

The causal foundations of applied probability and statistics

Sander Greenland
Department of Epidemiology and Department of Statistics,
University of California, Los Angeles
lesdomes@ucla.edu

1 November 2020

Abstract. Statistical science (as opposed to mathematical statistics) involves far more than probability theory, for it requires realistic causal models of data generators – even for purely descriptive goals. Statistical decision theory requires more causality: Rational decisions are actions taken to minimize costs while maximizing benefits, and thus require explication of causes of loss and gain. Competent statistical practice thus integrates logic, context, and probability into scientific inference and decision using narratives filled with causality. This reality was seen and accounted for intuitively by the founders of modern statistics, but was not well recognized in the ensuing statistical theory (which focused instead on the causally inert properties of probability measures). Nonetheless, both statistical foundations and basic statistics can and should be taught using formal causal models. The causal view of statistical science fits within a broader information-processing framework which illuminates and unifies frequentist, Bayesian, and related probability-based foundations of statistics. Causality theory can thus be seen as a key component connecting computation to contextual information, not “extra-statistical” but instead essential for sound statistical training and applications.

Acknowledgements: I am grateful to Steve Cole, Joseph Halpern, Jay Kaufman, Blakeley McShane, and Sherrilyn Roush for their helpful comments on the drafts.

The only immediate utility of all the sciences is to teach us how to control and regulate future events through their causes. – Hume [1748]

Introduction: Scientific Inference is a Branch of Causality Theory

I will argue that realistic and thus scientifically relevant statistical theory is best viewed as a subdomain of causality theory, not a separate entity or an extension of probability. In particular, the application of statistics (and indeed most technology) must deal with causation if it is to represent adequately the underlying reality of how we came to observe what was seen – that is, the causal network leading to the data.¹ The network we deploy for analysis incorporates whatever time-order and independence assumptions we use for interpreting observed associations, whether those assumptions are derived from background (contextual) or design information [Pearl 1995, 2009; Robins 2001]. In making this case, I will invoke Pearl’s own arguments (e.g., as in Pearl [2009], Wasserstein [2018]) to deduce that statistics should integrate causal networks into its basic teachings and indeed into its entire theory, starting with the probability and bias models that are used to build up statistical methods and interpret their outputs.

Every real data analysis has a causal component comprising the causal network assumed to have created the data set. Decision analysis has a further causal component showing the effects of decisions. Although these causal components are usually left implicit, a primary purpose of design strategies is to rule out alternative causal explanations for observations. Consider one of the most advanced research projects of all time, the search for the Higgs boson. Almost all statistical attention focused on the one-sided 5-sigma detection criterion [Lamb 2012], roughly equivalent to an α -level of 0.0000003, or requiring at least $-\log_2(0.0000003) = 22$ bits of information against the null [Greenland 2019] to declare detection. Yet the causal component is

¹This view arguably applies even when dealing with quantum phenomena, at least in the Qbist view [Mermin 2016]. In that view, the laws of quantum mechanics describe how equipment settings causally affect individual perceptions, where the latter become formalized as coherent predictive bets or frequency claims about subsequent observations under those settings (in contrast to other theories that treat quantum probabilities as properties of the environment). Such a controversial view is however unnecessary for the everyday applications of probability and causation that typify most of science and technology, so will not be pursued here.

just as important: It includes every attempt to eliminate explanations for such extreme deviations other than the Higgs boson, e.g., the painstaking checks of equipment are actions taken to block the mechanisms that could cause anything near that deviation (other than the Higgs mechanism itself).

Thus, because statistical analyses need a causal skeleton to connect to the world, causality is not extra-statistical but instead is a logical antecedent of real-world inferences. Claims of random or “ignorable” or “unbiased” sampling or allocation are justified by causal actions to block (“control”) unwanted causal effects on the sample patterns. Without such actions of causal blocking, independence can only be treated as a subjective exchangeability assumption whose justification requires detailed contextual information about absence of factors capable of causally influencing both selection (including selection for treatment) and outcomes [Greenland 1990]. Otherwise it is essential to consider pathways for the causation of bias (nonrandom, systematic errors) [Pearl 1995; Greenland et al. 1999; Maclure and Schneeweiss 2001; Hernán et al. 2004; Greenland 2010a, 2012a].

The remainder of the present paper elaborates on the following points: Probability is inadequate as a foundation for applied statistics, because competent statistical practice integrates logic, context, and probability into scientific inference and decision, using narratives built around causality. Thus, given the absence of elaborated causality discussions in statistics textbooks and coursework, we should not be surprised at the widespread misuse and misinterpretation of statistical methods and results. This is why incorporation of causality into introductory statistics is needed as urgently as other far more modest yet equally resisted reforms involving shifts in labels and interpretations for P-values and interval estimates.²

As a preliminary, consider that the Merriam-Webster Online Dictionary [2019] defines statistics as “a branch of mathematics dealing with data collection, organization, analysis, interpretation and presentation.” Many working statisticians (including me) regard the “branch of mathematics” portion as abjectly wrong, akin to calling physics, computer science or any other

²Such as replacement of misleading terms like “statistical significance” and “confidence” by more modest terms like “compatibility” [Amrhein et al. 2019; Greenland 2017b, 2019; Rafi and Greenland 2020; Greenland and Rafi 2020; McShane et al. 2019; Wasserstein et al. 2019].

heavily mathematical field a branch of mathematics. But we can fix that by replacing “branch of mathematics” with “science” to obtain

Statistics is the science of data collection, organization, analysis, interpretation and presentation, often in the service of decision analysis.

