LES TESTS STATISTIQUES SUR HYPOTHÈSES ANCIENNES ET NOUVELLES PRATIQUES

Marc Bourdeau

École Polytechnique de Montréal, C.P. 6079 Centre-ville
Montréal, Québec, H3C 3A7
Louis.Marc.Bourdeau@Gmail.com
http://wikistat.ca/

Résumé. Depuis sa mise au point dans les années 1930, le paradigme des tests sur hypothèses nulles (TSHN) a fait l'objet de controverses récurrentes. La dernière provient de la non-reproductibilité d'un grand nombre d'études en sciences humaines et de la santé. Plusieurs ont vu dans ce fait des failles logiques rédhibitoires aux TSHN. Après de longues consultations commencées en 2016, l'*American Statistical Association (ASA)* a fini par définir en mars 2019 les paramètres d'une nouvelle pratique des TSHN pour remédier quelque peu à la situation. Même s'ils ne sont pas totalement satisfaisants pour les fins de l'inférence, il n'y a pas d'erreur logique dans le paradigme des TSHN, les problèmes venant plutôt de l'institution scientifique elle-même qui incite aux tricheries de toutes sortes. Nous mentionnerons les recommandations radicales de l'ASA, et ses nouvelles prescriptions pour l'utilisation légitime des TSHN. Au passage nous illustrerons la fausseté de la solution préconisée par bon nombre de praticiens de réduire l'erreur de première espèce, comme si c'était la seule en jeu, à α=0,005 plutôt qu'au α=0,05 conventionnel. Selon l'ASA, le principal problème cependant réside dans l'enseignement des TSHN, où, justement, on fait trop souvent l'impasse sur l'erreur de seconde espèce, concept crucial pour déterminer les tailles échantillonnales requises, en tenant compte des tailles de l'effet requis. Nous présenterons le paradigme complet de NP des TSHN, et ce sans les équations qui font toujours problème dans notre enseignement du sujet. Nous montrerons comment présenter ce sujet en se fondant, dès les premières présentations de l'analyse des données, sur la rareté des réalisations d'événements incertains, et sur la pratique des parieurs en présence d'incertitude. Nous référons également à un site de documents courts et motivants.

Mots-clés. Fisher, Neyman-Pearson, tests statistiques sur hypothèses nulles (TSHN), directives de l'ASA, nouvelles pratiques, rareté, puissance, taille de l'effet.

Abstract. The replication crisis is what prompted the ASA to go back to the drawing board and give some new directions for the use of the standard Neyman-Pearson paradigm for statistical inference. We will indicate some new ways to present the basic concepts for Probability and Statistics, based on our own teaching experience.

Keywords. Null hypothesis statistical testing, Fisher and Neyman-Pearson paradigms, ASA 2019 statement, power, effect size.

There is no true value of anything.

(W. Edwards Deming[1])

1 Introduction. Le paradigme standard des TSHN.

Fisher. De tout temps, on a cherché à prédire l'avenir, à inférer ou induire l'inconnu à partir du connu. Dans les années trente du siècle précédent, la statistique fut mise à contribution pour les questions d'inférence scientifique. R.A. Fisher d'abord a inventé la statistique moderne, elle repose sur la théorie des probabilités, pour ne pas rejeter ou rejeter une hypothèse (et une seule), essentiellement sur la valeur d'un paramètre d'une loi en se fondant, grâce au théorème central de la limite, fondamental en inférence statistique, sur la rareté d'une réalisation d'une loi échantillonnale, i.e. construite sur un échantillon d'une variable aléatoire appropriée. Cette loi définit un test statistique. La rareté en *question* est définie par une probabilité, dite en anglais la *p-value,* qu'on veut assez petite pour être suffisamment sceptique sur l'hypothèse postulée pour ne pas l'admettre. Cette probabilité est celle d'obtenir un faux positif, i.e. rejeter à tort l'hypothèse. Elle est définie invariablement donc par une probabilité de dépassement qu'on veut petite, d'une statistique de dépassement. Plutôt que d'utiliser le calque de l'anglais p-valeur[2], nous préférons noter *cette probabilité de dépassement* par $p_{dép}$. Pour la rareté de la statistique, Fisher, Fisher a proposé $p_{dép}<0,05$.

> *« Personally, the writer prefers to set a low standard of significance at the 5 percent point. A scientific fact should be regarded established only if a properly designed experiment* rarely fails *to give this level of significance. »* (Fisher, 1926, p.504)

On verra plus loin que le 0,05 est devenu un nombre quasi-sacré, mais qu'il est presque impossible de respecter la reproduction des études… Tout le problème vient de là.

Neyman-Pearson. Le paradigme de Fisher fut complété par Jerzy Neyman et Egon Pearson. Ils ont introduit une deuxième hypothèse, dite l'hypothèse alternative, la première devenant l'hypothèse nulle. Cela en réalité est tout à fait naturel. Mais Fisher n'y a pas pensé. Ainsi pour prendre l'exemple simple d'une hypothèse nulle sur une moyenne μ, il y a 3 alternatives possibles :

$$H_0 : \mu=\mu_0 \text{ vs } H_1 : \mu< \mu_0 \text{ (ou } \mu>\mu_0 \text{ ou } \mu\neq\mu_0).$$ [3]

Les deux hypothèses fixées dépendent des contextes et sont tout à fait naturelles : si on n'admet pas la première c'est en faveur de la seconde et réciproquement,[4] donnent lieu maintenant à deux erreurs possibles de jugement par rareté : la première dite maintenant l'erreur de première espèce, celle de rejeter à tort H_0, admettre H_1, sur la base d'un *faux positif* observé, ou de ne pas rejeter à tort H_0, et sur la base d'un *faux négatif* observé. Les calculs donnent des propriétés d'optimalité au paradigme de NP, qui prescrit la statistique du test grâce au fameux lemme de Neyman-Pearson. On retombe souvent sur les statistiques intuitives, parfaitement raisonnables et naturelles, celles utilisées par Fisher.

