Discussion générale et conclusion

A travers cette thèse, nos objectifs de départ étaient (1) d'identifier des manquements dans les pratiques actuelles des chercheurs, via des analyses d'articles publiés dans des revues de psychologie ; (2) de réaliser des simulations, en vue de montrer l'impact de ces pratiques et (3) de proposer des recommandations pour les améliorer.

Dans un premier temps, nous nous sommes focalisés sur l'usage des tests t de Student et F de Fisher, soit des tests communément utilisés par les chercheurs en psychologie, en vue de comparer les moyennes de deux ou plusieurs groupes de sujets indépendants, et qui reposent sur les conditions que des résidus, indépendants et identiquement distribués, soient extraits d'une distribution normale et que les variances des populations dont sont extraits chaque groupe soient identiques (soit la condition d'homogénéité des variances). Bien que les enjeux des conditions statistiques de ces tests aient déjà été largement explorés par le passé, ils semblaient toujours largement ignorés par de nombreux chercheurs appliqués. Notre principale motivation à aborder cette thématique était dès lors d'ordre pédagogique: il nous semblait nécessaire de combler le fossé entre les méthodologistes et la majorité des chercheurs appliqués. Il nous est d'abord apparu que la litérature manquait d'articles expliquant de manière compréhensible les raisons pour lesquelles les conditions statistiques des tests t de Student et F de Fisher étaient peu réalistes. Nous avons dès lors mis en évidence toute une série d'arguments qui permettent de remettre en cause la crédibilité, dans de nombreux domaines de la psychologie, des conditions statistiques de normalité (comme la présence de sous-populations définies par des facteurs non identifiés dans le design, l'étude de mesures bornées, comme le temps qui ne peut prendre des valeurs négatives, ou encore le fait qu'un traitement est susceptible de modifier la forme des distributions étudiées) et d'homogénéité des variances (comme l'étude de groupes pré-existants à l'expérience, définis par des variables telles que le genre ou l'origine ethnique[1], ou encore le fait qu'un traitement, qu'il soit expérimental ou quasi-expérimental, est susceptible d'agir sur tous les paramètres d'une distribution, incluant sa variance). Ensuite, grâce aux avancées informatiques récentes, nous ont pu étendre les travaux déjà engagés par de nombreux auteurs avant nous [voir par exemple @harwell_summarizing_1992], en vue de montrer les conséquences réelles de la violation des conditions de normalité et d'homogénéité des variances pour respectivement les tests statistiques t de Student (chapitre 2) et F de Fisher (chapitre 3). Nous avons à cette fin réalisé des simulations intensives, avec 1,000,000 d'itérations pour un nombre très vaste de scénarios, variant en fonction d'un ensemble de paramètres connus pour jouer un rôle clé sur les taux d'erreur de type I et II des test t de Student et F de Fisher. Il est ressorti de nos simulations que de manière consistante avec nos attentes théoriques, lorsque les deux échantillons comparés sont de même taille, le test t de Student est robuste aux violations de la condition d'homogénéité des variances. Par contre, il en est différemment lorsque les échantillons sont de tailles différentes: sur le long terme, la probabilité de rejeter l'hypothèse nulle avec ce test est supérieure aux attentes théoriques lorsque le plus petit échantillon est extrait de la population ayant la plus grande variance, et est inférieure aux attentes théoriques lorsque le plus petit échantillon est extrait de la population ayant la plus petite variance. Dans la mesure où l'ANOVA F de Fisher est une généralisation du test t de Student[2], il n'est pas surprenant que nos simulations relatives à l'ANOVA F de Fisher aient amené à des constats semblables à ceux obtenus sur base de nos simulations relativement au test t de Student. En outre, ces simulations nous ont permis de faire deux constats supplémentaires: d'abord, lorsqu'on compare plus de deux groupes, l'ANOVA F de Fisher est affectée par les écarts à la condition d'homogénéité des variances, même lorsque tous les échantillons sont de tailles identiques. Dans ce cas, le test devient plus liberal, ce qui signifie qu'il amène à rejeter l'hypothèse nulle plus souvent qu'attendu théoriquement, sur le long terme. Ensuite, plus le nombre d'échantillons comparés est important, plus le test est affecté par les violations de la condition d'homogénéité des variances. Si la prise de conscience des limites d'une méthode est un premier pas très important, il est tout aussi important de savoir comment pallier ces limites. C'est pour cette raison que nos simulations incluaient également les résultats de tests théoriquement jugés comme constituant de bonnes alternatives, plus robustes en cas de violation de la condition d'homogénéité des variances, à savoir les tests t de Welch, F de Fisher et F^* de Brown-Forsythe. De plus, il est souvent

[1] Dans ce cas, les sujets ne sont pas répartis aléatoirement entre les groupes. Les variances inégales entre les groupes sont dès lors le résultats de la violation de la condition méthodologique d'indépendance des résidus.