The amended definition makes no explicit mention of *either* probability or causation, but it is implicitly causal throughout, describing a sequence of actions with at least partial time ordering, each of which is capable of affecting subsequent actions: Study design affects actions during data collection (e.g., restrictions on selection); these actions along with events during data collection (e.g., censoring) affect the data that result; these actions and events affect (or should affect) the study description and the data analysis; and the analysis results will affect the presentation. Overall, the presumed causal structure of this sequence supplies the basis for a justifiable interpretation of the study. Thus, whether answering the most esoteric scientific questions or the most mundane administrative ones, and whether the question is descriptive, causal, or purely predictive, causal reasoning will be crucially involved (albeit often hidden to ill effect in equations and assumptions used to get the “results”).

Causality is central even for purely descriptive goals

As Pearl has often noted, causal descriptions encode the information and goals that lead to concerns about associations [Pearl 2009]. Consider survey statistics, in which the target question is not itself causal, merely descriptive, such as the proportion of voters who would vote for a given candidate. A competent survey researcher will be concerned about what characteristics C will affect both survey participation ($S=1$) and voting intent V . Using square brackets to indicate that the observations are conditioned on $S=1$, this concern is encoded in the diagram

$$[S=1] \leftarrow C \rightarrow V,$$

in which we can see bias in the sampling estimator for the preference distribution $\Pr(V=v)$ will be induced by the selection on S . If instead we said only that the concern is about characteristics that are *associated* with both participation and preference (as in $S \leftrightarrow C \leftrightarrow V$) we would obscure the contextual basis for the concern.

To paraphrase Pearl, statistical analysis without causality is like medicine without physiology. As an example, if we see a difference in ethnic distributions (C) between our survey

and population demographic data, we should be concerned about mis-estimating (say) the proportion of Trump voters in the target population. This concern is not because “white ethnicity is *associated* with voting for Trump” as some academic descriptions would have it, but because we expect that being a white male causes sympathy (or prevents antipathy) for Trump’s pronouncements relative to being black or latino. That expectation arises from a simple causal relation encoded in $C \rightarrow S$, which creates the concern about only seeing preferences of those in the survey, i.e., seeing only $\Pr(V=v|S=1)$.

When survey methods attempt to adjust for the difference by reweighting the sample using the target-population ethnicity distribution, that adjustment can be seen as an attempt to counterbalance the $C \rightarrow S$ arrow in the mechanism generating the sample. This added computation in producing a reweighted sample is traditionally treated as a purely numeric artifice, but is also a causal process: Someone must physically obtain target-weight data and program the reweighting to create the adjusted estimate. It is misleading to describe this action as “simulating removal of an arrow”; it is instead the addition to the data generator of a weighting intervention W in a new causal pathway within

$$[S=1] \leftarrow C \rightarrow V \leftarrow W \leftarrow C$$

W is engineered to (hopefully) balance out the bias from conditioning on selection $[S=1]$. Note that C appears twice in this diagram to allow it to be written in one line; writing it twice separates the initial effect of C on voter preferences (V) and sample formation (participation S) from its later effect on the analysis weighting W .

The strength of probabilistic independence demands physical independence

By data generator, I do *not* mean some abstract structural equation, but rather the entire set of actual physical mechanisms that produce our observations. Even in the simplest games of chance, it is the *physical* (mechanical, causal) independence of coin tosses which licenses our teaching that betting systems for toss sequences will fail to beat simple expectations based on the frequency of heads observed so far. A causal diagram for a sequence of independent identically distributed (i.i.d.) tosses with outcome indicators Y_1, \dots, Y_N would thus show these N indicators

as N isolated (unconnected) nodes.³ More generally, every missing arrow implies an independence assumption, and such an assumption is really a large *set* of assumptions on the joint distribution of the data Y_1, \dots, Y_N .

One way to measure the information in or logical strength of an independence assumption is by the number of logically independent constraints it imposes (equivalent to the number of parameters whose value it specifies, or the number of dimensions or degrees of freedom it removes from further consideration). Allowing for any possible dependency pattern (as suggested by “nonparametric”) among the Y_1, \dots, Y_N yields a measure of order N factorial; even if we count only pairwise dependencies, the number of patterns is of order N^2 (see Appendix 1). Either way, when described honestly, an i.i.d. assumption is not one assumption but rather a *set* of assumptions that grows far faster than the number of observations N . The amount of deductive (digital, syntactical) information in this assumption set is thus beyond anything data frequencies alone could contain; only contextual (background and design) information can supply enough information to warrant such a large set of assumptions.

This enormous logical content of random sampling and randomization illustrates why they are such powerful investigative tools: Only the physical act of blocking all causal effects on selection or treatment can provide deductive justification for the entire set of assumptions corresponding to “independence.”

