DES PETITS EFFETS PASSÉS AU CRIBLE

Il est difficile de mettre en évidence un petit effet, car les données n'apportent toujours qu'une information limitée. Plus l'effet que l'on cherche à estimer est petit, plus la quantité de données nécessaire pour le mettre en évidence est grande pour qu'il ressorte des fluctuations dues au hasard. N'y a-t-il pas d'autre issue qu'une simple augmentation de la quantité de données, coûteuse et parfois impossible ? Autrement dit : « Comment exploiter au mieux un ensemble de données ? » Donnons quelques exemples. Dans un essai thérapeutique concernant une nouvelle préparation que l'on souhaite comparer à une ancienne, les patients sont souvent très hétérogènes. On constitue des paires de patients aussi proches que possible pour toutes les caractéristiques susceptibles d'influer sur le résultat (sexe, âge, catégorie socioprofessionnelle, origine ethnique...). Pour éviter tout biais, on choisit au hasard dans chaque paire le patient qui reçoit le traitement à tester et celui qui reçoit l'ancien. La comparaison des deux traitements se fonde ainsi sur un ensemble de comparaisons élémentaires beaucoup plus efficaces que si l'on avait choisi les patients sans tenir compte de leur statut. Dans les sondages d'opinion, on sait que l'âge, la catégorie socioprofessionnelle influencent le résultat. On répartit la population à sonder en différents groupes plus homogènes que l'on échantillonnera séparément (échantillonnage stratifié). Si l'effectif des groupes est déjà connu, on montre que l'on peut ainsi améliorer la précision par rapport à un échantillonnage aléatoire simple de même taille. On dispose en plus d'une information beaucoup plus riche sur la répartition des opinions. En outre, les statisticiens ont développé la théorie des plans d'expérience pour améliorer la collecte des données. Il s'agit, dans les exemples cités, de concevoir, par exemple, l'allocation des mesures ou le choix des unités expérimentales, de façon à ce que pour un coût donné la précision de l'analyse sur les quantités d'intérêt soit la meilleure possible. Cette optimisation repose sur le modèle d'analyse et en exploite les propriétés. Une analyse statistique n'est jamais menée sans une connaissance minimale du problème, et la prise en compte, dans le modèle statistique, de toute information pertinente déjà disponible, permet aussi un gain de précision. Les statisticiens ont développé beaucoup d'outils pour permettre une inférence efficace, mais sans jamais lever le caractère intrinsèquement probabiliste de la démarche. Il reste toujours une incertitude : statistique n'est pas divination. Un exemple historique est celui des élections présidentielles américaines de 1948. Les sondeurs d'opinion, rendus trop confiants par leur précédent succès, annoncent la victoire de Dewey, mais Truman l'emporte et le jour de son investiture, les sénateurs – déçus – de l'Indiana observeront une minute de silence « à la mémoire du Dr Gallup », le fondateur de l'institut de sondage qui porte son nom !

François Rodolphe
Laboratoire de mathématique, informatique et génome (INRA), *Jouy-en-Josas*

généralement de présenter des caractéristiques valorisées par la société. Elles auraient du pouvoir, ce qui d'après certaines théories sociologiques, serait plus bénéfique pour les hommes que pour les femmes. Il serait donc « naturel » que des parents attrayants aient plus de garçons. Nous ne prétendons pas que c'est vrai ; nous disons simplement que cela pourrait l'être, mais qu'on pourrait tout aussi bien imaginer une argumentation aboutissant au fait qu'ils ont plus de filles... Ce qui n'est pas sans poser quelques difficultés !

TROUVER DES DIFFÉRENCES LÀ OÙ IL N'Y EN A PAS

Comparez ces deux affirmations : « Les parents beaux ont plus de filles » et « Il n'est pas prouvé que des parents beaux aient plus ou moins de filles ». Nul doute que la première, sensationnelle, ferait davantage les gros titres ! Les éditeurs des revues sérieuses où l'affirmation de Kanazawa a été publiée ont-ils eux-mêmes été influencés ? Sans doute, et à cela deux raisons possibles. D'une part, les erreurs statistiques sont parfois difficilement détectables, même par des spécialistes. D'autre part, la signification statistique n'est pas directement liée à la taille des échantillons quand les effets testés sont petits. Avec un échantillon suffisamment grand, on peut presque toujours trouver de petits effets statistiquement significatifs. Mais quand les effets étudiés sont infimes, les études faites sur des cohortes trop petites conduisent à des interprétations abusives.

