Canadian Psychology / Psychologie canadienne

Pourquoi tout ce bruit autour de la valeur p? Quelques pistes de compréhension pour le non-expert

--Manuscript Draft--

Manuscript Number:	CAP-2022-0089R3
Full Title:	Pourquoi tout ce bruit autour de la valeur p? Quelques pistes de compréhension pour le non-expert
Abstract:	La valeur p (p-value) fait l'objet de vifs débats dans la communauté scientifique et influence l'évaluation que les intervenants font des recherches dans leur domaine de pratique. Certains souhaitent la proscrire, alors que d'autres veulent continuer à l'utiliser. Cet article présente des éléments de réflexion concernant les écarts entre sa réelle signification et l'usage commun qui en est fait. Il discute des définitions ou interprétations erronées qui lui sont associées, notamment par rapport aux seuils de la valeur p. Enfin, il présente des alternatives ou compléments à cette statistique. Son objectif est d'identifier et de proposer aux chercheurs en sciences sociales, humaines et de la santé les meilleures pratiques de l'analyse des données quantitatives et de développer en même temps une réflexion critique à l'égard de leurs résultats basés sur cette statistique, laquelle continue à attiser bien des passions.
Article Type:	Masked Article / Article anonyme
Keywords:	Valeur p; test d'hypothèse; précision; facteur de Bayes; comparaison de modèles
Corresponding Author:	Sébastien Béland, Ph.D. Université de Montréal: Universite de Montreal CANADA
Corresponding Author E-Mail:	sebastien.beland@umontreal.ca
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Université de Montréal: Universite de Montreal
Other Authors:	Michael Cantinotti
	Denis Cousinea
	Marco Barroca-Paccard
	Marie-Aude Boislard
	Christian Bourassa
	Julien Bureau
	Pier-Olivier Caron
	Lucile Chanquoy
	Christophe Chénier
	Jean-François Daoust
	Éric Dionne
	Stéphanie Forté
	Éric Frenette
	Stéphanie Girard
	Vincent Grenon
	Bradley Harding
	Éric Lacourse
	Daniel Lalande

	Florent Michelot
	Quoc Dinh Nguyen
	Jean-Sébastien Renaud
	Jérôme St-Amand
	Floris Van Vugt
Author Comments:	Nous souhaitons soulever la grande qualité du travail de l'évaluateur 3
Corresponding Author's Secondary Institution:	
First Author:	Sébastien Béland, Ph.D.
Order of Authors Secondary Information:	
Manuscript Region of Origin:	CANADA
Opposed Reviewers:	
Order of Authors:	Sébastien Béland, Ph.D.
	Michael Cantinotti
	Denis Cousinea
	Marco Barroca-Paccard
	Marie-Aude Boislard
	Christian Bourassa
	Julien Bureau
	Pier-Olivier Caron
	Lucile Chanquoy
	Christophe Chénier
	Jean-François Daoust
	Éric Dionne
	Stéphanie Forté
	Éric Frenette
	Stéphanie Girard
	Vincent Grenon
	Bradley Harding
	Éric Lacourse
	Daniel Lalande
	Florent Michelot
	Quoc Dinh Nguyen
	Jean-Sébastien Renaud
	Jérôme St-Amand
	Floris Van Vugt

Title page with All Author Information / Page titre contenant tous les renseignements sur l'auteur

Pourquoi tout ce bruit autour de la valeur p?
Quelques pistes de compréhension pour le non-expert

https://osf.io/prab2/?view_only=db563a9efa754b7d8c86defba5fc0416

1. Sébastien Béland, Université de Montréal, sebastien.beland@umontreal.ca: auteur de correspondance
2. Michael Cantinotti, Université du Québec à Trois-Rivières, Michael.Cantinotti@uqtr.ca
3. Denis Cousineau, Université d'Ottawa, Denis.Cousineau@uottawa.ca
4. Marco Barroca-Paccard, HEP Vaud, marco.barroca-paccard@hepl.ch
5. Marie-Aude Boislard, Université du Québec à Montréal, boislard-pepin.marie-aude@uqam.ca
6. Christian Bourassa, Université de Montréal, christian.bourassa.1@umontreal.ca<
7. Julien Bureau, Université Laval, julien.bureau@fse.ulaval.ca
8. Pier-Olivier Caron, TÉLUQ, Pier-Olivier.Caron@teluq.ca
9. Lucile Chanquoy, Université Nice Sophia Antipolis, lucile.chanquoy@univ-cotedazur.fr
10. Christophe Chénier, Université de Montréal, christophe.chenier@umontreal.ca
11. Jean-François Daoust, Université de Sherbrooke, JF.Daoust@USherbrooke.ca
12. Éric Dionne, Université d'Ottawa, Eric.Dionne@uOttawa.ca
13. Stéphanie Forté, Université de Montréal, steph.forte1@gmail.com
14. Éric Frenette, Université Laval, Eric.Frenette@fse.ulaval.ca
15. Stéphanie Girard, Université du Québec à Trois-Rivières, Stephanie.Girard3@uqtr.ca
16. Vincent Grenon, Université de Sherbrooke, Vincent.Grenon@USherbrooke.ca
17. Bradley Harding, Université de Moncton, bradley.harding@umoncton.ca
18. Éric Lacourse, Université de Montréal, eric.lacourse@umontreal.ca
19. Daniel Lalande, Université du Québec à Chicoutimi, daniel_Lalande@uqac.ca
20. Florent Michelot, Université de Moncton, florent.michelot@umoncton.ca
21. Quoc Dinh Nguyen, Université de Montréal, nguyen.quoc.d@gmail.com
22. Jean-Sébastien Renaud, Université Laval, jean-sebastien.renaud@fmed.ulaval.ca
23. Jérôme St-Amand, Université du Québec en Outaouais, jerome.st-amand@uqo.ca
24. Floris Van Vugt, Université de Montréal, floris.van.vugt@umontreal.ca

Notes :

Texte révisé à la suite des remarques de l'évaluateur 3

Pourquoi tout ce bruit autour de la valeur p? Quelques pistes de compréhension pour le non-expert

03/06/2024

Résumé

La valeur *p* (*p-value*) fait l'objet de vifs débats dans la communauté scientifique et influence l'évaluation que les intervenants font des recherches dans leur domaine de pratique. Certains souhaitent la proscrire, alors que d'autres veulent continuer à l'utiliser. Cet article présente des éléments de réflexion concernant les écarts entre sa réelle signification et l'usage commun qui en est fait. Il discute des définitions ou interprétations erronées qui lui sont associées, notamment par rapport aux seuils de la valeur *p*. Enfin, il présente des alternatives ou compléments à cette statistique. Son objectif est d'identifier et de proposer aux chercheurs en sciences sociales, humaines et de la santé les meilleures pratiques de l'analyse des données quantitatives et de développer en même temps une réflexion critique à l'égard de leurs résultats basés sur cette statistique, laquelle continue à attiser bien des passions.

Mots-clés

Valeur *p* ; test d'hypothèse ; précision ; facteur de Bayes ; comparaison de modèles

Déclaration d'importance publique

Un grand nombre d'études en psychologie rapportent des valeurs *p*. Malheureusement, cette statistique est souvent mal comprise et mal interprétée. Ce manuscrit vise à éclairer le non-expert sur la nature de la valeur *p* en plus de présenter quelques alternatives et stratégies pour mieux l'utiliser dans les écrits scientifiques.

Introduction

La valeur *p* (*p-value*) occupe une position importante dans les représentations entourant la recherche quantitative, au point où elle fait l'objet de nombreux commentaires humoristiques. Une recherche réalisée au début du mois de décembre 2023, avec l'expression « p value » et « même », génère ainsi plus d'un million de résultats sur Google.com. À titre d'illustration, nous avons observé un photomontage de ce type sur la porte d'un local à l'Université de Montréal (Figure 1).

[Insérer la Figure 1, ici]

Cela n'est pas étonnant, car des études ont montré qu'une valeur *p* inférieure à un seuil de 0,05 augmente les chances qu'un manuscrit soit évalué positivement par les pairs et publié (Easterbrook et al., 1991 ; Koletsi et al., 2009). D'autres études montrent aussi qu'il recevra davantage d'attention et qu'il sera plus souvent cité par la communauté scientifique (Greenberg, 2009). Ce seuil a reçu un nom, le *seuil alpha* (*α*, « the first source of error » selon Neyman et Pearson, 1928, p. 177), et les résultats inférieurs à ce seuil sont considérés par un grand nombre de chercheurs comme étant *statistiquement significatifs* alors que, dans le même temps, les limites qui sont associées à ce concept ont amené certains spécialistes à se positionner à l'encontre de son utilisation (Wasserstein, Schirm, et Lazar, 2019).

À titre d'exemple, Ioannidis (2019) a examiné 62 résumés d'articles dans le domaine de la santé. Sur un total de 141 valeurs *p* répertoriées, 86% étaient inférieures à 0,05. Devrait-on en conclure que les valeurs supérieures à 0,05 refléteraient des résultats moins dignes de mention que leurs contreparties présentant des valeurs *p* inférieures à

0,05? C'est ce que pense Melton (1965), éditeur du *Journal of Experimental Psychology*, quand il a introduit la politique consistant à refuser les articles dont les valeurs *p* ne se logeaient pas sous ce seuil (Loftus, 1993 ; Lykken, 1991).

Les étudiants sont au courant de cet enjeu. Plusieurs d'entre eux ont sans doute aussi lu les appels à la prudence concernant la valeur *p,* par exemple, ceux d'Amrhein et al. (2019) dans *Nature*. De nombreux articles provenant de toutes les disciplines scientifiques et ayant comme objectif de clarifier cette statistique sont parus (Dixon, 2003 ; Goodman, 2008 ; Nahm, 2017 ; Hubbard et Lindsay, 2008 ; Lakens, 2021) dont celui émanant de l'*American Statistical Association* (Wasserstein et Lazar, 2016) est probablement l'un des plus… significatifs.

