CAUSES AND COUNTERFACTUALS IN THE EMPIRICAL SCIENCES

Judea Pearl
University of California
Los Angeles
(www.cs.ucla.edu/~judea)

OUTLINE

• Inference: Statistical vs. Causal distinctions and mental barriers
• Formal semantics for counterfactuals: definition, axioms, graphical representations
• Inference to three types of claims:
  1. Effect of potential interventions
  2. Attribution (Causes of Effects)
  3. Direct and indirect effects

TRADITIONAL STATISTICAL INFERENCE PARADIGM

Data \xrightarrow{P} Joint Distribution \xrightarrow{Q(P)} (Aspects of P)
Inference

e.g., Infer whether customers who bought product A would also buy product B.
\[ Q = P(B \mid A) \]

FROM STATISTICAL TO CAUSAL ANALYSIS: 1. THE DIFFERENCES

Data \xrightarrow{P} Joint Distribution \xrightarrow{Q(P)} (Aspects of P)
Inference

What happens when P changes?
e.g., Infer whether customers who bought product A would still buy A if we were to double the price.

FROM STATISTICAL TO CAUSAL ANALYSIS: 1. THE DIFFERENCES (CONT)

Data \xrightarrow{P} Joint Distribution \xrightarrow{P'} Joint Distribution \xrightarrow{Q(P')} (Aspects of P')
Inference

What remains invariant when P changes say, to satisfy 
\[ P'(price=2) = 1 \]

Note: \[ P'(v) \neq P(v \mid price = 2) \]
P does not tell us how it ought to change
e.g. Curing symptoms vs. curing diseases
e.g. Analogy: mechanical deformation

FROM STATISTICAL TO CAUSAL ANALYSIS: 1. THE DIFFERENCES

What remains invariant when P changes say, to satisfy 
\[ P'(price=2) = 1 \]

Note: \[ P'(v) \neq P(v \mid price = 2) \]
P does not tell us how it ought to change
e.g. Curing symptoms vs. curing diseases
e.g. Analogy: mechanical deformation
FROM STATISTICAL TO CAUSAL ANALYSIS: 2. MENTAL BARRIERS

1. Causal and statistical concepts do not mix.

Causal
Spurious correlation
Randomization / Intervention
Confounding / Effect
Instrumental variable
Strong Exogeneity
Explanatory variables

Statistical
Regression
Association / Independence
"Controlling for" / Conditioning
Odd and risk ratios
Collapsibility / Granger causality
Propensity score

2. No causes in – no causes out (Cartwright, 1989)
statistical assumptions + data
causal assumptions
causal conclusions


4. Non-standard mathematics:
   a) Structural equation models (Wright, 1920; Simon, 1960)
   b) Counterfactuals (Neyman-Rubin \( Y_x \), Lewis \( xY \))

WHY CAUSALITY NEEDS SPECIAL MATHEMATICS

Scientific Equations (e.g., Hooke’s Law) are non-algebraic

Correct notation:
\[
\begin{align*}
Y &= 2X \\
X &= 1 \\
Y &= 2
\end{align*}
\]

Had \( X \) been 3, \( Y \) would be 6.
If we raise \( X \) to 3, \( Y \) would be 6.
Must “wipe out” \( X = 1 \).

THE STRUCTURAL MODEL PARADIGM

Data Distribution

Data Generating Model

\( Q(M) \) (Aspects of \( M \))

Inference

\( M \) – Invariant strategy (mechanism, recipe, law, protocol) by which Nature assigns values to variables in the analysis.

FAMILIAR CAUSAL MODEL

ORACLE FOR MANIPILATION

INPUT

\( X \)

\( Y \)

\( Z \)

OUTPUT
STRUCTURAL MODELS

Definition: A structural causal model is a 4-tuple \( \langle V, U, F, P(u) \rangle \), where
- \( V = \{ V_1, ..., V_n \} \) are observable variables
- \( U = \{ U_1, ..., U_m \} \) are background variables
- \( F = \{ f_1, ..., f_n \} \) are functions determining \( V \), \( v_i = f_i(v, u) \)
- \( P(u) \) is a distribution over \( U \)

\( F \) and \( P \) induce a distribution \( P(v) \) over observable variables.