[2] L'ANOVA F de Fisher peut être utilisée lorsqu'on compare deux ou plus de deux échantillons indépendants sur base de leur moyenne. Lorsqu'on compare exactement deux groupes, le test t de Student et l'ANOVA F de Fisher sont strictement équivalents. En effet, ils entretiennent la relation mathématique suivante: $F(1, x) = t^2(x)$.

recommandé aux chercheurs de tester préalablement la condition d'homogénéité des variances et ensuite d'utiliser soit le test t de Student (ou F de Fisher) soit une alternative plus robuste aux écarts à la condition d'homogénéité des variances, suivant que cette condition soit ou non respectée. Nous avons dès lors expliqué et illustré une faille importante du test de Levene, le test d'égalité des variances le plus susceptible d'être utilisé par les chercheurs en psychologie, de par son accessibilité dans les logiciels conviviaux tels qu'SPSS et Jamovi : la puissance du test de Levene à détecter les écarts à la condition d'homogénéité des variance est souvent très faible, si bien qu'il conduira le plus souvent à privilégier le test t de Student (ou l'ANOVA F de Fisher) aux alternatives plus robustes. En ce qui concerne la comparaison des taux d'erreur de type I et II des tests t de Student et t de Welch, il est apparu que le test t de Welch est pratiquement aussi puissant que le test t de Student lorsque la condition d'homogénéité des variances est respectée, et contrôle bien mieux les taux d'erreur de type I et II lorsqu'elle ne l'est pas. De même, le test W de Welch est très légèrement inférieur aux tests F^* de Brown-Forsythe et F de Fisher en cas d'homogénéité des variances, tant en termes de contrôle des erreurs de type I et II qu'en termes de consistances entre les puissances théoriques et observées. Par contre, il leur est bien supérieur en cas de violation de la condition d'homogénéité des variances. Après avoir décrit tous ces résultats, il nous semblait indispensable de résumer le message clé de ces deux premiers articles par des recommandations claires et précises. Cela nous a semblé d'autant plus important que bien souvent, les chercheurs appliqués sont noyés sous les articles dans leur domaine d'expertise si bien que cela limite le temps dont ils disposent pour se consacrer aux articles méthodologiques [@mills_quantitative_2010]. La formulation de directives précises nous semblait être un moyen opportun de limiter ce temps. Compte tenu du fait que la condition d'homogénéité des variances est plus souvent l'exception que la norme, qu'il est parfois très difficile (voire impossible) de détecter les écarts à cette condition à travers des tests, et que la très légère perte de puissance des tests t et F de Welch lorsque la condition d'homogénéité des variances est respectée est largement compensée par le gain que constitue leur usage (en termes de contrôle des erreurs de type I et II) lorsque la condition d'homogénéité des variances n'est pas respectée, nous recommandons l'usage de ces tests par défaut. Cette recommandation s'applique au moins au cas où les échantillons sont de taille différente, lorsqu'on ne compare que deux groupes, et s'applique dans tous les cas lorsqu'on compare plus de deux groupes. Les choix de comparer les tests t de Student et F de Fisher respectivement aux tests de Welch et de Brown-Forsythe et finalement de recommander l'usage des tests de Welch par défaut étaient fortement guidés par le désir de proposer des stratégies qui pourraient être facilement comprises et appliquées par la grande majorité des chercheurs. Comme nous l'avions déjà mentionné en introduction, il existe des tests qui sont plus robustes simultanément aux violations des conditions de normalité et d'homogénéité des variances, tels que les tests où l'on compare des moyennes trimmées [@wilcox_results_1994; @wilcox_how_1998] ou encore les tests non paramétriques. Cependant, ces tests étaient à nos yeux moins susceptibles de provoquer l'adhésion de la majorité des chercheurs, pour deux raisons essentielles. Premièrement, ces tests ne reposent pas sur la même hypothèse nulle que les tests t de Student et F de Fisher, puisqu'on n'y compare plus les moyennes de chaque groupe. L'usage des tests de Welch, au contraire, constitue un moyen simple d'améliorer les pratiques sans pour autant obliger à repenser la manière de définir l'hypothèse nulle. Deuxièmement, les tests de Welch sont déjà implémentés dans la plupart des logiciels courants tels qu'SPSS, Jamovi et R. C'est même la stratégie proposée par défaut dans Jamovi et R, ce qui est important compte tenu de la propension des chercheurs à privilégier les méthodes proposées par défaut dans les logiciels [@counsell_reporting_2017]. Nous ne sous-entendons pas que les tests reposant sur les moyennes trimmées ou les tests non paramétriques sont à bannir (dans la section dédiée aux limites de cette thèse, nous parlerons notamment du test de Yuen que nous avons peut-être injustement sous-estimé au sein de l'article du chapitre 2) et encore moins qu'un outil statistique n'est pas digne d'intérêt s'il n'est pas déjà implémenté dans les logiciels courants (comme nous y reviendrons, il est de plus en plus abordable de proposer de nouveaux outils, par exemple via R). Nous pensons simplement qu'il était plus réaliste, dans un premier temps, de s'assurer que les hypothèses généralement définies par les chercheurs soient testées correctement, avant d'amener une réflexion sur la manière dont on peut améliorer leur définition. Finalement, afin d'assurer l'accessibilité de nos travaux et de permettre à chacun de disposer d'un maximum d'éléments pour les critiquer de manière éclairée, nous avons accordé une grande importance au fait de rendre gratuitement disponibles, en ligne, tant nos articles que l'ensemble des outils qui nous ont permis de les écrire. Nous avons effectué plusieurs démarches en ce sens: les articles présentés au sein des chapitres 2 et 3 ont été publiés dans *l'International Review of Social Psychology*, une revue Open Access. De plus, avant qu'ils n'aient été acceptés pour publications, nous avions diffusé des preprints de ces articles sur les réseaux sociaux (Facebook, Twitter...). Enfin, nous avons rendu

