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. >

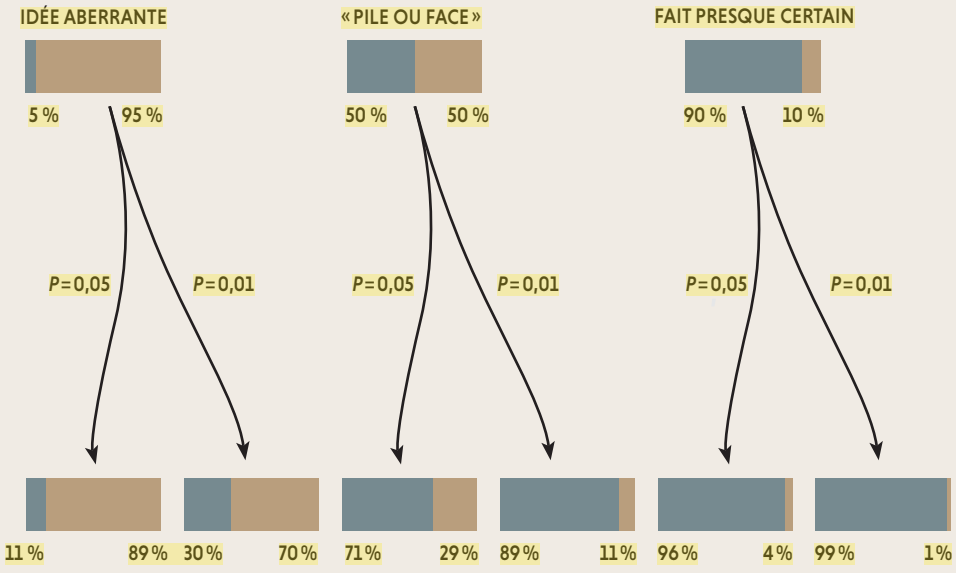
PLAUSIBLE OU NON ?

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.



Une information supplémentaire décisive manque : la probabilité que l'effet en question existe réellement (*voir l'encadré page précédente*). En faire abstraction est semblable à se réveiller le matin avec une migraine et en incriminer une tumeur rare au cerveau ; c'est possible, mais assez peu probable, car bien plus de preuves sont nécessaires pour exclure des explications plus courantes. Moins l'hypothèse est plausible et plus un résultat excitant, une fausse alerte en fait, surgira fréquemment, et cela complètement indépendamment de la valeur-*p*.

Pour le praticien, une telle affirmation est difficile à appréhender. Comment doit-il évaluer la probabilité que l'effet existe ? S'agit-il de la fréquence relative avec laquelle il est constaté au sein d'un ensemble d'études (on parle d'interprétation fréquentielle) ? Ou bien un tel nombre traduit-il les connaissances partielles du chercheur, qui seraient à améliorer par l'expérience seulement (interprétation bayésienne) ? Quoi qu'il en soit, avant une expérience, un chercheur estime grossièrement la probabilité que l'hypothèse étudiée soit vraie.

En établissant cette probabilité dès le départ, et en calculant avec des hypothèses additionnelles judicieuses comment les valeurs-*p* se comportent dans ces conditions, les résultats sont alors moins spectaculaires. Avec une probabilité préalable de 50 % attribuée à l'hypothèse de Motyl, la valeur-*p* de 0,01 signifie alors : dans un cas sur neuf, son expérience lui fait miroiter l'illusion qu'il existe un effet qui n'existe absolument pas.

Par ailleurs, la probabilité que ses collègues puissent reproduire son expérience n'est pas de 99 %, comme on pourrait le penser, mais se situe plutôt autour de 73 %, voire de seulement 50 %, si l'on souhaite à nouveau une valeur-*p* de 0,01. En d'autres termes, le fait que sa seconde expérience soit non concluante est à peu près aussi surprenant que s'il avait perdu au jeu de pile ou face.

