LA REPRODUCTIBILITÉ

L'INSTITUTION SCIENTIFIQUE EN CRISE

MARC BOURDEAU

« *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.* »
R.A. Fisher (1935)

Éléments essentiels de la théorie classique des tests statistiques sur paires d'hypothèse

Dans les tests sur hypothèse nulle, H_0, avec une contre-hypothèse, dite aussi l'hypothèse alternative H_a, unilatérale ou bilatérale, on distingue deux risques : ou rejeter à tort H_0 en misant sur H_a, c'est le *risque de première espèce*, ou accepter à tort H_0, c'est le *risque de seconde* espèce, donc ne pas miser à tort sur H_a.[1]

La procédure classique de Neyman-Pearson[2] consiste à déterminer, en considérant les deux hypothèses, une variable aléatoire construite sur un échantillon de la variable dont on veut tester un paramètre ; on appelle le *test*, cette *variable aléatoire* dite *échantillonnale*, car elle est une fonction d'un vecteur de variables aléatoires, un échantillon. Une fois réalisé l'échantillon on calcule le test sur les réalisations. Cette variable aléatoire, est une fonction des du vecteur échantillon détermine un critère de décision basé sur un risque maximal de première espèce α à ne pas dépasser pour décider du rejet de H_0. Le réalisation du test sur un échantillon donne une probabilité de dépassement[3] du test, $p_{dép}$, et le critère de décision est le suivant : on rejette H_0 lorsque $p_{dép} < \alpha$: $p_{dép}$ = P[*Rejeter* H_0 | H_0 est supposée vraie] < α. Plus $p_{dép}$ est petit, plus le risque de première espèce est petit. Le seuil admissible, α est le risque tolérable de rejeter à tort H_0 qui ne saurait être dépassé : $p_{dép}$ s'appelle la *rareté* de la réalisation du test. C'est pourquoi on rejette H_0 si $p_{dép} < \alpha$.

Tout test donc repose sur la paire α, déterminé *a priori* et *n* la taille échantillonnale pour l'échantillon de la variable dont on veut tester un paramètre. Mais on peut compléter en faisant jouer le risque de seconde espèce : β = P[Accepter H_0 | H_0 est fausse].

1

[1] Il y a un lien fondamental entre les concepts de probabilité, de cote d'un pari, et des mises et cotes à considérer pour qu'un pari soit honnête, c'est-à-dire d'espérance nulle. Ce qui n'est pas en général le but des parieurs… qui bien sûr parient pour gagner et non pour être 'honnête' ! On pourra trouver sous le lien suivant un fichier élémentaire sur les questions d'espérance de pari de Pascal et de l'appétence pour les loteries.

[2] On trouvera sous le lien suivant un court texte sur le paradigme de Neyman-Pearson, dit le paradigme classique ou fréquentiel des tests statistiques. Ce texte est presque sans équation.

[3] On appelle en anglais *p-value,* la probabilité $p_{dép}$. Mais bien sûr *p-value* est un terme sans signification claire. Le terme *value* n'as pas de sens clair, pas plus que le terme *cast* dans *pod-cast* (on dit 'balado', abréviation de balado-diffusion) au contraire de *broad-cast* sur lequel il est calqué, qui lui est bien nommé. $P_{dép}$ correspond bien au génie de clarté de la langue française.

Mais si H_0 précise la valeur du paramètre, ainsi $\mu = \mu_0$ un nombre donné, de sorte qu'on sait le sens précis de l'hypothèse nulle, partant à 'H_0 est vraie', on ne sait rien du sens à donner à 'H_0 est fausse'. Pour en donner un, il faut préciser la valeur du paramètre de l'hypothèse alternative : e.g. $H_a : \mu = \mu_1$: comme on a obtenu le critère de rejet de H_0 par la rareté tolérable α, on connaît par complémentation le critère d'acceptation de H_0 quel que soit *n*, on peut contrôler β par la taille échantillonnale. Si *n* est bien déterminé, on pourra contrôler β par l'équation :

