Commentaire

Le calcul de puissance en assignation aléatoire : un point de vue de statisticien

Hélène Erkel-Rousse

Dans Économie & prévision 2014/1-2 n° 204-205 , pages 145 à 161
Éditions Direction générale du Trésor

ISSN 0249-4744
DOI 10.3917/ecop.204.0145
Date de mise en ligne : 14/01/2016

Article disponible en ligne à l'adresse
https://shs.cairn.info/revue-economie-et-prevision-2014-1-page-145?lang=fr

CAIRN.INFO

Découvrir le sommaire de ce numéro, suivre la revue par email, s'abonner...
Scannez ce QR Code pour accéder à la page de ce numéro sur Cairn.info.

Distribution électronique Cairn.info pour Direction générale du Trésor.
Vous avez l'autorisation de reproduire cet article dans les limites des conditions d'utilisation de Cairn.info ou, le cas échéant, des conditions générales de la licence souscrite par votre établissement. Détails et conditions sur cairn.info/copyright.
Sauf dispositions légales contraires, les usages numériques à des fins pédagogiques des présentes ressources sont soumises à l'autorisation de l'Éditeur ou, le cas échéant, de l'organisme de gestion collective habilité à cet effet. Il en est ainsi notamment en France avec le CFC qui est l'organisme agréé en la matière.

Commentaire

Le calcul de puissance en assignation aléatoire : un point de vue de statisticien

Hélène Erkel-Rousse(*)

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

Ce commentaire résulte des échanges passionnants que j'ai eus avec Adrien Bouguen et Juliette Seban (dans la suite BS) autour de leur article "L'assignation aléatoire comme méthode d'évaluation des politiques publiques". Cet article me paraît constituer une introduction très utile et très pédagogique à l'évaluation par la méthode d'assignation aléatoire. En mettant bien en évidence les avantages de cette méthode, tout en en indiquant aussi les risques et les moyens de s'en prémunir, les auteurs transmettent des informations précieuses à tout économiste amené à réaliser sa première évaluation ou à utiliser de manière éclairée le travail d'évaluation d'autrui. Leur article est pour BS l'occasion de faire bénéficier le lecteur de leur expérience d'évaluateurs, tant théorique qu'opérationnelle, ce qui est particulièrement appréciable. Ainsi, l'article est bien équilibré entre présentations techniques et exemples illustratifs tirés d'évaluations réelles. En outre, il développe certains points sur lesquels la littérature antérieure passe souvent très rapidement, au risque d'être mal comprise par le lecteur non familier des notions utilisées. Pour autant, quelques interrogations peuvent subsister, qui méritent une clarification. Le présent commentaire apporte sur ces questions un regard de statisticien, complémentaire de celui de l'évaluateur, en espérant que ce soit utile au lecteur.

Parmi les nombreuses discussions que j'ai eues avec les auteurs, les principales sont nées d'interrogations qui me sont venues à l'esprit en lisant la partie de l'article consacrée au calcul de puissance, ainsi que les passages d'ouvrages de référence abordant cette question. C'est sur ce sujet que porte plus spécifiquement ce commentaire. Comme le montre très bien l'article de BS, le calcul de puissance revêt une très grande importance pratique. En effet, il conditionne largement la capacité d'un protocole expérimental à évaluer de manière satisfaisante l'effet d'une mesure de politique économique, en guidant l'évaluateur sur les choix des caractéristiques techniques du protocole les mieux adaptées (taille de l'échantillon, type de tirage des individus pour constituer l'échantillon, proportion d'individus traités, etc.). Dès lors, il est important de bien comprendre les raisons et la portée des options méthodologiques les plus fréquemment retenues, que décrit l'article de BS.

(*) Direction générale du Trésor.
E-mail : helene.erkel-rousse@dgtresor.gouv.fr

Je remercie Adrien Bouguen et Juliette Seban pour les échanges très intéressants que nous avons eus autour des questions liées au calcul de puissance en assignation aléatoire, ainsi que Bruno Crépon, pour notre discussion très éclairante à ce sujet. Je remercie aussi Sylvain Chabé-Ferret pour ses conseils et suggestions avisés, ainsi que pour sa lecture attentive d'une version antérieure de ce commentaire. Merci, enfin, à Lucile Romanello pour ses commentaires et son soutien. Je reste toutefois seule responsable des imperfections de ce texte.

Ce commentaire n'engage que son auteure et non les institutions auxquelles elle appartient. Il n'engage *a fortiori* ni la Direction générale du Trésor, ni le ministère des Finances et des Comptes publics, ni le ministère de l'Économie, de l'Industrie et du Numérique.

La question à laquelle je consacre la majeure partie de ce commentaire est celle du type de test à utiliser pour procéder au calcul de puissance. BS évoquent brièvement cette question dans l'annexe 2 de leur article. La pratique usuelle privilégie le test dit bilatéral, au moyen duquel on vérifie si la politique à évaluer a un effet ou pas. Certains articles recommandent plutôt l'utilisation de tests unilatéraux mais définissent ces derniers sous une forme réductrice, ce qui en limite probablement l'usage. Après avoir posé le cadre de raisonnement, la première partie de mon commentaire expose les différents types de tests d'hypothèses auxquels on peut penser *a priori* pour fonder un calcul de puissance. J'explique pourquoi, intuitivement, l'usage préférentiel d'un test plutôt qu'un autre devrait renvoyer à la question la plus adaptée au problème posé. À la lumière de cette première ébauche de discussion, il ne paraît pas exclu que l'usage privilégie un peu trop souvent le test bilatéral de la nullité de l'effet contre sa significativité au détriment d'autres tests envisageables.

Toutefois, de simples raisonnements intuitifs ne suffisent pas à justifier l'usage préférentiel de tel ou tel test. Il est important de bien maîtriser les conséquences pratiques du fait d'opter pour un test ou pour un autre en termes de choix des caractéristiques du protocole expérimental. Pour ce faire, la deuxième partie détaille le calcul de puissance pour chacun des tests envisagés dans la première partie. J'y montre en particulier que deux des arguments fréquemment utilisés pour écarter *a priori* les tests unilatéraux ne sont pas convaincants.

La troisième partie réexamine la question du choix du test à privilégier pour le calcul de puissance, en la relativisant très notablement au regard des enseignements de la deuxième partie. Ceux-ci suggèrent en effet que le choix de tel ou tel test est finalement assez secondaire, chaque test envisagé permettant d'aboutir à un calcul d'effet minimum détectable susceptible de répondre aux questions posées par l'ensemble de ces tests, à condition d'en maîtriser les propriétés. Cette partie est l'occasion de revenir sur un troisième argument fréquemment avancé pour justifier la pratique courante, mais qui ne me semble pas résister à l'examen. En revanche, celle-ci, au même titre que d'autres pratiques, ressort justifiée de cette analyse, mais pas pour les raisons pour lesquelles on la recommande habituellement.

La quatrième et dernière partie aborde une autre question relative au calcul de puissance. Il s'agit de rappeler l'importance pour ce calcul d'une bonne formulation de l'écart-type de l'estimateur de l'effet de la politique à évaluer (noté $\sigma_{\hat{\delta}}$ dans l'article de BS). Tout d'abord, une illustration de la manière dont est utilisé le calcul de puissance pour orienter le choix des caractéristiques du protocole d'expérimentation met bien en évidence le fait que c'est exclusivement à travers les composantes de cet écart-type que l'évaluateur peut peser sur l'effet minimum détectable *via* les caractéristiques du protocole expérimental et, donc, améliorer la puissance de l'évaluation par ce moyen. Or, l'expression de cet écart-type est conditionnée notamment par les hypothèses portant sur la matrice de variances-covariances des termes d'erreurs du modèle d'évaluation. L'hypothèse de termes d'erreurs indépendants et identiquement distribués (*i.i.d.*) est souvent retenue dans les calculs de puissance présentés dans la littérature. L'article de BS ne fait pas exception à cet égard. Toutefois, il donne un exemple éclairant et très utile de la manière dont la formulation de l'écart-type $\sigma_{\hat{\delta}}$ est affectée lorsqu'une des caractéristiques de l'hypothèse *i.i.d.* est levée, à savoir l'indépendance des observations, en raison du choix d'un mode de tirage par grappes plutôt qu'au niveau individuel. Cet exemple, de même que les autres développements de l'article sur le sujet, illustrent la nécessité de porter une grande attention en amont du calcul de puissance aux hypothèses sous-jacentes à ce calcul. Enfin, j'évoque la difficile estimation de la matrice de variances-covariances des termes d'erreur du modèle d'évaluation ainsi que d'autres paramètres nuisibles intervenant dans l'expression de $\sigma_{\hat{\delta}}$. Puis je conclus.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

Sur quel test d'hypothèses fonder un calcul de puissance préparatoire à l'évaluation d'une politique : ce que suggère l'intuition

Mettons-nous dans la situation de l'évaluateur s'apprêtant à estimer l'effet d'une mesure de politique économique au moyen de la méthode d'assignation aléatoire. En amont, l'évaluateur réalise un calcul de puissance destiné à orienter ses choix concernant les principales caractéristiques du protocole expérimental à mettre en place, de manière à lui donner un maximum de chances de détecter correctement l'effet de la politique à évaluer. Au moment où il réalise ce calcul de puissance, l'évaluateur n'a donc encore à sa disposition ni échantillon, ni *a fortiori* un groupe test et un groupe témoin. En revanche, il dispose d'un cadre d'analyse lui permettant d'anticiper les procédures statistiques (estimations, tests) qu'il mettra en œuvre pour évaluer l'effet de la politique. En se positionnant dans ce cadre, il procède *ex ante* aux calculs théoriques qui seront à la base de son évaluation chiffrée de la politique lorsque les données nécessaires seront disponibles. Ce faisant, il identifie les caractéristiques du protocole expérimental qui conditionneront la performance statistique future de son évaluation. Cela lui permet de les calibrer au mieux dans la perspective de cette dernière.

Le cadre d'analyse

Je me place dans un cadre d'analyse proche de celui de BS. Le modèle d'évaluation retenu est :

$$(1)\quad Y_i = c + \delta T_i + \varepsilon_i$$

où Y_i est l'indicateur de résultat pour l'individu i et T_i son groupe d'assignation (T_i est égal à 1 si l'individu i est affecté au groupe test et à 0 s'il est affecté au groupe témoin). c est une constante, δ le paramètre d'intérêt (l'effet moyen de la mesure) à estimer et ε_i le terme d'erreur, d'espérance nulle. Soit $\hat{\delta}$ l'estimateur des moindres carrés ordinaires (MCO) du paramètre d'intérêt δ et $\sigma_{\hat{\delta}}$ son écart-type conditionnel au traitement. Comme le montrent BS, $\hat{\delta}$ s'exprime comme l'écart $\overline{Y}^1 - \overline{Y}^0$ entre les moyennes empiriques de l'indicateur de résultat sur le groupe test et le groupe témoin. Quant à l'écart-type $\sigma_{\hat{\delta}}$, il dépend des composantes inconnues de la matrice de variances-covariances des termes d'erreur et des caractéristiques (à déterminer) de l'échantillon, d'une manière plus ou moins complexe selon les hypothèses retenues sur la distribution jointe des termes d'erreur $(\varepsilon_i)_{i=1,\ldots,N}$, où N est le nombre total d'observations.

