Des probabilités et des cotes

Philippe Marchal

Quel niveau de difficulté conceptuelle représente le fait de distinguer p et $p/(1-p)$? Lorsqu'on est habitué à l'usage des mathématiques, on pourrait penser que toute personne ayant suivi un cursus scientifique au lycée, et *a fortiori* toute personne ayant suivi des études scientifiques devrait n'avoir absolument *aucune* difficulté à ce propos. Il n'en est rien, nous l'allons montrer tout à l'heure.

Commençons par citer un chercheur en médecine à propos de cette subtile distinction : "C'est tellement complexe que malheureusement (...) Il y a aussi un usage pour nous d'utiliser certaines métriques, certaines mesures en fait par simplification. C'est voulu, on sait que ça distord un peu mais c'est l'usage consacré. On pourrait refaire avec les bonnes choses. Ça complexifierait beaucoup la méthode (...)" [1]. Ce chercheur a au moins le mérite de la franchise : pour lui, distinguer p et $p/(1-p)$ relève d'un niveau de complexité qui le dépasse.

Des chances et des paris

On ne va pas reprendre ici toute la théorie des probabilités mais rappeler brièvement ceci :

– la probabilité qu'un événement se produise est un nombre p compris entre 0 et 1,

– à une probabilité p on peut associer une cote ("odds" en anglais) qui est donnée par $p/(1-p)$. [2]

Le terme "cote" renvoie au vocabulaire des paris. Supposons que vous misiez sur la victoire d'un cheval à c contre 1. Cela veut dire que

– si le cheval perd, vous perdez votre mise,

– s'il gagne, on vous rembourse votre mise et en sus, on vous donne c fois votre mise.

Supposons qu'on connaisse la probabilité p que le cheval perde. À quelle condition le jeu sera-t-il équitable ? L'espérance de gain se calcule simplement comme suit : avec probabilité p vous perdez votre mise, qu'on note m ; avec probabilité $1-p$ vous gagnez votre mise multipliée par la cote, soit cm. L'espérance de gain est alors

$$E = -pm + (1-p)cm$$

Le jeu est équitable si l'espérance de gain est nulle, autrement dit si $-pm + (1-p)cm = 0$, soit

$$c = p/(1-p)$$

La cote représente donc le facteur multiplicatif de gain pour un pari équitable. Bien sûr, les maisons de jeu n'ont pas intérêt à proposer de paris équitables : leur but est de gagner de l'argent mais ceci est une autre question.

1. Voir à 1h10 sur la vidéo `https://www.youtube.com/watch?v=xm5GvREYQMY`
2. Si $p = 1$, la cote est infinie ; il s'agit d'un cas limite sur lequel on ne va pas s'attarder ici.

Si maintenant on a deux probabilités p et q, le rapport de ces probabilités, aussi appelé risque relatif, est p/q ; le rapport des cotes, "odds ratio" en anglais, est

$$\frac{\left(\frac{p}{1-p}\right)}{\left(\frac{q}{1-q}\right)}$$

Naturellement, si $p \neq q$, le risque relatif diffère du rapport des cotes :

$$\frac{p}{q} \neq \frac{\left(\frac{p}{1-p}\right)}{\left(\frac{q}{1-q}\right)}$$

Des risques et des cotes

L'article "odds ratio" de wikipedia anglophone nous apprend que selon une étude datée de 2001, environ 1/4 des articles publiés dans certaines revues de médecine confondaient les notions de risque relatif et de rapport des cotes. Un contributeur de wikipedia qualifie ceux qui commettent cette erreur du terme peu élogieux de "uncomprehending authors", qu'on pourrait traduire un peu familièrement par "auteurs malcomprenants".

Plus récemment, on retrouvait cette erreur dans un article scientifique censé évaluer la surmortalité due à l'usage de l'hydroxychloroquine pour traiter la covid-19 [3]. Interrogé sur les critiques contre cette étude, un des auteurs répondait par les propos cités plus haut. Cet article, coécrit par six chercheurs, publié après relecture minutieuse (ou pas) par d'autres chercheurs, a été abondamment relayé par de nombreux scientifiques (dont au moins un, une devrais-je dire, siège à l'Académie des sciences) sans que cette erreur de niveau lycée soit relevée. Il aura fallu attendre quelques semaines pour qu'elle soit enfin exposée publiquement, grâce au mathématicien Vincent Pavan [4].

