L1: Getting Started

Jean Morrison

University of Michigan

2021-01-05 (updated: 2022-01-04)

Motivation

- The idea that causality is interesting is easy to motivate.
- The question "Why?" motivates nearly all scientific endeavors.
- Causal inference formalizes what it means to answer this "Why?" question.
- The idea that we need this formalization is much newer than the idea that causal questions are interesting.

Example: Smoking and Lung Cancer

- In 1900 only 140 cases of lung cancer were known in published medical literature.
- By the 1920's incidence of lung cancer had increased dramatically.
- Smoking was a hypothesized cause of lung cancer as early as 1912...
- But so were:
    - Air pollution
    - Asphalt dust from new roads
    - Poison gas from WWI
    - Influenza pandemic of 1918
    - Increasing use of radiography
    - Increasing clinical awareness of lung cancer
    - Aging population

The Case Against Smoking

- Observational studies showed a strong association between smoking and lung cancer.
    - Earliest studies used a case/control design.
    - Followed by prospective studies matching healthy smokers and non-smokers by age, sex, and occupation and following them over time.
    - All studies observe a large and robust association between smoking and lung cancer (Doll and Hill estimate an OR of 40 in 1954).
- Supporting evidence comes from animal studies in the 1930's and 1940's: Exposing animals to tobacco products causes cancer.
- Even more evidence from chemical analysis in the 1940's and 1950's: Cigarette smoke contains known cancer causing chemicals.
    - Much of this work is done by tobacco companies themselves.
- By the 1950's tobacco companies also believe that tobacco use causes cancer. This information is kept secret.

Challenges to the Smoking-Lung Cancer Link

- Tobacco companies invested large amounts of money into research and advertising challenging the link between smoking and lung cancer.
- In the 1960's only one third of doctors believed smoking to be a major cause of lung cancer.
- RA Fisher was a famous challenger of the smoking $\rightarrow$ lung cancer hypothesis:
    - Fisher pointed to a genetic factor linked to both smoking and lung cancer, arguing this factor may be a common cause of both.
    - This concern about confounding is valid, but the effect size would need to be enormous.
    - The genetic hypothesis is also inconsistent with earlier low rates of lung cancer, animal studies, and reduced cancer rates in quitters.
    - It also disregards the possibility that smoking may be mediating the gene-cancer association.

Why did Fisher (and others) Get it Wrong?

- Some have suggested that Fisher had conflicts of interest -- he had done work as a tobacco industry consultant and was himself a smoker.
- Fisher's statement that the association between smoking and lung cancer could be explained by a common cause is correct.
- But that model is inconsistent with many other lines of evidence.

Lessons

- Causal inference cannot be achieved through only statistical procedures. It requires a model and interpretation provided by the practitioner.
- All causal analyses of observational data require un-provable assumptions.
- Many interesting and important questions cannot be answered in a randomized trial.
    - We need theory and language that allows us to test causal hypotheses in observational data.

Early Foundations of Causal Theory

- Neyman 1923: Notation and formalization of potential outcomes introduced.
- Fisher 1925: Physical randomization of units as the "reasoned basis" for inference.
- Wright 1921: Introduced graphical models and path analysis.

Languages of Causality

- Causality described using potential outcomes/counterfactuals:
  - Proposes the existence of unobserved outcomes for each unit under different possible exposures or treatments.
  - First formalized by Neyman in 1923 and further developed by Donald Rubin (1974 and onward) and others.
  - "Rubin causal model".
- Causality described using graphs:
  - Represents causal relationships between observed and unobserved variables as directed edges in a graph.
  - Causal interventions are represented as modifications of the graph.
  - Introduced by Wright (geneticist) in 1921, further developed by Judea Pearl (1988 and onward) and others.
- Causality described using structural equations:
  - Represents causal relationships as a series of equations describing conditional probability distributions.
  - Also introduced by Wright in 1921.
  - More work by Pearl, Haavelmo (econometrics), Duncan (social sciences), and many more.
- Under some conditions, all three languages are equivalent/mutually compatible.

Counterfactuals/Potential Outcomes

- What does it mean to say that A causes Y?
- A counterfactual value, $Y_i(A_i = a)$ is the value of Y the ith individual **would have had** if A had been **intervened on** and set to a.
- Many equivalent notations:
  - $Y(A = a)$, $Y(a)$: I will primarily use these notations
  - Y^a: This is the main notation in Hernán and Robins.
  - $Y|\ do(A = a)$: This notation explicitly emphasizes the action of setting A equal to a and is favored by Judea Pearl.
- A counterfactual fundamentally supposes a hypothetical intervention (treatment).

