Causal questions are questions that are answered by a function

David Johnston

September 8, 2021

Researchers in the field of causal inference will often choose a causal framework as one of the first steps of their investigations, or in some cases, one of the first steps of their careers. One could postulate that "causal inference" is what one does when one does work using a causal modelling framework. We argue that "causal inference" is better understood by the kind of questions people ask on rather than the kind of framework people use to answer them. Pearl and Mackenzie (2018) has proposed a three-level hierarchy for classifying causal questions: at the bottom are "seeing" questions followed by "doing" questions with "imagining" questions at the top. We propose an alternative characterisation: "ordinary statistical questions" are questions involving data that are answered by distributions on a given set while "causal statistical questions" are questions involving data that are answered by stochastic functions with given domain and codomain. Potential outcomes and graphical models are features of modelling frameworks, while interventions and counterfactuals are features of causal problems. We show how both potential outcomes and causal graphical models arise in *see-do models*, a generic modelling framework we introduce that addresses causal questions in general, as we define them. We hypothesise that some confusion about interventions and counterfactuals arises from assuming they are given by the modelling framework rather than by the problem under investigation.

or something like that; it could also be functions of a distribution like a maximum likelihood estimate or a p-value

Contents

0.1 Technical prerequesites

Many people are familiar with probability theory, but some may be less familiar with *Markov kernels*, which play a central role in the work developed in this paper. Markov kernels are measurable functions that map to probability distributions on some measurable set. Expressions like $\mathbb{P}(\mathsf{Y}|\mathsf{X})$ and $x \mapsto \mathbb{P}(\mathsf{Y}|do(\mathsf{X}=x))$ represent Markov kernels that map from the range of the random variable X to probability distributions on the range of the random variable Y. Conditional probabilities like $\mathbb{P}(\mathsf{Y}|\mathsf{X})$ are typically obtained by disintegrating a joint probability $\mathbb{P}(\mathsf{Y},\mathsf{X})$, but Markov kernels can also be things other than conditional probabilities, like "interventional maps" $x \mapsto \mathbb{P}(\mathsf{Y}|do(\mathsf{X}=x))$.

We will consider only discrete sets in this paper, as uncountable sets raise a number of difficulties we prefer to avoid in this paper. A discrete set is a set X which is at most countably infinite, equipped with the

Footnote?: These difficulties may be a general phenomenon - for example, letting $X := \mathrm{Range}(\mathsf{X})$, if X is real-valued, then for almost every $x \in X$ there are many choices for $\mathbb{P}(\mathsf{Y}|do(\mathsf{X}=x))$ that all satisfy the definition given by Pearl (2009) because $\mathbb{P}(\mathsf{X}=x)$ can be positive for an at most countable subset of X. Also, there are examples of theorems that hold for discrete sets only Heymann et al. (2021)

We will take advantage of the fact that we are working with discrete sets and define probability measures as vectors, measurable functions as covectors and Markov kernels as matrices.

Given a set X, a probability measure $\mathbb{P}$ on X is a vector in $\mathbb{R}^{|X|}$; $\mathbb{P} := (P_i)_{i\in X}$. We require that

$$0 \leq P_i \leq 1 \qquad \forall i \in X \tag{1}$$

$$\sum_i P_i = 1 \tag{2}$$

An *event* A is a subset of X, and we define $\mathbb{P}(A) := \sum_{i\in A} P_i$.

A measurable function $f : X \to Y$ is a covector in $Y^{|X|}$; $f := (f_i)_{i\in X}$ where Y is a vector space (i.e. it has addition and scalar multiplication). Defining a measurable function by a mapping $f := x \mapsto f(x)$ is equivalent to defining $f := (f(x))_{x\in X}$.

Given discrete sets X and Y, a Markov kernel $\mathbb{K} : X \to \Delta(Y)$ is a matrix in $\mathbb{R}^{|X|\times|Y|}$; $\mathbb{K} = (K_{ij})_{i\in X, j\in Y}$ where

$$0 \leq K_{ij} \leq 1 \qquad \forall i, j \tag{3}$$

$$\sum_{i\in X} K_{ij} = 1 \qquad \forall j \tag{4}$$

We use subscripts to refer to rows of a Markov kernel $\mathbb{K}_x := (K_{xj})_{j\in Y}$; these are all probability measures.

We use the usual associative notion of vector and matrix products for expressions like $\mathbb{P}\mathbb{K}$, $\mathbb{P}f$, $\mathbb{K}f$ and so on.

A probability measure is a Markov kernel where the domain set X has a single element.