disponible en ligne tous les scripts de nos simulations et analyses, en utilisant la plateforme de l'*Open Science Framework*.

Dans un deuxième temps, nous nous sommes intéressés à la significativité pratique des effets étudiés, au delà de leur significativité statistique. Plus spécifiquement, nous nous sommes focalisés sur le contexte de la comparaison de deux moyennes, dans la continuité de l'article présenté au sein du chapitre 2. En entamant ce chapitre, nous avions deux missions principales à l'esprit. Premièrement, nous voulions rappeler aux chercheurs qu'à l'instar des tests t de Student et F de Fisher, la mesure de taille d'effet la plus connue et la plus utilisée en vue de comparer la moyenne de deux groupes, à savoir le d de Cohen, n'est souvent pas appropriée. A travers l'article présenté au sein du chapitre 4, nous avons rappelé deux limites importantes de cette mesure. La première limite est que le d de Cohen est biaisé, même lorsque toutes les conditions dont il dépend sont respectées. Heureusement, il peut être transformé de sorte à annuler son biais lorsque la condition de normalité des résidus est respectée : la mesure transformée se nomme le g de Hedges, en référence à l'auteur ayant proposé cette transformation. La deuxième limite a été mise en évidence grâce à de nouvelles simulations intensives pour un nombre très vaste de scénarios : une violation de la condition d'homogénéité des variances amène à une forte augmentation de la variance des estimateurs d de Cohen et g de Hedges, et ce même lorsque les deux échantillons sont de taille identique. Deuxièmement, nous souhaitions apporter notre contribution scientifique, suite au constat d'un désaccord, de la part des méthodologistes, quant à la mesure de taille d'effet la plus appropriée à utiliser lorsqu'on compare deux groupes sur base de leur moyenne. A cette fin, nous avons inclus dans nos simulations l'étude de différents estimateurs qui ont été proposés dans la littérature en vue de remplacer le traditionnel d de Cohen (et le g de Hedges) en cas de violation de la condition d'homogénéité des variances. Par rapport aux simulations présentées au sein des chapitres 2 et 3, nous avons accordé plus d'importance au réalisme des scénarios envisagés, en nous appuyant sur l'investigation de @cain_univariate_2017, de sorte à définir des déviations de la condition de normalité qui semblent crédibles dans les domaines de la recherche en psychologie. Il en est ressorti quelques constats très intéressants. Parmi les estimateurs fréquemment proposés en vue de remplacer le traditionnel d de Cohen, on retrouve fréquemment le d de Glass. Celui-ci peut être transformé de sorte à obtenir le g de Glass, théoriquement non biaisé lorsque les résidus se distribuent normalement. Nos simulations ont révélé que la variance du g de Glass varie fortement en fonction de paramètres que l'on ne peut contrôler. Il en est de même pour son biais, lorsque les résidus sont extraits de populations qui ne se distribuent pas normalement. Il s'agit là d'un argument fort pour décourager l'usage de cette mesure. Or, ceci nous semble être un important apport théorique de notre article, dans la mesure où à notre connaissance, personne avant nous n'avait révélé aux psychologues les failles du g de Glass de manière aussi détaillée. Dans la litérature, on retrouve également la mesure d de Shieh, qui entretient une relation mathématique directe avec le t de Welch, ainsi que la mesure d^* de Cohen qui, contrairement au d de Cohen classique, implique le calcul de la moyenne *non poolée* des variances de chaque groupe. De même que pour les estimateurs précédemment cités, il est possible de transformer ces mesures en vue de supprimer le biais lorsque la condition de normalité des résidus est respectée. Cela donne respectivement lieu aux mesures g^* de Hedges et g de Shieh. Grâce à nos simulations, nous avons révélé que le g^* de Hedges est supérieur au g de Shieh, non seulement d'un point de vue inférentiel (contrairement au g de Shieh, le g^* de Hedges est consistant, ce qui signifie que sa variance diminue toujours lorsque les tailles d'échantillon augmentent, de même que son biais lorsque les résidus sont extraits d'une population anormale) que d'un point de vue interprétatif (sa valeur est constante, peu importe que les deux échantillons soient de taille identique ou non). Finalement, lorsqu'on compare les mesures g de Hedges et g^* de Hedges, on constate que le g^* de Hedges n'est très légèrement inférieur au g de Hedges, en termes de biais et de variance, que lorsque des échantillons de tailles différentes sont extraits de population aux variances identiques. Il est tout aussi efficace que le g de Hedges lorsque tant les tailles d'échantillons que les variances de population sont identiques. De plus, il reste valide lorsque la condition d'homogénéité des variances n'est pas respectée, contrairement au g de Hedges. Pour des raisons similaires à celles avancées précédemment, il nous semblait indispensable de conclure cet article par des recommandations très pratiques. C'est ce que nous avons fait en recommandant de privilégier le g^* de Hedges par défaut. Dans la mesure où cette solution n'est pas encore proposée dans la plupart des logiciels conviviaux tels que Jamovi et SPSS[3], notre article a été accompagné d'outils gratuits pour aider les chercheurs à mettre cette recommandation en