Le contenu de ce qui suit n'est pas nouveau et s'adresse aux chercheurs qui ne sont pas des experts en méthodes quantitatives. Bien qu'il existe de nombreux écrits critiques sur la valeur *p*, un grand nombre de chercheurs et d'intervenants continuent d'utiliser mécaniquement cette statistique comme nous pouvons le constater dans les écrits scientifiques et professionnels. L'objectif de ce texte consiste à clarifier ce qu'est la valeur *p* et son seuil d'interprétation alpha en présentant leurs forces et limites principales et quelques alternatives ou compléments intéressants. Nous nous insérons donc dans la mouvance faisant la promotion de meilleures pratiques dans l'analyse de données quantitatives en sciences sociales (voir Finkel et al., 2015, pour une discussion dans le contexte de la psychologie). L'idée est d'éviter d'utiliser ce que Sijtsma (2016) ainsi que Flake et Fried (2020) nomment les *pratiques de recherche discutables* (« *questionable research practices* »). L'article conclut sur des pistes à envisager en vue d'améliorer les pratiques en méthodologie de recherche et en interprétation des résultats de recherche.

Qu'est-ce que la valeur *p*?

Le test d'hypothèse enseigné dans le paradigme statistique dominant (souvent identifié par l'acronyme anglophone *NHST* pour « *null hypothesis significance testing* ») utilise la valeur *p* pour évaluer si l'hypothèse nulle (symbolisée par H_0) peut être rejetée ou non (Denis, 2004). Pour prendre cette décision, quatre étapes sont habituellement franchies.

Premièrement, H_0 doit être définie. Traditionnellement, elle caractérise une situation de *statu quo*, selon laquelle il ne se produit rien de systématique (par exemple, deux groupes expérimentaux ne diffèrent pas sur une variable dépendante ou deux variables ne sont pas corrélées entre elles). Cette hypothèse nulle décrivant un paramètre dans la population sert de référence pour tester à quel point les données observées concordent avec elle. Cela correspond, par exemple, à une différence de moyennes négligeable ou une corrélation presque nulle.

Afin d'illustrer ce test, imaginons que nous souhaitions comparer les résultats en mathématiques des élèves de deux écoles différentes en utilisant les scores obtenus à une même épreuve. Dans ce cas, H_0 caractérise une situation où les moyennes des deux populations, notées μ_1 et μ_2, ne diffèrent pas, ce qui peut s'écrire $H_0 : \mu_1 = \mu_2$, ou encore $H_0 : \mu_1 - \mu_2 = 0$. On cherche alors à savoir si cette hypothèse reste plausible alors que les moyennes observées de deux écoles seront vraisemblablement inégales. Autrement dit, la procédure vise à estimer la probabilité que la différence observée entre les moyennes soit le fruit d'une fluctuation d'échantillonnage et à décider qu'en dessous d'un certain seuil de probabilité (par exemple 5%), cette probabilité est trop faible pour qu'il soit plausible de l'attribuer à cette seule fluctuation.

Ceci étant dit, est-ce réaliste, plausible ou utile de spécifier une H_0 qui correspond à une différence exactement nulle entre deux groupes (certains auteurs anglophones parlent de « *nil hypothesis* »)? En effet, il est utopique de prédire la survenue d'une différence absolument nulle entre des groupes (Lykken, 1968, 1991). Par ailleurs, même si l'effet s'avère infime, il y aura toujours une taille d'échantillon suffisamment grande pour rendre cet effet significatif et rejeter l'H_0 (par exemple, Mitroff et Biggs, 2014, collectent un million de nouvelles observations par jour!). Pour cette raison Craver (1978) considère que « le test de signification statistique met en place un homme de paille, soit l'hypothèse nulle, et tente de le renverser [*knock down*] » (p. 381, traduction libre).

Deuxièmement, il est nécessaire de sélectionner un seuil de rejet nommé alpha (α). Ce seuil prend généralement la valeur de 0,05 en sciences sociales, mais parfois de 0,01 et, même, plus bas dans des domaines tels que la recherche en génétique et la physique ; nous y reviendrons plus loin. Puisque les statistiques inférentielles visent à évaluer la plausibilité d'un effet hypothétique dans la population, à partir d'échantillons observés, il est toujours possible de rejeter H_0 erronément à la suite de l'analyse des données. Ici, α représente un seuil définissant une probabilité d'erreur, l'erreur de type I, qui est établie *a priori* par le.la chercheur.se. C'est le seuil selon lequel H_0 n'est plus jugée plausible puisqu'il n'est pas raisonnable d'attribuer à la seule fluctuation d'échantillonnage la différence observée entre les échantillons. L'erreur de type I consiste donc à rejeter H_0 alors qu'il aurait fallu la conserver parce que la valeur *p* apparaît trop faible.

Troisièmement, il faut réaliser la collecte des données (la passation d'une épreuve en mathématiques dans notre exemple). Il faut pour cela idéalement choisir une méthode d'échantillonnage aléatoire, par exemple un échantillonnage par grappe (souvent utilisé dans les milieux scolaires où *des groupes d'étudiants* sont choisis aléatoirement, pas les élèves) ou encore un échantillonnage par strate (fréquent pour les sondages téléphoniques où l'on tente de préserver la représentativité des groupes d'âge, par exemple).

Quatrièmement, il faut choisir un test statistique pertinent. Dans notre cas, le test t pour échantillons indépendants est adéquat pour comparer les moyennes des scores provenant des élèves de deux groupes différents. La différence de moyennes observée est alors convertie sur une échelle standardisée qui correspond à une distribution théorique de référence nommée distribution t de Student. Celle-ci s'approche progressivement d'une distribution normale lorsque la taille de l'échantillon augmente.

Finalement, le test t permet de se prononcer sur le rejet ou non de H_0. Pour un test unidirectionnel, le rejet de H_0 survient seulement si le t calculé dépasse une valeur critique qui dépend du seuil α prédéfini. Par exemple, avec un seuil α de 0,05 et un n de 5 000 pour chacun des deux groupes (donc, un total de 10 000 élèves), cela correspond à la valeur critique t de 1,65. Si nous obtenons une valeur t observée, par exemple, de 2,20, la procédure commande de rejeter H_0 : on peut considérer que les moyennes des deux groupes d'étudiants sont différentes. Par contre, un t observé de 1,60 amène à conclure au « non-rejet de H_0 ». Cette phrase sibylline a généré énormément de confusion dès les débuts (Fisher le reconnait dès 1935, p. 16), au point que Howell (2010, p. 93) écrit « le problème de l'interprétation d'un non-rejet a tourmenté les étudiants de statistique depuis 75 ans » (traduction libre ; nous sommes maintenant presque rendus à

90 ans de tourments!). Il faut en retenir que l'absence de preuves n'est pas la preuve de l'absence.

Dans la pratique contemporaine où l'utilisation de logiciels est omniprésente, les chercheurs utilisent plutôt la valeur *p* pour décider du rejet ou non de H_0. Cette valeur indique la probabilité d'obtenir une différence de moyennes égale ou plus extrême que celle observée lorsque H_0 est présumée vraie ; cette valeur *p* correspond à une valeur du *t* pour une taille d'échantillon donnée. Elle permet de jauger directement, sur une échelle de probabilité, la plausibilité des données observées si l'on postule que H_0 est vraie, le chercheur pouvant la confronter ou non à une probabilité-seuil s'il le juge opportun. Krueger et Heck (2019) insistent sur le fait que cette statistique doit être écrite $p(D|H_0)$, soit la probabilité des données qui sont observées (D) postulant (le signe « | ») que l'hypothèse nulle H_0 est vraie dans la population d'où l'échantillon est extrait.

Qui es-tu, valeur p ?

Comme nous l'avons déjà mentionné, de nombreuses études ont discuté de l'idée que la valeur *p* est mal comprise (Cohen, 1994 ; Goodman, 2008 ; Greenland, et al., 2016 ; Haller et Krauss, 2002). Par exemple, plusieurs croient, à tort, que *p* représente la vraisemblance de H_0 étant donné les données observées (qu'on pourrait écrire $p(H_0 \mid D)$ dans la notation ci-dessus). Cette distinction est importante puisque la probabilité de A postulant B n'est pas égale à la probabilité de B postulant A. Ainsi, au Canada, si je vous présente une personne membre d'un ordre professionnel (A), la probabilité qu'elle soit psychologue en exercice (B) peut être faible, disons 4 sur 100, soit *p*(B|A). Mais si je vous présente une psychologue en exercice, la probabilité qu'elle soit membre d'un ordre

professionnel, soit p(A|B), est égale à 1 ou presque (si l'on exclut les personnes en situation d'infraction!).

Que sait-on, alors, sur cette statistique p? Demidenko (2016) montre que, toutes choses étant par ailleurs égales, la valeur p diminue lorsque la taille de l'échantillon s'accroît, lorsque l'effet qui est mesuré est de plus grande taille ou les deux. Pour illustrer cette idée, utilisons la différence de moyennes d'anxiété suivante qui est établie à partir de deux populations d'élèves : $\mu_1 - \mu_2 = 2{,}5$ et les écarts-types dans les deux populations qui sont égales : $\sigma_1 = \sigma_2 = 10$. Nous allons progressivement accroître la taille de l'échantillon, comme indiqué dans la colonne de gauche du Tableau 1.

[Insérer le Tableau 1 ici]

Même lorsque les moyennes et les écarts-types des groupes restent inchangés, le test t mène à une autre interprétation lorsque la taille de l'échantillon augmente : nous ne rejetons pas H_0, au seuil $\alpha = 0{,}05$, lorsque $n = 20$ et $n = 200$, alors que c'est le cas dès que n est égal ou supérieur à 250. Toutes choses étant égales par ailleurs, deux élèves de moins ($n = 248$) et nous aurions manqué cet effet. Cet exemple montre qu'adopter une posture catégorique, binaire, ne s'accorde pas avec la nature continue des phénomènes observés. Mentionnons également que la valeur p varie aussi en fonction d'autres variables telles que la taille d'effet[1], la variance et la forme de la distribution des données (Dahiru, 2008).