$$\beta = P[\text{miser sur } H_0 \mid H_a : \mu_1 - \mu_0].$$

On nomme $\delta = \mu_1 - \mu_0$ l'écart à H_0, $\mu_1 = \mu_0 + \delta$. On a le pouvoir de détection du test, on définit ainsi sa puissance $1 - \beta$, qui est une fonction de *n* et α, pour un δ donné. Ainsi on pourra vouloir miser sur H_a : $\mu_0 + \delta$, dès que δ dépasse une valeur critique avec probabilité, e.g. $1 - \beta = 0{,}80$, ou $\beta = 0{,}20$, en se donnant une taille échantillonnale suffisante : β devient une fonction de *n*, δ et α. On appelle $1 - \beta$, la puissance, ou puissance de détection, du test qui demande la considération de ces trois paramètres.

Une expérience statistique bien définie, selon les normes en vigueur (e.g. Ellis [2010], Bailar the III & Mosteller [1988], **AERA**, **Wilkinson [1999] pour l'APA**, etc.) demande de définir la taille échantillonnale *n*, pour un α donnée, typiquement $\alpha = 0{,}05$, avec une puissance prescrite à un δ précisé. Inutile de dire que ces normes conçues pour améliorer la qualité des recherches ne sont pas souvent respectées (Ellis, 2010 puis 2013). Nous voici donc dans ce paradigme classique avec les 4 paramètres : α, β, δ, n et on a $\beta(\alpha, \delta, n)$.

Quelques mythes à débusquer

On peut penser réduire le risque de première espèce tolérable, α, mais en évitant Charybde, on tombe sur Scylla, car alors $\beta(\alpha, \delta, n)$, augmente avec la décroissance de α, (δ, n) étant égaux par ailleurs. On ne peut à la fois éliminer l'un et l'autre risque.

β diminue lorsque *n* augmente, α et δ étant fixés par ailleurs, mais les grandes tailles échantillonnales sont rapidement impraticables.

Détecter des petits écarts δ, disons à l'intérieur d'une fraction de l'écart type de la variable dont on veut tester un paramètre, devient rapidement là aussi impraticable.

Il faut se contenter de risques et de puissance pas trop grande. D'où la citation de Fisher en exergue. La tradition se contente d'un $\alpha = 0{,}05$ (5%) et d'une puissance $1 - \beta = 0{,}20$ (20%) pour un δ approprié. Qu'en est-il alors de la reproductibilité des expériences, i.e.la possibilité de retrouver sur des expériences indépendantes pour le même paramètre des $p_{dép}$ presque tous sous α. On reprend ici les termes de Fisher.

Il est presque impensable de répliquer des expériences statistiques, c'est évident. Question de ressources. Qu'en est-il de seulement 2 réplications ? C'est ce que nous verrons à la section suivante.

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 2015 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 ! Cette étude jette tout un pavé dans la mare. On apporte 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

C'est ainsi que 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.

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 tests sur hypothèse nulle (THN), 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 déclarées dans les publications, mais surtout au préalable i.e. *ante hoc*, et les données devraient être accessibles (Bailar the III & Mosteller, 1988).

On pourrait penser aussi développer un site internet, à déclaration obligatoire, 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 leurs 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… Rien n'empêche les chercheurs de publier des données trafiquées au préalable pour obtenir des résultats significatifs mais pas très… question d'assurer une publication, et de ne pas passer pour un fraudeur advenant une étude à l'effet contraire …qu'on ne risque pas beaucoup de voir tant qu'on ne publie que des études significatives. Publier ou périr !

On remarque que la non reproductibilité ne sévit pas beaucoup en dehors des SHS&Santé. Dans l'industrie de fabrication 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. Et pas question de tricher. Attention toutefois aux impératifs commerciaux dénoncés par Marcia Angell ainsi que dans le Oreskes & Conway, pour le *big Pharma*, l'industrie du tabac, des énergies fossiles, des produits chimiques agricoles… Les vertus se perdent. L'institution scientifique est bien en crise.

Références

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