Cette deuxième hypothèse complique notablement les choses. Quand on se donne la probabilité (petite) maximale α du risque de première espèce : on rejette H_0 si $p_{dép} < \alpha$, et typiquement bien sûr, α =0,05, cette valeur hélas est quasi sacrée…, on peut alors contrôler l'erreur de seconde espèce, notée β, à un écart prescrit de la valeur du paramètre fixée dans H_0. C'est la taille de l'échantillon

[1] Dans la nouvelle préface au livre de Walter A. Shewhart, 1939 (1986), «Statistical method from the viewpoint of quality control», New York NY : Dover Publications.
[2] Qui ne signifie rien, pas plus en anglais d'ailleurs.
[3] Bien entendu, une seule des trois alternatives est possible dans cette paire d'hypothèses.
[4] À noter qu'on pratique beaucoup la litote en statistique : on ne parle pas d'admettre une hypothèse mais de ne pas la rejeter…

pour la statistique appropriée qui va permettre ce contrôle. L'écart admissible de détection de la valeur postulée du paramètre testé s'appelle la taille de l'effet : il est mesuré en termes de proportion de l'écart type (présumé connu ou estimé sur un échantillon préliminaire). Dépendamment de cet écart, ne pas commettre une erreur de seconde espèce, être assuré donc de pouvoir le détecter (i.e. ne pas accepter l'hypothèse nulle) avec une probabilité 1-β élevée, typiquement 1-β = 0,80 ou 0,90, prescrit la taille échantillonnale de l'expérience statistique. Le 1-β s'appelle la puissance du test, ou la probabilité de détection d'un certain écart à la valeur de l'hypothèse nulle. Le 1-α s'appelle l'efficacité du test. Ledit lemme de Neyman-Pearson prescrit la statistique test à utiliser. Contrairement à ce qui se passe pour le paradigme de Fisher où règne de ce côté, intuitif tant qu'on voudra, un certain arbitraire.

On remarquera aisément la complexité qui découle de NP en relisant quelques fois si nécessaire, les paragraphes qui précèdent qui manquent peut-être de clarté...[5]

Fisher n'a jamais admis les perfectionnements de NP, qui est rapidement devenu la norme en matière de TSHNEt comme il avait très mauvais caractère, la chose a dégénéré en querelle féroce qui ne s'est jamais apaisée (Bourdeau, 2015a ; Droesboeke & Tassi, 2015 ; Lehman, 2013). rapidement devenu la norme en matière de TSHN. D'autant plus qu'il avait mis au point aux mêmes fins les statistiques fiduciales, complètement oubliées aujourd'hui et guère jamais utilisées au profit du paradigme NP.

2 La querelle continuelle

La complexité du paradigme standard, celui de Neyman & Pearson (NP), a cependant donné lieu à toutes sortes de mauvaises interprétations. Notamment en ce qui concerne le $p_{dép} < \alpha = 0,05$ qui entraîne la non acceptation de l'hypothèse nulle.

La difficulté de la compréhension par les étudiants et les applicateurs, impose finalement des recettes simples. Et avec les valeurs socio-politiques de l'institution scientifique, notamment le nécessité de publier pour survivre comme chercheur ('publier ou périr'), avoir des promotions et des subventions, a imposé la dichotomie, non voulue par Fisher, marquée par α=0,05, devenue une norme quasi sacrée, favorise toutes les tricheries : on ne peut en effet publier si $p_{dép} > \alpha$ ...

Les controverses sur les TSHN ont commencé presque à la naissance du paradigme. Peut-être à la suite de la querelle féroce de la naissance. Il y a plus de 20 ans maintenant (amplifié par, e.g., Ioannidis, 2007) on s'est aperçu des difficultés à reproduire les études publiées (Nosek, 2015). Diverses études à cet égard ont été publiées depuis 20 ans. Bourdeau (2015b) en rapporte un bon nombre où on voit bien le problème : l'impératif du *rarely fails* de Fisher n'est pas toujours respecté... C'est un fait, on ne réplique quasiment jamais les études publiées …pas plus que les autres et pour cause, le $p_{dép}$ étant supérieur à 0,05, elles ne sont pas publiées! Remarquons bien sûr que $p_{dép}$ est une statistique, calculée sur résultat d'une variable aléatoire échantillonnale, obtenue *une seule fois* par échantillon : sans réplication, il n'y a forcément qu'une seule réalisation Sa loi, moyenne et variance, sont inconnues…

La dernière crise des TSHN se produisit quand, à la suite de nombreux appels, David Trafimow et Michael Marks (2015) dans un éditorial de la revue *Basic and Applied Social Psychology,* dont Trafimow est le rédacteur en chef, ont franchi le pas et annoncé qu'il refuserait dorénavant tous les articles où on utilisait les TSHN. Ça alors! Mais où va-t-on sans inférence? Quoi? Se limiter aux statistiques descriptives! Cette prise de position jeta la communauté des statisticiens dans une grande

[5] Les énoncés de proba-stat sont souvent des exercices de lecture fine !

perplexité. Plusieurs revues avaient pensé le faire, c'est Trafimow qui l'a fait, cela risquait d'être contagieux! L'*American Statistical Association (ASA)* directement interpelée fut forcée de réagir (Wasserstein & Lazar, 2017). Elle organisa à Washington un symposium sur la question en 2017, le SSI (*Symposium on Statistical Inference*), où une bonne partie des gens concernés, parmi lesquels d'éminents statisticiens, prirent position. On décida alors de former un comité élargi pour réfléchir et prendre position officielle. Le SSI ne se déroula pas sans quelques perturbations, mais cela était dû non pas à des statisticiens, gens plutôt pacifiques et très ouverts aux discussions, mais aux circonstances politiques où le parti Républicain, nouvellement élu aux USA, a voulu imposer ses vues et détourner la science de sa mission de recherche de la vérité![6] C'était ridicule, mais très significatif (sans jeu de mots…) du nouveau climat politique aux USA.

3 L'ASA intervient

Le résultat de ce brassage d'idées fut un numéro spécial du *The American Statistician,* avec une nouvelle Déclaration (en anglais : *Statement*) de Wassertein &al. (Wasserstein, Shirm, Lazar, 2019), suivi d'une quarantaine d'articles faisant le tour de la question (Ils sont tous en accès libre). Ce nombre imposant d'articles sur le seul sujet des TSHN montre que la situation est fort complexe…

Voici le résumé des recommandations de l'ASA aux chercheurs.

- Ne basez pas vos conclusions sur le seul fait qu'une probabilité de dépassement ait franchi un certain seuil arbitraire de signification, e.g. $p_{dép}<0,05$.
- Ne croyez pas qu'une association, un effet existe, du seul fait qu'un seuil de 'signification statistique' ait été atteint.
- Ne croyez pas qu'une association, un effet n'existe pas, du seul fait qu'un seuil de signification n'a pas été atteint.
- Ne croyez pas que le hasard uniquement vous donne une $p_{dép}$ significative ou non.
- Ne concluez rien sur l'importance scientifique ou pratique basé sur une pseudo signification statistique ou son absence.

Nous voilà prévenus. Il est hasardeux de se fier du hasard... Ne jamais oublier que $p_{dép}$ est la réalisation unique d'une loi échantillonnale. De plus, les conditions prérequises aux TSHN sont rarement vérifiées. Tout un chacun est en mesure de comprendre ces choses à éviter. Elles tombent sous le sens pour tous ceux qui ont une *pratique* de l'analyse des données. Avant tout, c'est la principale recommandation explicite de l'ASA, on ne doit jamais dichotomiser les tests entre 'significatifs' et 'non significatifs' avec la frontière du sacro-saint α=0,05. Ou quelque variante : ainsi, les notations *p* ou p**, etc.* Cet empêchement à la publication est une cause importante de tricherie et de non-reproductibilité des études. Fini!

