Recueil du Symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherché économique et sociale

CERTAINS GRANDS PROBLÈMES ET DÉFIS EN STATISTIQUE

Professeur David Cox[1]

Je suis toujours très heureux de venir au Canada. En un sens, cette visite est empreinte de nostalgie pour moi. La première fois que je suis venu dans votre pays il y a un peu plus de 39 ans, je suis descendu à cet hôtel; l'Institut international de la statistique y tenait ses séances en 1963. J'ai gardé de très bons souvenirs de ma visite sur le double plan personnel et scientifique.

Au fil des ans, je croyais avoir conçu avec succès toutes sortes de techniques pour trouver des titres neutres et non contraignants à mes exposés. Il y a déjà un certain temps, on m'a invité à venir ici et, relativement peu de temps après, on m'a demandé un titre. Je pensais avoir trouvé un de ces titres neutres. (J'ai en fait présenté un résumé qui l'est tout à fait, comme vous pourrez le constater si vous regardez votre programme.) J'ai toutefois eu tort d'agir ainsi, puisque mon propos sera certes de traiter d'un certain nombre de problèmes généraux, mais plus précisément du problème de la causalité en statistique. D'une certaine manière, cette question résume un certain nombre d'autres questions. En d'autres termes, posons-nous la question suivante : « Que signifient nos analyses en réalité une fois faites? »

Si nous regardons en général les études spécialisées en statistique, la chose curieuse à propos de la causalité est qu'elle ressemble au chien de Charlotte Home. Ce chien était intéressant parce qu'il n'aboyait pas la nuit. De même, on mentionne à peine la causalité : peut-être y voit-on de la philosophie, de la philosophie dont on doit se tenir loin? On peut aussi penser que la causalité est une question spécialisée; c'est vrai d'une certaine manière. Mon propos sera donc de tâcher de mieux lier les résultats des analyses statistiques à l'interprétation spécialisée qui leur est sous-jacente.

Si nous retenons cette façon plutôt négative de voir les choses, mais revenons aux origines de l'enseignement de la statistique, il est fort possible que nous retrouvions le mot d'ordre « Corrélation n'est pas causalité ». C'est peut-être la seule fois que le mot causalité ait jamais été employé.

Corrélation n'est pas causalité. Par inférence, on peut tirer un rapport de causalité d'expériences aléatoires, mais non pas d'études par observation. C'est la chose qui pourrait venir après, et c'est considérablement plus qu'une demi-vérité; c'est peut-être une vérité à 95 % ou 98 % dont nous parlerons un peu plus loin. C'est donc ce qu'on trouve dans les études statistiques en général.

Mais plus on se rapproche des sciences sociales ou plus on considère les études quantitatives en sociologie, en économétrie et sûrement en épidémiologie, et plus la causalité devient sujet de discussion. Les mots ont des usages qui varient grandement selon les gens. Il n'y a pas que les façons d'utiliser les mots qui varient, mais aussi le soin qu'on y met. Outre les domaines énumérés, je pourrais mentionner la discipline de l'intelligence artificielle où on fait grand usage de modèles de causalité. Il y a des gens qui emploient plutôt librement le terme « causal ».

En revanche, les épidémiologistes (les meilleurs du moins) distinguent très soigneusement ce qu'ils appellent les facteurs de risque (qui, pour une maladie, sont des indicateurs d'association) de tout jugement de causalité. Ainsi, les définitions et les niveaux d'utilisation du terme causalité abondent, et il importe de comprendre ce phénomène lorsqu'on regarde les études spécialisées.

Je donnerai de la causalité trois définitions liées les unes aux autres, mais qui sont aussi à distinguer.
Il y a d'abord ce que j'appellerai la « causalité de niveau un ». Nous avons ce que nous tenons pour une cause possible; par souci de simplicité, je supposerai que c'est un facteur binaire. Il peut s'agir de gens que

[1] Nuffield College, Oxford, Royaume-Uni

nous voulons étudier et de qui nous obtenons une réponse ou non. Il peut s'agir d'enfants éduqués par l'une ou l'autre de deux méthodes pédagogiques avec une mesure de réussite dans une matière. Il peut s'agir de criminels visés par l'une ou l'autre de deux méthodes punitives avec une mesure R de la récidive. Il y a aussi le cas de sujets d'essais cliniques traités par l'une ou l'autre de deux méthodes, C0 et C1, avec une réponse qui serait une mesure de leur survie, par exemple. L'idée est qu'une personne pourrait conceptuellement être C0 ou C1. Même si on enseigne à un enfant par la méthode C0, celui-ci aurait théoriquement pu avoir C1 comme méthode pédagogique.

Il est question de causalité dans ce contexte si la réponse pour la méthode C1 diffère systématiquement de la réponse pour la méthode C0, toutes choses étant égales. Je mets à dessein cette dernière expression en italique. En d'autres termes, ce qui est immédiatement en cause, c'est l'essai clinique aléatoire où un patient peut recevoir l'un ou l'autre de deux traitements possibles. Il en reçoit un en réalité, mais pourrait avoir reçu l'autre si le hasard en avait décidé autrement. Dans le cas d'une étude criminologique, un criminel peut être l'objet ou non de mesures de justice réparatrice. Il a peut-être droit à de telles mesures, mais il pourrait en être autrement. Le monde aurait pu être différent de ce qu'il est en réalité.

La causalité est l'idée, toutes choses étant égales (et c'est une condition primordiale), qu'il existe une différence systématique entre ces deux réponses. Plus précisément, nous pouvons parler de deux réponses possibles, R0 ou R1, des gens selon le traitement qui s'applique, C0 ou C1. Nous pouvons observer un des éléments de cette alternative, mais un seul. L'autre est ce que les philosophes appellent un « contrefactuel » ou un conditionnel contraire aux faits. C'est quelque chose qui peut se concevoir, mais qui ne peut directement s'observer. Il nous faut alors formaliser la notion que R1 est systématiquement différent de R0. Pour être plus précis, je parlerai de R1 plus grand que R0. On pourrait dire que, pour tous les gens, R1 est supérieur à R0. Quiconque subit C1 s'en tire mieux que s'il subissait C0, pour l'exprimer de cette façon. On peut aller plus loin et affirmer que la différence entre ces R est une constante. Une version plus réaliste dans bien des contextes consisterait à dire que R1 est stochastiquement supérieur à R0. En d'autres termes, la fonction de distribution cumulative de R1 est décalée à droite par rapport à la fonction de distribution cumulative de R0. C'est bien moins que de dire que, pour chaque personne, l'inégalité tient. Dans une version plus faible encore, ce serait les moyennes qui différeraient. Dans les deux derniers cas, c'est une question de probabilités avec, implicitement ou peut-être explicitement, l'idée d'une population cible à laquelle l'énoncé s'applique.