[3]Dans Jamovi, par exemple, lorsqu'on réalise un test t de Welch tout en demandant une mesure de taille d'effet, c'est la mesure d^* de Cohen sans correction du biais qui est proposée.

oeuvre. Nous avons créé un nouveau package qui se nomme *deffectsize* [4] et pour ceux qui ne sont pas familiers avec R, nous avons également créé une application shiny[5]. De même que pour les articles présentés au sein des chapitres 2 et 3, nous avons voulu assurer l'accessibilité et la transparence de notre recherche, et pour ce faire, nous avons été un cran plus loin que précédemment : alors que les articles antérieurs n'avaient été rendus accessibles que lorsqu'ils avaient atteint une forme relativement aboutie, via les preprints, la découverte de l'outil Github nous a permis de rendre cette recherche disponible dès le début de sa création[6]. Bien que l'article n'ait pas encore été accepté pour publication[7], le fait de le rendre disponible en ligne a permis la génération de diverses ressources théoriques et pratiques, grâce à la contribution d'autres chercheurs. D'un point de vue théorique, d'abord, le preprint a donné lieu à des échanges très enrichissants avec Geoff Cumming (voir Annexe C). Ceux-ci vont très certainement nous permettre d'améliorer la qualité de cet article, notamment grâce à la prise de conscience de certaines limites, dans sa version actuelle, qui seront décrites dans la section prévue à cet effet. D'un point de vue pratique, certains chercheurs se sont inspirés de notre article et des références que nous y citons pour améliorer des outils disponibles dans Jamovi et dans R. C'est le cas d'Aaron Caldwell[8], qui s'est appuyé sur nos travaux sur le g^* de Cohen pour améliorer la fonction du package "TOSTER" (disponible dans Jamovi) qui sert à réaliser un test d'équivalence dans le contexte de la comparaison des moyennes de deux échantillons indépendants.[9] C'est également le cas de Mattan S. Ben-Shachar[10] VOIR AVEC CHRISTOPHE SUIS PAS SURE DE MA TRADUCTION DANS LA NOTE DE BAS DE PAGE qui a pu constater des divergences entre la manière dont les bornes de l'intervalle de confiance autour de certains estimateurs de tailles d'effet étaient calculées dans le package *effectsize* dont il est le créateur (disponible sur le CRAN) et dans notre package *deffectsize*, et qui a pu corriger son package en conséquence.[11]