It can be verified that $\mathbb{P}\mathbb{K}$ is a probability measure, $\mathbb{K}f$ is a measurable function and $\mathbb{P}f := \mathbb{E}_{\mathbb{P}}[f]$ is a scalar which we define as the expectation of f under $\mathbb{P}$. Given another Markov kernel $\mathbb{L} : Y \to \Delta(Z)$, the matrix product $\mathbb{K}\mathbb{Y}$ is also a Markov kernel.

0.1.1 Cartesian and tensor products

The cartesian product $X \times Y := \{(x, y)|x \in X, y \in Y\}$.

Given kernels $\mathbb{K} : W \to Y$ and $\mathbb{L} : X \to Z$, the tensor product $\mathbb{K} \otimes \mathbb{L} : W \times X \to \Delta(Y \times Z)$ is given by $\mathbb{K} \otimes \mathbb{L} := (K_{wy}L_{xz})_{(w,x)\in X\times W,(y,z)\in Y\times Z)}$. Equivalently, it is given by the mapping $\mathbb{K} \otimes \mathbb{L} : (w, x, A \times B) \mapsto \mathbb{K}_w(A)\mathbb{L}_x(B)$ for $w \in W$, $x \in X$, $A \subset Y$ and $B \subset Z$.

Given functions $f : W \to Y$ and $g : X \to Z$, the tensor product $f \otimes g : W \times X \to yz|y \in Y, z \in Z$ is the covector $((f_w, g_x))_{w\in W,x\in X}$.

0.1.2 Indicator functions, delta measures and function-associated Markov kernels

The iverson bracket $[\![\cdot]\!]$ evaluates to 1 if $\cdot$ is true and 0 otherwise.

For any X and any $A \subset X$, $\mathbb{1}_A$ is the function $x \mapsto [\![x \in A]\!]$. According to the definition above $\mathbb{P}(A) = \mathbb{P}\mathbb{1}_A$.

For any X and any $x \in X$, δ_x is the probability measure $([\![x = i]\!])_{i \in X}$.

We can define the Markov kernel $\mathbb{F}_f : X \to \Delta(\mathcal{Y})$ associated with the function $f : X \to Y$ with the matrix that sends $x \mapsto \delta_{f_x}$. Alternatively, we can define it by its rows: $(\delta_{f_x})_{x \in X}$.

0.1.3 Copy maps and sequences

The copy map $\curlyvee : X \to \Delta(X \times X)$ is the Markov kernel mapping $x \mapsto \delta_x \otimes \delta_x$. Alternatively, its rows are given by $([\![x = i \;\;\&\;\; x = i']\!])_{x,i,i' \in X}$.

Given $X : E \to X$ and $Y : E \to Y$, the *sequence* random variable $(\mathsf{X}, \mathsf{Y}) : E \to X \times Y$ is defined as $\curlyvee(\mathsf{X} \otimes \mathsf{Y})$. This is the function given by $\omega \mapsto (\mathsf{X}(\omega), \mathsf{Y}(\omega))$ or equivalently $((\mathsf{X}(\omega), \mathsf{Y}(\omega))_{\omega \in E}$.

0.1.4 Generalised random variables

It is typical to define a probability space as a probability measure along with its underlying set and its σ-algebra: $(\mathbb{P}, (E, \mathcal{E}))$. Here where E is sometimes called the sample space and $\mathcal{E}$ is sometimes called the set of events; as we are considering discrete sets, in this paper we always have $\mathcal{E} := \mathcal{P}(E)$ and we will typically only talk about the set E.

Given a probability space $(\mathbb{P}, E)$, we can define *random variables* as measurable functions $\mathsf{X} : E \to X$. The *marginal distribution* of X is given by $\mathbb{P}^{\mathsf{X}} := \mathbb{P}\mathbb{F}_{\mathsf{X}}$.

Here we want to consider "Markov kernel spaces", which is a Markov kernel along with its domain and underlying set of its codomain: $(\mathbb{K}, D, F)$. Given such a triple, a *generalised random variable* is a function from $D \times F$ to $\mathbb{R}$. Unlike random variables, generalised random variables don't necessarily have a unique marginal distribution. For brevity, we will simply call them *variables* henceforth.

Instead, we define the *domain variable* $\mathsf{D} : D \times F \to D$ by the projection $(d, f) \mapsto d$ for all $d \in D$, $f \in F$. For each $d \in D$, any variable $\mathsf{X} : D \times F \to X$ has a unique marginal distribution $\mathbb{K}_d^{\mathsf{X}|\mathsf{D}} := \mathbb{K}_d \mathbb{F}_{\mathsf{X}}$.

To save space, we say that the marginal distribution of (X, Y) is $\mathbb{K}_d^{\mathsf{XY}|\mathsf{D}}$, and use a similar shorthand for sequences henceforth.

