Chapitre 6 : Discussion générale et conclusion

Objectifs de départ, résumé et apports de la thèse

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 littérature 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 préexistants 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 avons 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), dans des conditions qui se veulent les plus réalistes possibles dans le contexte de la recherche en psychologie (en termes d'hétéroscédasticité et d'écarts à la condition de normalité). 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 pour des échantillons 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 libéral, 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

[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 : $F(1, x) = t^2(x)$.

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 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 consistance entre les puissances théoriques et observées. Par contre, il leur est bien supérieur dans les cas les plus fréquents en psychologie, à savoir les 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 et al.,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 d'utiliser ces tests par défaut. Cette recommandation s'applique au moins au cas où les échantillons sont de tailles différentes, 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, 1994; Wilcox, 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 (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 publication, nous avions diffusé des preprints de ces articles sur les réseaux sociaux (Facebook, Twitter...). Enfin, nous avons rendu disponibles en ligne tous les scripts de nos simulations et analyses, en utilisant la plateforme de l'*Open Science Framework* dans un premier temps, et Github ensuite.

Dans un deuxième temps, nous nous sommes intéressés à la significativité pratique des effets étudiés, au delà de leur significativité statistique. Cela implique d'étudier la taille des effets étudiés, au delà de la p-valeur. 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 illustrée par 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 tailles identiques. 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 et al. (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 littérature, 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 tailles identiques 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 populations 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 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

[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.

[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

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] Le fait de le rendre rapidement 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. Un des points abordés dans cet échange sera décrit dans la section "limites" de cette thèse. 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[7] 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.[8] C'est également le cas de Mattan S. Ben-Shachar[9] 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.[10]

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. Nous avons conclu cette thèse par une comparaison du TOST et du SGPV (Second Generation P-Value), récemment proposé par Blume et al. (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

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] 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/

[8] 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

[9] 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/

[10] 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).

d'une correction sous-optimale) et ne sommes pas parvenus à mettre en évidence une 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 le comparer à des outils déjà existants afin d'en établir les forces et les faiblesses.

Limites

Les simulations que nous avons présentées au sein du chapitre 2 avaient pour ambition de comparer les tests t de Student, de Welch et de Yuen. Cependant, le plan de simulation tel que réalisé n'était pas adéquat pour juger des performances du test de Yuen. A travers ce test, 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 les 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 des tests étaient systématiquement des scénarios pour 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 sont identiques. Au contraire, lorsque la distribution d'une population est asymétrique, la moyenne et la moyenne trimmée diffèrent (la moyenne trimmée est plus proche du mode de la distribution que la moyenne). Dès lors, lorsque les échantillons sont extraits de populations qui suivent une distribution asymétrique, le fait que les moyennes de population soient identiques dans nos simulations ne garantit pas que les moyennes trimmées de populations le soient également (sauf bien sûr si les échantillons sont tous extraits de distributions identiques). In fine, à l'exception des cas où les échantillons étaient extraits de populations qui se distribuent symétriquement ou qui ont une distribution identique, nos scénarios n'étaient tout simplement pas aptes à évaluer le contrôle du taux d'erreur de type I du test de Yuen. Ayant pris conscience de cette erreur, au cours des simulations présentées au sein du chapitre 3, nous n'avons plus considéré de tests comparant les groupes sur base d'autres indicateurs de tendance centrale que la moyenne dans nos simulations et nous avons écrit ceci (p.18) : *"... additional tests exist that allow researchers to compare groups either based on other estimators of central tendency than the mean (see for example Erceg-Hurn & Mirosevich, 2008; Wilcox, 1998)..."*, et ceci dans la conclusion de ce même article (p.27) : *"For more information about robust alternatives that are based on other parameters than the mean, see Erceg-Hurn and Mirosevich (2008)."*

Bien que nous ayons relativement peu parlé de la puissance des tests t de Student et t de Welch dans l'article présenté au sein du chapitre 2, nous avons réalisé des simulations assez détaillées à ce sujet. Nous avons conclu de ces simulations que lorsque les données sont extraites de populations suivant une loi double exponentielle, cela conduit à une importante perte de puissance, tant pour le test t de Welch que pour le test t de Student. Cependant, cette conclusion provenait malheureusement 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. Lorsqu'on utilise cette fonction, on doit préciser la taille des échantillons à générer (n) ainsi que deux paramètres de population : un indicateur de tendance centrale (μ) et un indicateur de dispersion (λ). λ diffère de l'écart-type de la population (σ) mais entretient une relation mathématique directe avec ce dernier :

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

Si nous voulons générer un échantillon extrait d'une double exponentielle ayant un écart-type de 2 via la fonction *smoothmest*, par exemple, nous devons introduire la valeur "lambda = 2/sqrt(2)" dans la fonction. Or, nous avons confondu λ et σ dans nos scripts. Cela nous a amené à introduire "lambda = 2" au lieu de "lambda = 2/sqrt(2)" lorsque nous voulions que l'écart-type de la population soit égal à 2. Il en résulte que pour tous les échantillons générés extraits de populations qui suivent une loi double exponentielle, l'écart-type de population était supérieur à celui considéré pour toutes les autres lois de distribution (il valait $2\sqrt{2}$). Il va de soi que tout autre paramètre étant maintenu égal, la puissance d'un test va diminuer à mesure que l'écart-type de la population augmente. La 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 était conforme aux constats d'autres auteurs avant nous, dont fait partie Wilcox [voir @wilcox_how_1998;@wilcox_introduction_2011]. Nous avons heureusement pu éviter de reproduire la même erreur dans l'article présenté au sein du chapitre 3.

Après avoir soumis un preprint de l'article présenté au sein du chapitre 4, nous avons eu le plaisir de recevoir un feedback très détaillé de Geoff Cumming.[11] Ce feedback nous a fait prendre conscience de certaines limites de l'article. Il nous apparaît que sa principale limite est le fait d'avoir donné plus d'importance aux propriétés inférentielles des estimateurs étudiés qu'à leurs propriétés interprétatives dans le choix des estimateurs comparés via nos simulations. Au delà de la significativité *statistique*, les mesures de taille d'effet donnent une information relative à la significativité *pratique* des effets étudiés. Comme nous le rappelons en introduction du chapitre 4, l'un des objectifs des mesures de taille d'effet est de fournir une information qui aidera le chercheur à statuer sur la *pertinence* d'un effet en situation réelle. Il est important de pouvoir déterminer, dans un contexte donné, à partir de quelle valeur une mesure de taille d'effet présentera un intérêt aux yeux du chercheur (ou autrement dit, d'être capable d'*interpréter* la mesure obtenue). Dans le contexte des analyses de puissance a priori, par exemple, cela permettra de déterminer les tailles des échantillons qui nous assureront une puissance suffisante en vue de détecter des effets jugés dignes d'intérêt. Cela permettra également de définir des hypothèses de test plus informatives que la traditionnelle hypothèse de présence d'un effet non nul, comme nous l'avons abordé au sein du chapitre 5 de cette thèse. Bien sûr, cela ne retire en rien l'importance d'avoir de bonnes propriétés inférentielles. Il est difficile, par exemple, de concevoir qu'un estimateur puisse fournir une interprétation adéquate s'il est extrêmement biaisé et si ses propriétés dépendent de paramètres que l'on ne peut contrôler. C'est d'ailleurs la raison qui nous empêche de partager l'enthousiasme du Dr. Cumming à l'égard du d de Glass. Cependant, si avoir de bonnes propriétés inférentielles est requis pour un estimateur, cela ne suffit pas. Peut-être avons-nous un peu trop mis l'accent sur les propriétés inférentielles afin de contourner une difficulté. La question de l'interprétation n'est certes pas une question simple (elle peut même devenir parfois très compliquée) et ce n'est probablement pas pour rien que les chercheurs reportent fréquemment des mesures sans leur fournir d'interprétation ni les inclure dans les discussions [@funder_evaluating_2019;@thompson_statistical_1997]. Il n'en reste pas moins que ne pas suffisamment considérer la portée interprétative des estimateurs retenus pourrait réduire l'intérêt et la portée pédagogique de notre travail.