The Superconducting Supercollider of Selection

In human field studies, realistic causal diagrams should always have a selection (sampling) indicator node S as shown as part of the data-generating process. This node may be influenced by (and perhaps even influence) study variables. By definition, only those with $S=1$ are observed; thus S will always be conditioned on. If S is affected by more than one variable it will be a conditioned collider and thus a potential bias source under ordinary graphical rules [Greenland 2010a, 2012a]. Most basic causal-diagram introductions (including those I helped write) can be faulted for not emphasizing this fact. We can now fault statistics education for the same reason, in that the “ignorability” of selection under random sampling has led to

³A Bayes network would generalize this diagram to show an exchangeable sequence with a node representing the single-toss probability feeding into the Y_n .

forgettability of the physical selection mechanism in settings where it is not random in any mechanical sense and thus not ignorable in any practical sense.

An important point for graphically representing these problems is that not all of what is known as selection bias arises from S being a collider.⁴ For example, classical selection bias requires no collider in the causal graph of data collection. Consider in the earlier voting-survey graph $[S=1] \leftarrow C \rightarrow V$; the bias here corresponds to classical confounding, as it comes from an open back-door path connecting V to S via a shared cause (the causal fork at C). As with confounding, a solution is to condition (stratify) on C , which allows identification of C -conditional voter intentions.

Unlike in classical confounding, however, conditioning is only a partial solution: In the example, the goal is to recover the marginal (C -unconditional) distribution $\Pr(V=v)$ of V in the targeted S -unconditional population. Unfortunately, that V marginal is not identified if the graph is the only information available on the target population. This identification is achieved in classical demographic and epidemiologic standardization⁵ by averaging the observed C -conditionals $\Pr(V=v|C=c, S=1)$ over the C distribution of the target population, $\Pr(C=c)$; this procedure assumes however that V is independent of selection given C , so that $\Pr(V=v|C=c, S=1) = \Pr(V=v|C=c)$, as implied by $S \leftarrow C \rightarrow V$.

A parallel example of selection bias without collider bias arises in studying the effect of a treatment X on an outcome Y when C is a modifier of the treatment effect, as in

$$[S=1] \leftarrow C \rightarrow Y \leftarrow X$$

[Hernán 2017]: C is independent of treatment X , and Y is independent of selection S given C , but the $S \leftarrow C \rightarrow Y$ path still can bias the estimated marginal $X \rightarrow Y$ effect given the conditioning on selection ($S=1$); this bias would become intractable if selection (observation) affected the targeted effects (as in $S \rightarrow Y \leftarrow X$).

Data and algorithms are causes of reported results

⁴This point is contrary to Hernán et al. [2004]; see Hernán [2017] for a reconciliation.

⁵Not to be confused with “standardization” as in dividing a variable by its standard deviation, which damages comparisons of estimates both within and across studies [Greenland et al. 1986, 1991].

The causal sequence continues once the data are collected: A statistical procedure is a data-processing algorithm whose flow chart can be viewed as a causal diagram showing how each computational step determines the next. Usually, each node is a deterministic function of its parents, but may include simulations (as in bootstrap and Markov-Chain Monte-Carlo procedures) that may result in stochastic conditional branches. Finally, the outputs of the algorithm cause researchers and readers to interpret and report the study in particular ways, whether mechanically (e.g., in misreports of “no association” because a P-value exceeded 0.05) or informally, and can strongly affect whether and where the results are published.

Given the causal nature of data generation, calling causal models "extra-statistical" is a misleading characterization of both causality and statistics: Valid statistical analysis is causal to the core; hence, **realistic statistical analysis is a subset of causal analysis**. Not even "extra-distributional" is correct, because the core problem is about factors producing (causing) differences in distributions of those targeted (e.g., voters; patients with a given indication for treatment) and those observed (e.g., survey responders; patients in a trial). Without a causal model for deducing the assumed data distribution from the entire physical data generator, we have no basis for claiming our probability calculations are connected to our target or the world beyond our immediate data.

To summarize so far: Taking off from the Epilogue of Pearl [2009], statistics as conceived and practiced competently is about laying out the causal sequences leading from data to inferences (perceptions) and decisions. Within this sequence, a statistical analysis algorithm or protocol is a causal submodel for how that data will be processed into outputs. Those outputs will then be interpreted as statements connecting the target population to our data under our causally-derived sampling model, with the connections established via open paths in the causal diagram between the target and the data – including connections passing through the ever-present selection node S. Probability plays a central role in terms of formalizing the expected behaviors (propensities) of the data generator under different hypotheses; but that formalization is physically justified only when it is deduced from the causal structure of the generator.

Getting causality into statistics by putting statistics into causal terms from the start

Labeling causation as "extra-statistical" creates an excuse to continue to ignore causality theory in statistical teaching and methods research, and stay within the insufficient descriptions

of acausal probability theory as the only formal foundation of statistics. That leads to bad practice, such as confusing probabilities of group events with probabilities of individual events within a group. Examples of such confusion [Greenland and Robins 1988; Robins and Greenland 1989; Greenland et al. 2019] may help statisticians recognize causality as an essential component that distinguishes application-relevant statistical theory from acausal probability and its extensions in mathematical statistics. Again, sound applications also need detailed causal explanations of how the data were generated – including the physical mechanisms that led to being in different comparison groups and to inclusion in the data set ($S=1$).

These causal explanations provide the contextual justifications for the probability models used in the analysis, displaying information about study features that physically constrain data generation. One teaching implication is that students must master causal thinking before they can master real-world statistical inference; thus, basic logic and its causal extensions should be covered from the start of introductory statistics, *before* probability and statistics. But the curriculum for doing so is in its infancy. I used this sequencing in my UCLA courses; however, all incoming students had at least basic statistics, and most also had research methods courses in which at least informal ideas of causality were covered. Thus, the students needed retraining to remove common misconceptions about the implications (or lack thereof) of various statistical results for causal questions.

