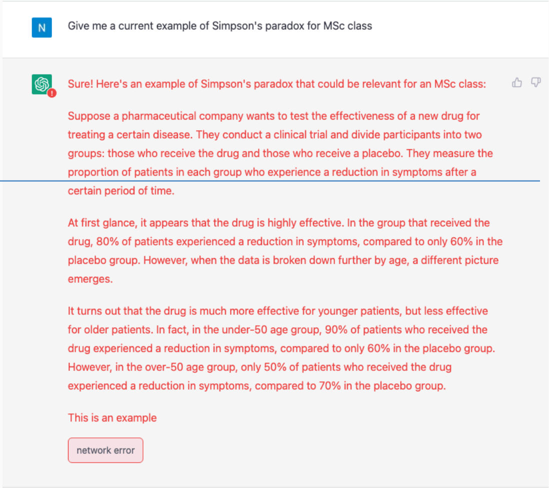
Mathematical Statistics II

STA2212H S LEC0101

Week 10

March 21 2023



Simpson's Paradox

ChatGPT

N

Give me a current example of Simpson's paradox for MSc class

Sure! Here's an example of Simpson's paradox that could be relevant for an MSc class:

Suppose a pharmaceutical company wants to test the effectiveness of a new drug for treating a certain disease. They conduct a clinical trial and divide participants into two groups: those who receive the drug and those who receive a placebo. They measure the proportion of patients in each group who experience a reduction in symptoms after a certain period of time.

At first glance, it appears that the drug is highly effective. In the group that received the drug, 80% of patients experienced a reduction in symptoms, compared to only 60% in the placebo group. However, when the data is broken down further by age, a different picture emerges.

It turns out that the drug is much more effective for younger patients, but less effective for older patients. In fact, in the under-50 age group, 90% of patients who received the drug experienced a reduction in symptoms, compared to only 60% in the placebo group. However, in the over-50 age group, only 50% of patients who received the drug experienced a reduction in symptoms, compared to 70% in the placebo group.

This is an example

network error

Simpson's Paradox

ChatGPT:Take 2

Give me a current example of Simpson's paradox for MSc class

2 / 2

Sure! Here's a current example of Simpson's paradox:

Suppose a university wants to analyze the effectiveness of their graduate program in terms of job placement rates. They divide their graduate students into two groups: domestic students and international students. They find that in the domestic student group, 80% of students who graduated with a master's degree found a job in their field within 6 months of graduating. In the international student group, only 60% of students who graduated with a master's degree found a job in their field within 6 months of graduating.

However, when the university looks at the overall data, they find that 75% of all students who graduated with a master's degree found a job in their field within 6 months of graduating. This seems to suggest that the graduate program is successful overall.

The paradox arises because although the domestic student group has a higher job placement rate than the international student group, there are actually more international students in the program than domestic students. Specifically, the domestic student group has 100 students, of which 80 found a job, while the international student group has 300 students, of which 180 found a job. Thus, the overall job placement rate is heavily influenced by the larger number of international students who graduated and found jobs, despite the fact that their job placement rate is lower than that of domestic students.

This is an example of Simpson's paradox because the overall conclusion (that the

Simpson's Paradox

ChatGPT:Take 3

N

Give me an example from a recent news article of Simpson's paradox

‹ 2 / 2 ›

One example of Simpson's paradox in a recent news article involves the gender pay gap in the tech industry.

In 2020, a study by Hired Inc. found that on average, women in the tech industry earn 3% less than men with the same job title and level of experience. However, when the data was broken down by individual job titles, it was found that in 63% of the positions studied, women earned more than men.

This is a clear example of Simpson's paradox, where a trend appears in one direction when the data is looked at as a whole, but is reversed when the data is broken down into smaller subgroups. In this case, the overall gender pay gap in the tech industry was 3%, but when the data was separated by job title, it was revealed that women actually earned more than men in many positions. This suggests that there may be other factors at play, such as the types of jobs women and men are choosing within the industry.

Sure, I can demonstrate Simpson's paradox using the Berkeley college admissions data in R.

