

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 quantitatives¹ [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 coeur de la grande majorité des programmes dans les domaines des Sciences Psychologiques et de l'Éducation [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 psychologie², 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 parmi 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 Fisher³. La statistique t de Student se calcule comme suit [student_probable_1908]⁴:

$$t_{Student} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{N-2}\right) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (1)$$

où N = le nombre total de sujets, et n_j , \bar{X}_j et S_j sont respectivement la taille, la moyenne et l'écart-type du $j^{ème}$ échantillon ($j = 1, 2$). La statistique F de Fisher se calcule comme suit:

$$F = \frac{\frac{1}{k-1} \sum_{j=1}^k \left[n_j (\bar{x}_j - \bar{x}_{..})^2 \right]}{\frac{1}{N-k} \sum_{j=1}^k \left[(n_j - 1) S_j^2 \right]} \quad (2)$$

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.

⁴ Dans l'article présenté au chapitre 2, nous nous contentons d'une formule simplifiée de la statistique t de Student: $t_{Student} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{N-2}\right) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$. Cette formule simplifiée est celle généralement présentée dans les articles et dans les

livres d'introduction aux statistiques, puisque la plupart du temps, on définit comme hypothèse nulle l'absence d'effet (avec $(\mu_1 - \mu_2)_0 = 0$)

où k est le nombre d'échantillons indépendants et S_j^2 est la variance du $j^{\text{ème}}$ échantillon ($1 \leq j \leq k$). Sous l'hypothèse de normalité, la statistique F suit loi de Fisher caractérisée par 2 paramètres:

$$df_1 = k - 1$$

$$df_2 = \sum_{j=1}^k n_j - k$$

Sous l'hypothèse de normalité, la statistique t de Student suit une distribution t avec $n_1 + n_2 - 2$ degrés de liberté.

Le test t de Student et l'ANOVA de Fisher consistent à comparer les scores moyens de deux (ou plusieurs) groupes indépendants de sujets. Les deux tests 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, requise pour pouvoir calculer le terme de variance poolée qui apparaît au dénominateur des équations (1) et (2)). 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.

Notons que 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 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]⁵. 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

⁵Depuis l'article de Hoekstra et al. (2012), la septième édition du manuel des normes APA est parue. La mention explicite des conditions d'application ne fait pas partie des mises à jours proposées dans cette nouvelle édition.

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 due 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 paraît, 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]. Tout ceci 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 que la solution la plus 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; kselman_statistical_1998]. Dans la mesure où une revue approfondie de la littérature démontre que le non respect de la condition d'homogénéité des variances affecte bien plus le taux d'erreur de type I ainsi que la puissance de tests t de Student et F de Fisher [grissom_heterogeneity_2000; erceg-hurn_modern_2008; hoekstra_are_2012; osborne_four_2002] que le non respect de la condition de normalité, 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 nous [voir, par exemple rasch_two-sample_2011; ruxton_unequal_2006; zimmerman_note_2004], pourtant, cela ne semble pas avoir eu d'impact sur les pratiques des chercheurs en psychologie. Afin de tenter de changer leurs pratiques, 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 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 recommandations, puisque c'est l'année où 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. Il semblerait même que cette dominance soit entretenue par certains articles méthodologiques, qui mentionnent cet estimateur sans envisager les autres alternatives appartenant à cette même famille [voir par exemple @tomczak_need_2014].

Le d de Cohen est une mesure de taille d'effet obtenue en divisant la différence de moyenne de chaque groupe par l'écart-type poolé:

$$d \text{ de Cohen} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1) \times S_1^2 + (n_2-1) \times S_2^2}{n_1 + n_2 - 2}}} = t_{Student} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (3)$$

Comme on le voit dans l'équation (3), le d de Cohen entretient une relation mathématique directe avec le t de Student et par conséquent, dépend des mêmes conditions d'application. 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]. Suite à ce constat, et en vertu des mêmes raisons qui nous ont amené à suggérer l'usage du test t de Welch à la place du test t de Student au sein du chapitre 2, nous avons décidé de réaliser des simulations Monte Carlo pour comparer le traditionnel d de cohen à plusieurs autres mesures appartenant à la famille d et qui sont jugées plus robustes aux écarts à la condition d'homogénéité des variances. Les résultats de ces simulations sont présentées au sein de l'article du chapitre 4⁶.