Je me situe dans la situation où l'estimateur $\hat{\delta}$ peut être considéré comme convergent et asymptotiquement normal :

$$(2)\quad (\hat{\delta} - \delta) / \sigma_{\hat{\delta}} \overset{loi}{\rightarrow} N(0,1)$$

Cette propriété est vérifiée sous les hypothèses retenues par BS en début d'annexe 2, d'un tirage effectué au niveau individuel et de termes d'erreur (ε_i) indépendants identiquement distribués (*i.i.d.*). Elle l'est aussi pour des distributions beaucoup plus générales des (ε_i), sous certaines conditions définies par les généralisations successives de la loi des grands nombres et du théorème central limite. Ainsi, la normalité asymptotique est vérifiée pour des (ε_i) indépendants mais de lois non identiques, à condition qu'aucun des termes d'erreur ε_i n'exerce une influence prépondérante sur les autres (ceci étant assuré par exemple par le critère de Lyapounov ou la condition de Lindeberg ou, encore, le théorème de Gnedenko et Kolmogorov[1]). La normalité asymptotique est aussi vérifiée dans des cas de non-indépendance, dès lors que les (ε_i) sont liés par une dépendance "faible". Bien évidemment, selon la forme de la distribution jointe des termes d'erreurs, l'écart-type de $\hat{\delta}$ n'a pas la même expression. C'est un point important que je détaille dans la quatrième partie du commentaire.

À une question posée un test *a priori* adapté

L'objectif du calcul de puissance est de se donner les moyens d'assurer la qualité de l'évaluation future. Il paraît donc essentiel de bien identifier la question à laquelle on va chercher à répondre lors de l'évaluation. Lorsqu'il s'agit d'évaluer une politique publique, ce qui intéresse le décideur au premier chef est : la politique envisagée est-elle efficace ou non au regard du but recherché ? Il peut aussi être prudent de vérifier si la politique ne risque pas d'avoir des effets contreproductifs. Il paraît plus secondaire *a priori* de savoir si la politique aura un effet (quel qu'il soit, sans qu'il soit possible de le caractériser plus avant) ou pas.

Intuitivement, le test d'hypothèses adapté au calcul de puissance est celui qui va permettre de répondre au mieux à la question posée. Dans le paragraphe précédent, trois questions différentes ont été évoquées, qui appellent en toute rigueur trois tests d'hypothèses différents, pouvant être réalisés dans un même cadre de raisonnement, c'est-à-dire pour un même modèle d'évaluation (ici, le modèle (1)).

Le but recherché par la mesure à évaluer guide le choix de la variable de résultat *Y*. La politique vise une diminution de la consommation de tabac d'une certaine population ? *Y* peut être définie en fonction du nombre de cigarettes fumées par personne et par

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

jour au sein de cette population. La politique cherche à augmenter le taux d'emploi d'une certaine population ? Une variable de résultat naturelle est ce taux d'emploi. Pour plus de clarté dans la suite du raisonnement, je suppose sans perte de généralité que l'indicateur de résultat est défini de manière à ce qu'une hausse de *Y* soit vue comme un résultat positif, recherché par la politique à évaluer. Dans l'exemple de la politique visant à augmenter le taux d'emploi d'une certaine population, la variable *Y* peut être définie comme ce taux d'emploi. Dans celui de la mesure visant une baisse de la consommation de tabac, *Y* est par exemple définie comme une fonction décroissante de la consommation journalière de tabac par tête. Cette convention permet d'interpréter systématiquement le signe du paramètre d'intérêt dans le même sens. Un coefficient δ positif correspond à une mesure allant dans la direction recherchée, tandis qu'un coefficient négatif va dans le sens opposé à l'objectif visé. Dès lors, il devient possible de traduire sous forme de tests d'hypothèses les trois questions évoquées plus haut.

Le test bilatéral (*b*0) de l'hypothèse nulle $H_0 : \delta = 0$ contre l'hypothèse alternative $H_a : \delta \neq 0$ utilisé par BS pour leurs calculs de puissance en annexe 2 et dans le corps du texte (application du calcul de puissance au projet "Groupements de Créateurs") est un test qui vise à répondre à la question : le programme a-t-il un effet ($\delta \neq 0$) ou pas ($\delta = 0$) ? Intuitivement, il est plus intéressant de savoir si la mesure a un impact strictement positif ($\delta > 0$) ou, au contraire, une incidence nulle ou contreproductive ($\delta \leq 0$), ce qui appelle un test unilatéral (*ud*0) de l'hypothèse nulle $H_0 : \delta \leq 0$ contre l'hypothèse alternative $H_a : \delta > 0$. Ce test est appelé test unilatéral à droite, par référence à la position du paramètre d'intérêt par rapport à zéro sous l'hypothèse alternative. Enfin, si l'on craint que le traitement puisse avoir un effet nocif ($\delta < 0$) et qu'on cherche à le vérifier, on peut être amené à procéder à un test unilatéral à gauche (*ug*0), de l'hypothèse nulle $H_0 : \delta \geq 0$ contre l'hypothèse alternative $H_a : \delta < 0$.

Notons que la réalisation de ces trois tests ne permet pas forcément de répondre à l'ensemble des questions que peut se poser le décideur à l'origine de la mesure à évaluer. Tout dépend de l'objectif recherché et de la nature de la mesure. Par exemple, il se peut qu'une mesure soit considérée comme efficace et, donc, digne d'être mise en œuvre si l'effet produit est supérieur à un impact minimal acceptable $\delta_0 > 0$ compte tenu du coût de la mesure. Dans ce cas, le test qui permettrait de répondre au questionnement du décideur et à lui donner l'information dont il a besoin pour juger s'il vaut mieux mettre en place la mesure envisagée ou pas serait un test unilatéral à droite (*ud*δ_0) de l'hypothèse nulle $H_0 : \delta \leq \delta_0$ contre l'hypothèse alternative $H_a : \delta > \delta_0$.

Pour résumer, cette première étape de discussion suggère qu'un calcul de puissance devrait non pas reposer automatiquement sur un test bilatéral (*b*0) de l'hypothèse nulle $H_0 : \delta = 0$ contre l'hypothèse alternative $H_a : \delta \neq 0$ mais être précédé par l'identification de la (ou des) question(s) à laquelle (auxquelles) l'évaluation est censée répondre pour que son résultat intéresse le décideur. Ainsi que l'écrit très justement Bloom (1995), « *The main goal of a program impact study should be to determine whether or not a program produced the results it was intended to produce* ». C'est l'une des raisons pour lesquelles Bloom se déclare plus favorable au test unilatéral à droite qu'au test bilatéral : « *One, therefore, should use a one-sided hypothesis test to determine the statistical significance of an impact estimate in this decision-making context* ». Au-delà, non seulement le test adapté au questionnement du commanditaire ne paraît pas être obligatoirement un test bilatéral mais, en outre, l'évaluation de l'efficacité d'une politique peut renvoyer à un impact minimal recherché δ_0, à déterminer en concertation avec le décideur.

L'utilisation prédominante du test bilatéral (*b*0) dans la littérature

Or, il semble que le test bilatéral (*b*0) soit utilisé de manière largement majoritaire, si ce n'est souvent systématique, dans les calculs de puissance. Pour expliquer cette prédominance du test bilatéral (*b*0), Bloom (1995) avance l'idée que cette convention trouve son origine dans la recherche empirique en sciences sociales, dont l'objectif premier est de tester des relations théoriques entre variables en s'intéressant à leur significativité ou non, indépendamment de leur direction. Il regrette cette convention : « *Because program evaluations should be conducted to inform program decisions, evaluators should use the logic of the decision they are attempting to inform rather than the strength of existing theory to formulate their impact hypothesis test* ».

Dans l'annexe 2 de leur article, BS mentionnent les arguments les plus souvent avancés en réponse à la position défendue par Bloom (1995).

Tout d'abord, les tests unilatéraux seraient en pratique moins largement utilisables que le test bilatéral, dans la mesure où, pour y avoir recours, il faudrait disposer d'une information *ex ante* quasi-certaine sur le signe de l'effet de la politique, ce dont on peut rarement être assuré. Cette affirmation repose sur deux arguments :
– le premier argument ne peut se comprendre qu'en remarquant que de nombreux auteurs (dont des partisans des tests unilatéraux comme Bloom, 1995) ne se réfèrent pas au type de test unilatéral tel que défini plus haut mais à un type de test unilatéral apparemment plus restrictif, avec hypothèse nulle

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

ponctuelle, c'est-à-dire, par exemple, au lieu du test (*ud*0), au test (*udp*0) de $H_0 : \delta = 0$ contre $H_a : \delta > 0$ ou, plus rarement, au lieu du test (*ug*0), au test (*ugp*0) de $H_0 : \delta = 0$ contre $H_a : \delta < 0$. Cette formulation des tests unilatéraux suggère qu'ils ne seraient utilisables que lorsque le signe de δ est connu *a priori*, positif dans le cas du test unilatéral à droite (négatif, respectivement, dans celui du test unilatéral à gauche), puisque l'ensemble des possibles défini par l'union de H_0 et de H_a exclut la plage de valeurs négatives (positives, respectivement) pour le paramètre d'intérêt. Toutefois, ce premier argument n'est pas convaincant dans la mesure où la théorie des tests montre qu'il n'y a en réalité aucune obligation technique à se restreindre à des tests avec hypothèse nulle ponctuelle. Certes, il est plus immédiat de définir la forme de la région critique (c'est-à-dire l'ensemble des observations conduisant à un rejet de l'hypothèse nulle) dans le cas de tests d'hypothèse nulle ponctuelle. Néanmoins, il n'existe aucune difficulté à le faire pour des tests dont l'hypothèse nulle est non ponctuelle, dès lors que celle-ci couvre une plage de valeurs pour le paramètre d'intérêt ayant la forme d'un intervalle fermé au niveau de la frontière avec l'hypothèse alternative. Dans les exemples des tests unilatéraux à droite (*ud*0) et à gauche (*ug*0), c'est bien le cas, puisque la frontière avec l'hypothèse alternative (la valeur zéro) appartient à l'intervalle des valeurs du paramètre d'intérêt constituant l'hypothèse nulle. Pour s'en convaincre, le lecteur se reportera à la partie suivante, où sont présentés des calculs de région critique pour les tests (*ud*0) et (*ug*0) ;

– le second argument, qui s'applique tant aux tests avec hypothèse nulle ponctuelle de type (*udp*0) ou (*ugp*0) qu'aux tests de type (*ud*0) ou (*ug*0), renvoie à une difficulté à laquelle l'évaluateur est parfois confronté : ayant prévu une mesure efficace (c'est-à-dire une valeur strictement positive de δ), il obtient *ex post*, lors de l'évaluation proprement dite, une estimation de δ négative sur la base des données qu'il a collectées. Dans ce cas, l'évaluateur n'aurait pas choisi le bon test en optant, lors du calcul de puissance préparatoire, pour le test unilatéral à droite (*ud*0), alors qu'il aurait fallu opter pour le test unilatéral à gauche (*ug*0). Dès lors, certains suggèrent de privilégier, dans la phase amont du calcul de puissance, le test bilatéral, qui ne présage pas de signe probable de δ sous l'hypothèse alternative. La réponse à ce second argument nécessite d'avoir en tête les résultats des calculs de puissance pour les deux types de tests (*cf. infra*).

Un troisième argument à l'appui de l'usage préférentiel du test bilatéral, étroitement lié au second argument, souligne le caractère plus conservateur du test bilatéral. À autres caractéristiques fixées, ce dernier conduirait l'évaluateur à opter pour une taille d'échantillon plus élevée que les tests unilatéraux. Compte tenu des difficultés et nombreux risques d'erreurs attachés au calcul de puissance, ceci constituerait une sécurité appréciable pour la suite de l'évaluation. Là encore, on ne peut bien comprendre et, *a fortiori*, discuter cet argument qu'après avoir détaillé le calcul de puissance pour chacun des tests d'hypothèse envisagés. C'est l'objet de la deuxième partie de ce commentaire. Je commenterai ces arguments à la lumière des résultats des calculs de puissance, au fur et à mesure qu'ils permettront de les discuter, au cours de la deuxième partie (deux premiers arguments) et dans la troisième partie (troisième argument).

