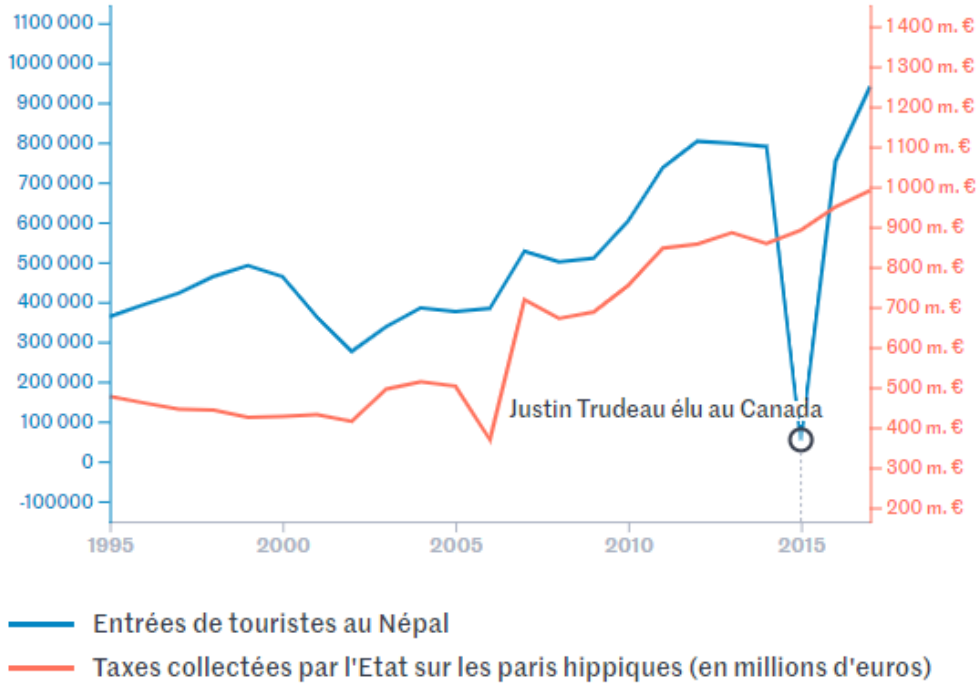
Corrélation ou causalité ?
Brillez en société avec notre générateur aléatoire de comparaisons absurdes

En France, 57 % des morts ont lieu à l'hôpital : la probabilité de mourir dans les établissements de santé est supérieure à celle de passer l'arme à gauche chez soi dans son lit. Alors, dangereux l'hôpital ? Non. Si la proportion de morts est plus élevée à l'hôpital, c'est parce qu'on s'y rend lorsqu'on est malade, et que c'est quand on est malade qu'on risque le plus de mourir.
Cette notion de corrélation, autrement dit quand deux données semblent liées, est tout à fait différente de celle de causalité, le lien de cause à effet. Ainsi, tenter de démontrer une théorie en additionnant des statistiques et en comparant des courbes ou des cartes peut être trompeur si la démonstration n'est pas accompagnée d'une étude rigoureuse.
Le risque ? Tomber dans un déterminisme comme celui de Pierre Simon de Laplace au XVIII^{e} siècle ou celui de certains géographes du XIX^{e} siècle qui défendaient que la géographie – physique, celle des vals et des collines – était responsable de l'ordre de la société. Il en irait de même pour le climat, qui déjà chez l'historien romain Tacite (I^{er} siècle), cité par le géographe Olivier Dollfus, façonnait le comportement des Germains, Tacite évoquant la « rudesse et sauvagerie des peuples venus du Nord, des pays aux hivers froids, qui sortent de la profondeur des forêts ».

De la différence entre corrélation et causalité

Ce graphique présente de manière aléatoire deux indicateurs – à partir d'environ 50 jeux de données historiques – selon que leur représentation se ressemblent, ainsi qu'un événement au hasard. Ses axes des ordonnées sont coupés en bas pour faire correspondre les courbes.



NB : D'autres exemples de corrélations aléatoires sont visualisables sur la version en ligne de l'article.

Dans le sillage des « sept conseils pour ne pas se faire avoir par les représentations graphiques », le graphe ci-dessus pourrait ressembler à un exemple de ce qu'il ne faut pas faire : les deux données n'ont ni la même échelle ni la même unité. En coupant les axes des ordonnées (à droite et à gauche), on peut superposer deux courbes qui n'ont rien à voir et laisser penser qu'elles ont une influence l'une sur l'autre, comme le fait depuis des années le site parodique Spurious Correlations.
Il en va de même pour l'apposition de cartes les unes à côté des autres ; ce n'est pas parce que deux cartes montrent une densité égale à deux indicateurs que ces deux indicateurs ont une influence l'un sur l'autre. Parfois, on se retrouve simplement avec deux cartes de France qui montrent la même chose : il y a plus de blocages, mariages, maraîchage, etc. là où il y a plus d'habitants.

Source : Pierre Breteau, Maxime Ferrer et Lucas Baudin, « Corrélation ou causalité ? Brillez en société avec notre générateur aléatoire de comparaisons absurdes », lemonde.fr, 2 janvier 2019.
https://www.lemonde.fr/les-decodeurs/article/2019/01/02/correlation-ou-causalite-brillez-en-societe-avec-notre-generateur-aleatoire-de-comparaisons-absurdes_5404286_4355770.html

Exploitation pédagogique

1. Qu'est-ce qu'une corrélation ? Quelles sont les deux formes qu'elle peut prendre ?
2. Montrez qu'il existe une corrélation entre le fait d'être à l'hôpital et de mourir. Caractérisez cette corrélation.
3. Peut-on en déduire que l'hôpital est dangereux ? Justifiez votre réponse.
4. Quelle est la variable cachée permettant d'expliquer la corrélation constatée ?
5. À l'aide du générateur de corrélation en ligne, générez aléatoirement une corrélation. Rédigez un court paragraphe détaillant la nature de la corrélation et expliquant pourquoi il n'y a pas de causalité entre les deux variables.
6. D'après l'article, quels sont les précautions à prendre lorsque l'on compare deux variables dans un graphique ?

Corrigé

1. Une corrélation consiste en un lien statistique observé entre deux variables. Une corrélation est dite positive lorsque les deux variables varient dans le même sens ; elle est dite négative lorsque les deux variables varient dans des sens opposés.

2. Il y a une corrélation positive entre le fait d'être à l'hôpital et de mourir. On constate en effet que 57 % des morts ont lieu à l'hôpital : un individu a donc plus de chances de mourir s'il est à l'hôpital que s'il est chez lui.

3. On ne peut pas déduire de cette corrélation que l'hôpital est dangereux. En effet, les hôpitaux sont a priori le lieu où l'on a le plus de chances d'être soignés et guéris.

4. La variable cachée permettant d'expliquer la corrélation est le fait d'être malade. On se rend à l'hôpital lorsqu'on est malade : cet état augmente les chances de mourir. Ce n'est donc pas l'hôpital en lui-même qui augmente la probabilité de mourir mais l'état de santé des patients qui s'y rendent. Il n'y a donc pas de causalité directe entre l'hôpital et la probabilité de mourir.

5. Chaque corrélation est différente. Il faudra déterminer si la corrélation trouvée est positive ou négative, puis montrer ensuite qu'il n'y a pas forcément de lien de causalité entre les deux variables.

6. D'après l'article, il faut éviter de superposer sur un graphique deux variables qui n'ont ni la même échelle ni la même unité. Cela pourrait en effet laisser penser qu'elles ont une influence l'une sur l'autre.