Review

Antoine Chambaz*, Isabelle Drouet and Jean-Christophe Thalabard

Causality, a Trialogue

Abstract: A philosopher, a medical doctor, and a statistician talk about causality. They discuss the relationships between causality, chance, and statistics, resorting to examples from medicine to develop their arguments. This debate gives rise to an original trialogue, a tribute to the famous conversation between d'Alembert and Diderot, two great French thinkers of the Enlightenment. The trialogue notably offers an introduction to the philosophy of causality and an initiation to statistics, including recent developments that should prove interesting to specialists and laypeople alike.

Keywords: statistics, philosophy, medicine

DOI 10.1515/jci-2013-0024

Otros cien pasos serían los que anduvieron, cuando al doblar de una punta pareció descubierta y patente la misma causa, sin que pudiese ser otra, de aquel horrísono y para ellos espantable ruido, que tan suspensos y medrosos toda la noche les había tenido; y eran (si no lo has, ! oh lector! por pesadumbre y enojo) seis mazos de batán que con sus alternativos golpes aquel estruendo formaban.

They went it might be a hundred paces farther, when on turning a corner the true cause, beyond the possibility of any mistake, of that dread-sounding and to them awe-inspiring noise that had kept them all the night in such fear and perplexity, appeared plain and obvious; and it was (if, reader, thou art not disgusted and disappointed) six fulling hammers which by their alternate strokes made all the din.

M. de Cervantes, *El Ingenioso Hidalgo de Don Quijote de la Mancha*
(English translation by J. Ormsby)

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*

Notation

- $\exists$, existential quantifier, mathematical symbol meaning “there exists.”
- $\Rightarrow$, mathematical symbol meaning “implies” (or, equivalently, “if (...) then (...)”).
- $0-1$, mathematical notation meaning “0 or 1.”
- $\{0, 1\}$, set consisting of the two numbers 0 and 1.
- $[a, b]$, interval consisting of all the real numbers equal to or larger than a and equal to or smaller than b .

***Corresponding author: Antoine Chambaz**, Modal'X (EA 3454), Université Paris Ouest Nanterre, 200 av de la République, Nanterre 92001, France, E-mail: achambaz@u-paris10.fr

Isabelle Drouet, SND (FRE CNRS 3593), Université Paris-Sorbonne, Paris, France, E-mail: isabelle.drouet@paris-sorbonne.fr

Jean-Christophe Thalabard, MAP5 (UMR CNRS 8145), Université Paris Descartes & Hôpital Cochin, AP-HP, Paris, France, E-mail: jean-christophe.thalabard@mathinfo.univ-paris5.fr

- W , random vector representing baseline covariates; A, A' , random variables representing treatment or exposure; L, L' , random variables representing intermediate covariates; Y , random variable quantifying the outcome of interest, called primary endpoint; Y_a and L_a , counterparts of Y and L under control $A = a$, for instance in the counterfactual world where the equality $A = a$ is guaranteed.
- $O \sim P$ an observation, seen as a random variable, whose law is $P \in \mathcal{M}$, where $\mathcal{M}$ is a set of candidate laws, also called “model.”
- $X \sim \mathbb{P}$ a full data, seen as a random variable, whose law is $\mathbb{P} \in \mathbb{M}$, where $\mathbb{M}$ is a set of candidate laws, also called “counterfactual model”.
- $\vartheta : \mathcal{M} \rightarrow \Theta$, a functional that associates every element $P \in \mathcal{M}$ with the statistical parameter $\vartheta(P)$. Note: ϑ is the Greek letter theta in its cursive form and Θ is that same letter in upper case.
- $\theta : \mathbb{M} \rightarrow \Theta$, a functional that associates every element $\mathbb{P} \in \mathbb{M}$ with the statistical parameter $\theta(\mathbb{P})$
- $P(\mathbb{P})$, the law of O when O is modeled as the incomplete observation of the full data $X \sim \mathbb{P}$.
- $\mathbb{P}(P)$, the law of the full data X in a counterfactual model synthetically built based on the observation $O \sim P$.
- $P\{O\}$, the mean value of $O \sim P$.
- $P\{Y|W\}$, the conditional mean value of Y given W for $O = (W, Y) \sim P$. If $Y \in \{0, 1\}$ then it coincides with the conditional probability $P(Y = 1|W)$ that Y be equal to 1 given the value of W .
- P_n^0 and P_n^k , initial and k times updated estimations of the law P of $O \sim P$ based on n observations.
- $\{P(\varepsilon) : \varepsilon \in [-1, 1]\} \subset \mathcal{M}$, parametric model, also called “path” because it is one-dimensional, subset of $\mathcal{M}$, the global model $\mathcal{M}$ of candidate laws.
- $\nabla\vartheta(P)$, a function of $O \sim P$, the “derivative” at P of a pathwise differentiable functional $\vartheta : \mathcal{M} \rightarrow \Theta$.
- s , a function of O , a “direction” of a path $\{P(\varepsilon) : \varepsilon \in [-1, 1]\}$ at $P = P(0)$.

Preamble

□ A philosopher, a medical doctor, and a statistician talk about causality. They discuss the relationships between causality, chance, and statistics, resorting to examples from medicine to develop their arguments. This debate gives rise to an original trialogue, a tribute to the famous conversation between d’Alembert and Diderot, two great French thinkers of the Enlightenment. The trialogue notably offers an introduction to the philosophy of causality and an initiation to statistics including recent developments that should prove interesting to specialists and laypeople alike.

◻ The driving force of the trialogue is not conflict. It is complementarity. The three actors come with their complementary backgrounds, questions, answers, which create a fruitful dynamic. Each one of the actors plays in turn the role of an ingenuous interlocutor. Ingenuous questions are at the core of the maieutic process. In this respect, the trialogue is closer to some of Plato’s dialogues than to Diderot and d’Alembert’s dialogue.

◻ We envision randomness at a time when the erosion of determinism has resulted in a slow revolution. Random variation and change are no longer scoria obscuring an immutable reality. Instead, variation and change are inherent in nature, they are not errors but rather the phenomena themselves. Thus it is legitimate, if not inescapable, for a scientific approach to reality to place variation and change at the core of the representation, not at its periphery. Although causality and determinism have long been intrinsically linked, the erosion of determinism did not imply an erosion of causality: in fact, the two notions have been separated.

◻ Systems of structural equations play a central role in the trialogue. Taking the form of a deterministic structure, represented by deterministic functions, powered by an indeterministic, random engine, they pertain both to a regularist and to a stochastic conception of reality. Using them does not commit one to either the Laplacian or the more stochastic view. Moreover, they enable one to define an apprehensible notion of causal effect revolving around that of intervention. In this context, the more general concept of counterfactual events is easily introduced.

☒ From this follows a counterfactual model for reality, where what happens in the real world is the projection of what concomitantly occurs in parallel, counterfactual worlds. Used for the sake of approaching causality, this induces a conceptual difficulty: there may be something more, or something else, in effects and causes than what actually takes place in our world. Or not, since a puzzling twist allows the statistician to build counterfactual worlds inducing by projection the actual world based on the observation of the sole actual world.

☒ The aforementioned projection is driven by possibly intertwined exposures that may not follow from interventions. From a statistical viewpoint, one may have access to experimental or observational data. This leads to a discussion about experimenting *versus* observing, confounding, the randomization trick and the randomization hypothesis.

☒☒ Throughout the trialogue, the idea will emerge that it is not necessary to adopt a clear-cut philosophical stance on the issue of determinism or on the nature of causality to tackle causal questions fruitfully. Moreover, the theory of statistics now provides the researcher with a fundamental *vademecum* for a sound statistical analysis of causal questions. The scientific question of interest may often be translated into a finite-dimensional feature of a possibly infinite-dimensional law representing how nature produces the data. The definition of the parameter and the choice of a model are decoupled, thus leaving room for the honest construction of a model including real knowledge, and nothing more, and the use of a commonsense parameter. Decoupling is also at play in the two-step elaboration of the model, conceived first as a statistical model, then extended to a causal one, at the price of possibly untestable assumptions. In the extended model, the parameter can be interpreted causally. The interpretation may collapse if the assumptions are wrong. However, the statistical parameter always makes sense. Finally, the concepts of consistency, valid confidence intervals, and efficiency, though abstract at first sight, deeply impact the characterization of the inference procedure that the statistician tailors and carries out to target the unknown truth. Embodying this philosophy, the targeted maximum likelihood estimation (TMLE) procedure is sketched and discussed.

1 A quantum Lucretius

AC: Isabelle, would you say that thinking in causal terms is universal?

ID: There is indubitably a cultural dimension to causality, which is extremely prevalent in the Western world, but not as much in other cultures. I would therefore be tempted to say – especially if I need to answer shortly – that thinking in causal terms is *not* universal.

JCT: Oh, but we have all our time! And it seems to me that Antoine's question, even restricted to Western thinking, is relevant, subtle, and long-standing. Did not Virgil write [1], at the dawn of the first millennium of our Western times:

Felix, qui potuit rerum cognoscere causas,

happy is he who gets to know the reasons for things?

ID: Yes, indeed, Jean-Christophe, and he was probably referring to Lucretius' work *De Rerum Natura* [2], *On the Nature of Things*, a poem that describes the world according to Epicurean principles. And we can go further back in time, when Plato has Timaeus say [3]:

But everything that is necessarily has a cause; as nothing that was created can have been created without a cause.

Timaeus, just as Virgil later on, refers to the creation of the universe, indeed at huge time and space scales, but starting from the primitive bodies that the atoms are.

AC: Does chance play a role in this poetic description of the world?

ID: I am afraid I must respond with another question: what do you mean by “chance”?!

AC: I know that the French word for “chance” is “hasard”, which is derived from the Arabic term “*al-zahr*”, meaning dice; I also know that S. Mallarmé tells us that a roll of the dice will never abolish chance [4], and that for Heraclitus, the fairest universe is but a heap of rubbish piled up at random!...

JCT: As far as I am concerned, I know that the French word “chance,” which means “good fortune” in English, comes from the Latin “*cadentia*,” falling things, a term which Cicero used to refer to jacks; I know that our French word “*aléatoire*”, “random” in English, is also derived from Latin, and more specifically from “*alea*,” game of dice, and “*aleatorius*,” regarding games of chance.

ID: There is a major conceptual gap between the dice and the jacks on the one hand, and Heraclitus’ heap of sweepings on the other. How does the shift operate, from the simplicity of rolling dice or throwing jacks, to the complexity of reality?

JCT: I can tell you how to get real numbers from “randomly obtained” 0–1 sequences.

ID: I am all ears.

JCT: You can for instance use a Galton board, such as the one you can admire at the Galerie de la Découverte in Paris. Sir Galton conceived the board so that he could visualize the random diffusion of balls of radius r . The device consists of a vertical board featuring nails in staggered rows, evenly spaced horizontally at a $r + \varepsilon$ distance, for $\varepsilon > 0$ small, one row being vertically spaced $r/2$ from the next. Balls are dropped from the top of the board, in the middle; they find a way down among the nails, and form piles of varying heights at the bottom of the board. The $r/2$ staggered rows warrant that, at each row of nails, the balls are as likely to bounce left as they are to bounce right. In addition, the width $r + \varepsilon$ and the spacing between the rows of nails warrant that past bouncing has virtually no influence on future bouncing.

AC: Let us follow the trajectory of one ball. If we number each one of the lower bins and note 0 or 1 according to where the ball bounces, left or right of each nail it gets across as it falls, then the number of the bin in which the ball ends up its course is indeed obtained from this sequence of 0–1.

ID: The number is random because the sequence of 0–1 is random too! I understand... How are the balls distributed when a large number is dropped successively? Does any specific pattern emerge?

JCT: When there is a large number of balls, one can empirically observe that the outer bins receive few balls whereas the biggest piles form in the central bins. The ball piles form a regular bell curve.

ID: Let us take a more global view, shall we? You have just explained how randomly drawing elements from a finite set can boil down to randomly drawing a sequence of 0–1. I can imagine that, if the board is huge, you can randomly obtain decimal numbers with great precision. Are there other ways to proceed?

JCT: This is a very interesting question. If we go to the limit, *i.e.* if we take an infinitely large board, the limit law we obtain is called a Gaussian law. This is one among an infinity of ways to randomly obtain numbers.

AC: The classic Bolzano–Weierstrass device would enable us to draw numbers from another law. Imagine that, from the same sequence of 0–1 as previously, and starting from the interval $[0, 1]$ of all the numbers comprised between 0 and 1, I successively divide the current interval in its middle and I choose its left half for a 0 and its right half for a 1. With 1,024 0–1, we can determine a random number in $[0, 1]$ with a precision of 308 digits after the decimal point.

JCT: With an infinitely large board, the limit law we obtain is called a uniform law on $[0, 1]$. It means that the probability to fall in an interval $[a, a + \ell]$ of size ℓ is dependent on ℓ , not on a .

ID: Very well. You have explained to me how to randomly draw a number following a Gaussian law or a uniform law. Your constructions, all in all, are nothing but series of randomly drawn 0–1...

AC: ... and to boot, these successive draws are independent, *i.e.* they are such that past values have no influence on future values, and they are equiprobable too, *i.e.* such that 0 and 1 are equally likely to be drawn! We say that a random variable with an equiprobable chance of taking the values 0 or 1 obeys Bernoulli's law of parameter $1/2$ or, equivalently, that it is a Bernoulli variable with parameter $1/2$.

ID: May I repeat my question? Is this what chance is for you?

JCT: Well, yes, although it might be counter-intuitive. Indeed, most of the random variables are elaborated from random variables drawn from Bernoulli's law of parameter $1/2$ and from the uniform law on $[0, 1]$, all independent.

ID: All right. But I doubt that the most rigorous way to generate random numbers involves boards spiked with nails!

AC: Indeed, nowadays, one of the ways to do so relies on the emission of photons on semi-transparent mirrors.

JCT: The laws of quantum mechanics indeed tell us that photons are offered two equiprobable choices: either crossing, or bouncing back. The same laws also tell us that the choices of successive photons are mutually independent.

ID: On second thoughts, and although it first seemed familiar to me, I believe the notion of two equiprobable choices had better be explained. Could you please do this for me?

AC: Certainly! This explanation, probabilistic in nature, is quite technical as it relies on the notion of limit. Stating that crossing or bouncing back are two equiprobable events for the photon comes back to saying that I am almost sure that, no matter how small the precision $\varepsilon > 0$, there is an integer n_0 that depends on ε such that, if I independently emit $n \geq n_0$ photons on a semi-transparent mirror, then the fraction n_t/n of photons crossing it gets distant from $1/2$ of at most ε . It is an example of the law of large numbers.

ID: You are only *almost* sure?!

AC: It is indeed the standard phrase! It means that if we repeat independently N times the experiment consisting in (i) arbitrarily choosing ε , (ii) emitting $n \geq n_0$ photons on the semi-transparent mirror, (iii) evaluating the fraction of photons crossing it, and (iv) evaluating the deviation of the fraction from $1/2$, then it will lead N times to the same conclusion, *i.e.* that the deviation is no bigger than ε , no matter how large N is.

JCT: I am not sure you have convinced Isabelle! Why not say, rather, that an event is almost sure to happen when its probability equals 1? *A contrario*, an event of probability 0 cannot be observed, but is nonetheless not impossible.

ID: Everything is clear now.

AC: The latest technologies thus permit to generate about 16 million of independent and equiprobable 0–1 per second. From which we can obtain 15,625 random numbers per second independently and uniformly on $[0, 1]$, with a precision of 308 digits after the decimal point.

ID: Then the circle is complete, and we may go back to your initial question, which started it.

AC: My question?... Oh, yes! Does chance play a role in Lucretius' poetic description of the world? And, since the question of causality seems to be involved, did Lucretius and Plato venture on the field of causality armed with the concept of chance, or did they do without it? Due to my education, and also I

believe out of inclination, it is difficult for me to envisage causality without relying on chance, at least partly.

JCT: In this respect you are like D. Diderot...

AC: This is flattering!

JCT: ... like Diderot, who argumented to J. d'Alembert [5, 6]:

(...) the cause undergoes too many particular vicissitudes which escape our observation, for us to be able to count with certainty upon the result that will ensue. Our certainty that a violent-tempered man will grow angry at an insult is not the same as our certainty that one body striking a smaller body will set it in motion.

