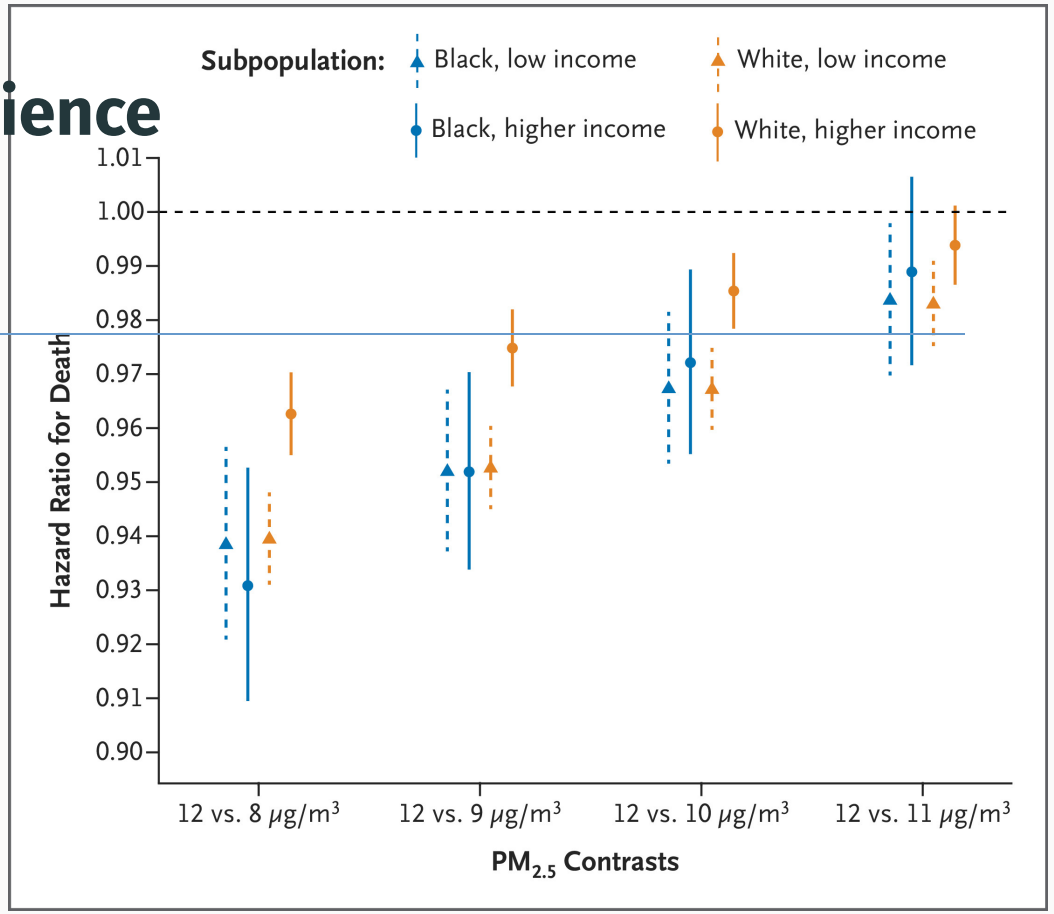
Statistical Theory for Data Science

STA2212H S LEC9101

Week 8

March 3 2026



SPECIAL ARTICLE

Air Pollution and Mortality at the Intersection of Race and Social Class

Kevin P. Josey, Ph.D., Scott W. Delaney, Sc.D., J.D., Xiao Wu, Ph.D., Rachel C. Nethery, Ph.D., Priyanka DeSouza, Ph.D., Danielle Braun, Ph.D., and Francesca Dominici, Ph.D.

ABSTRACT

BACKGROUND

Black Americans are exposed to higher annual levels of air pollution containing fine particulate matter (particles with an aerodynamic diameter of ≤2.5 μm [$PM_{2.5}$]) than White Americans and may be more susceptible to its health effects. Low-income Americans may also be more susceptible to $PM_{2.5}$ pollution than high-income Americans. Because information is lacking on exposure–response curves for $PM_{2.5}$ exposure and mortality among marginalized subpopulations categorized according to both race and socioeconomic position, the Environmental Protection Agency lacks important evidence to inform its regulatory rulemaking for $PM_{2.5}$ standards.

METHODS

We analyzed 623 million person-years of Medicare data from 73 million persons 65 years of age or older from 2000 through 2016 to estimate associations between annual $PM_{2.5}$ exposure and mortality in subpopulations defined simultaneously by racial identity (Black vs. White) and income level (Medicaid eligible vs. ineligible).

RESULTS

Lower $PM_{2.5}$ exposure was associated with lower mortality in the full population, but marginalized subpopulations appeared to benefit more as $PM_{2.5}$ levels decreased. For example, the hazard ratio associated with decreasing $PM_{2.5}$ from 12 μg per cubic meter to 8 μg per cubic meter for the White higher-income subpopulation was 0.963 (95% confidence interval [CI], 0.955 to 0.970), whereas equivalent hazard

From the Departments of Biostatistics (K.P.J., R.C.N., D.B., F.D.) and Environmental Health (S.W.D.), Harvard T.H. Chan School of Public Health, Boston; the Department of Biostatistics, Mailman School of Public Health, Columbia University, New York (X.W.); and the Department of Urban and Regional Planning, University of Colorado Denver, Denver (P.D.). Dr. Dominici can be reached at fdominic@hsph.harvard.edu or at the Department of Biostatistics, Harvard T.H. Chan School of Public Health, 655 Huntington Ave., Bldg. 2, 4th Flr., Boston, MA 02115.

Drs. Josey and Delaney and Drs. Braun and Dominici contributed equally to this article.

This article was published on March 24, 2023, at NEJM.org.