Une critique récurrente de la valeur p est que son interprétation dichotomique s'avère artificielle. Ainsi, pourquoi devrait-on se réjouir d'un $p = 0{,}049$ (ce qui mène à un rejet de H_0) et éprouver de la déception face à un $p = 0{,}051$? À l'évidence ces deux

[1] La taille de l'effet (*effect size*) mesure dans une population statistique la force de la relation entre deux ou plusieurs variables.

valeurs présentent des différences infimes. Une position rigide, tout-ou-rien, manque de nuances et pourrait nuire au développement des connaissances (McShane et Gal, 2016). Gelman et Stern (2006) rappelaient d'ailleurs qu'« une erreur statistique courante consiste à résumer les comparaisons par signification statistique, puis à établir une distinction nette entre les résultats significatifs et non significatifs » (p. 328, traduction libre) dans la discussion des résultats.

Pour reprendre un exemple en contexte scolaire, avec un échantillon total de 20 élèves, le chercheur mal avisé, décrit par Gelman et Stern (2006), écrira « après avoir collecté un échantillon par école, on trouve $p = 0{,}583$, ce qui montre une absence totale de différence entre les élèves des deux écoles ». Cette interprétation est erronée, car, d'une part, la valeur *p* ne permet pas de conclure à la véracité de H_0 (une absence de différence) et, d'autre part, la différence observée n'est pas totalement nulle. Le chercheur avisé fournira quant à lui plus de détails: « le premier échantillon présente une moyenne de 72,5 et le second, de 75,0, soit une différence de 2,5 points, avec un intervalle de confiance à 95% sur cette différence allant de -4,1 à 9,1 (écart-type regroupé $s_p = 10$). Cette différence est faible (*d* de Cohen de 0,25, intervalle de confiance à 95% allant de -0,63 à +1,13) et ne s'avère pas probante, $t(18) = 0{,}559$, $p = 0{,}583$. Obtenir ces résultats dans un contexte hypothétique où il n'y aurait pas de réelle différence entre les élèves des écoles est plausible ».

Enfin, mentionnons que Wasserstein et Lazar (2016) ont publié un article influent pour résumer la position de l'*American Statistical Association* sur ce qu'est la valeur *p* et ce qu'elle n'est pas. Nous reprenons intégralement ces six éléments : i) La valeur *p* indique à quel point les données sont compatibles avec un modèle statistique spécifique ;

ii) La valeur *p* ne mesure pas la probabilité qu'une hypothèse étudiée soit vraie ou la probabilité que des données soient produites seulement par la chance, soit une hypothèse non conditionnelle, $p(H_0)$; iii) Des conclusions scientifiques, d'affaires ou des décisions politiques ne devraient pas uniquement être basées sur la comparaison d'une valeur *p* avec un seuil prédéterminé ; iv) Une inférence adéquate requiert de rapporter toutes les informations sur le devis de recherche et la population sur laquelle l'on veut faire des inférences et de faire preuve de transparence sur les forces et les faiblesses de ce devis ; v) Une valeur *p* ne mesure pas la grandeur de l'effet ou l'importance dans un contexte appliqué d'un résultat ; et vi) La valeur *p*, à elle seule, ne présente pas un appui suffisant à l'égard d'un modèle ou d'une hypothèse. En statistique, une règle importante est que plus l'échantillon est grand, plus l'estimé sera précis. Or, en ce qui concerne la valeur *p*, plus l'échantillon est grand ($n = 500\,000$, voir plus), plus les résultats apparaîtront statistiquement significatifs pour un seuil donné.

Pourquoi un seuil de 0,05?

La valeur critique la plus communément utilisée est 0,05. Mais d'où vient ce seuil? Et pourquoi pas un autre seuil? Il semble que cette pratique se soit progressivement installée à la suite de la lecture d'un paragraphe qui provient de l'ouvrage phare du statisticien Fisher, *Statistical methods for research workers*, initialement publié en 1925. Nous reproduisons ici l'extrait traduit de ce paragraphe, où Fisher indiquait au sujet de l'interprétation du résultat à un test du khi carré:

> « Ce qui nous intéresse, c'est savoir si la valeur *p* remet une hypothèse en cause. Si *p* se situe entre 0,1 et 0,9, il n'y a pas de raison de remettre en cause l'hypothèse nulle. Si la valeur *p* est inférieure à 0,02, il y a alors de fortes raisons de croire que l'hypothèse nulle ne parvient pas à expliquer ce qui est observé. Ainsi, nous nous

> égarerons rarement si nous traçons un seuil à 0,05 » (Fisher, 1934, p. 82, traduction libre).

Le seuil critique de 0,05 semble s'être tranquillement installé dans les us et coutumes de la communauté scientifique en séparant cette idée du contexte de réflexion plus général proposé par Fisher.

Certains se sont demandé s'il fallait utiliser un seuil plus restrictif tel que 0,005 (e.g., Benjamin et al., 2017). La réponse a été assez rapide (e.g., Trafimov et al., 2018 ; Lakens, 2021): le problème émane de la façon dont la valeur *p* est conceptualisée plutôt que du seuil en soi. Pour illustrer ce point, nous vous suggérons de faire le test de la tache de café, à la Figure 2, avant de poursuivre votre lecture.

[Insérer la Figure 2, ici]

Dans le graphique de gauche, la valeur *p* est de 0,051. Selon le seuil alpha de 0,05, c'est un résultat non significatif, et Raphaëlle est déçue d'avoir obtenu ce résultat. Dans le graphique de droite, la valeur *p* est de 0,049, et la différence est significative. Or, bien malin est celui qui peut voir la différence entre les deux graphiques tellement ils sont quasiment identiques (pour le graphique de gauche, les moyennes sont de 77,00 et 71,77, pour un *d* de Cohen de 0,395, IC95% [-0,002, 0,789] ; pour le graphique de droite, les moyennes sont de 77,56 et 71,21, pour un *d* de Cohen de 0,397, IC95% [0,002, 0,792]). Comme le mentionnent les chercheurs étasuniens Rosnow et Rosenthal (1989) « Dieu aime presque autant le seuil 0,06 qu'il aime le seuil 0,05 » (p. 1277, traduction libre). Cette boutade montre qu'un seuil basé sur une valeur *p* rigide trace une distinction binaire entre des résultats dont la différence est presque totalement indiscernable. Un seuil basé sur une lecture critique de la valeur *p* permet de discriminer et de souligner un résultat vraisemblablement exceptionnel, surtout lorsqu'elle est appuyée par une méta-analyse ou

des reproductions qui ajoutent un poids considérable à la valeur des inférences statistiques.

Fisher, qui est le premier à avoir systématisé le calcul de la valeur *p* dans ses écrits des années 1910-1930, arrondit cette valeur à des ordres de grandeur faciles à appréhender et permettant de produire des tableaux statistiques dans les publications: 1 sur 10 (soit 0,10), 1 sur 20 (soit 0,05) ou encore 1 sur 50 (soit 0,02). Pour Fisher, ce sont des raccourcis simples à comprendre pour indiquer à quel point une valeur *p* obtenue est faible ou non (Lehman, 2011). Par ailleurs, il n'utilise pas ce résultat de façon dichotomique, mais plutôt afin de graduer le résultat des analyses. Pour lui, la valeur *p* est un *indice* de la force des évidences contre H_0 (bien que formellement, ce soit la probabilité des données quand on postule H_0 ; Berger, 2003). Nuzzo (2014) explique qu'aux yeux de Fisher, une valeur *p* représente une façon informelle de juger de l'importance du résultat: est-ce qu'il mérite qu'on s'y intéresse un peu plus? Autrement dit, une petite valeur *p* serait un drapeau qui s'agite pour attirer l'attention des chercheurs et des utilisateurs des recherches et non une fin en soi.

Suivant cette logique, nous mentionnons Rafi et Greenland (2020), qui suggèrent de transformer la valeur *p* en « indice de surprenance des données » avec cette formule :

$$s = -\log_2 p \quad (1)$$

où *s* indique à quel point les données sont surprenantes dans le cadre de H_0. Par exemple, lorsque *s* vaut 1, les données sont aussi surprenantes que d'obtenir pile en lançant une pièce de monnaie une fois (un cas pas du tout surprenant, d'où une valeur *p* de 0,50) ; lorsque *s* vaut 10, le résultat est aussi surprenant que d'obtenir 10 piles après 10 lancers (c.-à-d., très surprenant, et *p* vaut $0{,}50^{10}$, soit ≈0,001). Pour un *p* de 0,05, l'indice de

surprenance de 4,3 correspond approximativement à la probabilité d'obtenir quatre piles sur quatre lancers (en effet, cette dernière probabilité est de 6,25%, soit proche de 5%). Si pareil résultat n'est pas exceptionnellement fréquent, il n'est pas non plus exceptionnellement rare. La valeur *s* ne remplacera pas la valeur *p* puisque les deux sont interchangeables, mais *s* a l'avantage de possiblement rendre une valeur *p* plus intuitive.

Significatif… mais instable ?

Imaginez la situation suivante: vous avez réalisé une étude où le résultat d'intérêt s'avère « significatif » (disons que $p = 0,04$) et votre réaction ressemble à celle illustrée à la Figure 1. Or, quelques minutes plus tard, un de vos participants vous contacte et, après réflexion, il préfère que vous n'utilisiez pas ses réponses dans votre recherche. Vous avez bien entendu obtenu un certificat d'éthique, et donc, la démarche est claire: vous effacez les données liées à ce participant et recrutez un nouveau participant pour le remplacer. La nouvelle valeur *p* sera-t-elle identique? Il y a très peu de chance que cela soit le cas, sauf si le participant de remplacement obtenait exactement le même score que le participant remplacé. Une simulation qui se veut une simple illustration sans prétention de généralisabilité montre qu'avec un seul changement dans un groupe de 50 personnes, la nouvelle valeur *p* obtenue pourra être aussi petite que 0,02 (Figure 1), mais aussi grande que 0,08 (Figure 3).

