Intellectica, 2004/1, 38, pp. 193-245

Expérimentation, inférence statistique et analyse causale

Bruno Lecoutre[*1]

« Causal inference is, even more than other forms of inductive inference, only tentative. »

(Dawid, 2000)

Résumé : On se situe dans le cadre de l'analyse causale de données d'expériences « randomisées » (les traitements sont affectés à chaque unité expérimentale par tirage au sort). Les apports de quelques fondateurs de l'inférence statistique sont rapidement examinés. On considère ensuite les travaux récents, et notamment ceux sur les modèles graphiques structuraux de Pearl, qui visent à unifier sous une interprétation unique un certain nombre d'approches, incluant notamment les analyses contrefactuelles, les modèles graphiques, les modèles d'équations structurelles. La plupart de ces travaux reposent sur une approche contrefactuelle (invoquant des résultats potentiels : « si un autre traitement avait été affecté à l'unité expérimentale… » de l'inférence causale. Dans un article provocateur, Dawid (2000) soutient que cette approche est essentiellement métaphysique, et pleine de tentations de faire des inférences qui ne peuvent pas être justifiées sur la base de données empiriques. Concernant plus particulièrement les modèles graphiques structuraux, la critique de Dawid est que les « variables latentes » en jeu dans de tels modèles ne sont pas de véritables variables concomitantes (variables mesurables, qui peuvent être supposées non affectées par le traitement appliqué) et qu'il n'y a alors aucun moyen, même en principe, de vérifier les suppositions (« assomptions ») faites – qui affecteront néanmoins les inférences qui en découlent. Dawid qualifie en conséquence ces modèles de pseudo-déterministes et les considère comme non scientifiques. Les différents arguments et les solutions proposées sont examinés et discutés.

---

[*] ERIS, *Laboratoire de Mathématiques Raphaël Salem*, UMR 6085, C.N.R.S. et Université de Rouen, Mathématiques Site Colbert, 76821 Mont-Saint-Aignan Cedex
Courrier électronique: bruno.lecoutre@univ-rouen.fr
Internet : http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/

[1] Je remercie vivement Jacques Poitevineau pour ses commentaires et suggestions.

Mots-clés : Analyse de données expérimentales ; randomisation ; effets individuels ; inférence statistique ; inférence causale ; modèles graphiques structuraux ; approche contrefactuelle ; variables concomitantes ; méthodes bayésiennes.

Abstract: Experimentation, statistical inference and causal analysis. The causal analysis of « randomised » experimental data (treatments are randomly assigned to each experimental unit) is considered here. The contributions of some founders of statistical inference are briefly examined. Recent works, and especially Pearl's graphical structural models, are then considered. These models include counterfactual analyses, graphical models, structural equations models. Most of these models are based on a counterfactual approach (involving potential response : « if another treatment had been allocated to the experimental unit… ») to causal inference. In a provocative article, Dawid (2000) argues that this approach is essentially metaphysical, and full of temptations to make inferences that cannot be justified on the basis of empirical data. Regarding graphical structural models, Dawid's major criticism is that « latent variables » involved in such models are not genuine concomitant variables (measurable variables, that can be assumed unaffected by the treatment applied) and that there is no way, even in principle, of verifying the assumptions made – which will nevertheless affect the ensuing inferences. Dawid terms these models pseudodeterministic and regards them as unscientific. The arguments and solutions are reviewed and discussed.

Key Words: Experimental data analysis ; randomisation ; individual effects ; statistical inference ; causal inference ; graphical structural models ; counterfactuals ; concomitant variables; Bayesian methods.

INTRODUCTION

L'inférence statistique est un domaine hautement subtil et controversé. D'un point de vue formel on peut cependant en donner une définition axiomatique, reposant sur le concept de distribution de probabilité. Un exemple d'une telle définition est donnée dans le traité de statistique mathématique de Lehmann :

> *« The raw material of a statistical investigation is a set of observations; these are the values taken on by random variables X whose distribution P_θ is at least partly unknown. Of the parameter θ, which labels the distribution, it is assumed known only that it lies in a certain set Ω, the* parameter space. Statistical *inference is concerned with methods of using this observational material to obtain information concerning the distribution of X or the pa-*

> *rameter θ with which it is labelled. » (Lehmann, 1959, p. 1).*

Même si, pour la *mettre en pratique*, il faudra admettre l'usage de concepts hypothétiques (par exemple celui de population), on peut plus ou moins s'accorder sur une telle définition. Bien entendu, comme le souligne encore Lehmann, il faudra préciser les questions auxquelles on veut répondre :

> *« To arrive at a more precise formulation of the problem we shall consider the purpose of the inference. »*

mais là encore ces questions concernent des problèmes relativement bien cernés (estimation, test, décision, etc.) et on peut encore s'accorder sur leurs définitions (sinon sur leur intérêt dans un problème particulier).

Il est tentant pour les statisticiens de considérer que l'inférence sur les causes des phénomènes observés entre dans leur domaine. Mais alors il est sans doute illusoire de rechercher une définition formelle, qui pourrait satisfaire une majorité, de ce qu'est (ou devrait être) une *inférence statistique causale*, et il n'est pas du tout certain, sans même parler de la manière de les résoudre, que l'on puisse s'entendre sur les *questions* qui doivent être résolues. Face à de telles difficultés, il m'a paru essentiel de restreindre ici la problématique de l'analyse causale.

(1) Je limiterai d'abord le sujet au domaine de l'*expérimentation* contrôlée, et plus particulièrement au cadre des essais cliniques *randomisés* en médecine et pharmacologie.

(2) Je supposerai que l'on dispose de données suffisamment nombreuses permettant d'estimer les paramètres des modèles statistiques avec précision. Cela permettra de négliger l'incertitude purement statistique (liée aux échantillons de taille finie) – dont on verra qu'elle n'a qu'un rôle relativement secondaire dans les discussions théoriques – sans pour autant écarter les problèmes statistiques liés à la modélisation nécessaire pour une inférence causale. On évitera également ainsi les discussions sur le choix (sujet à controverse) d'une méthode d'inférence statistique particulière (tests de signification, inférence bayésienne…).

(3) Plutôt que de passer en revue tous les développements actuels sur l'analyse causale dans le domaine expérimental, j'accorderai une place prépondérante aux critiques récemment formulées par Dawid (2000) à l'encontre de la plupart de ces développements. Dawid argue que l'introduction dans ces développements d'entités « métaphysiques » (en particulier des éléments contrefactuels) intrinsèquement inobservables est à la fois *non nécessaire* et

indésirable. Il en résulte des critiques très sévères à l'égard des travaux de l'inférence statistique causale basés sur l'utilisation des *modèles graphiques structurels* (pour une synthèse de ces développements, *cf* Pearl, 2000). Même si ces critiques, volontairement provocatrices, et en un sens « destructrices », soulèvent sans doute plus de problèmes qu'elles n'en résolvent, elles n'en sont pas moins à mon avis un avertissement salutaire qui devrait permettre de reconsidérer un certain nombre de problèmes liés à l'usage même de la causalité dans les sciences expérimentales. Ce choix délibéré de privilégier ici les thèses de Dawid est en outre justifié par l'examen des commentaires des plus éminents spécialistes du domaine qui ont discuté son article. Ceux-ci ne trouvent en fait à lui opposer que des arguments de principe, s'en prenant le plus souvent à sa position philosophique, dont on verra qu'elle met clairement en avant une emphase *positiviste* sur les observables et sur la falsifiabilité des inférences. En particulier la discussion épistémologique apparaît tourner court. Ainsi Casella et Schwartz reprochent à Dawid sa référence explicite à Popper en arguant uniquement que celui-ci a maintenant entièrement perdu la faveur des philosophes (« Popper is out, counterfactuals are in », Dawid 2000, p. 426-427) ; mais on peut revendiquer fort justement avec Dawid que ses arguments soient jugés selon leurs propres mérites et non sur le fait qu'ils soient ou non à la mode.

Plan de l'exposé

Le plan de l'exposé sera le suivant. La section I permettra de préciser la problématique et d'introduire un certain nombre de concepts de base de l'expérimentation, à partir de l'exemple d'un essai clinique *randomisé*. Dans la section II j'envisagerai les approches les plus usuelles qui ont été utilisées par les statisticiens pour aborder les questions de causalité. Un rapide survol historique conduira à examiner brièvement les apports de quelques fondateurs de l'inférence statistique, et notamment ceux de Fisher, qui ont conduit à la caractérisation d'expérimentations rigoureuses, basées sur le principe de « randomisation » (affectation des traitements à chaque unité expérimentale par tirage au sort). En ce qui concerne les développements plus récents, j'introduirai essentiellement dans la section III les travaux de Pearl sur l'analyse causale qui, basés sur l'utilisation des *modèles graphiques structurels*, visent à unifier sous une interprétation unique un certain nombre d'approches précédentes, incluant notamment les *analyses contrefactuelles*, les *modèles graphiques* et les *modèles d'équations structurelles*. J'examinerai ensuite les critiques de Dawid à l'égard de l'approche contrefactuelle (invoquant des *résultats potentiels* hypothétiques) sur laquelle reposent précisément la plupart des développements sur l'inférence causale. Enfin, dans les trois dernières sections, plus techniques, après avoir introduit la formalisation statistique nécessaire à une discussion plus approfondie

(section IV), je développerai successivement les arguments de Dawid et en particulier la possibilité d'une analyse causale « sans contrefactuels » (section V), puis les solutions qu'il propose et les conclusions qu'il en tire (section VI). Un aspect essentiel de l'argumentation sera que, 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*.

Quelques remarques de terminologie

Il m'a paru difficile d'éviter les anglicismes suivants, qui se justifient toutefois par leurs origines. J'utiliserai ainsi les termes *randomisation* (randomization pour les américains) et *randomiser* dont l'usage est maintenant largement répandu en français, en me référant au fait que l'anglais *random* (« hasard ») a la même origine que l'ancien français *randon*, « mouvement impétueux », dont a été tiré *randonnée* (on a ainsi pu proposer randonisation et randoniser pour franciser ces termes). Dans le raisonnement statistique (comme d'ailleurs dans toute argumentation scientifique), il est essentiel de distinguer deux types d'hypothèses, celles qui sont des propositions que l'on met à l'épreuve et celles qui sont des propositions qu'il est nécessaire de postuler pour permettre l'inférence. Par exemple, dans le cas de la comparaison de deux moyennes par un test de signification (par exemple « le *t* de Student »), le premier type correspond à l'« hypothèse nulle » d'égalité des deux moyennes et le second type correspond aux suppositions qui doivent être tenues pour vraies pour assurer la validité de la procédure : indépendance des observations, normalité des distributions, égalité des variances, etc. Conformément à l'usage en anglais d'employer deux termes différents pour désigner ces deux types, j'utiliserai respectivement hypothèse et *assomption*. J'emploierai également *compliance*, qui dérive directement de l'ancien français *complir* (qui a donné accomplir). Enfin, pour alléger le texte, j'utiliserai « contrefactuel » comme un substantif, à l'instar de l'anglais.

I. L'EXPÉRIENCE DE BASE : L'ESSAI CLINIQUE RANDOMISÉ

Causalité et inférence statistique sont, au moins en principe, étroitement liées dans la méthodologie de l'expérimentation en médecine, comme le montre la caractérisation suivante d'un *essai clinique randomisé* dans l'ouvrage classique de Schwartz, Flamant et Lellouch, *L'essai thérapeutique chez l'homme* :

> *« L'essai vise à comparer plusieurs groupes, disons deux pour simplifier, ayant reçu des traitements A et B, de manière à savoir soit s'il existe entre eux une différence (dans un sens ou dans l'autre), soit si l'un d'eux, traite-*

> *ment nouveau, l'emporte sur l'autre, traitement classique […].*
>
> *La comparaison peut porter sur plusieurs critères, tenant compte du résultat sous différentes formes, et éventuellement des effets secondaires; pour simplifier, nous ne retiendrons ici qu'un seul critère, supposé qualitatif à deux classes : les malades sont 'guéris' ou non.*
>
> *Imaginons que les résultats de l'essai soient : 70 % de guéris pour le groupe A, contre 50 % pour le groupe B. Peut-on conclure à l'avantage du traitement A ?*
>
> *On rappelle que la réponse à cette question demande un jugement de signification et un jugement de causalité.*
>
> *a)* Le jugement de signification *permet de dire si la différence observée peut résulter des seules fluctuations d'échantillonnage ou si, au contraire, elle est réelle. Il est basé sur un* test statistique *(ici le test de* χ^2 *à un degré de liberté). La pratique de tels tests constitue l'essentiel des manuels de méthodologie statistique générale ; nous la supposerons connue dans ses grandes lignes.*
>
> *b)* Le jugement de causalité *permet, si la différence est significative, de l'imputer à la différence des deux traitements ; il n'en est bien ainsi que si les deux groupes considérés sont, à part l'administration du traitement, strictement comparables à tout point de vue. La constitution de deux groupes comparables est un problème également classique en méthodologie statistique : on sait que la solution correcte est le tirage au sort. » (Schwartz* et al.*, 1981, p. 11)*

Cette caractérisation de la problématique d'un essai clinique randomisé, faisant explicitement référence à un *jugement de causalité,* relève d'une conception traditionnelle qui remonte (au moins) aux travaux statistiques de Fisher et de Neyman sur les expériences agricoles. Elle comporte trois aspects essentiels : « la constitution de deux groupes comparables » ; « le tirage au sort » ; l'utilisation d'un « test statistique ».

I.1 La conception traditionnelle

I.1.1 La constitution de deux groupes comparables

Que signifie « strictement comparables à tout point de vue » ? Pour répondre à cette question, on invoque souvent une expérience *idéale* dans laquelle, en dehors du fait qu'ils reçoivent des traitements différents, les deux groupes seraient *semblables* dans tous les aspects. Idéalement, on voudrait comparer les résultats des patients obtenus pour le traitement A avec les résultats qu'on aurait obtenus pour les mêmes patients s'ils avaient reçu le traitement B, *toutes choses étant égales par ailleurs* (clause *ceteris paribus*). C'est ce qu'énoncent par exemple Angrist, Imbens & Rubin dans le cas particulier où le groupe B est un groupe contrôle (non traité) :

> *« The causal effect of a treatment on a single individual or unit of observation is the comparison (e.g., difference) between the value of the outcome* if the unit is treated *and the value of the outcome* if the unit is not treated. *» (Angrist* et al.*, 1996, p. 444 ; emphase ajoutée)*

Dans cette approche, la notion de base est celle de *résultat potentiel* : nous devons être capables d'imaginer les résultats que l'on aurait observés pour le groupe traité dans d'autres circonstances que celles auxquelles il a été réellement exposé. Dans la terminologie des philosophes, on invoque ainsi une interprétation *contrefactuelle*. Le terme *contrefactuel* renvoie à une notion subtile dont l'importance est particulièrement critique dans les discussions philosophiques de la causalité[2]. Un énoncé contrefactuel est une assertion de la forme « si *X* avait été, alors *Y* se serait produit » (si le nez de Cléopâtre avait été plus court…). Il est implicite dans une telle assertion que *X* n'a pas eu lieu et que la comparaison se fait entre un résultat réel et un résultat contrefactuel.

Pour rendre opérationnelle la comparaison des traitements, les statisticiens doivent préciser comment est mesuré *l'effet causal*. On sait que la plupart des analyses statistiques de données expérimentales se focalisent sur un effet « moyen », typiquement une différence de moyennes (ou de proportions). Ainsi les auteurs complètent leur définition de l'effet causal de la manière suivante :

> *« The target of estimation, the estimand, is typically the average causal effect, defined as the average difference between treated and untreated outcomes across all units*

[2] Voir la théorie contrefactuelle de Lewis (notamment Lewis, 1973, 1986), les controverses qu'elle a suscitées (par exemple, Menzies, 1989) et ses développements récents (Lewis, 2000).

> *in a population or in some subpopulation (e.g., males or females). » (Angrist* et al.*1996, p. 444)*

Je reviendrai ultérieurement sur cette mesure de l'effet causal, ainsi que sur le concept de *population* que l'on voit apparaître ici.

