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.

En 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…

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

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.

de ce qu'il veut prouver, par exemple que les individus qui ont reçu un médicament ne sont pas en meilleure santé que ceux ayant reçu un placebo. Le chercheur suppose ensuite que l'hypothèse nulle est vraie et calcule sous cette condition préalable la probabilité d'obtenir des résultats au moins aussi marqués que ceux qu'il a réellement observés. Cette probabilité est la valeur-*p*. Plus elle est petite, d'après Fisher, plus grande est la probabilité que l'hypothèse nulle soit fausse, et donc qu'elle doive être rejetée, ce que le chercheur souhaitait au début, car l'effet étudié est ainsi confirmé.

On peut calculer la valeur-*p* avec autant de chiffres après la virgule que l'on souhaite, mais cette précision apparente est trompeuse, car les hypothèses étudiées sont loin d'être aussi fines. Pour Fisher, la valeur-*p* ne représentait qu'un maillon dans une chaîne qui relie les observations et les connaissances de base pour parvenir à une conclusion scientifique.

Ces précautions ont été balayées par un mouvement dont le but était de rendre la prise de décision fondée sur les preuves aussi rigoureuse et objective que possible. Les pionniers de ce revirement étaient, à la fin des années 1920, le mathématicien polonais Jerzy Neyman et le statisticien britannique Egon Pearson, les rivaux les plus acharnés de Fisher. Leur système inclut des concepts comme « faux positif », « faux négatif » (*voir les Repères, page 6*) et « puissance » d'un test statistique, que l'on trouve aujourd'hui dans tous les cours de statistiques pour débutants. Neyman et Pearson mirent cependant volontairement de côté la valeur-*p*.

Pendant que les représentants des deux camps s'écharpaient, d'autres chercheurs perdirent patience et rédigèrent eux-mêmes des manuels de statistiques. Malheureusement, beaucoup de ces auteurs n'étaient pas suffisamment versés en la matière pour apprécier les subtilités philosophiques des deux visions. Ils intégrèrent donc la valeur-*p* de Fisher, facile à calculer, dans le système régulé, rigoureux et sécurisant de Neyman et Pearson. Depuis, du haut de son piédestal, la valeur-*p* trône toujours. La valeur de 0,05 devint le seuil séparant le significatif de ce qui ne l'est pas. Mais ce n'était pas son rôle.

GUEULE DE BOIS ET TUMEUR

En conséquence, il règne une grande confusion aujourd'hui sur ce que la valeur-*p* exprime réellement. L'étude de Motyl en est une bonne illustration. La plupart des scientifiques interpréteraient sa première valeur-*p* de 0,01 comme une probabilité de fausse alerte de 1 %. C'est pourtant faux. La valeur-*p* ne livre qu'un résumé sommaire des données en considérant comme vraie une hypothèse nulle spécifique. >

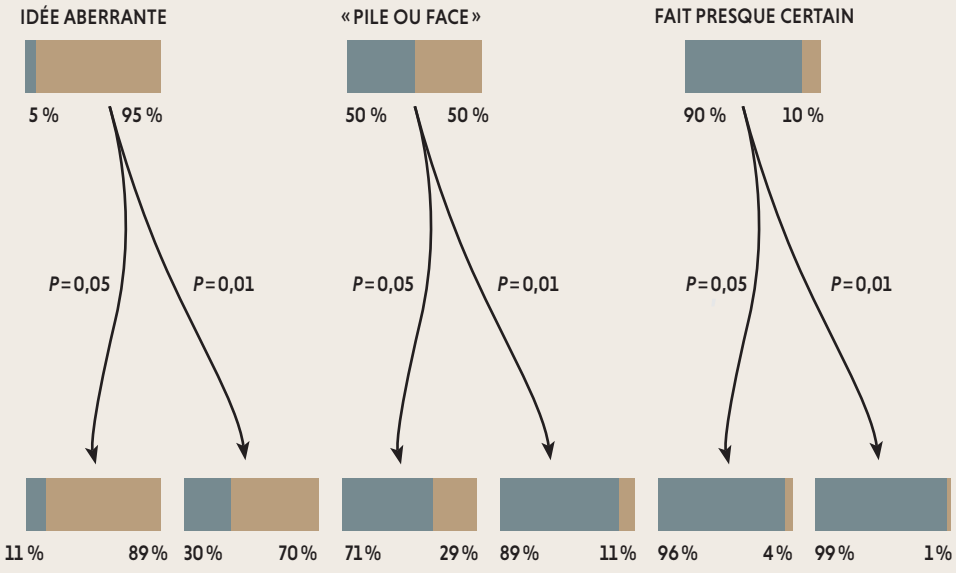
PLAUSIBLE OU NON ?