[Insérer la Figure 3, ici]

Murdoch et al. (2008) nous rappellent que la valeur *p* est de source aléatoire: elle dépend des données et si les données changent, la valeur *p* change elle aussi.[2]

Quelques stratégies pour embrasser l'ère « post $p < 0,05$ »

En plus d'être mal comprise, la valeur *p* ne serait pas sans faille (Goodman, 2008). Ce constat avait déjà été établi dès 1933 par Neyman et Pearson. Certains suggèrent de rejeter cette statistique (Hurlbert et Lombardi, 2009, qui pointent entre autres la dichotomisation arbitraire des résultats en « significatifs et « non-significatifs » et la perception erronée que *p* concerne la probabilité relative $\frac{p(\mathrm{H0})}{p(\mathrm{H1})}$. Ce rejet de *p* a été adopté par le journal *Basic and Applied Social Psychology* où les valeurs *p* sont proscrites depuis 2015 (Trafimow et Marks, 2015). La valeur *p* nous semble toujours pertinente. Nous comprenons que, pour les chercheurs et les chercheuses pour qui la valeur *p* est habituelle dans l'usage des procédures statistiques, il puisse être difficile d'envisager de ne plus l'utiliser. Une alternative à la démarche radicale appliquée par le journal *Basic and Applied Social Psychology* consisterait à inciter les chercheurs et les chercheuses à mieux comprendre ce qu'elle signifie et d'exiger que les résultats soient complémentés avec d'autres informations statistiques. Dans ce qui suit, nous allons présenter quelques stratégies d'intérêt à ce sujet.

Stratégies centrées sur la valeur p et le seuil α

[2] Notons que cette critique s'applique à tous les types de mesure, pas seulement la valeur *p*.

Est-ce vraiment utile de formuler une H_0 où les moyennes sont exactement égales, comme on le voit dans un grand nombre d'études comparatives? En réalité, la population des élèves de deux écoles n'a jamais exactement la même moyenne à un examen ; la question est plutôt de savoir quelle sera l'ampleur des différences observées et avec quels phénomènes ces différences sont associées. Il est douteux qu'il existe, dans nos disciplines, une situation dans laquelle une H_0 de zéro serait plausible. Or, puisque la valeur p est une probabilité conditionnelle à H_0 (donc $p(D|H_0)$) et que H_0 est très peu plausible par principe, que vaut la valeur p? Est-elle même interprétable? Plusieurs auteurs pensent que c'est une raison importante pour la laisser de côté, par exemple Hurlbert et Lombardi (2009), Cumming (2014), ainsi que McShane et al. (2019).

Cependant, les habitudes d'utilisation d'une statistique populaire sont souvent bien ancrées dans la pratique et, jusqu'à présent, les résistances individuelles et institutionnelles n'ont pas permis de concrétiser cet objectif.[3] Goodman (2019) mentionnait à ce sujet:

> « L'explication la plus simple relativement à l'utilisation de la valeur p n'est ni philosophique ni scientifique, mais bien sociologique : tout le monde l'utilise et tout le monde croit en sa valeur et c'est cette valeur qui détermine ce que constitue le savoir, et ce qui peut être avancé, publié, financé et promu. Peu importe la réelle signification de la valeur p, c'est elle qui fixe la valeur de ce qui est écrit » (p. 27, traduction libre).

Dans le contexte de la crise de reproductibilité (Pashler et Wagenmakers, 2012), certains auteurs suggèrent de réduire le seuil α. Par exemple, Benjamin et al. (2017),

[3] Fait intéressant, le même genre de remarque est apparue concernant le coefficient alpha de Cronbach, dans la communauté en psychométrie. Par exemple, Sijtsma (2009) écrivait : « la seule et unique raison de rapporter l'alpha est que les meilleures revues ont tendance à accepter les articles qui utilisent des méthodes statistiques qui existent depuis longtemps, telles que l'alpha » (p. 118, traduction libre).

proposent de le réduire à 0,005 pour rejeter H_0. L'impact de cette modification revient à baisser l'occurrence d'erreurs de type I (faux positifs), ce qui est l'effet recherché par ces auteurs. Néanmoins, ceci résulterait en une augmentation des erreurs de type II (faux négatifs), notée β. Ioannidis (2018) est toutefois très sceptique avec cette proposition. Nous estimons que ce type de solution mécanique comprend le même vice que l'heuristique habituelle: elle ne peut remplacer l'exercice de la pensée critique et mène à des pratiques fétichistes de la statistique plus qu'à un raisonnement nuancé et rigoureux concernant les données. Or, par l'effet de l'argument d'autorité (« c'est publié, donc cela doit être pertinent »), cette pratique influence également de nombreux utilisateurs des connaissances, allant des psychologues au grand public. De plus, utiliser un seuil plus restrictif tel que 0,005, toutes autres choses étant égales par ailleurs, se traduit par une perte de puissance statistique. Ceci peut impacter les études dont les tailles des échantillons sont plus faibles, par exemple dans le cas des populations difficiles à étudier ou pour des maladies ou troubles mentaux plus rares. Si les grands échantillons sont souhaitables, car ils génèrent une estimation plus précise des paramètres qui intéressent le chercheur, la qualité méthodologique d'une étude ne doit toutefois pas être réduite à ce seul enjeu.

Alternativement, plutôt que manipuler le seuil α, certains suggèrent de manipuler l'hypothèse[4]. Au lieu de tester une H_0, pourquoi ne pas tester l'hypothèse de la présence d'un effet d'intérêt substantiel? Les tests de non-équivalence ont pour but de répondre à ce genre de question (Schuirmann, 1987). La difficulté de cette approche est cependant de définir ce que représente un « effet intéressant ». Cette définition doit être établie avant l'étude pour éviter d'être biaisée par l'observation des résultats, et doit revêtir un

[4] Nous remercions un évaluateur anonyme de nous avoir rappelé cette alternative.

caractère consensuel. Au final, il apparaît toutefois que la valeur *p* de tels tests ne s'avère pas plus aisée d'interprétation.

D'autres auteurs déclarent qu'il faut devenir des utilisateurs de statistique avisés, tout spécialement concernant la valeur *p* (Wasserstein et Lazar, 2016). Ainsi, laisser les logiciels définir des valeurs arbitraires par défaut (par exemple, $\alpha = 0{,}05$, $\alpha = 0{,}01$, ou encore $\alpha = 0{,}005$) constitue un automatisme qui peut engendrer des conclusions erronées ou dénuées de mise en contexte.

Dans l'optique de rendre les chercheurs plus responsables, Lakens (2022) invite à garder à l'esprit quels sont les buts et les attentes des chercheurs (en particulier, quels sont les effets auxquels ils s'attendent, quels sont les effets qu'ils voudraient rejeter, quels sont les effets qu'ils trouveraient intéressants d'observer). Lakens (2022) suggère même qu'il serait acceptable d'accroître α si cela peut mieux servir les buts de l'étude et si ce choix est adéquatement argumenté.

Aguinis et al. (2010), de leur côté, suggèrent de contraindre le choix du α afin de mieux distinguer les erreurs de type I et les erreurs de type II acceptables dans un projet de recherche. Pour cela, ils proposent de calculer le niveau α avec une formule qu'ils nomment $\alpha_{Désiré}$, de la façon suivante:

$$\alpha_{Désiré} = \left[\frac{p(H_1)\beta}{1-p(H_1)}\right]\left(\frac{1}{DRS}\right) \quad (2)$$

où DRS est le « desired relative seriousness » qui représente le danger de commettre une erreur de type I par rapport au danger de commettre une erreur de type II, $p(H_1)$ est la probabilité que l'hypothèse alternative soit plausible (basé sur des connaissances obtenues *a priori*) et β est l'erreur de type II. La valeur critique $\alpha_{Désiré}$ est ainsi mieux justifiée et moins arbitraire qu'une valeur α « par défaut ». Supposons que nous fixons

$p(H_1) = 0{,}6$ en nous basant sur des études antérieures, avec $\beta = 0{,}15$ et $DRS = 2$, le $\alpha_{Désiré}$ serait alors égal à 0,11. Cela veut dire que 0,11 est l'erreur de type I jugée acceptable sélectionnée avant d'accomplir les analyses. En se référant aux résultats de la Figure 2, les deux tests rejetteraient H_0. Plus le DRS sera élevé, plus l'accent sera mis sur l'importance d'éviter une erreur de type I.

Quant à eux, Blume et al. (2019) proposent l'idée d'une valeur *p* de deuxième génération, notée p_δ. Lorsque H_0 n'est pas contenue dans l'intervalle de confiance de 95%, nous savons que la valeur *p* est plus petite que 0,05. Or, baser H_0 sur une seule valeur est artificiel et inintéressant en réalité. Voilà pourquoi ces auteurs utilisent un intervalle de valeurs défini par le chercheur (« zone d'indifférence » aux résultats), plutôt qu'une seule valeur, pour opérationnaliser H_0. Cela permet de mieux formaliser l'information du contexte d'analyse en calculant la proportion de l'intervalle de confiance qui chevauche (ou pas) l'intervalle de valeurs de H_0. Le lecteur curieux pourra en apprendre davantage sur cette approche en consultant l'interface visuelle suivante : https://lucy.shinyapps.io/sgpvalue/

Stratégie centrée sur les tailles de l'effet et leur précision

Des auteurs tel que Ranstam (2012) et Cumming (2014) proposent plus simplement de rapporter systématiquement les tailles d'effet (*effect size*) et leurs intervalles de confiance (IC). La taille de l'effet représente une estimation de la taille de la différence entre deux groupes, par exemple, qui n'est pas biaisée par la taille de l'échantillon comme la valeur *p* peut l'être. Comme le rappelle Glass, cité dans Sullivan et Feinn (2012, p. 279) : « la signification statistique est la partie la moins intéressante

des résultats d'une étude. C'est donc elle qui devrait être décrite et non sa simple présence/non présence. Autrement dit, nous ne devrions pas nous demander si un traitement génère un effet sur des individus, mais à quel point ce traitement présente un effet sur ceux-ci » (traduction libre).

