Méthodologie

La valeur-p – savoir ou hasard ? Un monde où le doute a sa place.

The p-value – knowledge or fortuity ? A world were doubt has its place.

PAUL VAUCHER[1] (DO, MSc, PhD)

1 Professeur en ostéopathie, Haute Ecole de Santé de Fribourg, Haute Ecole Supérieure Suisse Occidentale (HES-SO)

Source de financement : la rédaction de cette article a été financée par la HES-SO.

Conflits d'intérêts : l'auteur déclare aucun conflit d'intérêt en relation avec cet article.

Approbation éthique : cet article ne nécessite pas d'abrobation éthique.

Keywords
Methodology, Statistics, Probability, p-value

Mots clés
Méthodologie, statistique, probabilité, valeur-p

Abstract

Living well with uncertainties and the unknown is a common trait shared between clinicians and scientists. This article overviews basic principles related to the p-value and the place it holds in scientific reasoning. p-values have a meaning only if rigorous methodological efforts have been put into place to make it possible to reject the ideas that are been tested. Ideas then remain truthful as long as they keep resisting to the challenges of been falsified. The p-value is our indicator of the level of uncertainty we agree to have when accepting an idea.

Résumé

Embrasser l'incertitude pour l'intégrer dans son quotidien; un point que se partage les cliniciens et les scientifiques. Cet article survole les principes de base de la valeur-p et sa place dans le raisonnement scientifique. Un accent particulier est mis sur l'importance de la rigueur méthodologique et le devoir du scientifique de tout faire pour prouver que son idée est fausse. Le vraisemblable est ce qui résiste aux épreuves de falsification. La valeur-p sert d'indicateur pour quantifier le niveau d'incertitude qui accompagne une idée qu'on accepterait.

Introduction

Se rapprocher le plus possible du correcte et du vrai est une garantie qu'on souhaite offrir à nos patients.[1] Cette quête d'honnêteté devient une réelle leçon d'humilité qui forge une complicité dans l'inconnu avec nos patients. Avec le temps, la maturation de notre métier et notre réflexion critique nous ont poussé à reconnaître que la déduction clinique dans notre métier est une illusion – les certitudes absolues n'existent pas.[2, 3] Avec nos patients, nous construisons donc ensemble un plan thérapeutique qui nous paraît vraisemblable (i.e. compatible avec les théories connues, cohérent, réaliste et parlant pour le patient)[4] tout en conservant le doute sur l'exactitude de nos interprétations [3, 5]. L'art de notre métier est de tendre vers le vrai sans jamais vraiment le saisir. Le faux devient un allié qu'on cherche à reconnaître et à minimiser.

Cette humilité face à nos erreurs se trouve également au cœur de la science. Un des piliers centraux du savoir moderne repose en effet sur le principe de reconnaître et quantifier l'erreur [6]. La beauté philosophique et symbolique de la science est bien trop souvent cachée en profondeur derrière sa fourrure épaisse de complexes calculs, formules et lois. Ce premier article de méthodologie vise à entrer dans le cœur de cette bête afin que nous, cliniciens, puissions apprivoiser et apprécier le sens de la valeur-p. Un simple chiffre qui vise uniquement à quantifier la coïncidence lors de notre tentative de chercher l'erreur.

Développement

L'empirisme

Pendant longtemps nous nous sommes contentés de la raison pour constituer nos écoles de pensées sans accorder d'importance à l'expérience. La métaphysique permettait cependant des écarts par rapport aux faits. Les concepts des cinq sens [7] ou des quatre éléments [8] en sont des exemples. Ce n'est qu'au XVII^{e} siècle que le savoir fondé sur l'observation émerge. On reconnait alors la place de l'empirisme et de l'induction comme source de savoir. Hors, le savoir issu uniquement de l'induction pose un problème; celle-ci ne peut pas assurer la véracité des théories car on ne peut jamais exclure totalement l'existence d'une contradiction. Pour lier rationalisme et empirisme, *Popper* [6] pose donc les fondements de la pensée scientifique. Une idée peut être acceptée comme vraie tant qu'elle résiste à l'épreuve d'être réfutée (Tableau 1). On dispose alors d'arguments rationnels et empiriques pour la soutenir.