Ce sont là les choses à ne pas faire, c'est tout du négatif, mais les articles du numéro spécial du TAS proposent des choses positives en ce qui concerne l'inférence statistique, dans une quarantaine de papiers! L'article de Wasserstein, Shirm & Lazar (2019) en fait le résumé. Il se penche également sur ce qu'est la pensée statistique; les qualités d'ouverture aux applications et à leurs experts; sur la modestie que doivent manifester les analystes de données.

[6] Je relate dans « Un incident au symposium » (Bourdeau, 2017) ces quelques péripéties.

Il insiste en conséquence sur les changements que doivent apporter les revues à leurs règles de publication de la partie statistique des articles. Et il insiste surtout sur le fait que la transmission déjà fort complexe des techniques de l'inférence statistique devrait s'ajuster. Il faut que la compréhension des TSHN augmente : la compréhension fine du paradigme NP devra être assurée. Tout un programme!

Avant tout, l'ASA recommande aux revues de publier les résultats des expériences statistiques quelles que soient les valeurs des $p_{dép}$, pourvu qu'elles soient compétentes et bien menées, c'est entendu, et de bannir toute référence au α=0,05 (ou même un autre *fil rouge*). Le terme à éviter dans les publications: « statistiquement », ou « non statistiquement significatif. » Toute dichotomisation empêche de *penser statistiquement, scientifiquement*. C'est clair!

Comme l'a souligné D. Donoho, peut-être le plus éminent statisticien américain de l'heure : il n'y a pas d'erreur de logique dans le TSHN qui continuera d'être largement pratiqué, notamment en santé et en sciences appliquées, le problème vient de l'institution scientifique en elle-même qui incite à la tricherie (Donoho, 2015). C'est particulièrement le cas des compagnies pharmaceutiques actuelles, malgré les grandes avancées que la pharmacopée a permis par le passé (St-Onge, 2013, ou sa présentation récente « L'influence des lobbies sur la science »). La science moderne est malade. L'argent la pourrit. Mais comment changer tout ça ? La voie sera longue et ardue.

4 Quelques conséquences pratiques

Les nouvelles directives aux auteurs et aux périodiques. Il n'y a pas lieu de parler de signification statistique et d'une dichotomisation définie par le α, au-delà ou en deçà tout changerait pour la signification statistique. Fini ! Ne plus mentionner le terme de signification statistique. les revues devraient publier les articles, pourvu qu'ils bien construits, quelle que soit la valeur de $p_{dép}$ obtenue des expérimentations statistiques, elles devraient adapter leurs directives aux auteurs (ce qui semble bien être le cas : e.g. Amrhein &al, 2019 ; Colquhoun, 2017 ; Kraemer, 2019).

Reproductibilité — Répétabilité — Réplicabilité. Nous sommes ici au cœur de la science : pas de réplication, pas de science (Kenett & Shmueli, 2016 chap. 1 ; Stevens, 2017). Comme le soulignent Kenett & Shmueli, 2015; Stevens, 2017), il y passablement de confusion dans ces termes, dont les acceptions dépendent des contextes. Il convient de clarifier un peu les mots et les choses.

- *Reproductibilité.* C'est la capacité générale de répliquer une étude pour arriver aux mêmes conclusions et intuitions, pour d'autres populations, pas nécessairement avec le même schéma expérimental;
- *Réplicabilité.* C'est la possibilité de retrouver les résultats sur d'autres données avec le même schéma expérimental, sur la même population;
- *Répétabilité.* C'est la capacité de reproduire les mêmes résultats, sur les données originales, par d'autres analystes.

Ainsi ce qui diffère dans l'une et l'autre définition, c'est le degré de généralisation attendu.

Ce n'est pas d'hier que cet aspect de la statistique nous préoccupe : e.g. Bourdeau (2015b, de même que Bourdeau (2017), ou encore plus longuement, toujours 2017), où nous décrivons le temps le plus fort de la crise de la réplication des études qui a justifié la revue *Basic and Applied Social Psychology,* par son rédacteur en chef, à refuser dorénavant tout article utilisant la théorie des TSHN. Ce qui a fait trembler l'ASA. Par ailleurs, la prestigieuse revue Science (Nosek &al. 2015) a répliqué 100 études *statistiquement significatives* au sens du α=0,05, sans retrouver ce résultat dans plus de 60% des cas! Cette étude et d'autres ont fait trembler l'ASA !

Donoho (2015), de même que Stevens (*op.cit.*) remarquent judicieusement que c'est l'institution scientifique qui est en cause, et non le paradigme des TSHN tel que développé par Neyman & Pearson.

Quoi qu'il en soit, le *fil rouge* du α=0,05, la ritualisation de la pratique des tests et la malhonnêteté des chercheurs pris dans la nécessité de publication les soumettent à des dilemmes cornéliens, dont ils ne sortent que par la petite porte... Les directives de publication de la Déclaration de l'ASA (Wasserstein &al., 2019) sont susceptibles de régler en grande partie cette crise de la reproductibilité des études faisant appel aux TSHN. Si elles sont mises en actes…

La fonction du statisticien consultant. Le statisticien va collaborer avec les scientifiques impliqués dans les études pour réfléchir au sens à donner aux résultats… Ce sera une tâche ardue pour laquelle le statisticien devra avoir une écoute assidue et des connaissances au moins minimales des champs où il s'applique, avec une certaine …modestie, dit l'ASA. Des connaissances minimales de la statistique devront être transmises adéquatement aux scientifiques impliqués. Une révolution donc, c'est ce que définissent les nouveaux cursus de Science des données. Mais la conception des expériences statistiques ne doit pas vraiment changer, ce que ne souligne pas assez le *ASA-Statement* de 2019. On a toujours les erreurs maximales tolérables de première et seconde espèce comme contraintes. Les conceptions des expériences statistiques se font sur ces bases.

Ainsi le α=0,05 peut encore être considéré comme raisonnable (ne pas oublier que des réplications devraient être pratiquées !), et β, l'erreur de seconde espèce, pourrait encore être fixée à à 0,10 ou 0,20, donc le pouvoir de détection de 80% ou 90% pour une taille de l'effet désirée (Ellis, 2010), laquelle devra tenir compte de l'application, de la discipline de l'expérimentation. Cela relève des compétences de l'expert avec lequel travaille le statisticien d'applications.

La connaissance complète du paradigme des TSHN, notamment en ce qui concerne l'erreur de seconde espèce (la puissance) ne devra plus relever du mystère comme maintenant ! Les TSHN devront faire partie des cursus de base. Les calculs nécessaires de détermination de la taille échantillonnale requise pour satisfaire les contraintes restera en général du domaine du statisticien. De même que la détermination du modèle requis, mais le pourquoi devra faire partie des connaissances de base de tout utilisateur de la statistique.