Example

For each person we observe:

- If they wear a helmet when biking ($A_i = 1$) or not ($A_i = 0$).
- If they sustain a head trauma in a given year ($Y_i = 1$) or not ($Y_i = 0$).
- Below is the full table of counterfatual outcomes:

	$Y(A = 0)$	$Y(A = 1)$
1	1	1
2	1	0
3	0	0
4	0	0
5	0	1
6	1	1
7	0	1
8	1	0

Individual vs Average Treatment Effects

- Sharp causal null: No effect of treatment for any individual, $Y_i(A_i = 1) = Y_i(A_i = 0)$ for all i.
- Average causal null: The average causal effect is zero, $E[Y(A = 0)] = E[Y(A = 1)]$.
- In our example, the sharp null is false, but the average null is true.

	$Y(A = 0)$	$Y(A = 1)$
1	1	1
2	1	0
3	0	0
4	0	0
5	0	1
6	1	1
7	0	1
8	1	0

Measures of Treatment Effects

- Average treatment effect (ATE):
  $E[Y(1)] - E[Y(0)] = E[Y_i(A1) - Y_i(0)]$
- Risk ratio (RR): $E[Y(0)]/E[Y(1)]$
- Odds ratio (OR; for binary outcomes): $\frac{E[Y(1)]/(1-E[Y(1)])}{E[Y(0)]/(1-E[Y(0)]}$
- Null hypotheses ATE $= 0$, RR $= 1$, and OR $= 1$ are equivalent.
- The ATE can be interpreted either as the difference in average outcome between groups or as the average difference.
- This is not true for RR and OR -- RR is not equal to the average risk ratio over the population.

Counterfactuals and Missing Data

- The counterfactual framework turns causal inference into a missing data problem.
- Even if we were able to observe A and Y for the entire population, uncertainty would remain in our estimate of the ATE because we cannot observe both $Y_i(1)$ and $Y_i(0)$ for the same individual.
- This has been called the "fundamental problem of causal inference."

Sample vs Population Treatment Effect

- We very rarely sample the entire population of interest.
- Sample average treatment effect (SATE): $\frac{1}{n}\sum_{i=1}^{n} Y_i(1) - \frac{1}{n}\sum_{i=1}^{n} Y_i(0)$
- Population average treatment effect (PATE): $E[Y(1)] - E[Y(0)]$, with expectation is taken over a super-population.
- Identifying the population is a scientific (rather than a statistical) task.

Non-Deterministic Counterfactuals

- So far we have seen two sources of uncertainty in the ATE:
  - Missing counterfactual outcomes
  - Sampling variation
- A third source is randomness in the counterfactual outcome.
- Rather than thinking of $Y_i(A_i = a)$ as a deterministic value, we can think of it as a random variable with individual specific (random) cdf $F_{Y_i(a)}(y)$.
- We will abbreviate $F_{Y_i(a)}(y)$ to $F_{a,i}$.

Non-Deterministic Counterfactuals

- If $Y_i(A_i = a)$ is a random variable, the ATE is a double expectation:

$$E[Y(A = a)] = E\left\{E[Y_i(A_i = a)|F_{a,i}]\right\}$$

  with the inside expectation taken over the distribution of $Y_i(A_i = a)$ and the outer expectation taken over the population.

- Let $F_a = E[F_{a,i}]$ be the average counterfactual cdf.

$$E[Y(A = a)] = E\left[\int y\, dF_{a,i}(y)\right] = \int y\, dE[F_{a,i}(y)] = \int y\, dF_a(y)$$

- So, the average counterfactual value is the expectation w.r.t the average counterfactual cdf.

- The distinction between deterministic and non-deterministic counterfactuals doesn't matter for average effects.

