> suis sous la barre des 0,05, c'est bon." Et la science s'arrête. »

La question maintenant est de savoir si les mentalités vont changer. « Il n'y a rien de neuf au fond. Cela doit nous faire envisager la possibilité que les mauvaises pratiques perdureront », observe Daniel Benjamin, spécialiste en économie comportementale à l'université de Californie du Sud, pourtant l'un des promoteurs de la réforme. Toutefois, même s'ils ont du mal à s'entendre sur le remède, nombre de chercheurs s'accordent à dire, comme l'économiste américain Stephen Ziliak l'a écrit, que « la culture actuelle du test de significativité statistique, son interprétation et son partage doivent cesser ».

LE BOSON DE HIGGS : UN RÉSULTAT EXCEPTIONNELLEMENT ROBUSTE

Les chercheurs utilisent des modèles statistiques pour inférer la vérité – déterminer, par exemple, si un traitement thérapeutique est plus efficace qu'un autre ou si un groupe diffère d'un autre. Tout modèle statistique repose sur un ensemble de suppositions sur la façon dont les données sont collectées et analysées, et dont les chercheurs choisissent de présenter leurs résultats. Presque toujours, ces résultats sont fondés sur une approche dite « de l'hypothèse nulle », laquelle produit une valeur-*p* (*voir l'encadré page 42*). Problème : ce test n'aborde pas la vérité de front, mais de biais. « Ce que l'on veut connaître quand on fait une expérience, c'est la probabilité que l'hypothèse de départ soit vraie, explique Daniel Benjamin. Mais [l'approche de l'hypothèse nulle] répond à une autre question, alambiquée : si mon hypothèse était fausse, à quel point mes données seraient-elles improbables ? »

Parfois, cette démarche porte ses fruits. La quête du boson de Higgs, une particule élémentaire théorisée par les physiciens dans les années 1960, est l'exemple extrême. L'hypothèse nulle était que le boson de Higgs n'existe pas. Les équipes du Grand collisionneur de hadrons (LHC), au Cern, à Genève, ont réalisé quantité d'expériences et ont fini par obtenir une valeur-*p* si faible que la probabilité que leurs résultats se réalisent, si le boson n'existait pas, était d'une chance sur 3,5 millions. Cela rendait l'hypothèse nulle intenable. Puis les chercheurs ont reproduit leurs résultats pour balayer tout risque d'erreur.

« La seule façon pour eux de s'assurer de l'importance scientifique du résultat, et du prix Nobel, était de montrer qu'ils avaient déployé un effort incommensurable pour écarter tous les problèmes capables de produire une si petite valeur, analyse Sander Greenland. Un tel nombre affirme que le modèle standard

© Mark Allen Miller

sans le boson de Higgs est incorrect – et le crie haut et fort. »

Mais la physique autorise un niveau de précision inaccessible dans les autres disciplines. Dans des tests sur des humains, comme en psychologie, on n'atteint jamais des probabilités d'un sur trois millions. Une valeur-*p* de 0,05 signifie qu'il y a une chance sur 20 qu'une hypothèse correcte soit rejetée plusieurs fois lors d'une multitude de tests (et n'indique pas, comme on le croit souvent, que la probabilité d'erreur sur un test unique est de 5 %). C'est pourquoi, jadis, les statisticiens énonçaient leurs résultats en ajoutant des « intervalles de confiance » – une façon d'estimer l'erreur ou l'incertitude dont ils ne pouvaient se départir.

Les intervalles de confiance sont mathématiquement reliés à la valeur-*p*. La valeur-*p* est comprise entre 0 et 1. Si vous soustrayez 0,05 à 1, vous obtenez 0,95, ou 95 %, l'intervalle de confiance classique. Mais, au fond, un intervalle de confiance n'est qu'un moyen commode pour résumer les résultats des tests de certaines hypothèses. « Il n'y a rien [dans ces intervalles] qui devrait inspirer de la confiance », rappelle Sander Greenland. Pourtant, au fil du temps, la valeur-*p* et les intervalles de confiance ont tenu bon, offrant l'illusion de la certitude.