Le calcul de puissance pour les trois types de tests envisageables

Je détaille ici le calcul de puissance pour les trois tests (*b*0), (*ud*0) et (*ug*0), en commençant par les tests unilatéraux, en raison de la plus grande simplicité des calculs associés et de leur utilité pour la réalisation et la compréhension du calcul de puissance dans le cas du test bilatéral. J'introduis tout d'abord deux notations commodes pour la suite :

$$(3)\quad \hat{d} = \frac{\hat{\delta}}{\sigma_{\hat{\delta}}} \quad \text{et} \quad d = \frac{\delta}{\sigma_{\hat{\delta}}}$$

L'hypothèse (2) de normalité asymptotique de l'estimateur des MCO de l'effet de la mesure à évaluer se réécrit immédiatement :

$$(4)\quad \hat{d} - d \xrightarrow{loi} N(0,1)$$

Dans la suite, Φ désigne la fonction de répartition de la loi suivie par $\hat{d} - d$, c'est-à-dire la fonction de répartition de la loi normale centrée réduite dans la configuration asymptotique (4) dans laquelle nous nous plaçons (ce qui suppose un nombre d'observations *N* suffisamment grand).

Remarques

1) Dans la théorie des tests appliquée dans un cadre usuel (celui de tests effectués *ex post*, c'est-à-dire une fois la base de données constituée et disponible pour réaliser des analyses statistiques), on veille à remplacer en début de calcul l'écart-type inconnu $\sigma_{\hat{\delta}}$ par un estimateur convergent de $\sigma_{\hat{\delta}}$, afin que la statistique $\hat{d}$ soit observable. Comme l'illustre l'article de BS, dans un calcul de puissance réalisé en amont d'une assignation aléatoire, on effectue le calcul de puissance comme si la variance $\sigma_{\hat{\delta}}$ était connue, quitte à l'estimer en toute fin de calcul. Le point de vue est en effet spécifique. D'une part, en tout état de cause, la statistique $\hat{d}$ n'est pas encore observable, donc il ne s'agit pas, lors du calcul de puissance, de décider laquelle de l'hypothèse nulle et de l'hypothèse alternative retenir (ce qui ne peut se

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

faire que si $\hat{d}$ est mesurée – *cf. infra*). D'autre part, comme l'expliquent BS, on peut s'affranchir de l'estimation de la matrice de variances-covariances du vecteur des termes d'erreur dans les cas où le nombre de paramètres inconnus contenus dans cette matrice se réduit à un seul, σ^2, variance commune des termes d'erreur dans le cas *i.i.d.*, en raisonnant en unités de l'écart-type σ (*cf.* dernière partie du commentaire). Enfin, même si on remplace en toute fin de calcul de puissance les moments du second ordre inconnus de la distribution jointe des termes d'erreur par des estimations, cela ne change rien au calcul qui a précédé (car cela ne modifie pas la loi asymptotique de $\hat{d} - d$, centrale dans les calculs de puissance *infra*) dès lors qu'il s'agit d'estimateurs presque sûrement convergents. Je reviendrai brièvement sur cette question après avoir détaillé les calculs de puissance.

2) Les calculs présentés *infra* couvrent par simple translation les cas de tests d'hypothèses plus généraux, n'excluant pas l'existence de seuils d'efficacité ou de nuisance, à savoir :
– le test unilatéral à droite ($ud\delta_0$) de l'hypothèse nulle $H_0 : \delta \leq \delta_0$ contre l'hypothèse alternative $H_a : \delta > \delta_0$;
– le test unilatéral à gauche ($ug\delta_0$) de l'hypothèse nulle $H_0 : \delta \geq \delta_0$ contre l'hypothèse alternative $H_a : \delta < \delta_0$;
– le test bilatéral ($b\delta_0$) de l'hypothèse nulle $H_0 : \delta = \delta_0$ contre l'hypothèse alternative $H_a : \delta \neq \delta_0$.

Pour cela, il suffit de modifier légèrement les définitions (3) :

$$(3') \quad \hat{d} = \frac{\hat{\delta} - \delta_0}{\sigma_{\hat{\delta}}} \quad \text{et} \quad d = \frac{\delta - \delta_0}{\sigma_{\hat{\delta}}}$$

À la définition de $\hat{d}$ et d près, le calcul de puissance pour le test ($ud\delta_0$) (respectivement ($ug\delta_0$) et ($b\delta_0$)) est identique au calcul de puissance pour le test ($ud0$) (respectivement ($ug0$) et ($b0$)), les effets minimums détectables (EMD) obtenus s'interprétant alors en différence non plus à zéro mais à δ_0.

Au total, les tests d'hypothèses étudiés se reformulent en fonction de d (défini par (3) ou (3'), selon les types de tests) comme :
– test unilatéral à droite ($ud0$) ou ($ud\delta_0$) : test de l'hypothèse nulle $H_0 : d \leq 0$ contre l'hypothèse alternative $H_a : d > 0$;
– test unilatéral à gauche ($ug0$) ou ($ug\delta_0$) : test de l'hypothèse nulle $H_0 : d \geq 0$ contre l'hypothèse alternative $H_a : d < 0$;
– test bilatéral ($b0$) ou ($b\delta_0$) : test de l'hypothèse nulle $H_0 : d = 0$ contre l'hypothèse alternative $H_a : d \neq 0$.

On montrera en outre que les calculs relatifs au test ($udp0$) sont contenus dans ceux du test ($ud0$). Il en est de même pour les calculs correspondants au test ($ugp0$), contenus dans ceux du test ($ug0$). Ainsi, l'étude de trois tests suffit à couvrir l'ensemble des tests auxquels il a été fait référence jusqu'ici.

Test unilatéral à droite (*ud*0)

Il faut tout d'abord définir la région critique du test, c'est-à-dire l'ensemble des observations W conduisant à une estimation du paramètre d'intérêt trop élevée pour que l'hypothèse nulle puisse être acceptée, soit :

$$W = \{\hat{d} > t\}$$

La valeur du seuil t est déterminée de manière à assurer que le risque d'une erreur de type I (c'est-à-dire la probabilité de rejeter à tort l'hypothèse nulle quand celle-ci est vraie) soit borné supérieurement par une valeur α suffisamment petite, appelée le niveau du test :

$$\alpha = \max_{d \leq 0} [\Pr(\hat{d} - d > t - d)] = \max_{d \leq 0} [1 - \Phi(t - d)]$$

Comme $d \leq 0$ et Φ est une fonction croissante, la valeur maximale est obtenue pour $d = 0$, c'est-à-dire à la frontière supérieure de l'hypothèse nulle. Ce résultat est conforme à l'intuition : plus l'effet de la mesure est proche de 0, la frontière avec l'hypothèse alternative, plus le risque de rejeter à tort l'hypothèse nulle est élevé. Le seuil t est donc déterminé par :

$$\alpha = 1 - \Phi(t) \Leftrightarrow \Phi(t) = 1 - \alpha \Leftrightarrow t = t_{1-\alpha}$$

où $t_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de fonction de répartition Φ. Les valeurs usuelles choisies pour le niveau de test α vont de 1 % à 10 %, selon le prix accordé à la limitation du risque d'erreur de type I. Ainsi, par exemple, t vaut approximativement 1,64 pour $\alpha = 5$ % et 1,96 pour $\alpha = 2{,}5$ %.

On obtient donc comme région critique du test ($ud0$) de niveau α :

$$W = \{\hat{d} > t_{1-\alpha}\}$$

Ce calcul montre qu'il n'y a aucune raison de se priver de la formulation la plus générale du test unilatéral ($ud0$). En effet, comme on vient de le voir, la formulation ($ud0$) n'implique aucune difficulté de traitement[(2)]. En conséquence, le premier argument appuyant l'idée que l'utilisation d'un test unilatéral ne se justifie que si l'on a une très forte présomption *a priori* sur le signe de l'effet de la mesure (*cf. supra*) ne tient pas. En effet, dans le cas du test ($ud0$), l'union des hypothèses nulle et alternative englobe

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

l'ensemble des nombres réels, soit l'ensemble des valeurs possibles pour l'effet de la mesure, au même titre que dans le cas du test bilatéral (*b*0). Ces deux tests ne se distinguent donc pas par le degré de généralité ou de contrainte des valeurs des effets qu'ils envisagent *a priori*.

Passons maintenant au calcul de puissance. Par définition, la puissance Π d'un test est la probabilité de rejeter H_0 lorsque celle-ci est effectivement fausse (c'est-à-dire, ici, lorsque l'effet du traitement δ est strictement positif). La loi normale $N(0,1)$ étant symétrique par rapport à zéro, la puissance du test unilatéral à droite s'écrit :

$$\Pi = \Pr_{d>0}(\hat{d} - d > t_{1-\alpha} - d) = \Phi(d - t_{1-\alpha})$$

Pour que la puissance du test soit suffisante (au moins égale à une valeur fixée π jugée assez élevée), la fonction de répartition Φ étant croissante, il faut que :

$$\Phi(d - t_{1-\alpha}) \geq \pi \Leftrightarrow d - t_{1-\alpha} \geq t_\pi \Leftrightarrow d \geq t_\pi + t_{1-\alpha}$$

où t_π est le quantile d'ordre π de la loi de fonction de répartition Φ. En d'autres termes, il faut que l'effet du traitement δ vérifie :

$$\delta \geq (t_\pi + t_{1-\alpha})\sigma_{\hat{\delta}}$$

Pour des valeurs usuelles de α et π, $(t_\pi + t_{1-\alpha})$ est strictement positif. L'effet minimum détectable (EMD), c'est-à-dire l'ampleur minimale de l'effet du traitement nécessaire à sa détection par un test de niveau de significativité α et de puissance π, vaut donc, pour le test (*ud*0)[3] :

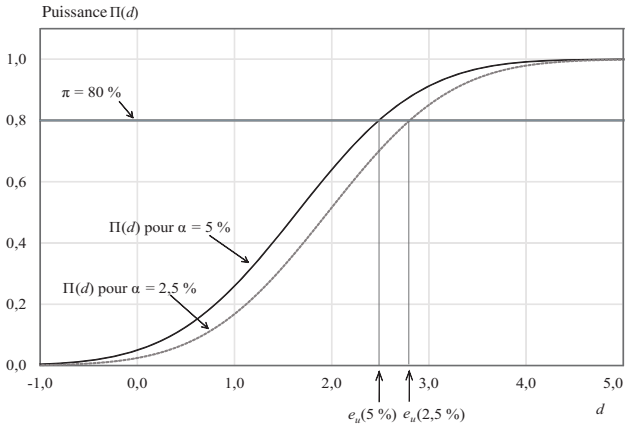
$$(5)\ EMD = (t_\pi + t_{1-\alpha})\sigma_{\hat{\delta}} > 0$$

Notons :

$$e_u = EMD / \sigma_{\hat{\delta}} = t_\pi + t_{1-\alpha}$$

Si l'on choisit $\alpha = 5\,\%$ et $\pi = 80\,\%$, pour une taille de l'échantillon suffisamment grande, les valeurs des quantiles sont $t_{1-\alpha} \approx 1{,}64$ et $t_\pi \approx 0{,}84$. On a alors $e_u \approx 2{,}49$. Pour un test de niveau $\alpha = 2{,}5\,\%$ (plus conservateur que le précédent, au sens où l'on se prémunit davantage contre l'abandon à tort de l'hypothèse nulle qu'en optant pour $\alpha = 5\,\%$) et, toujours, $\pi = 80\,\%$, $t_{1-\alpha} \approx 1{,}96$ et $e_u \approx 2{,}80$. Pour un test de niveau $\alpha = 1\,\%$ encore plus conservateur, à $\pi = 80\,\%$ inchangé, alors $t_{1-\alpha} \approx 2{,}33$ et $e_u \approx 3{,}17$. On retrouve l'arbitrage classique entre risque d'erreur de type I (probabilité de rejeter l'hypothèse nulle à tort) et risque d'erreur de type II (probabilité d'accepter l'hypothèse nulle à tort = 1 - Π) : à un test associé à un risque d'erreur de type I plus élevé

Graphique 1 : calcul de l'EMD dans le cas d'un test unilatéral à droite



Lecture : les valeurs de *d* supérieures à e_u sont détectables par le test unilatéral à droite. Plus le test est conservateur (c'est-à-dire plus le risque d'erreur de type I est limité ou, en d'autres termes, plus α est petit), plus la puissance Π est faible, donc plus e_u (et, avec lui, l'EMD) est élevé, toutes choses égales par ailleurs. Ce graphique vaut tant pour le test (*ud*0) que pour le test (*ud*δ_0) dès lors que la définition de *d* est adaptée au test en question : définition (3) pour le test (*ud*0) et (3') pour le test (*ud*δ_0). Il vaut aussi pour le test (*udp*0) (resp. (*udp*δ_0)) avec la définition (3) (resp. (3')) de *d*.