Dans le prolongement de cet examen de la définition, il convient de noter qu'on se trouve à restreindre le choix de variables admissibles susceptibles d'être considérées comme causales, puisque ce serait dire : « Nous avons observé quelqu'un à C1, mais il aurait pu être à C0. » Le monde aurait pu être différent de ce qu'il est.

On m'a déjà servi un exemple semblable. C'était presque au tout début de ma carrière en recherche. Je travaillais avec un médecin qui me disait que le temps qui passe n'était pas une cause. David Binder se rappellera sans doute que, plus jeune, j'avais les cheveux brun foncé. Je m'en souviens à peine, mais David se le rappellera probablement bien. Dans la perspective que nous nous sommes donnée, il serait faux de dire que le temps qui passe a fait tourner mes cheveux au gris : quel serait le sens de cette affirmation dans notre définition? Cela signifierait qu'en partant il y a longtemps de mes cheveux bruns, le temps n'a pas passé, mais nous voici aujourd'hui et mes cheveux sont gris. Cet énoncé n'a aucun sens, parce qu'il est conceptuellement impossible d'opérer ce transfert du passé au présent sans dire que le temps a passé.

On pourrait dire que le stress que me cause la préparation de mes conférences m'a fait grisonner. La question n'est pas de savoir si cet énoncé est vrai, mais plutôt s'il est significatif. Il est significatif, parce que le temps pourrait avoir passé. Je pourrais ne pas avoir préparé de conférences. Je pourrais avoir sombré dans l'oisiveté, avoir lu nos « Annals of Statistics » ou quelque chose d'autre et peut-être alors n'aurais-je jamais grisonné. À mon avis, cet énoncé est à la fois vrai et significatif. Voilà le point. Cela a une incidence sur les questions d'interprétation.

Les différences entre les sexes auront rarement cette vertu causale, par exemple. On peut dire : « Voyez cette femme. Quelle aurait été la réponse (toutes choses étant égales) si cette femme avait été un homme? »

Dans la plupart des contextes, cette phrase n'a aucun sens. Dans le contexte de la discrimination en matière d'emploi, elle pourrait fort bien avoir un sens, mais non si on la porte sur un plan plus général.

Il y a donc restriction quant aux causes admissibles. Cette façon d'aborder les choses a un caractère technologique implicite, mais non pas explicite. C'est l'idée que nous nous intéressons aux grandes décisions sur les modes de traitement des patients ou aux politiques économiques en nous demandant, par exemple, si une hausse du taux d'escompte ou du taux d'intérêt peut avoir un effet sur l'économie et quel pourrait être cet effet. C'est presque parler du monde des décisions, mais tel n'est pas nécessairement le cas. Un généticien pourrait fort bien dire qu'une combinaison particulière d'allèles du génome cause telle ou telle maladie. (Les généticiens ne s'expriment pas de cette façon.) Cela veut dire que, si l'haplotype d'une personne avait été différent, le résultat aurait été différent lui aussi. Ce n'est plus nécessairement le monde des décisions, mais il y a comme une décision qui se glisse là-dedans.

Je mettrai à votre disposition une liste de documents de référence. Ce sera plus que ce que je mentionne dans mon exposé. Le meilleur document à consulter si le sujet est nouveau pour vous est une recension, un document de travail qu'a publié Paul Holland il y a quelques années déjà dans le Journal of the American Statistical Association. L'idée générale remonte en réalité aux origines de la théorie des plans d'expérience dans les travaux de Neyman et Fisher. Cochran a produit, au sujet de l'analyse et de l'interprétation des études par observation, un très important rapport qui met en évidence un grand nombre des notions en cause, mais c'est Don Rubin en particulier – qui a travaillé avec Cochran, bien sûr – qui a développé le thème et l'a appliqué à toutes sortes de contextes en sciences sociales. Jamie Robins a travaillé plus ce thème dans un contexte épidémiologique, mais ses idées sont assez semblables.

La notion de « contrefactuel » désarçonne quelque peu bien des gens. Si vous considérez que les seules choses défendables qui importent vraiment sont celles que l'on peut directement observer, vous pouvez vous débarrasser de tout, variables et autres choses semblables. Si vous pensez devoir vérifier directement par expérimentation ce que vous observez, la notion de contrefactuel n'aura rien pour vous plaire en soi. Dans une très intéressante étude, Phil David a critiqué sous cet angle une telle conception de la causalité. Là encore, il y a un article intéressant dans le Journal of American Statistical Association où cette conception de la causalité est largement reprise. C'est l'idée que, si le monde était différent de ce qu'il est (C), la réponse (R) serait aussi différente. Si nous intervenions sur le C, nous changerions le monde pour le mieux, espérons-le. C'est ce que j'appelle la causalité de niveau un.

La causalité de niveau deux est, en un sens, quelque chose de fort différent, en apparence du moins. Elle a à voir avec l'idée que nous observons une certaine régularité, une certaine régularité statistique, voire même un rapport causal, au sens que je viens d'évoquer. Et on peut dire que la raison en est que certains processus, dont nous avons une compréhension au moins partielle, entrent en jeu. C'est donc une « compréhension par les faits » de tels processus générateurs.

Prenons un exemple. On peut recueillir une foule de données sur les tempêtes, leur durée, leur ampleur, le nombre d'éclairs, leur épicentre, etc.; on pourra peut-être déduire certaines régularités statistiques qui sont très nettes, très stables et très intéressantes. Quelqu'un se demandera alors inévitablement quel est le phénomène physique qui joue. On voudra savoir comment tout cela est lié aux équations fondamentales de la physique, de la dynamique des fluides et de l'électromagnétisme? Et quelqu'un comprendra alors la cause de ces rapports par une véritable compréhension des causes.

Un collègue à Oxford, le sociologue John Goldthorpe, maintient que c'est la bonne définition ou la définition la plus féconde pour la sociologie aussi, s'inspirant ainsi à l'extrême de la physique. Les régularités statistiques qu'établissent les sociologues devraient pouvoir trouver une explication dans la psychologie de la personne ou une théorie comme celle des choix rationnels. C'est là une notion de compréhension. La chose est inévitablement provisoire et, par sa nature même, elle appelle des études à divers niveaux.