L'approche scientifique du savoir empirique

1. Avoir une idée rationnelle (qui peut émerger de l'observation).
2. L'énoncer sous une forme réfutable.
3. Chercher par l'observation à démontrer que l'énoncé est faux (partie empirique).
4. Accepter l'énoncé comme vraisemblable tant qu'il résiste aux épreuves.

› Tableau 1 : l'empirisme, l'inférence, le scepticisme et le savoir

Nous allons maintenant survoler quelques concepts clefs vous permettant de mieux comprendre et interpréter les valeurs-p que vous trouverez dans des articles scientifiques.

L'hypothèse nulle

Le cerveau humain est fait pour donner un sens à des coïncidences [9]. Prenons un exemple fictif. Lors d'un cours de formation continue, *Evelyne* et *François* réalisent qu'ils sont tous les deux du 27 novembre. Intrigués, ils découvrent ensuite qu'ils ont également tous les deux un chat, ont voyagé en Amérique du Sud et ont travaillé dans la restauration durant leurs études. Pierre souligne alors que les traits, « aventuriers, spontanés, et curieux », sont caractéristiques des sagittaires. Il n'est donc pas étonnant que tous les deux soient ostéopathes et aient suivi le même cours. *Julie*, qui les a entendus, fait alors remarquer qu'à partir d'un groupe de 23 personnes, il devient totalement attendu que deux personnes partagent la même date d'anniversaire. Leur présence à ce cours est-elle donc réellement due à leur date de naissance commune ou n'est-elle qu'une simple coïncidence ?

La première épreuve de véracité qu'on s'impose est de vérifier que ce qu'on pense être une réalité ne résulte pas simplement du hasard. On va donc chercher à démontrer la véracité de l'hypothèse nulle, à savoir qu'il n'existe aucun lien entre le fait d'être sagittaire et de suivre des cours de formation continue en ostéopathie. On va ensuite quantifier la probabilité qu'une différence observée résulte simplement du hasard. Si cette probabilité devient très faible, on peut alors raisonnablement rejeter l'hypothèse nulle et accepter l'hypothèse alternative comme étant vraisemblable. L'avantage de cette approche est de pouvoir quantifier notre incertitude. Le doute de rejeter à tort l'hypothèse nulle s'appelle « erreur de première espèce » et il est quantifié par la valeur-p (Tableau 2) qui est comprise entre zéro et un. Plus cette valeur est faible, plus l'idée de base a résisté à l'épreuve. Dans notre exemple, si p=0.478, on aurait presque une chance sur deux que la différence observée résultait simplement du hasard.

	Hypothèse nulle	
Enoncé	Acceptée	Rejetée
Vrai	β Erreur de 2^{e} espèce (β<0.2)	α-1 Niveau de signification (α-1>0.95)
Faux	β -1 Puissance (β-1>0.8)	α (valeur-p) Erreur de 1ère espèce (α<0.05)

α = probabilité de rejeter à tort l'hypothèse nulle (≈faux-positif pour l'hypothèse alternative)
α - 1 = probabilité de justement rejeter l'hypothèse nulle (≈vrai-positif pour l'hypothèse alternative)
β = probabilité d'accepter à tort l'hypothèse nulle (≈faux-négatif pour l'hypothèse alternative)
β - 1 = probabilité de justement accepter l'hypothèse nulle (≈vrai-négatif pour l'hypothèse alternative)

› Tableau 2 : les quatre probabilités liées à un énoncé. Les valeurs entre parenthèses correspondent aux valeurs seuils habituelles que nous choisissons arbitrairement.