Causation vs Association

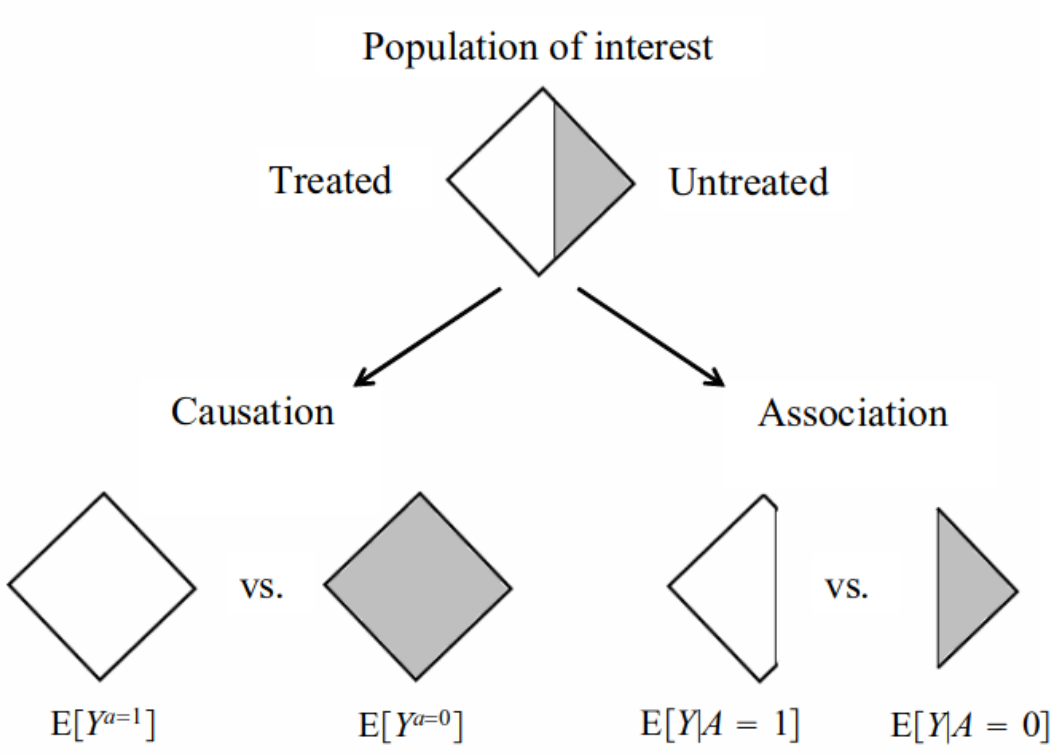

Fig 1.1 from HR

Bike Helmet Example Continued

	$Y(A=0)$	$Y(A=1)$
1	1	**0**
2	**0**	0
3	**1**	1
4	0	**0**
5	**0**	0
6	1	**0**
7	0	**0**
8	1	**1**
9	**1**	0
10	**0**	0

	$Y(A=0)$	$Y(A=1)$
11	**0**	0
12	**1**	0
13	0	**0**
14	**0**	0
15	**1**	1
16	**0**	0
17	**0**	0
18	1	**1**
19	**0**	0
20	0	**0**

Average causal effect: $E[Y(A=1)] - E[Y(A=0)] = 0.2 - 0.4 = -0.2$

Observed association: $E[Y|A=1] - E[Y|A=0] = 0.25 - 0.33 = -0.08$

Conditions for Identifying the Causal Effect

- Suppose we only observe A and Y. When is it possible to estimate the ATE?
- We must be able to estimate $E[Y(a)]$ for all values of a. Therefore, we must have

$$E[Y(a)] = E[Y|A = a]$$

- This is true under three assumptions:
  - Consistency
  - Stable Unit Treatment Value Assumption (SUTVA)
  - Exchangeability (exogeneity)

Consistency

- The observed value of Y_i is the same as the counterfactual value of Y_i under the treatment a_i, where a_i is the treatment individual i actually recieved.

$$Y_i = Y_i(A_i = a_i)$$

.

Bike example:

- A person's chances of head injury are the same if they wear a helmet of their own accord or if they are required to wear a helmet.
- Whether or not consistency holds may depend on the nature of the hypothesized intervention.

Stable Unit Treatment Value Assumption

1. There are no different versions of treatment available to an individual and the treatment level a is unambiguous for all values of a.
2. No interference: The counterfactual outcome for unit i, $Y_i(A_i = a)$ is independent of the treatment received by other units in the study.
    - Formally, let $\mathbf{A} = (A_1, \ldots, A_n)$ be the vector of treatment assignments for all units with $\mathbf{a}$ being a single realization of $\mathbf{A}$. Let a_i be the ith element of $\mathbf{a}$ and $\mathcal{A}^n$ be the sample space of $\mathbf{A}$. Let $Y_i(\mathbf{A} = \mathbf{a})$ be the counterfactual value of Y_i under the treatment vector $\mathbf{a}$. No interference is the condition that

$$Y_i(\mathbf{A} = \mathbf{a}) = Y_i(A_i = a_i) \quad \forall \ \mathbf{a} \in \mathcal{A}^n$$

Pair discussion:

- What do these assumptions look like in the bike helmet example?
- Can you think of a time "No interference" may not hold?

