L'ESSENTIEL

> L'utilisation de la valeur-*p*, qui sert depuis près d'un siècle à jauger la qualité statistique de résultats expérimentaux, a donné une fausse illusion de certitude et a mené à des crises de reproductibilité dans de nombreux champs scientifiques.

> Refonte totale ou simples aménagements ? Malgré le consensus grandissant pour une réforme de l'analyse statistique, les chercheurs sont en désaccord sur l'ampleur à lui donner.

> Diverses pistes sont envisagées, mais une chose est sûre : il est temps pour les scientifiques et le grand public d'accepter l'inconfort du doute.

L'AUTRICE

LYDIA DENWORTH
journaliste scientifique à New York, aux États-Unis

La valeur-*p* : un problème significatif

Les méthodes statistiques pour jauger la pertinence d'un résultat expérimental sont attaquées de toutes parts. Résisteront-elles ?

En 1925, le généticien et statisticien britannique Ronald Fisher publiait un livre intitulé *Statistical Methods for Research Workers* (*Les Méthodes statistiques adaptées à la recherche scientifique*). Il n'avait *a priori* rien d'un best-seller, pourtant il remporta un énorme succès et installa Fisher comme le père des statistiques modernes. Dans l'ouvrage, Fisher expliquait aux chercheurs comment utiliser les statistiques pour éprouver la robustesse de leurs conclusions, afin qu'ils puissent décider ou non de donner une suite à leurs travaux. Fisher y décrivait notamment un test qui examinait la compatibilité des résultats avec un modèle et produisait en sortie un nombre : la valeur-*p* ou *p*-valeur (*p-value* en anglais). Il fixa un seuil à $p = 0{,}05$: « Il est commode de prendre ce point comme limite pour juger si une déviation [par rapport au modèle] doit être considérée comme significative ou non. » Fisher conseillait aux chercheurs de poursuivre leurs travaux s'ils obtenaient des valeurs-*p* en deçà. Ainsi naissait l'idée qu'une valeur-*p* inférieure à 0,05 apposait le sceau « statistiquement significatif » sur des recherches.

Aujourd'hui, presque un siècle plus tard, une valeur-*p* inférieure à 0,05 est considérée comme la référence pour jauger la qualité statistique d'une expérience. Elle ouvre aux chercheurs les portes du monde académique – au financement et à la publication – et dès lors détermine en partie la majorité des résultats scientifiques publiés. Mais même Fisher avait conscience des limites de ce critère statistique. La plupart de ces travers sont connus depuis des décennies. « Le recours excessif aux tests de pertinence statistique, écrivait le psychologue américain Paul Meehl en 1978, est une façon appauvrie de faire de la science. » La valeur-*p* est fréquemment mal interprétée et l'on a tendance à oublier qu'un résultat statistiquement significatif n'est pas forcément significatif en pratique.

En outre, la méthodologie pratiquée dans les laboratoires donne la possibilité aux expérimentateurs, consciemment ou non, d'augmenter >

> ou diminuer artificiellement une valeur-*p*. « Comme on le dit souvent, vous pouvez prouver tout et son contraire avec les statistiques », explique le statisticien et épidémiologiste Sander Greenland, professeur émérite à l'université de Californie à Los Angeles et l'une des principales voix qui appellent à une réforme du système. Les études qui visent seulement à être statistiquement significatives produisent régulièrement des affirmations inexactes – elles donnent l'apparence du vrai au faux et *vice versa*. Quand Fisher a pris sa retraite, on lui a demandé si sa longue carrière lui inspirait un regret : « Avoir mentionné 0,05 », aurait-il confié.