Source : auteure.

($\alpha = 5\%$) est associé un EMD plus proche de zéro ou, de manière équivalente, une puissance plus élevée (graphique 1).

Test unilatéral à gauche (*ug*0)

En adaptant le raisonnement précédent au cas d'un test unilatéral à gauche et en utilisant les mêmes notations, on obtient comme région critique *W* du test au niveau α :

$$W = \{\hat{d} < t_\alpha\}$$

où t_α est le quantile d'ordre α de la loi de fonction de répartition Φ, égal à l'opposé de $t_{1-\alpha}$ en raison de la symétrie de la loi normale[4].

La puissance Π du test (*ug*0) est égale à :

$$\Pi = \Pr_{d<0}(\hat{d} - d < -d - t_{1-\alpha}) = \Phi(-d - t_{1-\alpha})$$

Pour que la puissance du test soit au moins égale à π, il faut que :

$$\Phi(-d - t_{1-\alpha}) \geq \pi \Leftrightarrow d \leq -(t_\pi + t_{1-\alpha})$$

$$\Leftrightarrow \delta \leq -(t_\pi + t_{1-\alpha})\sigma_{\hat{\delta}}$$

Pour des valeurs usuelles de α et π, $-(t_\pi - t_{1-\alpha})$ est strictement négatif. On définit alors l'EMD comme l'ampleur minimale en valeur absolue de l'effet du traitement nécessaire à sa détection par un test de

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

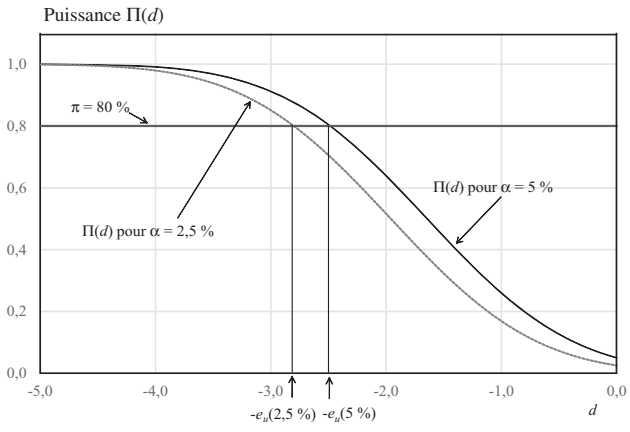
niveau de significativité α et de puissance π. Pour le test considéré, l'EMD vaut donc :

(6) $EMD = (t_{\pi} + t_{1-\alpha})\sigma_{\hat{\delta}} = e_u \sigma_{\hat{\delta}}$

En comparant (6) à (5), on voit que l'EMD du test unilatéral à gauche (*ug*0) est égal à l'EMD du test unilatéral à droite (*ud*0). Dans le cas du test (*ug*0), toutefois, la lecture de l'EMD est différente puisque ce sont les effets négatifs du traitement dont l'ampleur en valeur absolue est supérieure à l'EMD qui seront détectés (ces effets se situent dans la partie du graphique 2 située à gauche de $-e_u$).

Nous disposons maintenant des éléments nécessaires à la discussion du deuxième argument à l'encontre du test unilatéral, selon lequel mieux vaudrait procéder au calcul de puissance sur la base d'un test bilatéral du fait que l'hypothèse alternative de ce dernier ne fait aucune supposition sur le signe de l'effet de la mesure. Rappelons que cet argument est fondé sur la crainte de choisir à tort, lors du calcul de puissance, le test unilatéral à droite (qui permet de vérifier si l'effet δ est strictement positif ou non) puis, *ex post*, d'obtenir une estimation négative de δ. Cette crainte serait parfaitement justifiée si le besoin *ex post* de tester (en utilisant (*ug*0)) la négativité stricte de δ avec une puissance suffisante nécessitait de disposer de plus d'observations et, plus généralement, de caractéristiques de l'échantillon plus exigeantes que ce qui aurait été décidé lors du calcul de puissance sur la base du test (*ud*0). Les égalités (5) et (6) montrent que cette crainte peut être totalement levée dans la mesure où les tests unilatéraux à gauche et à droite aboutissent à la même valeur d'EMD, à α, π et $\sigma_{\hat{\delta}}$ donnés. Ainsi, le recours au test unilatéral à droite (*ud*0) lors du calcul de puissance aboutit à des décisions en termes de caractéristiques de l'échantillon identiques à celles auxquelles aurait abouti le recours au test unilatéral à gauche (*ug*0). L'évaluateur peut donc tester *ex post* avec une puissance adéquate l'inégalité stricte opposée de l'hypothèse alternative du test unilatéral choisi lors du calcul de puissance. Il ne me reste plus qu'à discuter le troisième argument mentionné à la fin de la première partie de ce commentaire. Ceci nécessite d'avoir réalisé le calcul de puissance pour le test bilatéral (*b*0).

Graphique 2 : calcul de l'EMD dans le cas d'un test unilatéral à gauche



Lecture : les valeurs de *d* inférieures à l'opposé de e_u sont détectables par le test unilatéral à gauche. Plus le test est conservateur (c'est-à-dire plus α est petit), plus la puissance Π est faible, donc plus e_u (et par conséquent plus l'EMD) est éloigné de zéro toutes choses égales par ailleurs. Ce graphique vaut tant pour le test (*ug*0) que pour le test ($ug\delta_0$) dès lors que la définition de *d* est adaptée au test en question : définition (3) pour le test (*ug*0) et (3') pour le test ($ug\delta_0$). Il vaut aussi pour le test (*ugp*0) (resp. ($ugp\delta_0$)) avec la définition (3) (resp. (3')) de *d*.

Test bilatéral (*b*0)

L'article de BS a le grand mérite d'expliciter davantage le calcul de puissance dans le cas du test bilatéral que les documents de référence antérieurs, notamment Duflo *et alii* (2006), dont la présentation (page 29), relativement elliptique, risque d'être mal interprétée. Toutefois, il n'est pas inutile de développer un peu plus ce cas pour bien comprendre comment il est lié aux deux cas précédents.

Toujours par le même raisonnement que *supra*, on détermine aisément la région critique du test bilatéral au niveau α :

$$W = \{|\hat{d}| > t_{1-\alpha/2}\}$$

où $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de fonction de répartition Φ. La région critique du test bilatéral (*b*0) de niveau α est donc l'union des régions critiques du test unilatéral à droite (*ud*0) et du test unilatéral à gauche (*ug*0) pris au niveau $\alpha/2$.

Comme le montrent BS, la puissance Π du test bilatéral est égale à :

$$\begin{aligned}
\Pi &= \Pr_{|d|>0}\left(|\hat{d}| > t_{1-\alpha/2}\right) \\
&= \Pr_{|d|>0}\left(\hat{d} - d > t_{1-\alpha/2} - d\right) \\
&+ \Pr_{|d|>0}\left(\hat{d} - d < -t_{1-\alpha/2} - d\right) \\
&= \Pr_{|d|>0}\left(\hat{d} - d < -t_{1-\alpha/2} + d\right) \\
&+ \Pr_{|d|>0}\left(\hat{d} - d < -t_{1-\alpha/2} - d\right) \\
&= \Phi\left(-t_{1-\alpha/2} + d\right) + \Phi\left(-t_{1-\alpha/2} - d\right)
\end{aligned}$$

La puissance du test bilatéral (*b*0) s'obtient donc comme la somme de deux termes. Le premier, $\Phi(-t_{1-\alpha/2} + d)$, est une fonction croissante de *d* dont la forme fonctionnelle est identique à celle de la puissance du test unilatéral à droite (*ud*0). Le second, $\Phi(-t_{1-\alpha/2} - d)$, est une fonction décroissante de *d* dont la forme fonctionnelle est identique à celle de la puissance du test unilatéral à gauche (*ug*0). La seule différence entre chacun de ces termes et la puissance

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

du test unilatéral correspondant est l'ordre du quantile : $1-\alpha/2$ au lieu de $1-\alpha$, ce qui renvoie au niveau de ces tests unilatéraux sous-jacents : non pas α mais $\alpha/2$ (*cf. supra*, lien entre la région critique du test bilatéral et les régions critiques des deux tests unilatéraux).

Pour que la puissance du test ($b0$) soit suffisante (au moins égale à une valeur fixée π jugée assez élevée, par exemple 80 %), il faut que :

$$(7)\ \Phi\left(-t_{1-\alpha/2}+d\right)+\Phi\left(-t_{1-\alpha/2}-d\right)\geq\pi$$

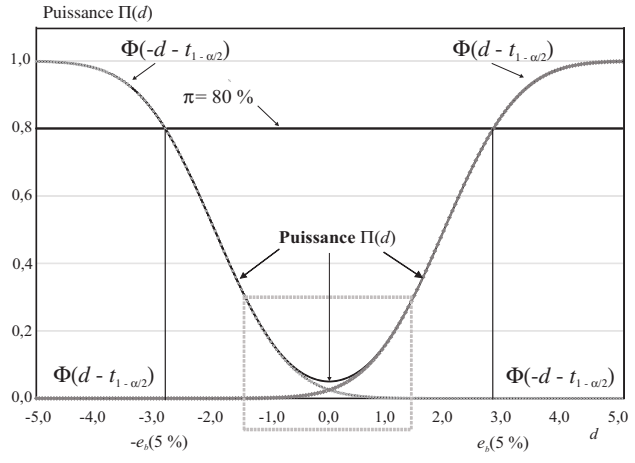
L'EMD peut s'obtenir rigoureusement par calcul numérique. Toutefois, on vérifie aisément, également par calcul numérique, que, pour toutes les valeurs usuelles de α et de π, un des deux membres de la somme constituant la puissance Π, à gauche de l'inégalité (7), est négligeable devant l'autre sur la plage des valeurs de d vérifiant (7) (graphique 3). Par exemple, pour tout α compris entre 1 % et 10 %, on montre aisément que :

– $\forall d$ tel que $d\leq-1{,}5$, $\Phi\left(-t_{1-\alpha/2}+d\right)\leq 0{,}002\Pi$, donc $\Pi\approx\Phi\left(-t_{1-\alpha/2}-d\right)$;

– $\forall d$ tel que $d\geq 1{,}5$, $\Phi\left(-t_{1-\alpha/2}-d\right)\leq 0{,}002\Pi$, donc $\Pi\approx\Phi\left(-t_{1-\alpha/2}+d\right)$;

– $\forall d$ tel que $-1{,}5<d<1{,}5$, $\Pi<0{,}45$: (7) n'est pas vérifiée pour les valeurs usuelles de π.

