S'en remettre à la reproductibilité, *Oui, je le veux !*

Joannie Gagnon
Étudiante en Écologie,
Rédactrice d'article à temps partiel,
Université de Sherbrooke

19 avril 2020

Résumé

La reproductibilité est un problème que les scientifiques connaissent depuis 2014 et peut-être même avant. Une reproductibilité adéquate peut éviter à beaucoup de chercheurs de se faire déclarer à tort comme fraudeur, exactement ce qui est arrivé à Jonathan Pruitt tout récemment. Les quelques solutions proposées ici seront d'une grande aide pour aider à sortir de cette boucle sans fin. La balle est dans votre camp maintenant...

C'est tout récemment que la communauté scientifique s'est penchée sur un cas de fraude de données scientifiques. L'article d'Elizabeth Pennisi décrit très bien la situation précaire du jeune écologiste comportemental, Jonathan Pruitt. Celui-ci s'est fait accusé de falsification de données pour son étude de la personnalité animale chez les araignées, car aucun autre chercheur était capable d'avoir les mêmes résultats. Étudiant avec précision les données et les résultats brutes, la communauté scientifique a déclaré Jonathan fraudeur, et a retiré tous articles écris ou collaborés par celui-ci. **Comment Jonathan peut-il s'en sortir maintenant ?**

FIGURE 1 – J.G. Roberto

La reproductibilité. Un terme simple désignant *« la capacité d'un chercheur ou d'une équipe à reproduire les informations d'une étude antérieure en s'appuyant sur les mêmes matériaux (texte, données, code de programmation) que ceux utilisés par le chercheur initial »* [Desquilbet et al., 2019, AFIS Science, 2019]. Autrement dit, faire une expérience déjà faite,

avec la même méthode et avec les mêmes outils. Une analogie toute simple pour exprimer ce concept serait l'arrivée d'un nouveau cuisinier dans un restaurant, devant faire pour la première fois une recette très connue dudit restaurant. Il serait attendu qu'il puisse reproduire la recette (obtenir le même résultat), avec les ingrédients habituels (les mêmes données), selon les étapes préétablies (la même méthode). Juste avec la définition de la reproductibilité, la problématique se fait déjà sentir.

La problématique

Ce n'est pas juste l'article de Mme Pennisi qui présente le problème d'incapacité à imiter un article déjà produit. Dans l'article de *Nature* (dans l'article de Baker, 2016) [Baker, 2016], il est démontré qu'environ 70% des chercheurs (sur un total de 1576) ont essayé de reproduire des résultats de recherche, sans y parvenir. De plus, la moitié d'entre eux le firent avec leur propre recherche.

Les quelques raisons du *pourquoi ne sont-ils point capable de copier les articles* :

1. Des dizaines d'approches alternatives (voire des centaines) existent pour analyser les mêmes données [Munafò et al., 2017]. Ce qui engendre une différence majeure dans l'interprétation des données.
2. Les données peuvent être mal interprétées dans des publications, et une fois publiés, cette interprétation n'est plus effaçable [Mills et al., 2015].
3. La majorité des études reproduites donnant des résultats non-significatifs sont mises de côté, car obtenir un résultat statique significatif indique automatiquement un résultat scientifique... [Lamy, 2017]
4. L'incapacité de reproduire les résultats publiés signifie que le résultat est automatiquement faux [Baker, 2016].
5. Il y a une forte pression de publier le premier et de sélectionner des recherches qui donneront un résultat significatif [Baker, 2016, Munafò et al., 2017].
6. Les chercheurs sont exposés à un risque plus élevé de voir leurs recherches ignorées simplement parce que d'autres non pas été capable de reproduire les expériences [Poisot et al., 2013].