J'ai déjà fait mention du travail fort important de Cochran sur les études par observation en 1965. Celui-ci y cite une conversation qu'il a eue avec R.A. Fisher où il lui a demandé : « Que pourrait-on faire pour rendre plus causales les études et les analyses par observation? » Son interlocuteur lui a fait comme d'habitude une

réponse de sphinx : « Développez vos théories. » Bien sûr, ce « Développez vos théories » est plutôt contraire au rasoir d'Occam et à la quête de simplicité. Développez vos théories. Ce que Fisher entendait par là, explique Cochran, c'est qu'il faut réunir toutes sortes de renseignements, toutes sortes d'études, toutes sortes de compréhensions dans un cadre cohérent de manière à mieux pouvoir cerner les rapports de causalité.

Le philosophe Bertrand Russell (pensant comme les philosophes des sciences l'ont très souvent fait, du moins à la première époque, en s'employant dans une très large mesure à faire correspondre la science à la physique) a écarté au départ la notion de causalité, du moins dans ce sens, comme étant inutile. « À mon avis, c'est comme la monarchie, devait-il écrire. Elle a survécu à sa propre utilité et, si elle s'est ainsi attardée, c'est parce qu'on croyait faussement qu'elle ne causait aucun tort. » Il a changé ses vues sur la causalité, je pense, mais j'ignore s'il s'est aussi ravisé au sujet de la monarchie. De toute manière, efforçons-nous de développer nos théories.

Voilà donc – du moins en apparence – une notion très différente de la causalité qui s'ancre dans la compréhension. J'ai l'impression que c'est l'idée même que la plupart des spécialistes des sciences naturelles s'en feraient s'ils employaient le terme. Le plus souvent, ils ne le font guère, mais c'est ce qu'ils entendraient par là, si on les pressait un peu sur ce point. Je souligne l'expression « compréhension par les faits », car il faut bien dire que, après tout, l'agile esprit humain peut penser à une explication pour tout phénomène, et ce, dans presque tous les domaines. Quelque bizarre que soit un phénomène, quelqu'un concoctera une explication. C'est peut-être très utile comme première étape, mais l'explication, en ce sens du moins, doit reposer sur des faits. Explication provisoire certes, mais fondée sur des faits. Voilà déjà deux définitions de la causalité.

J'en avance une troisième. Je ne veux rien écarter, mais c'est ce que j'ai appelé la causalité de niveau zéro. C'est parfois ce que l'on appelle la causalité probabiliste. Vous avez sans doute remarqué que, dans mes propos depuis un quart d'heure, j'ai rarement employé le terme « probabilités ». À un certain moment, j'ai parlé de fonctions de distribution. C'est bel et bien de la causalité probabiliste. L'idée est que la fonction de distribution de la réponse pour C1 est inférieure ou égale à la fonction de distribution pour C0. La fonction est décalée par rapport à la distribution vraie avec une stricte inégalité pour un certain jeu de probabilités positives. Les fonctions de distribution sont donc séparées. Entre les deux intervient un ordre stochastique. C'est seulement dire que C et R sont en association. Comme cela ne suffit absolument pas à une définition acceptable de la causalité : on doit aller plus loin. Une certaine information extérieure doit nous indiquer que C précède R, et non le contraire. Il faut donc établir d'une manière quelconque le sens de cette causalité, par ordre temporel ou par hypothèse notamment.

Ma collègue Nanny Wermuth a une très intéressante étude en sociologie médicale qui porte essentiellement sur le rapport entre la connaissance et la maîtrise qu'a le diabétique de sa maladie et dégage une relation positive entre cette connaissance et cette maîtrise, les gens qui connaissent mieux le diabète réussissant mieux à maîtriser leur maladie. On a voulu avancer comme explication que cette connaissance apporte en un sens une meilleure maîtrise. C'est là une hypothèse de travail. Il s'agissait de données transversales, et on peut sûrement concevoir une relation inverse où les gens qui maîtrisent mieux leur maladie seraient enclins à plus en apprendre à son sujet. Ce sens pourrait être le bon. L'autre sens de la causalité n'était établi que par une hypothèse de travail.

De toute manière, il nous faut un ordre et, aspect des plus importants, nous devons nous assurer qu'il n'existe aucune autre explication possible, qu'un conditionnement par une autre variable admissible, quelle qu'elle puisse être, ne détruira pas cette relation.

La question de la nature de cette variable admissible dans notre contexte vise la nature des variables que l'on devrait introduire dans l'équation de régression par laquelle on étudie la relation entre R et C. Mon ton devient un peu plus statistique pour le moment.

La nature des variables admissibles a étroitement à voir avec la notion de « toutes choses étant égales » dans notre causalité de niveau un. C'est une définition qui, je pense, remonte à I.J. Good et qu'a énoncée plus systématiquement le philosophe américain des sciences, Patrick Suppes. En économétrie, la notion

s'applique aux séries chronologiques. Il existe une généralisation plutôt évidente de cette notion lorsque R et C sont tous deux des séries chronologiques. C'est ce qu'on appelle en économétrie la « causalité de Granger ». Nous savons que le mathématicien emploie une notion tout à fait analogue dans un contexte général.

Voilà donc nos trois notions de causalité, nos trois définitions si vous voulez. Le monde serait différent si C passait de C0 à C1 pour la réponse R, toutes choses étant égales. Nous comprenons en un certain sens pourquoi les choses sont, ce qui est la causalité de niveau deux. La causalité de niveau zéro a un caractère plus directement probabiliste. Nous pouvons alors établir une relation statistique entre C et R. Nous savons que la dépendance va de C à R, et non le contraire, et qu'il est impossible de tout ficeler avec une explication quelconque. Ce sont les définitions de la causalité, plus particulièrement de niveau zéro et de niveau un. Ce sont des définitions. L'essentiel est de savoir comment nous pourrions vérifier ces définitions dans des cas particuliers, comment nous pourrions acquérir la certitude que la causalité entre véritablement en jeu dans une situation. La chose est bien plus difficile, dois-je préciser. La dernière définition n'est pas directement applicable, c'est sûr, en dehors du cadre des expériences aléatoires, parce que nous devons nous assurer qu'il n'y a pas d'autres explications possibles. Il se peut toujours que, dans une étude par observation, il y ait une autre explication et que nous n'ayons tout simplement pas mesuré les variables qui permettraient de l'évaluer. On peut donc dire que beaucoup de ces choses restent un but à atteindre, mais qu'elles ne sont pas directement réalisables.