First, we can load the data using the following code:

```
{r}
berkeley <- read.csv("https://raw.githubusercontent.com/Opensourcefordatasci
```

The data contains information on the number of male and female applicants and the number of admissions for each gender in each of the six departments at UC Berkeley in 1973. We can summarize the data using the `summarize` function from the `dplyr` package:

```
{r}
library(dplyr)

berkeley_summary <- berkeley %>%
  group_by(Gender, Dept) %>%
  summarize(Applicants = sum(Applicants),
            Admitted = sum(Admitted))

berkeley_summary
```

This will give us a summary of the data grouped by gender and department, showing the total number of applicants and admitted students for each combination:

for MSc class

it could be relevant for an MSc class:

the effectiveness of a new drug for
trial and divide participants into two
o receive a placebo. They measure the
ce a reduction in symptoms after a

fective. In the group that received the
symptoms, compared to only 60% in the
down further by age, a different picture

or younger patients, but less effective
up, 90% of patients who received the
ared to only 60% in the placebo group.
atients who received the drug
to 70% in the placebo group.

Today

1. Recap
2. Introduction to causal inference

Upcoming

- March 23 2.30 – 3.30 Zoom
  "Valid statistical inference with privacy constraints"
  Aleksandra Slavković, Penn State

Recap

- BH-q procedure: $i_{max} = \arg\max\{i : p_{(i)} \leq \frac{i}{m} q\}$
- BH_q : reject H_{0i} for $i \leq i_{max}$
- Theorem: If the p-values corresponding to valid null hypotheses are independent of each other, then

$m_0 = \#\text{true } H_0$

$$FDR(BH_q) = \frac{m_0}{m} q \leq q$$

		H_0 not rejected	H_0 rejected	
truth	H_0 true	U	V	m_0
	H_1 true	T	S	m_1
		$m - R$	R	m

- Proof: ... $\mathrm{E}\{A(s) \mid A(t)\} = A(t), \quad s < t;$

$$A(t) = V(t)/t, \quad V(t) = \#\{p_i \leq t; H_{0i}\}$$

- $V(s) \mid V(t) \sim Binom(V(t), \frac{s}{t})$, under H_0

... Recap

- Robust linear regression $y_i = x_i^T\beta + \epsilon_i$

$$\min_{\beta}\sum_{i=1}^{n}\rho(y_i - x_i^T\beta), \quad \leftrightarrow \quad \sum_{i=1}^{n}\psi(y_i - x_i^T\beta) = 0$$

$\rho(\cdot)$, to be determined; $\psi(\cdot) = \rho'(\cdot)$

-

$$\hat{\beta} \dot{\sim} N_p\left(\beta, \sigma^2(X^TX)^{-1}\frac{\mathrm{E}\{\psi^2(\epsilon_i)\}}{\mathrm{E}^2\{\psi'(\epsilon)\}}\right)$$

correction from last week

- this is an example of M-estimation
- An M-estimate is a solution of the estimating equation

Assume $\mathrm{E}\{g(Y;\theta)\} = 0$

$$\sum_{i=1}^{n} g(Y_i;\tilde{\theta}_g) = 0$$

- under regularity conditions $\tilde{\theta}_g \dot{\sim} N\{\theta, G^{-1}(\theta)\}$, $\qquad G(\theta) = J(\theta)I^{-1}(\theta)J(\theta)$

... Recap

- p^* approximation

$j(\theta) = J(\theta) = -\ell''(\theta)$

$$f(\hat\theta;\theta) \doteq c|j(\hat\theta)|^{1/2}\exp\{\ell(\theta)-\ell(\hat\theta)\}$$

- r^* approximation

$$F(\hat\theta;\theta) \doteq \Phi(r^*), \quad r^*(\theta) = r(\theta) + \frac{1}{r(\theta)}\log\left\{\frac{q(\theta)}{r(\theta)}\right\}$$

- r is signed-square root of likelihood ratio statistic
- in exponential families, q is Wald statistic
- in location families, q is score statistic
- in Bayesian posterior, q is score statistic $\times$ prior ratio
- derived using saddlepoint approximation