N Engl J Med 2023;388:1396-404.
DOI: 10.1056/NEJMsa2300523

Today

1. Midterm 2 March 10
2. Recap: missing data assumptions, IPW estimators, meta-analysis and publication bias, funnel plots centered at ??
3. Introduction to causal inference
4. Project papers
5. March 17 class

Upcoming

- Toronto Data Workshop, Wednesday 4 March, 10.00 (EST) on Zoom
  Lisa Oswald, Goethe University: "Disentangling participation in online political discussions with a collective field experiment"
- Statistics Colloquium, Thursday 5 March, 11am Hydro 9195/9
  Nicholas Polson, U Chicago: "Generative Bayes, *E*-values, and conformal prediction"

Recap: Missing data

- Missing responses: completely at random, at random, not at random

$$\text{pr}(R = 1 \mid Y, X) \begin{cases} = \text{pr}(R = 1), & \text{MCAR} \\ = \text{pr}(R = 1 \mid X) & \text{MAR} \\ = \text{pr}(R = 1 \mid Y, X) & \text{?NIN?} \end{cases}$$

- under MCAR, MAR, complete case likelihood function can be used for inference about $f(y \mid x)$.
- use observed Fisher information to estimate variability
- use of missing data models for meta-analysis
- funnel plots centered at estimated average effect (over studies)
- multiple imputation (often based on MV Normal) for missing covariates
- another approach for missing covariates includes an indicator covariate to indicate missing-ness

Cox & Donnelly

... Recap: Missing data

X in AoS

- AoS example: $\psi = \mathrm{pr}(Y = 1) = \sum_z \mathrm{pr}(Y = 1 \mid Z = z)\mathrm{pr}(Z = z)$
- estimated by

$$\widehat{\psi} = \sum_{i=1}^{n} \frac{Y_i R_i}{\mathrm{pr}(R_i = 1)}$$

inverse prob. weighted
IPW
Horvitz–Thompson
estimator

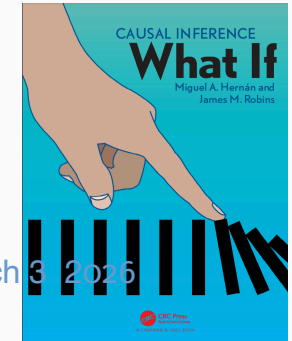
- R_i is an indicator of (non)-missingness

- EM algorithm introduces missing data (latent variables) to make likelihood calculations easier

Causality

AoS 16,17; SM 9.1; Cox & Donnelly 9.2

- randomization; confounding; observational studies; experiments; "correlation is not causation", Simpson's 'paradox' — informal
- counterfactuals; average treatment effect; conditional average treatment effect; ... — potential outcomes
- graphical models; directed acyclic graphs; causal graphs; Markov assumptions...
- structural equation models
- The Book



Very helpful Isabelle

Journal of Data Science 24 (1), 53–85
January 2026
DOI: 10.6339/25-JDS1211
Data Science Reviews

Causal Inference: A Tale of Three Frameworks

Linbo Wang[1,*], Thomas S. Richardson[2], and James M. Robins[3]

[1]*Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada*
[2]*Department of Statistics, University of Washington, Seattle, WA, U.S.A.*
[3]*Harvard T.H. Chan School of Public Health, Boston, MA, U.S.A.*

Abstract

Causal inference is a central goal across many scientific disciplines. Over the past several decades, three major frameworks have emerged to formalize causal questions and guide their analysis: the potential outcomes framework, structural equation models, and directed acyclic graphs. Although these frameworks differ in language, assumptions, and philosophical orientation, they often lead to compatible or complementary insights. This paper provides a comparative introduction to the three frameworks, clarifying their connections, highlighting their distinct strengths and limitations, and illustrating how they can be used together in practice. The discussion is aimed at researchers and graduate students with some background in statistics or causal inference who are seeking a conceptual foundation for applying causal methods across a range of substantive domains.

Keywords *directed acyclic graphs; identification; potential outcomes; structural equation models; SWIGs*

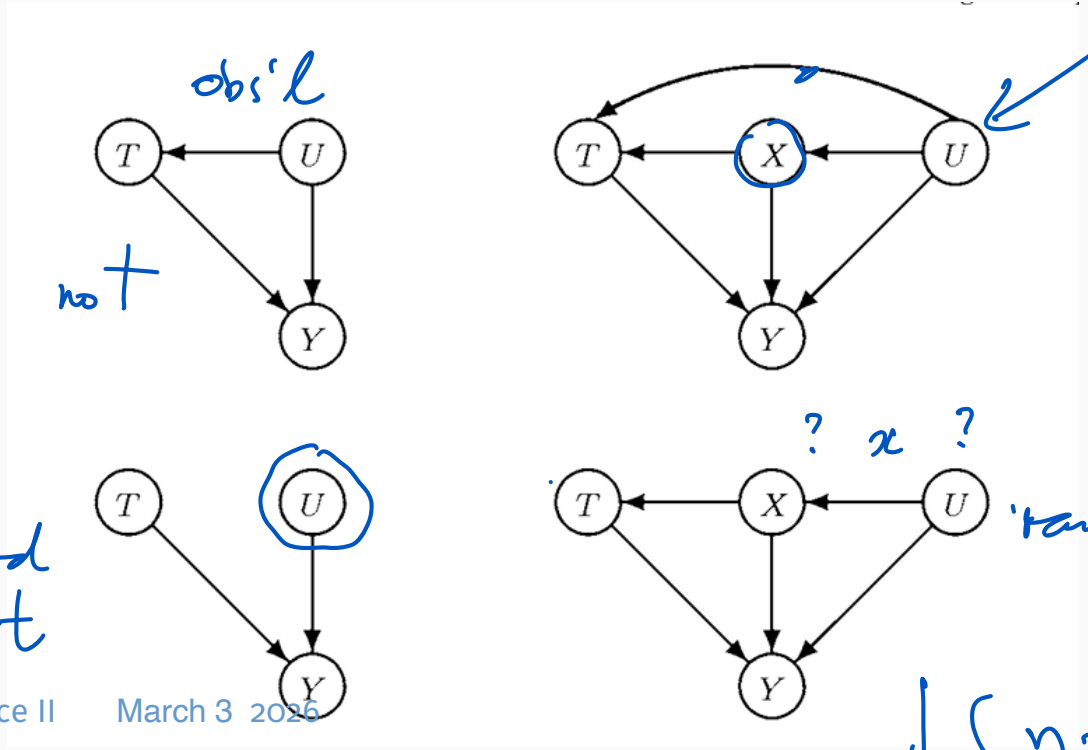
1 Introduction

Randomization

- designed experiment: treatments are assigned to units, a response is measured
- treatments are assigned at random to ensure balance between groups
- units may be blocked, with randomization applied within each block

Randomization

- designed experiment: treatments are assigned to units, a response is measured
- treatments are assigned at random to ensure balance between groups
- units may be blocked, with randomization applied within each block