Long before Diderot, Lucretius, along with his mentor, Epicurus, placed at the heart of Epicurean physics the notion of *clinamen*, i.e. the spontaneous deviation of atoms, not of photons, from their vertical free fall, a random variation that accounts, one thing leading to another, for the existence of bodies and human free will.

ID: Are you suggesting that the notion of randomness has been playing a prominent role in the description of reality since Antiquity? On the contrary, for a very long time, increasingly complex deterministic descriptions of the world have been elaborated. As explained by R. Starmans [7], the related concepts of variation and change had a fairly pejorative connotation. What I. Hacking calls the “erosion of determinism” [8] was a long emancipation process in which C. Darwin’s theory of evolution and the accumulation and analyses of administrative data sets were pivotal.

AC: You are talking about deterministic descriptions of the world. Does that imply that reality itself is deterministic for those people? Most importantly for us, what was their impact on causality?

ID: These are very difficult questions. What I can say is that this deterministic view profoundly influenced the way causality was envisioned. In particular, D. Hume in his *Treatise on Human Nature* [9] develops a concept of cause the repercussions of which have been formidable, and which, placing *regularity* at the heart of causality, seems to exclude randomness. We must not conclude precipitously that it is necessary to relate causality and randomness.

2 Inflamed Hume

AC: What is regularity?

ID: Constant conjunction...

JCT: Which means?!

ID: Let me read the key passage [9]:

‘Tis therefore by EXPERIENCE only, that we can infer the existence of one object from that of another. The nature of experience is this. We remember to have had frequent instances of the existence of one species of objects; and also remember, that the individuals of another species of objects have always attended them, and have existed in a regular order of contiguity and succession with regard to them. Thus we remember to have seen that species of object we call *flame* and to have felt that species of sensation we call *heat*. We likewise call to mind their constant conjunction in all past instances. Without any farther ceremony, we call the one cause and the other effect, and infer the existence of the one from that of the other. In all those instances, from which we learn the conjunction of particular causes and effects, both the causes and effects have been perceived by the senses, and are remembered. But in all cases, wherein we reason concerning them, there is only one perceived or remembered, and the other is supplied in conformity to our past experience.

Inspiring, isn’t it?!

JCT: This definition indeed leaves no room to randomness or to statistics.

AC: Why do you say this? It seems to me that, on the contrary, this passage suggests that the notion of causality is intrinsically linked to that of statistics. Isn't statistics the art of extracting information from observations, from what is *experienced*, or *experimented*?

JCT: Beware! There is only one term in French, "expérience," for both "experience" and "experiment." The terms "experienced" and "experimented" are not synonyms.

ID: And by "experience." D. Hume refers to sensitive experience. In addition, when you associate causality with statistics, you basically refute the vision of K. Pearson, one of the founding fathers of statistics [10], creator of the correlation coefficient and of the notion of regression, who, along with B. Russell [11], battled against the very notion of causality at the beginning of the twentieth century.

JCT: In the observations that you mention, as well as in the laws that govern their production, K. Pearson saw no less than reality itself reduced to its very essence.

ID: This view was, quoting R. Starmans [7], "the crown" that K. Pearson put on the emancipation process that we evoked. Concomitantly, causality was simply denied any existence [12]. But your misreading of D. Hume may be interesting. Can you elaborate further?

JCT: I would be glad to help him do this, patiently reflecting on the possibility of a relationship between causality and statistics. Let us isolate, in D. Hume's quotation, the notion of "contiguity," which I would willingly characterize as spatial and temporal. I understand there is a necessity, on the one hand, that the action of the cause and the measure of its effect apply to one single, coherent system, or experimental unit, to use the jargon of statistics; on the other hand, that the observation of the candidate cause and effect takes place on a time scale the characterization of which depends on their very nature.

AC: I agree with this interpretation. The condition of temporal succession D. Hume formulates also appears as *a priori* knowledge...

JCT: ... or as an *a priori* constraint...

AC: Right, *a priori* knowledge or *a priori* constraint imposed on the statistic models...

ID: ... or on the concepts of cause and effect!...

AC: Agreed, on the statistic model designed to shed light, from observations, on a candidate cause–effect relationship, or on the concepts of cause and effect themselves, so that an effect may not precede its cause or be simultaneous to it.

JCT: This seems natural on the human scale, but it may be debatable on a quantum scale, as suggested by the theoretical and experimental solution of the EPR paradox [13–15]. Isabelle, could you please explain to us what the third condition of constant conjunction is?

ID: Regularity, or constant conjunction, is the idea that an event we name "cause" is *always* followed with the element we name "effect". Each time you place your hand above the flame, you get burnt...

AC: ... And if you take your hand away, the flame does not burn you any longer.

ID: D. Hume does not say this, but I believe that you are right and that we can add this.

AC: Well, a cause is always relative to a situation from which it is absent.

JCT: At any rate, D. Hume's main idea is that causes are always followed with their effects. Nonetheless, it seems obvious to me that such is not always the case. Striking a match on a coarse surface causes it to catch fire. Yet, if the match is wet, or if there is no oxygen, it will *not* catch fire. How do you evade this trap?

ID: This is a very basic trap. I will tell you that what causes the match to catch fire is not merely its being struck on a coarse surface, but a whole set of conditions among which, in addition to striking it, also feature the dryness of the match, the presence of oxygen in the air – and probably other additional conditions. It is the whole set of these conditions that would *always* be followed by the match catching fire. We may say the set is “sufficient” for the match to catch fire, and then we do go back to D. Hume’s regularist condition.

AC: To summarize, striking a match on a coarse surface is a cause of the match catching fire insofar as striking the match belongs to a set of conditions which, when they are all met, are always followed by the match catching fire, and striking the match is indispensable, a necessary step.

ID: Absolutely. It is the theory of J.S. Mill [16], which dates back to the middle of nineteenth century, and then of J.L. Mackie [17], in the second half of the twentieth century: a given cause is an INUS condition, *i.e.* a condition that is not sufficient to produce the effect, but is a nonredundant part of an unnecessary but sufficient condition for the effect.

AC: What does “INUS” stand for?

ID: It is the acronym of “Insufficient but Nonredundant part of a condition that is itself Unnecessary but Sufficient” for the result.

JCT: A similar conception is conveyed in K. Rothman [18, 19]’s model, which epidemiologists have been using since the late 1970s. This model constitutes a set of guidelines devised to establish cause-to-effect relationships. As early as in the nineteenth century, physicians began to use such sets of guidelines. Thus, for instance, J. Henle and his disciple R. Koch defined four criteria capable of evidencing a causal relationship between a microbe and a disease [20, 21]. R. Koch used these criteria to characterize the etiology of tuberculosis and anthrax. Though periodically tailored to fit current advances in the field, “Koch’s postulates” are still being used in microbiology [22].

AC: In a similar vein, A.B. Hill’s criteria [23] were devised in the mid-1960s in the field of occupational epidemiology. Although they are neither necessary nor sufficient, they still contribute to structuring the causal interpretation of epidemiological studies.

JCT: Absolutely, and we owe R. Doll and A.B. Hill – with his criteria – the elucidation of the causal relationship between tobacco and lung cancer [24]. However, I would like to go back to the notion of regularity. Can we always use subterfuge, as some might say, to support a regularist definition of causality, as in the case of the match? For instance, can we consider that, despite the obvious fact that not all smokers die of lung cancer, there still exists a set of conditions, among which tobacco use, which is sufficient to cause lung cancer? Such set of assumptions certainly constitutes a solid methodological tenet. But is this tenet credible? Is it always possible, *in fine*, to reduce causality to regularity?

3 *Ceteris paribus sic standibus*

ID: Causality and determinism have been intrinsically linked for a very long time. This is why the aforementioned erosion of determinism led K. Pearson to deny the existence of causality. However, the erosion of determinism did not imply an erosion of causality; the two notions have been separated. The philosophical analyses of causality in terms of probability have been developed in the second half of the twentieth century precisely to counter the possibility of reducing causality to regularity.

AC: Along which lines?

ID: The idea that underpins these analyses is the following: C causes E if and only if C increases the probability of E , *ceteris paribus*. I.J. Good, P. Suppes, N. Cartwright, B. Skyrms, all tried to confer precise meaning to this idea [25–29].

AC: Probabilists and statisticians like D.B. Rubin, J. Robins, P.W. Holland, D.A. Freedman, J. Pearl, S. Greenland and P. Dawid, among others [30–37] also reflected on the relationship between causality and probability. Randomness is back into play! But what does your “*ceteris paribus*” mean?

ID: It is a short version of the Latin phrase *ceteris paribus sic stantibus*, which translates into “all things being equal.” In other words, C causes E if and only if the presence of C increases the probability of E compared to its absence, all other things – *ceteris paribus* – being equal.

AC: Thus, in this paradigm, in order to cause E , C does not necessarily have to be an INUS condition! And if C is an INUS condition then, in the simultaneous presence of the other elements of the sufficient condition to which C belongs, the presence of C almost surely, *i.e.* with a probability equal to 1, brings about the presence of E , whereas in its absence the presence of E is not almost sure, *i.e.* E is likely to be absent.

JCT: It seems to me that a shift in difficulty has been taking place: what does “*ceteris paribus sic stantibus*” actually mean here? Isn’t it difficult to specify these things which are such that, when kept unchanged, the presence of C increases the probability of E compared to its absence, and in this respect C causes E ?

AC: If they are to be specified, I would personally start by saying that they are characteristics, or features, of laws.

ID: Do you mean physical laws? Or more generally what philosophers label “laws of nature”?

AC: No, I mean laws in the probabilistic sense. In other words, rules characterizing the production of random variables. What are your opinions about this?

JCT: My feeling is that if we mean by “laws” what the philosophers do, then the characterization of these “things kept unchanged” is causal in nature! We are going full circle.

ID: Yes, and this is precisely why philosophers have concluded that causality cannot be reduced to probability.

AC: I would like us to go back to the phrase “*ceteris paribus sic stantibus*.”

JCT: For example: is it always possible to define *something* such that the phrase “conditionally on *something*” be equivalent to “*ceteris paribus sic stantibus*”?

AC: We could, sometimes, but not in general!

ID: What are the circumstances in which we could?

AC: Here is the most simple example that comes to my mind. We try to define the possible effect of a treatment, which I note $a = 1$, on a given disease, for instance in terms of survival, seriousness, or duration, compared to the absence of any treatment, which I note $a = 0$.

JCT: Why on earth do you choose to use the letter “ a ”?

AC: Let us say that I choose it because it is the first letter in the word “action.” The variable A testifies to the presence of the cause, when $A = 1$, or to its absence, when $A = 0$, as soon as we mean by “cause” being under treatment *versus* not taking any treatment. The effect of the treatment is expressed by the variable noted Y , which is posterior to A in time. If we introduce a little bit of formalism, it can look like this (*cf* the left-hand side of Blackboard 1).

$\exists f_W, f_A, f_Y$ deterministic functions, $\exists U_W, U_A, U_Y$ independent sources of randomness such that	
“natural system”	“controlled system”
$\begin{cases} W = f_W(U_W) \\ A = f_A(W, U_A) \in \{0, 1\} \\ Y = f_Y(W, A, U_Y) \\ O = (W, A, Y) \end{cases}$	$\begin{cases} W = f_W(U_W) \\ A = a \\ Y_a = f_Y(W, A = a, U_Y) \\ O_a = (W, A = a, Y_a) \end{cases}$
how nature produces O	how nature produces O_a when we impose that $A = a$, for $a \in \{0, 1\}$ here, <i>ceteris paribus</i> is nearly equivalent to “conditionally on W ”

Blackboard 1 Modeling how nature produces the random variable $O = (W, A, Y)$ without intervention (left) and under the intervention $A = a$ (right). Here, “*ceteris paribus sic standibus*” is nearly equivalent to “conditionally on” W .

ID: There is something I do not understand. I thought you were going to present a probabilistic model for causality in the specific example, and yet you start with a set of deterministic functions. What am I missing?

JCT: Antoine is using the so-called structural equations model. It originates in the works of S. Wright and T. Haavelmo [38, 39] and was recently brought up-to-date by J. Pearl [35]. Let me point to the fact that randomness is present in the model, through the sources of randomness U_W, U_A, U_Y . So the model is partly probabilistic indeed.

ID: But what do the sources of randomness represent? Are they only here to account for our current ignorance? In which case, a deterministic view of reality, or nature, would hide behind this partly probabilistic representation.

JCT: This is a Laplacian conception toward which J. Pearl has expressed preference. From this stance, imagine that in a distant future all the facets of a physical phenomenon are known. According to this conception, the corresponding structural equations model would be free of random inputs.

AC: To me, there is more to these sources of randomness than ignorance. The physical phenomenon itself is random, no matter how long you study it. But this is not in contradiction with the existence of a deterministic structure, represented by the deterministic functions, powered by an indeterministic, random engine.

JCT: I like the structure-engine metaphor. It is noteworthy that using the formalism of structural equations system does not commit us to either the Laplacian or the more stochastic view.

ID: Shall we go back to the description of the model? You said that Y is posterior to A . By analogy, I deduce that W chronologically precedes A , and therefore Y as well. What does this variable correspond to? »

JCT: The variable W represents pieces of information that are anterior to both the cause and its effect. These pieces of information are crucial to the determination, or rather to the *realization*, of A and Y .

ID: Let us recapitulate: A represents the nature of the cause, Y its effect... So, by process of elimination, I imagine that W corresponds to these things that are referred to by “*ceteris paribus*.” Thus, would you say, in this case, that “*ceteris paribus sic standibus*” and “conditionally on the realization of W ” are equivalent?

AC: Nearly equivalent (*cf* the right-hand side of Blackboard 1), because the potential effect of the cause on the disease is naturally expressed in terms of a comparison of Y_1 with Y_0 , *i.e.* the evolution of the disease when we *impose* treatment on the one hand, or when we *impose* the absence of treatment on the other...

JCT: And because Y_1 and Y_0 are, intrinsically, functions of the same W !

ID: I see! But why do you say “nearly”?

AC: Because in fact Y_1 and Y_0 are, intrinsically, functions of the same W and of U_Y , according to f_Y ! Thus for me, “*ceteris paribus sic standibus*” precisely refers to keeping unchanged the marginal law of W and the conditional law of Y given (A, W) . Or, in other words, to keeping unchanged f_W, f_Y , as well as the way the sources of randomness U_W and U_Y are produced. These are the features of the law I was referring to earlier.

ID: You have convinced me! And I can see at last the fundamental reason why you do not want to consider “*ceteris paribus*” and “conditionally on *something*” as similar: the first phrase refers to features of laws, whereas the second one refers to variables that these laws produce!

AC: Exactly. It is the distinction I had in mind, which also drives the elaboration of a second scenario that sheds light on the impossibility to use one phrase for the other.

ID: Could you please, first, present me a real-life situation corresponding to your second scenario?...

AC: Here is one, again from the medical field (*cf* Blackboard 2). It focuses on a treatment, once again, but a dynamic one this time. On the basis of initial information gathered in W , a physician prescribes either a weak dose $a = 0$ or a strong dose $a = 1$ of a given active principle. After one week of treatment, an examination allows the physician to gather information, which I note L , on the physiological reaction of the patient to the dose that he or she was initially prescribed. In accordance with the nature of the information that he or she gathers, the physician prescribes either a weak dose $a' = 0$ (renewal of the initial dose if $a = 0$, decrease if $a = 1$) or a strong dose $a' = 1$ (renewal of the initial dose if $a = 1$, increase if $a = 0$). The variable Y quantifies the effect of the dynamic treatment (A, A') on the disease, for example in terms of survival, seriousness, or duration. The effect of the static treatment $(a, a') = (1, 1)$ compared to the static treatment $(a, a') = (0, 0)$ is naturally expressed by comparing $Y_{1,1}$ to $Y_{0,0}$. Here, $Y_{1,1}$ and $Y_{0,0}$ are the primary endpoints when we *impose* two successive strong doses or two successive weak doses, respectively.