Si l'expérience idéale était réalisable, *tous* les facteurs causaux seraient considérés, et on constituerait des groupes semblables pour tous ces facteurs excepté un seul – le traitement – qui serait manipulé. La procédure d'inférence reposerait alors sur ce que les philosophes appellent l'*induction par élimination* et serait particulièrement directe : on pourrait inférer par un processus d'élimination que toute différence est *causée* par le traitement. C'est ce qu'énonce Urbach :

> *« If we also ignore the remote possibility that the causal mechanism is irreducibly indeterministic, such as might perhaps operate at the quantum level, then clearly the treatment must be the cause of the difference. This sort of inference, where every potential causal factor is laid out and all but one is excluded by the experimental information, is a form or what is traditionally called eliminative induction. » (Urbach, 1993, p. 1422)*

Mais, même si on affaiblit l'exigence de groupes *semblables* en *aussi semblables que possible*, l'expérience idéale semble inatteignable. Malgré toutes les précautions que l'on pourrait prendre, les groupes différeront toujours selon d'innombrables aspects ; c'est ce que soulignait Fisher :

> *« …it would be impossible to present an exhaustive list of such possible differences appropriate to any one kind of experiment, because the uncontrolled causes which may influence the results are always strictly innumerable. » (Fisher, 1990/1935, p. 18)*

Les conditions permettant une induction par élimination telle que décrite ci-dessus sont pour le moins difficiles à établir. Pourtant il s'agit là d'un processus de la démarche expérimentale qui, comme le souligne Evans paraît essentiel :

> *« Eliminative induction, though it is not usually referred to in those terms, is a vital component of peer review and discussion of all experimental findings. Critics will attempt to show that the author has failed to take account of some factor which gives a more convincing basis for the interpretation of the results. »*

et, indépendamment de toute considération philosophique, correspond à une pratique claire :

> *« Whether this is done following Popper or not, the practice is clear. » (Evans,* in *Urbach, 1993, p. 1433)*

On pourrait alors envisager une *approximation* de l'expérience idéale. Dans ce sens, une conception déjà plus réaliste pourrait être que les groupes devraient être semblables en ce qui concerne les facteurs que l'on appelle habituellement les *facteurs de pronostic* – c'est-à-dire des caractéristiques d'un patient qui lui sont propres et qui ont un rôle pronostic connu ou plausible, par exemple l'âge, le sexe, et évidemment l'état de la maladie avant le traitement. Comme le suggère Urbach (1993), il pourrait sembler en accord avec le bon sens de contrôler au moins les facteurs de pronostic connus, en appariant les groupes expérimentaux en ce qui les concerne. Cependant, un tel contrôle a encore de sérieuses limites, comme le soulignait Fisher :

> *« …whatever of care and experimental skill is expended in equalising the conditions, other than the one under test, which are liable to affect the result, this equalisation must always be to a greater or less extent incomplete, and in many important practical cases will certainly be grossly defective. » Fisher (1990/1935, p. 19)*

Dans un essai clinique les facteurs de pronostic sont effectivement généralement innombrables et la plupart ne sont même pas imaginables, alors que d'autres peuvent être considérés à tort comme étant non pertinents. Plus encore, un tel appariement des groupes expérimentaux est tout simplement *irréalisable* en pratique. C'est ce que Evans objecte à Urbach, en soulignant le fait que dans de nombreux essais on ne cherche même pas à identifier les facteurs de pronostic, lesquels ne sont pas en général sous le contrôle d'un expérimentateur :

> *« Patients do not arrive for treatment in any matched order, and one cannot delay their treatment until a patient with similar values for all prognostic factors turns up, or even allocate them provisionally to one group and allocate the next patient with those same factors to the other group. » (Evans,* in *Urbach, 1993, p. 1433)*

I.1.2 Le tirage au sort (randomisation)

Puisqu'une approximation de l'expérience idéale, qui consisterait à contrôler *au mieux* les facteurs de pronostic connus, est généralement infaisable en pratique, il fallait lui trouver un *substitut*. On uti-

lise pour cela un processus d'*affectation au hasard* – en anglais *randomisation* – des patients aux différents traitements. Ce processus vise à jouer pour les facteurs que l'on ne contrôle pas (ou que l'on ne peut pas contrôler) le même rôle que le processus d'appariement, comme l'énoncent Schwartz, Flamant et Lellouch :

> *« L'affectation aux malades du traitement A ou B, par tirage au sort, constitue deux groupes de malades aussi semblables que possible pour l'ensemble de tous les caractères,* connus ou inconnus, *des sujets. » (Schwartz* et al., *1981, p. 12)*

On peut alors étendre l'argument reposant sur le processus d'induction par élimination, en en gardant la simplicité essentielle, comme le décrit Cox (même s'il admet qu'en pratique il existe toutes sortes de complications) :

> *« Any difference between treatment groups either arises via the accidents of random assignment, and probability calculations of various kinds can be made about this, or is produced or explained or caused by the treatments: there are no other possible explanations. » (Cox,* in *Schaffner, 1993, p. 1495)*

Il en résulte que la randomisation est généralement considérée comme un apport essentiel, qui procure à l'expérimentateur un confort intellectuel indéniable. Ainsi Cox conclut :

> *« It seems to me that the notion of a carefully controlled randomized clinical trial may well be statistics' greatest contribution to the welfare of mankind and that suggestions that the underlying notions are faulty are quite mistaken. » (Cox,* in *Schaffner, 1993, p. 1495)*

I.1.3 L'utilisation d'un test statistique

Dans le contexte présent, un « test statistique » est utilisé pour rejeter l'hypothèse (« nulle ») que la différence observée est due « aux seules fluctuations d'échantillonnage » et par suite, si cette hypothèse peut être rejetée, admettre qu'il y a un effet de la manipulation expérimentale. Les tests statistiques, tels qu'ils sont habituellement pratiqués, sont sans nul doute la procédure d'inférence statistique la plus utilisée, mais ils sont aussi la procédure la plus controversée, tant sur le plan théorique que sur le plan méthodologique. Dans la mesure où l'on supposera ici que les données sont suffisamment nombreuses pour pouvoir ignorer les fluctuations d'échantillonnage et en particulier considérer que la

différence observée estime avec précision la différence « vraie » (dans la population), ces controverses pourront être éludées. En particulier la question (qui serait essentielle autrement) de savoir s'il convient d'utiliser d'autres procédures statistiques (notamment des procédures bayésiennes) sera ici évitée, sans perte de généralité (le lecteur intéressé pourra se référer à Lecoutre et Poitevineau, 2000, Rouanet *et al.*, 2000 et Lecoutre *et al.*, 2001).

I.2 Difficultés et complexifications

Bien entendu, même si elle paraît parfaitement bien définie, la situation de base précédente peut donner lieu à toutes sortes de difficultés en pratique, tant dans la réalisation de l'expérience que dans l'analyse des résultats (même si on exclut ici les problèmes liés au choix d'une procédure d'inférence statistique). Je me contenterai ici d'esquisser quelques-unes de ces sources de difficultés qui seront pertinentes pour la suite de l'exposé. Mais, même en restant dans le domaine de l'expérimentation contrôlée, il en existe bien entendu beaucoup d'autres, en particulier celles liées à l'utilisation de plans d'expériences plus complexes, à l'existence de facteurs confondus, à la considération de variables *censurées* (dans le cas où l'on mesure par exemple la durée de survie d'un patient après le traitement et où le décès ne s'est pas produit au terme de la période d'observation), etc.

I.2.1 Le non-respect de la randomisation

Pour que la randomisation puisse remplir le rôle qui lui est assigné, il faut bien entendu que l'essai clinique soit soigneusement contrôlé, que les procédures expérimentales soient suivies fidèlement, qu'une procédure « à l'aveugle » soit réellement mise en œuvre (à l'évidence le patient ne devrait pas savoir lequel des traitements il reçoit, surtout s'il s'agit d'un placebo), que les observations soient enregistrées « honnêtement », etc. Les conditions d'un respect *parfait* de la randomisation n'existent sans doute pas. Ne serait-ce que parce qu'en pratique on cherche souvent à constituer deux groupes de mêmes effectifs, on introduit une contrainte sur le tirage au sort ; ceci peut être négligeable en pratique mais d'autres sources de perturbation ne peuvent être ignorées et devraient être prises en compte explicitement dans la modélisation et l'analyse statistique. Je mentionnerai simplement ici le problème de la « compliance » au traitement, qui est la façon dont le patient respecte le traitement qui lui est affecté (il peut par exemple ne pas suivre le traitement, le suivre sous une forme différente, prendre des médicaments supplémentaires…), et sur lequel je reviendrai.

I.2.2 Notions d'unité expérimentale, de population

Une source de difficultés importante est que l'inférence statistique met en jeu un certain nombre de notions, pour lesquelles on

trouve différentes conceptions, mais qui en pratique peuvent difficilement être regardées autrement que comme des notions hypothétiques. C'est notamment le cas des notions de population et d'unité expérimentale. En pratique, et en dehors des situations de « tirage dans une urne », il paraît difficile de concevoir que le concept de population corresponde à une réalité objective ; c'est ce qu'énonce Fisher :

> *« Whereas, the only populations that can be referred to in a test of significance have no objective reality, being exclusively the product of the statistician's imagination through the hypotheses which he has decided to test, or usually indeed of some specific aspects of these hypotheses. » (Fisher, 1990/1956, p. 81)*

Pour rendre possible le raisonnement *inductif* de l'inférence statistique, on doit néanmoins supposer l'existence d'une population, c'est-à-dire un ensemble *U* d'« individus » *u*, que l'on appellera unités. L'expérience consiste à sélectionner deux sous-ensembles disjoints – ou *groupes* – d'unités U_{t1} et U_{t2} (dans le cas de l'expérience de base), auxquels on applique respectivement le traitement *t1* et le traitement *t2* et à observer les réponses qui en résultent. Il est important de prendre en compte le fait qu'une fois que le traitement a été appliqué à une unité *u*, l'autre traitement ne peut plus lui être appliqué, du moins *dans des conditions semblables*. Pour assurer cette propriété, il faut concevoir une définition appropriée de l'unité expérimentale (et par suite de la population), qui devra donc correspondre à *un certain état* du sujet (du patient) plutôt qu'au sujet lui-même. On voit donc bien ici que, même dans le cas où on s'intéresserait à un ensemble bien défini de sujets, la population qu'on devrait considérer, c'est-à-dire l'ensemble des états possibles des sujets, serait un ensemble infini et hypothétique. Cela implique encore que, sauf situation très particulière, la réplication des observations est *impossible*.

I.2.3 Unités expérimentales indistinguables, population homogène

Un concept essentiel dans l'expérimentation est celui d'unités *indistinguables*, ou du moins considérées comme telles par l'expérimentateur, avant les observations. En particulier il est essentiel que le traitement affecté à une unité particulière *u* ne dépende (délibérément ou non) *d'aucune information* qui pourrait permettre de la distinguer des autres unités. Ceci est en accord avec le fait que, dans la plupart des études expérimentales, les unités particulières utilisées dans l'expérience n'ont pas en elles-mêmes d'intérêt particulier, mais fournissent seulement une base pour l'inférence sur des propriétés génériques des unités sous l'effet des différents traite-

ments. Il s'agit là encore d'un concept intuitif, qui peut être formalisé de nombreuses façons. D'un point de vue fréquentiste, les individus sont généralement regardés comme *tirés au sort et indépendamment* de la population (même si en pratique cela n'est pas réalisable) ; d'un point de vue Bayésien, on peut les regarder comme *échangeables* (voir par exemple Press, 1989). On dit souvent dans le cas où les unités expérimentales sont indistinguables que la population est *homogène*, ce qu'il faut sans doute entendre comme « les individus ne peuvent être distingués sur la base des informations disponibles ».

Bien entendu, dans le cas où l'expérimentateur dispose *avant* l'affectation du traitement d'une information qui permet de distinguer les individus, il peut considérer une (ou plusieurs) sous-population d'unités dont les éléments sont indistinguables par cette information, ce qui le ramène à la situation précédente (on suppose bien entendu que l'information n'identifie pas l'unité de manière unique). Ceci correspond à une pratique courante dans les essais cliniques, où on impose souvent des *critères d'inclusion* des patients, qui conduisent par exemple à exclure *a priori* ceux d'entre eux qui auraient des valeurs extrêmes pour un certain critère. Mais un effet pervers d'une telle pratique est que les unités expérimentales peuvent être rendues *non représentatives* de la population à laquelle on cherche réellement à généraliser les résultats (à moins d'admettre que le traitement ne sera jamais appliqué aux patients ne satisfaisant pas les critères).

I.2.4 Prise en compte d'informations supplémentaires : covariables, variables concomitantes

Nous avons vu que des informations supplémentaires pouvaient être prises en compte dans la construction de l'expérience ; il est également possible (et souhaitable si cela est pertinent) de les considérer dans l'analyse des résultats, ce qui bien entendu compliquera la modélisation et les procédures. Il convient ici de préciser les définitions suivantes. On appellera *covariable* une variable mesurée *avant* l'application du traitement et *variable concomitante* une variable qui *n'est pas affectée* par le traitement appliqué. Par construction une *covariable* est un cas particulier de variable concomitante ; un exemple de variable concomitante qui n'est pas une covariable est, dans les expériences agricoles, le temps qu'il a fait entre le moment de la plantation (donc après application du traitement) et la récolte.

II. LA CAUSALITÉ ET LES STATISTICIENS

Comme le remarque Holland, la notion même de causalité paraît le plus souvent aux statisticiens très éloignée de leur problématique[3] :

[3] Sans parler des éminents statisticiens (notamment Pearson, 1911 et Jeffreys, 1998/1939) qui ont rejeté le principe de causalité (ou de déterminisme, ou d'uniformité de la nature),

> *« The reaction of many statisticians when confronted with the possibility that their profession might contribute to a discussion of causation is immediately to deny that there is any such possibility. "That correlation is not causation is perhaps the first thing that must be said" (Barnard, 1982, p. 387). » (Holland, 1986, p. 945)*

En dehors éventuellement de la mise en garde rituelle « l'association n'est pas la causalité », la plupart des manuels de statistique appliquée à l'usage des expérimentateurs, consacrés aux plans d'expériences et aux méthodes d'inférence statistique (l'analyse de variance notamment), accordent effectivement peu de place (sinon aucune) aux questions de causalité ; c'est ce que souligne Pearl :

> *« Statistical analysis which, traditionally, has excluded causal vocabulary both from its mathematical language and from its mainstream educational program. » (Pearl, 2001, p. 1)*

Mais on notera que ces manuels excluent plus généralement *tout aspect controversable*, évitant ainsi également dans la plupart des cas les controverses sur les tests de signification, qui à l'instar de la causalité font seulement l'objet d'avertissements rituels comme « la significativité statistique n'est pas la significativité clinique ».

On trouve pourtant dans le manuel du logiciel statistique (largement utilisé) SPSS (*Statistical Package for the Social Sciences*) une définition (au moins approximative) de la causalité :

> *« We propose the following 'operational' definition as an initial approximation to the idea of causality:* X_1 *is a cause of* X_0 *if and only if* X_0 *can be changed by* manipulating X_1 *and* X_1 alone. *We note first that the notion of causality implies prediction but prediction of a particular kind. It implies the notion of* possible *manipulation.*
>
> *The preceding definition of causality suggests both the criterion of causality and the means to measure causal effects. First to establish conclusively that* X_1 *is a cause of* X_0*, one must perform an "ideal" experiment in which all the other* relevant *variables are held constant while the causal variable is being manipulated. Second there should*

---

sous toute forme telle que « des antécédents exactement semblables conduisent à des conséquences exactement semblables ».

> *be some accompanying change in the dependent variable. We will use such validation as the ultimate criterion that X_1 is the cause of X_0. » (cité par Schaffner ; 1993, p. 1478)*

Cette caractérisation montre que l'approche de la causalité que les philosophes appellent l'approche de la « manipulation » (ou intervention) est largement répandue chez les statisticiens (au moins implicitement). La seconde partie de la définition fait intervenir explicitement la notion d'expérience idéale et donc sous-entend l'interprétation « contrefactuelle » de cette approche, dont l'objet est précisément de permettre de distinguer la causalité de la simple corrélation.