Randomization and covariate

SM §9.1.1

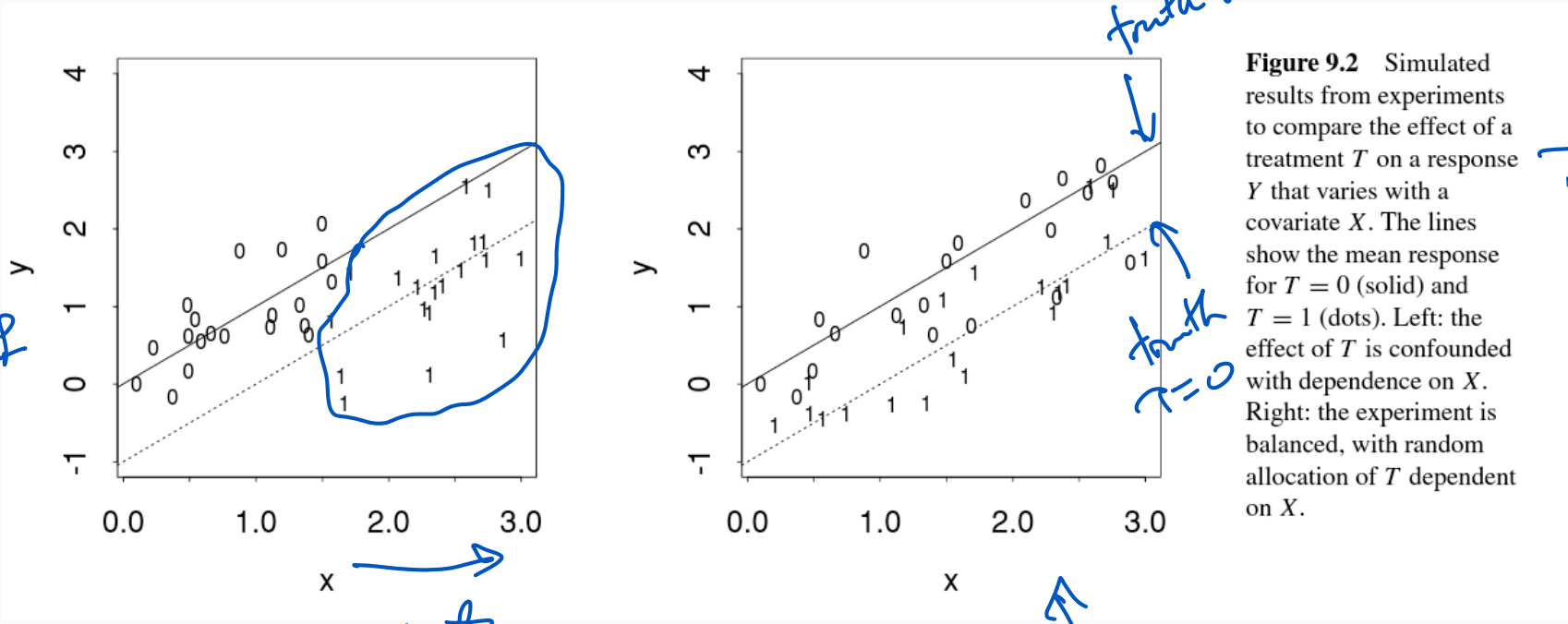


Figure 9.2 Simulated results from experiments to compare the effect of a treatment T on a response Y that varies with a covariate X. The lines show the mean response for $T = 0$ (solid) and $T = 1$ (dots). Left: the effect of T is confounded with dependence on X. Right: the experiment is balanced, with random allocation of T dependent on X.

resp

covariate

truth $T=1$

truth $T=0$

$T = 1, 0$

$T=1 \Rightarrow x$ large-ish

$T=0 \Rightarrow x$ small-ish

$T=1$ is randomized within levels of x

… Randomization and covariate

SM §9.1.1

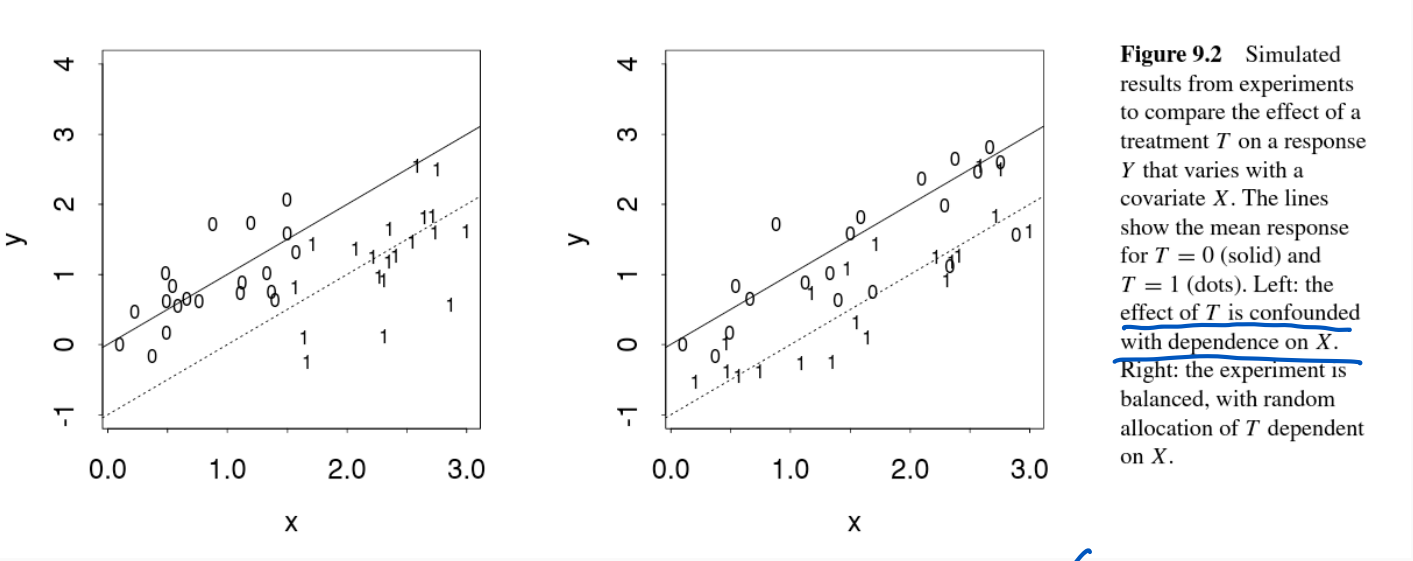


Figure 9.2 Simulated results from experiments to compare the effect of a treatment T on a response Y that varies with a covariate X. The lines show the mean response for $T = 0$ (solid) and $T = 1$ (dots). Left: the effect of T is confounded with dependence on X. Right: the experiment is balanced, with random allocation of T dependent on X.

Causal effect ≡ 1

Left: $\bar{y}_1 - \bar{y}_0 = 0.2 \pm 0.2$

Right: $\bar{y}_1 - \bar{y}_0 = -1.2 \pm 0.3$

benefit of rand = of tx

→ adjust for covariate: $y = \beta_0 + \beta_1 x + \delta t + \epsilon$ analysis of covariance

Left: $\hat{\delta} = -0.7 \pm 0.3$ Right: $\hat{\delta} = -1.25 \pm 0.16$

≈ 0.05

right randomized within pairs; matched on x

$\text{Corr}(\hat{\beta}_1, \hat{\delta})$

-.7 (nfbc)

What is a causal effect?

- philosophical: deterministic Internet Encyclopedia of Philosophy "*A* always follows *B*"
- statistical: probabilistic "In a large population, *T* is expected to cause an average change in *Y* ... "

What is a causal effect?

- philosophical: deterministic Internet Encyclopedia of Philosophy "*A* always follows *B*"
- statistical: probabilistic "In a large population, *T* is expected to cause an average change in *Y* ... "

- strong: there may be a well-understood mechanism that links *T* to *Y*

- weak: there has been observed a stable association between *T* and *Y* that cannot be otherwise explained
  no unmeasured confounding

- intermediate: an approach based on counter-factuals

Confounding variables

Bickel et al. 1975

	Men			Women		
Major	Number of applicants	Number admitted	Percent admitted	Number of applicants	Number admitted	Percent admitted
A	825	512	62	108	89	82
B	560	353	63	25	17	68
C	325	120	37	593	202	34
D	417	138	33	375	131	35
E	191	53	28	393	94	24
F	373	22	6	341	24	7
Total	2691	1198	44	1835	557	30

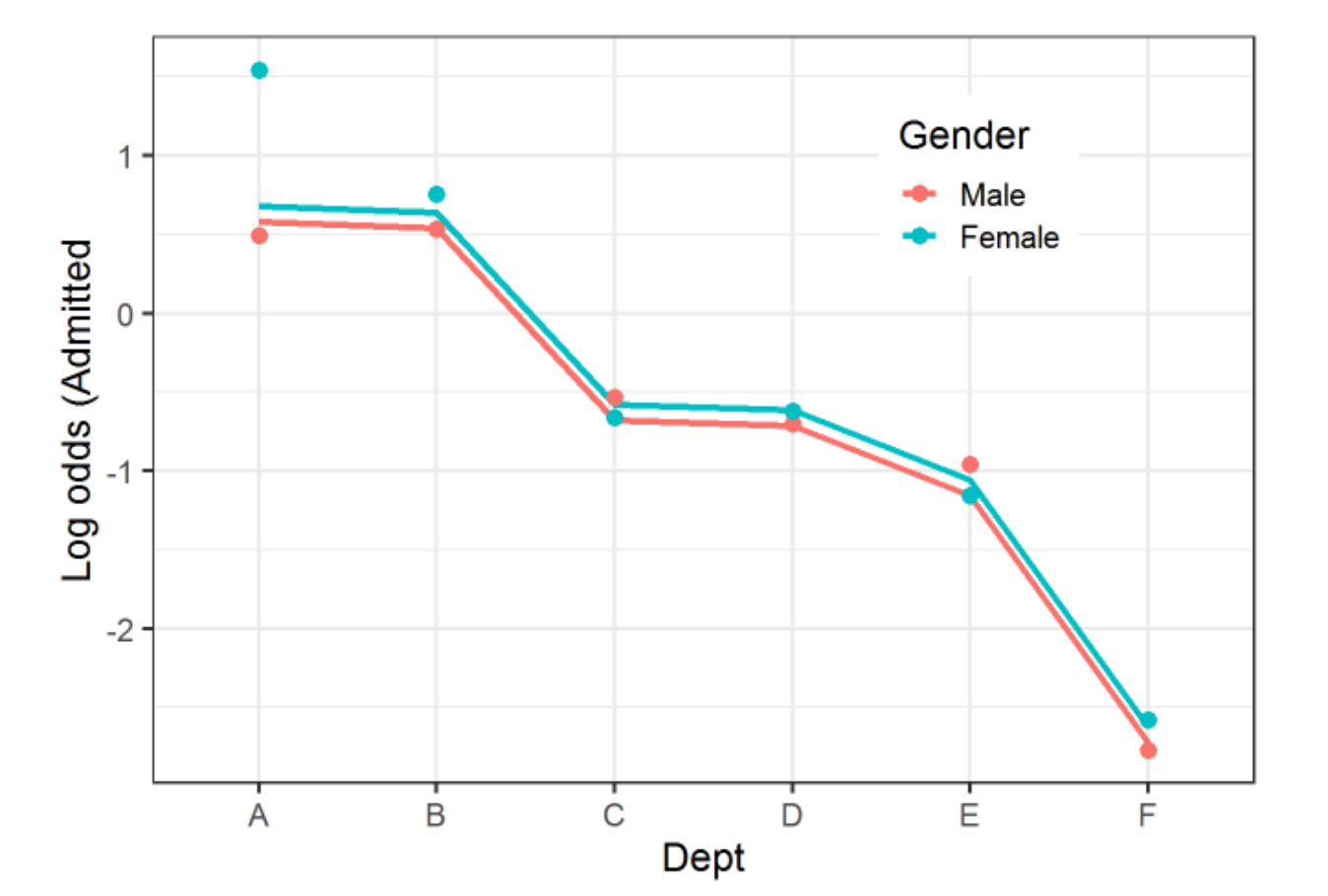
```
data(UCBAdmissions)
```

Bickel et al.

choice of major confounded w/ gender

ratio of proportions of # of applicants varies with gender

… Confounding variables



... Confounding variables

Radelet 1981

race of defendant	death penalty imposed	death penalty not imposed	percentage
white	19	141	11.88%
black	17	149	10.24%

... Confounding variables

Radelet 1981

race of defendant	death penalty imposed	death penalty not imposed	percentage
white	19	141	11.88%
black	17	149	10.24%

white victim	race of defendant	death penalty imposed	death penalty not imposed	percentage
	white	19	132	12.58%
	black	11	52	17.46%

black victim	race of defendant	death penalty imposed	death penalty not imposed	percentage
	white	0	9	0%
	black	6	97	5.83%

Age (years)	Smokers	Non-smokers
Overall	139/582 (24)	230/732 (31)
18–24	2/55 (4)	1/62 (2)
25–34	3/124 (2)	5/157 (3)
35–44	14/109 (13)	7/121 (6)
45–54	27/130 (21)	12/78 (15)
55–64	51/115 (44)	40/121 (33)
65–74	29/36 (81)	101/129 (78)
75+	13/13 (100)	64/64 (100)

Table 6.8 Twenty-year survival and smoking status for 1314 women (Appleton *et al.*, 1996). The smoker and non-smoker columns contain number dead/total (% dead).

