Causal questions are questions that are answered by a function

David Johnston

September 15, 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

1 Technical prerequesites **2**

1.1 Markov kernels 2

1.2 Cartesian and tensor products 3

1.3 Delta measures, erase maps, copy maps 3

1.4 Products 4

1.5 Labeled Markov kernels, conditional probabilities 4

1.6 Modelling context, extension 4

1.7 Sequences 5

1.8 Disintegration 6

1.9 Conditional independence 6

2 See-do models **6**
2.1 See-do models for data-driven decision problems 8

3 Potential outcomes and counterfactuals **10**
3.1 Potential outcomes in see-do models 11
3.2 Parallel potential outcomes representation theorem 12

1 Technical prerequesites

Our theory makes heavy use of *Markov kernels* or *stochastic functions*, which are taken from probability theory. However, the manner in which we use them is non-standard. The usual way to apply probability theory to model building is to assume we have a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ with random variables defined as functions with domain Ω, and all aspects of the model of interest are supposed to be captured by this. Under our approach, we instead consider components, represented by Markov kernels $\mathbb{K} : E \to \Delta(F)$ along with labeled inputs and outputs. The labels do the same job that random variables do in the usual formulation. These components can be composed or broken apart, but we do not assume that there is an overarching probability space from which all components can be derived.

In addition, we introduce a graphical notation for Markov kernels that is the subject of a coherence theorem: two Markov kernels represented by pictures that differ only by planar deformations are identical (Selinger, 2010).

1.1 Markov kernels

Markov kernels can be thought of as measurable functions that map to probability distributions. A conditional probability $\mathbb{P}(\mathsf{Y}|\mathsf{X})$, which maps from values of X to probability distributions over Y, and an interventional map $x \mapsto \mathbb{P}(\mathsf{Y}|do(\mathsf{X} = x))$ that likewise maps values of X to probability distributions on Y, are both Markov kernels.

Our theory is susbtantially simplified by restricting our attention to discrete sets – that is, sets X with at most a countable number of elements endowed with the σ-algebra made up of every subset of X, also called the discrete σ-algebra.

In the discrete setting, we can represent probability distributions as covectors, Markov kernels as matrices and measurable functions as vectors.

Given a set X, a probability distribution $\mathbb{P}$ on X is a covector in $\mathbb{R}^{|X|}$, which we will write $\mathbb{P} := (\mathbb{P}^i)_{i\in X}$. To be a probability distribution we require

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

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

Given discrete sets X and Y, a Markov kernel $\mathbb{K} : X \to \Delta(Y)$ is a matrix in

$\mathbb{R}^{|X|\times|Y|}$; $\mathbb{K} = (K_i^j)_{i\in X, j\in Y}$ where

$$0 \leq K_i^j \leq 1 \qquad \forall i,j \tag{3}$$

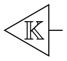
$$\sum_{i\in X} K_i^j = 1 \qquad \forall j \tag{4}$$

Rows of Markov kernel are probability distributions: $\mathbb{K}_x := (K_x^j)_{j\in Y}$. Alternatively, we can consider probability distributions to be Markov kernels with one row.

Graphically, we represent a Markov kernel as a box and a probability distribution as a triangle:

$$\mathbb{K} := \;-\boxed{\mathbb{K}}- \tag{5}$$

$$\mathbb{P} := \tag{6}$$



1.2 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 defined by $(\mathbb{K}\otimes\mathbb{L})_{(w,x)}^{(y,z)} := K_w^y L_x^z$.

Graphically, the tensor product is represeted by parallel juxtaposition:

$$\mathbb{K}\otimes\mathbb{L} := \begin{array}{c} -\boxed{\mathbb{K}}- \\ -\boxed{\mathbb{L}}- \end{array} \tag{7}$$

1.3 Delta measures, erase maps, copy maps

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

For any X and any $x \in X$, $\delta[x]$ is the probability measure defined by $\delta[x]^i = [\![x = i]\!]$. The identity map $\mathrm{Id}[X] : X \to \Delta(X)$ is given by $x \mapsto \delta[x]$.

Graphically, the identity map is a bare line:

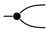
$$\mathrm{Id}[X] := \;- \tag{8}$$

The erase map $\ast[A] : A \to \{1\}$ is the map $\ast[A]_i = 1$. It is the unique Markov kernel with domain A and only one column.

Graphically, the stopper is a fuse:

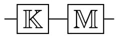
$$\mathrm{Id}[X] := \;-\!\ast \tag{9}$$

The copy map $\curlyvee[X] : X \to \Delta(X\times X)$ is the Markov kernel defined by $\curlyvee_x := \delta_x \otimes \delta_x$. Graphically it is a fork with a dot at the point where it splits:

$[X] :=$  (10)

1.4 Products

Two Markov kernels $\mathbb{L} : X \to \Delta(Y)$ and $\mathbb{M} : Y \to \Delta(Z)$ have a product $\mathbb{L}\mathbb{M} : X \to \Delta(Z)$ given by the usual matrix-matrix product: $\mathbb{L}\mathbb{M}_x^z = \sum_y \mathbb{L}_x^y \mathbb{M}_y^z$. Graphically, we write represent products by joining kernel wires together:

