– TRAVERSES STATISTIQUES –

TONNERRE ET TREMBLEMENTS
La crise existentielle de la statistique en cet automne 2015

Marc Bourdeau[1]

La chronique « Traverses statistiques » souhaite s'appuyer sur l'actualité et présenter des éléments susceptibles d'agrémenter les cours de statistique et de probabilités

« *There is no true value of anything.* »
William Deming[2]

Coup de tonnerre en février dernier dans le ciel statistique. David Trafimow, le nouveau rédacteur en chef de la revue *Basic and Applied Social Psychology (BASP)*, après avoir prévenu (2014) ses lecteurs et auteurs dans un éditorial, a mis sa menace à exécution (2015)—serait-il le premier à le faire? Désormais, il n'acceptera plus de *p-values*[3] dans sa revue. Aucune inférence statistique n'y aura donc sa place : la théorie des tests statistiques pour les hypothèses nulles (TSHN, en anglais *NHSTP*) est logiquement invalide, de même que les intervalles de confiance. Ils ont trop d'interprétations fausses. Trafimow veut augmenter «*the quality of submitted papers by liberating authors of from the stultified structure of null hypothesis of statistical testing procedures thereby eliminating an important obstacle to creative thinking* («la qualité des papiers soumis à sa revue en libérant les auteurs de procédures fossilisées, et par voie de conséquence éliminer un obstacle à la pensée créatrice») (2015, p.2). Il fonde cette interdiction sur une série de papiers qu'il a publiés depuis plus d'une décennie.

On en sera réduit dorénavant à la BASP à des « […] *strong descriptive statistics including effect sizes.* » (« […] des statistiques descriptives robustes, comprenant l'ampleur des effets ») Le terme *strong* est intéressant… Plus d'inférence statistique (au sens classique du terme), plus de modèles, plus de prévision. En fait de progrès on aura vu mieux. Il n'est guère plus tolérant pour l'inférence bayésienne.

La statistique tremble sur ses fondements. On réagit presque du jour au lendemain dans bien des secteurs de notre discipline, souvent avec violence. Soit avec approbation, soit avec dédain, voire avec dérision. Voir le fichier sous-jacent pour un aperçu.

Nous savons tous que le ciel de l'inférence statistique est tout sauf serein. Les nuages lourds de cette controverse parfois assez acerbe l'ont peuplé depuis ses tout débuts

1

[1] Professeur associé, École Polytechnique de Montréal, Canada, Louis.Marc.Bourdeau@gmail.com
[2] Dans sa préface à la réédition du livre de W. A. Shewhart (1986, orig. 1939). On peut télécharger le fichier des références citées et quelques autres ici.
[3] Dans tous les cas de figure, le '*p-value*' est la probabilité que la statistique échantillonnale réalisée sur un échantillon ait donné une valeur absolue au moins aussi grande. *P-value* en anglais n'a aucun sens manifeste. Comme souvent l'anglais est plutôt nébuleux dans ses appellations. Nous avons proposé dans nos cours le terme de 'probabilité de dépassement', $p_{dép,}$ Les étudiants ont besoin de notations claires, celle-ci fait référence au concept de rareté.

© Société Française de Statistique (SFdS), Décembre/December 2015

(Harlow *&al.*, 1997; Morrison & Kenckel, 1970; Krantz, 1999). Fisher lui-même qui a inventé les méthodes de l'inférence statistique moderne (Christensen, 2005) a rappelé à l'ordre ses collègues sur un grand nombre de sujets dont celui qui nous occupe (e. g. Inman, 1994). Mais la controverse n'a jamais pris fin. L'avant dernière attaque en date est celle initiée par Ioannidis au début du siècle (2001 à 2007). Ioannidis, qui est cité partout, s'adresse plus particulièrement à la non-reproductibilité des études en sciences de la santé avec inférence statistique *significative* (les seules qu'on publie), donc à la valeur du critère fondé sur $p < 0,05$, qui est préalable à toute publication. Selon lui, plus de la moitié des études publiées ne seraient pas reproductibles.

Dans ce qui suit, nous passons en revue les deux arguments principaux des négationnistes (!), et abordons rapidement la question de la reproductibilité et celle des conflits d'intérêt qui lui est reliée, avant de proposer quelques suggestions pour limiter les dégâts.

