

L'ERREUR EST HUMAINE

San Francisco, 22 décembre 1972, une infirmière, nommée Diana Sylvester, est trouvée morte poignardée dans son appartement. La police relève des traces de sperme, mais à cette époque l'analyse génétique n'existait pas. L'enquête de voisinage se dirigea sur un dénommé Robert Baker qui semblait le suspect idéal, mais l'affaire fut classée faute de preuves et le dossier dormit trente ans aux archives des affaires non résolues.

En 2003, la police de San Francisco reçut une bourse destinée à l'usage de nouvelles technologies. L'idée était de rouvrir les anciens dossiers contenant un prélèvement d'ADN et de confronter cet ADN à une grande base de données qui venait d'être constituée et qui rassemblait les profils génétiques de criminels californiens.

On rouvrit le dossier de Diana qui contenait une lame avec un peu de sperme prélevé sur le corps de Diana. Le sperme en question était endommagé et ne faisait apparaître que partiellement le code génétique de l'assassin. Pour les cinq *locus* identifiables, la RMP (*Random match probability*) est estimée à 1 pour 1,1 millions. La probabilité qu'un individu possède les mêmes caractéristiques que les caractéristiques lisibles de l'assassin est estimée à 1/1.100.000.

La base de données des délinquants californiens comportait 338.000 profils et on trouva un individu, du nom de John Puckett, dont l'ADN avait les cinq caractères de l'ADN de l'assassin.

Joli procès en perspective, on cherchait un individu ayant le même ADN que l'assassin, on en trouve un, et c'est quand même plutôt rare d'avoir cet ADN : une chance sur un million. Qu'est-ce que ça nous dit sur la probabilité de la culpabilité de Puckett ? Il n'y a qu'une personne sur un million qui a cet ADN. Je répète comme un procureur.

Le biais du procureur

C'est la lecture du livre de Vincent Berthet qui m'a fait rouvrir le livre des *Maths au Tribunal* où le cas de John Puckett est développé. Vincent Berthet étudie divers processus de raisonnements irrationnels, les biais de raisonnement. Il cite et reprend les travaux de Kahneman et Tversky et il ajoute le *biais du procureur*. C'est une inversion des probabilités conditionnelles : au lieu de se demander quelle est la probabilité que l'accusé soit innocent, compte tenu de son dossier, on se demande quelle est la probabilité qu'il ait ce dossier s'il est innocent. Dans le cas du crime de San-Francisco, c'est d'ailleurs le seul chiffre que nous ayons à notre disposition.

L'analyse de l'ADN de John Puckett, c'est analogue à un test médical : avez-vous l'ADN de l'assassin ? Il y a des faux positifs. Pour continuer l'histoire, il va falloir que je parle de probabilités conditionnelles et de tests biologiques.

Pour tenir le lecteur en haleine, je précise tout de suite qu'il n'y avait pas d'impossibilité technique à ce que Puckett fût l'assassin. Agé de soixante-dix ans en 2003, il avait la trentaine et habitait la région au moment du meurtre de Diana.

Proportions et probabilités conditionnelles¹

Commençons par un exemple. Dans un lycée (petit), il y a trois classes X1, X2 et X3, et il y a aussi une chorale C qui est composée de quelques élèves de chaque classe. Par exemple les élèves de la classe X1 qui participent à la chorale sont au nombre de $N(X1 \& C)$. Dans la chorale qui compte $N(C)$ choristes, les élèves de la classe X1 sont en proportion $P(X1 | C)$. On a donc l'égalité

$$N(X1 \& C) = P(X1 | C) \times N(C).$$

En faisant un raisonnement dans l'autre sens, on a aussi l'égalité

$$N(X1 \& C) = P(C | X1) \times N(X1),$$

où $P(C | X1)$ est la proportion de choristes dans la classe X1.

On en déduit l'égalité suivante, connue sous le nom de **formule de Bayes**² :

$$P(X1 | C) \times N(C) = P(C | X1) \times N(X1).$$

Supposons que je sois intéressé à la proportion $P(X1 | C)$, je peux faire le calcul

$$P(X1 | C) = P(C | X1) \times N(X1) / N(C).$$

Il me suffit de connaître le nombre d'élèves de la classe X1 et la proportion de choristes de cette classe. En l'absence du chef de chœur, pour connaître le nombre $N(C)$ des choristes, il suffit d'additionner ce qui se passe dans chaque classe

$$N(C) = P(C | X1) \times N(X1) + P(C | X2) \times N(X2) + P(C | X3) \times N(X3).$$

Tests biologiques

Dans un cours de Probabilités, on remplace les nombres (d'élèves) par des probabilités et les proportions sont appelées **probabilités conditionnelles**. Lorsque j'enseignais les probabilités, je prenais comme premier exemple d'application de la formule de Bayes, l'interprétation d'un test biologique.

