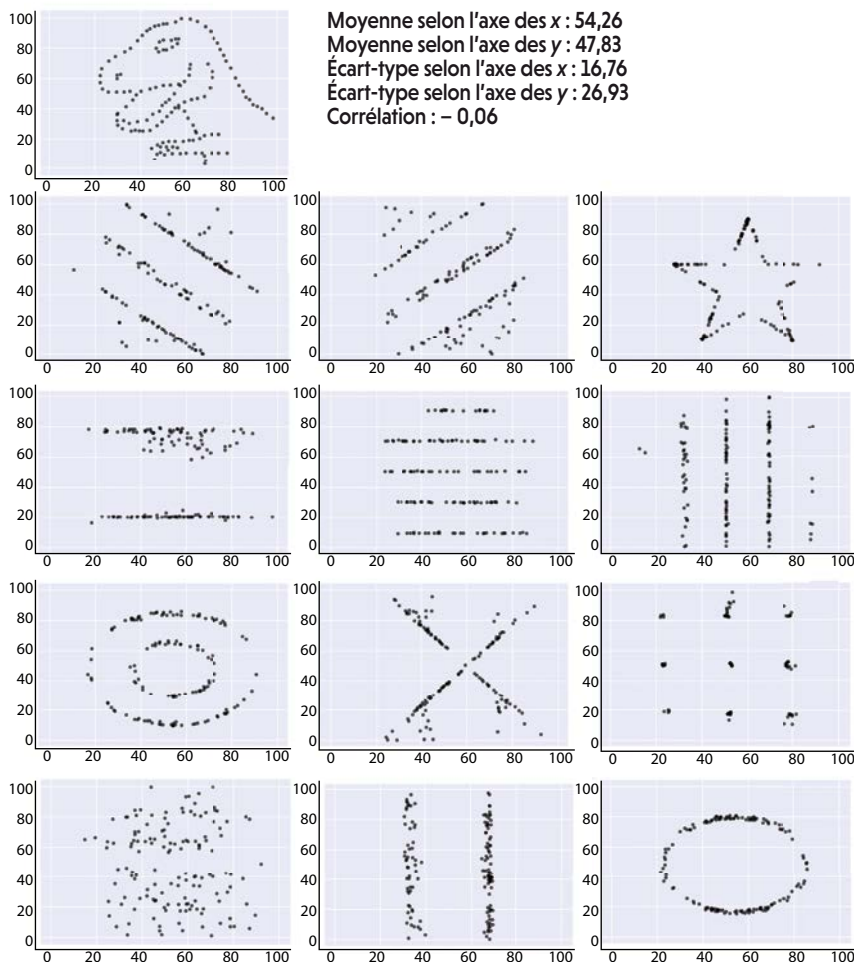


MÊMES PARAMÈTRES... ET POURTANT SI DIFFÉRENTS

4

Nous pensons en général que si des données ont des paramètres statistiques proches (moyennes, écarts-types, etc.), elles se ressemblent nécessairement. Les paramètres statistiques, pensons-nous, sont comme des versions compressées des séries de données. Justin Matejka et George Fitzmaurice prouvent graphiquement que c'est faux. À l'aide d'une méthode qui leur permet de créer facilement des séries ayant exactement les mêmes paramètres statistiques, ils ont notamment produit ces 13 schémas, utiles à tous ceux qui souhaitent prendre conscience de notre fragile compréhension intuitive du hasard et des statistiques.

Il s'agit de graphiques de séries de couples de nombres (x, y) , qui apparaissent très différents. Pourtant, d'une série à l'autre, les 5 paramètres suivants sont les mêmes : moyenne des x , écart-type des x , moyenne des y , écart-type des y , corrélation entre les x et les y . Lorsque la corrélation entre les x et les y est parfaite, ce type de schéma est une droite. (Voir J. Matejka et G. Fitzmaurice, « Same stats, different graphs : Generating datasets with varied appearance and identical statistics », *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1290-1294, ACM, 2017.)



> 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 (voir l'encadré 3 page 111).