Une fois le design expérimental déterminé, il faut savoir déterminer quelle est la population étudiée, le schéma d'échantillonnage, randomiser les sujets de l'expérimentation, saisir les données, les analyser, ce qui est tout un art, et conclure… Et penser aux réplications ! Rien de simple dans tout cela (voir par exemple Hahn & Meeker, 1993). Le statisticien professionnel a encore de beaux jours devant lui.

La prépublication des devis et des données. Il y a longtemps qu'on parle de prépublier les devis expérimentaux ainsi que les données une fois la cueillette réalisée (e.g. Nosek &al, 2018) . Bien sûr dans des sites infalsifiables. Au premier regard cela semble relever du bon sens et de l'honnêteté.

Mais la réalité est plus complexe. Va pour les devis expérimentaux, mais les données peuvent toujours être trafiquées, après analyses préliminaires non encore publiées, pour les amener à réaliser des probabilités de dépassement, $p_{dép}$ 'raisonnables'. D'autre part, quel chercheur voudrait laisser ses données souvent fort complexes à cueillir, de plus en plus lourdes, entre de mains éventuellement peu professionnelles ou même ennemies ? Les entreprises et industries ne sont pas toujours des parangons d'honnêteté… Les données massives demandent un travail considérable, des soins infinis. On parle maintenant d'entrepôts de données de téraoctets!

Les données 'sensibles' sont en effet souvent sujettes à des controverses impossibles à arbitrer, dépendent pour leur établissement (cueillette et mise en forme) de bien des valeurs politico-idéologiques (voir la note 4 de cet article). Les vertus se perdent, les valeurs d'argent sont souvent prépondérantes… Non! Il faut répliquer, i.e. recommencer *ab nihilo.* Vaste programme!

La réduction du alpha à 0,005. Voilà qui peut sembler une solution évidente au problème de la reproductibilité : réduire la probabilité conventionnelle de trouver un faux positif, i.e. par exemple réduire le α à 0,005 préconisé par Benjamin & al. (2018). Cet article a circulé sous diverses formes, telle une traînée de poudre : qui ne l'a pas reçu ? Ce serai la panacée ! Mais le problème principal de cette solution, est que la réduction de l'erreur limite acceptable de première espèce, α, entraîne automatiquement, la taille échantillonnale étant égale par ailleurs de même que la contrainte de détection d'un écart à l'hypothèse nulle fixée (soit une valeur prédéterrminée de la taille de l'effet), à une augmentation de l'erreur de seconde espèce, β, donc une diminution de la puissance 1-β du test. Si on veut conserver la même puissance, il faut augmenter la taille échantillonnale… jusqu'à des tailles souvent impossibles pratiquement.

Ainsi pour prendre un exemple fréquent et facile à calculer : on veut tester une hypothèse nulle sur une moyenne avec variance connue ou estimée, avec alternative unilatérale, et β=0,1 (la puissance à 90%), le passage de α=0,05 à α=0,005 demande une augmentation de près de 50% de la taille échantillonnale. Ainsi pour un *d* de Cohen moyennement exigeant, par exemple *d*=1/4 (détecter un écart d'un quart de σ de la moyenne) de N=168 à N=249 échantillons ou sujets de l'expérience, et pour *d*=1/5, un peu plus exigeant de N=265 à N=380.[7] Ce qui augmente considérablement les coûts d'une étude, ralentit la recherche. Quand on a aidé de nombreux thésards, on comprend que cela est trop exigeant. Nombre d'études sur questionnaires sont faites avec des échantillons de tailles aux environs de N=100 dans les cas des études en sciences infirmières que nous avons souvent pilotées, ce qui déjà apparaît assez rédhibitoire. Il faut convaincre les chercheurs pour arriver à ce nombre. Beaucoup de champs d'études, sinon la plupart, commencent par des thésards qui sont pressés dans tous les sens du terme, comme on sait ! Ils sont dans l'exploratoire des études préliminaires qui demanderont à être éventuellement confirmées, raffinées …répliquées.

De surcroît donc, dans cette éventualité, la 'nécessité' de tricher augmente aussi ! Comment la suggestion de Benjamin aiderait-il à reproduire les études, pour en augmenter la fiabilité ?

Non : quant à nous, nous nous rapportons à la citation de R.A. Fisher rapportée plus haut : un α peu exigeant à 0,05, mais des réplications. Pas de réplications, pas de science !

[7] Notons que les contraintes sur les tailles des effets, dépendent du contexte des études. Elles reposent sur la signification dite pratique plutôt que statistique, et demande des connaissances assez profondes des disciplines et de l'étude concernées.

Élaborons quelque peu sur le problème de réduction de l'erreur de première espèce tolérable, α. Dans notre enseignement élémentaire à de futurs ingénieurs, nous développons complètement les TSHN/NP sur l'exemple paradigmatique du test unilatéral à droite sur une moyenne avec variance σ connue, et on fixe les idées sur un exemple simple (contexte). Tous les tests sur paramètres se conjuguent plus ou moins selon les mêmes lignes :

$$H_0 : \mu=\mu_0 \quad vs \quad H_a : \mu > \mu_0.$$

Le calcul (simple) donne pour l'erreur de seconde espèce β à contrôler, qui dépend de trois paramètres, $\mu=\mu_1$, N, α, où N est la taille échantillonnale, α l'erreur tolérable de première espèce et μ_1 la moyenne à détecter avec une certaine probabilité, i.e. la puissance du test 1-β désirée à cette valeur, qui peut être obtenue en contrôlant la taille N :

$$\beta(\mu_1, N, \alpha)= P[Z\leq (\mu_0 - \mu_1)/ \sigma * \sqrt{N} + z_\alpha],$$

puis, comme on écrit maintenant, surtout dans les SHS et de la santé, l'écart normalisé et non signé $|(\mu_0 - \mu_1)| / \sigma = d$ (soit le d de Cohen [Ellis, 2010]), la taille de l'effet minimal détectable désiré, soit $|\mu_0 - \mu_1|$ en proportion de l'écart type.[8] On lira au Tableau 1, les effets de la variation un à la fois des paramètres sur l'erreur de seconde espèce, β, partant de la puissance 1-β. Pour Cohen, rapporté dans Ellis (2010), un effet est considéré petit si d≤0,2. On constatera à la Fig. 1 que ces détecter de tels petits effets, des faux négatifs donc, est assez difficile avec les échantillons courants d'une centaine de sujets dans bien des études en SHS/Santé…

Tab. 1— Pour un test sur une moyenne d'une variable aléatoire avec variance connue, test unilatéral à droite. Les erreurs de seconde espèce et la puissance en fonction des paramètres N, la taille de l'échantillon ; d, la taille de l'effet à détecter ; α l'erreur de première espèce tolérable. En fonction des variations d'un seul des trois paramètres à la fois.