La valeur-*p* mesure la probabilité qu'un résultat observé soit dû au hasard. Mais elle néglige la question essentielle : quelle est la probabilité que l'hypothèse testée soit correcte ? La réponse dépend de la robustesse du résultat observé et, plus encore, de la plausibilité de l'hypothèse testée.

Avant l'expérience, la plausibilité de l'hypothèse testée (la probabilité qu'elle soit vraie) peut être estimée à partir d'études antérieures, de mécanismes supposés… Ici, on envisage trois cas.

Une valeur-*p* égale à 0,05 traduit conventionnellement un résultat « statistiquement significatif ». À 0,01, c'est « très significatif ».

Après l'expérience, une faible valeur-*p* peut rendre l'hypothèse plus plausible, mais pas nécessairement de façon… significative.



> Une information supplémentaire décisive manque : la probabilité que l'effet en question existe réellement (*voir l'encadré page précédente*). En faire abstraction est semblable à se réveiller le matin avec une migraine et en incriminer une tumeur rare au cerveau ; c'est possible, mais assez peu probable, car bien plus de preuves sont nécessaires pour exclure des explications plus courantes. Moins l'hypothèse est plausible et plus un résultat excitant, une fausse alerte en fait, surgira fréquemment, et cela complètement indépendamment de la valeur-*p*.

Pour le praticien, une telle affirmation est difficile à appréhender. Comment doit-il évaluer la probabilité que l'effet existe ? S'agit-il de la fréquence relative avec laquelle il est constaté au sein d'un ensemble d'études (on parle d'interprétation fréquentielle) ? Ou bien un tel nombre traduit-il les connaissances partielles du chercheur, qui seraient à améliorer par l'expérience seulement (interprétation bayésienne) ? Quoi qu'il en soit, avant une expérience, un chercheur estime grossièrement la probabilité que l'hypothèse étudiée soit vraie.

En établissant cette probabilité dès le départ, et en calculant avec des hypothèses additionnelles judicieuses comment les valeurs-*p* se comportent dans ces conditions, les résultats sont alors moins spectaculaires. Avec une probabilité préalable de 50 % attribuée à l'hypothèse de Motyl, la valeur-*p* de 0,01 signifie alors : dans un cas sur neuf, son expérience lui fait miroiter l'illusion qu'il existe un effet qui n'existe absolument pas.

Par ailleurs, la probabilité que ses collègues puissent reproduire son expérience n'est pas de 99 %, comme on pourrait le penser, mais se situe plutôt autour de 73 %, voire de seulement 50 %, si l'on souhaite à nouveau une valeur-*p* de 0,01. En d'autres termes, le fait que sa seconde expérience soit non concluante est à peu près aussi surprenant que s'il avait perdu au jeu de pile ou face.

L'EFFET DE LA TAILLE DE L'EFFET

De nombreux critiques estiment aussi que la valeur-*p* nuit au raisonnement, en particulier parce qu'elle détourne l'attention de la taille réelle de l'effet. Prenons un exemple. En 2013, une étude de John Cacioppo, de l'université de Chicago, portant sur plus de 19 000 participants montra que les mariages nés d'une rencontre en ligne étaient plus robustes ($p < 0{,}002$) que les autres. En outre, les individus encore mariés avaient tendance à être plus satisfaits de leur mariage que ceux qui s'étaient rencontrés hors ligne ($p < 0{,}001$). Ces valeurs impressionnent, mais l'effet mesuré était infime : une rencontre sur Internet faisait baisser le taux de séparation de 7,67 à 5,96 % et augmentait la satisfaction du couple de 5,48 à 5,64 sur une échelle de 7.