Les tailles d'effets sont de plusieurs sortes (voir le Tableau 2). Certaines sont « brutes », c'est-à-dire tirées directement des données, comme la différence entre deux moyennes. D'autres sont standardisées et sont plus abstraites, comme le *d* de Cohen, le η-partiel et la corrélation, mais ont l'avantage d'être interprétables sur une échelle universelle (XXXXX). Par exemple, la valeur du *d* de Cohen est de 0,395, IC95% [-0,002, 0,790] pour la comparaison à gauche de la Figure 2 et de 0,397, IC95% [0,001, 0,792] pour la comparaison qui est à sa droite.

[Insérer le Tableau 2, ici]

Pour chaque statistique observée, il est important d'estimer sa précision en lui assignant un intervalle de confiance (e.g., XXXXX). Rappelons que l'intervalle de confiance représente une mesure de la précision d'une statistique (par exemple, l'intervalle de confiance de la moyenne ou l'intervalle de confiance de la médiane). Si une étude est conduite à plusieurs reprises et de la même façon, nous pouvons estimer pour chaque échantillon tiré au hasard, une moyenne et un intervalle de confiance à 95%. L'interprétation de l'intervalle de confiance à 95% indique que 95% de ces intervalles de confiance contiennent la vraie moyenne de la population (le paramètre). Quand un sondeur rapporte la moyenne des votes qui irait à un parti politique en vue d'une élection, omettre l'intervalle de confiance revient à ignorer la précision du sondage (ou indiquer subrepticement que l'échantillonnage était non probabiliste), ce dont le lecteur devrait

être informé directement. De plus, il est intéressant de constater que, contrairement à la valeur *p* qui n'offre pratiquement aucune information utile dans une perspective de réplication des résultats d'une étude, un intervalle de confiance à 95% pour une moyenne représente une probabilité de 83% que la moyenne de l'étude répliquée dans des conditions similaires se trouve entre le seuil inférieur et le seuil supérieur de cet intervalle (Cumming, 2008). Enfin, l'intervalle de confiance permet de s'extraire d'une vision dichotomique des résultats d'une procédure statistique. À ce sujet, l'intervalle de confiance ne doit pas être perçu comme une simple duplication de la valeur *p* (dans le sens où l'on peut se demander si l'intervalle inclut une différence de moyenne nulle), mais bien comme un intervalle dont les bornes inférieures et supérieures apportent une information pertinente pour développer une pensée méta-analytique. Par exemple, en agrégeant les résultats de différentes études au sein d'un domaine de recherche, les intervalles de confiance offrent une idée de l'ampleur des effets observés. Ces informations permettent donc de réfléchir à la portée appliquée et théorique des résultats, plus qu'en termes de présence ou d'absence d'effets « statistiquement significatifs » (Dienes, 2008).

Les tailles d'effet les moins abstraites sont sans doute celles qu'il faut recommander le plus, comme les tailles d'effets brutes. Baguley (2009) avance deux éléments pour soutenir cette proposition. Tout d'abord, bien que la standardisation permette la comparaison de mesures provenant d'échelles différentes, elle entraîne plusieurs enjeux, entre autres un manque de robustesse (ex. : différentes manières de calculer un même indice, manque de fidélité en raison des fluctuations de la variance dans l'échantillon). Ensuite, l'utilisation d'indices faisant sens sur les échelles de mesure

d'origine facilite le jugement critique concernant l'interprétation et la portée des résultats, alors que la standardisation entraîne un niveau d'abstraction qui peut complexifier l'exercice d'un jugement critique sur les résultats. Cependant, Borenstein et coll. (2016, p. 25) rappellent à juste titre que les mesures doivent être toutes sur la même échelle, sinon, la standardisation est la seule façon de comparer ces mesures. *A contrario*, l'utilisation de tailles d'effets standardisées ouvre la porte à l'invocation rituelle de seuils arbitraires permettant de démarquer des tailles d'effet petites, moyennes et grandes pour le *d* de Cohen ou η^2 sans considérer si ces seuils sont porteurs d'un sens par rapport à un domaine de recherche spécifique ou par rapport à des résultats publiés antérieurement. Cela représente finalement un enjeu similaire à l'utilisation dichotomique de la valeur *p*. Afin que ces seuils ne soient pas arbitraires, Gignac et Szodorai (2016) réalisent une vaste méta étude afin d'identifier les quartiles des corrélations généralement rapportées en psychologies. Ces quartiles deviennent alors les frontières pour définir les tailles d'effets faibles et larges. Ils trouvent ,1 en deçà duquel une corrélation peut être considérée faible, et ,3 au-delà duquel une corrélation peut être considérée forte. Ce dernier nombre est bien inférieur au ,5 suggéré par Cohen (1992).

Stratégies centrées sur la comparaison de modèles

Le processus statistique expliqué jusqu'ici se base sur H_0 et son rejet (ou pas), en faveur de H_1. Selon cette perspective, on ne peut pas soutenir H_0 avec des données, mais seulement la rejeter (ou ne pas être en mesure de la rejeter). Pour le dire autrement, on ne peut pas, avec les tests de H_0, soutenir l'absence de différence entre deux conditions

pertinentes, même s'il est parfois scientifiquement pertinent de montrer que deux interventions sont équivalentes.

En guise d'illustration, reprenons l'exemple des écoles et imaginons que la seconde école a reçu un enseignement qui est présumé plus efficace que celui de la première école (qui sert de point de référence). Après réflexion par rapport aux connaissances sur les effets de taille minimalement pertinente associées à la mesure, il apparaît que dans la seconde école, les scores devraient augmenter minimalement de 5 points en moyenne pour que le résultat soit jugé utile. Il serait alors intéressant de tester cela en formulant H_0: $\mu_1 - \mu_2 = 5$. Sur le site OSF https://osf.io/prab2/ se trouve un script R qui montre comment il est aisé de tester une hypothèse nulle différente de zéro. Ceci revient toutefois à soumettre à l'épreuve des données uniquement cette H_0, ce qui peut entraîner les inconvénients mentionnés au paragraphe précédent lorsqu'elle est rejetée.

Comme autre stratégie, nous pourrions aussi opposer H_0: $\mu_1 - \mu_2 \leq 0$ à l'hypothèse telle que H_1: $\mu_2 - \mu_1 \geq 5$, c'est-à-dire que nous envisagerions dans ce cas que la seconde classe aura 5 points de plus en moyenne que la première. Ainsi, en plus de tester des hypothèses nulles non nulles, il est également possible de tester des hypothèses de non-nullité (les tests d'équivalence).

Dans la suite de cette section, un duo d'hypothèses sert à représenter des modèles permettant d'utiliser des approches basées sur le ratio de vraisemblance ou le facteur de Bayes. Wasserstein et Lazar (2016) écrivent que certaines personnes préfèrent remplacer les valeurs *p* par d'autres approches « comme des méthodes qui mettent en avant l'estimation plutôt que les tests, tels les intervalles de confiance, les intervalles de crédibilité ou les intervalles de prédiction, des méthodes bayésiennes ; des ratios de

vraisemblance ou des facteurs de Bayes, et d'autres approches comme la théorie de la décision et des faux positifs » (p. 132, traduction libre).

Le ratio de vraisemblances RV constitue le rapport entre le maximum de vraisemblance des données avec le modèle correspondant à l'hypothèse 1 donnée ci-dessus et le maximum de vraisemblance des données avec le modèle correspondant à l'hypothèse 0, que l'on exprime comme suit dans le cadre de la comparaison entre H_0 et H_1:

$$RV = \frac{p_{MV}(D|H_1)}{p_{MV}(D|H_0)} = \frac{Maximum\ de\ vraisemblance\ de\ H_1}{Maximum\ de\ vraisemblance\ de\ H_0} \quad (3)$$

La vraisemblance des données indique à quel point des données telles que celles observées auraient pu se produire lorsque le modèle est postulé. Sans surprise, un bon modèle est celui qui suggère que les données observées s'avèrent probables. Par extension, la vraisemblance des données devient la vraisemblance du modèle une fois les données obtenues. Cette statistique présente l'avantage d'être intuitive à interpréter. Ainsi, un $RV = 5$ indiquera que H_1 est cinq fois mieux soutenue par les données que H_0. À l'inverse, un RV de 0,2 (1 sur 5) indiquera que l'hypothèse nulle est cinq fois mieux soutenue par les données que l'hypothèse alternative (ce ratio est donc réversible). Évidemment, ce ratio ne fournit aucune information sur la comparaison avec n'importe quel autre modèle qui n'a pas été estimé.

Le facteur de Bayes (FB)[5] a aussi été proposé pour comparer des hypothèses. Il a en outre l'avantage de contourner certaines limites de l'approche fréquentiste que Cohen

[5] Nous ne prétendons pas offrir une quelconque introduction aux modèles bayésiens. Notre objectif vise plutôt à mentionner l'existence du FB comme alternative (ou complément) et de donner des éléments de compréhension minimaux à son sujet.

(1994) discute de manière particulièrement éloquente (c'est pourquoi nous laissons le passage en anglais) :

> "What's wrong with NHST? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is "Given these data, what is the probability that H_0 is true?" But as most of us know, what it tells us is "Given that H_0 is true, what is the probability of these (or more extreme) data?" These are not the same, as has been pointed out many times over the years…" (1994, p. 997)

En réalité, ce qui intéresse le chercheur est surtout de connaître la probabilité d'un modèle (ou d'une hypothèse) à partir des données qui ont été analysées. Et c'est une information qu'il est possible d'obtenir avec les statistiques bayésiennes, par exemple $p(H_0|D)$.

Mathématiquement, cette statistique prend la forme suivante :

$$FB = \frac{p_{FB}(D|H_1)}{p_{FB}(D|H_0)} = \frac{Vraisemblance\ marginale\ de\ H_1}{Vraisemblance\ marginale\ de\ H_0}. \quad (4)$$

Le facteur de Bayes implique aussi un rapport de vraisemblance (similaire à celui de l'équation 3, mais dans certains cas seulement) auquel on ajoute l'information *a priori* que nous avons sur chacun des modèles, laquelle est *souvent* de l'information qui doit être formalisée (par exemple, des résultats provenant d'études précédentes ou certaines connaissances pertinentes du chercheur). Cette information est de nature subjective et le FB indique comment ce chercheur devrait réviser ces connaissances.

