Chapitre 5: les tests d'équivalence

Lorsqu'on applique un test d'hypothèse, l'hypothèse nulle la plus couramment définie est celle d'absence d'effet ou de différence entre les groupes (**nickerson_null_2000?**). Il arrive également parfois que les chercheurs définissent un intervalle de valeur comme hypothèse nulle, mais le plus souvent, cet intervalle est borné par la valeur 0 (**nickerson_null_2000?**), on parle alors d'hypothèse unilatérale. Avec cette stratégie, le rejet de l'hypothèse nulle constitue un soutien en faveur de la présence d'un effet non nul, par contre, le non rejet de l'hypothèse nulle ne peut être interprété comme un soutien en faveur de l'absence d'effet. Pourtant, il arrive souvent que des chercheurs l'interprètent de la sorte (**anderson_theres_2016?**). (**finch_reporting_2001?**), par exemple, ont reporté que parmis 150 articles publiés entre 1940 et 1999 dans le *JAP* (*Journal of Applied Psychology*), 38% interprétaient un résultat non significatif comme une acceptation de l'hypothès nulle. Plus récemment, (**lakens_equivalence_2017?**) a noté que l'expression "pas d'effet" a été utilisée dans 108 articles publié dans *Social Psychological and Personality Science* avant août 2016 et que dans presque tous les cas, c'était sur base du non rejet de l'hypothèse nulle que cette conclusion était tirée. Cette erreur d'interprétation est également fréquemment commise dans le cadre des études de réplication. (**anderson_theres_2016?**), par exemple, ont analysé 50 réplications d'études publiées en 2013 dans PsycINFO. Ils ont noté que 14 études affirmaient avoir obtenu des effets "nuls" (interprété comme un échec à la réplication), et tous l'ont fait sur base de l'acceptation d'une hypothèse nulle d'absence d'effet. C'est par exemple de cette manière qu'on été réalisées la plupart des tentatives de réplications de la célèbre étude de Bem (Ritchie, Wiseman & French, 2012, cités par **anderson_theres_2016?**).

A travers ce chapitre, notre premier objectif sera d'expliquer pourquoi interpréter le non rejet de l'hypothèse d'absence d'effet comme un soutien en faveur d'une absence d'effet n'est pas une bonne stratégie. Nous introduirons ensuite les tests d'équivalence qui

permettent d'obtenir un soutien en faveur d'un effet jugé non pertinent, et plus particulièrement le TOST (Two One-sided test). Nous verrons que l'aspect le plus compliqué de la réalisation du TOST est la définition des bornes d'équivalence. Pour cette raison, notre troisième objectif sera de fournir quelques pistes en vue de définir ces bornes. Pour finir, nous présenterons un article dans lequel nous comparons le TOST à la SGPV (Second Generation *P*-Value), une stratégie récemment développée par (**blume_second-generation_2018?**).

Limites de l'approche traditionnelle

Lorsqu'on teste une hypothèse nulle, il y a deux conclusions possibles: soit ont la rejette, soit on ne la rejette pas. Si rejeter l'hypothèse nulle amène à conclure en faveur de l'hypothèse alternative, ne pas la rejeter ne permet pas de conclure en faveur de l'hypothèse nulle. Au mieux, cela nous montre que les données ne sont pas incompatibles avec l'hypothèse nulle, mais cela ne veut en aucun cas dire qu'elles ne sont compatibles avec aucune autre hypothèse. Afin de l'illustrer, la Table 1 résume les résultats de simulations Monte Carlo pour un ensemble de 42 scénarios qui varient en fonction de la taille des échantillons (n_j) et de la différence entre les moyennes des deux populations dont sont extraits les échantillons ($\mu_1 - \mu_2$). Pour chaque scénario, à 100,000 reprises, nous avons généré aléatoirement une paire d'échantillons indépendants, réalisé un test t de Student pour échantillons indépendants et extrait la p-valeur du test. Ensuite, nous avons calculé la proportion d'itérations associées à une p-valeurs supérieures à .05, nous amenant à ne pas rejeter l'hypothèse nulle lorsqu'on travaille avec un risque alpha de 5% (ce risque alpha étant communément accepté par la majorité des chercheurs, **meyners_equivalence_2012?**). Lorsque l'hypothèse nulle est fausse (toutes les colonnes de la Table 1, à l'exception de la première), cette proportion correspond au taux d'erreur de type II (communément appelé β).

Table 1.

Proportion de p-valeurs supérieures à .05 en fonction de la taille des échantillons (n_j) et de la différence entre les moyennes de chaque population ($\mu_1 - \mu_2$).

	Différence de moyennes dans la population ($\mu_1 - \mu_2$)					
n_j	0	.1	.2	.3	.4	.5
100	0.949	0.891	0.709	0.441	0.196	0.060
200	0.949	0.832	0.483	0.151	0.021	0.001
300	0.950	0.768	0.312	0.043	0.001	< .001
400	0.949	0.708	0.194	0.011	< .001	< .001
500	0.949	0.646	0.116	0.003	< .001	< .001
600	0.950	0.591	0.066	0.001	< .001	< .001
700	0.948	0.536	0.037	< .001	< .001	< .001

Note. Pour chaque scénario, les deux échantillons sont toujours de même taille ($n_1 = n_2 = n$) et sont extraits de populations se distribuant normalement et ayant la même variance ($\sigma_1 = \sigma_2 = \sigma$). La moyenne de la première population (μ_1) vaut systématiquement 0, et celle de la deuxième population (μ_2) varie de sorte à obtenir la différence de moyenne $\mu_1 - \mu_2$ désirée. Par ailleurs, σ vaut systématiquement 1, si bien que la différence de moyenne brute est égale au δ de Cohen.

Pour les scénarios de la première colonne, l'hypothèse nulle est vraie: il n'y a pas de différence entre les moyennes de population. Puisque les conditions d'application du test t de Student sont toutes rencontrées, on est amené à rejeter l'hypothèse nulle, c'est-à-dire à commettre une erreur de type I dans une proportion d'itération égale à $\alpha = 5\%$. Par conséquent, on est amené à *ne pas* rejeter l'hypothèse nulle 95% du temps (ce qui correspond à $(1 - \alpha)\%$)[1]. Pour les scénarios envisagés dans toutes les autres colonnes de la Table 1, une vraie différence entre les moyennes de population existe, si bien que le rejet de l'hypothèse nulle est la bonne décision. Pourtant, pour plusieurs scénarios, le nombre d'itérations amenant à conclure au non rejet de l'hypothèse nulle est bien supérieur au nombre d'itérations amenant à conclure au rejet de l'hypothèse nulle, comme on peut le voir à travers les valeurs β. Par exemple, avec 100 sujets par groupes et considérant $\sigma_1 = \sigma_2 = 1$, on ne détectera pas une différence de moyenne de .1 dans près de 90% des cas. Avec 700 sujets par groupe, cette différence ne sera toujours pas détectée plus d'une fois sur deux ($\approx 54\%$ des itérations). En présence d'un effet non nul, cela se justifie par un manque de puissance des tests réalisés, ce qui démontre bien qu'un non rejet de l'hypothèse nulle peut en fait signifier deux choses: soit qu'il n'y a vraiment pas de différence entre les moyennes des populations (ou autrement dit, que les différences observées sont dûes au hasard), soit que le test n'est pas suffisamment puissant pour détecter la différence. Or, le manque de puissance des tests est récurrent dans la littérature, comme tendent à le montrer diverses méta-analyses (**button_power_2013?**; **bakker_rules_2012?**; **funder_improving_2014?**).

Pour éviter d'interpréter un test peu puissant comme un soutien en faveur de l'hypothèse nulle, l'approche de la puissance est devenue l'approche par défaut dans les années 80 pour tester l'équivalence (**meyners_equivalence_2012?**). A travers cette approche qui est

[1] Nous observons en réalité des proportions qui varient de 94.8% à 95%, à cause du hasard d'échantillonnage, mais sur le long terme, lorsque le nombre d'itérations tend vers l'infini, toutes les proportions ne non rejet de l'hypothèse nulle quand l'hypothèse nulle est vraie vont tendre vers $(1 - \alpha)\%$.

restée très populaire (**quertemont_how_2011?**), dans un premier temps, on définit ce qu'on considère comme étant la plus petite valeur d'intérêt (en anglais, le "SESOI" pour "Smaller Effect Size of Interest"), c'est-à-dire la taille d'effet minimale requise pour considérer qu'un effet est pertinent. Ensuite, on estime la puissance de notre test à détecter un effet de cette taille[2], et si cette estimation atteind une valeur jugée satisfaisante (en général, 80%), alors on considère que l'on peut interpréter le non rejet de l'hypothèse nulle d'absence d'effet comme soutien en faveur de l'équivalence (**quertemont_how_2011?**; **meyners_equivalence_2012?**; **schuirmann_comparison_1987?**). L'idée sous-jacente est que si l'effet est au moins aussi grand que les bornes de la zone d'équivalence, sur le long terme, on devrait le plus souvent rejeter l'hypothèse nulle. Par conséquent, un non rejet de l'hypothèse nulle devrait généralement signifier que l'effet n'atteint pas le SESOI et donc, que l'effet observé n'est pas pertinent. Bien que ce raisonnement puisse sembler tentant, de prime abord, il présente d'importantes limites.