0.1.5 Disintegration

Conditional probabilities are *disintegrations* of probability measures. Given a probability space $(\mathbb{P}, E)$ and random variables $\mathsf{X} : E \to X$ and $\mathsf{Y} : E \to Y$, the probability of X given Y is any Markov kernel $\mathbb{P}^{\mathsf{Y}|\mathsf{X}}$ such that $\mathbb{P}^{\mathsf{XY}} = (P_i^{\mathsf{X}} P_{ij}^{\mathsf{Y}|\mathsf{X}})_{i \in X, j \in Y}$. Note that this is generally non-unique. However, wherever $P_i^{\mathsf{X}} > 0$, $P_{ij}^{\mathsf{Y}|\mathsf{X}}$ must be equal to $\frac{P_{ij}^{\mathsf{XY}}}{P^{\mathsf{X}_i}}$.

We define disintegrations of kernels analogously. Given a Markov kernel space $(\mathbb{K}, D, F)$, domain variable D and variables X, Y, $\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}$ is any Markov kernel such that $\mathbb{K}^{\mathsf{XY}|\mathsf{D}} = (\mathbb{K}^{\mathsf{X}|\mathsf{D}}_{ij}\mathbb{K}^{\mathsf{Y}|\mathsf{XD}}_{ijk})_{i\in D, j\in X, k\in Y}$.

Note that we have at this point offered no definition of expressions like $\mathbb{K}^{\mathsf{X}|\mathsf{Y}}$ which lack dependence on the domain variable D. As mentioned, there is in general no unique marginal distribution of (X, Y). However, such a distribution might exist if X and Y are independent of

0.1.6 Conditional independence

Given a Markov kernel space $(\mathbb{K}, D, F)$, and random variables $\mathsf{X}, \mathsf{Y}, \mathsf{Z}$ we say X is independent of Y given Z, notated $\mathsf{X} \perp\!\!\!\perp_{\mathbb{K}} \mathsf{Y}|\mathsf{Z}$ iff a version of $\mathbb{K}^{\mathsf{X}|\mathsf{YZ}}$ exists such that $\mathbb{K}^{\mathsf{X}|\mathsf{YZ}}_{(y,z)} = \mathbb{K}^{\mathsf{X}|\mathsf{YZ}}_{(y',z)}$ for all $y, y' \in Y$.

A version of $\mathbb{K}^{\mathsf{X}|\mathsf{Z}}$ exists iff $\mathsf{X} \perp\!\!\!\perp_{\mathbb{K}} \mathsf{D}|\mathsf{Z}$ and in this case is given by $z \mapsto \mathbb{K}^{\mathsf{X}|\mathsf{DZ}}_{(d,z)}$ for any version of $\mathbb{K}^{\mathsf{X}|\mathsf{DZ}}$ and any $d \in D$.

0.2 See-do models

We will first introduce *see-do models* as a type of model that functions as the basic kind of thing which we will use to examine questions in the decision theoretic, potential outcomes and graphical models appraoch.

See-do models can be understood as generalisations of statistical models. Statistical models are a ubiquitous type of model in statistics and machine learning that consist of a set of *states* S, and for each state the model prescribes a single probability distribution on a given set of *outcomes* O.

Definition 0.2.1 (Statistical model)**.** A statistical model is a set of states S, a set of outcomes O and a Markov kernel $\mathbb{T} : S \to \Delta(O)$.

For example, a potentially biased coin can be modelled with a statistical model. Suppose the coin has some rate of heads $\theta \in [0, 1]$, and we furthermore suppose that for each θ the result of flipping the coin can be modeled (in some sense) by the probability distribution Bernoulli(θ). The statistical model here is the set of states $S = [0, 1]$ (corresponding to *rates of heads*), the observation space $O = \{0, 1\}^n$ with the discrete sigma-algebra (where n is the number of flips observed) and the stochastic map $\mathbb{B} : [0, 1] \to \Delta(\mathcal{P}(0, 1))$ which is given by $\mathbb{B} : \theta \to$ Bernoulli(θ).

This example actually goes beyond our formal definitions here in that θ is real-valued between 0 and 1. Extending probability theory to real-valued spaces is well understood, see for example Çinlar (2011), but in that setting the existence of disintegrations on kernel spaces (section 0.1.5) is a problem to which we presently only have a partial solution. Discrete sets allow us to discuss see-do models without going into this difficulty. The price we pay is that to properly model the above problem we require θ to take on discrete values, for example restricting it to the rationals.

A see-do model adds the following structure to a statistical model:

- The state is a pair consisting of a *hypothesis* $h \in H$ and a *decision* $d \in D$; $S = H \times D$
- The outcome is a pair consisting of an *observation* $x \in X$ and a consequence $y \in Y$
- The observation is conditionally independent of the decision given the hypothesis

We can use see-do models to model situations where we have some hypotheses and we have the opportunity to make an observation that takes values in X. Depending on what we see, we can select a decision from a set of possibilities D, and the ultimate consequence depends probabilistically on the decision we selected as well as whichever hypothesis turns out to best describe the world.