La valeur-*p* elle-même n'est pas nécessairement le problème. C'est un outil utile quand il est manié à bon escient. Le problème survient quand la significativité statistique des

résultats est exagérée, une situation qui se produit facilement avec de petits échantillons. C'est l'origine de la crise actuelle de reproductibilité des expériences. En 2015, Brian Nosek, cofondateur du Center for open science, un organisme américain qui vise à favoriser l'intégrité de la recherche scientifique, a dirigé un effort collectif destiné à reproduire cent expériences majeures en psychologie sociale. Le résultat ? Seulement 36,1 % de ces expériences ont pu être reproduites sans ambiguïté. En 2018, le *Social sciences replication project*, un projet mené par un consortium international de chercheurs, a dressé le bilan de la répétition de 21 études en sciences sociales qui avaient été publiées dans *Nature* et *Science* entre 2010 et 2015. Dans 13 études seulement (62 % des cas), les nouveaux résultats allaient dans le sens de la conclusion de la publication originale, et l'« importance de l'effet » était en moyenne deux fois moindre. Cette grandeur statistique caractérise l'ampleur de l'écart obtenu par rapport à l'hypothèse nulle (*voir l'encadré page 42*).

La génétique a aussi connu sa crise de reproductibilité dans la première moitié des années 2000. Après un long débat, le seuil de significativité statistique a ainsi été considérablement déplacé. « Quand vous découvrez un variant génétique lié à une maladie ou à un autre phénotype, le standard de significativité statistique est 5×10^{-8}, ce qui équivaut à 0,05 divisé par 1 million, décrit Daniel Benjamin, qui mène aussi des recherches en génétique. La nouvelle génération d'études génétiques est réputée comme très robuste. »

Hélas, on ne peut pas en dire autant de la recherche biomédicale, où le risque de faux négatifs (le fait de rapporter qu'un effet n'est pas significatif alors qu'il est par ailleurs bien réel) est élevé. L'absence de preuve n'est pas la preuve de l'absence. De telles méprises conduisent parfois devant les tribunaux lorsque la responsabilité des entreprises et la sécurité des consommateurs sont en jeu.

Jusqu'à quel point la science vacille-t-elle ? Globalement, les scientifiques s'accordent sur le fait que le malentendu autour de la valeur-*p* et son utilisation à tout-va sont de réels problèmes dans certaines disciplines. Toutefois, des chercheurs se montrent moins durs dans leur diagnostic. « Je vois les choses à long terme, déclare Blair Johnson, spécialiste en psychologie sociale à l'université du Connecticut. La science fonctionne par phases. Le pendule balance entre deux extrêmes, il faut bien vivre avec. » Selon lui, cette crise a le bénéfice de rappeler aux chercheurs qu'ils doivent rester prudents dans leurs déductions.

DES PISTES DE SOLUTIONS

Pour avancer, toutefois, les scientifiques devront s'accorder sur des solutions – un défi aussi difficile que la pratique des statistiques elle-même. « La crainte est qu'en renonçant à cette pratique de longue date qui permettait de déclarer des résultats comme statistiquement significatifs ou non, on introduise une forme d'anarchie dans le processus », dépeint Ronald Wasserstein. Pourtant, les suggestions ne manquent pas. Elles incluent des changements dans les méthodes statistiques, dans le langage utilisé pour décrire ces méthodes et dans la façon dont les analyses statistiques sont utilisées. Les idées les plus marquantes ont été mises en avant dans une série d'articles publiés à la suite de la déclaration de l'Association américaine de statistique en 2016, où plus d'une vingtaine de statisticiens s'étaient mis d'accord sur plusieurs principes de réforme.

Ainsi, en 2018, 72 scientifiques ont publié un commentaire intitulé « Redéfinir la significativité statistique » dans la revue *Nature Human Behaviour*, où ils recommandaient d'abaisser le seuil de *p* de 0,05 à 0,005 pour qualifier un résultat de « découverte ». Des résultats entre 0,05 et 0,005 seraient juste considérés comme « indicatifs ». Daniel Benjamin, le premier auteur de cet article, voit dans cette mesure une solution imparfaite, à court terme, mais exploitable immédiatement.