Premièrement, le test n'a pas de bonnes propriétés asymptotiques. Ceci est illustré au sein de la Table 2, dans laquelle nous envisageons les mêmes scénarios que dans la Table 1 et ajoutons une contrainte de puissance: nous décidons qu'on ne peut conclure à l'équivalence que si l'on atteind une puissance de 80% pour détecter une différence de moyenne de .3. On constate qu'avec 100 sujets par groupes, aucune itération n'amènera à conclure à l'équivalence, pas même lorsque la différence entre les moyennes de population vaut 0. Cela s'explique par le fait que l'on n'atteind jamais la puissance minimale de 80%.[3] Par contre,

---

[2] On parle d'estimation et non de mesure, car la puissance du test dépend de σ, l'écart-type de la population, qu'on ne connait pas et devra donc estimer sur base de S, l'écart-type de l'échantillon (Schuirmann,1987).

[3] Avec 100 sujets par groupe, on estime la puissance du test à 80% lorsque l'estimation d de Cohen vaut .3981. Par conséquent, un test sera susceptible de conclure à l'équivalence si les bornes de la zone d'équivalence, exprimée en mesure standardisée d de Cohen, sont supérieures ou égales à .3981. Lorsqu'on fixe les bornes aux différences de moyennes $\pm.3$, cela n'est possible que si S est inférieur ou égal à .7535. En effet, $d = \frac{\theta}{S} \leftrightarrow .3981 = \frac{.3}{S} \leftrightarrow S = \frac{.3}{.3981} = .7535$. Or, avec 100 sujets par groupe, aucune estimation S ne

une fois les échantillons assez grands pour s'assurer une puissance de 80% pour détecter une différence de moyennes de population de .3, lorsque la différence entre les moyennes de populations est non nulle, la proportion d'itérations qui amènent à conclure à l'équivalence diminue à mesure que la taille des échantillons augmente. Par exemple, lorsque la différence de moyennes vaut .1 au niveau des populations, on conclura à l'équivalence dans 81% des itérations avec 200 sujets par groupe, contre seulement 54% des itérations avec 700 sujets par groupe [4].

. on sera amené à ne pas rejeter l'hypothèse nulle dans une proportion de cas inférieure à $(1 - \alpha)\%$ pour le scénarios dans lequel la différence de moyennes de population est nulle. *C'est vraisemblablement dû à la dispersion de la distribution d'échantillonnage de S, l'estimation de l'écart-type: pour un petit nombre d'itérations, l'estimation de S serait telle que la puissance estimée serait inférieure à* 80%.

car on n'atteint pas la puissance minimale requise. Une fois la puissance minimale requise atteinte, par contre, on constate que la proportion d'itérations qui amènent à conclure à l'équivalence diminue à mesure que la taille des échantillons augmente. Par exemple, si la probabilité de conclure à l'équivalence est d'environ 81% lorsque la vraie différence de moyenne vaut .1 et qu'il y a 200 sujets par groupe, cette probabilité tombe à approximativement 54% avec 700 sujets par groupe.

---

sera inférieure ou égale à .7535 lorsque σ vaut 1.

[4] En comparant les Tables 1 et 2, on constate qu'avec 200 sujets par groupes, les proportions d'itérations de chaque scénario qui amènent à conclure à l'équivalence, dans la Table 2, sont inférieures aux proportions d'itérations de chaque scénarios qui amènent à ne pas rejeter l'hypothèse nulle, dans la Table 1. Plus les échantillons sont grands, plus la valeur maximale de S permettant d'assurer la puissance des 80% sera élevée. Par exemple, avec 200 sujets par groupes, la valeur maximale autorisée pour S sera de $\frac{.3}{.2808} = 1.07$. Avec 300 sujets par groupes, la valeur maximale autorisée pour S sera de $\frac{.3}{.2291} = 1.31$

Table 2.

Proportion d'itérations qui amènent à conclure à l'équivalence en fonction de la taille des échantillons (n_j) et de la différence entre les moyennes de chaque population ($\mu_1 - \mu_2$), quand on exige une puissance minimale de 80% pour détecter une différence de moyenne de .3.

	Différence de moyennes dans la population ($\mu_1 - \mu_2$)					
n_j	0	.1	.2	.3	.4	.5
100	< .001	< .001	< .001	< .001	< .001	< .001
200	0.923	0.809	0.469	0.146	0.020	0.001
300	0.950	0.768	0.312	0.043	0.001	< .001
400	0.949	0.708	0.194	0.011	< .001	< .001
500	0.949	0.646	0.116	0.003	< .001	< .001
600	0.950	0.591	0.066	0.001	< .001	< .001
700	0.948	0.536	0.037	< .001	< .001	< .001

Note. Pour chaque scénario, les deux échantillons sont toujours de même taille ($n_1 = n_2 = n$) et sont extraits de populations se distribuant normalement et ayant la même variance ($\sigma_1 = \sigma_2 = \sigma$). La moyenne de la première population (μ_1) vaut systématiquement 0, et celle de la deuxième population (μ_2) varie de sorte à obtenir la différence de moyenne $\mu_1 - \mu_2$ désirée. Par ailleurs, σ vaut systématiquement 1, si bien que la différence de moyenne brute est égale au δ de Cohen.

Deuxièmement, pour une taille d'échantillon donnée, plus l'erreur (la variabilité des scores au sein de chaque groupe) sera plus grande (**meyners_equivalence_2012?**; **schuirmann_comparison_1987?**), plus la probabilité de conclure à l'équivalence augmentera. Ce dernier point est illustré au sein de la Figure 1, dans le contexte de la comparaison de deux moyennes. Sur l'axe des abscisses, on représente différentes estimations de la différence de moyenne ($\bar{X}_1 - \bar{X}_2$) et sur l'axe des ordonnées, la précision des estimations $\bar{X}_1 - \bar{X}_2$ ($S\sqrt{\frac{2}{n}}$ correspond à l'estimation de l'erreur standard de $\bar{X}_1 - \bar{X}_2$, avec S étant l'écart-type poolé et n la taille de chaque échantillon, lorsque les échantillons ont tous les deux la même taille et sont extraits de population ayant la même variance)[5]. Le triangle grisé représente l'ensemble des combinaisons estimation/précision qui vont amener à conclure à l'équivalence, avec l'approche de la puissance, lorsqu'on travaille avec

[5] Par facilité, à l'instar de Schuirman (1987), on envisage le cas où les échantillons sont de même taille et que l'on suppose que la condition d'homogénéité des variances est respectée. Notons cependant que d'après Schuirman, ce raisonnement peut être généralisé aux scénarios où les deux échantillons n'ont pas la même taille et sont extraits de population n'ayant pas la même variance.

des échantillons de taille 50, en acceptant un risque α de 5% et en exigeant une puissance minimale de 80% pour détecter une différence de 20 unités ($|\theta_j = 20|,\ j = 1, 2$). Dans cet exemple, pour toutes les valeurs de $S\sqrt{\frac{2}{n}}$ supérieures à 7.07 aucune estimation de différence de moyennes ne permettra de conclure à l'équivalence (pas même 0) puisque la puissance du test à détecter une différence de 20 unités est inférieure à 80%. Pour toutes les valeurs de $S\sqrt{\frac{2}{n}}$ inférieures à 7.07, on constate que plus notre estimation de $\bar{X}_1 - \bar{X}_2$ est précise (lorsqu'on se déplace du haut vers le bas, sur l'axe des ordonnées), plus l'estimation doit être proche de 0 pour pouvoir conclure à l'équivalence. Comme on peut le voir à travers le triangle hachuré sur la Figure 1, cette propriété peu désirable n'est pas partagée par le TOST, un test d'équivalence que nous allons décrire ci-dessous (**schuirmann_comparison_1987?**).

Les tests d'équivalence

Avec les tests d'équivalence, il n'est pas possible de démontrer qu'un effet vaille exactement zéro (**meyners_equivalence_2012?**). Il est par contre possible de montrer que l'effet observé est suffisamment petit pour être jugé non pertinent. Or, cela peut s'avérer précieux dans de nombreuses situations, par exemple pour justifier la décision de regrouper plusieurs groupes de sujets ensemble (**rogers_using_1993?**), pour contrôler qu'il n'y ait pas de différence trop importante entre les groupes sur base de critères autres que le (ou les) facteur(s) d'intérêts en cas de quasi-expérience (**seaman_equivalence_1998?**) ou encore pour falsifier une théorie qui prônerait en faveur d'un effet dépassant une certaine taille (**lakens_equivalence_2017?**; **anderson_theres_2016?**).