Je ne sais pas pour combien de temps je dois m'étendre sur la question, puisque j'ignore dans quelle mesure cela peut aider. Citons un exemple pris dans les journaux il y a quelques semaines. C'est une étude – j'avoue ne pas en avoir lu tous les détails, et ce que j'en dirai sera plutôt hypothétique – où on disait que l'exercice réduisait les risques de certains cancers. Comment cela pourrait-il s'interpréter causalement? Selon la première définition, on dirait : voilà des gens qui ont fait beaucoup d'exercice; s'ils n'en avaient pas fait mais sans changer autre chose, plus d'entre eux auraient attrapé ces cancers, et le contraire est vrai pour ceux qui n'ont pas fait beaucoup d'exercice et qui auraient pu en faire.

C'est une définition plutôt précise. La question est de savoir comment on peut la vérifier. Si on étudiait des animaux, des souris par exemple, on pourrait songer à une expérience aléatoire où on répartirait en deux groupes un nombre approprié de souris, l'un qui bougerait beaucoup et l'autre qui serait restreint dans ses mouvements. On pourrait peut-être aussi utiliser des souris transgéniques et peut-être les irradier pour élever les taux de cancer. On espérerait dégager une certaine différence entre les deux groupes. Il ne peut y avoir aucune autre explication. En cas de différence importante, on aurait été berné par le hasard à sa très grande honte ou la réduction du cancer serait attribuable à l'exercice, parce qu'il n'y aurait aucune autre explication. Les deux groupes sont comparables sur les autres plans.

Dans une étude par observation où on tenterait d'appliquer la définition de la causalité de niveau zéro, on devrait pouvoir dire que les gens qui font de l'exercice et ceux qui n'en font pas ont la même structure par âge et sexe, la même classe socio-économique, etc. Quand les intéressés ont-ils fait de l'exercice? Se peut-il que les gens qui viennent de contracter un cancer n'aient pas ce qu'il faut pour faire de l'exercice et que la raison en soit justement ce début de cancer. Manifestement, l'application de cette définition est hérissée de difficultés d'interprétation.

Dans l'application de notre définition de la causalité de niveau deux (compréhension), il faudrait démontrer l'existence de processus biochimiques et physiologiques qui, stimulés par l'exercice, auraient bel et bien empêché la formation et le développement d'un cancer. C'est donc essayer de comprendre les processus qui sont sous-jacents au phénomène observé. Sans expérimentation, il est difficile d'établir les choses empiriquement pour la causalité de niveau zéro ou de niveau un.

Des complications s'ajoutent, bien sûr. En préface d'une de ses œuvres de dramaturge, George Bernard Shaw a écrit quelque chose sur l'exercice : « Si vous êtes en forme, vous n'en avez pas besoin et, si vous n'êtes pas en forme, il peut être dangereux. » En d'autres termes, il évoquait une interaction entre un état de santé initial et les effets de l'exercice. C'est ajouter une complication à la chose.
J'ai déjà souligné, je pense, que presque tout ce que j'ai pu dire jusque-là de la causalité ne nous révèle presque rien, à bien des égards, sur une question bien plus pressante : comment s'y prendre, tant dans le

contexte d'une enquête que sur un plan plus général, pour établir quelque chose comme la causalité? De très importants travaux récents viennent notamment de l'école philosophique de Carnegie Mellon, de Glymour et alia et, plus particulièrement, d'un ouvrage inspirant de Judith Pearl à UCLA. J'aimerais parler de cela et du rapport entre ce que propose Judith Pearl et ce que je vois comme une réflexion statistique type.

Pour ce faire, j'évoquerai une situation très idéale, une étude où il existe foncièrement quatre types de variables.

Il y a une variable qui pourrait être une cause C et une autre qui pourrait être une réponse R. Il existe aussi une variable de base B et une variable intermédiaire I. B et I en particulier pourraient être hautement multidimensionnelles, tout comme R d'ailleurs, peut-on supposer. Pour le moment, je vois la variable C comme unidimensionnelle. Dans tout ce travail, la première hypothèse posée est que ce système forme ce qu'on appelle un graphe acyclique orienté. Plus concrètement, nous connaissons le sens de toute paire de variables dans ce système. Des variables peuvent être associées ou non, mais si elles le sont, nous savons dans quel sens. La chose serait tout à fait concrète si on disait que B, C, I et R se présentent longitudinalement dans le temps. (C'est tout simplement l'idée que l'avenir est une réponse au présent si on peut s'exprimer ainsi, et non le contraire.)

Dans un graphe acyclique orienté, les arêtes sont orientées entre les sommets et les arêtes manquantes représentent une certaine indépendance conditionnelle. Je n'entre pas dans les détails, mais retenons que le sens de la dépendance est censé être connu.

Nous nous demandons si C influe sur R. Selon la première définition de C, nous voudrions pouvoir dire, si nous passions de C0 à C1 ou si C devait varier par pas – s'il s'agissait d'une mesure d'exposition à une échelle continue –, ce qu'il adviendra de R. Apparemment, nous n'avons qu'à procéder à une régression de R et à examiner le coefficient de C puisque les coefficients de régression, du moins dans une régression par les moindres carrés, nous indiquent la variation de R par variation unitaire de C. Quelle régression devrions-nous effectuer?

Voici ce que fait Pearl en particulier : d'abord, nous oublions I : I est marginalisé et quelque chose d'intermédiaire, en ce sens qu'il est une réponse à C et une explication à R. Donnons peut-être un simple exemple, celui de la tension artérielle d'un patient. Disons que C est un médicament pris au début de l'étude à diverses doses, I la tension artérielle après trois mois et R le décès par maladie du cœur dans les cinq ans, oui ou non. Nous procéderions à une régression logistique de la réponse en fonction des variables précédentes et de la tension artérielle après trois mois. Comme I aurait probablement subi l'influence du traitement, cette variable serait exclue de l'étude de l'incidence directe de C. Examinons donc la distribution conditionnelle de R compte tenu de C et B.

Dans un système linéaire où tout est en courbe normale à plusieurs variables ou qu'on peut raisonnablement caractériser par des équations de régression par les moindres carrés, il existe un lien entre le coefficient total de régression de R sur C et le coefficient partiel de régression de R sur C compte tenu de B. En d'autres termes, la relation totale entre C et R est à la fois le chemin direct entre ces deux variables et le chemin indirect qu'indique le diagramme. L'idée remonte aux généticiens qui ont écrit dans les années 1920, en fait vers la fin des années 1910. Une autre personne qui a produit une brillante idée dans son travail doctoral ou prédoctoral – comme nous l'avons entendu hier – aura cette relation.

Ce que Pearl fait, c'est marginaliser sur C. Elle considère la relation entre R et C compte tenu de B, puis intègre sur la distribution de B, mais notre intérêt n'est pas la distribution conditionnelle de B compte tenu de C, c'est la distribution marginale de B puisque, avec cette variation théorique de C, notre point de mire est l'effet de la manipulation de C. Lorsque nous manipulons C, nous n'agissons pas sur B, car ce qu'on fait aujourd'hui ne change pas le monde comme il était hier. Pearl joue beaucoup avec ce qu'elle appelle le calcul causal par opposition au calcul probabiliste, mais cela se ramène en réalité à une intégration sur une distribution de B qui est marginale plutôt que conditionnelle.

Les statisticiens ne feraient rien de cela de toute manière, peut-on penser. Ils feraient une régression. Ils oublieraient I pour le moment, j'espère; ils opéreraient la régression de R sur C et B. Ils rechercheraient des interactions pour les effets de B sur C (l'effet de C peut ne pas être le même chez les hommes que chez les femmes). Ils regarderaient donc les interactions. Si aucune ne paraissait présente, ils tenteraient de trouver une spécification de l'incidence de C en invariance par rapport à la variable de base B.

Quel est le rapport avec les autres définitions de la causalité? Dans une étude par observation et, bien sûr aussi, dans une expérience, il peut fort bien y avoir des variables de base inobservées. Ce sont des variables partiellement explicatives de C mais inobservées, soit qu'on n'ait pas pris la peine de les observer par souci d'économiser, soit que nous ignorions même en quoi elles consistent. Dans le cas de l'exercice, ce pourrait fort bien être une certaine variable, peut-être même latente, de l'état général de santé au début de la période considérée, par exemple.

Nous pouvons reprendre notre argument. Nous voulons opérer la régression de R sur C, B et U. C'est impossible, parce que nous ignorons ce qu'est U. Ce que nous devons faire, c'est effectuer la régression de R sur C et B et nous demander s'il importe vraiment que nous n'ayons pas observé la variable de base. Comme il apparaît plus généralement, la condition requise est que ce terme soit nul ou du moins très petit, ce qui peut se produire de deux manières. Avec une valeur nulle, R est indépendant de U compte tenu de C et B et, dans ce cas où U, la variable inobservée, influe sur C, elle influe sur B et elle influe sur R. En fait, il n'y a pas d'arête de C à R dans notre graphe. C'est tout simplement dire d'une manière assez évidente que, si la variable inobservée n'a pas sur la réponse un effet autre que celui dont rend déjà compte notre modèle, elle n'a pas d'importance, du moins dans une relation linéaire. (Il y a une erreur dans le transparent : on devrait dire que C est indépendant de U compte tenu de B ou que cette arête manque.) Il se peut donc que la variable de base influe sur la réponse, mais elle n'a aucune incidence sur la cause sauf par des choses que nous avons déjà observées.

Dans une expérience aléatoire, cela se vérifie par définition, puisque nous choisissons C, en un sens. Cela peut dépendre de caractéristiques observées, de strates et ainsi de suite. J'évoque ici une attribution aléatoire, et non pas une sélection aléatoire au sens de l'échantillonnage. Dans une expérience aléatoire, cette arête manque par définition. Dans un tel formalisme, on établit le rôle de l'expérience aléatoire en maintenant que des arêtes comme celle-là manquent par définition. Dans une étude par observation, nous devons essayer d'interpréter causalement le coefficient de régression de R sur C compte tenu de B, c'est-à-dire de déterminer ce que serait l'incidence réelle d'une variation de C. Dans une interprétation causale, nous devons nous assurer que soit cette arête manque soit certaines des arêtes sont petites. Bien sûr, la chose est difficile.

Quelques conséquences générales maintenant. J'ai exprimé tout cela par des représentations tirées de la théorie des graphes, mais on peut aussi se reporter à des formules, bien que les graphes aident peut-être. Une observation d'un caractère bien plus mathématique en passant est que l'on sait que, en principe, tous les résultats en théorie des graphes – au sens algébrique du terme – d'ensembles de sommets reliés par des arêtes orientées ou non avec des arêtes manquantes peuvent toujours être mis en corrélation avec les résultats d'un traitement algébrique matriciel. Il s'agit essentiellement de spécifier la présence d'arêtes par des matrices d'incidence. Le professeur Wermuth et moi avons récemment démontré assez en détail le rapport entre les opérations en théorie des graphes qui reposent sur ce qu'on appelle les théorèmes de séparation, d'une part, et celles de l'algèbre matricielle, d'autre part.

L'aléatorisation, j'en ai parlé. Comment aborder cette hypothèse dont nous avons besoin? Tout ce que je dis – et vous pourriez penser que j'insiste trop – est que l'interprétation causale de coefficients de régression au sens général demande des données d'observation et une hypothèse au sujet de l'effet des variables inobservées. Un mode d'évaluation est l'analyse de sensibilité. Le livre de Paul Rosenbaum sur les études par observation (la deuxième édition vient de paraître) en traite très bien. Je ne m'étendrai pas sur ce dernier point, car je risque de manquer de temps.

J'ai décrit sommairement comment Pearl et d'autres (et une grande partie des auteurs spécialisés) s'attaquent à ces problèmes par des graphes acycliques orientés. Pour un traitement plus réaliste de ces questions, ce que nous devons faire est, malheureusement, de préciser bien plus la nature des variables que

nous mesurons dans chaque étude. (Je passerai rapidement sur ce point.) En un sens, c'est certainement formaliser des idées relativement évidentes. Pour chaque sujet d'une étude, nous avons des variables intrinsèques qui définissent ce qu'est cette personne. Nous avons aussi des variables explicatives de base auxquelles nous ne prêterons peut-être pas un caractère causal, mais qui demeurent importantes. Il y a également les variables qui pourraient avoir ce caractère causal. Il y a des variables intermédiaires, puis des variables de réponse. Toutes ces variables seront généralement multidimensionnelles, bien sûr, dans une situation réaliste de complexité des choses. Ce qu'il faut, c'est réfléchir à la relation entre deux variables quelconques. À propos de chaque paire de variables, il faut dire que l'une est peut-être explicative de l'autre ou que les deux doivent être traitées en toute égalité. Vous pourriez avoir mesuré la tension diastolique et systolique de quelqu'un : ce n'est donc pas là une question spécialisée. Il y a peut-être des contextes où les choses sont différentes, mais elles ne le seront pas en général. Nous décrivons donc les variables comme étant égales dans une paire ou comme explicatives l'une de l'autre par ordre temporel ou par hypothèse de travail. Une généralisation est possible du graphe acyclique orienté où les variables seraient disposées par blocs de sorte que les arêtes lient deux blocs, c'est-à-dire les variables que comprennent ces blocs. L'orientation serait peut-être de gauche à droite, mais dans un bloc les arêtes demeureraient sans orientation. Il faut donc un développement de la théorie qui porte sur ce que l'on appelle les graphes groupés à chaîne. À mon avis, c'est presque essentiel si on songe à des progrès en tout réalisme dans ce domaine.