II.1 Les travaux fondateurs

C'est aux années 1920-1930 que l'on peut faire remonter la prise en compte des notions précédentes dans un raisonnement statistique *formalisé*. Le processus de randomisation est déjà sous-jacent au début des années vingt, notamment chez Student (1923) et chez Neyman (1923) qui introduit dans la modélisation statistique la notion de résultats potentiels et par suite des quantités contrefactuelles. Il a été explicitement développé comme une technique de planification des expériences par Fisher (Fisher, 1990/1935) et Yates (1935), à l'origine pour les expériences agricoles. Il permet d'assurer que les différences observées sont effectivement dues, avec une probabilité élevée, *à l'intervention*, comme l'énonce Barnard :

> *« Its crucial function is in fact, to assure, with high probability, that differences in the output variables associated with changes in the input variables really are due to these changes and not to other factors. » (Barnard, 1982, p. 688)*

Dès 1918, Fisher apparaît considérer la possibilité d'établir un lien de causalité comme un apport important de la statistique. En effet, son article paru cette année là, intitulé « The causes of human variability », débute ainsi :

> *« The great service which the modern development of statistics has rendered to eugenics is that it supplies a definite method of measuring and analysing variability. Only so can the* true causes *of variability be ascertained and the factors which are of no effect, the false claimants to importance in this regard, be excluded from consideration. » (Fisher, 1918, p. 213 ; emphase ajoutée)*

Et plus loin il explicite ce qu'il entend par cause :

> « *In the same way it should be clearly understood what we mean by a* cause of variability. *If we say, "This boy has grown tall because he has been well fed," we are not merely tracing out cause and effect in an individual instance ; we are suggesting that he might quite probably have been worse fed, and that in this case he would have been shorter. We are, in fact, suggesting that existing differences of nutrition can account for differences of stature comparable to the standard deviation of stature. Now this is just what is meant when we speak of nutrition as a cause of variability ; we thereby mean that in* a population absolutely uniform in regard to other causes, *such as breeding and exercise, existing differences of nutrition would produce a certain variability – in fact, that* a certain percentage *of the variance must be ascribed to nutrition* » *(Fisher, 1918, p. 214; deuxième emphase ajoutée)*

On voit bien apparaître ici le concept d'une population dont les éléments sont indistinguables (avant l'observation) et la clause *ceteris paribus* dans la formulation « *a population absolutely uniform in regard to other causes* ».

C'est en appliquant ces idées dans le cadre de l'expérimentation, que Fisher a permis de cristalliser, en quelque sorte, les notions sous-jacentes aux pratiques déjà en usage – et notamment le test de signification et la randomisation, dont on a vu le rôle primordial dans la méthodologie des essais cliniques (et bien entendu dans les autres disciplines expérimentales) – et à leur conférer un statut de méthode d'inférence incontournable. C'est en 1925 qu'est parue la première édition de son livre *Statistical Methods for Research Workers*, puis en 1935 celle de *The Design of Experiments* qui ont connu un succès considérable (14 éditions pour le premier ouvrage, 8 pour le second). Dans la conception de Fisher, la randomisation est la « base physique » *suffisante* qui permet de dépasser la corrélation pour obtenir – avec le test de signification – une interprétation causale (au sens expérimental) :

> « *…correlation is not causation. The fact is that if two factors, A and B, are associated – clearly, positively,* with statistical significance *as I say – it may be that A is an important cause of B, it may be that B is an important cause of A, it may be that something else, let us say X, is an important cause of both.* If, now, A the supposed cause has been randomized – *has been randomly assigned to the material from which the reaction is seen – then one may*

> *exclude at a blow the possibility that B causes of A, or that X causes A. We know perfectly well what causes A – the fall of the dice or the chances of the random sampling numbers, and nothing else. » (Fisher, 1959, p. 14 ; emphase ajoutée)*

et la validité du test de signification est elle-même assurée par le processus de *randomisation* :

> *« …by which the validity of the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated » (Fisher, 1990/1935, p. 19)*

II.2 Les approches récentes

Ce n'est que dans les années 1970 qu'est véritablement apparu un renouveau d'intérêt des statisticiens pour l'analyse causale, parallèlement au renouveau de l'inférence bayésienne[4]. On peut trouver son origine dans les travaux de Rubin (1974, 1978) qui s'inscrivent directement dans la ligne de Fisher, en ce qu'ils étendent la formulation des plans d'expériences aux études d'observations, sur la base des concepts de randomisation et de résultats potentiels. Des références qui se situent dans cette même ligne d'analyse contrefactuelle sont notamment Rosenbaum et Rubin, 1983 [ordre des auteurs inverse dans la biblio] ; Holland, 1986; Robins, 1986 ; Manski, 1995 ; Greenland *et al.*, 1999. D'autres travaux sont basés sur des modèles formels qui ont également fait l'objet de nombreux développements dans d'autres domaines, notamment les *modèles graphiques* (Pearl, 1988 ; Lauritzen, 1996 ; Cowell *et al.*, 1999) et les *modèles d'équations structurelles* (Angrist *et al.*, 1996 ; Heckman et Smith, 1998). Ces derniers ont eu un rôle prédominant pour les inférences sur les effets causaux dans les sciences économiques au cours des quarante dernières années ; ils reposent sur la spécification de systèmes d'équations avec des paramètres et des variables qui visent à prendre en compte les relations du comportement et spécifient les liens causaux entre les variables. Les inférences dans les modèles d'équations structurelles exploitent souvent la présence de *variables instrumentales,* qui sont des variables qui sont explicitement exclues de certaines équations et incluses dans d'autres, et qui sont par conséquent corrélées avec certaines réponses seulement à travers leur effet sur d'autres variables. Il n'est pas question de développer ici tous ces travaux, et je me contenterai d'introduire dans la section suivante une formulation plus récente proposée par Pearl (1995), qui

[4] S'il est indéniable et donne lieu de nos jours à un nombre important de travaux, cet intérêt des statisticiens reste toutefois essentiellement une affaire de spécialistes.

vise précisément à les unifier sous une interprétation unique, la formulation des *modèles graphiques structurels*. On trouvera dans Pearl (2000) une synthèse des travaux probabilistes et statistiques sur la causalité, et dans Pearl (2001) une présentation de sa formulation plus particulièrement adaptée aux sciences médicales et donc à notre propos. Pearl, au moins dans ses présentations récentes, adopte explicitement une approche contrefactuelle de l'inférence causale. Dans cette section, j'examinerai donc également les critiques de Dawid (2000), qui soutient que l'approche contrefactuelle est essentiellement *métaphysique*, et pleine de tentations de faire des inférences qui ne peuvent pas être justifiées sur la base de données empiriques et qu'il considère en conséquence comme étant *non scientifiques*.

III. LES MODÈLES GRAPHIQUES STRUCTURELS (PEARL, 2001) ET LES ANALYSES CONTREFACTUELLES

Pearl considère que les espérances de pouvoir réduire la causalité à la probabilité sont à la fois intenables et injustifiées, et que les philosophes qui ont fait de telles tentatives il y a une vingtaine d'années (notamment dans la ligne des travaux de Suppes, 1970) ont été forcés d'admettre des concepts extra-probabilistes (tels que « contrefactuels » ou « pertinence causale ») dans l'analyse causale. Il y voit la raison très simple que la théorie des probabilités traite des croyances sur un monde incertain, mais *statique*, tandis que la causalité traite des *changements* qui se produisent dans le monde lui-même. Selon lui, la théorie des probabilités, même si nous nous donnons une distribution conjointe complètement spécifiée sur toutes les variables de l'espace, ne peut nous dire comment cette fonction changerait sous des interventions externes. Pearl en tire deux conséquences.

(1) Il convient de distinguer les concepts statistiques et les concepts causaux, d'où les définitions suivantes qui reposent sur l'idée que les concepts causaux ne peuvent être définis à partir de la seule distribution (de probabilité) des variables observées :

> *« A statistical concept is any concept that can be defined in terms of a distribution (be it personal or frequency-based) of observed variables, and a causal concept is any concept concerning changes in variables that cannot be defined from the distribution alone. Examples of statistical concepts are: mean, variance, correlation, regression, dependence, conditional independence, association, likelihood, collapsibility, risk ratio, odd ratio, marginalization, conditionalization, "controlling for", and so on. Examples of causal concepts are: randomization, influence,*

> *effect, confounding, "holding constant", disturbance, spurious correlation, instrumental variables, intervention, explanation, attribution, and so on. » (Pearl, 2001, p. 3)*

(2) Le vocabulaire et la syntaxe du calcul des probabilités sont insuffisants pour exprimer l'information causale. Notamment ils ne permettent pas de distinguer la *dépendance statistique* – quantifiée par exemple par la probabilité conditionnelle Pr(maladie|symptômes) – de la *dépendance causale*, pour laquelle il n'existe pas d'expression dans le calcul des probabilités standard. Pour chercher à exprimer des relations causales, il faut donc *compléter* le langage des probabilités par un vocabulaire de la causalité (avec de nouvelles notations), dans lequel on aura des représentations symboliques distinctes pour « les symptômes causent la maladie » et pour « les symptômes sont associés avec la maladie ». Une telle distinction devrait permettre de qualifier la première assertion de « fausse » et la seconde de « vraie », afin d'incorporer de façon appropriée l'information causale dans le plan et l'interprétation des études statistiques.

On notera que les tentatives déjà faites dans les années 1980 pour enrichir le calcul des probabilités avec le vocabulaire causal, en utilisant des notations contrefactuelles (par exemple Robins, 1986, 1987 ; Greenland et Robins, 1986 ; Robins et Greenland, 1989), ont soulevé des difficultés et donné lieu à d'intenses controverses avec les défenseurs des notations statistiques conventionnelles. Mais pour Pearl ces difficultés étaient essentiellement dues à l'absence d'une conceptualisation appropriée ; il propose comme solution d'utiliser le « langage des graphes et des équations structurelles », qui selon lui offre à la fois une approche mathématique à l'analyse de l'effet causal et une base formelle pour l'analyse contrefactuelle.

III.1 Les modèles d'équations structurelles et leurs graphes associés

La figure 1 donne deux exemples de graphes associés à un modèle structurel simple.

III.1.1 Les graphes orientés

Dans le cadre des essais cliniques les variables observées peuvent représenter des symptômes, des traitements, etc. A titre d'illustration, supposons que T représente une variable « traitement », Y une variable « réponse », et K une certaine covariable qui agit sur la quantité de traitement reçue. Sur le graphe une flèche (ou arc orienté) – par exemple de T à Y – représente à la fois l'existence d'une influence causale (directe) de T sur Y et l'absence d'une influence causale de Y sur T. Les variables U, V et W représentent des facteurs, généralement non observés, qui exercent une influence sur les autres variables mais ne sont pas influencés par elles dans le modèle. On les

appelle variables « exogènes » par opposition aux variables « endogènes » *T*, *Y* et *K*. La variable *V* par exemple représente des facteurs qui influencent *T*, qui peuvent ou non être corrélés avec *U* (les facteurs qui influencent *Y*) et avec *W* (les facteurs qui influencent *K*). La possibilité d'une corrélation entre deux variables exogènes est figurée par une double flèche en pointillés ; c'est le cas pour *U* et *V* dans la figure 1b, alors que la figure 1a exclut cette possibilité. Il faut souligner que d'une certaine façon c'est l'absence d'une flèche et non sa présence qui exprime les assomptions causales dans le graphe. En effet une flèche indique seulement la possibilité d'un lien causal, dont l'intensité reste à déterminer à partir de données, alors que l'absence d'une flèche implique définitivement un lien d'intensité nulle.

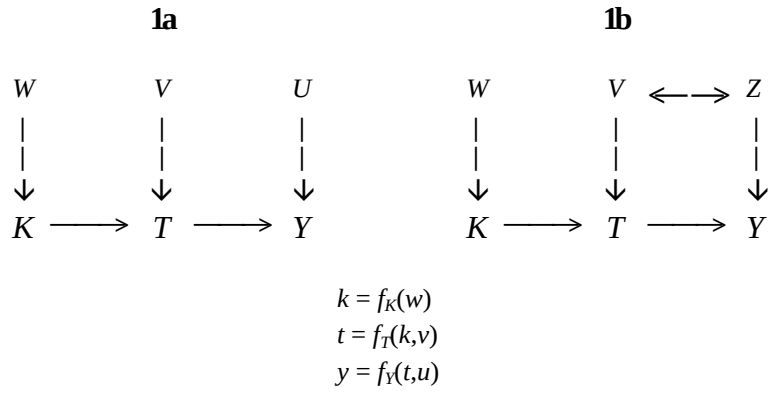

Figure 1 – Exemples de graphes associés à un modèle structurel simple

Bien que chaque assomption causale prise isolément ne puisse être testée, l'ensemble de toutes les assomptions causales dans un modèle a souvent des implications testables. Ainsi par exemple, dans la figure 1a, le modèle exprime sept assomptions causales, chacune correspondant pour un couple de variables soit à une seule flèche manquante soit à une double flèche manquante. L'ensemble de ces assomptions implique que *K* est non-associée avec *Y* pour chaque modalité de *T*. De telles implications testables peuvent être déduites à partir du graphe en utilisant un critère appelé *d-séparation* entre variables (Pearl, 1995).

III.1.2 Les modèles d'équations structurelles

Les modèles d'équations structurelles ont été à l'origine développés dans un cadre linéaire (remontant au moins aux travaux du généticien Wright, 1921, 1934). Alors que, comme on l'a vu, le

graphe n'exprime que l'existence possible ou au contraire la non-existence d'un lien causal, les équations structurelles expriment les relations quantitatives entre les variables considérées. Considérons par exemple, pour la figure 1a, l'équation

$$y = \beta t + u$$

Le paramètre β quantifie « l'effet causal » direct de T sur Y : étant donné la valeur numérique β, l'équation affirme qu'une augmentation de T d'une unité entraîne une augmentation de Y de β unités. Ceci suppose bien entendu que T et Y sont des variables numériques.

Pearl s'appuie sur une extension des modèles d'équations structurelles pour traiter les situations mettant en jeu des variables dichotomiques et des dépendances non linéaires. Dans cette généralisation, basée sur la simulation d'interventions hypothétiques dans le modèle, l'« effet » est défini, non plus simplement comme un coefficient dans une équation (β dans l'exemple précédent), mais comme une « capacité à transmettre les *changements* entre les variables ». L'objectif est d'ouvrir une nouvelle voie pour définir et estimer les effets causaux dans des modèles non linéaires et plus généralement dans des modèles dans lesquels la forme fonctionnelle des équations est inconnue (modèles généralement qualifiés de « non-paramétriques »). Ainsi par exemple on fera correspondre au graphe de la figure 1a un ensemble de trois fonctions, chacune correspondant à l'une des trois variables observées

$$k = f_K(w),\ t = f_T(k,v) \text{ et } y = f_Y(t,u)$$

où *W*, *V* et *U* sont supposées conjointement indépendantes, mais arbitrairement distribuées. Chacune de ces fonctions représente un processus (ou mécanisme) causal qui détermine la valeur de la variable de gauche (*sortie*) à partir de la valeur des variables de droite (*entrée*). L'absence d'une variable à la droite d'une équation exprime l'assomption qu'elle n'a pas d'effet direct sur la variable de gauche (ce qu'on peut « lire » sur le graphe). Un système de telles fonctions est dit *structurel* si chaque fonction est invariante sous des changements possibles de forme des autres fonctions.

III.2 La modélisation des effets causaux et des contrefactuels

III.2.1 La représentation des interventions

C'est cette propriété d'invariance qui permet de proposer une modélisation des effets causaux et des contrefactuels basée sur les équations structurelles. Pour cela, Pearl introduit un opérateur mathématique appelé « *do*( ) » qui simule des interventions physiques en supprimant certaines fonctions du modèle, les remplaçant par une constante, mais en gardant le reste du modèle inchangé. Par exemple, dans le modèle de la figure 1a, pour simuler une intervention

« $do(t_0)$ » qui fixe le traitement T à t_0, l'équation $t = f_T(k,v)$ est remplacée par « $t = t_0$ », ce qui conduit au nouveau modèle

$$k = f_K(w),\ t = t_0 \text{ et } y = f_Y(t,u)$$

dont la description graphique est montrée dans la figure 2.

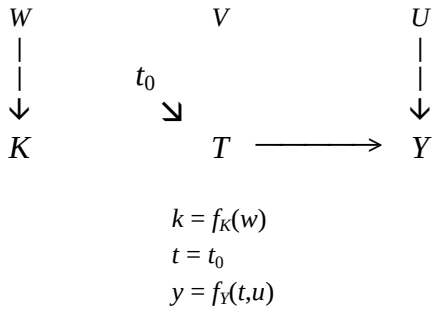

Figure 2 – Graphe associé au modèle modifié représentant l'intervention $do(t_0)$

La distribution de probabilité conjointe associée à ce modèle modifié, notée $\Pr(k,y \mid do(t_0))$ décrit la distribution des variables Y et K *après intervention*. Dans notre exemple $\Pr(k,y \mid do(t_0))$ donne la proportion d'individus qui auraient pour la réponse la modalité $Y = y$ et pour la covariable la modalité $K = k$ sous le traitement (hypothétique) $T = t_0$ administré uniformément à la population. On peut alors estimer l'efficacité du traitement, en considérant différentes modalités t. $\Pr(Y=y \mid do(t_0))$ décrit la distribution *marginale* de la réponse Y après intervention et est appelée « effet causal » par Pearl. On l'obtient comme

$$\Pr(Y=y \mid do(t_0)) = \Sigma_K \ \Pr(k,y \mid do(t_0))$$

Cette distribution fournit des mesures de l'efficacité du traitement. Un critère habituel pour la comparaison de deux modalités du traitement t_0 et t_1 est la différence (ou le rapport) des moyennes correspondantes, mais on peut également considérer d'autres caractéristiques des distributions (par exemple leurs variances).