Des barèmes ont été proposés pour interpréter le facteur de Bayes, dont le plus connu a été fourni par Kass et Raftery (1995). Ici, nous allons plutôt rapporter celui de Kelter (2020) qui nous semble plus nuancé :

[Insérer la Figure 4, ici]

Par exemple, un FB égal à 20 montre que les données sont 20 fois mieux représentées par H_1 que par H_0 lorsque les distributions *a priori* sont pris en compte. Les données observées corroborent de manière probante l'hypothèse H_1. Un FB de 0,05 offrirait, à l'opposé, un élément de preuve relativement fort en faveur de H_0.

Lorsque les moyennes présentées à la Figure 2 sont analysées à l'aide du test *t* bayésien, nous observons un FB égal à 1,173, donc faiblement en faveur de H_1. Ce résultat s'avère toutefois anecdotique selon le barème de la Figure 4. Un FB de 1,195, également en faveur de H_1, est calculé à la droite de cette même figure. Encore une fois, ce résultat est considéré comme anecdotique. Pour ces analyses, la distribution Cauchy est utilisée en guise de distribution *a priori* (il serait possible d'envisager une autre loi de distribution, au besoin). Notez que le facteur de Bayes est de plus en plus intégré dans les logiciels, dont R, JASP (https://jasp-stats.org/) et SPSS, parmi d'autres.

Il est possible d'estimer un facteur de Bayes à partir d'une valeur *p*. Benjamin et Berger (2019) ainsi que Held et Ott (2018) proposaient d'utiliser une façon de calculer la limite supérieure du facteur de Bayes (notée « LFB »), ce qu'ils ont appelé le « Bayes factor bound ». Selon Benjamin et Berger (2019), le *LFB* :

> « indique l'inférence la plus forte qui puisse être effectuée à partir des données. La rapporter pourrait informer les chercheurs des cas où une preuve apparemment forte (par exemple, une valeur *p*) n'est, en fait, pas très convaincante; ce qui éviterait à des chercheurs de conclure trop

> rapidement, et à tort, sur la base d'une valeur *p* » (p. 188, traduction libre).

Le tableau qui suit présente des exemples de valeurs *p* et leur équivalent *LFB*.

[Insérer le Tableau 3, ici]

Par exemple, pour une valeur *p* de 0,05, *LFB* vaut 2,44, donc en faveur de H_1 et lorsque $p = 0{,}01$, *LFB* vaut 8,13, également en faveur de H_1. Le premier reste anecdotique ; le second commence à être plus convaincant. Le *LFB* représente ainsi la valeur la plus élevée du *FB* qui soit en adéquation avec une valeur *p* observée. Selon cette stratégie, les valeurs *p* obtenues à la Figure 2, soit 0,051 et 0,049, génèrent une valeur $LFB = 2{,}44$ en faveur de H_1.

Quelques pistes pour améliorer les pratiques d'analyse

Nous suggérons quatre pistes pour entrer dans l'ère « post $p < 0{,}05$ » qui nous semblent réalistes, afin d'aider les chercheurs à accomplir cette transition.

La simplicité des statistiques descriptives pour « sonder le terrain »

Un article scientifique raconte une histoire. Les statistiques descriptives et des graphiques pertinents (Tufte, 2001) constituent la première étape nécessaire pour illustrer l'effet qui nous intéresse. Ces informations amènent à réfléchir sur la crédibilité des résultats par rapport aux attentes du chercheur et des connaissances antérieures sur le sujet. Pour le dire autrement, les statistiques descriptives poussent à une réflexion conceptuelle plutôt qu'à l'application mécanique d'heuristiques ou de conventions.

Dans la majorité des cas, la valeur p n'est pas suffisante

Betensky (2019) rappelle que la valeur *p* nécessite un contexte plutôt qu'un seuil α, qui simplifie à outrance la prise de décision lors d'un test d'hypothèse. Ainsi, il faut plutôt contextualiser la valeur *p* en fonction de l'échantillon et de la valeur de la taille de l'effet minimalement pertinente (par exemple, le *d* de Cohen lors d'une comparaison entre deux moyennes), soit celle qui a du sens pour le chercheur ou pour les personnes visées par une intervention. Comme le titraient Gelman et Stern, dans leur article de 2006 : « The Difference Between "Significant" and "Not Significant" is not itself statistically significant » (voir aussi le test de la tache de café, en guise d'illustration). Nous mettons l'accent sur le fait que la valeur *p* ne s'avère pas, à elle seule, suffisante pour conclure à la supériorité d'une hypothèse sur une autre.

Le chercheur doit avoir dans sa boîte à outils un faisceau de stratégies, dont plusieurs sont présentées dans cet article. Ainsi, le rapport de vraisemblance et le facteur de Bayes offrent une interprétation plus nuancée qu'un seul et unique seuil α. Mais il y a plus : il est important de contextualiser la statistique utilisée par le chercheur dans une perspective de décision pertinente plutôt que comme une conclusion scientifique isolée. Pour le dire autrement, est-ce que ce qui a été trouvé est utile de façon appliquée? Et de plus, comme « le doute est l'oreiller du savant » (Héger, 1895, p. 169), et ajouterions-nous, également celui de tout bon lectorat scientifique, quels sont les aspects méthodologiques qui viennent nuancer les résultats obtenus et qui doivent être transmis au lectorat pour lui permettre de juger des résultats dans leur contexte ? Il faut accepter de se mettre « à nu »

méthodologiquement pour offrir une inspection détaillée de la démarche scientifique dont les résultats découlent.

Intégrer systématiquement des mesures de précision dans les écrits qui relèvent d'une approche quantitative

Une carence méthodologique revient souvent dans les sondages rapportés dans certains journaux : l'absence d'intervalle de confiance fréquentiste ou d'intervalle de crédibilité bayésien. Or, la mesure de la précision est fondamentale pour bien contextualiser une statistique telle qu'une moyenne. De plus, cette information est nécessaire lorsque vient le temps de comparer des hypothèses sur des paramètres statistiques.

Adopter un langage plus nuancé

Plusieurs auteurs ont critiqué l'utilisation de la terminologie « statistiquement significatif ». En effet, « $p < 0{,}05$ » a parfois été considéré à tort comme équivalent de « Découverte scientifique ». Or, au moment où l'un d'entre nous écrit ces lignes, les hirondelles volent une nouvelle fois plus bas qu'à l'habitude et l'orage gronde. Après de multiples observations, cette relation apparaît statistiquement significative ($p < 0{,}001$), mais faut-il pour autant en faire une théorie scientifique qui suggère que les hirondelles volent plus bas pour se protéger de l'orage? La valeur p ne permet aucunement de suggérer ceci. Un biologiste affirmera plutôt que c'est la pression d'une masse d'air qui

influence la position des insectes et donc la position du garde-manger des hirondelles… (Cabaret, 2014).

Benjamin et Berger (2019) suggèrent d'utiliser le qualificatif « suggestif » plutôt que « significatif » ; en français, il est possible d'utiliser « notable » pour un grand effet (dans le sens *il vaut la peine de le noter*) et « cohérent » (c.-à-d., *cohérent avec cette hypothèse*). Nous abondons dans le même sens en estimant qu'il faut adopter une écriture prudente au moment de rapporter des résultats découlant de comparaisons d'hypothèses ou de modèles. La « significativité » statistique peut constituer un guide approximatif sur ce qu'il faut rapporter dans le texte, sur là où il faut mettre l'accent, mais n'est certainement pas un guide pour développer une théorie.

Dans un contexte bayésien, lorsqu'un résultat incohérent apparaît (incompatible avec une hypothèse), cette hypothèse perd en crédibilité au détriment d'une autre qui est plus crédible. Contrairement à la perspective poppérienne qui présume que la « vérité » scientifique peut à tout moment être réfutée, ici une seule réfutation (basée sur un critère statistique) ne constitue pas un poids suffisant pour rejeter une théorie. Une réfutation basée sur un critère statistique représente plutôt une invitation à identifier une explication alternative.

Enfin, nous considérons que les analyses statistiques ne remplacent pas un bon devis de recherche et une approche méthodologique rigoureuse. Ainsi, une avenue qui semble prometteuse est le pré-enregistrement des devis de recherche, soit faire évaluer par les pairs la question de recherche et la méthodologie avant de connaître les résultats d'une étude. En effet, une bonne méthodologie peut se passer de statistiques, mais les statistiques ne peuvent jamais se passer d'une bonne méthodologie.

Conclusion

Cet article semble ouvrir un débat essentiellement actuel, mais il n'en est rien : des critiques à l'égard de la valeur *p* sont publiées depuis plus de 90 ans. Il est maintenant temps d'agir et d'améliorer les pratiques d'analyse afin que les décideurs et la population puissent avoir accès à une science de plus grande qualité.

La place des analyses statistiques en sciences sociales, humaines et en santé fait actuellement l'objet de débats. N'oublions pas que ces disciplines sont critiquées par plusieurs et que certaines universités ont même fermé des départements en sciences sociales, comme cela a été le cas au Japon il y a quelques années (Grove, 2015). Nous avons remarqué que les personnes qui ne sont pas au courant des récentes avancées entretiennent des idées stéréotypées et souvent dépassées sur la façon de bien analyser des données. Pire, elles commettent des erreurs d'analyse, ce qui nuit grandement à la réputation de la science dans nos disciplines. Il n'y a pas de doutes pour nous : il est grand temps d'entrer de plain-pied dans l'ère « post $p < 0{,}05$ »!

Remerciement

Nous remercions sincèrement les évaluateurs de ce manuscrit pour la qualité de leurs commentaires.

Référence

XXXXX : 4 références sont cachées pour préserver l'anonymat des auteurs, en conformité avec les instructions de la revue.