Graphique 3 : décomposition de la puissance d'un test bilatéral de niveau α = 5 % en la somme de puissances de deux tests unilatéraux de niveau $\alpha/2$



Lecture : la puissance Π est une fonction de d en forme de cloche inversée, obtenue par la somme d'une fonction croissante $\Phi(d-t_{1-\alpha/2})$ et d'une fonction décroissante $\Phi(-d-t_{1-\alpha/2})$. La première fonction, identique à la fonction puissance d'un test unilatéral à droite de niveau $\alpha/2$, tend rapidement vers zéro quand d décroît en dessous de -1,5. La seconde fonction, identique à la fonction puissance d'un test unilatéral à gauche de niveau $\alpha/2$, tend rapidement vers zéro quand d croît au-dessus de 1,5. L'intervalle [-1,5 ; 1,5] des valeurs de d pour lesquelles la puissance ne peut être approchée par l'une de ces ceux fonctions correspond à la zone où la puissance est la plus faible, bien inférieure à π (partie encadrée, proche de zéro, la frontière entre l'hypothèse nulle et l'hypothèse alternative). Ce graphique vaut tant pour le test ($b0$) que pour le test ($b\delta_0$) dès lors que la définition de d est adaptée au test en question : définition (3) pour le test ($b0$) et (3') pour le test ($b\delta_0$).

Source : auteure.

Ainsi, pour $\Pi\geq\pi$, π prenant des valeurs usuelles (de l'ordre de 0,80 à 0,90), on peut poursuivre le calcul en appliquant l'une des formules approchées de Π.

Premier cas : $d\geq 1{,}5$

Sur cette plage de valeurs, on se ramène au calcul de puissance d'un test unilatéral à droite ($ud0$) de niveau $\alpha/2$. En effet, (7) équivaut alors à :

$$\Phi\left(-t_{1-\alpha/2}+d\right)\geq\pi\Leftrightarrow d\geq(t_\pi+t_{1-\alpha/2})$$

$$\Leftrightarrow\delta\geq(t_\pi+t_{1-\alpha/2})\sigma_{\hat{\delta}}$$

D'où :

$$(8)\ EMD=(t_\pi+t_{1-\alpha/2})\sigma_{\hat{\delta}}\equiv e_b\sigma_{\hat{\delta}}$$

Par exemple pour α = 5 % et π = 80% (valeurs fréquemment choisies) et pour un échantillon suffisamment grand, on a : $t_{1-\alpha/2}\approx 1{,}96$ et $t_\pi\approx 0{,}84$, d'où $e_b\approx 2{,}80$. Autour de cette valeur nettement supérieure à 1,5, le calcul approché de la puissance qui a été utilisé est extrêmement précis.

Second cas : $d\leq-1{,}5$

Sur cette plage de valeurs, on se ramène au calcul de puissance d'un test unilatéral à gauche ($ug0$) de niveau $\alpha/2$. En effet, (7) équivaut alors à :

$$\Phi\left(-t_{1-\alpha/2}-d\right)\geq\pi\Leftrightarrow d\leq-(t_\pi+t_{1-\alpha/2})$$

$$\Leftrightarrow\delta\leq-(t_\pi+t_{1-\alpha/2})\sigma_{\hat{\delta}}$$

Comme *supra*, l'EMD est alors défini comme la valeur absolue de ce seuil, ce qui correspond exactement au résultat du calcul effectué pour $d\geq 1{,}5$:

$$EMD=(t_\pi+t_{1-\alpha/2})\sigma_{\hat{\delta}}=e_b\sigma_{\hat{\delta}}$$

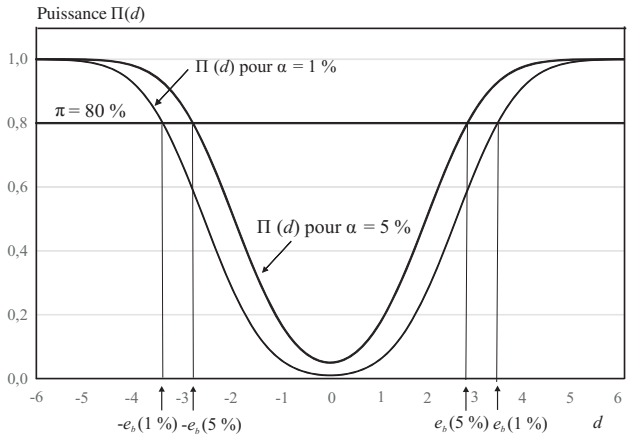
En pratique, il suffit donc de faire le calcul de l'EMD dans le premier cas et d'en interpréter le résultat en disant qu'un effet de la mesure, de quelque signe qu'il soit, sera détecté comme significatif si sa valeur absolue est supérieure à l'EMD ainsi calculé :

$$\Pi\geq\pi\Leftrightarrow|d|\geq t_\pi+t_{1-\alpha/2}\Leftrightarrow|\delta|\geq(t_\pi+t_{1-\alpha/2})\sigma_{\hat{\delta}}$$

En raisonnant en unités d'écart-type $\sigma_{\hat{\delta}}$, ces effets détectables correspondent à la partie de l'axe des abscisses du graphique 4 extérieure à l'intervalle $]-e_b,+e_b[$ (page suivante). Là encore, on montre aisément que, plus on opte pour un test conservateur (c'est-à-dire plus α est petit, autrement dit plus on se prémunit contre l'abandon à tort de l'hypothèse nulle), plus la puissance Π est faible ou, ce qui revient au même, plus l'EMD est élevé, toutes choses égales par ailleurs.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

Graphique 4 : calcul de puissance dans le cas d'un test bilatéral



Lecture : les valeurs de d situées en dehors de l'intervalle $]-e_b, +e_b[$ sont détectables par le test bilatéral. Plus le test est conservateur (c'est-à-dire plus α est petit), plus la puissance Π est faible, donc plus e_b et l'EMD sont éloignés de zéro toutes choses égales par ailleurs. Ce graphique vaut tant pour le test ($b0$) que pour le test ($b\delta_0$) dès lors que la définition de d est adaptée au test en question : définition (3) pour le test ($b0$) et (3') pour le test ($b\delta_0$).

Source : auteure.

Quels enseignements tirer de ces calculs de puissance au regard de la question posée

Des calculs d'EMD ont été présentés pour trois tests différents : le test bilatéral ($b0$) et les deux tests unilatéraux ($ud0$) et ($ug0$). Comme on l'a vu, ces calculs sont immédiatement transposables par simple translation de l'interprétation de l'EMD aux tests plus généraux ($b\delta_0$), ($ud\delta_0$) et ($ug\delta_0$) et ils renseignent également sur les tests ($udp0$) et ($ugp0$). À la lumière de ces calculs, quel test convient-il de privilégier ? En principe, ceci a son importance puisque le calcul d'EMD conduit à opter pour certaines caractéristiques de l'expérience. Le plus souvent dans la littérature en sciences sociales, les tests bilatéraux sont privilégiés, avec un niveau de test $\alpha = 5\ \%$ (donc $t_{1-\alpha/2} \approx 1{,}96$) et une valeur de puissance acceptable $\pi = 80\ \%$ (donc $t_\pi \approx 0{,}84$). Ceci revient à calculer l'EMD suivant :

$$EMD_b = e_b \sigma_{\hat{\delta}} \approx 2{,}80\, \sigma_{\hat{\delta}}$$

Cependant, ainsi que l'explique Bloom (1995), l'utilisation d'un test unilatéral à droite se justifierait dans de nombreux cas d'évaluation de politique publique, où l'on cherche à savoir si cette dernière a l'effet attendu. Ceci conduirait à l'EMD suivant (le même que pour un test unilatéral à gauche), pour $\alpha = 5\ \%$ (donc $t_{1-\alpha} \approx 1{,}64$) et $\pi = 80\ \%$ (donc $t_\pi \approx 0{,}84$) :

$$EMD_u = e_u \sigma_{\hat{\delta}} \approx 2{,}49\, \sigma_{\hat{\delta}}$$

À α, π et $\sigma_{\hat{\delta}}$ donnés, le test unilatéral à droite (respectivement à gauche) est donc plus puissant que le test bilatéral sur les plages de valeurs de d positives (respectivement négatives) correspondant à l'hypothèse alternative du test unilatéral à droite (respectivement à gauche). C'est ce qui fait dire à Bloom (1995) qu'utiliser un test unilatéral à droite est un moyen d'augmenter la puissance de l'évaluation qui ne devrait pas être négligé : « *This issue has important practical consequences, because limited statistical power is a major weakness of most program impact studies […]. There are well-established ways to increase power, such as increasing sample size, increasing the reliability of outcome measures, blocking the experimental sample, and statistically controlling for relevant covariates. But shifting from a two-sided hypothesis test to a more powerful one-sided test is not usually prescribed, although it should be in many cases* ».

Corrélativement, le test unilatéral est moins conservateur au sens où, pour assurer le même EMD que le test bilatéral toutes choses égales par ailleurs dont α et π, il conduit au choix d'un échantillon de plus petite taille. C'est là la traduction concrète d'un test plus puissant. Nous disposons maintenant de tous les éléments pour réexaminer le troisième argument souvent cité en appui de l'utilisation du test bilatéral que j'avais évoqué à la fin de la première partie de ce commentaire. De nombreux évaluateurs déclarent préférer utiliser le test bilatéral pour pouvoir disposer d'une marge de sécurité supérieure en termes de taille d'échantillon, afin de pallier toute éventualité, sachant à quel point un calcul de puissance est délicat en pratique (*cf.* quatrième partie de mon commentaire). Ce choix a un coût et peut être perçu comme quelque peu paradoxal d'un point de vue purement statistique. En effet, il revient à préférer un test moins puissant au motif que ce dernier conduit à un EMD plus élevé, ce qui nécessite, pour le réduire jusqu'à un certain seuil donné, une correction supérieure, se traduisant par le choix d'un nombre d'observations plus grand. Toutefois, le souci de se protéger contre les aléas de l'expérimentation en visant un nombre d'observations suffisamment élevé pour pallier toute éventualité est, lui, parfaitement légitime.

Cependant, même si l'on comprend bien ce souci, ceci ne rend pas l'argument convaincant. En réalité, cette question du choix entre test bilatéral et test unilatéral à droite s'apparente assez largement à un faux problème, compte tenu de la grande similarité des formules d'EMD associées à chacun des deux tests. En effet, comme le montrent les deux formules d'EMD (6) et (8) et comme cela a été souligné *supra*, il est équivalent de procéder à un calcul d'EMD à partir d'un test bilatéral de niveau α ou à partir d'un test unilatéral à droite (ou à gauche) de niveau $\alpha / 2$ (graphique 5) : on aboutit exactement au même résultat et, donc, aux mêmes décisions en termes de

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

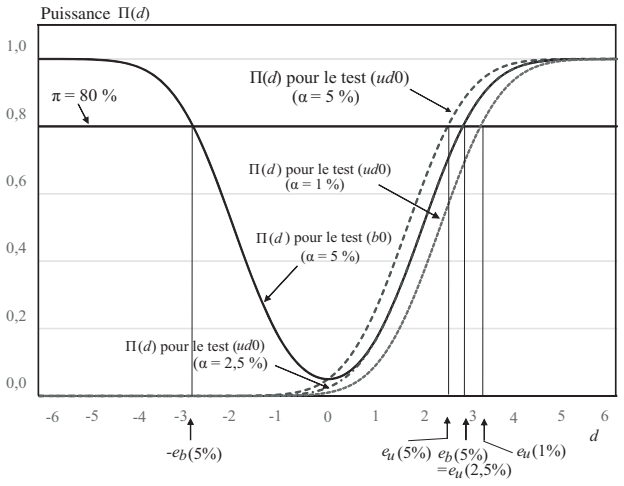
caractéristiques de l'expérience (*cf.* quatrième partie).