Fixés	Si	Alors	Alors
d, α	N ↑	β ↓ 0	Puissance ↑ 1
N, α	d ↑	β ↓ 0	Puissance ↑ 1
N, d	α ↓ 0	β ↑ 1	Puissance ↓ 0

Et on peut tirer de l'équation pour β la figure suivante (Fig 1).

[8] On notera que sur le paradigme illustré ici, le $(\mu_0 - \mu_1)/\sigma$ est négatif, car $\mu_1 > \mu_0$, et ce de plus en plus avec la croissance de *d*.

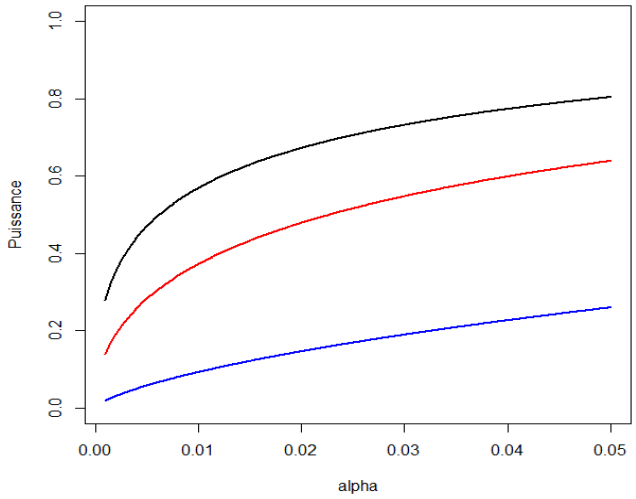

Fig. 1— Test sur une moyenne, unilatéral à droite, variance connue. La puissance du test pour les alpha (α) de 0,001 à 0,05. Pour N=100, et de haut en bas pour *d*: *d*=0,25, 0,2, 0,1.

En général, ce sont les tailles échantillonnales qui contrôlent la puissance d'un test à un *d* de Cohen et un α donné. Selon Ellis (2010, Tableau 2.1, p.41 ; d'après Cohen, 1988), les petits effets de taille sont pour des d≤0,20, i.e. à des écarts d'un cinquième d'écart type de la valeur de l'hypothèse nulle. Évidemment cela dépend des contextes et des objectifs des expérimentations. Pour obtenir une probabilité de 80% de détection des faux négatifs à *d*=0,20, (une puissance de 80% qui semble la norme dans bien des applications), il faut des tailles échantillonnales conséquentes… On trouvera à la Figure 2, l'effet de la croissance de N, pour *d*=0,20 selon la plage [0,005 ; 0,05] de α.

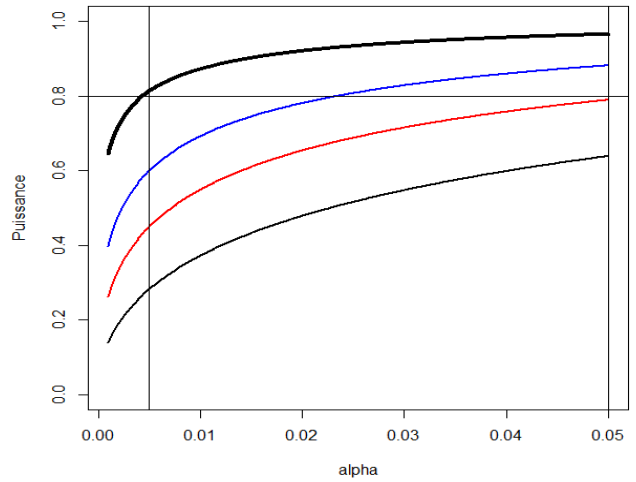

Fig. 2 — Pour *d*=0,20, les puissances pour N=100, 150, 200, 300, de bas en haut : on retrouve le graphique en rouge de la Fig. 1 avec ici le graphique du bas. Le croisement des lignes verticale à α=0,005 et horizontale à la puissance 80% montre qu'avec N=300, on atteint les objectifs de 80% de détection des faux négatifs lorsque la probabilité de détection des faux positifs (1-α) est de 99,5%. Alors que le graphique en rouge où N=150 montre que la puissance de 80% est atteinte pour un α=0,05, ou (1-α) = 0,95 qui est la probabilité de détection de faux positifs.

Chez les ingénieurs, on trouve des abaques de puissance ou d'erreur de seconde espèce pour déterminer sans calculer les tailles échantillonnales suffisantes pour détecter les faux négatifs avec des puissances déterminées avec des α, ou des probabilités de détection des faux positifs (1-α) conventionnels de 95% ou 99%. On remarque que le paradigme de NP est parfaitement logique et couvre l'ensemble des *desiderata* raisonnables pour une bonne théorie des TSHN. À cette nuance près qu'il n'y a pas de certitudes… mais que des probabilités. À commencer par la valeur de $p_{dép}$ qui est la réalisation unique d'une variable aléatoire échantillonnale sur chaque échantillon. On se rapportera aux mises en garde de l'ASA (Wasserstein &al., 2019) citées plus haut. Pas d'erreur de logique, soit, mais cela renvoie encore et toujours à des problèmes dans l'institution socio-politique de la science…

On remarquera aussi dans ce qui précède la difficulté technique pour les utilisateurs 'lambda' de la statistique. Et pourtant ces choses sont élémentaires.[9] Comme on sait par ailleurs, que la statistique est de plus en plus nécessaire dans toutes les sciences, humaines, sociales et de la santé en particulier, et qu'elle se fait avec de moins en moins de statisticiens — les machines ne font-elles pas tout maintenant, on a de plus en plus affaire en effet à une statistique robotisée ! —, il y a de quoi être assez sceptique sur la qualité des devis, de leurs réalisations et de leurs analyses !

C'est pourquoi le dernier *ASA-Statement* (Wasserstein, Shirm & Lazar, 2019) insiste beaucoup sur la transmission des compétences statistiques de base. Et plusieurs des articles du supplément du TAS s'y adressent. Nous traitons trois aspects de la question.

5 L'éducation à la pensée statistique

Quiconque a enseigné la statistique élémentaire, notamment aux sciences appliquées comme moi, quiconque a travaillé pratiquement sur des données avec des chercheurs dans les domaines d'applications perçoit sans peine la difficulté à enseigner les éléments les plus simples de l'inférence statistique. On parle de l'opacité cognitive des TSHN (Calin-Jageman & Cummings, 2019) : c'est très contre-intuitif d'admettre que des $p_{dép}$ petits sont désirables; et pourquoi espérer que l'hypothèse nulle soit fausse; comment expliquer ce mystérieux nombre sacré de 0,05 ; comment d'ailleurs poser les deux hypothèses ? ; les deux types d'erreurs sont sujets à beaucoup de confusion. Bien difficile à comprendre tout cela!

