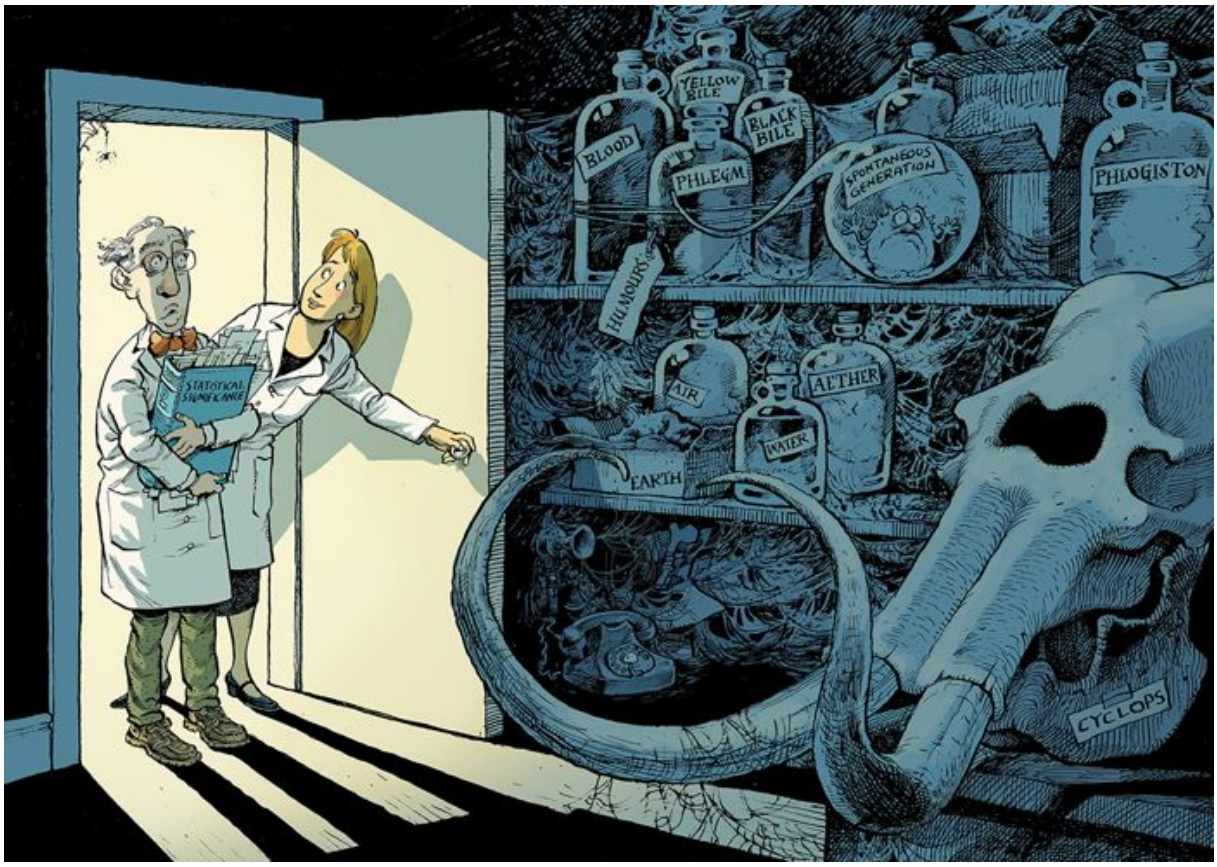

Quand est-ce que la dernière fois que vous avez entendu un orateur du séminaire affirmer qu'il n'y avait «pas de différence» entre deux groupes, parce que la différence était «statistiquement non significative»?

Si votre expérience correspond à la nôtre, il y a de fortes chances que cela se soit produit lors du dernier entretien auquel vous avez assisté. Nous espérons qu'au moins une personne de l'auditoire a été perplexe si, comme cela se produit souvent, un complot ou une table a montré qu'il y avait réellement une différence.