Au total, ce qu'il faut donc retenir, me semble-t-il, c'est qu'en procédant à un seul calcul d'EMD, par exemple à celui associé au test bilatéral comme c'est le plus fréquent, ou bien à celui du test unilatéral à droite, d'interprétation plus intuitive au regard de l'évaluation des politiques publiques et de résolution plus simple, on obtient les EMD associés à l'ensemble des tests envisagés[5], par la formule générale unique :

$$(9)\quad EMD = [t_{\pi} + t(\alpha)]\sigma_{\hat{\delta}}$$

On peut, pour se donner une marge de manœuvre en termes, notamment, de taille de l'échantillon, à toutes fins utiles, poser $t(\alpha) = 1{,}96$ quel que soit le test d'hypothèses d'intérêt retenu. Ce faisant, il faut juste se souvenir que le niveau des tests unilatéraux correspondants est deux fois moindre que le niveau du test bilatéral (2,5 % au lieu de 5 %). Il est évidemment immédiat, si l'on préfère gagner en puissance, d'opter pour des tests unilatéraux de même niveau que le test bilatéral. En effet, pour passer d'un test bilatéral de niveau α à un test unilatéral de même niveau, il suffit de remplacer $t(\alpha) = t_{1-\alpha/2}$ par $t(\alpha) = t_{1-\alpha}$, qui est inférieur à $t_{1-\alpha/2}$. On pourrait aussi opter pour un test unilatéral plus conservateur (et, donc, moins puissant) que le test bilatéral de niveau 5 % en optant pour un niveau de 1 % pour le test unilatéral. En effet, la valeur $e_u(1\%)$ est approximativement égale à 3,17, donc supérieure à $e_b(5\ \%) \approx 2{,}80$ (*cf.* deuxième partie, sous-partie "test unilatéral à droite (*ud*0)" et voir graphique 5). En d'autres termes, le test bilatéral n'est plus conservateur (et moins puissant) que le test unilatéral que si l'on s'enferme dans un niveau α donné, ce qui n'est en rien une obligation. Le choix du niveau d'un test doit refléter le prix que l'évaluateur accorde à la plus ou moins grande limitation du risque d'un abandon à tort de l'hypothèse nulle : il s'agit donc d'un paramètre qui relève de son choix, dans l'éventail de valeurs considérées comme admissibles [1 % ; 10 %], et qu'il doit effectuer en fonction du contexte de l'expérimentation, des contraintes auxquelles il est confronté et de ses préférences. En tout état de cause, selon le test d'hypothèses que l'on souhaite privilégier (test bilatéral, unilatéral à droite ou unilatéral à gauche), il suffit de se souvenir de la manière dont interpréter l'EMD, qui est différente selon les tests, comme cela a été montré plus haut et illustré par les graphiques 1, 2 et 4. Enfin, un évaluateur non statisticien qui ne maîtriserait pas les calculs qui viennent d'être présentés peut, sans risque de se tromper, suivre l'approche, courante, décrite par BS, tout comme il pourrait suivre l'approche suggérée par Bloom (1995). Il me paraît néanmoins souhaitable de maîtriser suffisamment ces questions pour pouvoir faire un choix éclairé du couple (type de test, niveau du test).

Graphique 5 : équivalence des calculs d'EMD tirés des tests (*b*0) et (*ud*0)



Lecture : à niveau α donné (ici 5 %), le test unilatéral à droite (*ud*0) est plus puissant que le test bilatéral (*b*0) pour toutes les valeurs positives de *d* : la courbe de puissance du test unilatéral se situe au-dessus de la courbe de puissance du test bilatéral ou, de manière équivalente, l'EMD du test unilatéral est inférieur à l'EMD du test bilatéral. En revanche, le test bilatéral (*b*0) de niveau 5 % et le test unilatéral (*ud*0) de niveau 2,5 % ont une puissance équivalente sur la plage des valeurs de *d* supérieures à 1,5 : les deux courbes de puissance sont confondues, de même que l'EMD des deux tests. Enfin, le test bilatéral (*b*0) de niveau 5 % est plus puissant que le test unilatéral (*ud*0) de niveau 1 %. Ce graphique vaut pour les tests ($b\delta_0$) et ($ud\delta_0$) en optant pour la définition (3') de *d*. On obtiendrait le même type de résultat entre le test bilatéral (*b*0) et le test unilatéral à gauche (*ug*0) (ou ($b\delta_0$) et ($ug\delta_0$) en optant pour la définition (3') de *d*) sur les valeurs négatives de *d* en renversant les courbes relatives à (*ud*0) par symétrie par rapport à l'axe des ordonnées et en remplaçant e_u par $-e_u$ et *vice versa*.

Source : auteure.

Remarquons qu'il existe une autre manière d'opter pour une décision plus ou moins conservatrice en jouant sur les valeurs du quantile $t(\alpha)$ à α donné qu'en optant pour un test bilatéral ou unilatéral. En abandonnant l'approche asymptotique au profit de l'approche à distance finie, quel que soit le test retenu, on adopte de fait un comportement plus conservateur que dans l'approche asymptotique, comme on va le voir. L'inconvénient majeur de l'approche à distance finie est d'obliger à retenir une hypothèse assez forte sur la loi à distance finie des termes d'erreur du modèle d'évaluation (1) pour pouvoir procéder au calcul de puissance. On suppose en général que les (ε_i) sont *i.i.d.* et normaux, de variance commune inconnue σ^2, sans pouvoir toujours justifier la normalité, du fait qu'on ne se situe pas dans le champ d'application de la loi des grands nombres, du théorème central limite et de leurs généralisations. Si, dans ce cadre, on utilise une estimation de l'écart-type σ, qui constitue une composante de $\sigma_{\hat{\delta}}$ (*cf.* partie suivante), pour calculer l'EMD, alors il faut se souvenir qu'à distance finie, le remplacement de σ par son estimateur modifie la loi suivie par $\hat{\delta}$. Par exemple, si l'on remplace σ par son estimateur sans biais optimal $\hat{\sigma}$ lors du calcul de puissance[6], (2) devient :

$$(2')\quad (\hat{\delta} - \delta) / \hat{\sigma}_{\hat{\delta}} \rightarrow T(N-2)$$

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

où $T(N\text{-}2)$ est la loi de Student à N-2 degrés de liberté. Cette loi étant symétrique par rapport à zéro, le calcul de puissance s'effectue ensuite de la même façon que sous l'hypothèse de normalité (2). En revanche, les valeurs des quantiles t_π et $t(\alpha)$ sont légèrement modifiées. Si l'on raisonne au moyen d'un test bilatéral, pour $\alpha = 5$ % et $\pi = 80$ %, au lieu des quantiles de la loi normale centrée réduite $t_{1-\alpha/2} \approx 1{,}96$ et $t_\pi \approx 0{,}84$ (conduisant à $e_b \approx 2{,}80$), on doit utiliser les quantiles de la loi de Student à N-2 degrés de liberté, qui sont légèrement supérieurs. Par exemple, pour $N - 2 = 60$, $t_{1-\alpha/2} \approx 2{,}00$ et $t_\pi \approx 0{,}85$, soit $e_b \approx 2{,}85 > 2{,}80$. Toutefois, au fur et à mesure que le nombre d'observations augmente, les quantiles de la loi de Student se rapprochent de ceux de la loi normale centrée réduite. Le caractère moins conservateur de l'approche asymptotique n'est donc pas flagrant, au moins pour ce qui concerne les tirages au niveau individuel. En effet, le nombre total d'observations N est en général suffisamment élevé pour que les quantiles de la loi de Student à N-2 degrés de liberté soient très proches de ceux de la loi normale centrée réduite. Toutefois, il se peut que le cadre à distance finie s'impose plus souvent dans un contexte de tirage par grappes, particulièrement si le nombre de grappes est limité[7].

Dans la suite, je me place à nouveau dans le cas d'un nombre d'observations suffisant pour que l'approche asymptotique puisse être considérée comme valide. Compte tenu de l'équivalence des calculs d'EMD sur les différents tests, je me concentre sur l'EMD du test bilatéral ($b0$), sachant que l'EMD des tests unilatéraux s'en déduit immédiatement en remplaçant $t_{1-\alpha/2}$ dans les formules d'EMD par $t_{1-\alpha}$.

Finalité du calcul de puissance et importance de la formulation de l'écart-type $\sigma_{\hat{\delta}}$

Illustration de la finalité du calcul de puissance

BS présentent un graphique très éclairant de la manière dont fonctionne le calcul de puissance dans l'annexe 2 de leur article. Les graphiques 6 de ce commentaire développent une illustration très proche, afin d'aider à la visualisation de l'utilité du calcul de puissance pour le choix des caractéristiques du protocole expérimental. Ces graphiques se rapportent au cas où le calcul d'EMD est effectué sur la base d'un test bilatéral et où la vraie valeur de l'effet δ est positive. Toutefois, le même raisonnement pourrait être aisément tenu dans le cas d'un effet négatif et/ou d'un test unilatéral, comme l'explique la note de lecture sous ces graphiques.

Sur ces graphiques 6 sont représentées les densités de $\hat{d}$ sous l'hypothèse nulle $d = 0$ (densité ϕ_0 de la loi normale centrée réduite) et sous l'hypothèse alternative (densité ϕ_d de la loi normale de moyenne d et de variance unitaire, conformément à (4)[8]). Le niveau α du test est représenté par l'aire sous la courbe de densité ϕ_0 extérieure à l'intervalle $]-t_{1-\alpha/2}, t_{1-\alpha/2}[$ (exprimée en pourcentage de l'aire totale sous la courbe de densité ϕ_0). La puissance est représentée par l'aire située sous la courbe de densité ϕ_d extérieure à ce même intervalle (exprimée en pourcentage de l'aire totale sous la courbe de densité ϕ_d). $\Phi(-d - t_{1-\alpha/2})$ correspond à l'aire sous la densité ϕ_d à gauche du seuil $-t_{1-\alpha/2}$. Elle peut être considérée comme infinitésimale au regard de la partie hachurée à droite du seuil $t_{1-\alpha/2}$, qui représente $\Phi(d - t_{1-\alpha/2})$.

À un niveau donné (par exemple $\alpha = 5$ %, soit $t_{1-\alpha/2} = 1{,}96$), le calcul de puissance consiste à jouer sur plusieurs facteurs conditionnant la valeur de l'EMD *via* la variance $\sigma_{\hat{\delta}}$ de l'effet estimé $\hat{\delta}$ (taille de l'échantillon, proportion d'individus traités, degré d'hétérogénéité des observations, etc. – *cf.* article de BS et voir *infra*), afin d'atteindre un niveau de puissance acceptable π. Dans le graphique 6.a, il faut réduire l'EMD *via* une diminution de $\sigma_{\hat{\delta}}$ jusqu'à que la partie hachurée représente $\pi = 80$ % de l'aire sous la courbe de densité ϕ_d, c'est-à-dire jusqu'à ce que la configuration du graphique 6.b soit atteinte, par déplacement de $d = \delta / \sigma_{\hat{\delta}}$ vers la droite (à δ évidemment constant) jusqu'à égaliser la valeur de e_b, fixée par le choix de α et de π. Dans cette configuration, l'effet de la mesure δ peut être détecté par l'expérience.