On comprend donc pourquoi tant d'enseignants font l'impasse sur l'erreur de seconde espèce cruciale dans tout *design* expérimental. La puissance des expérimentations statistiques laisse souvent à désier.

Chez les enseignants on aime couvrir beaucoup de surface, enseigner du presse-boutons sur des méthodes complexes (et ne pas beaucoup en tester les apprentissages!), ça fait savant! plutôt que de faire comprendre profondément les éléments simples, qui mènent au reste.

Notre expérience d'enseignant et de consultant nous a amené à quelques procédés pédagogiques dont nous vous faisons part ici. Il importe de faire comprendre les éléments simples en vue de l'apprentissage fin des éléments des TSHN.

[9] N'est-il pas vrai que tout est toujours plus compliqué qu'il n'y paraît ? Et ce même pour les choses qui à première vue sont très élémentaires. À feuilleter le supplément du TAS, on a un bel exemple de la pertinence de ce dicton… Tout cours de base de statistique devrait traiter de ces choses en profondeur, quitte selon les publics à ne pas faire les calculs, seulement illustrer/interpréter les résultats.

L'incertitude de tous les systèmes de mesure. La rareté. En fait, l'incertitude sur tout… L'enseignement de notre cours de base commence par une expérience réalisée par l'ensemble de la classe: mesurer des clous (!) « Quelle est la longueur des clous de 6,35cm? ». En même temps que se déroule cette expérience, nous enseignons les principaux outils descriptifs de données, et celles de l'expérience sont rapidement accessibles.

On peut remplacer cette expérience par la détermination de la précision obtenue par nos doigts pour arrêter le chronomètre d'un téléphone intelligent à précisément 10 secondes (cette loi n'est pas symétrique, contrairement à la précédente).

L'étudiant conçoit ainsi, viscéralement, l'incertitude de toute mesure ; à percevoir les facteurs dont dépendent leur précision, soit les sources d'erreurs de tout système de mesures ; que tout ensemble de données donne naturellement lieu à une fonction de densité (de poids), ici sur la droite, qui est une introduction au concept de variable aléatoire ; que pour les densités unimodales (continues ou non) la rareté d'une réalisation se définit non pas en l'atteinte d'une valeur donnée, mais à la probabilité/proportion des observations qui dépasse cette valeur sur la densité présumée. On prépare ainsi les étudiants aux concepts fondamentaux des TSHN. En fait, au concept intuitif de ce qu'est une variable aléatoire, au concept de rareté, à celui de données suspectes.

Les paris ont la cote. Un parieur invétéré a l'habitude de prendre des paris avec des cotes : supposons qu'il offre X unités de monnaie à son contre-parieur s'il perd, dont il obtiendra Y unités de monnaie s'il gagne : il prend un pari avec la cote X/Y.[10] Ainsi il offre du 3 pour 1 sur le pari donné. Normalement dans les cas les plus fréquents Y=1. On trouve rarement du 3 contre 2 par exemple…

Faisons le lien qu'un parieur fait rarement avec les probabilités. Supposons que le parieur sait que son pari a une probabilité p d'être gagnant. Son espérance de gain est alors de $Yp - X(1-p)$. Il sera gagnant si cette espérance est non négative :

$$Yp - X(1-p) \geq 0.$$

Le pari est dit *honnête* si l'égalité est satisfaite. On tire de cette équation que le pari sera honnête si $X/Y = p/(1-p)$. Pour un probabiliste, cette valeur s'appelle la cote de p (*odds* en anglais). Un bon parieur est forcément malhonnête, il veut gagner ses paris. S'il raisonne en termes de probabilités, il cherchera à prendre des paris tels que $X/Y<p/(1-p)$. Ainsi si $p=4/5$, il cherchera un contre-parieur prêt à accepter une cote inférieure à 4 pour 1.

Dans le cas des TSHN, le $\alpha=0,05$ correspond à une cote de 19 pour 1. À se tromper, à parier avec cette cote donc, sur l'hypothèse alternative, il prend un risque (de 1re espèce) inférieur à $p=\alpha$. Avec $\alpha=0,005$, la cote est de ≈200 pour 1. Pour prendre un pari avec une pareille cote, il faut être passablement certain de ne pas se tromper à rejeter sans erreur l'hypothèse nulle! Tout parieur qui pense à un pari avec une telle cote n'aura pas de difficulté à offrir un alléchante cote, et à trouver un bon poisson…

Il arrive que montrer à parier sur des probabilités aux étudiants peut leur donner des envies d'être plus rationnel, plus savants dans leurs paris… car qui ne prend pas de paris? Parier avec des cotes est tout de même plus savant que de s'en tenir à la cote habituelle de 1 pour 1… Cet apprentissage est motivant pour beaucoup.

[10] Rares, évidemment, sont les parieurs qui raisonnent en termes de probabilité. Ils intuitionnent leurs cotes intuitivement et cherchent à gagner leurs paris… Ils sont forcément malhonnêtes! Du filoutage—*a con game, that's the name of the game* —. De toute façon, la définition fréquentielle de probabilité d'un événement donné parmi un nombre fini de possibilités demande de réaliser une 'expérience' aléatoire de nombreuses fois et de calculer la proportion des fois où on a réalisé l'événement donné. Ce qu'on n'a pas toujours la possibilité de faire avant de prendre un pari !..

Un véritable défi : enseigner les proba/stat sans équations. Les progrès de la technologie ont rendu les étudiants paresseux : beaucoup plus qu'auparavant sont devenus très ignorants en ce qui concerne la manipulation mathématique, voire la simple compréhension du sens à donner à une équation! C'est le cas même des futurs ingénieurs à qui j'ai enseigné pendant des décennies. A fortiori pour les disciplines des science humaines et sociales qui ont constitué mes champs d'application!

Les techniques de visualisation sont évidemment indispensables (Grant, 2019). Et les logiciels ont permis bien des développements de ce côté. Mais tout de même, vient un moment où on doit trouver des façons de faire comprendre les choses en mots plutôt qu'en équations. Nous avons écrit plusieurs textes avec un minimum d'équations au sujet des TSHN. Voici des liens pour certains de ces textes écrits à l'occasion de consultations : le dernier en date ; un autre aussi un peu différent. Ils pourraient être utiles.

Le cours de l'histoire et de la société. Motiver les étudiants. Il est très difficile de motiver les étudiants à comprendre la finesse de proba-stat, surtout dans les cours d'applications qui comptent pour plus de 80% de notre enseignement. Les logiciels sont très puissants, les statisticiens professionnels sont plutôt rares. La statistique est robotisée !