Definition 0.2.2. A *see-do model* $(\mathbb{T}, \mathsf{H}, \mathsf{D}, \mathsf{X}, \mathsf{Y})$ is a Markov kernel space $(\mathbb{T}, H \times D, O)$ along with four variables: the *hypothesis* $\mathsf{H} : H \times D \times O \to H$, the *decision* $\mathsf{D} : H \times D \times O \to D$, the *observation* $\mathsf{X} : H \times D \times O \to X$ and the *consequence* $\mathsf{Y} : H \times D \times O \to Y$, all given by the projections onto the respective spaces. In addition, a see-do model must observe the conditional independence:

$$\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D} | \mathsf{H} \tag{5}$$

See-do models feature variables D and H that act like Dawid's "non-stochastic regime indicators" described by Dawid (2002, 2012, 2020). In particular, see-do models induce a collection of probability measures indexed by the elements of $H \times D$, just as regime indicators induce collections of indexed probability measures. Dawid's regime indicators seem to typically do a similar job to a decision variable D rather than a decision-hypothesis pair.

The hypothesis set is similar to the parameter set described by Lattimore and Rohde (2019) that relates pre- and post-interventional distributions. Lattimore and Rohde consider models with a prior distribution over this parameter set. A similar type of model can be created by taking the prodcut of a prior over the hypothesis set and a see-do model.

Finally, see-do models are somewhat similar to the models proposed by Savage (1954) for decision problems if we identify *states* with *hypotheses* and *acts* with *decisions*. Savage's models consider deterministic rather than stochastic functions from acts to outcomes, and did not explicitly distinguish observations from consequences.

0.2.1 See-do models are motivated by data-driven decision problems

- You get an observation x and may return some mixtures of decisions in $\Delta(\mathcal{D})$; that is, you can choose some function $x \to \Delta(\mathcal{D})$

- For any function $Q : x \to \Delta(\mathcal{D})$ we have a forecast of the observations, decisions and consequences, and it is appropriate to model the forecast with some probability over observations decisions and consequences $\mathbb{P}_Q \in \Delta(\mathcal{X} \otimes \mathcal{D} \otimes \mathcal{Y})$
  - You can probably make this a collection of probabilities/Markov kernel, it's just simpler to consider one probability for this outline
- For all $Q, R : x \to \Delta(\mathcal{D})$, $\mathbb{P}_Q^{\mathsf{X}} = \mathbb{P}_R^{\mathsf{X}}$, $\mathbb{P}_Q^{\mathsf{Y}|\mathsf{XD}} = \mathbb{P}_R^{\mathsf{Y}|\mathsf{XD}}$ and $\mathbb{P}_Q^{\mathsf{D}|\mathsf{X}} = Q$
  - That is: I should expect the same observations whatever function I end up choosing and I expect the same consequences holding the observations and decision fixed (even more informally: it doesn't matter how I choose decisions provided I end up choosing the same one)
  - Finally, the decision function we choose is the relation between X and D
- There exists some Q such that $\mathbb{P}_Q$ has full support

References

Erhan Çinlar. *Probability and Stochastics*. Springer, 2011.

A. P. Dawid. Influence Diagrams for Causal Modelling and Inference. *International Statistical Review*, 70(2):161–189, 2002. ISSN 1751-5823. doi: 10.1111/j.1751-5823.2002.tb00354.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2002.tb00354.x`.

A. Philip Dawid. Decision-theoretic foundations for statistical causality. *arXiv:2004.12493 [math, stat]*, April 2020. URL `http://arxiv.org/abs/2004.12493`. arXiv: 2004.12493.

Philip Dawid. The Decision-Theoretic Approach to Causal Inference. In *Causality*, pages 25–42. John Wiley & Sons, Ltd, 2012. ISBN 978-1-119-94571-0. doi: 10.1002/9781119945710.ch4. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119945710.ch4`.

Benjamin Heymann, Michel de Lara, and Jean-Philippe Chancelier. Causal Inference Theory with Information Dependency Models. August 2021. URL `https://arxiv.org/abs/2108.03099v2`.

Finnian Lattimore and David Rohde. Replacing the do-calculus with Bayes rule. *arXiv:1906.07125 [cs, stat]*, December 2019. URL `http://arxiv.org/abs/1906.07125`. arXiv: 1906.07125.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect.* Basic Books, New York, 1 edition edition, May 2018. ISBN 978-0-465-09760-9.

Leonard J. Savage. *The Foundations of Statistics.* Wiley Publications in Statistics, 1954.

Appendix: