# COÏNCIDENCES SURPRENANTES

Coïncidences surprenantes, mais banales



https://www.lemonde.fr/blog/binaire/2018/04/09/coincidences-surprenantes-mais-banales/

*Nous avons publié récemment un article expliquant pourquoi les pics de naissances se rencontrent en septembre. Les chercheurs avaient exploité, de façon pertinente, des corrélations entre évènements en utilisant des informations issues de Google trends. Cependant, il nous arrive de voir dans certaines coïncidences des phénomènes incroyables et à leur rechercher d'impossibles explications. L'introduction généralisée des méthodes et des outils d'extraction d'information a amplifié ce risque et il convient donc de les utiliser avec clairvoyance. Jean-Paul Delahaye va nous aider à décrypter ce phénomène. Ce texte est extrait d'un article plus long publié dans Pour la Science (novembre 2017, n° 481, pp. 108-113) et que nous vous invitons à lire.*

**Pascal Guitton & Thierry Viéville**

Il est vrai que, en moyenne dans une école primaire, plus les élèves lisent rapidement, plus ils sont grands. En conclurons-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 déduirons-nous que les chercheurs doivent manger du chocolat pour augmenter leurs chances de se voir attribuer le fameux prix ?

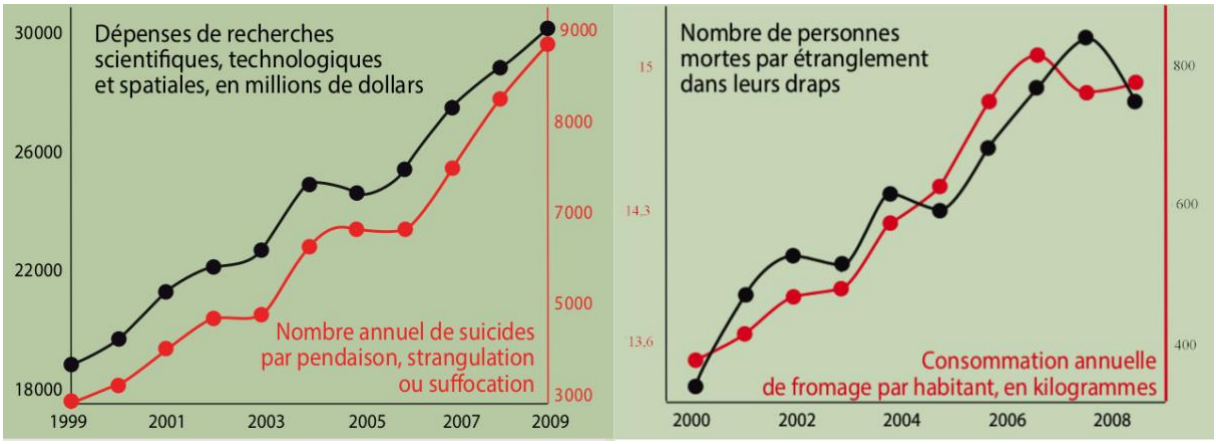
L'explication ne serait-elle pas plutôt que le système social et éducatif des pays riches favorise la bonne recherche et donc l'attribution des prix Nobel, et que cette même richesse favorise l'achat par tous de chocolat ? Les deux faits sont bien liés, mais seulement parce qu'ils sont la conséquence 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 illusoires ont été 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 http://www.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 ?




Des corrélations étonnantes mais fortuites, aux États-Unis © Pour la Science
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 illusoires : 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 trompe en croyant qu'une observation est significative et doit donc être expliquée, alors qu'elle devait se produire (elle ou une autre du même type) du fait du grand nombre d'hypothèses envisagées, parfois collectivement par l'ensemble des équipes de recherche, dans la quête de régularités statistiques.

***Data snooping***

Cette illusion est proche de celle dénommée *data snooping* (« furetage discret de données ») que l'on peut utiliser pour des tromperies délibérées. Pour en illustrer le fonctionnement, Halbert White, de l'université de Californie à San Diego, a suggéré la méthode suivante permettant à un journal financier d'augmenter le nombre de ses abonnés.

La première semaine, le journal envoie un exemplaire gratuit à 20 000 personnes ; dans la moitié des numéros envoyés, il est écrit que l'indice boursier (par exemple le Dow Jones) va monter, dans l'autre moitié le journal affirme que l'indice va baisser. Selon que l'indice a effectivement monté ou baissé, le journal envoie la semaine suivante, aux 10 000 adresses qui ont reçu la bonne prévision, un second numéro gratuit, avec dans la moitié des exemplaires l'annonce pour la semaine suivante que l'indice va monter et dans l'autre moitié qu'il va baisser. La troisième semaine, le journal envoie 5 000 numéros gratuits à ceux qui ont reçu la bonne anticipation deux semaines de suite, etc.

