Tutorials in Quantitative Methods for Psychology
2007, Vol. 3(2), p. 35-42.
DOI: 10.20982/tqmp.03.2.p035

Inventer ou estimer la puissance statistique ? Quelques considérations utiles pour le chercheur

Louis Laurencelle
Université du Québec à Trois- Rivières

La puissance statistique est un concept dont la définition mathématique et l'utilisation par les chercheurs soulèvent encore des difficultés. Nous revisitons différents aspects du problème dans cet article polémique et formulons enfin quatre propositions argumentées dans le but d'éclairer le chercheur : 1. Le recours à une valeur d'effet prescrite, selon un argument clinique ou de portée pratique, produit une estimation de puissance délibérément fausse, et qui contrevient de ce fait aux raisons déontologiques sur lesquelles il repose. 2. Il est illusoire et peut être mensonger de calculer une puissance qui n'est pas basée sur des données du domaine de recherche. 3. Le chercheur qui possède de l'information numérique préalable sur son projet peut s'en servir utilement pour en estimer la puissance. 4. Le calcul de la puissance d'un test d'hypothèses bidirectionnel ne doit pas inclure la contrepartie de probabilité.

The mathematical definition and practical use of statistical power by researchers raise logical as well as computational difficulties. In this polemical paper, we revisit some aspects of the problem and formulate four backed up propositions intended for the user : 1. Use of a prescribed effect size, on the basis of its clinical or practical significance, entails a deliberately false estimation of power, which contravenes also with its deontological justification. 2. It is illusory and may be deceitful to calculate and to propound a power value that is not grounded upon relevant data in the research area. 3. If the researcher has previous numerical information pertaining to his research venture, he may usefully apply it for estimating the anticipated statistical power. 4. It is absurd to incorporate the alternate-tail probability when computing the power of a two-tailed test.

Le chercheur qui planifie une expérimentation prochaine se voit souvent confronté au besoin d'évaluer la *puissance statistique* associée à son projet ou, équivalemment, de déterminer le nombre de sujets qu'il lui faudra recruter pour atteindre une puissance suffisante. Il est en effet peu rentable pour le chercheur d'entreprendre un travail de recherche dont la puissance, c.-à-d. la probabilité d'obtenir un résultat significatif, serait d'emblée insuffisante, sans parler des enjeux déontologiques qui en découlent. Ce besoin d'évaluer la puissance est aussi une obligation qu'imposent plusieurs organismes bailleurs de fonds, et la justification donnée, une condition d'attribution budgétaire. Besoin ou obligation, la question de puissance constitue un pivot pour toute entreprise de recherche : quelle que soit la valeur de l'hypothèse de recherche ou la qualité du protocole mis en oeuvre, le travail ne sera pas fructueux si la puissance n'y est pas.

Dans cet article, en partie polémique, nous rappelons d'abord les définitions et quelques éléments de notation mathématique liés à la puissance statistique. Nous présentons ensuite quatre arguments traitant chacun d'un aspect distinct du calcul de puissance, des arguments de controverse qui heurtent certaines conceptions et pratiques courantes en recherche et dont nous tenterons d'établir le bien-fondé.

Définitions et notation

La question de puissance est étroitement liée à la pratique des tests d'hypothèses statistiques, une pratique dont le cadre conceptuel a naguère été campé par J. Neyman et E. S. Pearson (voir Kendall et Stuart, 1979) : il s'agit des procédures de calcul grâce auxquelles le chercheur peut décider et affirmer que les différences observées dans ses données expérimentales sont significatives ou non. Schématiquement, ce cadre conceptuel inclut :

l'hypothèse nulle, symbolisée par H_0, et qui, en ignorant les effets expérimentaux possibles, exprime la valeur attendue des résultats sous le seul effet du hasard ;

la contre-hypothèse (ou hypothèse de recherche, ou hypothèse alternative), symbolisée par H_1, par laquelle le chercheur indique comment les effets expérimentaux se manifesteront dans les résultats ou comment l'hypothèse nulle sera contredite ;

la règle de rejet, qui met en jeu une statistique (p. ex. un *t*, un χ^2, un *z*, un *F*), un seuil de probabilité symbolisé par α et une (ou deux) valeur(s) critique(s), délimitant les zones d'acceptation et de rejet de H_0.

Prenons l'exemple suivant pour concrétiser ces notions. Un chercheur veut déterminer si l'ajout d'un supplément vitaminique à la diète de souris de laboratoire peut augmenter leur niveau d'activité. Pour cela, il utilise 20 souris, qu'il partage au hasard en deux groupes de n = 10 souris chacun, la diète du groupe Expérimental étant enrichie durant une semaine avant la prise de mesure, au contraire du groupe Témoin. Les souris sont placées dans une roue d'activité durant une heure : le nombre de tours comptés constitue la variable dépendante. Le chercheur anticipe des résultats plus élevés dans le groupe Expérimental.

Nous pouvons donc écrire, symboliquement :

$$H_0 : \mu_E = \mu_T \ (\text{ou } \mu_E - \mu_T = 0)$$
$$H_1 : \mu_E > \mu_T \ (\text{ou } \mu_E - \mu_T > 0) \qquad (1)$$

Rejet de H_0 au profit de H_1 si $t_{\overline{X}_E - \overline{X}_T} \geq t_{18[0,95]} = 1{,}734$.

La procédure de test, avec la formule de calcul indiquée par $t_{\overline{X}_E - \overline{X}_T}$ et ses conditions de normalité, additivité des effets et égalité des variances, est réputée respectée.