Dans un troisième temps, nous nous sommes concentrés sur la tendance des chercheurs à définir par défaut, comme hypothèse nulle, une hypothèse d'absence d'effet. Nous avons souligné que cette tendance persiste même lorsque l'objectif est de prouver une absence d'effet: c'est alors sur base d'un non rejet de l'hypothèse nulle que les chercheurs affirment pouvoir valider leur hypothèse. Pourtant, nous avons vu que ce n'est pas une stratégie adéquate puisque non seulement le test utilisé de cette manière présente de faibles propriétés asymptotiques, mais en plus, la probabilité que le test amène à conclure à l'absence d'effet augmente à mesure que l'erreur de mesure augmente. Nous avons également souligné qu'en réalité, il n'existe aucun test d'hypothèses qui permette de démontrer l'absence totale d'effet. Par contre, il est possible de démontrer qu'un effet observé ne s'éloigne pas de l'absence d'effet d'une quantité supérieure à une valeur définie (dit autrement, qu'il est *équivalent*), à condition de comprendre qu'il est théoriquement possible de définir n'importe quelle différence (ou intervalle de différences) entre les groupes comme hypothèse nulle. C'est le principe sur lequel repose le TOST (Two One-Sided Tests), à travers lequel on conclut à l'équivalence à condition que l'intervalle de confiance à $(1-2\alpha)\%$ autour de l'effet étudié soit entièrement inclus à l'intérieur de la zone d'équivalence.

[4] Ce package n'a pas été sousmis sur le CRAN de R. Pour pouvoir l'utiliser via la console R, vous devez appliquer le code suivant :
install.packages("devtools")
library(devtools)
install_github("mdelacre/deffectsize").

[5] L'application est disponible à l'adresse suivante: https://effectsize.shinyapps.io/deffsize/

[6] Le draft de l'article ainsi que l'ensemble des scripts et outputs générés sont disponibles via le lien suivant: https://github.com/mdelacre/Effect-sizes.

[7] Nous avons soumis une première fois l'article au *British Journal of Mathematical and Statistical Psychology* mais celui-ci a été rejeté.

[8] Aaron Caldwell est un chercheur qui a obtenu un doctorat en Sciences de la Santé, du Sport et de l'Exercice à l'université d'Arkansas et qui réalise actuellement un post-doctorat axé sur la performance humaine dans les environnements extrêmes (chaleur, froid et altitude). Voici sa page: https://aaroncaldwell.us/

[9] Le package amélioré n'est pas encore disponible sur le CRAN et dans Jamovi, mais il est possible d'en avoir un aperçu via ce fil d'actualité Twitter: https://twitter.com/ExPhysStudent/status/1400861069048958981.

[10] Mattan S. Ben-Shachar est un chercheur qui réalise actuellement un doctorat, au sein du laboratoire d'ERP neurocognitif développemental, à l'université Ben Gourion du Néguev (Israël). Voici sa page: https://sites.google.com/view/mattansb/.

[11] Les modifications qu'il a apporté de la sorte sont les suivantes:
1) utilisation de la correction gamma exacte, plutôt qu'une approximation, en vue de supprimer le biais de l'estimateur g de Hedges;
2) utilisation de la méthode basée sur les distributions t non centrales pour définir les bornes de l'intervalle de confiance autour du g de Glass, alors que celles-ci étaient précédemment définies via la méthode du bootstrapping;
3) correction d'un bug dans le calcul de l'intervalle de confiance autour des traditionnels d de Cohen et g de Hedges (lié à une erreur dans la définition de la relation mathématique unissant le t de Student et le d de Cohen).

Nous avons conclus cette thèse par une comparaison du TOST et du SGPV (Second Generation P-Value), récemment proposé par @blume_second-generation_2018 et défini par ses auteurs comme un nouvel outil permettant de calculer la proportion des valeurs de l'intervalle de confiance à $(1 - \alpha)\%$ qui sont également compatibles avec l'hypothèse nulle (ou autrement dit, qui se situent à l'intérieur de la zone d'équivalence). Cette comparaison nous semblait pertinente, dans la mesure où les deux stratégies reposent sur un principe similaire, à savoir la comparaison de l'intervalle de confiance de l'effet observé avec la zone d'équivalence. A travers notre investigation, nous avons révélé de nombreuses failles à l'usage du SGPV (par exemple, dans la mise en place d'une correction sous-optimale) et ne sommes pas parvenus à mettre en évidence de réelle plus-value de cet outil, par rapport à l'usage du TOST. Cette investigation démontre bien à quel point il est important, lorsqu'on propose un nouvel outil, de la comparer à des outils déjà existants afin d'en établir les forces et les faiblesses.

