Discussion générale et conclusion

Résumé 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 statistiques t de Student, F de Fisher et d de Cohen, soit des mesures 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). Comme nous l'avons théoriquement décrit, il existe toute une série d'arguments qui permettent de remettre en cause la crédibilité 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, telle(s?) que 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), dans de nombreux domaines de la psychologie. Conformément à nos objectifs de départ, nous avons réalisé des simulations Monte Carlo en vue de montrer les conséquences réelles de la violation des conditions de normalité et d'homogénéité des variances, et de comparer respectivement les statistiques t de Student (chapitre 2), F de Fisher (chapitre 3) et d de Cohen (chapitre 4) à des alternatives plus robustes en cas de violation de la condition d'homogénéité des variances.[2] Les avancées récentes en termes d'informatique nous ont permis d'étendre les travaux déjà engagés en ce sens par de nombreux auteurs, grâce à 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 du test t de Student. Il en est principalement ressorti 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. Au contraire, le test t de Welch ne dépend pas de la condition d'homogénéité des variances. Il est souvent recommandé aux chercheurs de tester préalablement la condition d'homogénéité des variances, via un test de Levene par exemple, et ensuite d'utiliser soit le test t de Student soit le test t de Welch, suivant que cette condition soit ou non respectée. Cependant, dans la mesure où la condition d'homogénéité des variances est plus souvent l'exception que la norme et qu'il est parfois très difficile (voire impossible) de détecter les écarts à cette condition à travers des tests, nous recommandons l'usage du test t de Welch par défaut, au moins lorsque les échantillons sont de taille différente. En effet, ce test 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. Par ailleurs, il est disponible dans presque tous les logiciels statistiques courants (R, Minitab, Jamovi, SPSS, etc.). Dans la mesure où l'ANOVA F de Fisher est une généralisation du test t de Student[3], il n'est pas surprenant que

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

[2] Comme expliqué en introduction, nous nous sommes principalement focalisés sur la condition d'homogénéité des variances compte tenu, d'abord, de la forte résistance de la part des chercheurs à l'égard des tests comparant d'autres indicateurs de tendance centrale que la moyenne et ensuite, du fait qu'un écart à la condition d'homogénéité des variances affectera bien plus les taux d'erreur de type I et II des statistiques t de Student, F de Fisher et d de Cohen qu'un écart à la condition de normalité.

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

nos simulations relatives à l'ANOVA F de Fisher amènent 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 la présence d'hétérogénéité 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. En cas d'homogénéité des variances, le test W de Welch est très légèrement inférieur aux tests F^* de Brown-Forsythe que F de Fisher, 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 d'hétérogénéité des variances. Pour les mêmes raisons que celles qui nous amènent à privilégier le test t de Welch par défaut, nous recommandons de privilégier systématiquement le test W de Welch lorsqu'on compare plus de deux groupes de sujets indépendants sur base de leur moyenne. Tout comme le test t de Welch, le test W de Welch est disponible dans la plupart des logiciels statistiques fréquemment utilisés par les chercheurs en psychologie (R, Minitab, Jamovi, SPSS, etc.). En ce qui concerne la mesure d de Cohen, nous avons mis deux éléments principaux en évidence. D'abord, il s'agit d'une mesure toujours biaisée, mêmelorsque toutes les conditions dont elle dépend sont respectées. Heureusement, une transformation de cette mesure existe telle que le biais devient nul lorsque la condition de normalité des résidus est respectée. Cette transformation a été proposée par Hedges et porte dès lors son nom: la mesure g de Hedges. Ensuite, une violation de la condition d'homogénéité des variances amènera à une forte augmentation de la variance des estimateurs d de Cohen et g de Hedges, même lorsque les deux échantillons sont de taille identique. Différents estimateurs ont été proposés dans la littérature en vue de remplacer le d de Cohen (et le g de Hedges) en cas de violation de la condition d'homogénéité des variances. Parmis ceux-ci, on retrouve fréquemment le d de Glass, qui 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 de même que son biais (lorsque les résidus sont extraits de populations qui ne se distribuent pas normalement) dépendent fortement de paramètres que l'on ne peut contrôler, ce qui nous amène à décourager l'usage de cette mesure. 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 les mêmes raisons que celles qui nous amènent à privilégier les tests t de Welch et F de Welch par défaut, nous recommandons de privilégier systématiquement le g^* de Hedges.

