partage équitable de toutes les valeurs de la série.

EXEMPLES :
La première série est qualitative, la moyenne n'a pas de sens pour cette série.
La seconde série a pour moyenne :

$$\frac{10+5+15+20+11+15+15+3+17}{9} = \frac{111}{9} \approx 12,33 \text{ à } 0,01 \text{ près.}$$

Pour la troisième série, il faut calculer la moyenne des centres des intervalles pondérée par l'effectif.

$$\frac{2,5 \times 26 + 7,5 \times 54 + 12,5 \times 60 + 17,5 \times 16}{26+54+60+16} = \frac{1500}{156} \approx 9,62 \text{ à } 0,01 \text{ près.}$$

ÉTENDUE

L'étendue d'une série statistique est l'écart entre la valeur maximale et minimale de la série.
L'étendue donne une information sur la dispersion des valeurs de la série : plus l'étendue est petite moins la série est dispersée.

EXEMPLE :
L'étendue de la deuxième série est $20 - 3 = 17$
Pour la deuxième série on peut seulement dire que l'étendue est inférieure ou égale à 20.

MÉDIANE

La **médiane** d'une série statistique est une valeur du caractère qui partage la série en deux séries ayant le même effectif.
La moitié des valeurs sont inférieures à la médiane, l'autre moitié est supérieure.
La médiane donne une information sur la dispersion des valeurs de la série. Son écart avec la moyenne est souvent intéressant.

MÉTHODE :
Pour calculer la médiane d'une série statistique il faut classer les valeurs du caractère dans l'ordre croissant puis déterminer la valeur centrale.

- si l'effectif est impair, $2n+1$, la médiane est la $n+1^{\text{e}}$ valeur ;
- si l'effectif est pair, $2n$, la médiane est la moyenne de la n^{e} et $n+1^{\text{e}}$ valeur.

EXEMPLES :
Pour la deuxième série, l'effectif total est impair : $9 = 2 \times 4 + 1$, la médiane est la $4+1 = 5^{\text{e}}$ valeur soit 15.
Pour la troisième série, l'effectif total est pair : $156 = 2 \times 78$, la médiane est la moyenne de la 78^{e} et 79^{e} valeurs.
D'après le tableau cette médiane se situe dans l'intervalle [5;10[.