$\mathbb{L}\mathbb{M} :=$  (11)

1.5 Labeled Markov kernels, conditional probabilities

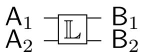
A labeled Markov kernel $(\mathbb{K}, \mathsf{A}_C, \mathsf{B}_D)$ is a Markov kernel $\mathbb{K} : X \to \Delta(Y)$ along with a sequence of *domain labels* $\mathsf{A}_C := (\mathsf{A}_i)_{i \in C}$ and *codomain labels* $\mathsf{B}_D := (\mathsf{B}_i)_{i \in D}$ such that

- Each label A_i has an associated discrete space A_i (and similarly B_i is associated with B_i)
- $X = \bigtimes_{i \in C} A_i$ and $Y = \bigtimes_{i \in D} B_i$

A labeled probability distribution $\mathbb{P} \in \Delta(Y)$ comes with a sequence of codomain labels $(\mathsf{B}_i)_{i \in D}$ only, satisfying $Y = \bigtimes_{i \in D} B_i$.

A conditional probability $\mathbb{L}[\mathsf{A}_C|\mathsf{B}_D]$ is a labeled kernel $(\mathbb{K}, \mathsf{A}_C, \mathsf{B}_D)$ along with a *background kernel* $\mathbb{L}$.

Graphically, we place the labels on the wires of a conditional probability and the name of the ambient kernel in the centre of the box:

$\mathbb{L}[\mathsf{B}_1\mathsf{B}_2|\mathsf{A}_1\mathsf{A}_2] :=$  (12)

1.6 Modelling context, extension

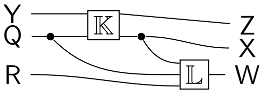
A *modelling context* $\mathcal{M}$ is a collection of conditional probabilities. It can be thought of as a namespace for the modelling work we do. A *model* is a subset of $\mathcal{M}$.

Given two conditional probabilities from a modelling context, we can extend conditional probabilities by matching labels on inputs of one with the labels on the inputs and outputs of the other. To extend conditional probabilities, we must be able to declare that one conditional probability comes before the other.

Given two conditional probabilities $\mathbb{K}[\mathsf{B}_C|\mathsf{A}_D]$ and $\mathbb{L}[\mathsf{F}_H|\mathsf{E}_G]$, we say $\mathbb{K}[\mathsf{B}_C|\mathsf{A}_D]$ is after $\mathbb{K}[\mathsf{B}_C|\mathsf{A}_D]$ if there is some label in A_D that matches a label in F_H. $\mathbb{K}[\mathsf{B}_C|\mathsf{A}_D]$ is before $\mathbb{L}[\mathsf{F}_H|\mathsf{E}_G]$ iff it is not after $\mathbb{L}[\mathsf{F}_H|\mathsf{E}_G]$.

For example, $\mathbb{K}[\mathsf{ZX}|\mathsf{YQ}]$ is before $\mathbb{L}[\mathsf{W}|\mathsf{XQR}]$. The extension of $\mathbb{K}[\mathsf{ZX}|\mathsf{YQ}]$ by $\mathbb{L}[\mathsf{W}|\mathsf{XQR}]$ is given by

$$\mathbb{K}[\mathsf{X}_1|\mathsf{Y}_1\mathsf{Q}] \rightrightarrows \mathbb{L}[\mathsf{W}_1|\mathsf{X}_1\mathsf{Q}] = \tag{13}$$



A collection of conditional probabilities from a modelling context can be joined. , let $I \subset C$ be the indices $\{i|\mathsf{A}_i \in \mathsf{E}_G\}$ and $J \subset H$ be the indices $\{i|\mathsf{B}_i \in \mathsf{F}_H\}$. $\mathbb{K}[\mathsf{B}_C|\mathsf{A}_D]$ can be joined with $\mathbb{L}[\mathsf{F}_H|\mathsf{E}_G]$ if and only if either $I = \emptyset$ or $J = \emptyset$. If $I = \emptyset$ then, letting $K = \{i|\mathsf{B}_i \in \mathsf{E}_G\}$ the result of joining is

$$(\mathbb{K}[\mathsf{B}_C|\mathsf{A}_D] * \mathbb{L}[\mathsf{E}_G|\mathsf{F}_H])_{b_J b_{JC}} \tag{14}$$

A conditional probability $\mathbb{K}[\mathsf{B}_D|\mathsf{A}_C]$ is a labeled Markov kernel $(\mathbb{K}, \mathsf{A}_C, \mathsf{B}_D)$

1.7 Sequences

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 $\curlyvee(\mathsf{X} \otimes \mathsf{Y})$. That is, $(\mathsf{X}, \mathsf{Y})_i = (\mathsf{X}_i, \mathsf{Y}_i)$.

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 $\mathcal{E}$ is the power set of E and we will henceforth only mention 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 $\mathbb{P}[\mathsf{X}] := \mathbb{P}\underline{\mathsf{X}}$.

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 *random variable* is a function $F \to Y$ for some vector space Y and a *state variable* is a function $D \to Y'$ for some vector space Y'. The *complete state variable* D is the identity function on D. Probabilities and conditional probabilities that we can define on the space $(\mathbb{K}, D, F)$ usually have to be conditioned on D, but there are some exceptions.

