Causalité, trilogue

Antoine Chambaz, Isabelle Drouet, Jean-Christophe Thalabard

Causalité, trilogue

A. Chambaz[1], I. Drouet[2], J-C. Thalabard[3]
[1] Modal'X (EA 3454), Université Paris-Ouest Nanterre
[2] SND (FRE CNRS 3593), Université Paris-Sorbonne
[3] MAP5 (UMR CNRS 8145), Université Paris Descartes

3 avril 2013

Résumé

Une philosophe, un médecin et un statisticien discutent de la causalité. Ils débattent des rapports que la causalité entretient avec le hasard et la statistique, en articulant leurs arguments autour d'exemples principalement empruntés à la médecine. De cette confrontation originale découle un trilogue enrichissant et accessible à un large public. Cet hommage au célèbre entretien entre d'Alembert et Diderot, deux grands penseurs français du 18ème siècle, offre en particulier une introduction à la philosophie de la causalité, une initiation à la statistique incluant des développements récents, le tout motivé par des questionnements médicaux.

A philosopher, a medical doctor and a statistician talk about causality. They discuss the relationships between causality, chance and statistics, borrowing examples from medicine to develop their arguments. This original confrontation gives rise to an enriching and large audience trialogue. This tribute to the famous conversation between d'Alembert and Diderot, two great French thinkers of the 18th century, notably offers an introduction to philosophy of causality, an initiation to statistics including recent developments, the whole being motivated by medical issues.

Ils n'avaient pas fait cent pas lorsque, au détour d'un rocher, apparut dans toute son évidence la cause, seule possible, de ce bruit épouvantable et pour eux terrifiant, qui les avait tenus en émoi toute la nuit. C'étaient — j'ose espérer, ami lecteur, que tu ne m'en tiendras pas rigueur — six maillets d'un moulin à foulon qui, en frappant alternativement, faisaient ce vacarme effrayant.

M. de Cervantes, *L'ingénieux Hidalgo Don Quichotte de la Manche*

In relating what follows I must confess to a certain chronological vagueness. The events themselves I can see in sharp focus, and I want to think they happened that same evening, and there are good reasons to suppose they did. In a narrative sense they present a nice neat package, effect dutifully tripping along at the heels of cause. Perhaps it is the attraction of such simplicity that makes me suspicious, that along with the conviction that real life seldom works this way.

R. Russo, *The risk pool*

Table des matières

Notations

- $\exists$, quantificateur existentiel, notation mathématique signifiant "il existe".
- $\Rightarrow$, connecteur binaire, notation mathématique signifiant "seulement si" (ou, de façon équivalente, "si $(\dots)$ alors $(\dots)$").
- 0-1, notation mathématique signifiant"0 ou 1".
- $\{0,1\}$, l'ensemble constitué des nombres 0 et 1.
- $[a,b]$, l'intervalle de tous les nombres réels compris entre a et b.
- W, vecteur aléatoire constitué de covariables initiales ; A, A', variables aléatoires d'exposition ; L, L', variables aléatoires constituées de covariables intermédiaires ; Y, variable aléatoire quantifiant une issue d'intérêt, appelée critère de jugement ; Y_a et L_a, contreparties de Y et L sous contrôle de $A = a$, par exemple dans le monde contrefactuel où l'égalité $A = a$ est garantie.
- $O \sim P$ une observation, conçue comme une variable aléatoire, dont la loi est $P \in \mathcal{M}$, où $\mathcal{M}$ est un ensemble de lois candidates aussi appelé "modèle".
- $X \sim \mathbb{P}$ une donnée dite complète, conçue comme une variable aléatoire, dont la loi est $\mathbb{P} \in \mathbb{M}$, où $\mathbb{M}$ est un ensemble de lois candidates aussi appelé "modèle contrefactuel".
- $\vartheta : \mathcal{M} \to \Theta$, une fonctionnelle qui associe à tout élément $P \in \mathcal{M}$ le paramètre statistique $\vartheta(P)$. Note : ϑ est la lettre grecque thêta dans sa version cursive et Θ est cette même lettre dans sa version majuscule.
- $\theta : \mathbb{M} \to \Theta$, une fonctionnelle qui associe à tout élément $\mathbb{P} \in \mathbb{M}$ le paramètre statistique $\theta(\mathbb{P})$.
- $P(\mathbb{P})$, la loi de O lorsqu'elle est modélisée comme l'observation incomplète de la donnée complète $X \sim \mathbb{P}$.
- $\mathbb{P}(P)$, la loi de la donnée complète X dans un modèle contrefactuel construit synthétiquement à partir de l'observation de $O \sim P$.
- $P\{O\}$ est la valeur moyenne de $O \sim P$.
- $P\{Y|W\}$ est la valeur moyenne conditionnelle de Y sachant W pour $O = (W, Y) \sim P$. Si $Y \in \{0,1\}$, elle coïncide avec la probabilité conditionnelle $P(Y = 1|W)$ que Y vaille 1 sachant la valeur prise par W.
- P_n^0 et P_n^k, estimations initiale et itérée k fois de la loi P de $O \sim P$ sur la base de n observations.
- $\{P(\varepsilon) : \varepsilon \in [-1,1]\} \subset \mathcal{M}$, modèle paramétrique, de dimension un et par conséquent aussi appelé "chemin", sous-ensemble de lois candidates, toutes éléments du modèle englobant $\mathcal{M}$.
- $\nabla\vartheta(P)$, "dérivée" en P d'une fonctionnelle $\vartheta : \mathcal{M} \to \Theta$ différentiable sur les chemins.

1 Lucrèce quantique

AC: « Isabelle, saurais-tu nous dire s'il est universel de penser en termes causaux ? »

ID: « Il y a sans aucun doute une dimension culturelle à la causalité, qui s'impose de façon prégnante dans les cultures occidentales mais point dans d'autres cultures. En cela, je répondrais par la négative si le temps nous était compté : penser en termes causaux n'est pas universel. »

JCT: « Mais nous avons tout notre temps ! Et la question d'Antoine circonscrite à l'occident n'en est pas moins je crois pertinente, délicate et ancienne. N'est-ce pas Virgile qui écrivait [31] à l'orée du premier millénaire de notre histoire occidentale

> Felix, qui potuit rerum cognoscere causas,

heureux qui a pu pénétrer la raison des choses ? »

ID: « Oui, en effet Jean-Christophe, et il faisait probablement en cela référence à l'œuvre de Lucrèce *De rerum natura* [22], de la nature des choses, sa description poétique du monde selon les principes d'Epicure. Et nous pouvons remonter plus avant le cours du temps : c'est Platon qui fait dire à Timée [23]

> Or, tout ce qui naît, procède nécessairement d'une cause ; car rien de ce qui est né ne peut être né sans cause.

Les propos de Timée d'abord puis de Virgile se rapportent à la création de l'univers, à une échelle gigantesque de temps et d'espace mais partant des corps premiers que sont les atomes. »

AC: « Est-il question de hasard dans cette description poétique du monde ? »

ID: « Je renvoie ta question à une autre : qu'entends-tu par le "hasard" ?! »

AC: « Je sais pour ma part que le mot "hasard" vient de l'arabe *al-zahr*, dé, que S. Mallarmé nous dit qu'un coup de dés jamais ne l'abolira, et que pour Héraclite, un tas de gravats déversé au hasard est le plus bel ordre du monde !... »

JCT: « Je sais quant à moi que le mot "chance", qui signifie hasard en anglais, nous vient du latin "cadentia", ces choses qui tombent, que Cicéron l'employait pour faire référence aux osselets, et que c'est aussi au latin qu'on doit le mot "aléatoire", tiré de ses "alea", jeu de dés, et "aleatorius", qui concerne les jeux de hasard. »

ID: « Un gouffre sépare ces dés et ces osselets du tas de gravats d'Héraclite. Comment glissons-nous de l'un à l'autre, de la simplicité du lancer d'un dé ou d'un osselet à la complexité du réel ? »

JCT: « Je peux t'expliquer comment engendrer des nombres réels à partir de suites de 0-1 "tirés au hasard". »

ID: « Je t'écoute. »

JCT: « Il y a par exemple la quinconce de F. Galton, dont on peut admirer un exemplaire à la Galerie de la Découverte, à Paris. Il l'a imaginée pour visualiser l'évolution d'une diffusion au hasard de billes de rayon r. La quinconce consiste en une planche verticale sur laquelle sont plantés des clous disposés en quinconce, régulièrement espacés d'une distance $r + \varepsilon$, pour $\varepsilon > 0$ petit, les rangées successives étant décalées de $r/2$. Introduites au milieu de la partie supérieure de la planche, les billes descendent en empruntant un chemin entre les clous, pour se répartir sur la partie inférieure de la planche selon des tas de tailles variables. Le décalage de $r/2$ garantit que, à chaque étape, les billes ont autant de chance de basculer sur la gauche que sur la droite. Par ailleurs, la largeur $r + \varepsilon$ et l'écartement entre les rangées de clous garantissent que les basculements passés n'influencent pour ainsi dire pas les basculements futurs. »

AC: « Suivons la trajectoire d'une bille. Si nous numérotons chacun des créneaux inférieurs et notons 0 ou 1 selon que la bille tombe à gauche ou à droite de chaque clou qu'elle rencontre au cours de sa chute, alors le numéro du créneau dans lequel aboutit la bille est bien obtenu à partir de cette suite de 0-1. »

ID: « Le numéro est aléatoire parce que la suite de 0-1 l'est ! Voilà qui est clair… Comment se répartissent les billes lorsque nous en lançons un grand nombre successivement ? Un motif particulier se distingue-t-il ? »

JCT: « Lorsque le nombre de billes est important, nous constatons empiriquement que les créneaux les plus proches des extrémités de la planche présentent des tas de petites tailles tandis que les créneaux centraux reçoivent les tas les plus importants. L'ensemble des hauteurs des tas dessine une courbe régulière en cloche. »

ID: « Prenons un peu de recul, voulez-vous ? Tu viens de m'expliquer comment tirer au hasard dans un ensemble fini d'éléments en tirant au hasard des 0-1. Je conçois que si la planche est gigantesque cela te permet de tirer au hasard des nombres décimaux avec une très grande précision. Y a-t-il d'autres façons de procéder ? »

JCT: « Ta question est avisée ! En passant à la limite, c'est-à-dire en prenant une planche infiniment grande, la loi limite que nous obtenons ainsi est appelée loi gaussienne. C'est l'une parmi une infinité de façons de tirer des nombres au hasard. »

AC: « Le procédé, classique, de Bolzano-Weierstrass permettrait de tirer des nombres selon une autre loi. Imagine que, à partir de la même suite de 0-1 que précédemment et en partant de l'intervalle $[0, 1]$ de tous les nombres compris entre 0 et 1, je découpe successivement l'intervalle courant en son milieu et je choisis sa moitié gauche pour un 0 et sa moitié droite

pour un 1. Avec 1024 0-1, nous déterminons ainsi un nombre aléatoire dans $[0, 1]$ avec une précision de 308 chiffres après la virgule. »

JCT: « Avec une planche infiniment grande, la loi limite ainsi obtenue est appelée loi uniforme sur $[0, 1]$. Cela signifie que la chance de tomber dans un intervalle $[a, a + \ell]$ de longueur ℓ ne dépend que de ℓ, pas de a. »

ID: « Fort bien. Vous m'avez expliqué comment tirer au hasard un nombre selon une loi gaussienne ou une loi uniforme. Vos constructions ne sont, en somme, que des échafaudages de 0-1 tirés au hasard... »

AC: « ... et pour enfoncer le clou, ces tirages successifs sont indépendants, c'est-à-dire tels que les valeurs passées n'influencent pas les valeurs futures, et équiprobables, c'est-à-dire tels que 0 et 1 ont autant de chance d'être tirés l'un que l'autre ! Nous disons qu'une variable aléatoire prenant équiprobablement les valeurs 0 ou 1 obéit à la loi de Bernoulli de paramètre $1/2$. »

ID: « Permettez-moi de renouveller ma question : résumez-vous le hasard à cela ? »

JCT: « Eh bien oui, et cela, peut-être, contre-intuitivement. Car la plupart des variables aléatoires se construisent à partir de variables aléatoires de loi Bernoulli de paramètre $1/2$ et de loi uniforme sur $[0, 1]$ toutes indépendantes. »

ID: « Entendu. Mais je doute que la génération la plus rigoureuse de nombres au hasard mobilise des planches criblées de clous ! »

AC: « En effet, la génération la plus rigoureuse de nombres au hasard repose, à ce jour, sur l'émission de photons sur une lame semi-réfléchissante. »

JCT: « Les lois de la mécanique quantique nous enseignent effectivement que deux choix équiprobables s'offrent aux photons : traverser ou rebondir. Ces mêmes lois nous enseignent que les choix de photons successifs sont mutuellement indépendants. »

ID: « A la réflexion, et bien qu'elle m'ait semblé familière au premier abord, je crois que cette notion de deux choix équiprobables mériterait une explication. Pourriez-vous m'en proposer une ? »

AC: « Avec plaisir ! Cette explication, de nature probabiliste, est plus technique car elle repose sur la notion de limite. Dire que traverser ou rebondir sont deux événements équiprobables pour le photon est équivalent à dire que je suis presque sûr qu'aussi petite que soit la précision $\varepsilon > 0$ eh bien il existe un entier n_0 qui dépend de ε tel que, si j'émets indépendamment $n \geq n_0$ photons sur une lame semi-réflechissante alors la fraction n_t/n de photons qui la traversent s'éloigne de $1/2$ d'au plus ε. C'est un exemple de la loi des grands nombres. »

ID: « Tu n'es que *presque* sûr ? ! »

AC: « C'est bien là l'expression consacrée ! Elle signifie que si cette expérience (choisir arbitrairement ε, envoyer $n \geq n_0$ photons sur la lame semi-réfléchissante, évaluer la fraction de ceux-ci qui la traversent, évaluer l'écart de la fraction à $1/2$) est répétée indépendamment N fois alors elle aboutira N fois à la même conclusion (l'écart est d'au plus ε), aussi grand fût-ce N. »

JCT: « Je ne suis pas sûr que tu aies convaincu Isabelle ! Pourquoi ne pas dire plutôt qu'un événement est presque sûr quand sa probabilité égale 1 ? *A contrario*, un événement de probabilité 0 n'a aucune chance d'être observé mais n'est pas pour autant impossible. »

ID: « C'est clair maintenant. »

AC: « Les technologies les plus récentes permettent d'engendrer ainsi environ seize millions de 0-1 équiprobables indépendants par seconde. Peuvent en découler 15625 nombres aléatoires tirés indépendamment uniformément sur $[0, 1]$ par seconde, pour une précision de 308 chiffres après la virgule. »

ID: « Alors la boucle est bouclée, et nous pouvons revenir à ta question qui l'avait ouverte. »

AC: « Ma question ?... Ah oui ! Le hasard, entre-t-il dans la description poétique du monde que fait Lucrèce ? Et puisqu'il semble y être question de causalité, Lucrèce et Platon s'aventuraient-ils sur le territoire de la causalité armés du concept de hasard ou bien s'en passaient-ils ? De par ma formation, et aussi je crois par inclination, il m'est difficile de penser à la causalité sans m'en remettre en partie au hasard. »

JCT: « Tu es en cela comme D. Diderot, qui argumentait auprès de J. d'Alembert [10] :

> (...) la cause subit trop de vicissitudes particulières qui nous échappent, pour que nous puissions compter infailliblement sur l'effet qui s'ensuivra. La certitude que nous avons qu'un homme violent s'irritera d'une injure n'est pas la même que celle qu'un corps qui en frappe un plus petit le mettra en mouvement.

Quant à Lucrèce, lui et son maître à penser, Epicure, plaçaient au cœur de la physique dite épicurienne la notion de *clinamen*, cette déviation spontanée des atomes (pas des photons !) relativement à leur chute verticale dans le vide, variation aléatoire permettant d'expliquer de fil en aiguille l'existence des corps et la liberté humaine. »

ID: « Vous concluez ma foi un peu précipitamment à la nécessité de lier causalité et hasard ! Dans son *Traité de la nature humaine* [17], D. Hume élabore un concept de cause dont les répercussions ont été magistrales, et qui en mettant la *régularité* au cœur de la causalité semble en exclure le hasard. »

2 Hume enflammé

AC: « Qu'est-ce que la régularité ? »

ID: « La conjonction constante... »

JCT: « Mais encore ?! »

ID: « Laissez-moi vous lire le passage clef [17] :

> C'est donc seulement par EXPERIENCE que nous pouvons inférer l'existence d'un objet de celle d'un autre. La nature de l'expérience est celle-ci : nous nous souvenons d'avoir eu de fréquents exemples de l'existence d'objets d'une espèce, et nous nous souvenons aussi que les individus d'une autre espèce d'objets les ont toujours accompagnés, et ont existé dans un ordre régulier de contiguïté et de succession par rapport à eux. Ainsi, nous nous souvenons d'avoir vu cette espèce d'objet que nous nommons *flamme* et d'avoir senti cette espèce de sensation que nous nommons *chaleur*. Nous rappelons également à l'esprit leur constante conjonction dans tous les cas passés. Sans plus de cérémonie, nous nommons l'une *cause* et l'autre *effet*, et inférons l'existence de l'un de l'existence de l'autre. Dans tous les cas à partir desquels nous sommes instruits de la conjonction de causes particulières et d'effets particuliers, à la fois les causes et les effets ont été perçus par les sens et nous nous en souvenons mais dans tous les cas où nous raisonnons sur eux, c'est seulement l'un que nous percevons ou dont nous nous souvenons, et l'autre se donne en conformité avec notre expérience passée.

Inspirant n'est-ce pas ?! »

JCT: « Cette définition ne laisse en effet de place ni au hasard ni à la statistique. »

AC: « Pourquoi dis-tu cela ? La statistique n'est-elle pas l'art d'extraire de l'information à partir d'observations elles-mêmes issues d'*expériences* ? Il me semble au contraire que ce passage suggère que la notion de causalité est intrinsèquement liée à celle de statistique. »

ID: « Ton interprétation est hétérodoxe car, par "expérience", D. Hume fait plutôt référence à l'expérience sensible. En outre, en associant causalité et statistique, tu t'inscris en faux contre la vision de K. Pearson, l'un des fondateurs de la statistique, père du coefficient de corrélation et de la régression, grand pourfendeur avec B. Russell au début du 20è siècle de la notion même de causalité. Dans ces observations que tu évoques et dans les lois qui régissent leur production, K. Pearson voyait rien moins que la réalité elle-même réduite à son essence, la causalité n'ayant tout simplement pas droit de cité. Mais peut-être que ton interprétation est intéressante. Penses-tu pouvoir la pousser plus loin ? »

JCT: « Je suis prêt à l'y aider, et à creuser, patiemment, le sillon d'un lien entre causalité et

statistique. Dans la citation de D. Hume, isolons l'expression "contiguïté" ; je la qualifierais volontiers de spatiale et temporelle. J'en comprends une nécessité d'une part que l'action de la cause et la mesure de son effet s'appliquent à un même système cohérent, ou unité expérimentale dans le jargon statistique ; d'autre part que l'observation de la cause et de son effet potentiels aient lieu à une échelle de temps dont la caractérisation dépend de la nature de ceux-ci. »

AC: « J'adhère à cette interprétation. La condition de succession temporelle énoncée par D. Hume apparaît aussi comme une connaissance *a priori*... »

JCT: « ... ou une contrainte *a priori*... »

AC: « entendu, une connaissance *ou* une contrainte *a priori* imposée au modèle statistique... »

ID: « ... ou aux concepts de cause et d'effet !... »

AC: « entendu, au modèle statistique dédié à la mise en lumière d'une potentielle relation de cause à effet à partir d'observations, *ou* aux concepts mêmes de cause et d'effet, pour écarter la possibilité qu'un effet puisse précéder sa cause ou même lui être simultané. »

JCT: « Cela semble naturel à l'échelle humaine ; c'est peut-être discutable à l'échelle quantique, comme le suggère la résolution théorique et expérimentale du paradoxe EPR [12, 2, 1]. Isabelle, pourrais-tu enfin nous expliquer ce qu'est cette troisième condition de constante conjonction ? »

ID: « La régularité, ou constante conjonction, c'est cette notion qu'un événement que nous nommons "cause" est *toujours* suivi de l'élément que nous nommons "effet". Chaque fois que tu présentes ta main au-dessus de la flamme, elle te brûle... »

AC: « ... et si tu l'ôtes elle ne te brûle plus. »

ID: « Cela n'est pas chez D. Hume, mais je pense que tu as raison et que nous pouvons l'ajouter. »

AC: « C'est qu'une cause est toujours relative à une situation dans laquelle elle est absente. »

JCT: « Il n'en reste pas moins que l'idée principale de D. Hume est que les causes sont toujours suivies de leurs effets. Néanmoins il me semble clair que ce n'est pas toujours le cas. Frotter une allumette contre un grattoir cause son embrasement et pourtant, si nous pensons à une allumette mouillée ou à un milieu ne contenant pas d'oxygène, nous comprenons que frotter une allumette contre un grattoir n'est pas toujours suivi de l'embrasement de l'allumette. Comment te tires-tu de ce piège ? »

ID: « Ton piège est rudimentaire ! Je te réponds que la cause de l'embrasement de l'allumette ce n'est pas seulement son frottement contre un grattoir, mais ce frottement en tant qu'il appartient à un ensemble de conditions auquel appartiennent aussi la sécheresse de l'allumette, la présence d'oxygène dans l'air — et sans doute d'autres conditions. L'ensemble de ces conditions serait, lui, toujours suivi de l'embrasement de l'allumette. Nous pouvons dire qu'il est "suffisant" pour cet embrasement, et nous retrouvons bien la condition régulariste de D. Hume. »

AC: « En somme, frotter l'allumette contre un grattoir est une cause de l'embrasement de l'allumette dans la mesure où frotter l'allumette appartient à un ensemble de conditions qui, quand elles sont toutes réunies, sont toujours suivies de l'embrasement de l'allumette et que le frottement est indispensable pour cela. »

ID: « Absolument, c'est la conception de J.S. Mill, qui date du milieu du 19è siècle, puis de J.L. Mackie, dans la seconde moitié du 20è siècle : une cause est une condition INUS (c'est un acronyme de l'anglais "Insufficient but Nonredundant part of an Unnecessary but Sufficient"), c'est-à-dire une condition qui n'est pas suffisante pour l'effet, mais qui est une partie non-redondante d'une condition suffisante quoique non nécessaire pour l'effet. »

JCT: « Une conception de ce type est véhiculée par le modèle de K. Rothman [24, 25] qu'utilisent les épidémiologistes depuis la fin des années soixante-dix. Ce modèle est un guide pour établir des relations de cause à effet. Dès le 19è siècle, les médecins se sont munis de tels guides. Ainsi par exemple, J. Henle et son élève R. Koch ont élaboré quatre critères pour mettre en évidence une relation causale entre un microbe et une maladie [19, 20]. Ces critères ont permis à R. Koch d'établir l'étiologie de la tuberculose et de la maladie du charbon. Régulièrement adaptés, les "postulats de Koch" sont toujours utilisés en microbiologie [13]. »

AC: « Dans le même ordre d'idée, les critères de A.B. Hill [16] ont été conçus au milieu des années soixante dans le contexte de l'épidémiologie du travail. Bien qu'ils ne soient ni nécessaires ni suffisants, ils continuent de structurer utilement l'interprétation causale des études épidémiologiques. »

JCT: « C'est bien vrai, et nous devons à R. Doll, A.B. Hill et ses critères la mise en lumière du rapport causal entre tabac et cancer du poumon. J'aimerais cependant que nous revenions à la régularité. Pouvons-nous toujours échafauder ce que d'aucun appellerait un subterfuge pour soutenir une définition régulariste de la causalité, comme dans le cas de l'allumette ? Par exemple, pouvons-nous considérer que, s'il est vrai que tous les fumeurs ne meurent pas du cancer du poumon, il existe néanmoins un ensemble de conditions, au nombre desquelles figure le tabagisme, et qui soit suffisant pour le cancer du poumon ? En faire l'hypothèse est sans doute un bon principe méthodologique. Mais ce principe est-il crédible ? Est-il toujours possible, *in fine*, de réduire la causalité à la régularité ? »

3 Ceteris paribus sic standibus

ID: « Les analyses philosophiques de la causalité en termes de probabilités se sont justement développées, dans la seconde moitié du vingtième siècle, contre la possibilité d'une telle réduction. L'idée qui se trouve au fondement de ces analyses est la suivante : C cause E si et seulement si C augmente la probabilité de E, *ceteris paribus*. I.J. Good, P. Suppes, N. Cartwright, B. Skyrms... tous essaient de donner un sens précis à cette idée. »

AC: « Le hasard entre de nouveau dans la danse ! Mais que signifie ton "*ceteris paribus*" ? »

ID: « C'est une forme réduite de la locution latine *ceteris paribus sic stantibus*, qui se traduit par "toutes choses étant égales par ailleurs". Autrement dit C cause E si et seulement si la présence de C augmente la probabilité de E relativement à son absence toutes choses égales par ailleurs – *ceteris paribus*. »

AC: « Ainsi, dans ce paradigme, pour causer E, C ne doit pas nécessairement être une condition INUS ! Et si C en est une alors, en la présence concomitante des autres éléments de la condition suffisante à laquelle C appartient, la présence de C entraîne celle de E presque sûrement (c'est-à-dire avec une probabilité égale à 1) tandis qu'en son absence la présence de E n'est pas presque sûre, c'est-à-dire que E est susceptible d'être absente. »

JCT: « Il me semble qu'un glissement de difficulté s'est opéré : que faut-il vraiment entendre ici par "*ceteris paribus*" ? N'est-il pas difficile de spécifier ce que sont ces choses telles que, lorsqu'elles sont maintenues inchangées, la présence de C augmente la probabilité de E relativement à son absence, et en ce sens C cause E ? »

AC: « S'il faut les spécifier, je commencerais par dire que pour moi, il s'agit d'éléments, ou traits, de lois. »

ID: « De lois physiques ? ou plus généralement de ce que les philosophes appellent des "lois de la nature" ? »

AC: « Non, de lois au sens probabiliste. C'est-à-dire de règles caractérisant la production de variables aléatoires. Qu'en pensez-vous ? »

JCT: « Mon intuition est que si nous entendons "lois" au sens des philosophes, alors la caractérisation de ces "choses maintenues inchangées" est une entreprise de nature causale ! La tâche devient circulaire. »

ID: « Oui, et l'analyse de la causalité en termes probabilistes n'est pas réductive. Cela signifie, sur le plan méthodologique, qu'il faut déjà avoir beaucoup de connaissances causales pour identifier de (nouvelles) relations causales. »

AC: « J'aimerais que nous revenions à l'expression "*ceteris paribus*". »

JCT: « Par exemple : emploierions-nous l'expression "conditionnellement à" de façon équivalente à "*ceteris paribus*" ? »

AC: « Je dirais oui parfois, non en général ! »

ID: « Dans quel scénario répondre par l'affirmative ? »

AC: « Voici l'exemple le plus simple qui me vienne à l'esprit. Nous nous intéressons à l'effet éventuel sur une maladie, par exemple en termes de survie, de gravité ou de durée, d'un traitement, que je note $a = 1$, relativement à l'absence de tout traitement, que je note $a = 0$. »

JCT: « Pourquoi diable utilises-tu ici la lettre "a" ? »

AC: « Disons que je la choisis parce que c'est la première lettre du mot "action". La variable A témoigne de la présence de la cause, quand $A = 1$, ou de son absence, quand $A = 0$, dès lors que nous entendons par "cause" la prise du traitement relativement à l'absence de tout traitement. L'effet du traitement est résumé par la variable notée Y, qui est postérieure à A dans le temps. En introduisant un peu de formalisme, cela peut par exemple se présenter ainsi (*cf* la partie gauche du Tableau noir 1). »

ID: « J'en déduis par analogie que W précède chronologiquement A, et donc Y. A quoi correspond cette variable ? »

JCT: « La variable W représente bien un ensemble d'informations antérieures à la cause et à son effet. Ces informations participent à la détermination, nous dirions plutôt à la *réalisation*, de A et Y. »

ID: « Résumons : A représente la nature de la cause, Y son effet... donc, par élimination, j'imagine que W correspond à ces choses que désigne le "*ceteris paribus*". Ainsi, dirais-tu dans ce cas que "*ceteris paribus*" et "conditionnellement à la réalisation de W" sont équivalentes ? »

AC: « Presqu'équivalentes (*cf* la partie droite du Tableau noir 1), car l'effet éventuel de la cause sur la maladie s'exprime naturellement en termes de comparaison de Y_1 et Y_0, c'est-à-dire des issues de la maladie lorsque nous *imposons* le traitement d'une part ou l'absence de traitement d'autre part... »

JCT: « et parce que Y_1 et Y_0 sont fonctions, par construction, *du même* W ! »

ID: « Je vois ! A quoi tient ton "presque" ? »

AC: « A ce qu'en fait Y_1 et Y_0 sont fonctions, par construction, du même W *et de* U_Y, *selon* f_Y ! Ainsi pour moi, "*ceteris paribus*" fait exactement allusion au maintien inchangé des lois marginales de W et conditionnelle de Y sachant (A, W), c'est-à-dire de f_W, f_Y

et de la façon dont les sources d'aléa U_W et U_Y sont produites. Voici les traits de la loi auxquels je faisais allusion plus tôt. »

ID: « Tu m'as convaincue ! Et je discerne enfin la raison fondamentale pour laquelle tu ne veux pas assimiler "*ceteris paribus*" et "conditionnellement à" : la première expression se rapporte à des traits de lois et la seconde à des variables que ces lois produisent ! »

AC: « Absolument. C'est cette distinction que j'avais à l'esprit et qui permet de construire un second scénario mettant en lumière l'impossibilité de substituer en général une expression à l'autre. »

ID: « Peux-tu, pour commencer, me proposer une situation réelle correspondant à ton second scénario ?... »

AC: « En voici une, toujours dans la veine médicale (*cf* Tableau noir 2). Il y est encore question d'un traitement, mais celui-ci est dynamique. Sur la base d'informations initiales contenues dans W, un médecin prescrit une dose faible $a = 0$ ou élevée $a = 1$ d'un certain principe actif. Après une semaine de traitement, un examen permet de recueillir des informations, que je note L, sur la réaction physiologique du patient à la dose initialement prescrite. En fonction de leur nature, le médecin prescrit une dose faible $a' = 0$ (reconduction de la dose initiale si $a = 0$, diminution si $a = 1$) ou élevée $a' = 1$ (reconduction de la dose initiale si $a = 1$, augmentation si $a = 0$). La variable Y quantifie l'effet du traitement dynamique (A, A') sur la maladie, par exemple en termes de survie, de gravité ou de durée. L'effet du traitement statique $(a, a') = (1, 1)$ relativement au traitement statique $(a, a') = (0, 0)$ sur la maladie s'exprime naturellement en termes de comparaison de $Y_{1,1}$ et $Y_{0,0}$. Ici, $Y_{1,1}$ et $Y_{0,0}$ sont les critères de jugement lorsque nous *imposons* deux dosages élevés ou deux dosages faibles successifs, respectivement. »

JCT: « Ce scénario se distingue du précédent en ce que la cause est déterminée séquentiellement. Ce que nous appellons "effet" de $(a, a') \in \{0, 1\}^2$ sur Y pourrait s'exprimer en termes de comparaison de $Y_{1,1}$ et $Y_{0,0}$ par exemple, dont les valeurs sont fonctions *du même* W mais *pas du même* L ! C'est effectivement, je crois, une illustration flagrante de ce que les expressions ne sont pas interchangeables. Il ne ferait pas sens de conditionner selon W et/ou L pour parler de l'effet de (a, a') sur Y. »

AC: « Ici, "*ceteris paribus*" fait clairement référence aux fonctions f_W, f_L, f_Y et à la façon dont les sources d'aléa U_W, U_L, U_Y sont produites et, partant, encore à certains traits seulement de la loi probabiliste du phénomène d'intérêt. »

Tableau noir 1: *Modélisation de la façon dont la nature produit la variable aléatoire* $O = (W, A, Y)$ *sans intervention (gauche) et sous l'intervention* $A = a$ *(droite). Ici,* ceteris paribus *est équivalent à "conditionnellement à* W*".*

$\exists\, f_W, f_A, f_Y$ fonctions déterministes,
$\exists\, U_W, U_A, U_Y$ sources d'aléa indépendantes telles que

"système naturel"	"système contrôlé"
$\begin{cases} W = f_W(U_W) \\ A = f_A(W, U_A) \in \{0,1\} \\ Y = f_Y(W, A, U_Y) \\ O = (W, A, Y) \end{cases}$	$\begin{cases} W = f_W(U_W) \\ A = a \\ Y_a = f_Y(W, A = a, U_Y) \\ O_a = (W, A = a, Y_a) \end{cases}$
la façon dont la nature produit O	la façon dont la nature produit O_a quand nous lui imposons que $A = a$, pour $a \in \{0,1\}$ ici, *ceteris paribus* similaire à "conditionnellement à W"

Tableau noir 2: *Modélisation de la façon dont la nature produit la variable aléatoire* $O = (W, A, L, A', Y)$ *sans intervention (gauche) et sous l'intervention* $(A, A') = (a, a')$ *(droite). Ici,* ceteris paribus *n'est pas équivalent à "conditionnellement à" quoi que ce soit.*

$\exists\, f_W, f_A, f_L, f_{A'}, f_Y$ fonctions déterministes,
$\exists\, U_W, U_A, U_L, U_{A'}, U_Y$ sources d'aléa indépendantes telles que

"système naturel"	"système contrôlé"
$\begin{cases} W = f_W(U_W) \\ A = f_A(W, U_A) \in \{0,1\} \\ L = f_L(W, A, U_L) \\ A' = f_{A'}(W, A, L, U_{A'}) \in \{0,1\} \\ Y = f_Y(W, A, L, A', U_Y) \\ O = (W, A, L, A', Y) \end{cases}$	$\begin{cases} W = f_W(U_W) \\ A = a \\ L_a = f_L(W, A = a, U_L) \\ A' = a' \\ Y_{a,a'} = f_Y(W, A = a, L_a, A' = a', U_Y) \\ O_{a,a'} = (W, A = a, L_a, A' = a', Y_{a,a'}) \end{cases}$
la façon dont la nature produit O	la façon dont la nature produit $O_{a,a'}$ quand nous lui imposons que $(A, A') = (a, a')$, pour $a, a' \in \{0,1\}$
	ici, *ceteris paribus* non similaire à "conditionnellement à quoi que ce soit"

4 Post hoc, ergo propter hoc

ID: « Votre approche est interventionniste. Pour vous, réfléchir causalement c'est être capable d'imposer par une intervention la nature de la cause que nous considérons, et elle seule, puis de raisonner *ceteris paribus*. »

JCT: « Et notre "*ceteris paribus*" ne suffit pas à lui seul à tirer des conclusions causales. Je trouve cela assez caractéristique d'une approche statistique de la causalité… »

AC: « La notion d'intervention apparaît donc comme l'un des artifices techniques qui permettent de formaliser mathématiquement les notions de cause et d'effet. »

ID: « L'intervention n'est pas seulement un artifice technique mathématique. C'est aussi une notion que les philosophes ont utilisée pour tenter de mieux cerner ce qu'est la causalité. »

AC: « Et où cela les a-t-il menés ? »

ID: « Tout d'abord à la conclusion qu'il est conceptuellement pertinent de s'appuyer sur la notion d'intervention, ou de manipulation, pour définir une cause potentielle. Accède ainsi au statut de cause potentielle un facteur pour lequel je peux concevoir une intervention permettant de le modifier *ceteris paribus*. »

AC: « Une intervention réelle ? »

ID: « Pas nécessairement, sans quoi nous rendrions la causalité dépendante de nos capacités d'intervention. Or la causalité est une notion objective. »

AC: « Mais notre capacité à même concevoir des interventions n'est-elle pas le reflet de nos connaissances à l'instant où nous entreprenons cette tâche ? »

JCT: « Si, et voici un exemple déontologiquement délicat. Dans la lignée des travaux des tératologistes du 19è siècle qui ont inscrit les monstres dans le développement de l'humain normal, les travaux pionniers d'E. Wolff ont jeté les bases d'une tératogenèse expérimentale, avec notamment la compréhension des mécanismes des ambiguités sexuelles. Les progrès de la génétique moléculaire ont permis l'étude fine d'observations de la nature (mâle XX, femelle XY), où sexe génétique, sexe apparent et sexe vécu pouvaient être dissociés [4, 15, 5, 14]. Dès lors, du moins chez l'animal, il est devenu possible d'imaginer des modifications de ce déterminisme sexué très tôt dans le développement, voire de proposer des interventions précoces *in utero* pour prévenir une virilisation anormale d'un fœtus féminin. »

AC: « Ainsi, si je comprends bien, le genre devient une cause potentielle dans le cadre de cette analyse philosophique ! Partant, l'ensemble des causes potentielles évolue dynamiquement avec nos connaissances et ce que nous jugeons plausible. La plausibilité biologique

est d'ailleurs l'un des critères de A.B. Hill. »

JCT: « Comme le dit Sherlock Holmes au docteur Watson [7],

> Lorsque vous avez éliminé l'impossible, ce qui reste, fût-il improbable, doit être la vérité.

