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!)

Elements qui n'entrent pas dans ma structure de base mais voir si et comment les intégrer dans l'intro

Problème lié au stat en général

J'ai l'impression que ça parle surtout des stat plus complexes mais au cas où ça servirait plus tard...

Parmi les milliers d'étudiants diplomés chaque année en psychologie au Canada, la plupart n'ont pris qu'un seul cours de statistiques et ont des connaissances très limitées en ce qui concerne l'analyse de données [@golinski_expanding_2009]. De plus, parmi les doctorants, beaucoup n'ont pas les connaissances requises pour faire face à la complexité des méthodes statistiques modernes qui apparaissent en psycho et doivent mobiliser des ressources pour les aider...
(1) les psys qui se spécialisent dans une méthode quanti (ex.: modèles d'équation structurée)
(2) des consultants (appartenant souvent à (1))

(3) des articles statistiques. Mais 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_quantitative_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_quantitative_2010, p.56].

–> Qu'est-ce qui va pousser les chercheurs à lire des articles méthodos?

RELIRE CECI POUR VOIR QUOI FAIRE DE CETTE IDEE: les tests conventionnels (tests t, ANOVA et régression) sont faussement simples, car ils reposent sur des notions importantes et complexes telles que la distribution d'échantillonnage et la p-valeur, notions malheureusement encore mal comprises par nombreux chercheurs. Bien comprendre celle-ci requiert du temps et de la patience. C'est un travail de longue haleine. Il m'a fallu des années de simulations et d'études pour réellement comprendre des tests qui semblent aussi simples que le test t, alors que je suis passionnée. Cela montre à quel point il est important pour les psychologues que des personnes prennent le temps de vulgariser le travail pour eux. Des outils comme Jamovi sont mis en place afin de faciliter la vie des chercheurs.

On sait que les chercheurs tendent à priilégier les méthodes qui sont proposées par défaut dans des logiciels de clique bouton (comme SPSS). C'est en tout cas ce que dit @counsell_reporting_2017 dans le contexte de la gestion des données manquantes (mais je crois que c'est vrai pour tout). Pour continuer à alimenter et mettre à jour ces outils le mieux possible, d'une manière pas trop douloureuse pour le chercheur lambda, il est important que des méthodologistes continuent à jouer le pont entre psychologues et statisticiens comme j'essaye de le faire.

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_expanding_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_expanding_2009).

Pourquoi un article ne suffit pas et que les logiciels comptent (ex. Jamovi, R...)

On est en droit de questionner l'impact réel des publications méthodologiques, pour 2 raisons, d'après @mills_quantitative_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).

Recommandations générales

@mills_quantitative_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 –> interesting, mais réaliste? La proportion de méthodologistes parmis les psychologues n'est pe pas assez élevée... Ou alors il faut vraiment de l'interdisciplinarité!

Début de la vraie intro

Les statistiques font partie intégrante du cursus de formation des psychologues et jouent un rôle très important dans le parcours des chercheurs en psychologie, et des psychologues de manière générale. On attend d'eux 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 [@counsell_reporting_2017] [1], les connaissances statistiques s'avèrent fondamentales pour comprendre, planifier et analyser une recherche [@howitt_understanding_2017; @everitt_statistics_2001].

Au cours des 50 dernières années, on a assisté à un développement inouï de méthodes statistiques de plus en plus complexes: méta-analyses, équations structurelles, modèles linéaires hiérarchiques... [@sharpe_why_2013]. Pourtant, les tests de comparaison de moyennes (suivies par les régressions avec prédicteurs continus) restent dominants en psychologie [@counsell_reporting_2017]. Ces tests restent, depuis plus de 50 ans [@nunnally_place_1960], les tests les plus fréquemment cités dans la littérature scientifique [@golinski_expanding_2009; @delacre_why_2017; @delacre_taking_2019]. @golinski_expanding_2009, par exemple, ont analysé 486 articles publiés en 2000 dans des journaux populaires en psychologie (*Child Development; Journal of Abnormal Psychology; Journal of Consulting and Clinical Psychology; Journal of Experimental Psychology: General; Journal of Personality and Social Psychology*) et ont relevé que pas moins de 140 d'entre eux ($\approx$ 29%) 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. JE SAIS PAS SI JE PEUX ME CITER MOI DANS L INTRO?