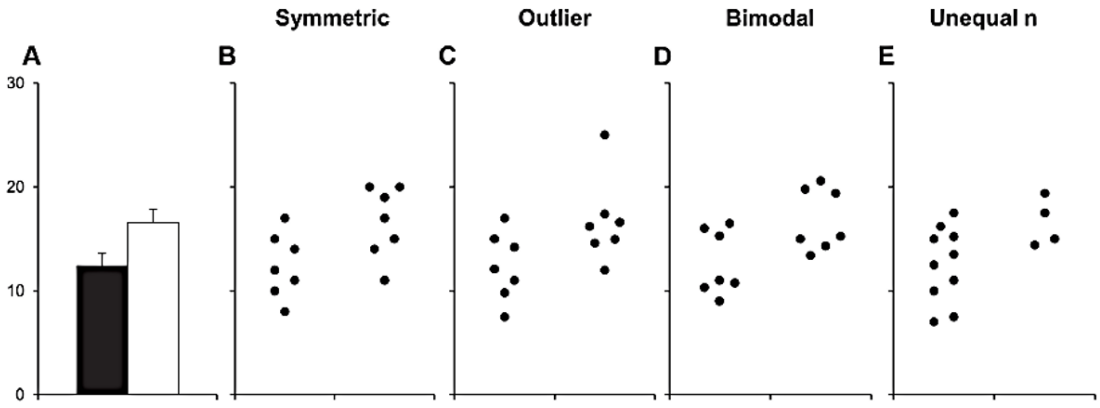
Il en existe encore plusieurs autres. Toutefois, « ces réflexions (...) ont été noyées dans un discours qui associe les problèmes de reproductibilité avec tout ce qui ne fonctionne pas dans le monde scientifique [Lambert-Chan, 2019]». La combinaison entre l'apophénie (dans ce contexte, c'est la tendance à voir des structures dans des données aléatoires), le biais de confirmation (la tendance à se concentrer sur des preuves qui correspondent à nos attentes, donc que des résultats positifs) et le biais de rétrospection (la tendance à considérer qu'un événement n'a été prévisible qu'après s'être produit) peut facilement conduire à de fausses conclusions [Munafò et al., 2017], en plus d'avoir une pression pour publier le plus tôt possible, ou encore l'envie de réussir en absolu [Dupage, 2019, Lamy, 2017], des interprétations statistiques permissifs, un biais de confirmation d'hypothèse, la manipulation des données [Lambert-Chan, 2019], etc. Est-ce une combinaison de plusieurs de ces facteurs qui a mené à un manque de confiance envers monsieur Pruitt ? Est-ce un effet « boule de neige » ?

Les solutions

N'aurait-il dont pas un moyen d'éviter, ou du moins de réduire, cette problématique ? Bien sûr. Il faut d'abord que le monde scientifique soit ouvert d'esprit et qu'il soit prêt à s'investir dans ces nouvelles procédures. En voici quelques solutions.

1. Donner une formation

Avec une formation adéquate auprès des scientifiques sur la présentation des données, ceux-ci peuvent éviter des fausses interprétations. Par exemple, les diagrammes à barres sont conçus pour des variables catégorielles, et non avec des données continues comme certains aiment beaucoup faire. Les graphiques à lignes sont des tableaux visuels montrant entre autres la moyenne, l'erreur-type ou l'écart-type. De nombreuses distributions de données différentes peuvent conduire au même graphique (image ci-dessous). Les lecteurs déduisent donc à tort que ces données sont distribuées normalement, faussant ainsi les résultats. Dans le cas des petits échantillons, par exemple, ce sont les diagrammes de dispersion univariée qui représente le meilleur choix, car ils permettent aux lecteurs d'examiner leur distribution [Weissgerber et al., 2015]. Puisque la présentation des données est essentielle, donner une formation auprès des scientifiques, des enquêteurs de publication et des étudiants pourrait permettre une standardisation et ainsi éviter des lacunes.



Test	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

FIGURE 2 – Plusieurs ensembles de données différentes conduisant au même graphique à barre [Weissgerber et al., 2015]

2. Centraliser les données

La centralisation des données dans une base unique, située à un seul endroit éviterait la fragmentation de celles-ci et assurerait leur totale sécurité et mise à jour [Mills et al., 2015]. Cela peut être sous la forme d'un site internet en lien avec l'industrie de publication ou encore d'une plateforme accessible pour tout le monde. Par exemple, DRYAD (référentiel international en libre accès) ou encore TREEBASE (référentiel de données phylogénétiques) sont des sites où les données brutes sont déposées dans des archives permanentes en libre accès [Mills et al., 2015]. En partageant les données, Poisot et son équipe (2013) ont démontré qu'il y a une augmentation à la fois sur la qualité et la visibilité de la science que les chercheurs tentent de produire. Cette disponibilité améliore la reproductibilité et la communication adéquate des résultats.

