> 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 >

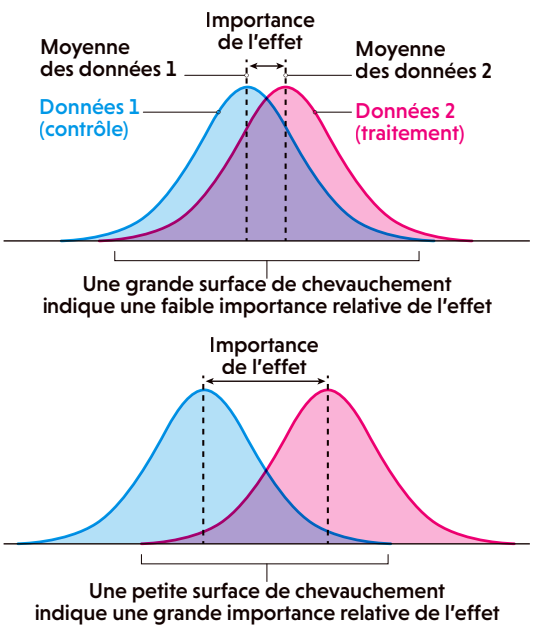
LA SIGNIFICATIVITÉ STATISTIQUE

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

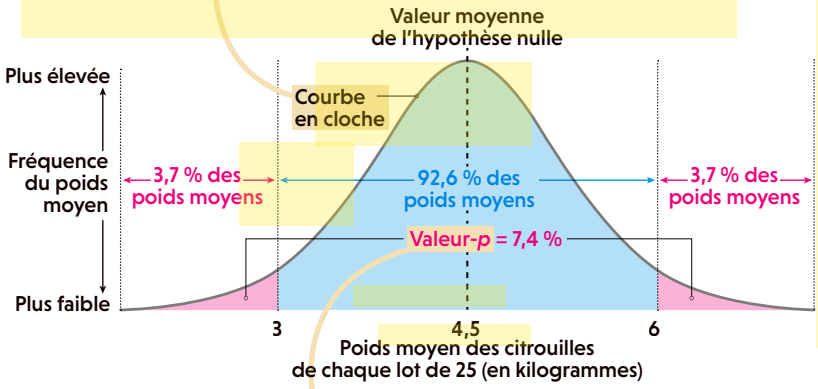
L'importance de l'effet d'un traitement est une grandeur statistique qui évalue l'écart entre les résultats moyens obtenus avec traitement et sans (expérience contrôle). Dans l'exemple des citrouilles, l'importance de l'effet se mesure en kilogrammes. Mais dans de nombreuses situations – comme des réponses 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.



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.

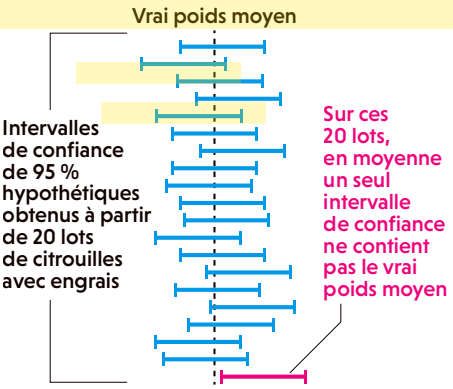


La valeur-*p* est ici la probabilité d'obtenir un lot de 25 citrouilles de poids moyen aussi éloigné du poids standard (4,5 kg) que celui que vous avez obtenu avec engrais (6 kg). Cela revient à calculer la probabilité que le poids moyen des citrouilles d'un lot soit supérieur à 6 kg ou inférieur à 3 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.

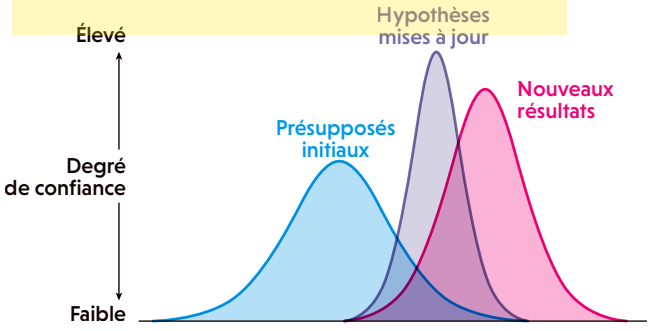
L'INTERVALLE DE CONFIANCE

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