Causal Inference: a Bayesian Perspective

Fabrizia Mealli

Department of Statistics, Computer Science, Applications
University of Florence
mealli@disia.unifi.it

BAYSM, Florence, June 19-21, 2016

Introduction

- Research questions that motivate most studies in statistics-based sciences are causal in nature
- What can statistics say about causation?
- How can Bayesian inference help in questions about causation?
- The usual motto is "correlation is not causation"
- Dominant methodology has excluded causal vocabulary both from its mathematical language and from its educational programs
- Yet, statisticians invented randomized experiments, universally recognized as a powerful aid in investigating causal relationships

- Statistics has a great deal to say about certain problems of causal inference
- Statistical models used to draw **causal inferences** are different from those commonly used to draw **associational inferences**
- Variety of questions under causality heading
  - ✓ the philosophical meaningfulness of the notion of causation
  - ✓ deducing the causes of a given effect
  - ✓ understanding the details of a causal mechanism
- I will focus on measuring the effects of causes because this seems to be a place where statistics, which is concerned with measurement, has major contributions to make

- The purpose is to present a model that is complex enough to allow us to formalize basic intuitions concerning causes and effects, to define causal effects and to make assumptions allowing estimation of such effects clear and explicit
- A statistical framework for causal inference is the one based on potential outcomes.
  - ✓ It is rooted in the statistical work on randomized experiments by Fisher (1918, 1925) and Neyman (1923), as extended by Rubin (1974, 1976, 1977, 1978, 1990) and subsequently by others to apply to nonrandomized studies and other forms of inference
- This perspective was called "Rubin's Causal Model" by Holland (1986) because it viewed causal inference as a problem of missing data, with explicit mathematical modeling of the assignment mechanism as a process for revealing the observed data.

Associational Inference vs Causal Inference

- Standard statistical models for associational inference relate two (or more) variables in a population
- The two variables, say Y and A, are defined for each and all units in the population and are logically on equal footing
- Joint distribution of Y and A
- Associational parameters are determined by this joint distribution: for example, the conditional distribution of Y given A describes how the distribution of Y changes as A varies
- A typical associational parameter is the regression of Y on A, that is, the conditional expectation $E(Y|A)$
- Associational inference is simply descriptive
- Role of time

Associational Inference vs Causal Inference

- Causal inference is different
- Use of language of experiments
- Model for causal inference starts with a population of units (persons, places, or things at a particular point in time) upon which a cause or a treatment may operate or act
- A single person, place, or thing at two different times comprises two different units
- The terms **cause** and **treatment** will be used interchangeably
- The effect of a cause is almost always relative to another cause: "A causes B" means relative to some other condition that may include "not A"
- The language of experiments: "treatment" vs "control"
- The key notion in causal inference is that each unit is potentially exposable to any one of the causes.
  - ✓ "She did well in the math test because she received good teaching"
  - ✓ "She did well in the math test because she is a girl"

Introducing Model and Notation

- Let W be the variable that indicates the treatment, 0 or 1, to which each unit is exposed
- The critical feature of the notion of a cause is that the value of W for each unit **could have been different**
- W must be a variable that is, at least in principle, manipulable
- Role of time: the fact that a unit is exposed to a cause or treatment must occur at a specific time
- Pre-exposure or pre-treatment variables, sometimes labelled covariates, X, whose values are determined prior to exposure to the cause
- Post-exposure or response variables, Y, on which to measure the effect of the cause

Introducing Model and Notation

- To represent the notion of causation, we postulate the existence of two variables, $Y(1)$ and $Y(0)$ for each unit, which represent the potential responses or potential outcomes associated with the two treatments
- These are the values of a unit's measurement of interest after (a) application of the treatment and (b) non-application of the treatment (i.e., under control)
- A causal effect is, for each unit, the comparison of the potential outcome under treatment and the potential outcome under control
- For example, we can say that treatment 1 (relative to treatment 0) causes the effect $Y_i(1) - Y_i(0)$ for unit i