Limites

Au sein du chapitre 2, nous avons comparé les tests t de Student, de Welch et de Yuen. Dans notre investigation, nous avons rapidement statué en défaveur du test de Yuen, en affirmant qu'il contrôle moins bien le taux d'erreur de type I que le test t de Welch. Nous l'affirmons notamment à la page 15, lorsque nous écrivons ceci: *"Yuen's t-test is not a good unconditional alternative because we observe an unacceptable departure from the nominal alpha risk of 5 percent for several shapes of distributions [...] particularly when we are studying asymmetric distributions of unequal shapes"*, ou encore à la page 16, lorsque nous écrivons: *"As it is explained in the additional file, Yuen's t-test is not a better test than Welch's t-test, since it often suffers high departure from the alpha risk of 5 percent"*. Ceci n'est pas exact d'un point de vue purement statistique. En utilisant le test de Yuen, on ne compare plus les moyennes de chaque groupe, mais les moyennes *trimmées* (soit les moyennes calculées sur les données après avoir écarté les 20% des scores les plus faibles ainsi que les 20% des scores les plus élevés). Autrement dit, l'hypothèse nulle classiquement définie pour ce test est que le moyennes *trimmées* de chaque groupe sont identiques. Or, les scénarios de nos simulations Monte Carlo créés en vue de tester le taux d'erreur de type I (risque α) des tests étaient systématiquement des scénarios dans lesquels les moyennes de chaque population étaient identiques. Lorsque la distribution des données est parfaitement symétrique au sein de la population, la moyenne et la moyenne trimmée de cette population seront identiques. Au contraire, lorsque la distribution d'une population est asymétrique, la moyenne et la moyenne trimmée diffèreront (la moyenne trimmée sera plus proche du mode de la distribution et donc, représentera mieux cette dernière). Dit autrement, à l'exception des cas où les échantillons étaient extraits de population qui se distribuent symétriquement, nos scénarios n'étaient tout simplement pas aptes à évaluer le contrôle du taux d'erreur de type I du test de Yuen. D'un point de vue méthodologique, nous avons déjà relevé que la plupart du temps, les chercheurs définissent l'absence de différence entre les moyennes comme hypothèse nulle. Or, dans ce contexte, le test de Yuen n'est pas approprié. Plutôt que d'inclure le test de Yuen dans nos simulations et d'affirmer qu'il contrôle moins bien le taux d'erreur de type I que le test de Welch, il aurait été plus judicieux d'aborder l'usage de ce test en affirmant qu'il ne devrait être utilisé que par des chercheurs ayant pleinement conscience du fait qu'il ne teste pas la même chose que les tests t de Student et t de Welch.

Dans ce même chapitre, nous avons relativement peu parlé de la puissance des tests t de Student et t de Welch. Par contre, nous avons abordé la puissance de manière plus détaillée dans le document fourni en supplément qui se nomme "Additionnal file to 'Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test.' " (DOI: https://doi.org/10.5334/irsp.82.s1). Nous pouvons lire ceci dans ce document, à la page 52: *"In general, departure from the normality assumption leads to a loss in power that can be relatively high, for example, with two double exponential distributions (see Tables A5.2 and A6.2), with one normal distribution and one double exponential distribution (See Tables A5.3 and A6.3), or with one uniform and one double exponential distribution (see Table A5.6 and A6.6)."* Ce constat est erroné, car il vient d'une erreur dans le codage de nos scripts. Pour générer des données provenant d'une distribution double exponentielle, nous avons utilisé la fonction *rdoublex* du package "Smoothmest" dans R. Cette fonction implique de spécifier trois arguments: la taille des échantillons à générer (n), un indicateur de tendance centrale (μ) et un indicateur de dispersion (λ). L'indicateur de dispersion λ diffère de l'écart-type

de la population (σ) mais entretien uen relation mathématique directe avec ce dernier. Cette relation se définit comme suit :

$$\lambda = \sqrt{2} \times \sigma \leftrightarrow \sigma = \frac{\lambda}{\sqrt{2}}$$

Par exemple, si nous voulons générer un échantillon extrait d'une double exponentielle ayant un écart-type de 2, il convient d'introduire la valeur "lambda = 2/sqrt(2)" dans notre fonction. Or, nous avons confondu λ et σ dans nos scripts. Dit autrement, nous avons cru à tort que $\lambda = \sigma$. Cela nous a amené à introduire "lambda = 2" au lieu de "lambda = 2/sqrt(2)" dans les simulations correspondant à l'exemple fourni précédemment. Il en résulte que pour toutes les distributions double exponentielles générées, l'écart-type de population était supérieur à celui considéré pour toutes les autres distributions. Il va de soi que tout autre paramètre étant maintenu égal, la puissance d'un test va diminuer lorsque la dispersion des scores augmente. Cette perte de puissance n'est donc pas due au kurtosis, contrairement à ce que l'on croyait. Nous n'avons pas réalisé cette erreur au moment d'analyser nos données, parce que cette perte de puissance correspondait à nos attentes théoriques [voir par exemple @wilcox_how_1998]. Nous avons corrigé cette préconception dans l'article présenté au sein du chapitre 3. On y mentionne qu'on affirme souvent dans la littérature que les distributions ayant une forte densité au niveau des extrêmités sont automatiquement associées à un écart-type plus élevé [@wilcox_introduction_2011], alors qu'en réalité, kurtosis et écart-type sont deux paramètres indépendants [@decarlo_meaning_1997].