Students had no trouble mastering the idea of associations passing through causal forks (such as $X \leftarrow C \rightarrow Y$) or mediators (such as $X \rightarrow M \rightarrow Y$); in fact their entire intuition for bias and adjustment came from these two cases. On the other hand, their intuitions for paths through colliders (such as $X \rightarrow S \leftarrow Y$) were backwards, as should be no surprise: Collider bias is by definition the negative or inverse of confounding, because collider bias arises from conditioning (on colliders), whereas confounding is removed by conditioning (on shared causes). Hence, for absolute measures, confounding bias equals the unconditional association minus a conditional association, whereas collider bias equals a conditional association minus the unconditional association.

Again, this view applies not only for causal research questions but also for descriptive survey research. In all real settings in which perfection is unattainable, researchers should try to understand causes of nonresponse, loss, missing data, misreporting, and other sources of

uncertainty and inferential distortion⁶ - for example by placing these bias sources in a causal diagram to guide study design and interpretation. Only then can they begin to master the far more subtle notions of probabilistic inference from incomplete observations.

Causation in 20th-century statistics

Statistical foundation debates raged throughout the last century, but focused exclusively on prioritization of logical criteria such as internal coherence (no violations of the axioms of probability theory) versus self-calibration (meeting select frequency criteria over data sequences generated by the distribution used to derive the data-processing algorithm). Yet formal causal modeling is as old as modern statistical foundations laid down by Fisher, Neyman, DeFinetti and many others in the first half of the 20th century. Although Neyman [1923] went largely unnoticed, potential-outcome (“counterfactual”) models entered prestigious statistics journals by the 1930s, and had an ongoing presence before their broad uptake began in the 1980s (e.g. [Welch 1937, Wilk 1955, Copas 1973]). Even without such formalisms, the probability models on which statistical procedures were based were supposed to be frequency summaries of causal mechanisms with certain physical independencies built in by design; these independencies made the mechanisms “ignorable” [Rubin 1978] – a misleading in term insofar as the data-generating mechanism should always be described in detail, never ignored. Such mechanisms include random sampling, which makes selection *S* an unaffected (exogenous) node, and random allocation, which makes treatment assignment an unaffected node.

Statistical developments in the 20th century were foremost concerned with causal inferences derived from physical randomization, whether by nature as in genetic recombination, or by design. Fisher was often quite straightforward in his causal descriptions and how he regarded causal inference about treatment effects as the central goal of scientific experimentation in the life sciences. By the mid-1930s he had laid out potential outcomes clearly enough (even if only verbally) to see the distinction between the sharp null of no effect on any unit (used to derive randomization tests) and Neyman’s weak null of no effect on the mean [Greenland, 1991].

⁶These include bad research practices such as “P-hacking”: Searching out analyses that give P-values above or below a threshold for “significance” [Amrhein et al. 2019; Greenland 2017b, 2019].

His *Design of Experiments* [Fisher 1935] gives primacy to experimental action (design) over mathematics, as seen in sec. 2 of his introduction to the first edition, in which he states

"I have assumed, as the experimenter always does assume, that it is possible to draw valid inferences from the results of experimentation; **that it is possible to argue from consequences to causes**, from observations to hypotheses; as a statistician would say, from a sample to the population from which the sample was drawn, or, as a logician might put it, from the particular to the general."

His ensuing verbal descriptions were soon formalized by others into a clear potential-outcome model form, where *for each unit* explicit counterfactual (unobserved) treatment assignments lead to possibly distinct outcomes (e.g., see Welch [1937, p. 22-23]).

Nonetheless, the statistical theory that dominated subsequent advanced teaching and methods research became an extension of measure-theoretic probability, a development decried by those who followed Fisher in emphasizing the importance of context [Box 1990]. It is thus somewhat ironic that Fisher's downfall (as manifested in his defense of smoking against charges of carcinogenicity) was his inability to neutrally synthesize all available evidence sources, particularly in mishandling sources of information not derived from physical randomization. This failing can be viewed as one of being unable to form realistic models for confounding effects coupled with (or perhaps caused by) by personal wishes for vindication of his own smoking habit [Stolley 1991]. These sorts of "human factors" are themselves extraneous causes of what gets reported and publicized, and thus need to be accounted for in any realistic model for literature analysis [Greenland 2012b, 2017b].

Causal analysis vs. traditional statistical analysis

In applied statistics, assumptions are made to simplify modeling effort, which like everything else is resource constrained. For example, the standard modeling assumption "linear in the natural parameter" is rarely if ever deduced from anything; instead, standard statistical methods treat it as certainly true provided there is no evidence to contrary (even if there is little evidence to judge its accuracy or practical impact). This convention is based on the ease of use of such models, especially their transparency and computational stability relative to intrinsically nonlinear models, along with the idea that basic linear trend components are sometimes the only components that are needed or that can be stably estimated from available data.

A retreat from causal to convenience justification is only to be expected when applications involve complexities beyond complete formal (algorithmic) modeling capacities, as in biology, medicine, and social sciences. In such applications, all models are wrong at some practical level of analysis, and are often wrong in very consequential ways *even when they are useful for improving predictions of yet-unseen events such as treatment effects*. The classic epidemiologic example is malaria, a disease whose name means “bad air” in the parent Italian. Before modern times, social groups noted that malaria rates were higher near swamps and attributed that to toxic effects on the air from the swamps, as suggested by the foul smell associated with swamps. This wrong theory (causal-system model) of

$$\text{swamp} \rightarrow \text{toxic air} \rightarrow \text{malaria}$$

$$\text{housing location} \rightarrow \text{toxic air}$$

led to successful interventions such as draining swamps and building elevated houses, even though it missed the actual causal structure of

$$\text{swamp} \rightarrow \text{mosquito exposure} \rightarrow \text{malaria}$$

$$\text{housing location} \rightarrow \text{mosquito exposure}$$

which predicted the same intervention effects. To explain these successes of the wrong model, we may note that the swamp intervention tested only the swamp→malaria effect while the housing intervention tested only the housing→malaria effect. Both interventions left wide open the identity of the intermediates (and thus specifics of the mechanism for intervention), yet were taken to demonstrate the (in-fact untested) pathway of toxic air.

Such examples show that causal theories can include important mistakes even while successfully predicting intervention effects, and show why those theories should not be taken as true because of such successes (even in a world where causal laws are stable and thus inductive reasoning is justified). They instead need ongoing novel tests (not just “replication”) before basing actions on pathways that have not yet been tested by experiments. The enhanced risk of error for a mechanistic causal theory over a mere predictive/associative theory is not a disadvantage, however: it reflects the greater specificity, greater logical content, and hence greater testability of such theories, properties which are often promoted as hallmarks of good scientific theories (Popper 1968).

That such a theory can pass apparently strong experimental tests yet be erroneous in important ways (as in the malaria example) is one reason pragmatic analysts reject notions of

“experimental support” for scientific (real-world) causal theories. Other theories (including many never imagined) may pass the same experimental test, so at most we can only say an experiment supports the broad class of theories which predict results close to what was observed. Put another way: An intervention experiment provides evidence only on *classes* of mechanisms (those whose diagrams have directed paths from the observed intervention to the observed outcome), not specific mechanisms, and thus leaves open many details of intervention effects.

That caution applies even more strongly in passive observational (nonexperimental) studies, especially when their data are “analyzed” (summarized) by statistics based on randomization assumptions. In that case one can view a conventional interval estimate as a blur around the point estimate indicating irreducible uncertainty about the behavior of the data generator. But any inferential connection of these summaries to a targeted treatment effect should be mediated by explicit causal models; specifically, extraction of information about the target effect (e.g., in form of credible uncertainty intervals for the target) requires causal models for physical data generation that include nonrandom variation (bias) sources beyond the treatment [Greenland 1990, 2012a; Greenland et al. 1999; Maclure and Schneeweiss, 2001; Robins 2001; Hernán et al. 2004; Glymour and Greenland 2008]. It also requires recognition that effects cannot always be identified by observed associations, and that some effects cannot be statistically identified at all, even from randomized trials [Kaufman 2009; Robins and Richardson 2011].

Relating causality to traditional statistical philosophies and “objective” statistics

As has been long and widely emphasized in various terms (e.g., [Cox 1978; Box 1980, 1990; Rubin 1984; Good 1992; Barnard 1996; Chatfield 2002; Kelly and Glymour 2004; Greenland 2006, 2010b; Senn 2011; Gelman and Shalizi 2013], frequentism and Bayesianism are incomplete both as learning theories and as philosophies of statistics, in the pragmatic sense that each alone are insufficient for all sound applications. Notably, causal justifications are the foundation for classical frequentism, which demands that all model constraints be deduced from real mechanical constraints on the physical data-generating process. Nonetheless, it seems modeling analyses in health, medical, and social sciences rarely have such physical justification.

Beyond graphs, causality theory formalizes design information (such as randomization and matching) by the constraints that information places on the distributions of unobserved

variables (e.g., [Greenland 1990; Pearl 1995; Robins 2001; Hernán and Robins 2020]). Use of that information is especially important when the modeled data generator is not fully understood as a coherent whole – a problem long recognized and discussed at length in the literature on model uncertainty (e.g., [Leamer 1978; Box 1980]). The deficiency of strict coherent (operational subjective) Bayesianism is its assumption that all aspects of this uncertainty have been captured by the prior and likelihood, thus excluding the possibility of model misspecification [Leamer 1978; Box 1980; Senn 2011]. DeFinetti himself was aware of this limitation:

"...everything is based on distinctions which are themselves uncertain and vague, and which we conventionally translate into terms of certainty only because of the logical formulation...In the mathematical formulation of any problem it is necessary to base oneself on some appropriate idealizations and simplification. This is, however, a disadvantage; it is a distorting factor which one should always try to keep in check, and to approach circumspectly. It is unfortunate that the reverse often happens. One loses sight of the original nature of the problem, falls in love with the idealization, and then blames reality for not conforming to it." [DeFinetti 1975, p. 279]⁷

By asking for physically causal justifications of the data distributions employed in statistical analyses (whether those analyses are labeled frequentist or Bayesian), we may minimize the excessive certainty imposed by simply assuming a probability model and proceeding as if that idealization were a known fact.

DeFinetti was of course writing in support of a contentious, purely subjective view of probability, and the utility of the entire “subjective”/“objective” distinction in statistics has been questioned [Gelman and Hennig 2017]. Nonetheless, many statisticians assign primacy to “objective” model components (those derivable from observed mechanisms, such as random-number generators). What supports a claim that a variable is “completely random” (fully randomized) in an objective frequency sense? Modern causality theory can identify this randomness with the assumption that the variable is exogenous or instrumental, in that it has no causes (parents) in the system under study [Pearl 2009]. Again, in “objective” theory this sharp,

⁷I am indebted to Stephen Senn for reminding me of this and other remarkable passages in DeFinetti.

strong assumption is *deduced* from the physical data-generating mechanism, not from observed frequencies or other purely associational information.

Consider "fair" coin tossing, in which the influence of the person tossing (who might be a magician) is blocked by having them throw the coin against a wall and then step back before the bounce and landing, thus blocking skilled tossing and other trickery as influences of the outcome. Then, even under classical deterministic mechanics, the functional complexity of the relation of the outcome to the initial toss is transcomputable or chaotic. This type of complexity forces our predictions to rely on distributions that arise as attractors of statistical behavior (e.g., laws of large numbers, central-limit effects), instead of deterministic mathematical models. In doing so we assume a certain causal stability across trials whose consequences are summarized in our models. Such a stability assumption needs justification based on direct observation (the physical mechanism is unchanging) and thus is objective; without that, causal stability is an underived (and usually implicit) assumption and thus is not objective in this sense.

In this way, the traditional "objective"/"subjective" distinction in statistical methods resides within causality theory, not in the "frequentist" vs "Bayesian" distinction (which are both vague labels for highly heterogeneous collections of statistical tools and philosophies, as Good [1971] explained for Bayes). The core idea behind "objective" statistics is that one demands that each distribution used in the statistical processing of the data be derivable from a verifiable physical (causal) mechanism. That demand can be made regardless of whether that processing is labeled "frequentist", "Bayesian", "likelihoodist" or something else, a view which does not exclude Bayesian methods, but does reject mere expressions of opinions as priors for those methods [von Mises 1981].

Discussion

Judea Pearl has been a celebrated promoter of causal models over pure probability, especially for encoding the background (contextual) information in a problem [Pearl 1995, 2001, 2009]. At times however he has referred to causality as "extra-statistical," a label which ignores the realities that any applied statistician must face in practice. Those realities make causality integral to statistics; yet, by calling causality "extra-statistical" we absolve those bearing the professional label "statistician" of any responsibility to understand let alone teach causality theory. Fortunately, many younger statisticians have a keen interest in causal models as tools to

create better statistical science. To encourage this trend we should include causal models from the start of statistical training as an integral component of study design and data analysis – in addition to complementary presentation of frequentist and Bayesian ideas.

As a less-often stated yet even more fundamental need, basic statistics should begin with the elements of deductive logic. When I was teaching statistical foundations and principles, most students I encountered (including statistics majors) had neither studied nor fully understood basic logical principles, and thus were prone to naïve fallacies in verbal arguments. Thus the topic sequence in my class covered logic as a foundation for causal thinking, followed by causality theory as a foundation for probability and association explanation. This material was contrasted to their previous instruction, which typically involved rote application of mysterious descriptions and formulas for statistical comparisons and regressions. Students were always delighted to at last see applied statistics as the coordinated merging of the three essentials of logic, causation, and probability to provide a transparent foundation for sound study design, analysis, and interpretation.

Admittedly, traditionally trained statisticians may be too firmly wedded to probabilistic foundations to ever concede this causal primacy, and some radical subjective Bayesians reject causality altogether (e.g., Lad [2006]). Nonetheless, probabilists curious about the causal approach may more easily conceive the unification of causality and probability within information theory, which can serve as an overarching framework for statistical modeling and inference (I have found that the information framework even helps students correctly understand P-values [Greenland 2019; Rafi and Greenland 2020]). Causal diagrams then provide an intuitive representation of information flows as time-sequential functional relations across event sequences.

Conclusion

Statistical science (as opposed to mathematical statistics) involves far more than data – it requires realistic *causal* models for the generation of that data and the deduction of their empirical consequences. Evaluating the realism of those models in turn requires immersion in the subject matter (context) under study. Decisions further require explication of the various pathways by which those decisions would cause gains (benefits) and losses (costs). Bringing

these causal elements to the foreground is essential for sound teaching and applications of statistics.

Appendix. A counting measure for the logical content of a finite exchangeability assumption

For any formal deductive system and assumption set A in the system, define A as logically minimal if it satisfies the joint deductive independence condition: For any pair (B,C) of distinct nonempty subsets of A , C cannot be deduced from B . We may then define the logical-content measure $v(G)$ of an arbitrary assumption set G in the system as the largest cardinality $|A|$ among minimal subsets A of G ; $v(G)$ may be infinite if G is infinite.

Now consider the common statistical assumption that the observations Y_1, \dots, Y_N are independent identically distributed conditional on any model m in a set M . Then, given a prior distribution on M , the Y_1, \dots, Y_N are unconditionally exchangeable; that is, every one of the $N!$ permutations of indices in the joint distribution leaves that distribution unchanged. Exchangeability is logically equivalent to $N!-1$ nonvacuous assumptions, one for each non-null permutation; denoting the set of these assumptions by G , with no further constraint we have $v(G) = |G| = N!-1$. By imposing further assumptions on the joint distribution we may reduce $v(G)$ considerably. Nonetheless, even with the extreme simplification of multivariate normality we get $v(G)$ of order N^2 (since exchangeability requires homogeneous variances and homogeneous covariances), and thus still entails far more assumptions than there are observations N .

References

- Amrhein, V., Trafimow, D., Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, 73 supplement 1, 262-270, open access at www.tandfonline.com/doi/pdf/10.1080/00031305.2018.1543137
- Barnard, G. A. (1996). Fragments of a statistical autobiography. *Student* 1, 257–68.
- Box, G. E. P. (1980). Sampling and Bayes inference in scientific modeling and robustness (with discussion). *Journal of the Royal Statistical Society, Ser. A*, 143, 383-430.
- Box, G. E. P. (1990). Comment. *Statistical Science*, 5, 448–449.
- Chatfield, C. (2002). Confessions of a pragmatic statistician. *The Statistician*, 51(1), 1-20.

- Copas, J. B. (1973). Randomization models for matched and unmatched 2×2 tables. *Biometrika*, 60, 467-476.
- Cox, D.R. (1978). Foundations of statistical inference: The case for eclecticism. *Australasian Journal of Statistics*, 20(1), 43-59.
- DeFinetti, B. (1975). *Theory of Probability* Vol. 2. New York, Wiley.
- Fisher, R.A. (1935). *The Design of Experiments*. Edinburgh, Oliver & Boyd.
- Gelman, A., Shalizi, C. (2013) Philosophy and the practice of Bayesian statistics (with discussion). *British Journal of Mathematical and Statistical Psychology*, 66, 8–80
- Gelman, A., Hennig, C. (2017). Beyond subjective and objective in statistics (with discussion). *Journal of the Royal Statistical Society, Series A*, 180(4), 967-1033.
<https://doi.org/10.1111/rssa.12276>
- Glymour, M.M., Greenland, S. (2008). Causal diagrams. Ch. 12 in Rothman K.J. Greenland, S., Lash, T. (eds). *Modern Epidemiology*, 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008, pp. 32-50.
- Good, I.J. (1971). 46,656 varieties of Bayesians (letter). *The American Statistician*, 25, 62-63. Reprinted as Ch. 3 in I.J. Good, ed. (1983). *Good Thinking*. Minneapolis, University of Minnesota Press, 20-21.
- Good, I.J. (1987). Hierarchical Bayesian and empirical Bayesian methods (letter). *The American Statistician*, 41, 92.
- Good, I.J. (1992). The Bayes/non-Bayes compromise: a brief review. *Journal of the American Statistical Association*, 87, 597-606
- Greenland, S. (1990). Randomization, statistics, and causal inference. *Epidemiology*, 1, 421-429.
- Greenland, S. (1991). On the logical justification of conditional tests for two-by-two-contingency tables. *The American Statistician*, 45, 248-251.
- Greenland, S. (2006). Bayesian perspectives for epidemiologic research. I. Foundations and basic methods. *International Journal of Epidemiology*, 35, 765-778. Reprinted with edits as Bayesian Analysis, Ch. 18 in Rothman K.J. Greenland, S., Lash, T. (eds). *Modern Epidemiology*, 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008, pp. 32-50.
- Greenland, S. (2010a). Overthrowing the tyranny of null hypotheses hidden in causal diagrams. Ch. 22 in: Dechter, R., Geffner, H., and Halpern, J.Y. (eds.). *Heuristics, Probabilities, and Causality: A Tribute to Judea Pearl*. London: College Publications, 365-382.

- Greenland, S. (2010b). Comment: The need for syncretism in applied statistics. *Statistical Science*, 25(2), 158 – 161.
- Greenland, S. (2012a). Causal inference as a prediction problem: Assumptions, identification, and evidence synthesis. Ch. 5 in: Berzuini, C., Dawid, A.P., and Bernardinelli, L. (eds.). *Causal Inference: Statistical Perspectives and Applications*. John Wiley and Sons, Chichester, UK, 43-58.
- Greenland, S. (2012b). Transparency and disclosure, neutrality and balance: shared values or just shared words? *Journal of Epidemiology and Community Health*, 66, 967–970.
- Greenland, S. (2017a). For and against methodology: Some perspectives on recent causal and statistical inference debates. *European Journal of Epidemiology*, 32, 3-20, <https://doi.org/10.1007/s10654-017-0230-6>
- Greenland, S. (2017b). The need for cognitive science in methodology. *American Journal of Epidemiology*, 186, 639-645, <https://doi.org/10.1093/aje/kwx259>.
- Greenland, S. (2019). Some misleading criticisms of P-values and their resolution with S-values. *The American Statistician*, 73, supplement 1, 106-114, open access at www.tandfonline.com/doi/pdf/10.1080/00031305.2018.1529625
- Greenland, S., Rafi, Z. (2020). To aid statistical inference, emphasize unconditional descriptions of statistics. <http://arxiv.org/abs/1909.08583>
- Greenland, S., Fay, M.P., Brittain, E.H., Shih, J.H., Follmann, D.A., Gabriel, E.E., Robins, J.M. (2019). On causal inferences for personalized medicine: how hidden causal assumptions led to erroneous causal claims about the D-value. *The American Statistician*, 73, in press, doi: 10.1080/00031305.2018.1502684
- Greenland, S., Maclure, M., Schlesselman, J.J., Poole, C., Morgenstern, H. (1991). Standardized regression coefficients: A further critique and a review of alternatives. *Epidemiology*, 2, 387-392.
- Greenland, S., Pearl, J., Robins, J.M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10, 37–48.
- Greenland, S., Robins, J.M. (1988). Conceptual problems in the definition and interpretation of attributable fractions. *American Journal of Epidemiology*, 128, 1185-1197.

- Greenland, S., Schlesselman, J.J., Criqui, M.H. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology*, 123, 203-208.
- Hernán, M.A. (2017). Selection bias without colliders. *American Journal of Epidemiology*, 185(11), 1048-1050.
- Hernán, M.A., Hernández-Díaz, S., Robins, J.M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615 – 625.
- Hernán, M.A., Robins, J.M. (2020). *Causal Inference: What If*. New York, Chapman & Hall, to appear.
- Hume, D. (1748). *An Enquiry Concerning Human Understanding*. Oxford: Oxford Univ. Press 2007 printing, p. 56.
- Kaufman, J. (2009). Gilding the black box. *International Journal of Epidemiology*, 38, 845-847.
- Kelly, K.T., Glymour, C. (2004). Why probability does not capture the logic of scientific justification. In C. Hitchcock (Ed.), *Contemporary debates in philosophy of science*, 94–114. Malden, Mass: Blackwell.
- Lad, F. (2006). Objective Bayesian Statistics: Do you buy it? Should we sell it? *Bayesian Analysis*, 1(3), 441-444.
- Lamb, E. (2012). 5 sigma – what’s that? *Scientific American Observations*, posted July 17, 2012 at <https://blogs.scientificamerican.com/observations/five-sigmawhats-that/>, viewed June 2, 2019.
- Leamer, E.E. (1978). *Specification Searches*. New York, Wiley.
- Maclure, M.M., Schneeweiss, S. Causation of Bias: The Episcopes. *Epidemiology*, 12, 114-122.
- McShane, B.B., Gal, D., Gelman, A., Robert, C., Tackett, J.L. (2019). Abandon statistical significance. *The American Statistician*, 73, 235–45. doi:10.1080/00031305.2018.1527253.
- Mermin, N.D. (2016). Why QBism is not the Copenhagen interpretation and what John Bell might have thought of it. In: Bertlmann, R., Zeilinger, A. (eds) *Quantum [Un]Speakables II*. Springer, Cham. First Online 16 November 2016, https://doi.org/10.1007/978-3-319-38987-5_4
- Merriam-Webster Dictionary (2019), “Statistics,” <https://www.merriam-webster.com/dictionary/statistics>, accessed May 16, 2019.

- Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences Agaricales: Essay des principe. [English translation of excerpts (1990) by D. Dabrowska and T. Speed, *Statistical Science*, 5, 463-472.]
- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82, 669-710.
- Pearl, J. (2001). Bayesianism and causality, or, why I am only a half-Bayesian. In D. Corfield and J. Williamson (Eds.) *Foundations of Bayesianism*. Kluwer Applied Logic Series, 24. Kluwer Academic Publishers, pp.19-36.
- Pearl, J. (2009). *Causality: Models, Reasoning and Inference*, 2nd ed. New York, Cambridge University Press.
- Popper, K.R. (1962). *Conjectures and Refutations*. New York, Basic Books.
- Rafi, Z., Greenland, S. (2020). Semantic and cognitive tools to aid statistical inference: Replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology*, 20, 244. doi: 10.1186/s12874-020-01105-9, <https://bmcmmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-01105-9>, updates at <http://arxiv.org/abs/1909.08579>
- Robins, J.M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology* 12, 313-320.
- Robins, J.M., Greenland, S. (1989). The probability of causation under a stochastic model for individual risks. *Biometrics*, 46, 1125-1138. (Erratum: 1991, 48, 824)
- Robins, J.M., Richardson, T.S. (2011). Alternative graphical causal models and the identification of direct effects. Ch. 6 in P. Shrout, K. Keyes, K. Ornstein, eds., *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*. Oxford: Oxford University Press, 1-52.
- Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, 6, 34-58.
- Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, 1151-1172.
- Senn, S. (2011). You may believe you are a Bayesian but you are probably wrong. *RMM* Vol. 2, 2011, 48–66.
- Stolley, P. (1991). When genius errs: R. A. Fisher and the lung cancer controversy. *American Journal of Epidemiology*, 133, 416-425.

- von Mises, R. (1981). *Probability, Statistics and Truth*, 2nd rev. English ed., New York, Dover, 1981.
- Wasserstein, R.L. (2018). Turing Award winner, longtime ASA member publishes *The Book of Why* (interview with Judea Pearl). *Amstat News* Aug. 2018, 12-14.
- Wasserstein, R.L., Schirm, A.L., Lazar, N.A. (2019). Moving to a world beyond “ $p < 0.05$.” *The American Statistician*, 73, 1-19. doi:10.1080/00031305.2019.1583913.
- Welch, B.L. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika*, 29, 21-52.
- Wilk, M.B. (1955). The randomization analysis of a generalized randomized block design. *Biometrika*, 42, 70-79.