The Science

Units	Covariates X	Potential $Y(1)$	Outcomes $Y(0)$	Unit-level Causal Effects	Summary Causal Effects
1	X_1	$Y_1(1)$	$Y_1(0)$	$Y_1(1)$ vs $Y_1(0)$	Comparison of
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$Y_i(1)$ vs $Y_i(0)$
i	X_i	$Y_i(1)$	$Y_i(0)$	$Y_i(1)$ vs $Y_i(0)$	for a common
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	set of units
N	X_N	$Y_N(1)$	$Y_N(0)$	$Y_N(1)$ vs $Y_N(0)$	

- "The fundamental problem of causal inference": each potential outcome is observable but we can never observe all of them
- Summary causal effects: the critical requirement is that for a comparison to be causal it must be a comparison of $Y_i(1)$ and $Y_i(0)$ on a common set of units

SUTVA

- The table for the Science requires the Stable Unit Treatment Value Assumption (SUTVA, Rubin, 1990) to be adequate
- No interference between units, that is, neither $Y_i(1)$ nor $Y_i(0)$ is affected by what action any other units received
- No hidden version of treatments: no matter how unit i received treatment 1, the outcome that would be observed would be $Y_i(1)$
- Also implicit in the representation is that the Science is not affected by how or whether we try to learn about it, whether by randomized block designs, observational studies or other methods

SUTVA and Other Assumptions

- Without these assumptions causal inference using potential outcomes is not impossible, but it is far more complicated
- SUTVA is commonly made, or studies are designed to make SUTVA plausible
- Bayesian inference can help addressing causal questions in the presence of interference (Forastiere et al., 2016)
- Nothing is wrong with making assumptions and causal inference is impossible without making assumptions; assumptions are the strand that links statistics to science
- It is the scientific quality of the assumptions, not their existence, that is critical
- In causal inference assumptions are always needed, and they typically do not generate testable implications, so it is imperative that they are explicated and justified

Scientific and Statistical Solutions

- Because at least half of the potential outcomes are always missing, as such, the fundamental problem of causal inference is not solved by observing more units
- The notation explicitly representing both potential outcomes is an exceptional contribution to causal inference
- Despite its apparent simplicity it did not arise until 1923 with the work of Neyman and only in the context of completely randomized experiments
- We had to wait until the seventies with the work of Rubin to use the notation of potential outcomes to describe causal effects in any setting, including observational studies
- Despite the fundamental problem of causal inference, there are some solutions to the fundamental problem

Scientific and Statistical Solutions

- The scientific solution exploits various homogeneity or invariance assumptions
  - ✓ $Y_i^{t-1}(0) = Y_i^t(0)$
  - ✓ Then, expose units to 1 and measure $Y_i(1)$
  - ✓ The scientist has made an untestable homogeneity assumption
- Science has made enormous progress using this approach, and it is the approach that we informally use often in our lives
- The statistical solution uses the observed values of W and $Y(W)$, together with assumptions about the way units where exposed to either $W = 1$ or $W = 0$ to address the problem

The Role of the Assignment Mechanism

- The key in Rubin's work is to see randomization as just one way to create missing and observed data in the potential outcomes
- There are many other processes for creating missing data and those were called **assignment mechanisms** (Rubin, 1978)
- The assignment mechanism gives the probability of each vector of assignments, W, given the Science:

$$Pr(W \mid X, Y(1), Y(0))$$

- Before Rubin (1975), there were written descriptions of assignment mechanisms, but no formal mathematical statement or notation showing the possible dependence of treatment assignments on BOTH potential outcomes
- Y_{obs}: the collection of observed potential outcomes, with $Y_{obs,i} = W_i Y_i(1) + (1 - W_i) Y_i(0)$
- Y_{mis}: the collection of missing or unobserved potential outcomes, with $Y_{mis,i} = (1 - W_i) Y_i(1) + W_i Y_i(0)$

