Interprétation, intuition et probabilités

Arthur Charpentier

Professeur d'actuariat à l'Université du Québec, Montréal

Selon le dictionnaire – en l'occurrence Romeuf [1956] –, un actuaire est un « mathématicien spécialisé dans le calcul des probabilités, dont les services sont utilisés, soit par des services financiers, pour des calculs d'amortissement, soit par des établissements dont l'activité comporte un risque calculable ou est basée sur la couverture d'un risque, comme les assurances ». Cela dit, même en étant spécialiste du calcul de probabilités, les raisonnements fallacieux ou les intuitions erronées sont nombreuses. Et peuvent parfois aboutir à des conclusions fausses. C'est précisément le but de la formation des actuaires : il faut non seulement être à l'aise pour faire des calculs, mais aussi développer une intuition robuste quand il s'agit de manipuler des tableaux statistiques, voire des données plus riches encore. Et avec l'explosion du nombre de données à disposition des actuaires, il est à la fois plus facile de se tromper sur les interprétations d'une corrélation, et de plus en plus simple de contrôler les données pour rendre les modèles plus robustes. Les actuaires, et les assureurs en général, se doivent de connaître quelques paradoxes classiques afin d'affiner leurs intuitions et de rendre leurs interprétations plus justes (et moins sujettes à des corrélations fallacieuses (1)).

Le paradoxe de Simpson

En 1951, Edward Simpson avait décrit un phénomène étrange, observé dans des tableaux de contingence [Simpson, 1951], reprenant alors un paradoxe évoqué par Karl Pearson dès 1899 [Pearson, 1899], et surtout par George Udny Yule en 1903 [Yule, 1903]. L'exemple le plus connu est le « biais du genre » dans les admissions aux programmes de cycle supérieur, à Berkeley, évoqué dans Bickel, Hammel et O'Connell [1975]. Mais considérons ici un exemple encore plus simple pour illustrer ce point (voir tableau 1 p. 102).

L'interprétation de ces statistiques est assez simple. Pour le premier tableau, si on doit aller à l'hôpital, mieux vaut aller à l'hôpital B, car il semble plus sûr (avec 90 % de taux de survie, contre 80 % pour l'autre). Maintenant, si on tient compte de l'état de santé :

- si on est en bonne santé, on a intérêt à choisir l'hôpital A, qui a 98 % de taux de survie (97 % pour l'autre) ;

Tableau 1 - Statistiques de décès dans deux hôpitaux

Total				
Hôpital	Admissions	Survivants	Décès	Taux de survie
A	1 000	800	200	80 %
B	1 000	900	100	90 %

Personnes en bonne santé				
Hôpital	Admissions	Survivants	Décès	Taux de survie
A	600	590	10	98 %
B	900	870	30	97 %

Personnes en mauvaise santé				
Hôpital	Admissions	Survivants	Décès	Taux de survie
A	400	210	190	53 %
B	100	30	70	30 %

Source : Exemple fictif. Arthur Charpentier.

- si on est en mauvaise santé, on a intérêt à choisir l'hôpital A, qui a 53 % de taux de survie (30 % pour l'autre hôpital).

Autrement dit, sans cette information relative à la santé, on choisit B, mais avec cette information supplémentaire, on choisit (toujours) A. Comme l'aurait dit Pierre Desproges : « Étonnant, non ? »

Sur l'interprétation mathématique

Pour comprendre ce paradoxe, nul besoin de comprendre les subtilités de la théorie de la mesure ou tout autre concept avancé de probabilité. En fait, la raison peut être expliquée à un enfant de dix ans. En effet, quand on conditionnait par la connaissance de l'état de santé, on préférait A car :

$$\frac{590}{600} \geq \frac{870}{900} \quad \text{et} \quad \frac{210}{400} \geq \frac{30}{100}$$

Et pourtant, quand on regarde globalement, on préfère B car :

