

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 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 pré-existants à l'expérience, définis par des variables telles que le genre ou l'origine ethnique¹, 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), 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 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 Student², 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 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

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

²L'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 : $F(1, x) = t^2(x)$.

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 consistances 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_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 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_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 (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 avons 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 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 avons 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 souhaitons 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 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 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 SPSS³, 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 application shiny⁵. 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 création⁶. 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 (voir Annexe C). Un des points abordés dans cet échange sera décrit dans la section “limites et perspectives” 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 Caldwell⁷, 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-Shachar⁹ 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),

³Dans 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.

⁴Ce package n’a pas été soumis 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").
```

⁵L’application est disponible à l’adresse suivante : <https://effectsize.shinyapps.io/deffsize/>

⁶Le 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>.

⁷Aaron 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/>

⁸Le 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>.

⁹Mattan 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/>.

¹⁰Les 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).

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. À 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 et perspectives

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 déclarant qu'il contrôle moins bien le taux d'erreur de type I que le test t de Welch. Nous l'affirmons notamment lorsque nous écrivons ceci (p.15) : *“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 lorsque nous écrivons ceci (p.16) : *“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”*. Une telle conclusion ne pouvait découler de nos simulations : 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 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 (risque α) 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 populations 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. 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 rappelant 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. La prise de conscience de cette erreur nous a amené à écrire ceci dans l'article présenté au sein du chapitre 3 (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 encore 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)”*.

Nous avons relativement peu parlé de la puissance des tests t de Student et t de Welch dans l'article présenté au sein du chapitre 2. Par contre, nous avons abordé la puissance de manière plus détaillée dans le document fourni en supplément de cet article, qui se nomme *“Additional 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 y lire ceci (p.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 provient 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). L’indicateur de dispersion utilisé dans cette fonction, λ , est un indicateur qui diffère de l’écart-type de la population (σ) mais qui 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.

Il nous est également arrivé d’écrire certaines phrases dans nos articles qui, avec du recul, nous semblent pouvoir porter à confusion. D’abord, dans l’article présenté au sein du deuxième chapitre, nous avons écrit ceci (p.12) : *“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).* Cette phrase peut donner l’impression que les deux statistiques mentionnées, ainsi que les degrés de liberté et p -valeurs qui leur sont associées sont identiques lorsqu’on travaille avec des échantillons de tailles identiques et que la condition d’homogénéité des variances est respectée, ou autrement dit, lorsque les variances de population ainsi que les tailles d’échantillon sont identiques, or, ce n’est nullement vrai. Pour calculer les statistiques t de Student et t de Welch ainsi que leurs degrés de liberté, on utilise les *estimations* des variances de chaque groupe, et non les variances de population. C’est donc chaque fois que l’on obtiendra des *estimations* identiques pour les variances de chaque groupe, sur base d’échantillons de taille égale, que les statistiques, leur degré de liberté et leur p -valeur seront identiques. Or, ceci n’est pas une information très pertinente en soi, puisqu’il arrive très fréquemment d’obtenir des estimations de variance différentes pour chaque groupe lorsque la condition d’homogénéité des variances est respectée, et qu’à l’inverse, il est possible (bien que peu probable) d’obtenir des estimations de variance identiques pour chaque groupe lorsque la condition d’homogénéité des variances n’est pas respectée. Ensuite, dans les articles présentés au sein des chapitres 2 et 3, nous avons décrit trois arguments en défaveur de l’usage du test de Levene. Le troisième argument est le manque de puissance de ce test. Par exemple, nous pouvons lire ceci dans la conclusion de l’article sur le test t de Welch (p.15) : *“Because the statistical power for this test is often low, researchers will inappropriately choose Student’s t-test instead of more robust alternatives”.* Cette phrase peut amener à comprendre que si le test de Levene était toujours très puissant, il serait approprié de l’utiliser en vue de choisir entre les tests t de Student et t de Welch. Pourtant, privilégier le test t de Student lorsque l’on ne peut rejeter l’hypothèse d’égalité des variances (autrement dit, lorsque les résultats du test de Levene sont non significatifs) reviendrait 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 dédié aux tests d’équivalence, nous avons montré que même lorsqu’on s’assure d’avoir une puissance suffisante pour détecter un effet de taille donnée, 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.

Après avoir soumis un preprint de l’article présenté au sein du chapitre 4, nous avons eu l’immense plaisir de recevoir un feedback très détaillé de Geoff Cumming (l’essentiel des échanges que nous avons eu avec cet au-

teur est reporté dans l'annexe C). 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. 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 amener à tirer 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. Pourtant, par moments, nous avons pu donner l'impression qu'un estimateur ayant de bonnes propriétés inférentielles pouvait être utile, même s'il pose de vrais soucis en terme d'interprétation. Par exemple, nous avons inclus le g de Shieh dans nos simulations, sans fournir d'explication réelle relative à ce choix, alors que nous avons très clairement mentionné à quel point cette mesure était difficile à interpréter (p.45) : *"The lack of generality caused by taking this specificity of the design into account has led Cumming (2013) to question its usefulness in terms of interpretability : when the mean difference $\bar{X}_1 - \bar{X}_2$, S_1 et S_2 remain constant, Shieh's d will vary as a function of the sample sizes allocation ratio ..."*. Il en est de même en ce qui concerne le g^* de Hedges. Pour ce dernier estimateur, nous avons même presque explicitement écrit que les propriétés inférentielles suffiraient à compenser les difficultés en termes d'interprétation (p.46) : *"This estimator has been widely criticized, because it results in a variance term of an artificial population [...] and is therefore very difficult to interpret (Grissom & Kim, 2001) [...]. However, we will show throughout the simulation section that this estimator exhibits very good inferential properties..."*. 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 très compliquée parfois) 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 de telles formulations sont très maladroites et pourraient réduire l'intérêt et la portée pédagogique de notre travail.

Perspectives futures

Perspectives méthodologiques

Etude des échelles de Likert Au cours de nos dernières simulations présentées dans l'article du quatrième chapitre de cette thèse, nous avons tenté d'augmenter le réalisme des distributions générées. Pour ce faire, nous nous sommes appuyés sur l'investigation de @cain_univariate_2017, et avons simulé des distributions continues dont les valeurs du kurtosis (G_2) et du coefficient d'asymétrie (G_1) étaient inspirées de cette investigation. Une piste que nous n'avons pas explorée, par contre, est celle des échelles de Likert. Or, cette piste pourrait s'avérer très utile compte tenu de la haute fréquence de leur usage dans la recherche en Psychologie [joshi_likert_2015;croasmun_using_2011]. @hartley_thoughts_2014, par exemple, soulignait que 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. Ces échelles sont constituées d'un ensemble d'affirmations (ou items) liées les unes aux autres. Le principe 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", etc.), et de combiner ensuite l'ensemble des scores obtenus. Le score composite résultant est censé refléter la position

