# **Anticipation de questions pour ma thèse**

## Tailles d’effets non standardisées : réflexion (résumé de l’article de Pek & Flora)

1. Il y a des situations où les tailles d’effet non standardisées sont très pertinentes. Par exemple, quand on crée un modèle, on va souvent utiliser les paramètres du modèle estimé comme mesure de taille d’effet (ex. : la pente non standardisée d’une droite de régression).

***Ex****. : le score d’IAT est souvent mesuré comme suit :*

* *On calcule le temps de latence moyen des items « cohérents » pour un raciste (blancs associés aux mots positifs et noirs associés aux mots négatifs), idem pour les items « incohérents » (blancs négatifs et noirs positifs).*
* *On calcule le log de ces deux moyennes, après winsorization (= remplacement des scores extrêmes par d’autres si j’ai bien compris)*
* *On calcule la différence des deux logs (= log(temps moyen de latence items* ***in****compatibles) – log (temps moyen de latence des items compatibles) 🡪 plus c positif, plus ça veut dire qu’ils buggent sur les items incompatibles et donc a priori ont du mal avec les noirs.*
* *On pourrait s’intéresser à la relation entre le score et l’IAT et le score à une échelle de racisme, et se dire par exemple que’ «quand le score à l’IAT augmente de 10 ms, on peut prédire une augmentation du score sur l’échelle de racisme de 0.3 points ». La taille d’effet ici est une pente de droite de régression = mesure non standardisée.*

1. On prétend souvent que les mesures standardisées sont « mieux » parce qu’on utilise en psycho des mesures « arbitraires ». Mais même ces mesures arbitraires peuvent avoir un intérêt, en termes de score brut, à condition que des experts aient trouvé un consensus.

*Ex. : Imaginons qu’on calcule la différence de scores de lecture entre les filles et les garçons adolescents en utilisant les scores de lecture de PISA (Programme for International Student Assessment reading scores). Avec ce programme, on définit 5 niveau de lecture (environ 70 points entre les niveaux): Niveau 1 (335-407), Niveau 2 (408-480), Niveau 3 (481-552), Niveau 4 (553-625), Niveau 5 (> 625). Imaginons qu’on trouve une différence de score de 60 points, ça veut dire qu’il y a une différence de ¾ de niveau. ce score de 60 peut être directement comparé aux guidelines prévue pour cette échelle.*

*Lire exemple de l’alcool « C:\Users\mdelacre\Dropbox\Préparation défense thèse\tailles d'effet brutes\ex alcohol.pdf »*

Au contraire, si on n’a pas de guidelines dans un contexte donné et pour un certain groupe cible pour les mesures standardisées, elles sont difficiles à interpréter.

*Pour reprendre notre exemple, déterminer l’écart en termes de « niveaux » sur les score brut est plus parlant que de dire, par exemple, qu’on trouve un d de Cohen de 1.2, qu’on aura plutôt tendance à interpréter en utilisant les balises arbitraires de Cohen peu adaptées au Cohen, parce qu’on ne dispose pas encore de guidelines précises dans ce contexte* (Gignac & Szodorai, 2016, cités par Pek et Flora).

Autre exemple : dire qu’en soirée, les hommes boivent en moyenne 2 bières de plus que les femmes parle plus que de dire que le petit d de Cohen vaut 0.2.

1. Même quand le but est in fine de calculer une mesure standardisée précise, il est souvent mieux de disposer des descriptives dans un article, pour pouvoir calculer LA mesure qui nous intéresse. Du coup, dans certains domaines, on conseille de privilégier les mesures non standardisées (tel que la différence de moyenne) :

* Dans le contexte des méta-analyses, on conseille souvent de se fier plutôt aux descriptives pour ensuite calculer la mesure standardisée qui correspond à ce dont on a besoin plutôt que de se fier directement à une mesure standardisée proposée (ex. : d de Cohen) ;
* Dans le contexte des analyses de puissance, il est recommandé d’utiliser des mesures non standardisées de manière à incorporer toutes les informations disponibles et ainsi calculer plus précisément la puissance en prenant directement en compte les particularités de l'étude conçue.

