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…

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.