Université Libre de Bruxelles

Faculté des Sciences Psychologiques et de l'Education

TITRE,

TITRE SUITE

Par

MARIE DELACRE

En vue de l'obtention du grade de docteur

Septembre 2021

Abstract

Abstract (voir nombre de mots requis)

Table des matières

Remerciements

Thank you for following this tutorial!

I hope you'll find it useful to write a very professional dissertation.

1 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 & McGarty, 2014). Or, dans un domaine dominé par les analyses quantitatives [1] (Counsell & Harlow, 2017), les connaissances statistiques s'avèrent fondamentales pour comprendre, planifier et analyser une recherche (Everitt, 2001 ; Howitt & Cramer, 2017). Les statistiques font dès lors partie intégrante du cursus de formation des psychologues et jouent un rôle très important dans leur parcours (Hoekstra, Kiers, & Johnson, 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, West, & Millsap, 2008 ; Curtis & Harwell, 1998 ; Golinski & Cribbie, 2009) et des livres d'introduction aux statistiques pour psychologues [Field (2013); autres exemples?]. Cela pourrait vraisemblablement expliquer pourquoi ils sont si persistants dans la recherche en psychologie (Counsell & Harlow, 2017). Ces tests sont les plus fréquemment cités dans la littérature scientifique depuis plus de 60 ans(Byrne, 1996 ; Golinski & Cribbie, 2009 ; Nunnally, 1960). Dans une revue de 486 articles publiés en 2000 dans des journaux populaires en psychologie [2], Golinski & Cribbie (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 & Harlow (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 et al., 2008 ; Hoekstra et al., 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. La statistique t de Student se calcule

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"

comme suit (Student, 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:

$$df_1 = k - 1$$

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

Comme le révêlent les équations (1) et (2), on compare les moyennes de chaque groupe à travers ces deux tests. De plus, la variance poolée intervient au dénominateur des statistiques t de Student et F de Fisher. Or, utiliser ce terme n'a théoriquement de sens que si les scores de chaque groupe sont extrait de distributions ayant la même variance (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 des conditions de normalité et d'homogénéité des variances. Osborne & Christianson (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 et al. (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 & Harlow, 2017), et la condition d'homogénéité des variances est encore moins fréquemment citée que celle de normalité. Parmi les 61 articles analysés par Keselman et al. (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 & Cribbie (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 ne pas mentionner les conditions d'application 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 & Harlow, 2017 ; Golinski & Cribbie, 2009). Golinski & Cribbie (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 & Harlow, 2017), les auteurs soient tentés de se limiter aux informations explicitement demandées par les éditeurs et les reviewers des journaux (Counsell & Harlow, 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 et al., 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, Zhang, & Yuan, 2017 ; Erceg-Hurn & Mirosevich, 2008 ; Grissom, 2000 ; Micceri, 1989 ; Yuan, Bentler, & Chan, 2004). 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 soutenue par Keselman et al. (1998) il y a plus de 20 ans, d'après laquelle la majorité des chercheurs applique des tests paramétriques indépendamment du fait que leurs conditions soient ou non respectées, nous semble plus probable. 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, Hoekstra et al. (2012) 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. 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

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.

variances et de normalité reposent sur les paramètres de *population* et non sur les paramètres d'échantillon. Or, ces paramètres de population ne sont pas connus (Hoekstra et al., 2012). A cause de cela, les chercheurs qui vérifient les conditions d'application des tests le font généralement à travers l'usage de tests statistiques, ce qui n'est pas approprié, puisqu'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 & Tony Ng, 2006). Tout ceci ne constituerait pas 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, comme nous l'élaborerons dans les chapitres 2 et 3 de cette thèse.

Compte tenu de tous les éléments précités, il semblerait donc que la seule alternative viable soit d'utiliser des tests qui ne reposent pas sur ces conditions. Il existe, par exemple, des tests qui reposent sur la comparaison d'autres indicateurs de tendance centrale que la moyenne (comme la moyenne trimmée, par exemple), mais ces dernières font très souvent face à une forte résistance de la part des chercheurs, qui persistent à vouloir comparer les moyennes (Erceg-Hurn & Mirosevich, 2008 ; Keselman et al., 1998 ; Wilcox, 1998).

Nous noterons, à travers une revue approfondie de la litérature, qu'un non respect de la condition d'homogénéité des variances affecte particulièrement le taux d'erreur de type I ainsi que la puissance de ces tests (Erceg-Hurn & Mirosevich, 2008 ; Grissom, 2000 ; Hoekstra et al., 2012 ; Osborne & Waters, 2002), bien plus que la condition de normalité. Pour cette raison, au sein des chapitres 2 et 3 de cette thèse, nous proposons de remplacer l'usage des tests t de Student et F de Fisher par d'autres tests de comparaisons de moyennes qui ne reposent pas sur la condition d'homogénéité des variances, à savoir les tests de Welch.

Dernier paragraphe: nous ne sommes pas les premiers à le faire, pourtant la recommandation ne semble pas avoir pris jusque là. Afin d'éviter de produire de nouveaux articles qui n'auront pas plus d'impacts que les autres, nous nous sommes particulièrement appliqués à parler de langage des psychologues. Nous fournirons de nombreux exemples issus de la psychologie et illustrons concrètement les conséquences des violations. Enfin, nous conclusions en donnant des recommandations facilement applicables, càd des solutions disponibles dans les logiciels les plus utilisés par les psychologues. *Relire les recommandations de mes deux articles.*

Keselman et al. (1998) écrit ceci: "Despite these repeated cautionary notes, behavioral science researchers have clearly not taken this mesasge to heart. It is strongly recommended that test procedures that have been desiged specifically for use in the presence of variance heterogeneity and/or

nonnormality be adopted on a routine basis" (p.358). Rem.: ils parlent d'un article de Lix et al. (1996) qui mentionne des packages qui permettent de le faire mais l'article est introuvable sur google scholar. L'open access est une des clés pour moi. alternatives robustes peu ou pas utilisées, et ce malgré de nombreuses tentatives pour changer cela (Keselman et al., 1998)

2) Par manque de connaissances, les chercheurs se contentent souvent des informations fournies dans les logiciels clic/bouton. *for example, if software does not report a CI on Cohen's d, it is unlikely that a researcher will calculate one his or herself* (Counsell & Harlow (2017)).

1.0.1 Limite 2: hypothèse nulle

Effect sizes are an important outcome of empirical research. Moving beyond decisions about statistical significance, there is a strong call for researchers to report and interpret effect sizes and associated confidence intervals. This practice is highly endorsed by the American Psychological Association (APA) and the American Educational Research Association (American Educational Research Association, 2006; American Psychological Association, 2010).

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 & Harlow, 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 (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 & Harlow (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 & Harlow (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 et al. (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, 2021).

1.0.2 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.

D'aucun on fait le constat d'un fossé entre les méthodes inférentielles recommandées dans la littérature scientifique et les techniques réellement utilisées par les chercheurs appliqués [keselman_statistical_1998].

PARLER DES DIFFERENTES REVUES DE LITTERATURE QUI LE DISENT.

Qu'est-ce qui pourrait expliquer cela? 1) Sharpe (2013): lack of awareness (p.573) Manque de conscience des développements dans le domaine?
2) Sharpe (2013): journal edotors (p.573) Les éditeurs ne poussent pas assez? –> Pas convaincue que ça m'intéresse
3) Sharpe (2013): Publish or perish? (p.574) je ne comprends mm pas en quoi c'est un argument
4) Sharpe (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 (2013): inadequate education (p.574)
6) Sharpe (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.

2 Conclusion

The conclusion section should specify the key findings of your study, explain their wider significance in the context of the research field and explain how you have filled the knowledge gap that you have identified in the introduction. This is your chance to present to your reader the major take-home messages of your dissertation research. It should be similar in content to the last sentence of your summary abstract. It should not be a repetition of the first paragraph of the discussion. They can be distinguished in their connection to broader issues. The first paragraph of the discussion will tend to focus on the direct scientific implications of your work (i.e. basic science, fundamental knowledge) while the conclusion will tend to focus more on the implications of the results for society, conservation, etc.

2.0.1 Usage des articles méthodo

Ces derniers semblent assez peu utilisés par les chercheurs. En tout cas s'ils les utilisent, ils les citent très peu dans leurs références pour justifier leurs choix (cf. article de Mills, Abdulla, & Cribbie (2010): le mode du nombre de citation méthodo dans les articles de recherche appliqués est 0, et la médiane vaut 1...). Dans l'autre sens, on constate que les articles méthodos sont généralement peu cités, et ils le sont encore 3 fois moins par les chercheurs appliqués que par les autres méthodologistes (Mills et al., 2010, p. 56).

On est en droit de questionner l'impact réel des publications méthodologiques, pour 2 raisons, d'après Mills et al. (2010):
(1) 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.
(2) malgré que des nouvelles méthodes sont disponibles, les chercheurs continuent à opter pour des tests traditionnels et familiaux (mais souvent inappropriés).
–> Qu'est-ce qui va pousser les chercheurs à lire des articles méthodos? –> Si je trouve la réponse à ceci, j'ai mon intro.

2.0.2 Importance des simulations et des logiciels mordernes pour enseigner les statistiques fréquentistes

On sait que les chercheurs tendent à privilégier les méthodes qui sont proposées par défaut dans des logiciels de clique bouton (comme SPSS). C'est en tout cas ce que dit Counsell & Harlow (2017) dans le contexte de la gestion des données manquantes (mais je crois que c'est vrai pour tout). Une manière d'améliorer les pratiques serait d'améliorer les options proposées par défaut dans les logiciels de

clic-bouton. C'est à ce genre de choses que j'aspire à travers mes articles.

Malgré tout, 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.

2.0.3 Comment écrire/transmettre l'info aux psys

Un consultant doit pouvoir parler de langage des psycs, c'est-à-dire décrire et expliquer les méthodes requises d'une manière compréhensible pour les clients (Golinski & Cribbie, 2009). Est-ce bien de demander à des mathématiciens/Statisticiens d'enseigner les stat aux psy's? Par forcément, car un psychologue spécialisé en méthodo quanti sera plsu à mˆmee de comprendre les procédures et méthodes requises par les psys (ex. de la question de la taille d'effet qui n'intéresse pas vraiment les statisticiens; Golinski & Cribbie (2009)).