Deux arguments

Trafimow et les négationnistes mentionnent bien que ce qu'il leur faut, c'est, après avoir cueilli les données, une appréciation de l'hypothèse nulle, soit $P(H_0 \mid data)$, alors que la théorie de l'inférence statistique classique ne permet de connaître que $P(data \mid H_o)$, sous la forme de sa rareté, soit la probabilité '$p_{dép}$', dite de dépassement. Beaucoup d'applicateurs ont plutôt mal appris leur probabilité-statistique et interprètent la seconde comme la première. Défaut d'éducation finalement, vice très répandu en notre époque presse-boutonnière, et non un vice logique de la théorie. Fisher (Inman, 1994) avait d'ailleurs fait valoir ce point dès 1934, rien de nouveau ici. Personne n'est dupe chez les statisticiens d'application.

Le statisticien monte un dossier à charge et à décharge, comme on voit partout dans les procès juridiques, pour ou contre la 'culpabilité' de l'hypothèse nulle, rien de plus. Aucune certitude sur la conclusion ne peut s'en dégager. Ce qui n'empêche pas de porter des jugements… Le fondement statistique des TSHN est correct, on verra maintenant que les problèmes sont ailleurs.

L'autre difficulté, plus délicate à contrer, vient du fait que les applicateurs amateurs des TSHN sont souvent tentés d'augmenter leurs tailles échantillonnales (*a posteriori*) jusqu'à obtenir le rejet tant désiré de leur hypothèse nulle. En effet, la région critique s'agrandit avec la taille échantillonnale, et la puissance du test augmente. À la limite, avec une taille échantillonnale assez grande on peut rejeter toute hypothèse nulle. Et alors les chercheurs peuvent publier, alimenter leur CV, obtenir des fonds pour poursuivre leur recherche, sauver leur emploi souvent précaire… Tout cela est un problème de l'organisation socio-politique de la recherche scientifique sur laquelle nous reviendrons.

Cette propriété de l'inférence statistique lorsque N grandit est bien connue, c'est pourquoi il faut rappeler qu'on doit bien définir au moment de la définition des protocoles expérimentaux[4], i. e. *ante hoc,* l'*ampleur des effets* qui sont pratiquement intéressants, i.e. essentiellement les écarts à l'hypothèse nulle qu'il faut détecter avec des probabilités fixées d'avance, i.e. la puissance espérée à cet écart de l'hypothèse nulle. Ce qu'on ne trouve pas souvent en Sciences humaines et sociales (SHS). Et ainsi définir une taille échantillonnale optimale, ni trop petite, ni trop grande en vue des objectifs : Verrill & Durst (2005) expliquent bien les tenants et aboutissants de ces étapes cruciales.

2

[4] À noter aussi qu'en toute rigueur, il faut définir des expériences randomisées pour que l'inférence statistique soit valide, ce qui est bien difficile à réaliser en SHS où les échantillons de convenance, non contrôlés sont légion.

© Société Française de Statistique (SFdS), Décembre/December 2015

La reproductibilité et les conflits d'intérêt

La reproductibilité est la règle d'or du progrès scientifique (Popper, 2002, [orig. 1959], Kenett & Schmuel, à paraître) et surtout Ioannidis (2001, 2005, 2007) qui a attaché le grelot depuis plus de dix ans.
À cet égard, la première chose à remarquer est que toute probabilité de dépassement, '$p_{dép}$', est une variable aléatoire qu'on réalise une seule fois dans toute expérience statistique. Elle a donc une loi, une moyenne, un écart type, etc. Ces propriétés ne sont pas simples à déterminer (Boos & Stefansky, 2011; Sackrowitz & Samuel-Cahn, 1997). Cependant, sous des hypothèses assez générales, Boos & Stefansky (p.221) déduisent un taux de non-reproductibilité aux environs de 33%, cela en admettant l'honnêteté des chercheurs…

Ce taux contraste fortement avec les résultats de la gigantesque étude du *Reproducibility Project,* rapportée en septembre par le *NY Times* ainsi que par Le Monde, à la suite de la revue Science (Nosek *&al.*, 2105) qui estime cette proportion aux deux-tiers! Autre coup de tonnerre! On jette ici un sérieux discrédit sur les Sciences humaines et sociales & les Sciences de la santé (SHS&Santé). Ioannidis l'estime à environ 60%. Young (2008) en annonce au moins 80%...