Comment les statistiques amènent-elles si souvent les scientifiques à nier les différences que ceux qui ne sont pas instruits en statistique peuvent constater? Depuis plusieurs générations, les chercheurs ont été avertis qu'un résultat non significatif sur le plan statistique ne «prouvait» pas l'hypothèse nulle (l'hypothèse selon laquelle il n'y a pas de différence entre les groupes ni d'effet d'un traitement sur certains résultats mesurés) [1] . Les résultats statistiquement significatifs ne «prouvent» pas une autre hypothèse. Ces idées fausses ont, de manière célèbre, faussé la littérature en affirmant des affirmations exagérées et, ce qui est moins célèbre, en alléguant des conflits entre des études lorsqu'il n'en existe aucune.

Nous avons des propositions pour empêcher les scientifiques de devenir la proie de ces idées fausses.

Problème omniprésent

Soyons clairs sur ce qui doit cesser: nous ne devrions jamais conclure qu'il n'y a pas de différence ou d'association simplement parce qu'une valeur *P*est supérieure à un seuil tel que 0,05 ou, de manière équivalente, car un intervalle de confiance inclut zéro. Nous ne devrions pas non plus conclure que deux études sont en conflit, l'une ayant un résultat statistiquement significatif et l'autre non. Ces erreurs gaspillent les efforts de recherche et faussent les décisions politiques.

Par exemple, considérons une série d'analyses des effets non voulus des médicaments anti-inflammatoires [2] . Du fait que leurs résultats étaient statistiquement non significatifs, un groupe de chercheurs a conclu que l'exposition aux médicaments n'était «pas associée» à une nouvelle fibrillation auriculaire (la perturbation la plus courante du rythme cardiaque) et que les résultats contrastaient avec ceux d'une fibrillation auriculaire. étude antérieure avec un résultat statistiquement significatif.

Maintenant, regardons les données réelles. Les chercheurs décrivant leurs résultats statistiquement non significatifs ont trouvé un rapport de risque de 1,2 (c'est-à-dire un risque 20% plus élevé chez les patients exposés par rapport aux patients non exposés). Ils ont également constaté un intervalle de confiance de 95%, qui allait de la diminution minime du risque de 3% à une augmentation considérable du risque de 48% ($p = 0,091$; notre calcul).Les chercheurs de la précédente étude, statistiquement significative, ont trouvé exactement le même rapport de risque de 1,2. Cette étude était simplement plus précise, avec un intervalle allant de 9% à 33% de risque en plus ($P = 0,0003$; notre calcul).

Il est ridicule de conclure que les résultats statistiquement non significatifs ont montré «aucune association», lorsque l'estimation de l'intervalle incluait des augmentations de risque sérieuses; il est également absurde de prétendre que ces résultats sont en contraste avec les résultats antérieurs montrant un effet observé identique. Pourtant, ces pratiques courantes montrent à quel point le recours à des seuils de signification statistique peut nous induire en erreur (voir «Attention aux conclusions erronées»).

Erreur ! Nom du fichier non spécifié.

Source: V. Amrhein *et al.*

Celles-ci et les erreurs similaires sont répandues. Des enquêtes portant sur des centaines d'articles ont révélé que les résultats statistiquement non significatifs sont interprétés comme indiquant «pas de différence» ou «pas d'effet» dans

environ la moitié (voir «Mauvaises interprétations» et «Informations supplémentaires»).

En 2016, l'American Statistical Association a publié une déclaration dans *The American Statistician* mettant en garde contre l'utilisation abusive de la signification statistique et des valeurs de p . Le numéro comprenait également de nombreux commentaires sur le sujet. Ce mois-ci, un numéro spécial du même journal tente de pousser plus loin ces réformes. Il présente plus de 40 articles sur "L'inférence statistique au XXIe siècle: un monde au-delà de P <0,05". Les éditeurs présentent la collection avec la prudence suivante: «Ne dites pas« statistiquement significatif »» [3] . Un autre article [4]avec des dizaines de signataires appelle également les auteurs et les éditeurs de revues à désavouer ces termes.

Nous sommes d'accord et demandons que le concept de signification statistique soit abandonné.

Erreur ! Nom du fichier non spécifié.