Nous n'avons malheureusement pris conscience de cette erreur qu'après publication de l'article. Cela nous a cependant permis de tirer une leçon importante: il est très important, lorsqu'on simule des données, de prendre le temps de vérifier que les paramètres de population considéré correspondent bien à ceux désirés. Cela peut se faire très simplement en simulant un échantillon de très grande taille (par exemple, un échantillon contenant 1,000,000 de données) et en vérifiant que la moyenne de cette échantillon se rapproche bien de la moyenne de la population. Il est également une plus value de systématiquement extraire des paramètres descriptifs des échantillons générés (au minimum, leur moyenne et leur écart-type) et d'étudier la distribution d'échantillonnage des paramètres extraits. Cela permet de prendre conscience d'éventuelles anomalies (parmis d'autres avantages, tels que la possibilité de constituer une information supplémentaire qui aide souvent à comprendre certains résutlats surprenants). Pour tous les articles qui ont suivi, nous avons effectué cette démarche. Nous avons également insisté sur la divergentre entre le kurtosis et les écart-types dans l'article présenté au sein du chaptre 3 "en vue de corriger le tir".

Revoir si je parle du fait que le kurtosis impacte la puissance du test de Welch dans l'article en tant que tel (je le fais en tout cas dans les annexes). Expliquer que j'avais fait une erreur en confondant kurtosis et sd (j'ai cru à tort que la mesure de dispersion de la double expo était le sd alors qu'en fait non). Ca m'a fait prendre conscience qu'il est SUPER important de toujours demander, dans les simulations, le calcul des descriptives, afin de vérifier que tout s'est bien passé (si la variance moyenne n'est pas égale à la variable théorique, par exemple, en tt cas qd la condition de normalité est ok, c'est qu'il y a eu un couac). En plus de permettre un contrôle des erreurs, ça peut être utile comme aide à l'interprétation. A partir de l'article 2, je l'ai systématiquement fait. Et pour me "rattraper", j'ai expliqué en détail la différence entre le kurtosis et le SD dans le 2ème article.

Dans le chapitre 3

Dans le chapitre 4

Dans le chapitre 5

nous avons porté un jugement trop sévère à l'égard du test de Yuen qui repose sur le principe de comparaison des moyennes trimmées. Ce jugement reposait sur une mauvaise compréhension de notre part des objectifs du test.

S'il est certain que les tests de Welch sont bien plus robustes que les tests t de Student et F de Fisher lorsque la condition d'homogénéité des variances n'est pas respectée, il est tout aussi certain qu'ils ne résolvent pas

tous les problèmes. Par exemple, comme le révèlent les figures de l'article présenté au sein du chapitre 3, le test de Welch est sensible à certaines violations de la condition de normalité [12]

L'apport théorique de ces échanges a été très riche, essentiellement parce que nous partions d'avis très divergents. Alors que nous mettions essentiellement l'accent sur les propriétés inférentielles des estimateurs à comparer, Cumming accordait beaucoup plus d'importance à la dimension interprétative. Cela nous a fait prendre conscience des certaines limites de l'article, tel qu'il est écrit actuellement, et également de pistes possibles d'amélioration.

Il est finalement ressorti de ces échanges qu'il n'est pas approprié de donner plus de poids à la dimension inférentielle qu'à la dimension interprétative. Nous semblons pourtant sous-entendre par endroit que de bonnes propriétés inférentielles suffisent à compenser des difficultés en termes d'interprétation (notamment en ce qui concerne la présentation de la mesure d de Shieh, puisque nous écrivons "..."). De tels propos risquent de réduire l'intérêt et la portée pédagogique de notre travail. Un estimateur non biaisé et présentant une faible variance n'a que peu d'intérêt s'il est difficilement interprétable. Cela devient surtout apparent lorsqu'on souhaite définir des hypothèses nulles autres que l'absence de différence, soit des hypothèses prenant en compte la significativité pratique des estimateurs, tel qu'abordé au sein du chapitre 5. Admettre cela ne signifie cependant pas pour autant qu'il ne faille accorder aucune importance aux propriétés inférentielles d'un estimateur: un estimateur ne peut être interprété correctement si sa valeur varie en fonction de facteurs non contrôlables, tels que la vraie corrélation entre la différence de moyenne et le standardiseur, cette dernière survenant lorsque les données sont extraites de populations qui se distribuent asymétriquement. Cela permet d'ouvrir le débats sur la manière de rendre un estimateur interprétable. *parler des auteurs qui proposent des benchmarks plus appropriées que celles de Cohen.*