La surmortalité due à l'hydroxychloroquine est calculée dans l'article par la formule

$$H \times T \times M \times (Odd - 1)$$

où H est le nombre de patients hospitalisés, T est le taux de patients traités avec l'hydroxychloroquine, M est le taux de mortalité sans usage de l'hydroxychloroquine et Odd est le rapport des cotes. Cette formule est évidemment fausse [5]. De fait cette surmortalité est donnée par

$$H \times T \times (M' - M)$$

où M' est le taux de mortalité quand on utilise l'hydroxychloroquine. Les quantités M et M' sont des probabilités et le rapport des cotes correspondant est

$$Odd = \frac{\left(\frac{M'}{1-M'}\right)}{\left(\frac{M}{1-M}\right)}$$

3. "Deaths induced by compassionate use of hydroxychloroquine during the first COVID-19 wave : an estimate", *Biomedicine and Pharmacotherapy*, 2024.

4. J'avais signalé l'erreur quelques jours auparavant sur le forum (semi-public) de mon laboratoire, avec des commentaires peu amènes que la décence m'interdit de reproduire ici.

5. "Demander à un chercheur en médecine une formule mathématique correcte, c'est comme demander à un sumotori de faire du saut à la perche", proverbe japonais

Des manipulations algébriques élémentaires permettent d'obtenir la formule correcte pour la surmortalité :

$$H \times T \times M \times (Odd - 1) \times \frac{1 - M}{1 + M(Odd - 1)}$$

Le calcul serait sans doute considéré comme plutôt difficile pour le brevet des collèges mais plutôt facile au niveau baccalauréat pour un élève ayant suivi l'option mathématiques. Le fait qu'un professeur en médecine considère comme "tellement complexe" la multiplication par le facteur

$$\frac{1 - M}{1 + M(Odd - 1)}$$

pourra susciter, au choix, un profond embarras ou un rire homérique. En tout cas, avec des valeurs de M qui peuvent être de l'ordre de 0,2, on voit que négliger ce facteur est tout à fait incorrect.

Plus généralement, que la distinction entre ces deux notions de risque relatif et de rapport des cotes soit considérée comme aussi difficile par de nombreux chercheurs n'étonnera que les personnes superstitieuses persuadées que de longues études couronnées par un doctorat garantissent des capacités intellectuelles supérieures à celles d'un bon lycéen.

Des contrôles d'identité

La pratique des contrôles d'identité par la police est un sujet de polémique récurrent et sans surprise, la confusion mentionnée plus haut s'y retrouve. Une étude très médiatisée à sa sortie et encore abondamment citée de nos jours est celle de Jobard et Lévy. Dans un premier rapport [6], les auteurs y expliquaient : "l'odds-ratio est le ratio de 2 probabilités (...)". Autrement dit, sans même parler de compétence mathématique, on constate que les auteurs se montrent incapables de recopier une définition.

Dans un texte ultérieur [7], les deux sociologues prétendent effectuer une régression logistique et calculer des rapports des cotes. Or leurs données ne permettent pas de faire ces calculs. N'importe qui ayant un minimum de compréhension des mathématiques devrait le voir facilement mais bien sûr, cela a complètement échappé à toute une communauté universitaire fort peu à l'aise avec l'algèbre de niveau lycée : l'article a été publié dans une des revues les plus prestigieuses du domaine et n'a jamais suscité la moindre critique méthodologique.

Pour résumer, les auteurs ont procédé en deux étapes :

– Une étape d'étalonnage ("benchmarking" en anglais) évaluant la composition démographique (sexe, âge, origine ethnique...) de la population présente en certains lieux par échantillonnage des flux entrants [8].

– Une étape relevant la composition démographique de tous les individus contrôlés par la police en ces lieux à certaines périodes.

La comparaison des proportions de personnes présentes et de personnes contrôlées permet de calculer le risque relatif, *mais pas* le rapport des cotes. Pour prendre un exemple, supposons que les femmes représentent la moitié des personnes présentes mais un dixième des personnes contrôlées. On en déduit que si N est le nombre de personnes présentes et si k femmes ont

6. "Police et minorités visibles : les contrôles d'identité à Paris", Open Society Justice Initiative

7. "Mesurer les discriminations selon l'apparence : une analyse des contrôles d'identité à Paris", *Population*, 2012

8. Les auteurs oublient au passage de mentionner une hypothèse cruciale : les flux sont proportionnels aux stocks. Toute personne ayant eu l'occasion dans sa vie de comparer la longueur de la queue aux toilettes hommes/femmes appréciera à sa juste valeur cet oubli.

été contrôlées, le risque d'être contrôlé pour une femme est $k/(N/2)$. Le nombre d'hommes contrôlés est alors $9k$ puisque les femmes représentent une personne contrôlée sur 10. Le risque d'être contrôlé pour un homme est donc $9k/(N/2)$ et le risque relatif est donné par

$$\frac{k/(N/2)}{9k/(N/2)} = \frac{1}{9}$$

On remarque que si on ne connaît pas N, on ne peut calculer aucun des deux risques d'être contrôlé (pour les femmes et les hommes) mais on peut calculer le rapport de ces risques puisque la quantité N disparaît dans l'expression du risque relatif.