Source: V. Amrhein *et al.*

Nous sommes loin d'être seuls. Lorsque nous avons invité d'autres personnes à lire un brouillon de ce commentaire et à signer leur nom si elles partageaient notre message, 250 l'ont fait dans les 24 premières heures. Une semaine plus tard, nous avions plus de 800 signataires - tous vérifiés pour une affiliation universitaire ou une autre indication d'un travail actuel ou passé dans un domaine qui dépend de la modélisation statistique (voir la liste et le décompte final des signataires dans les Informations supplémentaires). Ceux-ci comprennent des statisticiens, des chercheurs cliniques et médicaux, des biologistes et des psychologues de plus de 50 pays et de tous les continents, à l'exception de l'Antarctique. Un défenseur l'a qualifié de «frappe chirurgicale contre des tests inconsidérés de signification statistique» et «une occasion de faire entendre votre voix en faveur de meilleures pratiques scientifiques».

Nous ne demandons pas une interdiction des valeurs de p . Nous ne disons pas non plus qu'ils ne peuvent pas être utilisés comme critère de décision dans certaines applications spécialisées (telles que la détermination d'un processus de fabrication répondant à certaines normes de contrôle de la qualité). Et nous ne préconisons pas non plus une situation quelconque, dans laquelle de faibles preuves deviennent soudainement crédibles. Au lieu de cela, et comme de nombreux autres au cours des décennies, nous demandons de mettre fin à l'utilisation des valeurs de P de manière classique et dichotomique - pour décider si un résultat réfute ou soutient une hypothèse scientifique [5] .

Quitter la catégorisation

Le problème est davantage humain et cognitif que statistique: les résultats obtenus avec les compartiments deviennent «statistiquement significatifs» et

«statistiquement non significatifs», ce qui donne à penser que les éléments attribués de cette manière sont catégoriquement différents [6 - 8] .Les mêmes problèmes sont susceptibles de se produire dans toute alternative statistique proposée impliquant une dichotomisation, qu'elle soit fréquente, bayésienne ou autre.

Malheureusement, la fausse croyance selon laquelle le seuil de signification statistique est dépassé suffit à montrer qu'un résultat est «réel» a conduit les scientifiques et les éditeurs de revues à privilégier de tels résultats, faussant ainsi la littérature. Les estimations statistiquement significatives sont biaisées à la hausse et potentiellement dans une large mesure, alors que les estimations non significatives sur le plan statistique sont à la baisse biaisées.Par conséquent, toute discussion portant sur des estimations choisies pour leur importance sera biaisée. De plus, la focalisation rigide sur la signification statistique incite les chercheurs à choisir des données et des méthodes produisant une signification statistique pour un résultat souhaité (ou simplement publiable), ou produisant une signification statistique non significative pour un résultat indésirable, tels que les effets secondaires potentiels de médicaments - invalidant ainsi les conclusions.

Le pré-enregistrement des études et l'engagement de publier tous les résultats de toutes les analyses peuvent beaucoup contribuer à atténuer ces problèmes. Cependant, même les résultats d'études préenregistrées peuvent être biaisés par des décisions invariablement laissées en suspens dans le plan d'analyse [9] . Cela se produit même avec les meilleures intentions.

Encore une fois, nous ne préconisons pas une interdiction des valeurs de p , des intervalles de confiance ou d'autres mesures statistiques, mais seulement que nous ne devrions pas les traiter de manière catégorique. Cela inclut la dichotomisation statistiquement significative ou non, ainsi que la catégorisation sur la base d'autres mesures statistiques telles que les facteurs de Bayes.

Une des raisons d'éviter une telle "dichotomanie" est que toutes les statistiques, y compris les valeurs de P et les intervalles de confiance, varient naturellement d'une étude à l'autre et le font souvent à un degré surprenant. En fait, la seule variation aléatoire peut facilement conduire à de grandes disparités dans les valeurs de P , bien au-delà d'une chute juste de part et d'autre du seuil de 0,05. Par exemple, même si les chercheurs pouvaient réaliser deux études de réplication parfaite ayant un effet réel, chacune avec 80% de puissance (chance) d'atteindre $P < 0,05$, il ne serait pas très étonnant que l'une obtienne $P < 0,01$ et l'autre $P > 0,30$. . Que la valeur Psoit petite ou grande, la prudence est de mise.