$\exists f_W, f_A, f_L, f_{A'}, f_Y$ deterministic functions, $\exists U_W, U_A, U_L, U_{A'}, U_Y$ independent sources of randomness such that	
“natural system”	“controlled system”
$\begin{cases} W = f_W(U_W) \\ A = f_A(W, U_A) \in \{0, 1\} \\ L = f_L(W, A, U_L) \\ A' = f_{A'}(W, A, L, U_{A'}) \in \{0, 1\} \\ Y = f_Y(W, A, L, A', U_Y) \\ O = (W, A, L, A', Y) \end{cases}$	$\begin{cases} W = f_W(U_W) \\ A = a \\ L_a = f_L(W, A = a, U_L) \\ A' = a' \\ Y_{a,a'} = f_Y(W, A = a, L_a, A' = a', U_Y) \\ O_{a,a'} = (W, A = a, L_a, A' = a', Y_{a,a'}) \end{cases}$
how nature produces O	how nature produces $O_{a,a'}$ when we impose that $(A, A') = (a, a')$, for $a, a' \in \{0, 1\}$
	here, <i>ceteris paribus</i> is not equivalent to “conditionally on anything”

Blackboard 2 Modeling how nature produces the random variable $O = (W, A, L, A', Y)$ without intervention (left) and with an intervention $(A, A') = (a, a')$ (right). Here, “*ceteris paribus sic standibus*” is not equivalent to “conditionally on anything”.

JCT: This scenario is different from the previous one insofar as the cause is sequentially determined. What we call an “effect” of $(a, a') \in \{0, 1\}^2$ on Y could be expressed in terms of a comparison between $Y_{1,1}$ and $Y_{0,0}$ for example, the values of which are functions of *the same* W but *not the same* L ! This constitutes indeed, I should think, a blatant demonstration of the fact that the phrases are not interchangeable. Conditioning on W and/or L to address the effect of (a, a') on Y would not make any sense.

AC: Here, “*ceteris paribus*” is an obvious reference to the functions f_W, f_L, f_Y and to the way the sources of randomness U_W, U_L, U_Y are produced and, therefore, only to certain features of the probability law of the phenomenon of interest.

4 *Post hoc, ergo propter hoc*

ID: We started with probabilistic theories of causality. They led us to discussing what must be held fixed in order to think causally, and whether or not this holding fixed is equivalent to conditioning. But it occurred to me that, in fact, you have an interventionist approach to causality. For you, thinking causally means being capable of imposing, via an intervention, the nature of the cause we are considering, and then reasoning *ceteris paribus*.

JCT: And our “*ceteris paribus*” is not sufficient for us to draw causal conclusions. I suppose this is quite typical of a statistical approach to causality...

AC: The notion of intervention thus appears as one of the technical devices involved in the mathematical formalization of the notions of cause and effect.

ID: Intervention is not a mere mathematical, technical device, though.

JCT: Of course not! For most people, interventions are very concrete operations. I think of C. Bernard who inserted curare under the skin of the back of a frog to study the effects of this substance. Interventions of this type are very different from your interventions.

ID: Not so much, insofar as, in both cases, intervention is a method of scientific investigation. We more frequently speak of experiments, but it is one and the same thing.

AC: And what can experiment be opposed to, then?

JCT: To observation. C. Bernard, in particular, differentiated observational sciences from experimental sciences. He considered the latter as superior to the former.

AC: And what makes him claim that experiment is superior to observation?

JCT: The fact that experiment yields far more interesting results. Consider, for example, this passage [40, Second part, Chapter 2, VIII, p. 114]:

Proof that a given condition always precedes or accompanies a phenomenon does not warrant concluding with certainty that this condition is the proximate cause of a phenomenon; it must still be established that when this condition is removed, the phenomenon will no longer appear. If we limited ourselves to the proof of presence alone, we might fall into error at any moment and believe in relations of cause and effect where there was nothing but simple coincidence. As we shall later see, coincidences form one of the most dangerous stumbling blocks encountered by experimental scientists in complex sciences like biology. It is the *post hoc, ergo propter hoc* of the doctors, into which we may very easily let ourselves be led, especially if the result of an experiment or an observation supports a preconceived idea.

The Latin phrase *post hoc, ergo propter hoc* means “after this, therefore because of this”.

ID: In other words: experiments constitute the best way to identify causal relations and, conversely, it is quite difficult to establish causal relations when we can only rely on observing. It is an idea that is already present in J.S. Mill’s [16].

AC: Thus, in the scenario of Blackboard 1, noting, at the end of the experiment, that the A of observation $O = (W, A, Y)$ that results from it equals $a \in \{0, 1\}$ (left-hand side of the Blackboard) is an event of a different nature from the one which consists in noting that the A of observation $O_a = (W, A = a, Y_a)$ that results from the experiment under the intervention $A = a$ equals, by definition, a (right-hand side of the Blackboard). Consequently, the primary endpoints Y and Y_a do not have the same interpretation.

ID: Because intervention is so efficient to reach causal knowledge, it has long been a central topic in the philosophy of science. This experimental tradition has culminated with the idea that intervention could do more than merely helping us find causal relations: it could actually provide us with a definition of causality itself. This move was explicitly and famously made by J. Woodward [41].

AC: And where did this move lead philosophers?

ID: First of all, to the conclusion that it is conceptually relevant to rely on the notion of intervention, or manipulation, to define a candidate cause. Candidate causes thus are factors for which I can devise an intervention to modify them *ceteris paribus*.

AC: But isn't our very ability to devise interventions reflecting what we know when we set to the task?

JCT: It is, and here comes a deontologically delicate example. In line with the nineteenth century work of the teratologists who inscribed monsters in the development of the normal human being, the pioneering works of E. Wolff and A. Jost set the bases of an experimental teratogenesis that notably made it possible to understand the mechanisms of sexual ambiguities [42, 43].

ID: Can you elaborate further?

JCT: The progress of molecular genetics made it possible to define the notion of genetic sex, XX for males and XY for females, which can differ from the apparent sex and the perceived sex. And therefore, at least for animals, it became possible to imagine modifications of sex determinism very early in the development, or even to conduct early *in utero* interventions to prevent an abnormal masculinization of a female fetus, or the opposite [44–47].

AC: So, if I understand correctly, within the framework of this philosophical analysis, gender becomes a candidate cause! Hence, the set of candidate causes dynamically evolves along with what we know and what we consider plausible. Incidentally, biological plausibility is one of A.B. Hill's criteria.

JCT: As Sherlock Holmes told Doctor Watson [48]:

When you have eliminated the impossible, whatever remains, however improbable, must be the truth.

This way of thinking paved the way to defining a bacterial origin to peptic ulcer, long unimaginable because considered to be out of the realm of possibilities, leading to a paradigm change and to the 2005 Nobel Prize in Physiology or Medicine for B.J. Marshall and J.R. Warren [49, 50].

AC: Now, I wonder... When we talk of interventions, do they have to be *real* interventions?

ID: Not necessarily, or else we would make causality dependent of our intervention capacities, but it is in fact an objective notion. To prevent the definition of causality from depending on what we know, philosophers widened the notion of intervention, including these fictitious but imaginable interventions, going as far as to accept “metaphysically possible” interventions.

AC: Does the analysis of causality in terms of imaginable intervention present flaws?

ID: I can see at least one! The example of the electromagnetic spinning-top used by M. Kistler [51] to show that sometimes analysis through intervention does not make it possible to differentiate between causality and certain types of regular association. The heart of the matter is that, if we follow this analysis, each one of the two candidate causes for the spinning of the top on its axis is a cause of the other one. Since a causal

relationship cannot be symmetrical, there is a contradiction. To this day, 11 objections have been raised against this counter-example, and they have all been defused.

JCT: In the end, what is M. Kistler's opinion?

ID: According to him, the notion of intervention is not sufficient to define causality. For him, interventionist theories of causality fail to capture the difference between causal relations and association laws.

JCT: That makes sense. Yet, this has no bearing on the unquestionable methodological importance of interventions.

ID: As a matter of fact, even though he studied in depth the notion of experimentation, J.S. Mill was no interventionist, but was more of a Humean, at least as far as the definition of causality is concerned.

5 From population to individuals

AC: As a physician, Jean-Christophe, you often focus on individuals rather than on populations. Aren't you therefore confronted to the delicate question of deciding what it is that you can really learn from the statistical analysis of causal problems?

JCT: Questioning statistics, or rather debating on the potential impossibility for medicine to learn anything from statistics, is nothing new. It is present in C. Bernard of course, for instance in the following passage [40, Second part, Chapter 2, IX, p. 243]:

A great surgeon performs lithotomy by a single method; later he makes a statistical summary of deaths and recoveries, and he concludes from these statistics that the mortality law for this operation is two out of five. Well, I say that this ratio means literally nothing scientifically and gives us no certainty in performing the next operation; for we do not know whether the next case will be among the recoveries or the deaths.

C. Bernard thus negates any and all external validity to observations.

AC: What is a lithotomy?

JCT: It is a surgical method to remove bladder stones. The Hippocratic Oath referred to it back in the late fifth century BC. Famous Flemish painters like J. Bosch, J.S. van Hemessen, P. Huys, and P. Bruegel the Elder have represented lithotomies, and French composer M. Marais paid tribute to it with a piece of music, his *Tableau de l'opération de taille*.

ID: In 1835, the French Academy of sciences had already addressed the question of deriving individual results from results at the population level, when dealing with Doctor J. Civiale's work comparing two different therapeutic approaches to treat bladder stones, again. We can read in the minutes [52]:

In statistics (...) the first task is to lose sight of the individual seen in isolation, to consider him only as a fraction of the species (...). In applied medicine, on the contrary, the problem is always individual, facts to which a solution must be found only present themselves one by one (...). For us, the masses are quite irrelevant to the issue.

The French controversy lasted for several years, with important contributions by M. Gavarret contradicted by M. Valleix [53, 54]. This also reminds me of the earlier debate on inoculation, which inflamed intellectuals during the Enlightenment.

AC: What is this debate about?

ID: It is the debate that was pursuant to Bernoulli's work – Daniel, one of the inventors of the statistical theory, nephew of Jacques, himself considered as one of the inventors of the probability theory. In 1760, he tried to determine what effect the inoculation of smallpox would have, if generalized to all young children, in the prevention of variola. A probabilistic reasoning involving comparisons in life expectancy led him to promote preventive inoculation as a salutary measure of collective disease prevention, in spite of the individual risk incurred.

JCT: A long mathematical and philosophical debate ensued, fueled among others by J. d'Alembert, who developed a detailed analysis of D. Bernoulli's theory, and wrote [55]:

I suppose, along with Mister Bernoulli, that the risk of dying from the inoculation [at the age of 30] is of 1 to 200. This being established, I believe that, in order to appreciate the benefit of inoculation, one should compare, not the average life of 34 years to the average life of 30, but the 1 to 200 risk of dying within a month of the inoculation (and this at the age of thirty, while still healthy and young) compared to the faraway benefit of living four additional years after 60 years of life, at an age when enjoying life is not so easy any longer... Here is, indubitably, what makes so many people, and especially so many mothers, little inclined to inoculation.

One of the conclusions is that collective benefit is different from individual benefit.

AC: It does correspond to the reality of our condition. Formally, within the framework of Blackboard 1 with $W = \emptyset$ reduced to nothing, noting $a = 1$ the realization of inoculation and $a = 0$ its contrary on the one hand, and $Y_a = 1$ the development of variola and $Y_a = 0$ its contrary under the intervention $a \in \{0, 1\}$ on the other hand, then four scenarios are conceivable for a given individual, whether he or she belongs to one among four groups of combinations of the possible values 0–1 for Y_0 and Y_1 respectively: $Y_0 = Y_1 = 1$ (group $\mathcal{G}_1$), $Y_0 = 1, Y_1 = 0$ (group $\mathcal{G}_2$), $Y_0 = 0, Y_1 = 1$ (group $\mathcal{G}_3$) and $Y_0 = Y_1 = 0$ (group $\mathcal{G}_4$).

JCT: I see what you are aiming at! Let p_k be the proportion of the whole population covered by the group $\mathcal{G}_k$...

ID: What do you mean exactly?

JCT: I mean that if I randomly pick a person in the general population without knowing which group he or she belongs to, then the probability he or she might belong to group $\mathcal{G}_k$ equals p_k . In the model Antoine mentioned, in which $Y_0 = Y_1 = 1$ for the members of group $\mathcal{G}_1$, $Y_0 = 1, Y_1 = 0$ for those of group $\mathcal{G}_2$, $Y_0 = 0, Y_1 = 1$ for those of group $\mathcal{G}_3$, and $Y_0 = Y_1 = 0$ for those of group $\mathcal{G}_4$, inoculation has a causal effect on the development of variola if $p_2 > 0$ or $p_3 > 0$.

AC: Here, $p_2 > 0$ and $p_3 > 0$ mean that the groups $\mathcal{G}_2$ and $\mathcal{G}_3$ are not empty.

ID: How does this induce a causal effect?

AC: Insofar as inoculation $A = 1$ or absence of inoculation $A = 0$ affects my future prospects $Y = Y_A$ regarding the development of variola if I belong to one of the groups $\mathcal{G}_2$ or $\mathcal{G}_3$...

JCT: However, if I belong to one of the groups $\mathcal{G}_1$ or $\mathcal{G}_4$, then inoculation or absence of inoculation does not change anything.

ID: Which closes the description on an individual scale. As for the description on a collective scale, we observe that inoculation has a *beneficial* statistical effect if and only if $p_2 + p_4 > p_3 + p_4$, i.e. if and only if $p_2 > p_3$, i.e. under the condition that the proportion of individuals who would benefit from the inoculation (those from $\mathcal{G}_2$) be larger than the proportion of those who would suffer from it (those from $\mathcal{G}_3$).

AC: At the collective level, the question is therefore to assess the difference $p_3 - p_2$. At the individual level, the question is to determine the group to which each individual belongs. In order to carry out these two statistical tasks, it is necessary to have additional, relevant information on each individual.

JCT: It all depends on what the adjective “relevant” means!

AC: We will get back to this.

6 From counterfactual worlds to the actual world

JCT: In this very special scenario, because Y_0 and Y_1 are deterministic in each one of the four groups, the indecision concerning group belonging is rigorously equivalent to the impossibility to observe both Y_0 and Y_1 . A given individual is either inoculated, or not inoculated. In the first case I do not know if the individual would develop the disease if he or she were not inoculated; in the second case I do not know if he or she would develop it if he or she were. It is a counterfactual model!

AC: Indeed, there is something that is conceptually difficult here: there seems to be something more, or something else, in causality, than what actually takes place in our world.

ID: We find a similar idea at the heart of the “counterfactual” philosophical theories of causality. These theories are based on the fundamental idea that A caused B if and only if B would not have been the case if A had not been the case – or else, in a probabilistic version, the probability of B would have been smaller if A had not been the case. Which means, essentially, that causality is related not only with what is the case, which actually takes place in our world, but also with what is not the case, which takes place in “another possible world” – to use the phrase of the philosophers.

JCT: All right. But how do you address the methodological difficulty I mentioned earlier? It seems quite difficult to determine what would have happened if things had been different. Do you have a conceptual framework and statistical tools at your disposal, that would help you to elaborate answers to such questions?

ID: If your question refers to what happens at the individual scale, the answer is no, we do not. But if it refers to the larger scale of the whole population, there are indeed such conceptual frameworks and statistical tools.

AC: Let us admit we adhere to the following counterfactual probabilistic model. I consider an individual randomly taken from my population. The individual is associated with a set of data X , known as the full data, decomposed into a finite superposition of parts X_i , $i \in I$, that may be redundant; I write it: $X = (X_i)_{i \in I}$. The i th data X_i has to be considered as the description of what happens for this individual in the i th counterfactual world. Knowing X implies simultaneously knowing the outcomes of the experiment for this individual in each one of the counterfactual worlds and thus, in particular, in the actual world, which is one of the counterfactual worlds. Observation in the actual world, O , is understood as a projection of the full data X in the actual world, with the loss of information this implies.

JCT: Thus, if I name $\mathbb{P}$ the law of the full data X and P the law of the observation O , the question I asked can be rephrased in the following way: can we infer features of $\mathbb{P}$ from observations made under P ? I am referring to features (i) that involve the comparison of counterfactual worlds, and (ii) that are expressed at the population level, not at the individual level.

Blackboard 3 Illustrating how a postulated counterfactual universe induces the actual universe.

AC: Absolutely (*cf* Blackboard 3). As a statistician, I conceive these features you are interested in as a functional $\theta : \mathbb{M} \rightarrow \Theta$, which associates the characteristic $\theta(\mathbb{P}) \in \Theta$ to each law $\mathbb{P} \in \mathbb{M}$ that the full data X may follow. Since we know how the observation O is deduced from X , we can elaborate a second functional $\vartheta : \mathcal{M} \rightarrow \Theta$ that associates to any law $P \in \mathcal{M}$ that O may follow a characteristic $\vartheta(P) \in \Theta$ such that, if I note $P(\mathbb{P})$ the law that O follows when X follows $\mathbb{P}$, then at the price of a hypothesis said of “randomization,” $\vartheta(P(\mathbb{P})) = \theta(\mathbb{P})$! The miracle, if I may say so, is that it is possible to infer $\vartheta(P(\mathbb{P}))$ from the observations made under $P(\mathbb{P})$ and therefore, indirectly, to infer $\theta(\mathbb{P})$ even though we have no observations made under $\mathbb{P}$!

JCT: The randomization hypothesis deserves more than a mere evocation (*cf* Sections 8 and 9). It is important to understand that it concerns the law $\mathbb{P}$ in all its complexity, *i.e.* it involves simultaneously all the counterfactual worlds, so that it is, by essence, impossible to test from the observations made under P .

ID: It would be a good idea, before we go any further, to put this discussion in the context of the two previous examples.

AC: Let us go back to the scenario of Blackboard 1. The full data X in the counterfactual world is written $X = (W, A, Y_0, Y_1)$, or $X = (X_0, X_1)$ with $X_0 = (W, A, Y_0)$ and $X_1 = (W, A, Y_1)$, and the observation O in the actual world is written $O = (W, A, Y_A)$, or X_A , to make things simpler.

ID: And what functionals θ and ϑ could we consider to address the scenario we evoked earlier, that of the evaluation of the potential effect of a treatment compared to the absence of any treatment on a disease, for instance in terms of survival or death?

JCT: The easiest is to choose the functional $\theta : \mathbb{M} \rightarrow \Theta$.

ID: Why is that?

JCT: Because θ associates a characteristic $\theta(\mathbb{P})$ to any law $\mathbb{P}$ that the full data X may follow. And it is conceptually easier to characterize an effect measure when we know the counterfactual outcomes! So, let us think as statisticians and see that comparing $\mathbb{P}\{Y_1\}$, the mean value of the primary endpoint Y_1 quantifying the outcome of the disease when we impose a treatment, with the mean value $\mathbb{P}\{Y_0\}$ of the primary endpoint Y_0 quantifying the outcome of the disease when we impose the absence of any treatment, gives access to the heart of the causal mechanism by quantifying the potential effect of the cause on the disease under the assumption that the intervention does not change the behavior.

AC: The functional $\theta : \mathbb{M} \rightarrow \Theta = [-1, 1]$ characterized by $\theta(\mathbb{P}) = \mathbb{P}\{Y_1\} - \mathbb{P}\{Y_0\}$ can play this role. We call it the causal excess risk, which has values in the interval $[-1, 1]$.

ID: I am astounded by your boldness! I thought we were discussing what we mean by “cause” and “effect,” and here you are, quantifying this notion! Should not we first decide whether the “cause” is indeed a cause and its “effect,” an effect?

AC: The statistician will respond to your question by elaborating a procedure, called test procedure, based on this quantification!

JCT: Then, what about the ϑ , about which you said, Antoine, that you knew how to associate it with θ ?

AC: We can justify that, under the randomization hypothesis previously mentioned...

ID: ... which we will have to go back to! (cf Sections 8 and 9)...

AC: ... we naturally associate to the question the functional $\vartheta : \mathcal{M} \rightarrow \Theta = [-1, 1]$ characterized by $\vartheta(P) = P\{P\{Y|A = 1, W\}\} - P\{P\{Y|A = 0, W\}\}$ and called generalized excess risk.

ID: What does the expression $P\{P\{Y|A = a, W\}\}$ for $a \in \{0, 1\}$ represent?

JCT: The explanation is twofold. *Primo*, $P\{Y|A = a, W\}$ is a random variable that only depends on W , or equivalently such that $P\{Y|A = a, W\} = \varphi(W)$. Informally, $\varphi(\omega)$ is the mean value of Y under P when we observe $A = a$ and $W = \omega$. *Secundo*, in the same way that $\mathbb{P}\{Y_a\}$ is the mean value under $\mathbb{P}$ of the random variable Y_a , $P\{P\{Y|A = a, W\}\}$ is the mean value of $\varphi(W)$ under P .

ID: So... $P\{P\{Y|A = a, W\}\}$ is not equal in general to $P\{Y|A = a\}$, because the marginal law of W can differ from the conditional law of W given $A = a$. I see. What would you do in the second scenario (cf Blackboard 2)?

AC: Here is what we have (cf Blackboard 4).

full data:	$X = (W, A, L_0, L_1, A', Y_{0,0}, Y_{0,1}, Y_{1,0}, Y_{1,1}) \sim \mathbb{P} \in \mathbb{M}$
observation:	$O = (W, A, L_A, A', Y_{A,A'}) \sim P(\mathbb{P}) \in \mathcal{M}$
causal excess risk:	$\theta(\mathbb{P}) = \mathbb{P}\{Y_{1,1}\} - \mathbb{P}\{Y_{0,0}\}$
excess risk:	$\vartheta(P) = \vartheta_{1,1}(P) - \vartheta_{0,0}(P)$, with
	$\vartheta_{a,a'}(P) = P\left\{\sum_{\ell \in \mathcal{L}} P\{Y A' = a', L = \ell, A = a, W\} \times P(L = \ell A = a, W)\right\}$

Blackboard 4 *Elaborating a statistical parameter of excess risk based on a certain causal excess risk in the scenario of Blackboard 2.*

JCT: New perspectives now open up in our discussion regarding the *distinguo* individuals *versus* population and the historical example of inoculation.

ID: We are listening.

JCT: I go back within the framework of the scenario of Blackboard 1, but twist it, since I additionally suppose that the variable W identifies exactly the group to which each individual belongs.

AC: In short, if I observe $W = \omega$ then the individual belongs to the group $\mathcal{G}_\omega$, and therefore, $P(W = \omega) = p_\omega$ for $\omega \in \{1, 2, 3, 4\}$.

JCT: That is right. And now, let me draw your attention to the following fact: $\vartheta(P) = p_3 - p_2$, as is proven by this simple calculation (cf Blackboard 5).

to alleviate notation, set: $\Delta(W) = P\{Y|A = 1, W\} - P\{Y|A = 0, W\}$
 by definition:

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

and, on a case by case basis:

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

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

Blackboard 5 Proving the equality $\vartheta(P) = p_3 - p_2$ in the context of the scenario of Blackboard 1 when $W \in \{1, 2, 3, 4\}$ indicates to which of the groups $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \mathcal{G}_4$ one belongs.

ID: This is interesting indeed: if the counterfactual piece of information on group belonging were available, then the excess risk would have a causal interpretation.

AC: We mentioned earlier (cf Section 5) that it is necessary to have additional, “relevant” information on each individual in order to assess the difference $p_3 - p_2$ at the collective level, and the belonging to one of the four groups at the individual level. We have just concluded that if the counterfactual piece of information concerning group belonging were available, then it would be sufficient relevant information. But it is unavailable, due to its counterfactual nature, and one of the roles of the statistician is to find substitutes for it. We actually speak of “predictors,” because their function is to help the statistician predict the value of the counterfactual variable we would have liked to observe, or the probability to observe it.

7 From the actual world to counterfactual worlds

JCT: Where do we stand? We have shown how to elaborate a model of causality provided that we accept the possibility of counterfactual worlds from which our actual world would ensue, i.e. if we accept the possibility that what happens in our world is the projection of what happens in these counterfactual worlds. Which presents at least the following formal benefit: $\mathbb{P}$ induces $P(\mathbb{P})$ and therefore $\mathbb{M}$ induces $\mathcal{M}$, θ induces ϑ , X induces O , and so on. Does the formal structure collapse if we deny this conception of the actual world?

AC: No. It is formally possible to adopt a diametrically opposed point of view, provided we admit we can play heads or tails an infinity of times in total independence [56].

ID: Well, if we go back to the beginning of our exchange, you ask us to let you resort to as many independent random variables as you might wish.

Blackboard 6 Illustrating the construction, based on the actual universe, of a counterfactual universe that induces it. The construction requires flipping a coin independently an infinity of times.

AC: Absolutely! Let me first set the scene (*cf* Blackboard 6). Say that we focus on a certain question of interest relative to a random phenomenon that we observe in the actual world. Formally, say that this question pertains to the law of the observation O in the actual world. Let $\mathcal{M}$ be the set of candidate laws P , among which one is the true law of O . For the sake of the argument, let us accept temporarily the possibility of counterfactual worlds from which our actual world would ensue. Let us note $\mathbb{M}$ the set of laws $\mathbb{P}$ to which we would then have access, among which one would be the true law of the counterfactual variable X that would induce O . We would express the question of interest in terms of an *ad hoc* functional $\theta : \mathbb{M} \rightarrow \Theta$, which in turn would induce the functional $\vartheta : \mathcal{M} \rightarrow \Theta$.

JCT: There is nothing new so far, and you keep us wanting more! What are you preparing us for?

ID: I would say this preamble was necessary to introduce the functional $\vartheta : \mathcal{M} \rightarrow \Theta$. Is my intuition right?

AC: Yes, it is excellent! Now, the idea is the following. It is formally possible: *primo*, to elaborate a counterfactual variable $\tilde{X}$, of the same nature as X , which induces O just as X induces O ; *secundo*, to elaborate, for any candidate law $P \in \mathcal{M}$, a law $\tilde{\mathbb{P}}(P)$ of $\tilde{X}$, hence the set $\tilde{\mathbb{M}} = \{\tilde{\mathbb{P}}(P) : P \in \mathcal{M}\}$, so that if $\tilde{\mathbb{P}}(P)$ induces P' just as $\mathbb{P}$ induces $P(\mathbb{P})$ then $P' = P$; all this so that $\theta(\tilde{\mathbb{P}}(P)) = \vartheta(P)$! In short it is formally possible to elaborate, from O , $\mathcal{M}$, θ , and ϑ , a counterfactual world of the same nature as the one summed up by X , $\mathbb{M}$, and θ .

JCT: We could almost be puzzled by the fact that we can thus randomly pick a counterfactual variable $\tilde{X}$ compatible with the observation O , and therefore observe it, even though it is not possible in the actual world!

ID: Puzzling indeed, until we realize that causality, in this counterfactual, formally elaborated world, cannot coincide with any notion of causality in our world!

JCT: The technical artifice is indeed formally convenient, but we have no access to any causal interpretation.

8 The randomization trick

ID: Thus, by relying on a counterfactual formalism we can define causal parameters that quantify the causal questions we are dealing with. It also teaches us that the conditions to reason causally in *post hoc*,

ergo propter hoc terms are generally not met. It may be time to go back to the randomization trick, which we touched upon earlier (cf Section 6). What does it consist in?

JCT: It is a trick the purpose of which is to warrant that a *post hoc*, *ergo propter hoc* reasoning is causal in nature. In other words, the randomization trick consists in gathering conditions under which we can infer the causal parameters on the basis of the observation of the results of a repeated experiment. The governing idea is to control the exposition by drawing it independently from the consequences each intervention might have.

ID: Let us consider the easiest scenario, in which the objective is to determine the causal effect of a treatment on a disease compared to the absence of treatment – for example, taking a placebo. If I understand correctly, the sequence of events is as follows: *primo*, I recruit a patient among a clearly identified population of potential patients; *secundo*, I play heads or tails – for example with a balanced coin, *i.e.* according to Bernoulli's law of parameter $1/2$ – to randomly choose treatment or placebo, and I impose it to my patient – by which, the nature of the exposition, treatment or placebo, is intrinsically independent of the two outcomes, for instance either success or failure, that they might have on my patient, who is ignorant of his or her counterfactual status (cf Section 6); *tertio*, I follow the patient until I observe the outcome.

JCT: You are perfectly right. I would like to underline the extent to which the randomized procedure is different from the observational framework as criticized by C. Bernard when he advocated a controlled experimental framework. In the latter framework, the patient you recruited could consult his or her physician. The physician could, on the basis of the medical record and a physical examination, decide to prescribe either a treatment or a placebo. Thus, of course, the nature of the exposition, treatment or placebo, would be intrinsically dependent on the two outcomes, for instance either success or failure, that treatment or placebo would have.

AC: We can cast formal light on your two scenarios thanks to that of Blackboard 1. Jean-Christophe, what you have just described very much pertains to the “natural system” that Blackboard 1 presents. Isabelle, the patient you recruit is characterized by the variable W , which you do not need to take into account when randomly drawing the exposition. Randomly drawing the nature of the exposition finally comes back to substituting for the equation $A = f_A(W, U_A)$ of the “natural system” the alternative equation $A = U_A$ with U_A of Bernoulli's law of parameter $1/2$.

ID: I see! And rather than speaking of substitution, as you do, we could say that, *via* randomization, we manage to get $f_A(W, U_A) = U_A$ with U_A of Bernoulli's law of parameter $1/2$: in short, you impose the form of the function f_A !

AC: This is true. The point of such formalization is notably to make it clear that, according to the result of the randomized draw, we can observe either one or the other of the controlled systems, and can therefore causally interpret the result of a comparison of their two behaviors.

JCT: Do you mean causally interpret a comparison of the means of the primary endpoints quantifying the observed outcomes of either treatment or placebo?

AC: Yes, that is it. Listen to this. Let us say we consider the causal excess risk $\theta : \mathbb{M} \rightarrow [-1, 1]$ characterized by $\theta(\mathbb{P}) = \mathbb{P}\{Y_1\} - \mathbb{P}\{Y_0\}$. Let us admit the consistency hypothesis which says that $Y = Y_A$, this equality being interpreted as the coincidence of the outcome in the actual world with the outcome in the counterfactual world we explore. Let us note that, inherently, due to randomization, $A = U_A$ is independent of (Y_0, Y_1) . Well, $\theta(\mathbb{P})$ coincides with the difference of the conditional means of Y given $A = 1$ and given $A = 0$, respectively, which we call the naive excess risk: $P\{Y|A = 1\} = P\{Y|A = 0\}$. Formally: (i) $P\{Y|A = 1\} - P\{Y|A = 0\} = \mathbb{P}\{Y_A|A = 1\} - \mathbb{P}\{Y_A|A = 0\}$ by consistency, (ii) this difference is equal to $\mathbb{P}\{Y_1|A = 1\} - \mathbb{P}\{Y_0|A = 0\}$ if we replace Y_A by Y_1 or Y_0 , depending on $A = 1$ or $A = 0$, and (iii) it is equal to $\theta(\mathbb{P})$ because the independence between A and (Y_0, Y_1) entails $P\{Y_1|A = 1\} = P\{Y_1\}$ and $P\{Y_0|A = 0\} = P\{Y_0\}$!

JCT: I would like to emphasize that the presentation of the randomization trick we have just made summarizes its very substance.

ID: And how do you plan on doing this?

JCT: Let us go back to the more complex example we have already mentioned (*cf* Blackboard 2). Our goal is to determine the causal effect, not of a treatment compared to a placebo, but rather of a sequentially determined treatment, characterized by a couple (a, a') with $a, a' \in \{0, 1\}$, compared to the reference treatment, noted $(0, 0)$.

AC: Presented this way, the only difference with what we saw previously is the characterization of the exposition according to four different levels instead of two. The sequence of events as Isabelle suggested it still holds: *primo*, I recruit a patient from a population of well-identified potential patients; *secundo*, I draw at random, for example with two balanced coins independently flipped, the nature (a, a') of the treatment and I impose it to my patient – doing which, the nature of the exposition is intrinsically independent from the four outcomes the four possible prescriptions would have; *tertio*, I follow the patient until I can observe the outcome.

JCT: And, as you did earlier, we can cast formal light on the process with the scenario of Blackboard 2. What you describe pertains exactly to the “natural system” that it presents. The patient you recruit is first characterized by the variable W , which you do not have to take into account since drawing at random, as you do, the nature of A comes back to imposing that $f_A(W, U_A) = U_A$ with U_A a Bernoulli variable of parameter $1/2$. Similarly, you can neglect the intermediate information summarized by the variable L , because drawing at random, as you do, the nature of A' comes back to imposing that $f_{A'}(W, A, L, U_{A'}) = U_{A'}$ with $U_{A'}$ a Bernoulli variable of parameter $1/2$ independent from U_A . Following the patient until one can observe the outcome comes back to observing Y .

