Introduction

Truc qui n'ont rien à voir mais dont je pourrais avoir besoin à la défense

"A review by van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, and Depaoli (2017) revealed that 31% of articles in the psychological literature that used Bayesian analyses did not even specifiy the prior that was used, at least in part because the defaults by the software package were used. Mindless statistic are not limited to pvalues" (dans l'article de Daniel... j'adore cet argument!)

Début de la vraie intro

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

[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)

[(**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.

Limite 1: conditions d'application

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 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

[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:

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

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), les deux tests consistent à comparer les moyennes de chaque groupe (ce qui sous-tend l'hypothèse de normalité). De plus, la variance poolée intervient au dénominateur des statistiques t de Student et F de Fisher. Or, utiliser la variance poolée n'a théoriquement de sens que si la condition d'homogénéité des variances est respectée. 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_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 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_reporting_2017?**; **golinski_expanding_2009?**). (**golinski_expanding_2009?**), par exemple, ont relevé que parmi les 11 articles de leur revue de littérature 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 manuscripts (**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?**).*sûr pour la 6ème édition, à confirmer pour la 7ème, mais dans l'article qui résume les différence entre les 2 éditions, je n'ai rien vu à ce sujet.* Dans un tel contexte, il n'y a que peu d'intérêt pour les chercheurs à mentionner les conditions lorsque ces dernières ne semblent pas violées. Il est cependant peu probable que cela suffise à expliquer pourquoi les conditions d'application soient mentionnées 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?**). In fine, il est plus probable que les chercheurs appliquent des tests paramétriques sans vérifier s'ils sont dans de bonnes conditions pour le faire. C'était déjà l'hypothèse soutenue il y a plus de 20 ans par (**keselman_statistical_1998?**), et qui semble confirmée par l'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 auteurs 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. Ils ajoutent que lorsque les conditions sont vérifiées, elles le sont d'une manière inappropriée, soit souvent à travers l'usage de tests statistiques, alors qu'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?**).

Ce qui rend cela si problématique, c'est que le résultat du test t de Student et de l'$ANOVA$ ne sont pas toujours fiables lorsque leurs conditions d'application ne sont pas respectées. *aller relever toutes ce qui relève de ça dans mes intros des 2 premiers articles - (**erceg-hurn_modern_2008?**) (cité par moi): "if variances are not equal across groups and sample sizes differ across independent groups, Student's t-test can be severly biased and lead to invalid statistical inferences)."*

- *Il a été argumenté à de nombreuses reprises que dans certains cas, la violation des conditions de normalité et d'homogénéité des variances peut entraîner une augmentation du risque alpha mais également amener à une perte de puissance (**hoekstra_are_2012?; osborne_four_2002?**).*

On pourrait alors argumenter qu'il serait nécessaire d'améliorer la transparence dans la transmission des analyses de données (**counsell_reporting_2017?**), de s'assurer de toujours s'assurer de systématiquement faire mention des conditions des tests utilisés, afin de rassurer le lecteur sur la confiance qu'il peut accorder à la fiabilité des résultats

(**osborne_four_2002?**). Pourtant, même si cette recommandation était d'application, il resterait toujours deux problèmes majeurs. D'abord, les conditions d'homogénéité des 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. Ensuite, comme déjà mentionné, ces conditions ne seront souvent pas respectées au sens strict du terme, puisqu'elles sont définies de manière très stricte (par exemple, l'hypothèse d'homogénéité des variances est que la variance des deux populations sont sont extraits les échantillons soient exactement égales) (**hoekstra_are_2012?**).

Il semblerait donc que la seule alternative viable soit d'utiliser des tests qui ne reposent pas sur ces conditions. *Il existe des alternatives robustes, mais elles n'ont que très peu de succès. Il y a une forte résistance de la part des chercheurs blablabla.* Comment s'assurer alors que ces recommandations soient enfin prises en compte? Réponse: en parlant le langage des psychologues. En donnant des exemples issues de la psychologie, en réalisant des simulations et en illustrant concrètement les conséquences. Enfin en donnant des recommandations facilement applicables, càd des solutions disponibles dans les logiciels les plus utilisés par les psychologues. C'est ainsi qu'on en est venu à proposer la solution du Welch et à écrire les articles 1 & 2.

En proposant des switchs faciles. L'usage des tests de Welch est un bel exemple de switch facile. Ceci dit, ce n'est pas parcpe que le switch est facile qu'il est forcément fait: (**keselman_statistical_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. w *alternatives robustes peu ou pas utilisées, et ce malgré de nombreuses tentatives pour*

*changer cela (**keselman_statistical_1998?**)* ARTICLE1 ARTICLE2.

Lié au manque de connaissance liées aux conditions: [Le théorème central limite, par exemple, fait l'objet d'incompréhensions, si bien que même des chercheurs expérimentés ont l'intuition erronée que la loi des grands nombres s'applique également à de petits nombres (Braver et al., 2014).]

–> POURQUOI ELLES NE SONT PAS VERIFIEES? -> 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_reporting_2017?**)).

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_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.

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_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.