**Quelle est la fiabilité des chandelles standard ?**

C'est une question difficile. Les relations qui donnent la luminosité intrinsèque sont-elles réellement universelles ? Une céphéide brille-t-elle de la même façon dans toutes les galaxies ? En effet, les galaxies n'ont pas toutes la même métallicité (le contenu en éléments plus lourds que l'hydrogène, l'hélium et le lithium). La composition des étoiles y est donc différente, ce qui peut influer sur leur comportement. Même question pour les supernovæ : explosaient-elles de la même façon dans l'Univers plus jeune ?

Valider ces hypothèses est un défi majeur. Plus on regarde loin dans l'Univers et plus il est difficile de détecter des supernovæ. Il faut regarder dans la bonne direction, au bon moment. Les futurs grands instruments comme le télescope spatial *James-Webb*, l'*E-ELT* (le futur observatoire géant que l'Europe construit au Chili) ou l'observatoire *Vera-Rubin* (nouveau nom du *LSST*) amélioreront la situation, mais ne permettront probablement pas de confirmer l'universalité des chandelles standard.

**Existe-t-il d'autres méthodes qui soient indépendantes de ces chandelles standard ?**

Une toute nouvelle méthode consiste à utiliser les ondes gravitationnelles émises lors de la fusion de deux étoiles à neutrons, ondes que la collaboration *Ligo-Virgo* a détectées pour la première fois en 2017. On parle de « sirènes standard » ! L'amplitude des ondes gravitationnelles captées permet de déterminer la distance de l'événement. Et contrairement à la coalescence de deux trous noirs, l'événement s'accompagne d'une émission intense d'ondes électromagnétiques dont on déduit, grâce au *redshift*, la vitesse de récession de la paire d'étoiles. Avec seulement six fusions de ce type détectées pour l'instant, la collaboration *Ligo-Virgo* a obtenu pour $H_\circ$ une estimation de 68 kilomètres par seconde et par mégaparsec. Mais les barres d'erreur sont encore très grandes. Les spécialistes de cette méthode estiment qu'il faudrait une cinquantaine d'événements pour avoir une valeur compétitive.

Cette technique est encore assez jeune. Il se pose beaucoup de questions sur sa robustesse. Par exemple, il faut déterminer avec précision la masse des étoiles qui ont fusionné pour connaître l'amplitude initiale, ce qui est loin d'être évident et dépend des modèles. Autre difficulté, les ondes gravitationnelles peuvent être amplifiées sur leur chemin, par un effet de lentille gravitationnelle ; la coalescence peut alors paraître plus proche qu'elle ne l'est en réalité.

**Et vous, sur quelle méthode travaillez-vous ?**

Dans notre projet *HoLiCOW*, l'idée est d'utiliser ce phénomène bien connu en relativité générale qu'est la lentille gravitationnelle. Prenez un objet lointain comme un quasar (une galaxie au noyau très actif et émettant beaucoup de lumière) et imaginez qu'une galaxie très massive se trouve sur la ligne de visée. Celle-ci agit comme une lentille dans un dispositif optique. En déformant l'espace-temps dans son voisinage, la galaxie-lentille dévie la lumière du quasar qui passe près d'elle et la focalise vers nous. La lumière nous parvient en suivant différents chemins de longueurs différentes, ce qui se manifeste par des images multiples du quasar. Quand la lumière du quasar fluctue, celle de ses images aussi, mais en décalé dans le temps (de plusieurs mois ou plus), car les chemins parcourus par la lumière ne sont pas de même longueur. En 1964, Sjur Refsdal a montré qu'il est possible d'exprimer la constante de Hubble en fonction de ce délai temporel et du *redshift*.

L'analyse est complexe, car divers phénomènes perturbent la mesure du délai temporel. Il faut aussi modéliser de la façon la plus réaliste possible la distribution de masse de la galaxie-lentille pour reproduire correctement le champ gravitationnel de la lentille qui, par dilatation du temps, influe sur le délai. Enfin, il faut aussi tenir compte de la matière présente sur le parcours de la lumière.

Qui plus est, de telles configurations de lentilles pour un quasar sont rares. Nous en avons étudié sept avec précision. Nous avons obtenu des valeurs de la constante de Hubble toutes très voisines, ce qui nous donne confiance en notre méthode d'analyse. Notre estimation est d'environ 73,3 kilomètres par seconde et par mégaparsec. Ce résultat est très proche de celui de *SHoES*. Avec les instruments d'observation à venir, nous espérons exploiter 50 quasars d'ici à quelques années. C'est l'objectif du nouveau consortium international *TDCOSMO* (*Time delay cosmography*).

**Peut-on espérer que la crise cosmologique sera bientôt résolue ?**

Aujourd'hui, la diversité de méthodes indépendantes de mesure de $H_\circ$, avec des erreurs systématiques de natures différentes, enrichit le débat. Nous ne savons pas encore expliquer tous les écarts dans les mesures, mais, en comparant méticuleusement ces approches, nous devrions bientôt commencer à y voir plus clair.