Le grand pédagogue Neil Postman (dans une série impressionnante d'essais, e.g. 1995, 1999) a prôné un retour de l'enseignement aux valeurs de l'humanisme du XVIII^{e} siècle et des Lumières, surtout françaises.[11] Au vu de ce qu'on a vécu depuis le début du siècle, qui pourrait s'y opposer ? On pense que les étudiants sont parfaitement blasés, dans l'instant, détachés de toute considération historique. Rien n'est plus faux. Les capsules historiques les intéressent au plus point, comme nous avons pu le vérifier années après années. Nous en avons préservé un certain nombre dans notre site internet.

Par ailleurs, introduire la société dans nos cours intéresse aussi au plus haut point (Wallman, 1993; Rey, 2016): rares sont les semaines où les médias ne rapportent pas des résultats d'enquêtes, de sondages et d'études d'intérêt public.

Nous avons le devoir de nous y intéresser, ne fût-ce que pour motiver les étudiants, mais surtout pour obéir aux injonctions bien sensées de Postman et de Wallman : le monde contemporain a bien besoin des Lumières et d'humanisme. Voici le lien pour notre site internet où on trouvera bon nombre des ces capsules: Wikistat.ca. On pourra consulter en français avec profit les deux livres suivants de Droesbeke & Vermandele (2016, 2018) très riches à cet égard.

Deux références. Comme le soulignent Wasserstein &al. (2019), la tâche vient d'être amorcée. Les habitudes à changer sont très ancrées. La route sera longue et difficile. Déjà le numéro spécial de *The American Statistician* de plus de 400 pages dresse un panorama qui sera long à explorer. La question de la transmission des connaissances, fortement compliquée par la révolution technologique qui s'amorce, est primordiale.

[11] Tous les essais de Postman [1931-2003] sont à lire, l'éducation étant au centre de ses préoccupations. Nous aimerions mentionner aussi « *Technolopoly. The surrender of culture to technology.* » (1993). En français, on trouve la remarquable suite d'essais de Jacques Ellul [1912-1994], hélas trop oublié aujourd'hui (il l'est moins aux USA) : « Le système technicien » (1977) ; « Le bluff technologique » (1988) ; « La technique ou l'enjeu du siècle » (1954, 1990). Peut-être mis à jour, pour tenir compte des très graves problèmes pour notre civilisation dus aux dégradations de l'environnement par « *Homo sapiens technologicus* » (Michel Puech, Le Pommier, 2008), qui cite généreusement les deux auteurs précédents.

On nous permettra de mentionner, en rapport avec la transmission de notre discipline, deux articles qui nous semblent intéressants en tant qu'enseignant : « *Beyond calculations : a course in statistical thinking* », de Steel, Lierman & Guttorp (2019), ainsi que l'expérience menée par Weinberg, Wiesner & Pfaff (2010), « *Using informal inferential reasoning to develop formal concepts : analyzing an activity* ». Ou encore, plus récemment, la recension de Tobias-Lara & Gomez-Blancarte (2019) sur la question du passage en bas âge par l'inférence informelle pour aider à l'apprentissage de l'inférence statistique formelle. Cette dernière référence offre une bibliographie récente sur la question.

Proposons enfin quelques livres de vulgarisation statistique qui pourraient intéresser et enrichir nos cours : Droesbeke & Vermandele, 2016 ; Rey, 2016 ; Salsburg, 2001 & 2017 ; Senn, 2003. Des experts qui savent vulgariser : de ça aussi, on a besoin.

6. Discussion et conclusion

De plus en plus de statistique, de moins en moins de statisticiens, de plus en plus de presse-bouton incompétents : l'ASA s'attaque à un vaste problème. L'occasion lui est donnée par une crise profonde concernant le paradigme fréquentiel de l'inférence statistique, les TSHN dans la variante de Neyman et Pearson. La réplication des résultats, condition *sine qua non* de toute science n'y est pas garantie, ne va pas de soi. Pas de réplication, pas de science ; pas de statistique, pas de réplication.

La faute en revient avant tout à l'institution scientifique elle-même qui tolère une statistique ritualisée, laissant la place belle à toutes les tricheries (Gigerenzer, 1998 &2018). La déclaration de l'ASA (2019) devrait apporter un traitement au moins partiel à cette crise de croissance.

Mais, les statisticiens eux-mêmes ont une grande part de responsabilité dans cette situation. Ils ont négligé les apories des TSHN, négligé leurs devoirs de transmission des éléments de leur science. Nous avons offert ici quelques pistes pour aider à la transmission de notre discipline.

L'ASA a finalement réagi devant la crise; entrepris une vaste analyse de la situation; publié une déclaration solennelle (mise au point ou mise à jour) sur les nouvelles pratiques à mettre en place. La conception des études reste passablement identique à l'ancienne, ce sont les règles de publication qu'il faudra modifier. Surtout éviter les mots *statistiquement significatif* ou *non...* ; il n'y a plus de fil rouge à respecter, en conséquence on ne doit plus jamais parler de α=0,05 dans les publications, sauf éventuellement dans les conceptions des expériences statistiques.

Il faudra encore beaucoup de réflexion sur les conditions à remplir pour la publication des articles. Sans compter que les 'vertus se perdent' comme disaient les Romains de la fin de la République. Comment retrouver les vertus d'honnêteté? Il y a loin de la coupe aux lèvres.

Pour notre part, nous offrons un parcours dans la vaste forêt de la question de l'inférence statistique fréquentielle élémentaire. Comment concilier une pratique ancienne, logiquement correcte, avec les nouvelles normes ? Nous n'avons aucune prétention d'être le moindrement exhaustif, ni même d'avoir tracé un parcours parfaitement rationnel.

En particulier, nous n'avons traité en aucune façon, et c'est une lacune, du grand paradigme concurrent, qui relève d'une tradition plus ancienne que les TSHN : le paradigme bayésien qui vient plutôt compléter que concurrencer le paradigme fréquentiel. Il devient praticable à grande échelle

avec la progression des moyens de calcul. Mais il pose également des problèmes et là non plus la reproductibilité tant désirée n'est pas garantie. Senn (2001) en fait une étude très fine, et arrive à la conclusion qu'il reste encore beaucoup d'espace pour la version TSHN. Nous avons ignoré aussi la théorie statistique de la décision qui est le cadre naturel pour les décisions à prendre en présence d'incertitudes (Manski, 2019).

Quoi qu'il en soit, les lourds nuages qui se sont accumulés sur la statistique classique sont maintenant largement dissipés. Les statisticiens peuvent respirer. Reste la nouvelle pratique, alliée à l'ancienne, à entreprendre…

Références

Nous avons beaucoup écrit sur ces questions. Voir une autre liste de références sous ce lien.

La plupart des références citées plus bas sont en accès libre. Plusieurs ont été hyper-référencées.

Amrhein, V., Greenland, S., McShane, B., & al. (2019). Retire statistical significance. *Nature, 567*, pp. 305-307.

Benjamin, D. J., & 71co-auteurs. (2018). Redefine statistical significance. *Nature: Human Behaviour, 2*, pp. 6-10.