L'EFFET DE LA TAILLE DE L'EFFET

De nombreux critiques estiment aussi que la valeur-*p* nuit au raisonnement, en particulier parce qu'elle détourne l'attention de la taille réelle de l'effet. Prenons un exemple. En 2013, une étude de John Cacioppo, de l'université de Chicago, portant sur plus de 19 000 participants montra que les mariages nés d'une rencontre en ligne étaient plus robustes ($p < 0{,}002$) que les autres. En outre, les individus encore mariés avaient tendance à être plus satisfaits de leur mariage que ceux qui s'étaient rencontrés hors ligne ($p < 0{,}001$). Ces valeurs impressionnent, mais l'effet mesuré était infime : une rencontre sur Internet faisait baisser le taux de séparation de 7,67 à 5,96 % et augmentait la satisfaction du couple de 5,48 à 5,64 sur une échelle de 7.

Une autre erreur, sans doute la plus grave, se cache derrière ce que le psychologue Uri Simonsohn, de l'université de Pennsylvanie, nomme le « *p*-hacking ». Ce biais consiste à manipuler les données jusqu'à l'obtention du résultat souhaité (*voir la figure page 35*), même sans mauvaises intentions. Ainsi, un changement minime de méthode lors de l'analyse des données peut faire monter le taux de faux positifs d'une étude à 60 %.

Il est difficile d'évaluer à quel point ce problème est répandu, mais selon Simonsohn, le *p*-hacking se multiplierait, parce qu'il serait courant de rechercher de très faibles effets dans des données « bruitées ». En analysant des études en psychologie, il a découvert que les valeurs-*p* publiées étaient nombreuses à se situer aux environs de 0,05. Est-ce étonnant si dans les faits les chercheurs partent à la pêche aux valeurs-*p* significatives jusqu'à ce que l'une d'elles, qui les intéresserait, tombe dans leur filet ?

Avec toutes ces critiques, les choses ont-elles changé ? Peu. John Campbell, aujourd'hui chercheur en psychologie à l'université du Minnesota à Minneapolis, le déplorait déjà en 1982, lorsqu'il était éditeur du *Journal of Applied Psychology* : « Il est pratiquement impossible d'arracher les auteurs à leurs valeurs-*p*. Et plus il y a de zéros derrière la virgule, plus ils s'y accrochent. »

RECETTE POUR LA VALEUR-P

D'abord, on identifie les résultats attendus d'une expérience, par exemple à partir d'études précédentes. Imaginons qu'en France, les voitures rouges soient deux fois plus souvent verbalisées que les bleues. Vous souhaitez vérifier que votre région reflète bien l'échelon national. Votre échantillon consiste en 150 amendes : le résultat attendu est donc 100 amendes pour les rouges et 50 pour les bleues. Les observations sont de respectivement 90 et 60. Les différences sont-elles le fruit du hasard ? Le calcul de la valeur-*p* s'impose.

On détermine d'abord le degré de liberté qui rend compte de la variabilité dans la recherche. Il correspond à *n*-1, *n* étant le nombre de variables (ici, 2, rouge et bleu). Le degré de liberté est donc 1.

On calcule ensuite le χ^2 (prononcez *Khi-2*) qui mesure la différence entre la théorie et les observations :

$\chi^2 = \Sigma((o-e)^2/e)$,

o correspondant aux données observées et *e* à celles attendues ou théoriques. On obtient ici :

$\chi^2 = ((90-100)^2/100) + ((60-50)^2/50)$

$\chi^2 = ((-10)^2/100) + (10)^2/50)$

$\chi^2 = (100/100) + (100/50) =$

$\chi^2 = 1 + 2 = 3$

Enfin, dans des registres statistiques de référence (accessibles sur Internet) reliant le degré de liberté et le χ^2, on trouve la valeur-*p* associée. Elle est ici comprise entre 0,05 et 0,1. Supérieure à 0,05, elle ne permet pas de conclure. On sait juste que les résultats observés ont entre 5 et 10 % de chances d'être dus au hasard. Ce n'est pas significatif.

© D'après WikiHow