Si l'on veut plaquer l'exemple sur celui de la chorale du lycée, il y a seulement deux classes : les malades M et les sujets sains S ; chanter dans la chorale, c'est avoir un test positif, et n'y pas chanter, c'est un test négatif N. Pour Bayes, sa formule devait permettre d'évaluer les causes, ici la probabilité d'être malade si le test est positif.

$$P(M|P) = P(P|M) \times P(M) / P(P).$$

Les médecins attribuent deux caractéristiques à un test biologique :

- la **sensibilité** qui est la probabilité $P(P|M)$ que le test soit positif pour un malade,
- la **spécificité** qui est la probabilité $P(N|S)$ que le test soit négatif sur un sujet sain.

Une bonne sensibilité évite les faux négatifs pour un malade. La spécificité, c'est pour éviter les faux positifs qui alarment à tort un sujet sain. Les faux positifs sont plus fréquents lorsque la proportion de malades est faible, mais surtout la positivité du test n'est plus significative. En écrivant le dénominateur $P(P)$ comme dans la chorale du lycée,

$$P(P) = P(P|M) \times P(M) + P(P|S) \times P(S),$$

on obtient

¹Le lecteur qui connaît le sujet peut sauter ce paragraphe.

²Thomas BAYES (Londres 1702 - Tunbridge Wells 1761). Bayes est un pasteur de l'Église presbytérienne et habite dans le Kent. Il s'intéresse aux Mathématiques en amateur éclairé au contact de Moivre et il est reçu à la Royal Society en 1742.

$$P(M \vee P) = 1 - \frac{P(P|S)P(S)}{P(P)}.$$

Tout ce qui augmente la fraction fait baisser la fiabilité d'un résultat positif :

- $P(S)$ c'est la proportion de sujets sains : quand il y a peu de malades, le test donne trop de faux positifs.
- $P(P)$ de même est d'autant plus petit qu'il y a moins de malades.
- $P(P|S)$ c'est 1 moins la spécificité $P(N|S)$.

On peut aussi voir cela sur un dessin. Les tests positifs sont en couleur : rouge pour les malades, verts pour les sujets sains. Les faux positifs sont une faible proportion pour les sujets sains, mais ils peuvent être une forte proportion parmi les tests positifs si les sujets sains sont très majoritaires.

sains négatifs		malades négatifs
		malades positifs
sains positifs		

Lorsque j'enseignais dans les années 1980, ce genre de calculs expliquait pourquoi les autorités médicales ne recommandait pas une campagne générale de tests du SIDA alors que certains politiques la souhaitaient. On a restreint les tests à une population réputée à risques.

En 2020, les tests qui ont été déployés dès le mois de mars avaient une sensibilité parfois médiocre (de 70% à 95%) tandis que les spécificités étaient toujours supérieures à 95%. Il y avait alors très peu de malades du covid.

Voici un petit calcul pour un test de sensibilité 90%, de spécificité 97%, et une proportion de 1 malade pour 1000 sujets testés³

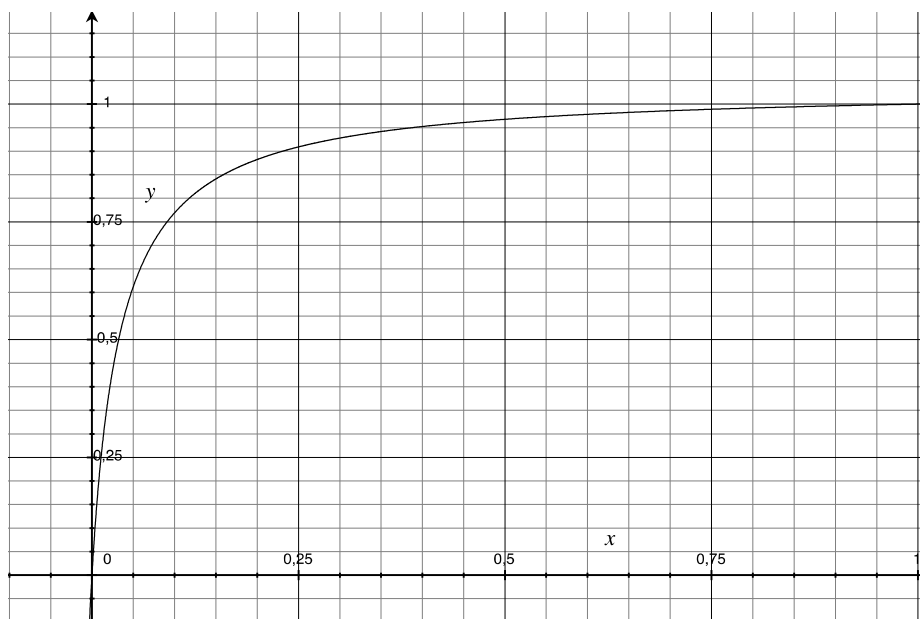
$$\begin{aligned} P(P) &= P(P|M)P(M) + P(P|S)P(S) \\ &= 0,9 \times 0,001 + 0,03 \times 0,999 = 0,031 \end{aligned}$$