Si le désaccord persiste, il faudra probablement repenser les modèles. Différentes pistes existent – les conditions dans le plasma primordial, la nature de l'énergie sombre, et bien d'autres. Nous avons même trop de solutions : pour l'instant, nous avons une serrure, mais toutes les clés fonctionnent. Nous ne savons pas encore laquelle est la bonne !

Propos recueillis par Sean Bailly

## BIBLIOGRAPHIE


K. C. Wong *et al.*, **H0LiCOW XIII. A 2.4 % measurement of $H_0$ from lensed quasars : 5.3 $\sigma$ tension between early and late-Universe probes,** à paraître dans *MNRAS*, 2020. (prépublication sur arXiv).

A. G. Riess *et al.*, **Large Magellanic Cloud cepheid standards provide a 1 % foundation for the determination of the Hubble constant and stronger evidence for physics beyond ΛCDM,** *Astrophys. J.*, vol. 876(1), article 85, 2019.

W. L. Freedman *et al.*, **The Carnegie-Chicago Hubble program. VIII. An independent determination of the Hubble constant based on the tip of the red giant branch,** *Astrophys. J.*, vol. 882(1), article 34, 2019.

**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 >

© Mark Allen Miller

> 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 >

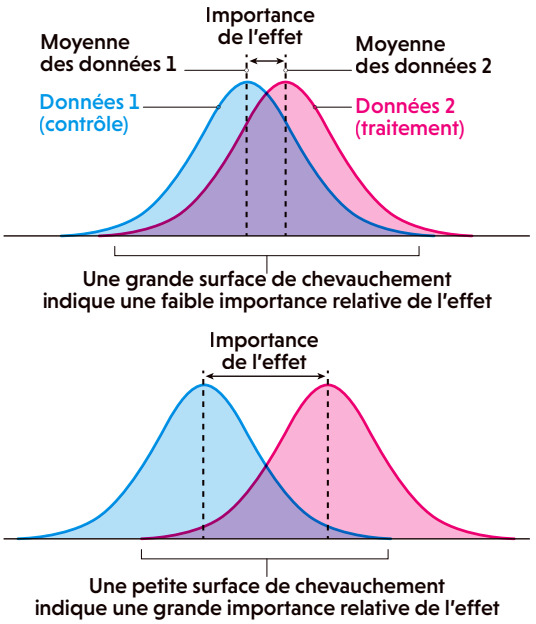
## LA SIGNIFICATIVITÉ STATISTIQUE

**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.

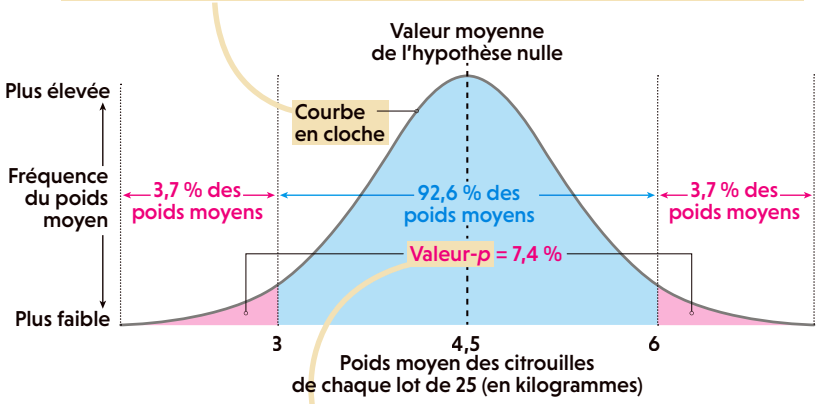



© Illustration d'Amanda Montañez (graphes) et Heather Krause

## LA VALEUR-*P*

Pour calculer la valeur-*p*, on compare le poids moyen des 25 citrouilles cultivées avec engrais (6 kg) avec la répartition des poids moyens que l'on aurait obtenue en cultivant de nombreuses fois 25 citrouilles sans fertilisant.

La répartition du poids moyen des citrouilles, obtenue après de multiples cultures sans engrais de lots de 25, est une courbe en cloche centrée sur la valeur de l'hypothèse nulle.

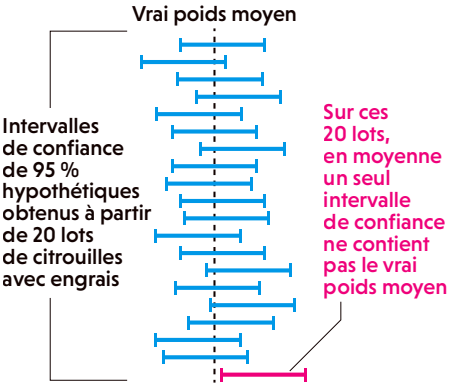



La valeur-*p* est ici la probabilité d'obtenir un lot de 25 citrouilles de poids moyen aussi éloigné du poids standard (4,5 kg) que celui que vous avez obtenu avec engrais (6 kg). Cela revient à calculer la probabilité que le poids moyen des citrouilles d'un lot soit supérieur à 6 kg ou inférieur à 3 kg. Ici, cette probabilité est 0,074. Cette valeur surpassant 0,05, votre résultat ne serait pas considéré comme significatif et l'engrais serait disqualifié.