Il faut garder à l’esprit qu’une mesure de taille d’effet doit être simple et compréhensible pour le lecteur, et qu’elle doit pouvoir être interprétée **en lien avec la question de recherche** (c’est une question plus importante à se poser que de savoir si on opte pour une mesure standardisée ou non). *Ex. : Si notre but est de savoir si une thérapie améliore le bien-être, le R² ne sera pas approprié (mais ne donne aucune info concrète sur le sens de l’effet ni sur l’amplitude en termes de score). Si au contraire notre but est de déterminer si un effet est fortement soutenu par les données ou pas, alors là le R² sera la mesure adéquate.*

*Ex.2 : si notre but est de voir si le score de lecture est stable à travers le temps chez les adolescents, le coefficient de corrélation intra classe (mesure standardisée) est plus top que la covariance (car balisée et donc plus facile à interpréter).*

1. Même si on se limite au *d* de Cohen, il existe plusieurs manières différentes de calculer le standardiser, et en fonction de celui qui est choisi, on peut avoir des résultats très différents. Les mesures non standardisées quand elle font sens sont moins simples et plus familières (ex. : nombre de cigarettes fumées, temps de latence, etc.)

## Pourquoi ne pas opter pour des transformations plutôt que des tests robustes?

D’après Erceg-Hurn et Mirosevich (2012), elles sont problématiques pour plusieurs raisons :

1. Les transformations échouent souvent à restaurer les conditions (on parle souvent de la transformation « log » pour restaurer la normale, mais ça ne marche pas toujours).
2. Elles ne permettent pas de traiter le problème des outliers
3. Elles peuvent faire perdre en puissance
4. Parfois, ça va modifier l’ordre des moyennes
5. Complique l’interprétation des résultats (on interprète des mesures transformées, plus les mesures d’ origine).

## Aide à l’interprétation d’une mesure standardisée

Interprétation objective quand il est possible d’établir des liens entre une différence de moyenne sur une échelle donnée et des « symptômes » concrets:

* Ex. : soit une échelle de Likert à 5 points (allant de 1 = jamais à 5 = très souvent) : on pourrait établir des liens entre ce score et des symptômes du style « *le nombre de cigarettes fumées par jour par des individus qui ont un score de 1.2* »).

## Bootstrapping : comprendre comment l’utiliser pour comparer les moyennes

Voir aussi « randomization test » (livre de Edgington, 1980)

## Quid du bayésien ? essayer de trouver une réponse-type au cas où (cf . article de Daniel éventuellement ?)

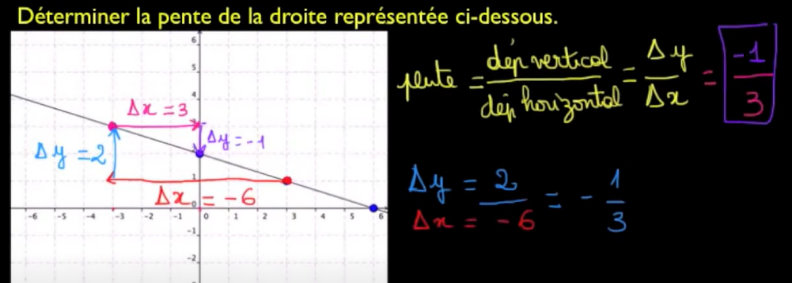
## Lien test de Welch et régression

[Common statistical tests are linear models (or: how to teach stats) (lindeloev.github.io)](https://lindeloev.github.io/tests-as-linear/#52_welch%E2%80%99s_t-test)

**Script R :** Welch\_modelelineaire.R

## L’intercept et la pente de régression quand on compare la moyenne de deux groupes :

De manière générale, voici comment on calcule la pente d’une régression :



Dans le contexte de la comparaison de deux moyennes… peu importe qu’on considère qu’il y a homoscédasticité ou pas, ça ne change rien à la définition des coefficients (ce qui va changer, c’est le test t réalisé sur le b1, cf. point 5.2).