Une autre erreur, sans doute la plus grave, se cache derrière ce que le psychologue Uri Simonsohn, de l'université de Pennsylvanie, nomme le «*p*-hacking». Ce biais consiste à manipuler les données jusqu'à l'obtention du résultat souhaité (*voir la figure page 35*), même sans mauvaises intentions. Ainsi, un changement minime de méthode lors de l'analyse des données peut faire monter le taux de faux positifs d'une étude à 60 %.

Il est difficile d'évaluer à quel point ce problème est répandu, mais selon Simonsohn, le *p*-hacking se multiplierait, parce qu'il serait courant de rechercher de très faibles effets dans des données «bruitées». En analysant des études en psychologie, il a découvert que les valeurs-*p* publiées étaient nombreuses à se situer aux environs de 0,05. Est-ce étonnant si dans les faits les chercheurs partent à la pêche aux valeurs-*p* significatives jusqu'à ce que l'une d'elles, qui les intéresserait, tombe dans leur filet ?

Avec toutes ces critiques, les choses ont-elles changé ? Peu. John Campbell, aujourd'hui chercheur en psychologie à l'université du Minnesota à Minneapolis, le déplorait déjà en 1982, lorsqu'il était éditeur du *Journal of Applied Psychology* : «Il est pratiquement impossible d'arracher les auteurs à leurs valeurs-*p*. Et plus il y a de zéros derrière la virgule, plus ils s'y accrochent.»

RECETTE POUR LA VALEUR-P

D'abord, on identifie les résultats attendus d'une expérience, par exemple à partir d'études précédentes. Imaginons qu'en France, les voitures rouges soient deux fois plus souvent verbalisées que les bleues. Vous souhaitez vérifier que votre région reflète bien l'échelon national. Votre échantillon consiste en 150 amendes : le résultat attendu est donc 100 amendes pour les rouges et 50 pour les bleues. Les observations sont de respectivement 90 et 60. Les différences sont-elles le fruit du hasard ? Le calcul de la valeur-*p* s'impose.

On détermine d'abord le degré de liberté qui rend compte de la variabilité dans la recherche. Il correspond à *n*-1, *n* étant le nombre de variables (ici, 2, rouge et bleu). Le degré de liberté est donc 1.

On calcule ensuite le χ^2 (prononcez *Khi-2*) qui mesure la différence entre la théorie et les observations :

$\chi^2 = \Sigma((o-e)^2/e)$,

o correspondant aux données observées et *e* à celles attendues ou théoriques. On obtient ici :

$\chi^2 = ((90-100)^2/100) + ((60-50)^2/50)$

$\chi^2 = ((-10)^2/100) + (10)^2/50)$

$\chi^2 = (100/100) + (100/50) =$

$\chi^2 = 1 + 2 = 3$

Enfin, dans des registres statistiques de référence (accessibles sur Internet) reliant le degré de liberté et le χ^2, on trouve la valeur-*p* associée. Elle est ici comprise entre 0,05 et 0,1. Supérieure à 0,05, elle ne permet pas de conclure. On sait juste que les résultats observés ont entre 5 et 10 % de chances d'être dus au hasard. Ce n'est pas significatif.

© D'après WikiHow

Chaque tentative de réforme devra combattre des habitudes fermement établies : la façon dont les statistiques sont enseignées dans les universités, celle dont les résultats des études sont exploités et interprétés et comment ils sont ensuite relayés dans les revues spécialisées. Mais au moins, de nombreux scientifiques ont admis l'existence d'un problème. Grâce à des chercheurs comme John Ioannidis, les préoccupations des statisticiens ne sont plus vues uniquement comme de la pure théorie.

Il est pratiquement impossible d'arracher les auteurs à leur sacro-sainte valeur-*p*

Pour améliorer la situation, les statisticiens ont quelques outils à leur disposition. Par exemple, les chercheurs pourraient toujours publier également les tailles d'effet et les intervalles de confiance (*voir les Repères, page 6*) obtenus. Ces valeurs expriment ce que ne peut exprimer la valeur-*p* seule : la portée et l'importance relative de l'effet.

