

Accromoth

Volume 17 • été-automne 2022



Des dames sur d'étranges échiquiers

Autres articles

- **L'utilisation d'indices pour combiner des informations**
- **Y a-t-il une relation de cause à effet? La science statistique de l'inférence causale**
- **Comparaison d'aires: 2. La méthode d'exhaustion et la méthode du levier**
- **(Ré)apprendre à multiplier avec la méthode de Karatsuba**

Rubrique des
Paradoxes

Le
déménagement
miraculeux

Éditorial α

Dans un article intitulé **Des dames sur d'étranges échiquiers**, Alexis Langlois-Rémillard et Charles Senécal s'intéressent aux nombres minimal et maximal de dames pour contrôler des échiquiers de différentes formes sans se menacer.

Les médias d'information nous présentent souvent des indices : indice des prix à la consommation, indice humidex, indice de Gini, indice de développement humain. Que représentent ces indices? Dans l'article **L'utilisation d'indices pour combiner des informations**, Christiane Rousseau nous présente la façon dont sont conçus certains indices.

Une corrélation entre deux phénomènes ne signifie pas automatiquement qu'il y a relation de cause à effet entre ceux-ci. La relation de causalité est plus difficile à établir, particulièrement lorsqu'il s'agit d'un traitement médical, comme nous en font part Christian Genest et Erica Moodie dans l'article **Y a-t-il une relation de cause à effet? La science statistique de l'inférence causale**.

Pendant plusieurs siècles, la méthode d'exhaustion fut considérée comme la seule méthode rigoureuse pour démontrer l'égalité de deux aires ou de deux volumes. Cette méthode ne permettait cependant pas de trouver le résultat à démontrer. Pour contourner cet écueil, Archimède avait recours à la méthode du levier. C'est le sujet abordé dans l'article **Comparaison d'aires : 2 La méthode d'exhaustion et la méthode du levier**.

Multiplier de grands nombres par la méthode traditionnelle est une tâche qui peut prendre un temps de réalisation et un coût importants, même pour un ordinateur. **(Ré)apprendre à multiplier avec la méthode de Karatsuba**, de Nadia Lafrenière, nous présente le premier algorithme visant à diminuer le temps et le coût d'une telle opération.

Jacques, qui habite la rue Gödel, a 35 ans; il déménage sur la rue Turing. Paradoxalement, après ce déménagement, la moyenne d'âge a diminué sur les deux rues. C'est le problème posé par Jean-Paul Delahaye dans le paradoxe **Le déménagement miraculeux**.

Bonne lecture!

André Ross

Rédacteur en chef

André Ross

Professeur de mathématiques

Comité éditorial

France Caron

*Professeure de didactique
des mathématiques
Université de Montréal*

Christian Genest

*Professeur de statistique
Université McGill*

Frédéric Gourdeau

*Professeur de mathématiques
Université Laval*

Bernard R. Hodgson

*Professeur de mathématiques
Université Laval*

Stéphane Laplante

*Enseignant de mathématiques
Collège de Montréal*

Christiane Rousseau

*Professeure de mathématiques
Université de Montréal*

Robert Wilson

*Professeur de mathématiques
Cégep de Lévis-Lauzon*

Production et Iconographie

Alexandra Haedrich

Institut des sciences mathématiques

Conception graphique

Pierre Lavallée

Néograf Design inc.

Illustrations de scientifiques et caricatures

Noémie Ross

Illustrations mathématiques

André Ross

Révision linguistique

Robert Wilson

*Professeur de mathématiques
Cégep de Lévis-Lauzon*

Accromath

*Institut des sciences mathématiques
Université du Québec à Montréal
Case postale 8888, succ. Centre-ville
Montréal (Québec)
H3C 3P8 Canada*

redaction@accromath.ca

www.accromath.ca

Accromath

Volume 17 • été – automne 2022

Sommaire

Dossier *Application des mathématiques*

Des dames sur d'étranges échiquiers

Alexis Langlois-Rémillard
Charles Sénécal

L'utilisation d'indices pour combiner des informations

Christiane Rousseau

Dossier *Probabilités et statistique*

Y a-t-il une relation de cause à effet?

La science statistique de l'inférence causale

Christian Genest
Erica Moodie

Dossier *Histoire des mathématiques*

Comparaison d'aires : 2. La méthode d'exhaustion et la méthode du levier

André Ross

Dossier *Nombres*

(Ré)apprendre à multiplier avec la méthode de Karatsuba

Nadia Lafrenière

Rubrique des **Paradoxes**

Le déménagement miraculeux

Jean-Paul Delahaye

Solution du paradoxe précédent

Jean-Paul Delahaye

Section problèmes

26

2

2

8

14

20

20

26

30

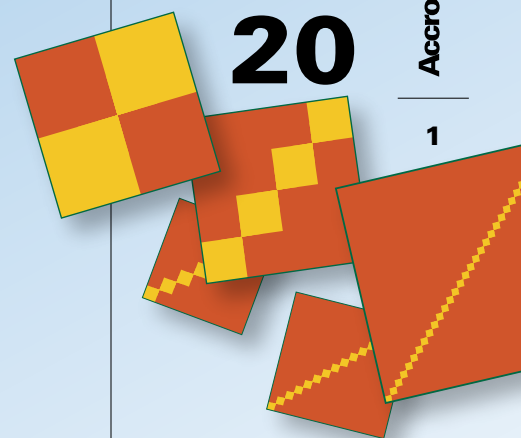
31

32

Vol. 17 • été – automne 2022

Accromath

1



Le jeu des échecs, avec ses règles simples régissant le déplacement des pièces sur son espace clos, a inspiré bien des problèmes ludiques. Certains de ceux-ci ont d'étonnants liens avec des domaines variés des mathématiques. Dans le dialogue entre échecs et mathématiques, ce fut parfois la recherche mathématique qui modifia les règles des échecs pour les faire concorder à des problèmes intéressants.

Alexis Langlois-Rémillard
Université de Gand
Charles Sénécal
Université de Montréal

Un des plus célèbres problèmes mathématico-échiquéens est le *Problème des 8 dames* : on demande à placer 8 dames sur un échiquier 8×8 de sorte qu'aucune dame n'en menace une autre. Un problème similaire

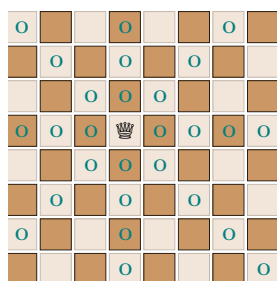


Figure 1: mouvement de la dame sur un échiquier normal.

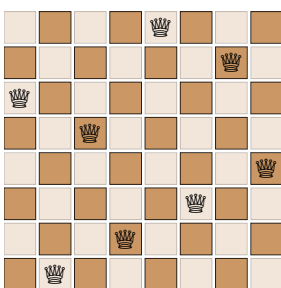


Figure 2: Une solution au problème des 8 dames.

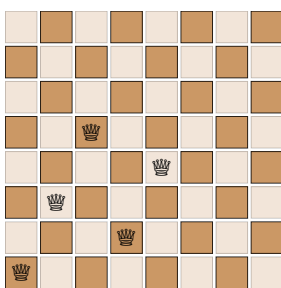


Figure 3: Domination de 5 dames.

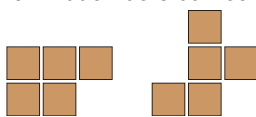


Figure 4: Deux polyominos de cinq cases.

est le problème de *Domination de l'échiquier par les dames* : on y demande le nombre minimal de dames permettant que toutes les cases de l'échiquier soient « surveillées », c'est-à-dire menacées ou occupées par une dame. Ces deux problèmes sont liés. Le premier a trait au nombre maximal de dames dominant l'échiquier sans se menacer; le deuxième, au nombre minimal de dames dominant l'échiquier.

Au fil du temps, les mathématiciennes et mathématiciens étudiant ces problèmes ont aussi tenté de les généraliser et d'ajouter ou enlever des contraintes afin de mieux les comprendre. Au-delà de la première généralisation

agrandissant l'échiquier à un échiquier $n \times n$, certaines contraintes changent profondément la nature des règles. Nous verrons deux de ces tentatives : une sur le problème des n dames faite par le mathématicien hongrois Georg Pólya en 1918 consistant à mettre l'échiquier, non pas sur un plan, mais sur un tore, ou un beigne; l'autre étudiée un siècle plus tard par les mathématiciennes Hannah Alpert et Érika Roldán sur le problème de domination des dames, en changeant l'échiquier pour des *polyominos*, des sortes de pièces de Tetris.

Le problème toroïdal des n dames

Quelles conditions n doit-il respecter pour qu'il soit possible d'avoir une solution au problème des n dames sur un échiquier $n \times n$ toroïdal?

Le problème de domination des polyominos

Quel est le nombre minimal de dames nécessaire pour dominer un polyomino de N cases?

Généralisation de Pólya, avec un petit détour modulaire

Considérons l'échiquier toroïdal. Comment les pièces bougent-elles sur un tel échiquier? Il n'est pas nécessaire d'avoir avec soi un

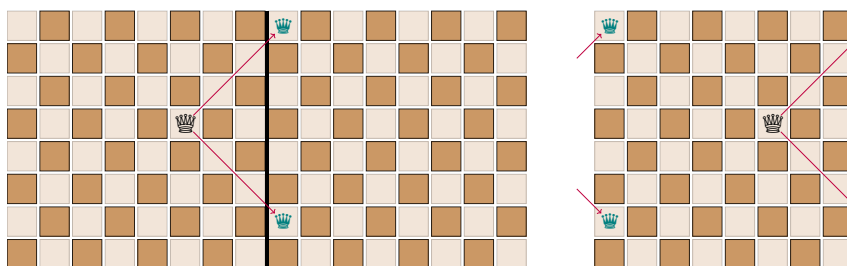


Figure 6: La dame peut atteindre les cases (9,8) et (9,2) sur le deuxième échiquier. Elle peut donc atteindre les cases (1,8) et (1,2) sur l'échiquier initial.



Figure 5 : échiquier toroïdal.

beigne quadrillé pour imaginer le mouvement, un simple échiquier plat suffit. On transforme un échiquier classique en échiquier toroïdal en éliminant les frontières :

une pièce qui atteint la frontière gauche ressort par la frontière droite, et de même pour celles du haut et du bas.

Afin de nous aider à visualiser, construisons une sorte d'échiquier « modulaire » qui étend l'échiquier initial durant le mouvement des pièces en mettant plusieurs échiquiers côte à côte. Durant leur déplacement, les pièces peuvent se déplacer sur tous les échiquiers et elles retournent à la case correspondante sur l'échiquier initial à la fin.

Notons la position d'une dame par ses coordonnées cartésiennes (c, r) pour c , la colonne et r , la rangée en comptant du bas vers le haut et de la gauche vers la droite. Par exemple, à la Figure 6, une dame en $(6, 5)$ atteint la case $(9, 8)$ sur un échiquier supplémentaire. Lorsqu'on rassemble les échiquiers, cette case devient la case $(1, 8)$.

Arithmétique modulaire

Y a-t-il une façon plus facile de donner le mouvement d'une dame? Bien sûr! Le procédé décrit ci-haut est un exemple d'*arithmétique modulaire*. Dans le monde modulaire, les égalités sont remplacées par des congruences liées à un nombre n . Deux nombres a, b sont dits congrus modulo n si les deux ont le même reste après division par n , autrement dit si $a = k \times n + b$ pour un certain nombre entier k ; on note alors $a = b \pmod n$.

Les heures sont un système modulaire pour $n = 24$ (ou $n = 12$). Quand on dort 8 heures après un coucher à 23h,

on se réveille à 7h, et non pas à 31h. Sans le savoir, on fait la congruence $31 = 7 \pmod{24}$.

L'échiquier « modulaire » que nous avons introduit est un exemple d'arithmétique modulaire pour $n = 8$. Tous les échiquiers supplémentaires ajoutés sont issus de translations horizontales ou verticales de 8 unités de l'échiquier initial. L'opération « revenir à l'échiquier initial » revient précisément à prendre le modulo de chaque coordonnée dans la paire (c, r) donnant la position d'une pièce.

Le problème de Pólya

La grande différence du problème toroïdal des n dames avec le problème classique survient sur les diagonales. Par exemple, dans un échiquier classique 8×8 , il y en a 15 de chaque sorte (Nord-Ouest et Sud-Est); sur un échiquier toroïdal, il n'y en a plus que 8 (voir Figure 7).

Il est connu que tous les échiquiers normaux $n \times n$ admettent une solution au problème des n dames dès que $n \geq 4$. Pólya se demandait s'il y a toujours une solution au problème des n dames sur un échiquier toroïdal.

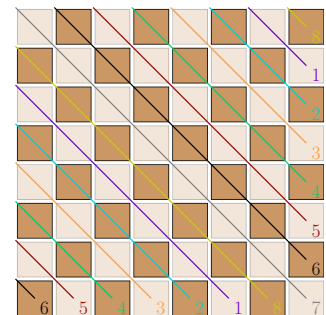


Figure 7 : Les 8 diagonales SE toroïdales.



1,8,7	2,8,6	3,8,5	4,8,4	5,8,3	6,8,2	7,8,1	8,8,8
1,7,8	2,7,7	3,7,6	4,7,5	5,7,4	6,7,3	7,7,2	8,7,1
1,6,1	2,6,8	3,6,7	4,6,6	5,6,5	6,6,4	7,6,3	8,6,2
1,5,2	2,5,1	3,5,8	4,5,7	5,5,6	6,5,5	7,5,4	8,5,3
1,4,3	2,4,2	3,4,1	4,4,8	5,4,7	6,4,6	7,4,5	8,4,4
1,3,4	2,3,3	3,3,2	4,3,1	5,3,8	6,3,7	7,3,6	8,3,5
1,2,5	2,2,4	3,2,3	4,2,2	5,2,1	6,2,8	7,2,7	8,2,6
1,1,6	2,1,5	3,1,4	4,1,3	5,1,2	6,1,1	7,1,8	8,1,7

Figure 8: La somme des chiffres c_i , r_i et d_i de chaque case est un multiple de 8.

Avant de donner la réponse complète de Pólya, étudions le problème pour un échiquier 8×8 .

Numérotons les 8 diagonales SE comme à la Figure 7. Écrivons, pour chaque case de l'échiquier, sa colonne c_i , sa rangée r_i et sa diagonale SE d_i ; chacun de ces nombres est compris entre 1 et 8. Remarquons que le choix de numérotation des diagonales fait en sorte que $c_i + r_i + d_i$ est toujours divisible par 8 (Figure 8). Supposons qu'il y ait une solution. Les huit dames seraient sur huit colonnes distinctes, huit rangées distinctes et huit diagonales SE distinctes. Additionnons

les sommes des colonnes, des rangées et des diagonales SE :

$$\begin{aligned} \sum_{i=1}^8 (c_i + r_i + d_i) &= \sum_{i=1}^8 c_i + \sum_{i=1}^8 r_i + \sum_{i=1}^8 d_i \\ &= 3 \sum_{i=1}^8 i = 3 \times 36 = 108. \end{aligned}$$

Mais, il y a un souci : 108 n'est pas divisible par 8. Donc notre supposition est impossible et il ne peut pas y avoir de solution pour l'échiquier 8×8 .

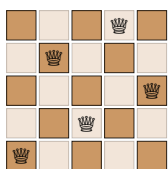
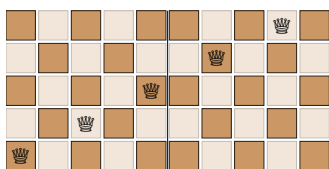


Figure 9: Placer les dames en bonds de cavalier.

La contrainte de Pólya est donc non triviale. Est-elle trop restrictive cependant? En vérifiant les petits n , on peut rapidement vérifier que non : l'échiquier 5×5 admet une solution toroïdale. On l'obtient en plaçant une dame à chaque bond de cavalier (voir Figure 9).

Qu'est-ce que l'échiquier 5×5 a que l'échiquier 8×8 n'a pas? Pólya donne un critère élégant pour déterminer la possibilité ou l'impossibilité du problème.

Théorème (Pólya 1918).

Soit $n \geq 4$. Une solution au problème des n dames sur un échiquier toroïdal $n \times n$ est possible si et seulement si n et 6 sont relativement premiers, c'est-à-dire si 2 et 3 ne divisent pas n .

Pour prouver le théorème, Pólya montre d'abord que la méthode des bonds de cavalier que nous avons présentée fonctionne quand n et 6 sont relativement premiers. Ensuite, il montre qu'il est impossible que 2 ou 3 divisent n s'il existe une solution, en manipulant des sommes sur les éléments de ces listes. Ces manipulations suivent la même idée que la démarche présentée plus haut pour l'échiquier 8×8 , mais utilisent l'arithmétique modulaire pour la généraliser.

Par exemple, déterminer si deux dames (c_i, r_i) et (c_k, r_k) partagent la même diagonale sur un échiquier $n \times n$ revient à vérifier une congruence : elles sont sur une même diagonale SE si $r_i + c_i = r_k + c_k \pmod n$ et sur la même diagonale NE si $r_i - c_i = r_k - c_k \pmod n$. En effet, elles seraient alors respectivement sur la même droite modulaire de pente -1 et de pente 1.

Pour illustrer, prenons la position de la Figure 10. Bien que cette position fonctionne sur un échiquier 8×8 normal, elle n'est pas une solution sur l'échiquier toroïdal. En effet, les dames (1, 6)

et (6, 3) sont sur la même diagonale NE, car $6 - 1 = 5 = -3 = 3 - 6 \pmod 8$, et les dames (4, 2) et (7, 7) sont sur la même diagonale SE puisque $4 + 2 = 6 = 14 = 7 + 7 \pmod 8$.