STRUCTURAL MODELS AND CAUSAL DIAGRAMS

The arguments of the functions \( v_i = f_i(v, u) \) define a graph \( v_i = f_i(pai, ui) \)

\[ PA_i \subseteq V \setminus Vi \]

Example: Price – Quantity equations in economics

\[ q = b_1 p + d_1 i + u_1 \]
\[ p = b_2 q + d_2 w + u_2 \]

STRUCTURAL MODELS AND INTERVENTION

Let \( X \) be a set of variables in \( V \).
The action \( \text{do}(x) \) sets \( X \) to constants \( x \) regardless of the factors which previously determined \( X \).
\( \text{do}(x) \) replaces all functions \( f_i \) determining \( X \) with the constant functions \( X = x \), to create a mutilated model \( M_x \).

\[ q = b_1 p + d_1 i + u_1 \]
\[ p = b_2 q + d_2 w + u_2 \]

CAUSAL MODELS AND COUNTERFACTUALS

Definition: The sentence: “\( Y \) would be \( y \) (in situation \( u \)), had \( X \) been \( x \),” denoted \( Y_x(u) = y \), means:
The solution for \( Y \) in a mutilated model \( M_x \), (i.e., the equations for \( X \) replaced by \( X = x \)) with input \( U = u \), is equal to \( y \).

The Fundamental Equation of Counterfactuals:

\[ Y_x(u) = Y_{M_x}(u) \]

CAUSAL MODELS AND COUNTERFACTUALS

Definition: The sentence: “\( Y \) would be \( y \) (in situation \( u \)), had \( X \) been \( x \),” denoted \( Y_x(u) = y \), means:
The solution for \( Y \) in a mutilated model \( M_x \), (i.e., the equations for \( X \) replaced by \( X = x \)) with input \( U = u \), is equal to \( y \).

- Joint probabilities of counterfactuals:
  \[ P(Y_x = y, Z_w = z) = \sum P(u) \]

The Fundamental Equation of Counterfactuals:

In particular:

\[ P(Y_x = y | \text{do}(x)) \overset{\Delta}{=} P(Y_x = y) = \sum_{u: X_x(u) = y} P(u) \]
\[ PN(Y_x = y | x, y) = \sum_{u: X_x(u) = y} P(u | x, y) \]
REGRESSION VS. STRUCTURAL EQUATIONS
(THE CONFUSION OF THE CENTURY)

Regression (claimless, nonfalsifiable):
\[ Y = ax + \epsilon_y \]

Structural (empirical, falsifiable):
\[ Y = bx + uy \]

Assumptions:
\[ E(Y|do(x)) = E(Y|do(x), do(z)) = bx \]

The mothers of all questions:
Q. When would \( b \) equal \( a \)?
A. When all back-door paths are blocked, or

Q. When is \( b \) estimable by regression methods?
A. Graphical criteria available

AXIOMS OF CAUSAL COUNTERFACTUALS

1. Definiteness
   \[ \exists x \in X \text{ s.t. } X_y(u) = x \]
2. Uniqueness
   \[ (X_y(u) = x) \land (X_y(u) = x') \Rightarrow x = x' \]
3. Effectiveness
   \[ X_{x_t}(u) = x \]
4. Composition
   \[ W_x(u) = w \Rightarrow Y_{x_t}(u) = Y_x(u) \]
5. Reversibility
   \[ (Y_{x_t}(u) = y) \land (W_x(u) = w) \Rightarrow Y_x(u) = y \]

INFERRING THE EFFECT OF INTERVENTIONS

The problem:
To predict the impact of a proposed intervention using data obtained prior to the intervention.

The solution (conditional):
Causal Assumptions + Data \( \rightarrow \) Policy Claims

1. Mathematical tools for communicating causal assumptions formally and transparently.
2. Deciding (mathematically) whether the assumptions communicated are sufficient for obtaining consistent estimates of the prediction required.
3. Suggesting (if (2) is negative) a set of measurements and experiments that, if performed, would render a consistent estimate feasible.

NON-PARAMETRIC STRUCTURAL MODELS

Given \( P(x, y, z) \), should we ban smoking?

Linear Analysis
\[ x = u_1, \]
\[ z = \alpha x + u_2, \]
\[ y = \beta z + \gamma u_1 + u_3. \]

Nonparametric Analysis
\[ x = f_1(u_1), \]
\[ z = f_2(x, u_2), \]
\[ y = f_3(z, u_1, u_3). \]

Find: \( x \cdot \beta \)
Find: \( P(y|do(x)) \)

EFFECT OF INTERVENTION
AN EXAMPLE

Given \( P(x, y, z) \), should we ban smoking?

Linear Analysis
\[ x = u_1, \]
\[ z = \alpha x + u_2, \]
\[ y = \beta z + \gamma u_1 + u_3. \]

Nonparametric Analysis
\[ x = \text{const.}, \]
\[ z = f_2(x, u_2), \]
\[ y = f_3(z, u_1, u_3). \]

Find: \( x \cdot \beta \)
Find: \( P(y|do(x)) \triangleq P(Y=y) \) in new model

EFFECT OF INTERVENTION
THE GENERAL CASE

Find the effect of \( X \) on \( Y \), \( P(y|do(x)) \), given the causal assumptions shown in \( G \), where \( Z_1, \ldots, Z_k \) are auxiliary variables.