3. Faire une prépublication

Vous arrive-t-il de demander un 2e avis à une tierce personne sur une action que vous voudriez faire ? Comme par exemple, voir un 2e médecin pour confirmer une maladie quelconque... Il s'agirait du même principe. En partageant partiellement l'article, les chercheurs peuvent avoir un retour sur l'étude de la part d'une communauté diversifiée. Cela peut se trouver sous forme de plateforme ou de site internet. Il y a entre autres le site web CONSORT *(Consolidated Standards of Reportings Trials)* qui fournit une orientation pour la rédaction d'un rapport clair, complet et précis. Quant à la plateforme ***Wellcome Open Research***, lancée par la fondation *Wellcome Trust*, elle offre une publication gratuite, immédiate et en libre accès avec une évaluation par les pairs après une publication. Ou encore il y a le site web PUBPEER qui est une autre alternative pour recevoir des commentaires sur une étude. Toutefois, ce site est public, et les commentaires anonymes [Munafò et al., 2017]. Tout ces exemples démontrent que la collaboration entre chercheurs améliore la qualité de leur rapport, et ainsi donc avoir une meilleure publication dans une revue littéraire.

Il existe encore plein d'autres solutions. Les revues peuvent fournir une section spéciale en indiquant qui a écrit l'article, qui a conçue l'étude ou encore qui a fourni les données [Poisot et al., 2013]. Ainsi, les chercheurs peuvent s'y référer sans souci et échanger au besoin. Les données peuvent être enregistrées sous forme Java Script (JSON) plutôt que sous le format CSV par Excel [Poisot et al., 2013]. Le script Java permet entre autres une standardisation des données qui est plus facile à manipuler que le CSV. La promotion d'une présentation des données complètes plutôt que partielles est une autre solution [Weissgerber et al., 2015]. Il est prouvé qu'avoir une plus grande taille d'échantillons donne des résultats plus significatives qu'une partie des données ou encore des données mises en commun dans un graphique.

Conclusion

Dans l'ensemble, si les chercheurs et les industries de publication se mettent tous d'accord sur les mêmes règles claires [Mills et al., 2015], précises et bien encadrées, le taux de reproductibilité pourra s'y voir doublé d'ici les prochaines années. L'impact du cas de Jonathan Pruitt a peut-être apporté beaucoup de remises en question auprès des schientifiques. Néanmoins, cela a aussi permis de faire avancer la science plus loin encore avec un regard différent.

Pour en savoir plus ou pour savoir par où commencer pour permettre une bonne reproductibilité, consultez le document de Desquilbet et ses collaborateurs. Vous y trouverez une mine d'or d'informations.

Références

[AFIS Science, 2019] AFIS Science (2019). Comment améliorer la reproductibilité de la recherche scientifique ? https ://www.pseudo-sciences.org/Comment-ameliorer-la-reproductibilite-de-la-recherche-scientifique. Consulté le 3 avril 2020.

[Baker, 2016] Baker, M. (2016). Reproducibility crisis ? *Nature*, 533(26) :353–66.

[Desquilbet et al., 2019] Desquilbet, L., Granger, S., Hejblum, B., Legrand, A., Pernot, P., Rougier, N. P., de Castro Guerra, E., Courbin-Coulaud, M., Duvaux, L., Gravier, P., et al. (2019). Vers une recherche reproductible.

[Dupage, 2019] Dupage, D. (2019). Comment améliorer la reproductibilité de la recherche scientifique ? https ://www.franceinter.fr/emissions/sante-polemique/sante-polemique-30-avril-2019. Consulté le 3 avril 2020.

[Lambert-Chan, 2019] Lambert-Chan, M. (2019). Reproductibilité : de l'impasse à l'espoir. https ://www.quebecscience.qc.ca/edito/reproductibilite-impasse-espoir/. Consulté le 3 avril 2020.

[Lamy, 2017] Lamy, E. (2017). Une crise de reproductibilité de la science ? non, c'est bien pire ! https ://theconversation.com/une-crise-de-reproductibilite-de-la-science-non-cest-bien-pire-85652. Consulté le 3 avril 2020.

[Mills et al., 2015] Mills, J. A., Teplitsky, C., Arroyo, B., Charmantier, A., Becker, P. H., Birkhead, T. R., Bize, P., Blumstein, D. T., Bonenfant, C., Boutin, S., et al. (2015). Archiving primary data : solutions for long-term studies. *Trends in Ecology & Evolution*, 30(10) :581–589.

[Munafò et al., 2017] Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature human behaviour*, 1(1) :1–9.

[Poisot et al., 2013] Poisot, T. E., Mounce, R., and Gravel, D. (2013). Moving toward a sustainable ecological science : don't let data go to waste ! *Ideas in Ecology and Evolution*, 6(2).

[Weissgerber et al., 2015] Weissgerber, T. L., Milic, N. M., Winham, S. J., and Garovic, V. D. (2015). Beyond bar and line graphs : time for a new data presentation paradigm. *PLoS biology*, 13(4).