Perspectives futures

Lorsque nous avons construit le plan expérimental des simulations présentées dans le quatrième chapitre de cette thèse, nous avons accordé plus d'importance que précédemment au réalisme des distributions générées. Il reste cependant des pistes à explorer pour se rapprocher encore plus de la réalité des données. Par exemple, une voie intéressante serait de créer des variables dépendantes construites au départ d'échelles de Likert lors de la réalisation de simulations Monte Carlo, ces dernières étant très fréquemment utilisées dans la recherche en Psychologie (Croasmun & Ostrom, 2011; Joshi et al., 2015). A titre d'illustration, 9 des 10 articles publiés dans l'*International Journal of Clinical and Health Psychology* (volume 13, numéro 3, Septembre 2013) décrivaient l'usage ou le développement d'échelles de Likert [@hartley_thoughts_2014]. Ces dernières sont constituées d'un ensemble d'affirmations (ou items) liées les unes aux autres. Le principe de base est de demander à un participant de statuer sur son degré d'accord pour chaque affirmation, de lui attribuer un score par affirmation en fonction de la réponse donnée (par exemple, 1 = "tout à fait d'accord", 2 = "d'accord"...), et de combiner ensuite l'ensemble des scores obtenus. Le score composite résultant est une manière de refléter la position du participant par rapport à un construit donné (Joshi et al., 2015).[12] Dans la mesure où rien ne garantit que les différentes modalités d'une échelle de Likert soient équidistantes (par exemple, rien ne garantit que l'écart entre les première et deuxième modalités "pas du tout d'accord" et "pas d'accord" soit le même que l'écart entre les deuxième et troisième modalités "pas d'accord" et "neutre"), plusieurs auteurs s'accordent à penser que les échelles de Likert sont de nature ordinale (voir par exemple H.N. Boone & D.A. Boone, 2012; Jamieson, 2004; Joshi et al., 2015). Cependant, il est également fréquemment

[11] Ce feedback a donné lieu à un échange et ensuite à un blog post, écrit par Geoff Cumming et disponible à l'adresse suivante : https://thenewstatistics.com/itns/2021/06/17/which-standardised-effect-size-measure-is-best- when-variances-are-unequal/.

[12] A l'origine, il n'est donc pas question d'analyser chaque item séparément [@boone_analyzing_2012], bien qu'il arrive que des chercheurs soient amenés à le faire (Boone & Boone, 2012; Joshi et al., 2015; Subedi, 2016).

admis que les scores composites résultant de la combinaison de plusieurs items de type Likert puissent être traités comme étant mesurés sur des échelles d'intervalle. D'aucuns considèrent que ce critère suffit à justifier l'usage d'outils reposant sur le calcul des moyennes et des écart-types (Joshi et al., 2015; Subedi, 2016). Si l'on partage cette conception, il fait tout à fait sens de tenter d'établir la robustesse des tests de Welch et de la mesure g^* de Cohen lorsque l'on simule des échelles de Likert. Soulignons qu'un tel projet serait de grande ampleur et qu'il impliquerait de se positionner par rapport à de nombreux degrés de liberté, tant ce qui concerne la manière de créer les items de type Likert (par exemple, combien de modalités à l'échelle? inclusion ou non d'une position neutre?) qu'en ce qui concerne la manière de les combiner pour former les échelles (par exemple, combien d'items forment l'échelle? Ont-ils tous une distribution de forme homogène ou pas?).

En ce qui concerne le choix des tests comparés à travers nos simulations, nous n'avons pas considéré le bootstrap. Or, un ouvrage récemment écrit par @wilcox_modern_2017 suggère de l'utiliser pour remplacer le test de Welch. Le bootstrap peut être utilisée pour tester l'hypothèse nulle d'égalité des moyennes (Efron et Tibshirani (1993) expliquent dans leur ouvrage comment utiliser la technique du bootstrap à cet escient) tout en présentant l'avantage de ne pas reposer sur l'hypothèse de normalité des résidus, contrairement au test de Welch. Cette stratégie est toutefois connue pour ne pas être appropriée en toute circonstance (elle peut notamment conduire à des résutlats peu fiables lorsqu'on utilise des échantillons de trop petite taille, Efron et Tibshirani, 1993). Reproduire toutes nos simulations en comparant ces deux tests pourrait dès lors constituer un intérêt, en vue de déterminer précisément dans quel cas le bootstrap constitue une réelle plus-value.

