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; autres exemples?]. 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 tendent à privilégier par défaut le test t de Student et l'$ANOVA$ de Fisher. *C'est tellement commun que les chercheurs ne précisent même plus toujours de quel test t ou de quel ANOVA il s'agit lorsqu'ils mentionnent faire référence à l'une de ces deux familes de tests [retrouver sur github la référence qui le dit]. Même dans certains articles méthodologiques on fait cela: dans l'article de @tomczak_need_2014, par exemple, ils parlent de l'ANOVA et du test $t sans préciser "Student" et "Fisher". Pourtant, on voit clairement que c'est de cela qu'ils parlent puisqu'ils associent des mesures de taille d'effet à ces statistiques qui reposent sur les mêmes conditions d'application que Student et Fisher.*

La statistique t de Student se calcule comme suit [@student_probable_1908]:

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

où $N =$ le nombre total de sujets, et n_j et $\bar{X}_j$ sont respectivement la taille et la moyenne du $j^{ème}$ échantillon ($j = 1, 2$). Sous l'hypothèse de normalité, la statistique t de Student suit une distribution t avec $n_1 + n_2 - 2$ degrés de liberté. La statistique F de Fisher se calcule comme suit:

$$F = \frac{\frac{1}{k-1}\sum_{j=1}^{k}\left[n_j\left(\bar{x_j} - \bar{x_{..}}\right)^2\right]}{\frac{1}{N-k}\sum_{j=1}^{k}\left[\left(n_j-1\right)S_j^2\right]} \tag{2}$$

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

$$df_1 = k - 1$$

[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"

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

Le test t de Student et l'$ANOVA$ 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][3]. Dans un tel contexte, il n'y a que peu d'intérêt pour les chercheurs à en discuter, si ce n'est pour justifier une décision inhérente à leur violation. 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 privilagiée 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

[3]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.

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;@keselman_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 Psychlogy*) afin d'assurer la diffusion la plus large possible de notre message.

Au là des tests d'hypothèses, il est de plus en plus fortement recommandé aux chercheurs de reporter une mesure de taille d'effet ainsi qu'un intervalle de confiance autour de cette mesure. Cette pratique est notamment approuvée dans la sixième édition du manuel de publication de l'American Psychological Association [APA; @american_psychological_association_publication_2010] ainsi que par l'American Educational Research Association [@duran_standards_2006]. Elle est également encouragée (voire même requise) par certains journaux de psychologie, tels que le *Journal of Consulting and Clinial Psychology* ou encore *Neuropsychology* [@cumming_statistical_2012]. Il s'agit là de réformes très encourageantes pour le devenir de la recherche en psychologie [@cumming_statistical_2012]. *@cumming_statistical_2012: le manuel de l'APA est utiilsé par plus de 1000 journaux dans pleins de disciplines et dans plein de pays à travers le monde. Ses conseils ont donc une influence énorme.*

1) On constate que c'est de plus en plus reporté
2) C'est une bonne chose car...
3) Cependant, le portrait n'est pas si optimiste car (limites), même si c'est en bonne voie.

Ces mesures sont-elles reportées/interprétées?

@tomczak_need_2014: - "the number of reports containing statistical estimates of effect sizes calculated after applyigng parametric tests is steadily increasing" (p. 19). - "By way of illustration, a meta-analysis of research accounts published in one prestigious psychology journal in the years 2009 and 2010 showed that

almost half of the articles reporting an ANOVA did not contain nay measure of effect size, and only a mere quarter of the surveyed research reports supplemented Student'es t-test analyses with information about the effect size" –> y'a une référence à un article [6] mais pas clair à quoi ça correspond

eFFECT SIZES

Pourquoi est-il important de les reporter?

@tomczak_need_2014:
(1) Alors que la p-valeur permet juste de déterminer si une différence existe entre les populations étudiées (autrement dit, que la différence entre les moyennes d'échantillonn n'est pas juste dûe au hasard), la taille d'effet permet de mesurer l'amplitude de la différence.
(2) Permet de comparer les effets observés dans différentes études (la p-valeur ne le permet pas, puisqu'elle dépend de la taille des échantillons)
(3) Permet de déterminer la taille d'échantillon requise pour avoir une bonne puissance statistique
(4) Dans les études pilotes, indicateur de futures attentes pour les recherches futures (forme d'analyse exploratoire?) –> j'utiliserais bien ce dernier argument pour embrayer sur les tests d'équivalence.

Quand oui, quelles mesures sont reportées?

Dans leur article, bien que @tomczak_need_2014 sugggèrent des tailles d'effet qu'on peut utiliser avec des tests non paramétriques, lorsqu'ils parlent des tests t et ANOVA, ils semblent (sans le préciser explicitement) faire référence exclusivement aux tests t de Student et à l'ANOVA de Fisher. Cela se voit au fait qu'ils suggèrent des mesures de taille d'effet qui sont mathématiquement liées à ces tests (tq le d de Cohen et g de Hedges, en fournissant les mesures qui impliquent la variance poolée).

Pour répondre à cette problématique, le chapitre 4 de cette thèse s'inscrit dans la continuité du chapitre 2.