Something that is either a random variable or a state variable is just a *variable*.

We can associate a random variable with any state variable by copying it from the input to the output, but I think with low confidence that not doing this is going to be simpler for this paper

For each $d \in D$, any random variable $\mathsf{X} : F \to X$ has a unique marginal distribution $\mathbb{K}[\mathsf{X}|\mathsf{D}]_d := \mathbb{K}_d\underline{\mathsf{X}}$.

To save space, we say that the marginal distribution of a sequence like (X, Y) is $\mathbb{K}[\mathsf{XY}|\mathsf{D}]_d$.

1.8 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 $\mathbb{P}[\mathsf{XY}]^{ij} = P[\mathsf{X}]^i P[\mathsf{Y}|\mathsf{X}]_i^j$. 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 $(\mathbb{K}, D, F)$, complete state variable D and variables X, Y, $\mathbb{K}[\mathsf{Y}|\mathsf{XD}]$ is any Markov kernel such that $\mathbb{K}[\mathsf{XY}|\mathsf{D}]_i^{jk} = \mathbb{K}[\mathsf{X}|\mathsf{D}]_i^j \mathbb{K}[\mathsf{Y}|\mathsf{XD}]_{ij}^k$.

As previously mentioned, in the kernel space $(\mathbb{K}, D, F)$ there is in general no unique marginal distribution of (X, Y) and similarly there is generally no unique distribution of X conditioned on Y. However, such a distribution might exist if X and Y are independent of D.

1.9 Conditional independence

Given a Markov kernel space $(\mathbb{K}, D, F)$, and 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 and $\mathbb{K}[\mathsf{X}|\mathsf{YZ}]_i^j = \mathbb{K}[\mathsf{X}|\mathsf{YZ}]_{i'}^j$ for all $i, i' \in Y$.

A version of $\mathbb{K}[\mathsf{X}|\mathsf{Z}]$ exists iff $\mathsf{X} \perp\!\!\!\perp_{\mathbb{K}} \mathsf{D}|\mathsf{Z}$ or $\mathsf{Z} = \mathsf{D}$, and in the former case is given by any kernel satisfying $\mathbb{K}[\mathsf{X}|\mathsf{Z}]_i^j = \mathbb{K}[\mathsf{X}|\mathsf{DZ}]_{ik}^j$ for any version of $\mathbb{K}[\mathsf{X}|\mathsf{DZ}]$ and all $k \in D$.

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

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 $\mathbb{B} : [0, 1] \to \mathit{\Delta}(\mathcal{P}(0, 1))$ 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 1.8) 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 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 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:

$$\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D} | \mathsf{H} \tag{15}$$

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. 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. Savage's models themselves form the basis of the “decision theoretic” approach to causal inference set out by Heckerman and Shachter (1995) (where I use quotes to indicate that there are several distinct “decision theoretic” approaches in existence).

2.1 See-do models for data-driven decision problems

We can be more precise about the type of decision problem see-do models are appropriate for.

Theorem 2.3 (See-do model representation)**.** *Suppose we have a decision problem that provides us with an observation $x \in X$, and in response to this we can select any decision or stochastic mixture of decisions from a set D; that is we can choose a "strategy" as any Markov kernel $\mathbb{S} : X \to \Delta(D)$. We have a utility function $u : Y \to \mathbb{R}$ that models preferences over the potential consequences of our choice. Furthermore, suppose that we maintain a denumerable set of hypotheses H, and under each hypothesis $h \in H$ we model the result of choosing some strategy $\mathbb{S}$ as a joint probability over observations, decisions and consequences $\mathbb{P}_{h,\mathbb{S}} \in \Delta(X \times D \times Y)$.*

Define X, Y and D such that $\mathsf{X}_{xdy} = x$, $\mathsf{Y}_{xdy} = y$ and $\mathsf{D}_{xdy} = d$. Then making the following additional assumptions:

1. *Holding the hypothesis h fixed the observations as have the same distribution under any strategy: $\mathbb{P}_{h,\mathbb{S}}[\mathsf{X}] = \mathbb{P}_{h,\mathbb{S}''}[\mathsf{X}]$ for all $h, \mathbb{S}, \mathbb{S}'$ (observations are given "before" our strategy has any effect)*

2. *The chosen strategy is a version of the conditional probability of decisions given observations: $\mathbb{S} = \mathbb{P}_{h,\mathbb{S}}[\mathsf{D}|\mathsf{X}]$*

3. *There exists some strategy $\mathbb{S}$ that is strictly positive*

4. *For any $h \in H$ and any two strategies $\mathbb{Q}$ and $\mathbb{S}$, we can find versions of each disintegration such that $\mathbb{P}_{h,\mathbb{Q}}[\mathsf{Y}|\mathsf{DX}] = \mathbb{P}_{h,\mathbb{S}}[\mathsf{Y}|\mathsf{DX}]$ (our strategy tells us nothing about the consequences that we don't already know from the observations and decisions)*

Then there exists a unique see-do model $(\mathbb{T}, \mathsf{H}', \mathsf{D}', \mathsf{X}', \mathsf{Y}')$ such that $\mathbb{P}_{h,\mathbb{S}}[\mathsf{XDY}]^{ijk} = \mathbb{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbb{S}^j_i \mathbb{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ijk}$.