Un problème essentiel est celui de *l'identification*, c'est-à-dire de la possibilité d'estimer la distribution après intervention à partir de données régies par la distribution avant intervention. Un théorème fondamental en analyse causale est qu'en général une telle identification pourra être faite chaque fois que le modèle est *markovien*, c'est-à-dire *acyclique* avec tous les termes d'erreurs conjointement indépendants. On peut également déterminer à partir de la structure du graphe les conditions qui permettent l'identification pour des

modèles non-markoviens, tels que ceux mettant en jeu des erreurs corrélées.

III.2.2 La dérivation des effets causaux

Considérons un exemple plus complexe de graphe avec cinq variables endogènes montré dans la figure 3.

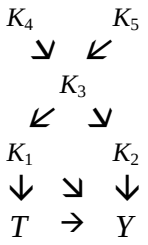

Figure 3 – Graphe associé à un modèle avec 5 variables endogènes

D'une manière générale, pour un tel graphe, la distribution conjointe peut être factorisée comme le produit

$$\Pr(v_1,v_2,\ldots,v_n) = \Pi_i \Pr(v_i|pa_i)$$

où les V_i (i=1,2…) sont les variables endogènes du modèle et pa_i représente les parents endogènes (s'il y en a) de V_i dans le graphe.

Par exemple, pour le graphe de la figure 3, on peut écrire directement la distribution conjointe, sans avoir besoin d'écrire les équations structurelles du modèle, comme

$$\Pr(t,k_1,k_2,k_3,k_4,k_5,y)$$
$$= \Pr(k_5)\Pr(k_4)\Pr(k_3|k_5,k_4)\Pr(k_2|k_3)\Pr(k_1|k_3)\Pr(t|k_1)\Pr(y|t,k_2,k_1)$$

(K_5 et K_4 n'ont pas de parent endogène, mais sont parents (endogènes) de K_3, etc.)

Supposons maintenant une intervention $do(t_0)$ qui fixe la variable T à t_0. Le modèle après intervention est également markovien et la distribution correspondante peut être factorisée comme le produit « tronqué »

$$\Pr(k_1,k_2,k_3,k_4,k_5,y|do(t_0))$$
$$= \Pr(k_5)\Pr(k_4)\Pr(k_3|k_5,k_4)\Pr(k_2|k_3)\Pr(k_1|k_3)\Pr(y|t_0,k_2,k_1)$$

soit d'une manière générale

$$\Pr(v_1,v_2,\ldots,v_n|do(t_0)) = \Pi_{i|V_i \notin T} \Pr(v_i|pa_i)\,|_{t=t_0}$$

où les $\Pr(v_i \mid pa_i)$ sont les probabilités conditionnelles avant intervention. L'effet causal de T sur Y, soit $\Pr(y \mid do(t_0))$ (suivant la définition de Pearl), est obtenu directement par marginalisation sur les varia-

bles K_j (j=1,2… 5), ce qui correspond à ce qui est appelé communément « l'ajustement pour les variables K_j ».

Notons que ces résultats sont encore applicables aux interventions, simultanées ou séquentielles, sur plusieurs variables.

III.2.3 L'analyse contrefactuelle dans les modèles structurels

L'opérateur *do*( ) ne permet pas d'exprimer toutes les questions de caractère causal. C'est notamment le cas pour celles qui concernent des relations contrefactuelles, telles que « *Y* aurait été *y* si *T* avait été *t* dans la situation $U = u$ » notée $Y_t(u) = y$. Néanmoins, dans le cadre des modèles d'équations structurelles, les contrefactuels peuvent être interprétés comme des solutions de systèmes d'équations modifiées, ce qui fournit le lien conceptuel et formel avec l'approche des *résultats potentiels* (en particulier, Robins, 1986 ; Holland, 1988 ; Angrist *et al.*, 1996). Pour cela Pearl propose d'interpréter la phrase « si *T* avait été *t* » comme une *instruction* pour modifier le modèle original et remplacer l'équation de *T* par une constante *t*, comme il a été fait plus haut pour représenter une intervention. Cette modification a pour but de permettre à la constante *t* de différer de la valeur réelle de *T* (dans le cas du modèle correspondant à la figure 1a, $f_T(k,v)$), sans créer de contradiction logique ; on peut par conséquent effectuer une inférence *en cascade* dans des modèles où l'antécédent d'un contrefactuel est une conséquence d'un autre.

A titre d'illustration, considérons à nouveau le modèle correspondant à la figure 2, obtenu en introduisant l'intervention $do(t_0)$. La solution de *Y* dans ce modèle, notée $Y_{t_0}(u,v,w)$, peut être appelée la *réponse potentielle* de *Y* à t_0. On peut donner à cette entité une interprétation contrefactuelle, car elle exprime la manière dont un individu présentant les caractéristiques (u,v,w) répondrait si le traitement avait été t_0, plutôt que le traitement $t = f_T(k,v)$ réellement reçu. Dans cet exemple, *Y* ne dépendant pas de *v* et *w*, on peut écrire

$$Y_{t_0}(u,v,w) = Y_{t_0}(u) = f_Y(t_0,u)$$

La distribution $\Pr(u,v,w)$ induit une probabilité bien définie sur l'événement contrefactuel $Y_{t_0} = y$, mais aussi sur des événements contrefactuels conjoints, tels que « $Y_{t_0} = y$ et $Y_{t_1} = y$' », qui sont, en principe, inobservables si $t_0 \neq t_1$.

III.2.4 La non-compliance dans un essai clinique randomisé

A titre d'illustration, considérons le problème de la non-compliance dans un essai clinique randomisé. Pearl envisage une approche similaire à celle utilisée pour les contrefactuels pour traiter ce problème. Pour estimer dans ce contexte l'effet causal du traitement *T* sur la réponse *Y*, il introduit la notion de « copie hypothétique » de la population. Considérons à nouveau l'exemple de la figure 1a, où on a, rappelons-le, trois variables observées *K*, *T*, *Y*, et le modèle

$k = f_K(w)$, $t = f_T(k,v)$ et $y = f_Y(t,u)$

Supposons maintenant que K représente une affectation randomisée des traitements, T est le traitement effectivement reçu et Y est la réponse observée. W représente le mécanisme aléatoire utilisé pour déterminer l'affectation ; V correspond à tous les facteurs (non observés) qui influencent la compliance du sujet, et U à tous ceux qui influencent la façon dont les sujets répondent aux traitements. On peut ajouter au modèle la possibilité d'une dépendance entre V et U pour permettre à certains facteurs (par exemple le statut socio-économique ou la prédisposition à la maladie) d'influencer à la fois la compliance et la réponse. La fonction f_T, qui dépend directement de k et v, représente le processus par lequel les sujets choisissent la modalité du traitement et f_Y, qui dépend directement de t et u, représente le processus qui détermine la réponse Y. Clairement, une compliance parfaite conduirait à poser $t = f_T(k,v) = k$ tandis que toute dépendance de T vis-à-vis de v représenterait une compliance imparfaite.

Le modèle graphique de la figure spécifie deux assomptions : (1) le traitement affecté K n'influence pas Y directement, mais à travers le traitement réel T ; ce type d'assomption est appelé « exclusion », car il *exclut* la variable K comme argument de la fonction f_Y ; (2) la variable K est indépendante de V et de U, ceci étant assuré par la randomisation de K qui élimine une cause commune pour K et V (de même que pour K et U). Dans cette situation Pearl définit l'effet causal comme la réponse de la population dans une expérience *hypothétique* où la modalité $T = t_0$ du traitement serait administrée uniformément à la population entière, en laissant t_0 prendre différentes valeurs dans des copies hypothétiques de la population. Une telle expérience hypothétique est régie par le modèle modifié exprimant l'intervention $do(t_0)$ considéré précédemment et la distribution correspondante $\Pr(y \mid do(t_0))$; mais cette distribution n'est plus identifiable. Notons cependant que dans ce cas il est possible (et c'est le mieux qu'on puisse faire) de dériver des bornes pour les quantités qui nous intéressent, c'est-à-dire un domaine de valeurs possibles qui représente notre ignorance sur le processus de génération des données et qui ne peut pas être amélioré en augmentant la taille de l'échantillon (voir Balke et Pearl, 1997 ; Chickering et Pearl, 1997).

III.2.5 En résumé

Les modèles graphiques structurels s'appuient sur des *relations fonctionnelles*, mettant en jeu les traitements et des variables *latentes* supplémentaires. Quand toutes les variables sont en principe observables, ils conduisent à la possibilité (au moins) d'inférences contrefactuelles bien définies. Ces modèles sont en apparence séduisants en semblant réduire les assomptions statistiques, grâce à l'utilisation de fonctions quelconques et non de distributions paramétriques d'une

forme particulière. Pour cette raison notamment ils peuvent apparaître bien adaptés aux données d'observation. Néanmoins dans ces modèles les inférences causales dépendent, en plus des propriétés purement liées aux distributions des variables manifestes, de la forme exacte des relations fonctionnelles (Balke, 1995 et Balke et Pearl, 1994, ont étudié cette dépendance). Ces relations fonctionnelles peuvent *en principe* être découvertes, mais il faut bien entendu qu'une telle structure déterministe puisse être prise au sérieux. C'est la condition pour que les modèles graphiques structurels puissent réellement prétendre fournir une solution de rechange à l'analyse statistique traditionnelle des données expérimentales.

III.3 Dawid (2000) et la critique des analyses contrefactuelles

III.3.1 Des moutons et des chèvres

Dawid (2000) argumente que tous les éléments d'une théorie qui n'ont pas de conséquences observables ou testables doivent être regardés comme métaphysiques. On ne devrait pas leur permettre d'avoir de conséquences sur l'inférence. Il invoque à ce propos la « loi de Jeffreys » que, dans un autre contexte (Dawid, 1984, sec. 5.2), il définit comme « l'exigence que des modèles mathématiquement distincts qui ne peuvent pas être distingués sur la base d'observations empiriques devraient conduire à des inférences indistinguables ». Cette propriété peut être démontrée mathématiquement dans les cas où ces inférences concernent des événements observables futurs. Dawid en vient ainsi à classer les analyses causales en « moutons » (celles obéissant à ce *dictum*) et en « chèvres » (les autres). Même s'il reconnaît que des modèles contrefactuels peuvent donner lieu à des utilisations particulières (acceptables) qui se trouvent être des moutons, il considère qu'il n'y a aucun avantage manifeste à de telles utilisations, devant le risque que ces modèles présentent d'engendrer des chèvres.

III.3.2 Le fatalisme

Pour Dawid, de nombreuses analyses contrefactuelles sont basées, explicitement ou implicitement, sur une attitude qui revient à considérer les différentes réponses potentielles $Y_t(u)$ (« si le traitement t avait été appliqué à l'unité u ») comme des attributs *prédéterminés* de l'unité u « qui n'attendent que d'être découverts » par une expérimentation appropriée ; il est implicite que l'unité u et ses propriétés et propensions existent indépendamment de tout traitement qui peut être appliqué, et ne sont pas altérées par ce traitement. Dawid qualifie cette attitude de *fataliste* et la regarde comme une assomption métaphysique parce que, chaque étiquette-unité u étant individuelle et non répétable, il n'y a jamais possibilité de tester empiriquement cette assomption. Il considère que la « vision fataliste » du monde s'oppose à la philosophie sous-jacente à la modélisation et l'inférence statistiques et nuit à leur mise en œuvre :

par exemple, elle exclut la possibilité d'introduire des effets stochastiques réalistes concernant les influences externes agissant entre le moment d'application du traitement et celui de la réponse. Dawid en conclut que cette approche devrait être traitée avec la plus grande défiance :

> *« Any account of causation that requires one to jettison all of the familiar statistical framework and machinery should be treated with the utmost suspicion, unless and until it has shown itself completely indispensable for its purpose. » (Dawid, 2000, p 413)*

Tout en reconnaissant que toutes les analyses contrefactuelles ne sont pas nécessairement fatalistes (il cite comme exemple d'exception Robins et Greenland, 1989), Dawid voit le fatalisme comme un compagnon très naturel de l'inférence contrefactuelle, au fonctionnement de laquelle il est souvent indispensable. A titre d'illustration, il considère qu'un usage fondamental du fatalisme est sous-jacent à certaines analyses contrefactuelles de la non-compliance au traitement, comme par exemple celles de Imbens et Rubin (1997). Ces analyses supposent que chaque patient peut être catégorisé comme « *compliant* » (prend le traitement s'il est prescrit, et ne le prend pas s'il n'est pas prescrit), « *défiant* » (ne le prend pas s'il est prescrit, le prend s'il n'est pas prescrit), « *prend toujours* » (qu'il soit prescrit ou non), ou « *ne prend jamais* » (qu'il soit prescrit ou non). La critique de Dawid est que des inférences causales basées sur la considération des réponses au traitement des différents « groupes » (les compliants, les défiants, etc.) ne peuvent avoir un contenu utile que si ces groupes ont une identité qui a un sens, et que cela n'est précisément réalisé que sous l'assomption irréaliste du fatalisme.

III.3.3 L'utilisation instrumentale de contrefactuels

Dawid voit cependant la possibilité d'accorder une place limitée pour un usage instrumental des contrefactuels dans le contexte de la *construction de modèles* causaux dans des problèmes complexes.

III.3.4 L'utilisation de contrefactuels pour l'inférence

Mais il reste persuadé de leur inutilité, même dans un rôle purement instrumental, pour l'inférence causale. Il affirme notamment que toutes les analyses contrefactuelles *acceptables* peuvent être facilement réinterprétées en termes non contrefactuels :

> *« However, I have not as yet encountered any use of counterfactual models for inference about the effects of causes that is not either (a) a goat, delivering misleading*

> *inferences of no empirical content, or (b) interpretable, or readily reinterpretable, in non-counterfactual terms. » (Dawid, 2000, p. 417)*

Dawid donne de cette possibilité de réinterprétation plusieurs exemples, dont je retiendrai les suivants. Robins et Wasserman 1997 ont reformulé en termes non contrefactuels des méthodes inférentielles causales que le premier auteur avait à l'origine développées sur la base d'un modèle contrefactuel. A l'inverse, Pearl avait précédemment introduit une sémantique pour les modèles graphiques des structures causales d'une manière qui évitait les contrefactuels ; Dawid ne voit aucun avantage évident à l'introduction ultérieure de contrefactuels, puisque les analyses spécifiques de Pearl (par exemple dans Pearl 1995, appendice) ne font aucun usage indispensable de cette structure supplémentaire.

L'approche contrefactuelle pourrait cependant (au moins) revendiquer un rôle heuristique ayant permis d'établir un grand nombre de résultats importants. Ainsi à Dawid qui note qu'il a été capable de dériver des résultats mathématiques identiques à ceux obtenus précédemment par Balke et Pearl (1994) sans recourir ni aux modèles fonctionnels ni aux contrefactuels, Wasserman (*in* Dawid, 2000, p. 443) objecte « peut-être, mais il connaissait déjà la réponse ! ». Cela sous-entend que « si Dawid n'avait pas connu la réponse il n'aurait pas été capable de la dériver » ; mais ce n'est à l'évidence qu'une assertion contrefactuelle invérifiable…

La conclusion de Dawid est que le résultat d'une analyse causale ne devrait mettre en jeu aucune assertion directe sur des contrefactuels. Je développerai plus loin ses arguments.

IV. FORMALISATION DE L'EXPÉRIENCE DE BASE

Une discussion plus approfondie nécessite une certaine formalisation statistique de la situation. Considérons toujours l'expérience de base randomisée dans laquelle on compare deux traitements possibles, que l'on notera *t1* (typiquement un nouveau traitement) et *t2* (typiquement un contrôle, un traitement de référence ou un placebo). Selon les notations introduites en 1.2 on considère deux sous-ensembles disjoints – ou *groupes* – d'unités U_{t1} et U_{t2}, auxquels on applique respectivement le traitement *t1* et le traitement *t2*, le choix du traitement appliqué s'effectuant par un tirage au sort.