En revanche, calculer le rapport des cotes nécessite de connaître aussi le rapport des probabilités de ne pas être contrôlé. Ce rapport est

$$\frac{1 - [k/(N/2)]}{1 - [9k/(N/2)]}$$

et ici, aucune simplification ne permet de terminer le calcul si on n'a pas la valeur de N. L'article repose donc sur une fabrication puisqu'il prétend calculer des quantités qu'on ne peut déterminer en l'absence d'une donnée cruciale. Il est vraisemblable que les auteurs ont fait comme si la population totale était égale au nombre d'individus relevés lors de la phase d'étalonnage. Il est facile de voir que cela sous-estime les probabilités de ne pas être contrôlé, surestimant par là les rapports des cotes.

On notera au passage qu'au sommet même du système universitaire français, à savoir au Collège de France, un professeur propriétaire de trois vaisseaux dessus la mer jolie (Héri Héran, Ranpataplan) [9] confond dans un de ses cours la population totale lors de la deuxième étape (la quantité N) avec la population totale échantillonnée lors de la première étape [10]. Rappelons qu'un professeur au Collège de France n'a qu'une vingtaine d'heures de cours à préparer chaque année. Errare humanum est, perseverare diabolicum : Ranpataplan a réitéré son erreur lors d'une conférence quelques mois plus tard [11]. Le pire est qu'en 2010, notre sommité universitaire, présidant un comité Théodule répondant au doux nom de Comedd, rédigeait un des ces innombrables rapports que personne ne lit mais dans lequel il était déjà question des travaux de Jobard–Lévy [12]. Au bout de dix ans, Ranpataplan n'aura donc toujours pas compris le protocole de l'étude, pourtant simplement explicable en quelques lignes. Ce faisant, l'impossibilité de calculer les rapports des cotes avec ce protocole lui sera passé largement par-dessus la tête.

Une autre étude illustre de manière encore plus caricaturale la confusion mentale autour des notions de risque relatif et de rapport des cotes. Un rapport du Défenseur des droits [13] relevait que 80% des jeunes hommes noirs ou arabes avaient fait l'objet d'un contrôle d'identité au cours des 5 années passées, contre 16% pour l'ensemble de la population. Les auteurs en concluaient : "ces profils ont ainsi une probabilité 20 fois plus élevée que les autres d'être contrôlés". Le lecteur attentif aura facilement identifié une nouvelle confusion entre le risque relatif et le rapport des cotes : on pourra calculer ces deux quantités pour $p = 0,8$ et $q = 0,16$.

À l'époque, cette grossière erreur avait été reprise en boucle par la presse - une recherche sur internet permettra de trouver des dizaines d'articles reprenant l'affirmation du Défenseur

9. `https://www.youtube.com/watch?v=uuyOfi7lKUk`
10. Voir à 1h34 sur la vidéo `https://www.youtube.com/watch?v=W01l6gigGFU`
11. Voir à 1h05 sur la vidéo `https://www.youtube.com/watch?v=Q55ljlPhiRI`
12. `https://www.vie-publique.fr/files/rapport/pdf/104000077.pdf`
13. "Relations police / population : le cas des contrôles d'identité", 2017

des droits sans la moindre forme d'esprit critique. Pis encore, les émeutes suite à la mort du prénommé Nahel ont donné l'occasion à de nombreux commentateurs de ressortir ce chiffre fantaisiste, toujours sans se poser la moindre question. À une époque où les journalistes se targuent de faire de la vérification factuelle, le fait qu'aucun grand média n'ait été capable de vérifier le résultat de la divison 80/16 est assez révélateur. Il y a un siècle, un élève incapable d'effectuer cette division - sans calculatrice naturellement - n'aurait jamais eu son certificat d'étude.

Des risques de contamination à la covid-19

Une série d'études sur la covid, baptisées comcor et abondamment relayées pendant la période pandémique, a été menée par un légionnaire d'honneur et son équipe de l'Institut Pasteur. Le premier volet de cette étude avait fait l'objet d'une prépublication [14] où il était écrit à propos des rapports des cotes, appelés odds ratios (OR) en bon franglais par les auteurs : "Les ORs représentent le ratio des risques d'être infecté par le SARS-CoV-2 entre exposés et non exposés pendant la période considérée".

Là encore, des chercheurs à bac+40 se montrent incapables de recopier une définition. Le plus remarquable est qu'à l'époque, le légionnaire en question, âgé de 59 ans, avait déjà derrière lui une carrière d'épidémiologiste de plusieurs décennies au cours de laquelle il avait été constamment confronté à la notion de rapport des cotes qu'il comprenait manifestement de travers. Cela ne l'a nullement empêché d'être nommé au Conseil "scientifique" ni d'être adoubé chevalier.