Cette démarche a permis d'établir une origine bactérienne de l'ulcère gastrique, alors que celle-ci a longtemps été inimaginable parce que ne faisant pas partie du champ des possibles. »

ID: « Tout cela est vrai, et pour éviter que la définition de la causalité ne dépende de l'état de nos connaissances, les philosophes ont étendu la notion d'intervention en incluant ces interventions fictives mais néanmoins concevables, allant jusqu'à accepter des interventions "métaphysiquement possibles". »

AC: « L'analyse de la causalité en termes d'intervention concevable présente-t-elle des failles ? »

ID: « Au moins une ! Il y a cet exemple de toupie électromagnétique que M. Kistler [18] met en avant pour montrer que parfois l'analyse par intervention ne permet pas de distinguer entre causalité et certaines formes d'association régulière. Le nœud de l'affaire c'est que chacune des deux causes potentielles de la rotation de la toupie sur son axe a pour effet l'autre cause potentielle si nous souscrivons à l'analyse. La relation causale ne pouvant être symétrique, il y a contradiction. A ce jour onze objections ont été élevées à ce contre-exemple, et toutes ont été contrées. »

JCT: « Nous avons discuté des dimensions mathématique puis philosophique de la notion d'intervention. Je souhaite ajouter qu'elle a aussi une dimension médicale essentielle d'intervention sur le réel présent ou à venir ! »

ID: « Oui ! Et, plus généralement, l'intervention est un mode d'investigation scientifique. Nous parlons alors plutôt d'expérimentation, mais il s'agit bien de la même chose. »

AC: « Et à quoi opposer alors l'expérimentation ? »

JCT: « A l'observation. C. Bernard, en particulier, distingue les sciences d'observation et les sciences d'expérimentation. Il considère que les secondes sont supérieures aux premières. »

AC: « Et sur quoi fonde-t-il son affirmation de la supériorité de l'expérience sur l'observation ? »

JCT: « Sur le fait que nous pouvons en attendre des résultats bien plus intéressants. Considère par exemple ce passage [3, Deuxième partie, Chapitre 2, VIII, p. 114] :

> Pour conclure avec certitude qu'une condition donnée est la cause prochaine d'un phénomène, il ne suffit pas d'avoir prouvé que cette condition précède ou accompagne toujours le phénomène ; mais il faut encore établir que, cette condition étant supprimée, le phénomène ne se montrera plus. Si nous nous bornions à la seule preuve de présence, on pourrait à chaque instant tomber dans l'erreur et croire à des relations de cause à effet quand il n'y a que simple coïncidence. Les coïncidences constituent, ainsi que nous le verrons plus loin, un des écueils les plus graves que rencontre la méthode expérimentale dans les sciences complexes comme la biologie. C'est le *post hoc, ergo propter hoc* des médecins auquel nous pouvons nous laisser très facilement entraîner, surtout si le résultat de l'expérience ou de l'observation favorise une idée préconçue.

La locution latine *post hoc, ergo propter hoc* signifie "à la suite de cela, donc à cause de cela". »

ID: « Autrement dit : l'expérimentation constitue la voie royale vers l'identification des relations causales et, à l'inverse, il est bien difficile de découvrir des relations causales quand nous en sommes réduits à observer. C'est une idée qui se trouve déjà chez J.S. Mill. »

AC: « Ainsi, dans le scénario du Tableau noir 1, constater, au terme de la réalisation de l'expérience, que le A de l'observation $O = (W, A, Y)$ qui en est issue vaut $a \in \{0, 1\}$ (partie gauche du tableau) est un événement de nature différente de celui qui consiste à constater que le A de l'observation $O_a = (W, A = a, Y_a)$ issue de l'expérience sous l'intervention $A = a$ vaut, par définition, a (partie droite du tableau). Par conséquent, les critères de jugement Y et Y_a n'ont pas la même valeur. »

5 De la population aux individus

JCT: « En tant que médecin, je m'intéresse aux individus plus qu'à des populations. Je suis ainsi de fait confronté à la délicate question de décider quels enseignements je puis véritablement tirer de l'anayse statistique de problèmes causaux. Qu'en pensez-vous ? »

ID: « Le questionnement sur les statistiques, ou plutôt le débat concernant l'éventuelle impossibilité d'en tirer quelqu'enseignement que ce soit pour la médecine, n'est pas nouveau. Nous le trouvons évidemment chez C. Bernard, par exemple dans un passage comme celui-ci [3, Deuxième partie, Chapitre 2, IX, p. 243] :

> Un grand chirurgien fait des opérations de taille par le même procédé ; il fait ensuite un relevé statistique des cas de mort et des cas de guérison, et il conclut, d'après la statistique, que la loi de la mortalité de cette opération est de deux sur cinq. Eh bien, je dis que ce rapport ne signifie absolument rien scientifiquement et ne donne aucune certitude pour faire une nouvelle opération, car nous ne savons pas si ce nouveau cas devra être dans les guéris ou dans les morts.

C. Bernard nie ainsi toute validité externe à des observations recueillies. »

JCT: « L'Académie des sciences s'était déjà penchée sur cette question en 1835, à propos des travaux du docteur J. Civiale consacrés à la comparaison de deux méthodes thérapeutiques concurrentes pour traiter les calculs de la vessie. Nous pouvons lire dans le rapport de séance [9] :

> En matière de statistique (...) le premier soin avant tout c'est de perdre de vue l'homme pris isolément pour ne le considérer que comme fraction de l'espèce (...). En médecine appliquée au contraire, le problème est toujours individuel, les faits ne se présentent à la solution qu'un à un (...). Pour nous les masses restent tout à fait en dehors de la question.

Cela me rappelle le débat sur l'inoculation qui a agité le milieu intellectuel au siècle des Lumières. »

AC: « De quel débat s'agit-il ? »

JCT: « Ce débat fit suite aux travaux de Bernoulli (Daniel, l'un des inventeurs de la théorie statistique, neveu de Jacques, considéré lui comme l'un des inventeurs de la théorie des probabilités), qui s'efforça en 1760 de déterminer quel effet aurait l'inoculation de la petite vérole, généralisée à tous les jeunes enfants, au titre de la prévention de la variole. Un raisonnement de nature probabiliste aboutissant à des comparaisons d'espérances de vie l'amena à se prononcer en faveur de l'inoculation préventive comme mesure salutaire de prophylaxie collective en dépit du risque individuel encouru. »

ID: « Un long débat mathématique et philosophique s'ensuivit, animé notamment par J. d'Alembert, qui développa une minutieuse analyse de l'argumentaire de D. Bernoulli, et écrivit notamment [8] :

> Je suppose avec monsieur Bernoulli que le risque de mourir de l'inoculation [à l'âge de 30 ans] soit de 1 sur 200. Cela posé, il me semble que pour apprécier l'avantage de l'inoculation, il faut comparer, non la vie moyenne de 34 ans à la vie moyenne de 30, mais le risque de 1 sur 200 auquel on s'expose de mourir en un mois par l'inoculation (et cela à l'âge de trente ans, dans la force de la santé et de la jeunesse) à l'avantage éloigné de vivre quatre ans de plus au bout de 60 ans lorsqu'on sera beaucoup moins en état de jouir de la vie... Voilà, il n'en faut point douter, ce qui rend tant de personnes, et surtout tant de mères, peu favorables parmi nous à l'inoculation.

L'un des constats est que l'intérêt collectif se distingue de l'intérêt individuel. »

AC: « C'est bien la vérité de notre condition. Formellement, dans le cadre du Tableau noir 1 avec $W = \emptyset$ réduit à rien, en notant $a = 1$ la pratique de l'inoculation et $a = 0$ son contraire d'une part, et $Y_a = 1$ le développement de la variole et $Y_a = 0$ son contraire sous l'intervention $a \in \{0, 1\}$ d'autre part, alors quatre cas de figures sont envisageables pour un individu donné, selon son appartenance à l'un parmi quatre groupes de combinaisons des

valeurs possibles 0-1 pour Y_0 et Y_1 respectivement : $Y_0 = Y_1 = 1$ (groupe $\mathcal{G}_1$), $Y_0 = 1, Y_1 = 0$ (groupe $\mathcal{G}_2$), $Y_0 = 0, Y_1 = 1$ (groupe $\mathcal{G}_3$) et $Y_0 = Y_1 = 0$ (groupe $\mathcal{G}_4$). »

JCT: « Je vois où tu veux en venir ! Soit p_k la proportion de la population totale couverte par le groupe $\mathcal{G}_k \ldots$ »

ID: « Qu'entends-tu précisément par cela ? »

JCT: « Cela signifie que si je désigne une personne au hasard dans la population sans avoir connaissance de son groupe d'appartenance, alors la probabilité qu'elle appartienne au groupe $\mathcal{G}_k$ égale p_k. Dans le modèle qu'Antoine évoquait où $Y_0 = Y_1 = 1$ pour les membres du groupe $\mathcal{G}_1$, $Y_0 = 1, Y_1 = 0$ pour ceux du groupe $\mathcal{G}_2$, $Y_0 = 0, Y_1 = 1$ pour ceux du groupe $\mathcal{G}_3$ et $Y_0 = Y_1 = 0$ pour ceux du groupe $\mathcal{G}_4$, eh bien l'inoculation a un effet causal sur le développement de la variole dès lors que $p_2 > 0$ *ou* $p_3 > 0$. »

AC: « Ici, $p_2 > 0$ et $p_3 > 0$ signifient que les groupes $\mathcal{G}_2$ et $\mathcal{G}_3$ ne sont pas vides. »

ID: « En quel sens cela induit-il un effet causal ? »

AC: « Au sens où l'inoculation $A = 1$ ou l'absence d'inoculation $A = 0$ affectent mon devenir $Y = Y_A$ relatif au développement de la variole si j'appartiens à l'un des groupes $\mathcal{G}_2$ ou $\mathcal{G}_3 \ldots$ »

JCT: « En revanche, si j'appartiens à l'un des groupes $\mathcal{G}_1$ ou $\mathcal{G}_4$, alors l'inoculation ou l'absence d'inoculation ne le changent en rien. »

ID: « Ce qui conclut la description à l'échelle individuelle. Quant à la description à l'échelle collective, nous constatons que l'inoculation n'a un effet statistique *bénéfique* que si, et seulement si, $p_2 + p_4 > p_3 + p_4$, c'est-à-dire si et seulement si $p_2 > p_3$, soit encore à condition que la proportion des individus qui bénéficieraient de l'inoculation (ceux issus de $\mathcal{G}_2$) soit plus grande que la proportion de ceux qui en pâtiraient (ceux issus de $\mathcal{G}_3$). »

AC: « Au niveau collectif, la question est donc d'estimer la différence $p_2 - p_3$. Au niveau individuel, la question est de déterminer à quel groupe appartient chaque individu. Pour entreprendre ces deux tâches statistiques, il faut disposer d'informations supplémentaires relatives à chaque individu qui soient pertinentes. »

JCT: « Tout se joue dans ton qualificatif de "pertinentes" ! »

AC: « Nous y reviendrons. »

6 Des mondes contrefactuels au monde actuel

JCT: « Dans ce scénario très particulier, au sens où Y_0 et Y_1 sont déterministes dans chacun des quatre groupes, l'indécision relative à l'appartenance à un groupe est rigoureusement équivalente à l'impossibilité d'observer à la fois Y_0 et Y_1. Un individu donné est soit inoculé, soit non inoculé. Dans le premier cas je ne sais pas si l'individu développerait la maladie s'il n'était pas inoculé, dans le second je ne sais pas s'il la développerait s'il ne l'était pas. Le modèle est contrefactuel ! »

AC: « Effectivement, il y a ici quelque chose de conceptuellement difficile : il apparaît qu'il y a plus ou autre chose, dans la causalité, que ce qui se passe effectivement dans notre monde. »

ID: « C'est bien une position de ce type qui se trouve au cœur des théories philosophiques "contrefactuelles" de la causalité. Ces théories reposent sur l'idée fondamentale selon laquelle A a causé B si et seulement si B n'aurait pas été le cas si A n'avait pas été le cas (ou alors, dans une version probabiliste, la probabilité de B aurait été moindre si A n'avait pas été le cas). Cela signifie que, essentiellement, la causalité a à voir non pas seulement avec ce qui est le cas, ce qui se passe effectivement dans notre monde, mais également avec ce qui n'est pas le cas, ce qui se passe dans un "autre monde possible" — pour parler comme les philosophes. »

JCT: « Entendu. Mais que faites-vous de la difficulté méthodologique que j'indiquais plus tôt ? Il semble bien difficile de déterminer ce qui se serait passé si les choses avaient été différentes. Disposez-vous d'un cadre conceptuel et d'outils statistiques qui vous permettent de répondre à de telles questions ? »

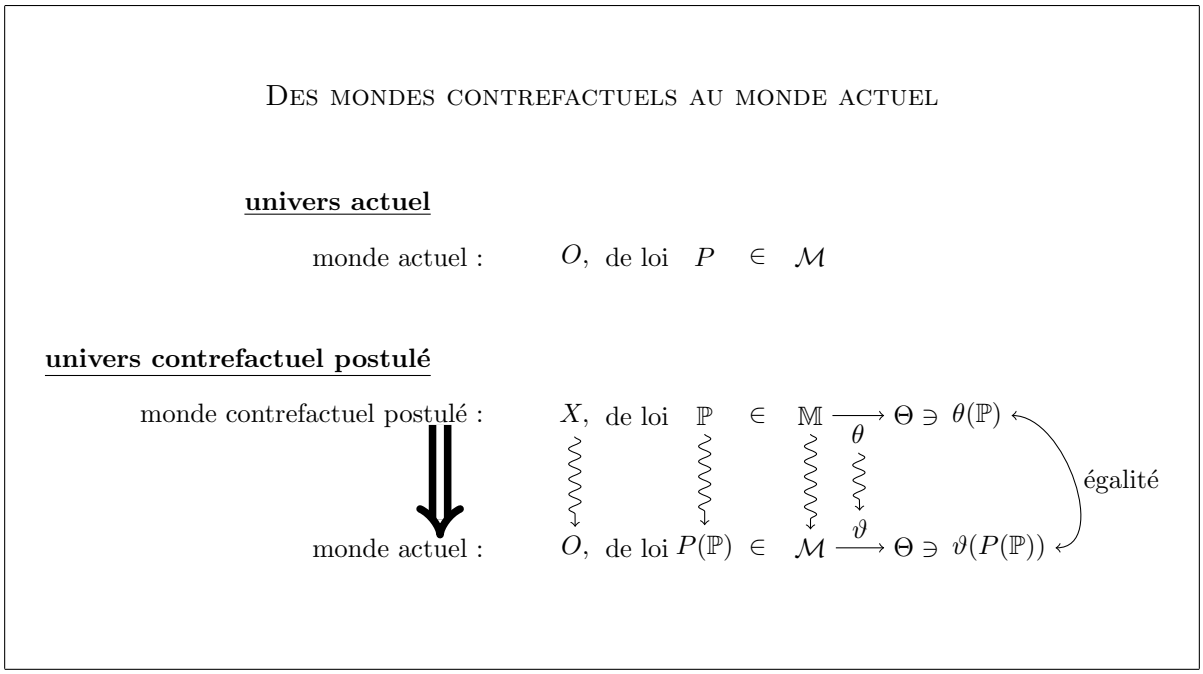
ID: « S'il faut entendre ta question à l'échelle de l'individu, la réponse est non. S'il faut en revanche l'entendre à l'échelle de toute la population, il existe bien de tels cadres conceptuels et outils statistiques. »

AC: « Admettons ainsi que nous adhérons au modèle probabiliste contrefactuel suivant. Je considère un individu tiré au hasard dans ma population. Il se voit associer une donnée X, dite complète, qui se décompose en une superposition finie de parties X_i, $i \in I$, qui sont éventuellement redondantes ; je l'écris $X = (X_i)_{i\in I}$. Il faut penser à X_i comme contenant la description de ce qui se passe pour cet individu dans le $i$ème monde contrefactuel. Connaître X, c'est disposer simultanément des issues de l'expérience pour cet individu dans chacun des mondes contrefactuels et donc, en particulier, dans le monde réel, qui se trouve être l'un d'eux. L'observation dans le monde réel, O, se conçoit comme une projection de la donnée complète X dans le monde réel, avec toute la perte d'information que cela implique. »

JCT: « Ainsi donc, si je nomme $\mathbb{P}$ la loi de la donnée complète X et P la loi de l'observation O, la question que je posais se reformule de la façon suivante : pouvons-nous inférer des traits de $\mathbb{P}$ à partir d'observations tirées sous P ? ! J'entends par "traits" des caractéristiques

(i) qui mettent en jeu la comparaison des mondes contrefactuels, et *(ii)* qui sont exprimées à l'échelle de la population et pas à celle des individus. »

Tableau noir 3: *Illustration de la façon dont un univers contrefactuel postulé induit l'univers actuel.*



AC: « Absolument (*cf* Tableau noir 3). En statisticien, je conçois ces traits qui t'intéressent sous la forme d'une fonctionnelle $\theta : \mathbb{M} \to \Theta$ qui associe à toute loi $\mathbb{P} \in \mathbb{M}$ que peut suivre la donnée complète X ce trait $\theta(\mathbb{P}) \in \Theta$. Puisque nous savons comment l'observation O est déduite de X, nous pouvons construire une seconde fonctionnelle $\vartheta : \mathcal{M} \to \Theta$ qui associe à toute loi $P \in \mathcal{M}$ que peut suivre O un trait $\vartheta(P) \in \Theta$ de telle sorte que si je note $P(\mathbb{P})$ la loi que suit O quand X suit $\mathbb{P}$ alors, moyennant une hypothèse dite de "randomisation", $\vartheta(P(\mathbb{P})) = \theta(\mathbb{P})$! Le miracle, si je puis dire, c'est qu'il est possible d'inférer $\vartheta(P(\mathbb{P}))$ à partir des observations tirées sous $P(\mathbb{P})$ et donc, par ricochet, d'inférer $\theta(\mathbb{P})$ quand bien même nous ne disposons pas d'observations tirées sous $\mathbb{P}$! »

JCT: « L'hypothèse de randomisation mérite plus qu'une simple évocation (*cf* Sections 8 et 9). Il faut bien comprendre qu'elle porte sur la loi $\mathbb{P}$ dans toute sa complexité, c'est-à-dire en mettant en jeu simultanément tous les mondes contrefactuels, de telle sorte qu'elle est, par essence, intestable à partir des observations tirées sous P. »

ID: « Il serait bienvenu, avant de creuser ce sillon, que nous replacions cette discussion dans le contexte des deux exemples précédents. »

AC: « Revenons sur le scénario du Tableau noir 1. La donnée complète X dans le monde contrefactuel s'y écrit $X = (W, A, Y_0, Y_1)$, ou $X = (X_0, X_1)$ avec $X_0 = (W, A, Y_0)$ et

$X_1 = (W, A, Y_1)$, et l'observation O dans le monde réel s'écrit $O = (W, A, Y_A)$, ou encore X_A pour simplifier. »

ID: « Et quelles fonctionnelles θ et ϑ pourrions-nous considérer pour aborder le scénario que nous évoquions de l'évaluation de l'effet éventuel d'une traitement relativement à l'absence de tout traitement sur une maladie, par exemple en termes de survie ou de décès ? »

JCT: « C'est le choix de la fonctionnelle $\theta : \mathbb{M} \to \Theta$ qui est le plus simple. »

ID: « Pourquoi donc ? »

JCT: « Parce que θ associe un trait $\theta(\mathbb{P})$ à toute loi $\mathbb{P}$ que peut suivre la donnée complète X. Or justement, il est conceptuellement aisé de caractériser une mesure d'effet lorsque nous disposons des issues contrefactuelles ! Ainsi, soyons statisticiens et convenons qu'une comparaison des valeurs moyennes $\mathbb{P}\{Y_1\}$ et $\mathbb{P}\{Y_0\}$ que prennent Y_1 et Y_0, qui sont les critères de jugement quantifiant les issues de la maladie quand nous imposons soit le traitement, pour Y_1, soit l'absence de tout traitement, pour Y_0, ouvre une fenêtre sur le cœur du mécanisme causal en quantifiant l'effet éventuel de la cause sur la maladie. »

AC: « La fonctionnelle $\theta : \mathbb{M} \to \Theta = [-1, 1]$ qui est caractérisée par $\theta(\mathbb{P}) = \mathbb{P}\{Y_1\} - \mathbb{P}\{Y_0\}$ joue bien ce rôle. On l'appelle l'excès de risque causal, qui est à valeurs dans l'intervalle $[-1, 1]$. »

ID: « Je suis abasourdie par votre hardiesse ! Il me semblait que nous devisions de ce que nous entendions par "cause" et son "effet", et vous voilà quantifiant cette notion ! Ne vaudrait-il pas mieux décider d'abord si la "cause" est bien cause et son "effet", un effet ? »

AC: « Le statisticien élaborera une réponse à ta question en mettant au point une procédure, dite de test, qui s'appuie sur cette quantification ! »

JCT: « Qu'en est-il alors du ϑ dont tu dis, Antoine, que tu sais l'associer à θ ? »

AC: « Nous pouvons justifier que, sous l'hypothèse de randomisation déjà évoquée… »

ID: « … et sur laquelle il nous faudra revenir ! (*cf* Sections 8 et 9)… »

AC: « … nous associons naturellement à la question la fonctionnelle $\vartheta : \mathcal{M} \to \Theta = [-1, 1]$ caractérisée par $\vartheta(P) = P\{P(Y|A = 1, W)\} - P\{P(Y|A = 0, W)\}$ et appelée excès de risque (généralisé). »

ID: « Que représente l'expression $P\{P(Y|A = a, W)\}$ pour $a \in \{0, 1\}$? »

JCT: « L'explication se développe en deux temps. *Primo*, $P(Y|A = a, W)$ est une variable aléatoire qui ne dépend que de W, soit encore $P(Y|A = a, W) = \varphi(W)$. Informellement, $\varphi(\omega)$ est la valeur moyenne de Y sous P quand nous *observons* $A = a$ et $W = \omega$. *Secundo*,

de manière analogue au fait que $\mathbb{P}\{Y_a\}$ est la valeur moyenne sous $\mathbb{P}$ de la variable aléatoire Y_a, $P\{P(Y|A=a,W)\}$ est la valeur moyenne de $\varphi(W)$ sous P. »

ID: « Et pour le second scénario ? »

AC: « Voilà ce que cela donne (*cf* Tableau noir 4). »

Tableau noir 4: *Elaboration d'un paramètre statistique d'excès de risque associé à un certain excès de risque causal dans le scénario du Tableau noir 2.*

donnée complète :	$X = (W, A, L_0, L_1, A', Y_{0,0}, Y_{0,1}, Y_{1,0}, Y_{1,1}) \sim \mathbb{P} \in \mathbb{M}$
observation :	$O = (W, A, L_A, A', Y_{A,A'}) \sim P(\mathbb{P}) \in \mathcal{M}$
excès de risque causal :	$\theta(\mathbb{P}) = \mathbb{P}\{Y_{1,1}\} - \mathbb{P}\{Y_{0,0}\}$
excès de risque :	$\vartheta(P) = \vartheta_{1,1}(P) - \vartheta_{0,0}(P)$, avec
	$\vartheta_{a,a'}(P) = P\Big\{ \sum_{\ell \in \mathcal{L}} P(Y\|A'=a', L=\ell, A=a, W) \times P(L=\ell\|A=a,W) \Big\}$

JCT: « Je réalise que nous pouvons désormais jeter un regard neuf sur notre discussion relative au *distinguo* individus *versus* population et à l'exemple historique de l'inoculation. »

ID: « Nous t'écoutons. »

JCT: « Je me place à cette fin de nouveau dans le cadre du scénario du Tableau noir 1, en lui tordant le cou, puisque je suppose de surcroît que la variable W permet d'identifier exactement à quel groupe chaque individu appartient. »

AC: « En somme, si j'observe $W = \omega$ alors l'individu appartient au groupe $\mathcal{G}_\omega$, et donc, $P(W = \omega) = p_\omega$ pour $\omega \in \{1, 2, 3, 4\}$. »

JCT: « En effet. J'attire alors votre attention sur le fait suivant : $\vartheta(P) = p_3 - p_2$, comme le prouve ce simple calcul (*cf* Tableau noir 5). »

ID: « Ainsi, nous retombons bien sur nos pieds : si l'information contrefactuelle de l'appartenance aux groupes était disponible alors l'excès de risque aurait une interprétation causale. »

AC: « Nous évoquions plus tôt (*cf* Section 5) la nécessité de disposer d'informations supplémentaires relatives à chaque individu qui soient « pertinentes » pour estimer la différence

Tableau noir 5: *Démonstration de l'égalité $\vartheta(P) = p_3 - p_2$ dans le cadre du scénario du Tableau noir 1 lorsque $W \in \{1, 2, 3, 4\}$ informe de l'appartenance à l'un des groupes $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_3$, $\mathcal{G}_4$.*

posons, pour alléger les notations : $\Delta(W) = P(Y|A = 1, W) - P(Y|A = 0, W)$
par définition :

$$\vartheta(P) = P\{\Delta(W)\} = \sum_{\omega \in \{1,2,3,4\}} \Delta(\omega) \times P(W = \omega) = \sum_{\omega \in \{1,2,3,4\}} \Delta(\omega) \times p_\omega$$

et au cas par cas :

ω	$\Delta(\omega)$
1	$1 - 1 = 0$
2	$0 - 1 = -1$
3	$1 - 0 = 1$
4	$0 - 0 = 0$

donc : $\vartheta(P) = p_3 - p_2$!

$p_2 - p_3$ au niveau collectif, et l'appartenance à l'un des quatre groupes au niveau individuel. Nous venons d'argumenter que si l'information contrefactuelle de l'appartenance aux groupes était disponible alors ce serait une information pertinente suffisante. Mais elle est indisponible, de par sa contrefactualité, et l'une des tâches du statisticien est de lui trouver des succédanés. Nous parlons d'ailleurs plutôt de « prédicteurs », car ils ont vocation à aider le statisticien à prédire ce que vaut la variable contrefactuelle que nous aurions aimé observer, ou bien la probabilité de l'observer. »

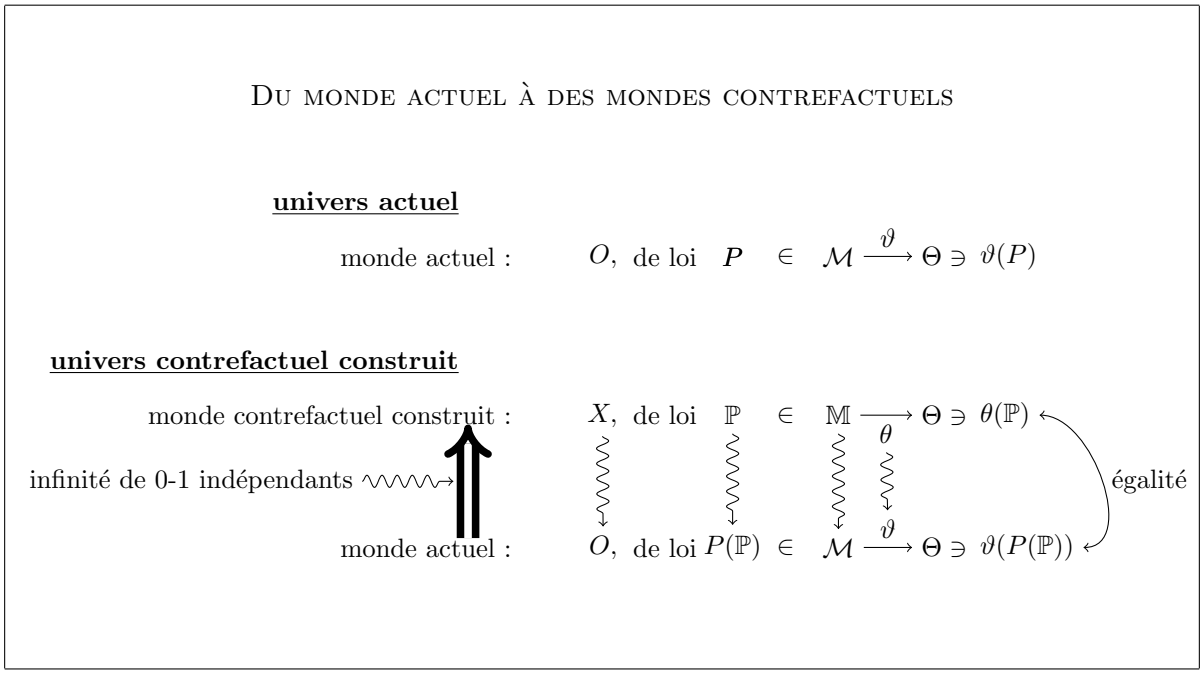
7 Du monde actuel à des mondes contrefactuels

JCT: « Si nous prenons un peu de recul, nous avons illustré comment modéliser la causalité si nous acceptons la possibilité de mondes contrefactuels dont notre monde actuel découlerait, au sens où ce qui se passe dans notre monde serait la projection de ce qui se passe dans l'un d'entre eux. Nous en avons notamment tiré un bénéfice formel : $\mathbb{P}$ induit $P(\mathbb{P})$ et donc $\mathbb{M}$ induit $\mathcal{M}$, θ induit ϑ, X induit O *etc.* L'édifice formel s'écroule-t-il si nous refusons cette conception du monde réel ? »

AC: « Non. Il est formellement possible d'adopter un point de vue diamétralement opposé, pour peu que nous admettions pouvoir tirer à pile ou face une infinité de fois en totale indépendance. »

ID: « En vertu du début de nos échanges, tu nous demandes ainsi de t'autoriser à recourir à autant de variables aléatoires indépendantes que tu le souhaiteras. »

Tableau noir 6: *Illustration de la construction, à partir de l'univers réel, d'un univers contrefactuel l'induisant. La construction nécessite de tirer une infinité de fois à pile ou face de façon indépendante.*



AC: « Absolument ! Je commence par poser le décor (*cf* Tableau noir 2). Soit $\mathcal{M}$ l'ensemble des lois P candidates à être celle de l'observation O dans le monde actuel. Notons $\mathbb{M}$ l'ensemble des lois $\mathbb{P}$ candidates à être celle de la variable contrefactuelle X auxquels nous accèderions si nous acceptions la possibilité de mondes contrefactuels dont notre monde actuel découlerait ; en particulier, X induirait O. La question d'intérêt nous conduirait à introduire une fonctionnelle *ad hoc* $\theta : \mathbb{M} \to \Theta$, qui à son tour induirait la fonctionnelle $\vartheta : \mathcal{M} \to \Theta$. »

JCT: « Rien de neuf pour le moment, tu te fais désirer ! A quel tour nous prépares-tu ? »

ID: « Je dirais que ce préambule était nécessaire pour introduire la fonctionnelle $\vartheta : \mathcal{M} \to \Theta$. Mon intuition est-elle bonne ? »

AC: « Elle est excellente ! Maintenant, l'idée est la suivante. Il est formellement possible :
– de construire une variable contrefactuelle $\tilde{X}$, de même nature que X, qui induit O de la même façon que X induit O ;
– de construire, pour toute loi candidate $P \in \mathcal{M}$, une loi $\tilde{\mathbb{P}}(P)$ de $\tilde{X}$, d'où l'obtention de l'ensemble $\tilde{\mathbb{M}} = \{\tilde{\mathbb{P}}(P) : P \in \mathcal{M}\}$, de telle sorte que si $\tilde{\mathbb{P}}(P)$ induit P' de la même façon que $\mathbb{P}$ induit $P(\mathbb{P})$ alors $P' = P$;
tout cela de telle sorte que $\theta(\tilde{\mathbb{P}}(P)) = \vartheta(P)$! En somme il est formellement possible de

construire, à partir de O, $\mathcal{M}$, θ et ϑ, un monde contrefactuel de même nature que celui résumé par X, $\mathbb{M}$ et θ. »

JCT: « Nous nous laisserions presque troubler par le fait que nous nous retrouvons ainsi en mesure de tirer au hasard une variable contrefactuelle $\tilde{X}$ compatible avec l'observation O, et donc d'observer celle-ci quand bien même cela n'est pas possible dans le monde actuel ! »

ID: « Troublant en effet, jusqu'à ce que nous réalisions que la causalité dans ce monde contrefactuel formellement construit ne peut coïncider avec aucune notion de causalité dans notre monde ! »

JCT: « L'artifice technique est certes formellement commode, mais nous n'accédons ainsi à aucune interprétation causale. »

8 De la ruse de randomisation

ID: « Ainsi le formalisme contrefactuel permet de définir des paramètres causaux quantifiant les questions causales qui nous préoccupent. Il nous enseigne aussi que les conditions pour raisonner causalement en termes *post hoc, ergo propter hoc* ne sont en général pas réunies. Il est peut-être temps de revenir sur la ruse de randomisation que nous avons effleurée plus tôt (*cf* Section 6). En quoi consiste-t-elle ? »

JCT: « C'est une ruse qui a vocation à garantir qu'un raisonnement *post hoc, ergo propter hoc* est de nature causale. En d'autres termes, la ruse de randomisation consiste à réunir des conditions sous lesquelles nous pouvons inférer les paramètres causaux sur la base de l'observation des résultats de la répétition d'une expérience. Le principe est de contrôler l'exposition en la tirant indépendamment des conséquences qu'aurait chaque intervention. »

ID: « Considérons le scénario le plus simple qui soit, où l'objectif est de déterminer l'effet causal d'un traitement sur une maladie relativement à l'absence de traitement (par exemple, la prise d'un placebo). Si je comprends bien, la séquence des événements est la suivante : *primo*, je recrute un patient dans une population de patients potentiels bien identifiée ; *deuxio*, je tire à pile ou face (par exemple avec une pièce équilibrée, c'est-à-dire selon la loi Bernoulli de paramètre 1/2) la prise du traitement ou celle du placebo et je l'impose à mon patient — ce faisant, la nature de l'exposition est par construction indépendante des deux issues que pourraient avoir chez mon patient, en absence de connaissance sur son statut contrefactuel (*cf* Section 6), la prise du traitement et celle du placebo ; *tertio*, j'accompagne le patient jusqu'à l'observation de l'issue. »

JCT: « Tu as tout à fait raison. J'aimerais mettre en exergue ce en quoi cette procédure randomisée s'écarte du cadre observationnel tel que critiqué par C. Bernard lorsqu'il mettait en avant le cadre expérimental contrôlé. Dans ce dernier, le patient que tu as recruté

pourrait consulter son médecin. Le médecin pourrait, sur la base du dossier médical et d'un examen, décider de prescrire la prise du traitement ou d'un placebo. Dans ce cas évidemment, la nature de l'exposition serait par construction dépendante des deux issues qu'auraient la prise du traitement et celle du placebo. »

AC: « Nous pouvons jeter un éclairage formel sur vos deux scénarios grâce à celui du Tableau noir 1. Jean-Christophe, ce que tu viens de décrire relève exactement du "système naturel" qui y est exposé. Isabelle, le patient que tu recrutes est caractérisé par la variable W dont tu n'as pas besoin de tenir compte lors du tirage. Tirer au hasard la nature de l'exposition revient en somme à substituer à l'équation $A = f_A(W, U_A)$ du "système naturel" l'équation alternative $A = U_A$ avec U_A de loi Bernoulli de paramètre $1/2$. »

ID: « Je vois ! Et plutôt que de parler de substitution comme tu le fais, on pourrait dire que nous faisons en sorte, *via* la randomisation, que $f_A(W, U_A) = U_A$ avec U_A de loi Bernoulli de paramètre $1/2$: en somme, tu imposes la forme de la fonction f_A ! »

AC: « C'est vrai. L'intérêt de cette formalisation est notamment de faire apparaître que, selon le résultat du tirage de randomisation, nous observons l'un ou l'autre des systèmes contrôlés, et donc que nous pouvons interpréter causalement le résultat d'une comparaison de leurs deux comportements. »

JCT: « Tu veux dire : interpréter causalement une comparaison moyenne des critères de jugement quantifiant les issues observées du traitement ou du placebo. »

AC: « C'est bien cela. Ecoutez plutôt. Disons que nous considérons l'excès de risque causal $\theta : \mathbb{M} \to [-1, 1]$ caractérisé par $\theta(\mathbb{P}) = \mathbb{P}\{Y_1\} - \mathbb{P}\{Y_0\}$. Admettons l'hypothèse de cohérence qui veut que $Y = Y_A$, cette égalité s'interprétant comme la coïncidence de l'issue dans le monde actuel avec l'issue dans le monde contrefactuel exploré. Notons que, par construction grâce à la randomisation, $A = U_A$ est indépendante de (Y_0, Y_1). Eh bien $\theta(\mathbb{P})$ coïncide avec la différence des valeurs moyennes conditionnelles de Y sachant $A = 1$ et sachant $A = 0$, respectivement, ce que nous appelons l'excès de risque naïf : $P(Y|A = 1) = P(Y|A = 0)$. Formellement : *(i)* $P(Y|A = 1) - P(Y|A = 0) = \mathbb{P}(Y_A|A = 1) - \mathbb{P}(Y_A|A = 0)$ en vertu de la cohérence, *(ii)* cette différence est égale à $\mathbb{P}(Y_1|A = 1) - \mathbb{P}(Y_0|A = 0)$ en remplaçant Y_A par Y_1 ou Y_0, selon que $A = 1$ ou $A = 0$, et *(iii)* celle-ci vaut enfin $\theta(\mathbb{P})$ en vertu de l'indépendance entre A et (Y_0, Y_1) ! »

JCT: « Je souhaiterais maintenant suggérer que la présentation de la ruse de randomisation que nous venons de proposer en retient bien la substantifique moelle. »

ID: « Et comment comptes-tu y parvenir ? »

JCT: « Revenons à l'exemple plus complexe déjà évoqué (*cf* Tableau noir 2). Nous souhaitons déterminer l'effet causal non pas d'un traitement relativement à un placebo mais plutôt d'un traitement établi en séquence et caractérisé par un couple (a, a') avec $a, a' \in \{0, 1\}$

relativement au traitement de référence noté $(0, 0)$. »

AC: « Présenté ainsi, la seule différence avec ce qui précède est la caractérisation de l'exposition selon quatre niveaux différents et non pas deux. La séquence des événements proposées par Isabelle tient toujours : *primo*, je recrute un patient dans une population de patients potentiels bien identifiée ; *deuxio*, je tire au hasard (par exemple avec deux pièces équilibrées lancées indépendamment) la nature (a, a') du traitement et je l'impose à mon patient — ce faisant, la nature de l'exposition est par construction indépendante des quatre issues qu'auraient les quatre prescriptions possibles ; *tertio*, j'accompagne le patient jusqu'à l'observation de l'issue. »

JCT: « Et, comme tu le faisais plus tôt, nous pouvons jeter un éclairage formel grâce au scénario du Tableau noir 2. Ce que tu décris relève exactement du "système naturel" qui y est exposé. Le patient que tu recrutes est d'abord caractérisé par la variable W dont tu n'as pas besoin de tenir compte puisque tirer au hasard comme tu le fais la nature de A revient à imposer que $f_A(W, U_A) = U_A$ avec U_A variable de loi Bernoulli de paramètre $1/2$. Tu peux, de façon similaire, négliger l'information intermédiaire résumée par la varible L car tirer au hasard comme tu le fais la nature de A' revient à imposer que $f_{A'}(W, A, L, U_{A'}) = U_{A'}$ avec $U_{A'}$ variable de loi Bernoulli de paramètre $1/2$ indépendante de U_A. Accompagner le patient jusqu'à l'observation de l'issue revient à attendre d'observer Y. »

ID: « Je prends la relève ! Disons que nous nous intéressons à l'excès de risque causal $\theta : \mathbb{M} \to [-1, 1]$ caractérisé par $\theta(\mathbb{P}) = \mathbb{P}\{Y_{1,1}\} - \mathbb{P}\{Y_{0,0}\}$ et admettons l'hypothèse de cohérence qui veut encore que l'issue dans le monde actuel, Y, coïncide avec celle dans le monde contrefactuel exploré, $Y_{A,A'}$. Le "miracle" de la randomisation, pour reprendre ton expression, Antoine, c'est que, dans le décor que nous venons de poser, le paramètre causal $\theta(\mathbb{P})$ soit égal à la différence des valeurs moyennes conditionnelles de Y sachant $(A, A') = (1, 1)$ et sachant $(A, A') = (0, 0)$, respectivement. »

JCT: « Cela conclut le premier élan de complexification. J'ambitionne de discuter un cas plus délicat dans lequel nous souhaitons comparer des régimes de traitement différents. »

ID: « Qu'entends-tu par là ? »

JCT: « Disons que je souhaite entreprendre la comparaison causale d'un régime de traitement statique, au sens où la dose A initialement prescrite est nécessairement reconduite après la visite intermédiaire, soit $A = A'$, relativement à un régime de traitement dynamique dans lequel la seconde dose A' peut être différente de la première en fonction des résultats de la visite intermédiaire consignés dans L. »

AC: « Formellement, dans le cadre du système naturel du Tableau noir 2, nous avons $f_A(W, U_A) = f_A(W)$, qui ne dépend que de W, les régimes statique et dynamique correspondant respectivement à $f_{A'}(W, A, L, U_{A'}) = A$ et $f_{A'}(W, A, L, U_{A'}) = f_{A'}(W, L)$, qui dépend de W et de L uniquement. »

ID: « La situation réelle que tu proposais Jean-Christophe pour illustrer le second scénario (fin de la Section 3) en apparaît donc comme un cas particulier. Pour les mêmes raisons que celles exposées précédemment, il n'est pas possible de tirer de conclusions causales de l'observation des deux systèmes naturels correspondant aux régimes statique et dynamique. En intervenant selon la ruse de randomisation, nous pouvons en revanche créer les conditions expérimentales qui transforment les deux systèmes naturels en deux systèmes contrôlés dont l'observation et la confrontation conduisent à des conclusions causales. »

JCT: « La randomisation doit porter sur l'assignation du régime, statique ou dynamique. Il nous faut donc faire apparaître une nouvelle variable dans le système dont nous n'avions jusque-là pas eu besoin. Appelons-la R pour "régime". Elle se situe chronologiquement entre la description du patient, W, et la première assignation de traitement, A. La variable R témoigne de la nature du régime que détermine le médecin sur la base de l'observation de W dans le système naturel. Disons que $R = 0$ pour le régime statique et $R = 1$ pour le dynamique. Cela aboutit aux systèmes naturel et contrôlé suivants (*cf* Tableau noir 7). »

Tableau noir 7: *Modélisation illustrant la ruse de randomisation pour l'étude de l'effet d'un régime de traitement dynamique, comme prolongement du scénario développé dans le Tableau noir 2.*

$\exists\, f_W, f_R, f_A, f_L, f_{A'}, f_Y$ fonctions déterministes,
$\exists\, U_W, U_R, U_A, U_L, U_{A'}, U_Y$ sources d'aléa indépendantes telles que

"système naturel"	"système contrôlé"
$\begin{cases} W = f_W(U_W) \\ R = f_R(W, U_R) \in \{0,1\} \\ A = f_A(W, R, U_A) \in \{0,1\} \\ L = f_L(W, R, A, U_L) \\ A' = f_{A'}(W, R, A, L, U_{A'}) \in \{0,1\} \\ Y = f_Y(W, R, A, L, A', U_Y) \\ O = (W, R, A, L, A', Y) \end{cases}$	$\begin{cases} W = f_W(U_W) \\ R = r \\ A_r = f_A(W, R = r, U_A) \\ L_r = f_L(W, R = r, A_r, U_L) \\ A'_r = f_{A'}(W, R = r, A_r, L_r, U_{A'}) \\ Y_r = f_Y(W, R = r, A_r, L_r, A'_r, U_Y) \\ O_r = (W, R = r, A_r, L_r, A'_r, Y_r) \end{cases}$
la façon dont la nature produit O	la façon dont la nature produit O_r quand nous lui imposons que $R = r$, pour $r \in \{0,1\}$

AC: « Formellement, la randomisation selon le régime R revient à faire en sorte que $f_R(W, U_R) = U_R$ avec U_R de loi de Bernoulli de paramètre $1/2$ par exemple. Plutôt que de laisser au médecin la responsabilité du choix du régime, c'est au hasard que nous nous remettons *via* le tirage de randomisation, dont l'issue détermine duquel des deux systèmes contrôlés nous observons une émanation. »

9 De la confusion et de l'hypothèse de randomisation

ID: « Fort bien ! Mais que se passe-t-il quand nous ne sommes pas en situation de recourir à la ruse de randomisation ? Y a-t-il un quelconque enseignement à tirer de ce que nous venons de voir pour le cas où nous sommes réduits à n'observer que le comportement du système naturel ? »

JCT: « Il y en a bien un, qui a abouti à la notion de confusion et généré l'hypothèse de randomisation qu'Antoine évoquait plus tôt. La formulation exacte de cette hypothèse de randomisation dépend de la question d'intérêt. Elle concerne la nature de la dépendance conditionnelle de la variable considérée comme cause relativement aux variables contrefactuelles d'effets sachant certaines autres variables, que nous qualifions de potentiels facteurs de confusion. »

ID: « Lisons-nous bien en creux ce que sont des facteurs de confusion ? Je dirais que ce sont ces variables qui créent une dépendance entre la variable de cause supposée et les variables contrefactuelles de ses effets supposés, dépendance qui ne renvoie pas à une relation de cause à effet. »

AC: « C'est très juste ! Sans doute serait-il pertinent de nous appuyer sur nos exemples précédents pour expliquer plus avant ces notions subtiles. »

JCT: « Replongeons-nous dans le cadre du Tableau noir 1. Le facteur de confusion y est W, la variable de cause supposée y est A et les variables contrefactuelles de ses effets supposées y sont Y_0, Y_1. L'excès de risque causal $\theta(\mathbb{P})$ quantifie l'effet causal d'intérêt. Si W est un facteur de confusion alors l'excès de risque naïf $P(Y|A=1) - P(Y|A=0)$ diffère de $\theta(\mathbb{P})$. L'existence de cet écart entre les deux quantités, appelé le biais de confusion, est un souci constant dans les études observationnelles en épidémiologie. »

ID: « Comment justifies-tu l'irruption du risque naïf ici ? »

AC: « Je pense que Jean-Christophe l'a fait entrer en scène parce que lorsque nous recourons à la ruse de randomisation cette grandeur coïncide avec $\theta(\mathbb{P})$. »

JCT: « C'est bien ça. En revanche si nous négligeons W sans recourir à cette ruse, c'est-à-dire si nous exploitons cet excès de risque naïf en lieu et place de $\theta(\mathbb{P})$ tandis que nous observons le système naturel, alors nous n'accédons pas à la relation de causalité. »

ID: « Et si nous ne le négligeons pas ? »

JCT: « Eh bien si W contient bien tous les facteurs de confusion, alors l'hypothèse de randomisation est satisfaite : nous avons A indépendante de (Y_0, Y_1) sachant W, et $\vartheta(P) = \theta(\mathbb{P})$ comme Antoine l'affirmait plus tôt (*cf* Section 6). »

ID: « En somme, l'hypothèse de randomisation se substitue à la ruse de randomisation

pour réunir des conditions permettant d'inférer des relations causales de l'observation du système naturel sans aucune intervention sur celui-ci. Forte de cette mise en perspective, nous pouvons concevoir inversement la ruse de randomisation comme un outil pour garantir la validité de l'hypothèse de randomisation. L'intervention sur le système naturel selon la ruse de randomisation (*cf* Section 8) apparaît en ce sens comme participant d'un procédé d'auto-validation comme le dit N. Cartwright [6]. »

AC: « Il se trouve que dans le cadre du Tableau noir 1 l'hypothèse de randomisation est satisfaite (cela découle essentiellement de l'indépendance des diverses sources d'aléa). Nous démontrons facilement l'égalité $\vartheta(P) = \theta(\mathbb{P})$. La preuve se développe ainsi (*cf* Tableau noir 8). »

Tableau noir 8: *Démonstration de l'égalité $\vartheta(P) = \theta(\mathbb{P})$ dans le cadre du scénario développé dans le Tableau noir 1, où l'hypothèse de randomisation est satisfaite.*

rappels :

(i) $P(Y|A = a, W = \omega)$: valeur moyenne de Y quand nous savons que $A = a$ et $W = \omega$
ici $Y \in \{0, 1\}$ donc : valeur moyenne et probabilité que $Y = 1$ coïncident

(ii) $\vartheta(P) = P\big\{P(Y|A = 1, W) - P(Y|A = 0, W)\big\}$

(iii) $\theta(\mathbb{P}) = \mathbb{P}\{Y_1\} - \mathbb{P}\{Y_0\}$

(iv) $P\big\{P(V|U)\big\} = P\{V\}$ pour tout couple (U, V) de variables aléatoires

nous pouvons écrire :

$$\begin{aligned} \vartheta(P) &= \mathbb{P}\{\mathbb{P}(Y_A|A = 1, W) - \mathbb{P}(Y_A|A = 0, W)\} \quad \text{par cohérence} \\ &= \mathbb{P}\{\mathbb{P}(Y_1|A = 1, W) - \mathbb{P}(Y_0|A = 0, W)\} \quad \text{par conditionnement} \end{aligned}$$

or $A = U_A$ indépendant de (Y_0, Y_1, W) donc : $\mathbb{P}(Y_a|A, W) = \mathbb{P}(Y_a|W)$ pour tout $a \in \{0, 1\}$

par conséquent : $\vartheta(P) = \mathbb{P}\big\{\mathbb{P}(Y_1|W) - \mathbb{P}(Y_0|W)\big\} = \theta(\mathbb{P})$

ID: « Que se passe-t-il si tu prends bien en compte W mais que, malgré tout, l'hypothèse de randomisation n'est pas satisfaite. En d'autres termes, que se passe-t-il si W ne contient pas tous les facteurs de confusion ? »

JCT: « Si l'hypothèse de randomisation n'est pas satisfaite alors *a priori* les excès de risque naïf et généralisé diffèrent tous deux de l'excès de risque causal. »

ID: « Dans ce cas, l'excès de risque généralisé présente-t-il un quelconque avantage sur sa contrepartie naïve ? »

AC: « De mon point de vue, l'excès de risque généralisé est préférable au sens où, contrairement à son concurrent, il intègre de la connaissance relative au phénomène d'intérêt *via* la prise en compte de tous les facteurs de confusion identifiés et observés, et qu'en ce sens il se rapproche autant que faire se peut de l'excès de risque causal étant donné nos connaissances et la nature de nos observations. »

JCT: « Peux-tu étayer mathématiquement ton affirmation sur la base des écarts à l'excès de risque causal ? »

AC: « Au prix d'hypothèses intestables sur ce que j'ai envie d'appeler "la loi causale" $\mathbb{P}$, dans lesquelles j'intégrerais que W contient effectivement certains facteurs de confusion, je le pourrais. Mais honnêtement, je pourrais tout aussi bien en formuler d'autres conduisant à la supériorité de l'excès de risque naïf. Il n'existe pas d'argument purement mathématique établissant la supériorité de l'un ou de l'autre. »

ID: « A propos d'hypothèses intestables, sommes-nous même en mesure de tester l'hypothèse de randomisation ? »

JCT: « Eh bien non, c'est là que le bât blesse : l'hypothèse de randomisation est par nature intestable sur des données. Nous pouvons au mieux rassembler un faisceau convergent d'indices de sa plausibilité, par exemple en invoquant les critères de A.B. Hill et les postulats de Koch, jamais la vérifier. »

10 Du paradoxe de Simpson

ID: « Prenons un peu de recul, voulez-vous ? Où nous trouvons-nous maintenant ? »

JCT: « Nous avons discuté différentes notions liées à celle de cause et qui ont été utilisées pour essayer de la définir. Nous avons en chemin formalisé mathématiquement un *vademecum* pour la quantification mathématique de questions causales dans un cadre probabiliste avec comme élément central la notion d'intervention. »

AC: « Ce que je trouve remarquable c'est la façon dont un problème causal aboutit à ce que j'identifie comme un problème statistique à part entière et comme émancipé de son origine, c'est-à-dire digne d'intérêt au-delà de la question causale qui l'a suscité ! »

ID: « C'est tout l'intérêt d'une approche scientifique des problèmes du monde réel. »

JCT: « Sans oublier qu'en chemin nous avons formulé des hypothèses, dont l'hypothèse de randomisation, qui permettent de passer de l'un à l'autre. C'est ce que nous appelons résoudre la question de l'identifiabilité. Dans notre formalisme, ces conditions permettent de garantir que $\vartheta(P) = \theta(\mathbb{P})$, où $\theta(\mathbb{P})$ est la quantification par un paramètre dit causal de la question d'intérêt et $\vartheta(P)$ est sa contrepartie statistique. »

ID: « Ici, $\mathbb{P}$ est ce qu'Antoine appelle la "loi causale" et P est la loi qui régit le système naturel. La ruse de randomisation dans le cadre par exemple d'un essai clinique a pour vocation de faire coïncider, idéalement, P et $\mathbb{P}$, tandis que les hypothèses de randomisation, quoiqu'intestables en pratique, garantissent que $\vartheta(P) = \theta(\mathbb{P})$. »

AC: « Ainsi donc, la question statistique de l'inférence du paramètre d'intérêt, $\vartheta(P)$, sur la base d'observations effectuées "sous P", s'impose enfin ! »

JCT: « Commençons par ce fameux exemple numérique appelé "paradoxe de Simpson" [26]. Ecoutez plutôt : je propose que nous nous placions de nouveau dans le cadre du système naturel du Tableau noir 1, pour une covariable W à valeurs dans $\{0, 1\}$ (décrivant le sexe), une variable d'exposition $A \in \{0, 1\}$ (codant pour l'exposition à un facteur de risque ou non), et un critère de jugement $Y \in \{0, 1\}$ (codant pour l'occurrence d'un événement délétère ou non). L'observation d'une population d'individus régis par ce système naturels aboutit à un jeu de données consistant en $O_1, \ldots, O_n$ où chaque observation $O_i = (W_i, A_i, Y_i)$ est une copie de $O = (W, A, Y)$ au sens où elle en suit la loi P, et supposons enfin que ces copies sont mutuellement indépendantes. »

ID: « N'y a-t-il pas contradiction entre le fait que les O_i sont toutes des copies de O et, simultanément, qu'elles sont mutuellement indépendantes ? »

AC: « Non. Ce que caractérisent ces deux propriétés c'est la loi jointe du jeu de données $(O_1, \ldots, O_n)$: la génération aléatoire de chaque O_i est régie par P, c'est-à-dire la loi de la variable générique O, et la réalisation de n'importe quel sous-groupe $(O_i : i \in I)$ n'apporte aucune information relative à celle du sous-groupe complémentaire $(O_i : i \notin I)$ — tout comme les comportements des photons envoyés successivement sur la lame semi-réfléchissante, la traverser ou y rebondir, ne dépendent des comportements ni passés ni futurs. »

JCT: « Résumer l'ensemble des données est élémentaire : en vertu de l'hypothèse d'indépendance, nul besoin de conserver l'information d'ordre des observations ; il nous suffit donc de compter combien d'individus émargent dans chacune des $2 \times 2 \times 2 = 8$ classes possibles. A titre d'illustration, imaginons que ce résumé exhaustif aboutit aux tableaux suivants (*cf* Tableau noir 9). Ainsi par exemple, parmi les $n = 80$ individus observés, une moitié sont caractérisés par $W = 1$ et, parmi ceux-ci, 8 individus exposés ($A = 1$) n'ont pas développé l'effet délétère ($Y = 0$). Le tableau le plus à droite est l'agrégation des deux tableaux situés à sa gauche ; nous y perdons l'information relative à W. Gardons-le présent à l'esprit pour le rôle qu'il va jouer dans la présentation du paradoxe. »

AC: « Voici ce que nous appelons des "tableaux de contingence". En son temps, K. Pearson y a vu la quintessence de la description numérique du réel. Le statisticien lit dans ces tableaux la version empirique de P que lui offrent les observations. Souvent notée P_n, le n en indice faisant référence à la taille du jeu de données, celle-ci est une approximation de la loi inconnue P élaborée sur la base de l'observation répétée n fois de la loi P vue comme

Tableau noir 9: *Illustration numérique du paradoxe de Simpson dans le cadre du scénario développé dans le Tableau noir 1 :* $\vartheta^{\emptyset}(P_n) = -\vartheta(P_n) = 1/10$*, les deux paramètres ne peuvent donc pas prétendre tous deux à une interprétation causale de l'effet de* A *sur* Y*.*