Puissance d'une étude

En clinique, on admet que la douleur thoracique à l'effort peut évoquer un angor instable [10]. Pour un patient donné, la présence de la douleur peut cependant nous mener à croire faussement qu'il souffre d'un angor (i.e. faux positif). On peut cependant aussi croire à tort qu'en absence de douleur, un patient n'a pas d'angor (i.e. faux négatif). Ce deuxième type d'erreur existe également lorsque l'on met à l'épreuve une hypothèse. On peut commettre une « erreur de deuxième espèce » en acceptant à tort l'hypothèse nulle (Tableau 2). Pour qu'une étude ait du sens, il est donc très important qu'elle ait suffisamment de puissance pour avoir un bon niveau de preuve pour rejeter l'idée de départ; on doit être capable de tester l'absence de ce qui serait cliniquement significatif. La puissance d'une étude reflète donc la capacité d'une étude à pouvoir réfuter une idée. Les petites études ou les études qui se focalisent uniquement sur ce qui est « statistiquement significatif » passent donc entièrement à côté du but premier qu'on est supposé se fixer. La valeur-p ne donne aucune indication sur cette puissance et ne permet donc pas à elle seule d'exclure l'existence d'un phénomène.

Erreur par chance (random error)

Une bonne étude est une étude qui minimise les erreurs de première et de deuxième espèce qu'on appelle aussi l'erreur par chance. Le choix du risque qu'on est prêt à prendre de commettre cette erreur est totalement arbitraire. Les facteurs déterminants pour minimiser les erreurs par chance sont la taille de l'échantillon et l'uniformité des résultats entre les participants.

Pour la valeur-p, on rejette l'hypothèse nulle généralement pour un $p<0.05$. On est donc disposé à accepter un énoncé comme vrai à tort une fois sur vingt (erreur de première espèce). Il est cependant tout à fait possible de choisir une autre valeur. Le seuil choisi pour accepter l'erreur par chance doit cependant être déterminé avant de commencer les observations. Il est aussi préférable de se fier à la valeur-p observée plutôt qu'uniquement au seuil ($p=0.004$ est bien plus informatif que $p<0.05$).

On est généralement un peu moins exigeant pour accepter des preuves qu'une idée est fausse. La puissance des études est souvent fixée à 0.8, soit une erreur de deuxième espèce de 0.2. Il s'en suit qu'une étude négative sur cinq considère à tort l'idée de départ comme étant fausse. On est donc bien plus méfiant envers des études négatives qui ont souvent de la peine à être publiées même si $p=0.976$. A l'inverse, même avec une puissance faible, une étude qui présente une petite valeur-p a bien plus de chance de se trouver dans un article scientifique. Ce biais de publication fait que les nouvelles idées publiées restent incertaines et sont bien souvent contredites par la suite [11, 12]. En clinique, il est donc préférable de se fier à des résultats qui ont été reproduits plusieurs fois dans différentes études (ex. revue systématique, méta-analyse) plutôt que d'appliquer et accepter comme vrai une idée mise à l'épreuve par une seule étude.

La valeur-p est une probabilité et elle peut perdre son sens dans plusieurs circonstances. On va maintenant voir deux exemples d'erreurs qu'on trouve fréquemment dans les publications lorsqu'on rapporte des valeurs-p.

Dans les essais cliniques, les participants sont répartis aléatoirement dans différents groupes. Si cette répartition est faite correctement, la probabilité que le hasard explique les différences entre les groupes au début de l'étude est de 100 %. En d'autres termes, la valeur-p est toujours égale à un. Il est donc insensé de calculer statistiquement des valeurs-p pour comparer l'équilibre des caractéristiques de départ entre des groupes randomisés.