Dans un deuxième 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) de moyennes 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. Récemment, @blume_second-generation_2018 ont proposé un nouvel outil qui se nomme le SGPV (Second Generation P-Value) qu'ils définissent comme 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). Il nous a semblé pertinent de comparer le SPGV au TOST, 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. Cependant, notre comparaison n'a pas permis de mettre en évidence de réelle plus-value du SGPV par rapport au TOST. Bien que @blume_second-generation_2018 présentent le SGPV comme un outil permettant de déterminer à quel degré les données sont compatibles avec l'hypothèse d'équivalence, nous avons révélé au moins deux situations pour lesquelles cette définition ne tient pas: lorsque l'intervalle de confiance autour de l'effet observé recouvre les deux bornes de la zone d'équivalence tout en ayant une largeur moins de deux fois supérieure à celle de la zone d'équivalence, et lorsque les intervalles de confiance sont asymétriques, ce qui est le cas, par exemple, lorsqu'on étudie une corrélation r de Pearson (tel que décrit dans l'article du chapitre 5) ou encore lorsqu'on étudie des mesures de taille d'effet standardisées de la famille d (ces dernières ayant fait l'objet du chapitre 4). In fine, les seules situations pour lesquelles le SGPV permet de tirer une conclusion claire sont celles où sa valeur vaut exactement 0 ou 1. Or, les conclusions tirées dans ce cas sont similaires, mais moins précises, à celles que permettent de tirer le TOST. *Blume et ses collaborateurs ont introduit un outil qui n'apporte pas grand chose*: *1) les seules valeurs que l'on peut facilement interpréter, ce sont 0 et 1 (mais ces valeurs correspondent à une p-valeur de TOST respectivement $> .975$ ou $< .025$ dc ça ne fait rien de plus que le TOST lorsqu'on l'utilise d'un point de vue Neyman-Pearson, où l'on compare la p-valeur au risque alpha, et fait moins que le TOST lorsqu'on l'utilise du point de vue de Fisher car la p-valeur différencie là où le TOST vaut tj 0 ou 1). 2) Il faut apporter une correction pour éviter une mauvaise interprétation quand l'IC est trop large (alors que pas besoin de correction avec le TOST). De plus, la correction exclut toute une série de situations (où l'IC chevauche les deux bornes de l'IC mais en étant moins que 2 fois plus grand que la zone d'équivalence). Et parfois, la correction apparaît quand ce n'est pas nécessaire, comme on le voit à travers la figure 13 de l'article du chapitre 5. 3) Blume et al. sous-entendent que le SGPV permet d'éviter les correction spour comparaison multiples. Mais c'est faux, vu la correspondance parfaite entre TOST et SGPV quand il s'agit de décider si on a un soutien en faveur de l'équivalence ou pas, ça démontre bien que dans les 2 cas, on peut avoir une déformation des taux d'erreur de type I et II.*

Apports (à faire article par article pour mettre mes idées en ordre)

Test t de student et ANOVA de Fisher

Le principal apport des articles présentés au sein des chapitres 2 et 3 était d'ordre pédagogique. D'un point de vue théorique, bien que la question de l'impact des violations des conditions d'application du test t de Student ait déjà été largement explorée par le passé [voir par exemple @harwell_summarizing_1992], les débats autours de cette question semblaient toujours ignorés par de nombreux chercheurs appliqués. Il nous est apparu que la litérature manquait d'articles expliquant de manière compréhensible les conséquences réelles des violations des conditions d'application des tests, en s'appuyant sur des exemples concrets et issus de la psychologie. C'est donc pour combler le fossé entre les méthodologistes et la majorité des chercheurs appliqués que nous avons écrit ce premier article. Cependant, si prendre conscience des limites d'une méthode est un premier pas très important, il est également important d'expliquer aux chercheurs comment réagir en conséquences.