One thing that is worth pointing out here is that the property defining observations – that they are independent of decisions given the hypothesis – does *not* imply that observations are probabilistically independent of decisions in the decision problem modeled in the manner described above. Instead, it reflects the set of assumptions made above.

This will eventually move to an appendix

Proof. Consider some probability $\mathbb{P} \in \Delta(X \times D \times Y)$. By the definition of disintegration (section 1.8), we can write

$$\mathbb{P}[\mathsf{XDY}]^{ijk} = \mathbb{P}[\mathsf{X}]^i \mathbb{P}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{P}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \tag{16}$$

Fix some $h \in H$ and some strictly positive strategy $\mathbb{S}$ and define $\mathbb{T} : H \times D \to \Delta(X \times Y)$ by

$$\mathbb{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbb{S}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbb{S}}[\mathsf{Y}|\mathsf{XD}]^l_{kj} \tag{17}$$

Note that because $\mathbb{S}$ is strictly positive and by assumption $\mathbb{S} = \mathbb{P}_{h,\mathbb{S}}[\mathsf{D}|\mathsf{X}]$, $\mathbb{P}_{h,\mathbb{S}}[\mathsf{D}]$ is also strictly positive. Therefore $\mathbb{P}_{h,\mathbb{S}}[\mathsf{Y}|\mathsf{D}]$ is unique and therefore $\mathbb{T}$ is also unique.

Define X' and Y' by $\mathsf{X}'_{xy} = x$ and $\mathsf{Y}'_{xy} = y$. Define H' and D' by $\mathsf{H}'_{hd} = h$ and $\mathsf{D}'_{hd} = d$.

We then have

$$\begin{aligned}
\mathbb{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj} &= \mathbb{T}\underline{\mathsf{X}}'^k_{hj} && (18)\\
&= \sum_l \mathbb{T}^{kl}_{hj} && (19)\\
&= \mathbb{P}_{h,\mathbb{S}}[\mathsf{X}]^k && (20)\\
&= \mathbb{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj'} && (21)
\end{aligned}$$

Thus $\mathsf{X}' \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}'|\mathsf{H}'$ and so $\mathbb{T}[\mathsf{X}'|\mathsf{H}']$ exists (section 1.9) and $(\mathbb{T}, \mathsf{H}', \mathsf{D}', \mathsf{X}', \mathsf{Y}')$ is a see-do model.

Applying Equation 16 to $\mathbb{P}_{h,\mathbb{S}}$:

$$\begin{aligned}
\mathbb{P}_{h,\mathbb{S}}[\mathsf{XDY}]^{ijk} &= \mathbb{P}_{h,\mathbb{S}}[\mathsf{X}]^i \mathbb{P}_{h,\mathbb{S}}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{P}_{h,\mathbb{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} && (22)\\
&= \mathbb{P}_{h,\mathbb{S}}[\mathsf{X}]^i \mathbb{P}_{h,\mathbb{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} && (23)\\
&= \mathbb{P}_{h,\mathbb{S}}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{T}[\mathsf{X}'\mathsf{Y}'|\mathsf{H}'\mathsf{D}']^{ik}_{hj} && (24)\\
&= \mathbb{S}^j_i \mathbb{T}[\mathsf{X}'\mathsf{Y}'|\mathsf{H}'\mathsf{D}']^{ik}_{hj} && (25)\\
&= \mathbb{S}^j_i \mathbb{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^i_{hj} \mathbb{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} && (26)\\
&= \mathbb{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbb{S}^j_i \mathbb{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} && (27)
\end{aligned}$$

Consider some arbitrary alternative strategy $\mathbb{Q}$. By assumption

$$\begin{aligned}
\mathbb{P}_{h,\mathbb{S}}[\mathsf{X}]^i &= \mathbb{P}_{h,\mathbb{Q}}[\mathsf{X}]^i && (28)\\
\mathbb{P}_{h,\mathbb{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} &= \mathbb{P}_{h,\mathbb{Q}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \text{ for some version of } \mathbb{P}_{h,\mathbb{Q}}[\mathsf{Y}|\mathsf{XD}] && (29)
\end{aligned}$$

It follows that, for some version of $\mathbb{P}_{h,\mathbb{Q}}[\mathsf{Y}|\mathsf{XD}]$,

$$\mathbb{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbb{Q}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbb{Q}}[\mathsf{Y}|\mathsf{XD}]^l_{kj} \tag{30}$$

Then by substitution of $\mathbb{Q}$ for $\mathbb{S}$ in Equation 22 and working through the same steps

$$\mathbb{P}_{h,\mathbb{S}}[\mathsf{XDY}]^{ijk} = \mathbb{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbb{Q}^j_i \mathbb{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} \tag{31}$$