Exchangeability

- Exchangeability: $Y(a) \perp\!\!\!\perp A$.
- Question: How is $Y(a) \perp\!\!\!\perp A$ different from $Y \perp\!\!\!\perp A$?
- Mean exchangeability: $E[Y(a)|A = a'] = E[Y(a)|A = a'']$ for all pairs $a', a'' \in \mathcal{A}$.
- Mean exchangeability is sufficient to prove $E[Y(a)] = E[Y|A = a]$.
- For dichotomous Y, mean exchangeability and exchangeability are equivalent.
- What about for non-dichotomous Y?
- Full exchangeability: Let $Y^{\mathcal{A}} = \{Y(a), Y(a'), \ldots\}$ be the set of all counterfactual outcomes ($Y^{\mathcal{A}} = \{Y(1), Y(0)\}$ for a dichotomous treatment). Full exchangeability states that

$$Y^{\mathcal{A}} \perp\!\!\!\perp A$$

Exchangeability in the Bike Helmet Example:

	$Y(A=0)$	$Y(A=1)$
1	1	**0**
2	**0**	0
3	**1**	1
4	0	**0**
5	**0**	0
6	1	**0**
7	0	**0**
8	1	**1**
9	**1**	0
10	**0**	0

	$Y(A=0)$	$Y(A=1)$
11	**0**	0
12	**1**	0
13	0	**0**
14	**0**	0
15	**1**	1
16	**0**	0
17	**0**	0
18	1	**1**
19	**0**	0
20	0	**0**

$E[Y(A=0)|A=0] = 0.33, E[Y(A=0)|A=1] = 0.5$

$E[Y(A=1)|A=0] = 0.17, E[Y(A=1)|A=1] = 0.25$

Identification Theorem

Theorem: If consistency, SUTVA, and exchangeability hold, then

$$Y(a) \stackrel{d}{=} Y|A = a,$$

where $\stackrel{d}{=}$ means equal in distribution.

$$\begin{aligned} P[Y(a) \leq y] &= P[Y(a) \leq y|A = a] \quad (\text{exchangeability}) \\ &= P[Y \leq y|A = a] \quad (\text{consistency}) \end{aligned}$$

Where did SUTVA come in?

- If the levels of A are not clearly defined, the causal question is ill-defined (i.e. $Y(a)$ is poorly defined).
- In the presence of interference, there is no single value of $Y_i(a_i)$. Instead, we must discuss $Y_i(\mathbf{a})$, the potential outcome given the entire vector of treatment assignments.

Simple Randomized Experiments

- Units are assigned a treatment value using a randomization procedure that is independent of all unit characteristics (e.g. flip a coin).
- For now, we assume full compliance (everyone assigned treatment a receives treatment a).
- Exchangeability holds by design.
- So, we can estimate the ATE as long as consistency and SUTVA hold.
- One estimator is just the difference in average outcomes (more on estimators in HW1):

$$\frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_2} \sum_{i:A_i=0} Y_i = \bar{Y}_1 - \bar{Y}_0$$

Conditionally Randomized Experiments

- Units are assigned a treatment with probability depending on a set of features, X.
- If X and $Y(a)$ are not independent, then exchangeability does not (generally) hold.
- However, we can still identify the ATE if we have access to the randomization features X.

Example

- We are doing an experiment of a new treatment for a disease. Some patients will receive the new treatment ($A = 1$), while the rest will receive standard of care ($A = 0$).
- We decide on a randomization scheme in which sicker patients are more likely to receive the new treatment. We set $P(A = 1|L = 0) = 0.5$ and $P(A = 1|L = 1) = 0.75$.
- Let $L = 0$ indicate that a patient is less sick and $L = 1$ indicate that they are more sick.
- We observe if each patient dies before a set time point ($Y = 1$ for death, $Y = 0$ for survival).