L'exemple montre un « test bilatéral », où la valeur-*p* représente la probabilité d'obtenir un poids moyen supérieur à 6 kg et inférieur à 3 kg. Dans certaines circonstances, un chercheur pourrait choisir de réaliser un « test unilatéral ». Dans ce cas, la valeur-*p* serait de 0,037, donc inférieure à 0,05, ce qui est considéré comme statistiquement significatif. Voici une situation typique où, selon son intention, l'expérimentateur peut déboucher sur plusieurs valeurs-*p* à partir des mêmes données.

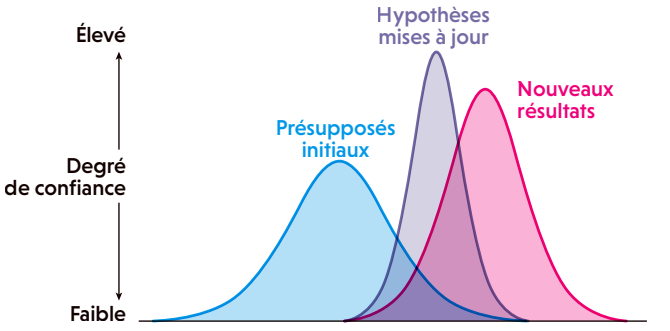
## L'INTERVALLE DE CONFIANCE

Il est possible de calculer un intervalle de confiance de 95 % à partir des données de notre échantillon de 25 citrouilles. Cela revient à deviner le poids moyen des citrouilles avec engrais. Pour cela, on inverse le calcul de la valeur-*p* afin de déterminer tous les poids moyens hypothétiques qui produisent une valeur-*p* supérieure ou égale à 0,05. Pour notre échantillon, l'intervalle de confiance de 95 % s'étend de 4,4 à 7,5 kilogrammes. Le « vrai » poids moyen des citrouilles avec engrais peut se situer ou non dans cet intervalle. Que signifie, alors, ce « 95 % » ? Imaginez ce qui se passerait si l'on cultivait de façon répétée des lots de 25 citrouilles avec engrais. Chaque lot fournirait un intervalle différent et aléatoire. Mais à long terme, 95 % de ces intervalles incluraient le « vrai » poids moyen et 5 % non. Et l'intervalle calculé à partir de notre propre échantillon de départ ? On ne peut savoir s'il se trouve parmi les 95 % qui ont fonctionné ou non. On sait juste que le procédé est correct 95 % du temps.




## LA MÉTHODE BAYÉSIENNE

Dans la méthode d'inférence bayésienne, l'état d'incertitude d'une personne au sujet d'une quantité inconnue est représenté par une loi de probabilité. On utilise le théorème de Bayes pour combiner les présupposés initiaux – la répartition supposée avant d'examiner les données – avec l'information déduite des résultats. On obtient ainsi une nouvelle répartition, qui devient la nouvelle hypothèse de départ pour une prochaine étude, et ainsi de suite. Le choix de critères « objectifs » pour produire les présupposés initiaux est un sujet controversé. L'enjeu est de trouver des façons de construire mathématiquement les hypothèses de départ – des répartitions *a priori* jugées comme raisonnables par une majorité de scientifiques.




## LA SURPRISE

La valeur-*p* traduit l'étonnement que suscitent nos données sur les citrouilles si l'on suppose qu'en réalité l'engrais n'a aucun effet. Or certains chercheurs pensent que la valeur-*p* ne souligne pas cette surprise de façon assez intuitive. Ils prônent sa conversion en une quantité mathématique nommée… « surprise » (ou valeur-*s* ou transformation de Shannon : $s = -\log_2(p)$), qui produit des « bits d'information », ci-dessous illustrés par une expérience de jets de pièces de monnaie. Notre échantillon de 25 citrouilles avec engrais, de poids moyen 6 kg et de valeur-*p* 0,074, produit ainsi entre 3 et 4 bits de surprise (plus exactement $-\log_2(0{,}074) = 3{,}76$).


**Deux faces d'affilée = 2 bits de surprise = valeur-*p* de $1/2^2 = 0{,}25$**


**Quatre faces d'affilée = 4 bits de surprise = valeur-*p* de $1/2^4 = 0{,}0625$**

**Cinq faces d'affilée = 5 bits de surprise = valeur-*p* de $1/2^5 = 0{,}03215$**

> 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 $5 \times 10^{-8}$, 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. »

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

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 >

> (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.

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

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. » ■

## 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.

ENGIE

# Résilience
# ET DIVERSITÉ ÉNERGÉTIQUE

© ENGIE / VINCENT BRETON

**DIDIER HOLLEAUX,**
Directeur Général
Adjoint d'ENGIE

**DOSSIER RÉALISÉ EN PARTENARIAT AVEC 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. ■