En parlant des tailles d'effets, on commence de plus en plus à les utiliser (j'ai une réf qui le dit) mais: - on les calcule sans vraiment les comprendres/interpréter - comme pour le test t de Student et l'ANOVA, on utilise un test qui dépend des mm conditions d'application. Utiliser des tests plus adéquats permettrait d'améliorer les pratiques et à termes, de déterminer des mesures de taille d'effets qui pourront être utilisées a priori dans des tests plus informatifs que ceux visant à détecter l'absence d'effet (cf. tests d'équivalence).

Un paragraphe relatif à la taille d'effet. EN EXPLORATOIRE, ce qui à termes pourrait servir à définir des hypothèses plus informatives pour d'autres chercheurs, qui pourraient être utilisées, soit dans des tests d'effets minimaux, soit pour des tests d'équivalence. Et that's it.

Rem.: "une violation des conditions d'application peut amener à une sous- ou sur-estimation des mesure de taille d'effet (Osborne & Waters, 2002, cités par Hoekstra!)

Le NHST fait l'objet d'énormément de critiques, si bien que certains recommandent de le remplacer par une mesure de taille d'effet accompagnée d'un intervalle de confiance autour de la taille d'effet. Le raisonnement est que si l'IC contient la valeur 0, on ne peut conclure à une différence significative [@counsell_reporting_2017].

Une des principales critiques des tests d'hypothèse est le fait que l'on compare la différence observée à l'absence totale de différence (= un effet de 0). C'est une question qui est peu intéressante, car peu surprenante. Mais pourquoi comparer à 0 et pas à une autre valeur?

D'après @lakens_practical_2021, un test d'hypothèse (selon l'approche de Nayman-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 rechercher, modifier une manipulation, ou de faire un certain type de déclaration ou revendication.

@counsell_reporting_2017:*the constant calls for reporting effect sizes appears to have had an effect on the Canadian psychology articles as just over 90% of the analyses that used a significance test also included a*

standardized or unstandardized effect size. Few articles presented an effect size without hypothesis testing, and few of the analyses' results included a CI.

Ca se fait apparemment de plus en plus de reporter la taille d'effet (dans leur analyse de 151 études, 90% des analyses incluaient une mesure de taille d'effet, standardisée ou non... mais très peu incluaient les IC et de plus, ils les donnaient mais sans vraiment en discuter... @@counsell_reporting_2017 dans la discussion).

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)

Dans la revue de @keselman_statistical_1998, ils mentionnent que les tailles d'effet ne sont pratiquement jamais reportées malgré les recommandations du panuel de l'APA (1994) (et qu'elles ne sont fournies qu'en cas d'effet significatif).

- 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-ypothesis tests (where the goal is to reject ann 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].

Pourquoi jusque là la sauce n'a pas pris?

Je suis loin d'être la première à signaler tt ça. Ce qui manque encore dans mon plan d'introduction, c'est que je dois encore trouver le moyen de montrer en quoi mes articles sont une plus-value, ce qu'ils apportent. 2) Parler des packages, des applications Shiny, etc.

PARLER DES DIFFERENTES REVUES DE LITTERATURE QUI LE DISENT.

Qu'est-ce qui pourrait expliquer cela? 1) @sharpe_why_2013: lack of awareness (p.573) Manque de conscience des développements dans le domaine?
2) @sharpe_why_2013: journal edotors (p.573) Les éditeurs ne poussent pas assez? –> Pas convaincue que ça m'intéresse
3) @sharpe_why_2013: Publish or perish? (p.574) je ne comprends mm pas en quoi c'est un argument
4) @sharpe_why_2013: Software (p.574) –> aaahh! Certaines pratiques comme les équations structurelles et les analyses de puissance ont été facilitées par des software comme gpower. Cela explique leur popularité. En ce qui concerne les statistiques plus robustes, par contre, elles ont moins de succès car non dispo dans les softwares dispo. Les gens veulent juste qu'on leur dise où cliquer pour avoir le test qu'ils veulent! C'est triste mais faut faire avec (à mon avis).
5) @sharpe_why_2013: inadequate education (p.574)
6) @sharpe_why_2013: mindset: facteurs psychologiques t.q. la peur de dévier des pratiques courantes (comme si on n'allait pas être publié si on ne faisait pas comme tlm).

Anecdote: les chercheurs font souvent l'erreur de croire qu'il faut vérifier la normalité de la VD en faisant une régression. Dans SPSS, il est assez complexe de le faire car il faut d'abord calculer les résidus, ce qui implique de comprendre que les tests t et ANOVA sont des cas particuliers de régression, puis ensuite a posteriori représenter graphiquement les résidus. C'est chronophage et complexe. Dans Jamovi, par contre, la vérification de la normalité des résidus est automatiquement réalisée lorsqu'on fait un test t. Le rôle des méthodologistes, à mon sens, est de prémacher le travail, pour permettre à d'autres de créer des outils conçus pour améliorer les pratiques de recherche. à partir du moment où c'est automatiquement fait correctement, il devient moins problématique que les psychologues maîtrisent le détail. Débarassés de ces questions, ils pourront peut-être alors plus se focaliser sur l'important pour mieux comprendre et interpréter les résultats de leur tests: càd comprendre la distribution d'échantillonnage, dont pratiquement tt découle.