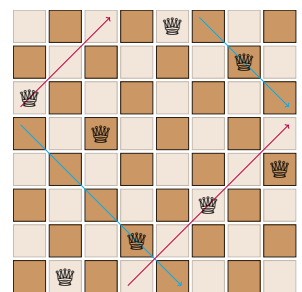


Figure 10: La solution classique ne fonctionne pas sur un tore.

La preuve du théorème de Pólya

Fixons tout d'abord une notation. Une position de dame avec une dame sur chaque colonne est notée par une liste (r_1, \dots, r_n) . C'est une solution si et seulement si :

- I. la liste (r_1, \dots, r_n) contient tous les nombres de 1 à n , on dit donc que c'est une permutation;
- II. la liste $((r_1 + 1) \bmod n, \dots, (r_n + n) \bmod n)$ est une permutation (avec $0 = n \bmod n$);
- III. la liste $((r_1 - 1) \bmod n, \dots, (r_n - n) \bmod n)$ est une permutation (avec $0 = n \bmod n$).

Démonstration

Supposons que 2 et 3 ne divisent pas n . Une solution est donnée en faisant des bonds de cavalier. Soit la liste (r_1, r_2, \dots, r_n) où $r_k = 2k \bmod n$; pour l'échiquier 5×5 , par exemple, cela donne $(2, 4, 1, 3, 5)$. Nous allons nous assurer que les points I-II-III sont vérifiés.

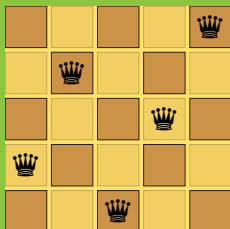


Figure 11:
La solution de Pólya
au problème 5×5
toroïdal.

Le point I est vérifié comme 2 ne divise pas n et que 2 est premier. En arithmétique modulaire, cela veut dire que 2 est inversible modulo n (par exemple $2^{-1} = 3 \bmod 5$ pour $n = 5$, car $3 \times 2 = 6 = 1 \bmod 5$). Cela veut dire que la liste $(2k \bmod n)$ est une permutation. En effet si $2k = 2p \bmod n$, on peut multiplier par l'inverse et alors

$2^{-1} \times 2k = 2^{-1} \times 2p \bmod n$, ce qui revient à $k = p \bmod n$. Ceci prouve que tous les $r_i = 2i$ sont distincts modulo n .

De même, pour le point II, remarquons que la somme donne $(3, 6, \dots, 3n) \bmod n$. Comme 3 ne divise pas n et qu'il est premier, 3 est inversible modulo n , et donc la liste est aussi une permutation.

Le point III se vérifie en remarquant que la différence des deux permutations est $2k - k = k \bmod n$, donc le résultat est simplement la permutation $(1, 2, \dots, n)$. Les trois points sont vérifiés : nous avons construit une solution.

Pour prouver l'autre direction du théorème, donc qu'une solution nécessite que n ne soit pas divisible par 2 ou 3, supposons que nous ayons une solution (r_1, \dots, r_n) . Elle respecte alors les points I-II-III. Le point III indique que

$$((r_1 - 1) \bmod n, (r_2 - 2) \bmod n, \dots, (r_n - n) \bmod n)$$

est une permutation. Donc, additionner ses éléments revient à additionner tous les nombres de 1 à n après un réarrangement des termes. On obtient donc la formule bien connue pour la somme de n nombres consécutifs :

$$\sum_{j=1}^n (r_j - j) = \sum_{k=1}^n k = \frac{n(n+1)}{2} \bmod n.$$

On peut calculer la somme autrement en la séparant. Comme (r_1, \dots, r_n) est une permutation par le point I, elle contient aussi tous les nombres de 1 à n , et donc la différence est :

$$\sum_{j=1}^n (r_j - j) = \sum_{j=1}^n r_j - \sum_{j=1}^n j = 0 \bmod n.$$

En combinant les deux égalités nous avons $n(n+1)/2 = 0 \bmod n$. Cela veut dire que n divise $n(n+1)/2$. Si n était pair, alors $n = 2^s r$ avec r impair et alors $n(n+1)/2 = 2^{s-1}(2^s + r)$. Or, comme ce nombre n'a que $s-1$ facteurs 2, il ne peut pas être divisible par n . Ceci implique que n est impair.

Pour montrer que 3 ne divise pas n , commençons par additionner le carré des éléments des deux listes $((r_1 + 1) \bmod n, \dots, (r_n + n) \bmod n)$ et $((r_1 - 1) \bmod n, \dots, (r_n - n) \bmod n)$.

Comme chaque liste est une permutation par les points II et III, la somme des carrés de ses éléments après réarrangement est la somme des carrés de 1 à n . On additionne ensuite les deux sommes en utilisant la formule de sommation des carrés, ce qui donne :

$$\begin{aligned} \sum_{j=1}^n (r_j - j)^2 + \sum_{j=1}^n (r_j + j)^2 &= 2 \sum_{j=1}^n j^2 \\ &= 2 \frac{n(n+1)(2n+1)}{6} \bmod n. \end{aligned}$$

Additionnons cette fois-ci en développant les deux carrés ce qui donne, comme (r_1, \dots, r_n) est une permutation par le point I,

$$\begin{aligned} \sum_{j=1}^n (r_j - j)^2 + \sum_{j=1}^n (r_j + j)^2 &= \sum_{j=1}^n (r_j^2 - 2r_j + j^2) + \sum_{j=1}^n (r_j^2 + 2r_j + j^2) \\ &= 4 \sum_{j=1}^n j^2 = \frac{2n(n+1)(2n+1)}{3} \bmod n. \end{aligned}$$

Et donc l'égalité suivante est obtenue :

$$\frac{n(n+1)(2n+1)}{3} = \frac{2n(n+1)(2n+1)}{3} \bmod n,$$

qui se simplifie en $n/3 = 2n/3 \bmod n$. Cette équation n'est possible que si 3 ne divise pas n . En effet, si 3 divise n , alors $n = 3^s r$ pour un r non-divisible par 3 et l'équation demande que $n/3 = 3^{(s-1)} r$ soit divisible par n , une contradiction.

Finalement, il y a une solution si et seulement si n et 6 sont relativement premiers. Nous avons donc prouvé le théorème de Pólya.

Polyominos et le problème de domination

Tournons-nous maintenant vers une seconde généralisation de l'échiquier : les polyominos. Dans notre usage, un polyomino sera un ensemble connexe quelconque de cases collées les unes aux autres. Un échiquier ne sera donc plus une grille carrée, mais n'importe quel agencement de cases avec pour seule contrainte à notre fantaisie, l'adjectif «connexe» utilisé ci-haut, qui demande que toutes les cases soient reliées les unes aux autres par un chemin formé de cases appartenant au polyomino. Voici deux polyominos et un non-polyomino :

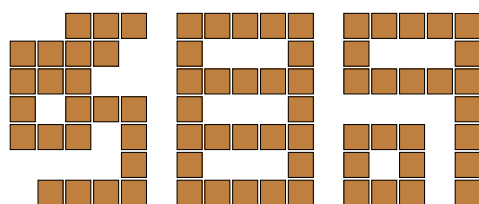


Figure 12: Deux polyominos de 23 et 26 cases et un non-polyomino.

Comme notre échiquier n'est maintenant plus une grille $n \times n$, on ne peut plus attribuer de sens au problème des n dames : quel n faudrait-il considérer ? On se tourne donc vers un problème plus général, dit «de domination», qui s'énonce de la façon suivante : étant donné un polyomino formé de N cases, combien faut-il placer de dames pour surveiller toutes les cases ? Si un certain ensemble de dames placées à certaines cases du polyomino surveille toutes les cases, on dit que cet ensemble «domine» le polyomino.

Remarquons d'abord qu'une contrainte a disparu par rapport aux problèmes précédents : les dames peuvent maintenant se menacer entre elles. Cette remarque nous indique alors une première solution : on n'a

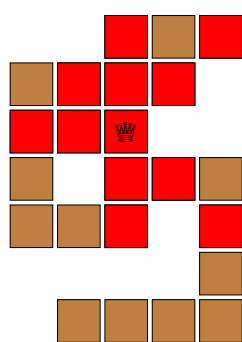


Figure 13 : Une dame et les cases qu'elle surveille.

qu'à placer une dame sur chaque case du polyomino ! Ceci fonctionne, mais n'est pas très intéressant. On cherchera donc plutôt le nombre minimal de dames nécessaire pour dominer un échiquier polyominal. Cette question dépend évidemment de la forme du polyomino. Par exemple, la domination des polyominos de 9 cases peut nécessiter une, deux ou trois dames.

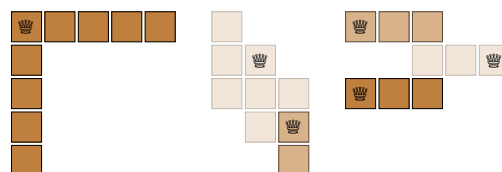


Figure 14 : Polyominos de 9 cases dont la domination nécessite respectivement 1, 2 et 3 dames.

Le problème de domination correctement formulé devient alors le suivant : étant donné un échiquier polyominal à N cases, quel est le plus petit nombre de dames suffisant pour le dominer, peu importe sa forme ?

La réponse à ce problème est donnée, au début de l'année 2021, par les mathématiciennes Hannah Alpert et Érika Roldán.

Théorème (Alpert-Roldán 2021)

Le nombre de dames qui est suffisant et parfois nécessaire pour dominer un polyomino à $N \geq 3$ cases est $\lfloor N/3 \rfloor$.

Commençons par montrer que ce nombre est parfois nécessaire. Il suffit d'exhiber des exemples de polyominos à N cases qui ne peuvent pas être dominés par moins de $\lfloor N/3 \rfloor$ dames. On les construit en empilant des lignes de trois cases comme dans les dessins ci-contre, selon que $N = 0, 1$ ou $2 \pmod 3$.

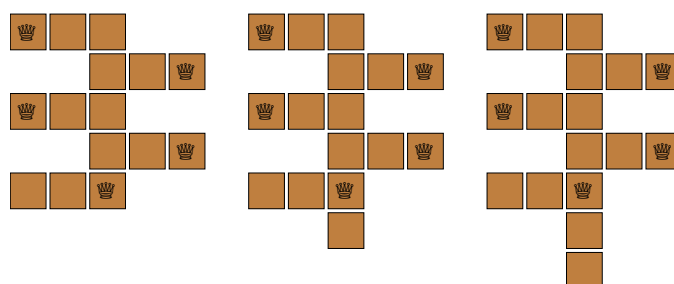


Figure 15: Polyominos nécessitant le nombre maximal de dames pour $N=15, 16, 17$.

Pour montrer que ce nombre est aussi suffisant, il nous faut introduire la notion de distance entre deux cases d'un polyomino. Cette dernière est définie comme étant la longueur du plus court chemin entre ces deux cases, où l'on se déplace d'une case à une autre à gauche, à droite, en haut ou en bas, mais pas en diagonale, et en restant toujours sur des cases appartenant au polyomino.

Alors, une dame placée sur une case d'un polyomino surveille toutes les cases qui sont à distance plus petite ou égale à 2. En effet, elle surveille évidemment sa propre case, qui est à distance 0, et toutes les cases qui sont à distance 1, car elle peut se déplacer d'une case dans toutes les directions. Les cases à distance 2 sont soit des cases directement en diagonale de sa position, auquel cas la dame les surveille, soit des cases à distance 2 en ligne droite dans une certaine direction, que la dame surveille, car elle peut se déplacer d'autant de cases qu'elle veut dans une direction donnée.

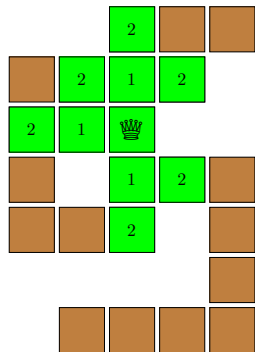


Figure 16 : Une dame et les cases à distance au plus 2 d'elle.

Soit maintenant un polyomino à N cases. La preuve suit l'idée suivante. Choisissons une case du polyomino, qu'on appelle racine, et notons, pour chacune des cases restantes, sa distance à la racine. Colorions ensuite les cases du polyomino avec trois couleurs selon que leur distance à la racine soit congrue à 0, 1 ou 2 mod 3. Chaque case a alors une couleur et la couleur la moins représentée apparaît au plus $\lfloor N/3 \rfloor$ fois. Plaçons alors une dame sur chacune des cases de cette couleur. Pour chaque case du polyomino, en empruntant le plus court chemin vers la racine, on croquera forcément¹ une dame en au plus deux pas. Par le résultat plus haut, comme cette dame se trouve à distance plus petite ou égale à deux, elle surveille la case; toutes les cases du polyomino sont donc surveillées par au moins une dame.

Conclusion

Avec un peu de créativité en mathématiques, on arrive souvent à explorer des avenues inattendues en modifiant un problème, peu importe la provenance de ce dernier. Pas besoin de plus que de se demander : et si...?

1. Un potentiel problème survient pour la racine et les cases à distance 1. On pourrait, pour ces cases, ne pas croiser de dames dans le plus court chemin vers la racine. En choisissant comme racine, si possible, une case qui ne touche qu'à une seule autre case, on peut vérifier que toutes ces cases sont bel et bien surveillées par une dame. Si aucune telle case n'est présente, on peut choisir n'importe laquelle comme racine.

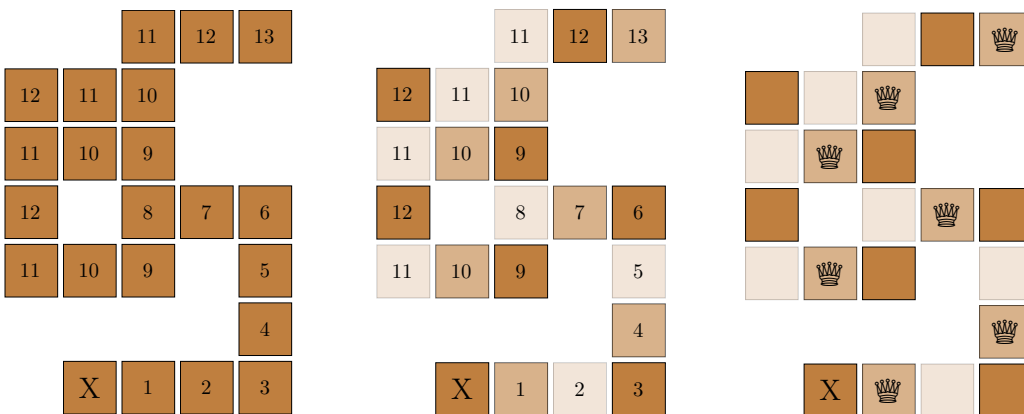


Figure 17:

Les trois étapes de la construction d'une domination par 7 dames d'un polyomino à 22 cases.

L'utilisation d'indices pour combiner des informations

Pour simplifier la présentation de phénomènes complexes dépendant de plusieurs variables, on utilise souvent un nombre, appelé indice, qui rassemble plusieurs informations. Ainsi, que signifie l'indice de refroidissement éolien, ou encore l'indice humidex ? Comment l'indice de Gini permet-il de comparer les inégalités dans la répartition de la richesse ? Comment les Nations unies classifient-elles les pays sur une échelle de développement humain ? Comment définir des indices pour mesurer la biodiversité d'écosystèmes ?

Christiane Rousseau
Université de Montréal

La première idée qui nous vient à l'esprit quand on veut condenser plusieurs informations en un seul nombre est la notion de moyenne. Et on peut vouloir la généraliser en donnant des poids aux différents éléments. On parle alors de *moyenne pondérée*. Elle fonctionne bien pour les notes dans les résultats scolaires. On peut aussi parler du salaire moyen. Par contre, si le taux d'intérêt varie chaque année et que l'on veut calculer le taux d'intérêt moyen, la moyenne usuelle, appelée *moyenne arithmétique*, ne donne pas le résultat cherché et on doit plutôt utiliser une *moyenne géométrique*. Regardons en effet le cas où, l'année i , un montant x placé en début d'année devient $r_i x$ en fin d'année. Après n années, le montant est devenu

$$y = r_1 \dots r_n x.$$

Si le rendement avait été constant égal à r , on aurait dû avoir $y = r^n x$, d'où l'on tire

$$r = (r_1 \dots r_n)^{1/n},$$

c'est-à-dire que r est la moyenne géométrique de r_1, \dots, r_n . Pour un épargnant, c'est équivalent d'avoir un rendement r pendant n années ou bien d'avoir des rendements consécutifs r_1, \dots, r_n . Donc si on considère la fonction $f(s_1, \dots, s_n) = s_1 \dots s_n$, alors on a

$$f(r_1, \dots, r_n) = f(r, \dots, r),$$

c'est-à-dire que (r_1, \dots, r_n) et (r, \dots, r) sont sur la même *surface de niveau* de f .

On peut généraliser cette idée : pour combiner plusieurs informations en une seule, on va vouloir utiliser des fonctions de ces informations dont la surface de niveau a une signification importante. Voyons des exemples.