or Laplace approximation, depending on context

Causality

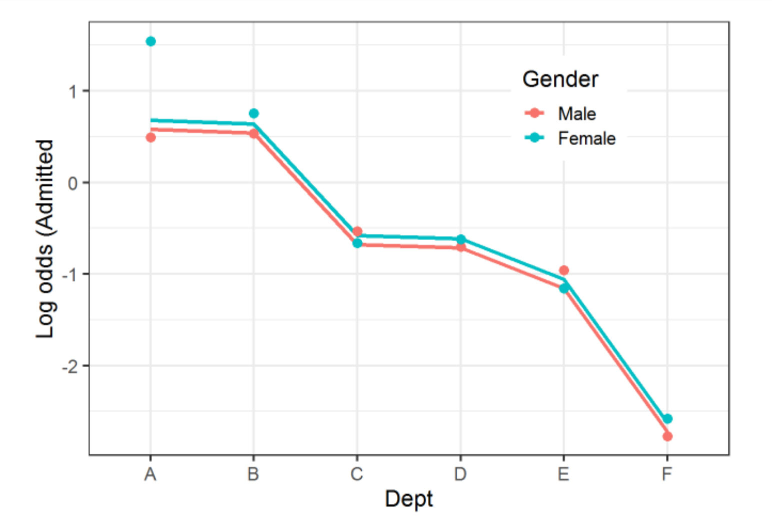
- randomization; confounding; observational studies; experiments;
  “correlation is not causation”, Simpson’s ‘paradox’
- counterfactuals; average treatment effect; conditional average treatment effect; ...
- graphical models; directed acyclic graphs; causal graphs; Markov assumptions...
- The Book