IV.1 Les tableaux d'observation physique et métaphysique

Supposons qu'une expérience a été réalisée, suivant ce plan. Désignons les unités ayant reçu (effectivement) le traitement t ($t \in T = \{t1,t2\}$) par $u\langle t\rangle$ ($u = 1,2 \ldots n_t$) (l'étiquetage des unités est bien entendu arbitraire puisque les unités sont indistinguables). L'expérience a permis d'associer à chaque unité $u\langle t\rangle$ une *réponse*

$X_t(u)$. La collection de toutes les observations, notée $\boldsymbol{X}_{U<T>}$ constitue un tableau, qu'à la suite de Dawid (2000) on appellera le *tableau physique*. En effet, pour pouvoir rendre compte de l'approche contrefactuelle de l'analyse causale pour ce problème, on doit considérer pour *chaque* unité *u* (quel que soit le traitement effectivement reçu) les *deux* réponses – notées $Y_{t1}(u)$ et $Y_{t2}(u)$ – associées à chacun des deux traitements. La collection de tous les résultats potentiels, notée $\boldsymbol{Y}_{U\times T}$ constitue le *tableau métaphysique* (par opposition au *tableau physique*). Dans la terminologie des plans d'expérience, le tableau physique $\boldsymbol{X}_{U<T>}$ correspond à un plan en « deux groupes indépendants », alors que le tableau métaphysique $\boldsymbol{Y}_{U\times T}$ correspond à un plan en « deux groupes appariés ». Les notations expriment ces deux structures, où les unités sont respectivement *emboîtées* dans les deux traitements (plan $U<T>$) et *croisées* avec les traitements (plan $U\times T$).

On notera qu'un grand nombre de variables dans le tableau métaphysique $\boldsymbol{Y}_{U\times T}$ sont, en adoptant encore la terminologie de Dawid empruntée à la physique quantique, *complémentaires*, en ce qu'elles ne sont pas observables simultanément. En particulier, pour toute unité *u* de la population, une seule (s'il y en a une) des deux variables complémentaires, soit $Y_{t1}(u)$ soit $Y_{t2}(u)$ (déterminée par l'affectation des traitements aux unités) donne lieu à observation. Bien que la collection complète $\boldsymbol{Y}_{U\times T}$ soit intrinsèquement *inobservable*, l'analyse contrefactuelle repose sur la considération de tous les $Y_{t1}(u)$ et $Y_{t2}(u)$ *simultanément*.

IV.2 Les modèles statistiques physiques et métaphysiques

Pour simplifier autant que possible l'exposé, je me limiterai ici à considérer les modèles statistiques usuels familiers aux utilisateurs de l'analyse de variance.

IV.2.1 Le modèle physique

Dans le *modèle physique* associé au tableau des réponses observées $\boldsymbol{X}_{U<T>}$ (plan en groupes indépendants $U<T_2>$), les réponses $X_{u<t>}$ sont modélisées comme des variables aléatoires indépendantes (à la fois à l'intérieur de chaque traitement et entre les traitements) identiquement distribuées, chacune suivant la distribution normale, de moyenne μ_t et variance σ^2, soit

$$X_t(u)\,|\,\mu_t,\sigma^2 \sim \mathrm{N}(\mu_t,\sigma^2) \qquad [1]$$

Dans le langage usuel de l'analyse de variance, ce modèle met en jeu deux sources de variation, un *effet traitement* fixe et un *effet unité* aléatoire. L'analyse usuelle repose uniquement sur ce modèle physique, mais une analyse contrefactuelle ne peut se justifier que par un *modèle métaphysique* pour le tableau des réponses potentielles $\boldsymbol{Y}_{U\times T}$.

IV.2.2 Le modèle métaphysique

Le modèle métaphysique est le modèle normal bivarié usuel pour un plan $U \times T_2$: les couples de variables $[Y_{t1}(u), Y_{t2}(u)]$ sont indépendants et identiquement distribués, de distribution normale bivariée, avec moyenne $[\mu_{t1}, \mu_{t2}]$, variance σ^2 pour chaque variable et corrélation ρ (ou covariance $\rho\sigma^2$)

$$\begin{bmatrix} Y_{t1}(u) \\ Y_{t2}(u) \end{bmatrix} \Big| \ \mu_{t1}, \mu_{t2}, \sigma^2, \rho \ \sim \mathbf{N}_2 \left(\begin{bmatrix} \mu_{t1} \\ \mu_{t2} \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \right) \quad [2]$$

Dans le langage usuel de l'analyse de variance, ce modèle met en jeu trois sources de variation : un *effet traitement* fixe et deux effets aléatoires, un *effet unité* et un *effet d'interaction unité-traitement* (dont on verra le rôle essentiel dans les analyses habituelles). On remarquera que le modèle physique pour le tableau $\mathbf{X}_{U<T>}$ peut être dérivé du modèle métaphysique ; je reviendrai également sur ce point.

IV.3 Définitions de l'effet causal et conséquences sur l'inférence

IV.3.1 l'effet causal moyen et l'effet causal individuel

La plupart des analyses de variance rapportées dans les articles expérimentaux se limitent à des comparaisons de moyennes. Ainsi, dans le cas présent, ces analyses se limiteraient à une inférence sur la différence $\delta = \mu_{t1} - \mu_{t2}$. Dans le cadre des analyses causales, δ correspond à l'*effet causal moyen* (ECM). Même si beaucoup de ces analyses se focalisent sur l'ECM, l'objet fondamental de l'inférence causale, tel qu'il apparaît dans l'approche contrefactuelle, devrait être typiquement l'*effet causal individuel* (ECI) associé à chaque unité *u*, qui est une comparaison des deux réponses potentielles, le plus souvent défini comme leur différence

$$D(u) = Y_{t1}(u) - Y_{t2}(u)$$

Bien entendu, d'autres définitions peuvent être envisagées pour l'ECI (et par suite pour l'ECM), en particulier la différence des logarithmes $\log Y_{t1}(u) - \log Y_{t2}(u)$ ou le rapport $Y_{t1}(u)/Y_{t2}(u)$, sans qu'il y ait de raison théorique pour choisir une définition plutôt qu'une autre. Mais, quelle que soit sa définition, l'ECI est intrinsèquement *inobservable* puisque, dans la terminologie introduite précédemment, il met en jeu une comparaison de quantités complémentaires.

Du modèle métaphysique [2], on dérive le modèle marginal relatif à l'effet causal individuel. $D(u)$ a une distribution normale avec moyenne δ et variance $\sigma_D^2 = \sigma^2 + \sigma^2 - 2\rho\sigma^2 = 2(1-\rho)\sigma^2$

$$D(u) \mid \delta, \sigma^2, \rho \sim \mathrm{N}(\delta, \sigma_D^2) \quad [3]$$

IV.3.2 Conséquences sur l'inférence

Conformément à ce qui a été dit dans l'introduction, on supposera que les deux groupes sont d'effectifs suffisamment élevés pour que l'on puisse considérer qu'il n'y a pas d'incertitude statistique liée aux observations. Dans ce cas, à partir du tableau et du modèle physiques, on peut clairement estimer avec précision les moyennes μ_{t1}, μ_{t2} (et par suite leur différence δ) et la variance σ^2. Mais la corrélation ρ n'est pas identifiable[5].

Mais, même si une analyse contrefactuelle superficielle, limitée à l'inférence sur l'effet causal moyen δ peut apparaître fondamentalement valide (un « mouton » au sens de Dawid) et donc conduire à des conclusions univoquement acceptables, cela peut ne plus être le cas avec un changement, même mineur en apparence, de définition. Ainsi si l'on définit l'effet causal individuel comme le rapport $Y_{t1}(u)/Y_{t2}(u)$, sa moyenne n'est plus déterminée par les paramètres du modèle physique ; de manière générale elle n'est plus déterminée par les distributions marginales des variables $Y_{t1}(u)$ et $Y_{t2}(u)$.

IV.3.3 L'inférence prédictive

Dawid (2000) propose d'imaginer une *unité test* nouvelle u_o, issue de la même population, qui n'a pas encore donné lieu à observation, mais 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 le traitement *t1*, alors on obtiendra la réponse $Y_{t1}(u_0)$; si on choisit *t2*, on obtiendra $Y_{t2}(u_0)$. Le problème est alors de comparer d'une façon ou d'une autre ces deux quantités. Dawid suggère que l'approche contrefactuelle pourrait se focaliser sur l'effet causal individuel pour l'unité test u_0, soit sur la différence $D(u_o) = Y_{t1}(u_0) - Y_{t2}(u_0)$ (ou une variante). $D(u_o)$, étant indépendant en probabilité de l'ensemble des données sur les unités de l'expérience, a la même distribution que $D(u)$, c'est-à-dire une distribution normale avec moyenne δ (estimée avec précision) et variance σ_D^2 (non estimable)

$$D(u_0) \mid \delta,\sigma^2,\rho,[\text{données}] \sim D(u_0) \mid \delta,\sigma^2,\rho \sim \mathrm{N}(\delta,\sigma_D^2)$$
$$\text{avec } \sigma_D^2 = 2(1-\rho)\sigma^2$$

L'inférence sur $D(u_o)$ se déduit donc de celle sur le couple de paramètres (δ,σ_D^2). Pour des effectifs suffisamment élevés, δ est estimé avec précision, mais on peut seulement inférer que σ_D^2 est compris entre 0 (pour ρ=1) et $2\sigma^2$ (pour ρ=0). Même si la borne supérieure $2\sigma^2$ peut être estimée avec précision, on ne peut donc pas en général obtenir une inférence précise sur σ_D^2, à moins de pouvoir montrer (ou de supposer) que σ^2 est nul, ou du moins négligeable. J'y reviendrai.

[5] On a ce que McCullagh et Nelder (1989, sec. 3.5) appellent un « *intrinsic aliasing* » de l'effet unité et de l'interaction unité-traitement.

V. L'INFÉRENCE CAUSALE SANS CONTREFACTUELS (DAWID, 2000)

V.1 Deux problèmes : les effets des causes et les causes des effets

Pour Dawid il convient de distinguer l'inférence sur les *effets des causes*, qui consiste à comparer les conséquences attendues de différentes interventions possibles dans un système, et l'inférence sur les *causes des effets*, où l'on cherche à comprendre la relation causale entre un résultat déjà observé et une intervention antérieure. Il considère que ces deux types d'inférences, qui correspondent à deux types de questions différentes, sont tous deux valides et importants, mais qu'ils nécessitent des analyses *différentes*, bien que liées : par exemple des résultats d'enquêtes épidémiologiques, qui sont directement pertinents pour les questions relevant des effets des causes, sont souvent utilisés de façon inappropriée pour répondre à des questions sur les causes des effets, sans prêter suffisamment d'attention à la différence entre les deux types de questions.

L'inférence prédictive permet précisément une distinction claire entre ces deux types de questions et les façons de les traiter. Supposons qu'une expérience (ou un ensemble d'expériences) a été réalisée. L'inférence sur les effets des causes correspondra à une prédiction sur les deux réponses $Y_{t1}(u_0)$ et $Y_{t2}(u_0)$ attendues pour l'unité nouvelle u_0 (ou pour un ensemble d'unités), *avant toute intervention*. Il faut souligner le fait que, bien que $Y_{t1}(u_0)$ et $Y_{t2}(u_0)$ soient complémentaires, aucune des deux n'est contrefactuelle tant que le traitement n'a pas été appliqué à u_0. Pour l'inférence sur les causes des effets, on suppose qu'un traitement particulier (disons *t1*) a été choisi et *déjà appliqué* à u_0, et que la réponse $Y_{t1}(u_0) = y_{t1}$ a été observée. La question sur laquelle on se focalise est de savoir si, pour l'unité spécifique u_0, l'application de *t1* a « causé » la réponse observée.

Un aspect important de l'inférence sur les effets des causes est la prise de la décision sur le traitement à affecter à l'unité nouvelle u_0. Dawid va même jusqu'à considérer l'objectif de l'expérience comme étant précisément d'aider à cette prise de décision.

V.2 L'inférence sur les effets des causes

V.2.1 La violation de la loi de Jeffreys

On montre facilement que l'inférence sur l'effet causal individuel peut contredire la loi de Jeffreys qui, comme cela a été énoncé en III.3, requiert que des modèles qui sont intrinsèquement empiriquement indistinguables conduisent à des inférences indistinguables. On peut obtenir des inférences très différentes selon les valeurs que l'on suppose pour la corrélation ; on a par exemple

pour $\rho = 0$ (indépendance de Y_{t1} et Y_{t2}), $\sigma_D^2 = 2\sigma^2$
pour $\rho = 1$, $\sigma_D^2 = 0$
pour $\rho = 1/2$, $\sigma_D^2 = \sigma^2$

Il en résulte une situation pour le moins embarrassante : comment choisir entre ces inférences ? En fait on peut seulement inférer l'inégalité $0 \leq \sigma_D^2 \leq 2\sigma^2$. 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 $\sigma_D^2 \approx 0$). 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 $\sigma^2 = 0$ peut bien entendu être étudiée empiriquement, et pour Dawid elle pourrait être regardée comme une caractéristique qui serait distinctive d'au moins certains problèmes dans les sciences « dures ». En effet, quand elle est satisfaite, il est possible d'observer *à la fois* $Y_{t1}(u_0)$ et $Y_{t2}(u_0)$ simultanément, en utilisant des unités *différentes*, ce qui permet donc une mesure directe des effets causaux. Mais à l'évidence cette propriété est complètement irréaliste dans des situations expérimentales telles que les essais cliniques.

V.2.2 Contraintes supplémentaires

Il est habituel dans les études basées sur les modèles contrefactuels d'imposer des contraintes supplémentaires pour pouvoir traiter le cas où les unités ne sont pas uniformes. Rappelons que, sous le modèle [2], $D(u_0)$ a la distribution $N(\delta, \sigma_D^2)$, et c'est tout ce que l'on peut dire sur $D(u_0)$ à partir des données, à moins d'imposer des contraintes supplémentaires. Ainsi il est usuel de faire l'assomption que l'effet causal individuel $D(u)$ est le même pour toutes les unités de la population, ce qu'on appelle traditionnellement l'assomption d'*additivité* (ou *absence d'interaction*) *unité-traitement* (Kempthorne, 1952). Cette contrainte est équivalente à $\sigma_D^2 = 0$, soit encore $\rho = 1$ qui est la forme sous laquelle elle est déjà présente dans Neyman (1990/1923). L'inférence pour l'unité test u_0 est alors particulièrement simple puisque l'on en déduit que $D(u_0)$ est égal à l'effet causal moyen δ qui à partir de données nombreuses est estimé avec précision par la différence moyenne observée. Mais, puisqu'il n'est

jamais possible d'observer les deux composants de chaque couple $(Y_{t1}(u),Y_{t2}(u))$, il n'y a donc aucun moyen de pouvoir tester cette assomption, c'est-à-dire de pouvoir démontrer *empiriquement* que la différence inobservable $Y_{t1}(u) - Y_{t2}(u)$ est la même pour tout u. Pour Dawid, il ne fait même pas sens de parler d'une telle propriété sans adopter une attitude fataliste.

Une autre assomption qui a souvent été considérée comme essentielle à des inférences causales utiles est celle de *valeur stable unité-traitement* (Rubin, 1980, 1986). Elle apparaît si on considère le fait que la réponse de l'unité u pourrait en principe dépendre de l'ensemble ζ des affectations des traitements à toutes les unités expérimentales, et pas seulement du traitement spécifique t affecté à u (cette réponse devant donc être notée $Y_{\zeta}(u)$). On doit alors envisager un modèle métaphysique plus général, ce qui complique la situation ; mais on peut imposer à ce modèle la contrainte que la réponse de l'unité u ne dépend en fait que de t, ce qui ramène à la situation considérée précédemment avec $Y_{\zeta}(u) = Y_t(u)$ pour chaque u. Ici encore Dawid considère qu'il n'est pas possible de donner un sens à cette assomption de *valeur stable unité-traitement*, sans une attitude fataliste consistant à supposer l'existence de valeurs préexistantes des réponses $Y_{\zeta}(u)$ pour tout ensemble d'affectations possibles ζ (on peut toutefois en donner une réinterprétation non fataliste, comme on le verra plus loin).

Un dernier exemple d'assomption également non testable communément faite est la *monotonie* (*cf* Imbens et Angrist, 1994), qui dans le cas de réponses binaires requiert que $Pr(Y_{t2}=1,Y_{t1}=0) = 0$ (où la réponse 1 représente un succès du traitement et 0 un échec).