En recherche, on apprend constamment de nos erreurs. Pour chaque article, j'ai pu identifier des éléments que je ne reproduirais plus à l'identique avec dû recul. Par exemple, dans l'article sur le test de Welch (le premier) j'aurais dû fragmenter. Commencer par faire des simultions avec des distributions identiques dans tous les groupes (car ça permet de moins bien visualiser l'effet de compensation, ex . quand asymétrie positive et négative en mm tps). J'ai également identifié des erreurs dans certains articles

Commentaires divers

Dans les deux articles sur le Welch: (même si les pages relevées concerne le test t, c vrai aussi pour le suivant):
- p.9: nous décrivons 3 arguments en défaveur de l'usage du test de Levene. En troisième argument, nous mentionnons le manque de puissance du test de Levenne. Ceci est rappelé en conclusion de l'article présenté au sein du chapitre 2: *"Because the statistical power for this test is often low, researchers will inappropriately choose Student's t-test instead of more robust alternatives"*. Nous aurions pu ajouter le fait qu'utiliser le test t de Student lorsque le test de Levene est non significatif revient à confondre le non rejet de l'hypothèse d'égalité des variances avec l'acceptation de l'hypothèse d'égalité des variances. Au sein du chapitre 5 sur les tests d'équivalence, il est démontré par simulation que même lorsqu'on s'assure d'avoir une puissance suffisante pour détecter une différence attendue, la stratégie qui consiste à interpréter le non rejet de l'hypothèse nulle comme un soutien en faveur de l'hypothèse nulle n'est pas appropriée.
- p.12: nous mentionnons ceci : *"When both variances and sample sizes are the same in each independent group, the t-values, degrees of freedom, and the p-values in Student's t-test and Welch's t-test are the same (see Table 1).* Avec du recul, cette phrase peut porter à confusion. Par "variances" il faut comprendre "*sample* variances" ou "variances* estimates*". Nous ne sommes donc* pas* en train de dire que les deux statistiques, ainsi que les degrés de liberté et p-valeurs qui leur sont associées seront identiques lorsque la condition d'homogénéité des variances sera respectée au niveau de la population, mais bien lorsque les estimations de chaque variance de population seront identiques.

Limites dans le chapitre 4: nous avons peut-êtr par moment légèrement perdu de vue l'importance de parler le langage des psychologues. *la question de la taille d'effet n'intéresse pas vraiment les statisticiens à la base; @golinski_expanding_2009)]. Et nous avons sans doute mis un peu trop l'accent sur les propriétés*

[12] Entre autre, la puissance du test est altérée (c'est-à-dire qu'elle est non conforme aux attentes théoriques) lorsque les données sont extraites de distributions fortement asymétriques ou qui ont des extrêmités très denses.

inférentielles, comme nous l'a judicieusement fait comprendre Cumming. Limites dans le chapitre 4, on compare essentiellement les estimateurs sur base de leurs propriétés inférentielles. Nous avons tenté de prendre la dimension interprétative en compte, mais c'est parfois très compliqué. Cette dimension est d'ailleurs rarement prises en compte par les chercheurs. On constate que même si les mesures de taille d'effet sont de plus en plus fréquemment reportées, elles ne sont que rarement interprétées et incluses dans les discussions [@funder_evaluating_2019;@thompson_statistical_1997] par les chercheurs. Dans un tel contexte, il est particulièrement important d'ouvrir les débats sur cette question.

Nous avons parfois pu donner l'impression, dans la manière dont nous avons écrit cet article, que de bonnes propriétés asymptotiques suffisaient (qu'un estimateur, même impossible à interpréter, pouvait être utile s'il avait de bonnes propriétés). Pourtant, les deux éléments sont extrêmement importants. Bien sûr, les propriétés inférentielles sont très importantes (il est difficile de concevoir qu'un estimateur puisse fournir une interprétation adéquate s'il est extrêmement biaisé, tel que le Glass, comme on le souligne dans nos échanges avec Cumming). Mais ça ne suffit pas. Il sera nécessaire de reformuler de sorte à mieux faire comprendre cela. Dire pour le Shieh par exemple, qu'on l'a inclu pour montrer que non seulement il est dur à interpréter, mais en plus, ses propriétés inférentielles ne sont pas aussi bonnes qu'on le croit.

Perspectives futures

t-test: "We do not include the bootstrapped t-test because it is known to fail in specific situations, such as when there are unequal sample sizes and standard deviations differ moderately"(p.8; Hayes & Cai, 2007): on s'est contenté de croire l'avis de machin qui dit que ça marche pas bien, mais on pourrait requestioner cela et dans les recherches futures le ré-investiguer la comparaison du test de welch classique avec sa version boostrappé. –> relire son artic pour voir dans quelles conditions ils ont étudié le t boostrappé. Pe pas les mêmes que nous! Dc ça pourrait être utile de refaire la même étude mais en comparant uniquement le t de Welch à sa version boostrappée.

- Je ne travaille pas avec des distributions discrètes.