Certaines études collectent une multitude de mesures différentes pour ensuite uniquement rapporter et se focaliser sur celles qui sont significatives. Cette approche s'appelle « p-hacking » ou « cherry picking » [13, 14]. Elle est perçue par les scientifiques comme une faute professionnelle grave. En effet, le but initial d'une étude est de démontrer que l'idée testée est fausse. Il est donc aberrant d'écarter les résultats qui soutiennent cette position. En plus, si on aligne une multitude de possibilités de réponses, inévitablement par simple probabilité au moins une sur vingt paraîtra comme étant significative ($p<0.05$). En se focalisant sur ce résultat, le risque réel de rejeter à tort l'hypothèse nulle est bien plus élevé que celui estimé par la valeur-p. Il est donc nécessaire d'interpréter les valeurs-p sur l'ensemble des mesures observées. Une manière alternative fréquemment utilisée pour contourner ce problème est de déterminer à l'avance quelle est la mesure principale et de se fier uniquement à ce résultat pour la valeur-p.

Une autre erreur qui est communément commise est de considérer qu'une association a un sens dès que la valeur-p est en dessous du seuil fixé. Une étude avec un nombre très important de participant aura une grande puissance statistique et sera capable de détecter même des effets ou associations de très faibles magnitudes qui n'ont plus vraiment de sens. Pour correctement interpréter la valeur-p, il est donc indispensable d'également avoir une petite idée de la taille de l'échantillon et de la magnitude de la différence observée.

Discussion

La valeur-p teste l'absence de différence. Elle ne donne donc aucune indication sur le sens d'une association. Il se peut donc qu'on ait une valeur $p<0.05$ mais pour une association inversée par rapport à celle qu'on attendait.

La valeur-p ne donne aucune indication sur la magnitude d'une association ou d'un effet. Pour de très grands échantillons, même les associations de faible magnitude peuvent être très significatives. De même, une étude en sous-puissance peut avoir une valeur-p élevée alors que la magnitude observée est cliniquement importante.

Le Tableau 3 donne un résumé des points importants à retenir et à vérifier pour vous aider à interpréter correctement les valeurs-p des publications.

- S'assurer que la méthode statistique a été prédéterminée avant que toutes formes d'analyse aient débuté et que celui qui effectue les analyses n'ai pas pu avoir une connaissance préalable des résultats.
- Connaître la taille de l'échantillon ou le nombre de mesures effectuées et se méfier des résultats de petits échantillons ($n<200$).
- Lorsque de multiples tests ont été effectués, comprendre qu'il est normal de trouver par chance une association qui n'existe pas. S'assurer que l'interprétation des résultats tient compte du fait qu'on ait pu effectuer une multitude de tests.
- S'assurer que la magnitude de l'effet a un sens clinique.

› Tableau 3 : règles d'or pour bien interpréter les valeurs-p

Conclusion

La valeur-p nous rappelle que l'incertitude est toujours là et qu'on l'accepte. Elle ne prend véritablement un sens que si les efforts nécessaires ont été faits pour planifier une étude capable de rejeter l'idée qu'on avait décidé de tester. L'esprit scientifique fait que l'on se permet uniquement de s'approprier comme connaissance ce qui a résisté aux épreuves de

falsification. Plus un énoncé est invraisemblable, plus la demande en preuves va être importante. Le clinicien se voit donc obligé de naviguer dans l'inconnu et de prendre une multitude de décisions qui ne sont pas fondées sur l'évidence. L'importance étant de se laisser guidé par ce qui l'est [15].

Implications pour la pratique

- L'incertitude est un allié une fois qu'on la reconnaît.
- Un principe peut être appliqué jusqu'à ce qu'on ait de l'évidence qu'il est faux.
- La remise en question est saine et permet à la profession d'évoluer.