V.2.3 Deux morales

D'une manière générale, quel que soit le modèle supposé, le modèle physique permet seulement d'identifier les distributions marginales P_{t1} et P_{t2}, associées à chacun des traitements ; mais la distribution d'un effet causal individuel dépendra en outre de la structure de dépendance de la distribution conjointe P (pour laquelle, si les distributions marginales sont connues, on peut cependant déduire un certain nombre de propriétés et inégalités). En conséquence, même quand on a recueilli des données très nombreuses, on ne peut faire *aucune* inférence sans ambiguïté sur l'effet causal individuel sans faire d'assomptions non testables, telle que l'additivité *unité-traitement*. 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 assomptions 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, qu'il juge plus révolutionnaire, est que si l'on ne peut pas obtenir une solution

raisonnable au problème, alors c'est peut-être que le problème lui-même, en se focalisant sur l'inférence pour $D(u_0)$, est mal posé.

V.2.4 Utilisation d'information supplémentaire

On peut envisager d'utiliser l'information supplémentaire apportée sur les unités individuelles par une *covariable* (ou éventuellement plusieurs). Rappelons que les covariables sont des caractéristiques des unités qui doivent être observées *avant* l'expérimentation. Soit donc une covariable K, déterminée par un protocole de mesure qui, quand il est appliqué à l'unité u, conduit à la mesure $K(u)$. Dawid distingue trois cas, selon que cette covariable est observée sur les unités expérimentales et/ou sur les unités tests.

(1) K est mesurée pour toutes les unités expérimentales, et aussi pour une unité test u_0 *avant* que la décision sur le traitement soit à prendre. Si K prend ses valeurs dans un ensemble fini, et si on se limite au sous-ensemble (supposé vaste) des unités expérimentales pour lesquelles $K(u) = K(u_0)$, alors on retrouve essentiellement le problème déjà analysé. Autrement, ou si le sous-ensemble considéré ci-dessus n'est pas suffisamment vaste, on peut recourir à une modélisation statistique appropriée. Mais une analyse contrefactuelle nécessiterait de modéliser la distribution conjointe de (Y_{t1},Y_{t2}) conditionnellement à K, et serait donc sensible à des assomptions supplémentaires concernant la distribution conjointe des réponses potentielles.

(2) Si la covariable est mesurée seulement pour les unités expérimentales, Dawid considère qu'il peut être approprié d'ignorer entièrement l'information qu'elle apporte, si on excepte le fait que, quand les effectifs ne sont pas élevés, modéliser cette information pourrait rendre plus précise l'estimation des distributions prédictives marginales pour chaque traitement.

(3) Le cas où la covariable est mesurée sur l'unité test seulement est plus problématique, parce que l'expérience ne donne pas d'information directe sur les distributions prédictives requises de la réponse étant donné la covariable et le traitement. Quelle que soit l'approche adoptée (contrefactuelle ou non), on ne peut échapper au fait que la solution sera fortement dépendante d'assomptions non testées (bien qu'en principe testables) sur ces distributions. Dans ce cas, il n'y a cependant rien à gagner à l'introduction de contrefactuels.

V.2.5 Une solution alternative à l'additivité unité-traitement

Un argument qui peut être donné pour la nécessité d'une assomption « métaphysique » telle que l'additivité *unité-traitement* est le suivant. Comme on l'a vu, un essai clinique a souvent des *critères* d'inclusion très spécifiques qui rendent les unités expérimentales non représentatives de la population à laquelle on cherche en fait à géné-

raliser les résultats. Même dans le cas où les unités à l'intérieur de l'expérience peuvent encore être considérées indistinguables, il risque alors d'être déraisonnable de considérer que l'unité test u_0 est échangeable avec les unités expérimentales. Toutefois, si on suppose l'additivité *unité-traitement* de sorte que $Y_{t1}(u) - Y_{t2}(u) \equiv \delta$ pour toutes les unités – expérimentales et tests – alors un estimateur de l'effet traitement δ obtenu à partir de l'expérience sera encore applicable à u_0 ; dans ce cas l'analyse contrefactuelle apparaît non affectée par cette modification du cadre expérimental due aux critères d'inclusion. Au contraire, dans l'approche décisionnelle, les inférences prédictives séparées requises sur la réponse $Y_t(u_0)$, étant donné chaque traitement, pour une unité test u_0 sont à la fois plus compliquées et moins fiables quand les unités expérimentales ne peuvent pas être regardées comme représentatives des unités tests.

Dawid propose une autre façon de procéder, qui évite l'assomption métaphysique d'additivité *unité-traitement*. Il introduit pour cela une variable Q qui rend compte du processus de planification de l'expérience. Pour chaque unité u, $Q(u)$ peut prendre l'une des trois valeurs $t1$, $t2$ et 0 : si $Q(u) = t1$ ou $Q(u) = t2$ on inclut l'unité u dans l'expérience et on lui applique le traitement correspondant ; si $Q(u) = 0$ on exclut l'unité u. Supposons que, pour une certaine covariable K, la distribution de $Q(u)$ étant donné $K(u)$ est la même pour toutes les unités u. Alors K est l'information que l'expérimentateur prend en compte en générant Q ; elle incorpore les critères d'inclusion et de traitement. La distribution de Q étant donné K est supposée non affectée par un conditionnement supplémentaire sur le traitement appliqué t et la réponse éventuelle Y. Utilisant la notation et les propriétés de l'indépendance conditionnelle (Dawid 1979)

$$Q \perp (t,Y) \mid K \text{ d'où } Y \perp Q \mid K,t \qquad [5]$$

où $\perp$ se lit « est indépendante de »

Considérons maintenant un modèle faisant l'assomption suivante sur l'espérance (la moyenne) de la réponse Y étant donné le traitement t et la covariable K

$$E(Y \mid K,t) = \mu_t + \gamma(K) \quad (t = t1, t2) \qquad [6]$$

pour des paramètres (inconnus) μ_{t1} et μ_{t2} et une fonction paramétrique $\gamma(\cdot)$. Si cela est réalisé, définissons $\delta = \mu_{t1} - \mu_{t2}$. Notons qu'en raison de l'indépendance [5], l'espérance [6] n'est pas modifiée par la donnée supplémentaire de Q, soit

$$E(Y \mid K,t,Q=t) = \mu_t + \gamma(K) \quad (t = t1, t2)$$

de sorte que pour tout k

$$E(Y|K=k,t1,Q=t1) - E(Y|K=k,t2,Q=t2) = \mu_{t1}-\mu_{t2} = \delta \qquad [7]$$

Inversement [7] avec l'indépendance [5] implique [6]. La quantité $E(Y|K=k,t,Q=t)$ peut être estimée directement à partir des mesures de la covariable K et des résultats Y sur l'ensemble des unités expérimentales à laquelle le traitement t a été appliqué. En conséquence la propriété [6] peut être testée à partir des données expérimentales ; si elle peut être tenue pour vraie le paramètre δ est estimable, un estimateur simple non biaisé de δ étant la différence des moyennes pour les deux groupes traités.

Pour une unité test nouvelle u_0, avec une valeur observée de la covariable $K(u_0) = k$ (et par construction $Q(u_0) = 0$), on peut comparer les réponses à des applications hypothétiques des traitements ; en utilisant les résultats précédents, on obtient encore (si $K(u_0)$ n'était pas observée, on considérerait une espérance supplémentaire sur K, ce qui ne changerait pas le résultat)

$$\begin{aligned} &E(Y(u_0)\,|K(u_0)=k,t1) - E(Y(u_0)\,|K(u_0)=k,t2) \\ &\quad = E(Y|K=k,t1,Q=0) - E(Y|K=k,t2,Q=0) \\ &\quad = E(Y|K=k,t1) - E(Y|K=k,t2) = \delta \end{aligned}$$

Dawid en conclut que cette approche, basée sur l'assomption testable [6] plutôt que sur l'assomption métaphysique de l'additivité *unité-traitement*, permet de généraliser aisément de l'expérience à la population cible, même en présence de critères de sélection qui rendent les unités expérimentales non représentatives.

V.2.6 Utilisation de contrefactuels pour la modélisation

Partant du modèle métaphysique [2], on peut dériver par marginalisation le modèle [1] pour le tableau physique. Cela peut être regardé comme un usage *instrumental* de contrefactuels dans un but de modélisation. Cependant, dans cet exemple simple on ne voit vraiment pas la nécessité d'un tel usage, le modèle [1] s'imposant directement et très naturellement pour le tableau physique. Ce n'est que dans des problèmes plus complexes, que Dawid voit un possible avantage véritable – une utilité ou au moins une commodité – à une modélisation au niveau métaphysique et en fournit un exemple pour le cas de plans en carré latin. Dans ce cas, modéliser le tableau métaphysique, dans le but purement instrumental de dériver un modèle approprié pour le tableau physique, apparaît être l'approche la plus fructueuse.

V.2.7 Compatibilité

L'approche consistant à modéliser chaque tableau physique possible par marginalisation à partir d'un *unique* modèle conjoint pour le tableau métaphysique confère aux modèles physiques une propriété – que Dawid appelle *compatibilité* – qui peut s'énoncer ainsi : pour deux dispositifs expérimentaux différents qui tous deux

conduisent à ce que l'unité *u* (et plus généralement une collection d'unités) reçoive le traitement *t*, les modèles marginaux pour les réponses associées à cette unité (à cette collection) sont identiques. Cette propriété peut être regardée comme une contrepartie *non contrefactuelle* de l'assomption contrefactuelle de valeur stable unité-traitement définie précédemment. Dawid distingue encore deux formes, *forte* et *faible*, de compatibilité pour une collection de modèles physiques. La compatibilité faible, qui semble la plus naturelle requiert simplement la propriété d'identité des modèles marginaux communs. Elle ne fait en aucune façon référence à des contrefactuels, contrairement à la compatibilité forte qui requiert également l'existence d'un modèle conjoint unique pour le tableau métaphysique permettant de générer les différents modèles physiques par une marginalisation appropriée.

V.3 L'inférence sur les causes des effets

Je n'ai considéré jusqu'à présent que l'inférence sur les « effets des causes ». La situation est encore plus problématique dans le cas de l'inférence sur les « causes des effets », pour laquelle il peut être impossible d'éviter un certain degré d'ambiguïté. Le nouvel élément important est que, en plus des données expérimentales, on a maintenant une unité supplémentaire u_0, présentant un intérêt particulier, à laquelle le traitement *t1* (disons) *a déjà été appliqué* et la réponse $Y_{t1}(u_0) = y_{t1}$ a été observée (on peut éventuellement avoir aussi une information supplémentaire pertinente sur u_0 ou sur son environnement, obtenue entre l'application du traitement et l'observation de la réponse ; je considérerai cette possibilité ultérieurement). L'inférence sur les causes des effets amène la question de savoir si, pour l'unité spécifique u_0, l'application de *t1* a « causé » la réponse observée. Pour aborder cette question, on ne peut faire autrement que de comparer d'une manière ou d'une autre la valeur observée y_{t1} avec la quantité *contrefactuelle* $Y_{t2}(u_0)$, la réponse 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 désirable que puisse être une telle inférence, elle n'est pas nécessairement possible. Dawid conclut même, à partir de l'exemple suivant, que quelqu'un d'entièrement sceptique pourrait soutenir que l'inférence sur les causes des effets, sur la base de l'évidence empirique, est impossible. Considérons encore le modèle métaphysique (contrefactuel) normal bivarié [2], et supposons qu'il n'y a aucune possibilité de mesurer une autre information pertinente sur aucune unité, en dehors de sa réponse au traitement. 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, de moyenne λ et de variance ζ_D^2

$$D(u_0) \mid \delta, \sigma^2, \rho, Y_{t1}(u_0) = y_{t1} \sim \mathrm{N}(\lambda, \zeta_D^2)$$ avec $\lambda = y_{t1} - \mu_{t2} - \rho(y_{t1} - \mu_{t1})$ et $\zeta_D^2 = (1 - \rho^2)\sigma^2$

au lieu de la moyenne $\delta = \mu_{t1}-\mu_{t2}$ et de la variance $\sigma_D^2 = 2(1 - \rho)\sigma^2$ pour la distribution non conditionnelle. Mais, même dans le cas de données très nombreuses, il reste un arbitraire résiduel, puisque, comme on l'a vu, la corrélation ρ ne peut pas être identifiée.

On a pour $\rho = 0$ (soit l'indépendance de Y_{t1} et Y_{t2})

$\lambda = y_{t1} - \mu_{t2}$ et $\zeta_D^2 = \sigma^2$

et, pour les valeurs extrêmes

$\lambda = \delta$et $\zeta_D^2 = 0$, pour $\rho = 1$ (soit l'additivité *unité-traitement*)

$\lambda = 2\, y_{t1} - \mu_{t1} - \mu_{t2}$ et $\zeta_D^2 = 0$, pour $\rho = -1$

Si on suppose $\rho \geq 0$ on peut seulement inférer les inégalités

$$\mu_{t1} - \mu_{t2} \leq \lambda \leq y_{t1} - \mu_{t2} \text{ et } \zeta_D^2 \leq \sigma^2 \qquad [8]$$

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 assomptions non testables empiriquement sur la corrélation ρ ; et c'est seulement quand 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_D^2 de $D(u_0)$.

Si on prend $\rho = 1$ (additivité *unité-traitement*), 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. Les inégalités [8] sont liées à l'assomption, elle-même non testable, de normalité conjointe : même si les données pouvaient corroborer la normalité marginale pour Y_{t1} et pour Y_{t2}, les autres caractéristiques de la distribution conjointe resteraient inconnues, et en principe la distribution de Y_{t2}, étant donné la valeur observée $Y_{t1} = y$, pourrait être quelconque pourvu que $\sigma^2 \neq 0$.

Notons que, si on suppose l'additivité *unité-traitement* 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 la discussion sur les effets des causes. Pour Dawid, c'est précisément parce que l'additivité *unité-traitement* est si dominante dans la littérature que la distinction essentielle entre l'inférence sur les effets des causes et l'inférence sur les causes des effets n'a généralement pas été relevée. Il en résulte qu'il y a une ambiguïté inhérente à l'inférence sur les causes des effets, ce qu'il résume ainsi :

> *« No amount of wishful thinking, clever analysis, or arbitrary untestable assumptions can license unambiguous inference about causes of effects, even when the model is simple and the data are extensive (unless one is lucky*

enough to discover uniformity among units). » (Dawid, 2000, p. 418)

VI. LES SOLUTIONS PROPOSÉES PAR DAWID (2000)

VI.1 L'inférence sur les effets des causes : Une approche décisionnelle

Rappelons qu'il s'agit de prendre une décision sur le traitement à affecter à l'unité nouvelle u_0. Cette décision fait intervenir les distributions *prédictives* marginales (relatives à cette unité) P_{t1} et P_{t2} associées à chacun des deux traitements, qui expriment l'incertitude sur la réponse associée à u_0 conditionnellement au fait qu'on lui affecte respectivement le traitement *t1* ou le traitement *t2*. Dans l'exemple précédent du modèle normal avec des données très nombreuses, la distribution P_t est simplement la distribution $N(\mu_t, \sigma^2)$. Avec des effectifs limités, il faudrait prendre en compte l'incertitude relative à μ_t et σ^2, ce qui nécessite une approche *bayésienne* permettant de dériver la distribution *a posteriori* pour ces paramètres à partir de l'expérience réalisée. Dans tous les cas les distributions P_{t1} et P_{t2} se déduisent du modèle physique, ce qui *évite complètement les contrefactuels*.

Une approche formelle du problème de décision utilise une fonction de coût $L(.)$, qui mesure la conséquence (la « pénalité ») de la décision d'affecter le traitement *t*. Si le traitement *t* est choisi, le coût moyen est l'espérance $E_{P_t}(L(Y))$. Les principes de la décision bayésienne permettent de choisir le traitement qui conduit au coût moyen le plus faible. Cette analyse s'étend aisément au cas où la décision porte sur le traitement à affecter à un ensemble d'unités nouvelles. Quelle que soit la fonction de coût utilisée la solution ne met en jeu que les distributions marginales identifiables P_{t1} et P_{t2} ; donc quelles que soient les suppositions que l'on peut faire sur la corrélation ρ, on parvient à la même décision. Dawid voit les avantages suivants à cette approche décisionnelle.