Le point de départ des tests d'équivalence est de définir θ_1 et θ_2, les bornes inférieures et supérieures de la zone d'équivalence, cette dernière contenant l'ensemble des valeurs jugées trop petites pour être susceptibles de nous intéresser. Ces bornes peuvent être exprimées soit dans l'unité des données brutes, soit en terme standardisé, mais doivent être définies avant la récolte des données (**anderson_theres_2016?**; **lakens_equivalence_2018?**). Il existe ensuite plusieurs approches pour démontrer que l'effet observé se situe dans la zone d'équivalence (voir **meyners_equivalence_2012?**, par exemple). Parmis celles-ci, une approche très simple est celle du "Two one-sided tests" (**schuirmann_comparison_1987?**; **lakens_equivalence_2017?**), plus communément appelé le TOST [6]. Le principe est de définir deux hypothèses nulles. La première est que l'effet observé est inférieur à la borne inférieure de la zone d'équivalence:

$$H0_1 : \theta < \theta_1, \ avec\ \theta_1 \neq 0$$

La deuxième est que l'effet observé est supérieur à la borne supérieure de la zone

[6] Il existe des alternatives au TOST qui sont très légèrement plus puissantes, mais le gain marginal en termes de puissance est contrebalancé par un niveau de complexité beaucoup plus élevé (Meyners, 2012).

d'équivalence:

$$H0_2 : \theta > \theta_2, \ avec \ \theta_2 \neq 0$$

Lorsque les deux hypothèses nulle peuvent être simultanément rejetées, on peut conclure à l'équivalence (**seaman_equivalence_1998?**). Cela équivaut, statistiquement parlant, à montrer que l'intervalle de confiance à $(1 - 2 \times \alpha)\%$ est entièrement inclus dans la zone d'équivalence (**seaman_equivalence_1998?**; **lakens_equivalence_2017?**). Notons qu'il n'est pas nécessaire de reporter les résulats des deux tests unilatéraux, lorsqu'on réalise le TOST: il suffit de reporter les résultats du test associé à la plus petite valeur de statistique (et par conséquent, à la plus grande p-valeur). En effet, si ce test amène à conclure au rejet de l'hypothèse nulle, le second test amènera automatiquement à la même conclusion (**rogers_using_1993?**; **lakens_equivalence_2018?**). Cette remarque reste vraie dans le cas particulier où les deux tests sont associés à la même valeur de statistique puisque dans ce cas, les deux tests mèneront à une conclusion identique (**rogers_using_1993?**). Notons également qu'il n'est pas nécessaire de procéder à une correction du risque alpha dûe à la réalisation simultanée de deux tests. En effet, une erreur de type I (rejeter à tort l'hypothèse nulle) ne peut être commise que si l'hypothèse nulle est vraie. Or, les deux hypothèses nulles testées sont mutuellement exclusives: il n'est pas possible que θ soit simultanément inférieur à θ_1 (ce qui correspond à $H0_1$) et supérieur à θ_2 (ce qui correspond à $H0_2$).

Jusqu'il y a peu, le TOST n'était pas disponible dans la plupart des logiciels, à l'exception de Minitab, ce qui constituait un frein important à son usage. Pour cette raison, (**lakens_20_2016?**) a créé le package R "TOSTER" et plus récemment encore, ce même package a été implémenté dans Jamovi [7]. Tant dans R que dans Jamovi, le package

[7] Jamovi est un logiciel clic-bouton entièrement gratuit qui gagne en popularité et qui présente, parmi ses nombreux avantages, le fait d'être particulièrement convivial. Dans la mesure où la plupart des chercheurs sont plus enclins à utiliser des procédures si elles sont implémentées dans ce type de logiciel (Fraas & Newman, 2000), cela constitue une excellente nouvelle pour le devenir du TOST dans la recherche en

compare simultanément l'effet observé à l'absence d'effet (cela correspond au test traditionnel) ainsi qu'aux deux bornes de la zone d'équivalence (cela correspond au TOST). Il en découle 4 conclusions distinctes possibles (**lakens_equivalence_2017?**), qui sont illustrées dans la figure 2 dans le contexte de la comparaison de deux moyennes indépendantes:

(1) La différence de moyenne observée diffère significativement des deux bornes d'équivalence, mais pas de 0 (scénario A, Figure 2): dans ce cas, on conclura à l'absence d'un effet au moins aussi grand que les bornes d'équivalence.

(2) La différence de moyenne observée diffère significativement des deux bornes d'équivalence ainsi que de 0 (scénario B, Figure 2): on conclura alors qu'il existe un effet non nul, mais qui ne dépasse pas une certaine taille fixée par les bornes. C'est ce qui arrive typiquement lorsqu'on travaille avec de très grands échantillons, si bien que le test traditionnel est très puissant, même pour détecter des effets très petits (**rogers_using_1993?**).

(3) La différence de moyenne observée diffère significativement de 0, mais ne diffère pas significativement d'au moins une des deux bornes d'équivalence (scénario C, Figure 2): on conclura alors à la présence d'un effet non nul (**rogers_using_1993?**).

(4) La différence de moyenne observée ne diffère significativement ni d'au moins une des deux bornes d'équivalence, ni de 0 (scénario D, Figure 2): c'est ce qui arrive lorsque les données sont si imprécises qu'on ne peut tirer aucune conclusion. Les données semblent compatibles tant avec un effet nul qu'avec un effet supérieur au SESOI.

---

psychologie.

Définir les bornes de la zone d'équivalence

L'aspect le plus compliqué dans la réalisation du TOST est la définition des bornes d'équivalence. Dans certains cas, il est possible de définir un critère objectif qui permettra de déterminer à partir de quand un effet est jugé pertinent (**lakens_equivalence_2018?**). Dans ce cas, établir l'équivalence revient à rejeter la présence d'un effet ayant un quelconque intérêt pratique (**rogers_using_1993?**). Par exemple, (**burriss_changes_2015?**) avaient émis l'hypothèse qu'une augmentation de la rougeur de la peau chez femmes les rendraient plus attractives pour les hommes en période d'ovulation. Or, une telle hypothèse n'est crédible que si le changement facial est visible à l'oeil nu. Dans ce contexte, le SESOI serait la plus petite variation dans la rougeur de la peau qu'il est possible de détecter à l'oeil nu (**lakens_equivalence_2018?**). Il est également parfois possible pour des experts de déterminer expérimentalement ce qui constitue un changement important, pour certaines échelles de mesures fréquemment utilisées en psychologie, à l'instar de (**button_minimal_2015?**) qui se sont penchés sur le BDI[8]. Ces auteurs ont interrogé un grand nombre de patients quant à leur ressenti subjectif en termes d'amélioration de leur dépression dans un certain laps de temps, et ont comparé leurs réponses à la différence de score obtenu à l'aide du BDI dans ce même laps de temps (**lakens_equivalence_2018?**).

Malheureusement, il n'est pas toujours possible d'établir un critère objectif en vue de définir les bornes d'équivalence. Dans ce cas, il existe diverses stratégies, plus subjectives, en vue d'établir ces bornes. En les utilisant, il faut cependant avoir conscience du fait que la question à laquelle nous répondons varie en fonction de la stratégie utilisée.

---

[8] Le BDI (Beck Depression Inventory) est une échelle auto-rapportée évaluant les symptômes cognitifs courants de la dépression. Cette échelle est constituée de 21 items évalués à l'aide des échelles de Likert allant de 0 à 3, ce qui donne un score total compris entre 0 et 63 qui sera d'autant plus élevé que la dépression sera sévère (Button et al., 2015).

Premièrement, il est possible de déterminer des bornes en s'inspirant de balises existantes, en vue d'exclure la présence d'un effet jugé petit, moyen ou grand par ces balises (**lakens_equivalence_2018?**). Notons que si cette stratégie est tentante de par sa simplicité, elle doit être utilisée avec prudence. D'abord, un effet ne devrait être qualifié de petit, moyen ou grand qu'en comparaison à d'autres effets connus, et non sur base d'impressions qualitatives (**gignac_effect_2016?**). Dit autrement, il est important d'avoir un cadre de référence pour juger de la taille d'un effet. Or, les balises de Cohen (en l'occurence, les balises les plus célèbres et les plus largement utilisées) sont dépourvues de ce cadre de référence, puisqu'elles ont été établies à une époque où très peu de chercheurs se préoccupaient de la taille des effets étudiés (**funder_evaluating_2019?**). Depuis Cohen, certains chercheurs ont déployé de gros efforts en vue d'établir de nouvelles normes sur base d'analyses systématiques quantitative de la litérature. (**gignac_effect_2016?**), par exemple, ont établi de nouvelles balises pour interpréter le r de Pearson, en définissant les quartiles d'une distribution de 708 mesures dérivées de méta-analyses issues de la psychologie sociale et de la personnalité. C'est de la sorte qu'ils ont proposé d'interpréter respectivement des mesures de 0.10, 0.20 et 0.30, dans ces domaines de la psychologie, comme représentant des effets relativement petits, typiques et relativement larges. Ces normes ont également été approuvées par (**funder_evaluating_2019?**). Ensuite, les balises ne prennent pas en compte le contexte de l'étude si bien que statuer sur la taille d'un effet ne fournit pas nécessairement d'information sur sa valeur. *Un effet même de très petite taille peut faire une différence pour les personnes concernées. Yzerbit donne l'exemple suivant: l'effet de l'aspirine sur l'espérance de vie est minimale, mais pour les gens qui survivent, ça fait une différence.. Autre exemple: Imaginons un médicament contre la dépression qui permettrait de réduire les symptômes un tout petit peu mais qui ne coûte presque rien. Une faible différence d'un point de vue statistique peut devenir une grande différence pour les personnes concernées. Idem pour un médicament qui sauve des vies.* C'est pour cette raison que les balises devraient toujours être utilisées en dernier recours,

lorsqu'on ne dispose d'aucune information contextuelle.

Deuxièmement, dans le contexte d'études de réplication, il est possible de se baser sur les résultats d'études antérieures pour définir les bornes d'équivalence. 2.1. (Levine et al. 2007): se baser sur les tailles d'effet suggérées dans la littérature (sur base de méta-analyses)[9].*Prenons l'exemple de l'augmentation de la pensée agressive quand on joue à des jeux violents. D'après une méta-analyse de Ferguson (2007), cette corrélation serait de r =.25 (ce qui correspond à un d de Cohen de .51). Je peux utiliser cette valeur comme borne pour définir l'intervalle d'équivalence. Si je parviens à montrer qu'il y a équivalence, je montre que l'effet étudié serait vraisemblablement plus petit que suggéré dans la littérature.* .

