L'effet *significatif* des Visualisations sur le Raisonnement Dichotomique

The *Significant* Effect of Visual Representations on Dichotomous Thinking

Jouni Helske, Matthew Cooper, Anders Ynnerman, Lonni Besançon

Résumé—Common reporting styles of statistical results for controlled experiments (confidence intervals and p-values) have been shown to lead to dichotomous inferences on the study outcome. We investigated in two controlled experiments whether several different visual representations could help limit such dichotomous interpretations. The results of our experiments suggest that some visual representations can help decrease binary interpretations of statistical results compared to textual representations of p-values and confidence intervals. All data analysis and scripts are available online.

1 Introduction

Dans de nombreux domaines scientifiques, l'une des questions de recherche les plus courantes est la suivante : "X a-t-il un effet sur Y ?", X étant, par exemple, un nouveau médicament et Y étant une maladie. Il s'agit souvent de déterminer si l'effet moyen de X sur un échantillon de la population diffère de 0 ou s'il est plus efficace qu'un autre médicament Z en moyenne. Souvent, les scientifiques cherchent à répondre à cette question sous la forme d'un test d'hypothèse nulle (NHST). Le NHST repose sur l'idée de postuler une hypothèse nulle "sans effet" que l'on cherche à rejeter, en calculant une statistique de test et la valeur p correspondante, définie comme la probabilité d'observer un résultat au moins aussi extrême que celui que l'on observe, supposant que cette hypothèse nulle est exacte. En supposant que nos hypothèses sont vraies, une valeur p indique que nos données sont incompatibles avec le modèle nul.

La "crise de réplication" actuelle a donné lieu à de nombreux commentaires critiques à l'encontre des seuils arbitraires de valeurs p et des tests d'hypothèse nulle en général [1], [2], [8], [10]–[12]. Les statisticiens et les praticiens réclament des analyses ou des stratégies de présentation des résultats statistiques différentes, basées sur des intervalles de confiance et des techniques d'estimation [3], [5]–[7], pour éviter les dangers habituels des NHST et des réflexions dichotomiques. Malgré ces critiques et les solutions proposées, il semblerait que la pensée dichotomique soit encore répandue dans les publications en Interface Homme Machine (IHM) [4]. Par conséquent, il est plus que probable que les tests d'hypothèse nulle resteront dans la boîte à outils scientifique pour les années à venir. Au lieu de plaider pour de meilleures solutions méthodologiques (comme la conversion aux approches bayésiennes [9]), nous étudions ici, avec deux expériences contrôlées, si différents styles de visualisation de résultats statistiques communs pourraient aider à atténuer certains des problèmes liés aux tests d'hypothèse nulle.

Tous les auteurs travaillent à Linköping University.
Correspondant: Lonni Besançon, lonni.besancon@gmail.com

2 Expériences

Nos expériences contrôlées imitent les expériences précédentes sur le *cliff effect*. Le cliff effect correspond à une chute soudaine de la confiance que les chercheurs accordent à leurs résultats lorsque la valeur p dépasse 0,05. Cependant, nous avons produit le premier ensemble d'expériences entièrement reproductibles sur le cliff effect et avons considéré plus de représentations visuelles que de simples intervalles de confiance et de valeurs p textuelles. Les deux expériences ont été menées sous forme de sondages en ligne. Les pré-enregistrements sont disponibles sur https://osf.io/v75ea/ et https://osf.io/brjzx/ et sont respectivement disponibles sur https://weber.itn.liu.se/~jouhe21/experiment1/ et 18~jouhe21/experiment2. Comme l'indique le pré-enregistrement, le nombre de participants n'a pas été décidé à l'avance, mais la date de fin de l'expérience a été fixée à l'avance à une durée de 21 jours. Pour recruter des participants, le sondage a été diffusé ouvertement sur Twitter, Reddit, LinkedIn, Facebook et par courriel aux universitaires dans plusieurs domaines (notamment l'interaction homme-machine, la visualisation, les statistiques, la psychologie et la sociologie analytique).

Nous avons utilisé un ensemble fixe de valeurs p (0,001, 0,01, 0,04, 0,05, 0,06, 0,1, 0,5, 0,8) et un écart type fixe de 3. Au cours de la première expérience, nous avons demandé aux participants : "Un échantillon aléatoire de 200 adultes Suédois s'est fait prescrire un nouveau médicament pendant une semaine. D'après l'information à l'écran, dans quelle mesure

êtes-vous confiant que le médicament a un effet positif sur le poids (augmentation du poids) ?". La réponse a été donnée sur une échelle continue (100 points, la valeur numérique n'était pas indiquée) à l'aide d'un curseur, avec des extrémités étiquetées ("Confiance zéro", "Confiance totale"), qui a été expliquée aux participants comme suit : "La position la plus basse du curseur correspond au cas "Je n'ai aucune confiance pour affirmer un effet positif", alors que la position la plus droite du curseur correspond au cas "Je suis pleinement convaincu qu'il y a un effet positif"". Le potentiomètre du curseur a été caché en premier afin d'éviter le biais possible dû à sa position initiale. Il est ensuite devenu visible lorsque le participant a cliqué sur le curseur. Enfin, tant que le curseur était masqué, les participants ne pouvaient pas passer à la question suivante. La deuxième enquête avait un cadre semblable, mais cette fois-ci, au lieu de comparer la valeur à une référence de 0 (pas d'effet), il s'agissait de comparer la moyenne de deux groupe : le groupe "traitement " et le groupe "témoin".