VI.1.1 L'approche décisionnelle et le fatalisme

L'approche décisionnelle ne nécessite pas de recours au fatalisme. Il n'y a aucune difficulté conceptuelle ou mathématique à considérer que les distributions marginales de probabilité des réponses P_{t1} et P_{t2} incorporent des influences supplémentaires incontrôlables en plus des effets directement attribuables au traitement. Une assomption telle que celle de *valeur stable unité-traitement* considérée dans la Section V.2 est donc inutile. Elle peut être remplacée par l'assomption beaucoup plus faible que l'application des traitements ne détruit pas le caractère indistinguable des unités (expérimentales ou futures), en dehors bien entendu du fait que certaines ont reçu l'un des traitements et certaines l'autre. On pourra par conséquent utiliser les données de l'expérience pour identifier la distribution P_t de la

réponse dans le groupe traité par *t*. Cette distribution exprime aussi l'incertitude sur la réponse $Y_t(u_0)$ d'une nouvelle unité u_0, si on lui affectait le traitement *t*. On est donc en mesure de poser – et de résoudre – le problème de décision pour u_0.

VI.1.2 L'approche décisionnelle et la prise en compte de covariables

Dans le cas où l'on prend en compte une covariable *K* mesurée pour toutes les unités expérimentales, ainsi que pour l'unité test u_0, l'approche décisionnelle nécessite seulement d'utiliser les données pour évaluer et comparer les distributions prédictives de $Y_t(u_0)$ étant donné $K(u_0)$, pour chaque traitement *t*. Ici encore cette approche, au contraire de l'approche contrefactuelle, est fondamentalement insensible à toute assomption supplémentaire concernant la distribution conjointe des réponses potentielles.

VI.1.3 L'approche décisionnelle et la compatibilité

Dans l'approche décisionnelle, la propriété de compatibilité, même si elle peut être très utile pour réduire la modélisation, n'a aucun rôle fondamental à jouer. Il est seulement nécessaire de construire des modèles appropriés reliant les résultats associés aux unités expérimentales avec les résultats des unités non traitées jusqu'à présent, sous différentes assomptions sur la façon dont ces unités pourraient être traitées. Ces unités peuvent alors être utilisées pour faire des inférences prédictives sous les différentes assomptions, et ainsi apprécier la valeur relative des interventions futures.

VI.2 L'inférence sur les causes des effets

Pour Dawid l'ambiguïté inhérente à l'inférence sur les causes des effets peut être réduite si on peut approfondir les mécanismes *cachés* des unités en observant des variables supplémentaires *concomitantes* appropriées ; c'est là pour lui « la base et l'objet de la recherche scientifique ». Ceci n'est pas essentiel pour l'estimation des effets des causes, qui peut essentiellement reposer sur une approche de « boîte noire » en modélisant simplement la dépendance des réponses au fait que l'information d'une covariable se trouve ou non être observée pour l'unité test. Mais c'est au contraire vital pour toute étude d'inférence sur les causes des effets, qui doit prendre en compte ce qui a été appris à partir des expériences sur les mécanismes cachés de la boîte noire. Ainsi supposons qu'il soit possible de mesurer des variables concomitantes. Il peut s'agir de *covariables*, comme cela a déjà été envisagé précédemment. Cependant, d'autres quantités peuvent aussi être considérées, pourvu que l'on puisse supposer qu'elles ne sont *pas affectées* par le traitement appliqué (bien que l'usage même du terme « pas affectées » amène beaucoup de questions causales et contrefactuelles : voir plus loin). Typiquement la variation de la réponse conditionnellement aux variables

concomitantes sera plus petite que la variation non conditionnelle. De cette manière il pourra être possible de réduire l'intervalle d'ambiguïté pour la corrélation ρ (voir un exemple en annexe). En conséquence on obtiendra pour la moyenne λ et la variance ζ_D^2 de la distribution conditionnelle de l'effet causal individuel $D(u_0)$ des bornes plus resserrées que celles fournies précédemment en 5.3 (sans variable concomitante). Si on a observé $K(u_0) = k$ pour l'unité nouvelle, le gain est d'autant plus important que la variance résiduelle de Y étant donné k est plus petite. Mais même si K ne peut pas être observée pour u_0 son identification pour les unités expérimentales, au contraire de ce qui se passait dans ce cas pour les effets des causes, affecte l'analyse et permet encore de réduire l'intervalle d'ambiguïté.

A partir des possibilités offertes par les résultats précédents, le but fondamental de la recherche scientifique peut être vu selon Dawid comme la découverte d'une variable concomitante – qu'il appelle *suffisante* – qui produit la plus petite variance résiduelle qu'on puisse obtenir et permet donc d'obtenir l'intervalle d'ambiguïté le plus court possible pour la corrélation ρ. Une difficulté est qu'il ne sera généralement pas possible de savoir si on a vraiment obtenu la plus petite variance résiduelle possible (bien entendu la collection de *toutes* les variables concomitantes est toujours suffisante, mais on espère être capable de la réduire sans perte explicative). Néanmoins on peut encore faire des inférences scientifiquement valides (bien qu'imprécises) à partir de toutes les variables explicatives concomitantes découvertes. Cela prendra en compte le fait qu'il y a une composante d'incertitude ou d'arbitraire non statistique dans les inférences, exprimée par les bornes d'intervalles sur les conclusions causales quantitatives.

Il a toujours été supposé que les expériences réalisées étaient suffisamment précises (avec des effectifs très élevés) pour que l'incertitude purement statistique puisse être ignorée. En pratique cela est évidemment rarement le cas. Mais, une méthodologie appropriée pour combiner l'incertitude statistique avec l'ambiguïté intrinsèque de l'inférence sur les causes des effets n'est pas encore disponible. Dawid en conclut que des techniques pour traiter ce problème sont nécessaires et qu'il est urgent de pouvoir disposer de nouvelles méthodes d'inférence statistique combinant ambiguïté et incertitude. Il suggère que ces techniques pourraient être basées sur l'*indépendance conditionnelle* et propose la démarche à suivre (Dawid, 2000, p. 419).

VI.3 Déterminisme et pseudo-déterminisme

A partir des considérations précédentes, Dawid en vient à un certain nombre de remarques sur le déterminisme.

VI.3.1 Problèmes déterministes et contrefactuels

Dans certains problèmes des sciences « dures » il peut arriver, en prenant en compte suffisamment de variables concomitantes, que l'on puisse faire disparaître complètement la variation résiduelle dans la réponse (au moins en ce qui concerne les applications pratiques) ; on induit ainsi à un niveau plus fin la situation d'*uniformité* – considérée dans la Section IV.2 – où tous les problèmes d'inférence causale et de prédiction disparaissent[6]. Pour Dawid de tels problèmes peuvent être appelés *déterministes*, parce que la réponse est alors donnée par une fonction $Y = f(t,K^*)$ du traitement t et de la *variable concomitante déterminante* appropriée K^* (qui est alors nécessairement *suffisante*), sans aucune variabilité supplémentaire. Cette propriété peut en principe être testée quand K^* est donnée (si elle est rejetée, il est en principe possible de réintroduire cette propriété à un niveau plus profond en affinant la définition de K^*). Cependant, même quand un tel déterminisme sous-jacent existe, découvrir que tel est le cas et identifier la variable concomitante déterminante K^* ainsi que la forme de f peut s'avérer difficile en pratique ou même impossible. Cela nécessiterait en tout état de cause de très nombreuses investigations scientifiques, détaillées et coûteuses, ainsi que des analyses statistiques sophistiquées.

Si on avait un modèle déterministe on pourrait l'utiliser pour *définir* les réponses potentielles comme $Y_t(u) = f(t,K^*(u))$. On a ici la propriété que K^*, étant une variable concomitante, n'est pas affectée par le traitement ; mais parce que K^* n'est pas nécessairement une covariable, ce modèle n'est pas nécessairement fataliste. On pourrait déterminer la valeur de toute réponse potentielle pour l'unité u en mesurant $K^*(u)$. Dans ce cas particulier (mais exceptionnel) on peut donner une signification empirique aux contrefactuels : on peut en effet considérer alors les variables complémentaires $Y_t(u) \equiv f(t,K^*(u))$, pour une unité fixée u mais différents traitements t, comme ayant une *existence simultanée réelle*. Mais, même dans ce cas, la modélisation causale n'est pas basée sur une notion primitive de contrefactuel ; mais plutôt les contrefactuels sont basés sur le modèle et prennent leur signification à partir de lui.

VI.3.2 Le pseudo-déterminisme des modèles graphiques structurels

Pour Dawid la popularité des modèles contrefactuels repose sur une conception implicite que tous les problèmes d'inférence causale peuvent être moulés dans le paradigme déterministe (ce qui selon lui est seulement rarement approprié), pour une variable concomitante déterminante (généralement non observée) appropriée K^*. S'il en

[6] Dans l'exemple donné en annexe ceci se produirait si on trouvait $\psi_K^2 = 0$, ce qui impliquerait une corrélation $\rho = 1$ et éliminerait ainsi toute ambiguïté.

était ainsi cela servirait à justifier l'assomption de l'existence simultanée de réponses potentielles complémentaires. C'est la voie que suivent par exemple Heckerman et Shachter (1995), à partir de Savage (1954), qui basait son exposé de la théorie bayésienne de la décision sur l'existence supposée d'un « état de la nature » – non affecté par toutes les décisions prises – qui, avec ces décisions, détermine toutes les variables. Shafer (1986) a mis en avant certaines des faiblesses de cette conception. Celle-ci paraît pourtant à la base des modèles graphiques structurels ; ainsi Pearl commente l'équation structurelle « $y = \beta x + u$ » (voir plus haut Section III.1) en ces termes :

> *« In interpreting this equation one should think of a physical process whereby Nature examines the values of x and u and, accordingly, assigns variable Y the value* $y = \beta x + u$. *» (Pearl, 2001, p. 9)*

La critique fondamentale de Dawid est que souvent les « variables latentes » en jeu dans les modèles graphiques structurels *ne sont pas de véritables variables concomitantes* (variables mesurables, non affectées par le traitement). Il n'y a alors aucun moyen, même en principe, de vérifier les assomptions faites qui affecteront néanmoins les inférences qui en découlent, au mépris de la loi de Jeffreys. En conséquence Dawid qualifie de tels modèles de *pseudo-déterministes* et les considère comme *non scientifiques* :

> *« I term such functional models* pseudodeterministic *and regard it as misleading to base analyses on them. In particular, I regard it as unscientific to impose intrinsically unverifiable assumed forms for functional relationships, in a misguided attempt to eliminate the essential ambiguity in our inferences. » (Dawid, 2000, p. 422)*

En fait, dans le cadre contrefactuel il est toujours possible de construire, *mathématiquement*, un modèle pseudo-déterministe : il suffit de définir $K^*(u)$ comme la collection complémentaire de tous les résultats potentiels pour l'unité u. Ainsi dans notre exemple de base, on prendrait $K^* = (Y_{t1},Y_{t2})$. On a alors la relation fonctionnelle déterministe triviale $Y = f(t,K^*)$, où f a la *forme canonique* $f(t,(Y_{t1}, Y_{t2})) = Y_t$. Si on fixe une distribution conjointe pour (Y_{t1},Y_{t2}), alors l'analyse présentée précédemment pour inférer les « causes des effets » dans les modèles déterministes pourrait formellement être appliquée. Mais ce n'est pas un vrai modèle déterministe : K^* n'est pas une véritable variable concomitante parce qu'elle n'est pas, même en principe, observable. La construction d'un tel modèle pseudo-déterministe ne fait absolument pas avancer la résolution des problèmes de non-unicité de l'inférence (exposés précédemment à

propos de la « situation embarrassante » dans la Section V.2). La conclusion de Dawid est qu'aucune somme d'investigations scientifiques ne suffira pour justifier toute structure de dépendance que l'on pourrait supposer pour (Y_{t1}, Y_{t2}), ou pour éliminer la sensibilité à cette structure des inférences sur les causes des effets. Cela ne peut être fait qu'en prenant en compte des variables concomitantes *véritables*.

VI.3.3 Le contexte de l'inférence

Dawid admet qu'en basant l'inférence sur les causes des effets sur des variables concomitantes, il semble déroger à son insistance sur le fait que des assomptions métaphysiques ne devraient pas pouvoir affecter les inférences. En effet dire qu'une variable est concomitante constitue une assomption clairement non testable empiriquement et les inférences causales dépendront des assomptions faites sur quelles variables doivent être traitées comme concomitantes. Cet arbitraire vient en sus de l'ambiguïté inférentielle essentielle identifiée précédemment (Section V.2). La position de Dawid est qu'il y a en effet un arbitraire dans les modèles que l'on peut raisonnablement utiliser pour faire des inférences sur les causes des effets, et de là dans les conclusions qui sont justifiées. Mais il considère cela comme lié, au moins en partie, aux différences dans la nature des questions traitées. L'essence d'une recherche causale spécifique est prise en compte dans la spécification largement conventionnelle de ce qu'il dénomme le contexte de l'inférence – c'est-à-dire la collection des variables que l'on considère qu'il est approprié de considérer comme concomitantes. Pour Dawid une spécification appropriée du contexte, adaptée à un objectif spécifique, est vitale pour donner véritablement sens aux questions causales et aux réponses ; elle peut être regardée comme apportant une clarification nécessaire de la clause *ceteris paribus* (« toutes choses étant égales par ailleurs ») souvent invoquée dans les tentatives pour expliquer l'idée de cause. Différents objectifs demanderont différentes spécifications du contexte, nécessitant des approches scientifiques et statistiques différentes et apportant des réponses différentes ; et les inférences causales devront être révisées quand la spécification du contexte pertinent devra être reformulée. En particulier, savoir s'il est raisonnable d'utiliser un modèle déterministe doit dépendre du contexte du problème traité, puisque cela spécifiera s'il est approprié de considérer une variable déterminante putative K^* comme étant véritablement concomitante, c'est-à-dire non affectée par le traitement. Pour différents contextes on pourrait avoir des modèles différents, certains déterministes (mettant en jeu différentes définitions de K^*) et certains non-déterministes.

VI.3.4 La conclusion de Dawid

Dawid (2000) présente une critique sévère, non seulement à l'égard de l'approche contrefactuelle, mais aussi à l'encontre des modèles graphiques structurels défendus par Pearl. Il considère que l'introduction de quantités contrefactuelles dans les modèles statistiques est dangereuse et doit être évitée. Il propose une approche alternative basée sur une analyse décisionnelle, qu'il considère naturellement séduisante et pleinement scientifique. Il affirme que sa solution est entièrement satisfaisante pour traiter le problème de l'inférence sur les *effets des causes*, et que l'approche « boîte noire » familière des statistiques expérimentales est parfaitement adaptée à cet objectif. L'inférence sur les *causes des effets* soulève de plus grandes difficultés. Une solution complètement non ambiguë ne peut être obtenue que dans les cas exceptionnels où il est possible de parvenir à une compréhension scientifique suffisante du système étudié permettant l'identification de mécanismes causaux essentiellement déterministes (reliant les réponses aux interventions et à des variables concomitantes définies de façon appropriée). Quand cela n'est pas réalisable (que les difficultés pour le faire soient fondamentales ou seulement pragmatiques) les inférences justifiées, même par des données très nombreuses, ne sont pas déterminées de manière unique et on doit se contenter d'inégalités. Cependant ces inégalités peuvent être affinées en modélisant le contexte pertinent et en réalisant des expériences dans lesquelles des variables concomitantes appropriées sont mesurées. Des investigations scientifiques importantes et détaillées peuvent toutefois être nécessaires pour réduire l'ambiguïté résiduelle à son minimum (sans que l'on ait nécessairement la garantie *a priori* de pouvoir y parvenir).

L'article de Dawid est suivi de nombreux commentaires de spécialistes du domaine. La plupart des opposants n'avancent que des arguments de principe, rejetant notamment la demande de Dawid de s'en tenir, dans les analyses statistiques, à des termes susceptibles d'être testés (au moins en principe) empiriquement. Un argument, avancé par Robins et Greenland (p. 434), est que pour Popper tout élément d'une théorie n'a pas à être testable pour que cette dernière puisse être qualifiée de scientifique. Pearl (p. 430) rappelle que la créativité et l'imagination sont essentielles au progrès scientifique et que des concepts qui ont un caractère essentiellement hypothétique à un moment donné du développement d'une science peuvent parfaitement être « validés » par la suite. Il donne notamment comme exemple, dans le domaine des sciences expérimentales, la notion d'atome en physique. Cependant, dans le cas des contrefactuels, cet argument n'apparaît pas pertinent, puisqu'ils sont de toute façon inobservables *par définition* (alors que dans le cas de l'atome rien ne permettait d'affirmer qu'un jour on ne pourrait pas l'observer s'il existait).