As $\mathbb{Q}$ was arbitrary, this holds for all strategies. □

3 Potential outcomes and counterfactuals

Potential outcomes is a widely used approach to causal modelling characterised by its use of "potential outcome" random variables. Potential outcome random variables are typically noted for being given counterfactual interpretations. For example, suppose have something we want to model, call it TYT ("The Y Thing"), which we represent with a variable Y. Suppose we want to know how TYT behaves under different regimes 0 and 1 under which we want to know about TYT, and we use a variable W to indicate which regime holds at a given point in time. A potential outcomes model will introduce the two additional "potential outcome" variables $(\mathsf{Y}(0), \mathsf{Y}(1))$. What these variables represent can be given a counterfactual interpretation like "$\mathsf{Y}(0)$ represents what TYT would be under regime 0, whether or not regime 0 is the actual regime" and similarly "$\mathsf{Y}(1)$ represents what TYT would be under regime 1, whether or not regime 1 is the actual regime". Note that we say "what TYT would be" rather that "what Y would be" as "what would Y be if W was 0 if W was actually 1" is not a question we can ask of random variables, but it is one that might make sense for the things we use random variables to model.

This is a key point, so it is worth restating: the assumption that potential outcome variables agree with "the value TYT would take" under fixed regimes regardless of the "actual" value of the regime seems to be a critical assumption that distinguishes potential outcome variables from arbitrary random variables that happen to take values in the same space as Y. However, this assumption can only be stated by making reference to the informally defined "TYT" and the informal distinction between the supposed and the actual value of the regime.

The potential outcomes framework features other critical assumptions that relate potential outcome variables to things that are only informally defined. For example, Rubin (2005) defines the *Stable Unit Treatment Value Assumption* (SUTVA) as:

> SUTVA (stable unit treatment value assumption) [...] comprises two subassumptions. First, it assumes that there is no interference between units (Cox 1958); that is, neither $Y_i(1)$ nor $Y_i(0)$ is affected by what action any other unit received. Second, it assumes that there are no hidden versions of treatments; no matter how unit i received treatment 1, the outcome that would be observed would be $Y_i(1)$ and similarly for treatment 0

"Versions of treatments" do not appear within typical potential outcomes models, so this is also an assumption about how "the thing we are trying to model" behaves rather than an assumption stated within the model.

Given informal assumptions like this, one may be motivated to "formalize" them. More specifically, one might be motivated to ask whether there is some larger class of models that, under conditions corresponding to the informal conditions above yield regular potential outcome models?

I have a vague intuition here that you always need some kind of assumption like "my model is faithful to the real thing", but if you are stating fairly specific conditions in English you should also be able to state them mathematically. Among other reasons, this is useful because it's easier for other people to know what you mean when you state them.

The approach we have introduced here, motivated by decision problems, has in the past been considered a means of avoiding counterfactual statements, which has been considered a positive by some (Dawid, 2000) and a negative by others:

> [...] Dawid, in our opinion, incorrectly concludes that an approach to causal inference based on "decision analysis" and free of counterfactuals is completely satisfactory for addressing the problem of inference about the effects of causes.(Robins and Greenland, 2000)

It may be surprising to some, then, that we can use see-do models to formally state these key assumptions associated with potential outcomes models. Furthermore, we will argue that potential outcomes are typically a strategy to motivate inductive assumptions in see-do models, and we will show that the counterfactual interpretation is unnecessary for this purpose.

3.1 Potential outcomes in see-do models

A basic property of potential outcomes models is the relation between variables representing actual outcomes and variables representing potential outcomes, which was stated informally in the opening paragraph of this section.

In the following definition, $\mathsf{Y}(W) = (\mathsf{Y}(w))_{w \in W}$.

Definition 3.1 (Potential outcomes)**.** Given a Markov kernel space $(\mathbb{K}, E, F)$, a collection of variables $\{\mathsf{Y}, \mathsf{Y}(W), \mathsf{W}\}$ where Y and $\mathsf{Y}(W)$ are random variables and W could be either a state or a random variable is a *potential outcome submodel* if $\mathbb{K}[\mathsf{Y}|\mathsf{W}\mathsf{Y}(W)]$ exists and $\mathbb{K}[\mathsf{Y}|\mathsf{W}\mathsf{Y}(W)]_{ij_1 j_2 \ldots j_{|W|}} = \delta[j_i]$.

We allow X to be a state or a random variable to cover the cases where potential outcomes models feature as submodels of observation models (in which case X is a random variable) or as submodels of consequence models (in which case X may be a state variable).

As an aside that we could define stochastic potential outcomes if we allow the variables $\mathsf{Y}(x)$ to take values in $\Delta(Y)$ rather than in Y, and then require $\mathbb{K}[\mathsf{Y}|\mathsf{X}\mathsf{Y}(X)]_{ij_1 j_2 \ldots j_{|X|}} = j_i$ (where j_i is an element of $\Delta(Y)$). This is more complex to work with and rarely seen in practice, but it is worth noting that Definition 3.1 can be generalised to cover models where $\mathsf{Y}(x)$ describes the value Y would take if X were x *with uncertainty.*

An arbitrary see-do model featuring potential outcome submodels does not necessarily allow for the formal statement of the counterfactual interpretation of potential outcomes. Here we use TYT ("the actual thing") and "regime" to refer to the things we are actually trying to model. We require that $\mathsf{Y} \overset{a.s.}{=} \mathsf{Y}(w)$ conditioned on $\mathsf{W} = w$. If we add an interpretation to this model saying Y represents TYT and W represents the regime, then we have "for all w, $\mathsf{Y}(w)$ is equal to Y which represents TYT under the regime w". However, this does not guarantee that our model has anything that reasonably represents "what TYT would be equal to under supposed regime w if the regime is actually w'".