–> Remarque: la méta-analyse elle donne .25, OK, mais en réalité, il y a une distribution autour de l'effet dans la littérature. Une solution plus conservatrice est alors de se baser sur les bornes inférieures de l'IC autour de la valeur de la méta-analyse. –> Comme dit Vincent, c'est la réalité de l'effet. C'est comme si tu faisais un test entre deux moyennes en prétendant que la myenne est une valeur et pas une distribution autorud e cette effet. Et donc c'est un peu délicat d'aller dire qu'on va tester un test d'équivalence pour contester l'effet de la violence en prenant la valeur obtenue dans la méta analyse comme une borne absolue et pas une distribution. Du coup, au minimum il essayerait d'intégrer le fait qu'il y a une distribution autour de la valeur, mais donc on ne peut pas décider que si on est en dessous de .50 ou au dessus de -.50.

–> Autre probleme, pourquoi prendre l'opposé comme borne inférieure? Alors que la méta analyse ne parle pas d'effet inverse. –> Pour l'heure, je serais plus d'avis de confirmer la méta-analyse en montrer que l'effet est bien situé à l'intérieur des bornes de l'IC autour de cet effet dans la méta-analyse. 2.2. Solution proposée par Simonsohn (2015) pour remettre

[9] Il est mieux de se baser sur des résultats de méta-analsye que d'étude isolée parce qu'à cause du biais de publication, les tailles d'effet observées sont souvent une sur-représentation de la réalité et donc si on a bcp de sujets dans notre étude, il y a vraiment bcp de chance qu'on démontre l'équivalence

en question la pertinence de l'outil qui a ét utilsié pour démontrer un effet dans une étude antérieure. Il dit que si on a pour un effet d'intérêt une puissance inférieure à 33% de le détecter, on a vraiment un gros problème de puissance (pourquoi 33% ça reste arbitraire, of course). il part de cette optique là et il se dit que du coup, ça pourrait être intérssant de définir les bornes de la zone d'équivalence en considérant un effet que l'étude d'origine aurait pu détecter avec une puissance de 33%. Ce faisant, si on parvient à démontrer que l'effet est encore plus petit qu'un effet que l'étude ne base n'aurait pu détecter qu'avec une puissance de 33%, il y a peu de chance pour que l'effet originellement proposé par l'étude d'origine soit vraiment basé sur un outil pertinent. *exemple concrêt: imaginons un test t de Student et qu'on a 21 sujets par groupe. On peut déterminer, en faisant une analyse de sensibilité dans Gpower, qu'on a une puissance 33% à détecter un effet d de COhen de .48. Du coup, on considererais .48 comme valeur pour notre zone d'équivalence.*

2.3. Lakens (2018): essayer de deviner implicitement ce que l'auteur de l'étude d'origine aurait pu considérer comme un effet pertinent (s'il n'a pas donné d'indication dans son article pour dire " je considère qu'un effet est pertinent à partir de telle valeur"). Cela peut se faire sur base de la taille d'échantillon utilisée par cette personne. *On ne pourra détecter un effet comme significatif que si la valeur de statistique observée dépasse une valeur seuil (la valeur critique). Il est possible, grâce à la relation qui existe entre la statistique t et la statistique d de Cohen, de déterminer à quelle "taille d'effet critique" correspond la statistique t critique. Par exemple, si le chercheur a utilisé 30 sujets par groupe: via gpower, on peut déterminer qu'il faudra une statistique observée t de minimum 2.045 pour pouvoir conclure au RH0. et compte tenu du lien entre la statistique t et la stat d, ça correspond à un d de Cohen de .373 (d_crit = t_crit/racine(n)). Concrètement, si on observe une taille d'effet supérieure ou égale à .373, on pourra conclure au rejet de l'H0. Si la taille d'effet est plus petite, on ne sera pas capable de conclure au rejet de l'hypothèse nulle. L'idée ce serait de démontrer qu'il y a équivalence, la personne qui a écrit l'article d'origine a utilisé une taille d'échantillon insuffisante pour étudier l'effet suggéré et donc,*

si on veut étudier ce même sujet d'étude, il faudrait nécessairement récolter des échantillons plus grands pour être capable de le faire correctement.

3) de se baser sur les ressources dont on dispose (analyse de sensibilité). Si moi je ne suis pas capable d'avoir un échantillon de plus de 2000 personnes, il y a certains effets que je ne serai pas capable de calculer. Et donc, je peux utiliser cette taille d'échantillon pour déterminer la taille d'effet dont je suis certain que je pourrai raisonnablement conclure au rejet de l'hypothèse nulle. Et donc là, si on démontre qu'il y a équivalence, on ne tire pas la ccl que l'effet n'est pas pertinent, mais que cet effet qu'on a envie d'étudier ne peut être l'être sur base des tailles d'échantillon qu'on a l'habitude d'utiliser.

Il est important de bien comprendre qu'en fonction de la stratégie utilisée, on ne se posera pas nécessairement la même question de recherche (et la réponse obtenue sera nécessairement liée à cette question de recherche).

1) définir comme limites la plus petite taille d'effet pour laquelle on peut atteindre une puissance de détection suffisante (déterminé par les ressources disponibles pour étudier l'effet, (**lakens_equivalence_2017?**)) –> voir la section "Setting equivalence bounds" p. 359 mais je crois que j'en parle aussi dans la vidéo SOCLAB.
2) le SESOI peut parfois être fixé objectivement
3) Idéalement basé sur une analyse coût-bénéfice). Attention: bien sûr une dimension subjective dans la définition des coûts et des bénéfices. Attention: le SESOI doit être déterminé AVANT et INDEPENDAMMENT des données.

Comparaison du TOST et du SGPV

Meta-Psychology, 2020, vol 4, MP.2018.933
https://doi.org/10.15626/MP.2018.933
Article type: Original Article

Open data: N/A
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: N/A

Edited by: S. R. Martin
Reviewed by: J. D. Blume, O. L. Olvera Astivia
Analysis reproduced by: André Kalmendal
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/ZP3KF

Equivalence Testing and the Second Generation P-Value.

Daniël Lakens
Eindhoven University of Technology, The Netherlands

Marie Delacre
Université Libre de Bruxelles, Belgium

Abstract

To move beyond the limitations of null-hypothesis tests, statistical approaches have been developed where the observed data are compared against a range of values that are equivalent to the absence of a meaningful effect. Specifying a range of values around zero allows researchers to statistically reject the presence of effects large enough to matter, and prevents practically insignificant effects from being interpreted as a statistically significant difference. We compare the behavior of the recently proposed second generation *p*-value (Blume, D'Agostino McGowan, Dupont, & Greevy, 2018) with the more established Two One-Sided Tests (TOST) equivalence testing procedure (Schuirmann, 1987). We show that the two approaches yield almost identical results under optimal conditions. Under suboptimal conditions (e.g., when the confidence interval is wider than the equivalence range, or when confidence intervals are asymmetric) the second generation *p*-value becomes difficult to interpret. The second generation *p*-value is interpretable in a dichotomous manner (i.e., when the SGPV equals 0 or 1 because the confidence intervals lies completely within or outside of the equivalence range), but this dichotomous interpretation does not require calculations. We conclude that equivalence tests yield more consistent *p*-values, distinguish between datasets that yield the same second generation *p*-value, and allow for easier control of Type I and Type II error rates.

Keywords: equivalence testing, second generation p-values, hypothesis testing, TOST, statistical inference

To test predictions researchers predominantly rely on null-hypothesis tests. This statistical approach can be used to examine whether observed data are sufficiently surprising under the null hypothesis to reject an effect that equals exactly zero. Null-hypothesis tests have an important limitation, in that this procedure can only reject the hypothesis that there is no effect, while scientists should also be able to provide statistical support for *equivalence*. When testing for equivalence researchers aim to examine whether an observed effect is too small to be considered meaningful, and therefore is practically equivalent to zero. By specifying a range around the null hypothesis of values that are deemed practically equivalent to the absence of an effect (i.e., 0 +- 0.3) the observed data can be compared against an *equivalence range* and researchers can test if a meaningful effect is absent (Hauck & Anderson, 1984; Kruschke, 2018; Rogers, Howard, & Vessey, 1993; Serlin & Lapsley, 1985; Spiegelhalter, Freedman, & Parmar, 1994; Wellek, 2010; Westlake, 1972).