du participant par rapport à un construit donné [joshi_likert_2015].¹¹ Bien qu'il soit généralement admis que les items type-Likert sont mesurés sur des échelles ordinales [rien ne garantit en effet que les différentes modalités de l'échelle soient équidistantes, jamieson_likert_2004], de nombreux auteurs considèrent que les scores composites résultant de la combinaison de plusieurs items de type Likert puisse être traités comme des échelles d'intervalle et dès lors, que l'usage d'outils reposant sur la moyenne et l'écart-type (tels que les tests de Welch ou la mesure d de Cohen) se justifie pleinement [subedi_using_2016; joshi_likert_2015]. (*Lire Carifio & Perla, 2007; @).

Il faut souligner que réfléchir à la manière de simuler des échelles de Likert pourrait constituer une thèse en soi, tant le nombre de degrés de liberté est grand. (1) D'abord, les items type likert sont très variables. A l'origine, les échelles de likert se composaient de 5 modalités sémantiques: "Tout à fait d'accord", "D'accord", "Sans opinion", "Pas d'accord", "Pas du tout d'accord" [boone_analyzing_2012]. Par la suite, de nombreuses variations des échelles de Likert ont été proposées: on peut à présent envisager de faire varier le nombre de modalités de réponse, d'inclure ou non la possibilité d'opter pour une position neutre ou une absence de position ou encore de proposer un nombre identique de modalités de réponse de part et d'autre de la position neutre (échelles symétriques) ou pas (échelles asymétriques, joshi_likert_2015). (2) Ensuite, il existe plusieurs manières de créer des variables discrétisées. voir article de Wu & Leung. (3) Enfin, il est possible de faire varier le nombre d'items constituant l'échelle de Likert (subedi_using_2016 recommande d'en combiner au moins 4 pour utiliser du paramétrique). Et bien sûr, tout cela doit se faire en garantissant le contrôle des paramètres de population de l'échelle finale (forme, indicateurs de tendance centrale et de dispersion). Il faut aussi réfléchir au degré d'homogénéité entre les items. C'est donc extrêmement complexe.

Lorsqu'on calcule un score composite, d'aucun admettent qu'on peut considérer le score résultant comme un score continu (voir Wu & Leung ou encore Joshi et al.). Il arrive cependant que les participants souhaitent analyser chaque item séparément (on ne parle alors plus d'échelle de Likert, mais d'items de type Likert, boone_analyzing_2012). Donner des exemples pour dire que c'est utilisé par les psy (voir ref Hartley par exemple). Pris séparément, les items correspondent à des mesures ordinales (). En effet, on ne peut affirmer que l'écart entre tous les points est régulier. Les outils tels que la moyenne et l'écart-type ne sont dès lors pas appropriés [boone_analyzing_2012, joshi_likert_2015]. Pourtant, il arrive fréquemment que les chercheurs traitent de telles échelles comme des échelles d'intervalles (Blaikie, 2003, cité par Jamieson, 2004), et cela fait controverse. Note: pour certains statisticiens, il n'y a aucun soucis à utiliser un test t pour des items type-likert (voir Sisson & Stocker, 1989, cités par Clason & Dormody, 1994) -> mais Clason & Dormody sont pas d'accord (et a priori moi non plus, mais aller lire l'article quand même).

- aborder les échelles de likert ça pourrait être intéressant mais c'est une thèse en soi parce que c'est ultra complexe. Et plus c pas si évident que ça soit très utile en psycho parce qu'en général on fait pas des single items. Voir si ça sert: cela implique une grande réflexion quant à la manière de construire les scores simulés.

Etude de tests pour échantillons intra

Déterminer l'équivalent "Régression" des tests de Welch

- trouver l'équivalent "régression" du test de Welch, pour pouvoir généraliser à des concepts plus complexes.

Perspectives pédagogiques

- En faire du matériel pédagogique sur base de la thèse.

¹¹A l'origine, il n'est donc pas question d'analyser chaque item séparément [boone_analyzing_2012], même s'il arrive que des chercheurs soient amenés à le faire [subedi_using_2016; joshi_likert_2015; boone_analyzing_2012].

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 requestionner cela et dans les recherches futures le ré-investiguer la comparaison du test de welch classique avec sa version boosttrappé. -> relire son artic pour voir dans quelles conditions ils ont étudié le *t* boosttrappé. 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 boosttrappée.

- Je ne travaille pas avec des distributions discrètes.
- parler du fait que l’usage des simulations devrait être plus exploité dans les cours? C’est déjà le cas dans certaines Fac (cf. Lisa Debruyne). Un logiciel ne fait pas tout et après avoir utilisé le test adéquat, il est important d’être capable de l’interpréter correctement. Les tests font appel à des notions faussement simples telles que les *p*-valeurs et les distributions d’échantillonnage. A mon sens, le seul moyen d’enseigner correctement ces notions, c’est à travers des simulations.