Conditional Exchangeability

- Notice that our trial looks like two fully randomized trials combined.
    - In one trial, the target population is patients with $L = 0$ and the treatment probability is 0.5.
    - In the other, the target population is patients with $L = 1$ and the treatment probability is 0.75.
- Conditional exchangeability captures the idea that data are randomized *within* levels of L.
- Conditional exchangeability holds with respect to a set of variables L, if

$$Y(a) \perp\!\!\!\perp A \mid L$$

Stratum Specific Causal Effects and Standardization

- Within values of L, $P[Y(a)|L = l]$ are identified because exchangeability holds within levels of L.
- If we want to estimate the population level marginal counterfactual value of Y under treatment $A = a$, we can simply weight our estimates by the population frequency of L:

$$E[Y(a)] = \sum_{l} E[Y(a)|L = l]P[L = l]$$

- This is called the standardized mean (standardized by whatevery population frequencies of L you choose).

Positivity

- Of course, the standardization trick doesn't work if, within one level of L, patients never (or always) receive treatment.
- The positivity condition states that at all individuals have some chance of receiving any treatment:

$$P[A = a|L = l] > 0 \ \ \forall a, l$$

Identification Theorem (Conditional)

Theorem: If consistency, SUTVA, conditional exchangeability, *and positivity* hold, then

$$Y(a)|X = x \stackrel{d}{=} Y|X = x, A = a.$$

We can use exactly the same proof conditioning on X in each step.

Inverse Probability Weighting

- Rather than using the standardization method, we can think of our trial as a weighted sampling from a larger, fully randomized trial with $2 * N$ participants.
- This **pseudo-population** contains two member for every member in our trial, one receiving each treatment.
- In the pseudo-population, $Y(a) \perp\!\!\!\perp A$ so the conditional mean, $E[Y|A = a]$ estimates the counterfactual mean $E[Y(a)]$.

Inverse Probability Weighting

- We can imagine that each individual in our trial was sampled from the larger population with probability conditional on L and A.
    - In our study we selected half of participants with $A_i = 0$ and $L_i = 0$
    - Half of participants with $A_i = 1$ and $L = 0$
    - One quarter of participants with $A_i = 0$ and $L = 1$
    - Three quarters of participants with $A_i = 1$ and $L = 1$
- We can recover the estimate from the pseudo-population by weighting each participant by the number of units in the larger study that they represent:
    - $A_i = 0, L_i = 0$: $1/0.5 = 2$
    - $A_i = 1, L_i = 0$: $1/0.5 = 2$
    - $A_i = 0, L_i = 1$: $1/0.25 = 4$
    - $A_i = 1, L_i = 0$: $1/0.75 = 1.33$

Inverse Probability Weighting

- Formally, let $f_{A|L}(a|l)$ be the conditional pdf of A given L.
- We assume that $f_{A|L}(a|l) > 0$ for all a and l s.t $P[L = l] > 0$.
- The IP weighting for individual i is $W_i^A = 1/f_{A|L}(a_i|l_i)$.
- The IP weighted mean for treatment level a is $E\left[\frac{I(A=a)Y}{f(A|L)}\right]$

Equivalence of IP Weighting and Standardization

- We will assume that A and L are discrete and $f(a|l) = P[A = a|L = l] > 0$ for all l with $P[L = l] > 0$.
- Use the iterated expectation formula:

$$E\left[\frac{I(A=a)Y}{f(A|L)}\right] = E_L\left\{E_A\left\{E_Y\left[\frac{I(A=a)Y}{f(A|L)}|A,L\right]\right\}\right\}$$

$$= \sum_l \frac{E[Y|A=a,L=l]}{f(a|l)} f(a|l)P[L=l]$$

$$= \sum_l E[Y|A=a,L=l]P[L=l]$$

- This proof extends to continuous L but not to continuous A (see HW 1).
- If conditional exchangeability holds, then both the IP weighted mean and the standardized mean estimate $Y(a)$.

Causal Effects from in an Observational Data

- None of our four identification conditions consistency, SUTVA, conditional exchangeability, and positivity require that our data is from a randomized trial.
- However, these are much stronger assumptions in observational data.
- We must also consider whether the causal question is well-defined:
    - Is the counterfactual outcome well defined? (Ex. what does it mean to say that obesity is a cause of heart disease)?
    - Are all levels of the treatment observed in the data?
- The target trial: Hernán and Robins suggest that researchers using observational data should imagine a hypothetical trial they are trying to emulate.