ID: My turn to take the helm! Let us say that we focus on the causal excess risk $\theta : \mathbb{M} \rightarrow [-1, 1]$ characterized by $\theta(\mathbb{P}) = \mathbb{P}\{Y_{1,1}\} - \mathbb{P}\{Y_{0,0}\}$, and let us admit the consistency hypothesis which states that the outcome in the actual world, Y , coincides with the outcome in the explored counterfactual world, $Y_{A,A'}$. The “miracle” of randomization, to use your words, Antoine, is that, in the scene we have just set, the causal parameter $\theta(\mathbb{P})$ is equal to the difference of the conditional means of Y given $(A, A') = (1, 1)$ and given $(A, A') = (0, 0)$, respectively.

JCT: This brings an end to the first zest for complexity. Indeed, I would like to discuss a more delicate case, in which we aim at comparing various treatment regimens.

ID: What are you saying?

JCT: To put things differently, I would like to causally compare a static treatment regimen, in that the dose A initially prescribed is necessarily renewed after the mid-course visit, or $A = A'$, with a dynamic treatment regimen in which the second dose A' may be different from the first one according to the results of the mid-course visit recorded in L .

AC: Formally, within the framework of the natural system of Blackboard 2, we have $f_A(W, U_A) = f_A(W)$, which only depends on W , the static and dynamic treatment regimens corresponding respectively to $f_{A'}(W, A, L, U_{A'}) = A$ and $f_{A'}(W, A, L, U_{A'}) = f_{A'}(W, L)$, which depend only on W and on L .

ID: The real situation you suggested to illustrate the second scenario, Jean-Christophe (end of Section 3), thus turns out to be a special case. For the same reasons as those presented previously, it is not possible to draw causal conclusions from the observation of the two natural systems corresponding to the static and dynamic treatment regimens. But with the randomization trick, we are able to create the experimental conditions that transform the two natural systems into two controlled systems, the observation of and the confrontation to which lead to causal conclusions.

JCT: Randomization bears on the assignment of the type of treatment regimen, either static or dynamic. We therefore have to bring out a new variable in the system, which had not been necessary so far. Let us call it R for

“regimen.” It is chronologically situated between the description of the patient, W , and the first treatment assignment, A . The variable R testifies to the nature of the treatment regimen chosen by the physician according to his or her observation of W in the natural system. Let us say that $R = 0$ for the static treatment regimen and $R = 1$ for the dynamic one. It leads to the following natural and controlled systems (cf Blackboard 7).

$\exists f_W, f_R, f_A, f_L, f_{A'}, f_Y$ deterministic functions, $\exists U_W, U_R, U_A, U_L, U_{A'}, U_Y$ independent sources of randomness such that	
“natural system”	“controlled system”
$\begin{cases} W = f_W(U_W) \\ R = f_R(W, U_R) \in \{0, 1\} \\ A = f_A(W, R, U_A) \in \{0, 1\} \\ L = f_L(W, R, A, U_L) \\ A' = f_{A'}(W, R, A, L, U_{A'}) \in \{0, 1\} \\ Y = f_Y(W, R, A, L, A', U_Y) \\ O = (W, R, A, L, A', Y) \end{cases}$	$\begin{cases} W = f_W(U_W) \\ R = r \\ A_r = f_A(W, R = r, U_A) \\ L_r = f_L(W, R = r, A_r, U_L) \\ A'_r = f_{A'}(W, R = r, A_r, L_r, U_{A'}) \\ Y_r = f_Y(W, R = r, A_r, L_r, A'_r, U_Y) \\ O_r = (W, R = r, A_r, L_r, A'_r, Y_r) \end{cases}$
how nature produces O	how nature produces O_r when we impose that $R = r$, for $r \in \{0, 1\}$

Blackboard 7 Model illustrating the randomization trick for the study of the effect of a dynamic treatment regimen, building upon the scenario developed in Blackboard 2.

AC: Formally, the randomization of treatment regimen assignment R boils down to setting $f_R(W, U_R) = U_R$ with U_R a Bernoulli variable of parameter $1/2$, for example. Instead of leaving the physician the full responsibility of choosing the treatment regimen, we leave it to chance *via* a randomization draw, the outcome of which defines which one of the two controlled systems we observe an emanation of.

9 Of confounding and the randomization hypothesis

ID: Very well! But what happens when we cannot resort to the randomization trick? Is there any lesson to learn from what we have just seen for the case in which we can do nothing but observe the behavior of the natural system?

JCT: There is one indeed, which has led to the notion of confounding and produced the randomization hypothesis that Antoine mentioned earlier. The exact formulation of the randomization hypothesis depends on the question of interest. It concerns the nature of the conditional dependence of the variable considered as a cause relative to the counterfactual effect variables, given some other variables which we label as potential confounding factors, or confounders.

ID: Do we get a clear enough picture of what confounders are? I would say they are the variables that create a dependence between the variable of supposed cause and the counterfactual variables of its supposed effects by affecting them.

AC: This is true. I would like to emphasize that the informal characterization that you suggest mixes probabilistic and causal considerations, and that it does not give a formal definition.

JCT: For decades, the literature has placed the conceptual emphasis on confounding rather than on confounders. Recently, attempts have been made to fill the resulting conceptual vacuum. For instance, T.J. VanderWeele and I. Shpitser [57] considered six candidate definitions proposed either formally or informally in the literature. Among them only one satisfies two properties that should be met.

ID: It might be a good idea to go back to the examples we used previously in order to explain further these subtle notions.

JCT: Let us return within the framework of Blackboard 1. There, the confounder is W , the variable of supposed cause is A and the counterfactual variables of its supposed effects are Y_0, Y_1 . The causal excess risk $\theta(\mathbb{P})$ quantifies the causal effect of interest. If W is a confounder, then the naive excess risk $P\{Y|A=1\} - P\{Y|A=0\}$ differs from $\theta(\mathbb{P})$. This difference between the two quantities, called confounding bias, is a constant concern in observational studies in epidemiology.

ID: How do you account for the presence of naive risk here?

AC: I think that Jean-Christophe referred to it because, when we use the randomization trick, this parameter coincides with $\theta(\mathbb{P})$.

JCT: This is correct. On the other hand, if we neglect W without resorting to this trick, *i.e.* if we exploit this naive excess risk in lieu of $\theta(\mathbb{P})$ while observing the natural system, then we do not have access to a causal relationship.

ID: What if we do not neglect it?

JCT: Well, if W does encompass all the confounders, then the randomization hypothesis is satisfied: we have A independent from (Y_0, Y_1) given W , and $\vartheta(P) = \theta(\mathbb{P})$, as Antoine said earlier (*cf* Section 6).

ID: In short, the randomization hypothesis is substituted for the randomization trick so that conditions might be gathered, allowing one to infer causal relationships from the observation of the natural system without any intervention whatsoever. Conversely, we can consider the randomization trick as a tool that warrants the validity of the randomization hypothesis. Intervening on the natural system via the randomization trick (*cf* Section 8) thus appears as part of a self-validation process, as N. Cartwright [58] says.

AC: As it happens, the randomization hypothesis is satisfied within the framework of Blackboard 1. It is mainly due to the independence of the various sources of randomness. It is easy to demonstrate the equality $\vartheta(P) = \theta(\mathbb{P})$. The proof is developed in Blackboard 8.

reminder:

(i) $P\{Y|A=a, W=\omega\}$: mean value of Y when we know that $A=a$ and $W=\omega$
(here $Y \in \{0, 1\}$ so: mean value and probability that $Y=1$ coincide)

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

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

(iv) $P\{P\{V|U\}\} = P\{V\}$ for every couple (U, V) of random variables

we can write:

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

now $A = U_A$ is independent from (Y_0, Y_1, W) so: $\mathbb{P}\{Y_a|A, W\} = \mathbb{P}\{Y_a|W\}$ for each $a \in \{0, 1\}$

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

Blackboard 8 *Proving the equality $\vartheta(P) = \theta(\mathbb{P})$ in the context of the scenario of Blackboard 1, where the randomization assumption is met.*

ID: What happens if you do take W into account, but the randomization hypothesis is nonetheless not satisfied? In other words, what happens if W does not contain all the confounders?

JCT: If the randomization hypothesis is not satisfied, then *a priori* the naive and generalized excess risks are both different from the causal excess risk.

ID: In this case, does the generalized excess risk present any advantage over its naive counterpart?

AC: From my point of view, the generalized excess risk is preferable insofar as, contrary to its competitor, it integrates knowledge on the phenomenon of interest, because it takes into consideration all the confounders identified and observed, thus approaching as much as possible the causal excess risk, given what we know and the nature of our observations.

JCT: Can you support your assertion mathematically on the basis of the deviations to the causal excess risk?

AC: I could, but at the cost of untestable hypotheses on what I fancy calling “the causal law” $\mathbb{P}$, among which I would integrate that W indeed contains some confounders. But, to be honest, I could just as well formulate other hypotheses establishing the superiority of the naive excess risk. There is no purely mathematical argument establishing the superiority of the one or the other.

ID: Speaking of hypotheses that cannot be tested, are we even able to test the randomization hypothesis? »

JCT: Well, no, we are not, and therein lies the rub: the randomization hypothesis is intrinsically impossible to test on data. We are at best able to gather a converging body of clues on its plausibility, for instance by using A.B. Hill’s criteria and Koch’s postulates, but not to verify it.

ID: Shall we go back now to the notion of confounders? You referred earlier to six candidate definitions built upon the notion of confounding among which a single one satisfies two properties. Can you elaborate?

JCT: Let me check... The two conditions are: one, controlling for all confounders suffices to control for confounding; two, each confounder in some context helps eliminate or reduce confounding bias. And the victorious definition is [57]:

A confounder may be defined as a pre-exposure covariate W for which there exists a set of other covariates W' such that effect of the exposure on the outcome is unconfounded conditional on $\{W, W'\}$ but such that for no proper subset of $\{W, W'\}$ is the effect of the exposure on the outcome unconfounded given the subset.

The expression “pre-exposure” is easily understood in the context of structural equations systems.

ID: This is food for thoughts.

AC: Indeed!

10 Of Simpson’s paradox

ID: Let us put things into perspective, shall we?

JCT: We have discussed various notions linked to that of cause and used the former in order to try and define the latter. As we proceeded, we have mathematically formalized a *vademecum* for the mathematical quantification of causal questions within a probabilistic framework that hinges on the notion of intervention.

AC: What I find remarkable is the way a causal problem leads to what I identify as a fully fledged statistical problem, as if freed from its origin, *i.e.* worthy of interest beyond the causal question that initiated it!

ID: This is what makes a scientific approach of real-life problems so interesting.

JCT: Not to forget that, as we proceeded, we formulated hypotheses, among which the randomization hypothesis, that lead us from the causal to the statistical problems. This is what we call “solving the question of identifiability.” In our formalism, these conditions warrant that $\vartheta(P) = \theta(\mathbb{P})$, in which $\theta(\mathbb{P})$ is the quantification (via a parameter said to be “causal”) of the question of interest, and $\vartheta(P)$ is its statistical counterpart.

ID: Here, $\mathbb{P}$ is what Antoine calls the “causal law” and P is the law that governs the natural system. The randomization trick, for instance within the framework of a clinical trial, is designed to, ideally, make P and $\mathbb{P}$ coincide, while the randomization hypotheses, although impossible to test in practice, warrant that $\vartheta(P) = \theta(\mathbb{P})$.

AC: And thus, the statistical question of how to infer the parameter of interest, $\vartheta(P)$, on the basis of observations made “under P ,” arises at last!

JCT: Let us start with this famous numerical example called “Simpson’s paradox” [59]. Listen: I suggest we place ourselves once again within the framework of the natural system of Blackboard 1, for a covariable W with values in $\{0, 1\}$, here denoting gender, an exposition variable $A \in \{0, 1\}$, here coding for the exposition to a risk factor, or absence thereof, and a primary endpoint $Y \in \{0, 1\}$, here coding for the occurrence of a deleterious event, or absence thereof. The observation of a population of individuals governed by this natural system yields a data set consisting in $O_1, \dots, O_n$ in which each observation $O_i = (W_i, A_i, Y_i)$ is a copy of $O = (W, A, Y)$ in so that it follows the law P , and let us suppose finally that these copies are mutually independent.

ID: Isn’t there a contradiction between the facts that the O_i all are copies of O and, simultaneously, that they are mutually independent?

AC: No, there is not. What these two properties characterize is the joint law of the data set $(O_1, \dots, O_n)$: the random generation of each O_i is governed by P , *i.e.* the law of the generic variable O , and the realization of any sub-group $(O_i : i \in I)$ fails to bring any information concerning that of the complementary sub-group $(O_i : i \notin I)$ – just as the behavior of the photons successively sent on the semi-transparent mirror, crossing it or bouncing back, does not depend on either past or future behaviors.

JCT: Summarizing all the data is elementary: under the independence hypothesis, it is not necessary to keep the order in which information accumulates; we merely have to count how many individuals appear in each one of the $2 \times 2 \times 2 = 8$ possible classes. As an example, let us imagine that this exhaustive summary leads to the following tables (*cf* Blackboard 9). Thus, for example, among the $n = 80$ observed individuals, 50% are characterized by $W = 1$ and, among these, eight exposed individuals ($A = 1$) have not developed the deleterious effect ($Y = 0$). The rightmost table is the aggregation of the other two tables; in this, we lose the information concerning W . Let us keep this in mind for the role it is going to play in the presentation of the paradox.

AC: This is what we call “contingency tables.” In his time, K. Pearson considered them to represent the quintessence of the numerical description of the actual world. The statistician reads in these tables the empirical version of P offered by observations. Often noted P_n , the subscript n referring to the size of the

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

on the one hand:

- $P_n(Y = 1|A = 1) = \frac{20}{20+20} = \frac{1}{2}$
- $P_n(Y = 1|A = 0) = \frac{16}{16+24} = \frac{2}{5}$
- hence:

$$\vartheta^0(P_n) = \frac{1}{2} - \frac{2}{5} = \frac{1}{10}$$

on the other hand:

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

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

Blackboard 9 Numerical illustration of Simpson's paradox in the context of the scenario developed in Blackboard 1: $\vartheta^0(P_n) = -\vartheta(P_n) = 1/10$, hence the two parameters cannot be simultaneously interpretable as measures of the causal effect of A on Y .

data set, it is an approximation of the unknown law P elaborated on the basis of observation, repeated n times, of the law P seen as a generation mechanism of the generic variable.

ID: What can the statistician get from such a source of information? As the exercise is purely rhetorical here, we may straightaway postulate that it is in terms of excess risk that we quantify the question of interest. Well, I am listening!

AC: To go back to the example Jean-Christophe used earlier, if the probability $P(Y = 0|A = 1, W = 1)$ is not known to us, its empirical counterpart $P_n(Y = 0|A = 1, W = 1)$ is equal to the ratio $8/(8 + 2) = 4/5$: on the $8 + 2 = 10$ individuals belonging to the class $W = 1$ for whom $A = 1$, 8 have not developed the deleterious effect.

JCT: We are in fact going to introduce two parameters of excess risk: the generalized excess risk, characterized by $\vartheta(P) = P\{P\{Y|A = 1, W\} - P\{Y|A = 0, W\}\}$, takes into account the covariable W whereas the naive excess risk, characterized by $\vartheta^0(P) = P\{Y|A = 1\} - P\{Y|A = 0\}$, neglects it.

AC: By the substitution principle, $\vartheta^0(P_n)$ and $\vartheta(P_n)$ are two estimators of $\vartheta^0(P)$ and $\vartheta(P)$.

ID: What do you mean by “substitution principle”?

JCT: It is the principle that says that, if we have a candidate estimator of the law P_0 , say P_n^0 , then it is natural to consider as estimators of $\vartheta^0(P_0)$ and $\vartheta(P_0)$ the estimators $\vartheta^0(P_n^0)$ and $\vartheta(P_n^0)$, obtained by substituting P_n^0 for P_0 . In Antoine's example, P_n^0 is simply the empirical measure P_n itself.

AC: By substitution, we thus obtain the pointwise estimations $\vartheta^0(P_n) = 1/10$ and $\vartheta(P_n) = -1/10$ (cf Blackboard 9). The theory of statistical inference teaches us that these two estimators are optimal.

ID: In what sense?

AC: In the sense that we cannot build more precise estimators when the number of observations n tends to infinity. Thus the confidence intervals, derived from the estimators by using the central limit theorem so as to contain the real unknown values with an arbitrarily high certainty, are as narrow as possible, when the number of observations n tends to infinity.

ID: You are speaking of a number of observations n that tends to infinity. What can we say when $n = 80$, as is the case here?