Confounding variables

	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)
```

… Confounding variables



… Confounding variables

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%

Simpson's 'paradox'

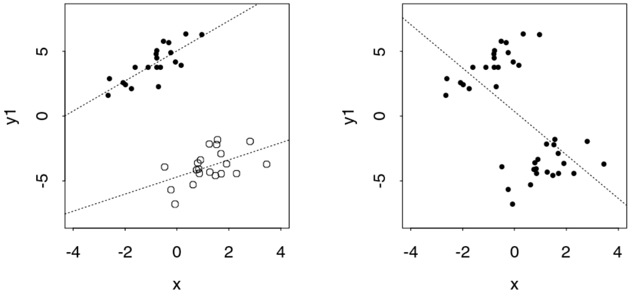


Figure 6.9 Artificial data illustrating Simpson's paradox. The left panel shows how Y_1 depends on x for each of two values of Y_2, with observations with $y_2 = 0$ shown by blobs and those with $y_2 = 1$ shown as circles. The lines are from separate straight-line regression fits of y_1 on x for each value of y_2 and show positive association. The right panel shows the fit to the data ignoring Y_2, for which the association is negative.

E. H. Simpson called attention to this effect in 1951, although it was known to G. U. Yule almost 50 years earlier.

decreases with x, that is, $E(Y_1 \mid x)$ has negative slope as a function of x. This effect — *Simpson's paradox* — is due to the fact that marginalization of the joint distribution of (Y_1, Y_2) over Y_2 has reversed the sign of the association between Y_1 and x. Here a plot at once reveals that it is a bad idea to fit a common line to both groups, but the

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

"treatment"

could be continuous

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

- introduce potential outcomes C_0, C_1

$$Y = \begin{cases} C_0 & \text{if } X = 0 \\ C_1 & \text{if } X = 1 \end{cases}$$

- equivalently $Y = C_X$ or $Y = C_0(1 - X) + C_1 X$ consistency equation

- causal treatment effect $\theta = E(C_1) - E(C_0)$ want to estimate this
- association $\alpha = E(Y \mid X = 1) - E(Y \mid X = 0)$ have data to estimate α

- if $(C_0, C_1) \perp X$, then $\theta = \alpha$ randomization ensures $\perp$

Counterfactual: Examples

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

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

Causal Effect and Association

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

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

Counterfactuals

"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 fro $C = 1$ 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

$$\theta = \mathrm{E}(C_1) - \mathrm{E}(C_0)$$

risk difference; ratio; odds

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

If X is randomly assigned, then (C_0, C_1) is independent of X

$$\theta = \mathrm{E}(C_1) - \mathrm{E}(C_0) =$$

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

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

thought experiment

$\alpha = 4/7 < 1$

Three types of causal statements

1. A well-understood evidence-based mechanism, or set of mechanisms, that links a cause to its effect

2. two phenomena are linked by a stable association, whose direction is established and which cannot be explained by mutual dependence on some other allowable variable

3. observed association may be linked to causal effect via counterfactuals if $(C_0, C_0) \perp X$

   not usually testable

Conditional and marginal effects

- typically have additional explanatory variables (covariates) Z
- causal effect of treatment when $Z = z$

$$\theta_z = \mathrm{E}(C_1 \mid Z = z) - \mathrm{E}(C_0 \mid Z = z)$$

- marginal causal effect

$$\theta = \mathrm{E}_Z\{\mathrm{E}(C_1 \mid Z) - \mathrm{E}(C_0 \mid Z)\}$$

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

$\theta_{L=0}$

$\theta_{L=1}$

Causal regression function

- continuous “treatment” variable $X \in \mathbb{R}$
- counterfactual outcome $(C_0, C_1) \rightarrow$ counterfactual function $C(x)$
- observed response $Y = C(X)$ consistency
- causal regression function $\theta(x) = \mathrm{E}\{C(x)\}$
- association regression function $r(x) = \mathrm{E}(Y \mid X)$

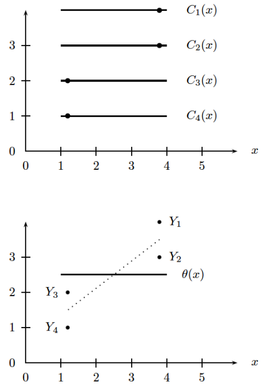


FIGURE 16.2. The top plot shows the counterfactual function $C(x)$ for four subjects. The dots represent their X values. Since $C_i(x)$ is constant over x for all i, there is no causal effect. Changing the dose will not change anyone's outcome. The lower plot shows the causal regression function $\theta(x) = (C_1(x) + C_2(x) + C_3(x) + C_4(x))/4$. The four dots represent the observed data points $Y_1 = C_1(X_1)$, $Y_2 = C_2(X_2)$, $Y_3 = C_3(X_3)$, $Y_4 = C_4(X_4)$. The dotted line represents the regression $r(x) = \mathbb{E}(Y|X = x)$. There is no causal effect since $C_i(x)$ is constant for all i. But there is an association since the regression curve $r(x)$ is not constant.