smoking status recorded 1972–4; survival status recorded +20 years

reversal of effect

Causality and Counterfactuals

- T – binary treatment indicator
- Y – binary outcome
- "T causes Y" to be distinguished from "T is associated with Y"

AoS uses X for tmt

could be continuous

Causality and Counterfactuals

AoS Ch. 16; Cox & D 9.2

- T – binary treatment indicator — AoS uses X for tmt
- Y – binary outcome — could be continuous
- "T causes Y" to be distinguished from "T is associated with Y"

- introduce potential outcomes $Y(0), Y(1)$ — AoS C_0, C_1; SM R_0, R_1
- causal treatment effect $\theta = E\{Y(1)\} - E\{Y(0)\}$ — want to estimate this
- association $\alpha = E(Y \mid T = 1) - E(Y \mid T = 0)$ — have data to estimate α

- Consistency assumption: $Y = Y(t)$ — we can learn about potential outcome from observed values

Counterfactual: Examples

Potential outcomes C_0, C_1

X	Y	C_0	C_1
0	4	4	*
0	7	7	*
0	2	2	*
0	8	8	*
1	3	*	3
1	5	*	5
1	8	*	8
1	9	*	9

treatment X, response Y

Potential outcomes Y^0, Y^1

Table 2.1

	A	Y	Y^0	Y^1
Rheia	0	0	0	?
Kronos	0	1	1	?
Demeter	0	0	0	?
Hades	0	0	0	?
Hestia	1	0	?	0
Poseidon	1	0	?	0
Hera	1	0	?	0
Zeus	1	1	?	1
Artemis	0	1	1	?
Apollo	0	1	1	?
Leto	0	0	0	?
Ares	1	1	?	1
Athena	1	1	?	1
Hephaestus	1	1	?	1
Aphrodite	1	1	?	1
Cyclope	1	1	?	1
Persephone	1	1	?	1
Hermes	1	0	?	0
Hebe	1	0	?	0
Dionysus	1	0	?	0

treatment A
response Y

Counterfactuals

Principles of App Stat CD 9.2.1

"For most statistical purposes an explanatory variable C, considered for simplicity to have just two possible values, 0 and 1, has a causal impact on the response Y of a set of study individuals if, for each individual:

- conceptually at least, C might have taken either of its allowable values and thus been different from the value actually observed; and
- there is evidence that, at least in an aggregate sense, Y values are obtained from $C = 1$ that are systematically different from those that would have been obtained on the same individuals had $C = 0$, other things being equal

The definition of the word 'causal' thus involves the counterfactual notion that, for any individual, C might have been different from its measured value.

A central point in the definition of causality ... concerns the requirement

other things being equal"

Causal treatment effect

AoS Eq. (16.2)

$\theta = \mathrm{E}\{Y(1)\} - \mathrm{E}\{Y(0)\}$

risk difference; ratio; odds

also called "ATE" and "ACE": average treatment/causal effect

$\alpha = \mathrm{E}(Y \mid T = 1) - \mathrm{E}(Y \mid T = 0)$ this can be estimated from the data

Thm 16.1: $\alpha \neq \theta$

Thm 16.3: If T is is independent of $\{Y(0), Y(1)\}$, then $\alpha = \theta$

Need $\mathrm{pr}(T = 1) > 0, \mathrm{pr}(T = 0) > 0$

positivity

If treatment is randomly assigned, then $T \perp \{Y(0), Y(1)\}$

$\perp$ = independent

breaking the arrow from $U \to T$

Example 16.2

X	Y	C_0	C_1
0	0	0	0*
0	0	0	0*
0	0	0	0*
0	0	0	0*
1	1	1*	1
1	1	1*	1
1	1	1*	1
1	1	1*	1

$\theta = 0$; $\alpha = 1$

(C_0, C_1) not independent of X (T)

X	Y	C_0	C_1
0	0	0	0*
1	0	0	0*
1	0	0	0*
1	0	0	0*
1	1	1*	1
1	1	1*	1
1	1	1*	1
1	1	1*	1

$\theta = 0, \quad \alpha = 4/7 < 1$

thought experiment

Potential outcomes

Table 1.1

	$Y^{a=0}$	$Y^{a=1}$
Rheia	0	1
Kronos	1	0
Demeter	0	0
Hades	0	0
Hestia	0	0
Poseidon	1	0
Hera	0	0
Zeus	0	1
Artemis	1	1
Apollo	1	0
Leto	0	1
Ares	1	1
Athena	1	1
Hephaestus	0	1
Aphrodite	0	1
Cyclope	0	1
Persephone	1	1
Hermes	1	0
Hebe	1	0
Dionysus	1	0

$\theta = 0$

$\| \quad E(Y^1) - E(Y^0)$

Observed outcomes

Table 1.2

	A	Y
Rheia	0	0
Kronos	0	1
Demeter	0	0
Hades	0	0
Hestia	1	0
Poseidon	1	0
Hera	1	0
Zeus	1	1
Artemis	0	1
Apollo	0	1
Leto	0	0
Ares	1	1
Athena	1	1
Hephaestus	1	1
Aphrodite	1	1
Cyclope	1	1
Persephone	1	1
Hermes	1	0
Hebe	1	0
Dionysus	1	0

$\alpha = E(Y|A=1)$

$- E(Y|A=0)$

- typically have additional explanatory variables (covariates) X

AoS uses Z; HR use L

- causal effect of treatment when $X = x$

$$\theta(x) = \mathrm{E}\{Y(1) \mid X = x\} - \mathrm{E}\{Y(0) \mid X = x\}$$

- marginal causal effect

$$\theta = \mathrm{E}_X[\mathrm{E}\{Y(1) \mid X\} - \mathrm{E}\{Y(0) \mid X\}]$$

- association function

$$r(x) = \mathrm{E}(Y \mid T = 1, X = x) - \mathrm{E}(Y \mid T = 0, X = x)$$

- marginal association

$$\mathrm{E}_X\{r(X)\}$$

d ábar obs'd assoc'