Can \( P(y|do(x)) \) be estimated if only a subset, \( Z \), of the auxiliary variables can be measured?
**Eliminating Confounding Bias: A Graphical Criterion**

\( P(y \mid \text{do}(x)) \) is estimable if there is a set \( Z \) of variables such that \( Z \) \( d \)-separates \( X \) from \( Y \) in \( G_x \).

Moreover, \( P(y \mid \text{do}(x)) = \sum_z P(y \mid x, z) P(z) \) ("adjusting" for \( Z \)).

**Effect of Intervention Beyond Adjustment**

Theorem (Tian-Pearl 2002)

We can identify \( P(y \mid \text{do}(x)) \) if there is no child \( Z \) of \( X \) connected to \( X \) by a confounding path.

**Inference Across Designs**

Problem:
Predict \( P(y \mid \text{do}(x)) \) from a study in which only \( Z \) can be controlled.

Solution:
Determine if \( P(y \mid \text{do}(x)) \) can be reduced to a mathematical expression involving only \( \text{do}(z) \).

**Effect of Intervention: Complete Identification**

- Complete calculus for reducing \( P(y \mid \text{do}(x), z) \) to expressions void of \( \text{do} \)-operators.
- Complete graphical criterion for identifying causal effects (Shpitser and Pearl, 2006).
- Complete graphical criterion for empirical testability of counterfactuals (Shpitser and Pearl, 2007).

**The Causal Renaissance: Vocabulary in Economics**

From Hoover (2004)
"Lost Causes"

1. Complete formal semantics of counterfactuals
2. Transparent language for expressing assumptions
3. Complete solution to causal-effect identification
4. Legal responsibility (bounds)
5. Imperfect experiments (universal bounds for IV)
6. Integration of data from diverse sources
7. Direct and Indirect effects,
8. Complete criterion for counterfactual testability
DETERMINING THE CAUSES OF EFFECTS
(The Attribution Problem)

• Your Honor! My client (Mr. A) died BECAUSE he used that drug.

• Court to decide if it is MORE PROBABLE THAN NOT that A would be alive BUT FOR the drug!

P(A | A is dead, took the drug) ≥ 0.50

THE PROBLEM

Semantical Problem:
1. What is the meaning of $PN(x,y)$:
   “Probability that event $y$ would not have occurred if it were not for event $x$, given that $x$ and $y$ did in fact occur.”

  Answer:
  \[ PN(x,y) = P(Y' = y' | x,y) \]
  Computable from $M$

Analytical Problem:
2. Under what condition can $PN(x,y)$ be learned from statistical data, i.e., observational, experimental and combined.

TYPICAL THEOREMS
(Tian and Pearl, 2000)

• Bounds given combined nonexperimental and experimental data
  \[ 0 \leq PN \leq \min \left\{ \frac{1}{P(x,y)} \left( \frac{P(y' | x') - P(y' | x)}{P(x,y)} \right) \right\} \]

• Identifiability under monotonicity (Combined data)
  \[ PN = \frac{P(y|x) - P(y|x')} + \frac{P(y|x') - P(y|x)}{P(x,y)} \]
  corrected Excess-Risk-Ratio
**CAN FREQUENCY DATA DECIDE LEGAL RESPONSIBILITY?**

<table>
<thead>
<tr>
<th></th>
<th>Experimental</th>
<th>Nonexperimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths (y)</td>
<td>16 14</td>
<td>2 28</td>
</tr>
<tr>
<td>Survivals (y')</td>
<td>984 986</td>
<td>908 972</td>
</tr>
</tbody>
</table>

- Nonexperimental data: drug usage predicts longer life
- Experimental data: drug has negligible effect on survival
- Plaintiff: Mr. A is special.
  1. He actually died
  2. He used the drug by choice
- Court to decide (given both data):
  Is it more probable than not that Mr. A would be alive but for the drug?

\[
PN \Delta P(Y' = y'| x, y) > 0.50
\]

**SOLUTION TO THE ATTRIBUTION PROBLEM**

- WITH PROBABILITY ONE \( 1 \leq P(y' | x, y) \leq 1 \)
- Combined data tell more that each study alone

**EFFECT DECOMPOSITION (direct vs. indirect effects)**

1. Why decompose effects?
2. What is the semantics of direct and indirect effects?
3. What are the policy implications of direct and indirect effects?
4. When can direct and indirect effect be estimated consistently from experimental and nonexperimental data?