The Role of the Assignment Mechanism

- The definition of the assignment mechanism states that probability of something that we *do now*, W, can depend, not only on things that we observe now, X, or even Y_{obs} in sequential experiments, but moreover on other things that will never even be realized, Y_{mis}. Yet, as a formal probability statement, it is mathematically coherent
- Understanding the assignment mechanism's possible dependence on values of the potential outcomes: think of unobserved - to the analyst of the data - covariates U that are associated with the future potential outcomes and are used by the assigner of treatments, hypothetical or real, in addition to X
- $Pr(W \mid X, Y(1), Y(0), U) = Pr(W \mid X, U)$
- When this expression is averaged over the values of U for fixed values of X, $Y(1)$, $Y(0)$ to calculate the assignment mechanism, the result yields dependence on $Y(1)$, $Y(0)$

Types of Assignment Mechanism

- The assignment mechanism is unconfounded (with the potential outcomes, Rubin, 1990) if:

$$Pr(W \mid X, Y(1), Y(0)) = Pr(W \mid X)$$

- An unconfounded assignment mechanism is probabilistic if all the unit-level probabilities, the propensity scores (Rosenbaum and Rubin, 1983), are strictly between zero and one:

$$0 < e_i = Pr(W_i \mid X) < 1$$

- An unconfounded probabilistic assignment mechanism is called strongly ignorable
- Classical randomized experiments are special cases of strongly ignorable assignment mechanisms
- In observational studies the assignment mechanism is not known and we need to make assumptions in order to be able to draw inference on causal effects
- Design stage of observational studies

Distinguishing between the Science and the Assignment Mechanism

- Using the potential outcomes notation maintains the critical distinction between **what we are trying to estimate**, the Science, and **what we do to learn about it**, the assignment mechanism
- This distinction was in the work of Neyman or Fisher, so that extensions to observational studies of classical methods of inference in randomized experiments, due to Fisher (1925) and Neyman (1923), are natural within the RCM framework

Modes of Inference: Causal Inference Based Solely on the Assignment Mechanism

- Both Fisher and Neyman proposed methods of causal inference based solely on the randomization distribution of statistics induced by classical randomized assignment mechanisms
- **Fisher's Exact p-values for Sharp Null Hypotheses**
- Fisher's method was essentially a stochastic proof by contradiction
- He wanted to prove that $H_0 = Y_i(1) = Y_i(0) \forall i$ is wrong using the randomization distribution under H_0

Modes of Inference: Causal Inference Based Solely on the Assignment Mechanism

- **Neyman's Randomization-Based Estimates and Confidence Intervals**
- Neyman (1923) showed that, in a completely randomized experiment, $\bar{y}_1 - \bar{y}_0$ is unbiased (averaging over all randomizations) for the average causal effect and propose a large-sample interval estimate for the average causal effect, which became the standard one in much of statistics and applied fields
- Neyman's approach has advantages over Fisher's in that it can deal with random sampling of units from a population; much of the theory behind propensity score methods is generalization of Neyman's approach
- Fisher's approach has the obvious advantage in not requiring large samples for the exactness of its probabilistic statements
- Fisher's and Neyman's approaches rarely addressed the real reasons we conduct studies: to learn about which interventions should be applied to future units
- **The third leg of the RCM is critical: pose a model on the Science and derives the Bayesian posterior predictive distribution of the missing potential outcomes**

Elements of the RCM

- The first leg is using potential outcomes to define causal effects no matter how we try to learn about them: *First define the Science*
- The second leg is to describe the process by which some potential outcomes will be revealed: *Second, posit an assignment mechanism*
- The third leg is placing a probability distribution on the Science to allow formal probability statements about the causal effects: *Third, incorporate scientific understanding in a model for the Science.*
- The Bayesian approach directs us to condition on all observed quantities and predicts, in a stochastic way, the missing potential outcomes of all units, past and future, and thereby makes informed decisions