Le refroidissement éolien

Quelle est la situation la plus dangereuse pour les engelures ? Quand il fait -25° sans vent ou quand il fait -20° et qu'il vente à 40 km/h ? On a l'impression de deux situations incomparables. Une information importante pour quiconque se promène dehors est le risque d'engelure lorsque le visage est exposé au grand froid. Ce risque augmente s'il y a du vent. L'*indice de refroidissement éolien*, R , mis au point par des chercheurs canadiens et américains et utilisé depuis 2001, vient combiner les deux informations que sont le vent et la température. Il est calculé pour une vitesse v du vent supérieure à 4,8 km/s. Cet indice est une fonction R de la température, T , et de la vitesse du vent, v . L'idée de sa définition est que si $R(T_1, v_1) = R(T_2, v_2)$, alors les risques d'engelure, lesquels sont mesurés par la vitesse à laquelle des engelures se produisent, sont les mêmes. Cet indice de refroidissement éolien est donné par une formule compliquée dérivée expérimentalement :

$$R = 13,12 + 0,6215T + (0,3965T - 11,37)v^{0,16},$$

où la vitesse est en km/h, et la température, en degrés Celsius.

Nous ne nous attarderons pas sur sa définition, mais ses courbes de niveau sont données ci-contre.

Ceci signifie que, sur chacune de ces courbes, le risque d'engelure sur le visage est le même, et on voit bien que la peau peut supporter des températures beaucoup plus froides lorsque le vent est faible. Mais, l'indice de refroidissement éolien n'est pas une vraie température, seulement une température ressentie. Le moteur d'une voiture qui a passé la nuit dehors à -20 est à -20, qu'il y ait du vent ou que ce soit le calme plat.

L'indice humidex

Cet indice calcule la capacité du corps humain à se refroidir lors de grosses chaleurs.

Puisque le corps se refroidit en transpirant, cette capacité diminue lorsque l'air ambiant est très humide. Au Canada on utilise depuis 1979 un indice humidex qui dépend de la température, T , et du point de rosée, t , tous deux évalués en degrés Celsius. Le point de rosée est une mesure de l'humidité : c'est la température à laquelle l'humidité présente dans l'air se condense. Le point de rosée est toujours inférieur ou égal à la température ambiante.

La formule de l'indice humidex est compliquée, et ici non plus nous ne discuterons pas de son élaboration :

$$H = T + 0,555 \left[6,11 e^{5417,7530 \left(\frac{1}{273,16} - \frac{1}{273,15+t} \right)} - 10 \right]$$

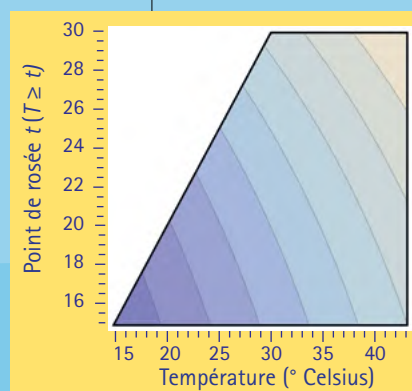
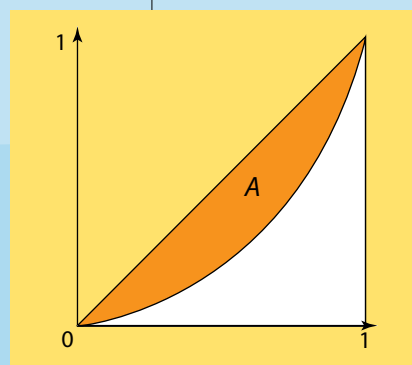
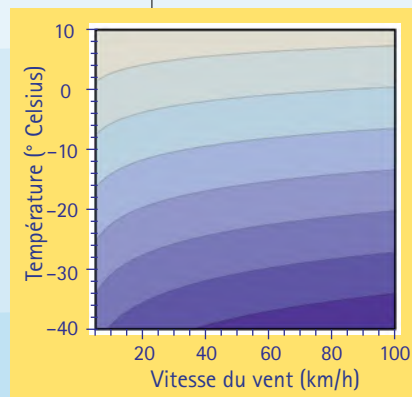
L'indice n'est pas défini dans la région $t > T$, et des courbes de niveau sont données dans la marge en bas à droite.

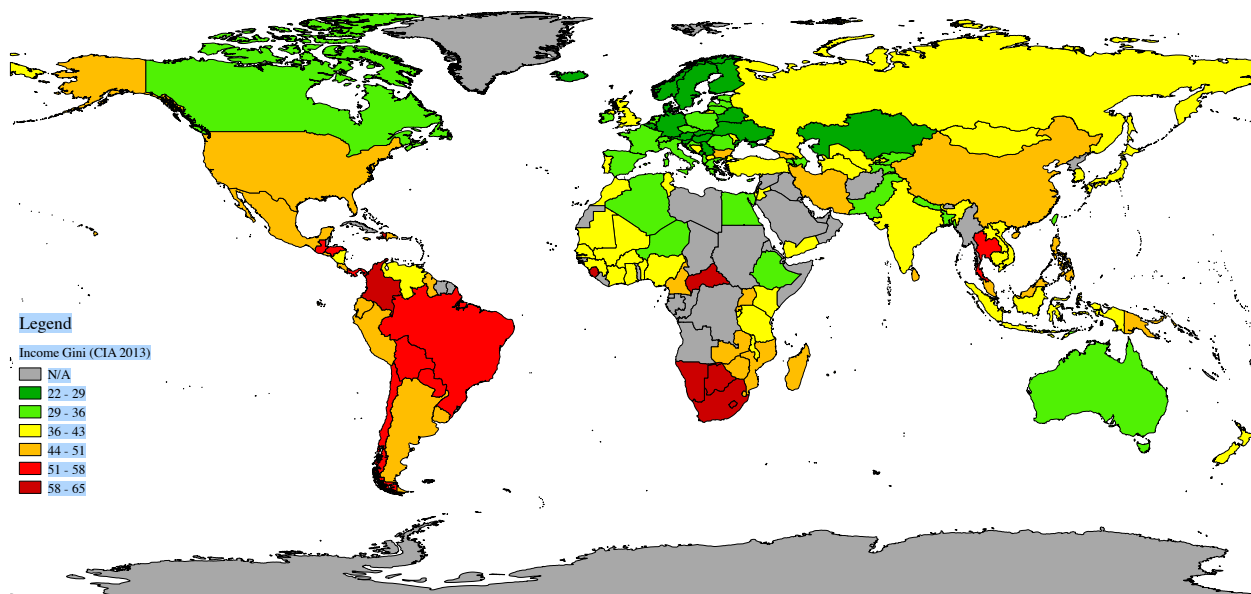
Ici encore, ce sont les courbes de niveau qui sont importantes : le long de chacune de ces courbes le corps a la même capacité de se refroidir. On voit que, pour un humidex donné, plus l'air est humide et moins la température est élevée.

Le coefficient de Gini en économie

Le coefficient de Gini est un indice qui a été introduit en économie pour comparer les différents pays en termes de distribution équitable de la richesse. Pour un pays donné, il mesure comment est répartie la richesse entre les individus de ce pays.

Regardons sa définition. Un pays pauvre pourrait avoir une répartition équitable de la richesse, et un pays riche, une distribution inéquitable. Donc, il faut un critère qui soit indépendant de la richesse du pays. Le genre de critère qui satisfait à cette contrainte est, par exemple, de se donner le pourcentage de la richesse du pays que possèdent les 25% plus pauvres de la population. Mais, il n'y a pas de raison de se limiter à 25%. Étant donné la proportion $p \in [0,1]$ des plus pauvres de la population, on peut se donner la fonction $L(p)$ mesurant la fraction totale de la richesse que cette proportion de la population possède. Alors, cette fonction L , appelée *fonction de Lorenz*, est croissante, $L(0)=0$ et $L(1)=1$. Si la richesse est équidistribuée, alors $L(p)=p$. Si une seule personne possède toute la richesse, alors $L(p)=0$ sauf quand on arrive à la dernière personne. Moins la richesse est équidistribuée, plus le graphe de L est loin de la diagonale. On mesure cette distance par l'aire entre les deux courbes en orange sur la figure ci-contre. En fait cette aire est un nombre de $[0,1/2]$. Donc, pour obtenir un nombre entre 0 et 1, le *coefficient de Gini* (aussi appelé *indice de Gini*) sera défini comme deux fois cette aire. Il prend la valeur 0 quand la richesse est équidistribuée, et une valeur à peu près égale à 1 quand toute la richesse est aux mains d'un seul individu.





Source : CIA Factbook (retrieved: 24/12/2013)

[https://fr.wikipedia.org/wiki/Coefficient_de_Gini#/media/Fichier:World_Income_Gini_Map_\(2013\).svg](https://fr.wikipedia.org/wiki/Coefficient_de_Gini#/media/Fichier:World_Income_Gini_Map_(2013).svg)

Voici ci-haut les coefficients de Gini des différents pays en 2013, là où la situation est connue.

L'indice de développement humain (IDH) des Nations unies

Cet indice a été introduit par les Nations unies en 1990 pour mesurer le taux de développement humain des pays de par le monde. La formule a été améliorée au cours du temps, et la formule actuelle s'est cristallisée en 2011. L'indice de développement humain (ou *IDH*) prend ses valeurs dans $[0, 1]$, où 0 est la plus mauvaise note, et 1, la meilleure note (voir carte page suivante).

Il rassemble trois sous-indices mesurant respectivement la santé-longévité (I_v), le niveau d'éducation (I_e), et le niveau de vie (I_n). Chacun des sous-indices est lui-même un nombre dans $[0, 1]$, et l'indice de développement humain est défini comme la moyenne géométrique de ces trois sous-indices :

$$IDH = \sqrt[3]{I_v \cdot I_e \cdot I_n}$$

Remarquons qu'il faut s'assurer que chacun des sous-indices soit non nul, sinon la moyenne géométrique sera nulle. Aussi, les

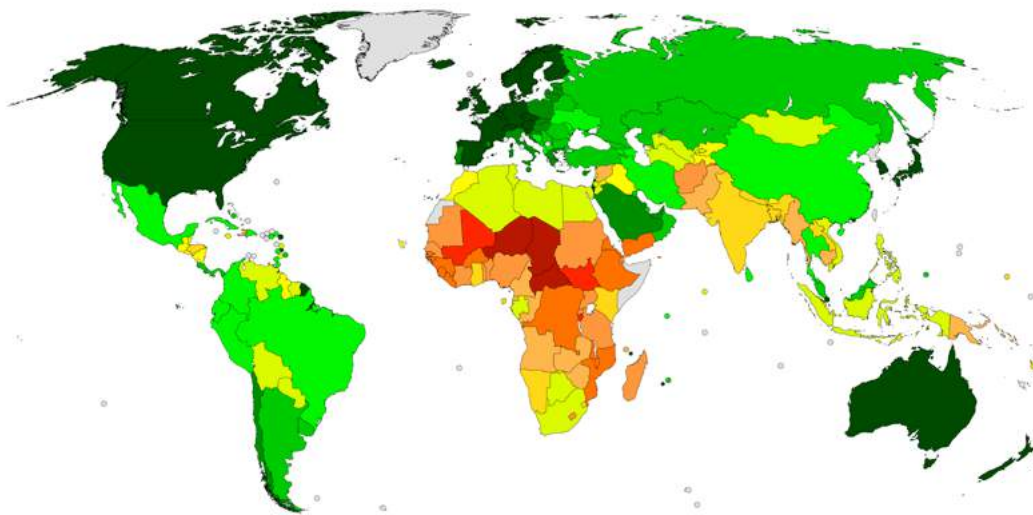
différentes quantités mesurées ont des unités de mesure non comparables. Idéalement, l'indice et les sous-indices ne devraient pas avoir d'unité de mesure. Pour définir un sous-indice prenant des valeurs dans $[0, 1]$, on choisit un indicateur, on se donne une valeur maximale et une valeur minimale de cet indicateur pour l'ensemble des pays et, pour chaque pays, on normalise l'indicateur de la manière suivante :

$$\text{sous-indice} = \frac{\text{valeur pour le pays} - \text{valeur minimale}}{\text{valeur maximale} - \text{valeur minimale}} \quad (*)$$

Ainsi, le sous-indice n'a pas d'unité.

Il est naturel de mesurer la santé-longévité par l'espérance de vie. En effet, si les conditions de santé ne sont pas bonnes dans un pays, que ce soit dû à un accès difficile à des soins de santé, ou à de mauvaises conditions dans les milieux de travail, ou encore à d'autres causes, cela se ressentira sur l'espérance de vie. On prend 20 pour valeur minimale, et la valeur maximale a varié selon les années : elle est de 85 dans le rapport 2020 des Nations unies.

Le choix des valeurs minimale et maximale n'est pas anodin. Si on prenait 0 comme valeur minimale on donnerait une note de 0,69 à un pays ayant une espérance de vie de 50 ans, alors qu'avec la valeur minimale de 20, sa note est de seulement 0,35.



Carte des pays du monde par IDH, selon l'ONU en 2019.
 ■ 0,900 et plus ■ 0,850 à 0,899 ■ 0,800 à 0,849 ■ 0,750 à 0,799 ■ 0,700 à 0,749 ■ 0,650 à 0,699
 ■ 0,600 à 0,649 ■ 0,550 à 0,599 ■ 0,500 à 0,549 ■ 0,450 à 0,499 ■ 0,400 à 0,449 ■ 0,350 à 0,399 et moins
 ■ Données indisponibles.

Attribution : JackintheBox, CC BY-SA 4.0, via Wikimedia Commons.

Pour le niveau d'éducation, on commence par se donner la durée moyenne de scolarisation que l'on compare aux durées moyennes minimale et maximale prises comme 0 et 15 en 2020 : c'est un premier sous-sous-indice, $I_{e,1}$, calculé par la formule (*). Mais, on veut ajouter une deuxième composante, qui mesure le progrès fait par le pays en termes d'éducation. Ainsi, deux pays peuvent avoir la même durée moyenne de scolarisation, par exemple 9 ans, mais le premier pourrait avoir une durée moyenne de scolarisation de 9 ans pour toute la population, alors que le deuxième serait passé d'une scolarité universelle de niveau élémentaire pour les 40 ans et plus à une scolarité universelle de niveau secondaire pour les jeunes générations et avoir instauré des politiques encourageant l'éducation supérieure. Une manière de quantifier cette distinction est de se donner la durée attendue de la scolarisation que l'on compare aux valeurs minimale et maximale de 0 et 18 utilisées en 2020, 18 ans correspondant au nombre d'années de scolarité pour une maîtrise. Ceci donne le deuxième sous-sous-indice, $I_{e,2}$, aussi calculé par la formule (*). Le sous-indice pour l'éducation est la moyenne arithmétique des sous-sous-indices :

$$I_e = \frac{1}{2}(I_{e,1} + I_{e,2}).$$

Le troisième sous-indice mesurant le niveau de vie est calculé par la formule (*) en utilisant le produit national brut per capita du pays (en dollars américains) avec des valeurs minimale et maximale de 100 et 75 000 en 2020. Le maximum de 75 000 est justifié par le fait que les quelques valeurs supérieures à 75 000 (3 pays en 2020) ne changent pas significativement le niveau de vie. Pour ces pays, on pose $I_n = 1$, plutôt que de permettre une valeur supérieure à 1 parce qu'on veut pas que I_n puisse compenser complètement une faiblesse du côté santé et/ou éducation.

Mesurer la biodiversité des écosystèmes

On considère qu'un écosystème est riche quand il abrite beaucoup d'espèces différentes. Donc, un indice naturel de biodiversité d'un écosystème est la *richesse spécifique*, donnée par le nombre absolu d'espèces y nichant. C'est un indice qui fournit peu d'information sur la santé d'un écosystème. Certaines espèces peuvent être en très petit nombre, peut-être même menacées d'extinction.

De plus, l'arrivée d'une espèce invasive augmente à court terme la valeur de l'indice, mais peut menacer la survie d'autres espèces à plus long terme. Regardons deux peuplements forestiers.



Crédit : Nicholas A. Tonelli, Jean-Pol Grandmont, U.S. Fish and Wildlife Service, Sue Sweeney, Famartin

Dans le premier peuplement, on a 30% d'érables, 30% de sapins baumiers et 40% de merisiers.

Dans le second, on a 10% d'érables, 10% de bouleaux gris, 10% de mélèzes laricins, 20% de merisiers et 50% de sapins baumiers. Le deuxième peuplement a une plus grande richesse spécifique, mais a-t-il vraiment une plus grande biodiversité? En effet, il abrite cinq espèces au lieu de trois, mais il compte une surreprésentation de sapins baumiers. Donc, pour quantifier la biodiversité on voudrait compléter le premier indice par un second indice qui mesure la répartition entre les espèces, appelé *équitabilité spécifique*, et qui tempère la richesse spécifique.

Il n'existe pas un unique indice mesurant l'*équitabilité spécifique*. Plusieurs sont utilisés dans la littérature scientifique. Chacun a ses caractéristiques propres, mais aussi ses

faiblesses. Et il est recommandé de prendre en compte le contexte particulier dans le choix d'un indice pour évaluer la biodiversité d'un écosystème. En particulier, plus le recensement est substantiel, plus on risque de découvrir de nouvelles espèces rares. Parmi ces indices, deux ont été introduits en 1949 : l'indice de diversité de Simpson a été introduit en écologie, alors que l'indice de Shannon-Wiener est emprunté à l'informatique théorique.