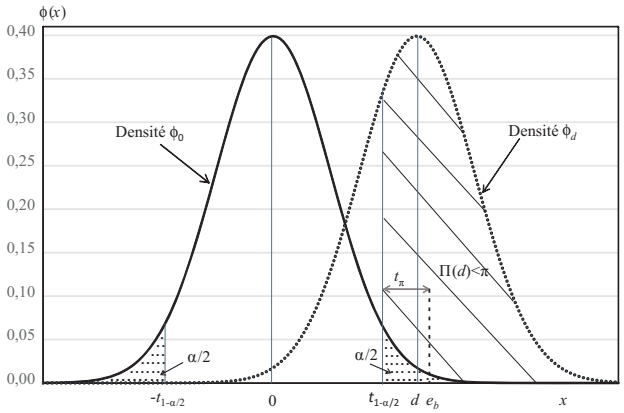
En pratique, la question de la valeur π la plus adaptée pour détecter correctement l'effet d'une politique se pose. Si l'on recommande en général de choisir une valeur comprise entre 80 % et 90 %, des contraintes budgétaires entrent souvent en ligne de compte, qui peuvent empêcher d'atteindre ces niveaux recommandés. Dans ce cas, le calcul de puissance s'apparente à une manière d'estimer, compte tenu des contraintes budgétaires, la puissance de l'évaluation future, une puissance trop basse pouvant conduire l'évaluateur à décider de ne pas se lancer dans l'évaluation dans ces conditions (Duflo *et alii*, 2006, page 39).

En tout état de cause, cette illustration souligne l'importance de l'écart-type $\sigma_{\hat{\delta}}$ dans l'utilisation du calcul de puissance, puisque c'est exclusivement à travers cet écart-type que les caractéristiques du protocole expérimental jouent sur la valeur de l'EMD à α et π fixés. Il est donc souhaitable d'utiliser la formulation de $\sigma_{\hat{\delta}}$ la plus juste possible, afin que le calcul de l'EMD soit le moins biaisé possible et que les choix qui en résultent permettent effectivement de détecter l'effet de la mesure à évaluer avec une puissance suffisante et un risque d'erreur de type I suffisamment contrôlé.

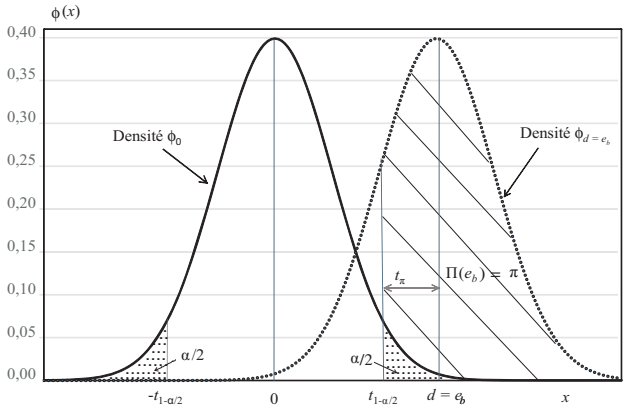
Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

Graphiques 6 : densités de la loi de $\hat{d}$ sous les hypothèses nulle et alternative dans le cas d'un test bilatéral

a) Puissance inférieure à π : l'effet de la mesure ne pourra être détecté par l'expérience



b) Puissance égale à π : l'effet de la mesure pourra être détecté par l'expérience



Lecture : les graphiques se rapportent au cas d'un test bilatéral, dans le cas où l'effet de la mesure est strictement positif. On obtiendrait le cas d'un test bilatéral avec effet strictement négatif en renversant les graphiques 6 par symétrie par rapport à l'axe des ordonnées et en remplaçant par e_b par $-e_b$. En outre, le cas d'un test unilatéral à droite se déduirait aisément des graphiques 6 en y remplaçant $t_{1-\alpha/2}$ par $t_{1-\alpha}$, $\alpha/2$ par α sous l'aire délimitée par la barre verticale surplombant $t_{1-\alpha/2}$ et la courbe de la densité ϕ_0, ainsi que e_b par e_u, et en supprimant $-t_{1-\alpha/2}$ et la barre verticale qui le surplombe ainsi que la référence à l'aire $\alpha/2$ qu'elle délimite avec la densité ϕ_0. Le cas d'un test unilatéral à gauche se déduirait tout aussi aisément du graphique correspondant au test unilatéral à droite par symétrie par rapport à l'axe des ordonnées et remplacement de e_u par $-e_u$ et de $t_{1-\alpha}$ par $-t_{1-\alpha}$.
Source : auteure.

Formulation de l'écart-type $\sigma_{\hat{\delta}}$

Dans la configuration retenue dans l'article de BS au début de l'annexe 2 et lors de leur illustration dans le corps du texte sur l'exemple des "Groupements de Créateurs", les termes d'erreurs sont *i.i.d.* de variance commune σ^2, inconnue. Comme le montrent BS, si la conformité est parfaite, l'écart-type estimé de l'estimateur des MCO $\hat{\delta}$ de l'effet δ de la mesure s'exprime alors comme suit :

$$(10)\ \sigma_{\hat{\delta}} = \frac{\sigma}{\sqrt{\tau(1-\tau)N}}$$

où τ désigne la proportion d'individus de l'échantillon (de taille N) assignés au groupe test. Dans un cadre *i.i.d.* avec conformité imparfaite, $\sigma_{\hat{\delta}}$ devient :

$$(11)\ \sigma_{\hat{\delta}} = \frac{\sigma}{\sqrt{\tau(1-\tau)N}}\ \frac{1}{(p^1 - p^0)}$$

où $p^1 - p^0$ est le taux de conformité (équation (6) de l'article de BS). Enfin, si l'hypothèse *i.i.d.* est levée pour cause d'un mode de tirage par grappes, il faut tenir compte d'une corrélation intra-grappe des termes d'erreurs et $\sigma_{\hat{\delta}}$ devient (équation (7) de l'article de BS, inspirée de Bloom, 2005) :

$$(12)\ \sigma_{\hat{\delta}} = \frac{\sigma}{\sqrt{G\tau(1-\tau)}}\ \sqrt{\rho + \frac{1-\rho}{n}}\ \frac{1}{(p^1 - p^0)}$$

où G désigne le nombre de grappes, n le nombre d'individus par grappe, ρ la corrélation intra-grappe et σ l'écart-type inter-grappes des termes d'erreur du modèle d'évaluation (1)[9].

Selon que l'on utilise une formulation ou une autre, le calcul de puissance ne débouche pas sur les mêmes caractéristiques de l'échantillon, notamment en termes de nombre d'observations N. Par exemple, prévoir (formulation (11)) ou non (formulation (10)) une configuration de conformité imparfaite n'est pas neutre pour la puissance *in fine* de l'évaluation. Négliger la possibilité de conformité imparfaite revient à opter pour des caractéristiques du protocole expérimental insuffisamment conservatrices. En effet, le taux de conformité étant inférieur à 1, la conformité imparfaite se traduit par une augmentation de la variance de l'estimateur $\hat{\delta}$. Il faut donc viser un nombre d'observations supérieur, toutes choses égales par ailleurs, pour aboutir à un EMD, donc une puissance de l'évaluation, identiques.

Ces exemples donnés dans l'article de BS permettent donc de sensibiliser utilement sur l'importance d'une bonne formulation de $\sigma_{\hat{\delta}}$ pour la suite de l'évaluation, en distinguant les causes d'influence de cette formulation liées aux choix de l'évaluateur lui-même (tirage au niveau individuel ou par grappes) de celles qui lui échappent au moins en partie (valeur du taux de conformité). Ils sensibilisent en outre sur la nécessité d'être suffisamment vigilant pour mettre en cause quand il le faut un corps d'hypothèses tellement usuel qu'on pourrait être tenté d'oublier d'en vérifier le réalisme dans le contexte de travail dans lequel on se situe.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

C'est le cas de l'hypothèse de termes d'erreur (ε_i) *i.i.d.*, dont BS rappellent qu'il convient de la lever partiellement si l'on opte pour un tirage par grappes, en raison des dépendances intra-grappe liées à ce type de tirage. L'hypothèse *i.i.d.* peut aussi devoir être levée partiellement dans d'autres configurations ne résultant pas de choix raisonnés de l'évaluateur. Le risque d'opter pour une formulation de $\sigma_{\hat{\delta}}$ inadéquate en raison d'hypothèses excessivement simplificatrices sur la matrice Σ de variances-covariances des termes d'erreur peut alors être plus élevé[10].

L'estimation de la matrice de variances-covariances des termes d'erreur

Pour terminer, remarquons que, dans les cas (10) et (11) de l'article de BS, $\sigma_{\hat{\delta}}$ (donc l'EMD) est proportionnel à σ et toutes les autres composantes de l'EMD (dans (10)) ou la plupart d'entre elles (dans (11)) sont des dimensions sur lesquelles peut jouer l'évaluateur. Comme l'indiquent BS, une première solution consiste alors à raisonner sur l'EMD en unités d'écart-type σ, c'est-à-dire à calculer EMD / σ Ceci permet de s'affranchir de l'obligation d'estimer σ, si cette estimation s'avère trop délicate. Ainsi que le font remarquer Duflo *et alii* (2006, page 39), cette méthode astucieuse a l'inconvénient, si l'on n'a aucune information sur σ, de laisser l'évaluateur dans l'incertitude totale concernant ce que signifie un effet petit, moyen ou grand en pratique. Une seconde solution, préférable si on peut l'appliquer, est donc de rechercher un estimateur de σ pour pouvoir raisonner dans l'absolu (*i.e.* sur *EMD* plutôt que sur EMD / σ). Dans leur exemple de calcul de puissance sur le cas « Groupements de Créateurs », BS utilisent une variable de résultat binaire prenant la valeur 0 ou 1, ce qui leur permet d'estimer la variance inconditionnelle théorique de *Y* en ayant une idée de la proportion des individus dans la population tels que $Y_i = 1$. Soit S_Y l'écart-type estimé de *Y* qui en résulte. BS remarquent ensuite que, sous leurs hypothèses, l'estimateur sans biais optimal de σ^2 est l'estimateur des MCO dans le modèle d'évaluation (1) et que cet estimateur s'exprime comme le produit de la variance empirique inconditionnelle de *Y* sur l'échantillon et de $1-R^2$, où R^2 est la part de la variance inconditionnelle de *Y* expliquée par le modèle (1). Dans la mesure où le R^2 des modèles d'évaluation est en général très petit, les auteurs optent pour S_Y comme estimation conservatrice de σ dans le calcul d'EMD.

Ces deux solutions sont effectivement envisageables dans les cas traités par les auteurs, sous condition que leur cadre d'hypothèses soit vérifié. Il reste alors à décider de la valeur de la proportion τ d'individus assignés au traitement : la valeur 0,5 maximisant $\tau(1-\tau)$ (et donc minimisant l'EMD) paraît indiquée dans les cas (10) et (11), à moins que des considérations de coût ne conduisent à opter pour un partage inégal entre individus tests et témoins (Duflo *et alii*, 2006, page 32). Ensuite, il reste à déterminer la taille *N* de l'échantillon permettant d'atteindre la valeur d'EMD souhaitée, en prenant garde à ne pas négliger le risque de non-conformité. La formulation (11) fait en effet apparaître une influence particulièrement forte de la non-conformité sur la valeur de $\sigma_{\hat{\delta}}$, soulignée dans Duflo *et alii* (2006, pages 33-34). Ceux-ci recommandent de limiter le risque de non-conformité par des conceptions d'échantillonnage adaptées. Toutefois, l'estimation du taux de conformité demeure un défi pour l'évaluateur.

Ces solutions simples ne peuvent évidemment pas être reproduites à l'identique dans tous les cas. Le calcul de l'EMD doit être adapté au contexte et, notamment, aux choix méthodologiques de l'évaluateur. Par exemple, dans le cas *i.i.d.* évoqué brièvement par BS dans leur article, où, pour réduire l'écart-type des termes d'erreur, l'évaluateur ajoute des variables de contrôle bien choisies au modèle d'évaluation, il serait discutable d'utiliser une estimation de la variance inconditionnelle de *Y* pour estimer σ^2, dans la mesure où l'ajout de variables de contrôle se traduit par une augmentation du R^2 du modèle qui peut être notable et, par conséquent, une diminution potentiellement non négligeable de σ^2 (c'est bien le but). On peut alors estimer le R^2 de la régression de *Y* sur les variables de contrôle choisies, par exemple au vu de travaux antérieurs.