À chaque fois, le journal insiste sur le fait qu'il a correctement prévu la tendance depuis plusieurs semaines et propose un bulletin d'abonnement. Au bout de 10 semaines, il ne restera qu'une vingtaine d'envois à faire, mais pour être persuadé que le journal sait prédire la bonne tendance de l'indice boursier, rares sont les lecteurs potentiels qui attendent une

prévision exacte 10 fois de suite. Par conséquent, un bon nombre de nouveaux abonnements auront été souscrits à mesure du déroulement des envois.

Une autre situation de *data snooping* dans le traitement des données financières conduit à une navrante désillusion. On cherche des règles du type « Si aujourd'hui le cours de l'action A monte et que le cours de l'action B baisse et que…, alors le cours de l'action X montera demain ». On en écrit un grand nombre, voire on écrit toutes les règles de ce type comportant 10 éléments dans leurs prémisses. On sélectionne ensuite, à l'aide d'une série de données provenant du passé, les meilleures règles. On élimine toutes les règles qui se sont trompées avec les données passées, et on retient seulement celles qui ont toujours fourni une bonne prédiction, ou celles qui ont eu raison le plus souvent. On disposera alors inévitablement d'une série réduite de règles qui, si elles avaient été appliquées sur les données du test, auraient permis de gagner beaucoup d'argent… et qui pourtant ne rapporteront rien dans les semaines qui suivront leur mise en œuvre pour faire des achats et des ventes.

Bien sûr, le piège est connu des chercheurs et des méthodes ont été mises au point pour ne pas être victime de l'illusion. On cherchera par exemple à évaluer, avant de mener la recherche des bonnes règles, la probabilité que lorsque les données sont tirées au hasard l'une des règles de la liste envisagée fonctionne avec ce hasard, et on s'assurera que cette probabilité est proche de zéro.

**Loterie truquée ?**

Qui peut considérer comme normal qu'à quelques jours de distance, la même série de 6 numéros sorte d'un tirage du Loto ? C'est pourtant ce qui se produisit le 10 septembre 2009 pour le Loto bulgare. La série 4, 15, 23, 24, 35, 42 n'a rien d'extraordinaire, sauf qu'elle avait déjà été tirée 6 jours auparavant, le 4 septembre, par le même Loto bulgare. Une coïncidence incroyable, au point que le ministre des Sports Svilen Neïkov demanda l'ouverture d'une enquête. Aucune tricherie n'a été détectée, et d'ailleurs les deux tirages avaient eu lieu devant les caméras de la télévision. La chose fut considérée si extraordinaire que la presse dans le monde entier rapporta l'événement.

Plus étonnant peut-être, un an après, à un mois d'intervalle, le 21 septembre et le 16 octobre 2010, le Loto israélien sortit la série 13, 14, 26, 32, 33, 36, provoquant là encore l'étonnement mondial. Comme l'expliquait David Hand dans son article « Des coïncidences pas si étranges » (Pour la Science n°438, avril 2014), si l'on prend en compte le nombre de jeux de Loto dans le monde, et le nombre de tirages que chacun d'eux opère, souvent plusieurs fois par semaine, notre étonnement doit cesser.

Plus précisément, considérons le Loto bulgare où l'on tire 6 numéros entre 1 et 49, ce qui donne une probabilité de gagner de 1/13 983 816. Demandons-nous combien de tirages sont nécessaires pour que l'événement « deux d'entre eux donnent le même résultat » ait une probabilité supérieure à 50 % de se produire. La réponse est 4 404.

Le calcul est analogue à celui fait pour le célèbre paradoxe des anniversaires, selon lequel dès que 23 personnes sont réunies, la probabilité que deux d'entre elles aient la même date

anniversaire dépasse 50 %. Cela montre que les événements ressentis comme extraordinaires pour les loteries bulgare et israélienne sont en fait nécessaires. Ce n'est pas la survenue des tirages identiques qui est improbable, mais l'inverse : si tous les tirages étaient toujours différents, nous devrions nous en étonner et en rechercher l'explication. Une autre source d'étonnements provient des séries rapprochées d'événements rares, tels que les accidents d'avion. Les journalistes aiment mentionner une prétendue « loi des séries », pourtant inconnue des mathématiciens, qui expliquerait ces rapprochements jugés à la fois fortement improbables et explicables par cette introuvable loi. Jacques Chirac disait simplement : « Les emmerdes, ça vole toujours en escadrille !» Certaines analyses tentent d'en identifier l'origine en parlant d'une « attente excessive d'étalement » (voir mon article « Notre vision du hasard est bien hasardeuse », Pour la Science n° 293, mars 2002) : notre intuition nous souffle à tort que, par exemple, les dates des accidents d'avion doivent être régulièrement espacées, alors que les statistiques nous montrent autre chose. Cette attente excessive d'étalement a été clairement analysée dans les cas précis des accidents d'avions par Élise Janvresse et Thierry de la Rue, de l'université de Rouen.