Contact

Paul VAUCHER, Professeur spécialisé en ostéopathie
PhD, MSc Clinical Trials, Ostéopathe CDS-GDK,
Haute Ecole de Santé de Fribourg,
Haute Ecole Supérieure Suisse Occidentale (HES-SO)
Rue des Cliniques 15, CH-1700 Fribourg, Suisse
+41 26 429 60 41

E-mail : paul.vaucher@hefr.ch

Références

1. Gillett G. Virtue and truth in clinical science. J Med Philos. 1995;20(3):285-98.
2. Banning M. A review of clinical decision making: models and current research. J Clin Nurs. 2008;17(2):187-95.
3. Thomson OP, Petty NJ, Moore AP. Clinical reasoning in osteopathy – More than just principles? International Journal of Osteopathic Medicine. 2011;14(2):71-6.
4. Mahr G. Narrative medicine and decision-making capacity. J Eval Clin Pract. 2015;21(3):503-7.
5. Flach PA, Kakas AC. Abductive and Inductive Reasoning: Background and Issues. In: Flach PA, Kakas AC, editors. Abduction and Induction: Essays on their Relation and Integration. Dordrecht: Springer Netherlands; 2000. p. 1-27.
6. Wilkinson M. Testing the null hypothesis: the forgotten legacy of Karl Popper? J Sports Sci. 2013;31(9):919-20.
7. Keeley BL. Making sense of the senses. J Philosophy. 2002;99(1):5-28.
8. Smith P. Alchemy and the Science of Matter. Science. 2007;315(5808):43-4.
9. Kray LJ, George LG, Liljenquist KA, Galinsky AD, Tetlock PE, Roese NJ. From what might have been to what must have been: counterfactual thinking creates meaning. Journal of personality and social psychology. 2010;98(1):106.
10. Vaucher P, Gencer B, Herzig L, Verdon F, Ruffieux C, Bosner S, et al. Ruling out coronary heart disease in primary care patients with chest pain: a clinical prediction. BMC medicine. 2010;8:9.
11. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. 2005;2(8):e124.
12. Ioannidis JP. Failure to Replicate: Sound the Alarm. Cerebrum. 2015;2015.
13. Morse JM. «Cherry picking»: writing from thin data. Qual Health Res. 2010;20(1):3.
14. Ulrich R, Miller J. p-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). Journal of Experimental Psychology-General. 2015;144(6):1137-45.
15. Shlonsky A, Mildon R. Methodological pluralism in the age of evidence-informed practice and policy. Scand J Public Health. 2014;42(13 suppl):18-27.

Quizz (tester vos connaissances)

Pour chacune des cinq questions suivantes, choisissez la réponse qui vous paraît la plus juste.

1. La valeur-p mesure

A. le degré de certitude qu'un énoncé est vrai.
B. le degré d'incertitude qu'un énoncé est vrai.
C. le degré de certitude qu'un énoncé est faux.
D. le degré d'incertitude qu'un énoncé est faux.
E. le degré de magnitude qu'un énoncé est vrai.

2. En acceptant de se tromper une fois sur vingt, un résultat deviendrait significatif pour accepter l'hypothèse alternative à partir de

A. P<0.05
B. P>0.05
C. P=0.2
D. P<0.95
E. P>0.95

3. La valeur-p n'a aucun lien avec

A. la présence d'une association réelle.
B. la magnitude de l'association réelle.
C. la direction de l'association réelle.
D. la taille de l'échantillon.
E. l'homogénéité des réponses.

4. Le biais de publication

A. augmente le risque qu'une étude soit véritablement positive.
B. diminue le risque qu'une étude soit véritablement négative.
C. augmente le risque qu'une étude soit faussement négative.
D. diminue le risque qu'une étude soit faussement positive.
E. augmente le risque qu'une étude soit faussement positive.

5. Lorsque la puissance d'une étude n'est pas assez élevée

A. On ne peut pas calculer la valeur-p.
B. On risque de ne pas pouvoir conclure.
C. On ne peut pas mener l'étude jusqu'au bout.
D. On risque de ne pas pouvoir suivre les participants jusqu'à la fin de l'étude.
E. On ne peut pas rejeter l'hypothèse nulle.

Réponses

1. B / 2. A / 3. C / 4. E / 5. B