JCT: We could build confidence intervals that would not be based on a passage to the limit in n , and so in particular not be based on a central limit theorem. To simplify, let us admit that the units used in Blackboard 9 are dozens of thousands of individuals, and that, therefore, our pointwise estimators and the associated confidence intervals are very precise. Isabelle, what does this inspire you?

ID: What surprises me is that $\vartheta^0(P_n)$ and $\vartheta(P_n)$ are so different from one another. One thing is certain though, it is that they cannot simultaneously acquire a causal interpretation in the actual world! Otherwise, first of all, the exposition of the whole population to the risk factor would cause a 10% increase in the proportion of the population developing the deleterious effect compared to the absence of exposition of the whole population, and secondly it would lead to the simultaneous 10% decrease of the proportion in the female population, in the male population and regardless of gender. It is extremely confusing...

AC: And yet the explanation is child's play: ϑ^0 and ϑ are two separate functionals and there is *a priori* no reason for the parameters to coincide.

ID: You answer as a mathematician, and that leaves me helpless. You do solve the paradox, but what is to be concluded, concretely?!

JCT: In fact, you highlight for us both the importance and the delicate nature of the choice of quantification.

AC: The process could be replicated: we could very well break up each one of the tables corresponding to the two strata of W into two sub-tables on the basis of another covariable W' , so that on the four resulting sub-strata we get $P_n(Y|A = 1, W, W') - P_n(Y|A = 0, W, W') = 1/10$ and so, globally, $P_n\{P_n(Y|A = 1, W, W') - P_n(Y|A = 0, W, W')\} = 1/10$.

JCT: I would like us to go back to the numerical example from the standpoint of the quantification of causal links. In the scenario of Blackboard 1, it is $\vartheta(P)$ that makes sense causally, since its definition revolves around the control of the confounding induced by the confounder W . Let us note that, in this system, the randomization hypothesis is satisfied.

ID: But we could just as well suppose that this very data set comes, in fact, from a natural system in which W is a joint effect of A and Y , as is summarized here (cf Blackboard 10). In this second scenario, ϑ^0 would be the appropriate quantification of the effect of A on Y , while ϑ would be a distorted quantification, W being considered as a confounder, which it is not. Formally, the randomization hypothesis is not satisfied in this second system.

JCT: I believe we have shed light on three major points! *Primo*, that intuition is misleading: the naive quantification of the dependence between A and Y , *i.e.* that which neglects W , does not tell us anything *a priori* concerning the naive quantifications limited to strata of W . Said differently, $\vartheta^0(P) = P\{Y|A = 1\} - P\{Y|A = 0\}$ does not tell us anything *a priori* concerning the values of $P\{Y|A = 1, W\} - P\{Y|A = 0, W\}$. Yet, we might have expected that $\vartheta^0(P)$ be some kind of average of the latter $P\{Y|A = 1, W\} - P\{Y|A = 0, W\}$ across W , which the example disproves. *Secundo*, the mistakes that

$\exists f_A, f_Y, f_W$ deterministic functions, $\exists U_A, U_Y, U_W$ independent sources of randomness such that	
“natural system”	“controlled system”
$\begin{cases} A = f_A(U_A) \in \{0, 1\} \\ Y = f_Y(A, U_Y) \\ W = f_W(A, Y, U_W) \\ O = (A, Y, W) \end{cases}$	$\begin{cases} A = a \\ Y_a = f_Y(A = a, U_Y) \\ W_a = f_W(A = a, Y_a, U_W) \\ O_a = (A = a, Y_a, W_a) \end{cases}$
how nature produces O	how nature produces O_a when we impose that $A = a$, for $a \in \{0, 1\}$

Blackboard 10 *Modeling how nature produces the random variable $O = (W, A, Y)$ without intervention (left) and under the intervention $A = a$ (right), to be compared to the model developed in Blackboard 1. Here, ϑ^0 can be interpreted causally, whereas ϑ cannot; a contrario, ϑ can be interpreted causally, not ϑ^0 , in the context of Blackboard 1.*

can stem from this misleading intuition are considerable. *Tertio*, it is impossible to spare oneself the trouble of thoroughly considering the nature of the phenomenon of interest prior to determining the statistical parameter of interest.

ID: We should formulate a practical rule. A discordance between the estimators obtained on embedded tables is tantamount to a warning on the nature of the relation we are studying.

11 Of Ockham’s razor

ID: I would like us to go back to the question of inference. We started discussing it in Section 10, then Simpson’s paradox came into play and reoriented our conversation. Did we examine the question of inference from all angles?

JCT: I suggest we focus on the particularly enlightening example of the inference of the generalized excess risk, characterized by $\vartheta(P) = P\{P\{Y|A = 1, W\} - P\{Y|A = 0, W\}\}$ for a primary endpoint $Y \in \{0, 1\}$. The estimator we introduced when discussing Simpson’s paradox, $\vartheta(P_n)$, is an excellent springboard.

ID: What are its upsides?

AC: First of all, it is a substitution estimator.

ID: How is that an asset?

JCT: First, it is a natural estimator, so that its candidacy is natural and straightforward for the statistician. Then, it is, in this precise case, very easy to construct. But its main upside probably resides in the fact that a substitution estimator automatically satisfies all the constraints that the parameter must meet.

ID: What else can you tell us? What are these constraints in the case of the excess risk?

JCT: Well, they notably include that $\vartheta(P_0) \in [-1, 1]$. More sophisticated inferential methods may require a final step putting the intermediate estimator under constraint so that it becomes an admissible final estimator, satisfying all the constraints. The substitution estimator does not require it.

AC: The second upside of this estimator is that it is consistent. The expression includes a number of situations. Heuristically, it means that the deviation between the true value and its estimation tends, in a sense, toward zero when the number of observations tends to infinity.

ID: We have identified two assets. Is the optimality of $\vartheta(P_n)$ a third one?

JCT: Well, absolutely!

ID: Do these three assets make $\vartheta(P_n)$ an unsurpassable estimator?

AC: I would say “no” out of principle, because no estimator is universally the best. We could imagine pitfalls precisely designed to disadvantage it! It is nonetheless an excellent estimator, hard to surpass in the conditions of our current discussion of Simpson’s paradox (cf Section 10).

ID: What do you mean by that?

AC: I am referring to the fact that W only takes a small number of different values.

ID: Have we examined the question from all angles now?

JCT: Far from it! To be convinced, suffice it to observe that our elaboration of the estimator $\vartheta(P_n)$ is essentially based upon the finite nature of the number of values that W can take. Thus, if W can take an infinite number of values, then the whole process collapses.

ID: I can indeed imagine that if W took were it only a very large number of values, which suffices for my argument, then it would be unreasonable to try and consider simultaneously all of the sub-tables of contingency corresponding to all the values W may take, which you called “strata” earlier on.

AC: Should we nonetheless try, a large number of these tables would be sparse, i.e. they would feature one or several null numbers.

ID: It also makes me think about Ockham’s razor [60]. An estimator $\vartheta(P_n)$ built as a weighted average of estimators restricted to strata is not economical in case there is a large number of strata, yet

Pluralitas non est ponenda sine necessitate,

multiplicity should not be posited without necessity. Are we in a dead end?!

12 Targeted inference – initialization

JCT: The solution that the statistician naturally considers consists in isolating the question of the estimation of the conditional law of Y given (A, W) and in addressing it as an intermediate problem.

ID: This is a little obscure! What does “estimate a law” mean?

JCT: Since $Y \in \{0, 1\}$, its conditional law given (A, W) is a Bernoulli law and, consequently, knowing this law boils down to knowing the probability $P_0(Y = 1|A, W)$. Thus, estimating the conditional law of Y given (A, W) is equivalent to estimating the function $(A, W) \mapsto P_0(Y = 1|A, W)$; this is called “regressing Y on (A, W) .” Let us note $P_n^0(Y = 1|A, W)$ the estimator of $P_0(Y = 1|A, W)$ that we build.

AC: The superscript “0” suggests that it is an initialization stage...

ID: And what is the link between the estimation of this conditional law $P_0(Y|A, W)$ with the estimation of $\vartheta(P_0)$?

JCT: The response might be a little bit arduous!

AC: By definition, if I denote $\Delta_0(W) = P_0(Y = 1|A = 1, W) - P_0(Y = 1|A = 0, W)$ then $\vartheta(P_0) = P_0\{\Delta_0(W)\}$ is the mean (relative to W) of the random variable $\Delta_0(W)$. Now, $\Delta_n^0(W) = P_n^0(Y = 1|A = 1, W) - P_n^0(Y = 1|A = 0, W)$ naturally appears as an estimator of $\Delta_0(W)$. Therefore, it suffices, in order to deduce from it an estimator of $\vartheta(P_0)$,

to estimate its mean $P_0\{\Delta_n(W)\}$, which now requires to estimate the marginal law of W . More concretely, I would recommend here to estimate the marginal law of W merely with its empirical version.

ID: And more concretely?!

JCT: It simply means that we estimate the marginal law of W via the law that gives to each observed value W_i of W a probability $1/n$ to be produced.

AC: I like to think of this in terms of simulation. Listen. We observe nature's work as it produces independent realizations of W that I denote $W_1, \dots, W_i, \dots, W_n$. Estimating the marginal law of W is equivalent to constructing an algorithm which also produces realizations of W under a law whose definition is aimed at mimicking nature. When we estimate the marginal law of W with its empirical counterpart, then this algorithm uniformly produces each one of the observed realizations, *i.e.* it produces each W_i with probability $1/n$.

ID: Let us recapitulate, please. We estimate $\Delta_0(W)$ via $\Delta_n^0(W)$ and the marginal law of W via its empirical version, our objective being to estimate $\vartheta(P_0) = P_0\{\Delta_0(W)\}$. If I am not mistaken, the resulting estimator is written $P_n\{\Delta_n^0(W)\} = n^{-1} \sum_{i=1}^n \Delta_n^0(W_i)$.

AC: You are perfectly right, and your concise presentation highlights the fact that this initial estimator is a substitution estimator!

JCT: This presentation also makes it easy to link with the case in which W only takes a small number of values. Indeed, in such a case, our best interest is to choose $P_n^0(Y=1|a, w) = P_n(Y=1|a, w)$, as the empirical probability to observe $Y=1$ in the line " $A=a$ " of the sub-table corresponding to the stratum " $W=w$." And then, surprise, $P_n\{\Delta_n^0(W)\} = \vartheta(P_n)$, the substitution estimator of the empirical measure.

ID: The estimator $P_n\{\Delta_n^0(W)\}$ is an alternative to $\vartheta(P_n)$ for the case in which W does not take a small number of values, and it extends $\vartheta(P_n)$, which is the optimal substitution estimator, when W takes a small number of values. In this capacity, does it inherit the latter's advantages?!

AC: Not necessarily, unfortunately, for a reason that is easy to grasp: what makes P_n^0 a good estimator of P_0 does not necessarily make $\vartheta(P_n^0)$ a good estimator of $\vartheta(P_0)$.

ID: I think I understand. But to make sure of it, I would like you to help me weave a metaphor. Rembrandt separately calls in two equally talented apprentices in his workshop. He tells the first one: "Learn how to paint like me," and the second one: "Learn how to paint hands like me." After a few weeks, the two apprentices appear before Rembrandt and declare their apprenticeship is over. Rembrandt then asks them to paint a hand like him. The hand painted by the second apprentice is more convincing than the hand painted by the first one.

AC: Rembrandt indeed penalized the first apprentice when he failed to tell him that he was interested in hands.

JCT: I like your metaphor, which I am going to use to build an analogy. Rembrandt's style, *i.e.* his ability to paint, is like the law P_0 , a highly complex object. Learning Rembrandt's style may be seen as estimating P_0 on the basis of the observation of the master's paintings, hence the acquisition by the first apprentice of a style P_n^0 approaching P_0 in all its complexity. Painting in the manner of Rembrandt then consists, for the first apprentice, in producing a painting under the law P_n^0 .

AC: In the same vein, and a little bit mischievously, although I am a poorly talented apprentice, I would know how to paint in the manner of Rembrandt under the empirical measure P_n ! I would merely need to draw one of his paintings at random and present it to him as is.

JCT: Rembrandt's style restricted to the representation of hands, which is only a fraction of his style, is of a much lesser complexity. I see it as a feature $\vartheta(P_0)$ of the style P_0 . When the first apprentice paints hands, he does it under $\vartheta(P_n^0)$.

AC: Meanwhile, the first apprentice, who now knows that it is hands he needs to paint for Rembrandt, goes back to work. He adapts his initial style, P_n^0 , into a style oriented toward hand production, which I note P_n^1 . Well, we can imagine that the hands he now paints under $\vartheta(P_n^1)$ surpass those he painted under $\vartheta(P_n^0)$, and maybe surpass even those the second apprentice paints, because he fathomed out the master's style as a whole.

13 Targeted inference – targeting

ID: What I conclude from this, going back to the excess risk, is that you know how to change the law P_n^0 into a law P_n^1 which, by targeting $\vartheta(P_0)$, makes the substitution estimator $\vartheta(P_n^1)$ a good estimator of $\vartheta(P_0)$.

AC: Absolutely! Indeed, the functional ϑ is endowed with an important property: it is pathwise differentiable.

ID: Differentiable as we would say “derivable” for a function defined on a set of real numbers?

AC: Yes, but the notion has to be extended, insofar as ϑ is a function that is defined, not on a set of real numbers, but on the set of laws $\mathcal{M}$. To do this, we consider the restrictions of ϑ to “paths” in $\mathcal{M}$, and since each point from such a path is unequivocally identified by a real number, just as a point on a road is identified by the distance that separates it from the starting point of the road, the study of the restriction of ϑ to the path pertains to the study of functions defined on a set of real numbers.

ID: And how do you do this?

AC: Formally, ϑ is differentiable if for each $P \in \mathcal{M}$, there exists a “direction” $O \mapsto \nabla \vartheta(P)(O)$ such that, whatever is the path $\{P(\varepsilon) : \varepsilon \in [-1, 1]\} \subset \mathcal{M}$ through $P = P(0)$ with a direction $O \mapsto s(O)$ at $P = P(0)$, the function $\varepsilon \mapsto \vartheta(P(\varepsilon))$ is differentiable at $\varepsilon = 0$, with a derivative equal to $P\{\nabla \vartheta(P)(O) \times s(O)\}$.

JCT: A path $\{P(\varepsilon) : \varepsilon \in [-1, 1]\} \subset \mathcal{M}$ is not an exotic object! It is nothing but a parametric model, the parameters of which are defined solely by $\varepsilon \in [-1, 1]$, and are therefore of dimension one.

ID: Is it right to consider the direction $\nabla \vartheta(P)$ as the derivative of ϑ in $P \in \mathcal{M}$?

AC: Yes it is!

ID: And what path do you then take, if you do not mind my playing on words, to exploit this property and adapt P_n^0 ?

AC: That is the question! We build a path that goes through P_n^0 , using the direction toward which $\nabla \vartheta(P_n^0)$ points.

ID: Does the updated law lie on this path?

AC: It does. We look for our adapted law P_n^1 on this path. Thus, identifying P_n^1 comes back to identifying the best parameter $\varepsilon = \varepsilon_n^0$ and to setting $P_n^1 = P_n^0(\varepsilon_n^0)$.

ID: And in what way can ε be the best?

AC: If the path points to the direction $\nabla \vartheta(P_n^0)$ in the sense of the likelihood – other choices could be made – then ε_n^0 is that value of the parameter which maximizes the likelihood along the path.

ID: Can I deduce from this that the targeted inference method is a kind of extension of the maximum likelihood inference principle?

AC: This is a good idea. This filiation is laudatory, as the maximum likelihood inference principle, coined by R.A. Fisher at the beginning of the twentieth century, has been playing a major role in statistics for nearly a century [61].

ID: I keep going. P_n^1 is associated with the substitution estimator $\vartheta(P_n^1)$: is it our final estimator of $\vartheta(P_0)$?

Blackboard 11 Illustrating the principle of targeted inference.

JCT: In the version we are presenting you, it is not the case. The procedure (*cf* Blackboard 11) has to be iterated. At the step $k \geq 1$, we determine the direction $\nabla \vartheta(P_n^k)$; we build the path $\{P_n^k(\varepsilon) : \varepsilon \in [-1, 1]\} \subset \mathcal{M}$ which points in this direction; we find the best update parameter, ε_n^k ; and finally we set $P_n^{k+1} = P_n^k(\varepsilon_n^k)$.