En août 2005, une série de cinq catastrophes aériennes s'est produite dans un intervalle de 22 jours, faisant plusieurs centaines de morts au total. Cela nous semble extraordinaire, mais l'est-ce vraiment ? Élise Janvresse et Thierry de la Rue proposent une belle analyse du problème dans leur petit livre La Loi des séries, hasard ou fatalité ? (Le Pommier, 2007). Ils concluent que la probabilité que, durant une année donnée, il se produise 5 accidents aériens graves ou plus dans une fenêtre de 22 jours est 11 %. C'est assez faible, mais cela rend la série malheureuse de 2005 peu étonnante, d'autant plus que le calcul mené ne prend pas en compte les variations saisonnières du trafic aérien qui, en concentrant les vols sur certaines périodes, augmentent le 11 % obtenu.

**Des séries très différents mais ayant la même statistique.**

Notre bon sens est défaillant pour traiter les probabilités et nous sommes surpris dans des cas où il n'y a pas lieu de l'être. Utiliser le mot « coïncidence » n'explique rien ou alors cela introduit des idées étranges comme la synchronicité de Carl Gustav Jung ou les champs morphiques de Rupert Sheldrake, dont les scientifiques ont vainement cherché à prouver l'existence (voir par exemple www.sceptiques. qc.ca/dictionnaire/).

**Simple et inattendu.**

Toujours afin de comprendre et de classifier les situations où notre esprit s'étonne alors qu'il ne devrait pas, détaillons à présent un cas moins connu, car lié à une théorie assez récente. La mauvaise compréhension de ce qui est probable, car simple, conduit à percevoir certains événements comme étonnants alors qu'ils ne le sont pas, et donc à croire être en présence de coïncidences miraculeuses sans bonne explication, alors que la situation est banale.

La notion la plus générale de simplicité est celle qui provient de la théorie algorithmique de l'information. Comme nous n'en avons pas toujours une bonne compréhension, cela nous conduit à voir des choses complexes et inattendues quand il n'y a en fait que des choses simples. Cette théorie algorithmique de l'information, ou théorie de la complexité de

Kolmogorov, mesure la complexité d'un objet par la taille du plus petit programme qui l'engendre. Elle s'applique aux objets numériques, ou susceptibles d'être représentés numériquement, tels que les images, les sons, les films, et à la plupart des objets du monde réel, si l'on ne prend en compte que leur apparence.

La théorie suggère que parmi les objets utilisant le même nombre de bits d'information (par exemple des images d'un million de pixels), les plus simples, c'est-à-dire ceux ayant la plus faible complexité de Kolmogorov, sont ceux qu'on rencontrera le plus fréquemment.

Une étoile est sphérique, comme le sont beaucoup de fruits. La section d'un tronc d'arbre est circulaire, comme le sont aussi les roues qui équipent nos véhicules. Une tige de blé est parfaitement rectiligne, comme les arêtes d'un cristal. La surface d'un lac est plane comme l'est, regardée de près, la peau de nombreux animaux. Tout cela correspond à des formes simples au sens de la théorie de Kolmogorov. Dans ces cas-là, nous ne cherchons donc pas à trouver une origine commune à deux objets sphériques, ou rectilignes, ou plans. Les formes simples n'ont pas nécessairement une origine commune, leur simplicité suffit à expliquer qu'on les retrouve partout. Jusqu'ici, tout va bien.

En revanche, certains objets ou formes que la théorie de la complexité de Kolmogorov identifie comme simples ne sont pas perçus comme tels par notre jugement immédiat. De nombreuses structures fractales (dont le fameux triangle de Siepinski), nous apparaissent complexes alors qu'elles ne le sont pas puisque des programmes courts les engendrent. Rencontrer ces formes dans la nature (c'est le cas du triangle de Sierpinski qu'on trouve à la surface de certains coquillages) ne doit donc pas nous étonner plus que de rencontrer dans la nature des droites ou des cercles.

Nous ne devons donc pas nous émerveiller de ces rencontres multiples sans lien, ni surtout imaginer qu'elles proviennent d'une sorte de fonctionnement secret de l'Univers qui resterait à comprendre. Comme pour la sphère, leur simplicité est l'explication de leur fréquente apparition.

Il est normal que les objets de faible complexité de Kolmogorov se retrouvent partout. Rechercher une explication profonde à la présence multiple de la suite de Fibonacci dans toutes sortes d'objets naturels ou artificiels est aussi naïf que rechercher une explication commune à la présence de longs segments rectilignes dans les arbres, dans les tracés dessinés par les couches géologiques, les stalagmites ou dans le ciel quand une météorite pénètre l'atmosphère terrestre.

Nous percevons facilement la simplicité de certaines formes, mais pour d'autres, il nous faut réfléchir à la théorie qui permet de comprendre la simplicité. Si nous réussissons, nous serons moins enclins à vouloir expliquer ce que nous percevons comme des coïncidences. Ici, comme pour les doubles tirages identiques au Loto, ou les courbes parallèles montrant des corrélations illusoires, nous devons éviter de rechercher des causes communes à ce qui, logiquement, n'en a pas besoin.

*** Jean-Paul Delahaye,** Professeur émérite à l'université de Lille et chercheur au Centre de recherche en informatique, signal et automatique de Lille (Cristal)