[- Sur un total de 282 études reportées dans *Social Psychological & Personality Science* entre avril 2015 et avril 2016, 97 ($\approx$ 34%) utilisaient un test-*t* [@delacre_why_2017].
- Quant aux ANOVA à un facteur, nous en avons trouvé dans 14% des 116 articles publiés en 2016 dans le *Journal of Personality and Social psychology* [@delacre_taking_2019]].

Les tests de comparaison de moyennes se trouvent également au choeur des programmes de statistiques au sein des pratiquement toutes les facultés de psychologie (vu pour le Canada et les USA en 2008, ESSAYER DE TROUVER UNE REF QUI LE DIT PR L EUROPE) [@aiken_doctoral_2008; @golinski_expanding_2009] et sont mentionnés dans pratiquement tous les livres d'introduction aux statistiques pour psychologues. ESSAYER DE TROUVER DES LIVRES QUI EN PARLENT THEORIQUEMENT ET D AUTRES QUI LES METTENT EN LIEN AVEC DES LOGICIELS CONNUS.

Compte tenu de tous ces éléments, nous serions en mesure d'espérer que ces tests soient parfaitement connus et maîtrisés, pourtant c'est nettement moins le cas qu'on ne pourrait le croire. Si d'aucun admettent assez

[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)

facilement que la plupart des étudiants diplômés en psychologie n'ont pas les connaissances suffisantes en analyse de données pour aborder des outils aussi complexes que les SEM et modèles linéaires hiérarchiques *Parmi les milliers d'étudiants diplomés chaque année en psychologie au Canada, la plupart n'ont pris qu'un seul cours de statistiques et ont des connaissances très limitées en ce qui concerne l'analyse de données [@golinski_expanding_2009]*, beaucoup croient que leurs étudiants ont une bonne maîtrise des tests t et des ANOVA. Par exemple, parmi les professeurs interrogés en 2008 par @aiken_doctoral_2008 au sein de 201 départements, professeurs conférant des programmes de doctorat en psychologie aux USA et au Canada, presque tous (entre 80% et 100%) affirmaient que la plupart (si pas tous) leurs étudiants étaient capable de réaliser une ANOVA à un ou plusieurs facteurs (que ce soit l'approche a priori ou post hoc). Mais n'est-ce pas un peu trop optimiste?

La manière dont l'ANOVA et les tests t sont utilisés en psychologie.

Les tests t et ANOVA ont la fausse réputation d'être simples et maîtrisés par la plupart. Pourtant, encore aujourd'hui, des nombreux manquements dans l'usage de ces tests peuvent être soulignés. C'est ce que nous tentons de faire à travers cette thèse.

Très classiquement, lorsqu'on compare 2 ou plus de 2 groupes définis par les modalités d'un ou plusieurs facteurs catégoriels, on "résume" la distribution des scores au sein de chaque groupe par un indicateur de tendance centrale (la moyenne), et on postule que tous les échantillons sont extraits de populations dont les moyennes sont identiques ($H_0 : \mu_1 = \mu_2 = ... = \mu_k$). On réaliste un test t (de Student) ou une ANOVA (de Fisher) et Si la p-valeur est inférieure au risque alpha, on conclut au rejet de l'hypothèse nulle.

Biensûr, la moyenne n'est pas le seul indicateur de tendance centrale et des tentatives ont été faites pour introduire d'autres indicateurs dans les tests (citer Wilcox par exemple), mais ces alternatives ont eu nettement moins de succès [@keselman_statistical_1998]. (Remarque: @keselman_statistical_1998 ajoute que les post hoc de Tukey sont les préférés aussi).

Et même en se limitant aux comparaisons de moyennes, il existe des alternatives au test t de Student et à l'ANOVA de Fisher (cf. tests robustes), mais leur adoption est très compliquée. @counsell_reporting_2017: il est commun que les chercheurs disent faire "une ANOVA ou un test t", mais sans préciseri de quel ANOVA ou de quel test t il s'agit. Manque de précision!

