Inférence statistique causale sur les effets individuels : Quelques éléments de réflexion

Bruno Lecoutre[1] et Jacques Poitevineau[2]

Introduction

Même si beaucoup d'analyses causales se focalisent sur l'*effet causal moyen* – typiquement une différence de moyennes entre deux conditions expérimentales – l'objet fondamental de l'inférence causale devrait être l'*effet causal individuel*. L'effet causal individuel est typiquement une comparaison de deux résultats *potentiels*, dont un seul a été effectivement observé. Mais en fait, même quand on a recueilli des données très nombreuses (d'où une incertitude *purement statistique* négligeable), et même dans le cas d'une expérimentation contrôlée (« randomisée »), on ne peut faire *aucune inférence sans ambiguïté* sur l'effet causal individuel sans faire de suppositions non testables sur le modèle d'échantillonnage. C'est ce que nous illustrerons ici, en nous appuyant sur les travaux de Dawid (2000) ; en conclusion nous résumerons brièvement les conséquences de ce constat (pour une présentation détaillée, *cf* Lecoutre, 2004).

Préliminaires : L'expérience de base randomisée

Dans l'expérience de base randomisée, on compare deux traitements possibles *t1* et *t2*, par exemple un nouveau traitement et un traitement de référence. Un concept (hypothétique) essentiel est celui de *population*, c'est-à-dire un ensemble *U* d'*unités* (« individus ») *u*. On suppose que l'on dispose de deux *groupes* (sous-ensembles) d'unités, auxquels on applique respectivement les traitements *t1* et *t2*. Le choix du traitement appliqué est effectué par tirage au sort (« *randomisation* »). L'expérience consiste à observer la réponse de chaque unité expérimentale au traitement. Il est fondamental de noter qu'une fois qu'un traitement a été appliqué à une unité, l'autre traitement *ne peut plus lui être appliqué* (du moins dans des

1 ERIS, LMRS, UMR 6085, C.N.R.S. et Université de Rouen, Site Colbert, 76821 Mont-Saint-Aignan Cedex. bruno.lecoutre@univ-rouen.fr, http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris.htm
[2] ERIS, LAM/LCPE, CNRS, 11 rue de Lourmel, 75015 Paris. jacques.poitevineau@ivry.cnrs.fr

conditions *semblables*). Cela nécessite une définition appropriée de l'unité expérimentale (et donc de la population), qui doit représenter *un certain état du sujet* et non le sujet lui-même. Il en résulte que la *réplication* des observations est *impossible* (sauf situation très particulière). Un concept essentiel est celui d'unités expérimentales indistinguables. Le traitement affecté à une unité ne doit dépendre d'aucune information qui pourrait permettre de la distinguer des autres unités. De fait, les unités utilisées n'ont pas d'intérêt particulier, mais on s'intéresse à une inférence sur des propriétés génériques des unités sous l'effet des traitements.

Les effets des causes et les causes des effets

On peut distinguer avec Dawid (2000) deux types de questions causales : (1) « J'ai mal à la tête. Est-ce que cela me soulagera si je prends de l'aspirine ? » et (2) « Mon mal de tête est parti. Est-ce parce que j'ai pris de l'aspirine ? ». Ces questions correspondent à deux problèmes valides et importants : (1) l'inférence sur les *effets des causes*, qui consiste à comparer des conséquences attendues de différentes interventions possibles dans un système - c'est le centre d'intérêt des analyses statistiques usuelles dans l'expérimentation ; (2) l'inférence sur les *causes des effets*, qui vise à comprendre la relation causale entre un résultat déjà observé et une intervention antérieure.

Ces deux types de questions nécessitent des *analyses différentes* (bien que liées). L'inférence *prédictive* permet d'apporter une distinction claire. Supposons qu'une expérience a été réalisée. Considérons une *unité test* nouvelle u_0, issue de la même population et n'ayant pas encore donné lieu à observation, à laquelle on se propose d'appliquer – sur la base des résultats de l'expérience – l'un des deux traitements *t1* ou *t2*. Si on choisit *t1* on obtiendra l'observation $Y_{t1}(u_0)$, et si on choisit *t2* on obtiendra l'observation $Y_{t2}(u_0)$. L'inférence sur les effets des causes consiste en une prédiction sur $Y_{t1}(u_0)$ et $Y_{t2}(u_0)$ *avant toute intervention* ; un aspect important est ici la prise de *décision* sur le traitement à affecter à u_0. L'inférence sur les causes des effets présuppose qu'un traitement particulier (disons *t1*) a été choisi et *déjà appliqué* à u_0 ; le résultat $Y_{t1}(u_0) = y_{t1}$ *a été observé*. Il s'agit alors de chercher à répondre à la question : est-ce que l'application de *t1* « *a causé* » le résultat observé pour u_0 ?