Le chercheur de notre exemple obtiendra-t-il un résultat significatif, c.-à-d. des données qui aboutiront à un *t* plus élevé que 1,734? En fait, il est impossible de répondre à cette question avec certitude, mais il arrive qu'on puisse en concocter une prédiction : c'est le calcul de puissance. En voici les grandes lignes théoriques.

Soit l'hypothèse nulle (H_0) est vraie. Dans ce cas, les niveaux de performance réels des deux populations (Expérimentale et Témoin) sont égaux, i.e. $\mu_E = \mu_T$, et la probabilité de rejeter par erreur H_0 est égale au seuil de signification, ici α = 0,05.

Soit l'hypothèse nulle (H_0) est fausse. Dans ce cas, sous H_1, le niveau de performance réel de la population expérimentale (μ_E) a *une* valeur spécifique[1], qui est plus grande que celle des Témoins (μ_T), avec une différence, ou effet, dénotée $\theta = \mu_E - \mu_T > 0$. La probabilité de rejeter H_0 est évidemment augmentée : elle dépend en fait de quatre paramètres clés : la différence réelle (θ) entre les populations, la variabilité intra-population (mesurée par l'écart-type σ), la taille échantillonnale (n) et le seuil de probabilité α. C'est cette probabilité qu'on appelle puissance statistique. En voici la définition technique usuelle :

(Définition usuelle de puissance statistique) (2)
La puissance statistique d'un test d'hypothèses est la probabilité qu'il rejette H_0 lorsque H_0 est fausse.

Mathématiquement, pour notre exemple, on peut en représenter le calcul par l'expression :

$$Pu = \Pr\{ t_{\overline{X}_E - \overline{X}_T} \geq t_{2n-2[1-\alpha]} \mid \theta, \sigma, n, \alpha \} \qquad (3)$$

Si les valeurs des quatre paramètres mentionnés (θ, σ, n, α) étaient connues, c.-à-d. connues du chercheur, il pourrait effectuer le calcul de puissance en se servant de l'un des nombreux outils mis à sa disposition à cette fin. L'obtention d'une puissance (*Pu*) de 0,50 lui indiquerait que, dans son expérience, il a une chance sur deux d'obtenir la conclusion significative qu'il souhaite. Avec *Pu* = 0,90, il serait presque sûr d'y parvenir alors que, avec *Pu* = 0,10, c'est presque du temps et de l'argent perdus pour lui que de persévérer. Qui plus est, la valeur de l'effet (θ) lui étant connue, à quoi lui servirait de réaliser l'expérience ou de faire le calcul, puisqu'il en connaîtrait déjà le résultat ultime?

En fait, pour la plupart des chercheurs et dans le plupart des cas, seul le facteur « seuil de signification α » est connu ou est prédéterminé dans une recherche, et la question du calcul estimatif de la puissance reste ouverte.

Nous proposons au lecteur d'examiner avec nous différents aspects de cette question, une question d'ailleurs peu ou pas débattue dans les grands traités de statistique. Sur la base de cet examen, nous serons à même de formuler quatre conclusions principales, des conclusions en partie polémiques parce qu'elles questionnent les conceptions et pratiques courantes des statisticiens appliqués et des

[1] Techniquement, l'hypothèse de recherche peut être simple (p.ex. $\mu_E = \mu_T + 4$) ou composée (p.ex. $\mu_E > \mu_T$), le premier cas aboutissant au calcul d'une puissance ponctuelle, le second à celui d'une courbe de puissance. Pour le chercheur, cependant, *son* hypothèse est intrinsèquement simple (en ce sens où la grandeur d'effet escomptée est une constante plutôt qu'un intervalle), mais il est ordinairement incapable de spécifier la valeur attendue de son résultat expérimental, d'où une certaine ambiguïté dans nos formulations.

chercheurs. Mais voyons d'abord les trois contextes d'information dans lesquels se présente le problème du calcul de puissance et de la détermination de la taille échantillonnale.

Contextes et enjeux du calcul de puissance

C'est à juste titre que les organismes publics ou privés qui financent la recherche se soucient de sa productivité et de ses chances de succès : pourquoi, en effet, accorder un important subside pour une entreprise dont la probabilité de conclure, c'est-à-dire la puissance, est trop faible? C'est la raison pour laquelle plusieurs organismes subventionnaires exigent d'entrée de jeu, dans la demande de subvention, un calcul de puissance et une justification du n, la taille d'échantillon proposée. C'est pourquoi aussi certains auteurs ont développé une méthodologie détaillée pour aider les chercheurs à mener à bien ce calcul : Cohen, avec son ouvrage « Statistical power analysis for the behavioral sciences » (1988), est sans doute le plus connu. D'autres auteurs se sont cantonnés dans l'autre volet de ce calcul, soit la détermination du nombre de sujets pour atteindre une puissance donnée : voir par exemple Desu et Raghavaro (1990) et Kraemer et Thiemann (1987). La méthodologie applicable suppose que l'on connaisse les valeurs réelles des facteurs θ, σ, n et α mentionnés, notamment l'effet θ et l'indice de variabilité σ, tels qu'ils se rapportent à la recherche envisagée. Que prévoit donc cette « méthodologie » pour le chercheur ? Desu et Raghavarao (1990), à l'instar des auteurs classiques en statistique (p. ex. Guenther, 1965 ; Hogg et Craig, 1978 ; Kendall et Stuart, 1979), posent comme connus les facteurs θ et σ (avec α) et, à partir de là, ils procèdent à la détermination de la taille n nécessaire pour atteindre une puissance donnée. Tout se passe comme si les auteurs voulaient faire une démonstration du calcul de puissance et de son utilité dans un contexte idéal, en faisant abstraction du contexte dans lequel le chercheur se trouve réellement. Où donc le chercheur trouvera-t-il les valeurs d'effet (θ) et de variabilité (σ) dont il a besoin?