Example

Table 2.2

	L	A	Y
Rheia	0	0	0
Kronos	0	0	1
Demeter	0	0	0
Hades	0	0	0
Hestia	0	1	0
Poseidon	0	1	0
Hera	0	1	0
Zeus	0	1	1
Artemis	1	0	1
Apollo	1	0	1
Leto	1	0	0
Ares	1	1	1
Athena	1	1	1
Hephaestus	1	1	1
Aphrodite	1	1	1
Cyclope	1	1	1
Persephone	1	1	1
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

cov. X T

$$\theta_{L=0} = E\{Y(1)\} - E\{Y(0)\}$$

$$= E(Y \mid T=1) - E(Y \mid T=0)$$

$$= \frac{1}{4} - \frac{1}{4} = 0$$

$$\theta_{L=1}$$

$$= \frac{6}{9} - \frac{2}{3} = 0$$

$L = 1$ $(X = 1)$ critical condition

$L = 0$ $(X = 0)$ stable condition

conditional randomization

$A \perp Y \mid L$ $(T \perp Y \mid X)$

No unmeasured confounding

AoS §16.3

- in observational studies treatment is not randomly assigned $\implies \theta(x) \neq r(x)$
- No unmeasured confounding:

$$\{Y(t); t \in \mathcal{T}\} \perp T \mid X$$

can learn about $Y(t)$ even if $T \neq t$ by using observed Y for 'similar' people from $T = t$ group

No unmeasured confounding

- in observational studies treatment is not randomly assigned $\implies \theta(x) \neq r(x)$
- No unmeasured confounding:

$$\{Y(t); t \in \mathcal{T}\} \perp T \mid X$$

can learn about $Y(t)$ even if $T \neq t$ by using observed Y for 'similar' people from $T = t$ group

- under the assumption of no unmeasured confounding, the marginal causal effect is

$$\mathrm{E}\{Y(t)\} = \int \mathrm{E}(Y \mid T = t, X = x) dF_X(x)$$

- this can be estimated by the association function

$$\widehat{\mathrm{E}}\{Y(t)\} = \frac{1}{n}\sum_{i=1}^{n} \hat{r}(t, X_i) = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 \bar{X}_n$$

causal reg function $\equiv$ adjusted treatment effect

"Bradford-Hill guidelines"

Evidence that an observed association is causal is strengthened if:

- the association is strong
- the association is found consistently over a number of independent studies
- the association is specific to the outcome studied
- the observation of a potential cause occurs earlier in time than the outcome
- there is a dose-response relationship
- there is subject-matter theory that makes a causal effect plausible
- the association is based on a suitable natural experiment

see also AoS §16.3

260 16. Causal Inference

	$Y=1$	$Y=0$	$Y=1$	$Y=0$
$X=1$	.1500	.2250	.1000	.0250
$X=0$	.0375	.0875	.2625	.1125
	$Z=1$ (men)		$Z=0$ (women)	

The marginal distribution for (X, Y) is

	$Y=1$	$Y=0$	
$X=1$	.25	.25	.50
$X=0$	.30	.20	.50
	.55	.45	1

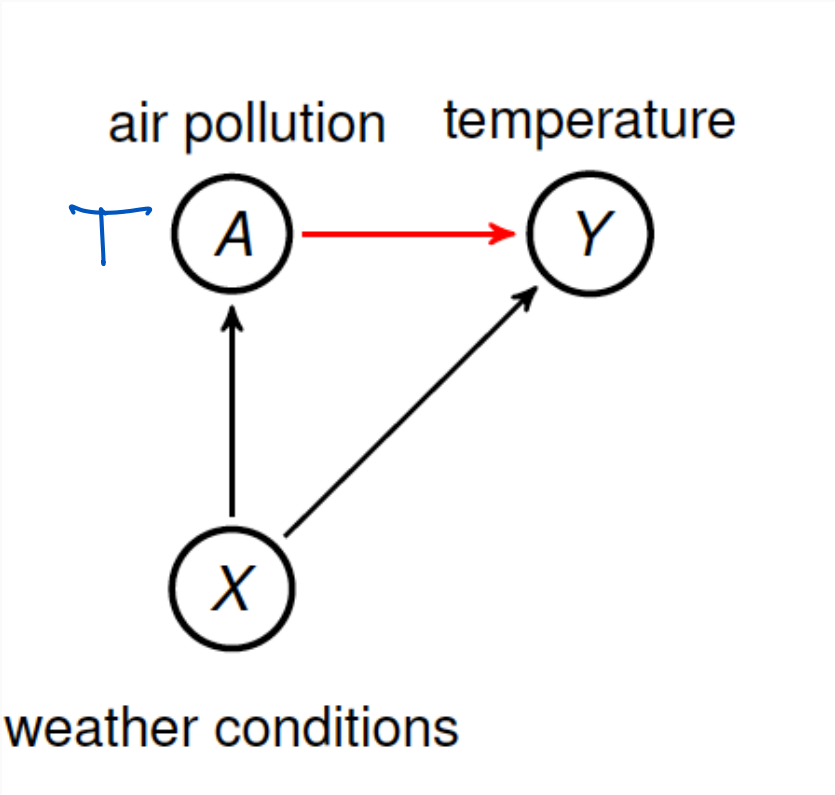
confusion of causal effect with association

From these tables we find that,

$$\begin{aligned}
\mathbb{P}(Y=1|X=1) - \mathbb{P}(Y=1|X=0) &= -0.1 \\
\mathbb{P}(Y=1|X=1, Z=1) - \mathbb{P}(Y=1|X=0, Z=1) &= 0.1 \\
\mathbb{P}(Y=1|X=1, Z=0) - \mathbb{P}(Y=1|X=0, Z=0) &= 0.1.
\end{aligned}$$

To summarize, we *seem* to have the following information:

Mathematical Statement	English Statement?



- assume no unmeasured confounding
- want to estimate
  $E\{Y(1) \mid X\} - E\{Y(0) \mid X\}$
  causal regression function
- or possibly $E_X[E\{Y(1) \mid X\} - E\{Y(0) \mid X)\}]$
  marginal effect of A
- regression model $A = T$

$$E(Y \mid X, A) = \beta_0 + \beta_1 A + \beta_2 X$$

- or something more complex

$$E(Y \mid X, A) = f(X, A)$$

neural network ...

- estimand average causal effect or average treatment effect (ATE)

$$E\{Y(1)\} - E\{Y(0)\}$$

estimand: something we estimate

- under the linear model $E(Y \mid X, A) = \beta_0 + \beta_1 A + \beta_2 X$, the ATE is β_1 if the linear model is correct

- estimated using

$$\widehat{\mathrm{E}}(Y(a)) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathrm{E}}(Y \mid A = a, X_i)$$

- recovers $\hat{\beta}_1$ in a linear model

- treat $Y_i(1)$ as missing data, if $A_i = 0$ (and v.v.)
- write

missing data weighted mean

$$\mathrm{E}(Y(a)) = \mathrm{E}\left\{\frac{1\{A=a\}Y}{\mathrm{pr}(A=a \mid x)}\right\}$$