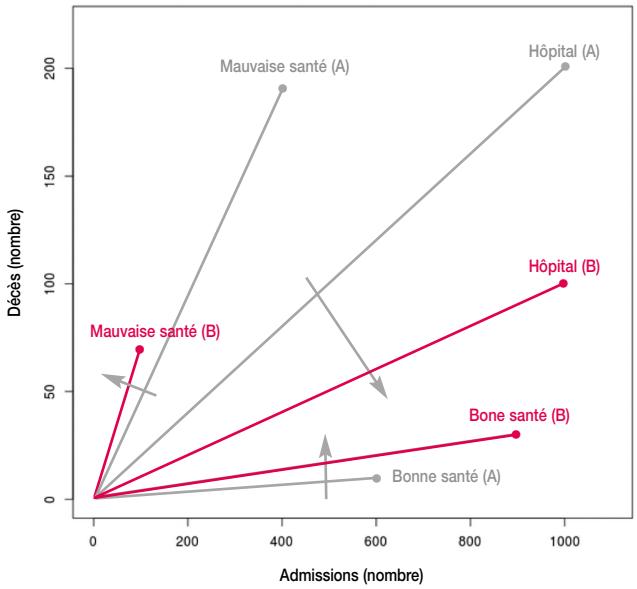
$$\frac{210 + 590}{400 + 600} \leq \frac{30 + 870}{100 + 900}$$

Graphiquement, ces fractions sont les pentes de la figure 1 (voir p. 103). Si on compare les admissions dans les hôpitaux indépendamment de l'état de santé (en haut à droite sur la figure), la pente associée à B est plus faible que celle de A, donc le taux de survie est meilleur avec B qu'avec A. En revanche, si on conditionne par l'état de santé, on va toujours préférer A à B.

L'explication de la possibilité d'un tel paradoxe est assez triviale d'un point de vue mathématique. Mais, comme souvent avec les paradoxes mathématiques, le plus important est l'interprétation et l'explication de l'origine du paradoxe. Dans l'exemple des hôpitaux, la raison vient du fait que le choix de l'hôpital et l'état de santé ne sont pas des variables indépendantes. En effet, l'hôpital B opère ici une sélection et privilégie les personnes en bonne santé. On ne peut alors pas

dire « toutes choses étant égales par ailleurs, si on choisit l'hôpital A… » : les variables explicatives sont ici corrélées, ce qui biaise complètement l'interprétation du modèle.

Figure 1 - Nombre de décès en fonction du nombre d'admissions (les taux de décès sont les pentes)



Source : Arthur Charpentier.

Les implications en théorie de la décision

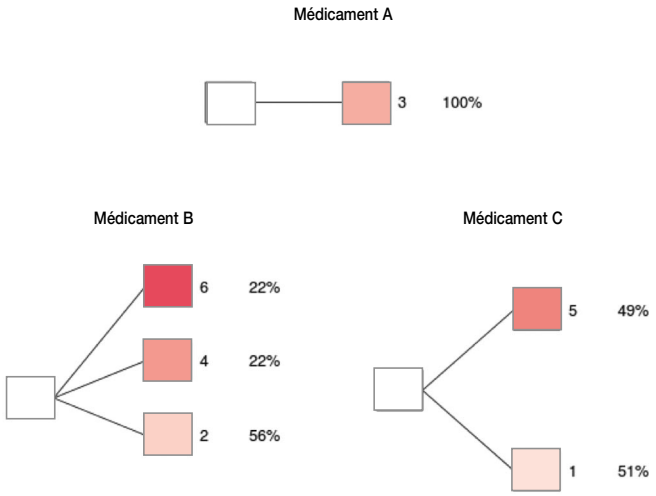
Dans le chapitre "Induction and Probabilities" de [Gardner, 1987], Martin Gardner considère l'exemple suivant, qui concerne l'utilisation d'un médicament.

L'état de santé est mesuré sur une échelle allant de 1 à 6 (de « souffrance atroce » à « parfait état de santé ») :

- le médicament A a un effet constant : en le prenant, les patients ont un état de santé de 3 ;
- le médicament B a un effet aléatoire : dans 56 % des cas, l'état de santé est de 2 et, dans 22 % des cas, l'état de santé est 4 ou 6 ;
- le médicament C a un effet aléatoire : dans 51 % des cas, l'état de santé est de 1, sinon il est de 5.

Les trois médicaments peuvent être visualisés sur la figure 2 (voir ci-dessous).

Figure 2 - État de santé – de 1 (souffrance atroce) à 6 (très bonne santé) – après consommation d'un médicament (A, B et C, de gauche à droite).



Source : Arthur Charpentier

Un médecin se demande quel médicament administrer à un patient. Le critère de choix est simple : un médicament est « meilleur » qu'un autre si la probabilité de se trouver dans un meilleur état de santé est plus grande.

- A sera préféré à B car, dans 56 % des cas, l'état de santé sera meilleur en prenant A.
- A sera préféré à C car, dans 51 % des cas, l'état de santé sera meilleur en prenant A.
- B sera préféré à C (en calculant les probabilités jointes, en supposant l'indépendance entre les états de santé, en prenant B et C) car, dans 61,78 % des cas, l'état de santé sera meilleur en prenant B.