Dans la plupart des cas, notamment en recherche expérimentale, c'est pour la première fois que le chercheur aborde une problématique particulière, ou qu'il l'aborde de cette façon particulière, avec son hypothèse et son type de mesure : rien n'existe à l'horizon qui puisse l'informer sur θ ou σ. Dans d'autres cas, cependant, le chercheur reprend une problématique déjà explorée, planifie une nouvelle étude sur la même question, et il peut disposer alors d'une information sérieuse sur les paramètres mentionnés. Par exemple, il a pu réaliser une première fois l'expérimentation suggérée plus haut sur l'effet d'un supplément vitaminique, et obtenir, disons, un test t = 1,200, non significatif (parce qu'inférieur à 1,734), à partir de deux groupes de n = 10 souris. Même s'il est non significatif, ce résultat l'informe néanmoins sur l'effet θ, par la différence empirique $\overline{X}_E$-$\overline{X}_T$ observée au numérateur du test t, et aussi sur la variabilité σ, par une partie de son dénominateur, $\sqrt{(s_E^2 + s_T^2)/2}$. Enfin, il existe aussi une dernière catégorie de cas, ceux des essais cliniques (EAEMP, 1998) et des applications industrielles, pour lesquels la valeur de l'effet θ est obtenue par prescription : cette valeur, notée $\tilde{\theta}$, est la valeur minimale exigible, en deçà de laquelle l'impact de l'effet θ dans le monde réel (biologique, physique, industriel) n'existe pas, et c'est cette valeur qui va servir pour calculer la puissance ou, équivalemment, la taille échantillonnale requise.

L'utilisation d'une valeur d'effet prescrite

Lenth (2001; voir aussi EAEMP, 1998) revoit en détail la procédure par laquelle une valeur de θ peut être prescrite (= $\tilde{\theta}$), en considérant d'une part les enjeux pratiques en cause dans le phénomène étudié et d'autre part les besoins de l'évaluation statistique de puissance. Dans un contexte biochimique, par exemple, il se peut qu'une augmentation d'un médiateur endocrinien de l'ordre de θ < 2 mmol soit insuffisante pour déclencher la réaction métabolique visée, alors qu'un effet plus grand, soit θ ≥ 2 mmol, y parviendrait. Le calcul de puissance peut alors tabler sur la quantité seuil $\tilde{\theta}$ = 2 mmol ; pour une puissance prédéterminée, le chercheur pourra fixer la taille minimale ($\tilde{n}$) exigible qui lui correspond.

Notons bien que la valeur prescrite $\tilde{\theta}$ ne permet pas d'établir la puissance réelle de l'expérience à faire : elle fait voir la puissance minimale que cette expérience aurait si l'effet réel θ rejoignait ou excédait la valeur prescrite. Supposons maintenant que l'effet réel existe mais qu'il est moins grand que l'effet minimal suffisant, soit $0 < \theta < \tilde{\theta}$. L'estimation de la taille échantillonnale, calculée d'après $\tilde{\theta}$ pour une puissance donnée, va fournir une valeur de taille $\tilde{n}$ *moindre* que celle qu'on aurait dû prendre pour faire apparaître l'effet réel θ ; la puissance effective (sous la vraie valeur θ) sera *moindre* que celle anticipée et l'expérience réalisée risque d'être décevante, non concluante.

L'utilisation d'une valeur d'effet prescrite ($\tilde{\theta}$) crée donc un paradoxe. D'une part, les auteurs cités plaident en faveur d'un tel procédé sur la base d'un argument réaliste (« Il est futile d'obtenir un résultat statistiquement significatif qui n'a pas d'impact réel, en pratique ») et d'un argument déontologique (« Il est inconvenant d'employer des sujets humains ou animaux dans une recherche qui n'a pas d'impact réel ») alors que, d'autre part, le recours à une puissance surestimée (si l'effet réel θ est moindre que l'effet prescrit $\tilde{\theta}$) rend toute l'expérimentation infructueuse. D'un autre côté, si l'effet réel (θ) se révélait être supérieur à la valeur minimale stipulée ($\tilde{\theta}$), la taille échantillonnale appliquée ($\tilde{n}$) serait excessive (avec une puissance plus

grande que celle prédéterminée), et l'argument déontologique (soit l'utilisation d'un nombre exagéré de sujets) s'appliquerait encore. En fait, il nous semble que c'est par erreur que le concept d'effet minimal, noté ici $\tilde{\theta}$, a été importé dans le domaine de l'évaluation de la puissance statistique, alors qu'il provenait d'un autre domaine, celui de la grandeur d'effet, et d'un autre débat, celui de la *différence expérimentalement significative* plutôt que *statistiquement significative* (Klerges 1982, Davidson 1994, Weber 1994, Rutledge et Loh 2004). Hoenig et Heise (2001) contestent d'ailleurs cette importation et re-justifient le concept d'effet (ou différence) cliniquement significatif, dans le contexte cette fois de la méthode des intervalles de confiance.

En conséquence, nous formulons ainsi notre proposition 1 : « Le recours à une valeur d'effet prescrite, selon un argument clinique ou de portée pratique, produit une estimation de puissance délibérément fausse, et qui contrevient de ce fait aux raisons déontologiques sur lesquelles il repose. »

L'invention pure et simple d'une grandeur d'effet

Quand le chercheur aborde pour la première fois une problématique, comme il arrive souvent en recherche expérimentale, le comportement de ses mesures (sa « variable dépendante ») ne lui est pas connu, non plus que la grandeur d'effet escomptée à partir des conditions d'expérience imposées (sa « variable indépendante ») : il n'a donc aucune connaissance préalable ni de l'effet θ, ni de la variabilité intrinsèque σ. Peut-il tout de même estimer la puissance?