n1 <- 10

n2 <- 12

Gr1 <- rnorm(n1)

Gr2 <- rnorm(n2)

y <- c(Gr1,Gr2)

x <- c(rep(1,n1),rep(2,n2))

#Représentation graphique des points + descriptives

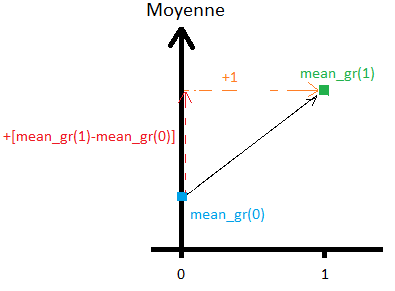
plot(y~x)

means <- tapply(y, x, mean) # Moyenne par groupe

m0 <- means[1]

m1 <- means[2]

… si j’utilise des dummy codes (0 ; 1)  :



Sachant que y = b0+b1x

* **b0 = ordonnée à l’origine** = valeur prédite pour y quand x = 0 (= moyenne du groupe codé 0)
* **b1 = pente** = (cf. schéma ci-dessus) = =

Intuitivement, la meilleure prédiction qu’on peut faire pour un point appartenant à un groupe donné correspond à la moyenne de ce groupe.

Sachant que

* Si x = 0 : = m0
* Si x = 1 : = m1 🡨🡪 = m1 - = m1-m0

# Utiliser le modèle linaire

# (a) avec le dummy coding

x <- c(rep(0,n1),rep(1,n2))

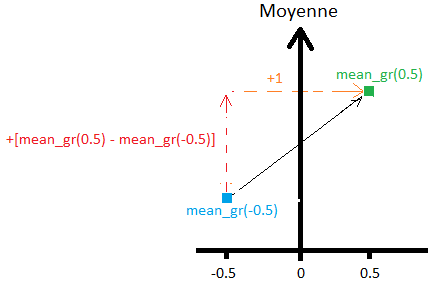
regr<-summary(lm(y~x))

round(regr$coefficients[1],5)==round(m0,5) # y = b0+b1x --> quand x = 0, y = b0

# b0 = moyenne du groupe codé 0

round(regr$coefficient[2],5)==round(m1-m0,5) # b1 = pente = delta\_y/delta\_x = m1-m0

… si j’utilise des codes contrastes (-0.5 ; 0.5) :



Sachant que y = b0+b1x

* **b0 = ordonnée à l’origine** = valeur prédite pour y quand x = 0 (= moyenne non pondérée des deux groupes, puisque l’intercept est exactement à égal distance des deux moyennes)
* **b1 = pente** = (cf. schéma ci-dessus) =

=

Intuitivement, puisque ni (= la différence de moyenne) ni (l’écart entre les 2 codes donnés) n’a changé, la pente est identique à celle obtenue avec le dummy coding. On n’a fait que « translater » la pente d’un demi-point vers la gauche. Donc seule l’ordonnée à l’origine () change.

# (b) Avec des prédicteurs centrés

x <- c(rep(-.5,n1),rep(.5,n2))

regr<-summary(lm(y~x))

round(regr$coefficients[1],5)==round((m0+m1)/2,5) # b0 est à mi-chemin entre les 2 moyennes,

# et correspond donc à la moyenne non pondérée

# des deux moyennes

round(regr$coefficient[2],5)==round(m1-m0,5) # b1 = pente = delta\_y/delta\_x = (m1-m0)/1 = (m1-m0)

* **CCL** : Puisque l’écart entre les deux valeurs sur l’axe des x est identique dans les deux cas, la pente sera la même dans les deux situations, mais par contre la valeur de l’intercept change 😊

## Welch : un cas particulier où l’on n’assume pas les variances identiques

Comme dit précédemment, si je réalise une régression n’impliquant pas qu’il y ait égalité des variances, je vais obtenir exactement les mêmes valeurs pour b0 et b1. Par contre, le test de significativité sur les coefficients va changer.