L'indice de diversité de Simpson

Cet indice, proposé par le statisticien Edward H. Simpson (1922-2019), est donné par la probabilité que deux spécimens choisis au hasard appartiennent à la même espèce. Supposons que l'on ait n espèces en proportions respectives p_1, \dots, p_n . On a donc

$$p_1 + \dots + p_n = 1, p_i \in [0,1], \quad (*)$$

qui est un simplexe dans l'espace (p_1, \dots, p_n) . La probabilité que deux spécimens choisis au hasard appartiennent à l'espèce i est p_i^2 . Alors, l'indice de Simpson est

$$S = p_1^2 + \dots + p_n^2.$$

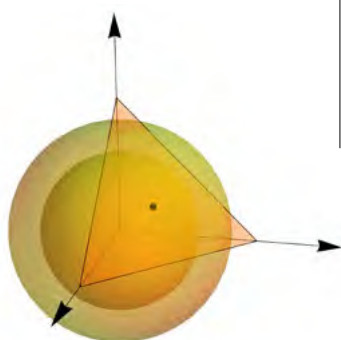
On peut montrer que, sous la contrainte (*), S est minimum pour

$$p_1 = \dots = p_n = 1/n \quad (**)$$

et vaut alors $S = 1/n$. Ceci se voit en faisant grossir une sphère $p_1^2 + \dots + p_n^2 = S$ centrée à l'origine. La première fois qu'elle touche au simplexe (*) est au point (**) (voir figure). La sphère cesse de couper le simplexe lorsqu'elle passe par les sommets du simplexe quand $S = 1$.

Donc, S prend ses valeurs dans $[1/n, 1]$. Et, plus S est petit, plus l'équitabilité spécifique est grande. Un problème avec cet indice est que la longueur de cet intervalle dépend du nombre d'espèces, n , ce qui rend plus difficile la comparaison entre les valeurs de S pour deux écosystèmes ayant des richesses spécifiques différentes.

Pour le premier peuplement, $S = 0,34$, et pour le second, $S = 0,32$. On pourrait dire que $0,32 < 0,34$, mais $0,34$ est beaucoup plus proche de $1/3$ que $0,32$ ne l'est de $1/5$: ceci reflète le fait que le premier peuplement est plus uniformément réparti que le second.



L'indice de diversité de Simpson varie peu si on découvre tout à coup de nouvelles espèces rares passées inaperçues lors des premières analyses de la structure du peuplement.

L'indice de Shannon-Wiener

Reprenons l'exemple de nos deux peuplements forestiers et, dans chacun, choisissons un arbre au hasard. On a plus d'incertitude dans le premier peuplement que dans le second où, dans 50% des cas, on va avoir choisi un sapin baumier. L'indice de Shannon-Wiener est une mesure de cette incertitude. La formule de cet indice est compliquée et sa justification apparaît dans l'encadré. On va étudier ses propriétés. Ici encore, on suppose que l'on ait n espèces en proportions respectives p_1, \dots, p_n , satisfaisant à (*). L'indice de Shannon-Wiener est défini comme

$$H = -p_1 \log_2 p_1 - \dots - p_n \log_2 p_n.$$

Remarquons que $H=0$ si on a une seule espèce, c'est-à-dire $n=1$ et $p_1=1$. Aussi, $H \geq 0$, et on peut montrer (avec les multiplicateurs de Lagrange) que H atteint son

maximum lorsque $p_1 = \dots = p_n = 1/n$, c'est-à-dire quand l'équitabilité spécifique est maximale. Ce maximum vaut alors $\log_2 n$. Donc, $H \in [0, \log_2 n]$ et, plus H est grand, plus l'équitabilité spécifique est grande. Remarquons maintenant que les dérivées partielles par rapport à p_i tendent vers l'infini quand $p_i \rightarrow 0$

$$\lim_{p_i \rightarrow 0} \frac{\partial H}{\partial p_i} = +\infty.$$

Ceci signifie que l'indice de Shannon-Wiener est très sensible aux espèces rares ou encore à l'introduction de nouvelles espèces.

Pour le premier peuplement, $H=1,57$, soit très proche de la valeur maximale de 1,585. Pour le deuxième peuplement $H=1,96$, alors que le maximum potentiel est 2,32. On peut se demander pourquoi les indices de Simpson et Shannon-Wiener n'ont pas été normalisés dans la littérature scientifique pour prendre des valeurs dans tout l'intervalle $[0,1]$, comme cela a été fait pour les sous-indices de l'indice de développement humain...

L'indice de Shannon

L'américain Claude Shannon (1916-2001) était à la fois mathématicien et ingénieur en génie électrique. Il est souvent considéré comme le père de la théorie de l'information. Dans la théorie de l'information, l'entropie de Shannon mesure la quantité d'information qu'un récepteur doit obtenir pour connaître un signal émis. Cette quantité d'information croît avec l'incertitude sur le signal. Voici l'idée derrière la formule de l'entropie de Shannon. Prenons le cas où on a un signal comprenant m bits et donc $N=2^m$ signaux possibles. Pour connaître complètement le signal, il faut connaître chacun de ces bits, donc $m = \log_2 N$ informations. Imaginons maintenant que chacun de ses N signaux (chaque signal est un vecteur de m bits) représente un spécimen d'une population et que cette population soit divisée en n espèces comprenant respectivement N_1, \dots, N_n individus.

La quantité moyenne d'information nécessaire pour identifier complètement un spécimen est égale à la quantité d'information pour identifier l'espèce du spécimen, que nous appellerons H et qui est inconnue, à

laquelle on ajoute la moyenne pondérée des quantités d'informations pour identifier chaque spécimen dans son espèce. La quantité d'information nécessaire pour identifier un spécimen dans l'espèce i est $\log_2 N_i$ si on refait le même raisonnement que ci-dessus. La moyenne pondérée de ces quantités est

$$\sum_{i=1}^n \frac{N_i}{N} \log_2 N_i = \sum_{i=1}^n p_i \log_2 N_i,$$

où $p_i = N_i/N$ est la probabilité qu'un spécimen choisi au hasard appartienne à l'espèce i .

On a alors $\log_2 N = H + \sum_{i=1}^n p_i \log_2 N_i$,

dont on tire $H = \sum_{i=1}^n p_i \log_2 N_i - \log_2 N \sum_{i=1}^n p_i$,

ou encore $H = \sum_{i=1}^n p_i \log_2 \frac{N_i}{N} = \sum_{i=1}^n p_i \log_2 p_i$.

Y a-t-il relation de cause à effet ?

La science statistique de l'inférence causale

Dans le langage courant, on appelle « cause » la raison ou l'origine d'un phénomène, mais il est plus difficile que l'on pense de définir le concept de causalité précisément. Dès lors, comment peut-on identifier une cause et en quantifier les effets, surtout si l'on ne peut pas recourir à une expérience contrôlée pour des raisons logistiques, matérielles ou éthiques ?

Christian Genest
Erica E. M. Moodie
Université McGill

Tel ou tel traitement, vaccin ou médicament est-il efficace ou non ? C'est une question centrale en médecine qui a d'ailleurs souvent fait la une des journaux depuis 2020, alors que divers moyens de se protéger contre la COVID-19 étaient mis de l'avant. Mais comment peut-on s'assurer qu'une thérapie est efficace si elle n'a pas toujours l'effet escompté ? En effet même une forte corrélation entre deux phénomènes n'est pas garante d'un lien de cause à effet : le nombre de consultations pour des coups de soleil a beau être plus grand quand les bars laitiers vendent beaucoup de cornets, la crème glacée n'est pas pour autant la cause des insulations !

La notion de causalité n'est pas aussi facile à définir qu'on pourrait le croire. Elle a passionné les philosophes et les théologiens bien avant qu'elle préoccupe les statisticiens. Les premiers écrits connus à ce sujet remontent à Platon, au 4^e siècle avant notre ère. La question a même amené de grands esprits tels Immanuel Kant et Friedrich Nietzsche à s'interroger sur la nature déterministe de l'univers et l'existence du libre arbitre.

Dans une perspective déterministe, tout événement est prévisible (au degré de précision voulue) à condition de disposer d'une connaissance parfaite du passé et des lois de la nature. Le hasard n'est alors qu'une perception attribuable à une connaissance partielle de l'état du monde. La théorie

statistique de l'inférence causale adopte implicitement ce point de vue en supposant qu'une fois tous les facteurs pertinents pris en compte, un traitement a toujours le même effet.

Ceci étant on peut espérer pouvoir mesurer l'effet d'un traitement si l'on connaît les conditions précises dans lesquelles il a été administré. Néanmoins, cet objectif est souvent illusoire parce que les conditions peuvent être difficiles à déterminer, mesurer ou contrôler. Il est alors possible que l'effet observé soit dû à des facteurs inconnus plutôt qu'au traitement. Pour parvenir à isoler un éventuel effet de traitement, on doit donc se résoudre à formuler des hypothèses et à faire appel à des approches d'analyse statistique dont nous allons décrire le fondement.



Comment faire des comparaisons équitables?

Pour mesurer l'efficacité d'un traitement, il suffit en principe d'administrer celui-ci à un certain nombre de sujets et de comparer leur évolution à celle des membres d'un groupe témoin. Pour que la comparaison soit équitable, il faut toutefois s'assurer que la manière dont les sujets sont assignés au groupe traitement ou au groupe contrôle ne fausse pas les résultats en s'appuyant, par exemple, sur leur état de santé initial ou sur leurs chances de bien répondre à la thérapie.

Cette idée n'est pas nouvelle. Dès 1364, le savant florentin Pétrarque estimait que pour pouvoir comparer deux traitements de façon adéquate, il fallait que les groupes soient constitués d'un nombre égal de sujets malades qui soient tous du même âge, de mêmes mœurs et de même tempérament en plus d'évoluer dans le même environnement. C'est dans ces conditions seulement, pensait-il, qu'une comparaison pouvait être équitable.

De nos jours, les essais cliniques cherchent

le plus souvent à assurer cette équité en assignant les traitements au hasard, c'est-à-dire essentiellement par un jet de dé (bien qu'en pratique, la randomisation soit plutôt faite par ordinateur). Cette façon de procéder garantit que les profils des groupes sont alors comparables en terme d'âge, de

sexe ou de toute autre caractéristique d'intérêt. Toute variation entre les groupes (hormis le traitement) n'est alors que le fruit du hasard.

Par exemple, supposons que l'on veuille étudier l'effet d'un régime à faible teneur en sel sur la prévention de l'hypertension artérielle (HTA) chez des sujets qui y sont prédisposés. Certains facteurs de risque déjà documentés doivent alors être pris en compte tels que l'âge et le sexe, puisque les hommes et les individus de plus de 60 ans sont davantage atteints. En déterminant au hasard qui suivra ou pas le régime faible en sel, on s'assure que les différents facteurs de risque tels l'âge, le sexe, les antécédents familiaux et autres sont présents en proportions à peu près égales au sein des deux groupes. Si une variation significative de la HTA est observée entre les groupes à la fin de l'étude, on pourra alors conclure que le traitement est en cause.

En somme, le recours à la randomisation permet de déduire qu'une corrélation est bel et bien le résultat d'une relation de cause à effet entre un traitement et un résultat observé. C'est le principe sur lequel reposent tous les essais cliniques et c'est ainsi, entre autres, que Santé Canada procède pour approuver un nouveau traitement ou un médicament.

Affaire classée?

S'il suffit de randomiser les traitements pour pouvoir établir un lien de cause à effet, pourquoi ne le fait-on pas toujours? Parce qu'il existe de nombreuses situations dans lesquelles cela s'avère impossible pour des raisons éthiques, matérielles ou logistiques. Si on soupçonne par exemple qu'un produit chimique est cancérigène, il serait contraire à l'éthique d'exposer volontairement des sujets à ses effets. Dans bien des cas, il s'avère aussi que l'exposition doit être prolongée pour qu'une augmentation de l'incidence du cancer soit observée. Avant de pouvoir conclure, il faudrait donc suivre les participants pendant des décennies, ce qui est coûteux et complexe.



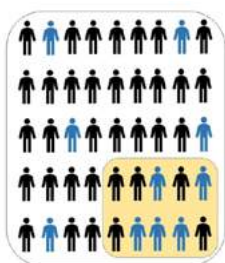


Figure 1

Au début de la pandémie, divers traitements ont été proposés par des médecins qui cherchaient des moyens d'atténuer les effets de la COVID-19. Pour juger de l'efficacité de ces approches, on ne disposait alors que de données disparates et anecdotiques, par opposition à celles émanant d'un essai clinique contrôlé. Bien que l'on ait pu observer que tel ou tel traitement semblait mieux fonctionner qu'un autre pour certains groupes de personnes (par exemple les fumeurs ou les gens obèses), il est possible que les caractéristiques de ces sujets aient été associées à des facteurs cachés qui augmentent les chances de succès a priori. À l'inverse, certains traitements non homologués sont utilisés en dernier recours chez des patients particulièrement mal-en-point ; pas étonnant alors que leur efficacité soit faible !

Le problème des essais cliniques non randomisés, c'est précisément que les traitements peuvent avoir été choisis, sciemment ou pas, en fonction de facteurs connus ou cachés qui sont susceptibles d'influencer le résultat.

Pour mieux comprendre en quoi cela pose problème, revenons à l'exemple portant sur l'évaluation d'un régime à faible teneur en sel comme moyen de réduction du risque de HTA chez les hommes de plus de 60 ans et examinons en quoi les antécédents familiaux peuvent influencer les résultats d'une étude non randomisée.

La population cible est représentée à la figure 1, où les sujets en bleu ont des antécédents familiaux de HTA et les autres pas. Supposons que les sujets du sous-ensemble en jaune aient décidé par eux-mêmes de suivre un régime faible en sel. Comme la proportion d'hommes à risque élevé de HTA est plus grande dans ce groupe (5/10) que dans le reste de la population-cible (5/35), il serait trompeur de juger de l'efficacité du traitement en comparant (disons l'année suivante) la proportion de sujets hypertendus dans les deux groupes.

En effet, s'il existe une prédisposition familiale à la HTA, on s'attendrait à ce que de nombreux participants à l'étude finissent par être hypertendus, même en l'absence de toute forme de traitement ou si celui-ci s'avérait inefficace. Comment tenir compte

d'un tel handicap, qui peut jouer contre le traitement ou dans certains cas, qui sait, en sa faveur ?

Avant de pouvoir répondre à cette question, il est nécessaire d'introduire un peu de formalisme.

Formulation mathématique

Pour un sujet donné, appelons Y la valeur future de la variable d'intérêt (par exemple la tension artérielle moyenne du sujet dans un an). Présument, la valeur de Y ne sera pas la même selon que ce sujet suive un régime à faible teneur en sel ou non. Si on dénote $Z = 1$ le fait que le sujet suive le traitement et $Z = 0$ le fait qu'il ne le suive pas, alors

$$Y = Z \times Y_1 + (1 - Z) \times Y_0,$$

où Y_z est la valeur future de la variable d'intérêt si $Z = z$. La distribution de Y_z (et en particulier sa moyenne) varie selon que $z = 0$ ou $z = 1$.

En terme d'espérance mathématique, l'effet moyen du traitement est alors égal à

$$E(Y_1) - E(Y_0).$$

Cette différence serait facile à estimer si, au lieu de comparer des groupes déséquilibrés comme dans la figure 2, on pouvait traiter tous les sujets et en même temps n'en traiter aucun de façon à comparer deux groupes identiques, comme dans la figure 3.

C'est toutefois impossible et

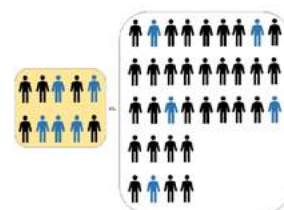


Figure 2

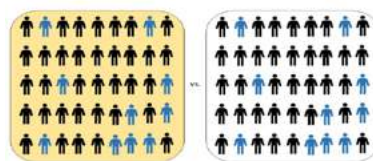


Figure 3

pour chaque sujet, on observera donc plutôt Y_0 ou Y_1 , mais pas les deux. Ce que les données permettent d'estimer, c'est alors une différence d'espérances conditionnelles, à savoir

$$E(Y_1 | Z = 1) - E(Y_0 | Z = 0).$$

Ceci ne pose pas problème dans le cadre d'une étude randomisée, car le fait d'avoir été assigné au groupe $z = 1$ ou au groupe $z = 0$ n'influence pas la distribution des variables Y_0 et Y_1 . On a donc

$$E(Y_0 | Z = 0) = E(Y_0), \quad (1)$$

$$E(Y_1 | Z = 1) = E(Y_1) \quad (2)$$

et l'effet de traitement peut être estimé par la différence entre les moyennes des deux groupes.

Dans une étude non randomisée, en revanche, la probabilité d'avoir $Z = 1$ peut dépendre de certaines caractéristiques du sujet, par exemple le fait qu'il ait des antécédents familiaux de HTA (dénote $X = 1$) ou qu'il n'en ait pas ($X = 0$). Les identités (1) et (2) ne sont alors plus valables et

$$E(Y_1 - Y_0) \neq E(Y_1 | Z = 1) - E(Y_0 | Z = 0).$$

Dans l'exemple portant sur la HTA, on peut résumer la situation au moyen des diagrammes des figures 4 et 5, dans lesquelles X représente la variable « antécédents familiaux » du sujet, Y dénote sa tension artérielle dans un an et Z ses chances de suivre le régime à basse teneur en sel. La direction de la flèche entre deux variables indique laquelle influence l'autre. Ainsi, dans la figure 4, X influence à la fois Y et Z , comme dans une étude non randomisée, alors que dans la figure 5 correspondant à une étude randomisée, X et Z ont un effet sur Y , mais X n'affecte pas Z .

Retour sur la randomisation

La clef de voûte d'une étude randomisée, ce sont les identités (1) et (2), c'est-à-dire le fait que l'assignation aux groupes traitement ($Z = 1$) et contrôle ($Z = 0$) ne dépend pas de facteurs influents ou « covariables » tels que X . Il suffit pour cela que l'assignation se fasse au hasard, mais il n'est pas nécessaire pour autant que la probabilité d'avoir $Z = 1$ soit égale à $1/2$.