L'étude du rapport des sexes à la naissance n'est pas neuve. Par exemple, dans son ouvrage *Probabilité, statistiques et vérité*, publié en 1957, Richard von Mises étudia ce rapport pour les naissances de 1907 et 1908 à Vienne, et trouva moins de variations qu'on en attendrait du simple hasard. Il l'attribua à des répartitions des sexes différentes selon les groupes ethniques. Pourtant, l'incertitude obtenue sur les mesures n'était ni plus ni moins que celle d'un hasard pur. Que faire face à cette volonté de trouver des différences là où il n'y en a pas ? Pour éviter ces biais, il faut montrer que les résultats observés représentent des effets réels indépendants de la sélection des échantillons, et trouver un argument biologique confirmant que des effets de l'ordre de 1 % sont importants.

Lorsque nous devons estimer des petits effets, les statisticiens doivent garder un regard critique sur les estimations obtenues. Mais les méthodes d'analyses ne sont pas exemptes de failles méthodologiques : les calculs fréquentistes ne tiennent pas compte des tailles des effets ; les analyses bayésiennes ne sont pas souvent meilleures en utilisant surtout des distributions a *priori* gaussiennes, et ignorent souvent les problèmes de puissance statistique. D'où l'importance d'estimer correctement les incertitudes, notamment en calculant les statistiques sur le signe des effets et sur leur amplitude. Nul doute que l'échange de méthodes et d'idées ouvrira la voie à une meilleure estimation des petits effets. ■

BIBLIOGRAPHIE

A. GELMAN *ET AL.*
Letter to the editor regarding some papers of Dr. Satoshi Kanazawa,
Journal of Theroretical Biology, vol. 245, pp. 597-599, 2007.

S. KANAZAWA
Beautiful parents have more daughters: a further implication of the generalized Trivers-Willard hypothesis,
Journal of Theoretical Biology, vol. 244, pp. 133-140, 2007.

R. VON MISES
Probability, Statistics and Truth, *Dover*, 2006.

L'ESSENTIEL

- La valeur-*p* désigne la probabilité qu'un résultat statistique ne soit pas le fait du hasard.
- Plébiscitée par les chercheurs, elle souffre néanmoins de nombreux travers. Ainsi, elle ne dit rien de la taille de l'effet que l'on étudie.
- Elle se laisse aussi facilement manipuler pour asseoir les résultats recherchés : c'est le «*p*-hacking».
- La recherche gagnerait en qualité et en reproductibilité si l'on tenait compte de la probabilité que l'effet étudié existe réellement.

L'AUTEUR

REGINA NUZZO
est journaliste indépendante et professeure de statistiques à l'université Gallaudet de Washington.

La malédiction de la VALEUR-P

Depuis des décennies, on mesure la qualité des résultats statistiques grâce à la valeur-*p*. Mais cet étalon-or n'est pas aussi fiable qu'on veut bien le croire. Il est temps de le remettre en question.

Un mauvais tireur incapable d'atteindre une cible *(à gauche)* peut néanmoins se déclarer doué en peignant la cible à l'endroit où se trouve par hasard la majorité des impacts de tir *(à droite)*. Cette métaphore illustre le «*p*-hacking», un comportement qui entache de nombreuses publications scientifiques.

Un court instant, Matt Motyl s'est tenu dans l'antichambre de la gloire scientifique. En 2010, il avait découvert lors d'une expérience que les extrémistes voyaient le monde en noir et blanc – littéralement. Ses résultats étaient «clairs comme de l'eau de roche», se souvient le doctorant en psychologie de l'université de Virginie, à Charlottesville. Ses travaux sur 2 000 participants avaient montré que parmi eux, les extrémistes de gauche ou de droite avaient plus de mal à différencier des nuances de gris que les personnes politiquement plus modérées. L'étudiant était confiant et croyait en la solidité de ses résultats. La publication dans une revue renommée semblait à portée de main. Mais rien ne s'est passé comme prévu.