**WHY DECOMPOSE EFFECTS?**

1. To understand how Nature works
2. To comply with legal requirements
3. To predict the effects of new type of interventions:
   Signal routing, rather than variable fixing

**LEGAL IMPLICATIONS OF DIRECT EFFECT**

Can data prove an employer guilty of hiring discrimination?

- (Gender) \( X \) \( \rightarrow \) \( Z \) (Qualifications)
- (Hiring) \( Y \) \( \rightarrow \) \( Z \)

What is the direct effect of \( X \) on \( Y \)?

\[
E(Y | do(x_1), do(z)) - E(Y | do(x_0), do(z))
\]

(averaged over \( Z \)) Adjust for \( Z \)? No! No!

**NATURAL SEMANTICS OF AVERAGE DIRECT EFFECTS**

Robins and Greenland (1992) – “Pure”

\[
X \rightarrow Z \quad z = f(x, u) \\
Y \leftarrow Z \quad y = g(x, z, u)
\]

Average Direct Effect of \( X \) on \( Y \): \( DE(x_0, x_1; Y) \)

The expected change in \( Y \), when we change \( X \) from \( x_0 \) to \( x_1 \) and, for each \( u \), we keep \( Z \) constant at whatever value it attained before the change.

\[
E[Y_{X_1Z_0} - Y_{X_0}]
\]

In linear models, \( DE = \) Controlled Direct Effect
SEMANTICS AND IDENTIFICATION OF NESTED COUNTERFACTUALS

Consider the quantity $Q = E_u[1_{X,Z}(u)u]$.

Given $(M, P(u))$, $Q$ is well defined.

Given $u$, $Z_{x^*}(u)$ is the solution for $Z$ in $M_{x^*}$; call it $z$.

$Y_{x_{z^*}(u)}(u)$ is the solution for $Y$ in $M_{xz^*}$.

Can $Q$ be estimated with experimental and nonexperimental data?

Experimental: nest-free expression
Nonexperimental: subscript-free expression

POLICY IMPLICATIONS OF INDIRECT EFFECTS

What is the indirect effect of $X$ on $Y$?

The effect of Gender on Hiring if sex discrimination is eliminated.

Blocking a link – a new type of intervention

RELATIONS BETWEEN TOTAL, DIRECT, AND INDIRECT EFFECTS

Theorem 5: The total, direct and indirect effects obey the following equality

$$\text{IE}(x_0, x_1, Y) = \text{TE}(x_0, x_1, Y) - \text{DE}(x_0, x_1, Y)$$

In words, the total effect on $Y$ associated with the transition from $x$ to $x^*$ is equal to the difference between the direct effect associated with this transition and the indirect effect associated with the reverse transition, from $x$ to $x^*$.

EXPERIMENTAL IDENTIFICATION OF AVERAGE DIRECT EFFECTS

Theorem: If there exists a set $W$ such that

$$Y_{xW} \perp Z_{x^*} | W$$

for all $z$ and $x$.

Then the average direct effect

$$\text{DE}(x, x^*; Y) = E[Y_{x}, Z_{x^*}] - E[Y_{x^*}]$$

is identifiable from experimental data and is given by

$$\text{DE}(x, x^*; Y) = \sum_{w \in W} [E(Y_{xW} | w) - E(Y_{x^*W} | w)]P(Z_{x^*} = z | w)P(w)$$

GRAPHICAL CONDITION FOR EXPERIMENTAL IDENTIFICATION OF DIRECT EFFECTS

Theorem: If there exists a set $W$ such that

$$(Y \perp Z | W)_{G_{\text{AG}}}$$

and $W \subseteq ND(X \cup Z)$

then,

$$\text{DE}(x, x^*; Y) = \sum_{w \in W} [E(Y_{xW} | w) - E(Y_{x^*W} | w)]P(Z_{x^*} = z | w)P(w)$$

Example:
GENERAL PATH-SPECIFIC EFFECTS (Def.)

Form a new model, $M'_g$, specific to active subgraph $g$

\[ f'_i(pa_i, u; g) = f_i(pa_i(g), pa_i(g), u) \]

Definition: $g$-specific effect

\[ E_g(x, x^*; Y)_M = TE(x, x^*; Y)_{M'_g} \]

Nonidentifiable even in Markovian models

SUMMARY OF RESULTS

1. Formal semantics of path-specific effects, based on signal blocking, instead of value fixing.
2. Path-analytic techniques extended to nonlinear and nonparametric models.
3. Meaningful (graphical) conditions for estimating direct and indirect effects from experimental and nonexperimental data.

CONCLUSIONS

Structural-model semantics, enriched with logic and graphs, provides:

- Complete formal basis for causal and counterfactual reasoning
- Unifies the graphical, potential-outcome and structural equation approaches
- Provides friendly and formal solutions to century-old problems and confusions.