$$P(M|P) = P(P|M) \times P(M) / P(P) = 0,9 \times 0,001 / 0,031 = 2,9\%$$

Voici la variation de $y=P(M|P)$ en fonction de $x=P(M)$ avec les mêmes paramètres de sensibilité et de spécificité :

$$y = \frac{(0,9 \cdot x)}{0,87 \cdot x + 0,03}$$

³Le nombre de 66.000 cas recensés en France est atteint (et dépassé) le 3 avril 2020.



Vous voyez le désastre ! Il n'y a pas plus de 3% de malades parmi les tests positifs. Et vous voyez pourquoi il ne faut pas se laisser tromper par les 90% de sensibilité $P(P|M)$.

C'est pourquoi, au printemps dernier, on recommandait un test uniquement aux personnes présentant des symptômes de l'infection. L'objectif était alors d'hospitaliser le plus tôt possible les personnes atteintes sans attendre un développement dramatique de la maladie.

En décembre 2020, la Région Rhône-Alpes-Auvergne a financé un dépistage massif dont l'objectif ni les résultats n'ont été clairement exprimés⁴. En janvier 2021, un dépistage massif à Saint-Etienne est piloté par une équipe hospitalière. Là, l'objectif avoué est clair, l'équipe le dit à la télé, elle veut faire une publication.

Les anniversaires

Revenons à San Francisco où une avocate a été désignée pour la défense de John Puckett. Bicka Barlow avait travaillé dans la médecine légale et elle était à présent membre du service de la défense publique de la ville de San Francisco. Elle avait depuis longtemps des doutes sur la méthode des coïncidences dans les bases de données et elle connaissait les travaux de Kathrine Trayer qui avait fait une gigantesque étude statistique sur une base de données génétiques de l'état de l'Arizona. La question était de distinguer entre « probabilité d'avoir un ADN donné » et « probabilité que deux individus aient le même ADN ».

C'est un petit jeu que l'on peut pratiquer en classe, par exemple dans une classe de 25 élèves.

1. Quelle est la probabilité que l'un des élèves ait son anniversaire le 23 janvier ?

Réponse : la proba qu'un élève donné n'ait pas son anniversaire le 23 janvier est $364/365$.

⁴La campagne est lancée avant Noël pour « protéger nos proches ». Bilan : 8 millions d'habitants, 626.000 tests et 29.000 tests positifs.

La probabilité qu'aucun élève de la classe n'ait son anniversaire le 23 janvier est $(364/365)^{25} = 0,9337$.

La probabilité qu'au moins un élève ait son anniversaire le 23 janvier est $1 - 0,9337 = 0,0663$ c'est-à-dire 6,63%.

2. Quelle est la probabilité que deux élèves aient leur anniversaire le même jour ?

Réponse : on compte d'abord le nombre de paires d'élèves $25 \times 24 / 2 = 300$.

Ensuite le nombre de paires de dates de l'année est $365 \times 366 / 2 = 66\,995$, parmi lesquelles 365 sont doubles, soit une proportion $p = 365/66\,995 = 0,00546$ de doubles.

La probabilité q d'éviter ces doubles est $q = 1 - p = 0,99454$.

La probabilité qu'il n'y ait pas de coïncidence d'anniversaires est $q^{300} = 0,193$.

Il y a donc une probabilité de 80% que deux élèves fêtent leur anniversaire le même jour.

Ce dernier raisonnement est un peu rapide car chaque élève appartient à deux paires et les paires de dates ne sont pas indépendantes. Mais il donne l'idée de la grande différence des deux probabilités.