Par ailleurs, nous nous sommes exclusivement focalisés sur l'usage de tests lorsqu'on postule que les erreurs sont toutes indépendantes les unes des autres. Autrement dit, nous nous sommes concentrés sur l'étude de designs inter-sujets. Cependant, dans certains domaines de recherche en psychologie tel qu'en psychophysiologie, les ANOVA sont fréquemment utilisées dans le cadre de designs à mesures répétées, c'est-à-dire lorsque les mêmes individus sont exposés à plusieurs conditions expérimentales [@vasey_continuing_1987]. Avec ce type de design, l'existence d'une corrélation entre les mesures répétées implique que l'ANOVA, traditionnellement utilisée, recquiert une condition supplémentaire par rapport aux designs inter-sujets pour être valide : le respect de la condition de sphéricité. Il s'agit d'une condition d'après laquelle les différences de scores entre toutes les combinaisons de conditions possibles doivent être extraites de populations ayant la même variance [@lane_assumption_2016].[13] Une violation de cette condition entraîne une augmentation du taux d'erreur de type I de l'ANOVA [@vasey_continuing_1987] qui sera d'autant plus forte que la violation est importante [@lane_assumption_2016]. De manière peu surprenante, compte tenu de ce que nous avons pu constater pour les designs inter-sujets, de nombreux chercheurs semblent ne pas avoir conscience de ces limites de l'approche classique. A titre d'illustration, Vasey et Thayer (1987) soulignaient que dans les volumes 21 et 22 de la revue *Psychophysiology*, parus en 1984 et 1985, plus de la moitié des études impliquant une ANOVA à mesure répétée ne discutaient pas de la condition de sphéricité [@vasey_continuing_1987].[14] Lorsqu'on se penche sur les solutions proposées par les méthodologistes, celles-ci consistent souvent en des procédures en plusieurs étapes: dans un premier temps, on tente de détecter une violation de la condition de sphéricité et dans un deuxième temps, on prend une décision en conséquence. [15] A l'instar de nos articles

[13]Par exemple, lorsqu'on réalise une ANOVA à mesures répétées avec 3 conditions nommées "A", "B" et "C", la condition de sphéricité sera respectée comme les différences "A-B", "A-C" et "B-C" sont toutes extraites de populations ayant la même variance [@lane_assumption_2016].

[14]Comme le soulignent Vasey et Thayer (1987), ce pourcentage a été calculé en ne prenant en compte que les études pour lesquelles le facteur intra-sujets contenait au moins 3 modalités, puisque le problème de la sphéricité ne se pose pas lorsqu'il n'y a que deux mesures répétées.

[15]Parmi les alternatives proposées, on distingue l'approche multivariée [MANOVA, @vasey_continuing_1987; @lane_assumption_2016] et une correction de l'approche univariée, consistant à corriger les degrés de liberté de la statistique F en les multipliant par une estimation de ϵ, un paramètre qui reflète le degré auquel la sphéricité est violée dans la population. $\epsilon = 1$ lorsque la condition de sphéricité est respectée et $\epsilon < 1$ quand elle est violée. L'argument est que la distribution de la stat univariée F serait mieux approximée par une distribution F avec un nombre réduit de degrés de liberté, d'après @box_theorems_1954. Les estimations les plus connues de ϵ sont $\hat{\epsilon}$ de Greenhouse & Geisser [@greenhouse_methods_1959] et $\tilde{\epsilon}$ de Huynh-Feld [@huynh_estimation_1976]. $\hat{\epsilon}$ est négativement biaisée, ce qui implique que la correction l'utilisant est légèrement conservatrice, et le sera d'autant plus que la violation de la condition de sphéricité est faible : lorsque la condition de sphéricité est respectée, $\hat{\epsilon} < 1$, les degrés de liberté diminuent donc, ce qui a pour conséquence de faire augmenter le seuil critique de la statistique F [@quintana_monte_1994]. A l'inverse, $\tilde{\epsilon}$ tend à être de plus en plus libérale à mesure que la violation de la condition de sphéricité augmente, au point d'observer une inflation du taux d'erreur de type I en cas de violation sévère

présentés au sein des chapitres 3 et 4, un article pédagogique montrant les limites d'une telle approche sur base de simulations Monte Carlo, et proposant des recommandations concrètes et facilement applicables par la majorité des psychologues pourrait s'avérer utile. Cela s'avèrait d'autant plus nécessaire que par le passé, plusieurs références ont déjà souligné non seulement les limites des tests existants pour détecter les écarts à la condition de sphéricité [@vasey_continuing_1987][16], mais également le caractère peu réaliste de la condition de sphéricité (Keselman et al., 2001; Keselman & Rogan, 1980; McCall & Appelbaum, 1973; O'Brien & Kaiser, 1985).

Pour finir, nous nous sommes exclusivement limités à l'étude de designs impliquant que conditions soient délimitées par un et un seul facteur catégoriel. Or, il arrive fréquemment que les chercheurs soient amenés à considérer des designs plus complexes, tels que des designs incluant plusieurs prédicteurs catégoriels, un (ou plusieurs) prédicteur(s) continu(s) ou encore simultanément des prédicteurs catégoriels et continus (cela arrive notamment lorsqu'on décide de prendre une covariable continue en compte dans notre design d'étude; Judd et al., 2018). Afin de faciliter la transition vers ces modèles plus élaborés, il nous semblerait intéressant d'écrire un court article expliquant comment le test de Welch peut être utilisé comme un outil servant à définir un modèle particulier de prédiction, conformément à l'approche proposée par Judd et al. (2010). Il s'agit d'une approche plus intégrée selon laquelle on se demande si un modèle incluant le groupe d'appartenance d'un sujet particulier (appelé le *modèle augmenté*) permet de prédire le score de ce sujet avec moins d'erreur qu'un modèle n'incluant pas cette information (appelé le *modèle compact*), plutôt que de se demander s'il existe des différences significatives entre les moyennes des différents groupes étudiés comme on le fait traditionnellement.[17]

Conclusion générale

A travers cette thèse et des articles qui la composent, mon ambition de départ était d'aider les chercheurs à améliorer leurs pratiques d'analyses de données. Le choix de me concentrer majoritairement sur la question des conditions d'application des tests paramétriques n'est pas venu par hasard. Lorsque j'ai commencé à enseigner les statistiques, à l'ULB puis à l'UCL, j'ai souvent observé que les étudiants appliquaient des stratégies apprises par coeur durant leurs cours. In fine, cela les amenaient à porter trop d'attention à la p-valeur, l'essentiel était que cette dernière soit inférieure au risque alpha. Les statistiques descriptives et l'allure générale des données n'étaient pas toujours sérieusement envisagées avant d'observer cette fameuse p-valeur. Plus tard, j'ai eu l'occasion de travailler avec plusieurs chercheurs dans les domaines des sciences psychologiques et du marketing. Il s'agissait pour la plupart de chercheurs très compétents et sérieux. Pourtant, en ce qui concerne l'analyse de données, j'observais chez eux des comportements assez similaires, même si ces comportements s'appliquaient parfois à des designs beaucoup plus complexes. Certains s'employaient à utiliser des techniques statistiques sophistiquées que je ne maîtrisais pas toujours moi-même, impliquant des équations structurelles, des régressions multiniveaux... Et lorsque je leur demandais un peu plus de détails, notamment sur les conditions d'application de ces tests, la réponse la plus courante était qu'ils ne les vérifiaient pas. Cette impression découlait simplement de mon vécu personnel ou reflétait-elle une réalité à plus large échelle? Si cela s'avérait être une pratique courante, était-elle dû à de la négligence ou à une non-prise de conscience quant à l'importance de cette question? Une revue de la littérature méthodologique m'a alors fait réaliser que non seulement, une multitude de références soulignait la fréquence de cette problématique, mais également que cela était dû à un manque de connaissance liée à l'impact d'une violation de ces conditions sur les tests. Il existe, encore aujourd'hui, un fossé considérable entre le monde des statisticiens, des

[@quintana_monte_1994]. Quelle que soit la correction retenue, dans la mesure où la même correction est apportée sur les degrés de liberté du numérateur et du dénominateur de la statistique F, cela impactera de manière proportionnelle les valeurs représentant la part de variabilité expliquée par le facteur ($CM_{Facteur}$, le numérateur dans le calcul de la F) et la part de variabilité résiduelle ($CM_{Résidus}$, le dénominateur dans le calcul de la F). In fine, la valeur de la statistique F est inchangée : seules les degrés de liberté sont impactés!

[16]Le test de Mauchly, par exemple, est sensible aux violations de la condition de normalité. Entre autres, il manquera de puissance avec des distributions à forte densité au niveau des extrémités ainsi que de manière plus générale, avec des petits échantillons [@obrien_manova_1985]. Le test de Box semble également manquer de puissance [@vasey_continuing_1987].