**Dans R**, en passant par le menu « gls », on a la possibilité de spécifier qu’on veut considérer dans notre régression un terme d’erreur qui n’implique pas une variance commune. De la sorte, on peut obtenir exactement la même valeur de statistique qu’en passant par le test t de Welch.

Cependant, le test de Welch se base sur l’approximation Satterwaithe pour estimer les degrés de liberté, et ce n’est pas le cas du package gls. Pour cette raison, on obtient des degrés de liberté différents avec les deux méthodes (et donc forcément des p-valeurs aussi différentes).

**Dans Jamovi**, en téléchargeant le package gamlj, on peut faire des modèles linéaires généralisés, mais rien d’équivalent au test de Welch. Via SPSS par contre, il parait que c’est possible.

# Exemple concret d’étude où il y a un souci d’hétéroscédasticité et que cela amène à mal estimer les choses

**Essayer de trouver un exemple d’une étude. Eventuellement calculer le d de Cohen (le classique) obtenu sur cette étude et le comparer à la valeur qu’on aurait obtenu avec le d non poolé… pour montrer que ça peut enraîner de grandes différences. Jf,**

## Exemple concret d’étude où on tente de montrer une absence de différence sur base d’un non-rejet de l’hypothèse nulle

**Résumé de l’article de Allison et al. (1993)[[1]](#footnote-1)** : ils ont étudié des produits alimentaires trouvés dans des commerces à Manhattan qui se déclaraient « produits sain » ou « produits de régime » (= produits contenant les mots « lite » « low-calorie », « reduced-calorie », « diet », »low-fat », « no-fat », ou « health » dans le descriptif), parmi lesquels :

* 12 sont distribués régionalement (n1=12)
* 20 sont commercialisés nationalement (n2=20)
* et 8 sont préparés localement (n3=8)

La VD de l’étude est l’écart entre les calories annoncées et le vrai nombre de calories, exprimé en % :

Ils ont déterminé le nombre de vraies calories en utilisant le calorimètre de Miller & Payne (1959), et masquaient le nombre de calories annoncées aux évaluateurs pour éviter de les influencer. Et ils faisaient le test 2 fois par aliment, et s’ils obtenaient un écart supérieur à 0.03kcal entre deux essais, ils recommençaient une 3ème fois et se fiaient aux 2 mesures les plus proches. Ils ont obtenu une corrélation test-retest de .9992 (donc super fiable apparemment).

La FDA (Food and Drug Administration) autorise une erreur de 20% mais en pratique, on constate que l’erreur dépasse parfois largement ces 20%.

Voici les descriptives trouvées[[2]](#footnote-2) :

|  |  |  |
| --- | --- | --- |
| Région : | Écart par « produit » : Moyenne(SD) | Écart par gramme : moyenne(SD) |
| Local | 85.42%(77.88%) | Valeur non donnée |
| Régional | 25.22%(15.58%) | 14.97%(17.95%) |
| National | 2.18(13.93%) | 0.01%(9.13%) |

Dans l’exemple de cette étude, à deux reprises, ils font l’erreur d’interpréter un NRH0 comme une AH0 (et en plus, ils le font sur base de tests t, et ce malgré des violations flagrantes de plusieurs conditions comme ils le signalent eux même) :

1. Ils utilisent plusieurs moyens de vérifier la fiabilité de la mesure « réelle » de calories. Parmi ces méthodes on retrouve ceci : en général, ils faisaient le test 2 fois par produit, et ils ont fait un test *t* pour échantillons dépendants (en bilatéral) : «**the first and second calorimetric measurements for the duplicate sample were tested for statistically significant mean differences by using a dependent-samples two-tailed t test. The mean caloric measurement was 944 (SD=731.9) for trial 1 and 942.7 (SD=730.3) for trial 2 (t=0.31; df=39;p=.76)**». Heureusement, ce n’était pas la seule méthode sur laquelle ils se fiaient pour montrer la fiabilité de la méthode (ils ont aussi calculé une corrélation très très proche de 1 entre les 2 essais).
2. Ils expliquent que pour les produits nationaux, il n’y avait pas de différence significative entre les calories annoncées et les calories réelles (en posant comme hypothèse nulle une absence de différence entre les deux) et concluent que « **nationally advertised food items were quite accurate** ». 🡪 Encore la même erreur ! C’est d’autant plus flashant qu’avec seulement 20 produits, la puissance du test est certainement très faible…

## Exemple méthodologique de contexte où l’on tente de montrer une absence de différence sur base d’un non-rejet de l’hypothèse nulle

Bien sûr, il y a la traditionnelle vérification des conditions d’application.

Mais aussi, en méta-analyses quand on tente d’étudier un effet donné, on tente parfois de voir s’il existe des variations dans les mesures analysées qui dépassent l’erreur de mesure, en calculant (voir le point sur les tailles d’effet hétérogène de Judd pour comprendre ce concept). Pour ce faire, on utilise le test chi-carré d’homogénéité (symbolisé Q de Cochran). 🡪 Lire article de Cochran (1954) que j’ai mis dans la section sur l’hétérogénéité des tailles d’effet pour voir s’ils l’interprètent parfois en AH0 (mais vu son nom, je dirais que oui 😊 ).

## Exemple concret d’étude où une petite taille d’effet a un intérêt pratique

**Ex. de Pek et Flora** : Imaginons qu’une thérapie contre l’anxiété permette à des coureurs de réduire leur temps de course de quelques millisecondes. Cette toute petite différence pourrait les faire monter sur le podium face à des compétiteurs de même niveau (donc grand intérêt pratique).

## Relire les taille d’effet hétérogène de Judd (pour Yzerbit)

Quand on réplique plusieurs études sur le même effet, on observe souvent des variations dans les estimations de taille d’effet qui dépassent la simple erreur d’échantillonnage (ce sont plus que des variations aléatoires ; en moyenne, l’hétérogénéité serait trois fois plus grande que celle due à l’erreur d’échantillonnage, d’après Stanley et al. 2018) 🡪 on conclut aux tailles d’effet hétérogènes à travers les études (même dans les réplications directes).

Cela a pour conséquence d’augmenter le taux d’erreur de type I relatif à l’hypothèse d’après laquelle la moyenne de la population de taille d’effet diffère de 0.

**Solution** : estimer le VRAI écart-type de la mesure de taille d’effet (ex. : = l’écart-type de la mesure Cohen’s d APRES avoir supprimé l’effet de l’erreur d’échantillonnage[[3]](#footnote-3) ; par opposition à qui est l’erreur standard)[[4]](#footnote-4).

A quoi serait dûe l’hétérogénéité ?

* À des modérateurs cachés (variations dans les expérimentateurs, participants, histoire, place, etc.)
* Des variations aléatoires (en fait il est impossible d’anticiper tous les modérateurs et donc il restera tj de l’hétérogénéité)

Vincent proposait d’envisager d’étudier la distribution des tailles d’effet plutôt que la tendance centrale via le TOST. Lire article de Hedges & Vevea (1998) pour comprendre un peu l’idée si j’ai le temps.

## Ok on observe des SDR différents de 1 dans la littérature, mais est-ce que les estimations de SDR faites sur base de la littérature s’écartent suffisamment de 1 :1 pour dire que ce n’est pas dû à l’erreur d’échantillonnage uniquement ?

Lire article de Ruscio & Roche (2011) : ils ont analysé un lot de 455 articles (en n’étudiant que la première étude à chaque fois), publiés dans 9 journaux de l’APA (incluant des journaux de neuropsycho, de psychologie de la santé, de psycho expérimentale, de psycho clinique, de psycho sociale, etc.). Pour chaque étude ils enregistraient entre autres le nombre de groupes, le nombre de sujets, les moyennes et écart-types de chaque groupe, ainsi que si l’étude était expérimentale (= attribution aléatoire au sein des groupes) ou corrélationnelle.

Pour chaque étude, ils ont mesuré l’hétéroscédasticité de trois manières distinctes :

1. Le ratio de variances
2. Le CVV = on calcule la somme d’écart de la variance de chaque groupe par rapport à la variance poolée, somme pondérée en fonction de la taille des groupes.

Ex. pour deux groupes :

* Du coup, si le plus petit groupe est celui qui est associé à la variance la plus proche de la variance poolée, le CVV diminue (au contraire, si le plus petit groupe est associé à la variance la plus loin de la variance poolée, le CVV augmente).
* Plus il y a de groupes, plus le CVV tend à grandir (car le numérateur sous la racine carrée grandit plus vite que le dénominateur).

1. SVH = mesure créée pour « annuler l’effet du nombre de groupes ». Le SVH peut varier de 0 (= homogénéité des variances) à 1 (quand la variance de tous les groupes sauf 1 = 0, signifiant que le ratio max/min = l’infini).

Box (1954) prétendait que le ratio de variances dépassait rarement 3 (or, un ratio de 3 correspond à un SDR de 1.73). Ruschio & Roche ont fait deux choses :

1. Ils ont déterminé, pour des designs ayant des tailles d’échantillon et un nombre de groupes similaires à ceux de leur investigation, dans quel pourcentage de cas le ratio de variances serait supérieur à 3 **si les échantillons étaient extraits de populations normales aux variances égales.**
2. Ils ont déterminé la proportion d’études dont le ratio de variances était supérieur à 3 dans leur investigation (pour voir si ce taux dépassait celui de l’erreur d’échantillonnage).

Ils ont trouvé que le ratio empirique était bien supérieur au ratio théorique (il dépassait 3 dans presque un quart des cas, alors que c’est bcp plus rare quand il y a homoscédasticité, d’après leurs estimations théoriques).

**Attention** : relativisons un peu leur constat :

* ils ont estimé le taux de VR>3, qu’on devrait estimer théoriquement, en faisant comme si les échantillons étaient extraits de populations normales. Or, on sait que c’est irréaliste en réalité (ils le disent eux même d’ailleurs). J’ai constaté qu’en considérant les valeurs de G1 et G2 que j’ai envisagé dans l’article sur les tailles d’effets que la proportion de VR supérieurs à 3 était bien plus élevée avec des distributions fortement asymétriques et heavy tailed.
* De même, si l’on compare la VR médiane et la VR maximum qu’ils trouvent empiriquement, les deux valeurs sont supérieures à celles qu’on aurait théoriquement si les données étaient extraites de populations normales aux variances identiques, mais pas forcément supérieures à ce qu’on aurait si les distributions étaient extraites de populations similaires à celles que j’ai envisagé dans mon article sur les tailles d’effet.

🡪 Tout ça me fait penser que leur étude est une très bonne idée et une très bonne piste pour estimer si vraiment les VR observées sont problématiques en réalité, mais qu’il faudrait envisager d’autres distributions théoriques que les distributions normales pour bien faire.

Attention : ce n’est pas si simple que ça parce que s’il est difficile de comparer les VR empiriques et théoriques, il est tout aussi difficile d’estimer des distributions de population réalistes (la distribution d’échantillonnage de G1 et G2 est aussi asymétrique, et donc les valeurs considérées dans mon article sont probablement aussi une large surestimation de la réalité !)

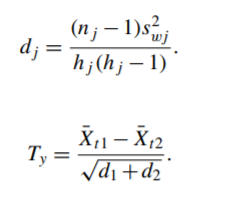
## A revoir dans la thèse avant la défense (pour être capable de le ré-expliquer)

### Test de Yuen : l’étudier et revoir pourquoi il est moins puissant que le test t classique quand les conditions sont respectées (cf. article de Wilcox, je crois que c’est « how many discoveries… »)

Le test de Yuen consiste à calculer des moyennes trimmées et la variance winsorizée.

* Moyenne trimmée = on enlève 20% des scores les plus élevés et les plus faibles (dans chaque groupe ? à vérifier)
* Winsorized variance = variance calculée sur la série dans laquelle, les scores « trimmés » sont remplacés par les scores les plus extrêmes non trimmés.

Ensuite, on calcule une statistique donnée dans les notes dans ma farde ainsi que dans mon script « Yuen calculations », qui revient identiquement à la statistique de Welch quand trimming = 0%.



**Remarque** : pas tout à fait lié, mais d’après Erceg-Hurn et MIrosevich, cette solution est préférable aux tests non paramétriques car peut être généralisée pour des designs puls complexes (designs factoriels par exemple).

**Reste à comprendre pourquoi il est moins bon que le test classique quand les conditions sont OK :**

* **Distribution d’échantillonnage de la moyenne vs. De la moyenne trimmée ?**
* **Comprendre pourquoi le test devient parfois trop libéral (cf. Supp Mat de mon article 1).**

### Explication rapide de l’IC autour des tailles d’effet standardisées (lien avec le paramètre de non-centralité, bref l’annexe de la thèse à ce sujet).

### Micceri

Le relire, dans la mesure où c’est l’article de référence pour dire que les distributions normales sont irréalistes. Il a utilisé des échantillons d’environ 400 personnes. Relire un peu sa méthodologie, et faire 2/3 simulations pour voir si de tels échantillons sont suffisants pour clamer que la normalité n’est pas ok (ex. : voir la distribution d’échantillonnage et G1 et G2 avec 400 sujets par échantillon).

### Chapitre sur le test d’équivalence : comment j’ai créé la figure 28 (p.96, partie relative à l’analyse de puissance)

**Approche de la puissance, explication du calcul pour la figure** :

1. Pourquoi on ne conclura JAMAIS à l’équivalence quand la SE dépasse 7.07 ?

**Réponse** : car il y a une contrainte de puissance ! Si on n’atteint pas la puissance de 80% pour détecter un effet aussi grand que le SESOI (qui correspond aux bornes de la zone d’équivalence), on ne peut conclure à l’équivalence.

* 1. Avec des analyses de puissance (sensitivity analyses) on est capable de déterminer la plus petite taille d’effet standardisée (mesure *d*) que l’on pourrait détecter avec une puissance de 80 % ( = dmin), compte tenu du n et du alpha.

*Dans l’exemple de la thèse, avec n = 50 par groupe et alpha = 5%, ce d vaut 0.5658… (illustra via R et gpower). Ça veut dire concrètement que .5658 correspond aux plus petites bornes standardisées qu’on peut fixer tout en assurant la puissance suffisante pour pouvoir conclure à l’équivalence.*

Une image contenant table

Description générée automatiquement

Une image contenant texte

Description générée automatiquementUne image contenant texte

Description générée automatiquement

* 1. Dans l’énoncé, on n’a pas fixé les bornes de la zone d’équivalence en unité standardisées, mais en unités brutes (en différences de moyenne). On a choisi de fixer ces bornes à theta = 20. Heureusement, on connait le lien entre le *d* standardisé et la différence brute de moyenne :

Dans notre exemple, . Si l’on connait la plus petite valeur que peut prendre *d* (on l’a déterminé au point précédent), on pourra en déduire la plus grande valeur que pourra prendre S (S reste en effet la seule inconnue de l’équation).

Remarque : Forcément, le plus petit *d* autorisé sera celui qui est associé au plus grand *S* autorisé (puisque theta est constant, plus S sera grand, plus d sera petit) :

🡨🡪

*Dans l’exemple de la thèse, 🡨🡪 35.3446*

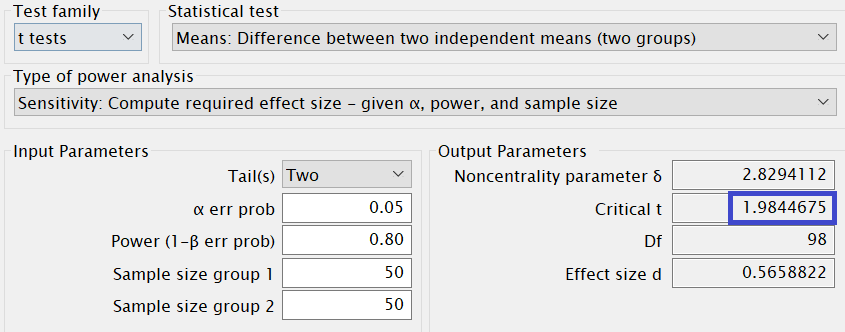
* 1. On connait le lien entre S et SE (quand n1 = n2 et S1=S2). Du coup, lorsqu’on a déterminé Smax, il est très facile d’en déduire le plus grand SE autorisé pour atteindre la puissance de 80%.

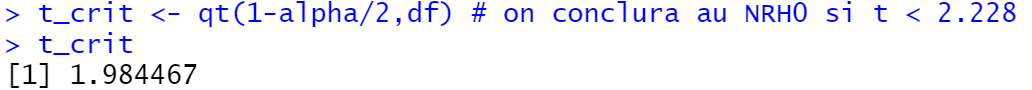
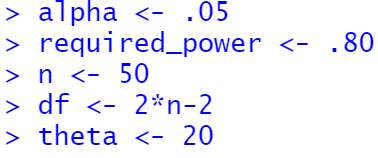
🡨 🡪

*Dans l’exemple de la thèse,*

1. Déterminer les valeurs de SE suffisamment grande pour conclure au non rejet de H0.
   1. Pour conclure à l’équivalence, il faut conclure au NRH0. On peut facilement, via R ou des tables, déterminer la t critique ( = la valeur de la statistique t en deçà de laquelle on ne conclura PAS au rejet de l’hypothèse nulle).

*Dans l’exemple de la thèse, avec n = 50 par groupe et alpha = 5%, ce t vaut 1.9845… (illustra via R et gpower). Ça veut dire que chaque fois qu’on trouvera une t inférieure à cette valeur, on conclure au NRH0*





* 1. Pour rappel, une statistique t est une différence de moyenne divisée par une SE :

🡨🡪

Donc une fois qu’on a tcrit, on peut déterminer SEcrit pour n’importe quelle différence de moyenne.

**En conclusion, pour une différence de moyennes donnée, chaque fois que , on conclura à l’équivalence avec l’approche de la puissance.**

Si on analyse , on comprend que pour une différence de moyennes donnée, plus SE est petite (= estimation précise), plus la statistique t sera grande. Donc avec un petit SE, il faudra une plus petite différence de moyennes absolue pour assurer que le *t* ne dépasse pas le seuil critique (et donc plus les résultats sont précis, plus difficilement on conclura à l’équivalence).

A revoir aussi pour le TOST

1. Allison, D. B., Heshka, S., Sepulveda, D., & Heymsfield, S. B. (1993). Counting calories–Caveat emptor. The Journal of the American Medical Association, 270, 1454–1456. http://dx.doi.org/10.1001/jama.1993 .03510120076034 [↑](#footnote-ref-1)
2. Etonnamment, même s’ils disent avoir privilégié des tests non paramétriques (= randomization tests) compte tenu des violations d’homogénéité des variances, d’indépendance des observations et de normalité de leur données, ils reportent quand même la moyenne et l’écart-type comme descriptives pour chaque groupe. Ils ne fournissent pas de mesure de taille d’effet dans leur article, mais si on voulait en calculer sur base des résultats fournis dans l’article, on ne pourrait partir que sur des mesures de comparaison de moyenne. [↑](#footnote-ref-2)
3. Dans la littérature sur les méta-analyses, la variance de cette mesure est souvent appelée . [↑](#footnote-ref-3)
4. Attention : forte corrélation entre la taille d’effet moyenne et l’écart-type de la taille d’effet, si bien qu’on trouve généralement que les effets sont très peu hétérogènes quand les tailles d’effets sont proches de 0. [↑](#footnote-ref-4)