We propose *parallel potential outcome submodels* as a means of formalising statements about what how TYT behaves under "supposed" and "actual" regimes:

Definition 3.2 (Parallel potential outcomes)**.** Given a Markov kernel space $(\mathbb{K}, E, F)$, a collection of variables $\{\mathsf{Y}_i, \mathsf{Y}(W), \mathsf{W}_i\}$, $i \in [n]$, where Y_i and $\mathsf{Y}(W)$ are random variables and W_i could be either a state or random variables is a *parallel potential outcome submodel* if $\mathbb{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]$ exists and $\mathbb{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]_{kj_1j_2\ldots j_{|W|}} = \delta[j_k]$.

A parallel potential outcomes model features a sequence of n "parallel" outcome variables Y_i and n "regime proposals" W_i, with the property that if the regime proposal $\mathsf{W}_i = w_i$ then the corresponding outcome $\mathsf{Y}_i \overset{a.s.}{=} \mathsf{Y}(w_i)$. We can identify a particular index, say $n = 1$, with the actual world and the rest of the indices with supposed worlds. Thus Y_1 represents the value of TYT in the actual world and Y_i $i \neq 1$ represents TYT under a supposed regime W_i. Given such an interpretation, the fact that $\mathsf{Y}_i \overset{a.s.}{=} \mathsf{Y}(w_i)$ can be interpreted as assuming "for all w, if the supposed regime W_i is w then the corresponding outcome will be almost surely equal to $\mathsf{Y}(w)$, regardless of the value of the actual regime W_1", which is our original counterfactual assumption.

We do not intend to defend this as the only way that counterfactuals can be modeled, or even that it is appropriate to capture the idea of counterfactuals at all. It is simply a way that we can model the counterfactual assumption typically associated with potential outcomes. We will show show that parallel potential outcome submodels correspond precisely to *extendably exchangeable* and *deterministically reproducible* submodels of Markov kernel spaces.

3.2 Parallel potential outcomes representation theorem

Exchangeble sequences of random variables are sequences whose joint distribution is unchanged by permutation. Independent and identically distributed random variables are one example: if X_1 is the result of the first flip of a coin that we know to be fair and X_2 is the second flip then $\mathbb{P}[\mathsf{X}_1\mathsf{X}_2] = \mathbb{P}[\mathsf{X}_2\mathsf{X}_1]$. There are also many examples of exchangeable sequences that are not mutually independent and identically distributed – for example, if we want to use random variables Y_1 and Y_2 to model our subjective uncertainty regarding two flips of a coin of unknown fairness, we regard our initial uncertainty for each flip to be equal $\mathbb{P}[\mathsf{Y}_1] = \mathbb{P}[\mathsf{Y}_2]$ and we our state of knowledge of the second flip after observing only the first will be the same as our state of knowledge of the first flip after observing only the second $\mathbb{P}[\mathsf{Y}_2|\mathsf{Y}_1] = \mathbb{P}[\mathsf{Y}_1|\mathsf{Y}_2]$, then our model of subjective uncertainty is exchangeable.

De Finetti's representation theorem establishes the fact that any infinite exchangeable sequence $\mathsf{Y}_1, \mathsf{Y}_2, ...$ can be modeled by the product of a *prior* probability $\mathbb{P}[\mathsf{J}]$ with J taking values in the set of marginal probabilities $\Delta(Y)$ and a conditionally independent and identically distributed Markov kernel $\mathbb{P}[\mathsf{Y}_A|\mathsf{J}]_j^{y_A} = \prod_{i\in A} \mathbb{P}[\mathsf{Y}_1|\mathsf{J}]_j^{y_i}$.

We extend the idea of exchangeable sequences to cover both random variables and state variables, and we show that a similar representation theorem holds for potential outcomes. De Finetti's original theorem introduced the variable J that took values in the set of marginal distributions over a single observation; the set of potential outcome variables plays an analagous role taking values in the set of functions from propositions to outcomes.

The representation theorem for potential outcomes is somewhat simpler that De Finetti's original theorem due to the fact that potential outcomes are usually assumed to be *deterministically reproducible*; in the parallel potential outcomes model, this means that for $j \neq i$, if W_j and W_i are equal then Y_j and Y_i will be almost surely equal. This assumption of determinism means that we can avoid appeal to a law of large numbers in the proof of our theorem.

An interesting question is whether there is a similar representation theorem for potential outcomes without the assumption of deterministic reproducibility. I'm reasonably confident that this is a straightforward corollary of the representation theorem proved in my thesis. However, this requires maths not introduced in this draft of the paper.

Extendably exchangeable sequences can be permuted without changing their conditional probabilities, and can be extended to arbitrarily long sequences while maintaining this property. We consider here sequences that are exchangeable conditional on some variable; this corresponds to regular exchageability if the conditioning variable is $*$ where $*_i = 1$.