[17]La question de l'équation de prédiction appropriée à utiliser en cas d'hétérogénéité des variances ainsi que de l'approche utile pour tester la significativité du (ou des) coefficient(s) associé(s) au prédicteur "groupe d'appartenance" est abordée dans le forum de discussion suivant : https://stats.stackexchange.com/questions/142685/equivalent-to-welchs-t-test-in-gls-framework).

méthodologistes et des chercheurs appliqués. A travers cette thèse, mon désir était de faire comprendre aux chercheurs du fait que la question des conditions d'application est bien plus qu'une considération intellectuelle et qu'elle ne concerne pas que les experts en statistiques. La prise en compte des conditions d'application des tests a des implications réelles sur les résultats de tout notre champ de recherche. Au cours des chapitres de cette thèse, nous avons longuement expliqué que l'hétérogénéité des variances, très fréquemment présente dans la recherche en psychologie, pouvait conduire à une perte de puissance ou à l'inverse à des résultats beaucoup trop libéraux lorsqu'on utilise les tests paramétriques traditionnellement d'usage. Concrètement parlant, nous risquons tant de passer à côté de découvertes (en obtenant des résultats non significatifs ou à des résultats qui semblent non pertinent dû à une sous-estimation de la taille d'effet) que de croire à tort avoir mis en évidence des résultats pertinents (en obtenant des résutlats faussement significatifs ou dont l'importance a été amplifiée par une surestimation de la taille d'effet). Dans un cas comme dans l'autre, cela peut s'avérer très problématiques: non seulement nous risquons d'orienter notre recherche vers une mauvaise voie ou encore d'abandonner des thématiques pourtant prometteuses mais en outre, cela pourrait conduire à des décisions cliniques, gouvernementales ou politiques infondées... Pour paraphraser Wilcox (1998), il est temps de cesser de laisser des années de traditions nous pousser à utiliser des tests qui ont pour effet de remplir la littérature scientifique de résultats peu fiables et de nous pousser à prendre de mauvaises décisions. Ceci explique pourquoi, tout au long du processus, nous avons accordé beaucoup d'importance au fait de fournir des recommandations très précises afin d'aider les chercheurs en psychologie à améliorer la fiabilité des résultats obtenus. Ce faisant, nous nous sommes efforcés de fournir des recommandations peu coûteuses, tant en termes de temps qu'en termes de flexibilité demandée. Les solutions proposées n'impliquent en effet pas de se former à de nouvelles techniques statistiques ni à l'usage de logiciels de traitement de données complexes.

Au delà d'un simple changement d'habitude, j'espère être parvenue à faire prendre conscience aux chercheurs de l'importance de bien considérer les enjeux des hypothèses de recherche définies et des tests utilisés pour tester ces dernières, quelle que soit le design étudié. En ce sens, je perçois notre travail sur les tests t de Student et F de Fisher comme une illustration d'une problématique plus large plutôt que comme une fin en soi. Une dérive possible de ces illustrations est d'amener certains chercheurs à remplacer une stratégie, anciennement utilisée par défaut, par une nouvelle stratégie qui deviendrait le défaut à son tour, sans critiquer la pertinence de ce nouveau choix. Ce serait dommage, car cela ferait passer les chercheurs à côté de mon objectif. Mon but n'était pas simplement d'induire ce changement de pratique, dans la contexte étroit que j'ai étudié, mais d'aider le plus de chercheurs possible à réaliser que la question des conditions d'application reste primordiale quelle que soit le design étudié : si deux conditions experimentales suffisent à introduire de l'hétérogénéité des variances, comment concevoir qu'il en soit autrement lorsqu'on augmente le nombre de variables considérées, et dès lors, potentiellement, le nombre de sources d'hétérogénéité des variances possibles? C'est ce désir qui m'amène à considérer l'importance de repenser le test de Welch selon l'approche par comparaison de modèles. Cela devrait permettre aux chercheurs de "concocter leurs propres recettes au fil de leurs analyses des données". Il est important de toujours faire preuve d'esprit critique et de garder à l'esprit qu'un test statistique n'est pas une fin en soi. Il ne s'agit que d'un outil servant à tenter de mettre en évidence un résultat qui nous intéresse, et plus l'outil est de bonne qualité, meilleurs sont nos chances de parvenir à nos fins.

d'un point de vue personnel, l'écriture de cette thèse m'a fait réaliser à quel point les simulations, et plus précisémen l'étude des distributions d'échantillonnages, m'ont aidé à mieux comprendre les statistiques et à réfléchir au sens de ce que je faisais. Cel'a ma convaincu que l'intérêt d'inclure ce genre de choses même dans l'enseigenemnt. Cela permet de mieux se rendre compte de ce qu'il y a en jeu, et de développer l'esprit critique, plutôt que d'appliquer en lciquant sur un bouton.