Tableaux et modèles physique et métaphysique

Notons $u\langle t\rangle$ (t=*t1*,*t2*) les unités ayant reçu (effectivement) le traitement t (l'étiquetage est arbitraire puisque les unités sont indistinguables). L'expérience associe à chaque unité $u\langle t\rangle$ une observation $X_t(u)$, d'où le tableau *physique*, qui est la collection des observations et qui correspond à une plan en deux groupes indépendants ($U\langle T\rangle$).

Dans la conception traditionnelle, qui remonte (au moins) aux travaux statistiques de Fisher et de Neyman sur les expériences agricoles, une notion de base est celles de *résultat potentiel*. Idéalement, on voudrait comparer les résultats obtenus pour le traitement *t* avec les résultats qu'on aurait obtenus pour *les mêmes sujets*, s'ils avaient été soumis *à l'autre traitement*, *toutes choses égales* par ailleurs (clause *ceteris paribus*). La procédure d'inférence repose sur ce que les philosophes appellent l'*induction par élimination* : toute différence est *causée* par le traitement. Elle invoque une interprétation *contrefactuelle* : « Si *A* avait existé,

alors *B* se serait produit », mais il est implicite que *B* n'a pas eu lieu. La procédure nécessite donc la comparaison d'un résultat *réel* et d'un résultat *contrefactuel* ; nous devons donc considérer le tableau *métaphysique*, qui est la collection de tous les résultats potentiels (que nous devons « imaginer »), et qui correspond à un plan en deux groupes appariés ($U \times T$). Pour *chaque* unité u (quel que soit le traitement effectivement reçu), ce tableau inclut les deux résultats possibles, $Y_{t1}(u)$ et $Y_{t2}(u)$, et comporte donc, en empruntant la terminologie de la physique quantique, un grand nombre de variables « complémentaires », c'est-à-dire non observables simultanément. Pourtant l'analyse contrefactuelle repose sur la considération de tous les $Y_{t1}(u)$ et $Y_{t2}(u)$ *simultanément*.

Nous considèrerons les modèles statistiques traditionnels de l'analyse de variance (les conséquences sur l'inférence seraient les mêmes pour tout autre modèle). Dans le modèle physique, les variables $X_{u<t>}$ sont indépendantes et équidistribuées, de distribution normale avec moyenne μ_t et variance σ^2 (égalité des variances). L'analyse usuelle repose uniquement sur ce modèle physique ; au contraire l'analyse contrefactuelle nécessite un modèle *métaphysique* pour le tableau des résultats potentiels : les couples de variables $(Y_{t1}(u), Y_{t2}(u))$ sont indépendantes et équidistribués, de distribution normale bivariée avec moyenne $[\mu_{t1}, \mu_{t2}]$, variance σ^2 (pour chaque variable) et corrélation ρ (ou covariance $\rho\sigma^2$). On remarquera que le modèle physique peut être *dérivé* du modèle métaphysique.

Effet causal moyen et effet causal individuel

La plupart des analyses de variance se limitent à des comparaisons de moyennes, c'est-à-dire ici à une inférence sur la différence des moyennes : $\delta = \mu_{t1} - \mu_{t2}$ (*Effet Causal Moyen* ou ECM). Or l'objet fondamental de l'inférence causale est (ou devrait être) l'effet causal individuel, ce qui nécessite une inférence sur la différence pour chaque unité u : $D(u) = Y_{t1}(u) - Y_{t2}(u)$ (*Effet Causal Individuel* ou ECI) ; d'autres définitions sont envisageables : $\log Y_{t1}(u) - \log Y_{t2}(u)$; $Y_{t1}(u) / Y_{t2}(u)$; etc. Mais, quelle que soit la définition, l'ECI est intrinsèquement *inobservable* puisqu'il repose sur la comparaison de quantités complémentaires.

Pour le modèle marginal relatif à l'effet causal individuel $D(u)$ (dérivé du modèle métaphysique), $D(u)$ a une distribution normale avec moyenne δ et variance $\sigma_D^2 = 2(1-\rho)\sigma^2$. Les conséquences sur l'inférence sont les suivantes (rappelons que nous supposons l'absence d'incertitude statistique liée aux observations). Le tableau et le modèle physiques fournissent des estimations précises des moyennes μ_{t1} et μ_{t2} (donc de l'ECM δ) et de la variance σ^2, mais la corrélation ρ est *non identifiable*. Même si une analyse contrefactuelle superficielle (limitée à l'ECM) peut apparaître ici fondamentalement valide et conduire à des conclusions univoquement acceptables, cela n'est plus cas avec un changement de définition : ainsi si l'ECI est défini comme le rapport $Y_{t1}(u)/Y_{t2}(u)$, sa moyenne n'est plus déterminée par les paramètres du modèle physique.

L'inférence sur les effets des causes

Rappelons que l'on considère une unité test nouvelle u_0 à laquelle on se propose d'appliquer *t1* ou *t2*, d'où l'ECI : $D(u_0) = Y_{t1}(u_0) - Y_{t2}(u_0)$ (ou une variante). On remarquera qu'aucune des deux réponses $Y_{t1}(u_0)$ et $Y_{t2}(u_0)$ n'est *contrefactuelle* tant que le traitement n'a pas été appliqué à u_0. $D(u_0)$ a la même distribution que

$D(u)$, c'est-à-dire une distribution normale avec moyenne δ (estimée avec précision) et variance $\sigma_D^2 = 2(1-\rho)\sigma^2$ (non estimable). Un principe qui paraît raisonnable et essentiel est que des modèles intrinsèquement *empiriquement indistinguables* devraient conduire à des *inférences indistinguables*. Mais ici on peut obtenir des inférences très différentes selon les valeurs que l'on suppose pour la corrélation : pour ρ=0 (indépendance de Y_{t1} et Y_{t2}), on a $\sigma_D^2=2\sigma^2$; pour ρ=1, on a $\sigma_D^2=0$; pour ρ=½, on a $\sigma_D^2=\sigma^2$; etc. Il en résulte une situation pour le moins embarrassante : comment choisir entre ces inférences ? En fait on peut seulement inférer l'inégalité $0\leq\sigma_D^2\leq2\sigma^2$. En général il ne sera pas possible d'obtenir d'inférence précise sur σ_D^2 , à moins de pouvoir montrer (ou de supposer) que σ^2 est nul, ou du moins est négligeable (ce qui impliquerait $\sigma_D^2\approx0$). Cela revient à supposer que les unités expérimentales sont non seulement indistinguables, mais *uniformes* : le résultat $Y_t(u)$ est *le même pour toutes les unités*.

La propriété σ^2=0 peut bien entendu être étudiée *empiriquement* ; elle peut être considérée comme une caractéristique distinctive d'au moins certains problèmes dans les « sciences dures ». Si elle est satisfaite, on dispose d'une mesure directe des ECI, puisqu'il est possible d'observer $Y_{t1}(u_0)$ et $Y_{t2}(u_0)$ simultanément, en utilisant des unités *différentes*. Mais cette propriété est complètement irréaliste dans les situations expérimentales en psychologie, et il faut donc pouvoir traiter le cas où les unités *ne sont pas uniformes*. Pour cela il est habituel dans les modèles contrefactuels d'introduire des contraintes supplémentaires sur le modèle métaphysique, et notamment l'hypothèse d'*additivité traitement-unité* (ou absence d'interaction). Celle-ci revient à supposer que l'ECI $D(u)$ est *identique pour toutes les unités* de la population et elle est équivalente à ρ=1. On obtient alors une inférence particulièrement simple : $D(u_0)=\delta$ (=ECM) et $\sigma_D^2=2(1-\rho)\sigma^2=0$; $D(u_0)$ est donc estimé (de façon précise) par la différence moyenne observée.

Mais il est impossible d'observer les deux composantes $Y_{t1}(u)$ et $Y_{t2}(u)$), et il n'y a donc *aucun moyen* de pouvoir tester l'additivité traitement-unité, c'est-à-dire de pouvoir démontrer *empiriquement* que l'ECI est le même pour tout u (une propriété en outre bien peu réaliste…) En conséquence aucune inférence sans ambiguïté sur l'effet causal individuel n'est possible sans suppositions non testables (même quand on a recueilli des données très nombreuses).

L'inférence sur les causes des effets

La situation est encore plus problématique dans le cas de l'inférence sur les « causes des effets ». Dans ce cas, on considère une unité supplémentaire u_0, qui présente un intérêt particulier et à laquelle le traitement *t1* (disons) *a déjà été appliqué* et l'observation $Y_{t1}(u_0)=y_{t1}$ a été faite. Pour aborder la question de savoir si, pour l'unité spécifique u_0, l'application de *t1* a « causé » la réponse observée, on ne peut faire autrement que de comparer d'une manière ou d'une autre la valeur observée y_{t1} avec la quantité *contrefactuelle* $Y_{t2}(u_0)$, qui aurait résulté de l'application de *t2* à u_0. Autrement dit, cela nécessite une inférence sur l'effet causal individuel : $D(u_0)=y_{t1}-Y_{t2}(u_0)$. Cependant, pour autant désirable que puise être une telle inférence elle n'est pas nécessairement possible.