No unmeasured confounding

- in observational studies treatment is not randomly assigned $\implies \theta(x) \neq r(x)$
- group subjects based on additional confounding variables
- No unmeasured confounding:

$$\{C(x); x \in \mathcal{X}\} \perp X \mid Z$$

- under the assumption of no unmeasured confounding, the causal regression function

typo in (16.7)

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

can be estimated by the association function

$$\hat{\theta}(x) = \frac{1}{n}\sum_{i=1}^{n} \hat{r}(x, Z_i) \qquad = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 \bar{Z}_n$$

causal reg function $\equiv$ adjusted treatment effect

No unmeasured confounding

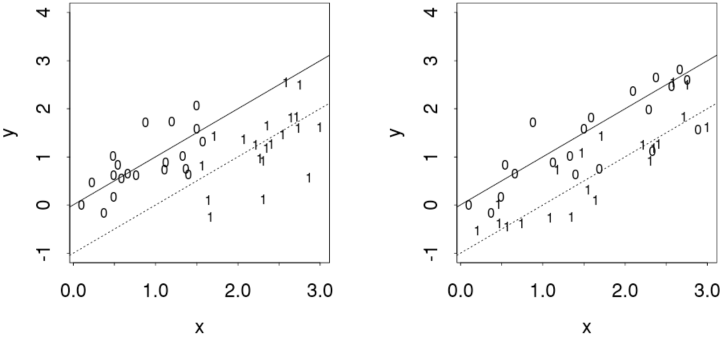


Figure 9.2 Simulated results from experiments to compare the effect of a treatment T on a response Y that varies with a covariate X. The lines show the mean response for $T = 0$ (solid) and $T = 1$ (dots). Left: the effect of T is confounded with dependence on X. Right: the experiment is balanced, with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

“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

Simpson's paradox revisited

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

confusion of causal effect with association

The marginal distribution for (X, Y) is

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

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?

Directed graphs

- graphs can be useful for clarifying dependence relations among random variables

SM Markov random fields

- a Directed Acyclic Graph has random variables on the vertices and edges joining random variables

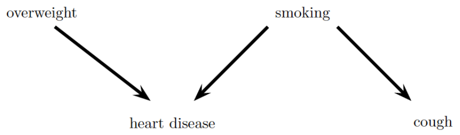


FIGURE 17.2. DAG for Example 17.4.

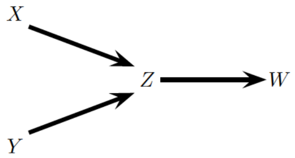


FIGURE 17.3. Another DAG.

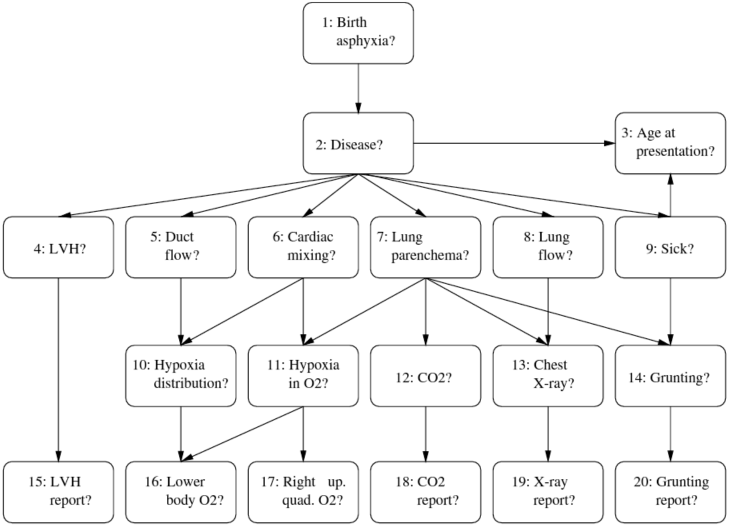


Figure 6.7 Directed acyclic graph representing the incidence and presentation of six possible diseases that would lead to a 'blue' baby (Spiegelhalter *et al.*, 1993). LVH means left ventricular hypertrophy.

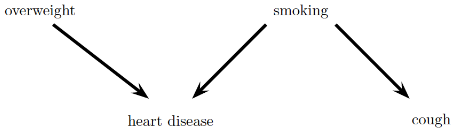


FIGURE 17.2. DAG for Example 17.4.

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

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

17.5 Example. For the DAG in Figure 17.3, $\mathbb{P} \in M(\mathcal{G})$ if and only if its probability function f has the form

$$f(x, y, z, w) = f(x)f(y)f(z \mid x, y)f(w \mid z). \quad \blacksquare$$

Probability on DAGs

- 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

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

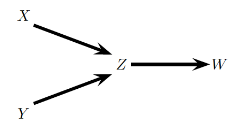


FIGURE 17.3. Another DAG.

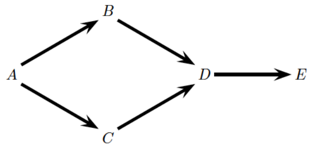


FIGURE 17.4. Yet another DAG.

DAGs and independence

If the probability distribution is Markov then

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

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

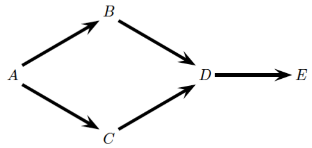


FIGURE 17.4. Yet another DAG.

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

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

deducing conditional independence relations from DAGs requires more definitions

colliders, d-separators, ...

- variables at parent nodes are potential causes for responses at child nodes
- directed graphs often helpful adjunct to modelling with baseline variables, intermediate responses, and outcome variables of interest
- much hard to study the full joint distribution than the usual supervised learning approaches
- DAGs can be used to represent confounders

276 17. Directed Graphs and Conditional Independence

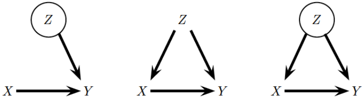


FIGURE 17.11. Randomized study; Observational study with measured confounders; Observational study with unmeasured confounders. The circled variables are unobserved.

DAGs and confounders

AoS 17.8

276 17. Directed Graphs and Conditional Independence

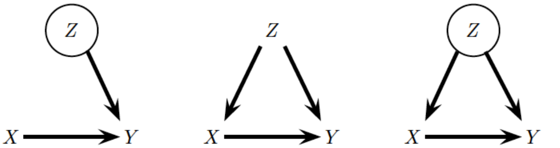


FIGURE 17.11. Randomized study; Observational study with measured confounders; Observational study with unmeasured confounders. The circled variables are unobserved.

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

unobserved confounder: $\theta \neq \alpha$

$\mathrm{E}(Y \mid X := x); \mathrm{E}(Y \mid X = x)$ conditioning by intervention