DES SÉRIES TRÈS DIFFÉRENTES, 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/).

Cette faiblesse de notre esprit à percevoir correctement les probabilités a encore récemment été mise en évidence. On est tenté de penser que si des données ont des paramètres statistiques proches, elles doivent nécessairement se ressembler; or Justin Matejka et George Fitzmaurice, de la société Autodesk Research à Toronto, ont créé délibérément des exemples magnifiques montrant que c'est faux.

Pour engendrer leurs séries, ils ont mis au point une technique qui leur permet de jouer facilement avec les formes des graphes créés. Ainsi, la figure de l'encadré 4 ci-contre visualise une collection de 13 séries de couples de nombres (x, y) qui, malgré leurs graphes très différents (chaque couple (x, y) représentant un point du plan), ont les mêmes 5 paramètres suivants : moyenne des x , écart-type des x , moyenne des y , écart-type des y , corrélation entre les x et les y .

SIMPLE ET INATTENDU

Toujours afin de comprendre et de classer 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

BIBLIOGRAPHIE

T. Vigen, **Spurious Correlations**, Hachette Books, 2015.

F. Messerli, **Chocolate consumption, cognitive function, and Nobel laureates**, *The New England Journal of Medicine*, vol. 367(16), pp. 1562-1566, 2012.

S. Young et A. Karr, **Deming, data and observational studies**, *Significance*, vol. 8(3), pp. 116-120, 2011.

H. Zenil et J.-P. Delahaye, **On the algorithmic nature of the world**, dans G. Dodig-Crnkovic & M. Burgin (éds.), *Information and Computation*, World Scientific, 2010 (<https://arxiv.org/abs/0906.3554>).

H. White, **A reality check for data snooping**, *Econometrica*, vol. 68(5), pp. 1097-1126, 2000.

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. Cela découle d'un théorème de Leonid Levin d'après lequel plus la complexité de Kolmogorov d'un objet ou d'une structure est élevée, moins il est probable que cet objet ou cette structure soit produite par une machine prise au hasard.

Si l'on pense que le monde physique est une sorte de grand système d'interactions menant une forme de calcul qui produit les objets et structures qu'on y rencontre, il est alors naturel de penser que les objets présents dans le monde y apparaissent avec une probabilité liée à leur mesure de Levin, et donc que les plus simples au sens de la complexité de Kolmogorov sont les plus fréquents. Des détails sur cette idée sont donnés dans un article d'Hector Zenil et moi-même (*voir la bibliographie*). Notons qu'il s'agit d'une idée conforme à l'intuition: dans une image numérique en noir et blanc représentant une photo prise du monde réel, on a plus de chances de rencontrer la série de 10 pixels blancs, *bbbbbbbbbb*, qui a une faible complexité de Kolmogorov, que la série précise *bnbbnbnbbb* dont la complexité de Kolmogorov est supérieure.

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. Une



Certains coquillages arborent à leur surface des motifs très semblables à la figure fractale dénommée triangle de Sierpiński, obtenue en enlevant d'un triangle un triangle central et en répétant indéfiniment l'opération (à droite). Ces motifs sont perçus comme complexes, alors qu'ils sont algorithmiquement simples.

structure comme celle du triangle de Sierpiński (*voir la figure ci-dessus*) nous apparaît complexe alors qu'elle ne l'est pas, puisque des programmes très courts l'engendrent. Si l'on rencontre souvent cette forme dans la nature, en particulier à la surface des coquillages, c'est parce que cette structure est simple, comme le sont une sphère ou un segment de droite.

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.

Ce qui est vrai pour le tapis de Sierpiński l'est aussi pour la suite de Fibonacci, dont certains s'émerveillent à tort qu'on la rencontre si souvent. C'est encore vrai pour le nombre d'or, qui suscite une superstition numérolgique. Cela l'est encore d'une multitude de formes fractales qui sont simples et dont on ne doit donc pas s'étonner de les rencontrer. 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 illusoire, nous devons éviter de rechercher des causes communes à ce qui, logiquement, n'en a pas besoin. ■