Pour la première expérience, nous avons utilisé quatre représentations différentes des résultats statistiques : valeurs textuelles, intervalles de confiances, dégradés et violin plots (les quatres premières figures visibles sur la Figure 1). Pour la deuxième espérience, les participants ont pu voir les trois visualisations précédentes mais les valeurs textuelles étaient remplacées par un violin plot discrétisé (les quatre dernières figures visibles sur la Figure 1).

Afin de minimiser les effets d'apprentissage, l'ordre des quatre conditions de chaque expérience (style de représentation) a été contrebalancé par un design Latin Square, et dans chaque condition, l'ordre des essais a été aléatoirement mélangé pour chaque participant. Nous avons recueilli les réponses de 114 participants pour la première expérience et de 39 participants pour la seconde.

3 Résultats et Conclusion

Nos résultats indiquent que, malgré les nombreuses mises en garde sur l'utilisation abusive des valeurs p et du test d'hypothèse nulle, les chercheurs sont encore susceptibles de faire des interprétations dichotomiques, même lorsqu'ils ont une formation statistique solide. Il semblerait également que la représentation visuelle, comme les intervalles de confiance et les violin plots, puisse aider à réduire les interprétations dichotomiques si ils sont bien expliqués. De plus, les participants ont mentionné que ces représentations étaient plus informatives parce qu'elles présentent plus d'information statistiques qu'une simple valeur textuelle. C'est notamment le cas pour le violin plot. L'un de leurs inconvénients, cependant, par rapport aux intervalles de confiance classiques, est qu'ils sont moins compacts et pourraient donc être problématiques pour des comparaisons multiples.

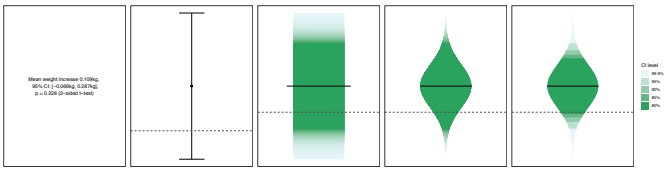

FIGURE 1. Visualisations comparées : version textuelle avec valeur p, visualisation classique d'un intervalle de confiance, intervalle de confiance degradé, violin plot et violin plot discrétisé.

Références

[1] V. Amrhein, S. Greenland, and B. McShane. Scientists rise up against statistical significance. *Nature*, 567(7748) :305–307, 2019.

[2] V. Amrhein, F. Korner-Nievergelt, and T. Roth. The earth is flat (p>0.05) : Significance thresholds and the crisis of unreplicable research. *PeerJ Preprints*, 5 :e2921v2, 2017.

[3] L. Besançon and P. Dragicevic. The significant difference between p-values and confidence intervals. In *Proc. IHM*, page 10, Poitiers, France, 2017. AFIHM.

[4] L. Besançon and P. Dragicevic. The continued prevalence of dichotomous inferences at CHI. In *CHI '19 - Proceedings of CHI Conference on Human Factors in Computing Systems Extended Abstracts*, Glasgow, United Kingdom, 2019.

[5] R. Calin-Jageman and G. Cumming. The new statistics for better science : Ask how much, how uncertain, and what else is known, 2018.

[6] G. Cumming. *Understanding the new statistics : effect sizes, confidence intervals and meta-analysis*. Routledge Taylor & Francis Group, 2012.

[7] P. Dragicevic. Fair statistical communication in HCI. In J. Robertson and M. Kaptein, editors, *Modern Statistical Methods for HCI*, chapter 13, pages 291–330. Springer International Publishing, Cham, Switzerland, 2016.

[8] P. Dragicevic, F. Chevalier, and S. Huot. Running an HCI experiment in multiple parallel universes. In *Extended Abstracts on Human Factors in Computing Systems*, pages 607–618, New York, 2014. ACM.

[9] M. Kay, G. L. Nelson, and E. B. Hekler. Researcher-centered design of statistics : Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the CHI 2016*, pages 4521–4532. ACM, 2016.

[10] D. McCloskey and S. Ziliak. *The Cult of Statistical Significance*. University of Michigan Press, 2008.

[11] B. B. McShane and D. Gal. Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, 112(519) :885–895, 2017.

[12] R. L. Wasserstein, A. L. Schirm, and N. A. Lazar. Moving to a world beyond "p<0.05". *The American Statistician*, 73(sup1) :1–19, 2019.