Pour d'autres, en revanche, redéfinir la pertinence statistique ne sert à rien, car le vrai problème est l'existence même d'un seuil. Dans un article publié en 2019 dans la revue *Nature*, Sander Greenland et deux autres chercheurs – Valentin Amrhein, zoologiste à l'université de Bâle, en Suisse, et Blakeley McShane, statisticien et expert en marketing à l'université Northwestern, aux États-Unis – prônent ainsi l'abandon pur et simple du concept de significativité statistique. Pour eux, la valeur-*p* doit être employée comme variable continue >

La probabilité que leurs résultats se réalisent si le boson n'existait pas était d'une chance sur 3,5 millions

> (c'est-à-dire non restreinte à un seuil) parmi d'autres éléments de preuve. Par ailleurs, ils plaident pour renommer les intervalles de confiance « intervalles de compatibilité » – un terme qui refléterait mieux leur sens : la compatibilité avec les données, et non la confiance dans le résultat. Après un appel lancé à la communauté scientifique sur Twitter, les chercheurs ont reçu le soutien de 800 collègues, dont Daniel Benjamin.

Par ailleurs, des méthodes statistiques meilleures ou au moins plus simples existent. Andrew Gelman lui-même n'emploie pas le test de l'hypothèse nulle. Il préfère la méthodologie bayésienne, une approche statistique plus directe dans laquelle on part de présupposés sur l'issue de l'expérience, on ajoute les nouveaux résultats et on met à jour les présupposés. Sander Greenland, lui, encourage l'utilisation de la « surprise », une quantité mathématique qui ajuste la valeur-*p* pour produire des bits (similaires aux bits des ordinateurs) d'information.

Dans cette approche, une valeur-*p* de 0,05 ne représente que 4,3 bits d'information par rapport à l'hypothèse nulle. « Si on lance une pièce de monnaie, cela revient à obtenir 4 faces d'affilée, explique Sander Greenland. Est-ce une preuve suffisante pour démontrer que la pièce est éventuellement truquée ? Non. Cette combinaison se produit souvent. C'est pourquoi 0,05 est une norme si faible. » Selon lui, si les chercheurs accompagnaient chaque valeur-*p* de sa traduction en bits, ils s'astreindraient à des standards plus élevés. Tenir plus compte de l'importance de l'effet aiderait également.

Enfin, améliorer l'éducation aux statistiques à la fois des chercheurs et du public rendrait davantage accessible le jargon des lois du hasard. Au début du xx^{e} siècle, quand Fisher a adopté le concept de significativité, ce mot avait une connotation moins forte qu'aujourd'hui. « Il relevait plus de la signification que de l'importance », rappelle Sander Greenland.

VIVRE AVEC L'INCERTITUDE

L'heure semble être venue de vivre dans l'inconfort de l'incertitude. Si l'on relève ce défi, la littérature scientifique changera de visage. Rapporter une découverte importante devrait « nécessiter un paragraphe, pas une phrase », déclare Ronald Wasserstein. Et ne pas se fonder sur une seule étude.

Les lignes commencent à bouger. « Nous sommes d'accord sur le fait que la valeur-*p* est parfois galvaudée ou mal interprétée, déclare Jennifer Zeis, porte-parole du *New England Journal of Medicine*. Conclure à l'efficacité d'un traitement si $p < 0,05$ et à son inutilité si $p > 0,05$ est une vision étroite de la médecine et ne reflète pas toujours la réalité. » Jennifer Zeis affirme que les travaux publiés dans son journal recourent moins aux valeurs-*p*, voire s'appuient sur des intervalles de confiance sans valeurs-*p*. Sa revue a aussi adopté les principes de la science ouverte – des règles de bonne conduite qui obligent les auteurs à détailler davantage leurs protocoles de recherche et à respecter certains canons d'analyse, tout écart devant être dûment signalé.