Limite 1: conditions d'application

Les tests statistiques conduiront à une interprétation valide à condition de respecter une ou plusieurs conditions d'application [@counsell_reporting_2017]. Un non respect des conditions d'application pourra sérieusement affecter le taux d'erreur de types I et II des tests. Pourtant, ces conditions ne sont que peu ou pas vérifiées [@hoekstra_are_2012].

- @keselman_statistical_1998: sur 61 articles contenant des designs dans lesquels des échantillons indépendants étaient définis par un seul facteur (between-subjects univariate designs, 13 n'ont pas donné d'information relatives aux écart-types. Parmi les autres, il y avait de l'hétéroscédasticité. En tout, moins d'un article sur 5 se préoccupait des violations possibles des conditions d'application, et lorsqu'ils le faisaient, ils s'inquiétaient surtout de la condition de normalité (moins de l'hétéroscédasticité), et ce malgré le fait que pourtant, des violations de cette conditions sont bien plus dommageables!

- voir dans mes 2 articles sur le Welch, je donne des exemples aussi de ça
- @golinski_expanding_2009: parmi 140 articles publiés dans plusieurs revues célèbres qui utilisaient l'ANOVA (voir plus haut), très peu vérifiaient les conditions d'application. *"It is interesting that, of the 11 articles that mentioned the normality assumption, 10 found distributions thate were nonnormal. Althought it is possible that the remaining articles taht did not mention th enormality assumptiona ll found noevidence of nonnormality, it seems highly unlikely given that Micceri (1989) who examined 440 variables from pubmished articles in educaiton and psychology, found that 84% showed moderate to extreme skew"*. En ce qui concerne l'homogénéité des variances, seulement 3 mentionnaient l'hypothèse d'homogénéité des variances, alors que parmis les 65 articles donnant l'info sur les variances de groupe,

27 avaient un SDR de plus de 2 (dans une étude, le SDR était même de 104), et que la plupart d'entre elles avaient des designs non balancés.

- @counsell_reporting_2017: sur 151 études analysées, 44 reportaient des informations relatives aux conditions d'application des tests utilisés. Parmi celles ci, seulement 2 étaient exhaustifs: toutes les autres soit reportaient seulement une partie des conditions, soit le faisaient de manière inappropriée (ex.: vérifier la normalité de la VD au lieu de celle des résidus, dans le cas d'une régression).

A cause du manque de transparence des chercheurs quant au respect ou non respect des conditions d'application, il est très compliqué de vérifier si le test qu'ils ont choisi est approprié ou non [counsell_reporting_2017].CA JE PEUX COMPLETER AVEC MON PROJET DE RECHERCHE: *le non report peut vouloir dire que la condition était ok mais ça semble très peu crédible qu'elle le soit dans autant d'études.*

–> POURQUOI ELLES NE SONT PAS VERIFIEES? -> 1) Leur non report peut être dû à un désir de limiter le nombre de pages (on se contente de reporter ce que les reviewers/éditeurs nous demandent de reporter; @counsell_reporting_2017) 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).

Ce constat est loin d'être récent. Par exemple, @keselman_statistical_1998 mettaient en évidence le fait que les chercheurs tendaient à utiliser des tests non robustes aux violations des conditions d'application, sans vérifier au préalable si ces conditions étaient respectées.

Même lorsqu'un chercheur souhaite vérifier les conditions d'application, il reste confronté à plusieurs problèmes.

1. Les conditions reposent sur les paramètres de *population* et non sur les paramètres d'échantillon. Or, ces paramètres de population ne sont pas connus (s'ils l'étaient, on n'aurait pas besoin des statistiques) [@hoekstra_are_2012].
2. Les conditions sont souvent très irréalistes.

Il existe des tests dit "tests robustes" qui ne sont théoriquement pas affectés par une violation des conditions d'application, mais ces derniers ne sont que peu ou pas utilisés [@sharpe_why_2013]

Comment améliorer les pratiques de recherche facilement, d'une manière qui assure que les chercheurs appliqueront les conseils? 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

ARTICLE1 ARTICLE2.

Limite 2: hypothèse nulle

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.