Si par exemple on décidait que $\Pr(Z = 1) = 1/3$ (en ne faisant suivre le régime à un sujet que si, disons, on obtient 5 ou 6 lors du jet d'un dé équilibré), le groupe contrôle serait alors à peu près deux fois plus grand que le groupe traitement, mais on s'attendrait quand même à ce que les caractéristiques des deux groupes soient les mêmes en moyenne, en raison de la randomisation.

Or, on peut pousser le raisonnement un peu plus loin et faire en sorte que la probabilité de $Z = 1$ dépende de la valeur de X . Supposons qu'en présence d'antécédents familiaux, il y ait une chance sur deux que le sujet adhère au traitement, dénoté $\Pr(Z = 1 | X = 1) = 1/2$, mais qu'en l'absence de tels antécédents, $\Pr(Z = 1 | X = 0) = 1/7$.

Dans pareil cas, les sujets ayant des antécédents familiaux seraient surreprésentés dans le groupe traitement, ce qui empêcherait toute comparaison équitable avec le groupe contrôle. En revanche, et c'est là le nœud de l'affaire, il serait encore possible de comparer de façon équitable les sujets traités et non traités parmi ceux pour lesquels $X = 1$, puisqu'il y aurait eu randomisation au sein de cette strate. De même, on pourrait encore comparer de façon équitable les sujets traités et non traités parmi ceux pour lesquels $X = 0$.

Comment s'y prendre pour imiter un essai randomisé stratifié ?

Si l'étude est non randomisée, on n'a généralement aucune idée de la valeur de $\Pr(Z = 1 | X = x)$ pour $x = 0$ ou 1 . On peut toutefois essayer de la déduire des données. Pour la population représentée à la figure 1, par exemple, on a

$$\Pr(Z = 1 | X = 1) = 5/10 = 1/2, \\ \Pr(Z = 1 | X = 0) = 5/35 = 1/7.$$

En supposant que l'assignation des sujets aux différents groupes ait été faite au hasard selon ces probabilités, on peut alors comparer strate par strate le groupe traitement au groupe contrôle et calculer un effet de traitement pour les sujets ayant des antécédents familiaux, soit

$$\mu_{oui} = E(Y | Z = 1, X = 1) - E(Y | Z = 0, X = 1)$$

ainsi que pour les sujets qui n'en ont pas, soit

$$\mu_{non} = E(Y | Z = 1, X = 0) - E(Y | Z = 0, X = 0).$$

On peut ensuite estimer l'effet de traitement global en faisant une moyenne pondérée par strates, à savoir

$$\mu = 10/45 \times \mu_{oui} + 35/45 \times \mu_{non},$$

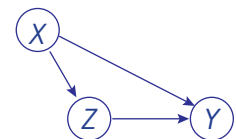


Figure 4

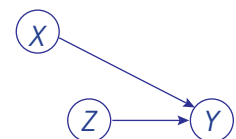


Figure 5

puisque 10 des 45 sujets représentés à la figure 1 avaient des antécédents familiaux et que 35 n'en avaient pas. On peut alors tirer des conclusions fiables quant à la valeur du traitement, sous réserve que l'assignation aux deux groupes ait été faite au hasard au sein de chacune des strates.

Évidemment, les choses se compliquent lorsque la variable X peut prendre plusieurs valeurs différentes, voire même une infinité, comme dans le cas de l'indice de masse corporelle, par exemple. Dans la pratique, il est aussi fréquent que le résultat (Y) et la probabilité d'être traité ($Z = 1$) soient influencées par un vecteur de covariables. Même si toutes les composantes de X sont discrètes, certaines combinaisons de catégories peuvent être rares ou n'avoir jamais été observées, ce qui rend encore plus périlleux le calcul du « score de pension » $\Pr(Z = 1 | X = x)$.

Approche par régression

La régression offre une autre façon de faire des comparaisons équitables. Cette méthode cherche à établir un lien entre la valeur moyenne de la variable d'intérêt, disons la tension artérielle Y mesurée dans un an, et plusieurs variables explicatives telles que le traitement reçu, l'âge, le sexe ou l'indice de masse corporelle.

Dans le cas le plus simple où il n'y a qu'une seule variable explicative X , disons l'âge, on peut supposer par exemple que l'espérance de Y est une fonction linéaire de X , c'est-à-dire

$$E(Y | X = x) = \beta_0 + \beta_1 x, \quad (3)$$

où β_0 est l'ordonnée à l'origine et β_1 est la pente de la droite. Pour estimer les valeurs de ces deux paramètres, on doit disposer d'observations $(x_1, y_1), \dots, (x_n, y_n)$ formant un échantillon de taille n .

Si toutes ces paires de points sont alignées, le calcul de β_0 et de β_1 est un jeu d'enfant. Dans la pratique, on obtient plutôt un graphique semblable à celui de la figure 6, parce que même si la relation (3) est un juste reflet de la réalité, la valeur Y_i observée chez le sujet i âgé de x_i années ne sera pas égale à $\beta_0 + \beta_1 x_i$. En effet, la tension artérielle fluctue

constamment et sera donc vraisemblablement différente de sa valeur moyenne au moment de la prise de mesure.

Pour tenir compte de cette variation individuelle, on suppose que la valeur future Y_i de la tension artérielle du sujet i d'âge $X = x_i$ s'exprime sous la forme

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

où ε_i représente l'écart à la moyenne spécifique à ce sujet.

Dans ces circonstances, estimer la pente et l'ordonnée à l'origine revient à trouver la droite qui s'ajuste le mieux aux données. Dans la figure 6, trois solutions possibles sont proposées : une droite en tiret (---), l'autre en pointillé (...), et la troisième en trait continu. Chacune des droites exprime une relation possible entre l'âge du sujet (abscisse, en années) et sa tension artérielle diastolique moyenne (ordonnée, en mmHg). La première est clairement un mauvais choix car elle passe en-dessous de la plupart des points. Cependant, il est plus difficile de choisir entre les deux autres.

Pour objectiver la recherche d'une solution, on a souvent recours au principe des moindres carrés, qui consiste à trouver les valeurs de β_0 et de β_1 pour lesquelles la fonction

$$L(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

est minimisée. Il se trouve ici que c'est la ligne pleine qui correspond à la solution.

L'approche par régression peut être étendue et adaptée au cas où l'espérance de Y dépend de plusieurs covariables. En particulier, si l'on soupçonne que la tension artérielle des sujets dépend à la fois de leur âge X et de leur adhérence à un régime à faible teneur en sel ($Z = 1$) ou non ($Z = 0$), on peut supposer que

$$E(Y | X = x, Z = z) = \beta_0 + \beta_1 x + \beta_2 z, \quad (4)$$

voire même

$$E(Y | X = x, Z = z) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz. \quad (5)$$

Après avoir estimé les paramètres par la méthode des moindres carrés, on peut alors tracer les droites correspondant aux groupes traitement ($z = 1$) et contrôle ($z = 0$). C'est ce qui a été fait pour le modèle (5) dans la

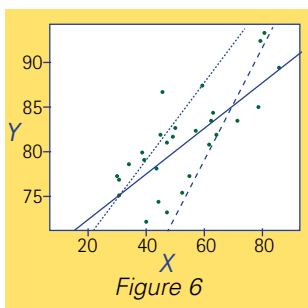


Figure 6

figure 7, où deux droites ont été tracées, une en bleu pour le groupe traitement ($z = 1$) et l'autre en rouge pour le groupe contrôle ($z = 0$).

Dans le modèle (4), les deux groupes ont la même pente, soit β_1 , mais leurs ordonnées à l'origine sont différentes, soit β_0 pour le groupe $z = 0$ et $\beta_0 + \beta_2$ pour le groupe $z = 1$. Dans le modèle (5), les deux groupes ont aussi des ordonnées à l'origine différentes, les mêmes que dans le modèle (4), mais en plus leurs pentes diffèrent : la pente est encore β_1 pour le groupe $z = 0$ mais elle vaut maintenant $\beta_1 + \beta_3$ pour le groupe $z = 1$.

Une fois que les paramètres ont été estimés pour chaque valeur $X = x$, on peut aisément calculer l'effet du traitement. Pour le modèle (5), on trouve

$$\begin{aligned} \mu_x &= E(Y | X=x, Z=1) - E(Y | X=x, Z=0) \\ &= \beta_2 + \beta_3 x. \end{aligned}$$

Si la variable X est le seul facteur susceptible d'influencer la valeur de Y , hormis le traitement lui-même, la relation peut alors être considérée comme causale. Si la variable X est vectorielle, on doit alors faire appel à une version plus complexe du modèle de régression et à la notion de score de propension.

Un exemple concret

Pour illustrer les méthodes présentées plus haut, considérons la greffe de cellules-souches hématopoïétiques allogéniques (AHCT), qui offre des perspectives de guérison aux hémopathies malignes. Cette procédure consiste à transplanter chez un patient des cellules-souches récoltées de la moelle osseuse, du sang périphérique ou du cordon ombilical du donneur. Ces cellules-souches engendrent la production de plaquettes et de globules blancs et rouges, ce qui permet à terme de restaurer le système immunitaire du receveur, dont l'intégrité est compromise. Malheureusement, la maladie aiguë du greffon contre l'hôte (MAGH) est observée chez certains receveurs de cellules-souches allogéniques, entraînant diverses complications pouvant provoquer la mort dans 20 à 40% des cas. Si cette maladie se développe et résiste aux tentatives de traitement précoces,

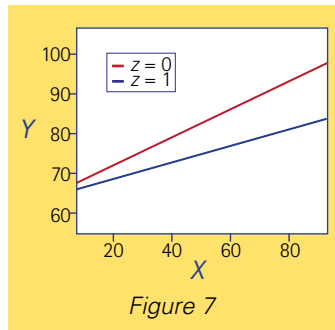


Figure 7

elle peut être combattue au moyen d'immunosuppresseurs standard ou de thérapies non spécifiques, dites NHTL, de déplétion massive des lymphocytes T (la lettre T est l'abréviation de thymus, nom de l'organe dans lequel leur développement s'achève). Des données non expérimentales suggèrent toutefois que les thérapies NHTL sont nocives.

La figure 8 indique la proportion de patients exempts de MAGH en fonction du temps, exprimé en mois depuis la greffe. La courbe rouge correspond aux patients ayant reçu la thérapie NHTL ; la courbe bleue correspond au traitement standard. La courbe rouge descend beaucoup plus rapidement que la bleue, de sorte qu'après 20 mois, par exemple, à peine 15% des patients sous thérapie NHTL restent exempts de MAGH, alors que près de 30% des patients sous traitement standard ne sont pas encore affectés.

À première vue, il semble donc que la thérapie NHTL n'ait aucun avenir. Toutefois, une analyse plus poussée des données révèle que le cancer des patients ayant reçu la thérapie NHTL était généralement plus avancé au moment de la greffe et que ces personnes avaient des risques de MAGH plus grands en raison d'un mauvais appariement donneur-receveur. Ces deux facteurs sont susceptibles d'influencer les résultats. En effet, compte tenu des différences entre les profils des patients assignés à l'un ou l'autre des deux traitements, on s'attendrait a priori à obtenir de moins bons résultats pour ceux qui sont sous thérapie NHTL, peu importe sa nature.

La régression linéaire s'avère ici insuffisante pour analyser adéquatement les données de cette étude, mais en faisant appel à des techniques plus avancées du même acabit, on peut montrer que si la thérapie NHTL conduit généralement à de moins bons résultats, elle s'avère néanmoins bénéfique pour certains types de sujets. Il appert aussi qu'en négligeant les facteurs mentionnés plus haut, on peut être amené à recommander le mauvais traitement chez environ 5% des patients.

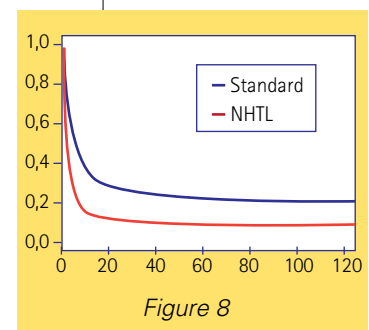


Figure 8

Comparaison d'aires :

2. la méthode d'exhaustion et la méthode du levier

La méthode d'exhaustion est une méthode de démonstration d'égalité d'aires et de volumes de figures géométriques.

Elle fut longtemps considérée comme la seule méthode de démonstration vraiment rigoureuse. Pour obtenir les résultats à démontrer, Archimède avait préalablement recours à la méthode du levier.

André Ross
Professeur retraité



Eudoxe de Cnide
(-408 à -355)

Les constructions à la règle et au compas étaient conformes au *monde des Idées*¹ du philosophe Platon (-427 à -347). Cependant, un de ses disciples, Eudoxe de Cnide (-408 à -355), a développé une méthode de démonstration, appelée depuis le XVII^e siècle *méthode d'exhaustion*, qui ne respecte pas les contraintes de la règle et du compas chères au philosophe.

Antiphon le sophiste (-480 à -411) avait énoncé que :

En doublant le nombre de côtés d'un polygone régulier inscrit dans un cercle et en répétant successivement l'opération, on peut rendre nulle la différence entre l'aire du cercle et l'aire du polygone.

Antiphon a tenté d'appliquer cette approche pour résoudre la quadrature du cercle par l'inscription successive de polygones réguliers dont les nombres de côtés formaient une progression géométrique. À partir du triangle, on obtient :

3, 6, 12, 24, ...

et à partir du carré :

4, 8, 16, 32, ...

1. Luc Brisson, dans son introduction à Platon, *Timée, Critias*, traduit le mot *Idées* par *Formes intelligibles*, parce que depuis Descartes le mot *Idées* désigne une représentation de la réalité, alors que chez Platon c'est la réalité.

À l'époque il était hasardeux de se prononcer sur le résultat d'un processus infini, comme l'avait illustré Zénon d'Élée (vers -495) dans ses paradoxes (voir encadré *Zénon d'Élée*). Pour rendre l'idée d'Antiphon utilisable dans une démonstration sans statuer sur le résultat infini, Eudoxe a adopté un postulat qu'on peut formuler comme suit :

Postulat

Si on soustrait d'une grandeur donnée une partie supérieure ou égale à sa moitié, et que du reste, on soustrait une partie supérieure ou égale à sa moitié et ainsi de suite, à la longue², la grandeur restante peut être rendue plus petite que n'importe quelle grandeur prédéfinie de même nature³.

Ainsi formulé, le postulat n'affirme pas que l'on peut rendre nulle la différence entre l'aire du cercle et celle du polygone inscrit. Il indique seulement que l'on peut rendre cette différence aussi petite que l'on voudra en lui soustrayant itérativement une partie supérieure ou égale à la moitié de la partie restante.

2. Dans cet énoncé, l'expression « à la longue » signifie après un nombre fini d'étapes.

3. Ce postulat se retrouve chez Euclide, proposition 1 du livre X des *Éléments*.



Zénon d'Élée (-490 à -430)

Le philosophe grec Zénon est né à Élée, une ville du sud de l'Italie, entre 495 et 480 avant notre ère. Comme son maître Parménide (vers le VI^e siècle avant notre ère), il fut probablement pythagoricien avant que Parménide ne fonde l'École d'Élée.

Zénon a énoncé divers paradoxes pour montrer l'inconsistance des enseignements des autres écoles. Quatre de ces paradoxes portent sur les enseignements de Pythagore (-569 à -475) et d'Anaxagore (environ -500 à -428). Pour Pythagore, le temps et l'espace sont constitués de parties indivisibles. Anaxagore rejetait cette représentation de l'espace et du temps en parties indivisibles. Il professait la divisibilité infinie de la matière, de l'espace et du temps.

Les paradoxes de Zénon portent sur ces deux représentations du temps et de l'espace. Zénon cherche à montrer qu'aucune de ces conceptions de l'univers n'est conforme à la réalité à l'aide de quatre paradoxes : la dichotomie, Achille et la tortue, la flèche, le stade.

La dichotomie

Le paradoxe de la dichotomie est formulé de la façon suivante :

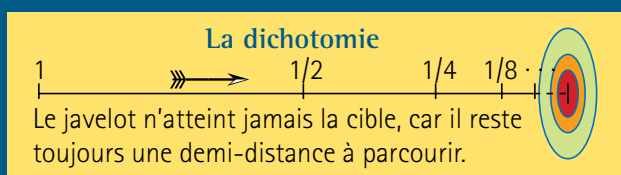
Achille lance un javelot vers une cible.

Pour atteindre la cible, le javelot doit d'abord parcourir la moitié de la distance, puis la moitié de la distance restante, et encore la moitié de la distance restante, ainsi de suite.

Puisque la longueur est infiniment divisible, il reste toujours une moitié de distance à parcourir et le javelot n'atteint jamais la cible.

Le mouvement est donc impossible.

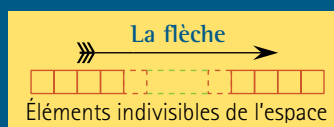
Si on accepte l'hypothèse de la divisibilité infinie, une longueur finie contient un nombre infini de parties. Pour que le mouvement soit possible, il faudrait parcourir un nombre infini de segments en un temps fini. Ce premier paradoxe, tout comme celui d'*Achille et la tortue* est construit en considérant que le temps est constitué d'instants indivisibles et que l'espace est infiniment divisible.



La flèche

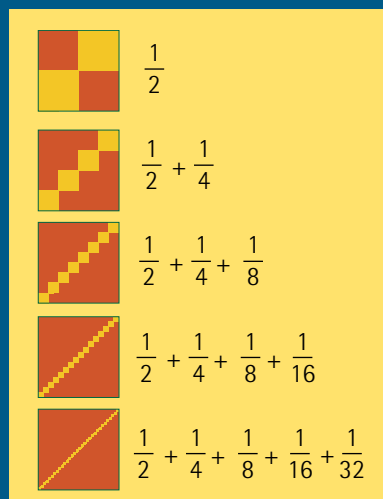
Les deux paradoxes, *la flèche* et *le stade* sont formulés en prenant comme hypothèse que le temps et l'espace sont constitués d'éléments indivisibles, conformément aux enseignements des Pythagoriciens. Le paradoxe de la flèche s'énonce comme suit :

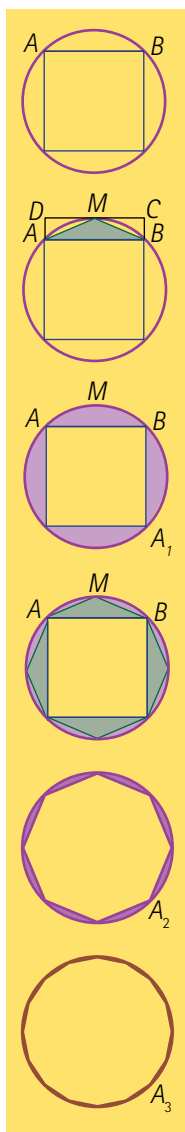
Si le temps est fait d'instants indivisibles, alors une flèche en mouvement est toujours arrêtée, car à tout instant la flèche est en une position donnée et occupe un espace égal à elle-même. Puisque cela est vrai en tout instant, il s'ensuit que la flèche ne se déplace jamais parce qu'un corps qui occupe toujours le même espace ne se déplace pas.



À cause de ces paradoxes, les mathématiciens et les philosophes grecs ont évité systématiquement l'usage de l'infini en raison des pièges que constitue le recours à des convictions intuitives fondées sur le fini, lorsqu'on traite de l'infini. Il n'était plus possible d'utiliser l'infini dans un raisonnement sans le rendre suspect. Ainsi, Euclide ne dit pas qu'il y a un nombre infini de nombres premiers, il considère qu'il y en a plus que tout nombre prédéterminé. La formulation de divers paradoxes fut, pour plusieurs siècles, la seule utilisation de l'infini dans les raisonnements.

De nos jours, les mathématiciens ont apprivoisé l'infini et reconnaissent qu'une somme comportant un nombre infini de termes peut donner un nombre fini. Ainsi, il est évident qu'en coloriant à chaque étape la moitié de la surface non encore coloriée, on ne peut couvrir plus que la surface du carré lui-même, qui est égale à 1.





En utilisant la notation moderne, on peut illustrer numériquement ce postulat. Soit a une grandeur dont on soustrait les deux tiers de la valeur. On a alors :

$$a - \frac{2}{3}a = \frac{3a}{3} - \frac{2a}{3} = \frac{a}{3}.$$

En soustrayant successivement du reste les deux tiers de cette valeur, on obtient :

$$\begin{aligned} \frac{a}{3} - \frac{2}{3} \times \frac{a}{3} &= \frac{3a}{9} - \frac{2a}{9} = \frac{a}{9}, \\ \frac{a}{9} - \frac{2}{3} \times \frac{a}{9} &= \frac{3a}{27} - \frac{2a}{27} = \frac{a}{27}, \dots \end{aligned}$$

Le postulat affirme qu'en poursuivant le processus, on peut rendre le reste plus petit que toute grandeur fixée d'avance.

Voyons comment utiliser ce postulat dans un contexte géométrique.

Aire du cercle et du polygone régulier inscrit

La méthode d'exhaustion d'Eudoxe repose sur l'idée d'approcher une figure donnée par d'autres figures connues, par exemple des polygones, qui viennent s'y « coller » de plus en plus près. Ces polygones, en quelque sorte, « épuisent » la figure donnée. En voici une illustration dans le cas du cercle.

La différence entre l'aire d'un cercle et l'aire d'un polygone régulier inscrit peut être rendue aussi petite que toute aire donnée.

Considérons un cercle, le carré inscrit et AB , un des côtés du carré. On construit le rectangle $ABCD$ tel que le côté CD soit tangent au cercle en M , point milieu de CD . L'aire du triangle AMB est alors la moitié de l'aire du rectangle $ABCD$; elle est donc supérieure à la moitié de l'aire du segment circulaire AMB . Posons A_1 , la différence de l'aire du cercle et de celle du carré. En soustrayant de A_1 le produit de l'aire du triangle AMB par le nombre de côtés du carré, on obtient A_2 , la différence entre l'aire du cercle et celle d'un octogone. En procédant ainsi, on peut rendre la différence entre l'aire du cercle et celle d'un certain polygone plus petite que toute grandeur d'aire donnée.

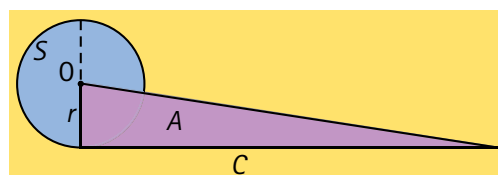
La mise en oeuvre de la méthode d'exhaustion pour établir l'égalité de deux aires se fait

habituellement par une double réduction à l'absurde en montrant que l'une d'elles ne peut être ni plus petite ni plus grande que l'autre. Il faut cependant avoir une connaissance préalable du résultat à démontrer, car la méthode ne permet que de valider ce résultat, et non pas de le découvrir.

Archimède et le cercle

La méthode d'exhaustion a été abondamment utilisée par Archimède (-287 à -212) pour établir des relations entre les aires de diverses figures géométriques. Il a ainsi démontré que :

L'aire d'un cercle est égale à l'aire d'un triangle dont la hauteur est égale au rayon et la base est égale à la circonférence.



En fait, il y a trois possibilités, l'aire S du cercle est soit plus petite, soit égale ou soit plus grande que l'aire A du triangle :

$$S > A, S = A \text{ ou } S < A.$$

En procédant par la méthode d'exhaustion, il faut montrer que les cas $S > A$ et $S < A$ ne peuvent être retenus.

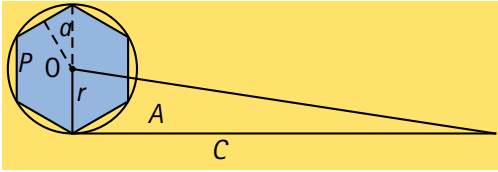
Idée de la preuve

Supposons que l'aire du cercle est plus grande que celle du triangle, c'est-à-dire $S > A$. Il existe alors une grandeur $e > 0$ tel que $S = A + e$, d'où $S - A = e$.

On peut construire un polygone inscrit dans le cercle de telle sorte que la différence entre l'aire S du cercle et l'aire P du polygone soit plus petite que e . On a alors

$$S - P < e = S - A,$$

d'où on conclut que $P > A$.⁴



Le polygone étant inscrit dans le cercle, son périmètre p est forcément plus petit que la circonférence et son apothème est plus petit que le rayon du cercle, d'où :

$$p < C \text{ et } a < r, \text{ et donc } pa < Cr.$$

Considérant l'aire P du polygone, on a alors $P = pa/2$. Or comme $A = Cr/2$ (voir la figure), on en conclut que $P < A$.

Mais cela vient en contradiction avec le fait que $P > A$. Il faut en conclure que l'hypothèse $S > A$ ne peut être retenue et que l'aire du

4. Il suffit, au besoin, de doubler autant de fois que nécessaire le nombre de côtés pour trouver le polygone qui satisfait à cette condition.

cercle ne peut être plus grande que celle du triangle. De façon analogue, Archimède complète sa démarche par double contradiction en montrant que l'aire du cercle ne peut être plus petite que celle du triangle. Il en conclut que ces deux aires sont forcément égales⁵.

Archimède et la parabole

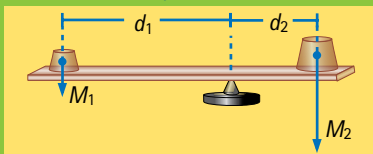
Nous avons déjà signalé que pour démontrer un résultat par la méthode d'exhaustion, il faut déjà connaître ce résultat. C'est ici qu'Archimède a recours à une méthode physique, la *méthode du levier*, pour déterminer quoi démontrer. Cette méthode consiste à considérer des segments de droites pris dans chacune des deux figures et à déterminer à quelles conditions ils seront en équilibre par rapport à un point choisi comme pivot.

5. Pour de plus amples informations sur ce résultat d'Archimède, voir Marie-France Dallaire et Bernard R. Hodgson, « Regard archimédien sur le cercle: quand la circonférence prend une bouffée d'aire. » *Accromath 8*, hiver-printemps 2013, pp. 32-37.

Archimède et l'étude des leviers

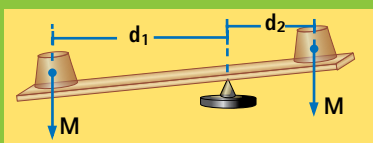
Dans son étude des leviers, Archimède adopte une approche analogue à celle de la géométrie en énonçant des principes sous forme de postulats :

Des masses inégales, à des distances inversement proportionnelles à ces masses, sont en équilibre.



Autrement dit, $\frac{M_1}{M_2} = \frac{d_2}{d_1}$.

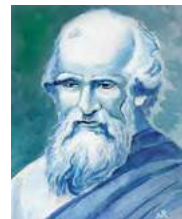
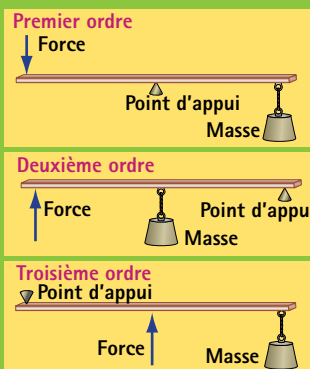
Des masses égales à des distances différentes ne sont pas en équilibre et penchent du côté de la masse qui est à la plus grande distance.



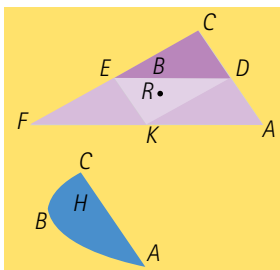
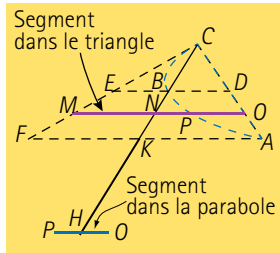
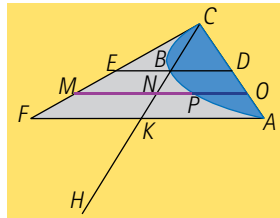
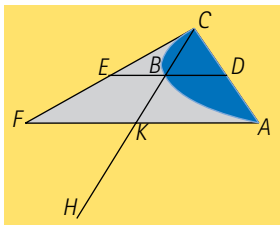
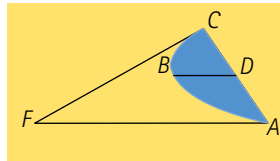
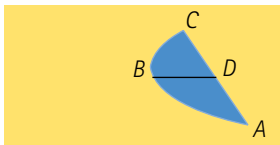
Des masses qui s'équilibrent à des distances égales sont égales.

Les leviers étaient utilisés depuis fort longtemps lorsqu'Archimède a fait une description mathématique de leurs caractéristiques fondamentales et utilisé cette abstraction mathématique pour démontrer d'autres propriétés.

Classification des leviers selon la position du point d'appui et des forces



Archimède
~287 à ~212



Illustrons cette façon de procéder en considérant un segment de parabole ABC et un segment de droite BD qui coupe la corde CA en deux parties égales. Du point C , Archimède trace la tangente au segment de parabole et du point A , une parallèle à l'axe BD jusqu'à leur rencontre en F . Puis, il trace CB qui coupe AF en K et qu'il prolonge jusqu'en H , de telle sorte que $CK = KH$ et il prolonge DB jusqu'en E sur CF .

Selon une proposition qu'Archimède attribue à Aristée l'Ancien (-360 à -300) et à Euclide (vers -320 à -260), $DB = BE$ puisque CE est tangente au segment de parabole et CD est la demi-longueur de sa corde. Il s'ensuit que CK est la médiane du triangle AFC . Il considère que la surface du triangle et celle du segment de parabole sont constituées de segments de droites parallèles à l'axe de celle-ci. Dans la figure ci-contre, le segment MO dans le triangle et le segment PO dans la parabole sont superposés.

Archimède cite alors un lemme qu'il a préalablement démontré à l'effet que :

$$\frac{CA}{AO} = \frac{MO}{OP}$$

Par le théorème de Thalès, $\frac{CK}{KN} = \frac{CA}{AO}$,

d'où $\frac{HK}{KN} = \frac{MO}{OP}$ puisque $CK = KH$.

Il utilise alors le segment CKH comme levier et K comme pivot. En suspendant le segment OP au point H , il équilibrera le segment MO en N .

Chaque segment de droite pris dans la parabole suspendue au point H équilibrera le segment de droite correspondant pris dans le triangle suspendu en son point milieu sur le levier.

L'aire de la parabole suspendue en H par son centre de gravité équilibrera l'aire du triangle suspendu par son centre de gravité sur KC . Or, ce centre de gravité est en un point R situé au tiers de KC . Le rapport de l'aire du triangle AFC , notée $\text{Aire}\Delta AFC$, sur l'aire du segment de parabole ABC , notée $\text{Aire}\widehat{ABC}$, est donc :

$$\frac{\text{Aire}\Delta AFC}{\text{Aire}\widehat{ABC}} = \frac{\overline{HK}}{\overline{KR}} = \frac{3}{1},$$

d'où $\text{Aire}\widehat{ABC} = \frac{1}{3} \text{Aire}\Delta AFC.$

Cependant, l'aire du triangle AFC est quatre fois l'aire du triangle ABC . En effet, les triangles ABD et CBD ont la même aire puisque leurs bases sont égales, D étant le point milieu de AC , et ils ont la même hauteur, la perpendiculaire abaissée de B sur AC . De plus, les triangles EBC et DBC ont même aire puisque B est le point milieu de ED et ils ont même hauteur, la perpendiculaire abaissée du sommet C sur ED . Par conséquent, l'aire du triangle DEC est égale à l'aire du triangle ABC .

De plus, les triangles DEC et AFC sont semblables puisque DE est tracée parallèlement à AF . Puisque D est le point milieu de AC , on a donc $FA = 2ED$.

Puisque les aires de figures semblables sont dans le rapport des carrés de leurs lignes homologues, l'aire du triangle AFC est quatre fois l'aire du triangle ABC inscrit, soit :

$$\begin{aligned} \text{Aire}\widehat{ABC} &= \frac{1}{3} \text{Aire}\Delta AFC \\ &= \frac{4}{3} \text{Aire}\Delta ABC. \end{aligned}$$

Grâce à la méthode du levier, Archimède sait donc que la proposition à démontrer est :

L'aire d'un segment de parabole est égale à une fois et un tiers l'aire du triangle ayant pour base la corde délimitant le segment de parabole et dont le troisième sommet est sur la parallèle à l'axe de la parabole passant par le point milieu de la corde.

Il peut alors donner une preuve de ce résultat reposant sur une démarche purement mathématique, sans l'emploi du levier (qui relève de la physique).⁶

6. À propos de l'aire du segment de parabole, voir aussi les trois textes de Marie Beaulieu et Bernard R. Hodgson dans *Accromath*, vol. 10 (hiver-printemps et été-automne 2015) et vol. 13 (hiver-printemps 2018).

Archimède et la sphère

Pour déterminer le volume de la sphère, Archimède a recours à nouveau à la méthode du levier. Voyons comment il a procédé, en utilisant une écriture et une nomenclature modernes qui nous sont familières. Il place la sphère (de rayon r) de telle sorte qu'un diamètre AB coïncide avec un axe horizontal et trace le diamètre perpendiculaire GH .

Dans le même plan que le diamètre GH , il construit le rectangle $ABED$ de telle sorte que $AD = r$. En prolongeant le segment AG jusqu'à sa rencontre avec le prolongement du côté BE , il construit le triangle ABC .

La révolution du rectangle $ABED$ autour de l'axe horizontal TB donne un cylindre et celle du triangle ABC autour du même axe engendre un cône.

Coupons ces trois solides en tranches d'épaisseur e , perpendiculaires à l'axe TB et à une distance x du point A considéré comme pivot du levier.

La tranche du cylindre est un disque de rayon est r , d'épaisseur e , et de volume

$$VT_{\text{cylindre}} = \pi r^2 e.$$

La tranche du cône est un disque dont le rayon est x , son volume est :

$$VT_{\text{cône}} = \pi x^2 e.$$

Dans la figure 4 ci-contre, on voit que la tranche de la sphère est un disque dont le rayon R est tel que :

$$r^2 = R^2 + (r - x)^2$$

d'où : $r^2 = R^2 + r^2 - 2rx + x^2$, qui donne $R^2 = 2rx - x^2$. Le volume de la tranche de la sphère est donc :

$$VT_{\text{sphère}} = \pi(2rx - x^2)e.$$

Suspendons les tranches de la sphère et du cône à l'extrémité T de l'axe où $TA = 2r$. On peut trouver le moment⁷ combiné de la tranche de la sphère de la sphère et de celle du cône par rapport à A , ce qui donne :

$$\begin{aligned} [VT_{\text{sphère}} + \Delta VT_{\text{cône}}]2r &= [\pi(2rx - x^2)e + \pi x^2 e] 2r \\ &= 4\pi r^2 e x \\ &= 4x VT_{\text{cylindre}} \end{aligned}$$

7. Le moment d'un volume par rapport à un point est le produit de la masse de ce volume par la distance entre le point et le centre de gravité du volume.

Le moment combiné des tranches de la sphère et du cône est donc égal au moment de la tranche du cylindre dans la position qu'elle occupe, à une distance x du point A . En additionnant les moments de toutes les tranches, on a :

$$2r(V_{\text{sphère}} + V_{\text{cône}}) = 4r V_{\text{cylindre}}$$

Cependant, le volume du cylindre est le produit de l'aire de sa base, πr^2 , par sa hauteur, $2r$, soit $V_{\text{cylindre}} = 2\pi r^3$. Le volume d'un cône est le tiers du volume du cylindre de même rayon et de même hauteur, on a donc $V_{\text{cône}} = 8\pi r^3/3$. En substituant,

$$2r \left[V_{\text{sphère}} + \frac{8\pi r^3}{3} \right] = 8\pi r^4$$

et

$$V_{\text{sphère}} = \frac{4\pi r^3}{3}.$$

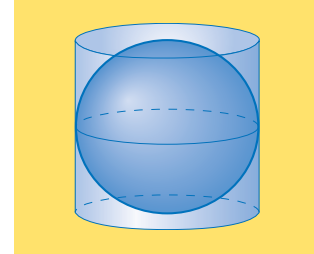
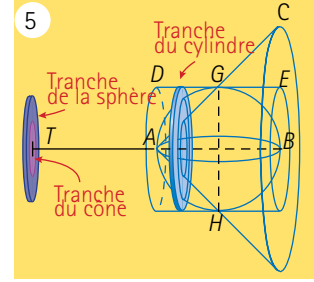
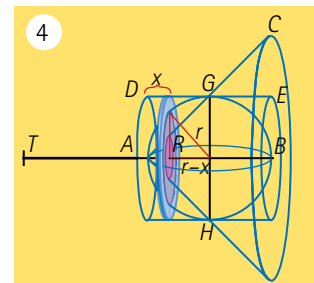
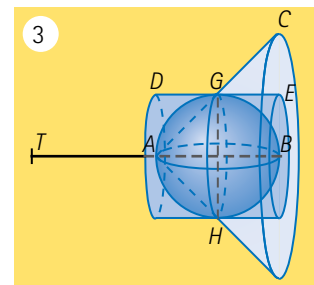
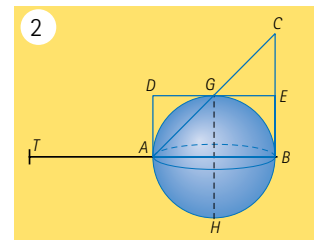
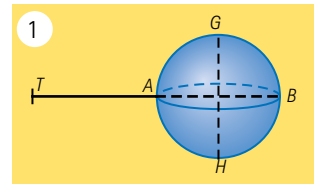
Archimède obtient alors :

$$\frac{V_{\text{cylindre}}}{V_{\text{sphère}}} = \frac{2\pi r^3}{4\pi r^3/3} = \frac{3}{2}.$$

Il établit le même rapport entre les surfaces du cylindre et de la sphère.

Conclusion

La méthode d'é exhaustion et la méthode du levier ont permis d'établir des résultats intéressants tant dans le plan que dans l'espace. Cependant, ces méthodes manquent de généralité, n'étant pas applicables à toutes les figures. Jusqu'au XVII^e siècle, elle ont été les seules méthodes de démonstration vraiment rigoureuses. L'étape suivante dans la comparaison d'aires sera la *méthode des indivisibles* développée par Bonaventura Cavalieri (1598-1647).



Lorsqu'un cylindre est circonscrit à une sphère avec un diamètre égal à celui de la sphère, le volume et la surface du cylindre sont une fois et demie le volume et la surface de la sphère.

(RÉ)APPRENDRE À M AVEC LA MÉTHODE D

On a tous appris à multiplier ensemble deux nombres entiers. Une partie de la communauté mathématique a même cru, au XX^e siècle, que la méthode enseignée dans nos écoles était la meilleure méthode qui puisse exister pour ce faire. Et s'il en était autrement ?

Nadia Lafrenière
Université Dartmouth
College

Si on vous demande de multiplier
45 758 780 × 96 803 528

Multiplication

```
  45 758 780
× 96 803 528
-----
 366 070 240
 915 175 600
22 879 390 000
137 276 340 000
      0
36 607 024 000 000
274 552 680 000 000
4 118 290 200 000 000
-----
4 429 611 340 975 840
```

sans calculatrice, comment feriez-vous ? Il y a fort à parier que vous écririez les nombres un au-dessus de l'autre, puis multiplieriez le nombre du haut par la dernière décimale du nombre du bas (les unités), un chiffre à la fois et à partir de la droite.

Après avoir obtenu le produit du nombre du haut par l'unité du bas, vous feriez des calculs similaires avec les dizaines du

bas, avant de procéder avec les centaines, etc. Vous vous rendriez facilement compte que le processus est bien long...

Avec cette méthode, on doit multiplier chaque chiffre du premier nombre avec chaque chiffre du deuxième nombre. En cours de route, si on multiplie deux nombres de n chiffres ensemble, ça revient à faire n^2 multiplications élémentaires – c'est-à-dire des multiplications de nombres entiers entre 0 et 9.

Ainsi, si on veut multiplier deux nombres de 4 chiffres chacun, ça fait 16 multiplications élémentaires. Si on désire multiplier deux nombres de 8 chiffres chacun, on doit exécuter 64 multiplications élémentaires, ce qui est beaucoup !

Bien que l'on ait souhaité faire une seule multiplication, on finit par faire n^2 multi-

plications élémentaires, ainsi que quelques additions. On verra plus tard que les additions sont beaucoup plus faciles que les multiplications.

Peut-on faire mieux ?

*Pourrait-on faire moins de multiplications élémentaires ?
Pourrait-on même imaginer en faire significativement moins ?*

Ce sont des questions que posa, en 1960, le mathématicien Andreï Kolmogorov à ses étudiants. Kolmogorov croyait profondément que c'était impossible, notamment parce qu'une façon aussi efficace de multiplier n'avait toujours pas été développée, même après des milliers d'années à multiplier. Il en était si persuadé qu'il organisa un séminaire afin de prouver sa conjecture. Après exactement une semaine, Anatoli Karatsuba, âgé de seulement 23 ans, décrivit un algorithme qui requiert beaucoup moins de multiplications élémentaires que l'algorithme qu'on a tous appris à l'école. C'était suffisant pour que Kolmogorov partage la bonne nouvelle et décide que le séminaire était terminé.

Alors, comment fait-on des multiplications plus efficacement ?

Le principe de l'algorithme de Karatsuba s'appelle « diviser pour régner ». Ça veut dire qu'on divise le problème initial (la multiplication de deux grands nombres), en plusieurs petits problèmes. Par exemple, au lieu de mul-

MULTIPLIER DE KARATSUBA

multiplier deux nombres à n chiffres ensemble, on pourrait faire quelques multiplications de nombres à $n/2$ chiffres. Pour obtenir un gain d'efficacité, il y a deux conditions :

- Le problème original est un processus relativement long. Mathématiquement, on dira que « sa complexité n'est pas linéaire ». C'est le cas de la multiplication, étant donné que faire deux multiplications de deux nombres de $n/2$ chiffres requiert moins de multiplications élémentaires que de faire une seule multiplication de nombres de n chiffres.
- Le nombre de petits problèmes doit être petit. Créer trop de petits problèmes pourrait même nous faire perdre de l'efficacité par rapport au problème original.

Dans le cas de la méthode de Karatsuba, on divise la multiplication de deux nombres de n chiffres en trois multiplications de $n/2$ chiffres. On doit aussi faire des additions et des soustractions pour recombinaison les résultats, mais celles-ci nécessitent peu d'opérations en comparaison (voir encadré).

Par exemple, pour multiplier ensemble deux nombres de huit chiffres, on doit les séparer en nombres de quatre chiffres. Pour chaque nombre de huit chiffres, on crée deux nombres de quatre chiffres: celui de gauche et celui de droite. Ensuite, on

multiplie ensemble les nombres de gauche et ceux de droite. Ainsi, au lieu de faire $45\,758\,780 \times 96\,803\,528$, on fait $4\,575 \times 9\,680$ et $8\,780 \times 3\,528$. On doit faire une multiplication supplémentaire pour combiner les deux. En multipliant ensemble des nombres de quatre chiffres, même trois fois, on sauve du temps. Là où le gain de temps est considérable, c'est qu'on applique aussi la même méthode pour faire les « plus petites » multiplications, c'est-à-dire celles avec moins de décimales.

Ceci permet un gain de temps considérable, comme expliqué plus bas.

Comment ça marche ?

Disons que l'on souhaite multiplier deux nombres, a et b , qui ont chacun un nombre pair de chiffres, par exemple $2k$ chiffres¹. On peut réécrire les nombres a et b comme

$$a = a_1 \times 10^k + a_2 \text{ et } b = b_1 \times 10^k + b_2,$$

dans lesquels a_1 , a_2 , b_1 et b_2 sont des nombres à k chiffres. Utiliser la distributivité pour obtenir le produit $a \times b$ nous donne

$$(a_1 \times 10^k + a_2) \times (b_1 \times 10^k + b_2) = a_1 \times b_1 \times 10^{2k} + (a_1 \times b_2 + b_1 \times a_2) \times 10^k + a_2 \times b_2.$$

Même si les multiplications par des puissances de 10 sont en quelque sorte gratuites (on « ajoute des 0 »), on n'atteint pas le gain d'efficacité promis. Outre les multiplications par des puissances de 10, on fait quatre multiplications de nombre à k chiffres... comme dans la méthode initiale !

1. L'algorithm peut facilement être adapté à la multiplication de deux nombres ayant un nombre impair de chiffres.



Anatoli Karatsuba (1937-2008)

Anatoli Alekseïevitch Karatsuba est un mathématicien russe né le 31 janvier 1937 à Grozny et mort le 28 septembre 2008 à Moscou. Il est notamment connu pour son algorithme de multiplication, qui est la première méthode de multiplication rapide : l'algorithme de Karatsuba.

En 1966, il a soutenu sa thèse *The method of trigonometric sums and intermediate value theorem*.

En 1975, il a publié *Foundations of Analytic Number Theory* qui fut réédité en 1983.

Il a été directeur du département de théorie des nombres à l'Institut de mathématiques Steklov de l'Académie des sciences de Russie.

Il a notamment travaillé sur les séries de Fourier et le théorème de Moore.

L'addition, bien plus simple que la multiplication!

Puisqu'on apprend à additionner deux grands nombres ensemble deux ans avant d'apprendre à les multiplier, l'addition doit être bien plus simple que la multiplication, n'est-ce pas? Certes, le résultat de la somme de deux entiers positifs est généralement beaucoup plus petit que celui du produit. Mais la simplicité peut aussi s'exprimer par le nombre d'opérations à faire pour obtenir le résultat désiré.

Même pour un ordinateur, l'addition est plus simple. Qu'est-ce que ça veut dire ?

Que beaucoup moins d'opérations sont nécessaires. En effet, lorsqu'on additionne deux nombres à n chiffres, le résultat a au plus $n + 1$ chiffres. Et pour y arriver, on ne fait qu'additionner les chiffres qui sont à la même position. Contrairement à la multiplication, le nombre d'opérations élémentaires correspond donc au nombre de chiffres de chacun des nombres. On dira alors que d'additionner deux nombres prend sensiblement le même temps que de lire ces nombres. On ne pourrait pas faire plus vite !

Exemple

Multiplier 2457 par 6819

1. Séparer les produits à effectuer

$$\begin{array}{r} 2457 \\ \times 6819 \\ \hline \end{array} \rightarrow \begin{array}{r} 24 \overset{!}{:} 57 \\ 68 \overset{!}{:} 19 \\ \hline \end{array} \rightarrow \begin{array}{r} 24 \\ \times 68 \\ \hline \end{array} \quad \begin{array}{r} 57 \\ \times 19 \\ \hline \end{array}$$

2. Multiplier les chiffres significatifs*

$$\begin{array}{r} 24 \\ \times 68 \\ \hline 1632 \end{array}$$

3. Multiplier les chiffres les moins significatifs*

$$\begin{array}{r} 57 \\ \times 19 \\ \hline 1083 \end{array}$$

4. Étapes de reconstitution du produit

- 4a. Pour chaque nombre, additionner les deux parties

$$\begin{array}{r} 24 \overset{!}{:} 57 \\ 68 \overset{!}{:} 19 \\ \hline \end{array} \rightarrow \begin{array}{r} 24 + 57 = 81 \\ 68 + 19 = 87 \end{array}$$

- 4b. Effectuer le produit des résultats de la partie 4a*

$$\begin{array}{r} 81 \\ \times 87 \\ \hline 7047 \end{array}$$

- 4c. Soustraire les résultats des parties 2 et 3

$$\begin{array}{r} 7047 \\ -1632 \\ \hline -1083 \\ \hline 4332 \end{array}$$

5. Additionner les résultats des étapes 2, 3 et 4c pour obtenir le produit recherché

$$\begin{array}{r} 1632 \\ 4332 \\ 1083 \\ \hline 16754283 \end{array} \rightarrow \begin{array}{r} 2457 \\ \times 6819 \\ \hline 16754283 \end{array}$$

* Les « petites multiplications » sont aussi effectuées avec la méthode de Karatsuba.

En fait, l'algorithme de Karatsuba utilise un truc supplémentaire : Il suffit de remarquer qu'on peut, au lieu de faire $(a_1 \times b_2 + b_1 \times a_2)$ dans l'exemple ci-dessus, écrire

$$(a_1 + a_2) \times (b_1 + b_2) - a_1 \times b_1 - a_2 \times b_2$$

Si cela semble encore plus compliqué qu'au départ, on réduit en fait le nombre de multiplications à faire, puisque

$$a_1 \times b_1 \text{ et } a_2 \times b_2$$

sont des expressions qu'on doit déjà calculer. Ainsi, le nombre de multiplications de nombres à k chiffres passe de quatre à trois. Certes, on doit faire des additions, mais celles-ci sont beaucoup moins coûteuses (voir encadré).

Pour multiplier deux nombres de $2k$ chiffres chacun, on devait faire $4k^2$ multiplications élémentaires avec la méthode apprise à l'école. Maintenant, on doit faire trois fois plus de multiplications élémentaires qu'il en faut pour multiplier des nombres à k chiffres. Si T est une fonction qui compte le nombre de multiplications élémentaires effectuées pour multiplier deux nombres à n chiffres², on a alors la récurrence

$$T(n) = 3T(n/2), \text{ quand } n > 1, \text{ et } T(1) = 1.$$

Celle-ci se résout par l'expression

$$T(n) = n^{\log_2(3)} \approx n^{1,58}.$$

2. Après avoir additionné les nombres de $n/2$ chiffres, il est possible qu'une des multiplications se fasse avec des nombres de $n/2 + 1$ chiffres. Ceci n'a pas d'impact sur l'exposant de la solution à la récurrence, donc on se permet d'omettre ce détail.

Une belle amélioration sur la méthode apprise à l'école, qui nous faisait faire n^2 multiplications élémentaires.

Pour avoir une idée du gain d'efficacité, on peut regarder le tableau ci-contre.

Y a-t-il plus efficace ?

Il y a une soixantaine d'années, la découverte de la méthode de Karatsuba a permis l'ouverture d'un tout nouveau champ de recherche. Depuis, d'autres méthodes encore plus efficaces ont été mises au jour. Certes, plusieurs d'entre elles sont difficiles à mémoriser, ce qui les rend peu réalistes pour un algorithme qui serait enseigné à l'école. En revanche, elles s'avèrent de formidables outils pour les ordinateurs qui doivent accomplir de nom-

Gain d'efficacité en utilisant la méthode de Karatsuba

n	n^2	$n^{1,58}$	Gain d'efficacité $n^2/n^{1,58}$
4	16	8,94	1,79
10	100	38,02	2,63
300*	90 000	8 200,71	10,97

* 300 correspond environ au nombre de chiffres des nombres utilisés dans les systèmes de cryptographie. La multiplication est au coeur de certains algorithmes de cryptographie.

breuses multiplications, ne serait-ce que pour crypter vos informations sensibles. Cela dit, si je vous demande de multiplier de grands nombres, peut-être gagneriez-vous à aller chercher la calculatrice !

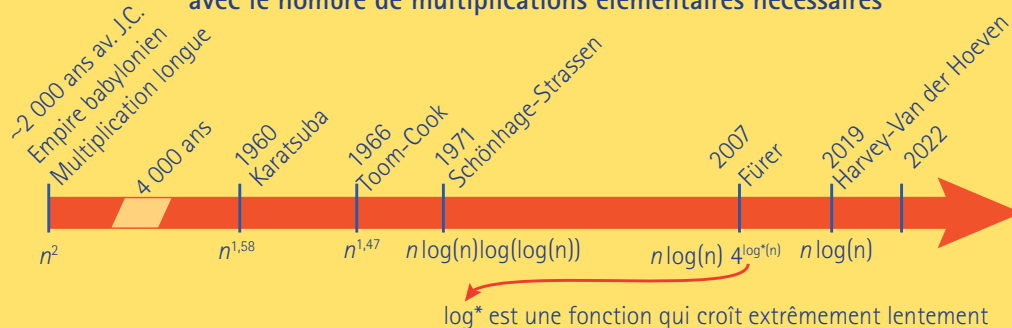
La façon parfaite de multiplier ?

Avec la découverte de son algorithme, Karatsuba a créé un nouveau domaine de recherche. Il était désormais possible de se demander quelle est la meilleure méthode pour multiplier, puisque ce n'est pas celle apprise à l'école. En 1971, Arnold Schönhage et Volker Strassen ont développé un algorithme de multiplication beaucoup plus rapide que la méthode de Karatsuba en utilisant la transformée de Fourier rapide. Cette approche a permis de réduire le nombre de multiplications élémentaires nécessaires pour multiplier deux nombres de n chiffres à, asymptotiquement, $n \log(n) \log(\log(n))$, ce qui est très significatif pour des grandes valeurs de n . Ils ont alors émis la conjecture que la parfaite façon de multiplier devrait nécessiter, asymptotiquement, $n \log(n)$ multiplications élémentaires. La justification derrière cette hypothèse ? Le nombre d'opérations nécessaires pour exécuter une opération aussi fondamentale que la multiplication devait être simple à exprimer.

Bien que plusieurs améliorations importantes aient eu lieu à travers les époques, ce n'est qu'en 2019 que David Harvey et Joris van der Hoeven ont trouvé un algorithme de multiplication avec le nombre désiré d'opérations élémentaires, c'est-à-dire $n \log(n)$. Pourrait-on encore trouver une méthode significativement plus rapide ? Ce n'est pas impossible, puisque démontrer qu'on ne pourrait trouver de meilleures méthodes est extrêmement difficile. Toutefois, une grande partie de la communauté mathématique voit dans les travaux d'Harvey et van der Hoeven la méthode parfaite pour multiplier... en théorie !

En pratique, un ordinateur choisira la méthode de Karatsuba pour multiplier de petits nombres, l'algorithme de Toom et Cook pour multiplier des nombres de taille intermédiaire et l'algorithme de Schönhage et Strassen pour les grands nombres.

Développement d'algorithmes de multiplication, avec le nombre de multiplications élémentaires nécessaires



Le déménagement miraculeux

Rubrique des **Paradoxes**

Jean-Paul Delahaye
Université des Sciences
et Technologies de Lille

Les âges des cinq habitants de la rue Kurt Gödel sont 8, 14, 20, 23 et 35 ; leur âge moyen est donc :

$$\frac{8 + 14 + 20 + 23 + 35}{5} = 20 \text{ ans.}$$

Les 6 habitants de la rue Alan Turing ont respectivement : 25, 30, 35, 40, 45 et 59 ans. Leur âge moyen est donc :

$$\frac{25 + 30 + 35 + 40 + 45 + 59}{6} = 39 \text{ ans.}$$

Jacques, qui habite la rue Gödel, a 35 ans. Il déménage et va habiter dans la rue Turing. Maintenant, l'âge moyen dans la rue Gödel est devenu :

$$\frac{8 + 14 + 20 + 23}{4} = 16,25 \text{ ans}$$

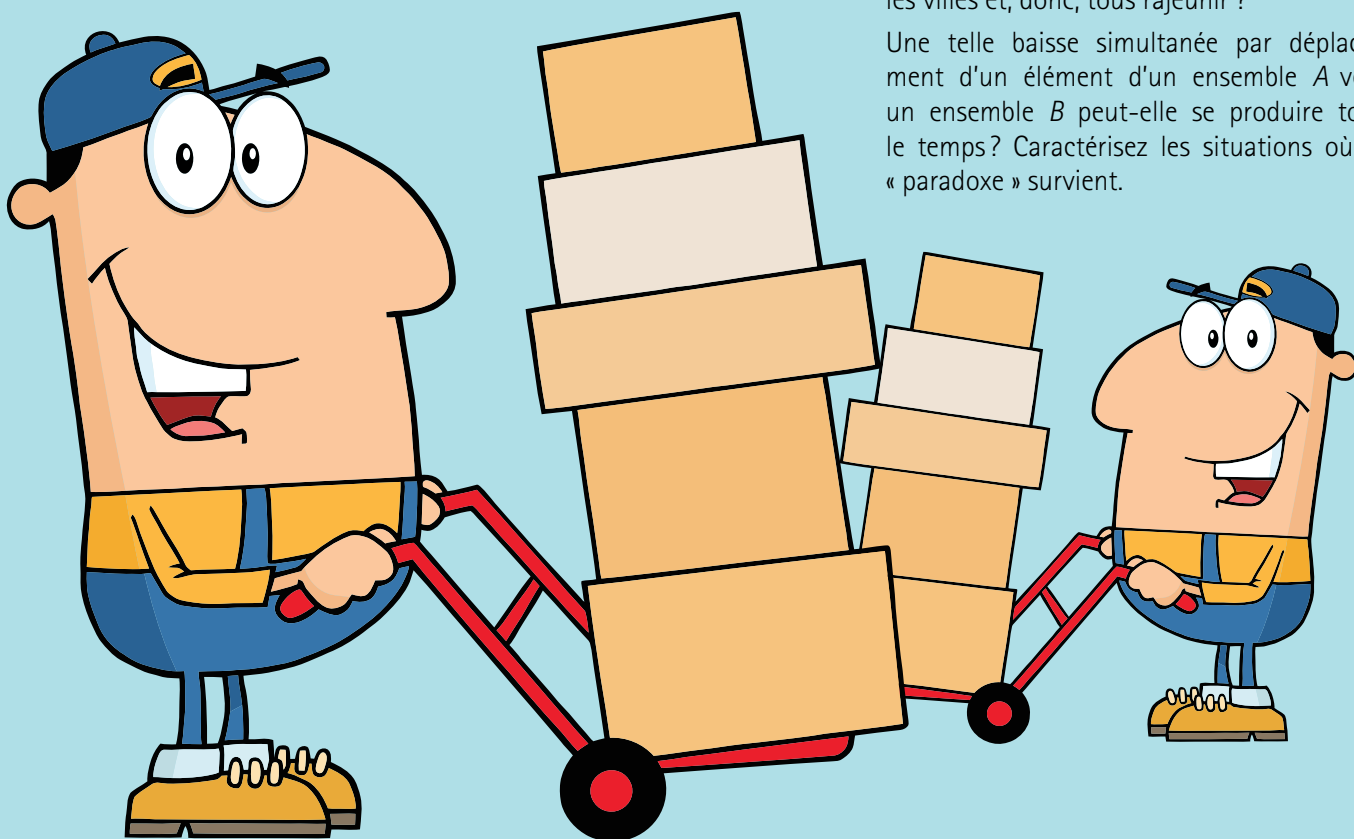
et l'âge moyen dans la rue Turing s'établit à :

$$\frac{25 + 30 + 35 + 35 + 40 + 45 + 59}{7} = 38,42 \text{ ans.}$$

Ne trouvez-vous pas paradoxal que les moyennes des âges dans les deux rues aient toutes les deux diminuées ?

En organisant des déménagements de ce type, ne pourrait-on pas alors faire baisser les âges moyens de toutes les rues dans toutes les villes et, donc, tous rajeunir ?

Une telle baisse simultanée par déplacement d'un élément d'un ensemble A vers un ensemble B peut-elle se produire tout le temps ? Caractériser les situations où le « paradoxe » survient.



Encore une histoire de chapeaux

Neuf joueurs portent des chapeaux dont la couleur est rouge, noire ou blanche. Chacun peut voir tous les autres chapeaux mais pas le sien. Les chapeaux ont été tirés au hasard à l'aide d'un dé (1 et 2 donnent noir, 3 et 4 donnent rouge, 5 et 6 donnent blanc). L'arbitre annonce que chaque joueur doit deviner la couleur de son chapeau en voyant les autres chapeaux, et que, si au moins trois d'entre eux donnent la bonne réponse, alors ils auront gagné un voyage à Londres tous ensemble. Les joueurs ont pu convenir d'une stratégie collective avant que les chapeaux soient disposés sur leurs têtes, mais ils donnent leur réponse simultanément sans avoir aucun échange entre eux une fois les chapeaux en place. En répondant au hasard, les joueurs auront une chance non négligeable de perdre. Précisément, ils perdent si 7, 8 ou 9 joueurs se trompent, ce qui se produit dans 37,7 % des cas. Même si cela vous semble paradoxal, ils peuvent réduire leur risque de perdre à 0, en convenant avant le jeu d'une stratégie astucieuse qui les fera gagner de manière certaine quelle que soit la répartition des chapeaux sur leur tête. Quelle est cette stratégie ?

Solution

Les joueurs se séparent en trois groupes de trois :

Groupe A : A_0, A_1, A_2 ,

Groupe B : B_0, B_1, B_2 ,

Groupe C : C_0, C_1, C_2 .

Dans chaque groupe, ils vont s'arranger pour que l'un d'eux donne la bonne réponse.

Attribuons un numéro à chacune des couleurs 0, 1 ou 2. Le joueur A_0 va jouer en proposant, pour son chapeau, la couleur telle que la somme des trois couleurs du groupe A fasse 0 ou 3 (c'est-à-dire $0 \pmod 3$). Si, par exemple, il voit 1 et 1 pour les deux autres joueurs de son groupe, il parie 1 pour la couleur de son chapeau ; s'il voit 2 et 1, il parie 0 pour lui, etc.

Le joueur A_1 va jouer en proposant, pour son chapeau, la couleur telle que la somme des trois couleurs du groupe A fasse 1 ou 4 (c'est-à-dire $1 \pmod 3$).

Le joueur A_2 va jouer en proposant, pour son chapeau, la couleur telle que la somme des trois couleurs du groupe A fasse 2 ou 5 (c'est-à-dire $2 \pmod 3$).

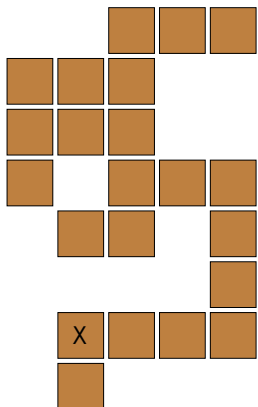
L'un des trois aura correctement deviné la couleur de son chapeau, car la somme des trois couleurs du groupe A vaut

- soit $0 \pmod 3$ (auquel cas A_0 aura bon),
- soit $1 \pmod 3$ (auquel cas A_1 aura bon),
- soit $2 \pmod 3$ (auquel cas A_2 aura bon).

Les joueurs du groupe B conviennent d'une méthode analogue, ainsi que ceux du groupe C. Dans chacun des groupes, un joueur devinera la couleur de son chapeau. Au total, trois joueurs (exactement) auront deviné la couleur et donc ils gagneront.



Section problèmes



Des dames sur d'étranges échiquiers

1. Prouver l'énoncé utilisé dans la preuve de Pólya. Généralement pour qu'une liste (r_1, \dots, r_n) soit une solution au problème des n dames sur un échiquier toroïdal, il faut et il suffit que :
 - I. la liste (r_1, \dots, r_n) contienne tous les nombres de 1 à n , on dit donc que c'est une *permutation*;
 - II. la liste $((r_1 + 1) \bmod n, \dots, (r_n + n) \bmod n)$ soit une permutation (avec $0 = n \bmod n$);
 - III. la liste $((r_1 - 1) \bmod n, \dots, (r_n - n) \bmod n)$ soit une permutation (avec $0 = n \bmod n$).

2. Quelle est la borne supérieure équivalente pour le problème de domination des tours sur un polyomino de M cases? (Indice : quel serait alors le pavage nécessaire?)
3. Que se passe-t-il si on essaie la même construction que dans l'article pour la figure ci-contre avec comme racine, le X ?
4. Que se passe-t-il si une des couleurs est vide dans la preuve du théorème d'Alpert-Roldán?
5. Le véritable théorème prouvé par Alpert et Roldán fonctionne en fait sur des polyominos de *dimension* d . Concentrons-nous sur $d=3$ question de ne pas perdre l'intuition géométrique. Faut-il changer quelque chose à la preuve présentée pour qu'elle soit valide alors?

Indices

1. Vous parcourez une distance d que l'on divise en n parties égales parcourues aux vitesses respectives v_1, \dots, v_n . Montrer que votre vitesse moyenne est

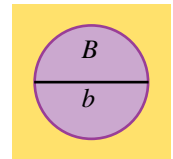
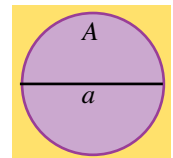
$$v = \frac{n}{\left(\frac{1}{v_1} + \dots + \frac{1}{v_n}\right)}$$

Ce nombre v est appelé la *moyenne harmonique* de v_1, \dots, v_n .

2. Dans le cas de deux nombres positifs x_1 et x_2 , montrer que leur moyenne arithmétique est toujours supérieure ou égale à leur moyenne géométrique, laquelle est supérieure à leur moyenne harmonique, et que ces trois moyennes sont toutes dans l'intervalle $[\min(x_1, x_2), \max(x_1 \text{ et } x_2)]$.

Comparaison d'aires

1. Expliquer en quoi consistait la méthode des leviers d'Archimède et indiquer quelques conjectures obtenus par cette méthode.
2. Quel est le principe fondamental de la méthode d'exhaustion selon la formulation d'Eudoxe? Illustrer numériquement ce postulat.
3. Pourquoi Archimède considérait-il que les résultats obtenus étaient simplement des conjectures qu'il fallait démontrer par exhaustion?
4. Décrire comment Archimède utilisait la méthode d'exhaustion pour démontrer des propriétés.
5. Archimède a montré que :



Lorsqu'un cylindre est circonscrit à une sphère, le volume et la surface du cylindre sont une fois et demie le volume et la surface de la sphère.

- a) En utilisant le symbolisme moderne, déterminer les formules du volume et de l'aire de la surface du cylindre considéré par Archimède.
 - b) En utilisant la relation établie par Archimède, déterminer les formules usuelles du volume et de l'aire de la surface de la sphère.
6. En supposant, comme l'a fait Archimède, que le rapport des aires est plus petit que le rapport des carrés des diamètres. Montrer que cela entraîne une contradiction.

Pour en savoir plus!

Applications des mathématiques

Des dames sur d'étranges échiquiers

- ALPERT, H. et ROLDÁN, É. *Art Gallery Problem with Rook and Queen*. *Vision Graphs and Combinatorics* 37:2 (2021), 621–642.
- Un étudiant d'Érika Roldán, Christoph Muessig, a mis en ligne un *programme* permettant de tester le problème de domination des tours sur des polyominos aléatoires. <https://mygame.page/art-gallery-with-rooks-guards>
- Pour plus de problèmes mathématico-échiquéens, le livre suivant est tout désigné. WATKINS, J.J. *The mathematics of chessboard problems*. Princeton University Press, 2004.

L'utilisation d'indices pour combiner des informations

- <https://hdr.undp.org/system/files/documents/hdr2016froverviewwebpdf.pdf> (voir notes techniques à partir de la page 185)
- https://fr.wikipedia.org/wiki/Refroidissement_%C3%A9olien
- https://fr.wikipedia.org/wiki/Indice_humidex
- <https://louernos-nature.fr/indices-de-diversite-ecologie-ecosystemes/>

Nombres

(Ré)apprendre à multiplier par la méthode de Karatsuba

- KARATSUBA, Anatolii. *The complexity of computations*, dans *Proceedings of the Steklov Institute of Mathematics*, n° 211, pages 169–183, janvier 1995. L'auteur décrit l'histoire de la découverte de son algorithme.
- HARTNETT, Kevin. *Mathematicians Discover the Perfect Way to Multiply*, dans *Quanta Magazine*, 11 avril 2019. En ligne (<https://www.quantamagazine.org/mathematicians-discover-the-perfect-way-to-multiply-20190411/>). L'article décrit les progrès récents dans la quête de l'algorithme parfait pour la multiplication.

Accromath est une publication de l'Institut des sciences mathématiques (ISM) et du Centre de recherches mathématiques (CRM). La revue s'adresse surtout aux étudiantes et étudiants d'école secondaire et de cégep ainsi qu'à leurs enseignantes et enseignants.

ISM

Institut des sciences mathématiques

L'Institut des sciences mathématiques est une institution unique dédiée à la promotion et à la coordination de l'enseignement et de la recherche en sciences mathématiques au Québec. En réunissant huit départements de mathématiques des universités québécoises (Concordia, Université Laval, McGill, Université de Montréal, UQAM, UQTR, Université de Sherbrooke, Bishop's), l'Institut rassemble un grand bassin d'expertises en recherche et en enseignement des mathématiques. L'Institut anime de nombreuses activités scientifiques, dont des séminaires de recherche et des colloques à l'intention des professeurs et des étudiants avancés, ainsi que des conférences de vulgarisation données dans les cégeps. Il offre également plusieurs programmes de bourses d'excellence.

L'ISM est financé par le Ministère de l'Enseignement supérieur et par ses huit universités membres.

CRM

CENTRE
DE RECHERCHES
MATHÉMATIQUES

Le Centre de recherches mathématiques est un centre national pour la recherche fondamentale en mathématiques et ses applications. Les scientifiques du CRM comptent plus d'une centaine de membres réguliers et de stagiaires postdoctoraux. Lieu privilégié de rencontre, le Centre est l'hôte chaque année de nombreux visiteurs et d'ateliers de recherche internationaux.

Les activités scientifiques du CRM comportent deux volets principaux : les projets de recherche qu'entreprennent ses laboratoires, et les activités thématiques organisées à l'échelle internationale. Ces dernières, ouvertes à tous les domaines, impliquent des chercheurs du CRM et d'autres universités. Afin d'assurer une meilleure diffusion des résultats de recherches de ses collaborateurs, le CRM a lancé en 1989 un programme de publications en collaboration avec l'American Mathematical Society et avec Springer.

Le CRM est principalement financé par le CRSNG (Conseil de recherches en sciences naturelles et en génie du Canada), le FQRNT (Fonds québécois de recherche sur la nature et les technologies), l'Université de Montréal, et par six autres universités au Québec et en Ontario.

Accromath bénéficie de l'appui de la Dotation Serge-Bissonnette du CRM.