Le lecteur un peu naïf pourra se demander comment cela est possible. Il faut savoir que l'usage des mathématiques par le chercheur non mathématicien et, disons, non physicien et non informaticien, se borne souvent à entrer des données dans un logiciel de statistiques et à interpréter, parfois de manière incorrecte, ce que le logiciel renvoie. Or dans la méthode de régression logistique couramment employée par de nombreux universitaires, le logiciel calcule un rapport des cotes et non un risque relatif.

En 2019, peu avant la pandémie, un article dans la revue *Nature* [15] mettait en garde contre une mauvaise utilisation des statistiques dans la recherche. Coécrit par trois personnes, il était cosigné par 800 chercheurs. Il émettait des recommandations en affirmant que si celles-ci étaient suivies, "people will spend less time with statistical software, and more time thinking" ("les gens passeront moins de temps sur leur logiciel de statistiques et plus de temps à réfléchir"). La question abordée était celle du seuil de significativité statistiques et non la confusion entre risque relatif et rapport des cotes. Mais le principe de réfléchir plus et de moins utiliser son logiciel de statistiques vaut de manière générale. Néanmoins avec la pandémie, non seulement ces recommandations n'ont pas été suivies mais au contraire, la ruée sur les données sans la réflexion nécessaire sur les bonnes méthodes mathématiques pour les utiliser a été la norme.

Le période du ridicule dans cet usage du logiciel de statistiques sans une once de réflexion est probablement atteint par un autre article sur les risques de contamination à la covid-19 [16]. Les auteurs s'y proposent de comparer les risques de contamination selon différents critères. Par exemple, le nombre d'hommes infectés (resp., non infectés) est 159 (resp. 462), ce qui

14. "Etude des facteurs sociodémographiques, comportements et pratiques associés à l'infection par le SARS-CoV-2 (ComCor)", 2020

15. `https://www.nature.com/articles/d41586-019-00857-9`

16. "Risk of SARS-CoV-2 Acquisition in Health Care Workers According to Cumulative Patient Exposure and Preferred Mask Type", *JAMA Network Open*, 2022

permet de calculer la probabilté d'infection

$$p = \frac{152}{152 + 469}$$

et la cote associée

$$\frac{\frac{152}{152+469}}{\frac{469}{152+469}} = \frac{152}{469}$$

pour un homme. On peut faire le même calcul pour les femmes et on trouve un rapport des cotes égal à 0,96. Cependant dans l'unique tableau de l'article, la première ligne nous apprend que l'âge médian des personnes contaminées (resp., non contaminées) est de 40,6 ans (resp. 43,2 ans). On peut évidemment effectuer le même calcul que plus haut

$$\frac{40,6}{43,2}$$

mais la quantité calculée ne s'interprète pas comme une cote ; en fait elle n'a absolument aucun sens statistique. Cela n'a pas empêché les auteurs de faire le calcul et de remplir la case correspondante. L'*âge du capitaine* était conçu à l'origine par Flaubert comme une plaisanterie, il est devenu une prophétie. Et un bonheur ne venant jamais seul, les auteurs ont fait le même calcul sans queue ni tête avec l'indice de masse corporelle.

Là encore, cet article a été abondamment relayé. Un personnage très suivi sur les réseaux sociaux a ainsi publié un message sur twitter aimé plus de 6000 fois et reposté plus de 2000 fois [17], notamment par cette académicienne déjà évoquée et que les esprits facétieux appellent madame Charlson [18].

Splendeur de la Science, misère des "scientifiques"

On pourrait bien sûr multiplier les exemples. Ceux qui sont passés en revue dans ce texte ont eu un écho particulier dans la presse ou sur les réseaux sociaux. Mais l'écosystème universitaire et médiatique qui a permis à ces erreurs grossières de se diffuser aussi facilement est globalement assez incompétent pour en avoir laissé passer bien d'autres. Par ailleurs, le présent texte ne mentionne que des erreurs concernant les rapports des cotes. Les autres occasions d'erreurs mathématiques dans le monde de la recherche sont innombrables. Tenter de les relever toutes permettrait de toucher du doigt cette notion topologique si pittoresque qu'on appelle *espace totalement inépuisable* [19].

17. `https://twitter.com/EricTopol/status/1559198721123921920`

18. Elle avait affirmé publiquement à trois reprises que la proportion d'hospitalisés de moins de 50 ans sans comorbidités lors de la première vague de covid était supérieure à 40%, interprétant de manière fautive l'indice de Charlson : voir `https://twitter.com/DgCostagliola/status/1480088447096401922`, ainsi que ses passages sur France Inter et France Culture autour de cette période. C'est un peu comme si un supposé spécialiste d'aviation affirmait qu'un airbus vole à 90 km/h.

19. Un espace topologique est inépuisable s'il n'est pas maigre relativement à lui-même. Il est totalement inépuisable si tout sous-espace fermé non vide est inépuisable. Bourbaki, *Topologie générale*