Bayesian Model-Based Imputation

- The benefits of modeling the science in causal inference include the ability to deal with more complex situations and to summarize results more logically
- We directly confront the fact that at least half of the potential outcomes are missing and create a posterior predictive distribution for them
- From a model on the science, $Pr(X, Y(1), Y(0))$, and the model for the assignment mechanism, we can find the posterior predictive distribution of Y_{mis}, given the observed values of W, X, and Y_{obs}

$$Pr(Y_{mis}|X, Y_{obs}, W) \propto Pr(X, Y(1), Y(0))Pr(W|X, Y(1), Y(0))$$

- We can calculate the posterior distribution of any causal estimand by multiply imputing Y_{mis}: draw a value of Y_{mis}, impute it, calculate the causal estimand, redraw Y_{mis}, and so on

Bayesian Model-Based Imputation

- Two critical facts simplify this approach

$$Pr(X, Y(0), Y(1)) = \int \prod f(X_i, Y_i(0), Y_i(1)|p(\theta)d\theta,$$

where $f(.|\theta)$ is an iid model for each unit's science given a hypothetical parameter θ with prior (or marginal) distribution $p(\theta)$

- This modelling task if far more flexible than specifying a regression model
- If the treatment assignment mechanism is ignorable then when the expression for the assignment mechanism is evaluated at the observed data, it is free of dependence on Y_{mis}.
- So the explicit conditioning on W can be ignored (hence the term ignorable assignment mechanism):

$$Pr(Y_{mis}|X, Y_{obs}, W) \propto Pr(Y_{mis}|X, Y_{obs})$$
$$Pr(Y_{mis}|X, Y_{obs}) = \int Pr(Y_{mis}|X, Y_{obs}, \theta)Pr(\theta, X, Y_{obs})d\theta$$

where $Pr(\theta|X, Y_{obs})$ is the posterior distribution of θ, equal to the prior distribution $p(\theta)$ times the likelihood of θ

Bayesian Model-Based Imputation

- Thus by supplementing the assignment mechanism with a model on the science, we can adopt, a Bayesian framework to inference for causal effects
- The Bayesian perspective is extremely flexible and is especially convenient for summarizing the current state of knowledge about the science in complex situations
- Assuming this summary of the current state of knowledge is accurate, this can be combined with various assessment of costs and benefits of various decisions to choose which decision to make (Deheija, 2003)

Extensions

- The potential outcome framework combined with Bayesian inference allowed us to make enormous progress in formalizing and solving problems in both randomized and observational studies
- The framework allowed to understand the meaning of IV estimation developed in Econometrics, by bridging randomized experiments with noncompliance with IV settings
- It provided insights into understanding causal mechanisms through *principal stratification*, an approach to handling intermediate variables within the RCM

Principal Stratification

Many scientific problems require that treatment comparisons be "adjusted" for post-treatment variables

- Treatment noncompliance
- Missing outcomes (dropout)
- Censoring (or truncation) by death
- Surrogate or biomarker endpoints
- Combinations of complications
- Understanding the causal pathways by which a treatment affects an outcome: Associate & Dissociative Effects - Direct & Indirect Effects

"Endogenous" selection problems

Principal Stratification

- Principal stratification gained used for defining causal estimands of interest
- Methods of inference differ a lot depending on assumptions one is willing to pose and the type of inference one wants
- Likelihood/Bayesian model-based inference as advocated in Imbens and Rubin (1997)
- Inference on the causal estimands of interest is complicated by the fact that we only observed mixture of distributions: we need to disentangle these mixtures