Nous devons apprendre à accepter l'incertitude. Une façon pratique de le faire consiste à renommer les intervalles de confiance en «intervalles de

compatibilité» et à les interpréter de manière à éviter toute confiance excessive. Nous recommandons en particulier aux auteurs de décrire les implications pratiques de toutes les valeurs comprises dans l'intervalle, en particulier l'effet observé (ou l'estimation ponctuelle) et les limites. Ce faisant, ils doivent se rappeler que toutes les valeurs comprises entre les limites de l'intervalle sont raisonnablement compatibles avec les données, compte tenu des hypothèses statistiques utilisées pour calculer l'intervalle [7,10] . Par conséquent, sélectionner une valeur particulière (telle que la valeur NULL) dans l'intervalle comme étant "indiqué" n'a aucun sens.

Nous en avons franchement marre de voir de telles «preuves irréfutables et absurdes» ainsi que des prétentions de non-association dans des présentations, des articles de recherche, des critiques et du matériel didactique. Un intervalle contenant la valeur null contient souvent aussi des valeurs non nulles d'une grande importance pratique. Cela dit, si vous considérez que toutes les valeurs à l'intérieur de l'intervalle sont pratiquement sans importance, vous pourrez alors dire quelque chose du type "nos résultats sont les plus compatibles sans effet important".

Lorsque vous parlez d'intervalles de compatibilité, gardez à l'esprit quatre choses. Premièrement, ce n'est pas parce que l'intervalle donne les valeurs les plus compatibles avec les données, compte tenu des hypothèses, que les valeurs situées en dehors de cette plage sont incompatibles. ils sont juste moins compatibles. En fait, les valeurs situées juste en dehors de l'intervalle ne diffèrent pas substantiellement de celles situées juste à l'intérieur de l'intervalle. Il est donc faux de prétendre qu'un intervalle indique toutes les valeurs possibles.

Deuxièmement, toutes les valeurs à l'intérieur ne sont pas également compatibles avec les données, compte tenu des hypothèses. L'estimation ponctuelle est la plus compatible et les valeurs proches sont plus compatibles que celles proches des limites. C'est pourquoi nous incitons les auteurs à discuter de l'estimation ponctuelle, même lorsqu'ils ont une valeur *P* élevée ou d'un intervalle large, ainsi que des limites de cet intervalle. Par exemple, les auteurs ci-dessus auraient pu écrire: «Comme dans une étude précédente, nos résultats suggèrent une augmentation de 20% du risque de fibrillation auriculaire d'apparition récente chez les patients recevant les médicaments anti-inflammatoires. Néanmoins, une différence de risque allant d'une diminution de 3%, d'une petite association négative à une augmentation de 48%, une association substantielle positive, est également raisonnablement compatible avec nos données, compte tenu de nos hypothèses. " En interprétant l'estimation ponctuelle, tout en reconnaissant son incertitude, vous éviterez de faire de fausses déclarations "aucune différence" et de faire des déclarations trop confiantes.

Troisièmement, à l'instar du seuil de 0,05 dont il est issu, la valeur par défaut de 95% utilisée pour calculer les intervalles est en soi une convention

arbitraire. Il repose sur la fausse idée selon laquelle il y a 95% de chances que l'intervalle calculé contienne lui-même la vraie valeur, couplé avec le sentiment vague que c'est un fondement pour une décision confiante. Un niveau différent peut être justifié, en fonction de l'application. Et, comme dans l'exemple des médicaments anti-inflammatoires, les estimations d'intervalle peuvent perpétuer les problèmes d'importance statistique lorsque la dichotomisation qu'elles imposent est traitée comme une norme scientifique.

Dernier point, et le plus important de tous, soyez humble: les évaluations de compatibilité reposent sur l'exactitude des hypothèses statistiques utilisées pour calculer l'intervalle. En pratique, ces hypothèses sont au mieux sujettes à une incertitude considérable [7, 8, 10]. Définissez ces hypothèses aussi clairement que possible et testez celles que vous pouvez, par exemple en traçant vos données et en ajustant des modèles alternatifs, puis en rapportant tous les résultats.

Quelles que soient les statistiques, il est bon de suggérer les raisons de vos résultats, mais discutez d'un éventail d'explications potentielles, et pas seulement de celles qui sont favorisées. Les inférences doivent être scientifiques, et cela va bien au-delà des simples statistiques. Des facteurs tels que les preuves de base, la conception de l'étude, la qualité des données et la compréhension des mécanismes sous-jacents sont souvent plus importants que les mesures statistiques telles que les valeurs de P ou les intervalles.