De fait, la robustesse des résultats statistiques avait été calculée par l'indice rituel, l'arme de tous les statisticiens, l'incontournable valeur-*p* (*voir l'encadré ci-contre*). Ici, il était de 0,01, ce qui indique un résultat «très significatif». Tout allait pour le mieux. Hélas, par précaution, Motyl et son encadrant, Brian Nosek, répétèrent l'expérience. Avec de nouvelles données, la valeur-*p* bondit à 0,59, bien au-delà du seuil de 0,05 sous lequel un résultat est considéré comme significatif. Ainsi disparut l'effet… et le rêve de gloire de Motyl. Le test de la valeur-*p* condamnait l'étude aux oubliettes. Et si nous avions affaire à une erreur judiciaire ? Et si la valeur-*p* n'était pas aussi fiable qu'on veut bien le penser ?

LA VALEUR-*P*, CE MOUSTIQUE

De fait, l'étude sur les extrémistes ne souffrait pas de problème de collecte de données ou d'erreur de calcul. La faute revenait bien plus à la nature trompeuse de la valeur-*p* elle-même. Cette dernière n'est en effet pas aussi fiable et objective que le pensent la plupart des scientifiques. Stephen Ziliak, économiste de l'université Roosevelt de Chicago et critique régulier de la façon dont les statistiques sont utilisées, va encore plus loin. Selon lui, «les valeurs-*p* ne font pas leur travail, car elles ne le peuvent pas».

C'est préoccupant, particulièrement au regard des inquiétudes sur la reproductibilité des résultats qui traversent la communauté scientifique, et que le cas de Motyl illustre. John Ioannidis, épidémiologiste à l'université Stanford, avait affirmé en 2005 que la majorité des résultats publiés étaient faux et avait avancé des explications à ce phénomène. Depuis, la reproduction d'études a échoué dans de nombreux cas célèbres, ce qui a contraint les scientifiques à reconsidérer leurs méthodes. En d'autres termes, la façon dont les résultats sont évalués est-elle fiable ? Existe-t-il des alternatives, c'est-à-dire des méthodes avec lesquelles toutes les fausses alertes sont identifiées, sans occulter un vrai résultat ?

La critique de la valeur-*p* n'a rien de nouveau. Depuis sa formalisation par le mathématicien britannique Karl Pearson au début du XX^{e} siècle, on l'a dénigrée en la comparant à un «moustique» (ennuyant et impossible de s'en débarrasser), aux «habits neufs de l'empereur» (il y a un problème, mais personne n'en parle)… Charles Lambdin, d'Intel Corporation, a même proposé de rebaptiser la méthode «Statistical Hypothesis Inference Testing», probablement pour l'acronyme que forment les initiales…

VALEUR-P

La valeur-*p* indique dans quelle mesure il est probable que le résultat présenté dans une étude soit vrai et ne résulte pas du hasard. Ainsi, une valeur-*p* inférieure à 0,05 signifie que nous aurions raison 95 fois sur 100 de croire qu'un effet observé n'est pas une coïncidence.

La valeur 0,05 devint le seuil séparant le significatif de ce qui ne l'est pas. Mais ce n'était pas son rôle

Ironie de l'histoire, lorsque Ronald Fisher (1890-1962), l'un des pères de la statistique moderne, introduisit la valeur-*p* dans les années 1920, il n'avait nullement en tête un test qui déterminerait tout de manière définitive. Il y voyait plus un moyen de juger si un résultat était significatif, ce terme étant à prendre dans un sens non scientifique : le résultat signifie quelque chose, suffisamment pour que l'on y regarde de plus près. Son idée était de mener une expérience et de vérifier ensuite si les données sont cohérentes avec ce que le hasard seul aurait produit.

Pour cela, un chercheur formule d'abord une hypothèse, dite nulle, qui exprime l'inverse