Si on résume, A est toujours le meilleur choix s'il est en vente. En revanche, C est toujours le pire des choix, car les deux autres sont « meilleurs ».

Supposons maintenant qu'on ne demande plus au patient de prendre un médicament précis, mais de garder parmi deux celui qui fait le meilleur effet. Par exemple (toujours en supposant les effets indépendants), en prenant B et C, il y a 28,56 % de chances (0,51 x 0,56) qu'un patient soit dans l'état 2 avec B et dans l'état 1 avec C. Dans ce cas le médecin prescrira B (et son état sera alors 2) ; il y a 27,44 % de chances (0,49 x 0,56) qu'un patient soit dans l'état 2 avec B et dans l'état 5 avec C. Dans ce cas, le médecin prescrira C.

- Dans 28,56 % des cas, l'état de santé sera meilleur avec A qu'avec (B + C).
- Dans 33,22 % des cas, l'état de santé sera meilleur avec B qu'avec (A + C).
- Dans 38,22 % des cas, l'état de santé sera meilleur avec C qu'avec (A + B).

Ici, C sera le médicament préféré, car il a plus d'une chance sur trois de procurer une meilleure santé que les deux autres. En revanche A sera le pire choix.

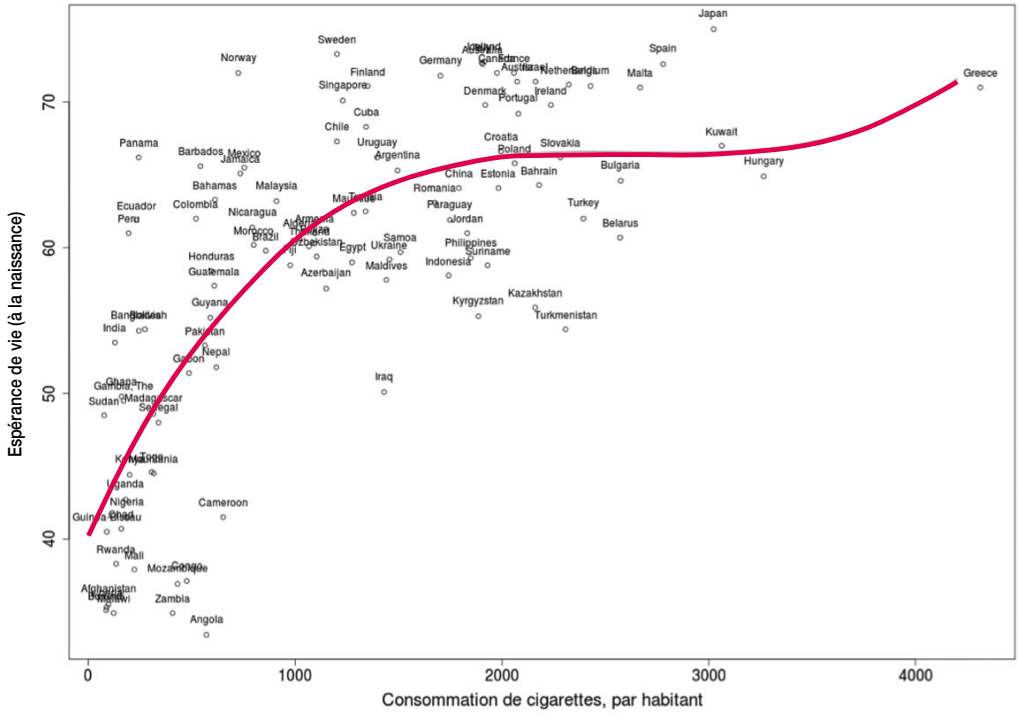
Au-delà du calcul de probabilités (dont les actuaires seraient des spécialistes), on arrive à une situation bien étrange. Si, sur le marché, seuls A et C sont en vente, alors tout médecin raisonnable recommandera à son patient de prendre A. Mais si B est en vente, alors il devrait lui recommander de prendre C.

Oublier des variables explicatives dans un modèle

Ce second exemple est plus compliqué à analyser (mais il est relativement classique dans les problèmes de choix lors d'élections, par exemple). En revanche, le premier est intéressant à plusieurs titres. On a vu que regarder les données agrégées conduisait à des conclusions étonnantes. De manière générale, oublier des variables importantes dans une régression donne des conclusions souvent fausses.