Beaucoup plaident aussi pour remplacer la valeur-*p* par des méthodes fondées sur la vision de Thomas Bayes, mathématicien britannique du XVIII^e^ siècle. Cette approche décrit une façon de penser les probabilités comme la plausibilité d'un résultat, plutôt que la fréquence potentielle de ce résultat. On introduit certes par là une subjectivité dans les statistiques, que les pionniers du début du XX^e^ siècle voulaient éviter à tout prix. Pourtant, la statistique bayésienne simplifie l'intégration des connaissances contextuelles sur le monde et permet de calculer comment les probabilités sont modifiées par des preuves nouvellement acquises.

D'autres soutiennent une approche plutôt pluraliste : les scientifiques devraient utiliser plusieurs méthodes sur un même ensemble de données. Lorsque les résultats différeront, les chercheurs devront être plus créatifs pour découvrir à quoi cela est dû. La compréhension de la réalité sous-jacente y gagnerait.

Aux yeux de Simonsohn, la meilleure protection pour un chercheur est de tout montrer. Les auteurs devraient garantir leur travail « sans *p*-hacking » en explicitant le choix de la taille de leurs échantillons, les données éventuellement mises de côté ainsi que les manipulations mises en œuvre. Aujourd'hui, aucune de ces informations n'est disponible ni vérifiable.

DES ÉTUDES EN DEUX ACTES

À l'instar de la mention « Les auteurs n'ont pas d'intérêts financiers dépendants du contenu de cet article » encore courante aujourd'hui, ces informations feraient la différence entre faute scientifique involontaire ou délibérée. Lorsqu'une telle déclaration sera monnaie courante, le *p*-hacking sera éliminé. Ou au moins son absence sera-t-elle remarquée par le lecteur et lui permettra un jugement plus éclairé.

L'analyse en deux étapes ou « *preregistered replication* » (réplication préenregistrée) est une idée qui va dans ce sens, et elle gagne du terrain. Dans cette approche, les études exploratoires et confirmatoires sont abordées différemment et clairement identifiées. Au lieu, par exemple, de conduire quatre petites expériences et de réunir les résultats dans un article, les chercheurs balaieraient d'abord le domaine pour récolter des observations intéressantes avec deux petites études exploratoires, sans trop se préoccuper des fausses alertes. C'est seulement après, sur la base de ces données, qu'ils concevraient une étude qui confirmera, peut-être, leurs résultats, et préenregistreraient publiquement leurs intentions dans une banque de données publique, telle l'Open Science Framework. Ils publieraient ensuite leurs résultats, accompagnés de ceux de l'étude exploratoire dans un article d'une revue habituelle. Une telle approche laisse beaucoup de liberté, explique le politologue et statisticien Andrew Gelman, de l'université Columbia à New York. Mais elle est aussi suffisamment rigoureuse pour diminuer le nombre de fausses découvertes.

Plus généralement, l'heure est venue pour les scientifiques de prendre conscience des limites des statistiques conventionnelles. Avant tout, une estimation scientifique sérieuse de la plausibilité des résultats devrait être effectuée dès leur analyse : quels résultats ont été apportés par des recherches similaires ? Existe-t-il un mécanisme qui pourrait les expliquer ? Les résultats se recoupent-ils avec l'expérience clinique ? Voilà les questions décisives ! En y répondant dans de prochains travaux, Motyl peut encore espérer accéder au succès qu'il espère ! ■

BIBLIOGRAPHIE

D. BENJAMIN *ET AL.*,
***Redefine statistical significance,* 2017**
https://psyarxiv.com/mky9j

B. NOSEK *ET AL.*,
Restructuring incentives and practices to promote truth over publishability,
***Perspect. Psychol. Sci.*,**
vol. 7, pp. 615-631, 2012.

C. LAMBDIN
Significance tests as sorcery : Science is empirical – significance tests are not,
***Theory & Psychology*,**
vol. 2, pp. 67-90, 2012.

Le *big data*, un flot de données venant de toutes parts.