Amrhein, V., Greenland, S. et McShane, B. (2019). Scientists rise up against statistical significance. *Nature. 567*, 305-307. https://doi.org/10.1038/d41586-019-00857-9

Baguley, T. (2009), Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*, 603-617. https://doi.org/10.1348/000712608X377117

Benjamin, D. J., Berger, J. O., Johannesson, M. et 69 autres auteurs (2017). Redefine statistical significance. *Nature Human Behavior, 2,* 6-10. https://doi.org/10.1038/s41562-017-0189-z

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman Have Agreed on Testing? *Statistical Science*, *18*, 1-12. https://doi.org/ 10.1214/ss/1056397485

Betensky, R. A. (2019) The *p*-Value Requires Context, Not a Threshold, *The American Statistician, 73*(sup1), 115-117. https://doi.org/10.1080/00031305.2018.1529624

Blume, J. D., Greevy, R. A., Welty, V. F., Smith, J. R. et Dupont, W. D. (2019). An Introduction to Second-Generation *p*-Values, *The American Statistician, 73*(sup1), 157-167. https://doi.org/10.1080/00031305.2018.1537893

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Cabaret, M. (2014). Pourquoi les hirondelles volent-elles bas avant le mauvais temps? [extrait audio]. Espace des sciences, Centre de culture scientifique technique et industrielle de Rennes. https://soundcloud.com/espacedessciences/pourquoi-les-hirondelles?utm_source=clipboard&utm_medium=text&utm_campaign=social_sharing

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*(3), 378-399. https://doi.org/10.17763/haer.48.3.t490261645281841

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155–159.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*(12), 997-1003. https://doi.org/10.1037/0003-066X.49.12.997

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7-29. https://doi.org/10.1177/0956797613504966

Cumming, G. (2008). Replication and *p* intervals : *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286-300. https://doi.org/10.1111/j.1745-6924.2008.00079.x

Dahiru T. (2008). P-Value, a true test of statistical significance? A cautionary note. *Annals of Ibadan postgraduate medicine*, *6*(1), 21–26. https://doi.org/10.4314/aipm.v6i1.64038

Demidenko, E. (2016). The *p*-Value You Can't Buy, *The American Statistician, 70*(1), 33-38. https://doi.org/10.1080/00031305.2015.1069760

Denis, D. J. (2004) The modern hypothesis testing hybrid: R. A. Fisher's fading influence. *Journal de la Société Française de Statistique*, *145*, 5-26.

Dienes, Z. (2008). Understanding psychology as a science: An introduction to scientific and statistical inference. Palgrave Macmillan.

Dixon, P. (2003). The p-value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 57*(3), 189-202. https://doi.org/10.1037/h0087425

Easterbrook, P. J., Berlin, J. A., Gopalan, R. et Matthews, D. R. (1991). Publication bias in clinical research. *Lancet (London, England)*, *337*(8746), 867-872. https://doi.org/10.1016/0140-6736(91)90201-y

Finkel, E. J., Eastwick, P. W. et Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology, 108*(2), 275-297. https://doi.org/10.1037/pspi0000007

Fisher, R.A. (1934). *Statistical methods for research workers* (5e éd.). Oliver and Boyd: Edinburgh.

Fisher, R. A. (1935). *The Design of Experiments*. Hafner Publishing Co.

Flake, J. K. et Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456-465. https://doi.org/10.1177/2515245920952393

Gelman, A. et Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant, *The American Statistician, 60*(4), 328-331. https://doi.org/10.1198/000313006X152649

Gignac, G. E., & Szodorai, E. T. (2016) Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74-78. https://doi.org/10.1016/j.paid.2016.06.069

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational researcher*, *5*(10), 3-8. https://doi.org/10.2307/1174772

Goodman S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of internal medicine*, *130*(12), 995-1004. https://doi.org/10.7326/0003-4819-130-12-199906150-00008

Goodman S. N. (2008). A dirty dozen: twelve p-value misconceptions. *Seminars in hematology*, *45*(3), 135-140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Goodman, S. N. (2019). Why is Getting Rid of *P*-Values So Hard? Musings on Science and Statistics, *The American Statistician, 73*(sup1), 26-30. https://doi.org/10.1080/00031305.2018.1558111

Greenberg S. A. (2009). How citation distortions create unfounded authority: Analysis of a citation network. *BMJ, 339*(b2680), 1-14. https://doi.org/10.1136/bmj.b2680

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., et Altman, D. G. (2016) Statistical tests, P values, confidence intervals, and power: A guide to misconceptions. *European Journal of Epidemiology, 31*, 337-350. https://doi.org/10.1007/s10654-016-0149-3

Grove, J. (2015). *Social sciences and humanities faculties 'to close' in Japan after ministerial intervention*. Times Higher Education. https://www.timeshighereducation.com/news/social-sciences-and-humanities-faculties-close-japan-after-ministerial-intervention

Haller, H., et Krauss, S. (2002). Misinterpretations of Significance: A Problem Students Share with Their Teachers? *Methods of Psychological Research Online, 7*(1), 1-20.

Héger, P. (1895). De l'idéal. *Revue de Belgique, 2*(14), 148-169.

Held, L., et Ott, M. (2018). On *P*-Values and Bayes Factors, *Annual Review of Statistics and Its Application, 5*, 393-419. https://doi.org/10.1146/annurev-statistics-031017-100307

Hubbard R, Lindsay RM. (2008). Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing. *Theory & Psychology, 18*(1), 69-88. https://doi.org/10.1177/0959354307086923

Hung, H. M., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The Behavior of the P-Value When the Alternative Hypothesis Is True. *Biometrics*, *53*, 11-22.

Hurlbert, S. H. et Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici, 46(5)*, 311-349. https://doi.org/10.5735/086.046.0501

Howell, D. C. (2010). *Statistical Methods for Psychology* (7th edition). Wadsworth.

Ioannidis J. (2018). The Proposal to Lower *P* Value Thresholds to .005. *JAMA, 319*(14), 1429-1430. https://doi.org/10.1001/jama.2018.1536

Ioannidis J. (2019). Publishing research with P-values: Prescribe more stringent statistical significance or proscribe statistical significance?. *European heart journal*, *40*(31), 2553-2554. https://doi.org/10.1093/eurheartj/ehz555

Kass, R. E. et Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*(430), 773-795. https://doi.org/10.1080/01621459.1995.10476572

Kelter, R. (2020). Bayesian alternatives to null hypothesis significance testing in biomedical research: a non-technical introduction to Bayesian inference with JASP. *BMC Medical Research Methodology, 20*(142). https://doi.org/10.1186/s12874-020-00980-6

Koletsi, D., Karagianni, A., Pandis, N., Makou, M., Polychronopoulou, A. et Eliades, T. (2009). Are studies reporting significant results more likely to be published?. *American journal of orthodontics and dentofacial orthopedics : official publication of the American Association of Orthodontists, its constituent societies, and the American Board of Orthodontics*, *136*(5), 632.e1-632.e5. https://doi.org/10.1016/j.ajodo.2009.02.024

Lakens, D. (2021). The Practical Alternative to the p Value Is the Correctly Used p Value. *Perspectives on Psychological Science, 16*(3), 639-648. https://doi.org/10.1177/1745691620958012

Lakens, D. (2022). Sample Size Justification. Collabra: Psychology. https://doi.org/10.1525/collabra.33267

Loftus, G. R. (1993). A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers, 25*, 250-256. https://doi.org/10.3758/BF03204506

Lykken, D. T. (1968). Statistical significance in psychological research. Psychological Bulletin, 70, 151-159. https://doi.org/10.1037/h0026141

Lykken, D. T. (1991). What's wrong with psychology anyway? Dans: Cicchetti, D. et Grove, W. M. (eds), *Thinking clearly about psychology [Vol. 1]: Matters of public interest*. Minneapolis: Minnesota University Press.

McShane, B. B., et Gal, D. (2016). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *ManagementScience,62*(6), 1707–1718. https://doi.org/10.1287/mnsc.2015.2212

McShane, B. B., Gal, D., Gelman, A., Robert, C. et Tackett, J. L. (2019). Abandon Statistical Significance, *The American Statistician, 73*(sup1), 235-245. https://doi.org/10.1080/00031305.2018.1527253

Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology, 64*(6), 553-557. https://doi.org/10.1037/h0045549

Mitroff, S. R., & Biggs, A. T. (2014). The ultra-rare-item effect: Visual search for exceedingly rare items is highly susceptible to error. *Psychological Science, 25*(1), 284-289. https://doi.org/10.1177/0956797613504221

Murdoch, D. J., Tsai, Y.-L., et Adcock, J. (2008). P-values are random variables. *The American Statistician, 62*(3), 242-245. https://doi.org/10.1198/000313008X332421

Nahm, F. S. (2017). What the *P* values really tell us. *The Korean journal of pain*, *30*(4), 241-242. https://doi.org/10.3344/kjp.2017.30.4.241

Nester, M. R. (1996). An applied statistician's creed. *Applied Statistics, 45(4)*, 401-410. https://doi.org/10.2307/2986064

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, *20A*(1/2), 175–240. https://doi.org/10.2307/2331945

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A, 231*(694-706), 289-337. https://doi.org/10.1098/rsta.1933.0009

Nuzzo, R. (2014) Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Nature, 506, 150-152. https://doi.org/10.1038/506150a

Pashler, H., & Wagenmakers, E.-J. (2012) Editor's introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528-530. https://doi.org/10.1177/1745691612465253

Rafi, Z. et Greenland, S. (2020). Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Med Res Methodol, 20*, 244. https://doi.org/10.1186/s12874-020-01105-9

Ranstam J. (2012). Why the P-value culture is bad and confidence intervals a better alternative. *Osteoarthritis and cartilage*, *20*(8), 805-808. https://doi.org/10.1016/j.joca.2012.04.001

Rosnow, R. L. et Rosenthal, R. (1989). Statistical Procedures and the Justification of Knowledge in Psychological Science. *American Psychologist, 44*(10), 1276-1284. https://doi.org/10.1037/0003-066X.44.10.1276

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120. https://doi.org/c76r6c

Schuirmann, D.J. A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics 15*, 657–680 (1987). https://doi.org/10.1007/BF01068419

Sijtsma K. (2016). Playing with Data--Or How to Discourage Questionable Research Practices and Stimulate Researchers to Do Things Right. *Psychometrika*, *81*, 1-15. https://doi.org/10.1007/s11336-015-9446-0

Sullivan, G.M., & Feinn, R. (2012). Using effect size-or why the *p* value is not enough. *Journal of Graduate Medical Education, 4*(3), 279-82. https://doi.org/10.4300/JGME-D-12-00156.1

Trafimow, D. et Marks, M. (2015) Editorial. *Basic and Applied Social Psychology, 37*(1), 1-2. https://doi.org/10.1080/01973533.2015.1012991

Trafimow, D., Amrhein, V., Areshenkoff, C. N. et 51 autres auteurs (2018). Manipulating the Alpha Level Cannot Cure Significance Testing. *Frontiers in Psychology*, *9(699)*, 1-7. https://doi.org/10.3389/fpsyg.2018.00699

Tufte, E. R. (2001). The visual display of quantitative information. Graphics Press.