Basic Notation

- Units: $i = 1, \ldots, N$
- Z_i = Binary treatment assignment:

$$Z_i = z \in \{0, 1\} = \{\text{Control Treatment}, \text{Active Treatment}\}$$

- Under SUTVA, each unit i has two potential outcomes for each post-treatment variable
- $S_i(z)$: Potential outcome for the intermediate post-treatment variable given assignment to treatment z
- $Y_i(z, S_i(z)) = Y_i(z)$: Potential outcome for the primary endpoint given assignment to treatment z with intermediate variable equal to $S_i(z)$
- Observed variables

$$\begin{aligned} X_i &: \text{ vector of pre-treatment variables} \\ Z_i &: \text{ actual treatment assignment} \\ S_i^{obs} &= S_i(Z_i) = Z_i S_i(1) + (1 - Z_i) S_i(0) : \text{ observed intermediate outcome} \\ Y_i^{obs} &= Y_i(Z_i) = Z_i Y_i(1) + (1 - Z_i) Y_i(0) : \text{ realized outcome} \end{aligned}$$

Principal Stratification: Definition

If the post-treatment variable S is binary, subjects can be classified into four groups according to the joint potential values of the intermediate potential variable, $(S_i(0), S_i(1))$:

$$\begin{aligned} 00 &= \{i : S_i(0) = 0, S_i(1) = 0\} \\ 10 &= \{i : S_i(0) = 1, S_i(1) = 0\} \\ 01 &= \{i : S_i(0) = 0, S_i(1) = 1\} \\ 11 &= \{i : S_i(0) = 1, S_i(1) = 1\} \end{aligned}$$

This cross-classification of units is the (basic) principal stratification with respect to the (binary) post-treatment variable S. Formally,

Principal Stratification and Principal Effects (Frangakis and Rubin, 2002)

The basic principal stratification P_0 with respect to post-treatment variable S is the partition of units $i = 1, \ldots, n$ such that, all units within any set of P_0 have the same vector $(S_i(0), S_i(1))$

A principal stratification P with respect to post-treatment variable S is a partition of the units whose sets are unions of sets in the basic principal stratification P_0.

Let P be a principal stratification with respect to the post-treatment variable S. Then a principal effect with respect to that principal stratification is defined as a comparison of potential outcomes under control versus active treatment within a principal stratum g in P, that is, a comparison between the ordered sets

$$\left\{ Y_i(0) : i \in g \right\} \qquad \text{and} \qquad \left\{ Y_i(1) : i \in g \right\}$$

Properties of Principal Strata and Principal Effects

- The principal stratum $G_i \equiv (S_i(0), S_i(1))$ to which unit i belongs, is not affected by treatment assignment for any principal stratification P
  - ✓ The value of the ordered pair $(S_i(0), S_i(1))$ is not affected by treatment assignment
  - ✓ Principal stratum membership only reflects subject's characteristics: it can be regarded as a pre-treatment variable
- Principal effects are properly defined causal effects because they are defined as comparison of potential outcomes $Y_i(0)$ and $Y_i(1)$ on a common set of units: the (union of) principal strata

Dissociative and Associative Principal Effects

- Dissociative Principal Effects: An effect on outcome that is dissociative with an effect on intermediate variable is defined as a comparison between the ordered sets

$$\left\{Y_i(0) : S_i(0) = S_i(1)\right\} \qquad \text{and} \qquad \left\{Y_i(1) : S_i(0) = S_i(1)\right\}$$

- Associative Principal Effects: An effect on outcome that is associative with an effect on intermediate variable is defined as a comparison between the ordered sets

$$\left\{Y_i(0) : S_i(0) \neq S_i(1)\right\} \qquad \text{and} \qquad \left\{Y_i(1) : S_i(0) \neq S_i(1)\right\}$$

The role of unconfounded assignment mechanisms in principal stratification analysis

If treatment assignment is uncounfouded:

$$P\left(Z_i|S_i(0), S_i(1), Y_i(0), Y_i(1), X_i\right) = P\left(Z_i|X_i\right),$$

then

- the principal stratum membership $G_i \equiv (S_i(0), S_i(1))$ is guaranteed to have the same distribution in both treatment arms (within cells defined by pre-treatment variables):

$$G_i \perp Z_i | X_i$$

- potential outcomes are independent of the treatment assignment given the principal strata:

$$Y_i(0), Y_i(1) \perp Z_i | G_i, X_i$$

  ✓ Treated and control units can be compared conditional on a principal stratum. It is this *latent* regularity that is exploited in IV settings