	$W=0$		$W=1$			$W \in \{0,1\}$	
	$Y=1$	$Y=0$	$Y=1$	$Y=0$		$Y=1$	$Y=0$
$A=1$	18	12	2	8	$\Rightarrow$	20	20
$A=0$	7	3	9	21		16	24

d'une part :
- $P_n(Y=1|A=1) = \frac{20}{20+20} = \frac{1}{2}$
- $P_n(Y=1|A=0) = \frac{16}{16+24} = \frac{2}{5}$
- d'où :

$$\vartheta^{\emptyset}(P_n) = \frac{1}{2} - \frac{2}{5} = \frac{1}{10}$$

d'autre part :
- $P_n(Y=1|A=1,W=1) - P_n(Y=1|A=0,W=1) = \frac{2}{2+8} - \frac{9}{9+21} = \frac{1}{5} - \frac{3}{10} = -\frac{1}{10}$
- $P_n(Y=1|A=1,W=0) - P_n(Y=1|A=0,W=0) = \frac{18}{18+12} - \frac{7}{7+3} = \frac{3}{5} - \frac{7}{10} = -\frac{1}{10}$
- $P_n(W=1) = P_n(W=0) = \frac{1}{2}$
- d'où :

$$\vartheta(P_n) = \frac{1}{2}\left(-\frac{1}{10} - \frac{1}{10}\right) = -\frac{1}{10}$$

mécanisme de génération de la variable générique. »

ID: « Qu'est-ce que le statisticien peut tirer d'une telle source d'information ? Comme l'exercice est ici purement rhétorique, nous pouvons postuler d'emblée que c'est en termes d'excès de risque que nous quantifions la question d'intérêt. Eh bien, je vous écoute ! »

AC: « Pour reprendre l'exemple qu'a pris Jean-Christophe il y a quelques instants, si la probabilité $P(Y = 0|A = 1, W = 1)$ nous est inconnue, sa contrepartie empirique $P_n(Y = 0|A = 1, W = 1)$ est égale au quotient $8/(8 + 2) = 4/5$: sur les $8 + 2 = 10$ individus relevant de la classe $W = 1$, 8 n'ont pas développé l'effet délétère. »

JCT: « Nous allons en fait introduire *deux* paramètres d'excès de risque : l'excès de risque généralisé, caractérisé par $\vartheta(P) = P\{P(Y|A = 1, W) - P(Y|A = 0, W)\}$, prend en compte la covariable W quand l'excès de risque naïf, caractérisé par $\vartheta^{\emptyset}(P) = P(Y|A = 1) - P(Y|A = 0)$, la néglige. »

AC: « En vertu du principe de substitution, $\vartheta^{\emptyset}(P_n)$ et $\vartheta(P_n)$ sont deux estimateurs de $\vartheta^{\emptyset}(P)$ et $\vartheta(P)$. »

ID: « Qu'entends-tu par "principe de substitution" ? »

JCT: « Il s'agit du principe qui suggère que, si nous disposons d'un estimateur de la loi P_0, disons P_n^0, alors il est naturel de considérer comme candidats estimateurs de $\vartheta^\emptyset(P_0)$ et $\vartheta(P_0)$ les estimateurs $\vartheta^\emptyset(P_n^0)$ et $\vartheta(P_n^0)$ obtenus en substituant P_n^0 à P_0. Dans l'exemple d'Antoine, P_n^0 est simplement la mesure empirique P_n elle-même. »

AC: « Par substitution, nous obtenons ainsi les estimations ponctuelles $\vartheta^\emptyset(P_n) = 1/10$ et $\vartheta(P_n) = -1/10$ (*cf* Tableau noir 9). La théorie de l'inférence statistique nous enseigne que ces deux estimateurs sont optimaux, au sens où nous ne pouvons pas construire d'estimateurs plus précis quand le nombre d'observations n tend vers l'infini. Ainsi les intervalles de confiance que ces estimateurs engendrent *via* le théorème de la limite centrale, intervalles dont la vocation est de contenir les vraies valeurs inconnues avec une certitude arbitrairement élevée, sont aussi étroits que possible, quand le nombre d'observations n tend vers l'infini. »

ID: « Tu parles d'un nombre d'observations n tendant vers l'infini ! Que pouvons-nous dire quand $n = 80$ comme ici ? »

JCT: « Nous pourrions contruire des intervalles de confiance ne reposant pas sur un passage à la limite en n, et donc en particulier pas sur un théorème de la limite centrale. Admettons pour simplifier que les unités en vigueur dans le Tableau noir 9 sont des dizaines de milliers d'individus, et que par conséquent nos estimateurs ponctuels et les intervalles de confiance associés sont très précis. Isabelle, que t'inspirent-ils ? »

ID: « Je m'étonne alors de ce que $\vartheta^\emptyset(P_n)$ et $\vartheta(P_n)$ soient si différents l'un de l'autre. Ce qui est sûr, c'est que les deux ne peuvent pas simultanément prétendre accéder à une interprétation causale dans le monde réel ! Car sinon, d'une part l'exposition au facteur de risque de toute la population serait la cause d'une augmentation de 10% de la proportion de la population développant l'effet délétère relativement à l'absence d'exposition de toute la population, et d'autre part elle conduirait simultanément à une diminution de 10% de la proportion chez les femmes, chez les hommes et indépendamment du sexe. J'y perdrais presque mon latin... »

AC: « Et pourtant l'explication est d'une simplicité enfantine : $\vartheta^\emptyset$ et ϑ sont deux fonctionnelles distinctes et il n'y a *a priori* aucune raison pour que les paramètres coïncident. »

ID: « Tu réponds en mathématicien, et me laisses désemparée. Certes, tu résous le paradoxe, mais qu'en conclure concrètement ? ! »

JCT: « Au fond, tu nous mets en garde quant à l'importance et la délicatesse du choix de la quantification. »

AC: « Le procédé pourrait être mis en abyme : nous pourrions très bien décomposer

chacun des tableaux correspondant aux deux strates de W en deux sous-tableaux sur la base d'une autre covariable W' de telle sorte que sur les quatre sous-strates résultantes nous ayons $P_n(Y|A=1,W,W') - P_n(Y|A=0,W,W') = 1/10$ et donc que globalement $P_n\{P_n(Y|A=1,W,W') - P_n(Y|A=0,W,W')\} = 1/10$. »

JCT: « Permettez que nous envisagions de nouveau l'exemple numérique sous l'angle de la quantification de liaisons causales. Dans le scénario du Tableau noir 1, c'est $\vartheta(P)$ qui fait sens causalement, puisque sa définition est articulée autour du contrôle de la confusion induite par le facteur de confusion W. Notez que, dans ce système, l'hypothèse de randomisation est satisfaite. »

ID: « Mais nous pourrions tout aussi bien supposer que ce même jeu de données provient en fait d'un système naturel dans lequel W est un effet joint de A et de Y, comme résumé ici (*cf* Tableau noir 10). Dans ce second scénario, ce serait $\vartheta^{\emptyset}$ la bonne quantification de l'effet de A sur Y, tandis que ϑ ne serait qu'une quantification distordue du fait de la prise en compte de W comme un facteur de confusion quand il n'en est pas un. Formellement, l'hypothèse de randomisation n'est pas satisfaite dans ce second système. »

Tableau noir 10: *Modélisation de la façon dont la nature produit la variable aléatoire $O = (W, A, Y)$ sans intervention (gauche) et sous l'intervention $A = a$ (droite), à comparer à celle développée dans le Tableau noir 1. Ici, $\vartheta^{\emptyset}$ accède à une interprétation causale, pas ϑ ;* a contrario*, c'est ϑ et non $\vartheta^{\emptyset}$ qui accède à une interprétation causale dans le Tableau noir 1.*

$\exists\, f_A, f_Y, f_W$ fonctions déterministes,
$\exists\, U_A, U_Y, U_W$ sources d'aléa indépendantes telles que

"système naturel"	"système contrôlé"
$\begin{cases} A = f_A(U_A) \in \{0,1\} \\ Y = f_Y(A, U_Y) \\ W = f_W(A, Y, U_W) \\ O = (A, Y, W) \end{cases}$	$\begin{cases} A = a \\ Y_a = f_Y(A = a, U_Y) \\ W_a = f_W(A = a, Y_a, U_W) \\ O_a = (A = a, Y_a, W_a) \end{cases}$
la façon dont la nature produit O	la façon dont la nature produit O_a quand nous lui imposons que $A = a$, pour $a \in \{0,1\}$

JCT: « Je dirais que nous avons mis en lumière trois points fort importants ! *Primo*, que l'intuition est trompeuse : la quantification naïve de la dépendance entre A et Y, c'est-à-dire celle négligeant W, soit encore $\vartheta^{\emptyset}(P) = P(Y|A=1) - P(Y|A=0)$, ne nous enseigne rien *a priori* quant aux quantifications naïves restreintes à des strates de W, c'est-à-dire aux $P(Y|A=1,W) - P(Y|A=0,W)$; nous nous attendrions plutôt à ce qu'elle en soit une forme de moyenne, ce que l'exemple démonte. *Deuxio*, les méprises qui peuvent

découler de cette trompeuse intuition sont considérables. *Tertio*, il est impossible de faire l'économie d'une réflexion approfondie sur la nature du phénomène d'intérêt en amont de la détermination du paramètre statistique d'intérêt. »

ID: « Il y a là matière à énoncer une règle pratique. Une discordance des estimateurs obtenus sur des tableaux emboîtés est une mise en garde sur la nature de la relation étudiée. »

11 Du rasoir d'Ockham

ID: « Je souhaite que nous revenions à la question de l'inférence que nous avons abordée en Section 10 et dont le paradoxe de Simpson nous a éloignés. Avions-nous fait le tour de la question si rapidement ? »

JCT: « Je propose que nous nous concentrions sur l'exemple, édifiant, de l'inférence de l'excès de risque généralisé, caractérisé par $\vartheta(P) = P\{P(Y|A=1,W) - P(Y|A=0,W)\}$ pour un critère de jugement $Y \in \{0,1\}$. L'estimateur que nous en avons introduit lors de la discussion du paradoxe de Simpson, $\vartheta(P_n)$, est un excellent tremplin. »

ID: « Quels sont ses atouts ? »

AC: « En premier lieu, c'est un estimateur de substitution. »

ID: « En quoi cela est-il avantageux ? »

JCT: « D'abord, c'est un estimateur naturel, c'est-à-dire que sa candidature s'impose d'elle-même au statisticien. Ensuite, il est, dans ce cas précis, d'une grande facilité d'accès. Mais son atout principal est sans doute qu'un estimateur de substitution satisfait automatiquement toutes les contraintes auxquelles le paramètre est soumis. »

ID: « Mais encore ? Quelles sont ces contraintes dans le cas de l'excès de risque ? »

JCT: « Eh bien que $\vartheta(P_0) \in [-1,1]$. Des méthodes inférentielles plus sophistiquées peuvent éventuellement requérir une étape finale de mise sous contraintes de l'estimateur intermédiaire pour en faire un estimateur final admissible, satisfaisant toutes les contraintes. L'estimateur de substitution échappe à cette nécessité. »

AC: « Le deuxième atout de cet estimateur est qu'il est consistant. L'expression recouvre un certain nombre de situations. Heuristiquement, cela signifie que l'écart, en un certain sens, entre la vraie valeur et son estimation tend, en un certain sens, vers zéro quand le nombre d'observations tend vers l'infini. »

ID: « Nous en sommes à deux atouts. L'optimalité de $\vartheta(P_n)$ en constitue-t-elle un troisième ? »

JCT: « Mais absolument ! »

ID: « Ces trois atouts font-ils de $\vartheta(P_n)$ un estimateur insurpassable ? »

AC: « Je répondrais "non" par principe, car nul estimateur n'est le meilleur universellement. Nous pourrions imaginer des chausses-trappes conçues tout exprès pour l'handicaper ! C'est néanmois un excellent estimateur difficile à surpasser dans les conditions de notre discussion du paradoxe de Simpson (*cf* Section 10). »

ID: « Qu'entends-tu par là ? »

AC: « Je fais référence à ceci que W ne prend qu'un petit nombre de valeurs différentes. »

ID: « Avons-nous fait le tour de la question ? »

JCT: « Loin s'en faut ! Pour s'en convaincre, il suffit de remarquer que notre élaboration de l'estimateur $\vartheta(P_n)$ repose de façon essentielle sur la finitude du nombre de valeurs que peut prendre W. Ainsi, si W peut prendre un nombre infini de valeurs alors le procédé s'écroule. »

ID: « Je conçois en effet que si W prend ne serait-ce qu'un très grand nombre de valeurs (cela suffit à mon argument) alors il serait déraisonnable de chercher à considérer simultanément l'ensemble des sous-tableaux de contingence correspondant à l'ensemble des valeurs que peut prendre W, ce que vous appeliez plus tôt des strates. »

AC: « Si nous essayions néanmoins, un grand nombre de ces tableaux seraient creux, c'est-à-dire présenteraient un ou plusieurs effectifs nuls. »

ID: « Cela me fait aussi penser au rasoir d'Ockham [11]. Un estimateur $\vartheta(P_n)$ construit comme une moyenne pondérée d'estimateurs restreints à des strates n'est pas économe en présence d'un grand nombre de strates, or

> Pluralitas non est ponenda sine necessitate,

les multiples ne doivent pas être utilisés sans nécessité. Nous trouvons-nous dans un cul-de-sac ? ! »

12 Inférence ciblée : initialisation

JCT: « La solution que le statisticien envisage naturellement consiste à isoler la question de l'estimation de la loi conditionnelle de Y sachant (A, W) et à la traiter comme un problème intermédiaire. »

ID: « C'est un peu abscons ! Que signifie "estimer une loi" ? »

JCT: « Puisque $Y \in \{0,1\}$, sa loi conditionnelle sachant (A,W) est une loi de Bernoulli et, par conséquent, la connaître est équivalent à connaître la probabilité $P_0(Y=1|A,W)$. Ainsi, estimer cette loi revient à estimer la fonction $(A,W) \mapsto P_0(Y=1|A,W)$; cela s'appelle "régresser Y sur (A,W)". Notons $P_n^0(Y=1|A,W)$ l'estimateur de $P_0(Y=1|A,W)$ que nous construisons. »

AC: « L'indice supérieur "0" suggère que ceci est une étape d'initialisation… »

ID: « Et quel est le lien entre l'estimation de cette loi conditionnelle $P_0(Y|A,W)$ avec l'estimation de $\vartheta(P_0)$? »

JCT: « La réponse risque de se révéler un peu ardue ! »

AC: « Par définition, si je pose $\Delta_0(W) = P_0(Y=1|A=1,W) - P_0(Y=1|A=0,W)$ alors $\vartheta(P_0) = P_0\{\Delta_0(W)\}$ est la moyenne (en W) de la variable aléatoire $\Delta_0(W)$. Or, $\Delta_n^0(W) = P_n^0(Y=1|A=1,W) - P_n^0(Y=1|A=0,W)$ apparaît naturellement comme un estimateur de $\Delta_0(W)$. Il suffit donc, pour être en mesure d'en déduire un estimateur de $\vartheta(P_0)$, d'estimer la moyenne $P_0\{\Delta_n(W)\}$, ce qui nécessite d'estimer la loi marginale de W. Plus concrètement, je préconise ici d'estimer simplement la loi marginale de W par sa version empirique. »

ID: « Et plus concrètement encore ? ! »

JCT: « Cela signifie simplement que nous estimons la loi marginale de W par la loi qui donne à chaque valeur observée W_i de W une probabilité $1/n$ d'être émise. »

AC: « J'aime me figurer cela en termes de simulation. Ecoutez plutôt. Nous observons la nature lorsqu'elle émet des réalisations indépendantes de W notées $W_1, \ldots, W_i, \ldots, W_n$. Estimer la loi marginale de W est équivalent à construire un algorithme qui émet lui aussi des réalisations de W sous une loi qui a vocation à imiter la nature. Lorsque nous estimons la loi marginale de W par sa version empirique, cet algorithme émet uniformément chacune des observations réalisées, c'est-à-dire qu'elle émet chaque W_i avec probabilité $1/n$. »

ID: « Résumons s'il vous plaît. Nous estimons $\Delta_0(W)$ par $\Delta_n^0(W)$ et la loi marginale de W par sa version empirique, notre objectif étant d'estimer $\vartheta(P_0) = P_0\{\Delta_0(W)\}$. Si je ne m'abuse, l'estimateur résultant s'écrit $P_n\{\Delta_n^0(W)\} = n^{-1}\sum_{i=1}^n \Delta_n^0(W_i)$. »

AC: « Tu as parfaitement raison, et ta présentation concise met bien en avant le fait que cet estimateur initial est un estimateur de substitution ! »

JCT: « Cette présentation a aussi le grand mérite de permettre aisément de faire le lien avec le cas où W ne prend qu'un petit nombre de valeurs. En effet, si tel est le cas, alors nous choisissons avantageusement $P_n^0(Y=1|a,w) = P_n(Y=1|a,w)$, la probabilité empirique d'observer $Y=1$ dans la ligne "$A=a$" du sous-tableau correspondant à la

strate "$W = w$". Et, ô surprise, $P_n\{\Delta_n^0(W)\} = \vartheta(P_n)$, l'estimateur de substitution de la mesure empirique. »

ID: « L'estimateur $P_n\{\Delta_n^0(W)\}$ est une alternative à $\vartheta(P_n)$ pour le cas où W ne prend pas un petit nombre de valeurs, et il étend l'estimateur de substitution optimal qu'est $\vartheta(P_n)$ lorsque W prend un petit nombre de valeurs. Hérite-t-il à ce titre des atouts de celui-ci ? ! »

AC: « Pas nécessairement malheureusement, et ce pour une raison facile à comprendre : on conçoit que ce qui fait de P_n^0 un bon estimateur de P_0 ne fait pas nécessairement de $\vartheta(P_n^0)$ un bon estimateur de $\vartheta(P_0)$. »

ID: « Je pense comprendre la raison que tu donnes. Pour m'en assurer, j'aimerais que vous m'aidiez à tisser une métaphore. Rembrandt convoque séparément deux des apprentis de son atelier, tous deux d'égal talent. Il dit au premier "Apprends à peindre à ma manière." et au second "Apprends à peindre des mains à ma manière." Après quelques semaines, les deux apprentis se présentent devant Rembrandt et lui annoncent qu'ils ont terminé leur apprentissage. Rembrandt leur demande alors de peindre une main à sa manière. La main peinte par le second apprenti est plus convaincante que celle peinte par le premier. »

AC: « Rembrandt a en effet désavantagé le premier apprenti en ne lui disant pas en amont que ce sont les mains qui l'intéressaient. »

JCT: « J'aime ta métaphore, sur laquelle je m'appuie pour bâtir une analogie. Le style de Rembrandt, c'est-à-dire sa capacité à peindre une toile, est comme la loi P_0, un objet de grande complexité. Apprendre le style de Rembrandt peut être envisagé comme estimer P_0 sur la base de l'observation des toiles du maître, d'où l'acquisition par le premier apprenti d'un style P_n^0 approchant P_0 dans toute sa complexité. Peindre à la manière de Rembrandt consiste alors, pour le premier apprenti, à produire une toile sous la loi P_n^0. »