Second generation *p*-values (SGPV) were recently proposed as a statistic that represents "the proportion of

2

data-supported hypotheses that are also null hypotheses" (Blume et al., 2018). The researcher specifies an equivalence range around a null hypothesis of values that are considered practically equivalent to the null hypothesis. The SGPV measures the degree to which a set of data-supported parameter values falls within the interval null hypothesis. If the estimation interval falls completely within the equivalence range, the SGPV is 1. If the confidence interval falls completely outside of the equivalence range, the SGPV is 0. Otherwise the SGPV is a value between 0 and 1 that expresses the overlap of data-supported hypotheses and the equivalence range. When calculating the SGPV the set of data-supported parameter values can be represented by a confidence interval (CI), although one could also choose to use credible intervals or Likelihood support intervals (SI). When a confidence interval is used, the SGPV and equivalence tests such as the Two One-Sided Tests (TOST) procedure (Lakens, 2017; Meyners, 2012; Quertemont, 2011; Schuirmann, 1987) appear to have close ties, because both tests compare a confidence interval against an equivalence range. Here, we aim to examine the similarities and differences between the TOST procedure and the SGPV. We limit our analysis to continuous data sampled from a bivariate normal distribution. The TOST procedure also relies on the confidence interval around the effect. In the TOST procedure the data are tested against the lower equivalence bound in the first one-sided test, and against the upper equivalence bound in the second one-sided test (Lakens, Scheel, & Isager, 2018). For an excellent discussion of the strengths and weaknesses of different frequentist equivalence tests, including alternatives to the TOST procedure, see Meyners (2012). If both tests statistically reject an effect as extreme or more extreme than the equivalence bound, you can conclude the observed effect is practically equivalent to zero from a Neyman-Pearson approach to statistical inferences. Because one-sided tests are performed, one can also conclude equivalence by checking whether the 1-2×α confidence interval (e.g., when the alpha level is 0.05, a 90% CI) falls completely within the equivalence bounds. Because both equivalence tests as the SGPV are based on whether and how much a confidence interval overlaps with equivalence bounds, it seems worthwhile to compare the behavior of the newly proposed SGPV to equivalence tests to examine the unique contribution of the SGPV to the statistical toolbox.

The relationship between *p*-values from TOST and SGPV when confidence intervals are symmetrical

The second generation *p*-value (SGPV) is calculated as:

$$p_\delta = \frac{|I \cap H_0|}{|I|} \times \max\left\{\frac{|I|}{2\,|H_0|}, 1\right\}$$

where I is the interval based on the data (e.g., a 95% confidence interval) and H_0 is the equivalence range. The first term of this formula implies that the second generation *p*-value is the width of the confidence interval that overlaps with the equivalence range, divided by the total width of the confidence interval. The second term is a "small sample correction" (which will be discussed later) that comes into play whenever the confidence interval is more than twice as wide as the equivalence range. To examine the relation between the TOST *p*-value and the SGPV we can calculate both statistics across a range of observed effect sizes. Building on the example by Blume et al. (2018), in Figure 1 *p*-values are plotted for the TOST procedure and the SGPV. The statistics are calculated for hypothetical one-sample *t*-tests for observed means ranging from 140 to 150 (on the x-axis). The equivalence range is set to 145 +- 2 (i.e., an equivalence range from 143 to 147), the observed standard deviation is assumed to be 2, and the sample size is 30. For example, for the left-most point in Figure 1 the SGPV and the TOST *p*-value is calculated for a hypothetical study with a sample size of 30, an observed standard deviation of 2, and an observed mean of 140, where the *p*-value for the equivalence test is 1, and the SGPV is 0.

Our conclusions about the relationship between TOST *p*-values and SGPV hold for second generation *p*-values calculated from confidence intervals, and assuming data is sampled from a bivariate normal distribution. Readers can explore the relationship between TOST *p*-values and SGPV for themselves in an online Shiny app: http://shiny.ieis.tue.nl/TOST_vs_SGPV/.

The SGPV treats the equivalence range as the null-hypothesis, while the TOST procedure treats the values outside of the equivalence range as the null-hypothesis. For ease of comparison we can plot 1-SGPV (see Figure 2) to make the values more easily comparable. We see that the *p*-value from the TOST procedure and the SGPV follow each other closely. When we discuss the relationship between the *p*-values from TOST and the SGPV, we focus on their correspondence at three values, namely where the TOST $p = 0.025$ and SGPV is 1, where the TOST $p = 0.5$ and SGPV $= 0.5$, and where the TOST $p = 0.975$ and SGPV $= 1$. These three values are important for the SGPV because they indicate the values at which the SGPV indicates the data should be interpreted as compatible with the null hypothesis (SGPV =

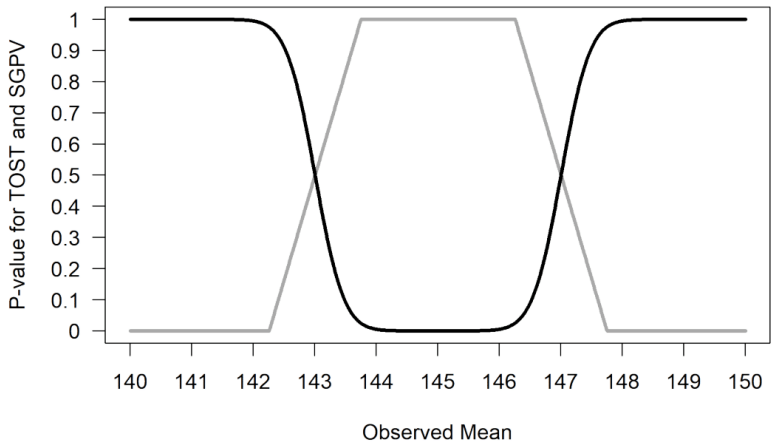

Figure 1. Comparison of *p*-values from TOST (black line) and SGPV (grey line) across a range of observed sample means (x-axis) tested against a mean of 145 in a one-sample *t*-test with a sample size of 30 and a standard deviation of 2, illustrating that when the TOST *p*-value = 0.5, the SGPV = 0.5, when the TOST *p*-value is 0.975, 1-SGPV = 1, and when the TOST *p*-value = 0.025, 1-SGPV = 0.

1), or with the alternative hypothesis (SGPV = 0), or when the data are strictly inconclusive (SGPV = 0.5).

These three points of overlap are indicated by the horizontal dotted lines in Figure 2 at TOST *p*-values of 0.975, 0.5, and 0.025.

When the observed sample mean is 145, the sample size is 30, and the standard deviation is 2, and we are testing against equivalence bounds of 143 and 147 using the TOST procedure for a one-sample *t*-test, the equivalence test is significant, $t(29) = 5.48$, $p < .001$. Because the 95% CI falls completely within the equivalence bounds, the SGPV is 1 (see Figure 1). On the other hand, when the observed mean is 140, the equivalence test is not significant (the observed mean is far outside the equivalence range of 143 to 147), $t(29) = -8.22$, $p = 1$ (or more accurately, $p > .999$ as *p*-values are bounded between 0 and 1). Because the 95% CI falls completely outside the equivalence bounds, the SGPV is 0 (see Figure 1).

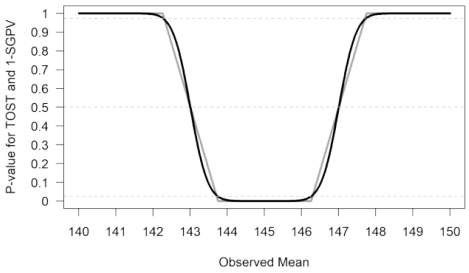

Figure 2. Comparison of *p*-values from TOST (black line) and 1-SGPV (grey line) across a range of observed sample means (x-axis) tested against a mean of 145 in a one-sample *t*-test with a sample size of 30 and a standard deviation of 2.

SGPV as a uniform measure of overlap

It is clear the SGPV and the *p*-value from TOST are closely related. When confidence intervals are symmetric we can think of the SGPV as a straight line that is directly related to the *p*-value from an equivalence test for three values. When the TOST *p*-value is 0.5, the SGPV is also 0.5 (note that the reverse is not true). The

4

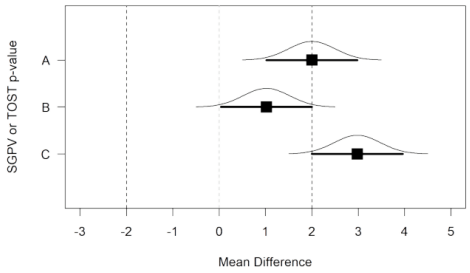

Figure 3. Means, normal distribution, and 95% CI for three example datasets that illustrate the relationship between *p*-values from TOST and SGPV.

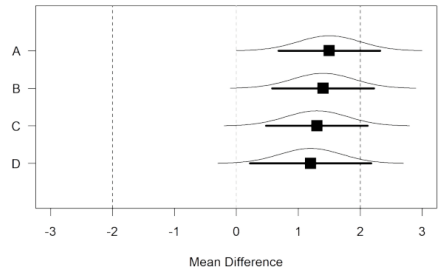

Figure 4. Means, normal distribution, and 95% CI for samples where the observed population mean is 1.5, 1.4, 1.3, and 1.2.

SGPV is 50% when the observed mean falls exactly on the lower or upper equivalence bound, because 50% of the symmetrical confidence interval overlaps with the equivalence range. When the observed mean equals the equivalence bound, the difference between the mean in the data and the equivalence bound is 0, the *t*-value for the equivalence test is also 0, and thus the *p*-value is 0.5 (situation A, Figure 3).

Two other points always have to overlap. When the 95% CI falls completely inside the equivalence region, and one endpoint of the confidence interval is exactly equal to one of the equivalence bounds (see situation B in Figure 3) the TOST *p*-value (which relies on a one-sided test) is always 0.025, and the SGPV is 1. Note that when sample sizes are small or equivalence bounds are narrow, small *p*-values for the TOST or a SGPV = 1 might not be observed in practice if too few observations are collected. The third point where the SGPV and the *p*-value from the TOST procedure should overlap is where the 95% CI falls completely outside of the equivalence range, but one endpoint of the confidence interval is equal to the equivalence bound (see situation C in Figure 3), when the *p*-value will always be 0.975, and the SGPV is 0. Note that this situation is in essence a minimum-effect test (Murphy, Myors, & Wolach, 2014). The goal of a minimum-effect is not just to reject a difference of zero, but to reject the smallest effect size of interest (i.e., the equivalence bounds). An equivalence test and minimum effect test against the same equivalence bound are complementary, and when a TOST *p*-value is larger than 0.975, the *p*-value for the minimum effect test is smaller than 0.05 (and therefore the minimum effect test provides no additional information that can not be derived from the *p*-value from the equivalence test). The SGPV summarizes the information from an equivalence test (and the complementary minimum-effect test). These can be two relevant questions to ask, although it often makes sense to combine an equivalence test and a null-hypothesis test instead (Lakens et al., 2018).

For example, in Figure 4 we have plotted four SGPV's. From A to D the SGPV is 0.76, 0.81, 0.86, and 0.91. The difference in the percentage of overlap between A and B (-0.05) is identical to the difference in the percentage of overlap between C and D as the mean gets 0.1 closer to the test value (-0.05). As the observed mean in a one-sample *t*-test lies closer to the test value, from situation A to D, the difference in the overlap changes uniformly. As we move the observed mean closer to the test value in steps of 0.1 across A to D the *p*-value calculated for normally distributed data are not uniformly distributed. The probability of observing data more extreme than the upper bound of 2 is (from A to D) 0.16, 0.12, 0.08, and 0.05. As we can see, the difference between A and B (0.04) is not the same as the difference between C and D (0.03). Indeed, the difference in *p*-values is the largest as you start at $p = 0.5$ (when the observed mean falls on the test value), which is why the line in Figure 1 is the steepest at $p = 0.5$. Note that where the SGPV reaches 1 or 0, *p*-values closely approximate 0 and 1, but never reach these values.

When different *p*-values for equivalence tests yield the same SGPV

There are three situations where *p*-values for TOST differentiate between observed results, while the SGPV does not differentiate. The first two situations were discussed before and can be seen in Figure 1. When

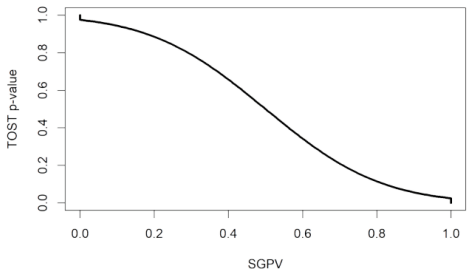

Figure 5. The relationship between *p*-values from the TOST procedure and the SGPV for the same scenario as in Figure 1.

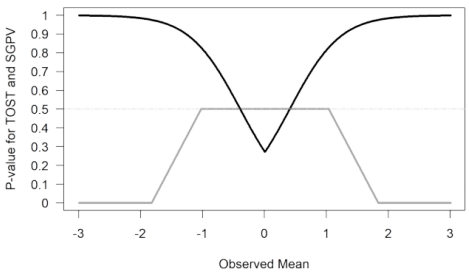

Figure 6. Comparison of *p*-values from TOST (black line) and SGPV (grey line) across a range of observed sample means (x-axis). Because the sample size is small (n = 10) and with a standard deviation of 2 the CI is more than twice as wide as the equivalence range (set to -0.4 to 0.4), the SGPV is set to 0.5 (horizontal light-grey line) across a range of observed means.

the SGPV is either 0 or 1, *p*-values from the equivalence test fall between 0.975 and 1 or between 0 and 0.025. Where the SGPV is 1 as long as the confidence interval falls completely within the equivalence bounds, the *p*-value for the TOST continues to differentiate between results as a function of how far the confidence interval lies within the equivalence bounds (the further the confidence interval is from both bounds, the lower the *p*-value). The easiest way to see this is by plotting the SGPV against the *p*-value from the TOST procedure. The situations where the *p*-values from the TOST procedure continue to differentiate based on how extreme the results are, but the SGPV is a fixed value are indicated by the parts of the curve where there are vertical straight lines at second generation *p*-values of 0 and 1.

A third situation in which the SGPV remains stable across a range of observed effects, while the TOST *p*-value continues to differentiate, is whenever the CI is wider than the equivalence range, and the CI overlaps with the upper *and* lower equivalence bound. When the confidence interval is more than twice as wide as the equivalence range the SGPV is set to 0.5. Blume et al. (2018) call this the "small sample correction factor". However, it is not a correction in the typical sense of the word, since the SGPV is not adjusted to any "correct" value. When the normal calculation would be "misleading" (i.e., the SGPV would be small, which normally would suggest support for the alternative hypothesis, but at the same time all values in the equivalence range are supported), the SGPV is set to 0.5 which according to Blume and colleagues signals that the SGPV is "uninformative". Note that the CI can be twice as wide as the equivalence range whenever the sample size is small (and the confidence interval width is large) *or* when then equivalence range is narrow. It is therefore not so much a "small sample correction" as it is an exception to the typical calculation of the SGPV whenever the ratio of the confidence interval width to the equivalence range exceeds 2:1 and the CI overlaps with the upper and lower bounds.

We can examine this situation by calculating the SGPV and performing the TOST for a situation where sample sizes are small and the equivalence range is narrow, such that the CI is more than twice as large as the equivalence range (see Figure 6). When the two statistics are plotted against each other we can see where the SGPV is the same while the TOST *p*-value still differentiates different observed means (indicated by straight lines in the curve, see Figure 7). We see the SGPV is 0.5 for a range of observed means where the *p*-value from the equivalence test still varies. It should be noted that in these calculations the *p*-values for the TOST procedure are *never* smaller than 0.05 (i.e., they do not get below 0.05 on the y-axis). In other words, we cannot conclude equivalence based on any of the observed means. This happens because we are examining a scenario where the 90% CI is so wide that it never falls completely within the two equivalence bounds.

As Lakens (2017) notes: "in small samples (where CIs are wide), a study might have no statistical power (i.e., the CI will always be so wide that it is necessarily wider than the equivalence bounds)." None of the *p*-values based on the TOST procedure are below 0.05, and thus, in the long run we have 0% power.

6

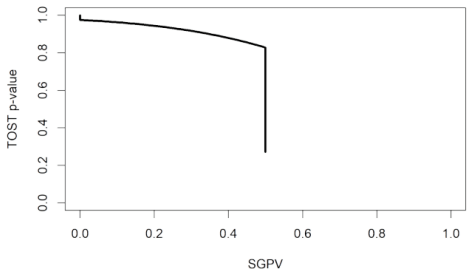

Figure 7. The relationship between *p*-values from the TOST procedure and the SGPV for the same scenario as in Figure 6.

The *p*-value from the TOST procedure still differentiates observed means, while the SGPV does not, when the CI is wider than the equivalence range (so the precision is low) and overlaps with the upper and lower equivalence bound, but the CI is *not* twice as wide as the equivalence range. In the example below, we see that the CI is only 1.79 times as wide as the equivalence bounds, but the CI overlaps with the lower and upper equivalence bounds (Figure 8). This means the SGPV is not set to 0.5, but it is constant across a range of observed means, while the TOST *p*-value is not constant across this range.

If the observed mean would be somewhat closer to 0, or further away from 0, the SGPV remains constant (the CI width does not change, and it completely overlaps with the equivalence range) while the *p*-value for the TOST procedure does vary. We can see this in Figure 9 below. The SGPV is not set to 0.5, but is slightly higher than 0.5 across a range of means. How high the SGPV will be for a CI that is not twice as wide as the equivalence range, but overlaps with the lower and upper equivalence bounds, depends on the width of the CI and the equivalence range.

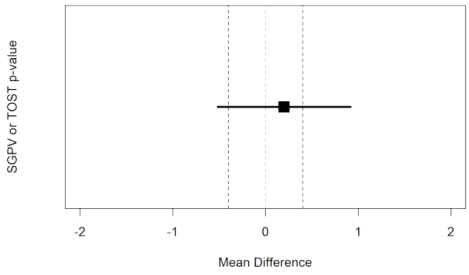

Figure 8. Example of a 95% CI that overlaps with the lower and upper equivalence bound (indicated by the vertical dotted lines).

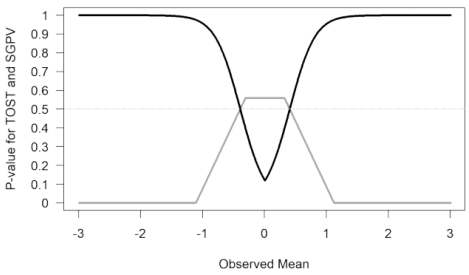

Figure 9. Comparison of *p*-values from TOST (black line) and SGPV (grey line) across a range of observed sample means (x-axis). The sample size is small (n = 10), but because the sd is half as big as in Figure 7 (1 instead of 2) the CI is less than twice as wide as the equivalence range (set to -0.4 to 0.4). The SGPV is not set to 0.5 (horizontal light grey line) but reaches a maximum slightly above 0.5 across a range of observed means.

If we once more plot the two statistics against each other we see the SGPV is 0.56 for a range of observed means where the *p*-value from the equivalence test still varies, as indicated by the straight section of the line (Figure 10).

To conclude this section, there are situations where the *p*-value from the TOST procedure continues to differentiate, while the SGPV does not. Therefore, interpreted as a continuous statistic, the SGPV is more limited than the *p*-value from the TOST procedure.

The relation between equivalence tests and SGPV for asymmetrical confidence intervals around correlations

So far we have only looked at the relation between equivalence tests and the SGPV when confidence intervals are symmetric (e.g., for confidence intervals around mean differences). For correlations, which are bound between -1 and 1, confidence intervals are only symmetric for a correlation of exactly 0. The confidence interval for a correlation becomes increasingly asymmetric

7

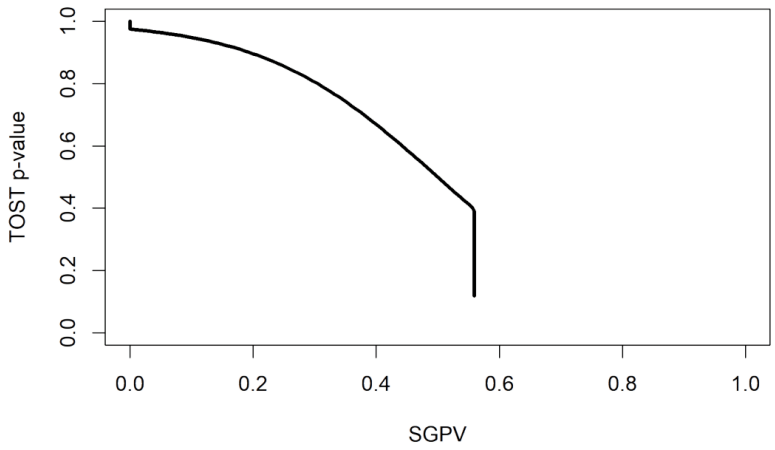

Figure 10. The relationship between *p*-values from the TOST procedure and the SGPV for the same scenario as in Figure 9.

as the observed correlation nears -1 or 1. For example, with ten observations, an observed correlation of 0 has a symmetric 95% confidence interval ranging from -0.63 to 0.63, while and observed correlation of 0.7 has an asymmetric 95% confidence interval ranging from 0.13 to 0.92. Note that calculating confidence intervals for a correlation involves a Fisher's z-transformation, which transforms values such that they are approximately normally z-distributed, which allows one to compute symmetric confidence intervals. These confidence intervals are then retransformed into a correlation, where the confidence intervals are asymmetric if the correlation is not exactly zero.

The effect of asymmetric confidence intervals around correlations is most noticeable at smaller sample sizes. In Figure 11 we plot the *p*-values from equivalence tests and the SGPV (again plotted as 1-SGPV for ease of comparison) for correlations. The sample size is 30 pairs of observations, and the lower and upper equivalence bounds are set to -0.45 and 0.45, with an alpha of 0.05. As the observed correlation in the sample moves from -.99 to 0 the *p*-value from the equivalence test becomes smaller, as does 1-SGPV. The pattern is quite similar to that in Figure 2. The *p*-value for the TOST procedure and 1-SGPV are still related as discussed above, with TOST *p*-values of 0.975 and 0.025 corresponding to a 1-SGPV of 1 and 0, respectively. There are two important differences, however. First of all, the SGPV is no longer a straight line, but a curve, due to the asymmetry in the 95% CI. Second, and most importantly, the *p*-value for the equivalence test and the SGPV do no longer overlap at $p = 0.5$.

The reason that the equivalence test and SGPV no longer overlap is due to asymmetric confidence intervals. If the observed correlation falls exactly on the equivalence bound the *p*-value for the equivalence test is 0.5. In the equivalence test for correlations the *p*-value is computed based on a z-transformation which better controls error rates (Goertzen & Cribbie, 2010). This transformation is computed as follows, where r is the observed correlation and ρ is the theoretical correlation under the null:

$$z = \frac{\frac{\log(\frac{1+r}{1-r})}{2} - \frac{\log(\frac{1+\rho}{1-\rho})}{2}}{\sqrt{\frac{1}{n-3}}}$$

Because the z-distribution is symmetric, the probability of observing the observed or more extreme z-score, assuming the equivalence bound is the true effect size, is 50%. However, because the r distribution is not symmetric, this does not mean that there is always a 50% probability of observing a correlation smaller or larger

8

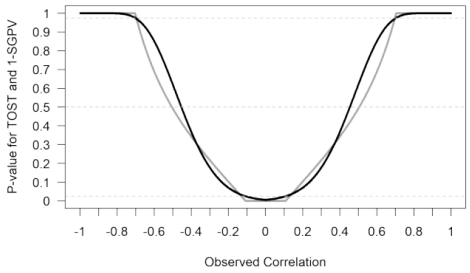

Figure 11. Comparison of *p*-values from TOST (black line) and 1-SGPV (grey curve) across a range of observed sample correlations (x-axis) tested against equivalence bounds of r = -0.45 and r = 0.45 with n = 30 and an alpha of 0.05.

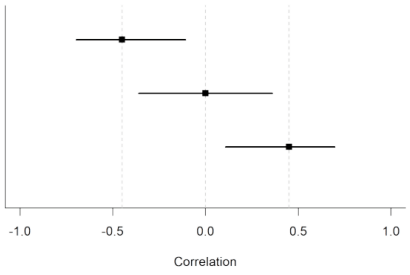

Figure 12. Three 95% confidence intervals for observed effect sizes of r = -0.45, r = 0, and r = 0.45 for n = 30. Only the confidence interval for r = 0 is symmetric.

than the true correlation. As can be seen in Figure 12, the proportion of the confidence interval that overlaps with the equivalence range is larger than 50% when the observed correlations are r = -.45 and r = .45, meaning that the two second generation p-values associated with these correlations are larger than 50%. Because the confidence intervals are asymmetric around the observed effect size of 0.45 (ranging from 0.11 to 0.70) according to Blume et al. (2018) 58.11% of the data-supported hypotheses are null hypotheses, and therefore 58.11% of the data-supported hypotheses are compatible with the null premise.

The further away from 0, the larger the SGPV when the observed mean falls on the equivalence bound. The SGPV is the proportion of values in a 95% confidence interval that overlap with the equivalence range, but not the probability that these values will be observed. In the most extreme case (i.e., a sample size of 4, and equivalence bounds set to r = -0.99 and 0.99, with a true correlation of 0.99) 97.60% of the confidence interval overlaps with the equivalence range, even though in the long run only 36% of the correlations observed in the future will fall in this range.

It should be noted that in larger sample sizes the SGPV is closer to 0.5 whenever the observed correlation falls on the equivalence bound, but this extreme example nevertheless clearly illustrates the difference between question the SGPV answers, and the question a *p*-value answers. The conclusion of this section on asymmetric confidence intervals is that a SGPV of 1 or 0 can still be interpreted as a $p < 0.025$ or $p > 0.975$ in an equivalence test, since the SGPV and *p*-value for the TOST procedure are always directly related at the values $p = 0.025$ and $p = 0.975$. Although Blume et al. (2018) state that "the degree of overlap conveys how compatible the data are with the null premise" this definition of what the SGPV provides does not hold for asymmetric confidence intervals. Although a SGPV of 1 or 0 can be directly interpreted, a SGPV between 0 and 1 is not interpretable as "compatibility with the null hypothesis" under the assumption of a bivariate normal distribution, and the generalizability of this statement needs to be examined beyond normal bivariate distributions. Indeed, Blume and colleagues write in the supplemental material that "The magnitude of an inconclusive second-generation *p*-value can vary slightly when the effect size scale is transformed. However definitive findings, i.e. a *p*-value of 0 or 1 are *not* affected by the scale changes."

What are the Relative Strengths and Weaknesses of Equivalence Testing and the SGPV?

When introducing a new statistical method, it is important to compare it to existing approaches and specify its relative strengths and weaknesses. Here, we aimed to compare the SGPV against equivalence tests based on the TOST procedure. First of all, even though a SGPV of 1 or 0 has a clear interpretation (we can reject effects outside or inside the equivalence range), intermediate values are not as easy to interpret (especially for effects that have asymmetric confidence intervals). In one sense, they are what they are (the proportion of overlap), but it can be unclear what this number tells us about the data we have collected. This is not too problematic, since the main use of the SGPV (e.g., in all examples provided by Blume and colleagues) seems to be to examine whether the SGPV is 0, 1, or inconclusive.

As already mentioned, this interpretation of a SGPV is very similar to the Neyman-Pearson interpretation of an equivalence test and a minimum effect tests (which are complementary). The difference is that where a SGPV of 1 can be interpreted as $p < .025$, equivalence tests provide exact *p*-values, and they continue to differentiate between for example $p = 0.024$ and $p = 0.002$. Whether this is desirable depends on the perspective that is used. From a Neyman-Pearson perspective on statistical inferences the main conclusion is based on whether or not $p < \alpha$, and thus an equivalence test and SGPV can be performed by simply checking whether the confidence interval falls within the equivalence range, just as a null-hypothesis test can be performed by checking whether the confidence interval contains zero or not. At the same time, it is recommended to report exact *p*-values (American Psychological Association, 2010), and exact *p*-values might provide information of interest to readers about how precisely how surprising the data, or more extreme data, is under the null model. Some researchers might be interested in combining an equivalence test with a null-hypothesis significance test. This allows a researcher to ask whether there is an effect that is statistically different from zero, and whether effect sizes that are considered meaningful can be rejected. Equivalence tests combined with null-hypothesis tests classify results into four possible categories, and for example allow researchers to conclude an effect is significant *and* equivalent (i.e., statistically different from zero, but also too small to be considered meaningful; see Lakens et al., 2018).

An important issue when calculating the SGPV is its reliance on the "small sample correction", where the SGPV is set to 0.5 whenever the ratio of the confidence interval width to the equivalence range exceeds 2:1 and the CI overlaps with the upper and lower bounds. This exception to the normal calculation of the SGPV is introduced to prevent misleading values. Without this correction it is possible that a confidence interval is extremely wide, and an equivalence range is extremely narrow, which without the correction would lead to a very low value for the SGPV. Blume et al. (2018) suggest that under such a scenario "the data favor alternative hypotheses", even when a better interpretation would be that there is not enough data to accurately estimate the true effect compared to the width of the equivalence range. Although it is necessary to set the SGPV to 0.5 whenever the ratio of the confidence interval width to the equivalence range exceeds 2:1, it leads to a range of situations where the SGPV is set to 0.5, while the *p*-value from the TOST procedure continues to differentiate (see for example Figure 6). An important benefit of equivalence tests is that it does not need such a correction to prevent misleading results.

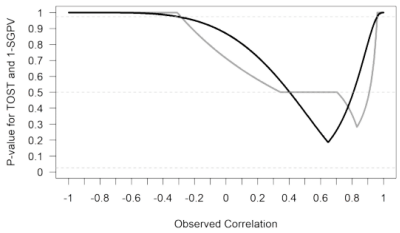

Figure 13. Comparison of *p*-values from TOST (black line) and 1-SGPV (grey curve) across a range of observed sample correlations (x-axis) tested against equivalence bounds of $r = 0.4$ and $r = 0.8$ with n = 10 and an alpha of 0.05.

As a more extreme example of the peculiar behavior of the "small sample correction" as currently implemented in the calculation of the SGPV, see Figure 13. In this figure observed correlations (from a sample size of 10) from -.99 to .99 are tested against an equivalence range from r = 0.4 to r = 0.8. We can see the SGPV has a peculiar shape because it is set to 0.5 for certain observed correlations, even though there is no risk of a "misleading" SGPV in this range. This example suggests that the current implementation of the "small sample correction" could be improved. If, on the other hand, the SGPV is mainly meant to be interpreted when it is 0 or 1, it might be preferable to simply never apply the "small sample correction".

Blume et al. (2018) claim that when using the SGPV "Adjustments for multiple comparisons are obviated" (p. 15). However, this is not correct. Given the direct relationship between TOST and SGPV highlighted in this manuscript (where a TOST $p = 0.025$ equals SGPV = 1, as long as the SGPV is calculated based on confidence intervals, and assuming data are sampled from a continuous bivariate normal distribution), not correcting for multiple comparisons will inflate the probability of concluding the absence of a meaningful effect based on the SGPV in exactly the same way as it will for equivalence tests. Whenever statistical tests are interpreted as support for a hypothesis (e.g., SPGV = 0 or SGPV = 1), it is possible to do so erroneously, and if researchers want to control error rates, they need to correct for multiple comparisons.

10

Conclusion

We believe that our explanation of the similarities between the TOST procedure and the SGPV provides context to interpret the contribution of second generation *p*-values to the statistical toolbox. The novelty of the SGPV can be limited when confidence intervals are asymmetrical or wider than the equivalence range. There are strong similarities with *p*-values from the TOST procedure, and in all situations where the statistics yield different results, the behavior of the *p*-value from the TOST procedure is more consistent. We hope this overview of the relationship between the SGPV and equivalence tests will help researchers to make an informed decision about which statistical approach provides the best answer to their question. Our comparisons show that when proposing alternatives to null-hypothesis tests, it is important to compare new proposals to already existing procedures. We believe equivalence tests achieve the goals of the second generation *p*-value while allowing users to easily control error rates, and while yielding more consistent statistical outcomes.

Authors Note

All code associated with this article, including the reproducible manuscript, is available from https://github.com/Lakens/TOST_vs_SGPV and https://osf.io/8crkg/. The preprint can be found at https://psyarxiv.com/7k6ay/.

Correspondence concerning this article should be addressed to Daniël Lakens, Den Dolech 1, IPO 1.33, 5600 MB, Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl

Open Science Practices

This article earned the Open Materials badge for making the materials openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, are published in the online supplement.

Conflict of Interest and Funding

No conflict of interest and no external funding. This work was supported by the NetherlandsOrganization for Scientific Research (NWO) VIDI grant 452-17-013.

Author Contributions

DL conceptualized the idea, both authors wrote and revised this manuscript.

References

American Psychological Association (Ed.). (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Blume, J. D., D'Agostino McGowan, L., Dupont, W. D., & Greevy, R. A. (2018). Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses. *PLOS ONE*, *13*(3), e0188299. doi:10.1371/journal.pone.0188299

Goertzen, J. R., & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 527–537. doi:10.1348/000711009X475853

Hauck, D. W. W., & Anderson, S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics*, *12*(1), 83–91. doi:10.1007/BF01063612

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. doi:10.1177/2515245918771304

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. doi:10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. doi:10.1177/2515245918770963

Meyners, M. (2012). Equivalence tests A review. *Food Quality and Preference*, *26*(2), 231–245. doi:10.1016/j.foodqual.2012.05.003

Murphy, K. R., Myors, B., & Wolach, A. H. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (Fourth edition.). New York: Routledge, Taylor & Francis Group.

Quertemont, E. (2011). How to Statistically Show the Absence of an Effect. *Psychologica Belgica*, *51*(2), 109–127. doi:10.5334/pb-51-2-109

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565. doi:http://dx.doi.org/10.1037/0033-2909.113.3.553

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. doi:https://doi.org/10.1007/BF01068419

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*(1), 73–83. doi:http://dx.doi.org/10.1037/0003-066X.40.1.73

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 357–416. doi:10.2307/2983527

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Boca Raton: CRC Press.

Westlake, W. J. (1972). Use of Confidence Intervals in Analysis of Comparative Bioavailability Trials. *Journal of Pharmaceutical Sciences*, *61*(8), 1340–1341. doi:10.1002/JPS.2600610845

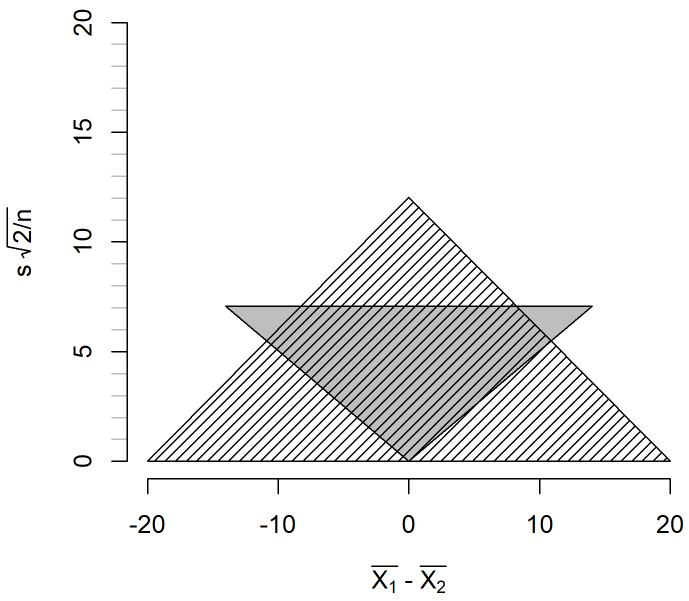

Figure 1. Région d'équivalence pour l'approche de la puissance (zone grisée) et pour le TOST (zone hachurée), pour l'exemple où $|\theta|$=20, n = 50 et $\alpha = .05$

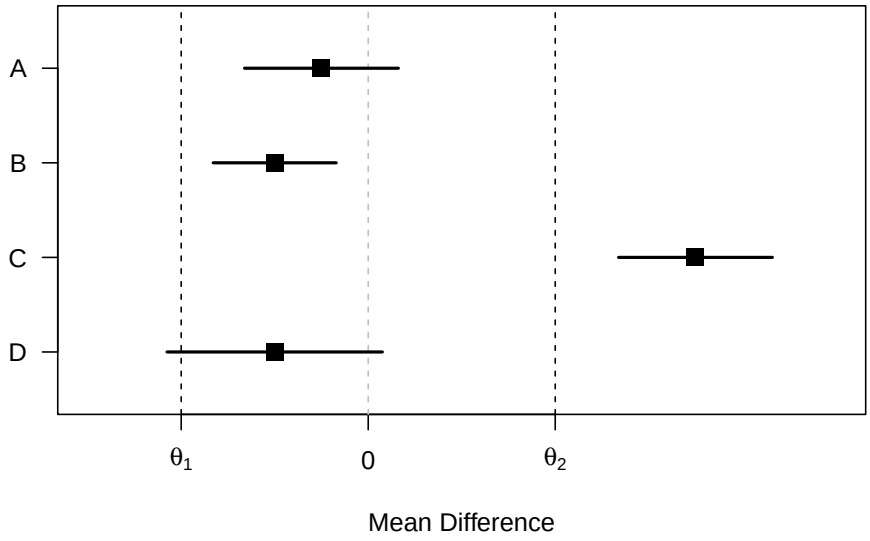

Figure 2. Différence de moyennes $(\bar{X}_1 - \bar{X}_2)$ et IC à $1 - 2\alpha\%$ autour de la différence de moyennes $(\bar{X}_1 - \bar{X}_2)$ pour 4 scénarios distincts.