Après cette instructive digression, revenons à San Francisco, l'avocate essaya de parler de coïncidence pour faire réviser le chiffre de 1 sur 1 million, mais le procureur la remit sur le droit chemin. La question, c'est bien une chance sur 1 million qu'une personne ait l'ADN de l'assassin et non le fait que deux individus aient le même ADN.

La base de données

L'accusation interprète le rapport de 1 sur 1 million comme la probabilité que Puckett **ne soit pas** celui dont provient l'échantillon. Ça ne lui fait pas un bon dossier : 999.999 chances sur un million qu'il soit coupable.

L'avocate Bicka Barlow, qui n'aime pas les bases de données, entreprend une autre démarche. L'ADN de Puckett figurait dans une base de données de 338.000 individus. Notre proportion 1 sur 1 million fait qu'il y a une chance sur 3 de trouver le bon ADN dans une base de 338.000. C'est ainsi que Puckett a été trouvé. Selon Bicka Barlow, cette extraction de Puckett de la base de données n'a rien à voir avec le meurtre de Diana ; il y avait statistiquement 1 chance sur 3 que l'on trouve quelqu'un avec l'ADN de l'assassin dans une base de 338.000 individus, comme dans n'importe quelle base de données de 338.000 individus.

Avant le début du procès, le juge décida qu'il serait trop déroutant de présenter devant la cour les différents arguments mathématiques qui, d'ailleurs, semblaient se contredire. Le jury a condamné Pluckett à la prison à vie. Il faut dire que Pluckett n'était pas par hasard dans la base de donnée de criminels de Californie : il avait à son actif plusieurs viols, commis dans des conditions voisines du viol (ou plutôt de la fellation contrainte) de Diana Sylvester. C'est une condamnation à la française sur des présomptions et, en plus, trente ans après les faits supposés.

Et après

Il y eut ensuite des experts juristes pour s'interroger sur le 1/3 et le 1/1 million, qui ne sont pas les probabilités de culpabilité ou d'innocence de Puckett. Les auteures des *Maths au tribunal* essayent de conclure.

En comptant 1 chance sur 1 million pour l'ADN et 300 millions d'américains, on arrive à 1 sur 300. Puis il y a le fait d'être adulte et en Californie en 1972. On arrive à 1 chance sur 70, mais de quoi ? Qu'un américain pris au hasard cumule toutes ces conditions.

Ensuite, les auteures cherchent la probabilité qu'une autre personne ait le même profil. Elles appliquent la formule de Bayes aux événements

A = deux personnes ont le même profil,

B = une personne au moins a ce profil.

Et elles ont la surprise de trouver $P(A)=1/70$. On pouvait s'en douter : c'est la probabilité qu'une personne ait le profil si l'on exclut Puckett.

S'il y a une seconde personne, ce peut être l'assassin.

On tourne en rond car on n'a que des probabilités d'avoir un profil donné et aucun lien avec l'innocence et la culpabilité.

Les tests ADN aujourd'hui

Ils sont devenus une routine pour les enquêteurs comme les empreintes digitales l'étaient devenues il y a un siècle. Le profil ADN, s'il est complet et non partiel comme à San Francisco en 1972, permet une identification certaine de l'individu.

BIBLIOGRAPHIE

SCHNEPS (Leila) & COLMEZ (Coralie), *Les Maths au Tribunal*, Seuil, 2015, titre original : *Math on Trial. How Numbers get Used and Abused in the Courtroom*, Basic books, 2013

BERTHET (Vincent), *L'erreur est humaine*, CNRS Éditions, 2018

KAHNEMAN (Daniel), *Système 1, système 2*, Flammarion, 2016.

Mes calculs

Le 23 janvier, j'avais moi-même tourné en rond avec la probabilité de 1 sur un million et en faisant des calculs. Comment calculer qu'une personne ait le profil de fréquence $1/N$ (où N est un grand nombre) dans une population de n personnes ?

La probabilité qu'aucune personne n'ait le profil est égale à

$$\left(1 - \frac{1}{N}\right)^n.$$

Lorsque N est grand, 1/e est une bonne approximation de $\left(1 - \frac{1}{N}\right)^N$, ensuite on élève à la puissance n/N et on passe au complémentaire.

Pour N = 1 million, on trouve une probabilité de 63% qu'il y ait au moins un individu dans une population d'un million de personnes. Mais toujours aucun lien avec culpabilité ou innocence.

Ensuite, pour les anniversaires, j'ai failli reculer ou rechercher ma vieille table de logarithmes avant de constater que mon téléphone calculait sans se plaindre 0,99454 à la puissance 300.

ooo-ooo