AC: « Dans le même esprit, et un peu malicieusement, bien qu'étant un apprenti de piètre talent, je saurais peindre à la manière de Rembrandt sous la mesure empirique P_n ! Il me suffirait pour cela de tirer au hasard l'une de ses toiles et de la lui présenter telle quelle. »

JCT: « Le style de Rembrandt limité à la représentation de mains, qui n'est qu'une fraction de son style, est d'une complexité bien moindre. Je le vois comme un trait $\vartheta(P_0)$ du style P_0. Lorsque le premier apprenti peint des mains, il le fait sous $\vartheta(P_n^0)$. »

AC: « Cependant, le premier apprenti, sachant maintenant que ce sont des mains qu'il doit réaliser pour Rembrandt, s'attelle de nouveau à la tâche. Il adapte son style initial, P_n^0, en un style ciblé vers la production de mains, que je note P_n^1. Eh bien, nous pouvons concevoir que les mains qu'il réalise dès lors sous $\vartheta(P_n^1)$ surpassent celles qu'ils réalisaient sous $\vartheta(P_n^0)$, voire celles que peint le second apprenti, parce que lui s'est imprégné de tout le style du maître. »

13 Inférence ciblée : ciblage

ID: « J'en conclus, en revenant à l'excès de risque, que vous savez adapter la loi P_n^0 en une loi P_n^1 qui, en ciblant $\vartheta(P_0)$, fait de l'estimateur de substitution $\vartheta(P_n^1)$ un bon estimateur de $\vartheta(P_0)$. »

AC: « Absolument ! C'est que la fonctionnelle ϑ jouit d'une importante propriété : elle est différentiable sur les chemins. »

ID: « Différentiable comme nous dirions "dérivable" pour une fonction définie sur un ensemble de nombres réels ? »

AC: « Oui, mais il faut étendre la notion dans la mesure où ϑ est une fonction définie non pas sur un ensemble de nombres réels mais sur l'ensemble de lois $\mathcal{M}$. Pour ce faire, on considère les restrictions de ϑ à des "chemins" dans $\mathcal{M}$, et puisque chaque point d'un tel chemin est identifié univoquement par un nombre réel, tout comme un point d'une route est repéré par la distance qui le sépare du début de la route, l'étude de la restriction de ϑ au chemin relève de l'étude de fonctions définies sur un ensemble de nombre réels. »

ID: « Et comment faites-vous cela ? »

AC: « Formellement, ϑ est différentiable si pour tout $P \in \mathcal{M}$, il existe une "direction" $\nabla\vartheta(P)$ telle que, quel que soit le chemin $\{P(\varepsilon) : \varepsilon \in [-1, 1]\} \subset \mathcal{M}$ passant par $P = P(0)$ et de direction s en $P = P(0)$, la fonction $\varepsilon \mapsto \vartheta(P(\varepsilon))$ est dérivable en $\varepsilon = 0$, avec une dérivée égale à $P\{\nabla\vartheta(P)(O) \times s(O)\}$. »

JCT: « Un chemin $\{P(\varepsilon) : \varepsilon \in [-1, 1]\} \subset \mathcal{M}$ n'est pas un objet exotique ! Ce n'est rien d'autre qu'un modèle paramétrique, paramétré par le seul $\varepsilon \in [-1, 1]$, et donc de dimension un. »

ID: « Est-il approprié de penser à la direction $\nabla\vartheta(P)$ comme à la dérivée de ϑ en $P \in \mathcal{M}$? »

AC: « Oui en effet ! »

ID: « Et quel chemin empruntez-vous alors, si vous me passez ce jeu de mots, pour exploiter cette propriété afin d'adapter P_n^0 ? ! »

AC: « Justement ! Nous construisons un chemin qui passe par P_n^0 en empruntant la direction vers laquelle pointe $\nabla\vartheta(P_n^0)$. »

ID: « La loi adaptée P_n^1 se situe-t-elle sur ce chemin ? »

AC: « En effet. Nous cherchons notre loi adaptée P_n^1 sur ce chemin. Ainsi, identifier P_n^1 revient à identifier le meilleur paramètre $\varepsilon = \varepsilon_n^0$ et à poser $P_n^1 = P_n^0(\varepsilon_n^0)$. »

ID: « Et en quel sens ε peut-il être le meilleur ? »

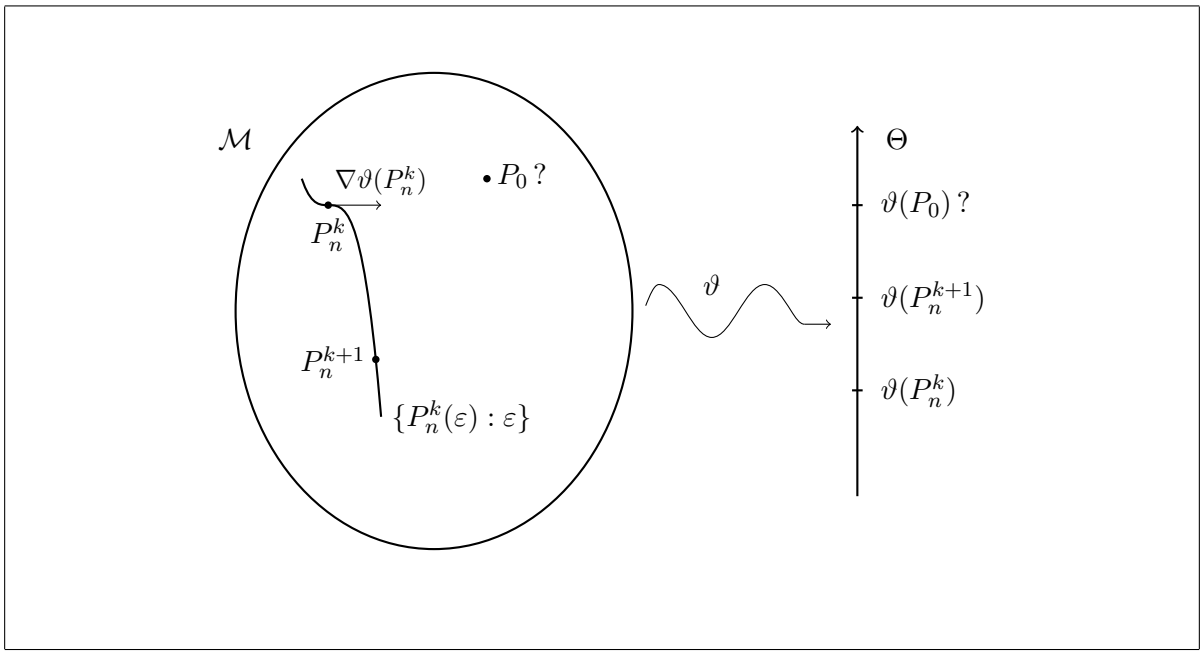
AC: « Si le chemin pointe dans la direction $\nabla\vartheta(P_n^0)$ au sens de la vraisemblance (d'autres choix sont possibles), alors ε_n^0 est la valeur du paramètre qui maximise la vraisemblance sur le chemin. »

ID: « Suis-je en droit d'en déduire que la méthode d'inférence ciblée est en quelque sorte une extension du principe d'inférence par maximisation de la vraisemblance ? »

AC: « C'est une bonne idée, et la filiation est flatteuse, tant le principe du maximum de vraisemblance joue un rôle important en statistique. »

ID: « Je continue sur ma lancée. Ce P_n^1 se voit associer l'estimateur de substitution $\vartheta(P_n^1)$: est-il notre estimateur final de $\vartheta(P_0)$? »

Tableau noir 11: *Illustration du principe de l'inférence ciblée.*



JCT: « Dans la version que nous te présentons, ce n'est pas le cas. Il faut itérer la procédure (*cf* Tableau noir 11). A l'étape $k \geq 1$, nous déterminons la direction $\nabla\vartheta(P_n^k)$; nous construisons le chemin $\{P_n^k(\varepsilon) : \varepsilon \in [-1, 1]\} \subset \mathcal{M}$ qui pointe vers cette direction ; nous trouvons le meilleur paramètre d'adaptation, ε_n^k ; puis nous posons enfin $P_n^{k+1} = P_n^k(\varepsilon_n^k)$. »

ID: « Ne s'arrête-t-on donc jamais ? ! »

JCT: « Nous disposons d'un certain nombre de critères d'arrêt qui nous informent de l'inutilité d'une éventuelle nouvelle adaptation. En notant P_n^* la dernière adaptation récursive

de P_n^0, nous disposons enfin de l'estimateur de substitution $\vartheta(P_n^*)$, appelé "estimateur de maximum de vraisemblance ciblée" de $\vartheta(P_0)$. »

AC: « Te voilà initiée au principe général de la procédure d'inférence ciblée que M. van der Laan et D. Rubin ont élaborée en 2006 [29]. Elle a été depuis largement développée et appliquée à une variété de problèmes statistiques [28]. »

ID: « Votre présentation de la procédure d'inférence ciblée me rappelle la méthode de Newton pour la recherche d'une racine d'une équation. Comme elle, la procédure est fondée sur une initialisation suivie d'une série de mises à jour dans la direction de la dérivée au point courant. »

AC: « Ta comparaison est très pertinente, même si la procédure d'inférence ciblée est une émanation de la théorie des "fonctions estimantes" [27, 30] caractérisée par la volonté de produire des estimateurs de substitution. Par ailleurs, c'est en pensant à la méthode de Newton que L. Le Cam a élaboré sa méthode d'estimation dite "à un pas" [21]. »

ID: « Les mêmes ressorts sont-ils à l'œuvre dans les méthodes d'estimation ciblée et à un pas ? »

AC: « C'est le même canevas, mais la méthode d'estimation à un pas procède aux mises à jour en travaillant directement sur l'estimateur dans l'espace des paramètres Θ, une unique fois, alors que la méthode d'estimation ciblée travaille sur les lois dans l'espace $\mathcal{M}$, éventuellement de façon itérée, les mises à jour dans $\mathcal{M}$ induisant par substitution des mises à jour dans Θ. »

14 Inférence ciblée : mérites

ID: « Fort bien, mais à quoi bon tout cela ? Quelles sont les bonnes propriétés de $\vartheta(P_n^*)$? »

AC: « Le statisticien attend *a minima* un résultat de consistance et un théorème de la limite centrale, pour construire des intervalles de confiance, sous des hypothèses aussi faibles que possible. »

JCT: « Concernant la consistance, il s'avère, sans grande surprise, que $\vartheta(P_n^*)$ converge vers $\vartheta(P_0)$ quand $\Delta_n^*(W) = P_n^*(Y = 1|A = 1, W) - P_n^*(Y = 1|A = 0, W)$ est un estimateur consistant de $\Delta_0(W)$. »

ID: « Je le conçois, dans la mesure où $\vartheta(P_n^*) = P_n^*\{\Delta_n^*(W)\}$ et où j'ai confiance dans l'estimation par P_n^* de la loi marginale de W sous P_0, qui n'est pas une tâche difficile. Quelle surprise me réserves-tu ? Je vois tes yeux briller ! »

JCT: « C'est que $\vartheta(P_n^*)$ converge aussi vers $\vartheta(P_0)$ quand $P_n^*(A = 1|W)$ est un estimateur

consistant de $P_0(A = 1|W)$. »

ID: « Cela me surprend, car enfin la loi conditionnelle de A sachant W n'apparaît pas dans la définition de ϑ, qui ne met en jeu que la loi conditionnelle de Y sachant (A, W) et la loi marginale de W... »

AC: « Cette propriété remarquable est appelée "double-robustesse". Elle n'est pas si surprenante que cela lorsque nous remarquons que $\vartheta(P)$ s'écrit de façon équivalente $\vartheta(P) = P\{AY/P(A = 1|W) - (1 - A)Y/P(A = 0|W)\}$! »

JCT: « Quant au théorème de la limite centrale, l'estimateur $\vartheta(P_n^*)$ en satisfait un sous un ensemble de conditions qui s'expriment notamment en termes de vitesse de convergence de $\Delta_n^*(W)$ et de $P_n^*(A = 1|W)$ vers leurs limites respectives. Heuristiquement, il faut que l'une de ces limites au moins coïncide avec $\Delta_0(W)$ ou $P_0(A = 1|W)$ et que le produit des vitesses de convergence soit en $\sqrt{n}$. »

AC: « Tu oublies de préciser que si d'aventure il y a coïncidence pour chacune des limites, alors $\vartheta(P_n^*)$ est efficace : il a la plus petite variance asymptotique possible, et donc les intervalles de confiance correspondants sont aussi étroits que possible. Sous des hypothèses moins contraignantes, nous savons construire un estimateur conservatif de la variance asymptotique de $\vartheta(P_n^*)$, c'est-à-dire un estimateur qui va sur-estimer la vraie variance limite, conduisant de ce fait à des intervalles de confiance certes un peu trop étendus mais valides. »

ID: « Vous évoquez des hypothèses qui vous sont favorables. Avons-nous les moyens de les vérifier, ou pour le moins de bonnes raisons de penser qu'elles sont satisfaites ? ! »

JCT: « Aïe ! Tu nous places face à un dilemme. Nous saurions déterminer un sous-ensemble $\mathcal{M}_0 \subset \mathcal{M}$ de lois tel que, si $P_0 \in \mathcal{M}_0$ alors nos hypothèses seraient satisfaites moyennant la constructionde P_n^0 sur la base de procédures statistiques adéquates. Mais nous ne savons pas si P_0 est bien un élément de $\mathcal{M}_0$ ou pas... »

AC: « Autre point de vue : si je précisais la nature des procédures statistiques qui régissent la construction initiale de P_n^0... »

JCT: « ... alors je saurais déterminer malicieusement une loi P_0 telle que les hypothèses ne soient pas satisfaites si P_0 s'avérait être la loi de la nature. »

ID: « A quelle aune jugeons-nous alors vos hypothèses ? »

AC: « A la taille de l'ensemble $\mathcal{M}_0$ qu'évoquait Jean-Christophe plus tôt : plus il est grand, moins les hypothèses sont contraignantes et plus convaincant est le résultat. »

ID: « Excusez ma question peut-être naïve, mais que pouvez-vous dire dans les cas où vos hypothèses ne sont pas satisfaites ? »

JCT: « Loin d'être candide, ta question est redoutable. Je tends à penser que ce sont les études de simulation qui nous permettent d'explorer ce territoire hostile où nos hypothèses ne sont pas satisfaites. »

ID: « Quand tu simules, ne simules-tu pas ? ! Autrement dit, déguises-tu un acte sous l'apparence d'un autre ou reproduis-tu artificiellement une situation réelle à des fins de démonstration ou d'explication ? ! »

JCT: « Le second bien sûr. L'étude de simulation consiste à construire une loi synthétique $P_0 \in \mathcal{M}$ dont nous maîtrisons tous les traits et qui a vocation à imitier la nature. En particulier, nous connaissons la valeur de $\vartheta(P_0)$. Nous pouvons aussi tirer sous P_0 des jeux de données virtuels de toute taille n. »

AC: « Le premier mérite d'une telle étude est qu'elle permet de vérifier que l'implémentation de la méthode d'inférence est correctement menée. Son deuxième mérite est illustratif, puisque nous pouvons constater que lorsque les hypothèses sont satisfaites alors l'estimateur jouit des propriétés attendues. »

JCT: « Enfin, pour répondre à ta question, elle permet de se faire une idée de ce qui se passe dans ces cas où les hypothèses ne sont pas satisfaites. »

ID: « Comme par exemple ? »

JCT: « Tous les résultats que nous évoquions sont asymptotiques. Ainsi, l'étude de simulation jette un éclairage sur le comportement de l'estimateur à horizon fini, c'est-à-dire pour des valeurs de n éventuellement petites. »

AC: « Ou bien, l'étude de simulation nous permet de mieux comprendre ce qui se passe lorsque les hypothèses ne sont que légèrement violées. Tout cela nous ouvre des horizons théoriques, computationnels et applicatifs passionnants. »

15 Epilogue

ID: « Avez-vous vu l'heure ? ! Il est grand temps je crois de laisser là notre discussion. »

AC: « Sans tirer de conclusion ? ! »

JCT: « Je propose de citer Lucrèce [22]

> Déterminer exactement celle de ces causes qui agit dans notre monde est chose difficile ; mais indiquer ce qui est possible, voilà ce que j'enseigne ; et je m'attache à exposer tour à tour les multiples causes qui peuvent être à l'origine du mouvement des astres : entre toutes, il ne peut y en avoir qu'une qui fasse mouvoir nos étoiles : mais laquelle ? L'enseigner n'est pas donné à notre science, qui n'avance que pas à pas.

Cette conclusion vous convient-elle ? »

ID: « Il est certes difficile de satisfaire notre curiosité, mais elle est un moteur formidable, tout comme l'est le partage des résultats que nous obtenons. »

AC: « Je suis bien d'accord. A cet égard, nous pouvons tirer de notre longue conversation un *vademecum* pour entreprendre de satisfaire notre curiosité avec rigueur. »

JCT: « Oui, et il faut toujours aller de l'avant sans perdre de vue les contributions passées. Voilà ce me semble une jolie conclusion. »

ID: « Adieu mes amis, bonsoir et bonne nuit. »

AC: « Vous plaisantez ; mais vous rêverez sur votre oreiller à cet entretien, et il continuera de prendre de la consistance. »

ID: « Eh bien curieuse je me serai couchée, curieuse je me lèverai. »

Glossaire

Bernoulli (loi de). La variable aléatoire A suit la loi de Bernoulli de paramètre $p \in [0,1]$ si A prend les valeurs 0 et 1 uniquement, de telle sorte que $A = 1$ avec probabilité p (et, donc, $A = 0$ avec probabilité $1 - p$).

Conditionnellement à. Voir "loi conditionnelle" et "indépendance conditionnelle".

Confiance (intervalle de). Un intervalle de confiance pour $\vartheta(P)$ de niveau $(1-\alpha) \in [0,1]$ est un intervalle aléatoire construit à partir de n observations tirées sous une loi P ayant vocation à contenir $\vartheta(P)$ avec probabilité supérieure ou égale à $(1-\alpha)$. A niveau $(1-\alpha)$ fixé, *(i)* le meilleur de deux intervalles de confiance est le plus étroit, et *(ii)* un intervalle de confiance est d'autant plus étroit que sa construction s'appuie sur un plus grand nombre d'observations. Plus le niveau est élevé, plus l'intervalle est étendu. Il est souvent intéressant de construire un intervalle de confiance en utilisant un estimateur ϑ_n de $\vartheta(P)$ comme pivot, c'est-à-dire sous la forme $[\vartheta_n - c_n, \vartheta_n + c_n]$ pour une demi-longueur c_n, éventuellement aléatoire, bien choisie.