Pour revenir à une question plus pratique, dans une étude par observation (non aléatoire) où il y a une cause possible C qui semble en corrélation positive avec une réponse R, on peut raisonnablement supposer qu'il s'agit d'un rapport causal selon la définition de la causalité de niveau un. Si nous faisons varier C, toutes choses étant égales, R variera aussi.

Bradford Hill, statisticien médical, a proposé un certain nombre de règles qui rendraient la chose vraie. Ces règles avancées dans le contexte de l'épidémiologie de l'environnement sont d'une application assez large. L'intéressé a souligné qu'il s'agissait de règles générales, et non pas de critères ni sûrement de conditions nécessaires et suffisantes. On peut faire le lien, dans une certaine mesure, avec mon propos semi-quantitatif d'il y a un moment. Si l'effet que nous constatons est important – il faut qu'il le soit –, nous avons à chercher une dépendance possible à l'égard d'autres variables de base. L'explication du moment ne suffit pas. Est-il possible qu'une variable inobservée explique effectivement ce que nous observons? Si l'effet est important, il est moins probable qu'une variable inobservée que nous ignorons soit l'explication réelle. En mathématiques, si le terme principal que nous étudions est important, il est moins probable que ce terme en soit l'explication réelle.

Si l'effet est monotone dans une mesure d'exposition – si nous avons au lieu d'une cause peut-être binaire un niveau quantitatif d'exposition à quelque danger de l'environnement –, il est plus probable qu'il ait un caractère causal. Si l'effet se reproduit dans diverses études pour diverses personnes, c'est-à-dire si on constate l'absence d'une interaction qualitative avec les variables de base, le rapport sera plus probablement causal.

S'il y a une explication spécialisée (pour modifier légèrement la version de Bradford Hill, je dirais plus volontiers : là où une explication spécialisée est obtenue de préférence a priori; j'ai déjà remarqué que presque tout peut recevoir une telle explication avec un peu d'imagination) et qu'elle découle de principes ayant une valeur absolue comme ceux de la mécanique quantique ou de la physique mathématique classique, tout est bien. Toutefois, si c'est une explication tirée d'une réflexion rétrospective, elle peut être vraie, et il peut être utile d'en trouver une, mais la valeur probante sera moindre.

On peut penser au contexte de ce que Bradford Hill appelle une expérience naturelle. Souvent, il s'agit d'une catastrophe ou d'un événement extrême dont les conséquences seront difficilement autres que l'effet direct de ce qui vient de se produire. Un exemple type en est une étude de santé sur les survivants d'Hiroshima. La relation est étroite entre leur état de santé et la dose de radiation qu'ils ont reçue : il est inconcevable que la bombe atomique n'en soit pas la cause.

Cette dernière question est bien plus controversée et certains épidémiologistes s'inscrivent en faux. Bradford Hill a dit que, si la relation est spécifique, elle sera plus probablement causale. Dans le contexte

des maladies, il voulait dire que, si une exposition quelconque provoquait le cancer à un certain siège et non pas à d'autres – il songeait, bien sûr, au tabagisme et au cancer du poumon –, le rapport était sans doute causal. Un siège était touché et l'action était des plus spécifiques. Lorsque les mesures de réponse sont très variables, un rapport de causalité est moins probable. La question demeure contestée et bien des épidémiologistes seraient d'un avis différent.

Un ou deux choses en conclusion. J'ai introduit les variables intermédiaires dans le tableau – elles sont intermédiaires entre la cause possible et la réponse – pour ensuite les écarter d'un revers de main. Si j'en ai fait mention, c'est pour établir en partie mon argument. Il ne s'agit pas d'opérer la régression de la réponse sur tout ce qu'on a mesuré. Les choses qui interviennent après la cause possible ne devraient pas être prises en compte. Pourquoi alors les évoquer du tout?

Il y a à cela un certain nombre de raisons. Les variables intermédiaires peuvent nous indiquer ce qui se produit réellement dans le processus. Elles peuvent aussi nous montrer des choses indépendantes du processus que nous observons et qui brouillent le tableau. Elles peuvent aussi servir de réponse substitutive. Il y a toute une problématique au sujet de ce qui peut constituer une réponse substitutive légitime. (Ce sont les situations où il est difficile, coûteux ou long de mesurer la réponse qui nous intéresse réellement.) Les variables intermédiaires peuvent enfin nous aider à tracer les voies qui vont de la cause C à la réponse R, ce qui nous rapproche de la causalité de niveau deux.

Pour illustrer mon second point, laissez-moi vous raconter une histoire qui n'est pas tout à fait hypothétique au sujet d'une expérience aléatoire. Dans un essai classique dans les champs où des traitements sont attribués aléatoirement à des parcelles, un engrais a donné une magnifique plantation luxuriante dont les oiseaux ont pu se repaître. Rappelons-nous que, dans de tels essais aléatoires, on prend toutes les précautions possibles pour éviter les oiseaux. Ceux-ci s'amènent de kilomètres à la ronde et dévorent la récolte. Lorsqu'on mesure la réponse finale, on constate que les parcelles ont été d'un très faible rendement. Notre engrais magique a-t-il fait baisser les rendements? En un sens, il l'a fait, puisque les parcelles qui ont eu droit à ce traitement hors du commun ont présenté un rendement inférieur aux parcelles qui auraient reçu tout autre traitement. Notre causalité de niveau un est ainsi vérifiée.