De son côté, l'Agence américaine des produits alimentaires et médicamenteux n'impose aucune modification aux essais cliniques, d'après John Scott, qui dirige sa division de biostatistiques. « Il est très improbable que la valeur-*p* disparaisse dans un futur proche du développement des médicaments, mais à long terme les approches alternatives se multiplieront », prédit-il. L'inférence bayésienne, en particulier, suscite un intérêt grandissant.

Blair Johnson, qui vient d'être nommé rédacteur en chef du *Psychological Bulletin*, y promet des changements : « J'ai l'intention d'imposer des règles strictes sur la façon de rapporter une découverte. Afin que les lecteurs sachent ce qui s'est passé et pourquoi. Ils jugeront ainsi plus facilement si les méthodes sont fiables ou présentent des failles. » Il souligne aussi l'importance des métaanalyses et des articles de revue (qui dressent un bilan des découvertes dans un domaine) comme moyens de limiter les répercussions des études uniques.

La valeur-*p* « ne devrait être la clé d'aucune découverte », selon Blakeley McShane, qui prône une approche plus holistique et nuancée, davantage fondée sur l'évaluation. Une position que même les contemporains de Ronald Fisher auraient soutenue. En 1928, deux autres géants des statistiques, Jerzy Neyman et Egon Pearson, écrivaient à propos de l'analyse statistique : « Les tests eux-mêmes ne livrent aucun verdict, mais sont des outils d'aide à la décision. » ■

Améliorer l'éducation aux statistiques rendrait davantage accessible le jargon des lois du hasard

BIBLIOGRAPHIE

F. Masquetiaux, **Le grand ménage de la psychologie,** *Pour la Science*, n° 504, pp. 46-51, 2019.

R. L. Wasserstein *et al.*, **Moving to a world beyond "p < 0.05",** *American Statistician*, vol. 73, Suppl. 1, pp 1-19, 2019.

C. F. Camerer *et al.*, **Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015,** *Nat. Hum. Behav.*, vol. 2, pp. 637-644, 2018.

Résilience
ET DIVERSITÉ ÉNERGÉTIQUE

© ENGIE / VINCENT BRETON

DOSSIER RÉALISÉ EN PARTENARIAT AVEC ENGIE

DIDIER HOLLEAUX,
Directeur Général
Adjoint d'ENGIE

Pendant des années notre système énergétique a su tirer sa force de la combinaison des vecteurs énergétiques et notamment de l'électricité et du gaz. Pourquoi au moment où se développe la vision d'une transition énergétique fondée sur la complémentarité des énergies progressivement décarbonées et décentralisées, devrions-nous opposer l'électricité et le gaz ?

En dehors de l'énergie contenue dans les liaisons moléculaires, les principes de la physique et de la chimie n'ont à ce stade pas identifié d'autres moyens de stocker l'énergie qui soient à la fois denses énergétiquement et aisément mobilisables et transportables. C'est pourquoi les produits pétroliers ont eu un tel succès au cours du siècle dernier. Mais, condamnés notamment par leur empreinte carbone, ils sont amenés à être progressivement remplacés par le gaz (naturel puis vert). L'essor des énergies renouvelables et intermittentes pour la production d'électricité verte accentue le besoin de développer des solutions de production, de stockage et de transport, flexibles et fiables pour répondre aux demandes du réseau électrique.

Si la neutralité carbone à l'horizon 2050 pousse naturellement à une plus forte électrification des usages et si la maîtrise de la demande (efficacité énergétique et pilotage dans le temps) a assurément un rôle à jouer, elles ne doivent pas écarter les différents vecteurs décarbonés (hydrogène, biogaz...) tout aussi indispensables à un mix énergétique à la fois décarboné et économique. Pour répondre aux différents besoins de transition vers le « zéro carbone » de nos clients (citoyens, industries et collectivités locales) ces solutions performantes sont également primordiales.

Le présent cahier a pour ambition de montrer toute la diversité des solutions qui vont rendre effective la réduction des émissions de gaz à effet de serre et rendre réaliste l'objectif de la neutralité carbone en tirant le meilleur parti de toutes les ressources et vecteurs disponibles. Comme la biodiversité, la diversité énergétique rend notre monde plus résilient : elle doit être préservée. ■