Puisque nous souhaitions nous adresser à tous les chercheurs en psychologie, nous avons mis des démarches en oeuvre en vue d'assurer une grande diffusion et visibilité de cet article: 1) Avant que l'article ne soit accepté pour publication, nous avons diffusé un preprint sur les réseaux sociaux (Facebook, Twitter...) 2) Nous avons soumis l'article dans une revue Open Access (*l'International Review of Social Psychology*).
3) Nous avons rendu disponible en ligne gratuitement et en open access tous les scripts de nos simulations et

analyses. Nous l'avons fait dans un premier temps en utilisant la plateforme de l'OSF (Open Science Framework) et ensuite via Github (à définir?) Ces démarches semblent avoir porté leur fruit, compte tenu du taux de citation de l'article (proche de 400 citations). **D'un point de vue appliqué**: *proposer des recommandations dont nous avons de bonnes raisons qu'elles seront appliquées par les chercheurs (cf.: pas introduire des nouveaux tests ultra complexe et révolutionner le truc, mais plutôt proposer des recommandations faciles à mettre en oeuvre). Là parler du Welch et des logiciels.*

Au delà de l'apport théorique, nous avons proposé des recommandations pratiques, pour aider les chercheurs à savoir comment réagir, afin de contrer les csq négatives des violations des conditions d'application. Pour assurer l'impact d'un article pédagogique, il est important de pouvoir comprendre les procédures et méthodes requises par les psychologues afin de s'adapter aux mieux à leurs besoins et attentes, et c'est pourquoi, à l'instar de @golinski_expanding_2009, il nous semblait important que ce travail soit réalisé par une équipe incluant des psychologues ayant une expérience dans la recherche appliquée, ce qui était le cas de deux des co-auteurs (Daniël Lakens et Christophe Leys). Cette conscience de la réalité des psychologues est vraisemblablemnet ce qui nous a amené à recommander l'usage du test de Welch plutôt que l'usage de tests bien plus complexes qui avaient peu de chances de susciter l'adhésion des chercheurs. Comme souligné en introduction, on aurait pu proposer des stratégies complexes tout à fait immunes aux conditions de normalité et d'homogénéité des variances. Wilcox, par exemple, a tenté de le faire pendant longtemps. Mais une prise de conscience de la forte résistance de la part des chercheurs à l'égard des tests comparant d'autres indicateurs de tendance centrale que la moyenne (doublée de la connaissance qu'un écart à la condition d'homogénéité des variances affectera bien plus les taux d'erreur de type I et II des statistiques t de Student, F de Fisher et d de Cohen qu'un écart à la condition de normalité) nous a amené à formulé des objectifs qui nous semblaient plus réalistes.

d de Cohen

Apport théorique: m'inspirer de tous les échanges avec Cumming. J'ai en gros bien montré les inconvénients du d de Glass. Et le faire en toute transparencea permis des échanges ultra utiles avec Cumming (en termes de réflexion liée à l'interprétabilité, etc.)

Apport pratique: Mettre toute la partie où je signale qu'un article méthodo suffit raremetn à lui seul. Dans les deux articles d'avant, il n'étai tpas nécessaire de créer des outils dans la mesure où le Welch est déjà dans pleins de logiciels très utilisés. Mais ce n'est pas le cas pour le g* de Cohen. D'où les packages et Shiny App.

TOST vs. SGPV

Apport essentiellement théorique puisqu'on critique le SGPV. Il n'y a pas eu vraiment d'apport pratique puisque le TOST a déjà fait l'objet d'outils concrets par le premier auteur de l'article (Lakens).