Pour la causalité de niveau deux, il y a toutes sortes d'essais aléatoires impossibles à expliquer. Il n'y a pas de variables de base qui expliquent le phénomène. Pourtant, ce serait une conclusion très trompeuse à tirer comme scientifique – car on se trouverait à se méprendre tout à fait sur ce qui s'est réellement passé – ou comme technologue. Si vous étiez agriculteur, vous vous demanderiez quel engrais vous allez employer dans une culture (je parle ici des jours où on considérait que cultiver davantage pour l'alimentation était une bonne chose). Quel engrais devrais-je mettre dans ma parcelle? Faut-il éviter l'engrais magique? Ce serait une fausse conclusion à tirer parce que, si vous répandez votre engrais magique sur une grande superficie, les oiseaux se seront empiffrés et votre haut rendement en qualité et en quantité ne sera plus le même.

On a l'habitude de dire que tout ce qui se passe après le moment de l'aléatorisation est une cause possible. Il faut bien dire que ce n'est pas là une vérité universelle. D'une certaine manière, cela illustre un des enseignements que David dit avoir tirés des exemples terribles que donne mon ouvrage avec David Hinkley sur les choses qui peuvent mal tourner. C'est l'illustration du principe selon lequel on ne doit pas oublier les variables qui interviennent entre la cause et la réponse. Ce n'est toutefois qu'un exemple. Si vous appliquez le principe à la lettre, vous vous retrouvez avec des réponses folles. Les variables intermédiaires ont justement cette utilité.

Comment résumer tout cela? On peut dire en un sens que la quête des causes est omniprésente. Si vous voulez comprendre ce que signifient vos résultats, vous vous mettez de ce fait à la recherche des causes. On peut employer ce terme de diverses façons. Je ne saurais trop insister sur la circonspection qui doit être la nôtre au moment d'utiliser ce mot. Dans certains domaines de la statistique appliquée, vous trouverez des modèles que l'on dit causaux simplement parce qu'on voudrait qu'ils le soient. Il y a un devoir implicite de prudence. Il faut privilégier les analyses qui visent au moins une certaine causalité, de préférence dans les trois sens que j'ai mentionnés. Il existe en un sens ce que j'appellerais non une divergence mais une tension entre ce que je viens de dire et ce que je qualifierais de devoir empirique de la statistique. C'est une façon

un peu pompeuse de dire que l'on doit coller aux données. Que nous disent les données? Il y a là une tension qui fait que nous devrions coller à ce que nous disent les données et laisser celles-ci nous suggérer des interprétations qui ont un certain poids spécialisé.

Pour résumer toute mon allocution, disons qu'on doit être optimiste avec prudence ou prudent avec optimisme, je n'en suis pas tout à fait sûr.

Merci beaucoup.

ACÉTATES DE LA PRÉSENTATION DU PROFESSEUR COX

Causalité (acétate # 1)

Attitudes des statisticiens à l'égard de la causalité

- Laissons les philosophes s'en occuper, c.-à-d. ignorons la question.
- Corrélation n'est pas causalité.
- On peut inférer la causalité à partir d'expériences randomisées, mais non à partir d'études par observation.
- Mais, intérêt important dans les milieux de la statistique sociale, de l'économétrie et de l'épidémiologie, et participation plus générale des statisticiens récemment.

Définition de niveau un (acétate # 2)

Cause possible C, considérée comme étant binaire par souci de simplicité, C_0, C_1.
Réponse R.

Pour chaque individu, on ne peut observer qu'une seule C et une seule R associée.

La réponse d'un individu sous C_1 diffère systématiquement de la réponse que l'on aurait observée pour cet individu sous C_0 *toutes choses étant égales par ailleurs*.

- Réponses possibles, R_0, R_1
- L'une d'elles est contrefactuelle.
- Divers moyens de formaliser une différence systématique, par exemple :
  - $R_1 > R_0$ pour tous les individus
  - $R_1 - R_0 = \Delta$
  - R_1 stochastiquement plus grand que R_0
  - $E(R_1) > E(R_0)$
  - Les deux dernières expressions concernent une population étudiée explicite ou implicite.
- Impose une contrainte sur les valeurs admissibles de C.
- La méthode présente un angle technologique (pas inévitable!).
- Holland (1986); Neyman et Fisher; Cochran (1965); Rubin (1974); Robins (1997), Dawid (2000).

Définition de niveau deux (acétate # 3)

Compréhension du processus de génération fondée sur des preuves

- Comprend des études à divers niveaux.
- Inévitablement provisoire.
- Maxime de Fisher : *Make your theories elaborate* (Faites entrer vos théories dans les détails).
- Bertrand Russell, Goldthrope (2000).

Définition de niveau zéro (acétate # 4)

Définition directement probabiliste

$$P\,(R \leq r \mid C_1) \leq P\,(R \leq r \mid C_0) \qquad (1)$$

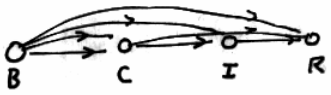
Plus la condition que

- C précède R temporellement ou autrement.
- Il n'existe aucun autre conditionnement par une *variable permise* qui détruira (1).
- I.J. Good, Suppes, Wiener, Granger

Un exemple (acétate # 5)

Exécution de l'exercice

Cas spécial simple (acétate # 6)



Pour l'évaluation causale, marginaliser sur *I* et conditionner sur *B*, puis examiner comment *R* dépend de *C* dans $f_{R|\,C,\,B}$.

Système linéaire

$$\beta_{RC} = \beta_{RC.B} + \beta_{RB.C}\beta_{BC}.$$

S'il existe une composante inobservée représentée par *U*, la marginalisation sur cette composante a lieu et n'a aucun effet sur la pente de la relation entre *R* et *C* ssi

$$R \perp\!\!\!\perp U \mid C, B \text{ ou } C \perp\!\!\!\perp U, B.$$

Quelques conséquences (acétate # 7)

Interprétation graphique
Matrices et graphes
Randomisation
Hypothèse et analyse de sensibilité
Préservation de la monotonicité de la relation

Une formulation générale (acétate # 8)

Pour chaque individu étudié, classer les variables observées comme étant

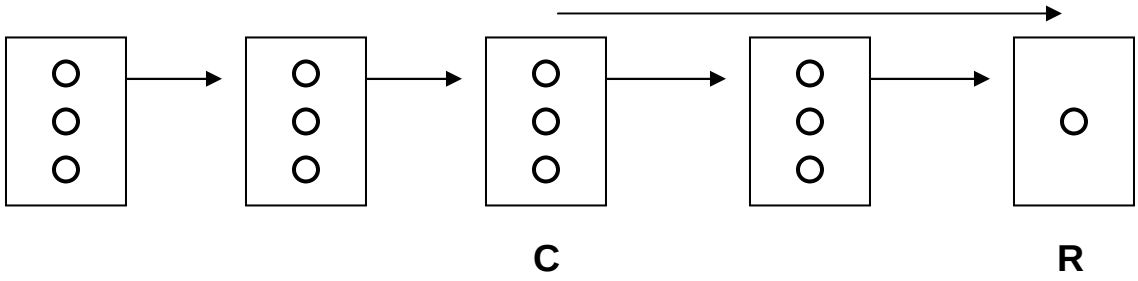
- des variables intrinsèques qui définissent l'individu étudié;
- des variables explicatives contextuelles qui ne sont pas traitées comme potentiellement causales;
- des variables causales possibles;
- des variables intermédiaires;
- des variables de réponse.

