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.