2.0.4 Recommandations générales

Mills et al. (2010):
- au moins un reviewer compétant pour analyser le caractère approprié des méthodes stat
- que les éditeurs/reviewers encouragent l'usage d'article de méthodo dans leur recherche –> interesting, mais réaliste? La proportion de méthodologistes parmis les psychologues n'est pe pas assez élevée... Ou alors il faut vraiment de l'interdisciplinarité!

3 Bibliographie

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist*, *63*(1), 32.

Byrne, B. M. (1996). The status and role of quantitative methods in psychology: Past, present, and future perspectives. *Canadian Psychology/Psychologie canadienne*, *37*(2), 76.

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior research methods*, *49*(5), 1716-1735.

Counsell, A., & Harlow, L. (2017). Reporting practices and use of quantitative methods in Canadian journal articles in psychology. *Canadian Psychology/psychologie canadienne*, *58*(2), 140.

Curtis, D. A., & Harwell, M. (1998). Training doctoral students in educational statistics in the United States: A national survey. *Journal of Statistics Education*, *6*(1).

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*(7), 591.

Everitt, B. S. (2001). *Statistics for psychologists: An intermediate course*. (S.l.) : Psychology Press.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. (S.l.) : sage.

Golinski, C., & Cribbie, R. A. (2009). The expanding role of quantitative methodologists in advancing psychology. *Canadian Psychology/Psychologie canadienne*, *50*(2), 83.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of consulting and clinical psychology*, *68*(1), 155.

Haslam, S. A., & McGarty, C. (2014). *Research methods and statistics in psychology*. (S.l.) : Sage.

Hoekstra, R., Kiers, H., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in psychology*, *3*, 137.

Howitt, D., & Cramer, D. (2017). *Understanding statistics in psychology with SPSS*. (S.l.) : Pearson London, UK:

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., ... Keselman, J. C. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of educational research*, *68*(3), 350-386.

Lakens, D. (2021). The practical alternative to the p value is the correctly used p value. *Perspectives on psychological science*, *16*(3), 639-648.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological bulletin*, *105*(1), 156.

Mills, L., Abdulla, E., & Cribbie, R. (2010). Quantitative methodology research: Is it on psychologists' reading lists?

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, *20*(4), 641-650.

Osborne, J. W., & Christianson, W. R. (2001). Educational Psychology from a Statistician's Perspective: A Review of the Quantitative Quality of Our Field.

Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical assessment, research, and evaluation*, *8*(1), 2.

Schucany, W. R., & Tony Ng, H. K. (2006). Preliminary goodness-of-fit tests for normality do not validate the one-sample Student t. *Communications in Statistics-Theory and Methods*, *35*(12), 2275-2286.

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological methods*, *18*(4), 572.

Student. (1908). The probable error of a mean. *Biometrika*, 1-25.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*(3), 300.

Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika*, *69*(3), 421-436.

4 Annexe(s)

4.1 Annexe A: erratum

4.1.1 Why psychologists Should by Default Use Welch's t-test Instead of Student's t-test

4.1.1.1 Erreurs conceptuelles

4.1.1.2 Mise en forme et Notations En termes de mises en forme, nous avons omis à plusieurs endroit d'italiser les notations mathématiques. Par exemple, à la page [93], nous avons indiqué "F-ratio test" au lieu de "F-ratio test" à plusieurs reprises. A la même page, nous avons également noté "$\mathrm{x_{ij}}$" au lieu de "x_{ij}", et $|\mathrm{x_{ij}} - \hat{\theta}_{\mathrm{j}}|$ au lieu de $|x_{ij} - \hat{\theta}_j|$.

En termes de notation, nous avons relevé des inconsistantes dans la notation de la variance de chaque groupes. Par exemple, dans l'équation 1, nous utilisons S_j^2, alors que nous utilisons s_j^2 dans l'équation 4 ou encore SDj lorsque nous définission la statistique du F-ratio à la page [93].

4.1.2 Taking parametric assumptions very seriously

4.1.2.1 Erreurs conceptuelles

4.1.2.2 Mise en forme et Notations Dû à un manque de connaissance de Latex lors de mes premières tentatives d'écritures d'articles via Rmarkdown, certaines majuscules sont manquantes dans les références bibliographiques. S'assurer qu'une lettre apparaissent en majuscule, via latex, implique de l'entourer des symboles {}, ce qui n'a pas été fait. Par exemple, dans le titre de l'article de Tiku(1971), il aurait fallu indiquer "Power function of the {F}-test.." au lieu de "Power function of the F-test...". Cela ne serait pas arrivé, si j'avais utilisé un outil comme Zotero, afin d'exporter directement un fichier au format Bibtex (puisque via ces outils, ce genre de détail est automatiquement inclu), mais je n'ai découvert cette possibilité que très récemment.

4.1.3 Effect sizes

4.1.4 Equivalence tests

4.2 Annexe B

Insert code (if any) used during your dissertation work here.