Bourdeau, M. (2015). Fisher a eu 125 ans ce 17 février 2015. *Statistique et enseignement, 6*(1), pp. 65-71. Consulté le 09 17, 2019, sur https://wikistat.mgi.polymtl.ca/tiki-download_file.php?fileId=128

Bourdeau, M. (2015). Tonnerre et tremblements. La crise existentielle de la statistique en cet automne 2015. *Statistique et enseignement, 6*(2), pp. 81-85. Consulté le 09 18, 2019, sur https://wikistat.mgi.polymtl.ca/tiki-download_file.php?fileId=147

Bourdeau, M. (2017). *Un incident au symposium SSI (ASA-Symposium on Statistical Inference).* Consulté le 25 août 2019, sur https://wikistat.mgi.polymtl.ca/tiki-download_file.php?fileId=438

Calin-Jageman, R. J., & Cummings, G. (2019). The new statistics for better science: ask how much, how uncertain, and what else is known. *The American Statistitian, 73sup:1*, pp. 271-280. Récupéré sur https://doi.org/10.1080/0031305.2018.15.18266

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2e éd.). New York NY: Lawrence Erlbaum Associates.

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science, 4*. Récupéré sur http://dx.doi.org/10.1098/rsos.171085

Donoho, D. (2015). *50 years of Data Science.* Presentation at the Tukey Centenial Workshop, Princeton NJ. Récupéré sur course.ccs.neu.edu/.../50YearsOfDataScience.pdf

Droesbeke, J.-J., & Tassi, P. (2015). *Histoire de la statistique* (2e éd. corrigée). Coll. Que sais-je? Paris F: Presses universitaires de France.

Droesbeke, J.-J., & Vermandele, C. (2016). *Les nombres au quotidien. Leur histoire, leurs usages.* Paris F: Éditions Technip.

Droesbeke, J.-J., & Vermandele, C. (2018). *Histoire(s) de(s) données numériques.* Paris : EdpScience.

Ellis, P. D. (2010). *The essential guide to effect sizes.* Cambridge UK: Cambridge University Press.

Fisher, R. (1926). *Statistical methods for research workers.* Edinburgh, Scotland: Oliver and Boyd.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Beharioral Brain Sciences*, pp. 199-200. doi:10.1017S0140525SX98281167

Gigerenzer, G. (2018). Statistical rituals: the replication delusion and how we got there. *Advances in Methods and Pratices in Psychological Science, 1*(2), pp. 198-208.

Grant, R. (2019). *Data visualization: charts, maps, and interactive graphics.* Boca Raton FL: CRC Press.

Hahn, G. J., & Meeker, W. Q. (1993). Assumptions for statistical inference. *The American Statistician, 47*(1), pp. 1-11.

Ioannidis, J. A. (2005). Why most published research findings are false. *PLoS-Medicine (Public Library of Science), 2*(8), pp. 696-601.

Kenett, R. S., & Shmueli, G. (2015, August). Clarifying the terminology that describes scientific reproductibility. *Nature Methods, 12*(8), p. 699.

Kenett, R. S., & Shmueli, G. (2016). *Applications of InfoQ (Information quality) and reproducible research.* New York NY: J.Wiley & Sons.
Kraemer, H. C. (2019, August 7). Is it time to ban the p-value? *JAMA Psychiatry*. Récupéré sur https://www.ncbi.nlm.nih.gov/pubmed/31389991
Lehmann, E. L. (2013). *Fisher, Neymman , and the creation of classical statistics.* New York, NY: Springer.
Madrid Casado, C. M. (2014). *Fisher et l'inférence statistique.* Paris F: RBA sciences.
Manski, C. F. (2019). Treatment choice with trial data: statistical decision theory should supplant hypothesis testing. *The American Statistician, 73:sup 1*, pp. 296-304.
Nosek, B. A. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). doi:10.17605/OSF.IO/EZCUJ
Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018, 13 mars). The preregistration revolution. *PNAS, 115*, pp. 2000-2006.
Postman, N. (1995). *The end of education. Redefining the value of school.* New York NY: Random House Inc/Vintage Books.
Postman, N. (1999). *Building a bridge to the 18th century. How the past can improve our future.* New York NY: Alfred. A. Knopf.
Rey, O. (2016). *Quand le monde s'est fait nombre.* Paris F: Stock.
Salsburg, D. S. (2001). *The lady tasting tea. How statistics revolutinized science in the twentieth century.* New York NY: W.H. Freeman and Co.
Salsburg, D. S. (2017). *Errors, blunders, and lies. How to tell the difference.* Boca Raton FL: CRC Press.
Senn, S. (2001). Two cheers for p-values. *Journal of Epidemiology and Biostatistics, 6*(2), pp. 193-204.
Senn, S. (2003). *Dicing with death. Chance, risk and health.* Cambridge UK: Cambridge University Press.
Steel, E. A., Lierman, M., & Guttorp, P. (2019). Beyond calculations: a course in statistical thinking. *The American Statistician, 73sup 1*, pp. 392-401. doi:10.1080/00305.2018.1505657
Stevens, J. R. (2017, May 26). Replicability and reproducibility in comparative psychology. *Frontiers in Psychology, 8, Article 862*. doi:10.3389/fpsycg.2017.00862
St-Onge, J.-C. (2013). *Tous fous? L'influence de l'industrie pharmaceutique sur la psychiatrie.* Montréal QC: Écosociété.
St-Onge, J.-C. (2019). L'influence des lobbies sur la science. *Présentation au JourStat2019 de la Société statistique de Montréal.* Montréal. Consulté le 25 août 2019, sur https://wikistat.mgi.polymtl.ca/tiki-download_file.php?fileId=511
Tobias-Lara, M. G., & Gomez-Blancarte, A. L. (May 2019). Assesment of formal and informal inferential reasoning: a critical research review. *Statistics Education Research Journal (SERJ),, 18*(1), pp. 8-25.
Wallman, K. K. (1993). Enhancing statistical education: enriching our society (ASA 1992 Presidential Address). *JASA (Journal of the ASA), 88*(421), pp. 1-8.
Wasserstein, R. L., & Lazar, N. P. (2016). The ASA-statement on p-values: context, process and purpose. *The American Statiscian, 70*(2), pp. 129-133.
Wasserstein, R. L., Shirm, A. L., & Lazar, N. L. (2019). Moving to a world beyond "p<0,05". *The American Statistician, 73sup:1*, pp. 1-19.
Weinberg, A., Wiesner, E., & Pfaff, T. J. (2010). Using informal inferential reasoning to develop formal concepts: analyzing an activity. *Journal of Statistics Education, 18*(2).
Xiaofeng, S. L. (2014). *Statistical power analysis for the social and behavioral sciences: basic and advanced techniques.* New York NY: Routledge.