- model $\mathrm{pr}(A=a \mid X)$, e.g. by logistic regression

$\beta_0 + \beta_1 A + \beta_2 X$

- doubly robust estimator

of $\mathrm{E}(Y(1))$

$$\hat{\mu}^{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{\widehat{\mathrm{pr}}(A=1 \mid X_i)} + \left\{1 - \frac{A_i}{\widehat{\mathrm{pr}}(A=1 \mid X_i)}\right\}\widehat{\mathrm{E}}(Y(1))$$

estimates $E\{Y(1)\}$

estimates $EY(1)$ if linear model

Link

$$\hat{\mathrm{E}}\{Y(1)\} = \frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{\pi(\boldsymbol{X}_i)}$$

$\pi(\underset{\sim}{X}_i) = P_a(A_i = 1 \mid \underset{\sim}{X}_i)$ propensity score

$A = 1$

$$\mathrm{E}\left\{\frac{AY}{\pi(\boldsymbol{X})}\right\} = \mathrm{E}\left[\mathrm{E}\left\{\frac{AY(1)}{\pi(\boldsymbol{X})} \mid Y(1), \boldsymbol{X}\right\}\right],$$

$$= \mathrm{E}\left[\frac{Y(1)}{\pi(\boldsymbol{X})}\mathrm{E}\{A \mid Y(1), \boldsymbol{X}\}\right]$$

$$= \mathrm{E}\left\{\frac{Y(1)}{\pi(\boldsymbol{X})}\mathrm{E}(A \mid \boldsymbol{X})\right\}$$

$$= \mathrm{E}\left\{\frac{Y(1)}{\pi(\boldsymbol{X})}\pi(\boldsymbol{X})\right\} = \mathrm{E}\{Y(1)\}$$

? $\xrightarrow{\text{n'tbc}}$ $A \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X$

A binary

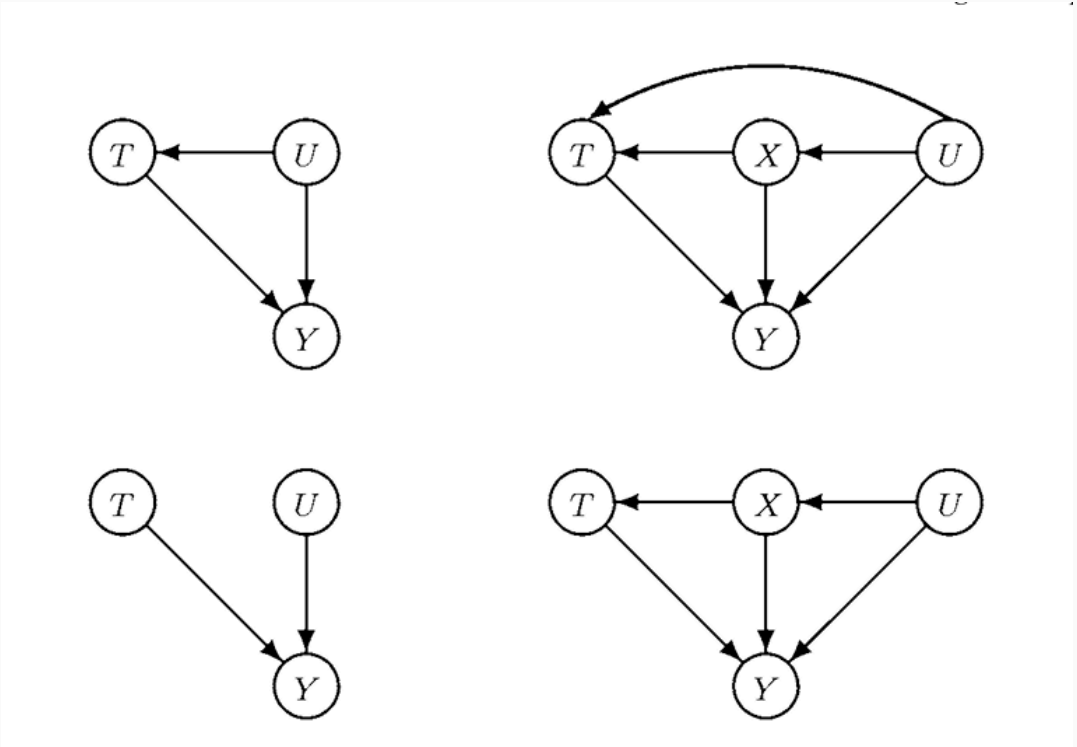
$A = 0$

$\mathrm{E}\{Y(0)\}$

Directed graphs and randomization

graphs can be useful for clarifying dependence relations among random variables

Fig 9.1 SM



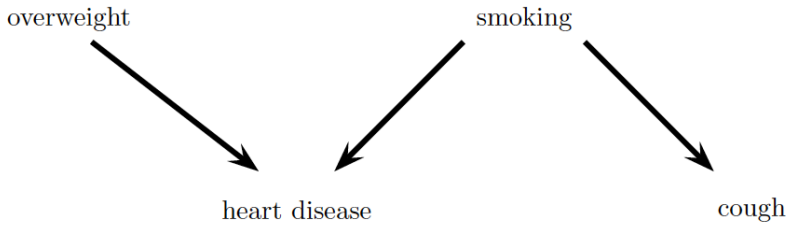


FIGURE 17.2. DAG for Example 17.4.

17.4 Example. Figure 17.2 shows a DAG with four variables. The probability function for this example factors as

$$\begin{aligned} f(\text{overweight}, \text{smoking}, \text{heart disease}, \text{cough}) \\ = \; & f(\text{overweight}) \times f(\text{smoking}) \\ \times \; & f(\text{heart disease} \,|\, \text{overweight}, \text{smoking}) \\ \times \; & f(\text{cough} \,|\, \text{smoking}). \quad \blacksquare \end{aligned}$$

- notation: $\mathcal{G}$ graph; $V = (X_1, \ldots, X_n)$ vertices
- The probability distribution on V is Markov if

π_i are parents of X_i

$$f(v) = \prod_{i=1}^{k} f(x_i \mid \pi_i)$$

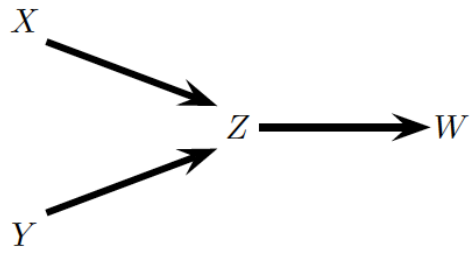


FIGURE 17.3. Another DAG.