DE NOMBREUSES FAILLES

Au cours de la dernière décennie, le débat sur les tests de robustesse s'est enflammé. Une publication a qualifié les fragiles fondations de l'analyse statistique de « secret le plus sale de la science ». Une autre a dénoncé de « nombreuses et profondes failles » dans ces tests. Une crise majeure a secoué l'économie expérimentale, la recherche biomédicale et surtout la psychologie lorsqu'on a découvert qu'une proportion substantielle des résultats publiés dans ces disciplines n'étaient pas reproductibles. Un des exemples les plus célèbres est l'idée qu'adopter une posture de victoire (les bras en V) donnerait non seulement du courage, mais altérerait la composition du sang en hormones, ce qui revient à affirmer que le langage corporel influe sur le fonctionnement biologique ; ce pseudo-résultat, qui reposait sur une unique publication, a depuis été invalidé par l'un de ses auteurs.

De même, un article sur l'économie du changement climatique (écrit par un climatosceptique) « s'est retrouvé avec presque autant d'erreurs corrigées que de données brutes – sans blague ! –, sans qu'aucune de ces erreurs amène l'auteur à revoir sa conclusion », dénonce Andrew Gelman, statisticien de l'université Columbia, à New York, sur son blog, où il pourfend la mauvaise science et dénonce la difficulté des chercheurs à reconnaître leurs erreurs.

Il est devenu clair que la notion de résultat significatif est l'une des sources du problème, même si ce n'est pas la seule. Ces trois dernières années, les scientifiques ont en masse appelé à une réforme urgente du système. L'association américaine de statistique, qui a publié une déclaration forte et inhabituelle sur la question en 2016, plaide pour le « passage au monde d'après $p < 0,05$ ». Ronald Wasserstein, le directeur de l'association, résume avec humour le problème : « La significativité statistique d'un résultat ne devrait pas avoir plus d'importance qu'un *swipe* vers la droite dans Tinder. Elle indique seulement le niveau d'intérêt qu'on lui porte. Malheureusement, elle est devenue tout autre chose. Les gens disent : "Je >

LA SIGNIFICATIVITÉ STATISTIQUE

Imaginez que vous cultivez des citrouilles dans votre jardin. L'emploi d'engrais améliorera-t-il leur croissance ? D'après votre longue expérience sans fertilisant, vous savez que le poids des citrouilles varie et en connaissez la valeur moyenne : 4,5 kg. Vous décidez de cultiver un échantillon de 25 citrouilles avec engrais. À maturité, leur poids moyen s'élève à 6 kg. Comment déterminer si l'écart de 1,5 kg par rapport au poids moyen sans engrais est dû au hasard – hypothèse dite « nulle » – ou au fertilisant ?

La solution du statisticien Ronald Fisher implique d'effectuer une expérience de pensée : supposez que vous cultiviez 25 citrouilles un grand nombre de fois. À chaque fois, le poids moyen des citrouilles différerait en raison de leur variabilité individuelle. Ensuite, vous traceriez la répartition de ces valeurs moyennes, soit la répartition des poids moyens des citrouilles, centrée sur la valeur correspondant à l'hypothèse nulle (4,5 kg). Il s'agit alors de considérer la probabilité – la valeur-*p* – d'obtenir le poids moyen issu de votre expérience avec engrais (6 kg) si l'engrais n'avait eu aucun effet. Par convention, le seuil de significativité a été fixé à 0,05. Ici, il s'agit donc de déterminer si, dans la répartition ainsi construite, la probabilité d'obtenir un poids moyen de 6 kg est inférieure à 0,05. Mais d'autres approches existent.

L'IMPORTANCE DE L'EFFET

L'importance de l'effet d'un traitement est une grandeur statistique qui évalue l'écart entre les résultats moyens obtenus avec traitement et sans (expérience contrôle). Dans l'exemple des citrouilles, l'importance de l'effet se mesure en kilogrammes. Mais dans de nombreuses situations – comme des réponses à des questionnaires en psychologie –, il n'existe pas d'unité naturelle. Dans ce cas, les chercheurs recourent souvent à l'importance relative de l'effet. Une façon de mesurer celle-ci est fondée sur l'ampleur du chevauchement des distributions des données avec et sans traitement.