Les distinctions dépendent fortement du contexte.

Supposer que toute paire de variables est telle que

- l'une est explicative de l'autre
  - d'après l'ordonnancement temporel
  - hypothèse de travail
- les variables sont sur pied d'égalité.

Représentation graphique (acétate # 9)



Arêtes orientées entre les blocs.
Arêtes non orientées à l'intérieur des blocs.
L'absence d'arête signifie l'indépendance conditionnelle.

Lignes directrices de Bradford Hill (acétate # 10)

Bradford Hill (1965) a proposé des lignes directrices (et non des critères) dans le contexte de l'épidémiologie environnementale quant à la probabilité qu'une association décelée lors d'une étude par observation soit causale.

- Quand l'effet est important.
- Quand l'effet est monotone par « dose ».
- Quand l'effet est reproduit dans le cadre de diverses études et pour divers types d'individus.
- Quand il existe une explication particulière au domaine.
- Dans le contexte d'une « expérience naturelle ».
- Quand l'effet de *C* est spécifique.

Résumé (acétate # 11)

- La recherche de la causalité est ubiquiste.
- Diverses utilisations du mot *causal*.
- Analyses qui pourraient être causales.
- En rapport avec l'éthos empirique de la statistique.

Morale (acétate # 12)

Prudence optimiste

Ou, mieux encore, peut-être

Optimisme prudent

Bibliographie (acétate # 13)

Box, G.E.P. (1966). Use and abuse of regression. *Technometrics* 8, 625-629.

Bradford Hill, A. (1965). The environment and disease: association or causation. *Proc. R. Soc. Medicine* 58, 295-300.

Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford University Press.

Cochran, W.G. (1938). The omission or addition of an independent variable in multiple linear regression. *Suppl. J.R. Statist. Soc*. 5, 171-176.

Cochran, W.G. (1965). The planning of observational studies in human population (with discussion). *J.R. Statist. Soc.* A 128, 234-265.

Cox, D.R. (1958). *Planning of experiments.* New York: Wiley.

Cox, D.R. (1968). Notes on some aspects of regression. *J.R. Statist. Soc*. A 131, 147-174.

Cox, D.R. (1984). Present position and future developments, some personal views: Design of experiments and regression. *J.R. Statist. Soc*. A 147, 306-315.

Cox, D.R. (1992). Causality: some statistical aspects. *J.R. Statist. Soc*. A 155, 291-301.

Cox, D.R. (1999). Some remarks on failure-times, surrogate markers, degradation, wear, and the quality of life. *Lifetime data analysis* 5, 307-314.

Cox, D.R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statistical Science* 8, 204-283.

Cox, D.R. and Wermuth, N. (1996). *Multivariate dependencies*. London: Chapman and Hall.

Cox, D.R. and Wermuth, N. (2002). On the qualitative effect on dependencies of marginalization. Submitted.

Dawid, A.P. (2000). Causal inference without counterfactuals (with discussion). *J. Amer. Statist. Assoc*. 95, 407-448.

Dawid, A.P. (2002). Influence diagrams for causal modeling and inference. *Int. Statist. Rev*. 70, 161-189.

Doll, R. (2003). Can epidemiological methods establish causality? Fisher memorial lecture. October 2001, in preparation.

Fisher, R.A. (1926). The arrangement of field experiments. *J. Ministry of Agric*. 33, 503-513.

Fisher, R.A. (1935). *Design of experiments*. Edinburgh: Oliver and Boyd. And subsequent editions.

Goldthorpe, J. (2000). *Causation, statistics and sociology*. Chapter 7 in On sociology. Oxford University Press.

Good, I.J. (1961). A causal calculus: I. Brit. *J. Philosoph. Sci.* 11, 305-318.

Granger, C.W.J. (1969). Investigating causal models by econometric models and cross-spectral methods. *Econometrica* 37, 424-438.

Holland, P.W. (1986). Statistics and causal inference (with discussion). *J. Amer. Statist. Assoc.* 81, 945-970.

Hoover, K.D. (2002). *Causality in microeconomics*. Cambridge University Press.

Lauritzen, S.L. (1996). *Graphical models*. Oxford University Press.

Lauritzen, S.L. (2000). Causal inference from graphical models. In *Complex stochastic systems*. Pp. 63-107. eds O.E. Barndorff-Nielsen et al. London: Chapman and Hall.

Lindley, D.V. (2002). Seeing and doing: the concept of causation (with discussion). *Int. Statist. Rev*. 70, 191-214.

McKim, V.R. and Turner, S.P., eds (1997). *Causality in crisis?* University of Notre Dame Press.

Pearl, J. (2000). Causality. Cambridge University Press.

Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* 8, 431-440.

Robins, J. (1997). Causal inference in complex longitudinal data. In *Latent variable modeling with applications to causality*. Pp. 69-117, ed. M. Berkane. New York: Springer.

Rosenbaum, P.R. (2002). *Observational studies.* Second ed. New York: Springer.

Rothman, K. and Greenland, S., eds (1999). *Modern epidemiology*, 2nd ed. Philadelphia: Lippincott-Raven.

Rubin, D.B. (1974). Estimating causal effect of treatments in randomized and nonrandomized studies. *J. Educational Psychol*. 66, 688-701.

Scheines, R. (1997). An introduction to causal inference. Pp. 185-199. In *Causality in crisis?* Editors V.R. McKim and S.P. Turner. University of Notre Dame Press.

Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, prediction and search*. New York: Springer.

Stone, R. (1993). The assumptions on which causal inference rest. *J.R. Statist. Soc*. B 55, 455-466.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: N. Holland.