Markov $\iff f(x, y, z, w) = f(x)f(y)f(z \mid x, y)f(w \mid z)$

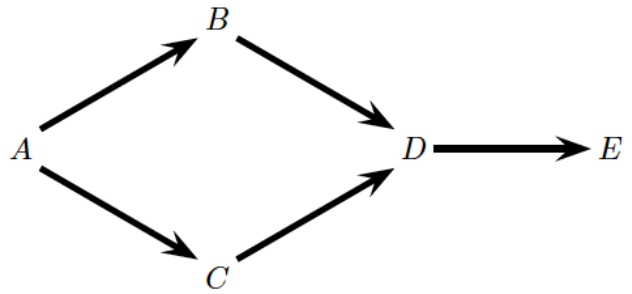


FIGURE 17.4. Yet another DAG.

$f(a, b, c, d, e) = f(a)f(b \mid a)f(c \mid a)f(d \mid b, c)f(e \mid d)$

If the probability distribution is Markov then

$\widetilde{W}$ other vars except parents and desc

$$W \perp \widetilde{W} \mid \pi_W$$

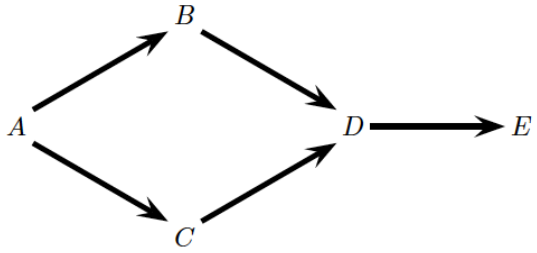


FIGURE 17.4. Yet another DAG.

$$f(a,b,c,d,e) = f(a)f(b \mid a)f(c \mid a)f(d \mid b,c)f(e \mid d)$$

$$D \perp A \mid \{B,C\}, \quad E \perp \{A,B,C\} \mid D, \quad B \perp C \mid A$$

deducing conditional independence relations from DAGs requires more definitions

colliders, d-separators, ...

distinguish $\mathrm{E}(Y \mid X = x)$ from $\mathrm{E}(Y \mid X := x)$ "$do(x)$"; set $X = x$

276 17. Directed Graphs and Conditional Independence

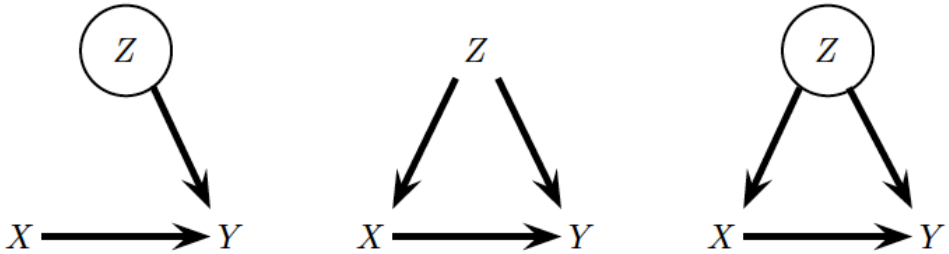


FIGURE 17.11. Randomized study; Observational study with measured confounders; Observational study with unmeasured confounders. The circled variables are unobserved.

randomized study observational study $\mathrm{E}(Y \mid x) = \int \mathrm{E}(Y \mid X, Z = z) dF_Z(z)$

$\mathrm{E}(Y \mid X := x) = \mathrm{E}(Y \mid X)$ $E(Y \mid X := x) = E(Y \mid x)$ unobserved confounder: $\theta \neq \alpha$

Journal of Data Science 24 (1), 53–85
January 2026
DOI: 10.6339/25-JDS1211
Data Science Reviews

Causal Inference: A Tale of Three Frameworks

Linbo Wang[1,*], Thomas S. Richardson[2], and James M. Robins[3]

[1]*Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada*
[2]*Department of Statistics, University of Washington, Seattle, WA, U.S.A.*
[3]*Harvard T.H. Chan School of Public Health, Boston, MA, U.S.A.*

Abstract

Causal inference is a central goal across many scientific disciplines. Over the past several decades, three major frameworks have emerged to formalize causal questions and guide their analysis: the potential outcomes framework, structural equation models, and directed acyclic graphs. Although these frameworks differ in language, assumptions, and philosophical orientation, they often lead to compatible or complementary insights. This paper provides a comparative introduction to the three frameworks, clarifying their connections, highlighting their distinct strengths and limitations, and illustrating how they can be used together in practice. The discussion is aimed at researchers and graduate students with some background in statistics or causal inference who are seeking a conceptual foundation for applying causal methods across a range of substantive domains.

Keywords *directed acyclic graphs; identification; potential outcomes; structural equation models; SWIGs*

1 Introduction

Causal inference is the science of understanding the consequences of interventions, requiring assumptions that extend beyond those needed for purely associational analysis. Its importance has grown rapidly in the era of machine learning and artificial intelligence, where the ability to draw reliable causal conclusions is central to building systems that are not only predictive but also trustworthy, transparent, and robust to distributional shifts (Peters et al., 2016; Wachter et al., 2017; Pearl, 2019; Arjovsky et al., 2020; Bühlmann, 2020; Tjoa and Guan, 2021; Schölkopf, 2022; Jiao et al., 2024). Over the past decades, three foundational frameworks have emerged to formalize causal reasoning: the potential outcomes framework, nonparametric structural equation models (NPSEMs), and directed acyclic graphs (DAGs). Each framework carries its own formal machinery, conceptual underpinnings, and historical roots. Although they originated in distinct disciplinary traditions, they are now increasingly recognized as complementary, and in many cases translatable into one another.

A substantial literature surveys causal inference from within a single framework or with an emphasis on identification and estimation approaches (e.g. Imbens and Wooldridge, 2009;

SPECIAL ARTICLE

Air Pollution and Mortality at the Intersection of Race and Social Class

Kevin P. Josey, Ph.D., Scott W. Delaney, Sc.D., J.D., Xiao Wu, Ph.D., Rachel C. Nethery, Ph.D., Priyanka DeSouza, Ph.D., Danielle Braun, Ph.D., and Francesca Dominici, Ph.D.

ABSTRACT

BACKGROUND

Black Americans are exposed to higher annual levels of air pollution containing fine particulate matter (particles with an aerodynamic diameter of ≤2.5 μm [$PM_{2.5}$]) than White Americans and may be more susceptible to its health effects. Low-income Americans may also be more susceptible to $PM_{2.5}$ pollution than high-

From the Departments of Biostatistics (K.P.J., R.C.N., D.B., F.D.) and Environmental Health (S.W.D.), Harvard T.H. Chan School of Public Health, Boston; the Department of Biostatistics, Mailman School of Public Health, Columbia University, New York (X.W.); and the Department of Urban and Regional Planning,

for $PM_{2.5}$ exposure and mortality among marginalized subpopulations categorized