

R

ENDEZ-VOUS

P.108 Logique & calcul
P.114 Idées de physique
P.118 Chroniques de l'évolution
P.122 Science & gastronomie

COÏNCIDENCES SURPRENANTES, MAIS BANALES

Des erreurs de jugement nous conduisent à voir dans certaines coïncidences des phénomènes incroyables et à leur rechercher d'impossibles explications.

Il est vrai que, en moyenne dans une école primaire, plus les élèves lisent rapidement, plus ils sont grands. En concluons-nous qu'apprendre à lire fait grandir? Plus sérieusement, une étude statistique de Franz Messerli, de l'université Columbia, menée avec toute la rigueur méthodologique nécessaire et publiée en 2012 dans la revue *New England Journal of Medicine*, a établi qu'il existe une étroite corrélation entre la consommation de chocolat par habitant d'un pays et le nombre de prix Nobel obtenus par ce pays par million d'habitants. En déduisons-nous que les chercheurs doivent manger du chocolat pour augmenter leurs chances de se voir attribuer le fameux prix?

L'explication ne serait-elle pas plutôt que le système social et éducatif des pays riches favorise la bonne recherche et donc l'attribution des prix Nobel, et que cette même richesse favorise l'achat par tous de chocolat? Les deux faits sont

bien liés, mais seulement parce qu'ils sont la conséquence d'une même cause, pas parce que l'un implique l'autre. Quand deux faits sont corrélés, cela ne signifie pas que l'un est la conséquence de l'autre, mais parfois seulement qu'un troisième facteur les entraîne tous les deux. La recherche des liens de causalité entre faits doit être menée avec précaution (voir l'article d'Isabelle Drouet, «Des corrélations à la causalité», *Pour la Science* n° 440, juin 2014).

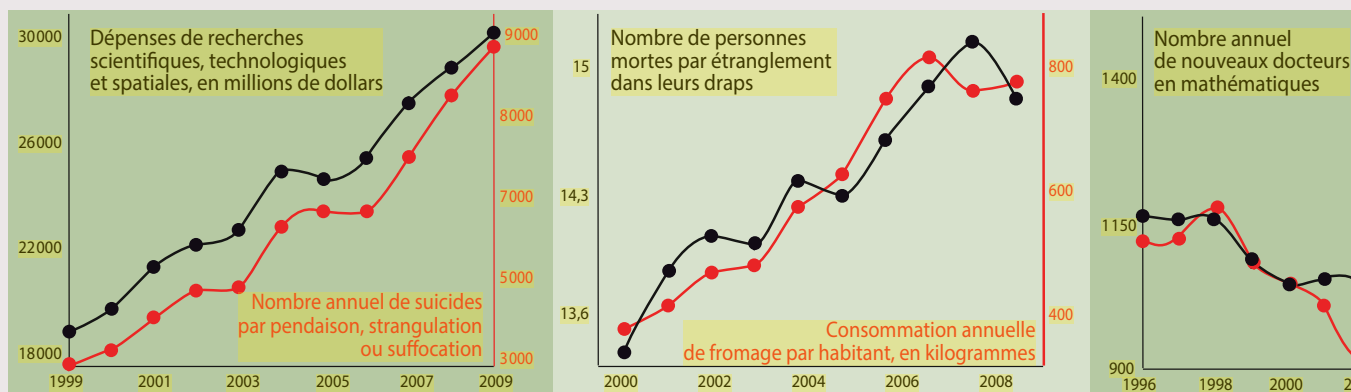
DU HASARD PUR?

Hormis ces situations où, malgré tout, la corrélation repérée a une explication satisfaisante faisant intervenir un troisième facteur, tel que l'âge ou la richesse économique, il existe des cas où, en définitive, la seule explication est le hasard.

Les quelques exemples donnés par la figure ci-dessous permettent de comprendre ce que cela signifie. Ces corrélations illusoire ont été

Source : d'après www.tylenigen.com

DES CORRÉLATIONS ÉTONNANTES MAIS FORTUITES, AUX ÉTATS-UNIS



minutieusement collectées par Tyler Vigen quand il était étudiant à la faculté de droit de Harvard, à Cambridge dans le Massachusetts. Il travaille aujourd'hui au Boston Consulting Group et a publié un livre avec ses amusantes découvertes (voir la bibliographie et les sites [www.tylervigen.com/spurious-correlations](http://tylervigen.com/spurious-correlations) ou <http://tylervigen.com/old-version.html>).

L'observation sur un même graphique de la courbe indiquant année après année le nombre de suicides aux États-Unis par pendaison ou suffocation et de la courbe indiquant année après année les dépenses de recherches scientifiques américaines est particulièrement troublante. Quand l'une des courbes baisse ou monte, l'autre la suit en parallèle. La synchronisation est presque parfaite. Elle se traduit en termes mathématiques par un coefficient de corrélation de 0,992082, proche de 1, le maximum possible. Cela établit-il qu'il existe un lien véritable entre les deux séries de nombres ?

Cela est si peu vraisemblable qu'il faut se rendre à l'évidence : c'est uniquement le fait du hasard, ou comme on le dit, une coïncidence. La seule façon raisonnable d'expliquer le parallélisme des deux courbes et des nombreux cas du même type proposés par Tyler Vigen est la suivante :

- Il a collecté un très grand nombre de séries statistiques.
- Il a su les comparer systématiquement, ce qui lui a permis d'en trouver ayant des allures similaires, d'où des coefficients de corrélation proches de 1.
- On ne doit pas s'en étonner : c'était inévitable du fait du très grand nombre de séries numériques pris en compte.

La quantité de données de base est l'explication et Tyler Vigen a détaillé sa méthode : après avoir réuni des milliers de séries numériques, il les a confiées à son ordinateur pour qu'il recherche systématiquement les couples de séries donnant de bons coefficients de corrélation ; il a ensuite fait appel aux étudiants de

sa faculté pour qu'ils lui indiquent, par des votes, les couples de courbes jugés les plus spectaculaires parmi ceux présélectionnés par l'ordinateur.

Cette façon de procéder porte en anglais le nom de *data dredging*, que l'on peut traduire par « dragage de données ». Il est important d'avoir conscience des dangers que créent de tels traitements, surtout aujourd'hui où la collecte et l'exploitation de quantités colossales d'informations de toutes sortes sont devenues faciles et largement pratiquées. La nouvelle discipline informatique qu'est la fouille de données (en anglais, *data mining*) pourrait être victime sans le savoir de ces corrélations illusoire : sans précaution, elle peut prendre ces dernières pour de véritables liens entre des séries de données en réalité totalement indépendantes.

UNE VICTIME : LA RECHERCHE MÉDICALE

La recherche médicale est fréquemment confrontée au problème de telles corrélations douteuses. Dans un article de 2011, Stanley Young et Alan Karr, de l'Institut américain d'études statistiques, citaient 12 études publiées qui établissaient de soi-disant liens entre la consommation de vitamine et la survenue de cancers. Ces études avaient parfois été réalisées en utilisant un protocole avec placebo et mesure en double aveugle, et semblaient donc sérieuses. Pourtant, quand elles ont été reprises, dans certains cas plusieurs fois, les nouveaux résultats ont contredit les résultats initiaux. Outre que la tentation est grande d'arrondir un peu les chiffres pour qu'ils parlent dans un sens bien clair et permettent la publication d'un article, il se peut simplement que les auteurs de ces travaux aux conclusions impossibles à reproduire aient été victimes d'une situation de mise en corrélation illusoire, comme Tyler Vigen en a repéré de nombreuses.

Ce hasard trompeur provenant de l'exploration d'un trop grand nombre de combinaisons est une erreur de jugement statistique. On se >

L'AUTEUR



JEAN-PAUL DELAHAYE
professeur émérite
à l'université de Lille
et chercheur au Centre
de recherche en
informatique, signal
et automatique de Lille
(Cristal)

Jean-Paul Delahaye
a récemment publié :
**Les mathématiciens
se plient au jeu,**
une sélection de ses
chroniques parues
dans *Pour la Science*
(Belin, 2017).