Les organismes bailleurs de fonds, au Québec et au Canada comme aux États-Unis, exigent de tous les candidats qui soumettent un projet pour subvention une estimation de puissance et la justification de la taille échantillonnale (n) proposée. Or, nonobstant l'ignorance de θ et σ, Cohen (1988), avec d'autres qui s'en inspirent (p.ex. Kraemer et Thiemann 1987), croient qu'une telle estimation est possible et ils se portent au secours du chercheur en mal de réponse. Pour le contexte statistique évoqué par notre exemple, soit la comparaison du taux d'activité chez deux groupes de souris, les souris du groupe expérimental recevant un supplément vitaminique, Cohen propose un indice d'écart standardisé **d**, défini par :

$$\mathbf{d} = \frac{\mu_E - \mu_T}{\sigma} = \frac{\theta}{\sigma} \; ;$$

tel qu'indiqué, les ingrédients de la formule sont les constantes de population (μ_E, μ_T, σ) plutôt que des estimations statistiques basées sur l'expérimentation. Cohen élabore l'interprétation de son indice **d** de différentes et intéressantes manières. Ainsi, quand **d** = 0 (un effet nul), il imagine que les distributions (normales) des deux populations comparées se recoupent complètement alors que, pour un écart égal à **d** = 0,50, environ 60% de la population supérieure (i.e. expérimentale) dépasse 60% de l'autre population (i.e. témoin). De plus, puisque l'indice **d** standardise l'effet (ou écart) θ en le divisant par σ, son impact sur la puissance est direct, ce qui permet à Cohen de présenter des tables de puissance basées d'une part sur l'indice **d** et d'autre part sur n.

L'approche de Cohen (1988), toute pédagogique qu'elle soit, n'a consisté jusqu'à présent qu'à remplacer deux paramètres, θ et σ, par un indice unique, **d** = θ / σ, basé encore sur les valeurs réelles de θ et σ, et *dont la valeur reste à trouver* (Lenth 2001). Pour « faciliter » la réflexion du chercheur, qui se croit obligé de calculer la puissance, et en raison du lien étroit entre l'indice standardisé **d** et la puissance, Cohen (p. 24-27) met de l'avant les catégories verbales suivantes, soit que le chercheur croit avoir affaire à un effet « petit » avec **d** = 0,20, « moyen » avec **d** = 0,50, ou « grand » avec **d** = 0,80. Que propose alors Cohen (1988) ? Il propose d'*inventer* purement et simplement **d**, à partir d'un « raisonnement » métaphorique, qui nous a paru saugrenu. En voici deux exemples.

« Une psychologue expérimentale planifie une expérience afin de mesurer l'impact que peut avoir pour des rats l'opportunité d'explorer sans récompense un labyrinthe sur l'apprentissage ultérieur de ce labyrinthe. Des échantillons de 30 animaux chacun sont formés et assignés au groupe expérimental (E), qui bénéficie d'une phase exploratoire, et au groupe témoin (T), sans chance d'exploration. Les rats sont ensuite testés, le nombre d'essais requis pour deux parcours consécutifs parfaits constituant la variable retenue. L'hypothèse H_0 (non-directionnelle) est : $|\mu_E - \mu_T| = 0$. **La chercheuse prévoit que l'effet sera tel que les meilleurs 60% d'une population dépasseront les 60% moins bons de l'autre**, soit l'indice $U_2 \approx 60\%$. Référant à la table 2.2.1 (en p. 22), elle trouve que $U_2 = 59{,}9\%$ équivaut à un effet moyen, **d** = 0,50. Donc, l'hypothèse de recherche (H_1) spécifie que les moyennes des deux populations s'écartent d'un demi écart-type (ou $\theta = \frac{1}{2}\sigma$) l'une de l'autre... D'où, pour un test bilatéral au seuil $\alpha = 0{,}05$ (avec $n = 30$ sujets par groupe),... la puissance égale 0,47. » [Cohen 1998, pp. 40-41 ; les caractères gras sont de nous.]

« Un chercheur en psychiatrie, intéressé à certains facteurs endocriniens impliqués dans la schizophrénie, fait une expérience pour comparer les échantillons d'urine de 500 schizophrènes et de 500 normaux équivalents, y mesurant un produit métabolique connu pour se distribuer à peu près normalement et avec variabilité constante dans la population. **Étant donné que le facteur endocrinien soupçonné n'est qu'indirectement lié au produit métabolique dans l'urine et peut-être pour d'autres raisons, le chercheur n'espère qu'un petit effet,**

spécifiquement d = 0,20. ... En conséquence, pour **d** = 0,20 et n = 500 (par groupe), *Pu* = 0,72. » [Cohen 1988, pp. 41-42 ; les caractères gras sont de nous.]

A-t-on déjà vu un chercheur qui puisse affirmer que 60% des sujets expérimentaux auront des résultats débordant ceux de 60% des sujets témoins, voire un chercheur qui réfléchisse à son expérience en ces termes? Quant à la seconde citation, le « petit effet » escompté en dit sans doute plus long sur la prudence ou la circonspection du chercheur que sur l'effet réel présent, un effet (θ), voire un effet standardisé (**d**) dont ni un chercheur ni l'autre ne connaît la valeur.

En résumé, hormis le cas d'une grandeur d'effet prescrite (sur la base de la signification clinique ou de sa portée pratique) discuté plus haut, les auteurs recensés traitent de puissance en termes de paramètres de population, notamment θ et σ (ou **d**) et leurs vraies valeurs. Or, dans le but évident de dépanner le chercheur qui, souvent, n'a pas d'idée précise sur les grandeurs numériques à espérer dans son projet expérimental, Cohen (1988) suggère de baser le calcul sur un effet imaginaire, catégorisé verbalement (petit, moyen, grand), c'est-à-dire sur l'appréhension que le chercheur a de ses résultats futurs. Une fois l'effet « verbo-numérique » (**d**) ainsi fixé, le reste est affaire de calcul, qui aboutit, comme nous venons de le démontrer, à une valeur totalement fictive de puissance.

D'où notre proposition 2 : « Il est illusoire et peut être mensonger de calculer une puissance qui n'est pas basée sur des données du domaine de recherche ».

L'estimation de puissance à partir de données expérimentales

Il arrive tout de même que, dans un projet donné, un chercheur reprenne une problématique ou une partie d'expérimentation qui a déjà été explorée et qui a produit des données empiriques crédibles. C'est notamment le cas des « confirmatory trials » en recherche médicale, pour lesquels l'Agence Européenne pour l'Évaluation des Médicaments indique : « assumptions should normally be based on published data or on the results of earlier trials » (EAEMP, 1998, p. 19). Hélas, et soulignons-le, la documentation recensée ne fait pas état de l'exploitation possible de tels résultats empiriques pour *estimer* la puissance. Seul, le classique Guilford (1965) y va d'une suggestion, en proposant d'estimer la taille n (par groupe) nécessaire *pour rendre le test obtenu tout juste significatif*, c'est-à-dire pour que le test égale tout juste la valeur critique. Notre exemple comportait au départ deux groupes égaux, avec n = 10 sujets chacun et *t* = 1,200. En manipulant n, nous aurions à résoudre :

$$\frac{\bar{X}_E - \bar{X}_T}{\sqrt{\frac{s_E^2 + s_T^2}{2}\left(\frac{2}{n}\right)}} \geq t_{2n-2[0.95]},$$

obtenant éventuellement $n \geq 20$, ce qui suggère ici de doubler la taille de chaque groupe. Évidemment, le test incluant les 10 sujets supplémentaires par groupe ne deviendra peut-être significatif que si les tendances observées sur les premiers sujets se maintiennent. Kendall et Stuart (1979), dans un autre contexte, mentionnent qu'une telle égalisation du test sur sa valeur critique (à partir, cette fois, des valeurs théoriques de θ et σ) correspond à une puissance (*Pu*) de 0,50 : nous revenons plus loin sur cette « correspondance médiane » entre puissance et paramètre θ (ou δ). Est-il possible d'aller plus loin et de proposer une méthodologie complète de l'estimation empirique de puissance à partir de données expérimentales?

Nous répondons par l'affirmative à la question posée ci-dessus. Ne fait-on pas de même lorsque, à partir d'un petit échantillon de données, on bâtit un intervalle de confiance pour une statistique d'intérêt, ou lorsqu'on exploite la statistique obtenue dans l'échantillon (p. ex. $\bar{X}$, s^2, *r*) comme une estimation de la valeur paramétrique correspondante (p. ex. μ, σ^2, ρ) ? C'est précisément à ce but que servent les données statistiques : estimer les paramètres de la population étudiée.

Dans le cas de la puissance, l'extrapolation d'information à partir des données préalables requiert les trois étapes que voici : 1) obtenir (réunir) des données numériques préalables ; 2) estimer le ou les paramètres de population appropriés ; 3) exploiter les estimations faites pour effectuer une projection de puissance. Même s'il est rigoureusement conçu et appliqué, ce processus comporte de l'incertitude. L'incertitude, c'est-à-dire la variabilité statistique dans ce processus, tient essentiellement aux données de l'étape 1, particulièrement à la taille d'échantillon utilisée : plus nombreux sera l'échantillon, plus sûres seront les estimations et les projections établies sur sa base. L'incertitude dans l'échantillon se répercute directement dans l'estimation, à l'étape 2 ; elle engendre une marge d'imprécision calculable autour de la valeur estimée. C'est la même incertitude qui transite des étapes 2 à 3, sans plus : rien n'y est ajouté, de sorte que, à incertitude égale, les trois étapes du calcul de puissance sont aussi valables et ni plus ni moins sûres que l'estimation d'une moyenne ou l'établissement d'un intervalle de confiance.

Logiquement donc, il est possible d'estimer la puissance à partir de données préalables. Or, nous l'avons déjà dit, la documentation recensée ne rapporte aucune méthode pour instrumenter cette estimation. Dans une recherche récente (Laurencelle, 2005), nous nous sommes penché sérieusement sur cette question, en considérant le problème d'estimation sous l'angle d'un calcul de puissance. Cette recherche s'appuyait sur une logique d'estimation en quelques étapes, dont voici la description sommaire :

1) Dans le cadre d'une situation de recherche et d'un test

d'hypothèses donnés, H_0 stipule que la statistique utilisée, p. ex. $t_{\overline{X}_E - \overline{X}_T}$, se distribue selon une loi centrale appropriée, p. ex. le *t* de Student, alors que H_1, indiquant une valeur différente d'un paramètre, fait implicitement appel à une loi ou distribution non-centrale correspondante, p. ex. le $t(\delta)$ non-central, avec le paramètre de non-centralité δ.

2) Le calcul de puissance supposant la vérité de H_1, nous posons que notre statistique (p. ex. $t_{\overline{X}_E - \overline{X}_T}$), obtenue à partir de données préalables, émane de la distribution non-centrale appropriée, avec une valeur particulière du paramètre de non-centralité (δ).

3) Il est possible d'estimer la valeur du paramètre de non-centralité (δ) à partir de la statistique obtenue ($t_{\overline{X}_E - \overline{X}_T}$).

4) Connaissant la valeur estimative du paramètre de non-centralité ($\hat{\delta}$), il est possible de calculer la puissance correspondante en invoquant la distribution non-centrale appropriée (Johnson, Kotz et Balakrishnan, 1995 ; Kendall et Stuart, 1979), soit p. ex. :

$$Pu = \Pr\{\, t_{2n-2}(\hat{\delta}) \geq t_{2n-2[1-\alpha]} \}\,; \qquad (4)$$

c'est l'expression (3) qui est reprise ici, dans sa forme calculable, et en appliquant le paramètre $\delta = \theta / \sigma \times \sqrt{n}$, le quotient θ/σ étant obtenu par estimation.

Le point charnière de cette approche tient à l'estimation du paramètre de non-centralité, à l'étape 3, estimation sur laquelle la littérature est restée pratiquement muette (Johnson et coll., 1995).

Dans l'étude rapportée, Laurencelle (2005) considère plusieurs types d'estimateurs, notamment un estimateur en espérance (la valeur de δ telle que la moyenne du t_δ soit égale à la statistique $t_{\overline{X}_E - \overline{X}_T}$ obtenue) et un estimateur en médiane (la valeur de δ telle que la *médiane* du t_δ soit égale à la statistique $t_{\overline{X}_E - \overline{X}_T}$ obtenue)[2]. L'auteur procède alors à des expérimentations Monte Carlo pour étudier la justesse et la précision des différents estimateurs et des puissances correspondantes. Grosso modo, le procédé Monte Carlo consiste à définir une situation paramétrique particulière sous H_1, avec différentes valeurs fixées de θ, σ et n (donc de puissance réelle), puis à générer des échantillons au hasard dont on tire ensuite des valeurs estimatives de $\hat{\delta}$ et de puissance.

Les conclusions de l'étude, qui portait sur quatre tests (le test *t* sur une moyenne, le test *F* d'analyse de variance sur 4 groupes, le test Khi-deux sur une variance et le test sur une corrélation), peuvent être résumées comme suit :

– tous les estimateurs (du paramètre de non-centralité δ ou θ) et leurs puissances associées covarient avec les valeurs réelles correspondantes ;

– tous les estimateurs (sauf un, l'estimateur modal pour la corrélation) sont consistants, c'est-à-dire que leur moyenne et leur médiane se rapprochent de la valeur cible quand la taille échantillonnale croît (notons que les estimateurs diffèrent généralement l'un de l'autre, étant donné que les distributions non-centrales dont ils proviennent, soit $t(\delta)$, $\chi^2(\lambda)$ et $F(\lambda)$, sont asymétriques) ;

– l'estimateur en médiane (voir Birnbaum, 1964) estime quasi exactement (sans biais médian) la valeur cible, de même que sa puissance associée estime quasi exactement la puissance réelle.

Bien sûr, il faut se rappeler la mise en garde faite d'entrée de jeu à propos de l'incertitude d'un calcul basé sur des données expérimentales : il s'agit d'une puissance *estimée* (puisque calculée à partir de valeurs estimées des paramètres θ et σ) et, comme telle, elle contient de l'incertitude (Laurencelle, 2005). Il reste que, répétons-le, la puissance *peut être estimée* à partir des données d'expérience et il nous paraît absurde de la calculer en dehors de cette base.

Nous pouvons donc formuler notre proposition 3 : « le chercheur qui possède de l'information numérique préalable sur son projet peut s'en servir utilement pour en estimer la puissance ».

La puissance d'un test statistique bilatéral

Dans la planification de son expérience et si on lui demande à quels résultats il s'attend, le chercheur peut généralement répondre de l'une des trois façons suivantes :

– soit il croit fermement que l'effet, s'il existe, ne peut se produire que dans un seul sens (p. ex. en augmentant comparativement les résultats du groupe expérimental), toute autre variation étant dépourvue d'intérêt ;

– soit il espère que l'effet se produise dans le sens prévu mais, vu la complexité du phénomène étudié, il accepte de considérer une variation sérieuse en sens contraire ;

– soit il n'en sait trop rien (comme c'est parfois le cas en recherche descriptive), et il acceptera joyeusement toute variation sérieuse, dans un sens ou dans l'autre.

Certains statisticiens prônent l'utilisation quasi exclusive des tests bidirectionnels (ou bilatéraux), parce que plus prudents, et d'autres font la nuance entre les trois situations présentées, en réservant le test unidirectionnel (ou unilatéral) à la première situation et en imposant les tests bidirectionnels dans les deux autres. Toujours est-il que les tests d'hypothèses bidirectionnels existent, qu'ils sont pratiqués et qu'on peut souhaiter en établir la puissance.

Pour illustrer notre propos, prenons un nouvel exemple, utilisant cette fois-ci le test *z* sur une moyenne, avec variance connue. Dans le but de planifier une campagne de publicité

[2] L'auteur considère aussi d'autres estimateurs, notamment les estimateurs « simples » tels que ceux utilisés pour quantifier la grandeur d'effet, p. ex. $\hat{\delta} = t$ (= 1,200), dans le cas de notre exemple.

nationale, une firme de consultants en marketing se propose d'évaluer le niveau d'intelligence moyen des personnes (adultes) qui circulent dans les centres commerciaux du pays entre 14 et 18 heures, en semaine. Pour ce faire, ils utilisent un test d'intelligence, le test Mon_QI, standardisé dans la population générale et doté d'une moyenne normative (μ) de 100, d'un écart-type (σ) de 15 et d'une distribution normale. L'enquêteur n'a ni préconception ni théorie sur la question ; il se demande seulement si la sous-population circulant dans les centres commerciaux se démarque ou non de la population générale.

Pour cet exemple, nous pouvons donc écrire :

$$H_1 : \mu_{[Clients]} < 100 \text{ ou } \mu_{[Clients]} > 100$$

$$H_0 : \mu_{[Clients]} = 100 \qquad (5)$$

Rejet de H_0 au profit de H_1 au seuil $\alpha = 0{,}05$ si

$$z(\bar{X}) \geq z_{[0{,}975]} = 1{,}960 \text{ ou } z(\bar{X}) \leq z_{[0{,}025]} = -1{,}960.$$

Tous auront reconnu le test sur une moyenne, $z(\bar{X}) = \sqrt{n}\left((\bar{X} - \mu)/\sigma\right)$, avec $\mu = 100$ et $\sigma = 15$. Supposons maintenant qu'on recrute et mesure quelque 25 ($= n$) clients, dans différents centres commerciaux. Comment, ici, évaluer la puissance de ce test bidirectionnel?

Attendu les considérations faites plus haut, nous pouvons reconnaître ici que soit nous n'avons pas d'information préalable sur l'effet θ escompté, c.-à-d. la différence μ[Clients] – 100, et alors il nous faudrait inventer un effet de toutes pièces, soit nous possédons des données préalables qui nous permettent d'*estimer l'effet* à partir de ces données. Admettons pour l'instant l'un et l'autre cas (même si le premier n'est pas scientifiquement admissible), et nous obtenons une valeur inventée ou estimée, disons $\theta = 2$ (en supposant que la clientèle touchée forme une sous-population de moyenne $\mu_{[Clients]} = 102$).

Une fois connue la valeur $\theta = 2$, nous pouvons procéder à l'évaluation de la puissance, en calculant d'abord le paramètre de non-centralité, $\delta = \theta / \sigma \times \sqrt{n} = 2 / 15 \times \sqrt{25} \approx 0{,}667$. Quelle est la puissance de ce test?

La méthode traditionnelle pour évaluer la puissance d'un test bidirectionnel (voir p. ex. Cohen 1988) consiste à observer rigoureusement la définition de puissance (2) déjà donnée, et à évaluer ici, par exemple (Φ dénote l'intégrale normale standard) :

$$\begin{aligned} Pu &= \Pr\{ z(\delta) \geq 1{,}960 \} + \Pr\{ z(\delta) \leq -1.960 \} \qquad (6) \\ &= \Phi(+\delta - 1{,}960) + \Phi(-\delta - 1{,}960) \\ &= \Phi(-1{,}293) + \Phi(-2{,}627) \\ &= 0{,}09801 + 0{,}00431 = 0{,}10232. \end{aligned}$$

La valeur obtenue, 0,10232, provient de deux parties : une partie plus importante (0,09801), celle située du côté favorable de l'effet (ou + δ), et une contrepartie (0,00431), moins importante, située du côté opposé (– δ). En combinant ces deux parties en un seul tout, la puissance résultante doit être interprétée comme la probabilité, quand H_0 est fausse, de se prononcer en faveur de la différence annoncée par H_1 *ou pour la différence exactement opposée*. Formellement, une telle *puissance bilatérale* serait définie comme :

(Définition explicite de puissance bilatérale) (7)

La puissance statistique d'un test d'hypothèses bidirectionnel est la probabilité qu'il rejette correctement ou incorrectement H_0 lorsqu'elle est fausse.

Or, pour tout chercheur comme pour tout organisme subventionnaire qui s'intéresse à la puissance, cette dernière est associée à l'idée et à la probabilité qu'un test d'hypothèses nous amène à prendre *la bonne décision* ; la part de probabilité qui concerne une décision exactement contraire à la réalité n'a pas de place dans ce concept.

Ces considérations nous conduisent à proposer la définition générale suivante de puissance, applicable aux tests unidirectionnels comme aux tests bidirectionnels :

(Définition universelle de puissance statistique) (8)

La puissance statistique d'un test d'hypothèses est la probabilité qu'il rejette correctement H_0 lorsque H_0 est fausse.

Cette nouvelle définition n'a de conséquence pratique que dans le cas des tests d'hypothèses bidirectionnels, et elle nous indique de ne retenir dans la puissance bilatérale que sa partie congrue, c'est-à-dire la probabilité que le test soit significatif dans la direction indiquée par la différence réelle. Pour le test donné en exemple ci-dessus, θ (= 2) étant positif, nous ne retiendrons donc que la portion positive, Pr{ $z(\delta) \geq 1{,}960$ }, obtenant une puissance effective de 0,09801.

Ainsi, dans le cas de tests bidirectionnels, l'utilisation de la seule partie congrue de probabilité fait en sorte que, pour un test effectué au seuil α, la puissance minimale (lorsque $\delta \to 0$) est de α/2 au lieu de α : cette perte initiale de puissance est le prix à payer pour la réalisation d'un test plus prudent. À toutes fins pratiques, cependant, la différence quantitative entre un mode de calcul et l'autre est quasi nulle (sauf pour $\delta < 0{,}5$) : cela devrait réconforter les chercheurs qui craindraient de « perdre de la puissance » en adoptant la définition universelle présentée ici.

Nous pouvons donner enfin notre proposition 4 : « le calcul de la puissance d'un test d'hypothèses bidirectionnel ne doit pas inclure la partie controlatérale de probabilité ».

Épilogue et recommandations

Dans cet article, qui comporte évidemment des points de polémique, nous avons voulu faire le tour de la question de puissance statistique telle qu'elle s'applique au chercheur et à la planification d'expérience plutôt que simplement comme concept mathématique abstrait. Le calcul de puissance est, comme tel, un procédé purement déductif. Cependant, lorsque le chercheur veut trouver la taille échantillonnale (n) à prévoir afin d'avoir de bonnes chances de démontrer un effet (θ) significatif, il se situe en mode

inductif, empirique, et les paramètres du calcul de puissance, p. ex. θ et σ, doivent alors être estimés.

Les travaux de Laurencelle (2005) rapportés ci-dessus ne sont qu'un premier pas dans l'étude de l'estimation du paramètre de non-centralité (θ/σ ou δ) et de la puissance, un premier pas encourageant et qui touche spécifiquement les tests utilisant les lois normale, *t*, χ^2 et *F* ainsi que l'estimation du coefficient de corrélation. Ces travaux ont montré que, pour les cas cités, il est possible au chercheur d'estimer la puissance pour une situation de recherche projetée à condition qu'il dispose de données préalables pertinentes, l'« estimateur médian » montrant à cet effet les caractéristiques les plus avantageuses.

Pour conclure, enfin, nous reprenons les arguments qui ont fait la trame de cet article polémique pour les reformuler, en ordre inverse, sous la forme de trois recommandations.

Recommandation 1

Le concept même de « puissance bilatérale » est aporétique, puisqu'il incorpore, dans la probabilité de prendre la bonne décision du rejet de H_0, une partie qui consiste à se prononcer en faveur d'un effet exactement contraire à l'effet réel. Dans tous les cas, notamment ceux des tests d'hypothèses bidirectionnels, la puissance est la probabilité de rejeter la valeur paramétrique associée à H_0 en faveur de celle associée à H_1 (sans y inclure la valeur opposée).

Recommandation 2

Le chercheur qui a en mains des données préalables peut s'en servir pour estimer la puissance statistique encourue dans son projet de recherche et en tirer parti, par exemple, en déterminant les tailles d'échantillon qui lui promettront une puissance satisfaisante. Pour ce faire, les indications trouvées dans Laurencelle (2005) pour estimer notamment le paramètre de non-centralité sont à considérer. Une fois la valeur de ce paramètre déterminée, les ouvrages publiés, et même celui très élaboré de Cohen (1988), peuvent prendre le relais.

Recommandation 3

Sans données préalables pertinentes qui l'informent numériquement sur sa recherche planifiée, le chercheur ne peut pas et ne doit pas s'essayer à un calcul de puissance, fût-ce sur la base d'un effet qui serait « cliniquement significatif » sans pour autant être empiriquement fondé. Agir ainsi serait résolument anti-scientifique, voire mensonger, et il est difficile de comprendre pourquoi des auteurs comme Cohen (1988) et Lenth (2001) et, faut-il le dire, certains organismes subventionnaires encouragent cette duperie.

Références

Birnbaum, A. (1964). Median-unbiased estimators. *Bulletin of mathematical statistics*, 11, 25-34.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2e édition). Hillsdale (NJ) : Lawrence Erlbaum Associates.

Davidson, R.A. (1994). Does it work or not? Clinical vs. statistical significance. *Chest*, *106*, 932-934.

Desu, M.M., Raghavarao, D. (1990). *Sample size methodology*. Boston : Academic Press.

EAEMP (1998). Statistical principles for clinical trials – ICH Topic 9. Document internet du European Agency for the Evaluation of Medicinal Products, à l'adresse : www.emea.eu.int/pdfs/human/ich/036396en.pdf.

Guenther, W.C. (1965). *Concepts of statistical inference*. New York : McGraw-Hill.

Guilford, J.P. (1965). *Fundamental statistics in psychology and education* (4e édition). New York : McGraw-Hill.

Hoenig, J.M., Heise, D.M. (2001). The abuse of power: the pervasive fallacy of power calculations in data analysis. *The American Statistician*, *55*, 19-24.

Hogg, R.V., Craig, A.T. (1978). *Introduction to mathematical statistics* (4e édition). New York : Macmillan.

Johnson, N.L., Kotz, S., Balakrishnan, N. (1995). *Continuous univariate distributions* (2 tomes) (2e édition). New York : Wiley.

Kendall, M.G., Stuart, A. (1979). The advanced theory of statistics. Volume 2 : Inference and relationship (4e édition). New York : Macmillan.

Klerges, R.C. (1982). Confusing clinical and statistical significance? A reply to Reynolds and Gutkin. *Journal of consulting and clinical psychology*, *50*, 772-774.

Kraemer, H.C., Thiemann, S. (1987). *How many subjects?* Newbury Park : SAGE.

Laurencelle, L. (2005). L'estimation de la puissance statistique à partir de données d'expérience. *Lettres Statistiques*, *12*, 15-36.

Lenth, R.V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*, 187-193.

Rutledge, T. et Loh, C. (2004). Effect sizes and statistical testing in the determination of clinical significance in behavioral medicine research. *Annals of behavioral medicine*, *27*, 138-145.

Weber, K. (1994). On contrasting clinical and statistical significance. *Pediatric pulmonology*, *18*, 64-65.

Manuscript received October 26th, 2007
Manuscript accepted September 13th, 2007