ID: Don't we ever stop?!

JCT: We have at our disposal a number of stopping criteria that tell us whether a potential additional adaptation is useful or not. By noting P_n^* the final recursive update of P_n^0 , we also have a substitution estimator $\vartheta(P_n^*)$, called “targeted maximum likelihood estimator” of $\vartheta(P_0)$.

AC: So, now, you are initiated to the general principle of the targeted inference procedure that was elaborated by M. van der Laan and D. Rubin in 2006 [62]. It has been extensively developed since, and applied to a whole range of statistical problems [63].

ID: Your presentation of the targeted inference procedure reminds me of Newton's method to determine the root of an equation. The procedure is similarly based on an initialization followed with a series of updates in the direction of the derivative at the current point.

AC: Your comparison is very relevant, even though the targeted inference procedure is an emanation of the theory of “estimating functions” [64, 65] characterized by the wish to produce substitution estimators. For that matter, it is with Newton's method in mind that L. Le Cam elaborated his own estimation method, said “one-step” [66].

ID: Are the same principles at work in both targeted and one-step estimation methods?

AC: It is the same framework, but the one-step estimation method proceeds with the updates directly on the estimator in the space of the parameters Θ , one single time, whereas the targeted inference method acts on the laws in the space $\mathcal{M}$, possibly iteratively, the updates in $\mathcal{M}$ inducing those in Θ by substitution.

14 Targeted inference – upsides

ID: It is all very well, but what purpose does it serve? What are the good properties of $\vartheta(P_n^*)$?

JCT: The statistician expects *a minima* a consistency result and a central limit theorem, so that he or she might build confidence intervals, under hypotheses as weak as possible.

AC: Regarding consistency, it turns out, unsurprisingly, that $\vartheta(P_n^*)$ converges toward $\vartheta(P_0)$ when $\Delta_n^*(W) = P_n^*(Y = 1|A = 1, W) - P_n^*(Y = 1|A = 0, W)$ is a consistent estimator of $\Delta_0(W)$.

ID: I am not surprised, insofar as $\vartheta(P_n^*) = P_n^*\{\Delta_n^*(W)\}$ and as I trust the estimation via P_n^* of the marginal law of W under P_0 , which is not a difficult task. But what surprise do you have in store for me? I can see your eyes sparkling!

AC: Well, $\vartheta(P_n^*)$ also converges toward $\vartheta(P_0)$ when $P_n^*(A = 1|W)$ is a consistent estimator of $P_0(A = 1|W)$.

JCT: Even if $\Delta_n^*(W)$ is not a consistent estimator of $\Delta_0(W)$, for instance if the model used to build Δ_n^* is mis-specified?

AC: Yes, even if the model is mis-specified, *i.e.* Isabelle, even if the model does not contain the truth, $\Delta_0(W)$.

ID: This is very strange. You're telling us that the unfortunate choice of a mis-specified model, *i.e.* of a model reflecting poorly how nature works, for the estimation of Δ_n^* , can be counterbalanced by a clever choice of a well-specified model for the estimation of the conditional law of A given W ?

AC: This is indeed what I am saying.

ID: This comes as a shock. After all, the conditional law of A given W does not appear in the definition of ϑ , which only involves the conditional law of Y given (A, W) and the marginal law of W ...

AC: This remarkable property is called “double robustness.” It is not that surprising once we notice that $\vartheta(P)$ can also be written $\vartheta(P) = P\{AY/P(A = 1|W) - (1 - A)Y/P(A = 0|W)\}$!

JCT: As for the central limit theorem, the estimator $\vartheta(P_n^*)$ satisfies one under a set of conditions that are notably expressed in terms of rates of convergence of $\Delta_n^*(W)$ and of $P_n^*(A = 1|W)$ toward their respective limits. Heuristically, at least one of these limits has to coincide with $\Delta_0(W)$ or $P_0(A = 1|W)$, and the product of the convergence rates has to be in $\sqrt{n}$.

AC: You forget to specify that if ever there is a coincidence for each one of the limits, then $\vartheta(P_n^*)$ is efficient: it has the smallest possible asymptotic variance, and therefore the corresponding confidence intervals are as narrow as possible. Under less restrictive hypotheses, we know how to build a conservative estimator of the asymptotic variance of $\vartheta(P_n^*)$, *i.e.* an estimator that is going to over-estimate the true variance limit, thus giving slightly too extended, but valid confidence intervals.

ID: You evoke hypotheses that are favorable to you. Can we verify them, or do we at the very least have good reasons to think that they are satisfied?!

JCT: Ouch! You confront us to a dilemma. We could determine a subset $\mathcal{M}_0 \subset \mathcal{M}$ of laws such that, if $P_0 \in \mathcal{M}_0$, then our hypotheses would be satisfied provided P_n^0 is built on the basis of adequate statistical procedures. But we do not know whether P_0 is an element of $\mathcal{M}_0$ or not...

AC: Another point of view: if I specified the nature of the statistical procedures that govern the initial construction of P_n^0 ...

JCT: ... then I could maliciously determine a law P_0 such that the hypotheses would not be satisfied if P_0 turned out to be the law of nature.

ID: By what standard do we then judge your hypotheses?

AC: By the size of the set $\mathcal{M}_0$ Jean-Christophe evoked earlier: the bigger it is, the less restrictive the hypotheses are, and the more convincing the result is.

ID: Please forgive my question, which may be naive, but what can you say in the cases in which your hypotheses are not satisfied?

JCT: Far from being naive, your question is terribly difficult. I tend to believe that simulation studies are a way to explore that hostile territory in which our hypotheses are not satisfied.

ID: But when you simulate, are you not simulating?! Said differently, do you disguise one act under the appearance of another, or do you artificially reproduce a real situation with a view to demonstrating or explaining something?!

JCT: The second one, of course. A simulation study consists in building a synthetic law $P_0 \in \mathcal{M}$, all the features of which are under control, and the purpose of which is to imitate nature. In particular, we know the value of $\vartheta(P_0)$. We can also draw from P_0 virtual data sets of any size n .

AC: The first advantage of such a study is to check that the inference method is adequately implemented. Its second advantage is illustrative, since we can see that when the hypotheses are satisfied, then the estimator is endowed with the expected properties.

JCT: Finally, to answer your question, it gives us an idea of what happens in the cases in which the hypotheses are not satisfied.

ID: What would this be?

JCT: All the results we evoked are asymptotic. Thus, the simulation study throws light on the behavior of the finite horizon estimator, *i.e.* for values of n that may be small.

AC: Or else, the simulation study allows us to understand better what happens when the hypotheses are only slightly violated. All this opens up fascinating theoretical horizons, both computational and applicative.

15 Epilogue

ID: My dear friends, that really was an intense and rich conversation. I must say my former understanding on statistics is dramatically revigorated.

JCT: And may we ask what your renewed view consists in?

ID: Let me gather my thoughts... Well, first and foremost, I am in awe of the way you statisticians translate the question of interest into a finite-dimensional feature of a possibly infinite-dimensional law representing how nature produces the data. By decoupling the definition of the parameter to target from the choice of a model, which I understand now are two different tasks where I used to see one only, you leave room for the honest construction of a model including real knowledge and nothing more, and for the use of a commonsense parameter. I reckon this is how things should be done, and TMLE is built on this very idea.

AC: TMLE is more than an inference procedure. It is an integrative approach.

ID: Speaking of decoupling, I adhere to this two-step procedure that consists in, first, making a statistical model and, second, extending it to a causal one, possibly at the price of untestable assumptions. In the extended model, the carefully designed parameter of interest can be interpreted causally. The interpretation may collapse if the assumptions are wrong, yet the statistical parameter keeps making sense.

JCT: Thus, we are not tilting at windmills! All our efforts were worth being undertaken.

ID: Precisely! And, this is my last discovery, I now realize the importance of consistency, valid confidence intervals, and efficiency. They are not abstract theoretical concepts, but very concrete notions that impact your practice. TMLE is exemplary in this respect too.

JCT: And you, Antoine, you look like you wish to say something!

AC: I agree with Isabelle on TMLE. Furthermore, I came to an important personal conclusion.

JCT: Would you like to share it with us?

AC: Absolutely. I have been wondering for some time if I had to commit myself to a specific philosophical stance on determinism or on the nature of causality. Our discussion licenses me not to.

JCT: Why would you have to commit yourself?

AC: Because it seemed to me that many philosophers tend to think that the way you seek for causality necessarily marries you to a philosophical school. This never convinced me. I feel entitled to conduct both observational and experimental statistical studies of causal questions, depending on the data and related scientific questions. What matters most is the scientific effort and energy that we put in the battle.

ID: And we strengthened your conviction?

AC: Indeed. To use Isabelle's word, it is now clearer to me that the methodological and philosophical issues are largely decoupled. The philosopher of science can shed light on where I come from, on the roots of my practice but, I think more importantly, he or she opens up my scientific understanding of the world, and of statistics.

ID: Have you seen how late it is?! It is high time, I should think, we end our discussion.

AC: Without drawing any conclusion?!

JCT: Let me cite Lucretius [2]

For it is hard to declare for certain which of these causes it is in this world; but what can happen and does happen through the universe in the diverse worlds, fashioned on diverse plans, that is what I teach, and go on to set forth many causes for the motions of the stars, which may exist throughout the universe; and of these it needs be one which in our world too gives strength to the motions of the heavenly signs; but to affirm which of them it is, is in no wise the task of one treading forward step by step.

Does this conclusion suit you?

ID: It is certainly difficult to satisfy our curiosity, but it constitutes a formidable driving force, just as sharing the results we obtain drives us forward.

AC: I totally agree. In this respect, we can draw from our long conversation a *vademecum* to set about satisfying our curiosity with rigor.

JCT: Yes, and it is necessary to always go forward without forgetting past contributions. Which is, I believe, a nice conclusion.

ID: Farewell, my friends, good evening and good night to you.

AC: You are joking; but you will dream on your pillows about this conversation, and it will take on substance there.

ID: Well, I shall go to bed a curious soul, and a curious soul I shall arise.

Glossary

Almost-sure. An event is almost-sure if its probability equals 1.

Example. If $O \sim N(0, 1)$ then the event $O \neq 0$ is almost-sure. Yet, by symmetry, the mass of the standard Gaussian law $N(0, 1)$ concentrates around 0: whatever is the length $\ell > 0$, the interval which has the largest probability to contain O is the interval $[-\ell/2, \ell/2]$ centered at 0 with radius $\ell/2$.

Example. The estimator ϑ_n of $\vartheta(P)$ is strongly consistent if the event $\lim_{n \rightarrow \infty} |\vartheta_n - \vartheta(P)| = 0$ is almost-sure for P .

Bernoulli law. The random variable A is drawn from the Bernoulli law with parameter $p \in [0, 1]$ if A can only take the values 0 and 1, in such a way that $A = 1$ with probability p (and, therefore, $A = 0$ with probability $1 - p$).

Central limit theorem. A central limit theorem is a theorem providing assumptions which guarantee that sums of a large number of random variables behave like a Gaussian random variable. Typically, if $O_1, \dots, O_n$ are real-valued and independent random variables such that $P\{O_i\} = 0$ for each $i \leq n$ and $\sum_{i=1}^n P\{O_i^2\} = 1$, and if moreover no O_i contributes too heavily to the sum, then $\sum_{i=1}^n O_i$ approximately follows the standard Gaussian law $N(0, 1)$.

Conditional independence. Consider a collection $\{O_i : i \in I\}$ of random variables indexed by I . Let $I_1, I_2, I_3 \subset I$ be subsets of I . We say that $\mathcal{O}_1 = \{O_i : i \in I_1\}$ is conditionally independent from $\mathcal{O}_2 = \{O_i : i \in I_2\}$ given $\mathcal{O}_3 = \{O_i : i \in I_3\}$ if the joint conditional law of $(\mathcal{O}_1, \mathcal{O}_2)$ given $\mathcal{O}_3$ is the product of the two conditional laws of $\mathcal{O}_1$ and $\mathcal{O}_2$ given $\mathcal{O}_3$. If $I_3 = \emptyset$ is empty, so that $\mathcal{O}_3 = \emptyset$ too, then conditional independence coincides with independence, which does not hold in general.

Conditional law. Consider a random variable $O \sim P$ which decomposes as $O = (W, Y)$. The conditional law of Y given W is the law of the random variable Y when the realization of W is given (known).

Example. $W \in [0, 1]$ and Y is drawn from the Bernoulli law with parameter $1/3$ if $W \leq 1/2$ and $3/5$ if $W > 1/2$.

Conditionally on. See “conditional law” and “conditional independence”.

Confidence interval. A confidence interval for $\vartheta(P)$ with level $(1 - \alpha) \in [0, 1]$ is a random interval, whose construction is based on n observations drawn from the law P , in order to contain $\vartheta(P)$ with probability at least $(1 - \alpha)$. At a fixed level $(1 - \alpha)$, (i) the better of two confidence intervals is the narrower, and (ii) the larger the number n of observations used to build a confidence interval the narrower it is. When the level increases, the resulting confidence interval gets wider. Confidence intervals are often built by using an estimator ϑ_n of $\vartheta(P)$ as a pivot, i.e. under the form $[\vartheta_n - c_n, \vartheta_n + c_n]$ for a well-chosen, possibly random, half-length c_n .

Confounding. The relationship between two variables is subject to confounding, or confounded, whenever their probabilistic dependence, possibly conditioned on a third variable, cannot be interpreted causally.

Consistent (estimator). Consistency is an asymptotic notion: an estimator ϑ_n of $\vartheta(P)$ is consistent if it converges in some sense to $\vartheta(P)$ when the number n of observations upon which its construction relies goes to infinity. The estimator is weakly consistent if, for every fixed error $\varepsilon > 0$, the probability that ϑ_n be at least ε -away from $\vartheta(P)$ goes to 0 when n goes to infinity: $\lim_{n \rightarrow \infty} P\{|\vartheta_n - \vartheta(P)| \geq \varepsilon\} = 0$. It is strongly consistent if ϑ_n converges to $\vartheta(P)$ almost-surely: $P\{\lim_{n \rightarrow \infty} |\vartheta_n - \vartheta(P)| = 0\} = 1$. Strong consistency implies weak consistency, but the reverse is not true.

Contingency table. A contingency table, term coined by K. Pearson in 1904, is a two (or more)-entry table where one reports the frequencies associated with two (or more) categorical variables of interest. The origin of contingency tables goes back to the research conducted by P.C.A. Louis to demonstrate the therapeutic inefficacy of bloodletting [67].

Example. Consider $O_1, \dots, O_n$ $n = 50$ variables such that each O_i contains $(A_i, Y_i) \in \{0, 1\}^2$. The following contingency table

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

teaches us that, among these n observations, 18 (respectively, 12, 7, and 13) feature a couple (A_i, Y_i) equal to $(1, 1)$ (respectively, $(1, 0)$, $(0, 1)$, and $(0, 0)$).

Correlation coefficient. The correlation coefficient of two real-valued random variable is a measure of their probabilistic dependence on a linear scale. If X and Y are independent then their correlation coefficient equals 0. The reverse is not true.

Empirical measure. Given n observations $O_1, \dots, O_n$, the empirical measure is the law P_n such that, if $O \sim P_n$ is drawn from P_n then $O = O_i$ with probability n^{-1} for each $1 \leq i \leq n$.

Estimator. An estimator is a random variable obtained by combining the observations yielded by an experiment for the sake of estimating a feature of interest of the experiment.

Example. Consider $O_1, \dots, O_n$ independent random variables drawn from a common law P . The empirical mean $n^{-1} \sum_{i=1}^n O_i$ is an estimator of the mean $\vartheta(P) = P\{O\}$ of $O \sim P$. If O is real-valued and if $P\{|O|\}$ is finite then the empirical mean is a strongly consistent estimator (by the strong law of large numbers).

Feature. See “parameter”.

Gaussian law. The real-valued random variable O is drawn from the standard Gaussian law $N(0, 1)$ if for all $a \leq b$, the probability that $O \in [a, b]$ equals the area under the Gauss curve of equation $t \mapsto \sqrt{2\pi}^{-1} \exp(-t^2/2)$. This law is particularly important because it naturally appears as a limit law of sequences of experiments in theorems referred to as “central limit theorems”.