Sur la figure 3 (voir ci-dessous) sont représentées, pour une centaine de pays, la consommation de cigarettes par habitant et l'espérance de vie (à la naissance). On

Figure 3 - Consommation de cigarettes par habitant et espérance de vie



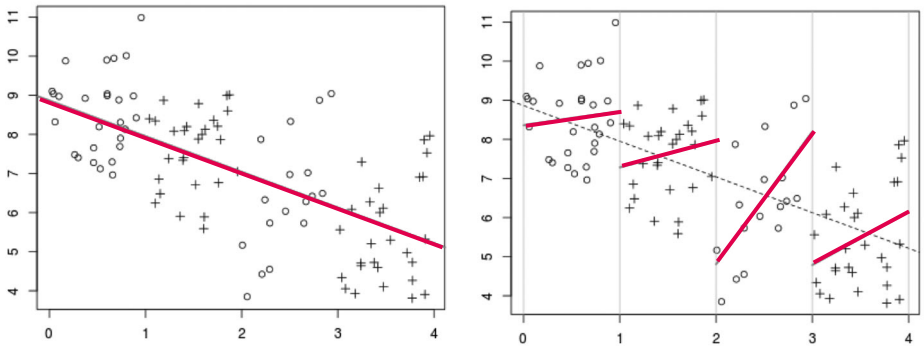
Source : Arthur Charpentier, données OMS.

y voit clairement une relation croissante, que l'on traduirait (hâtivement) par « plus on fume, plus on vit longtemps ». On travaille ici sur des données agrégées, et la lecture plus juste serait de dire que, dans un pays, plus la consommation de cigarettes est importante, plus l'espérance de vie est importante. C'est en fait un effet de richesse que l'on visualise ici (sans qu'il soit représenté explicitement). Cette variable cachée est l'analogue de l'état de santé du patient dans le premier exemple.

Sur la figure 4 (voir plus bas), on a quatre groupes (quatre états de santé pour reprendre les exemples précédents). Si on oublie de les prendre en compte, on observe un effet négatif de la variable explicative (en abscisse) sur la variable que l'on souhaite expliquer (en ordonnée). En revanche, si on raisonne ensuite par sous-groupe, on voit au contraire que l'effet est plutôt positif.

plus sollicités pour rechercher des tendances dans ces masses de chiffres, afin de « séparer un signal du bruit », pour reprendre l'expression de Nate Silver (2). Faire parler les données est finalement assez simple. Mais les faire parler pour extraire un message cohérent et consistant l'est beaucoup moins. On l'a vu, manipuler des tableaux croisés relativement simples peut amener à des conclusions fausses. Car oublier des variables explicatives importantes dans un modèle fait dire aux chiffres le contraire de ce qu'ils disent réellement ! La notion de *data driven* est de plus en plus présente, mais elle est dangereuse. Se laisser conduire par les données peut amener un peu n'importe où. Il est important d'apprendre à écouter ses données, tout en gardant un regard critique à leur sujet.

Figure 4 - Impact de l'oubli de variables explicatives dans un modèle de régression



Source : Arthur Charpentier.

Data science et interprétation statistique

Depuis plusieurs mois, on insiste sur le fait que les actuaires doivent avoir une culture des données, qu'ils doivent être des *data scientists*. Les assureurs collectent en effet beaucoup de données, et les actuaires sont de plus en

Notes

1. Notion de « spurious regression » *en anglais.*

2. Silver N., The Signal and the Noise, *Penguin Group, 2012.*

Bibliographie

BICKEL P. J. ; HAMMEL E. A. ; O'CONNELL J. W., "Sex Bias in Graduate Admissions: Data from Berkeley", *Science*, vol. 187 (4175), 1975, pp. 398-404.

GARDNER M., "Induction and Probabilities", *Time Travel and Other Mathematical Bewilderments*, W. H. Freeman & Company, 1987.

PEARSON K. ; LEE A. ; BRAMLEY-MOORE L., "Genetic (Reproductive) Selection: Inheritance of Fertility in Man", *Philosophical Transactions of the Royal Society*, série A, vol. 192, pp. 257-330, http://rsta.royalsocietypublishing.org/content/192/257.citation#related-urls

ROMEUF J. (dir.), *Dictionnaire des sciences économiques*, tome 1, PUF, 1956.

SIMPSON E. H., "The Interpretation of Interaction in Contingency Tables", *Journal of the Royal Statistical Society*, série B, vol. 13 (2), 1951, pp. 238-241.

YULE G. U., "Notes on the Theory of Association of Attributes in Statistics", *Biometrika*, vol. 2 (2), 1903, pp. 121-134.