Pour sa part, Cox (p. 424) critique l'approche *décisionnelle* de Dawid en mettant en avant qu'il lui semble que les expérimentateurs (dans les essais cliniques) cherchent à progresser dans la compréhension d'un phénomène, et non simplement à décider quel traitement attribuer. Mais Dawid insiste bien sur la nécessité, surtout en ce qui concerne les causes des effets, de progresser dans la connaissance scientifique du phénomène, sa conclusion étant que, si l'on veut obtenir des assertions véritablement pleines de sens et utiles sur les causes des effets, alors il faut être très clair sur la signification et le contexte des questions et faire de la « science véritable » :

> *« And then there is no magical statistical route that can bypass the need to do real science to attain the clearest possible understanding of the operation of relevant (typically nondeterministic) causal mechanisms. » (Dawid, 2000, p. 423)*

Il apparaît là rejoindre Fisher qui à la question qu'on lui posait dans une conférence sur ce qui pourrait être fait dans le cas de données d'observation, répondit: « make your theories elaborate »

> *« On being asked at a conference what could be done to make observational studies yield more nearly causal conclusions, Fisher replied: make your theories elaborate. » (propos mentionnés par Cochran dans un papier lu à la Royal Statistical Society et rapportés par Cox, in Schaffner, 1993, p. 1495)*

CONCLUSION

Même dans le cadre d'une expérimentation rigoureuse avec des données très nombreuses supprimant l'incertitude purement statistique, obtenir des inférences statistiques utiles sur les causes des phénomènes observés, notamment sur les effets individuels, peut nécessiter des investigations considérables. Cela nécessite en tout état de cause une analyse critique approfondie ; une question essentielle pour l'expérimentateur est de savoir si l'inférence statistique causale requiert des outils et des techniques autres que ceux disponibles dans le cadre usuel du calcul des probabilités. Pour Pearl (2001) ce cadre est insuffisant et il est nécessaire de l'étendre par une conceptualisation appropriée, apportée par les modèles graphiques structurels. Cette approche peut paraître séduisante à première vue, mais Dawid (2000) argumente au contraire qu'il n'y a rien à gagner à une telle extension qui est à la fois non nécessaire et indésirable, surtout en ce qui concerne l'introduction d'éléments contrefactuels (voir aussi Dawid, 2002a). Plus récemment il a démontré formellement les avantages d'une modélisation directement basée sur les

concepts probabilistes ; il a montré en particulier que même dans les cas où les modèles contrefactuels permettent d'obtenir des résultats importants, les mêmes résultats peuvent être exprimés plus clairement et dérivés plus facilement grâce à l'utilisation de diagrammes d'influence probabilistes appropriés, sans aucune référence à des relations fonctionnelles ou à des contrefactuels (Dawid, 2002b). Les seules analyses qui ne peuvent pas être exprimées de cette manière sont celles où la « réponse » dépend de la manière arbitraire par laquelle on choisit entre des modèles fonctionnels *indistinguables par l'observation*.

Pour Dawid nous devons précisément nous garder de poser des questions qui apparaissent bien formulées, mais sont en fait scientifiquement dénuées de sens. Avec les approches comme celles des modèles graphiques structurels qui laissent implicites beaucoup d'assomptions nécessaires à l'inférence, il y a un grand risque d'obtenir pour ces questions des réponses qui paraissent pleines de sens parce qu'elles peuvent être exprimées *mathématiquement* dans les termes des ingrédients que nous avons choisis d'inclure dans notre modèle ; mais ces réponses ne reposent en fait sur aucune base scientifique, puisque ces ingrédients sont eux-mêmes largement arbitraires et que cet arbitraire peut se retrouver dans nos conclusions. Il est important de pouvoir identifier ces questions, et ce n'est certainement pas un recul scientifique – bien au contraire – de reconnaître qu'il y a sans doute des questions auxquelles des données empiriques ne peuvent permettre de répondre de manière non ambiguë.

Références

Angrist, J.D., Imbens, G.W., Rubin, D.B. (1996). Identification of causal effects using instrumental variables (with comments). *Journal of the American Statistical Association*, 91, pp. 434-444.

Balke, A.A., Pearl, J. (1994). Probabilistic evaluation of counterfactual queries, in *Proceedings of the Twelth National Conference on Artificial Intelligence*, vol. 1, Menlo Park, CA: MIT press, pp. 230-237.

Balke, A.A., Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92, pp. 1172-1176.

Barnard, G.A. (1982). Causation, in Kotz, S., Johnson, N. et Read, C. (éds.), *Encyclopedia of Statistical Science, vol. 1*, New York: Wiley, pp. 387-389.

Chickering, D.M., Pearl, J. (1997) A clinician's tool for analysing non-compliance, *Computing Science and Statistics*, 29, pp. 424-431.

Cowell, R.G., Dawid, A.P., Lauritzen, S.L., Spielgelhalter, D.J. (1999). *Probabilistic Networks and Expert Systems*, New York, NY: Springer Verlag.

Dawid, A.P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society* B, 41, pp. 1-31.

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

Dawid, A.P. (2002). Counterfactuals: Help or hindrance? (invited discussion of « Estimating causal effects », by G. Maldonado and S. Greenland), *International Journal of Epidemiology*, 31, pp. 429-430. Corrigenda, ibid., p. 437.

Dawid, A.P. (2002b). Influence diagrams for causal modelling and inference, *International Statistical* Review, 70, pp. 161-189.

Evans, A.S. (1993). *Causation and Disease: A Chronological Journey*, New York: Plenum.

Fisher, R.A. (1918). The causes of human variability, *Eugenics Review*, 10, pp. 213-220.

Fisher, R.A. (1959). *Smoking. The cancer controversy*, Edindurgh: Oliver and Boyd.

Fisher, R.A. (1990/1935). *The Design of Experiments* (8ème édition de 1966), in Bennett, J.H. (éd.), *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford: Oxford University Press (1ère édition: 1935, London: Oliver and Boyd.).

Fisher, R.A. (1990/1956). *Statistical Methods and Scientific Inference* (3ème édition de 1973), in Bennett, J.H. (éd.), *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford: Oxford University Press (1ère édition: 1956, London: Oliver and Boyd.)

Greenland, S., Robins, J.M. (1986). Identifiability, exchangeability, and epidemiological confounding, *International Journal of Epidemiology*, 15, pp. 413-419.

Greenland, S., Robins, J.M., Pearl, J. (1999). Confounding and collapsibility in causal inference, *Statistical Science*, 14, pp. 29-46.

Heckerman, D., Shachter, R. (1995). Decision-theoretic foundations for causal reasoning, *Journal of Artificial Intelligence Research*, 3, pp. 405-430.

Heckman, J.J., Smith, J. (1998). Evaluating the welfare state, in Strom S. (éd.), *Econometric and Economic Theory in the 20th Century*, Cambridge: Cambridge University Press, pp. 1-60.

Holland, P.W. (1986). Statistics and causal inference (with comments and rejoinder), *Journal of the American Statistical Association*, 81, pp. 945-970.

Holland, P.W. (1988). Causal inference, path analysis, and recursive structural equations models, in Clogg, C. (éd.), *Sociological Methodology*, Washington, D.C.: American Sociological Association, pp. 449-484.

Imbens, G.W., Angrist, J.D. (1994). Identification and estimation of local average treatment effects, *Econometrica*, 62, pp. 467-475.

Imbens, G.W., Rubin, D.B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance, *The Annals of Statistics*, 25, pp. 305-327.
Jeffreys, H. (1998/1939). *Theory of Probability* (3ème édition), Oxford: Oxford University Press. (1ère édition : 1939)
Kempthorne, O. (1952). *The Design and Analysis of Experiments*, New York: Wiley.
Lauritzen, S.L. (1996). *Graphical Models*, Oxford: Clarendon Press.
Lecoutre, B., Lecoutre, M.-P., Poitevineau, J. (2001). Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? *International Statistical Review*, 69, pp. 399-417.
Lecoutre, B., Poitevineau, J. (2000). Aller au-delà des tests de signification traditionnels : Vers de nouvelles normes de publication, *L'Année Psychologique*, 100, pp. 683-713.
Lehmann, E.L. (1959). *Testing Statistical Hypotheses*, New York: Wiley.
Lewis, D. (1973). Causation, *Journal of Philosophy*, 70, pp. 556-67.
Lewis, D. (1986). *Philosophical Papers: Volume II*, Oxford: Oxford University Press.
Lewis, D. (2000). Causation as Influence, *Journal of Philosophy*, 97, pp. 182-97.
Lindley, D. V. (1972). *Bayesian Statistics: A Review*, Philadelphia: Siam, Regional Conference Series in Applied Mathematics.
Manski, C.F. (1995). *Identification Problems in the Social Sciences*, Cambridge, MA: Harvard University Press.
McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models* (2ème édition), London: Chapman and Hall.
Menzies, P. (1989). Probabilistic Causation and Causal Processes: A Critique of Lewis, *Philosophy of Science*, 56, pp. 642-663.
Neyman, J. (1990/1923). On the application of probability theory to agricultural experiments - Essay on principles, Section 9 (traduit et édité par D.M., Dabrowska et T.P., Speed à partir du texte polonais paru en 1923), *Statistical Science*, 5, pp. 465-480.
Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*, San Mateo, CA: Morgan Kaufmann.
Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82, pp. 669-710.
Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*, New York: Cambridge University Press.
Pearl, J. (2001). Causal inference in the health sciences: A conceptual introduction, *Health Services and Outcomes Research Methodology* (special issue on causal inference), 2, pp. 189-220.
Pearson, K. (1911). *Grammar of science*, London: A and C. Black.
Press, S.J. (1989). *Bayesian Statistics: Principles, Models, and Applications*, New York: Wiley.
Robins, J.M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period: Applications to control of the

healthy workers survivor effect, *Mathematical Modeling*, 7, pp. 1393-1512.

Robins, J.M. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods, *Journal of Chronic Diseases*, 40 (Suppl. 2), pp. 139S–161S.

Robins, J.M., Greenland, S. (1989). The probability of causation under a stochastic model for individual risk, *Biometrics*, 45, pp. 1125-1138.

Robins, J.M., Wasserman, L.A. (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs, Technical Report 654, Carnegie Mellon University, Department of Statistics.

Rosenbaum, P.R., Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, 70, pp. 41-55.

Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P., Le Roux, B. (2000). *New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference* (2ème édition), Bern, CH: Peter Lang.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies, *Journal of Educational Psychology*, 66, pp. 688-701.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization, *The Annals of Statistics*, 6, pp. 34-58.

Rubin, D.B. (1980). Discussion of « Randomization analysis of experimental data: The Fisher randomisation test » by D. Basu, *Journal of the American Statistical Association*, 75, pp. 591-593.

Rubin, D.B. (1986). Which Ifs have causal answers. Discussion of « Statistics and causal inference » by P.W. Holland, *Journal of the American Statistical Association*, 81, pp. 961-962.

Savage, L.J. (1954). *The Foundations of Statistics*, New York: Wiley.

Shafer, G. (1986). Savage revisited (with discussion), *Statistical Science*, 4, pp. 463-501.

Schaffner, K.F. (1993). Clinical trials and causation: Bayesian perspectives (with discussion), *Statistics in Medicine*, 12, pp. 1477-1499.

Schwartz, D., Flamant, R., Lellouch, J. (1981). *L'essai Thérapeutique chez l'Homme* (2ème édition), Paris: Flammarion.

Student (1923). On testing varieties of cereals, *Biometrika*, 15, pp. 271-293.

Suppes, P. (1970). *A Probabilistic Theory of Causality*, Amsterdam: North-Holland Publishing Company.

Urbach, P. (1993). The value of randomization and control in clinical trials (with discussion), *Statistics in Medicine*, 12, pp. 1421-1441.

Wright, S. (1921). Correlation and causation, *Journal of Agricultural Research*, 20, pp. 557-585.

Wright, S. (1934). The method of path coefficients, *Annals of Mathematical Statistics*, 5, pp. 161-215.

Yates, F. (1935). Complex experiments, *Journal of the Royal Statistical Society - Supplement*, 2, 181-223.

ANNEXE : Exemple de réduction de l'intervalle d'ambiguïté pour la corrélation ρ dans le cas de l'inférence sur les « causes des effets »

[Le traitement *t1* a déjà été appliqué et la réponse $Y_{t1}(u_0) = y_{t1}$ a été observée]

Supposons que

- on a réalisé une expérience (toujours avec des données très nombreuses) plus détaillée, où on a mesuré une variable concomitante K ;

- on a montré que, conditionnellement à $K(u) = k$ et à l'application du traitement t, la réponse $Y_t(u)$ a la distribution

$$Y_t(u) \,|\, \mu_t, \psi_K^2, K(u) = k \sim \mathrm{N}(\mu_t + k, \psi_K^2)$$

[Il s'agit ici d'un modèle très simple mais la logique essentielle du raisonnement ci-après continuerait à s'appliquer pour des modèles plus complexes]

Cette expérience a permis d'estimer avec précision les valeurs des paramètres μ_{t1}, μ_{t2} et ψ_K^2.

Conséquences

Définissons $\sigma_K^2 = \mathrm{var}(K)$ et posons $\psi_0^2 = \sigma^2 = \sigma_K^2 + \psi_K^2$. Alors $\mathrm{cov}(K, Y_{t2}) = \mathrm{cov}(K, Y_{t1}) = \sigma_K^2$.

En combinant ceci avec la structure de covariance pour le couple complémentaire (Y_{t1}, Y_{t2}) impliquée par le modèle métaphysique [2], on obtient la matrice de variances-covariances complète de (K, Y_{t1}, Y_{t2}) :

$$\begin{bmatrix} \sigma_K^2 & \sigma_K^2 & \sigma_K^2 \\ \sigma_K^2 & \sigma^2 & \rho\sigma^2 \\ \sigma_K^2 & \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

Donc la corrélation conditionnelle entre Y_{t1} et Y_{t2} étant donné K est :

$$\rho_{t1,t2|K} = (\rho\sigma^2 - \sigma_K^2) / (\sigma^2 - \sigma_K^2) = 1 - (1 - \rho)(\psi_0^2 / \sigma_K^2)$$

Le paramètre arbitraire $\rho_{t1,t2|k}$ ne peut pas être identifié à partir de l'expérience plus détaillée (bien qu'il puisse être raisonnable de prendre $\rho_{t1,t2|K} \geq 0$) ?

Considérons maintenant l'inférence sur les « causes des effets » sur une unité test u_0.

(1) On a effectivement observé $K(u_0) = k$

On peut mener une analyse analogue à celle effectuée dans le cas où il n'y avait pas de variable concomitante :

$$D(u_0) \mid \delta,\sigma^2,\rho,Y_{t1}(u_0)=y_{t1},K(u_0)=k \sim \mathrm{N}(\lambda, \zeta_D^2)$$

avec

$$\lambda = (y_{t1} - \mu_{t2} - k) - \rho_{t1,t2|K}(y_{t1} - \mu_{t1} - k)$$

[au lieu de $y_{t1} - \mu_{t2} - \rho(y_{t1} - \mu_{t1})$ précédemment]

et

$$\zeta_D^2 = (1 - \rho_{t1,t2|K}^2)\sigma_K^2$$

[au lieu de $(1-\rho^2)\sigma^2$ précédemment]

La moyenne λ, parce que le terme final entre parenthèses est maintenant d'ordre $\sqrt{\psi_K^2}$, plutôt que $\sqrt{\psi_0^2}$ comme précédemment, devrait être moins sensible à l'arbitraire dans la corrélation, maintenant $\rho_{t1,t2|K}$.

De manière similaire, la variance ζ_D^2 est maintenant bornée supérieurement par $\sigma_K^2 < \psi_0^2$, au lieu de $\sigma^2 = \psi_0^2$

Clairement ces gains sont plus importants avec une plus petite variance résiduelle σ_K^2 de Y étant donné k.

(2) On n'a pas observé $K(u_0) = k$

L'analyse *est affectée* par les résultats plus détaillés de l'expérience.

Définissons $\gamma_K = \sigma_K^2/\sigma^2 = 1 - \psi_K^2/\psi_0^2$

La corrélation conditionnelle entre Y_{t1} et Y_{t2} étant donné K implique :

$$2\gamma_K - 1 \le \rho \le 1$$ [au lieu de $-1 \le \rho \le 1$ précédemment]

ou si on suppose $\rho_{t1,t2|K} \ge 0$:

$$\gamma_K \le \rho \le 1$$ [au lieu de $-1 \le \rho \le 1$ précédemment]