Supposons qu'il n'y a aucune possibilité de mesurer une autre information pertinente sur aucune unité, en dehors de sa réponse au traitement. Sous le modèle métaphysique (contrefactuel), la distribution conditionnelle de

$D(u_0)=y_{t1}-Y_{t2}(u_0)$, étant donné la réponse observée $Y_{t1}(u_0) = y_{t1}$, est normale, avec moyenne $\lambda=y_{t1}-\mu_{t2}-\rho(y_{t1}-\mu_{t1})$ et variance $\zeta^2=(1-\rho^2)\sigma^2$. Comme nous l'avons déjà souligné, à partir de données très nombreuses, seuls μ_{t1}, μ_{t2} et σ^2 peuvent être estimés, mais la corrélation ρ ne peut pas être identifiée. Par suite, même dans ce cas il reste un arbitraire résiduel. On a : (1) pour ρ=0 (soit sous le modèle normal l'indépendance de Y_{t1} et Y_{t2}), $\lambda=y_{t1}-\mu_{t2}$ et $\zeta^2=\sigma^2$; (2) pour ρ=1 (soit l'additivité traitement-unité), $\lambda=\mu_{t1}-\mu_{t2}$ et ζ^2=0 ; (3) pour ρ=-1, $\lambda=2y_{t1}-\mu_{t1}-\mu_{t2}$ et ζ^2=0. Si nous supposons $\rho\geq 0$ nous pouvons seulement inférer les inégalités : $\mu_{t1}-\mu_{t2}\leq\lambda\leq y_{t1}-\mu_{t2}$ et $\zeta^2\leq\sigma^2$. C'est donc seulement quand y_{t1} est suffisamment proche de μ_{t1} que l'on pourra obtenir une conclusion sans ambiguïté sur la moyenne λ de $D(u_0)$, insensible à des suppositions non testables empiriquement sur la corrélation ρ ; et c'est seulement quand la variance σ^2 est suffisamment petite que l'on sera capable de dire quelque chose qui soit empiriquement fondé et sans ambiguïté sur la variance ζ^2 de $D(u_0)$. Si on prend ρ=1, ce qui est équivalent à l'additivité traitement-unité, alors on obtient une inférence apparemment déterministe : $D(u_0)=\mu_{t1}-\mu_{t2}$; mais cela est sans grande valeur réelle puisque les données ne peuvent donner aucune raison de choisir une valeur particulière de ρ plutôt qu'une autre. Notons que, si on suppose l'additivité traitement-unité et rien d'autre, alors l'inférence rétrospective sur $D(u_0)$ n'est pas affectée par l'information supplémentaire $Y_{t1}(u_0)= y_{t1}$ sur la nouvelle unité et est donc la même que dans le cas de l'inférence sur les effets des causes. On peut y voir la raison du fait que la distinction essentielle entre l'inférence sur les effets des causes et l'inférence sur les causes des effets n'est généralement pas faite.

Conclusion

S'il est possible de traiter l'inférence sur les effets des causes par une approche bayésienne décisionnelle (Dawid, 2000), Il y a une ambiguïté inhérente à l'inférence sur les causes des effets. Dawid en tire deux morales, toutes les deux fondées sur le principe que l'on devrait prendre soin de ne pas faire d'« inférences métaphysiques » sensibles à des hypothèses qui ne peuvent pas être testées empiriquement. La première morale est que l'inférence sur les effets causaux individuels devrait être soigneusement restreinte. La seconde morale est que si l'on ne peut pas obtenir une solution raisonnable au problème, alors c'est peut-être que le problème lui-même est mal posé. On peut ainsi arguer que l'introduction d'éléments contrefactuels est à la fois *non nécessaire* et *indésirable* ; il en résulte une critique très sévère à l'égard des développements récents de l'inférence statistique causale basés sur l'utilisation des *modèles graphiques structurels* (pour une synthèse de ces développements, *cf* Pearl, 2000).

Bibliographie

Dawid A. P. (2000). Causal inference without counterfactuals (with comments and rejoinder), *Journal of the American Statistical Association*, *95*, 407-448.

Lecoutre B. (2004). Expérimentation, inférence statistique et analyse causale, *Intellectica*, 38, 193-245.

Pearl J. (2000). *Causality: Models, Reasoning, and Inference*, New York: Cambridge University Press.