General Structure of Bayesian Inference

- The quantities associated with each unit are, $Y_i(1), Y_i(0), S_i(1), S_i(0), X_i, Z_i$, four of which, $Y_i^{obs} = Y_i(Z_i), S_i^{obs} = S_i(Z_i), Z_i, X_i$ are observed and the rest two $Y_i^{mis} = Y_i(1 - Z_i)$, and $S_i^{mis} = S_i(1 - Z_i)$ are unobserved
- Assuming unit exchangeability and by appealing to de Finetti's theorem, we can write the probability density function of all random variables as

$$\Pr(Y(0), Y(1), S(0), S(1), Z, X) \quad = \quad \int \prod_i \Pr(Y_i(0), Y_i(1), S_i(0), S_i(1), Z_i, X_i, \theta) p(\theta) d\theta,$$

  where θ is the global parameter with prior distribution $p(\theta)$
- The posterior predictive distribution of the missing potential outcomes is:

$$\begin{aligned}
&\Pr(Y^{mis}, S^{mis} | Y^{obs}, S^{obs}, Z, X) \\
&\propto \int \prod_i \Pr(Z_i | Y_i(0), Y_i(1), G_i, X_i, \theta) \Pr(Y_i(0), Y_i(1) | G_i, X_i, \theta) \Pr(G_i | X_i, \theta) \Pr(X_i | \theta) p(\theta) d\theta \\
&\propto \int \prod_i \Pr(Y_i(0), Y_i(1) | G_i, X_i, \theta) \Pr(G_i | X_i, \theta) p(\theta) d\theta
\end{aligned}$$

- This suggests that, to conduct Bayesian inference for ignorable assignment mechanisms with intermediate variables, one need to specify two models:
  - ✓ $\Pr(Y_i(0), Y_i(1) \mid G_i, X_i, \theta)$: the distribution of potential outcomes $Y(0)$ and $Y(1)$ conditional on principal strata (and covariates), and
  - ✓ $\Pr(G_i \mid X_i, \theta)$: the distribution of principal strata conditional on the covariates

  and the prior distribution, $p(\theta)$
- The posterior distribution of θ is generally not tractable. Instead one can use a Gibbs sampler to simulate from the joint posterior distribution $\Pr(\theta, S^{mis} \mid Y^{obs}, S^{obs}, Z, X)$ by iteratively drawing between the two conditional distributions $\Pr(S^{mis} \mid Y^{obs}, S^{obs}, Z, X, \theta)$ and $\Pr(\theta \mid Y^{obs}, S^{obs}, S^{mis}, Z, X)$, the latter of which is proportional to the complete intermediate data likelihood
- These random draws provide posterior inference for the estimands

- From a Bayesian perspective, PCEs are always identified, because with proper prior distributions of the parameters posterior distributions of the causal estimands are always proper
- But some estimands are weakly identified, with substantial regions of flatness in their posterior distributions
- Bayesian inference for causal estimands can be sharpened by additional assumptions, such as monotonicity or exclusion restrictions

- Inference for PCEs can also be sharpen (reduce posterior variance) by secondary outcomes and covariates (Mealli and Pacini, 2013; Mattei et al., 2013) in weakly identified models. The inherent mixture structure of principal strata analysis underpins this improvement
- Extensions to
  - ✓ Multiple intermediate variables (Mattei and Mealli, 2007; Frumento et al., 2012)
  - ✓ Multivalued intermediate variables (Feller et al., 2016)
  - ✓ Continuous intermediate variables, using Bayesian semiparametric models (Schwartz et al., 2011)
- The field is still very fertile and I strongly to encourage students to engage in applied and theoretical work in causal inference