De même, dans le cas de la formule (12), il faut se souvenir que σ n'a pas la même interprétation que dans (10) et (11). Par ailleurs, lorsque la matrice Σ de variances-covariances des termes d'erreurs (ε_i) ne se réduit pas à un seul paramètre inconnu σ^2, comme c'est le cas si l'évaluateur opte pour un tirage par grappes, la solution qui consiste à raisonner sur l'EMD en unités d'écart-type ne constitue plus une panacée. En effet, même si l'on peut exhiber un paramètre de la matrice Σ avec lequel $\sigma_{\hat{\delta}}$ est proportionnel (σ dans (12)), le raisonnement en unités de cet écart-type particulier est sans doute moins parlant que lorsqu'on raisonne en unités d'écart-type total des termes d'erreurs du modèle d'évaluation. Surtout, il n'affranchit pas l'évaluateur de l'estimation des autres paramètres "nuisibles" de la matrice Σ (ρ dans (12)). L'estimation de ces paramètres peut être très délicate (Duflo *et alii*, 2006, page 38). En pratique, dans le cas d'un tirage par grappes, l'évaluateur peut tenter d'avoir recours à des sources préexistantes pour estimer les paramètres nuisibles σ et ρ, en général à partir des variances inter et intra-grappe(s) de la variable de résultat, par un raisonnement analogue à celui développé par BS dans le cas du tirage au niveau individuel. Ensuite, il peut décider quel nombre de grappes *G* et nombre d'individus par grappe *n* retenir en fonction, notamment, de considérations de coût.

Au total, en matière d'estimation de $\sigma_{\hat{\delta}}$, il n'y a pas de recette systématique, tout dépend des choix

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

méthodologiques de l'évaluateur, du contexte de l'évaluation et de l'ensemble d'informations *ex ante* que l'évaluateur peut se procurer en amont du calcul de puissance. BS donnent cependant des pistes utiles, qui peuvent être adaptées à d'autres contextes que ceux dans lesquels ils se situent. En tout état de cause, l'attention consacrée par l'évaluateur à estimer les composantes de $\sigma_{\hat{\delta}}$ au plus juste (ou, à tout le moins, en veillant à ne pas le sous-estimer) et les efforts qu'il consent pour s'en donner les moyens ont une influence notable sur la qualité du calcul de puissance et, par suite, sur la puissance ultérieure de l'évaluation.

Conclusion

L'article de BS a le grand mérite de présenter le calcul de puissance dans la méthode d'assignation aléatoire de manière beaucoup plus rigoureuse et pédagogique que dans nombre de documents de référence antérieurs. Pour autant, certains points de questionnement subsistent inévitablement, que j'ai cherché à développer et clarifier ici.

Le cœur du commentaire s'est intéressé à la question du test à utiliser de manière préférentielle pour fonder un calcul de puissance (trois premières parties). La littérature est en effet allusive ou non unanime sur cette question. En outre, les arguments qui y sont avancés pour justifier l'emploi de tel ou tel test au détriment d'un autre ne s'avèrent pas toujours convaincants. J'espère avoir clarifié cette question, en montrant l'équivalence des calculs d'EMD quel que soit le test utilisé parmi l'ensemble de tests envisagés, dès lors que l'on maîtrise bien l'interprétation de l'EMD pour chacun de ces tests ainsi que la question du niveau de test à utiliser pour assurer l'égalité stricte des EMD tirés des différents tests. Un seul calcul théorique d'EMD suffit : tous les autres s'en déduisent. Il faut ensuite faire un choix concernant le quantile $t(\alpha)$ ou, en d'autres termes, le risque d'erreur de type I à privilégier, selon ce à quoi on est le plus sensible, en fonction du contexte. Si l'on vise le gain en puissance, le choix du quantile d'ordre 95 % de la loi normale centrée réduite, correspondant à un test unilatéral, peut être le plus approprié. Si l'on préfère opter pour un test plus conservateur afin de s'assurer *ex post* une marge de manœuvre suffisante en termes de nombre d'observations en cas de mauvaise surprise, le choix du quantile d'ordre 97,5 % de la loi normale centrée réduite, correspondant à un test unilatéral de niveau 2,5 % ou à un test bilatéral de niveau 5 %, peut-être plus indiqué. En tout état de cause, ce choix, contrairement à ce qui se dit souvent, ne consiste pas à opter pour un type de test (test bilatéral) ou un autre (test unilatéral) mais pour un risque d'erreur de type I adapté, quel que soit le test sur lequel est fondé le calcul de puissance.

L'autre question abordée dans ce commentaire est celle de l'importance opérationnelle du paramètre $\sigma_{\hat{\delta}}$, unique canal par lequel l'évaluateur peut influencer la puissance de l'évaluation en jouant sur les caractéristiques de l'échantillon. Encore faut-il que le paramètre $\sigma_{\hat{\delta}}$ soit correctement formulé, pour que cette étape préparatoire importante de l'évaluation soit réalisée dans de bonnes conditions. Cette bonne formulation suppose notamment de bien mesurer les conséquences du mode de tirage retenu et de prendre en compte au mieux certains paramètres "nuisibles" difficilement maîtrisables. Au-delà de cette formulation, se pose la question de l'estimation *ex ante* des composantes du paramètre $\sigma_{\hat{\delta}}$, notamment des coefficients de la matrice de variances-covariances des termes d'erreurs du modèle d'évaluation. Comme on l'a vu, cette estimation ne procède pas de recettes valables dans tous les cas. Toutefois, de grands principes utiles ont été formulés dans l'article de BS, qui peuvent être adaptés selon les besoins et le contexte.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

Notes

(1) Voir, par exemple, les articles “Théorème Central Limite” et “Central Limit Theorem” de Wikipedia.

(2) Remarquons que la région critique du test (*ud*0) est identique à celle du test unilatéral à droite avec hypothèse nulle ponctuelle (*udp*0) ($H_0 : d = 0$) contre $H_a : d > 0$). Le calcul de puissance qui suit est également identique pour les deux tests. En réalité, hormis dans les cas où, effectivement, d est d'un signe bien déterminé, l'écriture du test avec hypothèse nulle ponctuelle n'a en général pas grand intérêt sinon de constituer une étape pédagogique à l'introduction du test (*ud*0), en rendant immédiat le calcul de la région critique.

(3) Si l'on s'intéresse au test ($ud\delta_0$) de l'hypothèse nulle $H_0 : \delta \leq \delta_0$ contre l'hypothèse alternative $H_a : \delta > \delta_0$, alors la relation précédente devient : $\delta - \delta_0 \geq (t_\pi + t_{1-\alpha})\sigma_{\hat{\delta}} > 0$, ce qui veut dire que l'EMD donné ensuite doit s'interpréter en écart à δ_0 (plutôt qu'en écart à zéro) : c'est l'écart à δ_0 minimum détectable.

(4) Comme dans le cas des tests unilatéraux à droite, la région critique du test (*ug*0) est identique à celle du test unilatéral à gauche avec hypothèse nulle ponctuelle (*ugp*0) ($H_0 : d = 0$ contre $H_a : d < 0$). Le calcul de puissance qui suit est également identique pour les deux tests (*ug*0) et (*ugp*0).

(5) y compris des tests d'hypothèses à frontière non nulle ($b\delta_0$), ($ud\delta_0$) et ($ug\delta_0$), en interprétant l'EMD en écart non plus à zéro mais à δ_0.

(6) Il s'agit là d'une option assez logique pour le calcul théorique de puissance, même si, *ex post*, l'estimation de σ retenue pour le chiffrage proprement dit sera plus frustre compte tenu des informations disponibles limitées au moment du calcul de puissance. À tout le moins la prise de conscience de l'impact sur la loi de $\hat{\delta}$ de l'estimation de σ à distance finie permet-elle, comme on va le voir, d'affiner le calcul de puissance à distance finie et d'opter pour un choix de quantiles plus conservateur, à défaut de pouvoir être très précis sur l'estimation de σ. Comme le montrent BS, c'est souvent un majorant de σ qui est choisi *in fine*, ce qui revient à être plus conservateur que si l'on disposait de l'estimateur sans biais optimal utilisé dans le calcul théorique de puissance.

(7) Intuitivement, on s'attend à ce que la dimension du nombre de grappes soit celle qui compte le plus pour établir la normalité asymptotique de l'estimateur $\hat{\delta}$ du fait de l'indépendance inter-grappes. À l'opposé, on s'attend à ce que, plus les individus d'une même grappe sont fortement dépendants, moins ils contribuent à cette normalité asymptotique. Toutefois, je n'ai pas connaissance de la démonstration de cette intuition.

(8) La traduction de (4) pour N grand est que $\hat{d}$ suit approximativement une loi normale de moyenne d et de variance 1, $N(d, 1)$. Sous l'hypothèse nulle, $d = 0$ (densité ϕ_0), sous l'hypothèse alternative, $d \neq 0$ (densité ϕ_d), les graphiques 6 prenant l'exemple $d > 0$.

(9) Le cas *i.i.d.* précédent s'obtient comme cas particulier de cette formule, en posant $\rho = 0$, l'écart-type σ étant alors identique à l'écart-type total commun des termes d'erreur (avec $G = N$ et $n = 1$). Le tirage par grappes est utilisé pour réduire le coût de l'expérimentation, non pour améliorer la précision de $\hat{\delta}$. La formulation (11) est en effet toujours inférieure ou égale à la formulation (12) à N fixé. Dans le cas d'un tirage par grappes, le nombre de grappes G influence la valeur de $\sigma_{\hat{\delta}}$ plus fortement que le nombre d'individus par grappe n, particulièrement si ρ est élevé. Intuitivement, cela s'explique par le fait qu'une augmentation du nombre n d'individus par grappe apporte moins d'information que l'ajout d'une grappe supplémentaire si les observations nouvellement introduites au sein des grappes préexistantes sont fortement corrélées aux observations préalablement présentes dans ces grappes (Duflo *et alii*, 2006, page 32). Le raisonnement limite conduit à minimiser (12) en posant $G = N$ et $n = 1$ à ρ et N donnés, ce qui se démontre aussi sans difficulté mathématiquement.

(10) Un exemple de telle configuration est celui où le modèle (1) est hétéroscédastique (Angrist et Pischke, 2009, chapitre 8, section 8.1). On peut montrer que, si l'homoscédasticité des termes d'erreur du modèle (1) est assurée pour un effet du traitement homogène sur l'ensemble des individus de l'échantillon ($Y_i^1 - Y_i^0 = \delta$), elle peut ne pas l'être sinon, où Y_i^1 (respectivement Y_i^0) désigne la variable de résultat pour l'individu i si i est traité (respectivement s'il ne l'est pas). Dans ce dernier cas, δ s'interprète comme l'effet moyen du traitement, non comme l'effet du traitement identique pour tous les individus de l'échantillon.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)

Références citées

Angrist J.D. et Pischke J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 392 pages.

Bloom H.S. (1995). "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs", *Evaluation Review*, vol. 19, n° 5, October, pp. 547-556.

Bloom H.S. (2005). "Randomizing Groups to Evaluate Place-Based Programs", NY: Russell Sage Foundation, chapitre "Learning More from Social Experiments", pp. 115-172.

Bouguen A. et Seban J. (2014). "L'assignation aléatoire comme méthode d'évaluation des politiques publiques", *Économie et Prévision*, n° 204-205, pp. 119-143, ce numéro.

Duflo E., Glennerster R. et Kremer M. (2006). "Using Randomization in Development Economics Research: A Toolkit", NBER Technical Working Paper, n° 333, 87 pages.

Direction générale du Trésor | Téléchargé le 27/10/2026 sur https://shs.cairn.info par Courtier Théophile (IP: 81.1.6.66)