Les sociétés, souvent multinationales, et autres organismes subventionnaires ont bien du pouvoir sur les chercheurs (Marsan & Laliberté, 2015; Steneck *&al.*, 2015; Foucart, 2015; Angell, divers articles du NYReview of Books). L'honnêteté est une vertu qui se perd.

Limiter les dégâts

La statistique est plus que jamais en état de crise. On a contesté sur une base philosophique l'inférence statistique qu'il ne fallait pas confondre avec l'inférence scientifique. Soit, mais c'est surtout l'organisation scientifique qui est en crise, et non le fondement théorique.

Pour les aspects statistiques, il y a déjà quelques décennies, les organisations scientifiques et les revues en SHS&Santé ont publié depuis longtemps des directives pour s'assurer que les résultats statistiques aient tous un certain standard de qualité (Bailar & Mosteller, 1988; Wilkinson & *Task Force on Statistical Inference*, 1999; American Educational Research Association; American Psychological Association Publication Manual, 6e édition; etc.). Elles sont assez peu suivies (Ellis, 2013).

On a surtout retenu l'importance de publier les intervalles de confiance associés aux paramètres testés, et pas seulement les probabilités de dépassement. Ces intervalles de confiance bilatéraux, équivalents aux THSN, donnent une meilleure idée de la valeur d'un effet obtenu (Krantz, Berkson, 2003). On trouvera d'autres conseils chez Gibbons & Pratt (1975), et chez Ellis (2010).

Cependant, les enquêtes sur la reproductibilité initiées par Ioannidis exigent impérativement d'autres importants ajustements.

On recommande de plus en plus de publier même les études non significatives, ce qui serait un grand progrès car c'est seulement par des études répétées qu'on peut établir les propriétés en Sciences humaines et sociales et en Sciences de la santé (SHS&Santé). L'utilisation de la théorie des méta-analyses est sans cesse handicapée par ce *biais de publication*… Sans compter que pareil biais de publication est un incitatif à la tricherie. Les définitions et propriétés des protocoles expérimentaux devraient bien sûr être

© Société Française de Statistique (SFdS), Décembre/December 2015

déclarées dans les publications, mais surtout au préalable i.e. *ante hoc,* et les données devraient être accessibles.

On pourrait penser, à cet égard, développer un site internet, non modifiable une fois les informations enregistrées, pour regrouper les protocoles *ante hoc* des projets d'expérimentation statistique, i.e. les descriptions complètes des expériences statistiques, ainsi que les données recueillies et les analyses complètes *post hoc*. Il faudrait de plus, une fois les protocoles évalués et agréés, garantir aux auteurs des articles la publication de leur résultats, même pour ceux qui ne «montrent pas d'effet», i.e. même si $p_{dép}$ est trop grand pour être 'significatif'.

En aucun cas toutefois, les fraudes ne deviennent-elles impossibles… C'est l'honnêteté intellectuelle qui est en jeu ici. Et l'honnêteté n'étant pas la chose du monde la plus répandue…

On remarque que la controverse ne sévit pas beaucoup en dehors des SHS&Santé. Dans l'industrie par exemple, on utilise énormément des plans d'expériences parfois très complexes pour garantir les connaissances hypothétiques et pour progresser. L'inférence statistique classique y est indispensable.

L'élément qui revient le plus souvent dans les réactions à l'interdiction du BASP est que les connaissances statistiques dans les SHS&Santé sont très superficielles et lacunaires. Tout statisticien qui a pratiqué le beau métier de consultant en a une expérience de première main. Presser des boutons, c'est très commode, indispensable même, mais cela ne dispense pas d'apprendre la rationalité statistique dans ses fondements. Comment parvenir à augmenter le niveau des connaissances de la statistique en SHS&Santé qui constituent aujourd'hui les principaux champs d'utilisation, c'est tout un défi !

Références

On peut télécharger le fichier des références citées et quelques autres ici.

Remerciements

L'auteur tient à remercier Corinne Hahn, Gilles Stoltz et Catherine Vermandele pour leurs lectures attentives du premier jet de ce texte.

De R.A. Fisher (1935), le mot de la fin

« *Personally, the writer prefers to set a low standard of significance at the 5 percent point. […] A scientific fact should be regarded as experimentally established only if a properly designed experiment* rarely *fails to give this level of significance.* »

© Société Française de Statistique (SFdS), Décembre/December 2015