Wasserstein, R.L. et Lazar, N.L. (2016) The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician, 70*(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R.L., Schirm, A.L., & Lazar, N.A. (2019) Moving to a world beyond "$p < 0.05$", *The American Statistician, 73*(S1), 1-19, 10.1080/00031305.2019.1583913

Table / Tableau

Tableau 1 : Résultats à un test *t* pour échantillons indépendants où $\mu_1 - \mu_2 = 2{,}5$, $ET_1 = ET_2 = 10$ avec six tailles d'échantillon totales

$n = 20$	$t(18) = 0{,}559$, $p = 0{,}583$
$n = 200$	$t(198) = 1{,}768$, $p = 0{,}079$
$n = 248$	$t(246) = 1.969$, $p = 0.0501$
$n = 250$	$t(248) = 1{,}976$, $p = 0{,}0492$
$n = 600$	$t(598) = 3{,}062$, $p = 0{,}002$
$n = 2000$	$t(1998) = 5{,}590$, $p < 0{,}001$

Note. *n* représente la taille des deux groupes combinés, lesquels sont égaux.

Tableau 2: Types de tailles d'effet, des plus simples et intuitives aux plus abstraits et complexes

Familles	Exemples
Tailles d'effets brutes	la moyenne, l'écart entre deux moyennes, la variance ou l'écart type
Tailles d'effets standardisées	le d de Cohen, le eta-partiel, la corrélation
Tailles d'effets pondérées	la valeur *t*, la valeur *F*
Tailles d'effet uniformisées	la valeur *p*

Tableau 3 : conversion de quelques valeurs *p* en *LFB*

p	0,1	0,05	0,01	0,005	0,001
LFB	1,60	2,44	8,13	13,9	52,9

Figure / Figure

Figure 1 : Exemple de photomontage référant à une valeur *p* significative

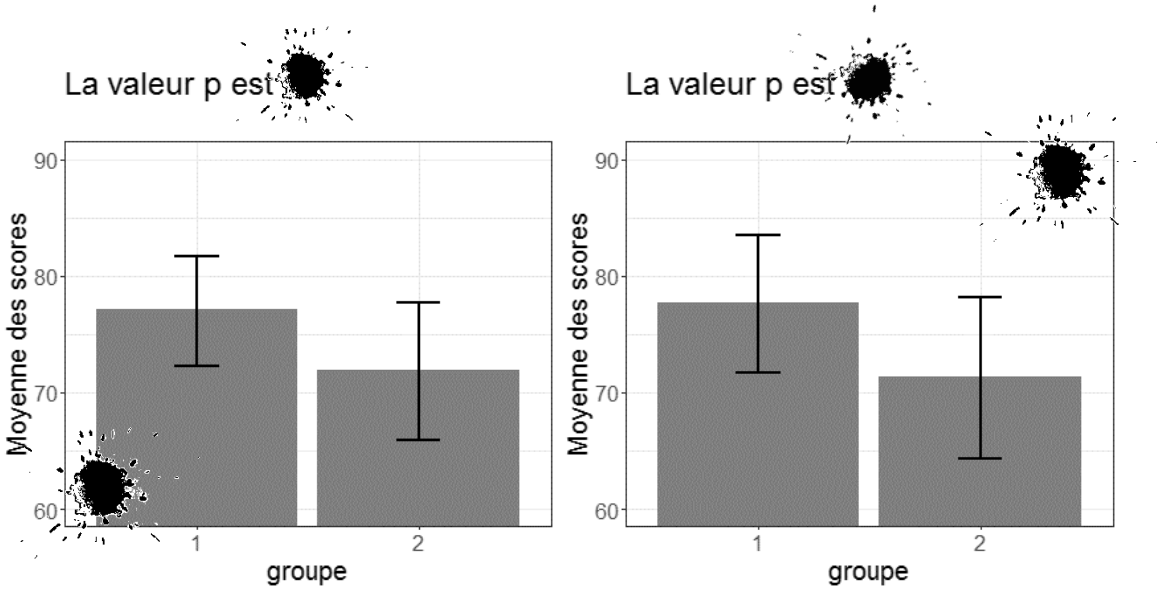

Figure 2: Le test de la tache de café. Benoit est très enthousiaste par ses résultats significatifs alors que Raphaëlle est déçue de ses résultats non significatifs. Leurs graphiques en main (moyennes et intervalles de confiance à 95%; il y a 50 participants par groupes dans les deux études), ils vont voir Guy pour recueillir son avis. Benoît et Raphaëlle sont tellement absorbés qu'ils ne voient pas ce dernier sortir de son bureau avec sa tasse de café noir et ils s'entrechoquent. Les deux graphiques s'envolent dans les airs et se tachent de café du côté des résultats. Pouvez-vous dire quel graphique (celui de gauche ou celui de droite) provient des analyses de Raphaëlle? (voir https://osf.io/prab2/ pour les données et le code utilisés pour réaliser ce graphique, XXXXX).

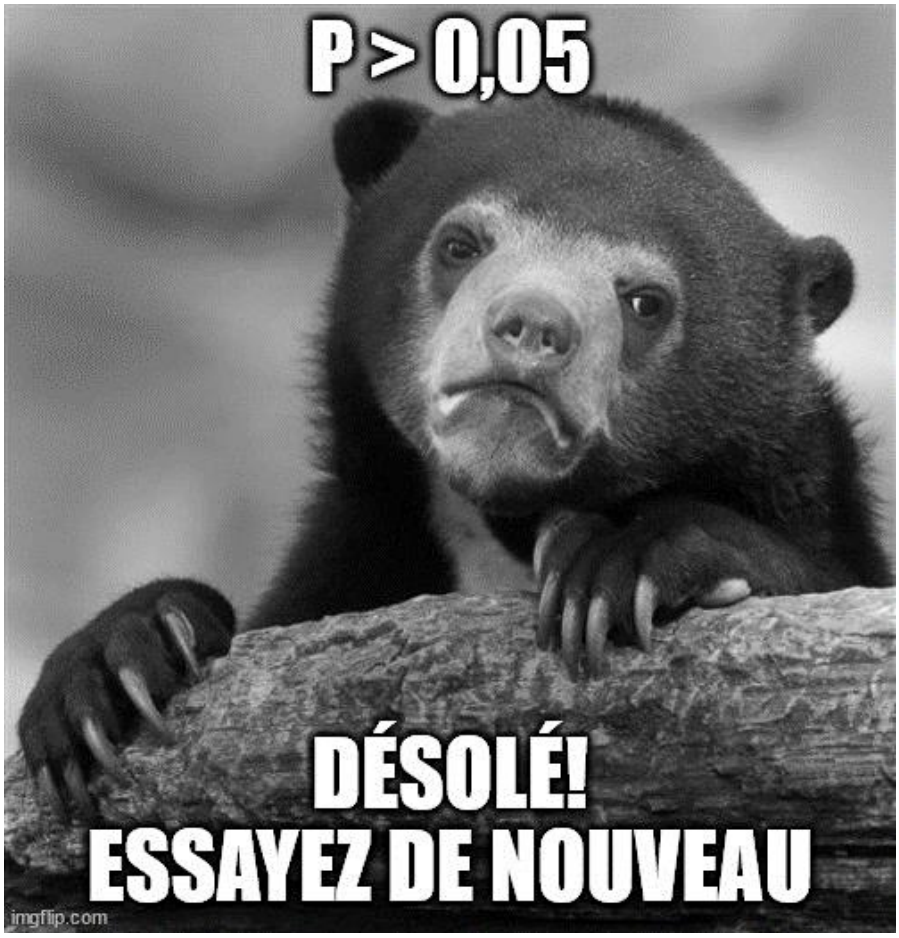

Figure 3 : Désolé! Essayez de nouveau

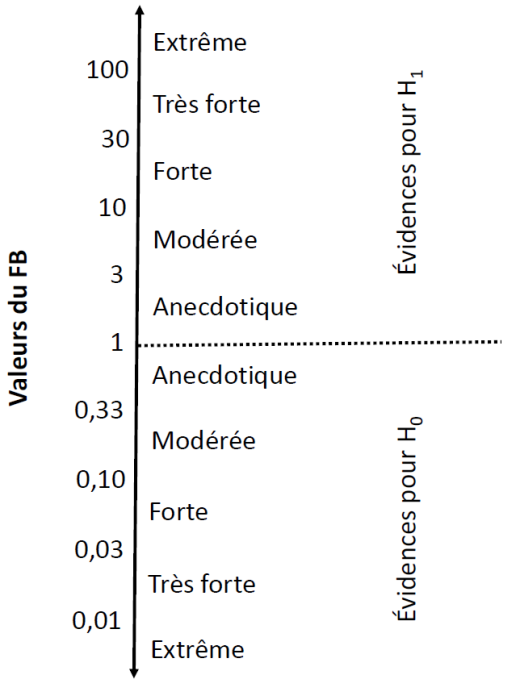

Figure 4 : Barème pour interpréter le facteur de Bayes pour H_0 et H_1 (traduit de la Figure 1 de Kelter, 2020)