Confusion. Une relation entre deux variables est soumise à confusion dès lors que la dépendance probabiliste entre celles-ci, éventuellement rapportée à des strates caractérisées par une troisième variable, ne peut pas être interprétée causalement.

Consistant (estimateur). La consistance est une notion asymptotique : un estimateur ϑ_n de $\vartheta(P)$ est consistant s'il converge en un certain sens vers $\vartheta(P)$ quand le nombre d'observations n sur lequel sa construction s'appuie tend vers l'infini. La consistance est dite faible si, pour toute erreur $\varepsilon > 0$ fixée, la probabilité que ϑ_n s'écarte d'au moins ε de $\vartheta(P)$ tend vers 0 quand n tend vers l'infini : $\lim_{n\to\infty} P\{|\vartheta_n - \vartheta(P)| \geq \varepsilon\} = 0$. La consistance est dite forte si ϑ_n converge vers $\vartheta(P)$ presque sûrement : $P\{\lim_{n\to\infty} |\vartheta_n - \vartheta(P)| = 0\} = 1$. Un estimateur fortement consistant est faiblement consistant, et la réciproque est fausse.

Contingence (tableau de). Un tableau de contingence, terme inventé par K. Pearson en 1904, est une table à deux (ou plus) entrées dans laquelle nous reportons les effectifs croisés associés à deux (ou plus) variables catégorielles d'intérêt. L'origine des tableaux de contingence remonte aux travaux menés par P.C.A. Louis pour démontrer l'inefficacité thérapeutique de la saignée.

Exemple. Soit $O_1, \ldots, O_n$ $n = 50$ variables telles que chaque O_i contient un couple $(A_i, Y_i) \in \{0,1\}^2$. Le tableau de contingence

	$Y=1$	$Y=0$
$A=1$	18	12
$A=0$	7	13

nous enseigne que sur ces n observations, 18 (respectivement, 12, 7 et 13) présentent un couple (A_i, Y_i) égal à $(1,1)$ (respectivement, $(1,0)$, $(0,1)$ et $(0,0)$).

Corrélation (coefficient de). Le coefficient de corrélation entre deux variables réelles quantifie leur dépendance probabiliste. Si X et Y sont indépendantes alors leur coefficient de corrélation est nul. La réciproque est fausse.

Empirique (mesure). Etant donné des observations $O_1, \ldots, O_n$, la mesure empirique est la loi P_n telle que, si $O \sim P_n$, alors $O = O_i$ avec probabilité n^{-1} pour tout $1 \leq i \leq n$.

Estimateur. Un estimateur est une variable aléatoire obtenue en combinant les observations issues d'une expérience afin d'en estimer un trait d'intérêt.

Exemple. Soit $O_1, \ldots, O_n$ indépendantes de loi commune P. La moyenne empirique $n^{-1} \sum_{i=1}^n O_i$ est un estimateur de la moyenne $\vartheta(P) = P\{O\}$ de $O \sim P$. Si O est à valeurs réelles et si $P\{|O|\}$ est finie alors la moyenne empirique est un estimateur fortement consistant (en vertu de la loi forte des grands nombres).

Gaussienne (loi). La variable aléatoire réelle O suit la loi gaussienne standard $N(0,1)$ si pour tout $a \leq b$, la probabilité que $O \in [a,b]$ égale l'aire sous la courbe de la courbe de Gauss d'équation $t \mapsto \sqrt{2\pi}^{-1} \exp(-t^2/2)$. Cette loi est particulièrement importante car elle apparaît naturellement comme loi limite de suites d'expériences dans des théorèmes dits de la limite centrale.

Grands nombres (loi des). La loi des grands nombres est un théorème probabiliste qui offre des hypothèses sous lesquelles la moyenne empirique $n^{-1} \sum_{i=1}^n O_i$ de n variables aléatoires $O_1, \ldots, O_n$ de loi commune P converge vers leur moyenne $P\{O\}$.

Nous qualifions de "faible" cette loi si la convergence a lieu en probabilité, c'est-à-dire si, quelle que soit la marge d'erreur préalablement fixée, la probabilité que l'écart séparant la moyenne empirique de la moyenne excède cette marge d'erreur tend vers 0 quand n tend vers l'infini. C'est J. Bernoulli qui le premier a formalisé cette loi, en 1690. La loi faible des grands nombres est garantie notamment lorsque que les O_i sont indépendantes, à valeurs réelles et telles que $P\{|O|\}$ soit finie.

Nous la qualifions de "forte" si la convergence a lieu presque sûrement, c'est-à-dire si la probabilité qu'elle n'ait pas lieu tend vers 0 quand n tend vers l'infini. Si la loi forte s'applique alors la loi faible s'applique nécessairement. A. Kolmogorov a prouvé en 1929 que la loi forte des grands nombres est garantie notamment lorsque que les O_i sont indépendantes, à valeurs réelles et telles que $P\{|O|\}$ soit finie.

Pour A. Kolmogorov,

> La valeur épistémologique de la théorie des probabilités est fondée sur le fait que les phénomènes aléatoires engendrent à grande échelle une régularité stricte, où l'aléatoire a, d'une certaine façon, disparu.

Indépendance. Soit $\{O_i : i \in I\}$ une collection de variables aléatoires indexées par I. Soit $I_1, I_2 \subset I$. Nous disons que $\mathcal{O}_1 = \{O_i : i \in I_1\}$ est indépendante de $\mathcal{O}_2 = \{O_i : i \in I_2\}$ si les valeurs des réalisations du premier ensemble ne sont pas affectées par les valeurs des réalisations du second ensemble. Plus formellement, $\mathcal{O}_1$ est indépendante de $\mathcal{O}_2$ si la loi jointe de $(\mathcal{O}_1, \mathcal{O}_2)$ est le produit des lois marginales de $\mathcal{O}_1$ et $\mathcal{O}_2$, ou encore si la loi conditionnelle de $\mathcal{O}_2$ sachant $\mathcal{O}_1$ coïncide avec la loi marginale de $\mathcal{O}_2$, et réciproquement.

Exemple. Soit $(W,Y) \in \{0,1\}^2$ telle que $P(W = Y = 1) = 1/10$, $P(W = 1, Y =$

$0) = 1/15$, $P(W = 0, Y = 1) = 1/2$ et $P(W = Y = 0) = 1/3$. La loi marginale de Y est la loi de Bernoulli de paramètre $P(Y = 1) = P(Y = 1$ et $(W = 1$ ou $W = 0)) = P(W = Y = 1) + P(W = 0, Y = 1) = 3/5$. La loi marginale de W est la loi Bernoulli de paramètre $P(W = 1) = P(W = 1$ et $(Y = 1$ ou $Y = 0)) = P(W = Y = 1) + P(W = 1, Y = 0) = 1/6$. Nous observons que $P(W = 1)P(Y = 1) = 1/10 = P(W = Y = 1)$, $P(W = 1)P(Y = 0) = 1/15 = P(W = 1, Y = 0)$, $P(W = 0)P(Y = 1) = 3/5 = P(W = 0, Y = 1)$ et $P(W = 0)P(Y = 0) = 1/3 = P(W = Y = 0)$. Ainsi, W et Y sont bien indépendantes sous P.

Indépendance conditionnelle. Soit $\{O_i : i \in I\}$ une collection de variables aléatoires indexées par I. Soit $I_1, I_2, I_3 \subset I$. Nous disons que $\mathcal{O}_1 = \{O_i : i \in I_1\}$ est conditionnellement indépendante de $\mathcal{O}_2 = \{O_i : i \in I_2\}$ sachant $\mathcal{O}_3 = \{O_i : i \in I_3\}$ si la loi jointe conditionnelle de $(\mathcal{O}_1, \mathcal{O}_2)$ sachant $\mathcal{O}_3$ est le produit des deux lois conditionnelles de $\mathcal{O}_1$ et $\mathcal{O}_2$ sachant $\mathcal{O}_3$. Si $I_3 = \emptyset$ se réduit au vide, de telle sorte que $\mathcal{O}_3 = \emptyset$, alors l'indépendance conditionnelle coïncide avec l'indépendance, ce qui est faux en général.

Inférence. L'inférence statistique est un ensemble de procédures mathématiques développées dans le cadre de la théorie de la statistique, qui s'appuie notamment sur celle des probabilités, afin d'analyser la structure d'une expérience aléatoire sur la base de son observation, typiquement en termes d'estimation ponctuelle ou par intervalles de confiance, de test d'hypothèses, ou de régression.

Limite centrale (théorème de). Les divers théorèmes de la limite centrale présentent des conditions sous lesquelles des sommes de nombreuses variables aléatoires suivent des lois approximativement gaussiennes. Typiquement, si $O_1, \ldots, O_n$ sont indépendantes telles que $P\{O_i\} = 0$ pour tout $i \leq n$ et $\sum_{i=1}^{n} P\{O_i^2\} = 1$, et si aucune des O_i ne contribue trop largement à la somme, alors $\sum_{i=1}^{n} O_i$ suit approximativement la loi gaussienne standard $N(0, 1)$.

Loi. La loi P d'une variable aléatoire O est la description exhaustive de la façon dont le hasard produit une réalisation de O. On note $O \sim P$ pour indiquer que O suit la loi P.

Loi conditionnelle. Soit $O \sim P$ une variable aléatoire qui peut se décomposer en deux parties, disons $O = (W, Y)$. La loi conditionnelle de Y sachant W est la loi de la variable aléatoire Y lorsque la réalisation de W est connue.
Exemple. $W \in [0, 1]$ et Y suit la loi de Bernoulli de paramètre $1/3$ si $W \leq 1/2$ et $3/5$ si $W > 1/2$.

Loi jointe. Soit O une variable aléatoire qui peut se décomposer en deux parties, disons $O = (W, Y)$. La loi jointe de O est la loi du couple.

Loi marginale. Soit $O \sim P$ une variable aléatoire qui peut se décomposer en deux parties, disons $O = (W, Y)$. La loi marginale de Y est la loi de la variable aléatoire Y lorsqu'elle est extraite de O. L'origine de cette expression est à trouver dans les tableaux de contingence.
Exemple. Soit $(W, Y) \in \{0, 1\}^2$ telle que $P(W = Y = 1) = 1/10$, $P(W = 1, Y =$

$0) = 1/5$, $P(W = 0, Y = 1) = 3/10$ et $P(W = Y = 0) = 2/5$. Alors $P(W = 1) = P(W = 1 \text{ et } (Y = 1 \text{ ou } Y = 0)) = P(W = Y = 1) + P(W = 1, Y = 0) = 3/10$, donc la loi marginale de W est la loi de Bernoulli de paramètre $3/10$. De façon similaire, $P(Y = 1) = P(Y = 1 \text{ et } (W = 1 \text{ ou } W = 0)) = P(W = Y = 1) + P(W = 0, Y = 1) = 2/5$, donc la loi marginale de Y est la loi Bernoulli de paramètre $2/5$.

Modèle. Un modèle est un ensemble de lois susceptibles d'être la loi d'une observation O. Un modèle est dit paramétrique si ses éléments sont identifiés par un paramètre de dimension finie.

Exemple. Soit $\mathcal{M}$ le modèle non-paramétrique de toutes les lois compatibles avec la définition de l'observation O. Un sous-ensemble $\{P(\varepsilon) : \varepsilon \in [-1, 1]\} \subset \mathcal{M}$ de lois candidates $P(\varepsilon)$ identifiées par le paramètre réel ε constitue un modèle paramétrique de dimension un, par conséquent appelé aussi "chemin".

Paramètre. Valeur d'une fonctionnelle définie sur un modèle et évaluée en une loi lui appartenant.

Exemple. $\theta(\mathbb{P})$ pour $\theta : \mathbb{M} \to \Theta$ ou $\vartheta(P)$ pour $\vartheta : \mathcal{M} \to \Theta$.

Régression. Etant donné des observations $O_1, \ldots, O_n$ d'une structure $O = (W, Y)$, régresser Y sur W consiste à inférer des observations des informations sur la façon dont Y dépend de W. Typiquement, régresser Y sur W s'entend au sens de chercher à expliquer la valeur moyenne de la variable aléatoire Y conditionnellement à W, c'est-à-dire en fonction de W.

Exemple. Si $Y \in \{0, 1\}$ alors régresser Y sur W revient à estimer ce que vaut la probabilité conditionnelle $P(Y = 1|W)$ que Y prenne la valeur 1 sachant la valeur que prend W. Cet exemple relève du sens typique cité plus tôt dans la mesure où $P(Y = 1|W)$ coïncide avec la valeur moyenne $P(Y|W)$ de Y sachant W.

Substitution (estimateur de). Etant donné une fonctionnelle d'intérêt $\vartheta : \mathcal{M} \to \Theta$, un estimateur ϑ_n du paramètre $\vartheta(P)$ est dit de substitution s'il s'écrit $\vartheta_n = \vartheta(P_n)$ pour P_n loi approchant P.

Exemple. Soit $\vartheta : \mathcal{M} \to \mathbb{R}$ telle que $\vartheta(P) = P\{O\}$ pour $\mathcal{M}$ un ensemble de lois qui admettent toutes une moyenne. Soit $O_1, \ldots, O_n$ indépendantes de loi commune P et soit P_n la mesure empirique. La moyenne empirique $n^{-1} \sum_{i=1}^{n} O_i = \vartheta(P_n)$ est un estimateur de substitution de la moyenne $\vartheta(P)$.

Sûrement (presque). Un événement est presque sûr si sa probabilité égale 1.

Exemple. Si $O \sim N(0, 1)$ alors $O \neq 0$ presque sûrement. Pourtant, par symétrie, c'est autour de 0 que se concentre la masse de la loi gaussienne $N(0, 1)$: pour toute longueur d'intervalle $\ell > 0$ fixée, l'intervalle qui a le plus de chance de contenir O est celui centré en 0, soit $[-\ell/2, \ell/2]$.

Exemple. L'estimateur ϑ_n de $\vartheta(P)$ est fortement consistant si l'événement $\lim_{n\to\infty} |\vartheta_n - \vartheta(P)| = 0$ est presque sûr pour P.

Trait. Voir "paramètre".

Uniforme (loi). La variable aléatoire réelle O suit la loi uniforme sur l'intervalle $[A, B]$ si pour tout $A \leq a \leq b \leq B$, la probabilité que $O \in [a, b]$ égale le rapport $(b-a)/(B-A)$.

Variable aléatoire. Description, éventuellement partielle, du résultat d'une expérience aléatoire, c'est-à-dire d'une expérience reproductible soumise à un aléa.
Exemple. L'expérience consistant à tirer à pile ou face une pièce équilibrée dans des conditions contrôlées est une expérience aléatoire (elle est reproductible et nous n'avons pas la certitude de ce que donnera le lancer). Le résultat de chaque lancer est décrit par la variable aléatoire prenant la valeur 1 si nous obtenons un pile et la valeur 0 sinon. La loi de cette variable aléatoire est la loi de Bernoulli de paramètre 1/2.

Vraisemblance. La vraisemblance d'une observation sous une loi susceptible de la produire quantifie l'adéquation entre l'observation et la loi. Plus l'adéquation est grande, plus la vraisemblance l'est. Le principe du maximum de vraisemblance s'inspire de cette interprétation : de deux lois d'égale complexité susceptibles de produire une observation dont nous disposons, il faut préférer celle maximisant la vraisemblance. Si les deux lois ne sont pas de même complexité, alors la comparaison des deux vraisemblances requiert un ajustement préalable fondé sur un principe de parcimonie, la loi la plus complexe bénéficiant au départ d'un avantage sur sa rivale.

Références

[1] A. Aspect. Bell's inequality test: more ideal than ever. *Nature*, 398:189–190, 1999.

[2] J. S. Bell. On the Einstein-Podolsky-Rosen paradox. *Physics*, 1, 1964.

[3] C. Bernard. *Introduction à l'étude de la médecine expérimentale*. Collection Classiques. Champs-Flammarion, 1865.

[4] G. Camerino and P. Goodfellow. A fragile understanding. *Trends Genet.*, 7(8):239–240, Aug 1991.

[5] G. Camerino, P. Parma, O. Radi, and S. Valentini. Sex determination and sex reversal. *Curr. Opin. Genet. Dev.*, 16(3):289–292, Jun 2006.

[6] N. Cartwright. The art of medicine. A philosopher's view of the long road from RCTs to effectiveness. *The Lancet*, 377, April 23 2011.

[7] A. I. Conan Doyle. *The sign of the four*. 1890.

[8] J. le R. d'Alembert. *Opuscules mathématiques*. 1761–1780. Tome 2, Onzième mémoire, Sur l'application du calcul des probabilités à l'inoculation de la petite vérole, p. 34.

[9] Académie des sciences. *Rapport de l'Académie des sciences*. Séance du 5 octobre 1835, p. 173.

[10] D. Diderot. *Entretien entre d'Alembert et Diderot*. 1769.

[11] G. d'Ockham. *Quaestiones et decisiones in quatuor libros Sententiarum cum centilogio theologico*. 1319. Livre 2.

[12] A. Einstein, B. Podolsky, and N. Rosen. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.*, 47(10):777–780, 1935.

[13] D. N. Fredericks and D. A. Relman. Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clin. Microbiol. Rev.*, 9(1):18–33, 1996.

[14] P. N. Goodfellow and G. Camerino. DAX-1, an 'antitestis' gene. *Cell. Mol. Life Sci.*, 55(6-7):857–863, Jun 1999.

[15] P. N. Goodfellow and G. Camerino. DAX-1, an "antitestis" gene. *EXS*, (91):57–69, 2001.

[16] A. B. Hill. The environment and disease: association or causation? *Proc. R. Soc. Med.*, 58:295–300, 1965.

[17] D. Hume. *Traité de la nature humaine*.

[18] M. Kistler. The interventionist account of causation and non-causal association laws. *Erkenntnis (submitted)*, 2012.

[19] R. Koch. Die aetiologie der Tuberculose. *Mitt. Kaiser. Gesundh.*, 2(1):1–88, 1884.

[20] R. Koch. Ueber bakteriologische Forschung. In *Verh. X. Int. Med. Congr. Berlin, 1890*, page 35, 1892.

[21] L. Le Cam and G. L. Yang. *Asymptotics in statistics*. Springer-Verlag, 2000.

[22] Lucrèce. *De rerum natura*.

[23] Platon. *Le Timée*.

[24] K. J. Rothman. Causes. *Am. J. Epidemiol.*, 104(6):587–592, 1976.

[25] K. J. Rothman and S. Greenland. *Modern Epidemiology*. second edition. Lippincott-Raven, Philadelphia, 1998.

[26] E. H. Simpson. The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B.*, 13:238–241, 1951.

[27] M. J. van der Laan and J. M. Robins. *Unified methods for censored longitudinal data and causality*. Springer-Verlag, 2003.

[28] M. J. van der Laan and S. Rose. *Targeted learning*. Springer, 2011.

[29] M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *Int. J. Biostat.*, 2:Art. 11, 40, 2006.

[30] A. W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 1998.

[31] Virgile. *Les Géorgiques*. Second livre.