Definition 3.3 (Exchangeability)**.** Given a Markov kernel space $(\mathbb{K}, E, F)$, a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ with Y_i random variables is *exchangeable* conditional on Z if, defining $\mathsf{Y}_{[n]} = (\mathsf{Y}_i)_{i\in[n]}$ and $\mathsf{D}_{[n]} = (\mathsf{D}_i)_{i\in[n]}$, $\mathbb{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$ exists and for any bijection $\pi : [n] \to [n]$ $\mathbb{K}[\mathsf{Y}_{\pi([n])}|\mathsf{D}_{\pi([n])}\mathsf{Z}] = \mathbb{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$.

Definition 3.4 (Extension)**.** Given a Markov kernel space $(\mathbb{K}, E, F)$, $(\mathbb{K}', E', F')$ is an *extension* of $(\mathbb{K}, E, F)$ if there is some random variable X and some state variable U such that $\mathbb{K}'[\mathsf{X}|\mathsf{U}]$ exists and $\mathbb{K}'[\mathsf{X}|\mathsf{U}] = \mathbb{K}$.

If $(\mathbb{K}', E', F')$ is an extension of $(\mathbb{K}, E, F)$ we can identify any random variable Y on $(\mathbb{K}, E, F)$ with $\mathsf{Y} \circ \mathsf{X}$ on $(\mathbb{K}', E', F')$ and any state variable D with $\mathsf{D} \circ \mathsf{U}$ on $(\mathbb{K}', E', F')$ and under this identification $\mathbb{K}'[\mathsf{Y} \circ \mathsf{X}|\mathsf{D} \circ \mathsf{E}]$ exists iff $\mathbb{K}[\mathsf{Y}|\mathsf{D}]$ exists and $\mathbb{K}'[\mathsf{Y} \circ \mathsf{X}|\mathsf{D} \circ \mathsf{E}] = \mathbb{K}[\mathsf{Y}|\mathsf{D}]$. To avoid proliferation of notation, if we propose $(\mathbb{K}, E, F)$ and later an extension $(\mathbb{K}', E', F')$, we will redefine $\mathbb{K} := \mathbb{K}'$ and $\mathsf{Y} := \mathsf{Y} \circ \mathsf{X}$ and $\mathsf{D} := \mathsf{D} \circ \mathsf{E}$.

I think this is a very standard thing to do – propose some X and $\mathbb{P}(\mathsf{X})$ then introduce some random variable Y and $\mathbb{P}(\mathsf{XY})$ as if the sample space contained both X and Y all along.

Definition 3.5 (Extendably exchangeable)**.** Given a Markov kernel space $(\mathbb{K}, E, F)$, a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ and a state variable Z with Y_i random variables is *extendably exchangeable* if there exists an extension of $\mathbb{K}$ with respect to which $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in\mathbb{N}}$ is exchangeable conditional on Z.

Here that we identify Z and $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ defined on the extension with the original variables defined on $(\mathbb{K}, E, F)$ while $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in\mathbb{N}\setminus[n]}$ may be defined only on the extension.

Deterministically reproducible sequences have the property that repeating the same decision gets the same response with probability 1. This could be a model of an experiment that exhibits no variation in results (e.g. every time I put green paint on the page, the page appears green), or an assumption about collections of "what-ifs" (e.g. if I went for a walk an hour ago, just as I actually did, then I definitely would have stubbed my toe, just like I actually did). Incidentally, many consider that this assumption is false concering what-if questions about things that exhibit quantum behaviour.

Definition 3.6 (Deterministically reproducible)**.** Given a Markov kernel space $(\mathbb{K}, E, F)$, a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ with Y_i random variables is *deterministically reproducible* conditional on Z if $n \geq 2$, $\mathbb{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$ exists and $\mathbb{K}[\mathsf{Y}_{\{i,j\}}|\mathsf{D}_{\{i,j\}}\mathsf{Z}]^{lm}_{kk} = [\![l = m]\!]\mathbb{K}[\mathsf{Y}_i|\mathsf{D}_i\mathsf{Z}]^l_k$ for all i, j, k, l, m.

Theorem 3.7 (Potential outcomes representation)**.** *Given a Markov kernel space $(\mathbb{K}, E, F)$ along with a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ with $n \geq 2$ and a conditioning variable Z, $(\mathbb{K}, E, F)$ can be extended with a set of variables $\mathsf{Y}(D) := (\mathsf{Y}(i))_{i\in D}$ such that $\{\mathsf{Y}_i, \mathsf{Y}(D), \mathsf{D}_i\}$ is a parallel potential outcome submodel if and only if $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ is extendably exchangeable and deterministically reproducible conditional on Z.*

Proof. If: Because $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ is extendably exchangeable, we can without loss of generality assume $n \geq |D|$.

Let $e = (e_i)_{i\in[|D|]}$. Introduce the variable $\mathsf{Y}(i)$ for $i \in D$ such that $\mathbb{K}[\mathsf{Y}(D)|\mathsf{D}_{[D]}\mathsf{Z}]_{ez} = \mathbb{K}[\mathsf{Y}_D|\mathsf{D}_D\mathsf{Z}]_{ez}$ and introduce X_i, $i \in D$ such that $\mathbb{K}[\mathsf{X}_i|\mathsf{D}_i\mathsf{Z}\mathsf{Y}(D)]^{x_i}_{e_i z j_1 \ldots j_{|D|}} = \delta[j_{e_i}]^{x_i}$. Clearly $\{\mathsf{X}_{[n]}, \mathsf{D}_{[n]}, \mathsf{Y}(D)\}$ is a parallel potential outcome submodel. We aim to show that $\mathbb{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}] = \mathbb{K}[\mathsf{X}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$.