Comme pour les articles précédents, notre objectif était de nous adresser à un public de psychologues. Nous avons, entre autres, réalisé un effort particulier pour comparer l'efficacité des différents estimateurs retenus dans des scénarios qui correspondent à des déviations réalistes de la condition de normalité. Pour ce faire, nous nous sommes appuyés sur l'investigation de @cain_univariate_2017, qui avaient calculé les indicateurs d'asymétrie et d'aplatissement 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). De plus, bien que la comparaison des estimateurs reposait essentiellement sur leur propriétés inférentielles, nous avons tenté de prendre en compte la dimension interprétative de ces estimateurs dans nos discussions. Cela nous semblait important, puisqu'il semblerait que cette dimension ne soit pas suffisamment considérée par les psychologues. Même si les mesures de taille d'effet sont de plus en plus fréquemment reportées, elles ne sont que rarement interprétées et incluses dans les discussions [funder_evaluating_2019; thompson_statistical_1997]. Pour finir, on constate que contrairement aux recommandations, les mesures de taille d'effet sont rarement accompagnées d'un intervalle de confiance, que ce soit lorsque les mesures sont utilisées pour accompagner un test d'hypothèse ou même lorsqu'elles sont utilisées seules [counsell_reporting_2017]. Dans le cas des mesures standardisées, le problème est probablement accru puisque le calcul des intervalles de confiance requière un niveau de connaissance plus poussé que généralement enseigné [vérifier la source, je crois que c'est funder_evaluating_2019]. Dès lors, afin d'aider les chercheurs, nous avons mis en place des outils permettant de calculer l'ensemble des estimateurs mentionnés dans l'article et les bornes de leur intervalle de confiance. Dans un premier temps, nous avons créé un package R. Ensuite, afin d'aider les chercheurs qui ne sont pas familiers avec R, nous avons réalisé une application Shiny. Ces différents outils seront fournis dans le chapitre 4.

Notre espoir est que les chercheurs se mettent à utiliser plus fréquemment des outils adéquats et qu'à terme, ceux-ci puissent permettre la création d'hypothèses plus informées que la simple nil-nill hypothesis → tests d'équivalence.

⁶Notons 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 & Harlow, 2017). Au sein de ce chapitre, nous nous focaliserons exclusivement sur les mesures de la famille d . Nous tenons cependant à rappeler qu'il est souvent très utile de reporter également des mesures non standardisées, telles que les différentes brutes de moyennes, et ce même si une emphase sur les tailles d'effet standardisées a pu donner l'impression que seules ces dernières étaient dignes d'intérêt. Nous recommandons l'article de Pek & Flora (2018) pour une discussion intéressante à ce sujet.

Tests d'équivalence

une des plus grosses critiques du NHST, c'est la tendance des chercheurs à interpréter un effet NS comme l'acceptation de l'hypothèse nulle (Schmidt, 1996, cité par Harris, 1997).

Jusque là, nous nous sommes focalisés sur des tests traditionnels pour lesquels l'hypothèse nulle est l'absence d'effet. Avec de tel test, le rejet de l'hypothèse nul veut juste dire qu'il y a un effet non dû au hasard, mais c'est tout. On a vu commencer accompagner ces tests de mesure de taille d'effet, mais on est resté dans une approche exploratoire, *a posteriori*.

Or, utiliser les tailles d'effet *a priori* permettrait la réalisation de tests plus informatifs que le classique test visant à détecter l'absence d'effet (cf. tests d'équivalence).

D'après @lakens_practical_2021, un test d'hypothèse (selon l'approche de Neyman-Pearson) vaut la peine à 2 conditions:

- 1) que l'hypothèse nulle soit assez plausible pour que son rejet puisse surprendre au moins certains;
- 2) le chercheur veut appliquer une procédure méthodol qui l'autorise à prendre des décisions quant à la manière d'agir, tout en contrôlant le taux d'erreur. Agir peut vouloir dire: adopter un traitement, une politique, une intervention, ou abandonner un domaine de recherche, modifier une manipulation, ou de faire un certain type de déclaration ou revendication.

Comme déjà mentionné, l'hypothèse nulle est l'absence d'effet. On en reste sur la *nil-hypothesis*. Du coup, un effet significatif n'a pas vraiment de valeur. En réponse à ce problème, on a écrit deux articles:

- On peut commencer par ajouter une information sur les tailles d'effets (mais du coup ça n'oblige pas à réfléchir à l'avance à l'effet qui nous intéresse)

On peut aussi faire des tests plus informatifs (tests d'équivalence et/ou tests d'effets minimaux). *One of the most widely suggested improvements of the use of p values is to replace null-hypothesis tests (where the goal is to reject an effect of exactly 0) with tests of range predictions (where the goal is to reject effects that fall outside of the range of effects that is predicted or considered practically important) [@lakens_practical_2021].