Chapitre 1: Introduction

On attend des chercheurs en psychologie, et des psychologues en général, qu'ils soient capables de produire des connaissances fondées sur des preuves scientifiques (et non sur des croyances et opinions), et également de comprendre et évaluer les recherches menées par d'autres [@haslam_research_2014]. Or, dans un domaine dominé par les analyses quantitatives[1] [@counsell_reporting_2017], les connaissances statistiques s'avèrent fondamentales pour comprendre, planifier et analyser une recherche [@howitt_understanding_2017; @everitt_statistics_2001]. Les statistiques font dès lors partie intégrante du cursus de formation des psychologues et jouent un rôle très important dans leur parcours [@hoekstra_are_2012].

Traditionnellement, depuis plus de 50 ans, les tests-t et les ANOVA se trouvent au choeur de la grande majorité des programmes dans les domaines des Sciences Psychologiques et de l'Education [@aiken_doctoral_2008; @golinski_expanding_2009;@curtis_training_1998] et des livres d'introduction aux statistiques pour psychologues [@field_discovering_2013; @judd_data_2011]. Cela pourrait vraisemblablement expliquer pourquoi ils sont si persistants dans la recherche en psychologie [@counsell_reporting_2017]. Ces tests sont les plus fréquemment cités dans la littérature scientifique depuis plus de 60 ans [@golinski_expanding_2009;@nunnally_place_1960; @byrne_status_1996]. Dans une revue de 486 articles publiés en 2000 dans des journaux populaires en psychologie [2], @golinski_expanding_2009 avaient relevé 140 articles ($\approx$ 29%) au sein desquels les auteurs avaient mené au moins une ANOVA à un ou plusieurs facteurs. Plus récemment, @counsell_reporting_2017 mentionnaient que parmis un ensemble de 151 études soumises dans 4 revues canadiennes en 2013, environ 40% incluaient une comparaison de moyennes. Peut-être est-ce en raison de leur grande fréquence d'usage, ajoutée à leur apparente simplicité, qu'on tend à croire que la plupart des chercheurs, si pas tous, ont une bonne maîtrise des tests de comparaisons de moyennes [@aiken_doctoral_2008; @hoekstra_are_2012]. Pourtant, certains indices semblent contredire cette conviction.

Bien qu'il existe plusieurs types de tests t et d'ANOVA, les chercheurs en psychologie privilégient souvent par défaut le test t de Student et l'ANOVA de Fisher [3]. Ces tests consistent à comparer les scores moyens de deux (ou plusieurs) groupes indépendants de sujets, et reposent sur les hypothèses que les résidus, indépendants et identiquement distribués soient extraits d'une population qui se distribue normalement et qui a la même variance au sein de chaque groupe (c'est ce qu'on appelle la condition d'homogénéité des variances). Pourtant, on constate que dans les articles publiés, il n'est que rarement fait mention de ces conditions. @osborne_educational_2001, par exemple, avaient trouvé que seulement 8% des auteurs reportaient des informations sur les conditions d'application des tests, soit à peine 1% de plus qu'en 1969. Plus récemment, @hoekstra_are_2012 ont montré que sur 50 articles publiés en 2011 dans *Psychological Science* utilisant au moins une ANOVA, test-t ou régression, seulement trois discutaient des questions de normalité et d'homogénéité des variances. Par ailleurs, les informations reportées sont souvent non exhaustives [@counsell_reporting_2017], et la condition d'homogénéité des variances est encore moins fréquemment citée que celle de normalité. Parmi les 61 articles analysés par @keselman_statistical_1998, seulement 5% des articles mentionnaient simultanément les conditions de normalité et d'homogénéité des variances (et en tout, la condition de normalité était mentionnée dans 11% des cas, contre seulement 8% pour la condition d'homogénéité des variances). @golinski_expanding_2009 ont fait un constat similaire: parmi les 140 articles qu'ils ont analysé, seulement 11 mentionnaient explicitement la condition de normalité, contre 3 qui mentionnaient celle d'homogénéité des variances.

La non mention des conditions d'application dans les articles ne veut pas forcément dire qu'elles n'ont pas été prises en compte dans les analyses. On pourrait imaginer que les auteurs vérifient les conditions

[1] Parmi 68 articles analysés en 2013 par Counsell et ses collaborateurs (2017) dans 4 revues canadiennes, 92.7% incluaient au moins une analyse quantitative (contre 7.3% incluant une analyse qualitative)

[2] Les revues analysées étaient les suivantes: "Child Development", "Journal of Abnormal Psychology", "Journal of Consulting and Clinical Psychology", "Journal of Experimental Psychology: General", "Journal of Personality" et "Social Psychology"

[3] Parfois, ils le font de manière implicite, en indiquant réaliser un test t (ou une ANOVA) mais sans préciser duquel (ou de laquelle) il s'agit [retrouver référence]. Cela arrive même avec des méthodologistes! Dans l'article de Tomczak & Tomczak (2014), par exemple, ils parlent de l'ANOVA et du test t, sans précision, et ce n'est qu'en lisant l'ensemble de l'article qu'on comprend qu'en réalité, ils font allusion exclusivement au test t de Student et à l'ANOVA de Fisher, entre autres, parce qu'ils proposent d'associer ces tests à des mesures de taille d'effet qui impliquent l'usage du terme de variance poolée, qui sera décrit juste après.

d'application des tests mais ne le mentionnent la plupart du temps que lorsqu'elles sont violées [@counsell_reporting_2017]. @golinski_expanding_2009, par exemple, ont constaté à travers leurs revue de littérature que parmi les 11 articles qui mentionnaient la condition de normalité, 10 montraient une violation de cette dernière. Il est possible que motivés par le désir de rentabiliser l'espace disponible dans les manuscrits [@counsell_reporting_2017], les auteurs soient tentés de se limiter aux informations explicitement demandées par les éditeurs et les reviewers des journaux [@counsell_reporting_2017]. Or, les informations relatives aux conditions d'application des tests en font rarement partie. Par exemple, leur report n'est pas explicitement demandé dans le manuel des normes APA [@hoekstra_are_2012][4]. Dans un tel contexte, il n'y a que peu d'intérêt pour les chercheurs à en discuter, si ce n'est pour discuter des violations des conditions (et éventuellement, se servir de cette information pour justifier une décision qui en découle). Il est néanmoins surprenant de constater que de telles discussions apparaissent dans un pourcentage si faible d'articles, puisqu'il a été argumenté à de nombreuses reprises que le respect des conditions de normalité et d'homogénéité des variances est plus l'exception que la norme dans de nombreux domaines de la psychologie [@cain_univariate_2017; @micceri_unicorn_1989; @yuan_structural_2004; @erceg-hurn_modern_2008; @grissom_heterogeneity_2000]. Bien que l'on ne puisse totalement écarter la possibilité que certains chercheurs prennent des décisions inhérentes aux violations des conditions d'application sans le mentionner dans leur article, l'hypothèse privilégiée par @keselman_statistical_1998 est que la majorité des chercheurs applique des tests paramétriques indépendamment du fait que leurs conditions soient ou non respectées. Cette hypothèse semble confirmée par une expérience de @hoekstra_are_2012: afin d'étudier les pratiques des chercheurs lorsqu'ils étaient confrontés à un scénario qui impliquait la réalisation d'un test t, d'une ANOVA ou d'une régression linéaire, ces chercheurs ont observé 30 doctorants qui travaillaient depuis au moins deux ans dans des départements de psychologie aux Pays-Bas et qui avaient dû pratiquer tous ces tests au moins une fois. Alors que tous ont opté pour un test paramétrique, les conditions d'application de ces tests n'ont été testées que dans un faible pourcentage de cas. Après l'expérience, les 30 doctorants ont été soumis à un questionnaire. Celui-ci a révélé que la non vérification des conditions d'application des tests était dûe à leur manque de familiarité avec les conditions d'application des tests, plutôt que par un choix délibéré de leur part. Il est à noter qu'en réalité, vérifier les conditions d'application des tests est bien plus complexe qu'il n'y parait, et tout chercheur désireux d'améliorer la transparence dans la transmission des analyses de données resterait confronté à un problème majeur: les conditions d'homogénéité des variances et de normalité reposent sur les paramètres de *population* et non sur les paramètres d'*échantillon*. Comme ces paramètres de population ne sont pas connus [@hoekstra_are_2012], on doit utiliser les paramètres de l'échantillon pour tenter d'inférer sur le respect des conditions d'application. Souvent, les chercheurs font cette inférence en utilisant des tests d'hypothèses, mais il a été démontré que l'application d'un test conditionnellement aux résultats d'un test statistique préliminaire a pour effet d'augmenter l'erreur de type I [@schucany_preliminary_2006]. La difficulté que représente la vérification des conditions d'application ne constituerait pas réellement un problème, en soi, si les test t de Student et F de Fisher étaient susceptibles de fournir des conclusions non biaisées et fiables même en cas d'écarts à ces conditions, or ce n'est malheureusement pas toujours le cas. Ces tests sont particulièrement sensibles aux violations de la condition d'homogénéité des variances, et cette sensibilité est accentuée lorsque les échantillons n'ont pas tous la même taille [@keselman_statistical_1998].

Compte tenu de tous les éléments précités, il semblerait donc qu'une solution viable serait d'utiliser des tests qui ne reposent pas sur les conditions de normalité et d'homogénéité des variances. Il existe, par exemple, des tests qui reposent sur la comparaison d'autres indicateurs de tendance centrale que la moyenne (comme la moyenne trimmée), mais ces derniers font très souvent face à une forte résistance de la part des chercheurs, qui persistent à vouloir comparer les moyennes [@wilcox_how_1998; @erceg-hurn_modern_2008;@keselman_statistical_1998]. Dans la mesure où une revue approfondie de la littérature démontre que les taux d'erreur de type I et II des tests t de Student et F de Fisher sont bien plus affectés par le non respect de la condition d'homogénéité des variances que par le non respect de la condition de normalité [@grissom_heterogeneity_2000; @erceg-hurn_modern_2008; @hoekstra_are_2012;@osborne_four_2002], nous recommandons aux psychologues de remplacer les tests t de Student et F de Fisher par le test de Welch, un test de comparaison de moyennes qui ne requiert pas la condition d'homogénéité des variances. Cette solution a été suggérée par de nombreux auteurs avant

[4]Depuis l'article de Hoekstra et al. (2012), la septième édition du manuel des normes APA est parues. La mention explicite des conditions d'application ne fait pas partie des mises à jours proposées dans cette nouvelle édition.

nous [voir, par exemple @rasch_two-sample_2011;@ruxton_unequal_2006; @zimmerman_note_2004], pourtant, cela semble avoir eu un impact limité sur les pratiques des chercheurs en psychologie. Pour tenter de les influencer, nous nous sommes particulièrement appliqués, au sein des articles présentés dans les chapitres 2 à 3, à nous adresser directement à ce public de chercheurs. Pour ce faire, nous avons tenté (1) d'expliquer concrètement pourquoi selon nous, la condition d'homoscédasticité n'est pas réaliste, en nous appuyant sur des exemples directement issus de la recherche en psychologie, (2) de définir certaines notions statistiques de la manière la plus simple possible, en limitant les explications mathématiques et (3) d'illustrer graphiquement l'impact des violations de la condition d'homoscédasticité, plutôt que de fournir des tableaux de chiffres lourds et complexes. De plus, nous avons conclu ces articles par des recommandations concrètes, afin d'aider les chercheurs à extraire le message clé de ces articles. Ajoutons que les deux articles ont été soumis et publiés dans une revue Open Access (l'*International Review of Social Psychology*) afin d'assurer la diffusion la plus large possible de notre message.

Au delà des tests d'hypothèse, de nombreux journaux de psychologies encouragent (voire même requièrent) de quantifier la taille des effets étudiés et de fournir un intervalle de confiance autour des estimations de taille d'effet [@cumming_statistical_2012]. L'année 1999 a joué un rôle clé dans la mise en oeuvre de ces recommendations, puisque l'*APA Task Force* a publié un rapport dans lequel elle soulignait l'importance de reporter des mesures de taille d'effet. Ce rapport a été suivi de recommandations précises de la part de l'American Psychological Association (APA) et de l'American Educational Research Association (AERA) quant à la manière de reporter ces mesures [@peng_impact_2013]. Or, il semblerait que ces diverses recommandations aient été associées à des modifications dans les pratiques des chercheurs. @peng_impact_2013 ont étudié l'évolution du taux moyen de report des mesures de taille d'effet en comparant ce taux moyen avant et après 1999, distinctement dans 19 revues consacrées à la recherche dans les domaines de la Psychologie et de l'Education. Ils ont noté une augmentation de ce taux variant de 5.2 % à 96.3 % dans chacun de ces journaux. Ils ont cependant également noté la persistance de pratiques inadéquates, telles que la dominance de la mesure du traditionnel d de Cohen. Le d de Cohen est une mesure de taille d'effet standardisée qui appartient à la famille d et qui entretient une relation mathématique directe avec le t de Student. Par conséquent, il dépend des mêmes conditions d'application que le test t de Student, c'est donc sans surprise qu'en cas de violation de ces conditions, son usage peut amener à une sous-représentation (ou au contraire à une sur-représentation) de la taille d'effet [@grissom_review_2001]. De même que pour le test t, il semblerait que ce soit essentiellement la violation de la condition d'homogénéité des variances qui soit problématique. De nombreux auteurs se sont demandés si le d de Cohen pourrait être remplacé par une autre mesure de la même famille lorsque les variances diffèrent d'une population à l'autre, mais ils n'ont pas trouvé de consensus quant à la solution la plus appropriée [@shieh_confidence_2013]. C'est pour cette raison que nous avons décidé de réaliser des simulations Monte Carlo pour comparer le traditionnel d de cohen aux mesures de la famille d les plus communément proposées pour remplacer le classique d de Cohen en cas d'hétéroscédasticité. Les résultats de ces simulations sont présentés au sein de l'article du chapitre 4. Si nous nous focalisons exclusivement sur les mesures de la famille d dans ce chapitre, c'est parce que les chercheurs utilisent très fréquemment le d de Cohen lorsqu'ils réalisent un test t. Nous tenons cependant à rappeler que la notion de taille d'effet est très vaste. Elle englobe toute mesure susceptible de fournir une information relative à l'ampleur d'un effet étudié, que ce soit à travers une mesure *non standardisée* [moyenne, médiane, coefficient de régression non standardisé...) ou à travers une mesure *standardisées* (R^2, coefficient de régression standardisé, différence de moyennes standardisée..., @counsell_reporting_2017]. Les mesures non standardisées peuvent également s'avérer très utiles, et ce même si une emphase sur les tailles d'effet standardisées a pû donner l'impression que seules ces dernières étaient dignes d'intérêt [pour une discussion intéressante sur l'intérêt des mesures non standardisées, nous recommandons l'article de @pek_reporting_2018]. Au sein de ce chapitre, nous avons tenté de comparer l'efficacité de différents estimateurs sous des déviations réalistes de la condition de normalité, en nous appuyant sur l'investigation de @cain_univariate_2017. Ces auteurs avaient en effet calculé les indicateurs d'asymétrie et d'aplatissement [5] de 1567 distributions univariées provenant de 194 articles publiés dans *Psychological Science* (de Janvier 2013 à juin 2014) et *American Education Research Journal* (de janvier 2010 à juin 2014). Nous nous sommes inspirés des résultats de

[5] Nous utilisons ce terme "aplatissement" parce que c'est ainsi que l'on traduit communément le terme anglais "kurtosis". Comme nous le préciserons cependant dans le chapitre 2, cette traduction commune ne représente que mal cette mesure, qui nous informe plus sur la densité des distributions, au niveau des extrêmités, que sur leur aplatissement

cette étude pour définir le degré d'asymétrie et d'aplatissement des distributions dont étaient extraits nos échantillons, au sein de nos simulations. De plus, au delà des simulations, nous proposons des outils afin d'aider les chercheurs à calculer différents estimateurs de taille d'effet ainsi que les bornes de l'intervalle de confiance autour de ces estimateurs. Cela nous a semblé être une nécessité puisque malgré les recommandations, il semblerait que les mesures de taille d'effet soient rarement accompagnées d'un intervalle de confiance dans la littérature [@peng_impact_2013; @counsell_reporting_2017]. C'est vrai même lorsque ces mesures sont utilisées seules [indépendamment d'un test d'hypothèse, @counsell_reporting_2017]. Dans le cas des mesures standardisées, le calcul des intervalles de confiance est particulièrement complexe puisqu'il requiert l'usage des distributions non centrales [@balluerka_controversy_2005]. Cela nous a motivé à créer un package R en vue d'aider les chercheurs, ainsi qu'une application Shiny pour ceux qui ne sont pas familiers avec R. Ces outils seront fournis dans le chapitre 4.

Qu'elles soient standardisées ou non, les mesures de taille d'effet ainsi que leur intervalles de confiance sont parfois vus comme des outils qui permettent de combler certaines limites des tests d'hypothèse. Une critique fréquemment avancée à l'égard des tests d'hypothèse est le fait qu'un rejet de l'hypothèse nulle ne fournit qu'une idée de la direction de l'effet, sans information relative à son ampleur. Cette critique repose implicitement sur la conception d'après laquelle l'hypothèse nulle doit être définie comme l'absence d'effet (ou l'absence de différence entre les groupes). Il est vrai que c'est l'hypothèse nulle la plus couramment définie par les chercheurs [@nickerson_null_2000; @lakens_equivalence_2018]. Pourtant, lorsque c'est pertinent, il est possible d'incorporer la significativité pratique dans les tests d'hypothèse. Cela implique de réfléchir *a priori* aux effets qui présentent un intérêt pratique aux yeux des chercheurs et des praticiens [@fraas_testing_2000], ce qui peut se faire sur base de diverses considérations, tel que des comparaisons coûts/bénéfices, par exemple [@fraas_testing_2000]. Dans ce contexte, l'hypothèse nulle n'est plus que l'effet soit nul, mais qu'il ne dépasse pas une certaine valeur ou autrement dit, dans le cadre d'un test de comparaison de moyenne, que la différence de moyennes entre les groupes ne dépasse pas une certaine valeur [@newman_testing_2001]. Un rejet de l'hypothèse nulle ne constituera alors plus un soutien en faveur de n'importe quel effet non nul, mais plutôt un soutien en faveur d'un effet jugé pertinent. Il est également possible de montrer un soutien en faveur de l'absence d'un effet jugé pertinent, en définissant comme hypothèse nulle que l'effet dépasse une certaine valeur [@lakens_equivalence_2018]. C'est le principe des tests d'équivalence, qui feront l'objet du cinquième et dernier chapitre de cette thèse.