Independence. Consider a collection $\{O_i : i \in I\}$ of random variables indexed by I . Let $I_1, I_2 \subset I$ be two subsets of I . We say that $\mathcal{O}_1 = \{O_i : i \in I_1\}$ is independent from $\mathcal{O}_2 = \{O_i : i \in I_2\}$ if the values of the realizations of the first set are not influenced by the values of the realizations of the second set. More formally, $\mathcal{O}_1$ is independent from $\mathcal{O}_2$ if the joint law of $(\mathcal{O}_1, \mathcal{O}_2)$ is the product of the two marginal laws of $\mathcal{O}_1$ and $\mathcal{O}_2$, or, equivalently, if the conditional law of $\mathcal{O}_2$ given $\mathcal{O}_1$ coincides with the marginal law of $\mathcal{O}_2$, and the other way around.

Example. Let $(W, Y) \in \{0, 1\}^2$ be such that $P(W = Y = 1) = 1/10$, $P(W = 1, Y = 0) = 1/15$, $P(W = 0, Y = 1) = 1/2$, and $P(W = Y = 0) = 1/3$.

The marginal law of Y is the Bernoulli law with parameter $P(Y = 1) = P(Y = 1 \text{ and } (W = 1 \text{ or } W = 0)) = P(W = Y = 1) + P(W = 0, Y = 1) = 3/5$. The marginal law of W is the Bernoulli law with parameter $P(W = 1) = P(W = 1 \text{ and } (Y = 1 \text{ or } Y = 0)) = P(W = Y = 1) + P(W = 1, Y = 0) = 1/6$.

Note that $P(W = 1)P(Y = 1) = 1/10 = P(W = Y = 1)$, $P(W = 1)P(Y = 0) = 1/15 = P(W = 1, Y = 0)$, $P(W = 0)P(Y = 1) = 3/5 = P(W = 0, Y = 1)$ and $P(W = 0)P(Y = 0) = 1/3 = P(W = Y = 0)$. Thus, W and Y are independent under P .

Inference. Statistical inference is the process of drawing inferences from data. Statistical inference relies on mathematical procedures developed in the framework of the theory of statistics, which builds upon the theory of probability, for the sake of analyzing the structure of a random experiment based on its observation. The analysis is typically expressed in terms of pointwise or confidence-interval-based estimation, or hypotheses testing, or regression.

Joint law. Consider a random variable O which decomposes as $O = (W, Y)$. The joint law of O is the law of the couple (W, Y) .

Law. The law P of a random variable O is the exhaustive description of how chance produces a realization of O . We note $O \sim P$ to indicate that O is drawn from the law P .

Law of large numbers. A law of large numbers is a probabilistic theorem providing assumptions which guarantee that the empirical mean $n^{-1} \sum_{i=1}^n O_i$ of n random variables $O_1, \dots, O_n$ sharing a common law

P converges to their common mean $P\{O\}$. We say that such a law is “weak” if the convergence takes place in probability, *i.e.* if whatever is a fixed margin of error, the probability that the gap separating the empirical mean and its theoretical counterpart exceed this error goes to 0 when n goes to infinity. It was J. Bernoulli who first formalized this law, back in 1690. A weak law of large numbers notably holds when the random variables $O_1, \dots, O_n$ are real-valued, independent, and such that $P\{|O|\}$ be finite.

We say that such a law is “strong” if the convergence takes place almost surely, *i.e.* if there is a 1-probability that the empirical mean converge to its theoretical counterpart when n goes to infinity. If a strong law holds then a weak law necessarily holds true. The reverse is not true. A. Kolmogorov proved in 1929 that a strong law of large numbers notably holds when $O_1, \dots, O_n$ are real-valued, independent, and such that $P\{|O|\}$ be finite.

For B. Gnedenko and A. Kolmogorov [68],

In fact, all epistemologic value of the theory of probability is based on this: that large-scale random phenomena, in their collective action, create strict non-random regularity.

Likelihood. The likelihood of an observation O under a law P susceptible to produce O quantifies how likely it is that O be actually drawn from P . The more likely it is, the larger the likelihood. The maximum likelihood principle builds upon this interpretation: given two laws of identical complexity, both susceptible to produce the observation O , one must prefer that law which maximizes the likelihood. If the two laws have differing complexities then the comparison of their likelihoods requires a preliminary adjustment based on a parsimony principle, the more complex law being naturally advantaged over the simpler law.

Marginal law. Consider a random variable $O \sim P$ which decomposes as $O = (W, Y)$. The marginal law of Y is the law of the random variable Y extracted from O . This expression originates from the vocabulary of contingency tables.

Example. Let $(W, Y) \in \{0, 1\}^2$ be such that $P(W = Y = 1) = 1/10$, $P(W = 1, Y = 0) = 1/5$, $P(W = 0, Y = 1) = 3/10$, and $P(W = Y = 0) = 2/5$. Then $P(W = 1) = P(W = 1 \text{ and } (Y = 1 \text{ or } Y = 0)) = P(W = Y = 1) + P(W = 1, Y = 0) = 3/10$, so the marginal law of W is the Bernoulli law with parameter $3/10$. Likewise, $P(Y = 1) = P(Y = 1 \text{ and } (W = 1 \text{ or } W = 0)) = P(W = Y = 1) + P(W = 0, Y = 1) = 2/5$, so the marginal law of Y is the Bernoulli law with parameter $2/5$.

Model. A model is a collection of laws from which the observation O may be drawn. A model is said parametric if its elements are identified by a finite-dimensional parameter.

Example. Let M be the non-parametric model consisting of all laws compatible with the definition of the observation O . A subset $\{P(\varepsilon) : \varepsilon \in [-1, 1]\} \subset \mathcal{M}$ of candidate laws $P(\varepsilon)$ identified by the real parameter ε is a parametric model. Since it is one-dimensional, it is often called a “path”.

Parameter. Value of a functional defined upon a model and evaluated at a law belonging to that model.

Example. $\theta(\mathbb{P})$ for $\theta : \mathbb{M} \rightarrow \Theta$ or $\vartheta(P)$ for $\vartheta : \mathcal{M} \rightarrow \Theta$.

Random variable. Description, possibly non-exhaustive, of the result of a random experiment, *i.e.* of a reproducible experiment subject to chance.

Example. The experiment consisting in flipping a balanced coin in a well-defined experimental setting is a random experiment (it is reproducible and we cannot be certain of its outcome). The result of each coin toss is described by the random variable which takes the value 1 for tail and 0 otherwise. The law of this random variable is the Bernoulli law with parameter $1/2$.

Regression. Given observations $O_1, \dots, O_n$ of a generic data $O = (W, Y)$, regressing Y on W consists in inferring from the observations information on how Y depends on W . Typically, regressing Y on W means explaining the mean of the random variable Y conditionally on W , *i.e.* expressing the mean of Y as a function of W .

Example. If $Y \in \{0, 1\}$ then regressing Y on W amounts to estimating the conditional probability $P(Y = 1|W)$ that Y equal 1 given the value of W . This example is an instance of regression in the

aforementioned typical sense since $P(Y = 1|W)$ coincides with the conditional mean $P\{Y|W\}$ of Y given W .

Substitution estimator. Given a functional $\vartheta : \mathcal{M} \rightarrow \Theta$ of interest, an estimator ϑ_n of the parameter $\vartheta(P)$ is a substitution estimator if it writes as $\vartheta_n = \vartheta(P_n)$ for a law P_n approaching P .

Example. Consider $\vartheta : \mathcal{M} \rightarrow \mathbb{R}$ such that $\vartheta(P) = P\{O\}$ where $\mathcal{M}$ is a set of laws which all admit a finite mean. Let $O_1, \dots, O_n$ be independent random variables with a common law P and let P_n be the empirical measure. The empirical mean $n^{-1} \sum_{i=1}^n O_i = \vartheta(P_n)$ is a substitution estimator of the mean $\vartheta(P)$.

Uniform law. The real-valued random variable O is drawn from the uniform law on $[A, B]$ if for all $A \leq a \leq b \leq B$, the probability that $O \in [a, b]$ equals the ratio $(b - a)/(B - A)$.

Acknowledgments: The authors wish to thank warmly K. Debbasch (MdL, Université Paris Descartes) for her help in translating the original French trialogue into English. The project benefited from the support of Paris Descartes University ATP “Biomathematics, Biostatistics, System Biology”. The authors are grateful for the reviewers’ comments that led to an improved text.

References

1. Lembke J. Virgil's Georgics, a new verse translation, Second book. New Haven, CT: Yale University Press, 2005.
2. Lucretius CT, Smith MF, Rouse WH. Lucretius, De rerum natura. Cambridge: Harvard University Press, 1982.
3. Gregory A, Waterfield R. Plato: Timaeus and Critias. New York: Oxford University Press, 2008.
4. Mallarmé S. Un coup de dé jamais n'abolira le hasard. Nouvelle Revue Française, 1914. Available at: <http://gallica.bnf.fr/ark:/12148/bpt6k71351c>.
5. Diderot D. Entretien entre d'alembert et diderot. Paris: Correspondance littéraire, 1782.
6. Diderot D, Kemp J, Stewart J. Diderot, interpreter of nature: selected writings. Westport: Hyperion Press. International Publishers, 1963.
7. Starman RJ. Models, inference, and truth: probabilistic reasoning in the information era. In: van der Laan MJ, Rose S, editors. Targeted learning. Springer series in statistics. New York: Springer, 2011:li–lxxi.
8. Hacking I. The taming of chance, vol. 17. Cambridge: Cambridge University Press, 1990.
9. Hume D. A treatise of human nature. London: John Noon, 1739.
10. Stigler SM. Karl Pearson's theoretical errors and the advances they inspired. Stat Sci 2008;23:261–71.
11. Russell B. On the notion of cause. In: Proceedings of the Aristotelian society, vol.13, 1912:1–26.
12. Pearson K. The grammar of science, vol. 20. London: W. Scott, 1892.
13. Einstein A, Podolsky B, Rosen N. Can quantum-mechanical description of physical reality be considered complete? Phys. Rev 1935;47:777–80.
14. Bell JS. On the Einstein-Podolsky-Rosen paradox. Physics 1964;1(3):195–200.
15. Aspect A. Bell's inequality test: more ideal than ever. Nature 1999;398:189–90.
16. Mill JS. System of logic: ratiocinative and inductive, being a connected view of the principles of evidence and the methods of scientific investigation, London: Longmans, Green and Co., 1843.
17. Mackie JL. The cement of the universe. Oxford: Clarendon Press, 1974.
18. Rothman KJ. Causes. Am J Epidemiol 1976;104:587–92.
19. Rothman KJ, Greenland S. Modern epidemiology, 2nd ed. Philadelphia, PA: Lippincott-Raven, 1998.
20. Koch R. Die aetiologie der tuberculose. Mitt Kaiser Gesundh 1884;2:1–88.
21. Koch R. Ueber bakteriologische forschung. In: Verhandlungen des X. Internationalen Medicinischen Congresses, Berlin, 1890, 1892:35.
22. Fredericks DN, Relman DA. Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. Clin Microbiol Rev 1996;9:18–33.
23. Hill AB. The environment and disease: association or causation? Proc R Soc Med 1965;58:295–300.
24. Doll R, Hill AB. Smoking and carcinoma of the lung. Bull World Health Organ 1999;77:84–93.
25. Cartwright N. Causal laws and effective strategies. Noûs 1979;13:419–37.
26. Good IJ. A causal calculus (I). Br J Philos Sci 1961;XI:305–18.
27. Good IJ. A causal calculus (II). Br J Philos Sci 1961;XII:43–51.
28. Skyrms B. Causal necessity: a pragmatic investigation of the necessity of laws. New-Haven, Londres: Yale University Press, 1980.

29. Suppes P. A probabilistic theory of causality. Amsterdam: North Holland Publishing Company, 1970.
30. Dawid AP. Causal inference without counterfactuals. *J Am Stat Assoc* 2000;95:407–24.
31. Freedman DA. Statistical models and shoe leather. *Sociol Methodol* 1991;21:291–313.
32. Freedman DA. Statistical models: theory and practice. Cambridge: Cambridge University Press, 2005.
33. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;10:37–48.
34. Holland PW. Causal inference, path analysis, and recursive structural equations models. *Sociol Methodol* 1988;18:449–84.
35. Pearl J. Causality: models, reasoning and inference, vol. 29. Cambridge: Cambridge University Press, 2000.
36. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Math Model* 1986;7:1393–512.
37. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974;66:688.
38. Haavelmo T. The statistical implications of a system of simultaneous equations. *Econometrica, J Econometric Soc* 1943;11(1):1–12.
39. Wright S. Correlation and causation. *J Agric Res* 1921;20:557–85.
40. Bernard C. Introduction à l'étude de la médecine expérimentale. Paris: Champs-Flammarion, 1865.
41. Woodward J. Making things happen: a theory of causal explanation. New York: Oxford University Press, 2003.
42. Jost A, Vigier B, Prépain J, Perchellet J-P. Studies on sex differentiation in mammals. *Recent Prog Horm Res* 1973;29:1–41.
43. Wolff E. Demonstration of a feminizing action of the right gonad in the female avian embryo by hemicastration experiments. *C. R. Séances Soc. Biol. Fil* 1951;145:1218–19.
44. Camerino G, Goodfellow P. A fragile understanding. *Trends Genet* 1991;7:239–40.
45. Camerino G, Parma P, Radi O, Valentini S. Sex determination and sex reversal. *Curr Opin Genet Dev* 2006;16:289–92.
46. Goodfellow PN, Camerino G. DAX-1, an “antitestis” gene. *Cell Mol Life Sci* 1999;55:857–63.
47. Goodfellow PN, Camerino G. DAX-1, an “antitestis” gene. In: Scherer G, Schmid M, editors. *Genes and mechanisms in vertebrate sex determination. Experientia supplementum*, vol. 91. Basel: Birkhäuser, 2001:57–69.
48. Conan Doyle AI. The sign of the four. Philadelphia: Lippincott's Monthly Magazine, 1890.
49. Marshall BJ, Warren JR. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* 1984;323:1311–15.
50. Warren JR, Marshall BJ. Unidentified curved bacilli on gastric epithelium in active chronic gastritis. *Lancet* 1983;321:1273–5.
51. Kistler M. The interventionist account of causation and non-causal association laws. *Erkenntnis* 2013;78:65–84.
52. Académie des sciences. Rapport de l'académie des sciences. Séance du 5 octobre 1835:173.
53. Gavarret J. Principes généraux de statistique médicale, ou développement des règles qui doivent présider à leur emploi. Paris: Béchot jeune et Labé, 1840.
54. Valleix FL. De l'application de la statistique à la médecine. In: Valleix, F. L. I. and Laségue, C. and Follin, E. editors. *Archives générales de médecine*, vol. 3. Paris: Béchot jeune et Labé, 1840:5–39.
55. d'Alembert JR. Opusculs mathématiques. Tome 2, Onzième mémoire, Sur l'application du calcul des probabilités à l'inoculation de la petite vérole, 1761–1780:34.
56. Yu Z, van der Laan MJ. Construction of counterfactuals and the G-computation formula. Technical report 122, U.C. Berkeley Division of Biostatistics, 2002.
57. VanderWeele TJ, Shpitser I. On the definition of a confounder. *Ann Stat* 2013;41:196–220.
58. Cartwright N. The art of medicine. A philosopher's view of the long road from RCTs to effectiveness. *Lancet* 2011;377:1400–01.
59. Simpson EH. The interpretation of interaction in contingency tables. *J R Stat Soc Ser B* 1951;13:238–41.
60. d'Ockham G. Quaestiones et decisiones in quatuor libros sententiarum cum centilogio theologico. Livre 2, 1319.
61. Aldritch J. R.A. fisher and the making of maximum likelihood 1912–1922. *Stat Sci* 1997;12:162–76.
62. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat* 2006;2:Article 11, 40.
63. van der Laan MJ, Rose S. Targeted learning. New York: Springer, 2011.
64. van der Laan MJ, Robins JM. Unified methods for censored longitudinal data and causality. New York: Springer, 2003.
65. van der Vaart AW. Asymptotic statistics, vol. 3. Cambridge: Cambridge University Press, 1998.
66. Le Cam L, Yang GL. Asymptotics in statistics. New York: Springer, 2000.
67. Louis PC. Recherches sur les effets de la saignée dans quelques maladies inflammatoires et sur l'action de l'émétique et des vésicatoires dans la pneumonie. Paris: J.-B. Baillière, 1835.
68. Gnedenko BV, Kolmogorov AN. Limit distributions for sums of independent random variables. Cambridge, MA: Addison-Wesley Publishing Company, 1954.