L'objection que nous entendons le plus contre le retrait de l'importance statistique est qu'il est nécessaire de prendre des décisions par oui ou par non. Toutefois, pour les choix souvent nécessaires dans les environnements réglementaire, politique et professionnel, les décisions basées sur les coûts, les avantages et les probabilités de toutes les conséquences potentielles dépassent toujours celles qui sont prises uniquement sur la base de la signification statistique. De plus, pour décider de poursuivre ou non une idée de recherche, il n'ya pas de lien simple entre une valeur P et les résultats probables d'études ultérieures.

À quoi ressemblera la signification statistique du départ à la retraite? Nous espérons que les sections sur les méthodes et la tabulation des données seront plus détaillées et nuancées. Les auteurs insisteront sur leurs estimations et leur incertitude - par exemple, en discutant explicitement des limites inférieure et supérieure de leurs intervalles. Ils ne s'appuieront pas sur des tests de signification. Lorsque les valeurs P sont rapportées, elles seront données avec une précision raisonnable (par exemple, $P = 0,021$ ou $P= 0,13$) - sans ornements tels que des étoiles ou des lettres pour indiquer la signification statistique et non comme des inégalités binaires ($P <0,05$ ou $P> 0,05$). Les décisions d'interprétation ou de publication des résultats ne seront pas fondées sur des seuils statistiques. Les gens passeront moins de temps avec les logiciels statistiques et plus de temps à réfléchir.

Notre appel à abandonner la signification statistique et à utiliser les intervalles de confiance comme intervalles de compatibilité n'est pas une panacée. Bien qu'il élimine de nombreuses mauvaises pratiques, il pourrait en introduire de nouvelles. Ainsi, surveiller la littérature sur les abus statistiques devrait être une priorité permanente pour la communauté scientifique. Toutefois, l'élimination de la catégorisation contribuera à mettre un terme aux affirmations trop confiantes, aux déclarations non fondées «aucune différence» et aux déclarations absurdes sur «l'échec de la réplication» lorsque les résultats des études originales et de la réplication sont hautement compatibles. L'abus de signification statistique a beaucoup nui à la communauté scientifique et à ceux qui s'appuient sur des avis scientifiques. *Les* valeurs *P* , les intervalles et d'autres mesures statistiques ont toutes leur place, mais il est temps que la signification statistique disparaisse.

Nature **567** , 305-307 (2019)

doi: 10.1038 / d41586-019-00857-9

Références

1. 1.

Fisher, RA *Nature* **136** , 474 (1935).

- **Article**
- **Google Scholar**

2. 2

Schmidt, M. et Rothman, KJ *Int. J. Cardiol.* **177** , 1089-1090 (2014).

- **PubMed**
- **Article**
- **Google Scholar**

3. 3

Wasserstein, RL, Schirm, A. et Lazar,
NA *Am. Stat* .https: / / doi.org / 10.1080 / 00031305.2019.1583913 (2019).

- **Article**
- **Google Scholar**

4. 4

Hurlbert, SH, Levine, RA et Utts,
J. *Am. Stat* .https: / / doi.org / 10.1080 / 00031305.2018.1543616 (2019).

- **Article**
- **Google Scholar**

5. 5

Lehmann, EL *Test d'hypothèses statistiques* 2e édition 70–71 (Springer, 1986).

-

6. 6

Gigerenzer, G. *Adv. Meth. Pract. Psychol. Sci.* **1** , 198-218 (2018).

-
- **Article**
- **Google Scholar**

7. 7.

Groenland, S. *Am. J. Epidemiol.* **186** , 639 à 645 (2017).

-
- **PubMed**
- **Article**
- **Google Scholar**

8. 8

McShane, BB, D. Gal, A. Gelman, C. Robert et Tackett, JL *Am. Stat* .https://doi.org/10.1080/00031305.2018.1527253 (2019).

-
- **Article**
- **Google Scholar**

9. 9

Gelman, A. et Loken, E. *Am. Sci.* **102** , 460 à 465 (2014).

-
- **Article**
- **Google Scholar**

10. dix.

Amrhein, V., Trafimow, D. et Groenland, S. *Am. Stat* .https://doi.org/10.1080/00031305.2018.1543137 (2019).