Let $y := (y_i)_{i\in|D|} \in Y^{|D|}$, $d := (d_i)_{i\in[n]} \in D^{[n]}$, $x := (x_i)_{i\in[n]} \in Y^{[n]}$.

$$\mathbb{K}[\mathsf{X}_n|\mathsf{D}_n\mathsf{Z}]^x_{dz} = \sum_{y\in Y^{|D|}} \mathbb{K}[\mathsf{X}_{[n]}|\mathsf{D}_n\mathsf{Z}\mathsf{Y}(D)]^x_{dzy}\mathbb{K}[\mathsf{Y}(D)|\mathsf{D}_{[n]}\mathsf{Z}]^y_{dz} \tag{32}$$

$$= \sum_{y\in Y^{|D|}} \prod_{i\in[n]} \delta[y_{d_i}]^{x_i}\mathbb{K}[\mathsf{Y}(D)|\mathsf{D}_n\mathsf{Z}]^y_{dz} \tag{33}$$

Wherever $d_i = d_j := \alpha$, every term in the above expression will contain the product $\delta[\alpha]^{x_i}\delta[\alpha]^{x_j}$. If $x_i \neq x_j$, this will always be zero. By deterministic reproducibility, $d_i = d_j$ and $x_i \neq x_j$ implies $\mathbb{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]_d z^x = 0$ also. We need to check for equality for sequences x and d such that wherever $d_i = d_j$, $x_i = x_j$. In this case, $\delta[\alpha]^{x_i}\delta[\alpha]^{x_j} = \delta[\alpha]^{x_i}$. Let $Q_d \subset [n] := \{i | \not\exists i \in [n] : j < i \;\&\; d_j = d_i\}$, i.e. Q is the set of all indices such that d_i is the first time this value appears in d. Note that Q_d is of size at most $|D|$. Let $Q_d^C = [n] \setminus Q_d$, let $R_d \subset D : \{d_i | i \in Q_d\}$ i.e. all the elements of D that appear at least once in the sequence d and let $R_d^C = D \setminus R_d$.

Let $y' = (y_i)_{i \in Q_d^C}$, $x_{Q_d} = (x_i)_{i \in Q_d}$, $\mathsf{Y}(R_d) = (\mathsf{Y}_d)_{d \in R_d}$ and $\mathsf{Y}(S_d) = (\mathsf{Y}_d)_{d \in S_d}$.

$$\mathbb{K}[\mathsf{X}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]^x_{dz} = \sum_{y \in Y^{|D|}} \prod_{i \in Q_d} \delta[y_{d_i}]^{x_i} \mathbb{K}[\mathsf{Y}(D)|\mathsf{D}_{[n]}\mathsf{Z}]^y_{dz} \tag{34}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbb{K}[\mathsf{Y}(R_d)\mathsf{Y}(R_d^C)|\mathsf{D}_{Q_d}\mathsf{D}_{Q_d^C}\mathsf{Z}]^{x_{Q_d}y'}_{d_{Q_d}d^C_{Q_d}z} \tag{35}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbb{K}[\mathsf{Y}_{R_d}\mathsf{Y}_{R_d^C}|\mathsf{D}_{Q_d}\mathsf{D}_{Q_d^C}\mathsf{Z}]^{x_{Q_d}y'}_{dz} \tag{36}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbb{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]^{x_{Q_d}y'}_{dz} \qquad \text{(using exchangeability)} \tag{37}$$

Note that

Only if: We aim to show that the sequences $\mathsf{Y}_{[n]}$ and $\mathsf{D}_{[n]}$ in a parallel potential outcomes submodel are exchangeable and deterministically reproducible. □

References

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

A. P. Dawid. Causal Inference without Counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424, June 2000. ISSN 0162-1459. doi: 10.1080/01621459.2000.10474210. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.2000.10474210. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.2000.10474210.

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`.

D. Heckerman and R. Shachter. Decision-Theoretic Foundations for Causal Reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, December 1995. ISSN 1076-9757. doi: 10.1613/jair.202. URL `https://www.jair.org/index.php/jair/article/view/10151`.

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 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.

James M. Robins and Sander Greenland. Causal Inference Without Counterfactuals: Comment. *Journal of the American Statistical Association*, 95 (450):431–435, 2000. ISSN 0162-1459. doi: 10.2307/2669381. URL `http://www.jstor.org/stable/2669381`. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Donald B. Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, March 2005. ISSN 0162-1459. doi: 10.1198/016214504000001880. URL `https://doi.org/10.1198/016214504000001880`.

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

Peter Selinger. A survey of graphical languages for monoidal categories. *arXiv:0908.3347 [math]*, 813:289–355, 2010. doi: 10.1007/978-3-642-12821-9_4. URL `http://arxiv.org/abs/0908.3347`. arXiv: 0908.3347.

Appendix: